

---

# A Systems Approach to Understanding Allergy, Asthma and Childhood Wheeze

---

Howard Ho-Fung Tang

ORCID: [0000-0001-6422-0270](https://orcid.org/0000-0001-6422-0270)

*Supervised by*

A/Prof. Michael Inouye and Prof. Kathryn Holt

*A dissertation submitted to*

**The School of BioSciences  
The University of Melbourne**

*in partial fulfillment of the requirements  
for the degree of Doctor of Philosophy.*

*Research conducted at*

**The Systems Genomics Laboratory  
Baker Heart & Diabetes Institute**

February, 2019



# Abstract

In Australia, asthma is a common respiratory disease with a significant health burden. Our current understanding of the biological mechanisms behind asthma is incomplete. It is not clear what makes a person more susceptible compared to another, nor is it clear how determinants of asthma susceptibility interact to cause disease. Childhood wheeze does not necessarily progress to asthma, and asthma itself is a heterogeneous condition that encompasses many different phenotypes, each with potentially different biology. However, we suspect that, for most affected individuals, the origins of asthma arise in early childhood, as embodied by the “hygiene hypothesis”. Events like microbial and allergen exposure in early life, as well as frequency and severity of respiratory infections, may steer the child on a course towards asthma and disease. Early prediction of disease susceptibility or severity is important because it may permit early intervention in young children, which may then limit the progression of asthma or prevent it altogether.

My research thesis had three general aims:

1. To uncover hidden subgroups or “*clusters*” of children who share similar trajectories of immune function and susceptibility to respiratory infection; and determine how these relate to asthma and other related phenotypes.
2. To describe *microbial communities in the upper respiratory tract* of infants, specifically distinct patterns of change or trajectories in the microbiome that emerge as the child ages; and to determine how these relate to respiratory health, asthma, and related phenotypes.
3. To identify novel *genetic determinants* of asthma and related phenotypes in early childhood (including immunorespiratory clusters and microbiome trajectories), and determine how these relate to each other.

Through this research, I hope to shed light on the complexity that is asthma pathogenesis. In particular, it may explain how the determinants of asthma are similar or different between individuals. With my research, it may be possible to better characterise the interlocking events that lead from disruption of normal physiology to eventual disease. Future studies can focus on the origins of asthma in specific subpopulations, as well as potential treatment targets within each subgroup. The results of this research may open up the potential for developing therapeutic and preventative measures for asthma, as well as allow earlier intervention for infants at risk of developing asthma later in life.



# Declaration of Authorship

I, Howard Ho-Fung Tang, declare that this thesis titled, “A Systems Approach to Understanding Allergy, Asthma and Childhood Wheeze” and the work presented in it are my own. I confirm that:

- This thesis is my own original work towards the degree of Doctor of Philosophy, at the University of Melbourne. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself in the preface.
- This thesis has not been previously submitted for any other degree or qualification at this University or any other institution.
- Due acknowledgement has been made in the text to all other material used.
- Exclusive of tables, bibliographies and appendices, the word count of this thesis is fewer than the maximum word limit of 100,000.

Signed:

---

Date: 15 February 2019

---



# Preface

## General orientation

This thesis is organised into six chapters:

1. Chapter 1 is a general introduction to the thesis.
2. Chapter 2 is a literature review.
3. Chapters 3, 4 and 5 are research chapters.
4. Chapter 6 is a final concluding chapter.

## Contributions to Chapters 1 and 6

Howard HF Tang (the candidate) was the sole contributor to the introductory and concluding chapters (Chapters 1 and 6).

## Contributions to Chapter 2

Chapter 2 of the thesis, titled **Systems biology and “big data” in asthma and allergy — recent discoveries and emerging challenges**, is a literature review of current systems-based research in asthma and allergy research. The chapter was written by the candidate under the guidance of supervisors Michael Inouye and Kathryn Holt. Further expert advice and commentary was provided by Patrick Holt and Peter Sly to guide the candidate in writing. There are plans to submit this chapter for review to the European Respiratory Journal in early 2019.

## Contributions to Chapter 3

Chapter 3, titled **Trajectories of childhood immune development and respiratory health relevant to asthma and allergy**, was written in collaboration with investigators of the Childhood Asthma Study (CAS), the Childhood Origins of Asthma Study (COAST) and the Manchester Allergy and Asthma Study (MAAS). Howard HF Tang (the candidate) was responsible for the majority (90%) of the biostatistical analysis and writing of the manuscript. This included: mixture-model-based cluster analysis of CAS, and derivation of the classifier model to apply to COAST and MAAS; analysis of demographics and summary statistics in CAS; and all other secondary analyses in CAS. Shu Mei Teo prepared the microbiome data in CAS, and edited the manuscript. Marta Brozynska also assisted in editing the manuscript. Danielle CM Belgrave analysed phenotypic data and performed replication analysis in MAAS. Michael Evans conducted the equivalent analysis in COAST. The primary investigators of COAST were Daniel J Jackson, James E Gern and Robert Lemanske; for MAAS they were Sebastian L Johnston, Angela Simpson and Adnan Custovic; for CAS they were Merci MH Kusel, Peter D Sly and Patrick G Holt. Each of these investigators made significant contributions in terms of: providing data for each cohort; providing insight into unique properties of each respective cohort; and assisting in interpretation of results. Kathryn E Holt and Michael Inouye provided supervision

to the candidate, and were the primary instigators of the CAS/COAST/MAAS cross-collaboration project, which integrated genomic, microbiomic and other phenotypic data. This chapter has been published in its entirety in eLife on 15 October 2018, available [here](#).

## Contributions to Chapter 4

The research for Chapter 4, titled **Diverging trajectories of nasopharyngeal microbiome during early childhood are associated with asthma and asthma-related traits**, was conducted in collaboration with CAS and COAST investigators. Howard Tang was responsible for the bulk (90%) of the analysis and writing — in particular, processing and analysis of the 16S microbiome data through the QIIME2 pipeline to derive amplicon sequence variants (ASVs); generation of microbiome profile groups (MPGs); derivation of microbiome trajectories; and all other secondary analyses, including the meta-analyses. Shu Mei Teo provided some of the initial methodology and protocol to prepare and analyse the microbiome data in CAS (specifically QIIME1), ran the initial FastSpar analyses, and advised on the use of generalised linear models and estimating equations. Stephen Watts was the author of the FastSpar software package which was used in the analysis of the microbiome data. Louise M Judd was responsible for preparing the microbiome data (16S V4 rRNA sequencing) for both CAS and COAST nasopharyngeal samples. Sebastian L Johnston, Yury A Bochkov, Kristine Grindle and James E Gern provided viral typing for both CAS and COAST samples. Michael D Evans and Anna (Ania) Lang provided phenotypic data from COAST. They also performed other COAST-based analyses *external* to this thesis. Merci MH Kusel, Danny Mok, Barbara J Holt, Michael Serralha and Niamh Troy provided phenotypic data from CAS. Kathryn E Holt and Michael Inouye provided supervision to the candidate. Other contributions were as given in the Chapter 3 description.

Preliminary results from the COAST-specific analyses were presented by Anna Lang at the American Academy of Allergy Asthma and Immunology (AAAAI)/ World Allergy Organization (WAO) Joint Congress, on 4 March 2018 at Orlando, Florida, United States of America. The COAST-specific elements of this chapter will be compiled along with ongoing research from COAST collaborators, and will be submitted in the near future as a research paper with equal first-authorship between Howard Tang and Anna Lang. The CAS-specific contents and meta-analysis results of this chapter will be submitted as a separate paper following the COAST-centric submission.

## Contributions to Chapter 5

The final research chapter of the thesis is titled **The link between genetics of asthma and allergic disease, and events in early childhood**. The research for this chapter was conducted in collaboration with CAS investigators. The candidate, Howard Tang, was responsible for the majority (90%) of the analysis and writing, including: processing and imputation of CAS genotype data using SHAPEIT/IMPUTE2 and 1000 Genomes reference; genome-wide association scans (GWAS) with FaST-LMM, longitudinal GWAS with RepeatABEL, and catalogue loci analyses; analyses with genome risk scores (GRS); and all other secondary analyses, including modelling associations between GRS and phenotypic traits. Shu Mei Teo formulated the GRS from the summary statistics data. Qinqin Huang performed imputation using the Michigan Imputation Server; the results of this imputation were then used for subsequent GRS calculations and analyses. Louise M Judd was responsible for preparing the microbiome data (16S V4 rRNA sequencing) for CAS nasopharyngeal samples. Oneil Bhalala and Lesley Raven provided preliminary



scripts and advice for conducting imputation with SHAPEIT/IMPUTE2 and genome-wide association analyses with FaST-LMM. Merci MH Kusel, Danny Mok, Barbara J Holt, Michael Serralha and Niamh Troy provided phenotypic data from CAS. Kathryn E Holt and Michael Inouye provided supervision to the candidate. Other contributions were as given in the Chapters 3 and 4 descriptions.

Preliminary results from genome-wide analyses for CAS-specific traits of respiratory health were presented by Howard Tang at the Lorne Genome Conference on February 2016 at Lorne, Victoria, Australia. At time of writing, there are plans to publish the GRS analyses of this chapter as part of a manuscript authored by Howard Tang. The GWAS analyses, if replicated in secondary independent cohorts (COAST, MAAS or others), may also be published.

## **Other details**

Prior to enrolment in the degree, the candidate performed some preliminary data checking, processing and quality control of CAS data, while working in the capacity of research assistant. Otherwise, none of the work presented in this thesis was carried out before enrolment.

None of the work in this thesis was submitted for other qualifications.

The candidate acknowledges that the research for this thesis was almost entirely funded by the NHMRC 2016 Clinical Postgraduate Research Scholarship, under the project title “Understanding the pathogenesis, phenotypic variation and risk prediction of childhood asthma using computational approaches” (NHMRC ID: 1114753).



# Acknowledgements

I, Howard Tang, will like to acknowledge the following individuals and entities for their support during the course of my PhD:

- Michael Inouye and Kathryn Holt for their supervision throughout the entire PhD. In particular, I thank Michael for his insightful advice on the general conduct of research, his keen eye towards seeing the forest from the trees, his approachability, and his limitless tolerance for my shortcomings. I am also grateful to Kathryn for her insight into all things related to microbiome, and for her advocacy in relation to getting my research recognised.
- My committee for their support. In particular, I thank Shu Mei Teo for her invaluable help with all things related to biostatistics in general. I am also grateful to Patrick Holt, Gary Anderson and Peter Sly for their scientific input and career advice. Finally, I thank Alex Andrianopoulos for chairing the committee and for keeping things running smoothly with the Faculty.
- The researchers and participants of the Childhood Asthma Study, without whom my research would have been impossible. In particular, I am grateful to Merci Kusel and Barbara Holt for curating the dataset, and to Louise Judd for her sequencing work.
- My collaborators at Wisconsin and Manchester. I especially thank Danielle Belgrave and Michael Evans for willingly consuming some of their own time to run additional analyses on my behalf. I also thank Jim Gern, Carole Ober, and Adnan Custovic for their cooperation in getting the cross-collaborative effort between CAS, COAST and MAAS running smoothly.
- The NHMRC for funding this project.
- The University of Melbourne and the Baker Institute for providing the academic resources that I needed to accomplish this doctorate degree.
- Fellow students, past and present, for their camaraderie and support — for being friendly and approachable; for filling me in on knowledge gaps related to things outside my immediate expertise; for being willing guinea pigs in my dungeon-keeping escapades; for stomaching my karaoke antics; and for generally not being annoyed at my idiosyncracies. I thank, in no particular order: Amy Hamilton, Jane Hawkey, Claire Gorrie, Jason Grealey, Qinqin Huang, Yang (Claire) Liu, Artika Nath, Owen Qin, Scott Ritchie, Yu Wan, and Stephen Watts.
- Post-docs and other research personnel, past and present, especially Gad Abraham, Oneil Bhalala, Marta Brożyńska, Sean Byars, Rodrigo Cánovas, Zoe Dyson, Liam Fearnley, Guillaume Méric, Lesley Raven, Alex Smith, and Tingting Wang. Thank you in particular to Oneil and Lesley for their help when I first started with genomics analysis.
- Friends. In no particular order: Jack Huang, Tan Kitipornchai, Eeshan Pang, Daniel Pham, Stephanie Wong, and Brian Yue. They have been instrumental in keeping me (somewhat) sane throughout this entire process.
- Family. Thank you to my parents for their understanding and their enduring support.



# Contents

<b>Abstract</b>	<b>iii</b>
<b>Declaration of Authorship</b>	<b>v</b>
<b>Preface</b>	<b>vii</b>
<b>Acknowledgements</b>	<b>xi</b>
<b>List of Figures</b>	<b>xvii</b>
<b>List of Tables</b>	<b>xxi</b>
<b>List of Abbreviations</b>	<b>xxiii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Overview of asthma and allergy . . . . .	1
1.2 The variable contribution of genetics and environment to asthma creates heterogeneity . . . . .	1
1.3 The arrival of big data and systems biology . . . . .	2
1.4 Current knowledge of asthma and allergy . . . . .	2
1.4.1 Theories of asthma and allergy . . . . .	2
1.4.2 The atopic march . . . . .	3
1.4.3 Exposures and the hygiene hypothesis . . . . .	3
1.4.4 Studies incorporating omics-level data . . . . .	3
1.4.5 Exploring asthma heterogeneity with clustering and classification . . . . .	4
1.5 Major gaps in knowledge . . . . .	4
1.6 Key research questions addressed by this thesis . . . . .	6
1.7 Key research aims and thesis structure . . . . .	7
References . . . . .	9
<b>2 Systems biology and “big data” in asthma and allergy — recent discoveries and emerging challenges</b>	<b>13</b>
2.1 Introduction . . . . .	13
2.2 Overview of omic findings in allergy and asthma . . . . .	15
2.2.1 Genomics . . . . .	15
2.2.2 Transcriptomics . . . . .	17
2.2.3 Epigenomics . . . . .	18
2.2.4 The microbiome . . . . .	19
2.2.5 The exposome and environmental exposures . . . . .	20
2.2.6 Proteomics, metabolomics, and lipidomics . . . . .	22
2.2.7 The phenome and physiome . . . . .	22
2.3 Integration of omics data . . . . .	23
2.3.1 Exploring intra- and inter-omic relations . . . . .	23
2.3.2 Machine learning, dimension reduction, and clustering . . . . .	25

2.3.3	Network analysis . . . . .	27
2.3.4	Mathematical modelling and prediction . . . . .	28
2.4	Pitfalls and challenges . . . . .	29
2.5	Future directions and concluding statements . . . . .	30
	References . . . . .	30
<b>3</b>	<b>Trajectories of childhood immune development and respiratory health relevant to asthma and allergy</b>	<b>41</b>
3.1	Introduction . . . . .	41
3.2	Results . . . . .	42
3.2.1	CAS1: low-risk, non-atopic cluster with transient wheeze . . . . .	42
3.2.2	CAS2: low-risk cluster susceptible to atopic and non-atopic wheeze . . . . .	48
3.2.3	CAS3: high-risk atopic cluster with persistent wheeze . . . . .	52
3.2.4	Comparison of measures of immunological response . . . . .	52
3.2.5	Comparison of clusters to existing criteria for atopy . . . . .	54
3.2.6	Comparison of clusters to time-dependent wheeze phenotypes and atopic disease . . . . .	54
3.2.7	Relationship with the nasopharyngeal microbiome . . . . .	54
3.2.8	External replication of clusters in MAAS and COAST . . . . .	55
3.2.9	Internal stability and validity of CAS clusters . . . . .	57
3.2.10	Decision tree analysis . . . . .	57
3.3	Discussion . . . . .	58
3.3.1	Cluster 3 is a high-risk, multi-sensitised, atopic phenotype . . . . .	58
3.3.2	Role of early-life HDM hypersensitivity . . . . .	59
3.3.3	Role of early-life food and peanut sensitization . . . . .	59
3.3.4	IgG4 separates individuals susceptible to atopic wheeze from those who are not . . . . .	59
3.3.5	The role of respiratory infection and nasopharyngeal microbiome in childhood wheeze differs across different clusters . . . . .	60
3.3.6	Implications for cluster analysis in asthma research . . . . .	61
3.3.7	Concluding statements . . . . .	62
3.4	Methods . . . . .	63
3.4.1	Patients and study design in CAS . . . . .	63
3.4.2	Measurements in CAS . . . . .	63
3.4.3	Identification of latent clusters and selection of clustering features . . . . .	63
3.4.4	Cluster analysis using non-parametric mixture modelling . . . . .	65
3.4.5	Classification of test datasets using mixture model densities . . . . .	65
3.4.6	Replication cohorts . . . . .	66
3.4.7	Cluster validity and stability . . . . .	67
3.4.8	Decision tree analysis . . . . .	67
3.4.9	Statistical analyses . . . . .	68
	References . . . . .	68
<b>4</b>	<b>Diverging trajectories of nasopharyngeal microbiome during early childhood are associated with asthma and asthma-related traits</b>	<b>73</b>
4.1	Introduction . . . . .	73
4.2	Methods . . . . .	74
4.2.1	Study design and overview of measurements . . . . .	74
4.2.2	Bacterial 16S profiling of nasopharyngeal samples, and annotation of OTUs and ASVs . . . . .	76

4.2.3	Comparison of CAS microbiome data generated with old versus new pipelines . . . . .	77
4.2.4	Generation of microbiome profile groups (MPGs) . . . . .	77
4.2.5	Virus detection . . . . .	78
4.2.6	Correlation among ASVs . . . . .	78
4.2.7	Diversity measures and ecological analyses . . . . .	78
4.2.8	Dimension reduction and clustering into microbiome trajectories . . . . .	78
4.2.9	Association analyses and meta-analyses . . . . .	79
4.3	Results . . . . .	80
4.3.1	Composition of the nasopharyngeal microbiome in CAS and COAST children . . . . .	80
4.3.2	Microbiome profile groups (MPGs) in CAS and COAST . . . . .	80
4.3.3	Correlation patterns between ASVs in CAS and COAST . . . . .	81
4.3.4	Associations of nasopharyngeal microbiome with respiratory illness in CAS and COAST . . . . .	83
4.3.5	Associations of nasopharyngeal microbiome with seasonal changes in CAS and COAST . . . . .	83
4.3.6	Viral detection patterns in CAS and COAST . . . . .	83
4.3.7	Combined association analysis for respiratory illness with multiple predictors and covariates . . . . .	85
4.3.8	Trends in MPGs and ASVs before and after respiratory infections . . . . .	85
4.3.9	Trajectory analysis of the nasopharyngeal microbiome in CAS and COAST . . . . .	86
4.3.10	Associations with later wheeze, asthma and related disease traits . . . . .	87
4.3.11	Associations with microbiome alpha diversity . . . . .	88
4.4	Discussion . . . . .	90
4.4.1	General trends in nasopharyngeal microbiome in early childhood – similarities across different populations . . . . .	90
4.4.2	Advantages and limitations of analyses using ASVs derived from 16S V4 region . . . . .	91
4.4.3	Contributions of bacteria to acute respiratory illness . . . . .	91
4.4.4	Contributions of season and viruses to acute respiratory illness . . . . .	93
4.4.5	Contributions of nasopharyngeal microbiota to later asthma outcomes . . . . .	94
4.4.6	Concluding statements . . . . .	95
	References . . . . .	96
<b>5</b>	<b>The link between genetics of asthma and allergic disease, and events in early childhood</b> . . . . .	<b>107</b>
5.1	Introduction . . . . .	107
5.2	Methods . . . . .	108
5.2.1	Samples, genotyping and imputation . . . . .	108
5.2.2	Single-timepoint association analyses . . . . .	109
5.2.3	Longitudinal “repeated-measures” association analyses . . . . .	110
5.2.4	Genomic risk scores (GRS) . . . . .	111
5.3	Results . . . . .	112
5.3.1	GWAS for early-life childhood traits relevant to asthma . . . . .	112
5.3.2	Association analyses of GWAS catalogue SNPs for childhood asthma-related traits . . . . .	116
5.3.3	Repeated-measures GWAS for early-life childhood traits, and meta-analysis for microbial traits . . . . .	116

5.3.4	Genomic risk scores (GRS) for asthma-related traits, and their relationships to early-life childhood events . . . . .	117
5.4	Discussion . . . . .	118
5.4.1	Some genetic associations with early-life traits were shared with known associations for asthma . . . . .	123
5.4.2	Genomic risk scores for later asthma-related traits were linked to early-life events, suggesting a route of pathology for genetic susceptibility to asthma . . . . .	124
5.5	Conclusions . . . . .	125
	References . . . . .	125
<b>6</b>	<b>Conclusions</b>	<b>131</b>
6.1	Summary of findings . . . . .	131
6.1.1	Chapter 3: Mixture-model clusters of asthma susceptibility . . . . .	131
6.1.2	Chapter 4: Nasopharyngeal microbiome and asthma . . . . .	131
6.1.3	Chapter 5: Genetics of asthma and early childhood events . . . . .	132
6.2	Overall contribution to knowledge . . . . .	133
6.2.1	Revisiting the key questions . . . . .	133
6.2.2	Insights from this thesis . . . . .	133
6.3	The future . . . . .	134
<b>A</b>	<b>ePrint of Chapter 3</b>	<b>137</b>
<b>B</b>	<b>Supplementary Figures and Tables for Chapter 3</b>	<b>169</b>
<b>C</b>	<b>Supplementary Figures and Tables for Chapter 4</b>	<b>205</b>
<b>D</b>	<b>Supplementary Figures and Tables for Chapter 5</b>	<b>235</b>



# List of Figures

1.1	Outline of research chapters in the thesis . . . . .	8
2.1	“Omics” in allergy, and their interrelationships . . . . .	14
2.2	Overview of systems-based approaches to tackling research questions in allergy and asthma . . . . .	16
2.3	Data-driven versus hypothesis-driven machine learning for integration of omic data . . . . .	26
3.1	Non-parametric mixture-model based clustering of CAS dataset, based on 174 features. . . . .	43
3.2	Graphical summary of proposed clusters. . . . .	44
3.3	Incidence of multiple phenotypes, including parent-reported wheeze (A), physician-diagnosed asthma (B), defined wheeze phenotypes (C), in relation to food and inhalant sensitisation (D), stratified by cluster and time in the CAS dataset. . . . .	44
3.4	HDM IgE (A), IgG (B) and IgG4 (C); and peanut IgE (D) and IgG4 (E) stratified by cluster and time, in the CAS dataset . . . . .	46
3.5	LRI frequency (A), wheezy LRI (wLRI) frequency (B), and HDM IgE (C), stratified by age-five wheeze status, cluster and time, in the CAS dataset. . . . .	51
3.6	PBMC expression of IL-5 (A) and IL-4 mRNA (B), as well as IL-13 protein (C), in response to stimulation HDM, stratified by cluster and time (CAS). . . . .	53
3.7	Description of npEM-derived clusters in external cohorts: in MAAS, incidence of wheeze (A), asthma diagnosis (B), and HDM IgE levels (C); in COAST, incidence of asthma diagnosis (D), proportion of individuals with detectable aeroallergen-specific IgE levels (E), and PBMC protein expression of IL-13 following HDM stimulation above unstimulated control (F). . . . .	56
4.1	Microbiome profile groups (MPGs) and the relative abundance of featured OTUs or ASVs, in (A) CAS and (B) COAST. . . . .	82
4.2	Distribution of MPGs (proportions of samples) in healthy and illness samples, within (A) CAS and (B) COAST . . . . .	100
4.3	SparCC correlations amongst ASVs in (A) CAS and (B) COAST. . . . .	101
4.4	Meta-analysis and forest plots of GEE associations between MPGs and respiratory health vs. illness status at time of sample collection. . . . .	102
4.5	Distribution of MPGs (proportions of samples) in healthy and illness samples, arranged by season and month of the year, within (A) CAS and (B) COAST (QIIME2 pipeline). . . . .	103
4.6	Average relative abundances of ASVs per healthy samples, per individual, in each microbiome trajectory as determined by MFA/ <i>k</i> -means analysis of microbiome from healthy samples; in (A) CAS and (B) COAST. . . . .	104
4.7	Alpha diversity of samples (measured by Shannon Diversity index) by age and illness status of samples in CAS and COAST . . . . .	105

5.1	Manhattan plots of genome-wide association scans for parent-reported wheeze at age 1 in CAS. . . . .	114
5.2	Manhattan plots of genome-wide association scans for parent-reported wheeze at age 4 in CAS. . . . .	115
5.3	Correlation patterns between the various GRS calculated for CAS. . . . .	119
5.4	Associations between GRS and early-life traits in CAS, as determined by GLMs. . . . .	120
5.5	Graphs showing relationships between GRS for allergic disease and (A) the high-risk npEM cluster from Tang et al; and (B) presence of illness-associated MPGs in healthy samples up to age two. . . . .	121
5.6	Prevalence of early-life traits versus timepoint, separated by GRS tertiles. . . . .	122
B.1	Scatterplot of principal components analysis (PCA) of the complete-case CAS dataset ( $N = 186$ , with points coloured by npEM clusters . . . . .	169
B.2	Silhouette widths of clusters generated by npEM. . . . .	170
B.3	Overview of study methodology. . . . .	171
B.4	Relationship of clusters to food sensitisation, eczema and wheeze. . . . .	172
B.5	Correlation patterns between IgE vs. IgG4 (A) and IgE vs. IgG (B) at age five. . . . .	173
B.6	Distinct biological signals of HDM IgE, IgG4, SPT, and Th2 cytokine (IL-13). . . . .	174
B.7	A “simple” decision tree generated by recursive partitioning from CAS data, with breakdown of tree clusters by actual CAS npEM-derived clusters (A); scatterplot showing separation of CAS clusters by decision split thresholds (B). . . . .	175
B.8	Decision tree generated by recursive partitioning from CAS data, excluding Phadiatop assay variables. . . . .	176
B.9	A “comprehensive” decision tree generated by recursive partitioning from CAS data, given CAS npEM-derived clusters and age-five wheezing status. . . . .	177
B.10	Comparison of predictors for age-five wheeze in CAS and COAST clusters. . . . .	178
C.1	Bioinformatic pipeline for processing and analyzing CAS and COAST 16S rRNA data, using QIIME2 and the “microbiome” R package . . . . .	206
C.2	Distribution of ASVs (average relative abundance) in healthy and illness samples, within (A) CAS and (B) COAST. . . . .	207
C.3	(next two pages) Heatmaps of MPGs with relative abundances of all common ASVs — complete versions of Figure 4.1. Samples clustered into MPGs in (A) CAS and (B) COAST. . . . .	208
C.4	(next two panels) Relative abundance of each OTU or ASV within all samples of each MPG in QIIME1 CAS (A), and QIIME2 CAS and COAST (left and right; B). . . . .	211
C.5	Meta-analysis and forest plots of GEE associations between ASVs and respiratory health vs. illness status at time of sample collection. . . . .	213
C.6	Distribution of ASVs (relative abundance) in healthy and illness samples, arranged by season and month of the year, within (A) CAS and (B) COAST. . . . .	214
C.7	Distribution of viruses in healthy and illness samples, arranged by season and month of the year; count of samples within (A) CAS (age 2-3y) and (B) COAST (age up to 3y); proportion of samples within (C) CAS (age 2-3y) and (D) COAST (age up to 3y) . . . . .	215
C.8	Correlation of viruses detected within nasopharyngeal samples during the first 3 years of life, in (A) CAS and (B) COAST. . . . .	216
C.9	Relative abundance of ASVs and proportion of viruses before, during, and after an acute respiratory infection (ARI), in (A) CAS and (B) COAST. . . . .	217

C.10 (next two pages) Trajectories in the nasopharyngeal microbiome as determined by Multiple Factor Analysis (MFA) and K-means clustering, in (A) CAS and (B) COAST. . . . .	218
C.11 Complete version of Figure 4.6: Average relative abundances of ASVs per healthy and illness-associated samples, per individual in each microbiome trajectory as determined by MFA/ <i>k</i> -means; in (A) CAS and (B) COAST (QIIME2). . . . .	221
D.1 Manhattan plots of genome-wide association scans for parent-reported wheeze at age 5 in CAS. . . . .	236
D.2 Manhattan plots of genome-wide association scan for any wheezy LRI at age 1 in CAS. . . . .	237
D.3 Manhattan plots of genome-wide association scan for any rhinovirus-C-associated LRI at age 1 in CAS. . . . .	238
D.4 Manhattan plots of genome-wide association scan for any rhinovirus-A-associated LRI at age 1 in CAS. . . . .	239
D.5 GWAS catalogue SNPs most frequently associated (at an unadjusted threshold) with an early-life trait in CAS, sorted by phenotypes. . . . .	240
D.6 Histogram of p-values of GWAS catalogue SNPs in association with eczema at age 6m in CAS. . . . .	241
D.7 Scatterplot of the first two principal components (PC)s from principal components analysis of GRS in CAS, coloured by npEM clusters as per Tang et al 2018. . . . .	242
D.8 Principal components analysis of GRS in CAS identified an “atopic vector” in PC1. . . . .	243
D.9 GLM associations with early-life traits vs. GRS, with membership in the high-risk npEM cluster “Cluster 3” and sex as potential covariates. . . . .	244
D.10 GLM associations with early-life traits vs. npEM cluster 3, with sex and GRS as potential covariates. . . . .	245



# List of Tables

3.1	Comparison of selected demographic and clinical variables in CAS clusters	45
3.2	Comparison of HDM-associated immunological variables in CAS . . . . .	47
3.3	Comparison of selected respiratory disease-related variables in CAS clusters	49
3.4	Analysis of selected predictors for age-five wheeze within each CAS cluster, with demographic covariates (sex, BMI, parental history of asthma) . . . . .	50
3.5	Key findings from cluster analysis. . . . .	62
4.1	Key differences between CAS and COAST birth cohorts . . . . .	75
4.2	Results of GEE models associating MPG of sample with illness status (well vs. unwell), with adjustments for child as subjects factor, and gender, age and season as covariates. . . . .	84
4.3	Results of GEE model associating respiratory illness with presence of virus, illness- or health-associated MPG, incorporating interaction effects. . . . .	86
4.4	GLM models associating wheeze and asthma outcomes with trajectories based on MFA/k-means of microbiome data, from routine healthy samples within first 2 years of life, and early sensitisation status, as represented by high-risk npEM cluster or early-life allergen sensitisation by age two; (A) in CAS, (B) in COAST. . . . .	89
5.1	Selected suggestive and significant SNPs for genome-wide association scans for early-life traits in CAS. . . . .	113
B.1	List of clustering features . . . . .	179
B.2	Terminology used to describe groupings produced by various clustering and classification methods on different datasets . . . . .	180
B.3	Comparison of selected demographic and clinical variables in CAS clusters	181
B.4	Repeated-measures ANOVA for selected predictors, in the first three years of life (timepoints at ages 6m, 1, 2, and 3) . . . . .	199
B.5	Comparison of the three clusters generated by npEM, with other clustering or classification schemes. . . . .	199
B.6	Correlation between Phadiatop vs. allergen-specific IgE and IgG4 in CAS.	200
B.7	Complete version of Table 4: Predictors for age-five wheeze within each CAS cluster, with demographic covariates (sex, BMI, parental history of asthma). . . . .	201
C.1	The top twenty common OTUs in CAS QIIME1 dataset, and their analogous ASVs in CAS QIIME2 dataset that have matching sequences . . . . .	222
C.2	The eighteen ASVs common to either CAS or COAST datasets (QIIME2) .	225
C.3	Results of GEE models associating common ASVs with respiratory illness status (well vs. unwell), with adjustments for child as subjects factor, and gender, age and season as covariates. . . . .	227

C.4	Results of GEE models associating MPGs and ASVs with winter season, with adjustments for child as subjects factor, and gender, age, season +/- respiratory illness as covariates. . . . .	228
C.5	Viruses detected in nasopharyngeal samples in the first 3 years of life, collected from CAS and COAST. . . . .	231
C.6	Results of GEE models associating viruses with respiratory illness status (well vs. unwell), with adjustments for child as subjects factor, and gender, age and season as covariates. . . . .	232
C.7	Results of GEE models associating viruses with winter season, with adjustments for child as subjects factor, and gender, age and respiratory illness as covariates. . . . .	232
C.8	Results of GEE models associating MPGs with the presence of any virus in the same sample, with or without adjustment for season and respiratory illness as covariates. . . . .	233
C.9	GLM models associating wheeze and asthma outcomes with proportion of illness-associated MPGs in routine healthy samples within first 2 years of life, stratified by early sensitization status and npEM clusters; (A) in CAS, (B) in COAST. . . . .	234
D.1	GWAS catalogue SNPs most frequently associated with an early-life trait in CAS, at the unadjusted p-value threshold . . . . .	246
D.2	Candidate SNPs associated with wheeze or asthma at age five in CAS, at the unadjusted p-value threshold . . . . .	248
D.3	Selected significant and suggestive SNPs for longitudinal genome-wide association scans for early-life traits in CAS, with repeatABEL. . . . .	249
D.4	Selected significant and suggestive SNPs for longitudinal genome-wide association scans for early-life microbiome-related traits in CAS and COAST, with repeatABEL. . . . .	250

# List of Abbreviations

ANOVA	ANalysis Of VAriance
ARI	Acute Respiratory Infection / Illness
ASM	Airway Smooth Muscle
ASV	Amplicon Sequence Variant
BAL	Bronchoalveolar Lavage
BH	Benjamini-Hochberg method
BIC	Bayesian Information Criterion
bp	Base Pairs (unit)
BY	Benjamini-Yekutieli method
CART	Classification and Regression Trees
CAS	Childhood Asthma Study
CBMC	(Umbilical) Cord Blood Mononuclear Cell
ChIP	Chromatin Immunoprecipitation
Chr	Chromosome
COAST	Childhood Origins of Asthma Study
COPD	Chronic Obstructive Pulmonary Disease
COPSAC	Copenhagen Prospective Study on Asthma in Childhood Study
DADA2	Divisive Amplicon Denoising Algorithm v2
DIGE	Difference Gel Electrophoresis
DNA	Deoxyribonucleic Acid
EBC	Exhaled Breath Condensate
ELISA	Enzyme-Linked Immunosorbent Assay
EM	Expectation-Maximisation
eQTL	Expression Quantitative Trait Locus
EWAS	Epigenome-Wide Association Study
FACS	Fluorescence-Assisted Cell Sorting
FaST-LMM	Factored Spectrally Transformed Linear Mixed Model
fLRI	Febrile Lower Respiratory Infection / Illness
GEE	Generalised Estimating Equation
GLM	Generalised Linear Model
GPCR	G-Protein-Coupled Receptor
GWAS	Genome-Wide Association Study
HDAC	Histone Deacetylase
HDM	House Dust Mite
HAT	Histone Acetyltransferase
HLA	Human Leukocyte Antigen
IgE	Immunoglobulin E
IgG	Immunoglobulin G
LC	Liquid Chromatography
LCA	Latent Class Analysis
LD	Linkage Disequilibrium
LMM	Linear Mixed Model
LOD	Limit Of Detection

LOO	Leave-One-Out
LPS	Lipopolysaccharide
LRI	Lower Respiratory Infection / Illness
MAAS	Manchester Asthma and Allergy Study
MeDALL	Mechanisms of the Development of Allergy Consortium
meQTL	Methylation Quantitative Trait Locus
MHC	Major Histocompatibility Complex
miRNA	Micro Ribonucleic Acid
MPG	Microbiome Profile Group
MS	Mass Spectrometry
NMR	Nuclear Magnetic Resonance
NPA	Nasopharyngeal Aspirate
NPS	(Nasopharyngeal Samples with) No Prior Sickness
NPV	Negative Predictive Value
npEM	(Model produced by) Non-Parametric Expectation-Maximisation-Like Algorithm
OTU	Operational Taxonomic Unit
PBMC	Peripheral Blood Mononuclear Cell
PCA	Principal Components Analysis
PCR	Polymerase Chain Reaction
PPI	Protein-Protein Interaction
PPV	Positive Predictive Value
PUFA	Polyunsaturated Fatty Acid
QIIME	Quantitative Insights Into Microbial Ecology
RFLP	Restriction Fragment Length Polymorphism
RNA	Ribonucleic Acid
RNA-seq	RNA sequencing
rRNA	Ribosomal Ribonucleic Acid
RSV	Respiratory Syncytial Virus
rtPCR	Reverse-Transcription Polymerase Chain Reaction
RV	(Human) Rhinovirus
SARP	Severe Asthma Research Program Study
SNP / SNV	Single-Nucleotide Polymorphism / Variant
SPT	Skin Prick / Sensitisation Test
STELAR	Study Team for Early Life Asthma Research Consortium
TDA	Topological Data Analysis
Th1	Type 1 Helper T cell
Th2	Type 2 Helper T cell
Th17	Type 17 Helper T cell
TLR	Toll-Like Receptor
Treg	Regulatory T cell
U-BIOPRED	Unbiased BIOMarkers in PREDiction of respiratory disease outcomes Study
URI	Upper Respiratory Infection / Illness
WES	Whole Exome Sequencing
WGCNA	Whole Genome Co-expression Network Analysis
WGS	Whole Genome Sequencing
wLRI	Wheezy Lower Respiratory Infection / Illness



## Chapter 1

# Introduction

### 1.1 Overview of asthma and allergy

Asthma is a common and chronic medical condition, characterised by recurrent episodes of respiratory wheeze, cough, and shortness of breath. These symptoms are caused by lower airway obstruction due to airway inflammation, bronchial hyperreactivity and mucus secretion [1, 2]. A feature that distinguishes asthma from other obstructive respiratory diseases, such as chronic obstructive pulmonary disease (COPD), is the fact that this obstruction is often partly-reversible, either spontaneously or with medication. Many cases of asthma also feature allergy.

Allergy, or type I hypersensitivity, is an exaggerated and pathological immune response where the body generates antibodies against what is usually a benign antigen. These antigens or “allergens” are typically exogenous, being ingested or inhaled from the environment. Allergy is a primary driver in many cases of asthma [1, 3, 4]; other conditions with an allergic basis include hayfever (allergic rhinosinusitis, rhinoconjunctivitis); atopic dermatitis or eczema; food allergy; eosinophilic oesophagitis; and anaphylaxis. Allergy is heritable, and the term “atopy” is commonly used to describe the familial predisposition to allergic responses. Atopic individuals often respond to multiple allergens [5, 6], and multiple allergic conditions often co-occur in these individuals.

Over recent years, the incidence and health impact of both asthma and allergy have steadily increased. Once a disease found predominantly in affluent populations, asthma has become more prevalent in developing nations, possibly due to urbanisation and other changes in living environment [7]. In Australia, asthma is the leading cause of disease burden in children under the age of fourteen [8]. Medical interventions may relieve symptoms or slow functional decline, but there is currently no cure for asthma, and treatment response varies from person to person. In severe cases, uncontrolled asthma can cause status asthmaticus, respiratory arrest and death [9, 10]. In spite of its health impact, there remain many unknowns about the condition – what causes it, how these causes interact with each other, how to best treat or prevent it, and why some people respond better to treatment than others.

### 1.2 The variable contribution of genetics and environment to asthma creates heterogeneity

To address these questions, one must explore the origins of disease. Most human diseases are caused by a combination of genetic and environmental factors, and asthma and allergy are no exceptions. There is a strong familial and heritable component to asthma [11], but acute exacerbations of asthma are often triggered by environmental exposure to an allergen or some other noxious stimulus [12, 13]. There is also evidence suggesting that the development of asthma is influenced by exposures during the perinatal period and early

infancy [14, 15]. Together, genetics and environment interact and contribute to biological dysfunction, which manifests as symptomatic disease.

Asthma can also exist without evident allergy, and there are alternative non-allergic mechanisms that contribute to recurrent respiratory wheeze: for example, airway inflammation due to microbial infection or inhaled irritants [13, 16]. Like many other diseases, the susceptibility and manifestation of asthma varies from individual to individual, due to differences in underlying genetic architecture, environmental exposure, and physiology. What these differences are remain unclear. Hence, a major research priority is discovering how and why there is such heterogeneity in disease — a better understanding of this can potentially allow scientists and clinicians to develop precise and personalised options for management of disease. To perform this in a quantified manner, comprehensive and well-powered datasets are needed — beyond the standard clinical and pathological investigations.

### 1.3 The arrival of big data and systems biology

Within the last two decades, the genomics revolution has swept in technological improvements that have allowed scientists to extract exhaustive amounts of “big data” from biological samples, such as genomic (DNA), transcriptomic (RNA) and proteomic data. It was hoped that, by having a sufficiently-comprehensive collection of biological information, one could definitively describe the functional and dysfunctional state of any biological system, up to the scale of the entire human body [17]. However, the complexity of biomolecular and cellular systems means that it is no trivial task to fully interpret these large, “omic-sized” datasets.

That said, new and useful biological information could still be derived from big data. Our traditional understanding of asthma and allergy was developed largely via experimental, clinical, and epidemiological work with human and animal populations. This understanding has recently been augmented by new methods that incorporate biostatistics, bioinformatics and systems biology. Modern systems-based research has focused on the genomics of disease; the transcriptomics of affected cells and tissue compartments; their interaction with environmental exposures; and attempts at disentangling the heterogeneity of diseases using clustering and classification methods.

### 1.4 Current knowledge of asthma and allergy

#### 1.4.1 Theories of asthma and allergy

Historically, pre-big data, there were a number of theories and paradigms on asthma pathogenesis. These ranged from a vague understanding of airway hyperreactivity in Ancient Greece and China; to theories of possible neurological involvement during the Renaissance; and finally to a modern understanding of asthma as having an inflammatory and allergic basis, driven partly by genetics and partly by environment [18]. Allergy was first used by von Pirquet to describe a change in reactivity of the immune system exposed to foreign antigen, with over-reactivity causing disease — a controversial concept at a time when immunity was thought to be exclusively protective [19]. The term now refers specifically to a type of hypersensitivity: increased activity of type 2 helper T (Th2) cells, which drive plasma cells to produce allergen-specific immunoglobulin E (IgE) [4]. However, what exactly triggers this increase in Th2 activity remains unclear; it may be related to priming of immunological responses during early childhood or even the perinatal period [20, 21]. Also, many but not all manifestations of asthma have an allergic

basis, and the relative importance of allergy to pathogenesis may vary from person to person. The definition of allergy or atopy is also troublesome: current clinical guidelines use a threshold of specific IgE (0.35 kU/L) as an indicator of positive sensitisation, and yet this threshold may not be ideal especially in younger age groups [22]. Current tests do not use variable thresholds that take into account age or type of allergen being tested. There is also a subtle distinction between mediators of chronic disease and acute exacerbations — chronic airway inflammation may be driven by underlying allergic inflammation, while acute insults such as respiratory infection and inhaled or ingested noxious agents may trigger acute exacerbations [13, 18].

### 1.4.2 The atopic march

Given the observation that various allergic diseases often co-occurred in the one patient, there were attempts to link the pathophysiology of these diseases together. The natural history of atopic individuals sometimes abides by the following timeline: food sensitisation and allergic eczema preceding wheeze and asthma, then asthma itself preceding allergic rhinitis which tends to occur in adolescence and adulthood. Therefore, the “atopic march” theory proposes that there must be some causal link amongst these three conditions occurring in sequence — perhaps disruption to epithelial barriers (skin, gut) promotes further allergen sensitisation, and accentuation of a systemic Th2 immune response, which then elicit allergic inflammation in other tissue compartments [23]. However, recent research has cast doubt on this theory as the disease progression described by the atopic march is not as common as initially believed [24].

### 1.4.3 Exposures and the hygiene hypothesis

The atopic march alludes to the importance of early childhood events in determining later disease. There is emerging evidence suggesting that other events in early childhood can influence the development of asthma many years later. For example, recurrent wheezy chest infections in the first year of life increases the risk of subsequent asthma [25–27]. Exposures (or lack thereof) to certain food products, and exposure to environmental endotoxin, may also have an impact on allergy [28, 29]. Related to endotoxin is the link to microbial exposure — the hygiene hypothesis posits that improved sanitation in developed nations reduced overall microbial exposure, thus resulting in inappropriate priming of a child’s developing immune system, and leading to increased prevalence of allergy in affluent populations [30]. However, latest evidence suggests that the hygiene hypothesis is an oversimplification of the narrative [31, 32], and it remains unclear how all these changes in early-life exposures interact with each other in driving the pathology of asthma.

### 1.4.4 Studies incorporating omics-level data

In recent years, omics-based approaches have been employed to explore the pathogenesis of asthma and allergy. The genetics of asthma and allergy have been thoroughly investigated, with seminal contributions from the likes of Ober [11] and the GABRIEL Consortium [33]. Further studies have also been conducted examining the genetics of risk factors for asthma, including early-life respiratory infections with rhinovirus [34]. Many significant gene associations involve the Th2 immune pathway, and have been found to regulate expression of genes and proteins relevant to immune function and inflammation. This is consistent with existing knowledge of asthma pathogenesis. Recently, interesting novel associations have also been identified, for example loci specific to certain ethnic groups (e.g. PYHIN1 in those of African descent) [35]. Furthermore, researchers have

explored transcriptomic, proteomic and metabolomic profiles for various tissues (e.g. airway) in disease and health, in an effort to identify predictive biomarkers as well as understand pathology. Finally, some studies have integrated measurements of environmental exposures, such as diet and microbial exposure. In relation to host microbiota, numerous recent studies have associated changes in the gut and airway metagenomes with active asthma and asthma risk [36].

#### 1.4.5 Exploring asthma heterogeneity with clustering and classification

The heterogeneity of asthma manifestations suggest that the disease may represent different states that all share wheeze as a common clinical feature, yet have fundamentally different biological mechanisms. Because of this, researchers began to look within subtypes of asthma, and assess them independently of one another. Asthma has often been dichotomised into “atopic” or “extrinsic” (those with evidence of allergic pathophysiology) vs. “non-atopic” or “intrinsic” asthma (those without). Distinctions have also been made for paediatric vs. adult-onset asthma, obesity-related asthma, exercise-induced asthma, mixed asthma-COPD phenotypes and others [37]. With the advent of big data, researchers have not only begun examining each of these sub-phenotypes for omic-level differences, but have also begun interrogating the omic data itself in an “unsupervised” systemic manner. This is the use of computerised, machine-learning methods to derive sub-phenotypes based on data structure, with minimal human input or pre-supposed expert knowledge. Given the molecular-level data from which they were derived, these groupings can be presumed to relate somehow to underlying disease pathophysiology — hence they have been given the label “endotype” [38–40]. With endotypes, we hope to gain a better understanding of how asthma presents differently in different people, thus making that first step towards personalised medicine. However, due to a number of limitations (described in later chapters), it has remained a challenge to corroborate and compare these endotypes.

### 1.5 Major gaps in knowledge

Given the above summary of asthma research, the major knowledge gaps that currently exist can be summarised as follows:

- **The nature and mechanisms of asthma heterogeneity remain poorly understood.** Although the distinction between allergic and non-allergic asthma is well-known, it is probable that other categories or subcategories exist, especially given that heterogeneity also exists within these subtypes [37, 41]. Unsupervised methods of clustering have uncovered some hidden or “latent” categories, and may aid in disentangling the pathophysiology behind each category. However, although researchers have performed both supervised and unsupervised derivations of asthma subphenotypes and endotypes, it is difficult to compare between them or apply them to clinical practice. Obstacles include inconsistent definitions of supervised categories, non-unified methods and variable types of input data used to establish categories, and unaccounted variation across study populations.
- **The existing definitions for allergy and atopy are imperfect.** We refer to the diagnostic or screening criteria from pathology tests used to determine an allergic predisposition such as IgE and skin prick or sensitisation tests (SPT). They have so far been helpful in identifying certain high-risk or severely-unwell individuals in a population. However, positive results are not always specific for disease, especially in young children. IgE levels also change with age, yet there are no validated

age-adjusted thresholds or measures. The sensitisation profile in infants and young children have been described by some studies [22, 42], but there has not yet been a universally-accepted method for using these profiles to predict long-term allergic disease. Also, the patterns or trajectories of IgE levels across different allergen specificities may be more relevant than the actual level itself, but such a hypothesis has not yet been explored in depth. Finally, current measures of allergic sensitisation also do not account for possible interactions between sensitisation and other environmental exposures (e.g. viral infection).

- The contributions of **events in early childhood** to asthma pathogenesis remain poorly understood — especially the contributions of the host's **airway microbiome**, in interaction with other factors. It is well-known that certain bacteria and viruses are associated with general respiratory illness in young infants [43, 44]. However, asymptomatic colonisation with these microbes may also impact on asthma risk and progression, via mechanisms dependent or independent of the susceptibility and severity of respiratory infections. Also, it is not clear how the microbiome relates to other environmental influences, such as delivery method, seasonal changes, breast-feeding and diet. Existing findings have so far been inconsistent or contradictory [36, 43, 44]. Finally, the possibility of interaction between microbial exposure and allergen sensitisation states has been hypothesised by many, but not fully elucidated.
- Most studies that examine the contributions of host microbiome to disease have used operational taxonomic units (OTUs) to describe bacterial taxa. Newer methods that surmise bacterial populations from metagenome samples generate amplicon sequence variants (ASVs), which have their advantages over OTUs [45]. There have far been no studies that use ASVs to describe airway microbiota and their associations with asthma and childhood wheeze.
- **Genetic contributions to asthma** have been well-characterised. However there remains much uncertainty as to how genetic risk relates to early-life risk factors of infection and sensitisation, as well as how they interact with environmental factors. In particular, measures that incorporate **repeated or serial measures** of a particular phenotype (e.g. an asthma risk factor) may be better-powered to address certain questions. In addition, methods that integrate multiple genetic signals, such as genomic risk scores (**GRS**) for asthma and allergy, may provide a better predictive or explanatory model of disease than individual disease-associated loci. While GRSs have been developed for other complex polygenic diseases, few have been constructed for asthma and allergy.
- It remains incredibly challenging to **translate results** from systems-based research into clinical practice and public health interventions. This is likely due to the relative nascency of systems- and omics-based approaches, as well as the underlying biological complexity being modelled. Numerous attempts have been made to derive predictive and risk-stratifying models from biomarkers (genomic, transcriptomic, and others), but these have been complicated by disease heterogeneity, incomplete availability of relevant biomarkers, and limited consideration of gene-environment interactions. Furthermore, we have not yet reached the stage of being able to formulate targeted management options for patients with asthma and allergy. Even if we could quantify the risk of later asthma in a young child, it is not clear whether early preventative use of bronchodilators or inhaled steroids will do anything to mitigate this risk. Precision or personalised medicine will require concrete evidence that an unwell individual has a specific flavour of disrupted pathophysiology that is being

precisely targeted by a particular treatment option — and this level of information is currently lacking.

## 1.6 Key research questions addressed by this thesis

To address some of the knowledge gaps listed above, I analysed data from the Childhood Asthma Study (CAS) [46], a prospective birth cohort from Western Australia of about 200 children with comprehensive and serial measurements in a wide range of parameters. These measurements include: records of respiratory disease (infections), measures of allergy and immunopathology (antibody levels and skin sensitisation tests), descriptions of nasopharyngeal microbiota (16S ribosomal RNA sequencing), lung function tests, and host genetics (DNA microarray). Members of our laboratory and CAS have conducted studies exploring the relationship of allergic diseases with early life events, including viral infections [47], allergen sensitisation [22], antibiotic use [48] and patterns in nasopharyngeal microbiome [43, 49]. We also collaborated with investigators from the Childhood Origins of Asthma Study (COAST) [50] and the Manchester Allergy and Asthma Study (MAAS) [51], who have collected similar types of data in US and UK populations respectively. These external datasets were used primarily for validation or replication of results from CAS.

With these datasets, we attempted to address some of the existing knowledge gaps by posing the following questions:

- Is it possible to use unsupervised clustering methods to derive clusters (presumed endotypes) of childhood asthma and asthma susceptibility from clinicopathological data? What do these clusters look like, and do they capture trajectories of childhood development relevant to immune or respiratory health and disease?
- How do these immunorespiratory clusters relate to existing subgroups of asthma susceptibility, or definitions of atopy and allergy? Do these clusters provide more information than existing criteria for allergy?
- Do similar clusters exist across different populations? How do these compare?
- Does characterisation of nasopharyngeal microbiota using ASVs differ much from using OTUs? Can similar findings be achieved to those of Teo et al [43] using OTU-based results? Does the bacterial composition of nasopharyngeal microbiota contribute to respiratory disease dependently or independently of other risk factors such as season and viral detection?
- Are there clusters of individuals who share similar patterns of nasopharyngeal microbiome that evolve with time and age? Do any of these “microbial trajectories” relate to asthma risk?
- Are these associations between nasopharyngeal microbiome and respiratory disease shared across different populations?
- Are there any loci in the genome that are associated with early-life risk factors for asthma (e.g. frequency of lower respiratory infections, allergen-specific IgE levels)? if so, have any of these been replicated?
- Does incorporating the longitudinal aspect of some GWAS phenotypes (e.g. repeated measurements) grant more biologically-relevant information and hence generate any new findings with longitudinal GWAS?

- Does the genetic signal for allergy disease later in life, represented by genomic risk scores (GRS), associate with early childhood traits such as allergic sensitisation, microbial colonisation, and wheezy respiratory infections? How do immunorespiratory clusters, microbiome, and genomics interact with each other when contributing to asthma risk?

## 1.7 Key research aims and thesis structure

Given all of the above, the aims or objectives of each research chapter are formulated as follows, and as summarized in **Figure 1.1**.

**Chapter 3** seeks to address questions pertaining to clusters based on clinicopathological data. Specifically, it sets out to:

1. use non-parametric mixture models to discover latent clusters that define early childhood trajectories of immune function and susceptibility to respiratory infection in the CAS dataset;
2. investigate how these immunorespiratory clusters relate to differential profiles of asthma susceptibility, and to existing definitions of atopy, in CAS;
3. identify risk factors for asthma within each cluster; and
4. externally validate the clusters in independent cohorts COAST and MAAS, by applying the CAS-derived mixture models as classifiers to these cohorts.

**Chapter 4** addresses questions relating to the nasopharyngeal microbiome of young children and their relationship to respiratory disease. In this chapter, we:

1. apply an ASV-based bioinformatic pipeline to nasopharyngeal microbiome data from CAS and COAST, to determine and compare profiles of microbial composition between cohorts.
2. build on OTU-based results from Teo et al 2018 [43] by conducting a meta-analysis of associations between microbial and asthma-related traits using both CAS and COAST data;
3. using clustering methods that account for repeated measures, determine and compare microbiome trajectories representing the evolving healthy nasopharyngeal microbiota in CAS and COAST; and
4. describe how these trajectories relate to asthma-related traits, together with other pathophysiologically-relevant factors.

The final research chapter, **Chapter 5**, explores the contribution of genetics and genomics to asthma and allergic disease in CAS. In brief, we:

1. perform a scan for genome-wide significant SNPs associated with early-life traits in CAS, such as frequency and severity of respiratory infections, or levels of allergen-specific antibodies;
2. test whether genome-wide significant loci known to be associated with asthma-related traits (in a curated GWAS catalogue) are also linked to early-life traits in CAS;
3. perform genome-wide analyses using longitudinal association models that incorporate the serial measurement of early-life CAS traits; and

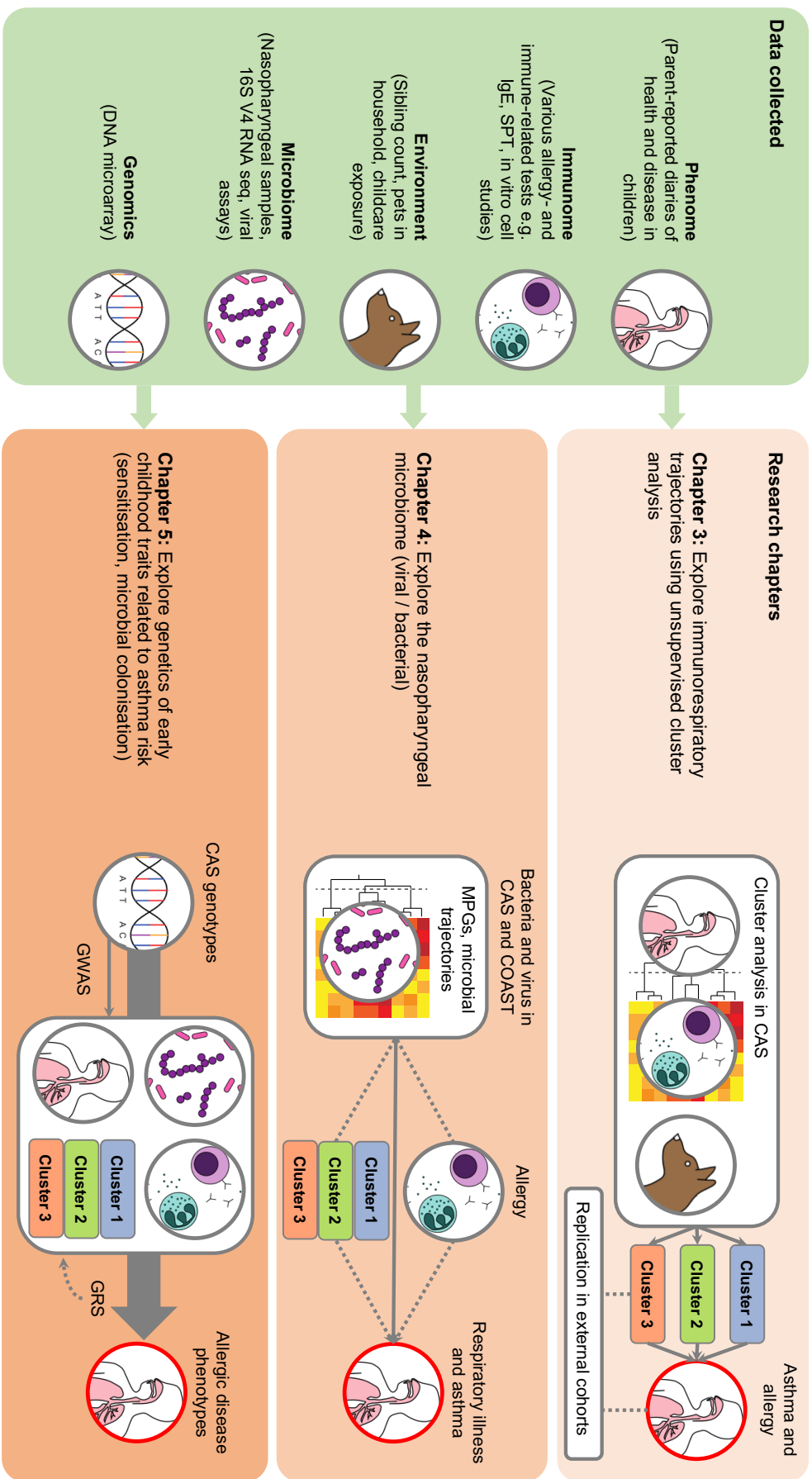


FIGURE 1.1: Outline of research chapters in the thesis

Refer to main body text for further details.



4. calculate GRS derived from larger meta-analyses for asthma and allergy-related traits, and explore the association between GRS and early-life traits in CAS.

At this point, I would like to highlight that there are two themes common to each research chapter:

1. The initial use of systems-based approaches (cluster analysis, dimension reduction) to find patterns and simplify complex data.
2. The subsequent application of “traditional” association analyses on this simplified data, to determine links between various pathophysiological entities (e.g. groups generated by cluster analysis; variables within each cluster).

Before describing my research output, we will first begin with **Chapter 2**: a literature review of systems-based approaches applied to allergy and asthma, including the examination of the major findings and limitations of such approaches.

## References

1. Global Initiative for Asthma. Global Strategy for Asthma Management and Prevention. 2015. URL: [https://ginasthma.org/wp-content/uploads/2016/01/GINA\\_Report\\_2015\\_Aug11-1.pdf](https://ginasthma.org/wp-content/uploads/2016/01/GINA_Report_2015_Aug11-1.pdf).
2. Martinez FD and Vercelli D. Asthma. *Lancet* 2013;382:1360–72.
3. Lung "H, Education BINA, and Program" P. Expert Panel Report 3: Guidelines for the Diagnosis and Management of Asthma. 2007. URL: <http://www.nhlbi.nih.gov/files/docs/guidelines/asthgdln.pdf>.
4. Galli SJ, Tsai M, and Piliponsky AM. The development of allergic inflammation. *Nature* 2008;454:445–54.
5. Coca AF and Cooke RA. On the Classification of the Phenomena of Hypersensitivity. *The Journal of Immunology* 1923;8:163–182.
6. Pawankar R, Holgate ST, and Rosenwasser LJ. *Allergy Frontiers: Classification and Pathomechanisms*. Springer Science & Business Media, 2009.
7. Global Initiative for Asthma. Global Strategy for Asthma Management and Prevention (Updated 2011). 2011. URL: <https://ginasthma.org/wp-content/uploads/2019/01/2011-GINA.pdf>.
8. Australian Institute of Health and Welfare. Asthma in Australia with a focus chapter on chronic obstructive pulmonary disease 2011. 2011. URL: <https://www.aihw.gov.au/getmedia/8d7e130c-876f-41e3-b581-6ba62399fb24/11774.pdf>.
9. Ortiz RA and Barnes KC. Genetics of allergic diseases. *Immunol Allergy Clin North Am* 2015;35:19–44.
10. Corbridge TC and Hall JB. The assessment and management of adults with status asthmaticus. *Am J Respir Crit Care Med* 1995;151:1296–316.
11. Ober C and Yao TC. The genetics of asthma and allergic disease: a 21st century perspective. *Immunol Rev* 2011;242:10–30.
12. Baxi SN and Phipatanakul W. The role of allergen exposure and avoidance in asthma. *Adolesc Med State Art Rev* 2010;21:57–71, viii–ix.
13. Holt PG and Sly PD. Viral infections and atopy in asthma pathogenesis: new rationales for asthma prevention and treatment. *Nat Med* 2012;18:726–35.

14. Von Mutius E. The microbial environment and its influence on asthma prevention in early life. *J Allergy Clin Immunol* 2016;137:680–9.
15. Lockett GA, Huoman J, and Holloway JW. Does allergy begin in utero? *Pediatr Allergy Immunol* 2015;26:394–402.
16. Holt PG and Sly PD. Non-atopic intrinsic asthma and the ‘family tree’ of chronic respiratory disease syndromes. *Clin Exp Allergy* 2009;39:807–11.
17. Thornton JM. From genome to function. *Science* 2001;292:2095–7.
18. Walter MJ and Holtzman MJ. A centennial history of research on asthma pathogenesis. *Am J Respir Cell Mol Biol* 2005;32:483–9.
19. Igea JM. The history of the idea of allergy. *Allergy* 2013;68:966–73.
20. Barrett EG. Maternal influence in the transmission of asthma susceptibility. *Pulm Pharmacol Ther* 2008;21:474–84.
21. Romagnani S. Immunologic influences on allergy and the TH1/TH2 balance. *J Allergy Clin Immunol* 2004;113:395–400.
22. Holt PG, Rowe J, Kusel M, et al. Toward improved prediction of risk for atopy and asthma among preschoolers: a prospective cohort study. *J Allergy Clin Immunol* 2010;125:653–9, 653–9.
23. Bantz SK, Zhu Z, and Zheng T. The Atopic March: Progression from Atopic Dermatitis to Allergic Rhinitis and Asthma. *J Clin Cell Immunol* 2014;5.
24. Belgrave DC, Granell R, Simpson A, et al. Developmental profiles of eczema, wheeze, and rhinitis: two population-based birth cohort studies. *PLoS Med* 2014;11:e1001748.
25. Pullan CR and Hey EN. Wheezing, asthma, and pulmonary dysfunction 10 years after infection with respiratory syncytial virus in infancy. *Br Med J (Clin Res Ed)* 1982;284:1665–9.
26. Lemanske R. F. J, Jackson DJ, Gangnon RE, et al. Rhinovirus illnesses during infancy predict subsequent childhood wheezing. *J Allergy Clin Immunol* 2005;116:571–7.
27. Jackson DJ, Gangnon RE, Evans MD, et al. Wheezing rhinovirus illnesses in early life predict asthma development in high-risk children. *Am J Respir Crit Care Med* 2008;178:667–72.
28. Lynch SV, Wood RA, Boushey H, et al. Effects of early-life exposure to allergens and bacteria on recurrent wheeze and atopy in urban children. *J Allergy Clin Immunol* 2014;134:593–601 e12.
29. Du Toit G, Roberts G, Sayre PH, et al. Randomized trial of peanut consumption in infants at risk for peanut allergy. *N Engl J Med* 2015;372:803–13.
30. Martinez FD. The coming-of-age of the hygiene hypothesis. *Respir Res* 2001;2:129–32.
31. Liu AH. Revisiting the hygiene hypothesis for allergy and asthma. *J Allergy Clin Immunol* 2015;136:860–5.
32. Brooks C, Pearce N, and Douwes J. The hygiene hypothesis in allergy and asthma: an update. *Curr Opin Allergy Clin Immunol* 2013;13:70–7.
33. Moffatt MF, Gut IG, Demenais F, et al. A large-scale, consortium-based genomewide association study of asthma. *N Engl J Med* 2010;363:1211–21.
34. Caliskan M, Bochkov YA, Kreiner-Moller E, et al. Rhinovirus wheezing illness and genetic risk of childhood-onset asthma. *N Engl J Med* 2013;368:1398–407.

35. Portelli MA, Hodge E, and Sayers I. Genetic risk factors for the development of allergic disease identified by genome-wide association. *Clin Exp Allergy* 2015;45:21–31.
36. Huang YJ and Boushey HA. The microbiome in asthma. *J Allergy Clin Immunol* 2015;135:25–30.
37. Hekking PP and Bel EH. Developing and emerging clinical asthma phenotypes. *J Allergy Clin Immunol Pract* 2014;2:671–80, quiz 681.
38. Lotvall J, Akdis CA, Bacharier LB, et al. Asthma endotypes: a new approach to classification of disease entities within the asthma syndrome. *J Allergy Clin Immunol* 2011;127:355–60.
39. Galowitz S and Chang C. Immunobiology of critical pediatric asthma. *Clin Rev Allergy Immunol* 2015;48:84–96.
40. Anderson GP. Endotyping asthma: new insights into key pathogenic mechanisms in a complex, heterogeneous disease. *Lancet* 2008;372:1107–19.
41. Wenzel SE. Asthma phenotypes: the evolution from clinical to molecular approaches. *Nat Med* 2012:716.
42. Lazic N, Roberts G, Custovic A, et al. Multiple atopy phenotypes and their associations with asthma: similar findings from two birth cohorts. *Allergy* 2013;68:764–70.
43. Teo SM, Tang HHF, Mok D, et al. Airway Microbiota Dynamics Uncover a Critical Window for Interplay of Pathogenic Bacteria and Allergy in Childhood Respiratory Disease. *Cell Host Microbe* 2018;24:341–352 e5.
44. Lynch JP, Sikder MA, Curren BF, et al. The Influence of the Microbiome on Early-Life Severe Viral Lower Respiratory Infections and Asthma—Food for Thought? *Front Immunol* 2017;8:156.
45. Callahan BJ, McMurdie PJ, and Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J* 2017;11:2639–2643.
46. Kusel MM, Holt PG, de Klerk N, and Sly PD. Support for 2 variants of eczema. *J Allergy Clin Immunol* 2005;116:1067–72.
47. Kusel MM, de Klerk NH, Holt PG, Kebadze T, Johnston SL, and Sly PD. Role of respiratory viruses in acute upper and lower respiratory tract illness in the first year of life: a birth cohort study. *Pediatr Infect Dis J* 2006;25:680–6.
48. Kusel MM, de Klerk N, Holt PG, and Sly PD. Antibiotic use in the first year of life and risk of atopic disease in early childhood. *Clin Exp Allergy* 2008;38:1921–8.
49. Teo SM, Mok D, Pham K, et al. The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development. *Cell Host Microbe* 2015;17:704–15.
50. Lemanske R. F. J. The childhood origins of asthma (COAST) study. *Pediatr Allergy Immunol* 2002;13 Suppl 15:38–43.
51. Custovic A, Simpson BM, Murray CS, et al. The National Asthma Campaign Manchester Asthma and Allergy Study. *Pediatr Allergy Immunol* 2002;13 Suppl 15:32–7.



## Chapter 2

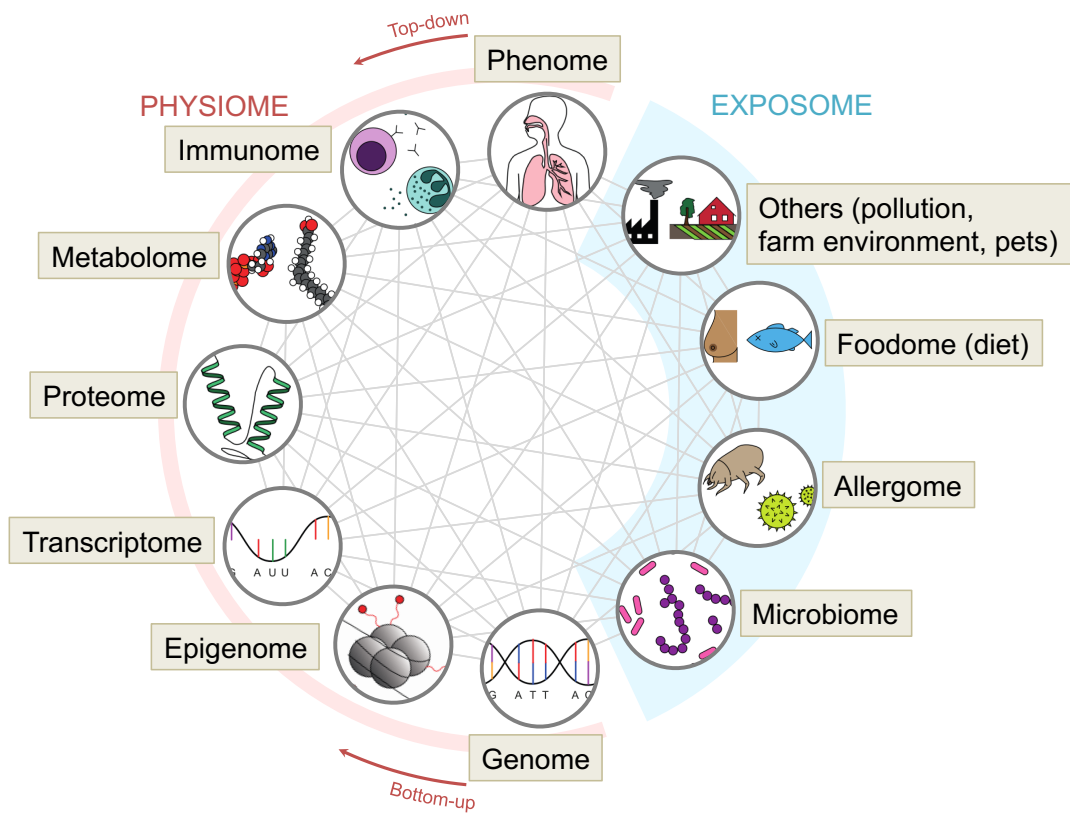
# Systems biology and “big data” in asthma and allergy — recent discoveries and emerging challenges

### 2.1 Introduction

The rising prevalence of asthma and allergy has been linked to changes in environment and lifestyle [1, 2]. But while we know of several genetic and environmental determinants of allergy and asthma, the potential interactions between these determinants remain unclear. Furthermore, asthma and allergy are umbrella terms that describe a spectrum of disease, with unexplained heterogeneity in clinical manifestations of disease. Finally, with the development of high-throughput technologies, we may be able to unravel some of this heterogeneity, but it remains challenging to process, analyse and interpret large volumes of biological data that emerge from these technologies. All these challenges have prompted researchers to search for new methods of inquiry more suited to these research problems.

Systems biology is a recent development that addresses the growing complexity of biomedical research questions. The term was coined in the 1960s to describe mathematical modelling of physiological systems [3]. Today it embodies expertise across multiple fields, including biology, mathematics, statistics, informatics and computer science. The “systems” community is diverse and as such there is no singular definition of the term “systems biology” [4]. However, it is commonly presented as the study of *biomedical* problems involving complex *systems* and their *interactions*, by surveying and *integrating high-volume data* that may cover wide *spatiotemporal scales* [3]. These “big datasets” typically originate from “omics”, fields of study involving high-throughput measurement of biomolecules: for instance, genomics for DNA, transcriptomics for RNA transcripts, and proteomics for translated proteins (**Figure 2.1**). Mathematical and computational expertise is then required to explore this high-volume data, using techniques such as dimension reduction; data- and text-mining; modified statistical analyses that account for spatiotemporal complexity and multiple testing burden; machine learning; and mathematical modelling. Therefore, systems biology is by its very nature *multi- and inter-disciplinary*.

The practice of systems biology follows two approaches: an unbiased, hypothesis-free *data-driven* approach, where few a priori assumptions are made and models are learnt from the data; and a *hypothesis-driven* approach, where model design and analysis are guided by previous experiments and expert knowledge [5]. The data-driven approach is becoming increasingly popular as it can uncover new knowledge on emergent behaviour from complex data. Systems biology can also be dichotomised into top-down versus bottom-up approaches, to describe the direction of enquiry in decreasing (big-to-small,



**FIGURE 2.1: "Omics" in allergy, and their interrelationships**

A depiction of the various "omics" that can be found in allergy and asthma research. Lines connecting the "omics" represent various biological relationships, associations or interactions that may exist. In systems biology, bottom-up approaches progress from the molecular scale to the macroscopic scale, and vice versa for top-down approaches.

long-to-short, system-to-components) or increasing spatiotemporal scales (small-to-big etc.), respectively (**Figure 2.1**) [6, 7].

On the surface, systems biology appears to be antithetical to the “reductionist paradigm” of old. However, systems-based approaches can produce new insights on how to proceed with reductionist experiments, and vice versa. Also, there are strengths and weaknesses attached to each; while reductionist methods can over-simplify problems, their tests are more appropriate in contexts such as causal inference. Nonetheless, systems approaches are becoming indispensable to biomedical research; they allow us to better understand disease phenomena, and form the basis for precision medicine, helping us improve the screening and management of disease.

Asthma and allergy, as biomedical problems, are well-suited to systems approaches. Allergic diseases have complex pathogenesis, with multiple tiers of biological complexity, polygenicity and gene-environment interactions. Systems approaches used in allergy research include: (1) discovery of disease associations within each omic field; (2) identification of relationships within and across omic fields; (3) examination of heterogeneity of disease states and phenotypes, typically by exploring the multidimensional structure of omic data via clustering or classification; (4) investigation of inter-connections between system components in omic data by network analysis; and (5) mathematical modelling to model physiological systems or disease states, and to generate and test predictions (**Figure 2.2**). Though the final approach is closest to the original formulation of systems biology, our review will take a high-level look at all approaches, with a focus on the first three.

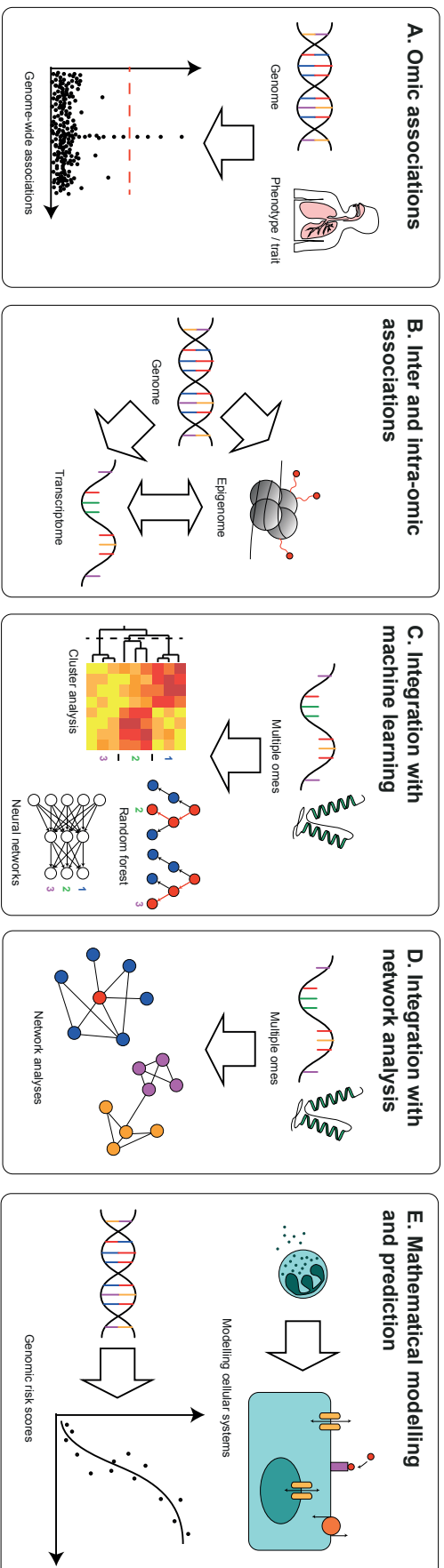
## 2.2 Overview of omic findings in allergy and asthma

We begin our examination of systems biology from bottom up – from the molecular level of genomes and transcriptomes, to the macroscopic level of observable phenotypes. We offer high-level summaries of recent findings at each level of profiling. Of all allergy-related diseases, asthma has come under the most scrutiny, so much of our discussion will revolve around it.

### 2.2.1 Genomics

In allergic disease, the genome is the most-studied of all the omic fields. Asthma and allergic disease are highly-heritable, with estimated heritability ranging from 35 to 95% [8]. In the last half-century, the quantitation of genetic variation has progressed from rough “ballpark” measurements such as restriction fragment length polymorphisms (RFLP), to precise single-nucleotide variants or polymorphisms (SNV/SNP) interrogated en masse using DNA microarrays. More recently, there has been a move towards whole exome (WES) and whole genome sequencing (WGS).

The complexity of genetic data analysis has grown in parallel, from candidate gene studies, to genome-wide linkage studies within pedigrees, to genome-wide association analyses (GWAS) involving case-cohort comparisons [9]. With GWAS and other omic-wide analyses, statistical adjustments are made for multiple testing burden. But, although some older candidate associations have been replicated by GWAS (e.g. *IL13/IL4* and *IL4R*), most have not. Furthermore, there is low concordance of significant results between genome-wide linkage studies and GWAS. These suggest that (1) older findings may be plagued by false positives; (2) each approach may have its own use: positional candidates from linkage studies may flag variants determining intra-family disease risk, while GWAS flag variants determining population-wide risk; and (3) rarer or weaker gene associations may



**FIGURE 2.2: Overview of systems-based approaches to tackling research questions in allergy and asthma**

The various ways in which systems biology of allergy can be interrogated: (A) discovery of disease associations within each omic field of enquiry; (B) identification of relations within and across omics; (C) examination of the heterogeneity of disease states or phenotypes, typically by exploring the structure of omic data via clustering or classification; (D) investigation of inter-connections between system components in omic data by network analysis; and (E) mathematical modelling to model physiological systems or disease states, and to generate and test predictions. Diagrams are for illustrative purposes only and do not convey real data.



be swamped by noise in GWAS, due to inadequate power from small sample sizes, and heterogeneity within cases and controls.

In the last 10 years, GWAS have identified many replicable loci shared across multiple allergic phenotypes, especially asthma, allergic rhinitis, atopic dermatitis and food allergy. These likely represent genetic contributors to general allergy, and include: the HLA locus, specifically *HLA-DQ/DRB1*, *HLA-DQA1/2* and *HLA-B/C* (6p21.32-33); *C11orf30/LRRC32* (11q13.5); *IL13/IL4/RAD50* (5q31.1); *IL1RL1/IL18R1* (2q12.1); and *TSLP/WDR36* (5q22.1) [9–12]. Some of these have plausible biological underpinnings; the HLA region encodes major histocompatibility complex (MHC) Class II molecules that are responsible for antigen presentation. *IL13*, *IL4* and *TSLP* encode cytokines related to the Th2 pathway, which is heavily implicated in allergy. Other associations remain uncertain or even dubious in terms of pathophysiology, and require further investigation.

There are also disease-specific loci, especially those specific for asthma, which may act independently of loci for general allergy. In particular, *ORMDL3/GSDMB/LRRC3C* (17q21.1) is linked with asthma, especially of childhood-onset [13–15]. More recently, there has been a focus on loci with ethnicity-specific effects – for instance, *PYHIN1* is significantly associated with asthma, but only in individuals of African ancestry [14]. There is also an increasing focus on using admixture to map risk loci [8]. Finally, associations distinct from disease-susceptibility loci have been identified for responsiveness to asthma therapy with  $\beta_2$ -agonist bronchodilators, leukotriene modifiers, and steroids [16, 17]. There are many other loci for related respiratory phenotypes, such as lung function and viral respiratory infections [18–21]. Some of these are shared with allergy and asthma phenotypes, suggesting shared mechanisms of pathogenesis.

Despite the large number of novel associations discovered using GWAS, these collectively explain only a small proportion of the total heritability of asthma and allergy. The use of significant SNPs as a predictive tool for disease is often limited [22]. The existing criteria for genome-wide significance may not be sensitive for so-called “mid-hanging fruit” [23]: loci that are not genome-wide significant but still have an incremental effect on the phenotype. More recently, alternative strategies such as genomic or polygenic risk scores have been employed to account for this missing signal. These use the summary statistics from existing large-scale GWAS to generate scores from either the entire genome-wide assortment of SNPs, or from a select few of high-predictive SNPs. Although it is difficult to surmise any pathophysiology from the scores themselves, they have shown promise as predictive or risk-stratifying tools for other chronic polygenic diseases such as cardiovascular disease [22].

### 2.2.2 Transcriptomics

The transcriptome represents the entire repertoire of genes expressed in an organism or cell. Mirroring the developments in genomics, there has been a move from investigation of single-gene transcripts via traditional methods (e.g. Northern blotting), to genome-wide methods involving oligonucleotide microarrays, and most recently to RNA sequencing (RNA-seq, which involves reverse transcription to cDNA followed by deep sequencing) [24]. Transcriptomes may be determined by aligning RNA-seq reads to transcripts annotated in a reference genome, or assembling transcripts de novo, followed by quantification based on abundance of reads per transcript.

Unlike the genome, the transcriptome varies across tissues and cell types, and changes dynamically during development and in response to external stimuli. Common tissue sources for transcriptomics include blood with or without cell sorting; bronchial epithelium, smooth muscle, or sputum cells for asthma; nasal epithelium for allergic rhinitis; and skin for atopic dermatitis. Different cell types may feature different associations, and

this provides insight into how various genes contribute to the many manifestations of allergy.

Recent studies of allergy have identified, across multiple tissue types, differential expression of genes involved in innate and adaptive immunity, inflammatory and repair responses, and epithelial integrity. Cytokines (Th2 and others), chemokines and their receptors, protease inhibitors (SERPINs) and other multifunctional regulatory proteins (S100 family) have been identified in multiple transcriptome analyses of allergic diseases [25–29]. Baines et al. [25] described transcriptional profiles specific for inflammatory phenotypes in adult asthma, and found that *IL1 $\beta$*  and *CXCR2* levels were predictive for neutrophilic inflammation. Differential gene expression between case and control may be explained by differences in relative numbers of immune cell populations, due to proliferation or contraction [25, 30]. Otherwise, they may relate to actual differences in behaviour of immune cell subsets, such as T cells (Th2, Th17, Treg), and cells of innate immunity (e.g. eosinophils) [31–35]. Specifically, allergic rhinitis is associated with transcriptome profiles representing exaggerated Th2 activity and aberrant Treg function, and treatment with steroids may reverse some of these changes [35].

Some significant loci in genetic association studies are also differentially expressed in allergic disease: for instance, genetic variants in Th2 cytokines (*IL5*, *IL13* etc.) have been associated with asthma in case-control studies, and *IL13* is upregulated in a Th2-dominated subphenotype of childhood asthma [31]. Other asthma-associated loci are expression quantitative trait loci or SNPs (eQTLs/eSNPs); these influence expression of nearby genes, for instance by altering the regulatory region of those genes (as cis-eQTLs) [36]. Local eQTLs associated with allergy have been found in *TSLP/WDR36* determining *TSLP* expression; in *ORMDL3/GSDMB/LRRC3C* for *GSDMA* and *GSDMB*; and in *IL1RL1* [37–40]. Yet many other genes are differentially expressed in asthmatics but have not yet been linked to allergy- or asthma-related variants from GWAS. These may represent manifestations of inflammatory pathology downstream of genetics; or they may represent drivers of interactions with environmental exposures.

### 2.2.3 Epigenomics

The epigenome is the set of heritable biochemical modifications that change gene expression, but are not coded in the DNA sequence. Epigenetics functions as a bridge between genome and transcriptome, providing mechanisms by which the micro- or macro-environment can influence gene expression within each cell; and by which transgenerational inheritance can occur after initial exposure to an epigenome-modifying environment [41, 42]. Epigenetic signals include: (1) DNA methylation at CpG islands, which silences expression of adjacent genes; (2) histone modifications (acetylation, methylation, and others), whose effects vary depending on type and position of modification; and (3) non-coding RNA such as microRNA (miRNA), which can silence genes by binding or degrading complementary mRNA [43]. Together, these epigenetic markers cause changes in accessibility of a local DNA segment to transcription or regulatory factors.

Low and high-throughput detection methods exist for each type of epigenetic signal. Methylation-sensitive restriction fingerprinting and microarrays for detecting 5-methylcytosine have been used to describe the DNA “methylome”. Genome-wide histone modifications can be detected using chromatin immunoprecipitation (ChIP). Next-generation sequencing options also exist (miRNA-seq, DNase-seq, FAIRE-seq, ChIP-seq, 3C-seq), which function by isolating DNA fragments that are accessible or inaccessible to a factor of interest, and sequencing those fragments to determine their identity [7]. Epigenome-wide association studies (EWAS) can then be performed to identify epigenetic features for a given trait or disease. Finally, like the transcriptome, the epigenome is

responsive to external stimuli and varies across cell types, and most epigenomic studies of allergy have so far examined blood, skin, or airway samples.

There is evidence that development and maturation of T cell lineages is partly determined by epigenetic changes [41]. Th2 differentiation is driven by *STAT6* and *GATA3*, resulting in epigenetic changes (DNA methylation, histone acetylation) that induce Th2-related (*IL4/IL13*), and suppress Th1-related (*TBET*, *IFNG*, *IL-12/STAT4* pathway) expression; conversely, Th1 differentiation is driven by *STAT4* and *TBET* to elicit the opposite epigenetic changes; finally, Treg differentiation is driven by *STAT5*, with associated epigenetic changes in *FOXP3* and the *IL10* locus [41, 43]. Given the role of epigenetics in T cell development, it is plausible that allergic disease may be linked to altered epigenetics affecting this process. Epigenetic signals have been observed across multiple tissue types in allergy. Changes to DNA methylation have been noted in loci related to Th2 function and T cell development (*IL4R*, *TSLP*, *IFNG*, *FOXP3*, *STAT5A*) [42, 44–48], while other significant loci control antigen presentation, eosinophil activity, lipid metabolism, and mitochondrial function [49, 50]. The relationship between histone modifications and allergy or asthma is less clear. Some studies have shown changes to global histone acetylation, with reduced deacetylating-to-acetylating (HDAC-to-HAT) activity in asthmatic lungs compared to normal [51–53]; while others suggest that HDAC inhibition can improve the suppressive function of Tregs [54]. Similarly, certain miRNAs are known to influence allergy risk. For example, Okoye et al. observed that miR-155 and miR-146 may be critical in determining T cell differentiation towards Th2 versus Th1/Th17 [55]. Other relevant miRNAs are reviewed elsewhere [56, 57]. Investigation of the full compendium of miRNA species is progressing rapidly, and may lead to new targeted therapeutics.

An important aspect of epigenetics is the link to environmental exposures. Because the development of the immune system begins in utero and continues through infancy, environmental modifiers of epigenetic signals may have a stronger impact earlier in life. Experimental and observational studies show that maternal exposures during pregnancy and exposures during early childhood can modify the child's epigenome. These exposures include changes to diet, macro- and micro-nutrition, farm environments, infections and microbes, animals, allergens, medications, pollutants, tobacco smoke, and even maternal stress [43, 58, 59]. In particular, folate and Vitamin B12 are methyl donors that have a global impact on DNA methylation [43]. Finally, genome associations have been identified for methylation patterns as quantitative traits (meQTLs). These include the *ORMDL3/GSDMB* locus, where a SNP behaves as both an eQTL and a meQTL [60], and others [49, 61, 62]. All these findings illustrate that certain perinatal exposures can act through genetics and epigenetics to influence disease risk.

#### 2.2.4 The microbiome

The microbiota is the community of microbes, including commensals and pathogens, that reside within a host or environment, while the microbiome is the genomic content that represents the microbiota. The “microbiota hypothesis”, a modern re-iteration of the hygiene hypothesis, suggests that perinatal microbial exposure is vital to proper development of immune functions, especially of tolerance [63–65]. Microbial exposures may modify allergy susceptibility by initiating different trajectories of immune development and function [58]. Epigenetic changes may also be involved in this process.

The primary interfaces for host-microbe interactions are the epithelial surfaces exposed to the external environment – in the skin, respiratory, and gastrointestinal tracts – so most studies on allergy microbiomes involve sampling at one of these sites directly (biopsy or surface samples) or indirectly (faecal or sputum samples). The gut is home to gut-associated lymphoid tissue (GALT), and its microbiome can influence disease at other

mucosal surfaces, such as the respiratory tract [66, 67]. The respiratory microbiome may also exert a direct influence on local inflammatory processes leading to asthma development [68]. The environmental microbiome may drive restructuring of host microbiomes, or modify allergy risk by other means; this may be particularly relevant in relation to the protective effect of farming environments [58]. Description of the microbiome relies mostly on quantification of DNA sequences encoding the 16S ribosomal RNA (rRNA) gene, which is common to all bacteria but contains variable regions used to differentiate taxa. The gene sequence is amplified using PCR and then examined using gel electrophoresis, terminal RFLP, microarrays, or sequencing. Recently there has been a transition to deep metagenomic sequencing, which captures the genomes of all organisms present in a sample, not just the 16S rRNA gene, and can be used to infer both taxonomic composition and function of the microbial community.

Microbiome studies are complicated by the fact that host microbiomes can change with age, season, time of day, site of sampling, and geographical location [69, 70]. However, a number of consistent findings have been established for asthma and allergy. Features of the gut microbiome associated with allergy include early-life reduction in microbial diversity; reduced populations of Bifidobacteria, Lactobacilli and Bacteroidetes; and increased coliforms and specific Firmicutes (Staphylococci, Enterococci) [66, 67, 71]. Reversing the above changes, for instance by oral administration of certain *Lactobacillus* and *Bifidobacterium* species, may offer some protection against both the initial development of allergy and further exacerbations of atopic disease [63]. Within the airway microbiome, asthma development, symptoms and exacerbation have all been associated with increased Proteobacteria populations (especially *Haemophilus*, *Moraxella*, *Streptococcus* and *Neisseria* spp.), and reduced Bacteroidetes and Fusobacteria commensals [63, 64, 67, 68, 72]. Remarkably, these associations begin during infancy: the detection of asthma-related bacteria in the first few months of life has been associated with developing allergic asthma by primary school age [64, 66]. Though it is unclear whether microbial changes represent a cause or effect of underlying immune dysfunction, there is evidence of altered gut and airway microbial communities preceding allergic sensitisation [68, 73, 74]. Ultimately, these findings suggest two independent processes at work: microbiota, especially of the gut, exerting systemic effects on immune maturation; and microbiota causing local inflammatory processes at the sites they inhabit, including those associated with asthma in the respiratory tract.

Other recent studies have uncovered the potential role for non-bacterial microbes, namely viruses such as human rhinovirus and respiratory syncytial virus (RSV), in causing early childhood wheeze which often precedes full-blown asthma [68, 75, 76]. There is evidence for the role of rhinovirus (RV), specifically RV-C, in causing severe respiratory illnesses that are associated with increased asthma risk later in life. The pathophysiology behind this may be related to chronic airway injury due to recurrent infection, possibly interacting with allergic mechanisms, to elicit and maintain sustained inflammation [75].

### 2.2.5 The exposome and environmental exposures

Researchers have frequently explored the relationship between environmental exposures and disease. The “exposome” builds on this idea by encapsulating all environmental exposures that contribute to human health and disease. The environmental microbiome, for instance, is just one type of exposure; even the host microbiome can be considered an exposure when describing microbes residing on the skin, or on luminal surfaces of hollow viscera exposed to the outside world. It is difficult to measure all exposures, let alone on a high-throughput scale, and there are other challenges related to correlation, confounding, and interaction amongst different exposures [77]. Instead, most studies have so far quantified a limited set of relevant exposures via questionnaires and environmental

sampling. North et al. is one of the first studies to adopt an exposomic approach to examine multiple types of exposures simultaneously, in their search for associations with childhood wheeze [78].

The environment can contribute to asthma and allergy pathogenesis in many ways. As mentioned in the epigenomics section, these include mechanisms acting through diet and nutrition, exposures to pets and animals, allergens, pollution, and tobacco smoke. For some of these, it is possible to measure and perform high-throughput analyses on proteomic and metabolomic data. Diet is one example: a protective effect against allergy has been reported for polyunsaturated fatty acids (PUFAs) found in fish oil, and for their metabolites [79]. So far, in allergy research, lipidomic analyses of fatty acids have been limited to biological tissue samples. There has been slow adoption of similar analyses in food [80], but in the future, it may be possible to scan the contents of an individual's diet in a high-throughput manner, construct a "foodome", and search for de novo associations with disease. Airborne pollutants may also be explored in a similar manner.

Environmental allergens can themselves be investigated by multiple omic approaches, in relation to quantity of exposure, geography of exposure, and allergenicity of protein structures. For instance, studies have identified that low environmental load of allergen can be a risk factor for disease [81, 82]. Timing and route of allergen exposure may also be relevant: early introduction of solids, including peanuts, may be protective, but only within a specific time window [83]. Also, early exposure to peanut allergen through the skin may promote sensitisation, while exposure through the gut may promote tolerance [84]. Other studies have overlaid geographical maps of exposure with maps of disease, as has been done for traffic-related air pollution and asthma [85]. Finally, it is still not clear why allergens behave as allergens. The term "allergome" is typically used to describe the proteomics-based discovery of allergenic protein structures within individual allergens – discussion of this is deferred to the proteomics section.

As previously alluded to, environmental exposures can act through interactions with host microbiome to modify disease risk [58, 66, 67]. Maternal and perinatal exposure to rural environments confers some protection, possibly due to contact with microbial products such as lipopolysaccharide (LPS), greater diversity in microbial exposure, or environmental modification of host microbiota. Caesarean deliveries and perinatal use of antibiotics may increase risk for allergy, possibly by disrupting neonatal microbial colonisation. The protective effect of oral probiotics with *Lactobacilli* and *Bifidobacterium* spp. has been reported, as noted previously, and they may also provide cross-organ protection, reducing the incidence and severity of respiratory infections [67]. The use of dietary fibre as prebiotics, with subsequent fermentation into short-chain fatty acids (SCFAs), may protect from allergy via TLR and GPCR signalling or epigenetic modifications [63, 64]. Vitamin D has potential immune and microbiome-modifying effects, and Vitamin D deficiency is a suspected risk factor for allergy [86, 87]. Breastmilk contains immunoactive molecules and may alter gut microbiota composition [63]. Finally, early-life viral respiratory infections contribute to onset and progression of asthma, and may act synergistically with allergic sensitisation to compound disease risk [75]. Colonisation with *Moraxella*, *Streptococcus* and *Haemophilus* spp. is associated with more frequent infections [68], and there are patterns of co-association between certain bacteria genera and viral pathogens [88]. Other microbe-specific omics such as the virome and the (fungal) mycobiome may also be relevant [58, 89]. Altogether, these findings offer a glimpse into how multiple environmental exposures may interact in a complex fashion to elicit disease.

### 2.2.6 Proteomics, metabolomics, and lipidomics

The proteome is the repertoire of proteins produced by cells or tissues, reflecting the molecular effectors and metabolic consequences of cell function. Common proteomic technologies can be grouped into antibody-based (ELISA), peak-profiling mass spectrometry (MS)-based (“fingerprinting”), gel-MS based (1D/2DG, 2D-DIGE), and LC-MS-based methods [90]. The general approach is to perform coarse separation of digested proteins into “bands” or “spots”, and then further investigate each spot by MS. The MS steps are often done in tandem (MS/MS) to achieve higher resolution. The information gained from MS can then be used to identify the peptide, or construct its amino acid sequence. Sources of proteomic samples include sites of pathology such as the airway, in the form of cellular or fluid content from bronchoalveolar lavage (BAL), induced sputum, biopsies, or in vitro cell cultures; or it may involve the usual blood or urine sample [91]. An accessible type of specimen unique to asthma research is exhaled breath condensate (EBC), which provides information on volatile compounds released from the airway.

In relation to allergy research, proteomic changes often depict non-specific pathology, as in a general elevated inflammatory state, as well as underlying pathological mechanisms. Therefore, while recent findings in allergy proteomics partly mirror transcriptomic changes, they also reflect altered functions in immunity, inflammation and anti-protease activity: affected proteins include defensins,  $\alpha$ -1 antitrypsin,  $\alpha$ -2 macroglobulin, SERPINs, S100-family proteins, apolipoproteins and complement proteins [90–93]. Interestingly, few recent studies have linked direct changes to Th2-specific cytokines on a proteome-wide scale, although associations have been identified in low-throughput in vivo studies in the past [94].

Another important contribution of proteomics to allergy research is allergen detection and discovery [95]. Recent studies have investigated a compendium of epitopes for aeroallergens such as house dust mite [96–98] and plant pollen [99–101], and for food allergens in seafood and processed foods [95]; these have served both to confirm existing epitopes and to identify new ones. Findings from these studies can be applied to non-clinical settings, such as food processing and safety [95].

Metabolomics is the systems-level study of metabolites – the non-peptide macromolecules representing the substrates and end-products of cellular activity. The two main technologies of measurement used in metabolomics are nuclear magnetic resonance (NMR) – which provides a spectral “fingerprint” of a system’s metabolite constituents – and mass spectrometry (MS). Like proteomics, most metabolomic studies focus on samples of blood serum, EBC and urine from asthmatic patients [102–104]. Lipidomics is a subset of metabolomics specifically dealing with lipid molecules, and lipidomic studies have shown that allergic disease is typically associated with elevation of arachidonic acid metabolites belonging to the LOX pathway, such as leukotrienes [105, 106]. Metabolomic associations with asthma involve immune and inflammatory functions, oxidative stress and hypoxia, cellular energy homeostasis, and lipid metabolism pathways [103]. These associations may not be specific for allergy or asthma, but rather may simply reflect general biological stress or inflammatory pathology. However, predictive and discrimination models based on metabolomic findings have shown some promise [103]. Finally, proteomic, metabolomic and lipidomic methods may be applied not just to host samples, but also to environmental samples, as alluded to in our previous discussion.

### 2.2.7 The phenome and physiome

Our final section on omics concerns phenomics, a broad term encompassing all physical or biochemical traits (phenotypes), observable in cells or individuals, that reflect states of

disease or health (“physiome”). In the case of allergy and asthma, possible phenotypes include cell types based on morphology and response (immunophenotyping); clinical biomarkers, such as antibody assays and cell counts; and the extensive physical manifestations of disease, embodied in clinical history, symptoms and signs, and investigation results. These traits may be quantified and described in detail, though not necessarily using high-throughput technologies.

Phenome-wide association analyses, where large sets of traits are screened for enrichment of allergy-related genetic loci [107, 108], have been performed in the past, but have yet to gain widespread popularity. Phenomes and phenotypes can also be analysed by machine learning, whether it be comparison of known phenotypes (via supervised classification) or construction of new phenotypes from omic or non-omic data (via unsupervised cluster analysis). This will be discussed in further detail later.

The immunome is a subset of the physiome that is highly-relevant to allergy, and where high-throughput technologies play a major role. Immunomics broadly describes the systemic quantification of immune function by examining immune cell populations and expression of immune mediators. It may use immunoglobulin [109–111] and cytokine (proteomic or transcriptomic) arrays [112, 113] to quantify immune responses such as sensitisation, *in vivo* or *in vitro*. It can also involve leukocyte immunophenotyping and high-dimensional or mass cytometry [114–117]. The immunome is complex and varies dramatically by sampled immune cell type, tissue or organ, age, and timing of sampling, especially before and after sensitisation. Although it is well-known that allergy is a Th2-driven phenomenon, it is still not clear how all the components interact to generate disease, nor is it clear how disease heterogeneity is explained by immunome heterogeneity. Future studies may be able to shed light on this.

## 2.3 Integration of omics data

Following our overview of the omics, we now discuss common techniques used to integrate and interpret omics data in allergy and asthma research.

### 2.3.1 Exploring intra- and inter-omic relations

To understand disease pathogenesis, it is natural to compare findings across different omics, and construct a multiomic model of pathophysiology that links these various elements together. This may be a simple sequential model of causality, or a complex network of interacting components. Many studies on omic associations with allergy and asthma also search for inter and intra-omic relationships. Relationships can take the form of direct associations, where one entity behaves as a trait for another; or an interactive effect between two entities in relation to a third entity as the trait of interest. The study of these relationships is the crux of modern systems biology.

Genomics, being the most well-studied system in allergy and asthma, features extensively in intra-omic and inter-omic analyses. GWAS can be found not only for clinical phenotypes (e.g. presence of allergic disease) as traits, but also for expression of transcripts (eQTL analyses), epigenetic markers (meQTLs), and intermediate phenotypes such as microbial exposures and immunomes. Recently, there has been a concerted move towards integrative genomics, and genetic effects on gene expression are a pervasive component of modern association studies – in the form of mandatory genome-wide eQTL analyses or targeted measurements of gene transcripts [57]. Also coming into vogue is the use of Mendelian randomisation, a technique which uses genomic information as instruments to infer causal links between one trait or phenomenon and another, based on the assumption

that allelic genotypes are randomly assigned as they are passed from parent to offspring [118]. The traits being linked may themselves be related to gene loci or expressed genes [119].

Analyses for interactive effects with other omics also feature heavily in allergy genomics. It is unlikely that genetic and environmental factors act independently in conferring risk, so modern genomic studies often include interaction terms with exposure variables. Scientists have explored interaction effects on asthma susceptibility between genetics and exposures such as air pollution and tobacco smoke [120, 121]. Another example is the impact of allergen exposure and genetics on immune cell gene expression [122]. Interaction analyses also extend beyond environmental effects. Gene-ethnicity interaction has been investigated via admixture mapping [8]. Genetic-epigenetic interactions have been reported; some genome-wide significant loci (e.g. *IL4R*) may interact with nearby epigenetic signals to alter disease risk [48]. While investigation of gene-gene interaction (epistasis) is of intense interest, the overwhelming number of active genes in the human genome means that such analyses have a large statistical burden and hence remain difficult. Therefore gene-gene interaction studies are so far limited to a few selected genes or SNPs. Finally, interactive effects can be explored by means beyond using interaction terms in regression models: for example, eQTL-weighted GWAS have been reported [123].

Given the strong links between environmental factors and asthma, interactions with environment exposures have been explored to a degree. In particular, microbial and pathogen exposures have been linked to differential gene expression; for instance, viral infections are associated with changes to airway epithelial transcriptomics in asthma [124, 125]. Unsurprisingly, the exposome and microbiome have been linked to epigenetic changes, and the various exposures are intricately entwined in complex interactions. For instance, a recent study has looked at the interaction between air pollution and the allergenicity of ragweed pollen [126]. Another recent study has identified that maternal phthalate exposure may promote allergy in subsequent generations via epigenetics [127]. Other examples concerning environmental interactions with diet and microbiome have already been discussed.

Finally, a common application of integrative omics is the use of gene ontology analysis to annotate discovered genes from genomic, transcriptomic, or epigenomic analyses [128]. This makes use of a pre-curated database of functional annotations for known genes, based on existing literature, to segregate discovered genes into groups or pathways with shared functions. An example is the Gene Ontology Consortium [129]. These databases of functional annotations convey phenomic information, where cell phenotypes, functions and behaviours are organised into discrete categories. In doing so, one aims to condense diverse genome-wide findings into concise summaries of biological function that may be easier to interpret when building a conceptual model of pathophysiology. Similar annotation analyses exist for proteomics [130, 131]. A limitation of such techniques is that the annotations may not always be certain, reliable, or up-to-date, and can often be vague or uninformative.

Inter- and intra-omic relationships may be explored either by low-throughput pairings, or by high-throughput assessment of larger networks [132, 133]. However, especially with the latter, it may be difficult to account for non-causal correlations or confounders. For example, despite the hygiene hypothesis, low socioeconomic status and impoverished environments remain risk factors for the development and severity of asthma [67]. This may be due to confounding factors that coexist with poverty, including urbanised environments, exposure to allergen and pollutants, dietary intake, and access to health care. There is no doubt that modifiers of allergy risk may co-occur together, but whether this represents a causal link is another matter. Methods such as Mendelian randomisation (MR, described previously) may be used to disentangle this, but one must be wary of violating



the numerous assumptions that underlie MR. Also, given the high dimensionality of inter- and intra-omic analyses, one may instead use dimension reduction and machine learning to identify potentially robust signals of relevance to pathogenesis.

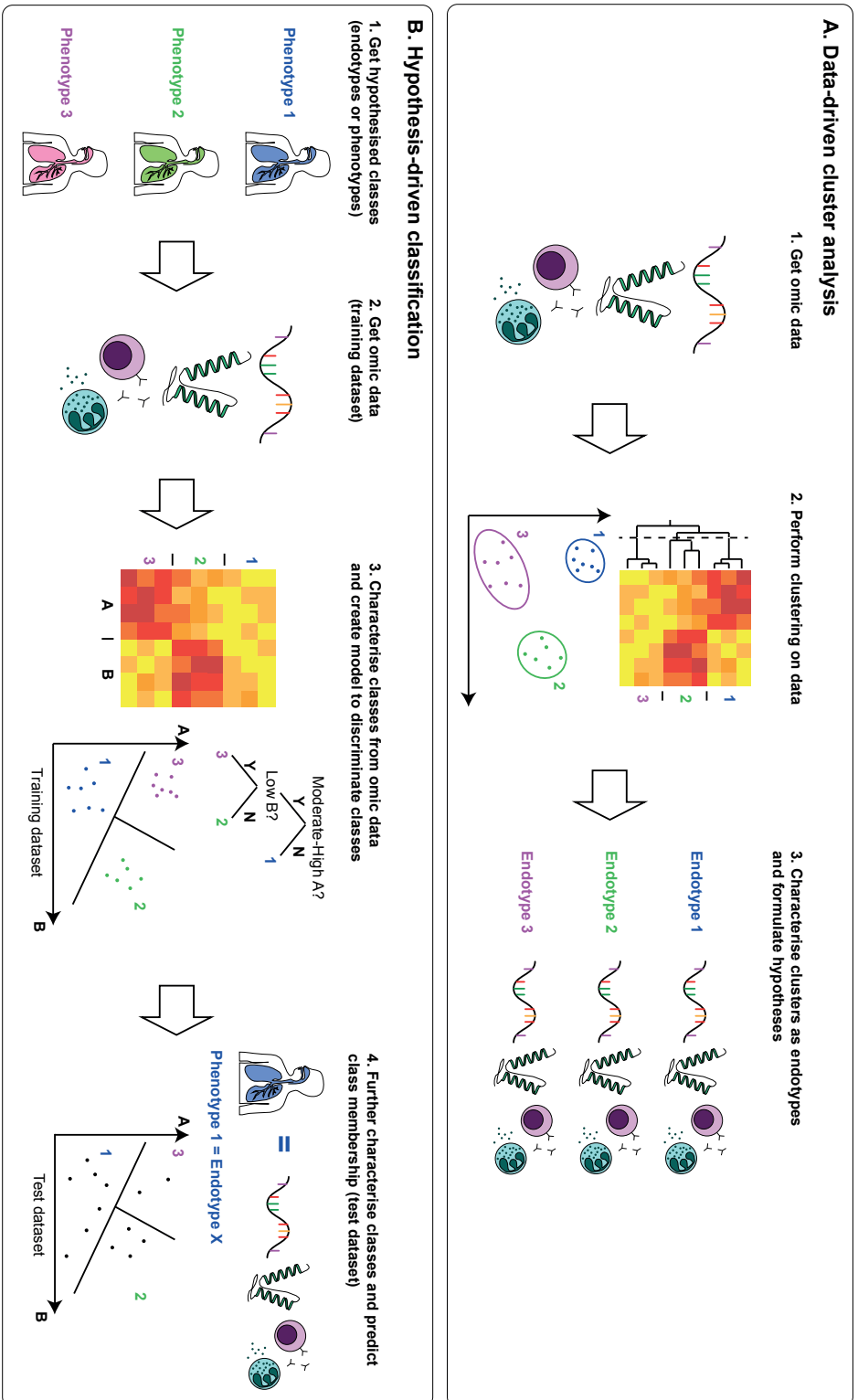
### 2.3.2 Machine learning, dimension reduction, and clustering

Machine learning is a set of methods that use computing to learn and formulate solutions from supplied data, with or without explicit human input. It is already in common use with various biomedical and ecological applications [134–136]; however, it is particularly useful when dealing with complex, high-throughput, and multidimensional data — especially in cases where pre-existing human knowledge may be unavailable or insufficient to decipher the data. Machine learning methods are often used to subset data samples or individuals into different groups or categories. This can either be done with cluster analysis or classification. Cluster analysis is a data-driven approach, where omic data is used to generate clusters in an unsupervised fashion. The clusters can then be interpreted for hypothesis generation and testing. On the other hand, classification is a hypothesis-driven approach: known phenotypes are used to determine a model of classification based on training data, which can then be applied to other datasets, or examined to look for further biological associations (**Figure 2.3**).

A drawback of machine learning is that there is little consensus or standardisation of optimal methods, although there are certainly favoured approaches for each problem. They are also intimidating for the regular clinician or biologist to adopt, and choice of method often depends on a specialist understanding of nuances in the data. As an example: when performing cluster analysis, many decisions need to be made prior to and during the procedure. This includes how to deal with missing data; select the variables or “features” for clustering; scale or normalise features; choose the algorithm to do the actual clustering; pick the number of clusters; control for overfitting; and validate or replicate results [137].

In exploring the correlation structure and confounders in a dataset, one can use principal components analysis (PCA), or similar methods, to transform the dataset into uncorrelated variables or “principal components”. In doing so, one can observe which of the original variables describe similar information (i.e. are highly-correlated with each other); and by plotting principal components, one can visualise the data in a way that maximises variability between samples or variables. By condensing our data to a limited selection of principal components, we can reduce the number of dimensions and simplify the input features for subsequent clustering or classification [138]. Feature selection can be limited to a single omic entity, or cover multiple omics simultaneously, depending on the question asked.

Cluster analysis involves separating samples in a dataset into discrete groups (*clusters*) based on what can be learnt from data structure, without specifying training examples for each group [137, 139]. Its objective is to minimise intra-group and maximise inter-group differences. Measures of difference or *dissimilarity* may be distance- or correlation-based. Common clustering techniques include hierarchical clustering, medoid-based methods, and latent variable modelling. Cluster analysis allows one to identify homogeneous groups within a heterogeneous dataset, and simplify analyses to comparisons between clusters rather than across entire cohorts. Clustering can also expose confounders without explicit adjustment for correlation, especially if clustering is “guided” by co-segregating omic variables. Using molecular omic-based features, cluster analysis may allow us to determine *endotypes* – subtypes of disease or health states – by common biomolecular interactions and pathophysiology [140]. These can be compared with known phenotypes to explore how variation in pathophysiological mechanisms are linked to variation in disease manifestations. Cluster analysis can also be applied to phenome data to deal with



**FIGURE 2.3: Data-driven versus hypothesis-driven machine learning for integration of omic data**

(A) Data-driven (unsupervised) cluster analysis to generate de novo groupings; reflective of shared pathophysiology (“endotypes”); (B) hypothesis-driven (supervised) classification to compare known phenotypes or endotypes, and to allow prediction of phenotype/endotype membership for additional samples. Diagrams are for illustrative purposes only and do not convey real data.

heterogeneity in phenotypes. Using “cleaner” sub-phenotypes for association analyses may improve the power and specificity of subsequent findings.

Classification methods determine a statistical model or decision-making algorithm that allocates individuals of a training dataset into known groups (*classes*) [137]. The learnt model or algorithm can then be applied to other test datasets for classification into classes. Methods include regression analysis, discriminant analysis, support vector machines, and partitioning or decision trees. The objective of classification varies with the method, but mainly involves achieving the “best fit” – minimising differences between predicted and actual class allocation for the training dataset – without compromising generalisability to external datasets. Classification can be used to design diagnostic or risk stratification algorithms from an omic dataset. Each sample is labelled as one of a predefined set of phenotypes (e.g. allergic versus non-allergic asthma, eosinophilic vs. neutrophilic, severe versus non-severe), then the algorithm seeks biomolecular or clinicophysiological features that best define the phenotype [33, 141]. In the absence of predefined phenotypes, one may combine both clustering and classification: generate clusters based on a training dataset, then devise a classifier which can classify test datasets into the discovered clusters.

Both cluster analysis and classification have been used extensively in asthma research. Major findings from such analyses include the discovery and characterisation of different subsets of childhood and adult asthma. Childhood wheeze has been categorised, by both traditional and machine-learning approaches, into persistent atopic wheeze of early onset, transient remitting viral wheeze, and a mixed atopic/non-atopic phenotype of variable onset [142–144]. Atopic wheeze appears to be characterised by Th2 activation, early sensitisation to allergens, greater severity of respiratory disease, greater likelihood of persistence to full-fledged allergic asthma, and concurrence of other atopic diseases. In terms of adult asthma, there are subtypes based on lung function [145], as well as atopic, non-atopic, mixed and other phenotypes [140]. Eosinophilic, neutrophilic and paucigranulocytic airway inflammation can also be distinguished from sputum samples, and accompanying transcriptomic, proteomic and immunomic data can provide some insight into underlying pathophysiology for each phenotype [25, 141, 146, 147]. Neutrophilic, Th1/Th17-dominant, and steroid-resistant asthma tend to co-occur together, suggesting a common endotype. Asthma, COPD and mixed asthma/COPD phenotypes have also been explored [148]. Other studies have looked at allergy phenotypes related to degree and pattern of allergic sensitisation (mono- versus poly-sensitised; early versus late-sensitised) [149, 150].

Clustering can be applied to other omic data, outside of phenotype data. In Teo et al [68], hierarchical clustering was used to generate the microbiome profile groups (MPGs) which categorised the infant nasopharyngeal microbiome into discrete clusters based on microbial abundance. This facilitated simpler analysis and interpretation of what was otherwise complex data.

### 2.3.3 Network analysis

Network analysis is the use of networks to model and investigate systems. Networks are represented by graphs consisting of nodes and edges, where nodes represent entities (e.g. biomolecule) and the edges between nodes indicate relationships between entities (e.g. correlation, transition probability, molecular interaction). Edges can be undirected (symmetrical) or directed (asymmetrical). Many types of network analyses involve use of machine learning to generate a best-fitting network for a given dataset.

Networks are used to discover and visualise how different components in a system relate to each other, whether they be abstract relations or actual molecular interactions. Bayesian network analysis involves probabilistic modelling of a network, where edges

are directed and annotated with a transition probability from one node to another. This technique has been used frequently in asthma research, for instance to identify candidate genes or SNPs associated with a bronchodilator response [151]; to quantify interactions between measured pathophysiological variables related to asthma and allergy [152]; and to describe gene regulatory networks using gene expression and GWAS data [38, 153].

Gene co-expression networks can be generated based on correlation between expression levels of different genes. High correlation reflects genes that are co-expressed and hence may be co-regulated or share a common biochemical pathway. Nodes represent genes, while edges represent correlation between them. Furthermore, edges can be weighted by degree of correlation, as in weighted gene co-expression network analysis (WGCNA); and highly-connected or proximate subgraphs can be interpreted as gene modules of functional importance. WGCNA has been used to identify co-expression networks underlying helper T cell responses to house dust mite stimulation [154]; transcription networks in whole blood of asthmatics [155]; an IgE-signalling gene network associated with blood lipids [133]; and co-methylation models that reflect asthma endotypes [60].

Other applications of network analysis exist. For example, Pillai et al. used bipartite network analysis of cytokine expression to sort patients into distinct endotypes [156]. Hinks et al. constructed a network of asthmatic individuals based on similarity in clinico-physiological parameters, then used topological data analysis (TDA) to assign nodes into clusters [152].

Finally, the term “network analysis” has also been used to describe the application of genomic, transcriptomic or proteomic data to existing networks stored in databases, specifically protein-protein interaction (PPI) networks, or networks representing biomolecular pathways. This is often done to generate subsets of the original interaction networks, which are then examined for biological interpretation [157, 158]. Network databases concerning other omics may be used to achieve a similar purpose (e.g. Ingenuity Pathway Analysis, innateDB) [154, 159].

### 2.3.4 Mathematical modelling and prediction

The ultimate goal of integrative analyses is to generate models that reliably explain biological phenomena. At the simplest level, one can use identified omic associations as biomarkers; generate a model consisting of the strongest biomarkers; and test the model on an external dataset. Many examples of such an approach exist in the literature [104, 160, 161]. At a deeper level, one can aggregate multiple biomarkers (potentially omic-wide) into a risk score, such as a genomic risk score [162]. Finally, the classification and network models discussed previously are themselves mathematical models that cover multiple omic domains and are testable on external datasets. In some of these applications, the models represent abstract attributions of risk, and strive to be useful as clinical predictive tools, rather than to be accurate or comprehensive representations of pathophysiology.

Mathematical modelling can generate *in silico* models to describe a complete biological subsystem in terms of components, interactions and functions; and then describe their perturbation during allergy or disease. Modelling biological systems in such a manner is challenging, as there are still many unknowns about its components. However, this has not stopped researchers from trying: for example, Hofer et al. modelled the IL4-dependent activation of GATA3 transcription in Th2 development [163]. Multi-scale approaches have been used to describe multiple levels of biological function, from intracellular molecular processes, to cell-to-cell communication, to organ-level function. For instance, Lauzon et al. formulated a model of airway hyperresponsiveness that accounted for actin-myosin mechanics; calcium signalling in airway smooth muscle (ASM) regulation; mechanical

forces of airway narrowing; and time-dependent distribution of ASM contraction throughout the lung [164]. Such approaches require knowledge of techniques that use differential equations and state diagrams; a review of these approaches is provided elsewhere [165]. However, since such models are usually generated with data from in vitro systems or animal models, it remains an ongoing challenge to test their relevance to in vivo human systems, and they should therefore be treated with caution [166]. Upcoming projects such as the Human Cell Atlas [167] seek to address some of these challenges, bridging the gap between cell biology and clinical medicine.

## 2.4 Pitfalls and challenges

Many challenges remain for systems biology. There are methodological challenges associated with statistical power, even in large consortia. This is due to the sheer scale of omic data, and the number of possible omic-omic comparisons or interactions. Next-generation technologies are becoming cheaper and more efficient, but the amount of data they generate will continue to pose a statistical challenge. Furthermore, the theory behind statistical and modelling methods still lags behind, and there is currently little consensus on the optimal systems-level pipelines (e.g. RNA-seq). Although research groups have recently been paying more attention to measurable environmental exposures in terms of their impact on biological systems [78], the lack of environmental data – and the uncertainty about which exposures actually matter – hinders examination of gene-environment interactions. Finally, even if we have a sufficiently powered sample, there is a so-called “Faustian bargain” [168], where large sample sizes introduce heterogeneity in cases and controls, thus obscuring findings. There is also a similar problem of the “Winner’s Curse” [169], where significant results in a one-wide study tend to exhibit larger effect sizes than what they are in reality.

Machine learning has been the go-to tool to handle phenotypic heterogeneity [170]. However, many biologists and clinicians remain sceptical of it, with concerns about its “hype” or fad-like status, its opaque “black box” nature, and the perceived lack of clear, consistent or immediately-applicable results [171]. Moore et al [145] were one of the first groups to apply unsupervised cluster analysis to an adult asthma cohort, and identify distinct clusters. While the clusters themselves proved useful in describing disease risk and severity profiles in the discovery population, subsequent studies attempting to replicate these clusters in other cohorts have had mixed results [172]. Numerous other studies have identified different sets of clusters based on different parameters and populations [173]. Results of machine learning methods may vary significantly depending on the nature of the input data, in terms of its quality; its relevance to the disease being studied; its depth (resolution of data – categorical vs. continuous) and breadth (single vs. multiple biological domains); and its balance (one domain prioritised over another vs. all treated equally). The variability in research outcomes may suggest to some that machine learning methods are ultimately unreliable; but the field is still growing, and we argue that it simply illustrates the immense complexity of biomedical systems – complexity that will remain impenetrable if we limit ourselves to traditional expert-driven approaches. Unsupervised machine learning can serve as a springboard for future hypotheses: recently, Lazic et al [174] used unsupervised latent class analysis to identify a high-risk multiple-sensitised subgroup, whose pathophysiological origins in early life may be worth exploring in further detail with hypothesis-driven approaches. Ultimately, a balance of human expertise and machine learning will be necessary to make the right decisions about data input and interpretation, and to transform big data into biomedically-relevant results.

Systems biology is multidisciplinary, and with this comes another challenge: communication and collaboration between the various disciplines. There is often a conflict of priorities: a clinician might be more interested in diagnosis, treatment and prognosis; an immunologist in the pathophysiology of allergy and asthma; the biostatistician in making sure that the statistics and modelling are sound; and the bioinformatician in generating clean data and writing problem-free code. There may be residual scepticism amongst some biologists or clinicians who perceive systems approaches as “data fishing” [7]. There is some evidence to suggest that multidisciplinary research projects have greater difficulty in getting funded or making a strong scientific impact [175], and this may reflect the challenge of balancing multiple priorities and conveying different perspectives to a broad audience, more so than the actual quality of the writing or research.

Multiple reviews have highlighted the ongoing inaccessibility of systems approaches to many biologists and clinicians, and have recommended the creation of “biologist-friendly” tools [114, 165, 176]. While this may indeed be helpful for common or simple analyses, there remains an ongoing need for specialist input in developing and using new tools. Tools are only useful if applied correctly, and a research group should not eschew specialist statistics or informatics input, simply to save costs or “to keep things simple”. Also, as one can clearly observe, systems biology is itself very diverse, covering multiple avenues of inquiry. Subspecialties will likely emerge within the field, each focussing on specific methodologies and their applications. It is likely that there will be a demand for specialists and generalists alike, and the movement of tertiary institutions towards incorporating mathematics, statistics, and informatics in undergraduate biomedical courses is certainly a welcome one.

## 2.5 Future directions and concluding statements

The recent developments in systems biology exemplify the global drive towards systems medicine [150, 177], and more broadly, “P4 medicine” – predictive, preventative, personalised and participatory [178]. Our ultimate objective is to achieve a critical level of biomedical understanding that permits development of precise and personalised interventions for individual patients. Worldwide, there has been a push by many groups to implement systems medicine, charting a path from wet lab to dry lab to bedside. Large consortia, such as MeDALL in Europe [150] and STELAR from the UK [170], have been established specifically to record and integrate multiomic data related to allergy and asthma, and conduct well-powered systems-based analyses. Other smaller groups are also involved in similar research via frequent cross-collaborations: these include CAS (Australia) [68], U-BIOPRED (European) [30], COAST (US) [76], COPSAC (Danish) [11], MAAS (UK) [149], SARP (US) [146] and others. In the modern age of systems biology, collaboration and data sharing is virtually mandatory when it comes to uncovering complex associations such as gene-environmental interactions.

Overall, systems biology has yielded fruitful outcomes in allergy research, and promises to deliver more in the future. At the moment, we are still a far way off from truly personalised medicine – being able to predict with reasonable accuracy the disease or prognosis of an individual based on well-sampled data. However, we can only expect the field to grow exponentially in the years to come.

## References

1. Gern JE. The Urban Environment and Childhood Asthma study. *J Allergy Clin Immunol* 2010;125:545–9.

2. Bousquet J, Anto J, Auffray C, et al. MeDALL (Mechanisms of the Development of ALLergy): an integrated approach from phenotypes to systems medicine. *Allergy* 2011;66:596–604.
3. Eberhardt M, Lai X, Tomar N, et al. Third-Kind Encounters in Biomedicine: Immunology Meets Mathematics and Informatics to Become Quantitative and Predictive. *Methods Mol Biol* 2016;1386:135–79.
4. Kirschner MW. The meaning of systems biology. *Cell* 2005;121:503–4.
5. Arazi A, Pendergraft WF, Ribeiro RM, Perelson AS, and Hacohen N. Human systems immunology: hypothesis-based modeling and unbiased data-driven approaches. *Semin Immunol* 2013;25:193–200.
6. Dada JO and Mendes P. Multi-scale modelling and simulation in systems biology. *Integr Biol (Camb)* 2011;3:86–96.
7. Bunyavanich S and Schadt EE. Systems biology of asthma and allergic diseases: a multiscale approach. *J Allergy Clin Immunol* 2015;135:31–42.
8. Gupta J, Johansson E, Bernstein JA, et al. Resolving the etiology of atopic disorders by using genetic analysis of racial ancestry. *J Allergy Clin Immunol* 2016;138:676–99.
9. Ober C and Yao TC. The genetics of asthma and allergic disease: a 21st century perspective. *Immunol Rev* 2011;242:10–30.
10. Ortiz RA and Barnes KC. Genetics of allergic diseases. *Immunol Allergy Clin North Am* 2015;35:19–44.
11. Bonnelykke K, Sparks R, Waage J, and Milner JD. Genetics of allergy and allergic sensitization: common variants, rare mutations. *Curr Opin Immunol* 2015;36:115–26.
12. Portelli MA, Hodge E, and Sayers I. Genetic risk factors for the development of allergic disease identified by genome-wide association. *Clin Exp Allergy* 2015;45:21–31.
13. Moffatt MF, Gut IG, Demenais F, et al. A large-scale, consortium-based genomewide association study of asthma. *N Engl J Med* 2010;363:1211–21.
14. Torgerson DG, Ampleford EJ, Chiu GY, et al. Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nat Genet* 2011;43:887–92.
15. Bonnelykke K, Sleiman P, Nielsen K, et al. A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. *Nat Genet* 2014;46:51–5.
16. Himes BE, Jiang X, Hu R, et al. Genome-wide association analysis in asthma subjects identifies SPATS2L as a novel bronchodilator response gene. *PLoS Genet* 2012;8:e1002824.
17. Farzan N, Vijverberg SJ, Arets HG, Raaijmakers JA, and Maitland-van der Zee AH. Pharmacogenomics of inhaled corticosteroids and leukotriene modifiers: a systematic review. *Clin Exp Allergy* 2017;47:271–293.
18. Imboden M, Bouzigon E, Curjuric I, et al. Genome-wide association study of lung function decline in adults with and without asthma. *J Allergy Clin Immunol* 2012;129:1218–28.
19. Yao TC, Du G, Han L, et al. Genome-wide association study of lung function phenotypes in a founder population. *J Allergy Clin Immunol* 2014;133:248-55 e1-10.

