The Phenotypic Consequences of Distinct Genetic Variation in an Electronic Medical Record

By

Rebecca Terrall Levinson

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in

Human Genetics
December, 2016
Nashville, Tennessee

Approved:

Melinda C. Aldrich, Ph.D., M.P.H.

Joshua C. Denny, M.D., M.S.

Bingshan Li, Ph.D.

Douglas P. Mortlock, Ph.D.

David C. Samuels, Ph.D.

I would like to dedicate my dissertation to my parents and my partner Bill Martin. Thank you for supporting me in graduate school and in life.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the assistance of many individuals. I would like to acknowledge the members of the Stein lab including QiPing Feng, Vivian Kawai, Abiodun Adefurin, Daniel Kurnik, Cecelia Chung, and Mike Stein for their scientific input and community. I would also like to thank Ryan Delehanty and Alicia Beeghly-Fadiel for helping me design some of the projects in this dissertation. Rich D'Aquila and Isabelle Clerc have been essential, offering me biological input on the role of the APOBEC3 gene family. I would also like to thank all the members of the Denny lab who have answered questions or provided me data, especially Lisa Bastarache and Robert Carroll.

Thanks to all the students in the Human Genetics program, specifically Sabrina Mitchell, Olivia Veatch, and especially Laura Wiley who have provided a sounding board for many of my not very well thought out ideas, advised me through my academic successes and frustrations, and provided me with helpful insights on what I could do in the future.

I would like to recognize my friends who have shared the moments of levity and/or alcohol consumption that have made my six years in graduate school possible. Jeannie Camarillo, Daniel LePage, Tim Shaver, Ashley Francis, and Samantha Mayden, thanks for always being there.

Last, and most importantly, I would like to thank my mentor David Samuels. My path through graduate school has not been as rigidly guided as many of my peers. Thank you for letting me interact with people across and outside the university to follow my interests and scientific questions, for being there to provide needed reality checks or to let me complain about

a project, and for never stopping me from trying something just because it was outside your area of expertise. I have learned a lot over the past years, and I think much of it was due to your willingness to let me step outside your comfort zone and find who or whatever I needed to get a project done. Thank you so much.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

LIST OF APPENDICES

INTRODUCTION

The Genomic Era and Genomics in Medical Records

In 2003, the conclusion of the human genome project officially brought a start to the "genomic era" [1]. With the start of the genomic era came the beginning of widespread and multiplex searches for genetic variants that caused diseases. Simultaneously, the presence of a human reference sequence allowed for the development of new genotyping platforms and new analysis methods, including the development of genome-wide association studies (GWAS).

GWAS tests the association of one phenotype with all genotypes available from GWAS chips. GWAS chips use tagging single nucleotide polymorphisms (SNPs) as a proxy for variants that were in linkage disequilibrium, so that not all variants had to be tested[2]. In 2005 the first successful GWASs were published[3–5]. Despite this initial success, and other success stories through the years, GWAS has failed to identify as much of the heritability of common diseases as expected[6]. The variants identified in GWAS as associated with a disease are not necessarily the causal variants, rather the associated SNP may be linked to the causal variant. Even in well powered studies, the process of identifying the causal gene and variant extends well after the performance of GWAS.

Importantly, GWAS studies frequently used ascertained case-control populations, which required significant time and money to assemble. Groups of cases with shared controls were developed, but this was not feasible for all phenotypes that investigators had found to be heritable and expected a genetic effect. Performing studies in an electronic medical record (EMR), also called an electronic health record (EHR), can be very cost effective compared to assembling a case control population[7].

The predecessors of electronic medical records appeared as early as the 1960s. Now, more than 50 years later, their uptake continues today. A study from 2009 estimated that only 1.5% of hospitals had a comprehensive electronic health record system [8]. The Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 directed the Office of the National Coordinator for Health Information Technology (ONC) to promote the adoption and meaningful use of electronic health records. More recent estimates indicate that 3 out of 4 hospitals have a basic EHR system[9]. A variety of information is available from EHRs. Billing data, laboratory data, vitals, medication lists, test results, and provider notes are all typically part of EMR data [10]. Billing data most often consists of International Classification of Disease (ICD) and Current Procedural Terminology (CPT) codes. EHRs also contain demographic information including age, sex, and physician reported race.

ICD codes are developed and maintained by the World Health Organization (http://www.who.int/classifications/icd/en/HistoryOfICD.pdf). They are a hierarchical way to classify diseases and symptoms. The original purpose was to obtain morbidity and mortality statistics from around the world. The ICD system has been adapted for use in billing by hospitals and insurance companies. ICD is currently in version 10. CPT codes are a way to identify clinical services and procedures for a medical encounter. They are designed and maintained by the American Medical Association[11].

Importantly, the information contained in medical records is not designed for research purposes. The information in medical records is also subject to provider preferences and habits, both at the physician and institutional level. Coding expectations and preferences may vary from one medical center to another. Despite this, phenotyping algorithms developed at one medical center have been adapted for use at others, though a skilled team is often necessary to

successfully deploy it[10]. Additionally, if incorrect information gets in to a medical record, it can be difficult to remove it. Medical records are not edited, rather they are amended, with a clarifying or correcting note added. This not may not be seen by providers, so incorrect information may be propagated through an individuals' record. Information in the medical record may also depend on the accuracy of information a patient may tell their provider. Patients may give an incomplete or incorrect medical history which can influence notes in a record. Despite this, there is evidence that studies performed in the EMR can robustly replicate studies performed in more traditionally assembled cohorts[12], and that this phenotyping is of sufficient quality for genetic analyses[13].

Many of the replication attempts using EMRs focus on associations discovered through GWAS. An EHR based study of 21 SNPs implicated in 5 common diseases found that the direction of effect of all SNPs tested was in the direction expected. In each disease at least one previously reported SNP association was replicated[11]. This study also manually reviewed their cases and found that their algorithms had greater than a 95% PPV for each of the 5 phenotypes studied. Another study on Rheumatoid Arthritis used a multi-ethnic EHR derived cohort to replicate a genetic risk score for rheumatoid arthritis[14]. Some reports also indicate that the longitudinal nature of the data in the EMR can assist in differentiating between related phenotypes[12].

BioVU and the Synthetic Derivative

The Synthetic Derivative (SD) is a de-identified mirror of the EMR from Vanderbilt University Medical Center. Several steps are taken to ensure de-identification of records. Records numbers go through a one-way hash to be assigned a number for that cannot be traced

back to the patient. Additionally, all Health Insurance Portability and Accountability Act (HIPAA) identifiers are removed from records. This includes names of individuals (both patients and health care providers). All dates in the record are shifted by 1-365 days, consistent within a record but variable between records. De-identified records are then available for research purposes. Due to the de-identification procedures, the SD has received a non-human subjects designation[15].

The Vanderbilt DNA Databank, BioVU, uses blood samples leftover from clinical testing as a source of DNA[16]. Samples banked are blood remnants that would otherwise be discarded. Initially, the DNA databank operated using an "opt-out" model. A statement was included in the consent to treat form describing the databank, and patients could opt out by checking a box. This resource was linked to de-identified medical records in the SD. In January of 2015, the consent process shifted to an opt-in model. Patients now have the option to consent to their samples being added to the DNA Databank. The blood samples are linked to de-identified entries in the SD. Individuals who had undergone hypertransfusion or bone marrow transplant are flagged as having compromised samples. Blood samples are available to be pulled for genotyping, and some samples have pre-existing genotype data available.

Genotyping has been performed on a subset of samples within BioVU. This genotyping is paid for by individual investigators for individual research projects, so specific phenotype groups are genotyped in specific platforms leading to potential phenotype differences in the populations with data available from any given genotyping platform. The existing genotype information is everything from Taqman on single SNPs to platforms targeting SNPs in specific pathways, to GWAS platforms. Investigators can apply to use existing data for new projects.

PheWAS: Theory and Methods

The underlying goal of PheWAS is to test the association of one variant with a range of phenotypes. In addition to the discovery of new disease genotype associations, PheWAS allows for the discovery of pleiotropy, the ability of a single gene or genetic variant to influence multiple traits, and an increased understanding of how phenotypes are related to each other[17]. PheWAS can also help differentiate between pleiotropy and comorbidities, and in some cases can differentiate between disease subtypes[18]. Many existing associations between genotypes and phenotypes have been replicated with PheWAS. PheWAS have also identified novel SNP-phenotype associations. Despite this, some argue that one of the major accomplishments of PheWAS is the establishment of workflows to analyze and visualize complex and multi-dimensional phenotype-genotype relationships[19].

PheWAS using ICD-9 codes for case control classification have repeatedly replicated or validated genotype-phenotype associations. The first PheWAS paper replicated seven SNP disease associations with p-values of at least 0.05 and odds ratios in the same direction as previous studies[20]. Another study focused on systematically comparing PheWAS results to associations previously found in GWAS, mapped existing significantly associated GWAS traits to PheWAS phenotypes, determined that they were able to replicate 28% of all tested associations with a p-value <0.05 and a consistent direction of effect[21]. Once the authors filtered this to only binary GWAS traits with an exact match in the PheWAS catalog that was adequately powered in their study, they were able to replicate 66% of SNP-phenotype associations tested.

Existing PheWAS have successfully shown variants to be pleiotropic. The idea of a single genetic factor resulting in multiple phenotypic outcomes is not new; studies going back many years have evaluated pleiotropy[22]. GWAS provides information about the correlation

between genomic variants, but analyzing only one phenotype at a time does not provide any information about the interconnected nature of phenotypes and disease outcomes. While pleiotropy has not been systematically studies in human complex trait genetics, it has been observed in human Mendelian disease genetics[23]. Several studies have looked at the influence of single SNPs on multiple related phenotypes such as immune phenotypes[24] or cancers[25,26]. A recent study evaluated 42 traits with GWAS data, and found 341 pleiotropic loci; some of which were associated with unexpected phenotypes given the gene function[27]. Pleiotropy can influence multiple diseases through distinct pathways or because one disease is in the pathway of another. Looking across the entirety of diseases in the medical record has the potential to enhance our understanding of biology by drawing connections between different phenotypes. What is unique about PheWAS is situating it in the breadth of phenotype data available in an individual's EMR.

As there are many types of information contained in the EMR, a PheWAS could test any type of information that could be obtained for a sufficient number of subjects. PheWAS have been published using aggregated International Classification of Disease version 9 (ICD-9) codes[20], but natural language processing (NLP) obtained strings, and lab values, or CPT codes are all possible starting points for phenotype determination. The PheWAS discussed in this dissertation will all use the most basic ICD-9 code aggregation method as a starting point.
A scheme of aggregating and further hierarchically clustering ICD-9 codes into PheWAS codes (Phecodes) was developed and published in the first PheWAS manuscript[20]. Over time, the aggregation has been update several times. Using instances of the ICD-9 codes and aggregation rules, an individual may be classified as a case, control, or exclusion for a specific PheWAS code (Figure 0-A). Individuals can be excluded for multiple reasons: they can have some, but fewer than the desired number of ICD-9 codes in their record; or they can appear to be a control, but be

a case for something highly related. These exclusions can be used to limit falsely classifying individuals as either a case or a control when there is uncertainty or messiness in the record. These case/control classifications are then used as the outcome in logistic regression. Alternatively, the number of ICD-9 codes in a record can be used as an outcome in linear regression.

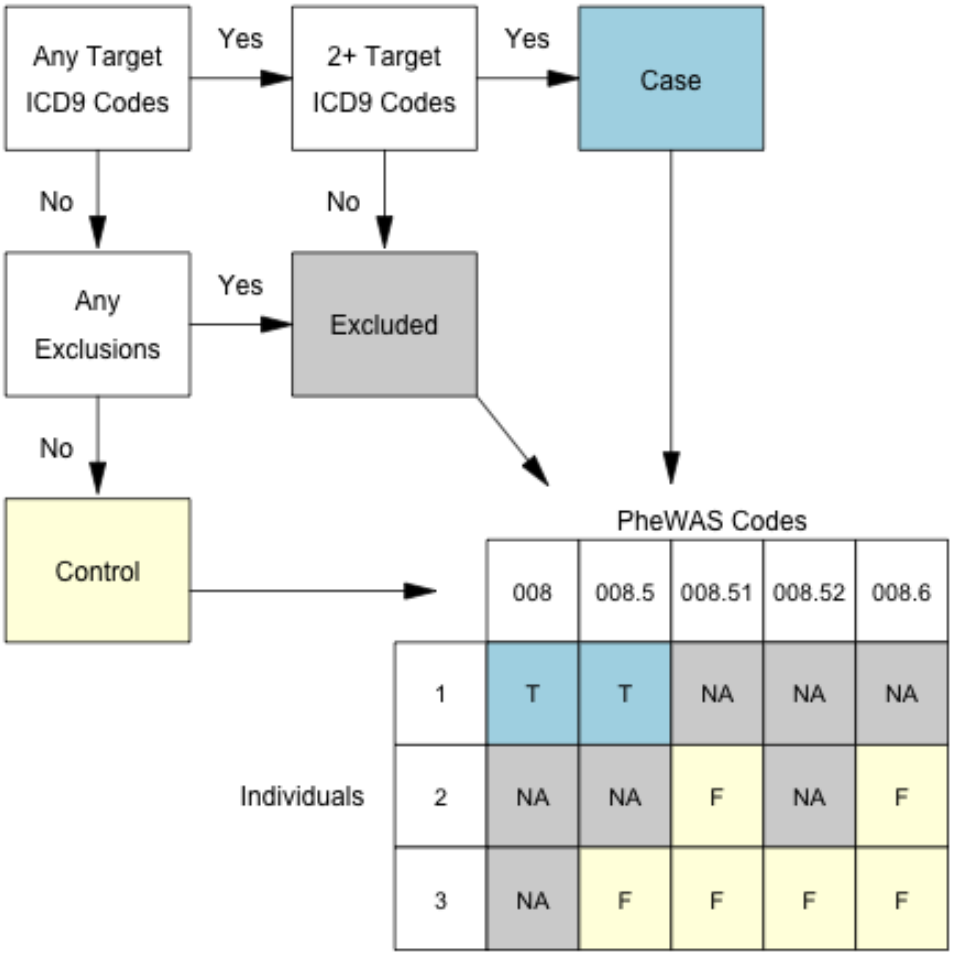


Figure 0-A. Decision flow for classifying an individual as a case, control, or exclusion for PheWAS. After classification, individuals are assembled into a PheWAS table where all cells in a row indicates whether that individual is a case, control, or exclusion for each PheWAS code, and all cells of a column indicate the classification of each individual for that code.

In our analyses individuals were required to have two different ICD-9 codes that mapped to a Phecode or one ICD-9 code on multiple days (therefore multiple times in their record) in order to be considered a case (Figure 0-A). Individuals who had a single instance of an ICD-9 code and no other ICD-9 codes that mapped to that Phecode were excluded. Individuals who had no ICD-9 codes in a Phecode and no exclusions for that Phecode became a control.

Phecodes can also be grouped into specific phenotype areas, such as "Neoplasms". This gives an added layer of possible aggregation. This aggregation may increase the difficulty of what is already the most difficult part of PheWAS, the interpretation of results. It could also potentially allow a more specific view. If only one group of phenotypes are of interest, one could analyze only that group and ignore the ICD-9 and Phecode record that groups outside the area of interest. It is important to note that Phecodes in the same group are often correlated, in some cases highly correlated. But Phecodes in separate code groups may be highly correlated or even indicative of the same underlying condition depending on the relationship of the ICD-9 contributing to each.

About this Dissertation

While the underlying theme of the work in this dissertation is PheWAS, it is important to note that each project described in this dissertation has a slightly different motivation that both allows for and requires different follow up to the PheWAS. PheWAS was chosen as an approach for a specific reason in each case; each chapter uses PheWAS with a goal in mind. We have undertaken PheWAS for many of the same reasons others have before us; to explore pleiotropy, identify the functional effects of deleterious variation, and uncover novel associations. Importantly, in all cases, PheWAS was used to inform our understanding of the biology

underlying genetics. The work presented herein also progresses through different types of genetic variation; starting with simple nuclear SNPs, moving to larger deletions, and concluding with mitochondrial variation. Each of these variants presents different challenges and expectations with PheWAS. While each project was performed as a stand-alone experiment, this dissertation aims to draw connections between different analyses to draw qualitative conclusions about the method as a whole. Lastly, in the assembled projects, PheWAS is rarely the end point. Something is always done after to understand the signal (or lack thereof) seen in the data, to try to return to and expand our underlying biological knowledge.

This chapter is adapted from a manuscript written with the assistance of C. Michael Stein, MBChB; Abiodun Adefurin, MBChB, Msc; Daniel Kurnik, MD; and David C Samuels, PhD.

I. SEX-SPECIFIC PLEIOTROPY OF THE THR164ILE VARIANT IN THE BETA-2-ADRENERGIC RECEPTOR

Introduction

The beta-2-adrenergic receptor (β_2AR) is one of three subtypes of beta-adrenoceptors, βARs[28]. These G-protein coupled receptors are expressed in a variety of tissues, including the heart and bronchial and vascular smooth muscle. The β_2AR is also the target of beta$_2$-agonists, which act through the receptor to mediate bronchodilation and are used clinically in patients with bronchoconstriction, e.g. asthma and chronic obstructive pulmonary disease[29]. There is a great variability in response to physiological and pharmacological stimulation or blockade of βARs, some of which is due to genetic variation among individuals in genes encoding βARs and their signal transduction proteins, including *ADRB2*, the gene encoding the β_2AR.

There are three polymorphisms in the coding region of *ADRB2* that have been shown to affect the functional properties of the receptor both *in vivo* and *in vitro*[30]. Two of these variants, Arg19Cys and Gln27Glu[31], form a haplotype and have been studied extensively, particularly in relation to asthma, heart failure, and hypertension, but it has been difficult to identify robustly associated clinical phenotypes[32]. The third variant, rs1800888, encodes the relatively uncommon *ADRB2* Thr164Ile variant that occurs at a frequency of only ~2% in individuals of European descent. This *ADRB2* variant is of particular interest because it is associated with profoundly reduced responses to agonist *in vitro* and has been associated with a five-fold reduction in

sensitivity to β_2AR agonist-mediated vasodilation *in vivo* in humans[33]. Thr164 is located in the upper part of transmembrane domain 4, and the nonsynonymous change to Ile has been predicted to cause a change in probability of transition of the receptor into the activated state[34,35]. Multiple experiments have shown that the presence of Ile at position 164 causes a decrease in receptor functionality[36,37].

Despite the strong *in vitro* and *in vivo* evidence of functional effects, few studies have addressed the functional consequences of the Thr164Ile transition in humans, likely because the variant is relatively infrequent. While some studies have reported no association between Thr164Ile and hypertension phenotypes[38,39], these studies had limited power. The two largest studies to date found associations of the Ile allele to increased blood pressure, but the effect was limited to females[40,41]. The Ile variant at position 164 has also been associated with a lesser response to beta2-agonist therapy in patients with asthma[42] and has been identified as a risk allele for chronic obstructive pulmonary disease, though other studies failed to detect these risks[43–45]. Moreover, Ile164 has been associated with adverse outcomes in patients with congestive heart failure[46]. However, the potential pleiotropy of the variant has not been systematically explored.

As the *ADRB2* gene has been associated with multiple phenotypes individually, and *in vivo* studies have shown that the Thr164Ile substitution causes a significant attenuation of receptor function, we hypothesized that individuals with this variant would manifest with other previously unrecognized clinical phenotypes in a systematic search using a PheWAS approach. We therefore conducted a PheWAS to investigate the association of Thr164Ile with many potentially interrelated phenotypes in an electronic health record (EHR)-based cohort.

Methods

Study Population

The study population consisted of adult individuals of European descent as identified by a third party. These individuals had both ICD.9 code data and genotyping on the Illumina Human Exome Bead Chip available in BioVU, Vanderbilt's DNA biobank[15]. The samples with genotyping in BioVU are linked to de-identified EHRs.

Genotypes

Genotypes for SNPs in *ADRB2* genotyped on the Illumina Human Exome Bead Chip were obtained[47] and the Thr164Ile allele was extracted. Population frequency was checked against 1000 Genomes. All quality control checks of genetic data were done in Plink[48].

PheWAS

Using all ICD.9 codes listed for adult patients (patient aged ≥18 years at time of code) with genotyping results in our dataset, we performed a phenome-wide association scan. PheWAS uses a predefined hierarchy to aggregate ICD.9 codes into PheWAS codes (Phecodes) which are then tested for association with a variant of interest[20]. Individuals with 2 or more ICD.9 codes that aggregate into a Phecode become a case for that Phecode. Individuals who have only one incidence of an ICD.9 code in the Phecode do not meet the definition of either a case or a control and are therefore excluded from analysis, as are individuals who could be a control for a code but are a case for a related Phecode. All other individuals become a control for that Phecode. Due to the exclusions, different Phecodes may have different numbers of total individuals tested.

We used logistic regression adjusting for age at last ICD.9 code and sex with an additive genetic model to test association of the *ADRB2* genotype with Phecodes. The study was restricted to individuals of European descent and to Phecodes with 50 or more cases. Limiting our analysis to Phecodes with 50 or more cases helped reduce our multiple testing threshold, while increasing the chance we would have power to see an association. Hardy-Weinberg equilibrium was checked in each regression. As a secondary analysis, we stratified by sex and ran the PheWAS in male and female cohorts separately. All sex-stratified PheWAS analyses only examined Phecodes with at least 25 cases. Following PheWAS, chi-square and Fisher's exact tests were used to check the allele distribution in cases and controls. The Bonferroni correction for our initial PheWAS was 4.08e-05 as 1224 Phecodes were populated with at least 50 cases. In the sex-stratified analyses, the level of statistical significance by Bonferroni correction was 4.06e-05 for females and 4.5e-05 for males. As Phecodes are often correlated, making a Bonferroni correction overly stringent, we also applied a 10% False discovery rate (FDR) correction to the data. Any biologically relevant signal that met the FDR level of significance was explored more deeply using laboratory data and existing literature.

Blood Pressure Analysis

As we saw an association between Thr164Il4 and Iatrogenic Hypotension in the PheWAS analysis we further explored the effect of the Thr164Ile variant on blood pressure. We studied blood pressure measurements first in all individuals, and second, in all individuals excluding those with the PheWAS code for Iatrogenic Hypotension. Date and time-stamped blood pressure measurements were downloaded and quality-controlled by removing any measurements that were not numeric, were negative or duplicate, or had incorrectly formatted dates. Any systolic

blood pressure measures greater than 300 mmHg and diastolic blood pressure measures greater than 240 mmHg or less than 5 mmHg were removed. Two separate blood pressure analyses were performed.

The first analysis was designed to test the previous association with hypertension and to see if our association with Iatrogenic Hypotension extended to general hypotension. Individuals with only one blood pressure measurement were removed. Individuals were categorized dichotomously as hypertensive if they had at least two measures greater than 140/90 mmHg in their record, and hypotensive if at least two measure less than 90/60 mmHg were present in their record. The same individual could therefore be categorized as both hypertensive and hypotensive.

We also performed an analysis to test the influence of the Thr164Il variant on non-dichotomized summary blood pressure measures. Median, 10^{th} percentile, and 90^{th} percentile summary measures were calculated for systolic blood pressure, diastolic blood pressure, and mean arterial pressure (MAP) using individuals who had more than 10 measures. Mean arterial pressure is a measure that combines both systolic and diastolic blood pressure measures (Equation 1). Linear regression for individuals with 10 or more blood pressure measures was performed.

Equation 1: $MAP = (\frac{1}{3})(SBP - DBP) + DBP$

Liver Enzyme Analysis

Since one of the Phecodes that passed the FDR in our analysis was Serum Enzyme abnormalities, we downloaded laboratory data for the levels of Aspartate transaminase (AST) and Alanine transaminase (ALT). These measures help quantity liver damage and can also

differentiate between possible causes. Non-numeric values, duplicate values, and values with an incorrect date and time stamp were removed. We required individuals to have an AST and ALT at the same time, so any measures where only one was present were removed. Median, minimum, and maximum values of these enzymes were calculated for each individual. The AST/ALT ratio was derived, and the median, minimum, and maximum values were obtained for each individual.

As obesity can impact these measures and has been previously associated with the The164Ile allele[49], we also obtained BMI measures and calculated a median BMI for each individual.

Statistical Analyses

All PheWAS analyses were performed using the PheWAS package for R[50]. Dichotomous blood pressure measures were analyzed using logistic regression analyses adjusting for median age of all blood pressure measures and sex. Summary blood pressure measures were analyzed using linear regression adjusting for median age across blood pressure measures and sex. Sex-stratified analyses for both types of blood pressure measure were performed adjusting for median age. Liver enzyme tests were analyzed using linear regression adjusting for median age over measures, sex, and BMI. Sex stratified analyses were also performed. All statistical analyses were performed using R 3.1.3[51].

Results

The population for the study of the *ADRB2* Thr164Ile variant included 23,854 individuals of European descent; the median age at last ICD.9 code was 64 years, and the median number of unique ICD.9 codes per person was 52 (Table 1-A).

Table 1-A. Demographic characteristics of the study population. P-values indicate statistical significance of comparisons between males and females.

	All	**Males**	**Females**	**p-value[1]**
n (%)	23,854	10,971 (46%)	12,883 (54%)	
Median age first ICD.9 code (IQR)	54.8 (41.7, 66.4)	57.0 (44.8, 67.0)	52.6 (39.1, 64.7)	<0.001
Median age at last code ICD.9 (IQR)	64.0 (51.9, 75.9)	65.3 (54.3, 76.1)	62.7 (49.8, 75.6)	<0.001
Median number of PheWAS codes (IQR)	23 (11, 44)	23 (10, 43)	24 (11, 44)	0.06
Median number of ICD.9 codes (IQR)	52 (27, 90)	51 (25, 90)	53 (28, 90)	0.002
Thr164Ile Minor Allele Frequency	0.012	0.013	0.011	0.22

[1]P-value shows statistical significance of Wilcoxon rank sum test or Fisher's exact test comparing male and females

Our initial PheWAS in the whole cohort showed a statistically significant risk effect for Thr164Ile with Iatrogenic Hypotension and Serum Enzyme Abnormalities (Figure 1-A, Appendix A). The T allele of Thr164Ile was significantly associated with the risk of Iatrogenic Hypotension, OR= 4.98 [95% confidence interval (CI) 2.37-10.47]; p=2.25e-05. This Phecode had a limited number of case patients (n=56), and 6 of these cases carried the risk allele (Table 1-B). By contrast, the Phecode for Serum Enzyme Abnormalities had 336 cases, with an odds ratio of 2.50 [95% CI, 1.59 - 3.91]; p= 6.5e-05.

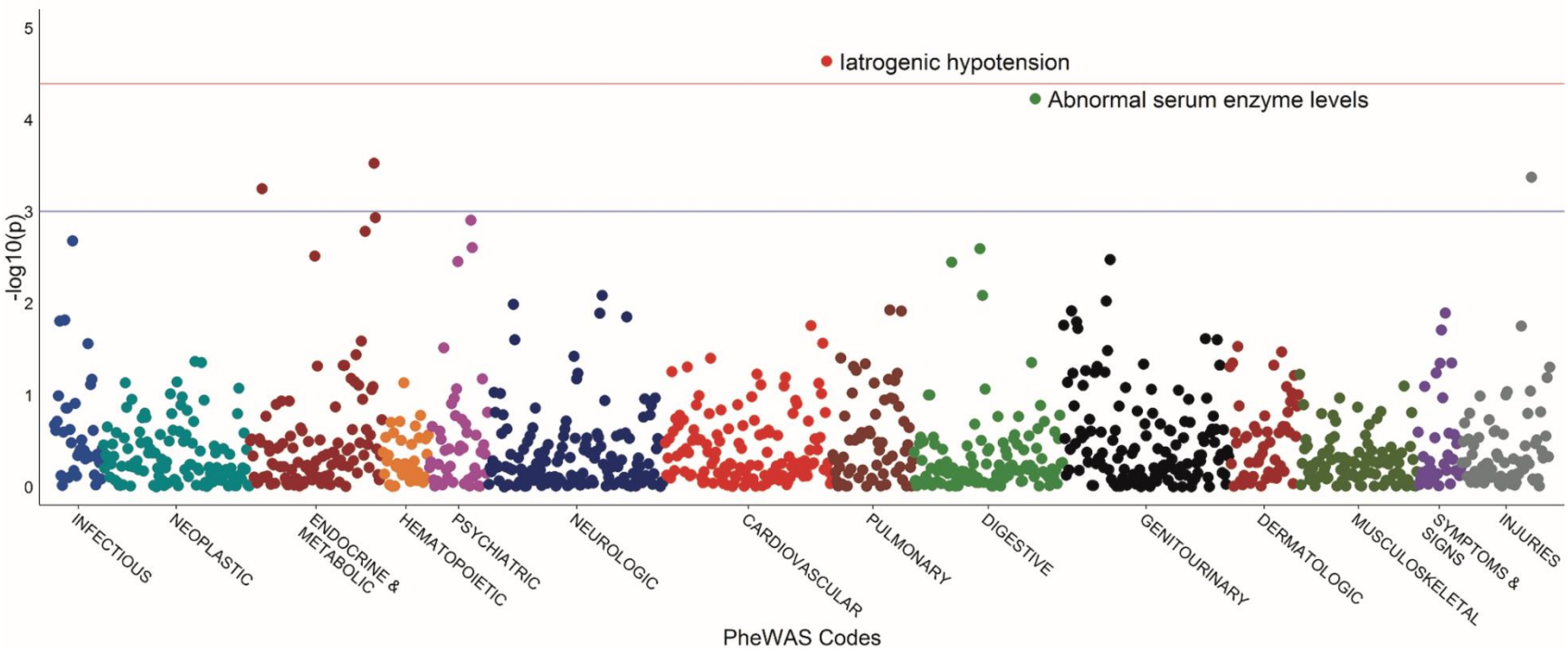


Figure 1-A. PheWAS Manhattan plot for the PheWAS of Thr164Ile in a mixed sex population of European descent. The red line is the Bonferroni correction for the number of tests, and the blue line is our level of suggestive significance (p=0.001). All hits that passed the FDR correction were annotated with their PheWAS description

Table 1-B. Genotype distribution across cases and controls for statistically significant PheWAS codes in males and females combined.

PheWAS Category and Case Status	Thr164Ile Minor Alleles		
	0	**1**	**2**
Iatrogenic Hypotension Cases	50	5	1
Iatrogenic Hypotension Controls	18664	451	9
Abnormal Serum Enzyme Level Cases	318	16	2
Abnormal Serum Enzyme Level Controls	18088	432	6

After sex stratification, no Phecodes were significant in males alone. Two Phecodes, Iatrogenic Hypotension and Hypothyroidism, were significant at the level of the Bonferroni correction (p-value of 4.06e-05), and an additional three were significant by FDR in females (Figure 1-B, Appendix B). Iatrogenic Hypotension remained significant (OR= 7.50 [95%CI, 3.25-17.30]; p=2.26e-06), but the case number was low in females only as was the number with a minor allele (Table 1-C). Acquired Hypothyroidism was the second most significant hit (OR= 4.64 [95% CI, 2.35-9.15]; p=9.57e-06). Abnormal Serum Enzyme Levels also remained significant in females alone (OR=3.14 [95% CI, 1.79-5.50]; p=6.57e-05), followed by Drug-Resistant Infection (OR= 2.91 [95% CI, 1.70-4.97]; p=9.10e-05). The Phecode for Ingrowing Nail was also significant by FDR, but we did not explore this code further. While Acquired Hypothyroidism was not significant at a Bonferroni corrected level or by FDR in the whole set it was beyond the level of suggested significance of $p< 0.001$ (Figure 1-C). The signal seen in women with Drug-Resistant Infection was present at a much attenuated level of significance in the whole cohort. None of our other expected signals were significant in all individuals or women alone (Appendix C).

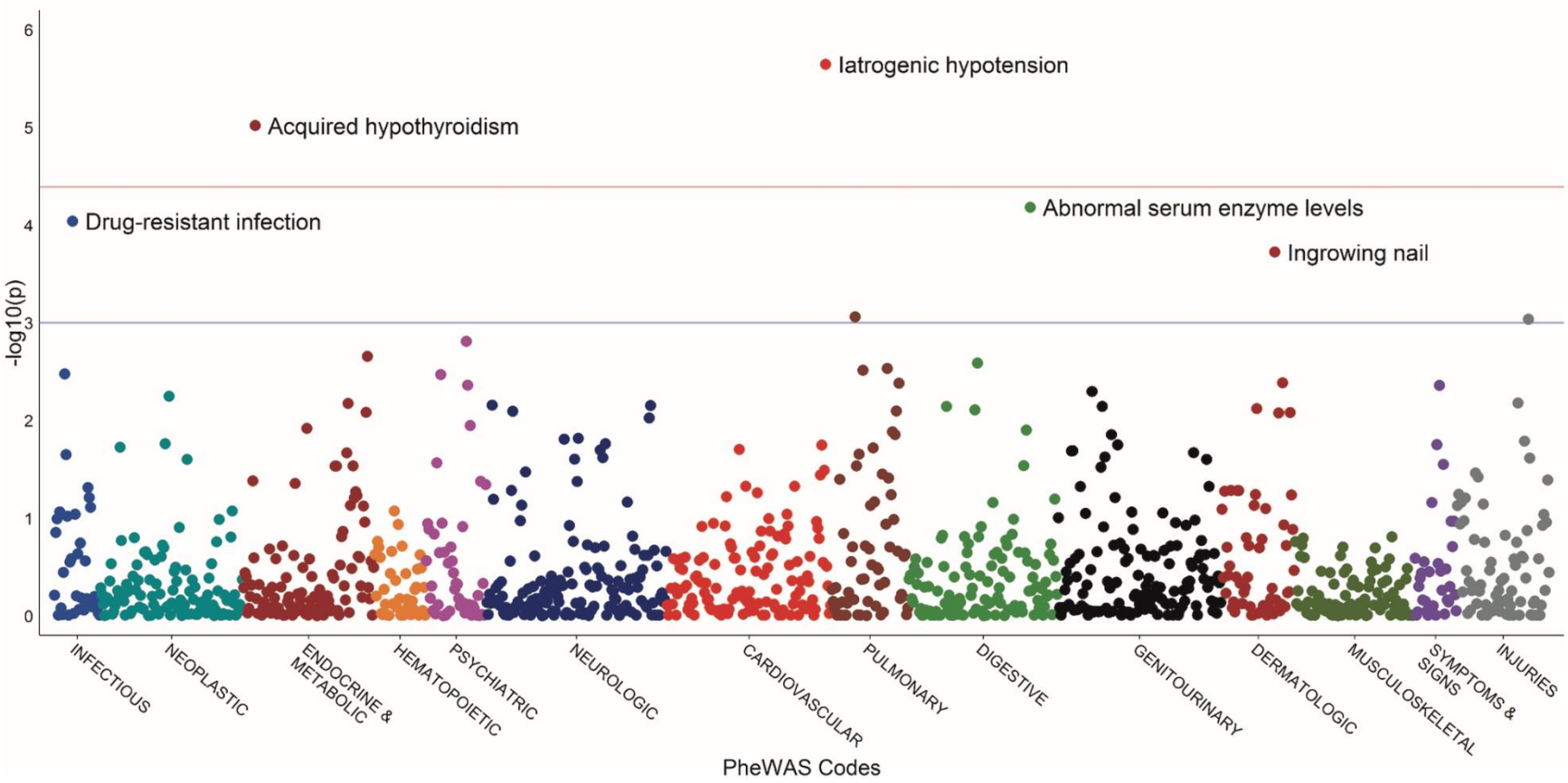


Figure 1-B. PheWAS Manhattan for the PheWAS of Thr164Ile in a female-only population of European descent. The red line is the Bonferroni correction for the number of tests, and the blue line is our level of suggestive significance (p=0.001). All hits that passed the FDR correction were annotated with their PheWAS description.

