UNIVERSITY OF CALIFORNIA

Los Angeles

Identification of rare-variant effect in complex human traits using whole-genome and whole-exome sequencing data

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Molecular Biology

by

Lingyu Zhan

2021

ABSTRAT OF THE DISSERTATION

Identification of rare-variant effect in complex human traits using whole-genome and whole-exome sequencing data

by

Lingyu Zhan

Doctor of Philosophy in Molecular Biology

University of California, Los Angeles, 2021

Professor Roel A Ophoff, Co-Chair

Professor Jae Hoon Sul, Co-Chair

For recent advancements in sequencing technologies, genetic information can be obtained from a large population at a relatively low cost. This provides an unprecedented opportunity to understand the role of genetic variability in association with complex human traits. One common strategy is to conduct genome-wide association studies to identify loci significantly associated with phenotypes of interest. However, the findings are usually limited to common variants with small effect sizes. Collectively, these identified loci can not fully explain the observed heritability, which is a problem commonly referred to as "the missing heritability." To uncover this problem, human genetic research has shifted more focus to other types of genetic variations, including rare variants, which is further capacitated and facilitated by the next-generation sequencing technique. These rare mutations are believed to harbor large effect sizes and, therefore to be one of the major contributors to complex traits.

Here, we describe our effort in analyzing the effect of rare variants in two complex human traits, Alzheimer's Disease and Tourette Syndrome, followed by conducting a genome-wide association study on human blood lipids. Exploring large whole-genome sequencing datasets, we have first demonstrated that rare variants were strongly associated with Alzheimer's Disease, neurofibrillary tangles, and age-related phenotypes within the endocytic pathway using a gene-set burden analysis framework. Subsequent gene-based analyses identified one AD-associated gene, *ANKRD13D*, and two e-Genes, *HLA-A* and *SLC26A7*. Leveraging bulk and scRNA-Seq data, we observed significant differential expression patterns in all three implicated genes. Secondly, we have explored a specific type of rare variants, *de novo* mutations, within Tourette Syndrome patients using a whole-exome sequencing trio dataset and identified a recurrent mutation in one gene, *FBN2*, previously implicated in TS. Comparing to the expected mutation rate, we demonstrated that the protein-truncating variants were enriched in probands. In addition, gene-set analysis displayed differential expression patterns across different tissue types and brain developmental stages. Lastly, we have performed a multi-population meta-analysis on blood lipid levels using electronic health records and genotyping information from the UCLA ATLAS database. We have observed genetic effects both specific to and shared across five different populations. Compared to previous large-scale GWASes, our results demonstrated consistent effect estimates while identifying one novel locus, rs72552763.

The dissertation of Lingyu Zhan is approved.

Jason Ernst

Nelson Freimer

Bogdan Pasaniuc

Matteo Pellegrini

Roel A Ophoff, Committee Co-chair

Jae Hoon Sul, Committee Co-Chair

University of California, Los Angeles

2021

DEDICATION

This dissertation is dedicated:

to my parents, grandparents, and all family members, without whom I would never have been able to pursue and succeed in my education, especially from a place thousands of miles away from home;

to my mentor Jae Hoon Sul, who provided great support, encouragement, patience, and mentorship throughout my journey of science;

to my co-chair, Roel Ophoff, who adopted me and provided the opportunity to start my next part of journey;

to my committee members, collaborators, and friends that have accompanied, enlighted, and warmed me throughout my graduate study.

Table of Contents

List of Figures and Tables

Figures

Tables

VITA

Education

- **University of California, Los Angeles (2016 – Now)**
    - **Ph.D.** in Molecular Biology Institute
    - **Home Area** in Gene Regulation, Epigenomics, and Transcriptomics (GREAT)
    - **Research focus**: bioinformatic and genetic analysis of complex traits
- **University of California, Los Angeles (2012 – 2016)**
    - **B.S.** in MCDB with specialization in Computing

Publications

- **First-Author**
    - **Zhan L.**, Li J., Jew B., Sul J., Rare variants in the endocytic pathway are associated with Alzheimer's disease, its related phenotypes, and functional consequences. PLoS Genet. 2021
- **Co-Author**
    - Liu Q., Bischof S, Harris CJ, Zhong Z, **Zhan L**. The characterization of Mediator 12 and 13 as conditional positive gene regulators in Arabidopsis. Nat Commun. 2020;11(1)
    - Li J., Jew B., **Zhan L.** et al. ForestQC: Quality control on genetic variants from next-generation sequencing data using random forest. PLoS Comput Biol. 2019;15(12)
    - Zhang Y, et al. Large-scale comparative epigenomics reveals hierarchical regulation of non-CG methylation in Arabidopsis. Proc Natl Acad Sci U S A. 2018;115(5)

Research Experience

- **Ph.D. Research**
    - Mentor Jae Hoon Sul (2017 summer to present)
    - Understand the contribution of rare variants to Alzheimer’s Disease using WGS/WES
    - Investigate *de novo* mutations in Tourette syndrome and comorbid disorders
    - GWAS of blood lipid phenotypes using EHR-linked biobank
    - Conduct quality control on large WGS/WES sequencing data, such as WGSPD
    - Analyze rare variants in Finnish WES data for clinical purposes
- **Lab Rotation**
    - Peter Tontonoz (2016 Fall): analyze immune-related genes in brown fat cells
    - Jingyi Jessica Li (2017 Winter): analyze RNA-Seq data
- **Undergraduate Research**
    - Steve Jacobsen (2015 winter to 2016 spring and 2017 spring): investigate epigenetic mediation in plants
    - Katsushi Arisaka (2014 winter to 2014 spring): analyze motion behaviors of C. elegans under various stringent conditions

Chapter 1 - Introduction

Genetic variation and association test

Genetic variability, or variation, refers to the difference in DNA sequence among individuals. A vast majority of the base pairs forming human DNA sequences are identical. In other words, genetic variants only represent about 0.1% to 0.4% of the human genome.(1) However, this small amount of differences is one of the major contributors to phenotypic variation. Therefore, one important goal of human genetic research is to understand whether and how any of the observed genetic variants is associated with a trait of interest. In return, the accumulation of our understanding forms the genetic architecture of the phenotype to which genetic studies are performed.

One of the most common types of genetic variants is single-nucleotide polymorphisms (SNPs) that are substitutions at a single nucleotide. On average, there are 10 to 11 million SNPs in the haploid human genome. One way to characterize SNPs is based on their population frequency which divides these variants into common and rare SNPs. Common SNPs are the variants that appear in a large proportion of a population and thus require a relatively small sample size to be identified. One proposed way to explore the effect of common variants was genome-wide association studies (GWASes). In this study design, the input genetic data are commonly obtained using microarray genotyping, a relatively low-cost technology. Common array designs are able to sequence 240K to 4M variants at once.(2) Subsequently, researchers then perform linear or logistic regression to estimate the association between each locus and the phenotype of interest. To date, over 5,700 GWASes have been conducted on over 3,300 traits and have yielded numerous significantly associated loci.(3) In addition to a deepened understanding of genetic architectures, insights into the genetic predisposition to common phenotypes, especially diseases, can substantially improve the health and life span in the general population.(4-6) However, the genotyping method and the study design limit the ability of GWASes within the scope of common variants that often have small effect sizes. Furthermore, these identified phenotype-associated common loci combinedly only explain a small amount of phenotypic variation, or genetic heritability, often addressed as the “missing heritability” problem.(7-9) Specifically, researchers have often observed a higher genetic heritability estimated from family/twin studies than that explainable from GWAS hits. This gap, therefore, points to the contribution of other genetic variations, such as rare variants.

Rare variants are low-frequency variants that appear only in a small fraction of the population and have been suggested to harbor large effect sizes. Hence, they are believed to be one of the major contributors to complex traits.(10-12) Previous genotyping methods, such as microarray genotyping, are difficult to efficiently and economically identify rare variants. But the recent advent in sequencing technologies, such as next-generation sequencing (NGS), enables us to explore almost the entire DNA sequence with a rapidly decreasing cost and sequencing time. Nowadays, many genetic studies have utilized sequencing data from NGS and have supported the notion that rare variants are one of the major risk factors. For instance, a recent study has demonstrated that rare variants in APP, PSEN1, PSEN2, and APOE genes were associated with the increased risk in late-onset AD.(13, 14) Two main NGS technologies, the whole-genome sequencing (WGS) and whole-exome sequencing (WES), have also led to the discovery of numerous rare variants accountable to complex traits.(15-19) However, the nature of rare variants, their low occurrence rate, continues to impede efficient identification of significant rare loci that are associated with phenotypes, especially disease traits.

Methods of rare-variant analysis —— rare-variant burden analysis

A few methods have been proposed to tackle this problem. One way is to combine the small individual effects of rare variants into a large aggregated effect, called a burden, which is then used to test for association. As opposed to traditional GWASes that test individual common variants, this burden method avoids the low statistical power limited by minor allele frequency of the tested loci and benefits from the wide range of potential grouping choices. In other words, when we aggregate rare-variant effects, we can either choose an arbitrary set of variants or collect those participating in the same biological context, such as genes or pathways. The latter is important as it bridges the gap between genetic variation and phenotype and points to the potential biological pathways involved in this process. Specifically, the rare-variant effects identified through this method involve can be directly understood in biological functions and thus facilitate further genetic and translational clinical studies. Furthermore, when we choose a set of genes involved in biological pathways, especially those previously implicated in certain traits, this can additionally reduce the burden of multiple testing that is commonly observed in gene-based rare-variant burden analysis and therefore harbor a larger detecting power for the rare-variant contribution to traits or disease susceptibility.

Application of rare-variant burden analysis to Alzhermer's disease

One complex disease on which genetic studies commonly focus is Alzheimer's Disease (AD). AD is a destructive and irreversible neurodegenerative disorder, predominantly targeting the elderly.(20) It accounts for 60 - 70% of dementia cases, characteristic of progressive disintegration of cognitive functions, language ability, and memory loss (20, 21) and has been reported by the National Institute on Aging (NIA) as the 6th, potentially the 3rd, leading cause of death in the US. However, while few studies were able to adequately depict the underlying biological pathways and conduct clinical trials with a satisfactory outcome,(22) studies have shown a substantial genetic component with an estimated heritability of 58 - 79%(23) and multiple modulating genes, including the strongest risk factor, *APOE*. Recent GWASes have identified over 50 risk loci accounting for over 33% of the overall estimated heritability(22-28). One recent study on AD has also suggested an oligogenic common variant architecture(29) where these risk loci fall in several known biological pathways, including the lipid metabolism, the immune system/response, the endocytic, the amyloid/tau processing, and the microglia-related pathways.(27, 28, 30, 31) This underlying genetic architecture poses a great subject to apply gene-set analysis. Given that little is know about the effect of rare variants within these biological contexts, we, therefore, have appropriated the benefit of gene-set analysis under a rare-variant burden framework. In particular, we selected the endocytic pathway that has been implicated in many clinical and genetic studies with multiple GWAS risk loci found, including *BIN1*, *CD2AP*, *PICALM*, *PLD3*, *EPHA1*, and *SORL1*(32-35), and meta-analyzed the rare-variant effects in this pathway across three large-scale WGS datasets. We showed that the effect of rare-variant could be identified within known pathways using gene-set rare-variant burden method with large WGS datasets. In particular, we found that the rare variants within the endocytic pathway were strongly associated with AD, neurofibrillary tangles (NFTs) (a histopathological indicator of AD progression)(36-38), and age-related phenotypes (age at onset and age of death). Furthermore, within the scope of one biological pathway, we also showed that single genes with large effect sizes could be filtered out when combining various related phenotypes, including gene expression data.

Methods of rare-variant analysis —— *de novo* mutations

Another commonly used method focuses on a specific type of rare variants, *de novo* mutations (DNMs). This refers to a genetic alteration that appears in a child but not the parents due to mutations within germ cells or the fertilized egg. The identification and research in DNMs have been limited in traditional genetic studies that mostly focus on inherited variations. Due to the recent advancement of unbiased WGS and WES techniques, we are now able to study the effect of DNMs in relationship to complex traits, especially diseases. These mutations represent an extreme version of rare variants as they have not been put under evolutionary selection pressure.(39, 40) As a result, on average, they are more deleterious and may confer more information of the disease susceptibility, such as the gene function they disrupt. Studies have shown that the mean germline *de novo* SNVs per individual is around 74(41) and have revealed major risk contributions from these DNMs to complex diseases including Autism Spectrum Disorder, Schinzel–Giedion syndrome, Kabuki syndrome, Bohring–Opitz syndrome, Proteus syndrome, intellectual disability, and schizophrenia.(42-53) Compared to traditional rare-variant analysis, one difference in DNM analysis is that we need to collect sequencing data from a proband and the two parents, namely a trio family, in order to pinpoint non-inherited genetic variations. This is challenging because it requires sequencing more samples with higher accuracy to effectively identify extremely rare variations. Facilitated by current NGS approaches, especially WES, nowadays, more genetic studies have been able to work with DNMs. There are a few steps to analyze DNMs. The first step is to simply pinpoint the location of DNMs, such as within a gene or known regulatory regions. The second one is to compare the observed DNM rates to the theoretical rates(54) at the gene level, partitioned by the types of mutations. And the third step is to compare the observed rates within probands to those computed from healthy samples. These comparison steps can reveal the underlying enrichment of the deleterious DNMs within the affected samples and thus point to their effect on disease susceptibility.

Analysis of *de novo* mutations in Tourette Syndrome

Recent studies on Tourette Syndrome (TS) have shown promising results using DNMs.(55, 56) TS is an early onset neurodevelopmental disorder with an estimated average prevalence rate of 0.6%.(57-62) The characteristic symptoms are chronic motor and vocal tic and show a higher prevalence rate in males. It is highly comorbid with other psychiatric disorders, including obsessive-compulsive disorder, attention deficit, and hyperactivity disorder, autism spectrum disorder (ASD).(60, 63-66) Many studies on these comorbid disorders have also indicated the contribution of DNMs.(50, 67, 68) Previous studies in our group have demonstrated enrichment of *de novo* Loss-of-Function (LoF) and damaging missense coding variants in TS probands.(55, 56) In this work, collaborating with The Tourette Association of America International Consortium for Genetics (TAAICG) and the Tourette International Collaborative for Genetics (TICGen), we extend our previous work to over 1200 TS trio families and identify additional TS susceptibility DNMs thought WES. Using nearly 900 trios currently available, we showed that DNMs could be efficiently called and analyzed in TS probands. We identified recurrent mutations in genes previously implicated in TS and demonstrated that the protein-truncating DNMs in TS probands were enriched in three genes while missense and synonymous DNMs are not, compared to the theoretical mutation rates of protein-truncating variants. Exploring functional enrichment aspects, we pointed out a differential regulation pattern throughout brain development tissues and stages. In addition, we are expecting to receive more samples to gain greater detecting power, as well as WES data of healthy individuals, which will allow us to analyze the effect of DNMs with greater statistical power.

Heterogeneity of genetic effects across different populations

Another problem of GWASes, on either common or rare variants, is sample heterogeneity. The statistical design of GWASes assumes that all samples analyzed have the same genetic background, namely the same population. If the assumption is violated, studies will often have a decreased power of detecting trait-associating loci or even spurious signals in extreme cases. A common strategy to increase detecting power in complex trait analysis is to use a large sample size, often through collaboration and a combination of smaller studies.(69) However, it becomes problematic as the increased sample size introduces sample heterogeneity and fails to achieve its original goal. A recent study has shown that simply increasing sample size causes the p-values of genetic effects in large-scale studies, such as those with over 100,000 samples, to increase rather than decrease.(70) Nowadays, researchers have increasingly focused on this issue of genetic heterogeneity and perform GWASes on specific populations.(71-76) However, as collecting large samples is already a difficult task in traditional GWASes, it is even more challenging to collect samples from various ancestral backgrounds with a sufficiently large number within each population. The emergence of large-scale biobank linked with electronic health records (EHR) databases has helped resolve this issue. The EHR-linked biobanks feature several advantages: 1. Most of them contain a large number of samples (over 50,000) with diverse ancestral backgrounds; 2. A wide range of phenotypes are readily available in the EHR database, enabling association analysis on multiple phenotypes at once; 3. They provide the potential to perform longitudinal analysis that is difficult with traditional GWAS design.(77, 78) Several examples of large-scale EHR-linked biobanks are UK Biobank, Million Veteran Program, and DeCODE Genetics. Many efforts have been made to analyze phenotypes provided by these EHR-linked biobanks.(79-86)

Multi-population meta-analysis of blood lipid phenotypes

One of the newly established large-scale EHR-linked biobanks is the UCLA ATLAS Precision Health Biobank. At the time of our study, this biobank has gathered genotyping data of over 26,000 individuals with diverse ancestral backgrounds and matching EHR data, and the sample size has been continued to grow. This dataset serves as a great platform to study genetic homogeneity and heterogeneity across different populations. As a result, we have selected one of the popular phenotypes, blood lipid concentrations, and performed population-specific GWAS, followed by meta-analysis. These plasma lipid levels are linked to common diseases, such as type 2 diabetes, fatty liver disease, and especially atherosclerotic cardiovascular disease.(87-91) Various types of blood lipid levels (high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, triglycerides, and total cholesterol) all have been reported with relatively high heritability, ranging from 33% to 51%.(92, 93) A large-scale GWAS by the Global Lipids Genetics Consortium (GLGC) has identified 444 independent hits.(94) However, as the UCLA ATLAS dataset has provided only microarray genotyping data, we are not able to perform analysis on rare variants as described in other parts of our work. Nonetheless, through population-specific GWAS, we have shown that there were indeed large heterogeneous genetic effects for blood lipid phenotypes across different populations. By meta-analyzing results from single populations, we have additionally demonstrated shared genetic effects common to all populations and identified novel loci significantly associated with triglyceride levels. We showed that our findings were consistent with previous studies on European and minority populations even though our current sample size was limited. In the future, once WES and WGS data are available, we will explore the effect of rare variants on the blood lipid levels in a population-specific manner and attempt to identify colocalized risk between common and rare variants, therefore providing a better understanding of the genetic architecture of human plasma lipid phenotypes. This will provide important insights into future genetic and clinical studies on blood-lipid-related diseases, especially atherosclerotic cardiovascular disease.

Figures

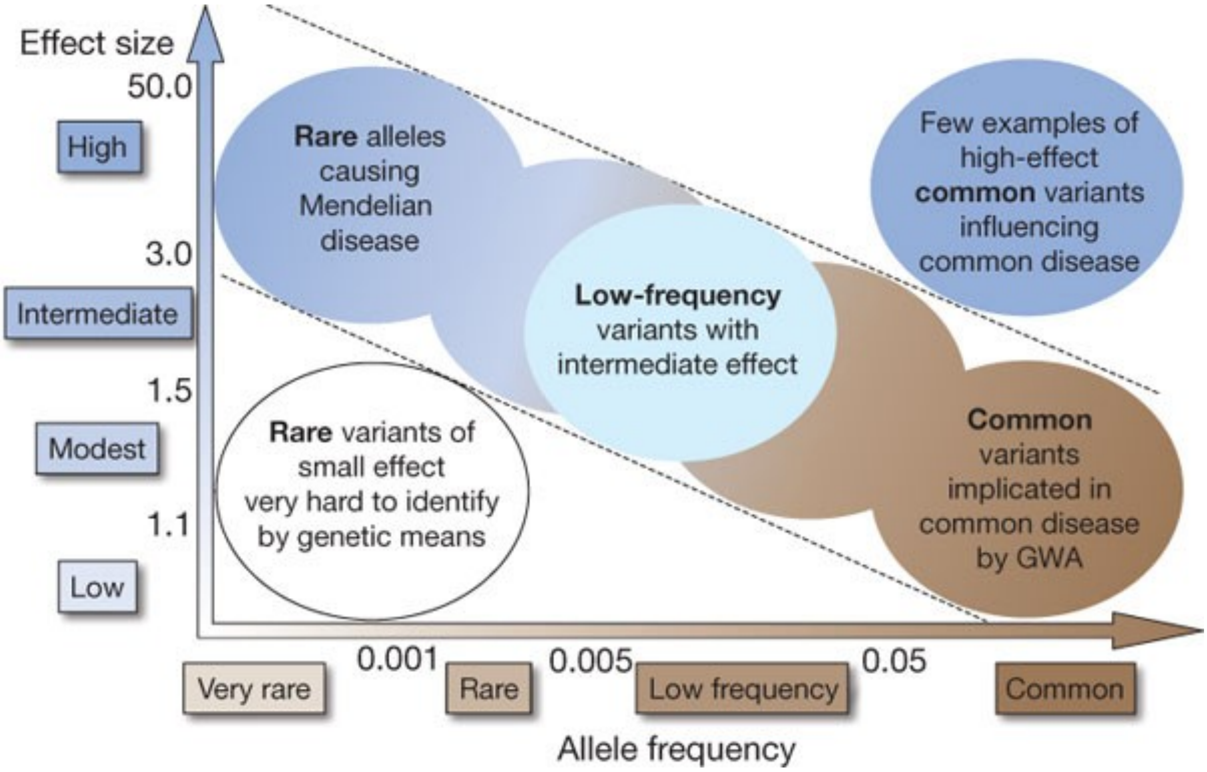

Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Few examples of high-effect common variants influencing common disease
Low-frequency variants with intermediate effect
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
Very rare
0.001
Rare
0.005
Low frequency
0.05
Common
Allele frequency

Figure 1-1. Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio).(8)

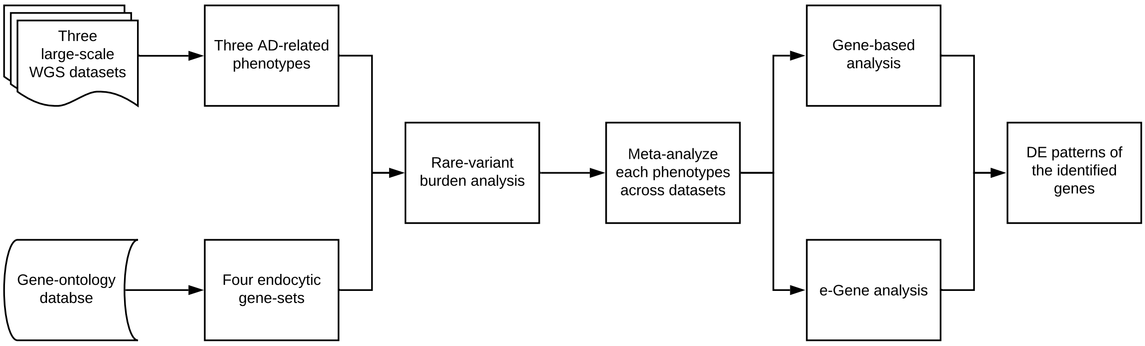

Three large-scale WGS datasets
Three AD-related phenotypes
Gene-ontology databse
Four endocytic gene-sets
Rare-variant burden analysis
Meta-analyze each phenotypes across datasets
Gene-based analysis
e-Gene analysis
DE patterns of the identified genes

Figure 1-2. Workflow of gene-set rare-variant burden analysis and its application in AD.

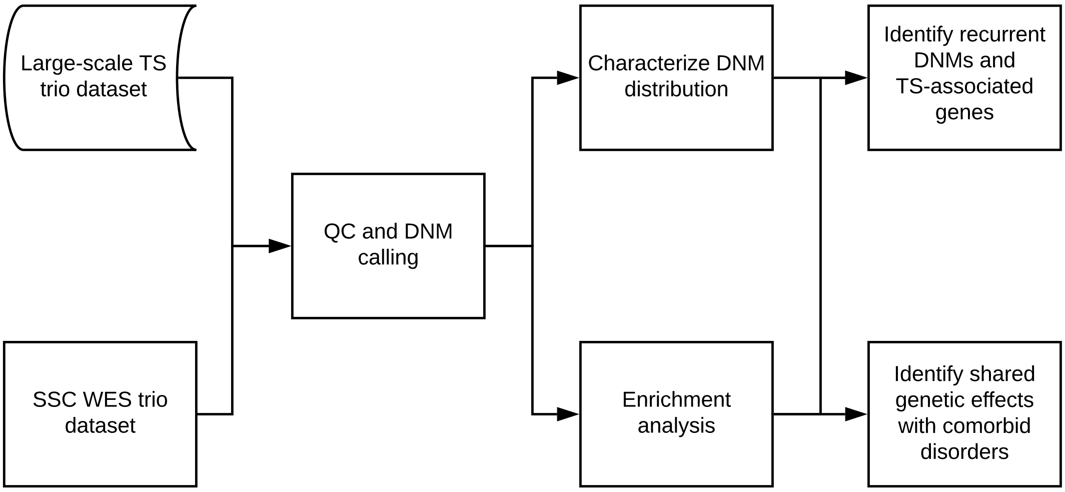

Large-scale TS trio dataset
SSC WES trio dataset
QC and DNM calling
Characterize DNM distribution
Enrichment analysis
Identify recurrent DNMs and TS-associated genes
Identify shared genetic effects with comorbid disorders

Figure 1-3. Workflow of DNM analysis and its application in TS.

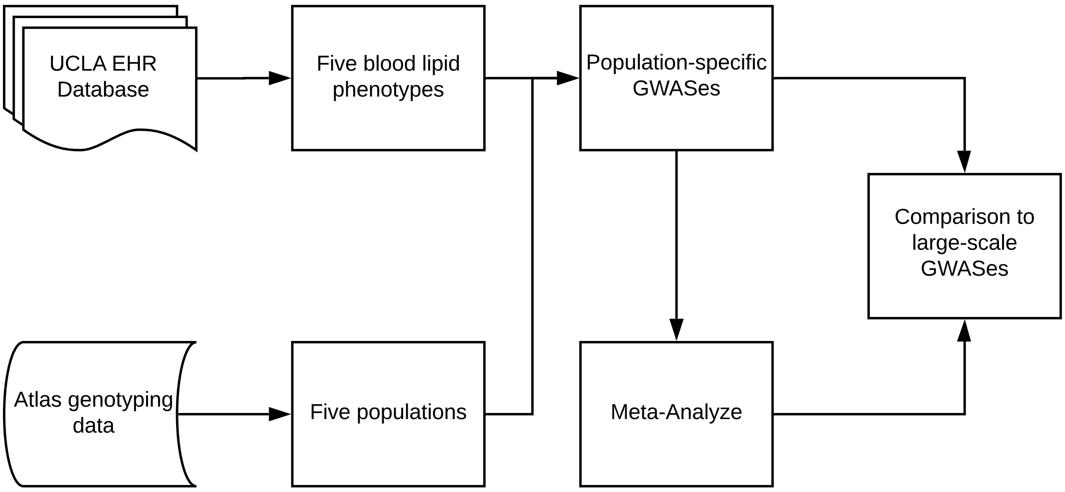

UCLA EHR Database
Five blood lipid phenotypes
Population-specific GWASes
Comparison to large-scale GWASes
Atlas genotyping data
Five populations
Meta-Analyze

Figure 1-4. Identify heterogeneous and shared genetic effects in different populations using blood lipid phenotype

Reference

1. Jorde LB, Wooding SP. Genetic variation, classification and 'race'. Nat Genet. 2004;36(11 Suppl):S28-33.

2. Verlouw JAM, Clemens E, de Vries JH, Zolk O, Verkerk A, Am Zehnhoff-Dinnesen A, et al. A comparison of genotyping arrays. Eur J Hum Genet. 2021.

3. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nature Reviews Methods Primers. 2021;1(1):59.

4. Sierra F, Hadley E, Suzman R, Hodes R. Prospects for life span extension. Annu Rev Med. 2009;60:457-69.

5. Olshansky SJ, Perry D, Miller RA, Butler RN. Pursuing the longevity dividend: scientific goals for an aging world. Ann N Y Acad Sci. 2007;1114:11-3.

6. Aging Well in the 21st Century: Strategic Directions for Research on Aging.

7. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, et al. Genome partitioning of genetic variation for complex traits using common SNPs. Nat Genet. 2011;43(6):519-25.

8. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009;461(7265):747-53.

9. Maher B. Personal genomes: The case of the missing heritability. Nature. 2008;456(7218):18-21.

10. Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev. 2009;19(3):212-9.

11. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. 2010;11(6):415-25.

12. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet. 2008;40(6):695-701.

13. Steinthorsdottir V, Thorleifsson G, Sulem P, Helgason H, Grarup N, Sigurdsson A, et al. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. Nat Genet. 2014;46(3):294-8.

14. Cruchaga C, Haller G, Chakraverty S, Mayo K, Vallania FL, Mitra RD, et al. Rare variants in APP, PSEN1 and PSEN2 increase risk for AD in late-onset Alzheimer's disease families. PLoS One. 2012;7(2):e31039.

15. Yu TW, Chahrour MH, Coulter ME, Jiralerspong S, Okamura-Ikeda K, Ataman B, et al. Using whole-exome sequencing to identify inherited causes of autism. Neuron. 2013;77(2):259-73.

16. Sims R, van der Lee SJ, Naj AC, Bellenguez C, Badarinarayan N, Jakobsdottir J, et al. Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. Nat Genet. 2017;49(9):1373-84.

17. Lange LA, Hu Y, Zhang H, Xue C, Schmidt EM, Tang ZZ, et al. Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. Am J Hum Genet. 2014;94(2):233-45.

18. Gudmundsson J, Sulem P, Gudbjartsson DF, Masson G, Agnarsson BA, Benediktsdottir KR, et al. A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. Nat Genet. 2012;44(12):1326-9.

19. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, et al. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet. 2015;47(5):435-44.

20. Mendez MF. Early-onset Alzheimer's disease: nonamnestic subtypes and type 2 AD. Arch Med Res. 2012;43(8):677-85.

21. Burns A, Iliffe S. Dementia. BMJ. 2009;338:b75.

22. Shen L, Jia J. An Overview of Genome-Wide Association Studies in Alzheimer's Disease. Neurosci Bull. 2016;32(2):183-90.

23. Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, et al. Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry. 2006;63(2):168-74.

24. Ridge PG, Mukherjee S, Crane PK, Kauwe JS, Alzheimer's Disease Genetics C. Alzheimer's disease: analyzing the missing heritability. PLoS One. 2013;8(11):e79771.

25. Naj AC, Jun G, Beecham GW, Wang LS, Vardarajan BN, Buros J, et al. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nat Genet. 2011;43(5):436-41.

26. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013;45(12):1452-8.

27. Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing. Nat Genet. 2019;51(3):414-30.

28. Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat Genet. 2019;51(3):404-13.

29. Zhang Q, Sidorenko J, Couvy-Duchesne B, Marioni RE, Wright MJ, Goate AM, et al. Risk prediction of late-onset Alzheimer's disease implies an oligogenic architecture. Nat Commun. 2020;11(1):4799.

30. Schwartzentruber J, Cooper S, Liu JZ, Barrio-Hernandez I, Bello E, Kumasaka N, et al. Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. Nat Genet. 2021;53(3):392-402.

31. Bellenguez C, Küçükali F, Jansen I, Andrade V, Moreno-Grau S, Amin N, et al. New insights on the genetic etiology of Alzheimer’s and related dementia. medRxiv. 2020:2020.10.01.20200659.

32. Van Acker ZP, Bretou M, Annaert W. Endo-lysosomal dysregulations and late-onset Alzheimer's disease: impact of genetic risk factors. Mol Neurodegener. 2019;14(1):20.

33. Tiwari S, Atluri V, Kaushik A, Yndart A, Nair M. Alzheimer's disease: pathogenesis, diagnostics, and therapeutics. Int J Nanomedicine. 2019;14:5541-54.

34. Karch CM, Goate AM. Alzheimer's disease risk genes and mechanisms of disease pathogenesis. Biol Psychiatry. 2015;77(1):43-51.

35. Heckmann BL, Teubner BJW, Tummers B, Boada-Romero E, Harris L, Yang M, et al. LC3-Associated Endocytosis Facilitates beta-Amyloid Clearance and Mitigates Neurodegeneration in Murine Alzheimer's Disease. Cell. 2019;178(3):536-51 e14.

36. Braak H, Braak E, Bohl J. Staging of Alzheimer-related cortical destruction. Eur Neurol. 1993;33(6):403-8.

37. Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol. 1991;82(4):239-59.

38. Abner EL, Kryscio RJ, Schmitt FA, Santacruz KS, Jicha GA, Lin Y, et al. "End-stage" neurofibrillary tangle pathology in preclinical Alzheimer's disease: fact or fiction? J Alzheimers Dis. 2011;25(3):445-53.

39. Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. Nat Rev Genet. 2007;8(8):610-8.

40. Crow JF. The origins, patterns and implications of human spontaneous mutation. Nat Rev Genet. 2000;1(1):40-7.

41. Conrad DF, Keebler JE, DePristo MA, Lindsay SJ, Zhang Y, Casals F, et al. Variation in genome-wide mutation rates within and between human families. Nat Genet. 2011;43(7):712-4.

42. Xu B, Roos JL, Dexheimer P, Boone B, Plummer B, Levy S, et al. Exome sequencing supports a de novo mutational paradigm for schizophrenia. Nat Genet. 2011;43(9):864-8.

43. Vissers LE, de Ligt J, Gilissen C, Janssen I, Steehouwer M, de Vries P, et al. A de novo paradigm for mental retardation. Nat Genet. 2010;42(12):1109-12.

44. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature. 2012;485(7397):237-41.

45. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature. 2012;485(7397):246-50.

46. O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, et al. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat Genet. 2011;43(6):585-9.

47. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet. 2010;42(9):790-3.

48. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, et al. Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature. 2012;485(7397):242-5.

49. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, et al. De novo gene disruptions in children on the autistic spectrum. Neuron. 2012;74(2):285-99.

50. Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, Krumm N, Levy D, et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature. 2014;515(7526):216-21.

51. Hoischen A, van Bon BW, Rodriguez-Santiago B, Gilissen C, Vissers LE, de Vries P, et al. De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. Nat Genet. 2011;43(8):729-31.

52. Hoischen A, van Bon BW, Gilissen C, Arts P, van Lier B, Steehouwer M, et al. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat Genet. 2010;42(6):483-5.

53. Girard SL, Gauthier J, Noreau A, Xiong L, Zhou S, Jouan L, et al. Increased exonic de novo mutation rate in individuals with schizophrenia. Nat Genet. 2011;43(9):860-3.

54. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014;46(9):944-50.

55. Willsey AJ, Fernandez TV, Yu D, King RA, Dietrich A, Xing J, et al. De Novo Coding Variants Are Strongly Associated with Tourette Disorder. Neuron. 2017;94(3):486-99 e9.

56. Wang S, Mandell JD, Kumar Y, Sun N, Morris MT, Arbelaez J, et al. De Novo Sequence and Copy Number Variants Are Strongly Associated with Tourette Disorder and Implicate Cell Polarity in Pathogenesis. Cell Rep. 2018;24(13):3441-54 e12.

57. Scharf JM, Miller LL, Gauvin CA, Alabiso J, Mathews CA, Ben-Shlomo Y. Population prevalence of Tourette syndrome: a systematic review and meta-analysis. Mov Disord. 2015;30(2):221-8.

58. Robertson MM. The prevalence and epidemiology of Gilles de la Tourette syndrome. Part 1: the epidemiological and prevalence studies. J Psychosom Res. 2008;65(5):461-72.

59. Knight T, Steeves T, Day L, Lowerison M, Jette N, Pringsheim T. Prevalence of tic disorders: a systematic review and meta-analysis. Pediatr Neurol. 2012;47(2):77-90.

60. Charania SN, Danielson ML, Claussen AH, Lebrun-Harris LA, Kaminski JW, Bitsko RH. Bullying Victimization and Perpetration Among US Children with and Without Tourette Syndrome. J Dev Behav Pediatr. 2021.

61. Centers for Disease C, Prevention. Prevalence of diagnosed Tourette syndrome in persons aged 6-17 years - United States, 2007. MMWR Morb Mortal Wkly Rep. 2009;58(21):581-5.

62. Bitsko RH, Holbrook JR, Visser SN, Mink JW, Zinner SH, Ghandour RM, et al. A national profile of Tourette syndrome, 2011-2012. J Dev Behav Pediatr. 2014;35(5):317-22.

63. Hirschtritt ME, Lee PC, Pauls DL, Dion Y, Grados MA, Illmann C, et al. Lifetime prevalence, age of risk, and genetic relationships of comorbid psychiatric disorders in Tourette syndrome. JAMA Psychiatry. 2015;72(4):325-33.

64. Ghanizadeh A, Mosallaei S. Psychiatric disorders and behavioral problems in children and adolescents with Tourette syndrome. Brain Dev. 2009;31(1):15-9.

65. Eapen V, Cavanna AE, Robertson MM. Comorbidities, Social Impact, and Quality of Life in Tourette Syndrome. Front Psychiatry. 2016;7:97.

66. Cravedi E, Deniau E, Giannitelli M, Xavier J, Hartmann A, Cohen D. Tourette syndrome and other neurodevelopmental disorders: a comprehensive review. Child Adolesc Psychiatry Ment Health. 2017;11:59.

67. Cappi C, Oliphant ME, Peter Z, Zai G, Conceicao do Rosario M, Sullivan CAW, et al. De Novo Damaging DNA Coding Mutations Are Associated With Obsessive-Compulsive Disorder and Overlap With Tourette's Disorder and Autism. Biol Psychiatry. 2020;87(12):1035-44.

68. Buja A, Volfovsky N, Krieger AM, Lord C, Lash AE, Wigler M, et al. Damaging de novo mutations diminish motor skills in children on the autism spectrum. Proc Natl Acad Sci U S A. 2018;115(8):E1859-E66.

69. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013;45(11):1274-83.

70. Kulminski AM, Loika Y, Culminskaya I, Arbeev KG, Ukraintseva SV, Stallard E, et al. Explicating heterogeneity of complex traits has strong potential for improving GWAS efficiency. Sci Rep. 2016;6:35390.

71. Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M. Genome-wide association studies in diverse populations. Nat Rev Genet. 2010;11(5):356-66.

72. Martin AR, Lin M, Granka JM, Myrick JW, Liu X, Sockell A, et al. An Unexpectedly Complex Architecture for Skin Pigmentation in Africans. Cell. 2017;171(6):1340-53 e14.

73. Hoffmann TJ, Van Den Eeden SK, Sakoda LC, Jorgenson E, Habel LA, Graff RE, et al. A large multiethnic genome-wide association study of prostate cancer identifies novel risk variants and substantial ethnic differences. Cancer Discov. 2015;5(8):878-91.

74. Hindorff LA, Bonham VL, Brody LC, Ginoza MEC, Hutter CM, Manolio TA, et al. Prioritizing diversity in human genomics research. Nat Rev Genet. 2018;19(3):175-85.

75. Giri A, Hellwege JN, Keaton JM, Park J, Qiu C, Warren HR, et al. Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. Nat Genet. 2019;51(1):51-62.

76. Akiyama M, Okada Y, Kanai M, Takahashi A, Momozawa Y, Ikeda M, et al. Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. Nat Genet. 2017;49(10):1458-67.

77. Wolford BN, Willer CJ, Surakka I. Electronic health records: the next wave of complex disease genetics. Hum Mol Genet. 2018;27(R1):R14-R21.

78. Collins R. What makes UK Biobank special? Lancet. 2012;379(9822):1173-4.

79. Nagai A, Hirata M, Kamatani Y, Muto K, Matsuda K, Kiyohara Y, et al. Overview of the BioBank Japan Project: Study design and profile. J Epidemiol. 2017;27(3S):S2-S8.

80. McCarty CA, Wilke RA, Giampietro PF, Wesbrook SD, Caldwell MD. Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank. Per Med. 2005;2(1):49-79.

81. Li R, Duan R, Kember RL, Rader DJ, Damrauer SM, Moore JH, et al. A regression framework to uncover pleiotropy in large-scale electronic health record data. J Am Med Inform Assoc. 2019;26(10):1083-90.

82. Kvale MN, Hesselson S, Hoffmann TJ, Cao Y, Chan D, Connell S, et al. Genotyping Informatics and Quality Control for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. Genetics. 2015;200(4):1051-60.

83. Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet. 2011;12(6):417-28.

84. Fritsche LG, Gruber SB, Wu Z, Schmidt EM, Zawistowski M, Moser SE, et al. Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. Am J Hum Genet. 2018;102(6):1048-61.

85. Cho SY, Hong EJ, Nam JM, Han B, Chu C, Park O. Opening of the national biobank of Korea as the infrastructure of future biomedical science in Korea. Osong Public Health Res Perspect. 2012;3(3):177-84.

86. Banda Y, Kvale MN, Hoffmann TJ, Hesselson SE, Ranatunga D, Tang H, et al. Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic

Epidemiology Research on Adult Health and Aging (GERA) Cohort. Genetics. 2015;200(4):1285-95.

87. Cardiovascular diseases (CVDs) fact sheets [Available from: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

88. Qi Q, Liang L, Doria A, Hu FB, Qi L. Genetic predisposition to dyslipidemia and type 2 diabetes risk in two prospective cohorts. Diabetes. 2012;61(3):745-52.

89. Oresic M, Hyotylainen T, Kotronen A, Gopalacharyulu P, Nygren H, Arola J, et al. Prediction of non-alcoholic fatty-liver disease and liver fat content by serum molecular lipids. Diabetologia. 2013;56(10):2266-74.

90. Kannel WB, Dawber TR, Kagan A, Revotskie N, Stokes J, 3rd. Factors of risk in the development of coronary heart disease--six year follow-up experience. The Framingham Study. Ann Intern Med. 1961;55:33-50.

91. Emerging Risk Factors C, Di Angelantonio E, Sarwar N, Perry P, Kaptoge S, Ray KK, et al. Major lipids, apolipoproteins, and risk of vascular disease. JAMA. 2009;302(18):1993-2000.

92. Weiss LA, Pan L, Abney M, Ober C. The sex-specific genetic architecture of quantitative traits in humans. Nat Genet. 2006;38(2):218-22.

93. van Dongen J, Willemsen G, Chen WM, de Geus EJ, Boomsma DI. Heritability of metabolic syndrome traits in a large population-based sample. J Lipid Res. 2013;54(10):2914-23.

94. Liu DJ, Peloso GM, Yu H, Butterworth AS, Wang X, Mahajan A, et al. Exome-wide association study of plasma lipids in >300,000 individuals. Nat Genet. 2017;49(12):1758-66.

Chapter 2 - Rare variants in the endocytic pathway are associated with Alzheimer’s disease, its related phenotypes, and functional consequences

Rare variants in the endocytic pathway are associated with Alzheimer's disease, its related phenotypes, and functional consequences

Rare variants, endocytic pathway, and Alzheimer's disease

Lingyu Zhan[1]*, Jiajin Li[2], Brandon Jew[3], Jae Hoon Sul[4]*

1. **Molecular Biology Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, USA**
2. **Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, USA**
3. **Interdepartmental Program in Bioinformatics, University of California, Los Angeles, Los Angeles, California, USA**
4. **Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, Los Angeles, California, USA**

* jaehoonsul@mednet.ucla.edu **(JHS);** * zhanly812@g.ucla.edu **(LZ)**

Abstract

Late-onset Alzheimer's disease (LOAD) is the most common type of dementia causing irreversible brain damage to the elderly and presents a major public health challenge. Clinical research and genome-wide association studies have suggested a potential contribution of the endocytic pathway to AD, with an emphasis on common loci. However, the contribution of rare variants in this pathway to AD has not been thoroughly investigated. In this study, we focused on the effect of rare variants on AD by first applying a rare-variant gene-set burden analysis using genes in the endocytic pathway on over 3,000 individuals with European ancestry from three large whole-genome sequencing (WGS) studies. We identified significant associations of rare-variant burden within the endocytic pathway with AD, which were successfully replicated in independent datasets. We further demonstrated that this endocytic rare-variant enrichment is associated with neurofibrillary tangles (NFTs) and age-related phenotypes, increasing the risk of obtaining severer brain damage, earlier age-at-onset, and earlier age-of-death. Next, by aggregating rare variants within each gene, we sought to identify single endocytic genes associated with AD and NFTs. Careful examination using

NFTs revealed one significantly associated gene, *ANKRD13D*. To identify functional associations, we integrated bulk RNA-Seq data from over 600 brain tissues and found two endocytic expression genes (eGenes), *HLA-A* and *SLC26A7*, that displayed significant influences on their gene expressions. Differential expressions between AD patients and controls of these three identified genes were further examined by incorporating scRNA-Seq data from 48 post-mortem brain samples and demonstrated distinct expression patterns across cell types. Taken together, our results demonstrated strong rare-variant effect in the endocytic pathway on AD risk and progression and functional effect of gene expression alteration in both bulk and single-cell resolution, which may bring more insight and serve as valuable resources for future AD genetic studies, clinical research, and therapeutic targeting.

Author summary

Late-onset Alzheimer's disease (LOAD) is the most common type of dementia and a leading cause of death in the world. Clinical and genetic studies have suggested the potential contribution of the cellular transportation pathway to AD with an emphasis on common variants. In this study, we investigated the effect of rare variants within the cellular transportation pathway and examined three large datasets with over 3,000 individuals with European ancestry. We reported enrichment of rare deleterious variants in the cellular transportation pathway in AD patients from all three datasets. We also observed an elevation of rare deleterious variants in this pathway was associated with individuals with severer brain damages (AD progression), earlier age-at-onset, and earlier age-of-death. By aggregating rare variants in each gene from the cellular transportation pathway, we revealed one gene in which rare variants were significantly associated with the progression of AD. By integrating gene expression data from brain tissues, we identified two additional genes whose rare-variant effect displayed significant influences on gene expression. Taken together, our results

demonstrated that rare-variant effect in the cellular transportation pathway is strongly associated with the risk and the progression of AD, which may serve as future clinical and therapeutic targets.

Introduction

Alzheimer's disease (AD) is a destructive and irreversible neurodegenerative disorder, predominantly targeting the elderly.[1] It accounts for 60 - 70% of dementia cases, characteristic of progressive disintegration of cognitive functions, language ability, and memory loss.[1,2] Late-onset Alzheimer’s Disease (LOAD) is a subcategory of AD that appears in persons aged 65 years or older, showing a greater incidence rate as age increases.[3] As the population of Americans age 65 and beyond is expected to reach 88 million by 2050, the number of new AD cases is predicted to double and the prevalence rate to quadruple.[4,5]

AD is known to have a substantial genetic component with multiple modulating genes. One of the strongest risk factors for LOAD is *APOE*. Recent GWASs have identified over 50 risk loci accounting for, together with all common SNPs, over 33% of the overall estimated heritability[6-12] that cohered into three major AD-related biological pathways: the cholesterol metabolism pathway, the immune response pathway, and the endocytic pathway.[13] While AD studies have mostly focused on the effect of common variants, such as in the lipid metabolism and immune system/response pathways implicated in recent GWASes, rare variants in genes related to these pathways have not yet been thoroughly investigated.[11,12,14-20] Among these implicated pathways, the endocytic pathway has been identified as one of the most prominent targets, where the earliest morphological changes can be observed as endosome enlargement in post-mortem brains from sporadic AD patients, as well as in some familial cases.[21,22] This phenomenon can be viewed as nearly diagnostic precision and served as blood-cellular markers.[23,24] These findings have also been supported by a recent genetic study showing the enrichment in clathrin-mediated/early endocytosis[25] and clinical research on the facilitation of Aβ clearance by LC3-associated endocytosis.[26] Previous studies using common variants have also identified several risk loci in the endocytic pathway, including *BIN1*, *PICALM*, *CD2AP*, *EPHA1*, and *SORL1*.[27]

However, despite being one of the histological hallmarks of AD, few studies have examined the effect of rare variants within this endocytic pathway on AD pathogenic progression.[13] It is thus of interest to study the rare-variant effect on AD in this pathway. One major challenge in the rare variant study is the lack of power due to their rarity. In this study, to overcome this issue, we analyzed large-scale whole-genome sequencing (WGS) datasets that were recently developed for the study of AD-related traits, including the Alzheimer's Disease Sequencing Project (ADSP) and the Accelerating Medicines Partnership-Alzheimer's Disease (AMP-AD). Another efficient tool we leveraged to increase the power was a gene-set burden analysis, where we focused on the collective rare variant effect within a set of genes of a known biological pathway, rather than the effect of single variants or single genes, and thus avoided the multiple testing burden required otherwise. This method has helped identify risk genes in various complex traits, such as in central nervous system pathways of schizophrenia.[28-39] In some studies, this method has led to the discovery of novel biological pathways and therapeutic targets through the identification of gene networks participating in the same functional processes.[40-45] Similar gene-set analyses focusing on biological pathways, as well as gene-ontology-based pathway/module analyses, have also been effectively demonstrated in AD studies.[11,46,47]

Therefore, in the current study, we included three large-scale WGS datasets with a total of 3,255 individuals of European ancestry, meta-analyzed under a gene-set rare-variant burden analysis framework. Phase 1 of this framework aimed to explore the effect of rare variants in the endocytic pathway as a whole and consisted of two stages followed by meta-analysis. Besides AD status, we additionally explored three AD-related phenotypes, neurofibrillary tangles (NFTs), age-at-onset (AAO), and age-of-death (AOD), along with the phase 1 analysis. NFT status was measured as Braak stages, first proposed by Braak and Braak in 1991, and served as a histopathological indicator of AD,[48-50] representing a finer progression of AD. Phase 2 of this framework was to identify single endocytic genes driving the rare-variant association we captured in phase 1. For each dataset, we examined each gene in the endocytic pathway using both AD and NFT status, followed by meta-analysis across all datasets. Finally, in phase 3, we sought to explore the functional consequences of the rare-variant effect identified in previous phases by examining both the bulk and single-cell expression of endocytic genes in relationship with AD status.

Methods

Study sample

To identify AD-associated rare-variant effects, we evaluated three publicly available large-scale WGS datasets collected for LOAD patients, downloaded as multi-sample VCF files. The Alzheimer's disease sequencing project (ADSP) Umbrella is a collection of sequencing data from the ADSP and other AD and Related Dementia studies. Under this Umbrella, the ADSP group sequenced a large number of well-characterized Alzheimer's Disease (AD) patients at three National Human Genome Research Institute Genome Centers (NHGRI) (Baylor College of Medicine Human Genome Sequencing Center, the Broad Institute, and the McDonnell Genome Institute at Washington University). The ascertainment methods and inclusion criteria are described in detail on the National Institute on Aging Genetics of

Alzheimer's Disease Data Storage Site (NIAGADS).[51,52] The sequencing results were mapped to the human reference genome (GRCh38) and processed using the VCPA 1.0 pipeline, which follows GATK best-practices pipeline.[53] Details of the variant calling pipeline can also be found on the NIAGADS. The ADSP discovery extension phase sequenced whole genomes of 1,466 cases and 1,534 controls from five cohorts provided by the Alzheimer's Disease Genetics Consortium (ADGC) and included samples with diverse ancestry backgrounds (Non-Hispanic White, Caribbean Hispanic, and African American). Another WGS project shared under the ADSP Umbrella is the Alzheimer's Disease Neuroimaging Initiative (ADNI), which is a longitudinal multi-center (63 sites across North America) study designed for early detection and tracking of AD. The ADNI WGS data contains 808 participants with 238 AD cases, 322 mild cognitive control (MCI) subjects, and 248 controls. A full list of the ascertainment methods and inclusion criteria can be found in detailed descriptions in the online ADNI protocol.[54] As of 2018, the ADNI was recalled under the same VCPA 1.0 pipeline as the ADSP discovery extension WGS data and mapped to the same human reference genome (GRCh38), which were then released together. This combined ADSP case-control dataset contained WGS data from a total of 3,896 individuals (accessed by us on Nov 20, 2018), which then underwent a sequence of quality control steps discussed later before including in our stage 1 analysis. Detailed demographic information of this dataset can be found in Table 1 and the distribution of age among AD cases and controls in S12 Fig. To note, we removed samples in the MCI category to ensure a strict bipartite definition of disease status from all our analyses.

WGS Datasets	Case-control		Family
Studies	ADSP	AMP-AD	ADSP
Total sample size	1,291	1,611	353

EUR Population (%)	41%	93%	50%
AD Patients	664	642	209
Controls	627	969	144
Males (%)	53.4%	35.4%	65.7%
APOE ε4 carriers (%)	43.5%	38.5%	44.8%
Reference genome	GRCh38	GRCh37	GRCh38

Table 2-1. Summary of clinical, demographic, and technical information of individuals from three large WGS datasets.

The numbers were counted only among the samples included in this current study. The percentages of EUR population were based on the total number of samples within each dataset and only the samples with EUR ancestry were included in this study, which served as the total input sample size in the first row. Abbreviations: AD: Alzheimer's disease; WGS: whole-genome sequencing; EUR: European; ADSP: the Alzheimer's disease sequencing project; AMP-AD: the Accelerating Medicines Partnership-Alzheimer's Disease.

Our stage 2 replication included 1,894 WGS samples from the Accelerating Medicines Partnership-Alzheimer's Disease (AMP-AD) Target Discovery and Preclinical Validation Project (accessed by us on Dec 13, 2018). The samples were separately sequenced at three centers: the Religious Orders Study and Memory and Aging Project (ROSMAP) (1,200 samples), the Mount Sinai Brain Bank (MSBB) study (354 samples), and the Mayo Clinic Brain Bank (Mayo) (350 samples). Previously reports have the detailed data collection scheme and sample inclusion and exclusion criteria.[55-58] This sequence data were mapped to the human reference genome (GRCh37) and were processed using the GATK best-practices workflow v3.4.0.[58] Another stage 2 replication was performed on the ADSP discovery extension phase family samples, which were released together with the ADSP case-control data. Therefore, this family WGS data were also mapped to the human reference genome (GRCh38) and processed using the VCPA 1.0 pipeline. The ADSP WGS family dataset contains 888 samples from 161 multiplex families. The inclusion criteria prioritized families loaded with LOAD with minimal *APOE* ε4 alleles. A detailed description of the study design and sample ascertainment methods can be found in previous reports.[59,60] For AMP-AD case-control study, this resulted in 642 AD patients and 969 controls after removing low-quality samples. For the ADSP family study, we obtained 545 AD patients and 285 cognitively normal older individuals (Table 1).

RNA-Seq data used for functional analysis were also obtained from the ROSMAP study of the AMP-AD Consortium[58]. The bulk RNA-Seq data were generated for 636 samples (254 AD cases, 368 controls, 12 other dementia, and two without annotation) from the dorsolateral prefrontal cortex (DLPFC) tissues by the Broad Institute's Genomics Platform and processed in an automatic and parallelized pipeline.[55] The ROSMAP group also selected 48 post-mortem samples (24 with severe AD pathology and 24 with low-to-no pathology) and conducted droplet-based single-nucleus RNA sequencing of the prefrontal cortex region.[61] Metadata of the RNA-Seq data were then used to map samples to cases and controls following the same rule as in stage 2 replication, as well as to merge with genotyping data.

Data processing and quality control of WGS data

Individual-level quality control of WGS data

We conducted stringent quality control (QC) to ensure that we only include high-quality samples. As the X chromosome was not available in the ADSP datasets, we did not include X chromosome for all analyses. Before checking the sequencing quality of each individual, we first removed variants failing Variant Quality Score Recalibration (VQSR) in the GATK pipeline and set all variants with genotyping quality (GQ) below 21 to missing. We included only bi-allelic variants for all future analyses. Within the remaining variants, for each individual, we evaluated the genotype missing rate, calculated theoretical relatedness to check for unexpected relationships by study design, and performed principal component analysis (PCA) to identify ancestral composition and population outliers. For the individual-level missing rate QC, we set the cutoff at 5% and removed all individuals beyond this threshold. For the relatedness check, we used PLink 1.9[62] and conducted identity by descent (IBD) analysis, which allowed us to compute a relatedness degree for each sample. For case-control studies, we retained only one in each cluster of samples estimated to be first- or second-degree relatives or duplicates within the corresponding cluster. For the ADSP family study, we compared the empirical kinship relationship record to our computed theoretical relatedness. For PCA, we used 1000 Genomes (1KG) phase 3 as a reference panel.[63] We used EIGENSTRAT[64] for PCA and included only independent common SNVs that were shared between 1KG and our dataset. To note, as PCA assumes unrelated individuals, when performing PCA for the ADSP family cohort, we restricted to only one sample in each family to avoid confounding ancestral relationship by kinship relationship. After having determined the ancestry of the included sample based on PCA, we then assigned that ancestry to the entire family of the included sample. PCA plots (PC1 vs. PC2) of all three datasets could be found in S4 and S6 Figs. As the X-chromosome was available for the AMP-AD study, we also performed sex-check for the AMD-AD and obsevered no sex-mismatched samples. In summary, after stringent sample-level quality control and the careful examination of ancestral backgrounds, we identified 1,291 (664 AD cases and 627 cognitively normal older controls), 1,611 (642 AD cases and 969 controls), and 353 (144 AD cases and 209 controls) high-quality European samples in the ADSP case-control, the AMP-AD case-control, and the ADSP family datasets, respectively, which then served as the primary objects of our study in both stages 1 and 2.

Variant-level QC of WGS data

We conducted stringent variant-level quality control to ensure keeping only high-quality SNVs. We included only variants that served as inputs for the individual-level QC while including only samples passing the individual-level QC. For each variant, we assessed the genotype missing rate, computed minor allele frequency (MAF) using all European samples, and calculated the Hardy-Weinberg Equilibrium (HWE) p-values using only unaffected European samples. For the variant-level missing rate QC, we set the cutoff at 2% and removed all variants beyond this threshold. For HWE, we set the cutoff at 0.001 for rare variants and removed all rare variants falling the HWE check where rare variants are defined in the following section. The number of variants passing the HWE filter could be found in S14 Table.

Identification and annotation of rare variants

To identify rare variants, we used both external and internal sources of allele frequency to avoid potential inflation of the allele frequency introduced by the study design. For the external sources, we looked at the Europeans (EUR) in 1KG[63] and Non-Finnish Europeans (NFE) in the gnomAD v2 database[65], which matched the ancestral backgrounds of our datasets. We used two different MAF thresholds (0.1% and 1%) to define rare variants, as there is no one consensus definition of rarity and we will correct for testing multiple MAF thresholds in future analysis. In practice, when a variant was present in either of 1KG EUR or gnomAD v2 NFE samples and below the aforementioned threshold, we would keep it for further analysis. For the internal sources, we retained only samples with European ancestry based on the previous PCA, as different ancestral groups would have different allele frequency distributions. Then when a variant was absent in both external databases, we would look at the MAF estimated from the European samples within our dataset and selected rare variants based on 0.1% and 1% MAF thresholds separately. We then annotated rare variants using Ensembl Variant Effect Predictor (VEP)[66]. We defined a variant to be 'deleterious' if it is within one of the following categories: stop-gain, stop-loss, frameshift, splice-donor, splice-acceptor, and missense variants. Particularly, for missense variants, we additionally consulted PolyPhen-2[67] and retained only confident missense variants predicted to be 'damaging.' This definition of deleteriousness focused on coding regions, primarily due to the fact that the effect of non-coding variants was challenging to predict.[68,69] A distribution of variant types and singletons among the selected set of rare deleterious variants could be found in S9 Fig and S11 Table. In an additional validation of the deleteriousness, we further introduced the CADD score[70] as a third deleterious criterion in phase 1 analysis combined with VEP and PolyPhen-2. The distribution of CADD scores among the set of rare deleterious variants could be found in S8 Fig. As suggested by the CADD documentation, variants with scaled CADD > 15 were retained as pathogenic variants and the set of rare deleterious/pathogenic variants passing all three annotation tools were used in this validation test.

Identification of genes in endocytic pathways

We identified genes involved in endocytic pathways using AmiGO 2[71,72] gene ontology database to select all genes participating in this pathway. We identified three specific GO terms related to the endocytic system in the Homo Sapiens category, which corresponded to three specific compartments in the endocytic system (endosome, lysosome, and trans-Golgi network). The endosome compartment is a membrane-bound vacuole in eukaryotic, participating in the endocytic trafficking from the trans-Golgi network to the plasma membrane and vice versa.[73] The trans-Golgi network serves as an interconnected tubular network and the final cisternal structure involved in packaging and transporting of cargos to the lysosome, endosome, and cell surface.[74] The lysosome, a small membrane-bound lytic vacuole, is one of the end-point in the endocytic transporting pathway, which contains hydrolytic enzymes to break down various biomolecules.[75] The combination of these three compartments formed the essential backbone of the endocytic system, which we named as “endo-system” and used this term throughout the paper. After removing duplicates, we obtained 1,435 genes in total in the endo-system, while the three compartmental gene-sets contained 899 (endosome), 678 (lysosome), and 236 (trans-Golgi network) genes, respectively. We confirmed their biological functions with a functional enrichment analysis using the Database for Annotation, Visualization, and Integrated Discovery (DAVID)[76], where the top enriched GO terms were indeed lysosome, endosome, and trans-Golgi network. (S13 Fig) A comparison of the endo-system gene-set to the findings in the recent AD GWASes[11,12] has been provided by checking the number of endocytic genes implicated in Jansen et al. and Kunkle et al. (S7 Fig). To note, some genes were related to multiple compartmental gene-sets and thus only one of the duplicated genes was included in the endo-system gene-set (S11 Fig).

Analysis of association between the burden of rare deleterious SNVs and AD status

To identify whether rare variants in the endocytic pathway are associated with AD, we compared the burden of rare deleterious SNVs between AD patients and controls. The burden was defined as the fraction of the alternative minor alleles that each individual carried for all rare deleterious SNVs, using the --score function in PLINK[62]. We additionally performed this procedure on the three compartmental gene-sets and obtained a burden score for each individual within each gene-set. To correct for potential confounding factors, for each gene-set, we first regressed the burden against the total number of rare SNVs and the top ten principal components (PCs). Due to randomness, the distribution of the number of rare SNVs might be naturally variable from sample to sample, in which case the distribution of rare deleterious SNVs would also be greatly affected. Similarly, the PCs helped to correct for potential population stratification within European ancestries. Both aspects could influence the burden score in ways unrelated to AD and thus need to be controlled. Once we had removed the confounding covariates, we performed three logistic regression models as proposed by Zhang et al.[77] using the residuals and AD status for all case-control studies. The three models differed in the covariates they corrected for. The minimal adjustment Model 0 (M0) controlled for the ten PCs and sequencing centers. This model has been previously reported to improve power for detecting variants whose effects are confounded with age and sex.[60] This phenomenon could be introduced by study design where the mean age between cases and controls are substantially disproportionate, as in the case of ADSP studies. Model 1 (M1) was built upon M0 by additionally including age and sex. Model 2 (M2) was further built upon M1 and included the count of *APOE* ε2 and ε4 alleles. For the ADSP family dataset, we generated kinship matrices and used a generalized linear mixed model (GLMM) to take kin relationships into consideration when calculating association p-values. In particular, we used the glmmkin function in the R package, GMMAT.[78] We computed odds ratio (OR) and p-values of association between the burden of rare deleterious SNVs and AD status in each model for European samples in each dataset (ADSP case-control study, AMP-AD case-control study, and ADSP family study). Our stage 1 analysis involved only the ADSP case-control dataset as the discovery set, while the AMP-AD case-control and the ADSP family study served as replication sets in our stage 2 analysis. We chose this analysis scheme because the ADSP case-control study encompassed the largest sample size, including non-European samples, even though we identified fewer samples with European ancestry compared to the AMP-AD case-control study. To note, the AMP-AD case-control study provided only the age-of-death for each individual, while the ADSP case-control and family studies provided only the age-at-onset. As a result, we used different definitions of age in analyzing different datasets. To validate our gene-set AD association analysis, we tested two additional methods provided by MAGMA[79] using the same set of rare deleterious variants. The first was the SNP-wise method applicable to both common and rare variants and the second was the burden method that MAGMA suggested to use for rare-variant-only analysis and was similar to the aforementioned gene-set AD association analysis using PLINK. We applied both methods to the set of rare deleterious variants previously defined and computed two types of p-values: a competitive p-value that tests whether the association within the gene-set is greater than in other genes and a self-contained p-value that tests whether there is an association within the gene-set of interest at all. The latter concept is the same as what our main analysis method aimed for. Due to our study design with multiple gene-sets and MAF thresholds, a Bonferroni correction was applied in accordance with the number of tests we performed in each analysis to define the study-wide significance threshold in each stage and each dataset. Although our analysis started with the whole endocytic pathway and then moved onto individual compartments, we, nonetheless, utilized a stringent multiple-testing correction threshold. Specifically, as we tested for four gene-sets (endo-system gene-set and three sub-compartmental gene-sets) and two MAF thresholds (1% and 0.1%), we set our significant threshold at α=0.05/8=0.00625 for both stage 1 discovery phase and stage 2 replication phase analyses. Accordingly, we set our nominal significance threshold at α=0.05.

To combine results from two stages (three studies) for each of the four gene-sets we tested previously, we performed meta-analyses on p-values using estimates from our best model, namely the model producing the smallest p-values among the three models tested. We used two meta-analysis methods to combine the results. The first was a fixed-effects inverse variance weighted method in METAL[80], which took ORs, standard deviations (SDs), and p-values for separate tests and combined them into one 'Gene-set level' p-value with an estimate of the unified effect. The second was Fisher's method which only required p-values and has been shown to be more robust to some situations where a small portion of p-values are very small.[81,82] In particular, we used the sumlog function from the R package, ‘metap,’[83] which took into account the direction of effects in each study and the corresponding p-values. It then computed a ‘Gene-set level’ p-value similar to METAL indicating the significance of rare variants' effect shared across studies but without an estimated effect size.

Analysis of association between the burden of rare deleterious SNVs and AD-related phenotypes

To test for association between the burden of rare deleterious SNVs and NFTs, we leveraged the Braak stages and followed a similar workflow as in testing AD status. As the sample size of patients with Braak staging information was limited in the ADSP family study, we tested for replication only in AMP-AD case-control study after analyzing the ADSP case-control study in stage 1. We obtained 626 and 1,399 individuals with Braak staging information in ADSP and AMP-AD case-control datasets, respectively. To note, even though the ADSP case-control study had fewer samples with Braak staging information, we, nonetheless, followed the same analysis scheme as in the previous AD analysis. In practice, after removing confounding effects from the burden score, we applied three ordinal logistic regression (OLR) models (M0, M1, M2) to account for multiple ordered categories present in the Braak staging (stage 0 to VI). The regular logistic regression only allows binary dependent variables, which is not feasible for Braak stages. In particular, we used the polr function from the R package, MASS[84], which fits a logistic regression model to an ordered factor response. Similar to the previous burden analysis, our M0 accounted for sequencing centers and the top 10 PCs; our M1 additionally controlled for sex and age; finally, our M2 further included the count of *APOE* ε2 and ε4 alleles. For analyses in all datasets, our significance threshold after the multiple-testing correction was still at α=0.00625 because we tested for two MAF thresholds and four gene-sets. Finally, the nominal significance threshold was also at α=0.05. To increase statistical strength and precision in estimating effects[85], we again performed meta-analyses and combined these two independent tests similar to what we did for AD association analyses

We additionally tested the age-specific risk of rare deleterious SNVs in the endocytic pathway. As aforementioned, the AAO and AOD information was provided by the ADSP studies and the AMP-AD study, respectively, which allowed us to test for two different age-specific risks within each gene-set. Different from AD risk, age-specific risk leveraged the information of age and estimated the association between the age-to-event (survival time) of patients and the rare-variant burden score. Therefore, we adopted a genetic epidemiological framework proposed by Desiken et al.[86], in which a Cox Proportional Hazard Regression (CPHR) was performed to account for age-to-event information. Specifically, we first used the Surv function from the R package, “survival”[87], and computed a survival time for each sample in each dataset. Then, we conducted CPHR using the coxph function from the R package, ‘survminer’[88], to estimate the hazard ratio, or the ratio of risk-to-event (onset or death), depending on the input age we used. We performed three CPHR models (M0, M1, and M2) similar to the previous burden analysis on AD status and Braak staging, except that age was not a covariate in either of the three models. Therefore, since we tested for two different MAF thresholds and four gene-sets (though in a stepwise fashion), we set a stringent significant threshold at α=0.05/8=0.00625 and our nominally significant threshold at α=0.05 for analyses in all three datasets. Finally, we combined the results of AAO in the same way as we did for AD and NFT association tests. The resulting p-value then indicated the shared rare-variant effect on AAO-specific risk across the ADSP case-control and family studies.

Single-gene analysis

To identify specific genes within the endocytic pathway associated with AD, we extracted rare deleterious SNVs as defined previously for each gene in the endo-system gene-set that were present in European samples for the ADSP case-control, the AMP-AD case-control, and the ADSP family study. Association test was performed for AD status by first building a null model using the SKAT_Null_Model function in the R package, SKAT,[89] followed by running the SKATBinary function using the SKAT-O feature to obtain association p-values for binary traits. We used a full model that included age, sex, sequencing center, the number of *APOE* ε2 and ε4 alleles, and top 10 PCs. To note, we also applied SKAT_Null_Model to the ADSP family dataset without incorporating kinship structure. This procedure could only be valid in the case where the family structure was relatively simple and did not contribute to a large effect in our analysis. By re-running the previous AD burden analysis with and without kinship information, we indeed observed only small deviations between these two tests. Specifically, for the full model of the endo-system gene-set, we observed an OR of 1.34 with kinship structure provided (p=0.035) while we observed a similar OR of 1.36 assuming an independent setup (p=0.02), which indicated that the family structure within the ADSP family study did not influence our analyses to a large extent.

To test for association with Braak stages, we first extracted only European samples with Braak staging information available for each dataset, before extracting rare deleterious SNVs for each gene within the endo-system gene-set. We leveraged the fact that it is a semi-quantitative trait and performed the association test with the SKAT function for continuous traits with the 'optimal' option after building null models as described for testing AD status. In the attempt to remove confounding factors and unbalanced sample distribution for Braak staging association test, we additionally included AD status in null models. Finally, we meta-analyzed variants across datasets and computed 'Gene-level' p-values for AD status as well as Braak staging. We combined genotyping matrices across three datasets for each gene using the R package, MetaSKAT.[90] Specifically, we first transformed our genotyping matrices into an SSD format for a single population and then analyzed all three populations at once using the function MetaSKAT_MSSD_ALL. This procedure increased the power to analyze the effects of rare variants that are shared across different studies. To correct for testing multiple genes within the endo-system gene-set, we obtained the number of genes we tested in each separate dataset and computed their corresponding Bonferroni corrected significance thresholds. Specifically, for the AD single-gene analysis, we tested 1,195, 1,228, and 683 genes in ADSP case-control, AMP-AD case-control, and ADSP family datasets, respectively, which corresponded to Bonferroni corrected significance thresholds of α=4.18*10^{-5}; 4.07*10^{-5}; 7.32*10^{-5}, respectively. In meta-analyses, we identified 642 genes in common and computed a Bonferroni corrected significance threshold of α=7.79*10^{-5}. For the Braak staging single-gene analysis, we retained only rare deleterious SNVs present in samples with Braak staging information available and tested for 1,035 and 1,176 genes for the ADSP and AMP-AD case-control studies, respectively. The corresponding Bonferroni corrected significance thresholds were then computed as α=4.83*10^{-5} for the ADSP case-control dataset and 4.25*10^{-5} for the AMP-AD case-control dataset. When performing meta-analyses, we examined 967 genes in common between these two datasets, which led to a Bonferroni corrected significance threshold of α=5.17*10^{-5}.

Functional analysis on AD

One approach to understanding how the effect of rare variants would influence the risk of AD status is to investigate how they regulate gene expression. A gene with a variation that is associated with its gene expression is called an eGene. Here, we obtained the bulk RNA-Seq data of DLPFC tissues of 636 individuals from the ROSMAP[55] study and performed an association test between the expression of a gene and rare variants in *cis* with the corresponding gene. In particular, for each gene within the endo-system gene-set, we included all variants within gene boundary and additionally all rare variants within 20kb up- and down-stream of the transcription start sites (TSS), which might potentially regulate the expression of a gene through *cis*-regulation, such as the effect of enhancer region. To overcome the problem of low power to detect the effect of single rare variants, we aggregated the effects of all rare variants within as well as near the TSS of each gene. We analyzed this aggregated effect on gene expression using the SKAT function to compute 'Gene level' p-values, while taking into account confounding covariates, including age, sex, sequencing locations, *APOE* ε2 and ε4 alleles, and top 10 PCs. To correct for testing multiple genes, we calculated false discovery rate for all tested genes and used FDR of 0.05 as the q-value threshold, following the suggestions of previous studies.[91,92] Follow-up validation was performed using genes previous identified from the burden and functional analyses, by directly comparing their expression levels between AD cases and controls using student t-test and computing the Pearson correlation between their expression levels and Braak stages. The multiple-testing issue was then addressed using the Bonferroni correction method.

The resolution of bulk RNA-Seq data may limit our capability of observing cell-type specific effects on AD.[55,61,93,94] To elucidate the underlying complexity of variation across cell types, we further obtained single-cell RNA-Seq (scRNA-Seq) of 48 samples (24 AD patients and 24 cognitively normal controls) from the ROSMAP study and investigated the pattern of expression for each of the six major cell types defined on a priori cell-type-specific gene-sets: excitatory neuron (Ex), inhibitory neuron (In), astrocyte (Ast), oligodendrocyte (Oli), oligodendrocyte-precursor-cell (Opc), and microglia (Mic)[61]. The six major cell types were further divided into sub-clustered cells based on the heterogeneity of gene expression within each cell type: 13 Exs, 12 Ins, 4 Asts, 5 Olis, 3 Opcs, and 4 Mics[61]. The whole dataset in

10X format was first processed using the R package, Seurat.[95] We followed the preprocessing steps as proposed by the Seurat developer by first filtering out cells with reads quantified for less than 200 or more than 2,500 genes, followed by filtering out cells with the percentage of mitochondrial gene counts over 5 percent. We then employed a global-scaling normalization method provided by the LogNormalize function, which normalized the feature expression measurements for each cell by the total expression, followed by a log-transformation. The major and sub-cell types were identified a priori for this scRNA-Seq data. Therefore, we extracted all significant genes identified in the previous single-gene and functional analyses for each specific cell type and conducted differential gene expression analysis using the student t-test method between cases and controls for each major cell type, as well as for each subcellular population within each major cell type.

Result

The burden of rare deleterious SNVs in endo-system gene-set for ADSP case-control study

To investigate whether rare deleterious SNVs in the endocytic pathway were associated with AD, we leveraged a gene-set method of burden analysis that collapsed individual effects of multiple variants into one 'gene-set level' effect, hence increasing the power of detecting rare variants' effect. We defined rare SNVs using both an external source of allele frequency and allele frequency observed in 1,291 European samples (664 AD cases and 627 controls) from the ADSP case-control study (see Methods). We focused on deleterious SNVs as defined in Methods, in which most were protein-altering variants. We identified rare deleterious SNVs in 1,133 of the 1,435 genes in our gene-set (see Methods). For each individual, we computed the burden of these rare deleterious SNVs. We then compared the genetic burden between AD cases and cognitively normal controls, while taking into account confounding covariates that can potentially influence the amount of burden. Such covariates include ancestral principal components, age, sex, the sequencing location, the number of *APOE* e2 and e4 alleles, and the total number of rare SNVs of each individual. The last procedure is necessary to account for individual differences in the total amount of variation; an individual is likely to carry more rare deleterious SNVs if she/he carries more rare SNVs overall. To note, we found that the total number of rare SNVs on the genome-wide scale has no statistically significant difference between cases and controls (p=0.67, student t-test). As described in Methods, we applied three logistic regression models to find associations between AD status and the burden scores while the three models were built on top of each other and tested for two MAF thresholds (1% and 0.1%). Looking at our best model in terms of the strongest association, we observed that the risk of AD, as indicated by the odds ratio (OR), increased by 1.24 for every one unit increase in residual burden score (p=0.00018 using GLM), which was a significant association after stringent multiple testing correction (α=0.00625) for all gene-sets (including sub-gene-sets we analyzed in next steps) (Fig 1).

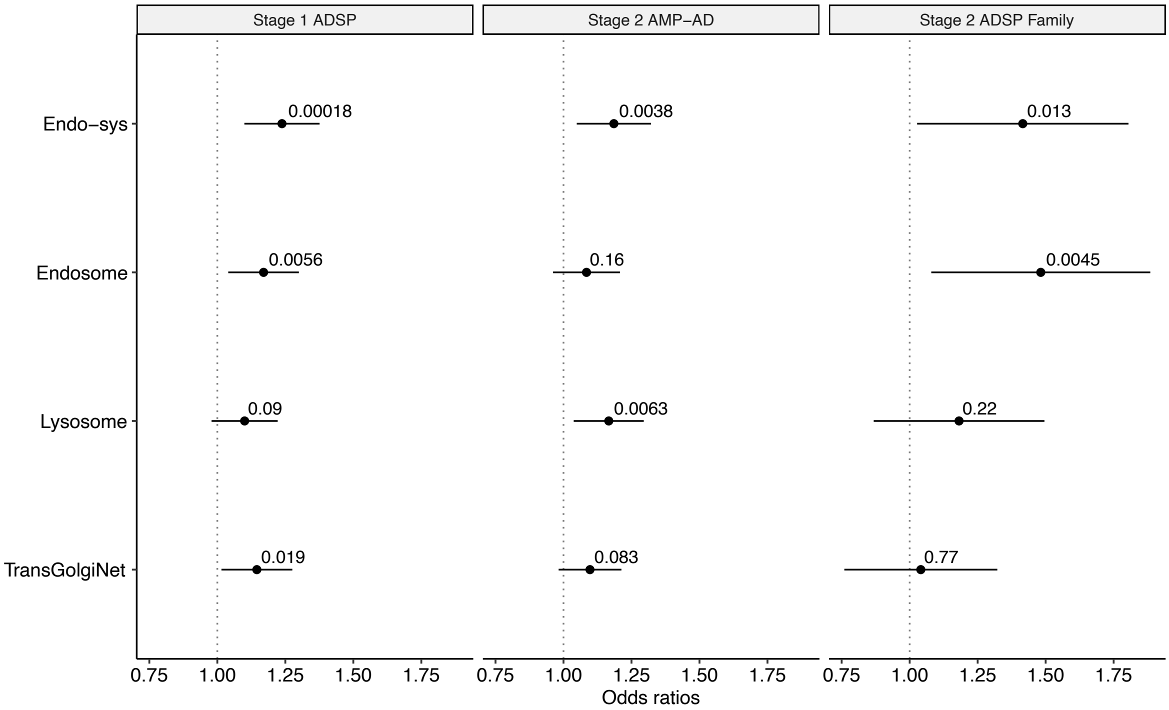


Figure 2-1. Rare deleterious variants are enriched in AD patients across the endocytic and corresponding compartmental gene-sets in stages 1 and 2.

We compared the burden of rare deleterious variants between AD patients and controls across the endo-system (endo-sys) gene-set and three compartmental sub-gene-sets (endosome, lysosome, and trans-Golgi network) in stage 1 ADSP case-control dataset (leftmost), which were then tested for replication in stage 2 AMP-AD case-control (middle) and ADSP family (rightmost) datasets. Enrichment (ORs) and p-value were computed using a linear regression model controlling for covariates, including the total count of rare variants (see Methods). P-values of enrichment in each gene-set are indicated above horizontal bars which represent 95% confidence intervals.

Additionally, we identified three major cellular compartments participating in the endocytic pathway and their corresponding genes, which constituted subsets of the endo-system gene-set. The first two compartmental gene-sets were endosome (n=811) and lysosome (n=620) gene-sets, which served as the major sorting station in the endocytic pathway and the final destination of proteolytic destinations[96], respectively. The third important compartment was the trans-Golgi network gene-set (n=208) which represented a pathway sorting station for retrograde trafficking. In summary, we identified 689 endosomal genes, 544 lysosomal genes, and 181 trans-Golgi network genes, respectively. We found the burden scores of rare deleterious SNVs were higher in cases than in controls for all three sub-gene-set. In our best model, the OR, representing the risk of AD, increased by 1.18 per unit for the endosome gene-set (p=0.0056 using GLM), 1.08 per unit for the lysosome gene-set (p=0.09 using GLM; Fig 1), and 1.14 per unit for the trans-Golgi network gene-set (p=0.019 using GLM). After the multiple-testing correction, we observed the endosome gene-set showed a gene-set-wide significant association with AD while the trans-Golgi network displayed a nominally significant association signal. In addition to exploring sub-gene-sets, we also checked the specificity of the association in the endocytic pathway by obtaining gene-sets unrelated to AD. Specifically, we explored two non-disease complex traits, BMI and height, and obtained related genes (212 and 78, respectively) from GeneRIF, a publically available database for functional annotations.[97] Indeed, we did not observe an enriched rare-variant burden in AD cases compared to controls in these gene-sets and the directions of effects were different across datasets, suggesting the observed rare-variant effect was specific to the endocytic pathway. (S13 Table)

Stage 2 replication of the burden analysis in two independent WGS datasets

The gene-set burden analysis in the ADSP case-control study demonstrated statistically significant enrichment of rare deleterious SNVs in cases in the endocytic pathway, indicating an increase of risk conferring AD. We further examined the endo-system gene-set in 1,611 European samples (642 AD cases and 969 controls) from the AMP-AD study. We obtained 1,198 endo-system-related genes and observed an elevated risk of AD in terms of OR of 1.19 (p=0.0038 using GLM; Fig 1), replicating the observation of a significantly higher burden of rare deleterious SNVs in the stage 1 analysis, using a multiple testing threshold of α=0.00625.

We performed additional gene-set burden analysis on the sub-gene-sets of the functional compartments in the AMP-AD study. We identified 735 endosomal genes, 576 lysosomal genes, and 187 trans-Golgi network genes, respectively. We again observed an increase in AD risk among cases for all three sub-gene-sets. A nearly significant signal was observed in the lysosome gene-set with an OR of 1.17 (p=0.0063 using GLM; Fig 1). For the other two gene-sets, we observed an OR of 1.08 (p=0.16 using GLM; endosome gene-set) and 1.10 (p=0.083 using GLM; trans-Golgi network). None of these gene-sets showed gene-set-wide significant association after multiple testing correction at α=0.00625, although the lysosome gene-set nearly reached the gene-set-wide significance threshold.

As described above, the AMP-AD study consisted of three sub-cohorts and the largest one, ROSMAP, contained around 71.5% of the total sample size. To avoid potential batch effect diluting the association signal, we re-performed the analysis on only the ROSMAP data. In fact, we observed slightly more significant results in nearly all gene-sets, where the endo-system and the lysosome gene-sets both reached gene-set-wide significance threshold. (S2 Table) Overall, the associations were similar between the AMP-AD and ROSMAP data, indicating a relatively low level of batch effect among the three sub-cohorts.

Given the observed risk in stage 1 ADSP case-control study and the stage 2 AMP-AD replication, we further examined the genetic burden in the ADSP Family study. We filtered and annotated rare deleterious SNVs based on the same workflow using 353 European samples (144 AD cases and 209 controls) of the ADSP family study. Due to the smaller sample size compared to the previous two case-control studies, we obtained 683 endo-system-related genes. To examine the AD risk, we performed GLMM using the burden of each individual. Due to family structure, we utilized the generalized linear mixed model to account for the relatedness between samples. We observed an OR of 1.42 (p=0.013 using GLMM), conferring an elevated AD risk among cases compared to controls. (Fig 1) This observation was not gene-set-wide significant using the Bonferroni correction threshold at α=0.00625. However, it displayed a nominally significant association with the same direction of effect as in the ADSP and AMP-AD case-control studies.

