Title
Atopy-bound Infant Microbiomes

ISBN
9798189294655

Author
Steininger, Holly Michelle

Publication Date
2026-09-03

Peer reviewed|Thesis/dissertation

Atopy-bound Infant Microbiomes

by
Holly Michelle Steininger

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biomedical Sciences

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Joseph Bondy-Denomy
D23522498F4247C...

Joseph Bondy-Denomy
Chair

DocuSigned by:
Dr. Sue Lynch
93D34C5...

Dr. Sue Lynch

DocuSigned by:
Spencer Diamond
AA40D4E0...

Spencer Diamond

DocuSigned by:
Tiffany Scharschmidt
EDFBCD1914E943B...

Tiffany Scharschmidt

Committee Members

Dedication

This thesis is dedicated to my family and the memory of Charles "Chuck" Kwok Fai Chan. My family for loving and believing in me no matter what dream I wanted to pursue, and Chuck for bringing my scientific dreams to reality.

Acknowledgements

Firstly, none of this work would have ever been possible without my advisor **Susan V. Lynch**. She has taught me about far more than the microbiome, and her lab was my home during my time at UCSF. It was her devotion to pursing research that is relevant to humans that attracted me to her lab, and it is that principle which I strive to center in my career.

Thank you to **Spencer Diamond** for taking on the role of co-advisor. I am proud that you are proud that I am your first graduate student. These manuscripts would never have happened without the time you dedicated to teaching me about computational work, manuscript writing, and statistics. One time you told me that as soon as I learn how to *really* use linear models I would never go back, and you were completely right as you often are.

Thank you to my committee members, for helping me to navigate my PhD during the stressful times, and the scientific advice **Tiffany Scharschmidt** helped me to view my work as a long career rather than a limited PhD, and **Joe Bondy-Denomy** as my chair was kind to me. Joe taught me that sometimes in manuscript you can't include all the data and encouraged me to share my excitement and passion when presenting helping me to bring my authentic self into my work.

Thank you to the Chan Lab who taught me scientific rigor and saw the passion I had before I knew I had it: **Lauren Koepke**, **Liming Zhao**, **Malachia Hoover**, **Mathew Murphy**, **Yuting Wang**, and **Thomas Ambrosi**. I did not know how special it was to learn mouse fracture, kidney capsule, destabilization of the meniscus, or microfracture surgeries from plastic surgeons who trained all around the world. To the leader and PI of the Chan lab, **Chuck**, your passing has been an immeasurable loss. You taught me to be a fearless scientist, and it is a tragedy that you are

not around to spread this scientific courage and fearlessness to the next generation. I will strive to keep this spirit alive.

Thank you to all the members of the Lynch lab for adding to my value as a person. To my fellow dungeon dwellers **Margôt Bacino**, **Chirstopher Bravo**, and **Margareta Mayer**, thank you for being there and reminding me I am not alone. Thank you to the postdocs who showed me other ways of living and thinking: **Jay Moltazau-Anderson**, **Elad Deiss-Yehiely,** and **Sophia Levan**. Thank you to all the other members of the Lynch Lab for bringing their individual strengths to our group: **Din Lin**, **Faith Vundla**, **Rebecca Knoll**, **Kameron Sugino**, and **Punit Sundaramurthy**.

Thank you to the wider BCMM community for the technical and mental support, specifically, **Abdur Rahim Kahn, Lu Yang**, **Kristin Dolan**, **Peter Turnbaugh**, and **Soumaya Zlitni**.

Thank you to those at the IGI who have supported me: **Lindsey Pieper-Zapien, Carlos Iglessia-Aguilar,** and the entirety of the **TED Audacious** project.

Thank you to my mentees, for making me a better scientist and teaching me that my compassion and sensitivity to others is a strength and enduring my insistent documentation practices. In particular, **Gabe Braga**, who has transitioned seamlessly from mentee to friend.

Thank you to my friends who helped me live life beyond the lab so that I wasn't consumed by my own obsessions. **Caroline Whitty** and **Anita Qualls** have been with me through every step of my PhD, to the other side of the world, and have become my favorite people to celebrate any holiday with. There *will be* a menu and a themed cocktail. You both inspire me. Thank you to **Ivy Ma** and **Arianna Haft** for keeping me connected to the worlds outside of academia and being the kind of friends who share girlhood by truly witnessing each

other. I think there is an idea that life stops during a PhD, and my best friends **Austin** and **Stephanie Philp** have made sure mine hasn't by sharing so many adventures these past 5 years. We've backpacked across the lakes of the Desolation Wilderness, watched an eclipse in Missouri, and you both brought my nephew, **Kenny**, into the world. I am so grateful to have people I love and who love me enough to share so many milestones and experiences together.

Thank you to my family for being proud of me just for being me. I have come to appreciate how great of a gift this is. I am, and have always been, proud to be a Steininger because of the legacy and love that comes with it. In this document I use my siblings' personal struggles with allergic diseases to paint a picture, and they all unanimously gave me approval prior to my exit talk. One of them even responded to my text request for his permission with "Holly I don't give a rip if you have to cut me open in front of the whole panel to finish your PhD lmao" when I asked. This is a level of support and grounding I think only my four brothers and mother can provide. To my Grandma and Grandpa, **Carole and Jim Schenold**. Grandma, you are my favorite person in the whole world, and I hope to embody your intelligent, kind, and humorous spirit. Grandpa, your love for Grandma has taught me to not settle for nothing less.

Lastly, thank you to my toughest scientific critic and my number one scientific supporter, my fiancée **Vanya Zheludev,** who has lovingly walked with me through every step of this journey. For reading my grad school application letters. For walking through my qualification exam chalk talk for hours with me. For making me confident that I can handle any question, because I am sure you have asked me a harder version of it already. For ubering up from Stanford to take me to pizza when I facetimed you crying on the lab floor at 6 PM on a Friday night (my pooled library tube flooded with ice water when I was adding the final reagent

just prior to loading the NextSeq). For reading every single bioRxiv ever posted and sharing anything relevant to my interests and work. I used to try to avoid mentioning that I learned something from my partner and would try to hide it because I feared it would undermine my position as a woman of science. Now everyone who knows me professionally, knows your name because your mark on my work is undeniable. You will have to wait for our wedding vows to hear the even greater impression on my non-scientific life you have made

Chapter 1 Acknowledgements

This is original work written for this dissertation by Holly M. Steininger and reviewed by Susan V. Lynch and Spencer Diamond.

Chapter 2 Acknowledgements

Chapter 2 is reprinted, with reformatting, from the following manuscript, which has been published to the *Journal of Allergy and Clinical Immunology*: Steininger HM†, McCauley KE†, Braga GE, Morin A, Lee KE, Visness CM, Fadrosh DW, Dapas M, O'Connor GT, Bacharier LB, Rivera-Spoljaric K, Kattan M, Lemanske RF Jr, Zoratti EM, Martinez FD, Gold DR, Hartert TV, Johnson CC, Miller RL, Seroogy CM, Wood RA, Togias A, Jackson DJ, Gern JE, Ober C, Lynch SV, for the CADRE investigators. "Upper Airway Microbiome Interacts with *GSDMB* and *ORMDL3* Asthma Risk SNPs to Influence Early-Life Wheeze Risk." † denotes co-first authorship.

Funding: American Lung Association (LTR DTD 4/12/16), National Heart, Lung, and Blood Institute (NHLBI; HL130045, HL132523, HL139054, HL147016), National Institute of Allergy and Infectious Disease (NIAID; HHSN272201000052I, P01 AI148104, R01 AI135108, R01 AI146131, R01 AI149754, R21 AI176305, 1U01AI178772, U19AI104317, U19 AI62310, U24AI179612,

UM1AI160040, UM1 AI114271, UM2 AI117870), Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHHD; HD10418), National Institute of Environmental Health Sciences (NIEHS; P30-ES000002), National Institutes of Health (NIH) Office of the Director (1UL1TR001430, UG3/UH3OD023282, UG3OD023290, UG3OD035509), OM PHARMA (UANLID46205). HMS and SVL are supported in part by Lyda Hill Philanthropies, Acton Family Giving, the Valhalla Foundation, Hastings/Quillin Fund – an advised fund of the Silicon Valley Community Foundation, the CH Foundation, Laura and Gary Lauder and Family, the Sea Grape Foundation, the Emerson Collective, Mike Schroepfer and Erin Hoffman Family Fund – an advised fund of Silicon Valley Community Foundation, the Anne Wojcicki Foundation through The Audacious Project at the Innovative Genomics Institute.

We thank all participants in the URECA and COAST birth cohorts for their biospecimens which made this study possible. This study was also supported by funding from the National Institutes of Health (NIH) Office of the Director (PO515267) and National Institute of Allergy and Infectious Disease (NIAID; UM1AI160040). Co-authorship by A. Togias on the underlying publication does not necessarily constitute endorsement by the NIAID, the NIH, or any other agency of the United States government.

Conflict of interest: LBB is a member of the GINA Science Committee; reports grants from NIH/NIAID/NHLBI, personal fees from GlaxoSmithKline, Genentech/Novartis, Merck, Teva, Boehringer Ingelheim, AstraZeneca, Avillion, WebMD/Medscape, Sanofi/Regeneron, Vectura, Circassia, OM Pharma, Recludix and Kinaset, for DSMB from AstraZeneca, DBV Technologies, Aravax and Vertex; and royalties from Elsevier outside the submitted work. FDM reports consultancy fees from OM PHARMA. TVH reports grants from NIH and the World Health

Organization during the conduct of the study and personal fees from the American Thoracic Society, the American Academy of Allergy, Asthma and Immunology; the Parker B. Francis Foundation Scientific Advisory Board, UpToDate as a content author, and Pfizer for service on DSMB outside the submitted work. DJJ reports grants from NIH during the conduct of the study and grants and/or personal fees from Areteia, Astra Zeneca, Apogee, GlaxoSmithKline, Sanofi, Regeneron, Upstream Bio, outside the submitted work. JEG reports grants from NIH during the conduct of the study, personal fees from Arrowhead Pharmaceuticals, Meissa Vaccines Inc., and Gilead Sciences, and stock options for Meissa Vaccines Inc. outside the submitted work. SVL is a co-founder, board member and consultant for Siolta Therapeutics, Inc, and holds stock in the company. All other authors declare no competing interests.

Chapter 3 Acknowledgements

Chapter 3 is reprinted, with reformatting, from a manuscript currently under review at *Microbial Genomics* and posted as a preprint on bioRxiv: Steininger HM†, Iglesias-Aguirre CE†, Panzer AR, Durack J, McKean M, Cabana MD, Diamond S‡, Lynch SV‡. "Carbohydrate Metabolism Differs in Infants by Asthma-risk Status and is Associated with the Functional Potential of Bacteroides cellulosilyticus." † denotes co-first authorship; ‡ denotes co-corresponding authorship.

Funding: This work was funded in part by The Audacious Project: Lyda Hill Philanthropies, Acton Family Giving, the Valhalla Foundation, Hastings/Quillin Fund – an advised fund of the Silicon Valley Community Foundation, the CH Foundation, Laura and Gary Lauder and Family, the Sea Grape Foundation, the Emerson Collective, Mike Schroepfer and Erin Hoffman Family Fund – an advised fund of Silicon Valley Community Foundation, and the

Anne Wojcicki Foundation through The Audacious Project at the Innovative Genomics Institute. Funding was also provided in part by a Research Award from the Shurl and Kay Curci Foundation to the Innovative Genomics Institute Genomic Tool Discovery Program at UC Berkeley, awarded to S. Diamond. SVL's research program is funded in part by NIH/NIAID awards AI128482, AI148104, AI160040 and AI089473.

Conflict of interest: SVL is a co-founder, board member and holds stock in Siolta Therapeutics Inc. All other authors declare no conflict of interest with respect to the work performed in this study.

Ethical approval: The Trial of Infant Probiotic Supplementation (TIPS – ClinicalTrials.gov ID: NCT00113659) clinical trial and Development of Infant Microbial Evolution (DIMES) birth cohort were both approved by the University of California, San Francisco IRB #2019-10582.

Chapter 4 Acknowledgements

Funding: This work was funded in part by The Audacious Project: Lyda Hill Philanthropies, Acton Family Giving, the Valhalla Foundation, Hastings/Quillin Fund – an advised fund of the Silicon Valley Community Foundation, the CH Foundation, Laura and Gary Lauder and Family, the Sea Grape Foundation, the Emerson Collective, Mike Schroepfer and Erin Hoffman Family Fund – an advised fund of Silicon Valley Community Foundation, and the Anne Wojcicki Foundation through The Audacious Project at the Innovative Genomics Institute. Funding was also provided in part by a Research Award from the Shurl and Kay Curci Foundation to the Innovative Genomics Institute Genomic Tool Discovery Program at UC Berkeley, awarded to S. Diamond. SVL's research program is funded in part by NIH/NIAID awards AI128482, AI148104, AI160040 and AI089473.

Conflict of interest: SVL is a co-founder, board member and holds stock in Siolta Therapeutics Inc. All other authors declare no conflict of interest with respect to the work performed in this study.

Contributions

Chapter 1 Contributions

This is original work written for this dissertation by Holly M. Steininger and reviewed by Susan V. Lynch and Spencer Diamond.

Chapter 2 Contributions

Holly M. Steininger, BS*,* Kathryn E. McCauley, MPH*, Gabe E. Braga, BS, Andréanne Morin, PhD, Kristine E. Lee, MS, Cynthia M. Visness, PhD, MPH, Douglas W. Fadrosh, MS, Matthew Dapas, PhD, George T. O'Connor, MD, Leonard B. Bacharier, MD, Katherine Rivera-Spoljaric, MD, Meyer Kattan, MD, Robert F. Lemanske, Jr, MD, Edward M. Zoratti, MD, Fernando D. Martinez, MD, Diane R. Gold, MD, Tina V. Hartert, MD, MPH, Christine C. Johnson, PhD, MPH, Rachel L. Miller, MD, Christine M. Seroogy, MD, Robert A. Wood, MD, Alkis Togias, MD, Daniel J. Jackson, MD, James E. Gern, MD, Carole Ober, PhD, Susan V. Lynch, PhD for the CADRE investigators. *co-first authorship. Corresponding author: Susan V. Lynch, PhD.

Chapter 3 Contributions

Holly M. Steininger, BS*, Carlos E. Iglesias-Aguirre, PhD*, Ariane R. Panzer, PhD, Juliana Durack, PhD, Michelle McKean, Michael D. Cabana, MD, MPH, *Spencer Diamond, PhD*, and *Susan V. Lynch, PhD*. *co-first authorship*. co-corresponding authorship*.

Author Contributions (CRediT):

HMS: Formal analysis; Visualization; Software; Writing - original draft; Writing - review & editing.

CEIA: Formal analysis; Validation; Visualization; Writing - review & editing.

ARP: Conceptualization; Investigation; Writing - original draft. JD - Investigation.

MM: Project administration.

MDC: Data curation; Project administration; Resources.

SD: Funding acquisition; Software; Supervision; Writing - review & editing.

SVL: Conceptualization; Funding acquisition; Supervision; Writing - review & editing.

Chapter 4 Contributions

Holly M. Steininger BS, Michelle McKean, Michael D. Cabana, MD, MPH, Faith Vundla, MD, Spencer Diamond, PhD, and Susan V. Lynch, PhD

Author Contributions (CRediT):

HMS: Conceptualization; Investigation; Formal analysis; Visualization; Writing - original draft; Writing - review & editing.

MM: Project administration.

MDC: Data curation; Project administration; Resources.

FV: Investigation

SD: Funding acquisition; Software; Supervision; Writing - review & editing.

SVL: Conceptualization; Funding acquisition; Supervision; Writing - review & editing.

Abstract

Atopy-bound Infant Microbiomes

Holly M. Steininger

In this dissertation, we demonstrate how clinical cohort studies of allergic disease can be paired with current sequencing and experimental technologies to identify microbial targets for disease prevention. The first year of life is a critical window during which the microbiome is shaped by, and shapes, the host immune and epithelial systems. This work is organized around the two organ systems most directly implicated in this process: the upper airway, where microbial interactions have a more direct influence on airway-specific disease, and the gut, where the microbiome shapes the baseline immune tolerance established in the first year of life and, with it, the overall capacity to develop allergic disease.

In the upper airway, we show that genetic risk alleles long associated with asthma, *GSDMB* and *ORMDL3*, interact with the airway microbiome at 1 year of age to influence early-life wheeze risk. Infants carrying these risk alleles who are dominated by *Moraxella*, *Streptococcus*, or *Haemophilus* at 1 year have increased wheeze risk (Pint = 0.016 for all), while dominance by *Corynebacterium*, *Dolosigranulum*, *Staphylococcus*, or *Bacillus* reduces risk. In airway epithelial cells CRISPR-edited to be homozygous for the GSDMB risk variant rs7216389^{TT}, we observed decreased expression of genes involved in antimicrobial response and neutrophil recruitment, along with increased microbial adherence, linking host genotype, epithelial function, and microbial colonization directly.

In the gut, we characterized stool from a longitudinal, high-risk infant cohort across the first year of life. In a 6-month subset, we identified *Bacteroides cellulosilyticus* and *Hungatella effluvii* as species anti-associated with asthmatic disease risk, with putative roles in carbohydrate and nitrogenous compound metabolism respectively. Extending this analysis across the full first year of life, we assembled a 563-genome catalog from shotgun metagenomic sequencing and identified bacterial genomes whose presence or absence within specific time windows tracks atopic, eczema, and asthma outcomes. The strongest genome-level association with asthma was *Akkermansia massiliensis*, enriched at birth ($q = 0.043$) and elevated in prevalence through 6 months of age before normalizing to levels seen in infants who do not develop asthma, underscoring the importance of timing in microbial-associated allergic disease risk. This genome catalog provides the foundation for ongoing work identifying bacterial enzymatic targets, particularly within the linoleic acid pathway that produces 12,13-diHOME, and for integrating taxonomic, functional, and metabolomic data to define the axes of variation that best explain allergic disease risk.

Together, these studies show that specific microbes, at specific times, in specific host genetic contexts, shape the trajectory toward allergic disease. The ultimate goal of this work is to design interventions that limit allergic disease development in future generations.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

The ABCs of the Allergic Infant Microbiome

I am the first scientist in my family, and as far as I am aware, I am the first Steininger in my family to receive a doctorate. As such, I wanted to gear this introduction as much as possible to a general audience by walking you through the historical context and basic concepts that underlie this body of work. Welcome to the ABCs of the allergic infant microbiome.

A: Atopy, allergy, and asthma. I expect the word allergy to be the most familiar to people. You might imagine a child having an anaphylactic reaction to food. Asthma is likely also ,connoting images of inhalers to calm the restriction in someone's lungs when they suddenly can't breathe. The reason I am confident in these being familiar diseases is due to how widespread and prevalent they are in America. Asthma is the most common chronic childhood disease, affecting 6.5% of U.S. children as of 2024.[1] These allergic diseases are connected by an umbrella term: atopy. Atopy now is commonly used as just another word for allergy, but it was initially coined by Coca and Cooke in 1923 because of frustration with existing definition of allergy.[2] Allergy at this time meant any increased immune activity, regardless of if it was appropriate or not. Therefore they based this term, atopy, on the Greek word atopia which means out of place or absurd to refer to diseases of hyperreactivity to inappropriate stimuli.[2] Now roughly 100 years later the words allergic and atopic are commonly used interchangeably in scientific literature. So, if you don't like the word atopic, replace it in your mind with the word allergic and know what I am referring to is an inappropriate response to an otherwise harmless stimulus. Roughly speaking, when this happens in your lungs it is diagnosed as

asthma, in your skin eczema, and in your nose allergic rhinitis. Although these allergic diseases affect various organs, they often share a common circulating blood molecule known as IgE.[3] High IgE levels are considered a signature defining feature of an atopic disease[4]; however we now know that there are individuals with high IgE but no clinical manifestations[5] and alternatively, strong clinical manifestations but relatively low IgE.[6] Partially because of this, we still rely heavily on patients' symptom reports and doctor visit records. So, in summary, A is for Atopy and Atopy is an Allergic Disease.

B: Bacteria. I was tempted to use B for babies and include photos of my adorable infant nephews as a figure, but unfortunately bacteria take precedence in this work. In the simplest sense, there are two conflicting views of bacteria. The germ theory of disease developed by Pasteur and Koch in the 19th century was transcendental.[7] It replaced the concept of humors driving disease to introduce the idea that microorganisms could drive infectious disease.[7] On the other hand, for all of human history we have been eating fermented foods, and some scientists in the 19th century such Metchnikoff and Shirota[8] believed that it was these bacteria which actually lead to health and longevity.[9] So I've chosen B to be for bacteria. Bacteria can drive disease, but they also likely important for protecting us from disease. How do we know which bacteria to trust? That brings me to C.

C: Cohort studies. Cohort studies are observational studies in which a cohort, or a group of individuals sharing some characteristic, are followed up over time, and outcomes are measured at one or more time points. These studies of what is happening in real world populations enable to understand what factors are associated with driving disease, and which are associated with protection from disease. These cohort studies are critical for providing the

context to interpret how microbiology impacts humans. One of the seminal cohort studies in the field of allergic disease was done by Strachan and published in 1988.[10] This was well before the term microbiome was popularized in 2001.[11] In this humble 1-page paper with a single table and no figures, Strachan reported in a cohort of 17,414 British children that the number of older children in the household was inversely related with hay fever prevalence.[10] So the more older siblings you have, the less likely you are to have hay fever.[10] It was this observation that is widely credited with giving rise to the hygiene hypothesis. The hygiene hypothesis proposes that the exposure to microbes, parasites, and environmental bacteria in early childhood is critical to the immune system's development, and that us youngest siblings often have far more of these exposures.

A real-life example of the ABCs

I believe that the closest thing to the truth in many human health questions often lies in the middle of all the data, and ultimately it is context that is critical for understanding human health outcomes. In large population studies, it is easy to lose the human context, so I'm going to share my family as a case example to illustrate the importance of context that may occur within the study of a larger population. I am the fifth and final child, and my siblings and I were born within 7 years of each other. First my oldest brother (#1) Eric, unfortunately, developed an allergic disease. He has struggled with severe eczema since his teenage years, and I remember as a child hearing him yell in pain in the shower because of the sensation of water on his skin. Next comes my brother Matt (#2), who has a food allergy, followed by Andrew (#3) who unfortunately also struggles with severe eczema. Thanks to the development of monoclonal antibodies to block the receptors on some of the immune cells that cause these reactions,[12]

both Andrew and Eric have found some relief in recent years. Then comes David (#4) who unfortunately received an eosinophilic esophagitis (EOE) diagnosis last year, which is also due an allergic immune response. Finally there was me. As if I didn't already owe my brothers enough for all the support, guidance, and personality they have given me, now I am thanking them for whatever they exposed me to in early life that may have kept me from developing an atopic disease. There are three points I hope to illustrate with this simplistic vignette: 1) A 20.4 % prevalence of hay fever with no older siblings in the home to a 6.5% prevalence with 4+ older siblings does not mean that any one individual has "20.4% hay fever", but rather that this is representative of the disease burden on a population level. 2) These population-level cohort studies on allergic disease are critical for capturing the healthcare realities happening in the homes of families. 3) These cohort studies are deeply precious and enable scientists to find trends, study them, and develop interventions to reduce the disease burden on those at the greatest risk.

In this dissertation, we demonstrate how clinical cohort studies of allergic disease can be paired with current sequencing and experimental technologies for insights into microbial applications for prevention of disease. This work is most naturally broken down into the two organ systems of interest due to their interactions with microbes: the upper airway and the gut. I view the upper airway microbial interactions with allergic disease as a more direct influence on airway-specific diseases,[13–15] and the gut microbiome as an influence on the overall capacity to develop allergic responses via the baseline immune tolerance level established in the first year of life.[16–19] The ultimate goal of this field is to determine what exposures (microbes, food, air particulates), when (first month, during weaning, first year, anytime), and for who (genetic

and environmental restraints) drive allergic disease in order to design interventions to limit the development of these diseases for future generations.

Chapter 2: Upper Airway Microbiome Interacts with *GSDMB* and *ORMDL3* Asthma Risk SNPs to Influence Early-Life Wheeze Risk

Abstract

Background: Single-nucleotide polymorphisms (SNPs) in the chromosome 17q12-q21 region and, independently, early-life nasal microbiota dominated by *Moraxella*, *Streptococcus*, or *Haemophilus* increase risk of chronic wheeze and asthma development.

Objective: Determine if 17q12-q21 risk SNPs and nasal microbiota interact to modulate childhood wheeze risk.

Methods: Nasal wash samples from 12-month-old infants in two birth cohorts, Childhood Origins of Asthma (COAST; n = 180) and Urban Environment and Childhood Asthma (URECA; n = 139), underwent *16S rRNA* V4 sequencing. Nasal microbiota dominated by *Moraxella*, *Streptococcus* or *Haemophilus* (MSH) or *Corynebacterium*, *Dolosigranulum*, *Staphylococcus* or *Bacillus* (CDSB) were assessed. Paired blood was genotyped for nine 17q12-q21 risk SNPs. Logistic regression tested interactions between 17q12-q21 SNPs and MSH or CDSB on wheeze risk in the first three years of life. A549 lung epithelial cells, CRISPR-edited to encode the rs7216389 risk genotype ($rs7216389^{TT}$) were compared to the heterozygous ($rs7216389^{CT}$) line using bulk RNA sequencing.

Results: SNPs, particularly those in the Orsomucoid-like sphingolipid biosynthesis regulator 3 (rs8076131; OR = 1.72, CI = 1.09-2.71, P_{int} = 0.031) and Gasdermin B (rs2305480, OR = 1.72, CI = 1.09-2.71, P_{int} = 0.042 and rs7216389, OR = 1.73, CI = 1.09-2.70, P_{int} = 0.047) genes,

interact with MSH microbiota to increase early-life wheeze risk (false discovery rate [FDR] P_{int} = 0.016 for all), while interactions with CDSB reduce risk. A549 airway epithelial cells homozygous for rs7216389^{TT} exhibited decreased expression of genes involved in antimicrobial responses and neutrophil recruitment and evidence increased microbial adherence compared with the heterozygous cell line.

Conclusion: Airway microbiota interact with SNPs at the 17q12-q21 locus in genes involved in sphingolipid metabolism and intracellular antimicrobial responses, to modulate wheeze risk.

Clinical Implication: Interventions aimed at decreasing pathogenic bacterial colonization of the airways may reduce wheeze burden in genetically susceptible children.

Capsule summary: Upper airway microbiota interact with asthma-risk genetic variants to influence early-life wheeze risk.

Keywords: Airway microbiota, Wheeze, 17q12-q21 risk loci.

Abbreviations:

COAST: Childhood Origins of Asthma

CDSB: Corynebacterium, Dolosigranulum, Staphylococcus, Bacillus

CREW: Children's Respiratory and Environmental Workgroup

GSDMB: Gasdermin B

MSH: *Moraxella*, *Streptococcus*, or *Haemophilus*

ORMDL3: Orsomucoid-like sphingolipid biosynthesis regulator 3

PCR: Polymerase chain reaction

SNP: Single-nucleotide polymorphism

SV: Sequence variant

URECA: Urban Environment and Childhood Asthma

Introduction

Frequent early-life wheezing illness and viral respiratory infections are amongst the most common risk factors for asthma development in later childhood.[20–22] Childhood asthma accounts for a significant proportion of pediatric morbidity in the United States and significant healthcare expenditure.[23] Parental asthma increases childhood asthma risk 2.2-fold compared to parents without asthma,[24] implicating a genetic component to disease development.

Variants at the 17q12-q21 locus are amongst the most highly replicated single-nucleotide polymorphisms (SNPs) in childhood-onset asthma genome-wide association studies .[25,26] Multiple independent studies have demonstrated that children are more likely to develop asthma if they experience one or more virus-associated wheezing illnesses in the first year of life and carry one or more risk SNPs in the Gasdermin B gene (*GSDMB*).[27,28] In addition, these SNPs have been associated with distinct wheeze trajectories (transient, late-onset and persistent, compared with infrequent) ascertained over the first 10 years of life,[29] and specific risk variants (e.g., rs2305480 in *GSDMB)* replicate across both European American and African American populations.[30] The latter population exhibits reduced linkage disequilibrium across the 17q12-q21 locus, offering the opportunity to identify potential causal SNPs underlying these associations.[30] More recently, in the European PASTURE cohort, rs7216389 in *GSDMB* was associated with childhood wheeze, specifically within a group of children who possessed a distinct gut microbiota composition,[31] suggesting that certain microbial exposures may enhance the pathogenic effects of 17q12-q21 risk variants on wheezing.

The functions of genes at the 17q12-q21 locus support this theory. For example, *ORMDL3 (*Orsomucoid-like sphingolipid biosynthesis regulator 3), primarily expressed in airway epithelial cells, controls sphingolipid homeostasis, which regulates cell membrane integrity.[32,33] Membrane integrity is key to microbial infection susceptibility.[34] *ORMDL3* also promotes autophagy and cell death by impairing intracellular calcium mobilization[35] and regulates epithelial repair via endoplasmic reticulum signaling. *GSDMB* has been linked to multiple processes involving microbes, including pyroptosis,[36–38] epithelial repair,[39] airway remodeling,[40] and regulation of interferon and antimicrobial responses, specifically against Gram-negative bacteria.[41]

Upper airway microbiota play a role in the development and clinical presentation of asthma.[13,14,42,43] Data from multiple studies indicate that in infancy, upper airway microbiota dominated by pathobionts of the *Moraxella*, *Streptococcus* or *Haemophilus* (MSH) genera, and in some studies, *Staphylococcus aureus*, are associated with increased risk for developing childhood asthma.[13,14,42,43] In a study of older, school-aged children with asthma, those with upper airway microbiota dominated by *Moraxella* were at increased risk of asthma exacerbations, while microbiota dominated by *Staphylococcus* or *Corynebacterium* associated with decreased exacerbation risk.[15] In that study, *Moraxella catarrhalis* isolated from children with asthma induced increased epithelial damage and IL-33 and IL-8 expression in cultured A549 cells, compared with isolates of *Staphylococcus epidermidis* or *Corynebacterium propinquum* obtained from relatively healthier children in the same cohort.[15]

Because *GSDMB* and *ORMDL3* encode functions linked to both epithelial repair and antimicrobial pathways, and both genotype and airway microbiota independently confer

respiratory disease risk in large pediatric cohorts, we hypothesized that interactions between risk variants in these genes and upper airway microbiota dominated either by MSH or by CDSB modulate the risk of wheeze in the first three years of life. More specifically, we hypothesized that interactions between *GSDMB* and *ORMDL3* asthma risk variants and MSH microbiota increase wheeze risk beyond their independent effects, while CDSB airway microbiota decrease risk, even in the context of genetic susceptibility.

