All of Us Genomic Research Data Quality Report: LDL Cholesterol GWAS Association Replication using the Whole Genome Sequencing dataset

Summary

The *All of Us* Research Program's Data and Research Center conducted a low-density lipoprotein (LDL) genome-wide association study (GWAS), and compared *All of Us* dataset association results with a recent multi-ethnic LDL GWAS in the NHLBI TOPMed (National Heart, Lung, and Blood Institute Trans-Omics for Precision Medicine) study. It revealed a very strong positive association between the two effect estimates, providing strong evidence for the ability of the current dataset to successfully reproduce prior results.

Background

In September 2021, the Data and Research Center (DRC) initiated a phased genome-wide association study (GWAS) and LDL Cholesterol Association Replication study to assess the "fitness for use" of the *All of Us* Research Program whole genome sequencing (WGS) data in preparation for the Controlled Tier and Genomics launch on March 17, 2022. This document describes the methods and results for this GWAS project using the first release of the *All of Us* Research Program's genomic data that contains 98,622 WGS samples.

Methods

The GWAS approach takes a set of genetic variants and a phenotype to find statistical associations between the phenotype and variants. The *All of Us* WGS Hail MatrixTable was used as the genotypic data in this study. The phenotype was a well-studied continuous human trait, LDL cholesterol levels.

The phenotypic data was extracted from the Curated Data Repository (CDR, Control Tier Dataset v5) in the *All of Us* Researcher Workbench. Then the*All of Us* Cohort and Dataset Builder were used to extract all LDL cholesterol measurements from the Lab and Measurements criteria in electronic health record (EHR)data for all participants who have WGS data. Afterwards, the most recent measurements were selected as the phenotype label and adjusted for statin use. A rank-based inverse normal transformation was applied for this continuous trait to increase power and deflate type I error [1]. (Z.R. McCaw, et al, 2020).

Using the Hail MatrixTable, a series of variant quality control (QC) were performed to minimize potential false findings, using Hail as the major tool. Figure 1 is the flowchart of the variant QC steps. The Hail MatrixTable was annotated with phenotypic data and component features (PCs) from the auxiliary data provided by the *All of Us* DRC, and filtered out samples that did not have phenotypic data. A linear regression was then performed, setting the first five PCs as covariates with 34, 924 participants and 3,453,783 variants.

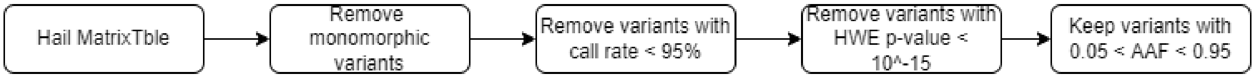
Hail MatrixTble
Remove monomorphic variants
Remove variants with call rate < 95%
Remove variants with HWE p-value < 10^-15
Keep variants with 0.05 < AAF < 0.95

Figure 1. Variant quality control (QC) steps. (HWE: Hardy–Weinberg Equilibrium, AAF: Alternative Allele Frequency).

Results

The expected loci were identified with a genome-wide significance including Apolipoprotein E (*APOE)* and Low Density Lipoprotein Receptor (*LDLR*) on chromosome 19 (chr19), Sortilin 1/Cadherin EGF LAG Seven-Pass G-Type Receptor 2 (*SORT1/CELSR2*) and Proprotein Convertase Subtilisin/Kexin Type 9 (*PCSK9)* on chr1, Apolipoprotein B (*APOB*) and ATP Binding Cassette Subfamily G Member 8 (*ABCG8*) on chr2, 3-Hydroxy-3-Methylglutaryl-CoA Reductase (*HMGCR*) on chr5, and Lipoprotein(a) (*LPA*) and human leukocyte antigen (*HLA*) on chr6. Figure 2 shows the Manhattan plot of the selected variants. The genome-wide significance cutoff p-value was 5 x 10^{-8} (red), and a suggestive threshold of 10^{-5} (blue) was plotted. Table 1 shows detailed information of the identified loci. The Q–Q (quantile-quantile) plot in Figure 3 shows that phenotype and genotype data was appropriately quality controlled prior to running association tests. The association results were also compared with the results from a recent multi-ethnic LDL GWAS in the NHLBI TOPMed study [2].(Selvaraj, biorxiv 2021). Figure 4 displays linear regression results of the two sets of effect estimates. It revealed a very strong positive association between the two effect estimates, providing strong evidence for the ability of the current dataset to successfully reproduce prior results. Differences between the Selvaraj analysis and our analysis are likely accounted for in part by differences in LDL phenotype ascertainment (the Selvaraj study primarily included cohort wide ascertainment of lipids rather than EHR derived lipid values in *All of Us*), differences in LDL lowering medication between the two analyses, and differences in the genetic ancestry composition of the two cohorts.