20. Larkin EK and Hartert TV. Genes associated with RSV lower respiratory tract infection and asthma: the application of genetic epidemiological methods to understand causality. *Future Virol* 2015;10:883–897.
21. Pasanen A, Karjalainen MK, Bont L, et al. Genome-Wide Association Study of Polymorphisms Predisposing to Bronchiolitis. *Sci Rep* 2017;7:41653.
22. Abraham G, Havulinna AS, Bhalala OG, et al. Genomic prediction of coronary heart disease. *Eur Heart J* 2016;37:3267–3278.
23. Ober C. Asthma Genetics in the Post-GWAS Era. *Ann Am Thorac Soc* 2016;13 Suppl 1:S85–90.
24. Fernandez TD, Mayorga C, Gueant JL, Blanca M, and Cornejo-Garcia JA. Contributions of pharmacogenetics and transcriptomics to the understanding of the hypersensitivity drug reactions. *Allergy* 2014;69:150–8.
25. Baines KJ, Simpson JL, Wood LG, et al. Sputum gene expression signature of 6 biomarkers discriminates asthma inflammatory phenotypes. *J Allergy Clin Immunol* 2014;133:997–1007.
26. Ewald DA, Malajian D, Krueger JG, et al. Meta-analysis derived atopic dermatitis (MADAD) transcriptome defines a robust AD signature highlighting the involvement of atherosclerosis and lipid metabolism pathways. *BMC Med Genomics* 2015;8:60.
27. Ghosh D, Ding L, Sivaprasad U, et al. Multiple Transcriptome Data Analysis Reveals Biologically Relevant Atopic Dermatitis Signature Genes and Pathways. *PLoS One* 2015;10:e0144316.
28. Himes BE, Koziol-White C, Johnson M, et al. Vitamin D Modulates Expression of the Airway Smooth Muscle Transcriptome in Fatal Asthma. *PLoS One* 2015;10:e0134057.
29. Suarez-Farinas M, Ungar B, Correa da Rosa J, et al. RNA sequencing atopic dermatitis transcriptome profiling provides insights into novel disease mechanisms with potential therapeutic implications. *J Allergy Clin Immunol* 2015;135:1218–27.
30. Bigler J, Boedigheimer M, Schofield JP, et al. A Severe Asthma Disease Signature from Gene Expression Profiling of Peripheral Blood from U-BIOPRED Cohorts. *Am J Respir Crit Care Med* 2016;195:1311–20.
31. Poole A, Urbanek C, Eng C, et al. Dissecting childhood asthma with nasal transcriptomics distinguishes subphenotypes of disease. *J Allergy Clin Immunol* 2014;133:670–8 e12.
32. Esaki H, Ewald DA, Ungar B, et al. Identification of novel immune and barrier genes in atopic dermatitis by means of laser capture microdissection. *J Allergy Clin Immunol* 2015;135:153–63.
33. Raedler D, Ballenberger N, Klucker E, et al. Identification of novel immune phenotypes for allergic and nonallergic childhood asthma. *J Allergy Clin Immunol* 2015;135:81–91.
34. Barnig C, Alsaleh G, Jung N, et al. Circulating Human Eosinophils Share a Similar Transcriptional Profile in Asthma and Other Hypereosinophilic Disorders. *PLoS One* 2015;10:e0141740.
35. Leaker BR, Malkov VA, Mogg R, et al. The nasal mucosal late allergic reaction to grass pollen involves type 2 inflammation (IL-5 and IL-13), the inflammasome (IL-1beta), and complement. *Mucosal Immunol* 2017;10:408–420.



36. Albert FW and Kruglyak L. The role of regulatory variation in complex traits and disease. *Nat Rev Genet* 2015;16:197–212.
37. Murphy A, Chu JH, Xu M, et al. Mapping of numerous disease-associated expression polymorphisms in primary peripheral blood CD4+ lymphocytes. *Hum Mol Genet* 2010;19:4745–57.
38. Hao K, Bosse Y, Nickle DC, et al. Lung eQTLs to help reveal the molecular underpinnings of asthma. *PLoS Genet* 2012;8:e1003029.
39. Li X, Hastie AT, Hawkins GA, et al. eQTL of bronchial epithelial cells and bronchial alveolar lavage deciphers GWAS-identified asthma genes. *Allergy* 2015;70:1309–18.
40. Gordon ED, Palandra J, Wesolowska-Andersen A, et al. IL1RL1 asthma risk variants regulate airway type 2 inflammation. *JCI Insight* 2016;1:e87871.
41. Begin P and Nadeau KC. Epigenetic regulation of asthma and allergic disease. *Allergy Asthma Clin Immunol* 2014;10:27.
42. DeVries A and Vercelli D. Epigenetics in allergic diseases. *Curr Opin Pediatr* 2015;27:719–23.
43. Harb H and Renz H. Update on epigenetics in allergic disease. *J Allergy Clin Immunol* 2015;135:15–24.
44. Brand S, Kesper DA, Teich R, et al. DNA methylation of TH1/TH2 cytokine genes affects sensitization and progress of experimental asthma. *J Allergy Clin Immunol* 2012;129:1602–10 e6.
45. Runyon RS, Cachola LM, Rajeshuni N, et al. Asthma discordance in twins is linked to epigenetic modifications of T cells. *PLoS One* 2012;7:e48796.
46. Stefanowicz D, Hackett TL, Garmaroudi FS, et al. DNA methylation profiles of airway epithelial cells and PBMCs from healthy, atopic and asthmatic children. *PLoS One* 2012;7:e44213.
47. Luo Y, Zhou B, Zhao M, Tang J, and Lu Q. Promoter demethylation contributes to TSLP overexpression in skin lesions of patients with atopic dermatitis. *Clin Exp Dermatol* 2014;39:48–53.
48. Soto-Ramirez N, Arshad SH, Holloway JW, et al. The interaction of genetic variants and DNA methylation of the interleukin-4 receptor gene increase the risk of asthma at age 18 years. *Clin Epigenetics* 2013;5:1.
49. Morales E, Bustamante M, Vilahur N, et al. DNA hypomethylation at ALOX12 is associated with persistent wheezing in childhood. *Am J Respir Crit Care Med* 2012;185:937–43.
50. Liang L, Willis-Owen SA, Laprise C, et al. An epigenome-wide association study of total serum immunoglobulin E concentration. *Nature* 2015;520:670–4.
51. Ito K, Caramori G, Lim S, et al. Expression and activity of histone deacetylases in human asthmatic airways. *Am J Respir Crit Care Med* 2002;166:392–6.
52. Cosio BG, Mann B, Ito K, et al. Histone acetylase and deacetylase activity in alveolar macrophages and blood monocytes in asthma. *Am J Respir Crit Care Med* 2004;170:141–7.
53. Grausenburger R, Bilic I, Boucheron N, et al. Conditional deletion of histone deacetylase 1 in T cells leads to enhanced airway inflammation and increased Th2 cytokine production. *J Immunol* 2010;185:3489–97.
54. Akimova T, Ge G, Golovina T, et al. Histone/protein deacetylase inhibitors increase suppressive functions of human FOXP3+ Tregs. *Clin Immunol* 2010;136:348–63.

55. Okoye IS, Czieso S, Ktistaki E, et al. Transcriptomics identified a critical role for Th2 cell-intrinsic miR-155 in mediating allergy and antihelminth immunity. *Proc Natl Acad Sci U S A* 2014;111:E3081–90.
56. Salam MT. Asthma epigenetics. *Adv Exp Med Biol* 2014;795:183–99.
57. Moheimani F, Hsu AC, Reid AT, et al. The genetic and epigenetic landscapes of the epithelium in asthma. *Respir Res* 2016;17:119.
58. Von Mutius E. The microbial environment and its influence on asthma prevention in early life. *J Allergy Clin Immunol* 2016;137:680–9.
59. Lockett GA, Huoman J, and Holloway JW. Does allergy begin in utero? *Pediatr Allergy Immunol* 2015;26:394–402.
60. Nicodemus-Johnson J, Myers RA, Sakabe NJ, et al. DNA methylation in lung cells is associated with asthma endotypes and genetic risk. *JCI Insight* 2016;1:e90151.
61. Morin A, Laviolette M, Pastinen T, Boulet LP, and Laprise C. Combining omics data to identify genes associated with allergic rhinitis. *Clin Epigenetics* 2017;9:3.
62. Xu CJ, Bonder MJ, Soderhall C, et al. The emerging landscape of dynamic DNA methylation in early childhood. *BMC Genomics* 2017;18:25.
63. Lynch JP, Sikder MA, Curren BF, et al. The Influence of the Microbiome on Early-Life Severe Viral Lower Respiratory Infections and Asthma-Food for Thought? *Front Immunol* 2017;8:156.
64. Stiemsma LT and Turvey SE. Asthma and the microbiome: defining the critical window in early life. *Allergy Asthma Clin Immunol* 2017;13:3.
65. Fu L, Wang C, and Wang Y. Seafood allergen-induced hypersensitivity at the microbiota-mucosal site: implications for prospective probiotic use in allergic response regulation. *Crit Rev Food Sci Nutr* 2017;58:1512–25.
66. West CE, Renz H, Jenmalm MC, et al. The gut microbiota and inflammatory non-communicable diseases: associations and potentials for gut microbiota therapies. *J Allergy Clin Immunol* 2015;135:3–13, quiz 14.
67. Huang YJ and Boushey HA. The microbiome in asthma. *J Allergy Clin Immunol* 2015;135:25–30.
68. Teo SM, Mok D, Pham K, et al. The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development. *Cell Host Microbe* 2015;17:704–15.
69. Sirisinha S. The potential impact of gut microbiota on your health: Current status and future challenges. *Asian Pac J Allergy Immunol* 2016;34:249–264.
70. Chotirmall SH, Gellatly SL, Budden KF, et al. Microbiomes in respiratory health and disease: An Asia-Pacific perspective. *Respirology* 2017;22:240–250.
71. Kang YB, Cai Y, and Zhang H. Gut microbiota and allergy/asthma: From pathogenesis to new therapeutic strategies. *Allergol Immunopathol (Madr)* 2016;45:305–309.
72. Yang X, Jiang Y, and Wang C. Does IL-17 Respond to the Disordered Lung Microbiome and Contribute to the Neutrophilic Phenotype in Asthma? *Mediators Inflamm* 2016;2016:6470364.
73. Azad MB, Konya T, Guttman DS, et al. Infant gut microbiota and food sensitization: associations in the first year of life. *Clin Exp Allergy* 2015;45:632–43.

74. Blazquez AB and Berin MC. Microbiome and food allergy. *Transl Res* 2017;179:199–203.
75. Holt PG and Sly PD. Viral infections and atopy in asthma pathogenesis: new rationales for asthma prevention and treatment. *Nat Med* 2012;18:726–35.
76. Caliskan M, Bochkov YA, Kreiner-Moller E, et al. Rhinovirus wheezing illness and genetic risk of childhood-onset asthma. *N Engl J Med* 2013;368:1398–407.
77. Siroux V, Agier L, and Slama R. The exposome concept: a challenge and a potential driver for environmental health research. *Eur Respir Rev* 2016;25:124–9.
78. North ML, Brook JR, Lee EY, et al. The Kingston Allergy Birth Cohort: Exploring parentally reported respiratory outcomes through the lens of the exposome. *Ann Allergy Asthma Immunol* 2017;118:465–473.
79. Miyata J and Arita M. Role of omega-3 fatty acids and their metabolites in asthma and allergic diseases. *Allergol Int* 2015;64:27–34.
80. Capozzi F and Bordoni A. Foodomics: a new comprehensive approach to food and nutrition. *Genes Nutr* 2013;8:1–4.
81. Lynch SV, Wood RA, Boushey H, et al. Effects of early-life exposure to allergens and bacteria on recurrent wheeze and atopy in urban children. *J Allergy Clin Immunol* 2014;134:593–601 e12.
82. Marchetti P, Pesce G, Villani S, et al. Pollen concentrations and prevalence of asthma and allergic rhinitis in Italy: Evidence from the GEIRD study. *Sci Total Environ* 2017;584-585:1093–1099.
83. Du Toit G, Roberts G, Sayre PH, et al. Randomized trial of peanut consumption in infants at risk for peanut allergy. *N Engl J Med* 2015;372:803–13.
84. Benede S, Blazquez AB, Chiang D, Tordesillas L, and Berin MC. The rise of food allergy: Environmental factors and emerging treatments. *EBioMedicine* 2016;7:27–34.
85. Shankardass K, Jerrett M, Dell SD, Foty R, and Stieb D. Spatial analysis of exposure to traffic-related air pollution at birth and childhood atopic asthma in Toronto, Ontario. *Health Place* 2015;34:287–95.
86. Clark A and Mach N. Role of Vitamin D in the Hygiene Hypothesis: The Interplay between Vitamin D, Vitamin D Receptors, Gut Microbiota, and Immune Response. *Front Immunol* 2016;7:627.
87. Wjst M. Linking vitamin D, the microbiome and allergy. *Allergy* 2017;72:329–330.
88. Mansbach JM, Hasegawa K, Henke DM, et al. Respiratory syncytial virus and rhinovirus severe bronchiolitis are associated with distinct nasopharyngeal microbiota. *J Allergy Clin Immunol* 2016;137:1909–1913 e4.
89. Man WH, de Steenhuijsen Piters WA, and Bogaert D. The microbiota of the respiratory tract: gatekeeper to respiratory health. *Nat Rev Microbiol* 2017;15:259–270.
90. Rossi R, De Palma A, Benazzi L, Riccio AM, Canonica GW, and Mauri P. Biomarker discovery in asthma and COPD by proteomic approaches. *Proteomics Clin Appl* 2014;8:901–15.
91. Terracciano R, Pelaia G, Preiano M, and Savino R. Asthma and COPD proteomics: current approaches and future directions. *Proteomics Clin Appl* 2015;9:203–20.
92. Teran LM, Montes-Vizuet R, Li X, and Franz T. Respiratory proteomics: from descriptive studies to personalized medicine. *J Proteome Res* 2015;14:38–50.

93. Tomazic PV, Birner-Gruenberger R, Leitner A, Spoerk S, and Lang-Loidolt D. Seasonal proteome changes of nasal mucus reflect perennial inflammatory response and reduced defence mechanisms and plasticity in allergic rhinitis. *J Proteomics* 2016;133:153–60.
94. Ferreira MA. Cytokine expression in allergic inflammation: systematic review of in vivo challenge studies. *Mediators Inflamm* 2003;12:259–67.
95. Di Girolamo F, Muraca M, Mazzina O, Lante I, and Dahdah L. Proteomic applications in food allergy: food allergenomics. *Curr Opin Allergy Clin Immunol* 2015;15:259–66.
96. Chan TF, Ji KM, Yim AK, et al. The draft genome, transcriptome, and microbiome of *Dermatophagoides farinae* reveal a broad spectrum of dust mite allergens. *J Allergy Clin Immunol* 2015;135:539–48.
97. Choopong J, Reamtong O, Sookrung N, et al. Proteome, Allergenome, and Novel Allergens of House Dust Mite, *Dermatophagoides farinae*. *J Proteome Res* 2016;15:422–30.
98. Oseroff C, Christensen LH, Westernberg L, et al. Immunoproteomic analysis of house dust mite antigens reveals distinct classes of dominant T cell antigens according to function and serological reactivity. *Clin Exp Allergy* 2017;47:577–592.
99. Campbell BC, Gilding EK, Timbrell V, et al. Total transcriptome, proteome, and allergome of Johnson grass pollen, which is important for allergic rhinitis in subtropical regions. *J Allergy Clin Immunol* 2015;135:133–42.
100. Ghosh N, Sircar G, Saha B, Pandey N, and Gupta Bhattacharya S. Search for Allergens from the Pollen Proteome of Sunflower (*Helianthus annuus* L.): A Major Sensitizer for Respiratory Allergy Patients. *PLoS One* 2015;10:e0138992.
101. Tiotiu A, Brazdova A, Longe C, et al. *Urtica dioica* pollen allergy: Clinical, biological, and allergomics analysis. *Ann Allergy Asthma Immunol* 2016;117:527–534.
102. Scrivo R, Casadei L, Valerio M, Priori R, Valesini G, and Manetti C. Metabolomics approach in allergic and rheumatic diseases. *Curr Allergy Asthma Rep* 2014;14:445.
103. Kelly RS, Dahlin A, McGeachie MJ, et al. Asthma Metabolomics and the Potential for Integrative Omics in Research and the Clinic. *Chest* 2017;151:262–277.
104. Villasenor A, Rosace D, Obeso D, et al. Allergic asthma: an overview of metabolomic strategies leading to the identification of biomarkers in the field. *Clin Exp Allergy* 2017;47:442–456.
105. Kasuga K, Suga T, and Mano N. Bioanalytical insights into mediator lipidomics. *J Pharm Biomed Anal* 2015;113:151–62.
106. Kunisawa J and Kiyono H. Sphingolipids and Epoxidized Lipid Metabolites in the Control of Gut Immunosurveillance and Allergy. *Front Nutr* 2016;3:3.
107. Pendergrass SA, Brown-Gentry K, Dudek S, et al. Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS Genet* 2013;9:e1003087.
108. Karaca S, Civelek E, Karaca M, et al. Allergy-specific Phenome-Wide Association Study for Immunogenes in Turkish Children. *Sci Rep* 2016;6:33152.
109. Vigh-Conrad KA, Conrad DF, and Preuss D. A protein allergen microarray detects specific IgE to pollen surface, cytoplasmic, and commercial allergen extracts. *PLoS One* 2010;5:e10174.

110. Mari A, Scala E, and Alessandri C. The IgE-microarray testing in atopic dermatitis: a suitable modern tool for the immunological and clinical phenotyping of the disease. *Curr Opin Allergy Clin Immunol* 2011;11:438–44.
111. Pomponi D, Bernardi ML, Liso M, et al. Allergen micro-bead array for IgE detection: a feasibility study using allergenic molecules tested on a flexible multiplex flow cytometric immunoassay. *PLoS One* 2012;7:e35697.
112. Hosoki K, Ying S, Corrigan C, et al. Analysis of a Panel of 48 Cytokines in BAL Fluids Specifically Identifies IL-8 Levels as the Only Cytokine that Distinguishes Controlled Asthma from Uncontrolled Asthma, and Correlates Inversely with FEV1. *PLoS One* 2015;10:e0126035.
113. Blanchard C, Stucke EM, Rodriguez-Jimenez B, et al. A striking local esophageal cytokine expression profile in eosinophilic esophagitis. *J Allergy Clin Immunol* 2011;127:208–17, 208–17.
114. Kidd BA, Peters LA, Schadt EE, and Dudley JT. Unifying immunology with informatics and multiscale biology. *Nat Immunol* 2014;15:118–27.
115. Biancotto A and McCoy JP. Studying the human immunome: the complexity of comprehensive leukocyte immunophenotyping. *Curr Top Microbiol Immunol* 2014;377:23–60.
116. Olnes MJ, Kotliarov Y, Biancotto A, et al. Effects of Systemically Administered Hydrocortisone on the Human Immunome. *Sci Rep* 2016;6:23002.
117. Yao Y, Welp T, Liu Q, et al. Multiparameter Single Cell Profiling of Airway Inflammatory Cells. *Cytometry B Clin Cytom* 2017;92:12–20.
118. Minelli C, van der Plaats DA, Leynaert B, et al. Age at puberty and risk of asthma: A Mendelian randomisation study. *PLoS Med* 2018;15:e1002634.
119. Porcu E, Rüeger S, Santoni FA, Reymond A, and Kutalik Z. Mendelian Randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *bioRxiv* 2018:377267.
120. Gref A, Kebede Merid S, Gruziova O, et al. Genome-wide Interaction Analysis of Air Pollution Exposure and Childhood Asthma with Functional Follow-up. *Am J Respir Crit Care Med* 2016;195:1373–83.
121. Vonk JM, Scholtens S, Postma DS, et al. Adult onset asthma and interaction between genes and active tobacco smoking: The GABRIEL consortium. *PLoS One* 2017;12:e0172716.
122. Sordillo JE, Kelly R, Bunyavanich S, et al. Genome-wide expression profiles identify potential targets for gene-environment interactions in asthma severity. *J Allergy Clin Immunol* 2015;136:885–92 e2.
123. Li L, Kabesch M, Bouzigon E, et al. Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma. *Front Genet* 2013;4:103.
124. Wagener AH, Zwinderman AH, Luiten S, et al. dsRNA-induced changes in gene expression profiles of primary nasal and bronchial epithelial cells from patients with asthma, rhinitis and controls. *Respir Res* 2014;15:9.
125. Wesolowska-Andersen A, Everman JL, Davidson R, et al. Dual RNA-seq reveals viral infections in asthmatic children without respiratory illness which are associated with changes in the airway transcriptome. *Genome Biol* 2017;18:12.

126. Zhao F, Durner J, Winkler JB, et al. Pollen of common ragweed (*Ambrosia artemisiifolia* L.): Illumina-based de novo sequencing and differential transcript expression upon elevated NO<sub>2</sub>/O<sub>3</sub>. *Environ Pollut* 2017;224:503–514.
127. Jahreis S, Trump S, Bauer M, et al. Maternal phthalate exposure promotes allergic airway inflammation over 2 generations through epigenetic modifications. *J Allergy Clin Immunol* 741-53 2017;141.
128. Nath AP, Ritchie SC, Byars SG, et al. An interaction map of circulating metabolites, immune gene networks and their genetic regulation. *bioRxiv* 2016.
129. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;25:25–9.
130. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Res* 2000;28:235–42.
131. Apweiler R, Bairoch A, Wu CH, et al. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* 2004;32:D115–9.
132. Inouye M, Kettunen J, Soininen P, et al. Metabonomic, transcriptomic, and genomic variation of a population cohort. *Mol Syst Biol* 2010;6:441.
133. Inouye M, Silander K, Hamalainen E, et al. An immune response network associated with blood lipid levels. *PLoS Genet* 2010;6:e1001113.
134. Sajda P. Machine learning for detection and diagnosis of disease. *Annu Rev Biomed Eng* 2006;8:537–65.
135. Tarca AL, Carey VJ, Chen XW, Romero R, and Draghici S. Machine learning and its applications to biology. *PLoS Comput Biol* 2007;3:e116.
136. Sommer C and Gerlich DW. Machine learning in cell biology - teaching computers to recognize phenotypes. *J Cell Sci* 2013;126:5529–39.
137. Tan PN, Kumar V, and Steinbach M. Introduction to data mining. Boston: Pearson Addison Wesley, 2005.
138. Clark NR and Ma'ayan A. Introduction to statistical methods to analyze large data sets: principal components analysis. *Sci Signal* 2011;4:tr3.
139. Everitt B. Cluster analysis. 3rd. London: Arnold, 1993.
140. Wenzel S. Severe asthma: from characteristics to phenotypes to endotypes. *Clin Exp Allergy* 2012;42:650–8.
141. Hastie AT, Moore WC, Meyers DA, et al. Analyses of asthma severity phenotypes and inflammatory proteins in subjects stratified by sputum granulocytes. *J Allergy Clin Immunol* 2010;125:1028–1036 e13.
142. Spycher BD, Silverman M, and Kuehni CE. Phenotypes of childhood asthma: are they real? *Clin Exp Allergy* 2010;40:1130–41.
143. Bisgaard H and Bonnelykke K. Long-term studies of the natural history of asthma in childhood. *J Allergy Clin Immunol* 2010;126:187–97, 187–97.
144. Lodge CJ, Zaloumis S, Lowe AJ, et al. Early-life risk factors for childhood wheeze phenotypes in a high-risk birth cohort. *J Pediatr* 2014;164:289-94 e1-2.
145. Moore WC, Meyers DA, Wenzel SE, et al. Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. *Am J Respir Crit Care Med* 2010;181:315–23.

146. Moore WC, Hastie AT, Li X, et al. Sputum neutrophil counts are associated with more severe asthma phenotypes using cluster analysis. *J Allergy Clin Immunol* 2014;133:1557–63 e5.
147. Chung KF and Adcock IM. Clinical phenotypes of asthma should link up with disease mechanisms. *Curr Opin Allergy Clin Immunol* 2015;15:56–62.
148. Weatherall M, Travers J, Shirtcliffe PM, et al. Distinct clinical phenotypes of airways disease defined by cluster analysis. *Eur Respir J* 2009;34:812–8.
149. Simpson A, Tan VY, Winn J, et al. Beyond atopy: multiple patterns of sensitization in relation to asthma in a birth cohort study. *Am J Respir Crit Care Med* 2010;181:1200–6.
150. Bousquet J, Anto JM, Akdis M, et al. Paving the way of systems biology and precision medicine in allergic diseases: the MeDALL success story: Mechanisms of the Development of ALLergy; EU FP7-CP-IP; Project No: 261357; 2010-2015. *Allergy* 2016;71:1513–1525.
151. Himes BE, Hunninghake GM, Baurley JW, et al. Genome-wide association analysis identifies PDE4D as an asthma-susceptibility gene. *Am J Hum Genet* 2009;84:581–93.
152. Hinks TS, Zhou X, Staples KJ, et al. Innate and adaptive T cells in asthmatic patients: Relationship to severity and disease mechanisms. *J Allergy Clin Immunol* 2015;136:323–33.
153. Chu JH, Weiss ST, Carey VJ, and Raby BA. A graphical model approach for inferring large-scale networks integrating gene expression and genetic polymorphism. *BMC Syst Biol* 2009;3:55.
154. Troy NM, Hollams EM, Holt PG, and Bosco A. Differential gene network analysis for the identification of asthma-associated therapeutic targets in allergen-specific T-helper memory responses. *BMC Med Genomics* 2016;9:9.
155. Kim YW, Singh A, Shannon CP, Gauvreau GM, and Tebbutt SJ. Transcriptional networks in whole blood of asthmatics. *Allergy, Asthma & Clinical Immunology* 2014;10:A58.
156. Pillai RR, Divekar R, Brasier A, Bhavnani S, and Calhoun WJ. Strategies for molecular classification of asthma using bipartite network analysis of cytokine expression. *Curr Allergy Asthma Rep* 2012;12:388–95.
157. Perkins JR, Barrionuevo E, Ranea JA, Blanca M, and Cornejo-Garcia JA. Systems biology approaches to enhance our understanding of drug hypersensitivity reactions. *Clin Exp Allergy* 2014;44:1461–72.
158. Dahlin A and Tantisira KG. Integrative systems biology approaches in asthma pharmacogenomics. *Pharmacogenomics* 2012;13:1387–404.
159. Lynn DJ, Winsor GL, Chan C, et al. InnateDB: facilitating systems-level analyses of the mammalian innate immune response. *Mol Syst Biol* 2008;4:218.
160. Sircar G, Saha B, Bhattacharya SG, and Saha S. Allergic asthma biomarkers using systems approaches. *Front Genet* 2014;4:308.
161. Zissler UM, Esser-von Bieren J, Jakwerth CA, Chaker AM, and Schmidt-Weber CB. Current and future biomarkers in allergic asthma. *Allergy* 2016;71:475–94.
162. Belsky DW, Sears MR, Hancox RJ, et al. Polygenic risk and the development and course of asthma: an analysis of data from a four-decade longitudinal study. *Lancet Respir Med* 2013;1:453–61.

163. Hofer T, Nathansen H, Lohning M, Radbruch A, and Heinrich R. GATA-3 transcriptional imprinting in Th2 lymphocytes: a mathematical model. *Proc Natl Acad Sci U S A* 2002;99:9364–8.
164. Lauzon AM, Bates JH, Donovan G, Tawhai M, Sneyd J, and Sanderson MJ. A multi-scale approach to airway hyperresponsiveness: from molecule to organ. *Front Physiol* 2012;3:191.
165. Narang V, Decraene J, Wong SY, et al. Systems immunology: a survey of modeling formalisms, applications and simulation tools. *Immunol Res* 2012;53:251–65.
166. Auffray C, Adcock IM, Chung KF, Djukanovic R, Pison C, and Sterk PJ. An integrative systems biology approach to understanding pulmonary diseases. *Chest* 2010;137:1410–6.
167. Regev A, Teichmann S, Lander ES, et al. The Human Cell Atlas. *bioRxiv* 2017.
168. Vercelli D and Martinez FD. The Faustian bargain of genetic association studies: bigger might not be better, or at least it might not be good enough. *J Allergy Clin Immunol* 2006;117:1303–5.
169. Huang QQ, Ritchie SC, Brozynska M, and Inouye M. Power, false discovery rate and Winner’s Curse in eQTL studies. *Nucleic Acids Research* 2018;46:e133–e133.
170. Belgrave D, Henderson J, Simpson A, Buchan I, Bishop C, and Custovic A. Disaggregating asthma: Big investigation versus big data. *J Allergy Clin Immunol* 2017;139:400–407.
171. Chen JH and Asch SM. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. *N Engl J Med* 2017;376:2507–2509.
172. Kupczyk M, Dahlen B, Sterk PJ, et al. Stability of phenotypes defined by physiological variables and biomarkers in adults with asthma. *Allergy* 2014;69:1198–204.
173. Deliu M, Sperrin M, Belgrave D, and Custovic A. Identification of Asthma Subtypes Using Clustering Methodologies. *Pulmonary Therapy* 2016;2:19–41.
174. Lazic N, Roberts G, Custovic A, et al. Multiple atopy phenotypes and their associations with asthma: similar findings from two birth cohorts. *Allergy* 2013;68:764–70.
175. Bromham L, Dinnage R, and Hua X. Interdisciplinary research has consistently lower funding success. *Nature* 2016;534:684–7.
176. Sparks R, Lau WW, and Tsang JS. Expanding the Immunology Toolbox: Embracing Public-Data Reuse and Crowdsourcing. *Immunity* 2016;45:1191–1204.
177. Sittka A, Vera J, Lai X, and Schmeck BT. Asthma phenotyping, therapy, and prevention: what can we learn from systems biology? *Pediatr Res* 2013;73:543–52.
178. Galli SJ. Toward precision medicine and health: Opportunities and challenges in allergic diseases. *J Allergy Clin Immunol* 2016;137:1289–300.



## Chapter 3

# Trajectories of childhood immune development and respiratory health relevant to asthma and allergy

NB: This chapter has been published in its entirety in eLife on 15 October 2018, available at <https://elifesciences.org/articles/35856>.

### 3.1 Introduction

Asthma is a global health problem, and there is a pressing need for better understanding of its pathogenesis [1]. Asthma is strongly associated with allergy, and both genetic and environmental factors may be involved [2, 3]. The “hygiene hypothesis” proposes that modern changes to hygiene, sanitation and living environment have modified human exposures to microbes, with subsequent effects on early-life immune development [4]. However, the clinical presentation and prognosis of childhood wheeze is highly variable: some children remit; others remit but relapse; and yet others have wheeze persisting into adult asthma [5]. These differences suggest that the underlying causes of disease also differ from person to person. For example, while asthma is commonly linked to allergy, not all individuals with wheeze are sensitised to allergen, and vice versa [6]. As such, childhood asthma is a heterogeneous condition [7, 8], and this greatly complicates the study of its pathogenesis [9]. We postulate that there are subpopulations in early childhood, each sharing similar patterns of pathophysiology, disease susceptibility and phenotype that permit categorisation into clusters. If we can agnostically identify these clusters, then we may explore the biological mechanisms that underlie them, and find targets for early intervention that are specific for different asthma subtypes.

Previous attempts at subtyping asthma susceptibility relied on supervised classification, using expert knowledge and cut-offs to define clusters. For example, criteria such as – specific immunoglobulin E (IgE)  $\geq 0.35$  kU/L; wheal diameter  $\geq 3$  mm in a skin prick test (SPT); or symptom score surpassing a threshold – may determine classification into a high-risk profile [10, 11]. However, these cut-offs vary with age, gender or other parameters, and may not accurately reflect true attribution of risk [12]. Hence, they often continue to produce heterogeneous groups. Furthermore, previous studies tended to focus on a single “domain”, for instance grouping only by immunological response [13], symptomatology or timing of disease [14, 15]. Recently, researchers have turned to unsupervised approaches, such as model-based cluster analysis and latent class analysis (LCA) [16–21]. These do not require experts to supply cut-offs, but can instead “learn” boundaries from the data. They can potentially uncover patterns of similarity not immediately obvious to the human eye. Finally, these methods can cover a broader range of domains, incorporating measurements

from multiple sources to determine clusters that are potentially informative of asthma risk.

Here, we use a data-driven unsupervised framework together with a comprehensively-phenotyped birth cohort, to define developmental trajectories during preschool years, a period known to be critical to asthma pathogenesis. Specifically, we: 1) use non-parametric mixture models to discover latent clusters that define early childhood trajectories of immune function and susceptibility to respiratory infection; 2) investigate how these clusters relate to differential profiles of asthma susceptibility, and to existing definitions of atopy; 3) identify risk factors for asthma within each cluster; and 4) externally validate the clusters in independent cohorts.

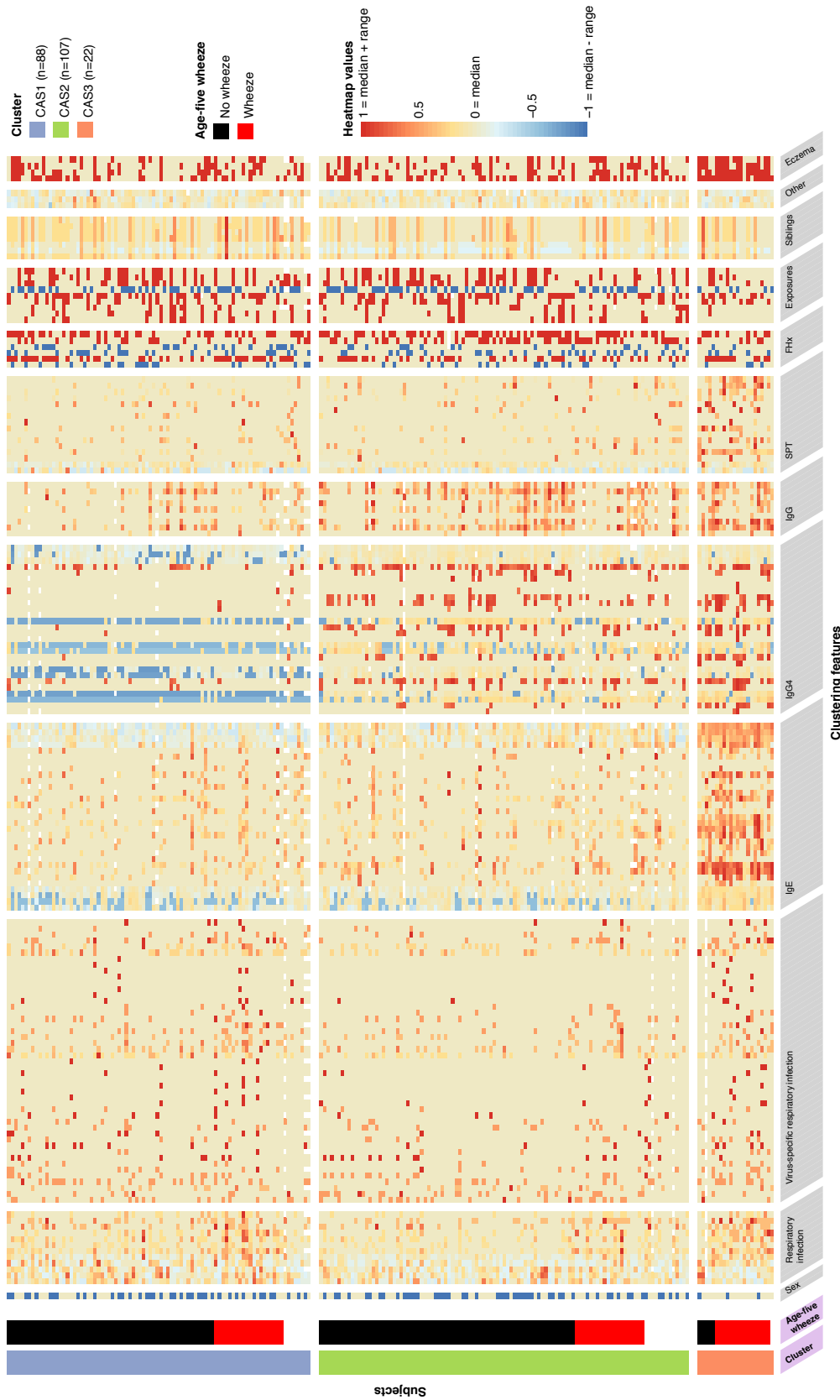
## 3.2 Results

To characterise the broad structure of an Australian dataset of young children (Childhood Asthma Study, CAS), we performed principal components analysis (**Supplementary Figure B.1**). Afterwards, to explicitly model the heterogeneous data types of the cohorts as well as explicitly identify clusters, we used non-parametric expectation-maximisation (npEM) mixture models (**Methods, Section 3.4**). By applying npEM-based clustering and classification to CAS, we identified three distinct clusters from 217 individuals and 174 clustering features (**Figure 3.1**): low-risk CAS1 ( $N=88$ , 25% wheeze at age 5), low-risk but allergy-susceptible CAS2 ( $N=107$ , 21% wheeze at age 5) and high-risk CAS3 ( $N=22$ , 76% wheeze at age 5). Forty-six individuals in CAS had excessive missing data and were not classifiable. The CAS clusters satisfied basic measures of internal stability, and were distinguishable on a PCA plot of the complete-case dataset (**Supplementary Figure B.1**). A graphical summary of results for the CAS clusters is presented in **Figure 3.2**.

### 3.2.1 CAS1: low-risk, non-atopic cluster with transient wheeze

CAS1 was a low-risk cluster with infrequent and transient respiratory wheeze. Rates of wheeze declined from 33% at age 1 to 12% by age 10 (**Table 3.1; Figure 3.3**). In this cluster, Th2 cytokine responses of peripheral blood mononuclear cells (PBMCs) to allergen stimulation were minimal; and rates of allergen sensitisation (as measured by IgE or skin prick test, SPT) were the lowest among all groups (**Table 3.2; Figure 3.4; Supplementary Table B.3B-D**). IgG and IgG4 were also low across all allergens.

Frequency of respiratory infection in CAS1 was low (**Table 3.3**). However, high frequency of lower respiratory infections (LRIs) in childhood, especially wheezy LRIs (wLRIs), was a risk factor for age-five wheeze – even after adjusting for sex, body mass index (BMI) and parental history of asthma as demographic covariates (**Table 3.4**). Repeated-measures ANOVA identified that LRI and wLRI frequency in the first three years were predictors for age-five wheeze (**Supplementary Table B.4**); however, timepoint-specific analyses showed that differences were only noticeable from age 3 onwards (**Table 3.4; Figure 3.4A-B**). A multiple regression model with stepwise elimination yielded three significant variables: age-three wLRI frequency (odds ratio OR 5.6 per unit increase,  $p = 0.0068$ ); age-four LRI frequency (OR 3.6,  $p = 0.018$ ); and a protective effect from proportion of infection-associated microbiome profile groups (MPGs; *Streptococcus*, *Haemophilus*, *Moraxella*) in age-two-to-four healthy nasopharyngeal aspirate samples (NPAs; OR 0.19 per quartile,  $p = 0.014$ ).



**FIGURE 3.1: Non-parametric mixture-model based clustering of CAS dataset, based on 174 features.**

SPT = skin prick test. White spaces within the heatmap indicate missing data. Rows represent individuals; columns represent clustering features with general categories as labelled on grey background. Variables with grey background are clustering features ordered by category or type of variable first (e.g. all HDM IgE-related variables grouped together), then by timepoint (earlier to later, from left to right). Variables with lilac background indicate resultant cluster membership and outcome variable age-five wheeze). Heatmap values are scaled relative to range and median values for each feature; the median is coloured beige-yellow, the median + range red, and median - range blue. For sex, -1/blue = female, 0/ yellow (median) = male.

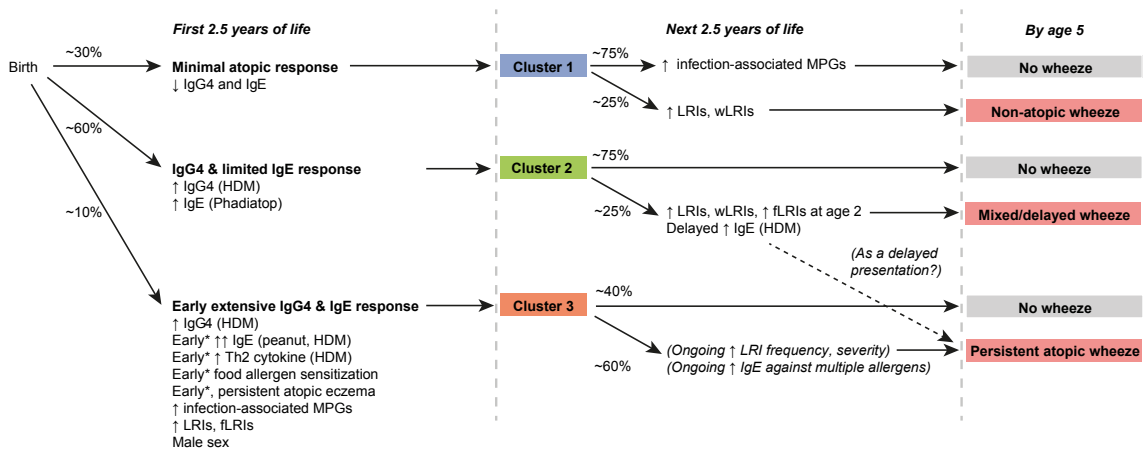


FIGURE 3.2: Graphical summary of proposed clusters.

\*“Early” specifically refers to “within the first 6 months of life”.

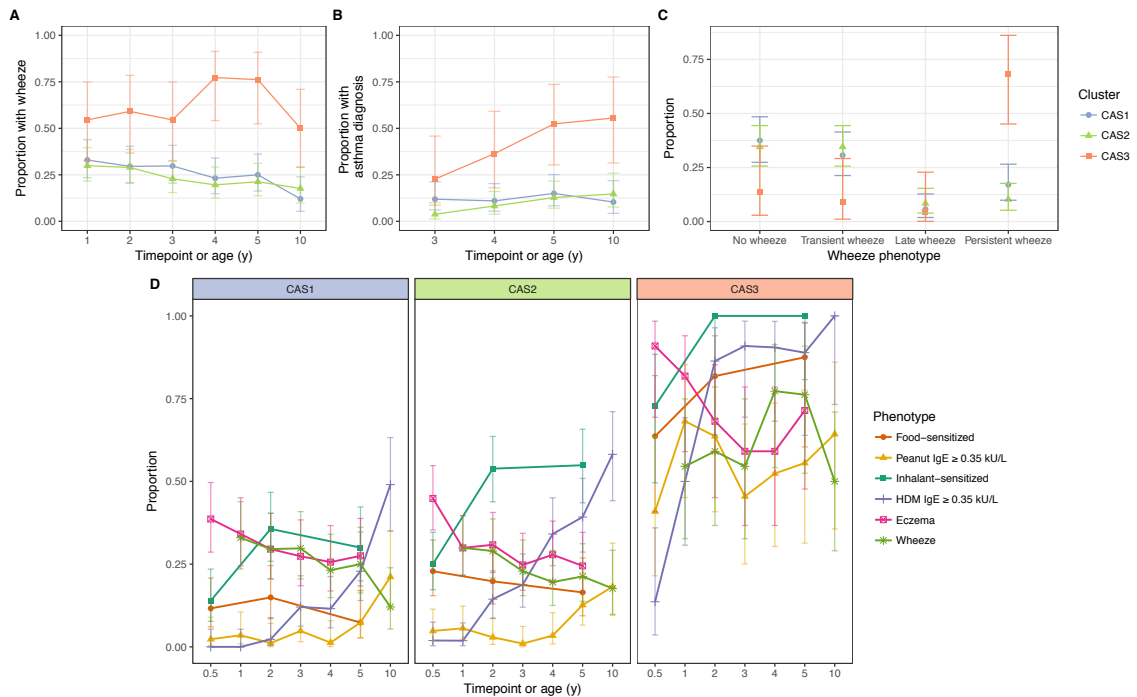


FIGURE 3.3: Incidence of multiple phenotypes, including parent-reported wheeze (A), physician-diagnosed asthma (B), defined wheeze phenotypes (C), in relation to food and inhalant sensitisation (D), stratified by cluster and time in the CAS dataset.

Points indicate observed proportion; bars indicate 95% CI (binomial distribution). Wheeze phenotypes defined as: no wheeze = no wheeze at ages 1 to 3, or age 5; transient wheeze = any wheeze at ages 1 to 3, but not age 5; late wheeze = wheeze at age 5, but not ages 1 to 3; persistent wheeze = any wheeze at both ages 1 to 3 and age 5. Food sensitization defined as peanut IgE  $\geq 0.35$  kU/L at any age, or cow’s milk, egg white, peanut SPT  $> 2$  or  $3$  mm for age  $\leq 2$  or  $> 2$  respectively. Inhalant sensitization defined as HDM, cat, couchgrass, ryegrass, mould or Phadiatop IgE  $\geq 0.35$  kU/L at any age, or mould SPT (*Alternaria* or *Aspergillus* spp.)  $> 2$  or  $3$  mm for age  $\leq 2$  or  $> 2$  respectively.

**TABLE 3.1: Comparison of selected demographic and clinical variables in CAS clusters**

BMI = body mass index; feature? = whether variable was used as a clustering feature or not; geom. mean = geometric mean; prop. = proportion. For categorical variables, associations were tested using Fisher exact test; for continuous variables, Kruskal-Wallis and Mann-Whitney-Wilcoxon. Bold text indicates statistical significance ( $p < 0.05$ ); italics indicate near-significance ( $p < 0.10$ ). \*Not used as clustering feature, as BMI is a derived variable. Height and weight at age 3 were used instead.

Variable	Age (y)	CAS1 (N=88)	CAS2 (N=107)	CAS3 (N=22)	P-value (unadjusted)			Feature?	
		Prop. (95% CI)	Prop. (95% CI)	Prop. (95% CI)	Overall	1 vs. 2	1 vs. 3		2 vs. 3
Sex = male		55% (44%-65%)	51% (42%-61%)	86% (71%-100%)	<b>7.3E-03</b>	0.67	<b>6.8E-03</b>	<b>3.7E-03</b>	Yes
Maternal asthma		51% (40%-62%)	41% (32%-51%)	59% (37%-81%)	0.19	0.19	0.63	0.16	Yes
Paternal asthma		22% (13%-30%)	44% (35%-54%)	23% (3.7%-42%)	<b>2.2E-03</b>	<b>1.3E-03</b>	1	<i>0.093</i>	Yes
Wheeze	1	33% (23%-43%)	30% (21%-39%)	55% (32%-77%)	<i>0.092</i>	0.76	<i>0.084</i>	<b>0.046</b>	No
	5	25% (15%-35%)	21% (13%-30%)	76% (56%-96%)	<b>7.1E-06</b>	0.59	<b>2.6E-05</b>	<b>3.4E-06</b>	No
	10	12% (3.4%-21%)	18% (8.4%-27%)	50% (24%-76%)	<b>3.1E-03</b>	0.46	<b>1.5E-03</b>	<b>0.011</b>	No
Asthma	5	15% (7%-23%)	13% (5.9%-20%)	52% (29%-76%)	<b>4.1E-04</b>	0.83	<b>7.7E-04</b>	<b>2.1E-04</b>	No
	10	10% (2.3%-18%)	15% (6.1%-23%)	56% (30%-81%)	<b>2.6E-04</b>	0.59	<b>1.8E-04</b>	<b>7.9E-04</b>	No
Eczema	6m	39% (28%-49%)	45% (35%-54%)	91% (78%-100%)	<b>2.4E-05</b>	0.47	<b>7.9E-06</b>	<b>9.0E-05</b>	Yes
	1	34% (24%-44%)	30% (21%-39%)	82% (64%-99%)	<b>2.5E-05</b>	0.54	<b>7.2E-05</b>	<b>1.4E-05</b>	Yes
	5	28% (18%-37%)	24% (16%-33%)	71% (50%-92%)	<b>2.1E-04</b>	0.73	<b>3.3E-04</b>	<b>7.9E-05</b>	No
Atopic rhinoconjunctivitis	5	30% (20%-40%)	39% (29%-49%)	76% (56%-96%)	<b>6.4E-04</b>	0.21	<b>2.7E-04</b>	<b>3.2E-03</b>	No
		Mean (95% CI)	Mean (95% CI)	Mean (95% CI)	Overall	1 vs. 2	1 vs. 3	2 vs. 3	
BMI (kg/m <sup>2</sup> )	3	16 (16-17)	16 (16-17)	16 (16-17)	0.86	0.65	0.68	0.8	No*
	4	16 (16-17)	16 (16-16)	17 (16-17)	0.59	0.76	0.32	0.39	No
	5	16 (16-16)	16 (16-16)	16 (15-17)	0.71	0.56	0.48	0.67	No
	10	18 (17-19)	18 (17-18)	18 (17-19)	0.89	0.75	1	0.62	No
Number of older siblings	0	0.93 (0.72-1.1)	0.53 (0.38-0.69)	0.77 (0.32-1.2)	<b>4.5E-03</b>	<b>1.0E-03</b>	0.37	0.25	Yes
	2	0.85 (0.66-1)	0.5 (0.34-0.65)	0.77 (0.32-1.2)	<b>2.8E-03</b>	<b>6.5E-04</b>	0.48	0.16	Yes
	5	0.68 (0.5-0.85)	0.39 (0.25-0.54)	0.67 (0.23-1.1)	<b>0.016</b>	<b>5.1E-03</b>	0.75	0.12	No
		Geom. mean (95% CI)	Geom. mean (95% CI)	Geom. mean (95% CI)	Overall	1 vs. 2	1 vs. 3	2 vs. 3	
Vitamin D (nmol/L)	1	60 (55-64)	59 (55-63)	59 (52-67)	0.93	0.98	0.76	0.7	No
	2	57 (54-61)	58 (55-61)	47 (40-55)	<b>0.012</b>	0.82	<b>5.4E-03</b>	<b>4.4E-03</b>	No
	5	89 (83-95)	84 (79-89)	77 (69-84)	<i>0.057</i>	0.46	<b>0.016</b>	<i>0.056</i>	No

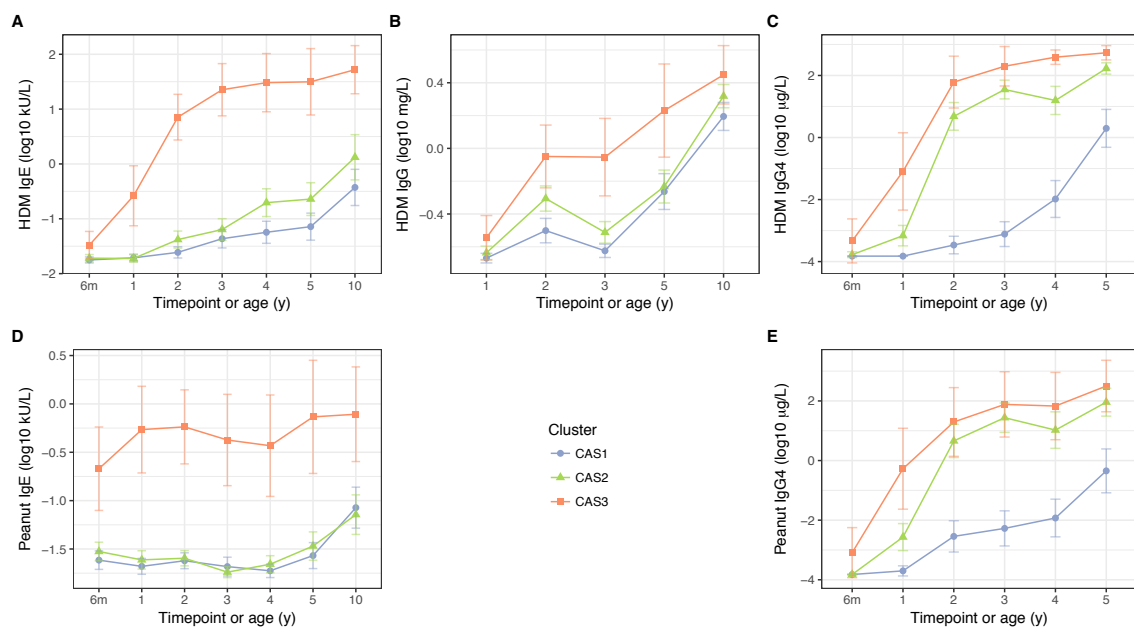


FIGURE 3.4: HDM IgE (A), IgG (B) and IgG4 (C); and peanut IgE (D) and IgG4 (E) stratified by cluster and time, in the CAS dataset

Points indicate means; bars indicate 95% CI (t-distribution).

**TABLE 3.2: Comparison of HDM-associated immunological variables in CAS**

Feature? = whether variable was used as a clustering feature or not; geom. mean = geometric mean; PBMC = peripheral blood mononuclear cells; prop. = proportion; SPT = skin prick or sensitisation test. For categorical variables, associations were tested using Fisher exact test; for continuous variables, Kruskal-Wallis and Mann-Whitney-Wilcoxon. Bold text indicates statistical significance ( $p < 0.05$ ); italics indicate near-significance ( $p < 0.10$ ). ^PBMC cytokine responses to HDM above unstimulated control; birth samples (age 0) taken from cord blood (CBMC). \*Not used as clustering features, as these were derived variables; the variables from which they were derived (HDM IgE and IgG4) were used instead.

Variable	Age	CAS1 (N=88)	CAS2 (N=107)	CAS3 (N=22)	P-value (unadjusted)			Feature?		
		Geom. mean (95% CI)	Geom. mean (95% CI)	Geom. mean (95% CI)	Overall	1 vs. 2	1 vs. 3		2 vs. 3	
<i>Total antibody</i> IgE (kU/L)	6m	1.2 (0.69-2)	2.2 (1.4-3.6)	21 (12-35)	<b>1.2E-07</b>	<b>0.044</b>	6.7E-08	2.2E-06	Yes	
	1	0.6 (0.29-1.3)	2 (1.1-3.7)	43 (17-109)	<b>2.0E-09</b>	<b>0.019</b>	4.3E-09	5.3E-08	Yes	
	2	6.6 (3.5-12)	17 (12-25)	187 (131-267)	<b>1.2E-11</b>	<b>0.044</b>	4.2E-11	1.4E-10	Yes	
	5	35 (23-55)	60 (46-80)	451 (278-731)	<b>2.2E-08</b>	<i>0.096</i>	1.9E-08	1.5E-07	No	
	10	85 (46-154)	150 (103-217)	800 (405- 1.6E+03)	<b>1.4E-04</b>	0.11	1.3E-04	2.8E-04	No	
<i>HDM antibody</i> IgE (kU/L)	6m	0.018 (0.016-0.02)	0.019 (0.016- 0.022)	0.033 (0.019- 0.059)	<b>1.9E-03</b>	0.47	7.9E-04	4.2E-03	Yes	
	1	0.019 (0.017- 0.023)	0.019 (0.016- 0.022)	0.26 (0.075-0.93)	<b>1.3E-09</b>	0.47	2.5E-07	4.5E-09	Yes	
	2	0.024 (0.019- 0.031)	0.042 (0.029-0.06)	7.1 (2.7-19)	<b>2.6E-16</b>	<i>0.078</i>	2.5E-15	3.5E-13	Yes	
	5	0.072 (0.041-0.13)	0.23 (0.12-0.45)	31 (7.8-127)	<b>4.2E-09</b>	<b>0.015</b>	3.8E-09	5.1E-07	No	
	10	0.37 (0.17-0.8)	1.3 (0.51-3.4)	52 (19-144)	<b>2.9E-06</b>	<i>0.068</i>	5.7E-07	9.7E-05	No	
IgG (mg/L)	1	0.21 (0.2-0.23)	0.23 (0.21-0.25)	0.29 (0.21-0.39)	<b>0.042</b>	<b>0.34</b>	<b>0.012</b>	<i>0.07</i>	Yes	
	2	0.32 (0.27-0.37)	0.49 (0.41-0.59)	0.89 (0.57-1.4)	<b>1.9E-06</b>	<b>2.1E-04</b>	3.8E-06	7.0E-03	Yes	
	5	0.55 (0.42-0.7)	0.59 (0.46-0.74)	1.7 (0.88-3.3)	<b>1.5E-03</b>	0.67	6.4E-04	9.0E-04	No	
IgG4 ( $\mu$ g/L)	10	1.6 (1.3-1.9)	2.1 (1.8-2.5)	2.8 (1.9-4.2)	<b>1.0E-02</b>	<b>0.023</b>	<b>0.011</b>	0.18	No	
	6m	1.5E-04 (1.5E-04- 1.5E-04)	1.7E-04 (1.3E-04- 2.1E-04)	4.6E-04 (9.0E-05- 2.4E-03)	<b>4.9E-03</b>	0.37	5.2E-03	<b>0.024</b>	Yes	
	1	1.5E-04 (1.5E-04- 1.5E-04)	6.9E-04 (3.2E-04- 1.5E-03)	0.081 (4.6E- 03-1.4)	<b>1.8E-10</b>	<b>5.2E-04</b>	6.6E-12	2.2E-05	Yes	
	2	3.4E-04 (1.8E-04- 6.6E-04)	4.8 (1.7-13)	61 (8.9-419)	<b>1.8E-25</b>	<b>1.5E-22</b>	8.6E-18	9.8E-05	Yes	
5	2 (0.48-8.1)	168 (111-256)	539 (317-917)	<b>1.1E-15</b>	<b>1.3E-12</b>	1.0E-08	1.9E-04	No		
<i>HDM cytokine response</i> <sup>^</sup>	IL-13 protein (pg/ml) <sup>^</sup>	0	0.22 (0.066-0.73)	0.22 (0.076-0.63)	0.085 (0.011-0.66)	0.68	0.76	0.41	0.45	No
		6m	0.064 (0.022-0.18)	0.06 (0.025-0.14)	19 (1.4-244)	<b>4.6E-06</b>	0.98	1.7E-05	4.1E-06	No
		5	0.13 (0.046-0.37)	0.32 (0.11-0.87)	12 (1.2-117)	<b>2.1E-04</b>	0.29	7.7E-05	5.1E-04	No
	IL-5 protein (pg/ml) <sup>^</sup>	0	0.043 (0.018-0.11)	0.026 (0.013- 0.052)	0.018 (5.0E- 03-0.068)	0.44	0.36	0.29	0.57	No
		6m	0.018 (9.2E- 03-0.034)	0.013 (8.9E- 03-0.02)	0.21 (0.012-3.7)	<b>7.9E-04</b>	0.4	8.1E-03	3.5E-04	No
		5	0.028 (0.014- 0.057)	0.042 (0.02-0.087)	2.3 (0.25-22)	<b>3.2E-06</b>	0.45	5.7E-06	2.0E-05	No
	IL-13 mRNA <sup>^</sup>	0	1.7E-03 (1.1E-04- 0.026)	6.0E-03 (4.8E-04- 0.075)	6.7E-03 (3.3E-05- 1.4)	0.85	0.6	0.68	0.94	No
		6m	1.0E-04 (8.8E-06- 1.1E-03)	3.2E-04 (3.8E-05- 2.6E-03)	2 (0.015-266)	<b>3.2E-04</b>	0.5	1.7E-04	3.8E-04	No
		5	0.036 (1.6E- 03-0.8)	0.11 (8.8E- 03-1.4)	2.9E+03 (742- 1.1E+04)	<b>6.8E-05</b>	0.59	9.9E-05	2.5E-05	No
	IL-4 mRNA <sup>^</sup>	0	1.4E-06 (6.9E-07- 3.0E-06)	1.9E-06 (7.8E-07- 4.4E-06)	1.0E-06 (1.0E-06- 1.0E-06)	0.71	0.65	0.6	0.47	No
6m		4.6E-06 (1.0E-06- 2.1E-05)	5.1E-06 (1.4E-06- 1.8E-05)	0.54 (6.5E-03-44)	<b>6.2E-09</b>	0.94	4.7E-07	1.0E-07	No	

Continued on next page

Continued from previous page

Variable	Age	CAS1 (N=88)	CAS2 (N=107)	CAS3 (N=22)	P-value (unadjusted)			Feature?	
		Geom. mean (95% CI)	Geom. mean (95% CI)	Geom. mean (95% CI)	Overall	1 vs. 2	1 vs. 3		2 vs. 3
IL-5 mRNA <sup>^</sup>	5	2.3E-04 (1.7E-05- 3.0E-03)	4.7E-04 (5.3E-05- 4.3E-03)	5.3 (0.082-345)	<b>4.9E-04</b>	0.72	<b>4.5E-04</b>	<b>3.2E-04</b>	No
	0	2.5E-04 (2.1E-05- 2.9E-03)	2.6E-04 (2.8E-05- 2.5E-03)	1.2E-05 (3.1E-07- 4.6E-04)	0.47	0.96	0.24	0.25	No
	6m	5.2E-05 (5.6E-06- 4.8E-04)	3.1E-05 (5.2E-06- 1.8E-04)	0.33 (1.3E-03-83)	<b>1.5E-04</b>	0.85	<b>2.3E-04</b>	<b>1.1E-04</b>	No
	5	0.021 (9.9E- 04-0.43)	0.07 (5.7E- 03-0.85)	246 (7-8.7E+03)	<b>1.3E-04</b>	0.49	<b>7.1E-05</b>	<b>1.1E-04</b>	No
		Prop. (95% CI)	Prop. (95% CI)	Prop. (95% CI)	Overall	1 vs. 2	1 vs. 3	2 vs. 3	
<i>HDM SPT past atopy threshold</i>									
Wheal $\geq$ 2mm	6m	2.3% (0%-5.4%)	1.9% (0%-4.5%)	14% (0%-29%)	<b>0.043</b>	1	<i>0.054</i>	<b>0.035</b>	No*
	2	10% (3.8%-17%)	15% (8.1%-22%)	86% (71%-100%)	<b>2.9E-12</b>	0.39	<b>8.2E-12</b>	<b>1.5E-10</b>	No*
Wheal $\geq$ 3mm	5	13% (5.2%-20%)	28% (18%-37%)	81% (63%-99%)	<b>1.5E-08</b>	<b>0.022</b>	<b>4.6E-09</b>	<b>1.0E-05</b>	No
	10	36% (23%-49%)	51% (38%-63%)	78% (57%-99%)	<b>7.4E-03</b>	0.11	<b>2.7E-03</b>	<i>0.06</i>	No

### 3.2.2 CAS2: low-risk cluster susceptible to atopic and non-atopic wheeze

Similar to CAS1, CAS2 was a low-risk cluster with infrequent allergic disease. Compared to CAS1, Phadiatop and house dust mite (HDM) IgE were elevated at most timepoints (**Table 3.2; Figure 3.4A; Supplementary Table B.3B**), with the exception of peanut IgE (Wilcoxon, adjusted  $p = 0.99$  at all timepoints; **Figure 3.4D**). CAS2 IgG and IgG4 were intermediate between CAS1 and CAS3 levels; CAS2 IgG was closer to CAS1, while CAS2 IgG4 was closer to CAS3 (**Table 3.2; Figure 3.4**). Despite these antibody differences, yearly rates of wheeze in CAS2 remained comparable to CAS1 (30% at age 1, declining to 18% at age 10; **Table 3.1; Figure 3.3**). Interestingly, compared to CAS1, individuals in CAS2 had fewer older siblings living in the household at age 2, as well as more frequent paternal history of asthma (adjusted  $p = 0.029$  and  $0.055$ , respectively; **Supplementary Table B.3A**).

Predictive factors for age-five wheeze in CAS2 included: LRI, wLRI and febrile LRI (fLRI) frequency (GLM;  $p = 2.7 \times 10^{-3}$ ,  $0.016$  and  $0.02$  at age 3, respectively); HDM IgE ( $p = 0.016$  and  $0.011$  at ages 2 and 4, respectively); and Phadiatop IgE ( $p = 0.01$  at age 4) (**Table 3.4**). Repeated-measures ANOVA showed that HDM IgE and LRI-related variables (LRI, wLRI, fLRI) from the first 3 years were significant predictors of age-five wheeze (**Supplementary Table B.4**). Timepoint-specific analyses showed that differences were observable in HDM IgE and fLRI from age 2 onwards, while in LRI and wLRI they were only noticeable from age 3 (**Table 3.4; Figure 3.5**). A multiple regression model with stepwise elimination identified three significant variables: age-two fLRI (OR 8 per unit increase,  $p = 0.0075$ ), age-four wLRI (OR 5.3,  $p = 0.0016$ ), and age-four Phadiatop IgE (OR 3.3,  $p = 0.0088$ ). But although both IgE-related and infection-related risk factors contributed to age-five wheeze, there was no significant evidence of interaction between them ( $p = 0.36$  within CAS2 alone,  $p = 0.92$  across entire cohort, for age-four wLRI frequency  $\times$  Phadiatop IgE). Overall, CAS2 represented a low-risk trajectory susceptible to, but not necessarily afflicted by, wheeze due to atopic and non-atopic risk factors. In this cluster, atopic determinants of age-five wheeze were only active from age two onwards, suggesting delayed atopic wheeze in this cluster. This duality of atopic and non-atopic risk factors for wheeze in this cluster was further supported by decision tree analysis, which identified that wheezy LRI frequency and HDM IgE best separated wheezers from non-wheezers in CAS2 (**Supplementary Figure B.9**).



**TABLE 3.3: Comparison of selected respiratory disease-related variables in CAS clusters**

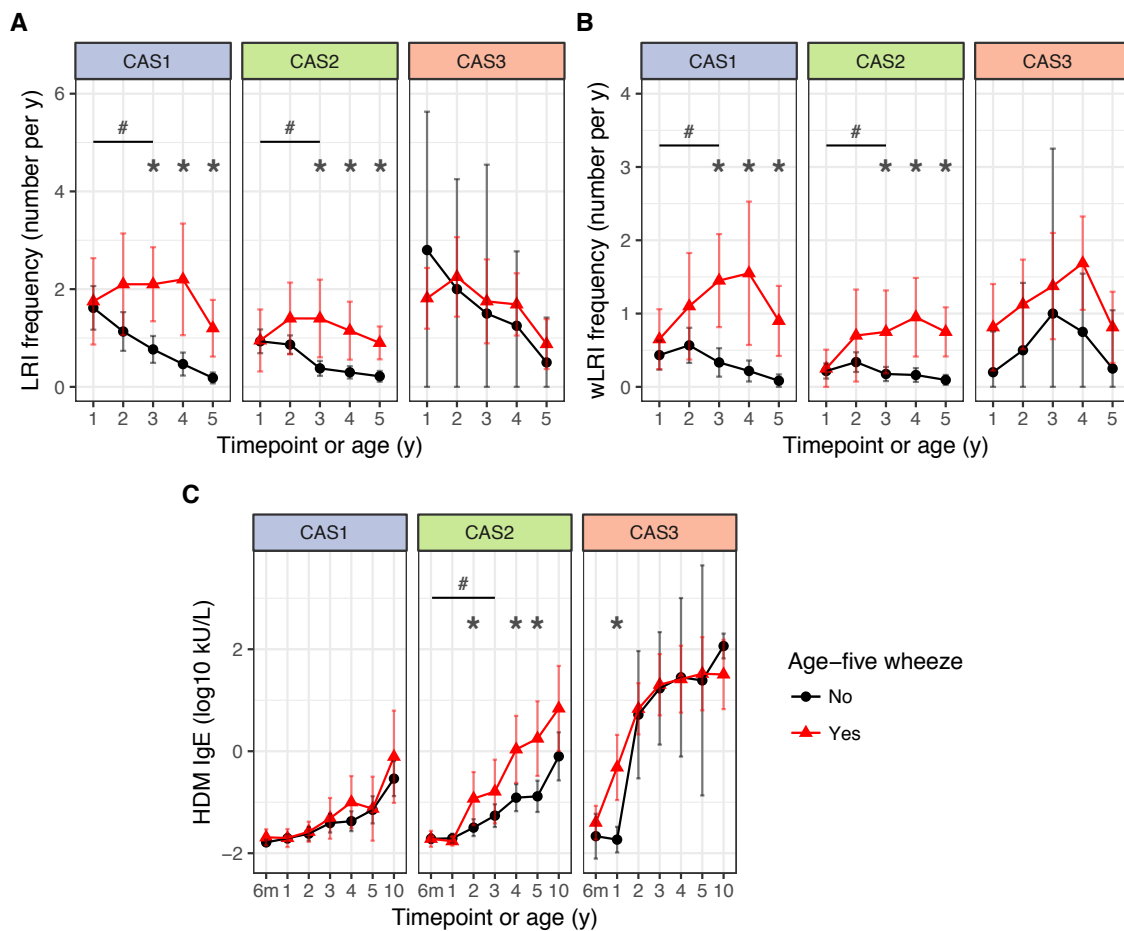
Feature? = whether variable was used as a clustering feature or not; geom. mean = geometric mean; ARI = acute respiratory infection (lower or upper); LRI = lower respiratory infection; MPG = microbiome profile group; NPA = nasopharyngeal aspirate; prop. = proportion; URI = upper respiratory infection; 7w = 7 weeks. For categorical variables, associations were tested using Fisher exact test; for continuous variables, Kruskal-Wallis and Mann-Whitney-Wilcoxon. Bold text indicates statistical significance ( $p < 0.05$ ); italics indicate near-significance ( $p < 0.10$ ). \*Not used as clustering features, as these were derived variables; the variables from which they were derived (URI, LRI, wLRI, fLRI) were used instead.

Variable	Age	CAS1 (N=88)	CAS2 (N=107)	CAS3 (N=22)	P-value (unadjusted)			Feature?	
		Mean (95% CI)	Mean (95% CI)	Mean (95% CI)	Overall	1 vs. 2	1 vs. 3		2 vs. 3
URI (events per y)	1	2.9 (2.4-3.3)	2.6 (2.2-3)	2.5 (1.7-3.3)	0.59	0.34	0.5	0.96	Yes
	2	3.2 (2.6-3.7)	2.6 (2.2-3)	2.5 (1.2-3.8)	0.19	0.19	0.12	0.34	Yes
	3	2.7 (2.2-3.2)	2.8 (2.4-3.3)	2.2 (1.3-3.2)	0.45	0.41	0.59	0.24	Yes
	4	2.1 (1.7-2.6)	2.2 (1.8-2.7)	1.7 (0.77-2.7)	0.5	0.94	0.26	0.27	No
	5	1.6 (1.1-2)	1.5 (1.2-1.9)	0.67 (0.2-1.1)	0.081	0.76	<b>0.047</b>	<b>0.026</b>	No
LRI (events per y)	1	1.6 (1.2-1.9)	0.98 (0.76-1.2)	2 (1.3-2.6)	<b>4.0E-03</b>	<b>0.021</b>	0.17	<b>2.6E-03</b>	Yes
	2	1.4 (0.98-1.7)	1 (0.81-1.2)	2.2 (1.6-2.9)	<b>2.5E-03</b>	0.83	<b>6.1E-03</b>	<b>2.0E-04</b>	Yes
	3	1 (0.76-1.3)	0.6 (0.4-0.8)	1.8 (1.1-2.6)	<b>6.1E-04</b>	<b>0.02</b>	<b>0.039</b>	<b>2.7E-04</b>	Yes
	4	0.87 (0.52-1.2)	0.46 (0.3-0.63)	2 (1.1-2.8)	<b>1.7E-05</b>	0.3	<b>3.5E-04</b>	<b>1.6E-06</b>	No
	5	0.42 (0.24-0.6)	0.36 (0.24-0.48)	0.86 (0.44-1.3)	<b>0.019</b>	1	<b>0.011</b>	<b>7.5E-03</b>	No
Wheezy LRI (wLRI, events per y)	1	0.47 (0.3-0.63)	0.24 (0.15-0.34)	0.64 (0.19-1.1)	<i>0.054</i>	<b>0.036</b>	0.61	<i>0.065</i>	Yes
	2	0.68 (0.45-0.91)	0.41 (0.26-0.56)	1 (0.56-1.5)	<b>5.2E-03</b>	<i>0.063</i>	<i>0.066</i>	<b>1.7E-03</b>	Yes
	3	0.59 (0.37-0.81)	0.3 (0.17-0.44)	1.4 (0.78-2.1)	<b>4.6E-05</b>	<i>0.065</i>	<b>2.5E-03</b>	<b>6.6E-06</b>	Yes
	4	0.52 (0.25-0.79)	0.32 (0.18-0.46)	1.9 (0.95-2.8)	<b>4.5E-08</b>	0.86	<b>9.3E-07</b>	<b>3.3E-08</b>	No
	5	0.28 (0.13-0.42)	0.23 (0.13-0.33)	0.76 (0.36-1.2)	<b>2.3E-03</b>	0.99	<b>2.0E-03</b>	<b>1.2E-03</b>	No
Febrile LRI (fLRI, events per y)	1	0.36 (0.22-0.51)	0.28 (0.16-0.4)	0.55 (0.28-0.81)	<b>0.025</b>	0.24	<i>0.071</i>	<b>6.4E-03</b>	Yes
	2	0.36 (0.23-0.5)	0.33 (0.22-0.43)	0.95 (0.46-1.4)	<b>0.01</b>	1	<b>6.1E-03</b>	<b>3.8E-03</b>	Yes
	3	0.38 (0.21-0.55)	0.16 (0.09-0.23)	0.52 (0.13-0.92)	<i>0.06</i>	<i>0.063</i>	0.44	<b>0.04</b>	Yes
	4	0.3 (0.13-0.47)	0.15 (0.064-0.24)	0.43 (0.16-0.7)	<b>0.021</b>	0.18	<i>0.091</i>	<b>4.9E-03</b>	No
	5	0.19 (0.082-0.3)	0.14 (0.06-0.21)	0.19 (0-0.42)	0.83	0.55	0.91	0.8	No
		<b>Prop. (95% CI)</b>	<b>Prop. (95% CI)</b>	<b>Prop. (95% CI)</b>	<b>Overall</b>	<b>vs. 2</b>	<b>1 vs. 3</b>	<b>2 vs. 3</b>	
>20% <i>Streptococcus</i> in first infection-naive NPA sample	7w	11% (0.34%- 23%)	15% (3.3%-26%)	44% (3.9%-85%)	<i>0.081</i>	0.75	<b>0.042</b>	<i>0.065</i>	No
	6m	7.6% (1.6%-14%)	18% (10%-26%)	14% (0%-31%)	0.12	<b>0.045</b>	0.39	1	No
% Healthy NPAs with infection- associated MPGs	0-2	49% (38%-59%)	32% (24%-39%)	62% (47%-76%)	<b>1.2E-03</b>	<b>0.013</b>	0.2	<b>5.5E-04</b>	No
	2-4	46% (37%-55%)	44% (37%-51%)	45% (29%-61%)	0.9	0.67	0.92	0.8	No

**TABLE 3.4: Analysis of selected predictors for age-five wheeze within each CAS cluster, with demographic covariates (sex, BMI, parental history of asthma)**

BMI = body mass index; HDM = house dust mite; LRI = lower respiratory infection. Association analyses performed via generalised linear models (GLM) with demographic covariates: age-five wheeze ~ predictor + sex (male) + BMI at age 3 + paternal history of asthma + maternal history of asthma. Bold text indicates statistical significance ( $p < 0.05$ ); italics indicate near-significance ( $p < 0.10$ ). \*Odds ratio (OR) is for every unit increase in log10 IgE, IgG4 or IgG (i.e.10-fold increase in IgE, IgG4 or IgG).

Selected predictors for age-five wheeze	Age	CAS1 (N=88)		CAS2 (N=107)		CAS3 (N=22)		All (N=261)	
		OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value
LRI (events per y)	1	0.97 (0.71-1.3)	0.84	1 (0.61-1.5)	0.99	0.48 (0.13-1.1)	0.16	1 (0.81-1.2)	0.92
	2	1.2 (0.88-1.6)	0.26	1.5 (0.97-2.5)	0.069	0.99 (0.34-2.6)	0.98	1.4 (1.1-1.7)	<b>5.3E-03</b>
	3	2 (1.3-3.2)	<b>2.3E-03</b>	2.6 (1.5-5.3)	<b>2.7E-03</b>	0.98 (0.4-2.6)	0.96	2 (1.5-2.7)	<b>3.8E-06</b>
	4	2 (1.4-3.4)	<b>2.0E-03</b>	3.6 (1.8-8.3)	<b>6.5E-04</b>	1.9 (0.57-8.4)	0.32	2.5 (1.8-3.6)	<b>1.5E-07</b>
Wheezy LRI (events per y)	1	1.3 (0.68-2.4)	0.43	1.1 (0.35-3)	0.83	2.6 (0.62-58)	0.34	1.5 (0.98-2.3)	0.06
	2	1.2 (0.8-2)	0.33	1.6 (0.89-2.9)	0.12	2.4 (0.67-16)	0.24	1.6 (1.2-2.2)	<b>5.6E-03</b>
	3	2.8 (1.6-5.6)	<b>1.3E-03</b>	3 (1.4-8)	<b>0.016</b>	1.2 (0.43-4.6)	0.76	2.7 (1.8-4.2)	<b>4.1E-06</b>
	4	2.5 (1.5-5)	<b>4.0E-03</b>	6.3 (2.5-21)	<b>6.8E-04</b>	7.1 (1.2-169)	0.1	3.9 (2.5-6.7)	<b>5.4E-08</b>
Febrile LRI (events per y)	1	1.6 (0.77-3.6)	0.21	0.84 (0.28-1.9)	0.71	7.3 (0.78-178)	0.12	1.5 (0.93-2.4)	0.098
	2	1 (0.44-2.2)	1	4.8 (1.8-15)	<b>3.9E-03</b>	1.6 (0.48-10)	0.5	2.3 (1.4-3.9)	<b>1.2E-03</b>
	3	2 (1-4.8)	0.08	4.3 (1.2-15)	<b>0.02</b>	4.2 (0.55-519)	0.37	2.4 (1.4-4.3)	<b>2.3E-03</b>
	4	1.8 (0.97-4.1)	0.092	2.6 (0.88-8.3)	0.082	1.1 (0.11-18)	0.93	2.2 (1.3-4)	<b>5.9E-03</b>
Quartile of % healthy NPAs with infection-associated MPGs	0-2	1 (0.54-1.8)	0.98	1.3 (0.72-2.4)	0.36	NA	NA	1.3 (0.89-1.8)	0.19
	2-4	0.45 (0.19-0.88)	<b>0.035</b>	1 (0.51-2.1)	0.9	NA	NA	0.8 (0.53-1.2)	0.24
HDM IgE (log10 kU/L)*	6m	8 (0.85-94)	0.074	0.93 (0.14-3.6)	0.92	3.4 (0.26-180)	0.4	2.3 (0.99-5.8)	0.054
	1	1.5 (0.22-7.8)	0.65	0.54 (0.039-2.3)	0.51	39 (2.5-22000)	0.082	2.7 (1.5-5)	<b>0.00089</b>
	2	0.93 (0.28-2.5)	0.89	2 (1.2-3.7)	<b>0.016</b>	1.4 (0.38-4.8)	0.62	2 (1.5-2.8)	<b>2.80E-05</b>
	3	1.4 (0.68-2.9)	0.32	1.5 (0.9-2.4)	0.12	1.5 (0.4-5.2)	0.55	1.7 (1.3-2.2)	<b>1.00E-04</b>
HDM IgG4 (log10 µg/L)*	4	1.9 (0.94-4.1)	0.086	1.9 (1.2-3.1)	<b>0.011</b>	1.4 (0.31-5.5)	0.64	1.9 (1.5-2.5)	<b>3.70E-06</b>
	6m	NA (NA-NA)	0.55	0.053 (NA-6.5e+24)	0.99	28 (1.7e-34-NA)	0.99	1.4 (0.88-2.6)	0.17
	1	NA (NA-NA)	0.61	1.1 (0.8-1.5)	0.5	0.9 (0.58-1.3)	0.6	1.2 (1-1.4)	0.053
	2	1.1 (0.71-1.6)	0.67	1.1 (0.85-1.4)	0.61	0.4 (0.038-1.2)	0.26	1.1 (1-1.3)	0.056
HDM IgG (log10 mg/L)*	3	1.1 (0.85-1.5)	0.35	1.1 (0.77-2)	0.64	0.94 (0.19-2.3)	0.9	1.1 (0.98-1.2)	0.1
	4	1.2 (0.98-1.5)	0.082	0.89 (0.7-1.1)	0.33	0.46 (0.031-5.4)	0.53	1.1 (1-1.3)	<b>0.034</b>
	1	25 (0.32-1.6E+04)	0.19	3.3 (0.16-46)	0.38	5.6E-03 (8.4E-06-0.57)	0.058	2 (0.31-11)	0.44
	2	0.8 (0.15-3.5)	0.78	0.97 (0.24-3.7)	0.96	0.79 (0.031-18)	0.88	1.3 (0.6-2.9)	0.48
3	2.3 (0.14-35)	0.54	0.48 (0.057-2.5)	0.43	3.9 (0.26-96)	0.34	2.1 (0.89-5)	0.089	



**FIGURE 3.5: LRI frequency (A), wheezy LRI (wLRI) frequency (B), and HDM IgE (C), stratified by age-five wheeze status, cluster and time, in the CAS dataset.**

Points indicate means; bars indicate 95% CI (t-distribution). # $p < 0.05$  for repeated-measures ANOVA across timepoints from the first 3 years of life (see Table 4). \* $p < 0.05$  for Mann-Whitney-Wilcoxon comparison within each timepoint.

### 3.2.3 CAS3: high-risk atopic cluster with persistent wheeze

CAS3 was a “high-risk” cluster, where persistent respiratory wheeze and atopic disease was seen in more than half the group throughout the first 10 years of life (**Table 3.1; Figure 3.3**). This cluster was dominated by males (86%, Fisher exact test, unadjusted  $p = 6.8 \times 10^{-3}$  compared to CAS1, **Table 3.1**), and appeared to represent an early- and multi-sensitised atopic phenotype with persistent wheeze. CAS3 had elevated IgE, IgG, and IgG4 responses to common allergens, especially Phadiatop, HDM and peanut IgE from 6 months onwards (**Table 3.2; Figure 3.4; Supplementary Table B.3B**). SPTs were also more frequently positive in CAS3, especially to HDM and food allergens (peanut, cow’s milk and egg white, **Supplementary Table B.3D**).

No strong predictors for age-five wheeze were identified within CAS3 (**Table 3.4**): only couch grass IgE at age 2 and acute respiratory infection (ARI) frequency at age 1 were weakly significant (both  $p = 0.046$ ). Neither of these reached statistical significance when incorporated in the same model. However, the prolific IgE response, and the frequency and severity of early-life LRIs in this cluster (**Table 3.3**), strongly suggest contribution from both atopic and non-atopic causes of wheeze. Hence, CAS3 primarily represented those with extreme levels of atopic sensitisation and infection. The relative paucity of identifiable predictors may be explained by the small size of CAS3 ( $N=22$ ), the intrinsically high rate of wheeze in the cluster (76% with age-five wheeze), and saturation of risk from high levels of IgE and frequent infections.

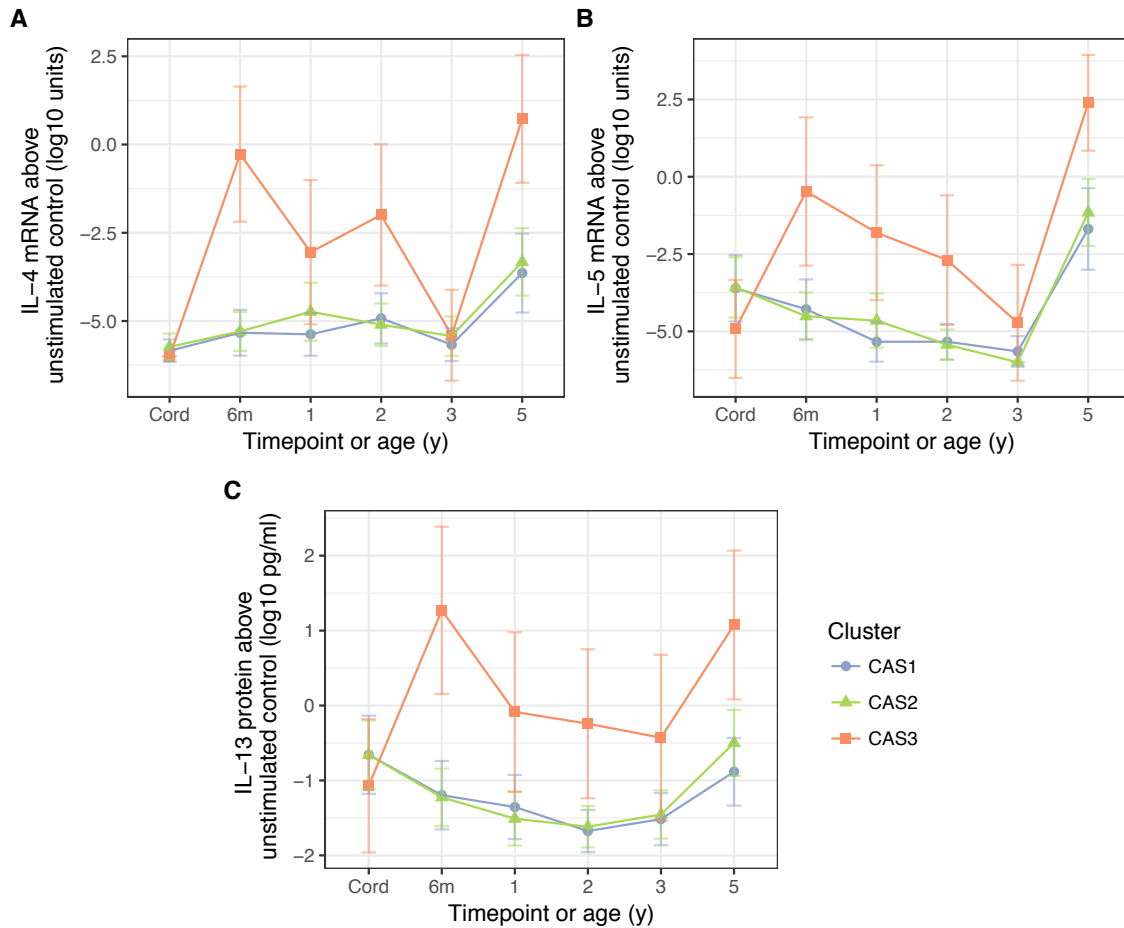
Unlike the antibody measurements, cytokine measurements were excluded as clustering features due to high missingness. Nonetheless, with post-hoc analyses, we found that in vitro stimulation of PBMCs with HDM antigen elicited stronger Th2 cytokine responses in CAS3 compared to other clusters (**Table 3.2, Figure 3.6**). These cytokines (IL-4, IL-5, IL-13) were elevated from a very young age (Wilcoxon, adjusted  $p = 4.6 \times 10^{-5}$  for IL-4 mRNA at age 6m, compared to CAS1), coinciding with increase in HDM IgE and IgG4 responses. Weaker but similar differences were observed for peanut- and ovalbumin-stimulated PBMCs at 6 months (unadjusted  $p < 0.05$  for all, **Supplementary Table B.3C**). There were no other significant differences for other non-Th2 cytokines (IFN- $\gamma$ , IL-10), nor were there specific differences for CAS1 or CAS2.

### 3.2.4 Comparison of measures of immunological response

Across all clusters, allergen-specific IgG4 and IgG were positively correlated with IgE for the same allergen (especially HDM, **Supplementary Figure B.5**). As noted previously, CAS2 and CAS3 were distinguished from CAS1 by high IgG4, and CAS3 had greater IgG4 than either CAS1 or CAS2 (**Supplementary Table B.3B; Figure 3.4**). Decision tree analysis (**Supplementary Figures B.7, B.8, B.9**) confirmed that IgG4-type variables strongly separated CAS2 and CAS3 from CAS1, while IgE-type variables separated CAS3 from the others.

Although previous literature suggests possible protection conferred by IgG4 [22] or IgG [23], in this study there was no clear evidence of such protection against later wheeze (**Table 3.4**). Furthermore, the protected status of CAS2 relative to CAS3 was unlikely to be driven by IgG4, given that CAS3 had greater quantities of both IgE and IgG4.

Although they were highly correlated, IgE, IgG, Th2 cytokine and SPT responses did not overlap perfectly. CAS3 was enriched for individuals with strong signals in all modalities, but there remained individuals within CAS3 and the rest of the cohort who were only responsive in some modalities and not others. Notably, the general direction of IgE, IgG4, SPT and Th2 cytokine signals did not always coincide (**Supplementary Figure B.6**).



**FIGURE 3.6: PBMC expression of IL-5 (A) and IL-4 mRNA (B), as well as IL-13 protein (C), in response to stimulation HDM, stratified by cluster and time (CAS).**

Cord = cord blood sample collected at birth. Points indicate means; bars indicate 95% CI (t-distribution).

### 3.2.5 Comparison of clusters to existing criteria for atopy

The npEM-derived CAS clusters were partially consistent with traditional atopy thresholds (i.e. any specific IgE  $\geq 0.35$  kU/L or SPT  $\geq 2$ mm at age 2). When we compared CAS clusters with supervised groups created using traditional thresholds (**Supplementary Table B.5**), we found that CAS1 most closely matched a non-atopic phenotype (58 of 84 had no specific IgE greater than 0.35 kU/L by age 2). Conversely, CAS2 and CAS3 partially matched traditional criteria for atopy, with CAS3 being an extreme phenotype (all 22 children in CAS3 had some specific IgE  $\geq 0.35$ kU/L by age 2).

However, the CAS clusters outperformed IgE/SPT-defined atopy in terms of predicting for age-five wheeze (likelihood ratio test for clusters vs. IgE/SPT, Chi-squared=23,  $p = 2.0 \times 10^{-6}$ ). In addition, at age 2, 68% of CAS3 were “sensitised” (any specific IgE  $\geq 0.35$ kU/L) to two or more allergens, compared to only 1% and 6% for CAS1 and CAS2 respectively. This emphasised CAS3 as an early- and multi-sensitised phenotype. Finally, fewer members of CAS1 and CAS2 who were IgE- or SPT-responsive prior to age 5 maintained atopic wheeze at age 5 (23% or 24 of 103), compared to CAS3 (76% or 16 of 21). Therefore, the association of IgE and SPT with disease risk varied across clusters. This suggests that fixed atopy thresholds are not sufficient to delineate risk profiles – instead, an unsupervised clustering approach may be more informative.

### 3.2.6 Comparison of clusters to time-dependent wheeze phenotypes and atopic disease

We mapped the npEM-derived clusters to pre-defined wheezing phenotypes (**Figure 3.3C**): no wheeze (in the first three years of life, or at age 5), transient wheeze (only in first three years), late wheeze (only at age 5), and persistent wheeze (both first three years and age 5). We found that CAS3 was enriched for persistent wheeze, while individuals in CAS1 or CAS2 tended to have transient or no wheeze. There were rarely any members of CAS with late wheeze (approximately 10%).

In addition to persistent wheeze, CAS3 was also enriched for persistent food sensitisation (peanut IgE  $\geq 0.35$  kU/L, or positive egg white or cow’s milk SPTs) and persistent eczema: 44% of CAS3 experienced all three (**Supplementary Figure B.4**). Almost all individuals in CAS3 had both eczema and food sensitisation from age 6m onwards, with rates of food sensitisation and wheeze increasing with time (**Figure 3.3D**). In contrast, CAS1 and CAS2 had low rates of food sensitisation, and declining rates of both eczema and wheeze. These trends lend credence to recent suggestions that the “atopic march” phenotype [24, 25] may only be present in a minority of the population (e.g. CAS3) [19].

### 3.2.7 Relationship with the nasopharyngeal microbiome

Previous studies suggest an association between asthma risk and early-life disruption of the respiratory microbiome, especially colonisation with *Streptococcus* spp. in the first 7 weeks of life [26]. In this study, using the same data and definitions, we found that CAS3 was overrepresented by individuals who had >20% relative abundance of *Streptococcus* in their first infection-naïve healthy NPA, within the first 7 weeks of life (44% versus 11% and 15% in CAS1 and CAS2, respectively; Fisher exact test, unadjusted  $p = 0.042$  and 0.065, respectively; **Table 3.3**).

Furthermore, Teo et al and others [26, 27] previously found that transient incursions with certain MPGs (*Streptococcus*, *Haemophilus*, *Moraxella* spp.) were associated with increased frequency and severity of subsequent LRIs and wheezing disease. Here, we found that proportion of these infection-associated MPGs in healthy samples from age 0 to 2 was greater in CAS3 (62% vs. 49% and 32% in CAS1 and CAS2, respectively; Fisher exact test,

unadjusted  $p = 0.2$  and  $5.5 \times 10^{-4}$ , respectively; **Table 3.3**). This finding was independent of LRI and wLRI frequency (GLM;  $p < 0.05$  for model predicting group membership, with age-two LRI and wLRI as covariates). On the contrary, there were no associations between cluster membership and health-associated MPGs (*Corynebacterium*, *Alloiooccus*, *Staphylococcus* spp.; **Supplementary Table B.3E**).

Recent work by Teo et al [28] suggested that infection-associated MPGs in early life were predictive for age-five wheeze in atopic children, while in non-atopic children they were predictive for transient wheeze. In this study, with the same cohort, we noted a similar trend for infection-associated MPGs from age 0 to 2, in relation to transient wheeze in “non-atopic” CAS1 (GLM, OR 3.6 per percent,  $p = 0.17$ , with demographic covariates). Surprisingly, there was evidence that infection-associated MPGs in later samples (from age 2 to 4) were *protective* against age-five wheeze in CAS1 (OR 0.086 per percent, 0.45 per quartile,  $p = 0.034$  and  $0.035$ , respectively; **Table 3.4**). Infection- and health-associated MPGs were otherwise not associated with age-five wheeze within the other clusters.

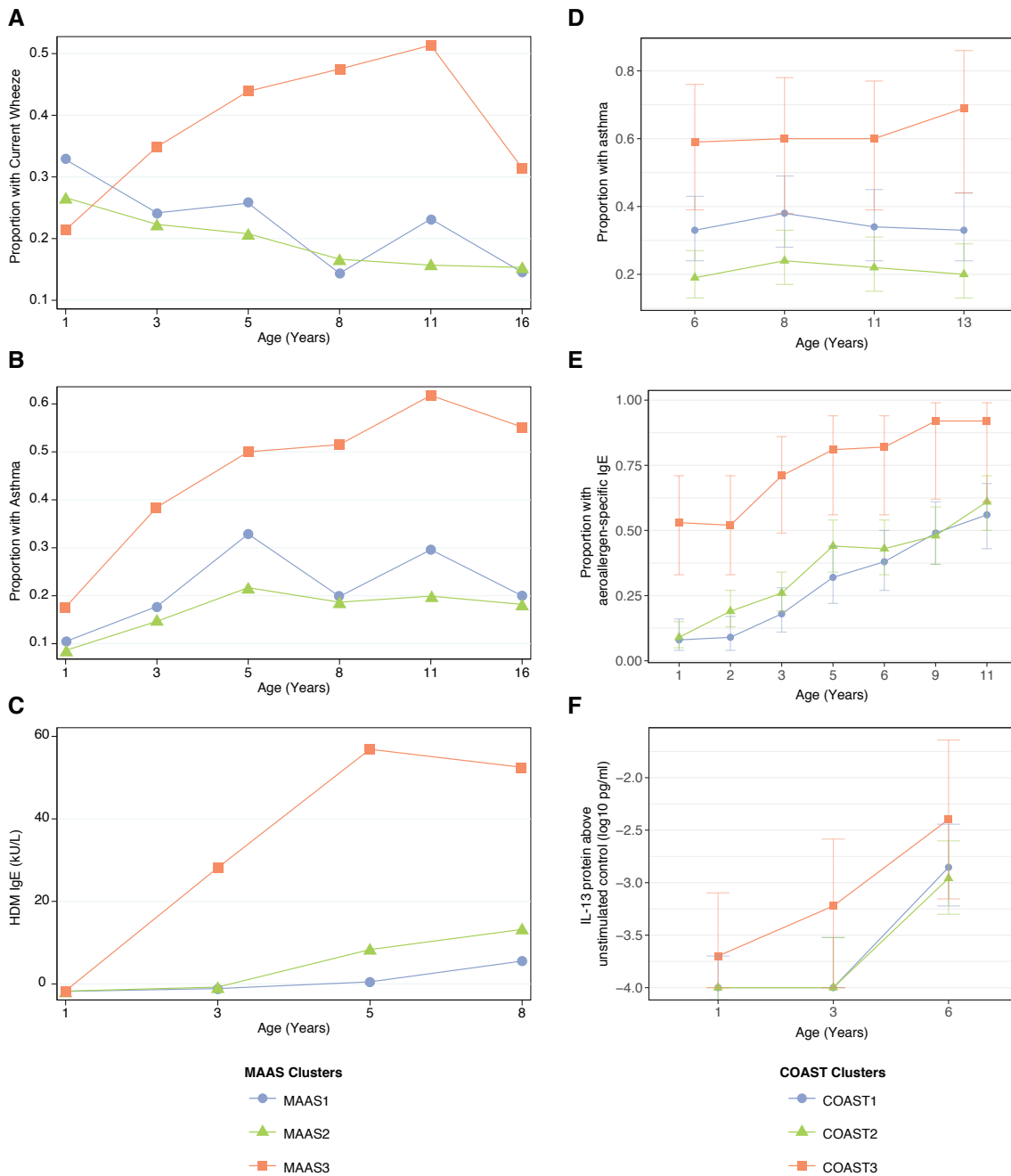
### 3.2.8 External replication of clusters in MAAS and COAST

The trajectories described by the CAS npEM clusters were replicated in two cohorts – the Manchester Asthma and Allergy Study (MAAS) ( $N = 1085$ ) [29] from Manchester, UK, and the Childhood Origins of Asthma Study (COAST) ( $N = 289$ ) from Wisconsin, USA [30]. After applying our npEM classifier to these external cohorts (**Methods, Section 3.4**), we found that individuals classified into “Cluster 3” (MAAS3/COAST3) had a persistent disease phenotype extending into late adolescence, with consistently high rates of parent-reported wheeze and physician-diagnosed asthma from birth to age 16. The other two clusters (Cluster 1 = MAAS1/COAST1; Cluster 2 = MAAS2/COAST2) appeared to be low-risk (**Figure 3.7A,B,D**).

MAAS3 and COAST3 exhibited stronger IgE expression (total, HDM, cat, dog) from ages 1 to 8 (**Figure 3.7C,E**), compared to other clusters in each dataset. Like CAS3, COAST3 demonstrated elevated PBMC expression of Th2 cytokine protein (IL-5 and IL-13) in response to HDM stimulation at age 3 (**Figure 3.7F**). This was not replicated in MAAS3, but previous work in MAAS had identified that a strong PBMC Th2 response (IL-5, IL-13) to HDM stimulation at age eight was associated with increased risk of HDM sensitisation and asthma [21]. Nonetheless, MAAS3 was overrepresented in “early-sensitised” and “multiple sensitised” phenotypes described by Lazic et al [17] from SPT and IgE data. Approximately 86% of individuals in MAAS3 belonged to either one of these two phenotypes, although only 13% of individuals in these two phenotypes were accounted for by MAAS3.

Furthermore, when we explored potential predictors of wheeze phenotypes and asthma diagnosis in later childhood, we found that the clusters in COAST were very similar to those in CAS. In COAST1, LRI and wLRI frequency at age 2 were predictive of asthma diagnosis at age six (GLMs with demographic covariates,  $p = 0.02$  and  $0.02$ , respectively), while in COAST2, HDM IgE at age 3, and LRI, wLRI and fLRI frequencies at age 2 were all predictive (GLMs,  $p < 0.05$  for all) (**Supplementary Figure B.10**). Although the timing and magnitude of associations differed between cohorts, this reaffirmed wheeze in Cluster 1 as being primarily non-atopic in origin, while wheeze in Cluster 2 appeared to be driven by both non-atopic and atopic factors.

We re-applied npEM classification to CAS using only those features present in MAAS or COAST. For MAAS and COAST features respectively, the subsequent clusters bore 79% and 72% concordance with the original CAS clusters. In both cases, concordance was excellent for Cluster 3 – all 22 members of the original CAS3 were correctly assigned to Cluster 3 after re-applying npEM. Therefore, CAS3, COAST3 and MAAS3 likely represent very similar phenotypes.



**FIGURE 3.7: Description of npEM-derived clusters in external cohorts: in MAAS, incidence of wheeze (A), asthma diagnosis (B), and HDM IgE levels (C); in COAST, incidence of asthma diagnosis (D), proportion of individuals with detectable aeroallergen-specific IgE levels (E), and PBMC protein expression of IL-13 following HDM stimulation above unstimulated control (F).**

MAAS cohort (N=934) was classified using npEM model from CAS, into MAAS1 (N=199, 21%), MAAS2 (N=692, 74%) and MAAS3 (N=43, 5%); these correspond to CAS clusters CAS1, 2 and 3, respectively. COAST cohort (N=285) was similarly classified into COAST1 (N=105, 37%), COAST2 (N=151, 53%) and COAST3 (N=29, 10%).



### 3.2.9 Internal stability and validity of CAS clusters

We checked the stability and validity of the CAS clusters with leave-one-out (LOO) analysis, Jaccard indices and silhouette widths. The average Jaccard indices from leave-one-individual-out analysis were 0.77, 0.76, and 0.85 for CAS1, 2 and 3, respectively. For leave-one-feature-out analysis, the average indices were 0.65, 0.60, and 0.74, respectively. This demonstrates that the clusters, especially CAS3, were relatively resilient to minor changes in sampling or feature selection.

In relation to internal validity of the CAS clusters, average silhouette widths were universally poor, at 0.05, 0.06 and 0.002 for CAS1, 2, 3, respectively, with an average for all three clusters of 0.05 (**Supplementary Figure B.2**). Silhouette widths were particularly suboptimal with CAS3, with at least half of those classified having negative values. The overall poor internal validity of the clusters may be due to the large-scale and exploratory nature of our approach – the metric may have been obscured by intra-cluster heterogeneity in other variables that were not particularly important for determining cluster membership. However, it must be noted that all clusters on average yielded positive silhouette widths, and as observed in the rest of the results, they were all relatively homogeneous in terms of the outcomes of interest (wheeze status, allergic disease phenotypes).

### 3.2.10 Decision tree analysis

Decision tree analysis on the CAS dataset, using all available predictors from all timepoints, created a “Simple Tree” with two decision nodes and three end nodes (**Supplementary Figure B.7**). This tree had 89% accuracy in retrieving cluster memberships from the original npEM model, where accuracy is calculated as percentage overlap of tree clusters with original CAS clusters. We found that membership in the CAS3-equivalent tree cluster was a better predictor for age-five wheeze (likelihood ratio test, Chi-squared=19,  $p < 1 \times 10^{-5}$ ) than traditional thresholds for atopy based on IgE and SPT measurements at age 2. IgG4-related variables best separated CAS1 from other clusters, while IgE-related variables best separated CAS2 and CAS3. Explicitly forcing the exclusion of Phadiatop variables from tree analysis caused these thresholds to be replaced with allergen-specific assays (HDM IgE for Phadiatop IgE, **Supplementary Figure B.8**) in a way that is consistent with correlation patterns amongst IgE and IgG4 variables (**Supplementary Table B.6**).

We also constructed a “Comprehensive Tree” that best split individuals into six groups, based on cluster membership crossed with age-five wheeze status (**Supplementary Figure B.9**). We thus identified nodes that were consistent with predictors for wheeze found in the previous regression analyses (**Table 3.4**), combined with nodes from the Simple Tree (**Supplementary Figure B.7**). The Comprehensive Tree had 77% accuracy in recovering both cluster membership and wheeze status. In terms of identifying pure wheeze status at age 5, the accuracy of the tree was 84%, with a positive predictive value (PPV, or precision) of 72%, negative predictive value (NPV) of 88%, sensitivity (recall) of 71% and specificity of 89%. The Comprehensive Tree was more successful in flagging age-five wheeze (likelihood ratio test, Chi-squared=60,  $p = 6.1 \times 10^{-13}$ ), compared to the traditional atopy thresholds described previously.

We attempted to validate a modified version of CAS-derived Simple Tree in the MAAS dataset, as that cohort contained measurements of both age-five HDM IgE and HDM IgG4, which we used as surrogates for age-three HDM IgE and HDM IgG4. These features comprised two decision-node features in the Phadiatop-free equivalent of the CAS Simple Tree (**Supplementary Figure B.8**). COAST did not have any IgG4 measurements, so tree validation was not attempted there. The performance of the Simple Tree when applied to MAAS was poor, with only 20% accuracy in terms of overlap between tree clusters and

npEM clusters, compared to 89% in CAS. Given the replication of the npEM model but not the decision trees, this suggests that multiple allergy-related measurements provide better prediction of disease state than singular measurements.

### 3.3 Discussion

We have used model-based cluster analysis to uncover clusters of children with differential asthma susceptibility. Specifically, there was a high-risk group (Cluster 3) characterised by very early allergen-specific Th2 activity; early sensitization to multiple allergens including food allergens; and concurrent frequent respiratory infections – resulting in high incidence of atopic persistent wheeze. We also found a lower-risk cluster (Cluster 2), with limited or delayed elevation in IgE – this resulted in a lower incidence of mixed (atopic and non-atopic) wheeze. Finally, there was a low-risk cluster (Cluster 1) which exhibited occasional and transient infection-related wheeze, with minimal allergen sensitisation. These clusters were replicated in external datasets, suggesting relevance across populations. Summaries of key findings are given in **Table 3.5** and **Figure 3.2**.

#### 3.3.1 Cluster 3 is a high-risk, multi-sensitised, atopic phenotype

Cluster 3 represented a multi-sensitive or polysensitised phenotype [31]. In CAS3, not only was total IgE elevated, but specific IgE were also raised for most allergens. Three in four CAS3 individuals were sensitised (specific IgE  $\geq 0.35$  kU/L) to two or more allergens. In our external replication with MAAS, we observed a large overlap between our predicted high-risk phenotype (MAAS3) and the multiple atopy phenotype from Lazic et al [17]. This was consistent with findings from other studies, where the severely atopic and polysensitised subpopulation was at greater risk of both wheezing disease and reduced lung function [32].

It is not currently known what is fundamentally producing the strong atopic predisposition in Cluster 3. It is possible that inherited (genetic/epigenetic) or environmental factors (including in utero or perinatal exposures) may be involved, and these should be targets for future investigations. The overrepresentation of males in CAS3 is consistent with the consensus that young boys are at greater risk for asthma than young girls; this was traditionally believed to be due to intrinsic sex differences in airway diameter [33]. However, our cluster analysis did not employ any clustering features related to airway size. This suggests that other sex-related factors could be involved, such as differences in immunity and allergic susceptibility. Allergic sensitisation is more frequent amongst prepubescent boys than girls [34, 35], and this may be linked to differences in cytokine responsiveness. However, not all boys were clustered into Cluster 3; and sex was not found to be a determinant for either IgE levels or cytokine response in CAS.

We did observe that CAS3 overlapped strongly with both persistent food sensitisation and eczema, and that persistent wheeze co-occurred with early sensitisation and eczema. This suggests that the “atopic march” may play a role in CAS3. Early disruption of the skin barrier and exposure to certain food allergens may act in concert to promote and entrench the atopic phenotype, through the activation of cytokine pathways involving TSLP, IL33 and IL25 [24, 25]. Although recent research has suggested that very few children actually follow the disease trajectory of the atopic march [19], we hypothesise that it remains relevant to a small but important high-risk subpopulation, who may potentially benefit from early interventions targeted at halting the progression of disease.

### 3.3.2 Role of early-life HDM hypersensitivity

In all three cohorts (CAS, MAAS, COAST), house dust mite (HDM) sensitivity was an important determinant of atopic disease risk. HDM was a strong predictor for both CAS3 membership and later childhood wheeze in CAS2, as well as being a “dominant” allergen in the Phadiatop Infant assays. CAS3 in particular exhibited early and extreme HDM hypersensitivity, with prematurely-elevated HDM IgE, as well as PBMC Th2 response (IL-4, 5, 9, 13) to HDM stimulation. Similar phenomena were seen with MAAS3 and COAST3. The importance of HDM hypersensitivity in driving allergic disease in some populations is well-described in the literature [36, 37]. Previous findings from MAAS and a similar cohort RAINE [21] have shown a confluence of high HDM IgE, as well as PBMC Th2 cytokine levels such as IL-13 and IL-5, in discrete subsets of the population. However, we did observe that in other clusters (CAS1 and CAS2), some individuals with purported HDM sensitisation (IgE > 0.35 kU/L) did not produce detectable Th2 responses; the reverse was also true, where Th2 response did not necessarily result in high IgE. It may be the case that there is high intra-individual variation in IgE and cytokine responses, or stochastic variation in detectability of IgE or cytokine, which may obscure association analyses. Regardless, early and strong Th2 cytokine responses against HDM indicate a high-risk phenotype.

### 3.3.3 Role of early-life food and peanut sensitization

Interestingly, early-life peanut IgE was a strong delineator between high-risk CAS3 and lower-risk CAS1 and 2. There is evidence in the literature for transmission of peanut allergen *in utero* or via breastmilk [38, 39], as well as early sensitisation via home environmental exposure, especially in those with concurrent eczema or a predisposing filaggrin (*FLG*) mutation that may allow transcutaneous infiltration of allergen [40, 41]. The strong correlation between Phadiatop and peanut IgE in the first year of life suggests that either peanut reactivity is significant at this earlier timepoint, or that “peanut-specific IgE” is cross-reactive and representative of some other allergen hypersensitivity. The fact that this correlation exists within each cluster (**Supplementary Table B.6**) suggests that it is not caused solely by differences between low and high-risk clusters (CAS1/CAS2 vs. CAS3). There is a possibility that peanut IgE is a marker for a broader phenotype of early and unremitting sensitisation to multiple food allergens (peanut, cow’s milk, eggwhite), as we had observed in CAS3. However, it is unlikely that premature exposure to food allergen is the lone driver for sensitisation and disease, given that well-timed oral exposures to common food allergens (e.g. within 4 to 6 months of age) may actually be protective [42]. There is some evidence that quantity (minute vs. abundant), route (skin vs. oral) and timing (early vs. late) of exposure are key modifiers of risk [25]. Ultimately, an underlying atopic predisposition linked to early-life exposure to food allergen may be driving the high-risk phenotype in Cluster 3.

### 3.3.4 IgG4 separates individuals susceptible to atopic wheeze from those who are not

In our study, neither IgG nor IgG4 were strong predictors or protectors of wheeze. However, IgG4 was a strong delineator of cluster membership in CAS, with individuals from CAS2 and CAS3 having elevated IgG4 across all specificities compared to CAS1. Vulnerability to early IgE-driven respiratory disease (“atopic wheeze”) can be seen in these same individuals –in CAS2 where HDM IgE is predictive for later wheeze, and in CAS3 where both wheeze frequency and IgE are elevated. Hence, although there had previously been

doubt about the efficacy of IgG4 as a marker for atopy [43], our study suggests that IgG4 is still relevant for determining atopic risk, especially when used in combination with IgE.

The underlying biology behind the association of IgG4 with susceptibility to “atopic wheeze” is unclear. Th2-related pathways drive production of both IgE and IgG4, with IgG4 predominating when modified by concurrent IL-10 signalling [44]. In susceptible individuals, IgG4 production likely precedes isotype switching to frank IgE production [45]. Multiple studies have reported that IgG4 is correlated with induced tolerance following desensitisation immunotherapy with high-dose allergen treatment [44]. However, based on this study alone, we cannot observe any protection from naturally-elevated IgG4 levels. Our group had previously suggested, using data from another cohort [23], that IgG and specifically IgG1 may provide endogenous protection against IgE-associated wheeze in children experiencing natural (low-level) exposure to aeroallergen. In this present study, IgG1 was not measured.

### 3.3.5 The role of respiratory infection and nasopharyngeal microbiome in childhood wheeze differs across different clusters

The co-occurrence of elevated IgE and LRI frequency in CAS3, as well as their predictive effect in CAS2, are consistent with previous findings from CAS [26, 46, 47]. They lend support to the theory that allergic and infective processes act additively to intensify airway inflammation during respiratory pathogen clearance, which in turn drives progression towards persistent wheeze [48]. In addition, our cluster analysis suggests that the pathologic effect of this interaction may be stratified in discrete subpopulations, rather than acting in a strictly dose-dependent fashion across the entire cohort. There may be subsets of children (CAS2 and CAS3) who are more susceptible to the effects of this viral-atopy interaction. On the other hand, pathogen clearance in infected non-atopic (CAS1) subjects may be more efficient, due to lack of susceptibility to the pro-inflammatory effects of atopic co-stimuli. This produces lower levels of “bystander” inflammatory damage to airway tissues, with opportunity for recovery, resulting in a less severe wheeze phenotype.

Of particular note is that, while both CAS1 and CAS2 have LRI and wLRI frequencies as predictors for age-five wheeze, CAS2 also has fLRI, particularly at age two. This, along with the general higher incidence of fLRI in CAS3, is consistent with previous findings from CAS [26, 47]. It suggests that symptomatically-severe infections, correlating with severe airway inflammation, may be more potent in causing persistence of wheeze, specifically among those who are “atopic” (CAS2 and CAS3).

In addition, even during periods of good health, the upper respiratory microbiome played a role in determining later childhood wheeze. Its effect interacted with cluster membership, as well as the age at which the microbiome changes occurred. CAS3 was enriched for early-life infection-associated MPGs (*Streptococcus*, *Moraxella*, and *Haemophilus*). This was consistent with the previous finding by Teo et al [28] that early-life infection-associated MPGs were predictive of age-five wheeze only within atopic individuals (as defined by IgE alone). Interestingly, in our current study, we found a protective effect of infection-associated MPGs from age two to four in CAS1. We hypothesise that those without atopy-related immune dysfunction are able to maintain a healthy trajectory by responding appropriately to stimuli from potential pathogens that colonise the respiratory tract, thus achieving protection against future (non-atopic) wheeze. This is akin to the “hygiene hypothesis”: exposure to a greater repertoire of pathogen-derived antigens may facilitate maturation of immune functions against said pathogens. Meanwhile, individuals with a predisposing immune dysfunction (i.e. “atopy” manifesting in early-life allergic sensitisation) may be responding in a maladaptive manner to these microbes [48]. This may

result in inability to clear potential pathogenic bacteria, or shaping of aberrant immune responses – with subsequent effects on airway inflammation and wheeze.

### 3.3.6 Implications for cluster analysis in asthma research

In this study, we applied mixture modelling to generate clusters from biological data. Similar methods such as latent class analysis (LCA) have previously been used in asthma research – for instance, LCA was applied to SPT and IgE measurements from MAAS to determine different patterns of allergen sensitisation and subsequent disease [17]. However, LCA is limited to categorical clustering features, so measures of sensitisation in that study were thresholded (e.g. IgE levels were split into  $<0.35$  kU/L, 0.35 to 100 kU/L, and  $>100$  kU/L). The method also assumed that these thresholds have the same relevance across all timepoints; that thresholds applied equally to all allergens; and that all allergens contributed equally to disease susceptibility profiles. Mixture modelling is an extension of LCA in that it does not require categorical variables or predetermined thresholds. Furthermore, non-parametric mixture modelling (npEM) does not require input features to have Gaussian distributions. Previous studies have used mixture models to explore phenotypes in adult asthma based on clinical measurements [49–52], and one of our own studies previously looked at cytokine expression patterns of PBMCs from children in response to HDM stimulation [21]. Our study is the first to apply non-parametric mixture modelling to data representing immune and respiratory health in early childhood, and to investigate possible predictors of disease within each cluster.

Currently, mixture models are limited by an unproven “track record”; a lack of consensus about best protocols for data processing and analysis; instability or inconsistency of clusters; difficulty in interpretation of results; and uncertainty regarding the validity of certain assumptions that accompany models [16]. Other methods of cluster analysis have similar problems, and while they have been applied frequently to asthma research, they have also produced a confusing myriad of phenotypes. The nature of cluster phenotypes is highly dependent on the type of features entered into the clustering algorithm. Clustering features that represent final clinical endpoints, such as markers of severity, may produce more heterogeneous clusters, as different pathological trajectories can arrive at similar endpoints. Some cluster phenotypes may contradict with each other, or may not be easily interpreted. Recently, Schoos et al [53] identified that, unlike our study, asthma was *not* as strongly associated with prominent HDM or peanut hypersensitivity in a Danish birth cohort (COPSAC) as other patterns of sensitisation (especially cat, dog and horse). However, we note that they used thresholded IgE  $>0.35$  kU/L to build their clusters. Other differences may emerge due to heterogeneity across different populations; geographical differences in environmental exposures and allergen sensitisation; and differences in testing procedures and phenotype definitions at different sites. COPSAC, CAS and COAST were cohorts enriched for high-risk individuals – each child had at least one parent with a history of atopic disease – while MAAS had no such recruitment criterion. Because of variability in findings, there has been wariness and scepticism among clinicians regarding the utility of mixture models and machine learning [54]. Ultimately, one may argue that discrepancies in our findings serve as a caution against the blind application of “algorithms” without due consideration of subtleties in target population and environment.

Nonetheless, what we have demonstrated here is the vast potential of cluster analysis. We have discovered clusters in an unsupervised and exploratory fashion, described them comprehensively, replicated our findings in multiple datasets, and compared our clusters with other existing phenotypes. In doing so, we have generated some new and interesting insights about the nature of atopy and asthma risk. Our results build on previous findings [11, 55] demonstrating that the concept of atopy, as an intrinsic or heritable predisposition

TABLE 3.5: Key findings from cluster analysis.

---

Certain childhood populations may be broadly split into three clusters, each representing a unique trajectory of immune function and susceptibility to respiratory infections: low-risk non-atopic Cluster 1 with transient wheeze; low-risk but allergy-susceptible Cluster 2 with mixed wheeze; and strongly-atopic high-risk Cluster 3 with persistent wheeze.

---

Cluster 3 is consistent with an early-sensitised and multi-sensitised phenotype.

---

HDM hypersensitivity is an important predictor of wheeze in allergic or allergy-susceptible individuals.

---

Food and peanut hypersensitivities are important contributors to membership in high-risk Cluster 3. This may be pathophysiologically related to eczema, multi-sensitisation and the atopic march.

---

In CAS, IgG4 flags for clusters with susceptibility to atopic disease (CAS2 and CAS3), while early and multiple-allergen elevation in IgE predicts frank atopic disease. The pathophysiological role of IgG4 remains unclear.

---

Allergic and infective processes act additively to intensify airway inflammation during respiratory pathogen clearance. Some (Cluster 3) may be more susceptible to this effect than others that lack strong allergic sensitisation (Cluster 1).

---

Tests for atopy (IgE, SPT, cytokines) do not overlap perfectly. Therefore, atopy may be better defined by the composite result from a battery of tests encapsulated in a predictive model, rather than just a single test or threshold.

---

The microbiome acts differently on asthma risk depending on cluster membership. In CAS, early-life asymptomatic colonisation with infection-associated MPGs is associated with risk of persistent wheeze in allergy-susceptible clusters (CAS2, CAS3), while it is potentially protective in non-atopic children (CAS1)

---

Different childhood populations may share similar trajectories of asthma susceptibility, but there may be subtle differences in terms of the types of tests, allergens, or biological signals that are most informative (SPT, IgE, cytokines, etc.).

---

to allergic disease, is more complicated than what could be described by dichotomies or thresholds. We have also demonstrated that addressing subgroup differences via cluster analysis allows for identification of intra-cluster disease predictors. In the future, clusters may be further characterised by other aspects of asthma pathophysiology, such as genomics, transcriptomics, and epigenomics.

### 3.3.7 Concluding statements

The results of our study strongly support the future use of predictive models with more precise and subgroup-driven representations of atopy or other relevant pathophysiology. We argue for ongoing collaboration between research groups in terms of refining methodology, answering questions unique to certain populations, and comparing cluster phenotypes arising from different algorithms and datasets. We believe that, as clustering methods become more frequently used, we will gradually develop better consensus on how such methods are best applied to biomedical phenomena. By continuing with these approaches, we can hopefully move away from fixed thresholds to more sophisticated formulations of risk, which will then improve future attempts at targeted screening, prevention and treatment of asthma. These approaches are already being applied to other heterogeneous diseases, and in the future computerised tools may be designed to embody the sum knowledge from these approaches. Such approaches can eventually help clinicians and scientists achieve a fuller understanding of pathophysiology, and hence better predict and manage human disease.

## 3.4 Methods

### 3.4.1 Patients and study design in CAS

Our discovery dataset was the Childhood Asthma Study (CAS), a prospective birth cohort ( $N=263$ ) operated by the Telethon Kids Institute from Perth, Western Australia [56]. The goal of CAS was to describe the risk factors and pathogenesis of childhood allergy and asthma. Further details of CAS have been reported previously [26, 47, 56–58].

In CAS, expectant parents were recruited from private paediatric clinics in Perth during the period spanning July 1996 to June 1998. Each child who was born and subsequently recruited had at least one parent with physician-diagnosed asthma or atopic disease (hayfever, eczema). The child was then followed from birth till age 10 at the latest, with routine medical examinations, clinical questionnaires, blood sampling at multiple time points (6–7 weeks, 6 months, 1 year, 2, 3, 4, 5, and 10 years) and collection of nasopharyngeal samples. Parents also kept a daily symptom diary for symptoms of respiratory infection in their child. The data extracted from these samples and measurements covered multiple “domains” of asthma pathogenesis, including respiratory infection, allergen sensitisation, and clinical or demographic background.

### 3.4.2 Measurements in CAS

For each child and visit, the investigators of CAS recorded metrics related to suspected or known modulators of asthma risk. These included markers of immune function: 1) IgG, IgG4, and IgE Phadiatop ImmunoCAP antibodies (ThermoFisher, Uppsala, Sweden), covering common allergens such as house-dust mite (HDM, *Dermatophagoides pteronyssinus*), mould, couch grass, ryegrass, peanut, cat dander; 2) IgE and IgG4 Phadiatop Infant and Adult assays (ThermoFisher, Uppsala, Sweden) that target multiple allergens simultaneously [59]; 3) skin prick or sensitisation tests (SPT) for HDM, mould, ryegrass, cat, peanut, cow’s milk and hen’s egg; and 4) cytokine responses (IL-4,5,9,13,10, IFN- $\gamma$ ) following in vitro stimulation of extracted peripheral blood mononuclear cells (PBMCs) by multiple antigen and allergen stimuli, including phytohaemagglutinin (PHA), HDM, cat, peanut and ovalbumin [47, 57].

In addition, nasopharyngeal samples (swabs or aspirates, NPAs) were taken from each child during healthy routine visits (healthy samples), and unscheduled visits where parents presented with their child if they have a suspected respiratory infection (disease samples). Frequency and severity of respiratory infections were measured accordingly. NPAs were then screened for viral and bacterial pathogens using rtPCR and 16s rRNA amplicon sequencing with Illumina MiSeq (San Diego, US), respectively [26]. These NPAs had previously classified by Teo et al [26, 28], based on clustering of bacterial composition, into microbiome profile groups (MPGs) that were associated with healthy respiratory states (health-associated MPGs, e.g. *Alloicoccus*-, *Staphylococcus*- or *Corynebacterium*-dominated) or infectious respiratory states (infection-associated MPGs, e.g. *Moraxella*-, *Haemophilus*-, or *Streptococcus*-dominated).

Other collected data included: sex, height and weight; paternal and maternal history of atopic disease; blood levels of basophils, plasmacytoid and myeloid dendritic cells as measured by fluorescence-assisted cell sorting (FACS); and levels of vitamin D (25-hydroxycholecalciferol, 25(OH)D) [58].

### 3.4.3 Identification of latent clusters and selection of clustering features

We adopted an exploratory approach to cluster analysis, whereby we attempted to interrogate as much of the existing dataset as possible, identifying latent clusters that arise from

the underlying data structure of CAS. We then assessed how these latent clusters correlate with risk of asthma or other markers of pathophysiology, such as degree of allergic sensitisation. All data processing and analysis was done in R v3.3.1 (RRID:SCR\_001905). A graphical overview of the analytic process is displayed in **Supplementary Figure B.3**.

To identify latent clusters, we applied non-parametric expectation-maximisation (“npEM”) mixture modelling to our discovery cohort CAS, using functions from the R package “mixtools” [60]. This method assumes that frequency distributions of each cluster can be represented by non-parametric density estimates learned from the data in an iterative process. npEM was used because: 1) it was plausible to consider a population as a mixture of subpopulations each with their own distributions; 2) it had advantages over other unsupervised approaches [61] – for example, with LCA, continuous variables cannot be handled appropriately; with hierarchical clustering, poor decisions made early in the classifying process are not easily amended; 3) many variables were categorical or non-Gaussian, so theoretically a non-parametric approach should be superior to a Gaussian mixture model or k-means approach; and 4) inherent within mixture models is an intuitive method for supervised classification of other datasets into similar clusters.

We used a largely non-selective approach to choosing features for cluster analysis, in that we attempted to retain as many CAS individuals and variables as possible. However, we did enforce certain quality-control measures such as excluding variables (“features”) that had missing data for >20% subjects (442 variables removed), and subjects with missing data for >30% of the remaining variables (39 subjects removed). Also excluded were features pertaining to our primary outcomes of interest: incidence of parent-reported wheeze, physician-diagnosed asthma and hayfever at all timepoints. We specifically excluded these from feature selection so we could determine how subsequent clusters differ in these outcomes, even when clustering was not explicitly driven by them. On the other hand, eczema was not excluded because of evidence that infantile eczema may itself influence the risk for subsequent sensitisation and asthma [62]. Frequency of wheeze in the context of respiratory infection was also included, as it was a symptomatic marker of infection severity. Variable reduction resulted in  $M = 174$  variables remaining out of an original 659. The complete list of variables included as clustering features is provided in **Supplementary Table B.1**, and importantly covers multiple domains including demographic (family history of atopy, household size), clinical (incidence of childhood eczema), immunological (IgE, IgG, IgG4, SPT) and microbiological (respiratory infections, viral pathogens associated with infection) features. By virtue of study design and quality control measures, many of the clustering features were related to immunological function or respiratory infection in *the first three years of life*.

Highly-skewed features, such as antibody and cytokine levels, were subjected to logarithmic (base 10) transformation. We also applied limited thresholding to some variables (cytokine responses, antibody assays), based on best practice for the reported limit-of-detection (LOD) of the measuring devices. The LOD for IgE was 0.03 kU/L; for IgG4, 0.0003  $\mu\text{g/L}$ ; for IgG, 0.4 mg/L. For these variables, we assigned any values below the LOD to half the LOD (i.e. 0.015 kU/L, 0.00015  $\mu\text{g/L}$ , and 0.2 mg/L, respectively). For stimulated cytokine expression above unstimulated control, any zero or negative values (i.e. unstimulated control had equal, or greater, expression than stimulated), were converted to 0.000001 units or 0.01pg/ml for mRNA and protein variables, respectively. Positional standardisation scaling was then applied across all variables, to equally weight the contributions of each feature to the mixture model. This involved replacing each value  $x_{ij}$  for individual  $i$  of feature  $j$ , by:

$$\frac{x_{ij} - \text{med}(x_j)}{\max(x_j) - \min(x_j)}$$



where functions  $\text{med}$ ,  $\text{max}$  and  $\text{min}$  refer to the median, maximum, and minimum for the complete-case dataset for feature  $j$ , respectively.

### 3.4.4 Cluster analysis using non-parametric mixture modelling

The processed and scaled CAS dataset was further split into those subjects with no missingness in the remaining variables (“complete-case”, 186 subjects, 174 variables); versus those who had limited missingness of <30% variables (“low-missingness”, 36 subjects, 174 variables). Cluster analysis was performed initially in the complete-case CAS subset to generate an npEM model.

The mathematical theory underpinning npEM has already been described extensively in other sources [63]. In brief, it involves three steps: 1) an expectation or E-step, which calculates the posterior probability of membership in cluster  $k$ , given the observed dataset, estimated mixing proportions  $\lambda_k$ , and probability distribution for  $k$ ; 2) a maximisation or M-step, which calculates the mixing proportions  $\lambda_k$  from the cluster memberships determined above; 3) a non-parametric kernel density estimation step, which calculates the probability distribution based on a kernel density function for each cluster  $k$  and clustering feature  $j$ . These steps were then iterated until the model converged to a point where log-likelihood values were maximised.