Methods

Study Populations

The current study utilized previously harmonized metadata generated from participants in two cohorts within the Children's Respiratory and Environmental Workgroup (CREW) consortium, COAST (Childhood Origins of Asthma) and URECA (Urban Environment and Childhood Asthma)(**Figure S2.1A**).[29] COAST is a high-risk asthma birth cohort which enrolled children (N = 289) November 1998-2000 from predominantly suburban neighborhoods in Madison, Wisconsin.[44] High risk was defined as having at least one parent with respiratory allergies (defined as one or more positive aeroallergen skin tests) and/or a history of physician-diagnosed asthma. URECA is a prospective multi-site high-risk asthma observational birth cohort recruited from urban neighborhoods in Baltimore, Boston, New York, and St Louis.[45] URECA children (N = 560) had at least one parent with an allergic disease or asthma, and all families lived in areas in which at least 20% of the population had incomes below the poverty line. The local institutional review board at each participating site approved the study protocol. Written informed consent or parent/guardian permission was obtained along with child assent

as appropriate for participation in specific cohorts and the CREW protocol. The CREW consortium within the Environmental influences on Child Health Outcomes Program was established in 2016. In 2023, CREW transitioned into the Children's Allergy and Asthma Data Repository (CADRE) to promote data sharing and reuse.

Clinical and demographic data, harmonized in a prior published study,[29] included child sex, smoking during pregnancy, and pets in the household amongst other variables (**Table 2.1**). The original demographic surveys collected parent-reported race, and race/ethnicity were harmonized as described in Hallmark et Al. 2021.[29] Wheeze status was ascertained via parent report annually. Early-life wheeze was defined as parent-reported wheeze during the first three years of life.

Sample Collection

Nasal samples collected at age 12 months from URECA (N=139) and COAST (N = 180) participants with paired 17q12-q21 genotyping data were used for the current study. URECA nasal samples were collected as previously described and stored in saline/gelatin solution (Special L#200 Gelita SOL KKA, lot 4) at −80°C.(29) COAST nasopharyngeal mucus samples (swab or aspirate) collected at 12-month clinic visits were flash frozen prior to storage at −80°C as previously described.[44]

CTAB DNA Extraction

DNA was extracted from URECA nasal wash samples, negative extraction controls, and three sterile aliquots of saline/gelatin solution (storage media controls) using a modified cetyltrimethylammonium bromide (CTAB)–polyethylene glycol (PEG) protocol as previously

described.[46] Briefly, 300 mL of nasal wash sample was added to 500 mL of CTAB extraction buffer (5% CTAB in 0.25 M phosphate buffer and 1M NaCl), followed by the addition of 500 µL phenol: chloroform: isoamyl alcohol (25:24:1). Cells were lysed by beat-beating in Lysis Matrix E tubes (MB Biomedicals) at 5.5 m/s for 30 sec; phases were separated by centrifugation at 16,000*g* for 5 min and the aqueous layer transferred to a new tube. To improve extraction efficiency, a second extraction was performed via addition of 500 µL CTAB to the original extraction tube, beat-beating and phase separation, after which the aqueous layer was mixed with that from the previous extraction. To remove excess phenol, the pooled aqueous layer from both extractions was extracted by vortexing in an equal volume of chloroform followed by centrifugation at 16,000*g* for 5 min. The resulting aqueous phase was added to 1 mL of polyethylene glycol precipitating solution (30% PEG 6000 in 1.6 M NaCl) and stored overnight at 4°C. Precipitated DNA was recovered by centrifugation for 10 min at 13,000*g*; DNA pellets were washed twice with 300 µl 70% ethanol, air-dried for 10 minutes, and resuspended in 50 µl of sterile water. DNA concentrations were quantified using the Qubit dsDNA HS Assay Kit (ThermoFisher Scientific, MA), diluted to 10 ng μl^{-1} and stored at −20°C.

16S rRNA Sequencing and Data Processing

The variable region 4 (V4) of the *16S rRNA* gene was amplified in triplicate from URECA DNA samples using 515F/806R primers and polymerase chain reaction (PCR) conditions previously described.[47] Triplicate amplicon reactions per sample were pooled before being purified using the SequalPrep Normalization Plate Kit (ThermoFisher Scientific, MA) according to the manufacturer's specifications. Amplicons were quantified using the Qubit dsDNA HS Assay Kit (ThermoFisher Scientific, MA) and pooled at equimolar concentrations. The amplicon

library was concentrated using the Agencourt AMPure XP system (Beckman-Coulter), quantified using the KAPA Library Quantification Kit (KAPA Biosystems), and diluted to 2 nM. Equimolar PhiX was added at 40% final volume to the amplicon library followed by sequencing on an Illumina NextSeq 500 Platform on a 2 × 150 bp sequencing run.

Demultiplexed raw sequence reads from both COAST and URECA underwent Divisive Amplicon Denoising Algorithm 2 (DADA2) in a run-specific manner.[48] The same quality parameters were applied for all runs, including a maximum expected error of 2 on the forward and reverse read. Downstream processing used the default parameters or those described in the DADA2 tutorial.[48] After a Sequence Variant (SV) table was generated, taxonomy was assigned using SILVA v138.[49] Negative controls in URECA were assessed on a plate-by-plate basis, with any sequence variants present in the negative control at 50 reads or more removed from sample data generated in the same 96-well plate.

COAST samples were extracted and sequenced at the University of Melbourne in Melbourne, Australia as previously published.[14] This current report used extant 16S rRNA data available from that study for analysis. In COAST, negative controls were not considered in background signal correction as per previous published work.[14]

Once the SV tables from the two cohorts were combined, SVs present in fewer than 0.001% of total reads were removed (SVs with fewer than 900 reads). The resulting SV table was representatively rarefied at 14,041 reads per sample, a threshold that maintained diversity and did not remove a significant number of samples. Representative rarefaction is a method that avoids profile bias that can occur when sequence data is sub-sampled only once. The approach randomly sub-samples amplicon sequence variant (ASV) data from each sample to

the desired read depth 100 times and ordinates the resulting community compositions using a distance matrix. For our study, for each sample, the bacterial profile at the centroid of a Bray-Curtis distance matrix was used for each sample in subsequent analyses.[50] This depth was additionally confirmed using Procrustes analysis of representatively rarefied read depths above and below this threshold to confirm that community structure was retained. The final sequence variant table comprised 1,292 SVs detected across 319 subjects. No sibling pairs were included in the study, and thus no relatedness adjustments or filtering were performed.

Genotyping

SNPs were genotyped as previously described using TaqMan assays (Applied Biosystems, CA) according to the manufacturer's instructions with standard quality control checks.[30] Nucleotide call rates were high (>95%) for all SNPs as previously described.[30] Nine SNPs described in **Table S2.1** formed the focus of this analysis: rs3859192 (*GSDMA*), rs8069202 (*GSDMA*), rs8076131 (*ORMDL3*), rs4065275 (*ORMDL3*), rs7216389 (*GSDMB*), rs2305480 (*GSDMB*), rs12936231 (*ZPBP2*), rs2517955 (*PGAP2*), and rs2941504 (*PGAP3*).

Data Analysis

Odds ratios (OR) and 95% confidence intervals (CI) were estimated using logistic regression. The dependent variable was wheezing during the first three years of life coded as a binary outcome (Yes/No). Odds ratios were obtained by exponentiating the estimated log-odds coefficients. Models were adjusted for sex, age in months at sample collection, cohort (if applicable) and the first three ancestry principal components (PCs), determined *a priori* in a previous publication using a genome-wide SNP array[35] using the following statistical model:

$$logit(P(Y = 1)) = \beta_0 + \beta_1 SNP + \beta_2 MSH + \beta_{int}(SNP\ x\ MSH) + \ldots$$

$$\ldots \beta_3 PC_{1,anc} + \beta_4 PC_{2,anc} + \beta_5 PC_{3,anc} + \beta_6 Sex + \beta_7 Age + \beta_8 Cohort$$

Where logit(P(Y = 1)) is the natural-log of the odds that a child experienced wheezing in the first three years of life (Yes = 1, No = 0). *SNP* is the genetic variant under study coded by the number of risk alleles. Risk alleles are defined in **Table S2.1**. *MSH* indicates the presence of *Moraxella*, *Streptococcus* or *Haemophilus*–dominated airway microbiota expressed as a binary outcome: *MSH* = 1, MSH dominated upper airway; *MSH* = 0, CDSB dominated upper airway, where CDSB represents upper airways dominated by *Corynebacterium*, *Dolosigranulum*, *Staphylococcus* or *Bacillus*. β_{int} (SNP × MSH) is the interaction term; the coefficient β_{int} quantifies how the effect of the SNP on wheezing differs when MSH is present versus absent (i.e., the difference in the SNP's log-odds effect between MSH strata). *PC1-3, anc*. are the first three principal components from genome-wide SNP data capturing major axes of population ancestry. *Sex* is the child's sex (binary, 0 = male), and *Age* is the child's age in months at the time of sample collection.

P-values for the interaction term (P_{int}) were adjusted for multiple testing using Benjamini-Hochberg[51] unless explicitly stated otherwise. Airway samples were classified by the dominant genus, i.e., the genus classification of the ASV with the greatest number of reads for a given sample. Genera of ASVs identified as dominant in only one sample were classified into an "Other" category, and these samples were excluded from analysis. Interaction models were fit between genotype and upper airway domination by either MSH or CDSB.

A549 rs7216389 CRISPR Editing

In our study, the rs7216389^{TT} risk variant exhibited significant interactions with upper airway microbiota to influence childhood wheeze outcomes and was thus chosen for further study. CRISPR Cas9 mediated knock-in cell clones of *GSDMB* rs7217389 SNP in A549 cells were generated by EditCo Bio Inc. (Redwood City, CA, USA). To generate these cells, Ribonucleoproteins containing the Cas9 protein and synthetic chemically modified guide RNA (5'-GTCACAAACATGCATGGACT-3') produced by EditCo Bio, Inc. were electroporated into the cells along with a single-stranded oligodeoxynucleotide (ssODN) donor using EditCo's optimized protocol. Editing efficiency was assessed upon cell recovery, 48 hours post electroporation. Genomic DNA was extracted from the edited cells, PCR amplified, and sequenced using Sanger sequencing. The resulting Sanger chromatograms were processed using EditCo's Inference of CRISPR edits software or NGS bioinformatics pipeline. To create monoclonal cell populations, edited cell pools were seeded at 1 cell/well using a single-cell printer into 96- or 384-well plates. All wells were imaged every 3 days to ensure expansion from a single-cell clone. Clonal populations were screened and identified using the PCR-Sanger-ICE genotyping strategy described above. Clone #1 used for the co-culture RNA-sequencing experiment described below had a KI-Score of 100. Clone #2 used for validation had a KI-Score of 98.

A549 Bacterial Co-culture

Cell lines (CT or TT at rs7216389) were cultured in F12-K media supplemented with 10% FBS and 5% Pen-Strep. Cells at passage number 13 and 15, respectively, were dissociated with 0.25% trypsin, counted with a Luna Cell Counting System, plated to confluency at 200,000 cells/well in a 24-well plate in media without antibiotic and incubated for 12 hours. One hour

prior to bacterial addition, media was changed to 400 µL fresh antibiotic-free media. *S. epidermidis* (isolated from a URECA subject's nasal swab) was cultured in BHI at 37°C under static conditions. Glycerol stocks (35% final glycerol concentration) were made and used to inoculate 5 ml of BHI media and grown overnight as described above before being diluted to an OD_{600} of 0.1. *S. epidermidis* was grown under the aforementioned conditions for 2.5 hours to reach the exponential phase of growth. Based on a linear regression of OD_{600} to colony forming unit (CFU) counts, bacterial cultures were harvested during mid-exponential phase for a target of 1×10^7 cells ml^{-1}; 134.2 µL of *S. epidermidis* were harvested at OD_{600} 0.0937 resulting in 7.8×10^6 cells mL^{-1}. Cells were centrifuged (5000*g*, 5 minutes), washed with 1 ml of sterile phosphate-buffered saline (PBS) and resuspended in 1 mL of antibiotic-free A549 Media. A total volume of 100 µL of *S. epidermidis* containing ~1×10^6 bacterial cells to achieve a goal multiplicity of infection (MOI) of 5 was added to triplicate wells of A549 cells (CT or TT at rs7216389) and co-cultured for 6 hours at 37°C. At the time of co-incubation with A549 cells, bacteria were also plated in parallel on blood agar plates and incubated under the strain-specific conditions to confirm via CFUs that the intended MOI had been achieved. The achieved MOI was 4.1 for *S. epidermidis*. Co-culture supernatant was also serially diluted and plated on blood agar post-incubation to determine bacterial CFUs at the end of experiment. Following three rounds of mammalian cell washing with sterile PBS, cells were thoroughly resuspended in Trizol and stored at 4°C overnight. Samples underwent RNA extraction the following day using the Zymo Direct-zol RNA Miniprep kit (Zymo Research; Irvine, CA).

RNA-Sequencing

RNA quality was assessed via bioanalyzer before rRNA depletion using the NEBNext rRNA Depletion Kit (Human/Mouse/Rat; cat #E7400). RNA-sequencing library preparation was performed with the SMARTer Stranded RNA-Seq Kit v2 (cat #634412), and sequencing was performed using an Illumina NextSeq 500/550 High Output Kit v2.5 (cat #20024908) on an Illumina NextSeq 500 instrument.

RNASeq Data Processing and Analysis

Raw RNASeq reads were trimmed and quality filtered using fastp,[52] retaining reads with an average quality score >30. Reads were aligned to the human genome (GRCh38.p14) with STAR,[53] and a gene count table was generated with featureCounts.[54] Libraries were filtered for a minimum read-depth of 500,000 reads, resulting in one sample lost from the analysis. Genes with low expression (max counts per million reads <6) were removed from the dataset before normalization with voom[55] to prevent skewing the normalization process, focus on reliably expressed genes, and reduce noise in differential expression analysis, aligning with recommendations for improved statistical modeling accuracy. This resulted in 16,750 genes in the final dataset.

For RNASeq datasets, plotMDS from limma[56] was used to approximate the distance between samples based on log-fold change in gene expression. This approach was applied before voom normalization. Voom estimates the mean-variance relationship of the log-counts, generates a precision weight for each observation, and enters these into the limma empirical Bayes analysis pipeline.[55] Benjamini-Hochberg was used for multiple comparison correction.[51] A multidimensional scaling (MDS) plot of leading $\log_2$ fold-change gene expression profiles based

on pairwise comparisons of the top 500 most variable genes was constructed before conducting multivariate analysis of variance using Pillai's Trace statistic to assess the influence of bacteria and genotype on sample distribution across MDS axes. The variance explained (R^2) for each factor was estimated by normalizing Pillai's Trace across the top two MDS dimensions. Differential expression analysis was performed with limma,[56] and gene set enrichment analysis was performed using Enrichr.[57] Limma employs a linear model used to make the following comparisons: 1. No bacterial exposure subset (Notx) of rs7216389^{TT} versus rs7216389^{CT}; 2. *S. epidermidis* co-cultured subset *(S. epi)* of rs7216389^{TT} versus rs7216389^{CT}; 3. rs7216389^{CT} *S. epi versus* rs7216389^{CT} Notx; 4. rs7216389^{TT} *S. epi versus* rs7216389^{CT} Notx.

RT-qPCR

Total RNA was extracted from A549 lung epithelial cells using the Direct-zol RNA Miniprep Kit (Zymo Research; Irvine, CA). RNA was stored at −80°C until being reverse-transcribed using the High-Capacity RNA-to-cDNA™ Kit (ThermoFisher Scientific, MA) according to the manufacturer's instructions. The cDNA was diluted to concentration of 1.25 ng/µL and stored at −20°C until use. Real-time quantitative PCR (qPCR) was performed in triplicate using SYBR Green Master Mix (Life Technologies) on the QuantStudio 6 Real-Time PCR System (Applied Biosystems) using standard cycling parameters. Data were generated in two independent qPCR experiments. The primer pairs used are provided in **Table S2.2**. The relative fold change in expression of each gene was normalized to GAPDH expression by the $2^{-\Delta\Delta CT}$ method.

Attachment Assay

A549-bacterial co-cultures were prepared as described above. Prior to bacterial inoculation of cell lines, A549 cells were quantified using a LUNA-II™ cytometer. Co-cultures of *S. epidermidis* were incubated at 37°C, 5% CO_2 for 2 hours prior to being washed three times with sterile PBS. After washing, cells were dissociated with 0.25% trypsin at 37°C for 5 minutes, plated on blood agar plates, and incubated overnight for CFU numeration of adherent bacteria. Adherent bacteria were calculated as CFU bacteria per A549 cell prior to the 2-hour co-culture experiment.

Code Availability

Code for 16S rRNA sequence analyses can be found on GitHub at https://github.com/lynchlab-ucsf/dada2_pipeline. Data analysis is available at https://github.com/hollysteininger/17qxmicrobiota.

Data Availability

The Sequence Read Archive contains URECA and COAST 16S rRNA sequence data under accession numbers PRJNA1391558 and PRJNA1391569. CADRE will share participant metadata upon reasonable request.

Results

Analysis of between-cohort differences indicated that children in the COAST cohort were primarily parent-reported White, while URECA participants were primarily parent-reported Black. URECA participants were slightly older when the 12-month sample collection occurred, were more likely to be exposed to maternal smoking during pregnancy and had fewer

cats or dogs in the household during infancy (Chi-Square < 0.05 for all; **Table 2.1**). No significant difference was observed in the rate of illness evident at the clinic visit, proportion of female children, or the occurrence of wheeze in the first three years of life across the two cohorts (P > 0.05).

The distribution of nasal microbiota dominant genera differed between cohorts (Chi-square test; P < 0.0001; **Table 2.1**). COAST subjects were primarily dominated by *Moraxella* (n = 78; 42.4%), *Streptococcus* (n = 38; 20.7%), *Dolosigranulum* (n = 32; 17.4%), *Haemophilus* (n = 16; 8.7%) and *Corynebacterium* (n = 13; 7.1%). URECA participants were also more likely to be dominated by *Moraxella* (n = 53; 37.3%), *Dolosigranulum* (n = 29; 20.4%), *Haemophilus* (n = 19; 13.4%) and *Streptococcus* (n = 11; 7.7%), but additionally evidenced a small number of samples with *Bacillus*-dominated microbiota (n = 15; 10.6%). *Bacillus* sequence variants were solely identified in URECA participants and given stringent negative control correction applied to the URECA samples; this difference is unlikely to have arisen from technical sources.

Data from the two cohorts (N = 314) were harmonized and tested for independent associations between each of the nine 17q12-q21 risk variants, either MSH or CDSB nasal microbiota colonization, and childhood wheeze risk reported in the first three years of life. Nasal microbiota groupings were based on previously reported associations with adverse (MSH) or improved (CDSB) respiratory outcomes. None of these relationships were found to be significant in this relatively small cohort (**Figure 2.1A; E1 B**).

We next examined interactions between risk variant SNPs and MSH or CDSB upper airway microbiota. Interactions between genotype and MSH microbiota associated with increased risk, while interactions with CDSB microbiota decreased risk (**Figure 2.1B**). The

rs8076131 SNP in *ORMDL3* and rs2305480 in *GSDMB* exhibited the strongest interactions with MSH (OR = 1.72, CI = 1.09-2.71, P_{int} = 0.031 and OR = 1.72, CI = 1.09-2.71, P_{int} = 0.042, respectively) with 5 other SNPs also exhibiting interactions (P_{int} = 0.035), likely due to strong linkage equilibrium in the 17q12-q21 region in populations with European ancestry. Conversely, CDSB microbiota interactions with these same SNPs decreased risk with rs8076131 (OR = 0.28, CI = 1.11-0.68, P_{int} = 0.031) and rs2305480 (OR = 0.33, CI = 0.14-0.79, P_{int} = 0.042) having the lowest odds ratios.

All SNPs tested for genotype by airway microbiota interactions associated with wheeze risk displayed a "flip-flop" pattern (**Figure 2.1C-I; Figure S2.1C-D**). This previously described phenomenon is associated with a gene by environment interaction in which exposure is the mediating factor for risk.[58] Thus, our data indicate that 17q12-q21 risk genotypes are associated with higher or lower risk of early-life wheeze, depending on the individual's upper airway microbiota colonization pattern. Specifically, children homozygous for risk alleles and dominated by MSH are at increased risk, whereas those dominated by CDSB have a lower probability of early-life wheeze, (**Figure 2.1C-I**).

When cohorts were considered independently, genotype by nasal microbiota interactions only reached FDR-corrected significance for *GSDMB* (rs7216389) and *ORMDL3 (*rs8076131) SNPs within the URECA cohort, though a similar trend remained evident in COAST children (**Figure 2.2A-B**). Within the URECA cohort (n = 136), the variants rs7216389 in *GSDMB* and rs8076131 in *ORMDL3* exhibited interactions with MSH and CDSB upper airway microbiota related with early-life wheeze outcome (MSH OR = 2.16, CI = 0.083-5.61, and OR = 2.34, CI =

0.88-6.22, respectively; CDSB OR = 0.23, CI = 0.05-1.14, and OR = 0.22, CI = 0.04-1.25, respectively; P_{int} = 0.027 for all; **Figure 2.2B**).

It is well-established that the 17q12-q21 region exhibits considerable linkage disequilibrium in European populations but less so in those of African American ancestry, and this is replicated within our cohorts (**Figure S2.1 E-F)**. Thus, to identify key SNP by microbiota interactions that modulate wheeze risk, we examined the African American population (n=97) exclusively, focusing on variants within *GSDMB* and *ORMDL3, genes that* govern epithelial integrity and anti-microbial responses. Within this population, interactions between MSH- or CDSB-dominated microbiota and *ORMDL3* rs8076131 (MSH OR = 3.62, CI = 1.07-12.20; CDSB OR = 0.18, CI = 0.02-1.49; P_{int} = 0.016;), *GSDMB* rs2305480 (MSH OR = 8.07, CI = 1.44-45.00; CDSB OR = 0.52, CI = 0.08-3.26; P_{int} = 0.016), and rs7216389 (MSH OR = 3.62, CI = 1.07-12.20; CDSB OR = 0.27, CI = 0.04-1.84; P_{int} = 0.016) remained significant (**Figure 2.2C**). These data indicate that associations with wheeze risk may be driven by interactions between the upper airway microbiota and one or more of these SNPs.

Given our interest in epithelial-microbial interactions and *GSDMB*'s previously reported bactericidal activity[41] we chose to focus on a single *GSDMB* SNP, rs7216389, for experimental investigation. *GSDMB* variants rs7216389 (intronic variant), rs2305480 (missense variant), and rs11078928 (splice variant not included in this study) are in strong linkage disequilibrium, even in individuals of African ancestry.[30] These SNPs are all expression quantitative loci (eQTLs) for *GSDMB* (but not *ORMDL3*) in nasal mucosal cells and for both genes in blood cells in the URECA cohort.[30] Additional findings from the PASTURE cohort indicate that *GSDMB* rs7216389 conferred risk of wheeze only in children with low cytokine response to bacterial

lipopolysaccharide.[31] Independently in the ALLIANCE cohort, the same SNP was identified as a key driver of lung mucosal *GSDMB* expression.[59] These reports coupled with our observations of a gene-by-microbiota interaction at this SNP in both the COAST and URECA cohort (**Figure 2.2D-E**), led us to examine the impact of *GSDMB* rs7216389 genotype on transcriptional activity in a lung A549 epithelial cell monolayer model. A549 lung epithelial cells that were either heterozygous at rs7216389 ($rs7216389^{CT}$) or CRISPR-edited to be homozygous for the risk allele ($rs7216389^{TT}$) were co-cultured with or without *S. epidermidis* as a representative commensal respiratory bacterial exposure before performing RNA sequencing.

An initial comparison of the $rs7216389^{CT}$ and $rs7216389^{TT}$ cell lines identified 25 differentially expressed genes (FDR-adjusted $P < 0.05$ and >2-fold change in gene expression). The majority of these (23/25) were downregulated in $rs7216389^{TT}$ epithelial cells **(Table S2.3; Figure 2.3A)**. Gene-set enrichment analysis revealed that the genes down-regulated in the $rs7216389^{TT}$ cell line were predominantly involved in expression of cell surface–associated molecules, while the nominal number of upregulated genes were involved in negative regulation of hydrogen peroxide production, an antimicrobial, with activity against bacteria, viruses and fungi (**Figure 2.3B)**.[60] Genes exhibiting the greatest degree of downregulation included those encoding for mucin production (*MUC5B, MUC5AC*),[61] an anti-inflammatory and asthma-protective neuropeptide (*NTS*)[62] and a cell adhesion molecule (*CEACAM6*) (**Figure 2.3C)**. This transcriptional profile was consistent irrespective of co-culture with *S. epidermidis* (**Figure 2.3C**), underscoring the dominant effect of this SNP on anti-microbial response genes.

Interestingly, no significant differences in expression of *GSDMB* or other 17q12-q21 genes were observed between the $rs7216389^{CT}$ and $rs7216389^{TT}$ cell lines cultured *in vitro*.

However, two genes (*COL11A1* and *PCDH9*) were downregulated both in our rs7216389TT cell line and in nasal mucosal cells from children with wheezing at preschool age (**Figure S2.2A)**.[59] Considering this lack of chromosome 17 cis-regulation within our transcriptional data, we validated the top differentially expressed genes of interest using RT-qPCR in a second rs7216389^{TT} clone and confirmed that *CEACAM6*, *MUC5AC*, and *MUC5B* were significantly downregulated in this cell line (**Figure S2.2B**). Due to the low likelihood of the same off-target effects being observed in separate clones, it is plausible that the observed transcriptional differentials are genotype-intrinsic and not off-target CRISPR effects.

Ordination of the top 500 most variably expressed genes across all experimental conditions confirmed that genotype represented the primary explanatory variable related to rs7216389^{CT} and rs7216389^{TT} transcriptional differences ($R^2 = 0.491$, $P = 6.72 \times 10^{-7}$), with bacterial exposure explaining a smaller but still significant degree of transcriptional variation ($R^2 = 0.341$, $p = 0.018$; **Figure 2.3D**). Sixteen genes were differentially expressed in the rs7216389^{CT} following *S. epidermidis* co-culture, with only nine of these reaching significance in the rs7216389^{TT} following *S. epidermidis* co-culture (**Table S2.3, Figure S2.2C-D**). CXCL1, a pro-inflammatory chemokine that recruits neutrophils[63] was among the seven genes whose expression was upregulated only in the rs7216389^{CT} cell line exposed to *S. epidermidis* and not in the rs7216389^{TT} cell line upon exposure to this microbe, indicating muted antimicrobial response to bacterial exposure in the risk variant rs7216389^{TT} cell line.

We hypothesized that the observed downregulation of antimicrobial gene expression in the rs7216389^{TT} cell line leads to heightened epithelial susceptibility to microbial colonization. To assess this, we performed a microbial adherence assay using *S. epidermidis*. Adherent

bacterial cell counts normalized to A549 cell numbers were significantly higher in the rs7216389^{TT} cell line compared with rs7216389^{CT} (Welch's T-test P = 0.007; **Figure 2.3E**), indicating that the risk genotype increases epithelial microbial attachment *in vitro*. Collectively, these data indicate that the rs7216389^{TT} genotype associates with reduced expression of epithelial genes that protect from microbial colonization and permit increased microbial adherence.

Discussion

The genetics of asthma susceptibility are complex, and environmental exposures are known to influence risk.[58,64] Findings from this study indicate that childhood wheezing risk is influenced by interactions between genotype at the 17q12-q21 locus and the upper airway microbiota. Combined and independent analyses of two birth cohorts with distinct genetic makeup and environmental exposures indicate that SNPs in *GSDMB* and *ORMDL3* increase childhood wheeze risk in the context of upper airway microbiota dominated by the asthmagenic genera *Moraxella, Streptococcus,* or *Haemophilus.* Conversely, wheeze risk is reduced in children with the same risk variants and airways colonized by *Corynebacterium, Dolosigranulum, Staphylococcus, or Bacillus* at age 12 months.

Notably a *flip-flop pattern was observed* for the 17q12-q21 SNPs tested. This phenomenon has been associated with genetic variants that associate with an increased risk in one environment and decreased risk in another.[58] Our data indicate that microbiota colonization of the upper airways underlie this observation at the 17q12-q21 locus. Childhood wheezing risk significantly increases in subjects carrying 17q12-q21 risk alleles when *MSH microbiota* colonize their airways. In sharp contrast, these same alleles are protective in the

context of airway colonization by *CDSB* microbiota. Members of the dominant genera of CDSB microbiota are known to have protective effects; *Dolosigranulum pigrim* inhibits growth of pathogenic *Staphylococcus aureus* and, in synergy with *Corynebacterium pseudodiphtheriticum*, robustly inhibits growth of *S. pneumoniae in vitro*.[65] This suggests that commensal bacterial colonization of the airway epithelium of children at genetic risk for wheezing in early life may reduce disease incidence by preventing epithelial-damaging microbial pathogen colonization of the upper airways.