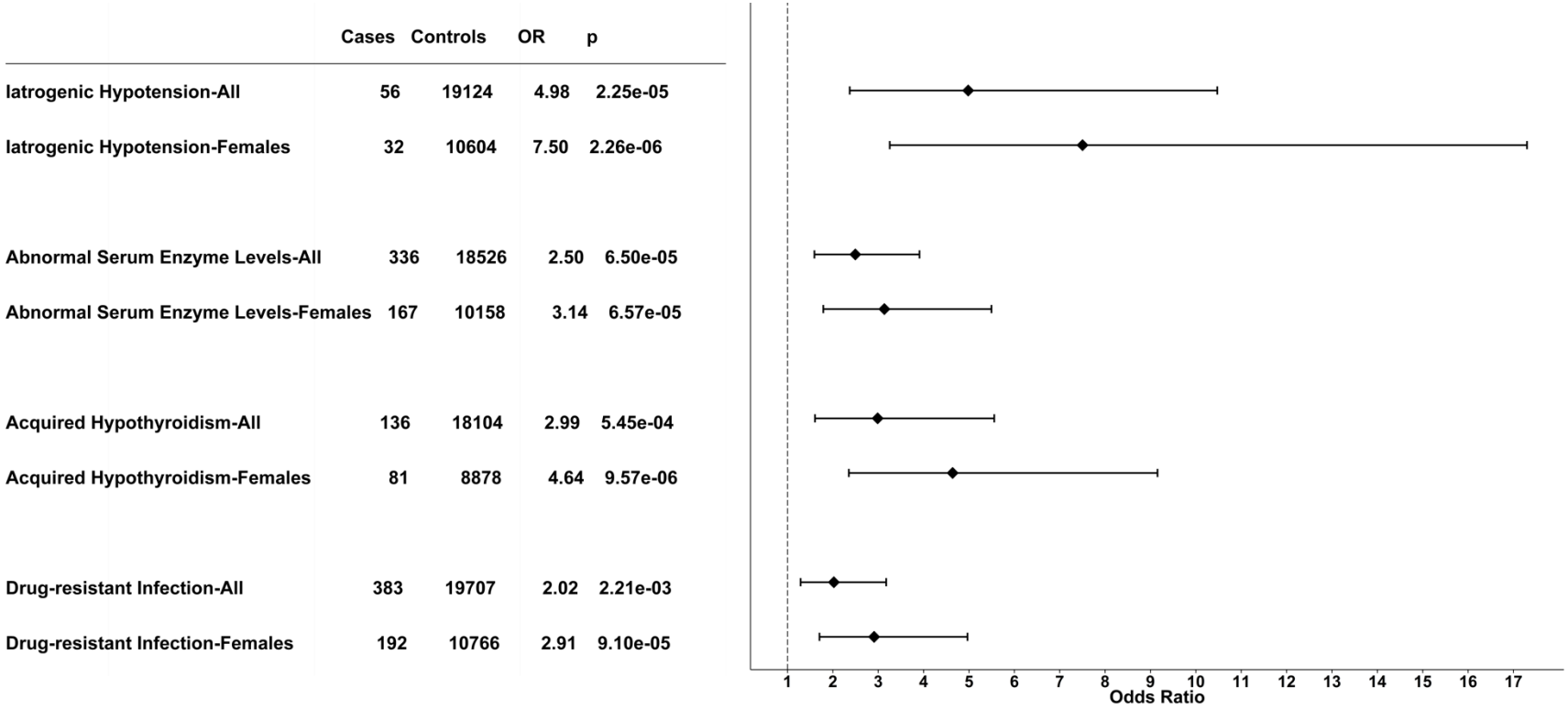


Figure 1-C. Forest plot of PheWAS hits for Thr164Ile is visible in all individuals and in females only.

Table 1-C. Genotype distribution across cases and controls for statistically significant PheWAS codes in females only.

PheWAS Category and Case Status	Thr164Ile Minor Alleles		
	0	**1**	**2**
Iatrogenic Hypotension Cases	27	4	1
Iatrogenic Hypotension Controls	10337	263	4
Acquired Hypothyroidism Cases	73	7	1
Acquired Hypothyroidism Controls	8662	214	2
Drug Resistant Infection Cases	179	12	1
Drug Resistant Infection Controls	10504	258	4
Abnormal Serum Enzyme Level Cases	156	9	2
Abnormal Serum Enzyme Level Controls	9909	246	3

22,314 of our study subjects had at least two blood pressure measures, including 12024 females and 10290 males. The median [IQR] number of blood pressure readings in study subjects was 154.0[47.0, 673.8]. Analysis of blood pressure measurements using all individuals found that Thr164Ile was associated with hypotension (at least two measure less than 90/60 mmHg) in males and females combined (OR=1.21 [95% CI, 1.01-1.44]) after adjustment for sex and median age, and in females alone after adjustment for median age (OR=1.29 [95% CI,1.02-1.64]). Sex and age were also significant predictors in several of the regressions; male sex and increasing age were associated with both hypertension and hypotension (Table 1-D). After the 55 Iatrogenic Hypotension cases in our blood pressure analysis were removed, the association between Thr164Ile and hypotension in all individuals, and in females, decreased and was no longer significant at the 0.05 level, and the odds ratio weakened slightly compared to when the Iatrogenic Hypotension cases were included (Table 1-E).

Table 1-D. Association of Thr164Ile with hypotension and hypertension in all individuals with a minimum of two blood pressure measures. Odds ratio (95% CI) and p-values are shown for the importance of each predictor in the regression. Individuals were classified as hypotensive if they had two measures less than 90/60 and hypertensive if they had two measures over 140/90. Median age was calculated using the ages at all blood pressure measures in the record.

	All (n=22,314)		**Males (n=10,290)**		**Females (n=12,024)**	
	Hypotensive OR (95% CI)	**Hypertensive OR (95% CI)**	**Hypotensive OR (95% CI)**	**Hypertensive OR (95% CI)**	**Hypotensive OR (95% CI)**	**Hypertensive OR (95% CI)**
Thr164Ile Allele	1.21 (1.01, 1.44) p=0.039	1.13 (0.96, 1.34) p=0.154	1.11 (0.85, 1.45) p=0.449	1.27 (0.98, 1.65) p=0.0764	1.29 (1.02, 1.64) p=0.034	1.04 (0.83, 1.30) p=0.761
Median Age	1.011 (1.009, 1.013) p<0.001	1.012 (1.011, 1.014) p<0.001	1.015 (1.013, 1.018) p<0.001	1.003 (1.000, 1.005) p<0.001	1.007 (1.005, 1.009) p<0.001	1.02 (1.017, 1.021) p<0.001
Sex==M	1.10 (1.01, 1.17) p<0.001	1.36 (1.29, 1.44) p=0.001	-	-	-	-

Table 1-E. Association of Thr164Ile with hypotension and hypertension in individuals without the Iatrogenic Hypotension Phecode and a minimum of two blood pressure measures. Odds ratio (95% CI) and p-values are shown for the importance of each predictor in the regression. Individuals were classified as hypotensive if they had two measures less than 90/60 and hypertensive if they had two measures over 140/90. Median age was calculated using the ages at all blood pressure measures in the record.

	All (n=22,259)		**Males (n=10,266)**		**Females (n=11,993)**	
	Hypotensive OR (95% CI)	**Hypertensive OR (95% CI)**	**Hypotensive OR (95% CI)**	**Hypertensive OR (95% CI)**	**Hypotensive OR (95% CI)**	**Hypertensive OR (95% CI)**
Thr164Ile Allele	1.17 (0.98, 1.40) p=0.091	1.12 (0.94, 1.33) p=0.203	1.10 (0.84, 1.44) p=0.475	1.26 (0.97, 1.64) p=0.083	1.23 (0.96, 1.57) p=0.096	1.02 (0.81, 1.28) p=0.897
Median Age	1.011 (1.009, 1.013) p<0.001	1.012 (1.011, 1.014) p<0.001	1.016 (1.013, 1.018) p<0.001	1.002 (1.000, 1.005) p<0.031	1.007 (1.004, 1.09) p<0.001	1.019 (1.017, 1.021) p<0.001
Sex==M	1.10 (1.04, 1.17) p<0.001	1.36 (1.29, 1.44) p<0.001	-	-	-	-

In our analysis of summary blood pressure measures, the Thr164Ile variant was not significantly associated with median systolic blood pressure, median diastolic blood pressure, or median MAP in both males and females combined. In the female only analysis, none of the measure were significantly associated with the variant, and while Thr164Ile was marginally ($p<0.1$) associated with increased median systolic blood pressure in males, none of the tests were significant at the $p=0.05$ threshold (Table 1-F). The same held true for our tests of the variant as a predictor of 10^{th} percentile BP (Table 1-G). When looking at the relationship between the 90^{th} percentile BP measures, we saw that the Ile allele was associated with increased systolic measures in males only ($p=0.04$), and was nearly statistically significant but did not reach $p=0.05$ for MAP (Table 1-H). The variant was not significant for any measures in either all individuals or females alone.

In order to determine if the pattern of PheWAS signals represented several discrete phenotypes or sub-phenotypes of a single condition, we identified how many females were cases for multiple of our top Phecodes. In females, individuals (including those carrying the Thr164Ile variant) who were cases for one of the top Phecode hits were unlikely to be cases for any of the other Phecode hits that passed significance thresholds (Figure 1-D-a), indicating that these are distinct and pleiotropic genotype-phenotype associations.

Table 1-F. Regression coefficients from linear regression for a mixed population, males, and females with median systolic blood pressure, diastolic blood pressure, and mean arterial pressure.

	All (n=21,339)			**Males (n=9,854)**			**Females (n=11,485)**		
	Median Systolic BP Beta (95%CI)	**Median Diastolic BP** Beta (95%CI)	**Median MAP** Beta (95%CI)	**Median Systolic BP** Beta (95%CI)	**Median Diastolic BP** Beta (95%CI)	**Median MAP** Beta (95%CI)	**Median Systolic BP** Beta (95%CI)	**Median Diastolic BP** Beta (95%CI)	**Median MAP** Beta (95%CI)
Thr164Ile Allele	0.13 (-0.95, 1.21) p=0.82	-0.12 (-0.85, 0.61) p=0.75	0.03 (-0.72, 0.77) p=0.95	1.38 (-0.25, 3.01) p=0.0961	0.61 (-0.49, 1.71) p=0.28	0.91 (-0.22, 2.04) p=0.12	-1.00 (-2.42, 0.42) p=0.17	-0.66 (-1.69, 0.23) p=0.14	-0.74 (-1.74, 0.25) p=0.14
Median Age	0.20 (0.19, 0.21) p<0.001	-0.14 (-0.14, -0.13) p<0.001	-0.01 (-0.03, -0.02) p<0.001	0.07 (0.07, 0.10) p<0.001	-0.16 (-0.17, -0.15) p<0.001	-0.08 (-0.09, -0.06) p<0.001	0.28 (0.27, 0.39) p<0.001	-0.12 (-0.13, -0.11) p<0.001	0.01 (0.01, 0.02) p=0.002
Sex==M	1.03 (0.68, 1.37) p<0.001	2.21 (1.97, 2.44) p<0.001	1.81 (1.57, 2.05) p<0.001	-	-	-	-	-	-

Table 1-G. Regression coefficients from linear regression for a mixed population, males, and females with 10th percentile systolic blood pressure, diastolic blood pressure, and mean arterial pressure.

	All (n=21,339)			**Males (n=9,854)**			**Females (n=11,485)**		
	10% Systolic BP Beta (95%CI)	**10% Diastolic BP** Beta (95%CI)	**10% MAP** Beta (95%CI)	**10% Systolic BP** Beta (95%CI)	**10% Diastolic BP** Beta (95%CI)	**10% MAP** Beta (95%CI)	**10% Systolic BP** Beta (95%CI)	**10% Diastolic BP** Beta (95%CI)	**10% MAP** Beta (95%CI)
Thr164Ile Allele	-0.13 (-1.17, 0.91) p=0.81	-0.03 (-1.07, 0.46) p=0.44	-0.12 (-0.87, 0.63) p=0.76	0.38 (-1.21, 1.97) p=0.64	0.24 (-0.92, 1.41) p=0.68	0.48 (-0.67, 1.62) p=0.41	-0.64 (-1.99, 0.72) p=0.36	-0.76 (-1.77, 0.26) p=0.14	-0.64 (-1.62, 0.35) p=0.21
Median Age	0.08 (0.07, 0.09) p<0.001	-0.14 (-0.15, -0.14) p<0.001	-0.05 (-0.06, -0.05) p<0.001	-0.01 (-0.03, 1.90) p=0.09	-0.16 (-0.17, -0.14) p<0.001	-0.10 (-0.10, -0.08) p<0.001	0.14 (0.13, 0.16) p<0.001	-0.14 (-0.15, -0.13) p<0.001	-0.02 (-0.03, -0.02) p<0.001
Sex==M	0.80 (0.46, 1.13) p<0.001	2.08 (1.84, 2.33) p<0.001	1.67 (1.43, 1.91) p<0.001	-	-	-	-	-	-

Table 1-H. Regression coefficients from linear regression for a mixed population, males, and females with 90th percentile systolic blood pressure, diastolic blood pressure, and mean arterial pressure.

	All (n=21,339)			**Males (n=9,854)**			**Females (n=11,485)**		
	90% Systolic BP Beta (95%CI)	**90% Diastolic BP** Beta (95%CI)	**90% MAP** Beta (95%CI)	**90% Systolic BP** Beta (95%CI)	**90% Diastolic BP** Beta (95%CI)	**90% MAP** Beta (95%CI)	**90% Systolic BP** Beta (95%CI)	**90% Diastolic BP** Beta (95%CI)	**90% MAP** Beta (95%CI)
Thr164Ile Allele	0.39 (-0.89, 1.67) p=0.55	0.07 (-0.66, 0.81) p=0.85	0.14 (-0.66, 0.94) p=0.74	2.00 (0.10, 3.89) p=0.04	0.80 (-0.32, 1.92) p=0.16	1.14 (-0.07, 2.34) p=0.07	-1.05 (-2.76, 0.67) p=0.23	-0.55 (-1.52, 0.41) p=0.26	-0.73 (-1.80, 0.33) p=0.18
Median Age	0.33 (0.32, 0.34) p<0.001	-0.09 (-0.10, -0.09) p<0.001	0.03 (0.02, 0.04) p<0.001	1.95 (0.18, 0.21) p<0.001	-0.14 (-0.15, -0.13) p<0.001	-0.04 (-0.05, -0.03) p<0.001	0.43 (-0.41, 0.44) p<0.001	-0.06 (-0.07, -0.06) p<0.001	0.08 (0.07, 0.09) p<0.001
Sex==M	1.02 (0.06, 1.43) p<0.001	2.00 (1.76, 2.23) p<0.001	1.70 (1.45, 1.96) p<0.001	-	-	-	-	-	-

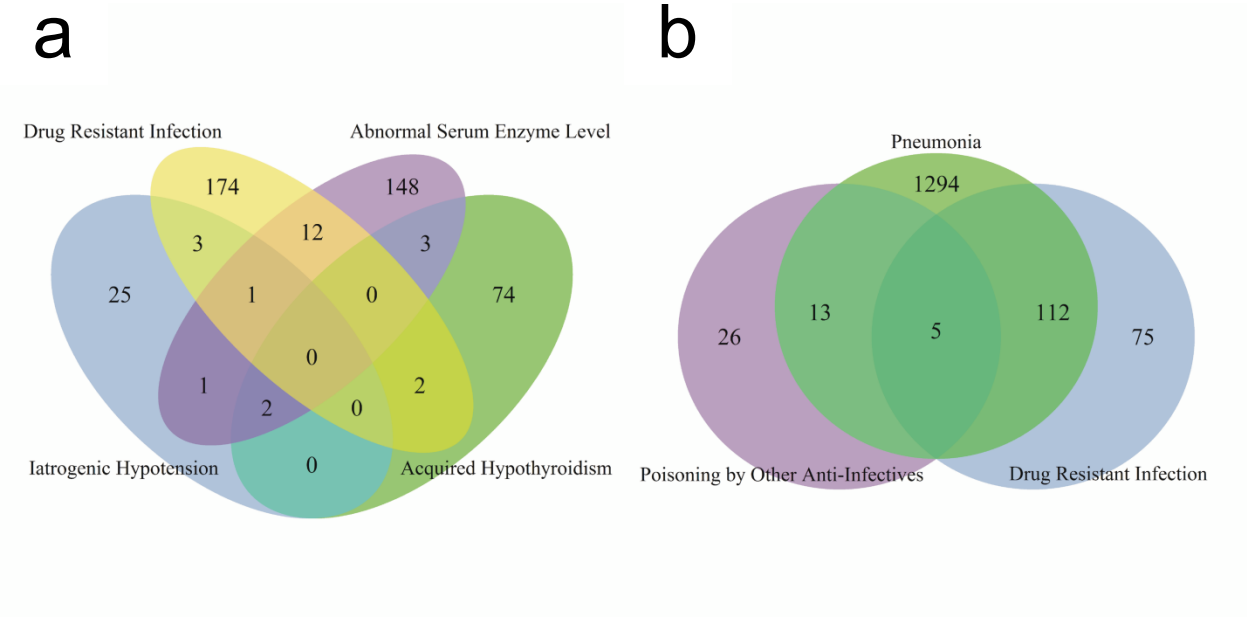


Figure 1-D. Venn diagram of the overlap of case individuals for PheWAS codes. a) Codes that were significant in either the analysis of all individuals or in females only were tested for overlap in females. b) The overlap of individuals coded for Drug-resistant Infection, Pneumonia, and Poisoning by Other Anti-Invectives in the female only population.

In PheWAS hits that passed the p-value threshold of 0.001 for suggestive significance in females, three codes (Drug-resistant Infection, Poisoning by Other Anti-infectives, and Pneumonia), were related to infection. Of females who had the Phecode for Drug-resistant Infection, 58% also had the Phecode for pneumonia (Figure 1-D-b). We also observed the Phecode for Poisoning by Other Anti-infectives, but the women with this code were largely distinct from those coded for Drug-resistant Infection. The Phecode for Drug-Resistant Infection is made up of 19 ICD-9 codes focused exclusively on drug resistance. The Phecode for Pneumonia comprised of 67 distinct codes (Appendix D). Most of these codes are immediately related to pneumonia, though some relate to congestion or difficulty breathing when an infection may be present. No ICD.9 codes are shared between the Phecode for Drug-resistant Infection and the Phecode for Pneumonia.

Our analysis of the liver function test laboratory values produced no significant associations after adjustment for BMI. The Thr164Ile allele was not significantly associated with median AST (Beta [95% CI] = -0.94[-2.48,0.60],p=0.23) , median ALT (-1.26[-2.96,0.44], p=0.15) or median AST/ALT ratio (0.02[-0.02, 0.06], p=0.31) after adjustment for BMI in all individuals. In females, the median AST/ALT ratio was not associated with the Thr164Ile allele (0.04[-0.01, 0.09], p=0.14), nor was median AST (-0.83 [-2.77, 1.11], p=0.40), but median ALT was associated (-2.09 [-4.11, -0.06], p=0.044) with each copy of the Ile allele resulting in decreasing ALT. BMI was the most significant predictor of the liver enzymes, and was also moderately associated with the Thr164Ile allele.

Discussion

The primary PheWAS association for the Thr164Ile variant in *ADRB2* was the code for Iatrogenic Hypotension. This association with hypotension was the opposite of what would be expected given the previous reports of an association with hypertension in other studies[40,41] and given the observation that the variant was associated with decreased agonist-induced vasodilation in translational studies[33].

Several reasons could explain our finding. Iatrogenic Hypotension describes a clinical scenario where a medical intervention inadvertently results in low blood pressure, i.e. an exaggerated hypotensive response to a medication or intervention. The population studied was a hospital-based cohort, and patients thus underwent medical interventions and received new medications; manual review of the records of the cases with a PheWAS diagnosis of Iatrogenic Hypotension revealed that 33 of 56 had undergone surgery shortly before the ICD.9 code was

entered into their record. The biology of the β_2AR suggests a mechanism whereby individuals with the Thr164Ile variant could be at increased risk for hypotension in such circumstances despite having impaired β_2AR-mediated vasodilation and thus a propensity towards hypertension. The β_2AR, in addition to mediating vasodilation, also increases heart rate and cardiac contraction, and thus cardiac output, a key determinant of blood pressure. Healthy Thr164Ile carriers have decreased chronotropic and inotropic responses to beta2-agonists[52], and this attenuated cardiac response was even more pronounced in patients with heart failure[53]. It is therefore possible that individuals with attenuated β_2AR-mediated signal transduction may become hypotensive in a postoperative setting where compensatory sympathetic stimulation and increased plasma catecholamine concentrations are critical to maintain blood pressure by increasing heart rate and contractility and thus cardiac output.

Consistent with the PheWAS code of Iatrogenic Hypotension, our analysis of all the blood pressure readings recorded also indicated that individuals with the Thr164Ile variant (in analyses that included and excluded those with the Iatrogenic Hypotension code) had increased odds of being hypotensive at some point in their medical record. Interestingly, this result was not visible when the median, 90^{th}, and 10^{th} percentile blood pressure measures were used. While other studies have shown that carriers of the variant are more likely to be hypertensive, that was not the case in this study. However, in males there was a trend in that direction, and in our summary analysis, the variant was associated with increased 90^{th} percentile measures in men. Many of the blood pressure measures in our dataset were not resting blood pressure measures, but rather measures obtained during a hospital stay or other intervention. Thus, the effects of stressors such as medications and illness may be magnified under these circumstances.

We observed the significant genotype-hypotension association in all individuals and in females only, consistent with other large studies of the Thr164Ile variant that found an effect on hypertension in females only[40,41]. As our study was conducted in an EHR dataset, it is possible that there was a difference in coding between males and females for certain traits, but as the case distribution was similar in both sexes for our primary hits, it is more plausible that there is a fundamental biological difference between the sexes in the effect of the variant. This is in keeping with the observation that there are differences among sexes in cardiac physiology, and women are more dependent than men on an increase in heart rate to maintain cardiac output under conditions of stress[54]. Therefore, an inability to increase heart rate to maintain cardiac output in Thr164Ile individuals under conditions of stress would be more likely to manifest as hypotension in women.

The lack of overlap of phenotypes in the same individuals suggests that Thr164Ile is associated with multiple distinct phenotypes. Alternatively, the different phenotypes could be related and coexist in the same patient, with only one being prominent enough to be listed as ICD diagnosis by the physician.

The code for Serum Enzyme Abnormalities encompassed only ICD.9 code 790.5 for "Other Nonspecific Abnormal Serum Enzyme Levels", representing predominantly an increase in serum liver enzymes. The increase in liver enzymes associated with Thr164Ile may be related to altered effects of catecholamines on liver cells[55] or effects on fat metabolism. Given the change in effect we saw after adjusting for BMI, it is also possible that the effect is related to BMI. The β2AR mediates lipolysis, and adipocytes from humans heterozygous for the Ile164 variant had a markedly decreased sensitivity to lipolysis induced by a beta-agonist[56]. Moreover, Ile164 was associated with obesity in population studies[49]. Obesity is the main risk factor for

non-alcoholic fatty liver disease, a common cause of elevated liver enzyme levels. Indeed, *ADRB2* variants have previously been associated with liver enzyme levels, and a link to non-alcoholic fatty liver disease was postulated[57].

The association of Thr164Ile with Acquired Hypothyroidism in females was unexpected but not biologically implausible. Thyroid hormone contributes to the regulation of βAR expression and has been associated with hypertension and other cardiovascular phenotypes[58–61]. Deficiency of thyroid hormone reduces responsiveness to catecholamines due to reduced βAR expression[62]; thus, it is possible that the presence of an *ADRB2* variant associated with reduced function could magnify the effects of hypothyroidism, facilitating its diagnosis.

The women who had the PheWAS code for Drug-resistant Infection overlapped substantially (58%) with those who were coded for Pneumonia. The Thr164Ile variant has been associated with several pulmonary phenotypes, including impaired bronchoconstriction, reduced lung function, and chronic obstructive pulmonary disease (COPD)[43], which are risk factors for recurrent pulmonary infections, reduced clearance of infections, and thus antibiotic resistance. Moreover, patients with underlying pulmonary diseases such as COPD could be more likely to be assigned a diagnosis of pneumonia. Alternatively, cross-talk between the immune and adrenergic systems through the β_2AR may be implicated[63].

The potential mechanisms underlying the sex-specificity of the PheWAS findings are of interest. One possibility is that coding practices among health care providers differ for men and women. However, more likely is that β_2AR responses are affected by sex, and there are many examples of such sex-related differences in other settings. For example, differences between men and women in the contribution of β_2AR-mediated responses have been noted for blood pressure regulation[64], cutaneous [65] and forearm blood flow[66,67], neutrophil function[68], and the anabolic

effects of an agonist[69], or among pre- and post-menopausal women. It is possible similar mechanism pertain in the setting of illness.

In conclusion, our study has shown that Thr164Ile has pleiotropic effects, and carriers are more likely to be assigned certain diagnosis codes by their physicians. Our PheWAS study, while testing for association with imperfectly phenotyped case and control groups, allowed us to identify genotype-Phecode associations that were plausible given the multiple physiological functions of the β_2AR. Replication in carefully phenotyped populations will be important for validating these associations.

II. THE APOBEC3G HIS186ARG VARIANT IMPACTS HUMORAL IMMUNITY IN CHILDREN

Introduction

The apolipoprotein B mRNA-editing catalytic polypeptide (*APOBEC*) family is a group of cytidine deaminases that were initially discovered due to the ability of *APOBEC3G* (A3G) to block the replication of Human Immunodeficiency Virus Type 1 (HIV-1)[70]. The *APOBEC3* sub-family is a group of 7 proteins (A3A-A3C, A3DE, A3F-H) encoded on chromosome 22. As cytidine deaminases, the members of this family mediate the change from C to T in single stranded DNA[71]. The A3s are thought to have evolved from *APOBEC1*, and the gene duplication events leading to the family are thought to have occurred during the expansion of the primate branch as a means of countering the exogenous and endogenous retroviruses[72]. A3G in particular has been subject to strong positive selection throughout primate evolution[73]. The *APOBEC3*s are known to function against exogenous retroviruses and human endogenous retroviruses, both through deamination dependent and deamination independent functions.

A3G and its variation has been thoroughly investigated for its effect in HIV. The A3G His186Arg variant is located in exon 4 of the APOBEC3G gene[74]. While this region of the protein in not directly involved in deamination, it is thought to be important for RNA binding which is involved in several deaminase independent functions and sliding along genetic material to promote deamination[75]. Some studies have shown that individuals of African descent with the A3G His186Arg variant progress from HIV to AIDs more quickly, but this effect of the variant does not seem present in individuals of European descent[74,76,77]. Biochemically, there is

increasing evidence that A3G proteins containing the His186Arg change exhibit less antiviral activity than wildtype A3G[78]. The effect of the His186Arg variant in populations infected with Hepatitis-B virus (HBV) has also been explored, but no effect was seen[79]. While the global allele frequency of A3G His186Arg is 14.6%, the variant is present at a frequency of around 3% in populations of European descent[80]. Despite the relatively common global occurrence of this variant, it has not been widely explored outside of the context of viral infection.

Given the limited knowledge of the A3H His186Arg variant and the conflicting reports of it importance in the context of HIV1 infection, we hypothesized that it was likely this variant manifested in one or more clinically relevant phenotypes in a hospital derived cohort. We specifically expected phenotypes of viral infection or other innate immune responses. To test this hypothesis, we performed a PheWAS in an EHR based dataset, exploring the impact of the A3G His186Arg variant on multiple clinically relevant phenotypes.

Methods

Study Population

The study population consisted of third-party identified white individuals with both ICD.9 code data and genotyping available on the Illumina Human Exome Bead Chip available in BioVU.

Genotypes

Genotypes for SNPs in A3G genotyped on the Illumina Human Exome Bead Chip were obtained. rs8177832, the SNP for His186Arg, was additively encoded and checked against published minor allele frequencies[80]. All quality control checks of genetic data were done in Plink [48].

PheWAS Aggregation and Regression

A record of all ICD-9 codes with date stamps were obtained for all individuals with genotyping. These were aggregated using a predefined hierarchy into PheWAS codes (Phecodes) as described in the Introduction. Briefly, individuals with 2 or more instances of any ICD-9 code that map to a Phecode become cases for that Phecode. Individuals who never have any instance of the ICD-9 codes within a Phecode become a control for that code. Individuals who have a single occurrence of one ICD-9 code within a Phecode become an "NA" and are excluded from analysis for that Phecode. Individuals can also be excluded from analysis if by aggregation they would be a control for a given Phecode, but are a case for a highly related Phecode. This exclusion only removes potential controls, never potential cases. ICD-9 codes that occur more than once on the same day in a patient's record are collapsed to one occurrence. Phecodes are then tested for association with a variant of interest.

Our initial PheWAS used logistic regression adjusting for age at last ICD-9 code in the record, and sex. We also performed a second PheWAS limiting our dataset to individuals who never reached the age of 20 in their ICD-9 based record. This logistic regression was also adjusted for age and sex. In PheWAS tests using all individuals, only Phecodes with at least 50 cases were tested. In age stratified PheWAS tests, Phecodes with at least 25 cases were tested.

ICD-9 Code Specific Analyses

Following PheWAS, we performed a series of logistic regressions using the ICD-9 codes within the Humoral Immunity Phecode as the outcome. The ICD-9 code records of all individuals were evaluated, and two instances of a specific ICD-9 code were considered a case for that ICD-9

code. Regressions performed separately in all individuals and individuals under 20, and were adjusted for first age at ICD-9 code and sex. Venn diagrams were made with Venny [81].

CPT Code Analysis

CPT codes for immune globulin infusions (82784) were obtained for white individuals with genotyping in our dataset. The number of codes each individual had were calculated. The minor allele frequency of His186Arg was calculated in individuals with at least one "82784" CPT code, and compared to the minor allele frequency in the whole population. A Wilcoxon rank sum test was used to see if the number of CPT codes on distinct days was different between the carriers and non-carriers of the His186Arg allele. The population overlap between individuals with the "82784" CPT code and those with the Deficiency of Humoral Immunity Phecode was also calculated.

Laboratory Value Analysis

Immunoglobulin G (IgG) laboratory values from the IgG Quantitative (IgG-Q) test were downloaded for individuals in the set. IgG values were quality controlled to remove non-numeric values and were then normalized using the age of the individual at the time of their measure and reference information from the Vanderbilt Pathology Laboratory Services (Appendix H).

Maximum, median, and minimum values were calculated using both raw and normalized IgG values all individuals with two or more measures. Wilcoxon rank sum tests were used to examine if these summary measures of IgG were different between homozygous reference

individuals and carriers of the His186Arg variant. Individuals were also stratified based on age at the time of measure, and adults and children were analyzed separately.

Results

27547 individuals were in our PheWAS dataset. 3,837 were younger than 20 at the last ICD-9 code in their record, and 23,710 had an age at last code greater than 20 (Table 2-A). 46.7% of the study individuals were male. The allele frequency of the His186Arg variant was not significantly different between individuals with an age at last code greater than 20 and an age at last code of less than 20, though it was trending towards significance.

Table 2-A. Demographics of the PheWAS analysis set combined and separated by age at last ICD-9 record.

	All	**Age at Last Code <20**	**Age at Last Code >=20**	**p-value**
N	27547	3837	23710	-
Male (%M)	12865 (46.7%)	1981 (51.6%)	10884 (45.9%)	0.004
Median Age First ICD-9 Code in Record (IQR)	51.4 (32.4, 64.7)	2.1 (0.2, 7.0)	55.2 (42.5, 66.6)	<2.2e-16
Median Age Last ICD-9 Code in Record (IQR)	60.7 (42.1, 73.6)	11.7 (6.9, 15.6)	64.1 (52.4, 75.8)	<2.2e-16
Deficiency of Humoral Immunity Cases (%)	108	27	81	0.002
His186Arg Allele Frequency	0.031	0.035	0.030	0.07

Our initial PheWAS tested 1215 Phecodes. A single Phecode, Deficiency of Humoral Immunity (Odds ratio (OR) = 2.93 [95% Confidence interval (CI) 1.83 -4.68]; p= 7.01E-06), passed the Bonferroni level of statistical significance (4.1e-05) (Figure 2-A). There were 108

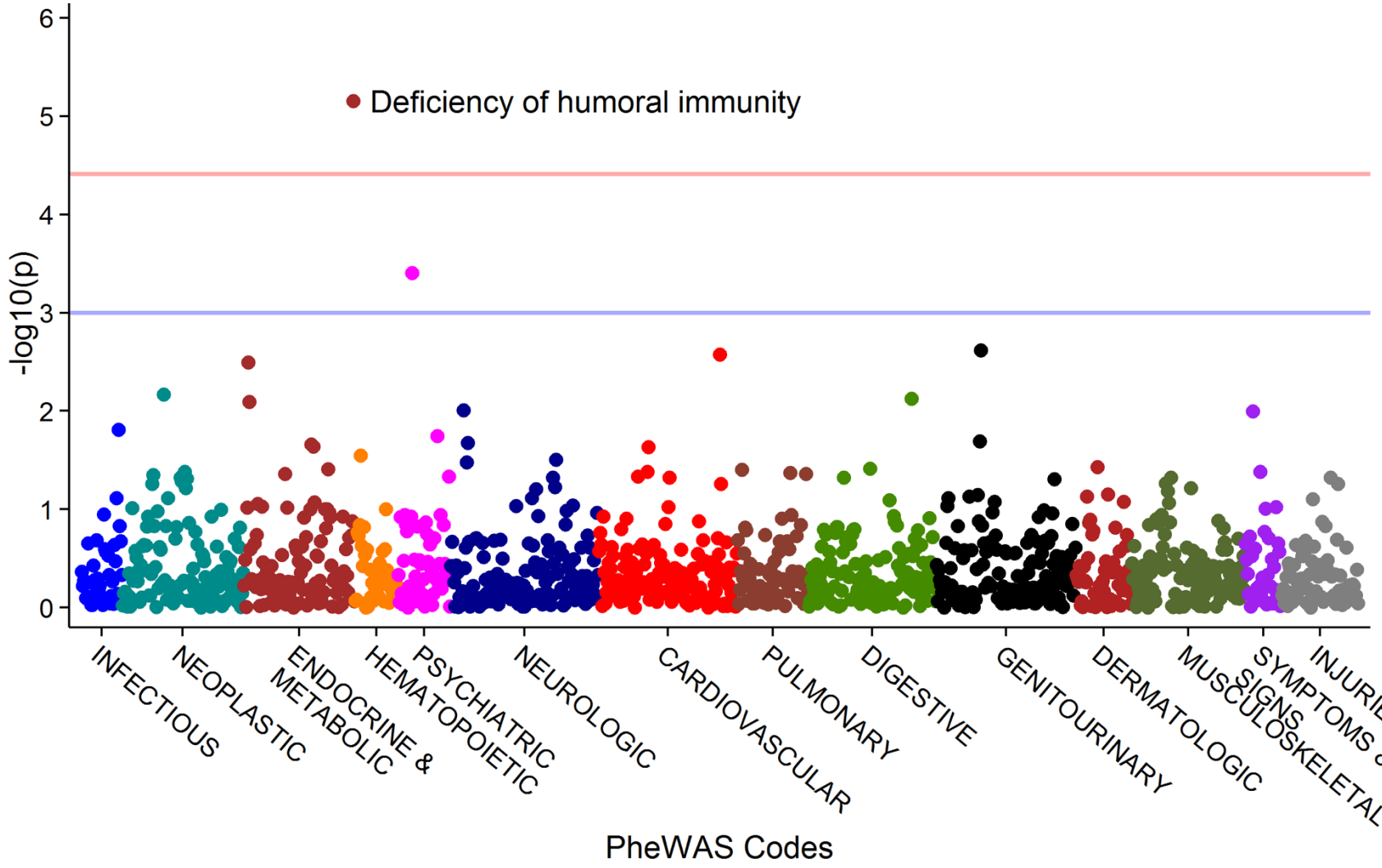


Figure 2-A. PheWAS manhattan plot showing the association of the Phecode for Deficiency of Humoral Immunity and the A3G His186Arg variant. Red line indicates Bonferroni correction and blue line indicates level of suggestive significance.

cases for Deficiency of Humoral Immunity, including 19 with the A3G 186R allele (Table 2-B). Only one other Phecode, Hallucinations, passed the level of suggestive significance (p=0.001) (Appendix E).