As with any EM algorithm, an initial state must first be set prior to commencing the iterative process. To do this, we used a constant seed state (“set.seed(1)”) to allow reproducibility of results. Based on these pseudo-random centroids for a set number of clusters  $L$ , the initial state was then determined by k-means clustering as in Benaglia et al [63]. The other options in npEM were set to defaults. These included the use of non-stochastic (deterministic) as opposed to a stochastic method; the use of a standard normal density function as the kernel function; and the use of default constant bandwidths for estimating kernel densities [63].

The ideal number of clusters  $L$  was determined by two methods. Firstly, we performed hierarchical clustering on the complete-case dataset, and scrutinised the dendrogram as well as a scree plot for an optimal cut-off using the “knee method” [61]. We observed that this occurred at around  $L = 3$  or  $4$ . Secondly, we repeated npEM clustering for values of  $L = 1, 2, \dots, 20$ , and calculated the Bayesian information criterion (BIC) for each of these, using the formula:

$$\text{BIC} = -2 \log(\hat{P}) + \nu \log(N)$$

where  $\hat{P}$  is the maximum likelihood,  $\nu = L \times M + (L - 1)$ , and  $L, M, N$  are total number of clusters, clustering features, and individuals respectively. The optimal number of clusters was again determined to be around  $L = 3$  or  $4$ , based on minimum BIC observed. For the sake of parsimony, we set the number of clusters to three.

### 3.4.5 Classification of test datasets using mixture model densities

The density functions generated by the resultant npEM model were then used to classify as many subjects of the low-missingness subset as possible. This method relied on the assumption that distributions observed in the “training” (complete-case) dataset were representative of distributions that existed in “test” (low-missingness or external) datasets.

Classification was performed as follows: consider individual  $i$  of  $N$ ; clustering feature or coordinate  $j$  of  $M$ ; and component or cluster  $k$  of  $L$ . For each individual  $i$  belonging to known cluster  $k = \mathbb{K}$ , let the kernel density function for coordinate  $j$  be  $f_{j\mathbb{K}}(x_{ij})$ . We now assume that the coordinates  $j$  were independent of each other. Although this was not truly

the case – for instance, weak correlation exists between IgE and IgG4 of different allergen specificities in the CAS dataset [57] – we believed the assumption was justified given our theory-naïve and exploratory approach. With this assumption, the joint distribution for individual  $i$  in cluster  $\mathbb{K}$  should be the product of density functions for all  $j$  given  $\mathbb{K}$ . and therefore the probability of individual with value  $x_{ij}$  belonging to cluster  $\mathbb{K}$  was:

$$P(k = \mathbb{K} | x_{ij}) = \frac{\lambda_{\mathbb{K}} \prod_{j=1}^M f_{j\mathbb{K}}(x_{ij})}{\sum_{k=1}^L \lambda_k \prod_{j=1}^M f_{jk}(x_{ij})}$$

In addition to this, we made two other assumptions: 1) if  $x_{ij}$  was missing, then the density was assumed to be one, or  $f_{j\mathbb{K}}(x_{ij}) = 1$ ; 2) else, if  $x_{ij} < \min(x_j)$ , the minimum value in feature  $j$  for which there was a non-zero density value, then the density was equal to that of the minimum value, i.e.  $f_{j\mathbb{K}}(x_{ij}) = f_{j\mathbb{K}}(\min(x_j))$ . Likewise, if  $x_{ij} > \max(x_j)$ , then  $f_{j\mathbb{K}}(x_{ij}) = f_{j\mathbb{K}}(\max(x_j))$ .

Individuals with membership probability greater than 90% for cluster  $\mathbb{K}$  were classified into  $\mathbb{K}$ . Using this method, an additional 31 individuals from 36 were successfully classified into one of three clusters, for a total combined dataset of 217 classified individuals in CAS.

Finally, we formally defined each CAS cluster using the composite of complete-case and low-missingness datasets, and described each cluster in terms of key characteristics and significant cluster-specific predictors for age-five wheeze. Importantly, variables that were initially excluded from feature selection were treated as subsequent outcomes for post-hoc comparison of clusters.

### 3.4.6 Replication cohorts

The study designs and measurements for the two replication cohorts – the Manchester Asthma and Allergy Study (MAAS) ( $N = 1085$ ) from Manchester, UK, and the Childhood Origins of Asthma Study (COAST) ( $N = 289$ ) from Wisconsin, USA – have been described elsewhere [19, 29, 30, 64]. COAST, like CAS, was comprised of high-risk individuals with a known family history of asthma or allergy; while MAAS included individuals without family history.

In terms of matching variables for replication, all cohorts had measurements that covered the three major “domains” of asthma pathogenesis: respiratory infection, allergen sensitisation, and clinical or demographic background. COAST had a comprehensive collection of respiratory infection and IgE-type measurements, but no IgG4 measurements. MAAS had multiple measurements of IgE and SPT-type variables. Following consultation with investigators from all three cohorts, clustering features were matched based on proximity of timepoint and phenotype. Respiratory infection phenotypes (ARI, LRI, URI, fLRI, wLRI) were generated in COAST and MAAS using recorded data, to approximate CAS infection phenotypes as closely as possible. Specifically, LRI was defined as respiratory infection with evidence of lower respiratory tract involvement in the form of chest sounds (wheeze, rattle, whistle), or increased respiratory effort (retractions, tachypnea, cyanosis); URI was defined as a cold-like infection limited to the upper respiratory tract, without signs of LRI. IgE and IgG4 assays for MAAS and COAST were performed using ImmunoCAP and UniCAP, respectively. Both replication cohorts recorded basic demographic data, and exposures to pets, childcare, and tobacco smoke. The complete list of clustering features and the matching scheme across cohorts is provided in **Supplementary Table B.1**.

The npEM clusters were described and validated in MAAS and COAST. This replication was performed by applying the density function-derived classifier used previously for the low-missingness CAS subjects. Because these external cohorts did not necessarily

share the same clustering features or variables as CAS ( **Supplementary Table B.1**), we assumed that the respective densities for these variables were  $f_{j\mathbb{K}}(x_{ij}) = 1$  for the  $j^{\text{th}}$  feature and  $\mathbb{K}^{\text{th}}$  cluster. In doing so, this was effectively the same as using a model where the missing features were excluded, and only those features common to both CAS and MAAS (or COAST) were used; or equivalently, where we assumed that each member of MAAS or COAST was missing values in those particular features. Because these “CAS-derived” npEM models were non-identical to the original npEM models in CAS, we tested whether “MAAS-like” and “COAST-like” algorithms (CAS-derived model as applied to MAAS or COAST, respectively) generated similar clusters to the original CAS clusters, when applied back onto CAS (**Results, Section 3.2**).

### 3.4.7 Cluster validity and stability

Internal validation of the clusters in the complete-case CAS dataset was performed by use of silhouette widths. Briefly, we calculated the silhouette widths for each cluster as per Rousseeuw et al [65]. For an individual, the closer the silhouette width is to one, the more appropriate the cluster membership; while the closer it is to negative one, the more likely it has been misclassified.

Cluster stability was assessed by performing leave-one-out (LOO) analysis – that is, we applied the npEM algorithm to a subset of the complete-case dataset – an  $N - 1$  by  $M$  dataset ( $N = 186$ ,  $M = 174$ ) for a total of  $N$  times, leaving out an individual each time. A similar process was repeated  $M$  times on an  $N$  by  $M - 1$  dataset, leaving out one clustering feature at a time. The Jaccard indices for each iteration were then calculated in comparison to known clusters from the original complete-case  $N$  by  $M$  dataset, and averaged across each assigned cluster. Cluster labels for each iteration were assigned based on whichever complete-case cluster yielded the smallest Jaccard index. This whole process was then repeated with 10 random seeds (“set.seed(1)” through to “set.seed(10)”) for determining the initial state for npEM. The final averaged Jaccard indices for each cluster thus represented the mean stability of each cluster.

### 3.4.8 Decision tree analysis

Decision tree analysis was performed using a number of different partitioning schemes. Classification trees with recursive partitioning were built from CAS clusters using the R package “rpart” [66], an open-source implementation of CART. The motivation for decision trees was to identify the variables that most *strongly separated* the clusters and wheezing status, and not necessarily variables that were most predictive.

For tree outcomes (end-nodes), we investigated both cluster membership and presence of age-five wheeze given cluster membership. That is, decision trees were generated to identify the biological features that most strongly distinguished each npEM cluster (“Simple Tree”), as well as npEM cluster  $\times$  age-five wheeze status (“Comprehensive Tree”).

We used two different schemes for selecting predictors on which to base the partitions: 1) include all predictors that were used as clustering features in the original npEM model; 2) include only predictors from one timepoint (variables from age 6m, 1, 2 or 3). The motivation for the latter was that we wanted to see whether measurements taken at a specific timepoint in early infancy could strongly distinguish between clusters. For the former scheme, we excluded all age-five features related to wheeze (e.g. LRIs, wheezy LRIs at age 5) as decision nodes, because of definitional overlap with our primary outcome of interest (age-five wheeze).

Decision trees were then pruned based on the complexity parameter that minimised cross-validated error. Final classification into tree clusters was manually performed based on the pruned tree, and not by automatic classification using the “predict” function for the “rpart” tree object – this was because, for the latter, individuals who are missing key variables were re-classified based on the next best, non-missing, surrogate variable [66]. Thus, it resulted in children being erroneously classified into a tree cluster even when they were missing key classifier variables.

The decision tree analyses generated thresholds which were then compared with existing thresholds for atopy (any specific IgE at age 2  $\geq$  0.35 kU/L, and/or any specific SPT at age 2  $\geq$  2mm) [11] in terms of predicting disease outcomes of interest.

### 3.4.9 Statistical analyses

We performed statistical analyses comparing clusters in terms of multiple variables, especially those not used as clustering features. Of interest to us were the primary outcomes of asthma diagnosis and parent-reported wheeze at each timepoint. Where appropriate, we used t-tests, Mann-Whitney-Wilcoxon tests, ANOVAs, Kruskal-Wallis tests, chi-squared and Fisher exact tests; and logistic and linear regression. For summary statistics, multiple testing adjustment was performed using the Benjamini-Yekutieli (BY) method, for all across-cluster tests (Cluster  $\times$  trait); and for all comparisons between clusters (CAS1 vs. 2, 1 vs. 3, and 2 vs. 3). The BY method was chosen as it accounted for positive dependency across the highly-correlated variables in the CAS dataset [67]. For variables that underwent logarithmic transformation for statistical analysis, we used geometric mean to describe central tendency.

We then determined the predictors for age-five wheeze within each cluster. Repeated-measures ANOVAs were performed for selected predictors of age-five wheeze. For each potential predictor, generalised linear regression models (GLMs) were generated with and without a base set of covariates (sex, family history of asthma, BMI where available). The pool of variables found to be statistically-significant (at least  $p < 0.05$ ) in the above analyses were further restricted, such that strongly-collinear predictors were avoided, and at most one timepoint was considered for each predictor type. Targeted multiple regression models were then built by selecting predictors from this constrained pool. Stepwise backward elimination was applied, in which the predictor with the largest  $p$ -value was eliminated at each step, until all remaining predictors have significant  $p < 0.05$ .

Using the “lrtest” function from the R package “Epidisplay” [68], likelihood ratios were examined to check how much cluster membership or classification improved upon prediction of age-five wheeze compared to traditional makers of atopy.

## References

1. Global Initiative for Asthma. Global Strategy for Asthma Management and Prevention. 2015. URL: [https://ginasthma.org/wp-content/uploads/2016/01/GINA\\_Report\\_2015\\_Aug11-1.pdf](https://ginasthma.org/wp-content/uploads/2016/01/GINA_Report_2015_Aug11-1.pdf).
2. Ober C and Yao TC. The genetics of asthma and allergic disease: a 21st century perspective. *Immunol Rev* 2011;242:10–30.
3. Dick S, Friend A, Dynes K, et al. A systematic review of associations between environmental exposures and development of asthma in children aged up to 9 years. *BMJ Open* 2014;4:e006554.
4. Okada H, Kuhn C, Feillet H, and Bach JF. The ‘hygiene hypothesis’ for autoimmune and allergic diseases: an update. *Clin Exp Immunol* 2010;160:1–9.

5. Morgan WJ, Stern DA, Sherrill DL, et al. Outcome of asthma and wheezing in the first 6 years of life: follow-up through adolescence. *Am J Respir Crit Care Med* 2005;172:1253–8.
6. Spycher BD, Silverman M, and Kuehni CE. Phenotypes of childhood asthma: are they real? *Clin Exp Allergy* 2010;40:1130–41.
7. Hekking PP and Bel EH. Developing and emerging clinical asthma phenotypes. *J Allergy Clin Immunol Pract* 2014;2:671–80, quiz 681.
8. Wenzel SE. Asthma phenotypes: the evolution from clinical to molecular approaches. *Nat Med* 2012:716.
9. Anderson GP. Endotyping asthma: new insights into key pathogenic mechanisms in a complex, heterogeneous disease. *Lancet* 2008;372:1107–19.
10. Castro-Rodriguez JA, Holberg CJ, Wright AL, and Martinez FD. A clinical index to define risk of asthma in young children with recurrent wheezing. *Am J Respir Crit Care Med* 2000;162:1403–6.
11. Frith J, Fleming L, Bossley C, Ullmann N, and Bush A. The complexities of defining atopy in severe childhood asthma. *Clin Exp Allergy* 2011;41:948–53.
12. Linden CC, Misiak RT, Wegienka G, et al. Analysis of allergen specific IgE cut points to cat and dog in the Childhood Allergy Study. *Ann Allergy Asthma Immunol* 2011;106:153–158 e2.
13. Prescott SL, Macaubas C, Smallacombe T, Holt BJ, Sly PD, and Holt PG. Development of allergen-specific T-cell memory in atopic and normal children. *Lancet* 1999;353:196–200.
14. Martinez FD, Wright AL, Taussig LM, Holberg CJ, Halonen M, and Morgan WJ. Asthma and wheezing in the first six years of life. The Group Health Medical Associates. *N Engl J Med* 1995;332:133–8.
15. Kurukulaaratchy RJ, Fenn MH, Waterhouse LM, Matthews SM, Holgate ST, and Arshad SH. Characterization of wheezing phenotypes in the first 10 years of life. *Clin Exp Allergy* 2003;33:573–8.
16. Deliu M, Sperrin M, Belgrave D, and Custovic A. Identification of Asthma Subtypes Using Clustering Methodologies. *Pulmonary Therapy* 2016;2:19–41.
17. Lazic N, Roberts G, Custovic A, et al. Multiple atopy phenotypes and their associations with asthma: similar findings from two birth cohorts. *Allergy* 2013;68:764–70.
18. Simpson A, Tan VY, Winn J, et al. Beyond atopy: multiple patterns of sensitization in relation to asthma in a birth cohort study. *Am J Respir Crit Care Med* 2010;181:1200–6.
19. Belgrave DC, Granell R, Simpson A, et al. Developmental profiles of eczema, wheeze, and rhinitis: two population-based birth cohort studies. *PLoS Med* 2014;11:e1001748.
20. Belgrave DC, Simpson A, Semic-Jusufagic A, et al. Joint modeling of parentally reported and physician-confirmed wheeze identifies children with persistent troublesome wheezing. *J Allergy Clin Immunol* 2013;132:575–583 e12.
21. Wu J, Prosperi MCF, Simpson A, et al. Relationship Between Cytokine Expression Patterns and Clinical Outcomes: Two Population-based Birth Cohorts. *Clinical & Experimental Allergy* 2015;45:1801–11.
22. Okamoto S, Taniuchi S, Sudo K, et al. Predictive value of IgE/IgG4 antibody ratio in children with egg allergy. *Allergy Asthma Clin Immunol* 2012;8:9.

23. Holt PG, Strickland D, Bosco A, et al. Distinguishing benign from pathologic TH2 immunity in atopic children. *J Allergy Clin Immunol* 2016;137:379–87.
24. Bantz SK, Zhu Z, and Zheng T. The Atopic March: Progression from Atopic Dermatitis to Allergic Rhinitis and Asthma. *J Clin Cell Immunol* 2014;5.
25. Han H, Roan F, and Ziegler SF. The atopic march: current insights into skin barrier dysfunction and epithelial cell-derived cytokines. *Immunol Rev* 2017;278:116–130.
26. Teo SM, Mok D, Pham K, et al. The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development. *Cell Host Microbe* 2015;17:704–15.
27. Bisgaard H, Hermansen MN, Buchvald F, et al. Childhood asthma after bacterial colonization of the airway in neonates. *N Engl J Med* 2007;357:1487–95.
28. Teo SM, Tang HH, Mok D, et al. Dynamics of the upper airway microbiome in the pathogenesis of asthma-associated persistent wheeze in preschool children. *bioRxiv* 2017.
29. Custovic A, Simpson BM, Murray CS, et al. The National Asthma Campaign Manchester Asthma and Allergy Study. *Pediatr Allergy Immunol* 2002;13 Suppl 15:32–7.
30. Lemanske R. F. J. The childhood origins of asthma (COAST) study. *Pediatr Allergy Immunol* 2002;13 Suppl 15:38–43.
31. Bousquet J, Anto JM, Wickman M, et al. Are allergic multimorbidities and IgE polysensitization associated with the persistence or re-occurrence of foetal type 2 signalling? The MeDALL hypothesis. *Allergy* 2015;70:1062–78.
32. Hose AJ, Depner M, Illi S, et al. Latent class analysis reveals clinically relevant atopy phenotypes in 2 birth cohorts. *J Allergy Clin Immunol* 2016;139:1935–45.
33. Almqvist C, Worm M, Leynaert B, and working group of GALENWPG. Impact of gender on asthma in childhood and adolescence: a GA2LEN review. *Allergy* 2008;63:47–57.
34. Gabet S, Just J, Couderc R, Seta N, and Momas I. Allergic sensitisation in early childhood: Patterns and related factors in PARIS birth cohort. *Int J Hyg Environ Health* 2016;219:792–800.
35. Kim HY, Shin YH, and Han MY. Determinants of sensitization to allergen in infants and young children. *Korean J Pediatr* 2014;57:205–10.
36. Thomas WR, Hales BJ, and Smith WA. House dust mite allergens in asthma and allergy. *Trends Mol Med* 2010;16:321–8.
37. Calderon MA, Linneberg A, Kleine-Tebbe J, et al. Respiratory allergy caused by house dust mites: What do we really know? *J Allergy Clin Immunol* 2015;136:38–48.
38. Vadas P, Wai Y, Burks W, and Perelman B. Detection of peanut allergens in breast milk of lactating women. *JAMA* 2001;285:1746–8.
39. DesRoches A, Infante-Rivard C, Paradis L, Paradis J, and Haddad E. Peanut allergy: is maternal transmission of antigens during pregnancy and breastfeeding a risk factor? *J Investig Allergol Clin Immunol* 2010;20:289–94.
40. Brough HA, Santos AF, Makinson K, et al. Peanut protein in household dust is related to household peanut consumption and is biologically active. *J Allergy Clin Immunol* 2013;132:630–8.

41. Brough HA, Simpson A, Makinson K, et al. Peanut allergy: effect of environmental peanut exposure in children with filaggrin loss-of-function mutations. *J Allergy Clin Immunol* 2014;134:867–875 e1.
42. Koplin JJ, Osborne NJ, Wake M, et al. Can early introduction of egg prevent egg allergy in infants? A population-based study. *J Allergy Clin Immunol* 2010;126:807–13.
43. Stapel SO, Asero R, Ballmer-Weber BK, et al. Testing for IgG4 against foods is not recommended as a diagnostic tool: EAACI Task Force Report. *Allergy* 2008;63:793–6.
44. Davies AM and Sutton BJ. Human IgG4: a structural perspective. *Immunol Rev* 2015;268:139–59.
45. Aalberse R. The role of IgG antibodies in allergy and immunotherapy. *Allergy* 2011;66 Suppl 95:28–30.
46. Kusel MM, de Klerk NH, Kebabze T, et al. Early-life respiratory viral infections, atopic sensitization, and risk of subsequent development of persistent asthma. *J Allergy Clin Immunol* 2007;119:1105–10.
47. Holt PG, Rowe J, Kusel M, et al. Toward improved prediction of risk for atopy and asthma among preschoolers: a prospective cohort study. *J Allergy Clin Immunol* 2010;125:653–9, 653–9.
48. Holt PG and Sly PD. Viral infections and atopy in asthma pathogenesis: new rationales for asthma prevention and treatment. *Nat Med* 2012;18:726–35.
49. Janssens T, Verleden G, and Van den Bergh O. Symptoms, lung function, and perception of asthma control: an exploration into the heterogeneity of the asthma control construct. *J Asthma* 2012;49:63–9.
50. Newby C, Heaney LG, Menzies-Gow A, et al. Statistical cluster analysis of the British Thoracic Society Severe refractory Asthma Registry: clinical outcomes and phenotype stability. *PLoS One* 2014;9:e102987.
51. Burte E, Bousquet J, Varraso R, et al. Characterization of Rhinitis According to the Asthma Status in Adults Using an Unsupervised Approach in the EGEA Study. *PLoS One* 2015;10:e0136191.
52. Sarstedt M and Mooi E. *A Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics*. Springer Berlin Heidelberg, 2014. URL: <https://books.google.com.au/books?id=r5QqBAAAQBAJ>.
53. Schoos AM, Chawes BL, Melen E, et al. Sensitization trajectories in childhood revealed by using a cluster analysis. *J Allergy Clin Immunol* 2017.
54. Chen JH and Asch SM. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. *N Engl J Med* 2017;376:2507–2509.
55. Klink M, Cline MG, Halonen M, and Burrows B. Problems in defining normal limits for serum IgE. *J Allergy Clin Immunol* 1990;85:440–4.
56. Kusel MM, Holt PG, de Klerk N, and Sly PD. Support for 2 variants of eczema. *J Allergy Clin Immunol* 2005;116:1067–72.
57. Hollams EM, Deverell M, Serralha M, et al. Elucidation of asthma phenotypes in atopic teenagers through parallel immunophenotypic and clinical profiling. *J Allergy Clin Immunol* 2009;124:463–70, 463–70.
58. Hollams EM, Teo SM, Kusel M, et al. Vitamin D over the first decade and susceptibility to childhood allergy and asthma. *J Allergy Clin Immunol* 2016;139:472–81.

59. Ballardini N, Nilsson C, Nilsson M, and Lilja G. ImmunoCAP Phadiatop Infant—a new blood test for detecting IgE sensitisation in children at 2 years of age. *Allergy* 2006;61:337–43.
60. Benaglia T, Chauveau D, Hunter DR, and Young DS. mixtools: An R Package for Analyzing Mixture Models. *Journal of Statistical Software, Articles* 2009;32:29.
61. Tan PN, Kumar V, and Steinbach M. Introduction to data mining. Boston: Pearson Addison Wesley, 2005.
62. Gustafsson D, Sjoberg O, and Foucard T. Development of allergies and asthma in infants and young children with atopic dermatitis—a prospective follow-up to 7 years of age. *Allergy* 2000;55:240–5.
63. Benaglia T, Chauveau D, and Hunter DR. An EM-Like Algorithm for Semi- and Nonparametric Estimation in Multivariate Mixtures. *Journal of Computational and Graphical Statistics* 2009;18:505–526.
64. Gern JE, Martin MS, Anklam KA, et al. Relationships among specific viral pathogens, virus-induced interleukin-8, and respiratory symptoms in infancy. *Pediatr Allergy Immunol* 2002;13:386–93.
65. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 1987;20:53–65.
66. Therneau TM and Atkinson EJ. An Introduction to Recursive Partitioning Using the RPART Routines. 2015. URL: <https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf>.
67. Benjamini Y and Yekutieli D. The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics* 2001;29:1165–1188.
68. Chongsuvivatwong V. epiDisplay: Epidemiological Data Display Package. 2015. URL: <https://cran.r-project.org/web/packages/epiDisplay/>.



## Chapter 4

# Diverging trajectories of nasopharyngeal microbiome during early childhood are associated with asthma and asthma-related traits

### 4.1 Introduction

There is increasing evidence for the involvement of host and environmental microbiota in the pathogenesis of asthma and childhood wheeze. The original “hygiene hypothesis”, formulated in 1989 by Strachan [1], has been largely supplanted by a “microflora” or “microbiota hypothesis”. This revised hypothesis suggests that the narrative is not as simple as “over-sanitation causes allergy” – but that changes to the living environment not related to sanitation or hygiene may also be driving changes to microbial exposure [2–4]. There is evidence that the impact of environmental microbiota on asthma or allergy development may be dependent on the timing of exposure, or that it may be modulated by genetics and other exposures [3]. Furthermore, early-life exposure to respiratory viruses and alterations to host microbiota play important roles in determining respiratory health [5, 6]. However, it remains unclear how viral and bacterial microbiota interact, with each other or with other features of host and environment, to elicit health or disease. There is also emerging interest in the host microbiome, not just in terms of microbial compositions at a fixed timepoint, but also in terms of its evolution and transformation as the host ages. A particular gap in our knowledge is the characterisation of certain patterns of change, or trajectories, that may influence the host towards respiratory health or disease

The host microbiome is the entire genomic content of a microbiotic sample from a host, typically at a host-environment interface such as the skin, gut or respiratory tract. As such, it produces large complex datasets that are challenging to process and analyse. To address this, techniques to analyse microbiome data have evolved rapidly within the last decade. Improvements have been made not just in sequencing technology, but also in quality control and extrapolation of reliable information from read sequences. For instance, Quantitative Insights Into Microbial Ecology (QIIME) is a bioinformatics platform for processing and analysing microbiome sequence data [7]. The developers of QIIME have recently transitioned to a new version (QIIME2) [8] that reimplements many features from the old platform, but distinctively facilitates and streamlines the use of denoising and quality control packages such as DADA2 [9]. DADA2 corrects for common sequencing errors, accurately inferring sequences down to a single nucleotide, and represents a move towards amplicon sequence variants (ASVs), which are effectively 100%-identity OTUs [10]. This produces read data with better biological interpretability and broader applicability across datasets compared to older “<100%” OTU-picking methods [10],

which rely on clustering of queried sequences into OTUs based on similarity to sequences in a reference database such as Greengenes [11]. Pipelines that call exact sequence variants may yield biological findings that build on those already discovered with traditional OTU-picking methods – for example, we can identify new taxa and compositional patterns, or more accurately pinpoint existing ones, that can then be linked with a particular health or disease state.

For this study we accessed two comprehensive birth cohorts with similar study designs – the Childhood Asthma Study (CAS) [12] and the Childhood Origins of Asthma Study (COAST) [13]. Both cohorts yielded nasopharyngeal samples which were subjected to microbiome-scale analyses. One of these cohorts had already been thoroughly investigated using the older bioinformatic pipeline [14, 15]. Given this context, the aims of this current study were four-fold: 1) to apply a QIIME2, DADA2/ASV-based pipeline to CAS and COAST, determine profiles of nasopharyngeal microbiome composition, and compare between the two cohorts; 2) to build on the results of Teo et al 2018 [15] by conducting a meta-analysis of associations between microbial and asthma-related traits, using both CAS and COAST data; 3) to determine microbiome trajectories in CAS and COAST using cluster analysis, and compare between cohorts; and 4) to describe how these trajectories relate to asthma-related traits, together with other pathophysiologically-relevant factors.

## 4.2 Methods

### 4.2.1 Study design and overview of measurements

The study employed data from two birth cohorts – the Childhood Asthma Study (CAS) ( $N = 263$ ) from Perth, Western Australia [12], and the Childhood Origins of Asthma Study (COAST) ( $N = 289$ ) from Wisconsin, USA [13]. Details on recruitment and sampling within each cohort have been described elsewhere [16, 17]; the following paragraphs provide a focused summary, with an emphasis on microbiome-related sampling and measurements, as well as phenotyping of respiratory infections. **Table 4.1** summarises the key differences between the two cohorts.

The CAS population was comprised of 263 children of primarily Caucasian ethnicity, recruited from private paediatric clinics around Perth, between July 1996 and June 1998. Children were only included if they had at least one parent with diagnosed asthma or other allergic disease (hayfever, eczema). Each child underwent routine collection of nasopharyngeal samples via aspirate or swab, at timepoints of 2m, 6m, 1y, then every half-year up to age 4y. Further non-routine samples were taken during any episode of clinician-confirmed respiratory infection up to age five, and the symptoms experienced during these infections were recorded.

The COAST cohort consisted of 283 children mostly of Caucasian or African-American ethnicity, recruited from Madison, Wisconsin via the University of Wisconsin Hospital and affiliated clinics, from November 1998 to May 2000. As for CAS, all recruited children had at least one parent with diagnosed allergic disease. Routine nasopharyngeal sampling was performed for children in COAST at timepoints of 2m, 4m, 6m, 9m, 1y, 1.5y and 2y. Unlike CAS, non-routine samples up to age three were collected only if the child had a symptom score exceeding a predefined threshold of five [16]. High scores were based primarily on symptoms and signs of severe respiratory illness – such as wheeze, chest retraction, dyspnoea or tachypnoea, and cyanosis. Because of this, non-routine samples in COAST were overrepresented by relatively severe infections. Some routine samples in COAST were also collected from mildly-unwell individuals with symptom scores  $<5$ , while in CAS all routine samples were strictly from well individuals.

TABLE 4.1: Key differences between CAS and COAST birth cohorts

	CAS	COAST
Population	263 children from Perth, Western Australia, Australia; each with at least one parent with history of asthma or allergic disease	283 children from Madison, Wisconsin, USA; each with at least one parent with history of asthma or allergic disease
Time range of birth dates	July 1996 - June 2003	November 1998 - May 2000
Ethnicity	Caucasian	Caucasian ( $N = 246, 87\%$ ), African-American ( $N = 11, 4\%$ ), mixed or other ( $N = 26, 9\%$ )
Number of samples sequenced	3439 across first 5yrs of life	3146 across first 3yrs of life
Number of samples after QC	3120	2926 (some episodes had multiple samples, and some samples had multiple extractions)
Timing of routine (usually healthy) samples	Routine sampling at ages 2m, 6m, 1y, 1.5, 2, 2.5, 3, 3.5, 4y; only collected if patient was healthy at routine follow-up	Routine sampling at ages 2m, 4m, 6m, 9m, 1y, 1.5, 2y; some routine samples may have been accompanied by mild infectious or respiratory symptoms
Timing of non-routine (sick / infectious) samples	Non-routine sampling of any episode of respiratory infection, confirmed by clinician, up to age 5y	Non-routine sampling of any episode of respiratory infection, up to age 3y; however, most of these were sampled only if symptom score $\geq 5$
Definition of harmonised infection phenotypes	<b>URI:</b> cough or rhinorrhea, without wheeze or rattle (i.e. not LRI); <b>LRI:</b> wheeze or rattle	<b>URI:</b> cough or rhinorrhea, but not LRI; <b>LRI:</b> wheeze or cyanosis or retractions or tachypnea (i.e. symptom score $\geq 5$ )
Viral typing	Performed only on samples from age 0 to 3y, by different laboratory (Johnston et al for Year 1; Gern et al for Years 2-3)	Performed for most samples
Rhinovirus subtyping	Performed only in LRI samples (Year 1) or in LRI samples that initially tested positive for RV (Years 2-3)	Performed for most samples

Children from both cohorts also had routine medical examinations and blood sampling performed, and numerous demographic, serological, and clinical measurements were conducted. CAS children were followed up to maximum age 10, while the maximum age of follow-up for COAST was 16 years. Variables common to both cohorts included: formal asthma diagnosis, respiratory infection severity and frequency; allergen-specific IgE, family history of allergic disease, number of older siblings, and environmental exposures (to tobacco smoke, childcare, pets) [17]. In CAS, we measured specific IgE against cat, couch grass, house-dust mite (*Dermatophagoides pteronyssinus*), mould, peanut, and ryegrass; in COAST, IgE against cat, dog, *D. pteronyssinus*, *D. farinae*, and *Alternaria* were measured. Also, specific respiratory infection phenotypes were defined in CAS and COAST such that they were comparable to CAS: an infection was defined as lower respiratory (LRI) if there was wheeze or rattle in CAS; or if there was wheeze, retractions, dyspnoea/tachypnoea, or cyanosis in COAST (Table 4.1). LRIs could be wheezy (wLRI), febrile (fLRI), or severe (wheezy or febrile, sLRI). An upper respiratory infection (URI) was defined in both cohorts if there was cough or rhinorrhea, without any other signs of LRI.

#### 4.2.2 Bacterial 16S profiling of nasopharyngeal samples, and annotation of OTUs and ASVs

We performed 16S rRNA amplicon sequencing of nasopharyngeal samples (swab or aspirate) using Illumina MiSeq (San Diego, US). Paired reads were sequenced from the 16S V4 subregion with 515F/806R primers. Each paired read (excluding primers and barcodes) was ~150 base pairs (bp) in length with ~46 bp overlap with its partner [18]. We applied the same sample-processing and sequencing protocol to both CAS and COAST samples; however, the cohorts were sequenced and processed separately across two distinct time periods i.e. CAS and COAST samples were not intermingled on the same sequencer. Also, for some COAST samples, we repeated runs or amplifications if the original run was of poor quality (with systematic, ambiguous base calls) or of very low yield.

Based on the CAS dataset, we compared results from our old pipeline using QIIME1 (v1.7) and closed-reference operational taxonomic unit (OTU) picking, with the new results using QIIME2 (v2017.10/12) and DADA2 with amplicon sequence variant (ASV) calling. Details of the old pipeline were described in Teo et al 2018 [15], and is summarised as follows: reads were merged; filtered by a set of quality criteria ( $\leq 3$  low-quality base pairs (bp),  $\geq 189$  consecutive high-quality bp, no N characters); and clustered into OTUs against Greengenes 99% 16S rRNA reference (v13\_05), for which OTUs with the same 16S V4 region sequence were merged as one OTU. Read counts were corrected for OTU-specific copy number. Reads that could not be matched to a Greengenes taxon were ignored, and samples with  $< 3000$  taxonomy-assigned reads were excluded. OTUs from CAS were named with the smallest Greengenes taxon at genus or above, followed by the Greengenes identifier (e.g. *Alloiococcus*.OTU886735).

Both CAS and COAST samples were processed using the new pipeline, albeit separately by cohort. Details of the new pipeline are summarised in **Supplementary Figure C.1**. In QIIME2, we determined a filtering and trimming protocol (5'-end trimmed at 10 bp, 3'-end truncated at 150 bp, for both CAS and COAST samples) based on visual inspection of average read quality by base pair length, then applied to the paired reads via DADA2. Subsequent joined reads were around 234 bp in length (both reads sum to ~300 bp, with 46 bp overlap, and 20 bp trimmed). We then used DADA2 to generate error models from the quality data from each run, which were then used to denoise, correct and merge paired sequences into ASVs. Finally, we used DADA2 to remove chimeras with the default consensus method.

Each unique ASV sequence was assigned a unique 32-digit string identifier, which was comparable across different runs and datasets. In addition, reads were assigned a taxonomy using a naïve Bayes classifier trained on the 16S V4 (515F/806R) region of Greengenes 99% OTU reference (v13\_08). Unlike the old OTU-based pipeline, reads without taxonomy assignment were *not* excluded. No correction was performed for ASV-specific copy number, because prediction of copy number information remains challenging for both ASVs and OTUs, especially those without pre-assigned Greengenes taxonomy [19]. For simplicity, each ASV was annotated with the lowest Greengenes taxon at genus level or above, followed by a suffix of the first 4 digits of its 32-digit identifier (e.g. *Alloiococcus*.dd2e). The exact FASTA sequence of each ASV was also parsed through the NCBI database using BLAST [20] to identify likely or potential candidates at the species level.

All data processing subsequent to the QIIME2 steps, as well as statistical analyses, were run on R v3.5.0 unless otherwise specified. Management of microbiome data was performed primarily with the “phyloseq” [21] and “microbiome” R packages [22]. An additional quality control pipeline was implemented within R as described in the lower half of **Supplementary Figure C.1**. For COAST extractions that came from the same

episode or were re-runs of the same amplified library (i.e. not reamplified), the reads were combined to give an aggregate sample per episode per individual. Further quality control measures were implemented in R as per Teo et al 2018 [15]: for each dataset, we removed samples with read counts less than a threshold that excluded ~80% negative controls (for CAS, <3000 reads; for COAST, <4000 reads). Finally, we ensured that samples were only categorised as “well” or “healthy” if they had no preceding illness episode for four weeks prior; otherwise they were excluded from analysis.

Certain ASVs were defined as “common” using pre-existing criteria applied to OTUs in Teo et al 2018 [15]: having mean relative abundance >0.1% across all samples; present in >20% samples; and dominating (>50%) at least one sample. “Non-rare” ASVs were defined as those present (with at least one read) in >1% samples. These subsets of ASVs were used for subsequent analyses: generation of microbiome profile groups (MPGs) and FastSpar correlation analyses, for “common” and “non-rare” respectively.

### 4.2.3 Comparison of CAS microbiome data generated with old versus new pipelines

The new ASV-based pipeline (Methods, Section 4.2) was applied to the CAS dataset, producing 3120 quality-controlled samples with read data for 23441 distinct ASVs. This yielded more samples and fewer unique taxa than the old OTU-based formulation of the CAS dataset (3014 samples and 28230 OTUs as per Teo et al 2018 [15]).

The overall results generated with the new pipeline remain similar to the old pipeline. Within CAS, the common QIIME1 OTUs and QIIME2 ASVs (as defined in the previous section) were concordant (Supplementary Table C.1), mapping to shared sequences of ~230 bp in length. Some common QIIME1 OTUs matched with multiple QIIME2 ASVs; however, there was always one “core” ASV which comprised the majority of that OTU, and which was also common in the QIIME2 dataset — we determined that these were analogous. Two of the 13 common ASVs in QIIME2 did not have a common analogous QIIME1 OTU (Supplementary Table C.2): these were Gemellaceae.d800 and *Escherichia.d2a4*. Relative abundances of common OTUs and their analogous common ASVs were roughly similar (Supplementary Table C.1, Supplementary Figure C.2A and B).

The remainder of this paper concerns results based on the new QIIME2 pipeline (CAS and COAST). The results from CAS QIIME1 have been described in Teo et al 2018 [15].

### 4.2.4 Generation of microbiome profile groups (MPGs)

The nasopharyngeal samples were clustered into microbiome profile groups (MPGs) using the method described in Teo et al 2015 and 2018 [14, 15]. This was performed separately for each cohort and pipeline. In brief: hierarchical clustering (complete-linkage, Bray-Curtis dissimilarity as distance metric) was conducted using R function “hclust”, on the relative abundance data of a reduced subset of features. This reduced feature set consisted of the following:

1. ASVs common in either CAS or COAST cohorts. The motivation for the latter was to generate MPGs that had better comparability between CAS and COAST.
2. Rarer ASVs that were aggregated into one of seven major genera or families (*Moraxella*, *Streptococcus*, *Haemophilus*, *Alloiococcus*, *Corynebacterium*, *Staphylococcus*, and Moraxellaceae family). These consisted of all other ASVs that belonged to a particular genus or family, but were not “common”. These were annotated with a suffix “.rare”

3. The aggregated read count of all other rare ASVs. This was named “others.rare”, and comprised the remaining reads such that the sum relative abundances of all considered features was equal to one for each sample.

The number of clusters or MPGs was chosen based on maximum average silhouette width. MPGs were named after the dominant OTU or ASV, where applicable. Cluster validity of MPGs was measured using average silhouette values and Bray-Curtis dissimilarity within and between MPG groups, using the “vegan” R package. Note that all quality-controlled samples from CAS (up to and including age 4) and COAST (up to and including age 2) were used for this analysis.

#### 4.2.5 Virus detection

In both CAS and COAST, nasopharyngeal samples up to age three were assessed for viral presence using reverse transcriptase polymerase chain reactions (rt-PCR). The materials and methods for this procedure have been described in previous publications [23–25]. A general screen was performed for common viral pathogens, including human rhinoviruses (RV), respiratory syncytial virus (RSV), and influenza. First-year CAS samples were analysed using assays and experimental conditions [23] which differed from the method used for second- and third-year CAS samples and all COAST samples (Respiratory MultiCode Assay, [24]) (Table 4.1). Some samples (LRI CAS samples, non-routine COAST samples) were further screened for RV subtypes (A,B,C) [25].

#### 4.2.6 Correlation among ASVs

We used FastSpar [26], an efficient implementation of SparCC [27] that calculates correlations in compositional data and evaluates statistical significance. Unlike traditional methods for correlation analysis, FastSpar and SparCC corrects for biases in compositional data that obscure true correlations and generate false ones. Further details of SparCC are found in Friedman et al 2012 [27]. As per Teo et al 2018 [15], we calculated the correlation among all ASVs that were sufficiently “non-rare” (present in >1% samples), using uncorrected ASV read counts. However, for this paper, we only present the results of significant correlations between the common ASVs. Statistical significance was determined based on bootstrap correlations from 1000 random permutations of the data, with significance level taken at 0.001.

#### 4.2.7 Diversity measures and ecological analyses

Alpha (within-sample) diversity was assessed using Shannon’s diversity index measure, which takes into account both the number of unique ASVs and their relative abundances. This was calculated using the in-built “diversities” function of the “microbiome” package [22]. GEE and GLM analyses were performed associating alpha diversity with age and illness status of sample, within each cohort (CAS, COAST).

#### 4.2.8 Dimension reduction and clustering into microbiome trajectories

We determined clusters of individuals who shared similar patterns of changing microbiome (“microbiome trajectories”) during healthy (asymptomatic, “baseline”) states. For both cohorts, we used the relative abundances of common ASVs from healthy routine nasopharyngeal samples, across the first two years of life, to determine distinct trajectories of change. As we had restricted ourselves to routine samples only, each subject usually yielded one sample at each timepoint of collection; if present, multiple samples were

averaged. The list of common ASVs was identical to that used to generate MPGs (i.e. including genera-level, family-level, and others.rare ASV groups). We created a relative abundance-per-timepoint matrix, with subjects as rows, and ASV  $\times$  timepoint as column variables. The 1.5y timepoint was excluded due to excessive missingness in both cohorts. For the remaining variables with missing values, mean imputation was conducted.

Then, we performed dimension reduction on this matrix using Multiple Factor Analysis (MFA) from the R package “FactoMineR” [28]. This method is similar to principal components analysis (PCA), except it considers user-defined groupings (e.g. timepoint of sample collection; e.g. 2m, 4m, up to and including 2y) of ASV  $\times$  timepoint variables while determining the dimensions. A standard PCA would have treated ASV abundances within each timepoint independently of each other, while MFA accounted for shared timepoints when weighting each variable to generate the dimensions. From the dimensions produced by MFA, we extracted the number of dimensions that accounted for at least 80% variance of the original dataset, and used this as the basis for clustering into trajectories.

K-means clustering was performed on the dimension-reduced dataset. We used the function “KMeans\_rcpp” function from R package “ClusterR (100 initialisations, 100 iterations, random seed for initialisation set at 1), which allows for selection of optimal initialisation centroids based on the best within-cluster sum-of-squares error (SSE) [29]. The number of clusters  $K$  was chosen by supervised judgement, based on: maximum average silhouette width from multiple hierarchical clusterings (Ward, Euclidean distance metric) with increasing  $K$ ; and the results of the function “Optimal\_Clusters\_Kmeans” from “ClusterR” which uses silhouette width, within-cluster SSE and Bayesian Information Criterion (BIC) to determine best  $K$ .

#### 4.2.9 Association analyses and meta-analyses

Basic statistical tests were used to compare MPGs and MFA+K-means trajectories — Fisher exact tests and Chi-square tests were used for categorical variables; while Kruskal-Wallis, t-tests and ANOVAs were used for continuous variables.

We then performed statistical analyses looking for associations between the newly-derived MPGs and important outcomes related to asthma and childhood respiratory health, including: healthy or ill respiratory status at the time of sample collection (i.e. absence or presence of respiratory infection); severity and symptoms of said infection; presence of virus during infection; measures of ecological diversity; early allergic sensitisation; and wheeze phenotype (early/transient, late, persistent). Most of these associations were performed as generalised estimating equation (GEE) models, adjusting for each child as subject, and gender, age and season as covariates (the latter two as repeated measures). Each MPG was modelled independently. Similar analyses were also performed at the ASV level for all common ASVs; for different diversity measures; and for trajectories of changing microbiome compositions (as described by the MFA+K-means results).

Due to the compositional nature of microbiome data, we expected the GEE models to give overestimated and biased results for analyses of relative abundance (i.e. ASV). Therefore, further ASV-level association analyses were performed using zero-inflated Gaussian mixture models, with the function fitZIG from the R package “metagenomeSeq” [30] — previously performed in Teo et al 2018 with the old pipeline data [15]. This method accounts for biases in differential analyses that result from possible undersampling, but, unlike GEE, does not account for subjects with repeated measures across timepoints. We only analysed those ASVs which were common (as defined previously). Furthermore, we applied log-transformation and cumulative-sum-scaling (CSS) to ASV read counts prior to analysis. Outcomes of interest were as described above, with special attention given to differences between healthy or illness status at the time of sample collection.

We also analysed for differences in microbial composition between cohorts (CAS versus COAST), although interpretation of these results was left open, due to the possibility of batch differences and unexplained confounders between cohorts.

For associations with respiratory health and asthma-related traits, meta-analyses (random-effects, inverse variance weights) were performed across CAS and COAST, using the function “metagen” from the R package “meta” [31]. Meta-analysis of correlations was performed with Fisher’s z transformation using the function “metacor” from the same package.

## 4.3 Results

### 4.3.1 Composition of the nasopharyngeal microbiome in CAS and COAST children

The QIIME2 pipeline was applied to the COAST dataset, producing 2922 quality-controlled samples with 12464 ASVs. There were substantially fewer unique ASVs in COAST compared to CAS (3120 samples, 23441 ASVs); this might be related to differences between batches, experimental procedure, geography, or some other uncontrolled difference. Nonetheless, there was significant overlap amongst the top common ASVs in either cohort (**Supplementary Table C.2**). In particular, both cohorts had the same common taxa belonging to the top six genera; these were named according to their Greengenes annotations as: *Alloioccus*.dd2e, *Corynebacterium*.cb50, *Haemophilus*.bc0d, *Haemophilus*.f579, *Moraxella*.d253, *Streptococcus*.4060, *Staphylococcus*.29eb, and *Streptococcus*.3575. According to NCBI BLAST, the sequences of each taxa most closely match: *Alloioccus otitis*, *Corynebacterium pseudodiphtheriticum*, subtypes of *Haemophilus influenzae*, *Moraxella catarrhalis*, *Streptococcus pneumoniae*, numerous species of *Staphylococcus*, and *Streptococcus mitis*, respectively (**Supplementary Table C.2**). Both cohorts shared the same top 3 ASVs (*Moraxella*.d253, *Streptococcus*.4060 and *Alloioccus*.dd2e) (**Table C.2**). Relative abundance of other ASVs were comparable across both cohorts (**Supplementary Table C.2, Supplementary Figure C.2A and B**), although there were some substantial differences: *Pseudomonas*.0925 was common in CAS not COAST; while *Moraxellaceae*.a5a0, *Neisseriaceae*.03f4, *Streptococcus*.b069, *Streptococcus*.be1b, and *Veillonella*.fb81 were common in COAST not CAS. Notably, although, *Streptococcus*.3575 was a common taxon in both CAS and COAST, it was substantially more so in COAST, and it comprised its own early-life microbiome trajectory (see **Section 4.3.9** later). Again, it was unclear whether these differences were due to true geographical or population effects, or because of variation in experimental procedure and design between the two cohorts.

### 4.3.2 Microbiome profile groups (MPGs) in CAS and COAST

Hierarchical clustering was performed on CAS and COAST separately, to group nasopharyngeal samples with similar profiles (microbiome profile groups, MPGs). This was performed on a subset of the data, comprising the following: 18 ASVs common to both CAS and COAST; 7 rarer ASVs belonging to a major genus or family; and a single “others.rare” feature consisting of all remaining ASVs combined (**Methods, Section 4.2**). This gave a total of 26 features for clustering into MPGs (**Figure 4.1**).

The MPGs derived from CAS using QIIME2 were largely consistent with those derived using QIIME1 (compared to MPGs described in Teo et al [15]). The COAST dataset yielded 13 MPGs, one fewer than CAS as it lacked the small *Streptococcus*.rare ASV-dominated MPG (**Supplementary Figure C.2B, Supplementary Figure C.4B**). The other MPGs were similar between CAS and COAST (**Supplementary Figure C.4B**). MPGs from both cohorts



had adequate cluster validity, with average silhouette widths in CAS and COAST of 0.44 and 0.39 respectively. The average within- and between-group Bray Curtis dissimilarities for CAS were 0.40 and 0.88; for COAST they were 0.43 and 0.87, respectively.

MPGs were named and colour-coded according to the dominant ASV within that cluster (**Supplementary Figure C.4**). All other non-dominant ASVs had zero-inflated distributions of relative abundance, and typically contributed sparingly to the overall read count of each MPG. There were noteworthy exceptions: in both cohorts, the *Alloiococcus*.dd2e MPG also contained a significant proportion of *Corynebacterium*.cb50 in addition to the dominant *Alloiococcus* ASV.

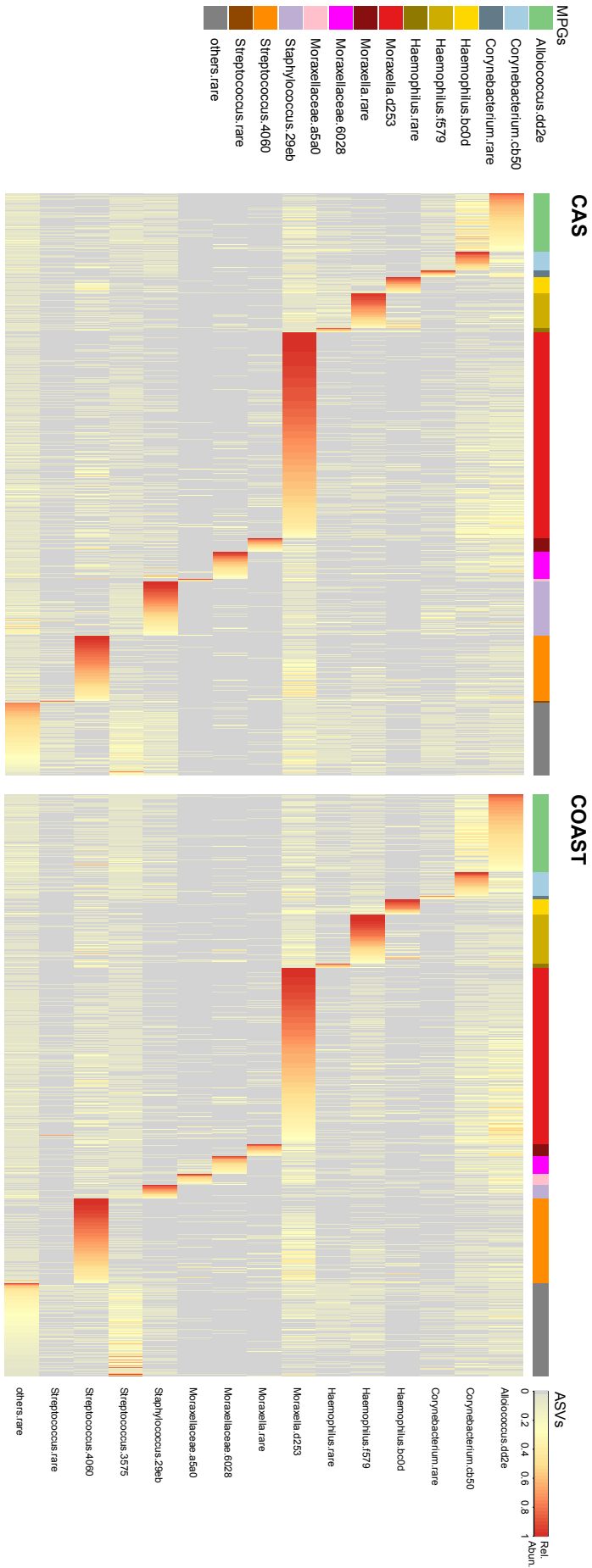
The others.rare MPG of the QIIME2 pipeline contained a mixed assortment of other rare ASVs, primarily taxa from *Neisseria*, *Prevotella*, *Veillonella*, and (for COAST) *Rothia* spp.. This was most consistent with the “others” or “Mixed 1” MPG in CAS QIIME1 as described in Teo et al 2018 [15]. However, in COAST, the others.rare MPG also consisted of a significant proportion of *Streptococcus*.3575 ASV (**Supplementary Figure C.4B**), the significance of which is discussed later.

Those taxa (OTUs/ASVs) and MPGs that were previously identified to be associated with respiratory illness [15] remained so with the QIIME2 pipeline. For both CAS and COAST, a number of specific taxa were overrepresented, at both the MPG and ASV levels, in sick samples compared to healthy samples (**Supplementary Figure C.2A-B, Figure 4.2**): these were *Moraxella*.d253 (*Moraxella*.OTU4398454), *Streptococcus*.4060 (*Streptococcus*.OTU1059655), *Haemophilus*.f579 (*Haemophilus*.OTU240051) and *Haemophilus*.bc0d (*Haemophilus*.OTU956702). In both cohorts, we observed that colonization with these illness-associated MPGs or ASVs increased with age, before plateauing from age 2y onwards (**Supplementary Figure C.2, Figure 4.2**).

### 4.3.3 Correlation patterns between ASVs in CAS and COAST

Correlation analysis using SparCC and FastSpar identified patterns of inter-ASV correlation in CAS (**Figure 4.3A**) that were similar to those discovered with QIIME1 (Teo et al [15], see Figure 4A for comparison). Patterns of correlation were also similar between CAS and COAST (**Figure 4.3B**). In CAS, illness-associated taxa (*Moraxella*.d253, *Haemophilus*.f579, *Haemophilus*.bc0d and *Streptococcus*.4060) were found to congregate together more frequently than expected (SparCC correlation coefficient  $\rho = 0.1$  to  $0.24$ ); while health-associated taxa (*Streptococcus*.3575, *Staphylococcus*.29eb) were negatively correlated with most of these ( $\rho = -0.15$  to  $-0.39$ ; **Figure 4.3A**). Similar correlations were identified in COAST, although to a lesser degree ( $\rho = 0.07$  to  $0.25$ ;  $-0.1$  to  $-0.43$  respectively; **Figure 4.3B**) with fewer linked ASVs and fewer significant correlations.

*Corynebacterium*.cb50 and *Alloiococcus*.dd2e were closely correlated in both CAS and COAST ( $\rho = 0.62$  and  $0.6$  respectively). This was consistent with the high relative abundances of both ASVs in *Corynebacterium*.cb50 and *Alloiococcus*.dd2e MPGs. Paradoxically, these two taxa were also found to be more frequently associated with illness-associated *Moraxella*.d253 in both cohorts ( $\rho = 0.21$  and  $0.24$  for CAS;  $\rho = 0.1$  and  $0.19$  for COAST). In CAS, this relationship was confined to the later two years of life; however this could also be observed in COAST samples which were predominantly collected at age  $<2y$ , albeit with weaker correlation. Specific to COAST was a positive correlation between the health-associated taxa *Staphylococcus*.29eb and *Streptococcus*.3575 ( $\rho = 0.39$ ).



**FIGURE 4.1: Microbiome profile groups (MPGs) and the relative abundance of featured OTUs or ASVs, in (A) CAS and (B) COAST.**

ASV = Amplicon sequence variant; MPG = Microbiome profile group; Rel. abun. = Relative abundance. Rows represent ASVs; columns represent samples. The coloured horizontal bar above the heatmap represents MPGs to which samples were assigned by hierarchical clustering. MPGs are named after the dominant ASV (highest mean relative abundance) of samples in each cluster. Only those fifteen ASVs that each dominated a particular MPG are shown. Please see [Supplementary Figure C.2](#) for a pair of more comprehensive heatmaps covering all common ASVs.

#### 4.3.4 Associations of nasopharyngeal microbiome with respiratory illness in CAS and COAST

To identify significant associations between MPGs and concurrent respiratory illness, we used generalised estimating equation (GEE) models – adjusting for each child as the subjects factor, and for gender, age and season as covariates (**Table 4.2**). These reaffirmed that members of *Moraxella*, *Streptococcus* and *Haemophilus* spp. were illness-associated in both COAST and CAS. Strikingly, the direction and magnitude of associations are consistent across both cohorts, as validated by meta-analysis of CAS and COAST MPGs (**Figure 4.4**). Of the 10 MPGs with a significant association in CAS, eight were also significantly-associated in COAST with the same direction of effect. Similar results were achieved with GEE models constructed at the level of individual ASVs (**Supplementary Table C.3, Supplementary Figure C.5**).

A point of difference between CAS and COAST was the association between “others.rare MPG” and respiratory illness. In COAST there was a significant negative association, while in CAS there was no significant association (**Table 4.2, Figure 4.4**). We previously noted that *Streptococcus.3575* ASV was somewhat dominant in COAST others.rare MPG, but not in CAS others.rare MPG. On further scrutiny, we identified that it was specifically this ASV that was negatively associated with respiratory infection in both cohorts (**Table C.3, Supplementary Figure C.5**). Further analyses using zero-inflated Gaussian models identified that the direction and degree of association between other ASVs and respiratory health status remained similar, even after adjusting for age, season, sex, and the two most common pathogenic ASVs as covariates (*Moraxella.d253, Streptococcus.4060*).

#### 4.3.5 Associations of nasopharyngeal microbiome with seasonal changes in CAS and COAST

In both cohorts, MPG proportions and ASV relative abundances tended to follow seasonal trends, with illness-associated taxa (particular *Moraxella.d253*) being more common during the winter months in both CAS and COAST (**Figure 4.5, Supplementary Figure C.6**). Specifically, *Moraxella.d253* and *Streptococcus.4060* colonisation were more frequent in winter months, while *Alloicoccus.dd2e* was less frequent (**Supplementary Table C.4A, C**). For *Moraxella.d253* and *Alloicoccus.dd2e*, this was partly independent of the increased propensity for respiratory infections during colder seasons, as verified in GEE models (season ~ MPG/ASV + gender + age +/- respiratory illness | subject; **Supplementary Table C.4B, D**). Within COAST, the others.rare MPG (and specifically, the *Streptococcus.3575* ASV) was more prevalent during summer and autumn, and less prevalent during winter and spring (**Supplementary Table C.4**).

#### 4.3.6 Viral detection patterns in CAS and COAST

The general distribution of virus samples in COAST was similar to CAS (**Supplementary Table C.5**; see also Figure S4 of Teo et al 2015 [14]). In both cohorts, rhinovirus (RV), parainfluenza virus and respiratory syncytial virus (RSV) were the most common viruses found in the nasopharyngeal samples. Among rhinovirus (RV) subtypes, RV-A and RV-C were equally common, while RV-B was rare. As expected, viruses of all varieties were more frequently found in unwell samples than healthy samples, (**Supplementary Figure C.7, Supplementary Table C.6**), although there was still a substantial number of cases of asymptomatic viral colonisation. In particular, we observed that the magnitude of association between RSV and respiratory illness appeared stronger in COAST compared

**TABLE 4.2: Results of GEE models associating MPG of sample with illness status (well vs. unwell), with adjustments for child as subjects factor, and gender, age and season as covariates.**

95% CI = 95% Confidence interval; LRI = Lower respiratory illness or infection; MPG = Microbiome profile group; OR = odds ratio; URI = upper respiratory illness or infection. See main text for definitions of LRI and URI. The model for analysis was a generalized estimating equation (GEE), of: respiratory illness status (well vs. unwell) ~ MPG + gender + age + season | subject. Unwell was defined as presence of URI or LRI. Separate models were created for each MPG (i.e. MPG of interest vs. all others). Note that COAST did not yield a *Streptococcus*-rare-dominated MPG. The table is sorted by descending odds ratio in CAS, and statistically-significant associations are bolded.

MPG	CAS					COAST				
	Well (%)	URI (%)	LRI (%)	OR (95% CI)	P-value	Well (%)	URI (%)	LRI (%)	OR (95% CI)	P-value
<i>Haemophilus</i> :bc0d	7 (0.63%)	41 (4.2%)	38 (3.7%)	<b>9.8 (3-31)</b>	<b>0.00013</b>	5 (0.47%)	54 (3.9%)	17 (3.6%)	<b>6.5 (2.3-19)</b>	<b>0.00051</b>
<i>Haemophilus</i> :rare	2 (0.18%)	11 (1.1%)	10 (0.97%)	7.3 (0.88-61)	0.066	2 (0.19%)	15 (1.1%)	6 (1.3%)	<b>4.6 (1.2-18)</b>	<b>0.029</b>
<i>Haemophilus</i> :f579	23 (2.1%)	73 (7.5%)	91 (8.8%)	<b>6.3 (3.5-12)</b>	<b>1.60E-09</b>	30 (2.8%)	161 (12%)	56 (12%)	<b>3.2 (2.1-4.9)</b>	<b>1.40E-07</b>
<i>Streptococcus</i> :4060	38 (3.4%)	138 (14%)	173 (17%)	<b>5 (3.4-7.2)</b>	<b>5.60E-17</b>	45 (4.2%)	279 (20%)	101 (21%)	<b>5 (3.6-7.1)</b>	<b>3.50E-20</b>
<i>Streptococcus</i> :rare	1 (0.089%)	2 (0.21%)	3 (0.29%)	2.5 (0.38-17)	0.33	NA	NA	NA	NA	NA
<i>Moraxella</i> :d253	318 (28%)	363 (37%)	425 (41%)	<b>1.5 (1.3-1.8)</b>	<b>1.50E-05</b>	223 (21%)	495 (36%)	164 (34%)	<b>1.5 (1.2-1.8)</b>	<b>0.00012</b>
others:rare	154 (14%)	114 (12%)	120 (12%)	0.96	0.76	291 (27%)	111 (8.1%)	59 (12%)	0.57	<b>3.30E-06</b>
<i>Moraxellaceae</i> :a5a0	5 (0.45%)	5 (0.52%)	3 (0.29%)	(0.73-1.3)	0.5	19 (1.8%)	28 (2%)	10 (2.1%)	(0.45-0.72)	0.26
<i>Corynebacterium</i> :rare	21 (1.9%)	9 (0.93%)	6 (0.58%)	(0.2-2.2)	<b>0.043</b>	15 (1.4%)	1 (0.073%)	0 (0%)	0.079	<b>0.016</b>
<i>Moraxella</i> :rare	40 (3.6%)	19 (2%)	14 (1.4%)	<b>0.36</b>	<b>0.00064</b>	24 (2.3%)	23 (1.7%)	10 (2.1%)	(0.01-0.62)	0.13
<i>Moraxellaceae</i> :6028	79 (7.1%)	35 (3.6%)	32 (3.1%)	(0.2-0.65)	<b>1.80E-07</b>	33 (3.1%)	46 (3.3%)	11 (2.3%)	0.71	0.17
<i>Corynebacterium</i> :cb50	64 (5.7%)	21 (2.2%)	16 (1.6%)	(0.24-0.52)	<b>3.10E-06</b>	73 (6.9%)	35 (2.5%)	13 (2.7%)	(0.44-1.2)	<b>6.60E-05</b>
<i>Alloiococcus</i> :dd2e	175 (16%)	82 (8.5%)	57 (5.5%)	0.35	<b>4.60E-15</b>	244 (23%)	117 (8.5%)	28 (5.9%)	<b>0.42</b>	<b>8.00E-21</b>
<i>Staphylococcus</i> :29eb	192 (17%)	57 (5.9%)	43 (4.2%)	(0.22-0.54)	<b>2.50E-17</b>	55 (5.2%)	11 (0.8%)	2 (0.42%)	(0.28-0.64)	<b>1.20E-05</b>
				(0.26-0.44)					(0.23-0.38)	
				0.28					0.23	
				(0.21-0.38)					(0.12-0.44)	

to CAS. There was otherwise no significant change to the frequency of virus detection as each child aged.

In both CAS and COAST, we observed seasonal patterns to virus detections across all samples (healthy or otherwise). The detection of RSV and influenza was more frequent in winter, while rhinovirus (RV) detection was more frequent in autumn (**Supplementary Figure C.7A-B**). Both of these trends were typically associated with illness samples in both cohorts. However, when we looked at proportion of samples, we see that RV was disproportionately more frequent in summer and autumn, but less frequent in winter compared to the other viruses (**Supplementary Figure C.7C-D**, **Supplementary Table C.7**).

For many nasopharyngeal samples, there was often more than one virus detected. However, in both cohorts, RV and RSV were less frequently found together compared to with other viruses (Spearman correlation,  $Rho = -0.14$  and  $-0.18$  for CAS and COAST respectively,  $p < 0.001$  for both, **Supplementary Figure C.8**). This may be partially independent of the seasonal patterns observed above, as these correlations remained significant when considering only the samples from winter ( $Rho = -0.20$  and  $-0.18$  for CAS and COAST respectively,  $p < 0.001$  for both).

Generally, illness-associated taxa (*Moraxella.d253*, *Streptococcus.4060*, *Haemophilus* spp.) more frequently co-occurred with viral colonization, and health-associated taxa (*Alloicoccus*, *Corynebacterium*) less frequently so (**Supplementary Table C.8**). This observation was partially independent of co-occurrence with illness episodes or colder seasons, especially for the co-association of viruses with *Streptococcus.4060*, as well as the negative co-association with *Alloicoccus.dd2e*. In particular, *Streptococcus.4060* was significantly co-associated with RSV and RV, even after adjusting for season and illness status (GEE model;  $p < 0.05$ ).

#### 4.3.7 Combined association analysis for respiratory illness with multiple predictors and covariates

In each cohort, we generated GEE models with multiple variables (MPG, presence of virus in sample, season of collection, gender, sex, age) predicting for respiratory illness as the outcome. To keep the analysis well-powered, we combined the presence of any illness-associated MPGs (*Moraxella.d253*, *Streptococcus.4060*, *Haemophilus.f579*, *Haemophilus.bc0d*, *Haemophilus.rare*) into one group, and any health-associated MPGs (*Staphylococcus.29eb*, *Corynebacterium.cb50*, *Alloicoccus.dd2e*) into another, as previously determined in **Table 4.2**. We also grouped the presence of any virus, irrespective of virus strain, into one group. By doing this, we found that colonisation with illness-associated MPGs, winter season, and presence of viruses all semi-independently contributed to risk of respiratory disease, even though they often co-occurred together (**Table 4.3**).

#### 4.3.8 Trends in MPGs and ASVs before and after respiratory infections

There was some evidence that certain changes in the nasopharyngeal microbiome – specifically, a subtle increase in *Moraxella.d253* MPG and ASV – preceded symptoms of respiratory illness. We previously observed that in CAS QIIME1, there was a prodromic increase in both MPG proportion and ASV relative abundance of *Moraxella.d253*, up to two weeks before any LRI (see Figure 5C of Teo et al 2018 [15]). We repeated this analysis in CAS QIIME2, and found similar results for the two-week period preceding either LRI or ARI (any URI or LRI) (ARI results shown in **Supplementary Figure C.9A**). There was also a mild elevation of *Moraxella.d253* in healthy samples up to 1-2 months after the original illness event.

**TABLE 4.3: Results of GEE model associating respiratory illness with presence of virus, illness- or health-associated MPG, incorporating interaction effects.**

95% CI = 95% confidence interval; MPG = Microbiome profile group; OR = Odds ratio. The model for analysis was a generalized estimating equation (GEE), of: respiratory illness ~ any illness-associated MPG × any health-associated MPG × any virus × winter season + age (in years, not days) + sex | subject. Statistically-significant associations are bolded.

Variable	CAS		COAST	
	OR (95% CI)	p-value	OR (95% CI)	p-value
Any illness-associated MPG	<b>7 (4.3-12)</b>	<b>1.20E-14</b>	<b>14 (8.5-22)</b>	<b>6.00E-27</b>
Any health-associated MPG	<b>0.45 (0.25-0.8)</b>	<b>0.0065</b>	1 (0.58-1.8)	0.94
Any virus	<b>2.6 (1.4-5)</b>	<b>0.0029</b>	<b>2.8 (1.3-6.2)</b>	<b>0.01</b>
Season = Winter	<b>3.1 (1.9-5)</b>	<b>2.60E-06</b>	<b>3.7 (2.1-6.3)</b>	<b>2.80E-06</b>
Gender = Male	1.2 (0.97-1.5)	0.085	1.1 (0.9-1.3)	0.38
Virus × illness MPG	1.5 (0.72-3.1)	0.28	0.62 (0.32-1.2)	0.15
Virus × health MPG	1.2 (0.49-2.9)	0.68	0.52 (0.2-1.3)	0.18
Virus × Season	0.99 (0.39-2.5)	0.98	0.52 (0.17-1.6)	0.26
Health MPG × Season	0.99 (0.53-1.8)	0.96	0.93 (0.5-1.7)	0.83
Season × illness MPG	0.46 (0.21-1)	0.057	0.91 (0.35-2.3)	0.84
Virus × protective MPG × Season	0.29 (0.076-1.1)	0.072	2.9 (0.72-12)	0.13
Virus × risk MPG × Season	1.3 (0.43-3.9)	0.64	0.92 (0.29-2.9)	0.88

We performed the same type of analysis in COAST, but did not replicate these results (**Supplementary Figure C.9B**). We note here that the general prevalence of *Moraxella.d253* was lower in COAST compared to CAS (mean relative abundance 25% in COAST vs. 31% in CAS, Kruskal  $p = 3.7 \times 10^{-7}$ , GLM  $p = 2.4 \times 10^{-9}$  adjusting for age and illness status). The COAST samples only covered the first two years of life, while *Moraxella.d253* was observed to remain at high abundance for CAS samples collected during the later years (age 3 to 5). It was noted that, in COAST, the relative abundance of *Moraxella.d253* was extremely elevated for those baseline samples which were more than 12 months after a preceding illness (**Supplementary Figure C.9B**). This may be explained by the fact that *all* of these samples were exclusively taken at ages 2 years or above. This suggests that hypothetical COAST samples beyond two years of age mirror those of CAS, with a reservoir of high *Moraxella.d253*.

#### 4.3.9 Trajectory analysis of the nasopharyngeal microbiome in CAS and COAST

For each cohort, we first used Multiple Factor Analysis (MFA, **Methods, Section 4.2**) to perform dimension reduction on relative abundance data within healthy routine samples, where mean relative abundances were given per common ASV, per timepoint of sampling up to age 2. The dimensions were generated with equal weighting for each timepoint. Then, we applied k-means clustering to this dimension-reduced dataset, to generate clusters or “trajectories” of subjects whose microbiome followed similar trajectories across time. These trajectories reflected patterns of change within the “healthy” (asymptomatic) nasopharyngeal microbiome over the first two years of life. Note that routine samples taken at age 1.5 were excluded from both cohorts due to low sampled number and high missingness.

Both cohorts generated four trajectories, and similar patterns could be observed across both cohorts (**Supplementary Figure C.10**). Generally, the trajectories followed one of four patterns (**Figure 4.6**):

- “Trajectory A”, which featured early dominance of *Alloiococcus.dd2e* and *Corynebacterium.cb50* ASVs that waned with time.

- “Trajectory B”, which featured persistent moderate-high abundance of *Moraxella.d253* ASV.
- “Trajectory C”, which featured very early domination of *Staphylococcus.29eb* ASV during the first two to six months of life.
- “Trajectory D”, which was found only in COAST. This featured early-life domination of *Streptococcus.3575* and other *Streptococcus* ASVs, before returning to a Trajectory A/B-like pattern from age 1 onwards.

Of the four trajectories in CAS, two shared similar patterns and were hence considered subtypes of one Trajectory (C). In Trajectory C.1, *Staphylococcus.29eb* ASV was dominant for only the first 2 months, while in Trajectory C.2 it remained dominant up to the age of one year. Trajectory D was not observed in CAS.

Interestingly, the patterns observed in the trajectories, which were originally derived from routine healthy samples, appeared to persist in illness samples (URI, LRI; **Supplementary Figure C.11**). Illness-associated taxa (*Moraxella.d253*, *Streptococcus.4060*, *Haemophilus* ASVs) were higher in the illness samples compared to the healthy samples; however, the key identifying characteristic of each trajectory was preserved in the illness samples: for example, illness samples from Trajectory C individuals continued to contain high *Staphylococcus.29eb* at 2mths, even though *Staphylococcus.29eb* was negatively-associated with illness status (see previous **Results**).

#### 4.3.10 Associations with later wheeze, asthma and related disease traits

We had previously found in CAS that colonization with illness-associated MPGs during the first 2 years of life was associated with increased risk of persistent wheeze, but only in those who were also early-sensitised [15]. Conversely, in individuals who were not early-sensitised, illness-associated MPGs was associated with increased transient wheeze (wheezing in the first 3 years of life, but not later). We repeated these analyses in CAS using the QIIME2-based pipeline, and confirmed these results (**Supplementary Table C.9A**). However, these findings were not replicated in COAST (**Supplementary Table C.9B**). We used the same criterion across all available IgE assays in determining early-life sensitisation (any specific IgE > 0.35kU/L at ages up to 2 years of age; **Methods**). Note however that only aeroallergen antibody assays were measured in COAST, whereas CAS also included food allergens; and the proportion of aeroallergen-sensitised individuals by age two in COAST was 23%, compared to 55% of individuals allergen-sensitised in CAS.

Besides analysing MPGs at early timepoints, we were also interested in assessing whether patterns of change in the nasopharyngeal microbiome, over the course of the first three years of life, were associated with respiratory health and asthma-related outcomes. The modelling of the microbiome as a transient, constantly-evolving entity rather than a static one may yield further insights. In COAST, we found that the early *Staphylococcus(.29eb)*-dominated trajectory (Traj. C) was associated with increased risk of later asthma (GLM,  $p = 0.034$  and  $0.0085$  for asthma at age 6 and 13 respectively). This association was mostly independent of high-risk npEM cluster but not of early sensitisation status (**Table 4.4B**). In CAS: whether by assessing Traj. C1 and C2 separately or by combining them into one group, there was no relationship between any one of these and wheeze or asthma diagnosis at age five or ten (GLM,  $p = 0.93$  and  $0.94$  for wheeze at age 5 and 10 respectively; see also **Table 4.4A** — results shown for combined Traj. C1+2 phenotype). Even when we specifically investigated the average proportion of *Staphylococcus.29eb* MPG amongst CAS samples up to age 6 months, we did not identify any significant associations with later wheeze outcomes (GLM models;  $p = 0.39$  for wheeze at age 5;  $p = 0.77$

for wheeze at age 10). None of the other trajectories had any significant relationship with risk of later asthma in either cohort.

We explored possible confounders for the association between early-life Staphylococcal colonisation and disease in COAST. Here, we observed that there was a positive association between membership in Traj.C and allergic sensitisation by age two (41% sensitised in Traj.C vs. 21% in other trajectories; Chi-square test,  $p = 0.04$ ). This explained the loss of association between Traj.C and asthma when accounting for sensitisation in COAST. Notably, there was no association between Traj.C and sensitisation in CAS ( $p = 0.99$  for Traj.C1 or C2). In both CAS and COAST, there were no clear associations between microbiome trajectory and the npEM clusters (immuno-respiratory trajectories) from **Chapter 3** (Chi-squared tests,  $p = 0.074$  and  $0.91$  for CAS and COAST respectively).

In COAST, there was no difference between vaginal delivery versus caesarean section in terms of subsequent microbiome trajectory (Chi-square,  $p = 0.87$ ). Traj. C was also not related to virus detection among samples in the first six months of life, nor to frequency of respiratory infections (GEE models of illness  $\sim$  trajectory + virus + age;  $p > 0.05$  all). However, Traj. C more frequently featured children who were born in winter months (December to February in COAST) as opposed to other months (46% in winter vs. 25% in other seasons; Chi-square  $p = 0.036$ ).

#### 4.3.11 Associations with microbiome alpha diversity

We described the alpha diversity of each nasopharyngeal sample in CAS and COAST using Shannon's diversity index. In doing so, we found that alpha diversity was lower in infection samples than in healthy control samples (**Figure 4.7**). This was true in both CAS and COAST (GEE of diversity  $\sim$  illness status  $\times$  age with subjects factor,  $p = 5.2 \times 10^{-4}$  and  $2.4 \times 10^{-10}$  respectively), and was consistent with our previous observation that infection samples tended to be overwhelmingly dominated by a single taxon or ASV — usually a suspected respiratory pathogen (*Moraxella.d253*, *Streptococcus.4060*, *Haemophilus.f579*, *Haemophilus.bc0d*).

We also observed that, in CAS, alpha diversity tended to dip slightly at age one, then increase with age (**Figure 4.7**). In COAST, few samples were collected past age two, and we were not able to identify any significant increase with age in the available early-life samples. Within the first three years, alpha diversity was positively-associated with increasing age in CAS (GEE model as above,  $p = 7.7 \times 10^{-7}$  for age association in samples up to age three), but not in COAST (GEE model as above,  $p = 0.93$ ). In addition, alpha diversity was significantly lower in healthy samples labelled with illness-associated MPGs (*Moraxella.d253*, *Streptococcus.4060*, *Haemophilus.f579*, *Haemophilus.bc0d*), compared to other MPGs (GEE  $p = 2.1 \times 10^{-39}$  in CAS;  $p = 7.4 \times 10^{-30}$  in COAST).

For each individual in CAS and COAST, the average Shannon index across all healthy early-life nasopharyngeal samples (up to age two) was calculated. We observed that this microbial diversity in early life (during periods of health) varied across microbial trajectories in a manner that was consistent with previous results — individuals in trajectories with higher prevalence of an illness-associated MPG such as *Moraxella.d253* (i.e. Traj.B) tended to have lower alpha diversity (Kruskal  $p = 0.002$  compared to Traj.A in CAS;  $p = 0.19$  in COAST with consistent trend). Interestingly, Traj.D in COAST tended to have higher alpha diversity (Kruskal  $p = 7.6 \times 10^{-7}$  vs. Traj.A), which was consistent with the higher relative abundance of multiple rare ASVs (*Streptococcus* and others) in these samples (**Figure 4.6**).

We then analysed whether early-life microbial diversity in healthy samples was associated with asthma-related outcomes, dependent or independent of other factors such as



**TABLE 4.4: GLM models associating wheeze and asthma outcomes with trajectories based on MFA/k-means of microbiome data, from routine healthy samples within first 2 years of life, and early sensitisation status, as represented by high-risk npEM cluster or early-life allergen sensitisation by age two; (A) in CAS, (B) in COAST.**

95% CI = 95% confidence interval; MPG = Microbiome profile group; OR = Odds ratio. The model for analysis was GLM of respiratory illness ~ microbiome trajectory × sensitisation status or NPEM cluster. Statistically-significant associations are bolded.

<b>A.1. Outcome ~ Traj. C × high-risk npEM cluster CAS3 in CAS</b>					
Outcome	Traj. C (C1+C2)		CAS3		Interaction
	OR (95% CI)	p-value	OR (95% CI)	p-value	p-value
Wheeze at age 5	0.71 (0.32-1.57)	0.39	<b>6.8 (2.0-23.8)</b>	<b>0.0026</b>	0.25
Asthma at age 5	0.93 (0.36-2.40)	0.88	<b>5.2 (1.6-17.5)</b>	<b>0.0070</b>	0.48
Wheeze at age 10	1.02 (0.37-2.80)	0.97	<b>4.7 (1.2-18.0)</b>	<b>0.023</b>	0.68
Asthma at age 10	0.76 (0.25-2.36)	0.64	<b>5.2 (1.4-20.1)</b>	<b>0.016</b>	0.25
<b>A.2. Outcome ~ Traj. C × early allergen sensitisation by age two in CAS</b>					
Outcome	Traj. C (C1+C2)		Sensitisation		Interaction
	OR (95% CI)	p-value	OR (95% CI)	p-value	p-value
Wheeze at age 5	1.3 (0.46-3.9)	0.6	1.9 (0.86-4.1)	0.11	0.39
Asthma at age 5	1.5 (0.45-5.3)	0.49	1.7 (0.66-4.3)	0.28	0.48
Wheeze at age 10	1.9 (0.44-8.7)	0.38	<b>3.9 (1.2-13)</b>	<b>0.027</b>	0.31
Asthma at age 10	1.1 (0.23-5)	0.92	2.2 (0.71-6.9)	0.17	0.96
<b>B.1. Outcome ~ Traj. C × high-risk npEM cluster COAST3 in COAST</b>					
Outcome	Traj. C		COAST3		Interaction
	OR (95% CI)	p-value	OR (95% CI)	p-value	p-value
Asthma at age 6	1.35 (0.43-4.19)	0.60	<b>2.97 (1.17-7.53)</b>	<b>0.022</b>	0.99
Asthma at age 8	<b>3.35 (1.09-10.3)</b>	<b>0.035</b>	<b>3.14 (1.15-8.55)</b>	<b>0.025</b>	0.99
Asthma at age 11	<b>4.14 (1.27-13.5)</b>	<b>0.019</b>	<b>3.56 (1.32-9.61)</b>	<b>0.012</b>	0.99
Asthma at age 13	<b>4.59 (1.35-15.5)</b>	<b>0.014</b>	<b>7.86 (2.56-24.2)</b>	<b>0.00032</b>	0.99
<b>B.2. Outcome ~ Traj. C × early aeroallergen sensitisation by age two in COAST</b>					
Outcome	Traj. C		Sensitisation		Interaction
	OR (95% CI)	p-value	OR (95% CI)	p-value	p-value
Asthma at age 6	1.1 (0.28-4.1)	0.91	<b>2.7 (1.4-5.3)</b>	<b>0.0039</b>	0.077
Asthma at age 8	1.7 (0.54-5.6)	0.35	<b>2.5 (1.3-5.1)</b>	<b>0.0093</b>	0.20
Asthma at age 11	2.1 (0.57-7.5)	0.27	<b>3.3 (1.6-6.8)</b>	<b>0.0016</b>	0.26
Asthma at age 13	2 (0.55-7.3)	0.29	<b>3 (1.4-6.3)</b>	<b>0.0044</b>	0.28

frequency of respiratory infection, prevalence of illness-associated MPGs, and npEM cluster. In doing so, we found that it was not associated with later wheeze or asthma diagnosis in either CAS or COAST, after accounting for the above covariates (GLM,  $p = 0.60$  for wheeze at age five in CAS;  $p = 0.94$  for asthma diagnosis at age six in COAST). Therefore, we did not find any evidence of microbial diversity (or lack thereof) influencing asthma outcomes in ways that were distinct from the other known effects of pathogen colonisation, inflammation, and allergy.

## 4.4 Discussion

### 4.4.1 General trends in nasopharyngeal microbiome in early childhood – similarities across different populations

In our study, we observed that the infant nasopharyngeal microbiome tended to be sparse but highly-structured: for most nasopharyngeal samples, there was one grossly-dominant taxon (ASV) present at high abundance compared to all other taxa. Nasopharyngeal samples could be segregated into discrete clusters (microbiome profile groups, MPGs) each dominated by a unique taxon. The advantage to analysis of categorical variables such as MPGs is that it is relatively less complex and better-powered than using continuous variables with zero-inflated or otherwise uncertain distributions. Also, as each MPG represents a general pattern of microbial composition, MPGs may allow us to more accurately model the composite biological signal of multiple taxa. The disadvantage of using MPGs is that biologically-relevant information may be discarded by reducing read abundances to yes/no categories: a rare taxon may have a significant biological effect even at low abundances. The aim of our dual approach (using both MPGs and ASVs in analysis) was to assess which findings were consistently identified with either approach, and therefore achieve a deeper understanding of how the microbiome contributes to health and disease.

In early childhood, there appear to be distinct patterns of nasopharyngeal microbiota which are similar across multiple populations. Nasopharyngeal samples from both CAS and COAST featured specific dominant taxa that did not only belong to the same genera, but were also represented by the same amplicon sequence variants (ASVs). When these were annotated with taxonomic terms, both the annotations provided by the Greengenes-based naïve Bayes classifier, and those provided by NCBI BLAST, reflected common or known inhabitants of the nasopharynx that had previously been reported for children in this age group [15, 32, 33]. In our study, the most common ASVs (and their best-matching BLAST species, in parentheses) were: *Alloiococcus*.dd2e (*A. otitis/Dolosigranulum pigrum*), *Corynebacterium*.cb50 (*C. pseudodiphthericum*), *Haemophilus*.bc0d (*H. influenzae*), *Haemophilus*.f579 (*H. influenzae*), *Moraxella*.d253 (*M. catarrhalis*), *Staphylococcus*.29eb (multiple *Staphylococcus* species including *S. aureus* and *S. epidermidis*), *Streptococcus*.4060 (*S. pneumoniae*), and *Streptococcus*.3575 (*S. mitis*). In turn, each of these common ASVs were dominant in their own MPG, and analogous MPGs shared similar compositional profiles between CAS and COAST, with a few notable exceptions. For instance, the *Moraxella*.d253 MPG had a similar profile of ASV relative abundances in both CAS and COAST; while the others.rare MPG had slightly different profiles between CAS and COAST, especially with regards to the relative abundance of *Streptococcus*.3575 ASV. We return to the significance of this finding later.

#### 4.4.2 Advantages and limitations of analyses using ASVs derived from 16S V4 region

Our study was among the first to derive ASVs (equivalent to 100%-identity OTUs) from nasopharyngeal microbiome data. Most other studies that used QIIME2 or DADA2 to extract ASVs did so in the context of the gut microbiome [34], although one other study also used derived ASV-level data from nasopharyngeal microbiota in adults [35]. Existing comparisons between different denoising and read-calling pipelines suggest that DADA2 tends to identify a lower number of unique taxa, but otherwise gives comparable results in terms of compositional profiles of each sample [34, 36]. ASV-guided methods offer a number of advantages. With older OTU-picking methods, different reference OTUs may share the same sequence at the 16S V4 region, and hence it is difficult to decide which of these OTUs should be used to label the query read (i.e. the OTUs are non-resolvable or ambiguous based on 16S V4 data alone). The reverse may also be true: because the classification is based on <100% sequence similarity, reads placed in the same OTU may not have identical sequences. By contrast, each ASV produced by DADA2 is recalled with high certainty after quality control, and represents a distinct sequence with a unique identifier that is universal across all studies and analyses (in this case, all studies involving 16S V4 region). The advantages of ASVs over traditional “x%” OTUs is that they encompass as much of the biological variation in the data as possible, and can be comparable across datasets even if they were derived independently [10]. The reliability of ASVs is demonstrated by our replication of key results from QIIME1-derived OTUs [15] using QIIME2-derived ASVs in the same cohort (CAS).

There remain some significant limitations of using ASVs derived from 16S rRNA sequences. One of these is inherent in the nature of 16S sequencing: certain bacterial species, especially *Staphylococcus* spp., cannot be distinguished from one another based on such data alone. We see in both this study and our previous one [15] that the representative sequence for *Staphylococcus.29eb* (or the equivalent OTU929976) is shared by multiple species of *Staphylococcus*. This is important as different Staphylococcal species often have very different effects on the biology and pathophysiology of their hosts: for example, *S. aureus* is both a common commensal and a pathogen in multiple infections, whereas *S. epidermidis* is a common skin commensal with occasional pathogenicity in the context of immunocompromise and medical instrumentation [37]. We cannot distinguish between such species based on 16S rRNA alone. In addition, ASVs do not give any information about antimicrobial resistance (e.g. methicillin-sensitive versus resistant *S. aureus*). Given that genes conferring resistance (e.g. *mecA*) are typically located outside the 16S region, methods to distinguish between species, subspecies and strains based on antimicrobial susceptibility must involve alternatives to 16S sequencing. One such emerging alternative is whole genome and metagenome sequencing. As such technologies become more prominent with future research, our understanding of microbial and antimicrobial effects on respiratory health and disease will also improve.

#### 4.4.3 Contributions of bacteria to acute respiratory illness

Our study replicated key results from previous studies [15, 32]: that nasopharyngeal samples collected during respiratory illness were associated with high abundances of known pathogenic taxa (*Moraxella.d253* = *M. catarrhalis*, *Streptococcus.4060* = *S. pneumoniae*, *Haemophilus.bc0d/f579* = *H. influenzae*), while the health-associated samples were linked to nasopharyngeal commensals (*Alloiococcus.dd2e* = *Alloiococcus/Dolosigranulum*, *Corynebacterium.cb50* = *C. pseudodiphtheriticum*) and the *Staphylococcus* genus (*Staphylococcus.29eb*). These associations accounted for sex, age and subjects as covariates. The findings were

confirmed using both MPG and ASV-level analyses; and they were replicated across both CAS and COAST with meta-analysis. The consistency observed between cohorts was striking, given that the two infant populations were geographically-distant, and were exposed to drastically different climates as well as potentially different microbial ecologies of both host and living environments.

Furthermore, in healthy samples, the relative abundance of *Moraxella.d253* increased with age, plateauing at around age one to two. This mirrored results from Biesbroek et al [32], where children were observed to migrate towards and remain within a *Moraxella*-dominated trajectory by age 2. In our study, all MFA/k-means-defined trajectories appeared to approach a *Moraxella*-dominated composition as children reach two years of age. Around half of all samples or reads belonged to *Moraxella.d253* MPGs or ASVs respectively, while the others were distributed amongst *Staphylococcus.29eb*, *Alloiococcus.dd2e*, *Corynebacterium.cb50*, and other rarer (others.rare) MPGs or ASVs. The biological significance of this trend towards increasing *Moraxella* remains unclear. It may be a consequence of microbial resilience granted by *Moraxella* biofilms, or persistence following *Moraxella*-associated respiratory infections, although we do not see this with other pathogenic taxa. Furthermore, the current literature reports that this trend reverses once more during later childhood and adolescence [38]. Upon reaching adulthood, nasopharyngeal microbiomes tend to have reduced *Moraxella* and increased *Dolosigranulum*, *Corynebacterium* and *Staphylococcus* [39, 40].

We did not identify any significant associations between the *Moraxella*-dominated Trajectory B and asthma outcomes. Biesbroek et al did identify that individuals with high abundance of *Moraxella*, *Dolosigranulum* and *Corynebacterium* were relatively protected from consecutive respiratory infections [32]. Conversely, we found that higher early abundance of illness-associated taxa (including *Moraxella.d253*) was associated with increased incidence of wheeze (transient or persistent) in later childhood, but only in CAS not COAST. Meanwhile, *Alloiococcus.dd2e* and *Corynebacterium.cb50* were both “protective” — or at least, associated with healthy nasopharyngeal samples but not with asthma outcomes. We suspect that the finding in Biesbroek et al was more reflective of an absence of the other pathogenic taxa (*Streptococcus*, *Haemophilus*) in peri-illness healthy samples, and the potential protective effect of health-associated taxa (discussed later), rather than an actual protective effect of *Moraxella*. However, we noted in our study that there were positive correlations amongst *Moraxella.d253*, *Corynebacterium.cb50* and *Alloiococcus.dd2e* (*Dolosigranulum*) in both CAS and COAST, and this trend was more prominent beyond the age of two [15]. In our trajectory analyses, those subjects with samples dominated by early *Alloiococcus/Corynebacterium* (Traj.A) eventually adopted microbiome profiles that also contained significant levels of *Moraxella*. The intrusion of *Moraxella* into later samples (after age two) may be related to the persistence of *Moraxella* colonisation following frequent respiratory infections before age two, as described above. Also, as hypothesised in our previous publication [15], the co-occurrence of *Moraxella*, *Alloiococcus* and *Corynebacterium* may be facilitated in part by *Moraxella*-derived biofilm protecting other commensals.

Interestingly, we observed that the *Streptococcus.3575* ASV was negatively associated with respiratory illness in both CAS and COAST, with the relative abundance in COAST being high enough to comprise a large proportion of the others.rare MPG (and in turn drive its negative association with illness at both ASV and MPG levels). By comparing the representative FASTA sequences, we found that this ASV was analogous to *Streptococcus* OTU1004451 reported in our previous paper [15], and was most similar to the NCBI reference genome of *S. mitis*. Another *Streptococcus* taxon (OTU509773 = *Streptococcus.a3a3* ASV = *S. salivarius* subsp. *thermophilus*) was previously reported to also have a negative association with respiratory illness across the first five years of life (Teo et al 2018, Figure 3 [15]). We note that *S. mitis* and *salivarius* are known commensals with potential but

uncommon pathogenicity (e.g. infective endocarditis, infections in immunocompromised individuals) [41]. On the contrary, *Streptococcus.4060* ASV was positively associated with disease, and was closely aligned with known pathogen *S. pneumoniae*. Hence, even within one genus or species, there are complex relationships between microbes and human health; the pathogenicity of a microorganism may depend on many factors, including the timing of colonization, the immune status of the host, and the presence or absence of other microbes.

Health-associated taxa may reflect absence of illness-associated taxa rather than actual protective effect. However there is some evidence that commensals such as *C. pseudodiphtheriticum* may inhibit pathogenic growth via competitive pressure or some other means [42, 43]. The associations of common taxa with respiratory illness (positive or negative) remained significant after adjusting for the abundances of prominent pathogenic ASVs (namely *Moraxella.d253*, and *Streptococcus.4060*). We had previously shown in Teo et al that for QIIME1 OTUs in CAS, accounting for *Moraxella* OTU 4398454 (equivalent to *Moraxella.d253*) as a covariate altered very few of the fold changes between illness and healthy samples for the other OTUs [15]. This suggests that the impact of individual microbial species on respiratory health, particularly dominant ones, is independent of the rest of the microbiome.

#### 4.4.4 Contributions of season and viruses to acute respiratory illness

It is well-known that respiratory infections tend to occur more frequently in winter. We reaffirmed seasonal variation in respiratory infection rates in both CAS and COAST (Table 4.3). Furthermore, we identified that both illness-associated bacteria (MPGs, ASVs) and viruses were more frequently found in nasopharyngeal samples collected in winter (Supplementary Table C.4, Supplementary Table C.6), irrespective of whether the samples were from healthy or illness episodes.

In both CAS and COAST, most viruses tended to be more frequent in Winter, even amongst healthy samples. In terms of raw frequency, rhinovirus (RV) was also more frequent in Autumn and Winter, but was disproportionately less frequent in winter compared to many of the other viruses. The epidemiology of RV infections in COAST was previously reported in Lee et al 2012 [44]. There it was noted that, while the patterns of RV colonization (symptomatic or asymptomatic) tended to peak in autumn, the rate of RV causing moderate-severe infections remained higher in winter months. Also, while illness samples often yielded more than one viral pathogen, the two most common viral pathogens (RV, RSV) tended to “oppose” each other, as they were less frequently found together even after taking into account seasonality. Overall, our findings suggest that: 1) rhinovirus was the most common viral pathogen in both cohorts, and had a more “perennial” pattern of colonization and infection than other viruses; 2) the pathogenicity of viruses is dependent on season, with many viruses being more prevalent and virulent in winter; and 3) there may be mild oppositional effects between viruses, especially between the two most prevalent pathogens RV and RSV.

Finally, when we combined all three risk factors (MPG, season and virus) in a model for prediction of respiratory illness, these three predictors were semi-independently associated with illness status (Table 4.3). That is, while illness-associated bacterial taxa, virus (RSV, RV) and winter season often co-occurred together, each also made independent contributions to disease risk. We found that certain MPGs, especially *Moraxella.d253* and *Streptococcus.4060*, were positively associated with viral colonization (Supplementary Table C.8). This was consistent with previous studies; Rosas-Salazar et al identified that RSV-infected samples were dominated by *Streptococcus*, *Moraxella*, and *Haemophilus* OTUs [45]. In terms of timing of events: it is possible that pre-colonisation with certain bacterial

pathogens (e.g. *Moraxella*) may increase the risk of superimposed viral infection, although we were only able to identify this in CAS not COAST (**Supplementary Figure C.9**). In addition, season may contribute to disease independently of microbes, and vice versa. For instance, it is probable that winter somehow enhances the pathogenicity of both viruses and bacteria; cold weather tends to promote congregation of human hosts in close proximity to one another (e.g. by spending time indoors), and thereby increase transmission risk of respiratory pathogens. However, while it is a promoter of pathogen virulence, winter season is not a *necessary* condition for respiratory illness. Similar statements can be made for the other predictors (viral or bacterial).

#### 4.4.5 Contributions of nasopharyngeal microbiota to later asthma outcomes

The nasopharyngeal microbiome may influence risk of later asthma and wheeze, and this may be dependent on allergic sensitisation (as represented by sensitisation test results or npEM cluster membership [17]). In CAS, colonisation with illness-associated MPGs during the first 2 years of life was associated with increased risk of persistent wheeze in those who were also early-sensitised, but not in those who were not. However, this was not replicated in COAST. The reasons for this may be related to differences in sampling; differences in allergen measurement, exposure or sensitisation; or to actual geographical or population-driven differences in the way allergy and microbiome interact to elicit disease. In CAS, ryegrass pollen was an important aeroallergen for which IgE was detected, while this was not the case for COAST. Food allergen sensitivity was also measured in CAS but not in COAST. This may have affected the power to detect microbiome-asthma associations stratified by sensitisation state in COAST.

In our exploration of the relationship between microbiota and disease, we hypothesised that the microbiota may contribute in more subtle ways – for instance, it may be the shifting patterns of early-life microbiota, rather than the actual micro-organisms themselves, that are relevant for disease. Therefore, we also described trends in the nasopharyngeal microbiome in terms of trajectories, and attempted to relate these trajectories to asthma outcomes. To our knowledge, we are the first researchers to attempt matching disease outcomes to subpopulations of children with similar trajectories of microbiota, rather than to individual microbial taxa or samples at single timepoints. By using a combined dimension-reduction/clustering method, we were able to summarise each cohort into subsets of individuals based on these microbial trajectories. In particular, three common patterns emerged in both CAS and COAST; one of each dominated by early *Alloisococcus/Corynebacterium* (Traj. A), persistent *Moraxella* (Traj. B); and early *Staphylococcus* (Traj. C). There was also one pattern (early *Streptococcus.3575* or Traj. D) which was found only in COAST. Each of these trajectories may have differential effects on later asthma risk, although the exact relationships remain unclear at this stage.

In COAST, the *Staphylococcus*-dominated trajectory (Traj.C) was associated with increased risk of later asthma; this was not replicated in CAS. We note that the effect of Traj. C on asthma in COAST was not independent of allergy-mediated effects, as represented by aeroallergen sensitisation — but it was independent of membership in the high-risk npEM cluster (**Table 4.4**). We also found a weak association with season of birth (with Traj. C children being more frequently born during the winter months). How all these elements are mechanistically-linked remains unclear. Staphylococci, especially *S. aureus*, are known to generate superantigen which have potentiating effects on T cell activation and potential immune-mediated disease [46]. Previous studies suggest that *Staphylococcus* spp., as skin commensals, may be more commonly found in neonates birthed via Caesarean as opposed to transvaginal deliveries [47]. We were unable to identify any association between Traj. C and delivery method in CAS or COAST. Recent evidence suggests that there may

be interaction with prematurity – in that preterm infants carry with them other risk factors that also promote allergic disease, or are more susceptible to persistent and aberrant microbiota colonisation post-Caesarean. Almqvist et al found that, while Caesareans on their own were not a risk factor for asthma diagnosis and medication, infants born from *emergency* Caesareans were at-risk [48]. Pattaroni et al (private correspondence) found that mode of delivery substantially impacts the respiratory microbiota of preterm infants only, and not infants born at term. At time of writing, we had yet to perform analyses looking at prematurity as a covariate in CAS or COAST, although this may be a future avenue of enquiry.

It is also interesting to note that while Traj. C was associated with later asthma disease, *Staphylococcus.29eb* itself was a *health*-associated microbe – in that it was found more frequently in healthy than illness samples. Similarly, *Alloicoccus.dd2e*, *Corynebacterium.cb50*, and *Streptococcus.3575* (in COAST) were previously found to be negatively-associated with respiratory illness, while *Moraxella.d253* was positively-associated; and yet Traj. A, D and B had no bearing on asthma outcomes (protective or risk-associated). Therefore, the contributions of microbes to respiratory infection or illness may be independent of their contribution to later asthma risk. This is surprising, given that frequent respiratory illness during early infancy, especially symptomatically-severe ones, are known to contribute to wheezing disease in later childhood [14]. It is therefore possible that microbes contribute to asthma in ways beyond simply respiratory infections – perhaps via differences in priming of host immunity, as was alluded to previously.

It is possible that dominance of one microbe to the deficit of other microbes may adversely skew the maturation of host mechanisms for immune surveillance. We examined whether differential microbial diversity, rather than specific microbial taxa, may be independently affecting respiratory health outcomes. We observed that diversity was overall reduced in illness-associated nasopharyngeal samples. A dip in diversity at around age one to two corresponded to the general increase in respiratory infections around this time in both CAS and COAST. We also found that for each individual, the average microbial diversity amongst all early-life samples did not associate with asthma, after accounting for previously-identified risk factors for asthma (number of infections, illness-associated MPGs, trajectories, npEM cluster membership). Notably, we could not distinguish the signal of diversity from the signal of early-life respiratory infections (a known risk factor for later wheeze). However, given the small sample sizes and limitations of our datasets, we cannot definitively argue that diversity does not play a role.

#### 4.4.6 Concluding statements

We applied a new bioinformatic pipeline to nasopharyngeal microbiome data collected from two separate childhood cohorts (CAS, COAST). In doing so we generated ASV-based results similar to those previously generated for CAS in Teo et al [15]. In particular, some microbes were associated with respiratory illness (*Moraxella.d253*, *Haemophilus.f579*, *Haemophilus.bc0d*, *Streptococcus.4060*) while others were health-associated (*Alloicoccus.dd2e* / *Dolosigranulum*, *Corynebacterium.cb50*, *Staphylococcus.29eb*, *Staphylococcus.3575*). Further analysis revealed that these associations were independent of viral co-colonisation or season, and a meta-analysis revealed strikingly consistent effect sizes across both cohorts. These microbial associations also interacted with allergic sensitisation — in that illness-associated microbes interacted with sensitisation to confer later asthma. We separated the children into distinct trajectories based on their evolving healthy microbiome in early life. Some of these trajectories bore similar patterns across both cohorts, although one type of trajectory with high early-life *Streptococcus.3575* was unique to COAST. A trajectory dominated by early-life *Staphylococcus.29eb* was associated with later asthma diagnosis

in COAST, but not in CAS. This association may be related to allergen sensitisation or seasonality of childbirth, but the exact mechanisms remain unclear. Attempting to disentangle these mechanisms and associations in greater detail may be an avenue for future research, but was currently not possible in this study due to the limiting sample sizes of both cohorts.

For this study we used data from two independent cohorts with similar properties. However, there were some key issues that may have limited the power of our analyses and interpretability of our results. Both cohorts had small samples, with the observational design of COAST having fewer cases of allergen sensitisation than the selected high-risk individuals in CAS. The tested allergens in COAST (aeroallergens) were slightly different from CAS (which included aeroallergens, food allergens and others). Geographical and demographic differences may also account for some of the discrepancies in findings between cohorts. Finally, differences in findings may reflect limitations in statistical power, as mentioned above. However, despite these differences, there remained extremely remarkable similarities especially in terms of key microbe-disease associations, and microbe-to-microbe relationships.

In conclusion, the early-life nasopharyngeal microbiota may play a role in asthma pathogenesis, in ways that are both dependent and independent of allergic sensitisation and respiratory infection. Some microbial associations are shared across multiple populations; these tend to reflect common or well-known patterns of microbial colonisation and pathogenicity. In addition, microbial mediators of acute respiratory disease (e.g. infection) may differ from mediators of chronic ones, as evidenced by the results of the trajectory analyses. Ultimately, the relationship between human microbiota and human disease is complex, and further research is needed before we can effectively take advantage of the microbiota as a potential avenue for treating and managing asthma and allergy.

## References

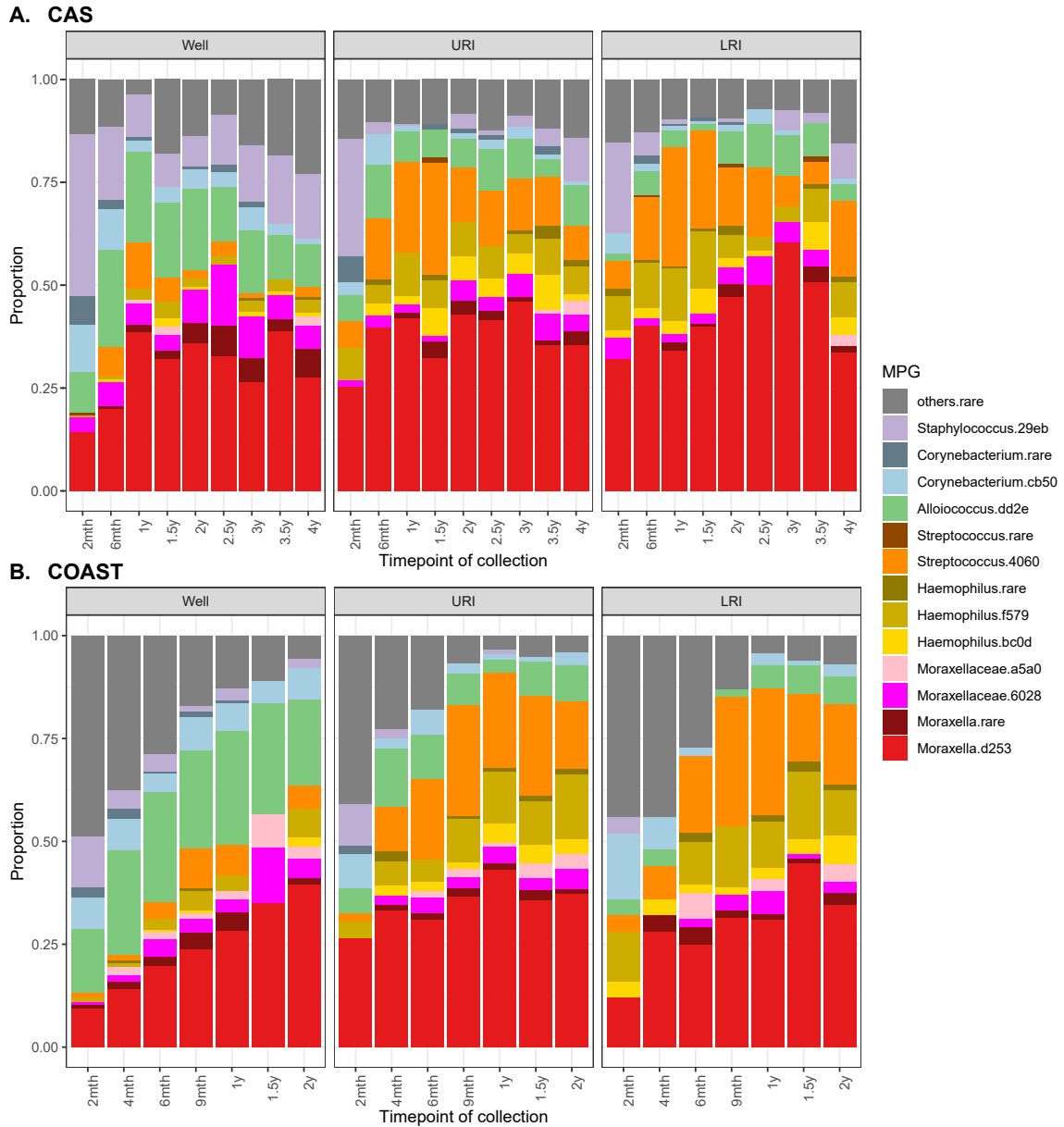
1. Strachan DP. Hay fever, hygiene, and household size. *BMJ* 1989;299:1259–60.
2. Brooks C, Pearce N, and Douwes J. The hygiene hypothesis in allergy and asthma: an update. *Curr Opin Allergy Clin Immunol* 2013;13:70–7.
3. Liu AH. Revisiting the hygiene hypothesis for allergy and asthma. *J Allergy Clin Immunol* 2015;136:860–5.
4. Stiemsma LT and Turvey SE. Asthma and the microbiome: defining the critical window in early life. *Allergy Asthma Clin Immunol* 2017;13:3.
5. Gern JE and Busse WW. Relationship of viral infections to wheezing illnesses and asthma. *Nat Rev Immunol* 2002;2:132–8.
6. Holt PG and Sly PD. Viral infections and atopy in asthma pathogenesis: new rationales for asthma prevention and treatment. *Nat Med* 2012;18:726–35.
7. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010;7:335–6.
8. QIIME 2 Development team. QIIME 2. 2018. URL: <https://qiime2.org/>.
9. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, and Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* 2016;13:581–3.
10. Callahan BJ, McMurdie PJ, and Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J* 2017;11:2639–2643.



11. DeSantis TZ, Hugenholtz P, Larsen N, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 2006;72:5069–72.
12. Kusel MM, Holt PG, de Klerk N, and Sly PD. Support for 2 variants of eczema. *J Allergy Clin Immunol* 2005;116:1067–72.
13. Lemanske R. F. J. The childhood origins of asthma (COAST) study. *Pediatr Allergy Immunol* 2002;13 Suppl 15:38–43.
14. Teo SM, Mok D, Pham K, et al. The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development. *Cell Host Microbe* 2015;17:704–15.
15. Teo SM, Tang HHH, Mok D, et al. Airway Microbiota Dynamics Uncover a Critical Window for Interplay of Pathogenic Bacteria and Allergy in Childhood Respiratory Disease. *Cell Host Microbe* 2018;24:341–352 e5.
16. Lemanske R. F. J, Jackson DJ, Gangnon RE, et al. Rhinovirus illnesses during infancy predict subsequent childhood wheezing. *J Allergy Clin Immunol* 2005;116:571–7.
17. Tang HH, Teo SM, Belgrave DC, et al. Trajectories of childhood immune development and respiratory health relevant to asthma and allergy. *Elife* 2018;7.
18. Illumina. High-Speed, Multiplexed 16S Microbial Sequencing on the MiSeq System. High-Speed, Multiplexed 16S Microbial Sequencing on the MiSeq System. 2014. URL: [https://www.illumina.com/documents/products/appnotes/appnote\\_miseq\\_16S.pdf](https://www.illumina.com/documents/products/appnotes/appnote_miseq_16S.pdf).
19. Louca S, Doebeli M, and Parfrey LW. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome* 2018;6:41.
20. Coordinators NR. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2018;46:D8–D13.
21. McMurdie PJ and Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 2013;8:e61217.
22. Lahti L, Shetty S, Salojärvi J, and Blake T. Tools for microbiome analysis in R. Microbiome package version 1.1.10013. 2017. URL: <http://microbiome.github.com/microbiome>.
23. Kusel MM, de Klerk NH, Holt PG, Kebabdz T, Johnston SL, and Sly PD. Role of respiratory viruses in acute upper and lower respiratory tract illness in the first year of life: a birth cohort study. *Pediatr Infect Dis J* 2006;25:680–6.
24. Gern JE, Pappas T, Visness CM, et al. Comparison of the etiology of viral respiratory illnesses in inner-city and suburban infants. *J Infect Dis* 2012;206:1342–9.
25. Bochkov YA, Grindle K, Vang F, Evans MD, and Gern JE. Improved molecular typing assay for rhinovirus species A, B, and C. *J Clin Microbiol* 2014;52:2461–71.
26. Watts SC, Ritchie SC, Inouye M, and Holt KE. FastSpar: Rapid and scalable correlation estimation for compositional data. *Bioinformatics* 2018:1–3.
27. Friedman J and Alm EJ. Inferring correlation networks from genomic survey data. *PLoS Comput Biol* 2012;8:e1002687.
28. Lê S, Josse J, and Husson F. FactoMineR: An R Package for Multivariate Analysis. 2008 2008;25:18.
29. Mouselimis L. ClusterR. 2018. URL: <https://github.com/mlampros/ClusterR>.

30. Paulson JN, Stine OC, Bravo HC, and Pop M. Differential abundance analysis for microbial marker-gene surveys. *Nat Methods* 2013;10:1200–2.
31. Schwarzer G. meta: An R package for meta-analysis. *R News* 2007;7:40–45.
32. Biesbroek G, Tsivtsivadze E, Sanders EA, et al. Early respiratory microbiota composition determines bacterial succession patterns and respiratory health in children. *Am J Respir Crit Care Med* 2014;190:1283–92.
33. Perez-Losada M, Alamri L, Crandall KA, and Freishtat RJ. Nasopharyngeal Microbiome Diversity Changes over Time in Children with Asthma. *PLoS One* 2017;12:e0170543.
34. Allali I, Arnold JW, Roach J, et al. A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome. *BMC Microbiol* 2017;17:194.
35. De Boeck I, Wittouck S, Wuyts S, et al. Comparing the Healthy Nose and Nasopharynx Microbiota Reveals Continuity As Well As Niche-Specificity. *Front Microbiol* 2017;8:2372.
36. Nearing JT, Douglas GM, Comeau AM, and Langille MGI. Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ* 2018;6:e5364.
37. Chessa D, Ganau G, and Mazzarello V. An overview of *Staphylococcus epidermidis* and *Staphylococcus aureus* with a focus on developing countries. *J Infect Dev Ctries* 2015;9:547–50.
38. Depner M, Ege MJ, Cox MJ, et al. Bacterial microbiota of the upper respiratory tract and childhood asthma. *J Allergy Clin Immunol* 2017;139:826–834 e13.
39. Cremers AJ, Zomer AL, Gritzfeld JF, et al. The adult nasopharyngeal microbiome as a determinant of pneumococcal acquisition. *Microbiome* 2014;2:44.
40. De Steenhuijsen Piters WA, Sanders EA, and Bogaert D. The role of the local microbial ecosystem in respiratory health and disease. *Philos Trans R Soc Lond B Biol Sci* 2015;370.
41. Krzysciak W, Pluskwa KK, Jurczak A, and Koscielniak D. The pathogenicity of the *Streptococcus* genus. *Eur J Clin Microbiol Infect Dis* 2013;32:1361–76.
42. Burkovski A. *Corynebacterium pseudodiphtheriticum*: Putative probiotic, opportunistic infector, emerging pathogen. *Virulence* 2015;6:673–4.
43. Kanmani P, Clua P, Vizoso-Pinto MG, et al. Respiratory Commensal Bacteria *Corynebacterium pseudodiphtheriticum* Improves Resistance of Infant Mice to Respiratory Syncytial Virus and *Streptococcus pneumoniae* Superinfection. *Front Microbiol* 2017;8:1613.
44. Lee WM, Lemanske R. F. J, Evans MD, et al. Human rhinovirus species and season of infection determine illness severity. *Am J Respir Crit Care Med* 2012;186:886–91.
45. Rosas-Salazar C, Shilts MH, Tovchigrechko A, et al. Nasopharyngeal Microbiome in Respiratory Syncytial Virus Resembles Profile Associated with Increased Childhood Asthma Risk. *Am J Respir Crit Care Med* 2016;193:1180–3.
46. Kowalski ML, Cieslak M, Perez-Novo CA, Makowska JS, and Bachert C. Clinical and immunological determinants of severe/refractory asthma (SRA): association with Staphylococcal superantigen-specific IgE antibodies. *Allergy* 2011;66:32–8.
47. Dominguez-Bello MG, Costello EK, Contreras M, et al. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc Natl Acad Sci U S A* 2010;107:11971–5.

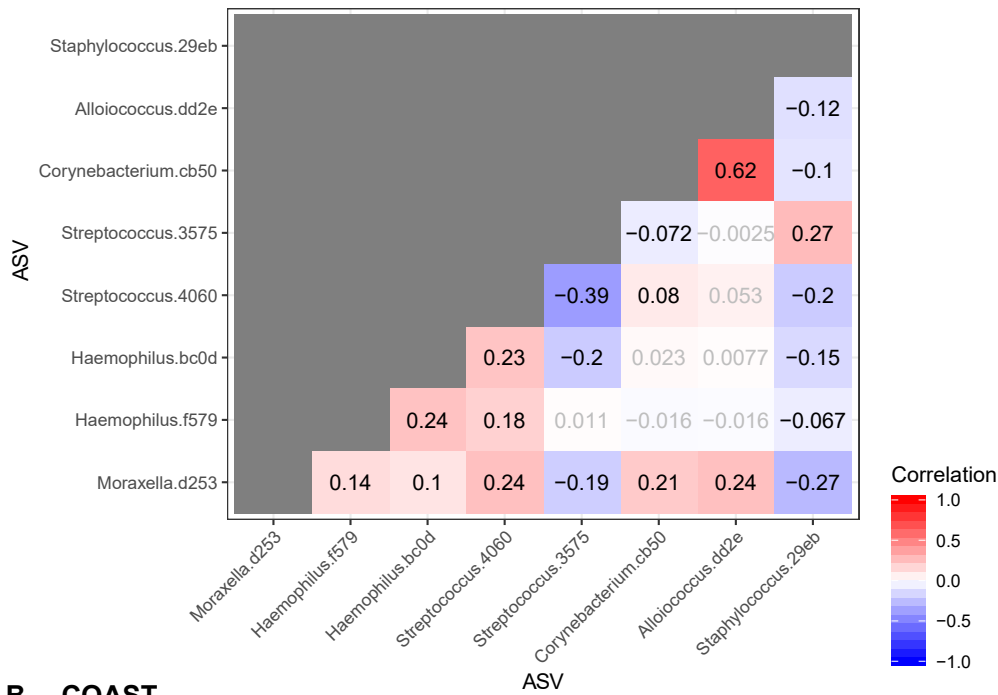
- 
48. Almqvist C, Cnattingius S, Lichtenstein P, and Lundholm C. The impact of birth mode of delivery on childhood asthma and allergic diseases—a sibling study. *Clin Exp Allergy* 2012;42:1369–76.



**FIGURE 4.2: Distribution of MPGs (proportions of samples) in healthy and illness samples, within (A) CAS and (B) COAST**

LRI = Lower respiratory illness or infection; MPG = Microbiome profile group; URI = upper respiratory illness or infection. See main text for definitions of LRI and URI. Note the different time scales for the timepoint of collection in CAS vs. COAST.

## A. CAS



## B. COAST

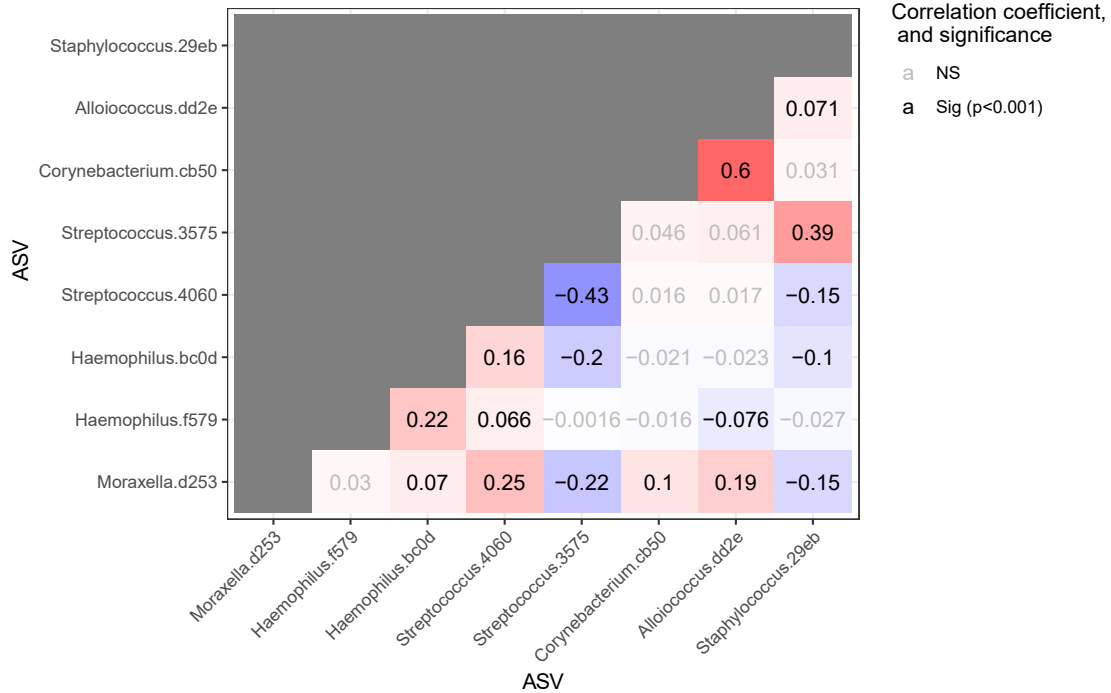
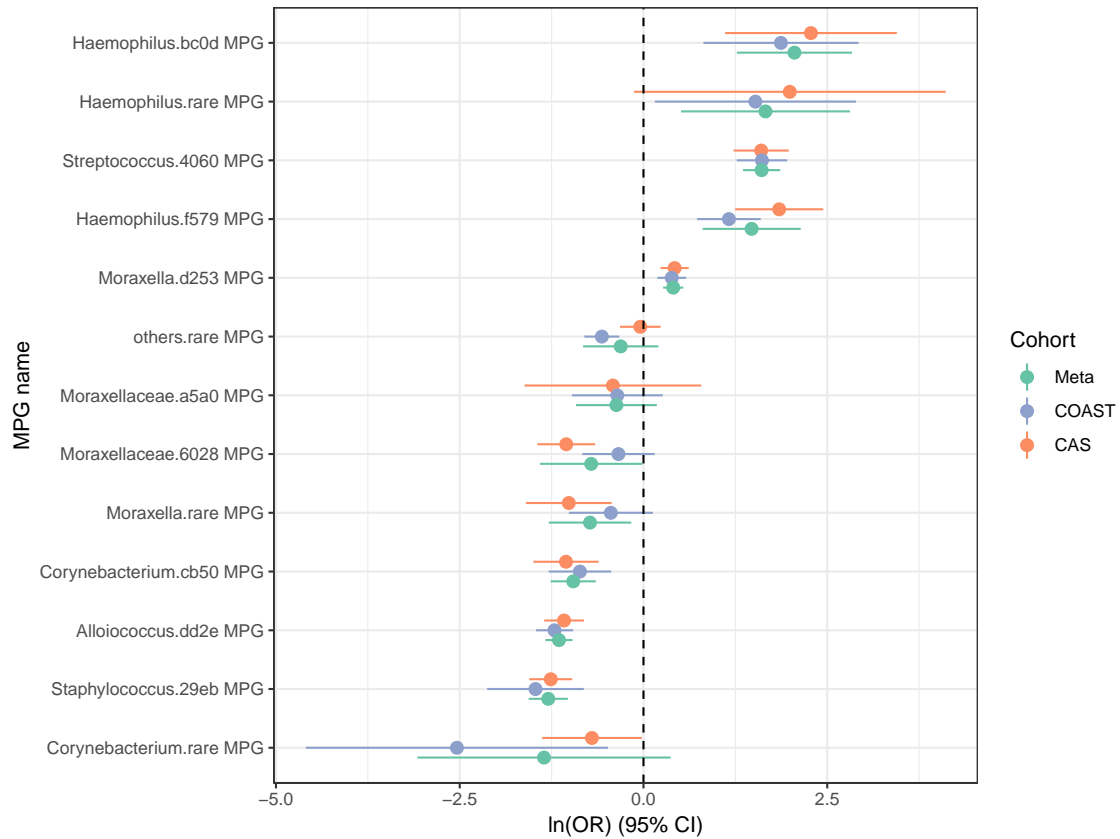


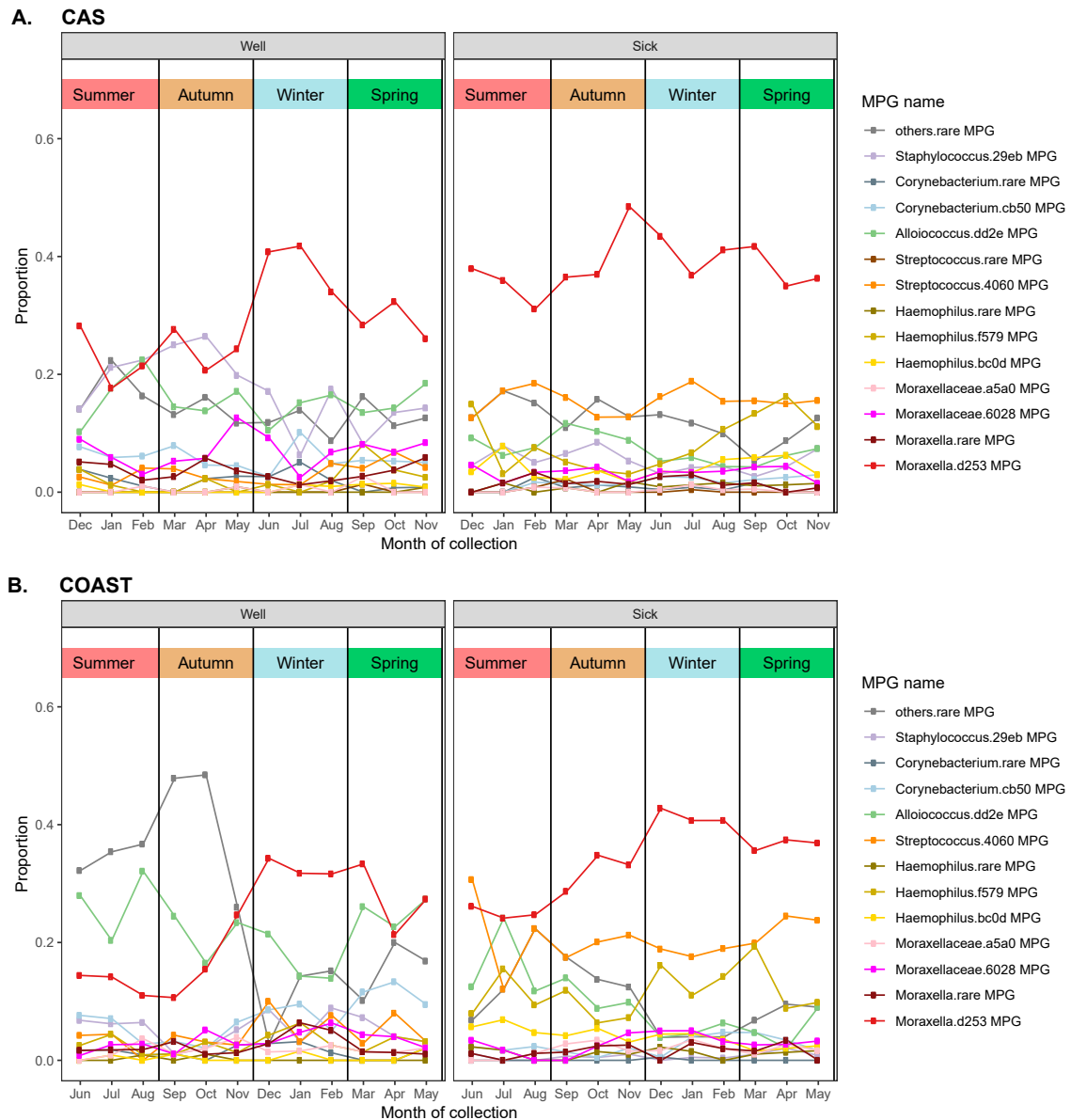
FIGURE 4.3: SparCC correlations amongst ASVs in (A) CAS and (B) COAST.

ASV = Amplicon sequence variant. Heat and number in each cell indicates magnitude of correlation coefficient (Rho); statistical significance of each correlation is indicated by bolded (significant at  $p < 0.001$ ) or greyed font (non-significant).