Transcriptional differentials between *GSDMB* rs7216389^{CT} and rs7216389^{TT} cell lines provide further evidence that this risk SNP confers susceptibility to pathogenic microbes and offers insights into how this variant may increase risk of childhood wheeze. The rs7216389^{TT} lung epithelial cell lines had decreased expression of genes encoding neurotensin (*NTS*), adhesion molecules (*CEACAM5* and *CEACAM6*)[66] and epithelial protective mucins (*MUC5B* and *MUC5AC*)[67] compared with the rs7216389^{CT} cell line. In addition to its anti-inflammatory and asthma-protective effects,[62] neurotensin stimulates neutrophil adherence to bronchial epithelial cells and modulates alveolar macrophage production of interleukin-1.[68] *CEACAM5 and CEACAM6* are both cell adhesion molecules, with *CEACAM6* having an established role in binding respiratory bacteria.[66] *MUC5B* and *MUC5AC* encode production of lung mucins that contribute to epithelial protection and control of microbial populations in the airways.[69,70]

Muc5b has been shown in a murine model to be essential for reducing the number of culturable bacteria in the lung, emphasizing its importance in regulating airway microbial burden.[69] More specifically, in the context of *S. epidermidis* co-culture, we observed reduced transcription of the neutrophil chemoattractant gene *CXCL1* in the rs7216389^{TT} cell line

compared with rs7216389^{CT}, compounding reduced capacity for neutrophil recruitment in response to microbes by the rs7216389^{TT} epithelium. *CXCL1* is produced in greater amounts by airway smooth muscle from non-asthmatic individuals upon stimulation,[71] and neutrophils are among the first cellular responders to microbial infection, facilitating bacterial pathogen killing via phagocytosis and reactive oxygen species generation.[63]

Thus, the association between the *GSDMB* rs7216389^{TT} genotype and lack of upregulation of multiple genes involved in epithelial protection and antimicrobial immunity indicate that these epithelia may be inherently susceptible to microbial colonization and pathogenesis. This finding is consistent with our observation of increased bacterial adherence to the rs7216389^{TT} cell line, indicating that parallel mechanisms related to muted microbial clearance and increased colonization susceptibility, relate with wheezing risk in genetically susceptible children. Thus, we propose that in the context of pathogenic respiratory microbes (e.g., *Moraxella*, *Streptococcus* or *Haemophilus* species), genotypic-mediated failure to protect the epithelium and recruit antimicrobial immune cell populations enhances microbial adherence, and susceptibility to wheezing episodes. Conversely, colonization of *GSDMB* rs7216389^{TT} epithelia by non-pathogenic, commensal organisms may compensate for SNP-mediated epithelial dysfunction by serving as a secondary protective barrier against microbial pathogen colonization, thus reducing wheeze risk.

Limitations of the study include evaluation of the impact of genotype on epithelial transcriptional profiles for only one of the risk SNPs and only using an *in vitro* A549 cell culture model. A549 cells predominantly exhibit alveolar epithelial (AT2-like) characteristics and therefore differ from pseudostratified ciliated nasal or bronchial epithelium.[72] Thus, future

studies leveraging upper airway cell lines that more faithfully recapitulate the pseudostratified ciliated nasal or bronchial epithelium, or indeed primary nasal epithelial cultures may offer further insights. Additionally, other SNPs, including those in *OMRDL3,* also merit similar dissection to better understand their impact on epithelial physiology and contribution to wheeze outcomes. Moreover, it is likely that more complex microbial-host-exposome interactions occur *in vivo*, leading to more expansive effects beyond those observed in our cell-line model. Nonetheless, the combined observations in humans and model systems suggest that genetically encoded risk of childhood wheeze is modulated by the upper airway microbiota and that decreasing respiratory pathogenic bacterial colonization may reduce wheeze incidence in genetically susceptible children.

Figures

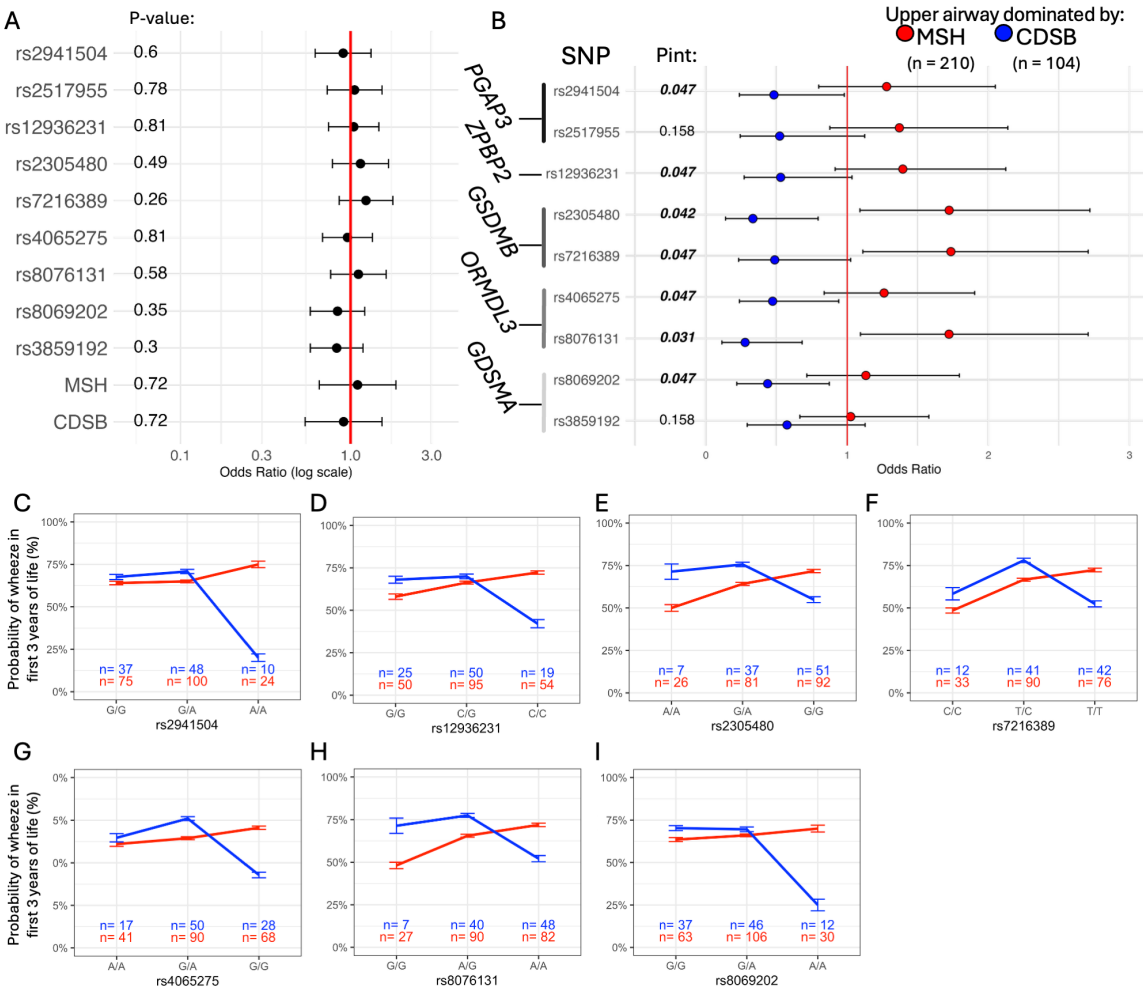

Figure 2.1. SNPs in the 17q12-q21 region interact with upper airway microbiota to influence risk of early-life wheeze.

A. Odds ratio (OR) of wheeze risk within the first three years of life based on each of the nine 17q12-q21 SNPS, and upper airway microbiota dominated by Moraxella, Streptococcus, or Haemophilus (MSH); or Corynebacterium, Dolosigranulum, Staphylococcus, or Bacillus (CDSB). 95% confidence intervals (CI), unadjusted Pint, and red line indicating the reference are displayed. SNPs are presented in the order in which they are present in the 17q12-q21 region. B. OR for wheeze risk within the first three years of life in the combined URECA and COAST cohorts based on interactions between 17q12-q21 and upper airway microbiota colonization by MSH (red) or CSBD (blue). Pints are presented to the left of the stratified effect estimates. C-I. Probabilities of wheeze in the first three years of life stratified by MSH (red) or CDSB (blue) (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) airway microbiota in the combined cohorts for each SNP of interest with 95% CI. Risk allele on the right of each plot. Sample sizes provided beneath each point.

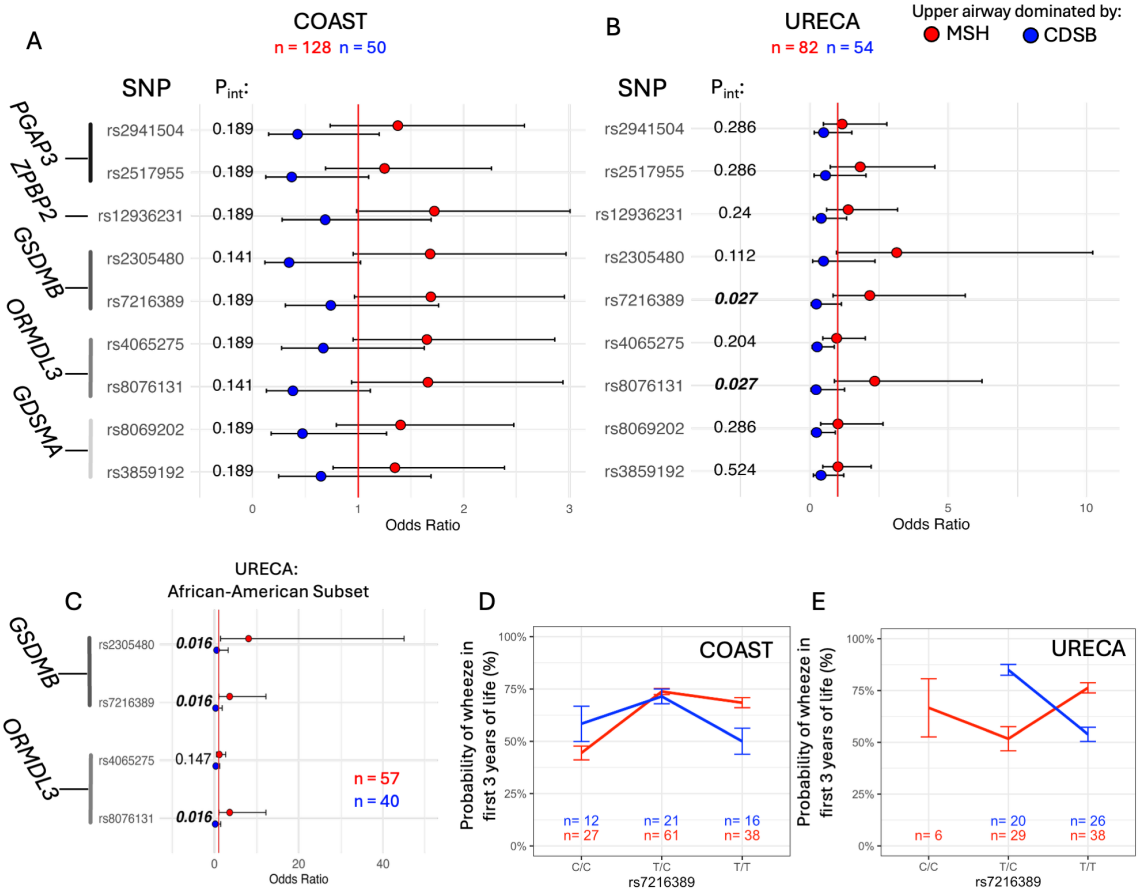

Figure 2.2. SNP by microbiota interactions related to increased risk of early-life wheeze are maintained in independent cohorts.

A-B. Odds ratio (OR) of wheeze risk within the first three years of life broken down by COAST (A) and URECA (B) cohorts and stratified by airway microbiota dominated by Moraxella, Streptococcus, or Haemophilus (MSH; red); or Corynebacterium, Dolosigranulum, Staphylococcus, or Bacillus (CDSB; blue). 95% confidence intervals (CI), P_{int}, and red line indicating the reference are displayed. (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) C. OR for variants within GSDMB and ORMDL3 in the African American subset of URECA. D-E. Probability of wheeze in the first three years of life by genotype for the GSDMB intronic SNP rs7216389 in COAST (D) and URECA (E).

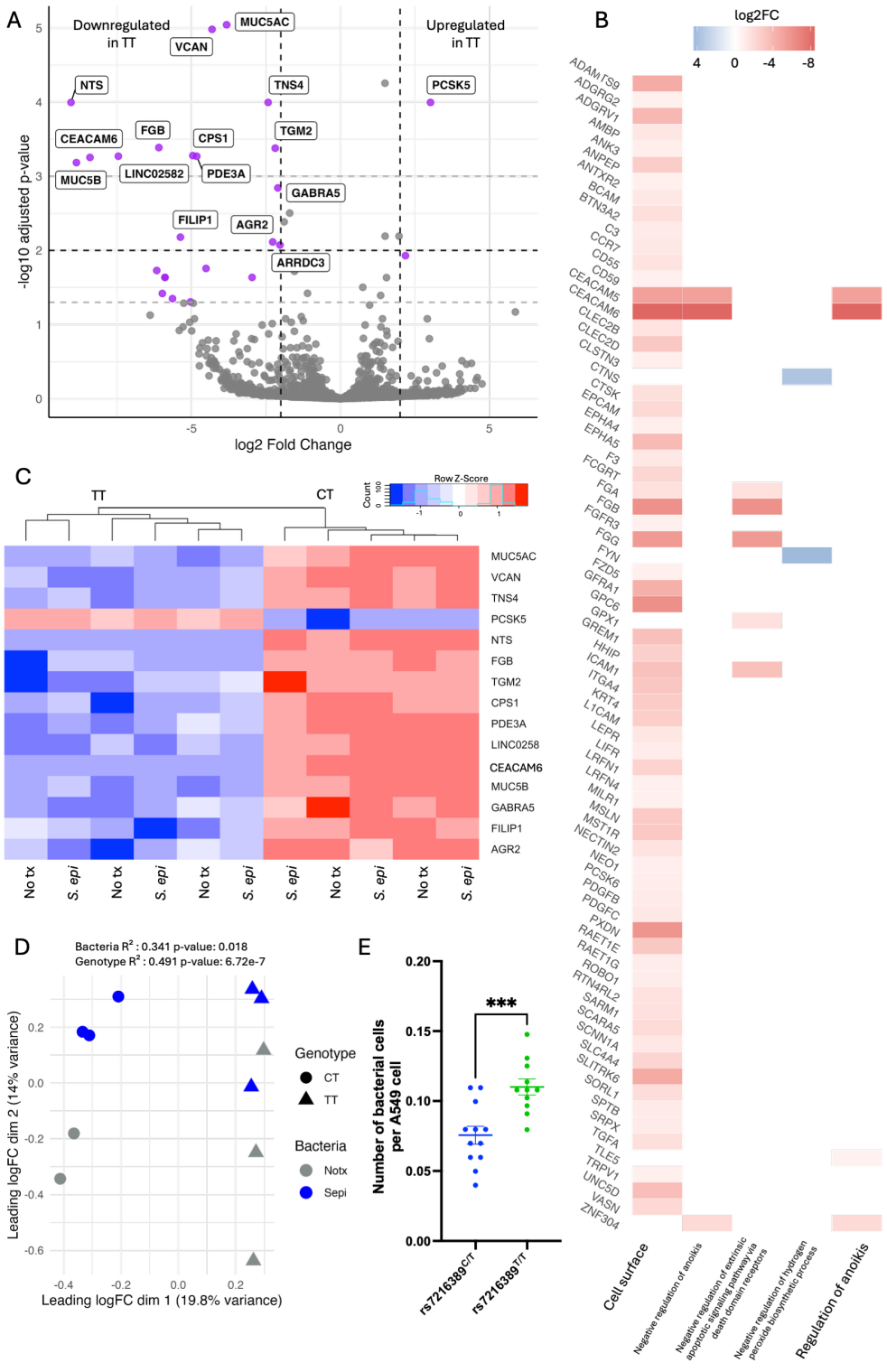

Figure 2.3. rs7216389^{TT} cell lines exhibit downregulation of genes involved in antimicrobial responses and increased susceptibility to microbial adherence.

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.)A. The majority of genes differentially expressed between rs7216389TT and rs7216389CT A549 cells are downregulated in the TT genotype. Genes with |logFC |>2 & adj. p-value < 0.05 shown in purple, and those with Padj < 0.001 labeled. B. Gene-set enrichment analysis displaying gene sets significantly enriched in rs7216389TT. GO terms: cell surface (0009986), negative regulation of anoikis (2000811), negative regulation of extrinsic apoptotic signaling pathway via death domain receptors (1902042), negative regulation of hydrogen peroxide biosynthetic process (0010730), regulation of anoikis (2000209). C. Heatmap of the top 15 most differentially expressed genes between rs7216389CT and rs7216389TT A549 cells ranked from top to bottom by Padj. No tx: No bacterial co-culture treatment, S. epi: Staphylococcus epidermidis. D. Multidimensional scaling (MDS) plot of leading log2 fold-changes based on pairwise comparisons of the top 500 most variable genes. E. S. epidermidis adherence to the rs7216389TT A549 cell line is significantly greater (Welch's T-test P = 0.007) than that of the rs7216389CT cell line.

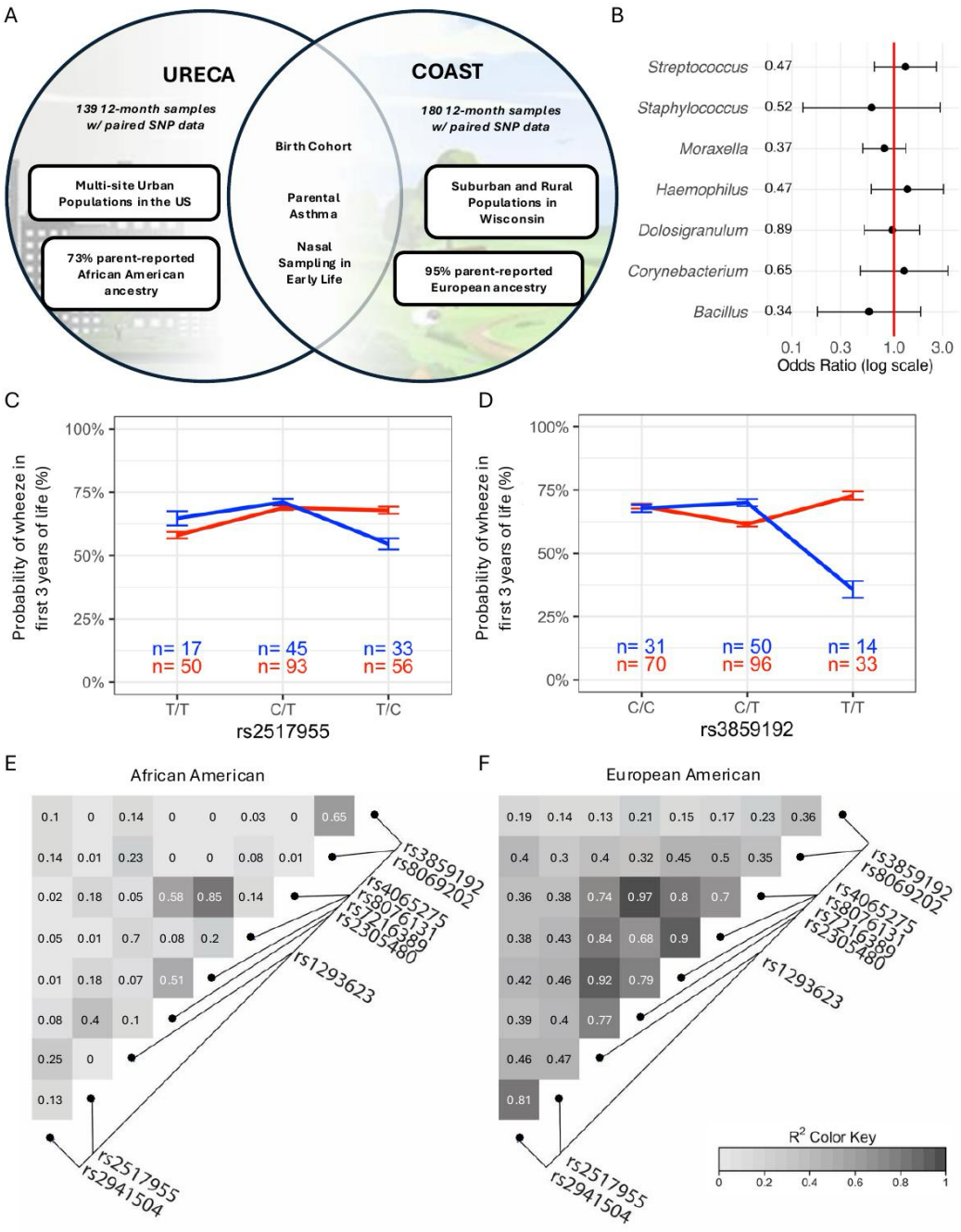

Figure S2.1. Childhood Origins of ASThma (COAST) and Urban Environment and Childhood Asthma (URECA) cohorts and additional SNP interactions.

A. COAST and URECA are high-risk birth cohorts which collected nasal samples at 1 year of life from of children born to parents with asthma; however, they have key demographic differences. (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) B. Odds ratio (OR) of wheeze risk within the first three years of life based on the dominant upper airway genera. 95% confidence intervals (CI), unadjusted P_{int}, and red line indicating the reference are displayed. C-D. Probability of wheeze in the first three years of life based on PGAP3 and GSDMA genotype (which did not reach significance in a logistical regression interaction analysis) stratified by MSH (red) and CDSB (blue) upper airway microbiota colonization. Risk allele on the right of each plot. Sample size beneath each point, 95% confidence intervals displayed. Correlation plot of the variant data present within this study demonstrating lower variant linkage in the African American population (E) compared to the European American population (F).

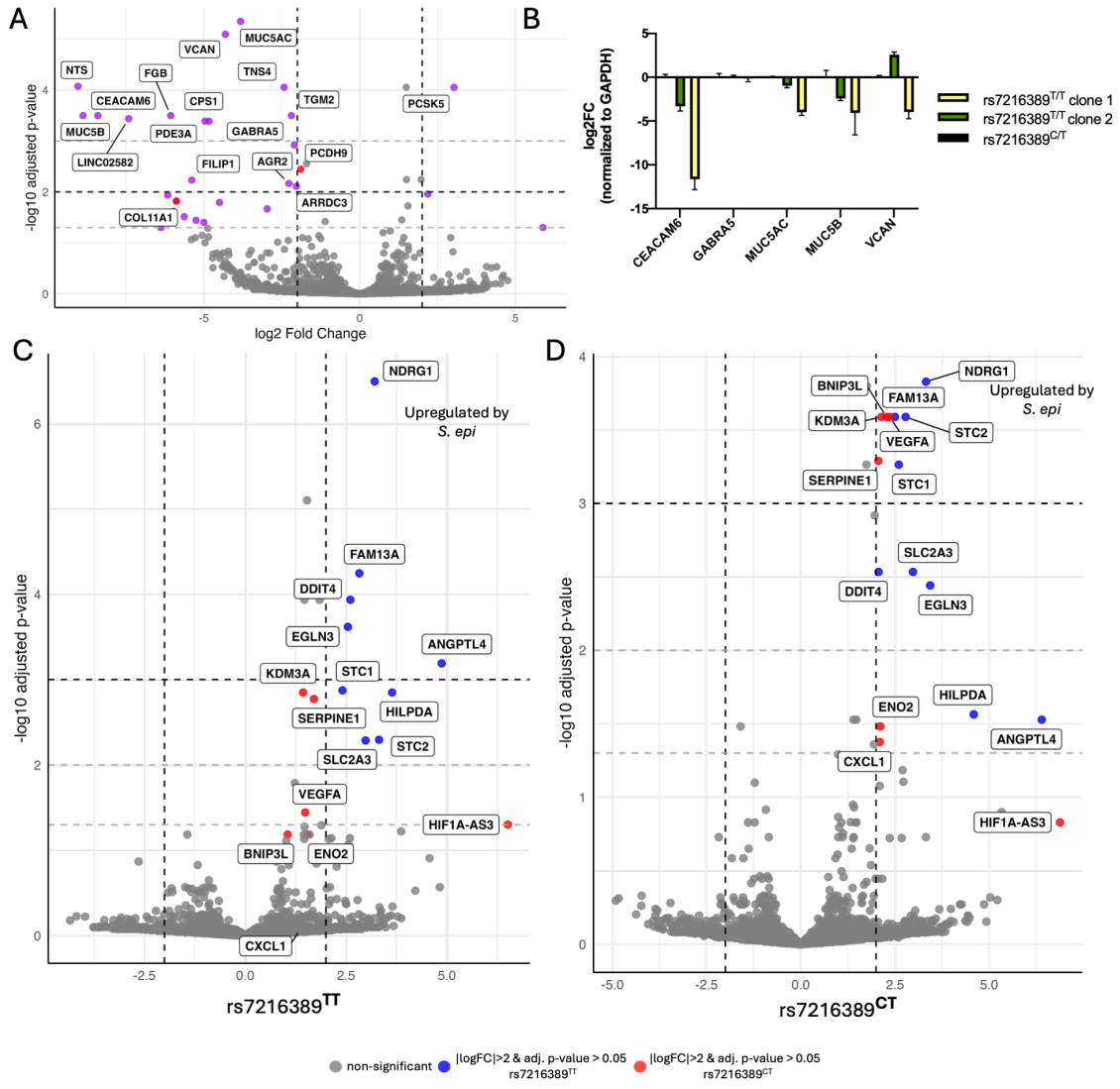

Figure S2.2. Further assessments of rs7216389^{TT} cell line.
A. Genes with |logFC|>2 & adj. p-value > 0.05 between rs7216389TT and rs7216389CT shown in purple, and those which were found to also be significant in the nasal mucosa of children
(Figure caption continued on the next page.)

(Figure caption continued from the previous page.) with wheezing at preschool age in in Jakwerth et. al[59] shown in red. B. RT-qPCR of 5 of the most differentially expressed genes in the electroporated unedited control CT, TT clone 1 (used for co-culture experiment, 100% editing efficiency) and TT clone 2 (98% editing efficiency) normalized to GAPDH and log2FC calculated with CT as reference. Genes significantly differentially expressed upon S. epidermidis exposure compared to no treatment in C. rs7216389TT or D. rs7216389CT cell lines. Genes which are uniquely significantly different in the rs7216389CT cell line are red, and those significantly different in both genotypes are blue.