Table 2-B. Genotype distribution in Deficiency of Humoral Immunity cases and controls for individuals of all ages.

PheWAS Category and Case Status	**His186Arg Minor Alleles**		
	0	**1**	**2**
Deficiency of Humoral Immunity Cases	89	19	0
Deficiency of Humoral Immunity Controls	23668	1463	32

Of the 9 ICD-9 codes (Appendix G) that comprised the Deficiency of Humoral Immunity PheWAS code only 2 of which were present twice in at least 5 individuals in our dataset. The ICD-9 code 279.00 contributes that largest signal, but that there is a second signal from 279.06 (Table 2-C). These signals contribute similar ORs but have substantially different p-values and case sizes. The signal from the union of these two groups is weaker than the signal from just 279.00 alone, indicating that 279.00 is the primary driver of our association (Figure 1-A).

Table 2-C. Association of His186Arg variant with individuals who have two or more incidences of the 279.00, 279.06, and 279.00 or 279.06 ICD-9 codes combined.

ICD-9 Code	**OR (95% CI)**	**p-value**	**Case Individuals**	**Control Individuals**
279.06	3.37(1.47, 7.77)	0.004274	30	27353
279.00	3.50 (2.12, 5.77)	8.66E-07	81	27353
279.00, 279.06, or 279.00 and 279.06 in combination	3.21 (1.98, 5.21)	2.28E-06	97	27353

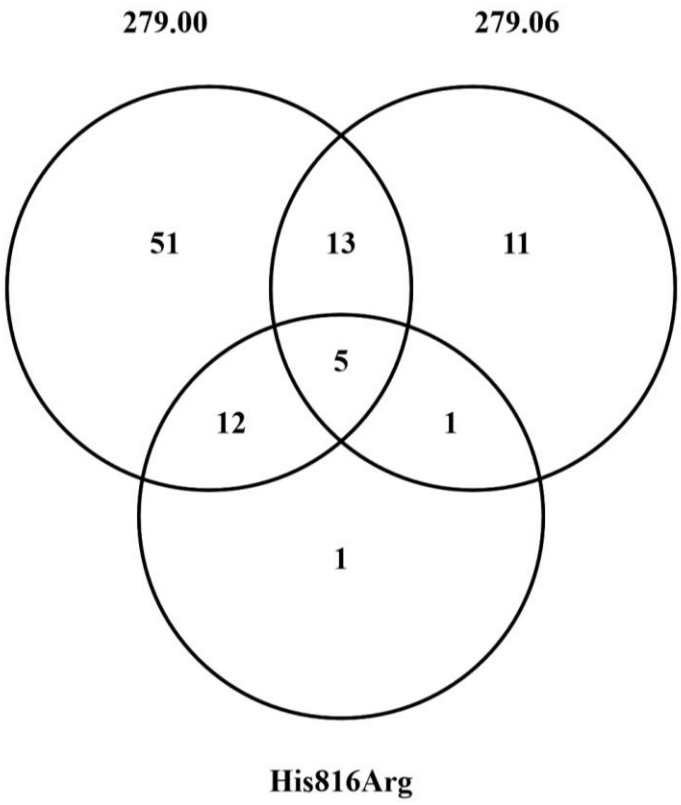


Figure 2-B. Venn diagram of distribution of His186Arg minor alleles in individuals with two incidences of the 279.00 ICD-9 code or the 279.06 ICD-9 code.

We then evaluated the median age at which people are first coded with 279.00. The median age at first code of heterozygotes (8.9 yrs) was significantly different, p=0.03, then the median age at first code of homozygous dominants (50.55 yrs) (Figure 2-C). The number of incidences of the 279.00 ICD-9 code in the two groups were not significantly different.

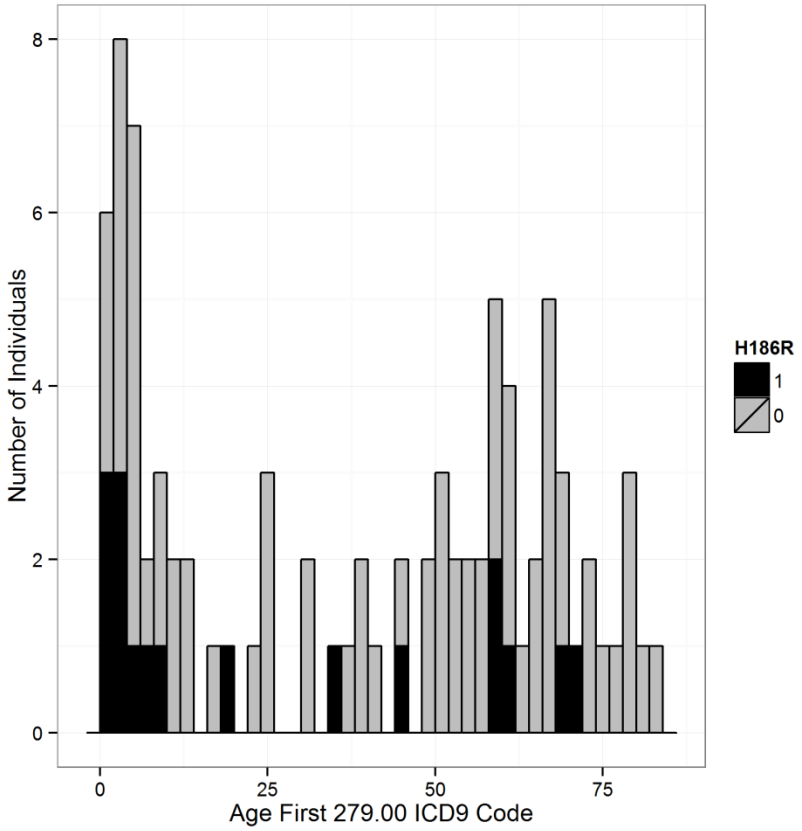


Figure 2-C. Histogram of the age at which Deficiency of Humoral Immunity individuals first have the 279.00 ICD-9 code in their record.

As the ICD-9 specific signal appeared to be driven by children, we wanted to see if other signals could be seen in a PheWAS of children alone. In individuals who never passed the age of 20 in their ICD-9 record, Deficiency of Humoral Immunity was again the only statistically significant hit, with an OR of 5.55 [95% CI 2.62 -11.76] and a p-value of 7.7E-06 (Figure 2-D, Appendix F). Our number of cases substantially decreased to 27 (9 with the 186R allele) reducing our power (Table 2-D), but we saw a stronger effect size in the younger individuals.

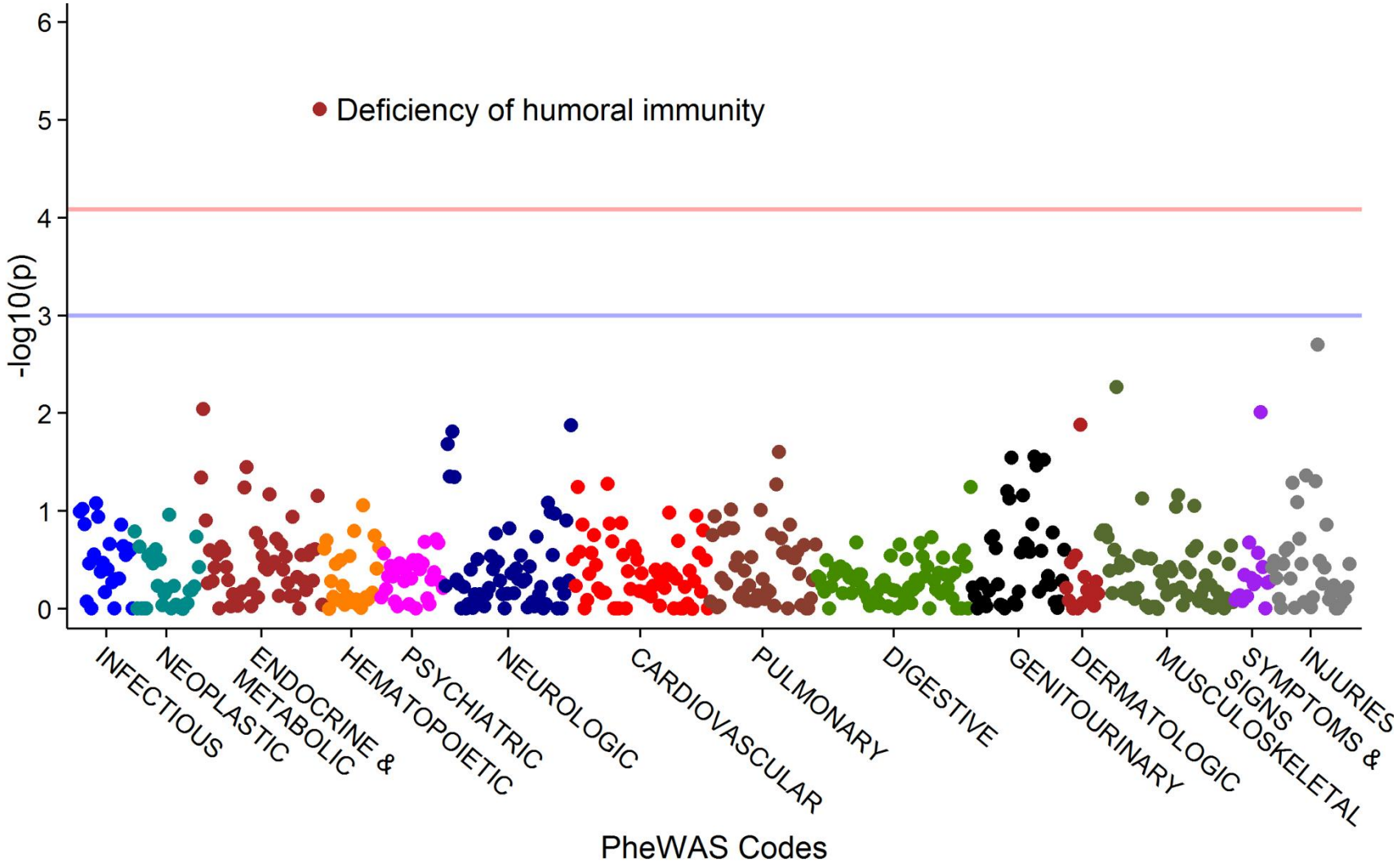


Figure 2-D. PheWAS manhattan plot for the A3G His186Arg variant in individuals under the age of 20. Red line indicates Bonferroni correction and blue line indicates level of suggestive significance.

Table 2-D. Genotype distribution in Deficiency of Humoral Immunity cases and controls for individuals under the age of 20 at their last ICD-9 code entry.

PheWAS Category and Case Status	His186Arg Minor Alleles		
	0	**1**	**2**
Deficiency of Humoral Immunity Cases	18	9	0
Deficiency of Humoral Immunity Controls	3295	229	6

Even when we further decreased the age cut off for individuals tested beyond an age at last code of 20, we still saw the association in the PheWAS. While we very quickly ran out of power to see a signal due to decreasing numbers of individuals with the His186Arg allele, even when we subset our data to those under the age of 8, the influence of the variant on Deficiencies of Humoral Immunity remained substantial (Figure 2-E). All of our younger groups showed an odds ratio of between 5.5 and 7.8, much stronger than in the whole population.

3544 individuals in our dataset had an '82784' CPT code in their record. The number of CPT codes for infusion was not different between carriers and non-carriers of the His186Arg allele (p=0.12). We also looked to see if the allele frequency in those with the CPT codes was higher than expected, but at 0.032 it was not significantly different than the allele frequency seen in the whole population used for our PheWAS. Almost all individuals who were PheWAS cases for Deficiency of Humoral Immunity (100 out of 108) had a '82784' CPT code in their record, including all 19 carriers of the 186Arg allele (Figure 2-F).

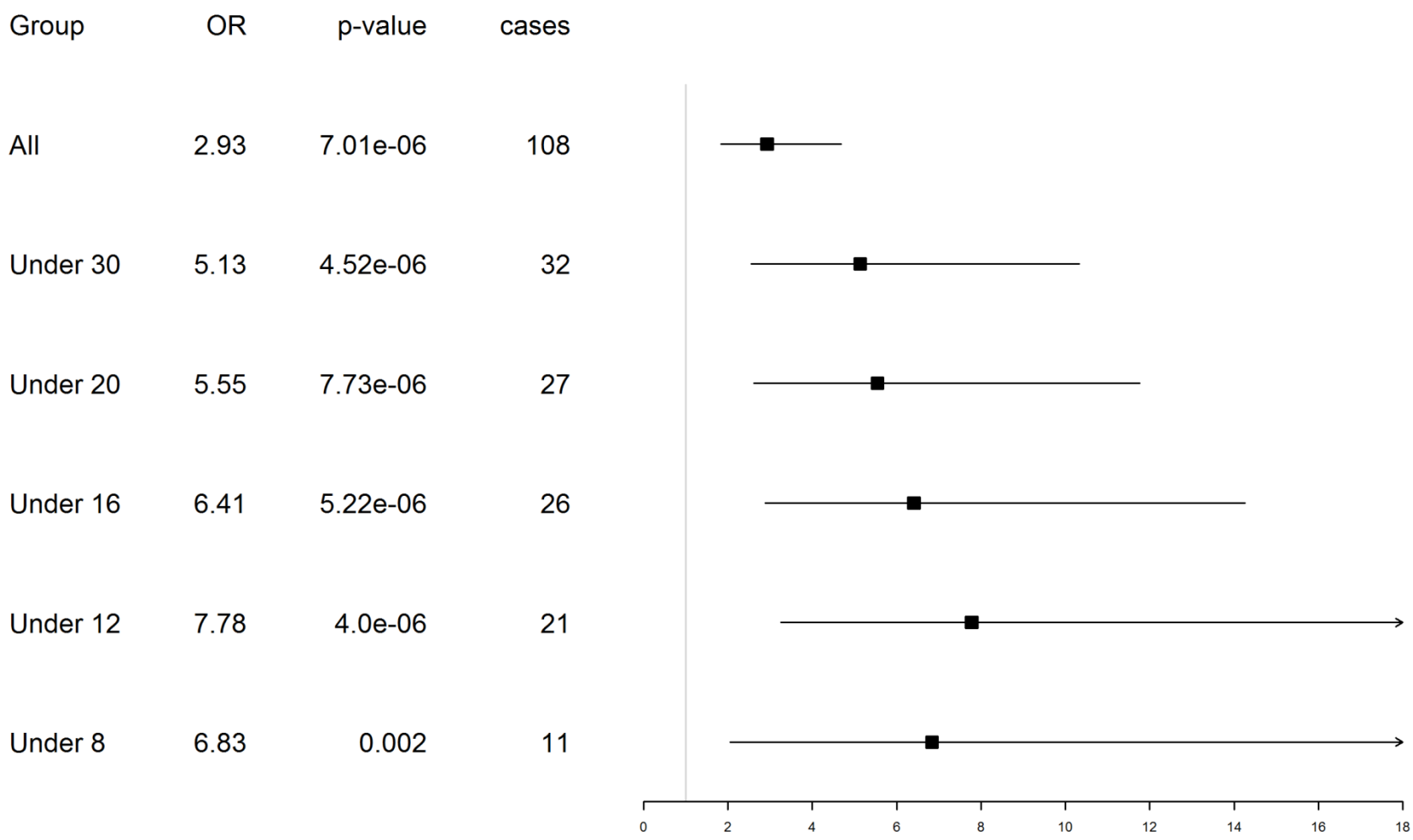


Figure 2-E. Forest plot for the association of His186Arg in the entire population, a population with an age at last code less than 30, a population with an age at last code under 20, 16, 12, and 8. Odd ratio (OR), p-value, and the number of cases in each of the PheWAS associations are shown.

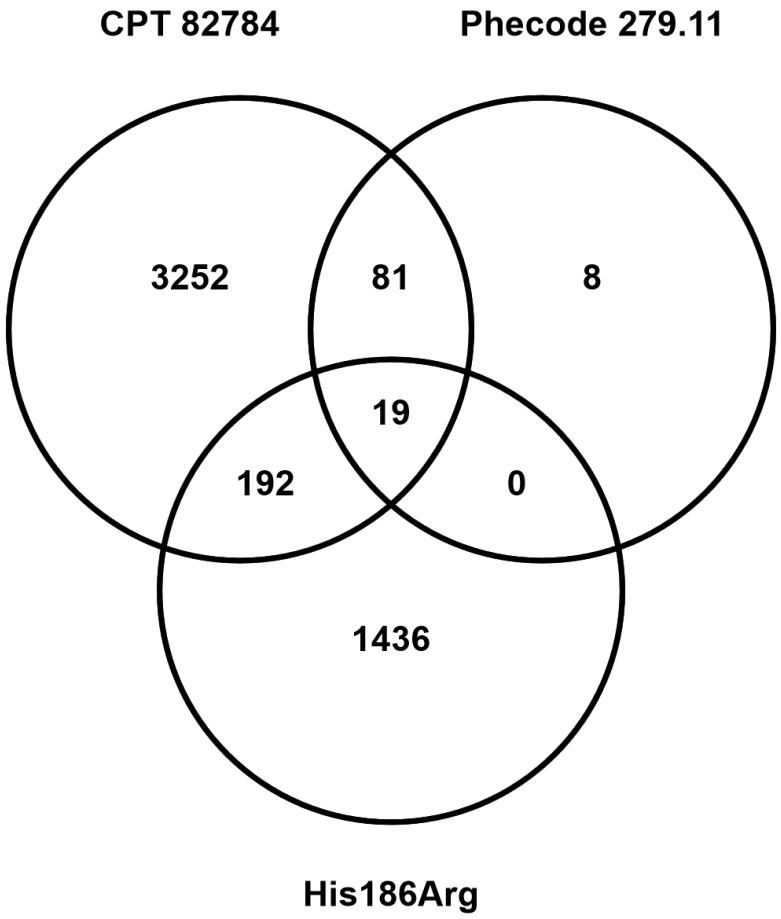


Figure 2-F. Venn diagram of overlap between individuals with an '82784' CPT code for Gammaglobulin Influsion and those who were cases for the Deficiency of Humoral Immunity PheWAS code. Distribution of individuals who are carriers for the His186Arg allele are also shown.

865 individuals with genotype data available had more than one IGG measure in their record. 198 of these were under the age of 19 at the time of their measure. None of the IGG summary measures for either the normalized or raw data were significantly different by genotype at the p=0.05 level. None of the summary measures was significant after stratification by age at test.

Discussion

In our analysis, the His186Arg variant in *APOBEC3G* was associated with the PheWAS code for Deficiency of Humoral Immunity. As our *a priori* hypothesis was that the variant would be associated with an immune related phenotype, this was not entirely unexpected, but an association that came mostly from young individuals was completely unexpected. We also anticipated more of an innate immune phenotype, rather than the adaptive immune phenotype of humoral immunity.

There is no established mechanism for how A3G might affect levels of immunoglobulins. A3G is widely expressed in hematopoietic cells, including B-cells, T-cells, and myeloid cells[82]. In these cells A3G protein localizes to the cytoplasm where it performs one of two functions; protect against exogenous retroviruses or prevent endogenous retroelements from completing retrotranscription and reintegrating into the genome. Within the cytoplasm, A3G exists in two forms called high molecular mass (HMM) A3G and low molecular mass (LMM) A3G. HMM A3G is A3G bound with ribonuclear protein complexes called P-bodies[83]. HMM A3G is less enzymatically active and functions to oppose retrotransposition of endogenous retroelements. Endogenous A3G expressed in H9 T-cells and mitogen-activated CD-4 T cells exists in a HMM state. The assembly into complexes is thought to be one mechanism of regulating the possibly mutagenic properties of A3G. By contrast, LMM A3G is much more enzymatically active, and is necessary for activity against exogenous retroviruses[83]. Interestingly, some RNA binding proteins involved in cell fate determination are thought to be part of these HMM complexes. Unpublished evidence from collaborators in the D'Aquila Lab at Northwestern University Medical School indicate that the A3G His186Arg change causes a

visible difference in levels of HMM and LMM complexes of A3G by Western Blot (data not shown). It is possible that a shift in the levels of HMM and LMM A3G could cause slight differences in the ability to deal with endogenous retroelements, which may alter how the adaptive immune system recognizes them. A change in the adaptive immune system could theoretically change the balance of different immunoglobulins. In the context of already sick children this could theoretically cause further immune dysfunction.

Despite the lack of an established mechanism to connect A3G and a Deficiency of Humoral Immunity, our evaluation of CPT codes shows that individuals are receiving intravenous immunoglobulin. Manual review of records for many carriers of the Arg allele indicates that many individuals in our dataset who receive ICD-9 codes indicating hypogammaglobulinemia or common variable immunodeficiency have additional medical problems. Many of the younger children in our dataset who were carriers and cases had notes indicating B-cell Acute Lymphoid Leukemia in their records previous to the Deficiency of Humoral Immunity ICD.9 codes, often by many years. While there is a well-established link between Chronic Lymphoid Leukemia (CLL) and hypogammaglobulinemia,[84–86] we did not see any association with CLL. In fact, when tested, the children who had ALL produced only a small signal, and those with both ALL and hypogammaglobulinemia produced a smaller signal than hypogammaglobulinemia alone. It is possible that our characterization of both CLL and ALL in the PheWAS was extremely poor, and our power to observe an association between His186Arg and either of these cancers was reduced or eliminated by poor phenotyping.

Interestingly, several of the other members of the APOBEC family, specifically Activation Induced Deaminase (AID) and APOBEC3B (A3B), have been associated with cancers[87,88]. One of the hypothesized reasons that A3G forms the HMM complex in the

cytoplasm is to stop it from entering the nucleus at certain times in the cell cycle where it could potentially damage single stranded DNA[83], though evidence shows that A3G is excluded from accessing chromatin at all stages of mitosis[89]. Normally, AID plays a role in hypermutation, class switch recombination, and gene conversion; the three processes for secondary antibody diversification in activated B-cells[90,91]. Mutations in AID have also been associated with hyper-IgM syndrome[92].

We did not see any signals specifically due to viral infection or any cancers. The only Phecode other than Deficiency of Humoral Immunity that even passed the suggestive significance line in our analysis was for Hallucinations. In the younger age datasets nothing else was even close to the threshold of suggestive significance. Given the multiple associations of this gene, and all others in the family, it was unexpected that the SNP was not pleiotropic in our analysis.

Given this association, the previous evidence of the role of His186Arg in HIV, and the unpublished data from our collaborator, we decided to search the GTEx[93] database to see if His186Arg was identified as an eQTL. In the relevant tissue, whole blood, rs1877832, the SNP responsible for the His186Arg change was an eQTL for A3G (t= -3.3, p=0.001) (Figure 2-G) though this was not significant once the total number of tissues tested was adjusted for. This effect was consistent across the vast majority of the GTEx tissues, where every copy of the Arg allele at A3G 186 decreased the expression of A3G.

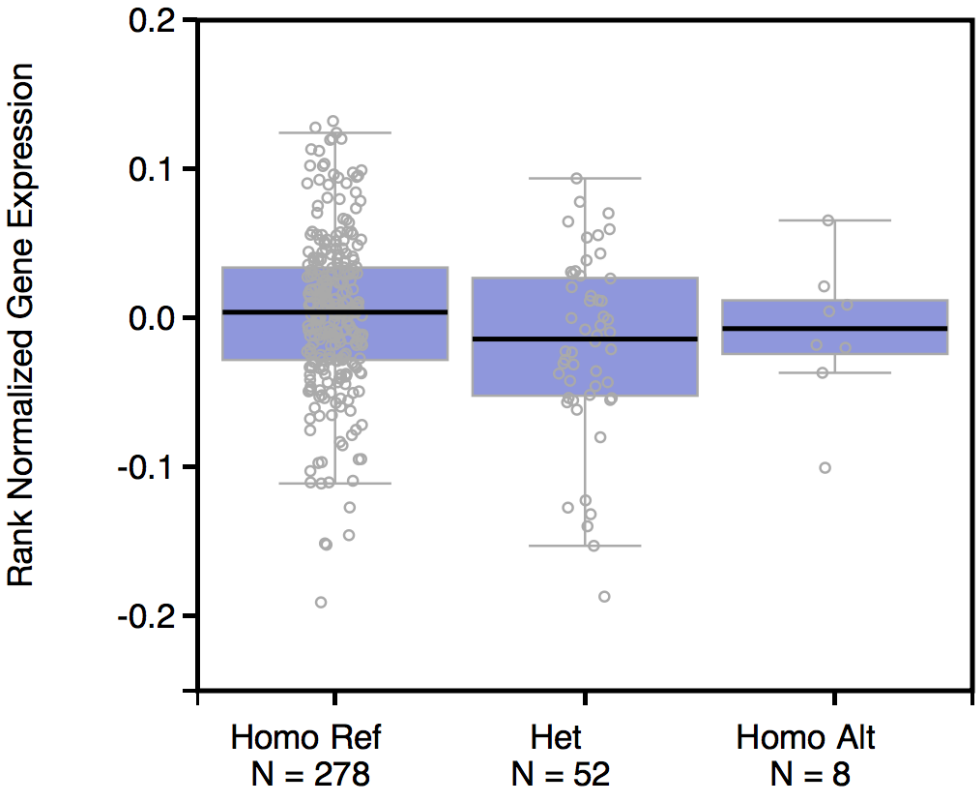


Figure 2-G. Rank normalized gene expression of *APOBEC3G* from whole blood for homozygous reference, heterozygotes, and homozygous alternate individuals for the His816Arg allele.

One of the main limitations of our study is the small number of children available, especially given the low minor allele frequency of the His186Arg variant. We attempted to explore this variant further in African American children as the variant is far more frequent in populations of African descent (~27% MAF), but in the more than 3000 African Americans in our dataset, there were only two PheWAS cases for Deficiency of Humoral Immunity. Despite this, we feel the association seen in Caucasian children is reasonable.

There is limited evidence of the function of A3G in children. Studies exploring the deamination ability of A3G in HIV-1 infected children found that the level of A3G did not differ statistically based on the level of A3G induced hypermutation[94]. Many studies in both children and adults with HIV-1 have found that the Arg genotype of A3G His186Arg is associated with a

significant decline in CD4 count[95]. Another study found that the Arg allele is associated with more rapid HIV-1 progression and central nervous system impairment in children[96]. While there is evidence that the adaptive immune systems of children are different than those of adults[97], how those differences might result in hypogammaglobinemia only in children remains unexplained.

In conclusion, we found the A3G His186Arg variant to be associated with a Deficiency of Humoral Immunity in a PheWAS and the ICD-9 code for Hypogammaglobulinemia in a targeted analysis. More analysis on the function of this variant in contexts outside of HIV-1 infection are necessary to understand its biological importance and further explore the association we saw. Further analysis of this variant in non-European descent populations will also be important if this association is generalizable to other racial and ethnic groups.

III. THE ASSOCIATION OF THE *APOBEC3B* DELETION WITH CARDIAC VALVE PHENOTYPES

Introduction

The APOBEC3s (A3) are a family of cytidine deaminases located on chromosome 22 (Chapter II –Introduction). The deletion of *APOBEC3B* (A3B) is a germline copy number variant that deletes the entirety of the coding region of A3B[98]. This deletion spans 29.5kb from the last exon of *APOBEC3A* (A3A) through the last exon of A3B, thereby removing the whole A3B gene and creating a fusion transcript that attaches the protein coding region of A3A to the 3' UTR of A3B[99] (Figure 3-A-a). This A3A_B fusion transcript is more stable than normal A3A, causing an increase in the level of A3A enzyme[99]. This deletion occurs in 6% of individuals of European descent, 9% of African Americans, and 37% in Asians[98].

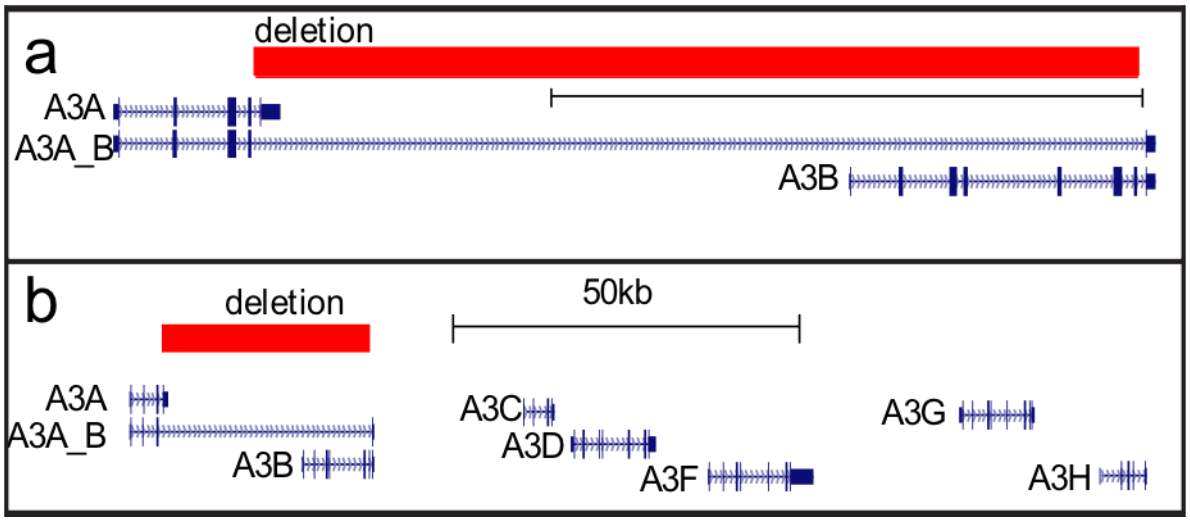


Figure 3-A. Location of the A3B deletion. a) The fusion transcript created spans from the last exon of A3A to the 3'UTR of A3B. b) Location of A3A, A3B, and the A3A_B fusion transcript in the context of the A3 gene family.

Recently, much attention has been devoted to understanding the role of the A3A and A3B in cancer. These proteins, part of the larger A3 family (Figure 3-A-b), have been established as the most likely cause of "mutation clusters" or "kataegis" seen in a variety of cancer types[100–102]. Both A3A and A3B are found in the nucleus[103,104], and these two proteins, like other members of the A3s cause deamination at cytosines on single-stranded DNA, resulting in C to T transitions[72,105,106]. The cytidine deaminase activity of A3B has been found to be a source of mutation in breast and other cancers[88]. The deletion of A3B has been found to increase the risk of breast cancer in both Asian and European women[107,108], possibly through the upregulation of A3A. In addition to cellular mutation and cancer, members of the A3s are involved in innate immunity to retroviral infection[109,110] and inhibition of endogenous retroelements[111,112].

The importance of this deletion has also been investigated in conjunction with infectious disease phenotypes. This deletion was found to be associated increased attenuation of the Hepatitis B Virus and hepatocellular carcinoma[113–115], and has been implicated in resistance to malaria[116]. The A3B deletion has been tested for an association with HIV, however results have been mixed[117,118]. A3A and A3B have also individually been tested for their importance in infectious disease phenotypes. **As studies in both cancer phenotypes and infectious diseases have shown the A3B deletion to be pleiotropic, we hypothesized that it would have additional disease associations. We tested this hypothesis with using a PheWAS in the hopes that we could validate existing phenotypes while discovering new ones.**

Methods

BioVU dataset

In our preliminary analysis, we used individuals for whom DNA had been collected in BioVU, Vanderbilt's de-identified DNA databank. These samples are linked to electronic medical records with all identifying information removed. A subset of individuals in BioVU have had genotyping performed on one or more of several platforms, the Illumina Omni-Quad, the Illumina 660W, and the Illumina Omni5-Quad . For our discovery analysis, we used data from 6332 individuals genotyped on the Illumina Omni-Quad, though data for individuals on all platforms was obtained. Age at last record was calculated for all individuals based on their date of birth and the last ICD.9 code represented in their medical record. Individuals with a third-party identified race that was not listed as white or an age at last ICD-9 record less than 18 were removed leaving 4948 individuals with phenotypic and demographic data.

Genetic Data Quality Control and Deletion Imputation

Called genotypes underwent QC before imputation. Each platform underwent quality control separately. Briefly, SNPs with a genotype efficiency of less than 98% were removed, as were individuals in whom fewer than 98% of SNPs were genotyped. Relatedness between individuals in the dataset was checked, and related individuals were removed. Sex and Mendelian inheritance checks were also performed. Strand alignment and pre-phasing of study genotypes was done with Shapeit[119].The reference panel used for phasing and imputation was the 1000Genomes Phase 1 version 3. Imputation was performed on each platform separately using

Impute2[120] in chucks of 5MB. Post-imputation genotypes were called based on a 90% threshold filter on the Impute2 values. Genotypes for the A3B deletion were extracted.

eMERGE dataset

Our replication analysis was performed in individuals genotyped as part of eMERGE 1[121]. These individuals were from one of four sites, Group Health Cooperative, Marshfield Clinic, Mayo Clinic, or Northwestern; each site has a DNA biorepository linked to an EMR, and individuals at these sites were genotyped on the Illumina 660W [27].Individuals from Vanderbilt University are also part of eMERGE, but these individuals were excluded from our study group to eliminate possible overlap. Genotype QC and imputation was performed as described above for the BioVU dataset.