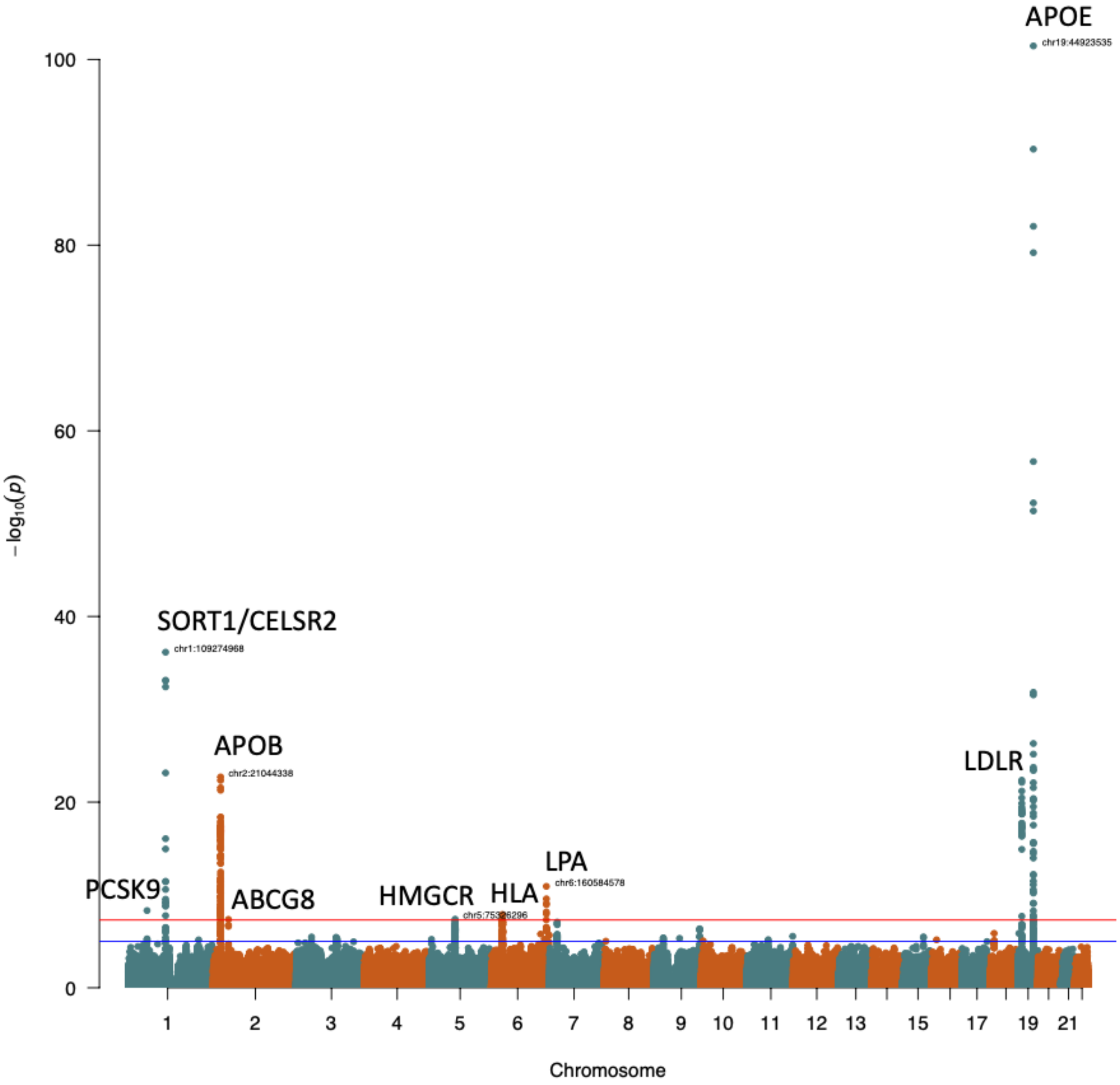

Figure 2. Manhattan plot (N = 34,924) from single variant GWAS study with LDL identifies the expected loci across the genome.

