

THE UNIVERSITY OF CHICAGO

THE ROLE OF RARE VARIATION AND HOST-MICROBE INTERACTIONS
IN COMPLEX TRAIT GENETICS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF HUMAN GENETICS

BY

CATHERINE IGARTUA

CHICAGO, ILLINOIS

JUNE 2017

To my mom

TABLE OF CONTENTS

LIST OF FIGURES	vi
LIST OF TABLES.....	viii
ACKNOWLEDGEMENTS	x
ABSTRACT.....	xii
CHAPTER 1 INTRODUCTION	1
1.1 MOTIVATION FOR STUDYING RARE VARIATION IN COMPLEX TRAITS	3
1.2 MOLECULAR EVOLUTION THEORY PREDICTS THAT DELETERIOUS VARIANTS ARE RARE	3
1.3 POPULATION GENETICS DATA SHOWS EVIDENCE OF PURIFYING SELECTION	4
1.4 MANY RARE VARIANTS OF LARGE EFFECT CAUSE RARE MONOGENIC FAMILIAL DISORDERS.....	4
1.5 CONSIDERATIONS FOR STUDIES OF RARE VARIATION.....	5
1.5.1 CHALLENGES OF RARE VARIANT STUDIES.....	5
1.5.2 VARIANT ANNOTATION AND SELECTION CRITERIA	6
1.5.3 VARIANT AGGREGATION METHODS CAN LEVERAGE POWER IN RARE VARIANT ASSOCIATION STUDIES.....	6
1.6 USE OF POPULATION ISOLATES FOR COMPLEX TRAIT MAPPING	7
1.6.1 THE HUTTERITE FOUNDER POPULATION	8
1.7 THE MICROBIOME AND COMPLEX DISEASE: AN INTERNAL HUMAN ENVIRONMENT.....	10
1.8 DISSERTATION OVERVIEW	11
CHAPTER 2 ETHNIC-SPECIFIC ASSOCIATIONS OF RARE AND LOW-FREQUENCY DNA SEQUENCE VARIANTS WITH ASTHMA.....	12
2.1 ABSTRACT	12
2.2 INTRODUCTION	13
2.3 RESULTS	14
2.3.1 SINGLE VARIANT ASSOCIATION ANALYSES.....	14
2.3.2 CONDITIONAL ANALYSES AT THE 17Q12-21 LOCUS	18
2.3.3 POWER ANALYSES FOR LOW-FREQUENCY VARIANTS.....	20
2.3.4 GENE-BASED ANALYSES.....	20
2.3.5 ASSESSING COVERAGE OF EXONIC VARIATION ON THE EXOME ARRAY.....	23
2.4 MATERIALS AND METHODS	25
2.4.1 STUDY SUBJECTS	25
2.4.2 EXOME ARRAY	26
2.4.3 GENOTYPING	26
2.4.4 SNP QUALITY CONTROL	27
2.4.5 SAMPLE QUALITY CONTROL.....	27

2.4.6	SINGLE VARIANT ANALYSES	28
2.4.7	POWER ANALYSES FOR SINGLE VARIANT ANALYSES.....	29
2.4.8	GENE-BASED ANALYSES.....	29
2.4.9	ASSESSMENT OF FUNCTIONAL VARIANTS ON THE ILLUMINA EXOME ARRAY	30
2.5	DISCUSSION.....	30
2.6	ACKNOWLEDGEMENTS	34
2.7	APPENDIX A: SUPPLEMENTARY DATA	35
CHAPTER 3 RARE VARIANTS ASSOCIATED WITH SERUM BLOOD LIPID TRAITS IN A FOUNDER		
POPULATION		47
3.1	ABSTRACT	47
3.2	INTRODUCTION	48
3.3	RESULTS	50
3.3.1	STUDY OVERVIEW	50
3.3.2	APPLICATION OF FGWAS TO BLOOD LIPID TRAITS	54
3.3.3	CONDITIONAL ANALYSES AND FOLLOW UP WITH GENE EXPRESSION IN LCLs	66
3.4	DISCUSSION.....	71
3.5	MATERIALS AND METHODS	74
3.5.1	STUDY SAMPLE.....	74
3.5.2	GENOTYPING	74
3.5.3	SINGLE VARIANT ANALYSES	75
3.5.4	VARIANT ANNOTATION	75
3.5.5	FGWAS	76
3.6	APPENDIX B: SUPPLEMENTARY TABLES AND FIGURES	77
CHAPTER 4 HOST GENETIC VARIATION IN MUCOSAL IMMUNITY PATHWAYS INFLUENCES THE		
UPPER AIRWAY MICROBIOME.....		96
4.1	ABSTRACT	96
4.2	INTRODUCTION	97
4.3	RESULTS	99
4.3.1	NASAL MICROBIOME COMPOSITION	99
4.3.2	NASAL MICROBIOME DIVERSITY	101
4.3.3	CORRELATION BETWEEN HOST GENETIC SIMILARITY AND MICROBIOME STRUCTURE.....	104
4.3.4	GENOME-WIDE ASSOCIATION STUDIES OF RELATIVE ABUNDANCE	105
4.3.5	MBQTL ASSOCIATIONS WITH MULTIPLE BACTERIA	117
4.3.6	PATHWAY ANALYSES OF GENES NEAR MBQTLs.....	119
4.3.7	COMPARISON OF GENES NEAR MBQTLs IN THE HUTTERITES TO PUBLISHED MICROBIOME QTL STUDIES	120
4.4	MATERIALS AND METHODS	122
4.4.1	SAMPLE COLLECTION	122
4.4.2	SAMPLE DNA EXTRACTION LIBRARY PREPARATION AND SEQUENCING	123

4.4.3	SEQUENCING AND TAXONOMIC CLASSIFICATION.....	123
4.4.4	DATA PROCESSING	124
4.4.5	GENOTYPE DATA.....	126
4.4.6	DIVERSITY METRICS	126
4.4.7	KINSHIP ASSOCIATIONS TO BETA DIVERSITY.....	127
4.4.8	CO-OCCURRENCE NETWORK ANALYSES	127
4.4.9	INGENUITY PATHWAY ANALYSIS (IPA) OF PROTEIN-PROTEIN INTERACTION NETWORKS	127
4.5	DISCUSSION.....	128
4.6	ACKNOWLEDGEMENTS	133
4.7	APPENDIX C: SUPPLEMENTARY DATA	133
CHAPTER 5	CONCLUSIONS	137
5.1	SUMMARY AND SIGNIFICANCE.....	137
5.2	FUTURE DIRECTIONS	139
5.2.1	FUNCTIONAL STUDIES OF RARE NON-CODING VARIATION	139
5.2.2	LINKING MICROBIAL STRUCTURE AND HOST GENETICS TO HEALTH AND DISEASE	141
REFERENCES		143
SUPPLEMENTARY FILE 4.1: RELATIVE ABUNDANCE (AVAILABLE ONLINE)		133
SUPPLEMENTARY FILE 4.2: QTL MAPPING RESULTS OF NASAL MICROBIOME RELATIVE ABUNDANCE (AVAILABLE ONLINE)		135
SUPPLEMENTARY FILE 4.3: NASAL VESTIBULE AND NASOPHARYNX SNP HERITABILITY ESTIMATES (AVAILABLE ONLINE)		136

LIST OF FIGURES

FIGURE 1.1: PRINCIPAL COMPONENT ANALYSIS (PCA) FOR HUTTERITE INDIVIDUALS IN OUR STUDIES	9
FIGURE 3.1: ALLELE FREQUENCIES OF VARIANTS DISCOVERED IN THE HUTTERITES	51
FIGURE 3.2: JOINT FGWAS MODELS FOR BLOOD LIPID TRAITS.....	56
FIGURE 3.3: RARE VARIANTS ON CHROMOSOME 19 ARE ASSOCIATED WITH LDL-C.....	57
FIGURE 3.4: RARE VARIANTS ON CHROMOSOME 11 ARE ASSOCIATED WITH TG LEVELS.....	60
FIGURE 3.5: RARE VARIANTS ON CHROMOSOMES 11 AND 19 ARE ASSOCIATED WITH HDL-C	63
FIGURE 3.6: LDL-C CONDITIONAL ANALYSIS ON TOP RARE VARIANT (rs557778817)	69
FIGURE 4.1: TAXONOMIC COMPOSITION OF BACTERIAL COMMUNITIES IN THE NASAL VESTIBULE AND THE NASOPHARYNX, SAMPLED IN SUMMER AND IN WINTER.....	100
FIGURE 4.2: DIVERSITY	102
FIGURE 4.3: HEAT SCATTERPLOTS OF EUCLIDEAN DISTANCE (BETA DIVERSITY) BY KINSHIP COEFFICIENT	105
FIGURE 4.4: ASSOCIATIONS WITH THE RA OF <i>DERMACOCUS</i> IN THE NASAL VESTIBULE IN THE SUMMER	112
FIGURE 4.5: SNP HERITABILITY ESTIMATES FOR 90 BACTERIA TESTED IN THE NASAL VESTIBULE AND NASOPHARYNX COMBINED SAMPLES.....	115
FIGURE 4.6: FIVE GENUS LEVEL BACTERIA ASSOCIATED WITH rs1543603 ARE HUBS IN A CO-OCCURRENCE MODULE OF 12 PROTEOBACTERIA AND ONE BACTEROIDETES	118
FIGURE 4.7: INGENUITY PATHWAY ANALYSIS (IPA) INTERACTION NETWORKS. NETWORKS SHOW GENES NEAR NASAL MBQTLs ARE ENRICHED FOR MUCOSAL IMMUNITY PATHWAYS	120
SUPPLEMENTARY FIGURE 2.1: QQPLOTS OF SINGLE VARIANT META-ANALYSES OF LOW FREQUENCY, FUNCTIONAL VARIANTS AND ASTHMA RISK	36
SUPPLEMENTARY FIGURE 2.2: POWER ANALYSES FOR LOW FREQUENCY SINGLE VARIANT META-ANALYSES.....	37
SUPPLEMENTARY FIGURE 2.3: GENE-BASED META-ANALYSES QQPLOTS FOR FUNCTIONAL VARIANTS DEFINED AS MISSENSE, NONSENSE AND SPLICING AND EQUAL VARIANT WEIGHTS	38
SUPPLEMENTARY FIGURE 2.4: GENE-BASED META-ANALYSES QQPLOTS FOR FUNCTIONAL VARIANTS DEFINED AS POLYPHEN-2 PROBABLY DAMAGING VARIANTS AND EQUAL VARIANT WEIGHTS	39

SUPPLEMENTARY FIGURE 2.5: GENE-BASED META-ANALYSES QQPLOTS FOR FUNCTIONAL VARIANTS DEFINED AS MISSENSE, NONSENSE AND SPLICING AND RARE VARIANT WEIGHTS	40
SUPPLEMENTARY FIGURE 2.6: GENE-BASED META-ANALYSES QQPLOTS FOR FUNCTIONAL VARIANTS DEFINED AS POLYPHEN-2 PROBABLY DAMAGING VARIANTS AND RARE VARIANT WEIGHTS.....	41
SUPPLEMENTARY FIGURE 2.7: PCA OF EUROPEAN AMERICAN STUDIES PROJECTED ONTO HAPMAP SUBJECTS	44
SUPPLEMENTARY FIGURE 2.8: PCA OF AFRICAN ANCESTRY STUDIES PROJECTED ONTO HAPMAP SUBJECTS	45
SUPPLEMENTARY FIGURE 2.9: PCA OF LATINO STUDIES PROJECTED ONTO HAPMAP SUBJECTS	46
SUPPLEMENTARY FIGURE 3.1: INTEGRATION OF FUNCTIONAL ANNOTATIONS TO GWAS OF BLOOD LIPID TRAITS FOR REVs	77
SUPPLEMENTARY FIGURE 3.2: QUANTILE-QUANTILE PLOTS FOR REV GWAS	78
SUPPLEMENTARY FIGURE 3.3: MANHATTAN PLOTS FOR REV GWAS.....	79
SUPPLEMENTARY FIGURE 3.4: RARE VARIANTS ON CHROMOSOMES 3 AND 1 ARE ASSOCIATED WITH INCREASED LDL-C	93
SUPPLEMENTARY FIGURE 4.1: NASAL MICROBIOME ALPHA DIVERSITY.....	134
SUPPLEMENTARY FIGURE 4.2. CORRELATION OF ALPHA DIVERSITY WITH AGE IN THE NASOPHARYNX IN THE SUMMER	135
SUPPLEMENTARY FIGURE 4.3. BOXPLOTS OF BETA DIVERSITY GROUPED BY NASAL SITE AND BETWEEN NASAL SITES.	135
SUPPLEMENTARY FIGURE 4.4: LOCUS AND GENOTYPE PLOTS FOR SELECTED MBQTLs.....	136

LIST OF TABLES

TABLE 2.1: SAMPLE COMPOSITION	16
TABLE 2.2: RESULTS OF META-ANALYSES OF SINGLE SNP ASSOCIATION STUDIES OF LOW-FREQUENCY FUNCTIONAL VARIANTS AND ASTHMA RISK WITH ASSOCIATION P VALUES $<10^{-4}$	17
TABLE 2.3: CONDITIONAL ANALYSIS RESULTS FOR rs12450091 (c.A365G, p.E122G) IN GSDMB AFTER CONDITIONING ON 17Q12-21 GENOTYPES AT rs7216389, rs2305480, AND rs907092	19
TABLE 2.4: GENES SIGNIFICANTLY ASSOCIATED WITH ASTHMA.....	22
TABLE 2.5: FUNCTIONAL VARIANT DISCOVERY IN WHOLE GENOME SEQUENCES OF 278 ASTHMATICS	24
TABLE 3.1: SAMPLE COMPOSITION AND CLINICAL DESCRIPTIONS	53
TABLE 3.2: GENOME-WIDE SIGNIFICANT RESULTS FOR THE REV GWAS OF BLOOD LIPID TRAITS.....	54
TABLE 3.3: SUMMARY RESULTS FOR CONDITIONAL ANALYSES	66
TABLE 4.1: SAMPLE COMPOSITION	101
TABLE 4.2: QTL MAPPING RESULTS OF NASAL MICROBIOME RELATIVE ABUNDANCE	107
SUPPLEMENTARY TABLE 2.1: NUMBER OF VARIANTS ON THE EXOME ARRAY THAT PASSED QC AND ARE VARIABLE IN 11,225 STUDY SUBJECTS	35
SUPPLEMENTARY TABLE 2.2: VARIANTS INCLUDED IN SINGLE VARIANT ANALYSES	35
SUPPLEMENTARY TABLE 2.3: REPRESENTATION OF GENES ON THE EXOME ARRAY WITH FUNCTIONAL VARIANTS IN THE WHOLE GENOME SEQUENCES OF 278 ASTHMATIC INDIVIDUALS	42
SUPPLEMENTARY TABLE 2.4: EXOME ARRAY VARIANTS REMOVED FROM ANALYSIS	42
SUPPLEMENTARY TABLE 2.5: TOTAL SUBJECT EXCLUSIONS.....	42
SUPPLEMENTARY TABLE 3.1: RARE VARIANTS IN EUROPEANS FGWAS COMBINED MODEL FOR LDL-C.....	79
SUPPLEMENTARY TABLE 3.2: RARE VARIANTS IN EUROPEANS FGWAS COMBINED MODEL FOR TG LEVELS.....	80
SUPPLEMENTARY TABLE 3.3: RARE VARIANTS IN EUROPEANS FGWAS COMBINED MODEL FOR HDL-C.....	82
SUPPLEMENTARY TABLE 3.4: RARE VARIANTS IN EUROPEANS FGWAS COMBINED MODEL FOR TOTAL CHOLESTEROL	82
SUPPLEMENTARY TABLE 3.5: SUMMARY OF ASSOCIATION RESULTS FOR RARE IN EUROPEAN VARIANTS.....	83

SUPPLEMENTARY TABLE 3.6: GWAS CATALOG VARIANTS FOR REGION ON CHROMOSOMES 3 AND 19 ASSOCIATED WITH LDL-C	88
SUPPLEMENTARY TABLE 3.7: GWAS CATALOG VARIANTS FOR REGION ON CHROMOSOME 11 ASSOCIATED WITH TRIGLYCERIDE LEVELS AND WITH HDL-C.....	89
SUPPLEMENTARY TABLE 3.8: GWAS CATALOG VARIANTS FOR REGION ON CHROMOSOME 16 ASSOCIATED WITH HDL-C	91

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Carole Ober. Carole's mentorship has not only allowed me to expand scientifically, but has also taught me a level of professionalism and ambition for personal growth that I will take forward in my career. Carole's encouragement, guidance and knowledge in the field were key motivators through my Ph.D. I couldn't have asked for a better advisor and I am very thankful for her mentorship.

I'd like to also thank my thesis committee members Dan Nicolae, Yoav Gilad, Vinny Lynch and John Novembre for their continuous support and insightful advice on my research projects. I am particularly grateful for the amount of time that Dan Nicolae has devoted to me throughout my graduate career, especially with regards to the design and statistical interpretation of my rare variant studies. In addition, I would like to thank Jay Pinto for his mentorship and insight in the airway microbiome study.

Most importantly, I would like to thank all present and past members of the Ober lab with whom I have interacted throughout these years. The Ober lab community has created an open and warm collaborative environment that has allowed me to succeed the past five years. I would like to thank Jessica Chong and Rachel Myers for their guidance during the beginning stage of my graduate career and Emily Davenport for her support in microbiome analysis. I'd also like to thank William Wentworth-Shields, Oren Livne and Tiratat Patana-anake for their computational support and Rebecca Anderson and Catherine Stanhope for their help with the Hutterite phenotype datasets. I would like to give a special thanks to Michelle Stein and Sahar Mozaffari for their continuous feedback in all my projects and friendship throughout these years.

I would like to acknowledge the Hutterite community for their continuous involvement in our studies. Without them, my research would have not been possible. The Hutterites have welcomed me into their homes multiple times through these years. I feel very fortunate to have been able to visit their communities and have made personal connections with such generous, positive people.

I would like to specially thank Alex Bobe, Jodie Laughlin, Katrina Webb and Kenneth Barr. Their friendship has supported me through this whole process and I'm particularly grateful for all their encouragement at times when this journey seemed impossible. I am especially grateful for all the scientific discussions I've had with Kenneth. His views have influenced the interpretation of my data and helped me gain confidence in myself as a scientist.

Last, but not least, I would like to thank my family who has always encouraged me to follow my dreams and has pushed and sacrificed for me to pursue my education. I always loved math as a kid, and my father would find algebra problems to keep me interested and ahead of my classroom. My mother's work-life balance has always been an inspiration and has taught me to work hard and never give up on my goals. My sisters and Jenny and Lula have always inspired to achieve personal growth and are always there to support me and remind me not to take life too seriously.

ABSTRACT

GWAS have been extensively used to identify genetic factors contributing to complex traits, but the variants identified thus far explain a small portion of the heritabilities of these traits. In my dissertation, I aimed to understand the role of rare variation in complex traits and identify host genotype influences on the microbiome, two understudied components of complex trait genetics. First, I searched for contributions from rare and low frequency putatively functional variants on asthma pathogenesis in a large sample of ethnically diverse populations (the EVE Consortium of Asthma Genetics). The few associations detected in this study were limited to single ethnic groups, suggesting that rare or low frequency variants that influence asthma are likely to be ethnic specific. Our results also indicate that rare and low frequency loss-of-function variants nearby coding regions are unlikely to explain a significant portion of the heritability for asthma. Second, I studied the effects of rare variants on blood lipid traits in the Hutterites, a founder population of European descent. I tested the genome-wide effects of variants that are absent or rare variants in Europeans but have drifted to higher frequencies in the Hutterites. I identified four novel variants associated with LDL-C or HDL-C, two of which were also at novel loci, and in addition replicated a known splicing variant in *APOC3*. Our results suggested that rare non-coding risk variants are likely to mediate their effects through gene expression. Lastly, I studied the upper airway microbiome, which has been associated with respiratory diseases, in the Hutterites. I identified airway bacteria whose abundance is influenced by host genotype and determined that host expression of innate and mucosal immune pathway genes plays a central role in structuring the airway microbiome (three results tables for the microbiome study are available online in supplementary files).

CHAPTER 1

INTRODUCTION

A major goal of medical genetics research is to identify variation that allows prediction of an individual's risk for disease and to uncover genes and molecular pathways that may serve as novel therapeutic targets. Mendelian traits are strongly influenced by highly penetrant mutations in a single gene, known as monogenic disorders, and are generally detected by familial patterns of inheritance. Traditional gene mapping approaches such as linkage analysis^{1,2}, homozygosity mapping³ and copy number variation(CNV) analysis⁴, and more recently exome sequencing⁵, have led to great success in identifying Mendelian disease genes. Complex diseases such as asthma, type II diabetes and cardiovascular disease have multifactorial etiologies that involve an interplay between genetic and environmental factors, and do not follow strict Mendelian patterns of inheritance⁶. Over the past decade, genome-wide association studies (GWAS) and sequencing studies have been extensively used to identify genetic factors contributing to common, complex traits. The many successes of GWAS of complex phenotypes for gene discovery reflects major advances in high-throughput technologies that have allowed simultaneous surveillance of millions of variants (single nucleotide polymorphisms [SNPs], and small insertions and deletions [indels]) in large numbers of individuals. As of 2/20/2017, nearly three thousand published GWAS reported the discovery of more than thirty thousand robust associations with complex phenotypes (<http://www.ebi.ac.uk/gwas/home>).

Despite the successes of GWAS, which largely interrogate common SNPs, the associated variants discovered thus far have very small effect sizes and only explain a small portion (at most half), of the heritabilities for complex traits⁷. This observation has been referred to as the “missing heritability”⁸. For example, blood lipid traits have heritability estimates suggesting that 40-70% of the phenotypic variance is attributed to genetic variance, but the discovered variants for these traits only account for only about 14% of the phenotypic variance⁹. To account for this missing heritability, large international efforts have been directed towards combining GWAS studies for meta-analyses, but these have still failed to account for much more of the heritability of complex traits. Although it is likely that many heritabilities are over-estimated⁸, another explanations for this could be that the genetic variants not interrogated by GWAS might be important for disease risk.

GWAS are predicated on ‘*the common disease, common variant*’ (CDCV) hypothesis¹⁰. This hypothesis states that the genetic architecture of common diseases is one of many contributing loci with low to intermediate effect sizes, and that associated alleles are present in the population at common frequencies. While GWAS findings support the CDCV hypothesis to a large extent, as the many variants associated with common diseases have small additive effects on disease risk (odds ratios typically around 1.2), this approach only interrogates common variants. Therefore, alternate approaches are needed to assess the potential effects of rare to low frequency variants with potentially larger effect sizes on disease risk.

1.1 Motivation for studying rare variation in complex traits

Rare variants with allele frequencies less than 1% in the population vastly outnumber common variants in the human genome⁶, but their effects on disease risk remain largely unknown. Rare variants are generally not tagged by markers present on genotyping arrays and therefore are not well interrogated by GWAS. In contrast to the CDCV hypothesis, the '*common disease, rare variant*' hypothesis (CDRV) predicts that a significant portions of the genetic susceptibility to common diseases are due to the collective effects of low frequency (1-5%) and rare (<1%) variants at many loci, each with moderate to high increase in relative risk¹¹. This hypothesis was motivated by the observations of highly penetrant rare variants that underlie Mendelian traits.

1.2 Molecular evolution theory predicts that deleterious variants are rare

The existence of rare alleles in the genome is a balance between new mutations arising in the population, recombination, random genetic drift and selection forces preventing deleterious mutations from drifting to high frequencies. The human mutation rate is sufficiently large that purifying selection cannot feasibly eliminate all deleterious mutations, particularly those with greater than zero fitness. Deleterious mutations that impact fitness will tend to stay at low frequencies. However, some disease-associated alleles can reach appreciable frequencies in the population if their effect on disease does not affect fitness, for example when the disease has a late onset or in carriers of recessive mutations¹².

1.3 Population genetics data shows evidence of purifying selection

Large-scale surveys of genome-wide genetic variation have provided empirical support for the demographic and evolutionary forces shaping human genetic variation. These surveys show a strong skew towards excesses rare variation in the allele frequency spectrum of protein coding alleles and reflect evidence for purifying selection against deleterious alleles^{13,14}. In fact, this notable excess of rare variation can be explained by signatures of rapid population growth and suggests many rare mutations are of recent age¹⁵⁻¹⁷. In addition, exome and whole genome sequencing studies in large numbers of individuals have identified i) a depletion of variation in coding compared to noncoding regions¹⁸, ii) an overabundance of common synonymous variants compared to nonsynonymous variants¹⁹, and iii) an enrichment of rare and low-frequency deleterious, nonsynonymous mutations over synonymous mutations^{14,20}. These observations suggest that rare variants may explain some of the missing heritability and disease risk for common diseases, and directly studying rare variants will provide a complementary approach to GWAS.

1.4 Many rare variants of large effect cause rare monogenic familial disorders

For many common diseases, there are rare familial monogenic forms or monogenic disorders with overlapping phenotypes. Causal variants identified for these monogenic traits are typically highly penetrant mutations segregating in families with a Mendelian pattern of inheritance. While these mutations are often rare within the population, or even private to the pedigree in which they are segregating, they identify novel causal genes and provide fundamental insight into the disease mechanisms and biology, which may be shared with the common forms of these diseases. Some well-known examples include the discovery of mutations in the *LDLR*, *APOB* and *PCSK9* genes in

familial hypercholesterimia²¹, mutations in *HNF1A*, *GCK*, *HNF4A* in families with Maturity-onset Diabetes of the Young (MODY²²), and the *BRCA1* and *BRCA2* mutations in early onset breast cancer²³.

1.5 Considerations for studies of rare variation

1.5.1 *Challenges of rare variant studies*

Because of the vast number of rare mutations present in the genome, a lot of which are likely to be recent and thus ancestry or population specific, sequencing large population samples is required to detect their presence in the population. Evaluating the effects of rare variants on complex disease requires even larger sample sizes because power depends heavily on allele frequency. Rare variant association studies seek to gain power from increased effects sizes in deleterious variants, however, sample sizes for rare variant studies have generally not reached sufficient numbers to detect their independent effects on disease risk²⁴. Instead, studies have focused on surveying the independent effects of low frequency variants, which are also unexplored and likely to play a role on disease risk, or using variant aggregation methods to identify regions (usually a gene) with accumulated rare variation that may be influencing a trait. While these approaches have had some success discovering new candidate genes and variants with large effect on complex disease²⁵⁻²⁷, early findings suggest most rare variants influencing complex phenotypes have modest to small effect sizes. However, this result should not undervalue rare variant studies; these studies identify novel genes and reveal fundamental insights into disease mechanisms.

1.5.2 Variant annotation and selection criteria

Rare variant association studies suffer from a significant multiple testing burden. One approach to minimize this burden is to filter out likely neutral variation. In fact, variants can be prioritized based on their functional annotations (e.g. missense, nonsense, splicing), predicted functional scores (e.g. PolyPhen2²⁸, SIFT²⁹), conservation scores (e.g. GERP³⁰, PhastCons³¹), or aggregates of multiple prediction scores (e.g., CADD³²). This approach can increase statistical power by reducing multiple testing burden, but the hypothesis tested and interpretability of the results rely heavily on the accuracy and assumptions of each annotation.

1.5.3 Variant aggregation methods can leverage power in rare variant association studies

Instead of testing each variant's association with the phenotype, variant collapsing methods combine the cumulative effects of multiple rare variants present within a gene or functional unit (reviewed in ³³⁻³⁵). These methods can be classified into three classes: i) burden tests (e.g. CMC³⁶), ii) variance component tests (e.g. C-alpha³⁷, SKAT³⁸), iii) combined burden and variance component tests (e.g. SKAT-O³⁹, EMMPAT⁴⁰) and iv) p-value combination methods (e.g. ADA⁴¹, Sigma P⁴²). Each of these methods relies on assumptions about the underlying disease architecture and can result in power differences that are dependent on the true (but unknown) disease model.

Burden tests summarize genotypes by counting the number of alleles across a set of variants and then testing the association between this summed value and the trait. Burden tests assume that the directionality and magnitude of effect is equal across each variant, providing the

most power when the proportion of causal variants in a region is large and all have effect sizes in the same direction (i.e., all deleterious or protective). Several adaptations of the burden test have been designed to increase power by up-weighting variants based on MAF (e.g. Madsen and Browning⁴³, beta density³⁸).

The variance component test evaluates the distribution of the aggregated test statistics for a set of variants. This test also allows for variant weights and is more powerful than a burden test when there is an increase number of neutral variants or when effects in opposite directions exist within a region. Because a mixture of both scenarios is likely to underlie complex traits, tests that combine and optimize the power of burden and variance component tests are typically more preferable.

An important disadvantage of variant aggregation tests is that it can be challenging to pinpoint the causal variants from all variants in the region or gene, and to validate the results in follow up translational studies. Summarizing over individual variant marginal statistics and backwards elimination procedures (e.g., remove a variant and re-test the new variant set) can help narrow down the individual effects and identify causal variants.

1.6 Use of population isolates for complex trait mapping

Founder populations provide considerable advantages for studying the clinical effects of rare variants on complex disease^{44,45}. Their unique demographic histories and founder effect, a recent bottleneck and rapid population expansion event, allowed genetic drift in the early generations after their founding to have profound effects on allele frequencies. Because both disease and neutral alleles are subject to random genetic drift, some rare disease alleles will be lost and some

will drift to higher frequencies. Alleles with phenotype effects that are present at sufficiently common frequencies in the founder population will be particularly informative and well suited for association studies of complex traits. Moreover, the small number of founders decreases the number of disease susceptibility alleles ultimately segregating in the population and increases the likelihood for affected individuals in the population to share the same genetic risk variants inherited from a common ancestor⁴⁶.

In addition to reduced genetic complexity, founder populations offer unique advantages that result from their shared environment and culture. Differences in environmental exposures and lifestyle are minimized in founder populations and can facilitate the detection of susceptibility loci that underlie complex disease^{45,46}. In addition, many founder populations keep genealogical records, making it possible to trace back the relationship between affected individuals.

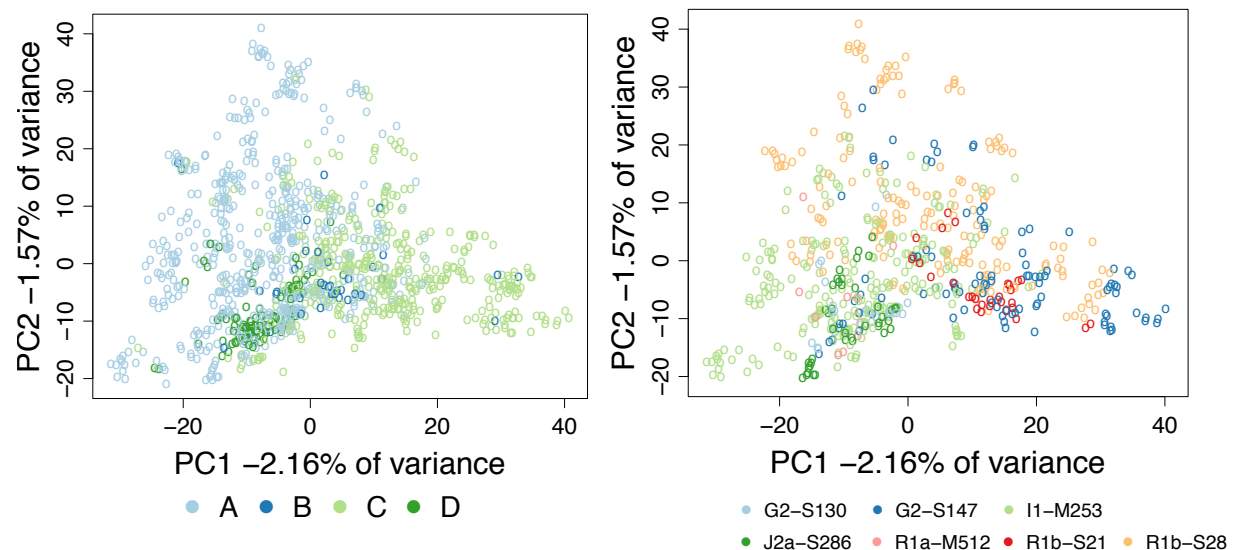
1.6.1 The Hutterite founder population

The Hutterites are an Anabaptist founder population that originated from approximately 90 founders in the Austrian province of Tyrol during the Radical Reformation in the 16th century. After more than 200 years of religious persecution throughout central and eastern Europe, this religious group immigrated to North America in the 1870s and settled in what is today South Dakota⁴⁷. Approximately 400 individuals settled into three communal farms (colonies) and gave rise to the three major Hutterite branches that exist today: the Schmiedeleut (S-leut), the Lehrerleut (L-leut), and the Dariusleut (D-leut). The Hutterites have since expanded to more than 400 colonies throughout the north-central plains states of the United States and the western provinces of Canada, with a population size approximately 42,000 individuals⁴⁸.

A distinct advantage of studying the Hutterite population, besides the fact that it is a founder population, is that their communal farming lifestyle that ensures environmental exposures are relatively similar among individuals. In Hutterite colonies, meals are cooked communally, individuals have similar access to exercise, wealth and education and exposures to infectious diseases are minimal. Moreover, their common language and religion, promotes a cultural and social unity. The Hutterites in our studies are from the S-leut branch and are related to each other in a 13-generation pedigree with 64 founders who were born between the early 1700s and the early 1800s. Our current genetic studies are based on 1,415 individuals from recent generations⁴⁹ (Figure 1.1).

Figure 1.1: Principal component analysis (PCA) for Hutterite individuals in our studies

The PCA on the left contains 1,415 Hutterite individuals. Color is based on the four colony lineages (A-D) representing descent from the four colonies in existence in 1910. The plot on the right contains only male Hutterite individuals and color is based on their Y chromosome haplogroup.



1.7 The microbiome and complex disease: an internal human environment

The human microbiome has been known for a long time to affect human health and disease. Our bodies house trillions of microbes composed of over a thousand species, with highest concentrations present in the gut and with significant numbers in the airways⁵⁰⁻⁵². These microbial communities differ in composition across body sites and depend on physical, chemical and immune environments. Under healthy conditions the relationship with the host is commensal; these microbes aid with metabolism and immunity. However, imbalances in human microbiota have been linked to diseases such as inflammatory bowel disease, diabetes, antibiotic resistance, sinusitis and asthma⁵¹⁻⁵⁵. In some cases, the diversity of the whole community rather than the abundance of a certain microbe is predictive of disease states. Supportive evidence for the role of a rich microbial environment in disease development includes the increased risk of asthma in babies born via cesarean section and for children treated with multiple courses of antibiotics⁴³.

Recent human population⁵⁶⁻⁵⁸ and mouse studies⁵⁹ have revealed microbial taxa whose abundance are partially regulated by host genotype across multiple body sites. Genome-wide association studies are well suited for finding novel associations between host genetics and the microbiome but a major concern for human studies is that environmental factors (such as diet and geographical location) can rapidly influence microbial composition and function. Furthermore, studies of related individuals and even twin pairs are largely confounded by the more similar environments among close relatives, making it impossible to completely disentangle the contributions of genes and environment. Founder populations offer reduced environmental heterogeneity are well positioned to study the genetic influences on the microbiome.

1.8 Dissertation overview

The main goal of my thesis was to study the role of rare variation on complex disease and explore host genetics influences on the upper airway microbiome. In chapter 2, I present the results of a study of the contributions of rare and low-frequency coding variation to asthma risk in the EVE Consortium of Asthma Genetics, a multi-center sample of ethnically diverse populations. The results of this study were published in *Nature Communications* (Ethnic-specific associations of rare and low-frequency DNA sequence variants with asthma. *Nat Commun* 6, 5965 (2015)). In chapter 3, I conduct functional GWAS of lipid traits (LDL cholesterol, HDL cholesterol, triglycerides and total cholesterol) for rare in European variants that have drifted to common frequencies in the Hutterite population. In chapter 4, I describe the results of study a genetic mapping studies of the microbiome in two accessible upper airway sites, the nasopharynx and the nasal vestibule, during two seasons in 144 adult Hutterites. These data were published in *Microbiome* (Host genetic variation in mucosal immunity pathways influences the upper airway microbiome. *Microbiome* 5, 16 (2017)). Finally, in chapter five, I discuss the overall significance and future directions for the work presented in this dissertation.

CHAPTER 2

ETHNIC-SPECIFIC ASSOCIATIONS OF RARE AND LOW-FREQUENCY DNA SEQUENCE VARIANTS WITH ASTHMA

2.1 Abstract¹

Common variants at many loci have been robustly associated with asthma but explain little of the overall genetic risk. Here we investigate the role of rare (<1%) and low-frequency (1-5%) variants using the Illumina HumanExome BeadChip array in 4,794 asthma cases, 4,707 non-asthmatic controls, and 590 case-parent trios representing European Americans, African Americans/African Caribbeans, and Latinos. Our study reveals one low-frequency missense mutation in the *GRASP* gene that is associated with asthma in the Latino sample ($p=4.31 \times 10^{-6}$; OR=1.25; MAF=1.21%) and two genes harboring functional variants that are associated with asthma in a gene-based analysis: *GSDMB* at the 17q12-21 asthma locus in the Latino and combined samples ($p=7.81 \times 10^{-8}$ and 4.08×10^{-8} , respectively) and *MTHFR* in the African ancestry sample ($p=1.72 \times 10^{-6}$). Our results suggest that associations with rare and low-frequency variants are ethnic specific and not likely to explain a significant proportion of the “missing heritability” of asthma.

¹ Citation for chapter: Igartua, C. *et al.* Ethnic-specific associations of rare and low-frequency DNA sequence variants with asthma. *Nat Commun* 6, 5965 (2015)

2.2 Introduction

Asthma is a common, chronic inflammatory disease of the airways with typical onset in childhood. Heritability estimates indicate that approximately half the variation in risk is attributable to genetic factors⁶⁰⁻⁶², yet the common variants identified by genome-wide association studies (GWAS) account for very little of the genetic risk⁶³⁻⁶⁶. This so called “missing heritability” in asthma and other common diseases has been attributed to many potential causes that generally fall into two categories⁶. First, the genetic variants interrogated by GWAS may not capture all relevant variation. In particular, rare variants, which comprise the bulk of genetic variation in the human genome and are predicted to have larger phenotypic effects than common variants, are not represented or tagged by the single nucleotide polymorphisms (SNPs) that are included on the genotyping arrays typically used in GWAS. Second, the statistical approaches used in GWAS may be overly simplistic and not adequately model the genetic architecture of asthma, which is likely polygenic and includes many gene-environment interactions⁶⁷.

In this study, we begin to address the first limitation of GWAS by investigating the role of rare (minor allele frequency [MAF] < 1%) and low-frequency (MAF 1-5%) variants in asthma. We conducted this meta-analysis study in the largest ethnically diverse sample of asthma cases, non-asthmatic controls, and case-parent trios, that include 13 studies^{63,68,69} from eight groups of investigators who compose the EVE Consortium on Asthma Genetics^{63,68}. We conducted analyses separately in each of the 13 study samples and used meta-analyses to first combine results within each of three major ethnic/racial groups (European American, African American/African Caribbean, Latino), and then across all ethnic/racial groups to yield measures of association for

the combined sample. Here we report associations with a low-frequency functional variant and lower risk for asthma in Latinos in a novel gene, *GRASP*, and with functional variants in the *MTHFR* gene in an African ancestry sample. We also report an association in Latinos and in the combined sample with functional variants in *GSDMB*, a gene at the 17q12-21 asthma locus^{63,64,70-72}, which is attributed to a putatively damaging common missense mutation.

2.3 Results

2.3.1 Single variant association analyses

We first evaluated the individual effects of low-frequency putatively functional exonic variants on asthma risk in 11,225 subjects representing European Americans, African Americans, African Caribbeans, Mexican Americans, Mexicans, and Puerto Ricans (Table 2.1). We tested each low-frequency variant, defined as having a MAF between 1% and 5% within each ethnic group. For the meta-analysis in the combined sample, we excluded variants that were either common (MAF $\geq 5\%$) in all three ethnic groups or rare ($<1\%$) in all three ethnic groups (Supplementary Table 2.1 and Supplementary Table 2.2). This definition allows us to include SNPs that may be common in one population but rare in the other(s). The quantile-quantile (QQ) plots for the analyses of low-frequency variants indicated that there was no inflation of test statistics (Supplementary Figure 2.1). One missense mutation (c.C1093G, p.L365V) in the *GRASP* gene was associated with asthma risk in the Latinos after correcting for multiple testing (9,519 tests; GEMMA for Genome-wide Efficient Mixed Model Association (GEMMA)⁷³/ Family Based Association Test (FBAT)⁷⁴ meta-analysis $p=4.31 \times 10^{-6}$; $N=4,554$; $OR=1.25$; $MAF = 0.012$). Six additional low-frequency variants were modestly associated at suggestive levels of significance ($p < 10^{-4}$) in at least one of the three groups

or in the combined sample (Table 2.2). Of these, five missense mutations (one each in *FAM134B*, *NPC1L1*, *IKZF1*, *SLC26A3* and *GSDMB*) showed evidence of association with asthma in the combined sample, two of which (*IKZF1* and *SLC26A3*) were more significant in the African ancestry sample. The low-frequency variants at *FAM134B* (c.G1145C, p.S382T; transcript accession NM_001034850), *NPC1L1* (c.C2920T, p.P974S), and *IKZF1* (c.G1009A, p.G337S; transcript accession NM_006060) were all associated with lower risk for asthma whereas the variants in *SLC26A3* (c.A241G, p.I81V) and *GSDMB* (c.A365G, p.E122G; transcript accession NM_001042471) were associated with increased risk for asthma. An additional nonsense mutation (c.C919T, p.R307X) in *PDE4DIP* was also associated with asthma risk in the African ancestry sample. The latter result is noteworthy for two reasons. First, an exome sequencing study in a European American, non-Hispanic white nuclear family with two asthmatic and two non-asthmatic children also reported a missense mutation in *PDE4DIP* (c.A907C, p.I303L) segregating with asthma¹¹. Second, this gene encodes a protein that interacts with *PDE4D*, a gene that was identified in a GWAS of asthma⁷⁵. Our results further implicate this pathway in asthma pathogenesis.

Table 2.1: Sample Composition

Sample	Case	Control	Male	Female	Total
European American					
ALHS	776	802	859	719	1,578
CAG	186	209	172	223	395
CAMP ^a	376	706	588	494	1,082
CHS	415	640	546	509	1,055
COAST	134	119	150	103	253
Total	1,511 376 ^a	1,770	2,315	2,048	4,363
African Ancestry					
CAG	274	200	170	304	474
GRADD-AA	236	288	199	325	524
GRADD-AC ^b	227	297	250	274	524
SAPPHIRE	553	233	278	508	786
Total	1,290	1,018	897	1,411	2,308
Latino					
CHS	581	775	713	643	1,356
GALA II-MA	568	604	568	604	1,172
GALA II-PR	844	540	731	653	1,384
MCAAS ^a	214	428	337	305	642
Total	1,993 214 ^a	1,919	2,349	2,205	4,554
Combined					
Total	4,794 590 ^a	4,707	5,561	5,664	11,225

ALHS, Agricultural Lung Heath Study; CAG, Chicago Asthma Genetics Study; CAMP, Childhood Asthma Management Program; CHS, The Children's Health Study; COAST, The Childhood Origins of Asthma Study; GALA II, Genes-Environment & Admixture in Latino Asthmatics (-MA, Mexican American; -PR, Puerto Rican); GRAAD, Genomic Research on Asthma in the African Diaspora and Barbados (-AA, African Americans; -AC, African Caribbeans); MCCAS, Mexico City Childhood Asthma Study; SAPPHIRE, The Study of Asthma Phenotypes and Pharmacogenomic Interactions by Race-Ethnicity. Descriptions of the individual study samples have either been published^{63,68} or are described in . ^aCase-parent trios. ^bSamples from 154 families; kinship matrix was used to address family structure as implemented in GEMMA⁷³.