PheWAS and Phecode correlation

For individuals in each of our datasets, complete ICD-9 code records were obtained, along with the date of each ICD.9 code entry. PheWAS case, control, and exclusion status was determined using a minimum of two ICD-9 codes in a category to be a case, as described previously. PheWAS analyses and meta-analysis were performed using the PheWAS package in R[50]. Bonferroni, false discovery rate (FDR), and Simple-M corrections were used. Correlations amongst PheWAS codes in individuals in our dataset were also determined using pairwise Pearson correlation tests. Correlations were plotted using the "correplot" package in R.

Ejection Fraction Analysis

Ejection fraction (EF) data was obtained for a subset of individuals in our dataset. Data had been previously extracted by the Denny group. Ejection fraction measures greater than 55 were censored to 55 as this was a common cut off for clinical practice. Once patients have an ejection fraction of 55, they clearly do not have a reduced ejection fraction, so many physicians will enter 55 for any number over 55. Summary measures including median, maximum, and minimum ejection fraction were calculated. Individuals were also classified dichotomously as having a high EF if their median EF was greater than 50. Median EF measures were split into three groups, those greater than 50, 35-50, and less than 35. These thresholds were chosen following consultation with physicians. Allele frequency for the A3B deletion was compared across these groups. A secondary analysis was performed where individuals that were Aortic valve malfunction cases from our PheWAS were removed, as we might expect them to have a different range of ejection fraction than other patients.

Statistical Analysis

We performed PheWAS using logistic regression on PheWAS codes with more than 20 individuals identified as cases. Age at last ICD.9 code in record, gender, and the first three principal components were included as covariates in the regression. For the replication in the eMerge dataset, site was included as an additional covariate in our logistic regression analysis.

Results

The A3B deletion imputed with an information score of 0.76 on the Illumina Omni-Quad used in our discovery analysis. From our BioVU population of 5198 white individuals we were able to impute an A3B deletion genotype for 4230 of them (Table 3-A). These 4230 individuals were our discovery population.

Table 3-A. Summary statistics of individuals genotypes on the Illumina Omni-Quad and used for PheWAS after QC measures were implemented. Shown in the whole dataset and by deletion status.

	All n=4829	A3BΔ==0 n=3471	A3BΔ==1 n=443	A3BΔ==2 n=16
Gender M (%M)	2581 (53.4)	1845 (53.2)	246 (55.5)	10 (62.5)
Median Age last record (IQR)	61 (49, 71)	61 (49, 71)	60 (47, 71)	58 (46, 67)

Phenotype categories were defined through sets of related ICD.9 codes. We required that a phenotype have at least 20 cases for analysis, and 908 PheWAS codes met that requirement, resulting in a Bonferroni corrected significance threshold of 0.05/908 = 5.5e-05, though this correction is overly stringent given that the tests are not truly independent. Three of the phenotype categories reached the Bonferroni corrected significance level (Figure 3-B). All of the phenotype categories that passed a suggestive significance level of 0.001 were related to abnormalities in cardiac function (Appendix I). Heart Failure NOS was the most significant hit (OR [95% CI] =2.03 [1.46, 2.82]; p= 2.63E-05), followed closely by Nonrheumatic Aortic Valve Disorders (1.98 [1.42, 2.76]; p=5.02E-05). The category for Systolic/diastolic Heart Failure also passed our Bonferroni correction threshold (1.65 [1.30, 2.11]; p=5.16E-05). Two more closely related phenotypes, Heart Valve Disorders and Heart Failure, also passed the FDR and Simple-M

corrections in our data (Table 3-B). All of these had substantial numbers of case individuals, well over the minimum 20 required by our analysis. These phenotype associations with the A3B deletion were used for replication in the eMERGE dataset.

As we saw a large number of cardiac related PheWAS codes appear as significant, we considered whether the same individuals might be driving the association of many different codes. Correlation tests amongst all Phecodes that reached the level of p=0.01 or better in analysis showed that while there is some overlap between all the cardiac codes, not all cardiac codes are correlated (Figure 3-C). Heart Failure NOS and Aortic Valve Disorders, our top two hits, while made up of different ICD-9 codes (Appendix J), have partially but not completely overlapping case populations. 158 out of the 299 individuals with Aortic Valve Disorders were also Heart Failure NOS cases (Figure 3-D). This amounted to 16.8% of Heart Failure NOS patients.

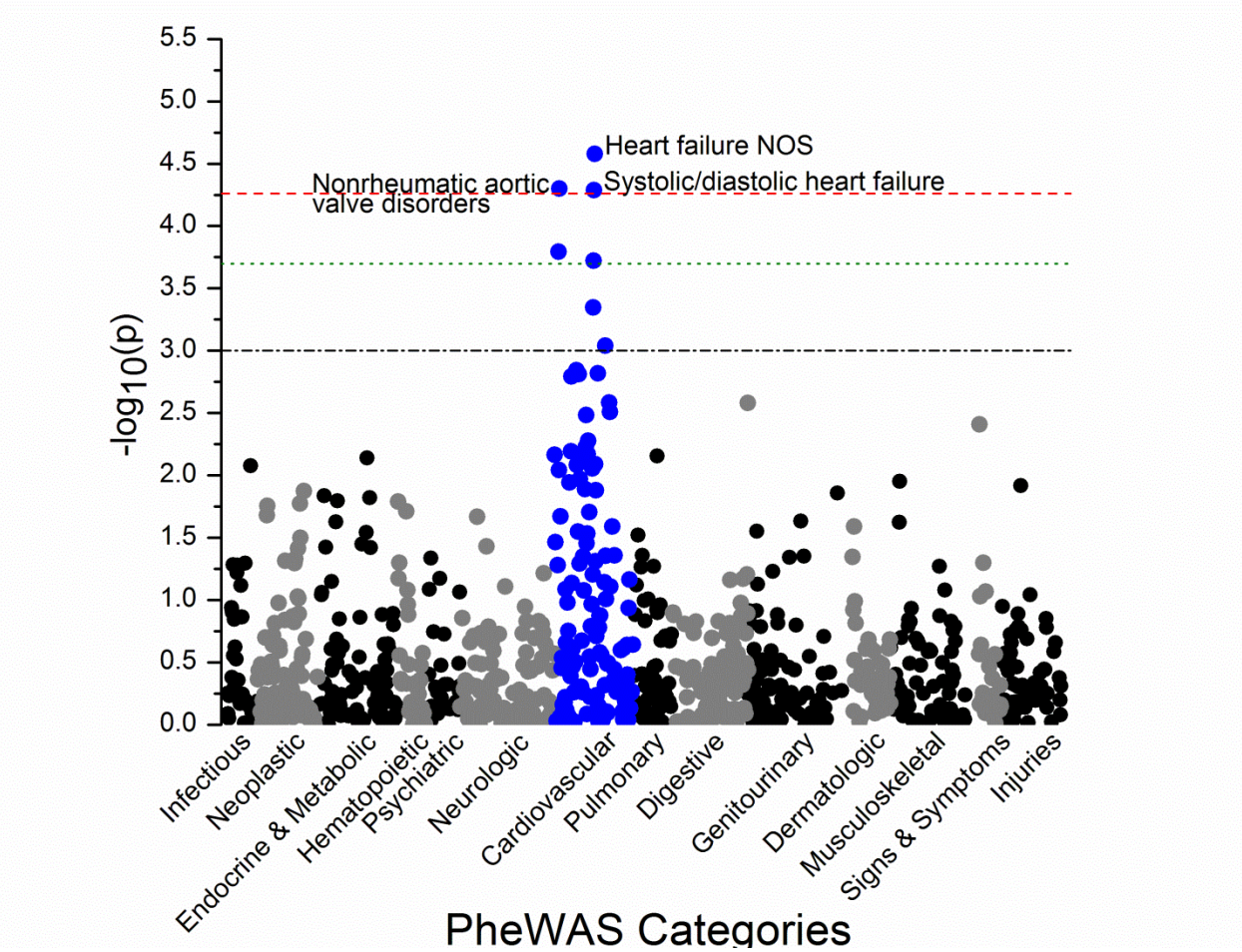


Figure 3-B.PheWAS manhattan plot. Groups of PheWAS categories are displayed on the x-axis. Red line indicates level of Bonferroni correction (5.45 e-5) and black line indicates line of suggestive significance (0.001). All dots above the green line are those that are significant according to the FDR and Simple-M phenotypes corrections implemented in PheWAS.

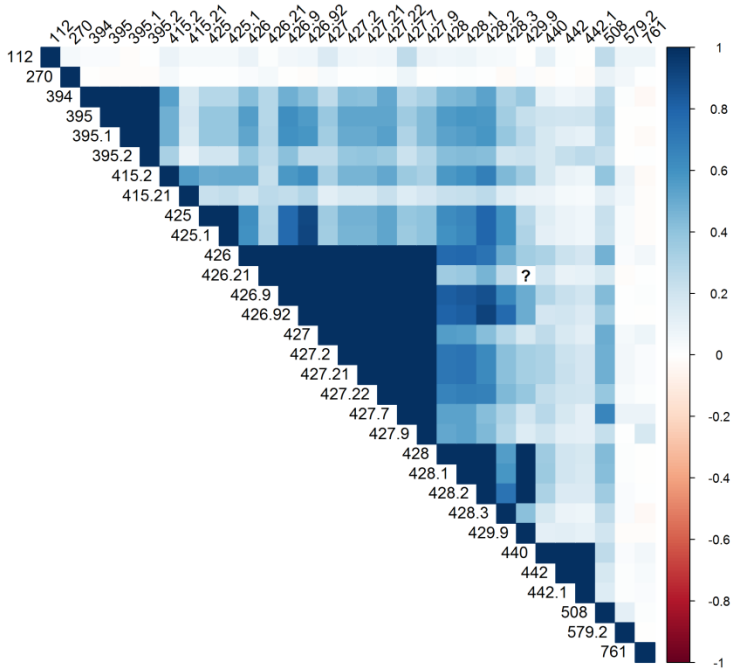


Figure 3-C. Pairwise correlations amongst PheWAS codes that had a p-value of less than 0.01 in our A3B deletion analysis. Cells marked "?" have no individuals with complete pairwise records.

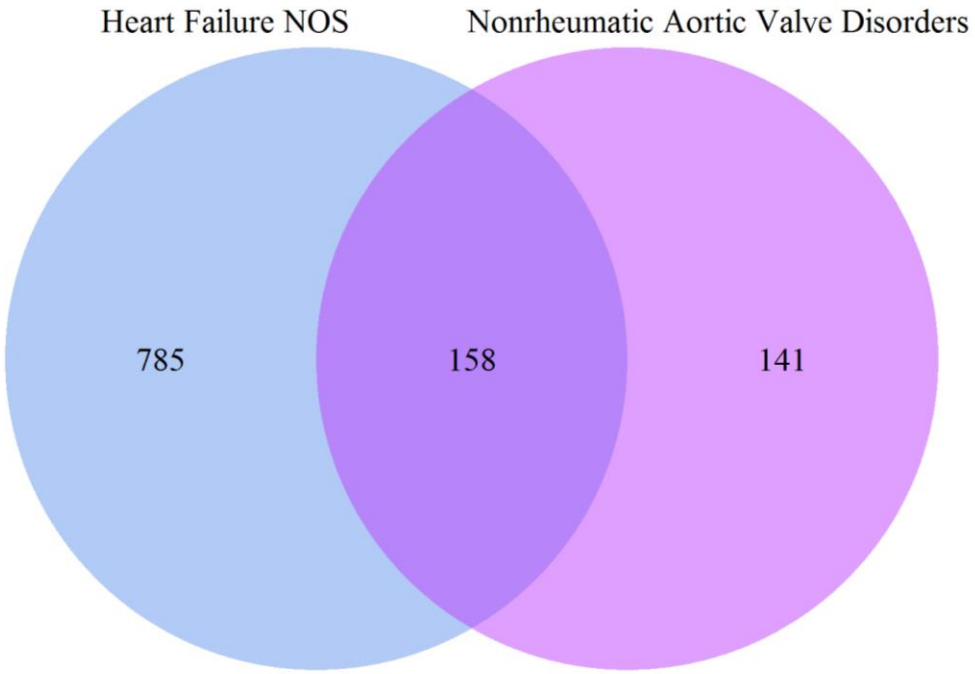


Figure 3-D.Venn diagram of overlap of individuals who are cases for Heart Failure NOS and Nonrheumatic Aortic Valve Disorders Phecodes.

To attempt to confirm our analysis, we performed a meta-analysis within BioVU. We performed PheWAS in the Omni5-Quad (1,599 individuals) and 660W populations (3,532 individuals) before performing a fixed-effect meta-analysis. The A3B deletion had an imputation score of 0.72 on the 660W and 0.9 on the Omni5-Quad. None of our top hits from the Omni-Quad were significant in the other two datasets (Figure 3-E), though the meta-analysis result was trending towards significance. The case distribution was quite different between the three sets, with far greater case number and case/control ratio for Heart Failure NOS on the Omni-Quad than on the other two platforms. The same was true for Nonrheumatic Aortic Valve Disorders. The meta-analysis resembled the Omni-Quad results far more than the other platforms, possibly due to the uneven case distribution.

In the eMERGE dataset, we only evaluated those categories that were significant in our preliminary analysis. The eMerge dataset consisted of 14,104 European American individuals from four different medical centers. The three heart failure codes (Heart Failure NOS, Systolic/diastolic Heart Failure, and Heart Failure) that were significant in our Omni-Quad discovery analysis were not significant within the eMERGE set. We only evaluated specific Phecodes seen in our discovery analysis in the eMerge set; we did not test other codes outside those targeted for replication. Both Nonrheumatic Aortic Valve Disorders (1.29 [1.00, 1.66], p=0.046)**,** and Heart Valve Disorders (1.26 [1.05, 1.51], p=0.01), were significant at the p=0.05 level in our replication (Table 3-C). We attempted to replicate five codes, only Heart Valve Disorders was significant at the level of the Bonferroni correction.

In our data, the presence of the A3B deletion was not associated with a lower minimum ejection fraction in all individuals, but once individuals with the Nonrheumatic Aortic Valve

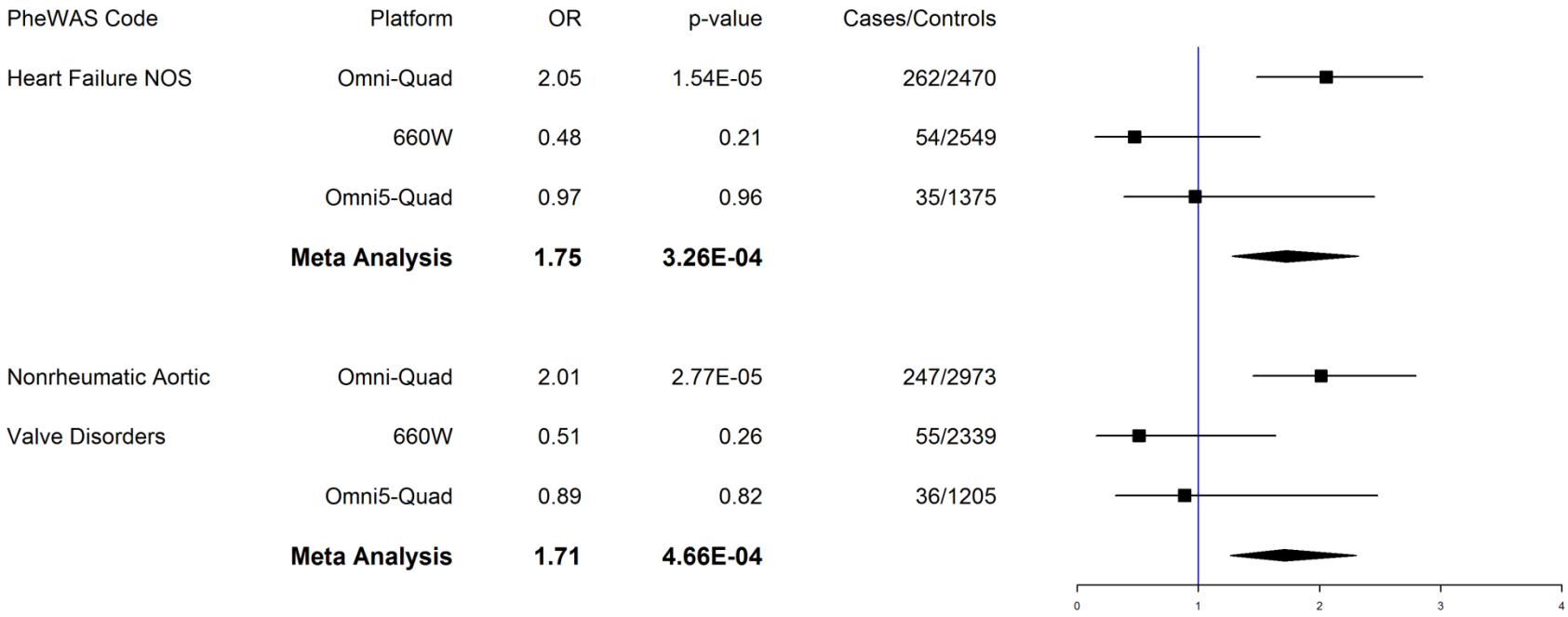


Figure 3-E. Forest plot of PheWAS results for Heart Failure NOS and Nonrheumatic Aortic Valve disorders codes in all BioVU sets and in a fixed effects meta-analysis.

Table 3-B.Results from the PheWAS that pass the FDR and simple-M correction thresholds.

PheWAS Code	PheWAS Code Description	OR (95% CI)	p-value	N cases	N controls
428.2	Heart Failure NOS	2.03 (1.46, 2.82)	2.63E-05	260	2389
395.2	Nonrheumatic Aortic Valve Disorders	1.98 (1.42, 2.76)	5.02E-05	244	2887
428.1	Systolic/diastolic heart Failure	1.65 (1.30, 2.11)	5.16E-05	772	2389
395	Heart Valve Disorders	1.61 (1.26, 2.05)	0.00016	586	2887
428	Heart Failure	1.56 (1.24, 1.98)	0.00019	865	2389

Table 3-C. Replication of hits from preliminary PheWAS in the eMERGE dataset.

PheWAS Code	PheWAS Code Description	OR (95% CI)	p-value	N cases	N controls
428.2	Heart Failure NOS	1.04 (0.68, 1.66)	0.78	271	8776
395.2	Nonrheumatic Aortic Valve Disorders	1.29 (1.00, 1.66)	0.046	898	10164
428.1	Systolic/diastolic heart Failure	1.05 (0.85, 1.29)	0.66	1994	8776
395	Heart Valve Disorders	1.26 (1.05, 1.51)	0.01	586	10164
428	Heart Failure	1.08 (0.88, 1.32)	0.46	865	8776

Disorders Phecode were removed, the we saw that the deletion was associated with lower minimum ejection fraction (B [95% CI] = -4.40 [-8.03, -0.77)]; p=0.02) (Table 3-D). Since the deletion was associated with valve disorders in our analysis, we were concerned that these individuals might have low ejection fraction levels and bias our analysis. As it is often not the exact measure, but whether individuals are below a certain threshold, that is clinically important, we also tested the impact of the variant on whether individuals had a low ejection fraction. In the whole population, the variant was not quite significant (OR [95% CI]=1.44 [0.98, 2.13], p=0.06), while in the group lacking the aortic valve malfunction, a copy of the deletion increased the odds of having a median ejection fraction considered low (2.26 [1.17, 4.36], p=0.02) (Table 3-E).

Table 3-D. Association of the A3B deletion with minimum ejection fraction in all individuals and all individuals except those that were Nonrheumatic Aortic Valve Disorders cases in our PheWAS.

	All (n=1,051)	All except Aortic Valve Malfunction (n=482)
	Minimum Ejection Fraction	Minimum Ejection Fraction
A3BΔ Allele	-1.71 (-4.12, 0.70) p=0.17	-4.40 (-8.03,-0.77) p=0.02
Age at Minimum	0.05 (-0.01, 0.10) p=0.09	0.005 (-0.07, 0.08) p=0.89
Sex==M	-4.48 (-6.20, -2.76) p=4.1e-07	-4.53 (-6.90, -2.15) p=2.1e-04

Table 3-E. The role of the A3B deletion in increasing the odds of having a median ejection fraction classified as low.

	All (n=1,051)	All except Aortic Valve Malfunction (n=482)
	Low Ejection Fraction	Low Ejection fraction
A3BΔ Allele	1.44 (0.98, 2.13) p=0.06	2.26 (1.17, 4.36) p=0.02
Age at Minimum	0.99 (0.98, 1.00) p=0.22	1.00 (0.99, 1.02) p=0.71
Sex==M	2.03 (1.49, 2.78) p=9.2e-06	2.77 (1.57, 4.90) p=4.6e-04

As breast cancer has been previously associated with the A3B deletion[88,107], we checked for an association in our population. The phenotype category Breast Cancer had 162 cases with a nonsignificant OR = 1.1 [0.72-1.83], p=0.56. Because the controls in our general analysis were both men and women, we performed a specific analysis using only females. In the female-only analysis the code for breast cancer was still not significant (OR = 1.2 [0.75-1.93], p=0.44), so we did not replicate the previously reported association[88,107]. Furthermore, no infectious disease phenotypes reached either significant or suggestive p-value levels, despite the known role of the A3 proteins in viral infection control[109,114,122]. The infectious diseases previously associated with the A3B deletion, HIV[117] and HBV[115], were tested in our PheWAS but had very few cases each (36 and 25 respectively).

Discussion

The association of the A3B deletion with phenotypes relating to heart function was unexpected. Previous associations have been to viral phenotypes and cancer risks and changes in the characteristics of enzymatic mutation in cancer[88,99,102,107,114,115]. One possible hypothesis for why the A3B deletion could be associated with heart phenotypes is that in heart tissue the deletion causes inflammation, which results in valve issues. However, endocarditis was not significantly associated with the A3B deletion in our discovery dataset (OR = 1.4 [0.74, 2.62], p = 0.30, 69 cases). If an inflammation related mechanism was at the base of this association, it is possible we would not see if despite testing for PheWAS codes including endocarditis because not all individuals who have effects from this process are severe enough to be diagnosed as such or the ICD.9 code for endocarditis is a poor marker for the actual phenotype. It is also possible

that a subset of individuals in this dataset have an infection to which the A3s would respond which is driving an association though it is not captured well by ICD.9 codes or we only capture it through secondary manifestations. Despite this, we were able to replicate the associations of valve phenotypes in an independent EMR dataset.

As we saw so many cardiac phenotypes in our dataset, it is not clear if the true association is to a phenotype represented by the Phecodes we saw or to some other condition for which people may be billed with one or more of these ICD-9 codes prior to or during diagnosis. There are clearly two distinct populations of cardiac phenotype patients that are enriched for the A3B deletion in the Illumina Omni Quad set in BioVU, though only one of those populations replicated. These populations seem to be quite different than those on the other BioVU genotyping platforms and those in eMerge. As each GWAS platform in BioVU was put together based on the presence and absence of individuals with specific phenotypes, it is not surprising that the results may be different across platforms. For example, the 660W was assembled partially to reduce the presence of cardiac patients on the platform; a vital distinction for our analysis. As all controls for PheWAS are potentially case patients for other phenotypes, it is possible that this ascertainment procedure causes some difference in association between the platforms. In the eMerge set we had to account for the different sites which contributed data, especially since each site's population is ascertained with different criteria.

The major limitation of this study is that we were unable to replicate previously known associations. Our inability to replicate associations with HIV is perhaps not surprising as well phenotyped cohorts have discordant results on whether that association is real[123–125]. Susceptibility to HBV is another previously existing association of the A3B deletion that we were not able to replicate, but case numbers were low, and the availability of a vaccine was not

accounted for within the PheWAS controls. For both of these diseases it is likely that PheWAS aggregations of ICD-9 codes alone provide inadequate information for ascertainment of true cases and controls. More concerning is that despite the known association of the A3B deletion with multiple cancers, we did not see a signal for any cancers in our analysis. One possible reason for this is that the aggregation approach used in PheWAS may not be ideal for capturing cancers. While some cancers may be ascertained well if one instance of an ICD-9 code is present, others may require 3 or even four instances of ICD-9 codes before they are well captured. By requiring two ICD-9 codes for case status, we would be excluding many true cases in the former scenario, and falsely classifying individuals as cases in the latter. Both these options would cause us to lose power through poor phenotype ascertainment. Also, some of the cancers the A3B deletion has been associated with predominantly or only occur in a single sex. As our initial PheWAS was mixed sex, we would have decreased our power to see these cancers. Despite this, we had assumed we would be able to see a cancer signal before we began our analysis.

Another potential limitation of the study is the genotype itself. The A3B deletion was imputed, and while we understand that deletions with reasonable imputation scores should be robust (unpublished communication with Evan Eichler), it does leave a certain amount of uncertainty in our data. Furthermore, deletions genotypes were assigned absolutely instead of using the probability of being a certain genotype for analysis. This means that we have far more individuals with missing genotypes than theoretically possible. It also means that more individuals are missing the genotype than are missing the SNPs used to impute it, resulting in a lower minor allele frequency than many reference populations and other studies report. While the imputed genotype does add a degree of uncertainty, there is only one SNP in A3B in strong LD

with the deletion, and it is not commonly placed on GWAS platforms. Using the imputed deletion allowed us to test a genetic variant that would not have otherwise been possible. This assignment of genotypes also creates an issue when analyzing multiple genotyping platforms together. Different platforms impute the A3B deletion with different imputation scores, and this affects the number of people confidently assigned a genotype. In BioVU, proportionally far more individuals on the 660W were missing genotypes than for the Omni5-Quad, because the SNP imputed with much greater confidence on the Omni5-Quad.

Despite these limitations, we were able to find and replicate a novel association of the A3B deletion with Aortic Valve phenotypes and provide substantial evidence that the A3B deletion is important in a number of cardiac related conditions. Future studies in carefully phenotyped cohorts will be important for identifying the condition underlying these PheWAS associations.

IV. CHARACTERIZATION OF COMMON DELETIONS ACROSS THE GENOME BY THEIR PHENOTYPIC IMPACT

Introduction

While SNPs are the most frequently analyzed genetic variants in the genome, they are far from the only variation present in the human genome. Deletions large enough to be classified as structural variants (SVs) or copy number variants (CNV) are present throughout the genome. These CNVs may be individually rare, but many occur in the population[126]. Deletions have long been associated with rare diseases[127,128], and more recently have been investigated for their impact in common disease[129].

While structural variants have traditionally been difficult to detect, the 1000 Genomes database provides a detailed view of both single nucleotide and large polymorphisms in humans The phase 1 version 3 research identified more than 14000 large deletions[130]. 1000 Genomes defines a SV as an insertion or deletion of 50 basepairs (bp) or larger. Despite a growing ability to examine the importance of these deletions in previously obtained data, their impact in disease is still not clear. Previous publications on genome-wide CNVs have concluded that common CNVs will not account for much of the unexplained heritability in diseases[131], and some have even tested all common CNVs in the genome with a limited group of phenotypes, and have found few or no associations[129]. Despite this, smaller studies have found that deletions may play an important role in some diseases including schizophrenia[132], autism spectrum disorders[133], autoimmune diseases, HIV[134], and risk for certain cancers[108]. CNVs have also been shown to be responsible for a portion of the differences in gene expression between individuals[135].

Deletions provide us with a "natural knock out" experiment, allowing us the clearest picture of what happens in a human when a piece of DNA is removed. Understanding the impact of removing areas of genes, areas of non-coding DNA that are regulatory elements, or areas that don't appear to be either, will enhance our understanding of how our genome contributes to disease. We can also examine the impact of the amount of the genome deleted, which will give us insight as to whether the size or position of the feature is generally more important. **We hypothesized that different deletions would have different likelihoods of causing a phenotypic consequence depending on deletion characteristics like size and location. For example, a deletion overlapping all or part of an exonic region of a gene would be more likely to result in a phenotypic effect than those that that did not.** We tested this hypothesis by performing PheWAS on imputable deletions in BioVU.

Methods

Deletion Calling

Initial data processing involved quality control of both the genotype data, and phenotype processing. Genotyping performed on the Illumina Omni-Quad GWAS chip was available for 5,198 self-identified white individuals in BioVU. SNP quality control and imputation is described in Chapter 3. While SVs have long been identified as potentially problematic for imputation[136,137], newer reference panels including 1000 Genomes allow imputation with reasonable accuracy (unpublished communication with Evan Eichler). A total of 13,805 deletions were imputed using 1000 Genomes Phase 1 version 3 as a reference panel. These deletions were put through a pipeline to prepare them for analysis (Figure 4-A).

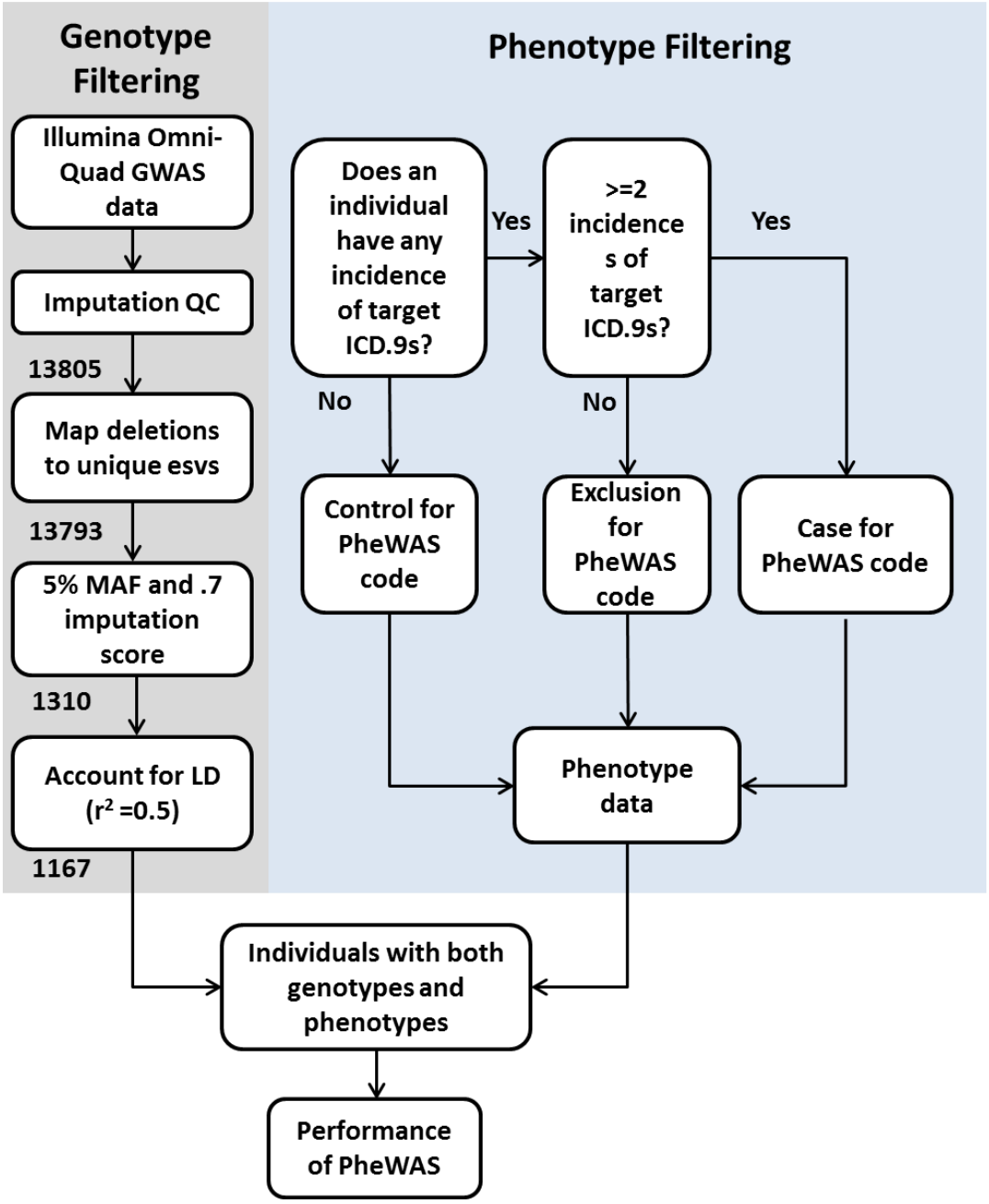


Figure 4-A. Flowchart of deletion imputation, PheWAS code aggregation, and data merging.

Deletion Mapping and Categorization

First we mapped each imputed deletion to the list of structural variants available through the database of genomic variants (DGV) [138] using tabix. For this project we used the "esv" annotation, which denotes structural variants that were submitted by the European Bioinformatics Institute (EBI). Any deletion that did not map to an esv or mapped to more than one esv was removed. Next we removed any deletions that did not have a minor allele frequency (MAF) of at least 5% or an imputation score of > 0.7 in our population. This left us with 1,255 deletions for analysis. LD was tested in these deletions and deletions with an r2 >=0.5 were identified. We were left with 1167 unique deletions.

BED files for structural variants were assembled using start and endpoints as determined from dbVar[139] dataset estd199, the structural variants submitted for 1,092 individuals sequenced by the 1000 Genomes Project (http://www.ncbi.nlm.nih.gov/dbvar/studies/estd199/#experdetailstab). The extent to which imputed deletions overlapped with genes (start and end positions) or individual exons was determined using BedTools[140] specifically the intersectBed function. Bed files containing all RefSeq transcripts[141] were downloaded from the UCSC Genome Browser (hg19, refflat table)[142]. In order to include each gene just once, we selected transcripts with the earliest transcriptional start point and the most distant transcriptional endpoint. In addition, BED files were created for the exon of each gene including all exons across multiple isoforms with one or more transcripts but with each exon represented only once. The dataset includes over 14 thousand larger deletions and captures 98% of variants at 1% frequency enabling the imputation of both common and low frequency structural variants. While the majority of variants had clearly defined start and endpoints, for several the exact base pair at which the structural variant begins could not be

determined with absolute certainty. In these cases, we used the widest region defined by the confidence interval, termed "outer-start" and "outer end" in this dataset.

Population Demographics and Phenotypic Data

Demographic data and ICD-9 code records were obtained. ICD-9 codes were aggregated into PheWAS codes as described previously and displayed in Figure 4-A. Sex was obtained from third party reporting, and age at last ICD-9 code entry was calculated from date of birth and the ICD-9 code record. Data was stratified to only include individuals of European descent with an age at last record over 18.

Statistical Analysis

PheWAS were performed as logistic regression adjusting for age at last record, sex, and the first three principal components using the PheWAS package in R. Statistics for the deletion predictor from PheWAS outputs were compared using Kolmogorov–Smirnov (ks) tests and Wilcoxon rank sum tests. Statistical analyses were performed with R[51].

Results

We imputed 1310 deletions. These deletions had info scores between 0.7 and 1, with 1 being the most common (Figure 4-B-a). 283 deletions had a minor allele frequency between 5% and 10%, and the number of deletions in each 5% bin decreased until 40% (Figure 4-B-b). The majority of these deletions did not overlap a gene, and only 9 overlapped more than 2 genes (Figure 4-B-c).