Tables

Table 2.1. Cohort Demographics of COAST and URECA Subjects

	COAST (n = 184)	URECA (n = 142)	P
Parent-Reported Race/Ethnicity			**<0.0001**
Black, African-American	6 (3.3%)	100 (70.4%)	
Caribbean-Hispanic	0 (0.0%)	20 (14.1%)	
European-American	154 (83.7%)	1 (0.7%)	
Hispanic-Other	7 (3.8%)	0 (0.0%)	
Mexican-American	0 (0.0%)	1 (0.7%)	
Multiple	10 (5.4%)	16 (11.3%)	
Other	3 (1.6%)	1 (0.7%)	
Missing	4 (2.2%)	3 (2.1%)	
Age at Sample Collection, months	12.1 (±0.4)	12.4 (±1.1)	**0.013**
Child Ill at Clinic Visit			0.16
No	111 (60.3%)	97 (68.3%)	
Yes	73 (39.7%)	45 (31.7%)	
Sex			0.22
Female	83 (45.1%)	74 (52.1%)	
Male	101 (54.9%)	68 (47.9%)	
Smoking During Pregnancy			**0.0002**
No	169 (91.8%)	111 (78.2%)	
Yes	11 (6.0%)	28 (19.7%)	
Missing	4 (2.2%)	3 (2.1%)	
Dogs in Infancy			**<0.0001**
No	113 (61.4%)	138 (97.2%)	
Yes	67 (36.4%)	1 (0.7%)	
Missing	4 (2.2%)	3 (2.1%)	

	COAST (n = 184)	URECA (n = 142)	P
Cats in Infancy			**<0.0001**
No	125 (67.9%)	136 (95.8%)	
Yes	55 (29.9%)	2 (1.4%)	
Missing	4 (2.2%)	4 (2.8%)	
Wheeze Diagnosis Between 0 and 3 Years			0.81
No	62 (33.7%)	46 (32.4%)	
Yes	118 (64.1%)	93 (65.5%)	
Missing	4 (2.2%)	3 (2.1%)	
Dominant Bacterial Genus			**<0.0001**
Other	2 (1.1%)	4 (2.8%)	
Bacillus	0 (0.0%)	15 (10.6%)	
Corynebacterium	13 (7.1%)	8 (5.6%)	
Dolosigranulum	32 (17.4%)	29 (20.4%)	
Haemophilus	16 (8.7%)	19 (13.4%)	
Moraxella	78 (42.4%)	53 (37.3%)	
Staphylococcus	5 (2.7%)	3 (2.1%)	
Streptococcus	38 (20.7%)	11 (7.7%)	
rs2941504			0.88
G/G	68 (37.0%)	55 (38.7%)	
G/A	95 (51.6%)	69 (48.6%)	
A/A	21 (11.4%)	18 (12.7%)	
rs2517955			**<0.0001**
T/T	56 (30.4%)	15 (10.6%)	
C/T	100 (54.3%)	53 (37.3%)	
C/C	28 (15.2%)	74 (52.1%)	
rs12936231			0.16
G/G	41 (22.3%)	44 (31.0%)	
C/G	90 (48.9%)	67 (47.2%)	
C/C	52 (28.3%)	31 (21.8%)	
Missing	1 (0.5%)	0 (0.0%)	
rs2305480			**<0.0001**
A/A	30 (16.3%)	6 (4.2%)	
G/A	87 (47.3%)	40 (28.2%)	
G/G	67 (36.4%)	96 (67.6%)	

	COAST (n = 184)	URECA (n = 142)	P
rs7216389			**<0.0001**
C/C	40 (21.7%)	10 (7.0%)	
T/C	87 (47.3%)	54 (38.0%)	
T/T	57 (31.0%)	78 (54.9%)	
rs4065275			0.82
A/A	37 (20.1%)	26 (18.3%)	
G/A	85 (46.2%)	71 (50.0%)	
G/G	62 (33.7%)	45 (31.7%)	
rs8076131			**<0.0001**
G/G	30 (16.3%)	7 (4.9%)	
A/G	89 (48.4%)	53 (37.3%)	
A/A	65 (35.3%)	82 (57.7%)	
rs8069202			**0.0009**
G/G	50 (27.2%)	61 (43.0%)	
G/A	98 (53.3%)	70 (49.3%)	
A/A	36 (19.6%)	11 (7.7%)	
rs3859192			**0.004**
C/C	52 (28.3%)	64 (45.1%)	
C/T	98 (53.3%)	62 (43.7%)	
T/T	34 (18.5%)	15 (10.6%)	
Missing	0 (0.0%)	1 (0.7%)	

Table S2.1. Single-Nucleotide Polymorphisms (SNPs) Assessed in This Study

SNP	Gene	SNP Type
rs2941504	PGAP3	Synonymous (c.465G>A, p.Val104)
rs2517955	PGAP2	Intron (C>T)
rs12936231	ZPBP2	Intron (G>C)
rs2305480	GSDMB	Missense (c.892G>A, p.Pro298Ser)
rs7216389	GSDMB	Intron (T>C)
rs4065275	ORMDL3	Intron (A>G)
rs8076131	ORMDL3	Intron (G>A)
rs8069202	GSDMA	Intron (G>A)
rs3859192	GSDMA	Intron (C>T)

Table S2.2. RT-qPCR Primers Used in This Study

Gene	Forward	Reverse
CEACAM6	GGGACGTTCCAGCAATCCAC	ACAGGAGCACTTCCAGAGACT
GABRA5	GTCCGACACGGAAATGGAGT	ATGGGCCCCTTAAACCGAAG
MUC5AC	TATACCGCCACAGAGACCTCGC	CGGGCTCGGAGGTGGATATTGA
MUC5B	GAATGAGAACTACGCCCGGCAC	TGGGGTTGATGATGGAGTGGCA
VCAN	ATTCAGGTGCCTCTGCCTTC	CCTTGGAATTTGTGCCAGCC
GAPDH	TTGGCTACAGCAACAGGGTG	GGGGAGATTCAGTGTGGTGG

Chapter 3: Carbohydrate Metabolism Differs in Infants by Asthma-Risk Status and us Associated with the Functional Potential of *Bacteroides Cellulosilyticus*

Abstract

Childhood atopic disease is linked to delayed gut microbiome development and metabolic dysfunction, however microbial drivers remain unclear. To explore microbial correlates of asthma risk during a time of active gut microbiome development, we analyzed stool from 6-month-old infants at high asthma risk (HR) or healthy controls (HC), using Genome-resolved metagenomics (HR=7; HC=12) and untargeted metabolomics (HR=11; HC=15). We recovered 82 bacterial species-level metagenomic-assembled genomes (MAGs). Global Taxonomic composition did not differ by asthma risk. Anticipating that key differences might associate with specific genomes, a machine-learning approach pinpointed *Bacteroides cellulosilyticus*, *Hungatella effluvii,* and *Enterocloster aldenensis* as linked with asthma risk status. All three species were more abundant in HC infants and the *B. cellulosilyticus* genome was enriched for carbohydrate metabolism genes relative to other MAGs. Metabolomic profiling revealed variance associated with asthma risk (PERMANOVA, R^2=0.069, p=0.016). HR fecal metabolomes were enriched in simple sugars, whereas HC contained more nitrogenous compounds. Integrative genome-metabolic modeling of compounds that significantly differentiate asthma-risk groups revealed risk-dependent interactions with community-encoded metabolic potential (CEP), for arabinose and agmatine, whose fecal concentrations are

linked with *B. cellulosilyticus* and *H. effluvii* functional traits respectively. These findings suggest that microbial-influenced metabolic differences associate with asthma risk at 6 months, with *B. cellulosilyticus* and *H. effluvii* emerging as candidate bacteria influencing this observed metabolic remodeling.

Impact statement

Leveraging a random forest classifier, we identified three bacterial species (*Bacteroides cellulosilyticus*, *Hungatella effluvii,* and *Enterocloster aldenensis)* as distinguishing features enriched in healthy 6-month old infant microbiomes compared to those at high risk of asthma development (HR). We developed an approach to integrate metabolomics and metagenomic-derived microbiome community encoded potential (CEP) with clinical outcomes to identify fecal metabolites whose concentrations are likely to be influenced by the microbiome. Fecal arabinose concentrations were positively associated with CEP in healthy infants, but not in HR subjects who exhibited elevated concentrations irrespective of CEP. These data implicate microbial activity as a contributor to the concentration of this metabolite in healthy but not HR infants. With a leave-one-out-cross-validation, we identified *B. cellulosilyticus* as a contributor to fecal arabinose concentrations. Our data indicate that microbial functional deficits in HR infants is associated with altered gut metabolic dysfunction during microbiome maturation.

Data summary

Durack et. al[73] is the source of the metabolomics data utilized in this study. The authors confirm that all other supporting data, code and protocols have been provided within the article or through supplementary data files.

Introduction

Atopic diseases, such as allergic asthma, are amongst the most prevalent chronic inflammatory diseases of childhood and are frequently characterized by Th2 inflammation and elevated circulating immunoglobulin E (IgE) concentrations.[74] Although risk alleles associated with atopy and asthma have been identified, the rapid increase in prevalence, particularly in industrialized nations, cannot be explained by host genetics alone.[75] Several studies have linked early life fecal microbiota perturbation and metabolic dysfunction with increased risk of atopy and asthma in childhood,[18,50,73,76] though little is known of the gut microbiome functional traits that characterize asthma risk, particularly in later infancy.

We previously compared fecal samples collected longitudinally over the first year of life from healthy control (HC) and high-risk for asthma (HR; placebo arm) infants in the Trial of Infant Probiotic Supplementation (TIPS).[73,77] HR infants exhibited delayed fecal microbiota diversification and a significantly distinct stool metabolic profile at 6 and 12 months of age. [73,77] Due to the increased number of significantly differential metabolites at 6 months compared to 12 months,[73] the rapid diversification that typically happens at this age when solid food introduction typically occurs,[78] and the absence of infant gut metagenomics at 6 months in the current asthma literature, we chose to focus on this critical timepoint. Comparative analysis of asthma-risk groups indicated a significant difference in carbohydrate metabolism characterized by enrichment of simple sugars in the feces of HR compared with HC subjects at 6 months of age. Altered carbohydrate metabolism has also been described in younger (1 month-old) infants at high-risk for atopy, indicating that it represents a consistent feature of those at higher risk of childhood atopy development.[50] However, which microbes contribute to carbohydrate

metabolism in the gut microbiome of HR and HC infants is largely unknown. Here we leverage fecal samples collected at 6 months of age from HR and HC infants and apply a multi-omics (genome-resolved metagenomics and metabolomics) approach to characterize the abundance and encoded genomic functions of the infant gut microbiome that differ between HR and HC infants with a specific focus on those fecal microbial species that contribute to altered carbohydrate metabolism.

Methods

Study Population

Stool samples for this study were obtained from the Trial of Infant Probiotic Supplementation (TIPS – ClincialTrials.gov ID: NCT00113659) clinical trial and Development of Infant Microbial Evolution (DIMES) birth cohort. Recruitment for TIPS was carried out in the San Francisco Bay Area and consisted of infants at high-risk of childhood asthma development based on an asthma diagnosis of one or both parents.[77] Infants were randomized in this double blind placebo controlled trial of daily oral *Lactobacillus rhamnosus* GG supplementation, and stool samples were collected at birth, 1, 3, 6, 9, and 12-months of age. Stool samples collected at 6 months of age from placebo treated infants in the TIPS trial served as our high-risk (HR) population for this study. DIMES also recruited healthy infants (e.g., no history of asthma, eczema or rhinitis) from the Bay Area. DIMES stool samples collected at 6 months of age served as our healthy control (HC) population for this study. This pilot study focused on fecal samples collected at 6-months of age because of paired extant untargeted metabolomics data available for many of these samples,[73] the increased number of significantly differential metabolites at

this timepoint compared to 12 months,[73] and the rapid diversification that typically happens at this critical age when solid food introduction typically occurs.[78] A total of 15 HC and 11 HR samples had available metabolomics for analysis, and as some samples had been fully consumed in earlier experiments, genome resolved metagenomic analysis could only be conducted on 12 HC and 7 HR samples. Across all samples evaluated in this work, 10 HC and 7 HR possessed paired metagenomic and metabolomic datasets (61% of samples).

DNA Extraction

DNA was extracted from 6-month old infant stool samples. Frozen samples were maintained on dry ice while a 4 mm punch biopsy with plunger (VWR International) was used to aliquot 0.3 grams of sample into a Lysing Matrix E tube (LME; MP Biomedicals) with 500 µL hexadecyltrimethylammonium bromide (CTAB, Sigma-Aldrich) and 500 µl phenol:chloroform:isoamyl-alcohol (25:24:1, Sigma-Aldrich). Samples were incubated at 65 ºC for 15 minutes. Following incubation, samples were homogenized in a Prep-24 homogenizer (MP Biomedicals) at 5.5 m/s for 30 seconds and centrifuged at 16,000 x g for 5 minutes at 4 ºC. The aqueous phase was transferred to sterile 1.5 mL tubes. An additional 500 µL of CTAB buffer was added to the LME tubes and incubation, homogenization, and centrifugation steps were repeated. The aqueous phases from both extractions were combined in a sterile 1.5 mL tube. An equal volume of chloroform was mixed with each sample, followed by centrifugation at 16,000 x g for 10 minutes at 4 ºC. The aqueous phase (600 µL) was transferred to a clean tube, combined with 2 volumes (1,200 µL) of polyethylene glycol (PEG, Sigma-Aldrich) and stored overnight at 4 ºC to precipitate DNA. Microcentrifuge tubes were centrifuged for 10 min at 16,000 x g, prior to washing the resulting DNA pellets with 300 µL of ice-cold 70% ethanol.

Pellets were air-dried for 10 minutes and re-suspended in 100 µL of sterile water. DNA from stool samples was quantified using the Qubit dsDNA Broad Range Assay Kit, diluted to 10 ng μL^{-1}.

Metagenomic Sequencing

Purified total DNA from the extracted 12 HC and 7 HR infant samples was sent to the Vincent J. Coates Genomic Sequencing Laboratory at the California institute for Quantitative Biosciences. DNA was fragmented and 150 bp x 2 (paired-end) libraries were constructed and sequenced on an Illumina HiSeq 4000 instrument to an average depth of 10M total reads per sample. (www.qb3.berkeley.edu/gsl).

Metabolomics

Extant untargeted metabolomics data previously generated in Durack *et al*[73] was used for this study. Briefly, 200 mg of stool samples from patients collected at 6 months (n=15 HC, n=11 HR) underwent ultrahigh performance liquid chromatography/tandem mass spectrometry (UPLC–MS/MS) and gas chromatography–mass spectrometry (GC–MS) by Metabolon (Durham, NC), using their standard protocol (http://www.metabolon.com/). Compounds were compared to Metabolon's in-house library of purified standards, which includes more than 3,300 commercially available compounds.

Metagenomics Analysis

Read Processing, Assembly and Binning. Raw metagenomic reads were processed using a Snakemake-based workflow. In brief, reads were first assessed using FastQC (v0.11.09),[79] followed by adapter removal and quality trimming to a mean of Q20 using BBDuk (v39.10)[80]

and Sickle (v1.33).[81] Assemblies for individual samples were generated with IDBA-UD (v1.1.3, r252)[82] using a k-mer range of 20-150 nucleotides. To improve contig recovery, three merged read sets constructed by concatenating all reads from (i) HC samples, (ii) HR samples, and (iii) all samples combined. Co-assemblies of these merged read sets were generated with IDBA-UD (v1.1.3, r252)[82] using a k-mer range of 30-150 nucleotides. A larger minimum k-mer was set for concatenated assemblies due to memory requirements. Following assembly, contigs < 1 kb were discarded. COBRA (v1.2.3)[83] was run on assembled contigs ≥ 5 kb to detect circular elements and extend contigs. Coverage of assembled contigs was evaluated by mapping sample reads back to contigs using Bowtie2 (v2.5.2).[84] Assemblies were independently binned using an automated pipeline integrating MetaBAT2 (v2.12.1),[85] MaxBin2 (v2.2.7),[86] CONCOCT (v1.1.0),[87] and VAMB (v3.0.9).[88] Contigs ≥ 2.5 kb were retained for binning; coverage of contigs in each sample vs all other samples was calculated by cross-mapping reads using SNAP (v2.0.1)[89] and a filtered coverage table was produced using jgi_summarize_bam_contig_depths from the MetaBAT2 package.[85] The best bins (MAGs) across all 4 binning methods in each sample were selected using DAS Tool (v1.1.1; score ≥0.3).[90]

Genome Aggregation, Coverage Profiling and Functional Annotation. MAGs from all samples were aggregated, quality-checked (CheckM v1.2.1; ≥60% completeness, ≤10% contamination),[91] and dereplicated with dRep (v3.4; ≥95% ANI, ≥10% overlap).[92] Taxonomy was assigned with GTDB-Tk (v2.1.1, r207).[93] Genetic codes were inferred with Codetta (v2.0),[94] ORFs predicted with Prodigal (v2.6.3),[95] and tRNAs with tRNAscan-SE (v2.0.12).[96] The final genome set consisted of 82 species-representative MAGs with no MAG sharing more than 95% average nucleotide identity (ANI) to any other representative MAG. This set of MAGs was then used as a

reference database for microbial abundance estimation. Predicted proteins were annotated using KofamScan (v1.3.0),[97] and only hits passing the adaptive thresholds were retained to build KO-based functional profiles. To assess microbial abundance and composition in each sample, reads from all samples were mapped to the species-representative MAGs database using Bowtie2 (v2.5.2)[98] to generate per-sample alignment files. Coverage statistics were then computed using CoverM (v0.7),[99] which summarized genome-level coverage and read counts within each sample. All statistical analysis were performed in R (v4.4.2) and visualizations generated using ggplot2 (v3.5.1).[100] We initially removed MAGs from our analysis that were present in ≤ 2 samples (n=20 MAGs) resulting in 62 MAGs being retained for analysis. Read counts were zero imputed with zero multiplicative replacement and normalized using center log-ratio transformation. Beta diversity was assessed using Aitchison distances calculated from CLR-transformed data. To test for differences in overall community composition between asthma risk groups, we performed a PERMANOVA (permutational multivariate analysis of variance) using the adonis2 function from the vegan R package (v2.7.1),[101] with 999 permutations.

Feature Selection, Co-occurrence, and Functional Enrichment. To identify microbes associated with asthma risk group classification, we applied the Boruta algorithm, a wrapper feature selection method built around Random Forests, using the Boruta R package (v8.0.0).[102] Feature importance scores were computed by comparing the relevance of each real feature to that of permuted shadow features across multiple randomized forests. To ensure stability of feature selection, Boruta was run across a grid of random seeds (n = 3: 11, 145, 224) and different numbers of decision trees (ntree = 1000 - 5000, in steps of 500), with a maximum of

200 iterations per run (maxRuns = 200). For each run, importance scores were extracted for all features and normalized as z-scores within each iteration. These z-scores were then aggregated across all seeds and tree counts into a unified dataset for visualization. Genomes confirmed to distinguish asthma risk group underwent a Wilcoxon rank-sum test, and p-values were adjusted for multiple testing using the Benjamini-Hochberg method. To account for class imbalance, Boruta was repeated using class-weighted Random Forests, with weights inversely proportional to class frequency implemented via the "class.weights" parameter. Microbial co-occurrence analysis was performed using the cooccur R package (v1.3).[103] A binary presence/absence matrix was generated by thresholding CLR-transformed genome abundances (presence defined as >0). Co-occurrence was assessed using a probabilistic null model (type="spp_site", thresh=TRUE, spp_names=TRUE), with species pairs having an expected co-occurrence <1 excluded from analysis. Of 1,891 possible species pairs, 700 (37.0%) were filtered, and 1,191 were tested for non-random associations. Pairs were classified as positive, negative, or random based on the statistical significance of their observed vs. expected co-occurrence ($p < 0.05$). Differential enrichment of KEGG Orthologs (KOs) between *Bacteroides cellulosilyticus* and the other 61 genomes evaluated was assessed using Fisher's Exact Test. KO count tables were filtered to exclude KOs observed in fewer than 10 genomes. *Bacteroides* KO counts were summed separately from background genomes, and 2×2 contingency tables were constructed for each KO. Fisher's Exact Test was applied to each table to test for significant differences in KO presence between groups, and p-values were adjusted for multiple testing using the Benjamini-Hochberg method.

Metabolomics Analysis

Of the 692 compounds detected in the untargeted metabolomics, 571 were present in at least half of the samples of each group (HC and HR) and retained for analysis. Metabolite intensities were normalized using probabilistic quotient normalization to correct for sample dilution, followed by $\log_2$ transformation. Missing values were imputed within phenotype groups using the quantile regression imputation of left-censored data (QRILC) algorithm under a left-censoring assumption. To control for batch effects while preserving biological variation associated with asthma risk, ComBat from the sva package (v3.54.0)[104] was applied using run day as the batch variable and the group (HC or HR) as a covariate in the design matrix.

Principal component analysis (PCA) was performed on the $\log_2$-transformed, batch-corrected data to visualize overall variance and group separation. The contribution of asthma-risk status to global metabolic variation was evaluated using PERMANOVA (adonis2, 9,999 permutations) on Euclidean distances derived from scaled metabolite intensities. Euclidean distances were chosen because scaling produces continuous, non-compositional features, and keeps the analysis similar to Aitchison distance metrics used in the metagenomic analysis.

Orthogonal Partial Least Squares Discriminant Analysis (OPLS-DA) was performed using the ropls R package (v2.38.0)[105] to identify metabolites that discriminate between HR and HC infant groups. Model performance was evaluated with 5-fold cross-validation and 9,999 label permutations (permI = 9999). The top 100 metabolites ranked by variable importance in projection (VIP) scores from the OPLS-DA model were tested for differential abundance between HR and HC groups using limma (v3.62.2).[56] Limma fits metabolite-wise linear models and employs empirical Bayes moderation to stabilize variance estimates across features.

Resulting p-values were adjusted for multiple testing using the Benjamini–Hochberg false discovery rate (FDR).

Metagenomic-Metabolomic Integrative Community Encoded Metabolic Potential (CEP)

Of the top 100 metabolites ranked by VIP, 60 had an assigned KEGG compound identifier. Among these, 39 were significantly different between HC and HR infants as tested via limma described above. 33 of these 39 significant compounds were linked to KOs represented in the metagenomic annotations. Of these 33 compounds, 26 had KEGG IDs which were present within the genomes analyzed in this study. For these 26 compounds, a genome-by-compound (GxC) weight matrix was generated by counting, per genome, the number of KO annotations linked to each compound. A genome-by-sample (GxS) matrix was generated by converting the CLR-normalized genome abundance table to sample-wise proportions by exponentiation and then rescaling each sample so that all genomes within that sample sum to 1. CEP was computed as the matrix product of GxS and GxC to cancel out genomes producing a samples-by-compound (SxC) matrix of community-encoded metabolic potential. The z-score of CEP was calculated across samples for standardization then fit to a robust rank-based regression (rfit) of measured metabolite abundance on CEP with asthma risk-group as an interaction below:

$$Y_i = \beta_0 + \beta_1 CEP_i + \beta_2 Risk_i + \beta_3(CEP_i \, x \, Risk_i) + \varepsilon_i$$

Y_i is the measured abundance for compound i. Risk is a binary indicator for the infant group (0= healthy, 1=high-risk). Each coefficient corresponds to a specific biological interpretation: β_0 is the intercept (expected metabolite abundance for healthy infants when CEP = 0), β_1 captures the CEP–metabolite association in healthy infants, β_2 captures baseline

differences between risk groups, and β_3 tests whether the CEP–metabolite relationship differs by asthma-risk group. ε_i is the error term. For each compound fit to the model above, the interaction effect (β_3) was FDR-adjusted using Benjamini-Hochberg. Given our small sample set, p_{int} CEPxRisk values <0.1 prior to FDR correction were investigated. To estimate the predictive influence of the individual genomes on the compound model, a leave-one-out-cross-validation (LOOCV) of the robust model using the full CEP predictor versus a CEP predictor missing the genome of interest was calculated and defined as Δr^2_{cv}. Genome and compound-level summaries were computed from the positive Δr^2_{cv} values.

Results

Bacteroides cellulosilyticus, Hungatella effluvii, and Enterocloster aldenensis distinguish asthma risk groups.

Principal coordinates analysis (PCoA) based on an Aitchison distance analysis of species-level genome counts did not identify significant microbiome species differences between the two risk groups (**Figure 3.1A**). However, using a permutative random forest algorithm for feature selection (boruta), several genomes significantly distinguished asthma risk groups (**Figure 3.1B**). Comparative analysis of the relative abundance of the genomes which distinguished the asthma group greater than the median of the shadow max identified three that exhibited a significant difference ($p < 0.01$) between risk groups; *Bacteroides cellulosilyticus, Hungatella effluvii*, and *Enterocloster aldenensis*. A class-weighted secondary analysis was highly concordant with 99.7% of the random forest decisions agreeing with the unweighted analysis indicating these results are robust to class imbalance (**Figure 3.1C**). All

three genomes are significantly more abundant in HC infant microbiomes (FDR p=0.012; **Figure 3.1D**). While no significant co-occurrence was detected between *B. cellulosilyticus* and *H. effluvii* or *E. aldenensis* across samples (**Figure 3.1E**), a positive association between *H. effluvii* and *E. aldenensis* was observed. This suggests that *B. cellulosilyticus* differentiates asthma risk groups independently of other species, whereas *H. effluvii* and *E. aldenensis* form a co-occurring pair with a common association pattern.

Fecal metabolomes are distinct in HR and HC infants

Unlike the species-resolved metagenomic data, fecal metabolomic profiles were significantly different in HR and HC infants (PERMANOVA; r^2=0.068, p=0.015), suggesting broad metabolic differences across these groups (**Figure 3.2A**). OPLS-DA identified the top 100 metabolites by variable importance projection (VIP) for subsequent statistical testing (**Figure 3.2B**). Sixty metabolites were found to be differentially abundant between HR and HC samples (FDR < 0.05; $|\log_2 FC| > 1$) falling into the following classes: lipids (20), nitrogenous compounds (12), carbohydrates (11), xenobiotics (12), cofactors and vitamins (3), and peptides (2). All 11 differentially abundant carbohydrates were significantly enriched in HR infant samples and 9 of 12 nitrogenous compounds were significantly enriched in the HC infant samples (**Figure 3.2C**).

Bacteroides cellulosilyticus and Hungatella effluvii influence distinct metabolite-CEP interactions based on asthma-risk

We next focused our analysis on the functional potential within the fecal microbiome which, we hypothesized, influences the infant gut metabolome. Glycan degradation genes (*FUCA*, *xynB*, and *bglX)* and glycan import genes (*SusC and SusD*) were significantly more

abundant (FDR < 0.05) in *B. cellulosilyticus* genomes compared to all other MAGs in this analysis (**Figure 3.2D**). To better understand which fecal metabolites may be influenced by these microbes we developed and deployed an integrative metagenomic-metabolomic analytical approach. Community encoded metabolic potential (CEP) for a particular metabolite was calculated from KO function abundance data (stratified by genome) and correlated with the observed metabolite abundance. We performed this analysis on the subset of data with paired metagenomic and metabolomic samples (HR=7, HC=10). Of the 26 metabolites assessed, interactions between agmatine (p_{int} = 0.0529) and arabinose (p_{int} = 0.0598) with asthma-risk passed the p_{int} significance threshold, however none of these findings remained significant following multiple testing correction ($p_{int.adj}$ = 0.6539 for all; **Figure 3.3A-D**).

Further analysis of these relationships indicated that agmatine is negatively associated with CEP in HC samples but positively associated with CEP in HR samples (**Figure 3.3A)**. Fecal arabinose concentrations were found to be positively associated with CEP in HC samples implicating microbial activity as a key contributor to the concentration of this metabolite in these infants. Notably, we found that arabinose is maintained at a relatively higher concentration in HR infants regardless of CEP, implicating dietary sources of this sugar in high-risk infants (**Figure 3.3C)**.

The primary contributors to the CEP x Compound models for both agmatine and arabinose are organisms that are present in both asthma risk groups (*B. fragilis* and *P. vulgatis*; **Figure 3.3B,D,H**). However, both *H. effluvii and B. cellulosilyticus*, rank highly amongst the genomes that contribute to the variance in fecal agmatine and arabinose concentrations. *H. effluvii* is the 8th most important contributor to agmatine x CEP model variance and *B.*

cellulosilyticus is the 5th most important contributor to arabinose x CEP model variance (**Figure 3.3B, D**). This implies that despite a small sample size and uneven distribution of these organisms, they are important contributors to the abundance of these fecal metabolites, particularly in healthy infants (**Figure 3.3A-D;H)**. Furthermore, we specifically evaluated which compound's abundance and CEP variance was most explained by the genomic capacity of *B. cellulosilyticus, H. effluvii*, and *E. aldenensis*. *B. cellulosilyticus* primarily explained carbohydrate variance (glycerolphosphoenthanolamine, lactose, glucose, arabinose, fructose; **Figure 3.3E**), while *H. effluvii* and *E. aldenensis* associated with nitrogenous compounds (**Figure 3.3F -** *H. effluvii*: xanthine, 5-aminovulinate, nicotinate ribonucleoside, ornithine, and agmatine; **Figure 3.3G -** *E. aldenensis*: isoleucine and leucine).

Discussion

The principal factors influencing infant gut microbial community structure at 6 months of age are dietary exposures, including breastfeeding, formula feeding, and the introduction of solid foods.[78] These dominant influences may mask more subtle ecological or functional differences associated with early-life risk factors for childhood asthma development. Consequently, large-scale shifts in overall microbiome composition are not expected to distinguish asthma risk groups, a pattern observed both in our study and in prior investigations.[73,106] Recognizing this, our study aimed to detect more nuanced early-life microbial determinants or correlates of asthma risk. This approach acknowledges that early-life microbial risk factors may operate through discrete microbes, metabolic outputs, or host-microbe interactions, rather than broad community-level restructuring.

Using a feature-selection approach to identify the organisms that most strongly differentiated HR and HC samples, we identified three microbes (*Bacteroides cellulosilyticus, Hungatella effluvii, and Enterocloster aldenensis)* which were associated with asthma risk and enriched in those infants at lower-risk suggesting they may play a role in asthma protection. *B. cellulosilyticus* has been reported to have anti-inflammatory effects in a gnotobiotic colitis mouse model reducing IL-6, TNF-α, and PI3K/Akt phosphorylation.[107] Furthermore, *B. cellulosilyticus* is capable of utilizing a wide array of carbohydrates including glucose, arabinose, N-acetylgalactosamine, N-acetylglucosamine, mannose, and fructose,[108] all of which were depleted in the stool metabolome of HC compared with HR infants in our study. Our integrative analysis of metabolomic data and CEP also points to *B. cellulosilyticus* as an important contributor to fecal carbohydrate relative abundances, as its largest related influences are on carbohydrate metabolites. More broadly, we observed the CEP had a positive association with arabinose abundance in low-risk infants and minimal association with the relative abundance of this sugar in those at high-risk. This suggests that the gut microbiome may differentially processes substrates in a health status-specific context, and that factors not captured in this study such as dietary differences may play a significant role in dictating these microbial interactions. *B. cellulosilyticus*, *H. effluvii,* and *E. aldenensis* were depleted in HR infant fecal microbiomes. *B. cellulosilyticus* is enriched for genes encoding enzymes that utilize the specific sugars found in higher concentrations in HR infants, indicating that the loss of these species and their encoded functional traits may contribute to the excess sugars observed in the feces of these infants.

H. effluvii and *E. aldenensis* are less well-studied organisms, and the associations between these species and nitrogenous compounds suggests may be linked to the variation in these compounds, several of which have known roles in inflammation. [109] Agmatine, which was relatively enriched in the HR infant stool metabolome is known to be produced by commensal organisms and has been implicated in inflammatory processes.[109] While the precise impact of agmatine on early-life immune development remains unclear, our findings pinpoint *H. effluvii* and *E. aldenensis* as contributors to nitrogenous compound metabolism in the gut microbiome of HC infants. Our data suggests that loss of these species is associated with increased concentrations of agmatine in the gut of infants at higher risk of childhood asthma development.

While this study moves beyond exclusively evaluating taxonomic or metabolite-level associations with asthma risk, and points to *B. cellulosilyticus, H. effluvii,* and *E. aldenensis* as organisms associated with lower asthma risk, several limitations should be noted. Due to utilizing previously processed samples, the sample size and sample overlap between metagenomic and metabolomic datasets was limited. Sample size limitations become apparent in CEP modeling where a large number of features and co-variates are evaluated and findings fail to pass our FDR threshold. Furthermore, the current annotations available in the KEGG database, which was used for both gene enrichment analysis and CEP modeling, poorly cover pathways such as microbial lipid metabolism that are known to be important in mediating inflammation and asthma risk.[110,111] Future work should include a significantly larger sample size of paired metagenomic and metabolomic datasets. Also, evaluating clinical incidence of disease rather than disease risk based on parental presence of asthma would be preferable.

Nonetheless, this proof-of-concept study identifies specific bacteria that both have the encoded potential to interact with the concentrations of several carbohydrate and nitrogenous compounds that differentiate HR and HC infants. Finally, our results from CEP modeling indicate that more complex interactions between encoded microbial metabolic potential and metabolite abundances exist, suggesting that additional factors such as diet and host metabolic capacity may influence metabolite concentrations, particularly in those at higher disease risk.