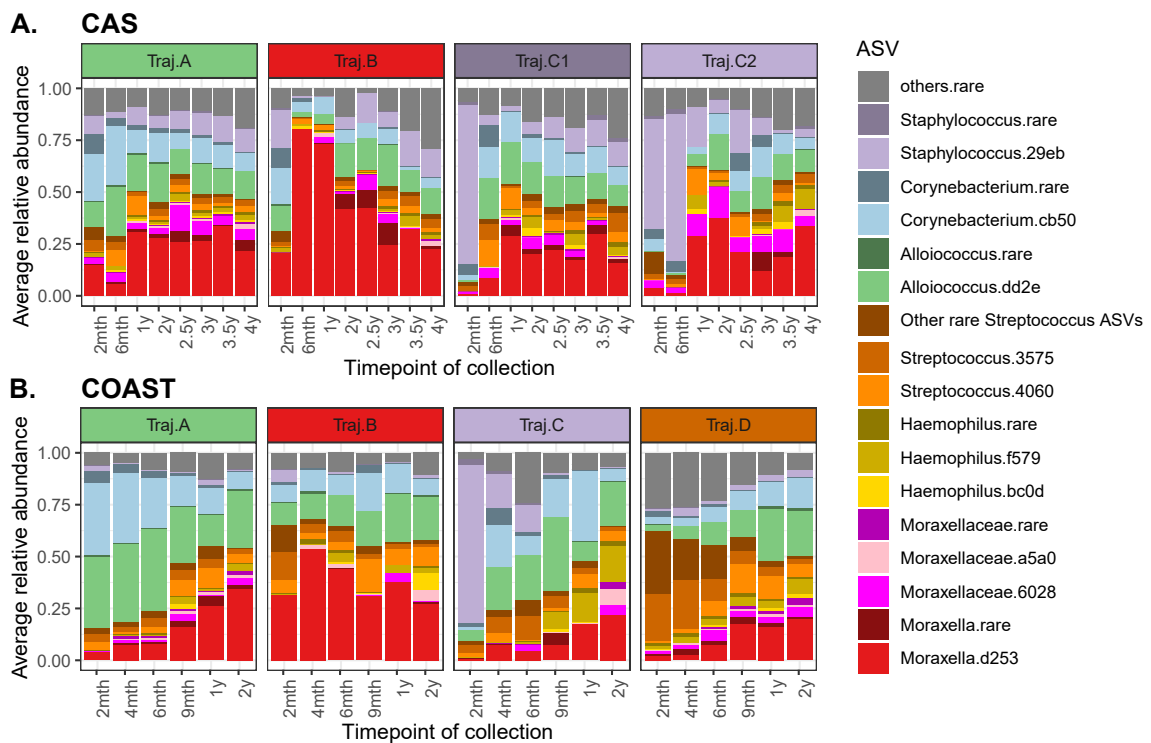
**FIGURE 4.4: Meta-analysis and forest plots of GEE associations between MPGs and respiratory health vs. illness status at time of sample collection.**

Meta = meta-analysis; MPG = Microbiome profile group; OR = odds ratio. Cohort-specific ORs and CIs were generated based on the following GEE model for each MPG: respiratory illness status (well vs. unwell)  $\sim$  MPG + gender + age + season | subject (See **Table C.1**). Meta-analysis was performed for each MPG with random effects and inverse variance weights. Above figure shows the subsequent forest plots, sorted in descending order of meta-analysis ORs. In each forest plot, points indicate natural log of OR, and error bars indicate 95% CI. Dotted line indicates null hypothesis (OR = 1). Observe the close proximity of CAS and COAST forest plot pairs for many MPGs. Also note that the plots for illness-associated MPGs (*Haemophilus.bc0d*, *Haemophilus.f579*, *Haemophilus.rare*, *Moraxella.d253*, *Streptococcus.4060*), are all contained on the right side of the dotted line; while many of the health-associated MPGs (*Alloiococcus.dd2e*, *Corynebacterium.cb50*, *Staphylococcus.29eb*) are placed on its left.



**FIGURE 4.5: Distribution of MPGs (proportions of samples) in healthy and illness samples, arranged by season and month of the year, within (A) CAS and (B) COAST (QIIME2 pipeline).**

LRI = Lower respiratory illness or infection; MPG = Microbiome profile group; URI = upper respiratory illness or infection. See main text for definitions of LRI and URI. Note the different ordering months and seasons in CAS (cohort from the southern hemisphere) versus COAST (northern hemisphere).



**FIGURE 4.6: Average relative abundances of ASVs per healthy samples, per individual, in each microbiome trajectory as determined by MFA/*k*-means analysis of microbiome from healthy samples; in (A) CAS and (B) COAST.**

ASV = Amplicon sequence variant. The microbiome trajectories were based on dimension reduction and clustering of healthy routine (“Well”) samples from the first two years of life, in either cohort.



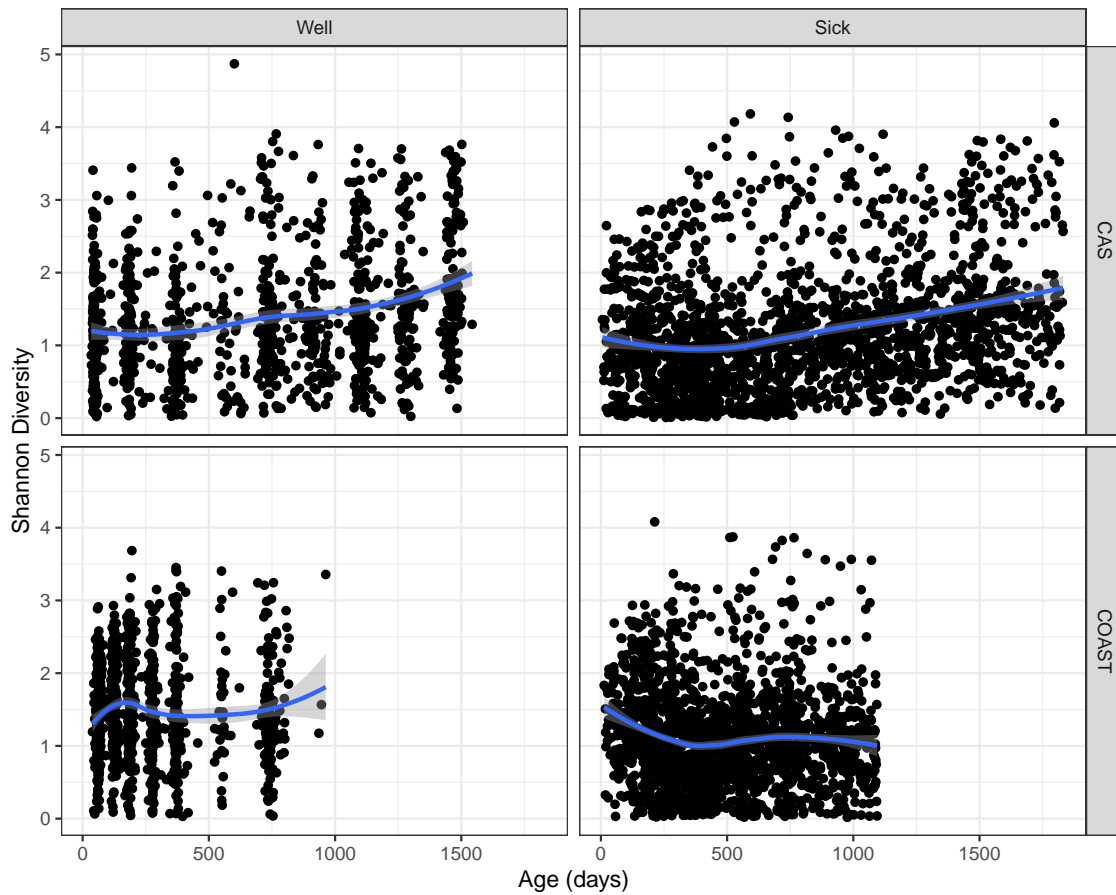


FIGURE 4.7: Alpha diversity of samples (measured by Shannon Diversity index) by age and illness status of samples in CAS and COAST

Note that COAST samples were only collected up to a maximum of age three. A single illness sample for COAST was collected at age greater than three — this outlier was removed from the above plot. Blue curves indicate curves of best fit (LOESS), and grey regions indicate 95% confidence intervals around these curves.



## Chapter 5

# The link between genetics of asthma and allergic disease, and events in early childhood

### 5.1 Introduction

Asthma and allergy are medical conditions with a known heritable component: an individual with a family history of allergy is at greater risk of having an allergic condition [1]. Using genome-wide association studies (GWAS), researchers have so far identified single nucleotide polymorphisms (SNPs) in the genome that are significantly-associated with asthma. Significant SNPs include those located near *IL13* (5q31) [2], *TSLP/WDR36* (5q22) [3, 4], *IL33* (9p24) [5], *CDHR3* (7q22) [6] and *ORMDL3/GSDMB* (17q21) [5]. The majority of these findings were derived from adult populations, although the latter two loci (*CDHR3* and *ORMDL3/GSDMB*) have been identified to be relevant for childhood-onset asthma. However, it is well-known that the origins of asthma (both childhood- and adult-onset) likely arise in early childhood, and many studies have linked early-life phenomena – such as sensitisation to both food [7] and aeroallergens [8], respiratory infections especially with certain viral pathogens [9], breast-feeding [10] and endotoxin exposure [11] – to later disease. There are currently only a few studies that have explored the genetic origins of these early-life risk factors, for instance viral bronchiolitis in early childhood [12, 13]. The gene *CDHR3* has been linked to facilitating rhinovirus-C infection [14], a known contributor to asthma and asthma severity [15–17]. Other studies have examined genome-wide associations with total IgE levels, finding loci such as *FCER1A*, in both adult and childhood populations [18, 19]. Nonetheless, for many potential risk factors in early childhood (such as microbial colonisation of the nasopharynx, as described in **Chapter 4**), it remains unclear how the genetic determinants for these are connected to the genetic basis for asthma itself.

Also, GWAS-derived SNPs account for only a small portion of the total genetic heritability of these diseases, which has been described to vary from about 35% to 95% depending on the study [1]. There remains a large proportion of “missing” or “hidden heritability” that is unexplained by the current list of genome-wide significant SNPs — in one study, the top 31 published associations for asthma accounted for only 2.5% of disease heritability [20]. Possible contributors to this missing heritability include: (1) the involvement of environment and epigenetics in modulating risk; the possibility of which is highlighted by the recent discovery of gene-environment interactions such as those relating to exposure to endotoxin [11] and viral pathogens [9]; (2) the overestimation of original heritability estimates, which may instead be attributable to shared environments especially with heritability estimates from twin studies; (3) the presence of epistasis, or

gene-gene interactions; (4) the contribution of rare variants not accounted for with SNP-based analyses; and (5) the fact that many SNP associations that fail the genome-wide significance threshold ( $5 \times 10^{-8}$ ) may still contribute to genetic risk – the so-called “mid-hanging fruit” described by Ober [21]. For other polygenic diseases such as cardiovascular disease and type-2 diabetes [22, 23], the aggregate contribution of SNPs spanning the entire genome, embodied by a genomic or polygenic risk score (GRS/PRS), may serve as a stronger predictor for disease than a limited selection of only genome-wide significant SNPs. It is likely that a similar phenomenon exists for asthma – however, most recent studies incorporating asthma GRS have constructed their scores from only a select number of genome-wide significant SNPs [24, 25], thus potentially missing out on contributions from the remaining parts of the genome [22]. Therefore we embarked on an approach that incorporated both non-significant and significant SNPs from previous large-scale GWAS analyses, to calculate GRS for asthma-related traits, and then relate these to early childhood traits that may be relevant to asthma pathogenesis.

The primary objective of this study was to use genotype data from a prospective birth cohort — the Childhood Asthma Study (CAS) — to identify whether the genetics of early-life childhood traits were shared or otherwise linked with genetic mediators of asthmatic or allergic disease. Specifically, there were four aims: (1) first we performed a scan for genome-wide significant SNPs associated with early-life traits in CAS, such as frequency and severity of respiratory infections, and allergen-specific antibodies; (2) we repeated these genome-wide analyses using longitudinal association models that also incorporated the serial measurement of early-life CAS traits; (3) we then tested whether genome-wide significant loci previously associated with asthma-related traits (from a curated GWAS catalogue) were linked to early-life traits in CAS; and finally (4) we calculated GRS derived from larger meta-analyses for asthma and allergy-related traits, and explored the association between GRS and early-life traits in CAS.

## 5.2 Methods

### 5.2.1 Samples, genotyping and imputation

The Childhood Asthma Study (CAS) was a prospective birth cohort (total N=263; genotyped N=215) from Perth, Western Australia. Details of sampling and data collection in CAS have been extensively described elsewhere [26, 27]. The salient details relevant to this study are as follows: each child was followed from birth to at most age 10, with routine medical examinations, clinical questionnaires and blood sampling at regular time-points (age 6 weeks, 6 months, then yearly). Serological tests were performed on blood samples for antibodies specific for common allergens, including house-dust mite (HDM) and peanut. Skin sensitisation tests (skin prick tests, SPT) were also conducted for these same allergens. Nasopharyngeal swabs were taken during both routine check-ups and periods of respiratory illness – these swabs were tested for viral content via viral PCR, and bacterial content via 16S V4 region rRNA sequencing with Illumina MiSeq (San Diego, US). Details of how microbial compositional data was derived from these samples is explained in Teo et al and Tang et al [26, 27] (also see previous **Chapter 4**). Virtually all individuals in this cohort were of Caucasian ethnicity.

Genotyping was conducted on the blood samples using an Illumina Omni 2.5 microarray (San Diego, US), to generate 2 391 739 genotyped SNPs per individual, across 22 autosomes. Pre-imputation SNP filtering was performed with the following inclusion criteria: individual missingness  $< 0.01$ , genotype SNP missingness  $< 0.01$ , minor allele frequency (MAF)  $> 0.01$ , and Hardy Weinberg Equilibrium (HWE) test p-value  $> 1 \times 10^{-6}$ . This produced a total of 1 395 154 SNPs for 215 individuals in CAS.

Imputation was then performed via two methods – one for GWAS and one for GRS:

- For GWAS (single-timepoint or repeated measures), we used genotype data imputed from 1000 Genome Phase 3 reference panels [28] using pre-phasing with SHAPEIT v2r644 [29] and IMPUTE2 v2.3.2 [30]. Post-imputation filtering was performed with INFO metric  $> 0.4$  and probability threshold 0.9, returning a total of 12 171 242 SNPs. After removal of monomorphic SNPs in CAS, the final sample size was 8 057 203 SNPs amongst 215 individuals.
- For GRS, we used data imputed by the Michigan Imputation Server [31] with Haplotype Reference Consortium (HRC) r1.1 as the reference panel [32]. Post-imputation filtering removed SNPs with low imputation accuracy ( $R^2 < 0.3$ ), yielding 12 684 109 SNPs in total. Actual number of SNPs used in the subsequent scoring of GRS in CAS ranged from 357 574 to 2 215 305, depending on the availability of SNPs employed for each GRS, as described later. The reason for this change in imputation scheme was primarily due to future plans to associate these GRS scores with RNA microarray data of cell-stimulated gene expression from CAS — the focus of a separate study not related to this thesis (Huang et al, to be published).

For both imputation schemes, cosmopolitan reference panels were used. All genotyping and imputation was performed in reference to the GRCh37/hg19 assembly.

### 5.2.2 Single-timepoint association analyses

We conducted case-control association analyses with single-timepoint early-life traits as outcomes, using Factored Spectrally-Transformed Linear Mixed Models (FaST-LMM v0.207) [33] for genome-wide analyses (GWAS), and PLINK v1.90 [34] for simple candidate SNP analyses.

FaST-LMM is a method that accounts for population and familial structure using a realized or genomic relationship matrix (RRM/GRM) in a linear mixed model. In calculating the GRM, Listgarten et al [35] recommended excluding SNPs that were near the target SNP to avoid “proximal contamination” – excessive deflation and loss of statistical power due to genetic linkage. Since calculating a distinct GRM each of 8 million SNPs was computationally-intensive, we adopted an alternate approach described by Lippert et al [33]: GRMs were calculated from SNPs in all other autosomes bar the one hosting the target SNP, granting a total of 22 unique GRMs, one for each autosome. Sex was a covariate for these analyses.

Outcomes of interest in these analyses included:

- Any occurrence (binary variable) and frequency (ordinal variable) of lower respiratory illness (LRIs), febrile LRIs and wheezy LRIs, in each year up to age 3
- Binary variable of yearly childhood wheezing as reported by the parent, and formal asthma diagnosis by a physician, up to age 5; as well as presence of other allergic diseases (eczema, rhinitis);
- HDM- and peanut-specific IgE levels and HDM SPT tests measured each year, both as binary (IgE  $> 0.35$  kU/L or similar threshold; SPT  $> 2$  or 3 mm depending on age) and continuous variables (IgE in log-transformed kU/L; SPT in mm).
- Average relative abundance of certain microbial taxa (represented by amplicon sequence variants, or ASVs) within nasopharyngeal samples as well as proportion of such samples dominated by such taxa (represented by microbiome profile groups,

MPGs), in each year up to age 2 (see **Chapter 4**); as continuous variables). Due to power limitations, only the top six common taxa with high relative abundances (*Alloiococcus*.dd23, *Corynebacterium*.cb50 *Staphylococcus*.29eb, *Moraxella*.d253, *Streptococcus*.4060, *Haemophilus*.f579; see **Chapter 4**) were examined.

- Membership in a particular subgroup, as determined by mixture-model cluster analysis of immunorespiratory features [27] (also see **Chapter 3**); as binary variable.

For both the single-timepoint and the subsequent longitudinal GWAS: statistical significance was set at the genome-wide threshold of  $5 \times 10^{-8}$ , while  $1 \times 10^{-5}$  was regarded as non-significant but suggestive. Manhattan plots, Q-Q plots and lambda statistics were calculated using the R packages “qqman” [36] and “GenABEL” [37]. Manhattan plots were also generated covering 500kb on either side of a significant or suggestive locus, using the online tool LocusZoom [38]. SNP annotation was performed semi-automatically using the Ensembl Variant Effect Predictor (VEP) [39] and UCSC Genome Browser (Feb 2009) with the GRCh37/hg19 assembly [40]. Only collections of SNPs (loci) with convincing peaks of association were reported; this was defined as the observation, on visual inspection of the Manhattan plots, that SNPs in moderate-to-high LD with the lead SNP were also significantly- or suggestively-associated. Otherwise, significant but lone SNPs were left unreported.

Furthermore, GWAS catalogue SNPs were chosen based on summary statistics of previous GWAS made available on the curated NHGRI-EBI GWAS catalogue [41] as of November 2018. Genome-wide significant SNPs ( $5 \times 10^{-8}$  in any reported study) were selected from published GWAS for asthma-related phenotypes, including: asthma including childhood-onset, allergy or atopy, allergic rhinitis including seasonal, allergic dermatitis or eczema, food allergy (peanut, egg, milk), IgE levels, COPD (including chronic bronchitis or emphysema), and measures of lung function (FEV1, FEV1:FVC ratios). For SNPs that were missing and not imputed in CAS, high-LD proxies were discovered using the R package “proxiesnps” [42], with a 1000 Genomes Phase 3 pan-European reference panel, window size of 500 bp, and  $R^2$  threshold of 0.8. These proxies were then used as surrogates for the missing SNPs. All SNPs were then LD-pruned using PLINK with a window size of 50 bp and pairwise  $R^2$  threshold of 0.8. The subsequent short-list of 416 unique SNPs was then tested for associations with early-life traits, using logistic and linear models applied in PLINK, with sex as a covariate, and multiple testing correction via false discovery rate (FDR-BH) [43] within each trait.

We also tested whether the p-values from the catalogue SNPs were significantly non-uniform, using a one-sample Kolmogorov-Smirnov test against a uniform distribution. Furthermore, we investigated for enrichment of lower p-values by visually inspecting histograms and performing Fisher exact tests for p-values lower than some appropriate threshold (e.g.  $p < 0.50$ ). With each early-life CAS trait, these analyses were performed twice, once for all catalogue SNPs, and once for similar-trait catalogue SNPs (e.g. eczema-only SNPs for early-life eczema in CAS).

### 5.2.3 Longitudinal “repeated-measures” association analyses

We also conducted longitudinal GWAS that incorporated repeated measures for a particular outcome trait across multiple timepoints. To achieve this, we used the function “rGLS” from the R package “repeatABEL” [44], which like FaST-LMM integrates a GRM in a linear mixed model. However, unlike FaST-LMM, rGLS also models repeated measurements as a random effect. rGLS calculates the GRM using a method distinct from Listgarten and Lippert, described elsewhere [44, 45]. Outcome traits for the longitudinal GWAS were as described in the single-timepoint analyses, except similar traits across multiple

timepoints were now incorporated into the model as a single outcome of interest, with sex and timepoint as covariates.

#### 5.2.4 Genomic risk scores (GRS)

Genomic risk scores were derived from publicly-available summary statistics of recent GWAS and meta-analyses conducted for asthma and asthma-related traits. Specifically, we constructed scores for each of the following phenotypes and studies:

- Any allergic disease, as determined by presence of asthma, hayfever or eczema, from a meta-analysis of 12 datasets primarily of European origin, excluding 23andMe ( $N = 9.7 \times 10^4$  cases vs.  $1.5 \times 10^5$  controls); from Ferreira et al [46]. Average number of scoring SNPs for each CAS individual was 1 625 705.
- Asthma, from a sub-study restricted to a European subpopulation ( $N = 1.9 \times 10^4$  cases vs.  $1.1 \times 10^5$  controls); from Demenais et al [47]. Average number of scoring SNPs was 357 574.
- Childhood-onset (age < 12) asthma, as well as adult-onset asthma (ages 26 to 65) from the UK Biobank ( $N = 2.2 \times 10^4$  childhood cases and  $9.4 \times 10^3$  adult cases vs.  $3.2 \times 10^5$  controls who were never diagnosed with either); from Pividori et al [48]. Each asthma subtype was treated independently. Average number of scoring SNPs was 820 627 and 820 603 for childhood and adult asthma respectively.
- Allergic rhinitis and non-allergic rhinitis from a meta-analysis of 17 studies, again restricted to those of European ancestry, and excluding 23andMe ( $N = 1.1 \times 10^4$  allergic cases vs.  $2.8 \times 10^4$  controls;  $2.0 \times 10^3$  non-allergic cases vs.  $9.6 \times 10^3$  controls); from Waage et al [49]. Here, the definition of allergy was based on questionnaire reports, and did not necessitate formal diagnosis. Each of these rhinitis subtypes was also treated independently. Average number of scoring SNPs was 1 461 794 and 2 215 305 for allergic and non-allergic rhinitis respectively.
- For comparison, we also derived scores for chronic obstructive airways disease or pulmonary disease (COAD/COPD) from the UK Biobank ( $N = 3.4 \times 10^5$  subjects in total) [50]. Average number of scoring SNPs was 820 646.

For all meta-analyses involving 23andMe [46, 49], the summary statistics made available to us were re-calculated with 23andMe samples excluded. Also, to reduce redundancy in contributing SNP signals, LD thinning was applied to all contributing SNPs from each set of summary statistics prior to scoring. Criterion for LD thinning was  $R^2 = 0.9$  based on the linkage patterns observed in the UK Biobank population [50]. In the absence of a training set, this threshold of  $R^2 = 0.9$  was chosen as it performed well for a wide range of other highly-polygenic traits which had been tested in UKB by our laboratory (not shown). Then, using PLINK v1.90 [34], we calculated phenotype-specific scores for each subject in CAS, based on their imputed genotype data (imputed with the Michigan Imputation Server [31], as described previously). The score for each individual was then standardised to produce a distribution of values within each phenotype with mean zero and standard deviation one.

We then analysed for associations between these standardised phenotype-specific scores and early-life traits. Due to correlation analysis showing moderate correlation amongst GRS (see later **Results, Section 5.3**), we also constructed scores based on the simple linear summation of all standardised GRS (“combined GRS”), as well as the first

principal component (PC1) from a principal components analysis (PCA) of the standardised GRS. We then associated these with early-life traits in CAS. As per the GWAS reported above, these outcome traits of interest included frequency of respiratory infections, parent-reported wheezing, diagnosis of allergic disease, IgE levels and bacterial compositions in nasopharyngeal samples. In addition we also searched for associations with membership in particular microbiome trajectories, as determined by MFA and k-means cluster analysis described in Tang et al (see **Chapter 4**). Analytic models included generalised linear models (GLMs) for single-timepoint outcome variables, and generalised estimating equations (GEEs) for mult-timepoint (longitudinal) outcome variables for which we wished to adjust for timepoint. All models included sex or gender as a covariate, and the significance threshold was set at  $p < 0.05$  for all tests unless otherwise specified.

## 5.3 Results

### 5.3.1 GWAS for early-life childhood traits relevant to asthma

Using a linear mixed model adjusting for sex and genetic relatedness, we identified several genome-wide significant or near-significant SNPs for a few single-timepoint early-life traits (**Table 5.1**). Most of these analyses yielded acceptable lambda inflation factors (close to 1.00). The presence of parent-reported wheeze at age one was significantly associated with an intronic locus in the gene encoding the enzyme mannosidase endo-alpha *MANEA* (lead SNP rs76781147, odds ratio (OR) 1.98 for effect allele C vs. T,  $p = 6.72 \times 10^{-9}$ ; **Figure 5.1**). Meanwhile, wheeze at age four was associated with a locus at the glutamate receptor subunit *GRIN2B* (lead SNP rs2268113, OR 1.28 for effect allele T vs. C,  $p = 1.1 \times 10^{-8}$ ; **Figure 5.2**). Wheeze at age five was suggestively-associated with a locus in *DPP10* (lead SNP rs9646928, OR 1.26 for effect allele T vs. C,  $p = 4.5 \times 10^{-7}$ ; **Supplementary Figure D.1**).

The occurrence of any wheeze during LRIs (wLRIs) in the first year of life was suggestively-associated with variants in the dynein subunit *DNAH5* (lead SNP rs7710301, OR 1.26 for effect allele T vs. C,  $p = 8.6 \times 10^{-8}$ ; **Supplementary Figure D.2**). Similarly, year-one febrile LRIs (fLRIs) were suggestively-associated with cullin-3 or *CUL3*, a key component of E3 ubiquitin ligase (lead SNP rs113820259, OR 1.34 for effect allele T vs. C,  $p = 3.6 \times 10^{-7}$ ). In the first year of life, the occurrence of any rhinovirus-C associated LRIs was significantly linked to a variant in the 3'UTR of the Frizzled-5 receptor gene *FZD5* (lead SNP rs74471859, OR 1.76 for effect allele T vs. A,  $p = 1.2 \times 10^{-9}$ ; **Supplementary Figure D.3**). The presence of year-one LRIs testing positive for rhinovirus-A was suggestively-associated with SNPs in apolipoprotein L3 or *APOL3* (lead SNP rs132651, OR 1.37 for effect allele A vs. C,  $p = 8.2 \times 10^{-8}$ ; **Supplementary Figure D.4**).

With the possible exception of *MANEA* [51] and *DPP10* [52], none of these loci had been previously identified in association analyses for asthma-related traits. We noted that the literature associations with *MANEA* and *DPP10* were not included in the NHGRI-EBI GWAS catalogue, and were not among the catalogue SNPs selected for further analysis in the next section. Huang et al 2015 [51] found the association with *MANEA* using a family-based integrative approach that incorporated gene expression and eQTL data; while the association with *DPP10* had been identified using positional cloning approaches [52].

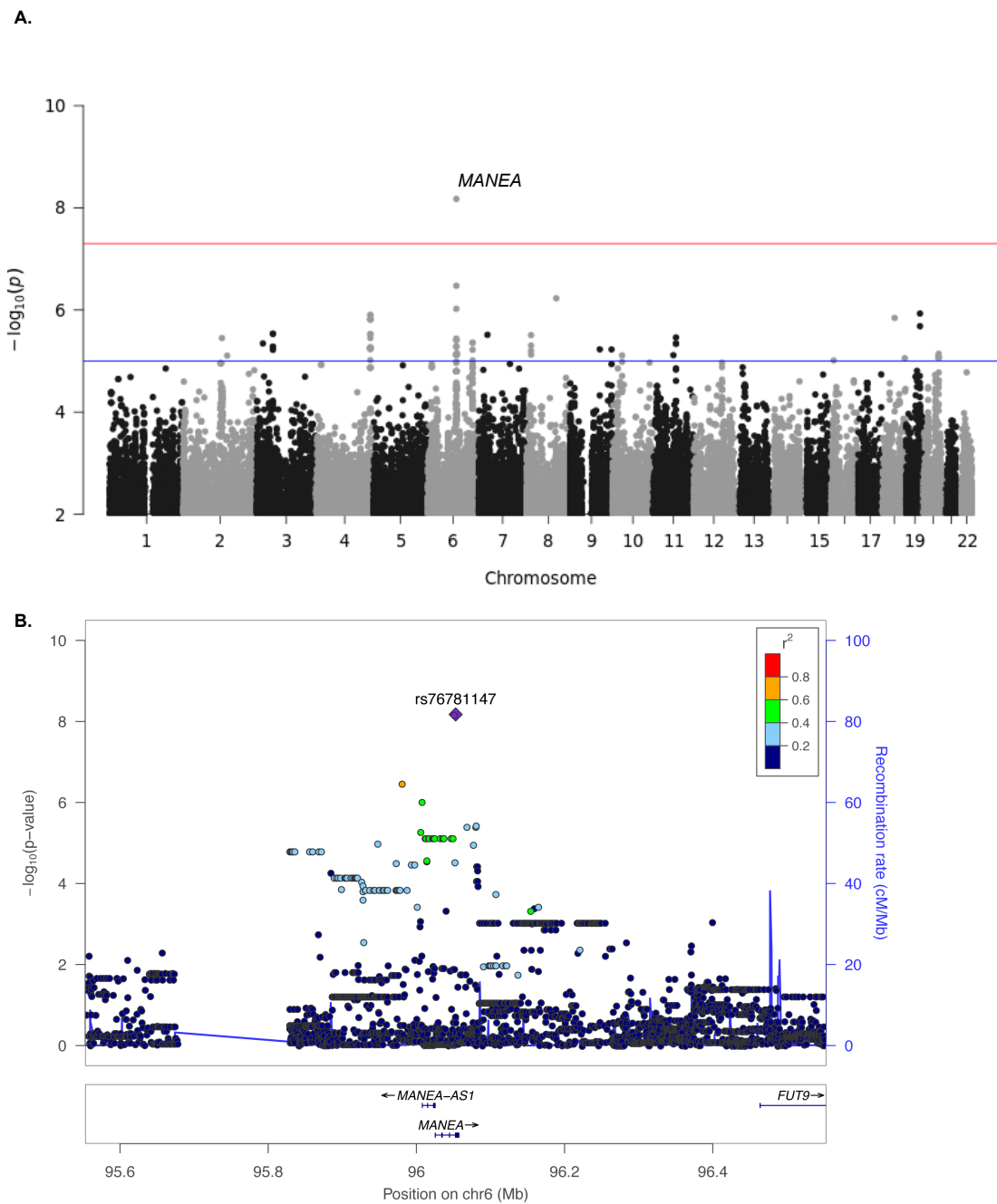
Analyses did not produce significant results for other early-life traits, including less-frequent traits such as infections associated with specific viral pathogens at later timepoints; membership in clusters as determined by Tang et al 2018 [27]; and quantitative traits such as IgE levels and microbial relative abundances.



TABLE 5.1: Selected suggestive and significant SNPs for genome-wide association scans for early-life traits in CAS.

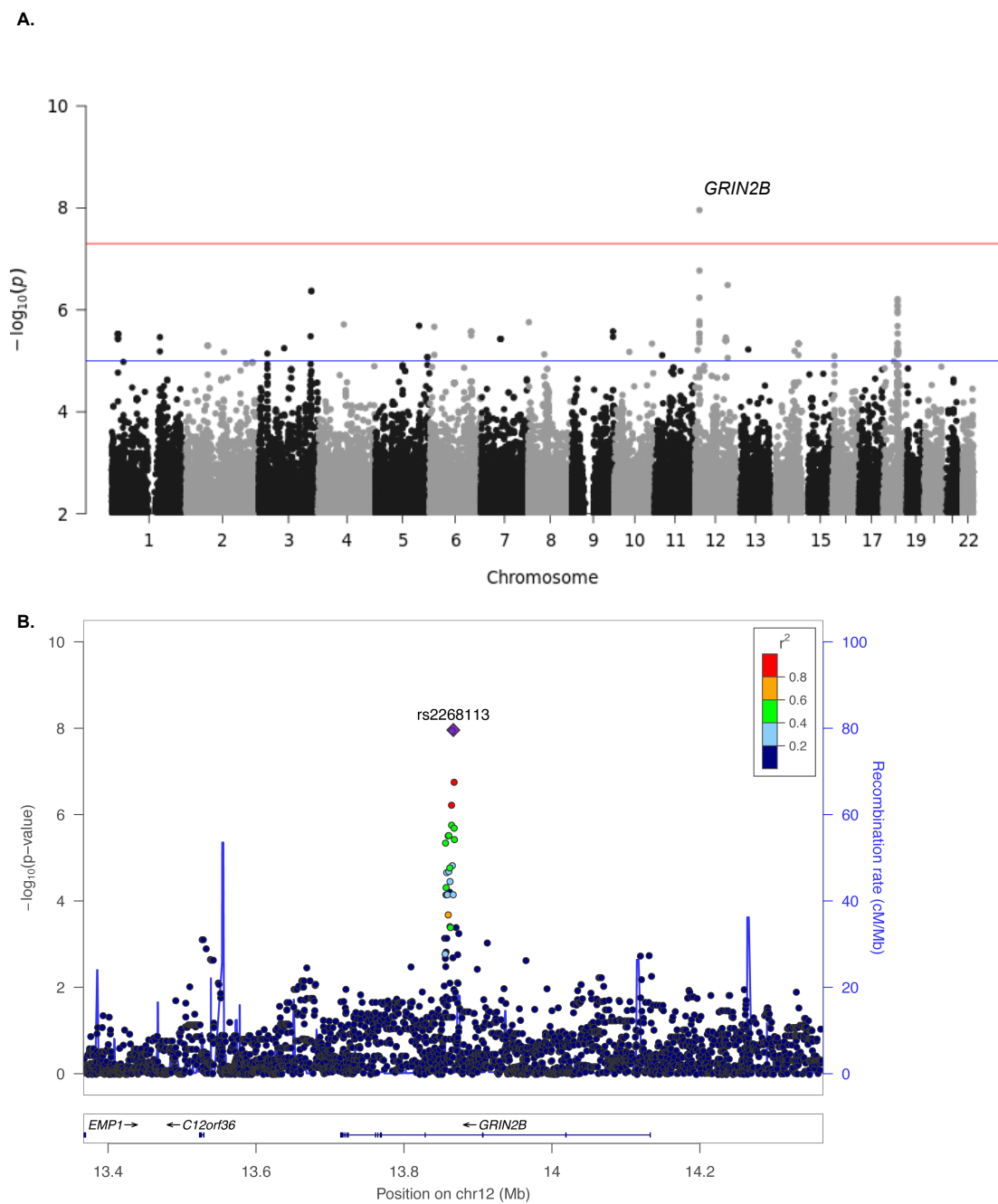
Eff. Alle. = Effect allele; Ref. Alle. = Reference allele. \*npEM clusters derived from mixture-model based cluster analysis, as described in Tang et al [27].

CAS phenotype	SNP	CHR	Position	Eff. Alle.	Ref. Alle.	Closest Gene	Effect of variant	OR	P	Lambda
Any wLRI at age 1	rs7710301	5	13738734	T	C	<i>DNAH5</i>	Intron	1.26	8.55E-08	1.00
Any wLRI at age 2	rs11858622	15	26156870	A	C	<i>RPL1-108419.1</i>	Intron	1.26	7.86E-07	1.00
Any fLRI at age 1	rs113820259	2	225464108	T	C	<i>CUL3</i>	Intergenic	1.34	3.59E-07	1.00
Any fLRI at age 2	rs7859416	9	74985705	T	C	<i>ZFAND5</i>	Intergenic	1.58	8.40E-08	1.00
	rs16856889	1	166269296	G	A	<i>FAM78B</i>	Intergenic	1.31	2.96E-07	1.00
	rs1903890	10	115967869	T	G	<i>TDRD1</i>	Intron	0.79	7.98E-07	1.00
Any RV-A LRI at age 1	rs132651	22	36555365	A	C	<i>APOL3</i>	Intron / non-coding	1.37	8.23E-08	1.00
Any RV-C LRI at age 1	rs74471859	2	208631211	T	A	<i>FZD5</i>	3' UTR	1.76	<b>1.23E-09</b>	1.00
Wheeze at age 1	rs76781147	6	96052987	C	T	<i>MANEA</i>	Intron / non-coding	1.98	<b>6.72E-09</b>	0.99
	rs11734697	4	181125964	A	G	-	Intergenic	1.28	1.26E-06	1.00
Wheeze at age 2	rs73027985	11	123438763	G	A	<i>GRAMD1B</i>	Intron / non-coding	1.77	1.70E-07	1.00
	rs7178909	15	90447946	G	A	<i>C15orf38 / AP3S2</i>	Intron / non-coding	0.79	3.41E-07	1.00
	rs116617169	4	24432684	A	G	<i>PPARGC1A / AC092846.2</i>	Intron / downstream	1.73	5.02E-07	1.00
Wheeze at age 4	rs2268113	12	13866739	T	C	<i>GRIN2B</i>	Intron	1.28	<b>1.10E-08</b>	1.00
Wheeze at age 5	rs9646928	2	115662988	T	C	<i>DPPI10</i>	Intron / non-coding	1.26	4.50E-07	1.00
Eczema at age 2	rs62040427	16	64839701	T	C	<i>CDH11</i>	Intergenic	0.75	<b>4.88E-08</b>	1.00
Eczema at age 5	rs199185	5	8680173	T	C	<i>SEMA5A</i>	Intergenic	1.32	3.74E-07	1.00
npEM cluster 1 membership*	rs35027739	10	99696003	A	G	<i>CRTAC1</i>	Synonymous	1.32	7.62E-07	0.99



**FIGURE 5.1: Manhattan plots of genome-wide association scans for parent-reported wheeze at age 1 in CAS.**

(A) General Manhattan plot; red line indicates threshold for genome-wide significance ( $5 \times 10^{-8}$ ); blue line indicates threshold for suggestive association ( $1 \times 10^{-5}$ ). (B) LocusZoom plot [38] focusing on the locus of interest at Chromosome 6 near *MANEA*. LD  $R^2$  values and recombination rates given as per hg19/1000 Genomes Nov 2014 EUR reference genome.



**FIGURE 5.2: Manhattan plots of genome-wide association scans for parent-reported wheeze at age 4 in CAS.**

(A) General Manhattan plot; red line indicates threshold for genome-wide significance ( $5 \times 10^{-8}$ ); blue line indicates threshold for suggestive association ( $1 \times 10^{-5}$ ). (B) LocusZoom plot [38] focusing on the locus of interest at Chromosome 12 near *GRIN2B*. LD  $R^2$  values and recombination rates given as per hg19/1000 Genomes Nov 2014 EUR reference genome.

### 5.3.2 Association analyses of GWAS catalogue SNPs for childhood asthma-related traits

We found that none of the GWAS catalogue SNPs or their proxies were associated with asthma in CAS at the significance threshold adjusted for multiple comparisons (FDR-BH threshold  $\alpha = 0.05$  for each trait; adjusted p-value  $p_{i.adj} = p_i \times 416/i$  for the  $i^{\text{th}}$  p-value in ascending order). At the unadjusted threshold ( $p < 0.05$ ), SNP or loci that most frequently featured across multiple early-life traits included those located in genes for asthma (*ZPBP2*, *SMAD3*, *GATA3*, *CLEC16A*), allergy (*PLCL1*, *RERE*), eczema (*LCE3E/CRCT1*), COPD (*CHRNA3/4/5*, *RAB4B*) and lung function (*SERPINA1*) (**Supplementary Table D.1**). We observed that for some SNPs, the significantly-associated early-life traits were somewhat correlated with the original phenotype for that SNP (**Supplementary Figure D.5**). SNPs associated with lung function (e.g. rs28929474 near *SERPINA1*) were associated with wheezing disease in CAS, but not with allergen-specific IgE measures. Conversely, SNPs associated with allergy (e.g. rs4908769 near *RERE*) were linked to allergen-specific sensitization but not respiratory infections.

Compared to uniform distribution, the combined p-value distribution of GWAS catalogue SNPs was significantly enriched for lower p-values in association with eczema at 6 months of age (Kolmogorov-Smirnov  $p = 0.0002$ , comparing catalogue p-value distribution with uniform, **Supplementary Figure D.6**). Enrichment was not observed for any other traits. When we performed similar analyses looking exclusively at catalogue SNPs associated with traits similar to the one studied in CAS (i.e. eczema-associated SNPs for early-life eczema in CAS; asthma-associated SNPs for asthma or wheeze in CAS; and atopy/IgE-associated SNPs for HDM and peanut IgE and SPT), we did not find any evidence for enrichment (Kolmogorov-Smirnov  $p > 0.05$  for all comparisons).

The development of later chronic respiratory disease may be influenced by genetics modifying one's susceptibility to early childhood events. The adenosine (A) allele at rs1663687 near *GATA3* was associated with fewer LRIs and less wheezing in CAS; this same allele has been shown to be protective for asthma in Demanais et al [47]. Some SNPs associated with later-onset chronic diseases (e.g. SNPs near *RAB4B* for COPD [53]) were also associated with early-life respiratory conditions (wheeze and respiratory infections at age three). Notably, we did not find any association between rs6967330 of *CDHR3*, previously implicated in the mechanism of RV-C cell invasion [14], and actual RV-C infection in CAS. For both parent-reported wheeze and asthma diagnosis at age five, the top catalogue SNPs that were associated at an unadjusted threshold ( $p < 0.05$ ) are listed in **Supplementary Table D.2**.

### 5.3.3 Repeated-measures GWAS for early-life childhood traits, and meta-analysis for microbial traits

Given the small sample size of CAS and the limitations of the previous analyses, we wished to see whether statistical power could be improved by combining phenotype information from multiple timepoints. Using repeated-measures GWAS, we examined the relationship between SNPs and longitudinal traits such as yearly presence of wheeze, frequency of respiratory infections, allergen-specific IgE levels, and relative abundances or proportion of samples with certain microbial taxa across the first three to five years of life. Furthermore, for microbial traits, we repeated similar analyses in another birth cohort (COAST), and performed meta-analyses combining results from both CAS and COAST.

The top hits for repeated-measures GWAS of early-life traits in CAS are shown in **Supplementary Table D.3**. We did not replicate many hits from the GWAS of single-timepoint phenotypes, or from the catalogue SNP analyses. One previously-identified

locus at *APOL3* was suggestively associated with both occurrence of RV-A LRI at age one year, and frequency of RV-A LRI across the first three years of life. This may be related to respiratory infections being far more frequent in the first year of life. Other genes with significant or suggestive loci in the repeated-measures GWAS included *ZBTB20* for asthma diagnosis from age three to five, and *LINGO2* and *FAM81A* for wheezy LRIs in the first three years of life. The same loci near *KBTBD7* and *EMCN* were associated with both LRI and wheezy LRI in the first three years of life – this is likely due to wheezy LRIs being a subset of the overall LRI phenotype.

We performed similar longitudinal GWAS for early-life traits related to the nasopharyngeal microbiome in CAS. We and others have previously reported that the microbiome of the upper respiratory tract during early infancy is intimately related to respiratory health later in life [26, 54]. In particular, as reported by Teo et al [26], colonization with certain illness-associated bacteria or pathogens (*Haemophilus*, *Moraxella*, *Streptococcus*) was associated with later asthma in allergen-sensitised individuals. We wanted to see whether genetic determinants for microbial colonization and asthma are distinct or shared, and also whether genetics and microbial exposure interact to elicit disease. Performing longitudinal GWAS using repeatABEL, we identified several loci for certain traits related to the early-life microbiome (**Supplementary Table D.4**). None of these overlapped with the loci identified previously. Notably, some associations were observed for HLA-related gene loci, such as Class I MHC molecule *HLA-A* for *Staphylococcus.29eB*, and Class II MHC molecule *HLA-DRB1* for *Streptococcus.4060*. A locus near *IFNG-AS1* was suggestively-associated with *Alloicoccus.dd23* colonisation. We also identified associations with a couple of genes that interact with the Ras superfamily of proteins (*TBC1D22A* for *Alloicoccus.dd23*, and *GDI2* for *Streptococcus.4060*).

#### 5.3.4 Genomic risk scores (GRS) for asthma-related traits, and their relationships to early-life childhood events

Finally, we generated standardised risk scores (GRS) from multiple GWAS for asthma- and COPD-related traits. As described in the **Methods**, these GWAS were manually-curated from large-scale studies and meta-analyses, and included studies for asthma (adult/childhood), allergic disease, rhinitis (allergic/non-allergic), and COPD. We calculated scores for each individual in CAS in relation to these GWAS traits, then tested for associations between GRS and our early-life childhood traits in CAS.

We found that many of the GRS for allergic phenotypes (presence of any allergic disease, adult and childhood asthma, allergic rhinitis) were slightly correlated in CAS (**Figure 5.3**). Interestingly, the GRS for allergic rhinitis was significantly correlated with that for non-allergic rhinitis (Pearson correlation  $Rho = 0.33$ ,  $p = 8.2 \times 10^{-7}$ ); we note that the summary statistics behind the scores for allergic and non-allergic rhinitis originated from the same study and population [49]. Also, the GRS for COPD was negatively-correlated with that for allergic sensitisation ( $Rho = -0.16$ ,  $p = 0.022$ ).

We performed multiple analyses using GLM (with sex as a covariate) to associate CAS phenotypes with GRS. Each GRS was tested independently. In doing so, we found that among the GRS, the GRS for any allergic disease was strongly associated with early-life allergic phenotypes, including allergic sensitisation and asthma diagnosis at both age 5 and 10 (**Figure 5.4**). The GRS for allergic disease was also associated with the high-risk npEM cluster, Cluster 3 or CAS3 as described in **Chapter 3** — with this high-risk cluster generally having a higher score compared to the others (**Figure 5.5A**). Furthermore, allergic disease GRS was associated with early asymptomatic colonisation of nasopharyngeal samples with illness-associated bacteria (*Haemophilus*, *Streptococcus* and *Moraxella* genera) in the first two years of life (**Figure 5.5B**). Meanwhile, few of the GRS for individual subtypes

of allergic diseases were associated with the examined early-life traits. The GRS for any particular allergic phenotype (e.g. childhood asthma, allergic rhinitis) was not significantly associated with the actual trait in CAS; this is most likely due to limitations in statistical power. However, adult and childhood asthma GRS were both negatively-associated with transient wheeze. The COPD GRS was not correlated with any respiratory-infection associated traits, and was weakly negatively-correlated with total IgE > 100 kU/L at age two ( $p = 0.048$ ). No significant associations were identified with the microbiome-based trajectories from **Chapter 4**. After performing a PCA on the collection of GRS (including non-allergic rhinitis and COPD), we found that there was no clear segregation of individuals into discrete patterns of genetic risk (**Supplementary Figure D.7**), and no clear separation between the npEM clusters discovered in **Chapter 3**.

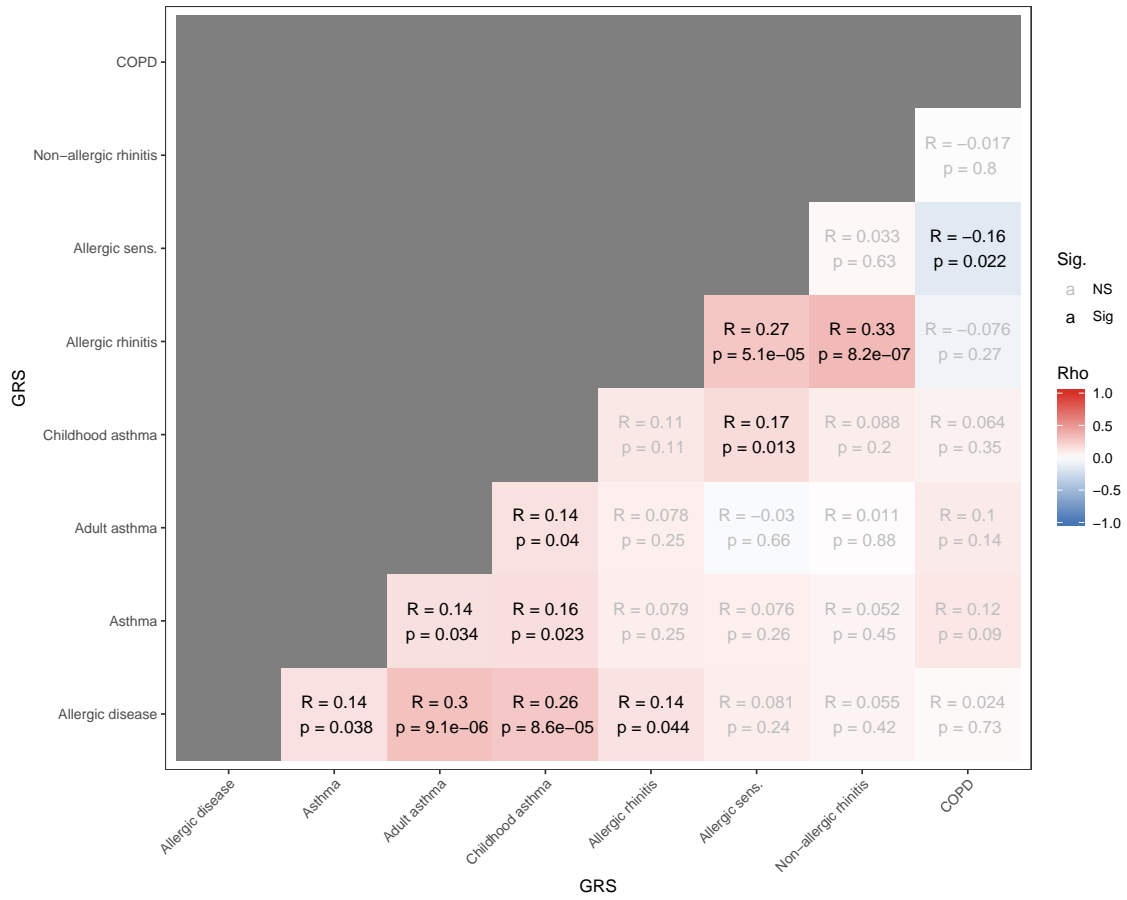
We found that PC1 of the GRS PCA most closely represented an “atopic vector”, with high loadings amongst all allergy-related GRS, but not COPD GRS (**Supplementary Figure D.8**). We interpreted this PC as a dimension-reduced signal that incorporated genetic risk for allergic disease from multiple studies. Like the allergic disease GRS of Ferreira et al [46], the PC1 of all GRS was strongly associated with allergic sensitisation, asthma diagnosis, and membership in the high-risk atopic npEM cluster (“Cluster 3”); while being negatively-associated with the transient wheeze phenotype (**Figure 5.4**).

In addition, we repeated the above GLM analyses with membership in the high-risk npEM cluster [27] as a covariate. In doing so, we found that many of the associations between GRS and early-life allergic traits (**Supplementary Figure D.9**) were diminished after accounting for this covariate. Conversely, the direction or effect sizes of associations between Cluster 3 and early-life traits remained relatively unchanged when GRS was a covariate (**Supplementary Figure D.10**).

Finally, we performed similar analyses using GEE models, accounting for repeated measures of certain early-life traits (e.g. parent-reported wheeze, asthma diagnosis, early-life respiratory infections; **Figure 5.6**). Again we found that the GRS for any allergic disease was most strongly associated with early-life allergy-related traits. Interestingly, the GRS for childhood asthma was associated with both the frequency of wLRIs ( $p = 0.016$ ) and specifically rhinovirus-associated wLRIs ( $p = 0.033$ ).

## 5.4 Discussion

We performed analyses with data from a pediatric cohort (CAS) to assess the link between genetics and early-life traits related to respiratory health, allergy, and asthma. From these analyses we identified a number of loci associated both with early-life traits in CAS, and asthma-related phenotypes in previous GWAS. However, due to the small size and limited power of our sample, traditional genome-wide methods yielded few significant results, even when using longitudinal methods with repeated measures. On the contrary, when we constructed genomic risk scores from GWAS summary statistics, then applied the scoring to CAS individuals, we were able to demonstrate that the genetic contributors to early-life respiratory and immune health are collectively shared with asthma and allergic disease. Although it remains unclear which genetic loci are implicated in both early-life events and later asthma, there is evidence that part of the genetic signal for multiple allergic conditions is associated with early-life allergic sensitisation, respiratory infection and nasopharyngeal microbiome composition.



**FIGURE 5.3: Correlation patterns between the various GRS calculated for CAS.**

Note the degree of correlation amongst atopy-related GRS. Note also the relatively-strong degree of correlation between allergic and non-allergic rhinitis GRS.

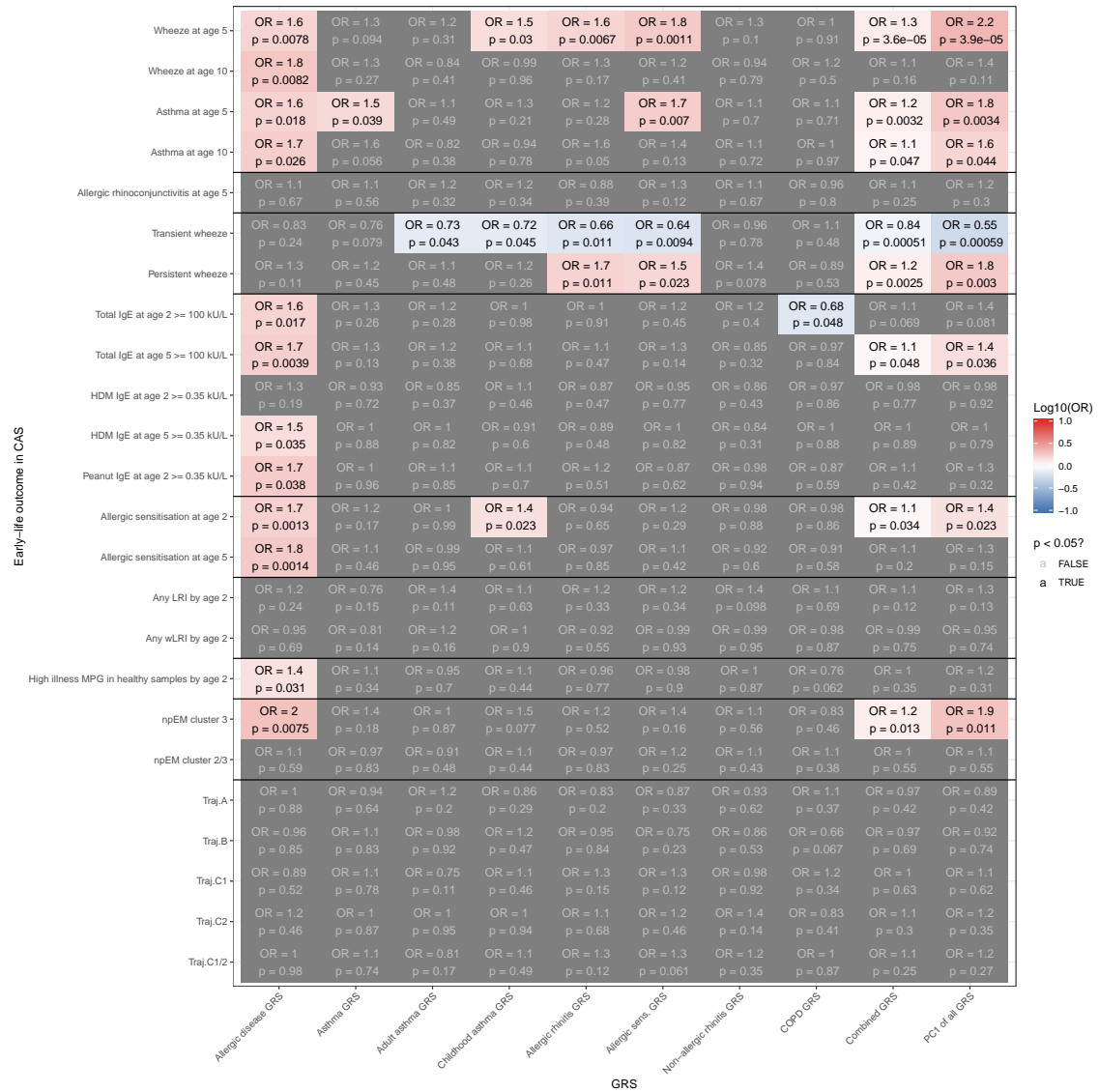
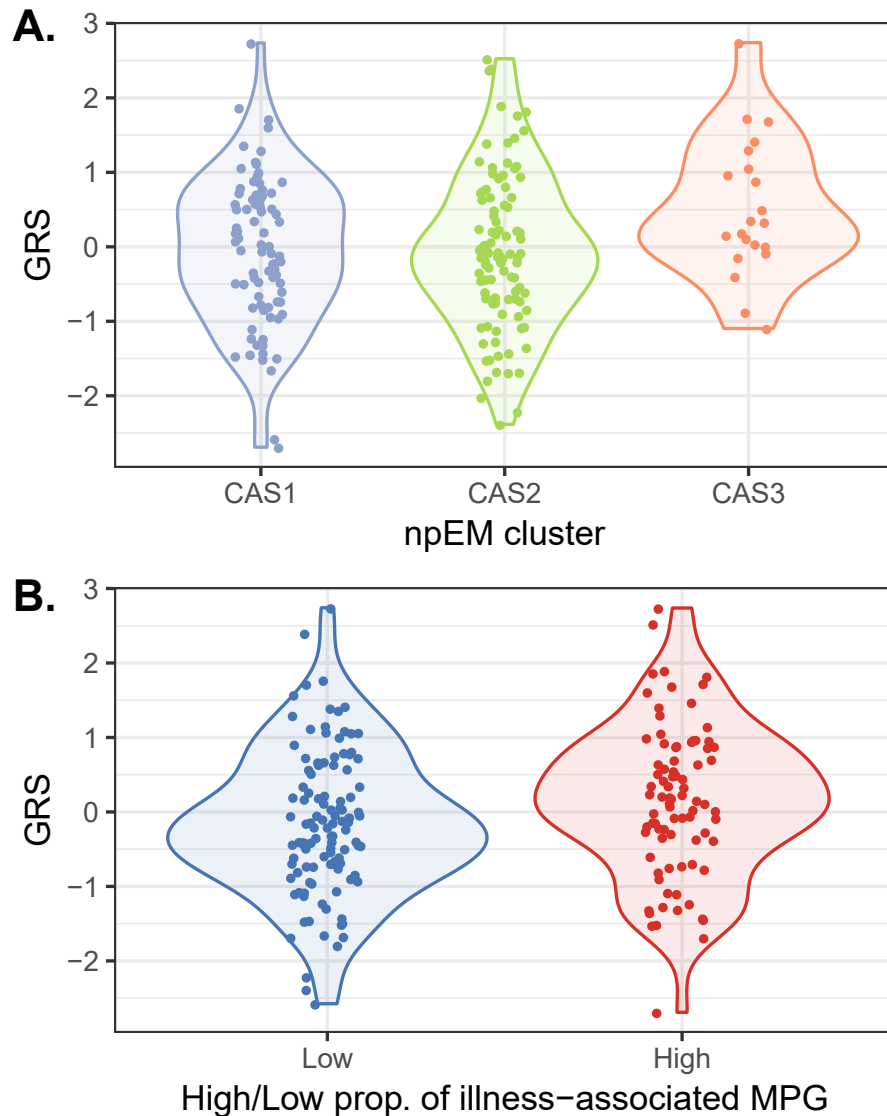


FIGURE 5.4: Associations between GRS and early-life traits in CAS, as determined by GLMs.

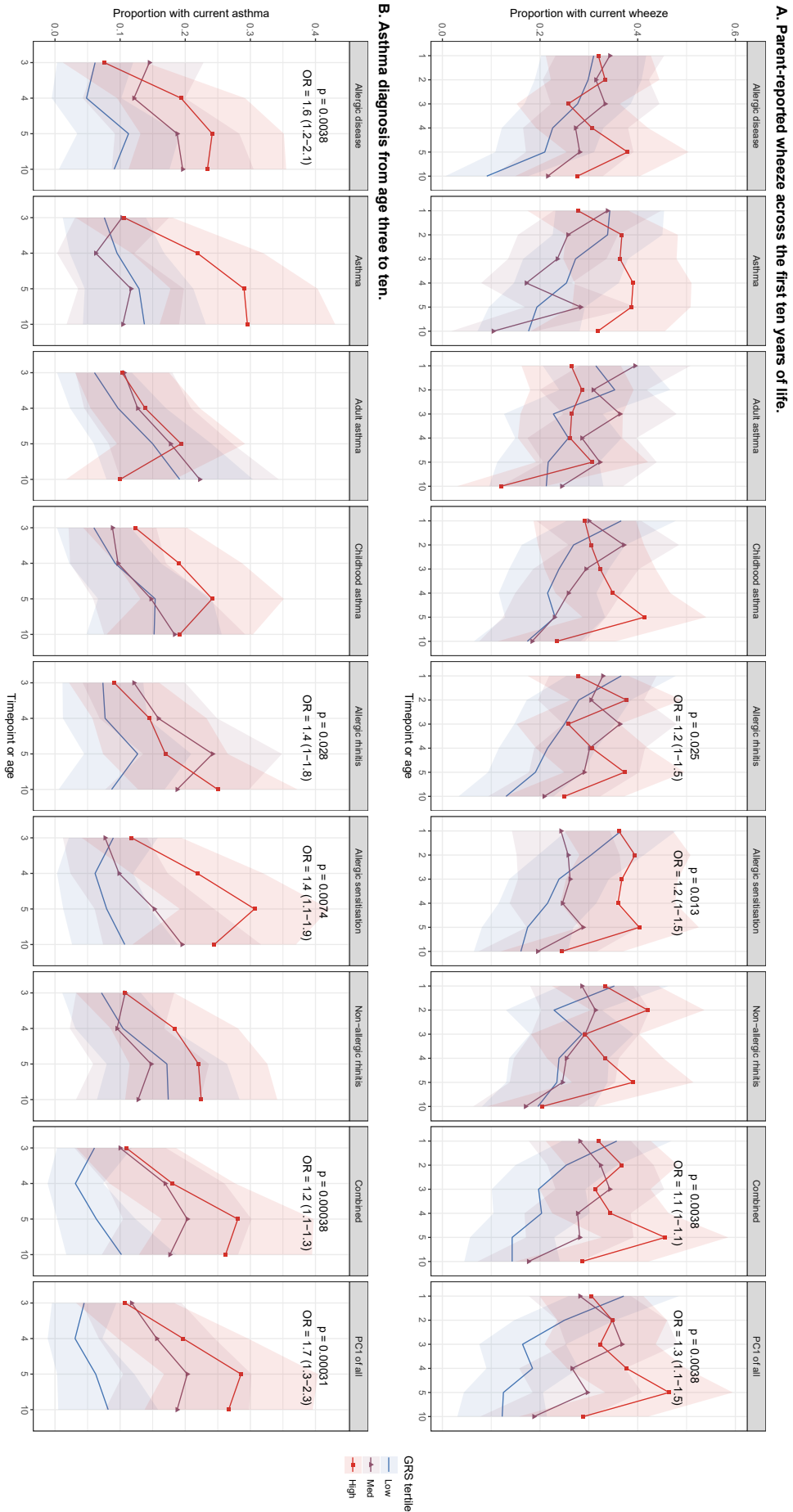
GLM of early-life trait ~ GRS + sex. Presented in each tile are the odds ratio of the association with each GRS; the heatmap fill colour represents the effect size of association (as log10(odds ratio)). Grayed tiles represent non-significant associations. The “npEM” clusters are clusters derived from non-parametric mixture models as in Tang et al [27]. Illness MPGs refer to those MPGs associated with respiratory illness (*Moraxella.d253*, *Streptococcus.4060*, *Haemophilus.bc0d*, *Haemophilus.f579*), as described in the previous chapter (Chapter 4). Trajectories (“traj.”) are clusters generated from nasopharyngeal microbiome data, also described in Chapter 4. Combined GRS refers to simple summation of all GRS, while PC1 of all GRS refers to the first principal component following PCA of all GRS.





**FIGURE 5.5:** Graphs showing relationships between GRS for allergic disease and (A) the high-risk npEM cluster from Tang et al; and (B) presence of illness-associated MPGs in healthy samples up to age two.

Violin plots and scatterplot of allergic disease GRS vs. specific phenotypes; for both panels, each point represents an individual from CAS. GRS = genomic risk score for allergic disease derived from Ferreira et al [46]. npEM clusters derived from Tang et al [27] (also **Chapter 3**); proportion of illness-associated MPGs treated as a binary variable of individuals that have low (<40%) vs. high ( $\geq 40\%$ ) proportion of all healthy samples clustered into illness-associated MPG, as reported in **Chapter 4**. Results of statistical analyses for both phenotypes (using GLMs) were statistically-significant ( $p = 0.0075$  for Cluster 3 and  $p = 0.031$  for illness-associated MPG, as previously reported in **Figure 5.4**).



**FIGURE 5.6: Prevalence of early-life traits versus timepoint, separated by GRS tertiles.**

(A) Parent-reported wheeze across the first five years of life; (B) Asthma diagnosis from age 3 to 5. Also indicated is statistical significance for associations between GRS and early-life traits, as determined by GEEs, with sex as covariate and age as within-subjects factor (repeated-measures). GEE model: early trait ~ GRS (standardised score, not tertile) + sex + timepoint | subject.

### 5.4.1 Some genetic associations with early-life traits were shared with known associations for asthma

Genome-wide scans of early-life traits – such as wheezy lower respiratory infections, rhinovirus-associated infections, and parent-reported wheeze – identified a few possible loci of interest. Some of these (*DPP10*, *MANEA*) had previously been linked to asthma traits, [51, 52] albeit not in studies with conventional GWAS experimental designs. Others were novel, and may be related to pathophysiology specific to infection or to bronchial obstruction due to inflammation. For instance, the link between *DNAH5* and wheezy LRIs may be explained by the role of dynein in providing motility to respiratory cilia, driving the mucociliary escalator and keeping the small airways clear of obstructions [55]. Other *DNAH5* variants, particularly those causing complete loss of function, have been implicated in primary ciliary dyskinesia (PCD), a disease similar to cystic fibrosis with mucus plugging and recurrent chest infections [56, 57]. It is possible that partial disruption to expression, localisation or function of *DNAH5* may be a common genetic variant that causes a mild wheezy phenotype in young infants, but without any of the serious detrimental effects seen in PCD. With regards to the suggestive link between *APOL3* and rhinovirus-A infection: various subtypes of apolipoprotein-L have been linked to immune functions, as well as inhibition of viral agents [58, 59]. *APOL1*, 2 and 3 are all highly-expressed in the lung, suggesting roles specialised to that organ [60]. Furthermore, it is known that some subtypes of rhinovirus-A rely on the LDL receptor for cell invasion [61], thus providing another possible link between mediators of lipid metabolism (apolipoprotein) and infection susceptibility.

Repeated-measures GWAS identified a few other new loci. A locus at *ZBTB20* was associated with asthma diagnosis in CAS: *ZBTB20* has been implicated in promoting plasma cell longevity and long-term antibody production [62]. *LINGO2* was associated with wLRIs in CAS, and elsewhere it has been linked to susceptibility for airway responsiveness in COPD [63]. Also associated with wLRIs in CAS was *FAM81A*, which had been found to be differentially expressed both in asthma and in response to RV-A16 infection [64]. Loci in the HLA region were associated with degree of nasopharyngeal colonisation by certain taxa, with *HLA-A* for *Staphylococcus.29eb* and *HLA-DRB1* for *Streptococcus.4060*. Similar relationships between MHC genes and microbial colonisation have been identified in the context of intestinal microbiota in mouse models, with subsequent links to altered physiological responses (IgA antibody production) and disease outcomes related to infection (*Salmonella* enteritis) [65] and autoimmunity (Type 1 diabetes) [66]. It is not improbable that similar relationships exist for nasopharyngeal microbiota and respiratory health, although it is beyond the scope of the current study to explore these ideas further.

We observed that many of the loci associated with asthma-related and microbiome-related traits were somehow related to proteins that interact with the Ras superfamily of proteins (Ras, Rho, Rab, Ran, Arf). These proteins are small GTPases that modulate a broad range of cellular and intracellular processes, including cell proliferation, movement, and vesicular trafficking [67]. For example: *TBC1D22A* was associated with *Alloiooccus.dd23* and encodes a GTPase-activating protein (GAP) for Rab GTPase [68]; and *GDI2* encodes a GDP dissociation inhibitor (GDI) that regulates Rab levels [69]. *CUL3* was associated with year-one febrile illnesses in the GWAS, while *KBTBD7* was associated with lower respiratory illness in the repeated-measures GWAS. Together the protein products *CUL3* and *KBTBD7* form part of a E3 ubiquitin ligase complex that regulates guanine nucleotide exchange factor (GEF) *TIAM1*, which in turn regulates the signalling of Rho GTPase *RAC1* [70]. The biological significance of these findings remain unclear, and may simply be related to the ubiquity of Ras superfamily proteins in physiological processes.

### 5.4.2 Genomic risk scores for later asthma-related traits were linked to early-life events, suggesting a route of pathology for genetic susceptibility to asthma

To work around the limited sample size of the CAS genetic data, we shifted our analysis to using genomic risk scores (GRS) derived from other larger GWAS. We selected specific GWAS that assessed phenotypes of allergic disease in large, (mostly) adult populations of similar ethnicity, such as the UK Biobank – hence our GRS represent cumulative genetic risk for later disease. The advantage of this method was that it leveraged genetic information from higher-powered studies, incorporating signal from the entire genome as well as possible epistatic effects between genes; and that it allowed us to temporally-link early-life events in CAS to risk of later disease in adulthood.

In doing so, we found that many GRS for allergic diseases and traits were associated with allergic sensitization and asthma diagnosis in early childhood, as expected. GRS for specific phenotypes (e.g. childhood asthma, allergic rhinitis) may not necessarily be correlated even with its own trait in CAS. This is likely due to a power limitation — the sample sizes of the GWAS used to generate GRS for these individual disease phenotypes were smaller than the “any allergic disease” GWAS of Ferreira et al [46]. Instead, a “composite GRS” — whether it be the allergic disease GRS from Ferreira et al, or the first PC of all GRS representing a condensed atopic signal — was associated with multiple early-life traits linked to asthma and allergy. The allergic disease GRS was negatively-associated with transient wheeze (wheezing up to age 3, but not at age 5), further reinforcing the transient wheeze phenotype as one distinct from entrenched asthma. The GRS was positively associated with later wheeze at ages five and ten. This was consistent with similar findings observed by Spycher et al, where a score based on the top 45 SNPs from the GABRIEL cohorts, applied to the ALPSAC cohort, yielded strong associations with persistent wheeze and not with transient early wheeze [25].

The allergic disease GRS was also associated with overabundance of illness-associated bacteria in healthy samples (*Haemophilus*, *Streptococcus*, *Moraxella*) – a known putative risk factor for later asthma. With a repeated-measures GEE model, we found that the GRS specifically for childhood asthma was associated with wLRI frequency in infancy, especially wLRIs positive for rhinovirus. All these findings painted a consistent portrait: that the genetic risk for asthma and allergic disease is likely acting in part through genetic susceptibility to complications from certain environmental exposures in early life, such as microbial colonization, respiratory infection, and contact with allergen.

Interestingly, we found a moderate correlation between GRS for allergic and non-allergic rhinitis. This correlation may be explained by the fact that: (1) tests commonly used to measure allergy (IgE levels, skin sensitisation tests) may produce false negatives; and (2) a majority of so-called “non-allergic” rhinitis may actually exhibit entropy – an allergic response due to localised IgE production in the nasal mucosa without systemic elevation in serum IgE [71].

The combined effect of genetics and environment may be represented by clusters discovered using non-parametric mixture models (npEM) in our previous publication [27] (see **Chapter 3**). Although principal components analysis failed to show any clear segregation between npEM clusters in terms of genetic risk for allergic disease, it was likely that some measure of this risk was still present within the npEM clusters. For instance, we found that the GRS for allergic disease was itself significantly-associated with membership in the high-risk npEM cluster (“Cluster 3”); the high-risk cluster corresponded to slightly higher allergic disease GRS. When we analysed for associations with early-life traits, with GRS and npEM cluster membership as predictors, we found that the inclusion of npEM cluster information diminished the effect of GRS, but not vice versa. This suggested that

the effect of GRS on early-life traits was partially-accounted for by cluster membership. It was likely that the npEM clusters included both polygenic and environmental signals, as the features used to determine the clusters incorporated susceptibility to respiratory infection and degree of allergen sensitisation [27]. Another possibility is that the npEM clusters embody epigenetic signals which have not been examined in this study, and are beyond our present scope.

## 5.5 Conclusions

Using data from a small but comprehensively-surveyed birth cohort, we examined the possible link between the genetics of asthma and allergic disease, and events in early childhood. In doing so, we found possible connections amongst early-life sensitization, respiratory infection (especially with rhinovirus), nasopharyngeal colonization with certain microbes, and genetic risk for allergy and asthma. Although we were unable to precisely pinpoint the genetic loci that contributed to these links, it is probable that asthma is the consequence of many genes interacting with each other; rather than a singular culprit as in Mendelian disease, or a small subset of genes acting in isolation. Based on analyses with genomic risk scores, we were able to conclude that asthma is likely the result of both polygenic and environmental factors, acting in concert to bring about early-life events, which then promote or protect against the development of disease as the child ages.

It is possible that certain environmental exposures in early life, such as tobacco smoke, diet, pets and other allergens, modulate the effect of genetics on disease pathology. However, it remains a challenge to integrate such information to study gene-environmental interactions in a well-powered manner. Further investigations of environmental covariates will likely require larger sample sizes, careful experimental design and specialized analytical methods.

## References

1. Ober C and Yao TC. The genetics of asthma and allergic disease: a 21st century perspective. *Immunol Rev* 2011;242:10–30.
2. Li X, Howard TD, Zheng SL, et al. Genome-wide association study of asthma identifies RAD50-IL13 and HLA-DR/DQ regions. *J Allergy Clin Immunol* 2010;125:328–335 e11.
3. Hirota T, Takahashi A, Kubo M, et al. Genome-wide association study identifies three new susceptibility loci for adult asthma in the Japanese population. *Nat Genet* 2011;43:893–6.
4. Torgerson DG, Ampleford EJ, Chiu GY, et al. Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nat Genet* 2011;43:887–92.
5. Moffatt MF, Gut IG, Demenais F, et al. A large-scale, consortium-based genomewide association study of asthma. *N Engl J Med* 2010;363:1211–21.
6. Bonnelykke K, Sleiman P, Nielsen K, et al. A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. *Nat Genet* 2014;46:51–5.
7. Simpson AB, Glutting J, and Yousef E. Food allergy and asthma morbidity in children. *Pediatr Pulmonol* 2007;42:489–95.

8. Sporik R and Platts-Mills TA. Allergen exposure and the development of asthma. *Thorax* 2001;56 Suppl 2:ii58–63.
9. Caliskan M, Bochkov YA, Kreiner-Moller E, et al. Rhinovirus wheezing illness and genetic risk of childhood-onset asthma. *N Engl J Med* 2013;368:1398–407.
10. Kull I, Melen E, Alm J, et al. Breast-feeding in relation to asthma, lung function, and sensitization in young schoolchildren. *J Allergy Clin Immunol* 2010;125:1013–9.
11. Simpson A, John SL, Jury F, et al. Endotoxin exposure, CD14, and allergic disease: an interaction between genes and the environment. *Am J Respir Crit Care Med* 2006;174:386–92.
12. Pasanen A, Karjalainen MK, Bont L, et al. Genome-Wide Association Study of Polymorphisms Predisposing to Bronchiolitis. *Sci Rep* 2017;7:41653.
13. Forton JT, Rowlands K, Rockett K, et al. Genetic association study for RSV bronchiolitis in infancy at the 5q31 cytokine cluster. *Thorax* 2009;64:345–52.
14. Bochkov YA, Watters K, Ashraf S, et al. Cadherin-related family member 3, a childhood asthma susceptibility gene product, mediates rhinovirus C binding and replication. *Proc Natl Acad Sci U S A* 2015;112:5485–90.
15. Kusel MM, de Klerk NH, Keadze T, et al. Early-life respiratory viral infections, atopic sensitization, and risk of subsequent development of persistent asthma. *J Allergy Clin Immunol* 2007;119:1105–10.
16. Jackson DJ, Gangnon RE, Evans MD, et al. Wheezing rhinovirus illnesses in early life predict asthma development in high-risk children. *Am J Respir Crit Care Med* 2008;178:667–72.
17. Bizzintino J, Lee WM, Laing IA, et al. Association between human rhinovirus C and severity of acute asthma in children. *Eur Respir J* 2011;37:1037–42.
18. Granada M, Wilk JB, Tuzova M, et al. A genome-wide association study of plasma total IgE concentrations in the Framingham Heart Study. *J Allergy Clin Immunol* 2012;129:840–845 e21.
19. Pino-Yanes M, Gignoux CR, Galanter JM, et al. Genome-wide association study and admixture mapping reveal new loci associated with total IgE levels in Latinos. *J Allergy Clin Immunol* 2015;135:1502–10.
20. Vicente CT, Revez JA, and Ferreira MAR. Lessons from ten years of genome-wide association studies of asthma. *Clin Transl Immunology* 2017;6:e165.
21. Ober C. Asthma Genetics in the Post-GWAS Era. *Ann Am Thorac Soc* 2016;13 Suppl 1:S85–90.
22. Abraham G, Havulinna AS, Bhalala OG, et al. Genomic prediction of coronary heart disease. *Eur Heart J* 2016;37:3267–3278.
23. Khera AV, Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* 2018;50:1219–1224.
24. Belsky DW, Sears MR, Hancox RJ, et al. Polygenic risk and the development and course of asthma: an analysis of data from a four-decade longitudinal study. *Lancet Respir Med* 2013;1:453–61.
25. Spycher BD, Henderson J, Granell R, et al. Genome-wide prediction of childhood asthma and related phenotypes in a longitudinal birth cohort. *J Allergy Clin Immunol* 2012;130:503–9 e7.

26. Teo SM, Tang HHH, Mok D, et al. Airway Microbiota Dynamics Uncover a Critical Window for Interplay of Pathogenic Bacteria and Allergy in Childhood Respiratory Disease. *Cell Host Microbe* 2018;24:341–352 e5.
27. Tang HH, Teo SM, Belgrave DC, et al. Trajectories of childhood immune development and respiratory health relevant to asthma and allergy. *Elife* 2018;7.
28. Consortium 1GP, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature* 2015;526:68–74.
29. Delaneau O, Marchini J, and Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods* 2012;9:179–81.
30. Howie B and Marchini J. IMPUTE2. 2014. URL: [https://mathgen.stats.ox.ac.uk/impute/impute\\_v2.html](https://mathgen.stats.ox.ac.uk/impute/impute_v2.html).
31. Das S, Forer L, Schonherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet* 2016;48:1284–1287.
32. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016;48:1279–83.
33. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, and Heckerman D. FaST linear mixed models for genome-wide association studies. *Nat Methods* 2011;8:833–5.
34. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;81:559–75.
35. Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, and Heckerman D. Improved linear mixed models for genome-wide association studies. *Nat Methods* 2012;9:525–6.
36. Turner SD. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *bioRxiv* 2014.
37. Aulchenko YS, Ripke S, Isaacs A, and van Duijn CM. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 2007;23:1294–1296.
38. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 2010;26:2336–7.
39. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, and Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 2010;26:2069–70.
40. Kent WJ, Sugnet CW, Furey TS, et al. The Human Genome Browser at UCSC. *Genome Research* 2002;12:996–1006.
41. MacArthur J, Bowler E, Cerezo M, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res* 2017;45:D896–D901.
42. Slowikowski K. proxysnps: Get proxy SNPs for a SNP in the 1000 Genomes Project. 2015. URL: <https://github.com/slowkow/proxysnps>.
43. Benjamini Y and Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995;57:289–300.

44. Ronnegard L, McFarlane SE, Husby A, Kawakami T, Ellegren H, and Qvarnstrom A. Increasing the power of genome wide association studies in natural populations using repeated measures - evaluation and implementation. *Methods Ecol Evol* 2016;7:792–799.
45. Valdar W, Holmes CC, Mott R, and Flint J. Mapping in structured populations by resample model averaging. *Genetics* 2009;182:1263–77.
46. Ferreira MA, Vonk JM, Baurecht H, et al. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nat Genet* 2017;49:1752–1757.
47. Demenais F, Margaritte-Jeannin P, Barnes KC, et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat Genet* 2018;50:42–53.
48. Pividori M, Schoettler N, Nicolae DL, Ober C, and Im HK. Shared and distinct genetic risk factors for childhood onset and adult onset asthma. *bioRxiv* 2018.
49. Waage J, Standl M, Curtin JA, et al. Genome-wide association and HLA fine-mapping studies identify risk loci and genetic pathways underlying allergic rhinitis. *Nat Genet* 2018;50:1072–1080.
50. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine* 2015;12:e1001779.
51. Huang YT, Liang L, Moffatt MF, Cookson WO, and Lin X. iGWAS: Integrative Genome-Wide Association Studies of Genetic and Genomic Data for Disease Susceptibility Using Mediation Analysis. *Genet Epidemiol* 2015;39:347–56.
52. Blakey JD, Sayers I, Ring SM, Strachan DP, and Hall IP. Positionally cloned asthma susceptibility gene polymorphisms and disease risk in the British 1958 Birth Cohort. *Thorax* 2009;64:381–7.
53. Cho MH, Castaldi PJ, Wan ES, et al. A genome-wide association study of COPD identifies a susceptibility locus on chromosome 19q13. *Hum Mol Genet* 2012;21:947–57.
54. Teo SM, Mok D, Pham K, et al. The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development. *Cell Host Microbe* 2015;17:704–15.
55. Roberts AJ, Kon T, Knight PJ, Sutoh K, and Burgess SA. Functions and mechanics of dynein motor proteins. *Nat Rev Mol Cell Biol* 2013;14:713–26.
56. Olbrich H, Haffner K, Kispert A, et al. Mutations in DNAH5 cause primary ciliary dyskinesia and randomization of left-right asymmetry. *Nat Genet* 2002;30:143–4.
57. Fliegauf M, Olbrich H, Horvath J, et al. Mislocalization of DNAH5 and DNAH9 in respiratory cells from patients with primary ciliary dyskinesia. *Am J Respir Crit Care Med* 2005;171:1343–9.
58. Taylor HE, Khatua AK, and Popik W. The innate immune factor apolipoprotein L1 restricts HIV-1 infection. *J Virol* 2014;88:592–603.
59. Wu X, Wang H, Bai L, et al. Mitochondrial proteomic analysis of human host cells infected with H3N2 swine influenza virus. *J Proteomics* 2013;91:136–50.
60. Duchateau PN, Pullinger CR, Cho MH, Eng C, and Kane JP. Apolipoprotein L gene family: tissue-specific expression, splicing, promoter regions; discovery of a new gene. *J Lipid Res* 2001;42:620–30.



61. Bochkov YA and Gern JE. Rhinoviruses and Their Receptors: Implications for Allergic Disease. *Curr Allergy Asthma Rep* 2016;16:30.
62. Recaldin T and Fear DJ. Transcription factors regulating B cell fate in the germinal centre. *Clin Exp Immunol* 2016;183:65–75.
63. Hansel NN, Pare PD, Rafaels N, et al. Genome-Wide Association Study Identification of Novel Loci Associated with Airway Responsiveness in Chronic Obstructive Pulmonary Disease. *Am J Respir Cell Mol Biol* 2015;53:226–34.
64. Bai J, Smock SL, Jackson G. R. J, et al. Phenotypic responses of differentiated asthmatic human airway epithelial cultures to rhinovirus. *PLoS One* 2015;10:e0118286.
65. Kubinak JL, Stephens WZ, Soto R, et al. MHC variation sculpts individualized microbial communities that control susceptibility to enteric infection. *Nat Commun* 2015;6:8642.
66. Silverman M, Kua L, Tanca A, et al. Protective major histocompatibility complex allele prevents type 1 diabetes by shaping the intestinal microbiota early in ontogeny. *Proc Natl Acad Sci U S A* 2017;114:9671–9676.
67. Wennerberg K, Rossman KL, and Der CJ. The Ras superfamily at a glance. *J Cell Sci* 2005;118:843–6.
68. Pan X, Eathiraj S, Munson M, and Lambright DG. TBC-domain GAPs for Rab GT-Pases accelerate GTP hydrolysis by a dual-finger mechanism. *Nature* 2006;442:303–6.
69. Sedlacek Z, Munstermann E, Mincheva A, Lichter P, and Poustka A. The human rab GDI beta gene with long retroposon-rich introns maps to 10p15 and its pseudogene to 7p11-p13. *Mamm Genome* 1998;9:78–80.
70. Genau HM, Huber J, Baschieri F, et al. CUL3-KBTBD6/KBTBD7 ubiquitin ligase cooperates with GABARAP proteins to spatially restrict TIAM1-RAC1 signaling. *Mol Cell* 2015;57:995–1010.
71. Powe DG, Jagger C, Kleinjan A, Carney AS, Jenkins D, and Jones NS. 'Entropy': localized mucosal allergic disease in the absence of systemic responses for atopy. *Clin Exp Allergy* 2003;33:1374–9.



## Chapter 6

# Conclusions

### 6.1 Summary of findings

The research output in this thesis can be summarised as follows:

#### 6.1.1 Chapter 3: Mixture-model clusters of asthma susceptibility

In this chapter, we applied an unsupervised mixture-model-based method of cluster analysis, to a prospective birth cohort (CAS) with featured data relating to immunorespiratory traits. By doing so, we found that:

- There were three groups of individuals with differential phenotypes, particularly in terms of prevalence or risk of later asthma, as well as predictors or potential pathophysiological drivers for asthma.
- In particular, there was a high-risk cluster (Cluster 3) with a phenotype consistent with very early sensitisation (particularly to house dust mite and peanuts), multiple allergen sensitisation, and persistent wheeze. This group had the greatest incidence of asthma as well as other allergic comorbidities. This high-risk cluster was also replicated in other cohorts (COAST and MAAS), for which the analogous clusters shared similar properties with the Cluster 3 observed in CAS.
- A low-risk cluster (Cluster 1) was characterised by low rates of wheeze. Those who had wheeze otherwise exhibited an early transient wheeze phenotype associated with respiratory infections.
- Another low-risk cluster (Cluster 2) was associated with IgG4 allergen sensitisation, with persistent wheeze being associated with both respiratory infections and IgE sensitisation to house dust mite (HDM).
- Allergic and infective processes acted additively to contribute to respiratory wheeze in early childhood.
- IgE levels fluctuated with age. Therefore, to describe or define atopy and disease risk, a method that incorporated multiple input variables such as one represented by a mixture model may be more appropriate than a single test that relied on fixed clinical thresholds for atopy (e.g. specific IgE).

#### 6.1.2 Chapter 4: Nasopharyngeal microbiome and asthma

In this chapter, we analysed the nasopharyngeal microbiome from two separate cohorts (CAS and COAST). From both healthy and illness-associated samples, we used an ASV-based bioinformatic pipeline to process 16S rRNA sequencing data into taxon-specific

relative abundances, which were then analysed with cluster analysis and association analyses. In doing so, we found that:

- The nasopharyngeal microbiome in young children was highly-structured, and was remarkably consistent between two independent populations (CAS and COAST). In both of these cohorts, the nasopharyngeal samples were dominated by taxa from six primary genera: *Alloiococcus/Dolosigranulum*, *Corynebacterium*, *Haemophilus*, *Moraxella*, *Staphylococcus*, and *Streptococcus*.
- In both CAS and COAST, illness-associated taxa representing the species *M. catarrhalis*, *S. pneumoniae*, and *H. influenzae*, were found more frequently in illness samples, confirming the findings from Teo et al [1]. These taxa tended to co-occur with winter season and colonisation with respiratory viruses. However, all three entities contributed semi-independently to risk of respiratory illness. These findings were shared between both CAS and COAST, with similar effect sizes for these associations.
- Patterns of correlation between individual bacterial taxa were also shared between CAS and COAST, with illness-associated bacteria tending to co-occur together, and likewise for health-associated bacteria.
- The nasopharyngeal microbiome followed distinct patterns or trajectories, each dominated by a unique taxa from an early age. One of these trajectories (e.g. early-life domination by *Staphylococcus.29eb*) was associated with later asthma, but only in COAST. The mechanism of association may be dependent on allergic sensitisation. In both cohorts, the healthy microbiome may be linked to later wheeze via abundance of illness-associated bacteria interacting with early allergic sensitisation.
- Despite frequent respiratory infections being a risk factor for asthma, the contributions of microbes to respiratory infection may be distinct from their contributions to later asthma risk.

### 6.1.3 Chapter 5: Genetics of asthma and early childhood events

Finally, in this research chapter, we performed a number of genome-related analyses, including GWAS on CAS data; and GRS from larger meta-analyses being calculated in CAS, followed by association tests between GRS and certain early-life traits in CAS. In doing so, we found that:

- A number of loci associated with lower respiratory infections or wheeze in early life included those previously associated with asthma traits, antibody production, and rhinovirus infection. Some loci in the HLA region were associated with nasopharyngeal colonisation by certain bacterial taxa (e.g. *HLA-A* for *Staphylococcus.29eb*, and *HLA-DRB* for *Streptococcus.4060*) There were a number of genome-wide associations with genes and proteins interacting with the Ras superfamily. The significance of this remains unclear. Further research with larger sample sizes may be needed.
- A genomic risk score (GRS) for any allergic disease was associated with multiple early-life traits linked to asthma and allergy in CAS.
- Patterns of association with GRS suggest that genetic risk for asthma and allergic disease is in part imposed via susceptibility to complications from certain environmental exposures in early life, such as microbial colonization, respiratory infection, and contact with allergen.

- The genetic risk embodied in allergic disease GRS is associated with, but does not account for all, the risk associated with the high-risk mixture-model cluster discovered in **Chapter 3**.

## 6.2 Overall contribution to knowledge

### 6.2.1 Revisiting the key questions

In **Chapter 1**, a number of key research questions were posed for this thesis. For ease of reference, these have been repeated below:

- Is it possible to use unsupervised clustering methods to derive clusters (presumed endotypes) of childhood asthma and asthma susceptibility from clinicopathological data? What do these clusters look like, and do they capture trajectories of childhood development relevant to immune or respiratory health and disease?
- How do these immunorespiratory clusters relate to existing subgroups of asthma susceptibility, or definitions of atopy and allergy? Do these clusters provide more information than existing criteria for allergy?
- Do similar clusters exist across different populations? How do these compare?
- Does characterisation of nasopharyngeal microbiota using ASVs differ much from using OTUs? Can similar findings be achieved to those of Teo et al using OTU-based results? Does the bacterial composition of nasopharyngeal microbiota contribute to respiratory disease dependently or independently of other risk factors such as season and viral detection?
- Are there clusters of individuals who share similar patterns of nasopharyngeal microbiome that evolve with time and age? Do any of these “microbial trajectories” relate to asthma risk?
- Are these associations between nasopharyngeal microbiome and respiratory disease shared across different populations?
- Are there any loci in the genome that are associated with early-life risk factors for asthma (e.g. frequency of lower respiratory infections, allergen-specific IgE levels)? if so, have any of these been replicated?
- Does incorporating the longitudinal aspect of some GWAS phenotypes (e.g. repeated measurements) grant more biologically-relevant information and hence generate any new findings with longitudinal GWAS?
- Does the genetic signal for allergy disease later in life, represented by genomic risk scores (GRS), associate with early childhood traits such as allergic sensitisation, microbial colonisation, and wheezy respiratory infections? How do immunorespiratory clusters, microbiome, and genomics interact with each other when contributing to asthma risk?

### 6.2.2 Insights from this thesis

We applied an unsupervised clustering method incorporating mixture models (npEM) and machine learning to CAS, a paediatric dataset with measurements related to immunorespiratory health. In doing so, we generated three clusters, each with a unique risk profile

for asthma as well as distinct pathophysiological mechanisms for each. These clusters, particularly the high-risk one, were replicated in other cohorts (COAST, MAAS), demonstrating a degree of universality to our findings. Overall, our mixture-model clusters were informative of the nature of asthma heterogeneity, and gave some insight into the cause of this heterogeneity — variable degrees of allergic sensitisation and respiratory infection, likely driven in part by genetics and environmental exposures. Because the mixture models represent an amalgamation of risk contributed by multiple inputs, they were better representations of disease risk than fixed thresholds of atopy (“positive” results on IgE or SPT). It remains a challenge to simplify or modify these mixture models into a condensed algorithm that is clinically useful; for instance, decision trees derived from CAS npEM clusters did not replicate well in external cohorts, and hence could not be applied broadly as simple decision-making algorithms for risk stratification. However, the mixture model does provide a flexible means to make predictions in external datasets based on parameters learned from a training set. The future may see the rise of machine-assisted diagnosis or screening methods, driven by complex models such as mixture model classification or neural networks rather than simple decision trees. These complex models incorporate multiple biometric inputs covering a broad range of domains (genomics, immunomics, microbiome) to produce a clinically-interpretable output (e.g. probability of disease, or membership in a particular risk profile).

Using an ASV-based taxonomy, we were able to describe the nasopharyngeal microbiome of children in COAST as being structured and very similar to CAS, with the same associations existing between illness-associated taxa (*Moraxella*, *Streptococcus*, *Haemophilus*) and respiratory infections in early childhood. Our findings were consistent with those of OTU-based CAS analyses reported in Teo et al 2018. Furthermore, we identified that although these taxa often co-occurred with viral infection (rhinovirus, RSV) and winter season, they all contributed semi-independently to the pathology of respiratory infection. There was also evidence to suggest that certain trajectories of healthy nasopharyngeal microbiome, as well as very early colonisation with certain taxa, were linked to later asthma outcomes. However, these associations with microbiome trajectory were inconsistent between CAS and COAST. Overall, we were able to at least confirm that the nasopharyngeal microbiome may be linked to respiratory health in young children, via mechanisms which may interact with early allergic sensitisation. Finally, we attempted to determine genetic contributions to asthma and asthma-related risk factors in the CAS dataset. Although we had limited statistical power, we were able to identify both previously-known loci as well as novel loci for asthma and related traits. There were suggestive links to loci near biologically-plausible genes: for example, genes that determine function of respiratory cilia (*DNAH5*), viral susceptibility (*APOL3*, *FAM81A*), asthma and airway responsiveness (*ZBTB20*, *LINGO2*), and microbial colonisation (HLA region). In addition, when we used summary statistics from large-scale GWAS to generate genomic risk scores (GRS) for allergic disease, and applied these to CAS, we found that these scores were associated with multiple early-life traits. These included both expected associations with allergic sensitisation, and unexpected ones involving illness-associated microbial taxa and membership in the high-risk npEM cluster.

### 6.3 The future

This thesis has provided a demonstration of system-based approaches at work, in uncovering the pathogenesis of complex diseases such as asthma and allergy. We were able to show that an unsupervised mixture model can provide informative risk profiling that is more nuanced than existing methods. We were also able to draw a connection to nasopharyngeal

microbiota, and demonstrate both similarities and differences in microbial-disease associations between two birth cohorts from different geographical regions. Finally, we were able to show that genetics plays an important role in asthma pathogenesis, by increasing susceptibility to early-life sensitisation as well as nasopharyngeal microbial colonisation. These genetic factors likely interact with environmental factors such as climate (geography, season), exposure to allergen and exposure to respiratory pathogens, to modify childhood susceptibility to persistent wheeze and asthma. As a whole, the thesis thoroughly explored the genetics, microbiomics, immunology and pathophysiology of early childhood asthma and allergy, and was thus able to identify key factors that significantly modified the risk of later disease. The contents of this thesis will hopefully aid future researchers in further examinations of the pathogenesis of asthma and allergy: for instance, one may begin by deriving similar npEM clusters in their own datasets using a classifier or similar tool, then comparing them in terms of other “omics”-based datasets not yet measured in CAS — such as the epigenomics and transcriptomics of specific cell compartments including immune cell lineages and airway epithelium.

The common theme amongst all three research chapters was the demonstrable utility and versatility of clustering methods to simplify data and potentially identify hidden information. This was shown by the use of unsupervised cluster analysis to generate clusters of children with differential disease risk (npEM clusters representing immunorespiratory trajectories); clusters of nasopharyngeal samples with similar patterns of microbial composition (MPGs); and clusters relating to child-specific patterns of nasopharyngeal microbiome that evolved with age (microbial trajectories). This was often followed by the more targeted use of association analyses to determine links between physiological entities, whether it be to quantify differences between clusters, or to search for disease associations within clusters. Similar methods may be applied to other biomedical problems beyond asthma and allergy, especially those related to diseases with complex aetiology and pathophysiology such as cardiovascular disease, diabetes and cancer.

Since the commencement of this thesis, there have been numerous recent developments in omics-based technology as well as biostatistical and bioinformatics methodology. In particular, there has been a shift towards whole-genome and exome sequencing, with its associated improvement in fidelity of bioinformatic signal. In addition, there has been an emphasis on exploring epigenomics and its impact on modifying gene expression and hence phenotype. There have also been further developments in inference of causality and drawing of connections between biological entities: methods such as mediation analysis and Mendelian randomisation have become more commonplace in recent years. In other words, there are now a host of new systems-based tools that we can use to build on the results of this thesis, and further our examination of complex disease.

In time, it is hoped that technological and methodological advancements can reach a stage where we may be able to accurately describe the “omic” status of an individual. Using clustering or dimension reduction methods, we may then be able to generalise or categorise the physiologic state of each individual as a “profile”. Further analyses within each profile or individual may allow us to determine relevant pathways of pathophysiology, including critical points of the pathway that may act as possible “railroad switches” for medical intervention. From these critical switches, we can then determine appropriate means of medical treatment and risk modification unique to each profile. All of these developments represent early steps towards the ultimate goal of precision or personalised medicine.





## Appendix A

# ePrint of Chapter 3

Chapter 3 was published in eLife on October 2018. The ePrint and its supplementary material can be found at <https://elifesciences.org/articles/35856>. The following pages are a facsimile of the paper published online.

Summary statistics for several analyses can be found at [https://figshare.com/articles/Supplementary\\_File\\_1\\_1/6934052](https://figshare.com/articles/Supplementary_File_1_1/6934052) and [https://figshare.com/articles/Supplementary\\_File\\_1\\_2/6934055](https://figshare.com/articles/Supplementary_File_1_2/6934055).

The rest of this page has been intentionally left blank.



# Trajectories of childhood immune development and respiratory health relevant to asthma and allergy

Howard HF Tang<sup>1,2\*</sup>, Shu Mei Teo<sup>1,3</sup>, Danielle CM Belgrave<sup>4</sup>, Michael D Evans<sup>3,5</sup>, Daniel J Jackson<sup>5</sup>, Marta Brozynska<sup>1,4</sup>, Merci MH Kusel<sup>6</sup>, Sebastian L Johnston<sup>7</sup>, James E Gern<sup>5</sup>, Robert F Lemanske<sup>5</sup>, Angela Simpson<sup>8</sup>, Adnan Custovic<sup>4</sup>, Peter D Sly<sup>6,9</sup>, Patrick G Holt<sup>6,9</sup>, Kathryn E Holt<sup>10,11</sup>, Michael Inouye<sup>1,3,12\*</sup>

<sup>1</sup>Cambridge Baker Systems Genomics Initiative, Baker Heart and Diabetes Institute, Victoria, Australia; <sup>2</sup>School of BioSciences, The University of Melbourne, Victoria, Australia; <sup>3</sup>Cambridge Baker Systems Genomics Initiative, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom; <sup>4</sup>Department of Paediatrics, Imperial College London, London, United Kingdom; <sup>5</sup>University of Wisconsin School of Medicine and Public Health, Madison, United States; <sup>6</sup>Telethon Kids Institute, University of Western Australia, Perth, Australia; <sup>7</sup>Airway Disease Infection Section, MRC & Asthma UK Centre in Allergic Mechanisms of Asthma, National Heart and Lung Institute, Imperial College London, London, United Kingdom; <sup>8</sup>Division of Infection, Immunity and Respiratory Medicine, The University of Manchester, Manchester, United Kingdom; <sup>9</sup>Child Health Research Centre, The University of Queensland, Brisbane, Australia; <sup>10</sup>Bio21 Molecular Science and Biotechnology Institute, The University of Melbourne, Victoria, Australia; <sup>11</sup>The London School of Hygiene and Tropical Medicine, London, United Kingdom; <sup>12</sup>The Alan Turing Institute, London, United Kingdom

**\*For correspondence:**

Howard.Tang@baker.edu.au (HHFT);  
mi336@medschl.cam.ac.uk (MI)

**Competing interests:** The authors declare that no competing interests exist.

**Funding:** See page 27

**Received:** 12 February 2018

**Accepted:** 05 October 2018

**Published:** 15 October 2018

**Reviewing editor:** M Dawn Teare, University of Sheffield, United Kingdom

© Copyright Tang et al. This article is distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use and redistribution provided that the original author and source are credited.

**Abstract** Events in early life contribute to subsequent risk of asthma; however, the causes and trajectories of childhood wheeze are heterogeneous and do not always result in asthma. Similarly, not all atopic individuals develop wheeze, and vice versa. The reasons for these differences are unclear. Using unsupervised model-based cluster analysis, we identified latent clusters within a prospective birth cohort with deep immunological and respiratory phenotyping. We characterised each cluster in terms of immunological profile and disease risk, and replicated our results in external cohorts from the UK and USA. We discovered three distinct trajectories, one of which is a high-risk 'atopic' cluster with increased propensity for allergic diseases throughout childhood. Atopy contributes varyingly to later wheeze depending on cluster membership. Our findings demonstrate the utility of unsupervised analysis in elucidating heterogeneity in asthma pathogenesis and provide a foundation for improving management and prevention of childhood asthma.

DOI: <https://doi.org/10.7554/eLife.35856.001>

## Introduction

Asthma is a global health problem, and there is a pressing need for better understanding of its pathogenesis (*Global Initiative for Asthma, 2015*). Asthma is strongly associated with allergy, and both genetic and environmental factors may be involved (*Ober and Yao, 2011; Dick et al., 2014*). The 'hygiene hypothesis' proposes that modern changes to hygiene, sanitation and living environment

**eLife digest** Asthma causes wheezy and troubled breathing, and can be life-threatening. Scientists and doctors understand that asthma begins in early childhood. Chest infections, exposure to bacteria, viruses, and allergies may cause or trigger asthma. One person with asthma may not have the same origins as another. But it is not yet clear how various triggers may interact to trigger or exacerbate asthma.

To disentangle how these factors contribute to asthma, experts have tried to group people with asthma into subgroups. Unfortunately, the groups often vary from expert to expert. Now, some scientists are using computers to sort patients with asthma. The scientists let the computers decide the best criteria for sorting patients. This way the machines may identify patterns that are not obvious to humans.

Using this computer-based approach, Tang et al. sorted Australian children with asthma into 3 groups based on their early life allergies and respiratory health. One group has high-risk asthma with frequent chest infections and strong allergic responses. The other two groups are low-risk, but they respond differently to allergy and infection. Common tests used by doctors to diagnose patients with allergy or asthma may not work the same with all three groups. The bacteria found in the nose influence the risk of asthma, even in patients who are well, and the way this occurs varies by group. Similar groups were also found among children with asthma in the United States and the United Kingdom.

Learning more about subgroups of patients with asthma may help other scientists and doctors design better ways to diagnose, treat, or prevent asthma. Working together with scientists around the world to determine how to best describe subgroups of people according to asthma type and risk is a critical step in the process. Tang et al. hope other scientist will test whether these three groups are also found in people from other parts of the world.

DOI: <https://doi.org/10.7554/eLife.35856.002>

have modified human exposures to microbes, with subsequent effects on early-life immune development (Okada et al., 2010). However, the clinical presentation and prognosis of childhood wheeze is highly variable: some children remit; others remit but relapse; and yet others have wheeze persisting into adult asthma (Morgan et al., 2005). These differences suggest that the underlying causes of disease also differ from person to person. For example, while asthma is commonly linked to allergy, not all individuals with wheeze are sensitised to allergen, and vice versa (Spycher et al., 2010). As such, childhood asthma is a heterogeneous condition (Hekking and Bel, 2014; Wenzel, 2012), and this greatly complicates the study of its pathogenesis (Anderson, 2008). We postulate that there are subpopulations in early childhood, each sharing similar patterns of pathophysiology, disease susceptibility and phenotype that permit categorisation into clusters. If we can agnostically identify these clusters, then we may explore the biological mechanisms that underlie them, and find targets for early intervention that are specific for different asthma subtypes.

Previous attempts at subtyping asthma susceptibility relied on supervised classification, using expert knowledge and cut-offs to define clusters. For example, criteria such as – specific immunoglobulin E (IgE)  $\geq 0.35$  kU/L; wheal diameter  $\geq 3$  mm in a skin prick test (SPT); or symptom score surpassing a threshold – may determine classification into a high-risk profile (Castro-Rodríguez et al., 2000; Frith et al., 2011). However, these cut-offs vary with age, gender or other parameters, and may not accurately reflect true attribution of risk (Linden et al., 2011). Hence, they often continue to produce heterogeneous groups. Furthermore, previous studies tended to focus on a single ‘domain’, for instance grouping only by immunological response (Prescott et al., 1999), symptomatology or timing of disease (Martinez et al., 1995; Kurukulaaratchy et al., 2003). Recently, researchers have turned to unsupervised approaches, such as model-based cluster analysis and latent class analysis (LCA) (Deliu et al., 2016; Lazic et al., 2013; Simpson et al., 2010; Belgrave et al., 2014; Belgrave et al., 2013; Wu et al., 2015). These do not require experts to supply cut-offs, but can instead ‘learn’ boundaries from the data. They can potentially uncover patterns of similarity not immediately obvious to the human eye. Finally, these methods can cover a broader

range of domains, incorporating measurements from multiple sources to determine clusters that are potentially informative of asthma risk.

Here, we use a data-driven unsupervised framework together with a comprehensively phenotyped birth cohort, to define developmental trajectories during preschool years, a period known to be critical to asthma pathogenesis. Specifically, we (1) use non-parametric mixture models to discover latent clusters that define early childhood trajectories of immune function and susceptibility to respiratory infection; (2) investigate how these clusters relate to differential profiles of asthma susceptibility, and to existing definitions of atopy; (3) identify risk factors for asthma within each cluster; and (4) externally validate the clusters in independent cohorts.

## Results

To characterise the broad structure of an Australian dataset of young children (Childhood Asthma Study, CAS), we performed principal components analysis (**Figure 1—figure supplement 1**). Afterwards, to explicitly model the heterogeneous data types of the cohorts as well as explicitly identify clusters, we used non-parametric expectation-maximisation (npEM) mixture models (Materials and methods). By applying npEM-based clustering and classification to CAS, we identified three distinct clusters from 217 individuals and 174 clustering features (**Figure 1**): low-risk CAS1 ( $N = 88$ , 25% wheeze at age 5), low-risk but allergy-susceptible CAS2 ( $N = 107$ , 21% wheeze at age 5) and high-risk CAS3 ( $N = 22$ , 76% wheeze at age 5). Forty-six individuals in CAS had excessive missing data and were not classifiable. The CAS clusters satisfied basic measures of internal stability and were distinguishable on a PCA plot of the complete-case dataset (**Figure 1—figure supplement 1**). A graphical summary of results for the CAS clusters is presented in **Figure 2**.

### CAS1: low-risk, non-atopic cluster with transient wheeze

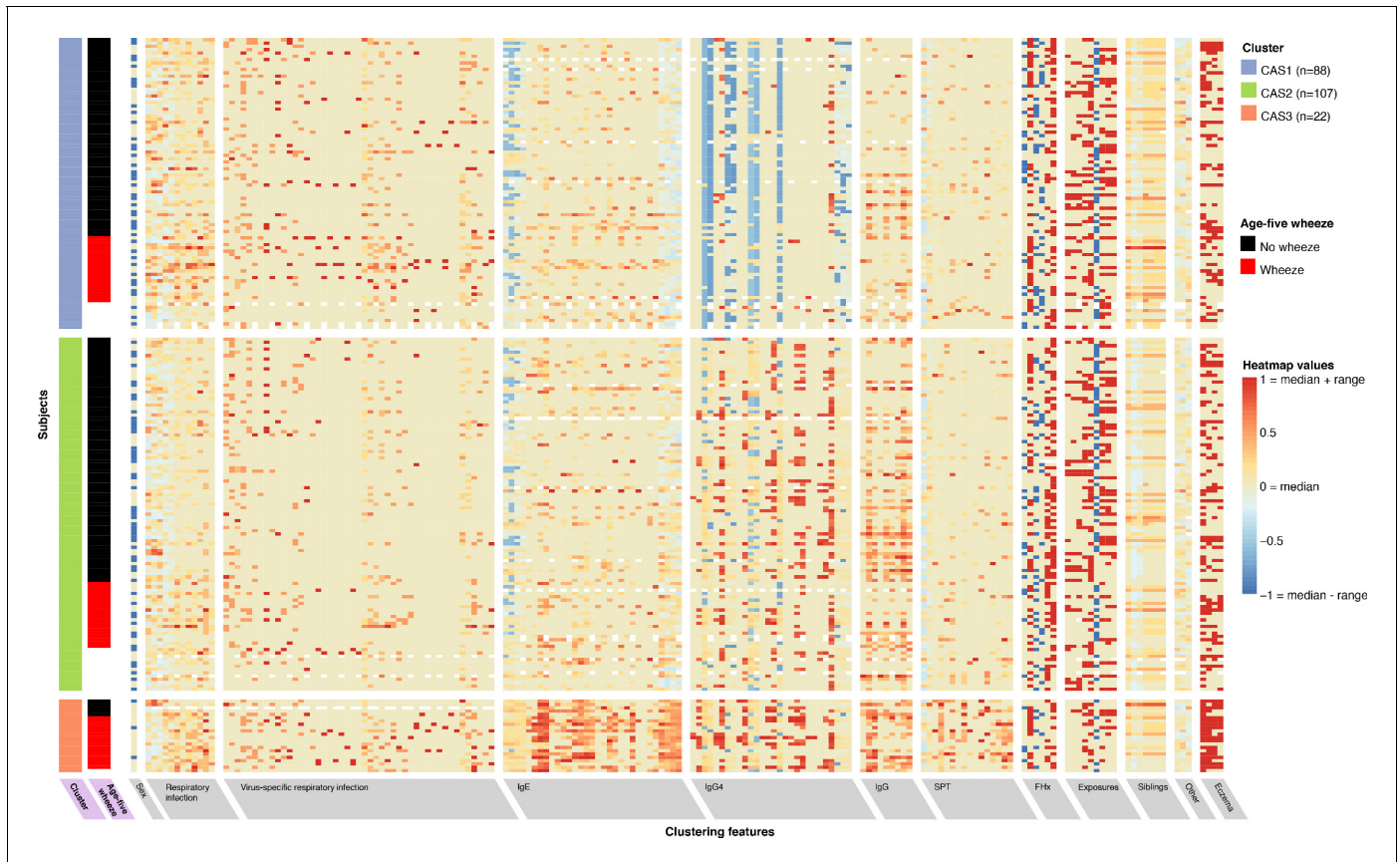
CAS1 was a low-risk cluster with infrequent and transient respiratory wheeze. Rates of wheeze declined from 33% at age 1 to 12% by age 10 (**Table 1**; **Figure 3**). In this cluster, Th2 cytokine responses of peripheral blood mononuclear cells (PBMCs) to allergen stimulation were minimal; and rates of allergen sensitisation (as measured by IgE or skin prick test, SPT) were the lowest among all groups (**Table 2**; **Figure 4**; **Supplementary file 1 – table supplement 3B-D**). IgG and IgG4 were also low across all allergens.

Frequency of respiratory infection in CAS1 was low (**Table 3**). However, high frequency of lower respiratory infections (LRIs) in childhood, especially wheezy LRIs (wLRIs), was a risk factor for age-5 wheeze – even after adjusting for sex, body mass index (BMI) and parental history of asthma as demographic covariates (**Table 4**). Repeated-measures ANOVA identified that LRI and wLRI frequency in the first 3 years were predictors for age-5 wheeze (**Supplementary file 1 – table supplement 4**); however, timepoint-specific analyses showed that differences were only noticeable from age 3 onwards (**Table 4**; **Figure 5A–B**). A multiple regression model with stepwise elimination yielded three significant variables: age-three wLRI frequency (odds ratio OR 5.6 per unit increase,  $p=0.0068$ ); age-four LRI frequency (OR 3.6,  $p=0.018$ ); and a protective effect from proportion of infection-associated microbiome profile groups (MPGs; *Streptococcus*, *Haemophilus*, *Moraxella*) in age-two-to-four healthy nasopharyngeal aspirate samples (NPAs; OR 0.19 per quartile,  $p=0.014$ ).

### CAS2: low-risk cluster susceptible to atopic and non-atopic wheeze

Similar to CAS1, CAS2 was a low-risk cluster with infrequent allergic disease. Compared to CAS1, Phadiatop and house dust mite (HDM) IgE were elevated at most timepoints (**Table 2**; **Figure 4A**; **Supplementary file 1 – table supplement 3B**), with the exception of peanut IgE (Wilcoxon, adjusted  $p=0.99$  at all timepoints; **Figure 4D**). CAS2 IgG and IgG4 were intermediate between CAS1 and CAS3 levels; CAS2 IgG was closer to CAS1, while CAS2 IgG4 was closer to CAS3 (**Table 2**; **Figure 4**). Despite these antibody differences, yearly rates of wheeze in CAS2 remained comparable to CAS1 (30% at age 1, declining to 18% at age 10; **Table 1**; **Figure 3**). Interestingly, compared to CAS1, individuals in CAS2 had fewer older siblings living in the household at age 2, as well as more frequent paternal history of asthma (adjusted  $p=0.029$  and 0.055, respectively; **Supplementary file 1 – table supplement 3A**).

Predictive factors for age-5 wheeze in CAS2 included: LRI, wLRI and febrile LRI (fLRI) frequency (GLM;  $p=2.7 \times 10^{-3}$ , 0.016 and 0.02 at age 3, respectively); HDM IgE ( $p=0.016$  and 0.011 at ages 2



**Figure 1.** Non-parametric mixture-model-based clustering of CAS dataset, based on 174 features. SPT = skin prick test. White spaces within the heatmap indicate missing data. Rows represent individuals; columns represent clustering features with general categories as labelled on grey background. Variables with grey background are clustering features ordered by category or type of variable first (e.g. all HDM IgE-related variables grouped together), then by timepoint (earlier to later, from left to right). Variables with lilac background indicate resultant cluster membership and outcome variable (age-5 wheeze). Heatmap values are scaled relative to range and median values for each feature; the median is coloured beige-yellow, the median +range red, and median - range blue. For sex, -1/blue = female, 0/yellow (median) = male.

DOI: <https://doi.org/10.7554/eLife.35856.003>

The following figure supplements are available for figure 1:

**Figure supplement 1.** Scatterplot of principal components analysis (PCA) of the complete-case CAS dataset ( $N = 186$ ), with points coloured by npEM clusters. Each point represents an individual.

DOI: <https://doi.org/10.7554/eLife.35856.004>

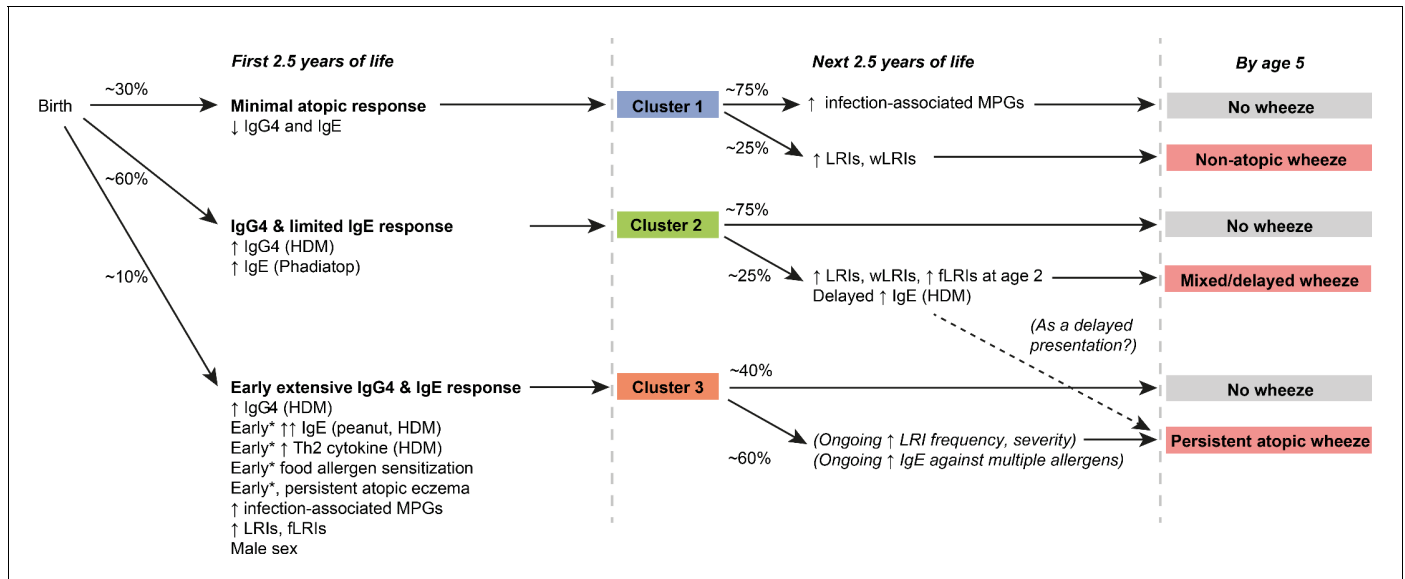
**Figure supplement 2.** Silhouette widths of clusters generated by npEM.

DOI: <https://doi.org/10.7554/eLife.35856.005>

**Figure supplement 3.** Overview of study methodology.

DOI: <https://doi.org/10.7554/eLife.35856.006>

and 4, respectively); and Phadiatop IgE ( $p=0.01$  at age 4) (**Table 4**). Repeated-measures ANOVA showed that HDM IgE and LRI-related variables (LRI, wLRI, fLRI) from the first 3 years were significant predictors of age-5 wheeze (**Supplementary file 1** – table supplement 4). Timepoint-specific analyses showed that differences were observable in HDM IgE and fLRI from age 2 onwards, while in LRI and wLRI they were only noticeable from age 3 (**Table 4**; **Figure 5**). A multiple regression model with stepwise elimination identified three significant variables: age-2 fLRI (OR eight per unit increase,  $p=0.0075$ ), age-4 wLRI (OR 5.3  $p=0.0016$ ), and age-4 Phadiatop IgE (OR 3.3,  $p=0.0088$ ). But although both IgE-related and infection-related risk factors contributed to age-5 wheeze, there was no significant evidence of interaction between them ( $p=0.36$  within CAS2 alone,  $p=0.92$  across entire cohort, for age-4 wLRI frequency  $\times$  Phadiatop IgE). Overall, CAS2 represented a low-risk trajectory susceptible to, but not necessarily afflicted by, wheeze due to atopic and non-atopic risk factors. In



**Figure 2.** Graphical summary of proposed clusters \*‘Early’ specifically refers to ‘within the first 6 months of life’.

DOI: <https://doi.org/10.7554/eLife.35856.007>

this cluster, atopic determinants of age-5 wheeze were only active from age 2 onwards, suggesting delayed atopic wheeze in this cluster. This duality of atopic and non-atopic risk factors for wheeze in this cluster was further supported by decision tree analysis, which identified that wheezy LRI frequency and HDM IgE best separated wheezers from non-wheezers in CAS2 (Figure 5—figure supplement 3).

### CAS3: high-risk atopic cluster with persistent wheeze

CAS3 was a ‘high-risk’ cluster, where persistent respiratory wheeze and atopic disease was seen in more than half the group throughout the first 10 years of life (Table 1; Figure 3). This cluster was dominated by males (86%, Fisher exact test, unadjusted  $p=6.8 \times 10^{-3}$  compared to CAS1, Table 1), and appeared to represent an early- and multi-sensitised atopic phenotype with persistent wheeze. CAS3 had elevated IgE, IgG, and IgG4 responses to common allergens, especially Phadiatop, HDM and peanut IgE from 6 months onwards (Table 2; Figure 4; Supplementary file 1 – table supplement 3B). SPTs were also more frequently positive in CAS3, especially to HDM and food allergens (peanut, cow’s milk and egg white, Supplementary file 1 – table supplement 3D).

No strong predictors for age-5 wheeze were identified within CAS3 (Table 4): only couch grass IgE at age 2 and acute respiratory infection (ARI) frequency at age 1 were weakly significant (both  $p=0.046$ ). Neither of these reached statistical significance when incorporated in the same model. However, the prolific IgE response, and the frequency and severity of early-life LRIs in this cluster (Table 3), strongly suggest contribution from both atopic and non-atopic causes of wheeze. Hence, CAS3 primarily represented those with extreme levels of atopic sensitisation and infection. The relative paucity of identifiable predictors may be explained by the small size of CAS3 ( $N = 22$ ), the intrinsically high rate of wheeze in the cluster (76% with age-5 wheeze), and saturation of risk from high levels of IgE and frequent infections.

Unlike the antibody measurements, cytokine measurements were excluded as clustering features due to high missingness. Nonetheless, with post-hoc analyses, we found that in vitro stimulation of PBMCs with HDM antigen elicited stronger Th2 cytokine responses in CAS3 compared to other clusters (Table 2, Figure 6). These cytokines (IL-4, IL-5, IL-13) were elevated from a very young age (Wilcoxon, adjusted  $p=4.6 \times 10^{-5}$  for IL-4 mRNA at age 6 m, compared to CAS1), coinciding with increase in HDM IgE and IgG4 responses. Weaker but similar differences were observed for peanut- and ovalbumin-stimulated PBMCs at 6 months (unadjusted  $p<0.05$  for all, Supplementary file 1 –

**Table 1.** Comparison of selected demographic and clinical variables in CAS clusters

Variable	Age (y)	Cas1 (N = 88)	Cas2 (N = 107)	Cas3 (N = 22)	P-value (unadjusted)			Feature?	
		Prop. (95% CI)	Prop. (95% CI)	Prop. (95% CI)	Overall	Cas1 vs. 2	Cas1 vs. 3		Cas2 vs. 3
Sex = male		55% (44–65%)	51% (42–61%)	86% (71–100%)	<b>7.3E-03</b>	0.67	<b>6.8E-03</b>	<b>3.7E-03</b>	Yes
Maternal asthma		51% (40–62%)	41% (32–51%)	59% (37–81%)	0.19	0.19	0.63	0.16	Yes
Paternal asthma		22% (13–30%)	44% (35–54%)	23% (3.7–42%)	<b>2.2E-03</b>	<b>1.3E-03</b>	1	<i>0.093</i>	Yes
Wheeze	1	33% (23–43%)	30% (21–39%)	55% (32–77%)	0.092	0.76	<i>0.084</i>	<b>0.046</b>	No
	5	25% (15–35%)	21% (13–30%)	76% (56–96%)	<b>7.1E-06</b>	0.59	<b>2.6E-05</b>	<b>3.4E-06</b>	No
	10	12% (3.4–21%)	18% (8.4–27%)	50% (24–76%)	<b>3.1E-03</b>	0.46	<b>1.5E-03</b>	<b>0.011</b>	No
Asthma	5	15% (7–23%)	13% (5.9–20%)	52% (29–76%)	<b>4.1E-04</b>	0.83	<b>7.7E-04</b>	<b>2.1E-04</b>	No
	10	10% (2.3–18%)	15% (6.1–23%)	56% (30–81%)	<b>2.6E-04</b>	0.59	<b>1.8E-04</b>	<b>7.9E-04</b>	No
Eczema	6m	39% (28–49%)	45% (35–54%)	91% (78–100%)	<b>2.4E-05</b>	0.47	<b>7.9E-06</b>	<b>9.0E-05</b>	Yes
	1	34% (24–44%)	30% (21–39%)	82% (64–99%)	<b>2.5E-05</b>	0.54	<b>7.2E-05</b>	<b>1.4E-05</b>	Yes
	5	28% (18–37%)	24% (16–33%)	71% (50–92%)	<b>2.1E-04</b>	0.73	<b>3.3E-04</b>	<b>7.9E-05</b>	No
Atopic rhinoconjunctivitis	5	30% (20–40%)	39% (29–49%)	76% (56–96%)	<b>6.4E-04</b>	0.21	<b>2.7E-04</b>	<b>3.2E-03</b>	No
		Mean (95% CI)	Mean (95% CI)	Mean (95% CI)	Overall	Cas1 vs. 2	Cas1 vs. 3	Cas2 vs. 3	
BMI (kg/m <sup>2</sup> )	3	16 (16–17)	16 (16–17)	16 (16–17)	0.86	0.65	0.68	0.8	No*
	4	16 (16–17)	16 (16–16)	17 (16–17)	0.59	0.76	0.32	0.39	No
	5	16 (16–16)	16 (16–16)	16 (15–17)	0.71	0.56	0.48	0.67	No
	10	18 (17–19)	18 (17–18)	18 (17–19)	0.89	0.75	1	0.62	No
Number of older siblings	0	0.93 (0.72–1.1)	0.53 (0.38–0.69)	0.77 (0.32–1.2)	<b>4.5E-03</b>	<b>1.0E-03</b>	0.37	0.25	Yes
	2	0.85 (0.66–1)	0.5 (0.34–0.65)	0.77 (0.32–1.2)	<b>2.8E-03</b>	<b>6.5E-04</b>	0.48	0.16	Yes
	5	0.68 (0.5–0.85)	0.39 (0.25–0.54)	0.67 (0.23–1.1)	<b>0.016</b>	<b>5.1E-03</b>	0.75	0.12	No
		Geom. mean (95% CI)	Geom. mean (95% CI)	Geom. mean (95% CI)	Overall	Cas1 vs. 2	Cas1 vs. 3	Cas2 vs. 3	
Vitamin D (nmol/L)	1	60 (55–64)	59 (55–63)	59 (52–67)	0.93	0.98	0.76	0.7	No
	2	57 (54–61)	58 (55–61)	47 (40–55)	<b>0.012</b>	0.82	<b>5.4E-03</b>	<b>4.4E-03</b>	No
	5	89 (83–95)	84 (79–89)	77 (69–84)	<i>0.057</i>	0.46	<b>0.016</b>	<i>0.056</i>	No

BMI = body mass index; feature?=whether variable was used as a clustering feature or not; geom. mean = geometric mean; prop. = proportion. For categorical variables, associations were tested using Fisher exact test; for continuous variables, Kruskal-Wallis and Mann-Whitney-Wilcoxon. Bold text indicates statistical significance ( $p < 0.05$ ); italics indicate near-significance ( $p < 0.10$ ). \*Not used as clustering feature, as BMI is a derived variable. Height and weight at age three were used instead.