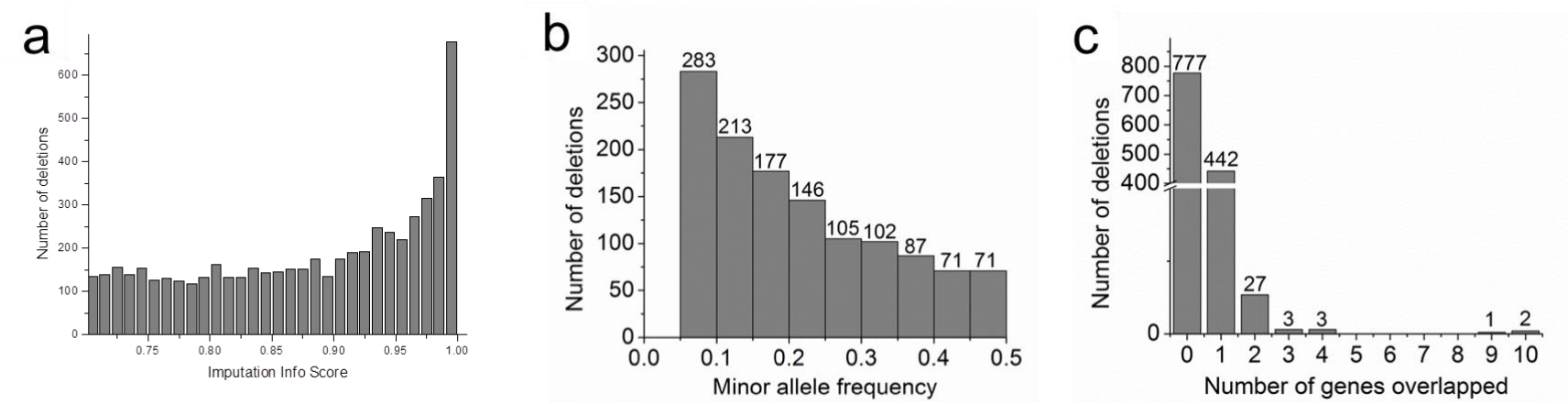


Figure 4-B. Characteristics of imputed deletions. Distribution of a) imputation info scores, b) minor allele frequencies, and c) number of genes overlapped.

Our population consisted of 5198 individuals of European descent previously described in Chapter 3(Table 3-A). These individuals had a median age at last code of 61 years old. While most of these individuals were PheWAS cases for between 20 and 30 PheWAS codes, some were PheWAS cases for over 200 codes (Figure 4-C-a). Individuals had an average of just under 60 deletions, and around 25 deletions that overlapped genes (Figure 4-C-b).

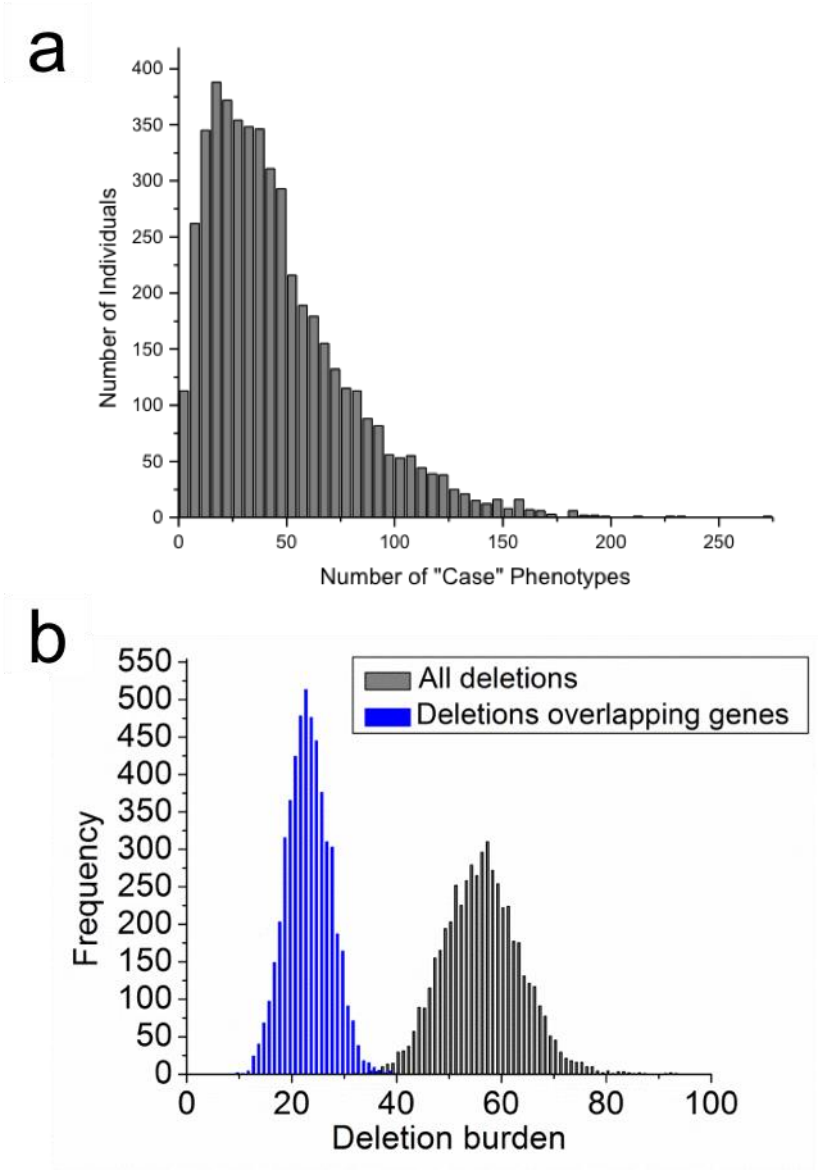


Figure 4-C. Distribution of a) PheWAS codes and b) deletions in individuals in the set.

We then compared deletions that overlapped genes with those that did not. While QQ plots show that p-values in deletions overlapping a gene did not substantially deviate from what was expected (Figure 4-D-a), and overall did not result in impressive p-values (Appendix K). By comparison, those that were not annotated as overlapping a gene visually appeared to have more significant p-values than what was expected (Figure 4-D-b). A ks test could not statistically determine that these two groups of p-values were drawn from different distributions ($p=0.076$) despite being visually distinct on QQ plots. QQ plots provide a mediocre comparison as we notice the inflation of a few specific points many of which are from a single deletion (Appendix L), while the lower left corner of the plot has far for data points in it. As deletions were assigned as overlapping a gene based on coordinates alone, we wanted to see if overlapping an exon made a difference in the best p-value resulting from testing each deletion with PheWAS. There was no significant difference ($p=0.26$) in the distribution of the best p-value of deletions overlapping an exon (Figure 4-D-c) and those annotated as overlapping a gene but not an exon (Figure 4-D-d). We also wanted to explore if the length of the deletion was correlated with the most significant p-value in PheWAS, but there was no correlation ($p= 0.69$).

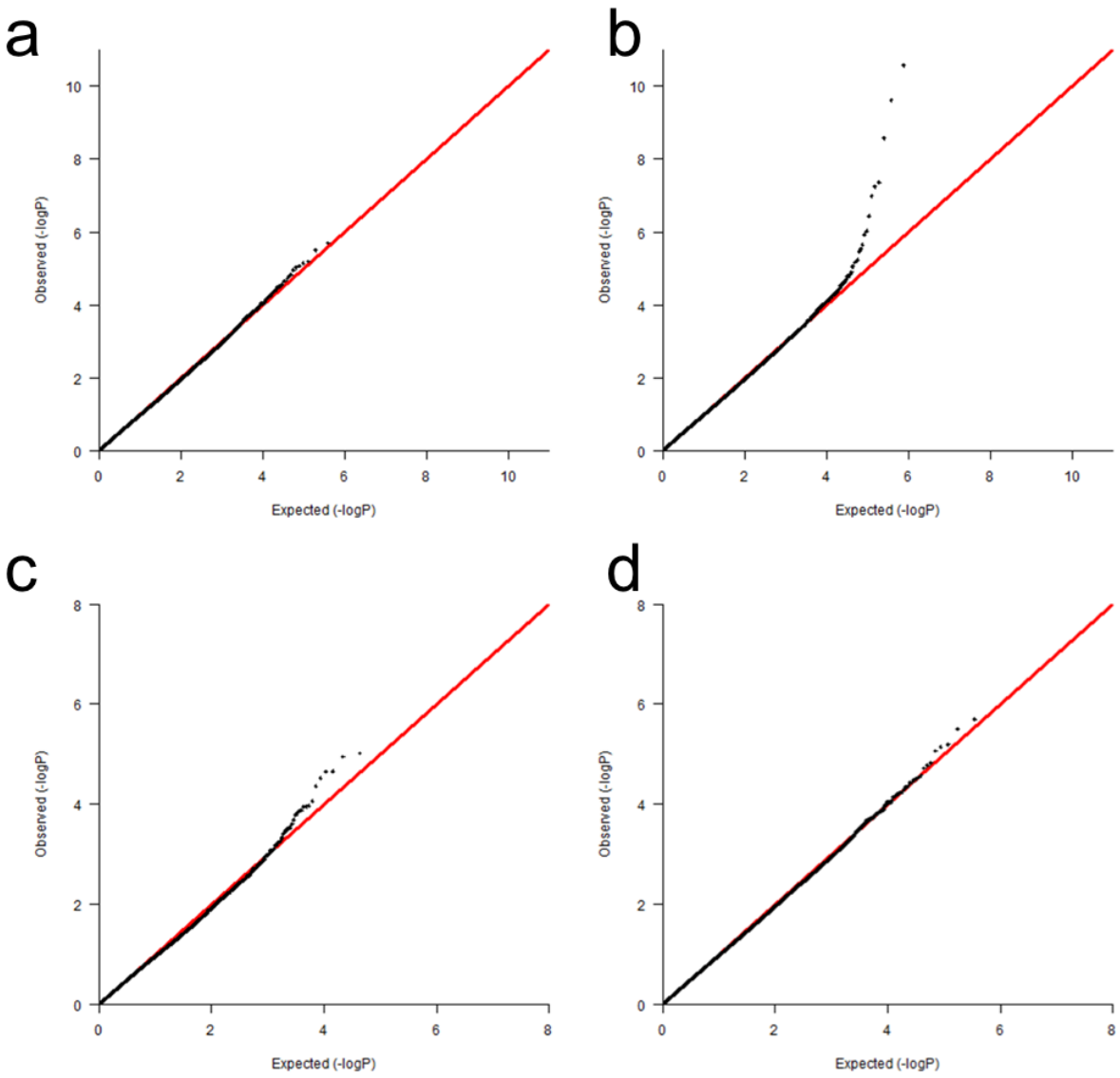


Figure 4-D. QQ Plots of deletions with different characteristics; a) deletions overlapping a gene, b)deletions not overlapping a gene, c) deletions overlapping a gene and an exon, d) deletions overlapping a gene but not an exon.

Discussion

Overall, we found that we could not successfully characterize deletions using PheWAS. The statistical significance of PheWAS hits did not seem to be different by whether a deletion overlapped a gene, exon or neither a gene nor an exon. If anything, deletions not annotated as overlapping a gene were the most significant in our analysis. Given that the majority of significant GWAS SNPs are in intergenic regions, this is not necessarily surprising. It is perhaps more surprising that deletions removing more of the genome are not more likely to give a phenotypic association with a lower p-value. One would assume that these deletions would be more likely to overlap something functional, whether coding or not. These analyses may have been limited by the correlation present in the phenotypic data of the PheWAS, and are certainly limited by the deletions we could reliably impute.

The length of deletions and its potential overlap has been determined by the outermost endpoints of the deletion and the gene. For many of the deletions there was one consensus start position and one consensus end position, but a small subset of the deletions had an "inner" and an "outer" start and end position listed. For these deletions with uncertain length we used the "outer" positions for both the start and the end. This may have artificially inflated the length of some of the deletions. It is unclear whether this had an impact on the PheWAS associations. For some deletions, the "outer" endpoint may have overlapped a gene where the "inner" endpoint would not have. To further complicate this, there may be other listed coordinates than the "outer" and "inner" that fall between the two, all of which may actually exist in individuals.

When we annotated a deletion as overlapping a gene, we did not mean that it had to touch the exonic region. As many as a single base pair of the deletion coordinates could overlap with the gene coordinates, and we would have counted it as overlapping. This is problematic, as a

deletion that overlaps five base pairs of a region annotated as a gene would be expected to have a different impact than a deletion that removes the entirety of a gene including all its exons. The exon overlap criteria helped with this, but some deletions were located in introns and could change the splicing of a gene making our reference coordinates inappropriate for the gene as it occurs. All these issues with overlap may have caused us to incorrectly classify deletions, reducing our ability to see an association.

RefSeq includes as “genes” many long non-coding RNAs and other transcripts with unknown function. One solution would be to limit our analysis to only protein coding genes. We could do this by using a second database such as Ensemb, to annotate protein-coding genes, and only use entries that are shared between the two databases. This may still leave us with some transcripts and RNAs, but at least we will know all "genes" are in standard locations. A third option is to manually curate the genes we use. Some of the entries are named as anti-sense transcripts or lnc RNAs. These should not have any entries in the RefSeq exon list, and we should be able to remove them. We could then re-annotate our deletions to ensure we are only including deletions that overlap protein coding genes.

Another potential issue is that we are using imputed deletions. While deletion imputation should be statistically the same as imputing SNPs, imputation of deletions is not widely published. A separate project in which we are participating is beginning to compare imputed and directly genotyped deletions to determine concordance. Additionally, calling the deletions from intensity data in the genotyping platforms may be a more appropriate way to capture them. Even deletions called from intensity data will be subject to a major limitation that is important for our study, we are limited by the SNPs on the genotyping platform. In this case, the distribution of SNPs will affect the deletions we capture as is the tagging nature of the GWAS platform. It is

quite possible that if we have ascertained all deletions throughout the human genome, our results would like quite different.

In conclusion, PheWAS does not seem an ideal way to classify imputed deletions in this study. This is not due to the nature of a deletion compared to other genetic variants; rather the difficulty arose from our use of PheWAS as a blanket classification scheme without considering how noise in both the genotypic and phenotypic data might complicate our efforts. Correlations in the PheWAS phenotypes complicated our ability to analyze results, as did the uneven distribution of our deletions across the genome and the different imputation quality present in different deletions. Perhaps given a completely random selection of deletions from all across the genome all called with the same error we would have been more successful. Despite the difficulties in this project, it is clear that analyzing a deletion instead of a SNP offers a far clearer biological hypothesis in the event of an association.

V. MITOCHONDRIAL HAPLOGROUP BACKGROUNDS MODIFY THE PHENOTYPIC IMPACT OF SNPS IN GENES RELEVANT TO MITOCHONDRIAL FUNCTION

Introduction

Mitochondria are double membrane bound organelles derived from an α-proteobacterial ancestor[143]. Mitochondria are dynamic, and one mitochondrion can split into two or multiple may merge into one, forming networks throughout the cell[144]. The outer membrane holds proteins necessary for intracellular signaling. The mitochondrial signaling pathways are involved in functions as diverse as regulation of metabolism and apoptosis. The inner mitochondrial membrane has folds called cristae that increase the surface area on which it can house the proteins in the electron transport chain. The electron transport chain facilitates oxidative phosphorylation to produce ATP in the cell[145]. Mitochondria have their own genome, but are still dependent on more than 1,000 genes encoded in the nucleus for proper function. Nuclear genes encode for proteins including Polγ, the mitochondrial polymerase, MAVS, the mitochondrial anti-viral signaling protein, and BCL2 which is important for apoptosis. In addition to many genes involved in mitochondrial signaling and immune response, nuclear encoded genes important for cytoskeletal formation and solute transport are essential for mitochondrial health. Polypeptides designated for the mitochondrial matrix have an N-terminal localization signal that directs them to pass through both membranes and into the mitochondria[146].

The mitochondria and its genome are maternally inherited[147]. The mitochondrial genome is comprised of a single circular chromosome. The majority of the mitochondrial DNA (mtDNA) is double stranded, except for a 1124 bp portion called the D-loop which is triple stranded and contains promoter and replication elements[148]. mtDNA encodes 37 genes; 13 proteins important

for oxidative phosphorylation, 22 tRNAs, and 2 rRNAs. Multiple copies of the genome can be found in each mitochondria, so a cell can contain many copies of the mitochondrial genome[144]. As mitochondria are haploid, mutations can accumulate at a much faster rate than they do in the nuclear genome[149]. Both mutations and population level variation are present in the mitochondrial genome (Figure 5-A).

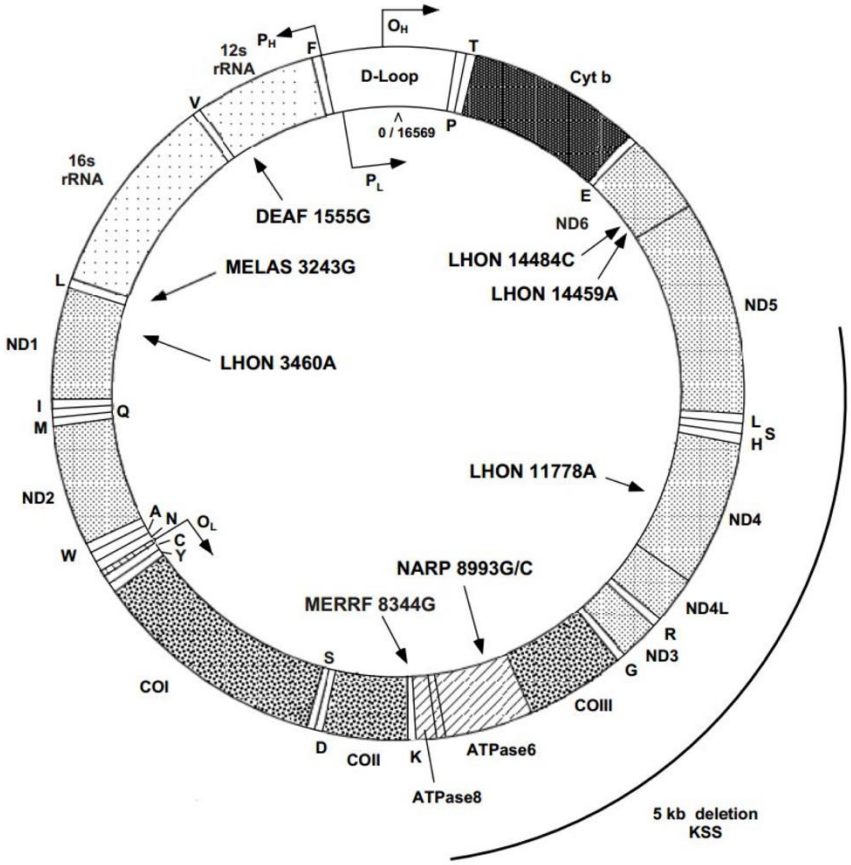


Figure 5-A. Map of the mitochondrial genome with genes and pathogenic mutations annotated. Figure was created by www.mitomap.org and is used under the creative commons license.

Population level variation in the mitochondrial genome can be phylogenetically grouped into haplogroups[150] (Figure 5-B). Mitochondrial haplogroups can be determined using SNPs commonly available on genotyping platforms. Different mt haplogroups function differently[151,152]. Haplogroups have been associated with numerous human diseases: haplogroup J has been associated with both susceptibility[153] to and protection[154,155] from Parkinson's Disease as well as susceptibility to Multiple Sclerosis[156,157], and haplogroup H has been associated with age related maculopathy[158].

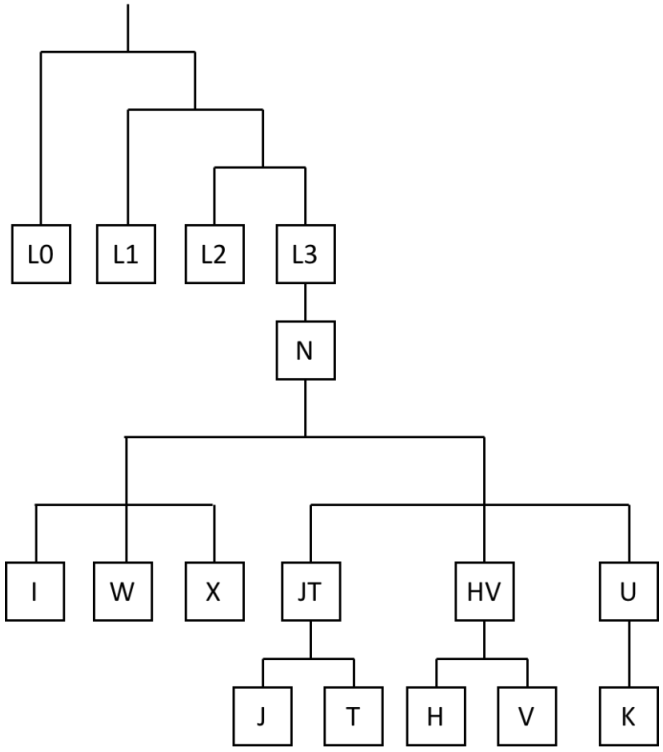


Figure 5-B. Mitochondrial Haplogroup tree showing relationship between mitochondrial haplogroups. European ancestry haplogroups include I, W, X, J, T, H, and Uk (from U and K combined).

Genetic variation in both the nucleus and mitochondria are responsible for mitochondrial health, but little is known about the relationship of the genetics in the two genomes and their collective impact on human disease. As mitochondrial diseases have a variety of phenotypes, we decided that a PheWAS would give us the opportunity to detect any possible outcome of nuclear and mitochondrial genetic variation. **We hypothesized that the phenotypic effect of nuclear SNPs relevant to mitochondrial function will be influenced by mitochondrial genetic variation.**

Methods

Genotypes and Haplogroup Determination

Nuclear and mtSNPs were obtained for individuals genotyped on the Human Exome Bead chip in BioVU. Heterozygous mtSNPs (due to either genotyping errors or potentially heteroplasmy) were set to missing for this analysis. Remaining mtDNA genotypes were classified into standard haplogroups using Haplogrep[159,160]. Fine level haplogroups were then assembled into analysis groups using racial information, since mitochondrial haplogroups are closely tied to continental ancestry. Individuals not of European descent by provider assigned race were removed from analysis.

Nuclear SNP genotypes underwent quality control. Briefly, SNPs with less than 95% genotyping efficiency were removed as were individuals with more than 5% missingness in the exome data. All nuclear SNPs with a minor allele frequency (MAF) over 10% present in genes on the MitoCarta2 Gene list[161], a list of genes involved in mitochondrial function, were extracted. The high MAF threshold was used since we would be further subdividing the analysis by mitochondrial haplogroups with population frequencies of ~10-50%. SNPs were additively encoded.

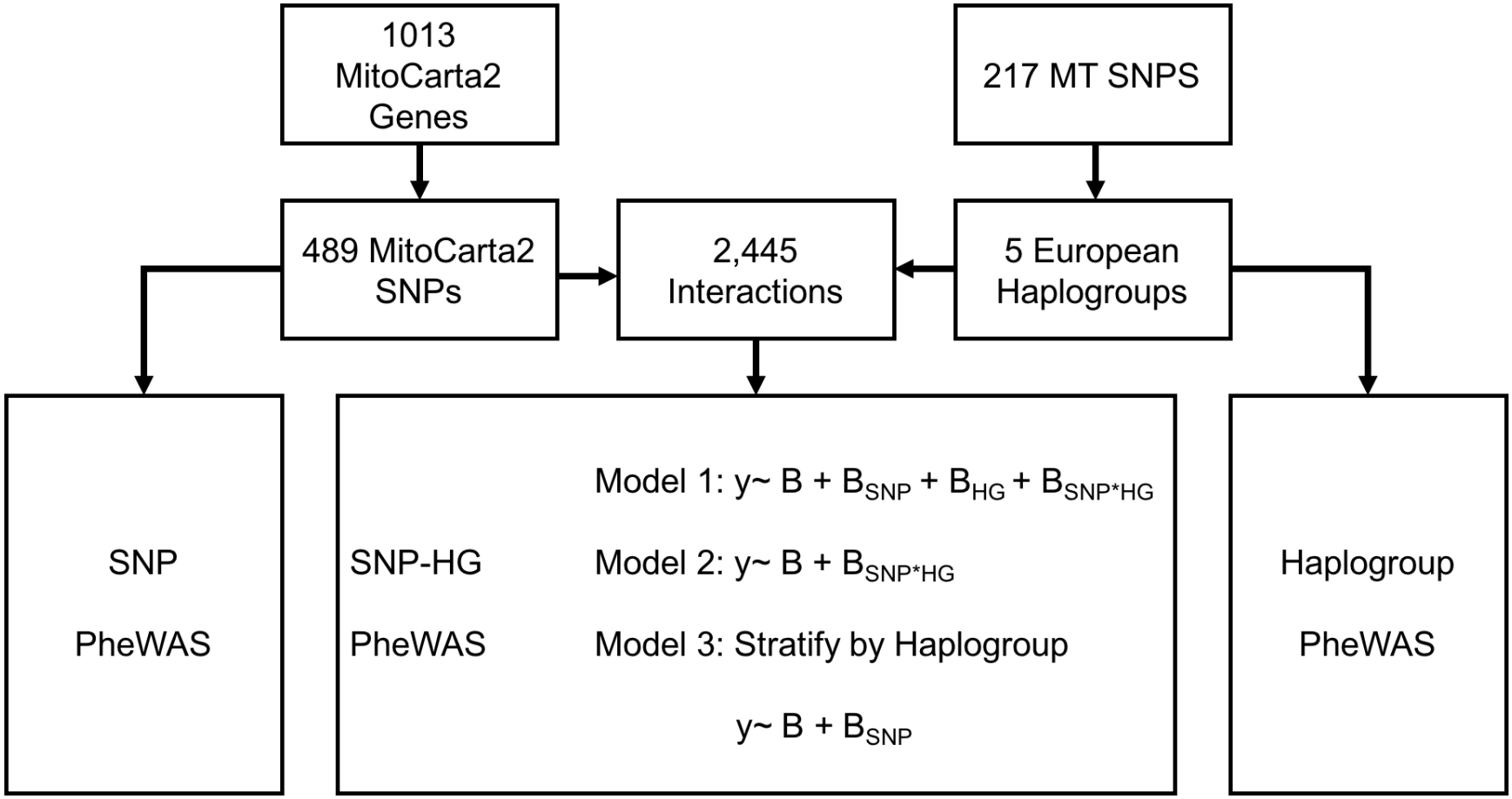


Figure 5-C. Flowchart of nuclear and mt SNP filtering and PheWAS performed on SNPs, haplogroups, and for SNP Haplogoup modification tests.

SNP-Haplogroup Regression Models

As limited studies have explored the relationship between nuclear and mitochondrial SNPs it is not clear exactly how to model the interaction or co-occurrence of the two. We decided to model the nuclear-mitochondrial SNP interaction in three ways (Figure 5-C). Model 1is based on the model traditionally used for SNP-SNP interactions. We will refer to this model as the "Main effects model" as it accounts for the main effects of the SNP and haplogroup in addition to the SNP-haplogroup effect, which is what we report here in our statistics. Model 2 models the nuclear-mito relationship without the individual effects of either the SNP or the haplogroup. We will call this the "mitochondrial haplogroup mediated SNP effect" model or "mtDNA mediated effect" model for short as we are testing individuals with both the SNP and haplogroup against those without either the SNP or the haplogroup. Model 3 stratifies by haplogroup and determines the effect of the SNP within each haplogroup. We will refer to this model as the "stratified model". All three models were run for all SNP-haplogroup combinations. To complement our models and allow us to look at all facets of the data, SNPs and haplogroups were each tested alone as well. Regression models were compared visually with QQplots and forest plots.

PheWAS

ICD-9 codes were obtained for all individuals. Duplicate entries, entries with corrupted dates, and non-numeric entries were removed. ICD-9 codes were aggregated to PheWAS codes. Demographics including sex, date of birth, and third-party assigned race, were obtained for all individuals in our set. Age at last ICD-9 code was calculated using the combined date of birth and ICD-9 information. Individuals with an age-at-last-ICD-9 code less than 18 or greater than

90 were removed. PheWAS was performed using each of the regression models and each variant from genes in the MitoCarta2 list, testing all phenotypes with more than 50 cases.

Result Filtering and Prioritization

PheWAS results were filtered in three major ways. First, we prioritized evaluation of two haplogroups, haplogroup H and haplogroup J. Haplogroup H is the most frequent of the European haplogroups (45% of the European descent population) and therefore would likely provide the best statistical power. Haplogroup J has the most existing disease associations, which provides us a base for which associations are biologically reasonable. The second filter, was statistical, and based on p-value, effect size, and the number of cases with the SNP-haplogroup combination. We predominantly viewed the results for each haplogroup separately, though we did compare the levels of significance of the top hits in different groups. Our last filter was an *a priori* prioritization of the disease categories in PheWAS. We thought that Phecodes in the Sensorineural, Neurological, and Musculoskeletal categories were most likely to be true associations if the Phecode reached the level of significance. While other disease categories were plausible, these would take priority if all else was the same.

Results

The population used to study the relationship between nuclear and mitochondrial genetics consisted of 20,064 adults of European descent. Haplogroups H, J, T, Uk, and haplogroup clade IWX were common enough to be evaluated; all other individuals were grouped into the category of Other. The haplogroup frequencies in our population were consistent with expected frequencies (Table 5-A). The median age at last code in our record was 65, and this metric was

Table 5-A. Demographic information for individuals of European descent used for our nuclear encoded mitochondria relevant SNP haplogroup analysis.

	All	**H**	**IWX**	**J**	**T**	**Uk**	**Other**
N (%)	20064	9419 (46.9)	1290 (6.4)	2100 (10.5)	2241 (11.2)	4548 (22.7)	466 (2.3)
N males (%)	9325 (45.5)	4325 (45.9)	586 (45.4)	981 (46.7)	1068 (47.7)	2157 (47.4)	208 (44.6)
Median age first ICD.9 code (IQR)	56 (43, 67)	56 (43, 67)	56 (42, 67)	56 (44, 67)	55 (42, 67)	56 (43, 67)	54 (40, 67)
Median age at last code ICD.9 (IQR)	65 (53, 77)	65 (53, 77)	65 (54, 77)	65 (54, 77)	65 (53, 77)	65 (53, 77)	63 (51, 76)
Median number of PheWAS codes (IQR)	24 (11, 43)	24 (11, 42)	24 (11, 45)	24 (12, 44)	23 (11, 43)	23 (11, 43)	23 (11,41)
Median number of ICD.9 code entries (IQR)	143 (68, 280)	143 (68, 279)	146 (69, 283)	147 (69, 282)	143 (68, 273)	141 (66, 282)	137 (70, 265)

consistent in all haplogroup subsets. The median number of PheWAS codes that individuals were cases for and the number of ICD.9 code entries were also consistent between haplogroups.

The PheWAS performed in mitochondrial haplogroup alone as a predictor (Figure 5-D-a) was much weaker than the signal we saw from SNPs in mitochondrial related genes (Figure 5-D-b). The PheWAS on the SNPs alone allowed us to see some known signals, including the association of SNPs in ARMS2 with age-related macular degeneration (OR [95% CI] ==3.06 [2.30, 4.07]; p=1.26E-14). The best signal in the haplogroup specific analysis was the Phecode for Intestinal Infection Due to C. difficile with haplogroup IWX (OR [95% CI] =2.00 [1.42, 2.81]; p= 6.4E-05). This was not quite below the Bonferroni correction for Haplogroup IWX (p=4.35E-5), and well below the correction when considering all haplogroups.

From our analysis of SNP- haplogroup co-occurence, one of the most consistent results was haplogroup J modifying the effect of rs3736032 in SLC25A37 on Other cerebral degenerations, which was the best signal seen in haplogroup J (Appendix M). This association was visible using all three of our models (Figure 5-E), though it was by far the strongest using the mtDNA mediated effect model (OR [95% CI]= 4.96[3.10, 7.92]., p=2.-3e-11 We saw a moderate effect of haplogroup J alone on the Phecode for Other Cerebral Degenerations (1.81 [1.32, 2.47], p=2e-04), but no signal from the SNP alone (1.27 [0.95, 1.70], p=0.12). In the main effects model, Other Cerebral Degenerations was still the best signal we saw (4.38[2.25, 8.55]; p=1.48e-05). In our stratified model, the Other Cerebral Degenerations signal was significant in haplogroup J (4.00[2.28, 7.01], p=1.28e-06), but not in the not-J group or individually in any of the other haplogroups (Figure 5-F). None of the other haplogroups had any significant signal at all. Other Phecodes with signals beyond suggestive significance thresholds for J and rs3736032 were also related to central nervous system complications.

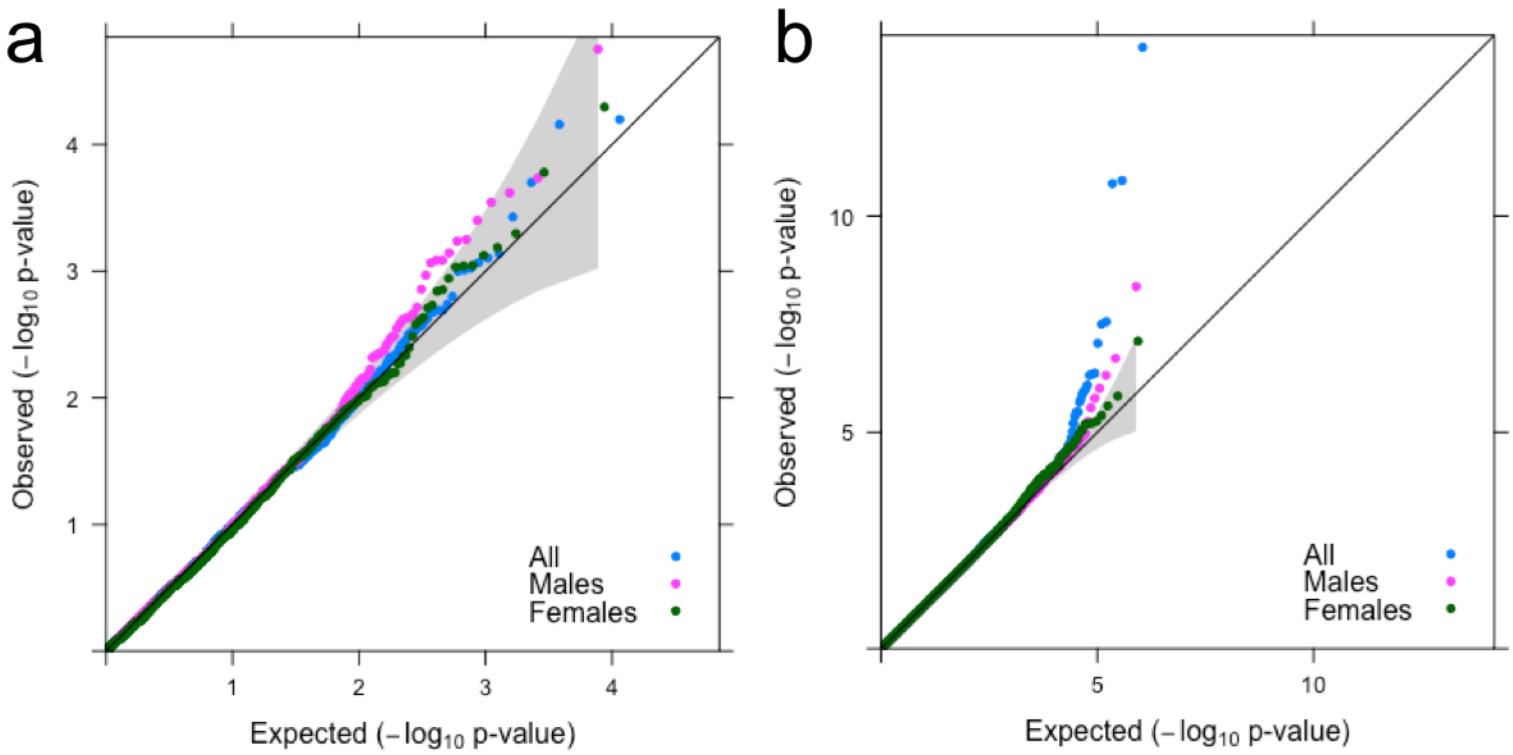


Figure 5-D. QQ plots of p-values from a) PheWAS using mitochondrial haplogroups as predictors and b) SNPs in mitochondria relevant genes as predictors. All individuals, males, and females are shown.