Table 2.2: Results of meta-analyses of single SNP association studies of low-frequency functional variants and asthma risk with association P values $<10^{-4}$

rsID (Chr: position)	Gene	Residue Change	Alleles	GEMMA/FBAT Meta-analysis p				Minor allele frequency			Odds Ratio (95% CI)			
				COMB (11,225)	EA (4,363)	AA (2,308)	LAT (4,554)	EA	AA	LAT	COMB	EA	AA	LAT
rs141783558 (1:144930790)	<i>PDE4DIP</i>	c.C919T, p.R307X	T/C	1.48×10^{-3}	-	9.76×10^{-5}	-	1.22×10^{-4}	0.0313	3.41×10^{-3}	1.12 [1.04-1.21]	-	1.17 [1.08-1.26]	-
rs61733811 (5:16475199)	<i>FAM134B</i>	c.G1145C , p.S382T (NM_001 034850)	G/C	1.30×10^{-5}	-	-	6.56×10^{-3}	7.30×10^{-4}	0.1175	0.0146	0.92 [0.89-0.95]	-	-	0.89 [0.81-0.97]
rs149782907 (7:44561344)	<i>NPC1L1</i>	c.C2920T, p.P974S	T/C	2.66×10^{-5}	-	1.13×10^{-4}	-	7.30×10^{-4}	0.0142	1.28×10^{-3}	0.8 [0.72-0.89]	-	0.8 [0.71-0.89]	-
rs148169768 (7:50467774)	<i>IKZF1</i>	c.G1009A , p.G337S (NM_006 060)	A/G	6.55×10^{-5}	-	1.68×10^{-5}	-	0	0.0109	5.33×10^{-4}	0.77 [0.68-0.87]	-	0.74 [0.65-0.85]	-
rs116793431 (7:107434217)	<i>SLC26A3</i>	c.A241G, p.I81V	G/A	7.89×10^{-5}	-	4.97×10^{-5}	-	4.87×10^{-4}	0.0475	4.15×10^{-3}	1.12 [1.06-1.19]	-	1.14 [1.07-1.21]	-
rs7974426 (12:52408888)	<i>GRASP</i>	c.C1093G , p.L365V	G/C	0.0543	0.99	-	4.31×10^{-6}	0.0337	0.0059	0.0121	1.04 [0.99-1.10]	0.98 [0.91- 1.04]	-	1.25 [1.14-1.38]
rs12450091 (17:38068621)	<i>GSDMB</i>	c.A365G, p.E122G (NM_001 042471)	G/A	5.56×10^{-5}	-	-	-	2.31×10^{-3}	3.11×10^{-3}	0.1945	1.07 [1.03-1.1]	-	-	-

Chr, chromosome; COMB, combined; EA, European American; AA, African Ancestry; LAT, Latino; CI, confidence interval. Positions are from NCBI build 37. For residue changes, the transcript accession number is presented in parenthesis when multiple transcripts exist. Alleles are presented as minor allele/major allele. Ethnic specific meta-analyses were performed for all low-frequency variants (MAF 1-5%; 8,249 in European American, 17,861 in African American, and 9,519 in Latinos); the combined meta-analysis included variants that were neither common ($\geq 5\%$) nor rare ($< 1\%$) in all three ethnic groups (32,681 variants). Odds ratios are calculated for case-control studies only (case-parent trio studies excluded) and presented for the minor allele. Only the association with the *GRASP* variant in Latinos was significant after Bonferroni correction (9,519 variants, GEMMA/FBAT meta-analysis $p=4.31 \times 10^{-6}$, $N=4,554$). Bold P-values and corresponding ORs represent the meta-analysis associations with $p < 10^{-4}$.

The missense mutation in *GSDMB* is also intriguing because this gene is at the well-replicated 17q12-21 asthma locus^{63,64}; this signal is largely attributable to the GALA II Mexican American sample (GEMMA $p=1.1 \times 10^{-4}$; $N=1,172$). This variant showed striking MAF differences between the groups, ranging from 0.19 in the Latinos to 0.0031 in the African Americans and 0.0023 in the European Americans but was included because it was neither rare nor common in all three groups (see Methods).

2.3.2 Conditional analyses at the 17q12-21 locus

To determine whether the association with this variant (c.A365G, p.E122G; rs12450091) in Latinos and the combined sample represents independent risk for asthma or is in linkage disequilibrium (LD) with the common variants at the 17q12-21 locus that have been associated with asthma in previous GWAS^{63,64,76,77}, we performed conditional analyses of rs12450091 in the case-control samples only (case-parent trio studies were excluded and meta p-value recalculated: Latino GEMMA meta-analysis $p=6.89 \times 10^{-5}$, $N=3,912$; Combined GEMMA meta-analysis $p=5.52 \times 10^{-5}$, $N=9,501$). For this analysis, we first identified three common variants (one synonymous, one missense, and one intronic) at this locus that were included on the exome array and associated with asthma in previous reports (G at rs907092^{63,64,76}, G at rs2305480^{64,77}, and T at rs7216389^{64,78,79}). In Latinos, the risk allele at rs12450091 (C) tags 28% of the high risk *GGT* haplotypes (*GGCT* frequency 18.09%, *GGTT* frequency 45.84%) and is not in LD with any of the known expression quantitative trait loci reported at this locus^{70,78-81}. We then repeated the association studies with rs12450091 (c.A365G, p.E122G) in *GSDMB* in three separate analyses, including each of the common variants as a covariate in each analysis to “remove” the effects of

the common variant on the association signal. In each of these three analyses, the evidence for association between rs12450091 and asthma was reduced from $p=6.89 \times 10^{-5}$ to $p \leq 0.022$ in Latinos and from $p=5.52 \times 10^{-5}$ to $p \leq 0.019$ in the combined sample (Table 2.3). These results suggest that while most of the evidence for association with this variant was shared with the associated common variants, the residual signal in conditional analyses indicates that this missense substitution or a variant in strong LD with the missense substitution contributes additional, independent risk to asthma in Latinos.

Table 2.3: Conditional analysis results for rs12450091 (c.A365G, p.E122G) in GSDMB after conditioning on 17q12-21 genotypes at rs7216389, rs2305480, and rs907092

rsID (Chr:Position)	Gene	Residue Change	Alleles	GEMMA meta- analysis p (COMB)	r^2/D' (LAT)	Conditional GEMMA meta- analysis p for rs12450091 (c.A365G, p.E122G)	
						COMB (no covariate $p=5.52 \times 10^{-5}$)	LAT (no covariate $p=6.89 \times 10^{-5}$)
rs907092 (17:37922259)	<i>IKZF3</i>	c.C1314T, p.S438S (NM_012481)	A/G	2.08×10^{-10}	0.102/0.97	0.0190	0.0215
rs2305480 (17:38062196)	<i>GSDMB</i>	c.C892T, p.P298S (NM_001042471)	T/C	1.09×10^{-8}	0.104/0.99	0.0139	0.0160
rs7216389 (17:38069949)	<i>GSDMB</i>	Intronic	C/T	1.23×10^{-7}	0.13/1	0.0131	0.0149

Chr, chromosome; COMB, combined; LAT, Latino. Three 17q12-21 common variants that were included on the array and associated with asthma in previous GWAS (rs907092, rs2305480, rs7216389) were included as covariates in association studies of rs12450091 in the case-control samples only (case-parent trio studies excluded and meta p -value recalculated). Conditional analyses included 3,912 subjects for the Latino study and 9,501 for the combined study. Results for rs12450091 with and without including the common variants are shown. Positions are from NCBI build 37. LD r^2 and D' between rs12450091 and each common variant were estimated in 4,696 unrelated Latino individuals in our study.

2.3.3 Power analyses for low-frequency variants

After testing for nearly 33,000 functional low-frequency exonic variants in 11,225 subjects, only one variant remained statistically significant in Latinos after correcting for testing of 9,519 of variants, and only six were modestly associated at p-values $<10^{-4}$ in the African ancestry or the combined sample. To assess whether the paucity of associations was due to lack of power, we performed simulations to determine the power of our studies to detect associations with asthma for low-frequency variants. We considered odds ratios (ORs) of 1.5, 2, 2.5, 3, and 4, within the expected ranges of effect sizes for low-frequency variants with moderate to high effects. In fact, we had at least 80% power to detect associations with variants with frequencies $\geq 1\%$ and ORs of 2.5 or greater in the combined sample. In the European American and Latino samples, we had 80% power to detect variants with frequencies $\geq 1\%$ and ORs of 4 or greater. We had lowest power in the African ancestry sample, in which there was 80% power to detect variants with frequencies $>2\%$ and ORs of 4 or greater (Supplementary Figure 2.2). Thus, the lack of significant associations with low-frequency variants in our study could be due to low power in the ethnic/racial groups, small effect sizes of variants in this frequency range, or both.

2.3.4 Gene-based analyses

Studies of rare variants have virtually no power to detect individual associations in samples of even many thousands of individuals^{50,82,83}. Rather, gene-based tests that combine the evidence for association among all variants within each gene provide more power to detect associations. To test the hypothesis that the overall burden of functional variants at each gene confers risk for asthma, we performed a gene-based association test as implemented in MetaSKAT-O⁸⁴ for each

of the 10,208 genes represented on the exome array that had three or more functional variants present in at least two case/control samples. We included in these analyses all putatively functional variants on the array regardless of allele frequency. We performed two sets of analyses. In the first we weighted all variants equally, and in the second we gave larger weights to rarer variants (MetaSKAT-O default). For each of those analyses, we considered all potentially functional variants (missense, nonsense, splice site) and then considered only variants predicted to be ‘probably damaging’ by Polymorphism Phenotyping v2 (PolyPhen-2)⁸⁵. In the analysis with equal weights and defining functional variants more broadly, associations with *GSDMB* and *ZBPB2*, two genes at the 17q12-21 asthma locus, were significant in the combined (MetaSKAT-O $p=6.10 \times 10^{-10}$, 16 variants; $p=1.34 \times 10^{-6}$, 7 variants; respectively; $N=9,501$) and Latino (MetaSKAT-O $p=1.02 \times 10^{-6}$, 12 variants; $p=1.73 \times 10^{-6}$, 7 variants; respectively; $N=3,912$) samples after Bonferroni correction for 9,534 and 10,439 genes, respectively. Additionally, the *MTHFR* gene was significant (MetaSKAT-O $p=1.72 \times 10^{-6}$, 11 variants; $N=2,308$) in the African ancestry sample after multiple testing correction for 10,342 genes. When we considered only variants with PolyPhen-2 prediction of ‘probably damaging’, the association with *GSDMB* remained significant (combined MetaSKAT-O $p=4.09 \times 10^{-8}$, 7 variants, $N=9,501$; Latino MetaSKAT-O $p=7.81 \times 10^{-8}$, 5 variants, $N=3,912$), but the associations with the *ZBPB2* and *MTHFR* genes did not. No genes were significant when larger weights were applied to rare variants in either of the analyses (Table 2.4 and Supplementary Figure 2.3-Supplementary Figure 2.6), suggesting that these gene-based associations are driven largely by the more common functional variants that are present on the array. In fact, when we remove one common missense SNP in *GSDMB*, (c.G844A, p.G282R; rs2305479; MAF 0.335 in Latinos), the

evidence for association of this gene in the test that considered only PolyPhen-2 predicted damaging variants becomes non-significant (equal weights; MetaSKAT-O $p=1.0$ in Latinos, $N=3,912$, and $p=0.276$ in the combined sample, $N=9,501$). Similar to *GSDMB*, the associations with *ZPBP2* were no longer significant after removing the one common missense variant (c.G518T p.S173I ; rs11557467; MAF 0.3576 in Latinos) and testing the association of the remaining 6 variants in the Latino (MetaSKAT-O $p=0.54$, $N=3,912$) and the combined sample (MetaSKAT-O $p=0.12$, $N=9,501$). This suggests that the association with *ZPBP2* is driven by this common variant. In contrast, removing the one common missense variant in *MTHFR* (c.C665T p.A222V; rs1801133; MAF=0.11 in African ancestry) does not eliminate the gene-based association signal (9 rare, 1 low-frequency variant; MetaSKAT-O $p=0.002$, $N=2,308$), indicating that the association with variants in the *MTHFR* gene is attributable to common, low-frequency and rare variants.

Table 2.4: Genes significantly associated with asthma

A. All missense, nonsense and splice site variants (regardless of MAF) were considered, resulting in analysis of 8,933 genes in European Americans, 10,342 genes in African Ancestry groups, 10,439 genes in Latinos and 9,534 genes in the combined sample. **B. Only variants that are predicted to be ‘probably damaging’ by PolyPhen-2⁸⁵** were considered, resulting in analysis of 1,977 genes in European Americans, 2,465 genes in African Ancestry, 2,427 genes in Latinos and 2,316 genes in the combined sample. Gene-based analyses of functional variants were conducted using equal variant weights. All genes with at least 3 functional variants present in at least two studies were included in this analysis. Genes presented in the table are significant after Bonferroni correction for the number of genes tested. Meta-SKAT-O⁸⁴ gene-based analyses included 3,281 subjects for the European American studies, 2,308 for the African ancestry studies, 3,912 subjects for the Latino studies and 9,501 subjects for the combined studies. Chr, chromosome; No, number.

Sample	Gene	Chr.	No. variants	MetaSKAT-O P
A. Functional variants: missense, nonsense and splice site				
Combined	<i>GSDMB</i>	17	16	6.10×10^{-10}

Table 2.4 (continued)

Sample	Gene	Chr.	No. variants	MetaSKAT-O <i>P</i>
Combined	<i>ZPBP2</i>	17	7	1.34×10^{-6}
Latino	<i>GSDMB</i>	17	12	1.02×10^{-6}
Latino	<i>ZPBP2</i>	17	7	1.73×10^{-6}
African Ancestry	<i>MTHFR</i>	1	11	1.72×10^{-6}
B. Functional variants: PolyPhen-2 probably damaging variants				
Combined	<i>GSDMB</i>	17	7	4.09×10^{-8}
Latino	<i>GSDMB</i>	17	5	7.81×10^{-8}

2.3.5 Assessing coverage of exonic variation on the exome array

Lastly, we assessed how well the variants on the exome array captured the low-frequency and rare functional variation present in the genomes of asthmatic individuals. We had available whole genome sequences (WGS) for 278 asthmatics (101 European Americans, 93 African Americans and 84 Latinos) as part of ongoing studies of the EVE Consortium^{63,68}, including 172 of the individuals in this study (43 European Americans, 44 African Americans, 85 Latinos). We selected all functional exonic variants with MAF<5% in any of the 3 groups from the WGS and compared the overlap between the missense, nonsense and splice site mutations present in the WGS with the variants on the exome array. Overall, there were 105,943 rare or low-frequency variants in 16,359 genes in the WGS of the 278 asthmatic individuals, of which 65,170 (61.51%) variants in 14,222 genes (86.93%) were on the array (Table 2.5).

Table 2.5: Functional variant discovery in whole genome sequences of 278 asthmatics

Ethnic group	WGS sample size	No. functional variants in WGS		No. genes with functional variants in WGS	
		Total	Present on exome array (%)	Total	Present on exome array (%)
European American	101	36,202	22,955 (63.40%)	12,568	10,044 (79.91%)
African Ancestry	93	59,851	38,227 (63.87%)	14,452	12,063 (83.47%)
Latino	84	37,308	23,420 (62.77%)	12,441	9,865 (79.29%)
Combined	278	105,943	65,170 (61.51%)	16,359	14,222 (86.93%)

WGS, whole genome sequences; No, number. Functional variants in this analysis include all missense, nonsense and splice site mutations with MAF<5% within any of the three ethnic groups. Functional variants in the WGS were excluded if they were absent in dbSNP137 or in the ESP variant server.

Surprisingly, the proportion of functional variants on the array was similar in the three ethnic groups (range: 62.77% to 63.87%), despite the fact that European Americans were the predominant source of variation in the array design (~9,000 of ~12,000 subjects)²⁶. Finally, we compared the number of genes in the WGS with functional exonic variants to the number of genes represented on the exome array. For this analysis, we stratified the 16,359 genes in the WGS with functional exonic variation into three groups: (i) genes not represented on the array, (ii) genes represented on the array but not all variants are included, and (iii) genes represented on the array and all functional variants are included. Overall, 2,137 genes (13.06%) were not represented on the array, and those genes included 5,283 functional exonic variants in the WGS. An additional 10,686 genes (65.32%) had incomplete sets of functional variants, with 35,490 variants present in the WGS but absent on the array. Finally, 3,536 genes (21.61%) genes had all 9,370 functional exonic variants present in the WGS of asthmatic individuals (Supplementary Table 2.3). At the

17q12-21 locus, *ORMDL3* had one rare missense mutation in Latinos (c.C287T, p.T96M; rs200735199; MAF 0.005) in the WGS and it was not included on the array, *GSDMB* had eight variants in the WGS and seven were on the array (one missense mutation missing, c.C689T, p.S230F; transcript accession NM_001165959; rs183724236; MAF 0.005 in African Americans), *IKZF3* had one missense variant in the WGS and it was on the array, and *ZPBP2* had five variants in the WGS and four were on the array (one missense mutation missing, c.G586T, p.G196C; transcript accession NM_198844; rs371561983; MAF 0.005 in Latinos). Thus, overall the rare and low-frequency functional variation at the 17q12-21 locus was reasonably well covered by the array.

2.4 Materials and Methods

2.4.1 Study Subjects

The subjects in the study were from 13 individual studies from eight investigators that comprise the EVE Consortium on the Genetics of Asthma. This sample has large overlap with our previous asthma meta-analyses^{63,68}, but includes three additional studies: GALA II, COAST, and ALHS. Descriptions of each of the studies can be found in prior published work^{63,68,69} or in Supplementary Note 2.1. The results reported here are based on analyses of 4,794 asthma cases, 4,707 asthma controls and 590 case-parent trios (Table 2.1). All participants, or the parents of minors, provided written informed consent. The relevant local institutional research ethics committee approved each contributing study. Albert Einstein College of Medicine of Yeshiva University, Baylor College of Medicine, Brigham and Women's Hospital, Children's Memorial Hospital, Columbia University, Henry Ford Health System, Johns Hopkins University, Kaiser Permanente, NIEHS (USA), National Institute of Public Health, Mexico (NIPH), Northwestern

University, San Juan VA Medical Center, University of California San Francisco, University of Chicago, University of Southern California, University of Wisconsin Madison, WIRB for Torre Medica Auxilio Mutuo.

2.4.2 Exome Array

The Illumina Infinium HumanExome BeadChip genotyping array interrogates 247,870 variants discovered from the exome and genome sequences of approximately 12,000 individuals of European American, African American, Latino and Asian subjects from a range of common disorders. Variants included on the array are predicted to have a functional effect on the protein and are enriched for missense mutations, nonsense mutations and variants within predicted splice sites, as previously described²⁶.

2.4.3 Genotyping

A total of 7,879 samples from seven studies (CAG, CAMP, COAST, GALA II, GRAAD, MCCAS, SAPPHIRE), including 107 blind duplicates, were genotyped at the Northwest Genomics Center at the University of Washington. An additional 2,440 CHS samples were genotyped at the Molecular Genomics Core at the University of Southern California, using a different version of the array, and 1,585 ALHS samples were genotyped at the Center for Inherited Disease Research at Johns Hopkins University. Genotype calling was performed using Genotyping Module version 1.9.4, GenTrain version 1.0 in GenomeStudio version 2011.1 for all projects except for CHS, which used GenTrain version 2.0. Genotype clustering was defined utilizing our study samples and was completed independently for each project. Because of the different array design used for the CHS samples, the union of 239,909 (96.79%) variants was genotyped in the 13 sample sets. An

additional 7,961 variants were genotyped in all the other samples and were subjected to quality control and meta-analyzed independent of CHS.

2.4.4 SNP quality control

Among the variants present on the Illumina HumanExome BeadChip array, 244,794 (98.76%) passed quality control (Supplementary Table 2.1). We excluded from analyses variants with genotype call rates less than 95%, Hardy-Weinberg equilibrium p-values less than 10^{-4} in any of the studied populations (African American, African Caribbean, European American, Mexican/Mexican American or Puerto Rican) and 333 caution sites reported as problematic by the exome array design group (http://genome.sph.umich.edu/wiki/Exome_Chip_Design; Supplementary Table 2.4). Among the variants passing quality control, 197,339 (79.61%) were variable in the individuals under study. In the combined sample, 23,206 (11.76%) variants were common ($MAF \geq 5\%$) to all ethnic groups, 36,413 (18.45%) were low-frequency (MAF 1-5%) and 137,720 (69.79%) were rare ($MAF < 1\%$) in all three groups. Private variants for each group and counts of variants tested are shown in Supplementary Table 2.1 and Supplementary Table 2.2. Additional probe quality and genotype calling evaluations were performed and are described in Supplementary Note 2.2.

2.4.5 Sample quality control

Six hundred and seventy-nine subjects were excluded from analysis (Supplementary Table 2.5) resulting in a final sample size of 11,225 individuals. Sample exclusions included gender discordances, sample duplications, Mendelian errors in trios, as implemented in PLINK⁸⁶, missing case/control status and incomplete trios. In addition, we completed principle component analysis

(PCA) using the R function `prcomp` to identify ancestry outliers (Supplementary Figure 2.7-Supplementary Figure 2.9). PCA was carried out utilizing the ancestry informative markers present on the exome array and including HapMap⁸⁷ CEU, YRI, CHB, JPT and MEX subjects. Fifty-six samples were excluded because one or both of the first two principle components were more than six standard deviations from the mean of the expected ethnic group. These inferred principle components were applied as covariates in the gene-based tests.

2.4.6 Single variant analyses

We subjected functional low-frequency variants to single variant association analysis for each of the 3 ethnic groups (European American, African American, and Latino) and for the combined sample. Low-frequency variants (1-5%) were included in each of the ethnic specific meta-analyses (8,249 in European American, 17,861 in African American, and 9,519 in Latinos) and the variants that were neither common ($\geq 5\%$) in all 3 ethnic groups nor rare ($<1\%$) in all three ethnic groups (32,681 variants) were meta-analyzed in the combined sample (Supplementary Table 2.2). To control for population stratification and relatedness, single SNP association tests for case-control studies were completed using the linear mixed model Genome-wide Efficient Mixed Model Association⁷³ (GEMMA). For the samples comprised of case-parent trios, we used the Family Based Association Test⁷⁴ (FBAT). Z-score statistics obtained from GEMMA and FBAT were combined in a meta-analysis as a weighted sum to reflect the power of each independent study. The weight (w) used for each SNP is a function of allele frequency (p), proportion of cases (v) and sample size (N) ($w = \sqrt{v(1-v)p(1-p)2N}$)⁶⁸ of each study. P-values were calculated using standard normal approximations and Bonferroni significance thresholds were applied ($\alpha = 0.05$) to correct for

the number of variants tested within each group and in the combined sample. Odds ratios were obtained for samples analyzed with GEMMA as a linear combination of log odds ratios with weights proportional to the standard errors of the log odds ratios using the R package *meta*.

2.4.7 Power analyses for single variant analyses

For each ethnic group and for the combined sample, we performed the following simulations. We randomly selected 1,000 variants with MAF between 0.5% and 5% and simulated genotypes 1,000 times for each of five odds ratios (OR = 1.5, 2, 2.5, 3 and 4). We utilized the same study structure (sample sizes and case/control or parent offspring designs) and observed MAF distributions that were present in our data. Z-scores were then meta-analyzed with the weighted approach described in the single variant meta-analysis methods and power was calculated with a predefined alpha of 0.05. We plotted power estimates vs. MAF and implemented a local linear regression fit with the R function *loess* (Supplementary Figure 2.2).

2.4.8 Gene-based analyses

Functional variants sets were defined two ways: 1) missense, nonsense or variants within splice sites or 2) coding variants with PolyPhen-2⁸⁵ prediction scores ≥ 0.957 (probably damaging variants). PolyPhen-2 predicts possible impact of an amino acid substitution on the structure and function of human proteins. In the first functional definition set, variant pruning for each gene was completed within each ethnic group by randomly selecting one variant for each pair of functional variants with linkage disequilibrium $r^2 > 0.9$ (r^2 and D' values were calculated with PLINK). Requiring each gene to comprise at least 3 functional variants, we considered a total of 140,986 functional coding variants (8,933 genes) in European Americans, 148,677 functional variants (10,342 genes)

in African Americans, 149,236 functional variants (10,439 genes) in Latinos and 151,006 functional variants (9,534 genes) in the combined sample. In the second functional definition set, we considered a total of 20,077 functional coding variants (1,977 genes) in European Americans, 22,625 functional variants (2,465 genes) in African Americans, 22,644 functional variants (2,427 genes) in Latinos and 22,248 functional variants (2,316 genes) in the combined sample. We performed a meta-analysis of gene-based association tests applying the first 5 principal components as implemented in MetaSKAT-O⁸⁴. Analyses were completed with two sets of variant weights: either larger weights on rare variants (MetaSKAT-O default parameter) or equal weights for all variants. Bonferroni significance thresholds were applied ($\alpha = 0.05$) to correct for the number of genes tested within each group and in the combined sample.

2.4.9 Assessment of functional variants on the Illumina exome array

Whole genome sequences (WGS) for the 278 asthmatics was carried out at Complete Genomics (CGI) and resulted in an average read depth of 95.4% at 20X or greater across the exome target and 91.4% across the genome. We excluded singletons from this analysis that were discovered in the 278 WGS but not present in dbSNP137 or in the Exome Sequencing Project (ESP) variant server. We selected functional variants (missense, nonsense and splice site mutations) that were low-frequency or rare ($MAF < 5\%$) in any of the ethnic groups in the WGS and identified the overlap with the variants present on the exome array.

2.5 Discussion

Our study is the largest to date to assess the effects of rare variants on asthma risk in ethnically diverse individuals. Although we were limited by the variation present on the array and by

relatively small samples within each ethnic/racial group, our results provide several insights into the genetic architecture of asthma. First, associations with rare or low-frequency variants are likely to be ethnic-specific. The few significant associations that we detected were all limited to a single ethnic group, despite having the greatest power to detect associations in the combined sample. Two novel associations, one with a missense mutation in the *GRASP* gene in Latinos and one with the many rare missense mutations in the *MTHFR* gene in African Americans/African Caribbeans, were revealed. The *MTHFR* gene harbors 11 missense mutations in African Americans, only one of which is common (rs1801133), and showed a gene-level association with asthma. Although most mutations in this gene were not predicted to be ‘damaging’ by PolyPhen-2 scores, the sheer number of potentially functional variants and the strong association with asthma in the African ancestry sample makes this gene a good candidate for further studies. Common variants in *MTHFR*, encoding methylenetetrahydrofolate reductase, have been associated with many common conditions, including neural tube defects, vascular disease, and pregnancy complications, among others^{75,88-90}. There have been conflicting reports as to whether the common C677T variant in this gene is associated with asthma or allergic diseases in European populations^{63,64,91-93}, although maternal folate levels in pregnancy have been associated with risk of early asthma outcomes in children^{94,95}. The missense mutation in *GRASP* that is associated with asthma in Latinos occurs at a frequency of 0.012 in this population, at higher frequency (0.034) in European Americans, and very low-frequency (0.006) in the African ancestry samples. The fact that the association was confined to Latinos even though we had greater power to detect associations in the European Americans suggests that the effects of this variant on asthma risk might involve

interactions with other variants or environmental exposures that are more common in Latinos, but this single variant association requires replication in an independent Latino sample or functional validation of this variant's role in asthma risk. *GRASP* encodes GRP1 (general receptor for phosphoinositides-1)-associated scaffold protein and has been shown to be highly expressed in the brain with lower levels of expression in lung, heart, embryo, kidney, and ovary⁹⁶. Although little is known about the function of this gene or its potential role in asthma, scaffold proteins have been implicated in asthma signaling pathways^{97,98}.

Second, among the six low-frequency variants showing suggestive evidence for association with asthma (Table 2.2), three variants were associated with decreased risk and three were associated with increased risk. The discovery of potentially protective alleles in three genes, *FAM134B* (family with sequence similarity 134, member B1), *NPC1L1* (Niemann-Pick disease, type C1, gene-like 1), and *IKZF1* (IKAROS family zinc finger 1), is intriguing because elucidating the function of these variants could identify novel therapeutic targets for asthma.

Third, the *GSDMB* (gasdermin B) locus on chromosome 17q12-21 is within a LD block that has been robustly associated with asthma^{63-65,70,71}, yet the causal gene and its function are still unclear. The SNPs discovered by GWAS that are associated with asthma are also associated with the expression of both *GSDMB* and its close neighbor *ORMDL3* in white blood cells⁷⁰, peripheral blood mononuclear cells (PBMCs)⁷⁸, lymphoblastic cells lines (LCLs)⁹⁹, CD4+ T cells^{80,81} and lung tissue⁷⁹, and it has been assumed both that the causal SNPs at this locus are expression quantitative trait loci (eQTLs) for these genes and that one or both of these genes are involved in asthma pathogenesis. Our study revealed several lines of evidence of association with *GSDMB* and

extensive functional variation within this gene, including a common missense mutation rs2305479 (c.G871A, p.G291R; rs2305479) that is predicted to be ‘probably damaging’ by PolyPhen-2. This common variant was identified in the previous EVE meta-analysis of GWAS⁶³ ($p=1.21 \times 10^{-12}$ in the combined sample) and is in strong LD with the GWAS-identified SNP but its potential as a causal candidate has not previously been discussed. Our study suggests rs2305479 as a good candidate for the causal variant at this important locus and implicates *GSDMB* more directly in asthma risk. We identified a second potential causal variant, a common SNP within a splicing site in *GSDMB* (rs11078928; c.661-2A>G; transcript accession NM_001165959). This SNP was previously shown to have a genotype-dependent effect on *GSDMB* transcript levels, whereby the asthma risk allele (A) was associated with increased expression of an alternate transcript lacking exons 5-8¹⁰⁰. Although not much is known about the function of *GSDMB* or of its alternative splice forms, other gasdermin genes are involved in apoptosis in epithelial cells¹⁰¹, processes and cells that are relevant to asthma. It is additionally notable that in contrast to *GSDMB*, there was only one rare missense mutation in *ORMDL3* in Latinos (rs200735199, c457T, p.T96M; transcript accession NP_644809; MAF 0.005) in the whole genome sequences from 278 asthmatics, and this variant was not included on the exome array.

In summary, we report significant associations of asthma with rare or low-frequency variants in two novel genes, *GRASP* and *MTHFR*, in Latinos and the African ancestry sample, respectively, and with a potentially damaging common missense mutation in *GSDMB* in Latinos and the combined sample. Based on the whole genome sequences from 278 asthmatic subjects, our study included over 60% of all putatively functional rare and low-frequency variation. Thus,

even if we discovered only half of all truly associated variants, it is unlikely that rare and low-frequency variants will account for a significant portion of the heritability of asthma.

2.6 Acknowledgements

This work was supported by NIH grants RC2 HL10165, P30 ES007048, P01 ES009581 and P01ES01162, R01 HL118267, and R01 AI079139, the Intramural Research Program of the NIH (NIEHS Z01 CP010119, Z01 ES049030 and ES102385), the American Asthma Foundation, the Hastings Foundation, the Fund for Henry Ford Hospital and through resources provided by the Computation Institute and the Biological Sciences Division of the University of Chicago and Argonne National Laboratory (S10 RR029030-01). Genotyping services were provided by the Northwest Genomics Center at the University of Washington, Department of Genome Sciences, under U.S. Federal Government contract number HHSN268201100037C from the National Heart, Lung, and Blood Institute, and the Molecular Genomics Core at the University of Southern California (P30ES007048, P01ES009581) and the Center for Inherited Disease Research at Johns Hopkins University. C.I. was supported by National Institutes of Health Grant T32 GM007197 and by the Ruth L. Kirschstein National Research Service Award (PA-11-112).

2.7 Appendix A: Supplementary Data

Supplementary Table 2.1: Number of variants on the exome array that passed QC and are variable in 11,225 study subjects

Ethnic group	Total variants	Private variants	Common variants	Low-frequency variants	Rare variants
EA (N=4,363)	133,748	17,227	26,967	9,470	97,311
AA (N=2,308)	139,694	14,704	31,803	19,496	88,395
LAT (N=4,554)	159,189	13,559	28,005	10,667	120,517
Total (N=11,225)	197,339	-	23,206	36,413	137,720
Total Functional	171,677	-	9,459	32,681 ^a	129,537

EA, European American; AA, African Ancestry; LAT, Latino. ^aVariants included in single-variant meta-analyses. Total counts for common and rare variants include SNPs with allele frequencies $\geq 5\%$ or $< 1\%$, respectively, in all three ethnic groups; the remaining variants are referred to as low frequency. Breakdown of low frequency variants by functional category is provided in Supplementary Table 2.2.

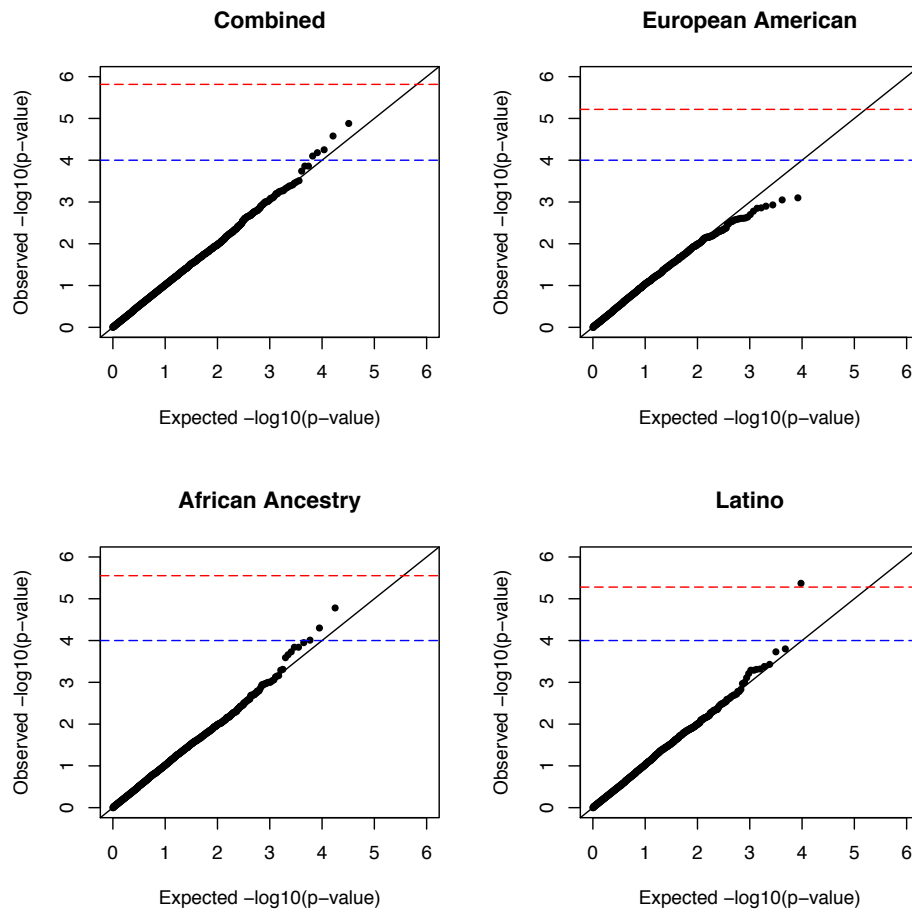
Supplementary Table 2.2: Variants included in single variant analyses

Variant type	Total (N=11,225)	MAF (1-5%)		
		EA (N=4,363)	AA (N=2,308)	LAT (N=4,554)
Missense	32,157	8,095	17,606	9,350
Nonsense	296	84	159	83
Splice site	228	70	96	86
Total	32,681	8,249	17,861	9,519

MAF Minor allele frequency; EA, European American; AA, African Ancestry; LAT, Latino.

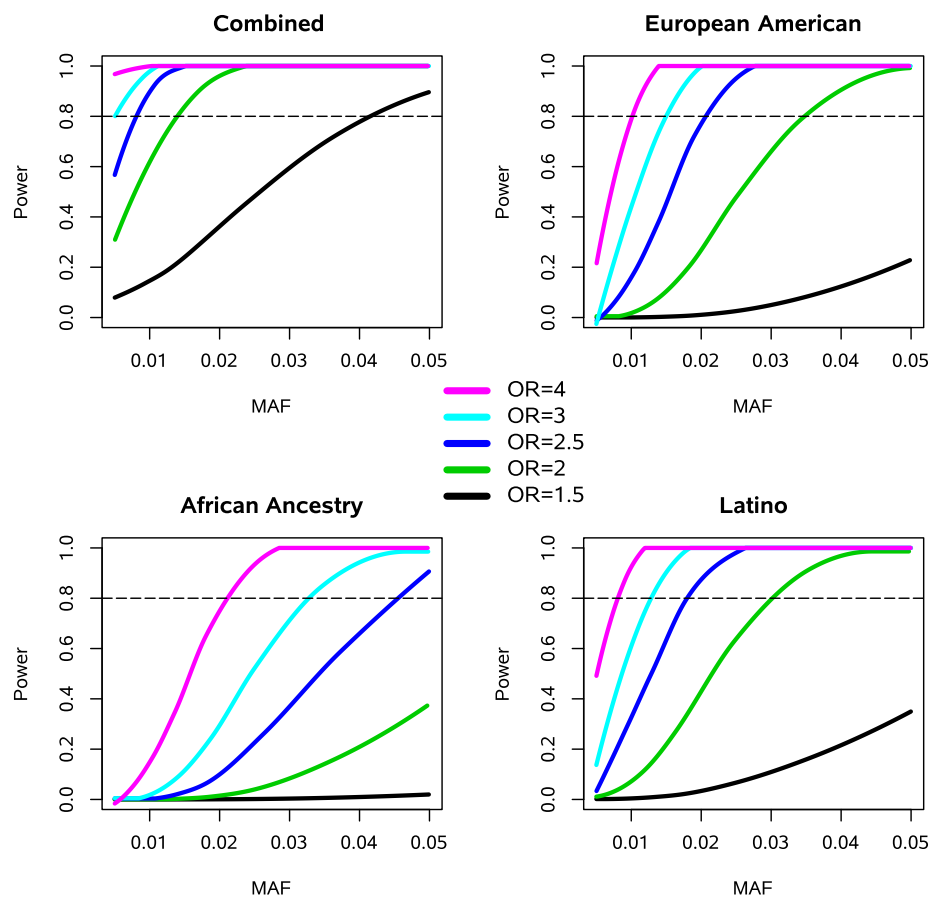
Supplementary Figure 2.1: QQplots of single variant meta-analyses of low frequency, functional variants and asthma risk

We considered 32,681 variants in the combined sample (N=11,225), 8,249 in European Americans (N=4,363), 17,861 in African Ancestry (N=2,308), and 9,519 in Latinos (N=4,554). Dashed red line shows the Bonferroni threshold of significance (corrected for number of variants tested). Dashed blue line shows $p=0.0001$.



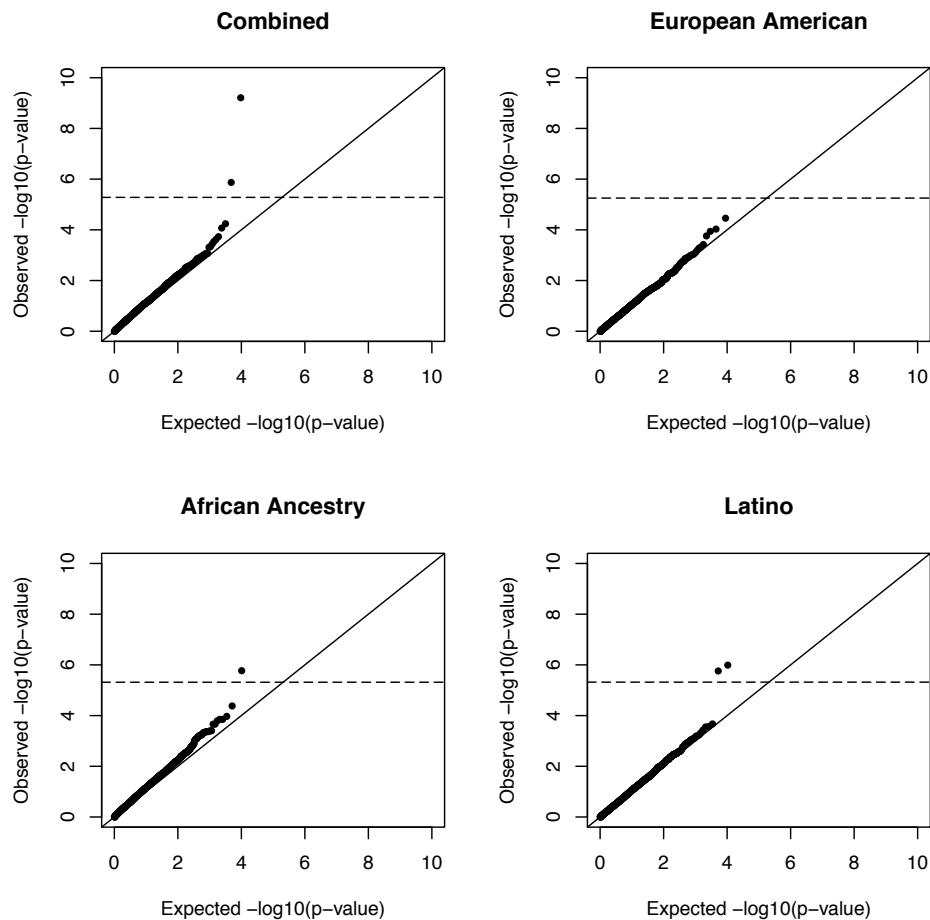
Supplementary Figure 2.2: Power analyses for low frequency single variant meta-analyses

For each ethnic group and the combined sample, 1,000 variants with MAF 0.5%-5% were randomly sampled for simulations; meta-analyses were performed using the same allele frequency structure in our data and the same study design of our study to maintain a 5% type I error. We performed simulations with ORs=1.5, 2, 2.5, 3, and 4, and fit a local linear regression curve. Dashed horizontal lines represents 80% power in the meta-analysis.



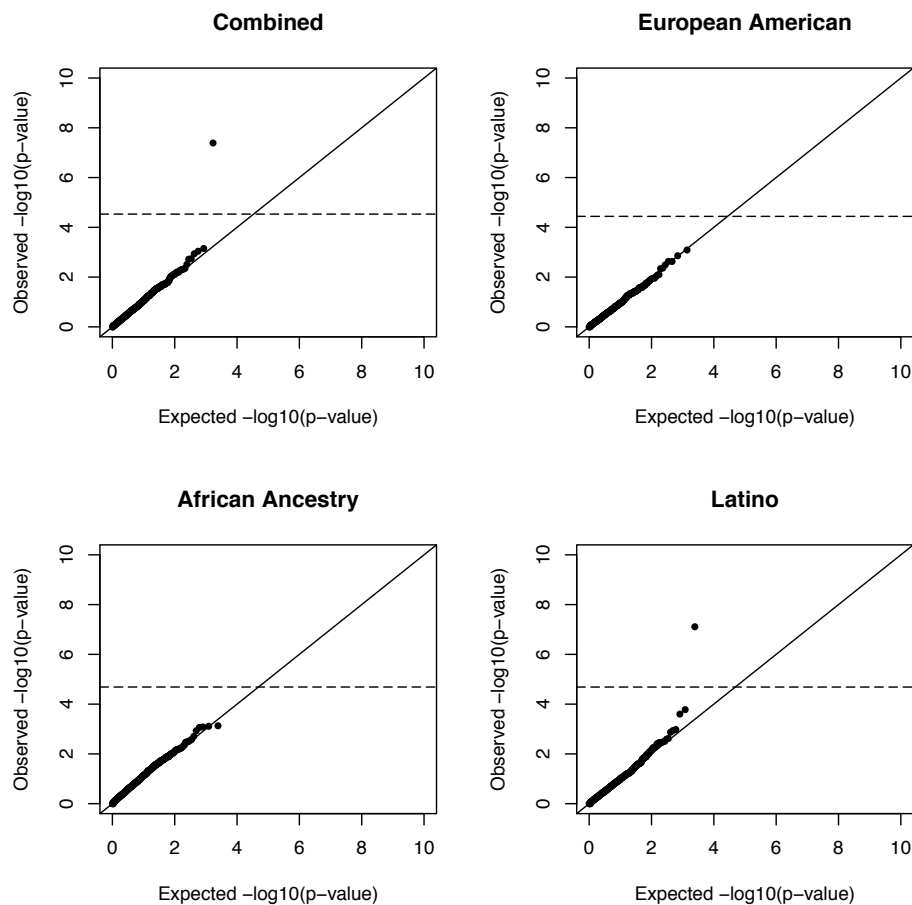
Supplementary Figure 2.3: Gene-based meta-analyses qqplots for functional variants defined as missense, nonsense and splicing and equal variant weights

Gene-based analyses of functional variants were conducted using equal variant weights (MetaSKAT-O⁸⁴). All genes with at least 3 functional variants that were present in at least two studies were included in this analysis. All missense, nonsense and splice site variants (regardless of MAF) were considered functional, resulting in analysis of 8,933 genes in European Americans (N=3,281), 10,342 genes in African Americans (N=2,308), 10,439 genes in Latinos (N=3,912) and 9,534 genes in the combined sample (N=9,501). Dashed line represents the Bonferroni threshold of significance (corrected for number of genes tested).



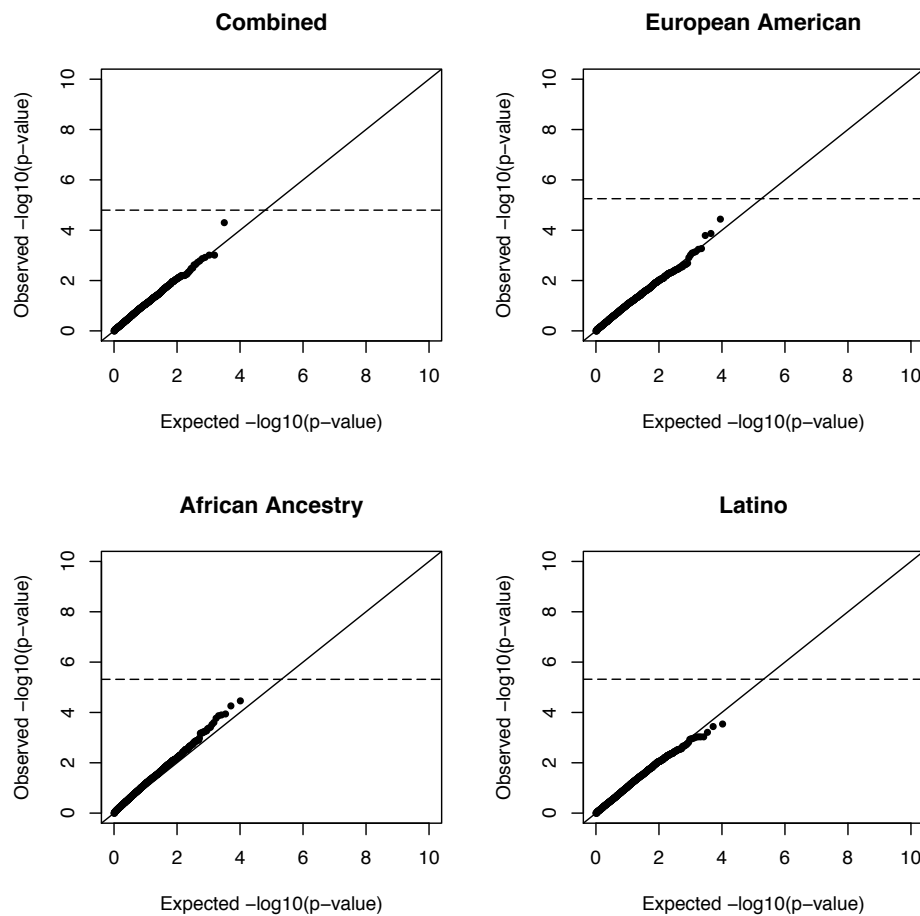
Supplementary Figure 2.4: Gene-based meta-analyses qqplots for functional variants defined as PolyPhen-2 probably damaging variants and equal variant weights

Gene-based analyses of functional variants were conducted using equal variant weights (MetaSKAT-O). All genes with at least 3 functional variants that were present in at least two studies were included in this analysis. All variants predicted to be 'probably damaging' by PolyPhen-2 were considered functional, resulting in analysis of 1,977 genes in European Americans (N=3,281), 2,465 genes in African Ancestry (N=2,308), 2,427 genes in Latinos (N=3,912) and 2,316 genes in the combined sample (N=9,501). Dashed line represents the Bonferroni threshold of significance (corrected for number of genes tested).