Nonetheless, we looked into the sub-gene-sets of the three functional compartments in the ADSP family dataset. We identified 402 endosomal genes, 342 lysosomal genes, and 106 trans-Golgi network genes, respectively. We observed a significant elevation of AD risk among cases for endosome gene-set with an OR of 1.48 (p=0.0045 using GLM). Similar increases were also observed in the lysosome and trans-Golgi network gene-sets, with OR of 1.18 (p=0.22 using GLMM) and 1.04 (p=0.77 using GLMM), respectively (Fig 1). Only the endosome gene-set remained gene-set-wide significant after multiple-testing correction, which was in concordance with our observation in the stage 1 ADSP case-control study.

A meta-analysis of stage 1 and 2 burden analysis

The stage 1 burden analysis using the ADSP case-control study demonstrated a significant increase in AD risk in the endo-system gene-set, which was replicated in one independent dataset, the AMP-AD case-control dataset, and displayed a nominal significance in the ADSP family study. We meta-analyzed the results using two different methods (see Methods) and computed a 'Gene-set level' p-value of $2.17*10^{-7}$ (by METAL; Fisher's method produced similar results; Table 2) for the endo-system gene-set, which was improved compared to stages 1 and 2. The same was also observed for sub-gene-sets where we computed a meta-analysis p-value of 9.78*10^{-5} for the endosome gene-set, 9.83*10^{-4} for the lysosome gene-set, and 1.19*10^{-2} for the trans-Golgi network gene-set. Except for the trans-Golgi network gene-set that has the smallest number of genes, all other gene-sets remained gene-set-wide significant after multiple-testing correction (α=0.00625), which strongly demonstrated a shared effect of rare deleterious variants within the endocytic pathway across multiple independent studies. To note, although we meta-analyzed the results from the best models, as proposed by Zhang et al. to improve the power of detection, the same pattern of rare-variant association could be observed using the same models for each gene-set across the three datasets. (S1 Table) For all models in the endo-system, endosome, and lysosome gene-sets except M2 of lysosome, we observed gene-set-wide significant p-values, regardless of the meta-analysis methods used, demonstrating a high consistency with the observations made using the best models.

Phenotype	AD		NFT		AAO	
	P	P*	P	P*	P	P*
Endo-system	2.17E-07	2.66E-07	1.16E-02	9.89E-03	2.47E-06	4.93E-07
Endosome	9.68E-05	6.05E-05	1.30E-01	9.34E-02	3.33E-05	2.04E-05
Lysosome	9.83E-04	1.15E-03	6.56E-03	6.11E-03	1.10E-02	3.11E-04
TransGolgiNet	1.20E-02	7.46E-03	5.71E-01	3.53E-01	2.10E-02	4.96E-03

Table 2. Meta-analysis of stages 1 and 2 gene-set burden analyses using AD, NFT, and AAO Abbreviations: AD: Alzheimer's disease; NFT: neurofibrillary tangle; AAO: age-at-onset; NFTs were analyzed using Braak stages. Gene-set-wide significant results were highlighted in bold. Displayed results of gene-set burden analyses were each meta-analyzed using

METAL (P) and Fisher's method (P*) (see Methods). Directions of effects were consistent across all tests.

A similar pattern of meta-analysis results was also observed in the additional validation tests from two aspects. Firstly, we wanted to check our results using different annotation tools. Given the set of deleterious variants used in previous phase 1 analyses, we additionally filtered by CADD scores (see Methods) and re-ran the gene-set AD association analyses with the resulting set of pathogenic/deleterious variants. In the meta-analysis, we observed that the endocytic, endosome, and lysosome gene-sets reached gene-set-wide significance threshold (see S5 Table), consistent with the rare-variant effect we observed in the endocytic pathway using the original set of rare deleterious variants.

The second aimed to validate our gene-set burden analysis using MAGMA with two different aggregation methods (see Methods). In the meta-analysis, both the SNP-wise and burden methods provided gene-set-wide significant self-contained p-values for nearly all gene-sets (S3 and S4 Tables; for endo-system, SNP-wise: $9.28*10^{-7}$; burden: $5.16*10^{-8}$), similar to the results shown above in Table 2. Compared to the MAGMA burden method, the SNP-wise method was not designed for rare-variant-only analysis and indeed showed weaker association signals. Especially for the competitive p-values, we observed gene-set-wide significant results for nearly all gene-sets using the MAGMA burden method, but not the SNP-wise method (for endo-system, SNP-wise: $2.41*10^{-2}$; burden: $1.90*10^{-3}$). We also attempted to compute a weighted burden score using pLI scores by PLINK and observed gene-set-wide significant associations in the endo-system gene-set in the meta-analysis. (S10 Fig, S12 Table) Compared to our main method above, the MAGMA methods and the weighted method displayed some fluctuations in individual datasets and models but consistent results in meta-analysis, indicating a robust rare-variant effect in the endocytic pathway under different statistical methods. Besides, as *APOE* was a major risk determinant in AD, in this validation, we also checked whether our observed rare-variant enrichment was mainly contributed from this gene, rather than the whole endo-system gene-set, by re-run the analysis with *APOE* excluded. Indeed, we observed nearly the same p-values in the meta-analysis, indicating a rare-variant enrichment in AD cases even without *APOE*.

The burden of rare deleterious SNVs on NFTs

NFT, measured in Braak staging, was one of the most important histopathological indicators of AD[48-50]. It is designed as an ordinal scale from 0 to VI of NFT pathology where AD patients with high Braak stages (V or VI) are diagnosed with high confidence.[98] Therefore, Braak stages may serve as a finer spectrum or proxy of AD severity and provide higher power in assessing the effect of rare variants in AD progression. Based on our previous AD analysis, we hypothesized that the burden of rare deleterious variants in the endocytic pathway would be higher in patients with later Braak stages. To test our hypothesis, we applied an ordinal logistic regression (OLR) method to Braak stages (see Methods). This method has been previously shown to be effective in studies of Braak staging as well as of other ordered phenotypes, such as oral cancers.[99,100] We obtained 626 individuals (475 AD cases and 151 cognitively normal controls) from the stage 1 ADSP case-control dataset and 1,399 individuals (533 AD patients and 866 controls) from AMP-AD case-control dataset with Braak staging information, which were used to fitted OLR models. In stage 1, We observed an OR of 1.16 (p=0.039 using OLR; S1 Fig) in the endocytic pathway, implicating a nominally significant association of rare-variant enrichment to later Braak stages. However, this result did not replicate in stage 2 with sufficient significance (OR=1.08, p=0.13 using OLR; S1 Fig). Comparing the stages 1 and 2 samples, we observed a distinct distribution of Braak stages. In particular, the stage 2 samples were concentrated in Braak stage III (23.1%), IV (28.1%), and V (23.3%), whereas most stage 1 samples were clustered in stage V (26.4%) and VI (34.8). (S3 Fig) We did not test for replication in the ADSP Family study due to limited samples with Braak staging information (n=38 individuals where only one sample had AD).

Our analyses of two independent datasets suggested a trend of increased risk of bearing later Braak stages with elevated rare-variant burden in the endocytic pathway. To improve power, we meta-analyzed the results from the ADSP and AMP-AD case-control studies, producing a 'Gene-set level' p-value between 0.0099 and 0.012, which did not pass our multiple-testing correction threshold of α=0.00625. (Table 2) Further Braak staging burden analysis using compartmental sub-gene-sets, however, revealed a gene-set-wide significant signal in the meta-analysis for lysosome gene-set (p=0.0066, Fisher's method). A full list of results for NFT burden analysis can be found in S6 Table.

Hazard analysis on population risk of AD age of onset and death

Previous gene-set burden analyses have demonstrated a significant correlation between the burden of rare deleterious variants within the endocytic gene-set and AD risk. One important aspect of AD development is its age-specific phenotypes, such as AAO. Previous studies on AD have shown a large genetic component in the heritability of AAO[101,102], with multiple risk loci associated with it. [103-107] It is thus of interest to also examine the genetic risk identified within the endocytic gene-set in this context. One approach is to evaluate whether AD patients with earlier AAO are associated with greater rare-variant burden within the endocytic gene-set. Previous studies have proposed a genetic epidemiological framework, where age-specific phenotypes were analyzed using a Cox Proportional Hazard Regression (CPHR) that considered a time-to-event probability, as opposed to the simple event probability estimated in logistic regression.[86,108] Therefore, we leveraged our previously computed burden score for each individual in the ADSP case-control study and constructed a cox proportional hazard (CPHR) model to estimate the instantaneous risk of developing AD, in consideration of genotype and AAO. A positive estimate of hazard in this model would indicate a higher risk of developing AD in early ages. We built three models as in the burden analysis and observed in our best model that an AAO-specific genetic risk increased by 1.14 per unit increase in the residual burden score (p=0.00083 using CPHR; Fig 2), which reached gene-set-wide significance after multiple testing correction (α=0.00625). We further examined the AAO-specific genetic risk within the functional sub-gene-sets. In our best model, we observed a gene-set-wide significant hazard ratio of 1.14 (p=0.00097 using CPHR) for lysosome gene-set and a nominal significant hazard ratio of 1.10 (p=0.011 using CPHR) for trans-Golgi network gene-set.

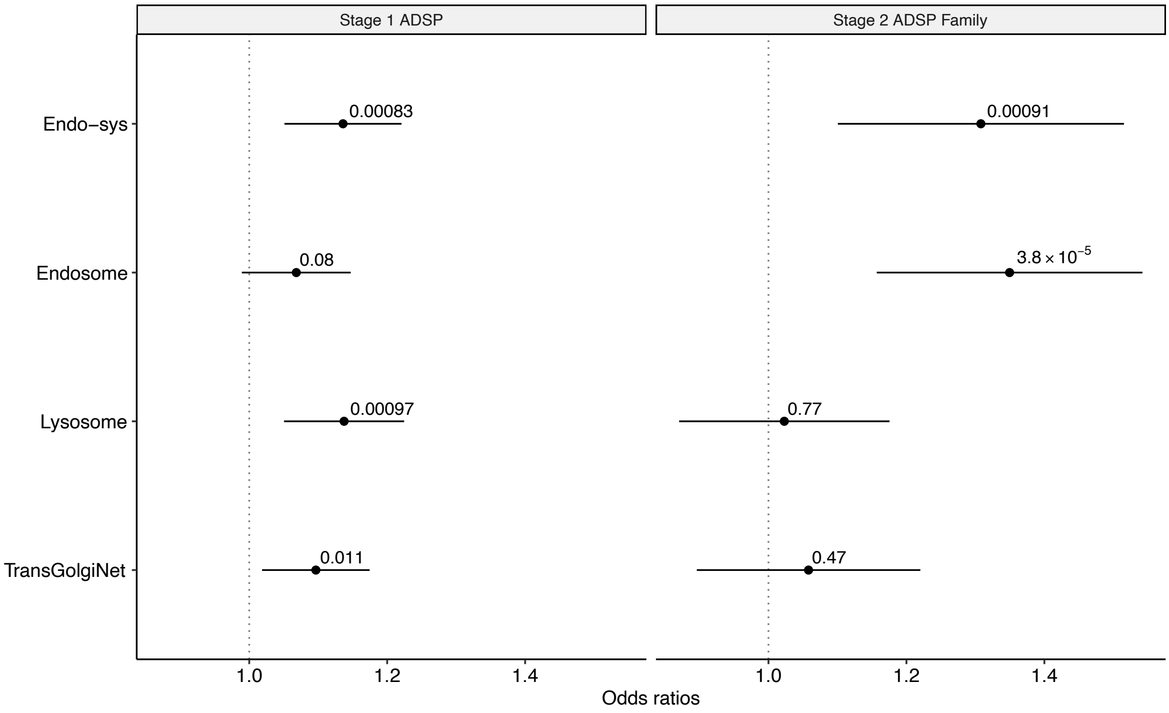


Figure 2-2. The enrichment of rare deleterious variants is associated with AD AAO across the endocytic and corresponding compartmental gene-sets in stages 1 and 2.

We computed a hazard ratio of obtaining AD in earlier ages using the burden of rare deleterious variants across the endo-system gene-set and three compartmental sub-gene-sets (endosome, lysosome, and trans-Golgi network) in stage 1 ADSP case-control dataset (left), which were then tested for replication in stage 2 ADSP family datasets (right). Enrichment (ORs) and p-value were computed using CPHR. P-values of enrichment in each gene-set are indicated above horizontal bars which represent 95% confidence intervals.

To test for replication, we examined the ADSP family study under the same statistical framework. Applying the CPHR models, we observed a gene-set-wide significant hazard ratio of 1.31 (p=0.00091 using CPHR; Fig 2) in the endo-system gene-set. Carefully examining the sub-gene-sets also revealed gene-set-wide significant AAO-specific risk within the endosome gene-set (HR=1.35, p=$3.83*10^{-5}$ using CPHR). We did not observe significant associations using the other two compartmental gene-sets (S7 Table).

To increase power, we performed meta-analyses to identify rare-variant effects shared across multiple studies. We combined the best results from ADSP case-control and family studies and observed a gene-set-wide significant p-value of $2.47*10^{-6}$ (by METAL; Fisher's method produced similar results; Table 2) for the endo-system gene-set, which was greatly improved compared to results in either stage. Similarly, the endosome gene-set also demonstrated an improved gene-set-wide significant p-value of $3.33*10^{-5}$. However, the lysosome and the trans-Golgi network gene-sets showed only nominally significant p-values in our meta-analysis, potentially due to the absence of signal in the ADSP family study. These findings strongly demonstrated that this AAO-specific rare-variant effect in the endocytic pathway was shared in European samples across different studies.

Another age-specific phenotype is the age of death (AOD), which has been shown to be affected by genetic groups implicated in AD AAO as well as in other dementia.[109,110] We thus followed the same analysis framework using the CPHR model and assessed whether

AD-affected patients with earlier AOD were associated with a higher rare-variant burden in the endocytic pathway. We looked at European samples in the AMP-AD case-control study, where the AOD information was available. We observed a hazard ratio of 1.10 (p=0.024 using CPHR; S2 Fig), indicating an increase of risk of death in AD patients as well as a worse prognosis along with an elevation in genetic burden. Further analysis using the lysosome sub-gene-set displayed a hazard ratio of 1.09 (p=0.036 using CPHR). Both endo-system and lysosome gene-sets demonstrated nominally significant associations with AOD but did not reach gene-set-wide significance after multiple-testing correction. Analysis using other sub-gene-sets did not provide significant hazard ratios.

Single-gene analysis on AD risk using AD and NFT status

From the previous analysis, the endo-system gene-set conferred a large rare-variant effect on AD and related phenotypes. Thus, we decided to examine the effect of rare variants in single endocytic genes, attempting to identify those associated with AD with large effect sizes. To increase power, we aggregated previously defined rare deleterious SNVs in each gene and tested for association with AD. We did not observe a single gene passing the Bonferroni corrected significance threshold in all three datasets, as well as in meta-analysis (See Methods; α=$4.18*10^{-5}$; $4.07*10^{-5}$; $7.32*10^{-5}$; $7.79*19^{-5}$, for ADSP case-control, AMP-AD case-control, ADSP family studies, and meta-analysis respectively; S15 Table).

As mentioned previously, NFT status may provide more detailed information of the pathological progression of AD and thus a greater power to detect signals of rare-variant effect. We, therefore, performed single-gene analysis using NFT status, as a proxy for AD status. For all datasets, we retained only rare deleterious SNVs that were present in samples with Braak staging information. We controlled for the same set of covariates as in previous analyses, except that we also included the AD phenotype (AD affected / unaffected) for each individual as one additional covariate (see Methods). The latter is necessary because the Braak staging and the AD phenotype are correlated, and the numbers of individuals with and without AD were vastly disproportionate among the samples with Braak staging information. For the ADSP case-control study, we observed six genes that reached Bonferroni corrected significance threshold (α=4.83*10^{-5}). None of the genes passed the Bonferroni corrected significance threshold (α=4.25*10^{-5}) in the AMP-AD study. Results of the top ten most significant genes can be found in S8 Table. We conducted meta-analyses for these two independent studies using MetaSKAT as before. In the combined results, we observed one gene, *ANKRD13D*, reached Bonferroni corrected significance threshold (p=3.56e-05; α=5.17*10^{-5}). This gene has been previously implicated in AD through RNA expression analysis[111] and protein interactome mapping[112].

The identification of functional effects of rare variants within the endocytic pathway

The hypothesis that the endo-system gene-set contains rare variants that are influential to AD development is endorsed by the previous gene-set burden analyses and single-gene analyses. One approach to investigating how the effect of rare variants takes place is to analyze how these rare variants are associated with gene expression. Such gene containing variations affecting its expression is often called an eGene.[91] To identify eGenes, we obtained bulk RNA-Seq data of DLPFC brain tissues of 636 individuals from the ROSMAP study as part of the AMP-AD study and tested for association of all variants in *cis* with a gene with its gene expression. Specifically, we grouped all variants within one gene, as well as those near the corresponding TSS, and assessed whether the aggregated rare-variant effect in an endocytic gene is associated with its expression level using SKAT (see Methods). Intersecting the bulk RNA-Seq and WGS data revealed 547 individuals with 224 AD patients and 323 controls. By taking an FDR of 5%, we discovered two genes, *HLA-A* and *SLC26A7*, whose rare variants were significantly associated with expressional changes. To note, previous studies have demonstrated that proteins from the same families of these two genes are associated with AD status. Specifically, two proteins from the HLA families and one from the SLC families have been implicated in AD through meta-analyses of large GWAS and brain DNA-methylation association analysis.[9,113] We first examined their single-gene analysis results and observed that none of them was significant using the AMP-AD dataset (p=4.74e-01; 2.14e-01, for *HLA-A* and *SLC26A7*, respectively). To validate our results and determine the direction of effects, we compared the expression of these two genes between cases and controls. Indeed, their expression levels were both significantly decreased in cases compared to controls (p=0.00073 *HLA-A*; Fig 3a; p=0.0054 *SLC26A7*; Fig 3b; student t-test; α=0.017;). We further examined the distribution of their expression levels across multiple Braak stages. Similarly, both expressions were strongly negatively correlated with greater Braak stages (r=-0.129, p=0.0024 *HLA*-A; r=-0.127, p=0.0029 *SLC26A7*; Pearson correlation; α=0.017).

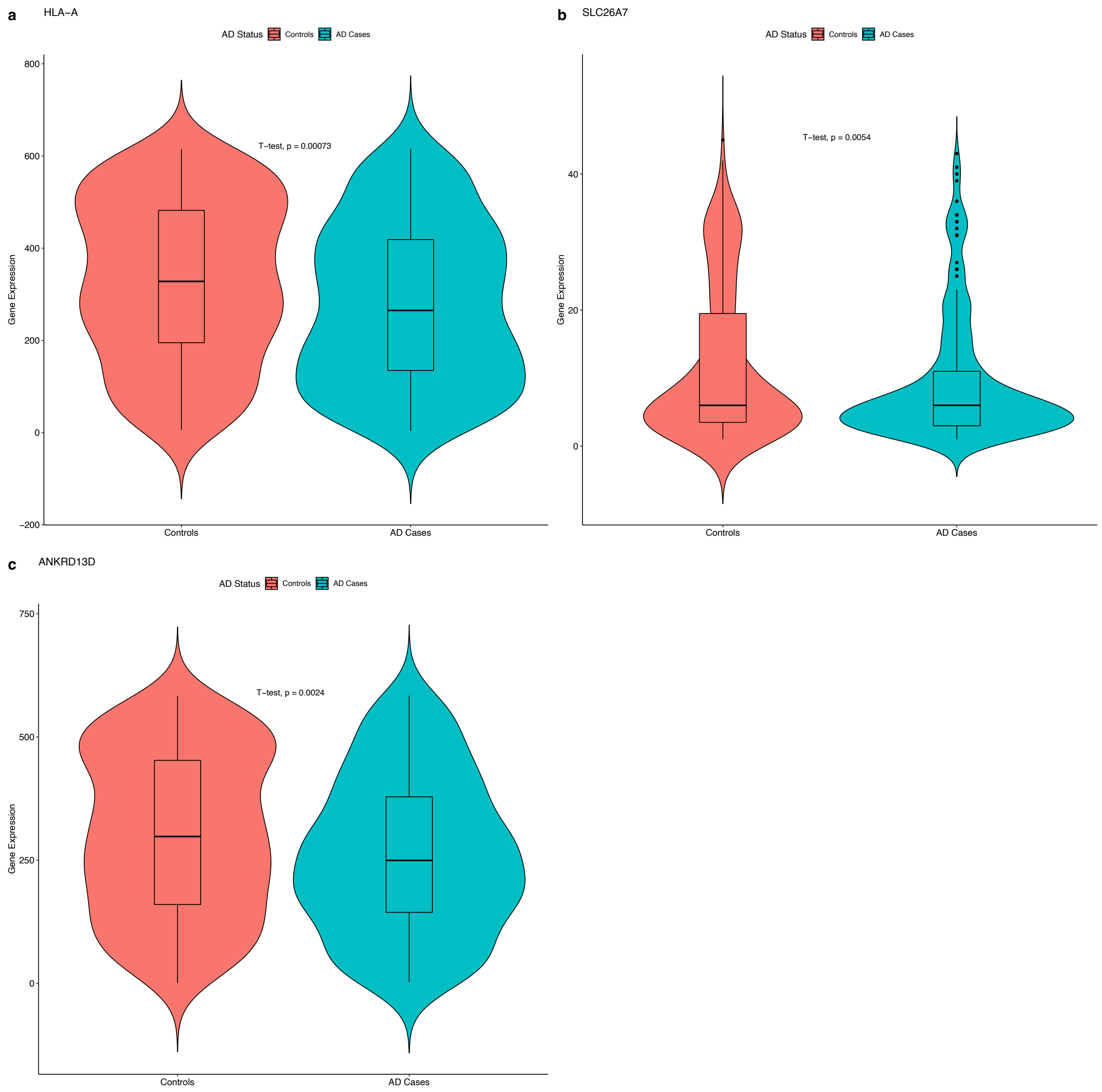


Figure 2-3. Comparison of the gene expression of HLA-A, SLC26A7, and ANKRD13D between AD cases and controls from the ROSMAP study.

Violin plots were used to represent the distribution of gene expression within each AD status, where a symmetric deviation from the middle line on both sides indicated a higher abundance of samples at the corresponding gene expression level. Comparisons between AD cases and controls were assessed using boxplots. P-values were computed using the student t-test. All three genes, *HLA-A, SLC26A7,* and *ANKRD13D,* are down-regulated in AD cases compared to controls.

We also investigated *ANKRD13D*, which we previously identified to be associated with Braak stages, in the context of gene expression. Although not an eGene, *ANKRD13D* exhibited a significant expressional decrease in cases compared to controls (p=0.0026; student t-test; Fig 3c). The analysis on Braak staging also revealed a strong negative correlation (r=-0.122, p=0.0042; Pearson correlation)

scRNA expression analysis

Recent advancement in analyzing gene expression in single-cell resolution has provided opportunities to uncover complex alterations across cell types and identify cell-type specific effects on AD.[55,61,93,94] For example, previous studies have pointed the imbalance of excitatory and inhibitory neurons could lead to overexcitability and early dysregulation in the development of AD [114]. Many other studies also demonstrated abnormalities in innate immune cells, primarily microglia, in the pathogenesis of AD.[115] Therefore, to investigate the potential cell-type specific effects of rare variants within the endocytic pathway, we obtained the single-cell RNA-seq data of 48 samples (24 AD patients and 24 cognitively normal controls) from the ROSMAP study. We focused on three genes we identified through the previous analysis, which demonstrated significant associations to AD progression. The scRNA-Seq data were labeled with six major cell types using a priori marker genes (Ex, In, Ast, Oli, Mic, and Opc), and sub-clustering within each cell type revealed cellular subpopulation (see Methods). We examined the expression of the three target genes in all major cell types and observed that *ANKRD13D* was up-regulated in Ex (p=1.92*10-18; student t-test), Ast (p=0.011; student t-test), and In (p=0.028; student t-test) (S9 Table). However, it exhibited a down-regulation in Oli (p=0.0018; student t-test). *SLC26A7* was observed to be up-regulated in Ex (p=0.0049; student t-test), while *HLA-A* displayed a pattern of down-regulation in both In and Mic (p=9.72*10-6and p=0.0031 respectively; student t-test). Four AD pathology-associated cellular subpopulations (Ex4, In0, Ast1, and Oli0) have been previous demonstrated for this scRNA-Seq data[61,91]. Our differential expression analysis within these four subpopulations showed a pattern of up-regulation of *ANKRD13D* in Ex4 and In0 (p=5.76*10-8 and p=0.036, respectively; student t-test; S10 Table). The other two genes, however, were not significantly differentially expressed in these four cell subpopulations.

Discussion

Using large publicly available WGS datasets, our study described here enabled us to assess the contribution of rare variants to AD. In our stage 1 discovery phase, we observed a significantly elevated burden of rare deleterious SNVs in affected individuals compared to cognitively normal older controls within the endocytic pathway. We chose this pathway because it represented one of the earliest morphological changes in AD development, and multiple AD risk factors, predominantly through common SNPs, have been implicated specifically in this pathway with genome-wide significance, including *BIN1*, *CD2AP*, *PICALM*, *RIN3*, and *SORL1*.[9,12,18,19,21,22,116] Our results demonstrated additional correlation between rare variants in the endocytic pathway and AD. Successful replication in the AMP-AD case-control study and improved meta-analysis association further strengthened this contribution of rare deleterious variants to AD risk. Our analysis using the ADSP family dataset showed a similar enrichment of rare deleterious SNVs in AD patients, although not reaching gene-set-wide significance. One possible explanation was that the sample size of this family study was relatively small (one third to one fifth) compared to the other two case-control studies. We additionally identified gene-set-wide-significant signals within the endosome and lysosome gene-sets using meta-analysis, implicating potential compartment-specific roles in AD pathology. One possibility that we did not observe significant results in separate stages for all three sub-gene-sets was because they contained a smaller number of genes compared to the endo-system gene-set and, therefore, smaller aggregated effects of rare variants, which required meta-analysis to combine signals in individual samples. As the smallest gene-set (one-third to one-fourth of the other two), the trans-Golgi network remained nominally significant even after meta-analysis.

In assessing the AD pathological progression, we examined the association of rare-variant effect to NFT pathology using Braak staging. We observed the gene-set-wide significant association within the lysosome gene-set, where individuals with higher Braak stages were enriched with rare deleterious SNVs. No significant association was found in other gene-sets, besides a nominally significant association in the whole endocytic pathway. Compared to the previous analysis using AD status, our analysis using Braak stages was largely limited in sample size. For example, only 626 out of 1,291 European samples in the ADSP case-control dataset had Braak staging information available. For the ADSP family study, only 38 out of 353 samples had Braak information available, which made analyzing Braak stages in this dataset infeasible. Additionally, this was further complicated by the disproportionate distribution of samples across different Braak stages. The ADSP case-control dataset contained 218 samples in stage VI while only 15 samples in stage 0. Such highly skewed distribution reduced our power to detect a significant association between rare variants' effects and Braak stages. The AMP-AD dataset was similarly skewed but also distributed largely differently from the stage 1 dataset. This distinction in distribution may explain why we observed different signals in our stage 1 and 2 analyses.

Based on the idea that rare variants within the endocytic pathway were associated with AD progression, we further tested age-specific phenotypes and leveraged a CHPR model previously proposed to be effective in assessing the effect of variants on age-to-event risk.[86] For AAO, we observed a gene-set-wide significant hazard ratio in the stage 1 analysis, indicating an association of rare-variant burden in the endocytic pathway to earlier AAO of AD, which was replicated in stage 2. A similar observation was found in the compartmental gene-sets, where endosome gene-set demonstrated a gene-set-wide significant signal in meta-analysis. Nonetheless, we did not replicate our stage 1 findings of the lysosome gene-set in stage 2, potentially due to the small sample size of the family dataset and the small size of the gene-set. For AOD, we examined the AMP-AD dataset and only observed nominally significant signals in the endocytic pathway and the lysosomal compartment. Previous analyses on AAO have demonstrated a substantial correlation of AAO between parents and their children, with multiple risk loci, such as *APOE, GRN, MPT,* and *C9orf72*.[101,109] Genetic studies using AOD from LOAD datasets have revealed additional associations of SNVs in these genes with human aging.[110] Consistent in the observation of significant genetic components, our results discovered an additional contribution of rare variants within the endocytic pathway to age-related phenotypes.

Our discovery of the increased burden of rare-variant effect in AD patients led us to explore the effect of individual genes within the endocytic pathway and attempt to identify specific ones with large effect sizes which might serve as potential clinical and therapeutic targets. We performed single-gene analysis using both AD status and Braak staging as the target phenotypes. When looking at the AD status, we did not observe a gene with a large enough effect to be detected in our analysis. Using Braak staging information, we were able to identify one gene, *ANKRD13D*, that showed robust signal across multiple studies after multiple-testing correction. This may be due to the fact that Braak stages provided a finer indication of AD progression. *ANKRD13D* encodes a member of the Ankyrin repeat domain 13 family, characterized by three ankyrin repeats at the N-terminal facilitating protein-protein interaction.[117] It has been experimentally shown to localize to endosomes and is known to regulate the rapid ubiquitin-dependent internalization and sorting of membrane-bound proteins within the endocytic pathway.[118] One of its main targets is the endocytosis of the epidermal growth factor receptor (EGFR) through the functional ubiquitin-interacting motif (UIM) of the ANKRD13 family proteins, which is then degraded in lysosomes.[118,119] EGFR is a transmembrane protein serving as a receptor epidermal growth factor (EGF) protein ligands. Multiple previous studies have reported abnormal plasma levels of EGF in AD patients,[120-122] and two recent studies on EGF have demonstrated its protective effects on AD by preventing amyloid-beta (Aβ)-induced angiogenesis deficit to brain endothelial cells in vitro and in vivo.[123,124] Recent studies have also described that the EGFR internalization after EGF binding was strongly inhibited when ANKRD13 proteins were over-expressed.[118] This mechanism implicates a potential regulatory effect of the *ANKRD13* family on AD pathology through the regulation of internalization of EGFR. Indeed, the link between *ANKRD13D* and AD is further bolstered by a recent RNA profiling where they identified an altered gene expression of *ANKRD13D* between the blood and brain tissue of AD patients.[111] In our analysis, we identified seven rare deleterious SNVs within *ANKRD13D*, where six were predicted to be missense damaging variants and one was predicted to be either missense damaging or splice region variant. These mutations could potentially alter its ubiquitin-binding ability, either through directly changing the sequence or indirectly through changing the 3D protein folding structure, and affect the normal protective function of EGF in AD development. Further functional studies of *ANKRD13D*, and in particular these seven variants, will be needed to specifically define its role in AD pathogenesis and evaluate the therapeutic and clinical importance of the EGFR pathway.

To investigate the functional effects of rare variants, we looked at the expression of genes in the endocytic pathway at both bulk tissue and single-cell resolutions. Leveraging bulk RNA-Seq data, we identified two significant eGenes, *HLA-A* and *SLC26A7*, in the ROSMAP study. Careful examination of these two eGenes in the context of AD status revealed a pattern of down-regulation in AD patients compared to cognitively normal controls. A similar negative correlation was found using Braak stages. *HLA-A* encodes a member of the human leukocyte antigen A (HLA) class I, also called the major histocompatibility complex (MHC) class I. It has been shown to participate in the important "cross-presentation" mechanism of T cell-mediated immune response, specifically efficient in dendritic cells.[125] This mechanism is part of the endocytic pathway that involves the internalization of HLA class I proteins from the cell surface through early endosomes and the loading of antigen peptides in lysosomes.[126] Previous studies have described an important role of HLA class I in maintaining the integrity of aging brains and have demonstrated significant dendritic atrophy with deficient HLA class I.[127] Moreover, recent GWA studies have identified specific alleles in *HLA-A* associated with AD in the Italian and Chinese population [128,129], as well as risk loci in other members of the HLA family.[9] The other identified eGene, *SLC26A7*, encodes a member of the solute carrier (SLC) family that localizes to subapical lysosomal membrane as well as endosomes, primarily serving as an exchanger and transporter of a broad spectrum of substrates in the endocytic pathway.[130,131] Disruption in the expression of SLC26 proteins has been shown to cause severe acid-base balance dysregulation, leading to disruption of anion homeostasis.[132] Multiple SLCs have been associated with AD, such as *SLC2A2*, which was linked to astrocyte activation leading to its elevation in AD patients,[130] and *SLC1A3*, whose expression has been associated with Aβ deposition[133]. Recent GWA studies have also identified risk loci in members of *SLCs*, such as *SLC24A4*.[9] Specific implication of *SLC26A7* has also been shown through gene co-expression network mining where STAT1, a transcription factor of *SLC26A7*, was differentially expressed between AD patients and cognitively normal controls.[47] In our analysis, we identified nine rare deleterious SNVs in *HLA-A* in which six were predicted to be damaging missense mutations, two were predicted to be splice acceptor variants, and one was predicted to be either damaging missense mutation or splice region variant. In *SLC26A7*, we also identified nine rare deleterious SNVs, which are all damaging missense mutations. As transporters, these two genes could potentially be altered in their affinities to ligands due to changes in primary or tertiary structures. Our results here supported these previous findings and provided additional evidence from the aspect of the rare-variant effect on gene expression. Further investigation will be required to elucidate specific variants conferring these effects as well as other participating proteins in the same signal relay mechanisms of *HLA-A* and *SLC26A7*.

In a single-cell resolution, we further explored the cell-type-specific functional effects of the significant genes identified in our previous analyses. Previous single-cell transcriptomic analyses have shown a large number of cellular subpopulations with cell type-specific associations with AD.[61] Our analysis supported this finding in *ANKRD13D*, *HLA-A*, and *SLC26A7*. For example, we observed an up-regulation of *ANKRD13D* in bulk tissue, but it was found to be regulated differently in different cell types: up-regulated in Ex, Ast, and In, while down-regulated in Oli. On the other hand, in single-cell RNA-Seq data, *SLC26A7* and *HLA-A* showed a pattern of down-regulation in AD patients, consistent with our findings using the bulk RNA-Seq data though with various effect sizes in different cell types.

Several strengths and limitations of our study warrant discussion. One of the major strengths is our study design to begin the analysis with pathways implicated in AD a priori. Our usage of the endocytic pathway provided us the power to identify rare-variant effects that would otherwise be missed in traditional association analysis of single variants. This design was further combined with the large sample sizes of the three independent datasets, which provided additional power. We separated these datasets into a discovery phase and a replication phase and were able to replicate our discovery phase results in two independent datasets of the replication phase, followed by meta-analyses of samples in all three studies. This procedure ensured us to identify and validate associations while retaining large power to identify small signals. Another strength of our study is the analysis of AD-related phenotypes, such as Braak stages, and provided additional power in identifying single genes with large aggregated rare-variant effect sizes. The analysis of AAO and AOD provided further information on the progression of AD, which is especially important in clinal AD prediction and intervention. One more strength in our analysis lies in our exploitation of bulk- and sc-RNA expression data in combination with AD genotyping data. Through this method, we were able to identify eGenes with large rare-variant effect, which would require a much higher sample size and greater power to be identified as eQTLs and suggested potential AD-regulating mechanisms.

One limitation of the study is that while we used WGS datasets, we only focused on analyzing rare SNVs within genic regions. Our analysis relied on knowing the deleteriousness of each variant contributing to the gene-set burden, and variant annotation is most reliably predicted for coding and splice site variants.[90,134] Including variants in intergenic regions or indels may result in the inclusion of variants with benign effects and decrease our power of detecting AD-associated genetic burden. Another limitation of our study is that even though we utilized WGS datasets of large sample size, they were not large enough to detect single genes where rare variants significantly influenced AD. Although our analyses displayed sufficient power to detect rare-variant effects within sets of genes, we nonetheless failed to directly identify direct gene-level associations with AD. To achieve this latter goal, we may need WGS datasets of larger sample sizes. A similar limitation on sample sizes was seen in those with expression data and Braak staging information. Our bulk RNA-Seq data is only available for 547 individuals from the ROSMAP study in which we have genotyping data for 1200 individuals. The scRNA-Seq data is further limited in that we have 48 samples from the ROSMAP study. These limitations in sample size decreased our capability of detecting functional effects of rare variants within the endocytic pathway. One more limitation in this study is that we primarily focused on European samples because we had a limited sample size for non-European ancestries across all three WGS datasets. Nonetheless, it may be of interest to check whether we would observe similar rare-variant effect in the endocytic pathway in non-European samples as we observed in European samples. Another limitation rooted in the potential batch effects among the ADSP datasets used in this study, as also mentioned in Holstege et al.[135], due to the fact that the samples were sequenced and called in different locations. In this study, we have addressed the potential batch effect from three aspects. Firstly, the version of the ADSP datasets used in this study has been quality controlled, where all samples from different centers were re-processed using the same VCPA 1.0 pipeline and corrected for many technical issues present in the previous version, including contaminations, mismatches, and duplicates.[136] Secondly, we conducted additional QC steps at variant-level and sample-level. These included many steps suggested by Holstege et al., such as sex-check, selecting European samples by PCA, removing unexpected related samples using IBD, checking for samples with aberrant Ti/Tv ratio or novel SNV/indel count, and filtering out variants failing VQSR, GQ, HWE, and missing rate thresholds. Thirdly, we included sequencing location as a covariate in all models (M0, M1, and M2) to account for potential batch effects. Therefore, in this study, we recognized and have carefully approached this limitation, as much as we could, to mitigate the potential batch effects.

In summary, our study demonstrated significant rare-variant effect within the endocytic pathway in European samples. Such effect was also associated with Braak stages and age-related phenotypes, suggesting a potential target for clinical and therapeutic studies. Further investigation within this pathway revealed one gene significantly associated with Braak stages and two eGenes with a pattern of differential expression between AD patients and cognitively normal controls. More functional studies will be necessary to gain a better understanding of their molecular mechanisms of how they participate in the processing and modification of AD-related proteins. In vitro and in vivo experiments on these genes will also provide further insights into the connections of genetic variants to their gene expression and elucidate protein signaling models that affect the pathogenic progression of AD.

Acknowledgments

We thank the ADSP, ADNI, and AMP-AD contributors who collected samples used in this study, as well as patients and their families, whose help and participation made this work possible. The data used in this study can be found using accession NG00067. Specific acknowledgment to each study is as follows:

Acknowledgments to ADSP:

The Alzheimer's Disease Sequencing Project (ADSP) is comprised of two Alzheimer's Disease (AD) genetics consortia and three National Human Genome Research Institute (NHGRI) funded Large Scale Sequencing and Analysis Centers (LSAC). The two AD genetics consortia are the Alzheimer's Disease Genetics Consortium (ADGC) funded by NIA (U01 AG032984), and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) funded by NIA (R01 AG033193), the National Heart, Lung, and Blood Institute (NHLBI), other National Institute of Health (NIH) institutes and other foreign governmental

and non-governmental organizations. The Discovery Phase analysis of sequence data is supported through UF1AG047133 (to Drs. Schellenberg, Farrer, Pericak-Vance, Mayeux, and Haines); U01AG049505 to Dr. Seshadri; U01AG049506 to Dr. Boerwinkle; U01AG049507 to Dr. Wijsman; and U01AG049508 to Dr. Goate and the Discovery Extension Phase analysis is supported through U01AG052411 to Dr. Goate, U01AG052410 to Dr. Pericak-Vance and U01 AG052409 to Drs. Seshadri and Fornage. Data generation and harmonization in the Follow-up Phases is supported by U54AG052427 (to Drs. Schellenberg and Wang).
The ADGC cohorts include: Adult Changes in Thought (ACT), the Alzheimer's Disease Centers (ADC), the Chicago Health and Aging Project (CHAP), the Memory and Aging Project (MAP), Mayo Clinic (MAYO), Mayo Parkinson's Disease controls, University of Miami, the Multi-Institutional Research in Alzheimer's Genetic Epidemiology Study (MIRAGE), the National Cell Repository for Alzheimer's Disease (NCRAD), the National Institute on Aging Late Onset Alzheimer's Disease Family Study (NIA-LOAD), the Religious Orders Study (ROS), the Texas Alzheimer's Research and Care Consortium (TARC), Vanderbilt University/Case Western Reserve University (VAN/CWRU), the Washington Heights-Inwood Columbia Aging Project (WHICAP) and the Washington University Sequencing Project (WUSP), the Columbia University Hispanic- Estudio Familiar de Influencia Genetica de Alzheimer (EFIGA), the University of Toronto (UT), and Genetic Differences (GD).
The CHARGE cohorts are supported in part by National Heart, Lung, and Blood Institute (NHLBI) infrastructure grant HL105756 (Psaty), RC2HL102419 (Boerwinkle) and the neurology working group is supported by the National Institute on Aging (NIA) R01 grant AG033193. The CHARGE cohorts participating in the ADSP include the following: Austrian Stroke Prevention Study (ASPS), ASPS-Family study, and the Prospective Dementia Registry-Austria (ASPS/PRODEM-Aus), the Atherosclerosis Risk in Communities (ARIC)

Study, the Cardiovascular Health Study (CHS), the Erasmus Rucphen Family Study (ERF), the Framingham Heart Study (FHS), and the Rotterdam Study (RS). ASPS is funded by the Austrian Science Fond (FWF) grant number P20545-P05 and P13180 and the Medical University of Graz. The ASPS-Fam is funded by the Austrian Science Fund (FWF) project I904),the EU Joint Programme - Neurodegenerative Disease Research (JPND) in frame of the BRIDGET project (Austria, Ministry of Science) and the Medical University of Graz and the Steiermärkische Krankenanstalten Gesellschaft. PRODEM-Austria is supported by the Austrian Research Promotion agency (FFG) (Project No. 827462) and by the Austrian National Bank (Anniversary Fund, project 15435. ARIC research is carried out as a collaborative study supported by NHLBI contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). Neurocognitive data in ARIC is collected by U01 2U01HL096812, 2U01HL096814, 2U01HL096899, 2U01HL096902, 2U01HL096917 from the NIH (NHLBI, NINDS, NIA and NIDCD), and with previous brain MRI examinations funded by R01-HL70825 from the NHLBI. CHS research was supported by contracts HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, and grants U01HL080295 and U01HL130114 from the NHLBI with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629, R01AG15928, and R01AG20098 from the NIA. FHS research is supported by NHLBI contracts N01-HC-25195 and HHSN268201500001I. This study was also supported by additional grants from the NIA (R01s AG054076, AG049607 and AG033040 and NINDS (R01 NS017950). The ERF study as a part of EUROSPAN (European Special Populations Research Network) was supported by European Commission

FP6 STRP grant number 018947 (LSHG-CT-2006-01947) and also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013)/grant agreement HEALTH-F4-2007-201413 by the European Commission under the programme "Quality of Life and Management of the Living Resources" of 5th Framework Programme (no. QLG2-CT-2002-01254). High-throughput analysis of the ERF data was supported by a joint grant from the Netherlands Organization for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR 047.017.043). The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, the Netherlands Organization for Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the municipality of Rotterdam. Genetic data sets are also supported by the Netherlands Organization of Scientific Research NWO Investments (175.010.2005.011, 911-03-012), the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), and the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) Netherlands Consortium for Healthy Aging (NCHA), project 050-060-810. All studies are grateful to their participants, faculty and staff. The content of these manuscripts is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the U.S. Department of Health and Human Services.

The four LSACs are: the Human Genome Sequencing Center at the Baylor College of Medicine (U54 HG003273), the Broad Institute Genome Center (U54HG003067), The American Genome Center at the Uniformed Services University of the Health Sciences (U01AG057659), and the Washington University Genome Institute (U54HG003079).

Biological samples and associated phenotypic data used in primary data analyses were stored at Study Investigators institutions, and at the National Cell Repository for Alzheimer's Disease (NCRAD, U24AG021886) at Indiana University funded by NIA. Associated Phenotypic Data used in primary and secondary data analyses were provided by Study Investigators, the NIA funded Alzheimer's Disease Centers (ADCs), and the National Alzheimer's Coordinating Center (NACC, U01AG016976) and the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS, U24AG041689) at the University of Pennsylvania, funded by NIA, and at the Database for Genotypes and Phenotypes (dbGaP) funded by NIH. This research was supported in part by the Intramural Research Program of the National Institutes of health, National Library of Medicine. Contributors to the Genetic Analysis Data included Study Investigators on projects that were individually funded by NIA, and other NIH institutes, and by private U.S. organizations, or foreign governmental or nongovernmental organizations.

Acknowledgments to ADNI:

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis

Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Acknowledgments to AMP-AD:

Mayo RNAseq Study- Study data were provided by the following sources: The Mayo Clinic Alzheimer's Disease Genetic Studies, led by Dr. Nilufer Ertekin-Taner and Dr. Steven G. Younkin, Mayo Clinic, Jacksonville, FL using samples from the Mayo Clinic Study of Aging, the Mayo Clinic Alzheimer's Disease Research Center, and the Mayo Clinic Brain Bank. Data collection was supported through funding by NIA grants P50 AG016574, R01 AG032990, U01 AG046139, R01 AG018023, U01 AG006576, U01 AG006786, R01 AG025711, R01 AG017216, R01 AG003949, NINDS grant R01 NS080820, CurePSP Foundation, and support from Mayo Foundation. Study data includes samples collected through the Sun Health Research Institute Brain and Body Donation Program of Sun City, Arizona. The Brain and Body Donation Program is supported by the National Institute of Neurological Disorders and Stroke (U24 NS072026 National Brain and Tissue Resource for Parkinson's Disease and Related Disorders), the National Institute on Aging (P30 AG19610 Arizona Alzheimer's Disease Core Center), the Arizona Department of Health Services (contract 211002, Arizona Alzheimer's Research Center), the Arizona Biomedical Research Commission (contracts 4001, 0011, 05-901 and 1001 to the Arizona Parkinson's Disease Consortium) and the Michael J. Fox Foundation for Parkinson's Research

ROSMAP- We are grateful to the participants in the Religious Order Study, the Memory and Aging Project. This work is supported by the US National Institutes of Health [U01 AG046152, R01 AG043617, R01 AG042210, R01 AG036042, R01 AG036836, R01 AG032990, R01 AG18023, RC2 AG036547, P50 AG016574, U01 ES017155, KL2 RR024151, K25 AG041906-01, R01 AG30146, P30 AG10161, R01 AG17917, R01 AG15819, K08 AG034290, P30 AG10161 and R01 AG11101.

ROSMAP (gene expression data) - The results published here are in whole or in part based on data obtained from the AD Knowledge Portal (https://adknowledgeportal.org). Study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA grants P30AG10161 (ROS), R01AG15819 (ROSMAP; genomics and RNAseq), R01AG17917 (MAP), R01AG30146, R01AG36042 (5hC methylation, ATACseq), RC2AG036547 (H3K9Ac), R01AG36836 (RNAseq), R01AG48015 (monocyte RNAseq) RF1AG57473 (single nucleus RNAseq), U01AG32984 (genomic and whole exome sequencing), U01AG46152 (ROSMAP AMP-AD, targeted proteomics), U01AG46161(TMT proteomics), U01AG61356 (whole genome sequencing, targeted proteomics, ROSMAP AMP-AD), the Illinois Department of Public Health (ROSMAP), and the Translational Genomics Research Institute (genomic). Additional phenotypic data can be requested at www.radc.rush.edu.

Mount Sinai Brain Bank (MSBB)- This work was supported by the grants R01AG046170, RF1AG054014, RF1AG057440 and R01AG057907 from the NIH/National Institute on Aging (NIA). R01AG046170 is a component of the AMP-AD Target Discovery and Preclinical Validation Project. Brain tissue collection and characterization was supported by NIH HHSN271201300031C.

Reference

1. Mendez MF. Early-onset Alzheimer's disease: nonamnestic subtypes and type 2 AD. Arch Med Res. 2012;43(8):677-85.

2. Burns A, Iliffe S. Dementia. BMJ. 2009;338:b75.

3. Hebert LE, Weuve J, Scherr PA, Evans DA. Alzheimer disease in the United States (2010-2050) estimated using the 2010 census. Neurology. 2013;80(19):1778-83.

4. 2021 Alzheimer's disease facts and figures. Alzheimers Dement. 2021;17(3):327-406.

5. Brookmeyer R, Johnson E, Ziegler-Graham K, Arrighi HM. Forecasting the global burden of Alzheimer's disease. Alzheimers Dement. 2007;3(3):186-91.

6. Naj AC, Jun G, Beecham GW, Wang LS, Vardarajan BN, Buros J, et al. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nat Genet. 2011;43(5):436-41.

7. Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, et al. Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry. 2006;63(2):168-74.

8. Shen L, Jia J. An Overview of Genome-Wide Association Studies in Alzheimer's Disease. Neurosci Bull. 2016;32(2):183-90.

9. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat Genet. 2013;45(12):1452-8.

10. Ridge PG, Mukherjee S, Crane PK, Kauwe JS, Alzheimer's Disease Genetics C. Alzheimer's disease: analyzing the missing heritability. PLoS One. 2013;8(11):e79771.

11. Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat Genet. 2019;51(3):404-13.

12. Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing. Nat Genet. 2019;51(3):414-30.

13. Small SA, Simoes-Spassov S, Mayeux R, Petsko GA. Endosomal Traffic Jams Represent a Pathogenic Hub and Therapeutic Target in Alzheimer's Disease. Trends Neurosci. 2017;40(10):592-602.

14. Guerreiro R, Wojtas A, Bras J, Carrasquillo M, Rogaeva E, Majounie E, et al. TREM2 variants in Alzheimer's disease. N Engl J Med. 2013;368(2):117-27.

15. Jonsson T, Stefansson H, Steinberg S, Jonsdottir I, Jonsson PV, Snaedal J, et al. Variant of TREM2 associated with the risk of Alzheimer's disease. N Engl J Med. 2013;368(2):107-16.

16. Cruchaga C, Karch CM, Jin SC, Benitez BA, Cai Y, Guerreiro R, et al. Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. Nature. 2014;505(7484):550-4.

17. Logue MW, Schu M, Vardarajan BN, Farrell J, Bennett DA, Buxbaum JD, et al. Two rare AKAP9 variants are associated with Alzheimer's disease in African Americans. Alzheimers Dement. 2014;10(6):609-18 e11.

18. Reitz C, Mayeux R, Alzheimer's Disease Genetics C. TREM2 and neurodegenerative disease. N Engl J Med. 2013;369(16):1564-5.

19. Sims R, van der Lee SJ, Naj AC, Bellenguez C, Badarinarayan N, Jakobsdottir J, et al. Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. Nat Genet. 2017;49(9):1373-84.

20. Wetzel-Smith MK, Hunkapiller J, Bhangale TR, Srinivasan K, Maloney JA, Atwal JK, et al. A rare mutation in UNC5C predisposes to late-onset Alzheimer's disease and increases neuronal cell death. Nat Med. 2014;20(12):1452-7.

21. Cataldo AM, Petanceska S, Terio NB, Peterhoff CM, Durham R, Mercken M, et al. Abeta localization in abnormal endosomes: association with earliest Abeta elevations in AD and Down syndrome. Neurobiol Aging. 2004;25(10):1263-72.

22. Cataldo A, Rebeck GW, Ghetri B, Hulette C, Lippa C, Van Broeckhoven C, et al. Endocytic disturbances distinguish among subtypes of Alzheimer's disease and related disorders. Ann Neurol. 2001;50(5):661-5.

23. Corlier F, Rivals I, Lagarde J, Hamelin L, Corne H, Dauphinot L, et al. Modifications of the endosomal compartment in peripheral blood mononuclear cells and fibroblasts from Alzheimer's disease patients. Transl Psychiatry. 2015;5:e595.

24. Cataldo AM, Peterhoff CM, Troncoso JC, Gomez-Isla T, Hyman BT, Nixon RA. Endocytic pathway abnormalities precede amyloid beta deposition in sporadic Alzheimer's disease and Down syndrome: differential effects of APOE genotype and presenilin mutations. Am J Pathol. 2000;157(1):277-86.

25. Schwartzentruber J, Cooper S, Liu JZ, Barrio-Hernandez I, Bello E, Kumasaka N, et al. Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. Nat Genet. 2021;53(3):392-402.

26. Heckmann BL, Teubner BJW, Tummers B, Boada-Romero E, Harris L, Yang M, et al. LC3-Associated Endocytosis Facilitates beta-Amyloid Clearance and Mitigates Neurodegeneration in Murine Alzheimer's Disease. Cell. 2020;183(6):1733-4.

27. Karch CM, Goate AM. Alzheimer's disease risk genes and mechanisms of disease pathogenesis. Biol Psychiatry. 2015;77(1):43-51.

28. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010;467(7317):832-8.

29. Menashe I, Maeder D, Garcia-Closas M, Figueroa JD, Bhattacharjee S, Rotunno M, et al. Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade. Cancer Res. 2010;70(11):4453-9.

30. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518(7538):197-206.

31. Eleftherohorinou H, Hoggart CJ, Wright VJ, Levin M, Coin LJ. Pathway-driven gene stability selection of two rheumatoid arthritis GWAS identifies and validates new susceptibility genes in receptor mediated signalling pathways. Hum Mol Genet. 2011;20(17):3494-506.

32. Wang K, Zhang H, Kugathasan S, Annese V, Bradfield JP, Russell RK, et al. Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. Am J Hum Genet. 2009;84(3):399-405.

33. Nurnberger JI, Jr., Koller DL, Jung J, Edenberg HJ, Foroud T, Guella I, et al. Identification of pathways for bipolar disorder: a meta-analysis. JAMA Psychiatry. 2014;71(6):657-64.

34. Askland K, Read C, Moore J. Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. Hum Genet. 2009;125(1):63-79.

35. Schizophrenia Working Group of the Psychiatric Genomics C. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511(7510):421-7.

36. Genovese G, Fromer M, Stahl EA, Ruderfer DM, Chambert K, Landen M, et al. Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. Nat Neurosci. 2016;19(11):1433-41.

37. Manshaei R, Merico D, Reuter MS, Engchuan W, Mojarad BA, Chaturvedi R, et al. Genes and Pathways Implicated in Tetralogy of Fallot Revealed by Ultra-Rare Variant Burden Analysis in 231 Genome Sequences. Front Genet. 2020;11:957.

38. Amanat S, Gallego-Martinez A, Sollini J, Perez-Carpena P, Espinosa-Sanchez JM, Aran I, et al. Burden of rare variants in synaptic genes in patients with severe tinnitus: An exome based extreme phenotype study. EBioMedicine. 2021;66:103309.

39. Sul JH, Service SK, Huang AY, Ramensky V, Hwang SG, Teshiba TM, et al. Contribution of common and rare variants to bipolar disorder susceptibility in extended pedigrees from population isolates. Transl Psychiatry. 2020;10(1):74.

40. Ripke S, O'Dushlaine C, Chambert K, Moran JL, Kahler AK, Akterin S, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat Genet. 2013;45(10):1150-9.

41. Qian DC, Byun J, Han Y, Greene CS, Field JK, Hung RJ, et al. Identification of shared and unique susceptibility pathways among cancers of the lung, breast, and prostate from genome-wide association studies and tissue-specific protein interactions. Hum Mol Genet. 2015;24(25):7406-20.

42. Manolio TA. Bringing genome-wide association findings into clinical use. Nat Rev Genet. 2013;14(8):549-58.

43. Sullivan PF, Daly MJ, O'Donovan M. Genetic architectures of psychiatric disorders: the emerging picture and its implications. Nat Rev Genet. 2012;13(8):537-51.

44. Schizophrenia Psychiatric Genome-Wide Association Study C. Genome-wide association study identifies five new schizophrenia loci. Nat Genet. 2011;43(10):969-76.

45. Manolio TA. Genomewide association studies and assessment of the risk of disease. N Engl J Med. 2010;363(2):166-76.

46. Xiao X, Jiao B, Liao X, Zhang W, Yuan Z, Guo L, et al. Association of Genes Involved in the Metabolic Pathways of Amyloid-beta and Tau Proteins With Sporadic Late-Onset Alzheimer's Disease in the Southern Han Chinese Population. Front Aging Neurosci. 2020;12:584801.

47. Xiang S, Huang Z, Wang T, Han Z, Yu CY, Ni D, et al. Condition-specific gene co-expression network mining identifies key pathways and regulators in the brain tissue of Alzheimer's disease patients. BMC Med Genomics. 2018;11(Suppl 6):115.

48. Abner EL, Kryscio RJ, Schmitt FA, Santacruz KS, Jicha GA, Lin Y, et al. "End-stage" neurofibrillary tangle pathology in preclinical Alzheimer's disease: fact or fiction? J Alzheimers Dis. 2011;25(3):445-53.

49. Braak H, Braak E, Bohl J. Staging of Alzheimer-related cortical destruction. Eur Neurol. 1993;33(6):403-8.

50. Braak H, Braak E. Neuropathological stageing of Alzheimer-related changes. Acta Neuropathol. 1991;82(4):239-59.

51. Beecham GW, Bis JC, Martin ER, Choi SH, DeStefano AL, van Duijn CM, et al. The Alzheimer's Disease Sequencing Project: Study design and sample selection. Neurol Genet. 2017;3(5):e194.

52. ADSP Discovery Extension Case-Control Sample Selection Criteria.

53. Leung YY, Valladares O, Chou YF, Lin HJ, Kuzma AB, Cantwell L, et al. VCPA: genomic variant calling pipeline and data management tool for Alzheimer's Disease Sequencing Project. Bioinformatics. 2019;35(10):1768-70.

54. ADNI procedue manual online protocol.

55. De Jager PL, Ma Y, McCabe C, Xu J, Vardarajan BN, Felsky D, et al. A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. Sci Data. 2018;5:180142.

56. Bennett DA, Schneider JA, Arvanitakis Z, Wilson RS. Overview and findings from the religious orders study. Curr Alzheimer Res. 2012;9(6):628-45.

57. Allen M, Carrasquillo MM, Funk C, Heavner BD, Zou F, Younkin CS, et al. Human whole genome genotype and transcriptome data for Alzheimer's and other neurodegenerative diseases. Sci Data. 2016;3:160089.

58. Wang M, Beckmann ND, Roussos P, Wang E, Zhou X, Wang Q, et al. The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. Sci Data. 2018;5:180185.

59. Blue EE, Bis JC, Dorschner MO, Tsuang DW, Barral SM, Beecham G, et al. Genetic Variation in Genes Underlying Diverse Dementias May Explain a Small Proportion of Cases

in the Alzheimer's Disease Sequencing Project. Dement Geriatr Cogn Disord. 2018;45(1-2):1-17.

60. Bis JC, Jian X, Kunkle BW, Chen Y, Hamilton-Nelson KL, Bush WS, et al. Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation. Mol Psychiatry. 2020;25(8):1859-75.

61. Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, et al. Single-cell transcriptomic analysis of Alzheimer's disease. Nature. 2019;570(7761):332-7.

62. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559-75.

63. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.

64. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904-9.

65. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-91.

66. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17(1):122.

67. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-9.

68. Saint Pierre A, Genin E. How important are rare variants in common disease? Brief Funct Genomics. 2014;13(5):353-61.

69. Todorovic V. Genetics. Predicting the impact of genomic variation. Nat Methods. 2016;13(3):203.

70. Rentzsch P, Schubach M, Shendure J, Kircher M. CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Med. 2021;13(1):31.

71. Gene Ontology C. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. 2021;49(D1):D325-D34.

72. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-9.

73. Mellman I. Endocytosis and molecular sorting. Annu Rev Cell Dev Biol. 1996;12:575-625.

74. Nakano A, Luini A. Passage through the Golgi. Curr Opin Cell Biol. 2010;22(4):471-8.

75. Settembre C, Fraldi A, Medina DL, Ballabio A. Signals from the lysosome: a control centre for cellular clearance and energy metabolism. Nat Rev Mol Cell Biol. 2013;14(5):283-96.

76. Jiao X, Sherman BT, Huang da W, Stephens R, Baseler MW, Lane HC, et al. DAVID-

WS: a stateful web service to facilitate gene/protein list analysis. Bioinformatics.

2012;28(13):1805-6.

77. Zhang X, Zhu C, Beecham G, Vardarajan BN, Ma Y, Lancour D, et al. A rare missense variant of CASP7 is associated with familial late-onset Alzheimer's disease. Alzheimers Dement. 2019;15(3):441-52.

78. Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, et al. Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. Am J Hum Genet. 2016;98(4):653-66.

79. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol. 2015;11(4):e1004219.

80. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190-1.

81. Chen Z, Yang W, Liu Q, Yang JY, Li J, Yang M. A new statistical approach to combining p-values using gamma distribution and its application to genome-wide association study. BMC Bioinformatics. 2014;15 Suppl 17:S3.

82. Chen Z, Huang H, Liu J, Tony Ng HK, Nadarajah S, Huang X, et al. Detecting differentially methylated loci for Illumina Array methylation data based on human ovarian cancer data. BMC Med Genomics. 2013;6 Suppl 1:S9.

83. Dewey M. metap: meta-analysis of significance values. 2020.

84. Ripley WNVaBD. Modern Applied Statistics with S. Fourth ed. New York: Springer; 2002.

85. Gotzsche PC. Why we need a broad perspective on meta-analysis. It may be crucially important for patients. BMJ. 2000;321(7261):585-6.

86. Desikan RS, Fan CC, Wang Y, Schork AJ, Cabral HJ, Cupples LA, et al. Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. PLoS Med. 2017;14(3):e1002258.

87. Terry M. Therneau PMG. Modeling Survival Data: Extending the Cox Model. New York: Springer; 2000.

88. A Kassambara MK, P Biecek. survminer: Drawing Survival Curves using 'ggplot2'. 2017.

89. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89(1):82-93.

90. Lee S, Teslovich TM, Boehnke M, Lin X. General framework for meta-analysis of rare variants in sequencing association studies. Am J Hum Genet. 2013;93(1):42-53.

91. Sul JH, Raj T, de Jong S, de Bakker PI, Raychaudhuri S, Ophoff RA, et al. Accurate and fast multiple-testing correction in eQTL studies. Am J Hum Genet. 2015;96(6):857-68.

92. Flutre T, Wen X, Pritchard J, Stephens M. A statistical framework for joint eQTL analysis in multiple tissues. PLoS Genet. 2013;9(5):e1003486.

93. Grubman A, Chew G, Ouyang JF, Sun G, Choo XY, McLean C, et al. A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. Nat Neurosci. 2019;22(12):2087-97.

94. De Strooper B, Karran E. The Cellular Phase of Alzheimer's Disease. Cell. 2016;164(4):603-15.

95. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36(5):411-20.

96. Hu YB, Dammer EB, Ren RJ, Wang G. The endosomal-lysosomal system: from acidification and cargo sorting to neurodegeneration. Transl Neurodegener. 2015;4:18.

97. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database (Oxford). 2016;2016.

98. Iwakiri M, Mizukami K, Ikonomovic MD, Ishikawa M, Hidaka S, Abrahamson EE, et al. Changes in hippocampal GABABR1 subunit expression in Alzheimer's patients: association with Braak staging. Acta Neuropathol. 2005;109(5):467-74.

99. Singh V, Dwivedi SN, Deo SVS. Ordinal logistic regression model describing factors associated with extent of nodal involvement in oral cancer patients and its prospective validation. BMC Med Res Methodol. 2020;20(1):95.

100. Beecham GW, Hamilton K, Naj AC, Martin ER, Huentelman M, Myers AJ, et al. Genome-wide association meta-analysis of neuropathologic features of Alzheimer's disease and related dementias. PLoS Genet. 2014;10(9):e1004606.

101. Li YJ, Scott WK, Hedges DJ, Zhang F, Gaskell PC, Nance MA, et al. Age at onset in two common neurodegenerative diseases is genetically controlled. Am J Hum Genet. 2002;70(4):985-93.

102. Daw EW, Payami H, Nemens EJ, Nochlin D, Bird TD, Schellenberg GD, et al. The number of trait loci in late-onset Alzheimer disease. Am J Hum Genet. 2000;66(1):196-204.

103. Marioni RE, Harris SE, Zhang Q, McRae AF, Hagenaars SP, Hill WD, et al. GWAS on family history of Alzheimer's disease. Transl Psychiatry. 2018;8(1):99.

104. Huang KL, Marcora E, Pimenova AA, Di Narzo AF, Kapoor M, Jin SC, et al. A common haplotype lowers PU.1 expression in myeloid cells and delays onset of Alzheimer's disease. Nat Neurosci. 2017;20(8):1052-61.

105. Naj AC, Jun G, Reitz C, Kunkle BW, Perry W, Park YS, et al. Effects of multiple genetic loci on age at onset in late-onset Alzheimer disease: a genome-wide association study. JAMA Neurol. 2014;71(11):1394-404.

106. Liu JZ, Erlich Y, Pickrell JK. Case-control association mapping by proxy using family history of disease. Nat Genet. 2017;49(3):325-31.

107. Kamboh MI, Barmada MM, Demirci FY, Minster RL, Carrasquillo MM, Pankratz VS, et al. Genome-wide association analysis of age-at-onset in Alzheimer's disease. Mol Psychiatry. 2012;17(12):1340-6.

108. Zhang Q, Sidorenko J, Couvy-Duchesne B, Marioni RE, Wright MJ, Goate AM, et al. Risk prediction of late-onset Alzheimer's disease implies an oligogenic architecture. Nat Commun. 2020;11(1):4799.

109. Moore KM, Nicholas J, Grossman M, McMillan CT, Irwin DJ, Massimo L, et al. Age at symptom onset and death and disease duration in genetic frontotemporal dementia: an international retrospective cohort study. Lancet Neurol. 2020;19(2):145-56.

110. Shi H, Belbin O, Medway C, Brown K, Kalsheker N, Carrasquillo M, et al. Genetic variants influencing human aging from late-onset Alzheimer's disease (LOAD) genome-wide association studies (GWAS). Neurobiol Aging. 2012;33(8):1849 e5-18.

111. Bai Z, Stamova B, Xu H, Ander BP, Wang J, Jickling GC, et al. Distinctive RNA expression profiles in blood associated with Alzheimer disease after accounting for white matter hyperintensities. Alzheimer Dis Assoc Disord. 2014;28(3):226-33.

112. Haenig C, Atias N, Taylor AK, Mazza A, Schaefer MH, Russ J, et al. Interactome Mapping Provides a Network of Neurodegenerative Disease Proteins and Uncovers Widespread Protein Aggregation in Affected Brains. Cell Rep. 2020;32(7):108050.

113. Yu L, Chibnik LB, Srivastava GP, Pochet N, Yang J, Xu J, et al. Association of Brain DNA methylation in SORL1, ABCA7, HLA-DRB5, SLC24A4, and BIN1 with pathological diagnosis of Alzheimer disease. JAMA Neurol. 2015;72(1):15-24.

114. Vico Varela E, Etter G, Williams S. Excitatory-inhibitory imbalance in Alzheimer's disease and therapeutic significance. Neurobiol Dis. 2019;127:605-15.

115. Leng F, Edison P. Neuroinflammation and microglial activation in Alzheimer disease: where do we go from here? Nat Rev Neurol. 2021;17(3):157-72.

116. Van Acker ZP, Bretou M, Annaert W. Endo-lysosomal dysregulations and late-onset Alzheimer's disease: impact of genetic risk factors. Mol Neurodegener. 2019;14(1):20.

117. Li J, Mahajan A, Tsai MD. Ankyrin repeat: a unique motif mediating protein-protein interactions. Biochemistry. 2006;45(51):15168-78.

118. Tanno H, Yamaguchi T, Goto E, Ishido S, Komada M. The Ankrd 13 family of UIM-bearing proteins regulates EGF receptor endocytosis from the plasma membrane. Mol Biol Cell. 2012;23(7):1343-53.

119. Burana D, Yoshihara H, Tanno H, Yamamoto A, Saeki Y, Tanaka K, et al. The Ankrd13 Family of Ubiquitin-interacting Motif-bearing Proteins Regulates Valosin-containing Protein/p97 Protein-mediated Lysosomal Trafficking of Caveolin 1. J Biol Chem. 2016;291(12):6218-31.

120. Humpel C, Hochstrasser T. Cerebrospinal fluid and blood biomarkers in Alzheimer's disease. World J Psychiatry. 2011;1(1):8-18.

121. Doecke JD, Laws SM, Faux NG, Wilson W, Burnham SC, Lam CP, et al. Blood-based protein biomarkers for diagnosis of Alzheimer disease. Arch Neurol. 2012;69(10):1318-25.

122. Bjorkqvist M, Ohlsson M, Minthon L, Hansson O. Evaluation of a previously suggested plasma biomarker panel to identify Alzheimer's disease. PLoS One. 2012;7(1):e29868.

123. Thomas R, Zuchowska P, Morris AW, Marottoli FM, Sunny S, Deaton R, et al. Epidermal growth factor prevents APOE4 and amyloid-beta-induced cognitive and cerebrovascular deficits in female mice. Acta Neuropathol Commun. 2016;4(1):111.

124. Koster KP, Thomas R, Morris AWJ, Tai LM. Epidermal growth factor prevents oligomeric amyloid-beta induced angiogenesis deficits in vitro. J Cereb Blood Flow Metab. 2016;36(11):1865-71.

125. Cresswell P, Ackerman AL, Giodini A, Peaper DR, Wearsch PA. Mechanisms of MHC class I-restricted antigen processing and cross-presentation. Immunol Rev. 2005;207:145-57.

126. Basha G, Lizee G, Reinicke AT, Seipp RP, Omilusik KD, Jefferies WA. MHC class I endosomal and lysosomal trafficking coincides with exogenous antigen loading in dendritic cells. PLoS One. 2008;3(9):e3247.

127. Lazarczyk MJ, Kemmler JE, Eyford BA, Short JA, Varghese M, Sowa A, et al. Major Histocompatibility Complex class I proteins are critical for maintaining neuronal structural complexity in the aging brain. Sci Rep. 2016;6:26199.

128. Ma SL, Tang NL, Tam CW, Lui VW, Suen EW, Chiu HF, et al. Association between HLA-A alleles and Alzheimer's disease in a southern Chinese community. Dement Geriatr Cogn Disord. 2008;26(5):391-7.

129. Guerini FR, Tinelli C, Calabrese E, Agliardi C, Zanzottera M, De Silvestri A, et al. HLA-A*01 is associated with late onset of Alzheimer's disease in Italian patients. Int J Immunopathol Pharmacol. 2009;22(4):991-9.

130. Liu Y, Liu F, Iqbal K, Grundke-Iqbal I, Gong CX. Decreased glucose transporters correlate to abnormal hyperphosphorylation of tau in Alzheimer disease. FEBS Lett. 2008;582(2):359-64.

131. Alper SL, Sharma AK. The SLC26 gene family of anion transporters and channels. Mol Aspects Med. 2013;34(2-3):494-515.

132. Yin K, Lei Y, Wen X, Lacruz RS, Soleimani M, Kurtz I, et al. SLC26A Gene Family Participate in pH Regulation during Enamel Maturation. PLoS One. 2015;10(12):e0144703.

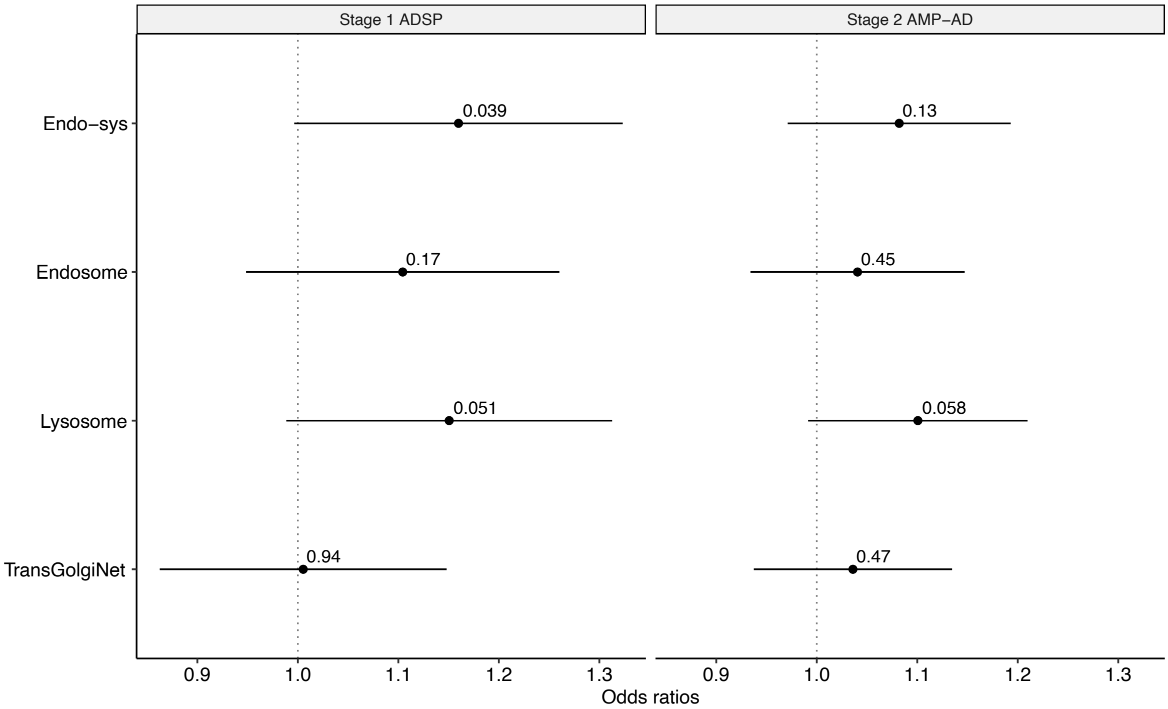
133. Hooijmans CR, Graven C, Dederen PJ, Tanila H, van Groen T, Kiliaan AJ. Amyloid beta deposition is related to decreased glucose transporter-1 levels and hippocampal atrophy in brains of aged APP/PS1 mice. Brain Res. 2007;1181:93-103.

134. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet. 2016;48(2):214-20.

135. Holstege H, Hulsman M, Charbonnier C, Grenier-Boley B, Quenez O, Grozeva D, et al. Exome sequencing identifies novel AD-associated genes. medRxiv. 2020:2020.07.22.20159251.

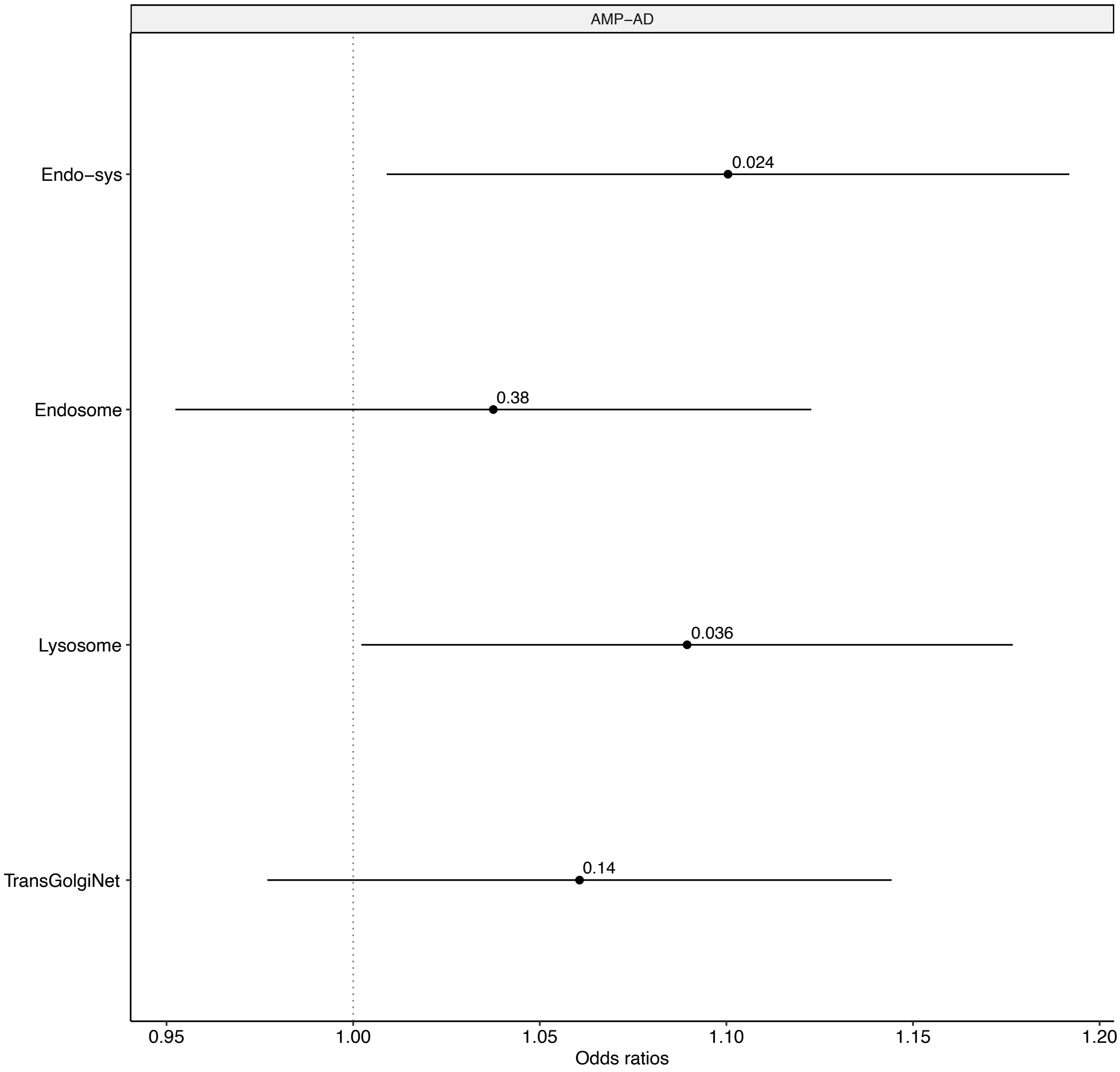
136. NIAGADS D. DSS Release Notes - NG00067.v6. 2021.

Supporting information



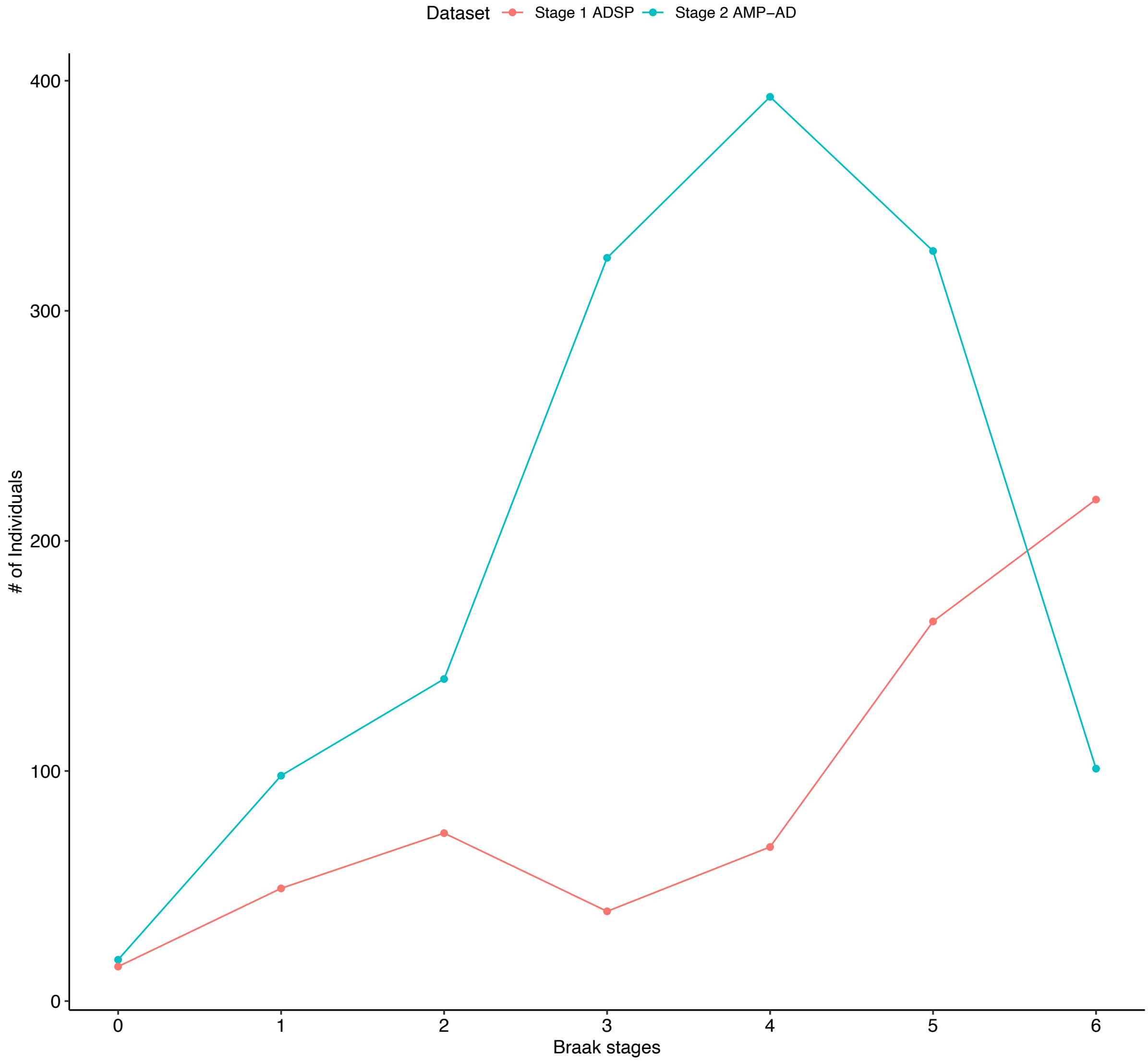
S1 Fig. Rare deleterious variants are enriched in patients with severe NFTs across the endocytic and corresponding compartmental gene-sets in stages 1 and 2.

We compared the burden of rare deleterious variants between patients with different severity of NFT across the endo-system gene-set and three compartmental sub-gene-sets (endosome, lysosome, and trans-Golgi network) in stage 1 ADSP case-control dataset (left), which were then tested for replication in stage 2 AMP-AD case-control dataset (right). Enrichment (ORs) and p-value were computed using OLR controlling for covariates, including the total count of rare variants (see Methods). P-values of enrichment in each gene-set are indicated above horizontal bars which represent 95% confidence intervals.