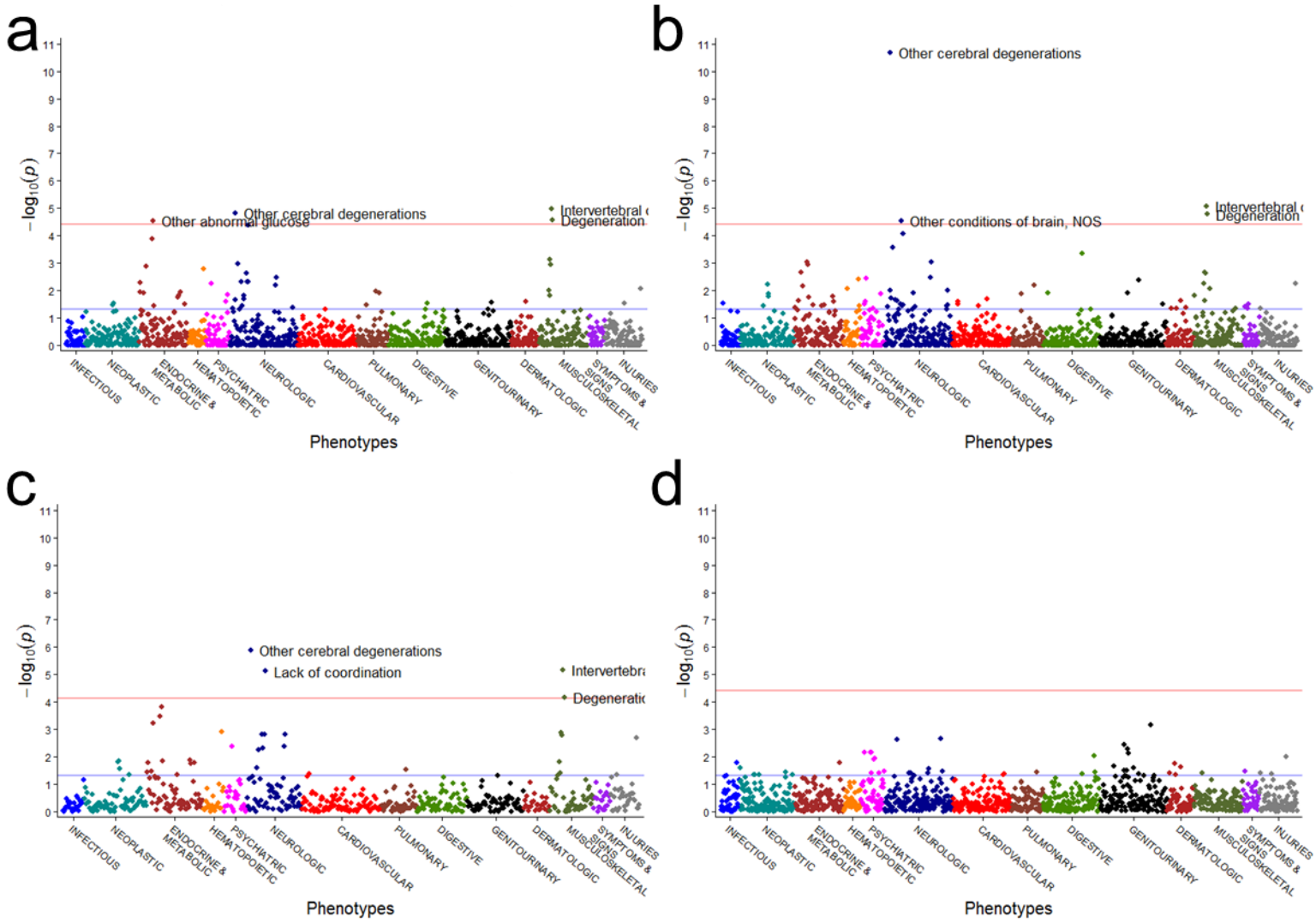


Figure 5-E. PheWAS Manhattan plots for the association of rs3736032 on Other Cerebral Degenerations modified by haplogroup J in a) the main effects model, b) the mtDNA mediated effect model, and the stratified model c) in haplogroup J individuals and d) in not J individuals.

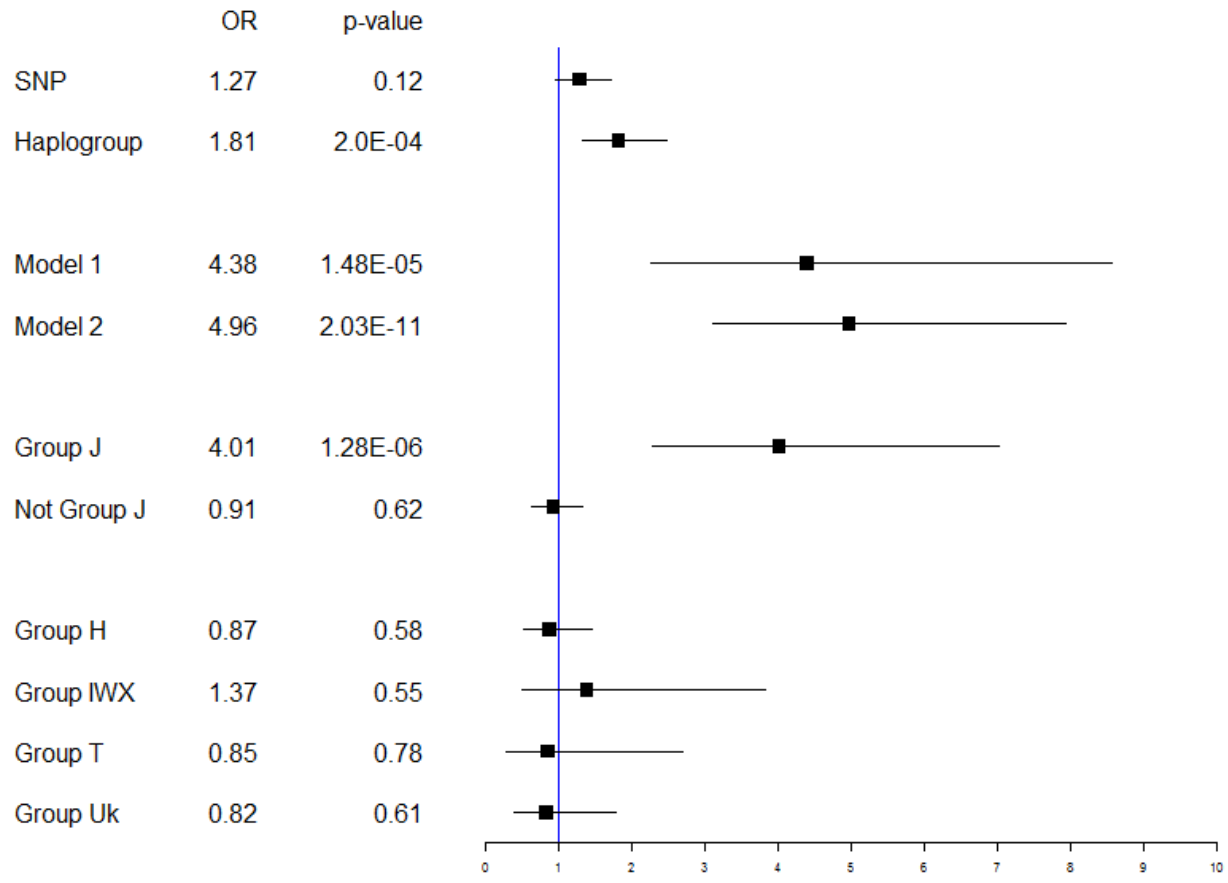


Figure 5-F. Forest plot of one of the most consistent results, haplogroup J modifying the effect of rs3736032 and haplogroup J on Other cerebral degenerations. Effects of SNP, haplogroup, and all three models are shown. The association of rs3736032 with Other cerebral degenerations is also shown in each non-J haplogroup for completeness.

A more unexpected association with a different pattern of expression was the association of rs17850652 with Tobacco Use Disorder modified by haplogroup H. This was not the most significant hit in SNPs tested for modification with group H (Appendix N), but was an interesting phenotype and was fairly consistent across models. While the mtDNA mediated effect model was still the most significant of the models tested (1.53[1.29, 1.83], p=1.12e-06) (Figure 5-G), the stratified analysis (1.54[1.29, 1.86], p=3.21e-06) looked more like the mtDNA mediated effect model, rather than the main effects model (1.68 [1.28, 2.21], p=1.7e-04). In this case, the SNP has a small effect alone (1.19 [1.04, 1.36], p=0.01), while the haplogroup alone is not significant (1.07 [0.96, 1.19], p=0.22). Despite the difference in p-values, all three models provided similar odds ratio estimates (Figure 5-H). No other Phecodes were significant in the analysis of this SNP-haplogroup combination.

As the examples we specifically investigated showed similar odds ratios despite different p-values, we wanted to know if this extended to all results. For all SNP-haplogroup-phenotype combinations present in both models with a p-value for the interaction term less than 0.01, we plotted the betas from the main effects model against those from the mtDNA mediated effect model. We found that the betas from the main effects models and the mtDNA mediated effect model were highly correlated ($p<2e-16$) (Figure 5-I). While the p-values from the mtDNA mediated effect model tended to be stronger, the betas from the main effects model are slightly stronger. The same trend also holds true even when no p-value threshold is placed on the SNP-haplogroup interaction term.

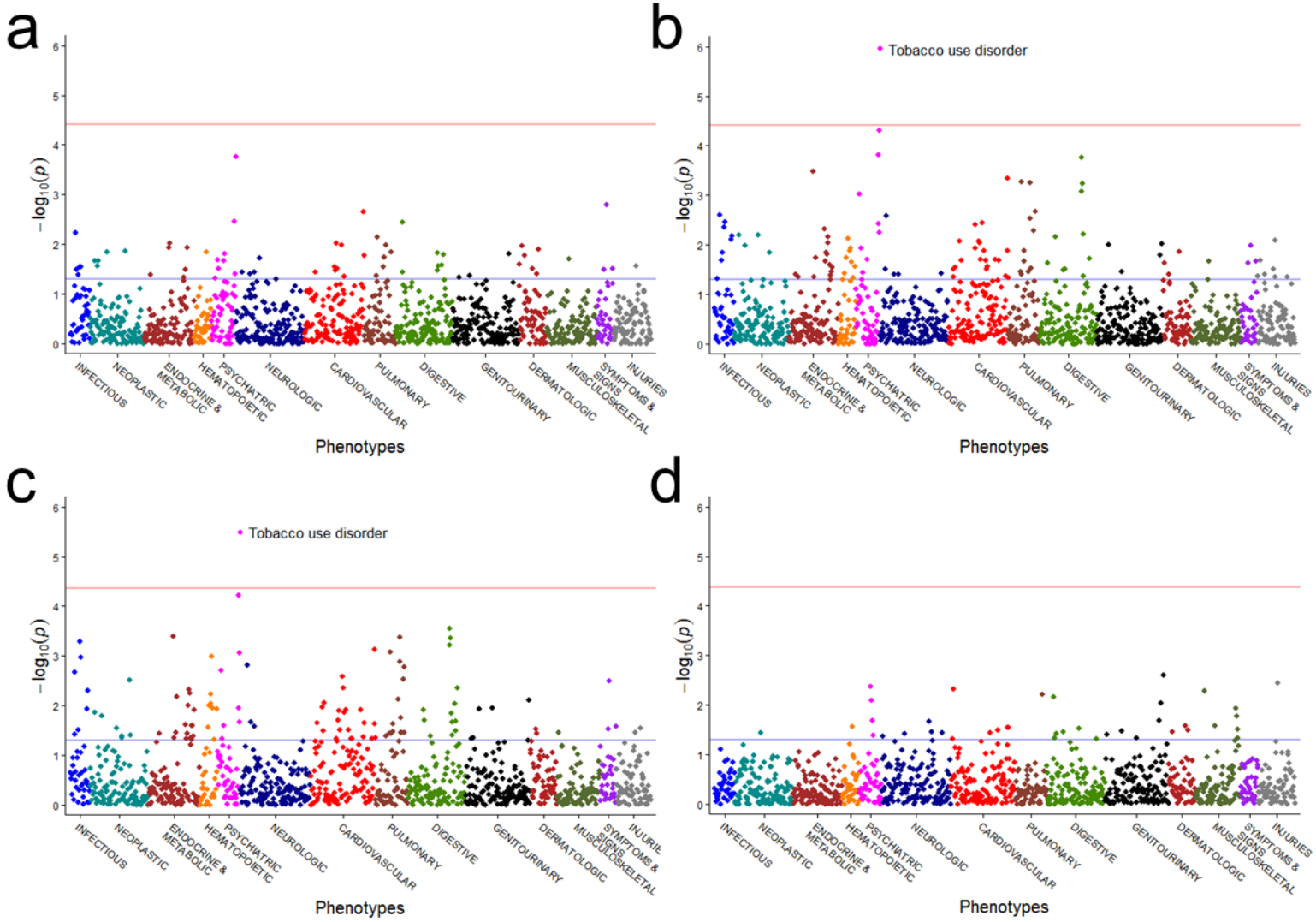


Figure 5-G. PheWAS Manhattans for the association of rs17850652 on Tobacco Use Disorder modified by haplogroup H in a) the main effects model, b) the mtDNA mediated effect model, and the stratified model c) in haplogroup H individuals and d) in not H individuals.

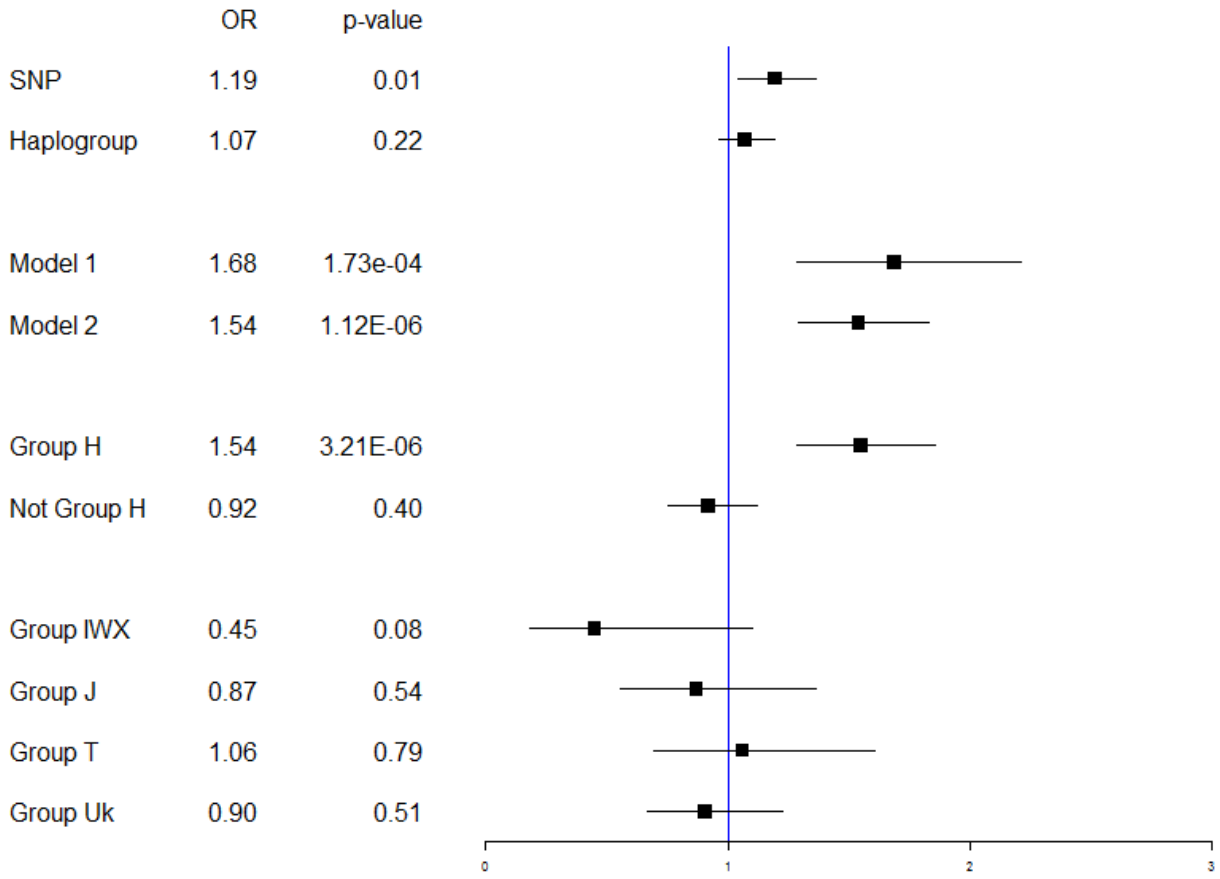


Figure 5-H. Forest plot of haplogroup H modifying the effect of rs17850652 on Tobacco Use Disorders. Effects of SNP, haplogroup, and all three models are shown. The association of rs17850652 on Tobacco Use Disorders in other European haplogroups is also shown.

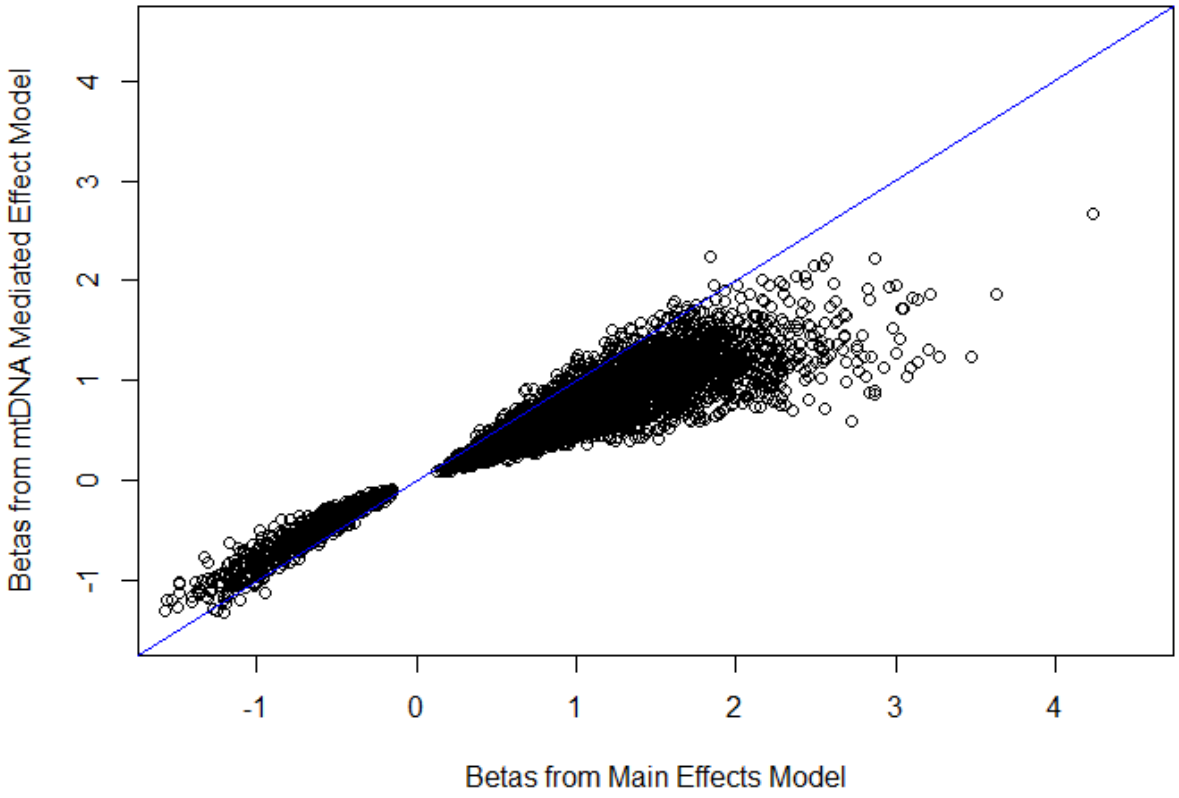


Figure 5-I. Correlation of betas from Main Effects model and mtDNA Mediated Effect model.

Discussion

We were able to see mitochondrial haplogroup background influence the effect of SNPs resulting in phenotypes relevant to mitochondrial function. The modification of association we are most confident about is the influence of rs3736032 on Other Cerebral Degenerations in haplogroup J. Haplogroup J has previously been associated with Parkinson's Disease and Multiple Sclerosis[154,156], so the cerebral degeneration related phenotype is reasonable. rs3736032 causes a Arg96Gln amino acid change in SLC25A37, which encodes an iron transporter on the inner mitochondrial membrane. This change is predicted to be tolerated by both SIFT and PolyPhen, though some transcripts are predicted to undergo nonsense mediated decay when this variant is present. Rs3736032 occurs with a MAF of 15%, so it is relatively common. Variants in SLC25A37 have been linked to iron imbalances[162,163]. Iron dysregulation occurs in Parkinson's disease[164] and has been related to other brain diseases[165]. Outside of neurodegenerative diseases, this SNP has been evaluated for gene environment interactions between iron intake and type 2 diabetes[166], but was not found to be significant.

Since haplogroup J has been previously associated with similar phenotypes, it is not clear if the effects reported in the literature are due independently to haplogroup J, are caused by group J individuals with this SNP, or some combination of the two. It is likely that both J and the SNP together exaggerate the effect, but J alone causes enough of an effect to see in targeted association studies. In our PheWAS (Appendix M), haplogroup J appeared to modify the effect of a number of different SNPs on this phenotype, but all three SNPs that were near the top of our results were in different genes on different chromosomes and appear to have different functions. Further exploration will be necessary to untangle this relationship.

While the association of rs17850652 and Tobacco Use Disorder modified by haplogroup H, was unexpected, it is not completely unreasonable. rs17850652 is a global 10% MAF variant that causes a Lys291Arg change in RARS2. RARS2 is the Mitochondrial Arginyl-TRNA Synthetase 2. This change is thought to be benign. As RARS2 is involved in translation of mitochondrial encoded proteins, any number of defects could result from a small change in protein functionality, but most likely oxidative phosphorylation would be altered slightly. Mutations in RARS2 have been associated with Pontocerebellar hypoplasia, a rare neurodegenerative disorder[167]. We cannot provide a specific mechanism for how haplogroup might affect Tobacco Use Disorder, though there is a small literature on smoking and mitochondria. We initially thought that this phenotype might be indicative of an underling predisposition to addiction, but no other signals for addiction are present in our PheWAS. Despite this, manual review of the records of patients with this PheWAS code indicate that they are smokers, consistent with reports about the reliability of smoking status in ICD-9 codes[168].

We used three different models to test whether mitochondrial haplogroups modify the effects of SNPs involved in mitochondrial function. As there are no known positive associations for mitochondrial modification of SNP effects, we could not evaluate our models against a known standard, so we instead compared them to each other. The similarity of effect in the three models is highly dependent on the specific SNP, haplogroup, and phenotype tested. Much of the time, the three models gave similar effect sizes for the SNP or SNP-haplogroup co-occurence, but not always. One of the things that seemed to effect this was how each of the haplogroups contributed to the signal. When testing a SNP, we limit our analysis to biallelic SNPs. For haplogroups, individuals not of the haplogroup being tested can be from any of the 4 other common European haplogroups or can be grouped into the "other" category. So not-H

individuals are not all the same, but belong to one of the other haplogroups. We envisioned two scenarios; one in which one haplogroup showed an effect and the others all showed no effect or the same opposite effect, and one where two haplogroup show effects in opposite directions and the rest have no effect. More specific examples will have to be extracted and interrogated to understand the variety of effects we will and will not be able to see with each of these models. While we tested traditional interaction models, we were worried that the models accounting for the main effects of both the SNP, haplogroup, and the SNP and haplogroup co-occurring together might sometimes cancel out the effect in the scenario where different haplogroups have different effects.

The main effects model accounts for the disease prevalence in individuals with the SNP-haplogroup combination, just the SNP, or just the haplogroup simultaneously. The mt mediated effect model compares disease prevalence in individuals with the SNP and haplogroup to those without the SNP-haplogroup combination. This is more similar to the mt haplogroup analysis methods comparing the haplogroup to the not-haplogroup population. Testing the co-occurrence of the SNP-haplogroup without the main effects of either might allow us to best evaluate scenarios where the SNP and haplogroup have no effect except together. We can still detect a signal when they each do have an effect, but then we are not properly accounting for it, and may inadvertently inflate the association of the combined occurrence. Our third model, stratifying by haplogroup and testing the SNP of interest in that group causes us to lose power, but provides the most intuitive and easily interpretable model.

An interesting side point from this analysis was the general lack of signal from haplogroups. Nothing we saw using any of the major European haplogroups as a predictors passed the Bonferroni correction for a single PheWAS. This was not completely unexpected as

mitochondrial haplogroups tend to have far more moderate significant effects than many SNPs, but we had assumed that as we were testing a wide variety of phenotypes, mitochondrial haplogroups would be significantly associated with at least one.

A major issue we encountered in this study was sample size. While haplogroup H seemed large enough to allow us sufficient power to see at least some associations, for lower frequency haplogroups we had limited power. Perhaps the only reason we were able to detect any signal in haplogroup J was the strong effect. In retrospect, combining haplogroups J and T may have allowed us more power to see an association, and as they are in the same clade, would be biologically reasonable. A higher cut off for the number of cases necessary for PheWAS to be performed would also have been helpful. We used a minimum of 50 cases, but perhaps scaling this number based on the frequency of the SNP-haplogroup combination would have been better.

Future directions include trying to replicate the results we have seen. We have begun looking within BioVU to validate the associations we have seen with more carefully phenotyped data. We would also like to replicate in an external dataset. A dataset used for GWAS, where a large number of individuals have been carefully phenotyped and have genetic data available would be ideal. We are currently evaluating the hits we think are biologically reasonable to find those that have Phecode associations that have existing GWAS proxies. As we were limited to Europeans by the nature of haplogroups, it would also be interesting to repeat this analysis in other population groups.

In vitro studies evaluating the impact of specific SNPs we think might be modified by mitochondrial haplogroup would be greatly helpful. Limited information is available for many of the genes and SNPs we have tested. In vitro studies to further explore the function of these genes and SNPs would help us evaluate our results for follow-up and also provide biological validation

for signals we think are real. Trying to decide if a SNP is reasonable is far more difficult when the gene has no annotated function. Additionally, for some of the genes on the MitoCarta2 list, it is not obvious how they might be relevant to mitochondrial function. The breadth of mitochondrial involvement in cell function provides us room to speculate in these cases, but our ability to draw biologically relevant connections to phenotypes is more complicated for these genes.

In conclusion, PheWAS seems to be an interesting and adequate way to evaluate the potential of mitochondrial haplogroup to modify SNP effects. We saw what we think are biologically reasonable associations between the SNP, haplogroup, and phenotype. More stringent evaluation of the models will be necessary to determine which is best, though it may be that different models are ideal depending on the hypothesis. Further exploration of mechanisms to correct for multiple testing in this scenario would assist us in prioritizing the signals we saw and guide efforts for statistical or biological validation.

CONCLUSION

My dissertation explores a variety of scenarios under which PheWAS is a reasonable technique to implement. The projects presented here begin with the most straight-forward scenario, directly genotyped single SNPs, and progress to imputed deletions before exploring ways to use PheWAS in multi-dimensional studies.

One of the things that has struck me during my time using this technique is how different the follow-up to a PheWAS can be depending on the goal of the study, the existing knowledge of the variant, and the expectations of those involved in the work. The majority of my studies immediately follow-up a PheWAS by moving down one level of aggregation in the data and focusing on the ICD-9 level data. Sometimes a Phecode is made up of many ICD9 codes and this provides no clarity, but occasionally looking at the ICD9 code data allows one to shift focus to something more specific. Chapter 2 of the work presented here is a great example of the latter, where focusing on ICD9 level data directed the rest of the follow-up we performed. Laboratory test data and CPT code data are two other data sources I have used for follow-up. Of these, CPT code data is by far the easier. Laboratory test data can be some of the most useful data in reassuring oneself of a PheWAS association, but can also be extremely complicated to quality control. I have encountered tests where the interpretation of results is highly dependent on the age of the individual at the time of the test. Tests often have unexpected values, such as a few instances when someone enters <30, while the majority of the entries are integers, or text entries indicated contaminated specimens will be mixed in with numeric measures. These problems have sometimes forced me to adapt my analysis plan to unexpected patterns in the data. Despite

these other options, manual review is the best mechanism to be sure that the PheWAS association you see truly represents a disorder in the data.

I feel that I did not really begin to appreciate all the nuance of the PheWAS method until I began to do higher dimensional analyses with the data and to think about how to compare different PheWAS outcomes to each other. I have found it difficult to identify how much of the correlation between codes in outcome is solely correlation in the ICD-9 codes used for aggregation and how much is actual or potential biology of the predictor tested. Studies using only SNPs of similar allele frequencies might be helpful to determine how much of the variability in PheWAS result is only due to allele frequency and how much is due to other factors.

One interesting connection between almost all the analyses performed was that age or sex or both were often stronger predictors than any of the PheWAS codes themselves. While we tried to limit the impact of this by stratifying data appropriately, we were not always successful, especially when age played an important role. An interesting illustration of how insignificant genetic predictors appear to be overall can be seen by performing a PheWAS using sex as the primary predictor (adjusting for median age over record). The best p-value signal we saw from a genetic predictor in any of our analyses was ~1e-14 from a SNP in *ARMS2* with age-related macular degeneration. Sex as a predictor, by contrast, results in many p-values less than 10e-30 (Figure 6-A). While some of this is surely due to power, it still clearly realigns our expectations about what we should see from genetic predictors.

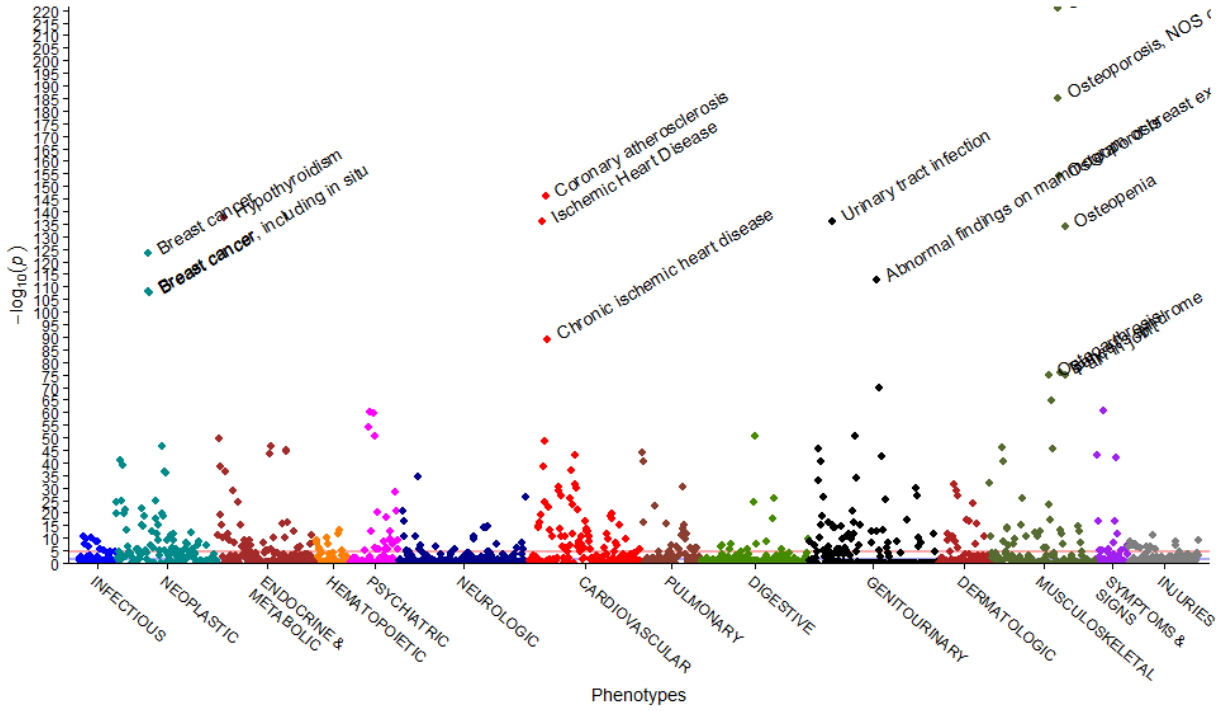


Figure 6-A. PheWAS manhattan plot using sex as the primary predictor. European descent individuals with genotyping on the Exome chip were used. Adjusted for median age over record.

As we chose each of the types of genetic variation and many specific genetic variants for analysis, one interesting dilemma was presented to us that we may not have noticed if we had just analyzed everything. How does one proceed when a variant that is known to have disease associations does not result in any signal in PheWAS? In our case, as we were primarily analyzing variants one at a time, we were able to explore the silent phenotypes in our PheWAS, those that we expected but did not see. We often went back to refine the population to optimize our chances of seeing the association. In cases where we were analyzing many deletions or SNPs, we never did this. There may be interesting reasons that we see some but not all known associations represented in our PheWAS outcomes and without detailed exploration of both the genotypic and phenotypic specifics in the dataset we would miss them. This does leave us with the question of whether we should value any of the PheWAS if we cannot see what we might

expect to use as positive controls. The potential development of agnostic methods for prioritizing results from multiplexed PheWAS studies provides an interesting thought experiment, though I remain unconvinced that any simple method will sufficiently do so. In our analyses, using p-values provided a preliminary, if occasionally insufficient, filter. Further filtering of the patterns of betas and the numbers of cases with the genetic predictor will most certainly be necessary to blindly evaluate many SNPs.

In conclusion, while I have found PheWAS to be a useful method for interrogating genotype-phenotype relationships, I think the merit of the approach lies in how it may direct you for future studies of the SNP, protein, or gene. While PheWAS can also be a useful tool for steering the learning about underlying biology of a genetic relationship, using it without any knowledge of the predictor makes it difficult to know how to interpret the results appropriately.