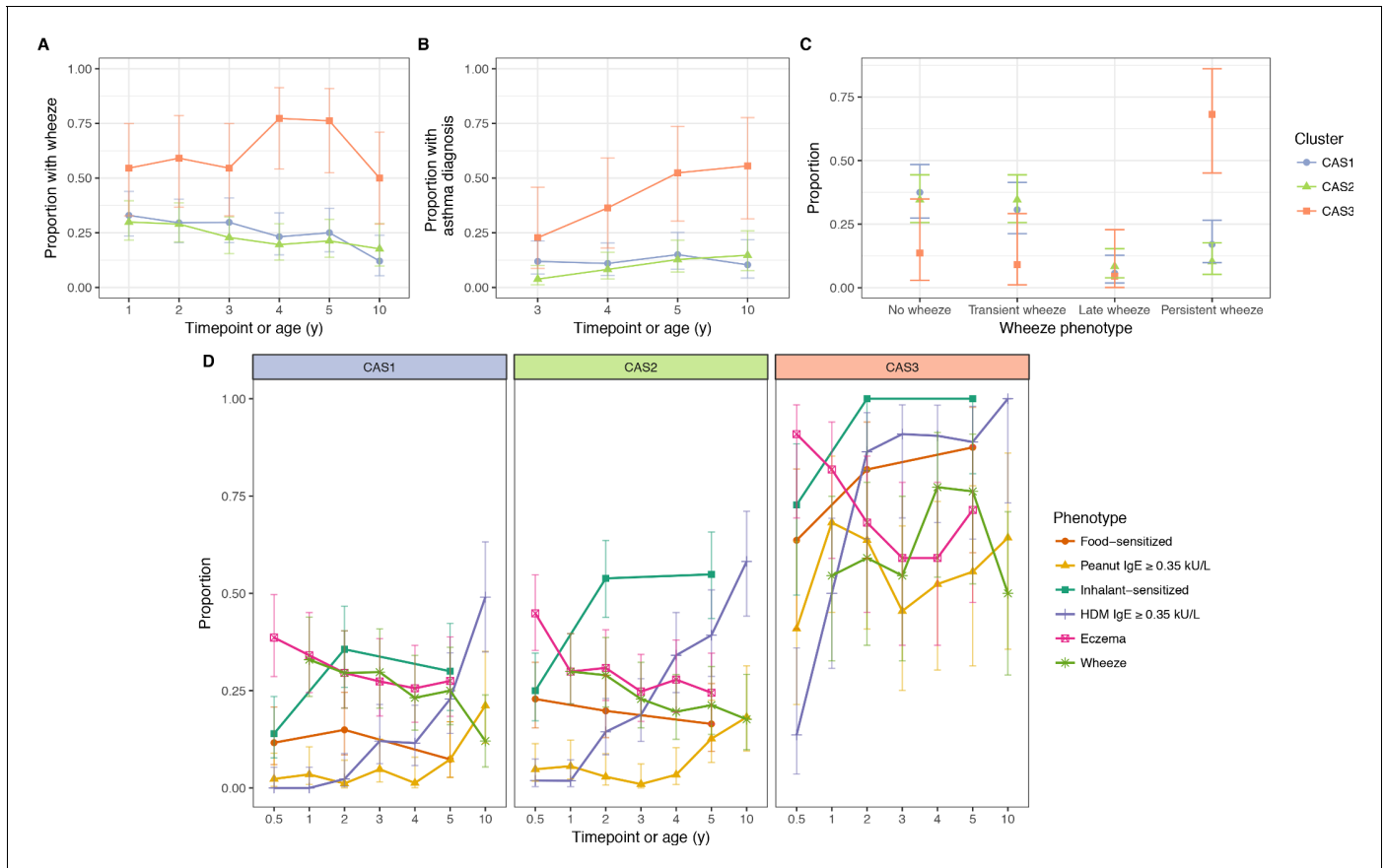
DOI: <https://doi.org/10.7554/eLife.35856.013>

table supplement 3C). There were no other significant differences for other non-Th2 cytokines (IFN- $\gamma$ , IL-10), nor were there specific differences for CAS1 or CAS2.

### Comparison of measures of immunological response

Across all clusters, allergen-specific IgG4 and IgG were positively correlated with IgE for the same allergen (especially HDM, **Figure 4—figure supplement 1**). As noted previously, CAS2 and CAS3 were distinguished from CAS1 by high IgG4, and CAS3 had greater IgG4 than either CAS1 or CAS2 (**Supplementary file 1 – table supplement 3B; Figure 4**). Decision tree analysis (**Figure 5—figure supplement 1 to 3**) confirmed that IgG4-type variables strongly separated CAS2 and CAS3 from CAS1, while IgE-type variables separated CAS3 from the others.

Although previous literature suggests possible protection conferred by IgG4 (**Okamoto et al., 2012**) or IgG (**Holt et al., 2016**), in this study there was no clear evidence of such protection against



**Figure 3.** Incidence of multiple phenotypes, including parent-reported wheeze. (A) Physician-diagnosed asthma (B) defined wheeze phenotypes (C) in relation to food and inhalant sensitisation (D) stratified by cluster and time in the CAS dataset. Points indicate observed proportion; bars indicate 95% CI (binomial distribution). Wheeze phenotypes defined as: no wheeze = no wheeze at ages 1 to 3, or age 5; transient wheeze = any wheeze at ages 1 to 3, but not age 5; late wheeze = wheeze at age 5, but not ages 1 to 3; persistent wheeze = any wheeze at both ages 1 to 3 and age 5. Food sensitisation defined as peanut IgE  $\geq 0.35$  kU/L at any age, or cow's milk, egg white, peanut SPT  $> 2$  or  $> 3$  mm for age  $\leq 2$  or  $> 2$  respectively. Inhalant sensitisation defined as HDM, cat, couchgrass, ryegrass, mould or Phadiatop IgE  $\geq 0.35$  kU/L at any age, or mould SPT (*Alternaria* or *Aspergillus* spp.)  $> 2$  or  $> 3$  mm for age  $\leq 2$  or  $> 2$ , respectively.

DOI: <https://doi.org/10.7554/eLife.35856.008>

The following figure supplement is available for figure 3:

**Figure supplement 1.** Relationship of clusters to food sensitisation, eczema and wheeze.

DOI: <https://doi.org/10.7554/eLife.35856.009>

later wheeze (**Table 4**). Furthermore, the protected status of CAS2 relative to CAS3 was unlikely to be driven by IgG4, given that CAS3 had greater quantities of both IgE and IgG4.

Although they were highly correlated, IgE, IgG, Th2 cytokine and SPT responses did not overlap perfectly. CAS3 was enriched for individuals with strong signals in all modalities, but there remained individuals within CAS3 and the rest of the cohort who were only responsive in some modalities and not others. Notably, the general direction of IgE, IgG4, SPT and Th2 cytokine signals did not always coincide (**Figure 4—figure supplement 2**).

### Comparison of clusters to existing criteria for atopy

The npEM-derived CAS clusters were partially consistent with traditional atopy thresholds (i.e. any specific IgE  $\geq 0.35$  kU/L or SPT  $\geq 2$  mm at age 2). When we compared CAS clusters with supervised groups created using traditional thresholds (**Supplementary file 1 – table supplement 5**), we found that CAS1 most closely matched a non-atopic phenotype (58 of 84 had no specific IgE greater than 0.35 kU/L by age 2). Conversely, CAS2 and CAS3 partially matched traditional criteria for atopy,



**Table 2.** Comparison of HDM-associated immunological variables in CAS clusters

Variable	Age	Cas1 (N = 88)	Cas2 (N = 107)	Cas3 (N = 22)	P-value (unadjusted)				Feature?	
		Geom. mean (95% CI)	Geom. mean (95% CI)	Geom. mean (95% CI)	Overall	Cas1 vs. 2	Cas1 vs. 3	Cas2 vs. 3		
<i>Total antibody</i>										
IgE (kU/L)	6m	1.2 (0.69–2)	2.2 (1.4–3.6)	21 (12–35)	1.2E-07	0.044	6.7E-08	2.2E-06	Yes	
	1	0.6 (0.29–1.3)	2 (1.1–3.7)	43 (17–109)	2.0E-09	0.019	4.3E-09	5.3E-08	Yes	
	2	6.6 (3.5–12)	17 (12–25)	187 (131–267)	1.2E-11	0.044	4.2E-11	1.4E-10	Yes	
	5	35 (23–55)	60 (46–80)	451 (278–731)	2.2E-08	0.096	1.9E-08	1.5E-07	No	
	10	85 (46–154)	150 (103–217)	800 (405–1.6E + 03)	1.4E-04	0.11	1.3E-04	2.8E-04	No	
<i>HDM antibody</i>										
IgE (kU/L)	6m	0.018 (0.016–0.02)	0.019 (0.016–0.022)	0.033 (0.019–0.059)	1.9E-03	0.47	7.9E-04	4.2E-03	Yes	
	1	0.019 (0.017–0.023)	0.019 (0.016–0.022)	0.26 (0.075–0.93)	1.3E-09	0.47	2.5E-07	4.5E-09	Yes	
	2	0.024 (0.019–0.031)	0.042 (0.029–0.06)	7.1 (2.7–19)	2.6E-16	0.078	2.5E-15	3.5E-13	Yes	
	5	0.072 (0.041–0.13)	0.23 (0.12–0.45)	31 (7.8–127)	4.2E-09	0.015	3.8E-09	5.1E-07	No	
	10	0.37 (0.17–0.8)	1.3 (0.51–3.4)	52 (19–144)	2.9E-06	0.068	5.7E-07	9.7E-05	No	
IgG (mg/L)	1	0.21 (0.2–0.23)	0.23 (0.21–0.25)	0.29 (0.21–0.39)	0.042	0.34	0.012	0.07	Yes	
	2	0.32 (0.27–0.37)	0.49 (0.41–0.59)	0.89 (0.57–1.4)	1.9E-06	2.1E-04	3.8E-06	7.0E-03	Yes	
	5	0.55 (0.42–0.7)	0.59 (0.46–0.74)	1.7 (0.88–3.3)	1.5E-03	0.67	6.4E-04	9.0E-04	No	
	10	1.6 (1.3–1.9)	2.1 (1.8–2.5)	2.8 (1.9–4.2)	1.0E-02	0.023	0.011	0.18	No	
IgG4 (µg/L)	6m	1.5E-04 (1.5E-04–1.5E-04)	1.7E-04 (1.3E-04–2.1E-04)	4.6E-04 (9.0E-05–2.4E-03)	4.9E-03	0.37	5.2E-03	0.024	Yes	
	1	1.5E-04 (1.5E-04–1.5E-04)	6.9E-04 (3.2E-04–1.5E-03)	0.081 (4.6E-03–1.4)	1.8E-10	5.2E-04	6.6E-12	2.2E-05	Yes	
	2	3.4E-04 (1.8E-04–6.6E-04)	4.8 (1.7–13)	61 (8.9–419)	1.8E-25	1.5E-22	8.6E-18	9.8E-05	Yes	
	5	2 (0.48–8.1)	168 (111–256)	539 (317–917)	1.1E-15	1.3E-12	1.0E-08	1.9E-04	No	
<i>HDM cytokine response</i>										
IL-13 protein (pg/ml)	0	0.22 (0.066–0.73)	0.22 (0.076–0.63)	0.085 (0.011–0.66)	0.68	0.76	0.41	0.45	No	
	6m	0.064 (0.022–0.18)	0.06 (0.025–0.14)	19 (1.4–244)	4.6E-06	0.98	1.7E-05	4.1E-06	No	
	5	0.13 (0.046–0.37)	0.32 (0.11–0.87)	12 (1.2–117)	2.1E-04	0.29	7.7E-05	5.1E-04	No	
IL-5 protein (pg/ml)	0	0.043 (0.018–0.11)	0.026 (0.013–0.052)	0.018 (5.0E-03–0.068)	0.44	0.36	0.29	0.57	No	
	6m	0.018 (9.2E-03–0.034)	0.013 (8.9E-03–0.02)	0.21 (0.012–3.7)	7.9E-04	0.4	8.1E-03	3.5E-04	No	

Table 2 continued on next page

Table 2 continued

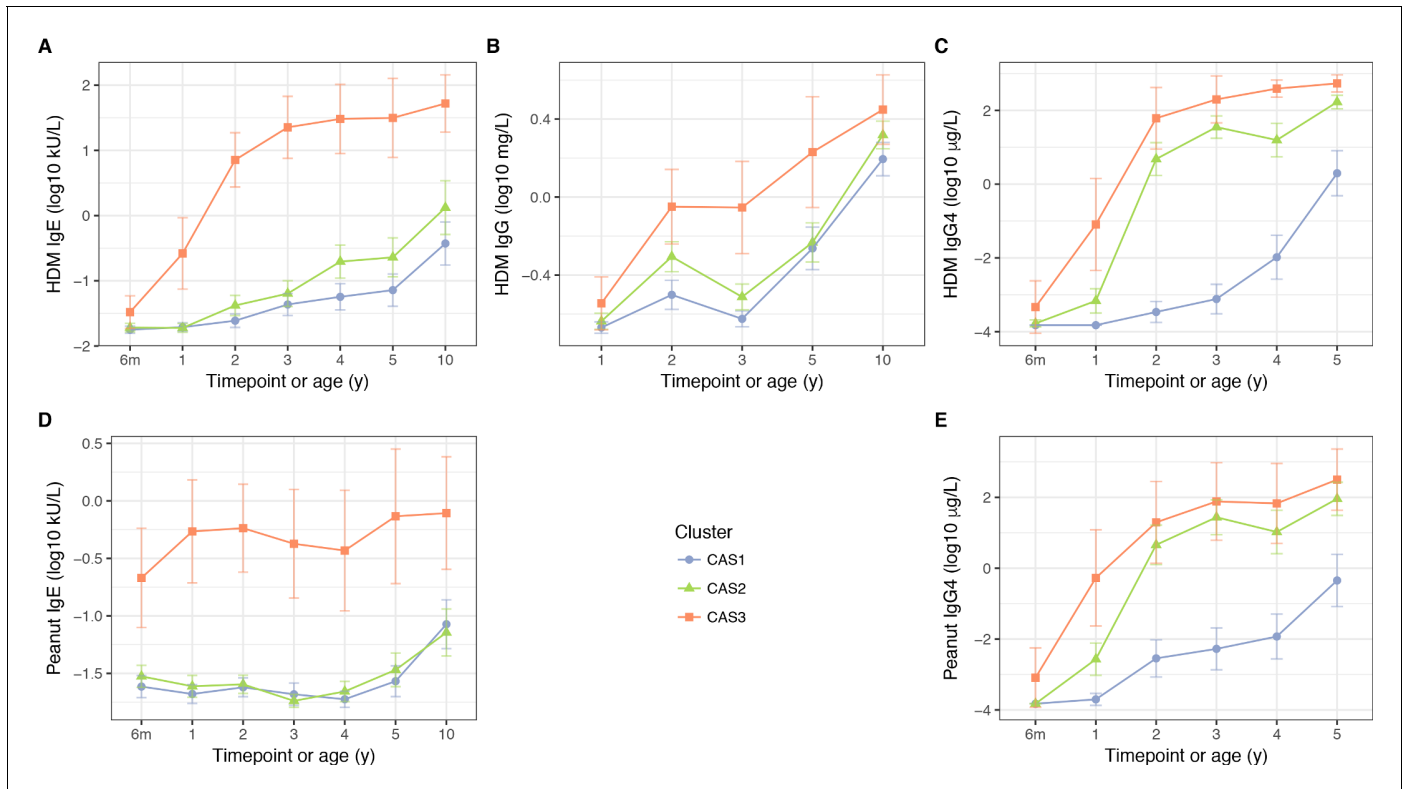
Variable	Age	Cas1 (N = 88)	Cas2 (N = 107)	Cas3 (N = 22)	P-value (unadjusted)				Feature?
		Geom. mean (95% CI)	Geom. mean (95% CI)	Geom. mean (95% CI)	Overall	Cas1 vs. 2	Cas1 vs. 3	Cas2 vs. 3	
IL-13 mRNA <sup>†</sup>	5	0.028 (0.014–0.057)	0.042 (0.02–0.087)	2.3 (0.25–22)	<b>3.2E-06</b>	0.45	<b>5.7E-06</b>	<b>2.0E-05</b>	No
	0	1.7E-03 (1.1E-04–0.026)	6.0E-03 (4.8E-04–0.075)	6.7E-03 (3.3E-05–1.4)	0.85	0.6	0.68	0.94	No
	6m	1.0E-04 (8.8E-06–1.1E-03)	3.2E-04 (3.8E-05–2.6E-03)	2 (0.015–266)	<b>3.2E-04</b>	0.5	<b>1.7E-04</b>	<b>3.8E-04</b>	No
IL-4 mRNA <sup>†</sup>	5	0.036 (1.6E-03–0.8)	0.11 (8.8E-03–1.4)	2.9E + 03 (742–1.1E + 04)	<b>6.8E-05</b>	0.59	<b>9.9E-05</b>	<b>2.5E-05</b>	No
	0	1.4E-06 (6.9E-07–3.0E-06)	1.9E-06 (7.8E-07–4.4E-06)	1.0E-06 (1.0E-06–1.0E-06)	0.71	0.65	0.6	0.47	No
	6m	4.6E-06 (1.0E-06–2.1E-05)	5.1E-06 (1.4E-06–1.8E-05)	0.54 (6.5E-03–44)	<b>6.2E-09</b>	0.94	<b>4.7E-07</b>	<b>1.0E-07</b>	No
IL-5 mRNA <sup>†</sup>	5	2.3E-04 (1.7E-05–3.0E-03)	4.7E-04 (5.3E-05–4.3E-03)	5.3 (0.082–345)	<b>4.9E-04</b>	0.72	<b>4.5E-04</b>	<b>3.2E-04</b>	No
	0	2.5E-04 (2.1E-05–2.9E-03)	2.6E-04 (2.8E-05–2.5E-03)	1.2E-05 (3.1E-07–4.6E-04)	0.47	0.96	0.24	0.25	No
	6m	5.2E-05 (5.6E-06–4.8E-04)	3.1E-05 (5.2E-06–1.8E-04)	0.33 (1.3E-03–83)	<b>1.5E-04</b>	0.85	<b>2.3E-04</b>	<b>1.1E-04</b>	No
HDM SPT past atopy threshold	5	0.021 (9.9E-04–0.43)	0.07 (5.7E-03–0.85)	246 (7–8.7E + 03)	<b>1.3E-04</b>	0.49	<b>7.1E-05</b>	<b>1.1E-04</b>	No
		Prop. (95% CI)	Prop. (95% CI)	Prop. (95% CI)	Overall	Cas1 vs. 2	Cas1 vs. 3	Cas2 vs. 3	
Wheal ≥ 2 mm	6m	2.3% (0–5.4%)	1.9% (0–4.5%)	14% (0–29%)	<b>0.043</b>	1	<i>0.054</i>	<b>0.035</b>	No*
	2	10% (3.8–17%)	15% (8.1–22%)	86% (71–100%)	<b>2.9E-12</b>	0.39	<b>8.2E-12</b>	<b>1.5E-10</b>	No*
Wheal ≥ 3 mm	5	13% (5.2–20%)	28% (18–37%)	81% (63–99%)	<b>1.5E-08</b>	<b>0.022</b>	<b>4.6E-09</b>	<b>1.0E-05</b>	No
	10	36% (23–49%)	51% (38–63%)	78% (57–99%)	<b>7.4E-03</b>	0.11	<b>2.7E-03</b>	<i>0.06</i>	No

Feature?=whether variable was used as a clustering feature or not; geom. mean = geometric mean; PBMC = peripheral blood mononuclear cells; prop. = proportion; SPT = skin prick or sensitisation test. For categorical variables, associations were tested using Fisher exact test; for continuous variables, Kruskal-Wallis and Mann-Whitney-Wilcoxon. Bold text indicates statistical significance ( $p < 0.05$ ); italics indicate near-significance ( $p < 0.10$ ). PBMC cytokine responses to HDM above unstimulated control; birth samples (age 0) taken from cord blood (CBMC). \*Not used as clustering features, as these were derived variables; the variables from which they were derived (HDM IgE and IgG4) were used instead.

DOI: <https://doi.org/10.7554/eLife.35856.014>

with CAS3 being an extreme phenotype (all 22 children in CAS3 had some specific IgE  $\geq 0.35$  kU/L by age 2).

However, the CAS clusters outperformed IgE/SPT-defined atopy in terms of predicting for age-5 wheeze (likelihood ratio test for clusters vs. IgE/SPT, Chi-squared = 23,  $p = 2.0 \times 10^{-6}$ ). In addition, at age 2, 68% of CAS3 were 'sensitised' (any specific IgE  $\geq 0.35$  kU/L) to two or more allergens, compared to only 1% and 6% for CAS1 and CAS2 respectively. This emphasised CAS3 as an early- and multi-sensitised phenotype. Finally, fewer members of CAS1 and CAS2 who were IgE- or SPT-responsive prior to age 5 maintained atopic wheeze at age 5 (23% or 24 of 103), compared to CAS3 (76% or 16 of 21). Therefore, the association of IgE and SPT with disease risk varied across clusters. This suggests that fixed atopy thresholds are not sufficient to delineate risk profiles – instead, an unsupervised clustering approach may be more informative.



**Figure 4.** HDM IgE (A), IgG (B) and IgG4 (C); and peanut IgE (D) and IgG4 (E) stratified by cluster and time, in the CAS dataset. Points indicate means; bars indicate 95% CI (t-distribution).

DOI: <https://doi.org/10.7554/eLife.35856.010>

The following figure supplements are available for figure 4:

**Figure supplement 1.** Correlation patterns between IgE vs IgG4 (A) and IgE vs IgG (B) at age five \* $p < 0.05$  for Spearman correlation with Holm correction for multiple testing.

DOI: <https://doi.org/10.7554/eLife.35856.011>

**Figure supplement 2.** Distinct biological signals of HDM IgE, IgG4, SPT, and Th2 cytokine (IL-13).

DOI: <https://doi.org/10.7554/eLife.35856.012>

## Comparison of clusters to time-dependent wheeze phenotypes and atopic disease

We mapped the npEM-derived clusters to pre-defined wheezing phenotypes (**Figure 3C**): no wheeze (in the first 3 years of life, or at age 5), transient wheeze (only in first 3 years), late wheeze (only at age 5), and persistent wheeze (both first 3 years and age 5). We found that CAS3 was enriched for persistent wheeze, while individuals in CAS1 or CAS2 tended to have transient or no wheeze. There were rarely any members of CAS with late wheeze (approximately 10%).

In addition to persistent wheeze, CAS3 was also enriched for persistent food sensitisation (peanut IgE  $\geq 0.35$  kU/L, or positive egg white or cow's milk SPTs) and persistent eczema: 44% of CAS3 experienced all three (**Figure 3—figure supplement 1**). Almost all individuals in CAS3 had both eczema and food sensitisation from age 6 m onwards, with rates of food sensitisation and wheeze increasing with time (**Figure 3D**). In contrast, CAS1 and CAS2 had low rates of food sensitisation, and declining rates of both eczema and wheeze. These trends lend credence to recent suggestions that the 'atopic march' phenotype (**Bantz et al., 2014; Han et al., 2017**) may only be present in a minority of the population (e.g. CAS3) (**Belgrave et al., 2014**).

**Table 3.** Comparison of selected respiratory-disease-related variables in CAS clusters

Variable	Age (y)	Cas1 (N = 88)	Cas2 (N = 107)	Cas3 (N = 22)	P-value (unadjusted)				Feature?
		Mean (95% CI)	Mean (95% CI)	Mean (95% CI)	Overall	Cas1 vs. 2	Cas1 vs. 3	Cas2 vs. 3	
URI (events per y)	1	2.9 (2.4–3.3)	2.6 (2.2–3)	2.5 (1.7–3.3)	0.59	0.34	0.5	0.96	Yes
	2	3.2 (2.6–3.7)	2.6 (2.2–3)	2.5 (1.2–3.8)	0.19	0.19	0.12	0.34	Yes
	3	2.7 (2.2–3.2)	2.8 (2.4–3.3)	2.2 (1.3–3.2)	0.45	0.41	0.59	0.24	Yes
	4	2.1 (1.7–2.6)	2.2 (1.8–2.7)	1.7 (0.77–2.7)	0.5	0.94	0.26	0.27	No
	5	1.6 (1.1–2)	1.5 (1.2–1.9)	0.67 (0.2–1.1)	<i>0.081</i>	0.76	<b>0.047</b>	<b>0.026</b>	No
LRI (events per y)	1	1.6 (1.2–1.9)	0.98 (0.76–1.2)	2 (1.3–2.6)	<b>4.0E-03</b>	<b>0.021</b>	0.17	<b>2.6E-03</b>	Yes
	2	1.4 (0.98–1.7)	1 (0.81–1.2)	2.2 (1.6–2.9)	<b>2.5E-03</b>	0.83	<b>6.1E-03</b>	<b>2.0E-04</b>	Yes
	3	1 (0.76–1.3)	0.6 (0.4–0.8)	1.8 (1.1–2.6)	<b>6.1E-04</b>	<b>0.02</b>	<b>0.039</b>	<b>2.7E-04</b>	Yes
	4	0.87 (0.52–1.2)	0.46 (0.3–0.63)	2 (1.1–2.8)	<b>1.7E-05</b>	0.3	<b>3.5E-04</b>	<b>1.6E-06</b>	No
	5	0.42 (0.24–0.6)	0.36 (0.24–0.48)	0.86 (0.44–1.3)	<b>0.019</b>	1	0.011	<b>7.5E-03</b>	No
Wheezy LRI (wLRI, events per y)	1	0.47 (0.3–0.63)	0.24 (0.15–0.34)	0.64 (0.19–1.1)	0.054	<b>0.036</b>	0.61	<i>0.065</i>	Yes
	2	0.68 (0.45–0.91)	0.41 (0.26–0.56)	1 (0.56–1.5)	<b>5.2E-03</b>	<i>0.063</i>	<i>0.066</i>	<b>1.7E-03</b>	Yes
	3	0.59 (0.37–0.81)	0.3 (0.17–0.44)	1.4 (0.78–2.1)	<b>4.6E-05</b>	<i>0.065</i>	<b>2.5E-03</b>	<b>6.6E-06</b>	Yes
	4	0.52 (0.25–0.79)	0.32 (0.18–0.46)	1.9 (0.95–2.8)	<b>4.5E-08</b>	0.86	<b>9.3E-07</b>	<b>3.3E-08</b>	No
	5	0.28 (0.13–0.42)	0.23 (0.13–0.33)	0.76 (0.36–1.2)	<b>2.3E-03</b>	0.99	<b>2.0E-03</b>	<b>1.2E-03</b>	No
Febrile LRI (fLRI, events per y)	1	0.36 (0.22–0.51)	0.28 (0.16–0.4)	0.55 (0.28–0.81)	<b>0.025</b>	0.24	<i>0.071</i>	<b>6.4E-03</b>	Yes
	2	0.36 (0.23–0.5)	0.33 (0.22–0.43)	0.95 (0.46–1.4)	<b>0.01</b>	1	<b>6.1E-03</b>	<b>3.8E-03</b>	Yes
	3	0.38 (0.21–0.55)	0.16 (0.09–0.23)	0.52 (0.13–0.92)	0.06	0.063	0.44	<b>0.04</b>	Yes
	4	0.3 (0.13–0.47)	0.15 (0.064–0.24)	0.43 (0.16–0.7)	<b>0.021</b>	0.18	<i>0.091</i>	<b>4.9E-03</b>	No
	5	0.19 (0.082–0.3)	0.14 (0.06–0.21)	0.19 (0–0.42)	0.83	0.55	0.91	0.8	No
>20% Streptococcus in first infection-naive NPA sample	7w	11% (0.34–23%)	15% (3.3–26%)	44% (3.9–85%)	0.081	0.75	<b>0.042</b>	<i>0.065</i>	No
	6m	7.6% (1.6–14%)	18% (10–26%)	14% (0–31%)	0.12	<b>0.045</b>	0.39	1	No
% Healthy NPAs with infection-associated MPGs	0–2	49% (38–59%)	32% (24–39%)	62% (47–76%)	<b>1.2E-03</b>	<b>0.013</b>	0.2	<b>5.5E-04</b>	No
	2–4	46% (37–55%)	44% (37–51%)	45% (29–61%)	0.9	0.67	0.92	0.8	No

Feature?=whether variable was used as a clustering feature or not; geom. mean = geometric mean; ARI = acute respiratory infection (lower or upper); LRI = lower respiratory infection; MPG = microbiome profile group; NPA = nasopharyngeal aspirate; prop. = proportion; URI = upper respiratory infection; 7w = 7 weeks. For categorical variables, associations were tested using Fisher exact test; for continuous variables, Kruskal-Wallis and Mann-Whitney-Wilcoxon. Bold text indicates statistical significance ( $p < 0.05$ ); italics indicate near-significance ( $p < 0.10$ ). \*Not used as clustering features, as these were derived variables; the variables from which they were derived (URI, LRI, wLRI, fLRI) were used instead.

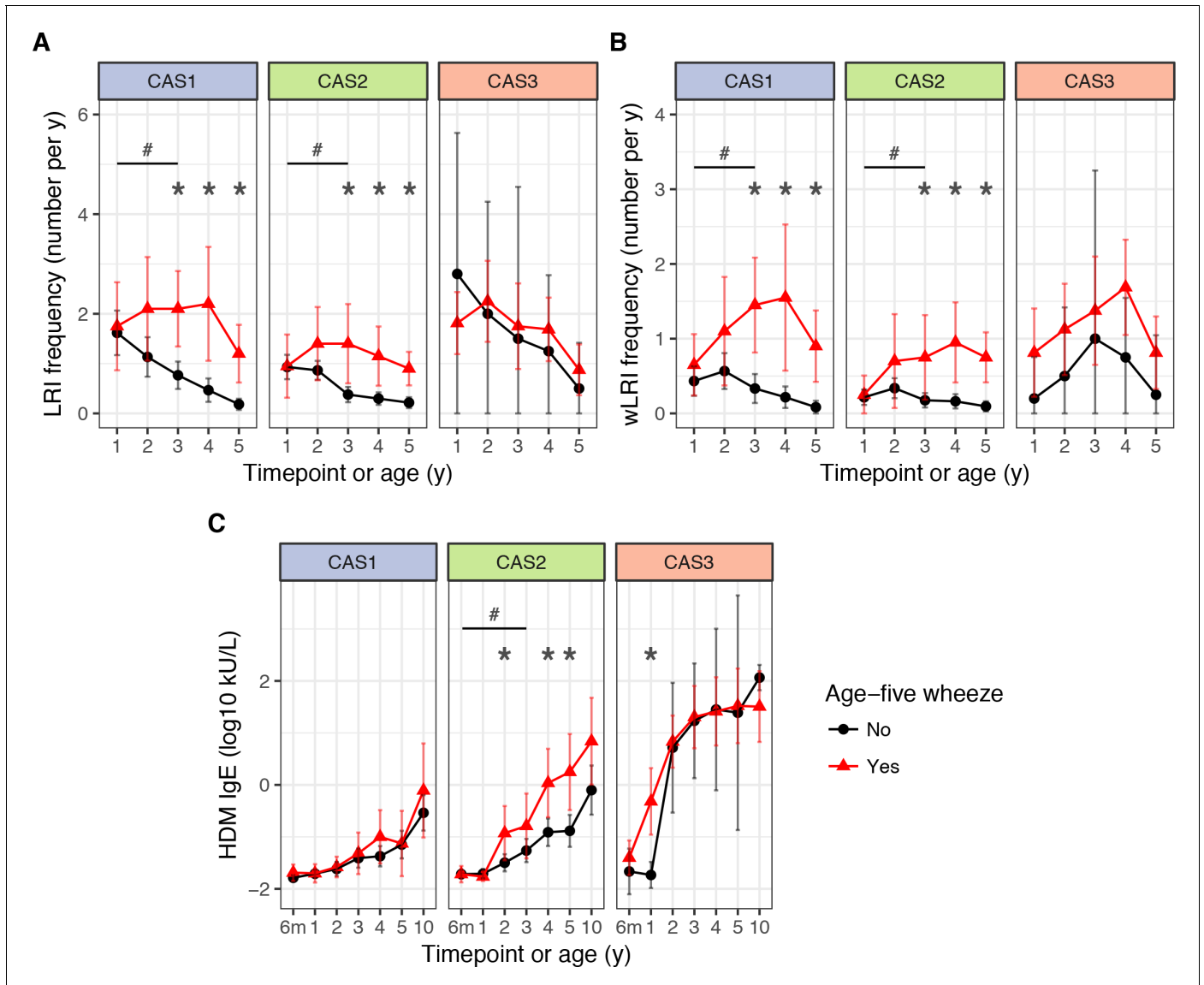
DOI: <https://doi.org/10.7554/eLife.35856.020>

**Table 4.** Analysis of selected predictors for age-5 wheeze within each CAS cluster, with demographic covariates (sex, BMI, parental history of asthma)

Selected predictors for age-5 wheeze		Cas1 (N = 88) Or (95% CI)	P-value	Cas2 (N = 107) Or (95% CI)	P-value	Cas3 (N = 22) Or (95% CI)	P-value	Or (95% CI)	P-value
LRI (events per y)	1	0.97 (0.71–1.3)	0.84	1 (0.61–1.5)	0.99	0.48 (0.13–1.1)	0.16	1 (0.81–1.2)	0.92
	2	1.2 (0.88–1.6)	0.26	1.5 (0.97–2.5)	<i>0.069</i>	0.99 (0.34–2.6)	0.98	1.4 (1.1–1.7)	<b>5.3E-03</b>
	3	2 (1.3–3.2)	<b>2.3E-03</b>	2.6 (1.5–5.3)	<b>2.7E-03</b>	0.98 (0.4–2.6)	0.96	2 (1.5–2.7)	<b>3.8E-06</b>
	4	2 (1.4–3.4)	<b>2.0E-03</b>	3.6 (1.8–8.3)	<b>6.5E-04</b>	1.9 (0.57–8.4)	0.32	2.5 (1.8–3.6)	<b>1.5E-07</b>
Wheezy LRI (events per y)	1	1.3 (0.68–2.4)	0.43	1.1 (0.35–3)	0.83	2.6 (0.62–58)	0.34	1.5 (0.98–2.3)	0.06
	2	1.2 (0.8–2)	0.33	1.6 (0.89–2.9)	0.12	2.4 (0.67–16)	0.24	1.6 (1.2–2.2)	<b>5.6E-03</b>
	3	2.8 (1.6–5.6)	<b>1.3E-03</b>	3 (1.4–8)	<b>0.016</b>	1.2 (0.43–4.6)	0.76	2.7 (1.8–4.2)	<b>4.1E-06</b>
	4	2.5 (1.5–5)	<b>4.0E-03</b>	6.3 (2.5–21)	<b>6.8E-04</b>	7.1 (1.2–169)	0.1	3.9 (2.5–6.7)	<b>5.4E-08</b>
Febrile LRI (events per y)	1	1.6 (0.77–3.6)	0.21	0.84 (0.28–1.9)	0.71	7.3 (0.78–178)	0.12	1.5 (0.93–2.4)	0.098
	2	1 (0.44–2.2)	1	4.8 (1.8–15)	<b>3.9E-03</b>	1.6 (0.48–10)	0.5	2.3 (1.4–3.9)	<b>1.2E-03</b>
	3	2 (1–4.8)	0.08	4.3 (1.2–15)	<b>0.02</b>	4.2 (0.55–519)	0.37	2.4 (1.4–4.3)	<b>2.3E-03</b>
	4	1.8 (0.97–4.1)	<i>0.092</i>	2.6 (0.88–8.3)	<i>0.082</i>	1.1 (0.11–18)	0.93	2.2 (1.3–4)	<b>5.9E-03</b>
Quartile of % healthy NPAs with infection-associated MPGs	0–2	1 (0.54–1.8)	0.98	1.3 (0.72–2.4)	0.36	NA	NA	1.3 (0.89–1.8)	0.19
	2–4	0.45 (0.19–0.88)	<b>0.035</b>	1 (0.51–2.1)	0.9	NA	NA	0.8 (0.53–1.2)	0.24
HDM IgE (kU/L)*	6m	8 (0.85–94)	0.074	0.93 (0.14–3.6)	0.92	3.4 (0.26–180)	0.4	2.3 (0.99–5.8)	0.054
	1	1.5 (0.22–7.8)	0.65	0.54 (0.039–2.3)	0.51	39 (2.5–22000)	0.082	2.7 (1.5–5)	<b>0.00089</b>
	2	0.93 (0.28–2.5)	0.89	2 (1.2–3.7)	<b>0.016</b>	1.4 (0.38–4.8)	0.62	2 (1.5–2.8)	<b>2.80E-05</b>
	3	1.4 (0.68–2.9)	0.32	1.5 (0.9–2.4)	0.12	1.5 (0.4–5.2)	0.55	1.7 (1.3–2.2)	<b>1.00E-04</b>
	4	1.9 (0.94–4.1)	<i>0.086</i>	1.9 (1.2–3.1)	<b>0.011</b>	1.4 (0.31–5.5)	0.64	1.9 (1.5–2.5)	<b>3.70E-06</b>
HDM IgG4 (µg/L)*	6m	NA (NA-NA)	0.55	0.053 (NA–6.5e + 24)	0.99	28 (1.7e-34-NA)	0.99	1.4 (0.88–2.6)	0.17
	1	NA (NA-NA)	0.61	1.1 (0.8–1.5)	0.5	0.9 (0.58–1.3)	0.6	1.2 (1–1.4)	0.053
	2	1.1 (0.71–1.6)	0.67	1.1 (0.85–1.4)	0.61	0.4 (0.038–1.2)	0.26	1.1 (1–1.3)	0.056
	3	1.1 (0.85–1.5)	0.35	1.1 (0.77–2)	0.64	0.94 (0.19–2.3)	0.9	1.1 (0.98–1.2)	0.1
	4	1.2 (0.98–1.5)	<i>0.082</i>	0.89 (0.7–1.1)	0.33	0.46 (0.031–5.4)	0.53	1.1 (1–1.3)	<b>0.034</b>
HDM IgG (mg/L)*	1	25 (0.32–1.6E + 04)	0.19	3.3 (0.16–46)	0.38	5.6E-03 (8.4E-06–0.57)	0.058	2 (0.31–11)	0.44
	2	0.8 (0.15–3.5)	0.78	0.97 (0.24–3.7)	0.96	0.79 (0.031–18)	0.88	1.3 (0.6–2.9)	0.48
	3	2.3 (0.14–35)	0.54	0.48 (0.057–2.5)	0.43	3.9 (0.26–96)	0.34	2.1 (0.89–5)	<i>0.089</i>

BMI = body mass index; HDM = house dust mite; LRI = lower respiratory infection. Association analyses performed via generalised linear models (GLM) with demographic covariates: age-5 wheeze ~predictor + sex (male) +BMI at age 3 + paternal history of asthma +maternal history of asthma. Bold text indicates statistical significance ( $p < 0.05$ ); italics indicate near-significance ( $p < 0.10$ ). \*Odds ratio (OR) is for every 10-fold increase in IgE, IgG4 or IgG.

DOI: <https://doi.org/10.7554/eLife.35856.021>



**Figure 5.** LRI frequency (A), wheezy LRI (wLRI) frequency (B), and HDM IgE (C), stratified by age-5 wheeze status, cluster and time, in the CAS dataset. Points indicate means; bars indicate 95% CI (t-distribution). # $p < 0.05$  for repeated-measures ANOVA across timepoints from the first 3 years of life (see **Table 4**). \* $p < 0.05$  for Mann-Whitney-Wilcoxon comparison within each timepoint.

DOI: <https://doi.org/10.7554/eLife.35856.015>

The following figure supplements are available for figure 5:

**Figure supplement 1.** A 'simple' decision tree generated by recursive partitioning from CAS data, with breakdown of tree clusters by actual CAS npEM-derived clusters.

DOI: <https://doi.org/10.7554/eLife.35856.016>

**Figure supplement 2.** Decision tree generated by recursive partitioning from CAS data, excluding Phadiatop assay variables.

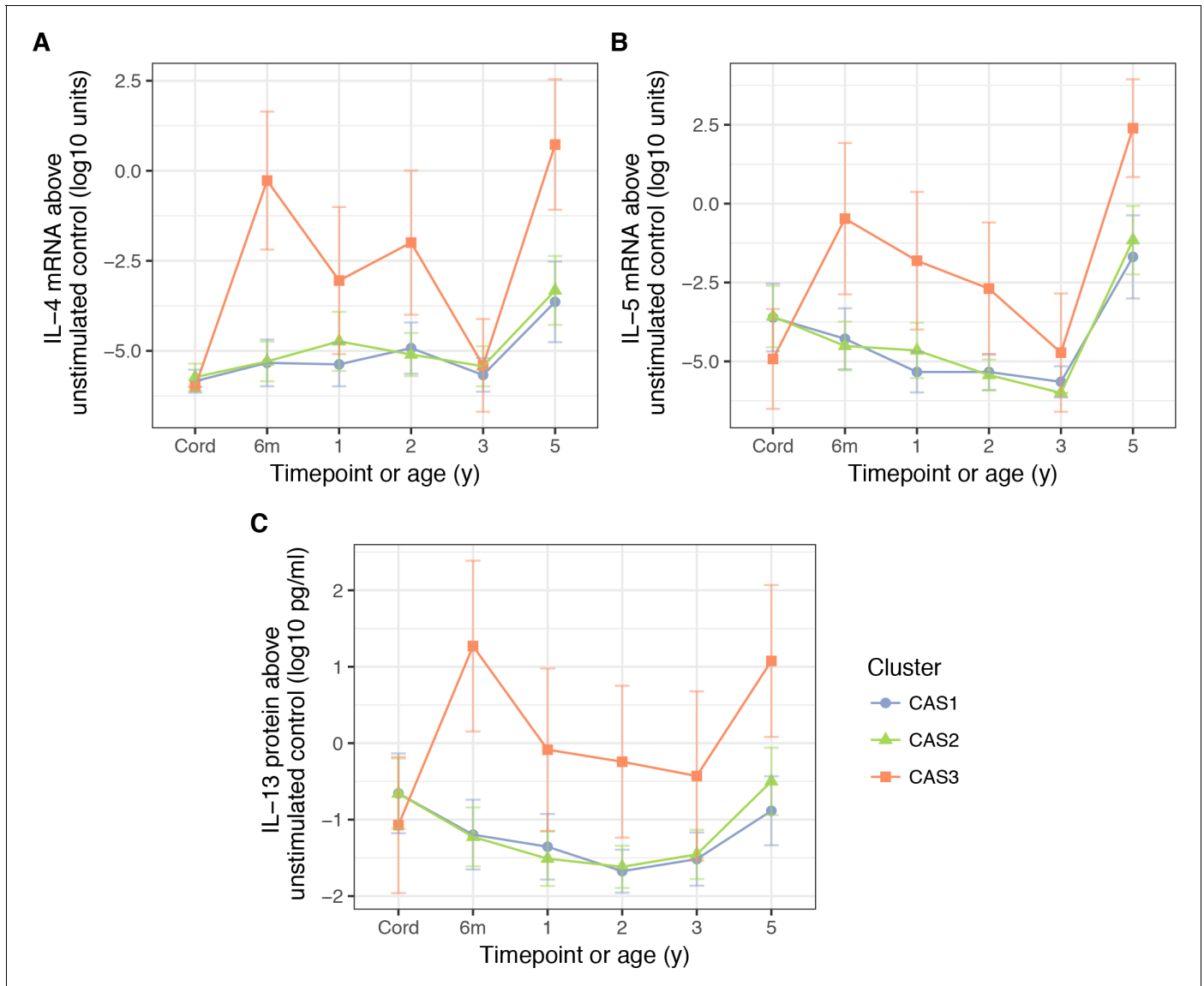
DOI: <https://doi.org/10.7554/eLife.35856.017>

**Figure supplement 3.** A 'comprehensive' decision tree generated by recursive partitioning from CAS data, given CAS npEM-derived clusters and age-5 wheezing status.

DOI: <https://doi.org/10.7554/eLife.35856.018>

**Figure supplement 4.** Comparison of predictors for age-5 wheeze in CAS and COAST clusters.

DOI: <https://doi.org/10.7554/eLife.35856.019>



**Figure 6.** PBMC expression of IL-5. (A) and IL-4 mRNA (B), as well as IL-13 protein (C), in response to stimulation HDM, stratified by cluster and time (CAS) Cord = cord blood sample collected at birth. Points indicate means; bars indicate 95% CI (t-distribution).

DOI: <https://doi.org/10.7554/eLife.35856.022>

## Relationship with the nasopharyngeal microbiome

Previous studies suggest an association between asthma risk and early-life disruption of the respiratory microbiome, especially colonisation with *Streptococcus* spp. in the first 7 weeks of life (Teo et al., 2015). In this study, using the same data and definitions, we found that CAS3 was over-represented by individuals who had >20% relative abundance of *Streptococcus* in their first infection-naive healthy NPA, within the first 7 weeks of life (44% versus 11% and 15% in CAS1 and CAS2, respectively; Fisher exact test, unadjusted  $p=0.042$  and  $0.065$ , respectively; Table 3).

Furthermore, Teo et al and others (Teo et al., 2015; Bisgaard et al., 2007) previously found that transient incursions with certain MPGs (*Streptococcus*, *Haemophilus*, *Moraxella* spp.) were associated with increased frequency and severity of subsequent LRIs and wheezing disease. Here, we found that proportion of these infection-associated MPGs in healthy samples from age 0 to 2 was greater in CAS3 (62% vs. 49% and 32% in CAS1 and CAS2, respectively; Fisher exact test,

unadjusted  $p=0.2$  and  $5.5 \times 10^{-4}$ , respectively; **Table 3**). This finding was independent of LRI and wLRI frequency (GLM;  $p<0.05$  for model predicting group membership, with age-2 LRI and wLRI as covariates). On the contrary, there were no associations between cluster membership and health-associated MPGs (*Corynebacterium*, *Alloicoccus*, *Staphylococcus* spp.; **Supplementary file 1 – table supplement 3E**).

Recent work by **Teo et al., 2017** suggested that infection-associated MPGs in early life were predictive for age-5 wheeze in atopic children, while in non-atopic children they were predictive for transient wheeze. In this study, with the same cohort, we noted a similar trend for infection-associated MPGs from age 0 to 2, in relation to transient wheeze in ‘non-atopic’ CAS1 (GLM, OR 3.6 per percent,  $p=0.17$ , with demographic covariates). Surprisingly, there was evidence that infection-associated MPGs in later samples (from age 2 to 4) were protective against age-5 wheeze in CAS1 (OR 0.086 per percent, 0.45 per quartile,  $p=0.034$  and  $0.035$ , respectively; **Table 4**). Infection- and health-associated MPGs were otherwise not associated with age-5 wheeze within the other clusters.

### External replication of clusters in MAAS and COAST

The trajectories described by the CAS npEM clusters were replicated in two cohorts – the Manchester Asthma and Allergy Study (MAAS) ( $N = 1085$ ) (**NAC Manchester Asthma and Allergy Study Group et al., 2002**) from Manchester, UK, and the Childhood Origins of Asthma Study (COAST) ( $N = 289$ ) from Wisconsin, USA (**Lemanske, 2002**). After applying our npEM classifier to these external cohorts (materials and methods), we found that individuals classified into ‘Cluster 3’ (MAAS3/COAST3) had a persistent disease phenotype extending into late adolescence, with consistently high rates of parent-reported wheeze and physician-diagnosed asthma from birth to age 16. The other two clusters (Cluster 1 = MAAS1/COAST1; Cluster 2 = MAAS2/COAST2) appeared to be low-risk (**Figure 7A,B,D**).

MAAS3 and COAST3 exhibited stronger IgE expression (total, HDM, cat, dog) from ages 1 to 8 (**Figure 7C,E**), compared to other clusters in each dataset. Like CAS3, COAST3 demonstrated elevated PBMC expression of Th2 cytokine protein (IL-5 and IL-13) in response to HDM stimulation at age 3 (**Figure 7F**). This was not replicated in MAAS3, but previous work in MAAS had identified that a strong PBMC Th2 response (IL-5, IL-13) to HDM stimulation at age 8 was associated with increased risk of HDM sensitisation and asthma (**Wu et al., 2015**). Nonetheless, MAAS3 was overrepresented in ‘early-sensitised’ and ‘multiple sensitised’ phenotypes described by **Lazic et al. (2013)** from SPT and IgE data. Approximately 86% of individuals in MAAS3 belonged to either one of these two phenotypes, although only 13% of individuals in these two phenotypes were accounted for by MAAS3.

Furthermore, when we explored potential predictors of wheeze phenotypes and asthma diagnosis in later childhood, we found that the clusters in COAST were very similar to those in CAS. In COAST1, LRI and wLRI frequency at age 2 were predictive of asthma diagnosis at age 6 (GLMs with demographic covariates,  $p=0.02$  and  $0.02$ , respectively), while in COAST2, HDM IgE at age 3, and LRI, wLRI and fLRI frequencies at age were all predictive (GLMs,  $p<0.05$  for all) (**Figure 5—figure supplement 4**). Although the timing and magnitude of associations differed between cohorts, this reaffirmed wheeze in Cluster 1 as being primarily non-atopic in origin, while wheeze in Cluster 2 appeared to be driven by both non-atopic and atopic factors.

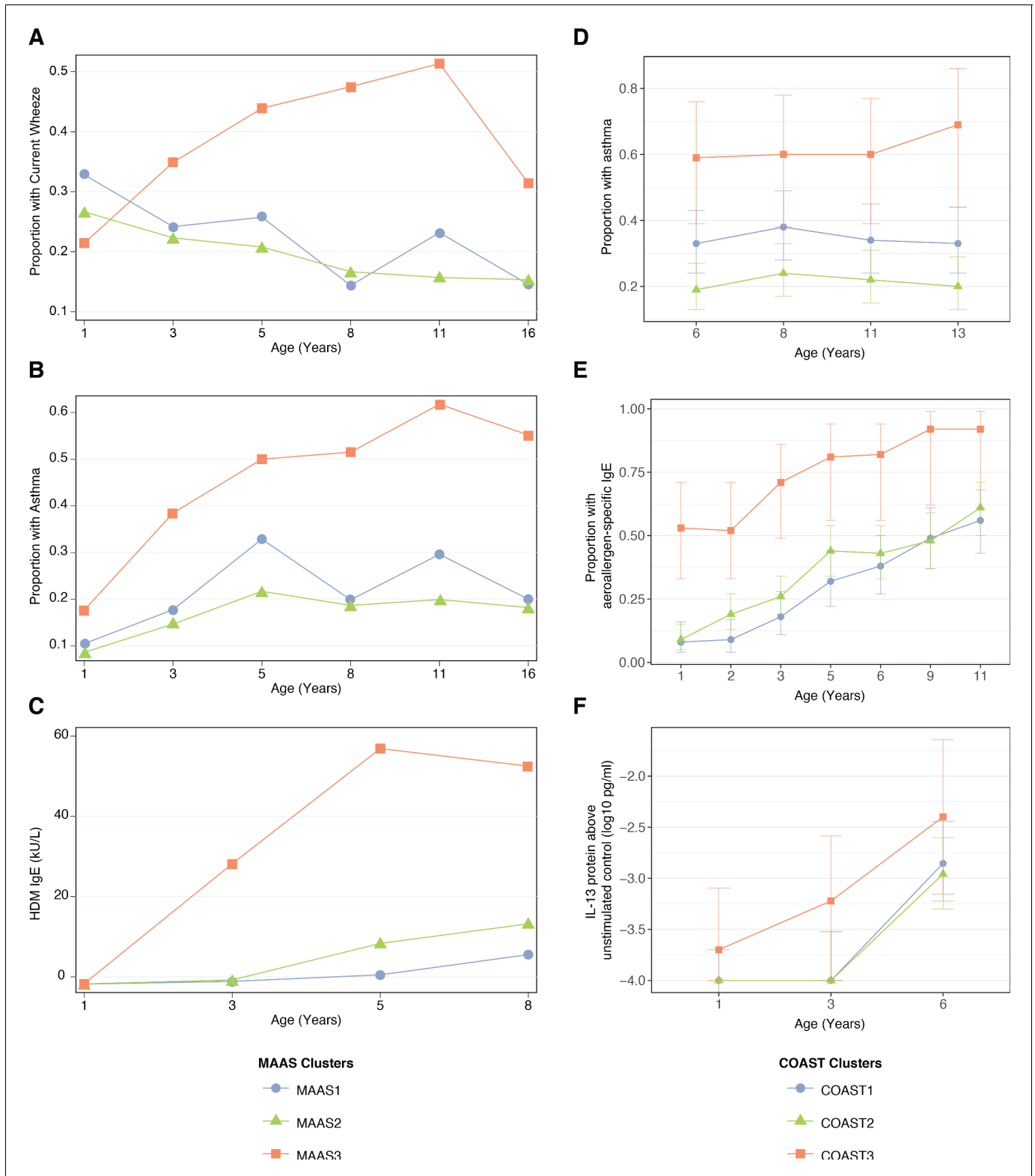
We re-applied npEM classification to CAS using only those features present in MAAS or COAST. For MAAS and COAST features, the subsequent clusters bore 79% and 72% concordance with the original CAS clusters, respectively. In both cases, concordance was excellent for Cluster 3 – all 22 members of the original CAS3 were correctly assigned to Cluster three after re-applying npEM. Therefore, CAS3, COAST3 and MAAS3 likely represent very similar phenotypes.

### Internal stability and validity of CAS clusters

We checked the stability and validity of the CAS clusters with leave-one-out (LOO) analysis, Jaccard indices and silhouette widths. The average Jaccard indices from leave-one-individual-out analysis were 0.77, 0.76, and 0.85 for CAS1, 2 and 3, respectively. For leave-one-feature-out analysis, the average indices were 0.65, 0.60, and 0.74, respectively. This demonstrates that the clusters, especially CAS3, were relatively resilient to minor changes in sampling or feature selection.

In relation to internal validity of the CAS clusters, average silhouette widths were universally poor, at 0.05, 0.06 and 0.002 for CAS1, 2, 3, respectively, with an average for all three clusters of 0.05





**Figure 7.** Description of npEM-derived clusters in external cohorts: in MAAS, incidence of wheeze (A), asthma diagnosis (B), and HDM IgE levels (C); in COAST, incidence of asthma diagnosis (D), proportion of individuals with detectable aeroallergen-specific IgE levels (E), and PBMC protein expression of IL-13 following HDM stimulation above unstimulated control (F) MAAS cohort (N = 934) was classified using npEM model from CAS, into MAAS1

Figure 7 continued on next page

Figure 7 continued

(N = 199, 21%), MAAS2 (N = 692, 74%) and MAAS3 (N = 43, 5%); these correspond to CAS clusters CAS1, 2 and 3, respectively. COAST cohort (N = 285) was similarly classified into COAST1 (N = 105, 37%), COAST2 (N = 151, 53%) and COAST3 (N = 29, 10%).

DOI: <https://doi.org/10.7554/eLife.35856.023>

(**Figure 1—figure supplement 2**). Silhouette widths were particularly suboptimal with CAS3, with at least half of those classified having negative values. The overall poor internal validity of the clusters may be due to the large-scale and exploratory nature of our approach – the metric may have been obscured by intra-cluster heterogeneity in other variables that were not particularly important for determining cluster membership. However, it must be noted that all clusters on average yielded positive silhouette widths, and as observed in the rest of the results, they were all relatively homogeneous in terms of the outcomes of interest (wheeze status, allergic disease phenotypes).

### Decision tree analysis

Decision tree analysis on the CAS dataset, using all available predictors from all timepoints, created a ‘Simple Tree’ with two decision nodes and three end nodes (**Figure 5—figure supplement 1**). This tree had 89% accuracy in retrieving cluster memberships from the original npEM model, where accuracy is calculated as percentage overlap of tree clusters with original CAS clusters. We found that membership in the CAS3-equivalent tree cluster was a better predictor for age-5 wheeze (likelihood ratio test, Chi-squared = 19,  $p < 1 \times 10^{-5}$ ) than traditional thresholds for atopy based on IgE and SPT measurements at age 2. IgG4-related variables best separated CAS1 from other clusters, while IgE-related variables best separated CAS2 and CAS3. Explicitly forcing the exclusion of Phadiatop variables from tree analysis caused these thresholds to be replaced with allergen-specific assays (HDM IgE for Phadiatop IgE, **Figure 5—figure supplement 2**) in a way that is consistent with correlation patterns amongst IgE and IgG4 variables (**Supplementary file 1 – table supplement 6**).

We also constructed a ‘Comprehensive Tree’ that best split individuals into six groups, based on cluster membership crossed with age-5 wheeze status (**Figure 5—figure supplement 3**). We thus identified nodes that were consistent with predictors for wheeze found in the previous regression analyses (**Table 4**), combined with nodes from the Simple Tree (**Figure 5—figure supplement 1**). The Comprehensive Tree had 77% accuracy in recovering both cluster membership and wheeze status. In terms of identifying pure wheeze status at age 5, the accuracy of the tree was 84%, with a positive predictive value (PPV, or precision) of 72%, negative predictive value (NPV) of 88%, sensitivity (recall) of 71% and specificity of 89%. The Comprehensive Tree was more successful in flagging age-5 wheeze (likelihood ratio test, Chi-squared = 60,  $p = 6.1 \times 10^{-13}$ ), compared to the traditional atopy thresholds described previously.

### Discussion

We have used model-based cluster analysis to uncover clusters of children with differential asthma susceptibility. Specifically, there was a high-risk group (Cluster 3) characterised by very early allergen-specific Th2 activity; early sensitisation to multiple allergens including food allergens; and concurrent frequent respiratory infections – resulting in high incidence of atopic persistent wheeze. We also found a lower risk cluster (Cluster 2), with limited or delayed elevation in IgE – this resulted in a lower incidence of mixed (atopic and non-atopic) wheeze. Finally, there was a low-risk cluster (Cluster 1) which exhibited occasional and transient infection-related wheeze, with minimal allergen sensitisation. These clusters were replicated in external datasets, suggesting relevance across populations. Summaries of key findings are given in **Table 5** and **Figure 2**.

### Cluster three is a high-risk, multi-sensitised, atopic phenotype

Cluster 3 represented a multi-sensitive or polysensitised phenotype (**Bousquet et al., 2015**). In CAS3, not only was total IgE elevated, but specific IgE were also raised for most allergens. Three in four CAS3 individuals were sensitised (specific IgE  $\geq 0.35$  kU/L) to two or more allergens. In our external replication with MAAS, we observed a large overlap between our predicted high-risk phenotype (MAAS3) and the multiple atopy phenotype from **Lazic et al., 2013**). This was consistent

**Table 5.** Key findings from cluster analysis

Certain childhood populations may be broadly split into three clusters, each representing a unique trajectory of immune function and susceptibility to respiratory infections: low-risk non-atopic Cluster 1 with transient wheeze; low-risk but allergy-susceptible Cluster 2 with mixed wheeze; and strongly-atopic high-risk Cluster 3 with persistent wheeze.

Cluster 3 is consistent with an early-sensitised and multi-sensitised phenotype.

HDM hypersensitivity is an important predictor of wheeze in allergic or allergy-susceptible individuals.

Food and peanut hypersensitivities are important contributors to membership in high-risk Cluster 3. This may be pathophysiologically related to eczema, multi-sensitisation and the atopic march.

In CAS, IgG4 flags for clusters with susceptibility to atopic disease (CAS2 and CAS3), while early and multiple-allergen elevation in IgE predicts frank atopic disease. The pathophysiological role of IgG4 remains unclear.

Allergic and infective processes act additively to intensify airway inflammation during respiratory pathogen clearance. Some (Cluster 3) may be more susceptible to this effect than others that lack strong allergic sensitisation (Cluster 1).

Tests for atopy (IgE, SPT, cytokines) do not overlap perfectly. Therefore, atopy may be better defined by the composite result from a battery of tests encapsulated in a predictive model, rather than just a single test or threshold.

The microbiome acts differently on asthma risk depending on cluster membership. In CAS, early-life asymptomatic colonisation with infection-associated MPGs is associated with risk of persistent wheeze in allergy-susceptible clusters (CAS2, CAS3), while it is potentially protective in non-atopic children (CAS1).

Different childhood populations may share similar trajectories of asthma susceptibility, but there may be subtle differences in terms of the types of tests, allergens, or biological signals that are most informative (SPT, IgE, cytokines, etc.).

DOI: <https://doi.org/10.7554/eLife.35856.024>

with findings from other studies, where the severely atopic and polysensitised subpopulation was at greater risk of both wheezing disease and reduced lung function (*Hose et al., 2017*).

It is not currently known what is fundamentally producing the strong atopic predisposition in Cluster 3. It is possible that inherited (genetic/epigenetic) or environmental factors (including in utero or perinatal exposures) may be involved, and these should be targets for future investigations. The overrepresentation of males in CAS3 is consistent with the consensus that young boys are at greater risk for asthma than young girls; this was traditionally believed to be due to intrinsic sex differences in airway diameter (*Almqvist et al., 2008*). However, our cluster analysis did not employ any clustering features related to airway size. This suggests that other sex-related factors could be involved, such as differences in immunity and allergic susceptibility. Allergic sensitisation is more frequent amongst prepubescent boys than girls (*Gabet et al., 2016; Kim et al., 2014*), and this may be linked to differences in cytokine responsiveness. However, not all boys were clustered into Cluster 3; and sex was not found to be a determinant for either IgE levels or cytokine response in CAS.

We did observe that CAS3 overlapped strongly with both persistent food sensitisation and eczema, and that persistent wheeze co-occurred with early sensitisation and eczema. This suggests that the 'atopic march' may play a role in CAS3. Early disruption of the skin barrier and exposure to certain food allergens may act in concert to promote and entrench the atopic phenotype, through the activation of cytokine pathways involving TSLP, IL33 and IL25 (*Bantz et al., 2014; Han et al., 2017*). Although recent research has suggested that very few children actually follow the disease trajectory of the atopic march (*Belgrave et al., 2014*), we hypothesise that it remains relevant to a small but important high-risk subpopulation, who may potentially benefit from early interventions targeted at halting the progression of disease.

### Role of early-life HDM hypersensitivity

In all three cohorts (CAS, MAAS, COAST), house dust mite (HDM) sensitivity was an important determinant of atopic disease risk. HDM was a strong predictor for both CAS3 membership and later childhood wheeze in CAS2, as well as being a 'dominant' allergen in the Phadiatop Infant assays. CAS3 in particular exhibited early and extreme HDM hypersensitivity, with prematurely-elevated HDM IgE, as well as PBMC Th2 response (IL-4, 5, 9, 13) to HDM stimulation. Similar phenomena were seen with MAAS3 and COAST3. The importance of HDM hypersensitivity in driving allergic disease in some populations is well-described in the literature (*Thomas et al., 2010; Calderón et al., 2015*). Previous findings from MAAS and a similar cohort RAINE (*Wu et al., 2015*) have shown a confluence of high HDM IgE, as well as PBMC Th2 cytokine levels such as IL-13 and IL-5, in discrete subsets of the population. However, we did observe that in other clusters (CAS1 and CAS2), some

individuals with purported HDM sensitisation (IgE >0.35 kU/L) did not produce detectable Th2 responses; the reverse was also true, where Th2 response did not necessarily result in high IgE. It may be the case that there is high intra-individual variation in IgE and cytokine responses, or stochastic variation in detectability of IgE or cytokine, which may obscure association analyses. Regardless, early and strong Th2 cytokine responses against HDM indicate a high-risk phenotype.

### Role of early-life food and peanut sensitisation

Interestingly, early-life peanut IgE was a strong delineator between high-risk CAS3 and lower-risk CAS1 and 2. There is evidence in the literature for transmission of peanut allergen in utero or via breastmilk (Vadas *et al.*, 2001; DesRoches *et al.*, 2010), as well as early sensitisation via home environmental exposure, especially in those with concurrent eczema or a predisposing filaggrin (*FLG*) mutation that may allow transcutaneous infiltration of allergen (Brough *et al.*, 2013; Brough *et al.*, 2014). The strong correlation between Phadiatop and peanut IgE in the first year of life suggests that either peanut reactivity is significant at this earlier timepoint, or that 'peanut-specific IgE' is cross-reactive and representative of some other allergen hypersensitivity. The fact that this correlation exists within each cluster (Supplementary file 1 – table supplement 6) suggests that it is not caused solely by differences between low- and high-risk clusters (CAS1/CAS2 vs. CAS3). There is a possibility that peanut IgE is a marker for a broader phenotype of early and unremitting sensitisation to multiple food allergens (peanut, cow's milk, eggwhite), as we had observed in CAS3. However, it is unlikely that premature exposure to food allergen is the lone driver for sensitisation and disease, given that well-timed oral exposures to common food allergens (e.g. within 4 to 6 months of age) may actually be protective (Koplin *et al.*, 2010). There is some evidence that quantity (minute vs. abundant), route (skin vs. oral) and timing (early vs. late) of exposure are key modifiers of risk (Han *et al.*, 2017). Ultimately, an underlying atopic predisposition linked to early-life exposure to food allergen may be driving the high-risk phenotype in Cluster 3.

### IgG4 separates individuals susceptible to atopic wheeze from those who are not

In our study, neither IgG nor IgG4 were strong predictors or protectors of wheeze. However, IgG4 was a strong delineator of cluster membership in CAS, with individuals from CAS2 and CAS3 having elevated IgG4 across all specificities compared to CAS1. Vulnerability to early IgE-driven respiratory disease ('atopic wheeze') can be seen in these same individuals – in CAS2 where HDM IgE is predictive for later wheeze, and in CAS3 where both wheeze frequency and IgE are elevated. Hence, although there had previously been doubt about the efficacy of IgG4 as a marker for atopy (EAACI Task Force *et al.*, 2008), our study suggests that IgG4 is still relevant for determining atopic risk, especially when used in combination with IgE.

The underlying biology behind the association of IgG4 with susceptibility to 'atopic wheeze' is unclear. Th2-related pathways drive production of both IgE and IgG4, with IgG4 predominating when modified by concurrent IL-10 signalling (Davies and Sutton, 2015). In susceptible individuals, IgG4 production likely precedes isotype switching to frank IgE production (Aalberse, 2011). Multiple studies have reported that IgG4 is correlated with induced tolerance following desensitisation immunotherapy with high-dose allergen treatment (Davies and Sutton, 2015). However, based on this study alone, we cannot observe any protection from naturally elevated IgG4 levels. Our group had previously suggested, using data from another cohort (Holt *et al.*, 2016), that IgG and specifically IgG1 may provide endogenous protection against IgE-associated wheeze in children experiencing natural (low-level) exposure to aeroallergen. In this present study, IgG1 was not measured.

### The role of respiratory infection and nasopharyngeal microbiome in childhood wheeze differs across different clusters

The co-occurrence of elevated IgE and LRI frequency in CAS3, as well as their predictive effect in CAS2, are consistent with previous findings from CAS (Holt *et al.*, 2010; Teo *et al.*, 2015; Kusel *et al.*, 2007). They lend support to the theory that allergic and infective processes act additively to intensify airway inflammation during respiratory pathogen clearance, which in turn drives progression towards persistent wheeze (Holt and Sly, 2012). In addition, our cluster analysis suggests that the pathologic effect of this interaction may be stratified in discrete subpopulations,

rather than acting in a strictly dose-dependent fashion across the entire cohort. There may be subsets of children (CAS2 and CAS3) who are more susceptible to the effects of this viral-atopy interaction. On the other hand, pathogen clearance in infected non-atopic (CAS1) subjects may be more efficient, due to lack of susceptibility to the pro-inflammatory effects of atopic co-stimuli. This produces lower levels of 'bystander' inflammatory damage to airway tissues, with opportunity for recovery, resulting in a less severe wheeze phenotype.

Of particular note is that, while both CAS1 and CAS2 have LRI and wLRI frequencies as predictors for age-5 wheeze, CAS2 also has fLRI, particularly at age 2. This, along with the general higher incidence of fLRI in CAS3, is consistent with previous findings from CAS (*Holt et al., 2010; Teo et al., 2015*). It suggests that symptomatically severe infections, correlating with severe airway inflammation, may be more potent in causing persistence of wheeze, specifically among those who are 'atopic' (CAS2 and CAS3).

In addition, even during periods of good health, the upper respiratory microbiome played a role in determining later childhood wheeze. Its effect interacted with cluster membership, as well as the age at which the microbiome changes occurred. CAS3 was enriched for early-life infection-associated MPGs (*Streptococcus, Moraxella, and Haemophilus*). This was consistent with the previous finding by *Teo et al., 2017* that early-life infection-associated MPGs were predictive of age-5 wheeze only within atopic individuals (as defined by IgE alone). Interestingly, in our current study, we found a protective effect of infection-associated MPGs from age 2 to 4 in CAS1. We hypothesise that those without atopy-related immune dysfunction are able to maintain a healthy trajectory by responding appropriately to stimuli from potential pathogens that colonise the respiratory tract, thus achieving protection against future (non-atopic) wheeze. This is akin to the 'hygiene hypothesis': exposure to a greater repertoire of pathogen-derived antigens may facilitate maturation of immune functions against said pathogens. Meanwhile, individuals with a predisposing immune dysfunction (i.e. 'atopy' manifesting in early-life allergic sensitisation) may be responding in a maladaptive manner to these microbes (*Holt and Sly, 2012*). This may result in inability to clear potential pathogenic bacteria, or shaping of aberrant immune responses – with subsequent effects on airway inflammation and wheeze.

### Implications for cluster analysis in asthma research

In this study, we applied mixture modelling to generate clusters from biological data. Similar methods such as latent class analysis (LCA) have previously been used in asthma research – for instance, LCA was applied to SPT and IgE measurements from MAAS to determine different patterns of allergen sensitisation and subsequent disease (*Lazic et al., 2013*). However, LCA is limited to categorical clustering features, so measures of sensitisation in that study were thresholded (e.g. IgE levels were split into <0.35 kU/L, 0.35 to 100 kU/L, and >100 kU/L). The method also assumed that these thresholds have the same relevance across all timepoints; that thresholds applied equally to all allergens; and that all allergens contributed equally to disease susceptibility profiles. Mixture modelling is an extension of LCA in that it does not require categorical variables or predetermined thresholds. Furthermore, non-parametric mixture modelling (npEM) does not require input features to have Gaussian distributions. Previous studies have used mixture models to explore phenotypes in adult asthma based on clinical measurements (*Janssens et al., 2012; Newby et al., 2014; Burte et al., 2015*), and one of our own studies previously looked at cytokine expression patterns of PBMCs from children in response to HDM stimulation (*Wu et al., 2015*). Our study is the first to apply non-parametric mixture modelling to data representing immune and respiratory health in early childhood, and to investigate possible predictors of disease within each cluster.

Currently, mixture models are limited by an unproven 'track record'; a lack of consensus about best protocols for data processing and analysis; instability or inconsistency of clusters; difficulty in interpretation of results; and uncertainty regarding the validity of certain assumptions that accompany models (*Deliu et al., 2016*). Other methods of cluster analysis have similar problems, and while they have been applied frequently to asthma research, they have also produced a confusing myriad of phenotypes. The nature of cluster phenotypes is highly dependent on the type of features entered into the clustering algorithm. Clustering features that represent final clinical endpoints, such as markers of severity, may produce more heterogeneous clusters, as different pathological trajectories can arrive at similar endpoints. Some cluster phenotypes may contradict with each other, or may not be easily interpreted. Recently, *Schoos et al. (2017)* identified that, unlike our study, asthma

was not as strongly associated with prominent HDM or peanut hypersensitivity in a Danish birth cohort (COPSAC) as other patterns of sensitisation (especially cat, dog and horse). However, we note that they used thresholded IgE >0.35 kU/L to build their clusters. Other differences may emerge due to heterogeneity across different populations; geographical differences in environmental exposures and allergen sensitisation; and differences in testing procedures and phenotype definitions at different sites. COPSAC, CAS and COAST were cohorts enriched for high-risk individuals – each child had at least one parent with a history of atopic disease – while MAAS had no such recruitment criterion. Because of variability in findings, there has been wariness and scepticism among clinicians regarding the utility of mixture models and machine learning (*Chen and Asch, 2017*). Ultimately, one may argue that discrepancies in our findings serve as a caution against the blind application of ‘algorithms’ without due consideration of subtleties in target population and environment.

Nonetheless, what we have demonstrated here is the vast potential of cluster analysis. We have discovered clusters in an unsupervised and exploratory fashion, described them comprehensively, replicated our findings in multiple datasets, and compared our clusters with other existing phenotypes. In doing so, we have generated some new and interesting insights about the nature of atopy and asthma risk. Our results build on previous findings (*Frith et al., 2011; Klink et al., 1990*) demonstrating that the concept of atopy, as an intrinsic or heritable predisposition to allergic disease, is more complicated than what could be described by dichotomies or thresholds. We have also demonstrated that addressing subgroup differences via cluster analysis allows for identification of intra-cluster disease predictors. In the future, clusters may be further characterised by other aspects of asthma pathophysiology, such as genomics, transcriptomics, and epigenomics.

## Concluding statements

The results of our study strongly support the future use of predictive models with more precise and subgroup-driven representations of atopy or other relevant pathophysiology. We argue for ongoing collaboration between research groups in terms of refining methodology, answering questions unique to certain populations, and comparing cluster phenotypes arising from different algorithms and datasets. We believe that, as clustering methods become more frequently used, we will gradually develop better consensus on how such methods are best applied to biomedical phenomena. By continuing with these approaches, we can hopefully move away from fixed thresholds to more sophisticated formulations of risk, which will then improve future attempts at targeted screening, prevention and treatment of asthma. These approaches are already being applied to other heterogeneous diseases, and in the future computerised tools may be designed to embody the sum knowledge from these approaches. Such approaches can eventually help clinicians and scientists achieve a fuller understanding of pathophysiology, and hence better predict and manage human disease.

## Materials and methods

### Key resources table

Reagent type (species) or resource	Designation	Source or reference	Identifiers
Biological sample ( <i>Homo sapiens</i> )	Childhood Asthma Study (CAS)	DOI: 10.1016/j.jaci.2005.06.038	Microbiome sequencing data: NCBI GenBank SRP056779
Biological sample ( <i>Homo sapiens</i> )	Childhood Origins of Asthma Study (COAST)	PMID:12688623	NA
Biological sample ( <i>Homo sapiens</i> )	Manchester Asthma and Allergy Study (MAAS)	PMID:12688622	NA
Software, algorithm	The R project for Statistical Computing	ISBN:3-900051-07-0	RRID:SCR_001905
Software, algorithm	ggplot2	ISBN:978-3-319-24277-4	RRID:SCR_014601

Continued on next page

Continued

Reagent type (species) or resource	Designation	Source or reference	Identifiers
Software, algorithm	mixtools	DOI: 10.18637/jss.v032.i06	NA
Software, algorithm	rpart	<b>Therneau and Atkinson, 2015.</b> Package 'rpart'. URL: <a href="https://cran.r-project.org/web/packages/rpart/rpart.pdf">https://cran.r-project.org/web/packages/rpart/rpart.pdf</a>	NA
Software, algorithm	epiDisplay	<b>Chongsuvivatwong, 2015.</b> Package 'epiDisplay' URL: <a href="https://cran.r-project.org/web/packages/epiDisplay/epiDisplay.pdf">https://cran.r-project.org/web/packages/epiDisplay/epiDisplay.pdf</a>	NA

## Patients and study design in CAS

Our discovery dataset was the Childhood Asthma Study (CAS), a prospective birth cohort ( $N = 263$ ) operated by the Telethon Kids Institute from Perth, Western Australia (**Kusel et al., 2005**). The goal of CAS was to describe the risk factors and pathogenesis of childhood allergy and asthma. Further details of CAS have been reported previously (**Kusel et al., 2005; Hollams et al., 2009; Holt et al., 2010; Teo et al., 2015; Hollams et al., 2017**).

In CAS, expectant parents were recruited from private paediatric clinics in Perth during the period spanning July 1996 to June 1998. Each child who was born and subsequently recruited had at least one parent with physician-diagnosed asthma or atopic disease (hayfever, eczema). The child was then followed from birth till age 10 at the latest, with routine medical examinations, clinical questionnaires, blood sampling at multiple time points (6–7 weeks, 6 months, 1 year, 2, 3, 4, 5, and 10 years) and collection of nasopharyngeal samples. Parents also kept a daily symptom diary for symptoms of respiratory infection in their child. The data extracted from these samples and measurements covered multiple 'domains' of asthma pathogenesis, including respiratory infection, allergen sensitisation, and clinical or demographic background.

## Measurements in CAS

For each child and visit, the investigators of CAS recorded metrics related to suspected or known modulators of asthma risk. These included markers of immune function: (1) IgG, IgG4, and IgE Phadiatop ImmunoCAP antibodies (ThermoFisher, Uppsala, Sweden), covering common allergens such as house-dust mite (HDM, *Dermatophagoides pteronyssinus*), mould, couch grass, ryegrass, peanut, cat dander; (2) IgE and IgG4 Phadiatop Infant and Adult assays (ThermoFisher, Uppsala, Sweden) that target multiple allergens simultaneously (**Ballardini et al., 2006**); (3) skin prick or sensitisation tests (SPT) for HDM, mould, ryegrass, cat, peanut, cow's milk and hen's egg; and (4) cytokine responses (IL-4,5,9,13,10, IFN- $\gamma$ ) following in vitro stimulation of extracted peripheral blood mononuclear cells (PBMCs) by multiple antigen and allergen stimuli, including phytohaemagglutinin (PHA), HDM, cat, peanut and ovalbumin (**Hollams et al., 2009; Holt et al., 2010**).

In addition, nasopharyngeal samples (swabs or aspirates, NPAs) were taken from each child during healthy routine visits (healthy samples), and unscheduled visits where parents presented with their child if they have a suspected respiratory infection (disease samples). Frequency and severity of respiratory infections were measured accordingly. NPAs were then screened for viral and bacterial pathogens using rtPCR and 16 s rRNA amplicon sequencing with Illumina MiSeq (San Diego, US), respectively (**Teo et al., 2015**). These NPAs had previously classified by **Teo et al. (2015); Teo et al. (2017)**, based on clustering of bacterial composition, into microbiome profile groups (MPGs) that were associated with healthy respiratory states (health-associated MPGs, for example *Alloicoccus*-, *Staphylococcus*- or *Corynebacterium*-dominated) or infectious respiratory states (infection-associated MPGs, for example *Moraxella*-, *Haemophilus*-, or *Streptococcus*-dominated).

Other collected data included: sex, height and weight; paternal and maternal history of atopic disease; blood levels of basophils, plasmacytoid and myeloid dendritic cells as measured by

fluorescence-assisted cell sorting (FACS); and levels of vitamin D (25-hydroxycholecalciferol, 25(OH) D) (Hollams *et al.*, 2017).

## Identification of latent clusters and selection of clustering features

We adopted an exploratory approach to cluster analysis, whereby we attempted to interrogate as much of the existing dataset as possible, identifying latent clusters that arise from the underlying data structure of CAS. We then assessed how these latent clusters correlate with risk of asthma or other markers of pathophysiology, such as degree of allergic sensitisation. All data processing and analysis were done in R v3.3.1 (RRID:SCR\_001905). A graphical overview of the analytic process is displayed in **Figure 1—figure supplement 3**.

To identify latent clusters, we applied non-parametric expectation-maximisation ('npEM') mixture modelling to our discovery cohort CAS, using functions from the R package 'mixtools' (Benaglia *et al.*, 2009a). This method assumes that frequency distributions of each cluster can be represented by non-parametric density estimates learned from the data in an iterative process. npEM was used because: (1) it was plausible to consider a population as a mixture of subpopulations each with their own distributions; (2) it had advantages over other unsupervised approaches (Tan *et al.*, 2005) – for example, with LCA, continuous variables cannot be handled appropriately; with hierarchical clustering, poor decisions made early in the classifying process are not easily amended; (3) many variables were categorical or non-Gaussian, so theoretically a non-parametric approach should be superior to a Gaussian mixture model or k-means approach; and (4) inherent within mixture models is an intuitive method for supervised classification of other datasets into similar clusters.

We used a largely non-selective approach to choosing features for cluster analysis, in that we attempted to retain as many CAS individuals and variables as possible. However, we did enforce certain quality-control measures such as excluding variables ('features') that had missing data for >20% subjects (442 variables removed), and subjects with missing data for >30% of the remaining variables (39 subjects removed). Also excluded were features pertaining to our primary outcomes of interest: incidence of parent-reported wheeze, physician-diagnosed asthma and hayfever at all timepoints. We specifically excluded these from feature selection so we could determine how subsequent clusters differ in these outcomes, even when clustering was not explicitly driven by them. On the other hand, eczema was not excluded because of evidence that infantile eczema may itself influence the risk for subsequent sensitisation and asthma (Gustafsson *et al.*, 2000). Frequency of wheeze in the context of respiratory infection was also included, as it was a symptomatic marker of infection severity. Variable reduction resulted in  $M = 174$  variables remaining out of an original 659. The complete list of variables included as clustering features is provided in **Supplementary file 1** – table supplement 1, and importantly covers multiple domains including demographic (family history of atopy, household size), clinical (incidence of childhood eczema), immunological (IgE, IgG, IgG4, SPT) and microbiological (respiratory infections, viral pathogens associated with infection) features. By virtue of study design and quality control measures, many of the clustering features were related to immunological function or respiratory infection in the first 3 years of life.

Highly skewed features, such as antibody and cytokine levels, were subjected to logarithmic (base 10) transformation. We also applied limited thresholding to some variables (cytokine responses, antibody assays), based on best practice for the reported limit-of-detection (LOD) of the measuring devices. The LOD for IgE was 0.03 kU/L; for IgG4, 0.0003 µg/L; for IgG, 0.4 mg/L. For these variables, we assigned any values below the LOD to half the LOD (i.e. 0.015 kU/L, 0.00015 µg/L, and 0.2 mg/L, respectively). For stimulated cytokine expression above unstimulated control, any zero or negative values (i.e. unstimulated control had equal, or greater, expression than stimulated), were converted to 0.000001 units or 0.01 pg/ml for mRNA and protein variables, respectively. Positional standardisation scaling was then applied across all variables, to equally weight the contributions of each feature to the mixture model. This involved replacing each value  $x_{ij}$  for individual  $i$  of feature  $j$ , by:

$$\frac{x_{ij} - \text{med}(x_j)}{\max(x_j) - \min(x_j)}$$



where functions med, max and min refer to the median, maximum, and minimum for the complete-case dataset for feature  $j$ , respectively.

### Cluster analysis using non-parametric mixture modelling

The processed and scaled CAS dataset was further split into those subjects with no missingness in the remaining variables ('complete-case', 186 subjects, 174 variables); versus those who had limited missingness of <30% variables ('low-missingness', 36 subjects, 174 variables). Cluster analysis was performed initially in the complete-case CAS subset to generate an npEM model.

The mathematical theory underpinning npEM has already been described extensively in other sources (Benaglia et al., 2009b). In brief, it involves three steps: (1) an expectation or E-step, which calculates the posterior probability of membership in cluster  $k$ , given the observed dataset, estimated mixing proportions  $\lambda_k$ , and probability distribution for  $k$ ; (2) a maximisation or M-step, which calculates the mixing proportions  $\lambda_k$  from the cluster memberships determined above; (3) a non-parametric kernel density estimation step, which calculates the probability distribution based on a kernel density function for each cluster  $k$  and clustering feature  $j$ . These steps were then iterated until the model converged to a point where log-likelihood values were maximised.

As with any EM algorithm, an initial state must first be set prior to commencing the iterative process. To do this, we used a constant seed state ('set.seed(1)') to allow reproducibility of results. Based on these pseudo-random centroids for a set number of clusters  $L$ , the initial state was then determined by k-means clustering as in Benaglia et al (Benaglia et al., 2009b). The other options in npEM were set to defaults. These included the use of non-stochastic (deterministic) as opposed to a stochastic method; the use of a standard normal density function as the kernel function; and the use of default constant bandwidths for estimating kernel densities (Benaglia et al., 2009b).

The ideal number of clusters  $L$  was determined by two methods. Firstly, we performed hierarchical clustering on the complete-case dataset, and scrutinised the dendrogram as well as a scree plot for an optimal cut-off using the 'knee method' (Tan et al., 2005). We observed that this occurred at around  $L = 3$  or 4. Secondly, we repeated npEM clustering for values of  $L = 1, 2, \dots, 20$ , and calculated the Bayesian information criterion (BIC) for each of these, using the formula:

$$\text{BIC} = -2\log(\hat{p}) + \nu \log(N)$$

where  $P$  is the maximum likelihood,  $\nu = L \times M + (L - 1)$ , and  $L$ ,  $M$ ,  $N$  are total number of clusters, clustering features, and individuals respectively. The optimal number of clusters was again determined to be around  $L = 3$  or 4, based on minimum BIC observed. For the sake of parsimony, we set the number of clusters to three.

### Classification of test datasets using mixture model densities

The density functions generated by the resultant npEM model were then used to classify as many subjects of the low-missingness subset as possible. This method relied on the assumption that distributions observed in the 'training' (complete-case) dataset were representative of distributions that existed in 'test' (low-missingness or external) datasets.

Classification was performed as follows: consider individual  $i$  of  $N$ ; clustering feature or coordinate  $j$  of  $M$ ; and component or cluster  $k$  of  $L$ . For each individual  $i$  belonging to known cluster  $k=K$ , let the kernel density function for coordinate  $j$  be  $f_{jK}(x_{ij})$ . We now assume that the coordinates  $j$  were independent of each other. Although this was not truly the case – for instance, weak correlation exists between IgE and IgG4 of different allergen specificities in the CAS dataset [23] – we believed the assumption was justified given our theory-naive and exploratory approach. With this assumption, the joint distribution for individual  $i$  in cluster  $K$  should be the product of density functions for all  $j$  given  $K$ . and therefore the probability of individual with value  $x_{ij}$  belonging to cluster  $K$  was:

$$P(k = K|x_{ij}) = \frac{\lambda_K \prod_{j=1}^M f_{jK}(x_{ij})}{\sum_{k=1}^L \lambda_k \prod_{j=1}^M f_{jk}(x_{ij})}$$

In addition to this, we made two other assumptions: (1) if  $x_{ij}$  was missing, then the density was assumed to be one, or  $f_{jK}(x_{ij}) = 1$ ; (2) else, if  $x_{ij} < \min(x_j)$ , the minimum value in feature  $j$  for which

there was a non-zero density value, then the density was equal to that of the minimum value, that is  $f_{jK}(x_{ij}) = f_{jK}(\min(x_j))$ . Likewise, if  $x_{ij} > \max(x_j)$ , then  $f_{jK}(x_{ij}) = f_{jK}(\max(x_j))$ .

Individuals with membership probability greater than 90% for cluster  $K$  were classified into  $K$ . Using this method, an additional 31 individuals from 36 were successfully classified into one of three clusters, for a total combined dataset of 217 classified individuals in CAS.

Finally, we formally defined each CAS cluster using the composite of complete-case and low-missingness datasets, and described each cluster in terms of key characteristics and significant cluster-specific predictors for age-5 wheeze. Importantly, variables that were initially excluded from feature selection were treated as subsequent outcomes for post-hoc comparison of clusters.

## Replication cohorts

The study designs and measurements for the two replication cohorts – the Manchester Asthma and Allergy Study (MAAS) ( $N = 1085$ ) from Manchester, UK, and the Childhood Origins of Asthma Study (COAST) ( $N = 289$ ) from Wisconsin, USA – have been described elsewhere (*Belgrave et al., 2014; Gern et al., 2002; NAC Manchester Asthma and Allergy Study Group et al., 2002; Lemanske, 2002*). COAST, like CAS, was comprised of high-risk individuals with a known family history of asthma or allergy; while MAAS included individuals without family history.

In terms of matching variables for replication, all cohorts had measurements that covered the three major ‘domains’ of asthma pathogenesis: respiratory infection, allergen sensitisation, and clinical or demographic background. COAST had a comprehensive collection of respiratory infection and IgE-type measurements, but no IgG4 measurements. MAAS had multiple measurements of IgE and SPT-type variables. Following consultation with investigators from all three cohorts, clustering features were matched based on proximity of timepoint and phenotype. Respiratory infection phenotypes (ARI, LRI, URI, fLRI, wLRI) were generated in COAST and MAAS using recorded data, to approximate CAS infection phenotypes as closely as possible. Specifically, LRI was defined as respiratory infection with evidence of lower respiratory tract involvement in the form of chest sounds (wheeze, rattle, whistle), or increased respiratory effort (retractions, tachypnea, cyanosis); URI was defined as a cold-like infection limited to the upper respiratory tract, without signs of LRI. IgE and IgG4 assays for MAAS and COAST were performed using ImmunoCAP and UniCAP, respectively. Both replication cohorts recorded basic demographic data, and exposures to pets, childcare, and tobacco smoke. The complete list of clustering features and the matching scheme across cohorts is provided in **Supplementary file 1** – table supplement 1.

The npEM clusters were described and validated in MAAS and COAST. This replication was performed by applying the density-function-derived classifier used previously for the low-missingness CAS subjects. Because these external cohorts did not necessarily share the same clustering features or variables as CAS (**Supplementary file 1** – table supplement 1), we assumed that the respective densities for these variables were  $f_{jK}(x_{ij}) = 1$  for the  $j^{\text{th}}$  feature and  $K^{\text{th}}$  cluster. In doing so, this was effectively the same as using a model where the missing features were excluded, and only those features common to both CAS and MAAS (or COAST) were used; or equivalently, where we assumed that each member of MAAS or COAST was missing values in those particular features. Because these ‘CAS-derived’ npEM models were non-identical to the original npEM models in CAS, we tested whether ‘MAAS-like’ and ‘COAST-like’ algorithms (CAS-derived model as applied to MAAS or COAST, respectively) generated similar clusters to the original CAS clusters, when applied back onto CAS (Results).

## Cluster validity and stability

Internal validation of the clusters in the complete-case CAS dataset was performed by use of silhouette widths. Briefly, we calculated the silhouette widths for each cluster as per *Rousseeuw (1987)*. For an individual, the closer the silhouette width is to one, the more appropriate the cluster membership; while the closer it is to negative one, the more likely it has been misclassified.

Cluster stability was assessed by performing leave-one-out (LOO) analysis – that is, we applied the npEM algorithm to a subset of the complete-case dataset – an  $N-1$  by  $M$  dataset ( $N = 186$ ,  $M = 174$ ) for a total of  $N$  times, leaving out an individual each time. A similar process was repeated  $M$  times on an  $N$  by  $M-1$  dataset, leaving out one clustering feature at a time. The Jaccard indices for each iteration were then calculated in comparison to known clusters from the original complete-

case  $N$  by  $M$  dataset, and averaged across each assigned cluster. Cluster labels for each iteration were assigned based on whichever complete-case cluster yielded the smallest Jaccard index. This whole process was then repeated with 10 random seeds ('set.seed(1)' through to 'set.seed(10)') for determining the initial state for npEM. The final averaged Jaccard indices for each cluster thus represented the mean stability of each cluster.

## Decision tree analysis

Decision tree analysis was performed using a number of different partitioning schemes. Classification trees with recursive partitioning were built from CAS clusters using the R package 'rpart' (Therneau and Atkinson, 2015), an open-source implementation of CART. The motivation for decision trees was to identify the variables that most strongly separated the clusters and wheezing status, and not necessarily variables that were most predictive.

For tree outcomes (end-nodes), we investigated both cluster membership and presence of age-5 wheeze given cluster membership. That is, decision trees were generated to identify the biological features that most strongly distinguished each npEM cluster ('Simple Tree'), as well as npEM cluster  $\times$  age-5 wheeze status ('Comprehensive Tree').

We used two different schemes for selecting predictors on which to base the partitions: 1) include all predictors that were used as clustering features in the original npEM model; 2) include only predictors from one timepoint (variables from age 6 m, 1, 2 or 3). The motivation for the latter was that we wanted to see whether measurements taken at a specific timepoint in early infancy could strongly distinguish between clusters. For the former scheme, we excluded all age-5 features related to wheeze (e.g. LRIs, wheezy LRIs at age 5) as decision nodes, because of definitional overlap with our primary outcome of interest (age-5 wheeze).

Decision trees were then pruned based on the complexity parameter that minimised cross-validated error. Final classification into tree clusters was manually performed based on the pruned tree, and not by automatic classification using the 'predict' function for the 'rpart' tree object – this was because, for the latter, individuals who are missing key variables were re-classified based on the next best, non-missing, surrogate variable (Therneau and Atkinson, 2015). Thus, it resulted in children being erroneously classified into a tree cluster even when they were missing key classifier variables.

The decision tree analyses generated thresholds which were then compared with existing thresholds for atopy (any specific IgE at age 2  $\geq$  0.35 kU/L, and/or any specific SPT at age 2  $\geq$  2 mm) (Frith et al., 2011) in terms of predicting disease outcomes of interest.

## Statistical analyses

We performed statistical analyses comparing clusters in terms of multiple variables, especially those not used as clustering features. Of interest to us were the primary outcomes of asthma diagnosis and parent-reported wheeze at each timepoint. Where appropriate, we used t-tests, Mann-Whitney-Wilcoxon tests, ANOVAs, Kruskal-Wallis tests, chi-squared and Fisher exact tests; and logistic and linear regression. For summary statistics, multiple testing adjustment was performed using the Benjamini-Yekutieli (BY) method, for all across-cluster tests (Cluster  $\times$  trait); and for all comparisons between clusters (CAS1 vs. 2, 1 vs. 3, and 2 vs. 3). The BY method was chosen as it accounted for positive dependency across the highly correlated variables in the CAS dataset (Benjamini and Yekutieli, 2001). For variables that underwent logarithmic transformation for statistical analysis, we used geometric mean to describe central tendency.

We then determined the predictors for age-5 wheeze within each cluster. Repeated-measures ANOVAs were performed for selected predictors of age-5 wheeze. For each potential predictor, generalised linear regression models (GLMs) were generated with and without a base set of covariates (sex, family history of asthma, BMI where available). The pool of variables found to be statistically significant (at least  $p < 0.05$ ) in the above analyses were further restricted, such that strongly collinear predictors were avoided, and at most one timepoint was considered for each predictor type. Targeted multiple regression models were then built by selecting predictors from this constrained pool. Stepwise backward elimination was applied, in which the predictor with the largest  $p$ -value was eliminated at each step, until all remaining predictors have significant  $p < 0.05$ .

Using the 'lrtest' function from the R package 'Epidisplay' (Chongsuvivatwong, 2015), likelihood ratios were examined to check how much cluster membership or classification improved upon prediction of age-5 wheeze compared to traditional makers of atopy.

## Additional information

### Funding

Funder	Grant reference number	Author
National Health and Medical Research Council	1049539	Michael Inouye
National Health and Medical Research Council	PhD Scholarship	Howard HF Tang

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.


### Author contributions

Howard HF Tang, Conceptualization, Data curation, Formal analysis, Validation, Investigation, Visualization, Methodology, Writing—original draft, Writing—review and editing; Shu Mei Teo, Data curation, Formal analysis, Methodology, Writing—review and editing; Danielle CM Belgrave, Resources, Software, Formal analysis, Validation, Methodology, Writing—review and editing; Michael D Evans, James E Gern, Resources, Data curation, Formal analysis, Validation, Methodology, Writing—review and editing; Daniel J Jackson, Robert F Lemanske, Resources, Data curation, Validation, Methodology, Writing—review and editing; Marta Brozynska, Conceptualization, Resources, Data curation, Formal analysis, Validation, Investigation, Visualization, Methodology, Writing—original draft, Project administration, Writing—review and editing; Merci MH Kusel, Resources, Data curation, Formal analysis, Methodology, Writing—review and editing; Sebastian L Johnston, Resources, Data curation, Software, Formal analysis, Funding acquisition, Validation, Methodology, Writing—review and editing; Angela Simpson, Resources, Validation, Methodology, Writing—review and editing; Adnan Custovic, Resources, Data curation, Validation, Writing—review and editing; Peter D Sly, Conceptualization, Resources, Data curation, Validation, Writing—original draft, Project administration, Writing—review and editing; Patrick G Holt, Conceptualization, Resources, Data curation, Supervision, Methodology, Writing—original draft, Writing—review and editing; Kathryn E Holt, Conceptualization, Resources, Formal analysis, Supervision, Funding acquisition, Validation, Methodology, Writing—original draft, Writing—review and editing; Michael Inouye, Conceptualization, Formal analysis, Supervision, Funding acquisition, Methodology, Writing—original draft, Project administration, Writing—review and editing

### Author ORCIDs

Howard HF Tang  <http://orcid.org/0000-0001-6422-0270>

Michael D Evans  <http://orcid.org/0000-0001-7449-3993>

Adnan Custovic  <http://orcid.org/0000-0001-5218-7071>

Michael Inouye  <http://orcid.org/0000-0001-9413-6520>

### Ethics

Human subjects: Ethics approval and consent requirements for each cohort were met as follows: The CAS study was approved by the ethics committees of the King Edward Memorial and Princess Margaret Hospitals in Western Australia; fully informed parental consent was obtained for all subjects. The COAST study was approved by the Human Subjects Committee of the University of Wisconsin. The MAAS study was approved by a Manchester Local Research Ethics Committee (ERP/94/032; SOU/00/258; 03/SM/400; Study registration ISRCTN72673620); fully informed parental consent was obtained for all subjects across all cohorts.

**Decision letter and Author response**Decision letter <https://doi.org/10.7554/eLife.35856.034>Author response <https://doi.org/10.7554/eLife.35856.035>**Additional files****Supplementary files**

- Supplementary file 1. All table supplements

DOI: <https://doi.org/10.7554/eLife.35856.025>

- Supplementary file 2. Comparison of variables (respiratory, immunological, clinical) across CAS clusters. Analogous to Table Supplement 3.

DOI: <https://doi.org/10.7554/eLife.35856.026>

- Supplementary file 3. Predictors for age-five wheeze within each CAS cluster, with demographic covariates (sex, BMI, parental history of asthma). Analogous to Table Supplement 7.

DOI: <https://doi.org/10.7554/eLife.35856.027>

- Transparent reporting form

DOI: <https://doi.org/10.7554/eLife.35856.028>**Data availability**

This study utilises extensive data from human subjects, specifically paediatric cohorts, for which eLife's policies recognise that there can be strong reasons to restrict access. For each of the cohorts involved in our study (CAS, COAST, MAAS), parents were consented on the use of biomedical data for allergy and asthma research, but not for the open sharing of their or their children's data. Studies were run in the late 1990s and early 2000s and we do not have ethics permission to attempt to recontact families to seek consent. Importantly, we note that key data features could risk re-identification of subjects (e.g. demographic data from small communities). However, we have provided public data at the summary level which can be used for subsequent studies, such as replication and meta-analysis. This is standard practice in sensitive data settings, such as genome-wide association studies. These data have been uploaded as Excel spreadsheets to FigShare for ease of data extraction: **Supplementary Table 3** [https://figshare.com/articles/Supplementary\\_File\\_1\\_1/6934052](https://figshare.com/articles/Supplementary_File_1_1/6934052); **Supplementary Table 7** [https://figshare.com/articles/Supplementary\\_File\\_1\\_2/6934055](https://figshare.com/articles/Supplementary_File_1_2/6934055)

The following datasets were generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Howard HF Tang, Michael Inouye, Kathryn E Holt	2018	Data from supplementary table 3	<a href="https://figshare.com/articles/Supplementary_File_1_1/6934052">https://figshare.com/articles/Supplementary_File_1_1/6934052</a>	Figshare, Supplementary_File_1_1/6934052
Howard HF Tang, Michael Inouye, Kathryn E Holt	2018	Data from supplementary table 7	<a href="https://figshare.com/articles/Supplementary_File_1_2/6934055">https://figshare.com/articles/Supplementary_File_1_2/6934055</a>	Figshare, Supplementary_File_1_2/6934055

**References**

- Aalberse R.** 2011. The role of IgG antibodies in allergy and immunotherapy. *Allergy* **66**:28–30. DOI: <https://doi.org/10.1111/j.1398-9995.2011.02628.x>, PMID: 21668848
- Almqvist C, Worm M, Leynaert B, working group of GA2LEN WP 2.5 Gender.** 2008. Impact of gender on asthma in childhood and adolescence: a GA2LEN review. *Allergy* **63**:47–57. DOI: <https://doi.org/10.1111/j.1398-9995.2007.01524.x>, PMID: 17822448
- Anderson GP.** 2008. Endotyping asthma: new insights into key pathogenic mechanisms in a complex, heterogeneous disease. *The Lancet* **372**:1107–1119. DOI: [https://doi.org/10.1016/S0140-6736\(08\)61452-X](https://doi.org/10.1016/S0140-6736(08)61452-X), PMID: 18805339
- Ballardini N, Nilsson C, Nilsson M, Lilja G.** 2006. ImmunoCAP Phadiatop Infant—a new blood test for detecting IgE sensitisation in children at 2 years of age. *Allergy* **61**:337–343. DOI: <https://doi.org/10.1111/j.1398-9995.2005.00936.x>, PMID: 16436143
- Bantz SK, Zhu Z, Zheng T.** 2014. The atopic march: progression from atopic dermatitis to allergic rhinitis and asthma. *Journal of Clinical & Cellular Immunology* **5**:67–73. DOI: <https://doi.org/10.4172/2155-9899.1000202>
- Belgrave DCM, Simpson A, Semic-Jusufagic A, Murray CS, Buchan I, Pickles A, Custovic A.** 2013. Joint modeling of parentally reported and physician-confirmed wheeze identifies children with persistent troublesome

- wheezing. *Journal of Allergy and Clinical Immunology* **132**:575–583. DOI: <https://doi.org/10.1016/j.jaci.2013.05.041>, PMID: 23906378
- Belgrave DC**, Granell R, Simpson A, Guiver J, Bishop C, Buchan I, Henderson AJ, Custovic A. 2014. Developmental profiles of eczema, wheeze, and rhinitis: two population-based birth cohort studies. *PLoS medicine* **11**:e1001748. DOI: <https://doi.org/10.1371/journal.pmed.1001748>, PMID: 25335105
- Benaglia T**, Chauveau D, Hunter DR, Young DS. 2009a. Mixtools: an R package for analyzing mixture models. *Journal of Statistical Software* **32**.
- Benaglia T**, Chauveau D, Hunter DR. 2009b. An EM-Like Algorithm for Semi- and Nonparametric Estimation in Multivariate Mixtures. *Journal of Computational and Graphical Statistics* **18**:505–526. DOI: <https://doi.org/10.1198/jcgs.2009.07175>
- Benjamini Y**, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29**:1165–1188.
- Bisgaard H**, Hermansen MN, Buchvald F, Loland L, Halkjaer LB, Bønnelykke K, Brasholt M, Heltberg A, Vissing NH, Thorsen SV, Stage M, Pipper CB. 2007. Childhood asthma after bacterial colonization of the airway in neonates. *New England Journal of Medicine* **357**:1487–1495. DOI: <https://doi.org/10.1056/NEJMoa052632>, PMID: 17928596
- Bousquet J**, Anto JM, Wickman M, Keil T, Valenta R, Haahtela T, Lodrup Carlsen K, van Hage M, Akdis C, Bachert C, Akdis M, Auffray C, Annesi-Maesano I, Bindslev-Jensen C, Cambon-Thomsen A, Carlsen KH, Chatzi L, Forastiere F, Garcia-Aymerich J, Gehrig U, et al. 2015. Are allergic multimorbidities and IgE polysensitization associated with the persistence or re-occurrence of foetal type 2 signalling? the MeDALL hypothesis. *Allergy* **70**:1062–1078. DOI: <https://doi.org/10.1111/all.12637>, PMID: 25913421
- Brough HA**, Santos AF, Makinson K, Penagos M, Stephens AC, Douiri A, Fox AT, Du Toit G, Turcanu V, Lack G. 2013. Peanut protein in household dust is related to household peanut consumption and is biologically active. *The Journal of allergy and clinical immunology* **132**:630–638. DOI: <https://doi.org/10.1016/j.jaci.2013.02.034>, PMID: 23608730
- Brough HA**, Simpson A, Makinson K, Hankinson J, Brown S, Douiri A, Belgrave DC, Penagos M, Stephens AC, McLean WH, Turcanu V, Nicolaou N, Custovic A, Lack G. 2014. Peanut allergy: effect of environmental peanut exposure in children with filaggrin loss-of-function mutations. *Journal of Allergy and Clinical Immunology* **134**:867–875. DOI: <https://doi.org/10.1016/j.jaci.2014.08.011>, PMID: 25282568
- Burte E**, Bousquet J, Varraso R, Gormand F, Just J, Matran R, Pin I, Siroux V, Jacquemin B, Nadif R. 2015. Characterization of rhinitis according to the asthma status in adults using an unsupervised approach in the EGEA study. *Plos One* **10**:e0136191. DOI: <https://doi.org/10.1371/journal.pone.0136191>, PMID: 26309034
- Calderón MA**, Linneberg A, Kleine-Tebbe J, De Blay F, Hernandez Fernandez de Rojas D, Virchow JC, Demoly P. 2015. Respiratory allergy caused by house dust mites: What do we really know? *Journal of Allergy and Clinical Immunology* **136**:38–48. DOI: <https://doi.org/10.1016/j.jaci.2014.10.012>, PMID: 25457152
- Castro-Rodríguez JA**, Holberg CJ, Wright AL, Martinez FD. 2000. A clinical index to define risk of asthma in young children with recurrent wheezing. *American Journal of Respiratory and Critical Care Medicine* **162**:1403–1406. DOI: <https://doi.org/10.1164/ajrccm.162.4.9912111>, PMID: 11029352
- Chen JH**, Asch SM. 2017. Machine learning and prediction in medicine - Beyond the peak of inflated expectations. *New England Journal of Medicine* **376**:2507–2509. DOI: <https://doi.org/10.1056/NEJMp1702071>, PMID: 28657867
- Chongsuvivatwong V**. 2015. *epiDisplay: Epidemiological Data Display Package*. <https://cran.r-project.org/web/packages/epiDisplay/>
- Davies AM**, Sutton BJ. 2015. Human IgG4: a structural perspective. *Immunological Reviews* **268**:139–159. DOI: <https://doi.org/10.1111/imr.12349>, PMID: 26497518
- Deliu M**, Sperrin M, Belgrave D, Custovic A. 2016. Identification of asthma subtypes using clustering methodologies. *Pulmonary Therapy* **2**:19–41. DOI: <https://doi.org/10.1007/s41030-016-0017-z>, PMID: 27512723
- DesRoches A**, Infante-Rivard C, Paradis L, Paradis J, Haddad E. 2010. Peanut allergy: is maternal transmission of antigens during pregnancy and breastfeeding a risk factor? *Journal of Investigational Allergology & Clinical Immunology* **20**:289–294. PMID: 20815306
- Dick S**, Friend A, Dynes K, Alkandari F, Doust E, Cowie H, Ayres JG, Turner SW. 2014. A systematic review of associations between environmental exposures and development of asthma in children aged up to 9 years. *BMJ Open* **4**:e006554. DOI: <https://doi.org/10.1136/bmjopen-2014-006554>, PMID: 25421340
- EAACI Task Force**, Stapel SO, Asero R, Ballmer-Weber BK, Knol EF, Strobel S, Vieths S, Kleine-Tebbe J. 2008. Testing for IgG4 against foods is not recommended as a diagnostic tool: eaaci task force report. *Allergy* **63**:793–796. DOI: <https://doi.org/10.1111/j.1398-9995.2008.01705.x>, PMID: 18489614
- Frith J**, Fleming L, Bossley C, Ullmann N, Bush A. 2011. The complexities of defining atopy in severe childhood asthma. *Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology* **41**:948–953. DOI: <https://doi.org/10.1111/j.1365-2222.2011.03729.x>, PMID: 21477182
- Gabet S**, Just J, Couderc R, Seta N, Momas I. 2016. Allergic sensitisation in early childhood: patterns and related factors in PARIS birth cohort. *International Journal of Hygiene and Environmental Health* **219**:792–800. DOI: <https://doi.org/10.1016/j.ijheh.2016.09.001>, PMID: 27649627
- Gern JE**, Martin MS, Anklam KA, Shen K, Roberg KA, Carlson-Dakes KT, Adler K, Gilbertson-White S, Hamilton R, Shult PA, Kirk CJ, Da Silva DF, Sund SA, Kosorok MR, Lemanske RF. 2002. Relationships among specific viral pathogens, virus-induced interleukin-8, and respiratory symptoms in infancy. *Pediatric Allergy and Immunology* **13**:386–393. DOI: <https://doi.org/10.1034/j.1399-3038.2002.01093.x>, PMID: 12485313

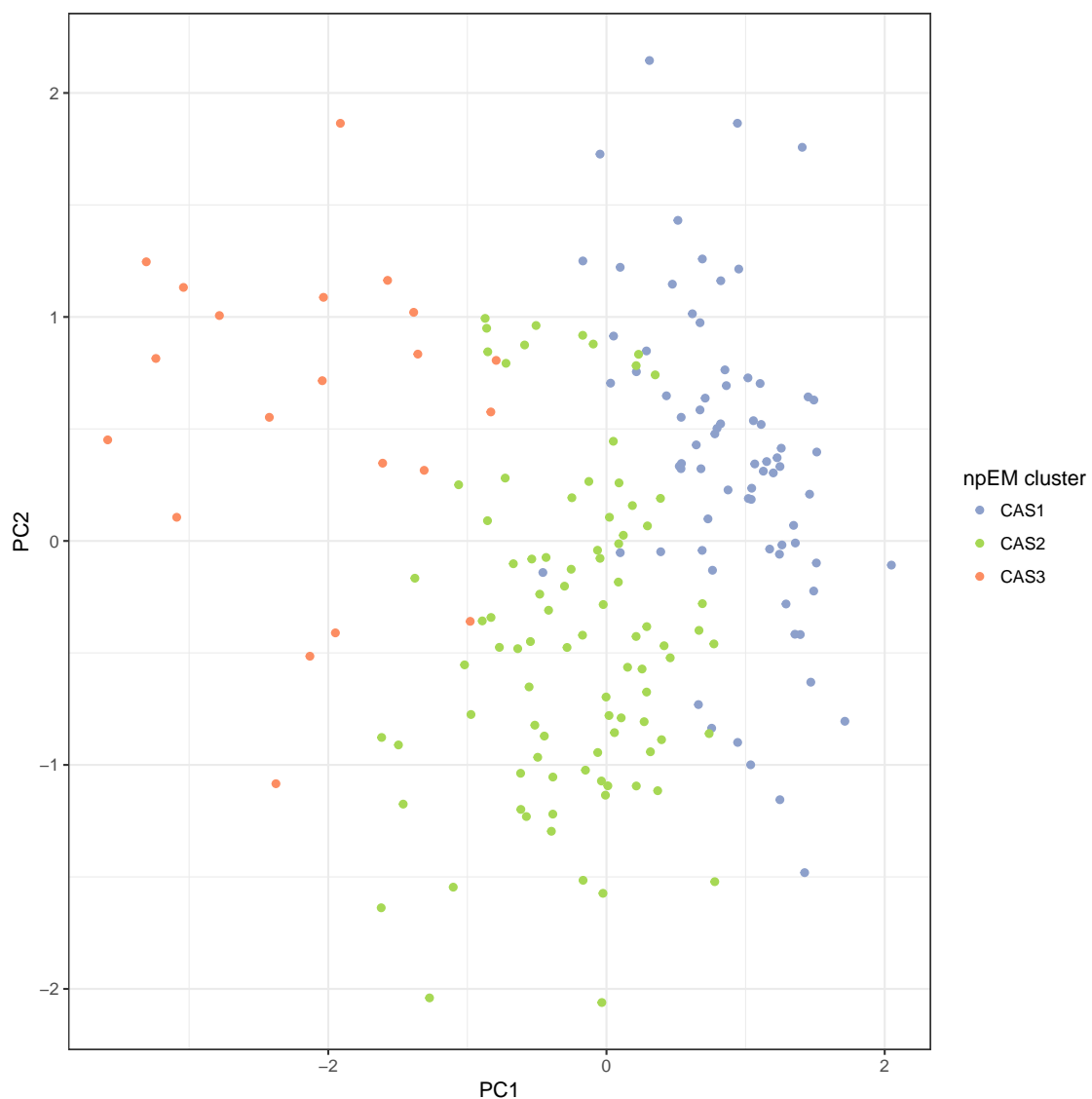
- Global Initiative for Asthma.** 2015. *Global Strategy for Asthma Management and Prevention*: ERS Publication.
- Gustafsson D, Sjöberg O, Foucard T.** 2000. Development of allergies and asthma in infants and young children with atopic dermatitis—a prospective follow-up to 7 years of age. *Allergy* **55**:240–245. DOI: <https://doi.org/10.1034/j.1398-9995.2000.00391.x>, PMID: 10753014
- Han H, Roan F, Ziegler SF.** 2017. The atopic march: current insights into skin barrier dysfunction and epithelial cell-derived cytokines. *Immunological Reviews* **278**:116–130. DOI: <https://doi.org/10.1111/imr.12546>, PMID: 28658558
- Hekking PP, Bel EH.** 2014. Developing and emerging clinical asthma phenotypes. *The Journal of Allergy and Clinical Immunology: In Practice* **2**:671–680. DOI: <https://doi.org/10.1016/j.jaip.2014.09.007>, PMID: 25439356
- Hollams EM, Deverell M, Serralha M, Suriyaarachchi D, Parsons F, Zhang G, de Klerk N, Holt BJ, Ladyman C, Sadowska A, Rowe J, Loh R, Sly PD, Holt PG.** 2009. Elucidation of asthma phenotypes in atopic teenagers through parallel immunophenotypic and clinical profiling. *Journal of Allergy and Clinical Immunology* **124**:463–470. DOI: <https://doi.org/10.1016/j.jaci.2009.06.019>, PMID: 19733295
- Hollams EM, Teo SM, Kusel M, Holt BJ, Holt KE, Inouye M, De Klerk NH, Zhang G, Sly PD, Hart PH, Holt PG.** 2017. Vitamin D over the first decade and susceptibility to childhood allergy and asthma. *Journal of Allergy and Clinical Immunology* **139**:472–481. DOI: <https://doi.org/10.1016/j.jaci.2016.07.032>, PMID: 27726947
- Holt PG, Rowe J, Kusel M, Parsons F, Hollams EM, Bosco A, McKenna K, Subrata L, de Klerk N, Serralha M, Holt BJ, Zhang G, Loh R, Ahlstedt S, Sly PD.** 2010. Toward improved prediction of risk for atopy and asthma among preschoolers: a prospective cohort study. *Journal of Allergy and Clinical Immunology* **125**:653–659. DOI: <https://doi.org/10.1016/j.jaci.2009.12.018>, PMID: 20226300
- Holt PG, Sly PD.** 2012. Viral infections and atopy in asthma pathogenesis: new rationales for asthma prevention and treatment. *Nature medicine* **18**:726–735. DOI: <https://doi.org/10.1038/nm.2768>, PMID: 22561836
- Holt PG, Strickland D, Bosco A, Belgrave D, Hales B, Simpson A, Hollams E, Holt B, Kusel M, Ahlstedt S, Sly PD, Custovic A.** 2016. Distinguishing benign from pathologic TH2 immunity in atopic children. *Journal of Allergy and Clinical Immunology* **137**:379–387. DOI: <https://doi.org/10.1016/j.jaci.2015.08.044>, PMID: 26518094
- Hose AJ, Depner M, Illi S, Lau S, Keil T, Wahn U, Fuchs O, Pfefferle PI, Schmauß-Hechfellner E, Genuneit J, Lauener R, Karvonen AM, Roduit C, Dalphin JC, Riedler J, Pekkanen J, von Mutius E, Ege MJ, MASPASTURE study groups.** 2017. Latent class analysis reveals clinically relevant atopy phenotypes in 2 birth cohorts. *Journal of Allergy and Clinical Immunology* **139**:1935–1945. DOI: <https://doi.org/10.1016/j.jaci.2016.08.046>, PMID: 27771325
- Janssens T, Verleden G, Van den Bergh O, Symptons VdenBO.** 2012. Symptoms, lung function, and perception of asthma control: an exploration into the heterogeneity of the asthma control construct. *The Journal of asthma : official journal of the Association for the Care of Asthma* **49**:63–69. DOI: <https://doi.org/10.3109/02770903.2011.636853>, PMID: 22121947
- Kim HY, Shin YH, Han MY.** 2014. Determinants of sensitization to allergen in infants and young children. *Korean Journal of Pediatrics* **57**:205–210. DOI: <https://doi.org/10.3345/kjp.2014.57.5.205>, PMID: 25045361
- Klink M, Cline MG, Halonen M, Burrows B.** 1990. Problems in defining normal limits for serum IgE. *Journal of Allergy and Clinical Immunology* **85**:440–444. DOI: [https://doi.org/10.1016/0091-6749\(90\)90153-U](https://doi.org/10.1016/0091-6749(90)90153-U), PMID: 2303647
- Koplin JJ, Osborne NJ, Wake M, Martin PE, Gurrin LC, Robinson MN, Tey D, Slaa M, Thiele L, Miles L, Anderson D, Tan T, Dang TD, Hill DJ, Lowe AJ, Matheson MC, Ponsonby AL, Tang ML, Dharmage SC, Allen KJ.** 2010. Can early introduction of egg prevent egg allergy in infants? A population-based study. *Journal of Allergy and Clinical Immunology* **126**:807–813. DOI: <https://doi.org/10.1016/j.jaci.2010.07.028>, PMID: 20920771
- Kurukulaaratchy RJ, Fenn MH, Waterhouse LM, Matthews SM, Holgate ST, Arshad SH.** 2003. Characterization of wheezing phenotypes in the first 10 years of life. *Clinical & Experimental Allergy* **33**:573–578. DOI: <https://doi.org/10.1046/j.1365-2222.2003.01657.x>, PMID: 12752584
- Kusel MM, Holt PG, de Klerk N, Sly PD.** 2005. Support for 2 variants of eczema. *Journal of Allergy and Clinical Immunology* **116**:1067–1072. DOI: <https://doi.org/10.1016/j.jaci.2005.06.038>, PMID: 16275378
- Kusel MM, de Klerk NH, Kebabdzte T, Vohma V, Holt PG, Johnston SL, Sly PD.** 2007. Early-life respiratory viral infections, atopic sensitization, and risk of subsequent development of persistent asthma. *Journal of Allergy and Clinical Immunology* **119**:1105–1110. DOI: <https://doi.org/10.1016/j.jaci.2006.12.669>, PMID: 17353039
- Lazic N, Roberts G, Custovic A, Belgrave D, Bishop CM, Winn J, Curtin JA, Hasan Arshad S, Simpson A.** 2013. Multiple atopy phenotypes and their associations with asthma: similar findings from two birth cohorts. *Allergy* **68**:764–770. DOI: <https://doi.org/10.1111/all.12134>, PMID: 23621120
- Lemanske RF.** 2002. The childhood origins of asthma (COAST) study. *Pediatric Allergy and Immunology* **13**:38–43. DOI: <https://doi.org/10.1034/j.1399-3038.13.s.15.8.x>, PMID: 12688623
- Linden CC, Misiak RT, Wegienka G, Havstad S, Ownby DR, Johnson CC, Zoratti EM.** 2011. Analysis of allergen specific IgE cut points to cat and dog in the Childhood Allergy Study. *Annals of Allergy, Asthma & Immunology* **106**:153–158. DOI: <https://doi.org/10.1016/j.anai.2010.11.004>
- Martinez FD, Wright AL, Taussig LM, Holberg CJ, Halonen M, Morgan WJ.** 1995. Asthma and wheezing in the first six years of life: the group health medical associates. *The New England Journal of Medicine* **332**:133–138. DOI: <https://doi.org/10.1056/NEJM199501193320301>, PMID: 7800004
- Morgan WJ, Stern DA, Sherrill DL, Guerra S, Holberg CJ, Guilbert TW, Taussig LM, Wright AL, Martinez FD.** 2005. Outcome of asthma and wheezing in the first 6 years of life: follow-up through adolescence. *American Journal of Respiratory and Critical Care Medicine* **172**:1253–1258. DOI: <https://doi.org/10.1164/rccm.200504-525OC>, PMID: 16109980

- NAC Manchester Asthma and Allergy Study Group**, Custovic A, Simpson BM, Murray CS, Lowe L, Woodcock A. 2002. The national asthma campaign manchester asthma and allergy study. *Pediatric Allergy and Immunology* **13**:32–37. DOI: <https://doi.org/10.1034/j.1399-3038.13.s.15.3.x>, PMID: 12688622
- Newby C**, Heaney LG, Menzies-Gow A, Niven RM, Mansur A, Bucknall C, Chaudhuri R, Thompson J, Burton P, Brightling C, British Thoracic Society Severe Refractory Asthma Network. 2014. Statistical cluster analysis of the british thoracic society severe refractory asthma registry: clinical outcomes and phenotype stability. *PLoS ONE* **9**:e102987. DOI: <https://doi.org/10.1371/journal.pone.0102987>, PMID: 25058007
- Ober C**, Yao TC. 2011. The genetics of asthma and allergic disease: a 21st century perspective. *Immunological reviews* **242**:10–30. DOI: <https://doi.org/10.1111/j.1600-065X.2011.01029.x>, PMID: 21682736
- Okada H**, Kuhn C, Feillet H, Bach JF. 2010. The 'hygiene hypothesis' for autoimmune and allergic diseases: an update. *Clinical & Experimental Immunology* **160**:1–9. DOI: <https://doi.org/10.1111/j.1365-2249.2010.04139.x>, PMID: 20415844
- Okamoto S**, Taniuchi S, Sudo K, Hatano Y, Nakano K, Shimo T, Kaneko K. 2012. Predictive value of IgE/IgG4 antibody ratio in children with egg allergy. *Allergy, Asthma & Clinical Immunology* **8**:9. DOI: <https://doi.org/10.1186/1710-1492-8-9>
- Prescott SL**, Macaubas C, Smallacombe T, Holt BJ, Sly PD, Holt PG. 1999. Development of allergen-specific T-cell memory in atopic and normal children. *The Lancet* **353**:196–200. DOI: [https://doi.org/10.1016/S0140-6736\(98\)05104-6](https://doi.org/10.1016/S0140-6736(98)05104-6), PMID: 9923875
- Rousseeuw PJ**. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**:53–65. DOI: [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- Schoos AM**, Chawes BL, Melén E, Bergström A, Kull I, Wickman M, Bønnelykke K, Bisgaard H, Rasmussen MA. 2017. Sensitization trajectories in childhood revealed by using a cluster analysis. *Journal of Allergy and Clinical Immunology* **140**:1693–1699. DOI: <https://doi.org/10.1016/j.jaci.2017.01.041>, PMID: 28347735
- Simpson A**, Tan VY, Winn J, Svendsén M, Bishop CM, Heckerman DE, Buchan I, Custovic A. 2010. Beyond atopy: multiple patterns of sensitization in relation to asthma in a birth cohort study. *American Journal of Respiratory and Critical Care Medicine* **181**:1200–1206. DOI: <https://doi.org/10.1164/rccm.200907-11101OC>, PMID: 20167852
- Spycher BD**, Silverman M, Kuehni CE. 2010. Phenotypes of childhood asthma: are they real? *Clinical & Experimental Allergy* **40**:1130–1141. DOI: <https://doi.org/10.1111/j.1365-2222.2010.03541.x>, PMID: 20545704
- Tan P-N**, Kumar V, Steinbach M. 2005. *Introduction to Data Mining*. Boston: Pearson Addison Wesley.
- Teo SM**, Mok D, Pham K, Kusel M, Serralha M, Troy N, Holt BJ, Hales BJ, Walker ML, Hollams E, Bochkov YA, Grindle K, Johnston SL, Gern JE, Sly PD, Holt PG, Holt KE, Inouye M. 2015. The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development. *Cell Host & Microbe* **17**:704–715. DOI: <https://doi.org/10.1016/j.chom.2015.03.008>, PMID: 25865368
- Teo SM**, Tang HH, Mok D, Judd LM, Watts SC, Pham K. 2017. Dynamics of the upper airway microbiome in the pathogenesis of asthma-associated persistent wheeze in preschool children. *bioRxiv*. DOI: <https://doi.org/10.1101/222190>
- Therneau TM**, Atkinson EJ. 2015. rpart: Recursive Partitioning and Regression Trees. <https://cran.r-project.org/web/packages/rpart/>
- Thomas WR**, Hales BJ, Smith WA. 2010. House dust mite allergens in asthma and allergy. *Trends in Molecular Medicine* **16**:321–328. DOI: <https://doi.org/10.1016/j.molmed.2010.04.008>, PMID: 20605742
- Vadas P**, Wai Y, Burks W, Perelman B. 2001. Detection of peanut allergens in breast milk of lactating women. *Jama* **285**:1746–1748. DOI: <https://doi.org/10.1001/jama.285.13.1746>, PMID: 11277829
- Wenzel SE**. 2012. Asthma phenotypes: the evolution from clinical to molecular approaches. *Nature Medicine* **18**:716–725. DOI: <https://doi.org/10.1038/nm.2678>, PMID: 22561835
- Wu J**, Prospero MC, Simpson A, Hollams EM, Sly PD, Custovic A, Holt PG. 2015. Relationship between cytokine expression patterns and clinical outcomes: two population-based birth cohorts. *Clinical & Experimental Allergy* **45**:1801–1811. DOI: <https://doi.org/10.1111/cea.12579>, PMID: 26061524



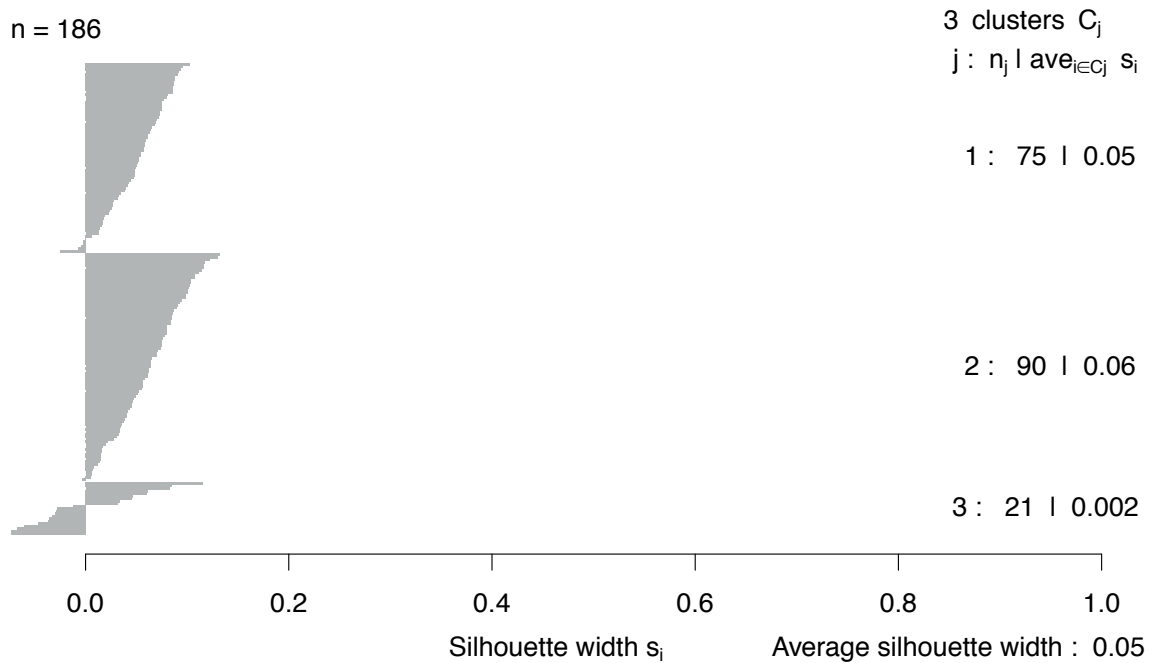
## Appendix B

# Supplementary Figures and Tables for Chapter 3



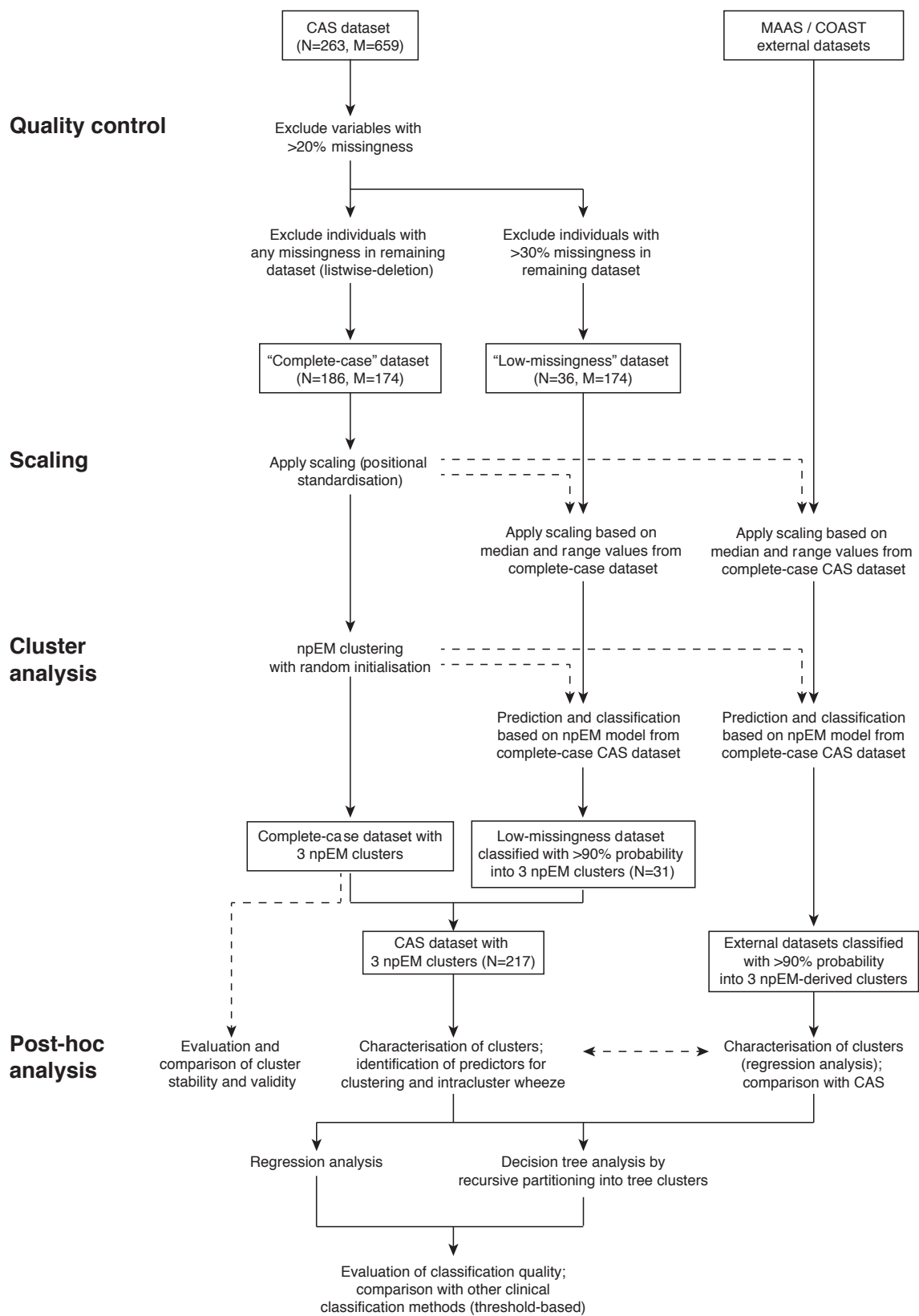
**FIGURE B.1:** Scatterplot of principal components analysis (PCA) of the complete-case CAS dataset ( $N = 186$ , with points coloured by npEM clusters)

Each point represents an individual. The first two PCs (shown) account for 16.7% of the total variance.