Figures

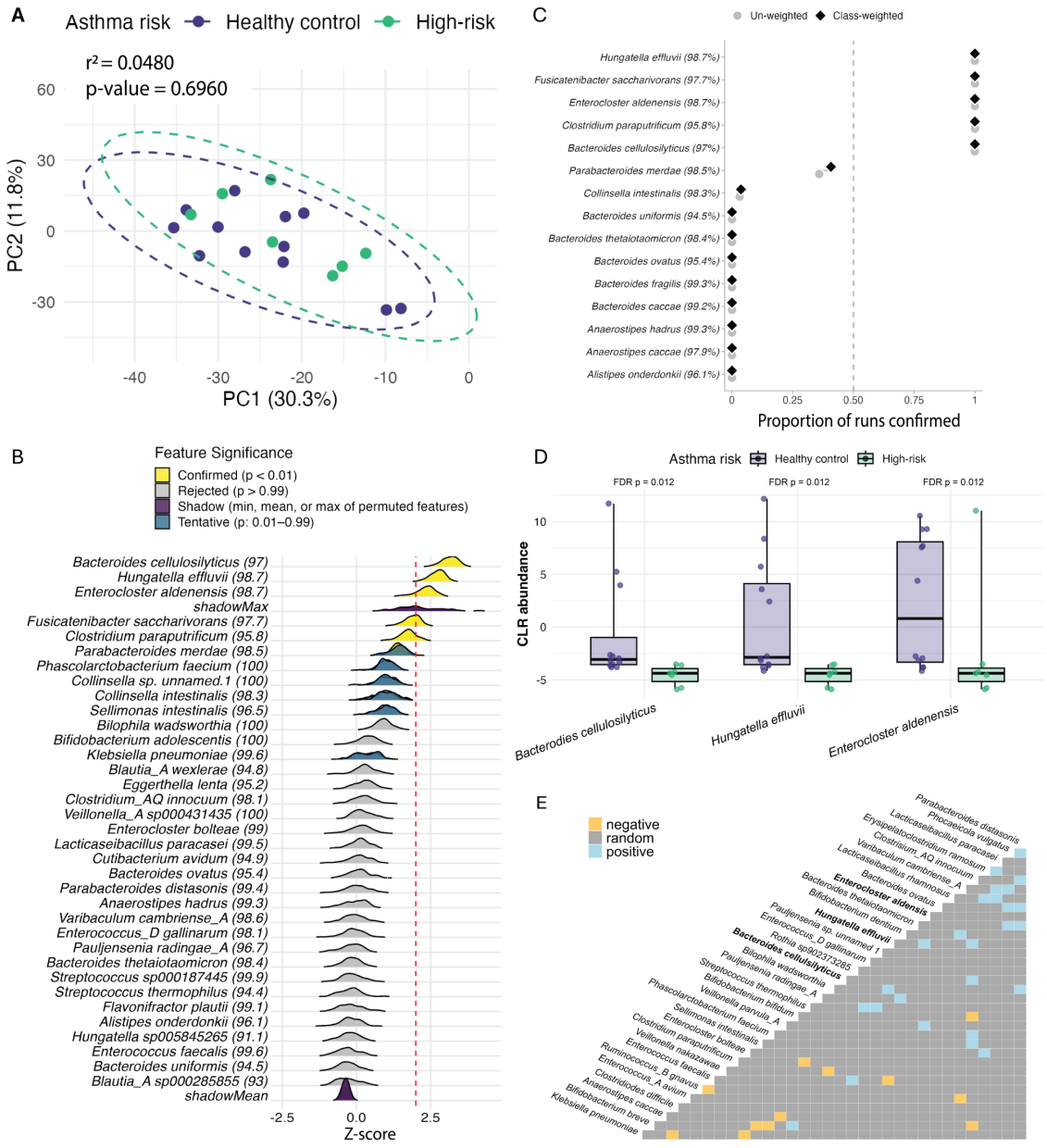

Figure 3.1. Bacteroides cellulosilyticus, Hungatella effluvii and Enterocloster Aldenensis are present at higher abundances in 6 month-old infants at low-risk for asthma.
A) PCoA of Aitchison distances of CLR-transformed species-level genome abundances. B) Z-scores of mean feature importance to asthma risk status across multiple Boruta runs with varying random seeds and numbers of trees (1,000–5,000) for each microbial species. Features are ranked by median importance and colored by their final Boruta decision. C) Proportion of runs confirming the genome as a distinguishing feature between the un-weighted and weighted Boruta analysis. (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) D) CLR abundance of Bacteroides cellulosilyticus, Hungatella effluvii, and Enterocloster aldenensis by asthma risk status. E) Pairwise co-occurrence matrix of genomes. Colors indicate positive, negative, and non-significant associations.

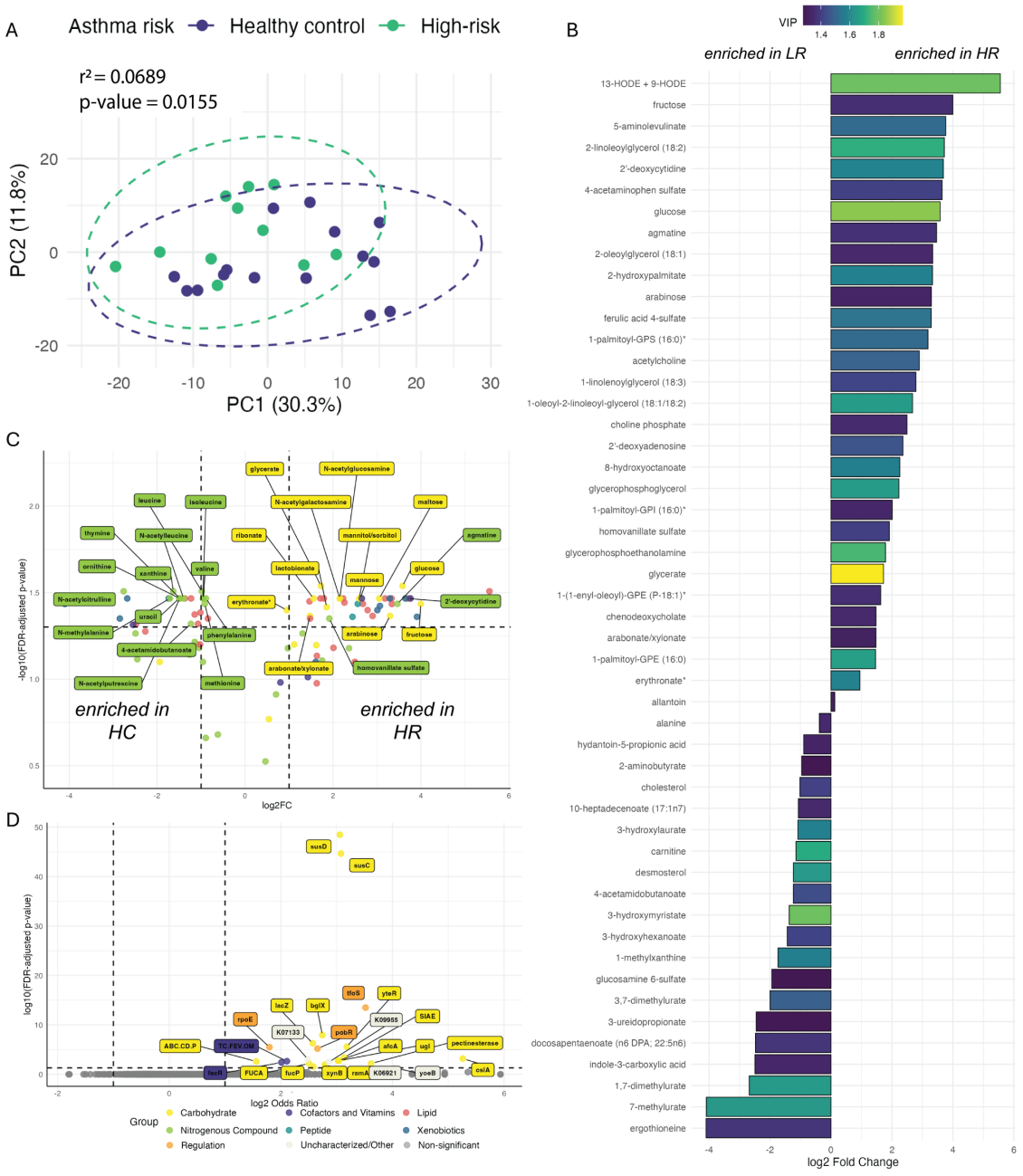

Figure 3.2. Infants at high-risk for asthma have a stool metabolome characterized by increased simple carbohydrates, and the Bacteroides cellulosilyticus genome is enriched for carbohydrate-related genes.

A) PCA of auto-scaled log-normalized stool metabolomics data. B) Top 50 metabolites predictive of asthma risk status colored by Variable Importance in Projection (VIP). C) Statistical significance and log-fold change of the top 100 compounds identified by OPLS-DA. Carbohydrates and nitrogenous compounds with significantly differential abundance (ANOVA FDR-adjusted P-value < 0.05) are labeled. (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) D) Genes significantly enriched in the *Bacteroides cellulosilyticus* genome compared to the 61 other species-level genomes evaluated in this study.

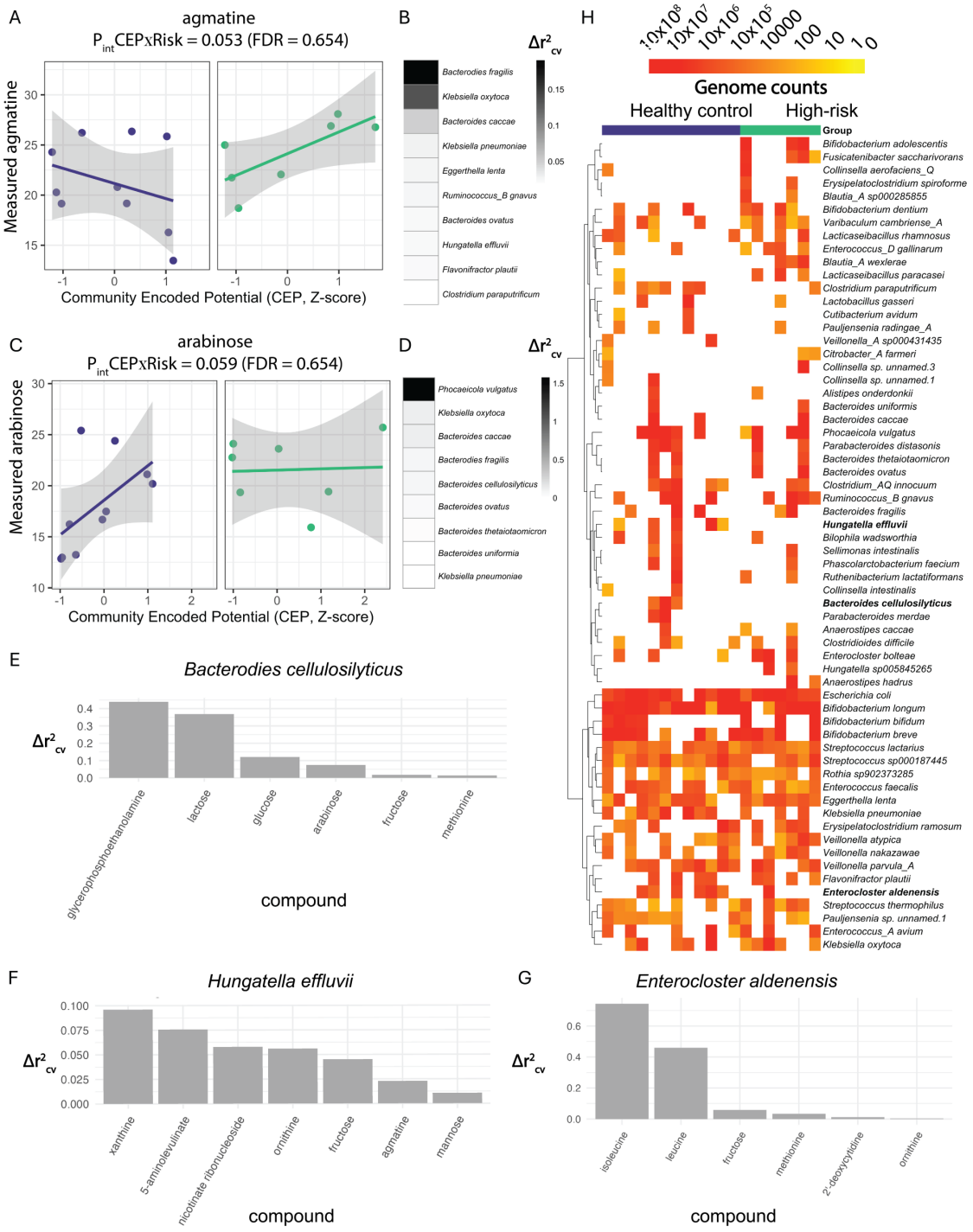

Figure 3.3. Risk-dependent metabolite–community encoded metabolic potential interactions reveal B. cellulosilyticus and H. effluvii as contributors to arabinose and agmatine abundance. A,C) Robust regression models evaluating the interaction between community-encoded metabolic potential (CEP) and asthma-risk group for agmatine (A) and arabinose (C). B,D) Shaded regions denote standard error of the regression. Leave-one-genome-out Δr^2_{cv} analyses (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) identifying genomes which influence the CEP–metabolite models for agmatine (B) and arabinose (D). E-G) Compound-specific $\Delta r^2_{cv)}$ profiles for B. cellulosilyticus (F), H. effluvii (G), and E. aldenensis (H). H) Log-transformed genome counts for all 62 species-level MAGs across healthy control (HC) and high-risk (HR) infant stool samples (n = 19).

Tables

Table 3.1. NCBI Sequence Accession Numbers for All Metagenomic Samples Sequenced in This Study

Sample	Accession	SRA Run
209164405_M06_stool	SAMN54466236	SRR36693160
209165190_M06_stool	SAMN54466235	SRR36693161
210156197_M06_stool	SAMN54466234	SRR36693162
209166191_M06_stool	SAMN54466233	SRR36693163
209154608_M06_stool	SAMN54466232	SRR36693164
209148606_M06_stool	SAMN54466231	SRR36693165
209138401_M06_stool	SAMN54466230	SRR36693166
212212999_M06_stool	SAMN54466229	SRR36693167
210204999_M06_stool	SAMN54466228	SRR36693168
210203999_M06_stool	SAMN54466227	SRR36693152
209102999_M06_stool	SAMN54466226	SRR36693153
208099999_M06_stool	SAMN54466225	SRR36693154
208098999_M06_stool	SAMN54466224	SRR36693155
208097999_M06_stool	SAMN54466223	SRR36693156
208095999_M06_stool	SAMN54466222	SRR36693157
208093999_M06_stool	SAMN54466221	SRR36693158
208090999_M06_stool	SAMN54466220	SRR36693159
208087999_M06_stool	SAMN54466219	SRR36693169
208083999_M06_stool	SAMN54466218	SRR36693170

Chapter 4: Atopy-bound Infants are Microecologically Distinguishable with Discrete Temporal Associations

Abstract

The early-life gut microbiome and metabolome are increasingly implicated in the development of allergic disease, but the timing and specificity of these associations across the first year of life remain poorly resolved. Using the Trial of Infant Probiotic Supplementation (TIPS) cohort, we performed genome-resolved metagenomics and paired untargeted and targeted metabolomics on stool collected at birth, 1, 3, 6, 9, and 12 months, alongside reports of atopy, asthma, and eczema. We assembled 7,261 genomes, dereplicated to 563 representative species genomes, and found that global microbial community structure tracked with age and breastfeeding duration, but not with disease. Cross-sectional differential abundance testing identified discrete, taxon-specific signatures across the first year. *Akkermansia massiliensis* was more prevalent in asthma-bound infants from birth through 6 months, marking an early window of divergence. At 3 months, *Staphylococcus aureus* and *Lactococcus lactis* were depleted in infants who developed atopy and asthma respectively. *Achromobacter aegrifaciens* was depleted in eczema-bound infants at 9 months. Longitudinal modeling of the full first year, rather than single visits, additionally identified *Phocaeicola massiliensis* and *Coprobacter fastidiosus*, taxa that diverged gradually rather than at any single timepoint. Untargeted metabolomics mirrored the 3-month genomic signal, with 171 features differentially abundant exclusively at that visit in asthma-bound infants. Twelve of these

features remained significant across three independently matched control subsets and were dominated by linoleic acid and related lipids. Targeted assay of 12,13-diHOME, a bacterial linoleic acid derivative, was elevated at 12 months in infants with atopy or asthma. Together, this study supports a model in which discrete early-life windows, rather than a diffuse first year, shape the trajectory toward atopic disease. We provide specific microbes, metabolites, and disease-relevant infant genomes that define these windows and offer a foundation for mechanistic studies of metabolic-immune tolerance in humans.

Introduction

Allergic disease, including asthma, atopic dermatitis, and food allergy, affects a substantial proportion of children in industrialized nations.[112] Asthma alone is classified as the most common chronic childhood disease, affecting 6.5% of U.S. children as of 2024.[1] The rapid rise in these conditions over the past several decades points to environmental rather than purely genetic drivers. A growing body of evidence identifies early life, particularly the neonatal period, as a critical window during which the developing immune system is uniquely receptive to microbial and environmental input.[16,17] Alteration in the gut microbiome during this window is consistently associated with later development of atopy and asthma.[73,106] Mice colonized in early life with specific bacterial taxa show reduced airway inflammation in adult progeny, while colonization delayed beyond this window fails to confer the same protection.[18]

Epidemiological studies of children raised in traditional farming environments reinforce this model. Farm-dwelling children show markedly reduced risk of physician-diagnosed asthma, an effect attributed to early and sustained exposure to a richer, more diverse environmental microbiome.[113,114] Together, these findings suggest that the composition and function of the

early-life gut microbiome, shaped by both host and environmental factors, may causally influence the trajectory toward allergic disease. However, the specific microbial and metabolic features that mediate this risk across the first year of life remain incompletely defined. Here, we leverage longitudinal metagenomic and metabolomic profiling of the Trial of Infant Probiotic Supplementation (TIPS) to characterize these dynamics and identify candidate microbial mediators of childhood asthma risk alongside their functional impacts.

Methods

Study Population

Stool samples for this study were obtained from the Trial of Infant Probiotic Supplementation (TIPS – ClincialTrials.gov ID: NCT00113659) clinical trial. Recruitment for TIPS was carried out in the San Francisco Bay Area, California, United States and consisted of infants at high-risk of childhood asthma development based on an asthma diagnosis of one or both parents.[77] TIPS was a double blind placebo controlled trial of daily oral *Lactobacillus rhamnosus* GG supplementation, where infants were randomized between treatment and control groups. Stool samples were collected from infants at birth, 1, 3, 6, 9, and 12 months of age. In this study only infants in the placebo arm of TIPS (N =83) were considered, such that *L. rhamnosus GG* supplementation was not a confounder in our study. All stool samples available for sequencing from the placebo arm were included, and every sample with material remaining after DNA extraction was additionally submitted for metabolomics. The parent study was approved by the Institutional Review Board of the University of California, San Francisco (IRB #10-02968);

oversight was transferred to the Children's Hospital at Montefiore Einstein (IRB #2019-10582) in 2019 for continued survey collection and data management.

Allergic outcomes were ascertained from clinical follow-ups and parental surveys. Asthma was defined based on the parent self-reporting the child's diagnosis at least twice by 7-year survey collection, and eczema was defined based on the parent self-reporting the child's diagnosis at least twice OR diagnosis once during the TIPS physical exam by 1 year of age. Total serum IgE was measured at up to three annual visits in the first three years of life by the UCSF Clinical Laboratories; high total IgE was defined as a measurement of at least 60 IU/mL at any of those visits, the upper limit of the reference interval for children aged 1 to 5 years. Two composite phenotypes were derived. *Atopic* was defined as eczema or high total IgE; asthma is deliberately excluded from this definition, so an infant with asthma alone is not an atopic case. *Atopic or asthmatic*, defined to match Levan et al.,[110] was defined as asthma, eczema or high total IgE.

Four contrasts were pre-specified as primary: 1) atopic or asthmatic compared to a healthy reference, 2) atopic compared to a healthy reference, 3) asthma compared to all infants without asthma, and 4) eczema compared to all infants without eczema. The healthy reference group comprised the 36 infants explicitly negative for asthma, eczema and high total IgE. Asthma and eczema were additionally contrasted against this healthy reference as supplementary sensitivity analyses, testing whether the primary findings persisted once infants carrying a different allergic phenotype were removed from the comparison group thereby separating loss of signal from loss of power in the smaller contrast. Across the 83 infants contributing samples, the case groups comprised 41 (atopic or asthmatic), 36 (atopic), 28

(eczema) and 11 (asthma) infants. As repeat visit compliance is variable in perspective cohort studies, we excluded infants from our analysis contrasts when data required for establishing a phenotype was missing. Total IgE was the most frequently missing data type, and the resulting per-contrast sample flow is shown in **Figure 4.1A**.

Twenty-four stool samples from a concurrently collected, low-asthma-risk cohort (DIMES) were processed in the same laboratory batches and are present in the sequencing and metabolomics runs. Because allergic outcomes were not ascertained in that cohort, these samples contributed only to metagenome assembly but not disease contrast.

Metagenomics

Stool DNA Extraction

Total nucleic acid was extracted using a combined chemical (cetyltrimethylammonium bromide, CTAB), mechanical (bead beating) and organic-solvent (phenol) protocol. Buffers were prepared in house: 5% CTAB extraction buffer (50 g hexadecyltrimethylammonium bromide, Sigma-Aldrich 52365, per liter of 1:1 1 M sodium chloride and 1 M phosphate buffer, pH 7.8) and 30% PEG/NaCl (300 g polyethylene glycol 6000, Sigma-Aldrich 81260, per litre of 1.6 M sodium chloride); both were autoclaved before use. All manipulations other than heat-block incubation, centrifugation and bead beating were performed in a biosafety cabinet decontaminated with 10% bleach, 70% ethanol and DNA Away (Fisher Scientific 21-236-28). Two negative extraction controls, processed identically except that no biological material was added, were included with every batch of samples.

Frozen stool samples were transferred directly from -80 °C storage onto dry ice. For each sample, 500 µL of pre-warmed 5% CTAB extraction buffer was aliquoted into a Lysing Matrix E tube (LME; MP Biomedicals 6914-100) and one stool punch (approximately 0.1–0.25 g of stool in total), taken from different regions of the sample where possible, were transferred into the tube using a sterile 4 mm biopsy punch (VWR International 95039-102). Samples were vortexed and incubated at 65 °C for 15 minutes with shaking at 400–500 rpm, then allowed to cool to room temperature. 500 µLs of the lower, room-temperature-equilibrated phase of phenol:chloroform:isoamyl alcohol (25:24:1; Sigma-Aldrich P2069) was added to each LME tube, and samples were homogenized in a FastPrep-24 homogenizer (MP Biomedicals) at 5.5 m/s for 30 seconds and centrifuged at 16,000 × g for 5 minutes at 4 °C to separate the aqueous and organic phases. 400 µLs of the aqueous phase was transferred to a 2 mL microcentrifuge tube pre-loaded with 800 µL of chloroform (Sigma-Aldrich 366919). A further 400 µL of 5% CTAB extraction buffer was added to the retained LME tube and the bead-beating and centrifugation steps were repeated; the second aqueous phase (400–500 µL) was pooled with the first in the chloroform-containing tube. The pooled extract was centrifuged at 16,000 × g for 5 minutes at 4 °C, and the upper aqueous phase was transferred to a fresh 2 mL tube containing 1,000 µL of 30% PEG/NaCl solution. Nucleic acid was precipitated overnight at 4 °C.

Precipitated nucleic acid was pelleted at 16,000 × g for 10 minutes at 4 °C, washed twice with 250 µL of freshly prepared, ice-cold 70% ethanol (each wash followed by 16,000 × g for 5 minutes at 4 °C), air-dried for 10 minutes and resuspended in 50 µL of TE buffer. DNA was quantified with the Qubit dsDNA Broad Range Assay Kit (Invitrogen Q33265) and stored at −20 °C until library preparation. A batch was rejected and re-extracted if any negative extraction

control produced visible high-molecular-weight DNA on a 0.8% agarose gel or exceeded 2 ng/µL by the Broad Range kit or 1 ng/µL by the High Sensitivity kit.

DNA Sequencing

Shotgun metagenomic library preparation and sequencing was performed by Novogene Corporation Inc. (Sacramento, CA, USA). 434 stool DNA extracts comprised of 407 TIPS stool aliquots with 20 replicates included for technical validation and 27 DIMES aliquots were submitted for sequencing. Libraries were prepared to a 350 bp insert size using the provider's standard, PCR-based shotgun metagenomic protocol and were sequenced on an Illumina NovaSeq platform in paired-end 150 bp mode to a target yield of 10 Gb of raw sequence per sample. 420 libraries (376 TIPS + 24 DIMES + 20 technical replicates) of the 434 returned sequences at a median depth of 10.3 Gb (69.4 million reads after quality trimming). Demultiplexing was performed by the provider prior to delivery; adapter removal and quality trimming were performed in house and are described below.

Read Processing, Assembly and Binning

Read pairs were pre-processed in house: read files belonging to the same sample were concatenated, quality was surveyed with FastQC,[79] adapters were removed with BBDuk (BBTools v38.98)[80] and read ends were quality-trimmed with sickle[81] at a minimum Phred score of 20. Processed reads were summarized with seqkit v2.5.1, confirming a minimum read length of 20 bp and a mean of approximately 149 bp.

Overlapping read pairs were merged with BBMerge (BBTools v38.98), and each sample was assembled independently with metaSPAdes (SPAdes v4.0.0)[115] in --meta mode at k-mer

lengths of 21, 33 and 55. Co-assembly of multiple infant samples was not performed, to avoid chimeric contigs. Scaffolds were filtered to a minimum length of 1,000 bp with Reformat (BBTools v38.98),[80] and per-sample read counts were re-attached after mapping with BBMap (BBTools v39.01)[80] at a minimum identity of 0.97. Open reading frames were predicted with Prodigal v2.6.3 (-p meta), transfer RNAs with tRNAscan-SE (v2.0.12)[96] and 16S rRNA genes with cmsearch (Infernal v1.1.5).[116] Predicted proteins were searched against KEGG, UniProt and UniRef100 with DIAMOND blastp for preliminary annotation.

Contigs of at least 2,500 bp were binned within each sample by five independent algorithms: MetaBAT2 (v2.12.1),[85] MaxBin2 (v2.2.7),[86] CONCOCT (v1.1.0),[87] MetaBinner (v1.4.4)[117] and VAMB (v3.0.9).[88] Results from the five binning algorithms were consolidated per sample with DAS Tool (v1.1.6)[90] at a score threshold of ≥ 0.3, yielding 7,391 metagenome-assembled genomes (MAGs) pooled across all samples of both cohorts.

Genome Aggregation, Coverage Profiling and Functional Annotation

Pooled MAGs were quality-assessed with CheckM2 (v1.0.1)[91] and dereplicated with dRep (v3.5.0).[92] Genomes were filtered to ≥ 70% completeness and ≤ 5% contamination, then clustered by MASH primary clustering at 90% average nucleotide identity (ANI) followed by ANImf secondary clustering at 95% ANI. Secondary clusters correspond to species level taxa. The highest-scoring genome, based on the dRep scoring scheme, in each cluster was retained as the species representative. Taxonomy was assigned to all genomes with GTDB-Tk (v2.4.0)[93] against GTDB release r220. Genetic codes were inferred for species representative genomes with Codetta,[94] after which open reading frames were re-predicted with Prodigal (v2.6.3)[95] in single-genome mode under the inferred code. Transfer RNAs were called with tRNAscan-SE

(v2.0.12).[96] The final catalogue comprised 563 species-representative MAGs spanning 41,612 scaffolds and served as the reference database for abundance estimation.

Reads from all 420 sequenced libraries were mapped to the concatenated MAG catalogue with Bowtie2 (v2.5.1)[98] and sorted with samtools (v1.17),[118] giving a median overall alignment rate of 84% (interquartile range 79–86%, full range 45–96%). Coverage was computed from the sorted BAM files with CoverM (v0.7.0; --min-read-percent-identity 95 --min-covered-fraction 0),[99] producing per-sample genome-level read counts and genome-size-normalized coverage. A filtered version of each table retained a genome in a sample only if it recruited at least 100 reads and its observed breadth of coverage did not deviate from that expected given its depth. A median of 75% of reads (IQR 69–79%) were assigned to a catalogue genome at this identity threshold. Integer read counts were used for all diversity and differential-abundance analyses. MAG abundances in DIMES samples, which were only included to enhance species representative MAG catalog construction were excluded at this point. Only MAG abundances from TIPS samples entered the downstream analyses.

Functional annotation was performed on the 1,479,029 proteins predicted from the species-representative MAG catalogue using three independent tracks. KEGG orthologues were called with KofamScan (v1.3; exec_annotation)[97] against KOfam release 2023_08_01, retaining only hits passing each family's adaptive score threshold. Protein domains were assigned with hmmscan (HMMER v3.4)[119] against the Pfam-A release 37.0 (21,979 families) using each family's curated gathering cutoff (--cut_ga). Carbohydrate-active enzymes were annotated with run_dbcan v5 (dbcan 5.2.8)[120] in protein mode using the DIAMOND, dbCAN-HMM, and dbCAN-sub methods, retaining a hit when supported by at least two of the three methods. For

community-level functional profiles, per-ORF annotations were aggregated to genome level and weighted by each genome's abundance in each sample, producing sample × KEGG-orthologue, sample × Pfam and sample × CAZy matrices.

Metabolomics

Metabolomics was performed by the Quantitative Metabolite Analysis Center (QMAC) at the University of California, San Francisco. The untargeted and targeted bioactive-lipid assays shared a single extraction but were acquired on separate instruments.