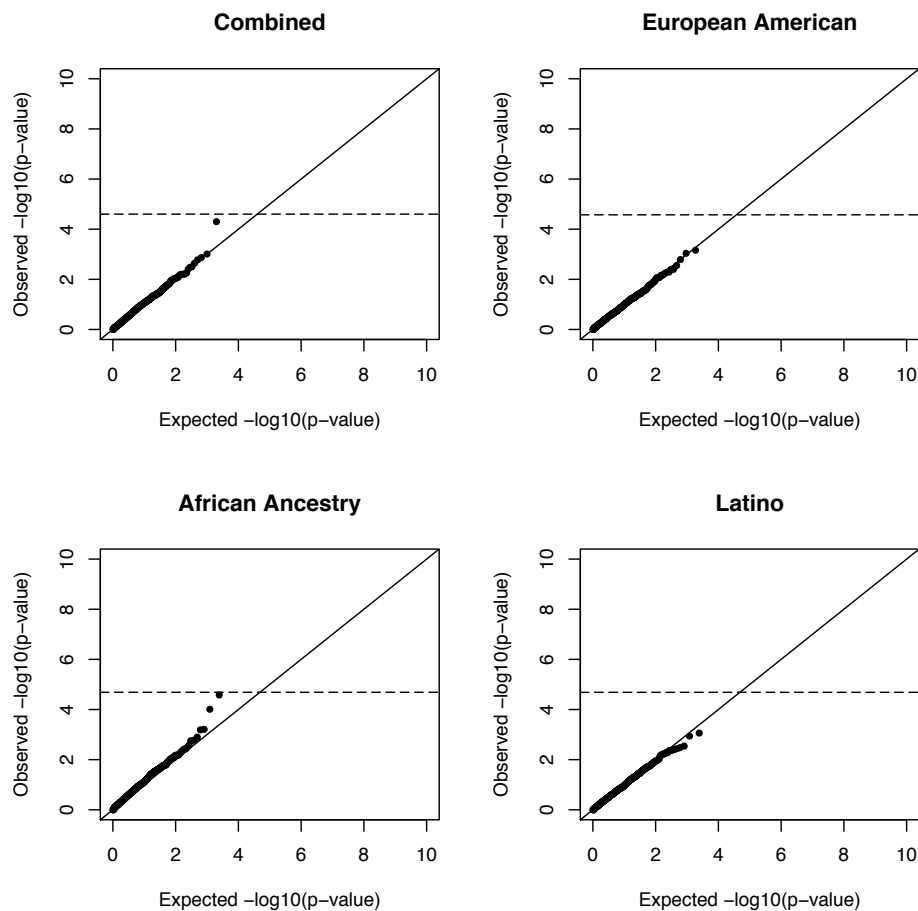
Supplementary Figure 2.5: Gene-based meta-analyses qqplots for functional variants defined as missense, nonsense and splicing and rare variant weights

Gene-based analyses of functional variants were conducted using rare variant weights (MetaSKAT-O default). All genes with at least 3 functional variants that were present in at least two studies were included in this analysis. All missense, nonsense and splice site variants (regardless of MAF) were considered functional, resulting in analysis of 8,933 genes in European Americans (N=3,281), 10,342 genes in African Americans (N=2,308), 10,439 genes in Latinos (N=3,912) and 9,534 genes in the combined sample (N=9,501). Dashed line represents the Bonferroni threshold of significance (corrected for number of genes tested).



Supplementary Figure 2.6: Gene-based meta-analyses qqplots for functional variants defined as PolyPhen-2 probably damaging variants and rare variant weights

Gene-based analyses of functional variants were conducted using rare variant weights (MetaSKAT-O default). All genes with at least 3 functional variants that were present in at least two studies were included in this analysis. All variants predicted to be 'probably damaging' by PolyPhen-2 were considered functional, resulting in analysis of 1,977 genes in European Americans (N=3,281), 2,465 genes in African Ancestry (N=2,308), 2,427 genes in Latinos (N=3,912) and 2,316 genes in the combined sample (N=9,501). Dashed line represents the Bonferroni threshold of significance (corrected for number of genes tested).



Supplementary Table 2.3: Representation of genes on the exome array with functional variants in the whole genome sequences of 278 asthmatic individuals

Ethnic group	No. genes not represented on the array	No. genes represented on the array but not all WGS variants included	No. genes represented on the array and all WGS functional variants included
EA	2,524 (20.08%)	4,604 (36.63%)	5,440 (43.28%)
AA	2,389 (16.53%)	7,198 (49.80%)	4,865 (33.66%)
LAT	2,576 (20.70%)	4,729 (38.01%)	5,136 (41.28%)
COMB	2,137 (13.06%)	10,686 (65.32%)	3,536 (21.61%)
No, number; WGS, whole genome sequence; EA, European American; AA, African American; LAT, Latino; COMB, combined.			

Supplementary Table 2.4: Exome array variants removed from analysis

Quality control exclusion category	Probes removed
Call rate <95%	2,280
SNPs not variable in our study	47,455
European American HWE	229
African American HWE	385
African Caribbean HWE	138
Mexican-American HWE	527
Puerto Rican HWE	242
Mitochondria SNPs	163
Caution site	333
Total unique number of SNPs removed	50,437
HWE, Hardy-Weinberg equilibrium.	

Supplementary Table 2.5: Total subject exclusions.

Exclusion Category	No. Subjects
Gender discordancy	27
Sample duplication ^a	107
Ancestry discordancy ^b	56
Mendelian error ^c	9
Missing case/control status	7
Incomplete trios ^c	473
Total subjects removed	679
No, Number; ^a Sample with lowest genotype call rate excluded. ^b Sample excluded if either of the top two inferred principle component was more than six standard deviations from the mean. ^c Case-parent trios only.	

Supplementary Note 2.1: Description of the Agricultural Lung Health Study

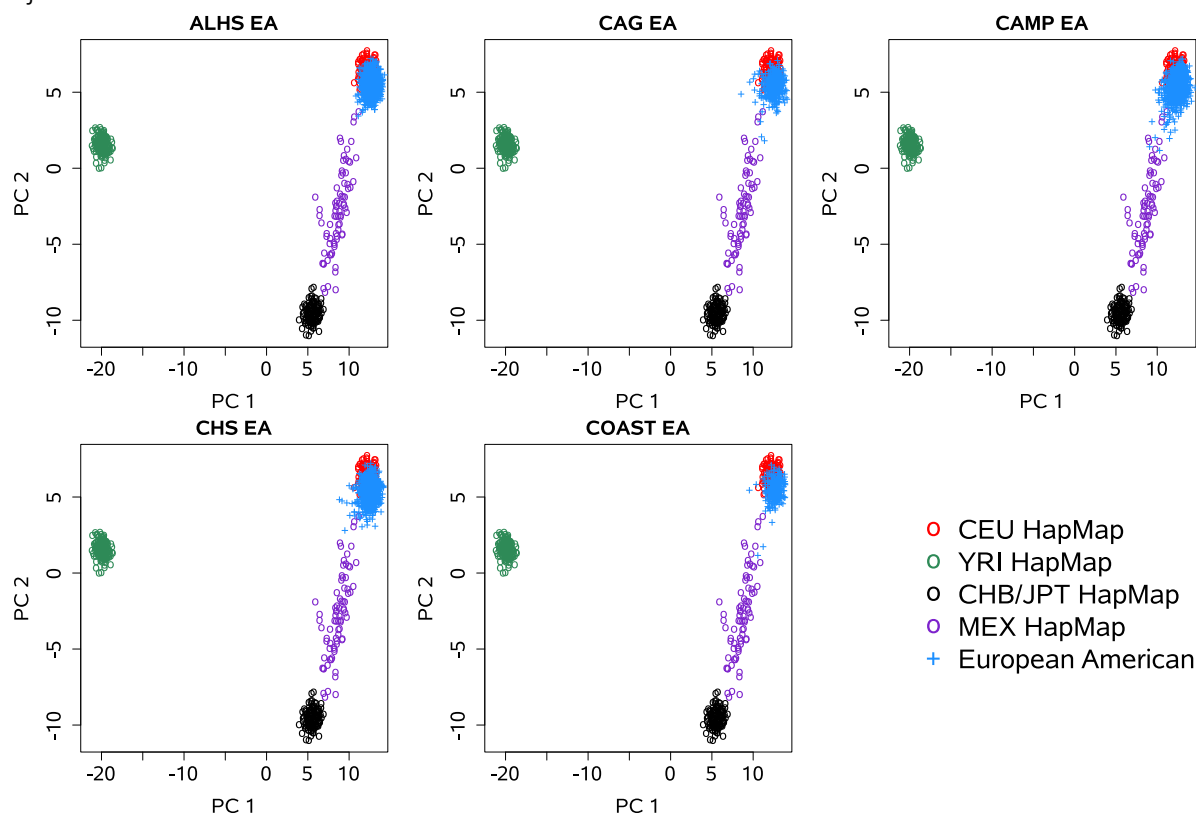
The Agricultural Lung Health Study (ALHS) is a case-cohort study nested within the larger Agricultural Health Study (AHS)⁶⁹. The AHS includes 52,394 subjects who are private pesticide applicators (mainly farmers) in North Carolina and Iowa and 32,345 spouses of farmers. Applicators enrolled in the study from 1993 to 1997 at pesticide licensing facilities by completing self-administered questionnaires. Subjects have been followed up by questionnaires. ALHS recruited participants based on their response to questions about current asthma and asthma symptoms reported during the AHS telephone follow-up interview conducted between 2005-2010. Current asthma cases who were selected for screening were defined as 1) presumptive current asthma (doctor diagnosed asthma with current symptoms or medication use, 81%) or 2) presumptive undiagnosed asthma (persistent wheeze or inhaler use for asthmatic symptoms, 19%). A random cohort sample was also generated from among respondents to the 2005-2010 telephone interview. DNA was extracted from blood for most subjects (approximately 95%), with the remainder from buccal cells collected using Oragene kits (DNA Genotek, Kanata, Ontario, Canada). Participants in the current study were subjects who self-reported as Caucasian enrolled from February 2009 through August 2011. Genotyping for the Illumina HumanExome BeadChip was conducted at the Genetic Resources Core Facility, Johns Hopkins Institute of Genetic Medicine, Baltimore, MD. From among 1,653 subjects genotyped, 31 failed quality control (3 sex discordances and 28 investigator reported exclusions), 44 were not of Caucasian American based on principal components analysis leaving 1,578 subjects for analysis including 776 cases and 802 controls.

Supplementary Note 2.2: Additional probe and genotyping quality evaluation

To further evaluate probe quality in the exome array and genotype calling, genotypes in the 7,879 EVE samples were re-called using optiCall¹⁰². Out of the SNPs that had passed QC filters, 5,244 SNPs (2.15%) resulted in optiCall call rates below 95% and 1,908 SNPs (0.78%) had concordance rates less than 99% between Illumina's genotype caller (GenCall) and optiCall (overall 6,302 unique SNPs). In addition, 82 of the 278 asthmatics with WGS were also genotyped on the exome array platform (including only samples genotyped at the NWGC). Genotype concordance rates for these 82 individuals were calculated for each of the two genotype callers and SNPs were flagged if 5 or more samples (6%) were identified as discordant. GenCall resulted in 3,079 (1.24%) discordant SNPs while optiCall had 3,567 SNPs (1.43%). Based on the slightly higher concordance rate for GenCall, we selected to continue with the original Illumina genotype caller algorithm.

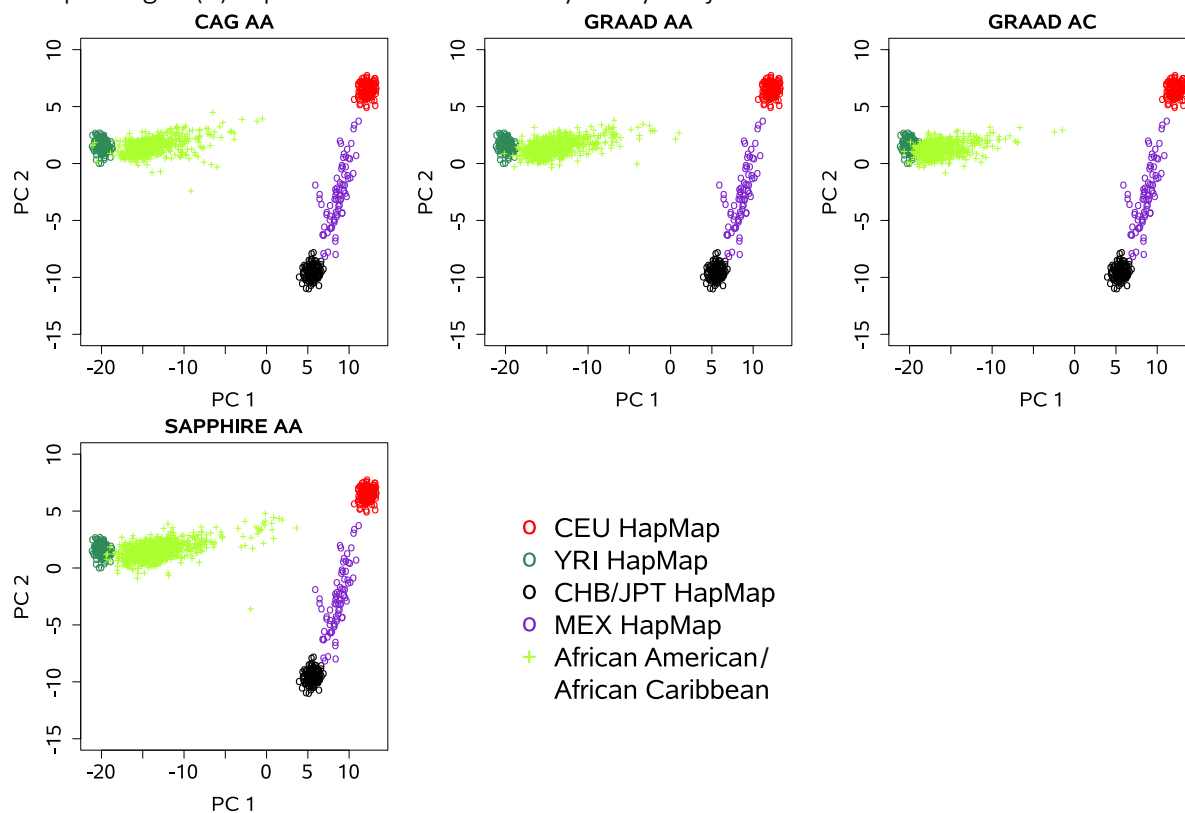
Supplementary Figure 2.7: PCA of European American studies projected onto HapMap subjects

Each panel represents one of the 5 European American studies included in the meta analysis^{63,68} (ALHS, Agricultural Lung Heath Study⁶⁹; CAG, Chicago Asthma Genetics Study; CAMP, Childhood Asthma Management Program; CHS, The Children's Health Study; COAST, The Childhood Origins of Asthma Study). Open circles represent HapMap⁸⁷ subjects of known ethnicities (red, CEU; green, YRI; black, CHB/JPT; purple, MEX). Blue plus signs (+) represent European American study subjects.



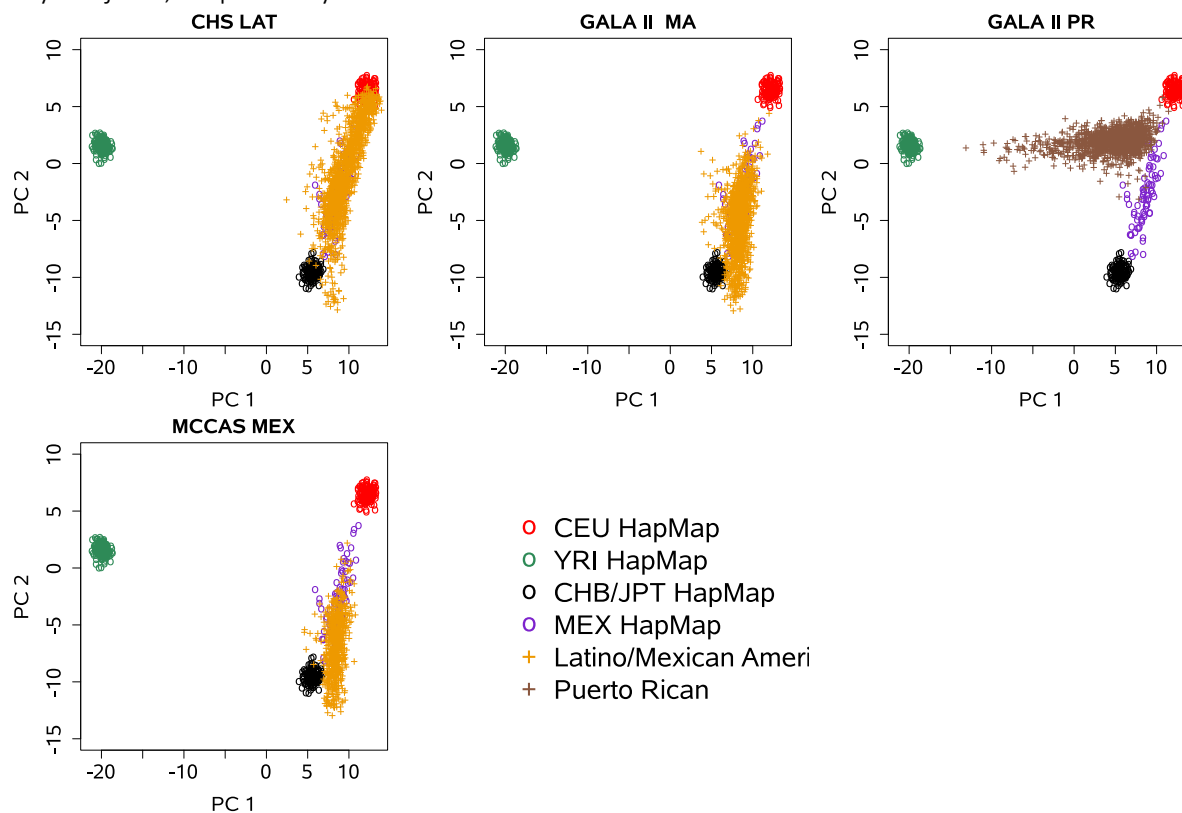
Supplementary Figure 2.8: PCA of African ancestry studies projected onto HapMap subjects

Each panel represents one of the 4 African ancestry studies included in the meta analysis^{63,68} (CAG, Chicago Asthma Genetics Study; GRAAD, Genomic Research on Asthma in the African Diaspora and Barbados [-AA, African Americans; -AC, African Caribbeans]; SAPPHIRE, The Study of Asthma Phenotypes and Pharmacogenomic Interactions by Race-Ethnicity). Open circles represent HapMap subjects of known ethnicities (red, CEU; green, YRI; black, CHB/JPT; purple, MEX). Yellow-green plus signs (+) represent African ancestry study subjects.



Supplementary Figure 2.9: PCA of Latino studies projected onto HapMap subjects

Each panel represents one of the 4 Latino studies included in the meta analysis^{63,68} (CHS, The Children's Health Study; GALA II, Genes-Environment & Admixture in Latino Asthmatics [-MA, Mexican American; -PR, Puerto Rican]; MCCAS, Mexico City Childhood Asthma Study [-MEX, Mexican]). Open circles represent HapMap subjects of known ethnicities (red, CEU; green, YRI; black, CHB/JPT; purple, MEX). Orange and brown plus signs (+) represent Latino and Puerto Rican study subjects, respectively.



CHAPTER 3

RARE VARIANTS ASSOCIATED WITH SERUM BLOOD LIPID TRAITS IN A FOUNDER POPULATION

3.1 Abstract

Founder populations are ideally suited for studies on the clinical effects of alleles that are rare in general populations but occur at higher frequencies in these isolated populations. Whole genome sequencing in 98 South Dakota Hutterites, a founder population of European descent, revealed 790,742 variants that are rare (<1%) or absent in European populations (783,057 [99%] outside the coding region), but occur at frequencies greater than 1% in the Hutterites. We examined the effects of these rare in European variants (REVs) on LDL cholesterol (LDL-C), HDL cholesterol (HDL-C), total cholesterol and triglycerides (TG) levels in 828 Hutterites and applied a Bayesian hierarchical framework to prioritize potentially causal variants based on functional annotations. We identified two novel rare variants within known genome-wide association study (GWAS) loci associated with LDL-C and HDL-C that are independent of known variation in these regions (rs17242388 in *LDLR* and rs141938931 in *SLC12A3*, respectively) and also replicated a previously discovered splicing variant in *APOC3* (rs138326449) associated with TG and HDL-C. In addition, we identified variants at two novel loci that are associated with LDL-C, intronic variants in *EPHA6* (rs191020975) and in *CNIH3* (chr1:225753576). Overall, our results provide evidence that rare non-coding variants influencing lipid traits are likely to mediate their effects by modulating

changes in gene expression. Moreover, our findings further highlight the advantages of population isolates for identifying novel loci and provide insights into the mechanisms regulating lipid traits and potentially new therapeutic targets.

3.2 Introduction

Blood lipid traits are under strong genetic control and are modifiable risk factors for cardiovascular disease, one of the leading causes of death¹⁰³. These traits include low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), total cholesterol and triglyceride (TG) levels and have estimated heritabilities of 40-60% across populations¹⁰⁴⁻¹⁰⁶. Although genome-wide association studies (GWAS) of lipid traits have been successful in identifying common variants with hundreds of robust associations, the variants at these associated loci account for only about 10-14% of the total phenotypic variance^{9,107}. While a portion of the unexplained genetic variance may result from overestimates of heritability and complex genetic architectures, such as those involving epistasis or gene-environment interactions⁸, the effect of rare loss-of-function variation in complex traits is understudied and likely to contribute to the heritability of blood lipid traits and risk for cardiovascular disease.

In fact, sequencing studies in families or patients with rare monogenic lipid disorders have uncovered many novel genes harboring rare coding mutations of large effect and revealed critical pathways for lipid metabolism¹⁰⁸⁻¹¹⁰. These studies have supported earlier observations suggesting that rare variants in the general population contribute significantly to lipid traits and possibly more generally to common, complex phenotypes. In particular, in a resequencing study of genes mutated in monogenic lipid disorders, Cohen *et al.* (2004) identified an enrichment of

nonsynonymous variants associated with lipid levels by sequencing unrelated individuals sampled from the tails of the HDL-C trait distribution¹¹¹.

While providing valuable insight into the mechanisms regulating lipid traits, rare variant studies in complex traits have focused on coding regions of the genome, largely due to recent explosion of exome sequencing studies and the relative ease in interpreting these findings¹¹². As a result, the non-coding portion of the genome has been largely unexplored. Similar to common non-coding variants, rare non-coding variants may also impact gene expression and protein abundance, but the sample sizes of most transcriptome studies are underpowered to identify the effects of rare variants on gene expression. Nonetheless, a few studies have provided strong support for the aggregated effect of multiple rare non-coding variants on gene expression, and identified excesses of rare regulatory variation near genes in individuals with extreme gene expression levels¹¹³⁻¹¹⁵. However, the broader questions of what functional features of rare non-coding variants characterize those that most influence on gene expression and how different functional classes of rare non-coding variations influence disease are unknown.

Founder populations offer the opportunity to study the effects of variants that are rare in general populations but have reached higher frequencies in these isolated populations due to the effects of random genetic drift in the early generations after their founding^{45,46,116}. In addition, their overall reduced genetic complexity and relatively homogenous environment and lifestyle can enhance the effects of rare genetic variants on phenotypic traits and thereby facilitate the detection of susceptibility loci that underlie complex disease, as elegantly illustrated in studies of the Amish and Icelandic populations^{44,117,118}.

In this study, we dissect the genetic architecture of blood lipid traits in South Dakota Hutterites, a founder population of European descent. The 828 individuals who participated in these studies are related to each other in a 13-generation pedigree with 64 founders. We performed GWAS using 790,742 “rare in European variants” (REVs) that occur at frequencies greater than 1% in the Hutterites and integrated functional and regulatory annotations to narrow down potential candidate variants despite the long-range linkage disequilibrium present in the population. Our studies revealed rare variants with large effects at two novel and three known blood lipid loci associated with LDL-C, HDL-C or TG, potentially yielding novel insights into the mechanisms regulating lipid traits and new therapeutic targets.

3.3 Results

3.3.1 Study overview

Approximately 7.6 million variants identified through whole genome sequencing in 98 Hutterite individuals were imputed using the known identity by descent (IBD) structure of the Hutterite pedigree⁴⁹ to the 828 individual in our studies. We selected 790,742 variants that were either absent or rare (<1%) in European databases (see Methods) and occurred at frequencies greater than 1% in the Hutterites (REVs) for association testing with fasting blood lipid measurements (LDL-C, HDL-C, total cholesterol and TG levels) (Table 3.1). Based on their RefSeq¹¹⁹ annotations the majority of these REVs were intergenic (431,061; 54.5%) or intronic (343,146; 43.4%); the remaining 2.1% were exonic (0.9%), 3' and 5' UTR (1.1%), indels and substitutions ($4 \times 10^{-4}\%$) and splicing ($7 \times 10^{-5}\%$; Figure 3.1).

Figure 3.1: Allele frequencies of variants discovered in the Hutterites

A. Bar plot of variants binned by maximum reported allele frequency in European databases, stratified by Hutterite minor allele frequency (MAF). B. Scatterplot of Hutterite MAF versus the maximum allele frequency in European databases. Variants presented in A-B include 7.6 million variants discovered in 98 Hutterite whole genome sequences. Maximum European allele frequencies were calculated from surveys of the Exome Sequencing Project (ESP), Exome Aggregation Consortium (ExAC) and the 1000 genomes project. **C. Boxplots of MAF by annotation category for the 790,742 REVs included in association studies.** Annotation categories were based on RefSeq¹¹⁹. Numbers in parenthesis correspond to the number of variants in each annotation class.

Supplementary Figure 3.1: Integration of functional annotations to GWAS of blood lipid traits for REVs (continued)

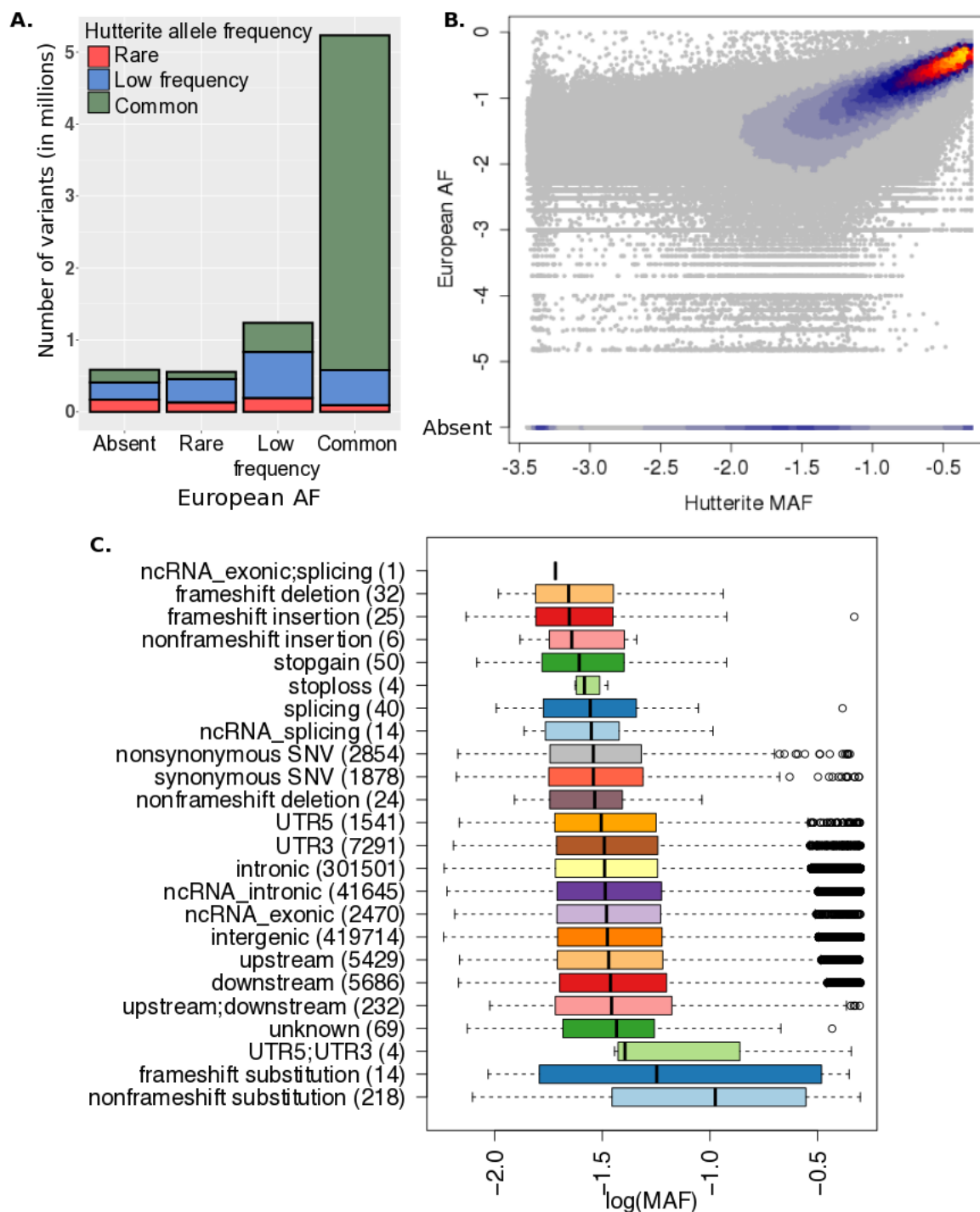


Table 3.1: Sample composition and clinical descriptions

Blood lipid	Sample size (Male/Female)	Mean age, years	Mean mg/dL
Low density lipoprotein cholesterol (LDL-C)	807 (367/440)	36.04 ± 0.546	125.1 ± 2.7
High density lipoprotein cholesterol (HDL-C)	828 (381/447)	36.35 ± 0.545	54.7 ± 1.1
Total cholesterol	828 (381/447)	36.35 ± 0.545	203.9 ± 2.8
Triglycerides (TG)	828 (382/446)	36.31 ± 0.542	121.5 ± 6.1

For each lipid trait, we performed association analyses in two stages. First, we performed single variant analyses using a linear mixed model (GEMMA⁷³), including age and sex as fixed effects and kinship as a random effect. Quantile-quantile plots showed no inflation of test statistics for any of the lipid traits (Figure 3.2) the most significant associations are shown in Table 3.2 and Manhattan plots in Supplementary Figure 3.3. Results for all variants tested are deposited in dbGaP (phs000185). We detected one genome-wide significant locus ($p < 5 \times 10^{-8}$) containing multiple rare variants of large effect associated with increased LDL-C and one genome-wide significant locus with reduced TG levels, both in regions previously linked to lipid traits in GWAS^{9,107}. These two loci include 78 REVs associated with increased LDL-C over a 7.6 Mb region flanking the LDL receptor gene (*LDLR*) on chromosome 19, and 39 REVs associated with reduced TG levels over a 3.8 Mb region flanking the Apolipoprotein C3 (*APOC3*) gene on chromosome 11. The latter includes a known rare splicing variant in *APOC3* (rs138326449)^{120,121}. No genome-wide significant associations with REVs were detected for HDL-C or total cholesterol.

Table 3.2: Genome-wide significant results for the REV GWAS of blood lipid traits

Variants presented are the most significant in each region

Trait	rsID (chr:location)	Minor/ major	MAF	1000G EUR AF	p	OR [95% CI]	Variant type (gene)	No. of significant REVs
LDL-C	rs557778817 19:11305534	T/C	0.032	0.003	1.48×10^{-17}	3.682 [2.75,4.93]	intronic (<i>KANK2</i>)	78
TG	rs184333869 11:117947268	T/C	0.023	0.001	5.41×10^{-13}	0.278 [0.20,0.39]	ncRNA intronic (<i>TMPRSS4-AS1</i>)	39

rsID was annotated with dbSNP142. Location is based on hg19. Direction of effect corresponds to the minor allele.

3.3.2 Application of fgwas to blood lipid traits

In second stage analyses, we annotated all variants discovered in the Hutterites using 25 sequence-based annotations and 332 functional annotations (see Methods) and applied the functional GWAS (fgwas)¹²² framework to evaluate the GWAS results with these annotations. We first investigated the enrichment of each functional category within each REV GWAS, and then performed hierarchical modeling to prioritize candidate rare variants based on prior functional knowledge. We split the genome into blocks averaging 125Kb (~50 REVs) and ran single annotation models against each of the four lipid traits. Annotations for enrichments where the 95% confidence intervals (CI) did not intersect 0 are shown in Supplementary Figure 3.1. For LDL-C, the annotation that most improved the probability of association between a genomic region and the trait was for regions of open chromatin defined by DNase I digital genomic footprinting (DGF)¹²³ ($\log_2$ fold increase 0.74 [0.23, 1.23]). For TG, the best annotation was within regions containing skeletal muscle cell type-specific histone marks (H3K4me1, H3K4me3, H3K9ac and H3K27ac) from the Epigenomics Roadmap Project merged by Finucane *et al.* (2015)¹²⁴ ($\log_2$ fold decrease -28.19 [-48.19, -0.86]). Lastly, For HDL-C and total cholesterol, the categories that best improved the

probability of association were enhancer regions within induced pluripotent stem cells (iPS DF 19.11 cells; $\log_2$ fold increase 34.57 [2.11, 54.58]) and synonymous variants ($\log_2$ fold increase 37.53 [0.32, 57.53]), respectively. It's important to mention that some of the maximum likelihood estimates (MLE) of the enrichment effects for some of the annotations have high uncertainty, with high point estimates and large confidence intervals. These enrichments are largely driven by a few number of very significant variants at a single locus.

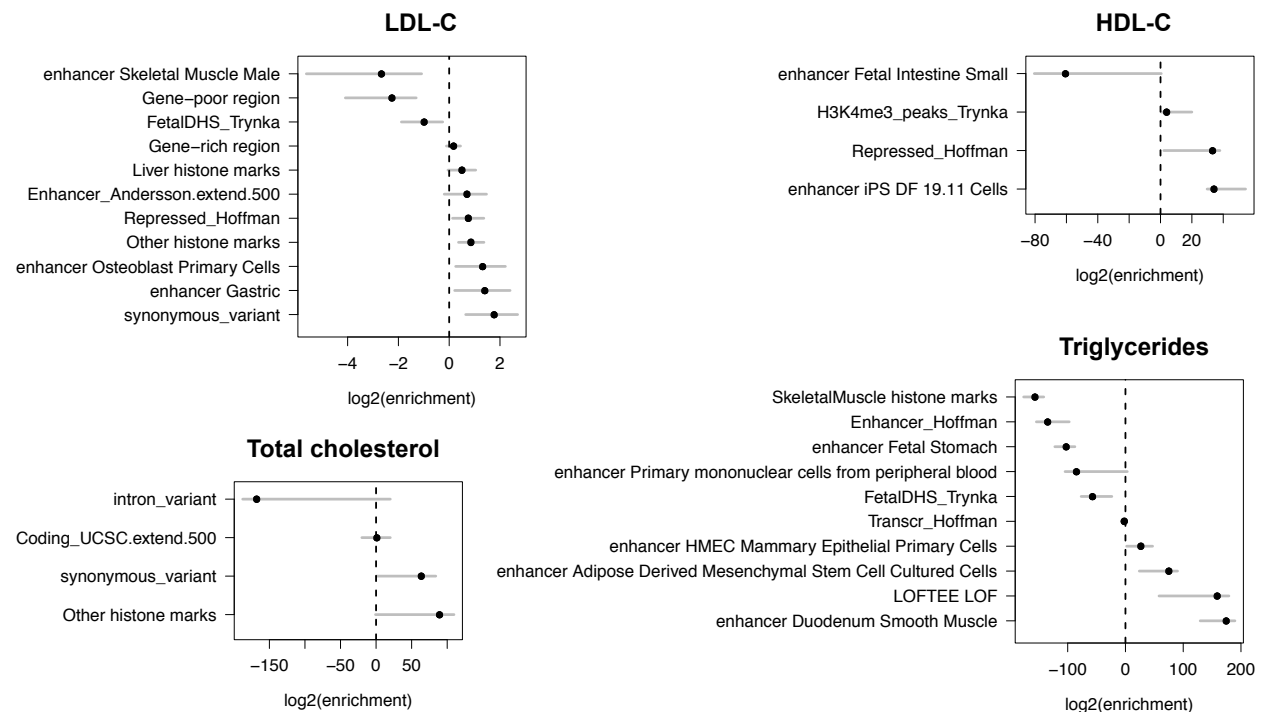
With this caveat in mind, we performed forward selection and built fgwas models that combined effects from multiple annotations and applied cross validation to avoid overfitting while maximizing the likelihood of each model. Figure 3.2 shows the maximum likelihood estimates (MLE) and 95% CI of the enrichment effects for the selected annotations in each of the final joint models for each of the blood lipid traits and annotation descriptions for each model are provided in Supplementary Table 3.1-Supplementary Table 3.4.

The joint model for each fgwas estimated both the probability that each block contains a causal variant and the posterior probability that each variant is causal conditional on the presence of one causal variant in the region. Variants with the largest posterior probabilities of causality will tend to have the most significant p-values and functional annotations that predict association genome-wide. We weighted our GWAS results based on the fgwas joint model and applied a regional prior probability of association (PPA) threshold of 0.75. Overall, we identified 95 consecutive blocks associated with LDL-C on chromosome 19 and 26 blocks on chromosome 11 associated with TG, all which harbor REVs that passed genome-wide significance in the original

GWAS ($p < 5 \times 10^{-8}$), plus additional blocks associated with HDL-C on chromosome 3, and with LDL-C in chromosomes 11 and 16 (Supplementary Table 3.5).

Figure 3.2: Joint fgwas models for blood lipid traits

Models were fit combining multiple annotations using the fgwas hierarchical methodology, as described in Pickrell *et al.* (2014)¹²². The MLE and 95% CI of the enrichment effects of each annotation in the final model (modeling performed using penalized likelihood) are shown.



LDL-C

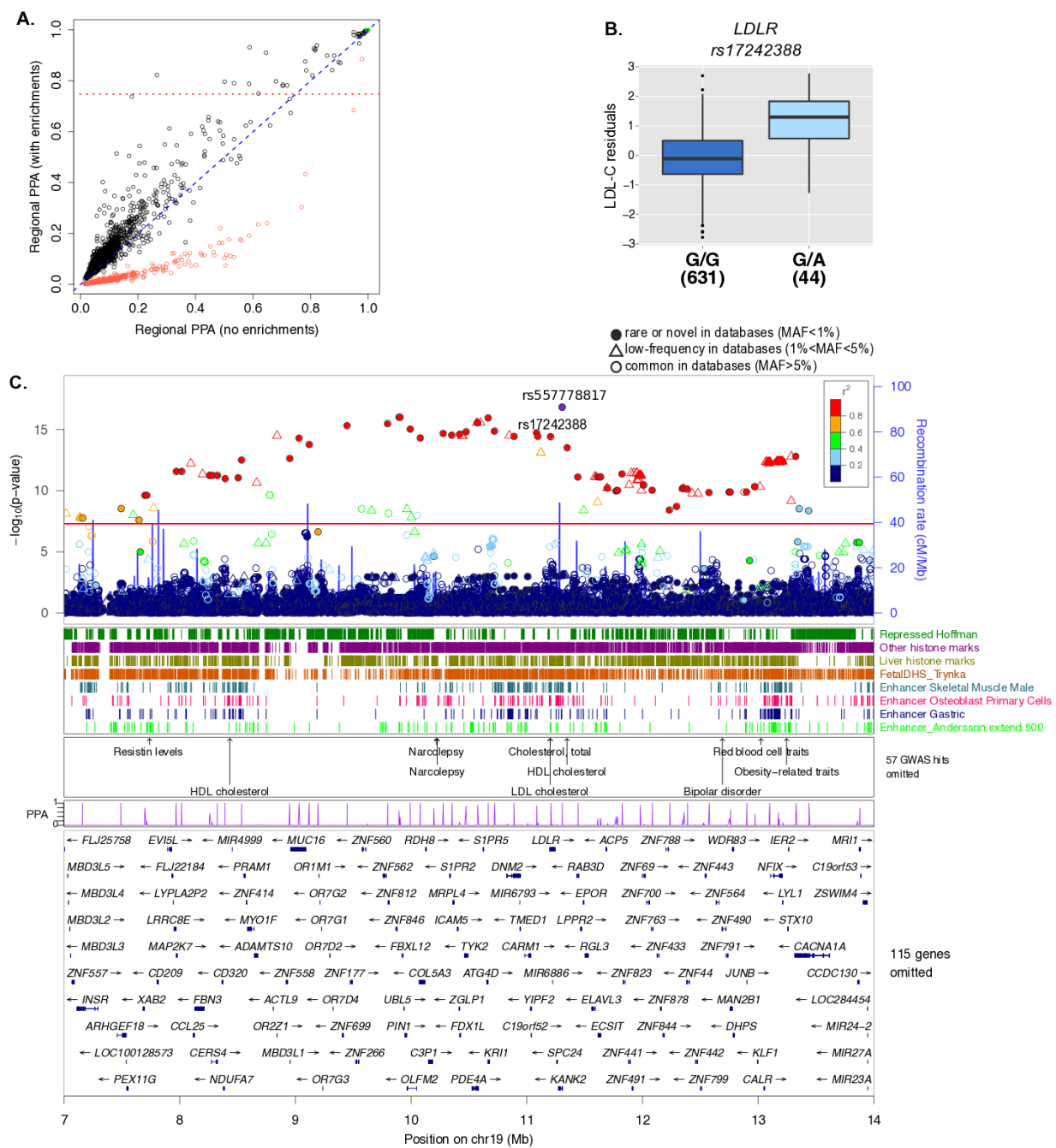
The GWAS of LDL-C identified 78 genome-significant REVs on a haplotype that spans a 7.6Mb region on chromosome 19p13.2. Despite the high LD and resulting long haplotype, we were able to prioritize 18 candidate variants that had fgwas posterior probabilities equal to 1 (PPA=1; Supplementary Table 3.5). Among these 18 variants was one located in the first intron on the *LDLR*

that was associated with increased levels of LDL-C (rs17242388; logBF=28.02; MAF=3%; 1000 genomes EUR MAF=0.6%). rs17242388 was selected as a candidate variant because it had the highest number of annotations (six) in the model, which included enhancer regions defined by Andersson *et al.* (2014)¹²⁵, putative enhancer regions in gastric tissue, skeletal muscle and osteoblast primary cells in the Roadmap Epigenomics¹²⁶, and cell type specific histone marks in liver and tissues labelled “other” by Finucane *et al.* (2015)¹²⁴ (see Supplementary Table 3.1 for model annotation descriptions).

Figure 3.3: Rare variants on chromosome 19 are associated with LDL-C

A. Reweighted GWAS. GWAS results were reweighted using the joint model presented in Figure 3.2. Each point represents a region of the genome and its corresponding posterior probability of association (PPA) in a model with and without enrichments. The dotted red line represents regional PPA > 0.75 in the enriched model. The red points correspond to regions in the bottom third of gene density and the green points are regions containing GWAS SNP(s) with $p < 5 \times 10^{-8}$. **B. LDL-C genotype boxplots for candidate LDLR intronic variant (rs17242388; $p = 3.89 \times 10^{-15}$).** **C. Locus plot.** The top panel shows the p-values of association with LDL-C for all variants discovered in the Hutterites regardless of allele frequency in Europeans. Symbols correspond to the maximum allele frequency in Europeans, with closed circles representing REVs (see legend), and are colored based on their LD r^2 with the most associated variant in the region (rs557778817). The next panel shows tracks for the annotations selected in the fgwas joint model. The next panel shows annotations of known GWAS loci from NHGRI (a full table of GWAS associations in this region is provided in Supplementary Table 3.6). The next panel illustrates the estimated PPA of being causal for each variant in the reweighted fgwas (conditional on there being one causal variant in the region). The bottom-most panel shows the genes in the region.