**FIGURE B.2: Silhouette widths of clusters generated by npEM.**

$j$  = cluster number;  $n_j$  = cluster size;  $\text{ave}_{i \in C_j} s_i$  = average silhouette width among members  $i$  of cluster  $C_j$ . Overall average silhouette width across all clusters is also given.



**FIGURE B.3: Overview of study methodology.**

Dashed arrows indicate non-critical elements of our method.



**FIGURE B.4: Relationship of clusters to food sensitisation, eczema and wheeze.**

Percentages denote proportion of cluster displaying phenotype (numbers in brackets denote actual sample numbers). Food sensitization defined as peanut IgE  $\geq 0.35$  kU/L at any age, or cow's milk, egg white, peanut SPT  $> 2$  or  $3$  mm for age  $\leq 2$  or  $> 2$  respectively. Subphenotypes defined for food sensitization, eczema and wheeze as: no phenotype = phenotype absent at all ages; transient = any incidence of phenotype at the earlier ages (1 to 3 for wheeze, 6m to 3 for eczema, 6m to 2 for sensitization), but not age 5; late = phenotype at age 5, but not the earlier ages; persistent = any incidence of phenotype at both earlier ages and age 5.

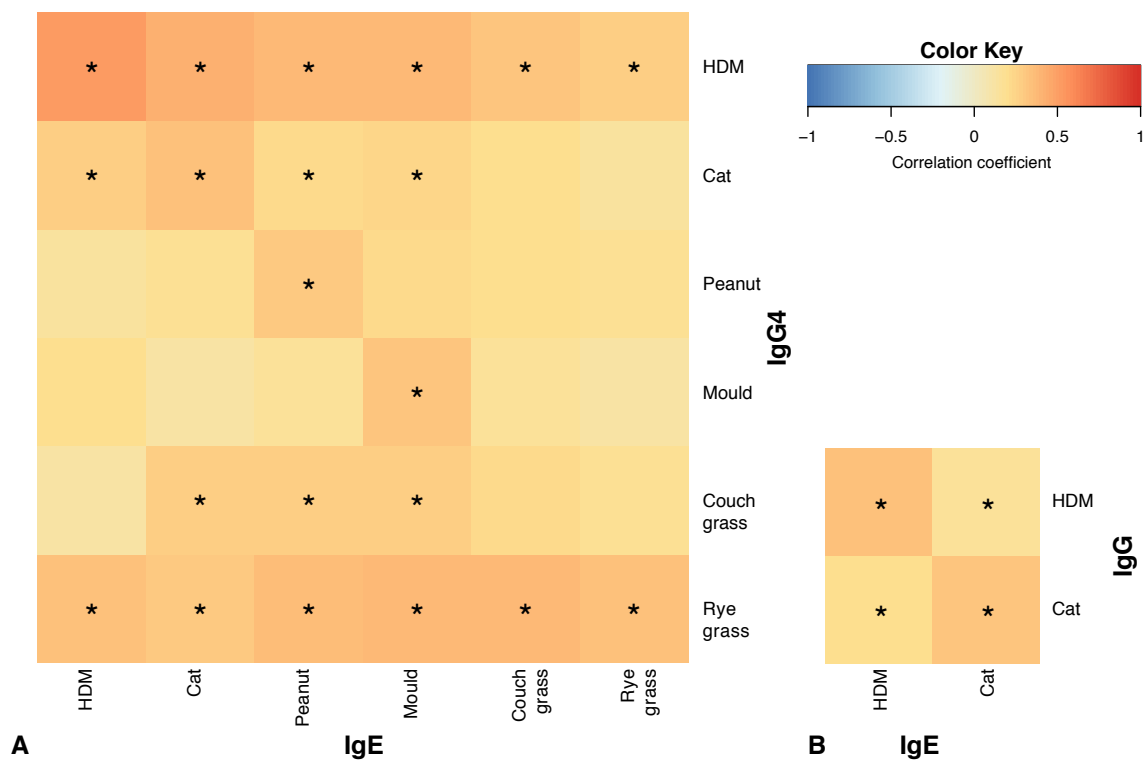
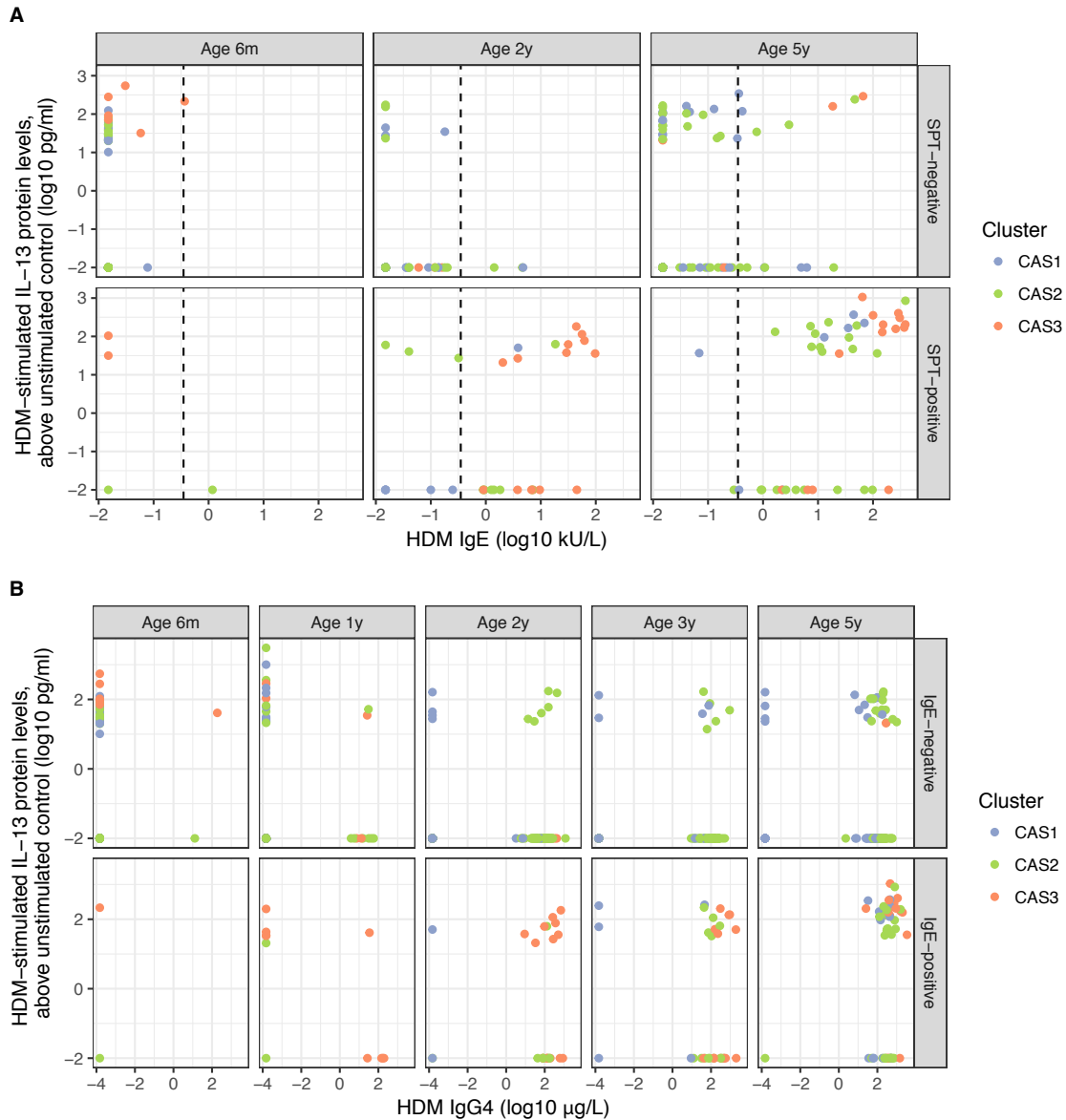


FIGURE B.5: Correlation patterns between IgE vs. IgG4 (A) and IgE vs. IgG (B) at age five.

\*  $p < 0.05$  for Spearman correlation with Holm correction for multiple testing. Note the slightly stronger heat along the main diagonals of both heatmaps, especially for HDM.

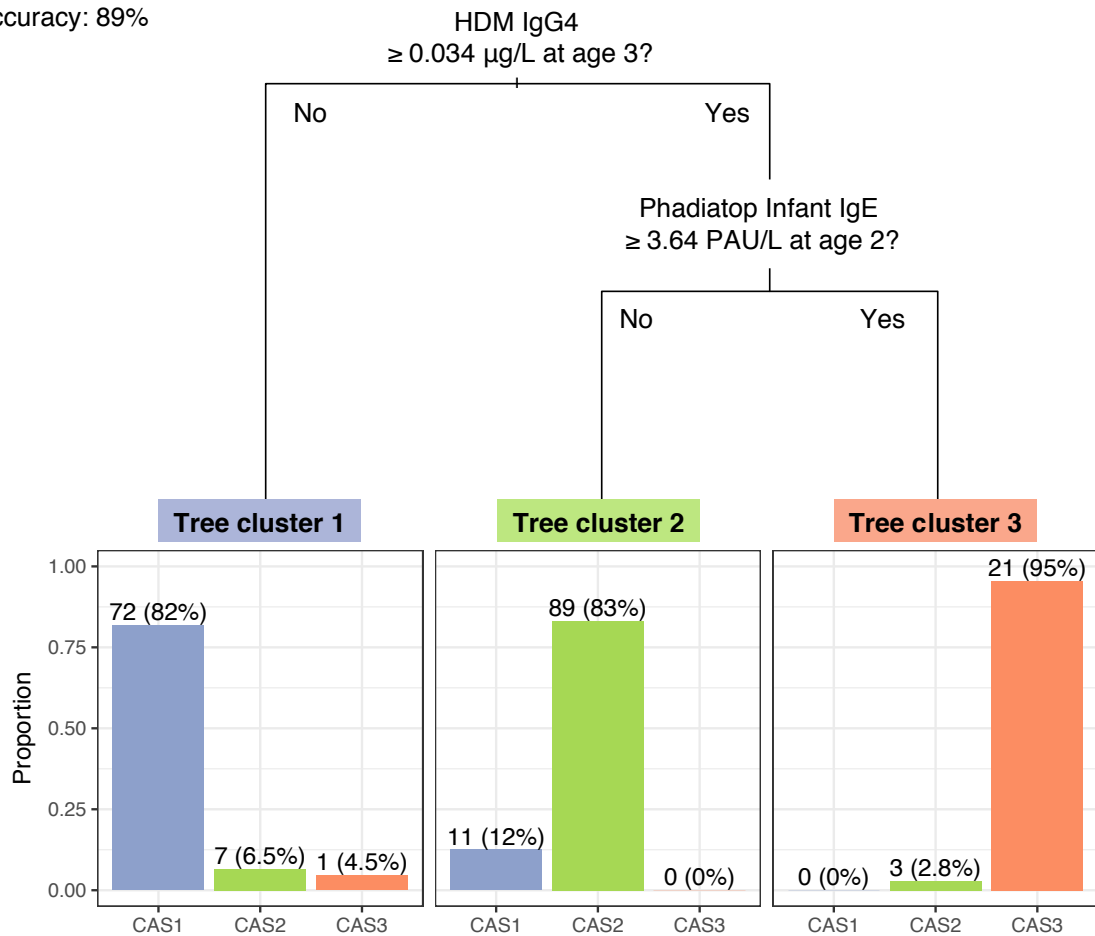


**FIGURE B.6: Distinct biological signals of HDM IgE, IgG4, SPT, and Th2 cytokine (IL-13).**

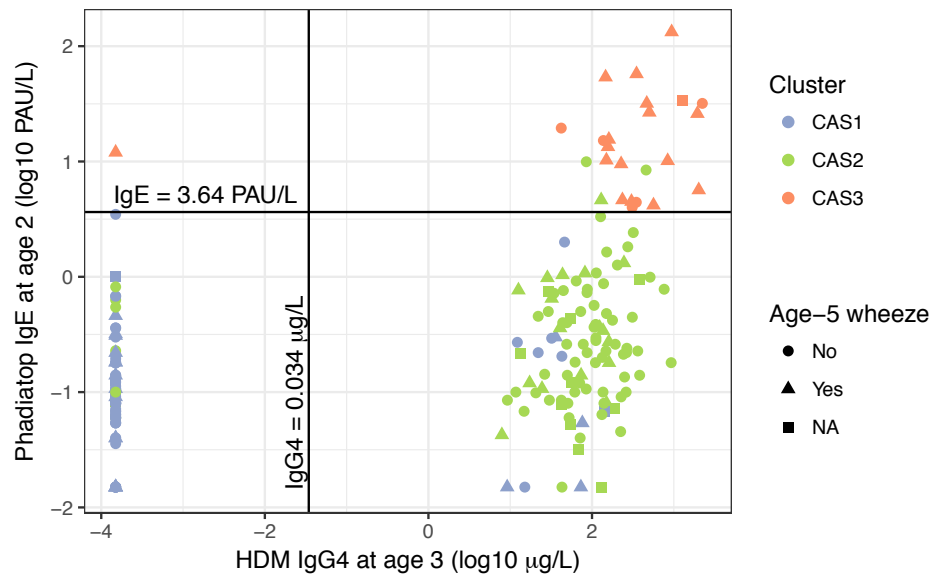
(A) Th2 cytokine (IL-13) to HDM stimulation of PBMCs *in vitro*, vs. HDM IgE responses *in vivo*, stratified by age of testing, and SPT result (positive denoted by  $\geq 2$ mm at age <2 or 3mm at age 5). Dotted line represents traditional threshold for HDM IgE positive result (0.35 kU/L). Note the significant number of individuals on either side of the dotted line for both HDM SPT-positive and negative subgroups. (B) Th2 cytokine (IL-13) to HDM stimulation of PBMCs *in vitro*, vs. HDM IgG4 responses *in vivo*, stratified by age of testing and HDM IgE result (positive denoted by  $\geq 0.35$  kU/L).

**A**

Accuracy: 89%



**B**



**FIGURE B.7:** A “simple” decision tree generated by recursive partitioning from CAS data, with breakdown of tree clusters by actual CAS npEM-derived clusters (A); scatterplot showing separation of CAS clusters by decision split thresholds (B).

Percentages in Panel A may not sum up to 100%, because some individuals have missing values for decision node variables, hence making them impossible to classify. In Panel B, note that left-most column of points represent values of HDM IgG4 that were less than the limit-of-detection (LOD) for that assay ( $0.0003 \mu\text{g/L}$ ), and were subsequently assigned to half the LOD ( $0.00015 \mu\text{g/L}$ ). Most of these points belonged to individuals from CAS1.

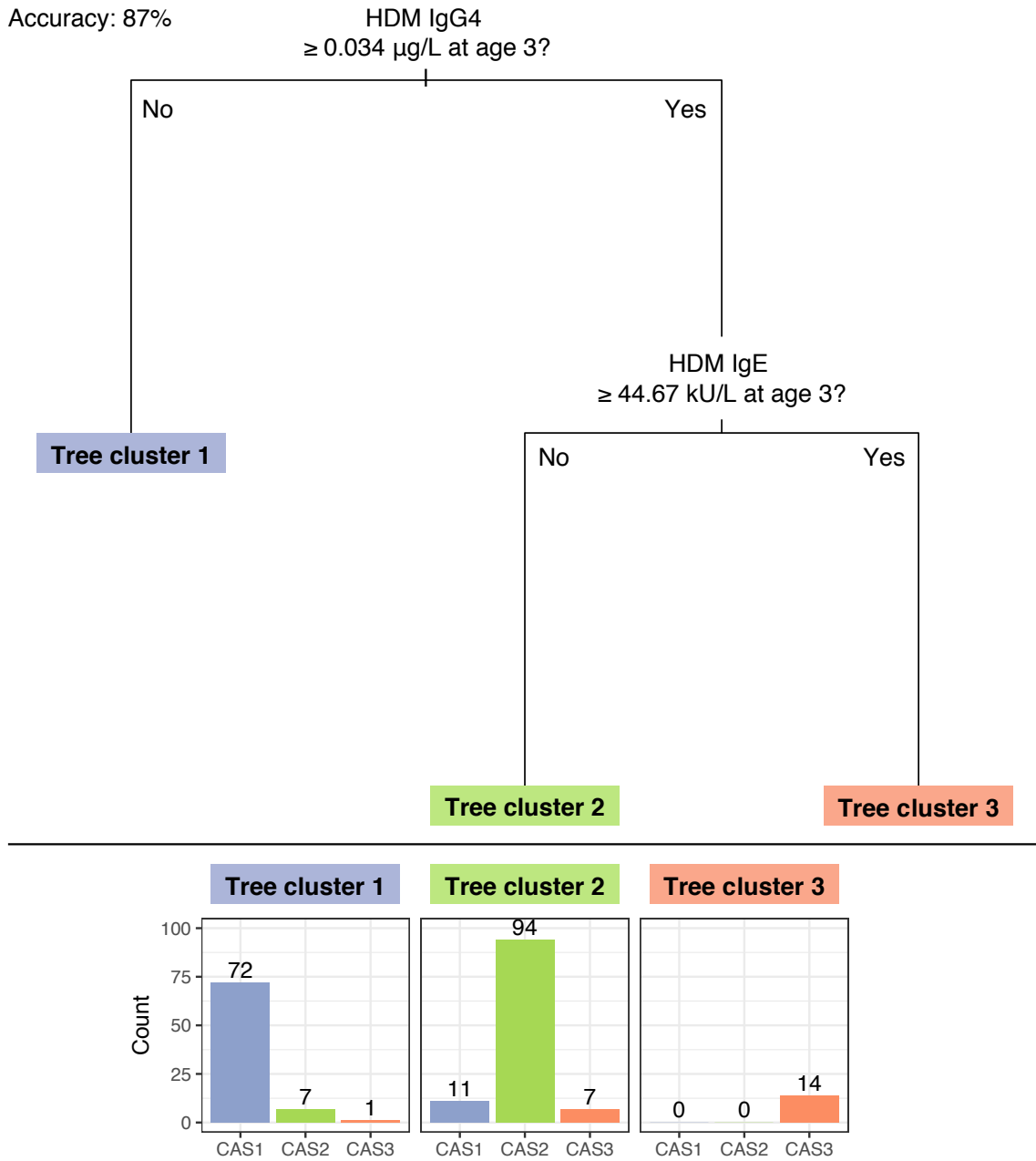


FIGURE B.8: Decision tree generated by recursive partitioning from CAS data, excluding Phadiatop assay variables.



Accuracy: 77%

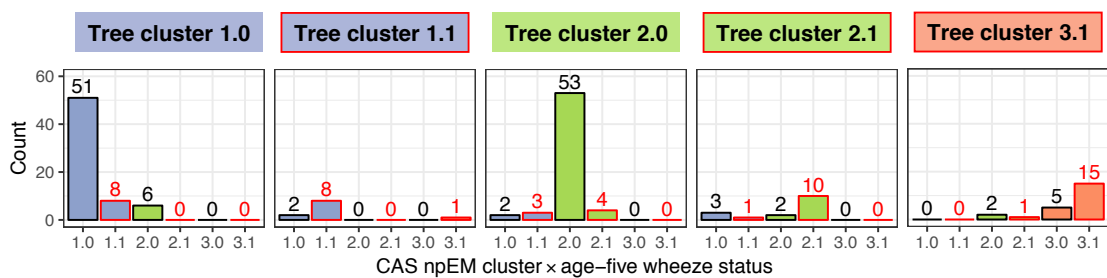
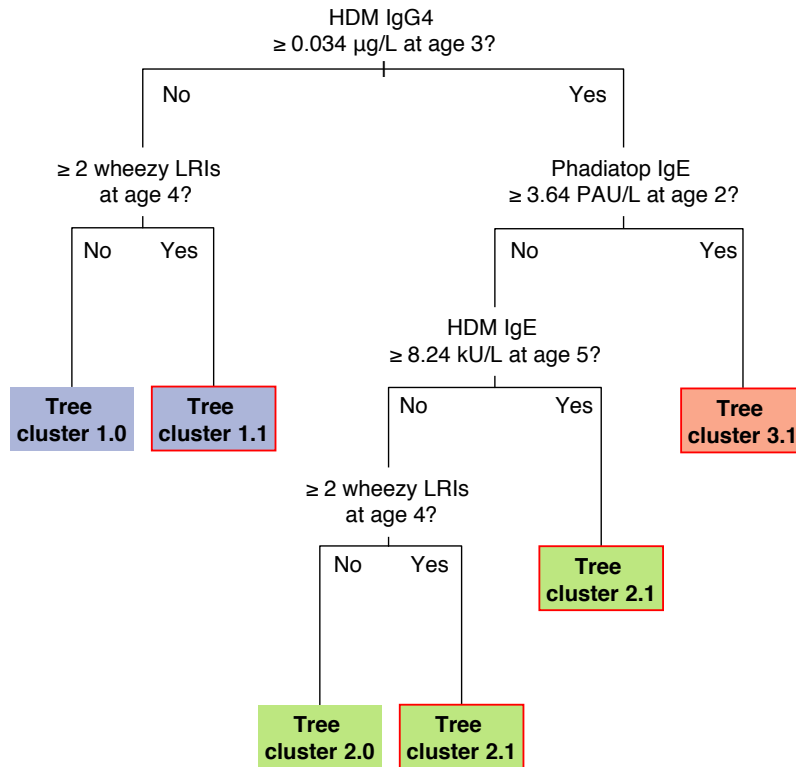
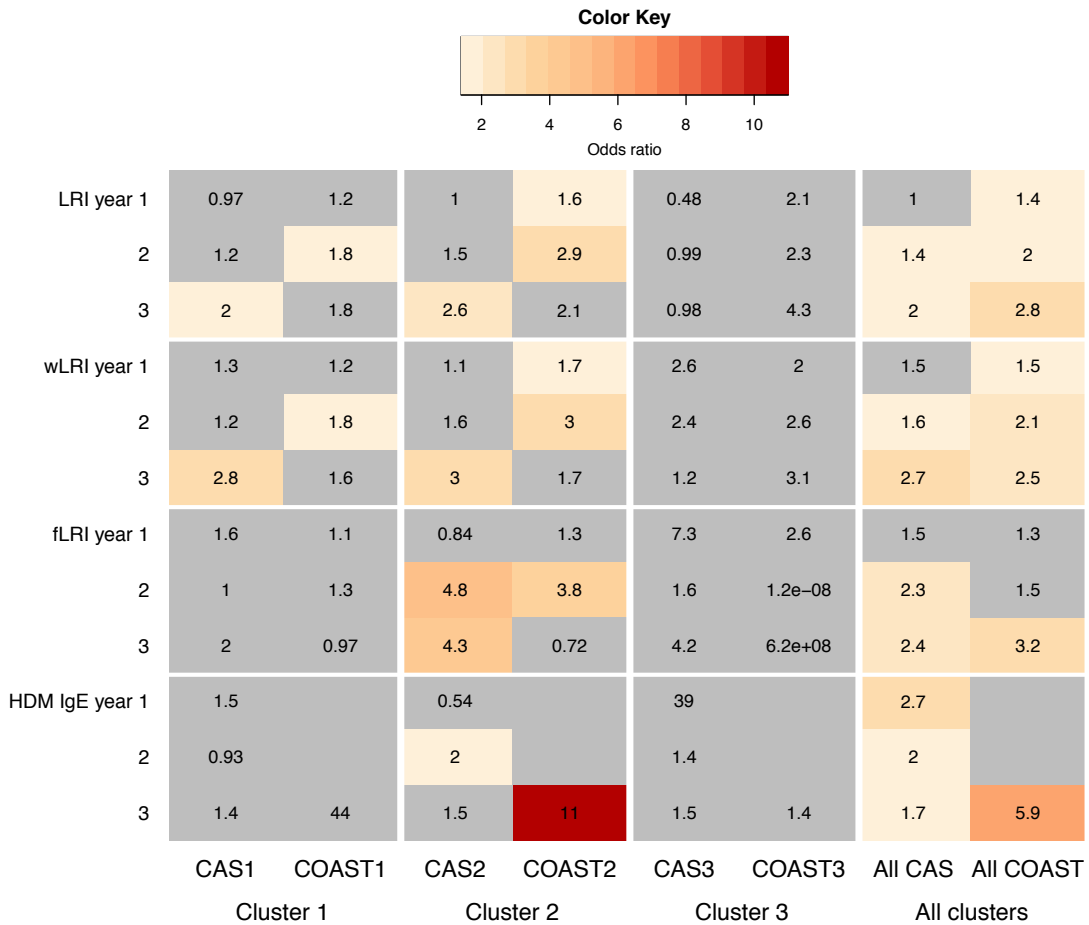


FIGURE B.9: A “comprehensive” decision tree generated by recursive partitioning from CAS data, given CAS npEM-derived clusters and age-five wheezing status.

CAS1.x, 2.x, and 3.x, and tree cluster 1.x, 2.x and 3.x, refer to the intersection of npEM-generated clusters and age-five wheeze status, and their analogous decision tree cluster, respectively. The second digit (x.0 or x.1) refers to age-five wheeze status, with “1” = present wheeze and “0” = no wheeze. Boxes, bars and digits with red outline indicate those with predicted (tree cluster) or actual (CAS npEM cluster) age-five wheeze. Note that the tree did not predict for non-wheezing CAS3, so there is no tree cluster 3.0, and all CAS3 individuals were automatically assigned to a wheezing tree cluster.



**FIGURE B.10: Comparison of predictors for age-five wheeze in CAS and COAST clusters.**

Colour coding and numbers in cells indicate odds ratio (OR) of predictor for age-five wheeze in GLM, with sex, maternal and paternal history of asthma, and (for CAS) BMI as covariates. Non-grey cell with number indicates statistically-significant association ( $p < 0.05$ ). Grey cell with number indicates non-significant ( $p > 0.05$ ); grey non-numbered cell indicates test not done due to lack of data.

TABLE B.1: List of clustering features

\*Used as surrogate measure for values at different timepoints. Age 5 values substituted for age 3; age 3 for age 2; age 1 for age 6m, as indicated. ^Mixed grass SPT used as surrogate for ryegrass SPT. Blue and Y indicates feature present in dataset; yellow and Y indicates present, but only for some time-points (as indicated by numbers in brackets); red and N indicates feature absent from dataset.

Feature	Present?		
	CAS	MAAS	COAST
Sex	Y	Y	Y
<i>Respiratory infection-related variables</i>			
Frequency of upper respiratory illnesses (URIs) at ages 1, 2, 3	Y	Y (1,3)	Y
Frequency of lower respiratory illnesses (LRIs) at ages 1, 2, 3	Y	N	Y
Frequency of wheezy LRIs (wLRIs) at ages 1, 2, 3	Y	Y (1,3)	Y
Frequency of febrile LRIs (fLRIs) at ages 1, 2, 3	Y	N	Y
Number of URIs, LRIs, wLRIs, and fLRIs with respiratory syncytial virus (RSV) detected, at ages 1, 2, 3	Y	N	Y
Number of URIs, LRIs, wLRIs, and fLRIs with influenza detected, at ages 1, 2, 3	Y	N	Y
Number of LRIs, wLRIs, and fLRIs with human rhinovirus A (HRVA) detected, at ages 1, 2, 3	Y	N	Y
Number of LRIs, wLRIs, and fLRIs with human rhinovirus B (HRVB) detected, at ages 1, 2, 3	Y	N	Y
Number of LRIs, wLRIs, and fLRIs with human rhinovirus C (HRVC) detected, at ages 1, 2, 3	Y	N	Y
<i>IgE variables</i>			
Total IgE (kU/L), with log10 transformation, at age 6m, 1, 2 and 3	Y	Y (1,3)	Y (1,2,3)
House dust mite (HDM)-specific IgE (kU/L), with log10 transformation, at age 6m, 1, 2 and 3	Y	Y (1,3)	Y (1,2,3)
Cat-specific IgE (kU/L), with log10 transformation, at ages 6m, 1, 2 and 3	Y	Y (1,3)	Y (1,2,3)
Peanut-specific IgE (kU/L), with log10 transformation, at ages 6m, 1, 2 and 3	Y	N	Y (1,2,3)
Couch grass-specific IgE (kU/L), with log10 transformation, at ages 6m, 1, 2 and 3	Y	N	N
Ryegrass-specific IgE (kU/L), with log10 transformation, at ages 6m, 1, 2 and 3	Y	N	N
Mould-specific IgE (kU/L), with log10 transformation, at ages 6m, 1, 2 and 3	Y	N	N
Phadiatop Infant IgE (kU/L), with log10 transformation, at ages 6m, 1, 2 and 3	Y	N	N
<i>IgG4 variables</i>			
HDM-specific IgG4 ( $\mu\text{g/L}$ ), with log10 transformation, at ages 6m, 1, 2 and 3	Y	Y (5*)	N
Cat-specific IgG4 ( $\mu\text{g/L}$ ), with log10 transformation, at ages 6m, 1, 2 and 3	Y	Y (5*)	N
Peanut-specific IgG4 ( $\mu\text{g/L}$ ), with log10 transformation, at ages 6m, 1, 2 and 3	Y	N	N
Couch grass-specific IgG4 ( $\mu\text{g/L}$ ), with log10 transformation, at ages 6m, 1, 2 and 3	Y	N	N
Ryegrass-specific IgG4 ( $\mu\text{g/L}$ ), with log10 transformation, at ages 6m, 1, 2 and 3	Y	N	N
Mould-specific IgG4 ( $\mu\text{g/L}$ ), with log10 transformation, at ages 6m, 1, 2 and 3	Y	N	N
Phadiatop Infant IgG4 ( $\mu\text{g/L}$ ), with log10 transformation, at ages 6m, 1, 2 and 3	Y	N	N
<i>IgG variables</i>			
HDM-specific IgG (mg/L), with log10 transformation, at ages 1,2, and 3	Y	Y (5*)	N
Cat-specific IgG (mg/L), with log10 transformation, at ages 1,2, and 3	Y	Y (5*)	N
(Timothy) Grass-specific IgG (mg/L), with log10 transformation, at ages 1,2, and 3	Y	N	N
<i>SPT variables</i>			
Histamine-specific skin sensitisation or skin prick test (SPT) response, diameter of wheal (mm) at ages 6m, 2	Y	N	N
HDM-specific SPT response, diameter of wheal (mm) at ages 6m, 2	Y	Y (1*,3*)	N
Cat-specific SPT response, diameter of wheal (mm) at ages 6m, 2	Y	Y (1*,3*)	N
Ryegrass-specific SPT response, diameter of wheal (mm) at ages 6m, 2	Y	Y (1*,3*)^	N
<i>Alternaria</i> -specific SPT response, diameter of wheal (mm) at ages 6m, 2	Y	N	N
<i>Aspergillus</i> -specific SPT response, diameter of wheal (mm) at ages 6m, 2	Y	Y (3*)	N
Cow's milk-specific SPT response, diameter of wheal (mm) at ages 6m, 2	Y	Y (1*,3*)	N
Egg white-specific SPT response, diameter of wheal (mm) at ages 6m, 2	Y	Y (1*,3*)	N
<i>Family history of asthma or atopy</i>			
Maternal and paternal history of atopy as determined by SPT $\geq$ 3mm to any allergen	Y	Y	Y
Maternal and paternal history of physician-diagnosed asthma	Y	Y	Y
Maternal and paternal history of physician-diagnosed atopic disease besides asthma (eczema, hayfever)	Y	Y	Y

Continued on next page

*Continued from previous page*

Feature	Present?		
	CAS	MAAS	COAST
<i>Other</i>			
Child ever exposed to cigarette smoke at ages 1, 2 and 3	Y	Y (1,3)	Y
Child ever attended childcare at ages 1, 2, and 3	Y	Y (2,3)	Y
Child ever exposed to cat at ages 1, 2 and 3	Y	Y	Y
Number of children (age<16) living in the same household at ages 1, 2, and 3	Y	Y	Y
Number of children older than the proband, living in the same household at ages 0, 1, 2, and 3	Y	Y	Y
Height (cm) at age 3	Y	Y	Y
Weight (kg) at age 3	Y	Y	Y
25-Hydroxy Vitamin D (nmol/L) in mothers's serum at 6 weeks postpartum	Y	N	N
Physician-diagnosed eczema or atopic dermatitis at ages 6m, 1, 2, and 3	Y	Y (1,3)	Y

**TABLE B.2: Terminology used to describe groupings produced by various clustering and classification methods on different datasets**

CAS = Childhood Asthma Study, Perth, Australia; COAST = Childhood Origins of Asthma Study, Wisconsin, US; MAAS = Manchester Asthma and Allergy Study, Manchester, UK; npEM = non-parametric expectation-maximisation mixture modelling.

Dataset	Method	Terminology		
		Cluster 1	Cluster 2	Cluster 3
CAS (discovery)	npEM clustering for complete-case subset; npEM-based classification for low-missingness subset	CAS1	CAS2	CAS3
MAAS (replication)	CAS npEM-based classification of MAAS, using only features common to MAAS and CAS	MAAS1	MAAS2	MAAS3
COAST (replication)	CAS npEM-based classification of COAST, using only features common to COAST and CAS	COAST1	COAST2	COAST3

**TABLE B.3: Comparison of selected demographic and clinical variables  
in CAS clusters**

See next pages.

## A. Clinical and demographic

Variable	Age	CAS1 (N=88)		CAS2 (N=107)		CAS3 (N=22)		P-value		P-value (adj.)		Feature?	
		Prop. (95% CI)	Prop. (95% CI)	Prop. (95% CI)	Prop. (95% CI)	Overall	1 vs. 2	1 vs. 3	2 vs. 3	Overall	1 vs. 2		1 vs. 3
Sex													
Male		55% (44%-65%)	51% (42%-61%)	86% (71%-100%)	7.3E-03	0.67	6.8E-03	3.7E-03	0.15	1	0.23	0.14	Yes
<i>Maternal medical history</i>													
Atopy		74% (65%-83%)	84% (77%-91%)	82% (64%-99%)	0.2	0.11	0.58	0.76	1	1	1	1	No
Atopic eczema		83% (75%-91%)	81% (74%-89%)	95% (86%-100%)	0.28	0.85	0.19	0.12	1	1	1	1	No
Asthma		51% (40%-62%)	41% (32%-51%)	59% (37%-81%)	0.19	0.19	0.63	0.16	1	1	1	1	Yes
Atopic disease besides asthma		72% (62%-81%)	74% (65%-82%)	77% (58%-96%)	0.88	0.75	0.79	1	1	1	1	1	Yes
<i>Paternal medical history</i>													
Atopy		74% (65%-83%)	85% (78%-92%)	91% (78%-100%)	0.075	0.072	0.15	0.74	1	1	1	1	No
Atopic eczema		58% (47%-68%)	77% (69%-85%)	59% (37%-81%)	0.01	5.0E-03	1	0.11	0.2	0.18	1	1	No
Asthma		22% (13%-30%)	44% (35%-54%)	23% (3.7%-42%)	2.2E-03	1.3E-03	1	0.093	0.051	0.055	1	1	Yes
Atopic disease besides asthma		48% (37%-58%)	54% (44%-63%)	41% (19%-63%)	0.49	0.47	0.64	0.35	1	1	1	1	Yes
<i>Atopy-associated diseases</i>													
Wheeze	1	33% (23%-43%)	30% (21%-39%)	55% (32%-77%)	0.092	0.76	0.084	0.046	1	1	1	1	No
	2	30% (20%-39%)	29% (20%-38%)	59% (37%-81%)	0.024	1	0.013	0.012	0.42	1	0.4	0.37	No
	3	30% (20%-40%)	23% (15%-31%)	55% (32%-77%)	0.015	0.32	0.044	7.6E-03	0.28	1	1	0.25	No
	4	23% (14%-32%)	20% (12%-28%)	77% (58%-96%)	9.3E-07	0.59	5.1E-06	5.2E-07	4.0E-05	1	3.9E-04	5.0E-05	No
	5	25% (15%-35%)	21% (13%-30%)	76% (56%-96%)	7.1E-06	0.59	2.6E-05	3.4E-06	2.6E-04	1	1.7E-03	2.8E-04	No
	10	12% (3.4%-21%)	18% (8.4%-27%)	50% (24%-76%)	3.1E-03	0.46	1.5E-03	0.011	0.069	1	0.063	0.34	No
Asthma	3	12% (4.8%-19%)	3.8% (0.087%-7.5%)	23% (3.7%-42%)	8.2E-03	0.049	0.3	7.9E-03	0.17	1	1	0.26	No
	4	11% (4.1%-18%)	8.2% (2.7%-14%)	36% (15%-58%)	4.2E-03	0.61	8.3E-03	2.0E-03	0.091	1	0.27	0.08	No
	5	15% (7%-23%)	13% (5.9%-20%)	52% (29%-76%)	4.1E-04	0.83	7.7E-04	2.1E-04	0.011	1	0.034	0.011	No
	10	10% (2.3%-18%)	15% (6.1%-23%)	56% (30%-81%)	2.6E-04	0.59	1.8E-04	7.9E-04	7.2E-03	1	9.6E-03	0.035	No
Allergic rhinoconjunctivitis	5	30% (20%-40%)	39% (29%-49%)	76% (56%-96%)	6.4E-04	0.21	2.7E-04	3.2E-03	0.017	1	0.014	0.12	No
Eczema	6m	39% (28%-49%)	45% (35%-54%)	91% (78%-100%)	2.4E-05	0.47	7.9E-06	9.0E-05	8.1E-04	1	5.8E-04	5.1E-03	Yes
	1	34% (24%-44%)	30% (21%-39%)	82% (64%-99%)	2.5E-05	0.54	7.2E-05	1.4E-05	8.4E-04	1	4.2E-03	9.6E-04	Yes
	2	30% (20%-39%)	31% (22%-40%)	68% (47%-89%)	2.7E-03	0.88	1.2E-03	1.5E-03	0.062	1	0.051	0.063	Yes
	3	27% (18%-37%)	25% (16%-33%)	59% (37%-81%)	7.1E-03	0.74	0.01	4.0E-03	0.15	1	0.32	0.15	Yes
	4	26% (16%-35%)	28% (19%-37%)	59% (37%-81%)	0.012	0.87	4.7E-03	0.011	0.23	1	0.17	0.34	No
	5	28% (18%-37%)	24% (16%-33%)	71% (50%-92%)	2.1E-04	0.73	3.3E-04	7.9E-05	6.1E-03	1	0.016	4.6E-03	No
<i>Exposure to tobacco smoke</i>													
Cigarette smoke exposure	1	20% (12%-29%)	20% (12%-27%)	0% (0%-0%)	0.041	1	0.021	0.023	0.65	1	0.6	0.64	Yes
	2	22% (13%-30%)	13% (6.6%-20%)	4.5% (0%-14%)	0.11	0.13	0.071	0.46	1	1	1	1	Yes
	3	19% (11%-28%)	19% (12%-27%)	4.5% (0%-14%)	0.24	1	0.11	0.12	1	1	1	1	Yes
	4	16% (7.8%-24%)	11% (4.9%-18%)	4.5% (0%-14%)	0.37	0.39	0.29	0.46	1	1	1	1	No
	5	15% (7%-23%)	16% (8.4%-23%)	4.8% (0%-15%)	0.49	1	0.29	0.3	1	1	1	1	No
	10	12% (3.4%-21%)	12% (3.9%-20%)	5.6% (0%-17%)	0.87	1	0.67	0.68	1	1	1	1	No
<i>Exposure to childcare</i>													
Childcare attendance	1	27% (18%-37%)	26% (18%-35%)	36% (15%-58%)	0.6	0.87	0.44	0.43	1	1	1	1	Yes
	2	45% (34%-55%)	44% (35%-54%)	41% (19%-63%)	0.96	1	0.81	0.82	1	1	1	1	Yes
	3	54% (43%-64%)	52% (43%-62%)	73% (53%-93%)	0.22	0.88	0.15	0.1	1	1	1	1	Yes

		4	85% (78%-93%)	86% (78%-93%)	100% (100%-100%)	0.15	1	0.066	0.07	1	1	1	No
<i>Exposure to pets</i>													
Exposure to cat		1	36% (26%-47%)	35% (25%-44%)	23% (3.7%-42%)	0.53	0.88	0.31	0.33	1	1	1	Yes
		2	35% (25%-45%)	36% (26%-45%)	23% (3.7%-42%)	0.55	1	0.32	0.32	1	1	1	Yes
		3	30% (20%-40%)	27% (18%-35%)	14% (0%-29%)	0.34	0.74	0.18	0.28	1	1	1	Yes
		4	28% (18%-38%)	26% (17%-35%)	18% (0.68%-36%)	0.7	0.74	0.42	0.59	1	1	1	No
		5	30% (20%-40%)	29% (19%-38%)	14% (0%-31%)	0.35	0.87	0.18	0.27	1	1	1	No
Exposure to dog		10	34% (22%-47%)	26% (16%-37%)	28% (4.9%-51%)	0.61	0.34	0.78	1	1	1	1	No
		10	38% (25%-51%)	32% (21%-44%)	22% (0.95%-43%)	0.48	0.58	0.27	0.57	1	1	1	No
Exposure to any furred pet		10	62% (49%-75%)	78% (68%-88%)	78% (57%-99%)	0.14	0.077	0.27	1	1	1	1	No
			<b>Mean (95% CI)</b>	<b>Mean (95% CI)</b>	<b>Mean (95% CI)</b>	<b>Overall</b>	<b>1 vs. 2</b>	<b>1 vs. 3</b>	<b>2 vs. 3</b>	<b>Overall</b>	<b>1 vs. 2</b>	<b>1 vs. 3</b>	<b>2 vs. 3</b>
<i>Anthropometry</i>													
Height (cm)		3	96 (95-97)	97 (96-97)	96 (95-97)	0.54	0.3	0.91	0.49	1	1	1	Yes
		4	104 (103-105)	104 (103-105)	103 (101-105)	0.82	0.83	0.64	0.52	1	1	1	No
		5	111 (110-112)	111 (111-112)	110 (108-112)	0.57	0.7	0.44	0.29	1	1	1	No
		10	1.4 (1.4-1.4)	1.4 (1.4-1.4)	1.4 (1.4-1.4)	0.26	0.25	0.15	0.41	1	1	1	No
Weight (kg)		3	15 (15-16)	15 (15-16)	15 (14-16)	0.8	0.61	0.78	0.6	1	1	1	Yes
		4	17 (17-18)	17 (17-18)	18 (16-19)	0.85	0.67	0.63	0.81	1	1	1	No
		5	20 (19-21)	20 (19-21)	20 (18-21)	0.77	0.49	1	0.7	1	1	1	No
		10	37 (35-40)	36 (34-38)	35 (32-39)	0.66	0.4	0.52	1	1	1	1	No
BMI (kg/m <sup>2</sup> )		3	16 (16-17)	16 (16-17)	16 (16-17)	0.86	0.65	0.68	0.8	1	1	1	No*
		4	16 (16-17)	16 (16-16)	17 (16-17)	0.59	0.76	0.32	0.39	1	1	1	No
		5	16 (16-16)	16 (16-16)	16 (15-17)	0.71	0.56	0.48	0.67	1	1	1	No
		10	18 (17-19)	18 (17-18)	18 (17-19)	0.89	0.75	1	0.62	1	1	1	No
<i>Household inhabitants</i>													
Number of children in household		1	1.8 (1.6-2)	1.5 (1.4-1.7)	1.8 (1.3-2.2)	0.013	3.5E-03	0.57	0.21	0.25	0.13	1	Yes
		2	2 (1.8-2.2)	1.6 (1.5-1.8)	2.1 (1.7-2.5)	8.2E-04	5.7E-04	0.68	0.014	0.021	0.026	1	Yes
		3	2.2 (2.1-2.4)	2 (1.8-2.1)	2.3 (1.9-2.7)	0.035	0.022	0.81	0.082	0.57	0.62	1	Yes
		4	2.4 (2.2-2.6)	2.1 (1.9-2.2)	2.5 (2.1-2.9)	0.03	0.02	0.8	0.061	0.5	0.57	1	No
		5	2.4 (2.2-2.6)	2.1 (2-2.3)	2.4 (2-2.8)	0.11	0.04	0.59	0.38	1	1	1	No
		10	2.6 (2.3-2.8)	2.2 (2-2.4)	2.3 (2-2.7)	0.077	0.026	0.35	0.51	1	0.71	1	No
Number of older children in household		0	0.93 (0.72-1.1)	0.53 (0.38-0.69)	0.77 (0.32-1.2)	4.5E-03	1.0E-03	0.37	0.25	0.097	0.043	1	Yes
		1	0.84 (0.65-1)	0.51 (0.36-0.67)	0.77 (0.32-1.2)	0.013	3.5E-03	0.57	0.21	0.25	0.13	1	Yes
		2	0.85 (0.66-1)	0.5 (0.34-0.65)	0.77 (0.32-1.2)	2.8E-03	6.5E-04	0.48	0.16	0.063	0.029	1	Yes
		3	0.81 (0.6-1)	0.5 (0.35-0.65)	0.77 (0.32-1.2)	0.032	0.01	0.74	0.19	0.53	0.32	1	Yes
		4	0.88 (0.66-1.1)	0.49 (0.34-0.65)	0.57 (0.23-0.91)	5.7E-03	1.4E-03	0.19	0.49	0.12	0.059	1	No
		5	0.68 (0.5-0.85)	0.39 (0.25-0.54)	0.67 (0.23-1.1)	0.016	5.1E-03	0.75	0.12	0.29	0.18	1	No
		10	0.76 (0.58-0.95)	0.44 (0.26-0.61)	0.59 (0.22-0.95)	0.028	7.5E-03	0.34	0.39	0.47	0.25	1	No
			<b>Geom. mean (95% CI)</b>	<b>Geom. mean (95% CI)</b>	<b>Geom. mean (95% CI)</b>	<b>Overall</b>	<b>1 vs. 2</b>	<b>1 vs. 3</b>	<b>2 vs. 3</b>	<b>Overall</b>	<b>1 vs. 2</b>	<b>1 vs. 3</b>	<b>2 vs. 3</b>
<i>Maternal Vitamin D</i>													
Vitamin D (nmol/L)		6wk	76 (69-83)	71 (66-77)	69 (59-80)	0.29	0.31	0.15	0.38	1	1	1	Yes
<i>Child Vitamin D</i>													
Vitamin D (nmol/L)		0	26 (23-29)	27 (24-30)	26 (21-32)	0.69	0.58	0.71	0.41	1	1	1	No
		6m	64 (59-69)	64 (59-68)	59 (50-70)	0.66	0.68	0.38	0.51	1	1	1	No
		1	60 (55-64)	59 (55-63)	59 (52-67)	0.93	0.98	0.76	0.7	1	1	1	No

		CAS1 (N=88)		CAS2 (N=107)		CAS3 (N=22)		P-value Overall	1 vs. 2	1 vs. 3	2 vs. 3	Overall	1 vs. 2	1 vs. 3	2 vs. 3	Feature?
		Age	Geom. mean (95% CI)	Geom. mean (95% CI)	Geom. mean (95% CI)	Geom. mean (95% CI)	Geom. mean (95% CI)									
"Deseasonalised" Vitamin D (nmol/L)	2	57 (54-61)	58 (55-61)	47 (40-55)	0.012	0.82	5.4E-03	4.4E-03	0.23	1	0.19	0.16	No			
	3	53 (49-58)	59 (56-62)	51 (46-57)	0.082	0.14	0.38	0.033	1	1	1	0.86	No			
	4	59 (56-62)	57 (53-60)	50 (45-56)	0.082	0.51	0.023	0.077	1	1	0.64	1	No			
	5	89 (83-95)	84 (79-89)	77 (69-84)	0.057	0.46	0.016	0.056	0.86	1	0.47	1	No			
	10	77 (72-82)	78 (73-83)	65 (56-75)	0.039	0.86	0.024	0.012	0.63	1	0.67	0.37	No			
	0	26 (23-29)	27 (25-30)	26 (21-31)	0.46	0.4	0.6	0.26	1	1	1	1	No			
	6m	65 (60-69)	64 (60-68)	60 (53-68)	0.46	0.61	0.21	0.36	1	1	1	1	No			
	1	60 (57-64)	60 (56-63)	59 (51-68)	0.91	0.89	0.67	0.71	1	1	1	1	No			
	2	57 (54-61)	58 (56-61)	48 (40-56)	0.032	0.77	0.011	0.015	0.53	1	0.34	0.45	No			
	3	52 (48-57)	60 (57-62)	51 (45-58)	0.018	0.052	0.27	9.8E-03	0.32	1	1	0.31	No			
4	59 (56-62)	57 (54-60)	51 (44-59)	0.31	0.52	0.13	0.26	1	1	1	1	No				
5	88 (83-94)	85 (81-89)	80 (73-87)	0.27	0.57	0.14	0.16	1	1	1	1	No				
10	76 (71-81)	78 (74-83)	69 (60-80)	0.21	0.44	0.23	0.074	1	1	1	1	No				

BMI = body mass index; feature? = whether variable was used as a clustering feature or not; geom. mean = geometric mean; P-value (adj.) = adjusted P-values (Benjamini-Yekutieli method); prop. = proportion. For categorical variables, associations were tested using Fisher exact test; for continuous variables, Kruskal-Wallis and Mann-Whitney-Wilcoxon. Bold text indicates statistical significance ( $p < 0.05$ ); italics indicate near-significance ( $p < 0.10$ ). \*Not used as clustering feature, as BMI is a derived variable. Height and weight at age 3 were used instead.

## B. Immunological (antibodies)

Variable	Age	CAS1 (N=88)		CAS2 (N=107)		CAS3 (N=22)		P-value Overall	1 vs. 2	1 vs. 3	2 vs. 3	Overall	1 vs. 2	1 vs. 3	2 vs. 3	Feature?
		Geom. mean (95% CI)	Geom. mean (95% CI)	Geom. mean (95% CI)	Geom. mean (95% CI)	Geom. mean (95% CI)	Geom. mean (95% CI)									
<i>Total antibody</i>																
IgE (kU/L)	6m	1.2 (0.69-2)	2.2 (1.4-3.6)	21 (12-35)	1.2E-07	0.044	6.7E-08	2.2E-06	5.9E-06	1	8.1E-06	1.9E-04	Yes			
	1	0.6 (0.29-1.3)	2 (1.1-3.7)	43 (17-109)	2.0E-09	0.019	4.3E-09	5.3E-08	1.3E-07	0.55	6.8E-07	6.6E-06	Yes			
	2	6.6 (3.5-12)	17 (12-25)	187 (131-267)	1.2E-11	0.044	4.2E-11	1.4E-10	1.4E-09	1	1.2E-08	3.6E-08	Yes			
	3	18 (13-27)	28 (22-35)	267 (177-401)	1.3E-11	0.11	2.8E-11	8.9E-11	1.4E-09	1	8.3E-09	2.5E-08	Yes			
	4	20 (13-29)	35 (26-47)	345 (212-563)	2.8E-10	0.064	2.0E-10	8.5E-09	2.2E-08	1	4.6E-08	1.2E-06	No			
	5	35 (23-55)	60 (46-80)	451 (278-731)	2.2E-08	0.096	1.9E-08	1.5E-07	1.2E-06	1	2.6E-06	1.7E-05	No			
10	85 (46-154)	150 (103-217)	800 (405-1.6E+03)	1.4E-04	0.11	1.3E-04	2.8E-04	4.2E-03	1	7.1E-03	0.014	No				
<i>HDM antibody</i>																
IgE (kU/L)	6m	0.018 (0.016-0.02)	0.019 (0.016-0.022)	0.033 (0.019-0.059)	1.9E-03	0.47	7.9E-04	4.2E-03	0.045	1	0.035	0.15	Yes			
	1	0.019 (0.017-0.023)	0.019 (0.016-0.022)	0.26 (0.075-0.93)	1.3E-09	0.47	2.5E-07	4.5E-09	9.1E-08	1	2.6E-05	7.1E-07	Yes			
	2	0.024 (0.019-0.031)	0.042 (0.029-0.06)	7.1 (2.7-19)	2.6E-16	0.078	2.5E-15	3.5E-13	8.0E-14	1	2.6E-12	2.4E-10	Yes			
	3	0.043 (0.029-0.064)	0.064 (0.04-0.1)	23 (7.5-68)	1.5E-13	0.49	8.8E-13	2.4E-12	2.2E-11	1	4.9E-10	1.0E-09	Yes			
	4	0.057 (0.036-0.09)	0.2 (0.11-0.35)	30 (8.9-103)	3.7E-11	2.8E-03	2.0E-10	1.7E-08	3.5E-09	0.11	4.6E-08	2.3E-06	No			
	5	0.072 (0.041-0.13)	0.23 (0.12-0.45)	31 (7.8-127)	4.2E-09	0.015	3.8E-09	5.1E-07	2.6E-07	0.45	6.1E-07	5.0E-05	No			
	10	0.37 (0.17-0.8)	1.3 (0.51-3.4)	52 (19-144)	2.9E-06	0.068	5.7E-07	9.7E-05	1.1E-04	1	5.4E-05	5.5E-03	No			
	1	0.21 (0.2-0.23)	0.23 (0.21-0.25)	0.29 (0.21-0.39)	0.042	0.34	0.012	0.07	0.66	1	0.37	1	Yes			
	2	0.32 (0.27-0.37)	0.49 (0.41-0.59)	0.89 (0.57-1.4)	1.9E-06	2.1E-04	3.8E-06	7.0E-03	7.8E-05	0.011	3.0E-04	0.24	Yes			
	3	0.24 (0.22-0.26)	0.31 (0.26-0.36)	0.88 (0.51-1.5)	2.5E-07	0.023	8.5E-08	3.7E-05	1.2E-05	0.64	1.0E-05	2.3E-03	Yes			
5	0.55 (0.42-0.7)	0.59 (0.46-0.74)	1.7 (0.88-3.3)	1.5E-03	0.67	6.4E-04	9.0E-04	0.036	1	0.029	0.039	No				
10	1.6 (1.3-1.9)	2.1 (1.8-2.5)	2.8 (1.9-4.2)	1.0E-02	0.023	0.011	0.18	0.2	0.64	0.34	1	No				
6m	1.5E-04 (1.5E-04-1.5E-04)	1.7E-04 (1.3E-04-2.1E-04)	4.6E-04 (9.0E-05-2.4E-03)	4.9E-03	0.37	5.2E-03	0.024	0.1	1	0.18	0.67	Yes				



1	1.5E-04 (1.5E-04-1.5E-04)	6.9E-04 (3.2E-04-1.5E-03)	0.081 (4.6E-03-1.4)	1.8E-10	5.2E-04	6.6E-12	2.2E-05	1.5E-08	0.024	2.5E-09	1.4E-03	Yes
2	3.4E-04 (1.8E-04-6.6E-04)	4.8 (1.7-13)	61 (8.9-419)	1.8E-25	1.5E-22	8.6E-18	9.8E-05	5.0E-22	1.4E-18	2.0E-14	5.5E-03	Yes
3	7.7E-04 (3.1E-04-1.9E-03)	35 (18-71)	198 (46-859)	2.9E-29	8.2E-26	1.0E-16	1.8E-06	1.6E-25	1.6E-21	1.6E-13	1.6E-04	Yes
4	0.01 (2.6E-03-0.041)	16 (5.5-44)	389 (228-664)	3.1E-18	2.0E-13	2.2E-12	1.5E-05	1.9E-15	1.5E-10	9.7E-10	1.0E-03	No
5	2 (0.48-8.1)	168 (111-256)	539 (317-917)	1.1E-15	1.3E-12	1.0E-08	1.9E-04	2.8E-13	6.5E-10	1.4E-06	0.01	No
Non-IgG4 IgG (mg/L)	1	0.21 (0.2-0.23)	0.28 (0.21-0.39)	0.042	0.34	0.012	0.07	0.66	1	0.37	1	No*
	2	0.31 (0.26-0.37)	0.46 (0.38-0.54)	5.7E-05	6.7E-04	8.6E-05	0.05	1.8E-03	0.03	4.9E-03	1	No*
	3	0.24 (0.22-0.26)	0.29 (0.25-0.33)	2.3E-04	0.061	5.8E-05	3.5E-03	6.5E-03	1	3.5E-03	0.13	No*
	5	0.52 (0.41-0.66)	0.47 (0.38-0.59)	9.4E-03	0.57	7.1E-03	3.3E-03	0.19	1	0.24	0.13	No
IgG-IgE (mg/L;mg/L)	1	4.6E+03 (3.9E+03-5.4E+03)	5.1E+03 (4.3E+03-6.0E+03)	2.9E-06	0.15	3.7E-05	4.7E-06	1.1E-04	1	2.3E-03	3.7E-04	No*
	2	5.4E+03 (4.0E+03-7.2E+03)	5.0E+03 (3.3E+03-7.6E+03)	4.9E-11	0.15	1.4E-11	5.4E-10	4.5E-09	1	4.7E-09	1.1E-07	No*
	3	2.3E+03 (1.5E+03-3.4E+03)	2.0E+03 (1.3E+03-3.2E+03)	3.9E-11	0.96	2.0E-11	2.7E-10	3.7E-09	1	6.2E-09	5.9E-08	No*
	5	3.8E+03 (2.2E+03-6.5E+03)	1.2E+03 (608-2.4E+03)	2.2E-08	0.029	8.8E-09	1.8E-06	1.2E-06	0.77	1.3E-06	1.6E-04	No
	10	1.7E+03 (830-3.5E+03)	656 (263-1.6E+03)	3.0E-05	0.18	3.0E-06	3.7E-04	9.9E-04	1	2.5E-04	0.018	No
IgG4-IgE (µg/L;µg/L)	6m	3.5E-03 (3.1E-03-4.0E-03)	3.6E-03 (2.8E-03-4.8E-03)	0.06	0.61	0.026	0.058	0.89	1	0.71	1	No*
	1	3.2E-03 (2.8E-03-3.8E-03)	0.015 (6.9E-03-0.033)	4.2E-03	9.2E-03	3.6E-03	0.15	0.091	0.3	0.13	1	No*
	2	5.8E-03 (3.0E-03-0.011)	53 (18-154)	2.2E-20	2.1E-19	8.8E-09	4.3E-04	2.4E-17	8.0E-16	1.3E-06	0.02	No*
	3	7.4E-03 (2.8E-03-0.02)	231 (101-528)	1.3E-23	9.4E-22	1.4E-07	1.1E-06	2.4E-20	6.0E-18	1.6E-05	1.0E-04	No*
	4	0.076 (0.021-0.28)	33 (11-103)	3.4E-09	7.7E-09	1.6E-03	1.5E-03	2.2E-07	1.1E-06	0.066	0.063	No
	5	11 (3-44)	307 (148-634)	6.3E-06	3.1E-04	0.13	1.2E-05	2.3E-04	0.015	1	8.3E-04	No
Cat antibody												
IgE (kU/L)	6m	0.018 (0.016-0.021)	0.021 (0.018-0.024)	0.029	0.21	7.3E-03	0.075	0.48	1	0.24	1	Yes
	1	0.017 (0.015-0.019)	0.017 (0.015-0.018)	1.7E-08	0.34	6.6E-06	3.9E-08	9.7E-07	1	4.9E-04	5.0E-06	Yes
	2	0.018 (0.016-0.021)	0.02 (0.017-0.023)	3.2E-09	0.66	4.6E-08	8.3E-08	2.1E-07	1	5.7E-06	1.0E-05	Yes
	3	0.021 (0.017-0.026)	0.019 (0.016-0.022)	2.2E-09	0.6	5.6E-07	6.8E-09	1.5E-07	1	5.3E-05	1.0E-06	Yes
	4	0.017 (0.015-0.018)	0.023 (0.018-0.029)	9.2E-12	0.071	1.3E-11	1.4E-07	1.1E-09	1	4.5E-09	1.6E-05	No
	5	0.018 (0.015-0.021)	0.027 (0.02-0.036)	2.4E-10	0.022	1.0E-10	1.2E-06	1.9E-08	0.62	2.7E-08	1.1E-04	No
	10	0.05 (0.029-0.084)	0.056 (0.033-0.095)	4.1E-06	0.71	3.8E-06	6.9E-06	1.5E-04	1	3.0E-04	5.1E-04	No
IgG (mg/L)	1	0.2 (0.2-0.21)	0.23 (0.21-0.25)	0.024	0.015	5.5E-03	0.49	0.42	0.45	0.19	1	Yes
	2	0.28 (0.24-0.32)	0.37 (0.31-0.45)	0.023	6.0E-03	0.19	0.71	0.4	0.21	1	1	Yes
	3	0.22 (0.21-0.24)	0.3 (0.26-0.35)	1.9E-03	8.6E-03	5.2E-04	0.14	0.045	0.28	0.024	1	Yes
	5	0.48 (0.38-0.62)	0.52 (0.41-0.66)	0.06	0.7	0.021	0.037	0.89	1	0.6	0.94	No
	10	0.72 (0.54-0.95)	0.89 (0.72-1.1)	0.3	0.18	0.24	0.72	1	1	1	1	No
IgG4 (µg/L)	6m	2.3E-04 (1.4E-04-3.9E-04)	4.8E-04 (2.4E-04-9.8E-04)	0.048	0.099	0.012	0.21	0.74	1	0.37	1	Yes
	1	7.5E-04 (3.2E-04-1.8E-03)	0.032 (9.5E-03-0.11)	2.2E-08	3.4E-06	1.6E-08	0.017	1.2E-06	2.8E-04	2.2E-06	0.5	Yes
	2	0.055 (0.015-0.21)	68 (36-126)	1.3E-19	2.6E-19	3.6E-08	0.34	1.2E-16	8.3E-16	4.7E-06	1	Yes
	3	0.32 (0.079-1.3)	110 (58-210)	7.0E-19	3.2E-17	1.1E-09	0.082	4.9E-16	6.8E-14	2.0E-07	1	Yes
	4	1.3 (0.31-5.5)	117 (56-245)	3.0E-14	1.9E-12	2.8E-08	0.054	5.9E-12	8.8E-10	3.8E-06	1	No
	5	32 (10-100)	449 (377-535)	2.7E-11	5.1E-10	2.1E-06	0.035	2.7E-09	1.1E-07	1.8E-04	0.9	No
Non-IgG4 IgG (mg/L)	1	0.2 (0.2-0.21)	0.23 (0.21-0.24)	0.025	0.015	5.5E-03	0.49	0.43	0.45	0.19	1	No*
	2	0.27 (0.24-0.32)	0.33 (0.28-0.38)	0.14	0.056	0.79	0.37	1	1	1	1	No*
	3	0.22 (0.21-0.24)	0.26 (0.23-0.29)	0.017	0.12	4.1E-03	0.078	0.31	1	0.15	1	No*
	5	0.43 (0.34-0.54)	0.37 (0.3-0.46)	0.54	0.48	0.53	0.33	1	1	1	1	No
IgG-IgE (mg/L;mg/L)	1	4.9E+03 (4.4E+03-5.5E+03)	5.7E+03 (5.0E+03-6.4E+03)	3.1E-05	0.03	8.4E-04	9.6E-05	1.0E-03	0.79	0.037	5.5E-03	No*
	2	6.2E+03 (5.0E+03-7.7E+03)	8.0E+03 (6.4E+03-9.9E+03)	3.0E-06	0.043	2.0E-05	7.1E-06	1.2E-04	1	1.3E-03	5.2E-04	No*

	3	4.5E+03 (3.6E+03-5.7E+03)	6.6E+03 (5.4E+03-8.1E+03)	1.8E+03 (724-4.6E+03)	2.8E-04	0.057	3.6E-03	2.8E-04	7.7E-03	1	0.13	0.014	No*
	5	1.1E+04 (8.4E+03-1.4E+04)	8.3E+03 (6.0E+03-1.2E+04)	1.8E+03 (686-4.7E+03)	1.6E-03	0.5	3.9E-04	2.2E-03	0.038	1	0.019	0.087	No
	10	6.0E+03 (3.6E+03-1.0E+04)	6.6E+03 (3.9E+03-1.1E+04)	326 (110-960)	2.3E-05	0.72	7.1E-06	2.5E-05	7.8E-04	1	5.2E-04	1.6E-03	No
IgG4:IgE (µg/L)	6m	5.4E-03 (3.2E-03-9.2E-03)	9.7E-03 (4.7E-03-0.02)	0.025 (1.9E-03-0.33)	0.72	0.94	0.41	0.51	1	1	1	1	No*
	1	0.018 (7.5E-03-0.044)	0.82 (0.24-2.8)	9.1 (0.48-171)	7.8E-06	4.4E-06	4.4E-04	0.54	2.8E-04	3.5E-04	0.021	1	No*
	2	1.3 (0.33-4.8)	1.4E+03 (752-2.7E+03)	191 (20-1.9E+03)	7.7E-18	1.6E-18	5.2E-04	0.011	4.3E-15	4.4E-15	0.024	0.34	No*
	3	6.5 (1.5-28)	2.4E+03 (1.2E+03-4.8E+03)	1.3E+03 (578-3.0E+03)	4.9E-15	8.1E-16	2.3E-03	0.018	1.1E-12	9.1E-13	0.097	0.52	No*
	4	33 (8-138)	2.1E+03 (1.0E+03-4.6E+03)	656 (151-2.8E+03)	1.9E-09	8.9E-10	0.44	8.4E-04	1.3E-07	1.6E-07	1	0.037	No
	5	732 (235-2.3E+03)	7.0E+03 (5.0E+03-9.8E+03)	580 (104-3.2E+03)	5.7E-07	7.0E-06	0.055	9.2E-05	2.5E-05	5.2E-04	1	5.2E-03	No
<i>Peanut antibody</i>													
IgE (kU/L)	6m	0.024 (0.019-0.03)	0.03 (0.024-0.037)	0.21 (0.079-0.58)	1.0E-06	0.11	4.0E-07	1.6E-05	4.3E-05	1	4.0E-05	1.1E-03	Yes
	1	0.021 (0.017-0.025)	0.024 (0.02-0.03)	0.54 (0.19-1.5)	8.9E-12	0.34	8.5E-11	5.6E-10	1.1E-09	1	2.4E-08	1.1E-07	Yes
	2	0.024 (0.02-0.029)	0.025 (0.021-0.03)	0.58 (0.24-1.4)	4.2E-14	0.56	7.3E-13	1.2E-12	7.8E-12	1	4.2E-10	6.3E-10	Yes
	3	0.021 (0.017-0.026)	0.018 (0.016-0.021)	0.42 (0.14-1.3)	3.8E-17	0.75	4.7E-12	9.8E-15	1.7E-14	1	1.9E-09	9.8E-12	Yes
	4	0.019 (0.016-0.022)	0.022 (0.018-0.027)	0.37 (0.11-1.2)	4.4E-13	0.17	6.4E-12	8.9E-10	6.1E-11	1	2.4E-09	1.6E-07	No
	5	0.027 (0.02-0.037)	0.034 (0.024-0.047)	0.73 (0.19-2.8)	6.9E-09	0.3	7.9E-09	2.2E-07	4.2E-07	1	1.2E-06	2.3E-05	No
	10	0.085 (0.052-0.14)	0.072 (0.045-0.11)	0.78 (0.25-2.4)	1.5E-04	0.58	2.6E-04	3.9E-05	4.5E-03	1	0.013	2.4E-03	No
IgG4 (µg/L)	6m	1.5E-04 (1.5E-04-1.5E-04)	1.5E-04 (1.5E-04-1.5E-04)	8.1E-04 (1.2E-04-5.6E-03)	2.1E-06	NA	5.7E-04	1.5E-04	8.5E-05	NA	0.026	8.2E-03	Yes
	1	2.0E-04 (1.3E-04-2.9E-04)	2.7E-03 (9.5E-04-7.7E-03)	0.53 (0.023-1.2)	3.3E-10	4.4E-05	3.8E-12	1.3E-04	2.5E-08	2.7E-03	1.5E-09	7.1E-03	Yes
	2	2.9E-03 (8.5E-04-9.6E-03)	4.6 (1.3-16)	20 (1.4-280)	4.9E-12	1.9E-11	5.7E-08	0.33	5.9E-10	6.0E-09	7.0E-06	1	Yes
	3	5.3E-03 (1.4E-03-0.021)	27 (8.8-85)	77 (6.1-953)	3.9E-14	1.7E-13	2.0E-08	0.2	7.5E-12	1.3E-10	2.7E-06	1	Yes
	4	0.012 (2.8E-03-0.051)	11 (2.6-43)	67 (5-907)	1.1E-10	8.5E-10	2.8E-07	0.31	9.8E-09	1.6E-07	2.9E-05	1	No
	5	0.45 (0.083-2.5)	91 (31-267)	316 (43-2.3E+03)	1.5E-07	8.4E-07	3.1E-05	0.23	7.3E-06	7.8E-05	2.0E-03	1	No
IgG4:IgE (µg/L)	6m	2.6E-03 (2.1E-03-3.2E-03)	2.1E-03 (1.7E-03-2.6E-03)	1.6E-03 (1.9E-04-0.013)	7.1E-03	0.11	3.8E-03	0.021	0.15	1	0.14	0.6	No*
	1	4.0E-03 (2.5E-03-6.2E-03)	0.046 (0.017-0.13)	0.41 (0.027-6.2)	9.9E-04	6.5E-03	2.2E-04	0.15	0.025	0.22	0.011	1	No*
	2	0.05 (0.015-0.17)	75 (21-268)	14 (1.2-160)	1.0E-10	6.7E-11	3.7E-03	3.8E-03	9.1E-09	1.9E-08	0.14	0.14	No*
	3	0.11 (0.028-0.39)	626 (204-1.9E+03)	75 (5.7-991)	4.6E-14	1.3E-14	3.7E-04	9.5E-03	8.1E-12	1.2E-11	0.018	0.31	No*
	4	0.26 (0.061-1.1)	199 (48-818)	76 (7-824)	1.3E-08	5.8E-09	3.0E-03	0.028	7.6E-07	9.0E-07	0.12	0.75	No
	5	7 (1.3-38)	1.1E+03 (389-3.2E+03)	179 (26-1.2E+03)	2.3E-05	1.2E-05	0.28	0.011	7.8E-04	8.3E-04	1	0.34	No
<i>Couch grass antibody</i>													
IgE (kU/L)	6m	0.02 (0.017-0.024)	0.021 (0.017-0.024)	0.024 (0.016-0.035)	0.47	0.88	0.28	0.24	1	1	1	1	Yes
	1	0.019 (0.016-0.022)	0.018 (0.016-0.02)	0.029 (0.017-0.049)	0.049	0.72	0.04	0.019	0.75	1	1	0.55	Yes
	2	0.022 (0.019-0.026)	0.019 (0.017-0.022)	0.087 (0.05-0.15)	7.2E-10	0.096	3.4E-07	2.7E-10	5.4E-08	1	3.4E-05	5.9E-08	Yes
	3	0.019 (0.015-0.023)	0.018 (0.015-0.021)	0.12 (0.047-0.28)	1.1E-12	0.23	5.7E-08	1.8E-11	1.5E-10	1	7.0E-06	5.8E-09	Yes
	4	0.02 (0.015-0.025)	0.023 (0.017-0.031)	0.32 (0.12-0.89)	5.8E-14	0.54	2.3E-11	2.0E-10	9.6E-12	1	7.1E-09	4.6E-08	No
	5	0.029 (0.02-0.041)	0.046 (0.029-0.072)	1.2 (0.34-4.3)	2.7E-08	0.17	1.5E-08	1.3E-06	1.5E-06	1	2.1E-06	1.2E-04	No
	10	0.19 (0.095-0.38)	0.12 (0.065-0.22)	2.8 (0.86-9.4)	3.3E-04	0.36	8.2E-04	7.6E-05	8.9E-03	1	0.036	4.5E-03	No
IgG4 (µg/L)	6m	1.5E-04 (1.5E-04-1.5E-04)	1.7E-04 (1.3E-04-2.1E-04)	2.8E-04 (7.6E-05-1.0E-03)	0.14	0.37	0.057	0.22	1	1	1	1	Yes
	1	1.7E-04 (1.3E-04-2.3E-04)	8.0E-04 (3.6E-04-1.8E-03)	0.01 (6.2E-04-0.18)	2.1E-05	1.3E-03	8.6E-07	0.015	7.2E-04	0.055	8.0E-05	0.45	Yes
	2	2.0E-04 (1.3E-04-3.0E-04)	0.02 (5.8E-03-0.067)	0.14 (6.1E-03-3.2)	7.3E-10	2.2E-09	5.3E-10	0.12	5.4E-08	3.8E-07	1.1E-07	1	Yes
	3	2.6E-03 (8.2E-04-8.4E-03)	13 (4.4-41)	18 (1.4-241)	3.4E-17	6.6E-17	3.3E-09	0.35	1.7E-14	1.1E-13	5.4E-07	1	Yes
	4	0.037 (8.1E-03-0.17)	42 (14-125)	86 (11-698)	6.4E-14	2.7E-13	1.8E-07	0.32	1.0E-11	1.9E-10	1.9E-05	1	No
	5	0.26 (0.055-1.3)	125 (59-265)	196 (32-1.2E+03)	9.1E-14	7.2E-13	3.0E-07	0.24	1.4E-11	4.2E-10	3.1E-05	1	No
IgG4:IgE (µg/L)	6m	3.1E-03 (2.6E-03-3.6E-03)	3.4E-03 (2.6E-03-4.5E-03)	4.9E-03 (1.2E-03-0.02)	0.7	0.77	0.51	0.43	1	1	1	1	No*
	1	3.8E-03 (2.8E-03-5.2E-03)	0.018 (8.1E-03-0.042)	0.15 (0.01-2.2)	0.014	0.032	7.1E-03	0.17	0.26	0.84	0.24	1	No*

2	3.7E-03 (2.4E-03-5.8E-03)	0.43 (0.12-1.5)	0.66 (0.031-1.4)	<b>5.7E-07</b>	<b>2.9E-08</b>	<b>0.038</b>	0.39	<b>2.5E-05</b>	<b>3.8E-06</b>	0.97	1	No*
3	0.058 (0.018-0.19)	316 (104-962)	65 (3.5-1.2E+03)	<b>3.7E-16</b>	<b>5.6E-17</b>	<b>1.1E-05</b>	0.23	<b>1.1E-13</b>	<b>1.1E-13</b>	<b>7.8E-04</b>	1	No*
4	0.77 (0.18-3.4)	759 (247-2.3E+03)	111 (11-1.2E+03)	<b>1.7E-11</b>	<b>6.0E-12</b>	<b>0.014</b>	<b>3.4E-03</b>	<b>1.8E-09</b>	<b>2.3E-09</b>	0.42	0.13	No
5	3.9 (0.74-20)	1.1E+03 (497-2.6E+03)	67 (9.6-472)	<b>1.1E-07</b>	<b>1.2E-07</b>	0.47	<b>5.7E-04</b>	<b>5.5E-06</b>	<b>1.4E-05</b>	1	<b>0.026</b>	No

*Ryegrass antibody*

IgE (kU/L)	6m	0.021 (0.018-0.025)	0.022 (0.018-0.026)	0.024 (0.016-0.035)	0.54	0.74	0.3	0.34	1	1	1	Yes	
	1	0.017 (0.015-0.02)	0.017 (0.016-0.019)	0.025 (0.015-0.042)	0.066	0.78	<b>0.033</b>	<b>0.042</b>	0.97	1	0.86	Yes	
	3	0.026 (0.02-0.035)	0.02 (0.017-0.025)	0.28 (0.1-0.78)	<b>2.0E-10</b>	0.14	<b>2.9E-07</b>	<b>1.2E-10</b>	<b>1.7E-08</b>	1	<b>3.0E-05</b>	Yes	
	4	0.024 (0.018-0.033)	0.03 (0.022-0.042)	0.86 (0.3-2.4)	<b>1.7E-11</b>	0.27	<b>1.1E-10</b>	<b>1.7E-09</b>	<b>1.8E-09</b>	1	<b>3.0E-08</b>	No	
	5	0.039 (0.026-0.059)	0.056 (0.035-0.09)	1.8 (0.49-6.4)	<b>2.5E-07</b>	0.37	<b>1.7E-07</b>	<b>1.4E-06</b>	<b>1.2E-05</b>	1	<b>1.9E-05</b>	No	
	10	0.3 (0.14-0.63)	0.23 (0.12-0.45)	4.3 (1.2-15)	<b>1.1E-03</b>	0.77	<b>1.2E-03</b>	<b>2.7E-04</b>	<b>0.027</b>	1	<i>0.051</i>	No	
IgG4 (µg/L)	6m	1.5E-04 (1.5E-04-1.5E-04)	1.5E-04 (1.5E-04-1.5E-04)	2.6E-04 (8.1E-05-8.6E-04)	<b>0.013</b>	NA	<i>0.051</i>	<b>0.031</b>	0.25	NA	1	0.81	Yes
	1	1.8E-04 (1.3E-04-2.5E-04)	2.6E-04 (1.6E-04-4.3E-04)	5.9E-03 (3.8E-04-0.09)	<b>8.4E-06</b>	0.17	<b>1.2E-05</b>	<b>3.9E-04</b>	<b>3.0E-04</b>	1	<b>8.3E-04</b>	<b>0.019</b>	Yes
	2	2.0E-04 (1.3E-04-2.9E-04)	5.9E-03 (1.8E-03-0.019)	0.12 (5.5E-03-2.6)	<b>4.0E-08</b>	<b>1.6E-06</b>	<b>6.3E-10</b>	<b>0.041</b>	<b>2.2E-06</b>	<b>1.4E-04</b>	<b>1.2E-07</b>	1	Yes
	3	1.5E-04 (1.5E-04-1.5E-04)	0.013 (3.7E-03-0.042)	0.2 (9.7E-03-4.3)	<b>2.0E-10</b>	<b>3.7E-09</b>	<b>1.3E-12</b>	<i>0.056</i>	<b>1.7E-08</b>	<b>6.0E-07</b>	<b>6.5E-10</b>	1	Yes
	4	2.7E-04 (1.5E-04-5.0E-04)	7.6E-03 (2.2E-03-0.027)	0.19 (7.8E-03-4.8)	<b>2.7E-07</b>	<b>1.7E-05</b>	<b>2.9E-08</b>	<b>0.023</b>	<b>1.2E-05</b>	<b>1.1E-03</b>	<b>3.8E-06</b>	0.64	No
	5	3.8E-03 (9.7E-04-0.015)	0.25 (0.056-1.1)	30 (3-303)	<b>1.5E-07</b>	<b>8.5E-05</b>	<b>1.2E-07</b>	<b>5.1E-03</b>	<b>7.3E-06</b>	<b>4.9E-03</b>	<b>1.4E-05</b>	0.18	No
IgG4:IgE (µg/L:µg/L)	6m	3.0E-03 (2.5E-03-3.6E-03)	2.9E-03 (2.4E-03-3.5E-03)	4.7E-03 (1.3E-03-0.017)	0.79	0.74	0.54	0.59	1	1	1	No*	
	1	4.2E-03 (3.0E-03-6.1E-03)	6.4E-03 (3.9E-03-0.011)	0.097 (7.1E-03-1.3)	<b>0.023</b>	0.5	<b>9.1E-03</b>	<b>0.027</b>	0.4	1	0.29	0.73	No*
	3	2.4E-03 (1.8E-03-3.1E-03)	0.26 (0.075-0.89)	0.3 (0.013-6.9)	<b>2.7E-06</b>	<b>1.7E-07</b>	<b>0.011</b>	<b>0.53</b>	<b>1.1E-04</b>	<b>1.9E-05</b>	0.34	1	No*
	4	4.7E-03 (2.5E-03-8.7E-03)	0.1 (0.03-0.36)	0.095 (3.9E-03-2.3)	<b>0.014</b>	<b>2.1E-03</b>	0.44	0.32	0.26	<i>0.084</i>	1	1	No
	5	0.041 (0.011-0.15)	1.9 (0.43-8.3)	7.1 (0.83-60)	<b>2.6E-04</b>	<b>5.5E-04</b>	<b>5.1E-04</b>	0.75	<b>7.2E-03</b>	<b>0.025</b>	<b>0.024</b>	1	No

*(Timothy) Grass antibody*

IgG (mg/L)	1	0.24 (0.21-0.28)	0.29 (0.25-0.34)	0.35 (0.24-0.52)	<b>0.029</b>	<b>0.054</b>	<b>9.5E-03</b>	0.22	0.48	1	0.31	1	Yes
	2	0.37 (0.31-0.46)	0.64 (0.52-0.79)	0.51 (0.34-0.77)	<b>1.4E-03</b>	<b>3.0E-04</b>	0.13	0.44	<b>0.034</b>	<b>0.015</b>	1	1	Yes
	3	0.25 (0.23-0.28)	0.38 (0.32-0.47)	0.57 (0.37-0.9)	<b>9.8E-05</b>	<b>2.3E-03</b>	<b>2.3E-05</b>	<i>0.059</i>	<b>3.1E-03</b>	<i>0.091</i>	<b>1.5E-03</b>	1	Yes
	5	0.77 (0.61-0.98)	0.77 (0.6-0.99)	1.8 (1.1-2.9)	<b>0.014</b>	0.88	<b>4.7E-03</b>	<b>6.9E-03</b>	0.26	1	0.17	0.23	No
	10	1.4 (1.1-1.6)	1.6 (1.3-1.9)	1.5 (1.1-2)	0.34	0.15	0.68	0.49	1	1	1	1	No

*Mould antibody*

IgE (kU/L)	6m	0.018 (0.016-0.02)	0.017 (0.015-0.02)	0.019 (0.013-0.027)	0.66	0.39	0.99	0.58	1	1	1	1	Yes
	1	0.016 (0.015-0.017)	0.016 (0.015-0.017)	0.015 (0.015-0.015)	0.77	0.82	0.48	0.53	1	1	1	1	Yes
	2	0.018 (0.016-0.02)	0.018 (0.016-0.021)	0.025 (0.018-0.034)	<b>3.4E-03</b>	0.94	<b>3.6E-03</b>	<b>2.5E-03</b>	<i>0.075</i>	1	0.13	<i>0.098</i>	Yes
	3	0.022 (0.017-0.028)	0.019 (0.016-0.023)	0.048 (0.026-0.09)	<b>2.3E-06</b>	0.49	<b>1.3E-04</b>	<b>1.2E-06</b>	<b>9.1E-05</b>	1	<b>7.1E-03</b>	<b>1.1E-04</b>	Yes
	4	0.017 (0.015-0.02)	0.026 (0.018-0.037)	0.042 (0.023-0.078)	<b>1.4E-05</b>	0.1	<b>2.4E-06</b>	<b>1.6E-03</b>	<b>4.9E-04</b>	1	<b>2.0E-04</b>	<i>0.066</i>	No
	5	0.018 (0.015-0.022)	0.029 (0.019-0.044)	0.085 (0.034-0.21)	<b>7.8E-07</b>	<i>0.071</i>	<b>1.6E-07</b>	<b>2.8E-04</b>	<b>3.4E-05</b>	1	<b>1.8E-05</b>	<b>0.014</b>	No
	10	0.057 (0.033-0.096)	0.087 (0.046-0.16)	1 (0.34-3.2)	<b>1.2E-04</b>	0.41	<b>2.1E-05</b>	<b>3.8E-04</b>	<b>3.7E-03</b>	1	<b>1.4E-03</b>	<b>0.018</b>	No
IgG4 (µg/L)	6m	2.3E-04 (1.4E-04-3.6E-04)	1.8E-04 (1.4E-04-2.5E-04)	2.8E-04 (7.7E-05-1.0E-03)	0.69	0.49	0.8	0.46	1	1	1	1	Yes
	1	1.5E-04 (1.5E-04-1.5E-04)	1.5E-04 (1.5E-04-1.5E-04)	2.7E-04 (8.1E-05-8.8E-04)	<b>0.012</b>	NA	<i>0.051</i>	<b>0.029</b>	0.23	NA	1	0.77	Yes
	2	1.5E-04 (1.5E-04-1.5E-04)	4.9E-04 (2.4E-04-9.9E-04)	2.4E-04 (8.8E-05-6.8E-04)	<b>0.011</b>	<b>3.1E-03</b>	<b>0.049</b>	0.41	0.22	0.12	1	1	Yes
	3	2.0E-04 (1.3E-04-3.1E-04)	1.1E-03 (4.6E-04-2.8E-03)	2.1E-03 (2.2E-04-0.019)	<b>2.9E-03</b>	<b>1.6E-03</b>	<b>8.6E-04</b>	0.6	<i>0.065</i>	<i>0.066</i>	<b>0.037</b>	1	Yes
	4	5.2E-04 (2.3E-04-1.2E-03)	1.8E-03 (6.3E-04-5.1E-03)	0.097 (5.4E-03-1.7)	<b>7.6E-05</b>	<i>0.076</i>	<b>1.2E-05</b>	<b>2.5E-03</b>	<b>2.4E-03</b>	1	<b>8.3E-04</b>	<i>0.098</i>	No
	5	2.9E-03 (8.2E-04-0.01)	0.3 (0.075-1.2)	3.3 (0.21-51)	<b>9.3E-07</b>	<b>8.4E-06</b>	<b>5.6E-06</b>	0.14	<b>4.0E-05</b>	<b>6.1E-04</b>	<b>4.3E-04</b>	1	No
IgG4:IgE (µg/L:µg/L)	6m	5.2E-03 (3.2E-03-8.5E-03)	4.4E-03 (3.2E-03-6.1E-03)	6.2E-03 (1.6E-03-0.024)	0.93	0.7	0.89	0.93	1	1	1	1	No*
	1	3.9E-03 (3.6E-03-4.3E-03)	4.0E-03 (3.8E-03-4.2E-03)	7.4E-03 (2.2E-03-0.024)	0.15	0.82	0.089	0.076	1	1	1	1	No*
	2	3.5E-03 (3.1E-03-3.9E-03)	0.011 (5.5E-03-0.023)	4.1E-03 (1.3E-03-0.013)	<b>3.1E-03</b>	<i>0.065</i>	<b>0.018</b>	<b>3.8E-03</b>	<i>0.069</i>	1	0.52	0.14	No*

<i>Food mix antibody</i>													
IgE (kU/L)	3	3.9E-03 (2.3E-03-6.5E-03)	0.025 (0.01-0.063)	0.018 (1.7E-03-0.19)	7.0E-03	4.5E-03	0.36	0.03	0.15	0.16	1	0.79	No*
IgG4 (µg/L)	4	0.013 (5.5E-03-0.03)	0.029 (0.011-0.079)	0.96 (0.046-20)	0.041	0.23	0.017	0.072	0.65	1	0.5	1	No
	5	0.054 (0.015-0.2)	4.3 (1.1-16)	16 (1.2-205)	1.6E-05	1.0E-05	7.0E-04	0.78	5.6E-04	7.2E-04	0.031	1	No
IgE (kU/L)	5	0.039 (0.028-0.054)	0.081 (0.055-0.12)	2.1 (0.66-6.9)	1.9E-09	6.9E-03	3.3E-09	3.6E-07	1.3E-07	0.23	5.4E-07	3.6E-05	No
IgG4 (µg/L)	5	758 (213-2.7E+03)	2.2E+04 (1.6E+04-2.8E+04)	1.1E+04 (1.2E+03-1.1E+05)	4.7E-14	2.3E-13	6.8E-07	0.5	8.1E-12	1.7E-10	6.4E-05	1	No
<i>Phadiatop Infant antibody</i>													
IgE (PAU/L)	6m	0.033 (0.025-0.044)	0.073 (0.053-0.1)	1 (0.36-2.9)	2.3E-10	2.8E-04	3.5E-10	2.4E-06	1.8E-08	0.014	7.6E-08	2.0E-04	Yes
	1	0.12 (0.092-0.15)	0.23 (0.18-0.3)	5.7 (3.2-10)	1.3E-14	1.7E-04	2.3E-12	1.8E-11	2.8E-12	9.2E-03	1.0E-09	5.8E-09	Yes
	2	0.088 (0.069-0.11)	0.29 (0.22-0.37)	15 (9.7-23)	1.6E-19	2.4E-09	4.7E-13	6.3E-13	1.3E-16	4.1E-07	3.0E-10	3.9E-10	Yes
	3	0.094 (0.066-0.13)	0.3 (0.21-0.43)	28 (16-49)	2.0E-16	5.7E-06	1.1E-12	2.2E-12	6.9E-14	4.3E-04	6.0E-10	9.7E-10	Yes
	4	0.088 (0.057-0.14)	0.46 (0.29-0.73)	43 (22-82)	4.9E-16	1.8E-07	8.7E-12	1.9E-10	1.3E-13	1.9E-05	3.1E-09	4.5E-08	No
	5#	0.085 (0.047-0.15)	0.35 (0.18-0.66)	35 (12-96)	2.1E-10	1.8E-03	9.0E-10	1.1E-07	1.7E-08	0.074	1.6E-07	1.3E-05	No
	10#	1.1 (0.5-2.6)	2.3 (1-5.4)	67 (36-125)	6.9E-06	0.15	2.0E-06	4.7E-05	2.5E-04	1	1.7E-04	2.9E-03	No
IgG4 (PAU/L)	6m	9.0E-03 (2.1E-03-0.039)	0.71 (0.16-3.2)	0.67 (0.019-23)	5.4E-04	1.7E-04	0.014	0.99	0.014	9.2E-03	0.42	1	Yes
	1	66 (21-203)	1.3E+03 (706-2.4E+03)	1.4E+03 (240-8.7E+03)	1.7E-09	1.9E-09	8.2E-05	0.58	1.2E-07	3.3E-07	4.7E-03	1	Yes
	2	64 (15-273)	1.0E+04 (5.7E+03-1.8E+04)	5.7E+03 (904-3.5E+04)	1.9E-16	1.2E-16	5.1E-06	0.69	6.9E-14	1.8E-13	3.9E-04	1	Yes
	3	677 (232-2.0E+03)	2.3E+04 (1.9E+04-2.9E+04)	2.5E+04 (1.4E+04-4.7E+04)	1.8E-20	3.9E-20	4.0E-08	0.96	2.4E-17	1.9E-16	5.0E-06	1	Yes
	4	575 (172-1.9E+03)	2.2E+04 (1.6E+04-2.8E+04)	2.4E+04 (1.3E+04-4.6E+04)	2.4E-16	5.2E-16	4.7E-07	0.81	7.8E-14	6.2E-13	4.6E-05	1	No
	5#	0.016 (3.0E-03-0.081)	59 (17-206)	643 (92-4.5E+03)	5.9E-14	3.7E-11	5.9E-10	3.6E-03	9.6E-12	1.1E-08	1.2E-07	0.13	No
IgG4:IgE (µg/L:µg/L)	6m	0.27 (0.063-1.2)	9.7 (2.1-44)	0.66 (0.023-19)	0.02	0.031	0.42	0.02	0.36	0.81	1	0.57	No*
	1	561 (179-1.8E+03)	5.6E+03 (3.0E+03-1.1E+04)	255 (44-1.5E+03)	4.3E-07	1.9E-04	0.012	2.0E-06	2.0E-05	0.01	0.37	1.7E-04	No*
	2	722 (170-3.1E+03)	3.5E+04 (1.9E+04-6.3E+04)	380 (57-2.5E+03)	1.8E-11	1.5E-06	2.0E-03	2.6E-10	1.8E-09	1.3E-04	0.08	5.8E-08	No*
	3	7.2E+03 (2.5E+03-2.1E+04)	7.6E+04 (5.0E+04-1.2E+05)	898 (347-2.3E+03)	1.7E-11	1.2E-05	1.7E-05	5.4E-10	1.8E-09	8.3E-04	1.1E-03	1.1E-07	No*
	4	6.5E+03 (2.0E+03-2.1E+04)	4.7E+04 (2.8E+04-7.8E+04)	564 (196-1.6E+03)	5.5E-08	0.016	4.2E-05	9.3E-09	2.9E-06	0.47	2.6E-03	1.3E-06	No
	5#	0.18 (0.036-0.88)	170 (47-612)	19 (1.9-178)	1.8E-07	2.0E-07	6.4E-03	7.0E-03	8.6E-06	2.1E-05	0.22	0.24	No
<i>Total antibody past atopy threshold</i>													
IgE ≥ 100 kU/L	6m	0% (0%-0%)	1.9% (0%-4.6%)	9.1% (0%-22%)	0.029	0.5	0.04	0.14	0.48	1	1	1	No*
	1	0% (0%-0%)	2.8% (0%-6%)	36% (15%-58%)	1.2E-07	0.26	9.1E-07	2.5E-05	5.9E-06	1	8.4E-05	1.6E-03	No*
	2	6.9% (1.5%-12%)	9.6% (3.9%-15%)	73% (53%-93%)	1.2E-10	0.6	6.3E-10	3.2E-09	1.1E-08	1	1.2E-07	5.3E-07	No*
	3	6% (0.8%-11%)	15% (7.8%-22%)	86% (71%-100%)	7.5E-14	0.061	1.5E-13	1.9E-10	1.2E-11	1	1.2E-10	4.5E-08	No*
	4	12% (4.3%-19%)	26% (17%-35%)	86% (69%-100%)	4.7E-10	0.019	1.8E-10	7.0E-07	3.6E-08	0.55	4.4E-08	6.6E-05	No
	5	27% (16%-38%)	35% (25%-46%)	94% (83%-100%)	4.8E-07	0.29	1.9E-07	6.1E-06	2.2E-05	1	2.0E-05	4.6E-04	No
	10	48% (34%-62%)	60% (47%-73%)	100% (100%-100%)	6.6E-04	0.25	3.7E-04	3.0E-03	0.017	1	0.018	0.12	No
<i>HDM antibody past atopy threshold</i>													
IgE ≥ 0.35 kU/L	6m	0% (0%-0%)	1.9% (0%-4.6%)	14% (0%-29%)	4.5E-03	0.5	7.5E-03	0.037	0.097	1	0.25	0.94	No*
	1	0% (0%-0%)	1.9% (0%-4.5%)	50% (27%-73%)	2.9E-11	0.5	2.0E-09	1.7E-08	2.8E-09	1	3.5E-07	2.3E-06	No*
	2	2.3% (0%-5.5%)	14% (7.6%-21%)	86% (71%-100%)	2.0E-16	3.9E-03	3.8E-16	1.2E-10	6.9E-14	0.14	4.8E-13	3.1E-08	No*
	3	12% (4.9%-19%)	19% (11%-27%)	91% (78%-100%)	2.1E-12	0.23	3.5E-12	2.1E-10	2.7E-10	1	1.4E-09	4.8E-08	No*
	4	12% (4.3%-19%)	34% (24%-44%)	90% (77%-100%)	2.0E-11	8.3E-04	1.1E-11	2.8E-06	2.0E-09	0.036	3.9E-09	2.3E-04	No
	5	23% (13%-33%)	39% (28%-50%)	89% (73%-100%)	1.1E-06	0.035	4.0E-07	1.5E-04	4.6E-05	0.9	4.0E-05	8.2E-03	No
	10	49% (35%-63%)	58% (45%-72%)	100% (100%-100%)	8.4E-04	0.44	3.7E-04	2.9E-03	0.021	1	0.018	0.11	No

<i>Cat antibody past atopy threshold</i>												
IgE ≥ 0.35 kU/L												
6m	1.2% (0%-3.5%)	1.9% (0%-4.6%)	9.1% (0%-22%)	0.17	1	0.1	0.14	1	1	1	1	No*
1	0% (0%-0%)	0% (0%-0%)	14% (0%-29%)	<b>9.4E-04</b>	1	<b>7.5E-03</b>	<b>4.4E-03</b>	<b>0.024</b>	1	0.25	0.16	No*
2	0% (0%-0%)	2.9% (0%-6.2%)	23% (3.7%-42%)	<b>9.8E-05</b>	0.25	<b>2.3E-04</b>	<b>4.2E-03</b>	<b>3.1E-03</b>	1	<b>0.012</b>	0.15	No*
3	3.6% (0%-7.7%)	2% (0%-4.7%)	23% (3.7%-42%)	<b>2.3E-03</b>	0.66	<b>9.6E-03</b>	<b>2.0E-03</b>	<i>0.053</i>	1	0.31	0.08	No*
4	0% (0%-0%)	5.7% (0.75%-11%)	24% (3.9%-44%)	<b>2.6E-04</b>	<i>0.061</i>	<b>2.8E-04</b>	<b>0.022</b>	<b>7.2E-03</b>	1	<b>0.014</b>	0.62	No
5	0% (0%-0%)	8.9% (2.5%-15%)	39% (14%-64%)	<b>3.3E-06</b>	<b>0.015</b>	<b>5.0E-06</b>	<b>3.8E-03</b>	<b>1.3E-04</b>	0.45	<b>3.9E-04</b>	0.14	No
10	17% (6.7%-28%)	20% (9.1%-31%)	71% (44%-98%)	<b>2.9E-04</b>	0.81	<b>2.3E-04</b>	<b>4.9E-04</b>	<b>8.0E-03</b>	1	<b>0.012</b>	<b>0.023</b>	No
<i>Peanut antibody past atopy threshold</i>												
IgE ≥ 0.35 kU/L												
6m	2.3% (0%-5.6%)	4.8% (0.63%-9%)	41% (19%-63%)	<b>1.5E-06</b>	0.46	<b>5.4E-06</b>	<b>3.5E-05</b>	<b>6.3E-05</b>	1	<b>4.2E-04</b>	<b>2.2E-03</b>	No*
1	3.5% (0%-7.4%)	5.6% (1.2%-10%)	68% (47%-89%)	<b>3.7E-12</b>	0.73	<b>1.3E-10</b>	<b>4.3E-10</b>	<b>4.6E-10</b>	1	<b>3.3E-08</b>	<b>9.1E-08</b>	No*
2	1.1% (0%-3.4%)	2.9% (0%-6.2%)	64% (42%-85%)	<b>1.8E-13</b>	0.63	<b>2.8E-11</b>	<b>1.3E-10</b>	<b>2.6E-11</b>	1	<b>8.3E-09</b>	<b>3.3E-08</b>	No*
3	4.8% (0.11%-9.5%)	0.99% (0%-3%)	45% (23%-68%)	<b>9.6E-09</b>	0.18	<b>1.4E-05</b>	<b>4.3E-08</b>	<b>5.7E-07</b>	1	<b>9.6E-04</b>	<b>5.4E-06</b>	No*
4	1.3% (0%-3.8%)	3.4% (0%-7.3%)	52% (29%-76%)	<b>1.3E-09</b>	0.62	<b>3.0E-08</b>	<b>2.5E-07</b>	<b>9.1E-08</b>	1	<b>3.9E-06</b>	<b>2.6E-05</b>	No
5	7.4% (0.99%-14%)	13% (5.2%-20%)	56% (30%-81%)	<b>3.7E-05</b>	0.41	<b>2.2E-05</b>	<b>2.6E-04</b>	<b>1.2E-03</b>	1	<b>1.4E-03</b>	<b>0.013</b>	No
10	21% (9.7%-33%)	18% (7.7%-29%)	64% (36%-93%)	<b>2.7E-03</b>	0.81	<b>3.4E-03</b>	<b>1.4E-03</b>	<i>0.062</i>	1	0.13	<i>0.059</i>	No
<i>Couch grass antibody past atopy threshold</i>												
IgE ≥ 0.35 kU/L												
6m	1.2% (0%-3.5%)	0.96% (0%-2.9%)	4.5% (0%-14%)	0.46	1	0.37	0.32	1	1	1	1	No*
1	0% (0%-0%)	0% (0%-0%)	4.5% (0%-14%)	0.1	1	0.2	0.17	1	1	1	1	No*
2	1.1% (0%-3.4%)	0.96% (0%-2.9%)	9.1% (0%-22%)	<i>0.054</i>	1	0.1	<i>0.079</i>	0.82	1	1	1	No*
3	2.4% (0%-5.8%)	2% (0%-4.7%)	36% (15%-58%)	<b>2.3E-06</b>	1	<b>3.9E-05</b>	<b>1.1E-05</b>	<b>9.1E-05</b>	1	<b>2.4E-03</b>	<b>7.8E-04</b>	No*
4	3.8% (0%-8.2%)	6.8% (1.4%-12%)	57% (34%-80%)	<b>2.0E-08</b>	0.5	<b>1.1E-07</b>	<b>1.0E-06</b>	<b>1.1E-06</b>	1	<b>1.3E-05</b>	<b>9.1E-05</b>	No
5	8.8% (1.9%-16%)	16% (8.1%-25%)	72% (49%-95%)	<b>2.7E-07</b>	0.22	<b>1.8E-07</b>	<b>8.6E-06</b>	<b>1.2E-05</b>	1	<b>1.9E-05</b>	<b>6.2E-04</b>	No
10	35% (21%-48%)	29% (17%-41%)	86% (65%-100%)	<b>3.6E-04</b>	0.68	<b>8.1E-04</b>	<b>1.7E-04</b>	<b>9.6E-03</b>	1	<b>0.036</b>	<b>9.2E-03</b>	No
<i>Ryegrass antibody past atopy threshold</i>												
IgE ≥ 0.35 kU/L												
6m	1.2% (0%-3.5%)	2.9% (0%-6.2%)	4.5% (0%-14%)	0.4	0.63	0.37	0.54	1	1	1	1	No*
1	0% (0%-0%)	0% (0%-0%)	4.5% (0%-14%)	0.1	1	0.2	0.17	1	1	1	1	No*
2	13% (5.5%-20%)	7.7% (2.5%-13%)	32% (11%-53%)	<b>0.012</b>	0.33	<i>0.05</i>	<b>5.0E-03</b>	0.23	1	1	0.18	No*
3	6% (0.8%-11%)	3% (0%-6.3%)	50% (27%-73%)	<b>5.6E-08</b>	0.47	<b>6.9E-06</b>	<b>1.3E-07</b>	<b>2.9E-06</b>	1	<b>5.1E-04</b>	<b>1.5E-05</b>	No*
4	5.1% (0.12%-10%)	10% (3.8%-17%)	71% (50%-92%)	<b>1.6E-10</b>	0.26	<b>7.4E-10</b>	<b>3.9E-08</b>	<b>1.4E-08</b>	1	<b>1.4E-07</b>	<b>5.0E-06</b>	No
5	10% (2.8%-17%)	22% (12%-31%)	78% (57%-99%)	<b>9.5E-08</b>	<i>0.074</i>	<b>3.8E-08</b>	<b>1.2E-05</b>	<b>4.9E-06</b>	1	<b>4.9E-06</b>	<b>8.3E-04</b>	No
10	40% (27%-54%)	40% (27%-53%)	86% (65%-100%)	<b>5.7E-03</b>	1	<b>5.3E-03</b>	<b>2.7E-03</b>	0.12	1	0.19	0.11	No
<i>Mould antibody past atopy threshold</i>												
IgE ≥ 0.35 kU/L												
6m	0% (0%-0%)	0.96% (0%-2.9%)	4.5% (0%-14%)	0.2	1	0.2	0.32	1	1	1	1	No*
1	0% (0%-0%)	0% (0%-0%)	0% (0%-0%)	1	1	1	1	1	1	1	1	No*
2	1.1% (0%-3.4%)	1.9% (0%-4.6%)	0% (0%-0%)	1	1	1	1	1	1	1	1	No*
3	7.2% (1.5%-13%)	4% (0.091%-7.8%)	14% (0%-29%)	0.16	0.35	0.39	0.11	1	1	1	1	No*
4	1.3% (0%-3.8%)	9.1% (3%-15%)	9.5% (0%-23%)	<b>0.047</b>	<b>0.037</b>	0.11	1	0.72	0.94	1	1	No
5	2.9% (0%-7.1%)	10% (3.3%-17%)	22% (0.95%-43%)	<b>0.027</b>	0.11	<b>0.016</b>	0.23	0.46	1	0.47	1	No
10	15% (5.2%-26%)	20% (9.1%-31%)	79% (54%-100%)	<b>1.9E-05</b>	0.62	<b>1.7E-05</b>	<b>7.9E-05</b>	<b>6.5E-04</b>	1	<b>1.1E-03</b>	<b>4.6E-03</b>	No

Feature? = whether variable was used as a clustering feature or not; geom. mean = geometric mean; HDM = house dust mite; P-value (adj.) = adjusted P-values (Benjamini-Yekutieli method). For categorical variables, associations were tested using Fisher exact test; for continuous variables, Kruskal-Wallis and Mann-Whitney-Wilcoxon. Bold text indicates statistical significance ( $p<0.05$ ); italics indicate near-significance ( $p<0.10$ ). \*Not used as clustering features, as these were derived variables. #Assay used at age 5 was the adult version, not Phadiatop infant. Therefore the standard unit (PAU) and specificities may differ between the two.

### C. Immunological (cytokines)

Variable	Age	CAS1 (N=88)		CAS2 (N=107)		CAS3 (N=22)		P-value		P-value (adj.)		Feature?		
		Geom. mean (95% CI)	Geom. mean (95% CI)	Geom. mean (95% CI)	Geom. mean (95% CI)	Geom. mean (95% CI)	Geom. mean (95% CI)	Overall	1 vs. 2	1 vs. 3	2 vs. 3	1 vs. 2	1 vs. 3	2 vs. 3
<i>HDM cytokine response</i>														
<i>above controls</i>														
IL-13 mRNA <sup>^</sup>	0	1.7E-03 (1.1E-04-0.026)	6.0E-03 (4.8E-04-0.075)	6.7E-03 (3.3E-05-1.4)	0.85	0.6	0.68	0.94	1	1	1	1	1	No
	6m	1.0E-04 (8.8E-06-1.1E-03)	3.2E-04 (3.8E-05-2.6E-03)	2 (0.015-266)	<b>3.2E-04</b>	0.5	<b>1.7E-04</b>	<b>3.8E-04</b>	<b>8.7E-03</b>	1	<b>9.2E-03</b>	<b>0.018</b>	<b>0.018</b>	No
	1	2.2E-05 (2.9E-06-1.6E-04)	4.9E-05 (5.4E-06-4.4E-04)	5.8E-03 (3.2E-05-1.1)	<b>0.026</b>	0.57	<b>8.4E-03</b>	<b>0.037</b>	0.45	1	0.27	0.94	No	
	2	1.9E-06 (7.7E-07-4.7E-06)	1.7E-06 (8.2E-07-3.3E-06)	2.2E-05 (6.4E-07-7.5E-04)	<b>0.04</b>	0.83	<i>0.054</i>	<b>0.023</b>	0.64	1	1	0.64	No	
	3	1.5E-06 (6.7E-07-3.3E-06)	1.4E-06 (7.2E-07-2.7E-06)	1.0E-06 (1.0E-06-1.0E-06)	0.87	0.9	0.62	0.65	1	1	1	1	No	
IL-4 mRNA <sup>^</sup>	5	0.036 (1.6E-03-0.8)	0.11 (8.8E-03-1.4)	2.9E+03 (742-1.1E+04)	<b>6.8E-05</b>	0.59	<b>9.9E-05</b>	<b>2.5E-05</b>	<b>2.2E-03</b>	1	<b>5.6E-03</b>	<b>1.6E-03</b>	No	
	0	1.4E-06 (6.9E-07-3.0E-06)	1.9E-06 (7.8E-07-4.4E-06)	1.0E-06 (1.0E-06-1.0E-06)	0.71	0.65	0.6	0.47	1	1	1	1	No	
	6m	4.6E-06 (1.0E-06-2.1E-05)	5.1E-06 (1.4E-06-1.8E-05)	0.54 (6.5E-03-44)	<b>6.2E-09</b>	0.94	<b>4.7E-07</b>	<b>1.0E-07</b>	<b>3.8E-07</b>	1	<b>4.6E-05</b>	<b>1.2E-05</b>	No	
	1	4.2E-06 (1.0E-06-1.7E-05)	1.8E-05 (2.8E-06-1.2E-04)	8.9E-04 (8.1E-06-0.099)	<b>9.3E-03</b>	0.22	<b>2.9E-03</b>	<b>0.035</b>	0.19	1	0.11	0.9	No	
	2	1.2E-05 (2.3E-06-6.1E-05)	7.9E-06 (2.0E-06-3.1E-05)	0.01 (1.0E-04-1)	<b>1.3E-04</b>	0.68	<b>5.5E-04</b>	<b>1.0E-04</b>	<b>4.0E-03</b>	1	<b>0.025</b>	<b>5.6E-03</b>	No	
IL-4R mRNA <sup>^</sup>	3	2.1E-06 (7.3E-07-6.2E-06)	3.7E-06 (1.0E-06-1.4E-05)	3.9E-06 (2.0E-07-7.7E-05)	0.8	0.53	0.65	0.96	1	1	1	1	No	
	5	2.3E-04 (1.7E-05-3.0E-03)	4.7E-04 (5.3E-05-4.3E-03)	5.3 (0.082-345)	<b>4.9E-04</b>	0.72	<b>4.5E-04</b>	<b>3.2E-04</b>	<b>0.013</b>	1	<b>0.021</b>	<b>0.016</b>	No	
	6m	0.011 (6.1E-04-0.19)	0.028 (3.0E-03-0.26)	0.23 (1.1E-03-51)	0.85	0.91	0.67	0.57	1	1	1	1	No	
	1	4.0E-03 (2.6E-04-0.062)	9.5E-03 (6.7E-04-0.14)	1.5E-03 (9.3E-06-0.25)	0.85	0.79	0.7	0.59	1	1	1	1	No	
	2	1.5E-05 (2.8E-06-8.1E-05)	1.2E-05 (2.8E-06-5.3E-05)	7.1E-06 (4.1E-07-1.2E-04)	0.91	0.92	0.66	0.75	1	1	1	1	No	
IL-5 mRNA <sup>^</sup>	3	3.2E-03 (1.9E-04-0.055)	0.012 (8.5E-04-0.17)	2.4 (9.0E-03-642)	<b>0.023</b>	0.54	<b>7.6E-03</b>	<b>0.019</b>	0.4	1	0.25	0.55	No	
	5	0.27 (0.013-5.5)	1 (0.1-11)	3.2 (5.7E-03-1.8E+03)	<b>0.095</b>	0.75	<b>0.035</b>	0.05	1	1	0.9	1	No	
	0	2.5E-04 (2.1E-05-2.9E-03)	2.6E-04 (2.8E-05-2.5E-03)	1.2E-05 (3.1E-07-4.6E-04)	0.47	0.96	0.24	0.25	1	1	1	1	No	
	6m	5.2E-05 (5.6E-06-4.8E-04)	3.1E-05 (5.2E-06-1.8E-04)	0.33 (1.3E-03-83)	<b>1.5E-04</b>	0.85	<b>2.3E-04</b>	<b>1.1E-04</b>	<b>4.5E-03</b>	1	<b>0.012</b>	<b>6.1E-03</b>	No	
	1	4.6E-06 (1.0E-06-2.0E-05)	2.2E-05 (2.9E-06-1.7E-04)	0.015 (1.0E-04-2.4)	<b>2.2E-04</b>	0.2	<b>7.2E-05</b>	<b>3.6E-03</b>	<b>6.3E-03</b>	1	<b>4.2E-03</b>	0.13	No	
IL-9 mRNA <sup>^</sup>	2	4.6E-06 (1.2E-06-1.7E-05)	3.7E-06 (1.2E-06-1.1E-05)	2.0E-03 (1.6E-05-0.25)	<b>1.7E-04</b>	0.8	<b>7.5E-04</b>	<b>2.0E-04</b>	<b>5.0E-03</b>	1	<b>0.033</b>	<b>0.011</b>	No	
	3	2.2E-06 (7.2E-07-7.0E-06)	1.0E-06 (1.0E-06-1.0E-06)	1.9E-05 (2.5E-07-1.4E-03)	<b>0.027</b>	0.13	0.16	<b>4.8E-03</b>	0.46	1	1	0.17	No	
	5	0.021 (9.9E-04-0.43)	0.07 (5.7E-03-0.85)	246 (7-8.7E+03)	<b>1.3E-04</b>	0.49	<b>7.1E-05</b>	<b>1.1E-04</b>	<b>4.0E-03</b>	1	<b>4.2E-03</b>	<b>6.1E-03</b>	No	
	0	0.021 (1.0E-03-0.43)	6.0E-03 (4.2E-04-0.085)	5.9E-04 (3.4E-06-0.1)	0.5	0.67	0.28	0.33	1	1	1	1	No	
	6m	7.0E-03 (3.8E-04-0.13)	3.1E-03 (3.1E-04-0.032)	2 (3.4E-03-1.1E+03)	<b>0.027</b>	0.6	<b>0.022</b>	<b>8.8E-03</b>	0.46	1	0.62	0.29	No	
IFN- $\gamma$ mRNA <sup>^</sup>	1	8.1E-05 (7.7E-06-8.5E-04)	4.9E-04 (3.5E-05-6.9E-03)	3.6 (0.072-175)	<b>2.3E-04</b>	0.31	<b>6.9E-05</b>	<b>1.6E-03</b>	<b>6.5E-03</b>	1	<b>4.1E-03</b>	<i>0.066</i>	No	
	2	2.0E-06 (7.6E-07-5.0E-06)	2.4E-05 (4.5E-06-1.3E-04)	1.3E-03 (7.6E-06-0.23)	<b>1.2E-03</b>	<b>0.016</b>	<b>2.1E-04</b>	<b>0.041</b>	<b>0.029</b>	0.47	<b>0.011</b>	1	No	
	3	5.8E-06 (1.0E-06-3.2E-05)	1.4E-05 (2.1E-06-9.3E-05)	2.5E-03 (4.6E-06-1.4)	<b>0.02</b>	0.5	<b>7.8E-03</b>	<b>0.029</b>	0.36	1	0.26	0.77	No	
	5	3.3E-03 (1.0E-04-0.11)	9.6E-03 (5.4E-04-0.17)	1.5E+03 (4.2-5.4E+05)	<b>3.5E-04</b>	0.77	<b>2.4E-04</b>	<b>2.5E-04</b>	<b>9.4E-03</b>	1	<b>0.012</b>	<b>0.013</b>	No	
	0	0.68 (0.031-15)	2.5 (0.18-34)	0.1 (2.4E-04-45)	0.67	0.9	0.48	0.37	1	1	1	1	No	
6m	0.46 (0.024-9)	0.15 (0.012-1.8)	23 (0.074-6.9E+03)	<b>0.043</b>	0.46	<b>0.045</b>	<b>0.014</b>	0.68	1	1	0.42	No		
1	1.2E-03 (7.0E-05-0.02)	6.9E-04 (4.3E-05-0.011)	0.054 (1.8E-04-16)	0.14	0.71	<i>0.086</i>	<i>0.066</i>	1	1	1	1	No		
2	0.09 (8.3E-03-0.97)	0.31 (0.04-2.4)	0.068 (4.8E-04-9.6)	0.46	0.24	0.43	0.88	1	1	1	1	No		
3	3.6E-06 (8.4E-07-1.5E-05)	5.7E-06 (1.3E-06-2.6E-05)	1.8E-05 (2.6E-07-1.3E-03)	0.63	0.66	0.33	0.55	1	1	1	1	No		

IL-13 protein (pg/ml) <sup>Δ</sup>	5	0.013 (5.4E-04-0.31)	0.056 (3.8E-03-0.84)	8.5 (0.034-2.1E+03)	0.1	0.33	<b>0.028</b>	0.15	1	1	0.75	1	No	
	0	0.22 (0.066-0.73)	0.22 (0.076-0.63)	0.085 (0.011-0.66)	0.68	0.76	0.41	0.45	1	1	1	1	No	
	6m	0.064 (0.022-0.18)	0.06 (0.025-0.14)	19 (1.4-244)	<b>4.6E-06</b>	0.98	<b>1.7E-05</b>	<b>4.1E-06</b>	<b>1.7E-04</b>	1	<b>1.1E-03</b>	<b>3.3E-04</b>	No	
	1	0.044 (0.016-0.12)	0.031 (0.014-0.07)	0.82 (0.071-9.5)	<b>2.8E-03</b>	0.53	<b>6.5E-03</b>	<b>1.5E-03</b>	<b>0.063</b>	1	0.22	<b>0.063</b>	No	
	2	0.021 (0.011-0.04)	0.024 (0.013-0.046)	0.57 (0.058-5.6)	<b>1.8E-04</b>	0.79	<b>3.1E-04</b>	<b>4.5E-04</b>	<b>5.3E-03</b>	1	<b>0.015</b>	<b>0.021</b>	No	
	3	0.03 (0.014-0.068)	0.035 (0.017-0.074)	0.37 (0.029-4.8)	<b>0.027</b>	0.78	<b>0.016</b>	<b>0.017</b>	0.46	1	0.47	0.5	No	
	5	0.13 (0.046-0.37)	0.32 (0.11-0.87)	12 (1.2-117)	<b>2.1E-04</b>	0.29	<b>7.7E-05</b>	<b>5.1E-04</b>	<b>6.1E-03</b>	1	<b>4.5E-03</b>	<b>0.024</b>	No	
	0	0.043 (0.018-0.11)	0.026 (0.013-0.052)	0.018 (5.0E-03-0.068)	0.44	0.36	0.29	0.57	1	1	1	1	No	
	6m	0.018 (9.2E-03-0.034)	0.013 (8.9E-03-0.02)	0.21 (0.012-3.7)	<b>7.9E-04</b>	0.4	<b>8.1E-03</b>	<b>3.5E-04</b>	<b>0.02</b>	1	0.27	<b>0.017</b>	No	
	1	0.012 (8.5E-03-0.016)	0.012 (8.3E-03-0.017)	0.017 (5.6E-03-0.049)	0.58	0.92	0.41	0.36	1	1	1	1	No	
IL-10 protein (pg/ml) <sup>Δ</sup>	2	0.011 (8.7E-03-0.015)	0.021 (0.012-0.039)	0.044 (8.0E-03-0.25)	<b>0.054</b>	<b>0.088</b>	<b>0.011</b>	0.28	1	0.34	1	1	No	
	3	0.016 (9.4E-03-0.026)	0.019 (0.011-0.033)	0.21 (0.018-2.4)	<b>2.1E-03</b>	0.65	<b>1.9E-03</b>	<b>3.9E-03</b>	<b>0.05</b>	1	<b>0.077</b>	0.14	No	
	5	0.028 (0.014-0.057)	0.042 (0.02-0.087)	2.3 (0.25-22)	<b>3.2E-06</b>	0.45	<b>5.7E-06</b>	<b>2.0E-05</b>	<b>1.2E-04</b>	1	<b>4.3E-04</b>	<b>1.3E-03</b>	No	
	0	0.089 (0.032-0.24)	0.19 (0.066-0.53)	0.35 (0.038-3.2)	0.32	0.26	0.17	0.57	1	1	1	1	No	
	6m	0.055 (0.02-0.15)	0.026 (0.013-0.051)	0.19 (0.012-3)	<b>0.09</b>	0.19	0.28	<b>0.032</b>	1	1	1	0.84	No	
	1	0.02 (0.01-0.04)	0.014 (8.8E-03-0.021)	0.01 (0.01-0.01)	0.33	0.3	0.24	0.46	1	1	1	1	No	
	2	0.011 (8.7E-03-0.015)	0.011 (8.9E-03-0.014)	0.01 (0.01-0.01)	0.87	0.91	0.61	0.64	1	1	1	1	No	
	3	0.01 (0.01-0.01)	0.011 (8.9E-03-0.014)	0.01 (0.01-0.01)	0.59	0.37	NA	0.66	1	1	NA	1	No	
	5	0.019 (0.011-0.032)	0.015 (0.01-0.023)	0.015 (6.2E-03-0.039)	0.82	0.55	0.75	1	1	1	1	1	No	
	0	0.085 (0.028-0.26)	0.081 (0.03-0.22)	0.057 (7.8E-03-0.42)	0.95	0.89	0.72	0.85	1	1	1	1	No	
IFN-γ protein (pg/ml) <sup>Δ</sup>	6m	0.094 (0.028-0.32)	0.075 (0.028-0.2)	0.46 (0.012-17)	0.43	0.82	0.28	0.21	1	1	1	1	No	
	1	0.038 (0.015-0.1)	0.019 (0.01-0.035)	0.055 (7.8E-03-0.38)	0.3	0.2	0.73	0.17	1	1	1	1	No	
	2	0.012 (8.4E-03-0.017)	0.01 (0.01-0.01)	0.01 (0.01-0.01)	0.47	0.28	0.61	NA	1	1	1	NA	No	
	3	0.029 (0.013-0.061)	0.022 (0.012-0.04)	0.038 (5.3E-03-0.27)	0.74	0.53	0.87	0.53	1	1	1	1	No	
	5	0.023 (0.012-0.045)	0.049 (0.022-0.11)	0.017 (5.5E-03-0.053)	0.23	0.16	0.65	0.23	1	1	1	1	No	
	<i>Cat cytokine response above control<sup>Δ</sup></i>													
	IL-13 protein (pg/ml) <sup>Δ</sup> to cat to Fel dl	0	2.8 (0.86-9.3)	1.6 (0.48-5.2)	1.3 (0.12-13)	0.78	0.63	0.52	0.77	1	1	1	1	No
		0	0.27 (0.057-1.3)	0.24 (0.04-1.4)	0.65 (2.9E-04-1.4E+03)	0.97	0.97	0.84	0.86	1	1	1	1	No
	IL-5 protein (pg/ml) <sup>Δ</sup> to cat to Fel dl	5	0.022 (0.01-0.049)	0.023 (0.011-0.047)	0.054 (4.3E-03-0.66)	0.61	0.98	0.38	0.37	1	1	1	1	No
		0	0.29 (0.092-0.93)	0.77 (0.25-2.4)	1.2 (0.12-12)	0.32	0.22	0.2	0.7	1	1	1	1	No
IL-10 protein (pg/ml) <sup>Δ</sup> to cat to Fel dl	5	0.012 (8.2E-03-0.018)	0.038 (0.01-0.14)	0.068 (1.5E-04-30)	0.89	0.67	0.97	0.77	1	1	1	1	No	
	0	14 (5.2-36)	8.5 (3.2-22)	20 (4.1-101)	<b>0.086</b>	0.73	<b>0.049</b>	<b>0.083</b>	1	1	1	1	No	
IFN-γ protein (pg/ml) <sup>Δ</sup> to cat to Fel dl	0	0.1 (0.026-0.42)	0.11 (0.021-0.52)	0.01 (0.01-0.01)	0.48	0.97	0.24	0.26	1	1	1	1	No	
	5	0.01 (0.01-0.01)	0.01 (0.01-0.01)	0.01 (0.01-0.01)	NA	NA	NA	NA	NA	NA	NA	NA	No	
IFN-γ protein (pg/ml) <sup>Δ</sup> to cat to Fel dl	0	4.5 (1.2-16)	4.5 (1.4-15)	0.98 (0.075-13)	0.44	0.87	0.21	0.25	1	1	1	1	No	
	0	0.091 (0.022-0.37)	0.17 (0.029-1)	0.66 (3.0E-04-1.5E+03)	0.56	0.56	0.35	0.48	1	1	1	1	No	
5	0.013 (7.7E-03-0.021)	0.014 (8.5E-03-0.024)	0.01 (0.01-0.01)	0.78	0.74	0.63	0.54	1	1	1	1	No		
<i>Peanut cytokine response above control<sup>Δ</sup></i>														
IL-13 protein (pg/ml) <sup>Δ</sup> to Ara h2 to peanut	6m	0.01 (0.01-0.01)	0.01 (0.01-0.01)	0.033 (1.8E-03-0.63)	<b>7.8E-03</b>	NA	<b>0.059</b>	<b>0.018</b>	0.16	NA	1	0.52	No	
	1	0.012 (8.2E-03-0.018)	0.014 (8.7E-03-0.022)	0.01 (0.01-0.01)	0.69	0.65	0.58	0.45	1	1	1	1	No	
IL-5 protein (pg/ml) <sup>Δ</sup> to Ara h2 to peanut	2	0.14 (0.04-0.47)	0.05 (0.021-0.12)	0.043 (4.9E-03-0.38)	0.27	0.13	0.31	0.92	1	1	1	1	No	
	5	0.02 (0.01-0.04)	0.033 (0.016-0.068)	0.12 (0.011-1.3)	0.15	0.36	<b>0.052</b>	0.18	1	1	1	1	No	
6m	0.01 (0.01-0.01)	0.012 (8.1E-03-0.019)	0.033 (1.8E-03-0.61)	0.12	0.43	<b>0.059</b>	0.18	1	1	1	1	No		

	1	0.01 (0.01-0.01)	0.01 (0.01-0.01)	0.01 (0.01-0.01)	NA	NA	NA	NA	NA	NA	NA	NA	No
to peanut	2	0.015 (8.6E-03-0.025)	0.018 (0.01-0.032)	0.023 (3.5E-03-0.16)	0.8	0.62	0.55	0.78	1	1	1	1	No
	5	0.01 (0.01-0.01)	0.011 (8.8E-03-0.014)	0.02 (4.5E-03-0.088)	0.18	0.39	0.073	0.25	1	1	1	1	No
IL-10 protein (pg/ml) <sup>^</sup> to Ara h2	6m	0.19 (0.04-0.89)	0.21 (0.057-0.74)	0.1 (2.6E-03-4.2)	0.94	0.82	0.94	0.76	1	1	1	1	No
to peanut	1	0.01 (0.01-0.01)	0.012 (8.3E-03-0.017)	0.01 (0.01-0.01)	0.56	0.37	NA	0.61	1	1	NA	1	No
	2	0.01 (0.01-0.01)	0.018 (0.01-0.031)	0.01 (0.01-0.01)	0.15	0.082	NA	0.38	1	1	NA	1	No
	5	0.028 (0.013-0.06)	0.017 (0.01-0.029)	0.018 (5.2E-03-0.06)	0.57	0.32	0.58	0.97	1	1	1	1	No
IFN- $\gamma$ protein (pg/ml) <sup>^</sup> to Ara h2	6m	3.3 (0.45-23)	10 (2.4-45)	1 (4.8E-03-212)	0.64	0.4	0.86	0.57	1	1	1	1	No
to peanut	1	0.012 (8.0E-03-0.02)	0.014 (8.7E-03-0.023)	0.01 (0.01-0.01)	0.7	0.67	0.58	0.45	1	1	1	1	No
	2	0.012 (8.3E-03-0.017)	0.042 (0.017-0.1)	0.022 (3.8E-03-0.13)	0.082	<b>0.026</b>	0.29	0.62	1	1	1	1	No
	5	0.014 (8.7E-03-0.023)	0.017 (0.01-0.027)	0.066 (7.5E-03-0.58)	0.074	0.66	<b>0.04</b>	0.063	1	1	1	1	No
<i>Ovalbumin cytokine response</i>													
<i>above control<sup>^</sup></i>													
IL-13 protein (pg/ml) <sup>^</sup>	0	1.2 (0.36-3.9)	1.2 (0.39-3.5)	0.13 (0.016-1.1)	0.21	0.84	0.098	0.097	1	1	1	1	No
	6m	0.021 (9.0E-03-0.049)	0.018 (9.3E-03-0.033)	0.49 (0.014-17)	<b>3.3E-03</b>	0.77	<b>9.3E-03</b>	<b>1.9E-03</b>	0.074	0.3	0.077	0.077	No
	1	0.033 (0.013-0.083)	0.03 (0.014-0.065)	0.065 (7.6E-03-0.56)	0.62	0.86	0.42	0.36	1	1	1	1	No
	2	0.11 (0.033-0.35)	0.055 (0.022-0.14)	0.13 (6.8E-03-2.5)	0.58	0.39	0.78	0.43	1	1	1	1	No
	5	0.14 (0.047-0.41)	0.12 (0.045-0.33)	0.052 (7.8E-03-0.34)	0.7	0.83	0.42	0.46	1	1	1	1	No
IL-5 protein (pg/ml) <sup>^</sup>	0	0.13 (0.045-0.39)	0.12 (0.043-0.34)	0.077 (0.011-0.55)	0.88	0.88	0.68	0.63	1	1	1	1	No
	6m	0.013 (7.8E-03-0.021)	0.01 (0.01-0.01)	0.055 (4.1E-03-0.74)	<b>5.1E-03</b>	0.27	0.058	<b>2.0E-03</b>	0.11	1	1	0.08	No
	1	0.01 (0.01-0.01)	0.01 (0.01-0.01)	0.01 (0.01-0.01)	NA	NA	NA	NA	NA	NA	NA	NA	No
	2	0.018 (9.2E-03-0.036)	0.015 (9.4E-03-0.025)	0.01 (0.01-0.01)	0.67	0.71	0.4	0.46	1	1	1	1	No
	5	0.014 (8.7E-03-0.023)	0.02 (0.011-0.037)	0.017 (5.5E-03-0.051)	0.61	0.33	0.65	0.81	1	1	1	1	No
IL-10 protein (pg/ml) <sup>^</sup>	0	1.5 (0.44-4.8)	6 (2.2-16)	6.8 (0.84-55)	0.12	0.069	0.13	0.67	1	1	1	1	No
	6m	0.07 (0.02-0.25)	0.14 (0.043-0.46)	0.18 (6.4E-03-5.3)	0.55	0.37	0.4	0.66	1	1	1	1	No
	1	0.2 (0.056-0.7)	0.33 (0.1-1.1)	0.25 (0.018-3.3)	0.73	0.42	0.8	0.85	1	1	1	1	No
	2	0.031 (0.013-0.075)	0.05 (0.021-0.12)	0.023 (3.4E-03-0.16)	0.6	0.41	0.83	0.48	1	1	1	1	No
	5	3 (0.97-9.5)	2.9 (0.95-8.7)	1.1 (0.087-14)	0.72	0.76	0.51	0.45	1	1	1	1	No
IFN- $\gamma$ protein (pg/ml) <sup>^</sup>	0	0.94 (0.23-3.8)	1.2 (0.33-4.5)	0.91 (0.074-11)	0.94	0.88	0.78	0.74	1	1	1	1	No
	6m	0.19 (0.041-0.88)	0.4 (0.11-1.5)	0.21 (6.2E-03-7.4)	0.79	0.48	0.82	0.94	1	1	1	1	No
	1	0.3 (0.072-1.3)	0.39 (0.11-1.4)	0.49 (0.034-6.9)	0.97	0.81	0.88	0.93	1	1	1	1	No
	2	0.024 (0.01-0.056)	0.082 (0.029-0.23)	0.074 (3.6E-03-1.5)	0.23	0.093	0.33	0.95	1	1	1	1	No
	5	0.14 (0.041-0.44)	0.25 (0.083-0.77)	0.058 (7.7E-03-0.44)	0.45	0.43	0.54	0.25	1	1	1	1	No
<i>Ryegrass cytokine response</i>													
<i>above control<sup>^</sup></i>													
IL-13 protein (pg/ml) <sup>^</sup>	2	0.52 (0.14-1.9)	0.076 (0.029-0.2)	1.1 (0.053-24)	<b>0.038</b>	<b>0.029</b>	0.54	<b>0.046</b>	0.62	0.77	1	1	No
	3	0.068 (0.023-0.2)	0.059 (0.024-0.14)	1.7 (0.12-25)	<b>5.2E-03</b>	0.88	<b>4.7E-03</b>	<b>2.7E-03</b>	0.11	0.17	0.11	0.11	No
	5	0.098 (0.036-0.26)	0.087 (0.035-0.22)	0.51 (0.042-6.1)	0.27	0.86	0.16	0.11	1	1	1	1	No
IL-5 protein (pg/ml) <sup>^</sup>	2	0.02 (0.01-0.041)	0.017 (0.01-0.027)	0.021 (4.0E-03-0.11)	0.88	0.67	0.95	0.73	1	1	1	1	No
	3	0.01 (0.01-0.01)	0.015 (9.5E-03-0.023)	0.062 (7.5E-03-0.52)	<b>6.1E-03</b>	0.15	<b>1.9E-03</b>	<b>0.039</b>	0.13	1	0.077	0.98	No
	5	0.02 (0.011-0.035)	0.024 (0.013-0.045)	0.045 (8.0E-03-0.25)	0.48	0.65	0.22	0.41	1	1	1	1	No
IL-10 protein (pg/ml) <sup>^</sup>	2	0.097 (0.033-0.29)	0.077 (0.031-0.19)	0.04 (5.0E-03-0.33)	0.68	0.79	0.37	0.48	1	1	1	1	No
	3	0.49 (0.13-1.8)	1.8 (0.61-5.1)	2.9 (0.23-38)	0.18	0.14	0.13	0.38	1	1	1	1	No
	5	0.14 (0.05-0.39)	0.18 (0.067-0.48)	0.27 (0.026-2.8)	0.81	0.73	0.58	0.62	1	1	1	1	No
IFN- $\gamma$ protein (pg/ml) <sup>^</sup>	2	0.18 (0.052-0.64)	0.28 (0.092-0.85)	0.11 (6.8E-03-1.9)	0.84	0.72	0.76	0.6	1	1	1	1	No



3	0.46 (0.12-1.8)	1 (0.32-3.3)	9.9 (0.66-148)	0.061	0.25	<b>0.017</b>	0.11	0.9	1	0.5	1	No
5	0.092 (0.035-0.24)	0.073 (0.029-0.18)	0.11 (0.011-1.1)	0.91	0.76	0.83	0.7	1	1	1	1	No

Feature? = whether variable was used as a clustering feature or not; geom. mean = geometric mean; HDM = house dust mite; PBMC = peripheral blood mononuclear cells; P-value (adj.) = adjusted P-values (Benjamini-Yekutieli method); prop. = proportion. For categorical variables, associations were tested using Fisher exact test; for continuous variables, Kruskal-Wallis and Mann-Whitney-Wilcoxon. Bold text indicates statistical significance ( $p < 0.05$ ); italics indicate near-significance ( $p < 0.10$ ). Note that none of these variables were used as clustering features. ^PBMC cytokine responses to HDM above unstimulated control; birth samples (age 0) taken from cord blood (CBMC).