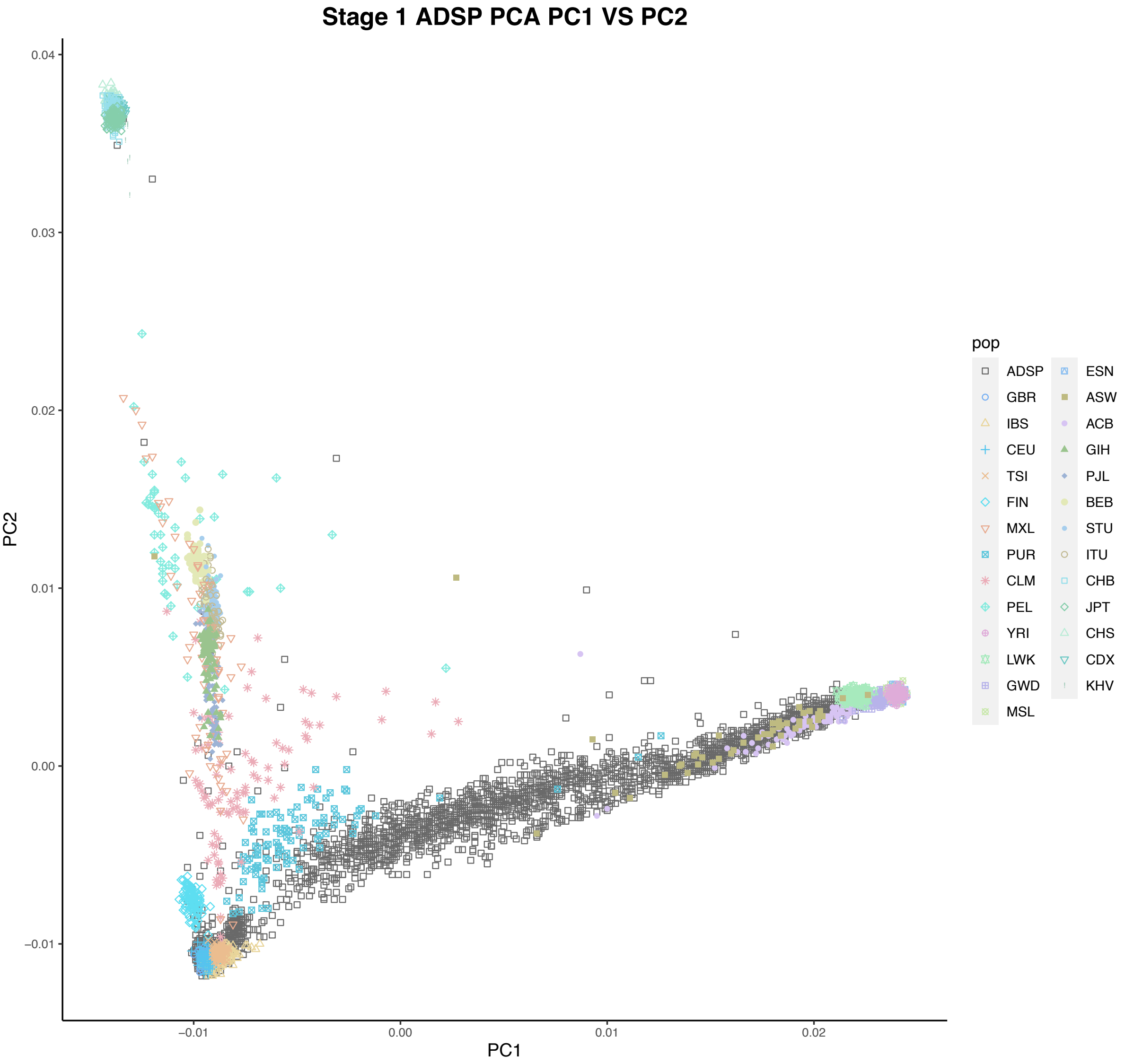


S2 Fig. The enrichment of rare deleterious variants is associated with AD AOD across the endocytic and corresponding compartmental gene-sets.

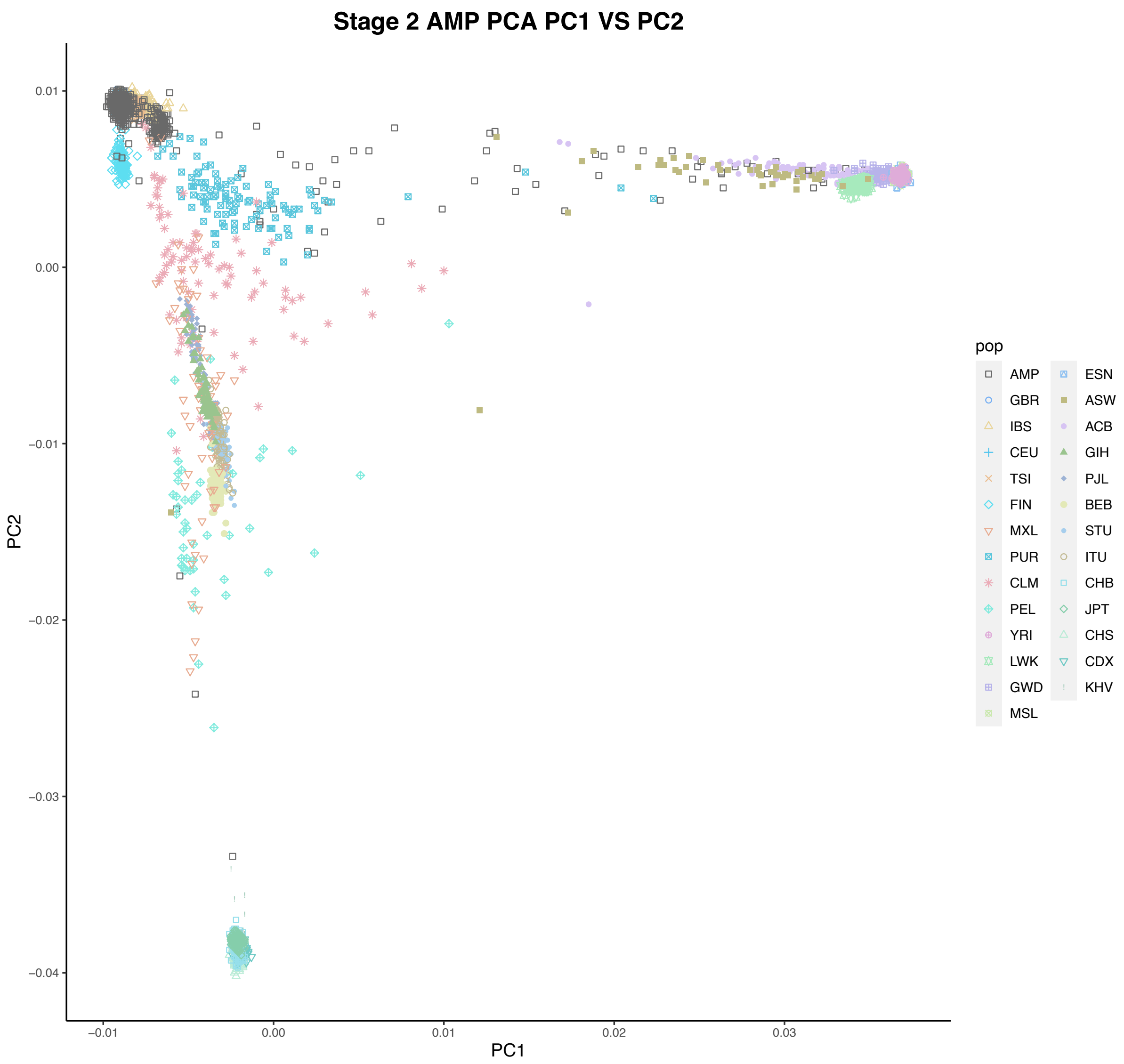
We computed a hazard ratio of earlier AOD with AD using the burden of rare deleterious variants across the endo-system gene-set and three compartmental sub-gene-sets (endosome, lysosome, and trans-Golgi network) in the AMP-AD study. Enrichment (ORs) and p-value were computed using CPHR controlling for covariates, including the total count of rare variants (see Methods). P-values of enrichment in each gene-set are indicated above horizontal bars which represent 95% confidence intervals.



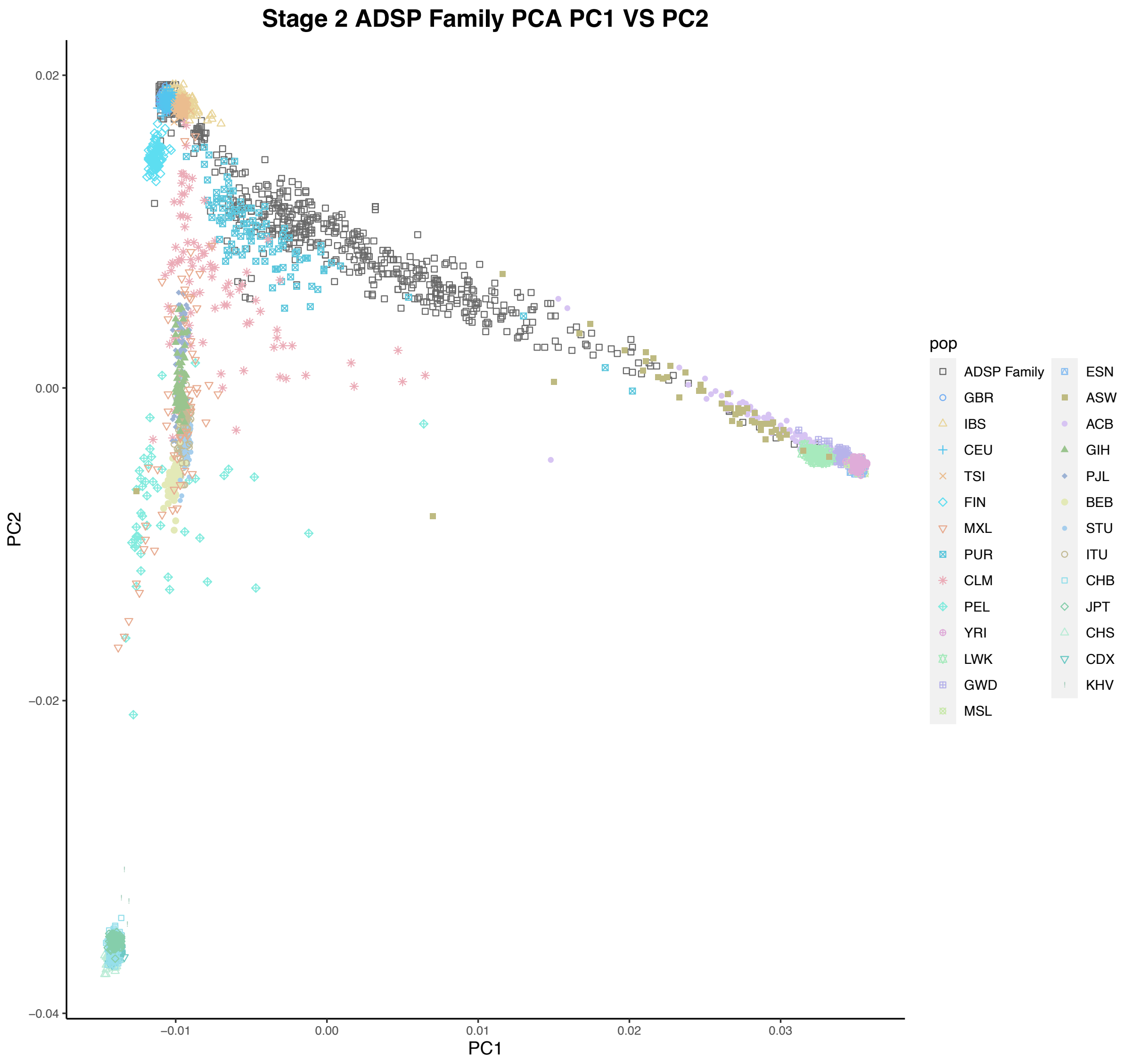
S3 Fig. Distribution of Braak stages in individuals from Stage 1 ADSP and Stage 2 AMP-AD datasets



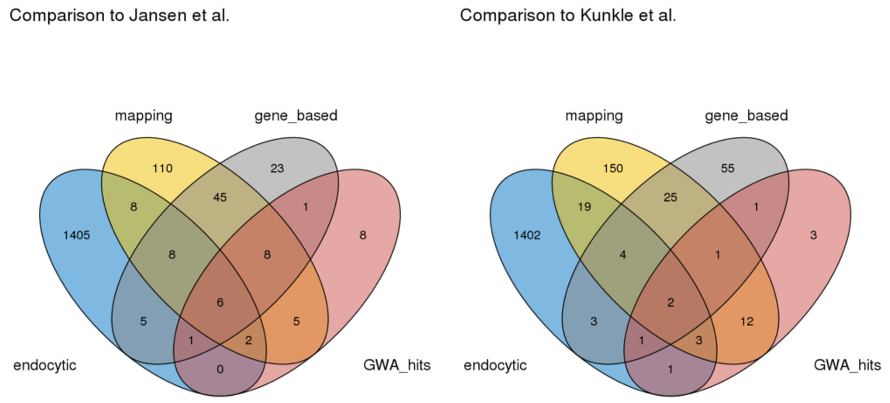
S4 Fig. PCA plots (PC1 vs. PC2) of the ADSP case-control dataset showing the distribution of ancestry backgrounds.

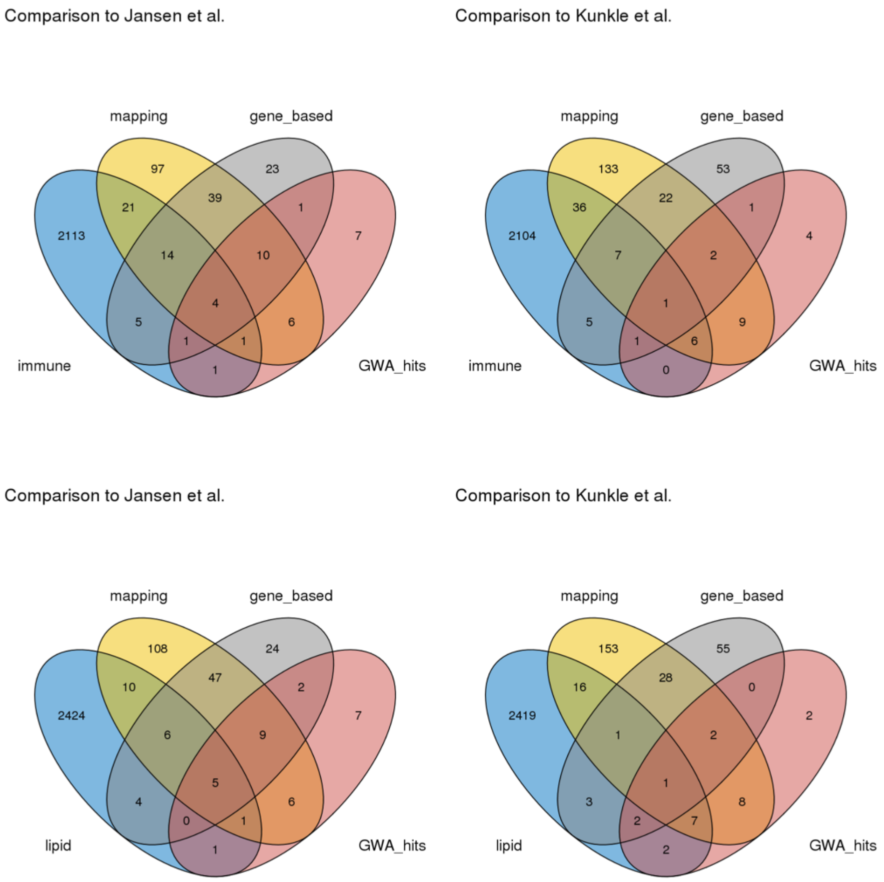


S5 Fig. PCA plots (PC1 vs. PC2) of the AMP-AD case-control dataset showing the distribution of ancestry backgrounds.



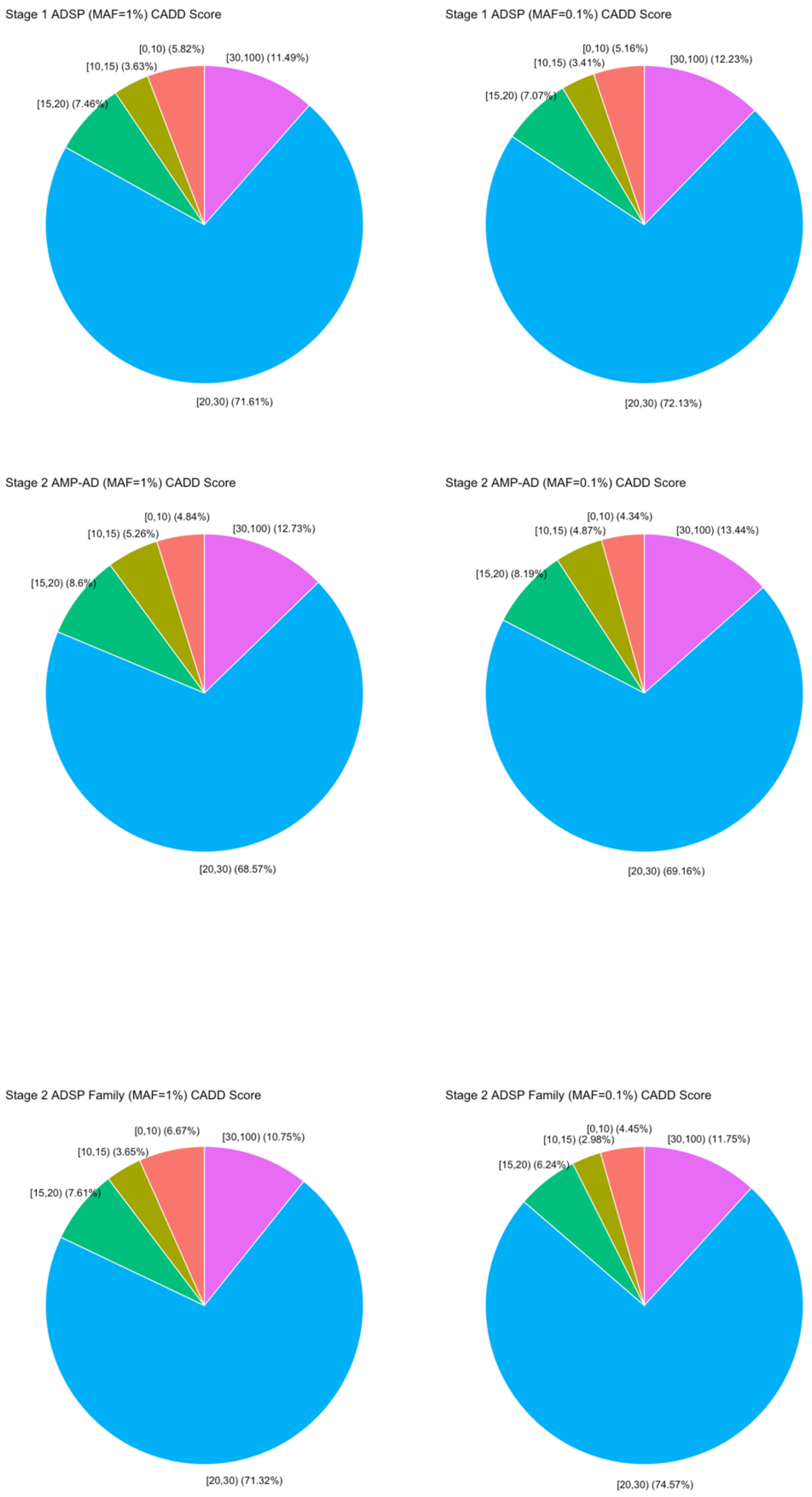
S6 Fig. PCA plots (PC1 vs. PC2) of the ADSP Family dataset showing the distribution of ancestry backgrounds.



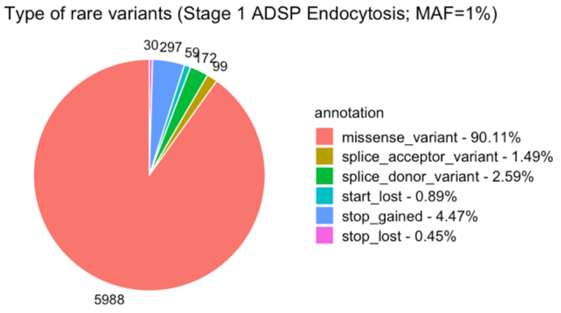


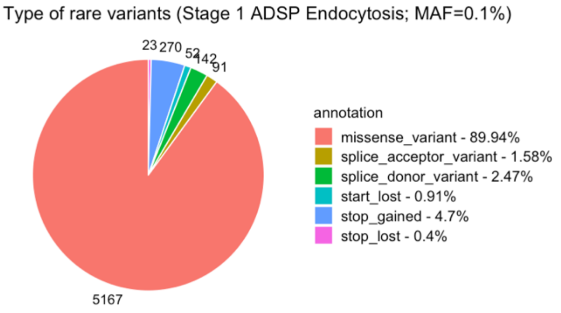
S7 Fig. Overlapping genes between gene-sets (the endocytic, the immune response, and the lipid metabolism pathways) and the findings in recent GWASes.

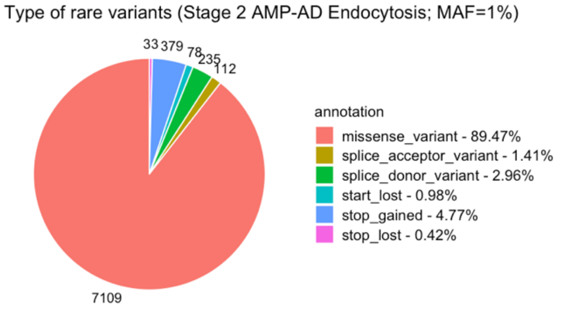
Gene-sets were defined through AmiGO 2 gene-ontology database. Two lists of genes implicated in AD were obtained from the two recent GWASes, Jansen et al.[11] (left) and Kunkle et al.[12] (right), and compared against the three defined gene-sets. The count of overlapping genes between each gene-set and the findings from recent GWASes were shown above. To note, AD-implicated genes were identified through a variety of ways in the GWASes and the overlapping counts in each category were shown.

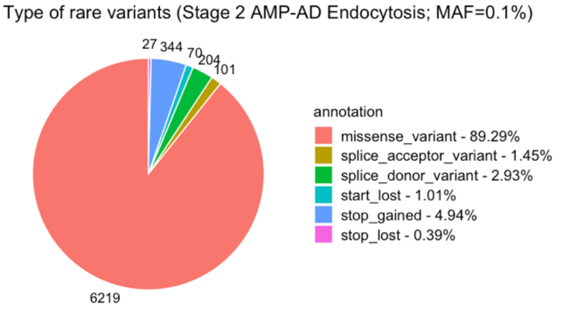


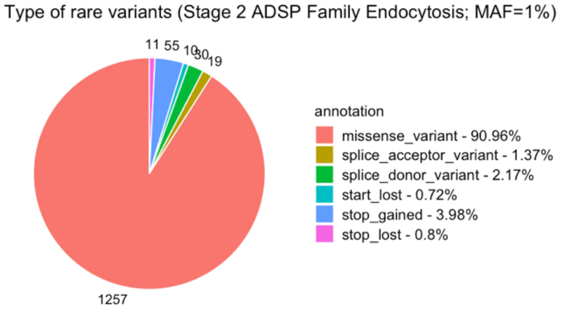
S8 Fig. Distribution of CADD scores among rare deleterious variants defined by VEP and PolyPhen-2.

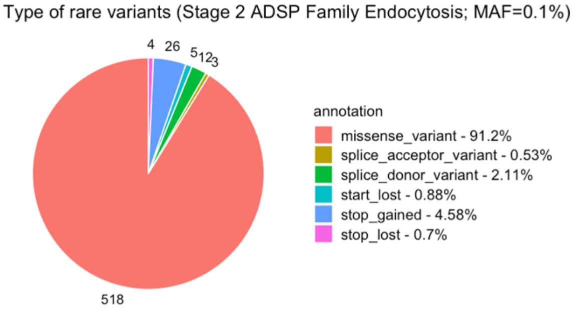

Type of rare variants (Stage 1 ADSP Endocytosis; MAF=1%)
30
297
59
172
99
5988
annotation
🟥 missense_variant - 90.11%
🟨 splice_acceptor_variant - 1.49%
🟩 splice_donor_variant - 2.59%
🟦 start_lost - 0.89%
🟦 stop_gained - 4.47%
🟪 stop_lost - 0.45%


Type of rare variants (Stage 1 ADSP Endocytosis; MAF=0.1%)
23
270
52
142
91
5167
annotation
🟥 missense_variant - 89.94%
🟨 splice_acceptor_variant - 1.58%
🟩 splice_donor_variant - 2.47%
🟦 start_lost - 0.91%
🟦 stop_gained - 4.7%
🟪 stop_lost - 0.4%

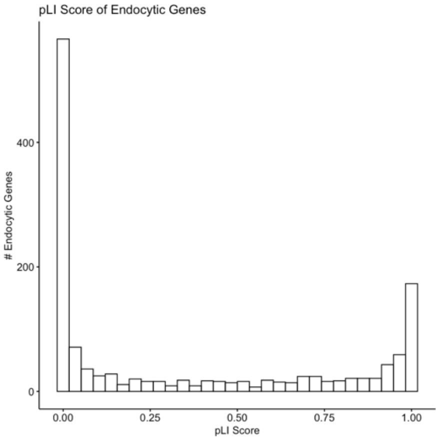

Type of rare variants (Stage 2 AMP-AD Endocytosis; MAF=1%)
33
379
78
235
112
7109
annotation
🟥 missense_variant - 89.47%
🟨 splice_acceptor_variant - 1.41%
🟩 splice_donor_variant - 2.96%
🟦 start_lost - 0.98%
🟦 stop_gained - 4.77%
🟪 stop_lost - 0.42%


Type of rare variants (Stage 2 AMP-AD Endocytosis; MAF=0.1%)
27
344
70
204
101
6219
annotation
🟥 missense_variant - 89.29%
🟨 splice_acceptor_variant - 1.45%
🟩 splice_donor_variant - 2.93%
🟦 start_lost - 1.01%
🟦 stop_gained - 4.94%
🟪 stop_lost - 0.39%

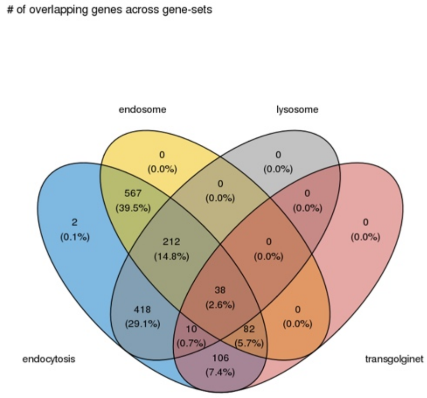

Type of rare variants (Stage 2 ADSP Family Endocytosis; MAF=1%)
11
55
10
30
19
1257
annotation
🟥 missense_variant - 90.96%
🟨 splice_acceptor_variant - 1.37%
🟩 splice_donor_variant - 2.17%
🟦 start_lost - 0.72%
🟦 stop_gained - 3.98%
🟪 stop_lost - 0.8%


Type of rare variants (Stage 2 ADSP Family Endocytosis; MAF=0.1%)
4
26
5
12
3
518
annotation
🟥 missense_variant - 91.2%
🟨 splice_acceptor_variant - 0.53%
🟩 splice_donor_variant - 2.11%
🟦 start_lost - 0.88%
🟦 stop_gained - 4.58%
🟪 stop_lost - 0.7%

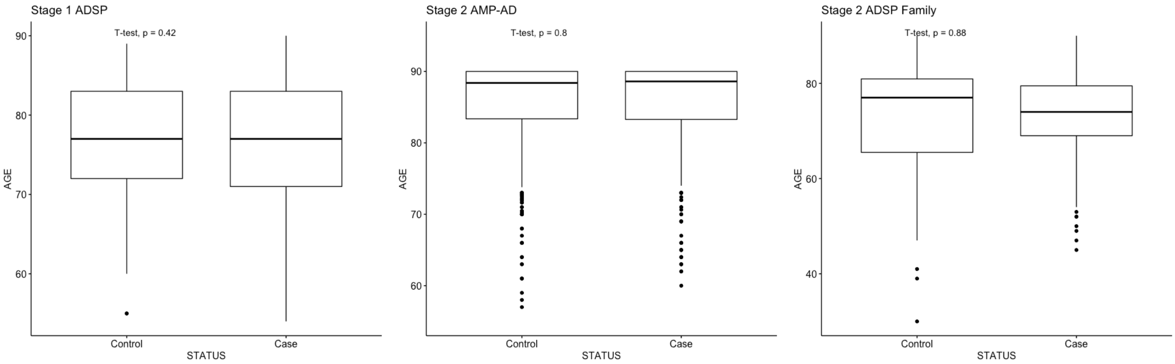
S9 Fig. Distribution of rare deleterious variants in different mutation categories.



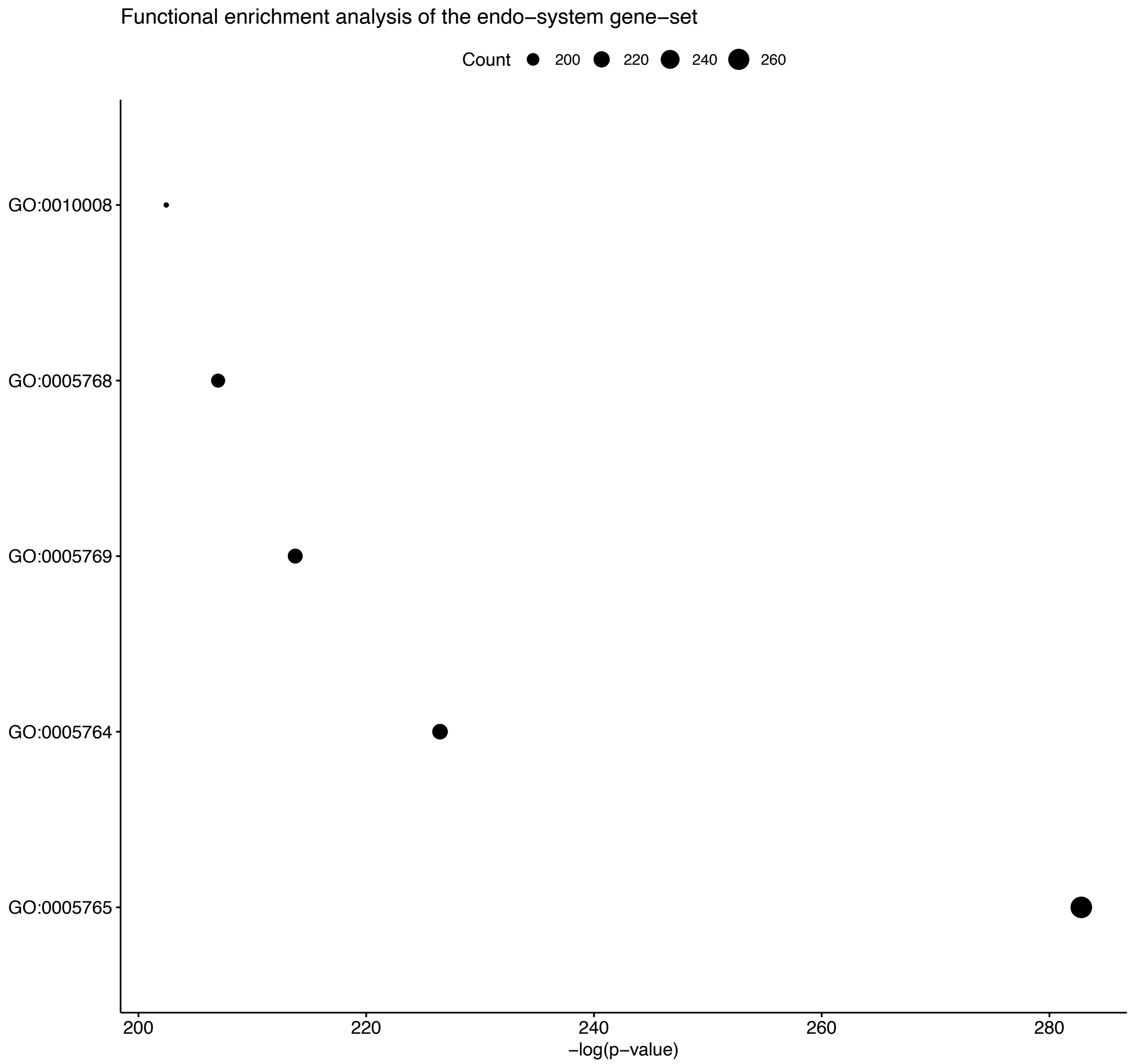
S10 Fig. Distribution of pLI scores among endocytic genes.



S11 Fig. Overlapping genes between the four gene-sets (endo-system, endosome, lysosome, and trans-Golgi network).



S12 Fig. Comparison of age distribution between AD cases and controls in the three datasets (ADSP case-control, AMP-AD case-control, and ADSP Family datasets).



S13 Fig. Functional annotation and confirmation of the biological functions of the endo-system gene-set.

Gene-set	Model	Stage 1 ADSP		Stage 2 AMP-AD		Stage 2 ADSP Family		Meta-analysis	
		OR	P	OR	P	OR	P	P	P*
Endosys	M0	1.24	1.83E-04	1.17	5.75E-03	1.38	1.90E-02	3.24E-07	5.42E-07
	M1	1.23	3.44E-04	1.19	3.83E-03	1.42	1.30E-02	2.34E-07	4.72E-07
	M2	1.20	3.07E-03	1.17	1.30E-02	1.35	4.20E-02	1.58E-05	2.83E-05
	# Variants	5,745		7,946		1,382			
Endosome	M0	1.17	5.56E-03	1.08	1.82E-01	1.41	1.00E-02	1.14E-04	1.35E-04
	M1	1.16	8.33E-03	1.08	1.60E-01	1.48	4.50E-03	7.07E-05	8.61E-05
	M2	1.13	5.07E-02	1.06	3.69E-01	1.48	7.50E-03	1.42E-03	1.26E-03
	# Variants	3,419		4,647		789			

Lysosome	M0	1.10	8.99E-02	1.15	1.05E-02	1.19	2.00E-01	1.39E-03	1.61E-03
	M1	1.08	1.54E-01	1.17	6.33E-03	1.18	2.20E-01	1.88E-03	1.79E-03
	M2	1.08	2.28E-01	1.16	1.32E-02	1.13	3.80E-01	8.42E-03	6.99E-03
	# Variants	2,959		4,076		733			
TransGolgiNet	M0	1.15	1.94E-02	1.09	1.06E-01	1.04	7.80E-01	1.46E-02	9.19E-03
	M1	1.14	2.10E-02	1.10	8.30E-02	1.04	7.70E-01	1.23E-02	7.95E-03
	M2	1.14	3.97E-02	1.09	1.41E-01	0.98	8.80E-01	3.36E-02	2.64E-02
	# Variants	894		1,215		204			

S1 Table. Rare-variant gene-set AD association analysis using PLINK.

The OR and P represented the estimated odds ratio and the p-value from the corresponding logistic regression model (or the generalized linear mixed model for family study). P-values were highlighted in red (if <0.05; nominally significant) or green (if <0.00625; gene-set-wide significant). M0 took into account the sequencing location, first ten PCs, total count of rare variants. M1 was M0 plus age and sex. M2 was M1 plus the count of APOE *ε*2 and *ε*4 alleles. The P and P* in the meta-analysis across two stages (three datasets) represented the p-values calculated using the fixed-effects inverse variance weighted method by METAL and the Fisher's method by 'meta-p,' respectively. # variants represented the number of rare deleterious variants identified in each dataset for each gene-set. The directions of effects were consistent across nearly all models.

Gene-set	Model	Stage 2 AMP-AD		Stage 2 AMP-AD*	
		OR	P	OR	P
Endosys	M0	1.17	5.75E-03	1.26	1.33E-03
	M1	1.19	3.83E-03	1.28	5.90E-04
	M2	1.17	1.30E-02	1.28	1.40E-03
Endosome	M0	1.08	1.82E-01	1.19	1.68E-02
	M1	1.08	1.60E-01	1.20	1.05E-02
	M2	1.06	3.69E-01	1.17	3.55E-02
Lysosome	M0	1.15	1.05E-02	1.24	2.02E-03
	M1	1.17	6.33E-03	1.27	6.84E-04
	M2	1.16	1.32E-02	1.26	1.89E-03
TransGolgiNet	M0	1.09	1.06E-01	1.08	2.40E-01

	M1	1.10	8.30E-02	1.08	2.51E-01
	M2	1.09	1.41E-01	1.08	2.88E-01

S2 Table. Comparison of stage 2 AMP-AD rare-variant gene-set AD association analysis.

The OR and P represented the estimated odds ratio and the p-value from the corresponding logistic regression model (or the generalized linear mixed model for family study). P-values were highlighted in red (if <0.05; nominally significant) or green (if <0.00625; gene-set-wide significant). M0 took into account the sequencing location, first ten PCs, total count of rare variants. M1 was M0 plus age and sex. M2 was M1 plus the count of *APOE* ε2 and ε4 alleles. The stage 2 AMP-AD* cohort represented the largest AMP-AD sub-cohort, ROSMAP.

Gene-set	Model	Stage 1 ADSP				Stage 2 AMP-AD				Stage 2 ADSP Family				Meta-analysis			
		Mu	P-self	Beta	P-comp	Mu	P-self	Beta	P-comp	Mu	P-self	Beta	P-comp	MU	P-self	Beta	P-comp
Endosys	M0	9.06E-02	1.02E-03	6.53E-02	1.30E-02	7.68E-02	4.88E-03	4.08E-02	7.41E-02	1.15E-01	1.35E-03	4.49E-02	9.28E-02	1.55E-01	5.16E-08	7.98E-02	1.90E-03
	M1	9.31E-02	7.64E-04	5.58E-02	2.58E-02	7.05E-02	8.85E-03	3.45E-02	1.12E-01	1.14E-01	1.58E-03	4.62E-02	8.64E-02	1.51E-01	1.12E-07	7.07E-02	4.90E-03
	M2	7.40E-02	5.88E-03	5.82E-02	2.23E-02	6.18E-02	1.87E-02	3.91E-02	8.59E-02	7.55E-02	2.49E-02	4.07E-02	1.25E-01	1.14E-01	4.44E-05	6.56E-02	9.13E-03
Endosome	M0	6.59E-02	3.94E-02	2.92E-02	2.15E-01	1.22E-01	6.11E-04	8.31E-02	9.66E-03	1.71E-01	3.22E-04	1.10E-01	5.78E-03	1.91E-01	1.13E-07	1.09E-01	7.33E-04
	M1	8.09E-02	1.54E-02	3.10E-02	1.95E-01	1.25E-01	4.79E-04	8.44E-02	9.02E-03	1.67E-01	4.47E-04	1.10E-01	5.87E-03	1.97E-01	4.31E-08	1.10E-01	6.31E-04
	M2	7.52E-02	2.25E-02	4.72E-02	9.77E-02	1.02E-01	3.33E-03	8.06E-02	1.25E-02	1.04E-01	1.88E-02	7.96E-02	4.02E-02	1.54E-01	1.55E-05	1.00E-01	1.89E-03
Lysosome	M0	1.30E-01	1.10E-03	1.11E-01	3.75E-03	6.92E-02	5.38E-02	3.87E-02	1.68E-01	8.72E-02	5.34E-02	1.36E-02	3.86E-01	1.70E-01	2.75E-05	9.48E-02	7.84E-03
	M1	1.18E-01	2.73E-03	8.91E-02	1.40E-02	6.87E-02	5.51E-02	4.09E-02	1.55E-01	1.17E-01	1.53E-02	4.38E-02	1.75E-01	1.71E-01	2.36E-05	9.40E-02	7.88E-03
	M2	1.33E-01	9.04E-04	1.27E-01	9.88E-04	7.52E-02	4.03E-02	5.71E-02	8.00E-02	9.52E-02	3.90E-02	5.79E-02	1.18E-01	1.66E-01	3.96E-05	1.19E-01	1.35E-03
TransGolgiNet	M0	1.35E-01	3.20E-02	9.67E-02	8.64E-02	1.49E-01	2.33E-02	1.19E-01	4.07E-02	2.00E-01	1.97E-02	8.68E-02	1.51E-01	2.58E-01	1.94E-04	1.83E-01	3.17E-03
	M1	9.34E-02	1.00E-01	4.44E-02	2.61E-01	1.26E-01	4.67E-02	9.19E-02	9.09E-02	1.30E-01	8.96E-02	2.38E-02	3.88E-01	1.98E-01	3.19E-03	1.18E-01	3.86E-02
	M2	6.35E-02	1.92E-01	2.82E-02	3.43E-01	1.05E-01	8.05E-02	7.84E-02	1.29E-01	6.79E-02	2.42E-01	-9.06E-03	5.41E-01	1.28E-01	3.96E-02	7.46E-02	1.35E-01
Endosys*	M0	9.15E-02	9.25E-04	6.63E-02	1.20E-02	7.71E-02	4.77E-03	4.11E-02	7.29E-02	1.14E-01	1.57E-03	4.27E-02	1.04E-01	1.55E-01	5.13E-08	7.98E-02	1.90E-03
	M1	9.46E-02	6.38E-04	5.75E-02	2.25E-02	7.10E-02	8.50E-03	3.50E-02	1.09E-01	1.12E-01	1.82E-03	4.41E-02	9.67E-02	1.51E-01	1.01E-07	7.13E-02	4.62E-03
	M2	7.36E-02	6.13E-03	5.77E-02	2.32E-02	6.29E-02	1.73E-02	4.02E-02	8.02E-02	7.55E-02	2.50E-02	4.02E-02	1.28E-01	1.14E-01	4.25E-05	6.59E-02	8.87E-03
Endosome*	M0	6.73E-02	3.63E-02	3.07E-02	2.03E-01	1.23E-01	5.89E-04	8.36E-02	9.35E-03	1.68E-01	4.01E-04	1.07E-01	7.30E-03	1.91E-01	1.12E-07	1.10E-01	7.31E-04
	M1	8.35E-02	1.31E-02	3.36E-02	1.76E-01	1.26E-01	4.46E-04	8.52E-02	8.50E-03	1.64E-01	5.50E-04	1.07E-01	7.32E-03	1.98E-01	3.77E-08	1.11E-01	5.75E-04
	M2	7.45E-02	2.36E-02	4.64E-02	1.02E-01	1.04E-01	2.92E-03	8.24E-02	1.10E-02	1.05E-01	1.87E-02	7.89E-02	4.15E-02	1.54E-01	1.45E-05	1.01E-01	1.81E-03

S3 Table. Rare-variant AD association analysis using the MAGMA burden method.

The starred (*) geneset are those excluding the *APOE* gene. The Mu and P-self represented the estimated mean association and the self-contained p-value testing whether an association existed within the tested gene-set. The Beta and P-comp represented the estimated effect size and the competitive p-value testing whether the association within the gene-set was greater than in other genes. P-values were highlighted in red (if <0.05; nominally significant) or green (if <0.00625; gene-set-wide significant). M0 took into account the sequencing location, first ten PCs, total count of rare variants. M1 was M0 plus age and sex. M2 was M1 plus the count of *APOE* ε2 and ε4 alleles. The directions of effects were consistent across nearly all models.

Gene-set	Model	Stage 1 ADSP				Stage 2 AMP-AD				Stage 2 ADSP Family				Meta-analysis			
		Mu	P-self	Beta	P-comp	Mu	P-self	Beta	P-comp	Mu	P-self	Beta	P-comp	Mu	P-self	Beta	P-comp
Endosys	M0	8.51E-02	2.60E-03	6.05E-02	1.96E-02	6.59E-02	1.74E-02	1.07E-02	3.54E-01	1.06E-01	3.38E-03	3.55E-02	1.46E-01	1.40E-01	1.94E-06	5.49E-02	2.41E-02
	M1	8.97E-02	1.63E-03	6.57E-02	1.12E-02	6.99E-02	1.27E-02	1.86E-02	2.55E-01	1.04E-01	3.97E-03	2.50E-02	2.31E-01	1.44E-01	9.28E-07	4.76E-02	4.07E-02
	M2	7.13E-02	9.72E-03	6.00E-02	1.94E-02	4.50E-02	7.49E-02	9.22E-03	3.74E-01	6.63E-02	4.56E-02	2.34E-02	2.53E-01	9.81E-02	5.91E-04	3.75E-02	8.98E-02
Endosome	M0	6.32E-02	5.00E-02	3.47E-02	1.74E-01	8.65E-02	1.37E-02	3.45E-02	1.68E-01	1.43E-01	2.39E-03	8.16E-02	2.97E-02	1.56E-01	2.06E-05	7.25E-02	1.86E-02
	M1	8.19E-02	1.64E-02	4.13E-02	1.27E-01	7.96E-02	2.13E-02	2.90E-02	2.05E-01	1.41E-01	2.84E-03	7.40E-02	4.47E-02	1.62E-01	9.75E-06	5.74E-02	4.60E-02
	M2	7.48E-02	2.58E-02	4.60E-02	1.04E-01	3.57E-02	1.81E-01	-4.16E-03	5.46E-01	8.98E-02	3.84E-02	4.78E-02	1.45E-01	1.08E-01	2.14E-03	4.00E-02	1.26E-01
Lysosome	M0	1.21E-01	2.59E-03	8.94E-02	1.56E-02	8.70E-02	2.52E-02	1.86E-02	3.23E-01	9.62E-02	3.95E-02	3.41E-02	2.32E-01	1.80E-01	1.46E-05	9.38E-02	8.73E-03
	M1	1.09E-01	5.97E-03	1.04E-01	5.24E-03	8.68E-02	2.55E-02	3.97E-02	1.59E-01	1.22E-01	1.29E-02	2.51E-02	2.96E-01	1.82E-01	1.26E-05	9.37E-02	7.80E-03
	M2	1.31E-01	1.25E-03	1.22E-01	1.47E-03	5.88E-02	9.29E-02	3.10E-02	2.22E-01	9.39E-02	4.32E-02	5.40E-02	1.34E-01	1.56E-01	1.51E-04	1.08E-01	3.16E-03
TransGolgiNet	M0	1.52E-01	1.93E-02	1.33E-01	3.01E-02	-1.06E-02	5.55E-01	-3.67E-02	7.00E-01	1.66E-01	4.39E-02	4.14E-02	3.10E-01	1.62E-01	1.35E-02	7.79E-02	1.24E-01
	M1	1.11E-01	6.57E-02	8.66E-02	1.07E-01	-2.79E-04	5.01E-01	-4.52E-02	7.44E-01	9.70E-02	1.60E-01	-2.38E-02	6.12E-01	1.24E-01	4.56E-02	3.27E-02	3.11E-01
	M2	8.45E-02	1.25E-01	6.51E-02	1.77E-01	3.54E-02	3.22E-01	-2.86E-03	5.16E-01	3.84E-02	3.47E-01	-5.16E-02	7.23E-01	8.25E-02	1.30E-01	1.30E-02	4.24E-01
Endosys*	M0	8.59E-02	2.40E-03	6.19E-02	1.74E-02	6.56E-02	1.79E-02	1.07E-02	3.54E-01	1.05E-01	3.88E-03	3.34E-02	1.61E-01	1.39E-01	2.03E-06	5.49E-02	2.41E-02
	M1	9.12E-02	1.39E-03	6.84E-02	8.70E-03	6.98E-02	1.28E-02	1.86E-02	2.54E-01	1.03E-01	4.44E-03	2.31E-02	2.48E-01	1.45E-01	8.90E-07	4.83E-02	3.85E-02
	M2	7.11E-02	9.87E-03	6.02E-02	1.91E-02	4.59E-02	7.12E-02	1.07E-02	3.55E-01	6.63E-02	4.57E-02	2.29E-02	2.58E-01	9.84E-02	5.73E-04	3.78E-02	8.80E-02
Endosome*	M0	6.46E-02	4.65E-02	3.70E-02	1.59E-01	8.59E-02	1.42E-02	3.46E-02	1.68E-01	1.40E-01	2.88E-03	7.83E-02	3.54E-02	1.55E-01	2.17E-05	7.26E-02	1.85E-02
	M1	8.44E-02	1.40E-02	4.56E-02	1.04E-01	7.95E-02	2.15E-02	2.91E-02	2.05E-01	1.38E-01	3.34E-03	7.11E-02	5.16E-02	1.62E-01	9.32E-06	5.86E-02	4.31E-02
	M2	7.43E-02	2.67E-02	4.63E-02	1.03E-01	3.71E-02	1.73E-01	-1.92E-03	5.21E-01	9.02E-02	3.79E-02	4.69E-02	1.50E-01	1.09E-01	2.07E-03	4.05E-02	1.23E-01

S4 Table. Rare-variant AD association analysis using the MAGMA SNP-wise method.

The starred (*) geneset are those excluding the *APOE* gene. The Mu and P-self represented the estimated mean association and the self-contained p-value testing whether an association has existed within the tested gene-set. The Beta and P-comp represented the estimated effect size and the competitive p-value testing whether the association within the gene-set was greater than in other genes. P-values were highlighted in red (if <0.05; nominally significant) or green (if <0.00625; gene-set-wide significant). M0 took into account the sequencing location, first ten PCs, total count of rare variants. M1 was M0 plus age and sex. M2 was M1 plus the count of *APOE* ε2 and ε4 alleles. The directions of effects were consistent across nearly all models.

Gene-set	Model	Stage 1 ADSP		Stage 2 AMP-AD		Stage 2 AMP-AD*		Stage 2 ADSP Family		Meta-analysis	
		OR	P	OR	P	OR	P	OR	P	P	P*
Endosys	M0	1.23	2.84E-04	1.14	1.84E-02	1.21	7.86E-03	1.38	1.40E-02	4.45E-07	8.12E-07
	M1	1.22	5.37E-04	1.16	1.26E-02	1.22	4.67E-03	1.41	8.90E-03	2.72E-07	5.99E-07
	M2	1.19	6.60E-03	1.16	1.41E-02	1.24	3.93E-03	1.35	1.90E-02	4.48E-06	9.56E-06
Endosome	M0	1.18	4.38E-03	1.07	2.13E-01	1.16	4.13E-02	1.46	2.90E-03	5.56E-06	1.01E-05
	M1	1.17	6.96E-03	1.08	1.79E-01	1.17	3.09E-02	1.54	8.70E-04	2.29E-06	4.05E-06
	M2	1.12	7.01E-02	1.08	2.16E-01	1.16	4.71E-02	1.55	1.10E-03	4.58E-05	5.56E-05
Lysosome	M0	1.09	1.33E-01	1.12	3.46E-02	1.19	1.09E-02	1.21	1.40E-01	1.43E-03	1.71E-03
	M1	1.07	2.17E-01	1.14	2.17E-02	1.22	4.19E-03	1.21	1.40E-01	1.29E-03	1.16E-03
	M2	1.06	3.42E-01	1.14	2.85E-02	1.22	6.57E-03	1.16	2.20E-01	4.72E-03	3.56E-03
TransGolgiNet	M0	1.16	1.00E-02	1.09	1.01E-01	1.08	2.27E-01	1.08	5.20E-01	1.06E-02	7.22E-03
	M1	1.16	1.17E-02	1.10	6.94E-02	1.08	2.18E-01	1.08	5.50E-01	1.21E-02	8.28E-03
	M2	1.15	2.69E-02	1.10	1.04E-01	1.09	2.17E-01	0.99	9.30E-01	4.12E-02	2.63E-02

S5 Table. Rare-variant AD association analysis using PLINK where rare variants were annotated by a combination of VEP, PolyPhen-2, and CADD (>15).

The OR and P represented the estimated odds ratio and the p-value from the corresponding logistic regression model (or the generalized linear mixed model for family study). P-values were highlighted in red (if <0.05; nominally significant) or green (if <0.00625; gene-set-wide significant). M0 took into account the sequencing location, first ten PCs, total count of rare variants. M1 was M0 plus age and sex. M2 was M1 plus the count of *APOE* ε2 and ε4 alleles. The stage 2 AMP-AD cohort was analyzed using all sub-cohorts and the largest ROSMAP sub-cohort (71.5% of the total sample size; marked in *). The P and P* in the meta-analysis across two stages (three datasets; AMP-AD* was used here) represented the p-values calculated using the fixed-effects inverse variance weighted method by METAL and the Fisher's method by 'meta-p,' respectively. Similar results could be obtained using the stage 2 AMP-AD. The directions of effects were consistent across nearly all models.

Gene-set	Model	Stage 1 ADSP		Stage 2 AMP-AD		Meta-analysis	
		OR	P	OR	P	P	P*
Endosys	M0	1.10	1.77E-01	1.03	6.21E-01	1.92E-01	1.26E-01
	M1	1.10	1.68E-01	1.03	5.48E-01	1.62E-01	1.10E-01
	M2	1.09	2.35E-01	1.04	4.46E-01	1.68E-01	1.22E-01
Endosome	M0	1.15	5.23E-02	1.09	7.61E-02	8.62E-03	7.87E-03
	M1	1.15	5.10E-02	1.10	7.10E-02	7.90E-03	7.26E-03
	M2	1.09	2.31E-01	1.10	5.83E-02	2.89E-02	2.26E-02
Lysosome	M0	1.00	9.68E-01	1.03	5.97E-01	6.88E-01	4.42E-01
	M1	1.01	9.42E-01	1.03	5.74E-01	6.53E-01	4.06E-01
	M2	0.93	3.51E-01	1.04	4.67E-01	2.40E-01	5.09E-01
TransGolgiNet	M0	1.16	4.37E-02	1.07	1.80E-01	1.76E-02	1.42E-02
	M1	1.16	3.92E-02	1.07	1.68E-01	1.50E-02	1.22E-02
	M2	1.13	9.34E-02	1.08	1.32E-01	2.43E-02	2.09E-02

S6 Table. Rare-variant gene-set Braak association analysis using PLINK.

The OR and P represented the estimated odds ratio and the p-value from the corresponding logistic regression model (or the generalized linear mixed model for family study). P-values were highlighted in red (if <0.05; nominally significant). M0 took into account the sequencing location, first ten PCs, total count of rare variants. M1 was M0 plus age and sex. M2 was M1 plus the count of *APOE ε*2 and *ε*4 alleles. The P and P* in the meta-analysis across two stages (two datasets) represented the p-values calculated using the fixed-effects inverse variance weighted method by METAL and the Fisher's method by 'meta-p,' respectively. The directions of effects were consistent across nearly all models.

Phenotype		AAO						AOD	
Gene-set	Model	Stage 1 ADSP		Stage 2 ADSP Family		Meta-analysis		Stage 2 AMP-AD	
		OR	P	OR	P	P	P*	OR	P

Endosys	M0	1.14	8.27E-04	1.31	9.06E-04	2.47E-06	3.09E-06	1.11	1.68E-02
	M1	1.13	1.92E-03	1.28	3.11E-03	1.83E-05	2.15E-05	1.10	2.40E-02
	M2	1.10	1.99E-02	1.19	3.50E-02	1.70E-03	1.68E-03	1.06	1.81E-01
Endosome	M0	1.07	8.00E-02	1.35	3.83E-05	3.33E-05	1.16E-05	1.04	3.79E-01
	M1	1.06	1.51E-01	1.33	1.44E-04	2.12E-04	7.10E-05	1.04	3.50E-01
	M2	1.03	4.56E-01	1.27	1.34E-03	5.19E-03	1.49E-03	1.01	8.07E-01
Lysosome	M0	1.14	9.71E-04	1.02	7.65E-01	1.10E-02	1.78E-03	1.09	3.60E-02
	M1	1.13	1.97E-03	1.01	9.22E-01	2.40E-02	3.94E-03	1.08	5.75E-02
	M2	1.15	4.68E-04	0.95	5.45E-01	3.71E-03	1.65E-03	1.05	2.00E-01
TransGolgiNet	M0	1.10	1.10E-02	1.06	4.70E-01	2.10E-02	9.91E-03	1.06	1.43E-01
	M1	1.09	1.82E-02	1.05	5.73E-01	3.86E-02	1.81E-02	1.05	2.01E-01
	M2	1.02	5.62E-01	1.02	7.95E-01	5.53E-01	3.57E-01	1.04	2.90E-01

S7 Table. Rare-variant gene-set AAO and AOD association analysis using PLINK.

The OR and P represented the estimated odds ratio and the p-value from the corresponding logistic regression model (or the generalized linear mixed model for family study). P-values were highlighted in red (if <0.05; nominally significant) or green (if <0.00625; gene-set-wide significant). M0 took into account the sequencing location, first ten PCs, total count of rare variants. M1 was M0 plus age and sex. M2 was M1 plus the count of *APOE* $\varepsilon2$ and $\varepsilon4$ alleles. The P and P* in the meta-analysis across two stages (two datasets) represented the p-values calculated using the fixed-effects inverse variance weighted method by METAL and the Fisher's method by 'meta-p,' respectively. The directions of effects were consistent across nearly all models.

Stage 1 ADSP		Stage 2 AMP-AD		Meta-analysis	
Gene	P	Gene	P	Gene	P
CD300LG	1.41E-06	HAVCR2	4.42E-04	ANKRD13D	3.56E-05
ANKRD13D	1.46E-06	LNPEP	1.12E-03	IL1B	5.34E-05
TJAP1	5.80E-06	ANKRD13A	1.56E-03	DYNC1H1	2.17E-03
PLBD1	9.16E-06	PCSK7	1.86E-03	HPS4	3.68E-03
IL1B	1.73E-05	CPNE1	2.28E-03	TLR3	4.15E-03
LLGL1	1.79E-05	TLR9	2.40E-03	KREMEN2	5.00E-03
HPS4	9.27E-05	AP2A2	2.58E-03	HAVCR2	5.30E-03
STAMBP	1.19E-04	OMD	2.96E-03	ARFGEF2	5.78E-03
LRRK2	3.58E-04	RAB17	3.87E-03	PLEKHA8	5.85E-03

DBNL	1.12E-03	ATP6V0A2	4.20E-03	EZR	6.76E-03

S8 Table. Top ten most significant genes in rare-variant single-gene NFT association analysis.

The genes were sorted in the descending order of p-values. The meta-analysis was performed using MetaSKAT. P-values below the Bonferroni threshold (α=4.83*10^{-5}; 4.25*10^{-5}; 5.17*10^{-5}, for ADSP, AMP-AD, and meta-analysis, respectively) were highlighted in red.

Gene name	Cell type	Effect	P-value
ANKRD13D	Ast	2.57	1.07E-02
	Ex	8.79	1.92E-18
	In	2.19	2.84E-02
	Mic	-1.75	8.26E-02
	Oli	-3.14	1.78E-03
	Opc	0.82	4.14E-01
HLA-A	Ast	-1.09	2.76E-01
	Ex	-0.03	9.79E-01
	In	-4.45	9.72E-06
	Mic	-2.98	3.07E-03
	Oli	-1.28	2.01E-01
	Opc	-1.46	1.45E-01
SLC26A7	Ast	-1.76	9.12E-02
	Ex	2.88	4.85E-03
	In	0.50	6.28E-01
	Mic	1.22	2.48E-01
	Oli	-0.40	6.90E-01
	Opc	0.12	9.16E-01

S9 Table. Differential expression analysis of three identified genes, *HLA-A*, *SLC26A*, and *ANKRD13D*, between AD cases and controls from the ROSMAP study using six major cell types.

Abbreviations: Ex: excitatory neuron; In: inhibitory neuron; Ast: astrocyte; Oli: oligodendrocyte; Opc: oligodendrocyte-precursor-cell; Mic: microglia. Effect: t-statistics calculated using student t-test, representing the direction of effect. P-values are computed using the same method.

Gene name	Cell type	Effect	P value
ANKRD13D	Ex4	5.49	5.76E-08
	In0	2.11	3.59E-02
	Ast1	1.37	1.78E-01
	Oli0	-1.59	1.12E-01
HLA-A	Ex4	0.53	5.99E-01
	In0	-1.59	1.14E-01
	Ast1	0.86	3.91E-01
	Oli0	-1.36	1.76E-01
SLC26A7	Ex4	2.11	8.99E-02
	In0	0.93	4.37E-01
	Ast1	-2.24	2.06E-01
	Oli0	-0.26	7.97E-01

S10 Table. Differential expression analysis of three identified genes, *HLA-A*, *SLC26A*, and *ANKRD13D*, between AD cases and control from the ROSMAP study using four cellular subpopulations implicated with AD pathology

Abbreviations: Ex: excitatory neuron; In: inhibitory neuron; Ast: astrocyte; Oli: oligodendrocyte; Opc: oligodendrocyte-precursor-cell; Mic: microglia. Effect: t-statistics calculated using student t-test, representing the direction of effect. P-values are computed using the same method.

	Stage 1 ADSP		Stage 2 AMP-AD		Stage 2 ADSP Family	
Endocytic gene-set	MAF 1%	MAF 0.1%	MAF 1%	MAF 0.1%	MAF 1%	MAF 0.1%
# of Singletons	5,113	5,011	5,878	5,803	618	305
# of Total variants	6,645	5,745	7,946	6,965	1,382	568
# of Private Doubletons	0	0	2	2	0	0
Percentage of Singletons	76.9%	87.2%	74.0%	83.3%	44.7%	53.7%

S11 Table. Count of singletons and private doubletons within the included rare deleterious variants.

The number of total variants represented all rare deleterious variants included under the corresponding MAF threshold.

Gene-set	Model	Stage 1 ADSP		Stage 2 AMP-AD		Stage 2 ADPS Family		Meta-analysis	
		OR	OR	OR	P	OR	P	P	P*
Endosys	M0	1.17	5.82E-03	1.11	1.50E-01	1.12	3.80E-01	3.38E-03	2.57E-03
	M1	1.17	6.59E-03	1.11	1.31E-01	1.15	2.60E-01	1.99E-03	1.86E-03
	M2	1.16	2.47E-02	1.12	1.16E-01	1.13	3.60E-01	6.26E-03	6.44E-03
Endosome	M0	1.12	5.18E-02	1.08	2.69E-01	1.02	9.00E-01	6.67E-02	4.44E-02
	M1	1.11	6.41E-02	1.09	2.27E-01	1.04	7.20E-01	4.84E-02	3.88E-02
	M2	1.08	2.03E-01	1.09	2.46E-01	1.05	6.70E-01	9.86E-02	8.97E-02
Lysosome	M0	1.07	2.80E-01	1.16	3.07E-02	1.19	1.70E-01	7.73E-03	8.52E-03
	M1	1.06	3.55E-01	1.18	1.87E-02	1.22	1.20E-01	5.29E-03	5.25E-03
	M2	1.04	5.92E-01	1.19	1.95E-02	1.19	1.90E-01	1.57E-02	1.17E-02
TransGolgiNet	M0	1.04	4.78E-01	0.99	8.93E-01	1.08	5.40E-01	4.00E-01	3.53E-01
	M1	1.05	4.48E-01	0.99	8.63E-01	1.08	5.20E-01	3.63E-01	3.38E-01
	M2	1.09	1.77E-01	0.98	8.05E-01	1.07	6.00E-01	2.20E-01	2.18E-01

S12 Table. Rare-variant AD association analysis weighted by pLI scores.

The OR and P represented the estimated odds ratio and the p-value from the corresponding logistic regression model (or the generalized linear mixed model for family study). P-values were highlighted in red (if <0.05; nominally significant) or green (if <0.00625; gene-set-wide significant). M0 took into account the sequencing location, first ten PCs, total count of rare variants. M1 was M0 plus age and sex. M2 was M1 plus the count of *APOE ε*2 and *ε*4 alleles. The P and P* in the meta-analysis across two stages (three datasets) represented the p-values calculated using the fixed-effects inverse variance weighted method by METAL and the Fisher's method by 'meta-p,' respectively.

Gene-set	Model	Stage 1 ADSP		Stage 2 AMP-AD		Stage 2 ADSP Family		Meta-analysis	
		OR	P	OR	P	OR	P	P	P*
BMI	M0	1.06	3.63E-01	0.99	8.49E-01	1.14	2.90E-01	2.13E-01	2.12E-01
	M1	1.06	3.20E-01	0.99	8.20E-01	1.12	3.70E-01	2.21E-01	2.31E-01
	M2	1.02	7.21E-01	0.98	7.04E-01	1.09	5.20E-01	4.26E-01	4.69E-01
Height	M0	0.93	2.20E-01	1.00	9.98E-01	1.01	9.10E-01	4.38E-01	7.85E-01
	M1	0.93	2.90E-01	1.01	9.11E-01	1.01	9.40E-01	4.72E-01	7.58E-01
	M2	0.96	6.08E-01	0.96	4.72E-01	1.03	8.40E-01	4.07E-01	8.09E-01

S13 Table. Rare-variant AD association analysis using gene-sets related to BMI and height. The OR and P represented the estimated odds ratio and the p-value from the corresponding logistic regression model (or the generalized linear mixed model for family study). M0 took into account the sequencing location, first ten PCs, total count of rare variants. M1 was M0 plus age and sex. M2 was M1 plus the count of *APOE ε*2 and *ε*4 alleles. The P and P* in the meta-analysis across two stages (three datasets) represented the p-values calculated using the fixed-effects inverse variance weighted method by METAL and the Fisher's method by 'meta-p,' respectively.

	Stage 1 ADSP		Stage 2 AMP-AD		Stage 2 ADSP Family	
HWE cutoff	MAF 1%	MAF 0.1%	MAF 1%	MAF 0.1%	MAF 1%	MAF 0.1%
Cutoff at 0.001	24,775,258	17,718,944	31,871,709	23,352,094	6,308,504	48,660
Cutoff at 5e-8	24,779,990	17,719,782	31,896,945	23,366,281	6,308,811	48,661
Percentage gained	0.019%	0.0047%	0.079%	0.061%	0.0049%	0.0021%

S14 Table. Number of rare variants passing different HWE cutoffs at different MAF thresholds.

Abbreviations: HWE: Hardy-Weinberg Equilibrium; MAF: minor allele frequency.

Stage 1 ADSP		Stage 2 AMP-AD		Stage 2 ADSP Family		Meta-analysis	
Gene	P	Gene	P	Gene	P	Gene	P
SLC17A3	2.38E-03	ZFYVE16	1.18E-03	SPNS1	3.36E-03	SORL1	2.02E-03
DDIT3	4.20E-03	MFSD8	2.00E-03	MPO	5.12E-03	ATP6V0A2	3.27E-03
SERPINB13	6.15E-03	PDIA3	2.82E-03	PRSS16	5.77E-03	PRSS16	5.47E-03
DAGLB	6.60E-03	CLCN7	3.82E-03	FLT1	6.15E-03	ATP6V1A	7.39E-03
SDC4	7.88E-03	EGFR	4.65E-03	TYRP1	1.33E-02	TMEM108	9.41E-03
VASN	8.39E-03	ABCB9	4.77E-03	ZFYVE9	1.39E-02	GPR137B	9.69E-03
VAC14	9.69E-03	ABCB9	4.77E-03	RAB27B	1.52E-02	CHST4	1.06E-02
HRNR	1.08E-02	CHMP2B	5.66E-03	AP5M1	1.78E-02	SNX17	1.33E-02
UNC93B1	1.18E-02	ATP6V1C2	7.51E-03	CDIP1	1.89E-02	AGRN	1.67E-02
TMEM127	1.29E-02	ATP8A2	7.83E-03	WDR48	2.26E-02	ARRDC3	1.90E-02

S15 Table. Top ten most significant genes in rare-variant single-gene AD association analysis.

The genes were sorted in the descending order of p-values. The meta-analysis was performed using MetaSKAT. The Bonferroni thresholds were α=4.18*10^{-5}; 4.07*10^{-5}; 7.32*10^{-5}; 7.79*10^{-5}, for ADSP, AMP-AD, ADSP Family, and meta-analysis, respectively.

Chapter 3 - Analysis of *de novo* sequence variants in Tourette syndrome

Introduction

Tourette Syndrome (TS) is an early onset neurodevelopmental disorder with an estimated average prevalence rate of 0.6%, ranging between 0.3 – 1%.(1-6) TS is characterized by chronic motor and vocal tics and is highly comorbid with other psychiatric disorders, including obsessive-compulsive disorder (OCD), attention deficit and hyperactivity disorder (ADHD), autism spectrum disorder (ASD).(5, 7-10) Studies have also reported sex differences for TS where males are more likely to be affected and bear comorbidities than females.(11, 12) However, given the current limited understanding of the pathophysiology, interventions, and treatments for tics and TS have demonstrated limited efficacy with long-term side effects.(13)

TS has been shown to have a substantial genetic component with a heritability of 70 – 80%(14), in which single nucleotide variants (SNVs) contributed 50 – 60% of the total estimated heritability.(15) Multiple genetic risk loci have been identified in TS, including loci found in genome-wide association studies (GWASes),(16) rare copy number variants,(17) as well as de novo mutations.(18) In fact, the recent findings by our group have revealed four potential risk genes associated with TS through de novo damaging SNVs and insertion/deletion variants (INDELs) and provided a powerful tool to discover TS-associated genes using recurrent de novo variants.

In this study, we conducted whole-exome sequencing of 2,720 samples and performed stringent quality control steps to ensure high quality in both sample and variant levels. We obtained a total of 858 complete high-quality trios and identified 987 high-quality de novo mutations (DNMs). We then performed functional annotations to the identified DNMs and observed an enrichment of protein-truncating variants (PTV) and missense mutations. We then compared the observed mutation rate per gene in TS samples with those in an expected baseline as described in Samocha et al.(19). Lastly, we conducted functional analyses across different tissue types and developmental stages to identify the potential biological connections.

Methods

Whole exome sequencing and variant calling

We performed whole-exome capture and sequenced the DNA of 2,720 samples. Fragmented DNA samples were supplied in equal amounts for hybridization, followed by PCR amplification. Then, all samples were normalized into 10 nM concentrations to 10 nM and sent for sequencing using Illumina HiSeq 4000 platform with 100 bp paired-end sequencing reads. GATK best practices(20) have been used for pre-processing and variant calling. Specifically, we ran the HaplotypeCaller in GVCF mode for each sample and conducted joint genotype calling to include only the common regions covered by all samples. This procedure resulted in a multi-sample VCF dataset for later quality control steps and analyses.

Quality control of WES data

Variant-level quality control of WES data

We conducted stringent quality control (QC) to ensure that only high-quality variants were included for analyses in this study. Before any QC was conducted, we first obtained WES target information and removed variants that fell into the non-targeted regions by design. This step ensured that all the remaining variants were captured by the design of the WES kit, which usually had higher quality and confidence than the flanking regions captured by chance. We additionally performed basic variant filtering to select variants meeting the following three basic criteria: 1. marked with PASS flags by the Variant Quality Score Recalibration (VQSR) in the GATK pipeline; 2. with genotyping quality (GQ) >= 21, and 3. bi-allelic variants. Then, we evaluated the sequencing metrics using the remaining variants from the first two filtering steps. We used the Single Nucleotide Polymorphism database (dbSNP) version 146 as the reference and measured the total numbers of SNVs/INDELs found, the amount of SNVs/INDELs found in dbSNP 146, and the novel SNVs/INDELs, followed by computing the Ti/Tv ratio within each category. We also counted the number of singletons within SNVs and INDELs. To note, the high-quality Ti/Tv ratio for WES was expected to be around 3.0(21) and thus served as the first check of the sequencing quality across the exome in our QC procedures. Subsequently, after confirming all sequencing metrics were within expectation, we then assessed the genotyping missing rate and set the threshold at 5%.

Individual-level quality control of WES data

We conducted stringent quality control (QC) to ensure that only high quality-samples were selected for analyses in this study. We included only variants that served as inputs for the variant-level QC, namely those that passed the first two sets of basic filterings. For each individual, we checked the sample-level missing rate, relatedness, population composition, expected sex, and concordance rate to microarray data. Specifically, we first computed the genotyping missing rate per individual for SNVs and INDELs separately and set the passing threshold at 5% and 10%, respectively. Secondly, we performed identity by descent (IBD) analysis by estimating the theoretical kinship relationship between all input samples using Plink 1.9(22), followed by comparing against known pedigree structures. We primarily focused on identifying duplications and unexpectedly related samples within each trio family. Thirdly, we performed principal component analysis (PCA) to examine population stratification within our samples. As PCA assumes independence across samples, we included only parents within our datasets while removing duplicated samples identified in the previous step. To identify specific ancestry groups, we used 1000 Genomes (1KG) phase 3(22) as the reference panel and performed the analysis using EIGENSTRAT(22) with only common SNVs shared across the 1KG reference panel and our dataset. To note, we only included the parents (samples labeled as parents or without parental information) into this analysis as PCA assumed independence among samples. The kinship relationship would be confounded with the ancestral relationship. The computed principal components (PCs) for 1KG then served as a guide to determine the ancestries among the TS dataset. PCA plots (PC1 v.s. PC2) were included in the supporting figures. (Fig S1.) Subsequently, we conducted sex-check using the –check-sex option from PLink 1.9 by estimating the theoretical sex from X-chromosomes and comparing it against the recorded empirical sex. In the final step, we obtained the microarray data for 3,215 individuals, of which 2,720 were within our WES dataset. We performed liftover for the microarray data from Human reference genome version 19 (Hg19) to Human reference genome version 38 (Hg38) to ensure consistency with the WES data. We then identified the shared variants and set the concordance rate cutoff at 98%.

De novo mutation calling and distribution checking

We filtered out low-quality variants and samples based on the aforementioned QC procedures. The remaining data then served as the input for calling de novo mutations (DNMs). We selected complete trios as both parent and child information was required for accurately detecting de novo variants. Additionally, for families with more than one probands, we split them into multiple families by duplicating the parents and suffixing their sample IDs with numbers. As a result, each family contained exactly one proband with two parents. As DNMs are rare events, we assumed all candidates were singletons with minor allele count (MAC) at 1. We used a Bayesian framework for DNM calling in trios (TrioDeNovo), developed by Wei et al..(23) To ensure a high quality of DNM calling, we set the minDQ and minDP parameters at 8 and 20, respectively, followed by additional filterings based on allele balance (AB). Specifically, we set the cutoff of homozygous AB at 0.99 and that of heterozygous AB between 0.3 and 0.7.

To examine the distribution of called DNMs among our samples, we first looked at the mean, median, and standard deviation of the number of DNMs per individual. According to previous studies, the distribution of DNMs per individual should follow a Poisson distribution.(18, 24) We used the poisson.test function in R(25) to test this hypothesis. Furthermore, we investigated the distribution of DNMs per gene as well as per chromosome.

Functional annotation of DNMs

To investigate the potential impact of the identified DNMs, we performed functional annotations using the ENSEMBL Variant Effect Predictor (VEP)(26) in combination with the Polymorphism Phenotyping v2 (PolyPhen-2)(27). To identify potential deleterious and pathogenic variants, we focused on the variants flagged with HIGH or MODERATE impact by VEP. In particular, the HIGH impact variants included transcript deletion or amplification, stop or start lost, stop gain, splice site, and frameshift mutations, while the MODERATE impact variants consisted of inframe insertion or deletion, missense, inframe protein-altering, and regulatory region deletion mutations. Additionally, we looked at the predicted score by PolyPhen-2 and set a cutoff at 0.446, which included the predicted possibly and probably damaging variants.

DNM enrichment analysis

Our hypothesis of the role of DNMs in TS samples was that DNMs were enriched in TS probands. Therefore, we tested the observed rate of mutations within our dataset and compared them against the estimated mutation rate of various types of mutations from external sources.(19) In particular, we used the high-quality DNMs called in the previous step and additionally removed individuals with over 10 DNMs, as they were considered outliers due to an excessive amount of DNMs. We summarized the DNMs into three categories, protein-truncating (PTV), missense, and synonymous variants. The PTV variants consisted of frameshift, stop gained, splice acceptor, and splice donor mutations. Then, we counted the number of DNMs from each category per gene and compared to the expectation under the null hypothesis using a Poisson distribution, previously estimated by Samocha et al.(19). To account for testing multiple genes, we considered two multiple-testing thresholds using Bonferroni correction, one corrected for the number of genes with at least one DNMs in a given category and one corrected for 18,226 genes with available, expected mutation rate. The resulting significance thresholds based on the first method were thus at α=1.28*10^{-3}; 8.20*10^{-5}; 2.22*10^{-4} for PTV, missense, and synonymous variants, respectively. The second method gave the same threshold of α=2.74*10^{-6} for tests.

Furthermore, we conducted a gene-set analysis using the group of 42 genes containing de novo loss of function (LoF) intolerant (pLI > 0.9) PTVs or de novo missense intolerant (Z-mis > 4) likely gene damaging (LGD) missense variants. The LoF and missense intolerances were estimated from gnomAD database.(28) We performed gene-set expression enrichment analysis using FUMA(29) across different human tissues available from GTEx version 9 gene-expression dataset(30). We also tested for differential expression of this gene-set spanning 11 human brain development stages provided by the BrainSpan project(31).

Results

Sample and variant QC outcome

We have successfully performed stringent QC on 2,720 samples, which resulted in 894 high-quality complete trio families. As described in Methods, we removed variants in non-target regions by the design of WES kit, which resulted in 638,345 variants. Subsequently, we applied basic filters and kept 580,720 SNVs and 22,036 INDELs. We then computed the sequencing metrics using these variants as inputs. Comparing to dbSNP version 146, we have found 16.61% and 27.49% novel SNVs and INDELs, respectively. The Ti/Tv ratio for the SNVs present in dbSNP 146 database was 3.16, which was close to the expected ratio of 3 for WES data(21) and indicated a good overall sequencing quality in this dataset. The novel SNVs, on the other hand, had a lower Ti/Tv ratio at 1.7515. (Fig S2.)

After checking the sequencing metrics and confirming an overall high quality in the database, we performed the QC steps at the variant and the sample level separately. In general, both SNVs and INDELs had low missing rates, where only 5.88% of the common variants and 2.91% of the rare variants fell above the missing rate threshold of 1%. Similar but slightly higher missing rates were observed in INDELs.

For sample-level QC, we first examined sample-level missing rates using SNVs and INDELs, separately. The maximum sample missing rates using SNVs and INDELs were at 3.1% and 6.45% and the mean at 0.95% and 2.46%, respectively. Secondly, to examine the kinship relatedness, we performed IBD analysis and identified 37 pairs of duplicates, 2,083 pairs of first-degree relatives, and ten pairs of second-degree relatives. All duplicates and second-degree relatedness were unexpected, which were thus examined first. For duplicated samples, we removed one in each pair. For second-degree relatives, we identified one trio family in which the parents were second-degree relatives. All other second-degree related samples were between different families. We thus did not remove any sample based on these criteria. Finally, we checked the first-degree related pairs. By study design, all parent-child pairs in trio families were expected and were thus removed from the identified first-degree related pairs. This resulted in 16 problematic first-degree related pairs where one in each pair was therefore removed from further analysis. Thirdly, to understand the population composition, we performed PCA using 1KG as the reference panel. We focused on identifying samples with European ancestry and used only parents to avoid confounding effects due to relatedness, which resulted in 1,595 theoretically determined European samples out of 1,694 parents. The fourth QC procedure was to check how well the empirical sex information was consistent with the theoretical sex estimation. We found 19 sex-mismatched samples, seven with ambiguous estimation and 45 with no empirical records. The confirmed 19 sex-mismatched samples were therefore removed from DNM calling. Lastly, we conducted genotyping concordance analysis using microarray data available to 3,215 individuals in total. In combination, 2,714 samples had both WES and microarray data with 37,720 joint bi-allelic SNVs. After examining these variants, we computed a mean concordance rate at 99.89% with only three individuals below 98%. Thus, we set the cutoff at 98% and removed all failed samples.

De novo mutation calling and functional annotation

After removing variants and samples that failed our stringent QC procedures, we additionally removed incomplete families and reformed other non-trio families. (See Methods) We kept only singletons as we assumed all de novo mutation events were rare in our dataset. This resulted in 141,555 singletons from 873 complete trio families. We used TrioDeNovo software to call DNMs. We tested for a range of possible minimum De Novo Quality parameters (minDQ = 5-15). (Table 1.) We observed a rapid decrease in the number of called DNMs when minDQ was over 11. Between minDQ = 6 and 9, there was nearly no change in the number of called DNMs, suggesting an optimal point to set this parameter. We, therefore, set minDQ at eight and minimum depth (minDP) at 20, where the latter was based on the recommended value by the TrioDeNovo documentation. This calling setup resulted in 3,519 DNMs. To further ensure the calling quality, we applied an additional check on allele balance for homozygous and heterozygous sites (AB Hom and AB Het), which gave us 2,929 high-quality DNMs identified in 588 families. A distribution of DNMs across probands, genes, and chromosomes could be found in Fig S3-S5. To note, we observed 15 samples with over ten DNMs and an additional four samples with over five DNMs. Excluding these outliers, the distribution of the DNMs per sample indeed followed a Poisson distribution with a mean of 1.12 and a standard deviation (SD) of 1.05. (Fig 1.) Out of 2,494 genes with at least one DNM, only 2.37% of genes contained three or more DNMs. To note, one gene contained a maximum of 12 DNMs.

MinDQ	**Mean**	**SD**	**Mean (<=10)**	**SD (<=10)**	**#Outlier (>10)**	**#Outlier2 (>5)**	**#Normal (>0)**
5	3.36	23.64	1.15	1.13	15	4	573
6	3.36	23.64	1.14	1.13	15	4	570
7	3.36	23.64	1.14	1.13	15	4	569
8	3.36	23.64	1.14	1.13	15	4	569
9	3.34	23.63	1.13	1.12	15	4	567
10	3.32	23.55	1.12	1.12	15	4	563
11	3.27	23.37	1.10	1.11	15	4	559
12	2.62	18.82	0.89	1.03	15	3	487
13	0.47	3.96	0.16	0.53	8	4	115
14	0.30	2.69	0.12	0.55	7	3	74
15	0.15	1.42	0.10	0.65	1	5	40

Table 3-1. DNM calling results under different minDQ setting

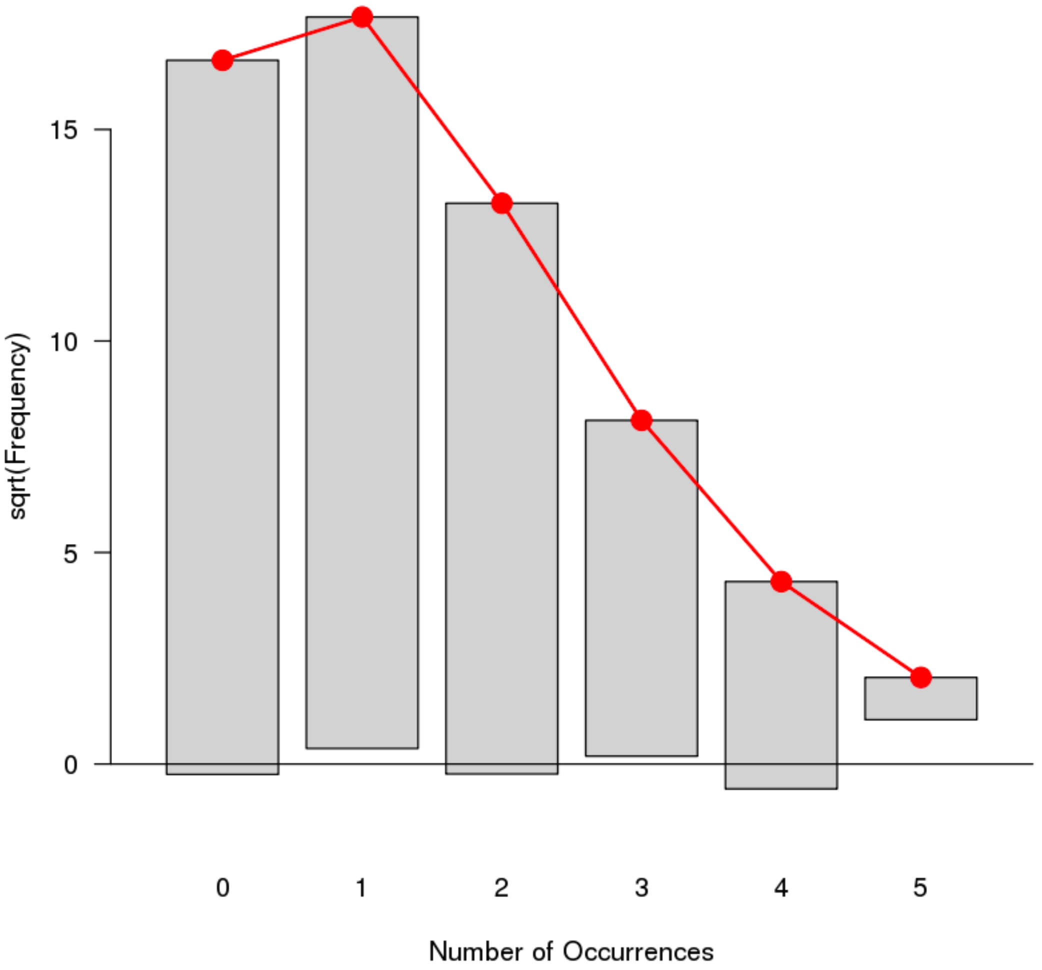


Figure 3-1. Distribution of DNMs per proband.