Sample preparation and mass spectrometry

Lyophilized stool was weighed to a target mass of 15 mg (achieved masses across the 364 TIPS aliquots submitted, 2.3–91.2 mg; median 23.7 mg) and homogenized in a Precellys 24 bead homogenizer (Bertin Technologies) pre-cooled with dry ice at 6,400 rpm for 3 × 20 s with 30 s rest intervals. Extraction buffer (acetonitrile:methanol:water, 2:2:1) was added to a stool concentration of 30 mg/mL, that is 15 mg in 500 µL; the 70 aliquots falling below the 15 mg target were extracted in the full 500 µL and are correspondingly more dilute, and a per-sample biomass ratio was carried forward to account for this. The buffer was supplemented with 2-amino-3-bromo-5-methylbenzoic acid and an isotope-labelled amino-acid mixture as untargeted internal standards (**Table S4.1**), together with deuterated eicosanoid standards spiked at lower concentrations for the targeted assay (**Table S4.2**). Samples were held on ice for 30 minutes, precipitated protein and cellular debris were pelleted by centrifugation at 21,000 ×g for 10 minutes, and supernatants were aliquoted in 500 µL volumes, dried under vacuum and stored at −80 °C until acquisition.

On the day of acquisition, dried aliquots were reconstituted in water, vortexed briefly and filtered through a 0.2 µm plate filter (3,400 rpm for 10 minutes) alongside an extraction control. Rehydration of the extracted stool clogged the first filter plate and gave variable recovered volumes; the reconstitution volume was therefore increased from 100 µL to 200 µL for the remaining plates, and three filter plates were used in total. Filtrates were diluted 6-fold in water before untargeted acquisition (15 µL filtrate plus 75 µL water on the first plate; 30 µL plus 60 µL followed by a further 2-fold dilution thereafter) and analyzed immediately. For the bioactive-lipid-mediator assay, filtrates were diluted a further 2-fold in methanol (12-fold in total), giving a final solvent composition of 50:50 water:methanol. Because the reconstitution volume differed between the first filter plate and the remainder, injected concentrations are not on a common absolute scale across filter plates; acquisition plate was included as a covariate in every targeted model, and all reported effects are within-compound contrasts on the log2 scale.

Untargeted metabolomics was performed by injecting a 5 µL aliquot of each sample onto a Shimadzu 30-AD UPLC coupled to a SCIEX TripleTOF 6600+ mass spectrometer operated in data-independent (SWATH) acquisition mode. Small molecules were separated on a Kinetex F5 column (2.6 µm, 2.1 × 150 mm; Phenomenex 00F-4723-AN) with mobile phase A of water with 0.1% formic acid and mobile phase B of acetonitrile with 0.1% formic acid. At a constant flow rate of 200 µL/min, the timed gradient was: 0 min, 0.2% B; 2 min, 0.2% B; 5 min, 2% B; 11 min, 25% B; 13 min, 98% B; 16 min, 98% B; 16.1 min, 0.2% B; 20 min, 0.2% B. Positive and negative ionization data were acquired in separate injections with source parameters CUR 35, GS1 60, GS2 80, ISVF 5500 V (positive) and −4500 V (negative), and TEM 400 °C. The method

surveyed ions from 50 to 1,000 m/z in MS1 followed by 20 SWATH windows for MS2 fragmentation at a collision energy of 15 V, with a total cycle time of 0.651 s. Samples were randomized and acquired in batches.

Targeted analysis of bioactive lipid mediators was performed by injecting a 5 µL aliquot of each sample onto a Shimadzu 30-AD UPLC coupled to a SCIEX 7500 triple-quadrupole mass spectrometer. Analytes were separated on a Kinetex Polar C18 column (2.6 µm, 100 Å, 100 × 3.0 mm; Phenomenex 00D-4759-Y0) with mobile phase A of water with 0.1% formic acid and mobile phase B of methanol with 0.1% formic acid, at a constant flow rate of 500 µL/min and a column temperature of 50 °C. The timed linear gradient was: 0 min, 10% B; 0.1 min, 45% B; 2 min, 45% B; 16.5 min, 80% B; 16.6 min, 98% B; 18.5 min, 98% B; 18.6 min, 10% B; 20.5 min, 10% B. Data were acquired in multiple-reaction-monitoring (MRM) mode with positive/negative polarity switching using source parameters CUR 40, GS1 60, GS2 70, CAD gas 12, temperature 350 °C, and ISVF 5500 V (positive) and −4500 V (negative). MRM transitions were optimized from authentic standards. Additional method parameters were a cycle time of 0.937 s, a dwell time of 1 ms per MRM pair, a settling time of 15 ms and a pause time of 5.007 ms. Targeted acquisition began approximately three months after extraction and proceeded one 96-well plate at a time, in both polarities, over approximately six days per plate; targeted extracts therefore underwent one freeze-thaw cycle and up to one week in the autosampler before acquisition, and questionable injections were reviewed and re-run. Although the method monitors lipid mediators derived from several polyunsaturated fatty acids, analysis here focused on linoleic-acid-derived metabolites.

Four classes of control injection were interspersed after every ten samples in the untargeted run: solvent blanks (100% final dilution solvent; background and carry-over), spiked isotope-labelled internal standards (instrument variability), extraction controls (the internal-standard mixture carried through the full extraction in 100% final solvent, capturing preparation bias) and pooled quality-control samples (4 µL drawn from every sample in the batch, used both to monitor reproducibility and to anchor drift correction).

Untargeted Data Processing

Raw untargeted data were processed by QMAC with MS-DIAL (RIKEN, version 4.9.221218) for peak detection, deconvolution, alignment, gap filling and quantification, with annotation by spectral matching against the MS-DIAL curated high-resolution MS2 library (version 19) and HMDB. Peak intensities were drift-corrected by locally weighted (LOWESS) regression anchored on the pooled quality-control injections. Zeros were replaced with a pseudo-count of 1×10^{-6}, and intensities were normalized first to sample biomass and then to the spiked isotope-labelled internal standards. QMAC removed background and non-biological features using five criteria (mean intensity above the mean blank; quality-control-to-blank ratio ≥ 2; quality-control coefficient of variation ≤ 50%; intraclass correlation ≤ 0.90; blank-versus-sample Mann-Whitney $p \leq 0.05$), reducing 36,675 positive-mode and 41,953 negative-mode aligned features to 4,716 and 10,032 respectively; blank and standard injections were consumed by this step, so no blank filter was reapplied downstream.

All subsequent processing was performed in R. A failure of the negative-mode normalization standard inflated affected values by many orders of magnitude; the 7 samples in which ≥ 20% of features were affected were removed, and the remaining affected cells (≤ 5% of

features in 12 samples) were replaced by the median of that feature across unaffected samples, with every repaired feature flagged through to the final result tables. Intensities were log2-transformed and reduced by a filter computed entirely from across-sample statistics, without reference to disease status, so that downstream false-discovery-rate control remains valid: features were required to have a quality-control coefficient of variation ≤ 0.30, detection in ≥ 50% of samples, and log2 variance above the 5th percentile. This retained 5,790 negative-mode features in 354 samples, 2,952 positive-mode features in 360 samples, and 8,742 features in the 352-sample combined matrix. Because false-discovery-rate correction was applied within each matrix's own feature set, the per-mode and combined analyses are reported separately and their adjusted *p*-values are not compared across matrices. Named metabolites throughout are MS-DIAL/HMDB spectral matches and are therefore putative: MS/MS library matches correspond to level 2 and accurate-mass or molecular-formula matches to level 4 on the scale of Schymanski et al.[121] Only the eight negative-mode features subsequently confirmed by QMAC against authentic standards, by exact mass within 5 ppm, retention time and MS2 spectrum, reach level 1; all other identifications are reported as putative and are not standard-confirmed.

Targeted Data Processing

Raw targeted data were processed by QMAC with SCIEX OS (SCIEX, version 2.1.6.59781) for peak picking, alignment and integration. Thirteen lipid mediators were monitored. Each analyte peak area was expressed as a ratio to its compound-matched deuterated internal standard. Ratios were normalized for biomass. All modelling was performed on the log2 scale. Each measurement was classified as detected, below the limit of detection, or unmeasured. The

limit of detection was computed per compound and per 96-well plate from blank wells as the mean blank plus three standard deviations, and values below it were imputed at LOD/√2. Blank levels differed by up to three orders of magnitude between plates. Plate was therefore verified to be independent of disease status and was included as a covariate in every model. Cells in which the internal standard itself failed were treated as unmeasured rather than as low values. In these cells a real analyte peak was frequently still present while the extraction control was simultaneously depleted 25 to 30 fold, indicating partial recovery through a clogged filter rather than a low analyte concentration.

Compounds were retained by an outcome-blind detection-prevalence filter (≥ 30%). Linoleic acid yielded a flat calibration curve with high background and was not quantifiable. Prostaglandin E2 and lipoxin A4 were run as positive controls and were not detected in any sample. 9-HODE was excluded on quality grounds, having unstable blank levels across plates, an unweighted calibration curve and 23.7% internal-standard failure. The arachidonic-acid-derived analytes were outside the scope of this analysis. The analysis set therefore comprised 5 compounds measured in 358 samples from 83 infants across 4 plates. 12,13-diHOME, 9,10-DiHOME and 13-HODE were quantified against their own calibration curves and analyzed on an absolute scale. The two EpOMEs were analyzed on a relative log2 area-ratio scale only, because their own deuterated analogues were present in too few samples to calibrate and were normalized by QMAC to deuterated 13-HODE. Results are reported as fold-changes with 95% confidence intervals.

Statistical Analysis

Software

All analyses were performed in R (v4.5.2).[122] Data handling and visualization used data.table (v1.18.2.1),[123] dplyr (v1.2.1),[124] tidyr (v1.3.2),[125] ggplot2 (v4.0.3),[100] cowplot (v1.2.0),[126] ggrepel (v0.9.8)[127] and scales (v1.4.0).[128] Community ecology used vegan (v2.7.3).[129] Differential abundance used limma/voom (v3.66.0),[55,56] edgeR (v4.8.2),[130] and variancePartition/DREAM (v1.40.2).[131,132] Linear mixed models used lme4 (v2.0.1)[133] and lmerTest (v3.2.1).[134] Matching and balance assessment used MatchIt (v4.7.2)[135] and cobalt (v4.6.3).[136] A single random seed was fixed at the top of every script, and each analysis script writes a report recording its parameters, the R version, the run log and all warnings alongside its result tables. Analysis code is available at https://github.com/hollysteininger/long_TIPS.

General approach

Two complementary designs were applied throughout. Cross-sectional analyses were fit independently within each visit (birth, 1, 3, 6, 9 and 12 months), and as there was only one sample per subject per visit, no random effect was required. Feature filtering was repeated within each visit so that each visit's analysis was conducted on its own feature set. Longitudinal analyses pooled all visits and modelled repeated measurements on the same infant with a subject-level random intercept, (1 | subject_id), with age encoded in months. Longitudinal models were fit in two arms: an additive arm testing a sustained mean difference between groups across the first year of life, and an interaction arm testing divergence of trajectories, in which the group × age coefficient is the test of interest.

Every disease contrast was adjusted for breastfeeding duration, as breastfeeding and diet are the strongest determinants of gut microbial composition in infancy.[137] Multiple testing was controlled by the Benjamini-Hochberg procedure[51] applied within each model (i.e. visit × disease × arm) and the scope of correction is stated with each table; p-values were corrected within, but not across visits, diseases or model arms.

Metagenomics

Beta diversity was assessed with the binary Jaccard distance (vegan::vegdist) and tested by PERMANOVA (vegan::adonis2, 999 permutations).[138] Covariates were screened individually within each visit and metric, and those associated with community composition were then fitted jointly with the disease term entered last, using marginal, order-independent sums of squares.

Differential abundance of metagenome-assembled genomes was tested on integer read counts. Cross-sectional analyses used limma-voom (TMM normalization within edgeR::DGEList, then voom, lmFit and eBayes) fit independently within each visit × disease cell after re-filtering to features present in at least three samples of that cell; longitudinal analyses used variancePartition::dream with a subject random intercept, in both arms.

Metabolomics

Untargeted beta diversity was assessed using the Canberra distance on linear intensities (vegan::vegdist), tested by PERMANOVA (vegan::adonis2, 999 permutations) and following the same adjusted, cross-sectional-only scheme described above. Differential abundance of untargeted features was tested with limma on the already-log2-transformed intensity matrix,

without voom, TMM or a DGEList, because the values are continuous intensities rather than counts. lmFit and eBayes were applied per visit × disease cross-sectionally, and variancePartition::dream with a subject random intercept longitudinally, in both arms. Because the 3-month asthma comparison was strongly imbalanced (8 cases versus 56 controls), its cross-sectional result was subjected to a matched-subset robustness battery in which controls were matched to cases on child sex (exact) and breastfeeding duration (nearest neighbour) using MatchIt without a propensity score given the small number of cases. The identical limma model was re-fit on each matched set; covariate balance before and after matching was assessed with cobalt. Three readouts were computed: significance in each of three sequentially drawn, non-overlapping 1:1 matched subsets; the fraction of 50 sex-stratified, breastfeeding-calipered (± 3 months) random 1:1 resamples in which each feature was again significant; and a leave-one-case-out analysis against the full control pool. Because Benjamini-Hochberg correction across approximately 8,700 features has almost no power at n ≈ 16 per matched subset, robustness was judged on the nominal resampling frequency and on the full-power leave-one-case-out analysis rather than on adjusted significance within the matched subsets.

The targeted panel was analyzed with per-visit comparisons fit as lm(log2(concentration) ~ group + breastfeeding duration + plate) and are reported descriptively as fold-changes with 95% confidence intervals without multiple-testing correction. Inference was drawn from longitudinal linear mixed models, lmer(log2(concentration) ~ group + age + breastfeeding duration + plate + (1 | subject_id)) for the additive arm and with a group × age interaction for the interaction arm, fitted with lme4 and lmerTest using Satterthwaite degrees

of freedom. Visit enters this model as a covariate, and Benjamini-Hochberg correction was applied across diseases within each compound and arm.

Results

The Trial of Infant Probiotic Supplementation (TIPS) collected stool at birth, 1, 3, 6, 9, and 12 months, along with reports of atopic disease (**Figure 4.1A**). To resolve the microbial signals in the first year of life, we assembled 7,261 genomes and dereplicated them to 563 representatives (**Figure 4.1B**). Microbial community structure was associated with age (Jaccard $R^2 = 0.11$, $p = 0.001$) and breastfeeding duration ($R^2 = 0.020$, $p = 0.001$; **Figure 4.1C–D**), but not with disease at any visit for any of the six contrasts (min $q = 0.23$**; Figure S4.1A**). Testing each visit separately additionally linked community structure to delivery mode at birth ($R^2 = 0.036$, $q = 0.064$), breastfeeding duration from 1 through 12 months ($R^2 = 0.023$–0.055, $q \leq 0.040$), and the number of children not yet toilet trained in the home at 12 months ($R^2 = 0.032$, $q = 0.016$) (**Figure S4.1A**).

Cross-sectional differential abundance, correcting for breastfeeding duration, identified four organisms of interest. *Staphylococcus aureus* was depleted at 3 months in infants who went on to develop atopy (log2FC = −7.23, $q = 0.012$), with a similar trend in those who developed atopy or asthma (log2FC = −6.77, $q = 0.055$, **Figure 4.1E–F**). Three taxa were enriched in infants who developed asthma; *Akkermansia massiliensis* at birth (log2FC = 13.25, $q = 0.043$), *and CAG-269 sp003525075* (log2FC = 4.34, $q = 0.046$) and *Faecalibacterium prausnitzii_J* (log2FC = 7.76, $q = 0.046$) at 3 months. Conversely, *Lactococcus lactis* was depleted at 3 months in asthma-bound infants (log2FC = −4.05, $q = 0.046$; **Figure 4.1G**). At 9 months,

Achromobacter aegrifaciens was strongly depleted in infants who developed eczema (log2FC = −8.67, q = 0.0041; **Figure 4.1H**).

Because the broad control arm pools infants with unrelated atopic outcomes, we refit each contrast against a clean healthy reference. This reduces the asthma fit at 3 months (67 to 39 samples) and no feature survives BH correction (minimum q = 0.085; **Figure S4.1B**), so we read this arm for effect stability rather than significance. On that basis *A. massiliensis* (log2FC = 11.00, 83% retained, p = 0.0021) and *L. lactis* (log2FC = −4.75, 117% retained, p = 0.0029) hold their direction and magnitude, whereas *CAG-269 sp003525075* (log2FC = 1.92, 44%, p = 0.22) and *F. prausnitzii_J* (log2FC = 2.85, 37%, p = 0.15) collapse toward zero (**Figure S4.1B,D**). *A. aegrifaciens* likewise persists against the healthy reference (log2FC = −7.08, 82% retained, q = 0.085). Against this reference *S. aureus* is the strongest signal at 3 months, reaching significance for both eczema (log2FC = −7.65, q = 0.0088) and atopy (log2FC = −7.23, q = 0.012) (**Figure S4.1C**).

Testing prevalence alongside abundance localized when each signal appears. The *S. aureus* and *L. lactis* differences were driven by abundance rather than prevalence, and only at 3 months. Neither prevalence gap survives correction (S. aureus, 16% vs 48% of infants, Fisher q = 0.081; L. lactis, 0% vs 37%, q = 0.27) (**Figure 4.1I–J**). *A. massiliensis* differed in abundance only at birth, but was more prevalent in infants who developed asthma from birth through 6 months (birth 60% vs 0%, q = 0.0033; 1 month q = 0.071; 3 months q = 0.031; 6 months q = 0.071) (**Figure 4.1K**). *A. aegrifaciens* was both less abundant and less prevalent in infants who developed eczema, and only at 9 months (0% vs 39%, q = 0.0092) (**Figure 4.1L**).

Modelling the full year rather than single visits highlighted two further taxa. A sustained offset between eczema and healthy identified *Phocaeicola massiliensis*, depleted in eczema across the first year as the only feature significant in that model (log2FC = −1.53, q = 0.014; **Figure S4.1E**); the trend is visible from birth to 6 months but reaches significance at individual visit for either abundance or prevalence (nominal p = 0.015 at 1 month and p = 0.032 at 3 months, all q ≥ 0.47; Supp. Figure 4.1F). A group × age interaction identified *Coprobacter fastidiosus*, as a feature with a significant interaction with age (0.68 log2 per month, q = 0.044; 376 samples). *Coprobacter fastidiosus* diverges in infants who develop asthma from 3 months onward (**Figure S4.1G**) without reaching per-visit significance (nominal p = 0.030 at 9 months and p = 0.024 at 12 months, all q ≥ 0.68; **Figure S4.1H**).

Untargeted metabolomic structure mirrored the genomic result. Canberra community structure was associated with age (R^2 = 0.11, p = 0.001) and breastfeeding duration (R^2 = 0.058, p = 0.001; **Figure 4.2A–B**). Testing each visit separately, breastfeeding duration was the only covariate associated with metabolome structure, and it was so at every visit from 1 through 12 months (R^2 = 0.042–0.180, q ≤ 0.032; **Figure S4.2A**). Its effect was strongest at 1 and 3 months (R^2 = 0.18 and 0.16) and decayed steadily through the first year, and at every visit breastfeeding retained more explanatory power for metabolome structure than it did for MAG composition (**Figure S4.1A; Figure S4.2A)**. No disease contrast reached significance at any visit, although atopy at 6 months approached it (R^2 = 0.032, q = 0.090).

Cross-sectional differential abundance across 8,742 features, six visits, and six disease contrasts found disease-specific abundances in a single window: 3 months, in infants who went on to develop asthma (8 cases, 56 controls; **Figure 4.2C**). One hundred seventy-one features

were differentially abundant at that visit ($q < 0.05$), 131 elevated and 40 depleted in asthma-bound infants, and no feature reached significance at any other visit or in any other contrast (**Figure 4.2C**). The most significant were depleted rather than enriched: pos_26893 (m/z 648.7 @ 8.9 min, log2FC = −17.6, $q = 7.1 \times 10^{-4}$) and pos_15846 (m/z 444.6 @ 8.7 min, log2FC = −17.3, q = 0.0075). These two were also the only differential features not detected in every sample (prevalence 0.90 and 0.89), so their signal carries a presence/absence component that the remainder do not (**Figure 4.2D-E**). Because the case group is small and the control pool large, we re-fit in three non-overlapping control subsets matched to the cases on sex and breastfeeding duration (8 vs 8 each). Twelve features remained nominally significant in all three subsets; ten of those were discovery hits and two emerged only under balancing, and eleven of the twelve stayed significant in at least half of the leave-one-case-out folds (**Figure 4.2C**).

The enriched side of the 3-month window was dominated by lipids. Linoleic acid (LA) was elevated in asthma-bound infants (log2FC = 3.01, q = 0.044; (**Figure 4.2C;F**). Ricinoleic acid (log2FC = 2.64, q = 0.044) and N1-acetylspermine (log2FC = 2.78, q = 0.044) showed effects of the same magnitude (**Figure 4.2F–H**). In the targeted assay, 12,13-diHOME was elevated roughly 1.8-fold at 12 months in infants with atopy or asthma relative to healthy children (p = 0.0089, q = 0.053, 34 vs 29 children), with concordant estimates in the atopy (1.83×, p = 0.016), eczema (1.79×, p = 0.031) and asthma (2.09×, p = 0.031) contrasts; no visit survived BH correction in any of the six visits (**Figure 4.2I**).

Discussion

In this study, the gut microbiome and metabolome show a discrete disease-associated signature depending on the age of the infant. This signature was comprised predominantly of

depletion of organisms in eczema and atopy, while asthma showed both enrichment and depletion signals. *Akkermansia massiliensis*, a recently classified *Akkermansia* species, is enriched in asthma-bound infants from birth through 6 months of age (**Figure 4.1K**). *Akkermansia muciniphila* has previously been associated with protection from allergic asthma in older children.[139] Our signal involves a distinct, more recently classified species. Species-level resolution afforded by genome-resolved metagenomic analysis may explain why our result diverges from prior 16S-based work, which could not distinguish between *Akkermansia* species as has occurred with *Bifidobacterium*.[140] Although it was not significant at any individual timepoint, the longitudinal nature of our data enabled the identification of a later age by asthma interaction of *Coprobacter fastidiosus* which appears to increase in asthmatic infants from 6 months onward (**Figure S4.1G-H**).

The bacteria that were depleted in infants who went on to develop allergic disease expand the previously established concept of allergic infant microbiomes lacking exposure at appropriate timing for immune tolerance development with specific genomes and timings.[16,17] *Staphylococcus aureus* was depleted at three months in infants who went on to develop eczema and atopy (**Figure 4.1E–F**). Commensal staphylococcal colonization of the skin at two months lowers risk of atopic dermatitis at one year,[141] however *S. aureus* skin colonization in particular is a risk factor for food allergy, independent of eczema severity.[142] These results highlight the allergic immune impacts of *S. aureus* and may reflect distinct roles for *S. aureus* at different body sites; tolerizing in the gut during early development and sensitizing on compromised skin. The other organism which was depleted at the same timepoint (3 months) in asthma-bound infants was *Lactococcus lactis* (**Figure 4.1G**), and interestingly a prior study by

Gołębiewski et. al. assessing infant gut microbiota around 3 months of age found *L. lactis* associated with allergy protection.[143] Our finding extends this association to asthma specifically and to a longitudinal, genome-resolved cohort. In the longitudinal model, *Phocaeicola massiliensis* was depleted in infants who developed eczema, an additive effect that highlights the value of longitudinal analysis for capturing disease-relevant divergence that accumulates gradually rather than as a sharp, single-timepoint perturbation (**Supp. Figure 4.1E-F**).

At three months the most differential genomic and metabolomic signals appear, unfortunately this is not a commonly sampled timepoint in other early-life allergic microbiome datasets. The convergence of signals at this timepoint suggests either a direct perturbation around three months, or a downstream consequence of founder effects established earlier in infancy. Our untargeted data show elevated linoleic acid (LA) metabolites in asthma-bound infants at three months (**Figure 4.2C-F**). This is consistent with prior work which defined microbial oxylipin production from LA, an essential fatty acid, as associated with asthmatic infant microbiome signatures at 1 month of life.[50,110,111] However, targeted 12,13-diHOME was elevated only at twelve months within asthmatic or atopic infants. It may reflect a delayed downstream consequence of the three-month microbial and metabolic perturbation or a distinct, later process. Given the causal mouse data already associating microbial oxylipin linoleic acid production with airway inflammation,[110] we treat this as biologically plausible despite the marginal statistics.

Several limitations should be considered. Case counts for individual outcomes were small, particularly the eight asthma cases at three months, which limits statistical power. TIPS represents a single cohort from a single site, which limits generalizability. Genome dereplication

to species-representative genomes trades some strain-level resolution for tractability. Two of the most significant untargeted metabolomic features, pos_26893 and pos_15846, remain unannotated. Finally, we acknowledge that while we can identify compelling statistically significant associations in this study, follow-up work will be required to establish mechanistic linkages. Work is ongoing to pair these assembled infant genomes directly with the enzymatic pathway that produce linoleic acid-derived oxylipins and validate this observation of altered asthmatic-bound LA metabolism at 3 months via stable isotope probing in minimal media stool suspensions. This will connect the microbial and metabolomic layers of this dataset more directly. Together, this study supports a model in which a narrow window in early infancy, rather than a diffuse first year, shapes the trajectory toward atopic disease. We also provide specifics microbes, metabolites, and disease relevant infant genomes to explore this window associated with metabolic-immune tolerance in humans.

Figures

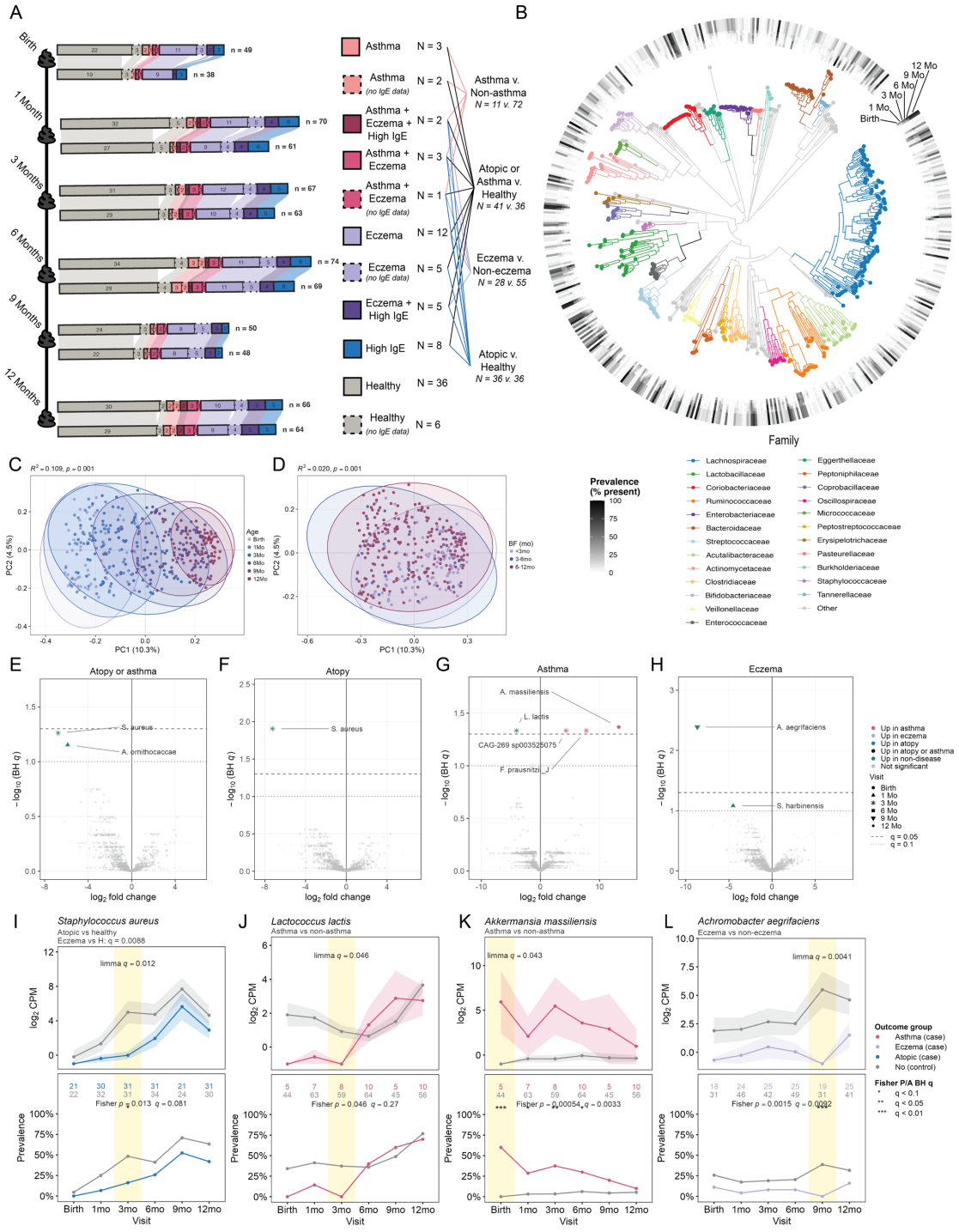

Figure 4.1 Cross-sectional genome-resolved analysis reveals organisms in early-life associated with allergic disease.

A) TIPS study design. (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) B) Phylogenetic tree of the 563 de-replicated genomes assembled with prevalence throughout the first year of life shown in the surrounding heatmap. Jaccard beta diversity metrics clustered by age (C) and breast-feeding duration (D). Cross-sectional differential abundance of children who go on to develop asthma by 7 years of age or any atopy evidenced by the data within this study (E), atopy (F), asthma (G), or eczema by 1 year of age (H). Datapoint shape indicates the independent timepoints tested. Abundance (top) and prevalence (bottom) across the first year of life for the top differential organisms: *Staphylococcus aureus* (I), *Lactococcus lactis* (J), *Akkermansia massiliensis* (K), and *Achromobacter aegrifaciens* (L). Highlighted visit indicates the significant timepoint in a cross-sectional differential abundance linear model correcting for breast-feeding duration. Q-value shown in the corresponding abundance plot. Fisher q-value <0.1 * <0.05 ** <0.01 **.