## D. Immunological (SPT)

Variable	Age	CAS1 (N=88)		CAS2 (N=107)		CAS3 (N=22)		P-value		P-value (adj.)			Feature?
		Mean	(95% CI)	Mean	(95% CI)	Mean	(95% CI)	Overall	1 vs. 2	1 vs. 3	2 vs. 3	Overall	
<i>Wheat size SPT</i>													
Histamine (mm)	6m	1.8 (1.7-2)	1.8 (1.7-1.9)	1.9 (1.6-2.2)	0.87	0.68	0.59	1	1	1	1	1	Yes
	2	3.4 (3.1-3.8)	3.4 (3.2-3.7)	4.1 (3.1-5.1)	0.4	0.66	0.24	1	1	1	1	1	Yes
	5	2.5 (2.3-2.6)	2.6 (2.5-2.8)	2.8 (2.2-3.4)	0.65	0.39	0.59	1	1	1	1	1	No
HDM (mm)	10	4.2 (3.9-4.4)	4.2 (3.9-4.5)	4.4 (3.8-5.1)	0.6	0.78	0.41	0.32	1	1	1	1	No
	6m	0.091 (0-0.18)	0.051 (0-0.12)	0.34 (0-0.75)	<b>0.04</b>	0.29	0.11	0.01	1	1	1	0.32	Yes
	2	0.44 (0.17-0.72)	0.67 (0.35-0.99)	6.2 (4.5-7.9)	<b>5.9E-15</b>	0.54	<b>3.9E-13</b>	<b>1.4E-12</b>	1	<b>2.6E-10</b>	<b>6.7E-10</b>	<b>6.7E-10</b>	Yes
Cat (mm)	5	1.1 (0.6-1.5)	1.9 (1.4-2.5)	5.5 (4.2-6.9)	<b>1.3E-08</b>	<b>0.026</b>	<b>5.9E-09</b>	<b>7.6E-07</b>	0.71	<b>9.1E-07</b>	<b>1.7E-04</b>	No	
	10	4.1 (3.4-4.8)	4.8 (4.2-5.3)	7 (4.9-9.1)	<b>2.2E-03</b>	<b>0.046</b>	<b>1.7E-03</b>	<b>0.016</b>	1	<i>0.051</i>	0.07	No	
	6m	0.091 (0-0.2)	0.11 (0.012-0.21)	0.32 (0-0.74)	0.15	0.67	0.064	0.12	1	1	1	Yes	
Ryegrass (mm)	2	0.35 (0.18-0.51)	0.4 (0.22-0.58)	1.4 (0.56-2.3)	<b>9.6E-03</b>	0.89	<b>4.5E-03</b>	<b>5.8E-03</b>	1	0.16	0.2	Yes	
	5	0.33 (0.16-0.49)	0.56 (0.32-0.79)	1.9 (0.6-3.2)	<b>4.0E-03</b>	0.3	<b>1.0E-03</b>	<b>9.7E-03</b>	1	<b>0.043</b>	0.31	No	
	10	3.5 (2.9-4.1)	4.6 (2.5-6.7)	6.5 (0-15)	0.36	0.79	0.27	0.2	1	1	1	No	
Aspergillus (mm)	6m	0.085 (0-0.2)	0 (0-0)	0.2 (0-0.47)	<b>1.8E-03</b>	<i>0.056</i>	<b>1.2E-04</b>	<b>0.043</b>	1	1	1	Yes	
	2	0.25 (0.1-0.4)	0.22 (0.11-0.33)	1.9 (0.82-2.9)	<b>5.0E-06</b>	1	<b>2.0E-05</b>	<b>7.4E-06</b>	1	<b>1.3E-03</b>	<b>5.4E-04</b>	Yes	
	5	0.37 (0.091-0.66)	0.63 (0.33-0.92)	3.1 (1.8-4.4)	<b>4.1E-09</b>	0.11	<b>2.8E-09</b>	<b>7.2E-07</b>	1	<b>4.7E-07</b>	<b>6.7E-05</b>	No	
Cow's milk (mm)	10	4.5 (3.8-5.2)	3.6 (3-4.2)	4.7 (3.8-5.7)	<b>0.025</b>	0.66	<b>0.033</b>	0.43	0.6	1	0.86	No	
	6m	0.057 (0-0.12)	0.014 (0-0.042)	0.11 (0-0.31)	<i>0.096</i>	0.23	0.27	0.27	1	1	1	Yes	
	2	0.14 (0.031-0.25)	0.089 (0-0.19)	0.091 (0-0.28)	0.59	0.34	0.59	1	1	1	1	Yes	
Egg white (mm)	5	0.13 (0-0.26)	0.41 (0.2-0.62)	0.36 (0-0.77)	<i>0.059</i>	0.14	0.82	0.89	0.52	1	1	No	
	6m	0.023 (0-0.054)	0.061 (5.4E-03-0.12)	0.45 (0.1-0.81)	<b>3.0E-05</b>	0.36	<b>4.6E-05</b>	<b>9.9E-04</b>	1	<b>2.8E-03</b>	<b>0.021</b>	Yes	
	2	0.27 (0.13-0.42)	0.14 (0.046-0.23)	0.52 (0-1.1)	0.2	0.15	0.56	0.13	1	1	1	Yes	
Peanut (mm)	5	0.18 (0.031-0.34)	0.35 (0.054-0.65)	0.6 (0-1.2)	0.24	0.57	<i>0.091</i>	0.2	1	1	1	No	
	6m	0.18 (0.014-0.35)	0.43 (0.13-0.74)	1.8 (0.65-2.9)	<b>2.4E-04</b>	0.26	<b>8.2E-05</b>	<b>1.3E-03</b>	1	<b>4.7E-03</b>	<i>0.055</i>	Yes	
	2	0.23 (0.097-0.37)	0.26 (0.093-0.42)	2.5 (0.98-3.9)	<b>1.8E-06</b>	0.8	<b>1.0E-05</b>	<b>3.5E-06</b>	1	<b>7.2E-04</b>	<b>2.8E-04</b>	Yes	
Peanut (mm)	5	0 (0-0)	0.022 (0-0.066)	0.78 (0-1.8)	<b>2.0E-06</b>	0.35	<b>5.7E-05</b>	<b>8.2E-05</b>	1	<b>3.4E-03</b>	<b>0.014</b>	No	
	6m	0.41 (0.085-0.73)	1.1 (0.54-1.7)	4.7 (2.5-6.9)	<b>1.1E-06</b>	0.2	<b>3.7E-07</b>	<b>4.7E-05</b>	1	<b>3.7E-05</b>	<b>2.9E-03</b>	Yes	
	2	0.24 (0-0.51)	0.78 (0.34-1.2)	5.8 (3-8.5)	<b>2.6E-09</b>	<b>0.021</b>	<b>1.3E-09</b>	<b>2.5E-06</b>	0.6	<b>2.3E-07</b>	<b>2.1E-04</b>	Yes	
Peanut (mm)	5	0 (0-0)	0.094 (0-0.2)	2.3 (0.97-3.6)	<b>3.9E-16</b>	0.1	<b>1.4E-11</b>	<b>1.0E-09</b>	1	<b>4.7E-09</b>	<b>1.8E-07</b>	No	
	5	0.013 (0-0.038)	0.26 (0.076-0.45)	1.5 (0.063-2.9)	<b>1.7E-04</b>	<b>0.026</b>	<b>1.2E-05</b>	<b>0.011</b>	0.71	<b>8.3E-04</b>	0.34	No	
		<b>Prop. (95% CI)</b>	<b>Prop. (95% CI)</b>	<b>Prop. (95% CI)</b>	<b>Overall</b>	<b>1 vs. 2</b>	<b>1 vs. 3</b>	<b>2 vs. 3</b>	<b>Overall</b>	<b>1 vs. 2</b>	<b>1 vs. 3</b>	<b>2 vs. 3</b>	

Histamine wheal $\geq 2$ mm	6m	57% (46%-67%)	55% (46%-65%)	64% (42%-85%)	0.78	0.89	0.63	0.49	1	1	1	1	No*
$\geq 3$ mm	2	89% (82%-95%)	97% (94%-100%)	95% (86%-100%)	<b>0.038</b>	<b>0.021</b>	0.69	0.53	0.62	1	0.6	1	No*
	5	30% (20%-41%)	39% (29%-49%)	29% (7.5%-50%)	0.44	0.26	1	0.46	1	1	1	1	No
HDM wheal $\geq 2$ mm	10	98% (95%-100%)	100% (100%-100%)	100% (100%-100%)	0.53	0.46	1	1	1	1	1	1	No
	6m	2.3% (0%-5.4%)	1.9% (0%-4.5%)	14% (0%-29%)	<b>0.043</b>	1	<i>0.054</i>	<b>0.035</b>	0.68	1	1	1	No*
	2	10% (3.8%-17%)	15% (8.1%-22%)	86% (71%-100%)	<b>2.9E-12</b>	0.39	<b>8.2E-12</b>	<b>1.5E-10</b>	<b>3.7E-10</b>	1	<b>3.0E-09</b>	<b>3.7E-08</b>	No*
	5	13% (5.2%-20%)	28% (18%-37%)	81% (63%-99%)	<b>1.5E-08</b>	<b>0.022</b>	<b>4.6E-09</b>	<b>1.0E-05</b>	<b>8.7E-07</b>	0.62	<b>7.2E-07</b>	<b>7.2E-04</b>	No
Cat wheal $\geq 2$ mm	10	36% (23%-49%)	51% (38%-63%)	78% (57%-95%)	<b>7.4E-03</b>	0.11	<b>2.7E-03</b>	<i>0.06</i>	0.15	1	0.11	1	No
	6m	3.4% (0%-7.3%)	4.7% (0.61%-8.7%)	9.1% (0%-22%)	0.4	0.73	0.26	0.34	1	1	1	1	No*
	2	14% (6.3%-21%)	14% (7.3%-21%)	41% (19%-63%)	<b>0.011</b>	1	0.012	<b>6.3E-03</b>	0.22	1	0.37	0.22	No*
	5	0% (0%-0%)	5.6% (0.73%-10%)	25% (4.2%-46%)	<b>1.9E-04</b>	<i>0.061</i>	<b>2.2E-04</b>	<b>0.017</b>	<b>5.5E-03</b>	1	<b>0.011</b>	0.5	No
	10	6.9% (0.18%-14%)	12% (4%-20%)	17% (0%-36%)	0.4	0.38	0.35	0.69	1	1	1	1	No
Ryegrass wheal $\geq 2$ mm	6m	1.1% (0%-3.4%)	0% (0%-0%)	4.5% (0%-14%)	<i>0.092</i>	0.45	0.36	0.17	1	1	1	1	No*
	2	10% (3.8%-17%)	7.5% (2.4%-13%)	45% (23%-68%)	<b>7.1E-05</b>	0.61	<b>4.3E-04</b>	<b>5.2E-05</b>	<b>2.3E-03</b>	1	<b>0.02</b>	<b>3.1E-03</b>	No*
	5	3.8% (0%-8.1%)	7.8% (2.1%-13%)	50% (26%-74%)	<b>2.2E-06</b>	0.34	<b>2.4E-06</b>	<b>3.9E-05</b>	<b>8.8E-05</b>	1	<b>2.0E-04</b>	<b>2.4E-03</b>	No
<i>Alternaria</i> wheal $\geq 2$ mm	10	21% (9.9%-31%)	24% (13%-34%)	56% (30%-81%)	<b>0.017</b>	0.83	<b>7.3E-03</b>	<b>0.019</b>	0.31	1	0.24	0.55	No
	6m	2.3% (0%-5.4%)	0% (0%-0%)	4.5% (0%-14%)	<i>0.078</i>	0.2	0.49	0.17	1	1	1	1	No*
	2	4.5% (0.11%-9%)	0.93% (0%-2.8%)	4.5% (0%-14%)	0.23	0.18	1	0.31	1	1	1	1	No*
	5	1.3% (0%-3.8%)	6.7% (1.4%-12%)	4.8% (0%-15%)	0.16	0.12	0.38	1	1	1	1	1	No
<i>Aspergillus</i> wheal $\geq 2$ mm	6m	0% (0%-0%)	0.93% (0%-2.8%)	18% (0.68%-36%)	<b>3.8E-04</b>	1	<b>1.3E-03</b>	<b>2.9E-03</b>	<b>0.01</b>	1	<i>0.055</i>	0.11	No*
	2	10% (3.8%-17%)	4.7% (0.61%-8.7%)	14% (0%-29%)	0.15	0.17	0.7	0.14	1	1	1	1	No*
	5	1.3% (0%-3.8%)	4.4% (0.1%-8.8%)	15% (0%-32%)	<b>0.032</b>	0.37	<b>0.025</b>	0.11	0.53	1	0.69	1	No
Cow's milk wheal $\geq 2$ mm	6m	3.4% (0%-7.3%)	7.5% (2.4%-13%)	36% (15%-58%)	<b>1.2E-04</b>	0.35	<b>8.7E-05</b>	<b>1.1E-03</b>	<b>3.7E-03</b>	1	<b>5.0E-03</b>	<b>0.047</b>	No*
	2	9.1% (3%-15%)	6.5% (1.8%-11%)	50% (27%-73%)	<b>4.7E-06</b>	0.59	<b>5.1E-05</b>	<b>4.5E-06</b>	<b>1.7E-04</b>	1	<b>3.1E-03</b>	<b>3.5E-04</b>	No*
	5	0% (0%-0%)	0% (0%-0%)	5% (0%-15%)	0.11	1	0.2	0.18	1	1	1	1	No
Egg white wheal $\geq 2$ mm	6m	6.8% (1.4%-12%)	15% (8.1%-22%)	55% (32%-77%)	<b>4.6E-06</b>	0.11	<b>1.9E-06</b>	<b>1.9E-04</b>	<b>1.7E-04</b>	1	<b>1.6E-04</b>	<b>0.01</b>	No*
	2	5.7% (0.75%-11%)	14% (7.3%-21%)	59% (37%-81%)	<b>1.8E-07</b>	<i>0.062</i>	<b>1.0E-07</b>	<b>2.4E-05</b>	<b>8.6E-06</b>	1	<b>1.2E-05</b>	<b>1.6E-03</b>	No*
	5	0% (0%-0%)	2.2% (0%-5.3%)	29% (7.5%-50%)	<b>1.1E-05</b>	0.5	<b>4.6E-05</b>	<b>5.2E-04</b>	<b>3.9E-04</b>	1	<b>2.8E-03</b>	<b>0.024</b>	No
Any SPT wheal $\geq 2$ mm	6m	16% (8.1%-24%)	21% (14%-29%)	64% (42%-85%)	<b>4.4E-05</b>	0.36	<b>2.1E-05</b>	<b>1.8E-04</b>	<b>1.4E-03</b>	1	<b>1.4E-03</b>	<b>9.6E-03</b>	No*
	2	34% (24%-44%)	36% (27%-46%)	100% (100%-100%)	<b>3.7E-09</b>	0.77	<b>3.6E-09</b>	<b>6.1E-09</b>	<b>2.3E-07</b>	1	<b>5.9E-07</b>	<b>9.2E-07</b>	No*
	5	16% (8.1%-25%)	38% (28%-48%)	90% (77%-100%)	<b>8.5E-10</b>	<b>2.1E-03</b>	<b>4.0E-10</b>	<b>1.3E-05</b>	<b>6.2E-08</b>	0.084	<b>8.6E-08</b>	<b>9.0E-04</b>	No
$\geq 3$ mm	10	52% (38%-65%)	60% (48%-72%)	89% (73%-100%)	<b>0.014</b>	0.37	<b>5.4E-03</b>	<b>0.026</b>	0.26	1	0.19	0.71	No

Feature? = whether variable was used as a clustering feature or not; geom. mean = geometric mean; HDM = house dust mite; P-value (adj.) = adjusted P-values (Benjamini-Yekutieli method); SPT = skin prick or sensitisation test. For categorical variables, associations were tested using Fisher exact test; for continuous variables, Kruskal-Wallis and Mann-Whitney-Wilcoxon. Bold text indicates statistical significance ( $p < 0.05$ ); italics indicate near-significance ( $p < 0.10$ ). \*Not used as clustering features, as these were derived variables.

## E. Microbiological and respiratory health-related

Variable	Age (y)	CASI (N=88)		CAS2 (N=107)		CAS3 (N=22)		P-value		P-value (adj.)			Feature?
		Mean	(95% CI)	Mean	(95% CI)	Mean	(95% CI)	Overall	1 vs. 2	1 vs. 3	2 vs. 3	Overall	
<i>Events in general</i>													
Any ARI (events per y)													
	1	4.4 (3.9-4.9)	3.6 (3.1-4.1)	4.5 (3.3-5.6)	<b>0.044</b>	<b>0.018</b>	0.86	0.16	0.69	1	0.52	1	No*
	2	4.5 (3.8-5.2)	3.6 (3.2-4)	4.7 (3.4-6)	0.13	0.13	0.56	<b>0.077</b>	1	1	1	1	No*
	3	3.7 (3.1-4.3)	3.4 (3-3.9)	4 (2.7-5.4)	0.74	0.56	0.77	0.53	1	1	1	1	No*
	4	3 (2.4-3.6)	2.7 (2.2-3.2)	3.7 (2.4-4.9)	0.28	0.48	0.28	0.11	1	1	1	1	No
	5	2 (1.5-2.5)	1.9 (1.5-2.3)	1.5 (0.9-2.1)	0.94	0.83	0.89	0.74	1	1	1	1	No
URI (events per y)													
	1	2.9 (2.4-3.3)	2.6 (2.2-3)	2.5 (1.7-3.3)	0.59	0.34	0.5	0.96	1	1	1	1	Yes
	2	3.2 (2.6-3.7)	2.6 (2.2-3)	2.5 (1.2-3.8)	0.19	0.19	0.12	0.34	1	1	1	1	Yes
	3	2.7 (2.2-3.2)	2.8 (2.4-3.3)	2.2 (1.3-3.2)	0.45	0.41	0.59	0.24	1	1	1	1	Yes
	4	2.1 (1.7-2.6)	2.2 (1.8-2.7)	1.7 (0.77-2.7)	0.5	0.94	0.26	0.27	1	1	1	1	No
	5	1.6 (1.1-2)	1.5 (1.2-1.9)	0.67 (0.2-1.1)	<i>0.081</i>	0.76	<b>0.047</b>	<b>0.026</b>	1	1	1	1	No
LRI (events per y)													
	1	1.6 (1.2-1.9)	0.98 (0.76-1.2)	2 (1.3-2.6)	<b>4.0E-03</b>	<b>0.021</b>	0.17	<b>2.6E-03</b>	0.087	0.6	1	0.1	Yes
	2	1.4 (0.98-1.7)	1 (0.81-1.2)	2.2 (1.6-2.9)	<b>2.5E-03</b>	0.83	<b>6.1E-03</b>	<b>2.0E-04</b>	0.058	1	0.21	<b>0.011</b>	Yes
	3	1 (0.76-1.3)	0.6 (0.4-0.8)	1.8 (1.1-2.6)	<b>6.1E-04</b>	<b>0.02</b>	<b>0.039</b>	<b>2.7E-04</b>	<b>0.016</b>	0.57	0.98	<b>0.014</b>	Yes
	4	0.87 (0.52-1.2)	0.46 (0.3-0.63)	2 (1.1-2.8)	<b>1.7E-05</b>	0.3	<b>3.5E-04</b>	<b>1.6E-06</b>	<b>5.9E-04</b>	1	<b>0.017</b>	<b>1.4E-04</b>	No
	5	0.42 (0.24-0.6)	0.36 (0.24-0.48)	0.86 (0.44-1.3)	<b>0.019</b>	1	<b>0.011</b>	<b>7.5E-03</b>	0.34	1	0.34	0.25	No
Wheezy LRI (wLRI, events per y)													
	1	0.47 (0.3-0.63)	0.24 (0.15-0.34)	0.64 (0.19-1.1)	<i>0.054</i>	<b>0.036</b>	0.61	<i>0.065</i>	0.82	0.93	1	1	Yes
	2	0.68 (0.45-0.91)	0.41 (0.26-0.56)	1 (0.56-1.5)	<b>5.2E-03</b>	<b>0.063</b>	<i>0.066</i>	<b>1.7E-03</b>	0.11	1	1	0.07	Yes
	3	0.59 (0.37-0.81)	0.3 (0.17-0.44)	1.4 (0.78-2.1)	<b>4.6E-05</b>	<i>0.065</i>	<b>2.5E-03</b>	<b>6.6E-06</b>	<b>1.5E-03</b>	1	0.098	<b>4.9E-04</b>	Yes
	4	0.52 (0.25-0.79)	0.32 (0.18-0.46)	1.9 (0.95-2.8)	<b>4.5E-08</b>	0.86	<b>9.3E-07</b>	<b>3.3E-08</b>	<b>2.4E-06</b>	1	<b>8.5E-05</b>	<b>4.3E-06</b>	No
	5	0.28 (0.13-0.42)	0.23 (0.13-0.33)	0.76 (0.36-1.2)	<b>2.3E-03</b>	0.99	<b>2.0E-03</b>	<b>1.2E-03</b>	<i>0.053</i>	1	0.08	<i>0.051</i>	No
Febrile LRI (fLRI, events per y)													
	1	0.36 (0.22-0.51)	0.28 (0.16-0.4)	0.55 (0.28-0.81)	<b>0.025</b>	0.24	<i>0.071</i>	<b>6.4E-03</b>	0.43	1	1	0.22	Yes
	2	0.36 (0.23-0.5)	0.33 (0.22-0.43)	0.95 (0.46-1.4)	<b>0.01</b>	1	<b>6.1E-03</b>	<b>3.8E-03</b>	0.2	1	0.21	0.14	Yes
	3	0.38 (0.21-0.55)	0.16 (0.09-0.23)	0.52 (0.13-0.92)	<i>0.06</i>	<i>0.063</i>	0.44	0.04	0.89	1	1	1	Yes
	4	0.3 (0.13-0.47)	0.15 (0.064-0.24)	0.43 (0.16-0.7)	<b>0.021</b>	0.18	<i>0.091</i>	<b>4.9E-03</b>	0.37	1	1	0.17	No
	5	0.19 (0.082-0.3)	0.14 (0.06-0.21)	0.19 (0-0.42)	0.83	0.55	0.91	0.8	1	1	1	1	No
Severe LRI (wLRI or fLRI, events per y)													
	1	0.69 (0.5-0.89)	0.44 (0.29-0.58)	1 (0.49-1.5)	<b>0.012</b>	<b>0.027</b>	0.25	<b>9.1E-03</b>	0.23	0.73	1	0.29	No*
	2	0.9 (0.62-1.2)	0.59 (0.43-0.75)	1.6 (1.1-2.2)	<b>7.9E-04</b>	0.22	<b>5.2E-03</b>	<b>1.2E-04</b>	<b>0.02</b>	1	0.18	<b>6.6E-03</b>	No*
	3	0.73 (0.49-0.97)	0.37 (0.23-0.51)	1.5 (0.85-2.2)	<b>1.6E-04</b>	<b>0.032</b>	<b>0.01</b>	<b>3.8E-05</b>	<b>4.8E-03</b>	0.84	0.32	<b>2.4E-03</b>	No*
	4	0.63 (0.32-0.94)	0.36 (0.21-0.52)	1.9 (1-2.8)	<b>2.8E-07</b>	0.56	<b>5.9E-06</b>	<b>8.4E-08</b>	<b>1.3E-05</b>	1	<b>4.5E-04</b>	<b>1.0E-05</b>	No
	5	0.36 (0.19-0.53)	0.27 (0.17-0.38)	0.76 (0.36-1.2)	<b>0.015</b>	0.88	<b>0.012</b>	<b>5.0E-03</b>	0.28	1	0.37	0.18	No
<i>Events with RSV detected in sample</i>													
URI (events per y)													
	1	0.22 (0.13-0.3)	0.24 (0.15-0.33)	0.14 (0-0.34)	0.41	0.84	0.22	0.19	1	1	1	1	Yes
	2	0.11 (0.032-0.2)	0.25 (0.16-0.34)	0.14 (0-0.31)	<b>0.033</b>	<b>0.01</b>	0.51	0.35	0.54	0.32	1	1	Yes
	3	0.15 (0.068-0.24)	0.18 (0.089-0.27)	0.14 (0-0.36)	0.82	0.82	0.63	0.54	1	1	1	1	Yes
LRI (events per y)													
	1	0.28 (0.18-0.39)	0.12 (0.059-0.18)	0.36 (0.15-0.58)	<b>5.6E-03</b>	<b>7.1E-03</b>	0.43	<b>5.3E-03</b>	0.12	0.24	1	0.19	Yes
	2	0.15 (0.072-0.22)	0.16 (0.084-0.23)	0.24 (0.039-0.44)	0.57	0.95	0.32	0.33	1	1	1	1	Yes
	3	0.13 (0.049-0.21)	0.076 (0.018-0.13)	0.29 (0.075-0.5)	<b>0.015</b>	0.23	<i>0.061</i>	<b>3.1E-03</b>	0.28	1	1	0.12	Yes
wLRI (events per y)													
	1	0.057 (7.5E-03-0.11)	0.028 (0-0.06)	0.18 (6.8E-03-0.36)	<b>0.016</b>	0.32	<i>0.058</i>	<b>4.0E-03</b>	0.29	1	1	0.15	Yes
	2	0.091 (0.03-0.15)	0.13 (0.066-0.2)	0.048 (0-0.15)	0.44	0.38	0.52	0.28	1	1	1	1	Yes
	3	0.047 (0-0.1)	0.029 (0-0.061)	0.24 (0.039-0.44)	<b>3.2E-04</b>	0.79	<b>2.1E-03</b>	<b>8.7E-03</b>	1	<i>0.084</i>	<b>0.017</b>	<b>0.017</b>	Yes
	1	0.1 (0.038-0.17)	0.037 (8.5E-04-0.074)	0.18 (6.8E-03-0.36)	<b>0.04</b>	<i>0.072</i>	0.31	<b>0.011</b>	0.64	1	1	0.34	Yes

2	0.08 (0.022-0.14)	0.093 (0.032-0.16)	0.24 (0.039-0.44)	0.075	0.9	<b>0.039</b>	<b>0.043</b>	1	1	0.98	1	<b>Yes</b>
3	0.082 (0.023-0.14)	0.019 (0-0.046)	0.095 (0-0.23)	0.097	<b>0.042</b>	0.86	0.072	1	1	1	1	<b>Yes</b>
<i>Events with influenza detected in sample</i>												
<i>URI (events per y)</i>												
1	0.12 (0.048-0.2)	0.13 (0.066-0.2)	0.045 (0-0.14)	0.53	0.74	0.34	0.26	1	1	1	1	<b>Yes</b>
2	0.091 (0.03-0.15)	0.075 (0.024-0.13)	0.19 (0-0.42)	0.57	0.69	0.45	0.29	1	1	1	1	<b>Yes</b>
3	0.059 (7.8E-03-0.11)	0.038 (8.7E-04-0.075)	0.048 (0-0.15)	0.8	0.51	0.85	0.85	1	1	1	1	<b>Yes</b>
<i>LRI (events per y)</i>												
1	0.057 (0-0.12)	0.028 (0-0.06)	0.091 (0-0.22)	0.4	0.51	0.42	0.17	1	1	1	1	<b>Yes</b>
2	0.034 (0-0.073)	0.037 (8.5E-04-0.074)	0.095 (0-0.23)	0.43	0.91	0.24	0.26	1	1	1	1	<b>Yes</b>
3	0.035 (0-0.075)	9.5E-03 (0-0.028)	0.048 (0-0.15)	0.38	0.22	0.8	0.21	1	1	1	1	<b>Yes</b>
<i>wLRI (events per y)</i>												
1	0.023 (0-0.054)	0 (0-0)	0.045 (0-0.14)	0.16	0.12	0.57	<b>0.029</b>	1	1	1	0.77	<b>Yes</b>
2	0.034 (0-0.073)	0 (0-0)	0.048 (0-0.15)	0.13	0.056	0.78	<b>0.025</b>	1	1	1	0.69	<b>Yes</b>
3	0.024 (0-0.056)	9.5E-03 (0-0.028)	0.048 (0-0.15)	0.47	0.45	0.56	0.21	1	1	1	1	<b>Yes</b>
<i>fLRI (events per y)</i>												
1	0.034 (0-0.073)	0.019 (0-0.045)	0.045 (0-0.14)	0.7	0.5	0.81	0.46	1	1	1	1	<b>Yes</b>
2	0.011 (0-0.034)	0.019 (0-0.045)	0.095 (0-0.23)	0.066	0.68	<b>0.037</b>	0.068	0.97	1	0.94	1	<b>Yes</b>
3	0.035 (0-0.075)	9.5E-03 (0-0.028)	0 (0-0)	0.35	0.22	0.39	0.67	1	1	1	1	<b>Yes</b>
<i>Events with HRV-A detected in sample</i>												
<i>LRI (events per y)</i>												
1	0.55 (0.33-0.76)	0.36 (0.24-0.49)	0.73 (0.27-1.2)	0.22	0.37	0.29	0.089	1	1	1	1	<b>Yes</b>
2	0.32 (0.2-0.44)	0.23 (0.14-0.33)	0.38 (0.11-0.65)	0.29	0.23	0.58	0.18	1	1	1	1	<b>Yes</b>
3	0.27 (0.14-0.41)	0.095 (0.032-0.16)	0.29 (0.031-0.54)	<b>0.048</b>	<b>0.031</b>	0.68	<b>0.042</b>	0.74	0.81	1	1	<b>Yes</b>
<i>wLRI (events per y)</i>												
1	0.14 (0.063-0.21)	0.15 (0.071-0.23)	0.23 (0-0.46)	0.79	0.95	0.54	0.51	1	1	1	1	<b>Yes</b>
2	0.17 (0.09-0.25)	0.084 (0.025-0.14)	0.19 (0-0.42)	0.13	<b>0.043</b>	0.83	0.3	1	1	1	1	<b>Yes</b>
3	0.14 (0.052-0.23)	0.057 (0.012-0.1)	0.24 (0-0.48)	0.098	0.13	0.38	<b>0.037</b>	1	1	1	0.94	<b>Yes</b>
<i>fLRI (events per y)</i>												
1	0.12 (0.048-0.2)	0.1 (0.044-0.16)	0.14 (0-0.34)	0.95	0.79	0.81	0.92	1	1	1	1	<b>Yes</b>
2	0.068 (5.6E-03-0.13)	0.047 (6.1E-03-0.087)	0.14 (0-0.31)	0.24	0.74	0.19	0.099	1	1	1	1	<b>Yes</b>
3	0.071 (5.8E-03-0.14)	0.038 (8.7E-04-0.075)	0.095 (0-0.29)	0.79	0.5	0.88	0.82	1	1	1	1	<b>Yes</b>
<i>Events with HRV-B detected in sample</i>												
<i>LRI (events per y)</i>												
1	0.068 (0.014-0.12)	0.019 (0-0.045)	0.091 (0-0.22)	0.15	0.084	0.72	0.078	1	1	1	1	<b>Yes</b>
2	0.011 (0-0.034)	0 (0-0)	0.048 (0-0.15)	0.11	0.27	0.28	<b>0.025</b>	1	1	1	0.69	<b>Yes</b>
3	0.024 (0-0.056)	0 (0-0)	0.095 (0-0.23)	<b>0.013</b>	0.12	0.13	<b>1.6E-03</b>	0.25	1	1	0.066	<b>Yes</b>
<i>wLRI (events per y)</i>												
1	0.011 (0-0.034)	0 (0-0)	0.045 (0-0.14)	0.12	0.27	0.29	<b>0.029</b>	1	1	1	0.77	<b>Yes</b>
2	0 (0-0)	0 (0-0)	0 (0-0)	NA	NA	NA	NA	NA	NA	NA	NA	No <sup>^</sup>
3	0.012 (0-0.035)	0 (0-0)	0.048 (0-0.15)	0.12	0.27	0.29	<b>0.027</b>	1	1	1	0.73	<b>Yes</b>
<i>fLRI (events per y)</i>												
1	0.034 (0-0.073)	0 (0-0)	0.045 (0-0.14)	0.13	0.056	0.81	<b>0.029</b>	1	1	1	0.77	<b>Yes</b>
2	0.011 (0-0.034)	0 (0-0)	0.048 (0-0.15)	0.11	0.27	0.28	<b>0.025</b>	1	1	1	0.69	<b>Yes</b>
3	0.012 (0-0.035)	0 (0-0)	0.048 (0-0.15)	0.12	0.27	0.29	<b>0.027</b>	1	1	1	0.73	<b>Yes</b>
<i>Events with HRV-C detected in sample</i>												
<i>LRI (events per y)</i>												
1	0.51 (0.35-0.67)	0.29 (0.18-0.4)	0.59 (0.14-1)	0.068	<b>0.027</b>	0.99	0.17	1	0.73	1	1	<b>Yes</b>
2	0.35 (0.19-0.51)	0.23 (0.14-0.33)	0.67 (0.3-1)	<b>0.021</b>	0.63	<b>0.027</b>	<b>5.1E-03</b>	0.37	1	0.73	0.18	<b>Yes</b>
3	0.25 (0.13-0.37)	0.14 (0.07-0.22)	0.57 (0.2-0.94)	<b>0.014</b>	0.26	<b>0.046</b>	<b>3.3E-03</b>	0.26	1	1	0.13	<b>Yes</b>
<i>wLRI (events per y)</i>												
1	0.15 (0.06-0.24)	0.075 (0.024-0.13)	0.23 (0-0.46)	0.23	0.23	0.48	0.11	1	1	1	1	No
2	0.14 (0.044-0.23)	0.084 (0.014-0.15)	0.38 (0.076-0.69)	<b>9.0E-03</b>	0.34	<b>0.03</b>	<b>2.1E-03</b>	0.18	1	0.79	0.084	No
3	0.16 (0.06-0.27)	0.067 (0.011-0.12)	0.48 (0.13-0.82)	<b>8.4E-04</b>	0.13	<b>0.016</b>	<b>1.2E-04</b>	<b>0.021</b>	1	0.48	<b>6.6E-03</b>	No
<i>fLRI (events per y)</i>												
1	0.091 (0.03-0.15)	0.075 (0.018-0.13)	0.091 (0-0.22)	0.8	0.52	1	0.69	1	1	1	1	<b>Yes</b>
2	0.057 (0-0.12)	0.047 (6.1E-03-0.087)	0.048 (0-0.15)	1	0.98	0.98	0.99	1	1	1	1	<b>Yes</b>
3	0.047 (1.1E-03-0.093)	0.029 (0-0.061)	0.14 (0-0.31)	0.08	0.5	0.12	<b>0.026</b>	1	1	1	0.71	<b>Yes</b>

		Prop. (95% CI)	Prop. (95% CI)	Overall	1 vs. 2	1 vs. 3	2 vs. 3	Overall	1 vs. 2	1 vs. 3	2 vs. 3	
<i>Proportion of ARI events</i>												
<i>wLRI / ARI</i>												
1		11% (7%-16%)	7.1% (4.1%-10%)	0.16	0.092	0.76	0.16	1	1	1	1	No*
2		15% (10%-19%)	12% (7.8%-17%)	<b>0.015</b>	0.11	0.07	<b>5.5E-03</b>	0.28	1	1	0.19	No*
3		16% (10%-23%)	10% (5.9%-15%)	<b>3.2E-04</b>	<b>0.082</b>	<b>7.2E-03</b>	<b>6.1E-05</b>	<b>8.7E-03</b>	1	0.24	<b>3.6E-03</b>	No*
4		14% (7.8%-21%)	13% (7.5%-19%)	<b>1.0E-07</b>	0.98	<b>5.2E-07</b>	<b>1.8E-07</b>	<b>5.1E-06</b>	1	<b>5.0E-05</b>	<b>1.9E-05</b>	No*
5		13% (6.2%-20%)	13% (6.7%-20%)	<b>2.4E-04</b>	0.92	<b>2.1E-04</b>	<b>1.8E-04</b>	<b>6.8E-03</b>	1	<b>0.011</b>	<b>9.6E-03</b>	No*
1	<i>fLRI / ARI</i>	7.3% (4.5%-10%)	7.3% (4%-11%)	<b>0.025</b>	0.43	<b>0.028</b>	<b>8.8E-03</b>	0.43	1	0.75	0.29	No*
2		7.9% (4.7%-11%)	11% (6.9%-15%)	<b>0.014</b>	0.69	<b>4.3E-03</b>	<b>0.01</b>	0.26	1	0.16	0.32	No*
3		11% (6.1%-16%)	5% (2.1%-8%)	<i>0.066</i>	<i>0.05</i>	0.6	<i>0.058</i>	0.97	1	1	1	No*
4		7.9% (4%-12%)	8.1% (2.8%-13%)	<b>0.046</b>	0.26	0.11	<b>0.013</b>	0.71	1	1	0.4	No*
5		11% (4.4%-17%)	7% (2.3%-12%)	0.68	0.38	0.74	0.86	1	1	1	1	No*
1	<i>sLRI / ARI</i>	16% (11%-21%)	12% (8%-16%)	<b>0.041</b>	<i>0.091</i>	0.2	<b>0.021</b>	0.65	1	1	0.6	No*
2		19% (14%-24%)	18% (13%-24%)	<b>2.3E-03</b>	0.56	<b>1.8E-03</b>	<b>7.9E-04</b>	<i>0.053</i>	1	<i>0.074</i>	<b>0.035</b>	No*
3		21% (14%-27%)	13% (7.7%-17%)	<b>1.0E-03</b>	<b>0.044</b>	<b>0.031</b>	<b>3.0E-04</b>	<b>0.025</b>	1	0.81	<b>0.015</b>	No*
4		17% (10%-24%)	15% (8.7%-21%)	<b>7.8E-07</b>	0.67	<b>4.0E-06</b>	<b>5.6E-07</b>	<b>3.4E-05</b>	1	<b>3.2E-04</b>	<b>5.3E-05</b>	No*
5		17% (9.4%-25%)	16% (8.8%-23%)	<b>1.5E-03</b>	0.74	<b>1.1E-03</b>	<b>7.1E-04</b>	<b>0.036</b>	1	<b>0.047</b>	<b>0.032</b>	No*
<i>Proportion of LRI events</i>												
<i>wLRI / LRI</i>												
1		33% (23%-44%)	28% (17%-38%)	0.54	0.27	0.82	0.63	1	1	1	1	No*
2		54% (42%-66%)	39% (28%-50%)	0.16	<i>0.068</i>	0.25	0.57	1	1	1	1	No*
3		52% (39%-66%)	52% (38%-67%)	<i>0.099</i>	0.95	<b>0.035</b>	<i>0.059</i>	1	1	0.9	1	No*
4		49% (33%-65%)	65% (49%-82%)	<b>1.1E-03</b>	<i>0.08</i>	<b>2.1E-04</b>	<b>0.026</b>	<b>0.027</b>	1	<b>0.011</b>	0.71	No*
5		62% (42%-82%)	66% (48%-84%)	0.28	0.75	0.11	0.19	1	1	1	1	No*
1	<i>fLRI / LRI</i>	24% (15%-33%)	25% (15%-35%)	0.21	0.81	0.11	0.11	1	1	1	1	No*
2		30% (20%-41%)	36% (26%-47%)	0.54	0.74	0.23	0.43	1	1	1	1	No*
3		36% (23%-48%)	32% (18%-45%)	0.77	0.55	0.56	0.92	1	1	1	1	No*
4		31% (18%-45%)	33% (17%-50%)	0.89	0.76	0.61	0.98	1	1	1	1	No*
5		45% (26%-65%)	34% (17%-51%)	0.2	0.37	0.08	0.28	1	1	1	1	No*
1	<i>sLRI / LRI</i>	50% (39%-61%)	42% (30%-53%)	0.17	0.23	0.32	0.085	1	1	1	1	No*
2		68% (58%-79%)	60% (49%-71%)	0.67	0.42	0.89	0.55	1	1	1	1	No*
3		67% (55%-80%)	65% (51%-78%)	0.28	0.93	0.13	0.14	1	1	1	1	No*
4		61% (45%-77%)	74% (58%-90%)	<b>3.6E-03</b>	<i>0.094</i>	<b>8.8E-04</b>	<i>0.057</i>	<i>0.08</i>	1	<b>0.038</b>	1	No*
5		82% (66%-98%)	77% (61%-92%)	0.67	0.57	0.75	0.42	1	1	1	1	No*
<i>&gt;20% Streptococcus in first infection-naive NPA</i>												
7w		11% (0.34%-23%)	15% (3.3%-26%)	<i>0.081</i>	0.75	<b>0.042</b>	<i>0.065</i>	1	1	1	1	No
6m		7.6% (1.6%-14%)	18% (10%-26%)	0.12	<b>0.045</b>	0.39	1	1	1	1	1	No
0-2	% Healthy NPAs with risk-associated MPGs	49% (38%-59%)	32% (24%-39%)	<b>1.2E-03</b>	<b>0.013</b>	0.2	<b>5.5E-04</b>	<b>0.029</b>	0.4	1	<b>0.025</b>	No
2-4	% Healthy NPAs with health-associated MPGs	46% (37%-55%)	44% (37%-51%)	0.9	0.67	0.92	0.8	1	1	1	1	No
0-2		35% (25%-44%)	46% (38%-55%)	<i>0.077</i>	<i>0.064</i>	0.61	<i>0.08</i>	1	1	1	1	No
2-4		33% (26%-41%)	28% (21%-35%)	0.34	0.24	0.23	0.66	1	1	1	1	No
<i>Quartile of % healthy NPAs with risk-associated MPGs</i>												
0-2		2.4 (2-2.7)	1.8 (1.6-2.1)	<i>0.081</i>	<b>0.015</b>	0.29	<b>2.6E-03</b>	<i>0.081</i>	0.45	1	0.1	No

Quartile of % healthy NPAs	2-4	2.2 (1.9-2.5)	2.2 (1.9-2.4)	2.2 (1.7-2.7)	0.99	0.99	0.89	0.87	1	1	1	1	No
with health-associated MPGs	0-2	2.1 (1.9-2.4)	2.5 (2.2-2.7)	2.1 (1.5-2.6)	0.15	0.087	0.83	0.18	1	1	1	1	No
	2-4	2.3 (2-2.5)	2 (1.8-2.3)	1.9 (1.3-2.4)	0.3	0.2	0.21	0.65	1	1	1	1	No

Feature? = whether variable was used as a clustering feature or not; geom. mean = geometric mean; ARI = acute respiratory infection (lower or upper); LRI = lower respiratory infection; MPG = microbiome profile group; NPA = nasopharyngeal aspirate; prop. = proportion; P-value (adj.) = adjusted P-values (Benjamini-Yekutieli method); URI = upper respiratory infection; 7w = 7 weeks. For categorical variables, associations were tested using Fisher exact test; for continuous variables, Kruskal-Wallis and Mann-Whitney-Wilcoxon. Bold text indicates statistical significance ( $p < 0.05$ ); italics indicate near-significance ( $p < 0.10$ ). \*Not used as clustering features, as these were derived variables; the variables from which they were derived (URI, LRI, wLRI, fLRI) were used instead. ^Not used as clustering feature due to no variation across entire cohort.

**TABLE B.4: Repeated-measures ANOVA for selected predictors, in the first three years of life (timepoints at ages 6m, 1, 2, and 3)**

Predictor	P-value for predictor, following repeated-measures ANOVA within each cluster		
	CAS1	CAS2	CAS3
HDM IgE	0.97	<b>8.1E-03</b>	0.059
Phadiatop IgE	0.24	0.78	0.47
LRI	<b>0.048</b>	<b>2.4E-03</b>	0.18
wLRI	<b>0.011</b>	<b>0.015</b>	0.97
fLRI	0.28	<b>0.011</b>	0.97

**TABLE B.5: Comparison of the three clusters generated by npEM, with other clustering or classification schemes.**

**A. npEM vs. atopy as defined by specific IgE or SPT past atopic threshold by age 2.** \*Any specific IgE  $\geq$  0.35kU/L, or any SPT  $\geq$  2mm, at any timepoint less than two years of age.

npEM	Atopic*	
	No	Yes
CAS1	39	46
CAS2	30	75
CAS3	0	22

**B. npEM vs. atopy as defined only by specific IgE past atopic threshold by age 2.** \*Any specific IgE  $\geq$  0.35kU/L at any timepoint less than two years of age.

npEM	Atopic*	
	No	Yes
CAS1	58	26
CAS2	40	65
CAS3	0	22

**TABLE B.6: Correlation between Phadiatop vs. allergen-specific IgE and IgG4 in CAS.**

**A. Phadiatop vs. allergen-specific IgE.** \*Assay used at age 5 was the adult version, not Phadiatop infant. Bold indicates allergen with strongest correlation in timepoint and cluster.

Age (y)	Allergen	All clusters		CAS1		CAS2		CAS3	
		Rho	P-value	Rho	P-value	Rho	P-value	Rho	P-value
6m	HDM	0.38	2.7E-09	0.27	0.012	0.37	1.2E-04	0.21	0.34
	Cat	0.37	5.3E-09	0.37	4.6E-04	0.38	7.3E-05	0.34	0.12
	Peanut	<b>0.67</b>	<b>2.1E-32</b>	<b>0.57</b>	<b>9.0E-09</b>	<b>0.5</b>	<b>4.7E-08</b>	<b>0.77</b>	<b>2.9E-05</b>
	Mould	0.18	5.3E-03	0.4	1.1E-04	0.25	0.012	-1.1E-01	0.61
	Couch grass	0.19	3.3E-03	0.33	2.2E-03	0.23	0.017	-1.4E-02	0.95
1	Ryegrass	0.21	9.8E-04	0.34	1.2E-03	0.29	3.1E-03	0.016	0.94
	HDM	0.52	6.3E-18	0.28	9.1E-03	0.29	2.4E-03	0.35	0.11
	Cat	0.43	8.4E-12	0.22	0.04	0.31	1.1E-03	0.23	0.31
	Peanut	<b>0.71</b>	<b>6.4E-37</b>	<b>0.54</b>	<b>1.1E-07</b>	<b>0.49</b>	<b>6.3E-08</b>	<b>0.69</b>	<b>3.5E-04</b>
	Mould	0.11	0.083	0.16	0.14	0.072	0.46	NA	NA
2	Couch grass	0.27	3.6E-05	0.27	0.011	0.18	0.069	0.14	0.54
	Ryegrass	0.21	1.3E-03	0.21	0.057	0.17	0.078	0.015	0.95
	HDM	<b>0.79</b>	<b>7.4E-47</b>	<b>0.64</b>	<b>2.3E-11</b>	<b>0.59</b>	<b>5.1E-11</b>	0.43	0.046
	Cat	0.46	7.5E-13	0.39	1.6E-04	0.29	2.8E-03	0.13	0.55
	Peanut	0.67	1.2E-29	0.4	1.5E-04	0.37	1.1E-04	<b>0.6</b>	<b>3.5E-03</b>
3	Mould	0.2	2.4E-03	0.11	0.29	0.21	0.032	0.13	0.57
	Couch grass	0.45	2.0E-12	0.33	1.8E-03	0.12	0.21	0.24	0.29
	Ryegrass	NA	NA	NA	NA	NA	NA	NA	NA
	HDM	<b>0.84</b>	<b>1.2E-57</b>	<b>0.81</b>	<b>1.5E-20</b>	<b>0.7</b>	<b>4.2E-16</b>	<b>0.81</b>	<b>4.2E-06</b>
	Cat	0.45	9.0E-12	0.5	1.5E-06	0.23	0.019	0.022	0.92
4	Peanut	0.61	8.3E-23	0.56	3.7E-08	0.26	8.8E-03	0.44	0.039
	Mould	0.36	7.7E-08	0.5	1.2E-06	0.13	0.2	0.46	0.031
	Couch grass	0.51	5.0E-15	0.5	1.3E-06	0.19	0.052	0.42	0.049
	Ryegrass	0.57	1.2E-19	0.59	3.3E-09	0.25	0.012	0.61	2.3E-03
	HDM	<b>0.88</b>	<b>6.1E-63</b>	<b>0.79</b>	<b>5.7E-18</b>	<b>0.83</b>	<b>6.5E-24</b>	<b>0.85</b>	<b>1.1E-06</b>
5*	Cat	0.52	3.0E-14	0.45	4.3E-05	0.4	9.5E-05	-8.0E-04	1
	Peanut	0.58	1.3E-18	0.54	3.1E-07	0.33	2.0E-03	0.38	0.09
	Mould	0.33	4.2E-06	0.26	0.023	0.28	8.3E-03	0.27	0.23
	Couch grass	0.57	6.5E-18	0.48	7.5E-06	0.35	7.1E-04	0.4	0.075
	Ryegrass	0.63	1.5E-22	0.55	2.1E-07	0.38	2.8E-04	0.51	0.018
5*	HDM	<b>0.93</b>	<b>1.1E-73</b>	<b>0.88</b>	<b>3.3E-23</b>	<b>0.9</b>	<b>8.0E-30</b>	<b>0.98</b>	<b>4.8E-12</b>
	Cat	0.58	1.9E-16	0.45	1.3E-04	0.52	1.1E-06	0.18	0.48
	Peanut	0.59	2.9E-17	0.61	3.7E-08	0.43	7.1E-05	0.24	0.34
	Mould	0.39	2.8E-07	0.37	1.8E-03	0.26	0.019	0.33	0.18
	Couch grass	0.66	5.2E-22	0.63	8.4E-09	0.5	3.3E-06	0.52	0.028
	Ryegrass	0.67	6.5E-23	0.59	1.2E-07	0.53	6.1E-07	0.56	0.016

**B. Phadiatop vs. allergen-specific IgG4.** \* Assay used at age 5 was the adult version, not Phadiatop infant.

Age (y)	Allergen	All clusters		CAS1		CAS2		CAS3	
		Rho	P-value	Rho	P-value	Rho	P-value	Rho	P-value
6m	HDM	0.14	0.029	NA	NA	0.053	0.59	0.38	0.081
	Cat	<b>0.38</b>	<b>2.7E-09</b>	<b>0.35</b>	<b>9.2E-04</b>	<b>0.29</b>	<b>2.8E-03</b>	<b>0.55</b>	<b>8.1E-03</b>
	Peanut	0.084	0.2	NA	NA	NA	NA	0.19	0.39
	Mould	0.064	0.33	0.18	0.1	-2.7E-02	0.78	0.28	0.21
	Couch grass	0.14	0.028	NA	NA	0.13	0.19	0.28	0.21
1	Ryegrass	0.1	0.12	NA	NA	NA	NA	0.28	0.21
	HDM	0.27	3.1E-05	NA	NA	0.31	1.1E-03	0.43	0.047
	Cat	<b>0.43</b>	<b>6.1E-12</b>	<b>0.29</b>	<b>6.1E-03</b>	<b>0.44</b>	<b>1.8E-06</b>	<b>0.58</b>	<b>4.5E-03</b>
	Peanut	0.22	5.8E-04	0.13	0.25	0.078	0.42	0.52	0.014
	Mould	0.085	0.19	NA	NA	NA	NA	0.25	0.26
2	Couch grass	0.3	3.2E-06	0.11	0.33	0.36	1.7E-04	0.44	0.04
	Ryegrass	0.15	0.023	0.056	0.61	0.088	0.37	0.33	0.13
	HDM	0.4	1.4E-09	0.057	0.6	<b>0.26</b>	<b>9.0E-03</b>	0.013	0.95
	Cat	<b>0.53</b>	<b>1.8E-17</b>	<b>0.43</b>	<b>3.5E-05</b>	0.2	0.04	<b>0.67</b>	<b>6.6E-04</b>
	Peanut	0.43	3.6E-11	0.28	9.4E-03	0.23	0.018	0.56	7.1E-03
	Mould	0.16	0.02	NA	NA	0.18	0.062	0.13	0.57

Continued on next page



Continued from previous page

Age (y)	Allergen	All clusters		CAS1		CAS2		CAS3	
		Rho	P-value	Rho	P-value	Rho	P-value	Rho	P-value
3	Couch grass	0.31	3.5E-06	0.13	0.23	0.21	0.035	0.44	0.039
	Ryegrass	0.2	2.4E-03	-1.1E-01	0.31	0.13	0.19	0.3	0.18
	HDM	0.49	2.6E-14	0.25	0.025	0.29	2.8E-03	-1.3E-02	0.95
	Cat	<b>0.56</b>	<b>1.9E-18</b>	<b>0.44</b>	<b>3.4E-05</b>	0.35	2.7E-04	<b>0.85</b>	<b>4.5E-07</b>
	Peanut	0.35	1.5E-07	0.14	0.21	0.019	0.85	0.23	0.31
	Mould	0.22	1.4E-03	0.14	0.22	0.36	2.1E-04	0.15	0.49
4	Couch grass	0.49	5.2E-14	0.3	5.2E-03	<b>0.49</b>	<b>2.7E-07</b>	0.64	1.5E-03
	Ryegrass	0.24	3.5E-04	NA	NA	0.18	0.071	0.042	0.85
	HDM	0.49	1.5E-12	0.34	2.2E-03	0.39	2.1E-04	0.17	0.47
	Cat	<b>0.58</b>	<b>2.0E-18</b>	<b>0.51</b>	<b>2.0E-06</b>	0.45	1.3E-05	0.62	2.7E-03
	Peanut	0.35	5.4E-07	0.19	0.087	0.19	0.072	0.61	3.4E-03
	Mould	0.25	6.5E-04	0.17	0.14	0.38	2.4E-04	0.49	0.024
5*	Couch grass	0.56	9.6E-17	0.44	6.7E-05	<b>0.6</b>	<b>5.1E-10</b>	<b>0.69</b>	<b>4.8E-04</b>
	Ryegrass	0.27	1.7E-04	0.13	0.25	0.28	8.0E-03	0.53	0.014
	HDM	0.56	2.1E-15	0.37	1.6E-03	0.53	5.7E-07	0.81	3.9E-05
	Cat	0.47	1.6E-10	0.34	3.9E-03	0.45	2.9E-05	0.97	1.5E-11
	Peanut	0.5	2.7E-12	0.4	6.6E-04	0.19	0.1	0.94	4.2E-09
	Mould	0.44	1.8E-09	0.19	0.11	0.34	2.0E-03	0.53	0.023
	Couch grass	0.57	8.1E-16	0.25	0.035	0.58	2.8E-08	<b>0.98</b>	<b>3.9E-13</b>
	Ryegrass	<b>0.68</b>	<b>4.5E-24</b>	<b>0.64</b>	<b>3.1E-09</b>	<b>0.59</b>	<b>1.2E-08</b>	0.74	4.6E-04

**TABLE B.7: Complete version of Table 4: Predictors for age-five wheeze within each CAS cluster, with demographic covariates (sex, BMI, parental history of asthma.**

BMI = body mass index; HDM = house dust mite; LRI = lower respiratory infection. Association analyses performed via generalised linear models (GLM) with demographic covariates: age-five wheeze ~ predictor + sex (male) + BMI at age 3 + paternal history of asthma + maternal history of asthma. Bold text indicates statistical significance ( $p < 0.05$ ); italics indicate near-significance  $p < 0.10$ ). \*Odds ratio (OR) is for every 10-fold increase in IgE, IgG4 or IgG.

Selected predictors for age-five wheeze	Age	CAS1 (N=88)		CAS2 (N=107)		CAS3 (N=22)		All (N=261)	
		OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value
ARI (events per y)	1	1.1 (0.88-1.5)	0.36	1.1 (0.87-1.3)	0.51	0.57 (0.29-0.93)	<b>0.046</b>	1 (0.89-1.2)	0.76
	2	1.1 (0.94-1.3)	0.22	1 (0.81-1.3)	0.82	0.43 (0.077-0.89)	0.12	1 (0.93-1.2)	0.44
	3	1.1 (0.87-1.3)	0.58	1.1 (0.91-1.4)	0.3	0.67 (0.36-1)	0.1	1 (0.93-1.2)	0.48
	4	1.2 (0.99-1.4)	<i>0.074</i>	1.2 (1-1.5)	<b>0.032</b>	0.63 (0.27-1.1)	0.15	1.2 (1-1.3)	<b>0.013</b>
LRI (events per y)	1	0.97 (0.71-1.3)	0.84	1 (0.61-1.5)	0.99	0.48 (0.13-1.1)	0.16	1 (0.81-1.2)	0.92
	2	1.2 (0.88-1.6)	0.26	1.5 (0.97-2.5)	<i>0.069</i>	0.99 (0.34-2.6)	0.98	1.4 (1.1-1.7)	<b>5.3E-03</b>
	3	2 (1.3-3.2)	<b>2.3E-03</b>	2.6 (1.5-5.3)	<b>2.7E-03</b>	0.98 (0.4-2.6)	0.96	2 (1.5-2.7)	<b>3.8E-06</b>
	4	2 (1.4-3.4)	<b>2.0E-03</b>	3.6 (1.8-8.3)	<b>6.5E-04</b>	1.9 (0.57-8.4)	0.32	2.5 (1.8-3.6)	<b>1.5E-07</b>
Wheezy LRI (events per y)	1	1.3 (0.68-2.4)	0.43	1.1 (0.35-3)	0.83	2.6 (0.62-58)	0.34	1.5 (0.98-2.3)	<i>0.06</i>
	2	1.2 (0.8-2)	0.33	1.6 (0.89-2.9)	0.12	2.4 (0.67-16)	0.24	1.6 (1.2-2.2)	<b>5.6E-03</b>
	3	2.8 (1.6-5.6)	<b>1.3E-03</b>	3 (1.4-8)	<b>0.016</b>	1.2 (0.43-4.6)	0.76	2.7 (1.8-4.2)	<b>4.1E-06</b>
	4	2.5 (1.5-5)	<b>4.0E-03</b>	6.3 (2.5-21)	<b>6.8E-04</b>	7.1 (1.2-169)	0.1	3.9 (2.5-6.7)	<b>5.4E-08</b>
Febrile LRI (events per y)	1	1.6 (0.77-3.6)	0.21	0.84 (0.28-1.9)	0.71	7.3 (0.78-178)	0.12	1.5 (0.93-2.4)	<i>0.098</i>
	2	1 (0.44-2.2)	1	4.8 (1.8-15)	<b>3.9E-03</b>	1.6 (0.48-10)	0.5	2.3 (1.4-3.9)	<b>1.2E-03</b>
	3	2 (1-4.8)	<i>0.08</i>	4.3 (1.2-15)	<b>0.02</b>	4.2 (0.55-519)	0.37	2.4 (1.4-4.3)	<b>2.3E-03</b>
	4	1.8 (0.97-4.1)	<i>0.092</i>	2.6 (0.88-8.3)	<i>0.082</i>	1.1 (0.11-18)	0.93	2.2 (1.3-4)	<b>5.9E-03</b>
% Healthy NPAs with infection-associated MPCs	0-2	0.9 (0.13-5.7)	0.91	2.6 (0.43-16)	0.3	NA	NA	2.3 (0.79-6.7)	0.13

Continued on next page

Continued from previous page

Selected predictors for age-five wheeze	Age	CAS1 (N=88)		CAS2 (N=107)		CAS3 (N=22)		All (N=261)	
		OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value
Quartile of % healthy NPAs with infection-associated MPGs	2-4	0.086 (6.8E-03-0.71)	<b>0.034</b>	0.8 (0.077-7.5)	0.85	4.4E+03 (2.1-2.5E+12)	0.13	0.49 (0.14-1.6)	0.24
	0-2	1 (0.54-1.8)	0.98	1.3 (0.72-2.4)	0.36	NA	NA	1.3 (0.89-1.8)	0.19
	2-4	0.45 (0.19-0.88)	<b>0.035</b>	1 (0.51-2.1)	0.9	NA	NA	0.8 (0.53-1.2)	0.24
HDM IgE (kU/L)*	6m	8 (0.85-94)	0.074	0.93 (0.14-3.6)	0.92	3.4 (0.26-180)	0.4	2.3 (0.99-5.8)	0.054
	1	1.5 (0.22-7.8)	0.65	0.54 (0.039-2.3)	0.51	39 (2.5-22000)	0.082	2.7 (1.5-5.5)	<b>0.00089</b>
	2	0.93 (0.28-2.5)	0.89	2 (1.2-3.7)	<b>0.016</b>	1.4 (0.38-4.8)	0.62	2 (1.5-2.8)	<b>2.80E-05</b>
	3	1.4 (0.68-2.9)	0.32	1.5 (0.9-2.4)	0.12	1.5 (0.4-5.2)	0.55	1.7 (1.3-2.2)	<b>1.00E-04</b>
	4	1.9 (0.94-4.1)	0.086	1.9 (1.2-3.1)	<b>0.011</b>	1.4 (0.31-5.5)	0.64	1.9 (1.5-2.5)	<b>3.70E-06</b>
Peanut IgE (kU/L)*	6m	2.5 (0.78-9)	0.13	1.5 (0.54-3.8)	0.41	1.1 (0.3-3.7)	0.92	2.3 (1.4-3.9)	<b>0.0014</b>
	1	1.7 (0.48-6.3)	0.39	2.2 (0.65-6.9)	0.19	0.47 (0.095-1.6)	0.27	2.2 (1.4-3.6)	<b>0.00098</b>
	2	0.51 (0.097-2)	0.37	3 (0.74-12)	0.12	2 (0.51-13)	0.37	2.7 (1.6-4.9)	<b>0.00046</b>
	3	1.7 (0.46-5.5)	0.37	0.53 (0.015-3.8)	0.61	3.3 (0.94-26)	0.13	2.6 (1.6-4.8)	<b>0.00068</b>
	4	0.2 (0.00073-2.9)	0.36	0.96 (0.19-3.2)	0.95	1.4 (0.49-6.5)	0.54	2.1 (1.3-3.7)	<b>0.006</b>
Cat IgE (kU/L)*	6m	6.6 (0.77-61)	0.079	2.2 (0.62-7.6)	0.2	0.24 (0.012-3.2)	0.29	2.3 (0.96-5.4)	0.061
	1	2.1 (0.13-30)	0.57	4 (0.54-32)	0.16	0.45 (0.053-2.8)	0.41	3.5 (1.4-9.5)	<b>0.0099</b>
	2	0.55 (0.042-3.7)	0.57	2.1 (0.59-7)	0.22	2.2 (0.42-26)	0.42	2.6 (1.3-5.5)	<b>0.0065</b>
	3	1.7 (0.49-5.6)	0.35	1.4 (0.21-6.7)	0.66	1.3 (0.29-6.9)	0.77	2.5 (1.3-4.9)	<b>0.0065</b>
	4	0.75 (0.0088-13)	0.86	1.5 (0.53-3.9)	0.4	0.83 (0.17-4.4)	0.81	2.4 (1.3-4.8)	<b>0.006</b>
Couch grass IgE (kU/L)*	6m	2.8 (0.51-14)	0.21	1.3 (0.3-4.5)	0.68	0.98 (0.048-59)	0.99	1.7 (0.71-3.9)	0.22
	1	0.38 (0.017-2.8)	0.42	0.33 (0.01-2.9)	0.41	0.15 (0.0058-1.5)	0.14	0.63 (0.19-1.7)	0.4
	2	0.085 (0.0034-0.7)	0.057	1.1 (0.14-6.3)	0.9	25 (1.6-1100)	<b>0.046</b>	2.1 (0.99-4.7)	0.053
	3	2 (0.44-8)	0.29	6.1e-06 (NA-8.1e+54)	0.99	2.3 (0.57-14)	0.29	2.5 (1.3-5.1)	<b>8.90E-03</b>
	4	8.4e-13 (NA-3.5e+172)	0.99	1.6 (0.55-4.1)	0.34	1.9 (0.54-10)	0.35	2 (1.3-3.4)	<b>4.30E-03</b>
Phadiatop IgE (PAU/L)*	6m	1.2 (0.44-2.9)	0.73	1.3 (0.65-2.6)	0.43	2.2 (0.66-12)	0.25	2 (1.3-2.9)	<b>0.00078</b>
	1	0.73 (0.2-2.5)	0.63	1.1 (0.41-2.8)	0.85	1.6 (0.23-18)	0.67	2.1 (1.3-3.4)	<b>0.0021</b>
	2	0.33 (0.091-1)	0.065	2.1 (0.81-5.9)	0.13	2.5 (0.18-70)	0.52	2 (1.3-3)	<b>0.0012</b>
	3	1.8 (0.8-4)	0.16	1.4 (0.72-2.8)	0.31	8.4 (0.53-380)	0.19	2 (1.4-2.9)	<b>8.00E-05</b>
	4	1.8 (0.91-3.8)	0.094	2.4 (1.3-4.8)	<b>0.01</b>	2.7 (0.16-66)	0.5	2.2 (1.6-3.2)	<b>2.20E-06</b>
HDM IgG4 (µg/L)*	6m	NA (NA-NA)	0.55	0.053 (NA-6.5e+24)	0.99	28 (1.7e-34-NA)	0.99	1.4 (0.88-2.6)	0.17
	1	NA (NA-NA)	0.61	1.1 (0.8-1.5)	0.5	0.9 (0.58-1.3)	0.6	1.2 (1-1.4)	0.053
	2	1.1 (0.71-1.6)	0.67	1.1 (0.85-1.4)	0.61	0.4 (0.038-1.2)	0.26	1.1 (1-1.3)	0.056
	3	1.1 (0.85-1.5)	0.35	1.1 (0.77-2)	0.64	0.94 (0.19-2.3)	0.9	1.1 (0.98-1.2)	0.1
	4	1.2 (0.98-1.5)	0.082	0.89 (0.7-1.1)	0.33	0.46 (0.031-5.4)	0.53	1.1 (1-1.3)	<b>0.034</b>
Peanut IgG4 (µg/L)*	6m	NA (NA-NA)	0.55	NA (NA-NA)	0.53	0.9 (0.42-1.9)	0.76	1.5 (0.94-2.6)	0.1
	1	0.075 (NA-3.5e+23)	0.99	0.89 (0.67-1.1)	0.35	0.96 (0.64-1.4)	0.84	1.1 (0.95-1.2)	0.22
	2	1.1 (0.85-1.3)	0.54	0.96 (0.8-1.2)	0.64	0.89 (0.48-1.4)	0.65	1 (0.95-1.2)	0.37
	3	1.1 (0.89-1.4)	0.37	1 (0.83-1.3)	0.87	0.68 (0.22-1.3)	0.37	1.1 (0.96-1.2)	0.27
	4	1.1 (0.92-1.4)	0.22	0.91 (0.76-1.1)	0.35	0.73 (0.19-1.4)	0.45	1.1 (0.96-1.2)	0.24
Cat IgG4 (µg/L)*	6m	0.057 (NA-2e+12)	0.99	0.99 (0.67-1.3)	0.95	24 (3.3e-30-NA)	1	1.1 (0.88-1.3)	0.41
	1	0.76 (0.43-1.1)	0.22	0.94 (0.78-1.1)	0.54	0.76 (0.42-1.2)	0.28	1 (0.9-1.1)	0.82

Continued on next page

Continued from previous page

Selected predictors for age-five wheeze	Age	CAS1 (N=88)		CAS2 (N=107)		CAS3 (N=22)		All (N=261)	
		OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value
Couch grass IgG4 ( $\mu\text{g/L}$ )*	2	1.4 (1.1-1.7)	<b>0.011</b>	0.92 (0.67-1.3)	0.59	0.96 (0.51-1.6)	0.88	1.1 (1-1.3)	0.053
	3	1.3 (1-1.6)	0.05	0.9 (0.63-1.4)	0.59	0.86 (0.054-13)	0.91	1.2 (1-1.4)	<b>0.033</b>
	4	1.4 (1.1-2)	<b>0.027</b>	0.89 (0.64-1.3)	0.49	0.54 (0.011-1.5)	0.58	1.2 (1-1.5)	<b>0.034</b>
	6m	NA (NA-NA)	0.55	0.062 (NA-1.3e+24)	0.99	19 (2.5e-57-NA)	1	1.3 (0.74-2.4)	0.32
	1	0.081 (NA-9.7e+23)	0.99	1 (0.77-1.3)	0.81	0.93 (0.6-1.4)	0.71	1.1 (0.92-1.3)	0.29
	2	0.071 (NA-2.1e+22)	0.99	0.88 (0.7-1.1)	0.22	0.91 (0.61-1.3)	0.61	1 (0.88-1.1)	0.96
	3	1.2 (0.99-1.6)	0.061	0.85 (0.7-1)	0.1	1.4 (0.88-2.2)	0.16	1.1 (0.96-1.2)	0.22
	4	1.1 (0.91-1.4)	0.28	0.72 (0.56-0.91)	<b>0.0074</b>	0.88 (0.24-1.9)	0.75	1 (0.91-1.2)	0.69
Phadiatop Infant IgG4 (PAU/L)*	6m	0.7 (0.45-0.91)	<b>0.03</b>	1 (0.88-1.2)	0.79	1.4 (0.96-2.4)	0.12	0.98 (0.89-1.1)	0.67
	1	0.91 (0.72-1.2)	0.4	0.73 (0.49-0.99)	0.057	0.83 (0.29-1.5)	0.64	0.93 (0.81-1.1)	0.35
	2	1.1 (0.89-1.3)	0.49	0.97 (0.68-1.6)	0.86	1.7 (0.93-7.7)	0.2	1.1 (0.96-1.3)	0.2
	3	2.3 (1.1-6.8)	0.091	0.23 (0.071-0.64)	<b>0.0076</b>	1 (0.17-7.3)	1	1.3 (0.96-1.8)	0.16
	4	1 (0.83-1.4)	0.71	0.3 (0.097-0.85)	<b>0.028</b>	0.42 (0.042-3.2)	0.4	1.1 (0.88-1.3)	0.61
HDM IgG (mg/L)*	1	25 (0.32-1.6E+04)	0.19	3.3 (0.16-46)	0.38	5.6E-03 (8.4E-06-0.57)	0.058	2 (0.31-11)	0.44
	2	0.8 (0.15-3.5)	0.78	0.97 (0.24-3.7)	0.96	0.79 (0.031-18)	0.88	1.3 (0.6-2.9)	0.48
	3	2.3 (0.14-35)	0.54	0.48 (0.057-2.5)	0.43	3.9 (0.26-96)	0.34	2.1 (0.89-5)	0.089
Cat IgG (mg/L)*	1	1.5E-15 (NA-1.2E+291)	0.99	6.5 (0.22-150)	0.24	4.6E-03 (1.4E-06-0.9)	0.082	1.7 (0.11-18)	0.68
	2	0.66 (0.077-3.5)	0.65	1.2 (0.28-4.3)	0.82	0.16 (4.0E-03-3.5)	0.26	0.87 (0.34-2.1)	0.75
	3	0.023 (8.2E-06-2)	0.18	0.52 (0.058-2.7)	0.49	3.7 (0.18-244)	0.44	1.1 (0.35-3)	0.9



## Appendix C

# Supplementary Figures and Tables for Chapter 4

The rest of this page has been intentionally left blank.

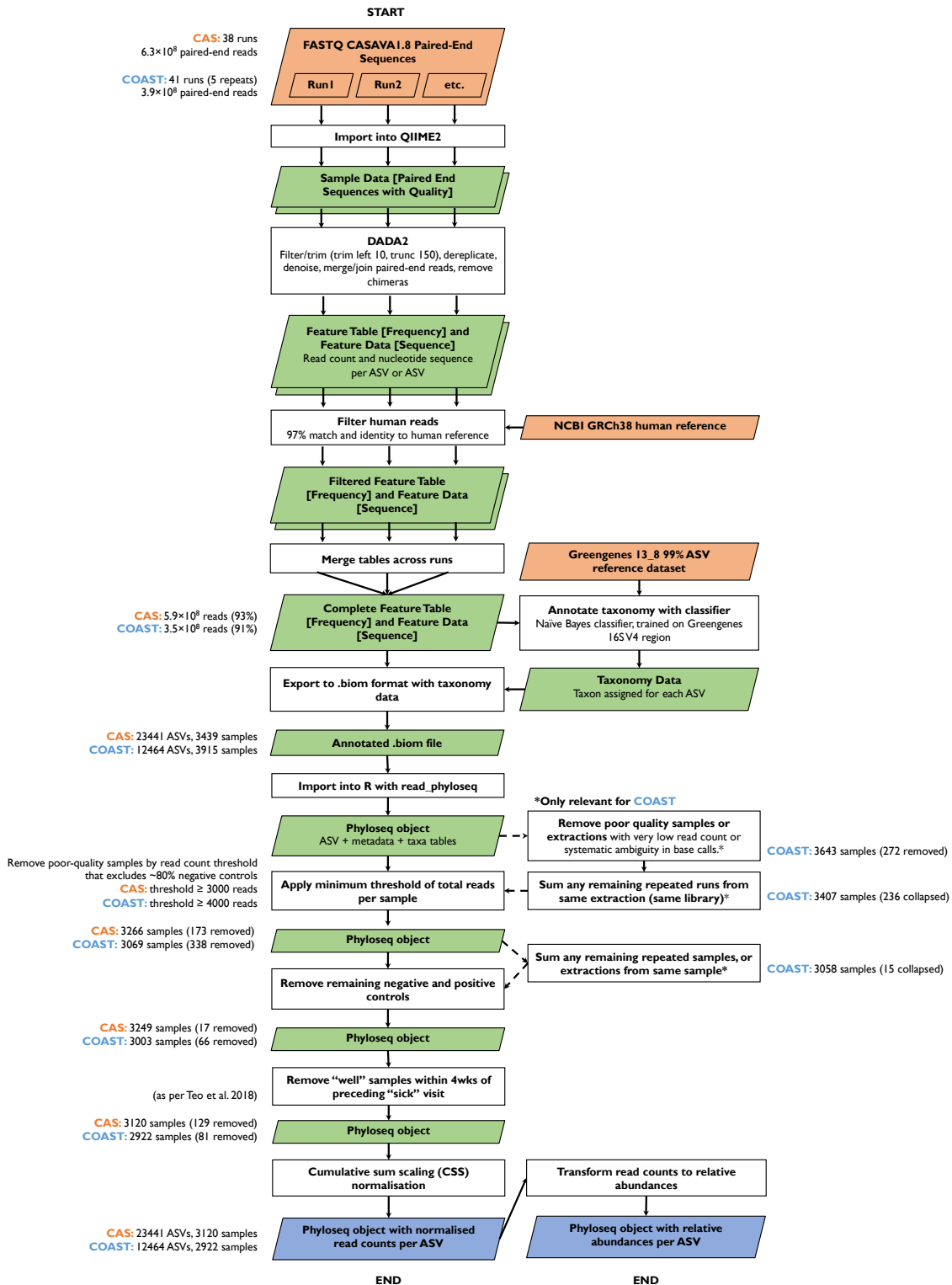
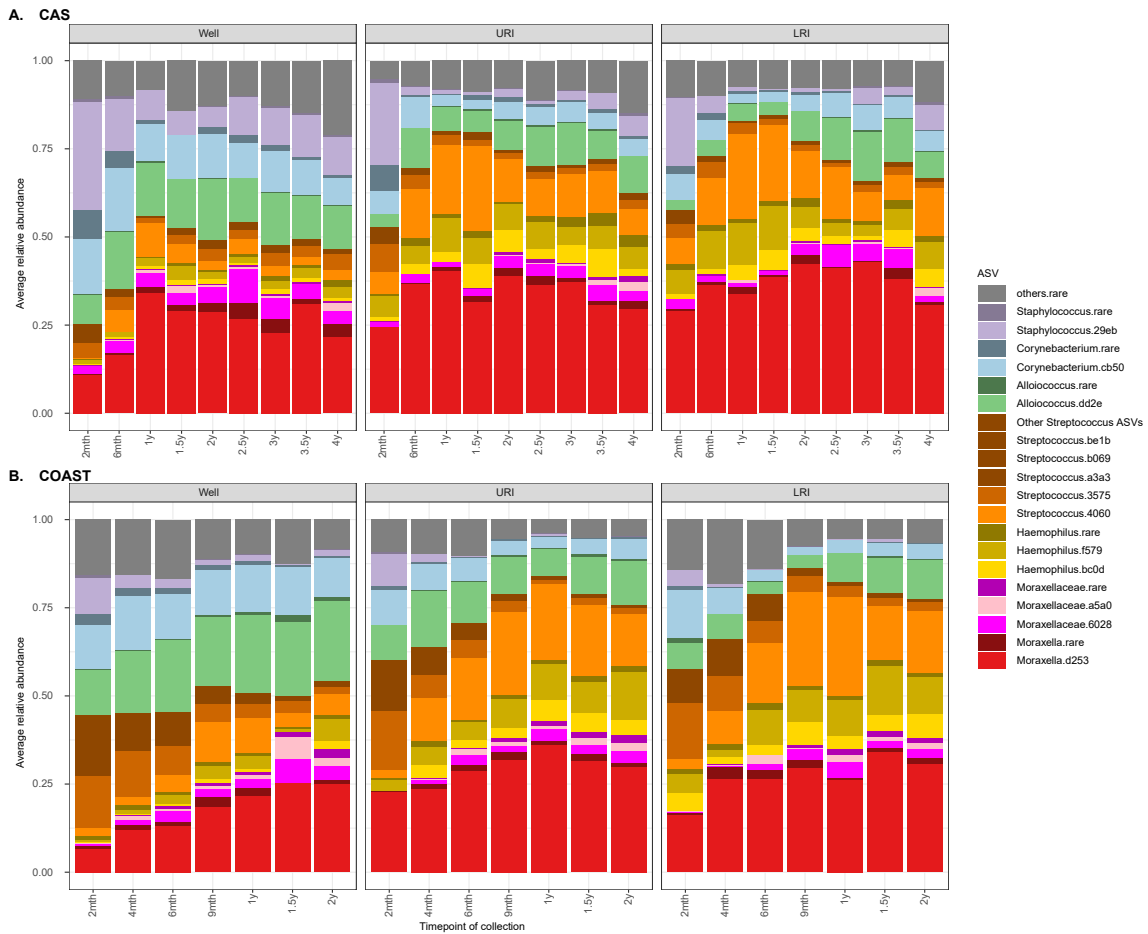


FIGURE C.1: Bioinformatic pipeline for processing and analyzing CAS and COAST 16S rRNA data, using QIIME2 and the “microbiome” R package



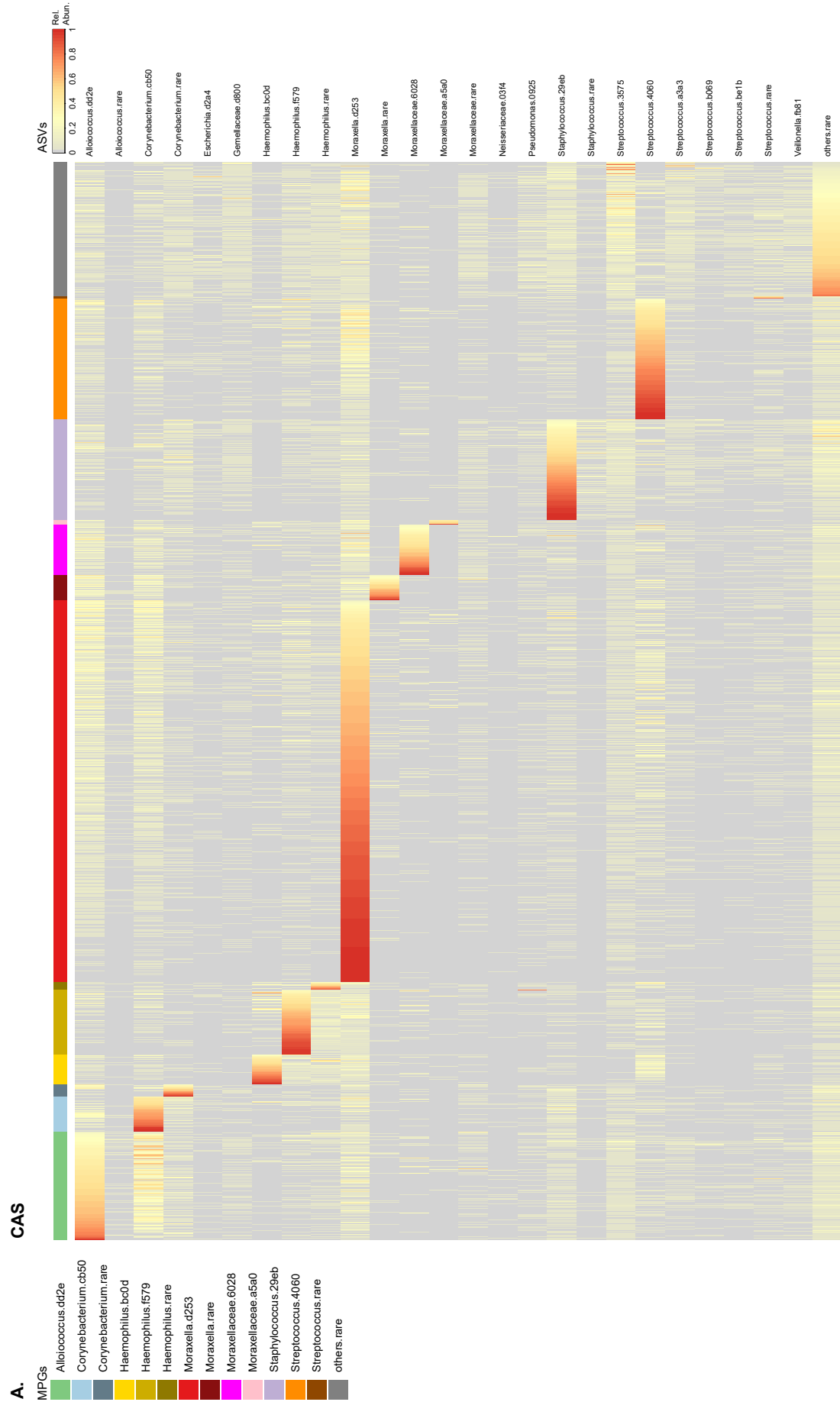
**FIGURE C.2: Distribution of ASVs (average relative abundance) in healthy and illness samples, within (A) CAS and (B) COAST.**

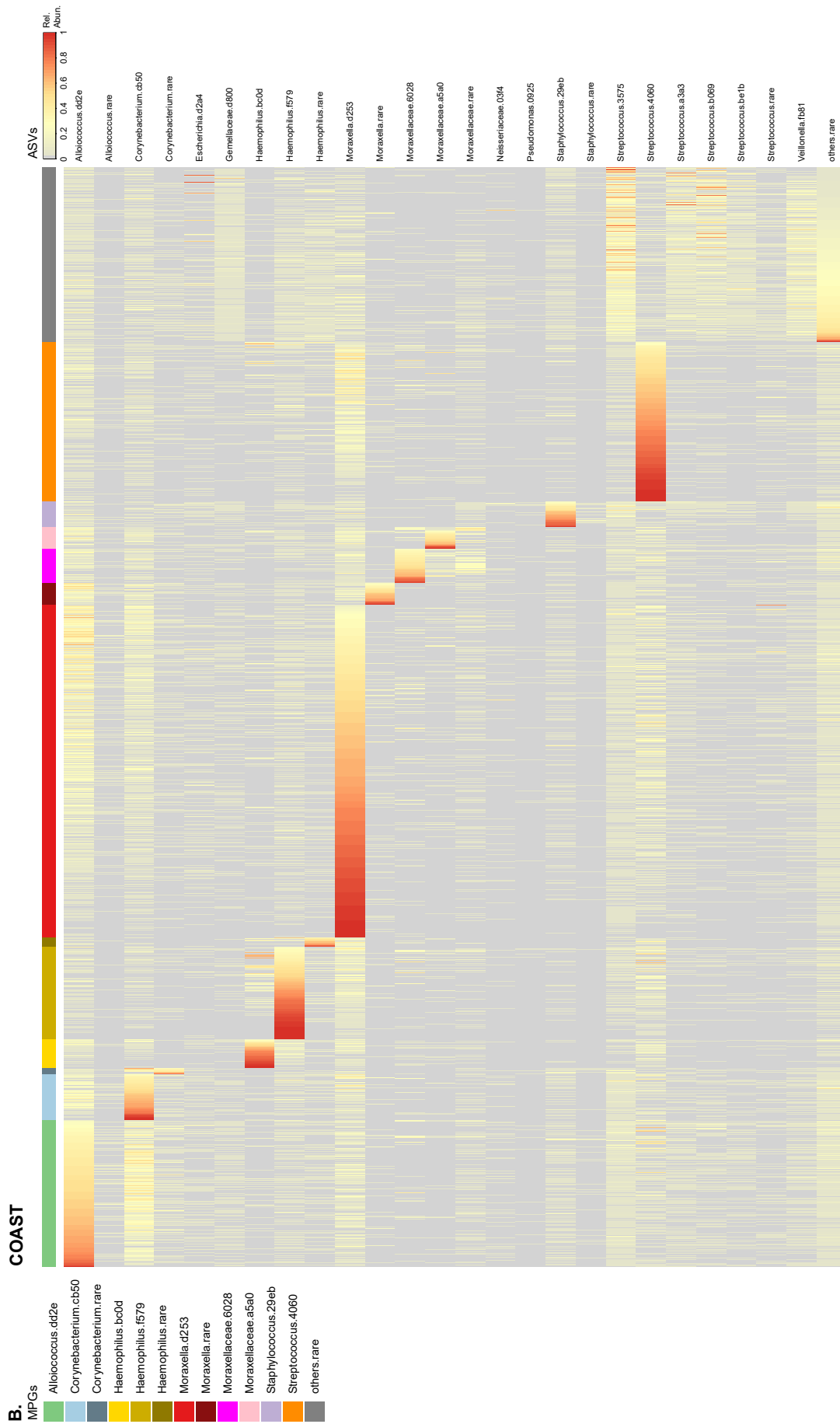
ASV = Amplicon sequence variant; LRI = Lower respiratory illness or infection; URI = upper respiratory illness or infection. See main text for definitions of LRI and URI. Note the different time scales for the timepoint of collection in CAS vs. COAST.

**FIGURE C.3: (next two pages) Heatmaps of MPGs with relative abundances of all common ASVs — complete versions of Figure 4.1. Samples clustered into MPGs in (A) CAS and (B) COAST.**

(See next two pages). ASV = Amplicon sequence variant; MPG = Microbiome profile group; Rel. abund. = Relative abundance. The coloured horizontal bar above the heatmap represents MPGs to which samples were assigned by hierarchical clustering.

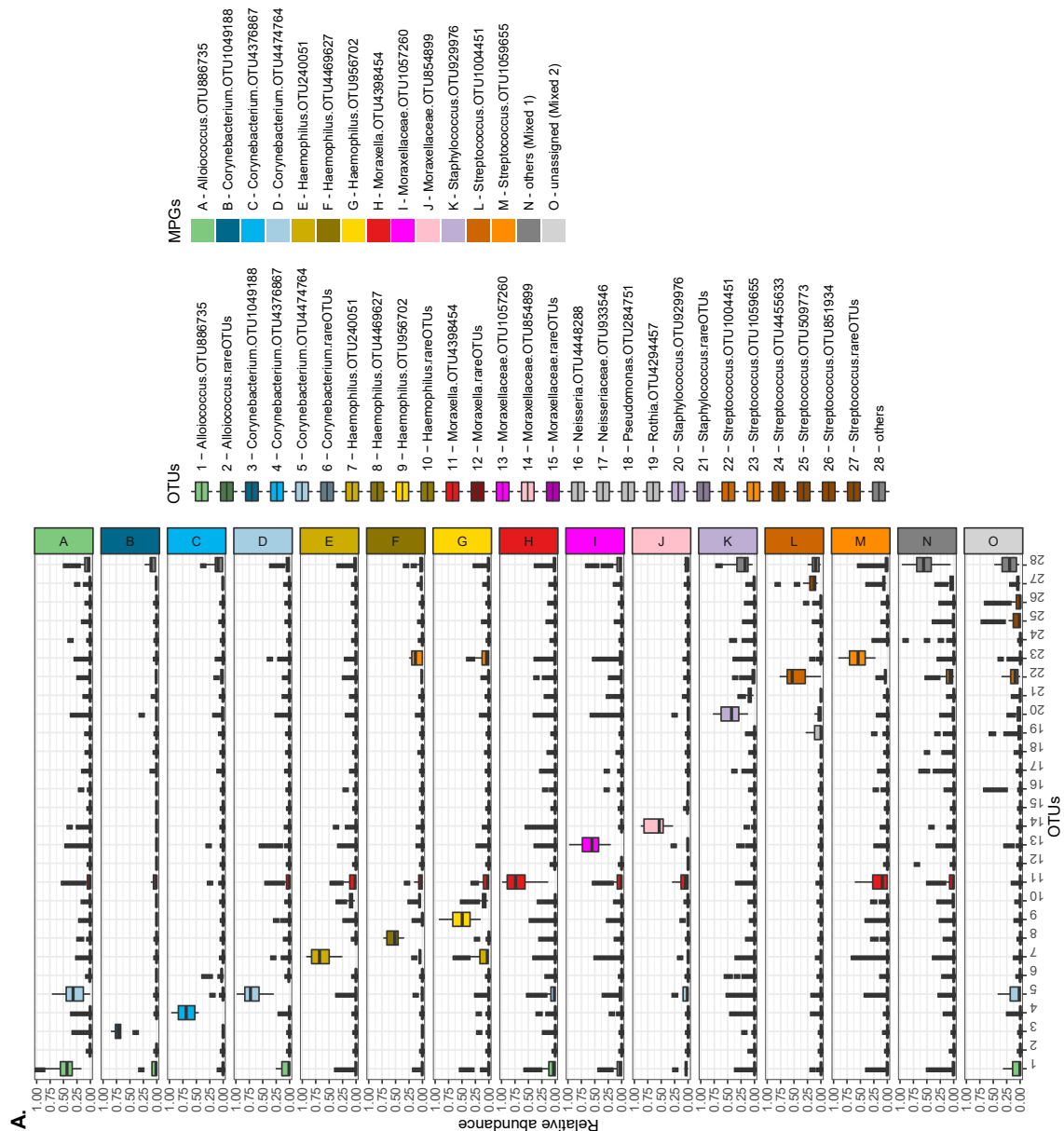


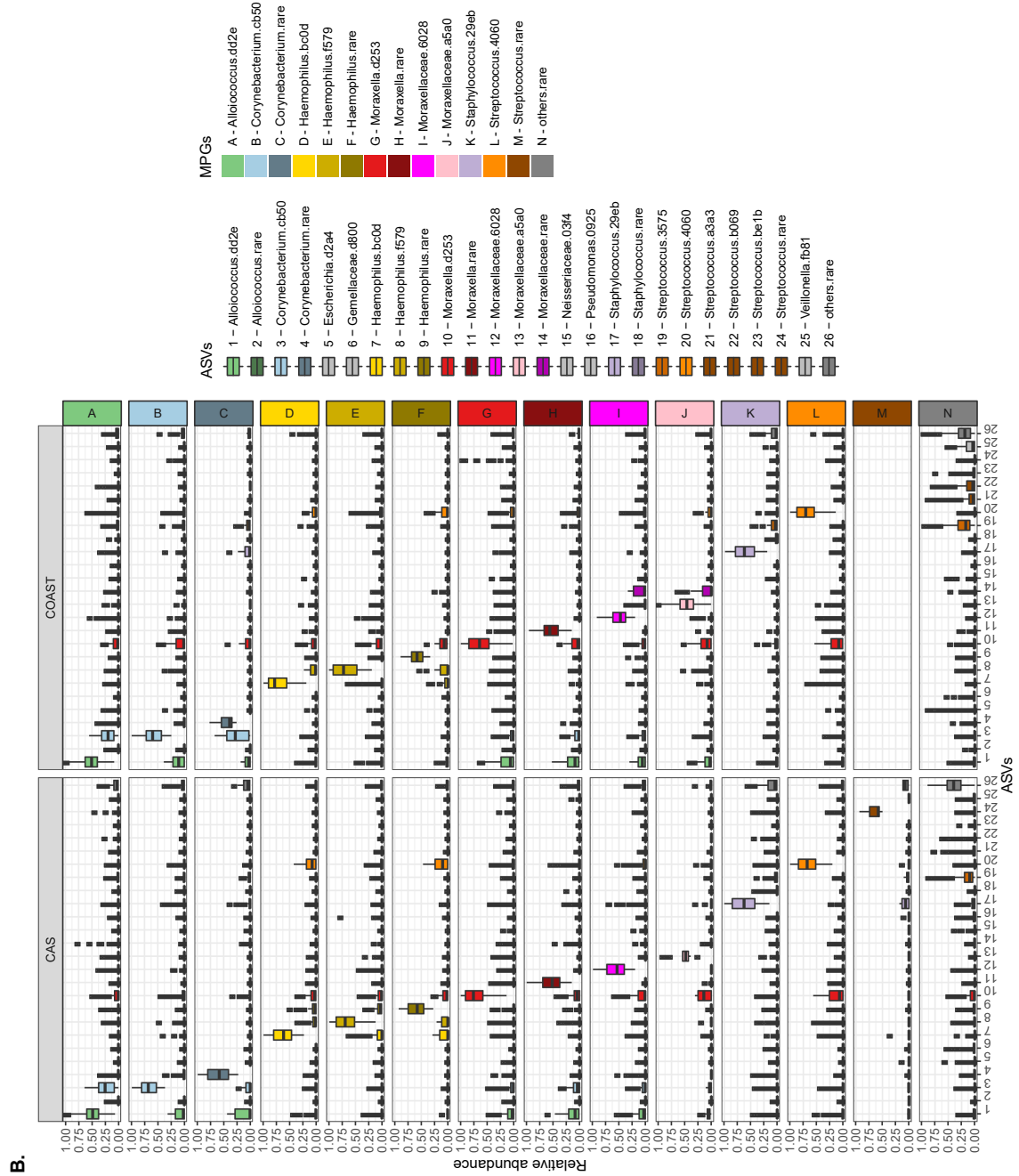


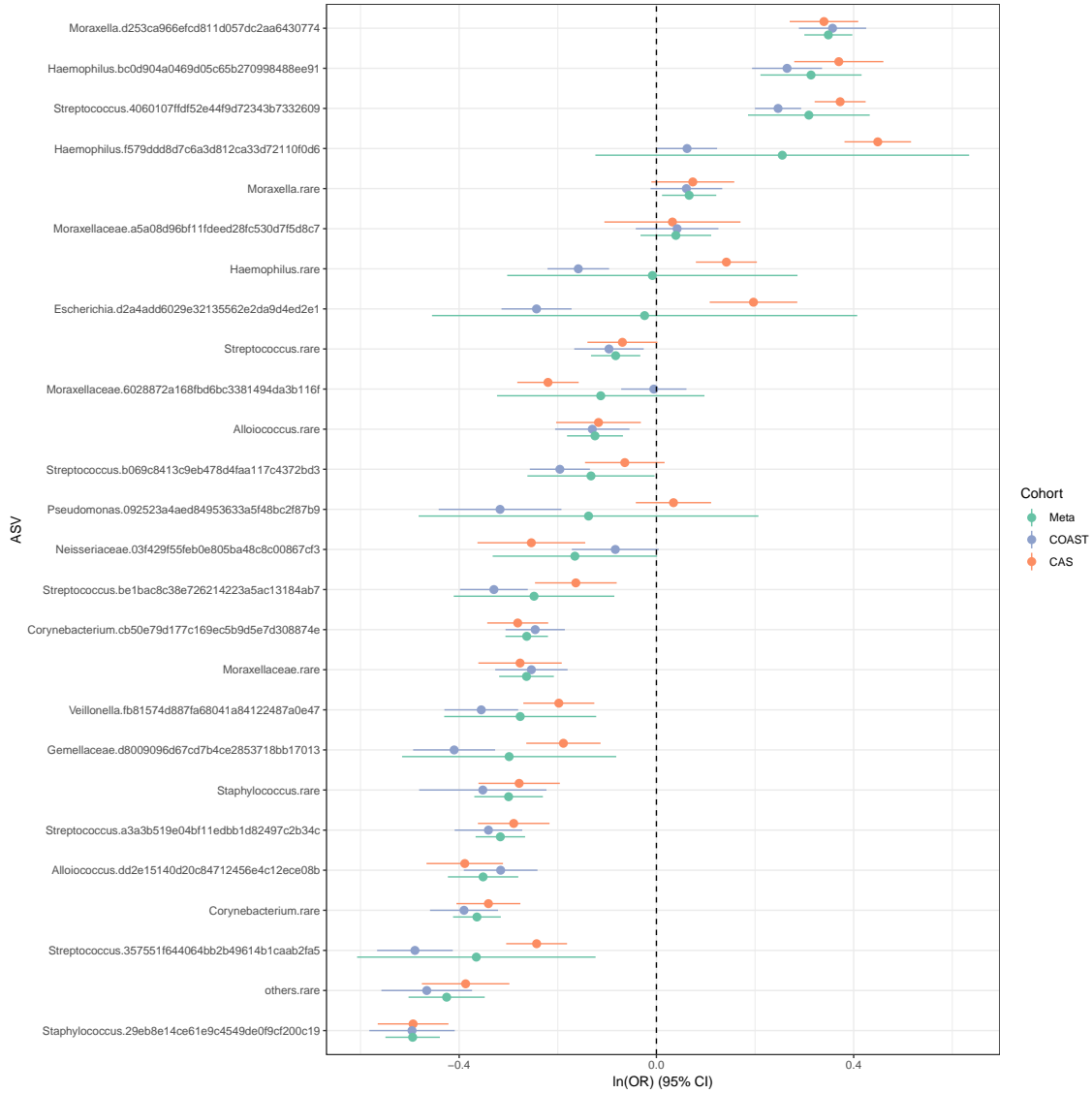


**FIGURE C.4: (next two panels) Relative abundance of each OTU or ASV within all samples of each MPG in QIIME1 CAS (A), and QIIME2 CAS and COAST (left and right; B).**

(See next two pages). ASV = Amplicon sequence variant; MPG = Microbiome profile group; OTU = Operational taxonomic unit; Rel. abun. = Relative abundance. Boxplots represent relative abundance of each OTU or ASV across all samples of each MPG. Each boxplot is in the style of Tukey i.e. horizontal line represents median, box represents IQR, whiskers represent  $1.5 \times \text{IQR}$ , points beyond whiskers represent outliers. Figure is organized with horizontal axis / numbers representing ASVs; vertical axis relative abundance; horizontal facets cohorts (CAS vs. COAST); and vertical facets / letters MPGs. It can be observed that for many MPGs, most of the non-dominant OTUs or ASVs are zero-inflated, with many visible outliers and barely-visible boxes. Note that COAST (QIIME2) did not have a *Streptococcus*.rare-dominated MPG.

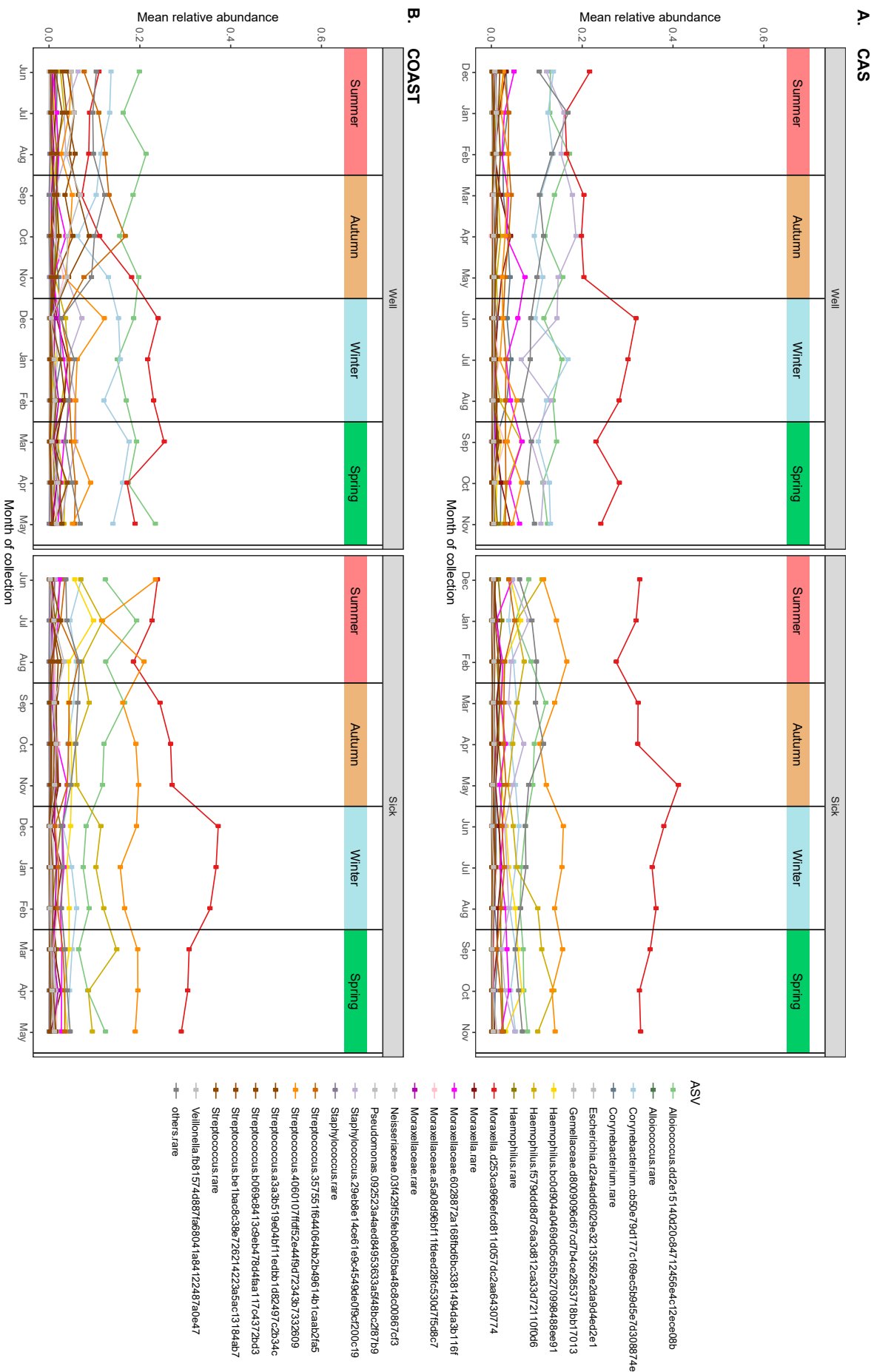






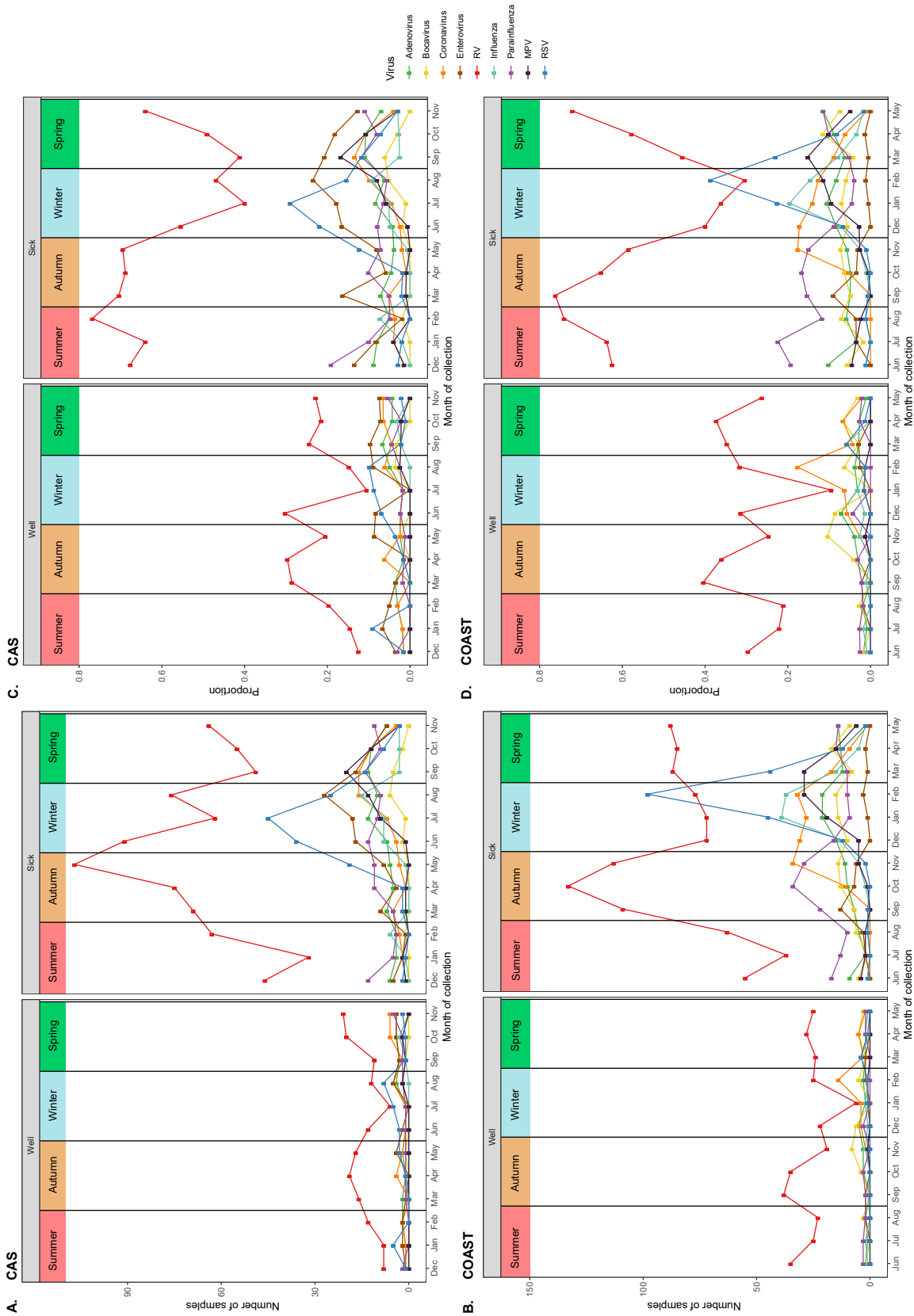
**FIGURE C.5: Meta-analysis and forest plots of GEE associations between ASVs and respiratory health vs. illness status at time of sample collection.**

ASV = Amplicon sequence variant; Meta = meta-analysis; OR = odds ratio. Cohort-specific ORs and CIs were generated based on the following GEE model for each ASV: respiratory illness status (well vs. unwell) ~ ASV + gender + age + season | subject (See **Supplementary Table C.1**). Meta-analysis was performed for each ASV with random effects and inverse variance weights. Above figure shows the subsequent forest plots, sorted in descending order of meta-analysis ORs. In each forest plot, points indicate natural log of OR, and error bars indicate 95% CI. Dotted line indicates null hypothesis (OR = 1).



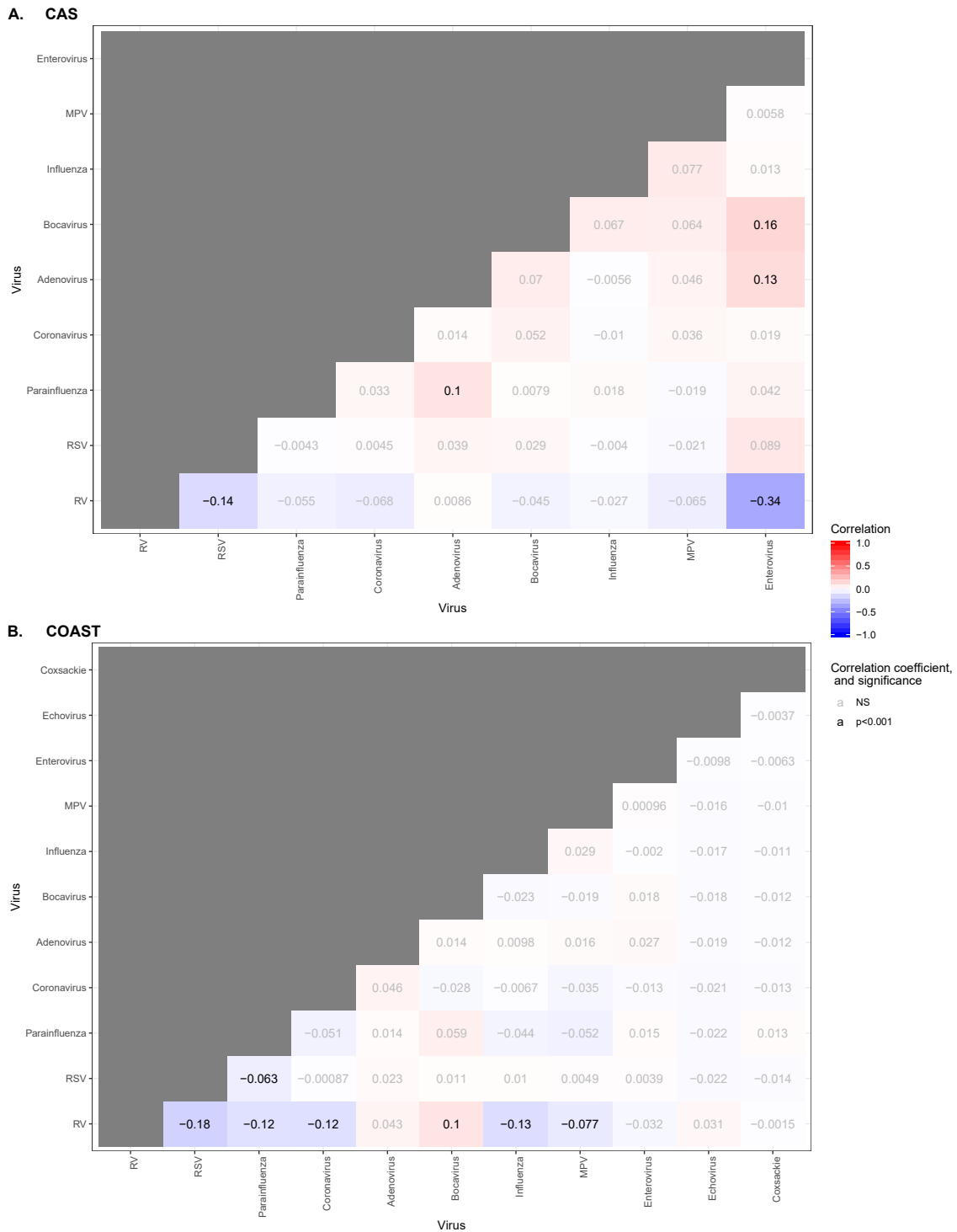
**FIGURE C.6: Distribution of ASVs (relative abundance) in healthy and illness samples, arranged by season and month of the year, within (A) CAS and (B) COAST.**

LRI = Lower respiratory illness or infection; MPG = Microbiome profile group; URI = upper respiratory illness or infection. See main text for definitions of LRI and URI. Note the different ordering months and seasons in CAS (cohort from the southern hemisphere) versus COAST (northern hemisphere).



**FIGURE C.7: Distribution of viruses in healthy and illness samples, arranged by season and month of the year; count of samples within (A) CAS (age 2-3y) and (B) COAST (age up to 3y); proportion of samples within (C) CAS (age 2-3y) and (D) COAST (age up to 3y)**

MPV = Human metapneumovirus; RSV = Respiratory syncytial virus; RV = Rhinovirus.

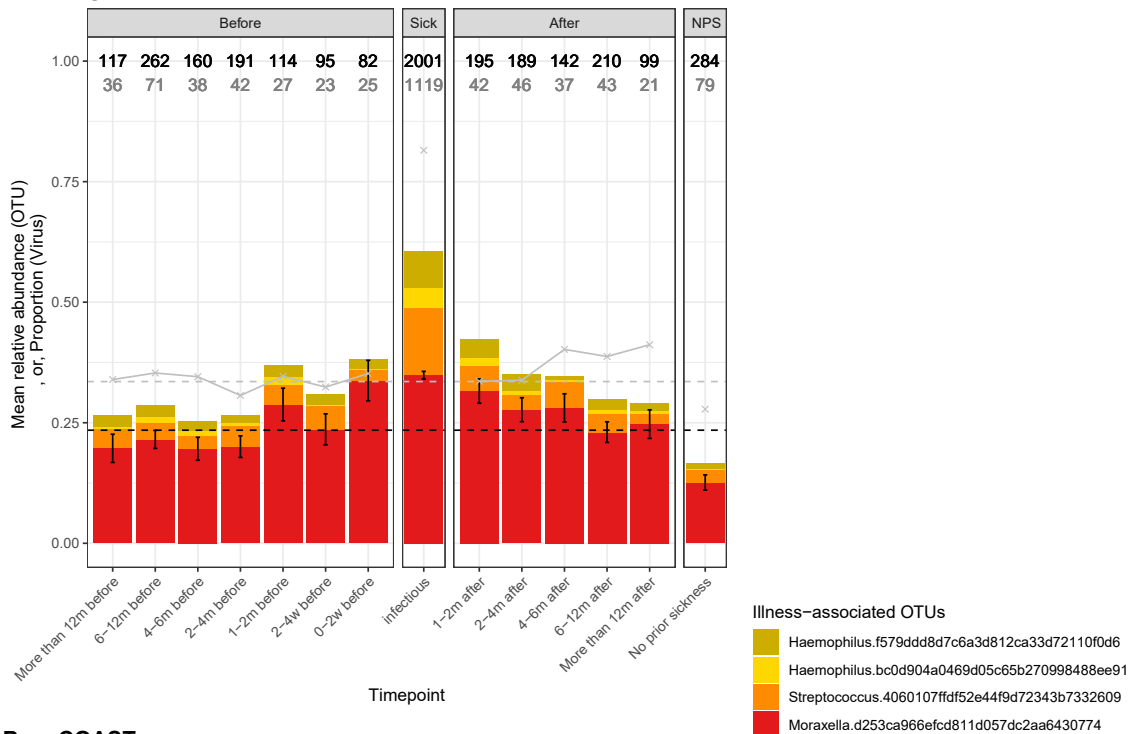


**FIGURE C.8: Correlation of viruses detected within nasopharyngeal samples during the first 3 years of life, in (A) CAS and (B) COAST.**

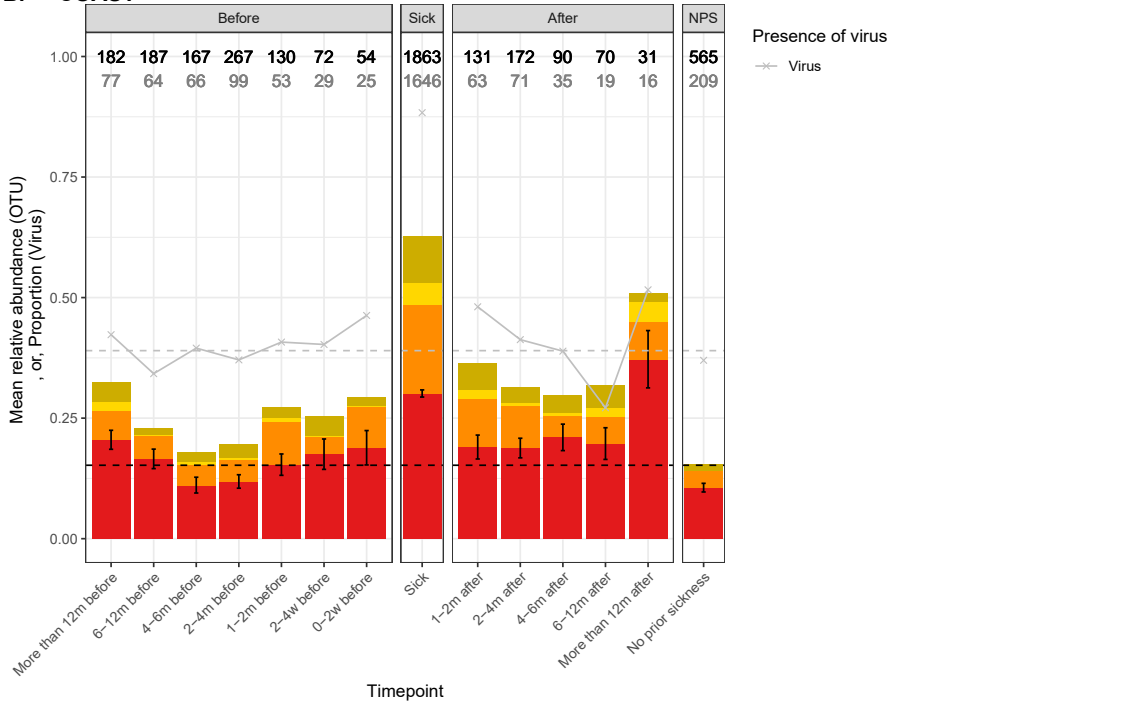
MPV = Human metapneumovirus; RSV = Respiratory syncytial virus; RV = Rhinovirus. Heat and number in each cell indicates magnitude of correlation coefficient (Spearman Rho); statistical significance of each correlation is indicated by bolded (significant at  $p < 0.001$ ) or greyed font (non-significant).



