All of Us Genomic Research Data Quality Report: Phenotype-Genotype Association Replication using the Whole Genome Sequencing Dataset

Summary

The *All of Us* Research Program's Data and Research Center tested replication rates of known phenotype-genotype associations in three of the largest predicted ancestry populations: European (EUR), African (AFR) and East Asian (EAS) using the extended "Phenotype / Genotype Reference Map" (PGRM) with the *All of Us* whole genome sequencing (WGS) data. These results show a breadth of replicated associations in the three selected populations. The overall replication rates were 68% and 46% for EUR- and AFR- ancestry variants, suggesting more work could be done to address the complexities of the data. For EAS, only one tested phenotype-genotype association was powered at 80%, which did not replicate. The sample sizes for EAS were still quite small here, though many variants were present and testable. This report provides evidence for known challenges using multi-site electronic health record (EHR) data, but does not present concerning evidence for unknown systemic biases at this time.

Background

In September 2021, the Data and Research Center (DRC) initiated a phased Phenotype-Genotype Association Replication study to assess the "fitness for use" of the *All of Us* Research Program WGS data in preparation for Controlled Tier and Genomics Launch on March 17, 2022. This document describes the methods and results for this Genotype-Phenotype Association Replication project using the *All of Us* WGS data.

This replication study specifically looks across many variants and phenotypes to see if there are data quality concerns that may not have been discovered by looking at a single variant or phenotype for association. The *All of Us* Research Program is recruiting participants who are typically Underrepresented in Biomedical Research (UBR) from across the country at specific clinics in different healthcare organizations with different information technology (IT) infrastructures. Because of this, the following challenges related to confounding by recruitment site were anticipated:

1) Sites are recruiting from different demographic groups
2) Sites are recruiting from different populations of genetic ancestry

3) Sites are recruiting in specific clinics which may affect the distribution of clinical attributes of the participants
4) Sites have different clinical specialities, i.e., may have data for some types of care and not others
5) Sites have different clinical practices and policies that impact standards of care and/or reporting
6) Sites have different EHR systems which may impact how/what data is recorded or extracted
7) Sites have different extract, transform, load (ETL) pipelines and curation practices to standardize their clinical data to share with the program

Methods

The Data and Research Center tested replication rates of known phenotype-genotype associations in the three of the four largest populations: EUR, AFR and EAS. The third largest population, Admixed American (AMR), was not included because they have no registered GWAS . This method was a conceptual extension of the original GWAS x PheWAS manuscript, Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data [1]. (Denny, J.C. et al, 2013). Denny, J.C. et al have expanded their original work to create a “Phenotype / Genotype Reference Map” (PGRM) based on associations in the GWAS catalog in June 2020. After directly matching the Experimental Factor Ontology (EFO) terms to phecodes, the team identified 8,085 unique loci and 170 unique phecodes which comprise the PGRM. For this analysis, the DRC used the EUR, AFR and EAS based maps, only considering catalog associations that were p<5e-8 or more significant.

The major tools used were Python package Hail and R package PheWAS. The phenotypes, sex at birth and year of birth were extracted from the *All of Us* Curation Data Repository (CDR, CDR, Control Tier Dataset v5). These phenotypes were then loaded into the *All of Us* WGS Hail MatrixTable from the Google Bucket, and related samples were removed using the relatedness data provided by the *All of Us* DRC.Only samples with EHR data were kept, filtered by selected loci, annotated with demographic and phenotypic information extracted from CDR , and ancestry prediction information provided by the *All of Us* DRC, resulting in 75,122 samples and 5,286 variants for downstream analysis. The variants were further filtered by minimum ancestry-specific allele frequency of 5% and a mean genotype quality score of 40, identifying 3,742 variants.

With the predicted ancestry labels, the three largest groups were studied: 42,156 (56%) predicted European ancestry (EUR) participants, 17,396 (23%) predicted African ancestry (AFR) participants, and 1,656 East Asian (EAS) participants. Three sets of analyses per ancestry were

performed, each sex-specific and common separately. Results where there were at least 20 cases in the analysis group were included. Then, a series of Firth logistic regression tests with phecodes as the outcome and variants as the predictor were performed, adjusting for age, sex (for non-sex-specific phenotypes), and the first three genomic principal component (PC) features as covariates. The PGRM was annotated with power calculations based on the case counts and reported allele frequencies. Power of 80% or greater was considered powered for this analysis.

Results

EUR population

The Firth logistic regression was run on 42,156 samples for 1,661 non-sex-specific phecodes, 25,192 samples for 119 female-specific phecodes, and 16,560 samples for 34 male-specific phecodes.

Replication rates

For EUR-predicted individuals, 3,534 phenotype-genotype associations were tested. 444 of those associations were powered at 80%. Among those powered associations, 300 of them were replicated, for an overall rate of 68%. Figure 1 shows the replication rates for EUR-ancestry variants across phecode categories, suggesting more work could be done to address the complexities of the data.