REFERENCES

1. Collins FS, Green ED, Guttmacher AE, Guyer MS. A vision for the future of genomics research. *Nature*. 2003;422(6934):835-847. doi:10.1038/nature01626.

2. Risch N, Merikangas K. The Future of Genetic Studies of Complex Human Diseases. *Science*. 1996;273(5281):1516-1517. doi:10.1126/science.273.5281.1516.

3. Edwards AO, Ritter R, Abel KJ, Manning A, Panhuysen C, Farrer LA. Complement Factor H Polymorphism and Age-Related Macular Degeneration. *Science*. 2005;308(5720):421-424.

4. Haines JL, Hauser MA, Schmidt S, et al. Complement Factor H Variant Increases the Risk of Age-Related Macular Degeneration. *Science*. 2005;308(5720):419-421.

5. Klein RJ, Zeiss C, Chew EY, et al. Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*. 2005;308(5720):385-389.

6. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747-753. doi:10.1038/nature08494.

7. Bowton E, Field JR, Wang S, et al. Biobanks and Electronic Medical Records: Enabling Cost-Effective Research. *Sci Transl Med*. 2014;6(234).

8. Jha AK, DesRoches CM, Campbell EG, et al. Use of Electronic Health Records in U.S. Hospitals. *N Engl J Med*. 2009;360(16):1628-1638. doi:10.1056/NEJMsa0900592.

9. Charles, D., Gabriel, M., Furukawa MF. Adoption of Electronic Health Record Systems among U . S . Non -federal Acute Care Hospitals : 2008-2013. *Heal Inf Technol*. 2014;2008(16):2-7.

10. Denny JC, Hindorff L, Sethupathy P, et al. Chapter 13: Mining Electronic Health Records in the Genomics Era. Lewitter F, Kann M, eds. *PLoS Comput Biol*. 2012;8(12):e1002823. doi:10.1371/journal.pcbi.1002823.

11. Thorwarth WT. CPT®: An Open System That Describes All That You Do. *J Am Coll Radiol*. 2008;5(4):555-560. doi:10.1016/j.jacr.2007.10.004.

12. Ritchie MD, Denny JC, Crawford DC, et al. Robust Replication of Genotype-Phenotype Associations across Multiple Diseases in an Electronic Medical Record. *Am J Hum Genet*. 2010;86(4):560-572. doi:10.1016/j.ajhg.2010.03.003.

13. Kho AN, Pacheco JA, Peissig PL, et al. Electronic Medical Records for Genetic Research: Results of the eMERGE Consortium. *Sci Transl Med*. 2011;3(79).

14. Kurreeman F, Liao K, Chibnik L, et al. Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. *Am J Hum Genet*. 2011;88(1):57-69. doi:10.1016/j.ajhg.2010.12.007.

15. Roden D, Pulley J, Basford M, et al. Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. *Clin Pharmacol Ther*. 2008;84(3):362-369. doi:10.1038/clpt.2008.89.

16. Pulley J, Clayton E, Bernard GR, Roden DM, Masys DR. Principles of human subjects protections applied in an opt-out, de-identified biobank. *Clin Transl Sci*. 2010;3(1):42-48. doi:10.1111/j.1752-8062.2010.00175.x.

17. Pendergrass SA, Ritchie MD. Phenome-Wide Association Studies: Leveraging Comprehensive Phenotypic and Genotypic Data for Discovery. *Curr Genet Med Rep*. 2015;3(2):92-100. doi:10.1007/s40142-015-0067-9.

18. Denny JC, Bastarache L, Roden DM. Phenome-Wide Association Studies as a Tool to Advance Precision Medicine. *Annu Rev Genomics Hum Genet*. 2016;17(1):353-373. doi:10.1146/annurev-genom-090314-024956.

19. Bush WS, Oetjens MT, Crawford DC. Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nat Rev Genet*. 2016;17(3):129-145. doi:10.1038/nrg.2015.36.

20. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*. 2010;26(9):1205-1210. doi:10.1093/bioinformatics/btq126.

21. Denny JC, Bastarache L, Ritchie MD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol*. 2013;31(12):1102-1110. doi:10.1038/nbt.2749.

22. Stearns FW. One hundred years of pleiotropy: a retrospective. *Genetics*. 2010;186(3):767-773. doi:10.1534/genetics.110.122549.

23. Visscher PM, Yang J. A plethora of pleiotropy across complex traits. *Nat Genet*. 2016;48(370):40133-40141. doi:10.1038/ng.3604.

24. Hebbring SJ. The challenges, advantages and future of phenome-wide association studies. *Immunology*. 2014;141(2):157-165. doi:10.1111/imm.12195.

25. Ghoussaini M, Song H, Koessler T, et al. Multiple loci with different cancer specificities within the 8q24 gene desert. *J Natl Cancer Inst*. 2008;100(13):962-966. doi:10.1093/jnci/djn190.

26. Fehringer G, Kraft P, Pharoah PD, et al. Cross-Cancer Genome-Wide Analysis of Lung,

Ovary, Breast, Prostate, and Colorectal Cancer Reveals Novel Pleiotropic Associations. *Cancer Res*. 2016;76(17):5103-5114. doi:10.1158/0008-5472.CAN-15-2980.

27. Pickrell JK, Berisa T, Liu JZ, et al. Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet*. 2016;48(7):709-717. doi:10.1038/ng.3570.

28. Bylund DB, Eikenberg DC, Hieble JP, et al. International Union of Pharmacology nomenclature of adrenoceptors. *Pharmacol Rev*. 1994;46(2):121-136.

29. Aziz I, Hall IP, McFarlane LC, Lipworth BJ. Beta2-adrenoceptor regulation and bronchodilator sensitivity after regular treatment with formoterol in subjects with stable asthma. *J Allergy Clin Immunol*. 1998;101(3):337-341. doi:10.1016/S0091-6749(98)70245-3.

30. Ahles A, Engelhardt S. Polymorphic variants of adrenoceptors: pharmacology, physiology, and role in disease. *Pharmacol Rev*. 2014;66(3):598-637. doi:10.1124/pr.113.008219.

31. Dewar JC, Wheatley AP, Venn A, Morrison JF, Britton J, Hall IP. Beta2-adrenoceptor polymorphisms are in linkage disequilibrium, but are not associated with asthma in an adult population. *Clin Exp Allergy*. 1998;28(4):442-448.

32. Kirstein SL, Insel PA. Autonomic nervous system pharmacogenomics: a progress report. *Pharmacol Rev*. 2004;56(1):31-52. doi:10.1124/pr.56.1.2.

33. Dishy V, Landau R, Sofowora GG, et al. Beta2-adrenoceptor Thr164Ile polymorphism is associated with markedly decreased vasodilator and increased vasoconstrictor sensitivity in vivo. *Pharmacogenetics*. 2004;14(8):517-522.

34. Warne T, Moukhametzianov R, Baker JG, et al. The structural basis for agonist and partial agonist action on a beta(1)-adrenergic receptor. *Nature*. 2011;469(7329):241-244. doi:10.1038/nature09746.

35. Green SA, Cole G, Jacinto M, Innis M, Liggett SB. A polymorphism of the human beta 2-adrenergic receptor within the fourth transmembrane domain alters ligand binding and functional properties of the receptor. *J Biol Chem*. 1993;268(31):23116-23121.

36. Turki J, Lorenz JN, Green SA, Donnelly ET, Jacinto M, Liggett SB. Myocardial signaling defects and impaired cardiac function of a human beta 2-adrenergic receptor polymorphism expressed in transgenic mice. *Proc Natl Acad Sci U S A*. 1996;93(19):10483-10488.

37. Bruck H, Leineweber K, Ulrich A, et al. Thr164Ile polymorphism of the human beta2-adrenoceptor exhibits blunted desensitization of cardiac functional responses in vivo. *Am J Physiol Hear Circ Physiol*. 2003;285(5):H2034-8. doi:10.1152/ajpheart.00324.2003.

38. Tomaszewski M, Brain NJ, Charchar FJ, et al. Essential hypertension and beta2-adrenergic receptor gene: linkage and association analysis. *Hypertension*. 2002;40(3):286-291. http://www.ncbi.nlm.nih.gov/pubmed/12215468.

39. Iaccarino G, Lanni F, Cipolletta E, et al. The Glu27 allele of the beta2 adrenergic receptor increases the risk of cardiac hypertrophy in hypertension. *J Hypertens*. 2004;22(11):2117-2122.

40. Sethi AA, Tybjaerg-Hansen A, Jensen GB, Nordestgaard BG. 164Ile allele in the beta2-Adrenergic receptor gene is associated with risk of elevated blood pressure in women. The Copenhagen City Heart Study. *Pharmacogenet Genomics*. 2005;15(9):633-645.

41. Thomsen M, Dahl M, Tybjaerg-Hansen A, Nordestgaard BG. beta2 -adrenergic receptor Thr164IIe polymorphism, blood pressure and ischaemic heart disease in 66 750 individuals. *J Intern Med*. 2012;271(3):305-314. doi:10.1111/j.1365-2796.2011.02447.x.

42. Ortega VE, Hawkins GA, Moore WC, et al. Effect of rare variants in ADRB2 on risk of severe exacerbations and symptom control during longacting beta agonist treatment in a multiethnic asthma population: a genetic study. *Lancet Respir Med*. 2014;2(3):204-213. doi:10.1016/S2213-2600(13)70289-3.

43. Thomsen M, Nordestgaard BG, Sethi AA, Tybjaerg-Hansen A, Dahl M. beta2-adrenergic receptor polymorphisms, asthma and COPD: two large population-based studies. *Eur Respir J*. 2012;39(3):558-566. doi:10.1183/09031936.00023511.

44. Turki J, Pak J, Green SA, Martin RJ, Liggett SB. Genetic polymorphisms of the beta 2-adrenergic receptor in nocturnal and nonnocturnal asthma. Evidence that Gly16 correlates with the nocturnal phenotype. *J Clin Invest*. 1995;95(4):1635-1641. doi:10.1172/JCI117838.

45. Reihsaus E, Innis M, MacIntyre N, Liggett SB. Mutations in the gene encoding for the beta 2-adrenergic receptor in normal and asthmatic subjects. *Am J Respir Cell Mol Biol*. 1993;8(3):334-339. doi:10.1165/ajrcmb/8.3.334.

46. Liggett SB, Wagoner LE, Craft LL, et al. The Ile164 beta2-adrenergic receptor polymorphism adversely affects the outcome of congestive heart failure. *J Clin Invest*. 1998;102(8):1534-1539. doi:10.1172/JCI4059.

47. Guo Y, He J, Zhao S, et al. Illumina human exome genotyping array clustering and quality control. *Nat Protoc*. 2014;9(11):2643-2662. doi:10.1038/nprot.2014.174.

48. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559-575. doi:10.1086/519795.

49. Thomsen M, Dahl M, Tybjaerg-Hansen A, Nordestgaard BG. beta2-adrenergic receptor

Thr164Ile polymorphism, obesity, and diabetes: comparison with FTO, MC4R, and TMEM18 polymorphisms in more than 64,000 individuals. *J Clin Endocrinol Metab*. 2012;97(6):E1074-9. doi:10.1210/jc.2011-3282.

50. Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics*. 2014;30(16):2375-2376. doi:10.1093/bioinformatics/btu197.

51. Team RC. R: A Language and Environment for Statistical Computing. 2015. http://www.r-project.org/.

52. Brodde OE, Buscher R, Tellkamp R, Radke J, Dhein S, Insel PA. Blunted cardiac responses to receptor activation in subjects with Thr164Ile beta(2)-adrenoceptors. *Circulation*. 2001;103(8):1048-1050. http://www.ncbi.nlm.nih.gov/pubmed/11222464.

53. Barbato E, Penicka M, Delrue L, et al. Thr164Ile polymorphism of beta2-adrenergic receptor negatively modulates cardiac contractility: implications for prognosis in patients with idiopathic dilated cardiomyopathy. *Heart*. 2007;93(7):856-861. doi:10.1136/hrt.2006.091959.

54. Wheatley CM, Snyder EM, Johnson BD, Olson TP. Sex differences in cardiovascular function during submaximal exercise in humans. *Springerplus*. 2014;3:445. doi:10.1186/2193-1801-3-445.

55. Oben JA, Yang S, Lin H, Ono M, Diehl AM. Acetylcholine promotes the proliferation and collagen gene expression of myofibroblastic hepatic stellate cells. *Biochem Biophys Res Commun*. 2003;300(1):172-177.

56. Hoffstedt J, Iliadou A, Pedersen NL, Schalling M, Arner P. The effect of the beta(2) adrenoceptor gene Thr164Ile polymorphism on human adipose tissue lipolytic function. *Br J Pharmacol*. 2001;133(5):708-712. doi:10.1038/sj.bjp.0704125.

57. Loomba R, Rao F, Zhang L, et al. Genetic covariance between gamma-glutamyl transpeptidase and fatty liver risk factors: role of beta2-adrenergic receptor genetic variation in twins. *Gastroenterology*. 2010;139(3):836-45, 845 e1. doi:10.1053/j.gastro.2010.06.009.

58. Klein I, Ojamaa K. Thyroid hormone and the cardiovascular system. *N Engl J Med*. 2001;344(7):501-509. doi:10.1056/NEJM200102153440707.

59. Hoit BD, Khoury SF, Shao Y, Gabel M, Liggett SB, Walsh RA. Effects of thyroid hormone on cardiac beta-adrenergic responsiveness in conscious baboons. *Circulation*. 1997;96(2):592-598. http://www.ncbi.nlm.nih.gov/pubmed/9244231.

60. Williams LT, Lefkowitz RJ, Watanabe AM, Hathaway DR, Besch Jr. HR. Thyroid hormone regulation of beta-adrenergic receptor number. *J Biol Chem*. 1977;252(8):2787-

2789. http://www.ncbi.nlm.nih.gov/pubmed/15999.

61. Fazio S, Palmieri EA, Lombardi G, Biondi B. Effects of thyroid hormone on the cardiovascular system. *Recent Prog Horm Res*. 2004;59:31-50.

62. Arioglu E, Guner S, Ozakca I, Altan VM, Ozcelikay AT. The changes in beta-adrenoceptor-mediated cardiac function in experimental hypothyroidism: the possible contribution of cardiac beta3-adrenoceptors. *Mol Cell Biochem*. 2010;335(1-2):59-66. doi:10.1007/s11010-009-0241-z.

63. Lorton D, Bellinger DL. Molecular mechanisms underlying beta-adrenergic receptor-mediated cross-talk between sympathetic neurons and immune cells. *Int J Mol Sci*. 2015;16(3):5635-5665. doi:10.3390/ijms16035635.

64. Joyner MJ, Wallin BG, Charkoudian N. Sex differences and blood pressure regulation in humans. *Exp Physiol*. 2015. doi:10.1113/EP085146.

65. Kneale BJ, Chowienczyk PJ, Brett SE, Coltart DJ, Ritter JM. Gender differences in sensitivity to adrenergic agonists of forearm resistance vasculature. *J Am Coll Cardiol*. 2000;36(4):1233-1238. http://www.ncbi.nlm.nih.gov/pubmed/11028476.

66. Greaney JL, Stanhewicz AE, Kenney WL, Alexander LM. Lack of limb or sex differences in the cutaneous vascular responses to exogenous norepinephrine. *J Appl Physiol*. 2014;117(12):1417-1423. doi:10.1152/japplphysiol.00575.2014.

67. Harvey RE, Barnes JN, Charkoudian N, et al. Forearm vasodilator responses to a beta-adrenergic receptor agonist in premenopausal and postmenopausal women. *Physiol Rep*. 2014;2(6). doi:10.14814/phy2.12032.

68. de Coupade C, Gear RW, Dazin PF, Sroussi HY, Green PG, Levine JD. Beta 2-adrenergic receptor regulation of human neutrophil function is sexually dimorphic. *Br J Pharmacol*. 2004;143(8):1033-1041. doi:10.1038/sj.bjp.0705972.

69. Lee P, Birzniece V, Umpleby AM, Poljak A, Ho KK. Formoterol, a highly beta2-selective agonist, induces gender-dimorphic whole body leucine metabolism in humans. *Metabolism*. 2015;64(4):506-512. doi:10.1016/j.metabol.2014.12.005.

70. Sheehy AM, Gaddis NC, Choi JD, Malim MH. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature*. 2002;418(6898):646-650. doi:10.1038/nature00939.

71. Neuberger M. Immunity through DNA deamination. *Tissue Antigens*. 2004;64(4):320. <Go to ISI>://000223876400003.

72. Conticello SG, Thomas CJF, Petersen-Mahrt SK, Neuberger MS. Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. *Mol Biol Evol*.

2005;22(2):367-377. doi:10.1093/molbev/msi026.

73. Sawyer SL, Emerman M, Malik HS. Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. *PLoS Biol*. 2004;2(9):E275. doi:10.1371/journal.pbio.0020275.

74. An P, Bleiber G, Duggal P, et al. APOBEC3G genetic variants and their influence on the progression to AIDS. *J Virol*. 2004;78(20):11070-11076. doi:10.1128/JVI.78.20.11070-11076.2004.

75. Feng Y, Chelico L. Intensity of deoxycytidine deamination of HIV-1 proviral DNA by the retroviral restriction factor APOBEC3G is mediated by the noncatalytic domain. *J Biol Chem*. 2011;286(13):11415-11426. doi:10.1074/jbc.M110.199604.

76. Do H, Vasilescu A, Diop G, et al. Exhaustive genotyping of the CEM15 (APOBEC3G) gene and absence of association with AIDS progression in a French cohort. *J Infect Dis*. 2005;191(2):159-163. doi:10.1086/426826.

77. Reddy K, Winkler CA, Werner L, et al. APOBEC3G expression is dysregulated in primary HIV-1 infection and polymorphic variants influence CD4+ T-cell counts and plasma viral load. *AIDS*. 2010;24(2):195-204. doi:10.1097/QAD.0b013e3283353bba.

78. Reddy K, Ooms M, Letko M, Garrett N, Simon V, Ndung'u T. Functional characterization of Vif proteins from HIV-1 infected patients with different APOBEC3G haplotypes. *AIDS*. 2016;30(11):1723-1729. doi:10.1097/QAD.0000000000001113.

79. Ezzikouri S, Kitab B, Rebbani K, et al. Polymorphic APOBEC3 modulates chronic hepatitis B in Moroccan population. *J Viral Hepat*. 2013;20(10):678-686. doi:10.1111/jvh.12042.

80. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393.

81. Oliveros JC. VENNY. An interactive tool for comparing lists with Venn Diagrams. *BioinfoGP of CNB-CSIC*. 2007:http://bioinfogp.cnnb.csic.es/tools/venny/index.ht. http://bioinfogp.cnnb.csic.es/tools/venny/index.html.

82. Koning FA, Newman ENC, Kim EY, Kunstman KJ, Wolinsky SM, Malim MH. Defining APOBEC3 Expression Patterns in Human Tissues and Hematopoietic Cell Subsets. *J Virol*. 2009;83(18):9474-9485. doi:10.1128/Jvi.01089-09.

83. Chiu Y-L, Greene WC. APOBEC3G: an intracellular centurion. *Philos Trans R Soc Lond B Biol Sci*. 2009;364(1517):689-703. doi:10.1098/rstb.2008.0193.

84. Hudson RP, Wilson SJ. Hypogammaglobulinemia and chronic lymphatic leukemia. *Cancer*. 1960;13(1):200-204. doi:10.1002/1097-0142(196001/02)13:1<200::AID-

CNCR2820130131>3.0.CO;2-Y.

85. Grey HM, Rabellino E, Pirofsky B. Immunoglobulins on the surface of lymphocytes. IV. Distribution in hypogammaglobulinemia, cellular immune deficiency, and chronic lymphatic leukemia. *J Clin Invest*. 1971;50(11):2368-2375. doi:10.1172/JCI106735.

86. Ezdinl EZ, Kucuk O, Chedid A, et al. Hypogammaglobulinemia and hemophagocytic syndrome associated with lymphoproliferative disorders. *Cancer*. 1986;57(5):1024-1037. doi:10.1002/1097-0142(19860301)57:5<1024::AID-CNCR2820570526>3.0.CO;2-H.

87. McCarthy H, Wierda WG, Barron LL, et al. High expression of activation-induced cytidine deaminase (AID) and splice variants is a distinctive feature of poor-prognosis chronic lymphocytic leukemia. *Blood*. 2003;101(12).

88. Burns MB, Lackey L, Carpenter MA, et al. APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature*. 2013;494(7437):366-370. doi:10.1038/nature11881.

89. Lackey L, Law EK, Brown WL, Harris RS. Subcellular localization of the APOBEC3 proteins during mitosis and implications for genomic DNA deamination. *Cell Cycle*. 2013;12(5):762-772. doi:10.4161/cc.23713.

90. Muramatsu M, Sankaranand VS, Anant S, et al. Specific Expression of Activation-induced Cytidine Deaminase (AID), a Novel Member of the RNA-editing Deaminase Family in Germinal Center B Cells*.

91. Muramatsu M, Kinoshita K, Fagarasan S, et al. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell*. 2000;102(5):553-563. doi:10.1016/S0092-8674(00)00078-7.

92. Revy P, Muto T, Levy Y, et al. Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2). *Cell*. 2000;102(5):565-575. doi:10.1016/S0092-8674(00)00079-9.

93. Lonsdale J, Thomas J, Salvatore M, et al. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45(6):580-585. doi:10.1038/ng.2653.

94. Amoêdo ND, Afonso AO, Cunha SM, Oliveira RH, Machado ES, Soares MA. Expression of APOBEC3G/3F and G-to-A hypermutation levels in HIV-1-infected children with different profiles of disease progression. *PLoS One*. 2011;6(8):e24118. doi:10.1371/journal.pone.0024118.

95. Bunupuradah T, Imahashi M, Iampornsin T, et al. Association of APOBEC3G genotypes and CD4 decline in Thai and Cambodian HIV-infected children with moderate immune deficiency. *AIDS Res Ther*. 2012;9(1):34. doi:10.1186/1742-6405-9-34.

96. Singh KK, Wang Y, Gray KP, et al. Genetic variants in the host restriction factor

APOBEC3G are associated with HIV-1-related disease progression and central nervous system impairment in children. *J Acquir Immune Defic Syndr*. 2013;62(2):197-203. doi:10.1097/QAI.0b013e31827ab612.

97. Simon AK, Hollander GA, McMichael A. Evolution of the immune system in humans from infancy to old age. *Proc R Soc London B Biol Sci*. 2015;282(1821).

98. Kidd JM, Newman TL, Tuzun E, Kaul R, Eichler EE. Population stratification of a common APOBEC gene deletion polymorphism. *PLoS Genet*. 2007;3(4):e63. doi:10.1371/journal.pgen.0030063.

99. Caval V, Suspene R, Shapira M, Vartanian JP, Wain-Hobson S. A prevalent cancer susceptibility APOBEC3A hybrid allele bearing APOBEC3B 3'UTR enhances chromosomal DNA damage. *Nat Commun*. 2014;5:5129. doi:10.1038/ncomms6129.

100. Roberts SA, Lawrence MS, Klimczak LJ, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet*. 2013;45(9):970-976. doi:10.1038/ng.2702.

101. Chan K, Roberts SA, Klimczak LJ, et al. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat Genet*. 2015;47(9):1067-1072. doi:10.1038/ng.3378.

102. Burns MB, Temiz NA, Harris RS. Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat Genet*. 2013;45(9):977-983. doi:10.1038/ng.2701.

103. Mussil B, Suspene R, Aynaud MM, Gauvrit A, Vartanian JP, Wain-Hobson S. Human APOBEC3A isoforms translocate to the nucleus and induce DNA double strand breaks leading to cell stress and death. *PLoS One*. 2013;8(8):e73641. doi:10.1371/journal.pone.0073641.

104. Lackey L, Demorest ZL, Land AM, Hultquist JF, Brown WL, Harris RS. APOBEC3B and AID have similar nuclear import mechanisms. *J Mol Biol*. 2012;419(5):301-314. doi:10.1016/j.jmb.2012.03.011.

105. Cullen BR. Role and mechanism of action of the APOBEC3 family of antiretroviral resistance factors. *J Virol*. 2006;80(3):1067-1076. doi:10.1128/JVI.80.3.1067-1076.2006.

106. Conticello SG. The AID/APOBEC family of nucleic acid mutators. *Genome Biol*. 2008;9(6):229. doi:10.1186/gb-2008-9-6-229.

107. Long J, Delahanty RJ, Li G, et al. A common deletion in the APOBEC3 genes and breast cancer risk. *J Natl Cancer Inst*. 2013;105(8):573-579. doi:10.1093/jnci/djt018.

108. Xuan D, Li G, Cai Q, et al. APOBEC3 deletion polymorphism is associated with breast cancer risk among women of European ancestry. *Carcinogenesis*. 2013;34(10):2240-2243.

doi:10.1093/carcin/bgt185.

109. Chiu YL, Greene WC. The APOBEC3 cytidine deaminases: an innate defensive network opposing exogenous retroviruses and endogenous retroelements. *Annu Rev Immunol*. 2008;26:317-353. doi:10.1146/annurev.immunol.26.021607.090350.

110. Cooper MD, Alder MN. The evolution of adaptive immune systems. *Cell*. 2006;124(4):815-822. doi:10.1016/j.cell.2006.02.001.

111. Bogerd HP, Wiegand HL, Doehle BP, Lueders KK, Cullen BR. APOBEC3A and APOBEC3B are potent inhibitors of LTR-retrotransposon function in human cells. *Nucleic Acids Res*. 2006;34(1):89-95. doi:10.1093/nar/gkj416.

112. Muckenfuss H, Hamdorf M, Held U, et al. APOBEC3 proteins inhibit human LINE-1 retrotransposition. *J Biol Chem*. 2006;281(31):22161-22172. doi:10.1074/jbc.M601716200.

113. Suspene R, Guetard D, Henry M, Sommer P, Wain-Hobson S, Vartanian JP. Extensive editing of both hepatitis B virus DNA strands by APOBEC3 cytidine deaminases in vitro and in vivo. *Proc Natl Acad Sci U S A*. 2005;102(23):8321-8326. doi:10.1073/pnas.0408223102.

114. Abe H, Ochi H, Maekawa T, et al. Effects of structural variations of APOBEC3A and APOBEC3B genes in chronic hepatitis B virus infection. *Hepatol Res*. 2009;39(12):1159-1168. doi:10.1111/j.1872-034X.2009.00566.x.

115. Zhang T, Cai J, Chang J, et al. Evidence of associations of APOBEC3B gene deletion with susceptibility to persistent HBV infection and hepatocellular carcinoma. *Hum Mol Genet*. 2013;22(6):1262-1269. doi:10.1093/hmg/dds513.

116. Jha P, Sinha S, Kanchan K, et al. Deletion of the APOBEC3B gene strongly impacts susceptibility to falciparum malaria. *Infect Genet Evol*. 2012;12(1):142-148. doi:10.1016/j.meegid.2011.11.001.

117. An P, Johnson R, Phair J, et al. APOBEC3B deletion and risk of HIV-1 acquisition. *J Infect Dis*. 2009;200(7):1054-1058. doi:10.1086/605644.

118. Imahashi M, Izumi T, Watanabe D, et al. Lack of association between intact/deletion polymorphisms of the APOBEC3B gene and HIV-1 risk. *PLoS One*. 2014;9(3):e92861. doi:10.1371/journal.pone.0092861.

119. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. *Nat Methods*. 2012;9(2):179-181. doi:10.1038/nmeth.1785.

120. Howie BN, Donnelly P, Marchini J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. Schork NJ, ed.

PLoS Genet. 2009;5(6):e1000529. doi:10.1371/journal.pgen.1000529.

121. Manolio TA. Collaborative genome-wide association studies of diverse diseases: programs of the NHGRI's office of population genomics. *Pharmacogenomics*. 2009;10(2):235-241. doi:10.2217/14622416.10.2.235.

122. Chiu YL, Greene WC. Multifaceted antiviral actions of APOBEC3 cytidine deaminases. *Trends Immunol*. 2006;27(6):291-297. doi:10.1016/j.it.2006.04.003.

123. An P, Johnson R, Phair J, et al. APOBEC3B deletion and risk of HIV-1 acquisition. *J Infect Dis*. 2009;200(7):1054-1058. doi:10.1086/605644.

124. Imahashi M, Izumi T, Watanabe D, et al. Lack of association between intact/deletion polymorphisms of the APOBEC3B gene and HIV-1 risk. *PLoS One*. 2014;9(3):e92861. doi:10.1371/journal.pone.0092861.

125. Itaya S, Nakajima T, Kaur G, et al. No evidence of an association between the APOBEC3B deletion polymorphism and susceptibility to HIV infection and AIDS in Japanese and Indian populations. *J Infect Dis*. 2010;202(5):815-6-7. doi:10.1086/655227.

126. Itsara A, Cooper GM, Baker C, et al. Population analysis of large copy number variants and hotspots of human genetic disease. *Am J Hum Genet*. 2009;84(2):148-161. doi:10.1016/j.ajhg.2008.12.014.

127. Lupski JR, de Oca-Luna RM, Slaugenhaupt S, et al. DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell*. 1991;66(2):219-232. doi:10.1016/0092-8674(91)90613-4.

128. Carlson C, Sirotkin H, Pandita R, et al. Molecular definition of 22q11 deletions in 151 velo-cardio-facial syndrome patients. *Am J Hum Genet*. 1997;61(3):620-629. doi:10.1086/515508.

129. The Wellcome Trust Case Control Consortium. Genome-wide association study of CNVs in 16 , 000 cases of eight common diseases and 3 , 000 shared controls. *Nature*. 2010;464(7289):713-720. doi:10.1038/nature08979.

130. 1000 Genomes Project Consortium T 1000 GP, Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56-65. doi:10.1038/nature11632.

131. Conrad DF, Pinto D, Redon R, et al. Origins and functional impact of copy number variation in the human genome. *Nature*. 2010;464(7289):704-712. doi:10.1038/nature08516.

132. Walsh T, McClellan JM, McCarthy SE, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*. 2008;320(5875):539-

543. doi:10.1126/science.1155174.

133. Marshall CR, Noor A, Vincent JB, et al. Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet*. 2008;82(2):477–488. doi:10.1016/j.ajhg.2007.12.009.

134. Gonzalez E, Kulkarni H, Bolivar H, et al. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*. 2005;307(5714):1434-1440. doi:10.1126/science.1101160.

135. Stranger BE, Forrest MS, Dunning M, et al. Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes. *Science*. 2007;315(5813):848-853. doi:10.1126/science.1136678.

136. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. *Nature*. 2006;444(7118):444-454. doi:10.1038/nature05329.

137. Pang AW, MacDonald JR, Pinto D, et al. Towards a comprehensive structural variation map of an individual human genome. *Genome Biol*. 2010;11(5):R52. doi:10.1186/gb-2010-11-5-r52.

138. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The Database of Genomic Variants: A curated collection of structural variation in the human genome. *Nucleic Acids Res*. 2014;42(D1). doi:10.1093/nar/gkt958.

139. Lappalainen I, Lopez J, Skipper L, et al. DbVar and DGVa: Public archives for genomic structural variation. *Nucleic Acids Res*. 2013;41(D1). doi:10.1093/nar/gks1213.

140. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-842. doi:10.1093/bioinformatics/btq033.

141. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*. 2007;35(SUPPL. 1). doi:10.1093/nar/gkl842.

142. Karolchik D, Barber GP, Casper J, et al. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res*. 2014;42(D1). doi:10.1093/nar/gkt1168.

143. Gray MW, Burger G, Lang BF. The origin and early evolution of mitochondria. *Genome Biol*. 2001;2(6):REVIEWS1018. doi:10.1186/gb-2001-2-6-reviews1018.

144. Friedman JR, Nunnari J. Mitochondrial form and function. *Nature*. 2014;505(7483):335-343. doi:10.1038/nature12985.

145. Gabaldón T, Huynen MA. Shaping the mitochondrial proteome. In: *Biochimica et Biophysica Acta - Bioenergetics*. Vol 1659. ; 2004:212-220.

doi:10.1016/j.bbabio.2004.07.011.

146. Neupert W, Herrmann JM. Translocation of Proteins into Mitochondria. *Annu Rev Biochem*. 2007;76(1):723-749. doi:10.1146/annurev.biochem.76.052705.163409.

147. Case JT, Wallace DC. Maternal inheritance of mitochondrial DNA polymorphisms in cultured human fibroblasts. *Somatic Cell Genet*. 1981;7(1):103-108.

148. Sharma H, Singh A, Sharma C, Jain SK, Singh N. Mutations in the mitochondrial DNA D-loop region are frequent in cervical cancer. *Cancer Cell Int*. 2005;5:34. doi:10.1186/1475-2867-5-34.

149. Brown WM, George M, Wilson AC. Rapid evolution of animal mitochondrial DNA. *Proc Natl Acad Sci U S A*. 1979;76(4):1967-1971.

150. Wallace DC, Brown MD, Lott MT. Mitochondrial DNA variation in human evolution and disease. *Gene*. 1999;238(1):211-230.

151. Gómez-Durán A, Pacheu-Grau D, López-Gallardo E, et al. Unmasking the causes of multifactorial disorders: OXPHOS differences between mitochondrial haplogroups. *Hum Mol Genet*. 2010;19(17):3343-3353. doi:10.1093/hmg/ddq246.

152. Brown MD, Trounce IA, Jun AS, Allen JC, Wallace DC. Functional analysis of lymphoblast and cybrid mitochondria containing the 3460, 11778, or 14484 Leber's hereditary optic neuropathy mitochondrial DNA mutation. *J Biol Chem*. 2000;275(51):39831-39836. doi:10.1074/jbc.M006476200.

153. Ross OA, McCormack R, Curran MD, et al. Mitochondrial DNA polymorphism: Its role in longevity of the Irish population. *Exp Gerontol*. 2001;36(7):1161-1178. doi:10.1016/S0531-5565(01)00094-8.

154. Gaweda-Walerych K, Maruszak A, Safranow K, et al. Mitochondrial DNA haplogroups and subhaplogroups are associated with Parkinson's disease risk in a Polish PD cohort. *J Neural Transm*. 2008;115(11):1521-1526. doi:10.1007/s00702-008-0121-9.

155. van der Walt JM, Nicodemus KK, Martin ER, et al. Mitochondrial polymorphisms significantly reduce the risk of Parkinson disease. *Am J Hum Genet*. 2003;72(4):804-811. doi:10.1086/373937.

156. Houshmand M, Sanati MH, Babrzadeh F, et al. Population screening for association of mitochondrial haplogroups BM, J, K and M with multiple sclerosis: interrelation between haplogroup J and MS in Persian patients. *Mult Scler*. 2005;11(6):728-730. doi:10.1191/1352458505ms1228sr.

157. Kalman B, Li S, Chatterjee D, et al. Large scale screening of the mitochondrial DNA reveals no pathogenic mutations but a haplotype associated with multiple sclerosis in

Caucasians. *Acta Neurol Scand*. 1999;99(1):16-25.