We annotated the 2,929 DNMs with VEP and PolyPhen-2 and identified a total of 920 deleterious variants defined in the Methods. Previous studies have identified six TS-associated genes with damaging variants. Among these implicated genes, we found one deleterious variant in FBN2. (Table 2)

Gene	#Damaging Var		#Deleterious Var Predicted In Our Data	
	Willsey et al.	Wang et al.	High Effect	Moderate Damaging

WWC1	2	2	0	0
CELSR3	2	3	0	0
OPA1	NA	2	0	0
NIPBL	2	2	0	0
FN1	2	2	0	0
FBN2	NA	2	0	1

Table 3-2. Number of deleterious DNMs identified in genes previously implicated in TS.

DNM enrichment analysis

To examine the enrichment of DNMs, we compared the observed number of DNMs against the expected mutation rate across all genes. A previous study by Samocha et al. estimated the expected mutation rate for all genes based on their characteristics for different types of mutations, including synonymous, missense, nonsense, essential splice site, and frameshift. We thus obtained the count of different types of DNMs per gene based on our previous functional annotation. In total, after excluding outlier samples with over ten DNMs, we had 987 DNMs from 858 complete trios, including families with zero DNMs. We grouped variants into three categories: protein-truncating (PTV), missense, and synonymous variants. Including frameshift, stop gained, splice acceptor, and splice donor mutations, 44 PTVs were found in 44 genes, among which 39 genes were provided with expected mutation rates. Running Poisson test, we observed three genes with significant enrichment using the lenient Bonferroni-corrected threshold, while none passed the stringent multiple-testing threshold. For missense and synonymous mutations, we discovered 683 variants from 654 genes and 246 variants from 241 genes, respectively. However, none of these genes passed our Bonferroni-corrected thresholds. Additionally, we checked whether any of the genes containing PTVs had other types of mutations. One gene, SIRT7, emerged with one additional missense mutation.

Furthermore, we selected 42 genes intolerant to loss-of-function or missense mutations and also bearing at least one PTV or missense variant. We conducted differential expression (DE) analysis in 54 adult human tissues from 8 gene-expression datasets using this gene-set. We tested for differential up-regulation, down-regulation, and both separately. Comparing each tissue type to others, we identified significant enrichment of this gene-set with up-regulation in ten tissue types, down-regulation in four tissue types, and both sides in 17 tissue types. Further analysis using 11 different human brain developmental stages revealed that these 42 genes were significantly up-regulated in early-mid prenatal brain development, late childhood, and significantly down-regulated in mid-adulthood. (Fig 2.)

Discussion

In our previous work, we established the notion that *de* novo damaging sequence variants significantly contributed to TS through observation of enrichment. Using a method based on recurrent DNMs, we identified one high-confident gene, *WWC1*, and three probable-confident genes. Based on these observations, we constructed a model that predicted a greater detecting power with a larger sample size of TS trio data.(18) Therefore, in this regard, we extended the total sample size in this work by sampling almost 900 TS trio families, compared to 674 high-quality trios in the previous study. Using WES data of these samples, we observed an enrichment of protein-truncating DNMs in probands compared to the expected mutation rates in three genes using a lenient Bonferroni-corrected threshold. This is consistent with the observations made in previous studies where de novo Mis3 and damaging variants were enriched in general in TS probands.(18) However, under a stringent significance threshold, we were unable to observe any gene with significant enrichment.

Nonetheless, compared to missense and synonymous variants, PTV displayed greater enrichment, suggesting the deleteriousness of the DNMs was correlated to TS. To note, a different DNM calling pipeline was used between the previous study by Willsey et al. and our study. We employed a stringent calling strategy that included only singletons as candidate DNMs. Although DNMs are very rare events, this strategy was known to be conservative and might lead to a shrinkage of the observed DNM rates compared to theoretical mutation rates. In the next phase of this study, we plan to extend this stringent pipeline and use a minor-allele-frequency-based pipeline that could potentially discover more DNMs in principal.

In terms of the sample size, we have analyzed around 900 TS trio families available at this phase. However, nearly one-third of the total expected samples, namely an additional 300 samples, were expected to arrive by the end of 2021. The sequencing of these samples was in part hindered by the 2020 COVID pandemic. Given the current observation of enriched DNM counts in TS probands, we expected to have a greater detecting power once the final one-third of trio data has arrived. In combination with a more lenient DNM calling pipeline, we would expect to detect more TS-associated genes in the next phase of the analysis.

Our current enrichment analysis was based on the theoretical mutation rates estimated by Samocha et al..(19) They provided detailed expected rates for different types of variants in the general population. However, it would be best to compare the observed number of DNMs in TS dataset to a control dataset. We have recently gained access to a trio WES dataset collected for ASD, the Simon Simplex Collection (SSC).(32) In addition to ASD probands, this dataset contained many families with healthy siblings of the probands, which provided a great opportunity to compare against our TS dataset. Furthermore, as aforementioned, TS is highly comorbid with other psychiatric diseases, including ASD. Analyzing TS probands together with ASD probands would provide additional insights into the shared genetic effects between these two diseases. To note, the SSC WES trio dataset was selected such that the deleterious mutations were depleted compared to the general population. This criterion provided added power to identify TS-associated genes with DNMs that would be missed otherwise.

In one gene, *FBN2*, we observed one missense DNM predicted to be deleterious by VEP and PolyPhen-2. This gene has been previously implicated by Wang et al.(24) with two damaging DNMs. This gene encodes a large protein, fibrillin-2, which is responsible for forming microfibrils and elastic fibers, especially during embryonic development.(33-36) It has been implicated in multiple diseases, including muscular degeneration, congenital contractural arachnodactyly, and TS in our previous work.(24, 34, 37) Given that recurrent mutations have suggested important TS-associated genes, this finding indicated a potential replication and further validation of the previous work. By expanding our sample size and relaxing the stringent workflow, we expect to observe more replicated genes, as well as novel genes, due to greater detecting power.

The functional enrichment analysis we performed indicated a pattern differential expression across different tissue types and brain developmental stages using a gene-set selected as genes with LoF-intolerant PTV and mis-intolerant missense mutations. Notably, enriched upregulation was observed in ten brain-derived tissues, which have been previously implicated in GWAS meta-analysis.(16) This colocalization of expressional regulation could suggest shared TS risk of common and rare variants in these tissues or genes participating in these pathways. Significant upregulation was also found in early-to-mid prenatal brain development, while this gene-set was down-regulated in the mid-to-adult stage, suggesting a parallel change with brain development. Furthermore, the gene, *FBN2*, identified with recurrent DNMs, was linked to embryo development and thus provided a potential connection to the observed expressional regulation in brain development.

Taken together, our study analyzed nearly 900 TS trio families, from which 2,929 high-quality DNMs were called. Functional annotation and enrichment analyses indicated recurrent mutations present in one gene *FBN2* and differential expression patterns across multiple tissue types and brain developmental stages. As more samples are expected, and an improved workflow will be employed, we anticipate a greater detecting power to discover further TS-associated genes. These findings provided important insights into the genetic architecture of TS and facilitated future genetic and clinical studies on TS and potentially other comorbid psychiatric disorders.

Reference

1. 1. Bitsko RH, Holbrook JR, Visser SN, Mink JW, Zinner SH, Ghandour RM, et al. A national profile of Tourette syndrome, 2011-2012. J Dev Behav Pediatr. 2014;35(5):317-22.

2. Robertson MM. The prevalence and epidemiology of Gilles de la Tourette syndrome. Part 1: the epidemiological and prevalence studies. J Psychosom Res. 2008;65(5):461-72.

3. Scharf JM, Miller LL, Gauvin CA, Alabiso J, Mathews CA, Ben-Shlomo Y. Population prevalence of Tourette syndrome: a systematic review and meta-analysis. Mov Disord. 2015;30(2):221-8.

4. Centers for Disease C, Prevention. Prevalence of diagnosed Tourette syndrome in persons aged 6-17 years - United States, 2007. MMWR Morb Mortal Wkly Rep. 2009;58(21):581-5.

5. Charania SN, Danielson ML, Claussen AH, Lebrun-Harris LA, Kaminski JW, Bitsko RH. Bullying Victimization and Perpetration Among US Children with and Without Tourette Syndrome. J Dev Behav Pediatr. 2021.

6. Knight T, Steeves T, Day L, Lowerison M, Jette N, Pringsheim T. Prevalence of tic disorders: a systematic review and meta-analysis. Pediatr Neurol. 2012;47(2):77-90.

7. Hirschtritt ME, Lee PC, Pauls DL, Dion Y, Grados MA, Illmann C, et al. Lifetime prevalence, age of risk, and genetic relationships of comorbid psychiatric disorders in Tourette syndrome. JAMA Psychiatry. 2015;72(4):325-33.

8. Ghanizadeh A, Mosallaei S. Psychiatric disorders and behavioral problems in children and adolescents with Tourette syndrome. Brain Dev. 2009;31(1):15-9.

9. Eapen V, Cavanna AE, Robertson MM. Comorbidities, Social Impact, and Quality of Life in Tourette Syndrome. Front Psychiatry. 2016;7:97.

10. Cravedi E, Deniau E, Giannitelli M, Xavier J, Hartmann A, Cohen D. Tourette syndrome and other neurodevelopmental disorders: a comprehensive review. Child Adolesc Psychiatry Ment Health. 2017;11:59.

11. Scharf JM, Yu D, Mathews CA, Neale BM, Stewart SE, Fagerness JA, et al. Genome-wide association study of Tourette's syndrome. Mol Psychiatry. 2013;18(6):721-8.

12. Freeman RD, Fast DK, Burd L, Kerbeshian J, Robertson MM, Sandor P. An international perspective on Tourette syndrome: selected findings from 3,500 individuals in 22 countries. Dev Med Child Neurol. 2000;42(7):436-47.

13. Quezada J, Coffman KA. Current Approaches and New Developments in the Pharmacological Management of Tourette Syndrome. CNS Drugs. 2018;32(1):33-45.

14. Mataix-Cols D, Isomura K, Perez-Vigil A, Chang Z, Ruck C, Larsson KJ, et al. Familial Risks of Tourette Syndrome and Chronic Tic Disorders. A Population-Based Cohort Study. JAMA Psychiatry. 2015;72(8):787-93.

15. Davis LK, Yu D, Keenan CL, Gamazon ER, Konkashbaev AI, Derks EM, et al. Partitioning the heritability of Tourette syndrome and obsessive compulsive disorder reveals differences in genetic architecture. PLoS Genet. 2013;9(10):e1003864.

16. Yu D, Sul JH, Tsetsos F, Nawaz MS, Huang AY, Zelaya I, et al. Interrogating the Genetic Determinants of Tourette's Syndrome and Other Tic Disorders Through Genome-Wide Association Studies. Am J Psychiatry. 2019;176(3):217-27.

17. Huang AY, Yu D, Davis LK, Sul JH, Tsetsos F, Ramensky V, et al. Rare Copy Number Variants in NRXN1 and CNTN6 Increase Risk for Tourette Syndrome. Neuron. 2017;94(6):1101-11 e7.

18. Willsey AJ, Fernandez TV, Yu D, King RA, Dietrich A, Xing J, et al. De Novo Coding Variants Are Strongly Associated with Tourette Disorder. Neuron. 2017;94(3):486-99 e9.

19. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014;46(9):944-50.

20. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43:11 0 1- 0 33.

21. Tuzov N. A framework for the estimation of the proportion of true discoveries in single nucleotide variant detection studies for human data. PLoS One. 2018;13(4):e0196058.

22. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559-75.

23. Wei Q, Zhan X, Zhong X, Liu Y, Han Y, Chen W, et al. A Bayesian framework for de novo mutation calling in parents-offspring trios. Bioinformatics. 2015;31(9):1375-81.

24. Wang S, Mandell JD, Kumar Y, Sun N, Morris MT, Arbelaez J, et al. De Novo Sequence and Copy Number Variants Are Strongly Associated with Tourette Disorder and Implicate Cell Polarity in Pathogenesis. Cell Rep. 2018;24(13):3441-54 e12.

25. Team RC. R: A language and environment for statistical computing. Vienna, Austria.: R Foundation for Statistical Computing,; 2013.

26. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17(1):122.

27. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-9.

28. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43.

29. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017;8(1):1826.

30. Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, et al. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. Biopreserv Biobank. 2015;13(5):311-9.

31. Hawrylycz MJ, Lein ES, Guillozet-Bongaarts AL, Shen EH, Ng L, Miller JA, et al. An anatomically comprehensive atlas of the adult human brain transcriptome. Nature. 2012;489(7416):391-9.

32. Fischbach GD, Lord C. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. Neuron. 2010;68(2):192-5.

33. Zhang H, Apfelroth SD, Hu W, Davis EC, Sanguineti C, Bonadio J, et al. Structure and expression of fibrillin-2, a novel microfibrillar component preferentially located in elastic matrices. J Cell Biol. 1994;124(5):855-63.

34. Ratnapriya R, Zhan X, Fariss RN, Branham KE, Zipprer D, Chakarova CF, et al. Rare and common variants in extracellular matrix gene Fibrillin 2 (FBN2) are associated with macular degeneration. Hum Mol Genet. 2014;23(21):5827-37.

35. Quondamatteo F, Reinhardt DP, Charbonneau NL, Pophal G, Sakai LY, Herken R. Fibrillin-1 and fibrillin-2 in human embryonic and early fetal development. Matrix Biol. 2002;21(8):637-46.

36. Lee B, Godfrey M, Vitale E, Hori H, Mattei MG, Sarfarazi M, et al. Linkage of Marfan syndrome and a phenotypically related disorder to two different fibrillin genes. Nature. 1991;352(6333):330-4.

37. Deng H, Lu Q, Xu H, Deng X, Yuan L, Yang Z, et al. Identification of a Novel Missense FBN2 Mutation in a Chinese Family with Congenital Contractural Arachnodactyly Using Exome Sequencing. PLoS One. 2016;11(5):e0155908.

Supporting information

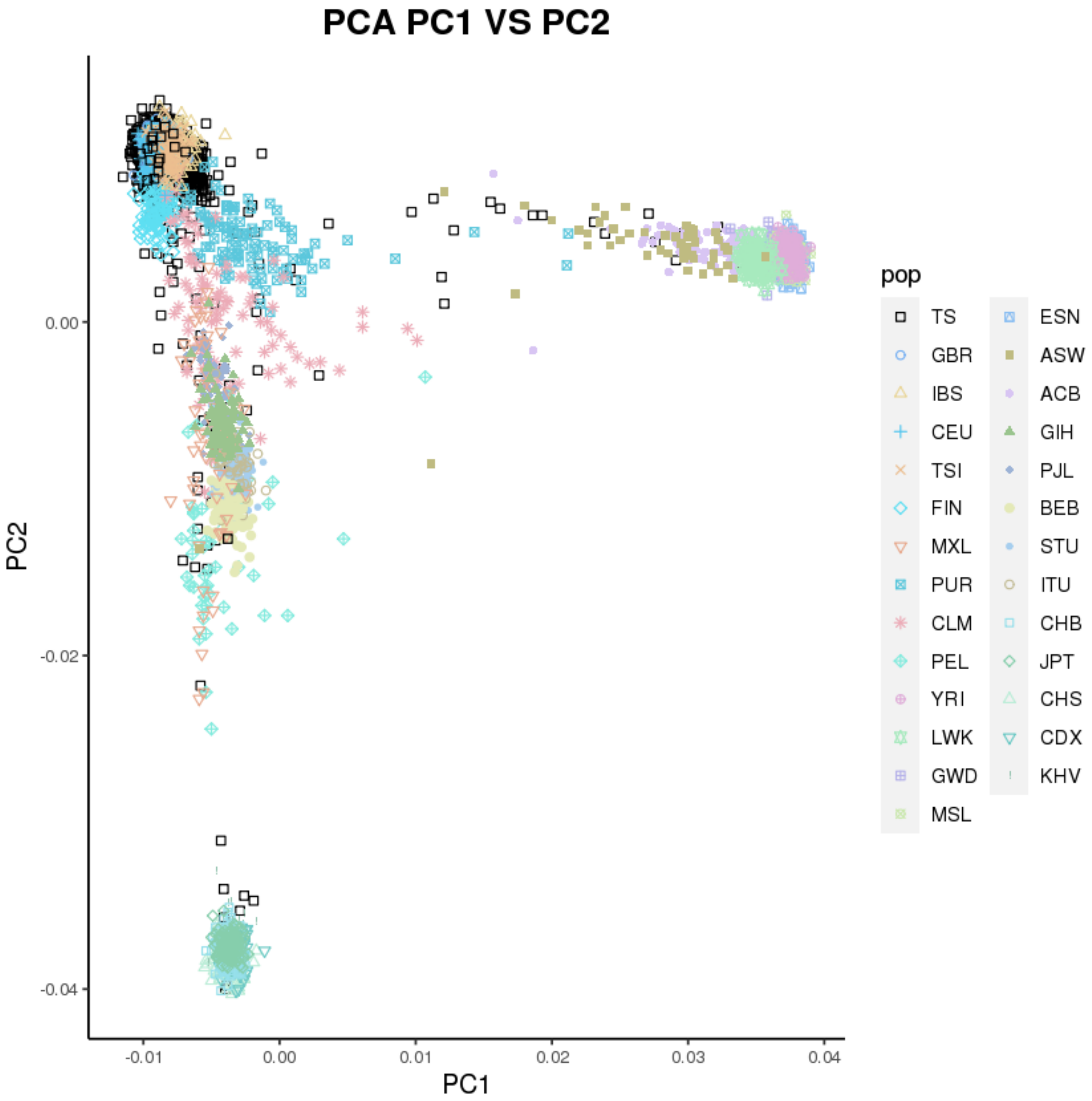


Fig S1. PCA plot (PC1 v.s. PC2) using all parents in TS trio dataset

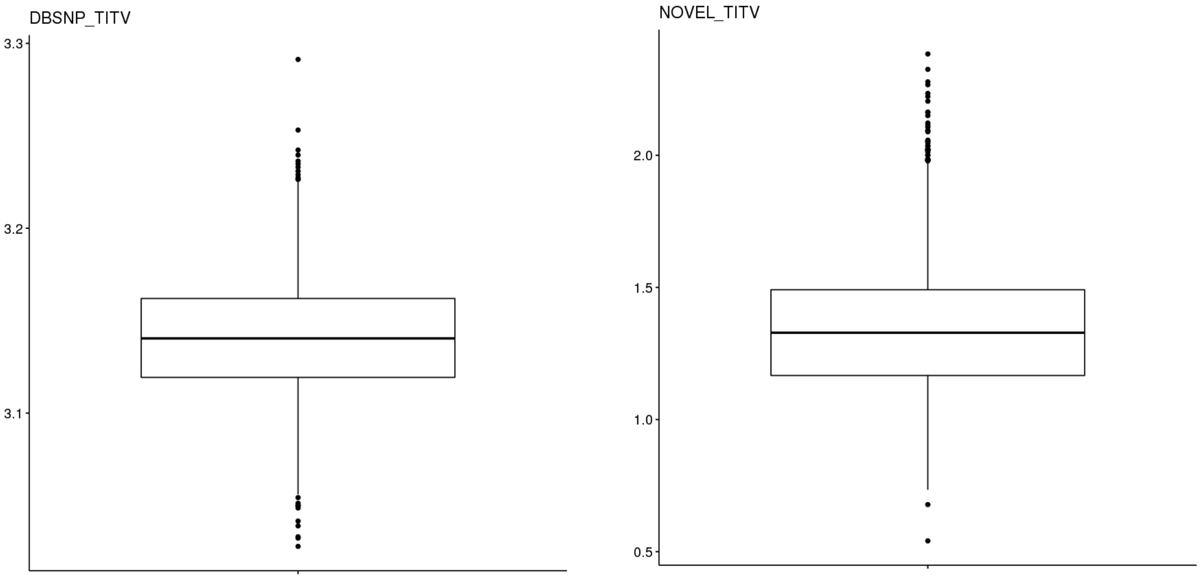


Fig S2. Ti/Tv ratio of all SNPs present in dbSNP database (left) and novel SNPs (right).

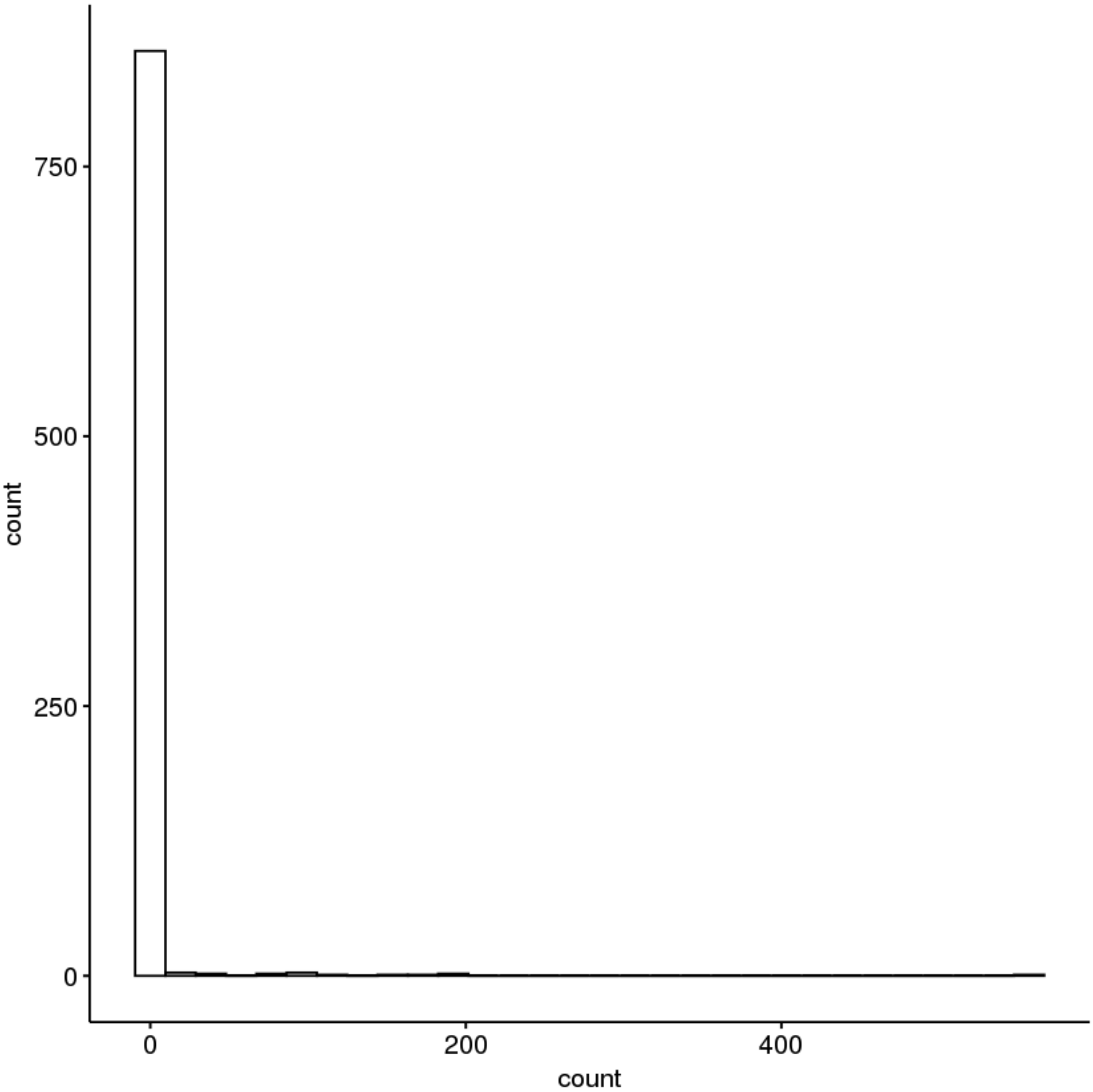


Fig S3. Distribution of DNMs per proband.

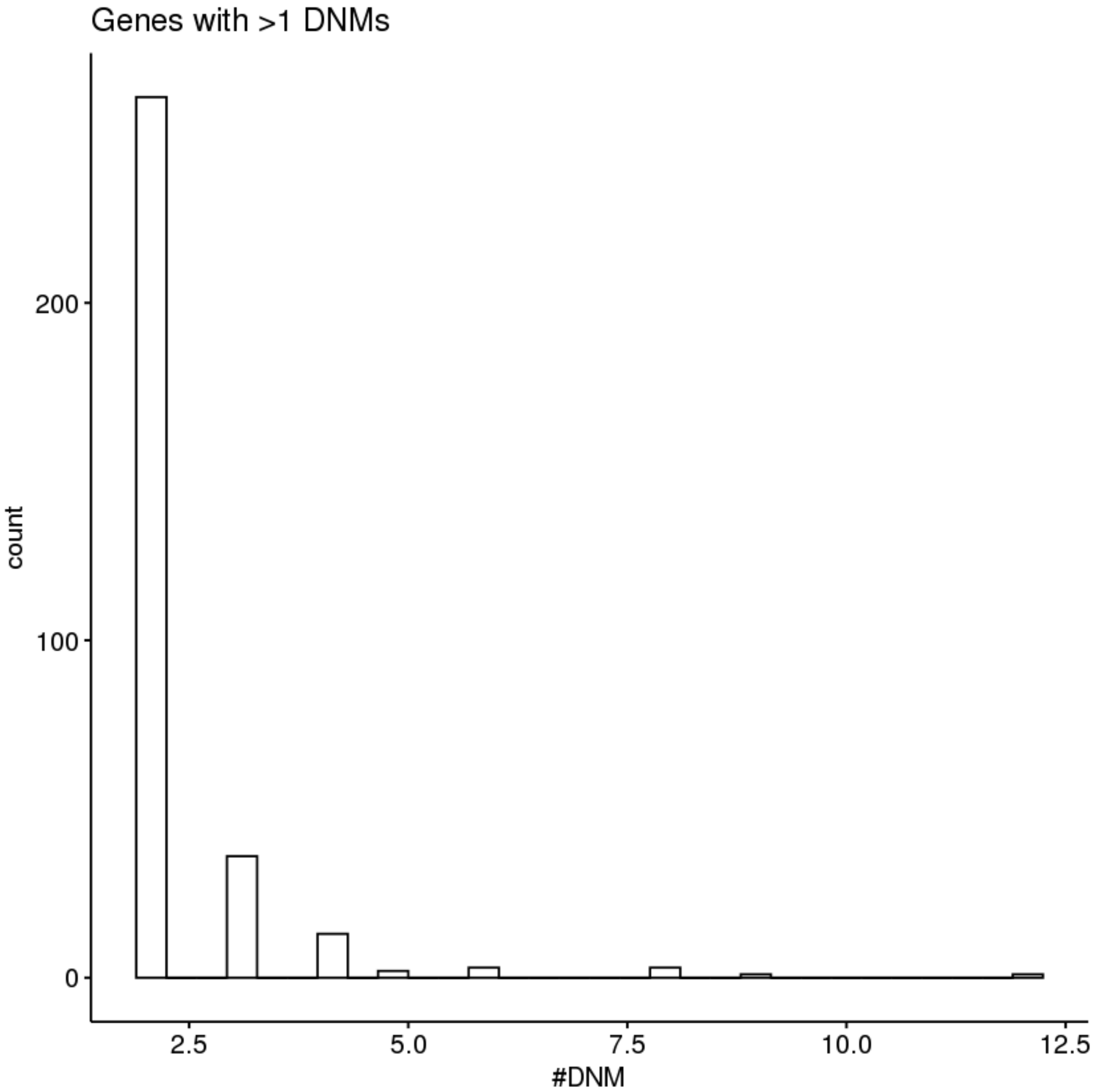


Fig S4. Distribution of DNMs per gene. Genes with no DNM called were excluded.

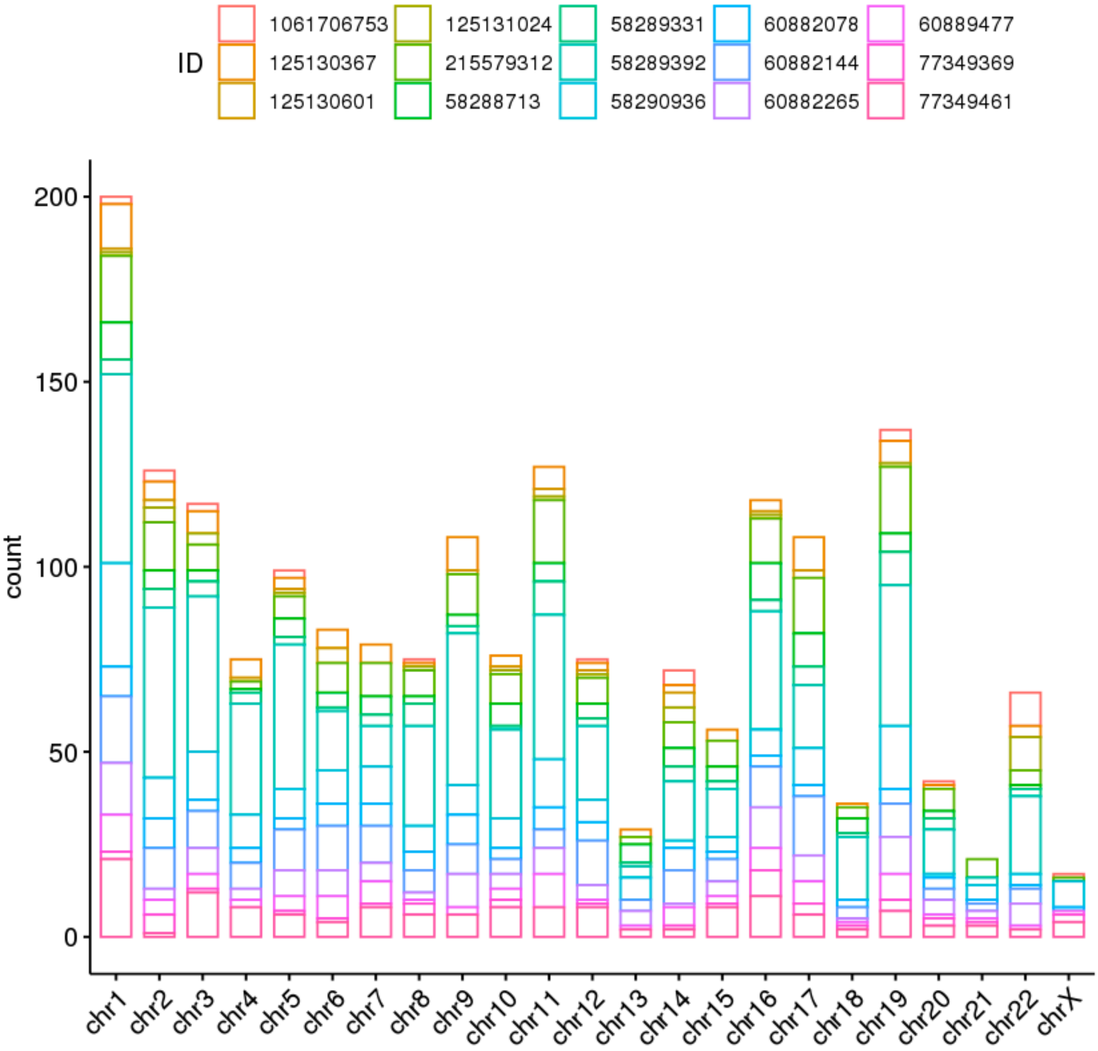


Fig S5. Distribution of DNMs per chromosome.

Chapter 4 – Multi-population meta-analysis of blood lipid levels identify one novel locus, rs72552763, in UCLA ATLAS Precision Health Biobank

Introduction

Blood lipids levels are a major contributor to various long-term health conditions, including type 2 diabetes, fatty liver disease, and especially atherosclerotic cardiovascular disease, an increasingly prevalent disease and the leading cause of death globally. (1-5) Research have suggested that blood lipid levels are highly heritable polygenic traits, with an estimation of heritability of 40%, 51%, 33%, and 51% for high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), triglycerides (TG), and total cholesterol (TC), respectively. (6, 7) To date, genome-wide association studies (GWAS) have revealed over 444 risk loci (8-17) associated with changes in blood concentrations of these lipid phenotypes, which further pointed to risk factors such as *ANGPTL4*, *LPL*, and *SVEP1* serving as potential therapeutic targets. (18) However, current findings through GWAS only explain about one-quarter to one-third of the heritability of these lipid phenotypes (12.8% HDL-C, 19.5% LDL-C, 9.3% TG, and 18.8% TC) (19-21), and a majority of these studies focused on European ancestry while underrepresenting other ancestries, such as African, Admdixed American, South Asian, and East Asian populations.

To address these issues would require very large sample sizes together with a diversity of ancestry backgrounds, which is both economically intense and practically time-consuming. However, health data documented in electronic health records (EHR) provide a resourceful and efficient means to overcome these limitations while presenting diverse and extensive health-related phenotypes for a large number of individuals potentially with a variety of ancestry backgrounds. (22) Recent advancements in building large-scale biobanks, such as UK Biobank and UCLA ATLAS Precision Health Biobank, and linking to EHR further facilitate genetic studies with unprecedented power and convenient rich resources of phenotypic data. (23) Previous studies have demonstrated the successful usage of the combination of biobank and EHR data in analyzing the genetic associations of multiple phenotypes, including obesity, breast cancer, and blood lipids (9, 24-26), and have provided important knowledge from real-world patients towards the genetic architecture and the etiology of common complex diseases.

In this study, we leveraged the combination of the genotyping data of 26,414 individuals from UCLA ATLAS Precision Health Biobank and four blood lipid phenotypes (HDL-C, LDL-C, TG, and TC) from UCLA EHR and developed a pipeline to perform GWAS on the selected phenotypes. In our pipeline, we first identified all individuals with available phenotypic measurements in EHR for each blood lipid category, followed by intersecting with the genotyping data available to us. In total, we identified 117,535, 120,186, and 139,454 individuals with available measurements of HDL-L, TG, and TC, respectively. For LDL-C, we observed two different measuring methods, one by calculation using HDL-L, TG, and TC (LDL-C Calc) and one by quantitation through direct measurements (LDL-C Quant). We considered them separately and recorded 104,260 and 19,201 individuals with LDL-C Calc and LDL-C Quant, respectively. Among these patients, we identified 17,226, 16,948, 4,599, 17,429, and 17,377 individuals with complete phenotypic (HDL-C, LDL-C Calc, LDL-C Quant, TG, and TC, respectively) and genetic information. To elucidate population stratification, we performed principal component analysis (PCA), which revealed five primary ancestries (European, Admixed American, African, East Asian, South Asian) among our samples. We then performed genome-wide association (GWA) analyses on common variants for each blood lipid phenotypes within each ancestry group, followed by meta-analyzing across all five populations. Our GWA analyses revealed 236 genome-wide significant hits across all populations where a majority of the signals came from the European population. To validate our findings, we extracted all variants with a p-value below 10^{-3} and compared and contrasted them to the GWAS results estimated using UK Biobank data, one of the largest biobank data worldwide. For all selected variants, our results displayed a relatively high correlation with UK Biobank results in terms of the p-values of the same variant in both datasets. Particularly for the genome-wide significant hits, we observed a nearly 100% replication between these two biobanks where we identified two novel hits for triglyceride among European and South Asian samples. In the meta-analysis, we demonstrated 26, 10, 2, 3, and 9 genome-wide significant signals for HDL-L, LDL-L Calc, LDL-L Quant, TG, and TC, respectively, where all hits were confirmed using UK Biobank data. In brief, our study represented one of the first GWA analyses using UCLA ATLAS Precision Health Biobank and UCLA EHR data and demonstrated the potentiality of EHR in providing ancestrally diverse and phenotypically abundant data for genetic studies. Our analyses on lipid phenotypes provided an easy pipeline to work with both UCLA data sources and revealed 236 population-specific genome-wide significant loci associated with blood lipid levels, where 50 of them remained significant after meta-analysis. Our results indicated a strong component of population-specific genetic effects in determining the levels of various blood lipid molecules and provided further understanding of their genetic architectures and potential targets for pharmaceutical and clinical research.

Results

Demographics of UCLA ATLAS population

A total of 26,414 patients were genotyped in high quality available in UCLA ATLAS Precision Health Biobank. Details of sample-level and variant-level quality control procedures can be found in Methods. Leveraging genetic information, we characterized these patients into five mutually exclusive ancestral groups, including European, African, Admixed American, South Asian, and East Asian, where over 67% were of European ancestry. (Table 1) We then obtained blood lipids phenotypes from the UCLA EHR database and identified

117,535, 104,260, 19,201, 120,186, and 139,454 individuals with HDL-C, LDL-C Calc, LDL-C Quant, TG, and TC lipid measurements, respectively. After merging with genotyping data, we retained 17,226, 16,948, 4,599, 17,429, and 17,377 individuals with a complete set of phenotypic information, including blood lipid measurements (HDL-C, LDL-C Calc, LDL-C Quant, TG, and TC, respectively), sex, age, and BMI. Across the identified five ancestral groups, we observed the lowest HDL-C level of 50.3 and the highest TG level of 141.7 in Admixed American, the highest calculated LDL-C level of 105.9 and TC level of 187.4 in Europeans, and the highest quantitated LDL-C level of 120.2 in South Asians. A detailed description of the demographic information can be found in Table 2.

Phenotype	EUR	AMR	AFR	EAS	SAS	Total	EUR (%)
HDL	11,652	2,637	959	1,700	278	17,226	67.64%
LDL Calculated	11,456	2,602	950	1,666	274	16,948	67.59%
LDL Quantitated	3,090	576	307	552	74	4,599	67.19%
Total Cholesterol	11,734	2,681	961	1,718	283	17,377	67.53%
Triglyceride	11,762	2,692	968	1,722	285	17,429	67.49%

Table 4-1. Number of samples with genotyping information identtified in the five popuations

Phenotype	HDL	LDL Calculated	LDL Quantitated	Total Cholesterol	Triglyceride
EUR	57.12	105.90	112.16	123.99	187.44
AFR	55.36	104.46	114.50	104.70	180.38
AMR	50.33	98.91	113.36	141.66	176.83
SAS	51.45	101.28	120.24	136.30	180.38
EAS	57.43	100.59	111.35	130.91	183.52
Age	51.77	52.07	55.70	51.78	51.67

Male (%)	46.25%	46.34%	42.29%	46.39%	46.27%
BMI	28.05	28.09	27.78	28.03	32.16

Table 4-2. Demographic information for five blood lipid levels

Row one to five are the mean concentration for each blood lipid measurement in each populaiton. Row six and eight are the mean age and BMI of all samples for a given blood lipid phenotype. Row seven is the percernage of males for a given blood lipid phenotype.

GWAS of the five blood lipid phenotypes

In order to identify population-specific effects, we performed genome-wide association tests for each of the five blood lipid phenotypes and for each ancestral group separately, adjusted for covariates including sex, age, and BMI. Results of the genome-wide association analyses for HDL-C in European ancestry were summarized in Fig 1 and 2. Manhattan plots and Q-Q plots for other populations were shown in Figures S1-S48, and the inflation factors for all combinations of blood lipid phenotypes and ancestry groups were shown in Table S1. Across all populations and phenotypes, we identified a total of 236 loci that surpassed the genome-wide significance threshold (p-value < $5*10^{-8}$). (Table 3) A majority of the GWAS hits were observed for European ancestry as it represented the largest ancestral group in our study. Nonetheless, we identified a number of significant signals in other populations, including the smallest South Asian population. Specifically, for HDL-C, we identified nine significant loci in Admixed Americans and six in East Asians. For LDL Calc, one and two significant loci were present in Africans and East Asians, respectively. Lastly, for TG, we found one significant locus in Admixed Americans, two in East Asians, and 1 in South Asians. The LDL Quant lipid measurements were available for the least number of patients and therefore did not deliver any significant loci for non-European populations.

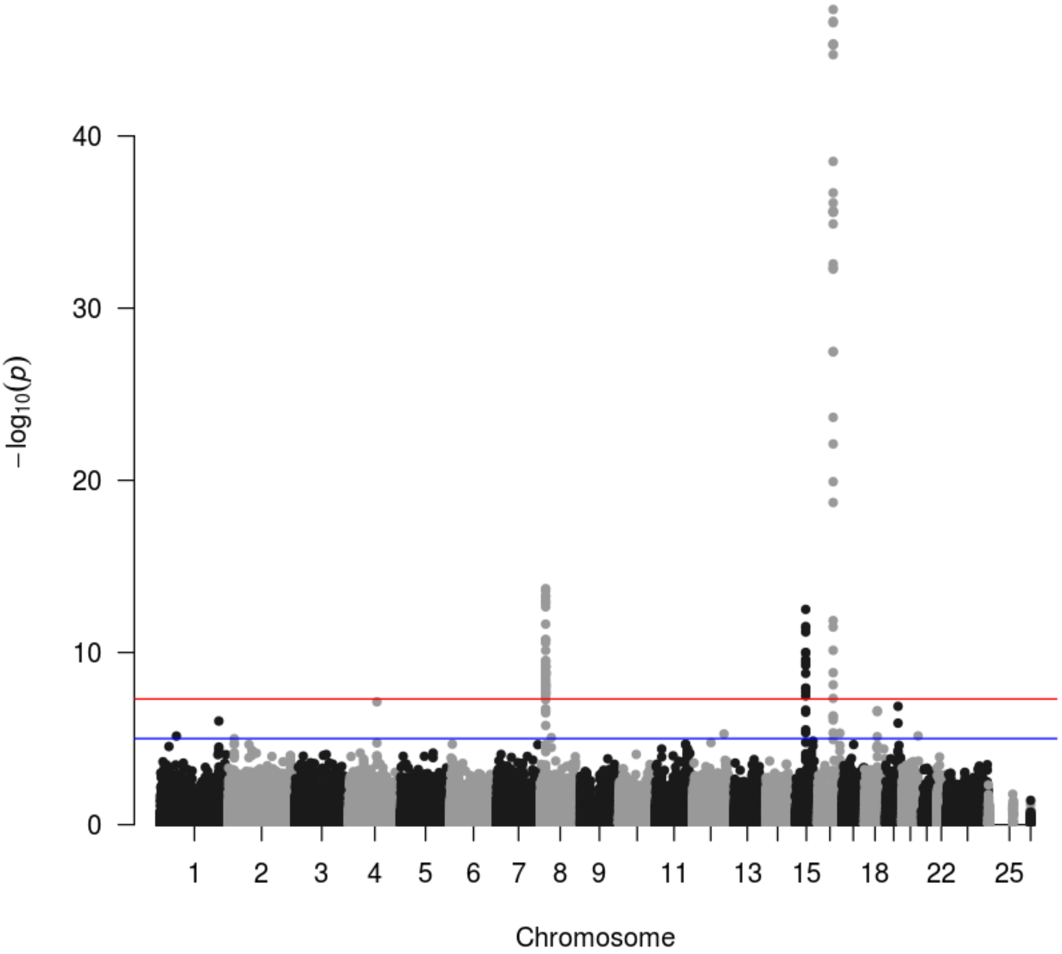


Figure 4-1. Manhattan plot of HDL-C in EUR population

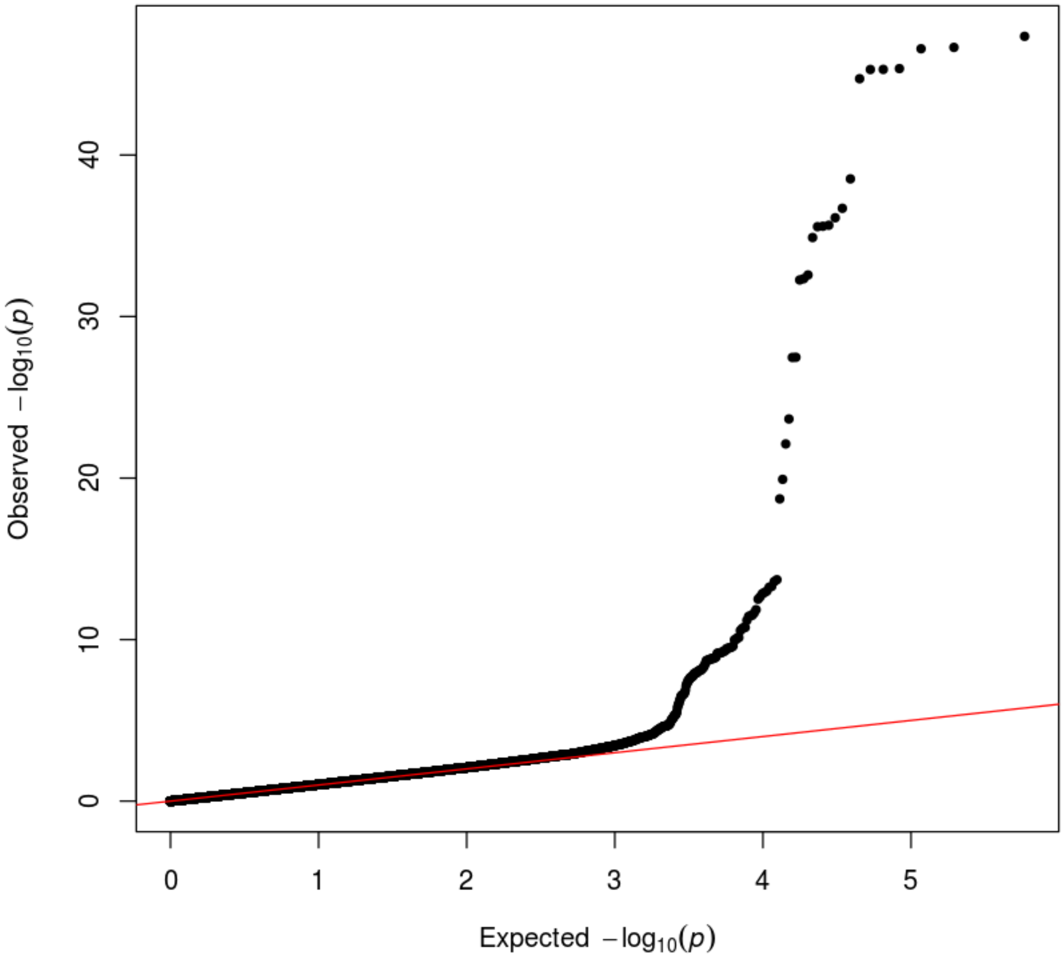


Figure 4-2. Q-Q plot of HDL-C in EUR population

Significant Hits	UKBB	EUR	AMR	AFR	EAS	SAS	Meta
HDL	52,619	95	9	0	6	0	95
LDL Calculated	33,488	30	0	1	2	0	34
LDL Quantitated	33,488	3	0	0	0	0	7
Total Cholesterol	41,960	13	0	0	0	0	18
Triglyceride	47,230	73	1	0	2	1	110

Table 4-3. Number of GWAS hits identified for each blood lipid phenotype and ancestry group in UCLA ATLAS dataset and in UK Biobank(27)

To incorporate the maximum possible number of samples and identify genetic effects common to all populations, we performed meta-analyses for 73,579 individuals across all five populations for each blood lipid phenotypes using an inverse-variance-weighted fixed-effects method. (See Methods) In total, we identified 264 genome-wide significant loci across all lipid traits meta-analyses. In particular, we identified five additional signals for HDL-C, four for LDL Calc, four for LDL Quant, 37 for TG, and six for TC. (Table 3) Manhattan plots of and Q-Q plots of meta-analysis results were shown in Fig 3-4 and Fig S49-S56. Overall, the meta-analysis was also able to reveal a greater number of genome-wide significant signals previously missing in population-specific GWA analyses. Notably, although LDL Quant contained the smallest number of measured patients in population-specific analyses, the meta-analysis identified two times more significant loci than those discovered previously in all five ancestry groups combined. The highest number of genome-wide significant loci were observed in TG, where a total of 110 signals were seen across the five ancestry groups. Given only 184 unique GWA loci were observed in the previous population-specific analyses, our meta-analysis was able to recover 9.23% more signals when the samples were combined into a meta-analysis with larger total sample size. In addition, some of the GWAS hits identified in specific populations were not observed in the meta-analysis, such as one genome-wide significant locus, which was only observed in the South Asian population but not in the meta-analysis. This observation suggested that these signals were potentially not shared across different ancestry groups but represented population-specific effects.

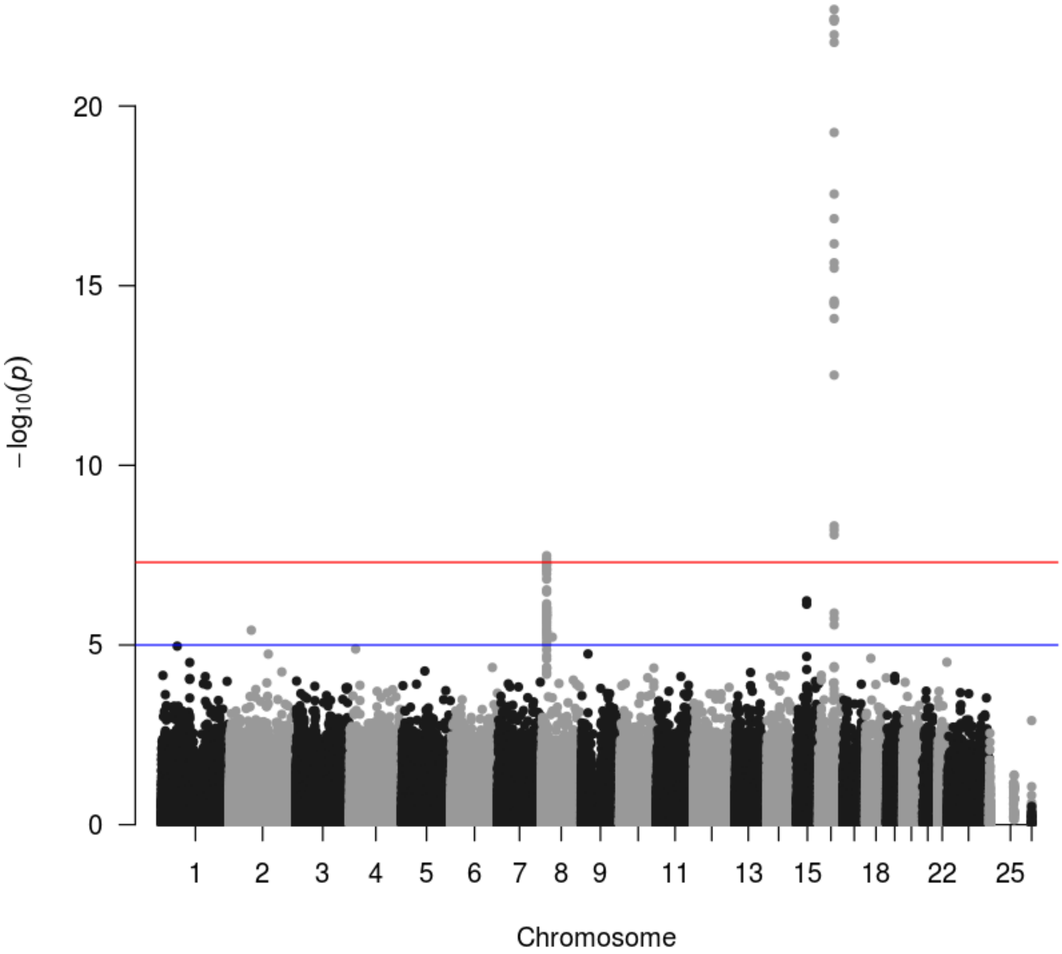


Figure 4-3. Manhattan plot of HDL-C in meta-analysis

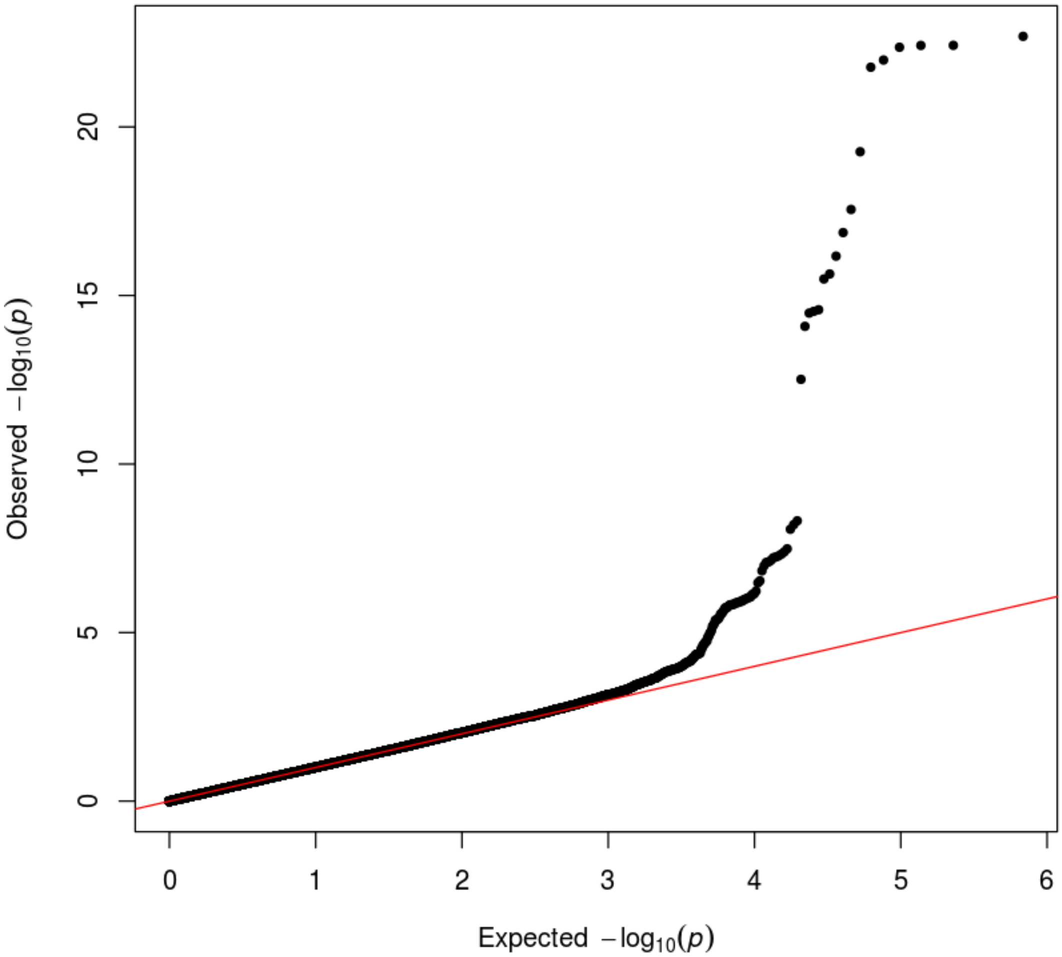


Figure 4-4. Q-Q plot of HDL-C in meta-analysis

Evaluation of previously established loci in UCLA ATLAS population

We first evaluated the overall trends of the effect estimates for all 444 known independent GWAS hits previously reported by the Global Lipids Genetics Consortium (GLGC)(11) within each lipid phenotype and ancestry group. Specifically, we obtained effect sizes estimated in the UCLA ATLAS population as well as from the GLGC study for these 444 previously identified loci and compared the strength of their estimates. To note, as the UCLA ATLAS genotyping data contained a fewer number of variants, we retained only variants present in both datasets for this evaluation. For LDL Calc, LDL Quant, and TC, we observed the strongest correlation within the effect sizes estimated from the European ancestry. (Table

4, Fig 5, and Fig S57-S80) In HDL-C and TG, the European samples demonstrated the second strongest correlated effect sizes to the GLGC estimates. This behavior was expected because the GLGC study consisted of primarily European samples (84%)(10). However, interestingly, the strongest correlations for HDL-C and TG were observed within the East Asian samples, which also showed the second strongest correlation in the other three lipid traits. Comparing across phenotypes, the LDL Calc demonstrated consistently strong correlations to the GLGC estimates, while the TG had the weakest correlations for all ancestry groups.

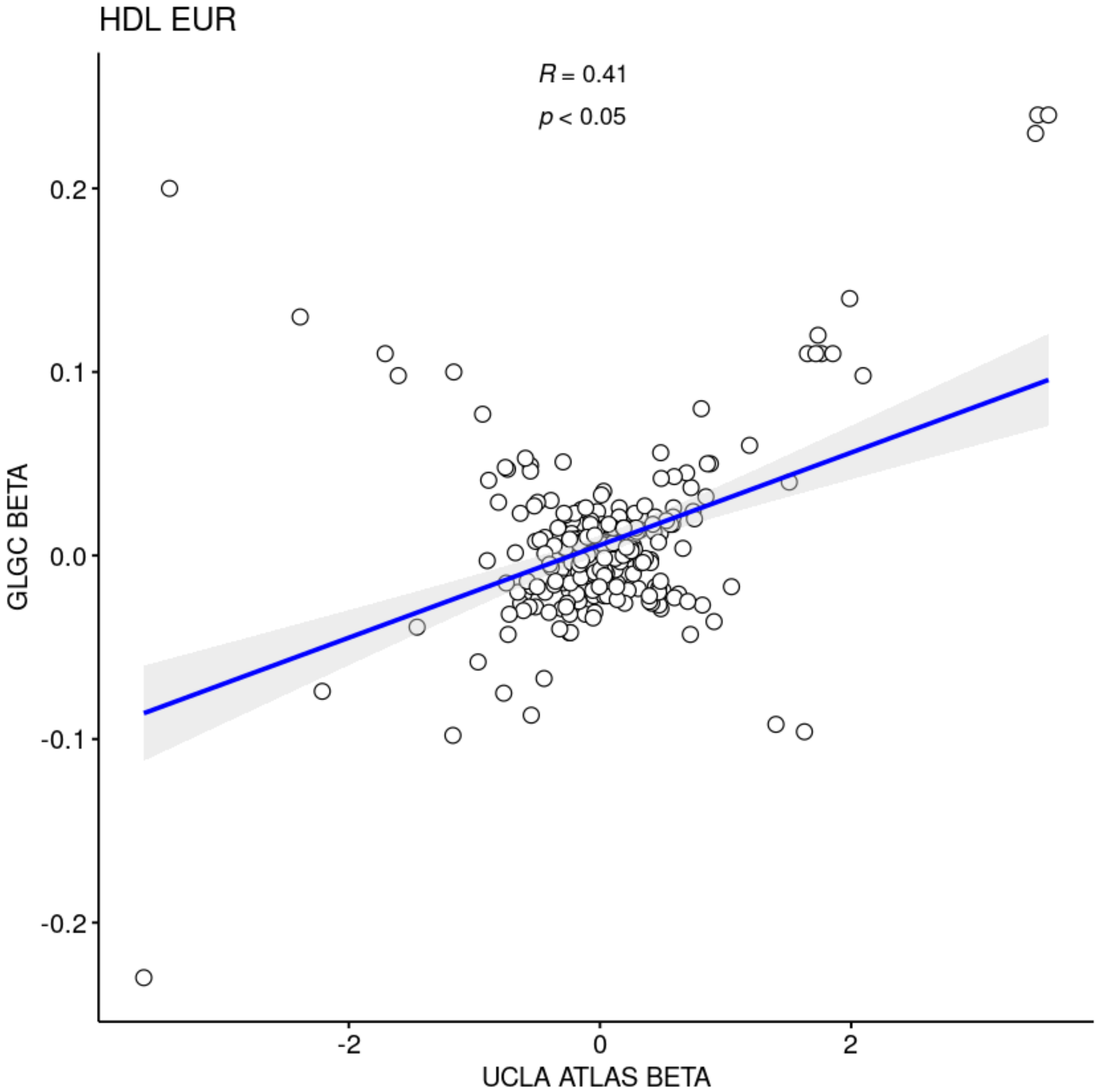


Figure 4-5. Correlation of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for HDL-C in EUR population.

Phenotype	EUR	AFR	AMR	SAS	EAS
HDL	0.41	0.19	0.32	0.24	0.45
LDL Calculated	0.72	0.36	0.29	0.28	0.47
LDL Quantitated	0.52	0.28	0.26	0.33	0.39
Triglyceride	0.14	0.04	0.04	0.03	0.21
Total Cholesterol	0.48	0.27	0.22	0.13	0.28

Table 4-4. Correlation coefficients of effect sizes of 444 known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset

Next, as our dataset was EHR-based, we explored an external large-scale EHR dataset, the UK Biobank dataset(27), to validate our observations across the same type of data. In particular, the UK Biobank dataset contained summary statistics estimated from 315,133, 343,621, 343,992, and 344,278 samples for HDL-C, LDL, TG, and TC, respectively. To note, there was only one type of LDL trait available in the UK Biobank dataset, the estimates from which were thus used to compare with both LDL Calc and LDL Quant results in our study. A summary of the number of SNVs and significant loci were shown in Tables 3 and 5. As the UK Biobank dataset was imputed, there were around 40 times more total SNVs than the UCLA ATLAS dataset to start with. Also, due to the larger sample size of the UK Biobank dataset, we observed much more significant loci across all lipid phenotypes compared to the UCLA ATLAS dataset. In our meta-analysis for HDL-C, we observed 95 genome-wide significant loci, while the UK Biobank revealed 52,619 significant hits. Thus, to compare results from these two datasets, we first selected SNVs under different p-value thresholds in the UCLA ATLAS dataset and compared their effect sizes in the two datasets. For European samples, the correlations of effect sizes were positive for all lipid phenotypes and increased with the p-value thresholds decreased. (Fig 6.) In other words, we observed more consistent effect sizes estimated using the UCLA ATLAS dataset with those from the UK Biobank dataset. For other ancestry groups, the trends were also similar with a few exceptions when the sample size and the number of SNVs under certain thresholds were both small, such as those estimated within the South Asian population for LDL Quant. (Fig S81-84.)

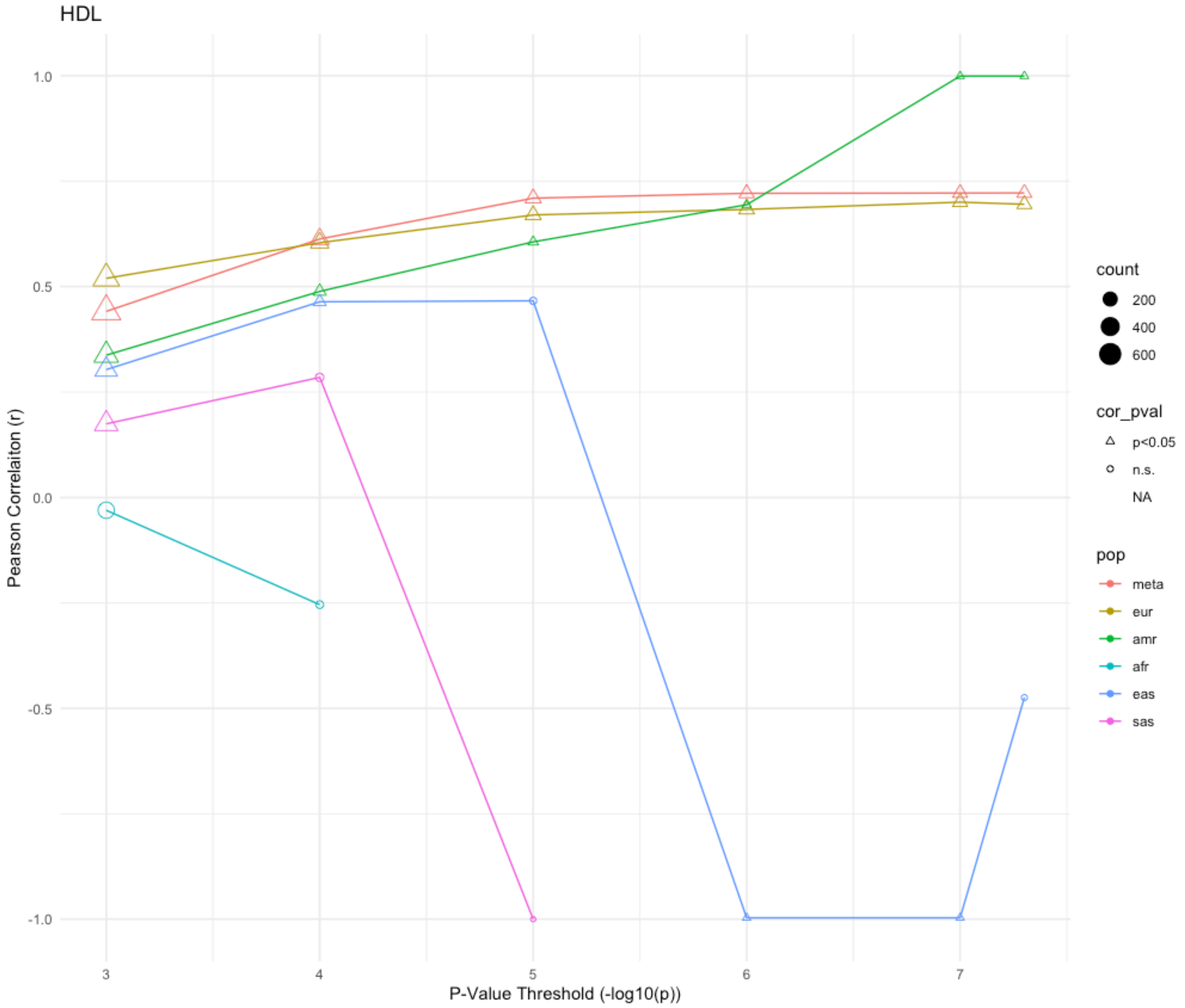


Figure 4-6. Correlation coefficients of the effect sizes of top SNPs identified in UCLA ATLAS and estimated in UK Biobank for HDL-C.

Top SNPs were selected from six p-value thresholds: 1e-3, 1e-4, 1e-5, 1e-6, 1e-7, and 5e-8. The number of top SNPs within each threshold that were also identified in UK Biobank was should represented by circle size. Significant correlation at each threshold was shown with triangle. Effect sizes estimated for all ancestry groups and in meta-analysis were shown. A missing point represented that none of the SNPs passed a given threshold in UCLA ATLAS dataset or less than three passing SNPs was found in UK Biobank.

SNPs	UKBB	Meta-Analysis	EUR	AMR	AFR	EAS	SAS
HDL	13,789,520	343,866	292,275	313,138	327,608	291,351	323,169

LDL Calculated	13,789,683	343,539	292,257	313,014	327,628	291,186	321,999
LDL Quantitated	13,789,683	344,332	293,609	314,678	327,381	290,914	320,160
Total Cholesterol	13,789,686	343,488	292,285	313,018	327,385	291,237	321,858
Triglyceride	13,789,685	343,604	292,243	312,965	327,523	291,082	322,345

Table 4-5. Number of SNPs analyzed for each blood lipid phenotype and ancestry group in UCLA ATLAS dataset and in UK Biobank(27)

Additionally, we performed a consistency evaluation to check how many of the observed signals in our dataset could also be found in the UK Biobank dataset using summary statistics(27). Specifically, we checked the percentage of overlapping SNVs under given p-value thresholds (1e-3, 1e-4, 1e-5, 1e-6, 1e-7, 5e-8) for both datasets. Similar to previous evaluation, in general, as the thresholds increased, a higher percentage of signals in the UCLA ATLAS dataset could be found in the UK Biobank dataset across all ancestry groups and lipid phenotypes. (Fig 7 and Fig S85-S88) For the meta-analysis, over 90% of the variants under the p-value threshold of 1e-5 were estimated with similar significance in the UK Biobank dataset as in the UCLA ATLAS dataset. Interestingly, for TG meta-analysis, we identified two genome-wide significant loci, rs72552763 and rs6589566, that were not present in UK Biobank. Further checking with the 444 GLGC GWAS hits and the GWAS Catalog revealed that the variant rs72552763 was indeed novel while the other loci have been reported for TG in previous studies.

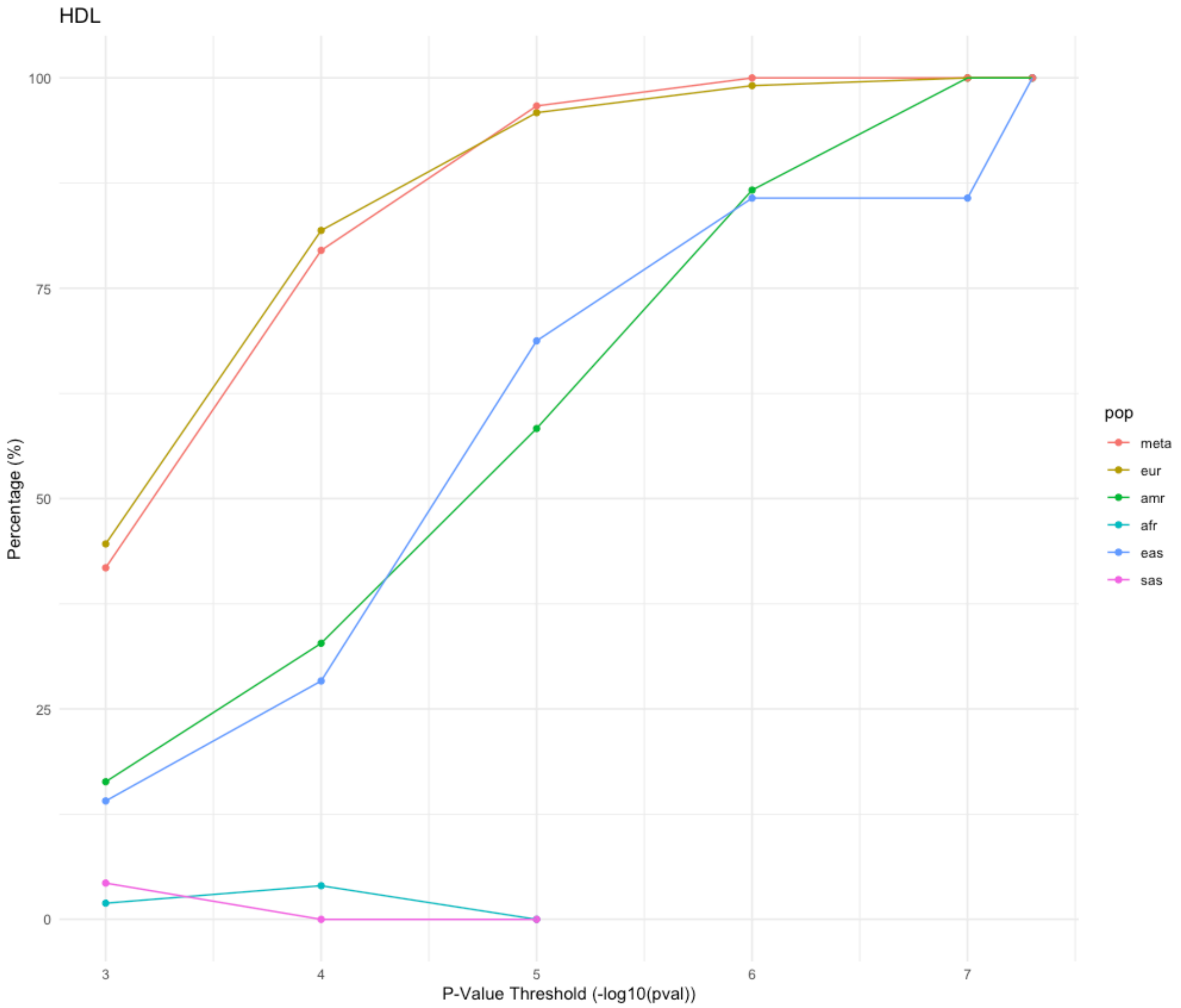


Figure 4-7. Percentage of overlapping SNPs between UCLA ATLAS dataset and UK Biobank(27) under given p-value thresholds for HDL-C.

SNPs were selected under six p-value thresholds: 1e-3, 1e-4, 1e-5, 1e-6, 1e-7, and 5e-8. The percentage of SNPs was computed based on the total number of SNPs passing a given threshold in UCLA ATLAS dataset and the number of SNPs among them that also passed the same p-value threshold in UK Biobank. A missing point represented that none of the SNPs passed a given threshold in UCLA ATLAS dataset.

The next evaluation was performed to check the amount of heterogeneity within the effect sizes estimated in the meta-analysis. We looked at the I^2 statistic and divided the variants into four categories based on this measurement (0 - 25%, 25% - 50%, 50% - 75% , and 75% - 100%). The I^2 statistic indicated the percentage of the variability in the effect estimates that could be explained by heterogeneity instead of sampling error, and thus, the categories ranged from low to moderate, high, and very high heterogeneity. The number of loci under different heterogeneous categories for each blood lipid phenotype was summarized in Table 6 and Table S2-S5. Focusing on only the genome-wide significant loci in the meta-analysis, we observed the lowest level of heterogeneity for TC where only 21.05% variants were in the high or very high category. (Table S4) The TG, however, represented the most heterogeneous phenotypes across populations where 68.76% of the variants were considered to be high or very high heterogeneous. (Table S5)

Threshold	Low	Medium	High	Very	Total	Low (%)	Medium (%)	High (%)	Very (%)
1.E-03	418	117	107	71	713	58.63%	16.41%	15.01%	9.96%
1.E-04	101	43	40	26	210	48.10%	20.48%	19.05%	12.38%
1.E-05	46	35	29	21	131	35.11%	26.72%	22.14%	16.03%
1.E-06	35	33	22	16	106	33.02%	31.13%	20.75%	15.09%
1.E-07	32	32	22	14	100	32.00%	32.00%	22.00%	14.00%
5.E-08	31	32	22	14	99	31.31%	32.32%	22.22%	14.14%

Table 4-6. Number and percentage of SNPs under different p-value thresholds and I^2 categories for HDL-C in the meta-analysis.

Finally, we conducted functional annotations for the 201 unique genome-wide significant loci identified in both population-specific analyses and meta-analyses. We first annotated the potential impact of variants on gene transcript and identified five missense, one inframe deletion, and one missense/splice region variant. (Table 7) Looking at the loss-of-function (LoF) score(28), we found 24 LoF-intolerant variants with a score > 0.9. We further checked the clinical significance using the ClinVar submitted records and identified one variant with at least one pathogenic record reported. To predict deleteriousness, we looked at SIFT(29), scaled CADD scores(30), and PolyPhen(29), which resulted in two deleterious variants (SIFT < 0.05), 79 deleterious variants (CADD > 15), and one probably deleterious variant (PolyPhen = 1). Investigating regulatory regions also revealed five variants located in transcription factor binding sites. When we focused on the novel GWAS loci, rs72552763, the variant displayed a relatively high level of deleteriousness, as it was predicted to be an inframe deletion in *SLC22A1* with an LoF score of 92 and a scaled CADD score of 11.48. Within the 444 GLGC-reported lipids-associated loci, five loci have been implicated in *SLC22A1*.

3_prime_UTR_variant	11
5_prime_UTR_variant	1
downstream_gene_variant	17
inframe_deletion	1
intergenic_variant	35
intron_variant	81
intron_variant,non_coding_transcript_variant	4
missense_variant	5
missense_variant,splice_region_variant	1
non_coding_transcript_exon_variant	1
regulatory_region_variant	8
splice_region_variant,intron_variant	2
splice_region_variant,non_coding_transcript_exon_variant	1
synonymous_variant	1
TF_binding_site_variant	5
upstream_gene_variant	27

Table 4-7. Functional annotation of significant loci identified in both population-specific GWASes and meta-analysis for all blood lipid phenotypes.

Discussion

In this study, we combined clinical and genetic information from UCLA ATLAS EHR and genotyping data to investigate the genetic architecture of five lipid phenotypes in over 26,000 patients. As the patients were from diverse ancestry backgrounds, we employed a two-stage meta-analysis, where we first conducted population-specific GWAS analysis for five different ancestry groups, followed by meta-analyzing them in one combined set. This strategy enabled us to identify 195 GWAS loci through the meta-analysis with one novel locus. Our population-specific analysis revealed six more GWAS hits that displayed potential population-specific effects. Our systematic evaluation of the 444 previously reported lipid-associated loci by GLGC in the UCLA ATLAS dataset demonstrated a shared genetic background across ancestry groups while also observing heterogeneity of effect sizes. We also compared our results to large external EHR-based GWAS results using the UK Biobank dataset, where a majority of our signals were replicated. We demonstrated that the effect sizes estimated from these two datasets were highly correlated as more significant variants were selected. To check the amount of heterogeneity across different lipid phenotypes, we explored the I^2 statistic in the meta-analysis and observed a wide range of heterogeneity, with TG being the most heterogeneous trait. Our functional annotation of the novel variant rs72552763 in *SLC22A1* revealed relatively high deleteriousness and loss-of-function intolerance, while other variants have been implicated in the same gene.

Our findings have provided three insights. Firstly, we have demonstrated the efficacy and usefulness of large-scale EHR in combination with genetic information to investigate the genetic basis of human complex traits. In this current study, we leveraged the UCLA ATLAS EHR and extracted all available laboratory measurements of lipid phenotypes. We retained the earliest measurements for each individual with multiple records, which resulted in a collection of over `00,000 patients ranging over ten years. Our subsequent GWAS of over 26,000 patients with genotyping data revealed many known GWAS loci to nearly 200, demonstrating the potential of combining EHR and genetic data in identifying genetic variants associated with complex traits. Furthermore, our findings suggested a novel locus, rs72552763, predicted to be an inframe deletion mutation in *SLC22A1*. Five independent GWAS loci have been reported in this gene by GLGC. This gene encodes an organic cationic transporter OCT1(31-33) and is primarily expressed in the liver as well as lung, kidney, and adrenal gland to a lesser degree. It functions to mediate the uptake of various organic cations and transport many commonly used drugs. For example, various polymorphisms of *SLC22A1* have been extensively studied in the clinical pharmacology of metformin exposure and responses for the treatment of type 2 diabetes. Previous in vivo and in vitro studies have also indicated the deletion of OCT1 would result in a disruption of the hepatic glucose-fatty acid cycle and, therefore, elevated total body adiposity with increased systemic glucose and lipids levels.(34) Many of the previously identified GWAS loci in this gene displayed reduced function in thiamine uptake (e.g., rs12208357) or lower expression (e.g., rs683369). In our analysis, the variant rs72552763 was predicted to be an inframe deletion with high CADD and LoF scores, suggesting that it could potentially alter the substrate-uptake ability of the encoded protein OCT1 and thus disrupt the lipids metabolism and peripheral energy homeostasis. Further functional studies and in vitro/in vivo experiments of *SLC22A1*, and in particular this variant, will be needed to elucidate its specific function and mechanism in lipids metabolism and assess the therapeutic and clinical importance to target this region or even provide personalized treatment for carriers of this variant.

Secondly, our study was built upon the UCLA ATLAS dataset with diverse ancestry backgrounds. This was enabled by the fact that the UCLA Hospital is located in one of the most ancestrally diverse regions and is able to obtain EHR data from patients with a wide range of backgrounds. As a result, we were able to estimate population-specific allelic effects, which helped to refine and provide accurate estimations for previously identified loci. Comparing between the effect sizes estimated in our study and the GLGC study., we observed a greater correlation in general among European samples and attenuated correlations among other ancestry groups, as the GLGC samples were composed of primarily Europeans, indicating both shared genetic effects between Europeans and other populations and various degrees of population-specific effects. Interestingly, the East Asian samples in our study also displayed a relatively high correlation with the GLGC estimates, suggesting a potentially higher degree of shared genetic basis among these two groups. Moreover, our population-specific GWAS revealed many loci that were not observed in the meta-analysis. For example, for HDL, we observed five loci specific to European samples that were absent in the meta-analysis; for TG and TC, we discovered one locus specific to European ancestry and one to South Asian ancestry, respectively. These findings suggested that the genetic effects of the same variant could vary greatly across different ancestry groups and, given a sufficiently large sample size, population-specific GWAS would reveal many significant loci that would otherwise obscure in a mixed population. To further refine the genetic estimates for specific ancestry groups and better explain the variation of blood lipid levels between different populations, we will still need larger sample sizes for single-ancestry cohorts as well as additional ancestries that are currently under-represented in our dataset.

Thirdly, we further checked and compared our results to the GWAS results from the UK Biobank dataset. Representing one of the larest EHR-based biobank, the UK Biobank dataset was a direct comparison and assessment to the UCLA ATLAS dataset. We evaluated the effect sizes of the top variants from both datasets and observed a relatively high correlation, especially for European ancestry, as the UK Biobank dataset consisted of patients with primarily European backgrounds. For other populations, such as East Asian and South Asian, the correlations decreased rapidly when variants with larger p-values were introduced. However, for the meta-analysis, we observed relatively high correlations for all lipid traits and ancestries, potentially as most of these estimates represented shared genetic effects across populations.

In addition to these insights gained, several limitations deserved to be mentioned. First, our lipid phenotypes were based entirely on the UCLA ATLAS EHR data where multiple entries were usually available for each individual. We decided to define the lipid levels using the earliest measurements, but the possibility of the introduction of noises in lipid levels due to circumstances such as time of entry during a day, diet, admission reason, or therapeutic status remains for participants entering the UCLA healthcare system. Second, the total sample size as well as the sample sizes for each ancestry group were limited, which might obstruct the power to detect novel loci in general. However, this was primarily due to the limited number of genotyped samples, while the total number of samples with lipid records in the EHR database exceeded 100,000, which was about one-third of the largest lipids GWAS reported to date (~300,000 samples) and was constantly increasing, suggesting a great potential of improvements as more patients being genotyped. Third, although over 9,000 females and around 8,000 males were included in our analysis, we did not attempt to detect sex-specific genetic associations due to suspected limited power. Nonetheless, it may be of interest to check the heterogeneity of genetic effects between sexes, as previous studies have demonstrated such differences among samples with European ancestry. Fourth, further functional studies will be required to gain a deeper understanding of the effect and underlying biological mechanism of the novel loci identified in this study.

In summary, in this study, we explored the population-specific and shared genetic effects on five lipid phenotypes among around 17,000 patients with existing EHR and genotyping data and identified one novel locus through meta-analysis. This result demonstrated the enormous potential of combining EHR and genetic data in the discovery of novel genetic associations for complex human traits. Comparing to previously reported results by GLGC and UK Biobank, we observed various degrees of differences in the effect sizes of the associated variants across ancestral groups, further suggesting the importance of refining our understanding of genetic effects for specific ancestral groups. The EHR database thus provided a rich resource for such ancestry-specific analysis and offered the platform for the development of novel therapeutic and clinical targets as well as personalized therapy.

Methods

Study sample and EHR-based lipid phenotypes

The UCLA ATLAS Precision Health Biobank and GenomicsDB is a central repository of de-identified records of patient Electronic Health Record (EHR) and genomic/genetic data from consented patients.(35) The database contained health records for a wide range of complex traits, laboratory measurements, and related information for patients from diverse ancestry backgrounds. We collected laboratory measurements for four phenotypes, HDL-C, LDL-C, TG, and TC. The LDL-C measurements were made from two different methods, an indirect calculation based on other three lipid traits (which is the more common method) (LDL Calc) and direct quantification (LDL Quant). Therefore, we extracted both records which resulted in a total of five different lipid traits. As multiple entries were available for each trait and each individual, we retained the first non-missing measurement available, which was the earliest record based on the timestamp associated with the laboratory measurement without a missing code. The range of the records was between 2005 and 2020. A detailed summary of the demographic information of the samples included in this study can be found in Table 2.