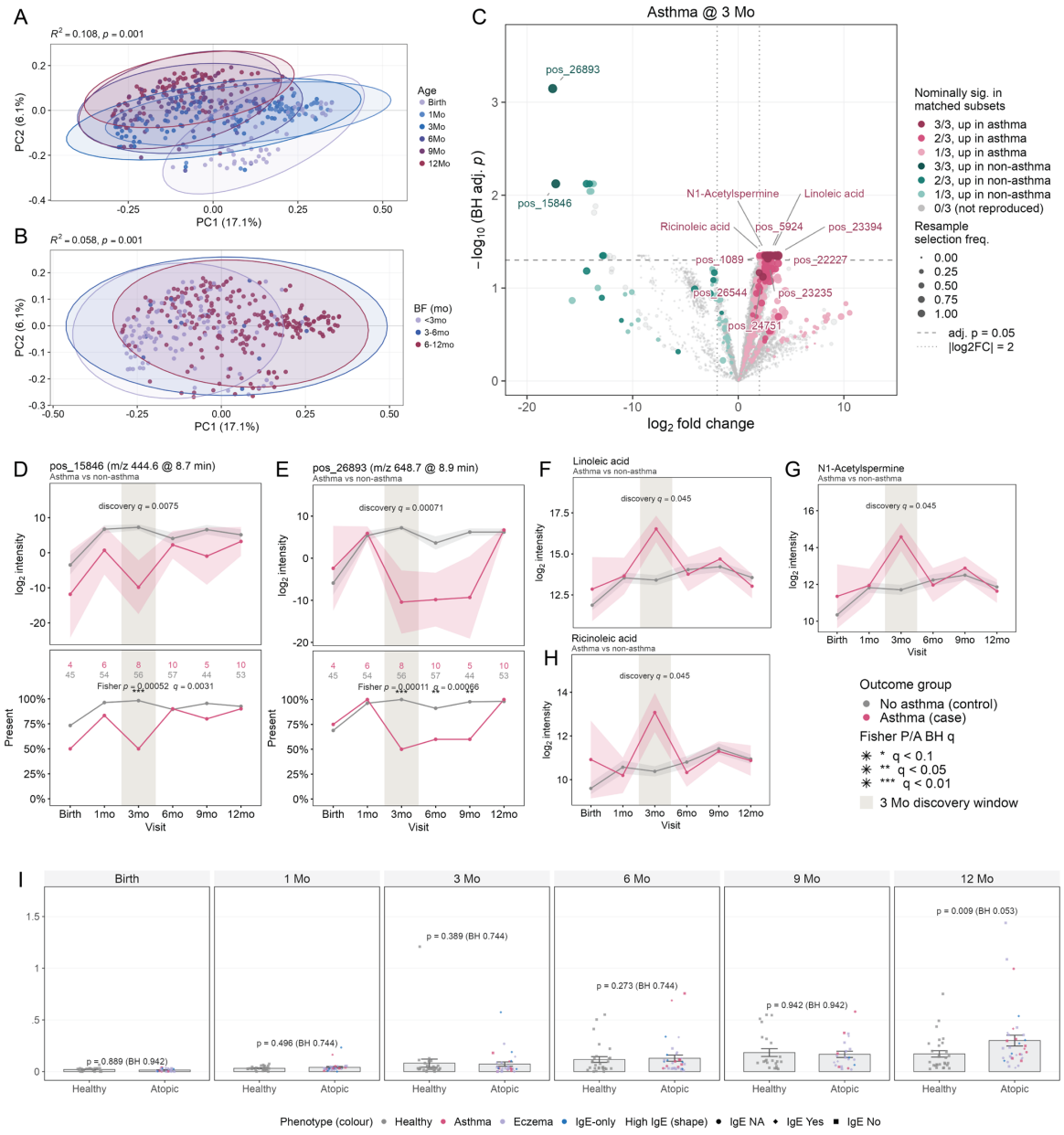

Figure 4.2 Cross-sectional untargeted metabolomic analysis reveals early-life metabolites associated with asthma development.

A-B) Canberra principal coordinates analysis of the filtered untargeted metabolome (8,736 features, 378 samples) ordinated once across all visits, clustered by age (A) and breast-feeding duration (B). Ellipses indicate 95% confidence regions; R^2 and p from cross-sectional PERMANOVA (999 permutations) on the pooled distance. C) Cross-sectional differential abundance of stool from children at 3 months of age who go on to develop asthma, the visit at which differential features were detected (8 cases v. 56 controls). Color indicates the direction of the effect and the number of sex and breast-feeding-matched case-control subsets in which the feature remained nominally significant; point size indicates selection frequency across

(Figure caption continued on the next page.)

(Figure caption continued from the previous page.) many randomly matched resamples (stability selection). The twelve features nominally significant in all three matched subsets are labelled. D–H) Abundance across the first year of life for the top differential features, with prevalence (bottom) shown for the two features that were not detected in all samples: pos_15846 (m/z 444.6 @ 8.7 min) (D), pos_26893 (m/z 648.7 @ 8.9 min) (E), linoleic acid (F), N1-acetylspermine (G), and ricinoleic acid (H). Highlighted visit indicates the 3-month discovery window; q-value shown in the corresponding abundance plot. Fisher P/A q-value <0.1 *, <0.05 **, <0.01 ***. I) Targeted 12,13-diHOME concentration (ng/mg lyophilized stool, mean ± SE with individual samples) in healthy children versus children with atopy or asthma at each visit, colored by phenotype and shaped by high-IgE status. Per-visit p-values from linear models of log2 concentration correcting for breast-feeding duration and plate, with BH correction across the six visits.

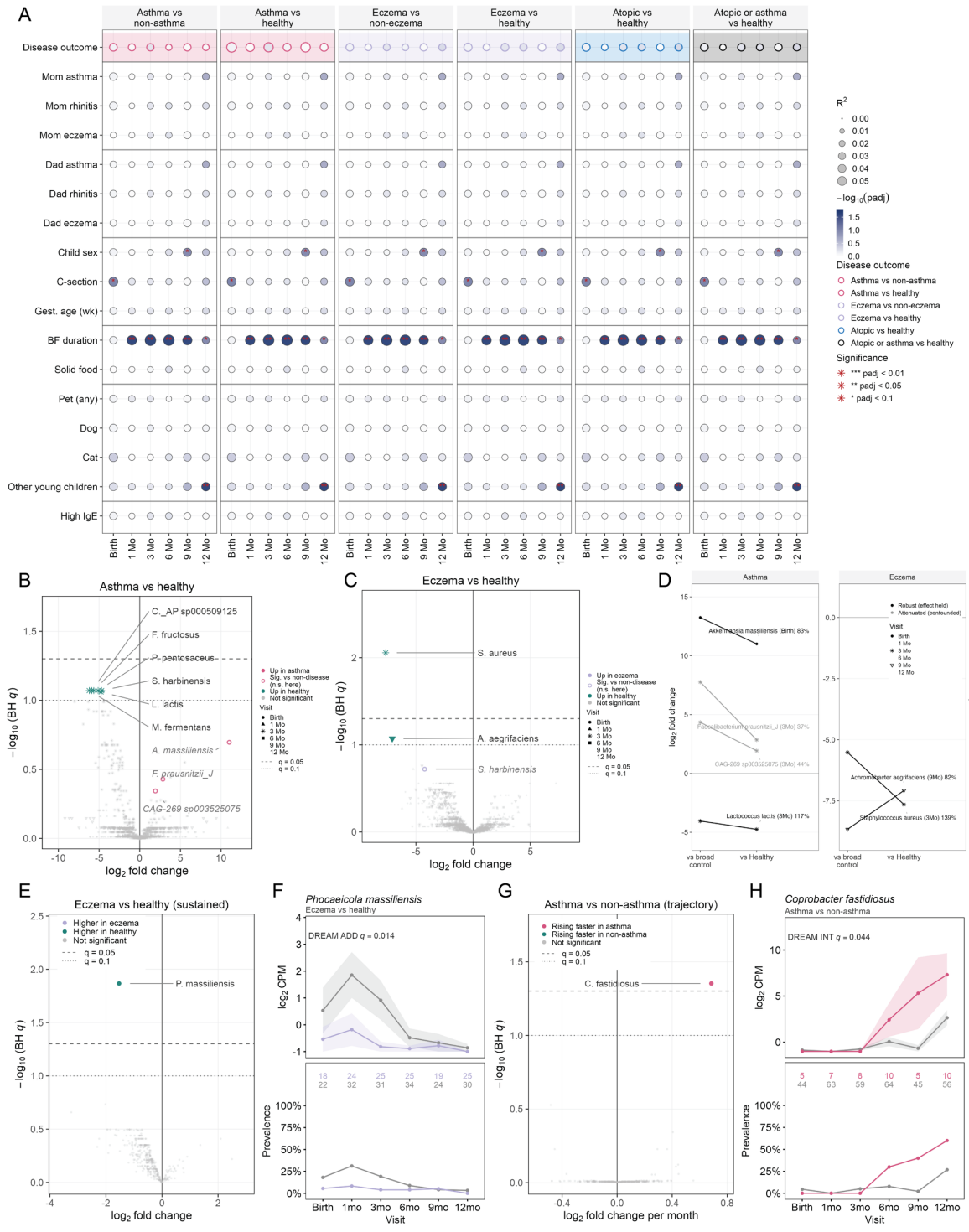

Figure S4.1 Reference group choice and longitudinal modelling refine genome-resolved associations with allergic disease.

A) Cross-sectional PERMANOVA of Jaccard community structure at each visit for each disease outcome (top row, tinted) and for 16 clinical and environmental covariates. Shown for all six disease contrasts. (Figure caption continued on the next page.)

(Figure caption continued from the previous page.)Bubble size indicates R^2; fill indicates log10 BH-adjusted p, corrected across all 22 variables within a visit. B–C) Cross-sectional differential abundance against the clean healthy reference for asthma (B) and eczema (C), with all six visits overlaid and indicated by point shape. Filled points are significant against healthy; ringed points reached significance only against the broad non-disease control. D) log2 fold change of each candidate taxon against the broad control and against healthy. A level line indicates an effect preserved under the change of reference group (black); a line collapsing toward zero indicates an effect that depended on the composition of the broad control (grey). Percentages indicate the fraction of the effect retained. E) Longitudinal differential abundance testing a sustained offset between eczema and healthy across the first year, adjusting for age and breast-feeding duration with a per-subject random intercept. F) Abundance (top) and prevalence (bottom) of Phocaeicola massiliensis across the first year of life. G) Longitudinal differential abundance testing a group × age interaction in children who go on to develop asthma. H) Abundance (top)

and prevalence (bottom) of Coprobacter fastidiosus. Dashed lines, q = 0.05; dotted lines, q = 0.1.

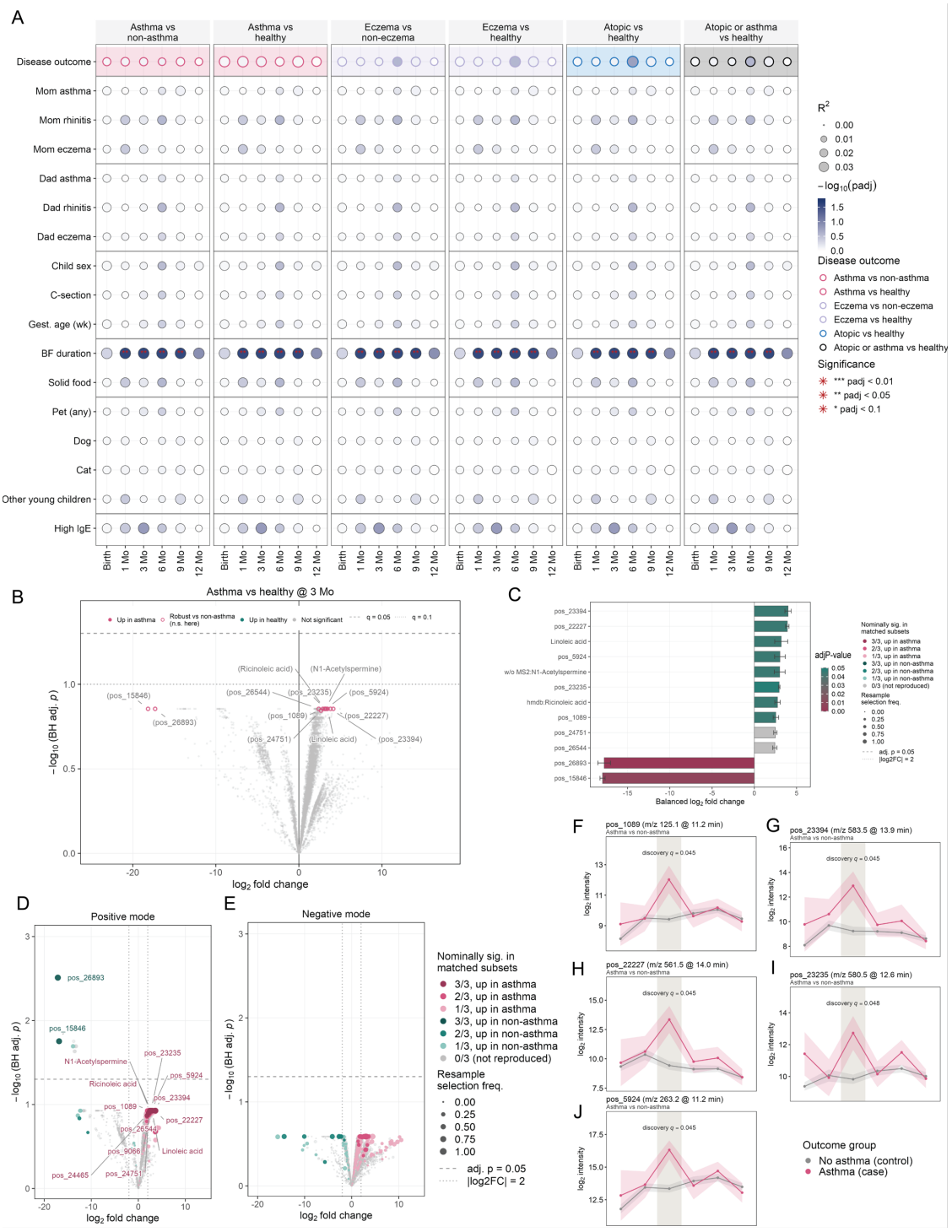

Figure S4.2 The early-life metabolomic asthma signal reproduces across reference groups, matched subsets, and ionization modes.

A) Cross-sectional PERMANOVA of Canberra metabolome structure at each visit for disease outcomes (top row, tinted) (Figure caption continued on the next page.)

(Figure caption continued from the previous page.) and for 16 clinical and environmental covariates. Shown for all six disease contrasts. Bubble size indicates R^2; fill indicates log10 BH-adjusted p, corrected across all 22 variables within a visit. B) Cross-sectional differential abundance at 3 months against the clean healthy reference rather than the broad non-asthma control. Rings mark the twelve features that were nominally significant in all three matched subsets of the broad analysis. C) Balanced log2 fold change of the twelve consistently selected features, sorted by effect size, calculated as the mean of the three matched-subset estimates; whiskers indicate ±1 SD across the three subsets. Bar fill indicates the discovery BH-adjusted p-value. D–E) Discovery differential abundance at asthma @ 3 months fitted and BH-adjusted independently within the positive-mode (D) and negative-mode (E) feature sets. Color indicates the direction of the effect and the number of sex and breast-feeding-matched case-control subsets in which the feature remained nominally significant; point size indicates selection frequency across many randomly matched resamples (stability selection). Fourteen features are nominally significant in all three matched subsets in positive mode and none in negative mode. F–J) Abundance across the first year of life for the remaining consistently selected features: pos_1089 (m/z 125.1 @ 11.2 min) (F), pos_23394 (m/z 583.5 @ 13.9 min) (G), pos_22227 (m/z 561.5 @ 14.0 min) (H), pos_23235 (m/z 560.5 @ 12.6 min) (I), and pos_5924 (m/z 263.2 @ 11.2 min) (J). Highlighted visit indicates the 3-month discovery window; q-value shown in each panel.

Tables

Table 4.1 Atopic/Asthma vs Healthy: Demographic and Exposure Comparison Table
Continuous variables: mean ± SD (Wilcoxon). Categorical variables: n (%) (Fisher's exact test). P < 0.05 shown in bold; unadjusted. 0 is no, 1 is yes.

Characteristic	Atopic/Asthma (n=41)	Healthy (n=36)	P
pet_exposure			**0.029**
0	18 (43.9%)	7 (19.4%)	
1	23 (56.1%)	29 (80.6%)	
dad_rhinitis_diagnosis_7			0.077
0	29 (70.7%)	20 (55.6%)	
1	8 (19.5%)	15 (41.7%)	
Missing	4 (9.8%)	1 (2.8%)	
pet_dog			0.077
0	18 (43.9%)	8 (22.2%)	
1	18 (43.9%)	22 (61.1%)	
Missing	5 (12.2%)	6 (16.7%)	
solid_food			0.137
0	2 (4.9%)	6 (16.7%)	
1	39 (95.1%)	30 (83.3%)	

Characteristic	Atopic/Asthma (n=41)	Healthy (n=36)	P
gestational_age_weeks	38.90 ± 1.37	39.11 ± 2.16	0.184
mom_asthma_diagnosis_12			0.448
0	10 (24.4%)	12 (33.3%)	
1	31 (75.6%)	23 (63.9%)	
Missing	0 (0.0%)	1 (2.8%)	
mom_eczema_diagnosis_21			0.448
0	31 (75.6%)	23 (63.9%)	
1	10 (24.4%)	12 (33.3%)	
Missing	0 (0.0%)	1 (2.8%)	
child_sex			0.496
1	22 (53.7%)	16 (44.4%)	
2	19 (46.3%)	20 (55.6%)	
c_section			0.604
1	32 (78.0%)	26 (72.2%)	
2	9 (22.0%)	10 (27.8%)	
bf_duration	7.68 ± 5.18	8.03 ± 4.86	0.649
num_child_not_toilet_trained	1.15 ± 0.48	1.17 ± 0.38	0.677
dad_eczema_diagnosis_13			0.761
0	33 (80.5%)	28 (77.8%)	
1	6 (14.6%)	7 (19.4%)	
Missing	2 (4.9%)	1 (2.8%)	
bf_6mo_min			0.805
FALSE	13 (31.7%)	10 (27.8%)	
TRUE	28 (68.3%)	26 (72.2%)	
dad_asthma_diagnosis_4			0.815
0	26 (63.4%)	21 (58.3%)	
1	15 (36.6%)	15 (41.7%)	
mom_rhinitis_diagnosis_15			1.000
0	21 (51.2%)	18 (50.0%)	
1	19 (46.3%)	16 (44.4%)	
Missing	1 (2.4%)	2 (5.6%)	
pet_cat			1.000
0	19 (46.3%)	21 (58.3%)	
1	10 (24.4%)	10 (27.8%)	
Missing	12 (29.3%)	5 (13.9%)	

Table 4.2 Atopic vs Healthy: Demographic and Exposure Comparison Table

Continuous variables: mean ± SD (Wilcoxon). Categorical variables: n (%) (Fisher's exact test). P < 0.05 shown in bold; unadjusted. 0 is no, 1 is yes.

Characteristic	Atopic (n=36)	Healthy (n=36)	P
pet_exposure			**0.013**
0	18 (50.0%)	7 (19.4%)	
1	18 (50.0%)	29 (80.6%)	
dad_rhinitis_diagnosis_7			0.076
0	25 (69.4%)	20 (55.6%)	
1	7 (19.4%)	15 (41.7%)	
Missing	4 (11.1%)	1 (2.8%)	
pet_dog			**0.037**
0	17 (47.2%)	8 (22.2%)	
1	14 (38.9%)	22 (61.1%)	
Missing	5 (13.9%)	6 (16.7%)	
solid_food			0.260
0	2 (5.6%)	6 (16.7%)	
1	34 (94.4%)	30 (83.3%)	
gestational_age_weeks	38.83 ± 1.40	39.11 ± 2.16	0.146
mom_asthma_diagnosis_12			0.443
0	9 (25.0%)	12 (33.3%)	
1	27 (75.0%)	23 (63.9%)	
Missing	0 (0.0%)	1 (2.8%)	
mom_eczema_diagnosis_21			0.300
0	28 (77.8%)	23 (63.9%)	
1	8 (22.2%)	12 (33.3%)	
Missing	0 (0.0%)	1 (2.8%)	
child_sex			0.346
1	21 (58.3%)	16 (44.4%)	
2	15 (41.7%)	20 (55.6%)	
c_section			0.786
1	28 (77.8%)	26 (72.2%)	
2	8 (22.2%)	10 (27.8%)	
bf_duration	7.44 ± 5.26	8.03 ± 4.86	0.497
num_child_not_toilet_trained	1.14 ± 0.49	1.17 ± 0.38	0.608
dad_eczema_diagnosis_13			1.000
0	28 (77.8%)	28 (77.8%)	

Characteristic	Atopic (n=36)	Healthy (n=36)	P
1	6 (16.7%)	7 (19.4%)	
Missing	2 (5.6%)	1 (2.8%)	
bf_6mo_min			0.798
FALSE	12 (33.3%)	10 (27.8%)	
TRUE	24 (66.7%)	26 (72.2%)	
dad_asthma_diagnosis_4			0.627
0	24 (66.7%)	21 (58.3%)	
1	12 (33.3%)	15 (41.7%)	
mom_rhinitis_diagnosis_15			1.000
0	19 (52.8%)	18 (50.0%)	
1	16 (44.4%)	16 (44.4%)	
Missing	1 (2.8%)	2 (5.6%)	
pet_cat			0.777
0	18 (50.0%)	21 (58.3%)	
1	7 (19.4%)	10 (27.8%)	
Missing	11 (30.6%)	5 (13.9%)	

Table 4.3 Asthma vs Non-asthma: Demographic and Exposure Comparison Table
Continuous variables: mean ± SD (Wilcoxon). Categorical variables: n (%) (Fisher's exact test). P < 0.05 shown in bold; unadjusted. 0 is no, 1 is yes.

Characteristic	Asthma (n=11)	Non-asthma (n=72)	P
pet_exposure			0.743
0	3 (27.3%)	25 (34.7%)	
1	8 (72.7%)	47 (65.3%)	
dad_rhinitis_diagnosis_7			0.319
0	9 (81.8%)	43 (59.7%)	
1	2 (18.2%)	24 (33.3%)	
Missing	0 (0.0%)	5 (6.9%)	
pet_dog			1.000
0	4 (36.4%)	25 (34.7%)	
1	6 (54.5%)	37 (51.4%)	
Missing	1 (9.1%)	10 (13.9%)	
solid_food			1.000
0	1 (9.1%)	7 (9.7%)	

Characteristic	Asthma (n=11)	Non-asthma (n=72)	P
1	10 (90.9%)	65 (90.3%)	
gestational_age_weeks	39.00 ± 1.55	39.00 ± 1.79	0.851
mom_asthma_diagnosis_12			1.000
0	3 (27.3%)	21 (29.2%)	
1	8 (72.7%)	50 (69.4%)	
Missing	0 (0.0%)	1 (1.4%)	
mom_eczema_diagnosis_21			0.277
0	6 (54.5%)	53 (73.6%)	
1	5 (45.5%)	18 (25.0%)	
Missing	0 (0.0%)	1 (1.4%)	
child_sex			0.756
1	5 (45.5%)	37 (51.4%)	
2	6 (54.5%)	35 (48.6%)	
c_section			0.264
1	7 (63.6%)	57 (79.2%)	
2	4 (36.4%)	15 (20.8%)	
bf_duration	7.27 ± 5.24	7.69 ± 5.09	0.733
num_child_not_toilet_trained	1.27 ± 0.65	1.15 ± 0.43	0.301
dad_eczema_diagnosis_13			0.680
0	10 (90.9%)	55 (76.4%)	
1	1 (9.1%)	14 (19.4%)	
Missing	0 (0.0%)	3 (4.2%)	
bf_6mo_min			0.743
FALSE	4 (36.4%)	23 (31.9%)	
TRUE	7 (63.6%)	49 (68.1%)	
dad_asthma_diagnosis_4			1.000
0	7 (63.6%)	43 (59.7%)	
1	4 (36.4%)	29 (40.3%)	
mom_rhinitis_diagnosis_15			0.057
0	3 (27.3%)	41 (56.9%)	
1	8 (72.7%)	28 (38.9%)	
Missing	0 (0.0%)	3 (4.2%)	
pet_cat			0.196
0	3 (27.3%)	42 (58.3%)	
1	4 (36.4%)	17 (23.6%)	
Missing	4 (36.4%)	13 (18.1%)	

Table 4.4 Eczema vs Non-eczema: Demographic and Exposure Comparison Table
Continuous variables: mean ± SD (Wilcoxon). Categorical variables: n (%) (Fisher's exact test). P < 0.05 shown in bold; unadjusted. 0 is no, 1 is yes.

Characteristic	Eczema (n=28)	Non-eczema (n=55)	P
pet_exposure			0.229
0	12 (42.9%)	16 (29.1%)	
1	16 (57.1%)	39 (70.9%)	
dad_rhinitis_diagnosis_7			0.123
0	20 (71.4%)	32 (58.2%)	
1	5 (17.9%)	21 (38.2%)	
Missing	3 (10.7%)	2 (3.6%)	
pet_dog			0.207
0	13 (46.4%)	16 (29.1%)	
1	12 (42.9%)	31 (56.4%)	
Missing	3 (10.7%)	8 (14.5%)	
solid_food			0.711
0	2 (7.1%)	6 (10.9%)	
1	26 (92.9%)	49 (89.1%)	
gestational_age_weeks	39.07 ± 1.21	38.96 ± 1.98	0.683
mom_asthma_diagnosis_12			1.000
0	8 (28.6%)	16 (29.1%)	
1	20 (71.4%)	38 (69.1%)	
Missing	0 (0.0%)	1 (1.8%)	
mom_eczema_diagnosis_21			1.000
0	20 (71.4%)	39 (70.9%)	
1	8 (28.6%)	15 (27.3%)	
Missing	0 (0.0%)	1 (1.8%)	
child_sex			0.817
1	15 (53.6%)	27 (49.1%)	
2	13 (46.4%)	28 (50.9%)	
c_section			0.786
1	21 (75.0%)	43 (78.2%)	
2	7 (25.0%)	12 (21.8%)	
bf_duration	7.79 ± 5.24	7.56 ± 5.04	0.923
num_child_not_toilet_trained	1.07 ± 0.38	1.22 ± 0.50	0.190
dad_eczema_diagnosis_13			1.000
0	22 (78.6%)	43 (78.2%)	

Characteristic	Eczema (n=28)	Non-eczema (n=55)	P
1	5 (17.9%)	10 (18.2%)	
Missing	1 (3.6%)	2 (3.6%)	
bf_6mo_min			1.000
FALSE	9 (32.1%)	18 (32.7%)	
TRUE	19 (67.9%)	37 (67.3%)	
dad_asthma_diagnosis_4			0.642
0	18 (64.3%)	32 (58.2%)	
1	10 (35.7%)	23 (41.8%)	
mom_rhinitis_diagnosis_15			1.000
0	15 (53.6%)	29 (52.7%)	
1	12 (42.9%)	24 (43.6%)	
Missing	1 (3.6%)	2 (3.6%)	
pet_cat			1.000
0	14 (50.0%)	31 (56.4%)	
1	6 (21.4%)	15 (27.3%)	
Missing	8 (28.6%)	9 (16.4%)	

Table 4.5 Asthma vs Healthy: Demographic and Exposure Comparison Table

Continuous variables: mean ± SD (Wilcoxon). Categorical variables: n (%) (Fisher's exact test). P < 0.05 shown in bold; unadjusted. 0 is no, 1 is yes.

Characteristic	Asthma (n=11)	Healthy (n=36)	P
pet_exposure			0.679
0	3 (27.3%)	7 (19.4%)	
1	8 (72.7%)	29 (80.6%)	
dad_rhinitis_diagnosis_7			0.172
0	9 (81.8%)	20 (55.6%)	
1	2 (18.2%)	15 (41.7%)	
Missing	0 (0.0%)	1 (2.8%)	
pet_dog			0.451
0	4 (36.4%)	8 (22.2%)	
1	6 (54.5%)	22 (61.1%)	
Missing	1 (9.1%)	6 (16.7%)	
solid_food			1.000
0	1 (9.1%)	6 (16.7%)	
1	10 (90.9%)	30 (83.3%)	

Characteristic	Asthma (n=11)	Healthy (n=36)	P
gestational_age_weeks	39.00 ± 1.55	39.11 ± 2.16	0.568
mom_asthma_diagnosis_12			1.000
0	3 (27.3%)	12 (33.3%)	
1	8 (72.7%)	23 (63.9%)	
Missing	0 (0.0%)	1 (2.8%)	
mom_eczema_diagnosis_21			0.722
0	6 (54.5%)	23 (63.9%)	
1	5 (45.5%)	12 (33.3%)	
Missing	0 (0.0%)	1 (2.8%)	
child_sex			1.000
1	5 (45.5%)	16 (44.4%)	
2	6 (54.5%)	20 (55.6%)	
c_section			0.710
1	7 (63.6%)	26 (72.2%)	
2	4 (36.4%)	10 (27.8%)	
bf_duration	7.27 ± 5.24	8.03 ± 4.86	0.585
num_child_not_toilet_trained	1.27 ± 0.65	1.17 ± 0.38	0.422
dad_eczema_diagnosis_13			0.658
0	10 (90.9%)	28 (77.8%)	
1	1 (9.1%)	7 (19.4%)	
Missing	0 (0.0%)	1 (2.8%)	
bf_6mo_min			0.710
FALSE	4 (36.4%)	10 (27.8%)	
TRUE	7 (63.6%)	26 (72.2%)	
dad_asthma_diagnosis_4			1.000
0	7 (63.6%)	21 (58.3%)	
1	4 (36.4%)	15 (41.7%)	
mom_rhinitis_diagnosis_15			0.177
0	3 (27.3%)	18 (50.0%)	
1	8 (72.7%)	16 (44.4%)	
Missing	0 (0.0%)	2 (5.6%)	
pet_cat			0.387
0	3 (27.3%)	21 (58.3%)	
1	4 (36.4%)	10 (27.8%)	
Missing	4 (36.4%)	5 (13.9%)	

Table 4.6 Eczema vs Healthy: Demographic and Exposure Comparison Table

Continuous variables: mean ± SD (Wilcoxon). Categorical variables: n (%) (Fisher's exact test). P < 0.05 shown in bold; unadjusted. 0 is no, 1 is yes.