Table 1. Genome-wide significant GWAS loci. These are well known loci that were identified as significant ($p<5 \times 10^{-8}$) across the genome.

Gene	rsID	Chr	Pos (hg38)	Pval	Beta

APOE	rs1332689064	19	44923535	3.5 x 10^{-102}	-12.8
PCSK9	rs1374703008	1	55055522	4.8 x 10^{-9}	-2.8
SORT1/CELSR2	rs611917	1	109274968	7.1 x10^{-37}	-4.4
APOB	rs1043768124	2	21044338	1.9 x 10^{-23}	3.7
ABCG8	rs4245791	2	43847292	4.3 x 10^{-8}	-1.8
HMGCR	rs147592913	5	75326296	4.1 x 10^{-8}	1.6
HLA	rs7773668	6	32410047	4.5 x 10^{-8}	-3.0
LPA	rs55730499	6	160584578	1.2 x 10^{-11}	4.4
LDLR	rs138294113	19	11081053	7.4 x 10^{-23}	-4.4

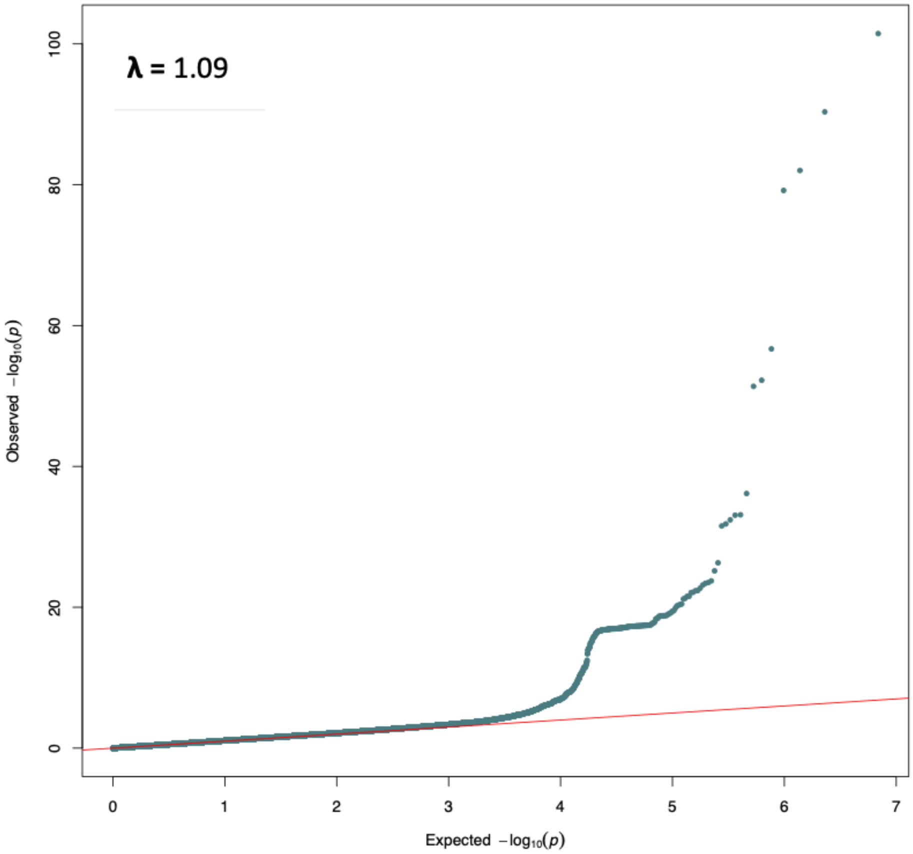

Figure 3. Q-Q plot identifies minimal test-statistic inflation, suggesting the phenotype and genotype data was appropriately quality controlled prior to running association tests. The genomic inflation factor was 1.09.