Data processing and quality control

Individual-level quality control

We conducted stringent quality control (QC) on the 26,414 genotyped samples to ensure we included only high-quality samples. We first removed known contaminated samples and performed sex-check using PLINK. Among the remaining samples, we then computed sample-level missing rate, estimated the theoretical relatedness using KING(36), and performed principal component analysis (PCA) using fastPCA. For sample-level missing rate, we set the cutoff at 0.05 and removed all individuals with higher missing rates. To avoid unexpected relatedness, we conducted identify-by-descent (IBD) analysis and removed one in each pair of predicted-to-be duplicated or first/second-degree related samples. For PCA, we used 1000 Genomes (1KG) phase 3(37) as a reference panel to determine the ancestry backgrounds within our dataset. Specifically, we included only common independent variants MAF > 15% that were shared between our dataset and 1KG and used the distribution of 1KG samples to assign the ancestry of our samples. PCA plots of all samples could be found in Fig 7. We also performed sex-check and observed no discrepancies between the predicted sexes and those reported in EHR. Lastly, we merged the genotyped data with EHR and retained only samples with at least one measurement for each lipid trait of interest. In summary, this resulted in 17,226, 16,948, 4,599, 17,429, and 17,377 high-quality samples for HDL-C, LDL Calc, LDL Quant, TG, and TC, respectively. A detailed count decomposition into different ancestry groups could be found along with the demographic information listed in Table 2.

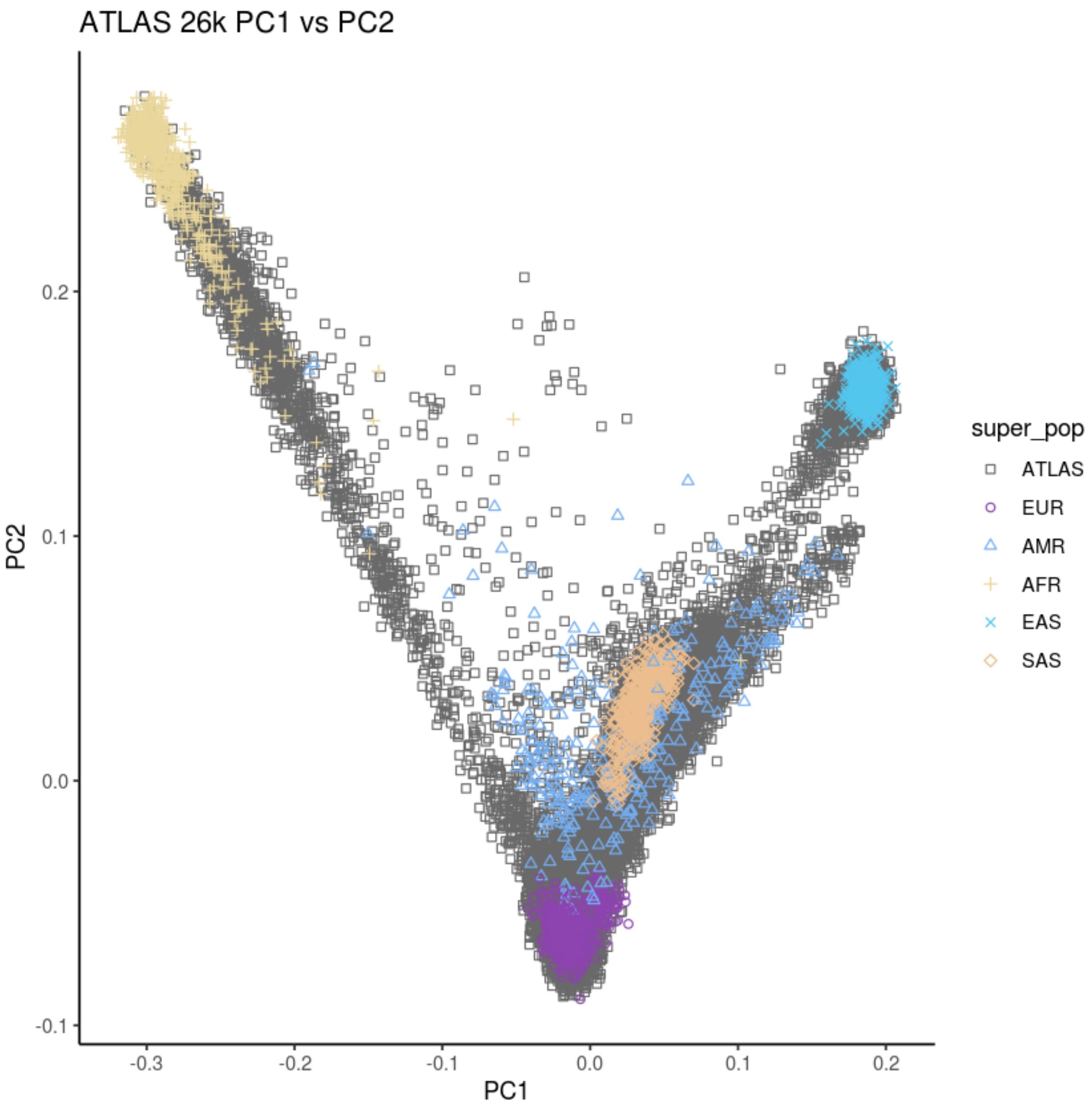


Figure 4-8. PCA plot (PC1 v.s. PC2) of the UCLA ATLAS genotyped samples.

Variant-level quality control

We conducted stringent variant-level quality control to ensure only high-quality SNVs were included in this study. We included only samples that passed the aforementioned sample-level QC steps before separating them into ancestry-matched groups. We first removed unmapped SNVs assigned to chromosome zero in our dataset, followed by removing monomorphic variants to retain only bi-allelic SNVs for all analyses in this study. Furthermore, low-quality SNVs meeting the criteria, including genotyping missing rates > 5%, Hardy-Weinberg Equilibrium (HWE) p-values < 1e-12, and ambiguous flip were also excluded. For each individual population, we then computed the ancestry-specific minor allele frequency (MAF) and kept only common variants with MAF > 5% for GWA analyses. No imputation has been performed on this dataset. In summary, this resulted in 343,866, 343, 539, 344,332, 343,488, and 343,604 high-quality SNVs for HDL-C, LDL Calc, LDL Quant, TG, and TC in the meta-analysis, respectively. The specific number of SNVs for different ancestry groups and lipid traits could be found in Table 5.

Genome-wide association analysis

The association analysis was composed of two stages where we first performed the population-specific analysis, then followed by meta-analyzing overall available ancestry groups. The stage 1 association test was performed using the --linear function in PLINK(38) for each lipid phenotypes and ancestries, with adjustment for age, sex, body mass index (BMI), and the first ten principal components (PCs), which accounted for potential confounders such as population stratification and sex differences. The estimated results were then combined in the stage 2 meta-analysis using METASOFT(39), which employed an inverse-variance-weighted fixed-effects method. To account for population structures, we additionally estimated genetic effects using Han and Eskin's random-effects model(39) which was optimized to detect associations under heterogeneity. Specifically, we followed the suggestions given by the documentation of METASOFT by first running the meta-analysis with default options and then re-running the analysis while specifying the inflation factors for the mean effect and heterogeneity effect of the random effect statistics computed in the first round. To note, we observed slightly more significant loci when we used Han and Eskin's random effect model in consideration of population heterogeneity. However, as the computed inflation factors were less than one, no correction was necessary, and the results using Han and Eskin's random-effects model were not reported here.

Evaluation of the results using previously reported meta-analysis and EHR-based GWAS results

Our evaluation consisted of two parts using GWAS results reported by GLGC and UK Biobank, respectively. For part 1, we obtained the 444 GWAS loci reported by GLGC that were independently and significantly associated with lipid traits.(11) We then identified these variants in the UCLA ATLAS dataset and compared their effect sizes estimated in our study and reported by GLGC. Over 200 reported loci were found in the UCLA ATLAS dataset with a maximum of 255 loci identified for HDL-C in the European samples. (Table 3) For the shared loci, we then computed the correlation of the effect sizes using a two-sided test of Pearson correlation and compared the strength of correlation across different ancestry groups for each lipid phenotype. For part 2, we sought to examine the UK Biobank GWAS results, which represented one of the largest EHR-based GWASes. To compare these two EHR-based results, we selected the top variants from our study with p-value cutoffs at 1e-3, 1e-4, 1e-5, 1e-6, 1e-7, and 5e-8. For all variants under each cutoff, we compared their effect estimates between the UCLA ATALS and UK Biobank datasets by computing two-sided Pearson correlation p-values, similar to part 1. Additionally, in order to check whether the same variants could reach a similar significance level in both EHR-based datasets, we extract variants from each dataset under p-value cutoffs at 1e-3, 1e-4, 1e-5, 1e-6, 1e-7, and 5e-8. At each cutoff, this resulted in two sets of variants and we counted the number of shared variants present in both sets. In our hypothesis, both the correlation and the percentage of shared variants should increase as the p-value threshold became more significant. Finally, for the genome-wide significant loci identified in our study but absent in the UK Biobank summary statistics, we manually examined them in the GWAS Catalog(40) to pinpoint potential novel loci.

Functional annotation of the genome-wide significant loci

We annotated the GWAS loci using the Ensembl Variant Effect Predictor (VEP) web interface(41), which predicted the potential consequences of a variant to the corresponding gene transcript. Additionally, we also checked the pathogenicity and deleterious using PolyPhen-2(42), SIFT(42), CADD(29, 30), and ClinVar(43). In summary, we identified five, six, 200, and 16 variants with available PolyPhen-2, SIFT, scaled CADD score, and ClinVar reports. We further searched whether the genes associated with the variants had loss-of-function (LoF) scores(28) and identified 129 variants with available LoF scores. Lastly, we checked if any variant fell into known regulatory regions by investigating potential overlap with transcription factor binding sites.