Figure 3.3: Rare variants on chromosome 19 are associated with LDL-C (continued)



Triglycerides

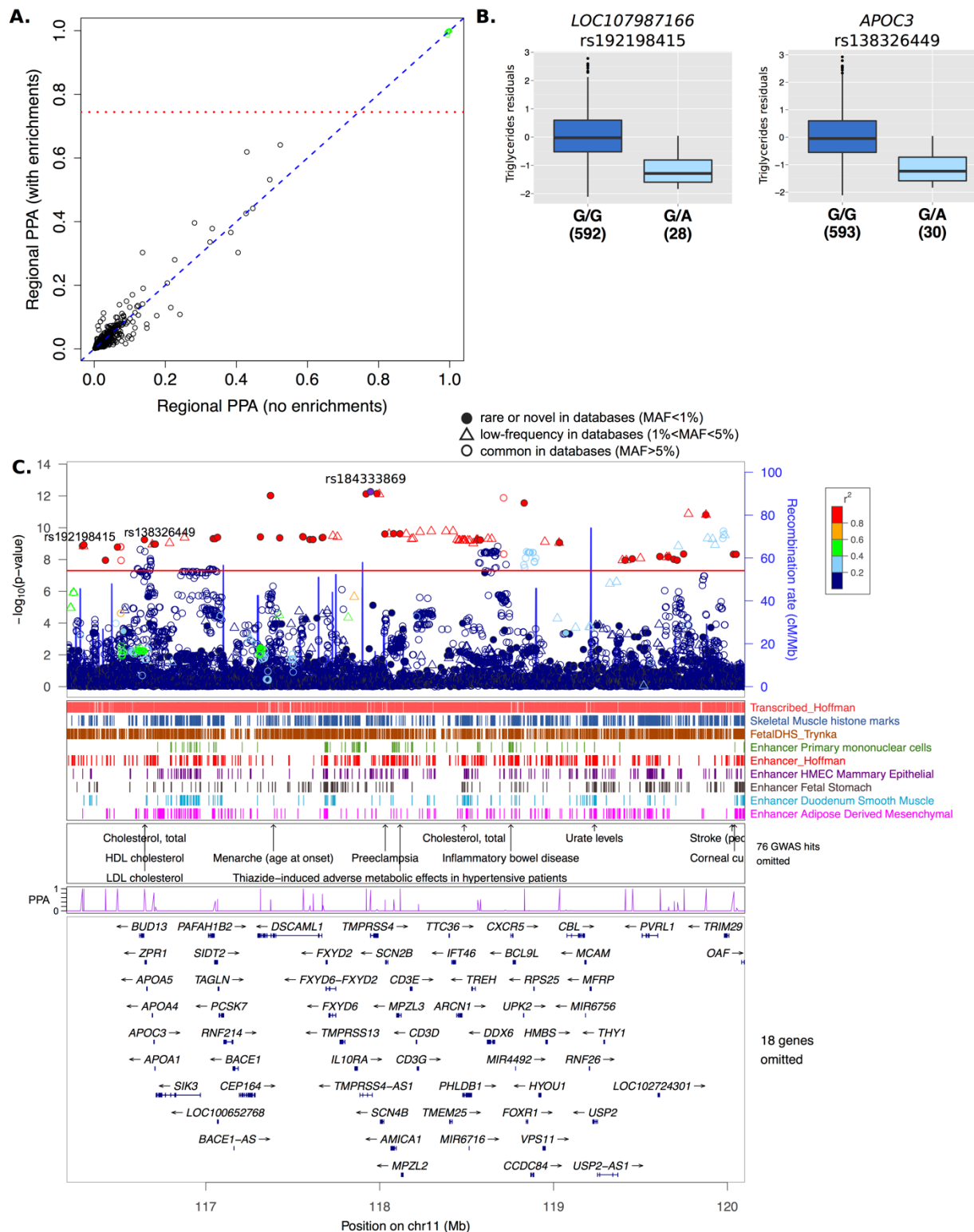
Out of the 39 REVs on a 3.8Mb haplotype on chromosome 11q23.3 associated with TG levels, 19 were potentially causal with posterior probabilities of association greater than 75%. The SNP with the highest posterior probability (PPA=1; logBF=22.49), rs149157643 (MAF=2.38%; 1000 genomes EUR MAF=0.7%) is an intronic variant within a non-coding RNA (ncRNA) gene, *TMPRSS4-AS1*, and is 25Kb away from the most associated variant in the region (rs184333869; PPA=0.93; MAF=2.40%; 1000 genomes EUR MAF=0.1%). Two of the 19 variants are within regions containing the maximum number of functional annotations selected by the joint model at this locus. The first variant, rs192198415, is an intronic variant in another noncoding RNA gene *LOC107987166* (PPA=0.999; logBF=15.26 ; MAF=2.06%; 1000 genomes EUR MAF=0.4%) that lies within a putative enhancer region in duodenum smooth muscle and within a DNase I hypersensitivity region in fetal tissues (FetalDHS_Trynka)¹²⁷. The second variant (349Kb downstream from rs192198415) is a novel 9 bp insertion private to the Hutterites, located between genes BUD13 Homolog (*BUD13*) and ZPR1 Zinc Finger (*ZPR1*; chr11:116646774/CCCTTCATA, PPA=0.99; logBF=15.98; MAF=2.24%). The *BUD13-APOC3* region is a well replicated GWAS locus for multiple lipid traits^{107,128} (Figure 3.4 and Supplementary Table 3.7) with several implicated rare variants, including loss-of-function mutations in *APOC3*^{44,120,121} associated with decreased TG levels. Among these is a splicing variant in *APOC3* (rs138326449; MAF=2.23%; 1000 genomes MAF=0.3%) that is associated with TG in the Hutterites, but with a slightly lower posterior probability of association in the model (PPA=0.82; logBF=15.38) compared to other variants in the region. To our knowledge, the role of this splicing *APOC3* polymorphism (rs138326449) in regulation of plasma lipids has not been characterized

thus far. Thus, while there may be compelling evidence that reduced plasma levels of APOC3 protein results in lower TG levels^{44,120}, there may be multiple rare variants within the haplotype influencing TG levels in the Hutterites.

Figure 3.4: Rare variants on chromosome 11 are associated with TG levels

A. Reweighted GWAS. GWAS results were reweighted using the joint model presented in Figure 3.2. Each point represents a region of the genome and its corresponding posterior probability of association (PPA) in a model with and without enrichments. The dotted red line represents regional PPA > 0.75 in the enriched model. **B. TG genotype boxplots for candidate non-coding RNA intronic variant in LOC107987166 (*rs192198415*; $p=1.19 \times 10^{-9}$) and previously reported splicing variant in APOC3 (*rs138326449*; $p=1.08 \times 10^{-9}$).** **C. Locus plot.** The top panel shows the p-values of association with TG for all variants discovered in the Hutterites regardless of allele frequency in Europeans. Symbols correspond to the maximum allele frequency in Europeans, with closed circles representing REVs (see legend), and are colored based on their LD r^2 with the most associated variant in the region (*rs184333869*). The next panel shows tracks for the annotations selected in the fgwas joint model. The next panel shows annotations of known GWAS loci from NHGRI (a full table of GWAS associations in this region is provided in Supplementary Table 3.7). The next panel illustrates the estimated PPA of being causal for each variant in the reweighted fgwas (conditional on there being one causal variant in the region). The bottom-most panel shows the genes in the region.

Figure 3.4: Rare variants on chromosome 11 are associated with TG levels (continued)



Although the HDL-C GWAS did not identify any genome-wide significant REVs, the fgwas revealed two loci associated with HDL-C with regional PPA greater than 90%. The first locus was associated with increased HDL-C levels and was the same haplotype on chromosome 11q23.3 associated with TG levels. The most likely variant was the most significant SNP in the HDL-C GWAS (rs184333869; logBF=9.03329; PPA=0.95; Figure 3.5) and was also the most significant signal in the TG GWAS. The second association was a 3.8Mb region on 16q13 associated with decreased HDL-C levels within one of the most replicated loci for HDL-C and increased cardiovascular disease risk^{107,129} (rs3764261, a variant upstream of Cholesteryl Ester Transfer Protein [*CETP*]). The variant with the highest posterior probability of association was an intronic variant in Solute Carrier Family 12 Member 3 (*SLC12A3*; rs141938931; MAF=5.2%; 1000 genomes EUR MAF=0.2%; logBF=6.58; PPA=0.75). While *SLC12A3* has not been directly linked to HDL-C levels, rs141938931 is located 91Kb upstream of the known *CETP* GWAS variant (rs3764261; LD $r^2=0.04$), a gene encoding a plasma protein that supports the transport of cholesteryl esters and triglycerides between the lipoproteins¹³⁰.

Figure 3.5: Rare variants on chromosomes 11 and 19 are associated with HDL-C

A. Reweighted GWAS GWAS results were reweighted using the joint model presented in Figure 3.2. Each point represents a region of the genome and its corresponding posterior probability of association (PPA) in a model with and without enrichments. The dotted red line represents regional PPA > 0.75 in the enriched model. **B. Genotype boxplots for HDL-C levels for candidate non-coding RNA intronic variant in TMPRSS4-AS1 (rs184333869; $p=1.10 \times 10^{-6}$).** **C. Genotype boxplots for HDL-C levels and for SLC12A3 expression in LCLs for candidate intronic variant in SLC12A3 (rs141938931; $p=2.26 \times 10^{-5}$ and $p=4.0 \times 10^{-4}$, respectively).** **D and E. Locus plots.** The top panel shows the p-values of association with TG for all variants discovered in the Hutterites regardless of allele frequency in Europeans. Symbols correspond to the maximum allele frequency in Europeans, with closed circles representing REVs (see legend), and are colored based on their LD r^2 with the most associated variant in the region (rs184333869 and rs189679427, respectively). The next panel shows tracks for the annotations selected in the fgwas joint model. The next panel shows annotations of known GWAS loci from NHGRI (a full table of GWAS associations in this region is provided in Supplementary Table 3.8). The next panel illustrates the estimated PPA of being causal for each variant in the reweighted fgwas (conditional on there being one causal variant in the region). The bottom-most panel shows the genes in the region.

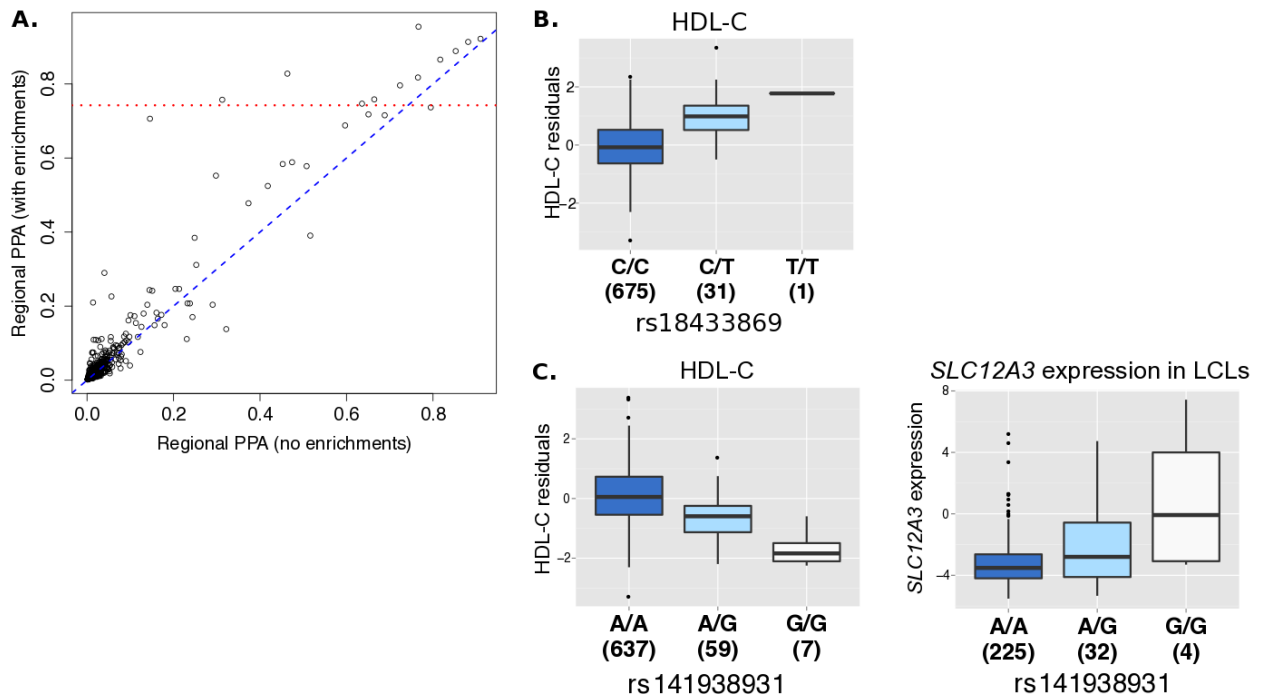


Figure 3.5: Rare variants on chromosomes 11 and 19 are associated with HDL-C (continued)

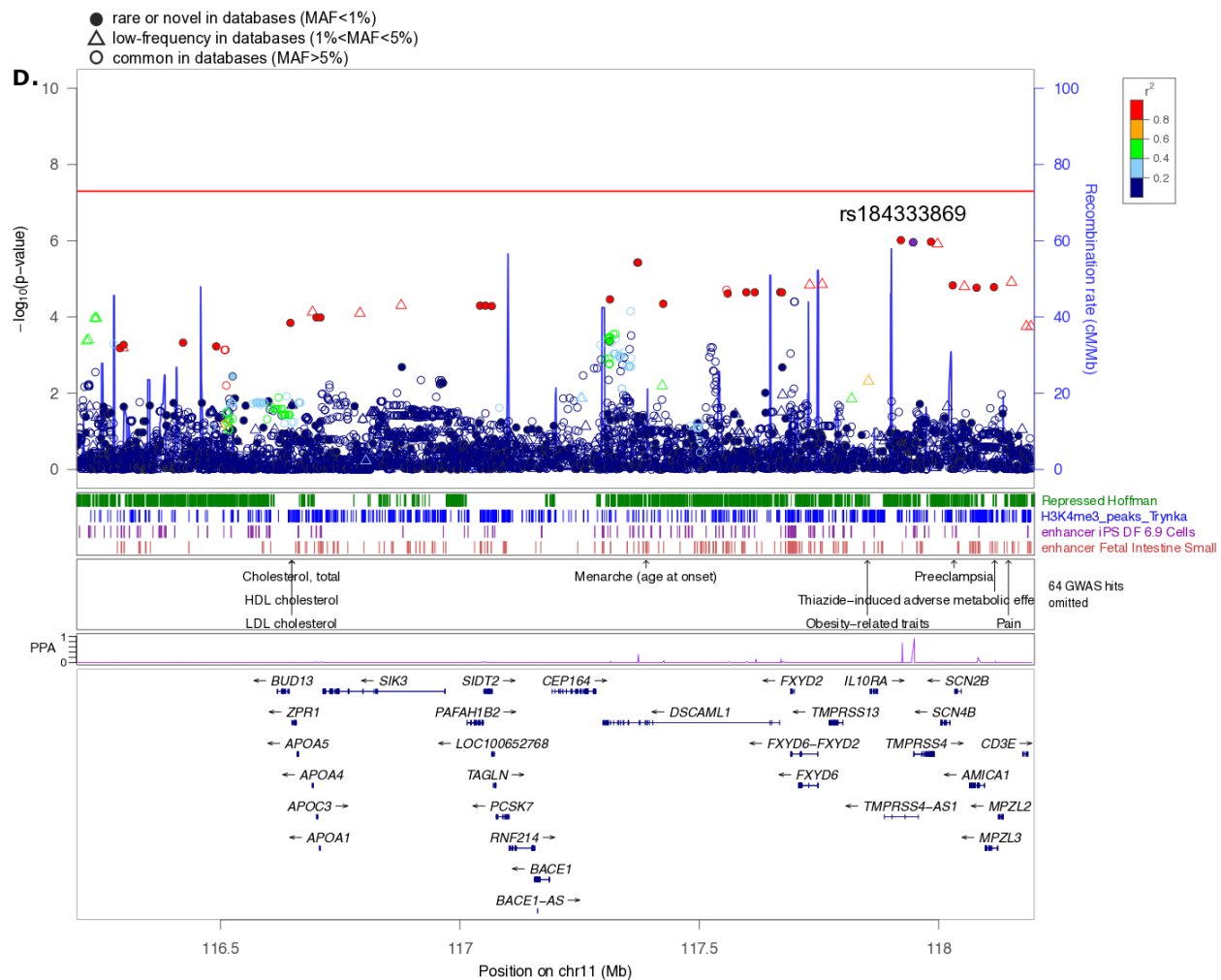
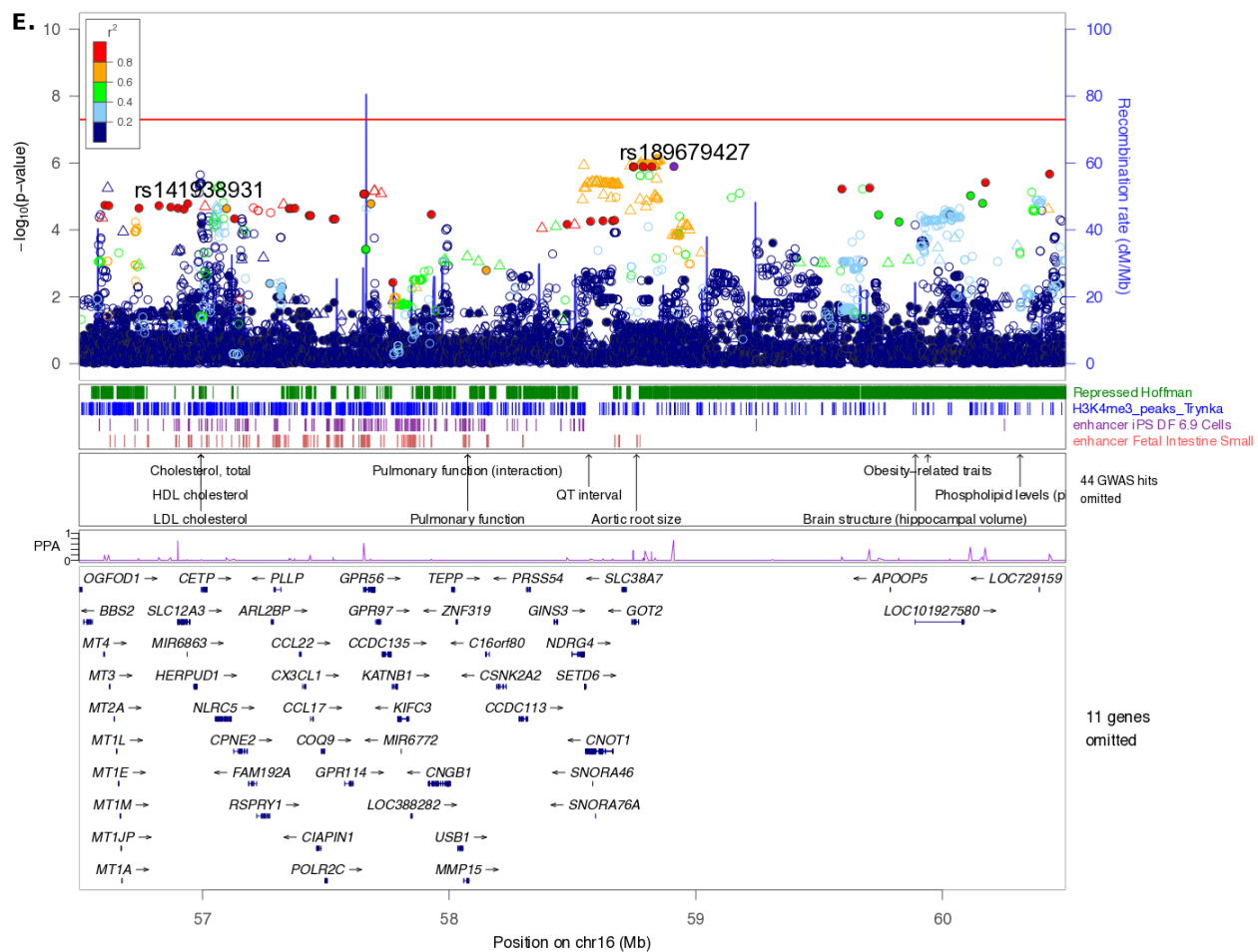


Figure 3.5: Rare variants on chromosomes 11 and 19 are associated with HDL-C (continued)



3.3.3 Conditional analyses and follow up with gene expression in LCLs

For each significant trait (LDL-C, TG, HDL-C), we performed two sets of conditional analyses including all variants present in the Hutterites genomes within each of the associated regions regardless of their allele frequencies in Europeans. First, we conditioned on the most significant rare variant both to assess whether other (rare or common) variants in the region contribute to the association signal, and to search for additional variants at this locus whose effects could have been masked by the rare variant. When available, we also conditioned on previous GWAS variants identified to verify that the rare variant signal in our study is independent of previous known associations at this locus (Table 3.3). Second, we checked for evidence of regulatory effects of associated variants using gene expression data in lymphoblastoid cell lines (LCLs) from the Hutterites¹³¹. Although LCLs have known limitations, shared genetic effects on gene expression across multiple tissues have been demonstrated¹³². Importantly, however, our focus on private or rare variation in the Hutterites makes it impossible to utilize publicly available eQTL databases in other tissues to assess our candidate variants.

Table 3.3: Summary results for conditional analyses

A. Conditional analysis on most significant rare variant identified in this study. Listed variants represent the lead variant among common variants that increased in significance after taking conditioning on the rare variants. **B. Conditional analysis on GWAS variant in the region.** Candidate causal variants nearby replicable GWAS variants remain significant after taking into account the effects from known common variation.

A. Conditional analysis on the most significant rare variant								
Trait	Conditioning variant	rsID	Minor/ major	MAF	p _{gwas}	p _{cond}	OR _{cond}	Variant type (gene)
LDL-C	rs557778817	rs513663	G/A	0.282	0.001	3.03 x10 ⁻⁶	1.30 [1.17,1.45]	intronic (PDE4A)

Table 3.3 (continued)

TG	rs184333869	rs11604424 ^a	C/T	0.225	0.0002	9.50 x10 ⁻⁹	1.43 [1.27,1.61]	downstream (<i>ZPR1</i>)
HDL-C	rs189664208	rs4783961 ^a	G/A	0.468	0.002	1.66 x10 ⁻⁵	0.77 [0.69,0.86]	upstream (<i>CETP</i>)
B. Conditional analysis on previously reported GWAS variant								
Trait	Conditioning variant (p_{gwas})	Candidate variant	Minor/ major	MAF	p_{gwas}	p_{cond}	OR _{cond}	Variant type (gene)
LDL-C	rs6511720 ^a ($p=0.009$)	rs17242388	A/G	0.03	3.89x10 ⁻¹⁵	1.31x10 ⁻¹⁵	3.57 [2.63,4.84]	intronic (<i>LDLR</i>)
TG	rs964184 ^{a,b} ($p=3.13 \times 10^{-9}$)	rs138326449	A/G	0.022	1.08x10 ⁻⁹	9.85 x10 ⁻⁹	0.35 [0.25,0.50]	splicing (<i>APOC3</i>)
HDL-C	rs964184 ^a ($p=0.2$)	rs184333869	T/C	0.024	1.10x10 ⁻⁶	1.63 x10 ⁻⁶	2.41 [1.69,3.46]	ncRNA intronic (<i>TMPRSS4-AS1</i>)
HDL-C	rs17231506 ^a ($p=3.85 \times 10^{-6}$)	rs141938931	G/A	0.052	2.26x10 ⁻⁵	6.55 x10 ⁻⁵	0.58 [0.46,0.76]	intronic (<i>SLC12A3</i>)
rsID was annotated with dbSNP142. Direction of effect corresponds to minor allele. Superscripts on rsIDs correspond to study where variant was reported as significant. ^a Global Lipids Genetics Consortium <i>et al.</i> Discovery and refinement of loci associated with lipid levels. <i>Nat. Genet.</i> 45 , 1274–1283 (2013). ^b Teslovich, T. M. <i>et al.</i> Biological, clinical and population relevance of 95 loci for blood lipids. <i>Nature</i> 466 , 707–713 (2010).								

Conditioning on the lead rare variant on chromosome 19 in the LDL-C analysis (rs557778817), revealed a novel common variant in the third intron of Phosphodiesterase 4A (*PDE4A*; rs513663) that reached suggestive significance after removing the effects of the rare variant ($p_{\text{gwas}}=0.001$; $p_{\text{rs557778817}}=3.0 \times 10^{-6}$; Figure 3.6A). *PDE4A* plays a key role in many physiological process by regulating levels of the cAMP, a mediator of response to extracellular signals¹³³, but to our knowledge variation in LD with this variant has not been previously linked to lipid traits. By phasing the *PDE4A* common (rs513663) and the *LDLR* rare variants (rs1724388) in the Hutterites, we show that the alleles at each variant associated with higher LDL-C are on separate haplotypes, with additive effects on lowering *LDLR* gene expression ($p=0.008$) and increasing levels of LDL-C (Figure 3.6B; $p=2.4 \times 10^{-8}$). In the second conditional analysis, we conditioned on a commonly replicated variant in the *LDLR* gene that is associated with decreased

LDL-C levels and lower risk for coronary heart disease (CHD)^{9,107,134} (rs6511720; p=0.009 in the Hutterites) and confirmed that the identified rare variants in this study have independent and opposite effects compared to the common GWAS variant at this locus.

Figure 3.6: LDL-C conditional analysis on top rare variant (rs557778817)

A. Locus plot. Conditional analysis on rs557778817 identified novel common variants with suggestive significance. **B. LDL-C and LCL gene expression haplotype boxplots.** Alleles were phased for the top common signal identified in conditional analysis (rs513663; top) and the candidate *LDLR* rare intronic variant identified in fgwas (rs17242388; bottom). The trends observed between each of five haplotype combinations present in our sample suggest these variants have additive effects that lower *LDLR* expression and increase LDL-C levels.

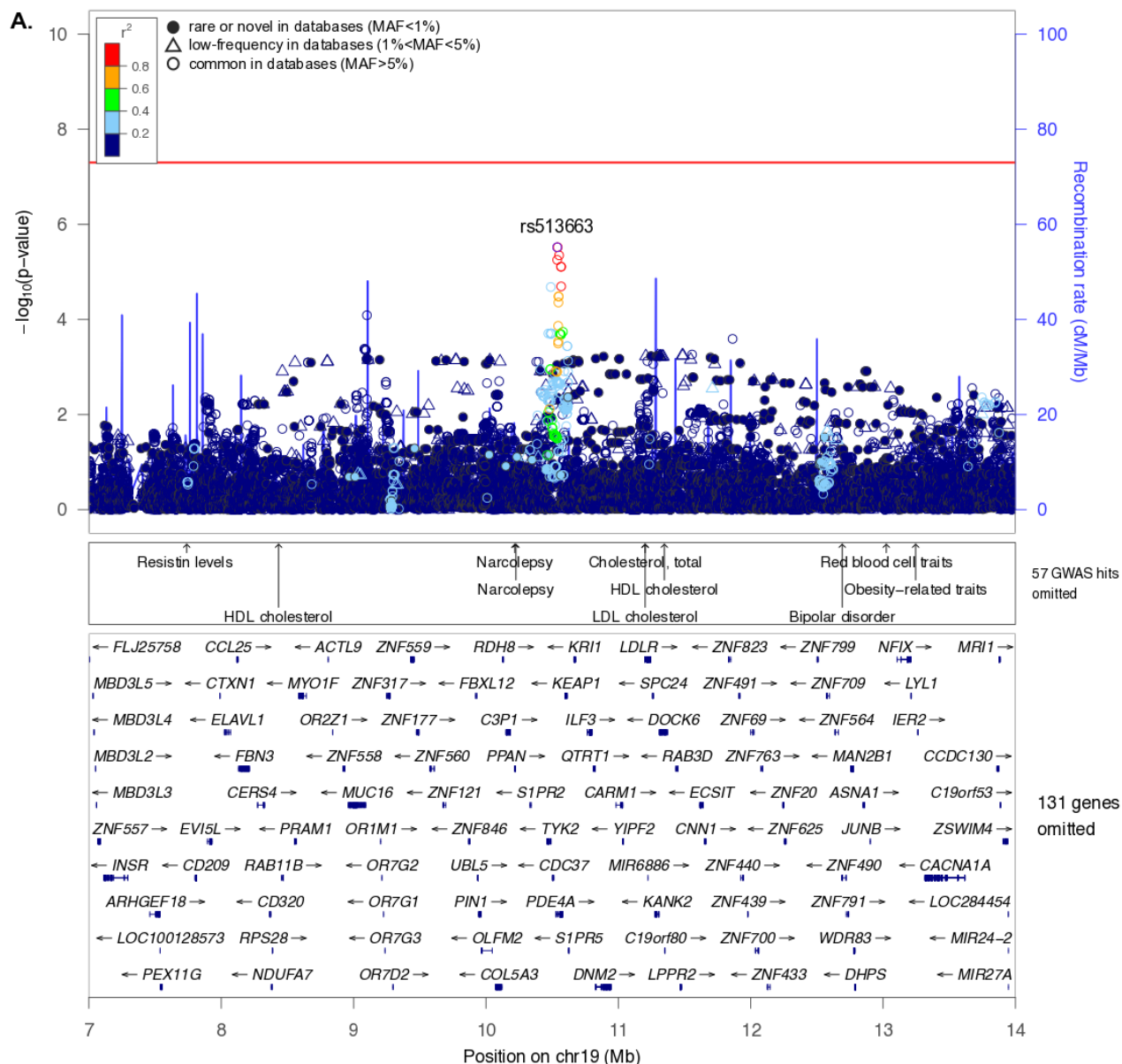
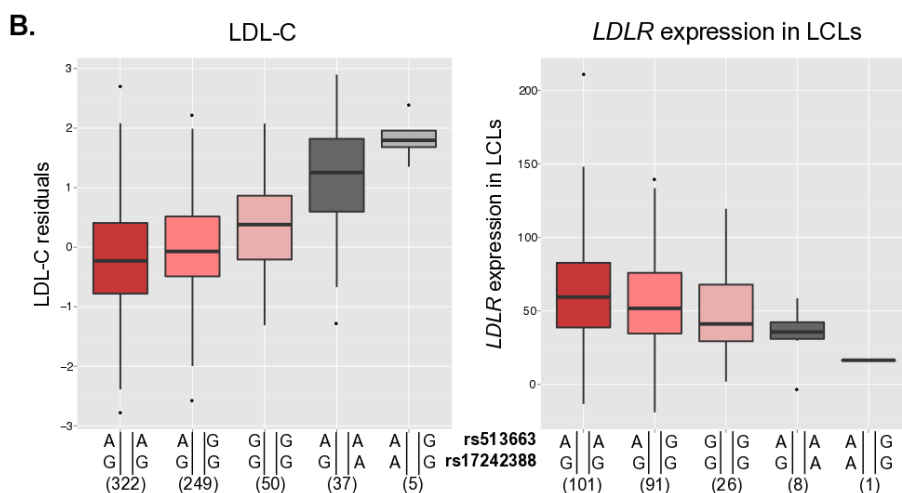


Figure 3.6: LDL-C conditional analysis on top rare variant (rs557778817) (continued)



Conditional analyses on the lead rare variant on chromosome 11 (rs184333869) associated with decreased TG levels uncovered associations with known common variation in the *BUD13*-*APOC3* locus that is associated with increased TG levels. The effects of this variant were masked by the opposite effects of the rare variant in the Hutterites ($p_{\text{gwas}}=0.0002$; $p_{\text{conditional}}=9.50 \times 10^{-9}$). Gene expression analyses showed a suggestive association between the top rare variant in *TMPRSS4-AS1* (rs184333869) with lower *TMPRSS13* expression ($p=0.01$) and between the splicing variant in *APOC3* (rs138326449) and higher expression of *BUD13* ($p=0.05$), but their direct effects on the apolipoprotein genes could not be assessed due to low detection of these gene in LCLs. Similarly, HDL-C conditional analyses on chromosomes 11 and 16 confirmed that the rare variant associations at these loci have independent effects from known common variants associated with HDL-C. Gene expression studies of chromosome 16 associations revealed that the candidate variant in *SLC12A3* (rs141938931) that is associated with lower HDL-C levels is associated with higher *SLC12A3* expression ($p=4 \times 10^{-4}$; Figure 3.5C), but showed no effect on the lipid metabolism gene *CETP*. Overall, our results provide evidence that rare variants associated with lipid traits are likely to mediate their effects by modulating changes in gene expression.

3.4 Discussion

We performed GWAS with ~790k rare in European variants that have drifted to higher frequencies in the Hutterites, identifying four novel associations with blood lipid traits as well as replicating the effects of the known *APOC3* splicing variant^{120,121}. While the increased frequencies of these alleles in the Hutterites provided sufficient power to identify these loci in GWAS, the long stretches of LD resulted in associations with many rare variants segregating on the same haplotype and posed challenges for pinpointing the causal variant. To increase resolution and prioritize candidate variants based on their likelihood to influence each trait, we applied a statistical fine mapping approach (fgwas¹²²) by jointly incorporating functional data to our marginal GWAS results. Application of the fgwas significantly improved our ability to narrow the subset of likely causal variants at each locus.

Only half of the six rare variants identified by GWAS in our study resided within known lipid loci identified by GWAS loci. However, even at those known loci, the associated rare variants in our study were independent of the known common variation in these regions. Among these was a variant in the first intron of the *LDLR* gene (rs17242388) that was associated with increased LDL-C levels in our study. Variation in the first intron of the *LDLR* has been previously implicated in regulating LDL-C levels in two separate studies, but with opposite effects to the rare variant in our study. The known variation at this locus includes a predominantly European variant 821 bp upstream associated with lower non-HDL-C levels in the Icelandic population (rs17248748; MAF=3.4%; and absent in the Hutterites) and a common variant 4.5 Kb upstream linked to multiple blood lipid traits^{9,135} (rs6511720; MAF= 15.3% and p=0.009 in the Hutterites). The *LDLR* variant in

the Hutterites drifted to a frequency five times higher than reported in 1000 genomes, and is located within a predicted enhancer region in a number of digestive tissues, including liver, small intestine and stomach mucosa. Conditional analyses revealed that the rare rs17242388-A allele and a common *PDE4D* allele (rs513663-G) occur on different haplotypes and have independent and additive effects on lowering expression of the *LDLR* gene and raising plasma LDL-C levels.

A second association with a novel rare variant also resides within a known lipid trait locus: an intronic variant in *SLC12A3* (rs141938931) on chromosome 16 is associated with lower levels of HDL-C in the Hutterites. *SLC12A3* is a sodium chloride solute transport protein expressed mainly in the kidney and located directly upstream of a well-established lipid metabolism plasma protein, *CETP*. *CETP* is known to mediate the transfer of esterified cholesterol from HDL-C to apoB-containing particles in exchange for TG¹³⁰. The rs141938931 allele associated with lower HDL-C was also associated higher expression of the *SLC12A3* gene, but not the *CETP* gene even though it was expressed in Hutterite LCLs. This suggests that *SLC12A3* may be directly involved in regulation of HDL-C levels, although gene expression studies in more relevant tissues are required to confirm this observation.

The association between variants at 16q13 and lower HDL-C provide support for a shared genetic architecture between Mendelian and complex traits. The variants associated at this locus are centered around *ARL6*, a causal gene for Bardet-Biedl syndrome¹³⁶, a highly penetrant oligogenic disorder that results in a number of clinical phenotypes, including childhood obesity and hyperlipidemia in the majority of cases¹³⁷. Many genes identified by GWAS of lipid traits have been shown to underlie Mendelian disorders of lipid metabolism¹³⁸⁻¹⁴⁰. Mutations within these

genes (coding and non-coding) provide complimentary viewpoints to the disease mechanisms influencing these trait, but further work is necessary to understand the molecular basis of how the variation associated in GWAS (rare and common) may be influencing these traits through gene expression. Overall, our results are consistent with previous findings that complex diseases are enriched for loci implicated in Mendelian traits and showed shared comorbidities between both trait classes¹⁴¹.

In summary, our findings validate the advantages of population isolates in the search for rare variants associated with complex traits. Importantly, all of the associated variation revealed in our study is within non-coding sequences and many were associated with gene expression, highlighting the importance of studying the effects rare non-coding variation on gene expression as well as on their relationship to common diseases. An inherent limitation of this study, and most studies of rare variants, is the challenge to replicate findings due to the very low frequency of the alleles under investigation in most populations. For example, the rare and low-frequency variants surveyed in the Global Blood Lipids Consortium meta-analysis⁹ are primarily loss-of-function coding variation and had no overlap with the significant variants in our study. Nonetheless, the discoveries revealed in this study, even those that may be private to Hutterites, uncovered potentially novel disease genes, and highlight new clinically relevant pathways that could point toward novel therapeutic targets for hypercholesterolemia and lowering risk for cardiovascular disease.

3.5 Materials and Methods

3.5.1 Study sample

Standard fasting blood lipid measurements were collected as part of a larger study of complex traits in the Hutterites (see Cusanovich *et al.* (2016)¹³¹ for details). Briefly, blood samples were collected after an overnight fast from 828 Hutterites (ages 14 to 85 years; Table 3.1) during field trips to Hutterite colonies in 1996–1997, and 2006–2009. Plasma levels for LDL-C, HDL-C, TG and total cholesterol were measured. Subjects receiving anti-hypercholesterimia medication, hormone replacement therapy, birth control or diagnosed with sitosterolemia¹⁴² were excluded from the study. For subsequent analyses, we applied a cubed root transformation to absolute LDL-C and HDL-C levels, a natural log transformation to TG and total cholesterol and applied a correction for age and sex.

3.5.2 Genotyping

We used PRIMAL⁴⁹, an in-house pedigree-based imputation algorithm, to phase and impute 7,605,123 variants discovered in 98 whole genome sequences to 1,317 Hutterites who were previously genotyped on Affymetrix arrays^{143–145}. The genotype accuracy of PRIMAL in the Hutterites was >99%, although the average genotype was 87.3% due to the variation in IBD sharing across the genome of individuals with the 98 sequenced Hutterites. Within individuals, genotype accuracy was uncorrelated with genotype call rates. See Livne *et al.* (2015)⁴⁹ for additional details.

3.5.3 Single variant analyses

We focused our studies on the 6,781,373 variants with genotypes called in at least 400 individuals and MAF > 1% in the Hutterites. From these, we selected the subset of 790,742 variants that were absent in European populations in ExAC¹⁴⁶, ESP¹⁴ or 1000 genomes¹⁴⁷ or occurred at frequencies <1%. We refer to these variants as REVs throughout the text. To test the effect of REVs on each of the blood lipid trait, we used a linear mixed model as implemented in GEMMA⁷³, applying age and sex as fixed effects and kinship as a random effect to correct for the relatedness between the individuals in our sample. Causal variants were then prioritized based on functional annotations as implemented in fgwas¹²². Follow-up conditional and LCL eQTL analyses were also performed using GEMMA. LCLs eQTL studies were performed in 441 Hutterite individuals (317 of which are in our lipid studies) that were collected as part of a separate study in the Hutterites¹³¹.

3.5.4 Variant annotation

We obtained variant annotations from dbSNP, ENSEMBL, LOFTEE, conservation and functional scores (e.g. CADD, GERP, PolyPhen, SIFT), and allele frequencies from European populations (ExAC¹⁴⁶, ESP¹⁴, 1000 genomes¹⁴⁷) using Variant Effect Predictor (VEP)¹⁴⁸. We downloaded promoter and enhancer annotations created by the Epigenomics Roadmap Project ($-\log_{10}(p) \geq 10$; http://www.broadinstitute.org/~meuleman/reg2map/HoneyBadger2_release/) for 127 cell types/tissues and removed from this analysis the seven cell types annotated with quality warnings. We directly annotated variants using the ClinVar database downloaded on 08/07/2016 and selected variants labelled as pathogenic or likely pathogenic. Lastly, we used the same set of 53 functional categories and 9 cell-type specific histone marks regions obtained from Finucane *et al.*

(2015) (<https://data.broadinstitute.org/alkesgroup/LDSCORE/>). Briefly, the annotations include annotations for RefSeq, digital genomic footprint and transcription factor binding sites from ENCODE¹²³, combined chromHMM and Segway annotations for six cell lines¹⁴⁹, processed DHSs data from ENCODE and Roadmap Epigenomics data and cell type specific H3K4me1, H3K4me3 and H3K9ac data from Roadmap Epigenomics¹²⁷, H3K27ac from Roadmap Epigenomics and from Hnisz *et al.* (2013)¹⁵⁰, super-enhancers from Hnisz *et al.* (2013)¹⁵⁰, processed conserved regions in mammals from Lindblad-Toh *et al.* (2011)^{151,152} and FANTOM5 enhancers¹²⁵. For each functional category Finucane *et al.* (2015), a 500-bp window was added as an additional category. For each DHS, H3K4me1, H3K4me3 and H3K9ac sites, a 100-bp window around the ChIP-seq peak was added as an additional category.

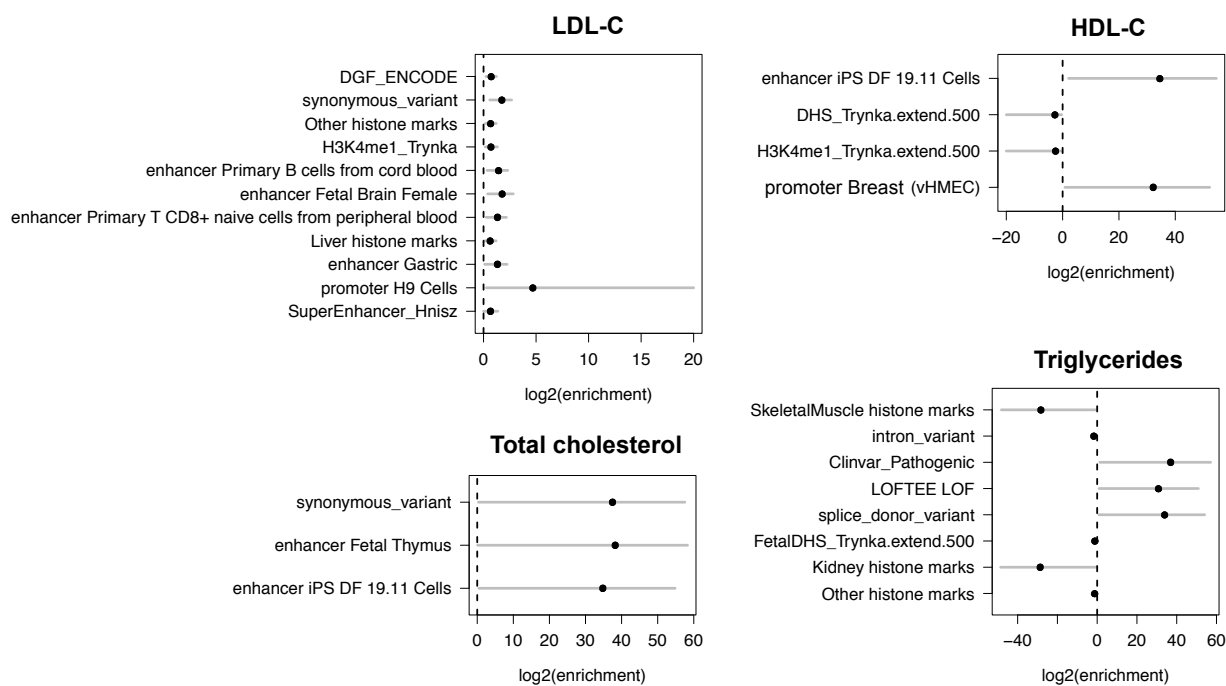
3.5.5 fgwas

Using the fgwas software¹²² and the genomic annotations described above, we applied a single annotation model to our GWAS results to investigate enrichment of each of the functional categories. Similar to the procedure performed by Pickrell (2014). First, we divided the genome into ~14,000 blocks of approximately 120 Kb each (~50 variants/block) and applied forward selection and built step-wise models including the combined effects from multiple annotations. Second, we followed by a cross validation step to avoid over fitting while maximizing the likelihood of each model. We present the final best fitting models in Figure 3.2.