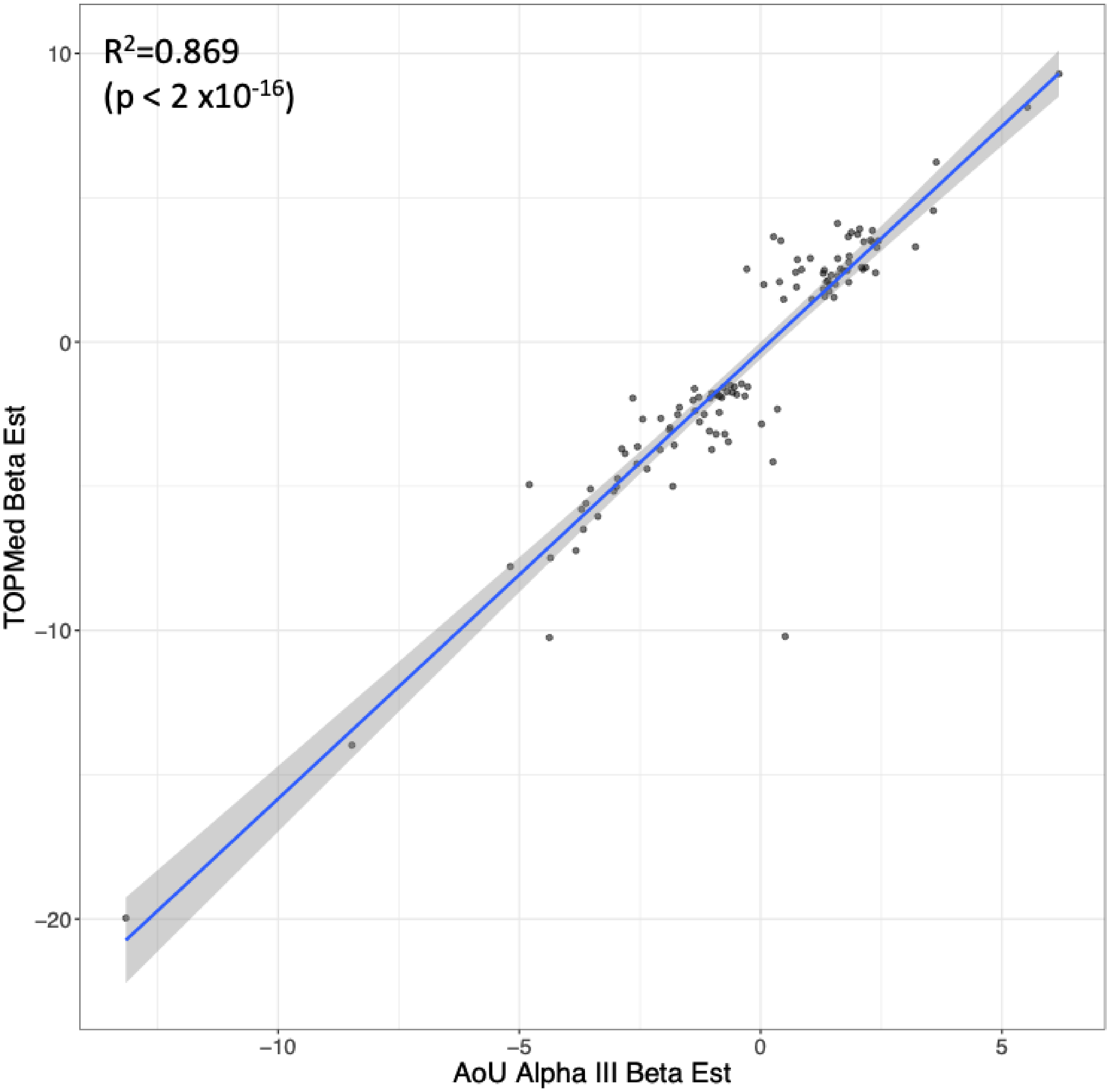

Figure 4. The *All of Us* LDL GWAS (N=34,924) was compared to a recent multi-ethnic LDL GWAS in the NHLBI TOPMed study (Selvaraj, biorxiv 2021) which included 66,329 ancestrally diverse (56% non-European ancestry) individuals, whole genome sequenced with 428M variants. The effect estimates (Beta) between the TOPMed genome-wide significant loci (clumped based on 250kb window, with a linkage disequilibrium threshold r^2 of .25) and the *All of Us* Beta variants for these loci were compared. This revealed a very strong positive association between the two beta estimates, providing strong evidence for the ability of the current dataset to successfully reproduce prior results. R^2 = 0.869 suggesting a very strong correlation with prior effect estimates (p < 2x10-16).

Supplemental Results

Phenotype and Demographics

Date of birth, LDL measurements and measurement date, race, ethnicity and sex at birth were extracted for all participants. Of the 98,622 WGS participants, 34,924 had LDL cholesterol measurements in the electronic health records (EHR) module. LDL cholesterol measurements