Supporting information

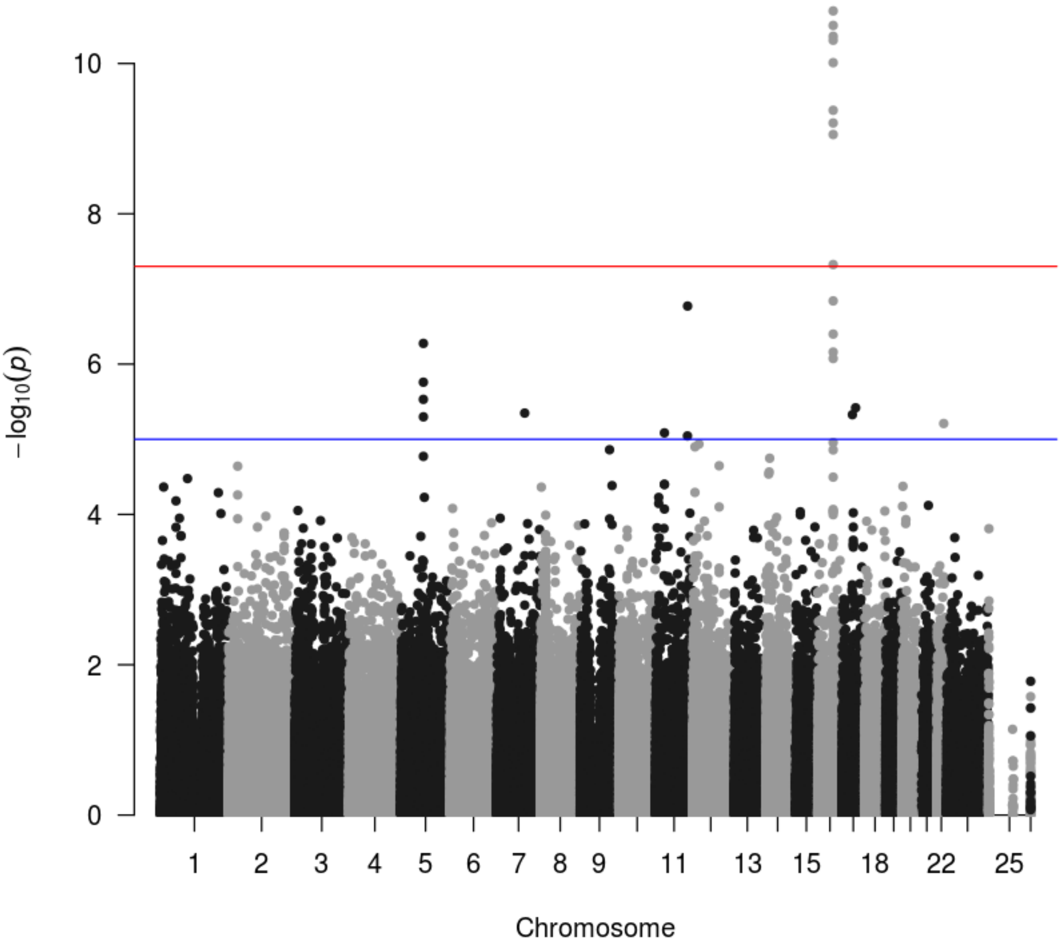


Fig S1. Manhattan plot of HDL-C in AMR population

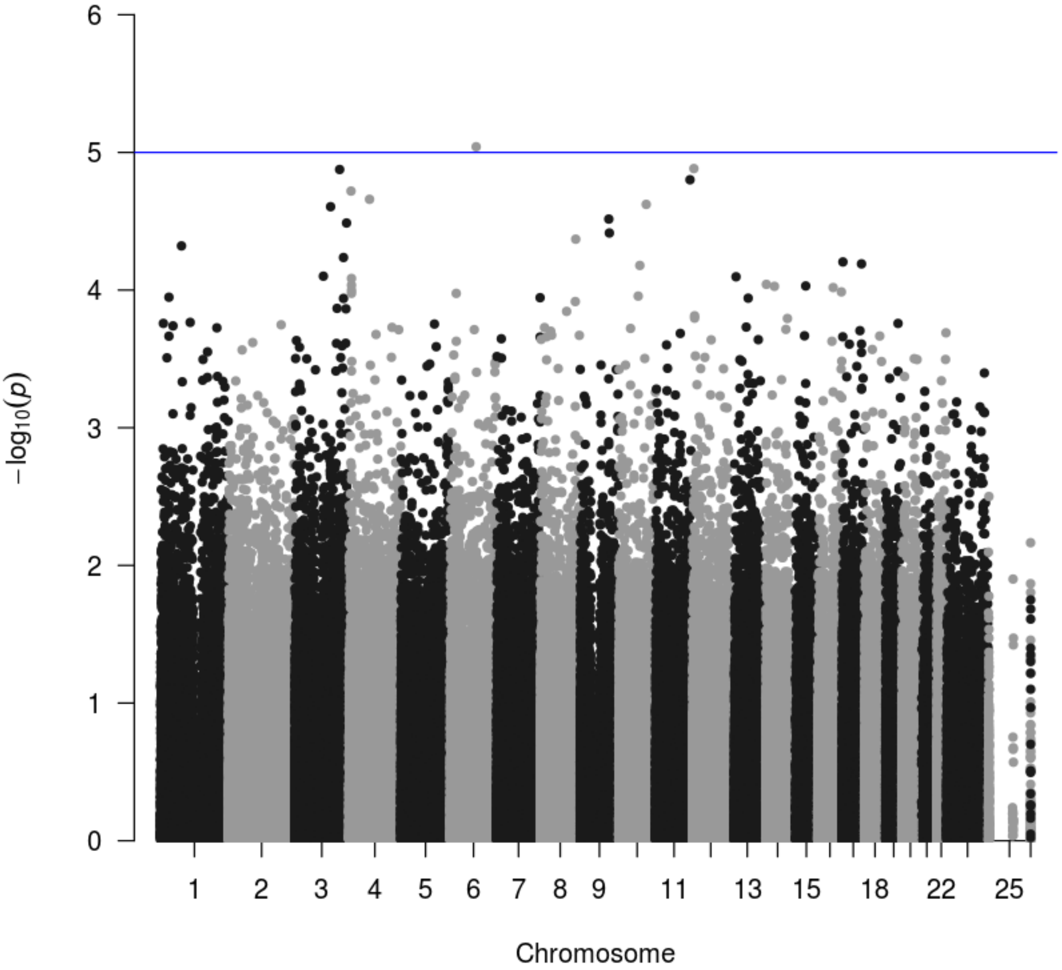


Fig S2. Manhattan plot of HDL-C in AFR population

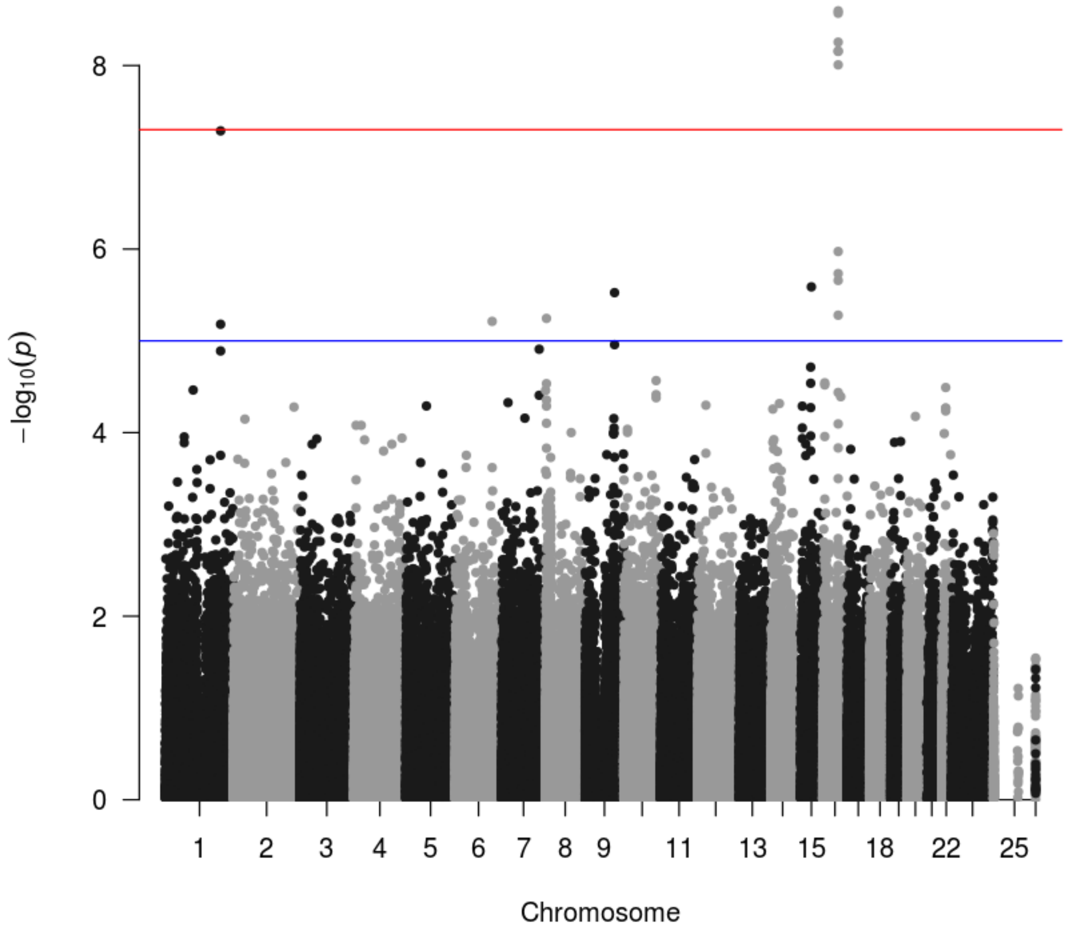


Fig S3. Manhattan plot of HDL-C in EAS population

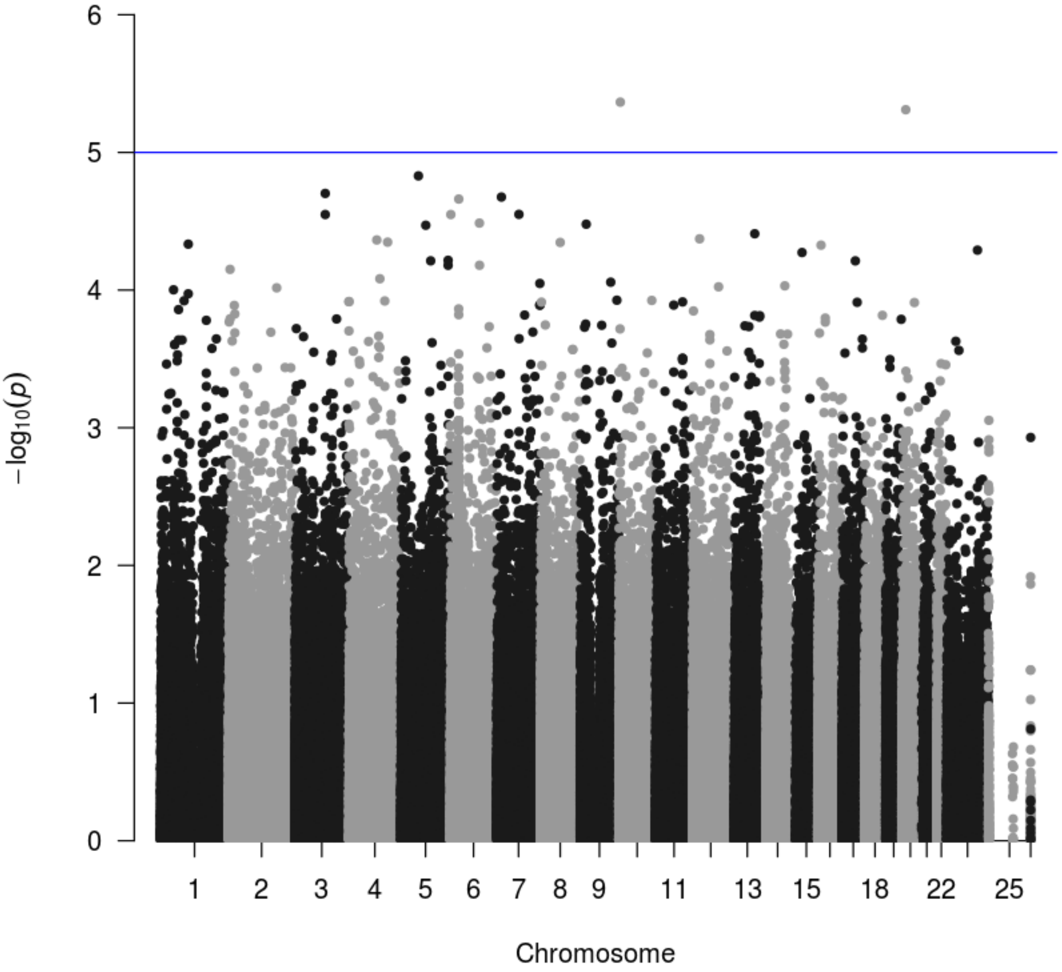


Fig S4. Manhattan plot of HDL-C in SAS population

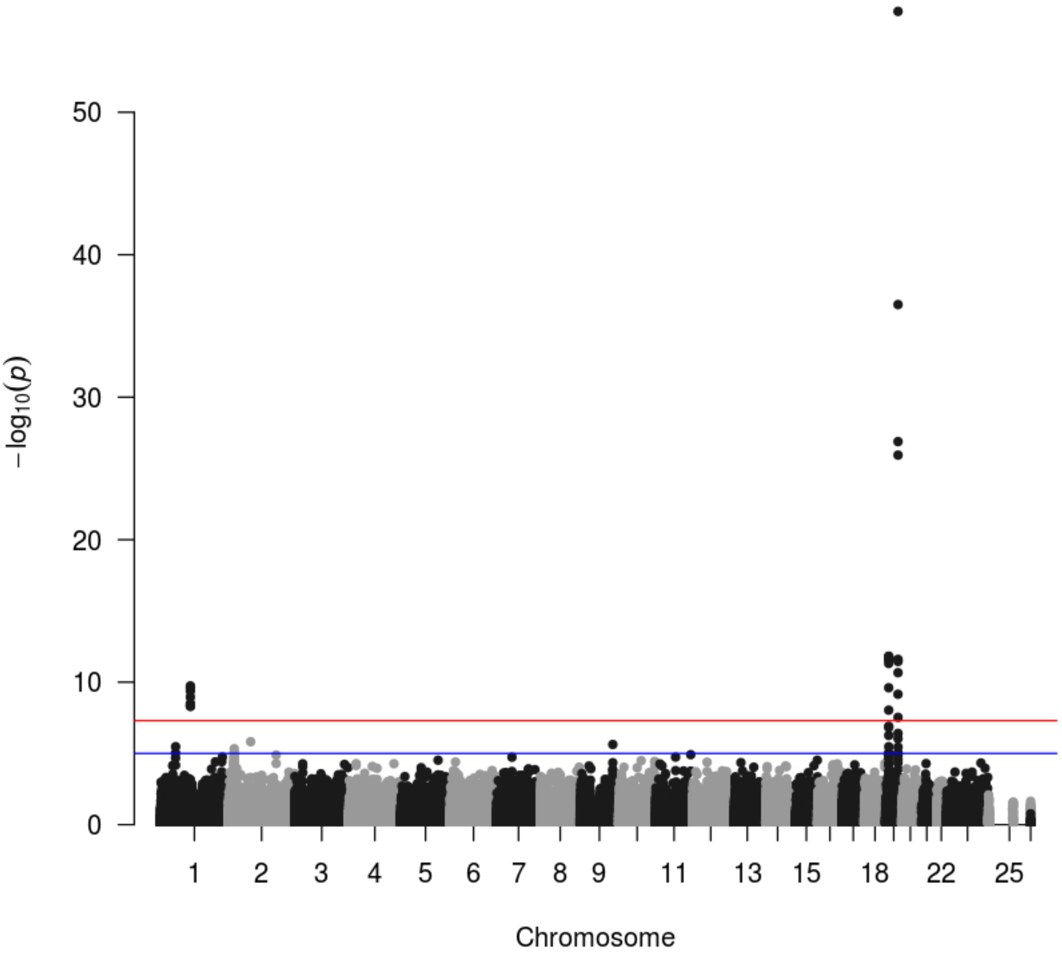


Fig S5. Manhattan plot of LDL Calc in EUR population

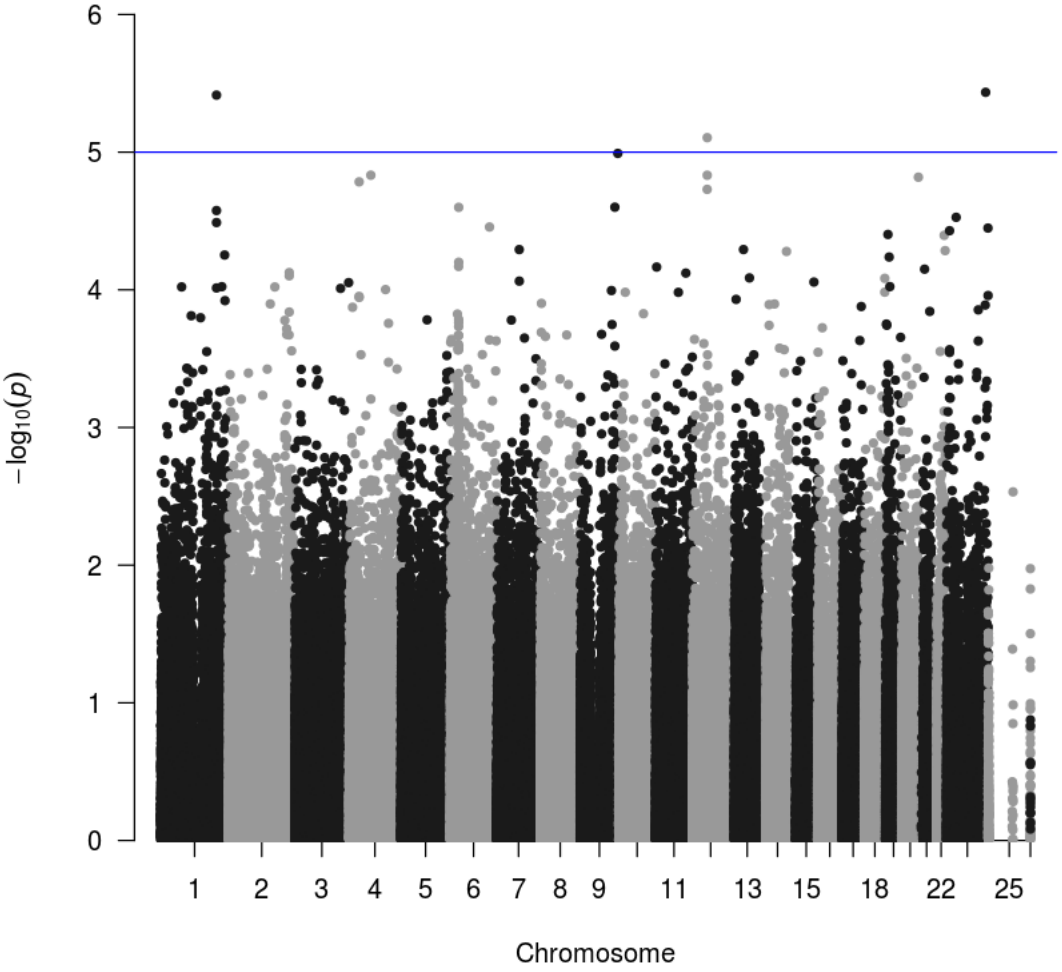

6
5
4
3
2
1
0
−log10(p)
1
2
3
4
5
6
7
8
9
11
13
15
18
22
25
Chromosome

Fig S6. Manhattan plot of LDL Calc in AMR population

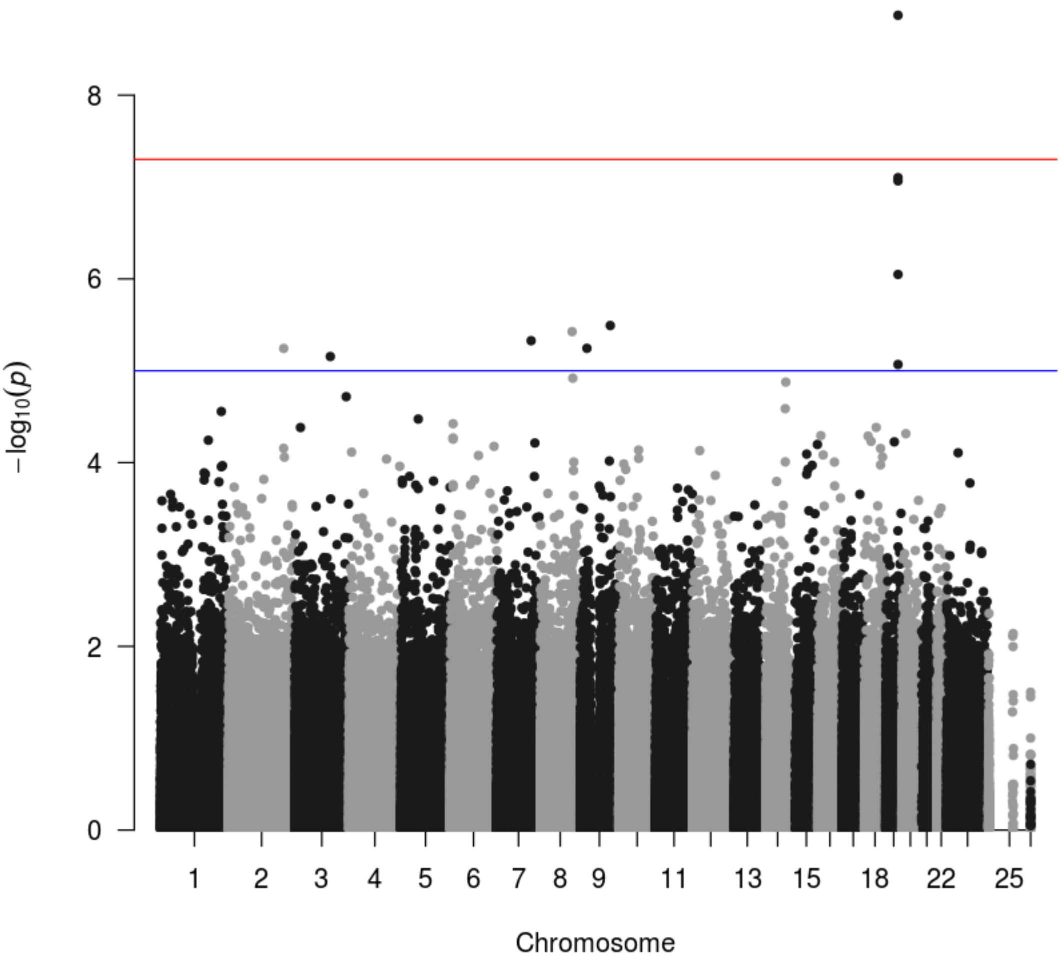


Fig S7. Manhattan plot of LDL Calc in AFR population

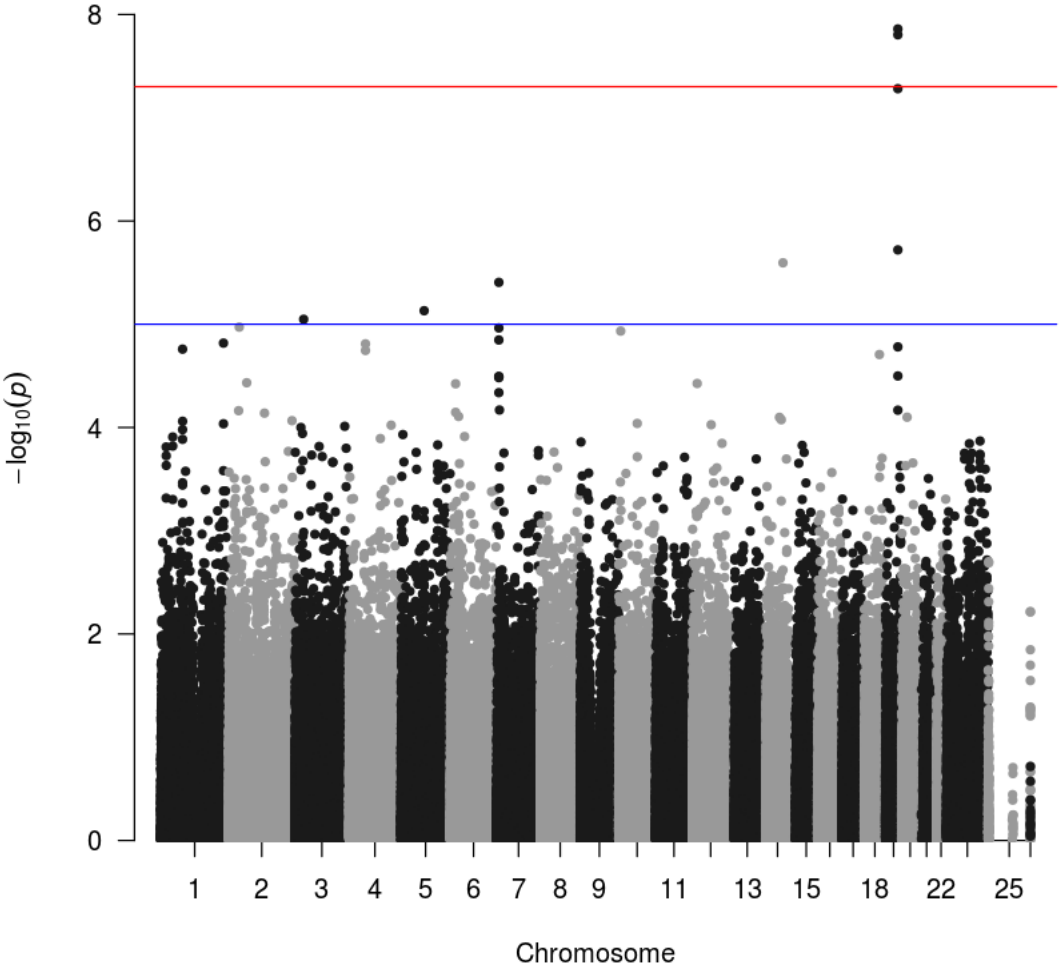


Fig S8. Manhattan plot of LDL Calc in EAS population

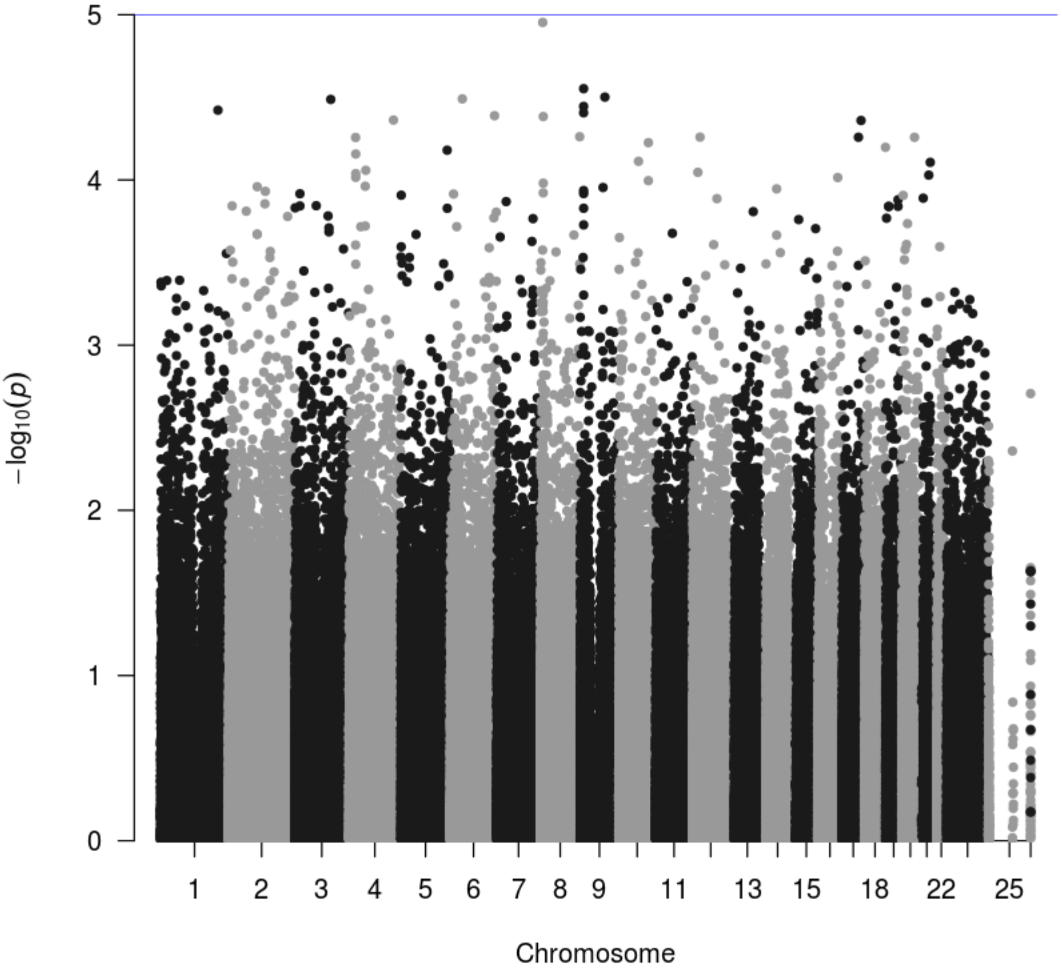


Fig S9. Manhattan plot of LDL Calc in SAS population

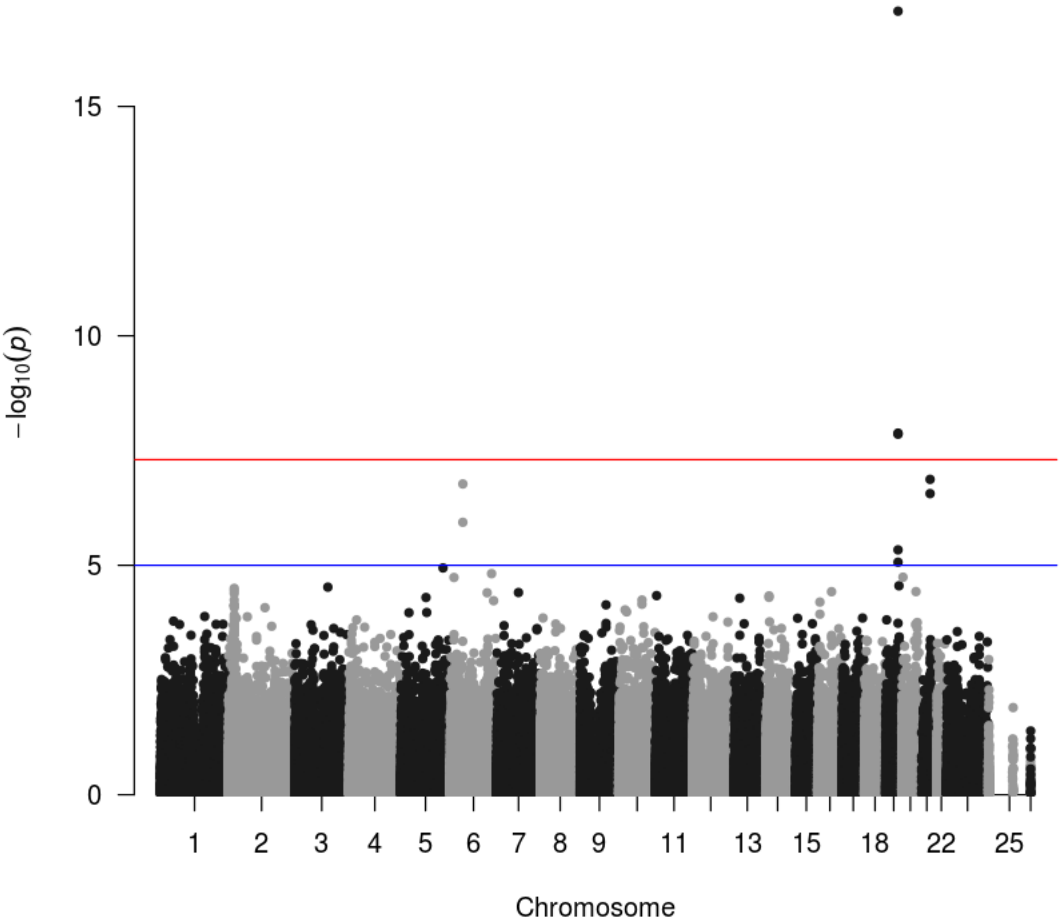


Fig S10. Manhattan plot of LDL Quant in EUR population

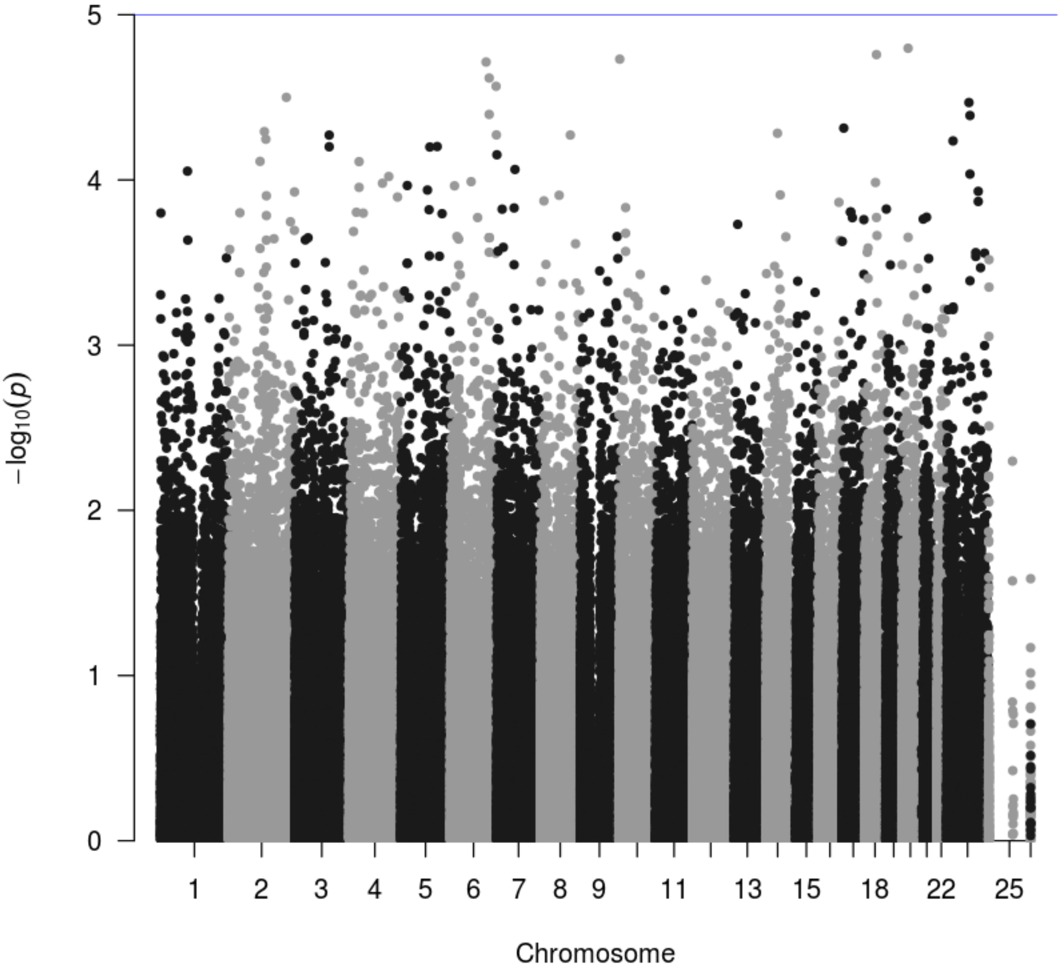


Fig S11. Manhattan plot of LDL Quant in AMR population

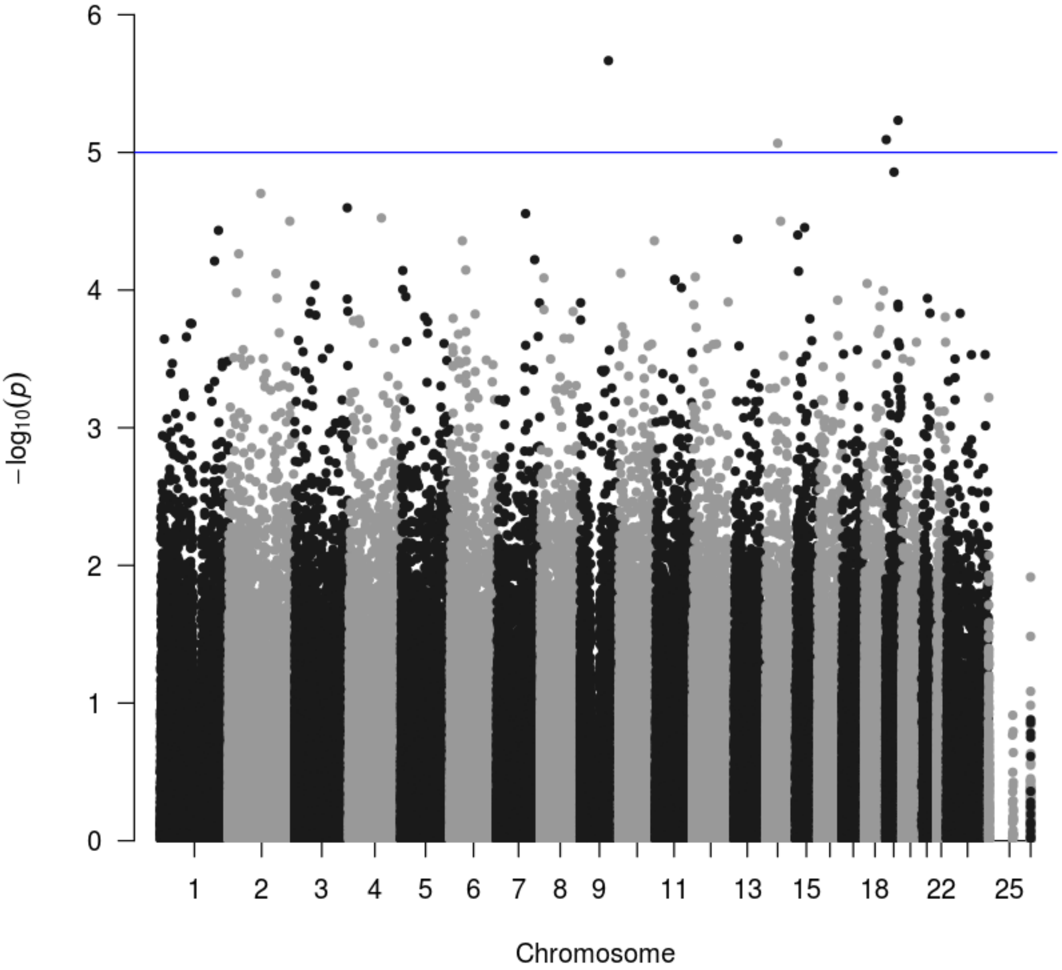


Fig S12. Manhattan plot of LDL Quant in AFR population

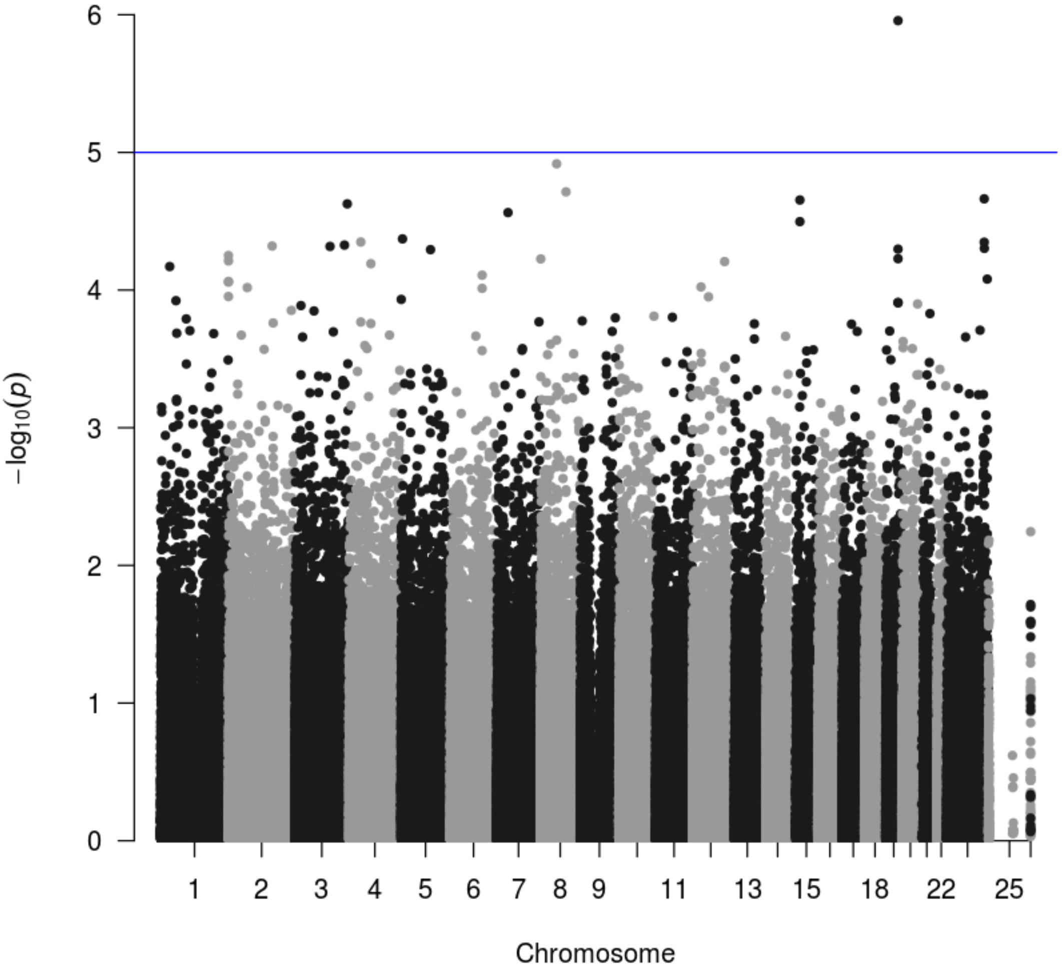


Fig S13. Manhattan plot of LDL Quant in EAS population

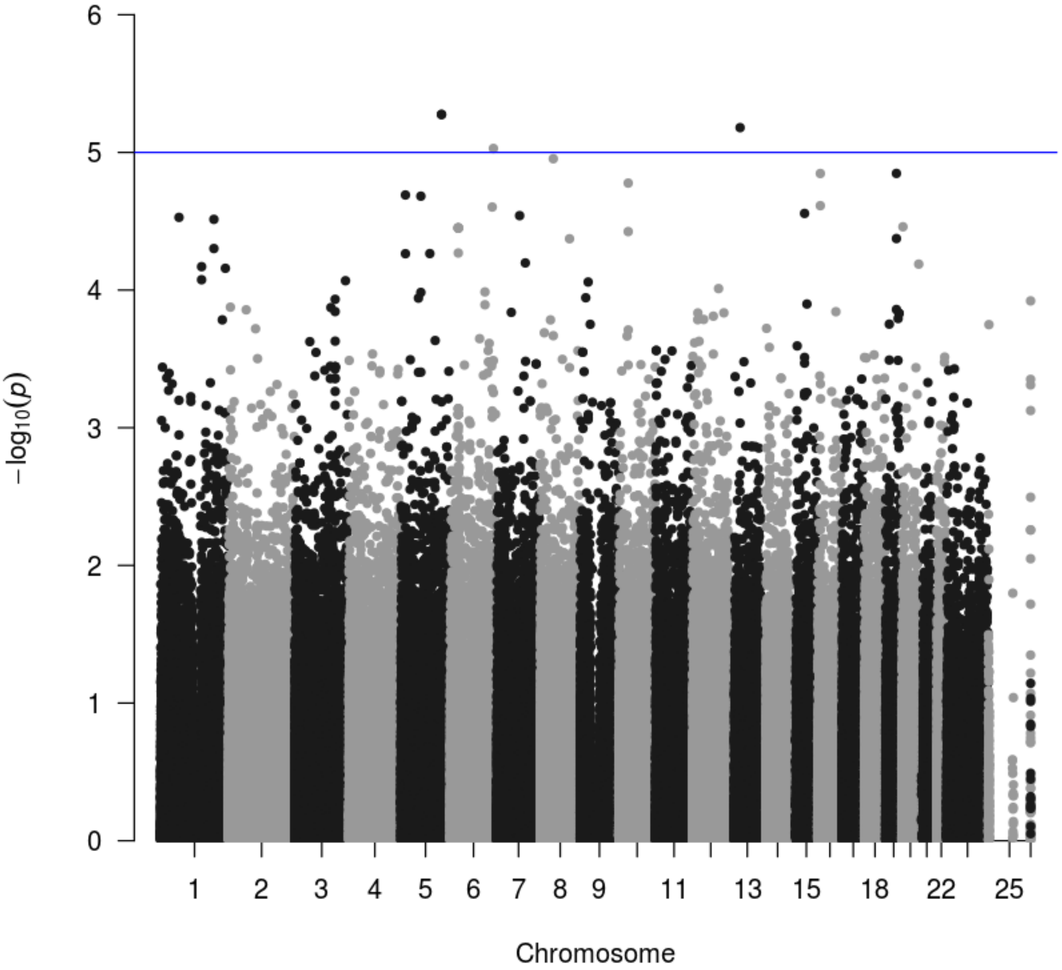


Fig S14. Manhattan plot of LDL Quant in SAS population

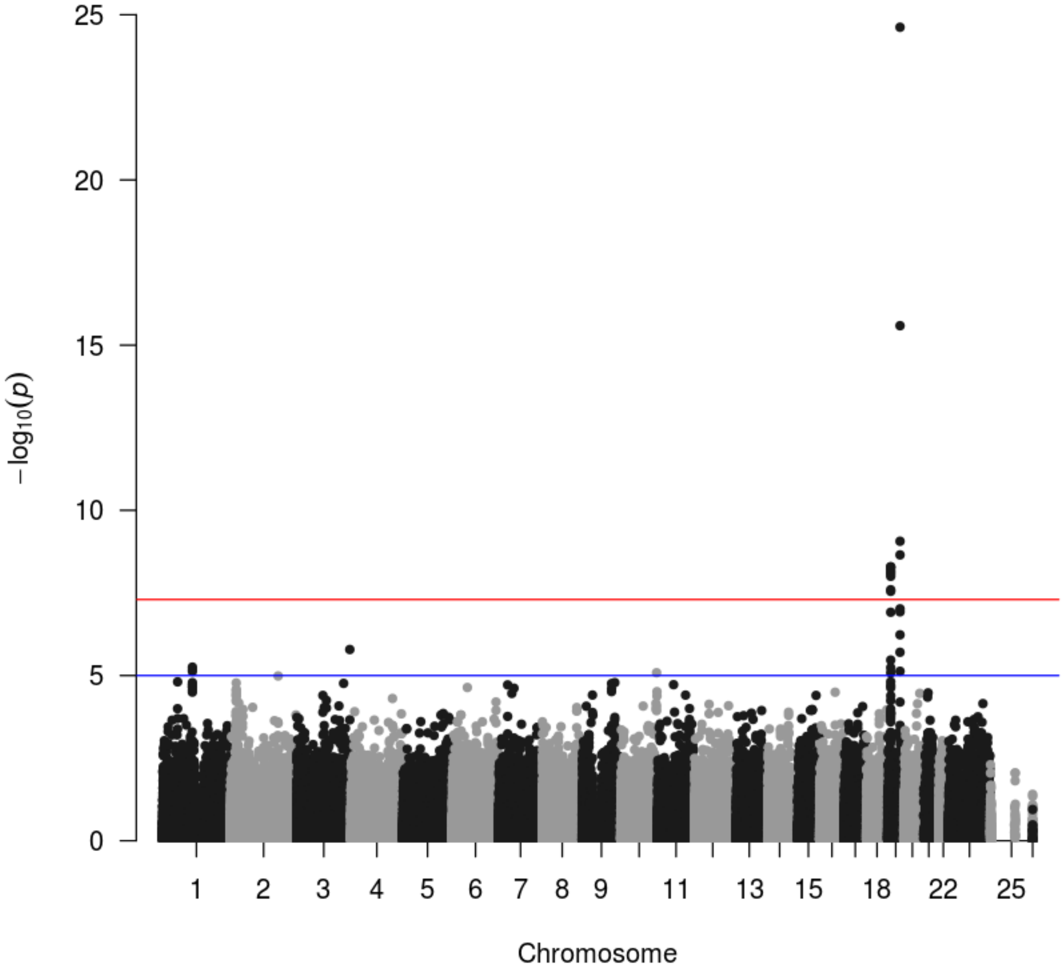


Fig S15. Manhattan plot of TC in EUR population

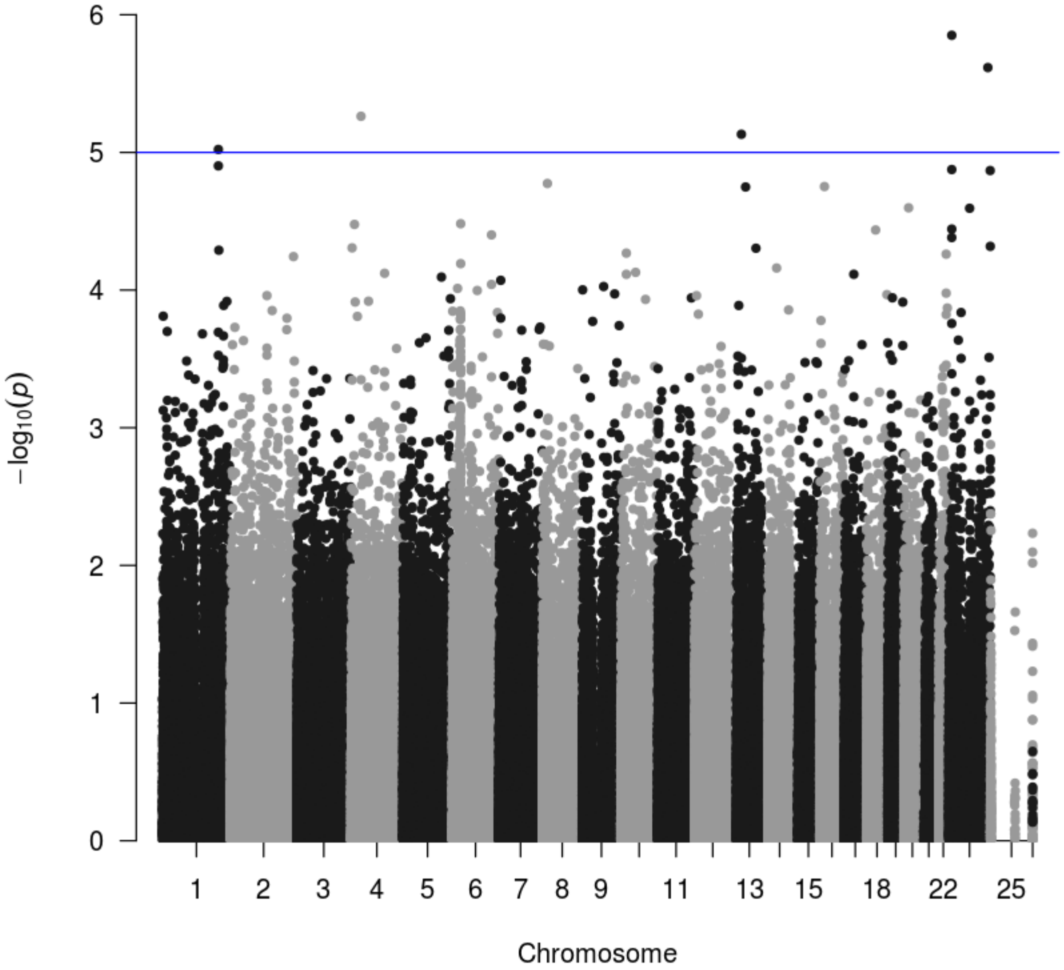


Fig S16. Manhattan plot of TC in AMR population

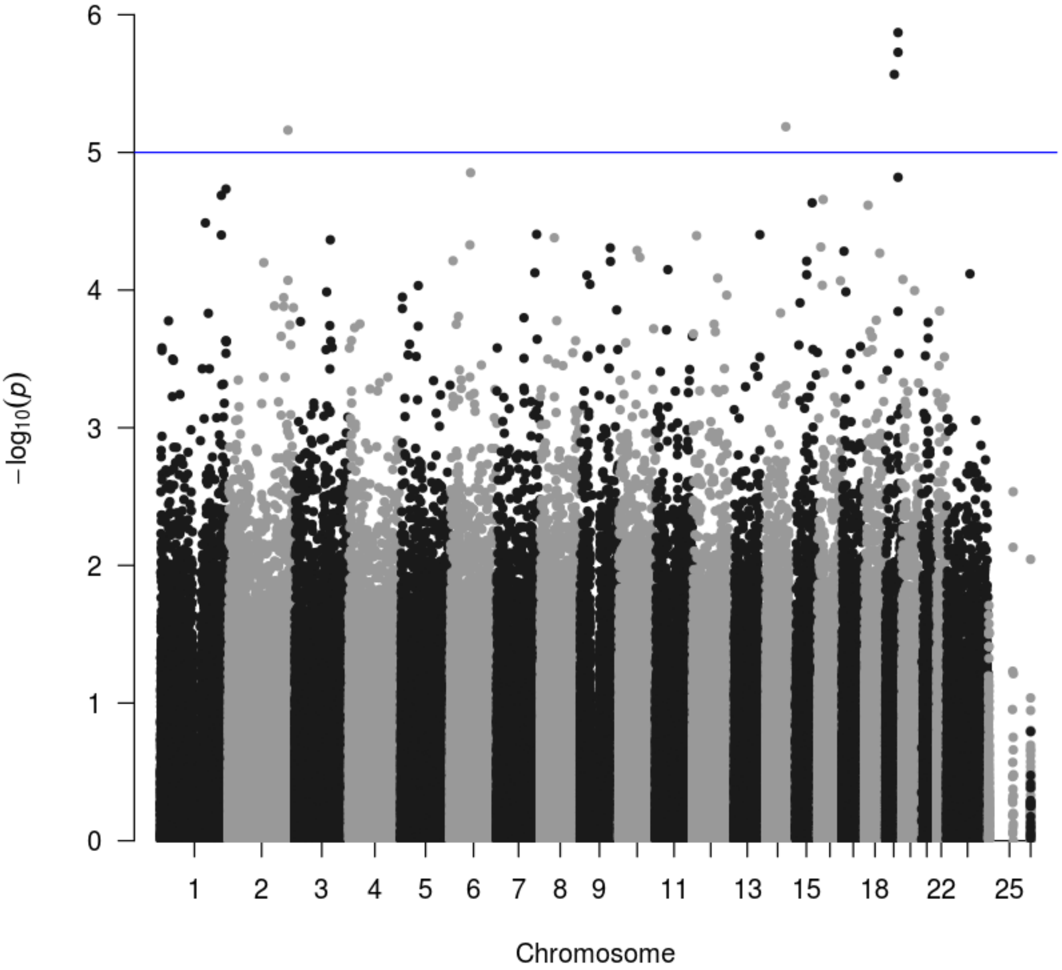


Fig S17. Manhattan plot of TC in AFR population

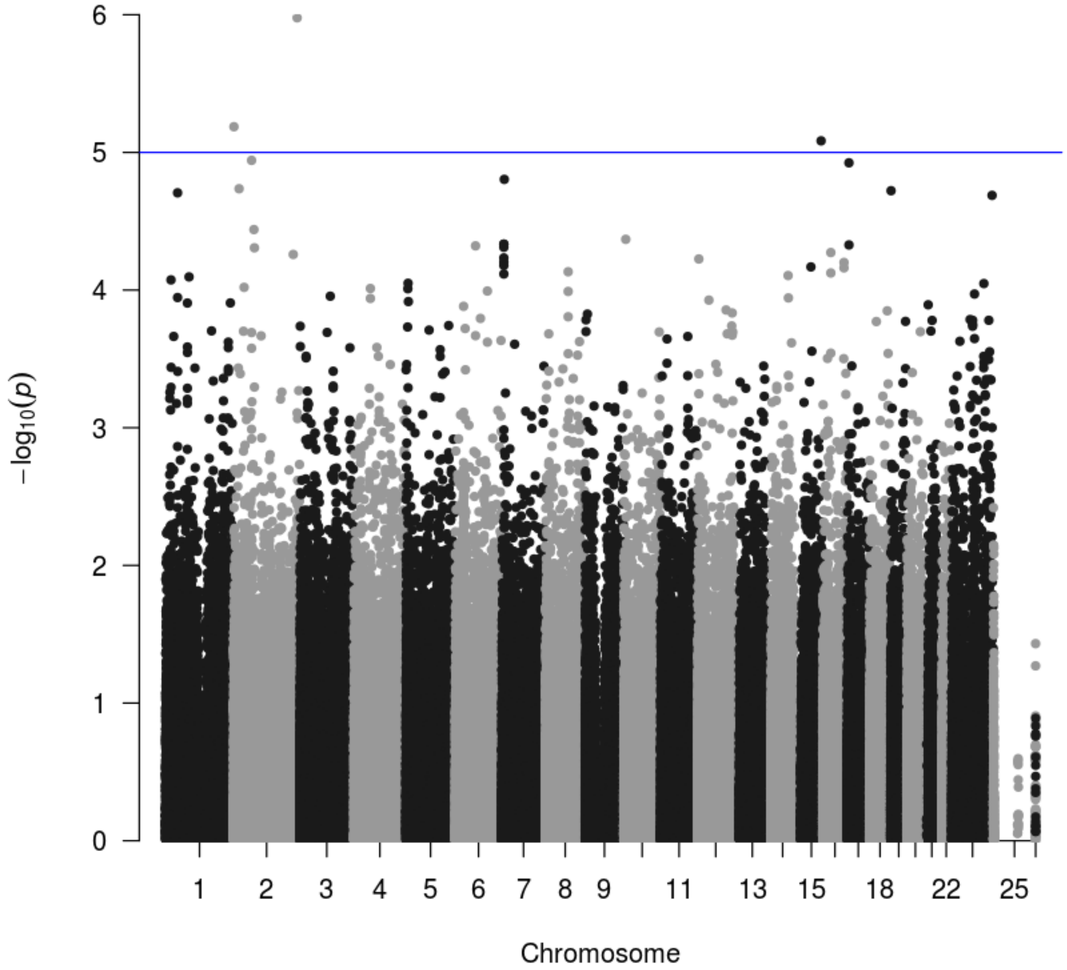


Fig S18. Manhattan plot of TC in EAS population

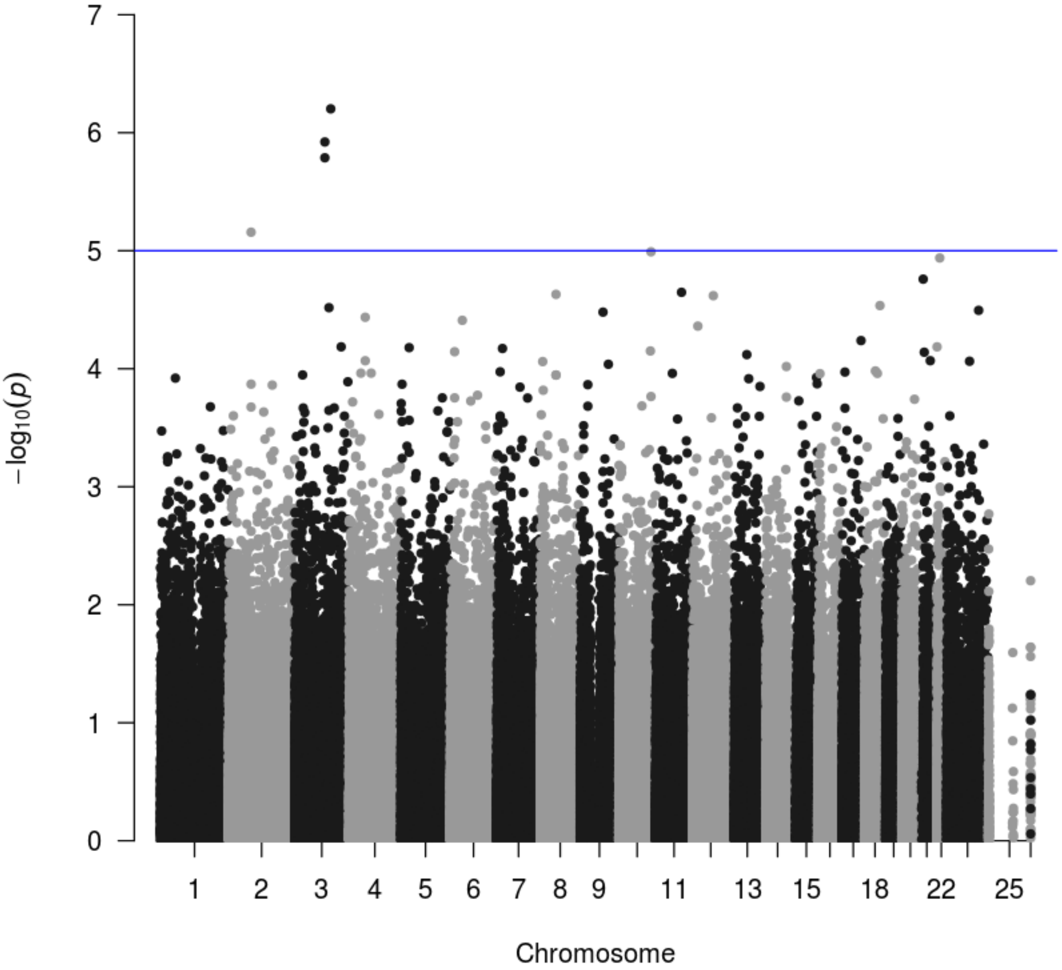


Fig S19. Manhattan plot of TC in SAS population

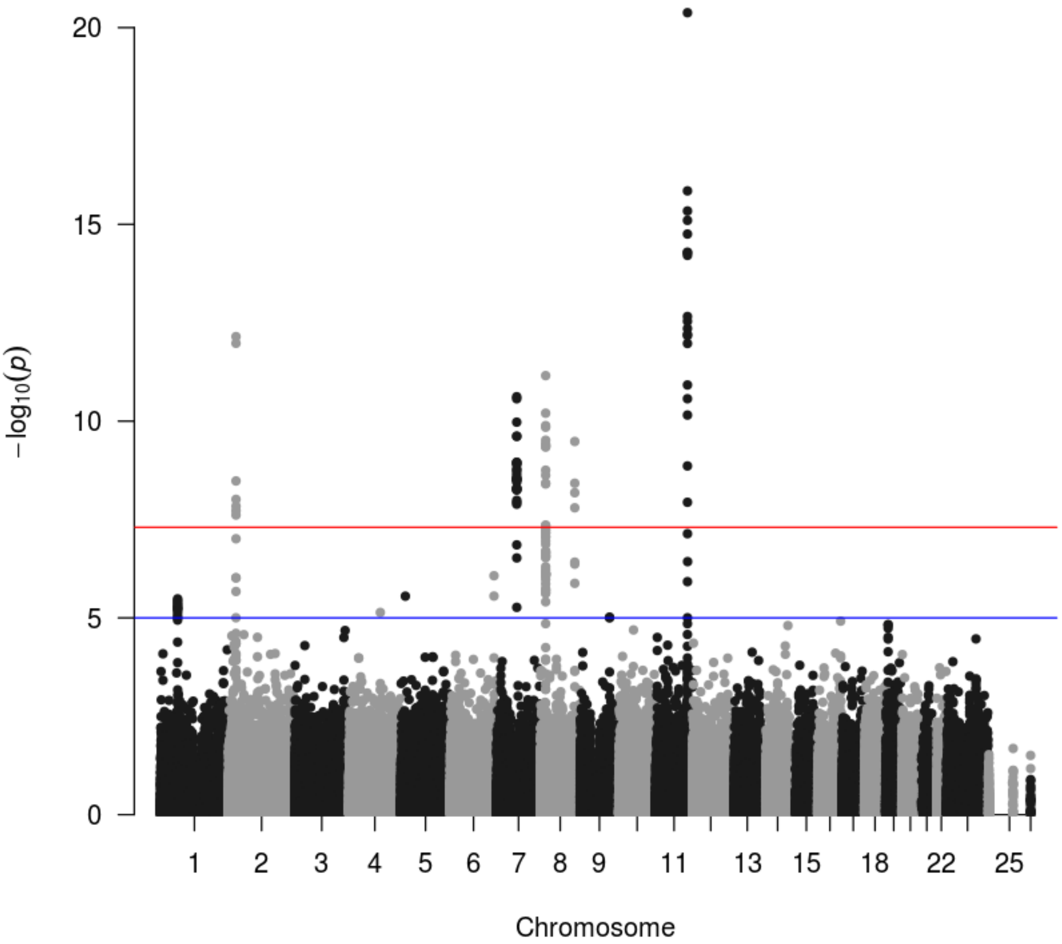


Fig S20. Manhattan plot of TG in EUR population

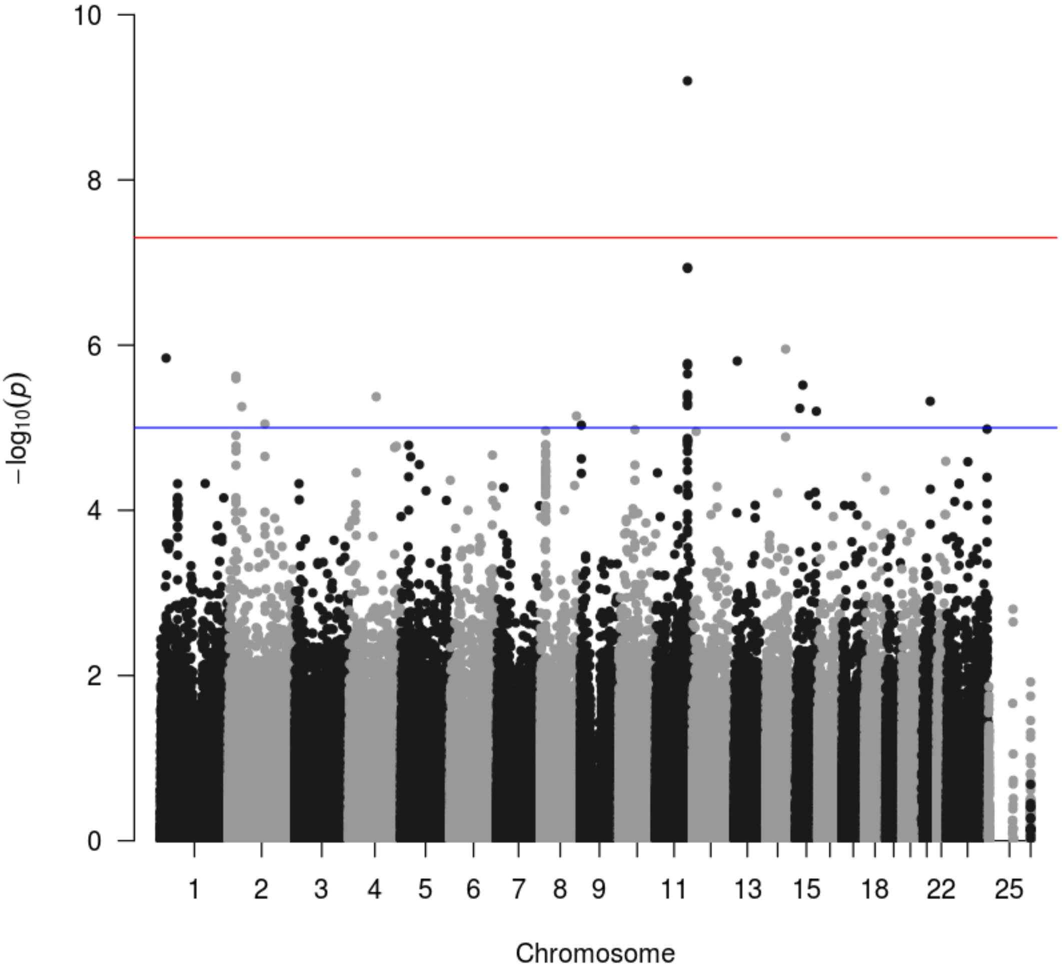


Fig S21. Manhattan plot of TG in AMR population

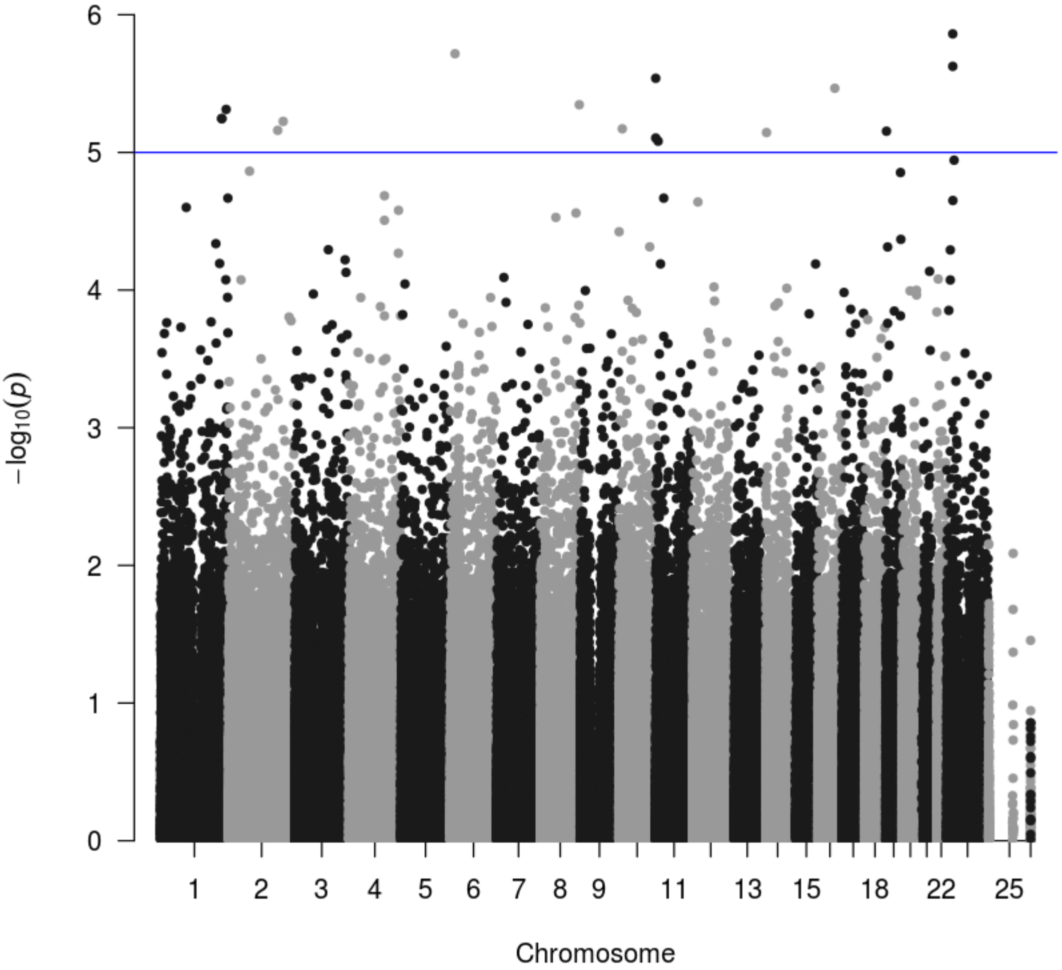


Fig S22. Manhattan plot of TG in AFR population

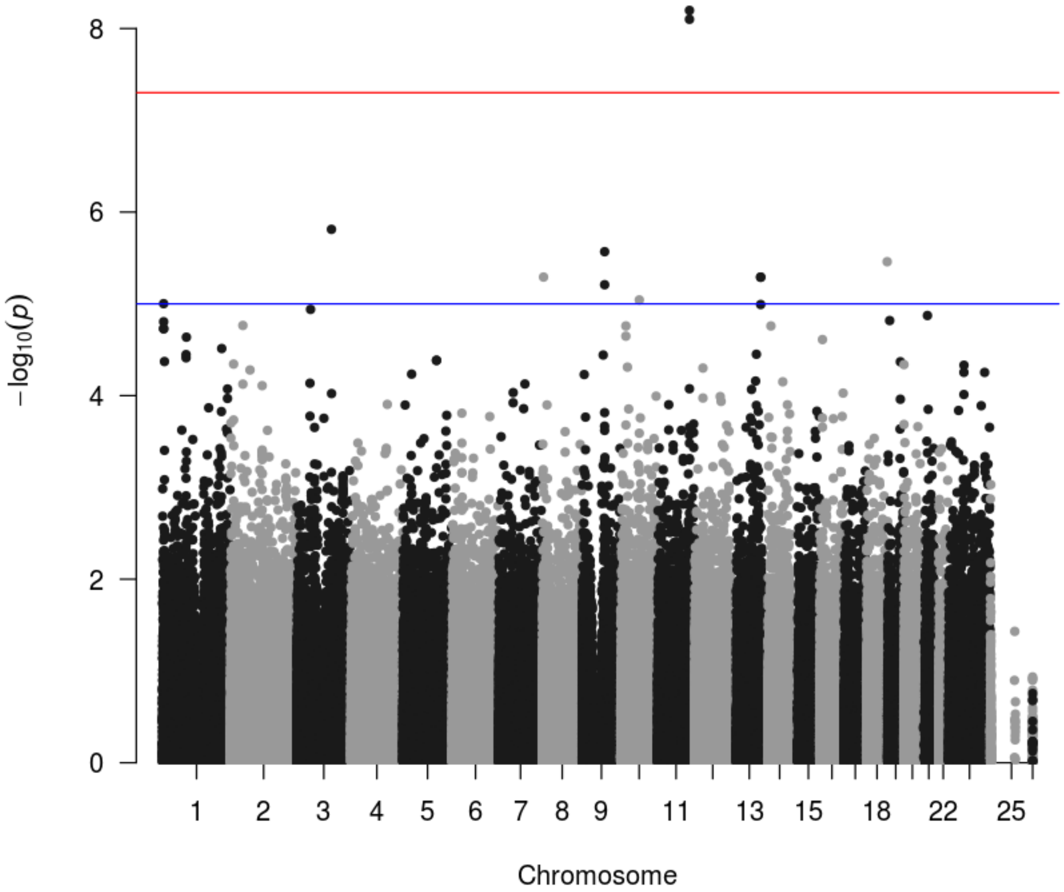


Fig S23. Manhattan plot of TG in EAS population

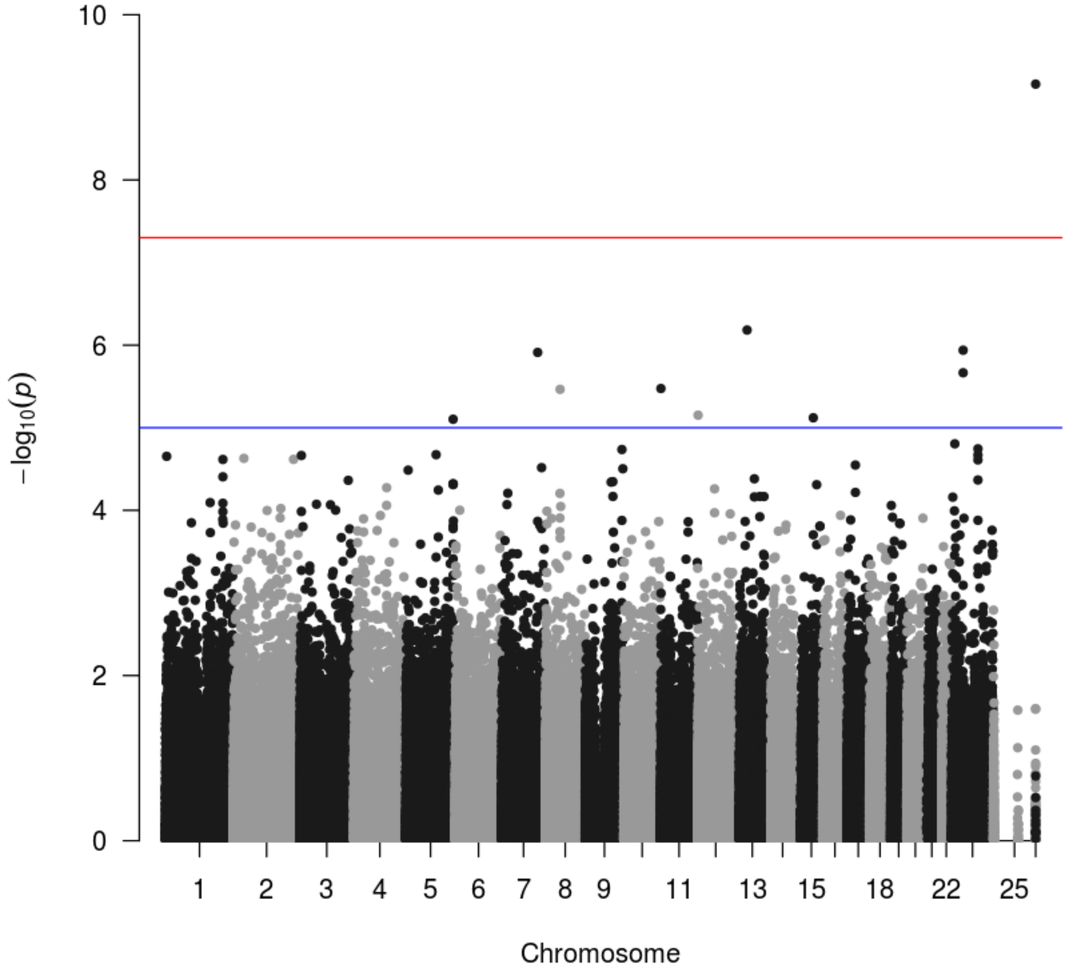


Fig S24. Manhattan plot of TG in SAS population

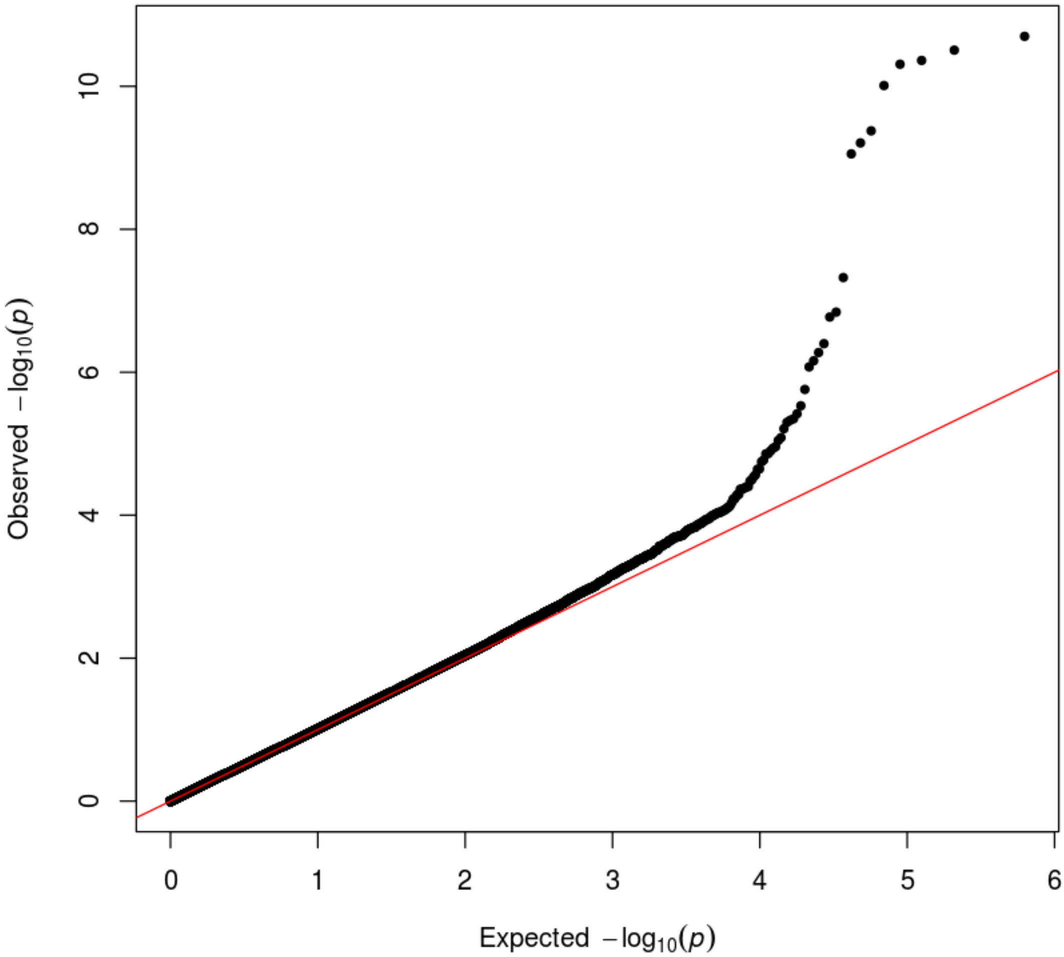


Fig S25. Q-Q plot of HDL-C in AMR populaiton

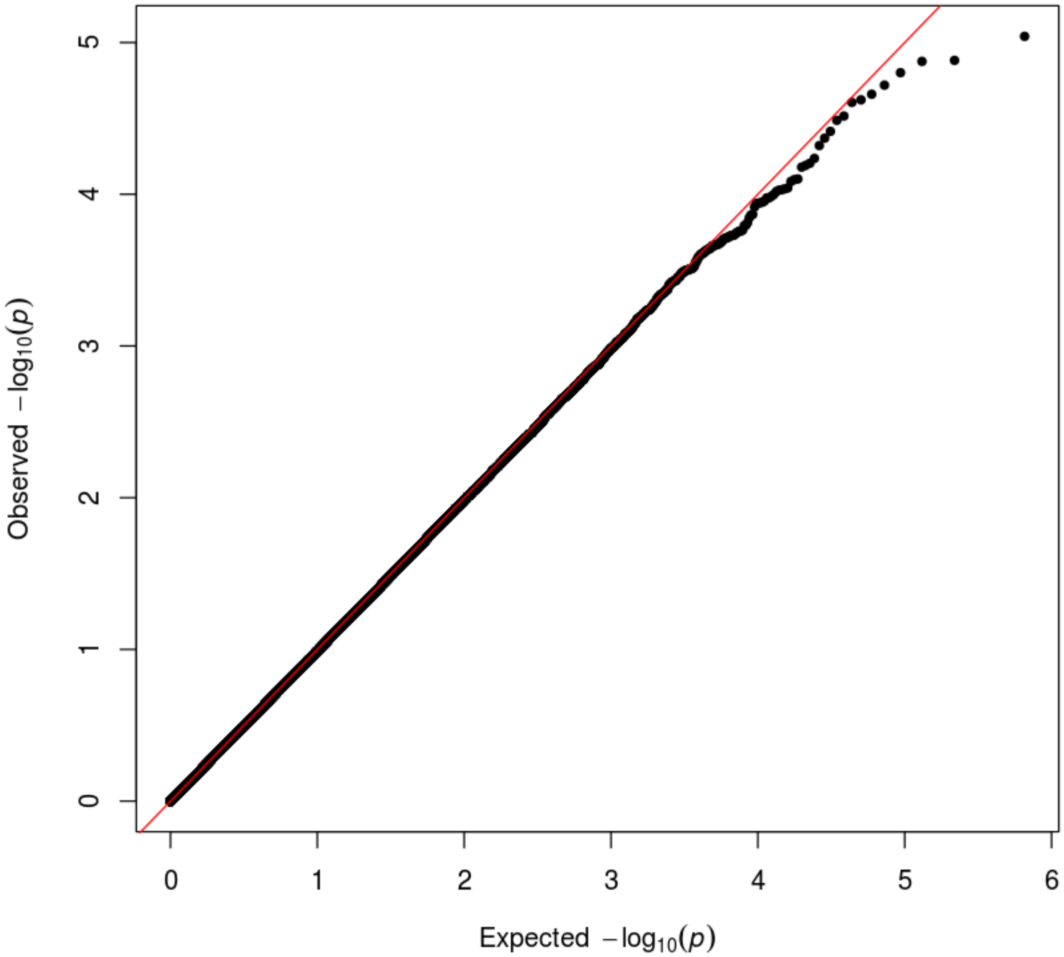


Fig S26. Q-Q plot of HDL-C in AFR populaiton

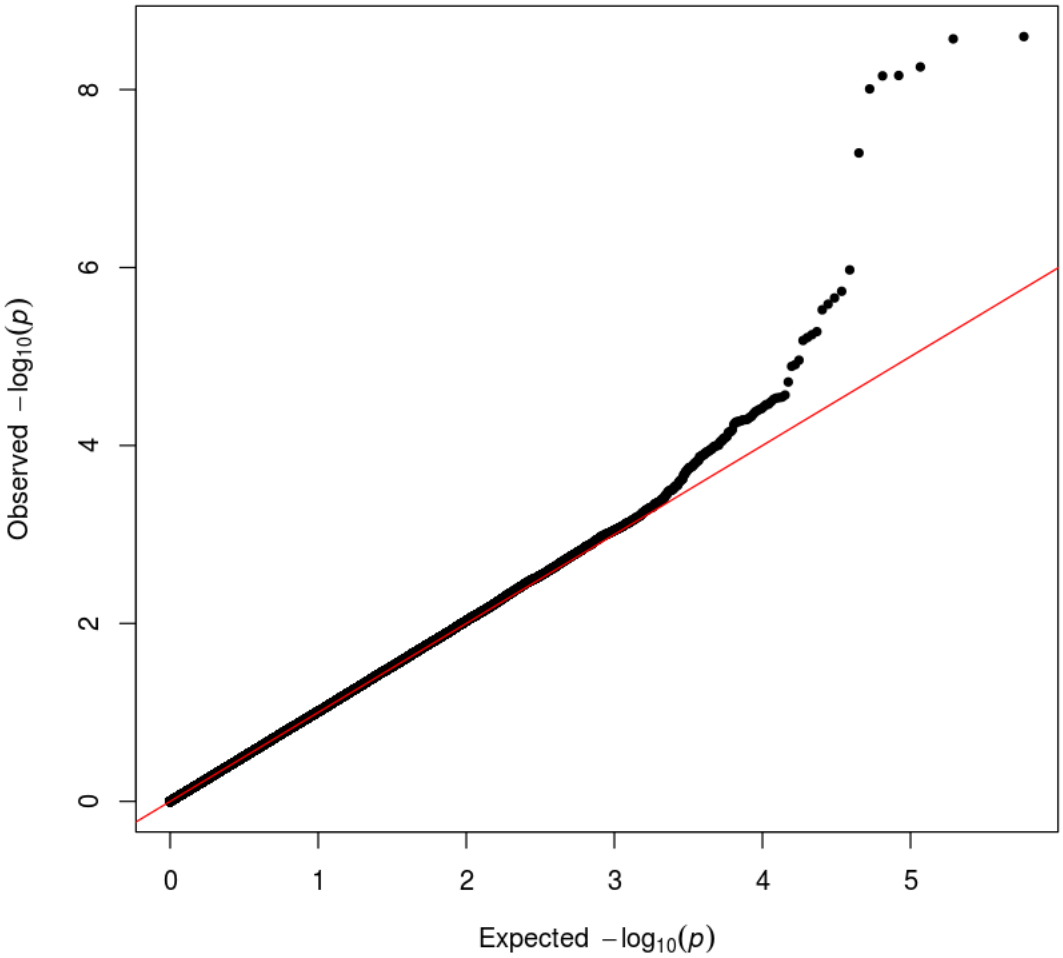


Fig S27. Q-Q plot of HDL-C in EAS populaiton

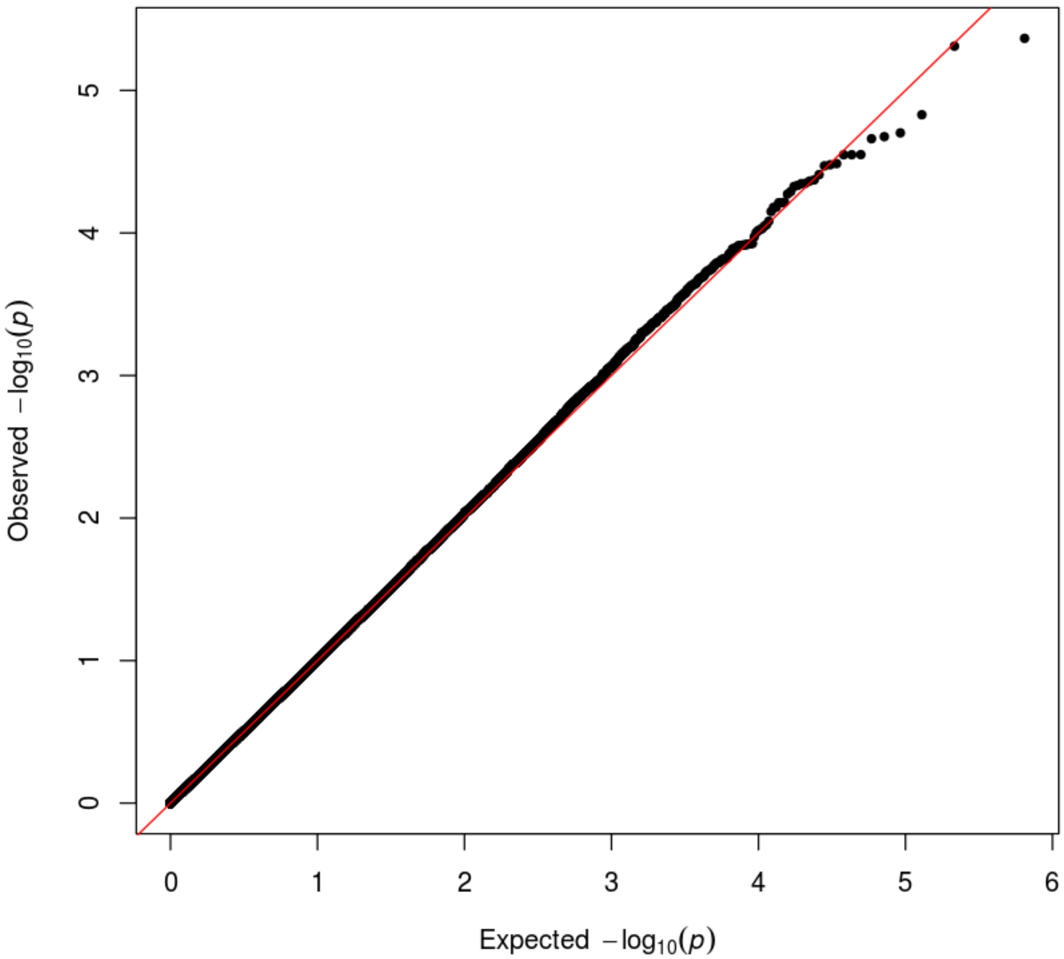


Fig S28. Q-Q plot of HDL-C in SAS populaiton

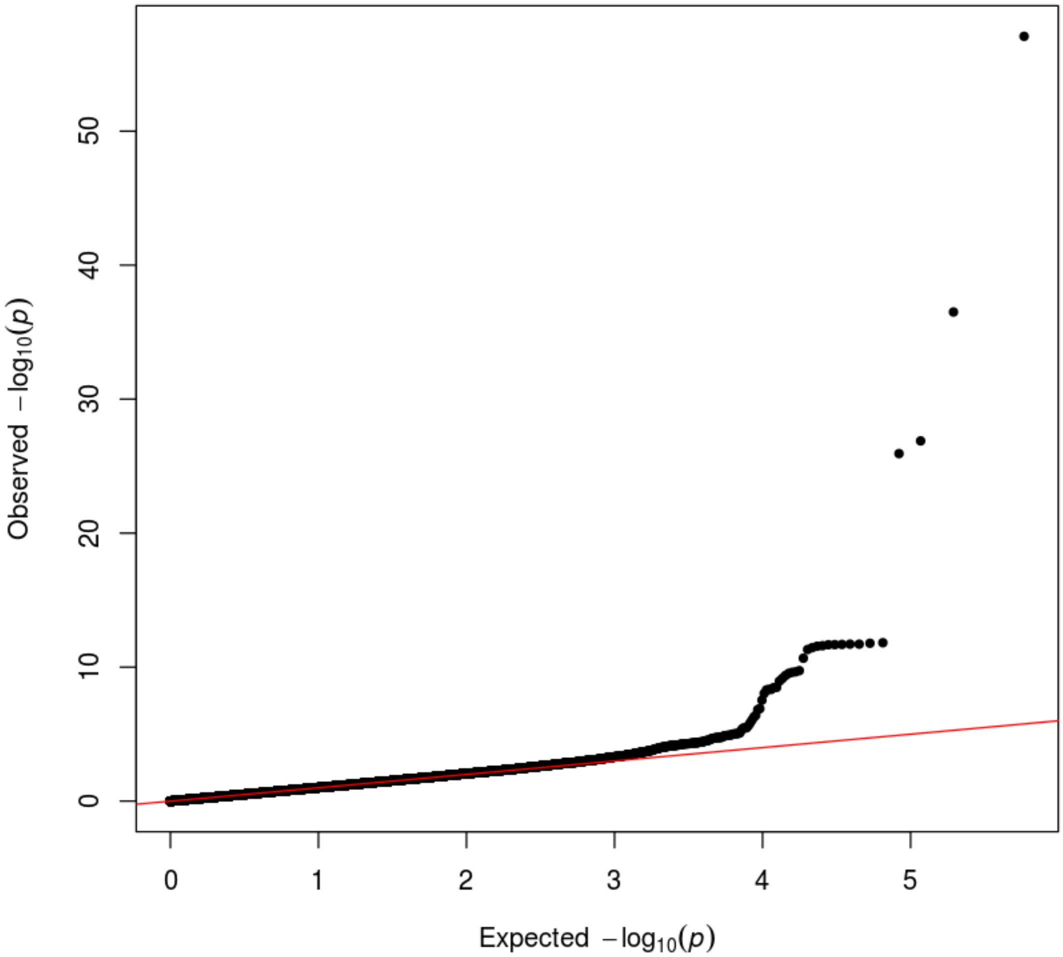


Fig S29. Q-Q plot of LDL Calc in EUR populaiton

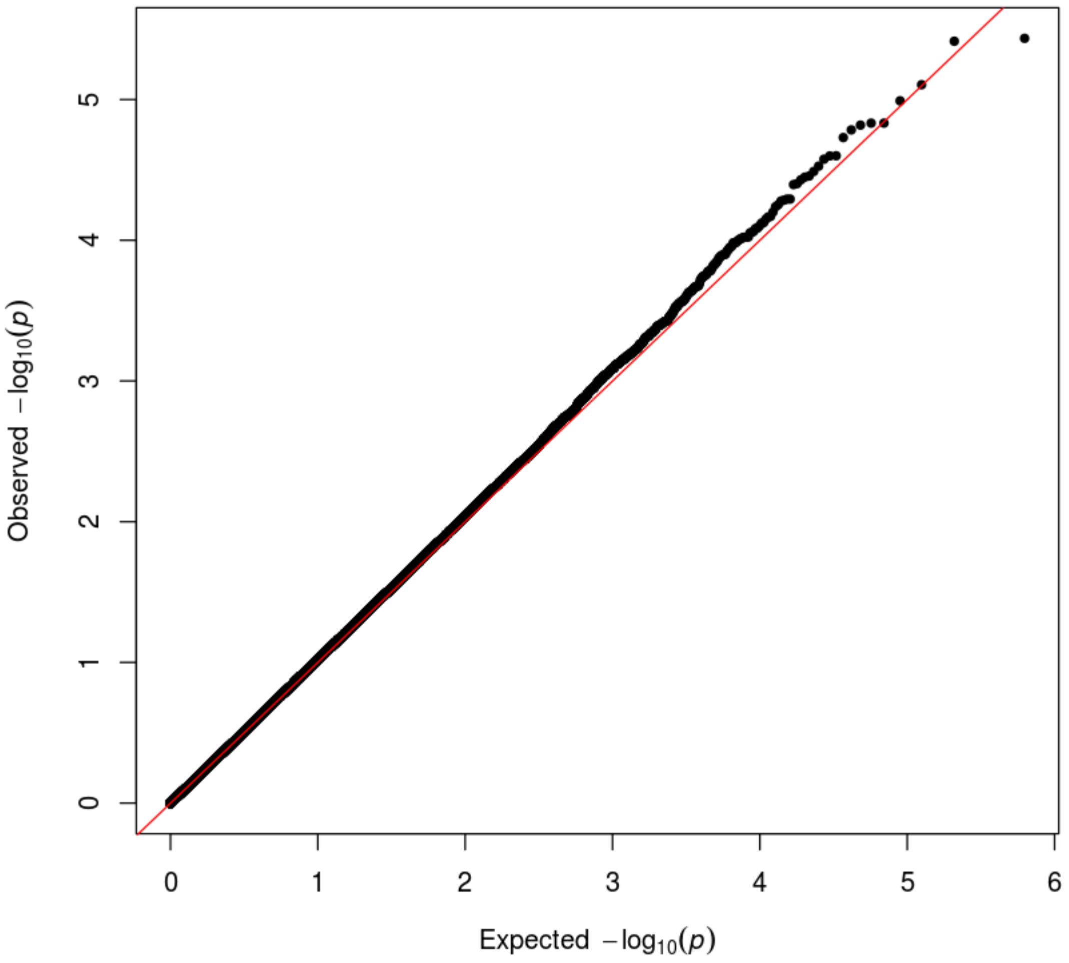


Fig S30. Q-Q plot of LDL Calc in AMR populaiton

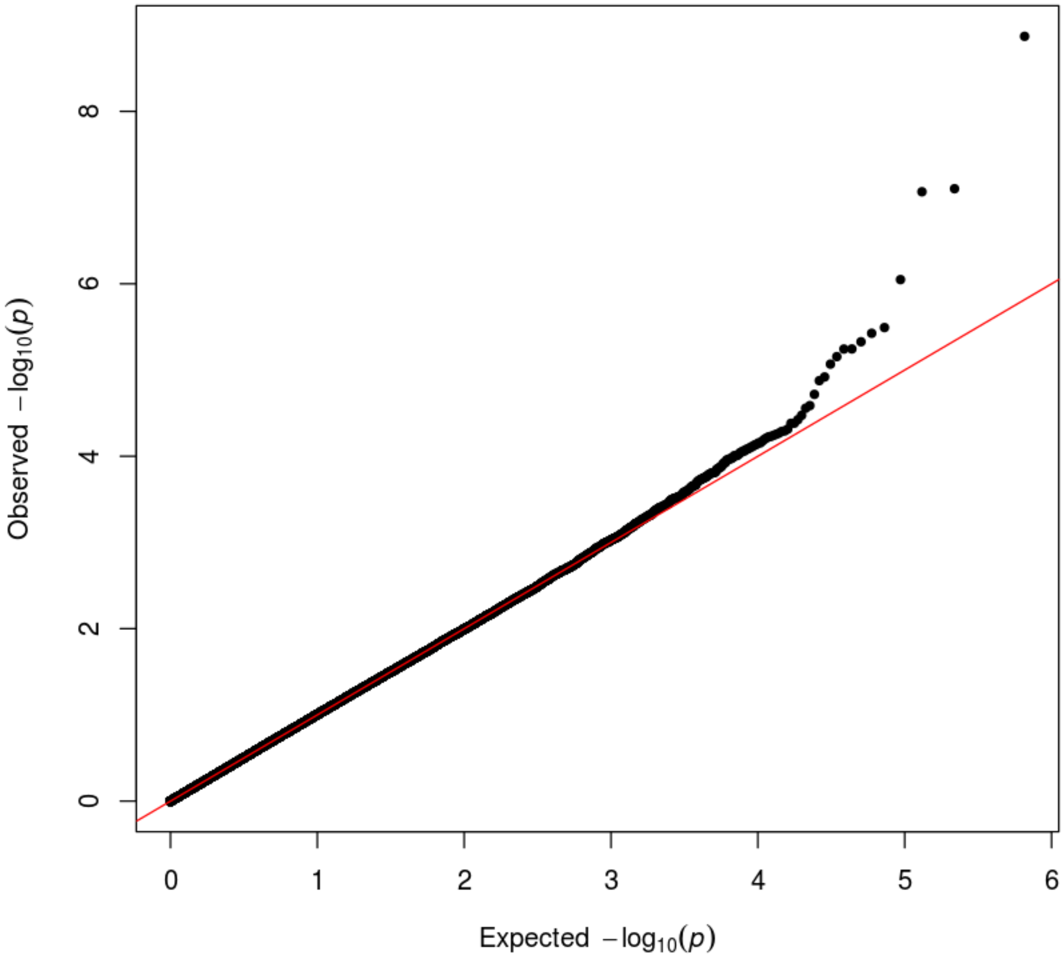


Fig S31. Q-Q plot of LDL Calc in AFR populaiton

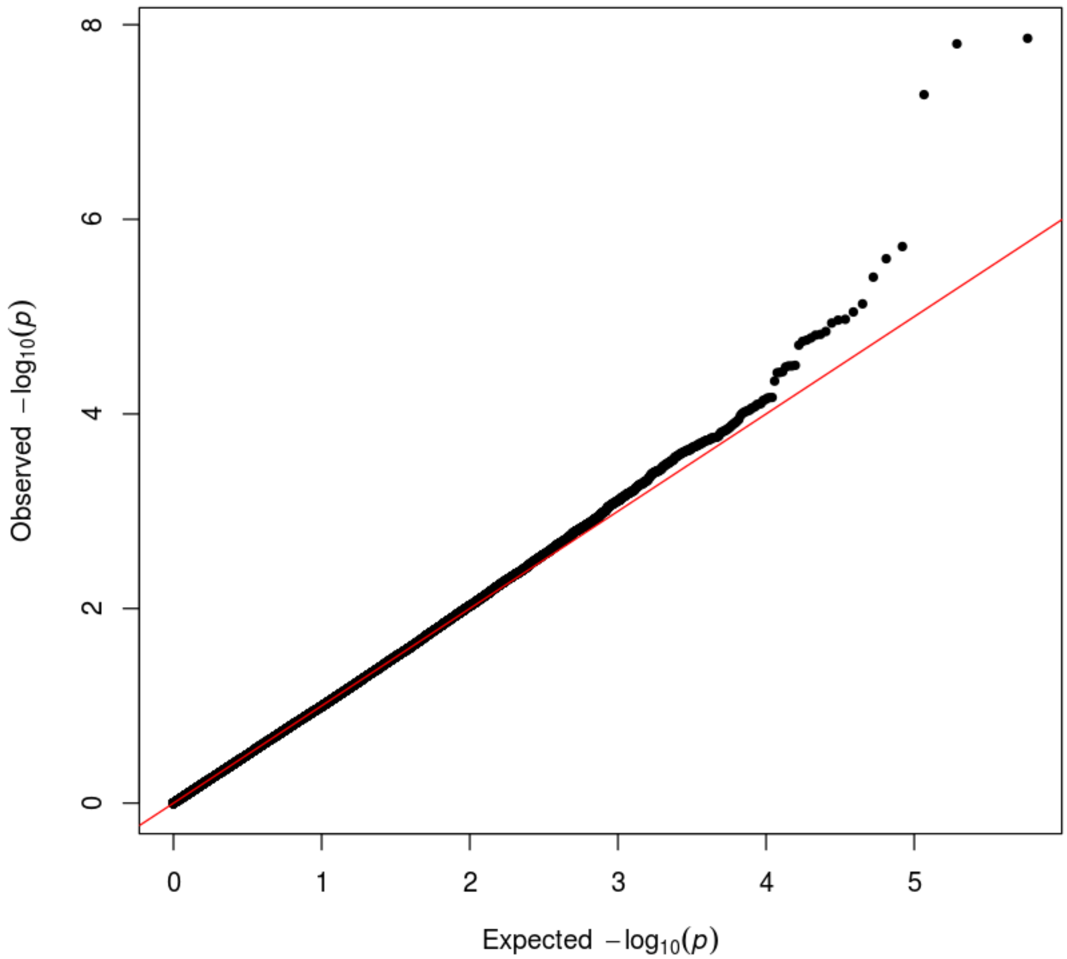


Fig S32. Q-Q plot of LDL Calc in EAS populaiton

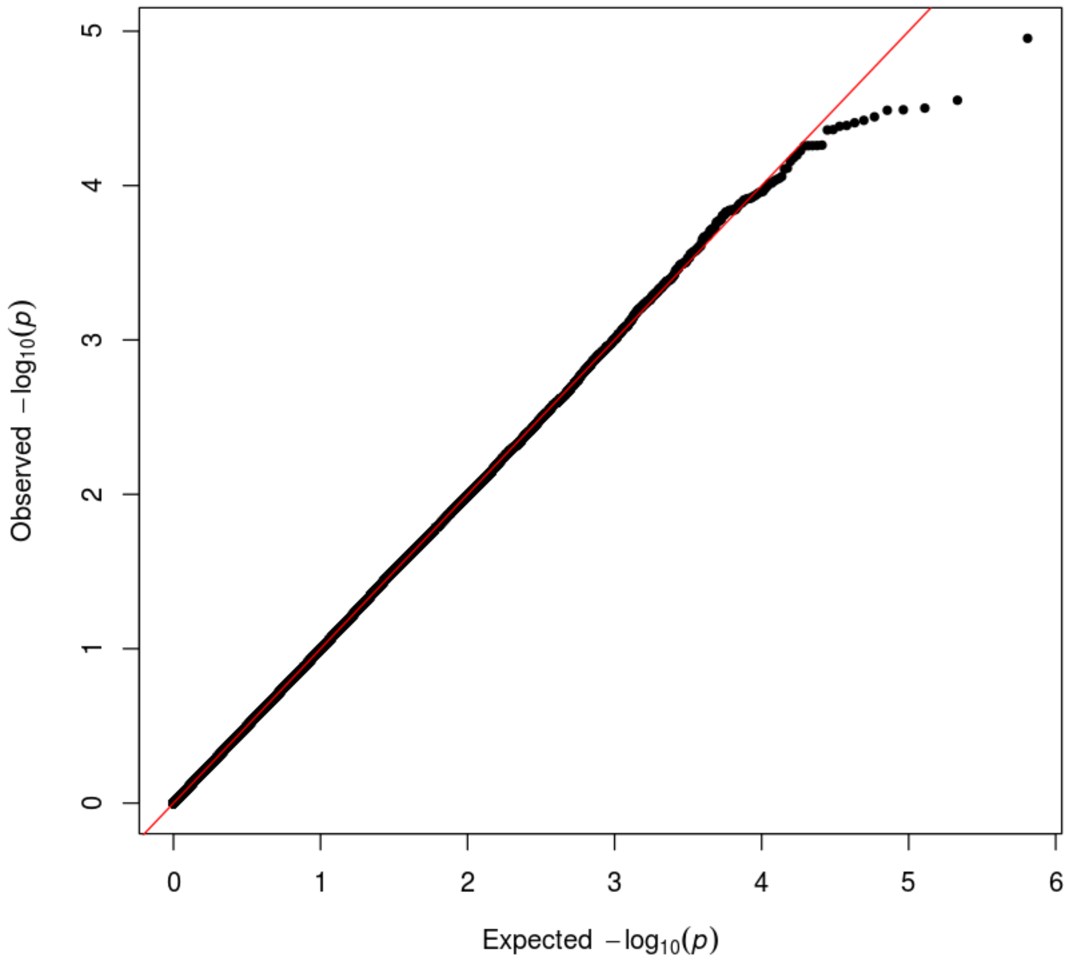


Fig S33. Q-Q plot of LDL Calc in SAS populaiton

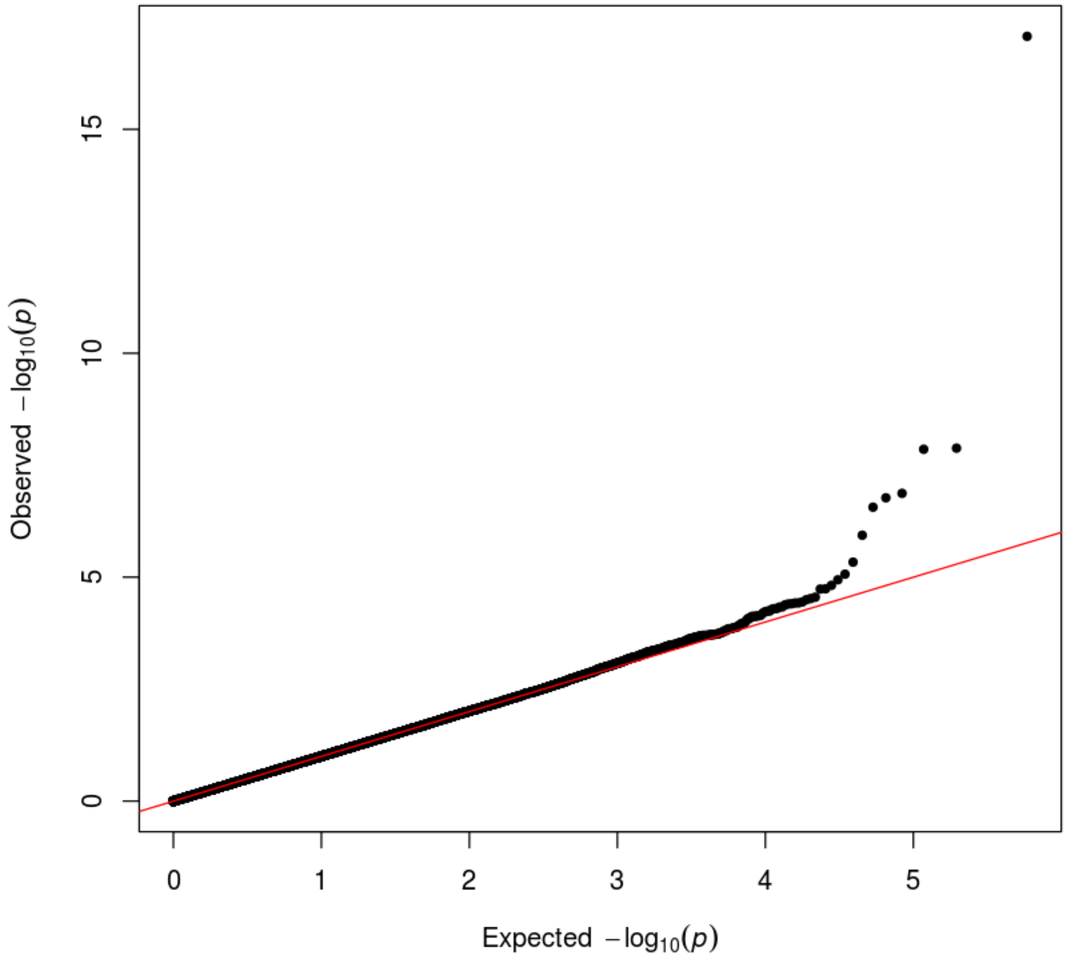


Fig S34. Q-Q plot of LDL Quant in EUR populaiton

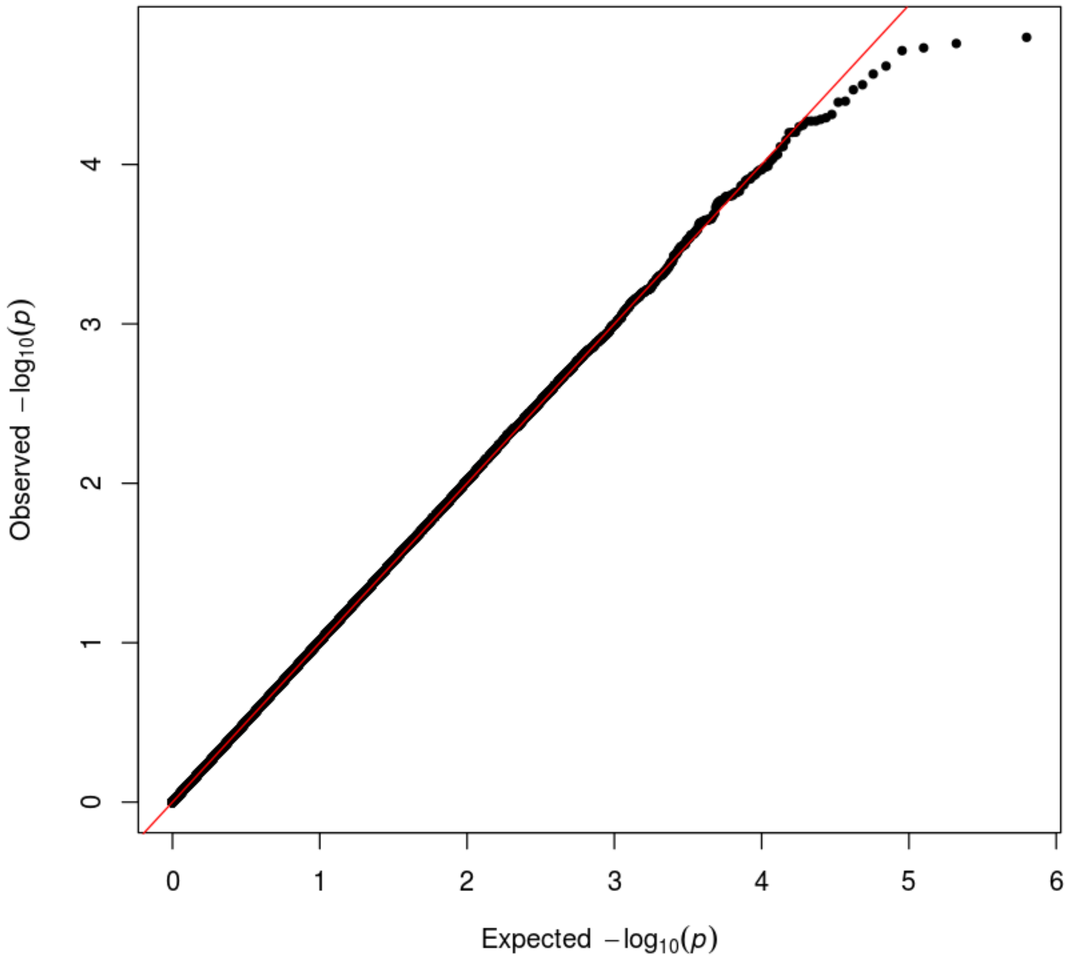


Fig S35. Q-Q plot of LDL Quant in AMR populaiton

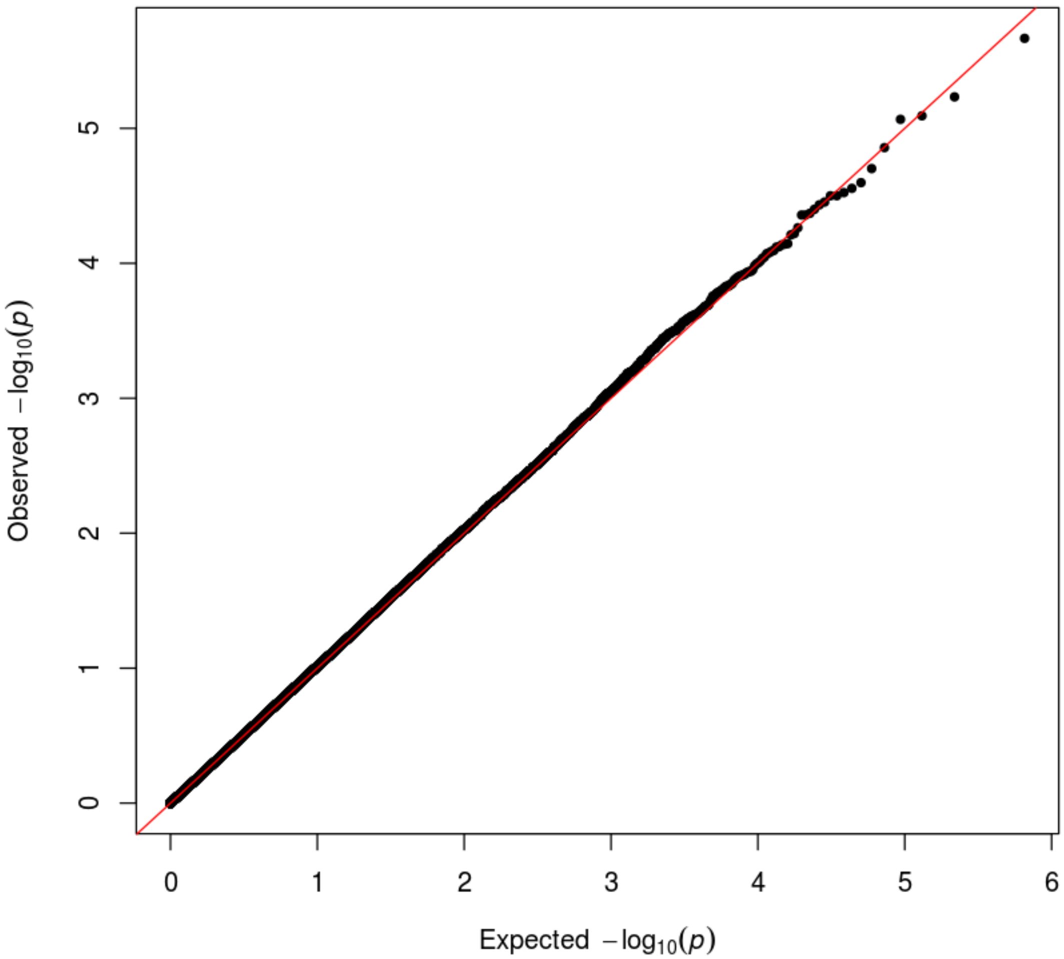


Fig S36. Q-Q plot of LDL Quant in AFR populaiton

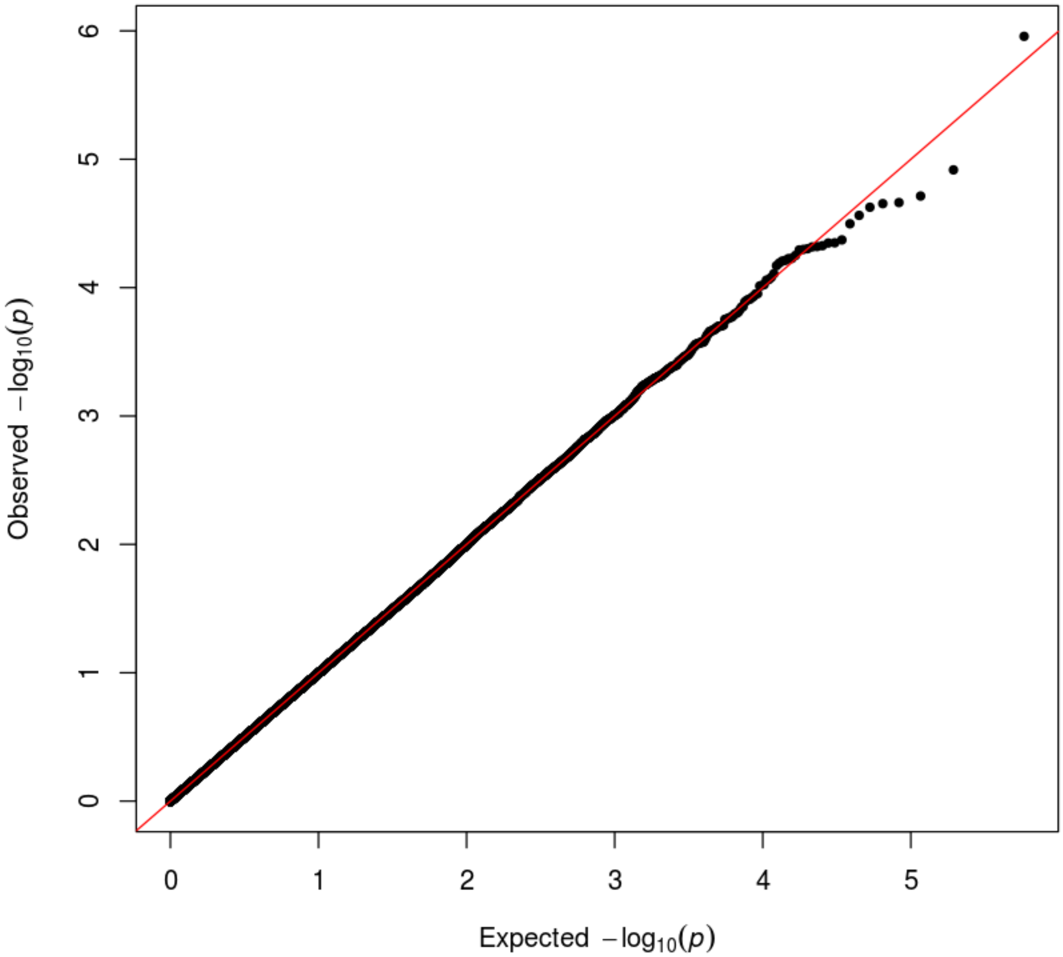


Fig S37. Q-Q plot of LDL Quant in EAS populaiton

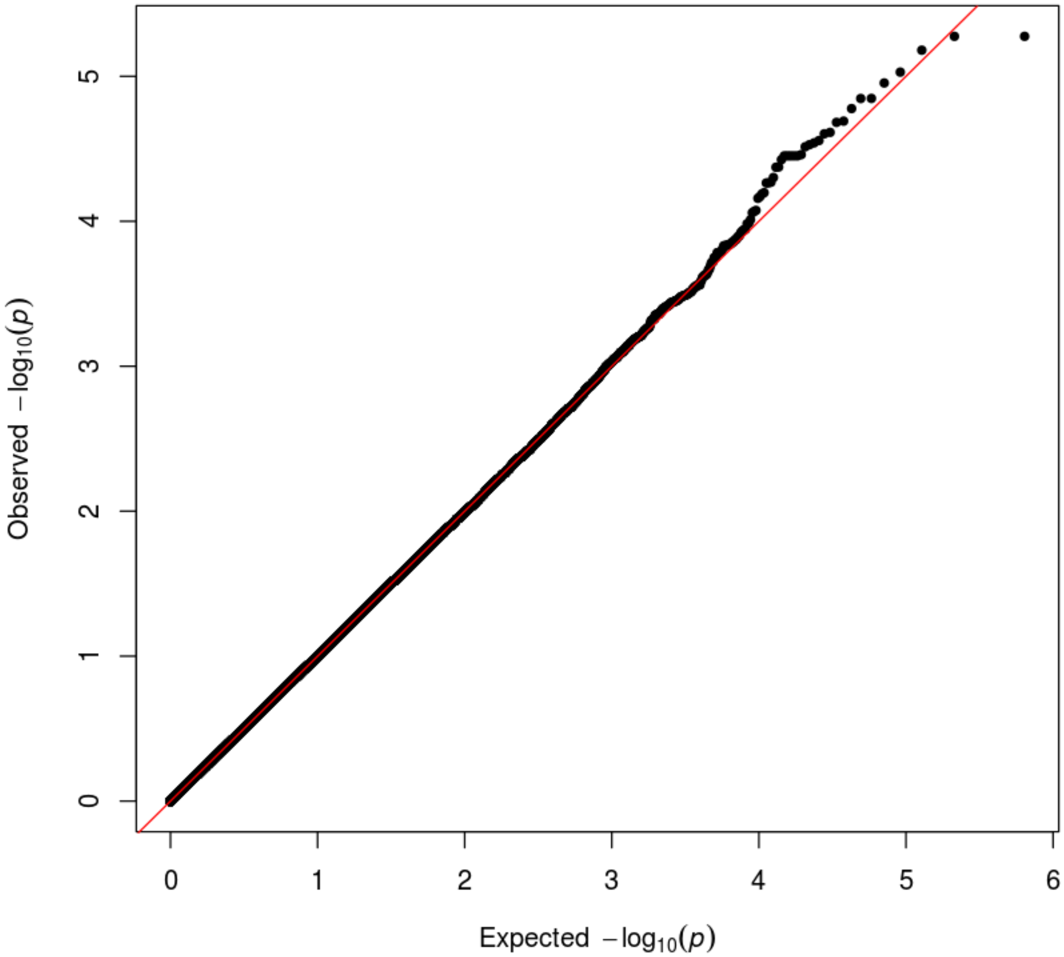


Fig S38. Q-Q plot of LDL Quant in SAS populaiton