Characteristic	Eczema (n=28)	Healthy (n=36)	P
pet_exposure			0.056
0	12 (42.9%)	7 (19.4%)	
1	16 (57.1%)	29 (80.6%)	
dad_rhinitis_diagnosis_7			0.096
0	20 (71.4%)	20 (55.6%)	
1	5 (17.9%)	15 (41.7%)	
Missing	3 (10.7%)	1 (2.8%)	
pet_dog			0.094
0	13 (46.4%)	8 (22.2%)	
1	12 (42.9%)	22 (61.1%)	
Missing	3 (10.7%)	6 (16.7%)	
solid_food			0.448
0	2 (7.1%)	6 (16.7%)	
1	26 (92.9%)	30 (83.3%)	
gestational_age_weeks	39.07 ± 1.21	39.11 ± 2.16	0.365
mom_asthma_diagnosis_12			0.786
0	8 (28.6%)	12 (33.3%)	
1	20 (71.4%)	23 (63.9%)	
Missing	0 (0.0%)	1 (2.8%)	
mom_eczema_diagnosis_21			0.786
0	20 (71.4%)	23 (63.9%)	
1	8 (28.6%)	12 (33.3%)	
Missing	0 (0.0%)	1 (2.8%)	
child_sex			0.615
1	15 (53.6%)	16 (44.4%)	
2	13 (46.4%)	20 (55.6%)	
c_section			1.000
1	21 (75.0%)	26 (72.2%)	
2	7 (25.0%)	10 (27.8%)	

Characteristic	Eczema (n=28)	Healthy (n=36)	P
bf_duration	7.79 ± 5.24	8.03 ± 4.86	0.767
num_child_not_toilet_trained	1.07 ± 0.38	1.17 ± 0.38	0.339
dad_eczema_diagnosis_13			1.000
0	22 (78.6%)	28 (77.8%)	
1	5 (17.9%)	7 (19.4%)	
Missing	1 (3.6%)	1 (2.8%)	
bf_6mo_min			0.786
FALSE	9 (32.1%)	10 (27.8%)	
TRUE	19 (67.9%)	26 (72.2%)	
dad_asthma_diagnosis_4			0.797
0	18 (64.3%)	21 (58.3%)	
1	10 (35.7%)	15 (41.7%)	
mom_rhinitis_diagnosis_15			1.000
0	15 (53.6%)	18 (50.0%)	
1	12 (42.9%)	16 (44.4%)	
Missing	1 (3.6%)	2 (5.6%)	
pet_cat			1.000
0	14 (50.0%)	21 (58.3%)	
1	6 (21.4%)	10 (27.8%)	
Missing	8 (28.6%)	5 (13.9%)	

Table S4.1. Stable isotope-labelled internal standards.
Added to the extraction buffer for untargeted metabolomics, with molecular formula, ionization mode, precursor m/z, and retention times. Precursor m/z values are monoisotopic $[M+H]^+$ (ESI+) and $[M-H]^-$ (ESI−) ions, calculated from the labelled molecular formula. Retention times are those reported by the Quantitative Metabolomics and Analytics Core (QMAC) for the acquisition method used in this study.

Compound	Isotope label	Molecular formula	Mode	Precursor m/z	RT (min)
Untargeted assay — stable isotope-labelled amino acids and aromatic standard					
L-Alanine	13C3,15N	13C3H715NO2	ESI+	94.0621	1.75
			ESI−	92.0475	1.72
L-Arginine	13C6	13C6H14N4O2	ESI+	181.1391	1.64
			ESI−	179.1245	1.62
L-Aspartic acid	13C4	13C4H7NO4	ESI+	138.0582	2.55
			ESI−	136.0437	1.74

Compound	Isotope label	Molecular formula	Mode	Precursor m/z	RT (min)
L-Glutamic acid	13C5	13C5H9NO4	ESI+	153.0772	1.79
			ESI−	151.0627	1.79
Glycine	13C2,15N	13C2H515NO2	ESI+	79.0430	1.72
			ESI−	77.0285	1.69
L-Histidine	13C6	13C6H9N3O2	ESI+	162.0969	1.62
			ESI−	160.0823	1.61
L-Isoleucine	13C6,15N	13C6H1315NO2	ESI+	139.1191	3.36
			ESI−	137.1045	3.36
L-Leucine	13C6,15N	13C6H1315NO2	ESI+	139.1191	3.74
			ESI−	137.1045	3.74
L-Lysine	13C6	13C6H14N2O2	ESI+	153.1329	1.58
			ESI−	151.1184	1.54
L-Methionine	13C5,15N	13C5H1115NSO2	ESI+	156.0721	2.55
			ESI−	154.0576	2.54
L-Phenylalanine	ring-13C6	C313C6H11NO2	ESI+	172.1064	7.81
			ESI−	170.0918	7.75
L-Proline	13C5	13C5H9NO2	ESI+	121.0874	1.96
			ESI−	119.0728	2.02
L-Serine	13C3,15N	13C3H715NO3	ESI+	110.0570	1.72
			ESI−	108.0424	1.69
L-Threonine	13C4	13C4H9NO3	ESI+	124.0789	1.76
			ESI−	122.0644	1.74
L-Tyrosine	ring-13C6	C313C6H11NO3	ESI+	188.1013	4.37
			ESI−	186.0867	4.37
L-Valine	13C5	13C5H11NO2	ESI+	123.1030	2.18
			ESI−	121.0885	2.15
L-Cystine	13C6,15N2	13C6H1215N2O4S2	ESI+	249.0453	1.70
			ESI−	247.0308	1.69
2-Amino-3-bromo-5-methylbenzoic acid	—	C8H8NO2Br	ESI+	229.9811	10.50
			ESI−	227.9666	10.50

Table S4.2. Deuterated internal standards for the targeted eicosanoid assay.
Added to the extraction buffer for the targeted bioactive-lipid-mediator assay, with molecular formula, ionization mode and precursor m/z. Precursor m/z values are monoisotopic $[M+H]^+$ (ESI+) and $[M-H]^-$ (ESI−) ions, calculated from the labelled molecular formula.

Compound	Isotope label	Molecular formula	Mode	Precursor m/z
Targeted eicosanoid assay — deuterated standards				
Arachidonic acid	d8	C20H24D8O2	ESI+	313.2977
			ESI−	311.2832
Linoleic acid	d4	C18H28D4O2	ESI+	285.2726
			ESI−	283.2581
Dihomo-γ-linolenic acid	d6	C20H28D6O2	ESI+	313.3008
			ESI−	311.2863
Prostaglandin E2	d4	C20H28D4O5	ESI+	357.2574
			ESI−	355.2428
5-HETE	d8	C20H24D8O3	ESI+	329.2926
			ESI−	327.2781
13-HODE	d4	C18H28D4O3	ESI+	301.2675
			ESI−	299.2530
9-HODE	d4	C18H28D4O3	ESI+	301.2675
			ESI−	299.2530
9,10-DiHOME	d4	C18H30D4O4	ESI+	319.2781
			ESI−	317.2635
12,13-DiHOME	d4	C18H30D4O4	ESI+	319.2781
			ESI−	317.2635
Lipoxin A4	d5	C20H27D5O5	ESI+	358.2636
			ESI−	356.2491

References

1. National Center for Health Statistics. *Percentage of Ever Having Asthma for Children under Age 18 Years, United States, 2019—2024*. https://wwwn.cdc.gov/NHISDataQueryTool/SHS_child/index.html (2024).
2. Coca, A. F. & Cooke, R. A. On the classification of the phenomena of hypersensitiveness. *J. Immunol.* **8**, 163–182 (1923).
3. Coombs, R. R. A. Classification of allergic reactions responsible for clinical hypersensitivity and diseases. *Clinical aspects of immunology* (1975).
4. Platts-Mills, T. A. The role of immunoglobulin E in allergy and asthma. *Am. J. Respir. Crit. Care Med.* **164**, S1-5 (2001).
5. Nemet, S. *et al.* Clinical significance of very high IgE levels (≥1000 IU/mL): Population-based study of 118,211 adults. *J. Allergy Clin. Immunol. Glob.* **4**, 100403 (2025).
6. Tokura, Y. Extrinsic and intrinsic types of atopic dermatitis. *J. Dermatol. Sci.* **58**, 1–7 (2010).
7. Blevins, S. M. & Bronze, M. S. Robert Koch and the “golden age” of bacteriology. *Int. J. Infect. Dis.* **14**, e744-51 (2010).
8. Chonan, O. Yakult’s research activities: Inheritance and practice of Shirota-ism. https://www.nature.com/articles/d42473-020-00368-6.
9. Gasbarrini, G., Bonvicini, F. & Gramenzi, A. Probiotics history. *J. Clin. Gastroenterol.* **50 Suppl 2, Proceedings from the 8th Probiotics, Prebiotics&New Foods for Microbiota and Human Health meeting held in Rome, Italy on September 13-15, 2015**, S116–S119 (2016).
10. Strachan, D. P. Hay fever, hygiene, and household size. *BMJ* **299**, 1259–1260 (1989).

11. Publications – LHNCBC: 'Ome Sweet 'Omics - A Genealogical Treasury of Words. https://lhncbc.nlm.nih.gov/LHC-publications/pubs/OmeSweetOmicsAGenealogicalTreasuryofWords.html.
12. Hamilton, J. D., Ungar, B. & Guttman-Yassky, E. Drug evaluation review: dupilumab in atopic dermatitis. *Immunotherapy* **7**, 1043–1058 (2015).
13. Teo, S. M. *et al.* The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development. *Cell Host Microbe* **17**, 704–715 (2015).
14. Tang, H. H. F. *et al.* Developmental patterns in the nasopharyngeal microbiome during infancy are associated with asthma risk. *The Journal of Allergy and Clinical Immunology* **147**, 1683–1691 (2021).
15. McCauley, K. *et al.* Distinct nasal airway bacterial microbiotas differentially relate to exacerbation in pediatric patients with asthma. *The Journal of Allergy and Clinical Immunology* **144**, 1187–1197 (2019).
16. Hornef, M. W. & Torow, N. "Layered immunity" and the "neonatal window of opportunity" - timed succession of non-redundant phases to establish mucosal host-microbial homeostasis after birth. *Immunology* **159**, 15–25 (2020).
17. Gensollen, T., Iyer, S. S., Kasper, D. L. & Blumberg, R. S. How colonization by microbiota in early life shapes the immune system. *Science* **352**, 539–544 (2016).
18. Arrieta, M.-C. *et al.* Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Sci. Transl. Med.* **7**, 307ra152 (2015).

19. Cahenzli, J., Köller, Y., Wyss, M., Geuking, M. B. & McCoy, K. D. Intestinal microbial diversity during early-life colonization shapes long-term IgE levels. *Cell Host Microbe* **14**, 559–570 (2013).

20. Jartti, T. & Gern, J. E. Role of viral infections in the development and exacerbation of asthma in children. *The Journal of Allergy and Clinical Immunology* **140**, 895–906 (2017).

21. Kusel, M. M. H. *et al.* Early-life respiratory viral infections, atopic sensitization, and risk of subsequent development of persistent asthma. *The Journal of Allergy and Clinical Immunology* **119**, 1105–1110 (2007).

22. Rosas-Salazar, C. *et al.* Respiratory syncytial virus infection during infancy and asthma during childhood in the USA (INSPIRE): a population-based, prospective birth cohort study. *Lancet* **401**, 1669–1680 (2023).

23. Serebrisky, D. & Wiznia, A. Pediatric asthma: A global epidemic. *Ann. Glob. Health* **85**, (2019).

24. Lim, R. H., Kobzik, L. & Dahl, M. Risk for asthma in offspring of asthmatic mothers versus fathers: a meta-analysis. *PLoS One* **5**, e10134 (2010).

25. Stein, M. M. *et al.* A decade of research on the 17q12-21 asthma locus: Piecing together the puzzle. *The Journal of Allergy and Clinical Immunology* **142**, 749-764.e3 (2018).

26. Kim, K. W. *et al.* Genome-wide association study identifies TNFSF15 associated with childhood asthma. *Allergy* **77**, 218–229 (2022).

27. Calışkan, M. *et al.* Rhinovirus wheezing illness and genetic risk of childhood-onset asthma. *N. Engl. J. Med.* **368**, 1398–1407 (2013).

28. Schoettler, N. *et al.* New insights relating gasdermin B to the onset of childhood asthma. *Am. J. Respir. Cell Mol. Biol.* **67**, 430–437 (2022).

29. Hallmark, B. *et al.* Chromosome 17q12-21 Variants Are Associated with Multiple Wheezing Phenotypes in Childhood. *Am. J. Respir. Crit. Care Med.* **203**, 864–870 (2021).

30. Ober, C. *et al.* Expression quantitative trait locus fine mapping of the 17q12-21 asthma locus in African American children: a genetic association and gene expression study. *Lancet Respir. Med.* **8**, 482–492 (2020).

31. Illi, S. *et al.* Immune responsiveness to LPS determines risk of childhood wheeze and asthma in 17q21 risk allele carriers. *Am. J. Respir. Crit. Care Med.* **205**, 641–650 (2022).

32. Luthers, C. R., Dunn, T. M. & Snow, A. L. ORMDL3 and asthma: linking sphingolipid regulation to altered T cell function. *Front. Immunol.* **11**, 597945 (2020).

33. Worgall, T. S. Sphingolipids and Asthma. *Adv. Exp. Med. Biol.* **1372**, 145–155 (2022).

34. Kunz, T. C. & Kozjak-Pavlovic, V. Diverse facets of sphingolipid involvement in bacterial infections. *Front. Cell Dev. Biol.* **7**, 203 (2019).

35. Guo, F. *et al.* Asthma susceptibility gene ORMDL3 promotes autophagy in human bronchial epithelium. *Am. J. Respir. Cell Mol. Biol.* **66**, 661–670 (2022).

36. Chen, Q. *et al.* GSDMB promotes non-canonical pyroptosis by enhancing caspase-4 activity. *J. Mol. Cell Biol.* **11**, 496–508 (2019).

37. Zhou, Z. *et al.* Granzyme A from cytotoxic lymphocytes cleaves GSDMB to trigger pyroptosis in target cells. *Science* **368**, (2020).

38. Panganiban, R. A. *et al.* A functional splice variant associated with decreased asthma risk abolishes the ability of gasdermin B to induce epithelial cell pyroptosis. *The Journal of Allergy and Clinical Immunology* **142**, 1469-1478.e2 (2018).

39. Rana, N. *et al.* GSDMB is increased in IBD and regulates epithelial restitution/repair independent of pyroptosis. *Cell* **185**, 283-298.e17 (2022).

40. Das, S. *et al.* GSDMB induces an asthma phenotype characterized by increased airway responsiveness and remodeling without lung inflammation. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 13132–13137 (2016).

41. Hansen, J. M. *et al.* Pathogenic ubiquitination of GSDMB inhibits NK cell bactericidal functions. *Cell* **184**, 3178-3191.e18 (2021).

42. Thorsen, J. *et al.* The airway microbiota of neonates colonized with asthma-associated pathogenic bacteria. *Nat. Commun.* **14**, 6668 (2023).

43. Sunde, R. B. *et al.* Bacterial colonisation of the airway in neonates and risk of asthma and allergy until age 18 years. *The European Respiratory Journal* **63**, (2024).

44. Lemanske, R. F. The childhood origins of asthma (COAST) study. *Pediatric Allergy and Immunology* **13**, 38–43 (2002).

45. Gern, J. E. *et al.* The Urban Environment and Childhood Asthma (URECA) birth cohort study: design, methods, and study population. *BMC Pulm. Med.* **9**, 17 (2009).

46. DeAngelis, K. M., Silver, W. L., Thompson, A. W. & Firestone, M. K. Microbial communities acclimate to recurring changes in soil redox potential status. *Environ. Microbiol.* **12**, 3137–3149 (2010).

47. Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl 1**, 4516–4522 (2011).

48. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).

49. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590-6 (2013).

50. Fujimura, K. E. *et al.* Neonatal gut microbiota associates with childhood multisensitized atopy and T cell differentiation. *Nat. Med.* **22**, 1187–1191 (2016).

51. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).

52. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

53. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

54. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

55. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).

56. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

57. Xie, Z. *et al.* Gene set knowledge discovery with Enrichr. *Curr. Protoc.* **1**, e90 (2021).

58. Kauffmann, F. & Demenais, F. Gene-environment interactions in asthma and allergic diseases: challenges and perspectives. *The Journal of Allergy and Clinical Immunology* **130**, 1229–40; quiz 1241 (2012).

59. Jakwerth, C. A. *et al.* 17q21 variants disturb mucosal host defense in childhood asthma. *Am. J. Respir. Crit. Care Med.* **209**, 947–959 (2024).

60. Abdelshafy, A. M., Neetoo, H. & Al-Asmari, F. Antimicrobial Activity of Hydrogen Peroxide for Application in Food Safety and COVID-19 Mitigation: An Updated Review. *J. Food Prot.* **87**, 100306 (2024).

61. Sheng, Y. H. & Hasnain, S. Z. Mucus and mucins: the underappreciated host defence system. *Front. Cell. Infect. Microbiol.* **12**, 856962 (2022).

62. Russjan, E. & Kaczyńska, K. Beneficial Effects of Neurotensin in Murine Model of Hapten-Induced Asthma. *Int. J. Mol. Sci.* **20**, (2019).

63. Kobayashi, Y. The role of chemokines in neutrophil biology. *Front. Biosci.* **13**, 2400–2407 (2008).

64. Hernandez-Pacheco, N., Kere, M. & Melén, E. Gene-environment interactions in childhood asthma revisited; expanding the interaction concept. *Pediatric Allergy and Immunology* **33**, e13780 (2022).

65. Brugger, S. D. *et al.* Dolosigranulum pigrum Cooperation and Competition in Human Nasal Microbiota. *mSphere* **5**, (2020).

66. Tchoupa, A. K., Schuhmacher, T. & Hauck, C. R. Signaling by epithelial members of the CEACAM family - mucosal docking sites for pathogenic bacteria. *Cell Communication and Signaling* **12**, 27 (2014).

67. Ehre, C. *et al.* Overexpressing mouse model demonstrates the protective role of Muc5ac in the lungs. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 16528–16533 (2012).

68. Robbins, R. A., Nelson, K. J., Gossman, G. L. & Rubinstein, I. Neurotensin stimulates neutrophil adherence to bronchial epithelial cells in vitro. *Life Sci.* **56**, 1353–1359 (1995).

69. Fletcher, A. A. & Evans, C. M. Regulation of microbial populations and immune functions by muc5b. *Ann. Am. Thorac. Soc.* **11**, S80–S80 (2014).

70. Cho, H.-Y. *et al.* Role for Mucin-5AC in Upper and Lower Airway Pathogenesis in Mice. *Toxicol. Pathol.* **49**, 1077–1099 (2021).

71. Alkhouri, H., Moir, L. M., Armour, C. L. & Hughes, J. M. CXCL1 is a negative regulator of mast cell chemotaxis to airway smooth muscle cell products in vitro. *Clinical and Experimental Allergy* **44**, 381–392 (2014).

72. Lieber, M., Smith, B., Szakal, A., Nelson-Rees, W. & Todaro, G. A continuous tumor-cell line from a human lung carcinoma with properties of type II alveolar epithelial cells. *Int. J. Cancer* **17**, 62–70 (1976).

73. Durack, J. *et al.* Delayed gut microbiota development in high-risk for asthma infants is temporarily modifiable by Lactobacillus supplementation. *Nat. Commun.* **9**, 707 (2018).

74. Ogulur, I. *et al.* Type 2 immunity in allergic diseases. *Cell. Mol. Immunol.* **22**, 211–242 (2025).

75. Ober, C. Asthma genetics in the post-GWAS era. *Ann. Am. Thorac. Soc.* **13 Suppl 1**, S85-90 (2016).

76. Stokholm, J. *et al.* Maturation of the gut microbiome and risk of asthma in childhood. *Nat. Commun.* **9**, 141 (2018).

77. Cabana, M. D., McKean, M., Wong, A. R., Chao, C. & Caughey, A. B. Examining the hygiene hypothesis: the Trial of Infant Probiotic Supplementation. *Paediatr. Perinat. Epidemiol.* **21 Suppl 3**, 23–28 (2007).

78. Shenhav, L., Azad, M. B. & Silverman, M. A. Human milk as a microbial pacemaker. *Cell Host Microbe* **33**, 762–765 (2025).

79. Andrews, S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*. (2010).

80. Bushnell, B. *BBMap: A Fast, Accurate, Splice-Aware Aligner*. (2014).

81. Joshi, N. A. & Fass, J. N. *Sickle: A Sliding-Window, Adaptive, Quality-Based Trimming Tool for FastQ Files*. (2011).

82. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).

83. Chen, L. & Banfield, J. F. COBRA improves the completeness and contiguity of viral genomes assembled from metagenomes. *Nat. Microbiol.* **9**, 737–750 (2024).

84. *Fast Gapp Fast Gapped-Read Alignment with Bowtie 2ed-Read Alignment with Bowtie 2*.

85. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

86. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).

87. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).

88. Nissen, J. N. *et al.* Improved metagenome binning and assembly using deep variational autoencoders. *Nat. Biotechnol.* **39**, 555–560 (2021).

89. Zaharia, M. *et al.* Faster and more accurate sequence alignment with SNAP. *arXiv [cs.DS]* (2011) doi:10.48550/arXiv.1111.5572.

90. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).

91. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).

92. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).

93. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).

94. Shulgina, Y. & Eddy, S. R. Codetta: predicting the genetic code from nucleotide sequence. *Bioinformatics* **39**, btac802 (2023).

95. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

96. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* **49**, 9077–9096 (2021).

97. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).

98. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

99. Aroney, S. T. N. *et al.* CoverM: read alignment statistics for metagenomics. *Bioinformatics* **41**, btaf147 (2025).

100. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer, New York, NY, 2009).

101. Caceres, D. *et al. Vegan: Community Ecology Package*. (2025).

102. Kursa, M. B. & Rudnicki, W. R. Feature Selection with theBorutaPackage. *J. Stat. Softw.* **36**, 1–13 (2010).

103. Griffith, D. M., Veech, J. A. & Marsh, C. J. cooccur: Probabilistic Species Co-Occurrence Analysis inR. *J. Stat. Softw.* **69**, 1–17 (2016).

104. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

105. ropls: PCA, PLS(-DA) and OPLS(-DA) for multivariate analysis and feature selection of omics data. https://bioc.r-universe.dev/articles/ropls/ropls-vignette.html.

106. Hoskinson, C. *et al.* Delayed gut microbiota maturation in the first year of life is a hallmark of pediatric allergic disease. *Nat. Commun.* **14**, 4785 (2023).

107. Mu, M. *et al.* Multiomics analysis of Bacteroides cellulosilyticus anticolitis via gut Microbiota metabolite-mediated PI3K-Akt signaling pathway. *J. Agric. Food Chem.* **73**, 16333–16347 (2025).

108. McNulty, N. P. *et al.* Effects of diet on resource utilization by a model human gut microbiota containing Bacteroides cellulosilyticus WH2, a symbiont with an extensive glycobiome. *PLoS Biol.* **11**, e1001637 (2013).

109. Lu, Y., Cui, A. & Zhang, X. Commensal microbiota-derived metabolite agmatine triggers inflammation to promote colorectal tumorigenesis. *Gut Microbes* **16**, 2348441 (2024).

110. Levan, S. R. *et al.* Elevated faecal 12,13-diHOME concentration in neonates at high risk for asthma is produced by gut bacteria and impedes immune tolerance. *Nat. Microbiol.* **4**, 1851–1861 (2019).

111. Lin, D. L. *et al.* 12,13-diHOME promotes inflammatory macrophages and epigenetically modifies their capacity to respond to microbes and allergens. *J. Immunol. Res.* **2024**, 2506586 (2024).

112. Dharmage, S. C., Perret, J. L. & Custovic, A. Epidemiology of asthma in children and adults. *Front. Pediatr.* **7**, 246 (2019).

113. Midodzi, W. K., Rowe, B. H., Majaesic, C. M. & Senthilselvan, A. Reduced risk of physician-diagnosed asthma among children dwelling in a farming environment. *Respirology* **12**, 692–699 (2007).

114. Deckers, J., Marsland, B. J. & von Mutius, E. Protection against allergies: Microbes, immunity, and the farming effect. *Eur. J. Immunol.* **51**, 2387–2398 (2021).

115. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).

116. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).

117. Wang, Z., Huang, P., You, R., Sun, F. & Zhu, S. MetaBinner: a high-performance and stand-alone ensemble binning method to recover individual genomes from complex microbial communities. *Genome Biol.* **24**, 1 (2023).

118. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

119. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29-37 (2011).

120. Zheng, J. *et al.* dbCAN3: automated carbohydrate-active enzyme and substrate annotation. *Nucleic Acids Res.* **51**, W115–W121 (2023).

121. Schymanski, E. L. *et al.* Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ. Sci. Technol.* **48**, 2097–2098 (2014).

122. R Core Team. R: A Language and Environment for Statistical Computing. Preprint at https://www.R-project.org/ (2025).

123. Barrett, T. *et al.* data.table: Extension of `data.frame`. Preprint at https://doi.org/10.32614/CRAN.package.data.table (2026).

124. Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. dplyr: A Grammar of Data Manipulation. Preprint at https://doi.org/10.32614/CRAN.package.dplyr (2026).

125. Wickham, H., Vaughan, D. & Girlich, M. tidyr: Tidy Messy Data. Preprint at https://doi.org/10.32614/CRAN.package.tidyr (2025).

126. Wilke, C. O. cowplot: Streamlined Plot Theme and Plot Annotations for "ggplot2." Preprint at https://doi.org/10.32614/CRAN.package.cowplot (2025).

127. Slowikowski, K. ggrepel: Automatically Position Non-Overlapping Text Labels with "ggplot2." Preprint at https://doi.org/10.32614/CRAN.package.ggrepel (2026).

128. Wickham, H., Pedersen, T. L. & Seidel, D. scales: Scale Functions for Visualization. Preprint at https://doi.org/10.32614/CRAN.package.scales (2025).

129. Oksanen, J. *et al.* vegan: Community Ecology Package. Preprint at https://doi.org/10.32614/CRAN.package.vegan (2026).

130. Chen, Y., Chen, L., Lun, A. T. L., Baldoni, P. & Smyth, G. K. edgeR v4: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. *Nucleic Acids Research* vol. 53 gkaf018 Preprint at https://doi.org/10.1093/nar/gkaf018 (2025).

131. Hoffman, G. E. & Schadt, E. E. variancePartition: Interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* vol. 17 Preprint at https://doi.org/10.1186/s12859-016-1323-z (2016).

132. Hoffman, G. E. & Roussos, P. dream: Powerful differential expression analysis for repeated measures designs. *Bioinformatics* Preprint at https://doi.org/10.1093/bioinformatics/btaa687 (2020).

133. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* vol. 67 1–48 Preprint at https://doi.org/10.18637/jss.v067.i01 (2015).

134. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* vol. 82 1–26 Preprint at https://doi.org/10.18637/jss.v082.i13 (2017).

135. Ho, D. E., Imai, K., King, G. & Stuart, E. A. MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software* vol. 42 1–28 Preprint at https://doi.org/10.18637/jss.v042.i08 (2011).

136. Greifer, N. cobalt: Covariate Balance Tables and Plots. Preprint at https://doi.org/10.32614/CRAN.package.cobalt (2026).

137. Savvina, M. S., Burtseva, T. E. & Munkhalov, A. A. Gut microbiome and breastfeeding of children. *Yakut Medical Journal* 107–109 (2021).

138. Anderson, M. J. A new method for non-parametric multivariate analysis of variance: NON-PARAMETRIC MANOVA FOR ECOLOGY. *Austral Ecol.* **26**, 32–46 (2001).

139. Demirci, M. *et al.* Reduced Akkermansia muciniphila and Faecalibacterium prausnitzii levels in the gut microbiota of children with allergic asthma. *Allergol. Immunopathol. (Madr.)* **47**, 365–371 (2019).

140. Myers, P. N. *et al.* Early-life colonization by aromatic-lactate-producing bifidobacteria lowers the risk of allergic sensitization. *Nat. Microbiol.* **11**, 429–441 (2026).

141. Kennedy, E. A. *et al.* Skin microbiome before development of atopic dermatitis: Early colonization with commensal staphylococci at 2 months is associated with a lower risk of atopic dermatitis at 1 year. *J. Allergy Clin. Immunol.* **139**, 166–172 (2017).

142. Tsilochristou, O. *et al.* Association of Staphylococcus aureus colonization with food allergy occurs independently of eczema severity. *J. Allergy Clin. Immunol.* **144**, 494–503 (2019).

143. Gołębiewski, M. *et al.* Mother's milk microbiome shaping fecal and skin Microbiota in infants with food allergy and atopic dermatitis: A pilot analysis. *Nutrients* **13**, 3600 (2021).

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
5E2737918AB1451...
Author Signature

08/02/2026
Date