were from 1 to 959 mg/dL, with a mean of 101.63 mg/ml. Adjusted LDL cholesterol measurements were from 10 to 470 mg/dL, with a mean of 110.92 mg/dL, and from -170.64 to 170.64 after normalization (rank-based inverse normal transformation). The ages at the most recent LDL measurement were from 7 to 99 with a mean of 56. And 21,469 (61%) of them were female. Figure S1 shows the distributions of selected participants' adjusted LDL cholesterol level before and after normalization. Figure S2 shows the distributions of selected participants' age at the most recent valid LDL measurement. For self-reported race and ethnicity, 21,701 (62%) of them were White and 28,579 (82%) of them were not Hispanic or Latino. Table S1 and Table S2 show selected participants self-reported race and ethnicity.

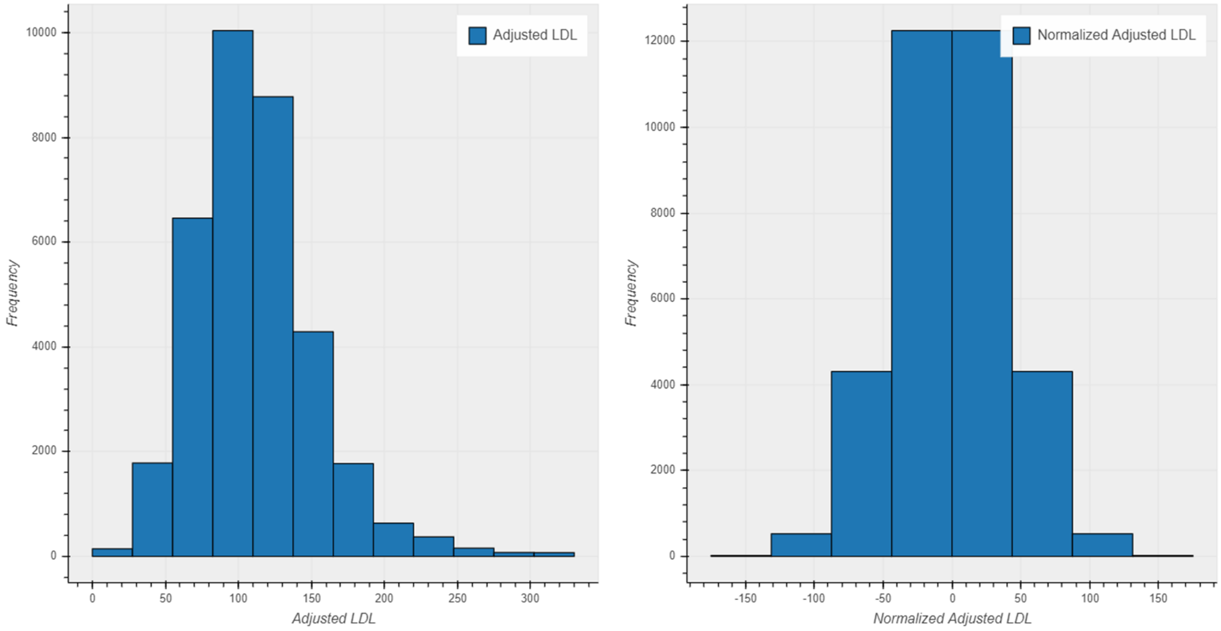

Figure S1. Distributions of selected participants' adjusted LDL cholesterol level before and after normalization. In order to comply with the All of Us Data and Statistics Dissemination Policy [3], we set those adjusted LDL values greater than 320 mg/dL to 320 mg/dL, and adjusted the number of bins.