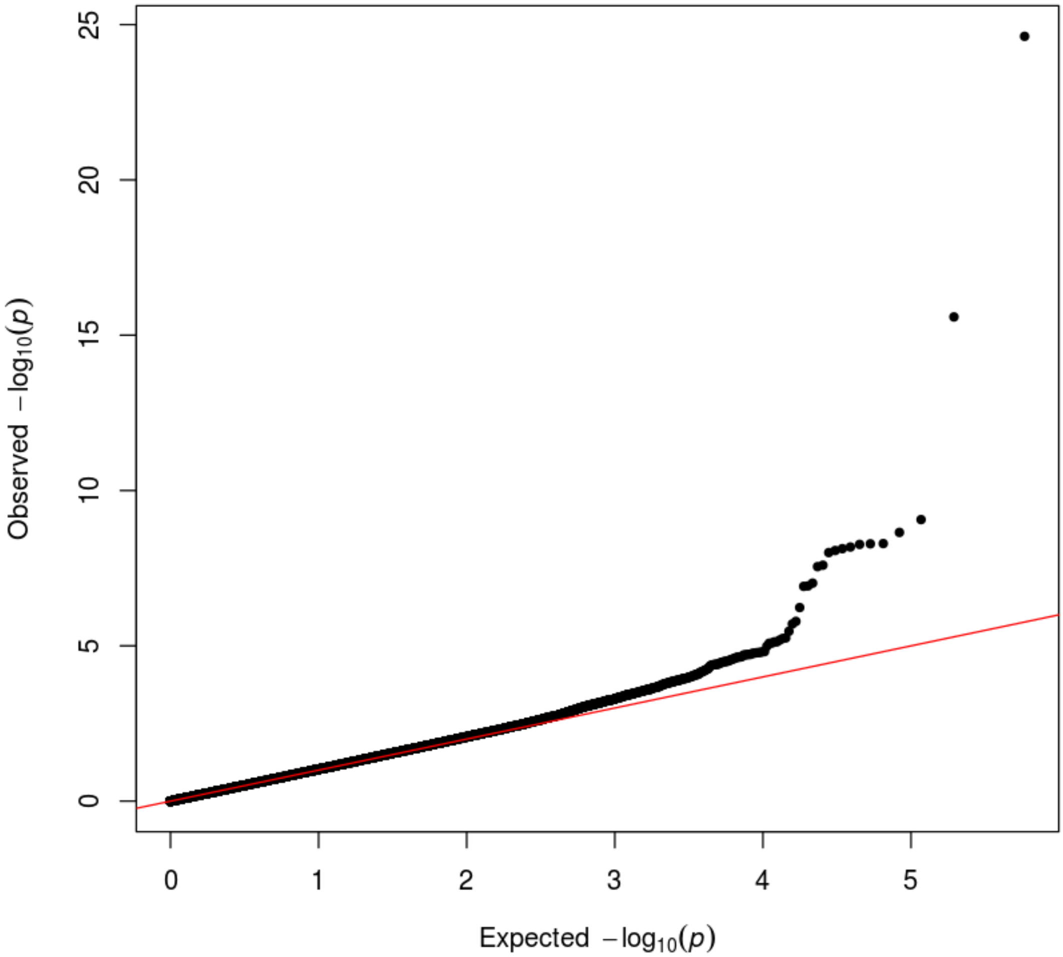


Fig S39. Q-Q plot of TC in EUR populaiton

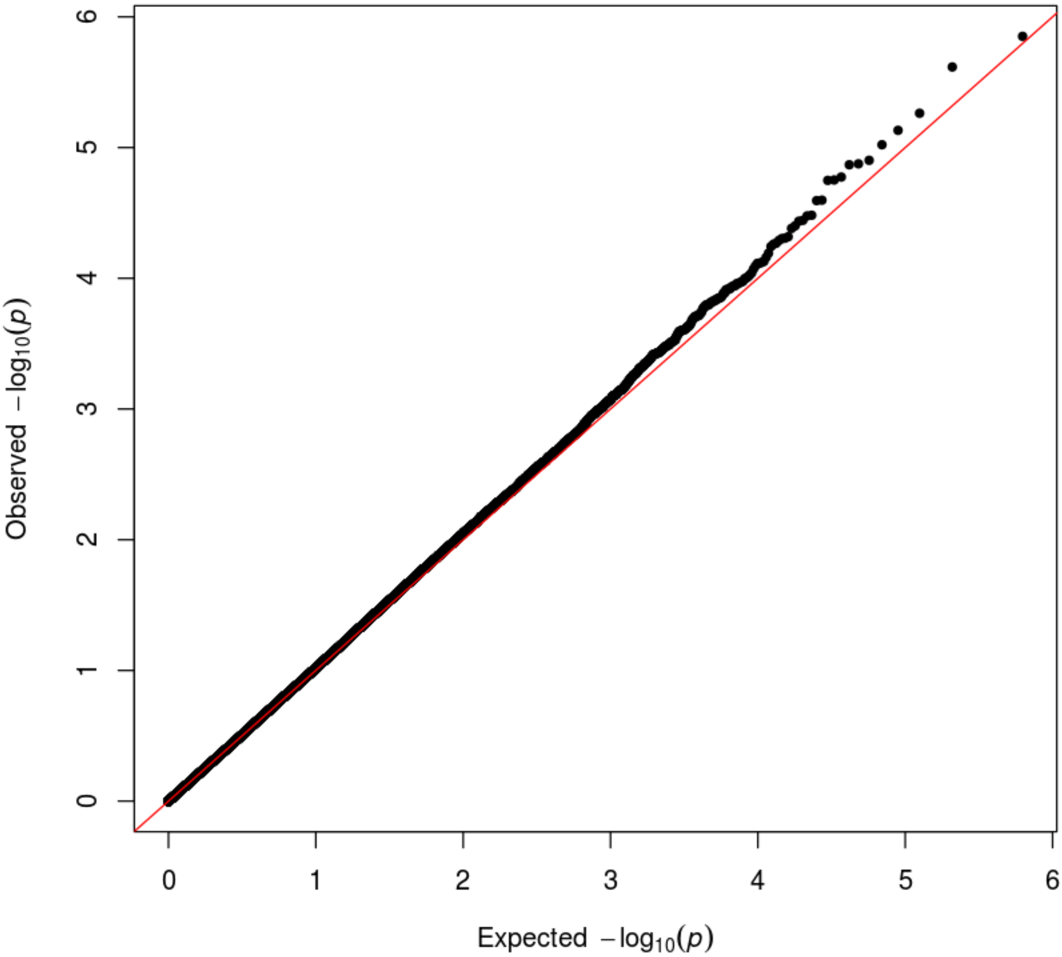


Fig S40. Q-Q plot of TC in AMR populaiton

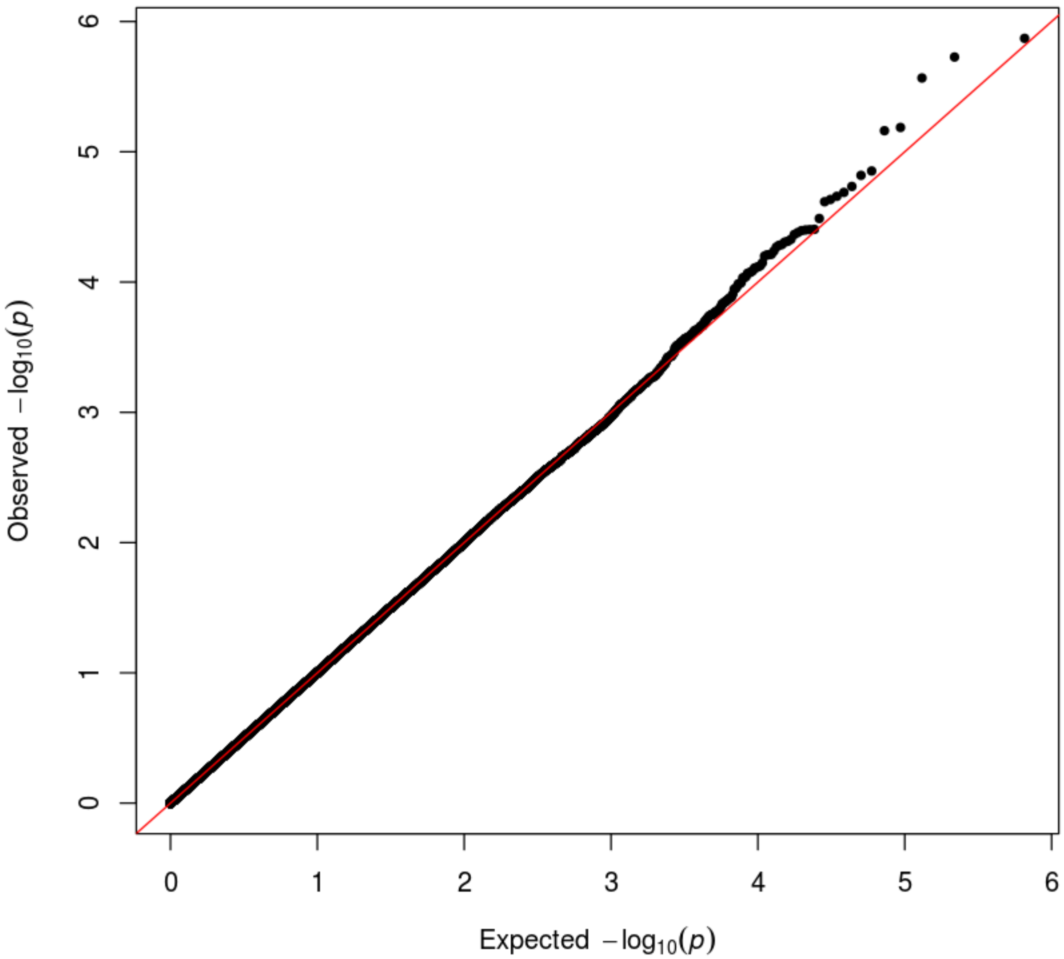


Fig S41. Q-Q plot of TC in AFR populaiton

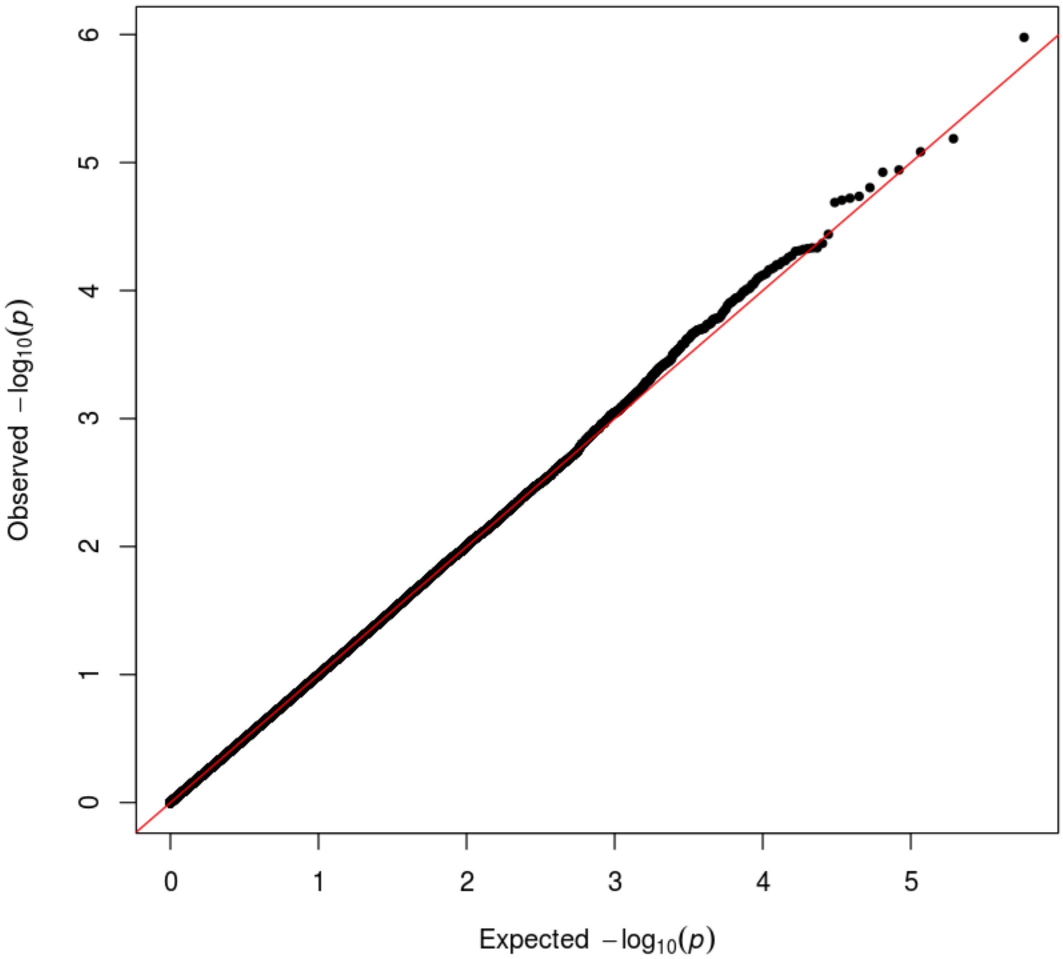


Fig S42. Q-Q plot of TC in EAS populaiton

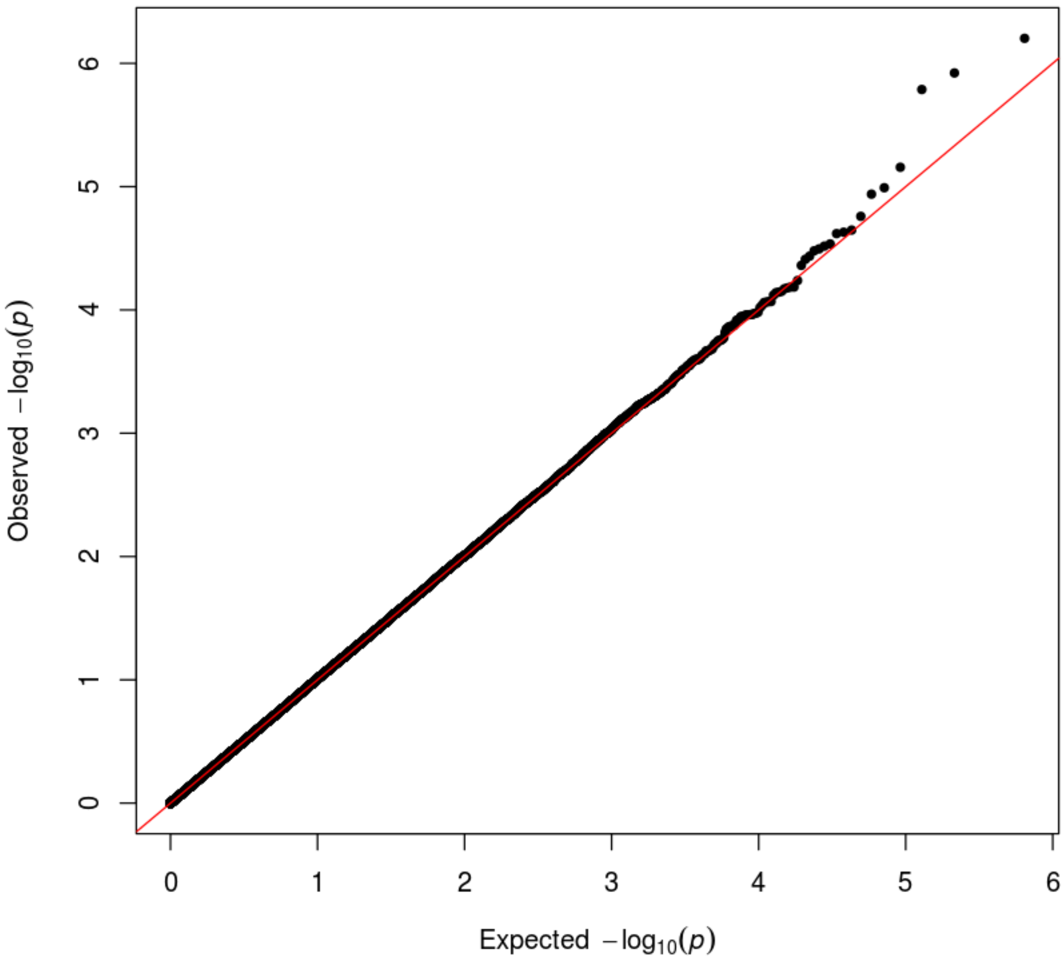


Fig S43. Q-Q plot of TC in SAS populaiton

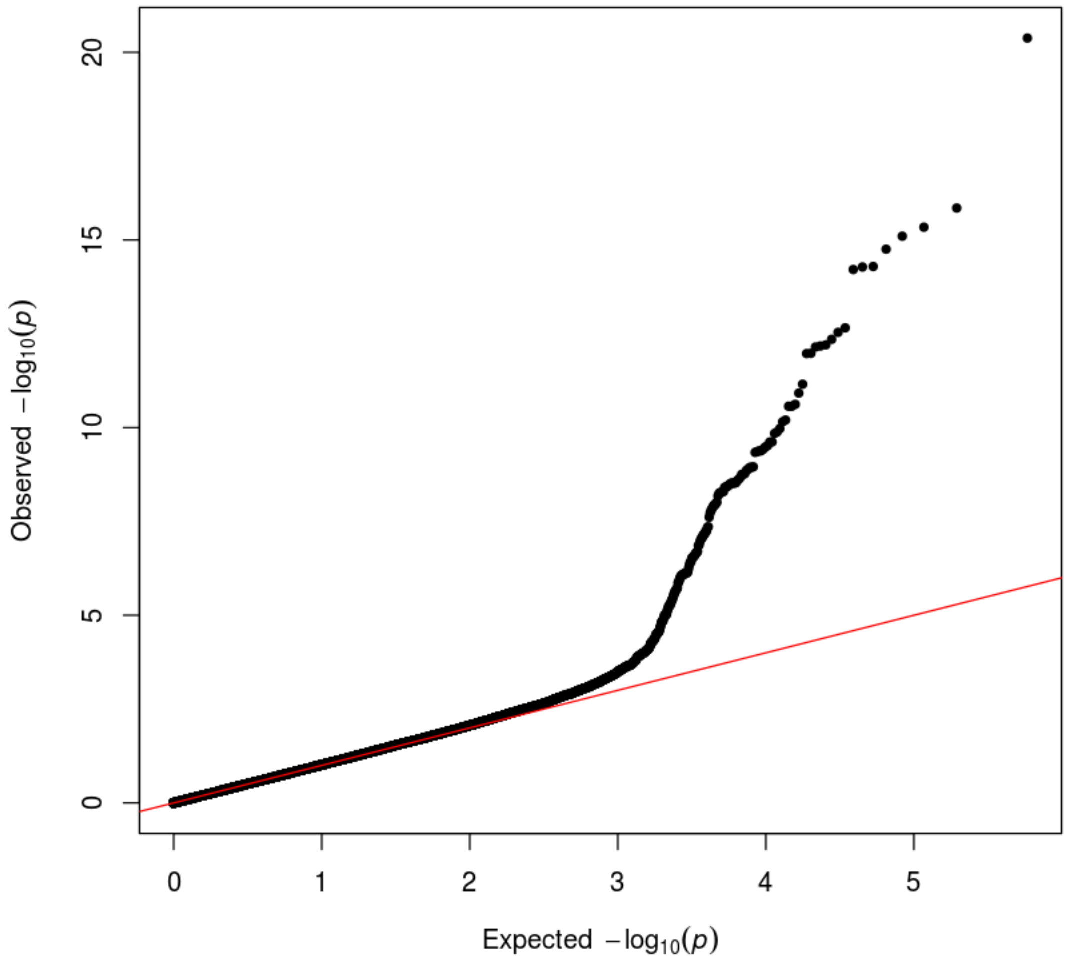


Fig S44. Q-Q plot of TG in EUR populaiton

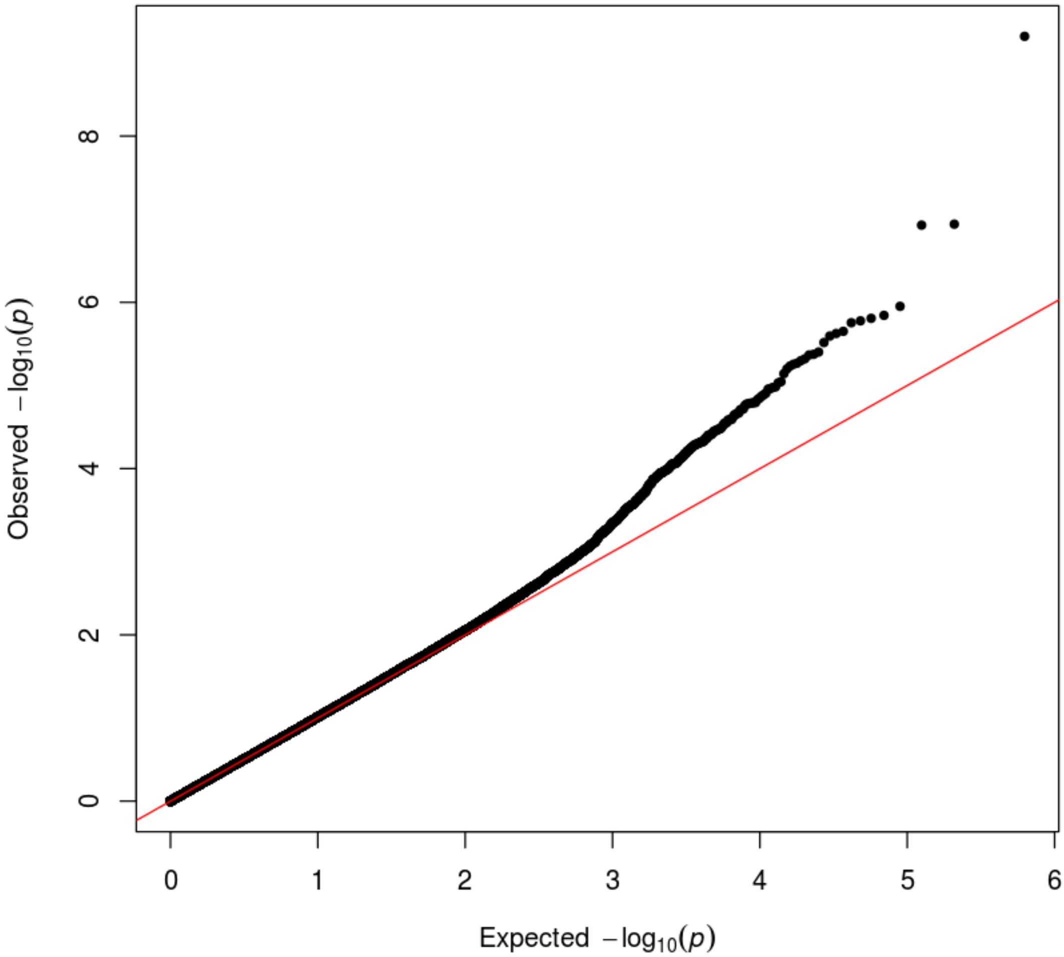


Fig S45. Q-Q plot of TG in AMR populaiton

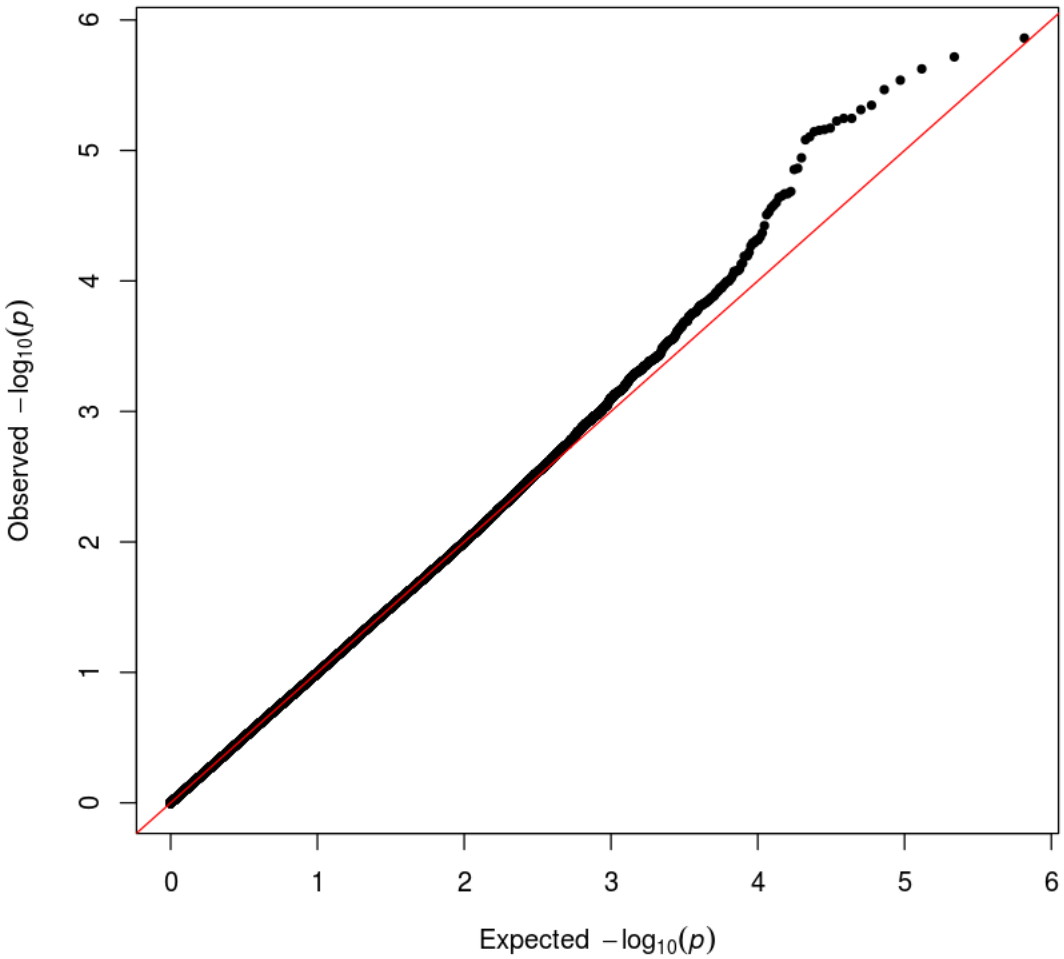


Fig S46. Q-Q plot of TG in AFR populaiton

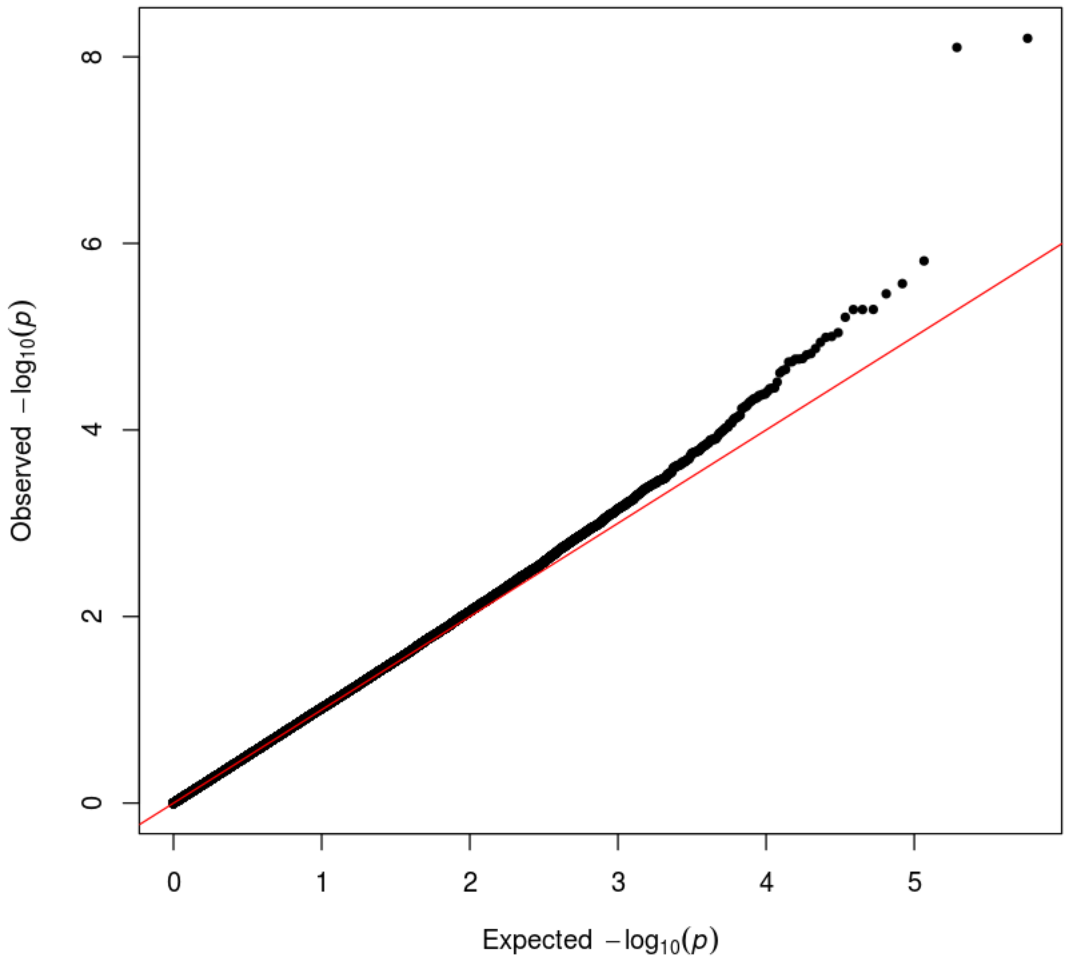


Fig S47. Q-Q plot of TG in EAS populaiton

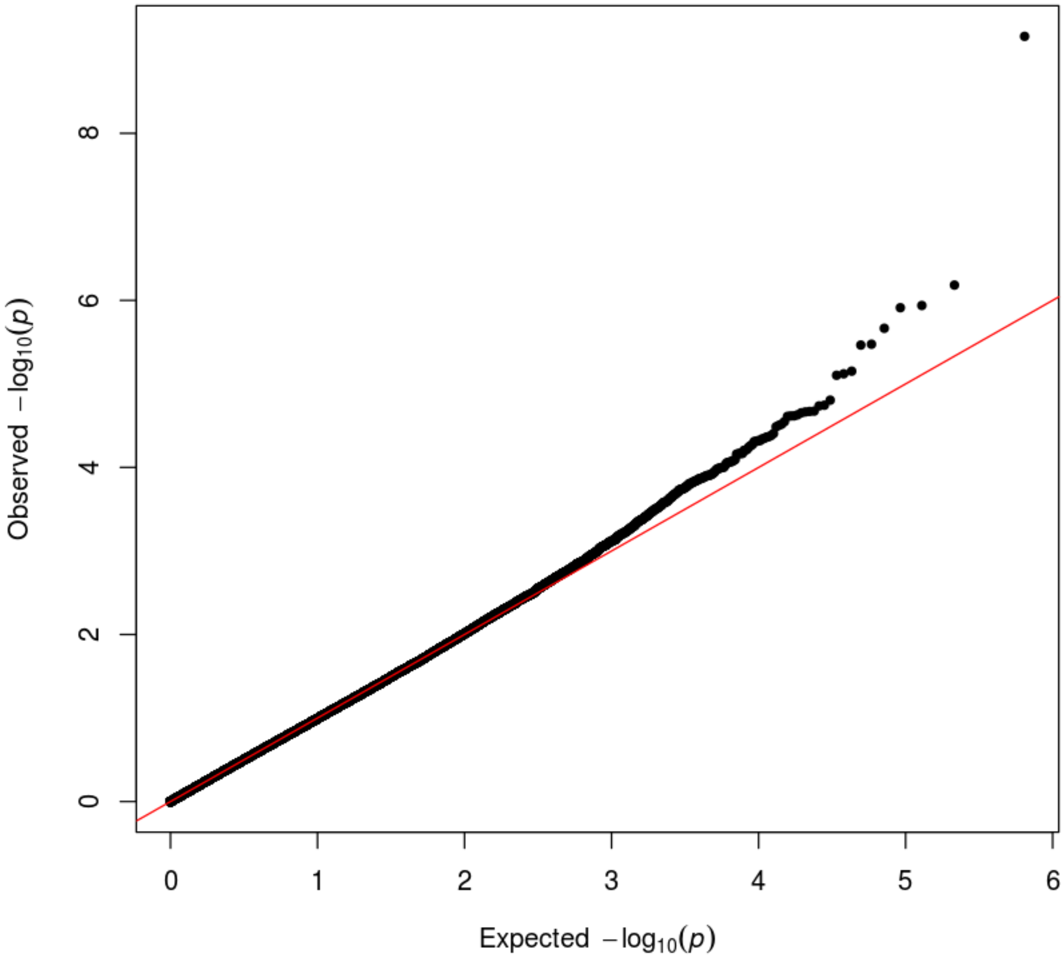


Fig S48. Q-Q plot of TG in SAS populaiton

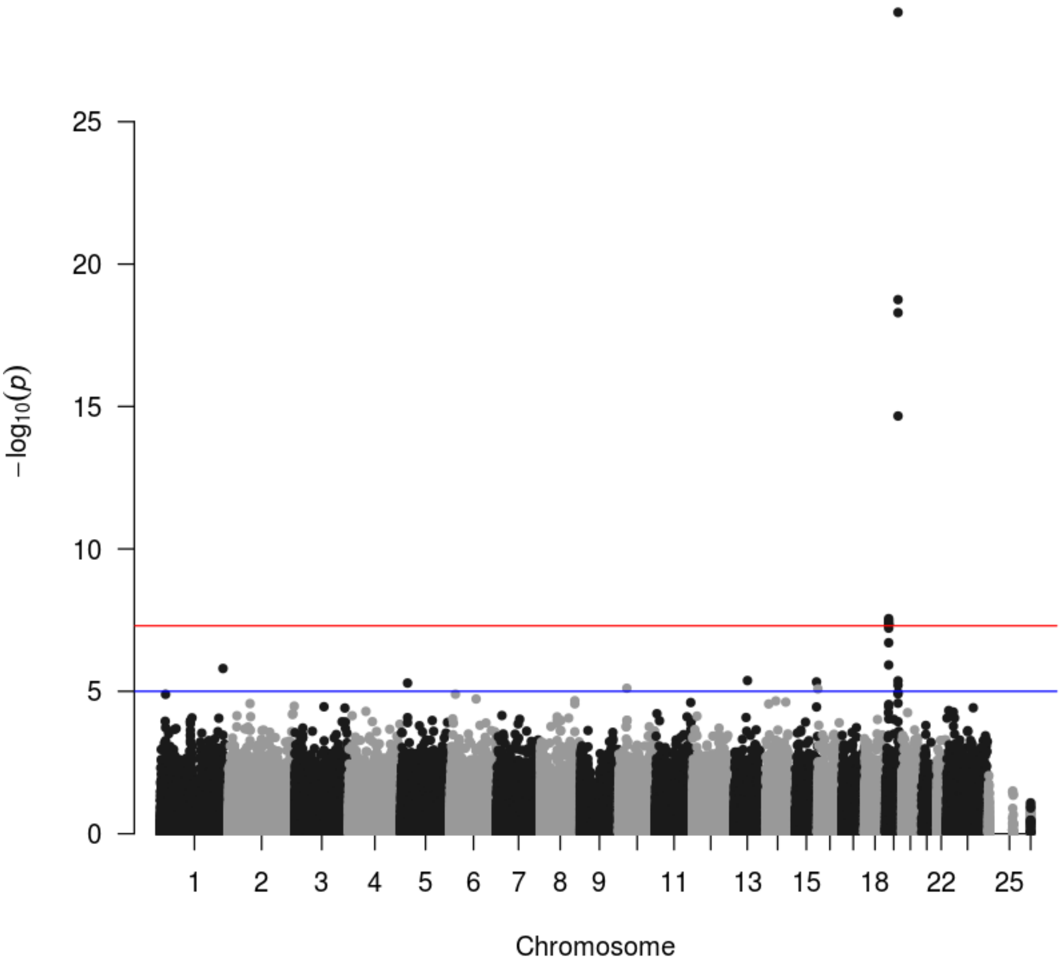


Fig S49. Manhattan plot of LDL Calc in meta-analysis

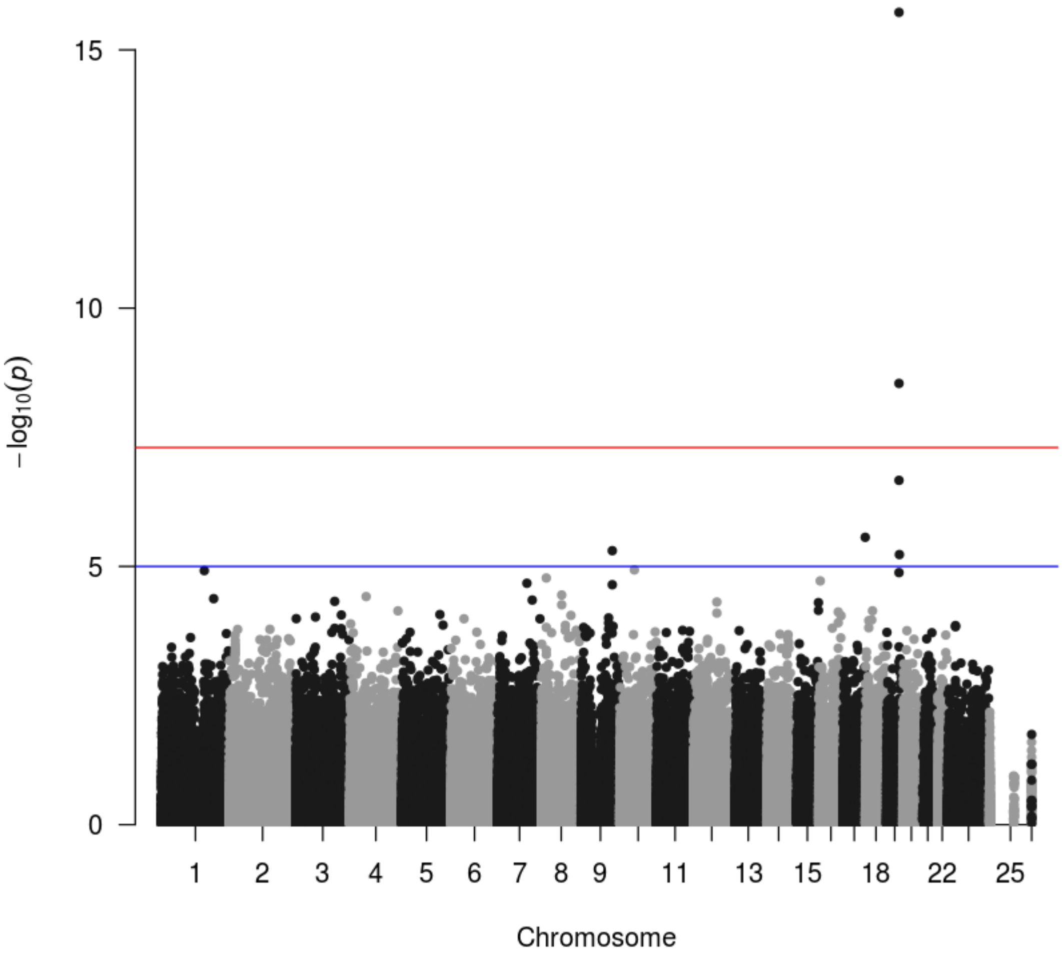


Fig S50. Manhattan plot of LDL Quant in meta-analysis

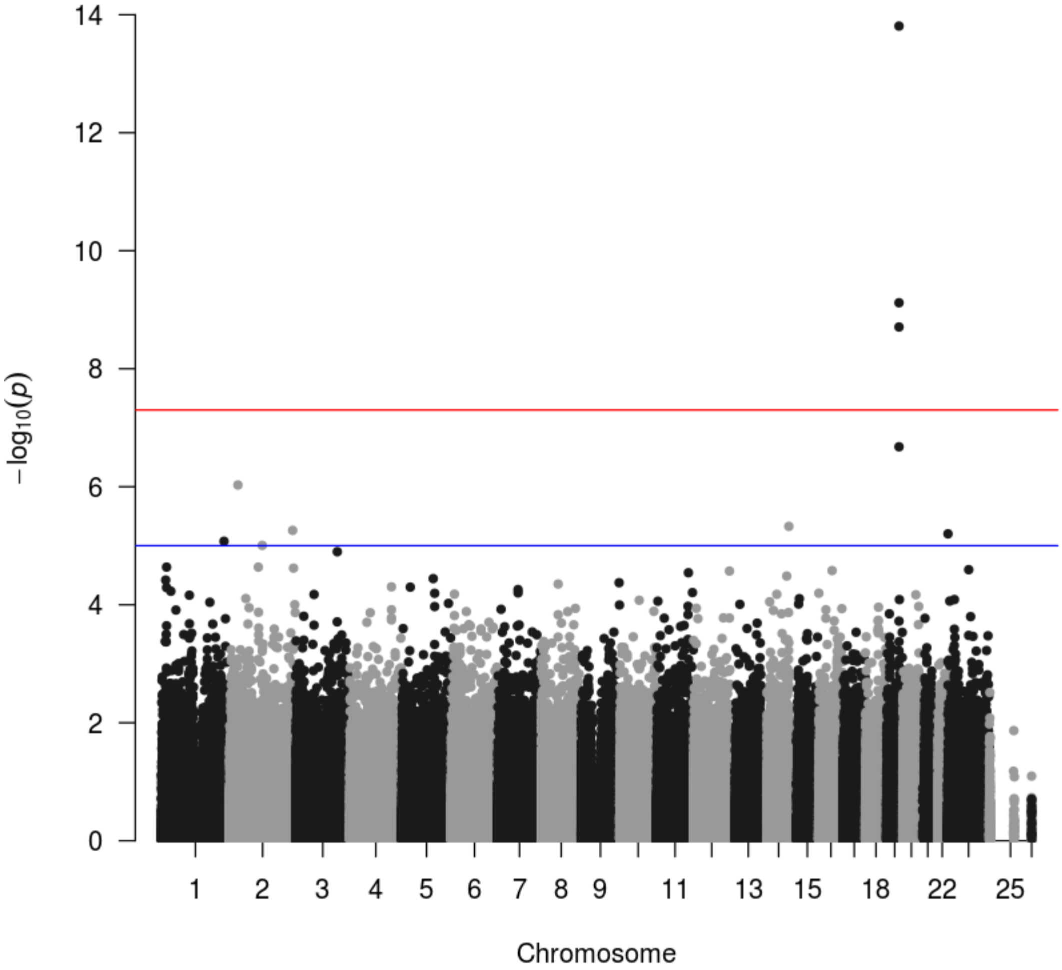


Fig S51. Manhattan plot of TC in meta-analysis

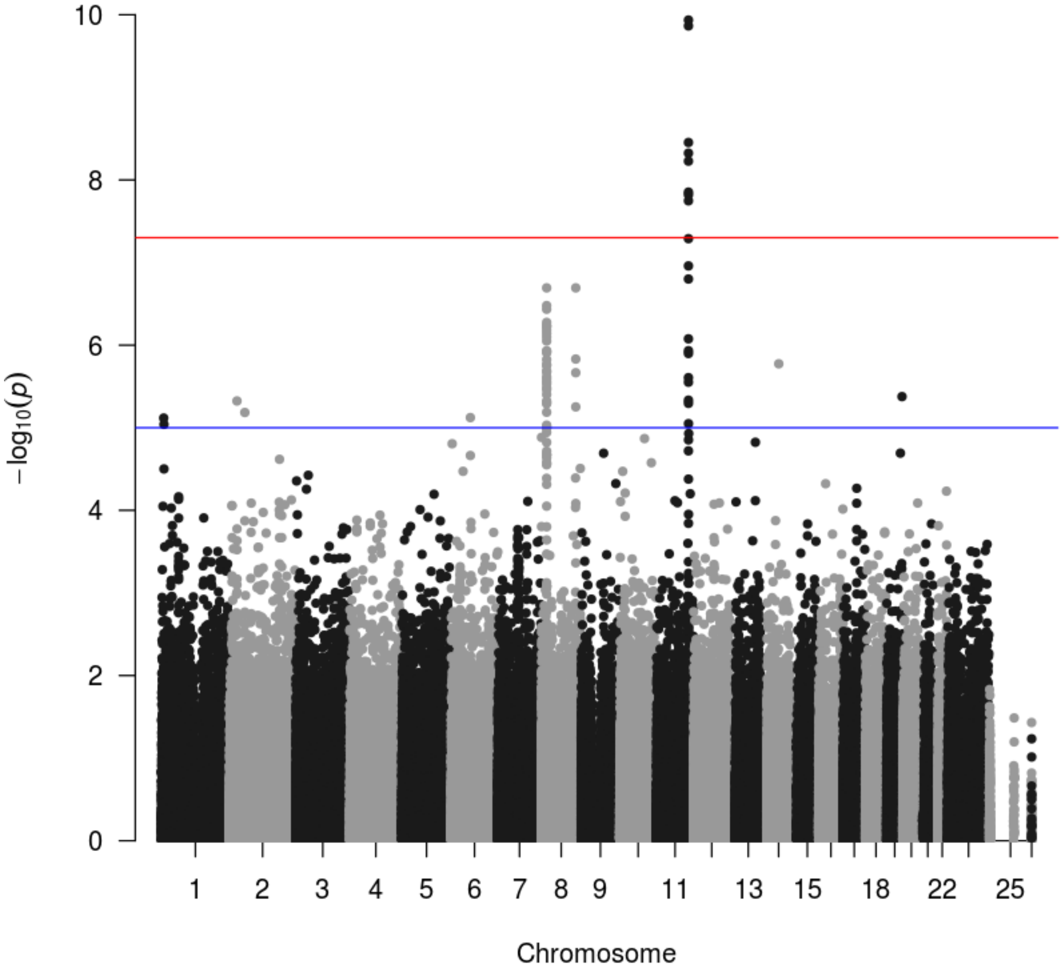


Fig S52. Manhattan plot of TG in meta-analysis

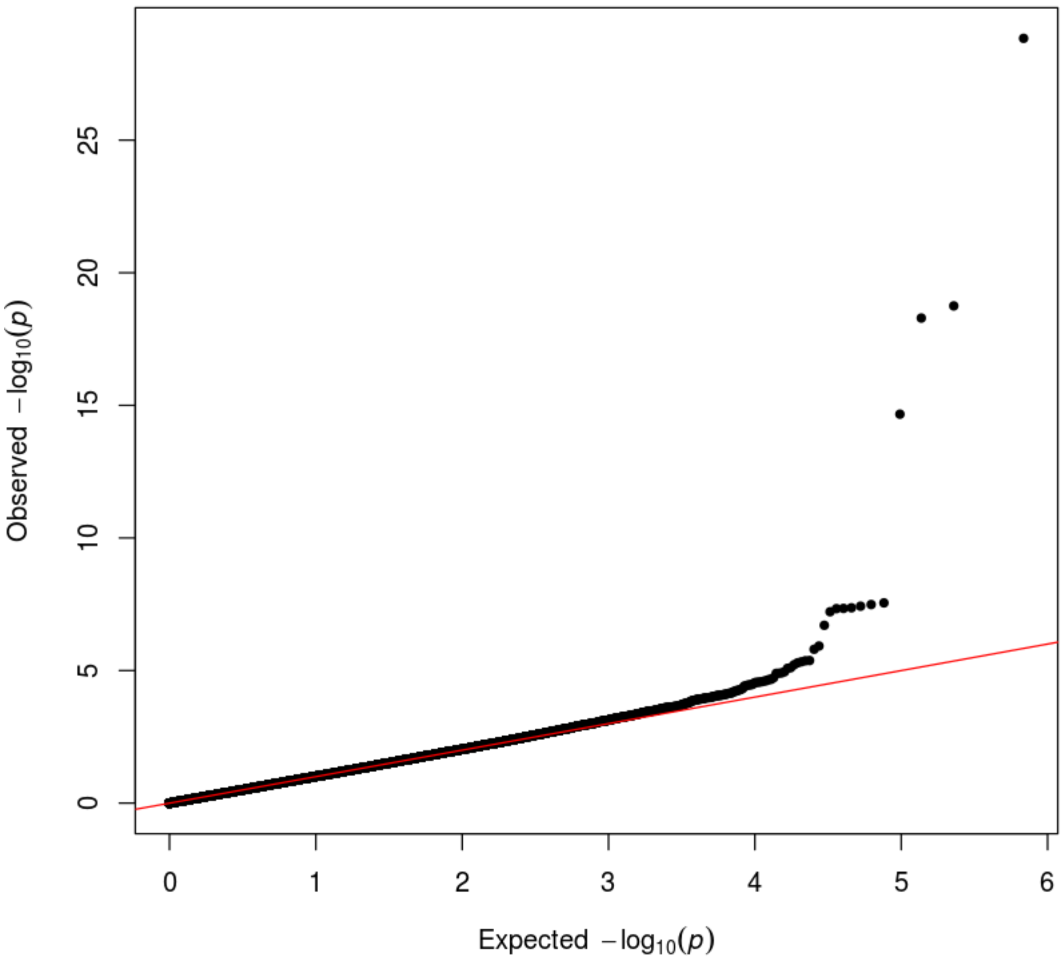


Fig S53. Q-Q plot of LDL Calc in meta-analysis

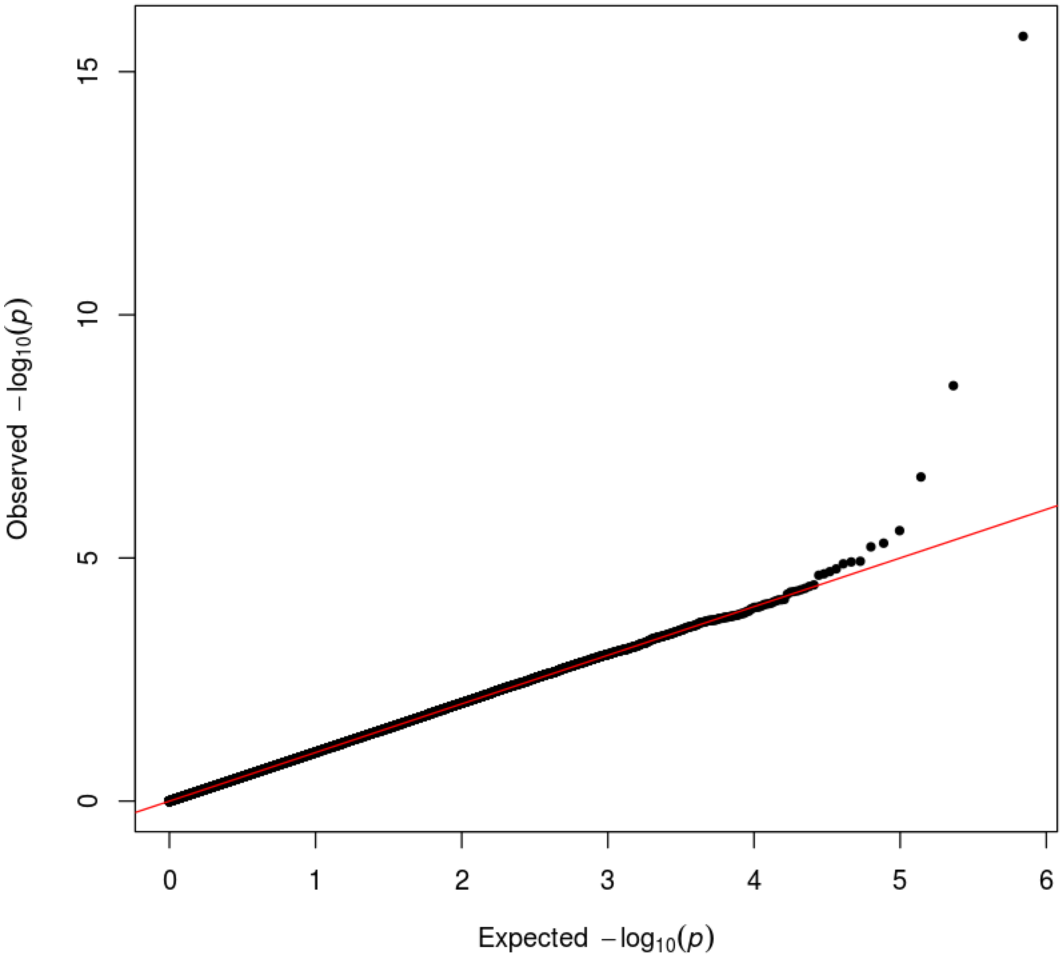


Fig S54. Q-Q plot of LDL Quant in meta-analysis

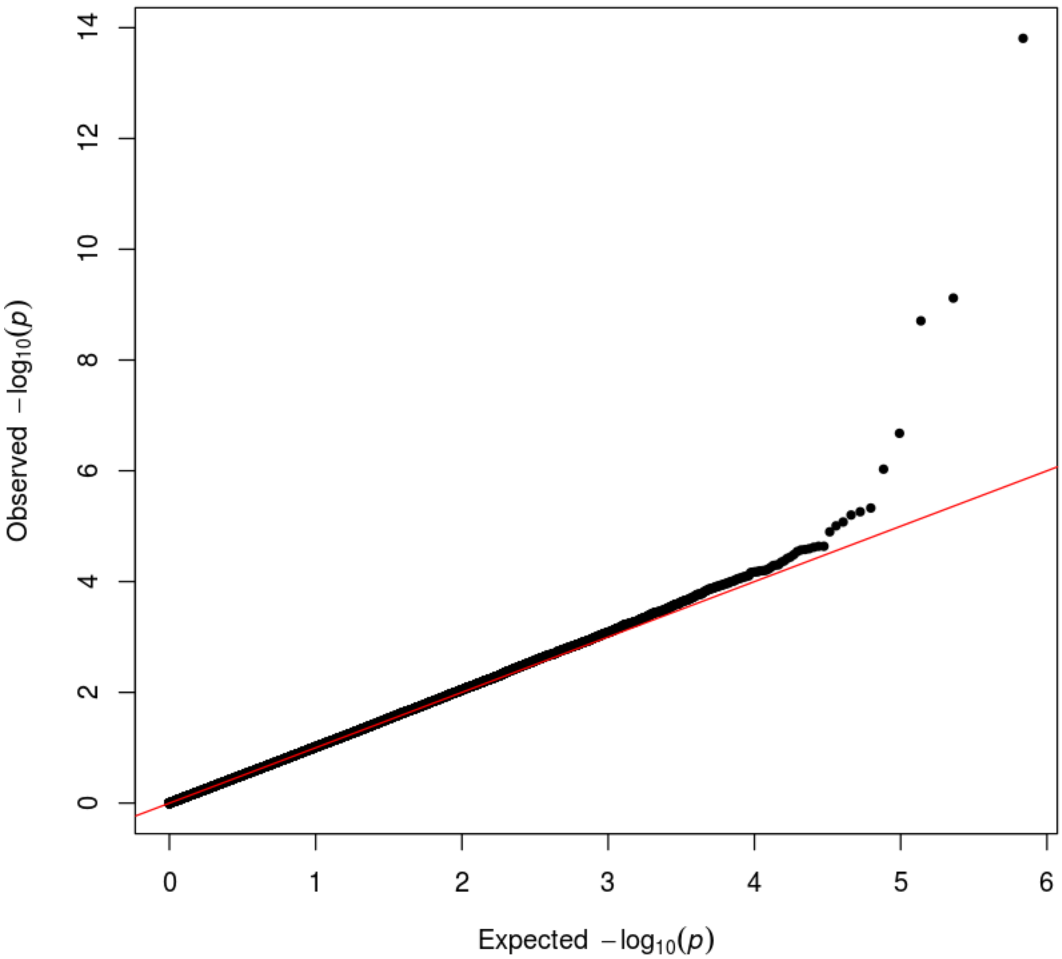


Fig S55. Q-Q plot of TC in meta-analysis

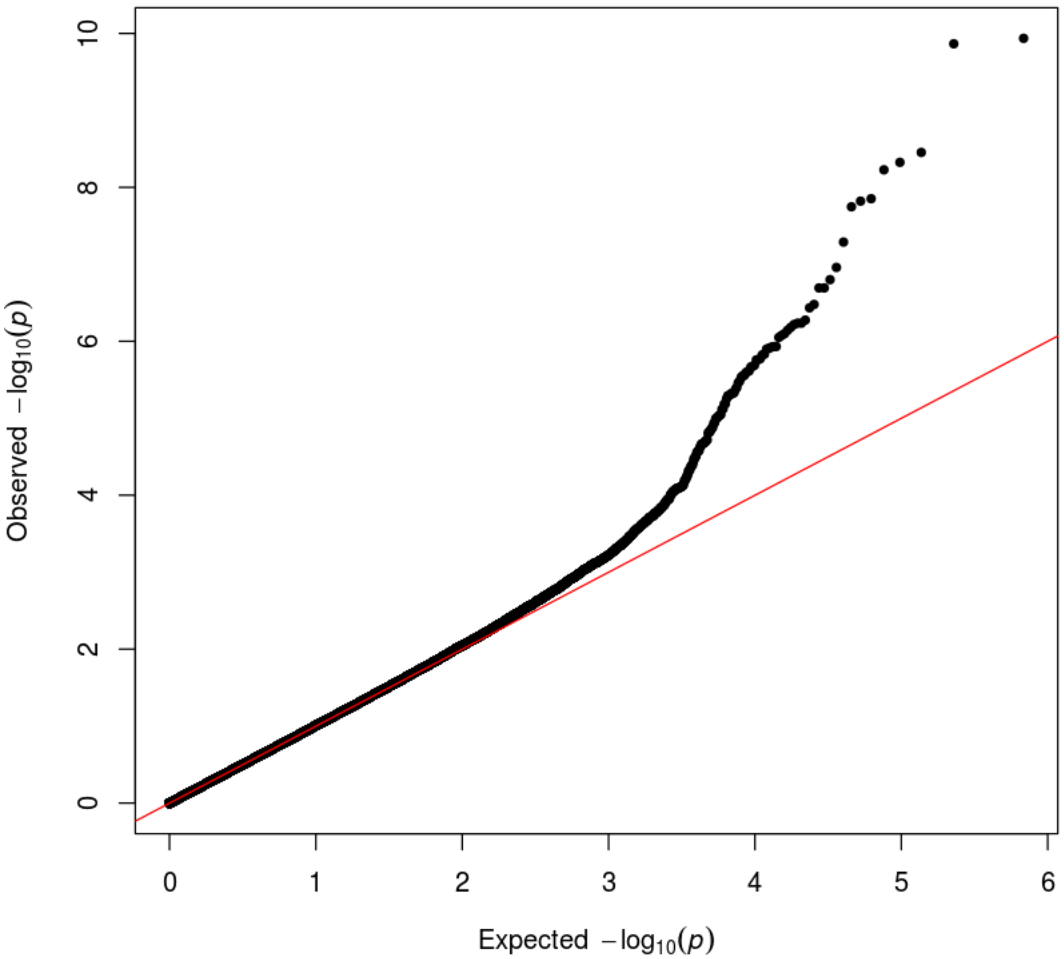


Fig S56. Q-Q plot of TG in meta-analysis

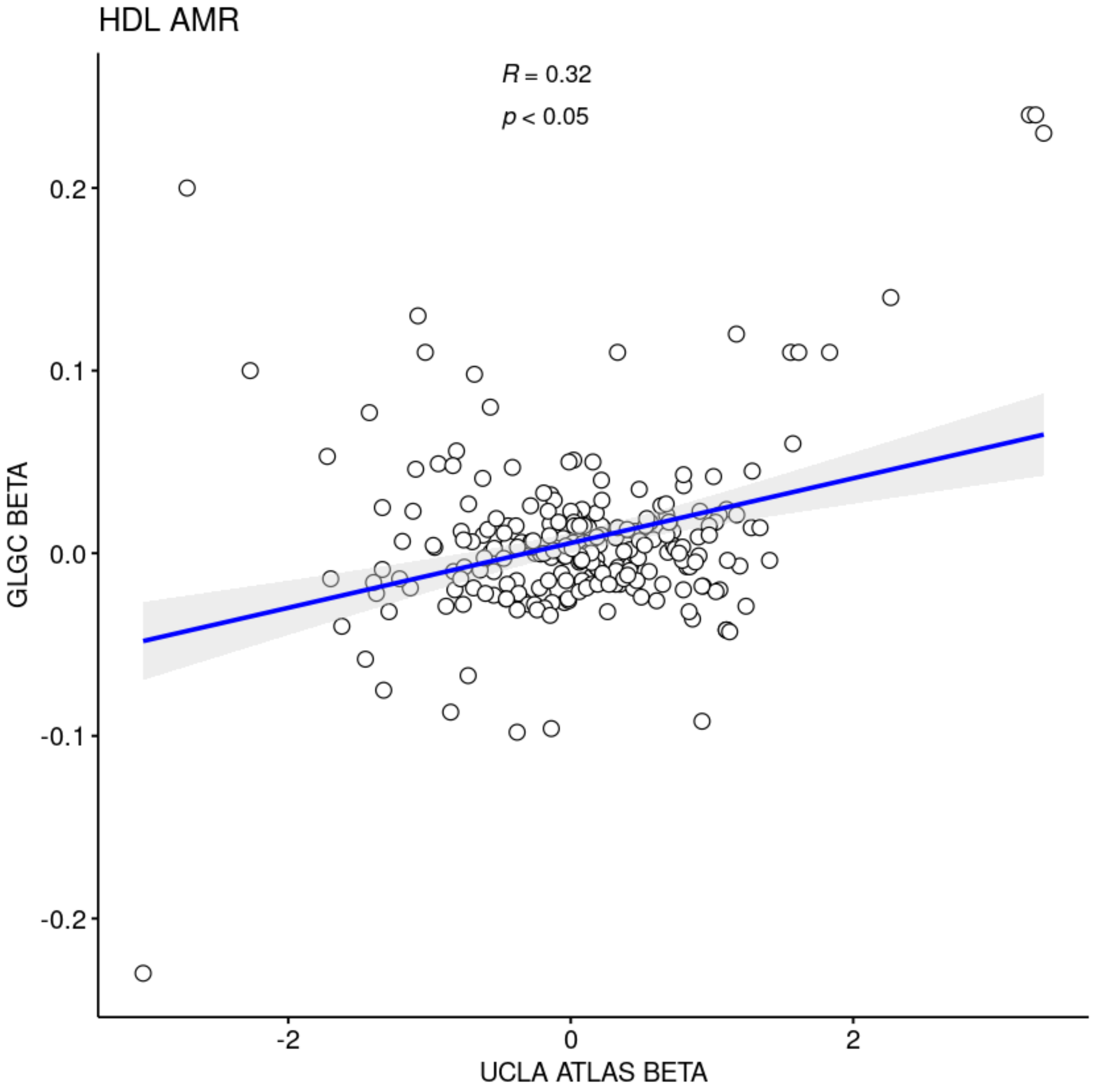


Fig S57. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for HDL-C in AMR population

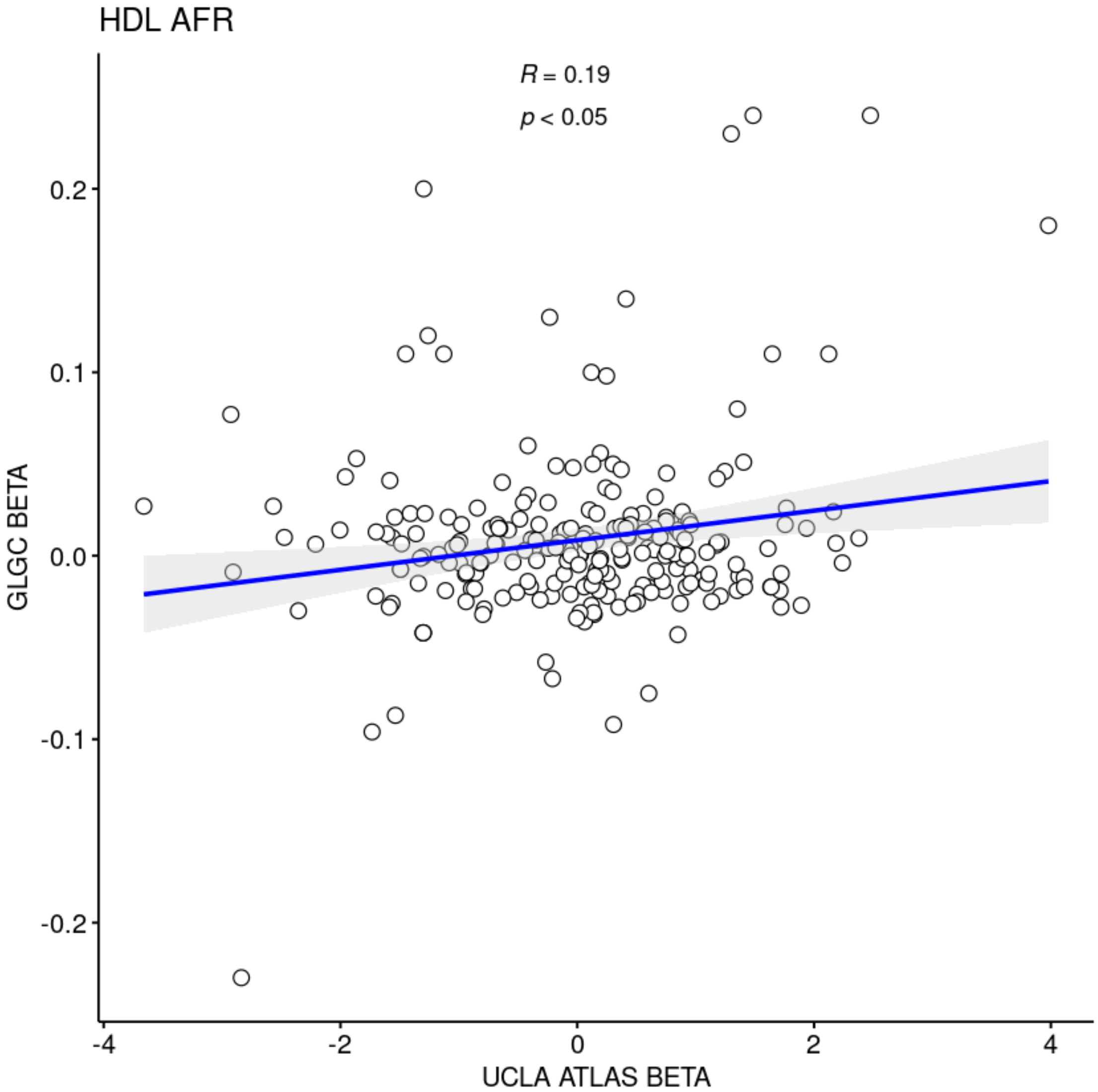


Fig S58. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for HDL-C in AFR population

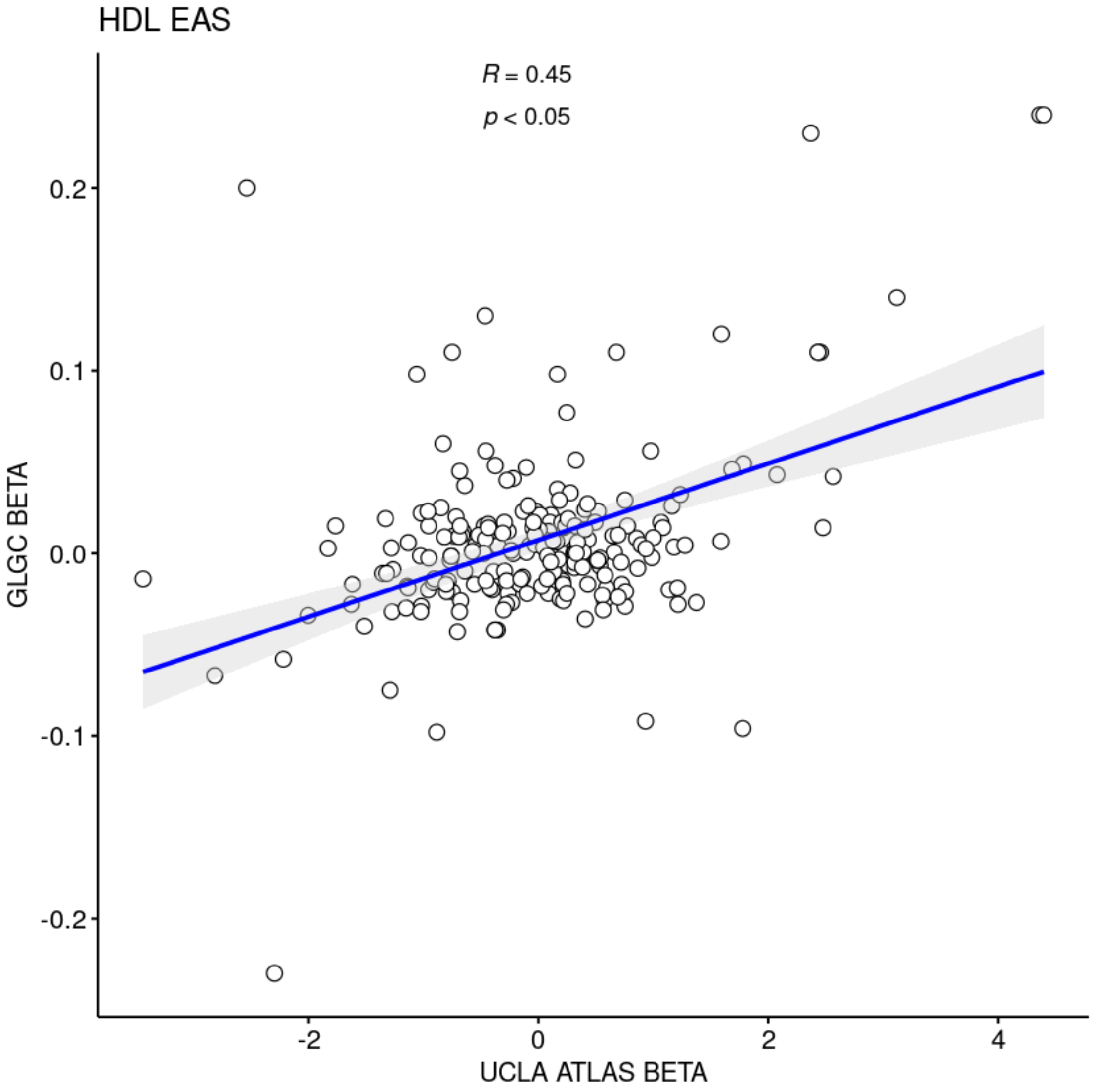


Fig S59. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for HDL-C in EAS population

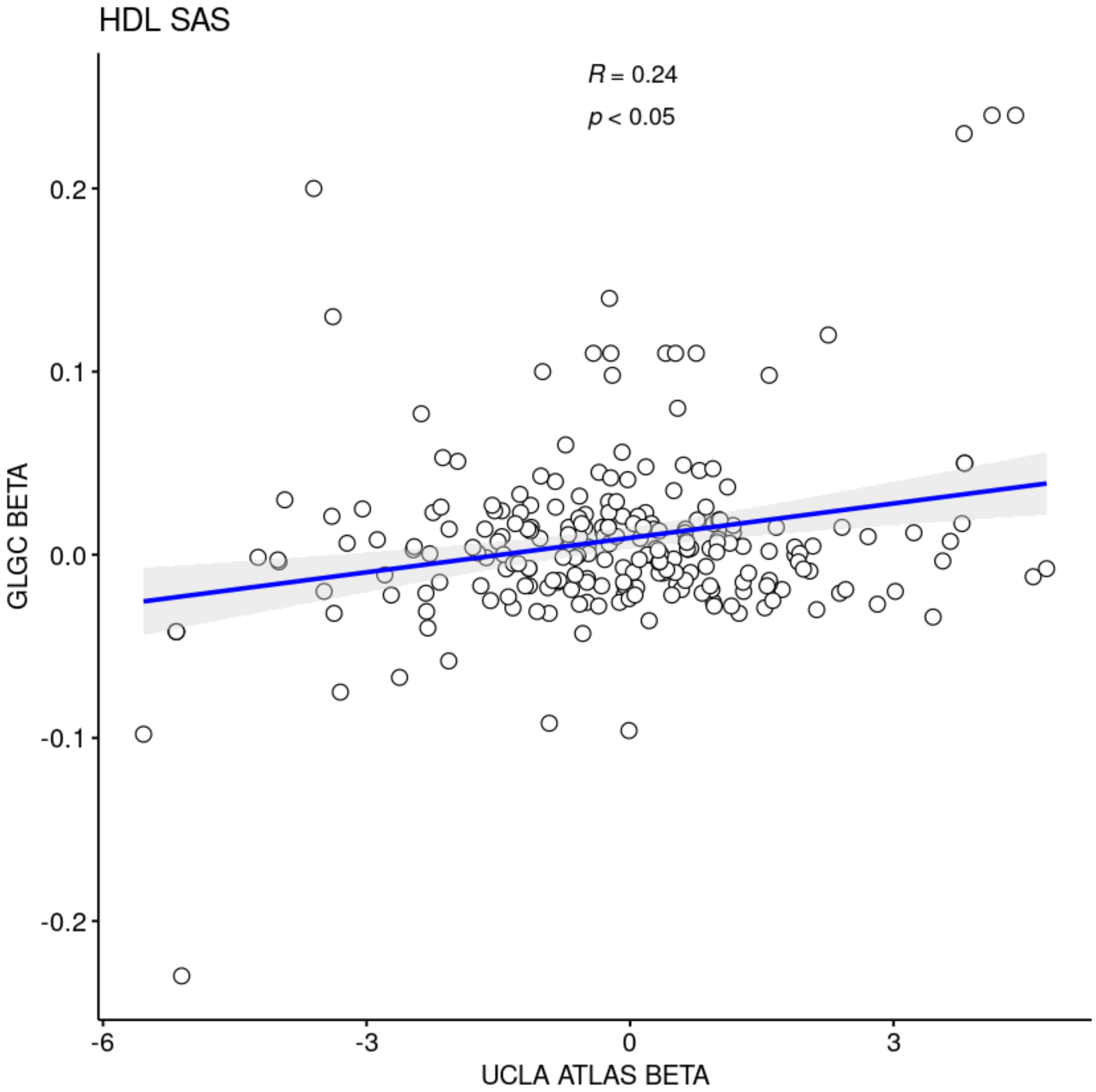


Fig S60. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for HDL-C in SAS population

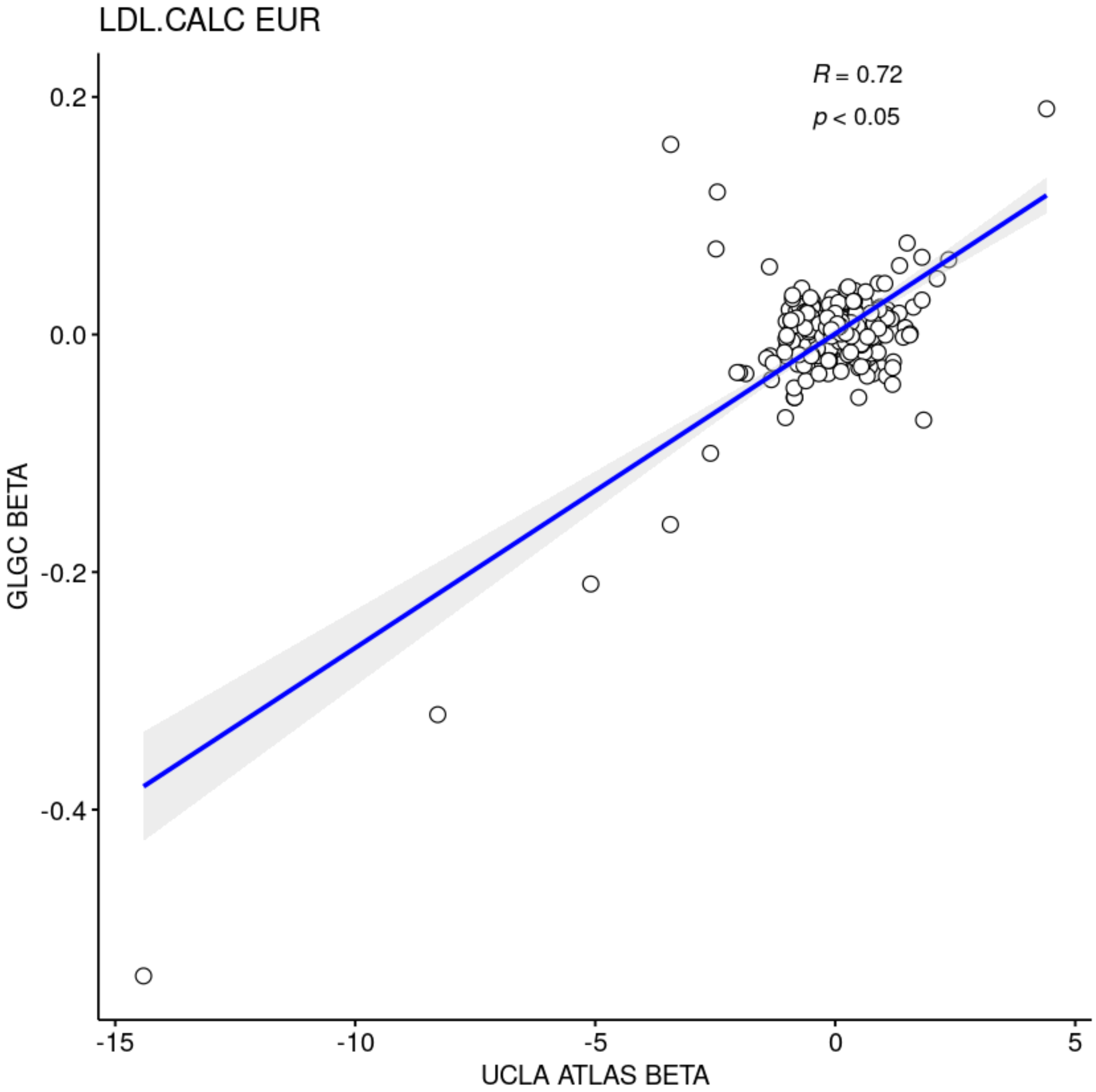


Fig S61. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for LDL Calc in EUR population

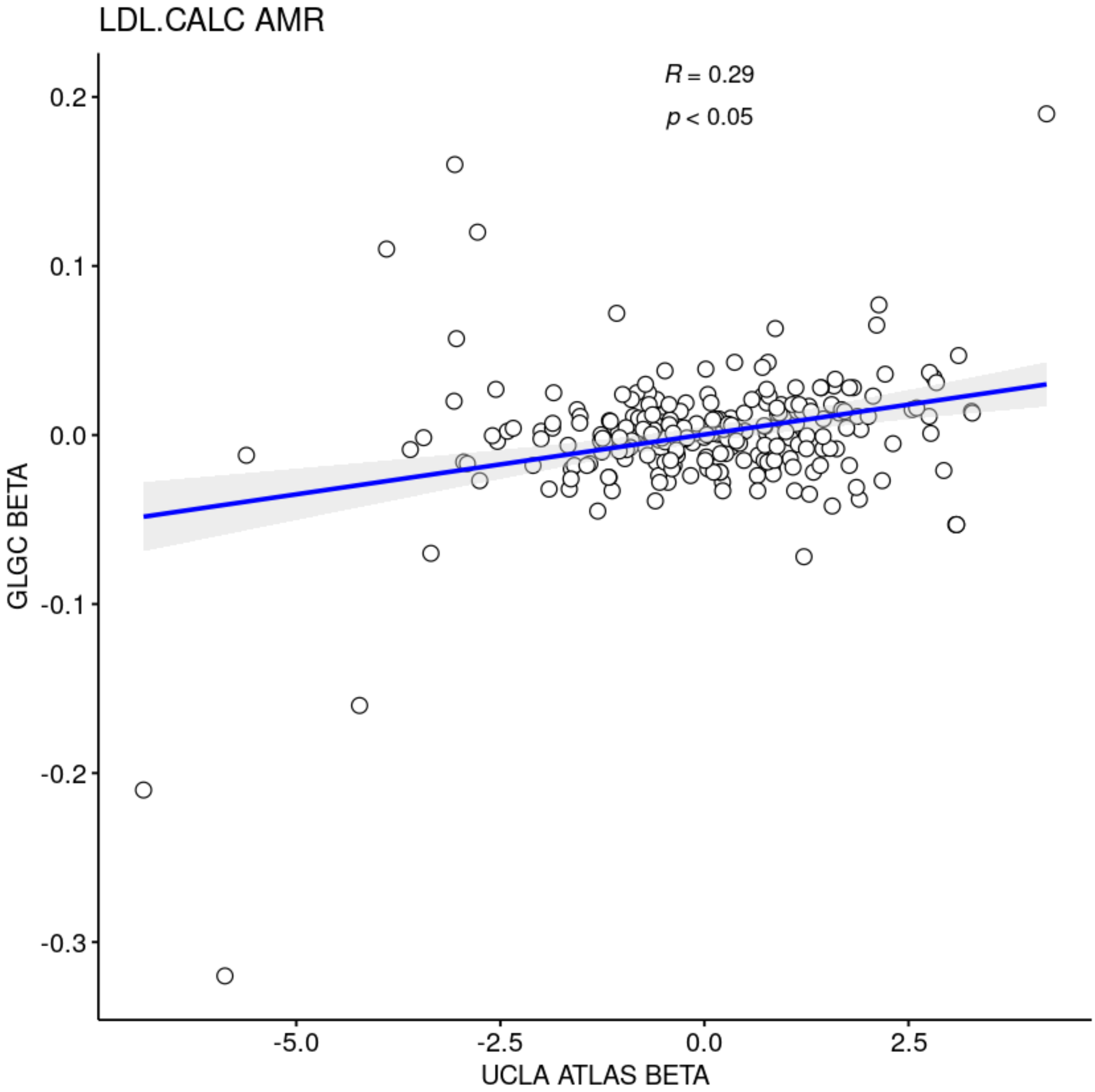


Fig S62. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for LDL Calc in AMR population

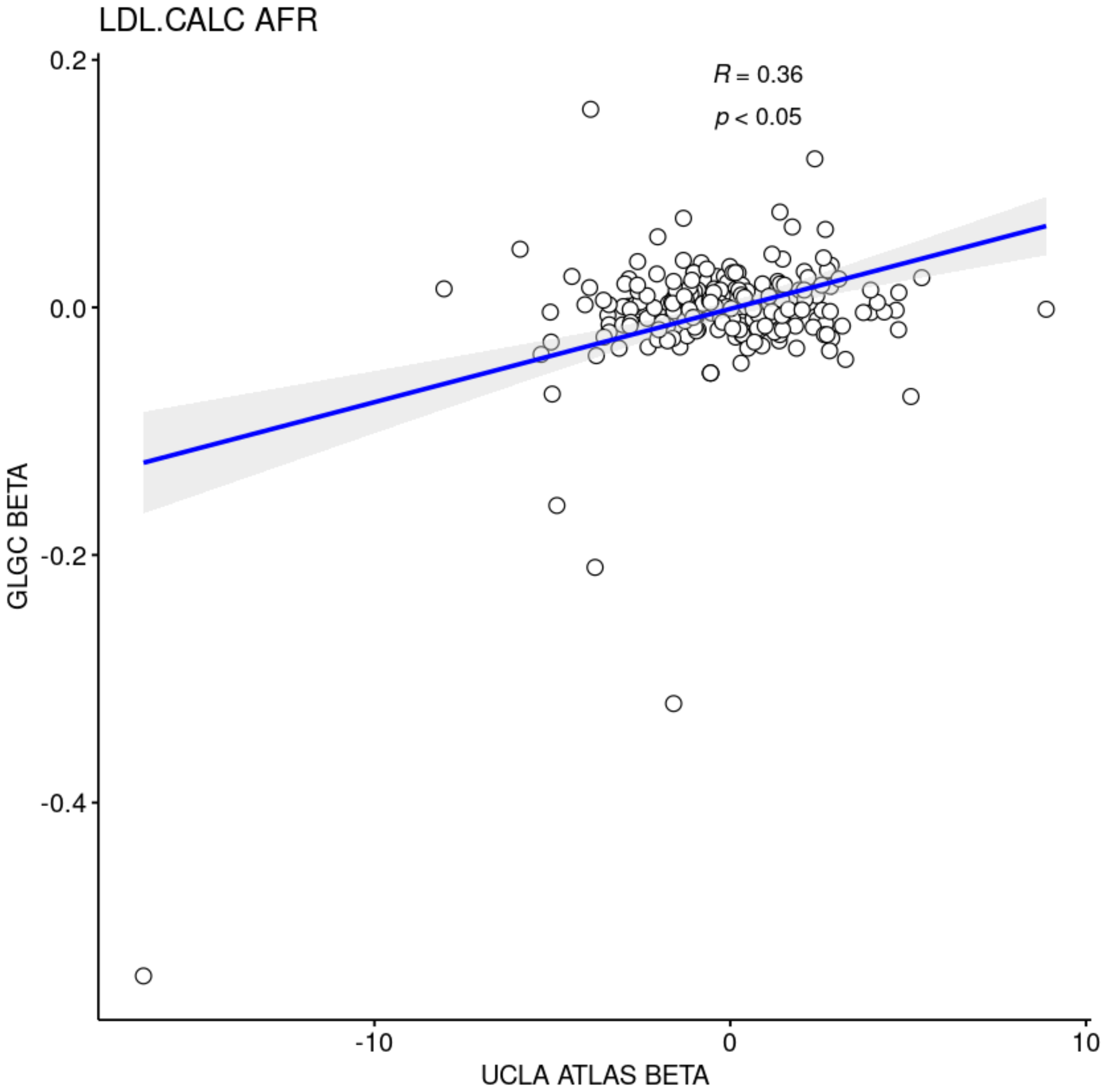


Fig S63. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for LDL Calc in AFR population

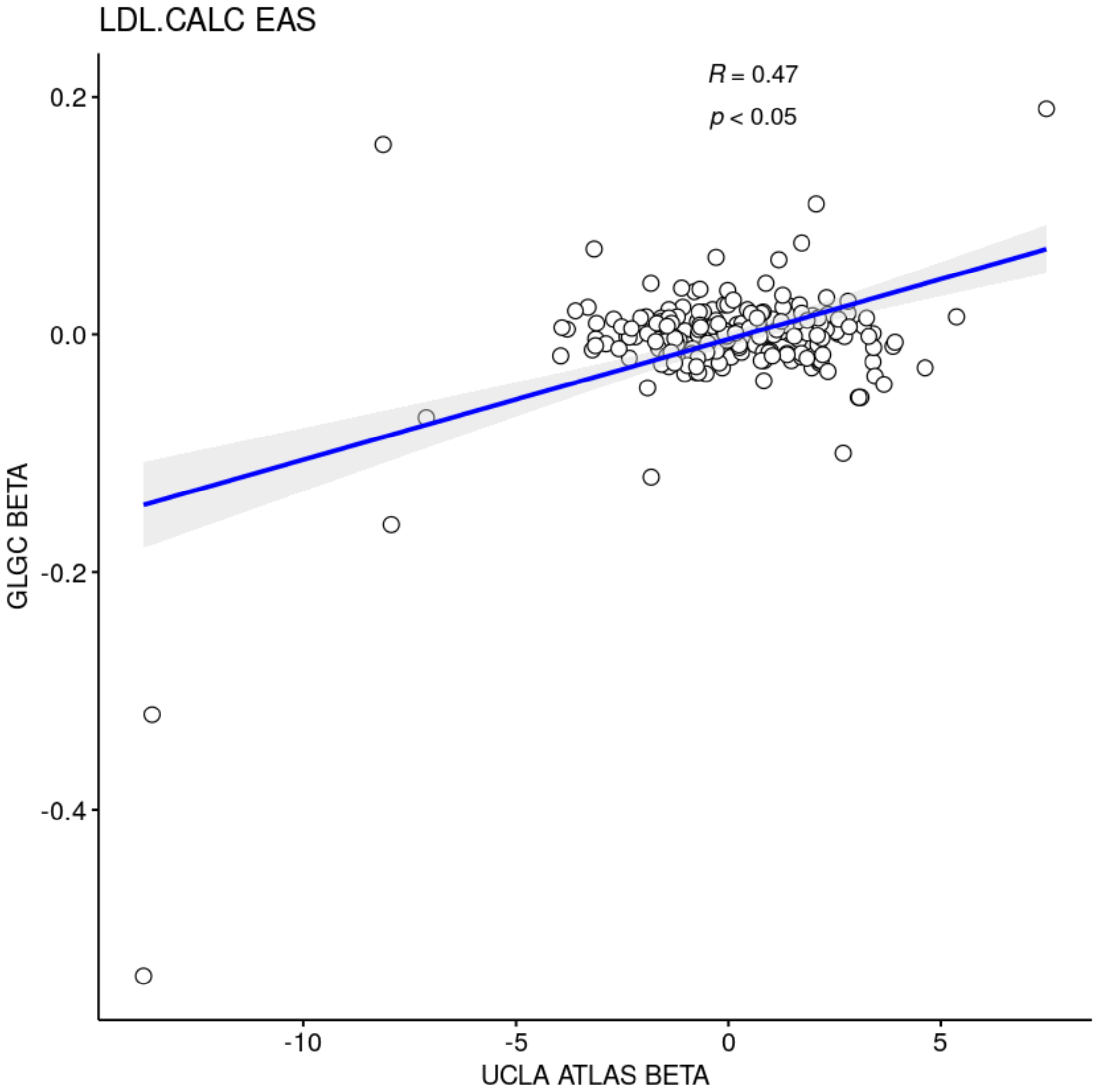


Fig S64. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for LDL Calc in EAS population

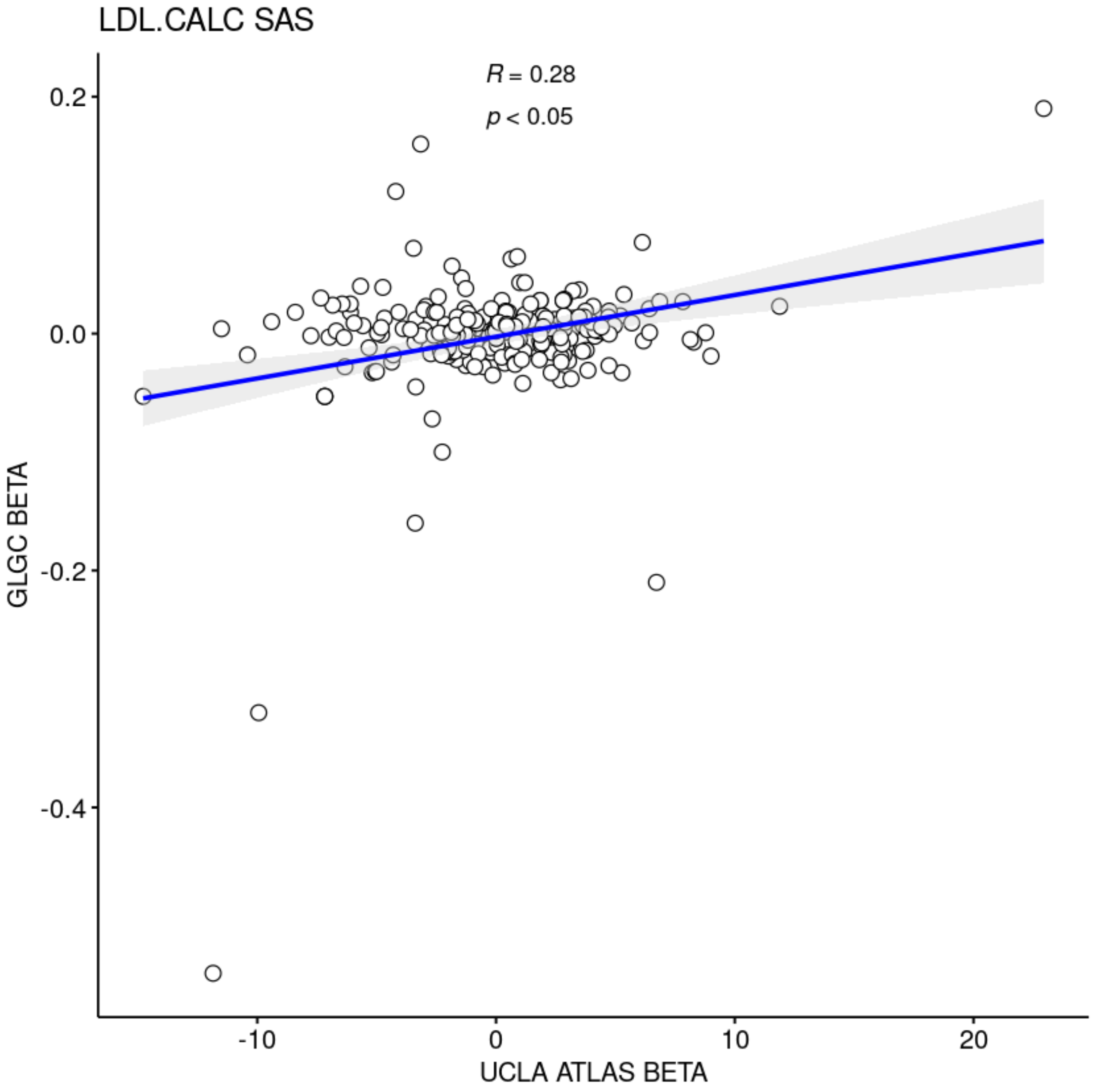


Fig S65. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for LDL Calc in SAS population

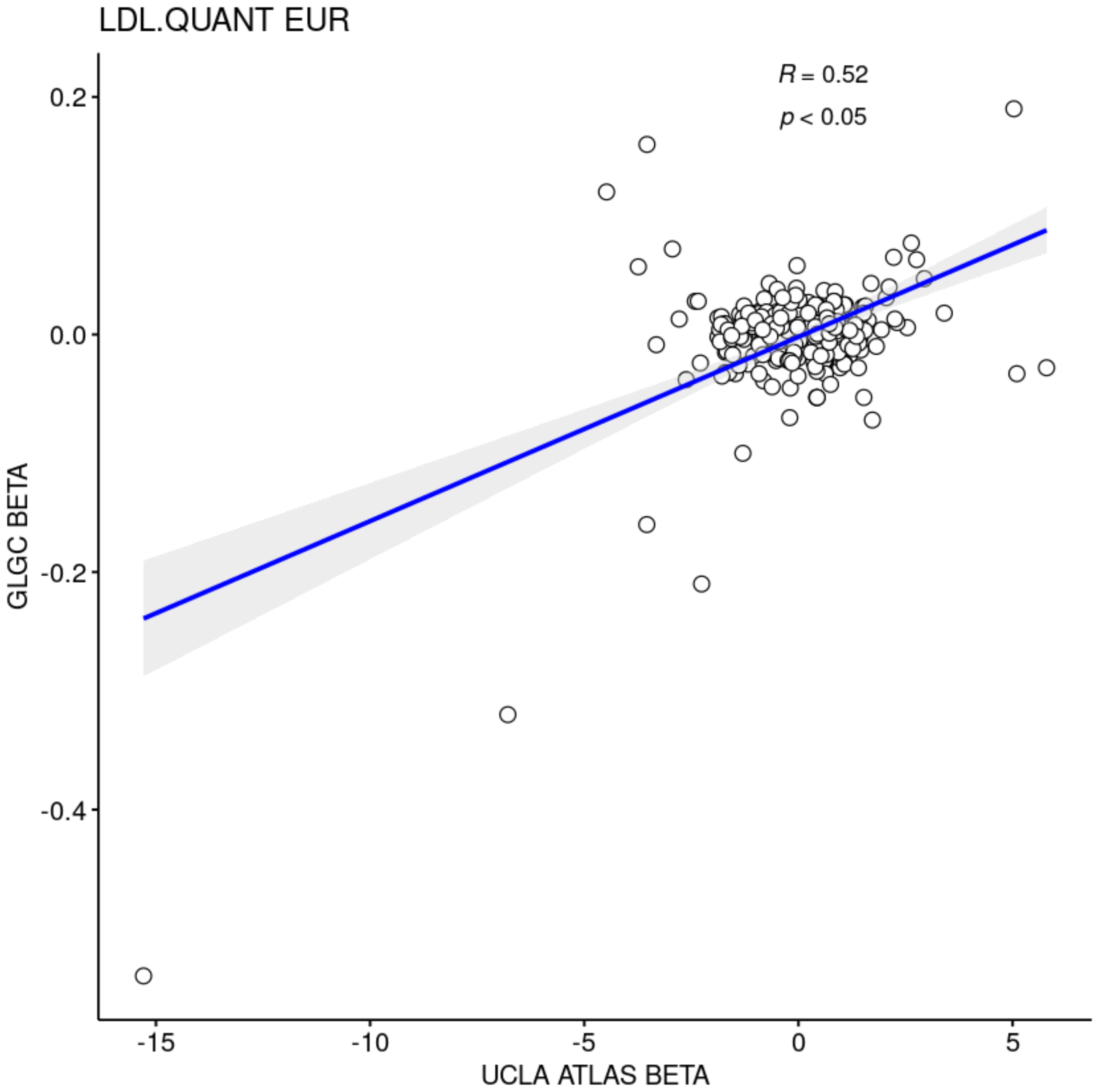


Fig S66. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for LDL Quant in EUR population

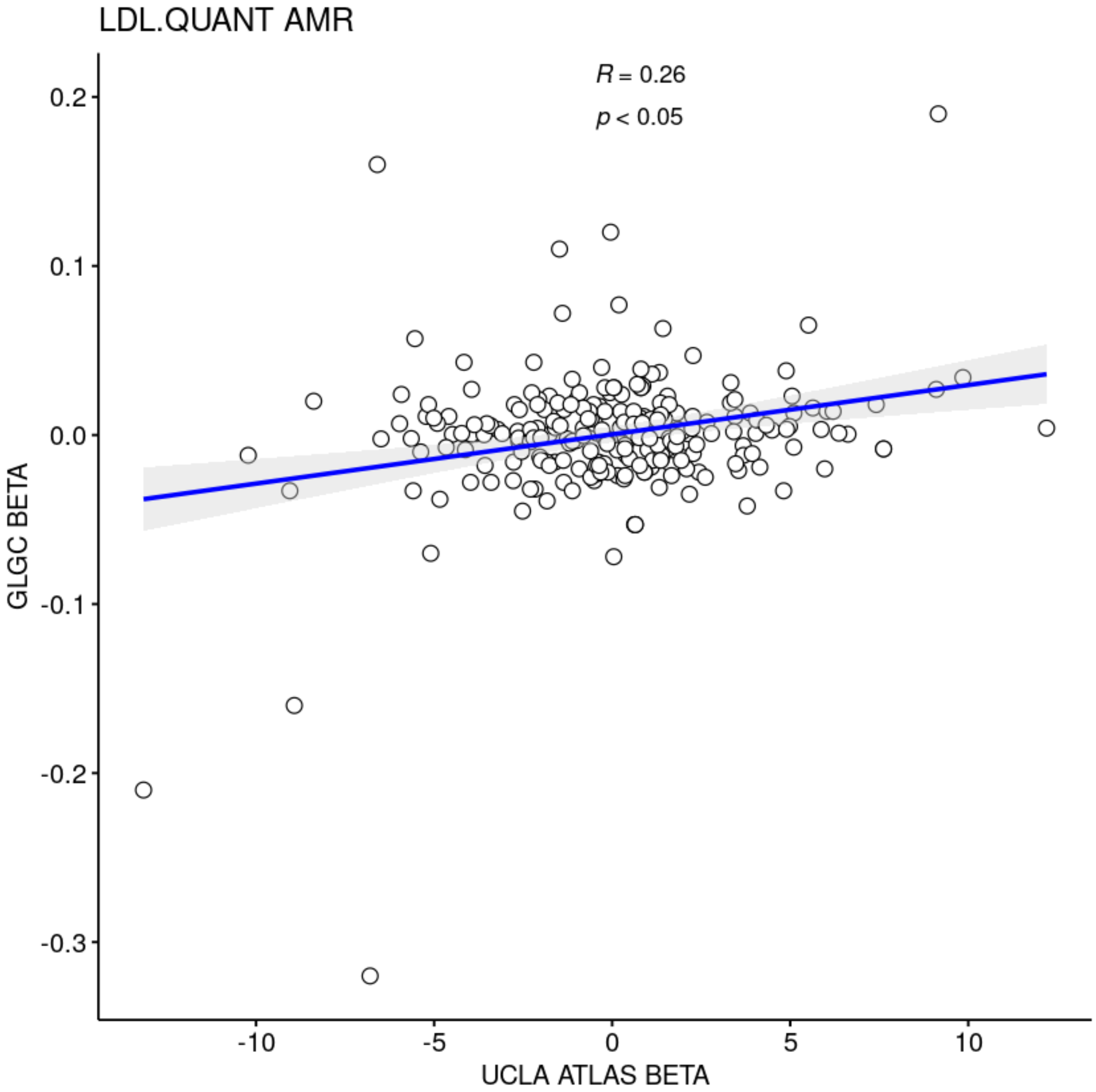


Fig S67. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for LDL Quant in AMR population

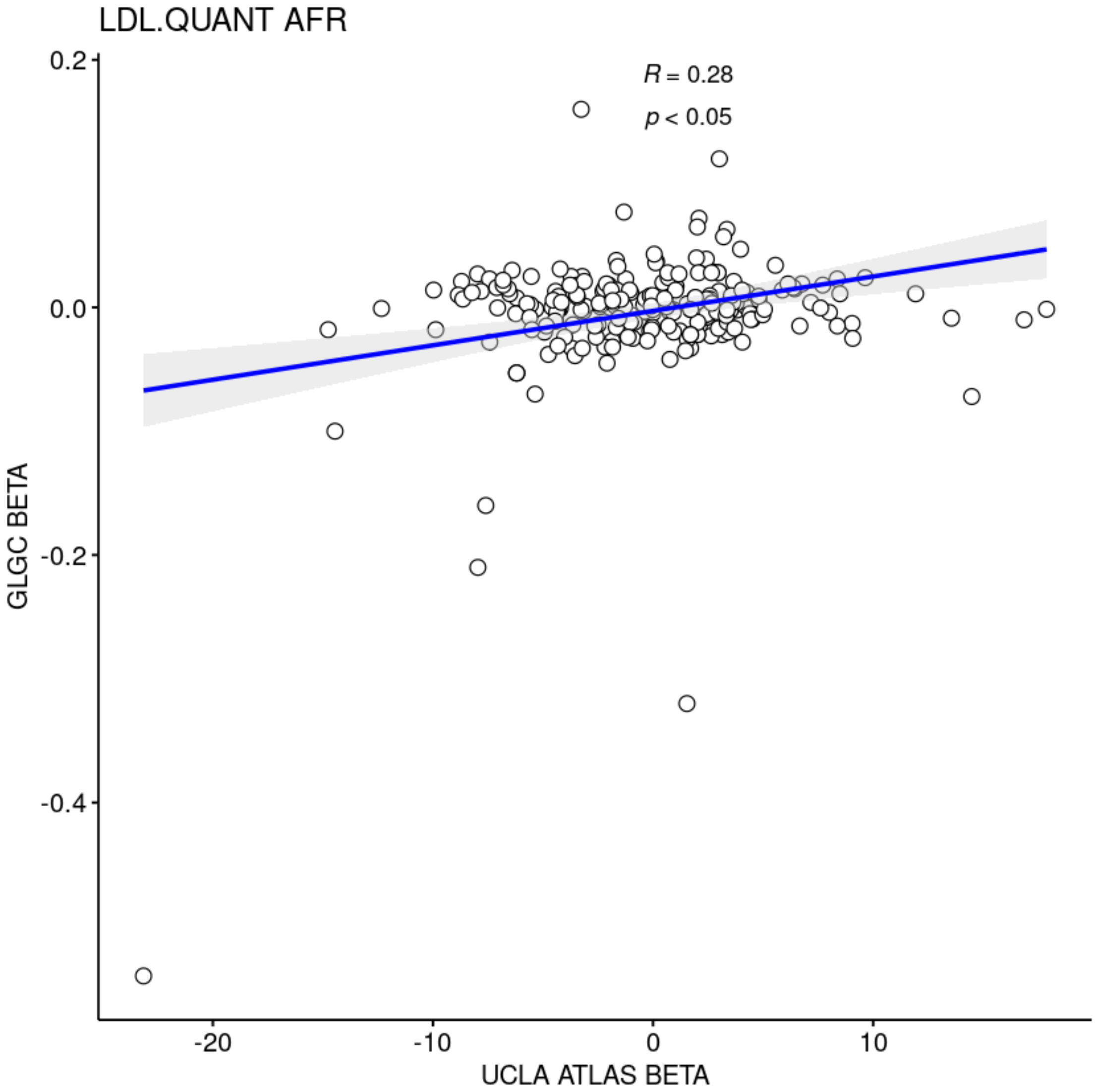


Fig S68. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for LDL Quant in AFR population

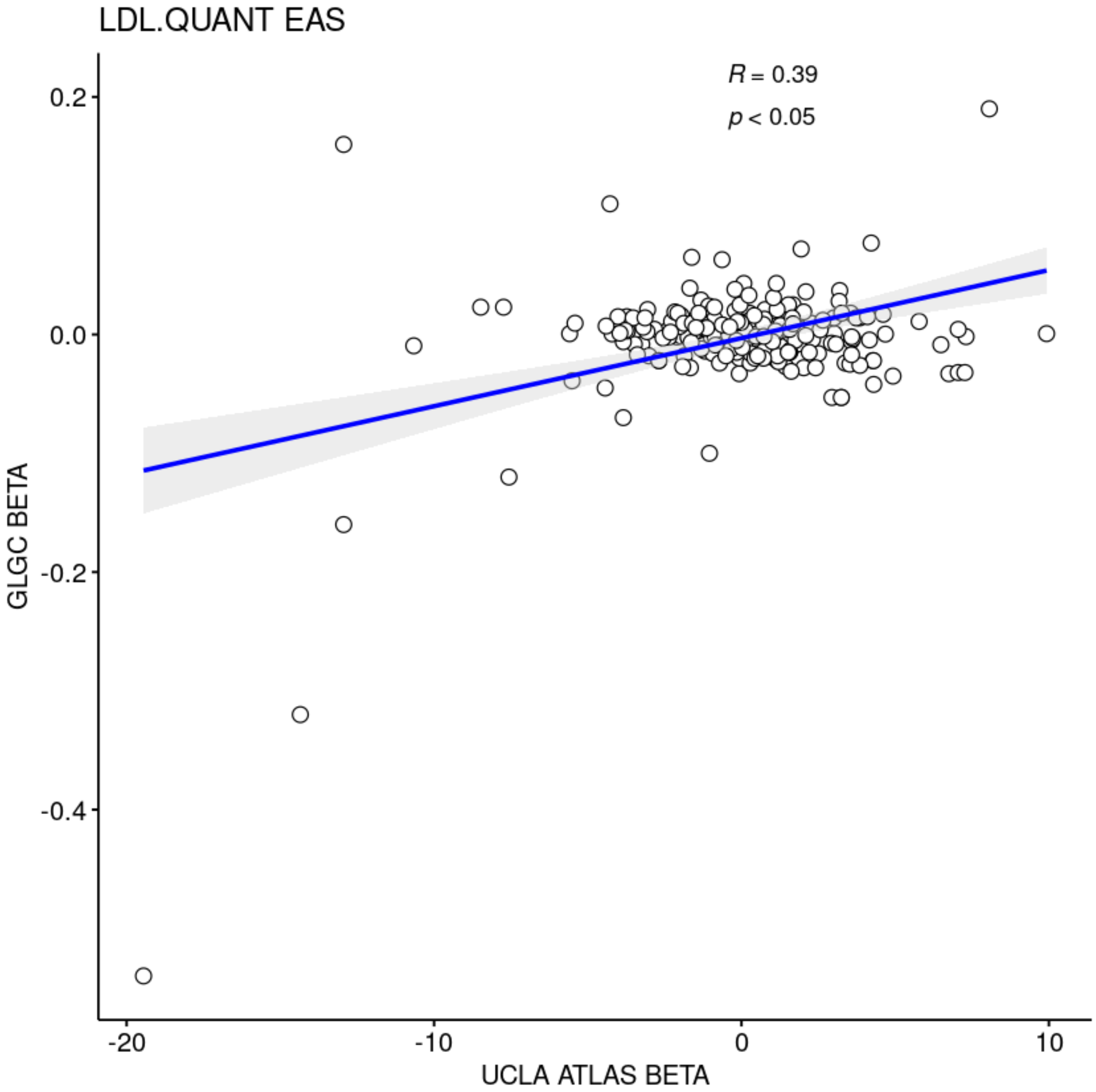


Fig S69. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for LDL Quant in EAS population

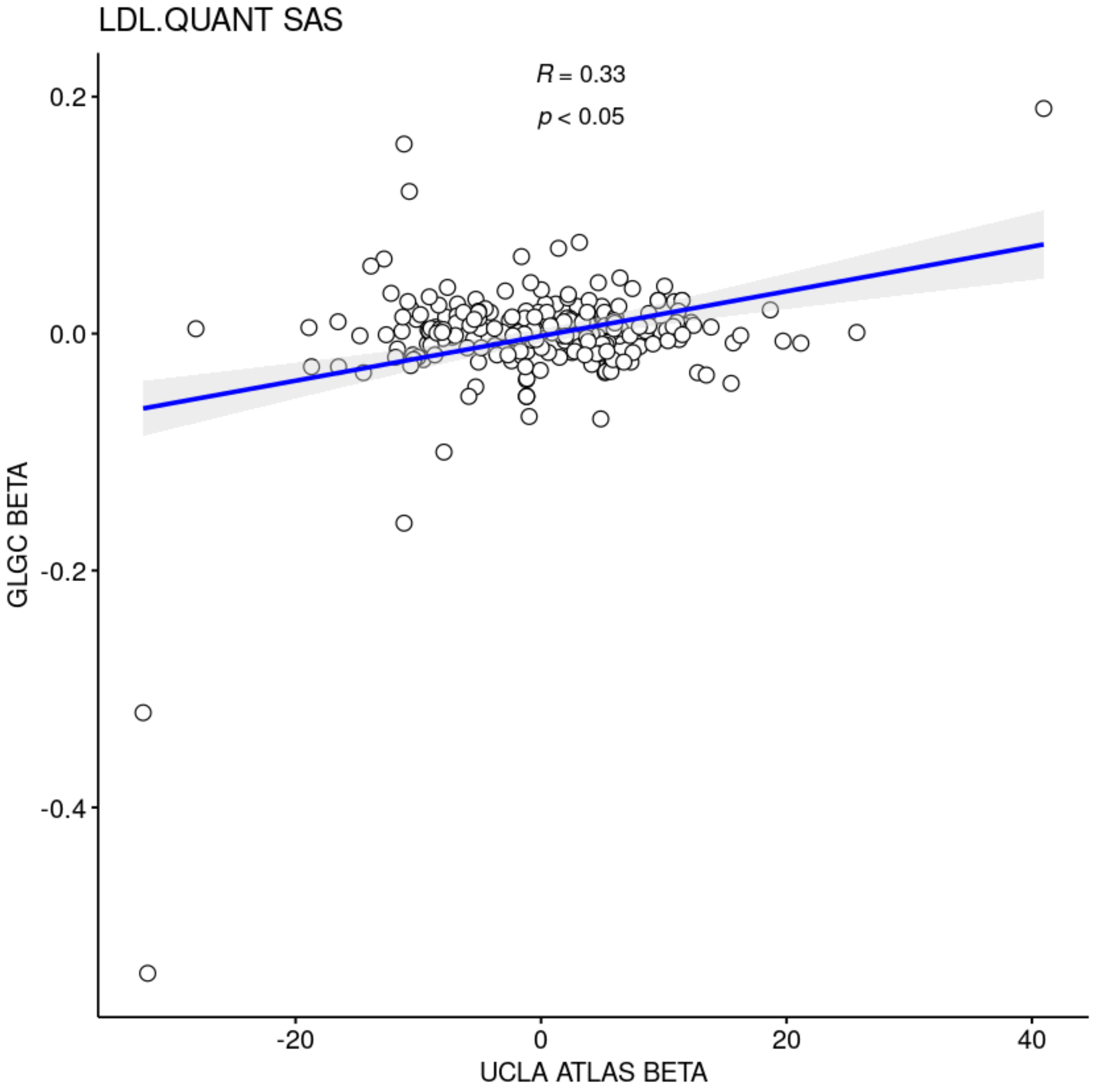


Fig S70. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for LDL Quant in SAS population

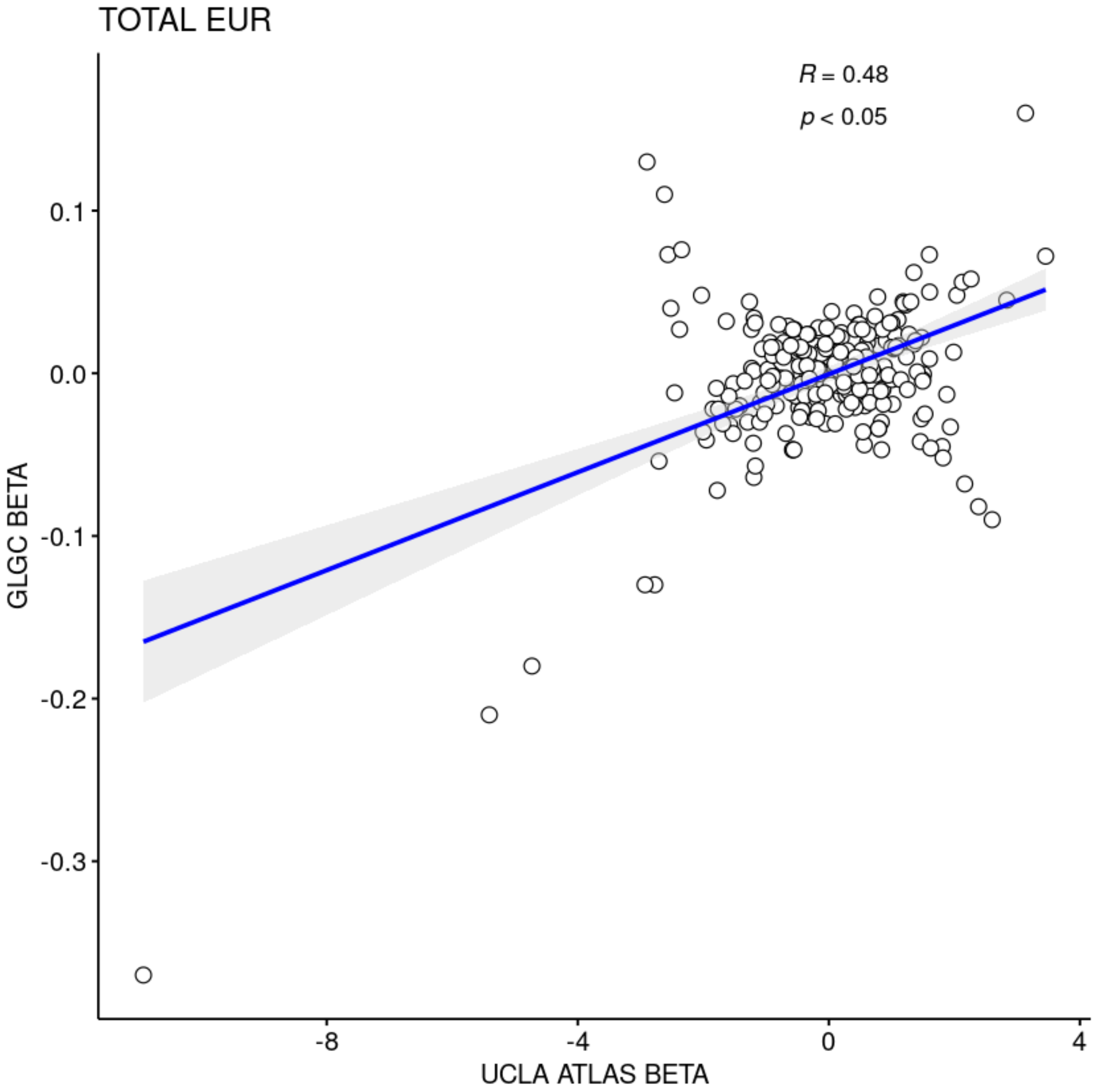


Fig S71. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for TC in EUR population