**A. CAS**



**B. COAST**



**FIGURE C.9: Relative abundance of ASVs and proportion of viruses before, during, and after an acute respiratory infection (ARI), in (A) CAS and (B) COAST.**

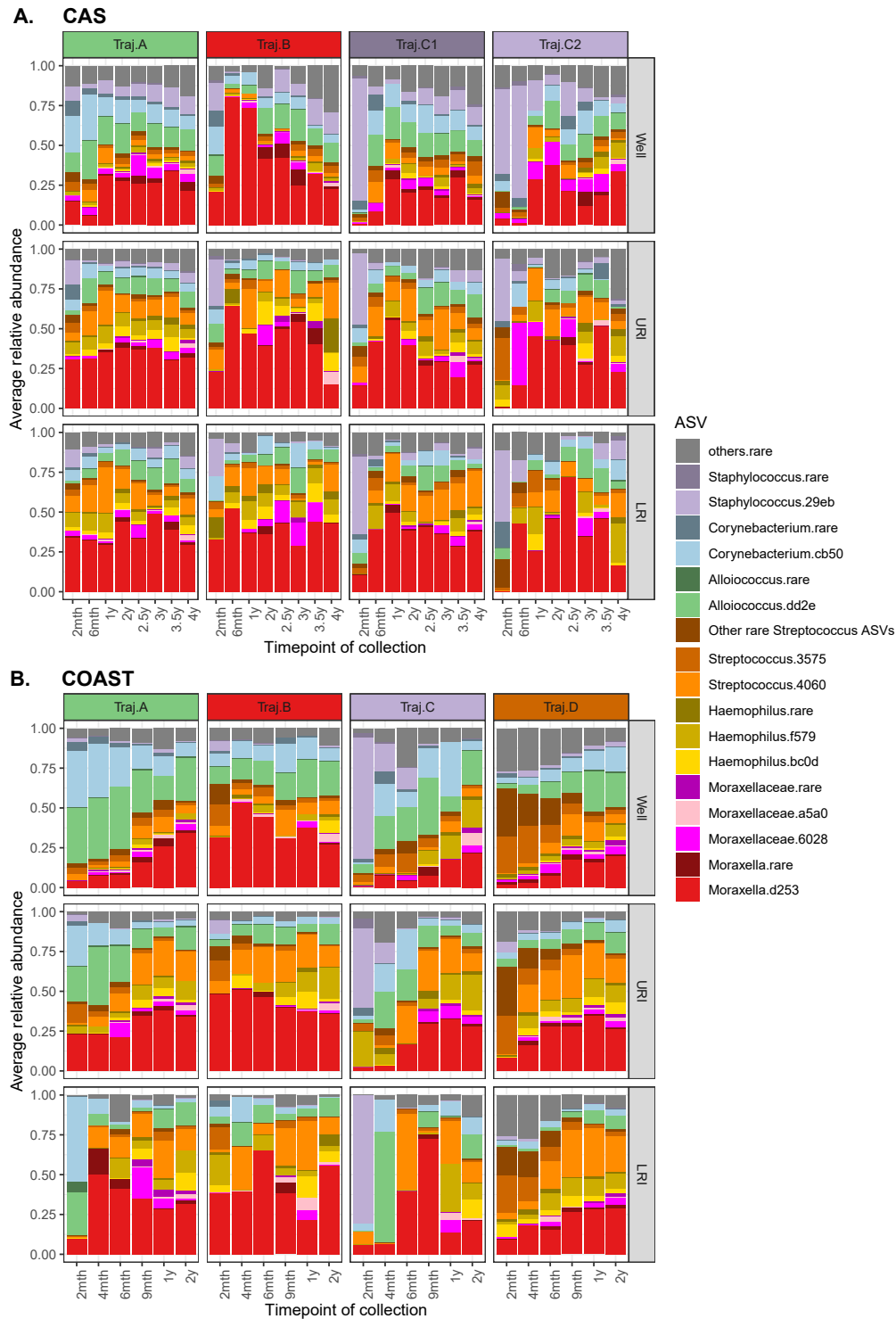
ARI = Acute respiratory illness; ASV = Amplicon sequence variant. NPS = Samples with no prior sickness. Grey line and crosses represent mean proportion of virus in each timepoint category. Coloured bars represent average relative abundances of ASV as indicated in legend. Dotted line represents mean relative abundance for *Moraxella.d253* ASV (black), or mean proportion of virus (grey) across all samples in the cohort. Black error bars represent standard errors for relative abundance of *Moraxella.d253*. Numbers at the top indicate total number of samples (black), and number of samples with virus (grey) respectively.

**FIGURE C.10: (next two pages) Trajectories in the nasopharyngeal microbiome as determined by Multiple Factor Analysis (MFA) and K-means clustering, in (A) CAS and (B) COAST.**

(See next two pages). Columns represent clustering features, arranged by ASV and timepoint (left-to-right: 2m to 2y within each ASV). Columns are labelled by the row of colours to the top of the heatmap, representing the colour coding of clustering features by ASV. Column of colours to the far left of heatmap represent Trajectories: Green=Traj.A (Early *Alloiooccus.dd2e* and *Corynebacterium.cb50*); Red=Traj.B (Persistent *Moraxella.d253*); Purple = Traj.C (Early *Staphylococcus.29eb*), with 2 variants for CAS – dark purple=Traj.C1 (Very early, in first 2mths), light purple=Traj.C2 (Up to age 1); Dark Orange=Traj.D (Early *Streptococcus.3575*).







**FIGURE C.11: Complete version of Figure 4.6: Average relative abundances of ASVs per healthy and illness-associated samples, per individual in each microbiome trajectory as determined by MFA/*k*-means; in (A) CAS and (B) COAST (QIIME2).**

ASV = Amplicon sequence variant; LRI = Lower respiratory illness; URI = Upper respiratory illness. Note that the original trajectories were based on dimension reduction and clustering of healthy routine (“Well”) samples from the first two years of life, in either cohort. Despite this, we can observe the patterns persisting into illness samples.

**TABLE C.1: The top twenty common OTUs in CAS QIIME1 dataset, and their analogous ASVs in CAS QIIME2 dataset that have matching sequences**

Abbr. = Abbreviated. ASV = Amplicion sequence variant; OTU = Operational taxonomic unit; Qi1 = QIIME1; Qi2 = QIIME2; rel. abund. = relative abundance. QIIME1 OTUs were assigned a unique identifier by closed-reference clustering using the Greengenes 99% reference (v13\_05); OTUs with the same V4 16S region were merged. QIIME2 ASVs were assigned after dereplication, denoising and merging of paired reads using DADA2. QIIME2 ASVs were also annotated with taxa using a naive Bayes classifier trained on the Greengenes 99% reference (v13\_08); these taxa were virtually identical to the analogous taxa from QIIME1 closed-reference OTU-picking (first column). Only the 20 most “common” OTUs in CAS QIIME1 pipeline are listed. ASVs that were also “common” in CAS QIIME2 are highlighted in bold. “Common” was defined for QIIME1 and QIIME2 separately, as described in the main text: mean relative abundance >0.1% across all samples; present in >20% samples; and dominating (>50%) at least one sample.

Qi1 Annotated taxon	Qi1 OTU	Qi2 ASV (abbr.)	Qi1 mean rel. abund.	Qi2 mean rel. abund.	Qi2 representative FASTA sequence (~230bp) for each ASV, V4 16S region	Common in CAS QI2?
<i>Allobacoccus</i>	OTU886735	d42e15140d20c8471 2456e4c12ece08b	0.096	0.099	ACAAGCGTTTCCGGATTTATTGGCGTAAAGGGAGCCGCGGCTGTGTTTATGTCTAATGTGAAGCCCGGGCTTAAC CGTGAACGGCATTTGAAACTGACAGACTGATGTAGAAAGAGAAATTCGAATTCGAAAGTGTAGCGGTGGAATCGGTAG ATAATTTGGAGAACGACGAGTGGCGAAGGCGATTCTTGTCTAACATTGACGCTGAGGCTCGAAAGGCTGGGG	Yes
		40eece21cd2ae3f632 7e91c95be05a4ad	0.096	8.60E-10	CCAGCAGCCCGGTTAATAGCTAGGTGACAGCGCTTCCGGATTTATTGGCGCTAAGGGAGCGGCGCTCTCTTTAG TCTMATGTGAAGCCACCGGCTTAACCGTGAACCGGATTTGAACTGACAGACTTGAATGATGTAGAGAGGAAATGGAAAT TCCAACTGTAGCGGTGGATCGCTAGATATTGGAGGAAACCAAGTGGCGAAGCGGATTTCTGTCTAACATTGACGCT GAGCTTCGAAAGCGCTGGGG	No
		48459e03615e19440 e483b108e685d10	0.096	1.80E-09	ACAAGCGTTTCCGGATTTATTGGCGTAAAGGGAGCCGCGGCTGTGTTTATGTCTAATGTGAAGCCCGGGCTTAAC CGTGAACGGCATTTGAAACTGACAGACTGATGTAGAAAGAGAAATTCGAATTCGAAAGTGTAGCGGTGGAATCGGTAG ATAT	No
<i>Corynebacterium</i>	OTU1049188	ca518512eb1662b21 963d6d32ed5b8c9	0.0018	0.0016	GCAMCGTTTCCGGATTTACTGGCGTAAAGCGCTAGCTAGTGGTGTGTCGCGCTCTCTGAAATTCACAGCTTAAC TGTGGCGTCCAGCGGATACGGGCTATACCTTGAGTACTGTAGGGTAAGTGGAAATTCCTGTGTAGCGGTGAATGGCGCA GATATCAGGAGAACACCGATGGCGAAGCGAGGTTACTGGCGAGTTACTGACGCTGAGGAGGCGAAGCGATGGGT	No
		b1b8738ea679c1266 fa0369c13932a61	0.0086	0.0095	CGAGCGTTTCCGGATTTACTGGCGTAAAGGGCTCGTAGGCTGTGTTGGCGCTCTCTGTGAATTCGGGGCTTAAC TCCGGCTCCAGCGGATACGGGCTAACCTGACTGTAGGGTAACTGGAATTCCTGTGTAGCGGTGGAATTCGGCGA GATATCAGGAGAACACCGATGGCGAAGCGAGGTTACTGGCGAGTTACTGACGCTGAGGAGGCGAAGCGATGGGT	No
<i>Corynebacterium</i>	OTU4474764	cb50e79d177c169ec 5b9d5e7d308874e	0.092	0.076	GCAMCGTTTCCGGATTTACTGGCGTAAAGCGCTCGTAGGCTGTGTCGCGCTCTCTGAAATTCACAGCTTAAC TGTGGCGTCCAGCGGATACGGGCTGACTTGAAGTACTGTAGGGTAAGTGGAAATTCCTGTGTAGCGGTGAATGGCGAG ATACTAGGAGAACACCGATGGCGAAGCGGCTTACTGGCGAGTTACTGACGCTGAGGAGGCGAAGCGATGGGT	Yes
		fc6b26865931614bd 16c05097112b7785	0.092	6.20E-09	CGAAGCGTTTCCGGATTTACTGGCGTAAAGGGCTCGTAGGCTGTGTTGGCGCTCTCTGTGAATTCACAGCTTAAC TGTGGCGTCCAGCGGATACGGGCTGACTTGAAGTACTGTAGGGTAAGTGGAAATTCCTGTGTAGCGGTGGAATTCGGCGAG ATACTAGGAGAACACCGATGGCGAAGCGGCTTACTGGCGAGTTACTGACGCTGAGGAGGCGAAGCGATGGGT AGGATTAGATACCC	No
<i>Haemophilus</i>	OTU240051	f579dd881c6a3481 2eae33d72110f0d6	0.049	0.057	CGAAGCGTTAATCGGAAATACTGGCGTAAAGGGCACGCGCGGTTATTAACTGAGCTGTGAAGCCCGCGGCTTAAC CTGGGAATTCGATTTGACATCGGTAAGTACTGTAGGGTAAGTGGAAATTCCTGTGTAGCGGTGGAATTCGGCGAG AGA TGTGGAAATACCGAAGCGGAGCGGCTTGGGAATGACTGACGCTCA TGTGGAAAGCGCTGGGG	Yes
<i>Haemophilus</i>	OTU4469627	1648c7f18abada521 7ae40437b7b60d5	0.0038	0.0048	CGGAGCGTTAATCGGAAATACTGGCGTAAAGGGCACGCGCGGTTATTAACTGAGGCTGTGAAGCCCGCGGCTTAAC CTGGGAATTCGATTTGACATCGGTAAGTACTGTAGGGTAAGTGGAAATTCCTGTGTAGCGGTGGAATTCGGCGAG AGA TGTGGAAATACCGAAGCGGAGCGGCTTGGGAATGACTGACGCTCA TGTGGAAAGCGCTGGGG	No
<i>Haemophilus</i>	OTU956702	bc04904a0469405c6 5b2709988488ee91	0.024	0.028	CGAAGCGTTAATCGGAAATACTGGCGTAAAGGGCACGCGCGGTTATTAACTGAGGCTGTGAAGCCCGCGGCTTAAC CTAGGAATTCGATTTGACATCGGTAAGTACTGTAGGGTAAGTGGAAATTCCTGTGTAGCGGTGGAATTCGGCGAG AGA TGTGGAAATACCGAAGCGGAGCGGCTTGGGAATGACTGACGCTCA TGTGGAAAGCGCTGGGG	Yes

Continued on next page



Continued from previous page

Q11 Annotated taxon	Q11 OTU	Q12 ASV	Q12 ASV (abbr.)	Q11 mean rel. abun.	Q12 mean rel. abun.	Q12 representative FASTA sequence (~230bp) for each ASV, V4 16S region	Common in CAS Q12?
<i>Streptococcus</i>	OTU4455633	15bf37a8413b8764bad9392e50bea0a	15bf	0.085	7.60E-09	CCGAGGCTTCCGGATTTATTGGCGTAAAGCGAGCCACGGCGGTTAGATAAGTCTGAAGTTAAAGGCTGTGGCTTAAC CATAGTAGGCTTTGGAAACTGTTAACTTGAGTCCACAGAGGGAGAGTGAATTCCATGTTAGCGGTGAATAATGCGTAGA TATATGGAGGAACACCGGTTGGCGAAAGCGGCTCTCTGGCTTCTAAGTGAAGCTTGAAGCTTGAAGTTAAAGGCAATTGGCTCAAC GGATTAGATACC	No
		c79e5cab23469318738bcffdea303b4	c79e	0.0014	0.0015	CCGAGGCTTCCGGATTTATTGGCGTAAAGCGAGCCACGGCGGTTTAAAGTCTGAAGTTAAAGGCAATTGGCTCAAC CATAGTAGGCTTTGGAAACTGAGAACTTGAGTCCACAGAGGGAGAGTGAATTCCATGTTAGCGGTGAATAATGCGTAGA TATATGGAGGAACACCGGTTGGCGAAAGCGGCTCTCTGGCTTCTAAGTGAAGCTTGAAGCTTGAAGTTAAAGGCAATTGGCTCAAC	No
<i>Streptococcus</i>	OTU509773	a3a3b519e04bf11edbb1482497c2b34c	a3a3	0.006	0.0075	CCGAGGCTTCCGGATTTATTGGCGTAAAGCGAGCCACGGCGGTTTGAATAAGTCTGAAGTTAAAGGCTGTGGCTCAAC CATAGTAGGCTTTGGAAACTGCAAACTTGAGTCCACAGAGGGAGAGTGAATTCCATGTTAGCGGTGAATAATGCGTAGA TATATGGAGGAACACCGGTTGGCGAAAGCGGCTCTCTGGCTTCTAAGTGAAGCTTGAAGCTTGAAGTTAAAGGCAATTGGCTCAAC	Yes
		b069c8413c9eb478d4faa117c4372bd3	b069	0.0034	0.0031	CCGAGGCTTATCCGGATTTATTGGCGTAAAGCGAGCCACGGCGGTTTGAATAAGTCTGAAGTTAAAGGCTGTGGCTTAAC CATAGTAGGCTTTGGAAACTGTTAACTTGAGTCCACAGAGGGAGAGTGAATTCCATGTTAGCGGTGAATAATGCGTAGA TATATGGAGGAACACCGGTTGGCGAAAGCGGCTCTCTGGCTTCTAAGTGAAGCTTGAAGCTTGAAGTTAAAGGCAATTGGCTCAAC	No



**TABLE C.2: The eighteen ASVs common to either CAS or COAST datasets (QIIME2)**

ASV = Amplicon sequence variant; rel. abund. = relative abundance. A “common” ASV was defined as having: mean relative abundance >0.1% across all samples; present in >20% samples; and dominating (>50%) at least one sample; in either CAS or COAST. NCBI BLAST performed September 2018.

QIIME2 ASV	Common in CAS?	Common in COAST?	CAS rel. abund.	COAST rel. abund.	Representative FASTA sequence, V4 16S region	BLAST best guess from NCBI reference	E-value of guess	% Identity of FASTA with NCBI reference
<i>Moraxella</i> .d253	YES	YES	0.31	0.25	GCAAGCGTTTAAATCGGAATTAAGTGGGCTAAGCCGCGCTAGTGGTGTATTTAAAGTCAGATGTGAAGCCGCGGCTTAAAC CTGGAACTCGATCTGTAATACTAGAGTAGGAGAGGGAGAGTGAATCCCAAGTCCAGGTTGCTAGCGGTGAATGGCGTAG AGATCTGGAGGAATACCGAATGGCGAAGCGACGCTCCGCTGCATCACTACACAGTGGAGTGGGAAGCGTGGGT	6E-118	100%	
<i>Streptococcus</i> .4060	YES	YES	0.1	0.14	CCGAGCGTTTGTCCGGAATTAATGGCGCTAAGCGAAGCCGAGCGGTTAGATMAGTCTGAAATTAAGGCTGTGGCTTAAAC CATAGTAGGCTTTGGAAATCTGTTTAACTGTAGTGCAGAGGGAGAGTGGAAATCCCATGTGTAGCGGTGAATGGGTAGA TATAATGGAGGAACACCGGTGGCAAGCGGCTCTCTGGCTTGTAACTGACGCTGAGAGCTCGAAGCGGTGGGG	2E-117	100%	
<i>Allobacoccus</i> .d42e	YES	YES	0.099	0.13	CAAGCGTTTGTCCGGAATTAATGGGCGTAAAGCGAGCGCAAGCGGCTGTGTTTACTCTAAATGTGAAGCCGCGGCTTAAAC GTGGAAAGCGCATTTGGAATCTGACAGACTTGAATGTAGAGAGAAATGGAATCCCAAGTGTAGCGGTGAATGGCTAGA TATTTGGAGGAACACCGAGTGGCAAGCGAATTTCTGGCTTAACATTTGACCTGAGGCTCGAAGCGTGGGG	1E-89	93%	
<i>Corynebacterium</i> .cb50	YES	YES	0.076	0.078	GCAAGCGTTTGTCCGGAATTAAGTGGGCTAAGCGGCTGCTAGTGGTGTGTGCGGCTGTGTGAAATTCGACAGCTTAAAC TGTGGGCTGCAGCGGCTGACTTGTAGTACTGTAGGTTAACTGGAAATTCCTGTGTAGCGGTGAATGGCGGAG ATATCAGGAGGAACACCGAATGGCGAAGCGGCTTCTGGGCAAGTACTGAGCGCTGAGGAGCGAAGCAATGGGT	6E-118	100%	
<i>Staphylococcus</i> .29eb	YES	YES	0.074	0.02	GCAAGCGTTTGTCCGGAATTAATGGGCTAAGCGGCTGAGCGGTTTTTAAAGTCTGATGTGAAGCCGCGGCTCAAC CGTGAAGGCTAATGGAATCTGAAACTGCAAACTGAGTGCAGAGAGAAAGTGAATCCCATGTGTAGCGGTGAATGGCGAG AGATTTGGAGGAACACCGAGTGGCAAGCGGCTTCTGGCTGTACTGAGCGGCTGATGTGGGAAGCGGTGGGG	6E-118 all	100% all	
<i>Haemophilus</i> .4579	YES	YES	0.057	0.071	GGAGCGTTTAAATCGGAATTAAGTGGGCTAAGGGCAGCGCGGTTTTTAAAGTGTAGTGTGAAGCCGCGGCTTAAAC CTGGAAATTCGATTTCCAGACTGGGTAACCTAGAGTACTGTAAGGAGGAGTGAATCCCAAGTCCAGGTAATGGCGTAG AGATCTGGAGGAATACCGAAGGGGGAAGCGCCCTTGGGAATGTACTGAGCGCTGATGTGGGAAGCGTGGGG	1E-114	99%	
<i>Moraxellaceae</i> .6028	YES	YES	0.034	0.025	GCAAGCGTTTAAATCGGAATTAAGTGGGCTAAGCGGAGCGTGTGGTGTATTTAAAGTCAGATGTGAATCCCTGGGCTTAAAC CTAGAACTGCATCTGATACTGATAAACTAGAGTAGGTAAGGAGGAGTGAATCCCAAGTCCAGGTTAGCGGTGAATGGCGTAG AGATCTGGAGGAATACCGAATGGCGAAGCGGCTTCTGGCTGCATCACTGAGCGGCTGCAAGCGTGGGT	1E-114	99%	
<i>Streptococcus</i> .3575	YES	YES	0.028	0.052	CCGAGCGTTTGTCCGGAATTAATGGCGCTAAGCGAAGCCGAGCGGTTAGATMAGTCTGAAATTAAGGCTGTGGCTTAAAC CATAGTAGGCTTTGGAAATCTGTTTAACTGTAGTGCAGAGGGAGAGTGGAAATCCCATGTGTAGCGGTGAATGGGTAGA TATAATGGAGGAACACCGGTGGCAAGCGGCTCTCTGGCTTGTAACTGACGCTGAGGCTCGAAGCGTGGGG	7E-113	99%	
<i>Haemophilus</i> .bc0d	YES	YES	0.028	0.032	GCGAGCGTTTAAATCGGAATTAAGTGGGCTAAGGGCAGCGCGGTTTTTAAAGTGTAGTGTGAAGCCGCTGGGCTTAAAC CTAGAAATTCGACTGGGTAACCTAGAGTACTTAAAGGAGGAGTGAATCCCAAGTCCAGGTTAGCGGTGAATGGCGTAG AGATCTGGAGGAATACCGAAGGGCAAGCGCCCTTGGGAATGTACTGAGCGCTGATGTGGGAAGCGTGGGG	6E-118	100%	

Continued on next page

Continued from previous page

QIIME2 ASV	Common in CAS?	Common in COAST?	CAS rel. abund.	COAST rel. abund.	Representative FASTA sequence, V4 16S region	BLAST best guess from NCBI reference	E-value of guess	% Identity of FASTA with NCBI reference
<i>Streptococcus.a3a3</i>	YES	YES	0.0075	0.016	CCGAGCGTTGTCCGGATTTATTGGGCGTAAAGCGAGCGAGCGGGTTGTAAGTCTGAAAGTAAAGCGTGTGGCTCAAC CATAGTTCCGCTTTGGAAACTGTCAAACTGTAGTCGACGAGGGAGAGTGGAAATTCATGTGTAGCGGTGAAATGCCTAGA TATATGGAGGAAACCCCGTGGCGAAGCGGCTCTCTGGTCTGTAACTGACGCTGAGGCTCGAAAGCGTGGGG	2E-117	100%	
<i>Pseudomonas.0925</i>	YES	NO	0.0061	0.00026	GCAAGCGTTAATCGGAAATTAATCGGCGTAAAGCGGCGTGTAGTGGTGGTTTAAAGTGTGGATGTGAATCCCGGGGCTCAAC CTGGAACTCGCTTAAAGAACTGTGACTAGTAGATGTGTAGAGGGTGGTGGAAATTCCTGTGTAGCGGTGAAATGCGGTAG ATATAGGAGGAAACACCGTGGCGAAGCGGCTTAATCTAGCTGAAATTCGAGGTGTAGCGGTGAAATGCGGTAG	6E-118	100%	
<i>Moraxellaceae.a5a0</i>	NO	YES	0.0057	0.014	GGAGCGTTAATCGGAAATTAATCGGCGTAAAGCGGCGTGTAGTGGTGGTTTAAAGTCAAGTGTGAATCCCTGGGCTTAAAC CTAGGAACTGCATCTGATTAATAACTAGTAGTGTAGAGGAAAGTGAATTCGAGGTGTAGCGGTGAAATGCGGTAG AGATCTGGAGGAATACCGATGGCGAAGCGAGCTTCTGGCATCATACTGACACTGAGGTTGCGAAGCGTGGT	3E-116	99%	
<i>Gemellaceae.d800</i>	YES	YES	0.0051	0.0048	GCAAGCGTTGTCCGGAAATTAATCGGCGTAAAGCGGCGGAGTGGTTTAAAGTCTGTATGTGAAGCGGCGCTCAAC CGTAGGGTCAATGGAAACTGTAAACTTTGAGTCGAGAGAAAGTGAATTCCTAGTGTAGCGGTGAAATGCGGTAG AGATTAGGAGGAAACACCGTGGCGAAGCGGCTTCTGGCCTGTAACTGACACTGAGCGGCGGCGAAGCGTGGGG	6E-118	100%	
<i>Veillonella.tb81</i>	NO	YES	0.0046	0.017	GCAAGCGTTGTCCGGAAATTAATCGGCGTAAAGCGGCGGAGTGGTTCAGTCTGTCTTAAAGTTCGGGGCTTAAAC CCCGTAGTGGATGGAACTGCCAATCTAGATATCGGAGGAAAGTGGAAATTCCTAGTGTAGCGGTGAAATGCGGTAGA TATTAGGAGGAAACACCGTGGCGAAGCGGACTTCTGGAGGAAACTGACGCTGAGGCGGCGAAGCGCGGGGG	1E-110	98%	
<i>Escherichia.d2a4</i>	YES	YES	0.0035	0.0087	GCAAGCGTTAATCGGAAATTAATCGGCGTAAAGCGGCGGAGTGGTGGTTTAAAGTCAAGTGTGAATCCCGGGGCTCAAC CTGGAACTGCATCTGTACTGCGAAGCTTGTAGTCTGTGTAGAGGGGTTAGATTCGCGGGTGTAGCGGTGAAATGCGGTAG AGATCTGGAGGAAATACCGGTGGGAGGCGGCGGCTGTGGAGGAACTGACGCTGAGGCTGCGAAGCGTGGGG	6E-118 all	100% all	
<i>Streptococcus.b069</i>	NO	YES	0.0031	0.02	CCGAGCGTTATCCGGATTTATTGGGCGTAAAGCGGCGGAGCGGTTAGATAAGTCTGAAAGTAAAGCGTGTGGCTTAAAC CATAGTCCGCTTTGGAACTGTAACTGTAGTCGAAAGGGGAGAGTGGAAATCCATGTGTAGCGGTGAAATGCGGTAGA TATATGGAGGAAACCCCGTGGCGAAGCGGCTCTCTGGCTTGTAACTGACGCTGAGGCTCGAAGCGTGGGG	1E-115	99%	
<i>Streptococcus.belb</i>	NO	YES	0.0019	0.0065	CCGAGCGTTATCCGGATTTATTGGGCGTAAAGCGGCGGAGCGGTTAGATAAGTCTGAAAGTAAAGCGTGTGGCTTAAAC CATAGTCCGCTTTGGAACTGTAACTGTAGTCGAAAGGGGAGAGTGGAAATCCATGTGTAGCGGTGAAATGCGGTAGA TATATGGAGGAAACCCCGTGGCGAAGCGGCTCTCTGGCTGTAACTGACGCTGAGGCTCGAAGCGTGGGG	5E-114	99%	
<i>Neisseriaceae.03f4</i>	NO	YES	0.0015	0.0023	GCAAGCGTTAATCGGAAATTAATCGGCGTAAAGCGGCGGAGCGGTTACTTAAAGCGAGATGTGAATCCCGCAAGCTTAAAC TTGGGACCTGCATTTGGAACTGTAGAGTGTGTAGAGGAGGAGTGGAAATTCGACATGTAGCGGTGAAATGCGGTAG AGATGTGGAGGAATACCGATGGCGAAGCGGCGCTCTGGGATAACACTGAGTGTAGCGGTGCGAAGCGTGGGG	2E-92	94%	
					<i>Neisseria mucosa C102</i>	1E-85	92%	

**TABLE C.3: Results of GEE models associating common ASVs with respiratory illness status (well vs. unwell), with adjustments for child as subjects factor, and gender, age and season as covariates.**

ASV = Amplicon sequence variant; 95% CI = 95% Confidence interval; OR = odds ratio. The model for analysis was a generalized estimating equation (GEE), of: respiratory illness status (well vs. unwell) ~ ASV + gender + age + season | subject. Separate models were created for each MPG (i.e. MPG of interest vs. all others). The table is sorted by descending odds ratio in CAS, and statistically-significant associations are bolded.

ASV (with full identifier)	CAS		COAST	
	OR (95% CI)	p-value	OR (95% CI)	p-value
Haemophilus.f579ddd8d7c6a3d812ca33d72110f0d6	<b>1.6 (1.5-1.7)</b>	<b>7.50E-39</b>	<b>1.1 (1-1.1)</b>	<b>0.045</b>
Streptococcus.4060107ffdf52e44f9d72343b7332609	<b>1.5 (1.4-1.5)</b>	<b>2.60E-45</b>	<b>1.3 (1.2-1.3)</b>	<b>5.30E-25</b>
Haemophilus.bc0d904a0469d05c65b270998488ee91	<b>1.4 (1.3-1.6)</b>	<b>1.10E-15</b>	<b>1.3 (1.2-1.4)</b>	<b>2.60E-13</b>
Moraxella.d253ca966efcd811d057dc2aa6430774	<b>1.4 (1.3-1.5)</b>	<b>9.10E-22</b>	<b>1.4 (1.3-1.5)</b>	<b>1.00E-24</b>
Escherichia.d2a4add6029e32135562e2da9d4ed2e1	<b>1.2 (1.1-1.3)</b>	<b>1.40E-05</b>	<b>0.78 (0.73-0.84)</b>	<b>1.90E-11</b>
Haemophilus.rare	<b>1.2 (1.1-1.2)</b>	<b>6.90E-06</b>	<b>0.85 (0.8-0.91)</b>	<b>6.50E-07</b>
Moraxella.rare	1.1 (0.99-1.2)	0.085	1.1 (0.99-1.1)	0.1
Pseudomonas.092523a4aed84953633a5f48bc2f87b9	1 (0.96-1.1)	0.37	<b>0.73 (0.64-0.82)</b>	<b>5.90E-07</b>
Moraxellaceae.a5a08d96bf11fdeed28fc530d7f5d8c7	1 (0.9-1.2)	0.64	1 (0.96-1.1)	0.33
Streptococcus.b069c8413c9eb478d4faa117c4372bd3	0.94 (0.87-1)	0.12	<b>0.82 (0.77-0.87)</b>	<b>3.00E-10</b>
Streptococcus.rare	0.93 (0.87-1)	0.057	0.91 (0.85-0.97)	0.0074
Alloiococcus.rare	<b>0.89 (0.82-0.97)</b>	<b>0.0074</b>	<b>0.88 (0.81-0.95)</b>	<b>0.00072</b>
Streptococcus.belbac8c38e726214223a5ac13184ab7	<b>0.85 (0.78-0.92)</b>	<b>0.00011</b>	<b>0.72 (0.67-0.77)</b>	<b>5.50E-21</b>
Gemellaceae.d8009096d67cd7b4ce2853718bb17013	<b>0.83 (0.77-0.89)</b>	<b>9.30E-07</b>	<b>0.66 (0.61-0.72)</b>	<b>3.90E-22</b>
Veillonella.fb81574d887fa68041a84122487a0e47	<b>0.82 (0.76-0.88)</b>	<b>7.10E-08</b>	<b>0.7 (0.65-0.76)</b>	<b>1.30E-20</b>
Moraxellaceae.6028872a168fbd6bc3381494da3b116f	<b>0.8 (0.75-0.85)</b>	<b>4.50E-12</b>	0.99 (0.93-1.1)	0.88
Streptococcus.357551f644064bb2b49614b1cab2fa5	<b>0.78 (0.74-0.83)</b>	<b>1.20E-14</b>	<b>0.61 (0.57-0.66)</b>	<b>6.80E-36</b>
Neisseriaceae.03f429f55feb0e805ba48c8c00867cf3	<b>0.78 (0.7-0.87)</b>	<b>5.40E-06</b>	0.92 (0.84-1)	0.064
Moraxellaceae.rare	<b>0.76 (0.7-0.83)</b>	<b>1.40E-10</b>	<b>0.78 (0.72-0.84)</b>	<b>1.30E-11</b>
Staphylococcus.rare	<b>0.76 (0.7-0.82)</b>	<b>3.70E-11</b>	<b>0.7 (0.62-0.8)</b>	<b>9.10E-08</b>
Corynebacterium.cb50e79d177c169ec5b9d5e7d308874e	<b>0.75 (0.71-0.8)</b>	<b>4.80E-19</b>	<b>0.78 (0.74-0.83)</b>	<b>1.20E-15</b>
Streptococcus.a3a3b519e04bf11edbb1d82497c2b34c	<b>0.75 (0.7-0.81)</b>	<b>5.60E-15</b>	<b>0.71 (0.66-0.76)</b>	<b>2.00E-22</b>
Corynebacterium.rare	<b>0.71 (0.67-0.76)</b>	<b>6.80E-25</b>	<b>0.68 (0.63-0.73)</b>	<b>1.40E-28</b>
others.rare	<b>0.68 (0.62-0.74)</b>	<b>1.30E-17</b>	<b>0.63 (0.57-0.69)</b>	<b>3.10E-23</b>
Alloiococcus.dd2e15140d20c84712456e4c12ece08b	<b>0.68 (0.63-0.73)</b>	<b>9.20E-23</b>	<b>0.73 (0.68-0.79)</b>	<b>1.20E-16</b>
Staphylococcus.29eb8e14ce61e9c4549de0f9cf200c19	<b>0.61 (0.57-0.66)</b>	<b>1.70E-41</b>	<b>0.61 (0.56-0.66)</b>	<b>3.30E-29</b>

**TABLE C.4: Results of GEE models associating MPGs and ASVs with winter season, with adjustments for child as subjects factor, and gender, age, season +/- respiratory illness as covariates.**

95% CI = 95% Confidence interval; ASV = Amplicon sequence variant; MPG = Microbiome profile group; OR = odds ratio. The model for analysis was a generalized estimating equation (GEE). Each sub-table presents a different modelling scheme: **(A)** winter season (winter vs. other seasons) ~ MPG + gender + age | subject; **(B)** winter season ~ MPG + gender + age + respiratory illness | subject; **(C)** winter season ~ ASV + gender + age | subject; **(D)** winter season ~ ASV + gender + age + respiratory illness | subject. Separate models were created for each MPG or ASV (i.e. MPG or ASV of interest vs. all others). Statistically-significant associations are bolded.

MPG or ASV	CAS		COAST		Meta	
	OR (95% CI)	p-value	OR (95% CI)	p-value	OR (95% CI)	p-value
<b>A. (Winter ~ MPG, without respiratory illness covariate)</b>						
Alloiococcus.dd2e MPG	<b>0.65</b> <b>(0.49-0.87)</b>	<b>0.0041</b>	<b>0.48</b> <b>(0.36-0.63)</b>	<b>1.80E-07</b>	<b>0.56</b> <b>(0.41-0.76)</b>	<b>0.00021</b>
Corynebacterium.cb50 MPG	1.1 (0.71-1.7)	0.66	1.1 (0.77-1.7)	0.5	1.1 (0.84-1.5)	0.43
Corynebacterium.rare MPG	1.1 (0.52-2.2)	0.85	1.7 (0.62-4.8)	0.3	1.3 (0.69-2.3)	0.45
Haemophilus.bc0d MPG	1.1 (0.72-1.8)	0.59	1.3 (0.8-2.1)	0.29	1.2 (0.87-1.7)	0.26
Haemophilus.f579 MPG	0.97 (0.7-1.4)	0.87	<b>1.6 (1.2-2.1)</b>	<b>0.0014</b>	1.2 (0.78-2)	0.36
Haemophilus.rare MPG	1.4 (0.59-3.2)	0.47	1 (0.41-2.4)	0.99	1.2 (0.64-2.2)	0.6
Moraxella.d253 MPG	<b>1.2 (1-1.4)</b>	<b>0.023</b>	<b>1.7 (1.4-2)</b>	<b>2.40E-09</b>	<b>1.4 (1-2)</b>	<b>0.032</b>
Moraxella.rare MPG	0.92 (0.52-1.6)	0.76	1.4 (0.78-2.4)	0.28	1.1 (0.75-1.7)	0.59
Moraxellaceae.6028 MPG	0.83 (0.55-1.2)	0.35	<b>1.9 (1.2-2.9)</b>	<b>0.005</b>	1.2 (0.56-2.7)	0.6
Moraxellaceae.a5a0 MPG	1.9 (0.55-6.3)	0.32	1 (0.53-1.9)	0.99	1.1 (0.65-2)	0.66
others.rare MPG	0.91 (0.71-1.2)	0.49	<b>0.25</b> <b>(0.18-0.34)</b>	<b>6.10E-17</b>	0.48 (0.13-1.7)	0.26
Staphylococcus.29eb MPG	<b>0.67</b> <b>(0.49-0.91)</b>	<b>0.011</b>	0.92 (0.52-1.6)	0.76	<b>0.72</b> <b>(0.55-0.95)</b>	<b>0.018</b>
Streptococcus.4060 MPG	<b>1.3 (1-1.6)</b>	<b>0.032</b>	1.1 (0.88-1.4)	0.41	<b>1.2 (1-1.4)</b>	<b>0.037</b>
Streptococcus.rare MPG	0.43 (0.054-3.4)	0.42	NA	NA	0.43 (0.054-3.4)	0.42
<b>B. (Winter ~ MPG, with respiratory illness covariate)</b>						
Alloiococcus.dd2e MPG	<b>0.74</b> <b>(0.55-0.99)</b>	<b>0.046</b>	<b>0.55</b> <b>(0.41-0.73)</b>	<b>3.40E-05</b>	<b>0.64</b> <b>(0.47-0.85)</b>	<b>0.0025</b>
Corynebacterium.cb50 MPG	1.3 (0.81-2)	0.29	1.3 (0.85-1.9)	0.24	1.3 (0.94-1.7)	0.11
Corynebacterium.rare MPG	1.2 (0.55-2.5)	0.68	2.3 (0.83-6.2)	0.11	1.5 (0.8-2.8)	0.21
Haemophilus.bc0d MPG	1 (0.64-1.6)	0.95	1.2 (0.72-1.9)	0.54	1.1 (0.78-1.5)	0.64
Haemophilus.f579 MPG	0.85 (0.61-1.2)	0.35	<b>1.4 (1.1-1.9)</b>	<b>0.012</b>	1.1 (0.67-1.8)	0.67
Haemophilus.rare MPG	1.3 (0.54-2.9)	0.59	0.9 (0.37-2.2)	0.82	1.1 (0.58-2)	0.81
Moraxella.d253 MPG	1.2 (0.98-1.4)	0.088	<b>1.6 (1.4-1.9)</b>	<b>1.00E-07</b>	1.4 (0.99-1.9)	0.06
Moraxella.rare MPG	1 (0.59-1.8)	0.92	1.4 (0.8-2.6)	0.22	1.2 (0.81-1.8)	0.37
Moraxellaceae.6028 MPG	0.93 (0.62-1.4)	0.73	<b>1.9 (1.2-3)</b>	<b>0.0037</b>	1.3 (0.66-2.7)	0.43
Moraxellaceae.a5a0 MPG	1.9 (0.58-6.5)	0.28	1 (0.55-1.9)	0.93	1.2 (0.68-2)	0.57
others.rare MPG	0.92 (0.72-1.2)	0.53	<b>0.27</b> <b>(0.19-0.38)</b>	<b>6.50E-15</b>	0.5 (0.15-1.7)	0.26
Staphylococcus.29eb MPG	0.78 (0.57-1.1)	0.13	1.1 (0.63-2)	0.68	0.87 (0.63-1.2)	0.38
Streptococcus.4060 MPG	1.1 (0.9-1.5)	0.28	0.95 (0.75-1.2)	0.67	1 (0.87-1.2)	0.67
Streptococcus.rare MPG	0.4 (0.049-3.3)	0.39	NA	NA	0.4 (0.049-3.3)	0.39
<b>C. (Winter ~ ASV, without respiratory illness covariate)</b>						
Alloiococcus.dd2e15140d	<b>0.92</b> <b>(0.86-0.98)</b>	<b>0.011</b>	<b>0.86</b> <b>(0.79-0.92)</b>	<b>5.10E-05</b>	<b>0.89</b> <b>(0.83-0.96)</b>	<b>0.0014</b>
20c84712456e4c12ece08b	1 (0.96-1.1)	0.31	0.97 (0.89-1.1)	0.48	1 (0.94-1.1)	0.85
Alloiococcus.rare	1 (0.99-1.1)	0.15	0.96 (0.91-1)	0.12	1 (0.92-1.1)	0.96
Corynebacterium.cb50e79d177c169ec5b9d5e7d308874e	0.96 (0.9-1)	0.13	<b>0.89</b> <b>(0.84-0.95)</b>	<b>9.00E-04</b>	<b>0.93</b> <b>(0.87-0.99)</b>	<b>0.022</b>
Escherichia.d2a4add6029e32135562e2da9d4ed2e1	<b>0.88</b> <b>(0.82-0.95)</b>	<b>0.00094</b>	0.97 (0.9-1)	0.43	0.93 (0.84-1)	0.11

Continued on next page

Continued from previous page

MPG or ASV	CAS		COAST		Meta	
	OR (95% CI)	p-value	OR (95% CI)	p-value	OR (95% CI)	p-value
Gemellaceae.d8009096d67cd7b4ce2853718bb17013	<b>0.89</b> (0.83-0.95)	<b>0.001</b>	<b>0.72</b> (0.66-0.78)	<b>2.20E-14</b>	<b>0.8 (0.65-0.99)</b>	<b>0.04</b>
Haemophilus.bc0d904a0469d05c65b270998488ee91	0.98 (0.93-1)	0.58	1 (0.97-1.1)	0.31	1 (0.96-1.1)	0.77
Haemophilus.f579ddd8d7c6a3d812ca33d72110f0d6	1 (0.97-1.1)	0.45	1 (0.94-1.1)	0.87	1 (0.97-1.1)	0.49
Haemophilus.rare	<b>0.94</b> (0.89-0.99)	<b>0.03</b>	<b>0.9 (0.85-0.96)</b>	<b>0.0017</b>	<b>0.92</b> (0.88-0.96)	<b>2.00E-04</b>
Moraxella.d253ca966efcd811d057dc2aa6430774	<b>1.2 (1.1-1.3)</b>	<b>7.00E-07</b>	<b>1.4 (1.3-1.5)</b>	<b>5.20E-15</b>	<b>1.3 (1.1-1.5)</b>	<b>0.00053</b>
Moraxella.rare	<b>1.2 (1.1-1.3)</b>	<b>6.90E-07</b>	1.1 (0.98-1.1)	0.15	1.1 (1-1.2)	0.06
Moraxellaceae.6028872a168fbd6bc3381494da3b116f	0.96 (0.91-1)	0.23	1 (0.97-1.1)	0.36	1 (0.93-1.1)	0.9
Moraxellaceae.a5a08d96bf11fdeed28fc530d7f5d8c7	0.98 (0.88-1.1)	0.74	1.1 (0.99-1.1)	0.09	1 (0.96-1.1)	0.36
Moraxellaceae.rare	<b>0.91</b> (0.85-0.99)	<b>0.02</b>	1.1 (1-1.2)	0.056	0.99 (0.85-1.2)	0.91
Neisseriaceae.03f429f55feb0e805ba48c8c00867cf3	1.1 (0.95-1.2)	0.29	0.97 (0.89-1.1)	0.53	1 (0.93-1.1)	0.83
others.rare	<b>0.83 (0.77-0.9)</b>	<b>1.40E-05</b>	<b>0.75</b> (0.68-0.84)	<b>3.90E-07</b>	<b>0.8 (0.72-0.88)</b>	<b>4.00E-06</b>
Pseudomonas.092523a4aed84953633a5f48bc2f87b9	1 (0.94-1.1)	0.87	1 (0.92-1.2)	0.61	1 (0.96-1.1)	0.69
Staphylococcus.29eb8e14ce61e9c4549de0f9cf200c19	<b>0.85 (0.79-0.9)</b>	<b>1.60E-07</b>	<b>0.88</b> (0.82-0.95)	<b>0.0012</b>	<b>0.86 (0.82-0.9)</b>	<b>1.10E-09</b>
Staphylococcus.rare	<b>0.92</b> (0.85-0.99)	<b>0.031</b>	0.98 (0.87-1.1)	0.73	<b>0.94 (0.88-1)</b>	<b>0.048</b>
Streptococcus.357551f644064bb2b49614b1caab2fa5	<b>0.91</b> (0.86-0.96)	<b>0.00081</b>	<b>0.8 (0.74-0.87)</b>	<b>6.10E-08</b>	<b>0.86</b> (0.76-0.97)	<b>0.013</b>
Streptococcus.4060107ffdf52e44f9d72343b7332609	<b>1.1 (1-1.1)</b>	<b>0.0017</b>	<b>1.1 (1-1.1)</b>	<b>0.0024</b>	<b>1.1 (1-1.1)</b>	<b>1.20E-05</b>
Streptococcus.a3a3b519e04bf11eddb1d82497c2b34c	<b>0.84 (0.78-0.9)</b>	<b>2.00E-07</b>	<b>0.77</b> (0.73-0.83)	<b>9.00E-15</b>	<b>0.81</b> (0.74-0.87)	<b>5.70E-08</b>
Streptococcus.b069c8413c9eb478d4faa117c4372bd3	0.93 (0.85-1)	0.079	<b>0.78</b> (0.73-0.83)	<b>7.50E-14</b>	0.85 (0.72-1)	0.058
Streptococcus.belbac8c38e726214223a5ac13184ab7	<b>0.87</b> (0.81-0.94)	<b>0.00072</b>	<b>0.81</b> (0.76-0.87)	<b>1.10E-08</b>	<b>0.84 (0.79-0.9)</b>	<b>7.30E-07</b>
Streptococcus.rare	0.97 (0.91-1)	0.4	<b>0.85</b> (0.79-0.92)	<b>1.70E-05</b>	0.91 (0.8-1)	0.16
Veillonella.fb81574d887fa68041a84122487a0e47	<b>0.84 (0.79-0.9)</b>	<b>1.20E-06</b>	<b>0.77</b> (0.71-0.83)	<b>1.70E-10</b>	<b>0.81</b> (0.73-0.89)	<b>8.40E-06</b>
<b>D. (Winter ~ ASV, with respiratory illness covariate)</b>						
Alloiococcus.dd2e15140d20c84712456e4c12ece08b	0.95 (0.89-1)	0.16	<b>0.88</b> (0.82-0.94)	<b>0.00029</b>	<b>0.92 (0.84-1)</b>	<b>0.039</b>
Alloiococcus.rare	1.1 (0.97-1.1)	0.21	0.97 (0.9-1)	0.45	1 (0.93-1.1)	0.8
Corynebacterium.cb50e79d177c169ec5b9d5e7d308874e	<b>1.1 (1-1.1)</b>	<b>0.011</b>	0.98 (0.93-1)	0.52	1 (0.94-1.1)	0.55
Corynebacterium.rare	0.99 (0.93-1.1)	0.79	<b>0.93</b> (0.87-0.99)	<b>0.035</b>	0.96 (0.9-1)	0.24
Escherichia.d2a4add6029e32135562e2da9d4ed2e1	<b>0.87 (0.8-0.93)</b>	<b>0.00022</b>	1 (0.93-1.1)	0.93	0.93 (0.81-1.1)	0.3
Gemellaceae.d8009096d67cd7b4ce2853718bb17013	<b>0.91</b> (0.85-0.97)	<b>0.0069</b>	<b>0.74 (0.68-0.8)</b>	<b>9.50E-14</b>	0.82 (0.67-1)	0.056
Haemophilus.bc0d904a0469d05c65b270998488ee91	0.96 (0.9-1)	0.15	1 (0.95-1.1)	0.87	0.98 (0.94-1)	0.43
Haemophilus.f579ddd8d7c6a3d812ca33d72110f0d6	0.98 (0.93-1)	0.44	0.99 (0.93-1.1)	0.85	0.99 (0.95-1)	0.48
Haemophilus.rare	<b>0.93</b> (0.87-0.98)	<b>0.0091</b>	<b>0.91</b> (0.86-0.97)	<b>0.002</b>	<b>0.92</b> (0.88-0.96)	<b>5.70E-05</b>
Moraxella.d253ca966efcd811d057dc2aa6430774	<b>1.1 (1.1-1.2)</b>	<b>0.00011</b>	<b>1.3 (1.2-1.4)</b>	<b>1.20E-11</b>	<b>1.2 (1.1-1.4)</b>	<b>0.0024</b>
Moraxella.rare	<b>1.2 (1.1-1.2)</b>	<b>1.80E-06</b>	1 (0.98-1.1)	0.17	1.1 (0.99-1.2)	0.063
Moraxellaceae.6028872a168fbd6bc3381494da3b116f	0.99 (0.93-1)	0.64	1 (0.98-1.1)	0.23	1 (0.96-1.1)	0.67

Continued on next page

Continued from previous page

MPG or ASV	CAS		COAST		Meta	
	OR (95% CI)	p-value	OR (95% CI)	p-value	OR (95% CI)	p-value
Moraxellaceae.a5a08d96bf11fdeed28fc530d7f5d8c7	0.98 (0.88-1.1)	0.75	1.1 (0.99-1.1)	0.1	1 (0.97-1.1)	0.34
Moraxellaceae.rare	0.94 (0.87-1)	0.1	<b>1.1 (1-1.2)</b>	<b>0.0066</b>	1 (0.87-1.2)	0.83
Neisseriaceae.03f429f55feb0e805ba48c8c00867cf3	1.1 (0.98-1.2)	0.12	0.98 (0.9-1.1)	0.7	1 (0.93-1.1)	0.57
others.rare	<b>0.86 (0.8-0.94)</b>	<b>0.00071</b>	<b>0.75 (0.69-0.82)</b>	<b>4.60E-11</b>	<b>0.81 (0.7-0.93)</b>	<b>0.0023</b>
Pseudomonas.092523a4aed84953633a5f48bc2f87b9	1 (0.94-1.1)	0.97	1.1 (0.94-1.2)	0.33	1 (0.96-1.1)	0.6
Staphylococcus.29eb8e14ce61e9c4549de0f9cf200c19	<b>0.88 (0.83-0.94)</b>	<b>2.00E-04</b>	<b>0.93 (0.86-1)</b>	<b>0.049</b>	<b>0.9 (0.86-0.95)</b>	<b>3.80E-05</b>
Staphylococcus.rare	0.94 (0.87-1)	0.16	1 (0.91-1.1)	0.75	0.97 (0.9-1)	0.39
Streptococcus.357551f644064bb2b49614b1caab2fa5	<b>0.93 (0.88-0.99)</b>	<b>0.013</b>	<b>0.81 (0.76-0.86)</b>	<b>6.50E-11</b>	<b>0.87 (0.76-1)</b>	<b>0.046</b>
Streptococcus.4060107ffdf52e44f9d72343b7332609	1 (0.99-1.1)	0.12	1 (0.99-1.1)	0.095	<b>1 (1-1.1)</b>	<b>0.022</b>
Streptococcus.a3a3b519e04bf11edbb1d82497c2b34c	<b>0.86 (0.81-0.92)</b>	<b>1.60E-05</b>	<b>0.8 (0.75-0.85)</b>	<b>8.40E-12</b>	<b>0.83 (0.77-0.89)</b>	<b>5.60E-07</b>
Streptococcus.b069c8413c9eb478d4faa117c4372bd3	0.94 (0.86-1)	0.13	<b>0.8 (0.75-0.85)</b>	<b>4.00E-12</b>	0.86 (0.74-1)	0.062
Streptococcus.be1bac8c38e726214223a5ac13184ab7	<b>0.89 (0.82-0.96)</b>	<b>0.0027</b>	<b>0.84 (0.78-0.9)</b>	<b>4.10E-07</b>	<b>0.86 (0.81-0.91)</b>	<b>5.90E-08</b>
Streptococcus.rare	0.98 (0.92-1)	0.53	<b>0.86 (0.8-0.92)</b>	<b>2.10E-05</b>	0.92 (0.81-1)	0.19
Veillonella.fb81574d887fa68041a84122487a0e47	<b>0.86 (0.8-0.92)</b>	<b>2.40E-05</b>	<b>0.78 (0.72-0.84)</b>	<b>6.90E-12</b>	<b>0.82 (0.74-0.91)</b>	<b>9.60E-05</b>

**TABLE C.5: Viruses detected in nasopharyngeal samples in the first 3 years of life, collected from CAS and COAST.**

AdV = Adenovirus; CMV = Cytomegalovirus; CoV = Coronavirus; HPV = Human papilloma virus; HSV = Herpes simplex virus; MPV = Human metapneumovirus; PIV = Parainfluenza virus; Prop. = Proportion; RSV = Respiratory syncytial virus; RV = Rhinovirus; TTV = Transfusion transmitted / Torque teno virus; WU = WU (Washington University) polyomavirus. Group indicates Baltimore virus classification. Year 1 CAS samples were performed at a different laboratory from Year 2-3 CAS and COAST samples (**Methods**). Subtypes of certain viruses do not necessarily add up to the same count as their genus or parent virus type, due to overlapping samples and the fact that not all subtypes were measured. \*"Enterovirus" covers other members of the *Enterovirus* genus besides rhinovirus (CAS, COAST), Coxsackie virus, or echovirus (COAST only). # "Picornavirus" covers other members Picornaviridae family besides rhinovirus. ^RV subtype proportion given as estimate of total proportion, by multiplying subtype-positive proportion within RV-positive samples, with overall proportion of RV-positive samples amongst all samples; number in brackets indicates original proportion of tested samples. Note that this assumes the prevalence rates of different RV subtypes in LRI samples is similar to that of all samples. For CAS samples, RV subtyping was only performed in Year 1 samples that were collected during an LRI (N = 227), or in Years 2-3 samples that were collected during an LRI *and* that had initially tested positive for RV (N = 226).

Virus	Group	CAS up to year.1			CAS years 2-3			COAST up to year.3		
		Positive	Tested	Prop.(%)	Positive	Tested	Prop.(%)	Positive	Tested	Prop.(%)
Adenovirus (AdV)	I	9	862	1%	111	1310	8.50%	166	2922	5.70%
AdV-B	I	-	-	-	9	1310	0.69%	3	2922	0.10%
AdV-C	I	-	-	-	102	1310	7.80%	103	2922	3.50%
CMV	I	-	-	-	1	-	-	1	2922	0.03%
HPV	I	-	-	-	-	-	-	1	2922	0.03%
HSV	I	-	-	-	-	-	-	3	2922	0.10%
WU	I	-	-	-	-	-	-	1	2922	0.03%
Bocavirus	II	-	-	-	21	1310	1.60%	157	2922	5.40%
TTV	II	-	-	-	-	-	-	3	2922	0.10%
Coronavirus (CoV)	IV	36	862	4.20%	69	1310	5.30%	197	2922	6.70%
CoV-NL63	IV	-	-	-	30	1310	2.30%	82	2922	2.80%
CoV-OC43	IV	-	-	-	39	1310	3%	88	2922	3% <sup>a</sup>
Coxsackie	IV	-	-	-	-	-	-	7	2922	0.24%
Echovirus	IV	-	-	-	-	-	-	17	2922	0.58%
Enterovirus*	IV	-	-	-	155	1310	12%	47	2922	1.60%
Picornavirus#	IV	64	862	7.40%	-	-	-	-	-	-
Rhinovirus (RV)	IV	346	862	40%	607	1310	46%	1296	2922	44%
RV-A	IV	110	227	19% (48%) <sup>^</sup>	98	226	20% (43%) <sup>^</sup>	592	2922	20%
RV-B	IV	14	227	2.5% (6.2%) <sup>^</sup>	6	226	1.2% (2.7%) <sup>^</sup>	91	2922	3.10%
RV-C	IV	100	227	18% (44%) <sup>^</sup>	124	226	25% (55%) <sup>^</sup>	599	2922	20%
Influenza	V	25	862	2.90%	27	1310	2.10%	134	2922	4.60%
Influenza-A	V	-	-	-	25	1310	1.90%	105	2922	3.60%
Influenza-B	V	-	-	-	2	1310	0.15%	24	2922	0.82%
MPV	V	13	862	1.50%	55	1310	4.20%	120	2922	4.10%
Parainfluenza (PIV)	V	30	862	3.50%	103	1310	7.90%	219	2922	7.50%
PIV-1	V	-	-	-	24	1310	1.80%	42	2922	1.40%
PIV-2	V	-	-	-	8	1310	0.61%	11	2922	0.38%
PIV-3	V	-	-	-	63	1310	4.80%	122	2922	4.20%
PIV-4b	V	-	-	-	9	1310	0.69%	15	2922	0.51%
RSV	V	80	862	9.30%	107	1310	8.20%	225	2922	7.70%
RSV-A	V	-	-	-	73	1310	5.60%	121	2922	4.10%
RSV-B	V	-	-	-	35	1310	2.70%	96	2922	3.30%

**TABLE C.6: Results of GEE models associating viruses with respiratory illness status (well vs. unwell), with adjustments for child as subjects factor, and gender, age and season as covariates.**

95% CI = 95% confidence interval; MPV = Human metapneumovirus; OR = Odds ratio; RSV = Respiratory syncytial virus; RV = Rhinovirus. The model for analysis was a generalized estimating equation (GEE), of: respiratory illness status (well vs. unwell) ~ virus + gender + age + season | subject. Separate models were created for each virus (i.e. virus of interest vs. all others including no virus). The table is sorted by descending odds ratio in CAS, and statistically-significant associations are bolded.

Virus	CAS		COAST	
	OR (95% CI)	p-value	OR (95% CI)	p-value
MPV	<b>6.6 (3.1-14)</b>	<b>8.10E-07</b>	<b>19 (5.8-63)</b>	<b>1.30E-06</b>
Influenza	<b>5.7 (2.7-12)</b>	<b>7.60E-06</b>	<b>3.3 (2-5.7)</b>	<b>1.00E-05</b>
RV	<b>5.5 (4.4-6.7)</b>	<b>2.30E-58</b>	<b>3.8 (3.2-4.7)</b>	<b>9.90E-43</b>
Parainfluenza	<b>3.8 (2.5-5.9)</b>	<b>5.80E-10</b>	<b>6.8 (4.2-11)</b>	<b>2.60E-14</b>
RSV	<b>2.6 (1.8-3.7)</b>	<b>5.00E-08</b>	<b>14 (6.6-30)</b>	<b>1.00E-11</b>
Enterovirus	<b>2.1 (1.4-3.1)</b>	<b>0.00029</b>	<b>2.7 (1.2-6.1)</b>	<b>0.02</b>
Adenovirus	<b>1.9 (1.3-2.9)</b>	<b>0.0019</b>	<b>2.9 (1.8-4.6)</b>	<b>8.80E-06</b>
Bocavirus	1.4 (0.47-4)	0.55	1.4 (0.96-2.1)	0.082
Coronavirus	1.3 (0.88-2)	0.17	1.8 (1.2-2.7)	0.0033

**TABLE C.7: Results of GEE models associating viruses with winter season, with adjustments for child as subjects factor, and gender, age and respiratory illness as covariates.**

95% CI = 95% confidence interval; MPV = Human metapneumovirus; OR = Odds ratio; RSV = Respiratory syncytial virus; RV = Rhinovirus. The model for analysis was a generalized estimating equation (GEE), of: winter season (winter vs. other seasons) ~ virus + gender + age + respiratory illness status | subject. Separate models were created for each virus (i.e. virus of interest vs. all others including no virus). The table is sorted by descending odds ratio in CAS, and statistically-significant associations are bolded. \*Note that the definition of Enterovirus differed between CAS and COAST – see **Supplementary Table C.5**.

Virus	CAS		COAST	
	OR (95% CI)	p-value	OR (95% CI)	p-value
RSV	<b>4.5 (3.3-6)</b>	<b>2.40E-23</b>	<b>5.4 (4-7.2)</b>	<b>5.70E-30</b>
Influenza	<b>3 (1.7-5.3)</b>	<b>0.00018</b>	<b>4.9 (3.1-7.6)</b>	<b>6.80E-12</b>
Bocavirus	2.4 (0.99-5.6)	0.053	1.1 (0.73-1.6)	0.68
Enterovirus*	<b>1.7 (1.2-2.3)</b>	<b>0.0012</b>	<b>0.32 (0.13-0.77)</b>	<b>0.011</b>
MPV	1.1 (0.7-1.8)	0.62	1.6 (1.1-2.4)	0.028
Coronavirus	1.1 (0.71-1.6)	0.77	2.7 (1.8-4.1)	1.00E-06
Adenovirus	0.91 (0.63-1.3)	0.61	1.4 (1-2.1)	0.047
Parainfluenza	0.72 (0.49-1.1)	0.089	0.49 (0.33-0.72)	0.00025
RV	<b>0.58 (0.48-0.7)</b>	<b>2.40E-08</b>	<b>0.4 (0.33-0.48)</b>	<b>2.90E-23</b>



**TABLE C.8: Results of GEE models associating MPGs with the presence of any virus in the same sample, with or without adjustment for season and respiratory illness as covariates.**

95% CI = 95% confidence interval; MPG = Microbiome profile group; OR = Odds ratio. The model for analysis was a generalized estimating equation (GEE), of: any virus ~ MPG + gender + age +/- season +/- respiratory illness status | subject. Separate models were created for each MPG. Statistically-significant associations are bolded.

MPG	CAS				COAST			
	No covar		+ Illness & Season		No covar		+ Illness & Season	
	OR (95% CI)	P	OR (95% CI)	P	OR (95% CI)	P	OR (95% CI)	P
Haemophilus.bcd MPG	<b>3.2 (1.4-7.1)</b>	<b>0.005</b>	1.5 (0.67-3.6)	0.31	<b>2.5 (1.2-5.2)</b>	<b>0.016</b>	1.4 (0.64-3.1)	0.39
Haemophilus.rare MPG	<b>4.2 (1-18)</b>	<b>0.047</b>	2.5 (0.36-17)	0.35	7.6 (0.94-61)	0.057	5.4 (0.51-58)	0.16
Haemophilus.f579 MPG	<b>2.6 (1.7-3.9)</b>	<b>4.80E-06</b>	1.3 (0.83-1.9)	0.28	<b>2.2 (1.5-3.2)</b>	<b>2.50E-05</b>	1.4 (0.92-2.2)	0.11
Streptococcus.4060 MPG	<b>2.6 (1.9-3.5)</b>	<b>5.00E-10</b>	<b>1.5 (1-2)</b>	<b>0.033</b>	<b>3.3 (2.4-4.6)</b>	<b>4.80E-14</b>	<b>1.8 (1.3-2.6)</b>	<b>7.00E-04</b>
Streptococcus.rare MPG	2.2 (0.2-25)	0.52	2 (0.26-15)	0.51	NA	NA	NA	NA
Moraxella.d253 MPG	<b>1.7 (1.4-2)</b>	<b>8.60E-08</b>	<b>1.4 (1.1-1.7)</b>	<b>0.0016</b>	<b>1.5 (1.2-1.8)</b>	<b>5.50E-05</b>	1.2 (0.92-1.4)	0.21
others.rare MPG	<b>0.67 (0.51-0.88)</b>	<b>0.0037</b>	<b>0.66 (0.49-0.9)</b>	<b>0.009</b>	<b>0.67 (0.54-0.83)</b>	<b>0.00028</b>	0.97 (0.76-1.3)	0.84
Moraxellaceae.a5a0 MPG	0.51 (0.033-7.9)	0.63	1.6 (0.11-25)	0.73	0.64 (0.36-1.1)	0.13	0.61 (0.31-1.2)	0.17
Corynebacterium.rare MPG	<b>0.75 (0.37-1.5)</b>	0.42	1.2 (0.48-3)	0.7	<b>0.21 (0.066-0.67)</b>	<b>0.0087</b>	0.43 (0.14-1.4)	0.16
Moraxella.rare MPG	<b>0.6 (0.35-1)</b>	<b>0.066</b>	0.91 (0.53-1.5)	0.72	1.6 (0.82-3)	0.17	<b>2.2 (1.1-4.6)</b>	<b>0.031</b>
Moraxellaceae.6028 MPG	<b>0.5 (0.33-0.76)</b>	<b>0.0011</b>	0.73 (0.47-1.1)	0.17	0.9 (0.56-1.5)	0.68	0.97 (0.57-1.7)	0.91
Corynebacterium.cb50 MPG	<b>0.51 (0.32-0.8)</b>	<b>0.0032</b>	0.76 (0.43-1.3)	0.34	<b>0.37 (0.39-0.84)</b>	<b>0.0043</b>	0.76 (0.5-1.2)	0.21
Alliotooccus.dd2e MPG	<b>0.46 (0.34-0.61)</b>	<b>6.40E-08</b>	<b>0.74 (0.54-1)</b>	<b>0.067</b>	<b>0.33 (0.26-0.41)</b>	<b>2.50E-22</b>	<b>0.53 (0.4-0.68)</b>	<b>1.50E-06</b>
Staphylococcus.29eb MPG	<b>0.34 (0.24-0.48)</b>	<b>5.50E-10</b>	<b>0.69 (0.47-1)</b>	<b>0.053</b>	<b>0.45 (0.27-0.74)</b>	<b>0.002</b>	<b>0.82 (0.47-1.5)</b>	0.5

**TABLE C.9: GLM models associating wheeze and asthma outcomes with proportion of illness-associated MPGs in routine healthy samples within first 2 years of life, stratified by early sensitization status and npEM clusters; (A) in CAS, (B) in COAST.**

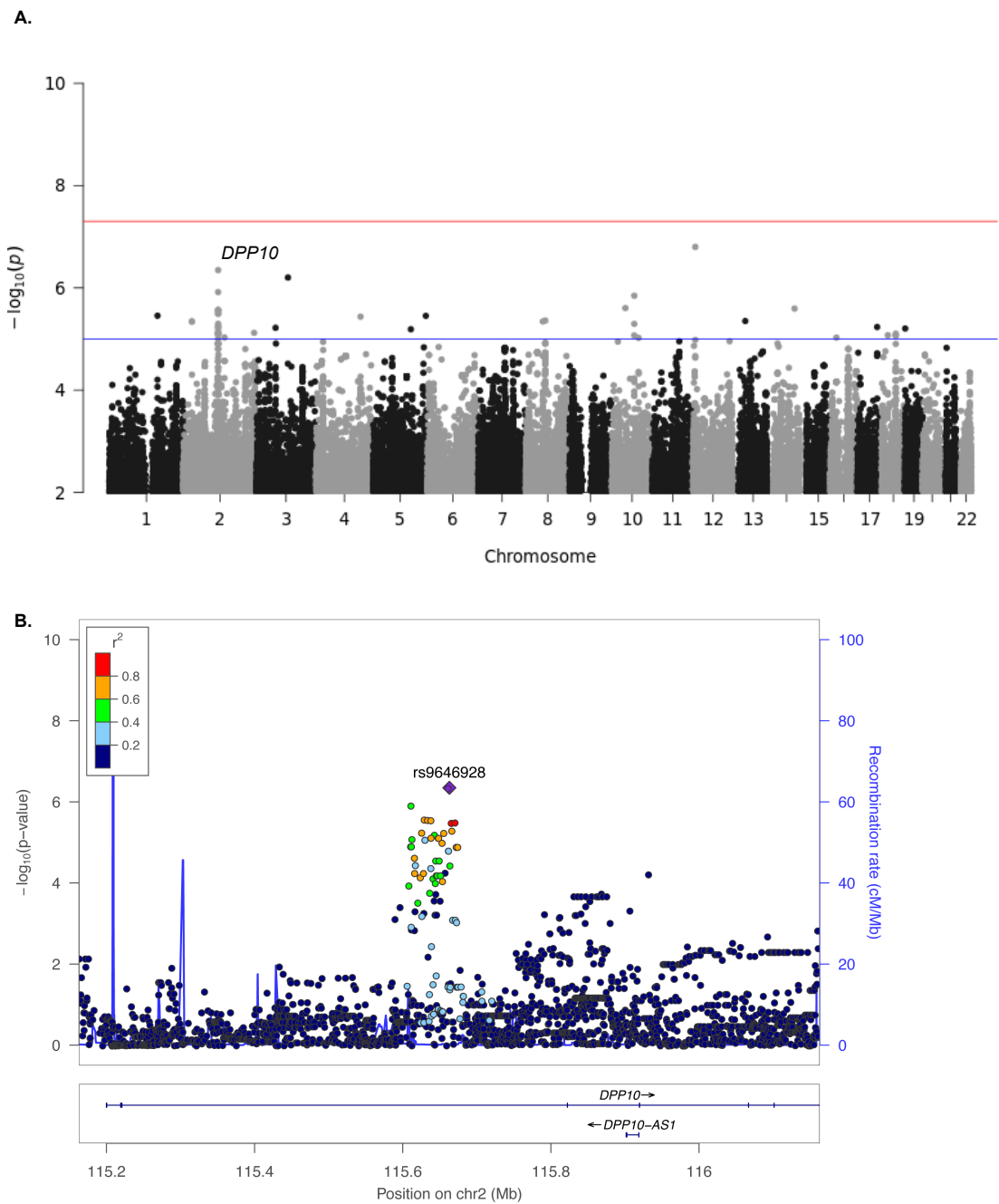
GLM model: outcome ~ proportion of illness-associated MPGs amongst all routine healthy samples in the first 2 years of life; separate model for each sensitisation status or npEM group. \*In COAST, we used any IgE against *Alternaria*, cat, dog, *Dermatophagoides pteronyssinus*, or *D. farinae* in the first two years of life, whereas in CAS we used IgE against cat, couch grass, *D. pteronyssinus*, mould (including *Alternaria* and *Aspergillus*), peanuts, and ryegrass (see **Methods**).

Outcome	Exposure	Stratified	OR (95% CI)	p-value
<b>A. Outcome ~ illness-associated MPGs, stratified by sensitisation status in CAS</b>				
Wheeze at age 5	Illness-associated MPGs in first 2 years of life	<i>All</i>	1.4 (0.99-2.1)	0.058
		<b>Early sensitised</b>	<b>2.2 (1.3-3.8)</b>	<b>0.003</b>
		Not early sensitised	0.8 (0.43-1.5)	0.47
Transient wheeze	Illness-associated MPGs in first 2 years of life	<i>All</i>	1.1 (0.8-1.6)	0.47
		<i>Early sensitised</i>	0.64 (0.38-1.1)	0.089
		<b>Not early sensitised</b>	<b>2 (1.1-3.5)</b>	<b>0.018</b>
Asthma at age 5	Illness-associated MPGs in first 2 years of life	<i>All</i>	1.5 (0.96-2.4)	0.071
		<b>Early sensitised</b>	<b>2 (1.1-3.8)</b>	<b>0.023</b>
		Not early sensitised	1.1 (0.51-2.3)	0.82
Asthma at age 10	Illness-associated MPGs in first 2 years of life	<i>All</i>	1.3 (0.76-2.3)	0.33
		<i>Early sensitised</i>	2 (0.99-3.9)	0.054
		Not early sensitised	0.83 (0.3-2.3)	0.73
<b>B. Outcome ~ illness-associated MPGs, stratified by sensitisation status in COAST*</b>				
Asthma at age 6	Illness-associated MPGs in first 2 years of life	<i>All</i>	0.83 (0.6-1.2)	0.28
		<i>Early sensitised</i>	0.84 (0.43-1.6)	0.62
		<i>Not early sensitised</i>	0.67 (0.43-1.1)	0.089
Transient wheeze	Illness-associated MPGs in first 2 years of life	<i>All</i>	0.9 (0.6-1.3)	0.6
		<i>Early sensitised</i>	1.1 (0.25-4.5)	0.93
		<i>Not early-sensitised</i>	0.89 (0.58-1.4)	0.62
Asthma at age 13	Illness-associated MPGs in first 2 years of life	<i>All</i>	1 (0.71-1.4)	0.95
		<i>Early sensitised</i>	0.77 (0.37-1.6)	0.49
		<i>Not early sensitised</i>	1 (0.65-1.6)	0.94

## Appendix D

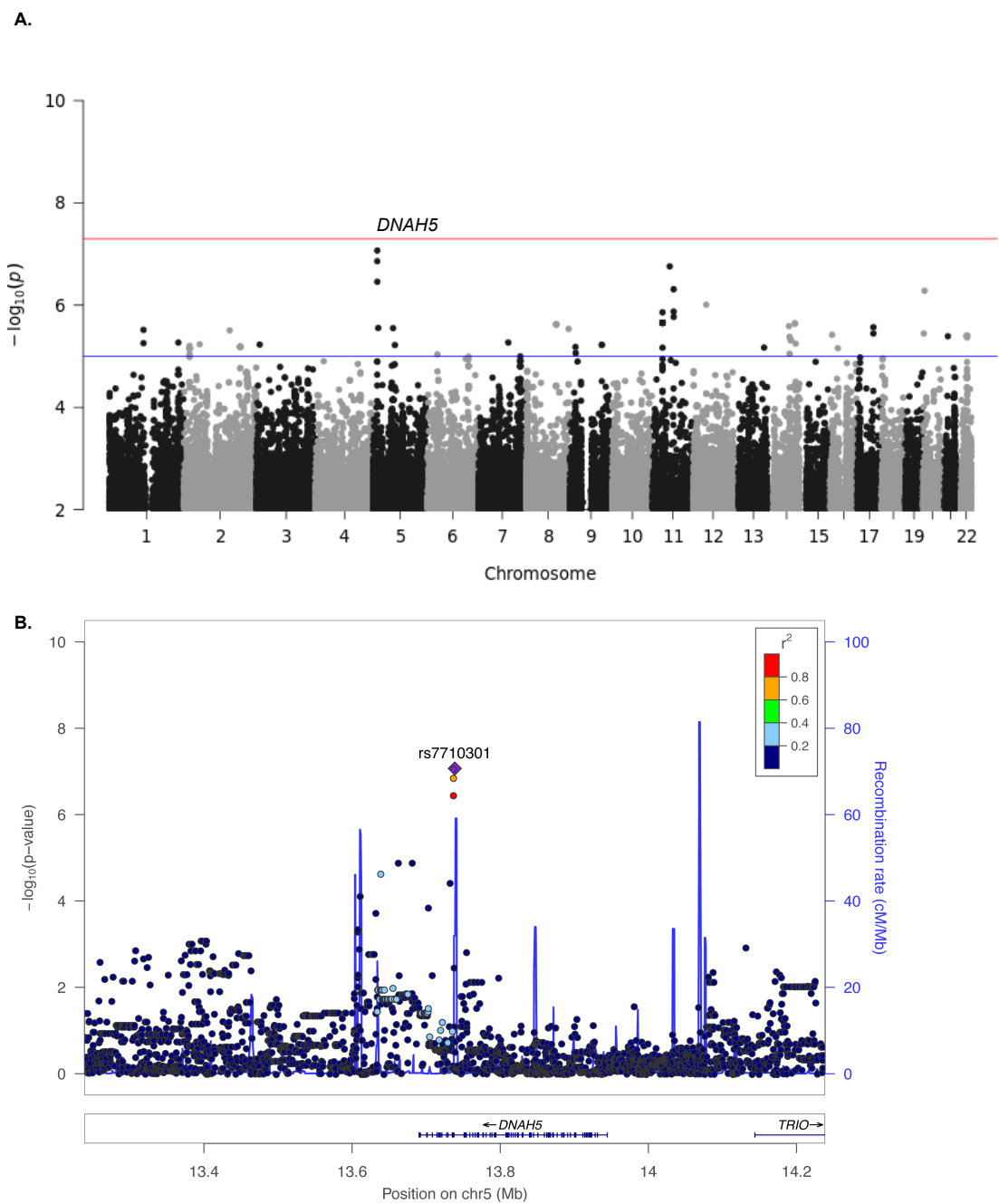
# Supplementary Figures and Tables for Chapter 5

The rest of this page has been intentionally left blank.



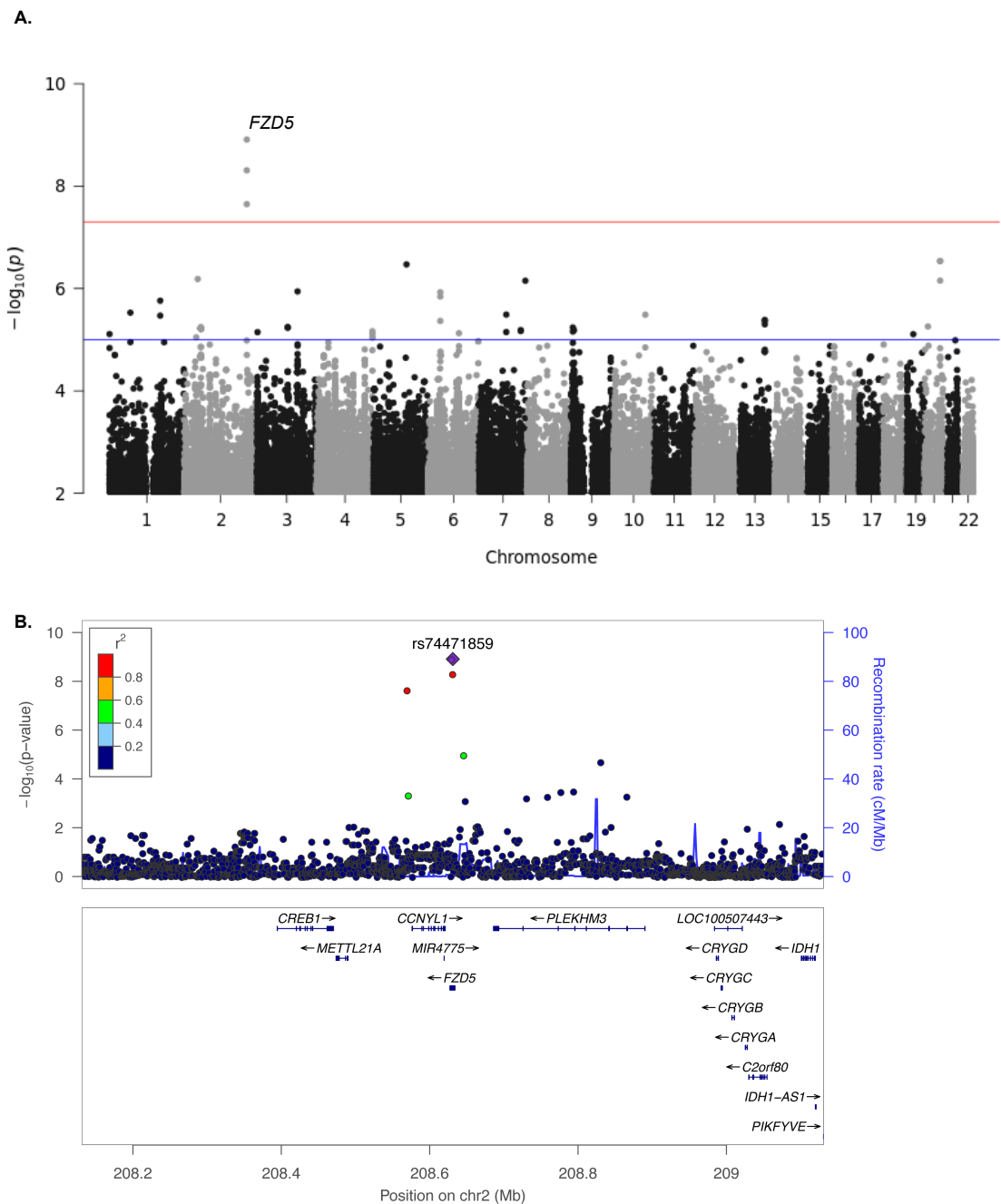
**FIGURE D.1: Manhattan plots of genome-wide association scans for parent-reported wheeze at age 5 in CAS.**

(A) General Manhattan plot; red line indicates threshold for genome-wide significance ( $5 \times 10^{-8}$ ); blue line indicates threshold for suggestive association ( $1 \times 10^{-5}$ ). (B) LocusZoom plot focusing on the locus of interest at Chromosome 2 near *DPP10*. LD  $R^2$  values and recombination rates given as per hg19/1000 Genomes Nov 2014 EUR reference genome.



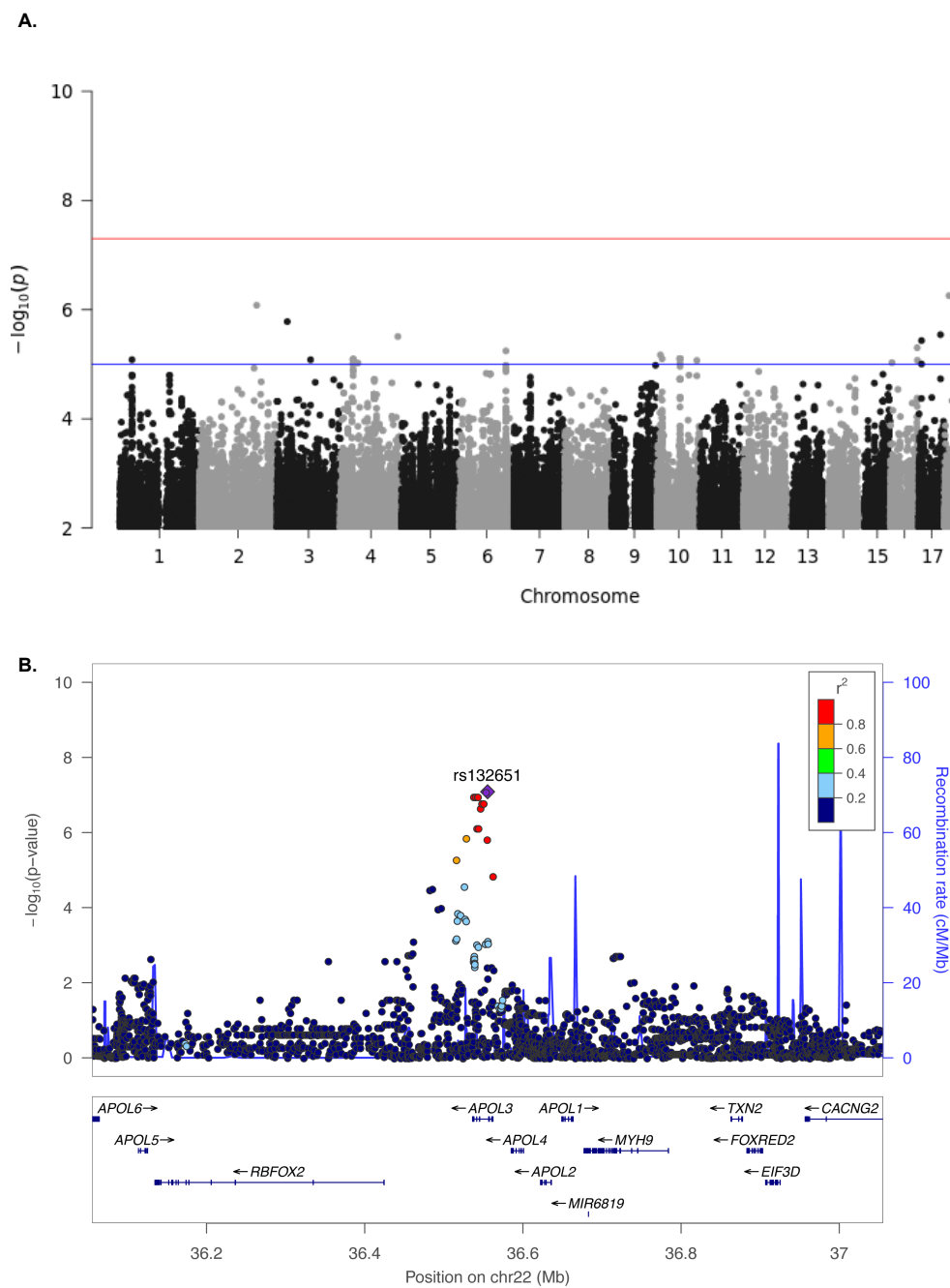
**FIGURE D.2: Manhattan plots of genome-wide association scan for any wheezy LRI at age 1 in CAS.**

(A) General Manhattan plot; red line indicates threshold for genome-wide significance ( $5 \times 10^{-8}$ ); blue line indicates threshold for suggestive association ( $1 \times 10^{-5}$ ). (B) LocusZoom plot focusing on the locus of interest at Chromosome 5 near *DNAH5*. LD  $R^2$  values and recombination rates given as per hg19/1000 Genomes Nov 2014 EUR reference genome.



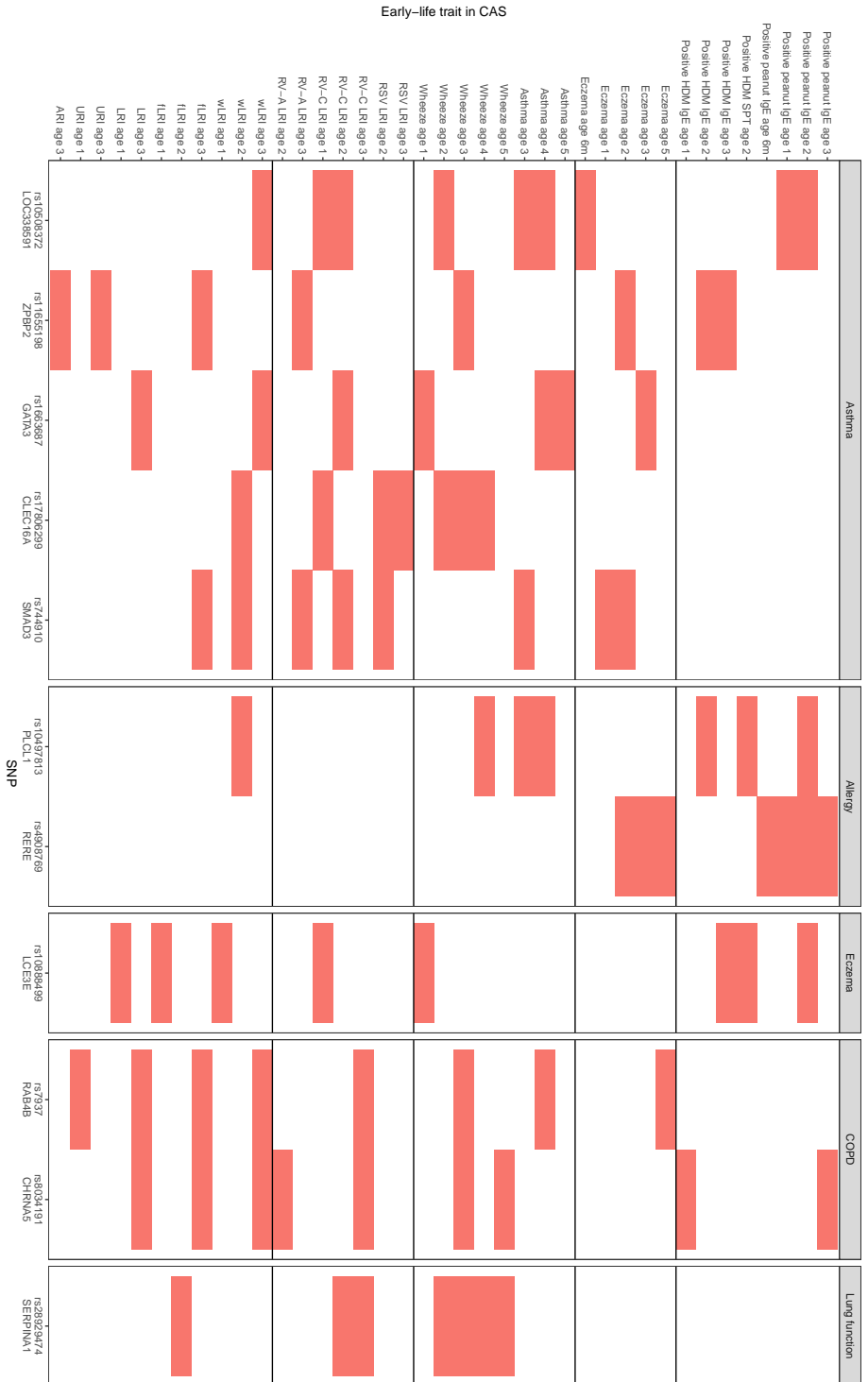
**FIGURE D.3: Manhattan plots of genome-wide association scan for any rhinovirus-C-associated LRI at age 1 in CAS.**

(A) General Manhattan plot; red line indicates threshold for genome-wide significance ( $5 \times 10^{-8}$ ); blue line indicates threshold for suggestive association ( $1 \times 10^{-5}$ ). (B) LocusZoom plot focusing on the locus of interest at Chromosome 2 near *FZD5*. LD  $R^2$  values and recombination rates given as per hg19/1000 Genomes Nov 2014 EUR reference genome.



**FIGURE D.4: Manhattan plots of genome-wide association scan for any rhinovirus-A-associated LRI at age 1 in CAS.**

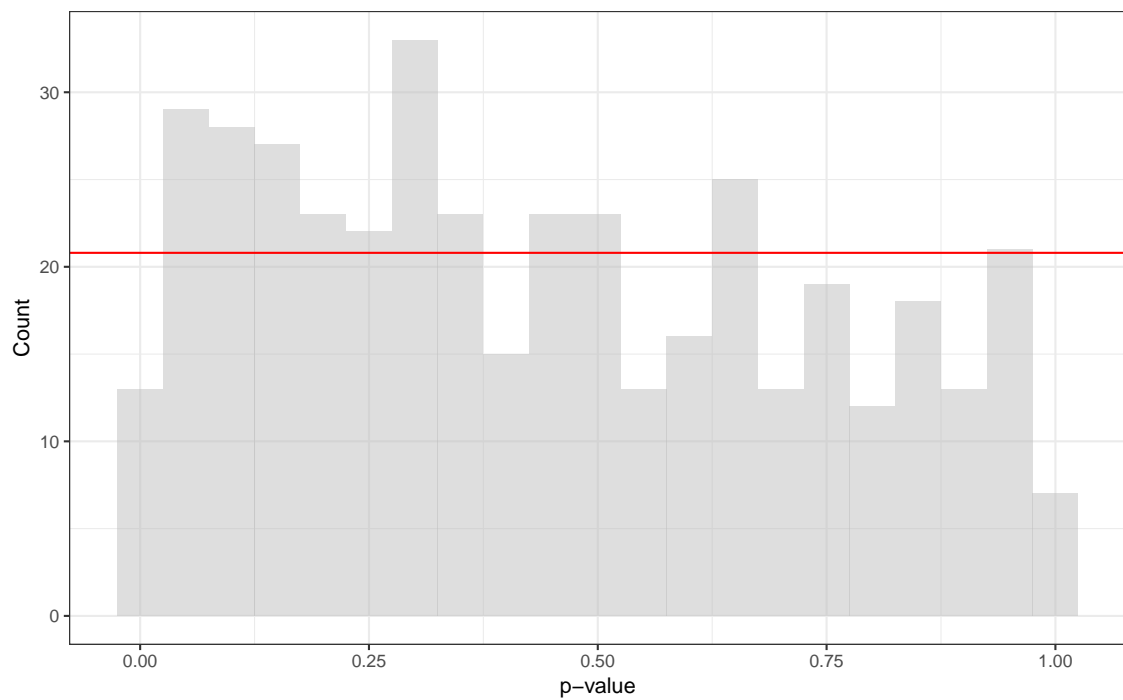
(A) General Manhattan plot; red line indicates threshold for genome-wide significance ( $5 \times 10^{-8}$ ); blue line indicates threshold for suggestive association ( $1 \times 10^{-5}$ ). (B) LocusZoom plot focusing on the locus of interest at Chromosome 5 near *DNAH5*. LD  $R^2$  values and recombination rates given as per hg19/1000 Genomes Nov 2014 EUR reference genome.



**FIGURE D.5: GWAS catalogue SNPs most frequently associated (at an unadjusted threshold) with an early-life trait in CAS, sorted by phenotypes.**

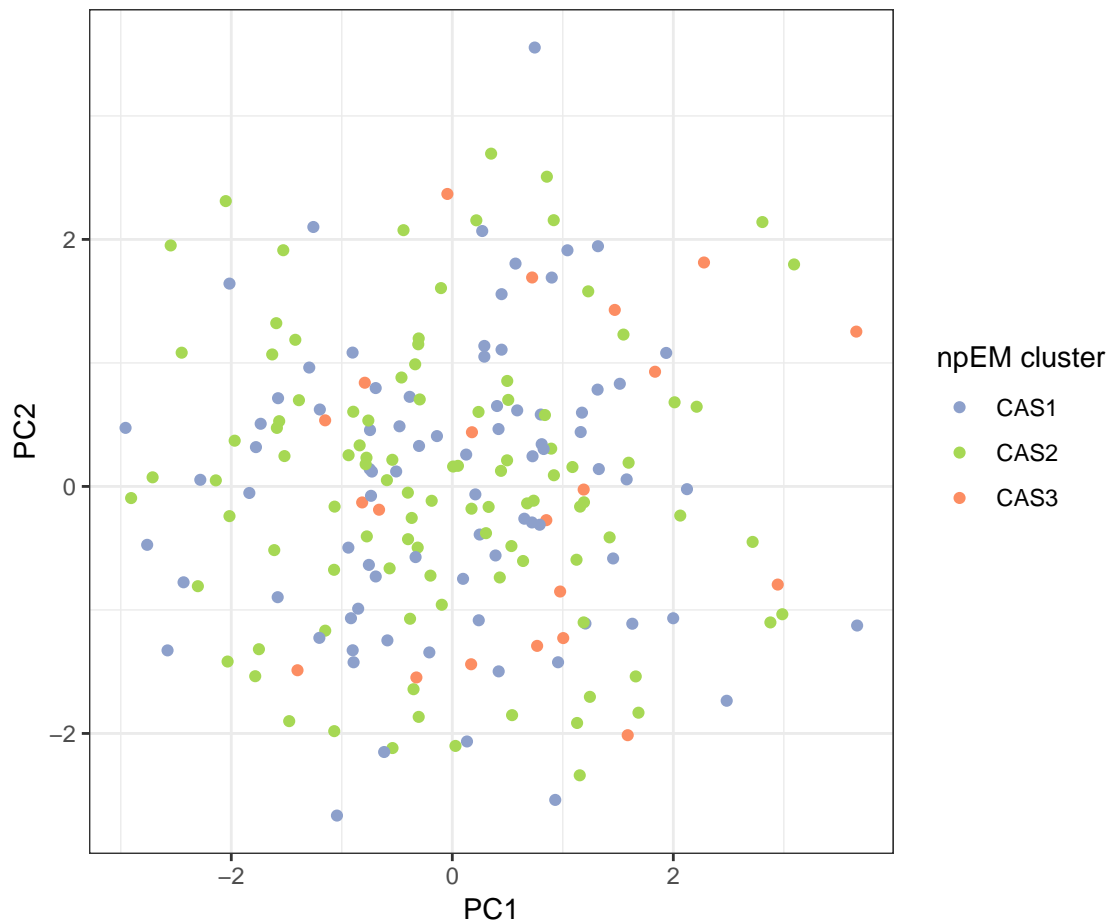
Shaded cells indicate significance at unadjusted threshold ( $p < 0.05$ ) for CAS early-life traits. Vertical axis traits are CAS early-life traits, while traits indicated in the top horizontal axis are GWAS catalogue traits for which the SNPs were originally significantly associated.





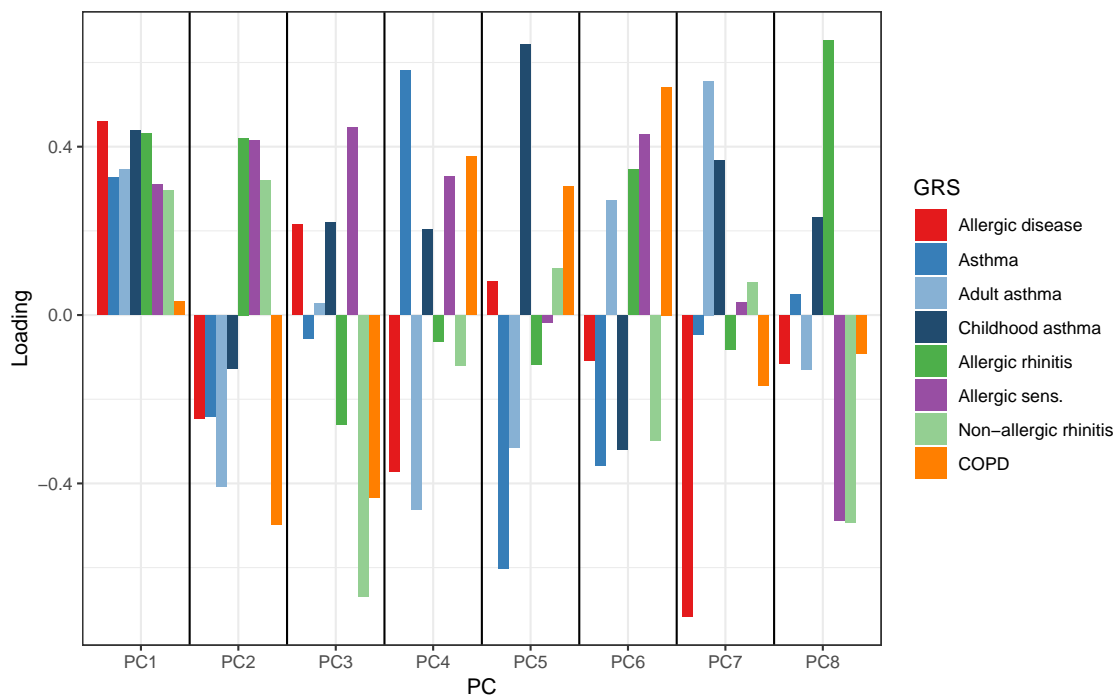
**FIGURE D.6: Histogram of p-values of GWAS catalogue SNPs in association with eczema at age 6m in CAS.**

Red line represents frequency of values from uniform distribution given histogram break size (0.05). Note overrepresentation of p-values on the left side, with 249 p-values less than 0.50 compared to 167 above (Fisher exact  $p = 0.005$ ).



**FIGURE D.7:** Scatterplot of the first two principal components (PC)s from principal components analysis of GRS in CAS, coloured by npEM clusters as per Tang et al 2018.

There was no clear segregation in genetic risk for allergy, asthma and COPD, and no clear separation of npEM clusters by genetic risk. Note however that members of CAS3 tended to have greater PC1.



**FIGURE D.8: Principal components analysis of GRS in CAS identified an “atopic vector” in PC1.**

Loadings of principal components by individual GRS. Note that PC1 was almost universally loaded with allergy-related GRS, and not by non-allergic GRS (COPD).

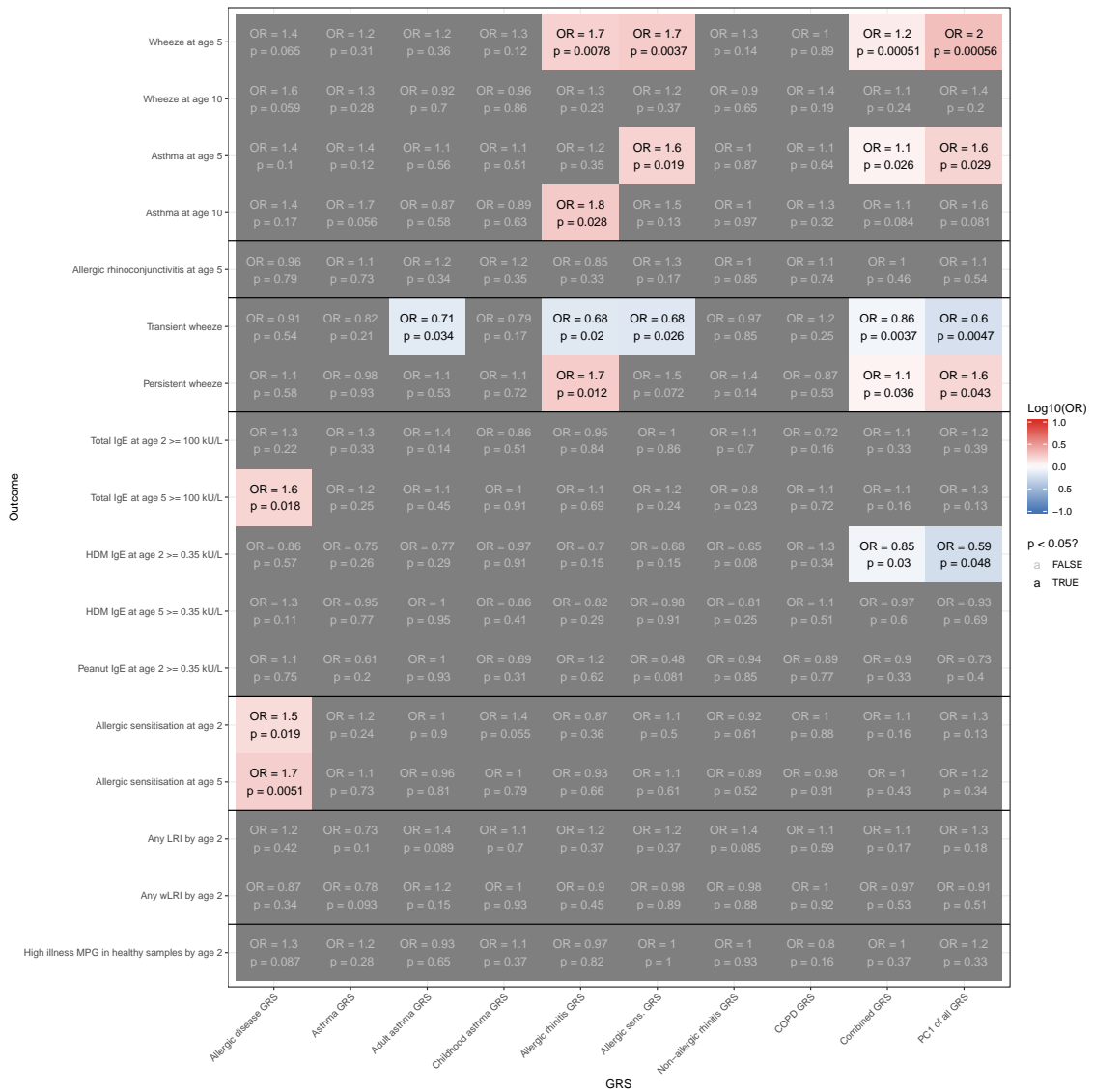


FIGURE D.9: GLM associations with early-life traits vs. GRS, with membership in the high-risk npEM cluster “Cluster 3” and sex as potential covariates.

GLM of early-life trait ~ GRS + npEM cluster 3 + sex. Note the diminished ORs compared to Figure 5.4.

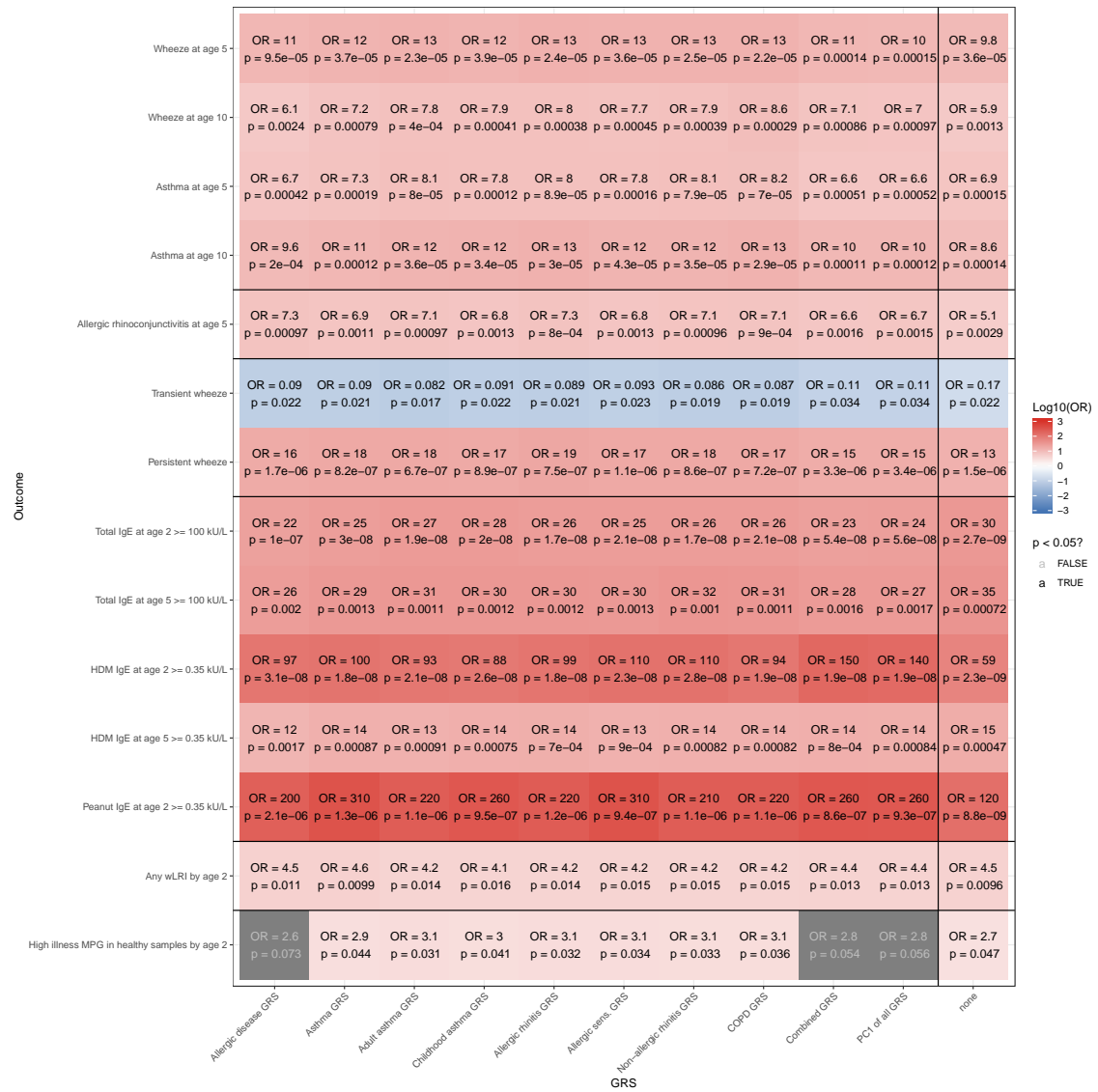


FIGURE D.10: GLM associations with early-life traits vs. npEM cluster 3, with sex and GRS as potential covariates.

GLM of early-life trait ~ npEM Cluster 3 + sex +/- GRS. Note the relatively unchanged effect sizes and statistical significance compared to the rightmost column.

**TABLE D.1: GWAS catalogue SNPs most frequently associated with an early-life trait in CAS, at the unadjusted p-value threshold**

Top candidate SNPs with nominally-significant ( $p < 0.05$ ) associations with seven or more phenotypes in CAS, sorted in order of associated phenotype in original GWAS, and descending number of associated CAS phenotypes. Eff. alle. = Effect allele

SNP	CHR	BP	Eff. alle.	Closest Gene	Effect of variant	Original effect in source GWAS	CAS phenotype	OR	P (unadj.)	P (FDR-BH)
rs10508372	10	8972018	A	LOC338391 / LOC105755953 / LOC101928272	Intergenic	Asthma (Hirota 2011) (risk allele C)	wLRI at age 3 RV-C LRI at age 1 RV-C LRI at age 2 Wheeze at age 2 Wheeze at age 3 Asthma at age 4 Positive peanut IgE at age 1 Positive peanut IgE at age 2 Eczema at age 6m	2.22 0.38 2.20 2.67 2.94 2.60 3.24 2.90 2.23	0.03 0.05 0.04 0.01 0.02 0.03 0.01 0.03 0.03	0.81 0.94 0.71 0.71 0.86 0.98 0.43 0.88 0.68
rs744910	15	67446785	A	SMAD3	Intron	Asthma (Moffatt 2010) (risk allele G)	fLRI at age 3 wLRI at age 2 RV-A LRI at age 3 RV-C LRI at age 2 RSV LRI at age 2 Asthma at age 3 Eczema at age 1 Eczema at age 2	1.78 1.51 1.82 1.80 1.77 2.37 0.52 0.64	0.02 0.04 0.04 0.01 0.03 0.02 0.002 0.03	0.82 0.93 0.67 0.71 0.85 0.86 0.59 0.59
rs11655198	17	38026169	T	ZFPB2	Intron	Asthma (Pickrell 2016)	ARI at age 3 URI at age 3 fLRI at age 3 RV-A LRI at age 3 Wheeze at age 3 Positive HDM IgE at age 2 Positive HDM IgE at age 3 Eczema at age 2	0.32 0.41 0.60 0.53 0.56 1.76 1.71 1.68	0.004 0.004 0.05 0.03 0.01 0.04 0.04 0.02	0.93 0.97 0.82 0.67 0.65 0.91 0.98 0.56
rs1663687	10	9054787	A	GATA3 / CELF2 / LOC105755953 / LOC101928272	Regulatory region	Asthma (Demanais 2017) (protective allele A)	LRI at age 3 wLRI at age 3 RV-C LRI at age 2 Wheeze at age 1 Asthma at age 4 Asthma at age 5 Eczema at age 3	0.68 0.67 0.63 0.63 0.50 0.57 1.49	0.04 0.05 0.03 0.02 0.03 0.04 0.05	0.76 0.85 0.71 0.88 0.98 0.89 0.84
rs17806299	16	11199980	A	CLEC16A / DEX1 / SOCS1	Intron	Asthma (Demanais 2017) (risk allele G)	wLRI at age 2 RV-C LRI at age 1 RSV LRI at age 2 RSV LRI at age 3 Wheeze at age 2 Wheeze at age 3 Wheeze at age 4	1.84 1.79 2.22 2.29 1.78 1.87 1.85	0.03 0.04 0.02 0.04 0.04 0.03 0.04	0.93 0.94 0.85 0.84 0.78 0.68 0.98
rs10497813	2	198914072	G	P1CL1	Intron	Allergy (Hinds 2013) (risk allele G)	wLRI at age 2 Wheeze at age 4 Asthma at age 3 Asthma at age 4 Positive HDM IgE at age 2 Positive peanut IgE at age 2	1.55 0.55 0.36 0.50 0.45 0.41	0.03 0.01 0.01 0.03 0.004 0.02	0.93 0.98 0.86 0.98 0.87 0.88

Continued on next page

Continued from previous page

SNP	CHR	BP	Eff. alle.	Closest Gene	Effect of variant	Original effect in source GWAS	CAS phenotype	OR	P (unadj.)	P (FDR-BF)
rs4908769	1	8701288	T	RERE	Intron	Allergy (Pickrell 2016)	Positive HDM SPT at age 2 Positive peanut IgE at age 6m Positive peanut IgE at age 1 Positive peanut IgE at age 2 Positive peanut IgE at age 3 Eczema at age 2 Eczema at age 3 Eczema at age 5	0.34 2.15 1.86 2.42 3.75 0.54 0.56 0.56	0.0001 0.03 0.04 0.02 0.003 0.01 0.02 0.02	0.05 0.73 0.90 0.88 0.97 0.56 0.84 0.74
rs10888499	1	152532742	C	LCE3E / CRCT1	Intergenic	Atopic eczema (Baurecht 2015)	LRI at age 1 fLRI at age 1 wLRI at age 1 RV-C LRI at age 1 Wheeze at age 1 Positive HDM IgE at age 3 Positive peanut IgE at age 2 Positive HDM SPT at age 2	0.49 0.42 0.49 0.54 0.55 0.52 0.14 0.48	0.01 0.004 0.02 0.03 0.03 0.04 0.01 0.03	0.75 0.86 0.59 0.94 0.88 0.98 0.88 1.00
rs8034191	15	78806023	C	CHRNA5 / CHRNA3 / CHRNA4 / HYKK / LOC123688	Intron	COPD (Pillai 2009) (risk allele C)	LRI at age 3 fLRI at age 3 wLRI at age 3 RV-A LRI at age 2 RV-C LRI at age 3 Wheeze at age 3 Wheeze at age 5 Positive HDM IgE at age 1 Positive peanut IgE at age 3	0.63 0.57 0.54 0.56 0.46 0.56 0.59 0.29 0.21	0.03 0.05 0.01 0.03 0.02 0.02 0.04 0.03 0.02	0.76 0.82 0.53 0.95 0.96 0.65 0.81 1.00 0.97
rs7937	19	41302706	C	RAB4B / EGLN2 / CYP2A6	3'-UTR	COPD (Cho 2011)	URI at age 1 LRI at age 3 fLRI at age 3 wLRI at age 3 RV-C LRI at age 3 Wheeze at age 3 Asthma at age 4 Eczema at age 5	1.98 1.68 2.45 1.77 1.79 1.78 1.98 2.05	0.05 0.02 0.0008 0.01 0.04 0.01 0.03 0.003	0.96 0.76 0.36 0.53 0.96 0.65 0.98 0.74
rs28929474	14	94844947	T	SERPINA1	Missense	Reduced FEV1:FVC, FEV1 (Li 2018) (risk allele T)	fLRI at age 2 RV-C LRI at age 2 RV-C LRI at age 3 Wheeze at age 2 Wheeze at age 3 Wheeze at age 4 Wheeze at age 5	4.52 4.10 5.29 4.43 4.35 8.86 4.40	0.04 0.04 0.02 0.04 0.05 0.01 0.05	0.90 0.71 0.96 0.78 0.69 0.98 0.81

TABLE D.2: Candidate SNPs associated with wheeze or asthma at age five in CAS, at the unadjusted p-value threshold

Eff. alle. = Effect allele

SNP	CHR	BP	Eff. Alle	Closest gene	Effect of variant	Original effect in source GWAS	OR	P (unadj.)	P (FDR-BH)
<b>A. Association with parent-reported wheeze at age five in CAS</b>									
rs1102705	1	172700668	G	FASLG / TNFSF18 / SLC25A38P1	Intergenic	Allergy (Ferreira 2017) (risk allele G)	2.20	0.023	0.81
rs1143633	2	115900467	T	IL1B	Intron	Allergy (Ferreira 2017) (risk allele C)	0.57	0.040	0.81
rs9815663	3	3614887	T	IL5RA / LOC100130207	Intron	Childhood onset asthma (Formo 2015) (protective allele T)	0.52	0.047	0.81
rs9687749	5	131955577	T	IL13 / RAD50	Intron	Allergic rhinitis (Waage 2018) (risk allele T)	0.53	0.034	0.81
rs546626089	6	31088145	A	PSORS1C1 / CDSN	Missense / Intron	Proxy for rs3095318; Adult asthma (Almoogura 2016)	1.73	0.048	0.81
rs2857595	6	31568469	A	NCR3 / UQCRRHP1	Intergenic	Decreased FEV1:FVC ratio (Wain 2017) (risk allele A)	2.42	0.003	0.81
rs204993	6	32155381	G	PBX2	Intron	Asthma (Hirota 2011) (risk allele A)	1.61	0.049	0.81
rs176095	6	32158319	G	GPSM3 / PBX2	Regulatory region	Atopic eczema (Hirota 2012) (risk allele T)	2.10	0.004	0.81
rs3129943	6	32338695	G	C6orf10 / LOC101929163	Intron	Asthma (Hirota 2011) (risk allele T)	1.86	0.021	0.81
rs7764819	6	32680576	G	HLA-DQA2 / HLA-DQB1 / MTCO3P1	Intergenic	Decreased FEV1:FVC ratio (Hancock 2012) (risk allele T)	0.33	0.024	0.81
rs7009110	8	81291879	T	ZBTB10 / LOC100216346 / RNJ6-1213P	Intron	Seasonal allergic rhinitis and asthma (Ferreira 2013) (risk allele T)	1.65	0.028	0.81
rs1251256	9	6231239	G	IL33 / RANBP6 / TPD52L3	Intron	Childhood onset asthma (Demenaïs 2017) (risk allele A)	1.66	0.046	0.81
rs10738626	9	22373457	T	DMRTA1 / CDKN2B-A51	Intergenic	Atopic eczema (Schaarschmidt 2015) (risk allele T)	2.02	0.008	0.81
rs7068966	10	12277992	C	CDIC123	Intron	Decreased FEV1:FVC ratio; FEV1 (Soler Artigas 2011) (protective allele T)	0.61	0.038	0.81
rs209961	11	62310909	C	AHNAK	Intron	Decreased FEV1 (Wain 2017) (protective allele C)	0.59	0.046	0.81
rs95226	12	84868673	G	SLC6A15 / TMTC2 / LOC105369875	Intergenic	Atopic march (Marenholz 2015) (risk allele G)	3.62	0.024	0.81
rs9540294	13	65564031	G	PCDH9 / LGMNPI / STARP1	Intergenic	Recalcitrant atopic dermatitis (Kim 2015) (risk allele G)	0.38	0.041	0.81
rs754388	14	93115410	G	RIN3	Intron	COPD (Hobbs 2017) (risk allele C)	0.42	0.014	0.81
rs28929474	14	94844947	T	SERPINA1	Missense	Decreased FEV1:FVC ratio; FEV1 (Li 2018) (risk allele T)	4.40	0.050	0.81
rs12914385	15	78898723	T	CHRNA3 / CHRNA5 / AGPHD1	Intron	COPD (Cho 2014) (risk allele T)	0.57	0.030	0.81
rs8034191	15	78806023	T	CHRNA5 / CHRNA3 / CHRNA4 / HYKK	Intron	COPD (Pillai 2009) (risk allele C)	0.59	0.039	0.81
rs72724130	15	41977690	T	MGA	Intron	Decreased FEV1:FVC ratio (Wain 2017) (risk allele T)	2.27	0.039	0.81
rs2286351	17	13928401	A	CDRT15P1	Intron	COPD (Burkart 2018)	2.84	0.049	0.81
<b>B. Association with physician-diagnosed asthma at age five in CAS</b>									
rs3126085	1	152300817	A	FLG / FLG-A51	Intron	Atopic eczema (Sun 2011) (risk allele A)	2.27	0.011	0.75
rs350729	2	52983773	T	ASB3 / IJND / SOCS / CEBPB	Intron	Response to bronchodilator in asthma (Israel 2015)	1.91	0.043	0.89
rs19973	3	187633268	A	BCL6 / LPP-AS2 / LOC105374264	Intron	Allergy (Ferreira 2017) (risk allele A)	0.45	0.020	0.75
rs6583203	3	197079586	C	DLG1 / LOC101926923 / LOC105374308	Regulatory region	Allergic rhinitis (Bunyavanch 2014)	1.84	0.032	0.88
rs9687749	5	131955577	T	IL13 / RAD50	Intron	Allergic rhinitis (Waage 2018) (risk allele T)	0.49	0.047	0.89
rs546626089	6	31088145	A	PSORS1C1 / CDSN	Missense / Intron	Proxy for rs3095318; Adult asthma (Almoogura 2016)	2.20	0.014	0.75
rs9368677	6	31272321	A	HLA-C / LOC105375015	Intron	Atopic eczema (Hirota 2012) (risk allele G)	4.00	0.017	0.75
rs9260772	6	31352113	C	MICA / HLA-C / HLA-S	Intergenic	Allergy (Hinds 2013) (risk allele C)	2.04	0.026	0.79
rs2857595	6	31568469	A	NCR3 / UQCRRHP1	Intergenic	Decreased FEV1:FVC ratio (Wain 2017) (risk allele A)	2.43	0.007	0.75
rs1251256	9	6231239	G	IL33 / RANBP6 / TPD52L3	Intron	Childhood onset asthma (Demenaïs 2017) (risk allele A)	2.04	0.018	0.75
rs1663687	10	9054787	A	GATA3 / CELF2 / LOC105755953 / LOC101928272	Regulatory region	Asthma (Demenaïs 2017) (protective allele A)	0.57	0.038	0.89
rs12413578	10	9049253	T	GATA3 / LOC105755953 / LOC101928272	Intergenic	Allergy (Ferreira 2017) (risk allele C)	0.23	0.049	0.89
rs7927044	11	127761666	A	NR / LOC107984373	Intron	Childhood onset asthma (Forma 2012) (protective allele A)	14.55	0.023	0.78
rs7137828	12	111932800	C	A1XX2	Intergenic	Allergy (Ferreira 2017) (risk allele T)	0.47	0.015	0.75
rs111371454	15	41760617	G	ITPKA / RTF1	Intron	Allergic rhinitis (Waage 2018) (risk allele G)	0.44	0.027	0.79
rs3743609	16	75467021	G	CFDP1	Intron	Decreased FEV1:FVC ratio (Wain 2017) (risk allele C)	2.35	0.004	0.75
rs9303280	17	38074031	T	GSDMB / IKZF3	Non-coding transcript	Allergy (Hinds 2013) (protective allele T)	0.58	0.049	0.89



**TABLE D.3: Selected significant and suggestive SNPs for longitudinal genome-wide association scans for early-life traits in CAS, with repeatABEL.**

Eff. alle. = Effect allele; Ref. alle. = Reference allele.

Phenotype	SNP	CHR	Position	Eff. alle.	Ref. alle.	Closest Gene	Effect of variant	Beta	P-value	Lambda
Wheeze in first 5 years of life	rs720267	17	36020981	C	G	<i>RP11-697E22.2/DDX52</i>	Intron / non-coding	0.31	1.11E-07	1.01
	rs6577351	1	103552920	A	G	<i>COL11A1</i>	Intron / upstream	0.30	7.14E-10	1.06
Asthma diagnosis at ages 3-5	rs138681209	3	114457040	T	C	<i>ZBTB20</i>	Intron / non-coding	0.60	5.59E-09	
	rs61823519	1	223689931	T	A	-	intergenic	0.43	6.04E-09	
LRI in first 3 years of life	rs201631417	12	105311431	T	G	<i>SLC41A2</i>	Intron	1.60	2.11E-09	1.02
	rs138414325	13	41759533	C	T	<i>KBTBD7</i>	Downstream	2.05	3.58E-09	
fLRI in first 3 years of life	rs72926646	6	93291617	C	A	-	Intergenic	1.02	2.06E-08	
	rs61619392	4	101472438	A	C	<i>EMCN</i>	Non-coding	1.53	5.57E-08	
wLRI in first 3 years of life	rs72733544	14	69419472	G	T	<i>ACTN1</i>	Intron	1.27	6.73E-11	1.01
	rs116210165	4	11718341	A	G	<i>RP11-281P23.2</i>	Intron / non-coding	1.24	2.74E-10	
RV-A LRI in first 3 years of life	rs140964702	12	61700025	C	T	-	Intergenic	1.21	7.25E-10	
	rs148332929	4	63737911	G	A	-	Intergenic	0.88	2.03E-09	
wLRI in first 3 years of life	rs117464855	15	93886876	T	A	<i>RP11-266O8.1</i>	Intron / non-coding	0.94	3.03E-09	
	rs28513694	4	70779093	T	C	<i>CSN1S1</i>	Intergenic	0.42	3.93E-09	
RV-C LRI in first 3 years of life	rs619999319	11	46456444	G	A	<i>AMBRA1 / DGKZ / LRP4</i>	Non-coding / missense	0.96	2.27E-08	
	rs185634998	9	28637676	T	C	<i>LINGO2</i>	Intron	1.43	1.45E-11	1.00
HDM IgE in first 5 years of life	rs76597228	17	8173771	C	T	<i>PEAS / SLC25A35 / RANGRF</i>	3' UTR / Intron / downstream	1.46	6.23E-10	
	rs117244162	15	59733543	G	A	<i>FAM81A</i>	Intron / non-coding / regulatory region	1.34	6.85E-09	
RV-A LRI in first 3 years of life	rs7727557	5	90616224	C	T	-	intergenic	0.70	8.70E-09	
	rs144177163	6	161164290	C	A	<i>PLG</i>	Intron / non-coding	0.81	1.33E-10	1.00
RV-C LRI in first 3 years of life	rs189442362	11	20633746	A	G	<i>SLC6A5</i>	Intron / NMD transcript / regulatory region	1.00	6.54E-09	
	rs142116093	3	124897746	A	G	<i>SLC12A8</i>	Intron / NMD transcript	0.81	7.97E-08	
HDM IgE in first 5 years of life	rs132651	22	36555365	C	A	<i>APOL3</i>	Intron / NMD transcript / non-coding	0.27	1.25E-07	
	rs138414325	13	41759533	C	T	<i>KBTBD7</i>	Downstream	0.85	5.13E-10	1.00
RV-A LRI in first 3 years of life	rs55642493	5	172311369	T	G	<i>ERGIC1</i>	Intron / NMD transcript / non-coding	0.67	8.45E-09	
	rs1782888	21	14836636	A	G	<i>GTF2IP2</i>	Intron / non-coding	0.70	2.72E-08	
RV-C LRI in first 3 years of life	rs73833422	4	101516584	A	T	<i>EMCN</i>	Intron / non-coding	0.66	6.63E-08	
	rs182267581	4	95232067	C	T	<i>HPGDS / SMARCAD1</i>	Intron / non-coding	1.82	3.83E-08	1.03

**TABLE D.4: Selected significant and suggestive SNPs for longitudinal genome-wide association scans for early-life microbiome-related traits in CAS and COAST, with repeatABEL.**

Eff. alle. = Effect allele; Prop. = Proportion (of all nasopharyngeal samples) Ref. alle. = Reference allele; Rel. abund. = Relative abundance (averaged per nasopharyngeal sample). All phenotypes reported relate to nasopharyngeal microbiome in the first two years of life.

Phenotype	SNP	CHR	Position	Eff. alle.	Ref. alle.	Closest Gene	Effect of variant	Beta	P-value	Lambda
Prop. <i>Altiococcus</i> .dd23 MPG	rs716680	4	157931688	G	A	<i>PDGFC/GLRB</i>	Intergenic	0.24	4.2E-07	1.03
	rs117437601	12	68454188	G	T	<i>IFNG-AS1</i>	Intron / non-coding	0.48	4.6E-07	
	rs58656950	22	47679506	G	A	<i>TBC1D22A</i>	Intergenic	0.32	8.0E-07	
	rs849258	2	206285824	T	C	<i>PAR3B</i>	Intron / non-coding	0.37	8.7E-07	
Prop. <i>Staphylococcus</i> .29eb MPG	rs9260059	6	29908440	G	A	<i>HLA-A/HCG4P5</i>	Downstream / upstream / missense	0.26	8.8E-08	1.00
	rs1693667	10	126865930	C	T	<i>CTBP2</i>	Intergenic	0.41	2.8E-07	
	rs145119122	16	77039776	C	A	<i>MON1B</i>	Intergenic	0.53	2.8E-07	
Rel. abund. <i>Streptococcus</i> .4060	rs187555021	10	5884866	T	G	<i>GD12</i>	Upstream	0.27	5.7E-09	1.01
	rs143167866	8	41216114	C	T	<i>SFRP1/GOLGA7</i>	Intergenic	0.20	2.2E-07	
	rs11709403	3	780175	G	A	<i>AC090044.1</i>	Intron / non-coding	0.07	3.0E-07	
	rs115170333	6	32552322	G	C	<i>HLA-DRB1</i>	Intron / regulatory region	0.27	3.8E-08	
Rel. abund. <i>Staphylococcus</i> .29eb	rs78089526	2	121309824	G	A	<i>AC073257.1</i>	Downstream / regulatory region	0.23	1.4E-08	1.01
	rs116643345	1	109502867	C	T	<i>CLCCI/AKNAD1</i>	Intron	0.25	1.5E-08	
	rs76170951	7	106915121	G	A	<i>COG5</i>	Intron	0.21	3.9E-08	
	rs143862608	7	102421956	C	T	<i>FAM185A</i>	Intron / non-coding / NMD transcript	0.25	8.1E-07	1.00
Rel. abund. <i>Moraxella</i> .d253	rs34074584	16	49299870	A	C	<i>CBLN1</i>	Intergenic	0.22	2.2E-09	1.02
	rs7014795	8	2368989	T	A	<i>ACT33633.2</i>	Intron / non-coding	0.12	2.8E-09	
Rel. abund. <i>Haemophilus</i> .f579	rs187011751	6	82291152	T	A	<i>FAM46A</i>	Intron	0.21	4.2E-09	
	rs182433605	11	14597052	G	A	<i>PSMA1</i>	Intron	0.21	1.5E-08	



**Minerva Access is the Institutional Repository of The University of Melbourne**

**Author/s:**

Tang, Howard Ho Fung

**Title:**

A systems approach to understanding allergy, asthma and childhood wheeze

**Date:**

2019

**Persistent Link:**

<http://hdl.handle.net/11343/223952>

**File Description:**

Complete thesis

**Terms and Conditions:**

Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.