158. Jones MM, Manwaring N, Wang JJ, Rochtchina E, Mitchell P, Sue CM. Mitochondrial DNA haplogroups and age-related maculopathy. *Arch Ophthalmol*. 2007;125(9):1235-1240. doi:10.1001/archopht.125.9.1235.

159. Kloss-Brandstätter A, Pacher D, Schönherr S, et al. HaploGrep: A fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum Mutat*. 2011;32(1):25-32. doi:10.1002/humu.21382.

160. van Oven M. PhyloTree Build 17: Growing the human mitochondrial DNA tree. *Forensic Sci Int Genet Suppl Ser*. 2015;5:9-11. doi:10.1016/j.fsigss.2015.09.155.

161. Calvo SE, Clauser KR, Mootha VK. MitoCarta2.0: An updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Res*. 2016;44(D1):D1251-D1257. doi:10.1093/nar/gkv1003.

162. Visconte V, Avishai N, Mahfouz R, et al. Distinct iron architecture in SF3B1 mutant myelodysplastic syndromes patients is linked to an SLC25A37 splice variant with a retained intron. *Leukemia*. 2014;29(April):188-195. doi:10.1038/leu.2014.170.

163. Del Rey M, Benito R, Fontanillo C, et al. Deregulation of genes related to iron and mitochondrial metabolism in refractory anemia with ring sideroblasts. *PLoS One*. 2015;10(5). doi:10.1371/journal.pone.0126555.

164. Belaidi AA, Bush AI. Iron neurochemistry in Alzheimer's disease and Parkinson's disease: targets for therapeutics. *J Neurochem*. October 2016:179-197. doi:10.1111/jnc.13425.

165. Heidari M, Gerami SH, Bassett B, et al. Pathological relationships involving iron and myelin may constitute a shared mechanism linking various rare and common brain diseases. *Rare Dis (Austin, Tex)*. 2016;4(1):e1198458. doi:10.1080/21675511.2016.1198458.

166. Pasquale LR, Loomis SJ, Aschard H, et al. Exploring genome-wide - dietary heme iron intake interactions and the risk of type 2 diabetes. *Front Genet*. 2013;4:7. doi:10.3389/fgene.2013.00007.

167. Edvardson S, Shaag A, Kolesnikova O, et al. Deleterious mutation in the mitochondrial arginyl-transfer RNA synthetase gene is associated with pontocerebellar hypoplasia. *Am J Hum Genet*. 2007;81(4):857-862. doi:10.1086/521227.

168. Wiley LK, Shah A, Xu H, Bush WS. ICD-9 tobacco use codes are effective identifiers of smoking status. *J Am Med Inform Assoc*. 2013;20:652-658. doi:10.1136/amiajnl-2012-001557.

APPENDIX

Appendix A. PheWAS result for top 25 hits in all individuals from ADRB2 Thr164Ile Analysis.

PheWAS Code	PheWAS Description	SNP	Beta	OR	SE	p	Total N	Case N	Control N	Passes Bonferroni	Passes FDR
458.2	Iatrogenic hypotension	Thr164Ile	1.606	4.983	0.379	2.25E-05	19180	56	19124	TRUE	TRUE
573.9	Abnormal serum enzyme levels	Thr164Ile	0.914	2.495	0.229	6.50E-05	18862	336	18526	FALSE	TRUE
278.11	Morbid obesity	Thr164Ile	0.597	1.817	0.162	2.23E-04	20356	905	19451	FALSE	FALSE
961	Poisoning by other anti-infectives	Thr164Ile	1.432	4.186	0.402	3.72E-04	19235	56	19179	FALSE	FALSE
244.2	Acquired hypothyroidism	Thr164Ile	1.094	2.987	0.317	5.45E-04	18240	136	18104	FALSE	FALSE
278.3	Localized adiposity	Thr164Ile	1.336	3.804	0.402	8.85E-04	19512	61	19451	FALSE	FALSE
303	Psychogenic and somatoform disorders	Thr164Ile	1.160	3.188	0.349	9.00E-04	15642	99	15543	FALSE	FALSE
276.42	Alkalosis	Thr164Ile	0.992	2.697	0.315	1.64E-03	14885	154	14731	FALSE	FALSE
303.4	Somatoform disorder	Thr164Ile	1.239	3.453	0.400	1.96E-03	15611	68	15543	FALSE	FALSE
41.9	Drug-resistant infection	Thr164Ile	0.704	2.023	0.230	2.21E-03	20090	383	19707	FALSE	FALSE
530.14	Reflux esophagitis	Thr164Ile	1.117	3.055	0.382	3.46E-03	14098	93	14005	FALSE	FALSE
550	Abdominal hernia	Thr164Ile	0.386	1.471	0.132	3.47E-03	22302	1899	20403	FALSE	FALSE
260	Protein-calorie malnutrition	Thr164Ile	0.400	1.491	0.137	3.57E-03	19088	1652	17436	FALSE	FALSE
297.1	Suicidal ideation	Thr164Ile	1.261	3.529	0.436	3.80E-03	15598	55	15543	FALSE	FALSE
594.3	Calculus of ureter	Thr164Ile	0.747	2.111	0.259	3.87E-03	22031	282	21749	FALSE	FALSE
369.5	Conjunctivitis, infectious	Thr164Ile	0.852	2.344	0.317	7.15E-03	20450	166	20284	FALSE	FALSE
550.2	Diaphragmatic hernia	Thr164Ile	0.514	1.672	0.193	7.60E-03	21141	738	20403	FALSE	FALSE
789	Nausea and vomiting	Thr164Ile	0.259	1.296	0.101	1.04E-02	21444	4090	17354	FALSE	FALSE
337	Disorders of the autonomic nervous system	Thr164Ile	0.852	2.345	0.334	1.06E-02	18266	154	18112	FALSE	FALSE
369	Infection of the eye	Thr164Ile	0.682	1.977	0.269	1.12E-02	20560	276	20284	FALSE	FALSE
512.2	Painful respiration	Thr164Ile	0.634	1.886	0.252	1.18E-02	17302	342	16960	FALSE	FALSE
594	Urinary calculus	Thr164Ile	0.433	1.543	0.173	1.22E-02	22657	908	21749	FALSE	FALSE
506	Empyema and pneumothorax	Thr164Ile	0.441	1.555	0.178	1.30E-02	16199	866	15333	FALSE	FALSE
377.1	Optic atrophy	Thr164Ile	1.119	3.060	0.453	1.35E-02	20619	67	20552	FALSE	FALSE

Appendix B. PheWAS result for top 25 hits in females only from ADRB2 Thr164Ile Analysis.

PheWAS Code	PheWAS Description	SNP	Beta	OR	SE	p	Total N	Case N	Control N	Passes Bonferroni	Passes FDR
458.2	Iatrogenic hypotension	The164Ile	2.015	7.503	0.426	2.26E-06	10636	32	10604	TRUE	TRUE
244.2	Acquired hypothyroidism	The164Ile	1.535	4.640	0.347	9.57E-06	8959	81	8878	TRUE	TRUE
573.9	Abnormal serum enzyme levels	The164Ile	1.143	3.135	0.286	6.57E-05	10325	167	10158	FALSE	TRUE
41.9	Drug-resistant infection	The164Ile	1.068	2.909	0.273	9.10E-05	10958	192	10766	FALSE	TRUE
703.1	Ingrowing nail	The164Ile	1.354	3.874	0.363	1.89E-04	11887	80	11807	FALSE	TRUE
480	Pneumonia	The164Ile	0.500	1.649	0.150	8.67E-04	11093	1427	9666	FALSE	FALSE
961	Poisoning by other anti-infectives	The164Ile	1.491	4.442	0.450	9.19E-04	10438	44	10394	FALSE	FALSE
303	Psychogenic and somatoform disorders	The164Ile	1.215	3.372	0.384	1.55E-03	7824	80	7744	FALSE	FALSE
278.3	Localized adiposity	The164Ile	1.361	3.900	0.445	2.21E-03	10351	47	10304	FALSE	FALSE
550.2	Diaphragmatic hernia	The164Ile	0.696	2.005	0.231	2.58E-03	11682	428	11254	FALSE	FALSE
506	Empyema and pneumothorax	The164Ile	0.697	2.007	0.234	2.93E-03	9160	393	8767	FALSE	FALSE
481	Influenza	The164Ile	1.144	3.139	0.386	3.06E-03	9759	93	9666	FALSE	FALSE
41	Bacterial infection NOS	The164Ile	0.457	1.579	0.156	3.34E-03	12016	1250	10766	FALSE	FALSE
292.4	Altered mental status	The164Ile	0.577	1.780	0.197	3.38E-03	10181	657	9524	FALSE	FALSE
705	Disorders of sweat glands	The164Ile	1.436	4.202	0.500	4.12E-03	11368	37	11331	FALSE	FALSE
512.2	Painful respiration	The164Ile	0.838	2.313	0.292	4.15E-03	9594	200	9394	FALSE	FALSE
303.4	Somatoform disorder	The164Ile	1.282	3.603	0.450	4.36E-03	7797	53	7744	FALSE	FALSE
783	Fever of unknown origin	The164Ile	0.406	1.501	0.143	4.36E-03	11650	1706	9944	FALSE	FALSE
591	Urinary tract infection	The164Ile	0.355	1.426	0.127	5.04E-03	10876	2782	8094	FALSE	FALSE
189.4	Malignant neoplasm of kidney and other urinary organs	The164Ile	1.414	4.113	0.511	5.65E-03	12481	37	12444	FALSE	FALSE
939	Atopic or contact dermatitis	The164Ile	0.531	1.701	0.196	6.63E-03	10899	689	10210	FALSE	FALSE
276.1	Electrolyte imbalance	The164Ile	0.362	1.437	0.134	6.68E-03	10569	2475	8094	FALSE	FALSE
327.3	Sleep apnea	The164Ile	0.566	1.761	0.210	6.98E-03	10809	555	10254	FALSE	FALSE
384.4	Perforation of tympanic membrane	The164Ile	1.372	3.943	0.509	7.05E-03	11269	39	11230	FALSE	FALSE
530.14	Reflux esophagitis	The164Ile	1.245	3.474	0.463	7.18E-03	7676	60	7616	FALSE	FALSE
594	Urinary calculus	The164Ile	0.650	1.915	0.242	7.18E-03	12306	363	11943	FALSE	FALSE

Appendix C. PheWAS results for phenotypes we had anticipated might be associated with ADRB2 Thr164Ile prior to analysis.

	PheWAS Description	SNP	Beta	OR	SE	p	Total N	Case N	Control N
All									
	Hypertension	Thr164Ile	0.114982	1.121854	0.094509	0.223747	21474	11444	10030
	Essential hypertension	Thr164Ile	0.129056	1.137754	0.094806	0.173431	21268	11238	10030
	Asthma	Thr164Ile	0.197687	1.218581	0.166332	0.234633	20053	1284	18769
	Chronic obstructive asthma	Thr164Ile	0.450507	1.569108	0.382303	0.238635	18948	179	18769
	Chronic obstructive asthma with exacerbation	Thr164Ile	NA	NA	NA	NA	18804	35	18769
	Asthma with exacerbation	Thr164Ile	0.438416	1.55025	0.37969	0.248225	18950	181	18769
Females									
	Hypertension	The164Ile	0.128784	1.137445	0.130667	0.324333	11702	5795	5907
	Essential hypertension	The164Ile	0.147349	1.158758	0.130903	0.260319	11612	5705	5907
	Asthma	The164Ile	0.256596	1.292523	0.196963	0.192655	11005	881	10124
	Chronic obstructive asthma	The164Ile	-0.00458	0.995428	0.585769	0.993758	10240	116	10124
	Chronic obstructive asthma with exacerbation	The164Ile	0.371146	1.449394	1.008659	0.712903	10151	27	10124
	Asthma with exacerbation	The164Ile	0.524153	1.689028	0.415364	0.20698	10264	140	10124
Males									
	Hypertension	Thr164Ile	0.094461	1.099067	0.138747	0.495986	9745	5634	4111
	Essential hypertension	Thr164Ile	0.103523	1.109071	0.139317	0.457436	9630	5519	4111
	Asthma	Thr164Ile	0.02469	1.024997	0.320546	0.938604	9030	402	8628
	Chronic obstructive asthma	Thr164Ile	0.964345	2.623069	0.500894	0.054199	8691	63	8628
	Chronic obstructive asthma with exacerbation	Thr164Ile	NA	NA	NA	NA	8636	8	8628
	Asthma with exacerbation	Thr164Ile	0.018239	1.018406	0.995608	0.985384	8669	41	8628

Appendix D. ICD-9 codes mapping to PheWAS codes discussed in Chapter 1. PheWAS Code Name and number are bolded followed by all ICD-9 codes in that mapping.

Iatrogenic Hypotension (458.2)	**Poisoning by Other Anti-infectives (961)**	**Pneumonia (480) Continued**	**Pneumonia (480) Continued**
458.2	961	112.4	482.89
458.21	961.1	114.0	482.9
458.29	961.2	114.4	483
	961.3	114.5	483.0
Serum Enzyme Abnormalities (573.9)	961.4	130.4	483.1
	961.5	136.3	483.8
790.5	961.6	480	484
	961.7	480.0	484.1
Drug-resistant Infection (041.9)	961.9	480.1	484.3
	E857	480.2	484.5
V09	E931	480.3	484.6
V09.0	E931.1	480.8	484.7
V09.1	E931.2	480.9	484.8
V09.2	E931.3	481	485
V09.3	E931.4	481.0	485.0
V09.4	E931.5	482	486
V09.5	E931.6	482.0	513
V09.50	E931.7	482.1	513.0
V09.51	E931.9	482.2	513.1
V09.6	V14.3	482.3	517.1
V09.7		482.30	V12.61
V09.70	**Pneumonia (480)**	482.31	
V09.71	003.22	482.32	
V09.8	020.3	482.39	
V09.80	020.4	482.4	
V09.81	020.5	482.40	
V09.9	021.2	482.41	
V09.90	022.1	482.42	
V09.91	031.0	482.49	
	039.1	482.8	
Acquired Hypotension (244.2)	052.1	482.81	
	055.1	482.82	
244.8	073.0	482.83	
	083.0	482.84	

Appendix E. Top 25 PheWAS hits for all individuals from the A3G His186Arg analysis.

PheWAS Code	PheWAS Description	SNP	Beta	OR	SE	p	Total N	Case N	Control N
279.11	Deficiency of humoral immunity	His186Arg	1.075	2.930	0.239	7.01E-06	25271	108	25163
292.6	Hallucinations	His186Arg	1.119	3.061	0.316	3.96E-04	20674	64	20610
597	Other disorders of urethra and urinary tract	His186Arg	0.502	1.652	0.166	2.42E-03	24887	400	24487
452	Venous embolism & thrombosis	His186Arg	-0.413	0.661	0.138	2.65E-03	22086	1396	20690
242	Thyrotoxicosis	His186Arg	0.471	1.602	0.160	3.20E-03	22032	441	21591
174.3	Neoplasm of uncertain behavior of breast	His186Arg	0.843	2.324	0.312	6.81E-03	23780	81	23699
573.5	Jaundice	His186Arg	0.461	1.585	0.172	7.52E-03	22159	370	21789
242.1	Graves' disease	His186Arg	0.557	1.745	0.210	8.02E-03	21821	230	21591
333	Extrapyramidal disease and abnormal movement disorders	His186Arg	-0.618	0.539	0.239	9.79E-03	21494	550	20944
771	Musculoskeletal symptoms referable to limbs	His186Arg	0.531	1.701	0.206	1.01E-02	24066	252	23814
117	Mycoses	His186Arg	-0.860	0.423	0.355	1.55E-02	23965	299	23666
306	Random mental disorder. Ignored for now	His186Arg	0.698	2.010	0.295	1.80E-02	18704	97	18607
596.5	Functional disorders of bladder	His186Arg	0.353	1.423	0.152	2.02E-02	25016	529	24487
334	Degenerative disease of the spinal cord	His186Arg	-0.701	0.496	0.304	2.12E-02	21302	358	20944
270	Protein plasma/amino-acid transport and metabolism disorder	His186Arg	-0.519	0.595	0.226	2.20E-02	26060	544	25516
270.3	Plasma protein metabolism disorder	His186Arg	-0.594	0.552	0.261	2.29E-02	25961	445	25516
426.91	Cardiac pacemaker in situ	His186Arg	0.287	1.333	0.127	2.34E-02	17613	948	16665
281.11	Pernicious anemia	His186Arg	0.544	1.722	0.248	2.84E-02	17501	170	17331
371.21	Allergic conjunctivitis	His186Arg	-0.689	0.502	0.320	3.13E-02	24018	322	23696
333.4	Torsion dystonia	His186Arg	-2.133	0.118	1.001	3.31E-02	21082	138	20944
695.41	Lupus erythematosus	His186Arg	0.676	1.966	0.324	3.72E-02	23093	85	23008
550.6	Incisional hernia	His186Arg	0.370	1.448	0.179	3.87E-02	24315	383	23932
275.5	Calcium/phosphorus disorders	His186Arg	-0.341	0.711	0.165	3.92E-02	25182	851	24331
362.3	NA	His186Arg	0.512	1.669	0.248	3.93E-02	27309	174	27135
470	Deviated nasal septum	His186Arg	-0.390	0.677	0.189	3.96E-02	20008	700	19308
279.11	Deficiency of humoral immunity	His186Arg	1.075	2.930	0.239	7.01E-06	25271	108	25163

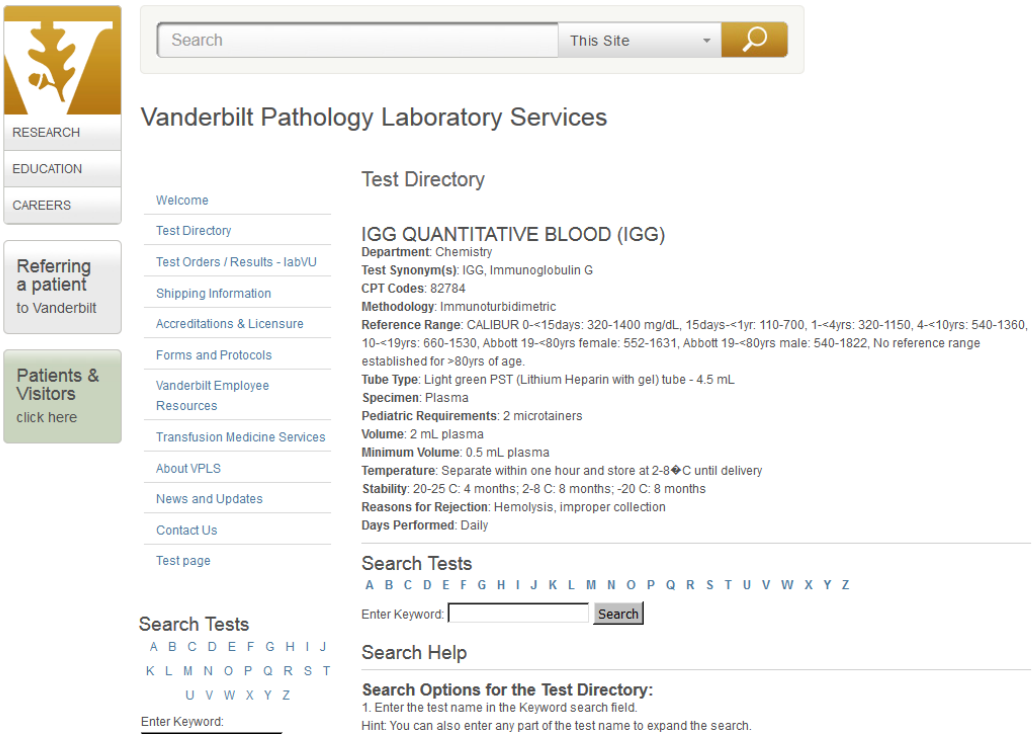
Appendix F.Top 25 PheWAS hits for individuals under the age of 20 at their last ICD9 code record from the A3G His186Arg analysis.

PheWAS Code	PheWAS Description	SNP	Beta	OR	SE	p	Total N	Case N	Control N
279.11	Deficiency of humoral immunity	His186Arg	1.714	5.549	0.383	7.73E-06	3557	27	3530
870.3	Other open wound of head and face	His186Arg	1.342	3.828	0.434	1.98E-03	3517	24	3493
714.1	Rheumatoid arthritis	His186Arg	1.021	2.775	0.366	5.35E-03	3579	43	3536
244	Hypothyroidism	His186Arg	0.687	1.987	0.263	9.01E-03	3684	123	3561
788	Syncope and collapse	His186Arg	0.809	2.245	0.313	9.73E-03	3696	76	3620
687.1	Rash and other nonspecific skin eruption	His186Arg	0.547	1.728	0.220	1.31E-02	3519	210	3309
389.2	Conductive hearing loss	His186Arg	0.580	1.786	0.234	1.31E-02	3560	176	3384
327.32	Obstructive sleep apnea	His186Arg	0.746	2.108	0.307	1.53E-02	3573	86	3487
327	Sleep disorders	His186Arg	0.568	1.765	0.245	2.05E-02	3649	162	3487
507	Pleurisy; pleural effusion	His186Arg	-0.775	0.461	0.345	2.48E-02	3266	263	3003
597	Other disorders of urethra and urinary tract	His186Arg	1.055	2.871	0.479	2.76E-02	3465	27	3438
588.2	Secondary hyperparathyroidism (of renal origin)	His186Arg	1.047	2.848	0.477	2.82E-02	3274	26	3248
599.4	Urinary incontinence	His186Arg	0.715	2.045	0.329	2.96E-02	3374	77	3297
244.4	NA	His186Arg	0.623	1.864	0.288	3.08E-02	3790	107	3683
599	Symptoms/disorders of the urinary system	His186Arg	0.491	1.634	0.232	3.41E-02	3496	199	3297
261.4	Vitamin D deficiency	His186Arg	0.807	2.242	0.383	3.53E-02	2665	47	2618
851	Complications of transplants and reattached limbs	His186Arg	0.939	2.558	0.464	4.30E-02	3384	29	3355
327.3	Sleep apnea	His186Arg	0.551	1.735	0.274	4.39E-02	3617	130	3487
327.4	Insomnia	His186Arg	1.029	2.797	0.513	4.48E-02	3509	22	3487
242	Thyrotoxicosis	His186Arg	1.013	2.755	0.506	4.53E-02	3583	22	3561
870	Open wounds of head; neck; and trunk	His186Arg	0.655	1.924	0.333	4.93E-02	3568	75	3493
809	Fracture of unspecified bones	His186Arg	-1.957	0.141	1.003	5.11E-02	3414	101	3313
416	Cardiomegaly	His186Arg	-0.618	0.539	0.318	5.24E-02	3650	257	3393
506	Empyema and pneumothorax	His186Arg	-1.947	0.143	1.006	5.29E-02	3096	93	3003
428.4	NA	His186Arg	-1.940	0.144	1.013	5.55E-02	3828	75	3753
579	Other symptoms involving abdomen and pelvis	His186Arg	-1.919	0.147	1.005	5.63E-02	3316	93	3223
395	Heart valve disorders	His186Arg	-0.972	0.378	0.510	5.66E-02	3374	140	3234

Appendix G. List of all ICD9 codes that map to the Deficiency of Humoral Immunity PheWAS code discussed in chapter 2.

Deficiency of Humoral Immunity (279.11)
279.0
279.00
279.01
279.02
279.03
279.04
279.05
279.06
279.09

Appendix H. Screenshot of Vanderbilt Pathology Laboratory Services Test Directory showing Reference ranges of IGG Quantitative Blood Test.



Appendix I. Top 24 hits from PheWAS of the A3B deletion in all individuals.

PheWAS Code	PheWAS Description	SNP	Beta	OR	SE	p	Total N	Case N	Control N	Passes Bonferroni	Passes FDR	Passes SimpleM
428.2	Heart failure NOS	A3BΔ	0.720	2.054	0.167	1.55E-05	2732	262	2470	TRUE	TRUE	TRUE
428.1	Systolic/diastolic heart failure	A3BΔ	0.518	1.678	0.123	2.42E-05	3250	780	2470	TRUE	TRUE	TRUE
395.2	Nonrheumatic aortic valve disorders	A3BΔ	0.700	2.013	0.167	2.77E-05	3220	247	2973	TRUE	TRUE	TRUE
395	Heart valve disorders	A3BΔ	0.490	1.633	0.124	7.81E-05	3564	591	2973	FALSE	TRUE	TRUE
428	Heart failure	A3BΔ	0.462	1.587	0.119	1.01E-04	3345	875	2470	FALSE	TRUE	TRUE
427.9	Palpitations	A3BΔ	0.577	1.781	0.164	4.22E-04	2059	350	1709	FALSE	FALSE	TRUE
440	Atherosclerosis	A3BΔ	0.556	1.744	0.161	5.61E-04	3094	302	2792	FALSE	FALSE	FALSE
425	Cardiomyopathy	A3BΔ	0.472	1.604	0.142	8.85E-04	3638	379	3259	FALSE	FALSE	FALSE
426	Cardiac conduction disorders	A3BΔ	0.462	1.587	0.141	1.02E-03	2364	655	1709	FALSE	FALSE	FALSE
429.9	Cardiac complications, not elsewhere classified	A3BΔ	1.177	3.244	0.371	1.51E-03	2502	32	2470	FALSE	FALSE	FALSE
415.21	NA	A3BΔ	0.857	2.357	0.271	1.59E-03	3956	62	3894	FALSE	FALSE	FALSE
427	Cardiac dysrhythmias	A3BΔ	0.324	1.383	0.107	2.49E-03	3371	1662	1709	FALSE	FALSE	FALSE
442	Other aneurysm	A3BΔ	0.618	1.855	0.205	2.57E-03	2952	160	2792	FALSE	FALSE	FALSE
426.92	Cardiac defibrillator in situ	A3BΔ	0.599	1.820	0.201	2.86E-03	1914	205	1709	FALSE	FALSE	FALSE
442.1	Aortic aneurysm	A3BΔ	0.672	1.958	0.227	3.10E-03	2914	122	2792	FALSE	FALSE	FALSE
579.2	NA	A3BΔ	0.817	2.263	0.281	3.62E-03	3920	59	3861	FALSE	FALSE	FALSE
761	Neck pain	A3BΔ	-0.695	0.499	0.239	3.63E-03	3747	301	3446	FALSE	FALSE	FALSE
426.9	Cardiac pacemaker/device in situ	A3BΔ	0.487	1.628	0.169	4.05E-03	2092	383	1709	FALSE	FALSE	FALSE
427.21	Atrial fibrillation	A3BΔ	0.412	1.509	0.145	4.48E-03	2470	761	1709	FALSE	FALSE	FALSE
395.1	Nonrheumatic mitral valve disorders	A3BΔ	0.433	1.541	0.152	4.50E-03	3349	376	2973	FALSE	FALSE	FALSE
427.22	Atrial flutter	A3BΔ	0.616	1.852	0.220	5.09E-03	1897	188	1709	FALSE	FALSE	FALSE
425.1	Primary/intrinsic cardiomyopathies	A3BΔ	0.424	1.528	0.151	5.09E-03	3595	336	3259	FALSE	FALSE	FALSE
428.3	NA	A3BΔ	0.447	1.564	0.160	5.17E-03	3904	290	3614	FALSE	FALSE	FALSE
394	Chronic rheumatic disease of the heart valves	A3BΔ	0.586	1.797	0.211	5.46E-03	3126	153	2973	FALSE	FALSE	FALSE

Appendix J. ICD-9 Codes that map to each of the PheWAS codes discussed in chapter 3.

Heart Failure NOS (428.2)	**Heart Failure (428)**
428.1	398.91
428.9	428
	428.0
Nonrheumatic Aortic Valve Disorders (395.2)	428.00
	428.1
424.1	428.2
	428.20
Systolic/Diastolic Heart Failure (428.1)	428.21
	428.22
398.91	428.23
428.00	428.3
428.0	428.30
	428.31
Heart Valve Disorders (395)	428.32
	428.33
424	428.4
424.0	428.40
424.1	428.41
424.2	428.42
424.3	428.43
424.91	428.9
V42.2	
V43.3	

Appendix K. Top 10 PheWAS hits from deletions annotated as overlapping genes. No deletions appear twice in the top 10.

PheWAS Code	PheWAS Description	ESV	Beta	OR	SE	p	Total N	Case N	Control N
704.8	Other specified diseases of hair and hair follicles	esv2673257	1.271	3.563	0.267	2.02E-06	4428	51	4377
255.2	Adrenal hypofunction	esv2670116	0.930	2.535	0.200	3.13E-06	3994	88	3906
367.2	Astigmatism	esv2664484	1.894	6.649	0.420	6.38E-06	4429	21	4408
41.12	Methicillin resistant Staphylococcus aureus	esv2663159	0.804	2.235	0.179	7.27E-06	3223	121	3102
530.12	Ulcer of esophagus	esv2667558	1.662	5.271	0.374	8.65E-06	2368	27	2341
355	Complex regional/central pain syndrome	esv2666201	1.385	3.995	0.313	9.53E-06	3924	29	3895
574.3	Cholecystitis without cholelithiasis	esv2671657	0.823	2.277	0.187	1.12E-05	4218	81	4137
348.1	NA	esv2677892	0.312	1.366	0.072	1.48E-05	3615	463	3152
117	Mycoses	esv2664490	0.892	2.440	0.207	1.68E-05	3940	124	3816
272.1	Hyperlipidemia	esv2673043	0.209	1.232	0.049	1.90E-05	4203	2219	1984

Appendix L. Top 10 PheWAS hits from deletions not annotated as overlapping genes. Color-coding represents unique signals.

PheWAS Code	PheWAS Description	ESV	Beta	OR	SE	p	Total N	Case N	Control N
250.1	Type 1 diabetes	esv2673441	-0.519	0.595	0.078	2.70E-11	3040	472	2568
250.12	Type 1 diabetes with renal manifestations	esv2673441	-1.014	0.363	0.160	2.46E-10	2696	128	2568
250.11	Type 1 diabetes with ketoacidosis	esv2673441	-0.731	0.481	0.123	2.69E-09	2759	191	2568
290.11	Alzheimer's disease	esv2659215	2.213	9.142	0.404	4.42E-08	3153	25	3128
272	Disorders of lipoid metabolism	esv2672272	-0.362	0.696	0.067	5.82E-08	2595	1361	1234
272.1	Hyperlipidemia	esv2672272	-0.356	0.701	0.067	1.03E-07	2587	1353	1234
250.13	Type 1 diabetes with ophthalmic manifestations	esv2673441	-1.043	0.353	0.205	3.62E-07	2646	78	2568
250.7	Diabetic retinopathy	esv2673441	-0.589	0.555	0.120	9.33E-07	2754	186	2568
272.11	Hypercholesterolemia	esv2672272	-0.420	0.657	0.086	1.16E-06	1842	608	1234
415.2	Chronic pulmonary heart disease	esv2657961	0.816	2.262	0.173	2.29E-06	3483	72	3411

Appendix M. Top 15 hits for the SNP-Haplogroup term in haplogroup J.

PheWAS Code	PheWAS Description	Exome SNP	Haplogroup	Beta	SE	P	Case N	Control N	Total N
331	Other cerebral degenerations	exm689825_A	J	1.601	0.239	2.03E-11	287	15349	15636
705.8	Generalized hyperhidrosis	exm381852_T	J	1.793	0.331	6.09E-08	69	17721	17790
612	Breast conditions, congenital or relating to hormones	exm1036183_A	J	1.958	0.362	6.31E-08	102	16954	17056
724	Other disorders of back	exm631536_C	J	1.808	0.337	8.07E-08	80	15609	15689
337	Disorders of the autonomic nervous system	exm1649321_T	J	1.165	0.223	1.86E-07	132	15347	15479
303.4	Somatoform disorder	exm1623703_C	J	1.746	0.338	2.37E-07	56	13106	13162
337	Disorders of the autonomic nervous system	exm1013046_T	J	1.622	0.315	2.51E-07	131	15298	15429
612.2	Hypertrophy of breast (Gynecomastia)	exm1036183_A	J	1.955	0.385	3.92E-07	85	16954	17039
331	Other cerebral degenerations	exm855988_T	J	0.615	0.123	5.43E-07	287	15339	15626
573.5	Jaundice	exm869579_T	J	0.910	0.182	6.06E-07	206	15875	16081
250.14	Type 1 diabetic neuropathy	exm117637_G	J	1.203	0.245	8.83E-07	194	13545	13739
281.12	Vitamin B12 deficiency anemia	exm1013046_T	J	1.579	0.326	1.28E-06	160	12181	12341
331	Other cerebral degenerations	exm1496803_G	J	0.576	0.119	1.30E-06	287	15349	15636
427.22	Atrial flutter	exm1013725_G	J	0.723	0.152	1.97E-06	534	11561	12095
703	Diseases of nail	exm137983_A	J	1.496	0.315	2.01E-06	98	18641	18739

Appendix N. Top 15 PheWAS hits for the SNP-Haplogroup term in haplogroup H.

PheWAS Code	PheWAS Description	Exome SNP	Haplogroup	Beta	SE	P	Case N	Control N	Total N
595	Hydronephrosis	exm473294_G	H	0.399	0.071	2.11E-08	379	18311	18690
153.3	Cancer of the lower GI tract	exm1523019_T	H	0.563	0.107	1.36E-07	384	15722	16106
153.3	Cancer of the lower GI tract	exm1522975_A	H	0.563	0.107	1.37E-07	384	15723	16107
595	Hydronephrosis	exm200699_A	H	0.388	0.076	4.06E-07	379	18291	18670
110.13	Dermatophytosis of the body	exm620674_C	H	1.203	0.240	5.45E-07	63	17059	17122
318	Tobacco use disorder	exm564892_C	H	0.430	0.088	1.12E-06	1456	16682	18138
172.1	Melanoma	exm720018_T	H	0.449	0.092	1.21E-06	1039	15302	16341
595	Hydronephrosis	exm60331_G	H	0.445	0.092	1.27E-06	379	18315	18694
357	Inflammatory and toxic neuropathy	exm1043475_T	H	0.960	0.200	1.55E-06	100	17700	17800
357	Inflammatory and toxic neuropathy	exm1043475_T	H	0.960	0.200	1.55E-06	100	17700	17800
153	Colorectal cancer	exm1523019_T	H	0.377	0.079	1.77E-06	879	15722	16601
599.3	Dysuria	exm869611_A	H	0.402	0.084	1.79E-06	843	14099	14942
153	Colorectal cancer	exm1522975_A	H	0.377	0.079	1.81E-06	879	15723	16602
276.42	Alkalosis	exm620881_T	H	1.032	0.218	2.26E-06	121	12470	12591
642	Hypertension complicating pregnancy	exm1065549_C	H	1.253	0.272	4.23E-06	93	19562	19655

Appendix O. Mapping of ICD9 codes to PheWAS codes discussed in chapter 5. PheWAS code description and number are listed above the mapped ICD9 codes.

OtherCerebral Degenerations (331)
330
330.8
331.81
331.89
331.9
331
330.3
330.1
330.9
331.6
330.2
330
331.3
331.8
331.4
331.7
331.5
Tobacco Use Disorder (318)
649
305.11
649.04
305.12
649
305.1
305.13
649.02
649.03
305.1
649.01