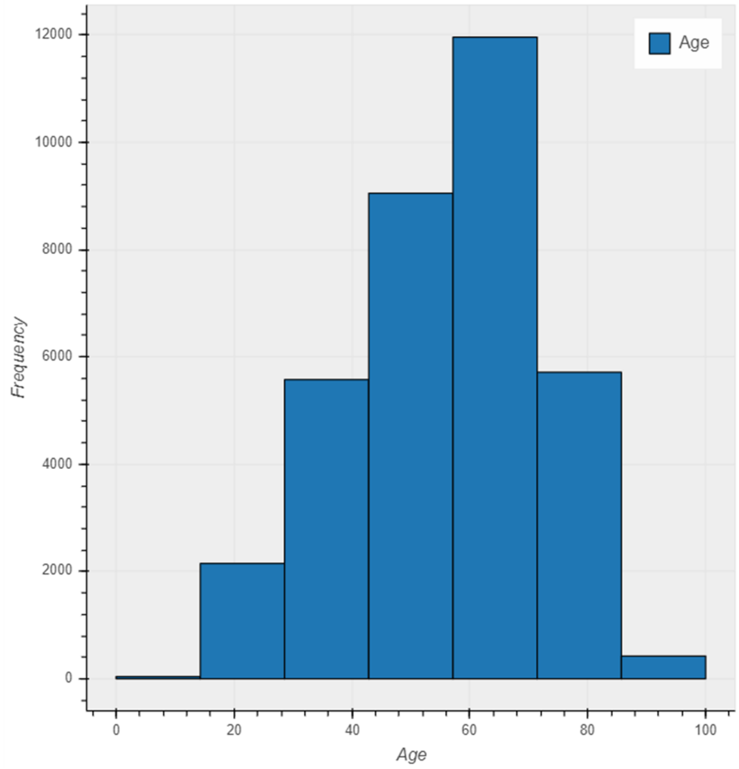

Figure S2. Distributions of selected participants’ age at the most recent valid LDL measurement.

Table S1. Selected participants’ self-reported race.

Race	Percent (%)
White	62.14
Black or African American	16.19
None Indicated	13.81
PMI: Skip	1.30
Asian	2.87
More than one population	1.50
None of these	1.01
Middle Eastern or North African	0.60
I prefer not to answer	0.53
Native Hawaiian or Other Pacific Islander	0.05

Table S2. Selected participants' self-reported ethnicity.

Ethnicity	Percent (%)
Not Hispanic or Latino	81.83
Hispanic or Latino	15.33
PMI: Skip	1.30
What Race Ethnicity: Race Ethnicity None of These	1.01
PMI: Prefer Not To Answer	0.53

Computing Environment:

Main node: 16 CPUs, 104GB RAM, 100GB Disk
Workers (400/800): 4CPU, 26GB RAM, 150GB Disk
Time / Cost: ~0.5hr / ~$99

Acknowledgement:

The DRC wishes to acknowledge the help provided by Nicole Deflaux and Margaret Sunitha Selvaraj for the phenotypic data extraction.

References:

[1] Z.R. McCaw, J.M. Lane, R. Saxena, S. Redline, X. Lin. **Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies**. Biometrics, 76 (2020), pp. 1262-1272

[2] Selvaraj, Margaret Sunitha, et al. "Whole genome sequence analysis of blood lipid levels in> 66,000 individuals." bioRxiv (2021).

[3] How to comply with the All of Us Data and Statistics Dissemination Policy, https://aousupporthelp.zendesk.com/hc/en-us/articles/360043016291-How-to-comply-with-the-All-of-Us-Data-and-Statistics-Dissemination-Policy