3.6 Appendix B: Supplementary Tables and Figures

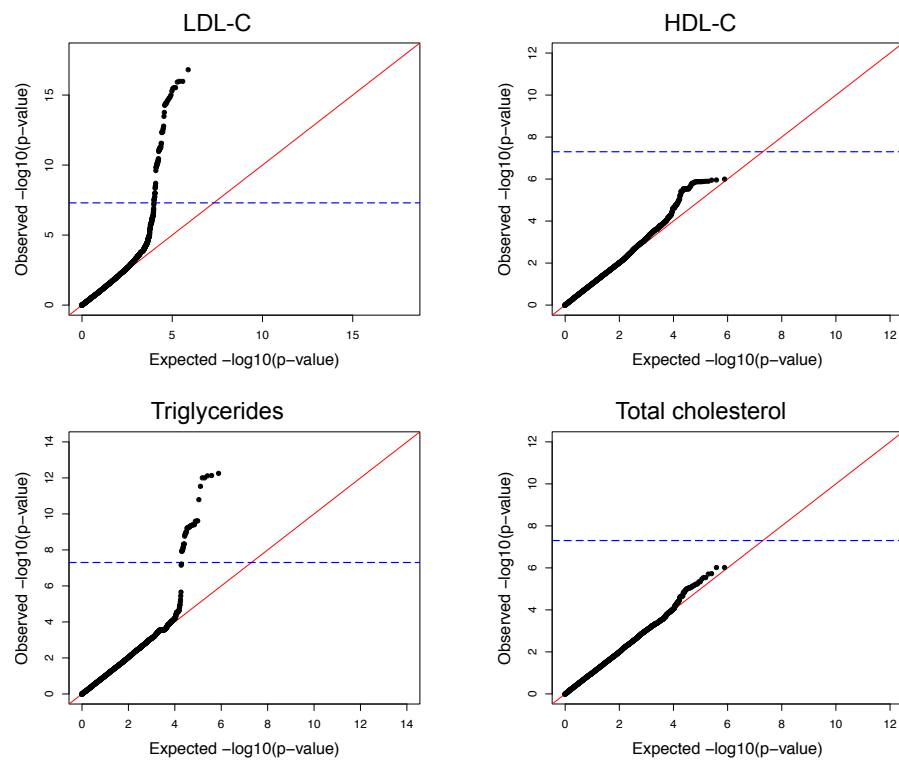
Supplementary Figure 3.1: Integration of functional annotations to GWAS of blood lipid traits for REVs

Models were fit for each individual annotation and a for regions in the top and bottom thirds of gene density using the fgwas framework. Maximum likelihood estimates (MLE) and the 95% CI of the enrichment parameter are plotted for each annotation with a CI that did not cross zero. Annotations are sorted (top to bottom) by how much they improved the likelihood of the model (AIC).



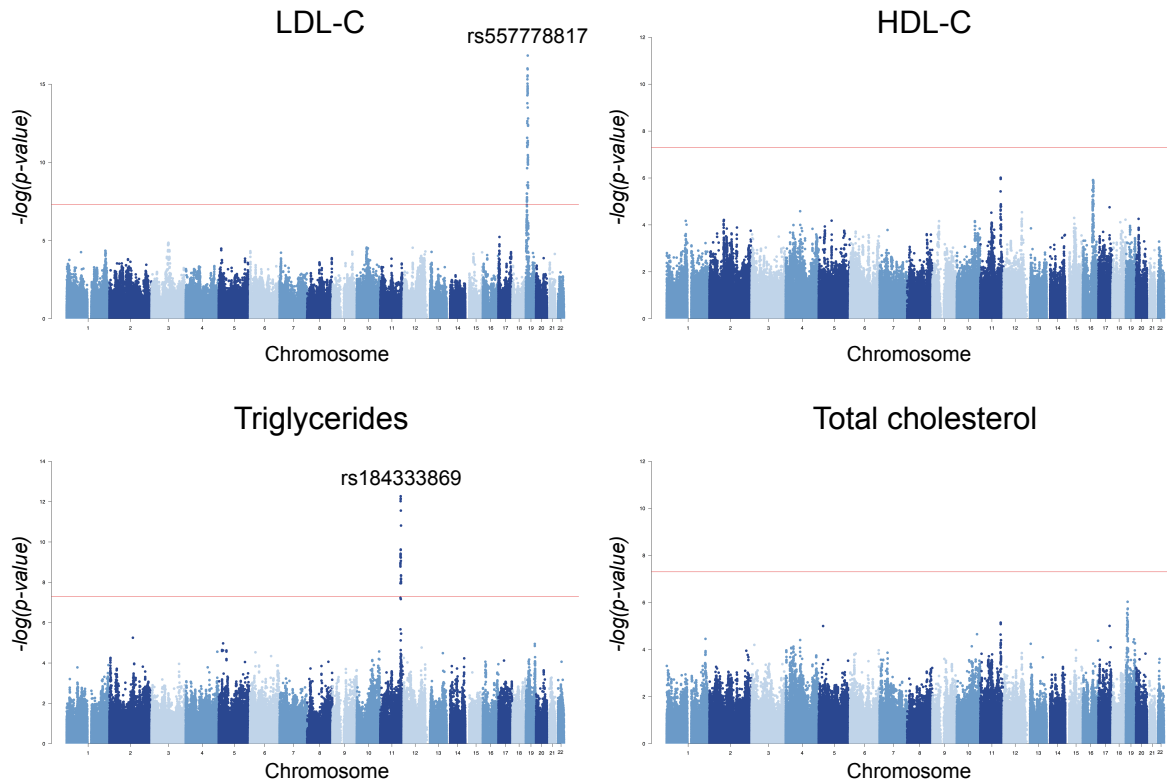
Supplementary Figure 3.2: Quantile-quantile plots for REV GWAS

Dashed blue line represents the genome-wide significant threshold ($p < 5 \times 10^{-8}$).



Supplementary Figure 3.3: Manhattan plots for REV GWAS

Red line shows the genome-wide significant threshold ($p < 5 \times 10^{-8}$). Labelled variants are the most significant variant at each genome-wide significant locus.



Supplementary Table 3.1: Rare variants in Europeans FGWAS combined model for LDL-C

Shown are the parameters inferred from a single annotation model for LDL-C along and the marginal effect of each annotation in the final model.

Parameter	Description	$\log_2(\text{effect})$ [95% CI]	Marginal(effect) [95% CI]
Gene-rich region	Regional annotation: top 1/3 of gene density	0.18 [-0.10, 0.44]	NA
Gene-poor region	Regional annotation: bottom 1/3 of gene density	-2.24 [-4.08, -1.29]	NA
synonymous_variant	Synonymous SNPs	1.77 [0.64, 2.69]	1.76 [0.59, 2.67]
Other cell type-specific histone marks	Union of cell type-specific histone marks (H3K4me1, H3K4me3, H3K9ac and H3K27ac) from Epigenome Roadmap merged by Finucane <i>et al.</i> (2015). ^{a,c}	0.86 [0.36, 1.36]	0.69 [0.18, 1.21]

Supplementary Table 3.1 (continued)

Parameter		Description				log ₂ (effect) [95% CI]	Marginal(effect) [95% CI]
enhancer	Osteoblast	Epigenome	Roadmap	putative	enhancer	1.32 [0.25,2.21]	0.94 [-0.32,1.85]
Primary Cells		Combination of ENCODE and Roadmap Epigenomics DHS data, post-processed by Trynka <i>et al.</i> (2013). ^c				-0.97 [-1.88,-0.25]	-0.05 [-0.90,0.61]
Repressed_Hoffman		Repressed regions defined as the intersection over the six cell lines from the combined chromHMM and Segway annotations in Hoffman <i>et al.</i> (2013). ^c				0.76 [0.13,1.36]	0.30 [-0.31,0.91]
Liver cell type-specific histone marks		Union of liver cell type-specific histone marks annotations (H3K4me1, H3K4me3, H3K9ac and H3K27ac) from Epigenome Roadmap merged by Finucane <i>et al.</i> (2015). ^c				0.51 [-0.04,1.04]	0.64 [0.04,1.19]
enhancer	Skeletal Muscle	Epigenome	Roadmap	putative	enhancer	-2.66 [-5.62,-1.08]	-1.01 [-3.98,0.55]
Male		Epigenome Roadmap putative enhancer region. ^b				1.40 [0.21,2.39]	1.341 [0.10,2.25]
enhancer Gastric		Epigenome Roadmap putative enhancer region. ^b				0.70 [-0.19,1.46]	0.72 [-0.32,1.53]
Enhancer_Andersson.exte		Epigenome Roadmap putative enhancer region. ^b					
nd.500							

^aOther cell type-specific histone marks corresponds to cell types not in adrenal and pancreas, CNS, cardiovascular, connective and bone, gastrointestinal, immune and hematopoietic, kidney, liver or skeletal muscle categories (includes adipose nuclei).
^bPutative enhancer region delineated using DNase I regions (-log(pvalue)>10), annotated with the 5 mark-15 state model on observed data across 127 epigenomes (ROADMAP + ENCODE). Data downloaded from Reg2Map Honeybadger2. ^cBed files downloaded from data published in Finucane *et al.* (2015).

Supplementary Table 3.2: Rare variants in Europeans FGWAS combined model for TG levels

Presented are the parameters inferred from a single annotation model for TG levels and the marginal effect of each annotation in the final model.

Parameter		Description				log ₂ (effect) [95% CI]	Marginal(effect) [95% CI]
Skeletal Muscle cell type-specific histone marks		Union of skeletal muscle cell type-specific histone marks annotations (H3K4me1, H3K4me3, H3K9ac and H3K27ac) from Epigenome Roadmap merged by Finucane <i>et al.</i> (2015). ^c				-156.23 [-176.23,-141.46]	-28.19 [-48.19,-0.86]
enhancer	Primary	Epigenome	Roadmap	putative	enhancer	-84.54 [-104.54,3.18]	-31.66 [-51.66,1.33]
mononuclear cells from peripheral blood		Epigenome Roadmap putative enhancer region. ^b					
enhancer	Duodenum	Epigenome	Roadmap	putative	enhancer	174.80 [129.39,189.56]	1.16 [-1.66,2.84]
Smooth Muscle		Epigenome Roadmap putative enhancer region. ^b				-102.26 [-122.26,-87.50]	-27.53 [-47.53, 0.32]
enhancer Fetal Stomach							

Supplementary Table 3.2 (continued)

Parameter	Description	$\log_2(\text{effect})$ [95% CI]	Marginal(effect) [95% CI]
FetalDHS_Trynka	Combination of ENCODE and Roadmap Epigenomics DHS data, post-processed by Trynka <i>et al.</i> (2013). ^c	-56.69 [-76.69,-23.56]	-1.50 [-4.53,0.08]
enhancer Mammary Primary Cells	HMEC Epithelial Epigenome Roadmap putative enhancer region ^b	26.99 [2.08,46.99]	0.50 [-2.17,1.97]
Transcr_Hoffman	Transcribed regions defined as the union over the six cell lines from the combined chromHMM and Segway annotations in Hoffman <i>et al.</i> (2013). ^c	-1.67 [-3.99,-0.08]	-1.11 [-2.68,0.14]
LOFTEE LOF	Stop gain of function, splicing and frameshift variants flagged as high confidence loss of function variants using the VEP plugin Loss-Of-Function Transcript Effect Estimator (LoFTEE).	159.14 [58.08,179.14]	30.90 [0.95,50.9]
Enhancer_Hoffman	Enhancer regions defined as the union over the six cell lines from the combined chromHMM and Segway annotations in Hoffman <i>et al.</i> (2013). ^c	-134.24 [-154.24,-97.37]	-27.42 [-47.43,0.09]
enhancer Adipose Derived Mesenchymal Stem Cell Cultured Cells	Predicted enhancer regions defined as the union over the six cell lines from the combined chromHMM and Segway annotations in Hoffman <i>et al.</i> (2013). ^c	75.45 [23.94,90.19]	0.39 [-2.91,2.01]

^bPutative enhancer region delineated using DNase I regions ($-\log(\text{pvalue}) > 10$), annotated with the 5 mark-15 state model on observed data across 127 epigenomes (ROADMAP + ENCODE). Data downloaded from Reg2Map Honeybadger2. ^cBed files downloaded from data published in Finucane *et al.* (2015).

Supplementary Table 3.3: Rare variants in Europeans FGWAS combined model for HDL-C

Presented are the parameters inferred from a single annotation model HDL-C and the marginal effect of each annotation in the final model.

Parameter	Description	$\log_2(\text{effect})$ [95% CI]	Marginal(effect) [95% CI]
enhancer Fetal Intestine Small	Epigenome Roadmap putative enhancer region. ^b	-60.49 [-80.49,0.48]	-30.11 [-50.11,2.11]
Repressed_Hoffman	Repressed regions defined as the intersection over the six cell lines from the combined chromHMM and Segway annotations in Hoffman <i>et al.</i> (2013). ^c	33.35 [2.12,37.88]	2.15 [-0.05,20]
H3K4me3_peaks_Trynka	H3K4me3n peaks post-processed by Trynka <i>et al.</i> (2013). ^c	4.06 [0.39,20]	1.63 [-20,4.24]
enhancer iPS DF 19.11 Cells	Epigenome Roadmap putative enhancer region. ^b	34.24 [29.72,54.24]	34.57 [2.11,54.57]

^bPutative enhancer region delineated using DNase I regions (-log(pvalue)>10), annotated with the 5 mark-15 state model on observed data across 127 epigenomes (ROADMAP + ENCODE). Data downloaded from Reg2Map Honeybadger2. ^cBed files downloaded from data published in Finucane *et al.* (2015).

Supplementary Table 3.4: Rare variants in Europeans FGWAS combined model for total cholesterol

Presented are the parameters inferred from a single annotation model for total cholesterol and the marginal effect of each annotation in the final model.

Parameter	Description	$\log_2(\text{effect})$ [95% CI]	Marginal(effect) [95% CI]
intron_variant	Intronic SNPs and indels	-167.54 [-187.54,20]	30.62 [-20,50.62]
Coding_UCSC.extend.500	USCD coding variants plus a 500bp window. ^c	1.36 [-20,20]	27.26 [-20,47.26]
synonymous_variant	Synonymous SNPs	63.92 [0.73,83.92]	37.531 [0.32,57.53]
Other cell type-specific histone marks	Union of cell type-specific histone marks (H3K4me1, H3K4me3, H3K9ac and H3K27ac) from Epigenome Roadmap merged by Finucane <i>et al.</i> (2015). ^{a,c}	89.516 [-0.80,109.51]	27.872 [-20,47.87]

^bPutative enhancer region delineated using DNase I regions (-log(pvalue)>10), annotated with the 5 mark-15 state model on observed data across 127 epigenomes (ROADMAP + ENCODE). Data downloaded from Reg2Map Honeybadger2. ^cBed files downloaded from data published in Finucane *et al.* (2015).

Supplementary Table 3.5: Summary of association results for rare in European variants

The table shows results from the association studies and from the fgwas for variants with posterior probability of association (PPA)>0.5 within regions with PPA>0.75. Variants listed are sorted by their PPA with the trait and then by the number of annotations in the model. Variants presented are the most significant in the region. rsID was annotated with dbSNP142. Location is based on hg19. Direction of effect corresponds to the minor allele.

rsID	Chr	Pos	Minor/ Major	MAF	1000g EUR MAF	Beta	SE	p	logBF	PPA	Func refGene	Gene refGene	CADD	GERP
LDL-C														
.	1	224811120	A/C	0.021	.	-0.82	0.21	7.99E-05	5.26	0.799	intronic	CNIH3	4.53	-2.98
rs2172423	1	225753576	A/G	0.023	0.005	-0.84	0.20	4.61E-05	5.74	0.528	intronic	ENAH	0.95	-3.26
rs191020975	3	96710644	G/A	0.019	0.001	0.86	0.20	1.88E-05	6.46	0.787	intronic	EPHA6	1.67	-1.84
rs186364673	3	97530464	T/G	0.018	0.007	0.87	0.20	1.45E-05	6.67	0.614	intergenic	ARL6, CRYBG3	0.75	-0.45
rs17242388	19	11206861	A/G	0.03	0.006	1.23	0.15	3.89E-15	28.02	1.000	intronic	LDLR	2.95	0.24
.	19	13096730	-/T	0.032	.	1.12	0.15	4.60E-13	23.25	1.000	intergenic	DAND5, NFIX	.	.
.	19	9992952	C/T	0.031	.	1.26	0.15	9.26E-16	29.47	1.000	intronic	OLFM2	1.65	0.36
rs574678076	19	9120550	A/G	0.029	0.001	1.23	0.16	1.66E-14	26.47	1.000	intergenic	MUC16, OR1M1	3.18	0.14
rs146413263	19	10352921	T/C	0.031	0.007	1.24	0.15	2.94E-15	28.31	1.000	intergenic	S1PR2, MRPL4	0.12	-0.32
rs149995389	19	9447570	T/C	0.03	0.001	1.27	0.15	4.91E-16	30.11	1.000	intronic	ZNF559, ZNF559-ZNF177	1.03	-1.23
rs142680117	19	10665828	T/C	0.031	0.002	1.29	0.15	1.13E-16	31.64	1.000	exonic	KRI1	16.19	-8.42
rs557493846	19	7970831	A/G	0.023	0.001	1.31	0.19	2.70E-12	21.10	1.000	intronic	MAP2K7	10.78	4.65
rs186246405	19	8018607	A/G	0.023	0.001	1.31	0.19	2.70E-12	21.10	1.000	intergenic	TIMM44, ELAVL1	1.63	0.02
.	19	11824001	T/C	0.04	.	0.97	0.14	4.24E-12	21.22	1.000	intergenic	ZNF833P, ZNF823	2.68	0.70
rs188348411	19	10889442	T/C	0.03	0.001	1.23	0.15	3.64E-15	28.10	1.000	intronic	DNM2	1.87	-2.12
.	19	11441687	T/G	0.039	.	0.97	0.14	7.62E-12	20.61	1.000	intronic	RAB3D	0.60	0.00
.	19	10276938	A/G	0.031	.	1.24	0.15	2.08E-15	28.66	1.000	intronic	DNMT1	3.65	-1.77
rs569073306	19	9798489	T/A	0.03	0.002	1.27	0.15	3.24E-16	30.52	1.000	intergenic	ZNF562, ZNF812	5.39	0.12
rs186853496	19	8950870	C/A	0.03	0.005	1.17	0.16	2.30E-13	23.88	1.000	intergenic	ZNF558, MBD3L1	0.01	-1.11
rs184956593	19	10721201	A/G	0.03	.	1.25	0.15	1.32E-15	29.11	1.000	intronic	SLC44A2	1.38	0.15
.	19	10079506	G/A	0.03	.	1.24	0.15	4.88E-15	27.75	1.000	intronic	COL5A3	1.09	0.47

Supplementary Table 3.5 (continued)

rsID	Chr	Pos	Minor/ Major	MAF	1000g EUR MAF	Beta	SE	p	logBF	PPA	Func refGene	Gene refGene	CADD	GERP
rs117064391	19	9032847	C/G	0.031	0.008	1.24	0.16	5.18E-15	27.71	1.000	intronic	<i>MUC16</i>	0.52	0.80
rs190268412	19	12387157	A/T	0.039	.	0.93	0.14	7.56E-11	18.36	1.000	intronic	<i>ZNF44</i>	12.33	0.24
.	19	8326153	-/C	0.028	.	1.15	0.16	5.60E-12	20.67	1.000	intronic	<i>CERS4</i>	8.32	.
rs139682067	19	8393494	A/T	0.027	0.003	1.15	0.17	1.03E-11	20.05	1.000	intronic	<i>KANK3</i>	4.99	-2.23
rs148951852	19	12574827	T/C	0.038	0.001	0.92	0.14	1.40E-10	17.76	1.000	exonic	<i>ZNF709</i>	0.00	-5.15
.	19	12899332	T/C	0.038	.	0.92	0.14	1.31E-10	17.83	1.000	intergenic	<i>HOOK2, JUNB</i>	9.83	-3.78
.	19	12083606	-/C	0.039	.	0.93	0.14	9.11E-11	18.18	1.000	intronic	<i>ZNF763</i>	2.15	.
.	19	5908883	A/C	0.03	.	0.97	0.17	9.84E-09	13.49	1.000	exonic	<i>VMAC</i>	9.65	1.78
rs140431175	19	13436237	C/T	0.087	0.005	0.57	0.10	4.33E-09	14.72	1.000	intronic	<i>CACNA1A</i>	1.29	0.19
rs568448676	19	13324259	A/G	0.033	.	1.14	0.15	1.57E-13	24.33	1.000	intronic	<i>CACNA1A</i>	1.93	0.65
rs566944285	19	7494350	A/G	0.023	.	1.06	0.18	2.89E-09	14.47	1.000	intronic	<i>ARHGEF18</i>	0.10	-0.47
rs183623185	19	12234039	A/G	0.038	0.0089	0.86	0.14	3.71E-09	14.58	1.000	intergenic	<i>ZNF788, ZNF20</i>	5.83	0.11
rs548788472	19	6940181	C/T	0.028	.	0.97	0.17	1.00E-08	13.45	1.000	UTR3	<i>ADGRE1</i>	0.95	-4.84
.	19	5851205	G/-	0.03	.	0.97	0.17	9.84E-09	13.49	0.999	ncRNA intronic	<i>LOC101928844</i>	.	.
.	19	7162104	A/G	0.026	.	0.96	0.17	1.70E-08	12.92	0.999	intronic	<i>INSR</i>	2.58	0.12
rs557778817	19	11305534	T/C	0.032	0.003	1.30	0.15	1.48E-17	33.78	0.998	intronic	<i>KANK2</i>	11.31	1.31
.	19	6385501	T/C	0.029	.	0.92	0.16	3.26E-08	12.37	0.997	intronic	<i>GTF2F1</i>	12.39	0.44
rs182067948	19	9196132	A/G	0.051	0.005	0.65	0.12	2.31E-07	10.78	0.992	intergenic	<i>MUC16, OR1M1</i>	1.80	0.30
rs8105775	19	4508945	A/G	0.028	0.003	0.81	0.16	9.59E-07	9.23	0.991	exonic	<i>PLIN4</i>	0.02	-10.20
rs146605171	19	6316687	T/C	0.028	0.008	0.89	0.17	1.46E-07	10.95	0.986	intronic	<i>ACER1</i>	0.94	0.00
.	19	4989452	A/G	0.027	.	0.84	0.17	5.95E-07	9.65	0.984	intronic	<i>KDM4B</i>	0.04	-0.47
rs149816565	19	5990960	C/T	0.028	0.003	0.88	0.17	2.28E-07	10.53	0.983	ncRNA intronic	<i>LOC100128568</i>	0.07	-0.30
rs143950469	19	5355165	G/T	0.027	0.005	0.84	0.17	5.14E-07	9.80	0.981	intergenic	<i>PTPRS, ZNRF4</i>	0.17	-0.94
rs571040577	19	4546637	-/T	0.028	.	0.81	0.16	9.59E-07	9.23	0.979	intronic	<i>SEMA6B</i>	3.31	0.23
.	19	8535995	T/C	0.03	.	1.16	0.16	3.11E-13	23.60	0.976	intronic	<i>HNRNPM</i>	3.30	1.47
.	19	5204448	C/A	0.027	.	0.85	0.17	4.73E-07	9.87	0.973	intergenic	<i>KDM4B, PTPRS</i>	0.35	-2.36

Supplementary Table 3.5 (continued)

rsID	Chr	Pos	Minor/ Major	MAF	1000g EUR MAF	Beta	SE	p	logBF	PPA	Func refGene	Gene refGene	CADD	GERP
rs189151917	19	14252905	G/A	0.072	0.004	0.52	0.10	7.61E-07	9.75	0.972	ncRNA intronic	<i>LOC100507373</i>	8.66	-0.64
rs186291988	19	6231974	A/G	0.028	0.0089	0.92	0.17	5.12E-08	11.93	0.957	intronic	<i>MLLT1</i>	3.13	0.15
rs143494927	19	6870759	T/A	0.028	0.001	0.97	0.17	2.93E-08	12.42	0.947	intergenic	<i>VAV1, ADGRE1</i>	0.22	-1.36
.	19	4789986	A/G	0.028	.	0.83	0.16	5.41E-07	9.75	0.943	intergenic	<i>MIR7-3HG, FEM1A</i>	0.21	-0.25
.	19	4910782	A/C	0.026	.	0.78	0.17	4.54E-06	7.80	0.933	UTR5	<i>UHRF1</i>	14.07	0.80
rs243408	19	4391219	T/C	0.071	.	0.42	0.10	2.39E-05	6.58	0.923	intronic	<i>SH3GL1</i>	0.28	-2.55
.	19	11978399	T/C	0.04	.	0.97	0.14	5.55E-12	20.95	0.920	exonic	<i>ZNF439</i>	16.65	0.58
rs138164292	19	10574510	T/C	0.031	.	1.28	0.15	2.79E-16	30.71	0.919	exonic	<i>PDE4A</i>	12.62	0.39
.	19	14905426	G/A	0.07	.	0.48	0.11	6.82E-06	7.71	0.916	intergenic	<i>ADGRE2, OR7C1</i>	10.87	0.26
rs144565099	19	4013262	T/C	0.025	0.0099	0.80	0.18	1.29E-05	6.85	0.906	exonic	<i>PIAS4</i>	13.17	-1.11
rs144299052	19	3704459	A/G	0.028	0.007	0.82	0.18	3.35E-06	8.07	0.892	intergenic	<i>PIP5K1C, TJP3</i>	4.25	0.13
rs183519608	19	11085893	G/A	0.03	0.003	1.25	0.15	1.86E-15	28.78	0.882	intronic	<i>SMARCA4</i>	0.67	-0.66
rs372169964	19	5756683	T/C	0.03	0.003	0.79	0.17	2.35E-06	8.44	0.866	intronic	<i>CATSPERD</i>	0.20	-0.41
.	19	10475942	G/C	0.031	.	1.25	0.15	1.50E-15	28.99	0.863	intronic	<i>TYK2</i>	2.46	0.18
rs61749998	19	5286119	T/A	0.027	0.005	0.85	0.17	4.38E-07	9.94	0.848	exonic	<i>PTPRS</i>	0.00	-4.27
rs144428474	19	4363665	A/G	0.029	0.008	0.78	0.16	2.19E-06	8.51	0.833	intronic	<i>SH3GL1</i>	3.05	-5.78
.	19	6508923	T/C	0.028	.	0.93	0.17	2.54E-08	12.59	0.823	intergenic	<i>TUBB4A, TNFSF9</i>	.	.
rs560044076	19	4123418	G/A	0.026	0.001	0.78	0.18	9.78E-06	7.12	0.819	intronic	<i>MAP2K2</i>	6.29	-0.49
.	19	15157649	-/T	0.07	.	0.49	0.11	3.99E-06	8.20	0.811	intergenic	<i>CCDC105, CASP14</i>	.	0.23
.	19	7701282	T/C	0.025	.	1.12	0.17	2.36E-10	16.90	0.788	upstream	<i>STXBP2</i>	6.48	-0.09
rs530101930	19	4656384	T/C	0.028	0.002	0.82	0.16	7.13E-07	9.50	0.782	downstream	<i>TNFAIP8L1</i>	2.06	-0.08
.	19	5103634	-/G	0.028	.	0.85	0.17	4.12E-07	10.01	0.780	intronic	<i>KDM4B</i>	.	-6.47
.	19	13875648	T/G	0.073	.	0.50	0.10	1.71E-06	8.99	0.777	intronic	<i>MRI1</i>	6.43	-0.80
rs138266407	19	10196115	A/G	0.086	0.005	0.41	0.10	2.35E-05	6.60	0.776	upstream	<i>C19orf66</i>	0.57	-2.93
rs142229220	19	5799928	T/C	0.029	0.003	0.80	0.17	1.55E-06	8.82	0.771	intergenic	<i>DUS3L, NRTN</i>	6.25	0.29
rs538957278	19	4086284	-/C	0.026	0.008	0.78	0.18	9.78E-06	7.12	0.767	intergenic	<i>ZBTB7A, MAP2K2</i>	4.69	.

Supplementary Table 3.5 (continued)

rsID	Chr	Pos	Minor/ Major	MAF	1000g EUR MAF	Beta	SE	p	logBF	PPA	Func refGene	Gene refGene	CADD	GERP
rs187206531	19	14959756	A/C	0.07	0.0099	0.48	0.11	7.67E-06	7.60	0.749	intergenic	OR7A10, OR7A17	1.62	-0.71
rs530966553	19	5587016	C/T	0.028	0.002	0.73	0.17	1.58E-05	6.74	0.737	UTR3	SAFB2	4.48	1.84
rs561129352	19	12758503	C/T	0.038	.	0.92	0.14	1.31E-10	17.83	0.697	intronic	MAN2B1	1.05	-4.99
rs574831801	19	4713209	A/G	0.028	0.002	0.82	0.16	6.49E-07	9.59	0.689	intronic	DPP9	0.58	-2.19
rs190144917	19	9895661	C/T	0.031	0.001	1.28	0.15	9.91E-17	31.79	0.651	intergenic	ZNF846, FBXL12	8.97	0.05
rs11569543	19	6684297	G/A	0.029	0.002	0.92	0.17	6.42E-08	11.72	0.632	intronic	C3	0.56	-0.23
rs181406089	19	14434363	C/G	0.074	0.007	0.51	0.10	1.06E-06	9.44	0.616	intergenic	ADGRL1, ADGRE5	0.15	0.00
.	19	12357787	T/C	0.039	.	0.94	0.14	6.58E-11	18.50	0.583	intergenic	LOC100289333, ZNF44	7.05	-0.56
rs386806954	19	13171585	G/T	0.032	.	1.12	0.15	4.09E-13	23.36	0.570	intronic	NFIX	.	.
rs79745793	19	8261478	G/A	0.028	0.0089	1.15	0.16	5.60E-12	20.67	0.535	intergenic	FBN3, CERS4	5.48	-0.96
rs576729806	19	3671995	A/G	0.028	.	0.82	0.18	3.35E-06	8.07	0.504	intronic	PIP5K1C	0.77	-2.39
TG														
rs149157643	11	117921732	G/A	0.023	0.007	-1.27	0.17	7.47E-13	22.49	1.000	ncRNA intronic	TMPRSS4-AS1	6.68	2.17
.	11	118830967	T/C	0.022	.	-1.27	0.18	2.79E-12	21.10	1.000	intergenic	UPK2, FOXR1	4.62	0.13
rs146127017	11	119875844	A/G	0.025	0.008	-1.12	0.16	1.54E-11	19.60	1.000	intergenic	LOC102724301, TRIM29	2.47	-0.35
rs554314590	11	118115935	A/G	0.021	0.001	-1.18	0.18	2.40E-10	16.81	1.000	intronic	MPZL3	1.84	0.11
rs150611309	11	117313707	T/C	0.021	0.004	-1.16	0.18	3.81E-10	16.39	1.000	intronic	DSCAML1	4.10	1.91
rs117119462	11	117559653	T/C	0.02	0.003	-1.18	0.19	3.80E-10	16.34	1.000	intronic	DSCAML1	2.20	-0.92
rs377657187	11	119033729	G/A	0.02	.	-1.18	0.19	8.63E-10	15.52	1.000	downstrea m	ABCG4	0.19	-0.30
rs192198415	11	116298324	A/G	0.023	0.004	-1.16	0.19	1.19E-09	15.26	1.000	intergenic	LOC107987166	5.01	1.24
.	11	116290343	A/T	0.023	.	-1.16	0.19	1.40E-09	15.10	0.999	intergenic	LINC00900, BUD13	.	.
rs112994980	11	116491788	A/G	0.023	0.003	-1.15	0.19	1.66E-09	14.93	0.999	intergenic	LINC00900, BUD13	7.29	-1.52
rs112909684	11	119611228	A/G	0.024	0.0089	-1.00	0.17	6.50E-09	13.78	0.999	downstrea m	LOC102724301	2.56	-0.56
.	11	119747886	A/G	0.023	.	-1.03	0.17	4.51E-09	14.09	0.998	intergenic	LOC102724301, TRIM29	0.15	-6.03
rs193288949	11	119452788	A/G	0.024	0.001	-0.99	0.17	9.09E-09	13.47	0.996	intergenic	USP2-AS1, PVRL1	0.85	-0.47

Supplementary Table 3.5 (continued)

rsID	Chr	Pos	Minor/ Major	MAF	1000g EUR MAF	Beta	SE	p	logBF	PPA	Func refGene	Gene refGene	CADD	GERP
.	11	116422626	T/C	0.022	.	-1.10	0.19	1.11E-08	13.10	0.994	intergenic	<i>LINC00900, BUD13</i>	4.94	-1.30
.	11	116646774	C/-	0.024	.	-1.17	0.19	5.63E-10	15.98	0.991	intergenic	<i>BUD13, ZPR1</i>	.	
.	11	119411377	T/A	0.024	.	-0.99	0.17	1.11E-08	13.28	0.987	intergenic	<i>USP2-AS1, PVRL1</i>	2.30	-0.85
rs184333869	11	117947268	T/C	0.023	0.001	-1.28	0.17	5.41E-13	22.79	0.935	ncRNA intronic	<i>TMPRSS4-AS1</i>	0.59	-5.54
rs193234129	11	120034709	A/C	0.024	0.008	-1.03	0.17	4.70E-09	14.08	0.865	intergenic	<i>TRIM29, OAF</i>	0.14	-0.30
rs138326449	11	116701354	A/G	0.024	0.003	-1.14	0.18	1.08E-09	15.38	0.816	splicing	<i>APOC3</i>	25.10	4.89
rs73585219	11	117616252	T/C	0.02	.	-1.17	0.19	5.53E-10	15.98	0.762	intronic	<i>DSCAML1</i>	2.94	0.08
rs148548901	11	117669384	C/T	0.02	0.001	-1.17	0.19	4.22E-10	16.24	0.755	intergenic	<i>DSCAML1, FXYD6-FXYD2</i>	1.42	-0.04
.	11	117067279	T/G	0.022	.	-1.15	0.18	4.03E-10	16.36	0.525	ncRNA intronic	<i>LOC100652768</i>	18.40	1.88
rs6589676	11	118578089	T/C	0.02	0.006	-1.19	0.19	5.90E-10	15.89	0.509	intergenic	<i>TREH, DDX6</i>	0.80	-0.67
.	11	118079744	T/G	0.021	.	-1.19	0.18	2.36E-10	16.82	0.508	intronic	<i>AMICA1</i>	1.03	1.09
HDL-C														
rs184333869	11	117947268	T/C	0.023	0.001	0.89	0.18	1.10E-06	9.03	0.926	ncRNA intronic	<i>TMPRSS4-AS1</i>	0.59	-5.54
rs141938931	16	56901445	G/A	0.052	0.002	-0.54	0.13	2.26E-05	6.58	0.745	intronic	<i>SLC12A3</i>	1.29	1.64
rs189679427	16	58911464	T/C	0.052	0.0099	-0.61	0.13	1.27E-06	9.24	0.722	intergenic	<i>GOT2, APOOP5</i>	7.27	0.00
rs189664208	16	57654462	T/C	0.051	0.007	-0.57	0.13	8.49E-06	7.46	0.629	intronic	<i>ADGRG1</i>	4.45	-8.00

Supplementary Table 3.6: GWAS catalog variants for region on chromosomes 3 and 19 associated with LDL-C

Variants were retrieved from the NGHRI catalog on 11/30/2016.

chr	pos(Mb)	trait	SNP
3	96.32883	Obesity-related traits	rs6796026
3	96.4006	Blood trace element (Zn levels)	rs6793516
3	96.68226	Obesity-related traits	rs7650621
3	97.16874	Blood trace element (Cu levels)	rs1458303
19	7.166109	Polycystic ovary syndrome	rs2059807
19	7.184762	Height	rs891088
19	7.196565	Diabetic retinopathy	rs2115386
19	7.199803	Urate levels	rs1035942
19	7.223848	Thyroid hormone levels,	rs4804416
19	7.223848	Hypothyroidism	rs4804416
19	7.224431	Triglycerides	rs7248104
19	7.424528	Sexual dysfunction (SSRI/SNRI-related)	rs6603109
19	7.734511	Resistin levels	rs3745367
19	7.739177	Resistin levels	rs1423096
19	7.909883	HIV1 control	rs558718
19	8.271721	Sphingolipid levels	rs7258249
19	8.433196	HDL cholesterol	rs7255436
19	8.469738	HDL cholesterol	rs2967605
19	8.644031	Height	rs4072910
19	8.672	Height	rs7249094
19	8.789381	Atopic dermatitis	rs2164983
19	10.000322	Menarche (age at onset)	rs1862471
19	10.213154	Protein quantitative trait loci	rs8109578
19	10.21803	Narcolepsy	rs1551570
19	10.226052	Narcolepsy	rs2305795
19	10.310392	Thiazide-induced adverse metabolic effects in hypertensive patients	rs2116940
19	10.394792	Soluble ICAM-1	rs1799969
19	10.395683	Soluble ICAM-1	rs5498
19	10.397238	Soluble ICAM-1	rs281437
19	10.397403	Soluble levels of adhesion molecules	rs3093030
19	10.469975	Crohn's disease	rs12720356
19	10.469975	Psoriasis	rs12720356
19	10.472933	Psoriasis	rs280519
19	10.475652	Type 1 diabetes autoantibodies	rs2304256
19	10.475652	Type 1 diabetes	rs2304256
19	10.512911	Inflammatory bowel disease	rs11879191
19	10.520064	Multiple sclerosis	rs8112449
19	10.798904	Obesity-related traits	rs2569507
19	10.838486	Coronary heart disease	rs11671653
19	11.092139	Coronary heart disease	rs11669133
19	11.1636	Coronary heart disease	rs1122608
19	11.1636	Myocardial infarction (early onset)	rs1122608
19	11.18815	Lipid metabolism phenotypes	rs55791371

Supplementary Table 3.6 (continued)

chr	pos(Mb)	trait	SNP
19	11.19503	LDL cholesterol	rs11668477
19	11.20001	LDL cholesterol	rs17249141
19	11.20231	Cholesterol, total	rs6511720
19	11.20231	LDL cholesterol	rs6511720
19	11.20231	Abdominal aortic aneurysm	rs6511720
19	11.20231	Metabolite levels	rs6511720
19	11.20231	Lipoprotein-associated phospholipase A2 activity and mass	rs6511720
19	11.20231	Cardiovascular disease risk factors	rs6511720
19	11.20231	Carotid intima media thickness	rs6511720
19	11.20231	Lipid metabolism phenotypes	rs6511720
19	11.21091	Cholesterol, total	rs2228671
19	11.21091	LDL cholesterol	rs2228671
19	11.22733	Metabolite levels	rs2738446
19	11.23847	LDL cholesterol	rs2738459
19	11.33094	Bipolar disorder and schizophrenia	rs17699030
19	11.33429	Coronary heart disease	rs4804155
19	11.3427	HDL cholesterol	rs12979813
19	11.34749	HDL cholesterol	rs737337
19	11.35049	HDL cholesterol	rs2278426
19	11.65501	IgG glycosylation	rs11880637
19	11.68249	Multiple sclerosis (severity)	rs7253363
19	12.14637	Multiple sclerosis	rs3745672
19	12.22537	Cognitive performance	rs17638629
19	12.69119	Bipolar disorder	rs7247513
19	12.93227	Mean corpuscular volume	rs7255045
19	13.00155	Mean corpuscular hemoglobin	rs11085824
19	13.02425	Red blood cell traits	rs741702
19	13.24597	Obesity-related traits	rs10408126

Supplementary Table 3.7: GWAS catalog variants for region on chromosome 11 associated with triglyceride levels and with HDL-C

Variants were retrieved from the NGHRI catalog on 11/30/2016.

chr	pos (Mb)	trait	snp
11	116.2984	Brain structure	rs490592
11	116.4004	Bipolar disorder and schizophrenia	rs7941534
11	116.5205	HIV-1 susceptibility	rs1240773
11	116.585	Triglycerides	rs4938303
11	116.5863	Hematological and biochemical traits	rs7350481
11	116.5863	Triglycerides	rs7350481
11	116.5994	Multiple sclerosis (severity)	rs180358
11	116.6037	Carotenoid and tocopherol levels	rs12272004
11	116.6037	Cholesterol, total	rs12272004

Supplementary Table 3.7 (continued)

chr	pos (Mb)	trait	snp
11	116.6037	LDL cholesterol	rs12272004
11	116.6037	Triglycerides	rs12272004
11	116.6074	LDL cholesterol	rs1558861
11	116.6074	Triglycerides	rs1558861
11	116.6137	Cardiovascular disease risk factors	rs12280753
11	116.6172	HDL cholesterol	rs11216126
11	116.6191	Lipid traits	rs28927680
11	116.6191	Triglycerides	rs28927680
11	116.6263	Triglycerides-Blood Pressure (TG-BP)	rs11825181
11	116.6339	Metabolic syndrome (bivariate traits)	rs11820589
11	116.6391	HDL Cholesterol - Triglycerides (HDL-C-TG)	rs10790162
11	116.6391	Metabolic syndrome	rs10790162
11	116.6391	Waist Circumference - Triglycerides (WC-TG)	rs10790162
11	116.6476	Triglycerides	rs2160669
11	116.6476	Obesity-related traits	rs2160669
11	116.6489	Cholesterol, total	rs964184
11	116.6489	HDL cholesterol	rs964184
11	116.6489	LDL cholesterol	rs964184
11	116.6489	Triglycerides	rs964184
11	116.6489	Hypertriglyceridemia	rs964184
11	116.6489	Metabolite levels	rs964184
11	116.6489	Response to Vitamin E supplementation	rs964184
11	116.6489	Metabolic syndrome	rs964184
11	116.6489	Phospholipid levels (plasma)	rs964184
11	116.6489	Lipoprotein associated phospholipase A2 activity and mass	rs964184
11	116.6489	Vitamin E levels	rs964184
11	116.6489	Coronary heart disease	rs964184
11	116.6491	Triglycerides - Blood Pressure (TG- BP)	rs11823543
11	116.6522	Metabolic syndrome (bivariate traits)	rs12286037
11	116.6522	Lipoprotein associated phospholipase A2 activity and mass	rs12286037
11	116.6522	Triglycerides	rs12286037
11	116.6524	Triglycerides	rs6589566
11	116.6524	LDL cholesterol	rs6589566
11	116.6533	HDL Cholesterol - Triglycerides (HDL-C - TG)	rs2075290
11	116.6533	Metabolic syndrome	rs2075290
11	116.6533	Waist Circumference - Triglycerides (WC-TG)	rs2075290
11	116.6544	Triglycerides	rs603446
11	116.6576	Obesity related traits	rs3741298
11	116.6607	HDL Cholesterol - Triglycerides (HDL-C-TG)	rs2266788
11	116.6607	Metabolic syndrome	rs2266788
11	116.6607	Triglycerides-Blood Pressure (TG, BP)	rs2266788
11	116.6607	Waist Circumference - Triglycerides (WC-TG)	rs2266788
11	116.6626	Lipid metabolism phenotypes	rs651821
11	116.6626	Triglycerides	rs651821
11	116.6637	Lipid traits	rs662799
11	116.681	Coronary heart disease	rs1263173

Supplementary Table 3.7 (continued)

chr	pos (Mb)	trait	snp
11	116.684	Quantitative traits	rs7396835
11	116.706	Infantile hypertrophic pyloric stenosis	rs12721025
11	116.722	Iron status biomarkers	rs10047462
11	116.7325	Triglycerides	rs2075292
11	116.783	Infantile hypertrophic pyloric stenosis	rs11216185
11	116.8297	Visceral adipose tissue adjusted for BMI	rs7120173
11	116.9739	Coronary heart disease	rs12269901
11	117.0374	Protein quantitative trait loci	rs7112513
11	117.0756	Cardiovascular disease risk factors	rs508487
11	117.0916	Iron status biomarkers	rs236918
11	117.1569	Cardiovascular disease risk factors	rs1047964
11	117.3889	Menarche (age at onset)	rs11216435
11	117.5317	Triglycerides	rs10892151
11	117.8511	Obesity related traits	rs58603829
11	118.032	Preeclampsia	rs11600901
11	118.117	Thiazide induced adverse metabolic effects in hypertensive patients	rs11216831
11	118.1457	Pain	rs17122021

Supplementary Table 3.8: GWAS catalog variants for region on chromosome 16 associated with HDL-C

Variants were retrieved from the NGHRI catalog on 11/30/2016.

chr	pos (Mb)	trait	SNP
16	56.72517	Gambling	rs8064100
16	56.96915	Metabolic syndrome	rs2217332
16	56.98514	Lipid traits	rs9989419
16	56.98514	HDL cholesterol	rs9989419
16	56.98804	Metabolite levels	rs173539
16	56.98804	Cardiovascular disease risk factors	rs173539
16	56.98804	HDL Cholesterol - Triglycerides (HDL-C-TG)	rs173539
16	56.98804	Metabolic syndrome	rs173539
16	56.98804	Metabolic syndrome (bivariate traits)	rs173539
16	56.98804	HDL cholesterol	rs173539
16	56.98959	Lipoprotein associated phospholipase A2 activity and mass	rs247616
16	56.98959	Cardiovascular disease risk factors	rs247616
16	56.99072	HDL cholesterol	rs247617
16	56.99072	Metabolic syndrome	rs247617
16	56.99332	Cholesterol, total	rs3764261
16	56.99332	HDL cholesterol	rs3764261
16	56.99332	LDL cholesterol	rs3764261
16	56.99332	Triglycerides	rs3764261
16	56.99332	Lipoprotein-associated phospholipase A2 activity and mass	rs3764261
16	56.99332	Lipid metabolism phenotypes	rs3764261