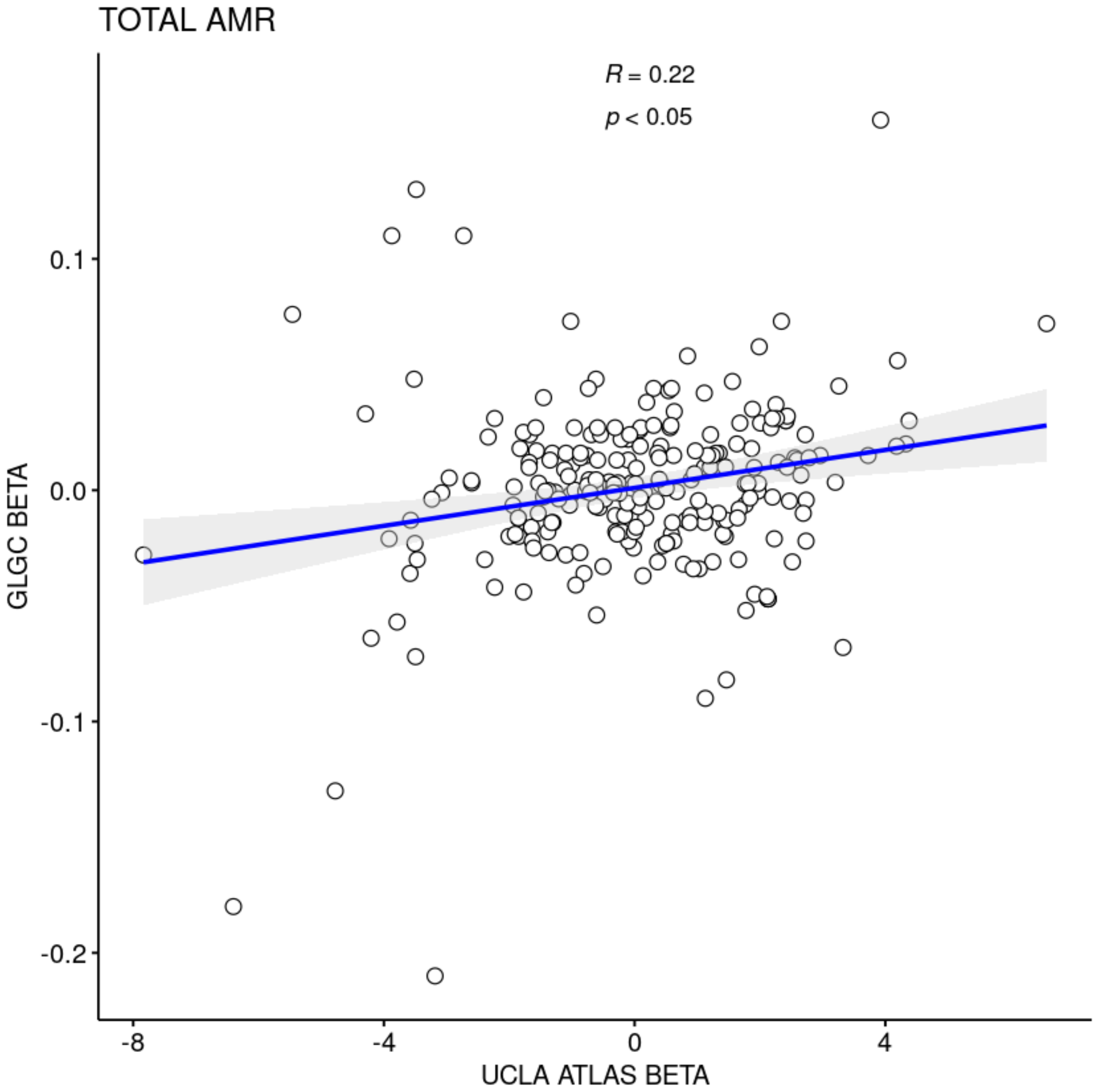


Fig S72. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for TC in AMR population

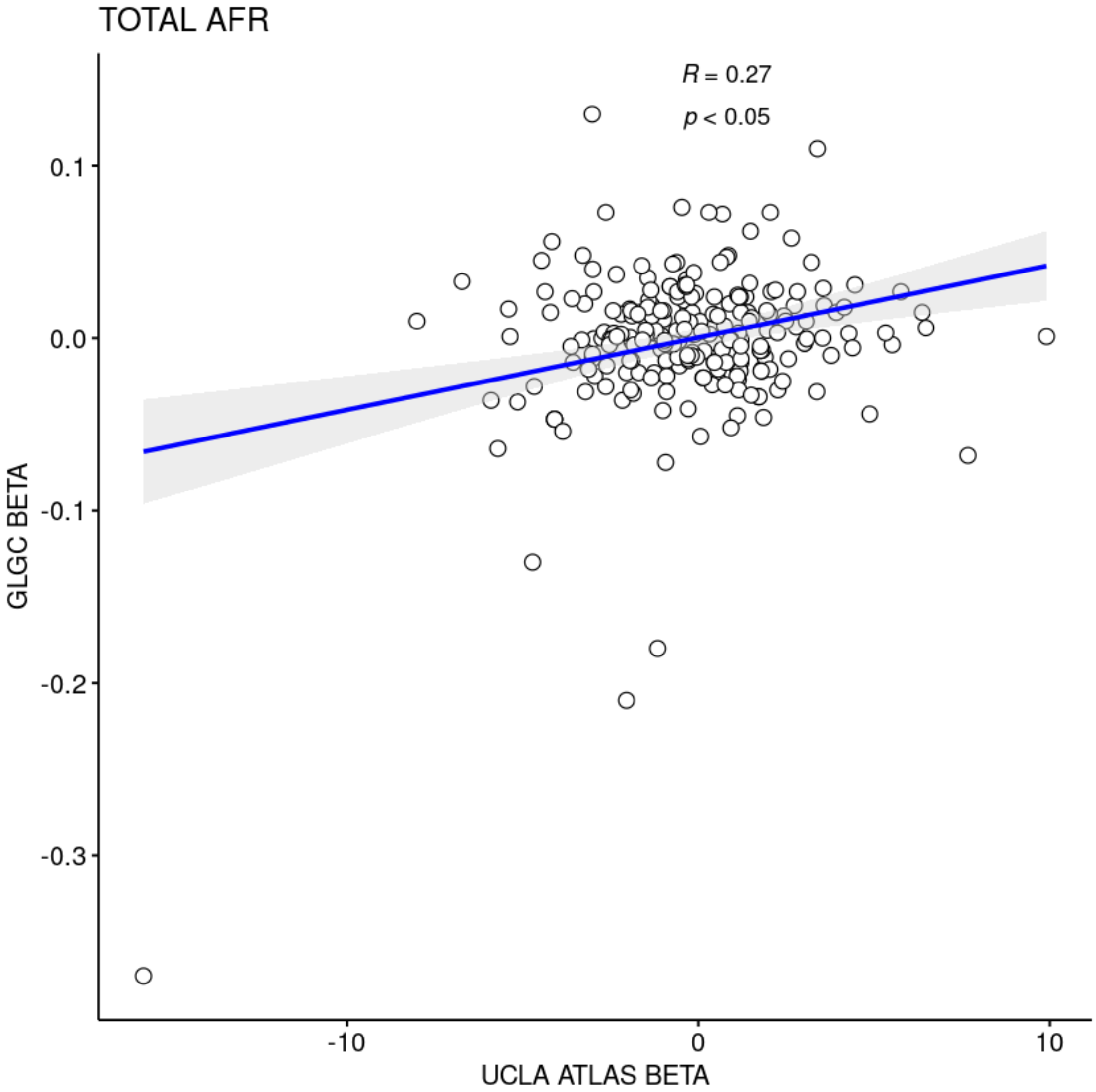


Fig S73. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for TC in AFR population

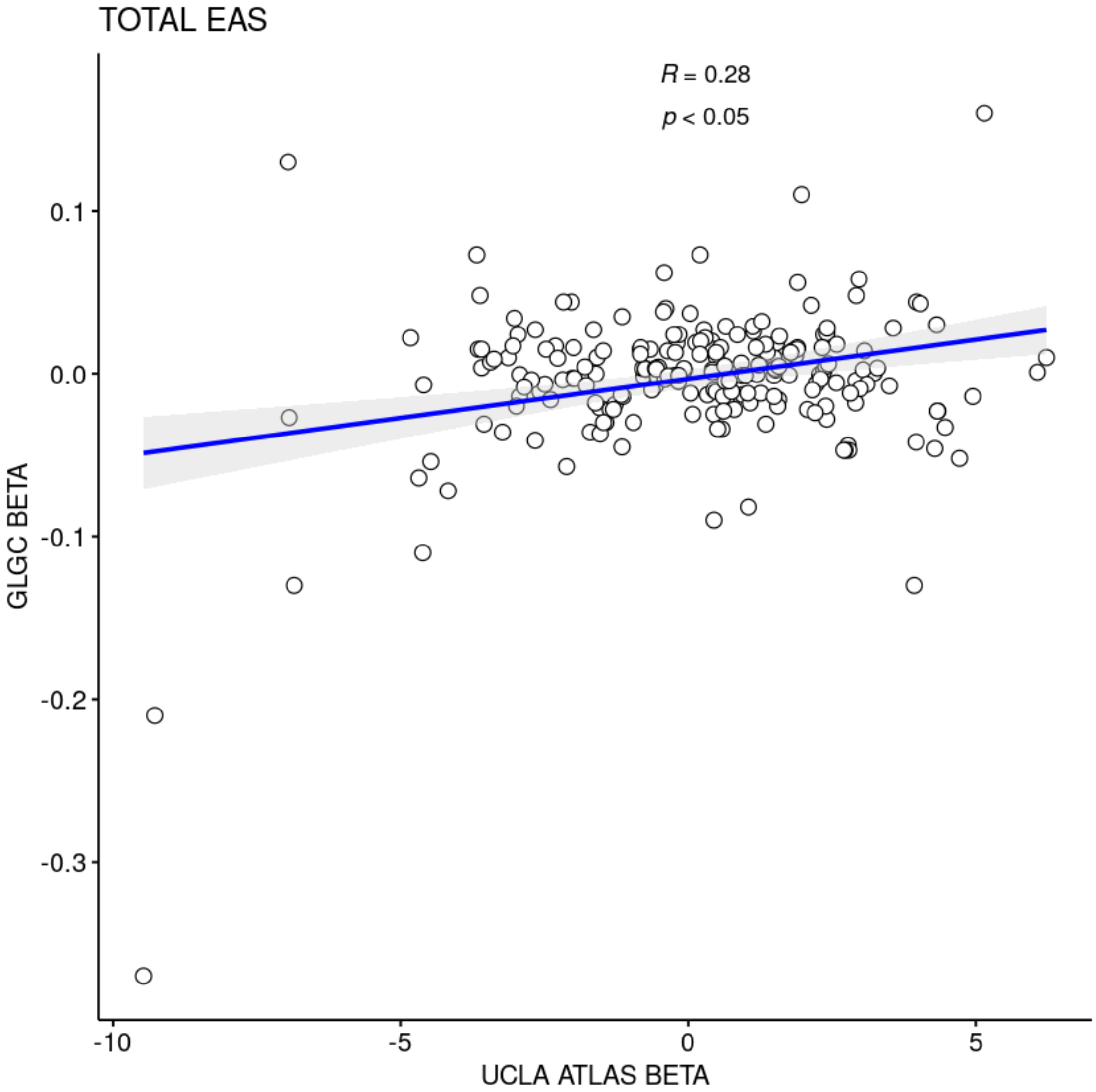


Fig S74. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for TC in EAS population

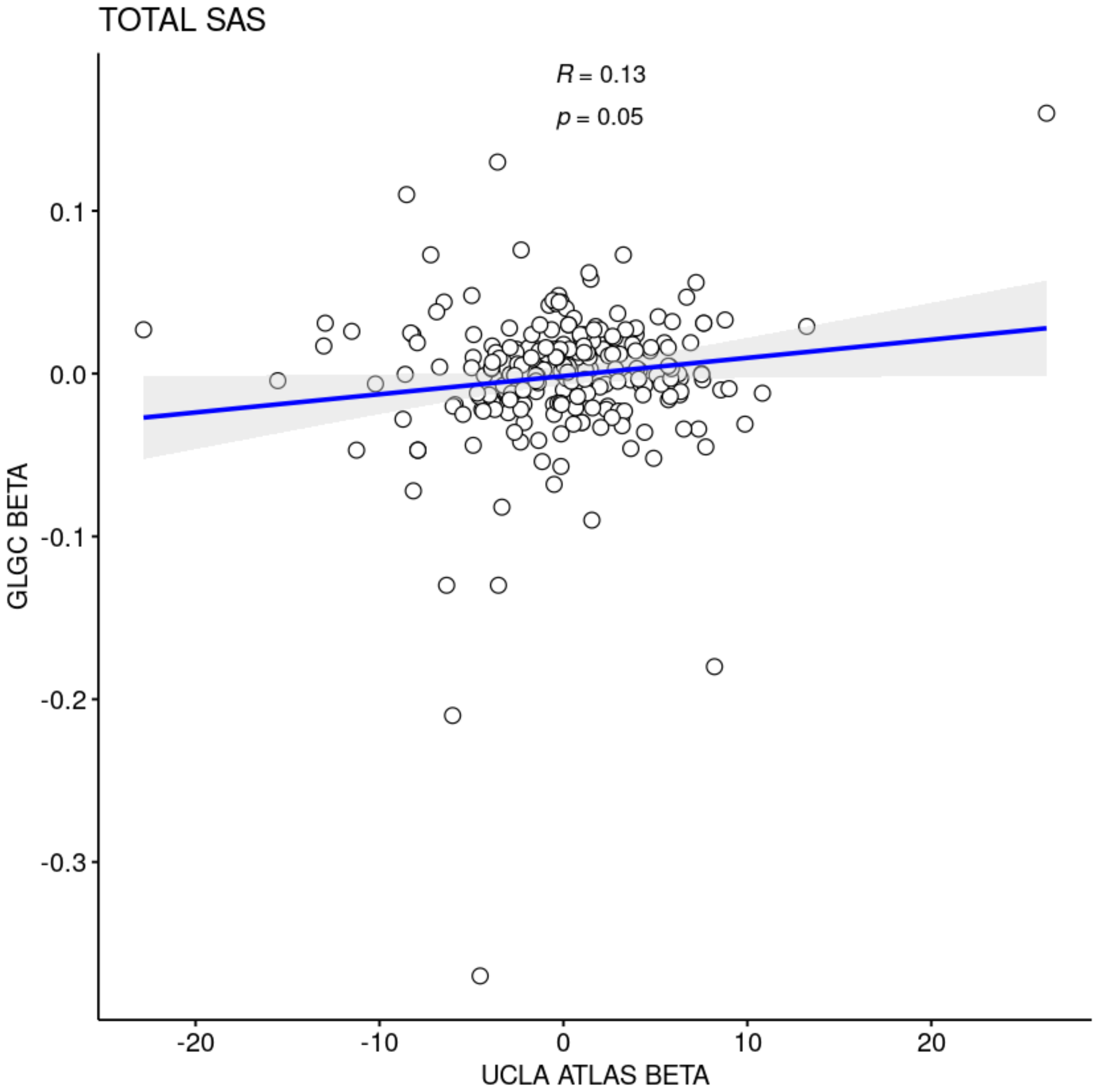


Fig S75. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for TC in SAS population

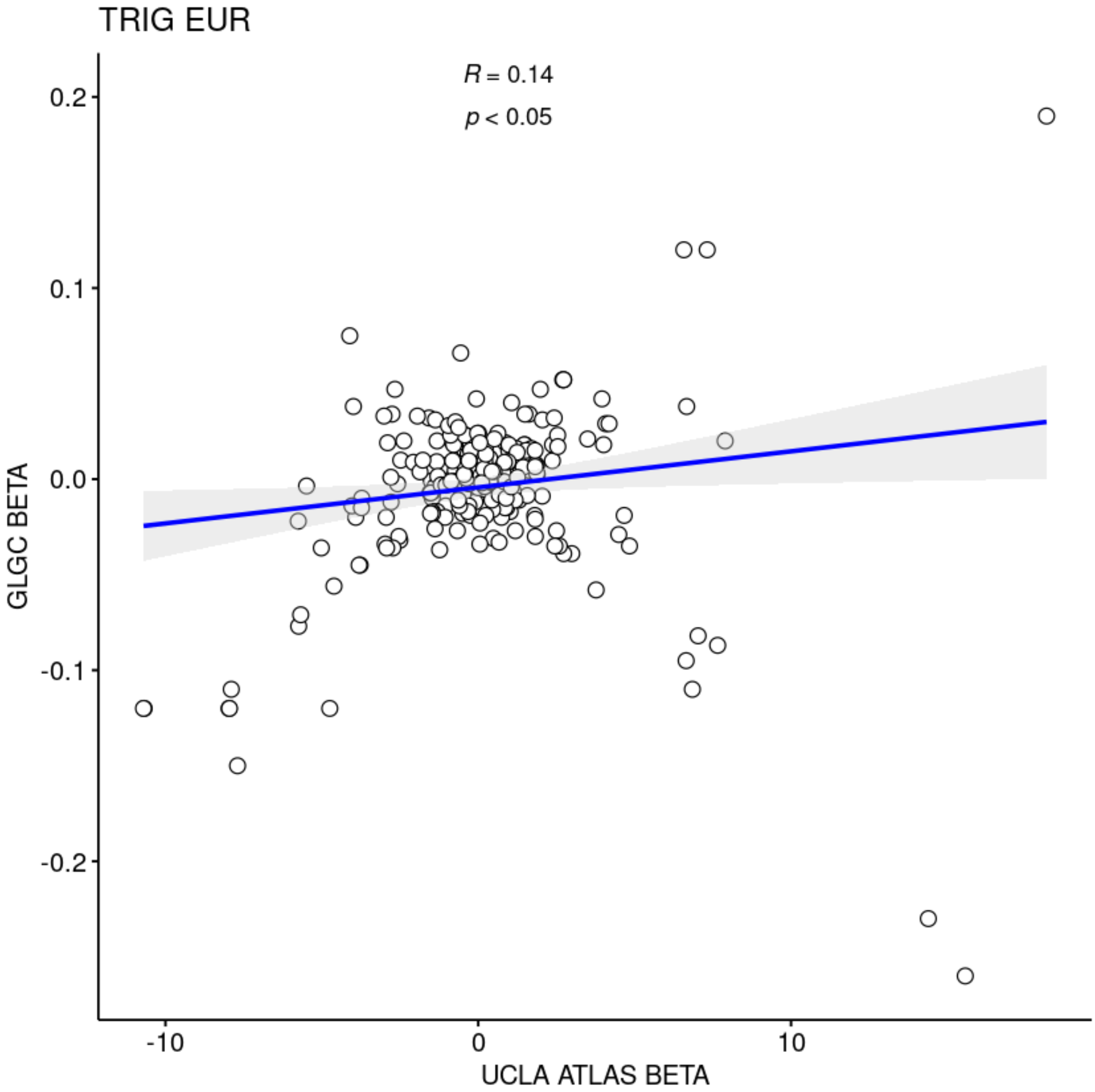


Fig S76. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for TG in EUR population

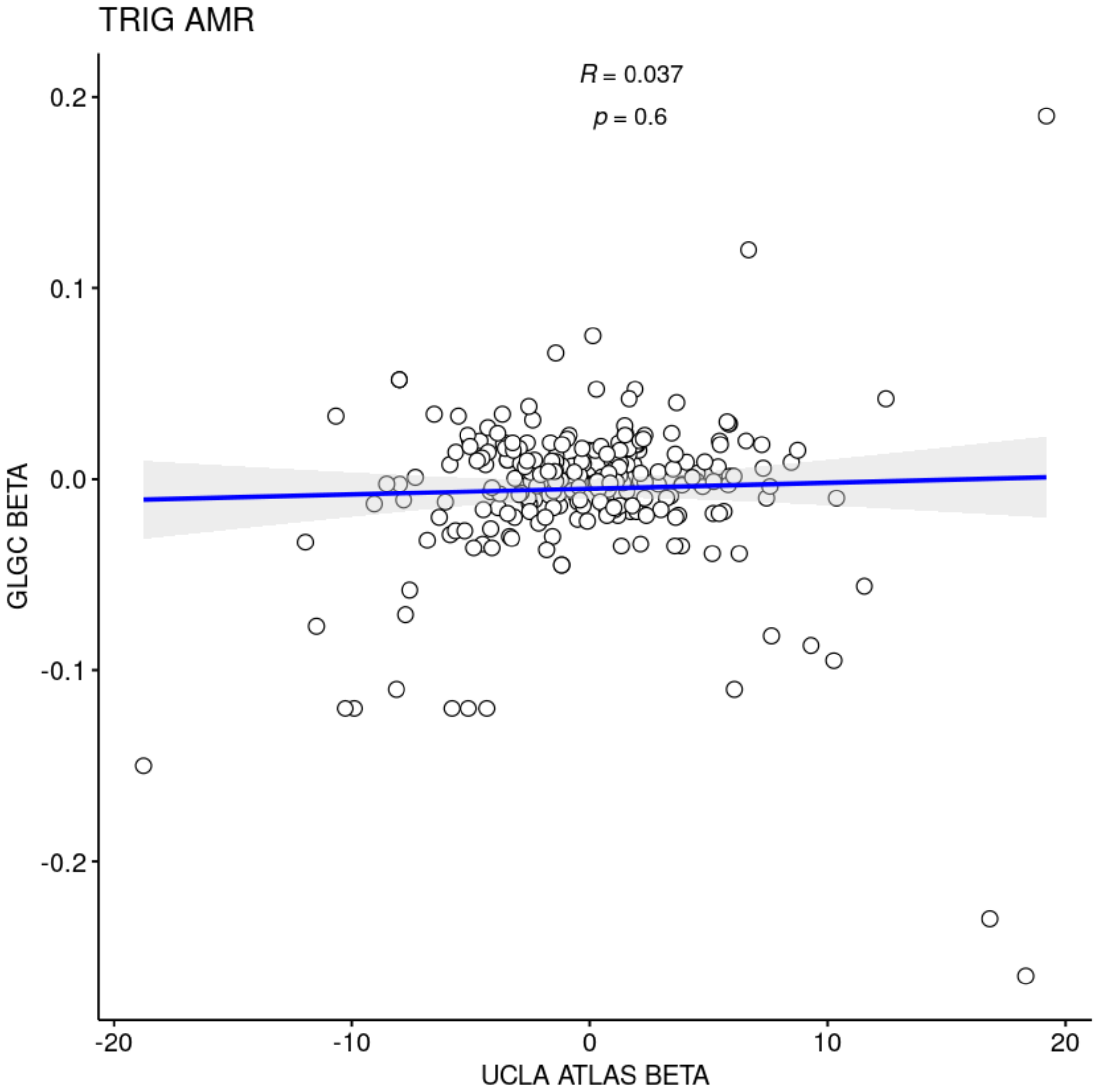


Fig S77. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for TG in AMR population

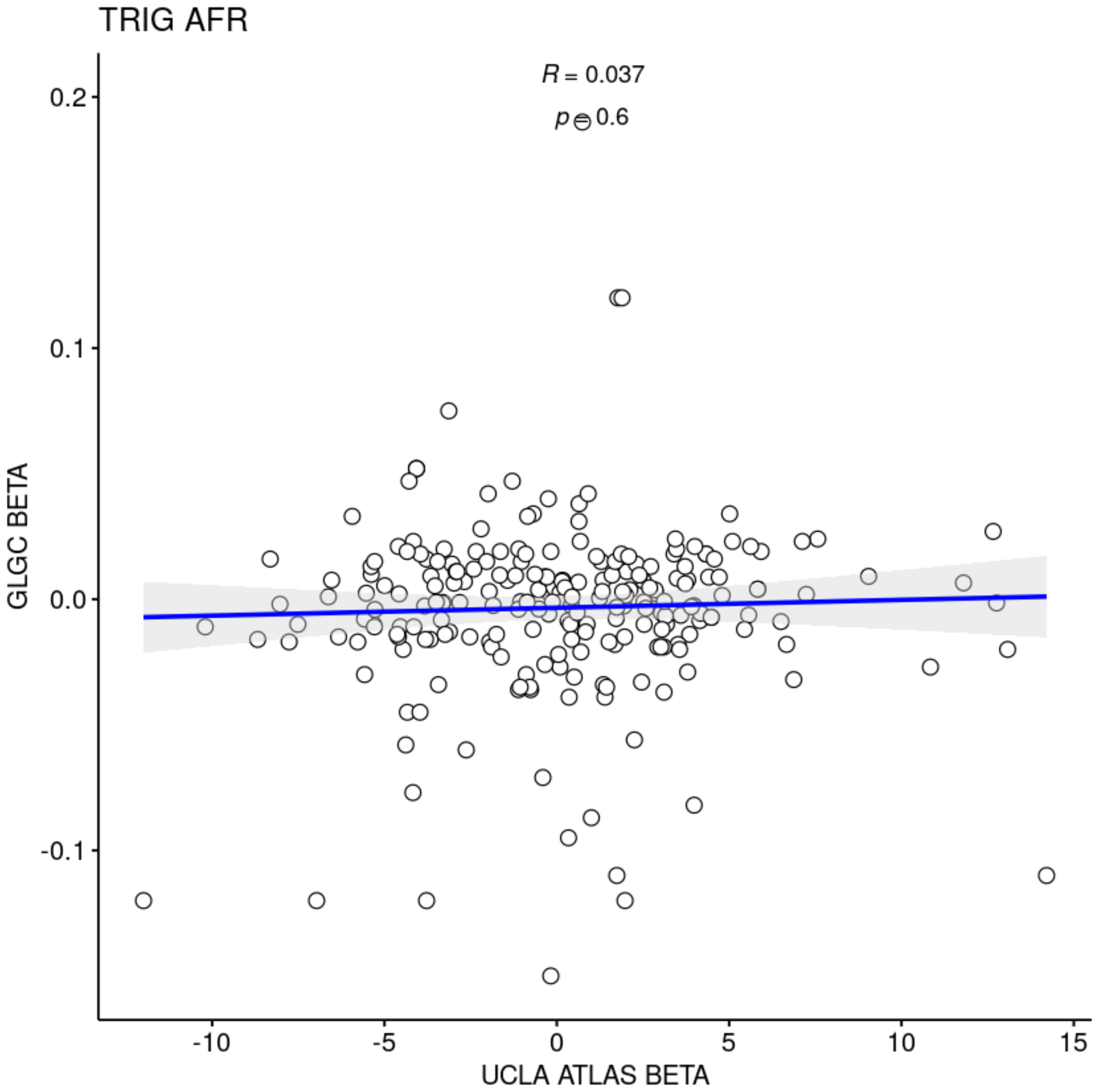


Fig S78. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for TG in AFR population

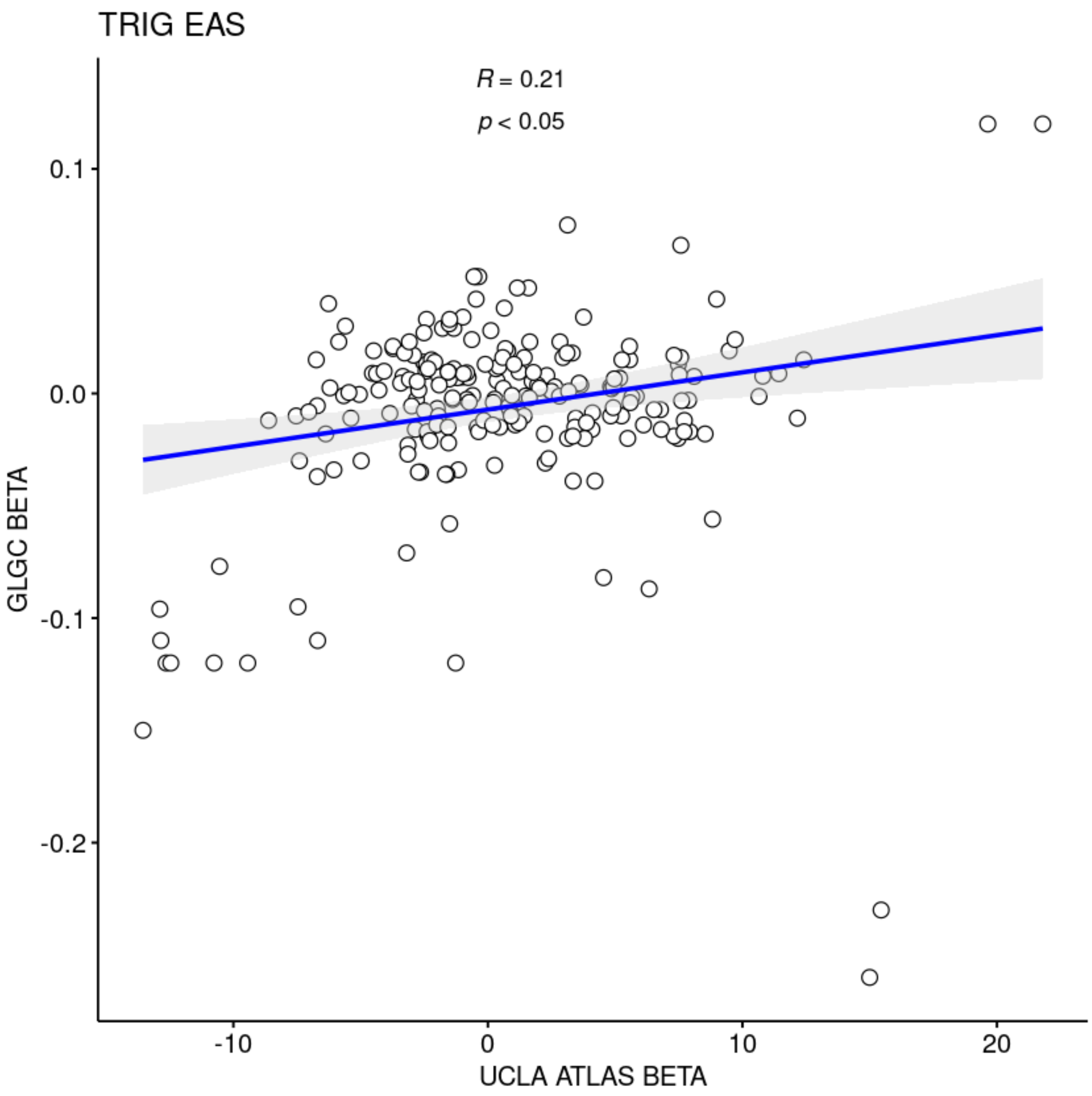


Fig S79. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for TG in EAS population

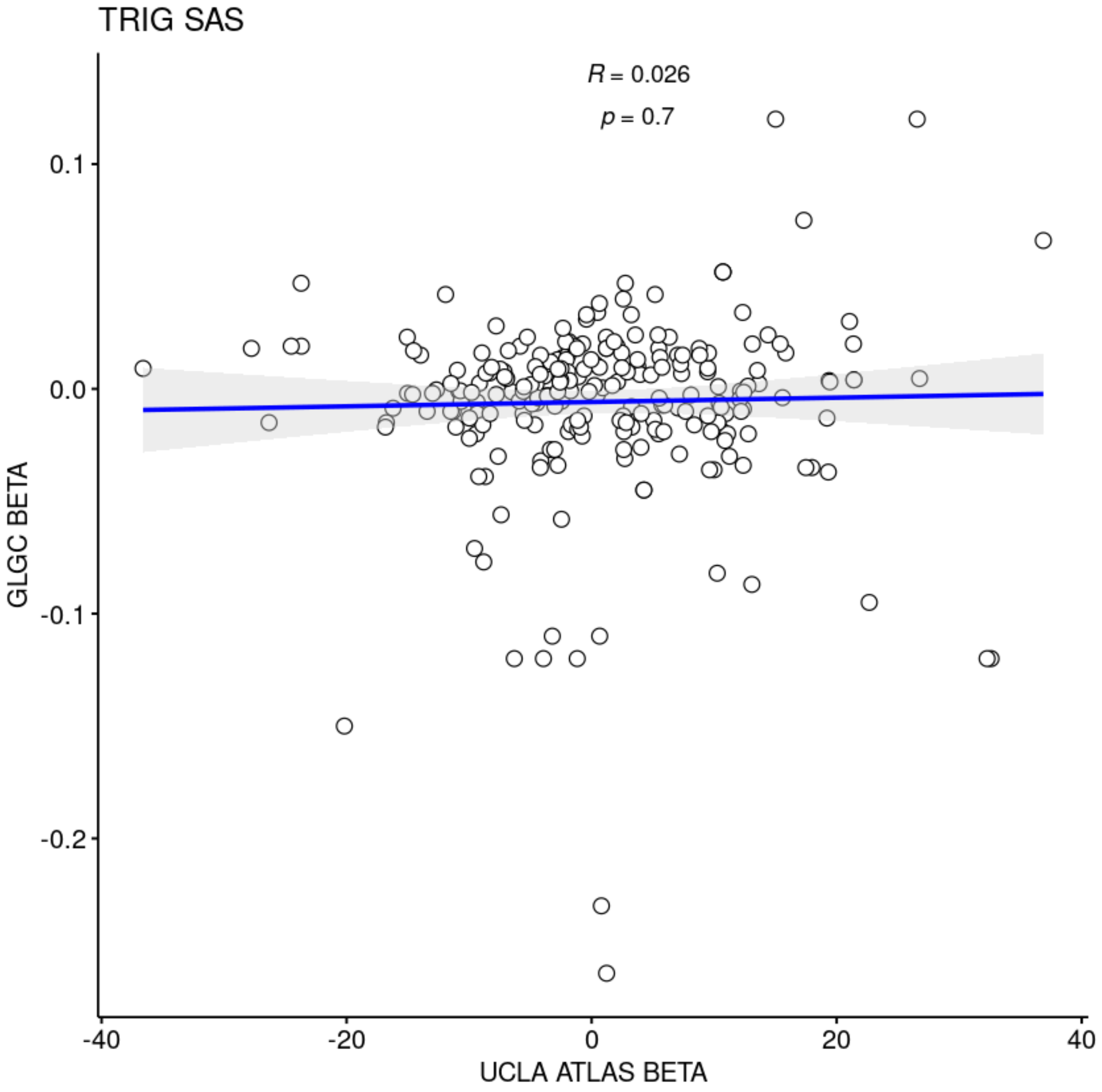


Fig S80. Distribution of effect sizes of known GWAS hits reported by GLGC(11) and estiamted in UCLA ATLAS dataset for TG in SAS population

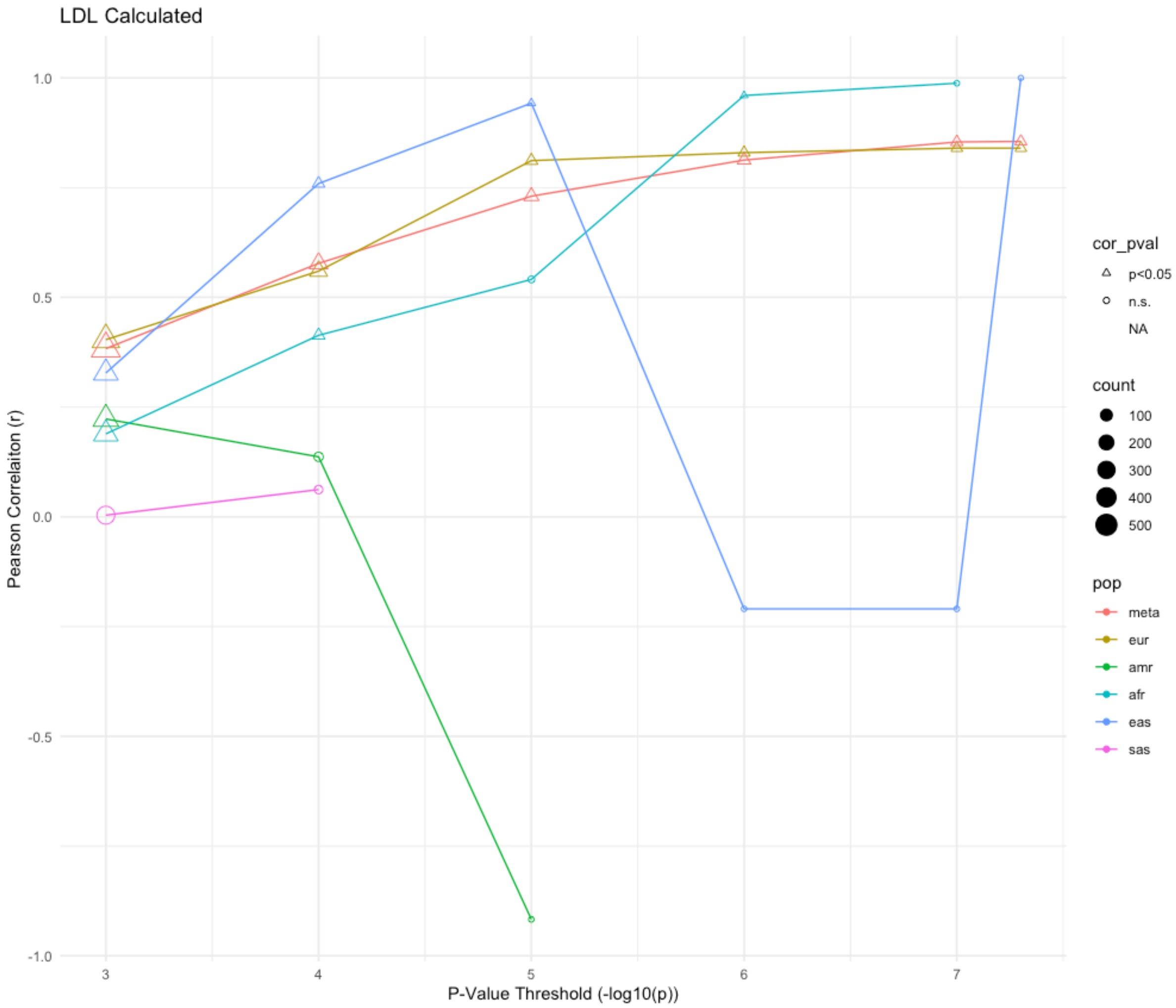


Fig S81. Correlation coefficients of the effect sizes of top SNPs identified in UCLA ATLAS and estimated in UK Biobank for LDL Calc.

Top SNPs were selected from six p-value thresholds: 1e-3, 1e-4, 1e-5, 1e-6, 1e-7, and 5e-8. The number of top SNPs within each threshold that were also identified in UK Biobank was should represented by circle size. Significant correlation at each threshold was shown with triangle. Effect sizes estimated for all ancestry groups and in meta-analysis were shown. A missing point represented that none of the SNPs passed a given threshold in UCLA ATLAS dataset or less than three passing SNPs was found in UK Biobank.

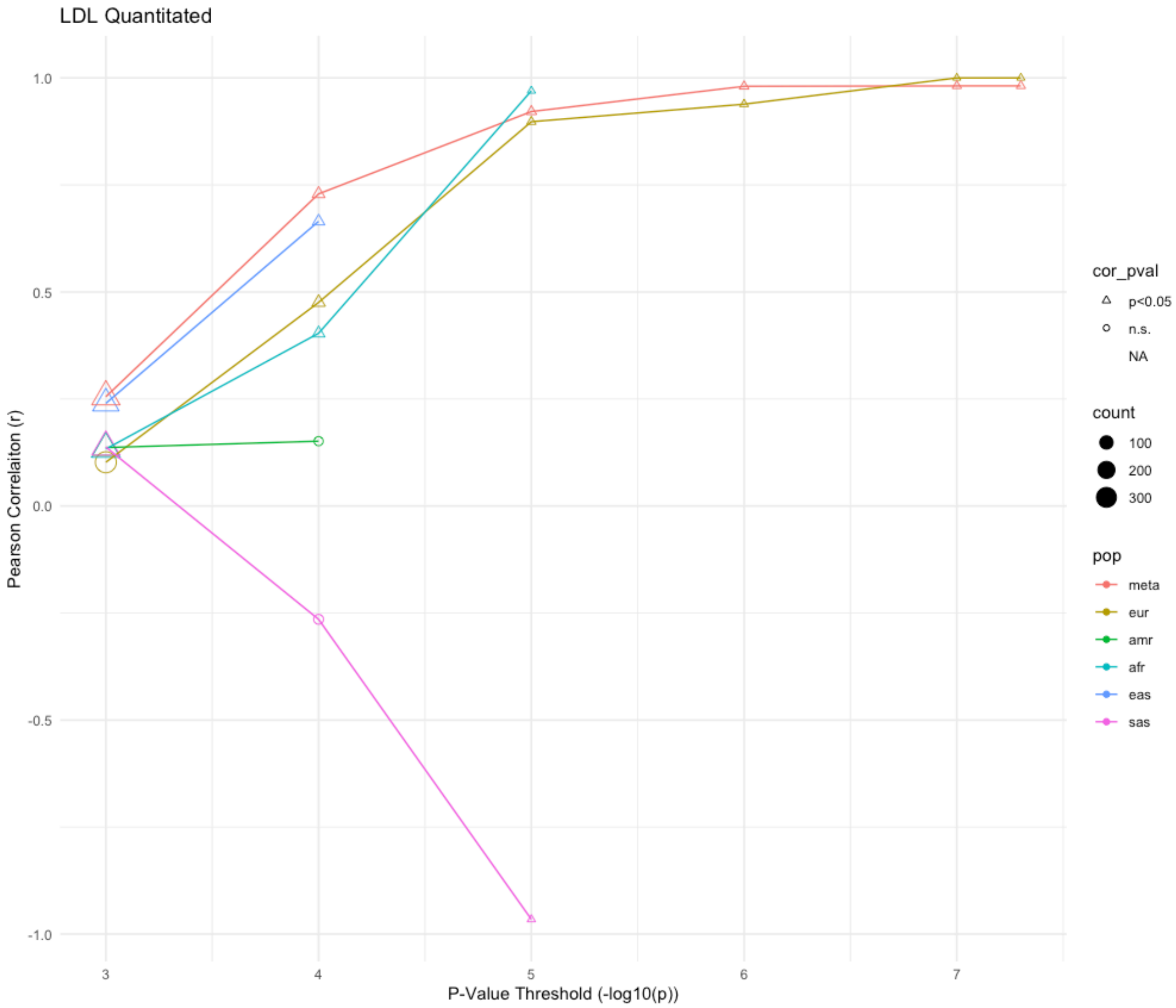


Fig S82. Correlation coefficients of the effect sizes of top SNPs identified in UCLA ATLAS and estimated in UK Biobank for LDL Quant.

Top SNPs were selected from six p-value thresholds: 1e-3, 1e-4, 1e-5, 1e-6, 1e-7, and 5e-8. The number of top SNPs within each threshold that were also identified in UK Biobank was should represented by circle size. Significant correlation at each threshold was shown with triangle. Effect sizes estimated for all ancestry groups and in meta-analysis were shown. A missing point represented that none of the SNPs passed a given threshold in UCLA ATLAS dataset or less than three passing SNPs was found in UK Biobank.

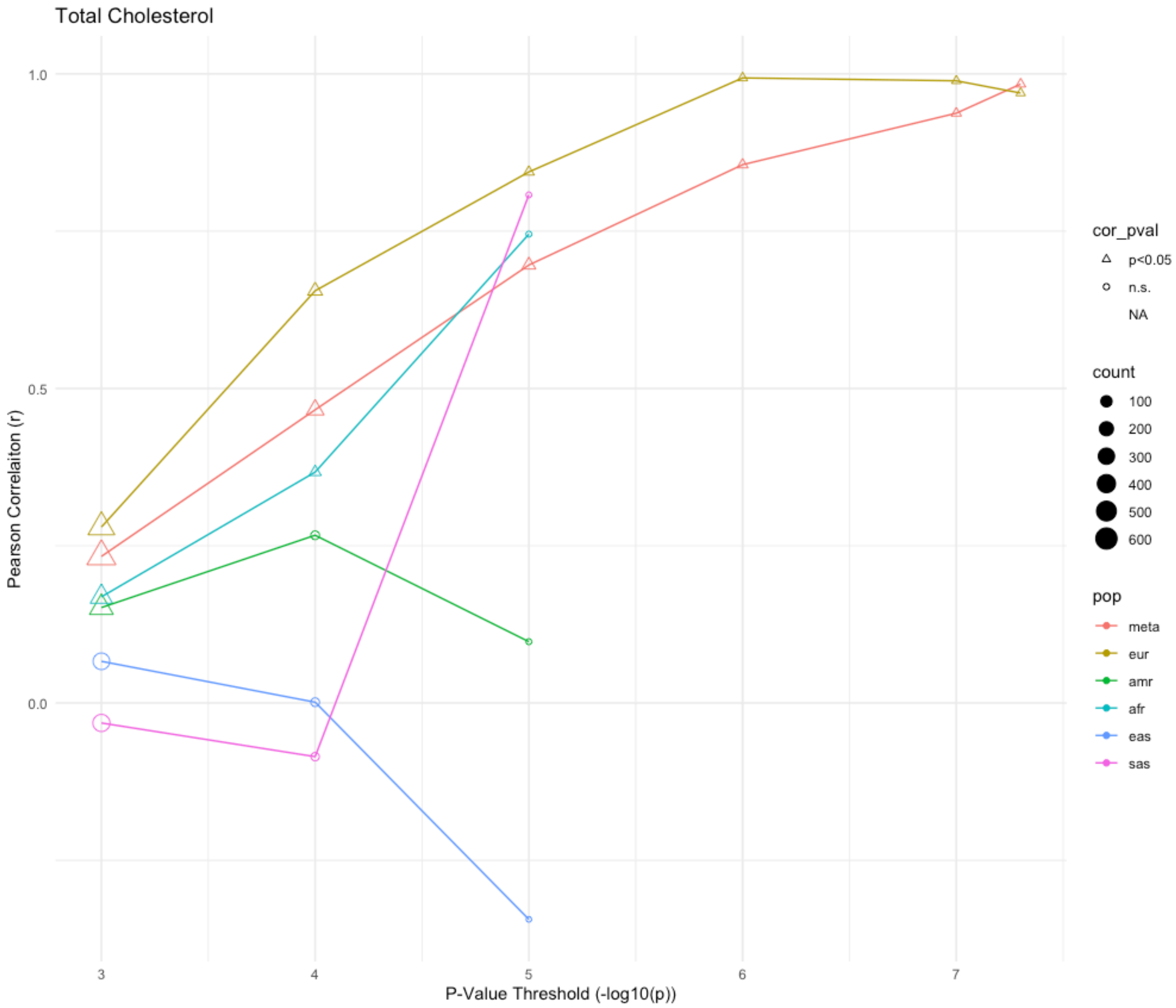


Fig S83. Correlation coefficients of the effect sizes of top SNPs identified in UCLA ATLAS and estimated in UK Biobank for TC.

Top SNPs were selected from six p-value thresholds: 1e-3, 1e-4, 1e-5, 1e-6, 1e-7, and 5e-8. The number of top SNPs within each threshold that were also identified in UK Biobank was should represented by circle size. Significant correlation at each threshold was shown with triangle. Effect sizes estimated for all ancestry groups and in meta-analysis were shown. A missing point represented that none of the SNPs passed a given threshold in UCLA ATLAS dataset or less than three passing SNPs was found in UK Biobank.

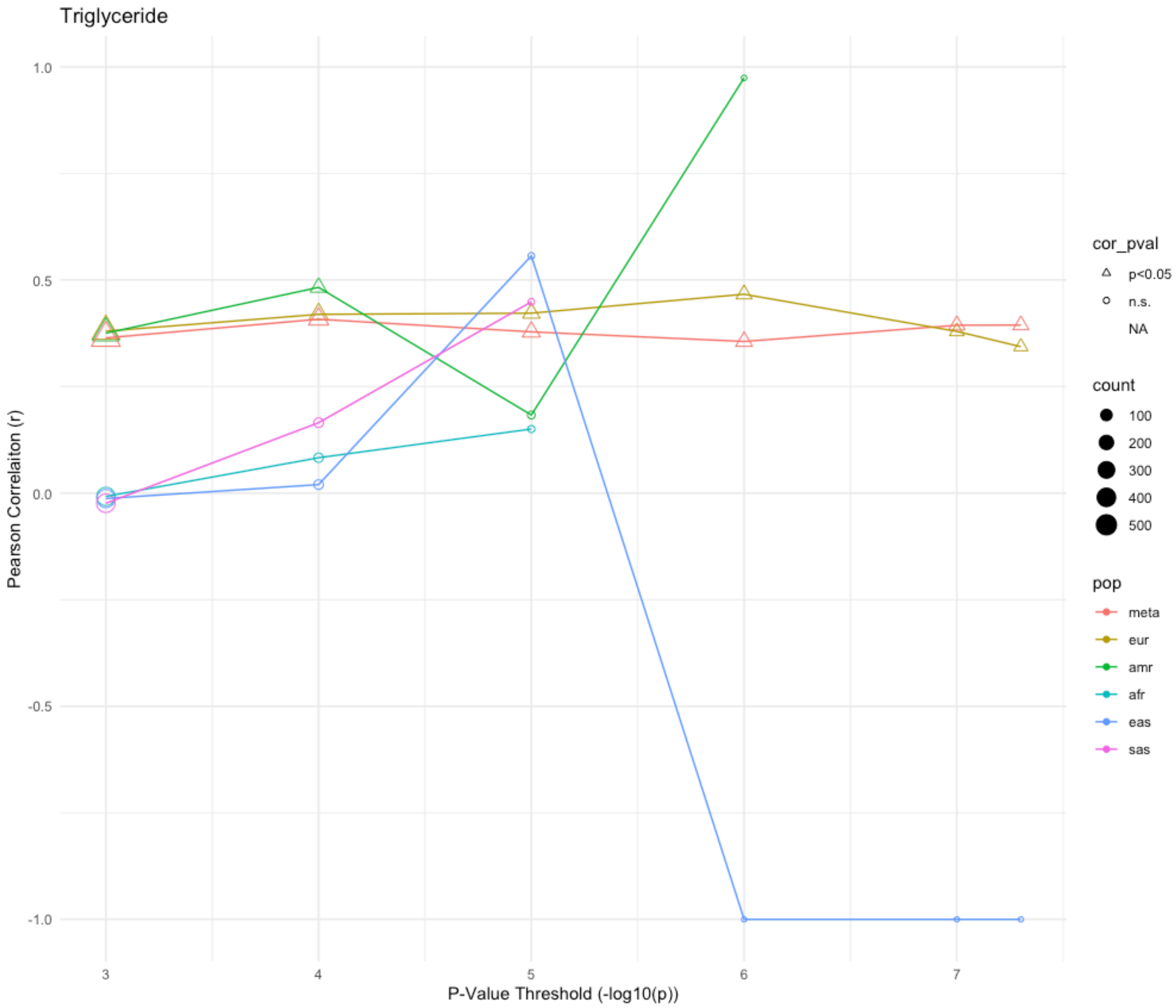


Fig S84. Correlation coefficients of the effect sizes of top SNPs identified in UCLA ATLAS and estimated in UK Biobank for TG.

Top SNPs were selected from six p-value thresholds: 1e-3, 1e-4, 1e-5, 1e-6, 1e-7, and 5e-8. The number of top SNPs within each threshold that were also identified in UK Biobank was should represented by circle size. Significant correlation at each threshold was shown with triangle. Effect sizes estimated for all ancestry groups and in meta-analysis were shown. A missing point represented that none of the SNPs passed a given threshold in UCLA ATLAS dataset or less than three passing SNPs was found in UK Biobank.

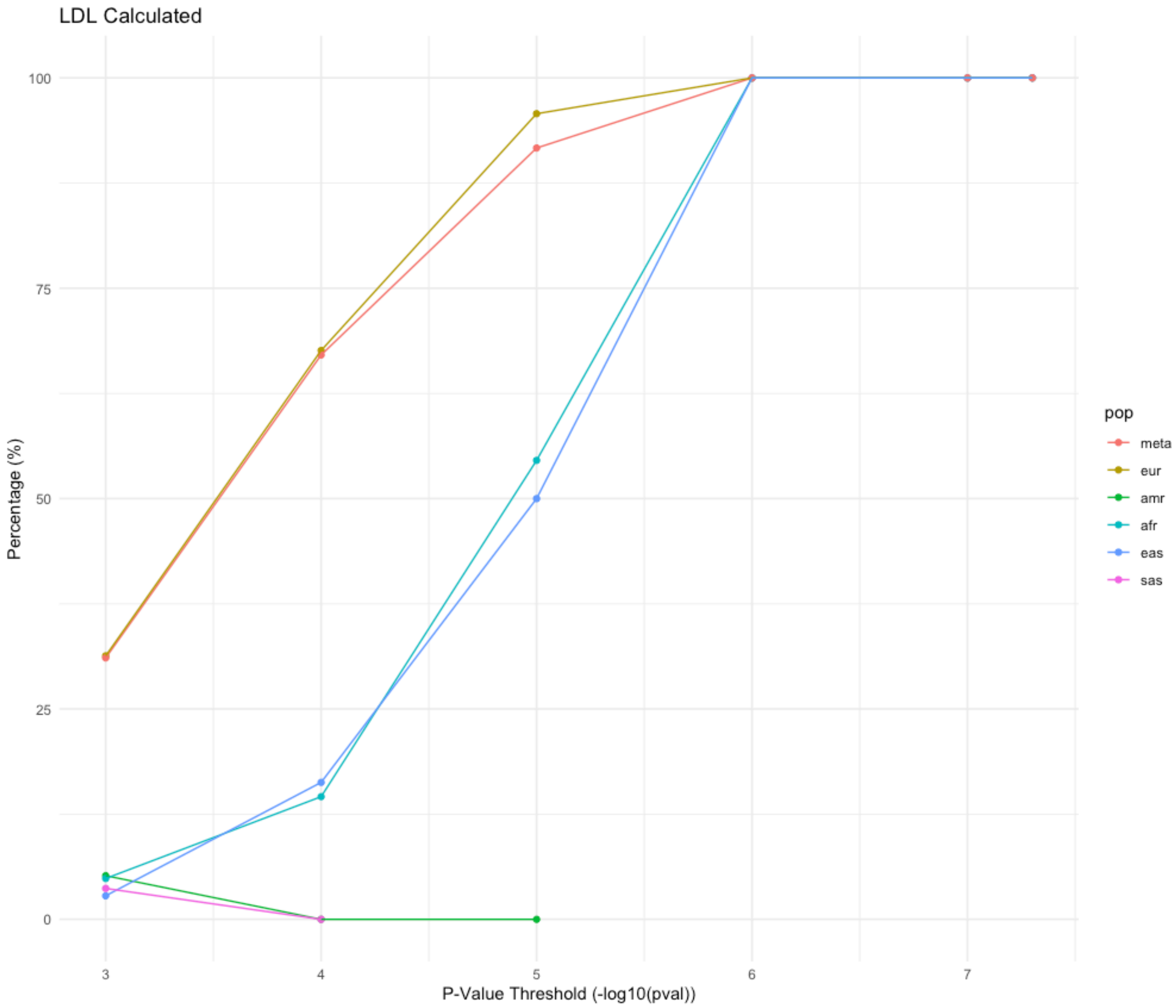


Fig S85. Percentage of overlapping SNPs between UCLA ATLAS dataset and UK Biobank(27) under given p-value thresholds for LDL Calc.
SNPs were selected under six p-value thresholds: 1e-3, 1e-4, 1e-5, 1e-6, 1e-7, and 5e-8. The percentage of SNPs was computed based on the total number of SNPs passing a given threshold in UCLA ATLAS dataset and the number of SNPs among them that also passed the same p-value threshold in UK Biobank. A missing point represented that none of the SNPs passed a given threshold in UCLA ATLAS dataset

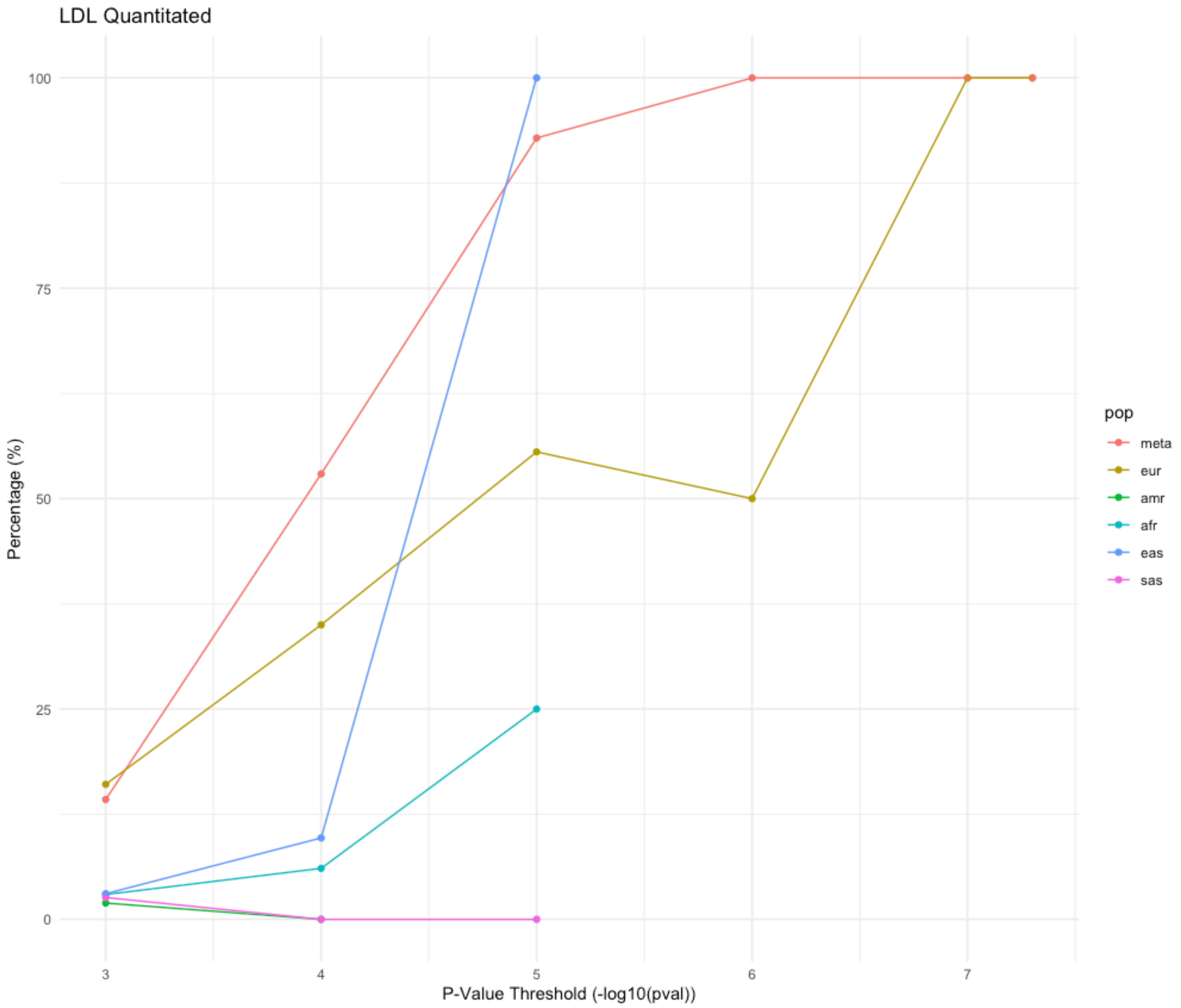


Fig S86. Percentage of overlapping SNPs between UCLA ATLAS dataset and UK Biobank(27) under given p-value thresholds for LDL Quant.

SNPs were selected under six p-value thresholds: 1e-3, 1e-4, 1e-5, 1e-6, 1e-7, and 5e-8. The percentage of SNPs was computed based on the total number of SNPs passing a given threshold in UCLA ATLAS dataset and the number of SNPs among them that also passed the same p-value threshold in UK Biobank. A missing point represented that none of the SNPs passed a given threshold in UCLA ATLAS dataset

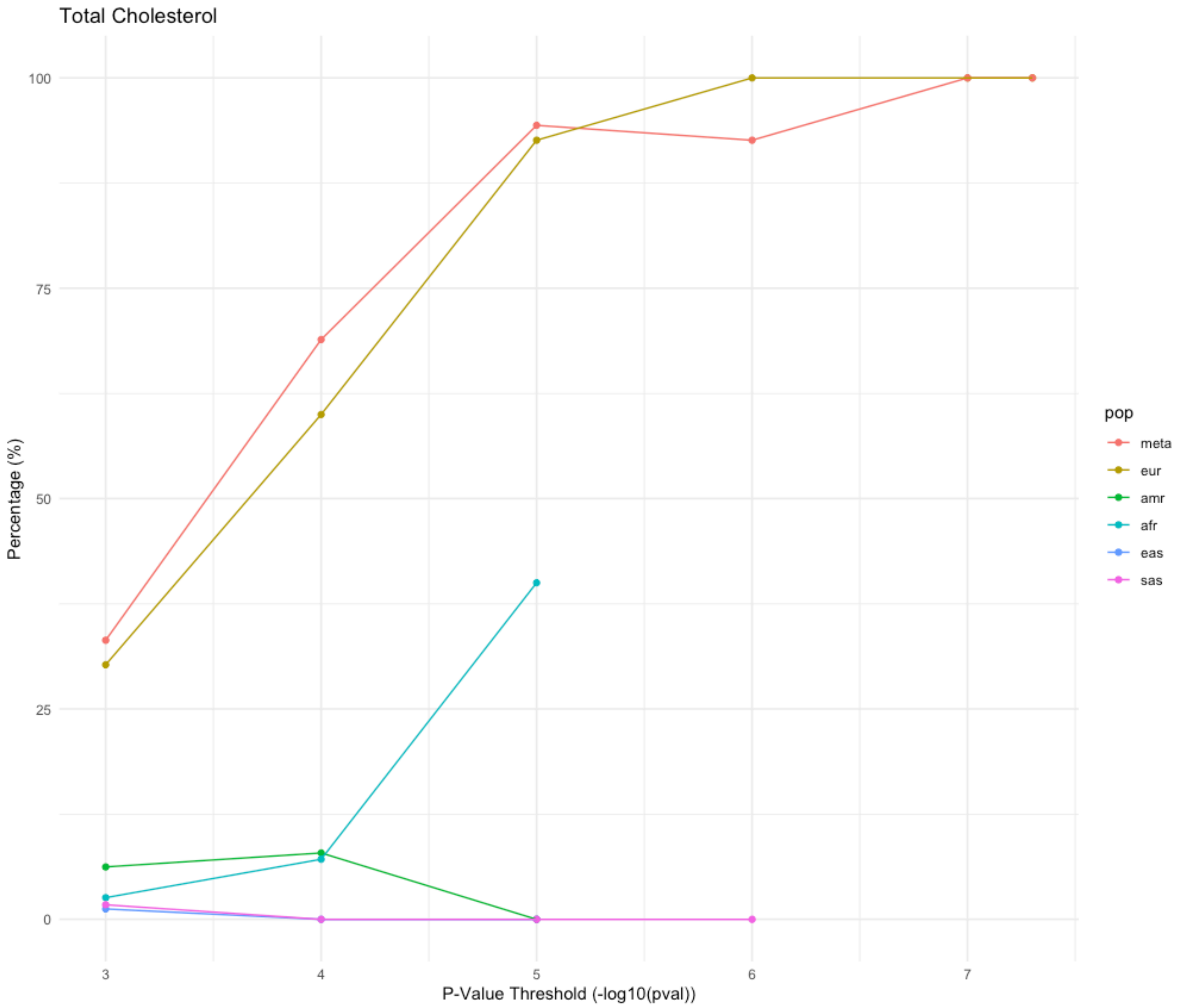


Fig S87. Percentage of overlapping SNPs between UCLA ATLAS dataset and UK Biobank(27) under given p-value thresholds for TC.

SNPs were selected under six p-value thresholds: 1e-3, 1e-4, 1e-5, 1e-6, 1e-7, and 5e-8. The percentage of SNPs was computed based on the total number of SNPs passing a given threshold in UCLA ATLAS dataset and the number of SNPs among them that also passed the same p-value threshold in UK Biobank. A missing point represented that none of the SNPs passed a given threshold in UCLA ATLAS dataset

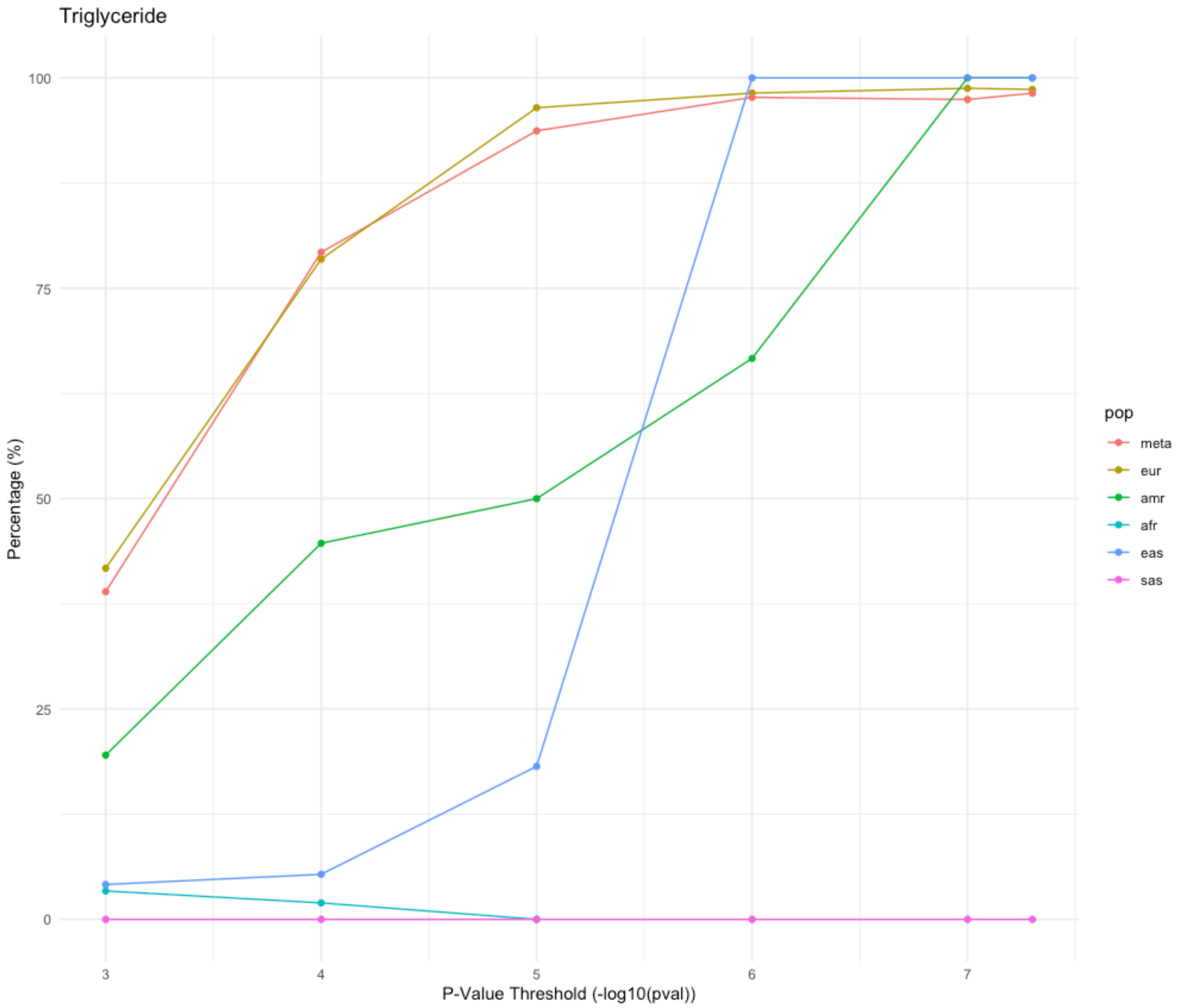


Fig S88. Percentage of overlapping SNPs between UCLA ATLAS dataset and UK Biobank(27) under given p-value thresholds for TG.
SNPs were selected under six p-value thresholds: 1e-3, 1e-4, 1e-5, 1e-6, 1e-7, and 5e-8. The percentage of SNPs was computed based on the total number of SNPs passing a given threshold in UCLA ATLAS dataset and the number of SNPs among them that also passed the same p-value threshold in UK Biobank. A missing point represented that none of the SNPs passed a given threshold in UCLA ATLAS dataset.

Phenotype	HDL	LDL Calculated	LDL Quantitated	Total Cholesterol	Triglyceride
EUR	1.04	1.01	1.00	1.03	1.02
AMR	1.02	1.02	1.01	1.02	1.01
AFR	0.99	1.01	1.01	1.01	1.00
EAS	1.01	1.00	1.00	0.99	1.02
SAS	1.01	0.99	1.01	1.01	1.01
Meta-analysis	1.04	1.01	1.00	1.03	1.01

Table S1. Inflation factor computed for each blood lipid phenotype and ancestry group

Threshol d	Lo w	Mediu m	Hig h	Ver y	Tota l	Low (%)	Medium (%)	High (%)	Very (%)
1.E-03	370	87	81	63	601	61.56%	14.48%	13.48%	10.48%

1.E-04	112	27	25	7	171	65.50%	15.79%	14.62%	4.09%
1.E-05	57	12	12	4	85	67.06%	14.12%	14.12%	4.71%
1.E-06	33	8	12	3	56	58.93%	14.29%	21.43%	5.36%
1.E-07	17	8	11	3	39	43.59%	20.51%	28.21%	7.69%
5.E-08	14	8	11	3	36	38.89%	22.22%	30.56%	8.33%

Table S2. Number and percentage of SNPs under different p-value thresholds and I^2 categories for LDL Calc in the meta-analysis.

Threshold	Low	Medium	High	Very	Total	Low (%)	Medium (%)	High (%)	Very (%)
1.E-03	239	53	63	45	400	59.75%	13.25%	15.75%	11.25%
1.E-04	12	10	11	4	37	32.43%	27.03%	29.73%	10.81%
1.E-05	3	5	6	3	17	17.65%	29.41%	35.29%	17.65%
1.E-06	2	3	4	0	9	22.22%	33.33%	44.44%	0.00%
1.E-07	2	1	4	0	7	28.57%	14.29%	57.14%	0.00%
5.E-08	2	1	4	0	7	28.57%	14.29%	57.14%	0.00%

Table S3. Number and percentage of SNPs under different p-value thresholds and I^2 categories for LDL Quant in the meta-analysis.

Threshold	Low	Medium	High	Very	Total	Low (%)	Medium (%)	High (%)	Very (%)
1.E-03	427	95	108	75	705	60.57%	13.48%	15.32%	10.64%
1.E-04	119	24	19	11	173	68.79%	13.87%	10.98%	6.36%
1.E-05	50	8	11	5	74	67.57%	10.81%	14.86%	6.76%
1.E-06	16	6	5	1	28	57.14%	21.43%	17.86%	3.57%
1.E-07	10	6	4	0	20	50.00%	30.00%	20.00%	0.00%
5.E-08	9	6	4	0	19	47.37%	31.58%	21.05%	0.00%

Table S4. Number and percentage of SNPs under different p-value thresholds and I^2 categories for TC in the meta-analysis.

Threshold	Low	Medium	High	Very	Total	Low (%)	Medium (%)	High (%)	Very (%)
1.E-03	335	100	123	138	696	48.13%	14.37%	17.67%	19.83%
1.E-04	76	30	80	64	250	30.40%	12.00%	32.00%	25.60%
1.E-05	36	24	66	43	169	21.30%	14.20%	39.05%	25.44%
1.E-06	25	19	54	37	135	18.52%	14.07%	40.00%	27.41%
1.E-07	24	17	52	36	129	18.60%	13.18%	40.31%	27.91%

5.E-08	23	17	52	36	128	17.97%	13.28%	40.63%	28.13%

Table S5. Number and percentage of SNPs under different p-value thresholds and I^2 categories for TG in the meta-analysis.

Reference

1. Kannel WB, Dawber TR, Kagan A, Revotskie N, Stokes J, 3rd. Factors of risk in the development of coronary heart disease--six year follow-up experience. The Framingham Study. Ann Intern Med. 1961;55:33-50.
2. Qi Q, Liang L, Doria A, Hu FB, Qi L. Genetic predisposition to dyslipidemia and type 2 diabetes risk in two prospective cohorts. Diabetes. 2012;61(3):745-52.
3. Emerging Risk Factors C, Di Angelantonio E, Sarwar N, Perry P, Kaptoge S, Ray KK, et al. Major lipids, apolipoproteins, and risk of vascular disease. JAMA. 2009;302(18):1993-2000.
4. Oresic M, Hyotylainen T, Kotronen A, Gopalacharyulu P, Nygren H, Arola J, et al. Prediction of non-alcoholic fatty-liver disease and liver fat content by serum molecular lipids. Diabetologia. 2013;56(10):2266-74.
5. Cardiovascular diseases (CDDs) Fact Sheet. World Health Organization; 2017.
6. Weiss LA, Pan L, Abney M, Ober C. The sex-specific genetic architecture of quantitative traits in humans. Nat Genet. 2006;38(2):218-22.
7. van Dongen J, Willemsen G, Chen WM, de Geus EJ, Boomsma DI. Heritability of metabolic syndrome traits in a large population-based sample. J Lipid Res. 2013;54(10):2914-23.
8. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010;466(7307):707-13.
9. Klarin D, Damrauer SM, Cho K, Sun YV, Teslovich TM, Honerlaw J, et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nat Genet. 2018;50(11):1514-23.
10. Lu X, Peloso GM, Liu DJ, Wu Y, Zhang H, Zhou W, et al. Exome chip meta-analysis identifies novel loci and East Asian-specific coding variants that contribute to lipid levels and coronary artery disease. Nat Genet. 2017;49(12):1722-30.
11. Liu DJ, Peloso GM, Yu H, Butterworth AS, Wang X, Mahajan A, et al. Exome-wide association study of plasma lipids in >300,000 individuals. Nat Genet. 2017;49(12):1758-66.
12. Below JE, Parra EJ, Gamazon ER, Torres J, Krithika S, Candille S, et al. Meta-analysis of lipid-traits in Hispanics identifies novel loci, population-specific effects, and tissue-specific enrichment of eQTLs. Sci Rep. 2016;6:19429.
13. Asselbergs FW, Guo Y, van Iperen EP, Sivapalaratnam S, Tragante V, Lanktree MB, et al. Large-scale gene-centric meta-analysis across 32 studies identifies multiple lipid loci. Am J Hum Genet. 2012;91(5):823-38.
14. Peloso GM, Auer PL, Bis JC, Voorman A, Morrison AC, Stitziel NO, et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. Am J Hum Genet. 2014;94(2):223-32.
15. Albrechtsen A, Grarup N, Li Y, Sparso T, Tian G, Cao H, et al. Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes. Diabetologia. 2013;56(2):298-310.

16. Chasman DI, Pare G, Mora S, Hopewell JC, Peloso G, Clarke R, et al. Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis. PLoS Genet. 2009;5(11):e1000730.
17. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013;45(11):1274-83.
18. Myocardial Infarction G, Investigators CAEC, Stitziel NO, Stirrups KE, Masca NG, Erdmann J, et al. Coding Variation in ANGPTL4, LPL, and SVEP1 and the Risk of Coronary Disease. N Engl J Med. 2016;374(12):1134-44.
19. Willer CJ, Mohlke KL. Finding genes and variants for lipid levels after genome-wide association analysis. Curr Opin Lipidol. 2012;23(2):98-103.
20. Hoffmann TJ, Theusch E, Haldar T, Ranatunga DK, Jorgenson E, Medina MW, et al. A large electronic-health-record-based genome-wide study of serum lipids. Nat Genet. 2018;50(3):401-13.
21. Parihar A, Wood GC, Chu X, Jin Q, Argyropoulos G, Still CD, et al. Extension of GWAS results for lipid-related phenotypes to extreme obesity using electronic health record (EHR) data and the Metabochip. Front Genet. 2014;5:222.
22. Pathak J, Kiefer RC, Chute CG. Using semantic web technologies for cohort identification from electronic health records for clinical research. AMIA Jt Summits Transl Sci Proc. 2012;2012:10-9.
23. Collins R. What makes UK Biobank special? Lancet. 2012;379(9822):1173-4.
24. Abul-Husn NS, Soper ER, Odgis JA, Cullina S, Bobo D, Moscati A, et al. Exome sequencing reveals a high prevalence of BRCA1 and BRCA2 founder variants in a diverse population-based biobank. Genome Med. 2019;12(1):2.
25. Tucker B, Sawant S, McDonald H, Rye KA, Patel S, Ong KL, et al. The association of serum lipid and lipoprotein levels with total and differential leukocyte counts: Results of a cross-sectional and longitudinal analysis of the UK Biobank. Atherosclerosis. 2021;319:1-9.
26. Wood GC, Still CD, Chu X, Susek M, Erdman R, Hartman C, et al. Association of chromosome 9p21 SNPs with cardiovascular phenotypes in morbid obesity using electronic health record data. Genomic Med. 2008;2(1-2):33-43.
27. Lab N. UK Biobank GWAS results.
28. Fadista J, Oskolkov N, Hansson O, Groop L. LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals. Bioinformatics. 2017;33(4):471-4.
29. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46(3):310-5.
30. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47(D1):D886-D94.
31. Koepsell H, Lips K, Volk C. Polyspecific organic cation transporters: structure, function, physiological roles, and biopharmaceutical implications. Pharm Res. 2007;24(7):1227-51.
32. Hilgendorf C, Ahlin G, Seithel A, Artursson P, Ungell AL, Karlsson J. Expression of thirty-six drug transporter genes in human intestine, liver, kidney, and organotypic cell lines. Drug Metab Dispos. 2007;35(8):1333-40.
33. Goswami S, Gong L, Giacomini K, Altman RB, Klein TE. PharmGKB summary: very important pharmacogene information for SLC22A1. Pharmacogenet Genomics. 2014;24(6):324-8.

34. Liang X, Yee SW, Chien HC, Chen EC, Luo Q, Zou L, et al. Organic cation transporter 1 (OCT1) modulates multiple cardiometabolic traits through effects on hepatic thiamine content. PLoS Biol. 2018;16(4):e2002907.
35. UCLA Precision Health Biobank [Available from: https://www.uclahealth.org/precision-health/ucla-biobank.
36. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867-73.
37. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
38. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559-75.
39. Han B, Eskin E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. Am J Hum Genet. 2011;88(5):586-98.
40. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47(D1):D1005-D12.
41. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17(1):122.
42. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-9.
43. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062-D7.

Chapter 5 - Conclusion

Understanding the genetic architecture of complex traits has provided an enormous amount of insights into their biology and has suggested numerous potential targets for clinical and therapeutic studies. As more efficient and economical sequencing approaches are available, researchers can analyze a larger number of samples with better resolution. Although many common risk loci with small to medium effect sizes have been discovered through GWASes, genetic studies are hoping to discover variants with medium to large effect sizes which could potentially serve as clinical or therapeutic targets. One category that fits this endeavor is the rare variant. The traditional difficulty of analyzing rare variants is two-fold. One is that rare variants are hardly available through microarray genotyping, which has been largely improved by the recent advent of WGS and WES technologies. The second is that their low occurrence rate limited the power of detecting significant loci. To facilitate this exploration, many study designs and statistical tools have thus been proposed to improve the accuracy and detecting power.

Therefore, in our work, we employed two rare-variant analysis approaches and applied to two complex traits whose genetic architecture has yet been fully understood. In our first application, we utilized a gene-set burden analysis framework to analyze the rare-variant effect in AD. Before our work, many GWASes have been performed and pinpointed over 50 risk loci participating in multiple biological pathways. However, little has been known for the effect of rare variants within these pathways. Hence, our analysis was one of the first works to identify the rare-variant contribution to AD with the endocytic pathway. Furthermore, we showed this contribution was not limited to AD status but also to related pathological phenotypes, such as NFT progression and age at onset. By leveraging on prior knowledge, we were able to avoid the unnecessary multiple-testing burden and focused on analyzing single genes with large effects. There are several other pathways implicated in AD, such as immune response and lipid metabolism pathways. We believe our analytic framework can be extended to these candidate pathways and identify potential contributions of rare variants to AD.

Our second application focused on a specific type of rare variants, DNMs, in TS. These mutations are believed to have large effect sizes as they were not negatively selected by evolutionary pressure and thus represent a probable source of risk leading to TS development. Cheaper and efficient sequencing techniques also largely facilitate studies on DNMs because at least a trio set, one proband and two parents, is needed to determine DNMs within the affected child. In our study, we analyzed nearly 900 trio families with an additional 300 in preparation. We showed that high-quality DNMs could be efficiently called within these TS probands. By partitioning DNMs into different levels of deleterious categories, we noted that the PTV category was significantly enriched compared to other categories, including missense and synonymous mutations. Recurrent mutations in *FBN2* was an important observation as it validated and provided additional evidence for our previous work on TS. Nonetheless, there are several points that could be improved in our next phase of analysis. First, we are currently expecting a third batch of TS trio data which would further increase our detecting power. Second, we plan to adjust our DNM calling pipeline to increase our calling rate. Third, we have recently been approved for the usage of an external trio dataset that could serve as the control for our study, which will provide added power to detect TS-associated enrich genes. Taken together, we believe more improvements are still needed in the genetic study of DNMs.

Lastly, one important aspect of genetic research is to understand the genetic heterogeneity across different populations. Because the association test assumes a homogeneous genetic background, researchers need to take extra care in collecting and analyzing large cohorts. If the assumption is violated, larger cohorts will not result in greater detecting power. In our work, we tried to tackle this problem by identifying underlying ancestry groups in our dataset and then perform an association test for each population individually. This method helps identify population-specific genetic effects. To find the shared effects, we followed with meta-analysis across all identified populations. Our dataset has a rich composition of ancestral backgrounds as we used UCLA EHR-linked biobank. This type of data is beneficial as they are not collected for a specific phenotype while containing larger cohorts with diverse genetic backgrounds. We showed that the common measurements, blood lipid concentrations, demonstrated both population-specific and shared effects in five identified populations. Although the sample size for each population is relatively small, we observed a high consistency compared to other large-scale GWASes and provided additional insights into novel risk loci. Unfortunately, as only genotyping data were available, we did not have enough power to detect the effect of rare variants. But as the sample size continues to grow and WES / WGS data will be available in the future, we will be able to analyze the effect of rare variants on blood lipid phenotypes. To note, thousands of phenotypes are available in UCLA EHR-linked biobank, and we recognize that there are still many opportunities in analyzing this dataset.

In brief, our work has focused on analyzing rare-variant effects in complex traits, including AD and TS. Our findings indicated additional risk loci to the understanding of the genetic architecture of these complex traits and facilitated related clinical and therapeutic studies in identifying candidate targets. We also explored EHR-linked biobank for heterogeneous genetic effects using common variants, which suggested its potentiality for rare-variant study. When sequencing data are available, a rare-variant analysis will be possible for thousands of different phenotypes and various ancestral groups, deepening our understanding of the connection between human genetics and phenotypes/disease predispositions. As an ultimate goal of genetic research, the life span and health expectation will be greatly improved.