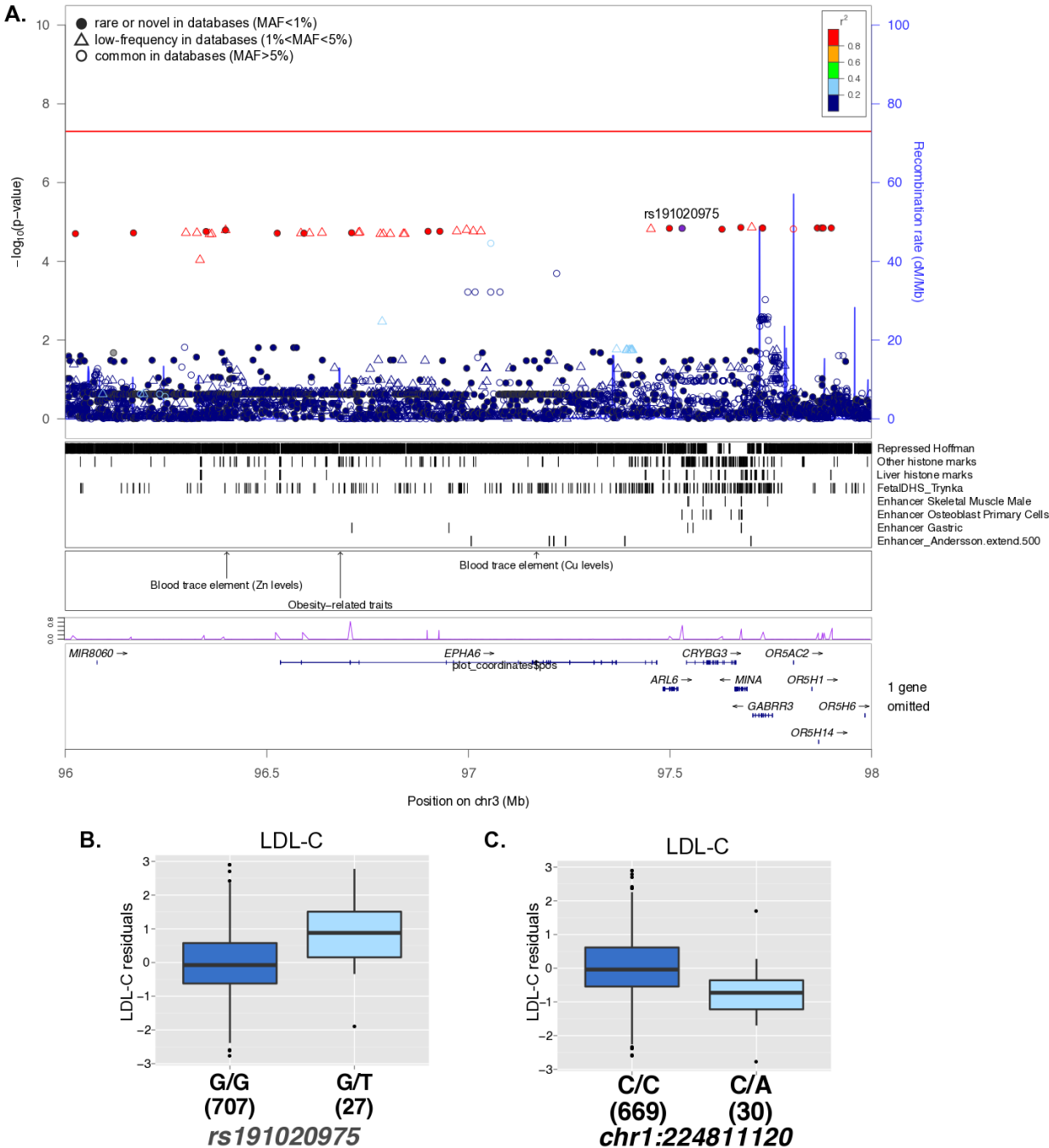
Supplementary Table 3.8 (continued)

chr	pos (Mb)	trait	SNP
16	56.99332	Age related macular degeneration	rs3764261
16	56.99332	Metabolic syndrome (bivariate traits)	rs3764261
16	56.99332	Metabolic syndrome	rs3764261
16	56.99332	Hematological and biochemical traits	rs3764261
16	56.99332	Waist circumference and related phenotypes	rs3764261
16	56.99524	Lipid traits	rs1800775
16	56.99524	Lipid metabolism phenotypes	rs1800775
16	56.99524	HDL cholesterol	rs1800775
16	56.99524	Triglycerides	rs1800775
16	56.99723	Age related macular degeneration	rs1864163
16	56.99723	Lipid metabolism phenotypes	rs1864163
16	56.99723	HDL cholesterol	rs1864163
16	57.00273	Metabolic syndrome (bivariate traits)	rs9939224
16	57.00548	HDL cholesterol	rs1532624
16	57.00548	Cholesterol	rs1532624
16	57.00548	Lipid metabolism phenotypes	rs1532624
16	57.00659	Lipid metabolism phenotypes	rs7499892
16	57.00659	Biochemical measures	rs7499892
16	57.01238	HDL cholesterol	rs12708980
16	57.02432	Schizophrenia	rs17290922
16	57.0473	Coronary heart disease	rs16965039
16	57.05858	Bipolar disorder and schizophrenia	rs821470
16	57.30319	Sleep duration	rs11640439
16	57.56367	Cataracts in type 2 diabetes	rs8052123
16	58.07528	Pulmonary function (interaction)	rs12447804
16	58.07528	Pulmonary function	rs12447804
16	58.5663	QT interval	rs37060
16	58.56724	Electrocardiographic traits	rs37062
16	58.56724	QT interval	rs37062
16	58.62218	QT interval	rs7188697
16	58.75933	Aortic root size	rs12600277
16	59.8907	Brain structure (hippocampal volume)	rs8056650
16	59.94019	Obesity-related traits	rs16961543
16	60.31466	Phospholipid levels (plasma)	rs9932186

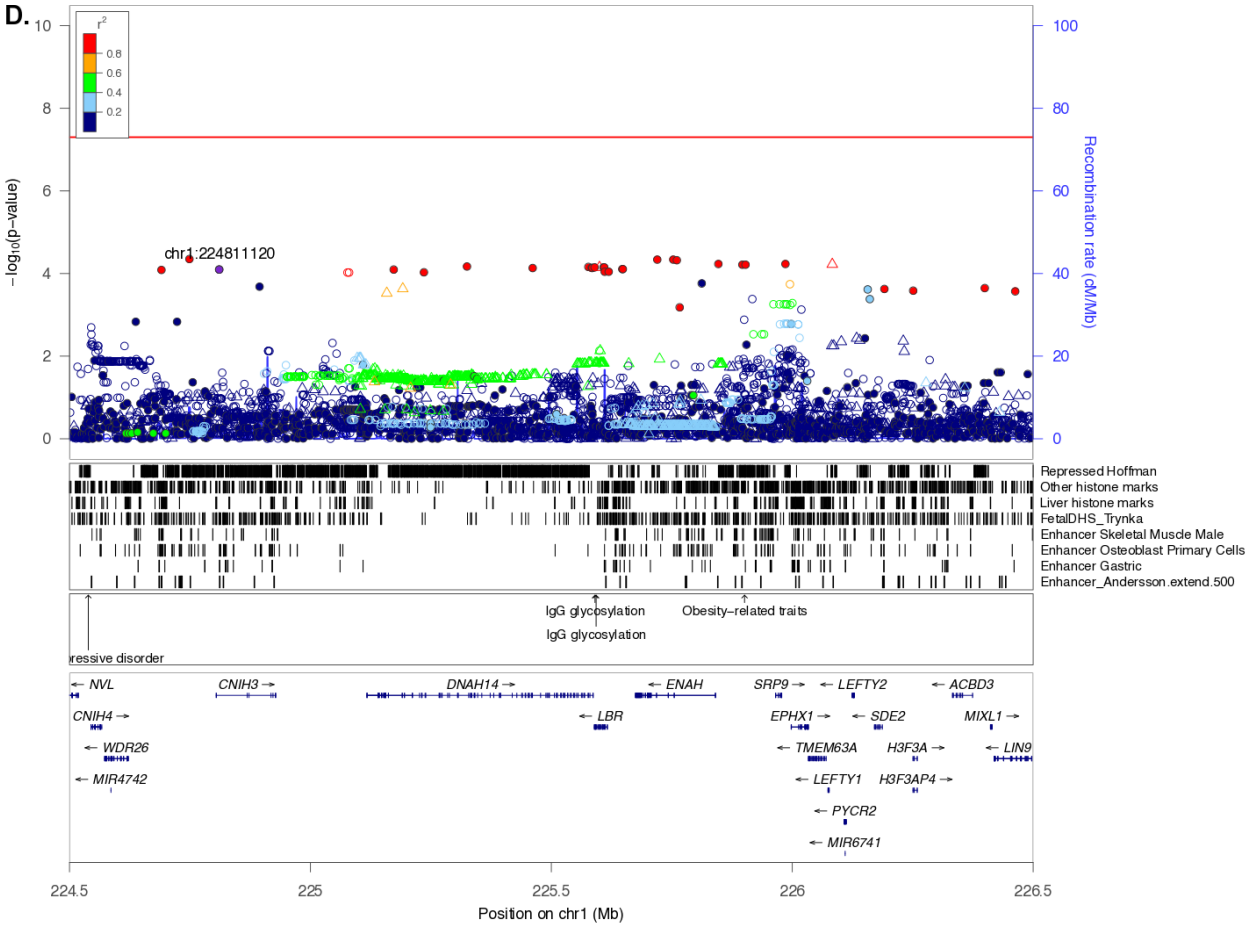
Supplementary Figure 3.4: Rare variants on chromosomes 3 and 1 are associated with increased LDL-C

A and D. Locus plots. The top panel shows the p-values of association with TG for all variants discovered in the Hutterites regardless of allele frequency in Europeans. Symbols correspond to the maximum allele frequency in Europeans, with closed circles representing REVs (see legend), and are colored based on their LD r^2 with the most associated variant in the region (rs191020975 and chr1:224811120, respectively). The next panel shows tracks for the annotations selected in the fgwas joint model. The next panel shows annotations of known GWAS loci from NHGRI. The next panel illustrates the estimated PPA of being causal for each variant in the reweighted fgwas (conditional on there being one causal variant in the region). The bottom-most panel shows the genes in the region. **B and C. Genotype boxplots for LDL-C for candidate intronic variants in EPHA6 (rs191020975; $p=1.88 \times 10^{-5}$) and CNIH3 (chr1:224811120; $p=7.99 \times 10^{-5}$).**

Supplementary Figure 3.4: Rare variants on chromosomes 3 and 1 are associated with increased LDL-C (continued)



Supplementary Figure 3.4: Rare variants on chromosomes 3 and 1 are associated with increased LDL-C (continued)



CHAPTER 4

HOST GENETIC VARIATION IN MUCOSAL IMMUNITY PATHWAYS INFLUENCES THE UPPER AIRWAY MICROBIOME

4.1 Abstract¹

The degree to which host genetic variation can modulate microbial communities in humans remains an open question. Here we performed a genetic mapping study of the microbiome in two accessible upper airway sites, the nasopharynx and the nasal vestibule, during two seasons in 144 adult members of a founder population of European descent. We estimated the relative abundances (RAs) of genus level bacteria from 16S rRNA gene sequences and examined associations with 148,653 genetic variants (linkage disequilibrium [LD] $r^2 < 0.5$) selected from among all common variants discovered in genome sequences in this population. We identified 37 microbiome quantitative trait loci (mbQTLs) that showed evidence of association with the RAs of 22 genera ($q < 0.05$), and were enriched for genes in mucosal immunity pathways. The most significant association was between the RA of *Demacoccus* (phylum Actinobacteria) and a variant 8kb upstream of *TINCR* (rs117042385; $p=1.61 \times 10^{-8}$; $q = 0.002$), a long non-coding RNA that binds to peptidoglycan recognition protein 3 (*PGLYRP3*) mRNA, a gene encoding a known antimicrobial protein. A second association was between a missense variant in *PGLYRP4* (rs3006458) and the RA of an unclassified genus of family Micrococcaceae (phylum Actinobacteria) ($p=5.10 \times 10^{-7}$; $q =$

¹ Citation for chapter: Igartua *et al.* Host genetic variation in mucosal immunity pathways influences the upper airway microbiome. *Microbiome* 5, 16 (2017)

0.032). Our findings provide evidence of host genetic influences on upper airway microbial composition in humans, and implicate mucosal immunity genes in this relationship.

4.2 Introduction

Diverse populations of microorganisms inhabit nearly every surface of the human body and these complex assemblies of microbes reflect host-microbe and microbe-microbe interactions. Collectively, these microorganisms constitute the human microbiome¹⁵³. Under healthy conditions, the relationship between microbes and the host is symbiotic with many physiologic benefits to the host¹⁵⁴. Imbalances or changes in the composition of bacterial communities can shift this relationship from symbiotic to pathogenic, a condition known as dysbiosis, which has been implicated in a variety of diseases¹⁵⁵. For example, altered composition of airway microbiota has been linked to important respiratory diseases such as sinusitis¹⁵⁶, chronic obstructive pulmonary disease (COPD)¹⁵⁷ and asthma^{53,158,159}. Similar to the traits it influences, the microbiome itself can be considered a complex phenotype with environmental and genetic factors contributing to its composition¹⁶⁰. Understanding how host genetic variation shapes the microbiome, and how the microbiome ultimately functions to modulate host immunity are fundamental questions that are central to fully characterizing the architecture of many common diseases that occur at mucosal surfaces, including those involving the airway.

Although knowledge of the airway microbiome lags behind that of the gut, important characteristics of the microbial communities in the airway are beginning to emerge. Similar to the gut, the community structure of an individual's airway microbiome is established early in life and plays a critical role in immune development¹⁶¹⁻¹⁶³. Many external factors influence the airway

microbiome, including mode of delivery at birth¹⁶⁴, breastfeeding¹⁶⁵, antibiotic use^{166,167}, and exposure to tobacco smoke¹⁶⁸ and pathogens⁵⁵. While the influences of environmental exposures on microbiome composition are well known, the degree to which host genetics plays a role in structuring microbial communities is less well understood. In fact, recent data suggest that host genetics may play an important role in shaping microbiome composition. For example, the heritability of the gut microbiome was recently investigated in 1,126 twin pairs⁵⁶. Out of 945 taxa examined, the RAs of 8.8% of taxa had non-zero heritability estimates suggesting that the abundances of those bacteria are influenced by host genetic variation. Moreover, more similar microbiome structures among related individuals compared to unrelated individuals^{169,170} further supports a role for genetics influencing inter-individual variability in microbiome profiles. In fact, quantitative trait locus (QTL) approaches have successfully identified variation in candidate host genes that influence the RA of specific bacteria not only in *Drosophila* and mice but also in humans¹⁷¹.

Studies of host genetic influences on the microbiome are particularly challenging due to the profound effects of environmental exposures on microbiome variability. It is not surprising, therefore, that two studies were unable to show host genotype effects on the human gut microbiome^{172,173}. Studies of related individuals and even twin pairs are confounded to a large extent by the more similar environments among close relatives, making it impossible to completely disentangle the relative roles of genes and environment. To address these challenges, we focused our studies on the Hutterites, a founder population that practices a communal, farming lifestyle that minimizes environmental variation between individuals¹⁷⁴, and should increase power to

identify genetic influences on complex traits, including the airway microbiome composition. For example, Hutterites prepare and eat all meals in communal kitchens, smoking is prohibited and rare, and individual family homes are nearly identical within each colony (communal farm) and very similar across colonies. Furthermore, the Hutterites in our studies are related to each other in a 13-generation pedigree and are descendants of only 64 founders. Finally, nearly all genetic variation in these individuals has been revealed through whole genome sequencing studies in 98 Hutterite individuals⁴⁹.

We previously reported studies of the gut microbiome in the Hutterites^{58,175}. Here we interrogated the interaction between host genetic variation and microbiome composition in two accessible sites in the upper airways, the nasal vestibule and the nasopharynx, which have important physiologic functions and relevance to airway diseases. While the nasal vestibule is located in the anterior nares and in direct contact with the environment, the nasopharynx is in the posterior nasal passage and continuous with the lower airway. Overall, our findings demonstrate that the airway microbiome is influenced by host genotype at many loci, and suggest that host expression of innate and mucosal immune pathway genes plays a significant role in structuring the airway microbiome.

4.3 Results

4.3.1 *Nasal microbiome composition*

To characterize the variation of the microbiome from the nasal vestibule and the nasopharynx, we first analyzed 16S rRNA V4 gene sequences from 322 samples collected from 144 Hutterite adults in summer and/or in winter months (Table 4.1). After applying quality control filters and

subsampling to 250,000 reads per sample, 83 million reads were assigned to 563 operational taxonomic units (OTUs) with 97% sequence identity. We identified sequences from eleven phyla, with three accounting for 98.94% of the sequences – Firmicutes (52.28%), Actinobacteria (29.81%) and Proteobacteria (16.85%). We then classified OTUs into 166 genera; six dominant genera accounted for 83.30% of the sequences (Figure 4.1 and Supplementary File 4.1A).

Figure 4.1: Taxonomic composition of bacterial communities in the nasal vestibule and the nasopharynx, sampled in summer and in winter

Genus level mean RA is shown for the 20 most abundant genera identified in the samples. The remaining 146 genera are grouped as “other”. Out of the six most abundant genera (83.30% of the sequences) in our sample, four (*Corynebacterium*, *Moraxella*, *Streptococcus*, *Staphylococcus*) were among the most abundant (RA > 1%) in adult participants in the HMP anterior nares sample¹⁵³ (adults) and five (*Corynebacterium*, *Moraxella*, *Alloiococcus*, *Streptococcus*, *Staphylococcus*) among infants in the Children Asthma Study nasopharynx sample¹⁷⁶. *Genus unclassified, family level presented. **Genus and family unclassified, order level presented.

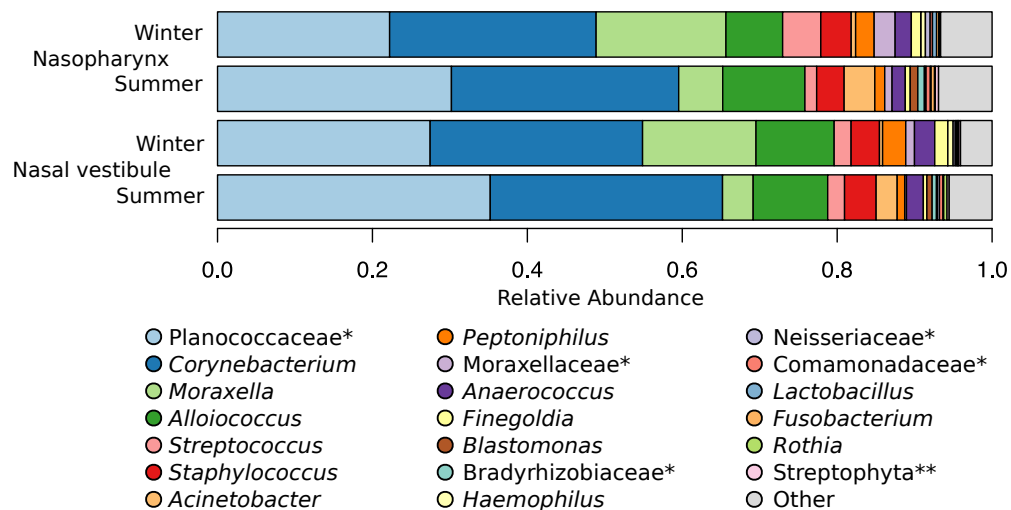


Table 4.1: Sample composition

A total of 332 samples were collected from 144 (58 male, 86 female) Hutterite adults (age 16 to 78 years).

Nasal site	Summer	Winter	Both seasons	Unique subjects
Vestibule	87	80	34	133
Nasopharynx	88	77	40	125
Both sites	72	60	23	144

In a prior study in a largely overlapping sample of adult Hutterites, we identified large seasonal variation in the gut (fecal) microbiome¹⁷⁵. To see if similar patterns were present in the nasal microbiome, we examined the genus level RAs for individuals studied in both seasons (n=34 for the nasal vestibule and 40 for the nasopharynx). The RA of 12 genera in the nasal vestibule and 15 in the nasopharynx differed by season after applying a Bonferroni correction (paired Wilcoxon rank sum test, $p < 0.0003$), nine of which were different between seasons at both nasal sites (Supplementary File 4.1B-C). Similarly, we looked for genus level RAs that differed between the nasal sites within each of the two seasons (N=72 for the summer and 60 for the winter) but did not identify statistically significant differences.

4.3.2 Nasal microbiome diversity

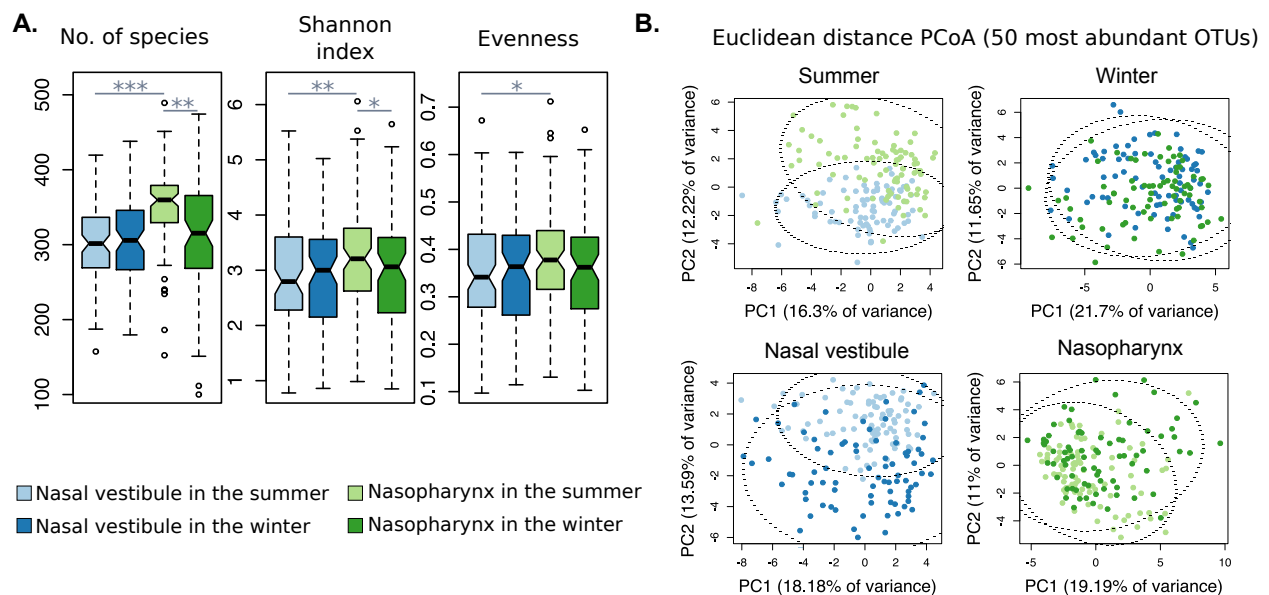
We used three diversity metrics to assess within sample (alpha) diversity for each of the four seasonal nasal site groups – the number of species (richness), Shannon index, and evenness. Overall, the highest alpha diversity was observed in the nasopharynx in the summer

Figure 4.2A and Supplementary Figure 4.1A), where the number of observed species and the Shannon index reflected higher diversity compared to the nasopharynx in the winter (paired Wilcoxon signed-rank test; $p = 0.002$ and 0.048 , respectively). Additionally, there was higher

diversity in the nasopharynx in the summer compared to the nasal vestibule in the summer (paired Wilcoxon signed-rank test; richness $p=4.6 \times 10^{-7}$, Shannon index $p=0.009$ and evenness $p=0.031$, respectively). Although higher diversity trends were observed in the nasopharynx in the summer compared to the winter, these associations were largely due to decreased alpha diversity among women compared to men in the winter (Wilcoxon signed-rank test; richness $p=0.03$, Shannon index $p=0.002$ and evenness $p=0.001$, respectively; Supplementary File 4.1B). Lastly, Shannon index and evenness decreased with increasing age only in the nasopharynx in the summer ($p=0.019$; Supplementary Figure 4.2).

Figure 4.2: Diversity

A. Alpha diversity. Alpha diversity measurements for microbial communities from the nasal vestibule and the nasopharynx by season. The nasopharynx in the summer (light green) shows the overall largest alpha diversity. *Paired Wilcoxon-rank sum test $p<0.5$; ** $p<0.01$; *** $p<0.001$. **B. Beta diversity.** Principal coordinate analysis (PCoA) of the 50 most abundant OTUs derived from Euclidean distance. Nasal site samples show separation in the summer ($PC2\ p<2.2 \times 10^{-16}$). Seasonal samples show separation in the nasal vestibule ($PC2\ p<2.2 \times 10^{-16}$) and in the nasopharynx ($PC2\ p<0.0002$). Ellipsoids contain 95% of the samples within each group.



We next analyzed community composition and structure between samples (beta diversity) by calculating Euclidean distances between all pairs of individuals. In the seasonal analyses, the summer samples for both the nasal vestibule and the nasopharynx had lower Euclidean distances compared to their respective winter samples (Wilcoxon signed-rank test, nasal vestibule $p < 2.2 \times 10^{-16}$ and nasopharynx $p < 2.2 \times 10^{-16}$), reflecting more similar microbiome diversity between pairs of individuals in the summer than in the winter. Moreover, Euclidean distances for the same individual paired with him/herself between seasons (separately within the nasal vestibule and nasopharynx samples) and between nasal sites (separately within the summer and the winter samples) were lower than the respective distances calculated between each individual with all other individuals (Wilcoxon signed-rank test nasal vestibule between seasons $p = 9.25 \times 10^{-8}$; nasopharynx between seasons $p = 3.97 \times 10^{-12}$; summer between nasal sites $p < 2.2 \times 10^{-16}$, winter between nasal sites $p < 2.2 \times 10^{-16}$; Supplementary Figure 4.3). These results reflect stability in microbiome structure between seasons and nasal sites within individuals, potentially reflecting a genetic component to microbiome composition and diversity.

To further examine the impact of anatomical location and season on beta diversity, we analyzed Euclidean distances derived from all OTUs with principal coordinate analysis (PCoA). Although PCoA did not show clear separation between the top two PCs and the four groups, analysis of similarity (ANOSIM) revealed significant differences in community structure ($p < 9.9 \times 10^{-5}$; $R = 0.16$). Because beta diversity was previously shown to have greater dissimilarities between anterior and other, more posterior nasal microbiomes¹⁷⁷, we performed PCoA again after selecting the 50 most abundant OTUs to determine if variation captured from composition of the

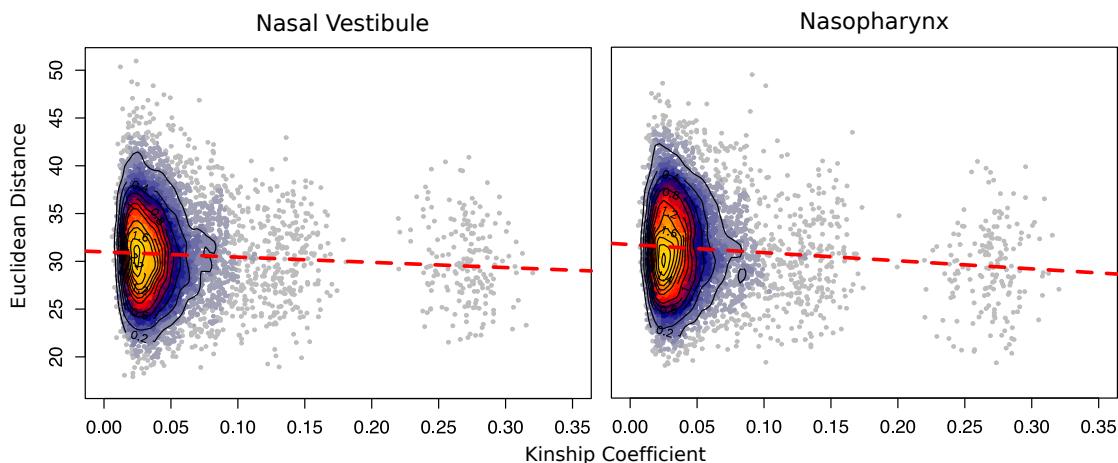
most predominant bacteria would separate between nasal sites. In this analysis, PCoA showed clear dissimilarities between the NV and the NP in the summer (PC2 $p < 2.2 \times 10^{-16}$; ANOSIM $p < 9.9 \times 10^{-5}$; $R = 0.24$; Figure 4.2B). We note that the lack of dissimilarity between nasal sites in the winter may result from decreased power due to the lower alpha diversity observed in the winter samples. In fact, the higher prevalence of respiratory viruses in temperate regions in this season¹⁷⁸ supports the possibility that viral-microbial interactions may synergistically affect microbial structure at both nasal sites in the winter. For example, studies of the airway microbiome after exposure to pH1N1¹⁷⁹ and rhinovirus¹⁸⁰ have demonstrated that viral infections can alter microbiome composition in a manner that decreases microbial diversity and enriches for pathogens associated with subsequent bacterial infections.

4.3.3 Correlation between host genetic similarity and microbiome structure

To evaluate the relationship between genetic similarity (or relatedness) among pairs of individuals and the similarity of their nasal microbiomes, we compared genetic distance, measured by the kinship coefficient and beta diversity between all pairs of individuals in the sample combined across seasons (see methods). We reasoned that if there was a genetic influence on bacterial composition and diversity, more related individuals should have lower measures of beta diversity, reflecting more similar microbiomes. To assess significance, we performed 10,000 permutations for each of the two nasal sites. This analysis revealed a significant negative Spearman correlation between kinship and beta diversity (NV $p < 1 \times 10^{-4}$, NP $p = 4.0 \times 10^{-4}$; Figure 4.3).

Figure 4.3: Heat scatterplots of Euclidean distance (beta diversity) by kinship coefficient

Individuals with larger kinship coefficients (more related) have more similar beta diversities (lower Euclidean distances). Red dashed represents the trend line from a linear model. Nasal vestibule $p < 1 \times 10^{-4}$; nasopharynx $p = 4.0 \times 10^{-4}$.



Although an individual's microbiome composition is highly sensitive to the household environment¹⁸¹, sharing of households by first degree relatives did not significantly affect the correlation between beta diversity and kinship in our sample. To examine this directly, we removed all first degree relatives who lived in the same household (three sibling pairs and their parents; 15 out of 175 first-degree relative pairs in the sample) and repeated the analysis. The correlation between kinship and Euclidean distance remained significant (NV $p < 1 \times 10^{-4}$ and NP $p = 5.0 \times 10^{-4}$), indicating that the significant effect of kinship on microbiome similarities between Hutterite adults is not likely due to shared environments. Instead, we attribute these correlations largely to shared genetic variation.

4.3.4 Genome-wide association studies of relative abundance

To directly test for host genetic effects on genus level bacteria in the nasal vestibule and in the nasopharynx, we performed microbiome quantitative trait locus mapping on the bacterial RAs for

each nasal site in the summer sample and winter samples separately, and in a larger sample combining both seasons. We tested for associations between 52 and 90 genera with 148,653 SNPs ($LD\ r^2 < 0.5$) using a linear mixed model as implemented in GEMMA⁷³, and included sex and age as fixed effects and kinship as a random effect to adjust for the relatedness between all pairs of individuals in our study. Due to the multiple testing burden inherent in microbiome studies, genome-wide significance thresholds were determined within each bacteria GWAS. We note, however, that none of the associations passed a study-wide significance threshold ($p < 1.1 \times 10^{-10}$) that would correct for the number of bacteria, sites, and seasons tested. Our analyses revealed 37 mbQTLs at $q < 0.05$, eight of which passed genome wide significance ($p < 5 \times 10^{-8}$), and three of which were associated with multiple bacteria (overall 37 variants associated with 22 genus level bacteria). Of the 37 mbQTLs, 14 were associated with 10 genera in the nasal vestibule and 23 were associated with 14 genera in the nasopharynx. The results for mbQTLs with $q < 0.05$ are shown in Table 4.2 and results for 108 mbQTLs with $q < 0.10$ are shown in Supplementary File 4.2.

Table 4.2: QTL mapping results of nasal microbiome relative abundance

A. *Nasal vestibule*. 14 host variants were associated at a $q < 0.05$ with the relative abundance of 10 genera. rs111354832 is associated with an unclassified genus of family Micrococcaceae in the summer and in the combined sample.

Bacteria genus (phylum/class/order/family)	Mean RA	rsID	Chr	Start	Alleles	Gene(s)	p	q	Beta
Summer									
<i>Dermacoccus</i> (Actinobacteria/ Actinobacteria/ Actinomycetales/ Dermacoccaceae)	7.9X10 ⁻⁵	rs67386870	5	126156219	A/C	<i>LMNB1</i>	2.46X10 ⁻⁷	0.016	-1.10
		rs77536542	5	168583325	G/A	<i>SLIT3 (1,2)</i>	6.35X10 ⁻⁸	0.005	-1.27
		rs117042385*	19	5530692	T/C	<i>ZNRF4, TINCR</i>	1.61X10 ⁻⁸	0.002	-1.16
		rs28362459	19	5844792	A/C	<i>FUT3</i>	9.47X10 ⁻⁷	0.047	-0.81
Unclassified genus (Actinobacteria/ Actinobacteria/ Actinomycetales/ Micrococcaceae)	4.1X10 ⁻⁴	rs111354832	4	7136504	-/CAT	<i>FLJ36777, SORCS2</i>	5.99X10 ⁻⁸	0.015	-0.74
Winter									
<i>Kocuria</i> (Actinobacteria/ Actinobacteria/ Actinomycetales/ Micrococcaceae)	5.2X10 ⁻⁴	rs12713689*	2	70427457	G/A	<i>C2orf42, TIA1</i>	2.10X10 ⁻⁸	0.005	-0.91
<i>Aerococcus</i> (Firmicutes/ Bacilli/ Lactobacillales/ Aerococcaceae)	2.2X10 ⁻⁴	rs10505338	8	119755490	A/G	<i>SAMD12-AS1, TNFRSF11B</i>	2.02X10 ⁻⁷	0.038	-0.81
<i>Lactobacillus</i> (Firmicutes/ Bacilli/ Lactobacillales/ Lactobacillaceae)	1.4X10 ⁻³	rs4142162	13	81127842	G/A	<i>SPRY2, LINCO0377</i>	1.45X10 ⁻⁷	0.017	-0.77
Combined									
Unclassified genus (Actinobacteria/ Actinobacteria/ Actinomycetales/ Intrasporangiaceae)	9.2X10 ⁻⁵	rs11085969	19	15792546	A/G	<i>CYP4F12</i>	5.84X10 ⁻⁸	0.005	-0.91
<i>Microbispora</i> (Actinobacteria/ Actinobacteria/ Actinomycetales/ Micrococcaceae)	5.5X10 ⁻⁴	rs2891405	12	113152097	G/A	<i>MIR1302-1, RPH3A</i>	1.50X10 ⁻⁷	0.038	-0.81
Unclassified genus (Actinobacteria/ Actinobacteria/ Actinomycetales/ Micrococcaceae)	3.5X10 ⁻⁴	rs111354832*	4	7136504	-/CAT	<i>FLJ36777, SORCS2 (2)</i>	3.45X10 ⁻⁸	0.017	-0.77
<i>Peptoniphilus</i> (Firmicutes/ Clostridia/ Clostridiales/ Tissierellaceae)	8.9X10 ⁻⁵	rs9865782	3	113652774	A/G	<i>GRAMD1C</i>	1.17X10 ⁻⁷	0.005	-0.91

Table 4.2 (continued)

Bacteria genus (phylum/class/order/family)	Mean RA	rsID	Chr	Start	Alleles	Gene(s)	p	q	Beta
<i>Paracoccus</i> (Proteobacteria/ Alphaproteobacteria/ Rhodobacterales/ Rhodobacteraceae)	6.5X10 ⁻⁵	rs9953410	18	29532946	C/A	<i>TRAPPC8</i> , <i>RNF125</i>	2.21X10 ⁻⁷	0.038	-0.81
<i>Enterobacter</i> (Proteobacteria/ Gammaproteobacteria/ Enterobacteriales/ Enterobacteriaceae)	6.4X10 ⁻⁵	rs11042877	11	10576232	A/C	<i>MRVI1-AS1</i>	3.86X10 ⁻⁷	0.017	-0.77
		rs12446497	16	7341674	A/G	<i>RBFOX1</i>	5.84X10 ⁻⁷	0.005	-0.91

B. *Nasopharynx*. 23 host variants were associated at a $q < 0.05$ with the relative abundance of 14 genera. At this site 2 SNPs (rs1653301 and rs7702475) are associated with more than one bacterium.

Bacteria species (phylum/class/order/family)	Mean RA	rsID	Chr	Start	Alleles	Gene(s)	p	q	Beta
Summer									
<i>Aerococcus</i> (Firmicutes/ Bacilli/ Lactobacillales/ Aerococcaceae)	6.5X10 ⁻⁴	rs7702475*	5	58088523	A/G	<i>RAB3C</i>	3.97X10 ⁻⁸	0.010	0.77
Unclassified genus (Firmicutes/ Bacilli/ Lactobacillales/ Aerococcaceae)	5.3X10 ⁻⁴	rs11888528	2	120118764	C/T	<i>C2orf76</i>	1.43X10 ⁻⁷	0.038	0.79
		rs308961	3	12150014	T/G	<i>SYN2</i>	8.12X10 ⁻⁷	0.034	-0.70
<i>Methylobacterium</i> (Proteobacteria/ Alphaproteobacteria/ Rhizobiales/ Methylobacteriaceae)	4.3X10 ⁻⁴	rs10547084	4	37753111	-/TCTC	<i>RELL1</i> , <i>PGM2</i>	4.93X10 ⁻⁷	0.031	0.77
		rs67737950	4	40260058	G/C	<i>RHOH</i> , <i>LOC101060</i> <i>498</i>	2.73X10 ⁻⁷	0.023	0.85
		rs7702475	5	58088523	A/G	<i>RAB3C</i>	2.46X10 ⁻⁷	0.023	0.77
		rs1278260	10	127731197	C/A	<i>ADAM12</i>	6.21X10 ⁻⁷	0.031	0.95
Unclassified genus (Proteobacteria/ Alphaproteobacteria/ Caulobacterales/ Caulobacteraceae)	8.5X10 ⁻⁴	rs927984	6	25412987	T/C	<i>LRRC16A</i>	2.78X10 ⁻⁷	0.036	0.94
		rs1543603*	6	25413922	A/G	<i>LRRC16A</i>	2.30X10 ⁻⁸	0.006	0.88
Winter									

Table 4.2 (continued)

Bacteria genus (phylum/class/order/family)	Mean RA	rsID	Chr	Start	Alleles	Gene(s)	p	q	Beta
<i>Mycobacterium</i> (Actinobacteria/ Actinobacteria/ Actinomycetales/ Mycobacteriaceae)	2.2X10 ⁻⁴	rs1802665	10	61788623	G/T	<i>ANK3</i>	8.73X10 ⁻⁸	0.007	- 0.98
<i>Gemella</i> (Firmicutes/ Bacilli/ Gemellales/ Gemellaceae)	3.7X10 ⁻⁴	rs17631306*	1	111072322	A/G	<i>KCNA10</i> , <i>KCNA2</i>	2.94X10 ⁻⁸	0.008	1.27
Unclassified genus (Firmicutes/ Bacilli/ Lactobacillales)	5.7X10 ⁻⁴	rs1153741	2	182860422	G/A	<i>PPP1R1C</i>	1.52X10 ⁻⁷	0.039	- 0.83
Combined									
<i>Gordonia</i> (Actinobacteria/ Actinobacteria/ Actinomycetales/ Gordoniaceae)	1.1X10 ⁻⁴	rs61925863	12	66694722	C/G	<i>IRAK3</i> , <i>HELB</i>	2.88X10 ⁻⁷	0.038	- 0.90
		rs12435212*	14	85483485	G/T	<i>LINC00911</i>	1.93X10 ⁻⁸	0.005	0.88
Unclassified genus (Firmicutes/ Clostridia/ Clostridiales/ Clostridiaceae)	1.0X10 ⁻²	rs9661504	1	205915667	A/T	<i>SLC26A9</i> , <i>FAM72C</i>	4.16X10 ⁻⁷	0.046	- 0.70
		rs10232599	7	46035291	G/A	<i>IGFBP3</i> , <i>TNS3</i>	1.14X10 ⁻⁷	0.029	- 0.67
Unclassified genus (Actinobacteria/ Actinobacteria/ Actinomycetales)	1.3X10 ⁻³	rs12156316	8	41706484	T/C	<i>ANK1</i>	8.16X10 ⁻⁷	0.042	0.58
		rs10901086	9	134635034	T/C	<i>RAPGEF1</i> , <i>MED27</i>	7.05X10 ⁻⁷	0.042	0.89
		rs12244238	10	6083239	G/A	<i>IL2RA</i>	6.82X10 ⁻⁷	0.042	- 0.69
Unclassified genus (Actinobacteria/ Actinobacteria/ Actinomycetales/ Micrococcaceae)	5.0X10 ⁻⁴	rs3006458	1	153320372	T/G	<i>PGLYRP4</i>	5.10X10 ⁻⁷	0.032	- 0.75
		rs4774283	15	58114121	T/G	<i>GCOM1</i> , <i>ALDH1A2</i>	2.56X10 ⁻⁷	0.032	- 0.57
		rs4814474	20	16322199	A/C	<i>KIF16B</i>	9.35X10 ⁻⁷	0.040	0.56
<i>Rhodococcus</i> (Actinobacteria/ Actinobacteria/ Actinomycetales/ Nocardaceae)	6.4X10 ⁻⁵	rs1653301*	2	201076401	A/G	<i>C2orf47</i> , <i>SPATS2L</i>	1.45X10 ⁻⁸	0.004	- 0.82
Unclassified genus (Actinobacteria/ Actinobacteria/ Actinomycetales/ Sporichthyaceae)	1.5X10 ⁻⁴	rs13128830	4	21455808	T/C	<i>KCNIP4</i>	3.03X10 ⁻⁷	0.022	0.64

Table 4.2 (continued)

Bacteria genus (phylum/class/order/family)	Mean RA	rsID	Chr	Start	Alleles	Gene(s)	p	q	Beta
Unclassified genus (Proteobacteria/ Alphaproteobacteria/ Sphingomonadales/ Sphingomonadaceae)	5.1X10 ⁻³	rs1653301	2	201076401	A/G	<i>C2orf47</i> , <i>SPATS2L</i> (3)	9.48X10 ⁻⁸	0.024	- 0.80
rsIDs presented for dbSNP142. *Genome-wide significant result ($p < 5 \times 10^{-8}$). Alleles presented as minor/major. Direction of effect is presented for the minor allele. When genus level is unclassified, highest classified taxonomic level is bolded. RA, relative abundance; Chr, chromosome. Gene labels 1-4 correspond to genes previously reported in either the Bonder [55] or Goodrich [19] GWAS. 1, Goodrich (fecal; unclassified genus of family Clostridiaceae); 2, Bonder (fecal; PWY- 6948 sitosterol degradation to androstenedione); 3, Goodrich (fecal; genus <i>Blautia</i>).									

The most significant association with relative abundance was with an intergenic SNP 8kb upstream of the *TINCR* gene on chromosome 19 and the abundance of *Dermaococcus* (phylum Actinobacteria) in the nasal vestibule in the summer (rs117042385; $p=1.61 \times 10^{-8}$; $q = 0.002$, Figure 4.4A-B). *Dermaococcus* abundance has been observed to be depleted in the skin of individuals with atopic dermatitis¹⁸², and microbes belonging to the Actinobacteria phylum are characterized by their production of metabolites with anti-inflammatory and anti-microbial properties¹⁸³. Interestingly, *TINCR* is a long non-coding RNA gene that controls human epidermal differentiation and directly binds to the peptidoglycan recognition protein 3 (*PGLYRP3*) transcript¹⁸⁴. These combined results reflect the important role of *Dermaococcus* in surface community structure and barrier integrity in both the skin and nasal epithelia¹⁸⁵. A second mbQTL (rs28362459), located 314 kb downstream from the *TINCR* mbQTL ($r^2 = 0.26$; $D'=0.76$), was also associated with the RA of *Dermaococcus* in the same site and season as *TINCR* ($p=9.47 \times 10^{-7}$, $q = 0.047$, Figure 4.4A-B). rs28362459 is a missense variant in fucosyltransferase 3 (*FUT3*), a gene essential for the synthesis of Lewis blood groups^{186,187}.

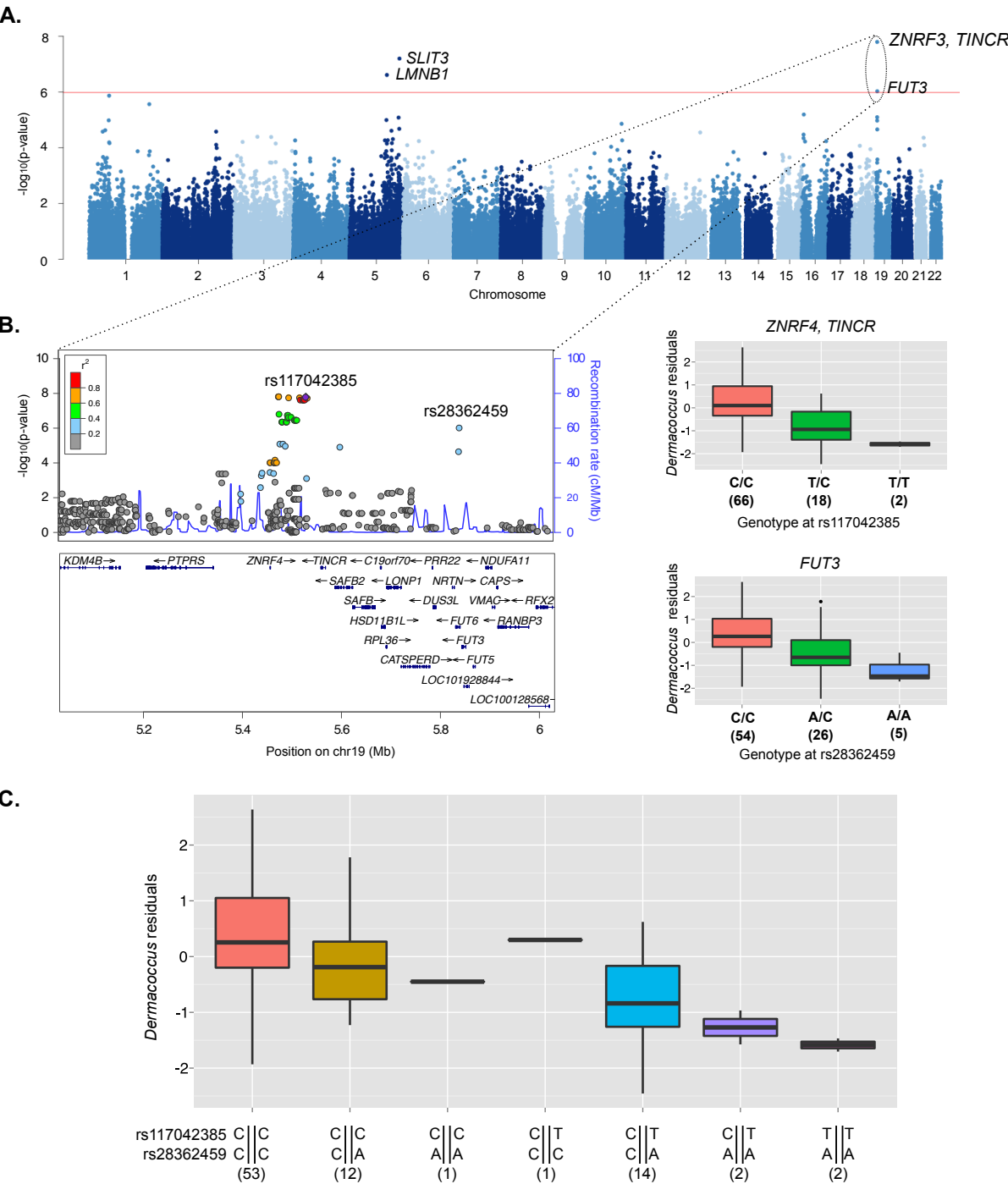
To determine if the association of increased *Dermaococcus* RA with the rs28362459-C allele in *FUT3* is independent of the association with the rs117042385-C allele upstream of *TINCR*, we phased the two variants and examined the four haplotypes (seven diplotypes) present in our sample. This revealed independent effects of genotypes at both SNPs contributing to the RA of *Dermaococcus* ($p=4.65 \times 10^{-9}$; Figure 4.4C). In particular, individuals who were homozygous for both alleles (rs117042385-CC/rs28362459-CC) had the highest RA of *Dermaococcus*, while one or two copies of the *FUT3* rs28362459-A allele on a homozygous *TINCR* rs117042385-CC background was

associated with decreased RA. Overall, the presence of a *FUT3* rs28362459-A allele was associated with lower RA regardless of genotype at *TINCR* rs117042385. The two individuals who were homozygous for both the *TINCR* rs117042385-T and *FUT3* rs28362459-A alleles did not have any *Demacoccus* sequences detected. Overall, these results suggest that the *FUT3* rs28362459 and *TINCR* rs117042385 variants (or variants in strong LD with them) are exerting independent effects on the RA of *Demacoccus*.

Figure 4.4: Associations with the RA of *Demacoccus* in the nasal vestibule in the summer

A. Manhattan plot. Association results are presented for variants pruned for LD ($r^2 < 0.5$). Four variants on chromosomes 5 and 19 are associated with the RA of *Demacoccus* at a $q < 0.05$ significance threshold (red line). **B. Locus and genotype plots for the 2 mbQTLs on chromosome 19.** Variants included in the locus plot are those with $MAF > 10\%$ in the Hutterites, prior to LD pruning. Genotype plots show both minor alleles (T at rs117042385 and A at rs28362459) are associated with lower *Demacoccus* RA. **C.** Boxplots of *Demacoccus* residuals for rs117042385 and rs28362459 phased haplotypes. Numbers underneath each boxplot represent the number of individuals in each genotype or haplotype class.

Figure 4.4: Associations with the RA of *Derma*coccus in the nasal vestibule in the summer
(continued)



Another mbQTL that linked PGLYRP genes more directly to host regulation of the microbiome is an association between a missense variant in *PGLYRP4* (rs3006458) on chromosome 1 and the RA of an unclassified genus of family Micrococcaceae (phylum Actinobacteria) in the combined season nasopharynx sample ($p=5.10 \times 10^{-7}$; $q = 0.032$, Supplementary Figure 4.4A). This same SNP was also associated with genus *Aerococcus* in the nasopharynx in the winter at a less stringent q value cutoff (phylum Firmicutes; $p=1.28 \times 10^{-6}$; $q = 0.06$). Peptidoglycan recognition proteins (PGRPs) are a conserved family of antibacterial pattern recognition molecules that directly bind peptidoglycan and other bacterial cell wall components, including lipopolysaccharide (LPS)¹⁸⁸. The genus *Aerococcus* is commonly found in a wide range of environments including air, dust, soil, meat-cutting brines, marine sources, and the human respiratory tract¹⁸⁹. Other lactic acid bacteria belonging to the order Lactobacillales (e.g. *Streptococcus pneumoniae*) colonize mucosal surfaces in the upper and lower airways of the host and cause important respiratory diseases including sinusitis, acute otitis media and pneumonia¹⁹⁰.

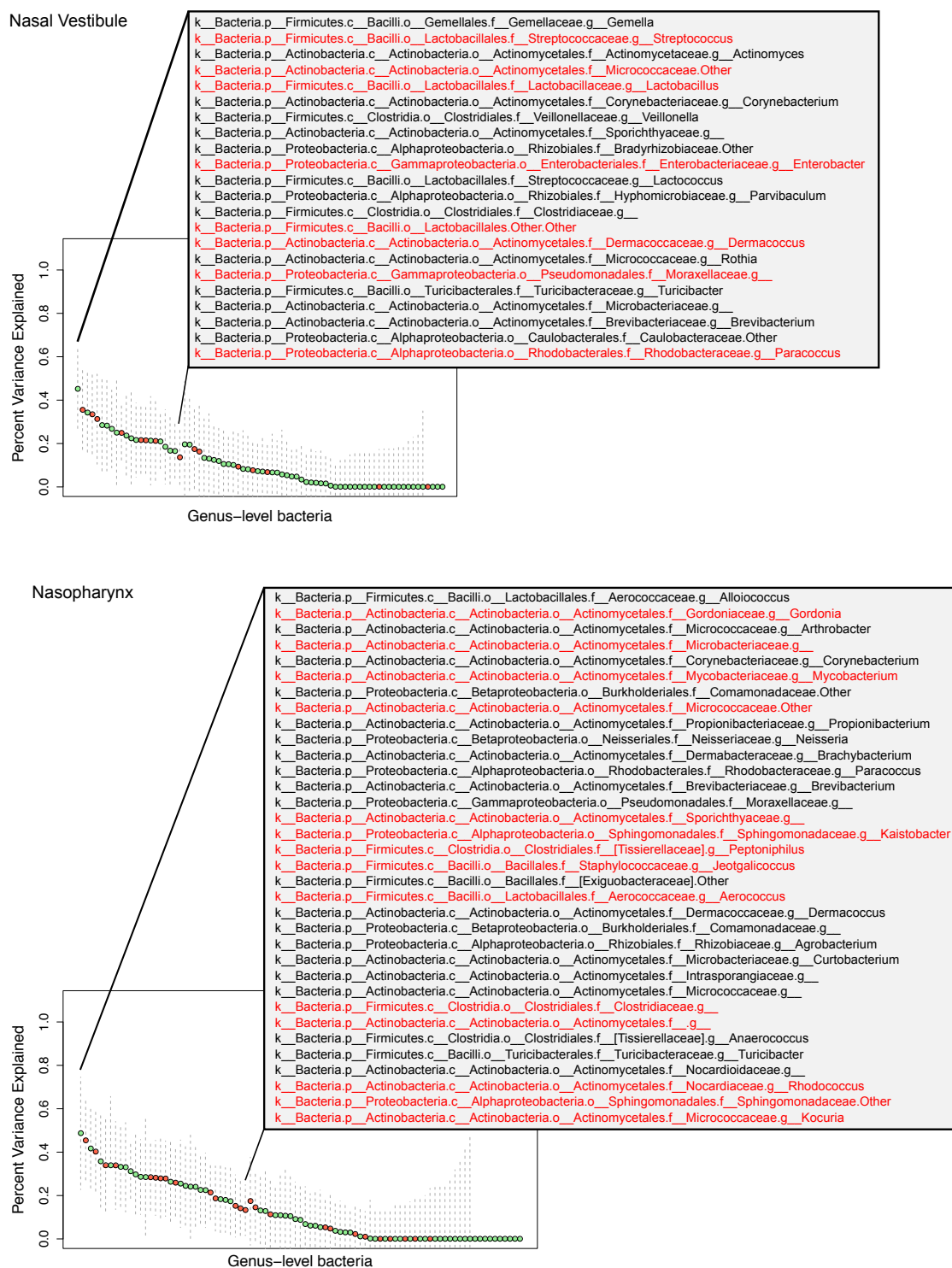
For each of the 90 bacteria tested, we calculated the proportion of relative abundance variance explained (PVE) by the joint effect of all SNPs tested in the mapping analyses (SNP heritability). PVE estimates the maximum phenotypic variance that can be explained by the variants interrogated in the analysis and provides a lower bound heritability estimate. In the combined analyses, there were 22 bacteria in the nasal vestibule (27%) and 34 bacteria in the nasopharynx (37%) that had SNP heritability estimates with standard errors not intersecting with zero. These estimates ranged between 13.25% and 48.69% (Figure 4.5 and Supplementary File

4.3). Overall, bacteria with identified mbQTLs in this study ($q < 0.10$) had higher estimates of SNP heritability than the remaining tested bacteria (Wilcoxon signed-rank test NV $p=0.009$; NP $p=0.05$).

Figure 4.5: SNP heritability estimates for 90 bacteria tested in the nasal vestibule and nasopharynx combined samples

Each point represents the percent variance explained (PVE) for each bacteria genus level relative abundance; dashed lines show the standard error. The bacteria listed in the inset have SNP heritability estimates with standard errors that do not span zero. Bacteria for which mbQTLs ($q < 0.10$) were identified in either the seasonal or the combined samples are shown in red. Heritability estimates for all bacteria are presented in Supplementary File 4.3.

Figure 4.5: SNP heritability estimates for 90 bacteria tested in the nasal vestibule and nasopharynx combined samples (continued)



4.3.5 *mbQTL associations with multiple bacteria*

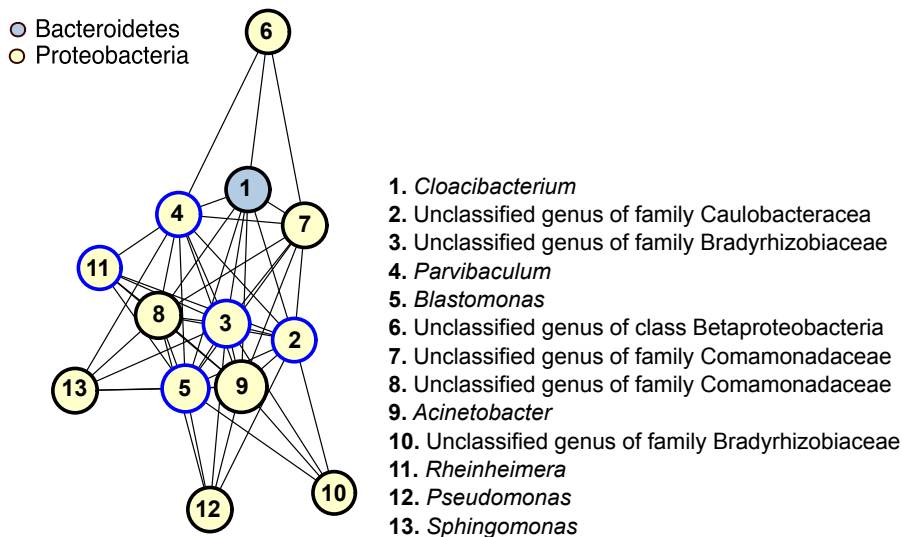
Five mbQTLs ($q < 0.05$), including the PGLYRP4 mbQTL discussed above, had associations with multiple bacteria within the same nasal site at a relaxed significance threshold ($q < 0.10$). The largest number of associations identified with a single mbQTL was an intronic variant in the Leucine Rich Repeat Containing 16A (*LRRC16A*; rs1543603) and the RAs of five Proteobacteria in the nasopharynx in the summer (unclassified genus of family Caulobacteraceae, unclassified genus of family Bradyrhizobiaceae, Parvibaculum, Blastomonas and Rheinheimera; average cumulative sum = 2.1% of all RA and 13.5% of all Proteobacteria), of which Caulobacteraceae was the most significant ($p=2.30 \times 10^{-8}$, $q = 0.006$; Supplementary Figure 4.4B). *LRRC16A* encodes CARMIL (capping protein, Arp2/3, and Myosin-I linker), a protein that plays an important role in cell shape and motility¹⁹¹.

The association between genotype at a single SNP, rs1543603, with the RAs of five genus level bacteria suggested potential functional community level relationships between these five Proteobacteria. Indeed, the RAs of all five Proteobacteria were correlated with each other (correlation coefficients > 0.773 ; median 0.924). Four of the Proteobacteria were classified as Alphaproteobacteria and one as Gammaproteobacteria, classes of bacteria with pathogenic species (e.g. *Inquilinus*, and *Moraxella* and *Haemophilus*, respectively) that have been linked to clinically important airway diseases such as cystic fibrosis¹⁹², COPD^{193,194}, sinusitis¹⁹⁵ or asthma⁵³. A co-occurrence network¹⁹⁶ assigned all five bacteria to a single network that included 13 bacteria (12 Proteobacteria [two Alphaproteobacteria, two Gammaproteobacteria and three Betaproteobacteria] and one Bacteroidetes) from among the 90 genera tested in the nasopharynx

in the summer (Figure 4.6). In this network, genera from families Bradyrhizobiaceae and Caulobacteraceae, two of the five bacteria associated with rs1543603 ($p=3.25 \times 10^{-7}$ and 2.30×10^{-8} , respectively) are the largest hubs with nine neighbors each. We next tested for the association of the sum of these 13 correlated bacteria with rs1543603 but this metric was not more significant ($p=7.72 \times 10^{-5}$) than any of the individual bacteria associations. These findings suggest that host genetic effects can act to modulate microbial community patterns, by directly affecting host-microbe interactions with only one or a few main drivers of the community.

Figure 4.6: Five genus level bacteria associated with rs1543603 are hubs in a co-occurrence module of 12 Proteobacteria and one Bacteroidetes

Co-occurrence networks built from correlation coefficients between all 90 genus level RAs determined in the nasopharynx summer sample. Nodes represent bacteria and are listed by number, colored by phylum and sized proportionally to the RA of each bacterium. Edges represent correlations greater than 0.75. Blue node borders represent the five bacteria associated with rs1543603, an intronic variant in *LRRC16A*.

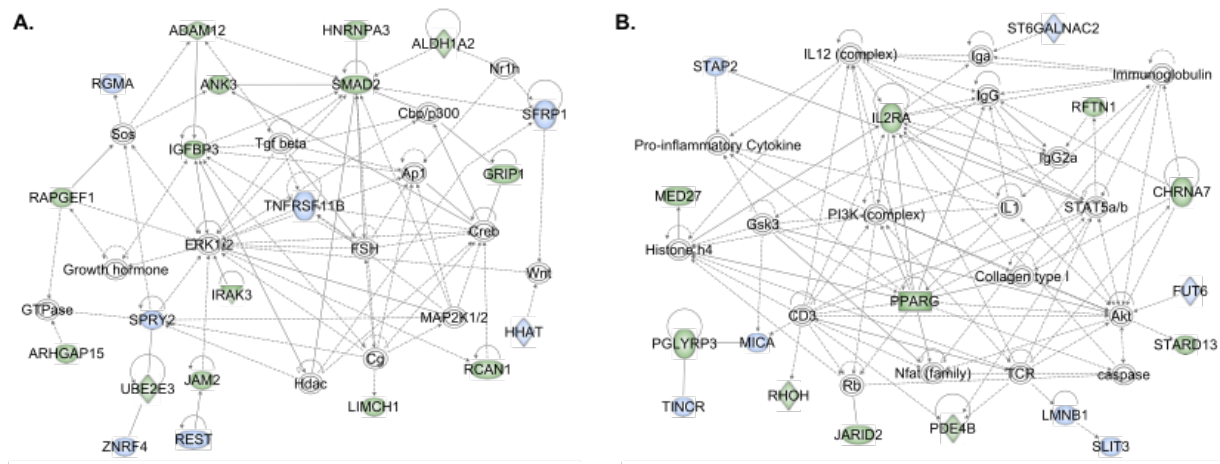


4.3.6 Pathway analyses of genes near mbQTLs

To further understand how host genetic variation regulates nasal microbiome composition and to identify shared pathways among the mbQTLs identified in this study, we selected the closest gene to each mbQTL ($q < 0.10$; 131 genes) and to all variants in LD ($r^2 > 0.8$) with each mbQTL, using LD estimates in the Hutterites. We then generated protein-protein interaction networks among these genes, using Ingenuity Pathway Analysis Knowledge Base (IPA®, QIAGEN Redwood City, CA), a curated database of biological interactions and functional annotations. IPA identified two networks with Fisher exact $p < 10^{-25}$. The most significant network included 21 of the 131 genes, nine of which were near mbQTLs with $q < 0.05$ (Fisher exact $p = 10^{-43}$; Figure 4.7A). This network contained many hubs including SMAD2, a gene that regulates the production of immunoglobulin A (IgA) by LPS-activated B-cells and activates immune response at other mucosal surfaces upon stimulation by pathogenic microbes¹⁹⁷. The second significant network (Fisher exact $p = 10^{-29}$; Figure 4.7B) contained 17 of the 131 genes, also with nine genes near mbQTLs with $q < 0.05$. Many of the hubs in this network represent important modulators of mucosal immunity, including immunoglobulins A and G (IgG and IgG2a), IL12/IL12RA, TCR and STAT5A/B¹⁹⁸⁻²⁰⁰.

Figure 4.7: Ingenuity Pathway Analysis (IPA) interaction networks. Networks show genes near nasal mbQTLs are enriched for mucosal immunity pathways

Two significant networks ($p < 10^{-25}$) are presented. **A. Network one is centered on SMAD2 and ERK1/2** ($p < 10^{-43}$; 21 genes). **B. Network two is a highly connected network centered on IL2RA, STAT5a/b and IL12, among others.** This network contains many of the key regulators of mucosal immunity ($p < 10^{-29}$; 17 genes). Node color represents genes near microbiome QTL associations in the nasal vestibule (blue) or in the nasopharynx (green); open symbols are genes added by IPA. Edges represent direct (solid) and indirect (dashed) interactions in the IPA Knowledge Base database. Node shapes correspond to functional classes of gene products: concentric circles for groups or complexes, diamonds for enzymes, rectangles for transcriptional regulators or modulators, ovals for trans-membrane receptors and circles for other.



4.3.7 Comparison of genes near mbQTLs in the Hutterites to published microbiome QTL studies

To determine if genes associated with the RA of bacteria in the upper airway also influence the RA of bacteria at other human body sites, we compared 53 genes near the 37 mbQTLs discovered in our study ($q < 0.05$) to genes reported in four published microbiome GWAS data^{56-58,201}. The TwinsUK study, the largest human gut microbiome QTL study to date, reported 17 genes within 10 Kb of 15 mbQTLs were associated with 15 different taxa (FDR < 5%). Blekhman *et al.* (2015) identified 34 exonic microbiome QTLs within 32 genes ($q < 0.05$) associated with taxa from within

12 Human Microbiome Project (HMP) body sites. Bonder *et al.* (2016) meta-analyzed three gut microbiome Dutch cohorts and identified nine loci associated with relative abundance and 33 with gene ontology terms and microbial pathways ($p < 5 \times 10^{-8}$). Lastly, Davenport *et al.* (2015) reported seven loci associated with six taxa ($q < 0.05$) in a gut microbiome GWAS in the Hutterites. Only one gene, Slit guidance ligand 3 (*SLIT3*) on chromosome 5, had multiple variants associated among these published studies (rs10055309 in TwinsUK and rs2163761 in Bonder *et al.* 2016). Overall, two genes with mbQTLs identified in this study at $q < 0.05$ have been previously reported in at least one other study: *SLIT3*^{56,201} and Sortilin related VPS10 domain containing receptor 2 (*SORCS2*)²⁰¹. An additional seven genes identified using more relaxed thresholds ($q < 0.10$ in our study and all reported genes in other GWAS studies) are noted in Table 4.2 and Supplementary File 4.2. Overall, there is little overlap between studies, likely reflecting the specificity of microbial communities to local environments and body sites, and the heterogeneity between studies with respect to sample size, geography, body sites, and the race, age and gender composition of the samples.

Nonetheless, three intronic variants in the gene *SLIT3* ($LD\ r^2 < 0.05$) on chromosome 5 were each associated with either the RA of an unclassified genus of family Clostridiaceae in TwinsUK (rs10055309) or the sitosterol degradation to androstenedione pathway (MetaCyc PWY-6948; pathway involved in plant derived steroid degradation) in Bonder *et al.* (2015; rs2163761), and with the RA of genus *Dermacoccus* in the nasal vestibule in the summer in our study (rs77536542; $p = 6.35 \times 10^{-8}$; Figure 4.4A). Both rs10055309 and rs2163761 were the most significant QTL reported in each study. *SLIT3* is a secreted protein that is widely expressed across many tissues

with highest expression in skin, brain cerebellum and lung²⁰². *SLIT3* hypermethylation has been reported in a number of human cancers²⁰² and *SLIT3* expression is increased in LPS stimulated macrophages in mice²⁰³.

A variant 57kb downstream of *SORCS2* on chromosome 4 (rs111354832) was associated in our sample with an unclassified genus of family Micrococcaceae (combined season nasal vestibule) and with *Kaistobacter* (nasal vestibule in the summer) while an intronic variant in *SORCS2* (rs10012347; LD $r^2=0.17$) was associated with the sitosterol degradation to androstenedione pathway (MetaCyc PWY-6948) in Bonder *et al.* (2016). Furthermore, an additional QTL in *SORCS3* (rs703462), a paralog of *SORCS2* on chromosome 10, was associated with *Peptoniphilus* in the combined nasopharynx sample ($q < 0.10$). *SORCS* genes are strongly expressed in the central nervous system and their patterns of expression have been implicated in organ development in mice, including the development of lung and nasal epithelial tissues²⁰⁴. The multiple associations of variants in *SLIT3* and *SORC* genes in the TwinsUK study, Bonder *et al.* (2016) and in our study suggest that these genes may play a role in modulating bacterial abundances in a genotype-specific manner across diverse body sites in humans.

4.4 Materials and Methods

4.4.1 Sample collection

Nasal brushings from Hutterites ages 16 to 78 from five colonies located in South Dakota, all within 14 miles of each other, were collected at two time points, winter (January/February 2011) and summer (July 2011), and from two nasal sites, the nasal vestibule and the nasopharynx. Samples from each of the two nasal sites were collected from opposite nares using sterile flock collection

swabs (Puritan© 25-3316). The HMP anterior nares collection protocol²⁰⁵ was used for the nasal vestibule and the Center of Disease Control and Prevention (CDC) protocol for collection of nasopharyngeal swabs (<http://www.cdc.gov/urdo/downloads/SpecCollectionGuidelines.pdf>) was used to sample the nasopharynx. After excluding samples with low DNA yield, low sequencing read depth, antibiotic history within the prior three months, current use of steroid nasal spray, or missing genotypes, our final data consists of 133 individuals with nasal vestibule samples (87 summer and 80 winter) and 125 individuals with the nasopharynx samples (88 summer and 77 winter; Table 4.1).

4.4.2 Sample DNA extraction library preparation and sequencing

Nasal brushes were immediately frozen at –20°C following collection, shipped on dry ice, and stored at –80°C. DNA extraction was carried out using the BiOstic® Bacteremia DNA Isolation Kit (12240-50). DNA concentration and purity were assessed using the Nanodrop 1000 spectrophotometer (Thermo Scientific, IL, USA). The 16S rRNA gene V4 region was amplified following conditions in Caporaso *et al.* protocol²⁰⁶ using 62 different region-specific primers labeled with a unique 12-base Golay barcode sequence in the reverse primer. Final libraries were quality controlled prior to pooling with the Agilent Bioanalyzer DNA 1000 (Agilent Technologies, CA, USA). Libraries were pooled into 8 pools of 62 samples and sequenced on the HiSeq2000 platform (Illumina Inc, CA, USA) under a single end 102 base pair protocol.

4.4.3 Sequencing and taxonomic classification

Data was pre-processed using CASAVA 1.8.1. Following sample de-multiplexing, 617,909,462 sequence reads were processed using the Quantitative Insights into Microbial Ecology (QIIME)

1.8.0 toolkit²⁰⁷. Quality controlled reads were required to have an exact match to an expected barcode, zero ambiguous base calls, less than three consecutive low quality base calls, and a minimum Phred quality score of 20 along the entire read. We used an open-reference OTU workflow where sequences were first clustered against the Greengenes May 2013 reference²⁰⁸ and reads that did not cluster with known taxa (97% identity) were subjected to de novo clustering. Representative sequences were aligned using PyNAST version 1.2.2²⁰⁹ and the taxonomy of each OTU cluster was assigned with the uclust classifier version 1.2.22q. We applied an OTU abundance filter of 0.005% to reduce spurious OTUs²¹⁰, removing 5.96% of total sequences and obtaining a final dataset of 563 OTUs.

4.4.4 Data processing

Seasonal

For each of the four seasonal groups (nasal vestibule in the summer and in the winter, and nasopharynx in the summer and winter), a genus level RA table was calculated after subsampling reads to 250,000 per sample. Each bacteria's RA was then quantile normalized using the qqnorm function in R. Next, PCA was performed using the prcomp function in R and each of the top 10 principal components (PCs, explaining ~76%-78% of the variance) were tested in a linear model against technical covariates. In at least one of the seasonal groups, we identified correlations between one of the PCs with DNA concentration prior to PCR, final base pair fragment size and date of sampling ($p < 0.001$). PCR adapter barcode, library batch, order within library batch, and final library concentration were not significant. After regressing out the identified technical covariates from the normalized RAs, we performed PCA on the residuals and tested for

associations between the top 10 PCs and biological covariates. Age and sex were significant and were therefore adjusted for in all subsequent analyses. Next, to reduce the burden of multiple testing in the seasonal mapping studies, we removed genera that were detected in fewer than 75% of individuals. This resulted in 78 genus level RAs in the nasal vestibule in the summer, 52 in the nasal vestibule in the winter, 90 in the nasopharynx in the summer and 59 in the nasopharynx in the winter.

Combined seasons

Although combining samples across seasons could introduce noise, it provides the largest possible sample size and consequently greatest power for genetic associations with bacteria that do not vary in abundance across seasons. Therefore, for each nasal site, we averaged the summer and winter genus level RA residuals obtained after quantile normalization and the regression of identified technical covariates for individuals with measurements during both seasons, or included the one season result for those with only one measurement (referred to as the combined sample). We selected all genus level bacteria present in at least 75% of individuals in either season, which resulted in 76 genus level RAs in the nasal vestibule and 90 in the nasopharynx. We performed PCA on the combined seasons matrix to verify variation among samples did not separate the combined samples from the samples with one seasonal measurement. Season of origin (summer, winter or averaged) was not correlated with any of the top 10 principle components in either nasal site.

4.4.5 Genotype data

The Hutterite individuals in our study are related to each other in a 13-generation pedigree that includes 3,671 individuals, all of whom originate from 64 founders. Using PRIMAL⁴⁹, an in-house pedigree-based imputation algorithm, whole genome sequences from 98 Hutterite individuals were phased and imputed to 1,317 Hutterites who were previously genotyped on Affymetrix arrays¹⁴³⁻¹⁴⁵. For mapping studies, we first selected 3,161,460 variants with genotype call rates greater than 95% in our sample and minor allele frequencies (MAF) > 0.10 in any of the 4 season/site subsamples. Next, we estimated LD in the Hutterite data using PLINK⁸⁶, and pruned variants for LD using an r^2 threshold of 0.5, to yield a final set of 148,653 variants for mapping studies.

4.4.6 Diversity metrics

Alpha diversity metrics at the species level (observed species, Shannon index and evenness) were calculated in QIIME²⁰⁷ using the `alpha_diversity.py` script after subsampling reads from 1,000 to 10,000 every 1,000 reads, from 10,000 to 100,000 every 10,000 reads and from 150,000 to 250,000 every 50,000 reads. Each subsampling series was completed 10 times and rarefaction curves were plotted. Diversity metrics were averaged from the 250,000 read subsamples using the `collate_alpha.py` script and this metric was compared across seasons and nasal sites.

To calculate beta diversity, we first obtained the OTU table using `phyloseq`²¹¹, quantile normalized OTU abundances using `qqnorm` in R and regressed out technical covariates (DNA concentration prior to PCR, final base pair fragment size and date of sampling). Next, we calculated pairwise Euclidean distance using the `vegdist` function in the R package `vegan`. We chose Euclidean

distance as our beta diversity metric because this metric does not depend on raw relative abundance values for distance calculation. We wanted to correct our data for significant technical covariates and Euclidean distance can be calculated on residuals after regressing out these technical covariates

4.4.7 Kinship associations to beta diversity

Pair-wise kinship coefficients were previously calculated by PRIMAL⁴⁹ using 271,486 variants genotyped on Affymetrix platforms. The average kinship coefficient between all pairs of individuals (N=144) in our study was 4.51% (range 0.60%-32.03%). We performed 10,000 permutations to assess the association between pairwise Euclidian distances and kinship coefficients in the combined seasons samples (nasal vestibule 8,778 pairs, nasopharynx 7,750). The p-value is the number of times out of 10,000 permutations that the Spearman correlation of the permuted sequence pair was more extreme than the observed pair.

4.4.8 Co-occurrence network analyses

We used SparCC¹⁹⁶ to calculate nasopharynx in the summer correlation coefficients between all 90 genera tested in our mapping studies. We applied default settings and assigned p-values calculated from 100 bootstraps. Co-occurrence networks were generated from the SparCC correlation matrix for genera with correlation $r^2 > 0.75$ and $p < 0.01$ (1/100 bootstraps). The network was generated using igraph R package, where nodes represent each genera and edges represent correlations between the genera above the applied threshold.

4.4.9 Ingenuity Pathway Analysis (IPA) of protein-protein interaction networks

We selected the closest gene to 1,413 variants (131 genes) with Hutterite linkage disequilibrium (LD) $r^2 > 0.8$ with the 108 mbQTLs ($q < 0.10$). To interrogate and visualize network associations, we used the Ingenuity Pathway Analysis Knowledge Base (IPA®, QIAGEN Redwood City, CA), limiting interactions to primary cells or tissues. The network scores generated by IPA are based on a right-tailed Fisher's Exact test comparing the observed and expected mbQTL genes present in a pathway relative to the IPA database.

4.5 Discussion

Our study is the first to assess the role of genome-wide host genetic variation in shaping the human microbiome at two upper airway sites. We first demonstrated reduced bacterial beta diversity between more closely related pairs of individuals, and then discovered associated genetic variants at functionally related genes. These combined results indicate a significant role for host genotype in patterning microbial diversity in the nose.

Our results further suggest that the upper airway may be the site of important gene-environment interactions. In this context, host genotype at many loci may ultimately impact health and disease by modulating particular members of the microbial community. For example, a missense variant in fucosyltransferase 3 (*FUT3*; rs28362459) was strongly associated with decreased abundance of *Demacoccus* (nasal vestibule in the summer), a bacteria that is depleted in the skin of individuals with atopic dermatitis¹⁸². This SNP is predicted to be deleterious by both Polymorphism Phenotyping (PolyPhen) v2⁸⁵ and Combined Annotation Dependent Depletion (CADD)³² scores (0.997 and 15.12, respectively). Interestingly, the non-secretor phenotype, characterized by a null variant in another FUT gene, *FUT2*, and the resulting absence of ABH

antigens in the mucosa in homozygotes for the null allele, influences the composition and diversity of the microbiome in the human intestinal tract^{212,213}. Moreover, variants in both *FUT2* and *FUT3* have been shown in GWAS to increase susceptibility to diseases associated with both mucosal surface pathobiology and microbiome composition, such as cystic fibrosis¹⁸⁶, Crohn's disease²¹⁴ and ulcerative colitis²¹⁵. Our study extends a role for fucosyltransferases to the nasal mucosal surface, and further implicates host genetic influences on bacterial diversity at this site.

Four mbQTLs show effects on phylogenetically diverse phyla and two were identified in different seasons and nasal sites. In particular, a missense variant in *PGLYRP4* (rs3006458) was associated with the abundance of genus *Aerococcus* (Firmicutes) in the nasopharynx in the summer and with family Micrococcaceae (Actinobacteria) in the nasopharynx in the combined sample. The RAs of *Aerococcus* and Micrococcaceae are only weakly correlated with each other indicating that these are likely independent associations. Moreover, the associations with phylogenetically distant bacteria and in different subsamples (summer vs. combined) suggest that *PGLYRP4* has pleiotropic effects over several organisms. Alternatively, the SNP identified in our study (rs3006458) could be tagging a haplotype with multiple variants that have independent effects on different bacterial abundances. The genomic region that includes the PGLYRP genes includes a cluster of genes implicated in epidermal barrier function²¹⁶ and SNPs in this region show extensive LD. However, independent evidence suggests that *PGLYRP4* may be the target gene of this association. The rs3006458-T allele, which is associated with lower RA of *Aerococcus* and Micrococcaceae in our study, was associated with increased gene expression of the *PGLYRP4* gene in epithelial and mucosal tissues (skin, small intestine and esophageal mucosa) in the Genotype-

Tissue Expression¹³², and in lung tissue in a separate eQTL study⁷⁹. These tissues serve as physical barriers and provide innate immune functions essential for antimicrobial defense²¹⁷. Collectively, these data suggest that host genotype at rs3006458 (or a variant in LD with rs3006458) regulates the expression in *PGLYRP4* in the skin, lung and airway mucosa and functions to modulate bacterial abundance, possibly beyond the two genera identified in this study. The link between genetic variation in host PGRPs and microbiome abundance revealed in this study indicates that at least some of the important role these proteins play in modulating communities of symbiotic organisms²¹⁸ is attributable to host genetic variation.

Eight of the mbQTLs identified in our study at a relaxed $q < 0.10$ influenced the abundance of more than one organism. Most were identified within the same season and nasal site and four influenced the abundance of multiple closely related bacteria. For example, an intronic variant in *LRRC16A* (rs1543603) was associated with the abundance of five highly correlated genera of phylum Proteobacteria in the nasopharynx in the summer and co-occur in a larger network of 12 Proteobacteria and one Bacteroidetes. One possible explanation for this correlated network may be that these bacteria are physically interacting and their overall community structure is influenced by host genotype, but there may be other metabolic or physiological reasons why these bacteria co-occur. Although not much is known about *LRRC16A*, other proteins with leucine-rich repeat (LRR) domains, such as nucleotide-binding oligomerization domain receptors (NODs) and toll-like receptors (TLRs)²¹⁹, function as recognition receptors in innate immunity.

Although our study provides novel insights into host genetic influences on the nasal microbiome, there are some limitations. In particular, the size of our sample is relatively small for

genetic mapping (77-88 individuals in seasonal analyses; 125 and 133 individuals in the combined). We reasoned that the reduced environmental heterogeneity among Hutterite individuals would enhance the effects of genetic variation and facilitate the detection of associated variants. While we were successful in identifying mbQTLs, we acknowledge that there are likely many more associations to be found in larger samples. A second limitation is the multiple testing burden that results from the high dimensionality of the microbiome. While we reduced the number of tests performed by mapping only genus level bacteria present in the majority of individuals, we only corrected for multiple testing within each study and did not correct for the 90 bacteria and six groups for which we performed mbQTL mapping. Although we used a fairly stringent threshold of genome-wide significance ($q < 0.05$), we acknowledge that some of our findings may be false positives. In addition, there is the possibility that samples from the nasopharynx may have been contaminated by the anterior nares; these regions are, indeed, anatomically and physiologically distinct. Our sampling method was motivated by the practicalities of field research and utilized standard clinical techniques. We note that only one study to date has surveyed the regional biogeography of the nasal microbiome; they included the anterior nares and two more posterior nasal sites (middle meatus, and sphenoethmoidal recess, which are distinct from the nasopharynx, though also lined by respiratory epithelium) using a similar sampling technique¹⁷⁷. Our findings of separation between the anterior and posterior nasal sites across some measures and variables (most prominently in the summer) are broadly consistent with their results, which show some differences and some overlap across sites. Nevertheless, precise spatial analyses of the nose require additional studies using alternative sampling techniques. Lastly, the genetic

effects revealed by our study are context specific due to the many environmental and stochastic factors that affect microbiome composition and, therefore, challenging to replicate. For example, even within our study of a relatively homogenous population, we detected significant effects of season, age, and gender. In fact, most of the mbQTLs that we identified were specific to one season and demonstrate that even small temporal changes (six months) in the RAs of bacteria within the same individuals can mask or enhance genetic effects. Although we did not formally replicate the results in independent populations, the identification of multiple intronic variants within the gene *SLIT3* in our study, the TwinsUK study⁵⁶, and Bonder *et al.* (2016)²⁰¹ bolsters confidence in the involvement of this gene in regulating microbial structure across multiple mucosal sites.

Our study provides evidence for genetic contributions to modulating variability of the nasal microbiome, a trait that has been linked to a number of airway diseases²²⁰. Importantly, our findings support the concept that host genetic variation directly influences the expression or function of genes that are specifically involved in innate mucosal immunity pathways. Such a framework is consistent with previous reports showing that antimicrobial peptides^{218,221} and host immunity²²² are key modulators of microbial defense in the mucosa. Our data further suggest that host genetic effects on immune genes modulate particular bacteria or the structure of whole microbial communities in the upper airways. We speculate that interactions between host genetics and microbiome structure or composition in the upper airway can influence dysbiotic tendencies that may predispose to respiratory disease and could be subject to intervention. Indeed, moving forward, more detailed analyses of the complex relationship between genetic

variation in host mucosal immunity and the microbiome — captured here in a snapshot in the upper airway — are required to fully characterize determinants of an inherently dynamic microbial ecosystem. Such work could potentially identify targets for novel therapeutic strategies useful across a wide range of respiratory diseases.

4.6 Acknowledgements

We would like to thank members of the Hutterite community for their continuous participation in our studies. We would also acknowledge members of the Ober, Gilbert and Gilad labs for useful discussions; Sally Cain and Rebecca Anderson for assistance with sample collection; Amy Mitrano for library preparation; James Lane for technical support; and Jack Gilbert for analysis advice.

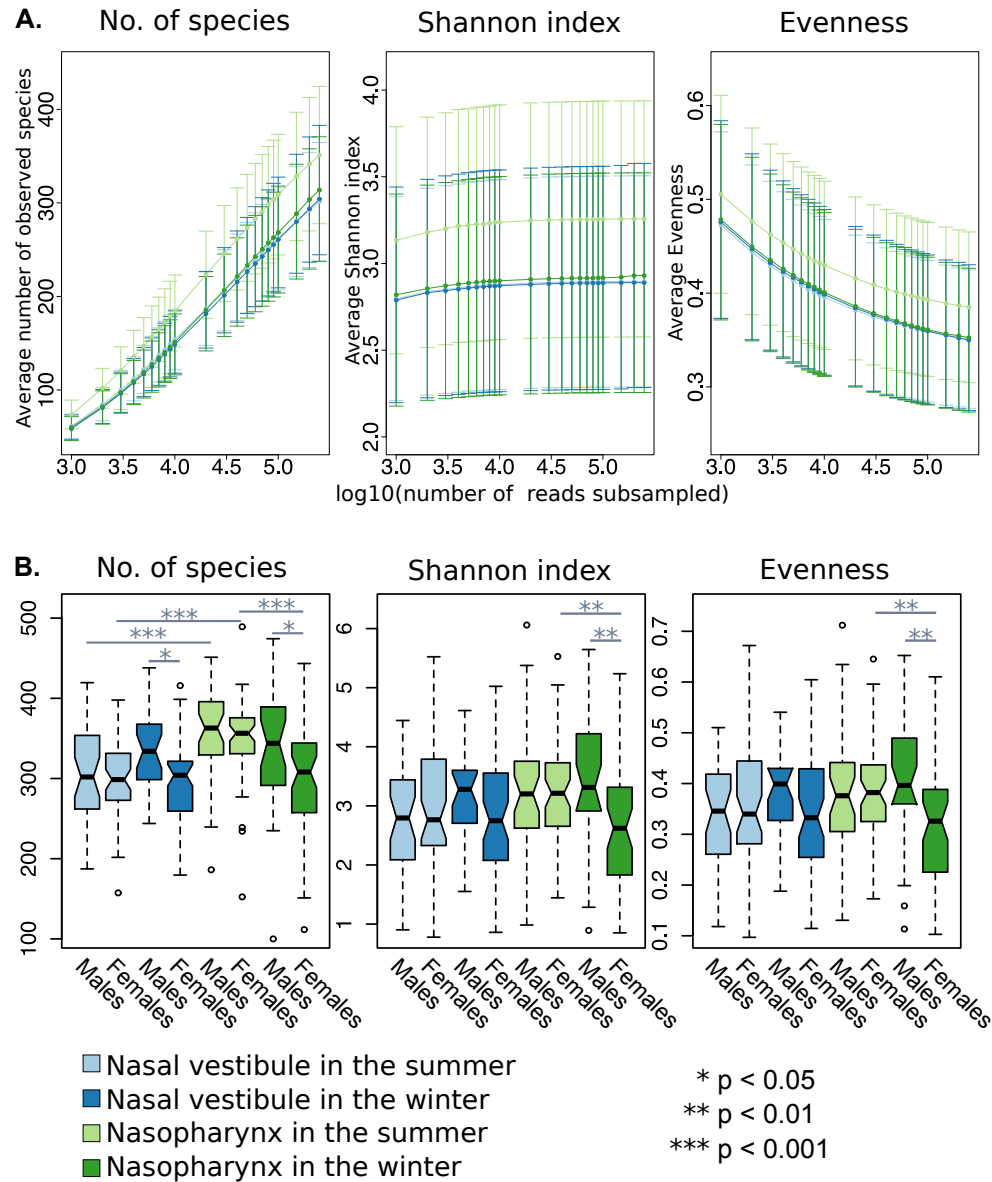
4.7 Appendix C: Supplementary Data

Supplementary File 4.1: Relative abundance (available online)

A. Distributions of 166 genus level nasal microbiome relative abundances. RA distribution summaries of are presented for the nasal vestibule and the nasopharynx for the summer and winter sample. Min, minimum; Qu, quantile; Max, maximum. **B-C. Results of seasonal differences in genus level relative abundances.** 12 genera in the nasal vestibule and 15 in the nasopharynx differed in relative abundance (RA) by season, nine of which were different at both nasal sites. The percent of individuals with detectible bacteria (RA > 0) is presented in parenthesis. *Represents bacteria significant after correcting for 166 tests.

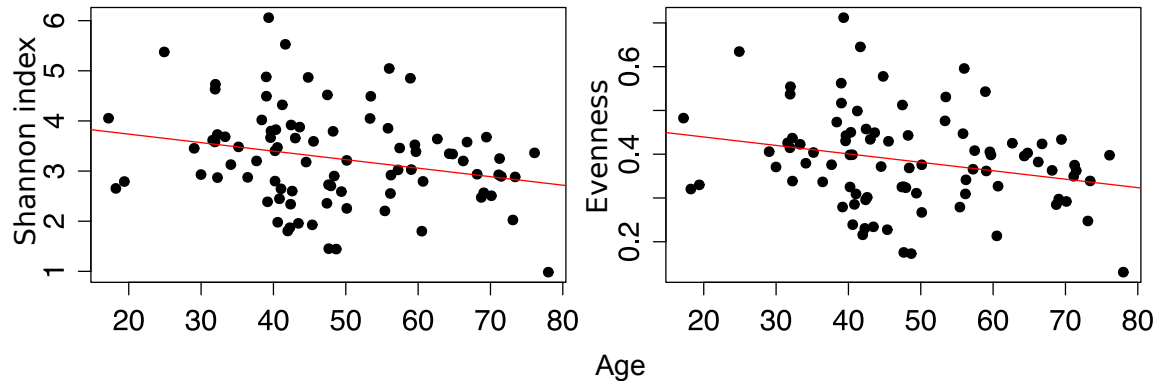
Supplementary Figure 4.1: Nasal microbiome alpha diversity

A. Rarefaction curves. Sequencing reads were subsampled 10 times from 1,000 to 10,000 every 1,000 reads, from 10,000 to 100,000 thousand every 10,000 reads and from 150,000 to 250,000 every 50,000 reads (x-axis). **B. Alpha diversity by sex.** For both sexes, the number of species is lower in the nasal vestibule in the summer (light blue) compared to the nasopharynx in the summer (light green). Within females, alpha diversity is lower in the nasopharynx in the winter (dark green) compared to the nasopharynx in the summer (light green). Compared to males, females have lower number of species in the nasal vestibule in the winter (dark blue) and overall lower alpha diversity in the nasopharynx in the winter (dark green). P values correspond to a Wilcoxon rank sum test.



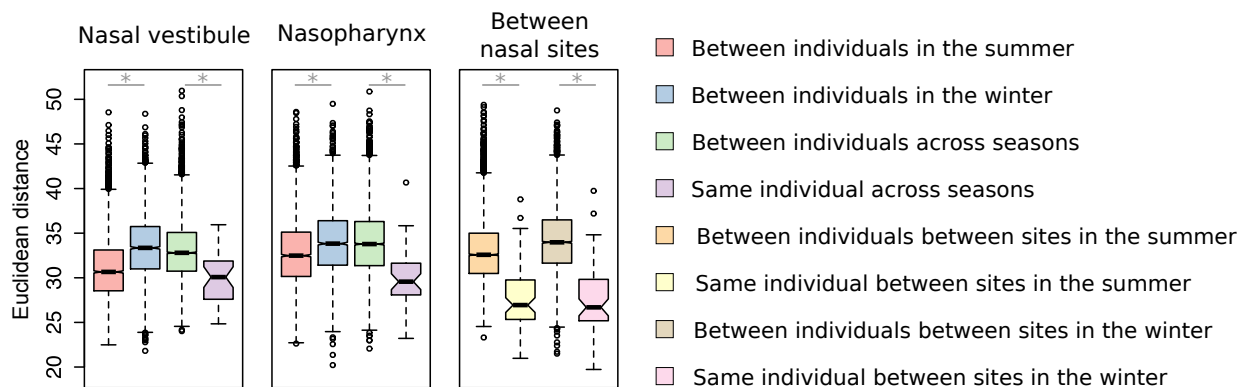
Supplementary Figure 4.2. Correlation of alpha diversity with age in the nasopharynx in the summer

Age is negatively correlated with Shannon index ($p=0.019$) and evenness ($p=0.019$), in the nasopharynx in the summer. Red line represents line of fit from a linear model.



Supplementary Figure 4.3. Boxplots of beta diversity grouped by nasal site and between nasal sites

For each of the nasal sites (panels 1 and 2), beta diversity (Euclidean distance) was lower in the summer (red) compared to the winter (blue) and between seasons for the same individual (purple) compared to all other pairs of individuals (green). In nasal site analyses (panel 3), beta diversity was lower for the same individual between nasal sites compared to all other pairs of individuals (summer: yellow vs. orange; winter: pink vs. brown). *Wilcox rank sum test $p < 10^{-7}$.

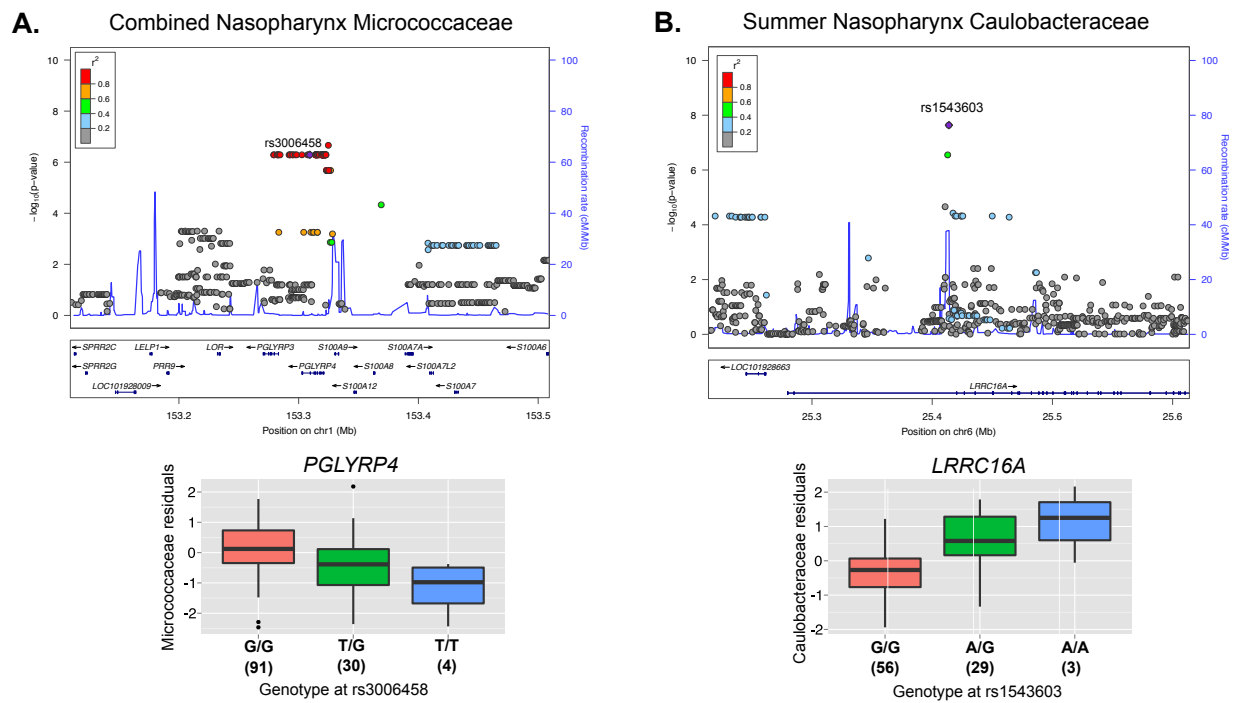


Supplementary File 4.2: QTL mapping results of nasal microbiome relative abundance (available online)

108 microbiome quantitative trait loci (mbQTLs) were associated at a relaxed significance threshold of $q < 0.10$ with the relative abundance of 37 microbial taxa. Direction of effect is presented for the minor allele. Significant seasonal results are bolded. RA, relative abundance; Chr, chromosome.

Supplementary Figure 4.4: Locus and genotype plots for selected mbQTLs

A. rs3006458 is a mbQTL associated with lower RA of an unclassified genus of family Micrococcaceae in the combined sample. B. rs1543603 is a mbQTL associated with higher RA of an unclassified genus of family Caulobacteraceae in the nasopharynx in the summer. Variants on locus plot represent Hutterite variants with MAF > 10% prior to LD pruning. Numbers underneath each boxplot represent the number of individuals in each genotype class.



Supplementary File 4.3: Nasal vestibule and nasopharynx SNP heritability estimates (available online)

Estimates for the combined season sample were calculated using GEMMA. PVE: percent variance explained; SE: standard error.

CHAPTER 5

CONCLUSIONS

5.1 Summary and Significance

Technology advances in recent years have allowed tremendous progress for complex trait mapping, with many efforts directed towards novel analytic approaches and rare variant studies. In this dissertation, I aimed to study genetic contributions to complex traits that are not traditionally considered in GWAS. More specifically, I explored the role of rare variation in asthma and blood lipid traits, and host genotype influences on the airway microbiome.

In chapter 2, I described a meta-analysis of rare and low frequency variants in asthma as part of the EVE Asthma Genetics Consortium. At the time of publication, this was the largest study assessing the effects of rare variants on asthma risk in ethnically diverse individuals, and only the second meta-analysis to use the Illumina HumanExome BeadChip (following the group that designed the array¹). In this study, I identified a low-frequency missense mutation in the *GRASP* gene (rs7974426) associated with increased asthma risk in Latinos, and six additional low frequency variants that were modestly associated ($p < 10^{-4}$) in at least one of the three ethnic groups or the combined sample. I then tested the aggregate effects of multiple loss-of-function variants present within single genes and revealed gene-level associations with genes at the known asthma locus on chromosome 17q21 (*GSDMB* and *ZPBP2*) in the Latino and combined sample, and with the *MTHFR* gene on chromosome 1 in the African ancestry sample. The few significant associations detected in our study were all limited to a single ethnic group, despite having the

greatest power in the combined sample, suggesting that rare or low frequency variants influencing asthma risk are likely to be ethnic specific. In addition, although we were limited by the variation present on the exome array and by relatively small sample sizes within each ethnic group, the few associations we did find had moderate to low effect sizes, indicating that rare and low frequency variants are not likely to account for a significant portion of the heritability of asthma. The results of this study were published in *Nature Communications* (Ethnic-specific associations of rare and low-frequency DNA sequence variants with asthma. *Nat Commun* 6, 5965 (2015)).

In chapter 3, I applied hierarchical Bayesian modeling to perform a functional GWAS of blood lipid traits (HDL-C, LDL-C, TG and cholesterol levels) using variants that are absent or rare in Europeans but drifted to higher frequencies in the Hutterites. I identified four novel rare variants associated with LDL-C or HDL-C levels, two of which were in novel loci, and replicated the effects of a known protective splicing variant in *APOC3*^{3,4} associated with TG and HDL-C levels. All of the putative candidate variants revealed in our study were in non-coding regions of the genome and many were associated with gene expression in LCLs from the Hutterites. Our results both highlight the importance of studying the genome-wide effects of rare non-coding variation on common disease, and suggest that these variants may influence disease risk through their effects on gene expression.

In chapter 4, I assessed the role of common host genetic variation in shaping the human microbiome at two upper airway sites (the nasal vestibule and nasopharynx) in two seasons (summer and winter). A key motivation for this study was that while the influences of environmental exposures on microbiome composition and their association to disease risk are well

studied, the role that host genetics plays a role in structuring microbial communities is less well understood. A few human studies have shown host genetics plays an important role in shaping microbiome composition⁵⁻⁷ but to date there has not been a comprehensive study in the airways. In our study, we first demonstrated reduced bacterial beta diversity between more closely related pairs of individuals and then discovered associated genetic variants at functionally related genes. The majority of the variants we identified were specific to one season or nasal site and demonstrate the results obtained in microbiome GWAS studies are context specific, and that even small temporal or spatial changes within the same individuals can influence bacterial abundance and mask or enhance genetic effects. Overall, our study showed that the structure of the airway microbiome is influenced by host genotype at many loci and revealed host variation in innate and mucosal immune pathway genes play a crucial role in this process. More generally, we showed that genome-wide association studies of the microbiome can capture multidimensional interactions between microbial communities and the host genome. This study was published in *Microbiome* (Host genetic variation in mucosal immunity pathways influences the upper airway microbiome. *Microbiome* 5, 16 (2017)).

5.2 Future Directions

5.2.1 Functional studies of rare non-coding variation

One of the fundamental challenges in complex trait genetics is translating known variant associations into disease mechanisms. For monogenic diseases, it has been relatively easier to identify target genes, as disease mutations commonly disrupt the function of a single protein.

However, functional prediction for non-coding variation is still in its infancy and the effects of genetic variation on gene expression are particularly lacking for rare variants.

In fact, the scarcity of functional annotations in non-coding regions was a limitation in the design and interpretation of the rare variant studies in this dissertation (chapters 2 and 3). In chapter 2, the studies of rare and low frequency variants on asthma focused on loss-of-function variation surrounding coding regions of the genome discovered in the Exome Sequencing Project (ESP)⁹ and were predicted to be functional based on allele frequency and consequence (missense, nonsense, spicing). In chapter 3, all the identified rare variants associated with lipid traits lied in non-coding regions and we were unable to use publicly available eQTL databases to assess their effects on gene expression in relevant tissues.

Our results suggest that rare non-coding variants mediate their effects on disease risk through gene expression, similar to common, non-coding variants, but few guidelines exist to link rare genetic variation to the target genes. Studies of allele-specific expression and emerging high throughput assays capable of quantifying the activity of candidate regulatory elements are a promising next step to understand what regulatory regions do rare non-coding variants have the most influence on gene regulation. As more comprehensive tissue specific functional and regulatory maps emerge, the annotations of non-coding regulatory variation will be refined and when integrated with disease association variants should reveal fundamental insight into disease mechanisms.

5.2.2 Linking microbial structure and host genetics to health and disease

Microbiome GWAS offer a means to untangle the mechanisms that link host genetics to communities of microbes, but how these interactions influence disease is not well understood. Current microbiome mapping studies have focused on the detection of host genetic influences on relative abundance for taxa (e.g. species, genus) detected in study samples, but the system is complex and therefore interpretations tend to be over simplified. For instance, relative abundance metrics do not reflect bacterial load, growth or decay because each bacteria level is measured relative to the total bacteria present in a sample and growth of a single species would therefore result in reduced relative abundance of all other species. In addition, the interpretation of a mbQTL association with a single taxon can be challenging because instead of a direct influence of host genotype on bacteria abundance, this result may be capturing a complex network between multiple bacteria within a community and the host (as seen in the associations presented in chapter 4 between a *LRRC16A* variant and 5 proteobacteria).

Fundamentally, the relevant measurement is not the relative abundance of single bacterial species, nor even the absolute load. Instead, the effect of the microbiome is a result of dynamically interacting communities of bacteria and this effect can be mediated by both the environment and host genotype. A complete understanding of the relationship between the microbiome and disease in the host will require deep characterization of the interactions between microbes, the environment and the host, as well as the expression profiles of both the host and the microbial communities.

Dissecting these interactions requires integrative 'omic studies. For example, studies of genetic variation integrated with microbial and host metabolomes could shed light on how different microbial communities influence metabolic and immune states in the host and how genetic variation may play a role regulating this relationship. Understanding these molecular states and the influences that perturb the feedback mechanisms between microbial communities and the host over time will start to reveal which microbe-microbe and host-microbe interactions are critical in the disease process.

REFERENCES

1. Kerem, B.-S. Identification of the cystic fibrosis gene: genetic analysis. *Trends in Genetics* 5, 363 (1989).
2. Corder, E. H. *et al.* Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* 261, 921–923 (1993).
3. Lander, E. S. & Botstein, D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* 236, 1567–1570 (1987).
4. Vissers, L. E. L. M. *et al.* Mutations in a new member of the chromodomain gene family cause CHARGE syndrome. *Nat. Genet.* 36, 955–957 (2004).
5. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* 42, 30–35 (2010).
6. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* 461, 747–753 (2009).
7. Gusev, A. *et al.* Quantifying missing heritability at known GWAS loci. *PLoS Genet.* 9, e1003993 (2013).
8. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. U.S.A.* 109, 1193–1198 (2012).
9. Global Lipids Genetics Consortium *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* 45, 1274–1283 (2013).
10. Lander, E. S. The new genomics: global views of biology. *Science* 274, 536–539 (1996).
11. DeWan, A. T. *et al.* Whole-exome sequencing of a pedigree segregating asthma. *BMC Med. Genet.* 13, 95 (2012).
12. Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* 13, 135–145 (2012).
13. Pritchard, J. K. Are Rare Variants Responsible for Susceptibility to Complex Diseases? *The American Journal of Human Genetics* 69, 124–137 (2001).
14. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 493, 216–220 (2013).

15. Keinan, A. & Clark, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336, 740–743 (2012).
16. Nelson, M. R. *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 337, 100–104 (2012).
17. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337, 64–69 (2012).
18. Mu, X. J., Lu, Z. J., Kong, Y., Lam, H. Y. K. & Gerstein, M. B. Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project. *Nucleic Acids Res.* 39, 7058–7076 (2011).
19. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073 (2010).
20. Li, Y. *et al.* Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat. Genet.* 42, 969–972 (2010).
21. Soutar, A. K. & Naoumova, R. P. Mechanisms of disease: genetic causes of familial hypercholesterolemia. *Nat Clin Pract Cardiovasc Med* 4, 214–225 (2007).
22. Weedon, M. N. & Frayling, T. M. Insights on pathogenesis of type 2 diabetes from MODY genetics. *Curr Diab Rep* 7, 131–138 (2007).
23. Easton, D. F. *et al.* A Systematic Genetic Assessment of 1,433 Sequence Variants of Unknown Clinical Significance in the BRCA1 and BRCA2 Breast Cancer–Predisposition Genes. *The American Journal of Human Genetics* 81, 873–883 (2007).
24. King, C. R. & Nicolae, D. L. GWAS to Sequencing: Divergence in Study Design and Analysis. *Genes (Basel)* 5, 460–476 (2014).
25. Igartua, C. *et al.* Ethnic-specific associations of rare and low-frequency DNA sequence variants with asthma. *Nat Commun* 6, 5965 (2015).
26. Huyghe, J. R. *et al.* Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat. Genet.* 45, 197–201 (2013).
27. Peloso, G. M. *et al.* Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am. J. Hum. Genet.* 94, 223–232 (2014).
28. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* Chapter 7, Unit7.20–7.20.41 (2013).

29. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812–3814 (2003).
30. Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* 6, e1001025 (2010).
31. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15, 1034–1050 (2005).
32. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46, 310–315 (2014).
33. Bansal, V., Libiger, O., Torkamani, A. & Schork, N. J. Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.* 11, 773–785 (2010).
34. Nicolae, D. L. Association Tests for Rare Variants. <http://dx.doi.org/10.1146/annurev-genom-083115-022609> 17, 117–130 (2016).
35. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* 95, 5–23 (2014).
36. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* 83, 311–321 (2008).
37. Neale, B. M. *et al.* Testing for an unusual distribution of rare variants. *PLoS Genet.* 7, e1001322 (2011).
38. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 89, 82–93 (2011).
39. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* 91, 224–237 (2012).
40. King, C. R., Rathouz, P. J. & Nicolae, D. L. An evolutionary framework for association testing in resequencing studies. *PLoS Genet.* 6, e1001202 (2010).
41. Lin, W.-Y., Lou, X.-Y., Gao, G. & Liu, N. Rare variant association testing by adaptive combination of P-values. *PLoS ONE* 9, e85728 (2014).
42. Cheung, Y. H., Wang, G., Leal, S. M. & Wang, S. A fast and noise-resilient approach to detect rare-variant associations with deep sequencing data for complex disorders. *Genet. Epidemiol.* 36, 675–685 (2012).
43. Madsen, B. E. & Browning, S. R. A Groupwise Association Test for Rare Mutations Using a

Weighted Sum Statistic. *PLoS Genet.* 5, e1000384 (2009).

44. Pollin, T. I. *et al.* A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. *Science* 322, 1702–1705 (2008).
45. Peltonen, L., Palotie, A. & Lange, K. Use of population isolates for mapping complex traits. *Nat. Rev. Genet.* 1, 182–190 (2000).
46. Ober & Cox. Mapping genes for complex traits in founder populations. *Clin. Exp. Allergy* 28, 101–105 (1998).
47. Hostetler, J. A., Opitz, J. M. & Reynolds, J. F. History and relevance of the Hutterite population for genetic studies. *American Journal of Medical Genetics Part A* 22, 453–462 (1985).
48. Boycott, K. M. *et al.* Clinical genetics and the Hutterite population: A review of Mendelian disorders. *American Journal of Medical Genetics Part A* 146A, 1088–1098 (2008).
49. Livne, O. E. *et al.* PRIMAL: Fast and accurate pedigree-based imputation from sequence data in a founder population. *PLoS Comput. Biol.* 11, e1004139 (2015).
50. Torgerson, D. G. *et al.* Resequencing candidate genes implicates rare variants in asthma susceptibility. *Am. J. Hum. Genet.* 90, 273–281 (2012).
51. Morgan, X. C. & Huttenhower, C. Chapter 12: Human microbiome analysis. *PLoS Comput. Biol.* 8, e1002808 (2012).
52. Phillips, M. L. Gut reaction: environmental effects on the human microbiota. *Environmental health perspectives* 117, A198–205 (2009).
53. Huang, Y. J. *et al.* Airway microbiota and bronchial hyperresponsiveness in patients with suboptimally controlled asthma. *J. Allergy Clin. Immunol.* 127, 372–381.e1–3 (2011).
54. Bogaert, D. *et al.* Variability and Diversity of Nasopharyngeal Microbiota in Children: A Metagenomic Analysis. *PLoS ONE* 6, e17035 (2011).
55. Bosch, A. A. T. M., Biesbroek, G., Trzcinski, K., Sanders, E. A. M. & Bogaert, D. Viral and bacterial interactions in the upper respiratory tract. *PLoS Pathog.* 9, e1003057 (2013).
56. Goodrich, J. K. *et al.* Genetic Determinants of the Gut Microbiome in UK Twins. *Cell Host Microbe* 19, 731–743 (2016).
57. Blekhman, R. *et al.* Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.* 16, 191 (2015).

58. Davenport, E. R. *et al.* Genome-Wide Association Studies of the Human Gut Microbiota. *PLoS ONE* 10, e0140301 (2015).
59. Benson, A. K. *et al.* Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc. Natl. Acad. Sci. U.S.A.* 107, 18933–18938 (2010).
60. Thomsen, S. F., van der Sluis, S., Kyvik, K. O., Skytthe, A. & Backer, V. Estimates of asthma heritability in a large twin sample. *Clin. Exp. Allergy* 40, 1054–1061 (2010).
61. Duffy, D. L., Martin, N. G., Battistutta, D., Hopper, J. L. & Mathews, J. D. Genetics of asthma and hay fever in Australian twins. *Am. Rev. Respir. Dis.* 142, 1351–1358 (1990).
62. Palmer, L. J., Burton, P. R., James, A. L., Musk, A. W. & Cookson, W. O. Familial aggregation and heritability of asthma-associated quantitative traits in a population-based sample of nuclear families. *Eur. J. Hum. Genet.* 8, 853–860 (2000).
63. Torgerson, D. G. *et al.* Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. *Nat. Genet.* 43, 887–892 (2011).
64. Moffatt, M. F. *et al.* A large-scale, consortium-based genomewide association study of asthma. *N. Engl. J. Med.* 363, 1211–1221 (2010).
65. Bønnelykke, K. *et al.* A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. *Nat. Genet.* 46, 51–55 (2014).
66. Hirota, T. *et al.* Genome-wide association study identifies three new susceptibility loci for adult asthma in the Japanese population. *Nat. Genet.* 43, 893–896 (2011).
67. Ober, C. & Vercelli, D. Gene–environment interactions in human disease: nuisance or opportunity? *Trends in Genetics* (2011).
68. Myers, R. A. *et al.* Further replication studies of the EVE Consortium meta-analysis identifies 2 asthma risk loci in European Americans. *J. Allergy Clin. Immunol.* 130, 1294–1301 (2012).
69. Alavanja, M. C. *et al.* The Agricultural Health Study. *Environ. Health Perspect.* 104, 362–369 (1996).
70. Halapi, E. *et al.* A sequence variant on 17q21 is associated with age at onset and severity of asthma. *Eur. J. Hum. Genet.* 18, 902–908 (2010).
71. Lluís, A. *et al.* Asthma-associated polymorphisms in 17q21 influence cord blood ORMDL3 and GSDMA gene expression and IL-17 secretion. *J. Allergy Clin. Immunol.* 127, 1587–94.e6 (2011).

72. Moffatt, M. F. *et al.* Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* 448, 470–473 (2007).
73. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44, 821–824 (2012).
74. Horvath, S., Xu, X. & Laird, N. M. The family based association test method: strategies for studying general genotype–phenotype associations. *Eur. J. Hum. Genet.* 9, 301–306 (2001).
75. Himes, B. E. *et al.* Genome-wide association analysis identifies PDE4D as an asthma-susceptibility gene. *Am. J. Hum. Genet.* 84, 581–593 (2009).
76. Galanter, J. M. *et al.* Genome-wide association study and admixture mapping identify different asthma-associated loci in Latinos: The Genes-environments & Admixture in Latino Americans study. *Journal of Allergy and Clinical Immunology* 134, 295–305 (2014).
77. Granell, R. *et al.* Examination of the relationship between variation at 17q21 and childhood wheeze phenotypes. *J. Allergy Clin. Immunol.* 131, 685–694 (2013).
78. Calışkan, M. *et al.* Rhinovirus wheezing illness and genetic risk of childhood-onset asthma. *N. Engl. J. Med.* 368, 1398–1407 (2013).
79. Hao, K. *et al.* Lung eQTLs to help reveal the molecular underpinnings of asthma. *PLoS Genet.* 8, e1003029 (2012).
80. Murphy, A. *et al.* Mapping of numerous disease-associated expression polymorphisms in primary peripheral blood CD4+ lymphocytes. *Hum. Mol. Genet.* 19, 4745–4757 (2010).
81. Sharma, S. *et al.* A genome-wide survey of CD4(+) lymphocyte regulatory genetic variants identifies novel asthma genes. *J. Allergy Clin. Immunol.* (2014). doi:10.1016/j.jaci.2014.04.011
82. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc. Natl. Acad. Sci. U.S.A.* 111, E455–64 (2014).
83. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* 11, 415–425 (2010).
84. Lee, S., Teslovich, T. M., Boehnke, M. & Lin, X. General framework for meta-analysis of rare variants in sequencing association studies. *Am. J. Hum. Genet.* 93, 42–53 (2013).
85. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249 (2010).

86. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575 (2007).
87. Gibbs, R. A. *et al.* The International HapMap Project. *Nature* 426, 789–796 (2003).
88. Kirke, P. N. *et al.* Impact of the MTHFR C677T polymorphism on risk of neural tube defects: case-control study. *BMJ* 328, 1535–1536 (2004).
89. Newton-Cheh, C. *et al.* Genome-wide association study identifies eight loci associated with blood pressure. *Nat. Genet.* 41, 666–676 (2009).
90. Wu, X. *et al.* Association between the MTHFR C677T polymorphism and recurrent pregnancy loss: a meta-analysis. *Genet Test Mol Biomarkers* 16, 806–811 (2012).
91. Husemoen, L. L. N. *et al.* The association between atopy and factors influencing folate metabolism: is low folate status causally related to the development of atopy? *Int J Epidemiol* 35, 954–961 (2006).
92. Thuesen, B. H., Husemoen, L. L. N., Fenger, M. & Linneberg, A. Lack of association between the MTHFR (C677T) polymorphism and atopic disease. *Clin Respir J* 3, 102–108 (2009).
93. Granell, R. *et al.* The association between mother and child MTHFR C677T polymorphisms, dietary folate intake and childhood atopy in a population-based, longitudinal birth cohort. *Clin. Exp. Allergy* 38, 320–328 (2008).
94. Håberg, S. E. *et al.* Maternal folate levels in pregnancy and asthma in children at age 3 years. *J. Allergy Clin. Immunol.* 127, 262–4– 264.e1 (2011).
95. Håberg, S. E., London, S. J., Stigum, H., Nafstad, P. & Nystad, W. Folic acid supplements in pregnancy and early childhood respiratory health. *Arch. Dis. Child.* 94, 180–184 (2009).
96. Nevriy, D. J. *et al.* Interaction of GRASP, a protein encoded by a novel retinoic acid-induced gene, with members of the cytohesin family of guanine nucleotide exchange factors. *J. Biol. Chem.* 275, 16827–16836 (2000).
97. Zhang, Y. *et al.* Roles of scaffolding protein Gab1 in regulation of myeloid dendritic cell-mediated allergic asthma (HYP7P.281). *The Journal of Immunology* 192, 119.4–119.4 (2014).
98. Walker, J. K. L. *et al.* Beta-arrestin-2 regulates the development of allergic asthma. *J. Clin. Invest.* 112, 566–574 (2003).
99. Verlaan, D. J. *et al.* Targeted screening of cis-regulatory variation in human haplotypes. *Genome Res.* 19, 118–127 (2009).

100. Morrison, F. S. *et al.* The splice site variant rs11078928 may be associated with a genotype-dependent alteration in expression of GSDMB transcripts. *BMC Genomics* 14, 627 (2013).
101. Saeki, N. *et al.* Distinctive expression and function of four GSDM family genes (GSDMA-D) in normal and malignant upper gastrointestinal epithelium. *Genes Chromosomes Cancer* 48, 261–271 (2009).
102. Shah, T. S. *et al.* optiCall: a robust genotype-calling algorithm for rare, low-frequency and common variants. *Bioinformatics* 28, 1598–1603 (2012).
103. Go, A. S., Mozaffarian, D., Roger, V. L. & Benjamin, E. J. AHA statistical update. *Circulation* (2013).
104. Kathiresan, S. *et al.* A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC Med. Genet.* 8 Suppl 1, S17 (2007).
105. Weiss, L. A., Pan, L., Abney, M. & Ober, C. The sex-specific genetic architecture of quantitative traits in humans. *Nat. Genet.* 38, 218–222 (2006).
106. Bielinski, S. J. *et al.* Genome-wide linkage scans for loci affecting total cholesterol, HDL-C, and triglycerides: the Family Blood Pressure Program. *Hum. Genet.* 120, 371–380 (2006).
107. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 466, 707–713 (2010).
108. Pérusse, L. *et al.* Familial resemblance of plasma lipids, lipoproteins and postheparin lipoprotein and hepatic lipases in the HERITAGE Family Study. *Arterioscler. Thromb. Vasc. Biol.* 17, 3263–3269 (1997).
109. Rao, D. C. *et al.* The Cincinnati Lipid Research Clinic family study: cultural and biological determinants of lipids and lipoprotein concentrations. *Am. J. Hum. Genet.* 34, 888–903 (1982).
110. Malhotra, A. *et al.* Meta-Analysis of Genome-Wide Linkage Studies of Quantitative Lipid Traits in Families Ascertained for Type 2 Diabetes. *Diabetes* 56, 890–896 (2007).
111. Cohen, J. C. Multiple Rare Alleles Contribute to Low Plasma Levels of HDL Cholesterol. *Science* 305, 869–872 (2004).
112. Zappala, Z. & Montgomery, S. B. Non-Coding Loss-of-Function Variation in Human Genomes. *Human Heredity* 81, 78–87 (2017).
113. Montgomery, S. B., Lappalainen, T., Gutierrez-Arcelus, M. & Dermitzakis, E. T. Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS*

- Genet.* 7, e1002144 (2011).
114. Zhao, J. *et al.* A Burden of Rare Variants Associated with Extremes of Gene Expression in Human Peripheral Blood. *The American Journal of Human Genetics* 98, 299–309 (2016).
 115. Li, X. *et al.* The impact of rare variation on gene expression across tissues. *bioRxiv* 074443 (2016). doi:10.1101/074443
 116. Peltonen, L. Molecular background of the Finnish disease heritage. *Ann. Med.* 29, 553–556 (1997).
 117. Jonsson, T. *et al.* A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature* 488, 96–99 (2012).
 118. Helgadottir, A. *et al.* Variants with large effects on blood lipids and the role of cholesterol and triglycerides in coronary disease. *Nat. Genet.* 48, 634–639 (2016).
 119. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–45 (2016).
 120. TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute *et al.* Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *N. Engl. J. Med.* 371, 22–31 (2014).
 121. Timpson, N. J. *et al.* A rare variant in APOC3 is associated with plasma triglyceride and VLDL levels in Europeans. *Nat Commun* 5, 4871 (2014).
 122. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* 94, 559–573 (2014).
 123. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* 95, 535–552 (2014).
 124. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* 47, 1228–1235 (2015).
 125. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* 507, 455–461 (2014).
 126. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330 (2015).
 127. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* 45, 124–130 (2013).

128. Below, J. E. *et al.* Meta-analysis of lipid-traits in Hispanics identifies novel loci, population-specific effects, and tissue-specific enrichment of eQTLs. *Sci Rep* 6, 19429 (2016).
129. Schierer, A. *et al.* Genetic variation in cholesterol ester transfer protein, serum CETP activity, and coronary artery disease risk in Asian Indian diabetic cohort. *Pharmacogenet. Genomics* 22, 95–104 (2012).
130. Boes, E., Coassin, S., Kollerits, B., Heid, I. M. & Kronenberg, F. Genetic-epidemiological evidence on genes associated with HDL cholesterol levels: a systematic in-depth review. *Exp. Gerontol.* 44, 136–160 (2009).
131. Cusanovich, D. A. *et al.* Integrated analyses of gene expression and genetic association studies in a founder population. *Hum. Mol. Genet.* 25, 2104–2112 (2016).
132. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648–660 (2015).
133. Huston, E., Houslay, T. M., Baillie, G. S. & Houslay, M. D. cAMP phosphodiesterase-4A1 (PDE4A1) has provided the paradigm for the intracellular targeting of phosphodiesterases, a process that underpins compartmentalized cAMP signalling. *Biochem. Soc. Trans.* 34, 504–509 (2006).
134. Fairoozy, R. H., White, J., Palmen, J., Kalea, A. Z. & Humphries, S. E. Identification of the Functional Variant(s) that Explain the Low-Density Lipoprotein Receptor (LDLR) GWAS SNP rs6511720 Association with Lower LDL-C and Risk of CHD. *PLoS ONE* 11, e0167676 (2016).
135. Linsel-Nitschke, P. *et al.* Lifelong reduction of LDL-cholesterol related to a common variant in the LDL-receptor gene decreases the risk of coronary artery disease--a Mendelian Randomisation study. *PLoS ONE* 3, e2986 (2008).
136. Badano, J. L., Leitch, C. C., Ansley, S. J. & May-Simera, H. Dissection of epistasis in oligogenic Bardet–Biedl syndrome. *Nature* (2006).
137. Imhoff, O., Marion, V., Stoetzel, C. & Durand, M. Bardet-Biedl syndrome: a study of the renal and cardiovascular phenotypes in a French cohort. *Clinical Journal of the ...* (2011).
138. Burnett, J. R. & Hooper, A. J. Common and rare gene variants affecting plasma LDL cholesterol. *Clin Biochem Rev* 29, 11–26 (2008).
139. Cohen, J. C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* 305, 869–872 (2004).
140. Cohen, J. C., Stender, S. & Hobbs, H. H. APOC3, coronary disease, and complexities of Mendelian randomization. *Cell Metab.* 20, 387–389 (2014).

141. Blair, D. R. *et al.* A Nondegenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk. *Cell* 155, 70–80 (2013).
142. Chong, J. X., Ouwenga, R., Anderson, R. L., Waggoner, D. J. & Ober, C. A population-based study of autosomal-recessive disease-causing mutations in a founder population. *Am. J. Hum. Genet.* 91, 608–620 (2012).
143. Ober, C. *et al.* Effect of variation in CHI3L1 on serum YKL-40 level, risk of asthma, and lung function. *N. Engl. J. Med.* 358, 1682–1691 (2008).
144. Yao, T.-C. *et al.* Genome-wide association study of lung function phenotypes in a founder population. *J. Allergy Clin. Immunol.* 133, 248–55.e1–10 (2014).
145. Cusanovich, D. A. *et al.* The combination of a genome-wide association study of lymphocyte count and analysis of gene expression data reveals novel asthma candidate genes. *Hum. Mol. Genet.* 21, 2111–2123 (2012).
146. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291 (2016).
147. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* 526, 68–74 (2015).
148. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* 17, 122 (2016).
149. Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* 41, 827–841 (2013).
150. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* 155, 934–947 (2013).
151. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478, 476–482 (2011).
152. Ward, L. D. & Kellis, M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* 337, 1675–1678 (2012).
153. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214 (2012).
154. Nicholson, J. K. *et al.* Host-Gut Microbiota Metabolic Interactions. *Science* 336, 1262–1267 (2012).
155. Chow, J., Lee, S. M., Shen, Y., Khosravi, A. & Mazmanian, S. K. Host-bacterial symbiosis in health and disease. *Adv. Immunol.* 107, 243–274 (2010).

156. Boase, S. *et al.* The microbiome of chronic rhinosinusitis: culture, molecular diagnostics and biofilm detection. *BMC Infectious Diseases* 13, 1 (2013).
157. Pragman, A. A., Kim, H. B., Reilly, C. S. & Wendt, C. The lung microbiome in moderate and severe chronic obstructive pulmonary disease. *PLoS ONE* (2012).
158. Hilty, M. *et al.* Disordered microbial communities in asthmatic airways. *PLoS ONE* 5, e8578 (2010).
159. Denner, D. R. *et al.* Corticosteroid therapy and airflow obstruction influence the bronchial microbiome, which is distinct from that of bronchoalveolar lavage in asthmatic airways. *Journal of Allergy and Clinical Immunology* (2015). doi:10.1016/j.jaci.2015.10.017
160. Marsland, B. J. & Gollwitzer, E. S. Host-microorganism interactions in lung diseases. *Nat. Rev. Immunol.* 14, 827–835 (2014).
161. Arrieta, M.-C., Stiemsma, L. T., Amenyogbe, N., Brown, E. M. & Finlay, B. The Intestinal Microbiome in Early Life: Health and Disease. *Frontiers in Immunology* 5, 355 (2014).
162. Gensollen, T., Iyer, S. S., Kasper, D. L. & Blumberg, R. S. How colonization by microbiota in early life shapes the immune system. *Science* 352, 539–544 (2016).
163. Thaiss, C. A., Zmora, N., Levy, M. & Elinav, E. The microbiome and innate immunity. *Nature* 535, 65–74 (2016).
164. Dominguez-Bello, M. G. *et al.* Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences* 107, 11971–11975 (2010).
165. Biesbroek, G. *et al.* The Impact of Breastfeeding on Nasopharyngeal Microbial Communities in Infants. *Am. J. Respir. Crit. Care Med.* 190, 298–308 (2014).
166. Noverr, M. C., Noggle, R. M., Toews, G. B. & Huffnagle, G. B. Role of Antibiotics and Fungal Microbiota in Driving Pulmonary Allergic Responses. *Infection and Immunity* 72, 4996–5003 (2004).
167. Suárez-Arrabal, M. C. *et al.* Nasopharyngeal bacterial burden and antibiotics: Influence on inflammatory markers and disease severity in infants with respiratory syncytial virus bronchiolitis. *Journal of Infection* 71, 458–469 (2015).
168. Morris, A. *et al.* Comparison of the respiratory microbiome in healthy nonsmokers and smokers. *Am. J. Respir. Crit. Care Med.* 187, 1067–1075 (2013).
169. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* 486, 222–227 (2012).

170. Tims, S. *et al.* Microbiota conservation and BMI signatures in adult monozygotic twins. *ISME J* 7, 707–717 (2013).
171. Goodrich, J. K., Davenport, E. R., Waters, J. L., Clark, A. G. & Ley, R. E. Cross-species comparisons of host genetic associations with the microbiome. *Science* 352, 532–535 (2016).
172. Yatsunenko, T., Rey, F. E., Manary, M. J. & Trehan, I. Human gut microbiome viewed across age and geography. *Nature* 486, 222–227 (2012).
173. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* 457, 480–484 (2009).
174. Ober, C., Abney, M. & McPeck, M. S. The genetic dissection of complex traits in a founder population. *Am. J. Hum. Genet.* 69, 1068–1079 (2001).
175. Davenport, E. R. *et al.* Seasonal variation in human gut microbiome composition. *PLoS ONE* 9, e90731 (2014).
176. Teo, S. M. *et al.* The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development. *Cell Host Microbe* 17, 704–715 (2015).
177. Yan, M. *et al.* Nasal Microenvironments and Interspecific Interactions Influence Nasal Microbiota Complexity and *S. aureus* Carriage. *Cell Host Microbe* 14, 631–640 (2013).
178. Stover, C. S. & Litwin, C. M. The Epidemiology of Upper Respiratory Infections at a Tertiary Care Center: Prevalence, Seasonality, and Clinical Symptoms. *Journal of Respiratory Medicine* 2014, 1–8 (2014).
179. Leung, R. K. K. *et al.* Modulation of potential respiratory pathogens by pH1N1 viral infection. *Clinical Microbiology and Infection* 19, 930–935 (2013).
180. Allen, E. K. *et al.* Characterization of the nasopharyngeal microbiota in health and during rhinovirus challenge. *Microbiome* 2, 22 (2014).
181. Lax, S. *et al.* Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science* 345, 1048–1052 (2014).
182. Chng, K. R. *et al.* Whole metagenome profiling reveals skin microbiome-dependent susceptibility to atopic dermatitis flare. *Nature Microbiology* 1, 16106 (2016).
183. L Shivlata, T. S. Thermophilic and alkaliphilic Actinobacteria: biology and potential applications. *Frontiers in Microbiology* 6, 753 (2015).

184. Kretz, M. *et al.* Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature* 493, 231–235 (2013).
185. AlMatar, M. *et al.* Are There Any Other Compounds Isolated From *Dermacoccus* spp at All? *Curr Microbiol* 74, 1–13 (2016).
186. Taylor-Cousar, J. L. *et al.* Histo-blood group gene polymorphisms as potential genetic modifiers of infection and cystic fibrosis lung disease severity. *PLoS ONE* 4, e4270 (2009).
187. Yamamoto, F. *et al.* An integrative evolution theory of histo-blood group ABO and related genes. *Sci Rep* 4, 6601 (2014).
188. Kashyap, D. R. *et al.* Peptidoglycan recognition proteins kill bacteria by activating protein-sensing two-component systems. *Nat. Med.* 17, 676–683 (2011).
189. Lawson, P. A., Falsen, E., Ohlén, M. & Collins, M. D. *Aerococcus urinaehominis* sp. nov., isolated from human urine. *International Journal of Systematic and Evolutionary Microbiology* 51, 683–686 (2001).
190. Kadioglu, A., Weiser, J. N., Paton, J. C. & Andrew, P. W. The role of *Streptococcus pneumoniae* virulence factors in host respiratory colonization and disease. *Nat. Rev. Microbiol.* 6, 288–301 (2008).
191. Yang, C. *et al.* Mammalian CARMIL inhibits actin filament capping by capping protein. *Dev. Cell* 9, 209–221 (2005).
192. Schmoldt, S. *et al.* Clonal analysis of *Inquilinus limosus* isolates from six cystic fibrosis patients and specific serum antibody response. *J. Med. Microbiol.* 55, 1425–1433 (2006).
193. Sze, M. A., Hogg, J. C. & Sin, D. D. Bacterial microbiome of lungs in COPD. *International Journal of COPD* 9, 229–238 (2014).
194. Sze, M. A. *et al.* The Lung Tissue Microbiome in Chronic Obstructive Pulmonary Disease. *Am. J. Respir. Crit. Care Med.* 185, 1073–1080 (2012).
195. Payne, S. C. & Benninger, M. S. *Staphylococcus aureus* Is a Major Pathogen in Acute Bacterial Rhinosinusitis: A Meta-Analysis. *Clinical Infectious Diseases* 45, e121–e127 (2007).
196. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* 8, e1002687 (2012).
197. Malhotra, N. & Kang, J. SMAD regulatory networks construct a balanced immune system. *Immunology* 139, 1–10 (2013).

198. Macpherson, A. J., McCoy, K. D., Johansen, F.-E. & Brandtzaeg, P. The immune geography of IgA induction and function. *Mucosal Immunol* 1, 11–22 (2008).
199. Holt, P. G., Strickland, D. H., Wikström, M. E. & Jahnsen, F. L. Regulation of immunological homeostasis in the respiratory tract. *Nat. Rev. Immunol.* 8, 142–152 (2008).
200. Mantis, N. J. & Forbes, S. J. Secretory IgA: Arresting Microbial Pathogens at Epithelial Borders. *Immunol. Invest.* 39, 383–406 (2010).
201. Bonder, M. J. *et al.* The effect of host genetics on the gut microbiome. *Nat. Genet.* 48, 1407–1412 (2016).
202. Dickinson, R. E. *et al.* Epigenetic inactivation of SLIT3 and SLIT1 genes in human cancers. *British Journal of Cancer* 91, 2071–2078 (2004).
203. Tanno, T. *et al.* Slit3 regulates cell motility through Rac/Cdc42 activation in lipopolysaccharide-stimulated macrophages. *FEBS Letters* 581, 1022–1026 (2007).
204. Boggild, S., Molgaard, S., Glerup, S. & Nyengaard, J. R. Spatiotemporal patterns of sortilin and SorCS2 localization during organ development. *BMC Cell Biol.* 17, 8 (2016).
205. Aagaard, K. *et al.* The Human Microbiome Project strategy for comprehensive sampling of the human microbiome and why it matters. *FASEB J.* 27, 1012–1022 (2013).
206. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* 6, 1621–1624 (2012).
207. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336 (2010).
208. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* 72, 5069–5072 (2006).
209. Caporaso, J. G. *et al.* PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* 26, 266–267 (2010).
210. Bokulich, N. A. *et al.* Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* 10, 57–59 (2013).
211. McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* 8, e61217 (2013).
212. Wacklin, P. *et al.* Secretor Genotype (FUT2 gene) Is Strongly Associated with the Composition of Bifidobacteria in the Human Intestine. *PLoS ONE* 6, e20113 (2011).

213. Rausch, P. *et al.* Colonic mucosa-associated microbiota is influenced by an interaction of Crohn disease and FUT2 (Secretor) genotype. *Proceedings of the National Academy of Sciences* 108, 19030–19035 (2011).
214. Hu, D.-Y. *et al.* Associations of FUT2 and FUT3 gene polymorphisms with Crohn's disease in Chinese patients. *J. Gastroenterol. Hepatol.* 29, 1778–1785 (2014).
215. Hu, D. *et al.* Association of Ulcerative Colitis with FUT2 and FUT3 Polymorphisms in Patients from Southeast China. *PLoS ONE* 11, e0146557 (2016).
216. Toulza, E. *et al.* Large-scale identification of human genes implicated in epidermal barrier function. *Genome Biol.* 8, R107 (2007).
217. Gallo, R. L. & Hooper, L. V. Epithelial antimicrobial defence of the skin and intestine. *Nat. Rev. Immunol.* 12, 503–516 (2012).
218. Royet, J., Gupta, D. & Dziarski, R. Peptidoglycan recognition proteins: modulators of the microbiome and inflammation. *Nat. Rev. Immunol.* 11, 837–851 (2011).
219. Ng, A. C. Y. *et al.* Human leucine-rich repeat proteins: a genome-wide bioinformatic categorization and functional analysis in innate immunity. *Proceedings of the National Academy of Sciences* 108, 4631–4638 (2011).
220. Dickson, R. P., Erb-Downward, J. R. & Huffnagle, G. B. The role of the bacterial microbiome in lung disease. *Expert Review of Respiratory Medicine* 7, 245–257 (2014).
221. Salzman, N. H. *et al.* Enteric defensins are essential regulators of intestinal microbial ecology. *Nat. Immunol.* 11, 76–83 (2010).
222. Hooper, L. V. & Macpherson, A. J. Immune adaptations that maintain homeostasis with the intestinal microbiota. *Nat. Rev. Immunol.* 10, 159–169 (2010).