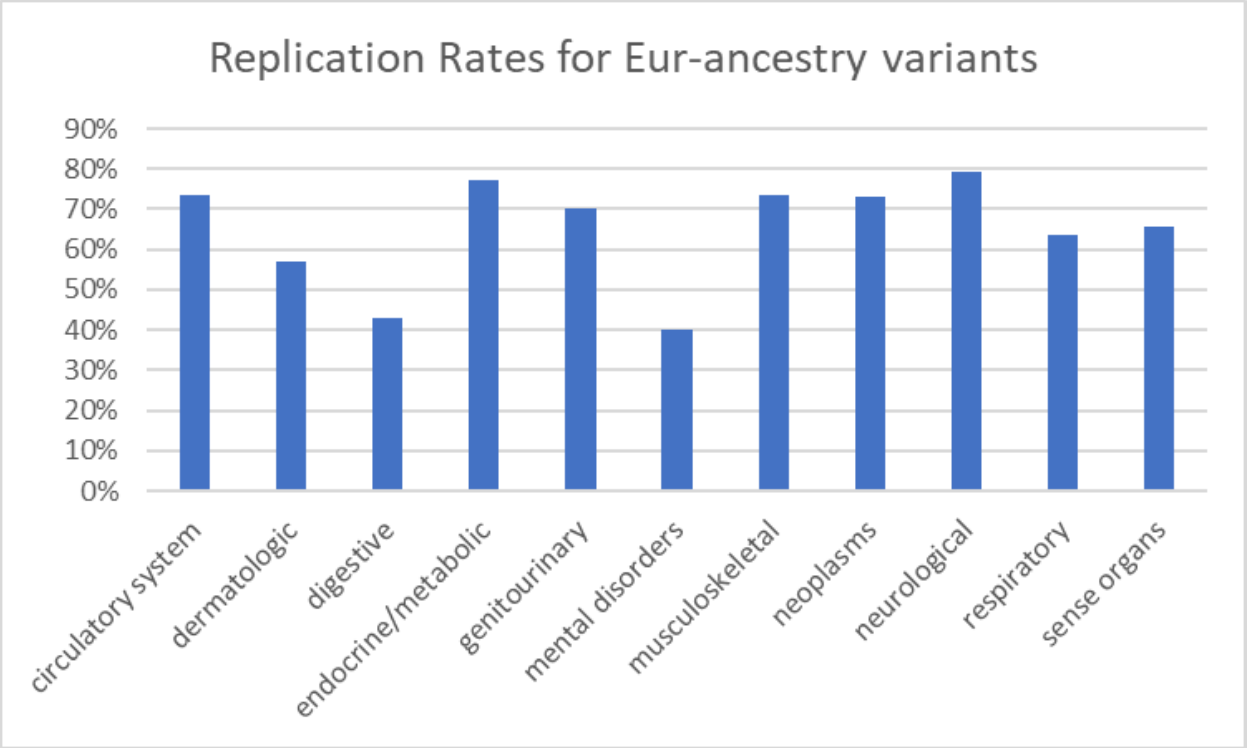

Figure 1. Replication rates for EUR-ancestry variants.

AFR population

The Firth logistic regression was run on 17,396 samples for 1,645 non-sex-specific phecodes, 17,396 samples for 119 female-specific phecodes, and 6,832 samples for 34 male-specific phecodes.

Replication rates

For AFR-predicted individuals, 39 phenotype-genotype associations were tested. 13 of those associations were powered at 80%. Among those powered associations, 6 of them were replicated, for an overall rate of 46%. Table 1 shows the replication rates for AFR-ancestry variants, suggesting more work could be done to address the complexities of the data.

Table 1. Replication Rates for AFR-ancestry Variants.

Phecode Group	Powered Associations	Replication Rate
circulatory system	4	0%

endocrine/metabolic	5	60%
genitourinary	1	0%
hematopoietic	1	100%
musculoskeletal	1	100%
neoplasms	1	100%

EAS population

The Firth logistic regression was run on 1,656 samples for 1,473 non-sex-specific phecodes, 1,106 samples for 113 female-specific phecodes, and 544 samples for 27 male-specific phecodes.

Replication rates

For EAS-predicted individuals, 517 phenotype-genotype associations were tested. one of these associations was powered at 80%, which did not replicate. The sample sizes were still quite small here, though many variants were present and testable.

Supplemental Results

Demographics

All figures were adjusted by bin size to comply with the *All of Us* data and statistics dissemination policy [2].

EUR population

Of all the 42,156 European-predicted individuals, ages ranged from 18 to 103, with a mean of 58. 25,192 (59.76%) of them were female. The case rate was calculated for 8,952,578 phenotype-variant pairs within the European cohort. The case rates for the phenotype-variant

pairs with at least 20 cases were from 0.05% to 41.08%, with a mean of 2.96%. Figure S1 shows the distribution of current age and case rate for the participants.

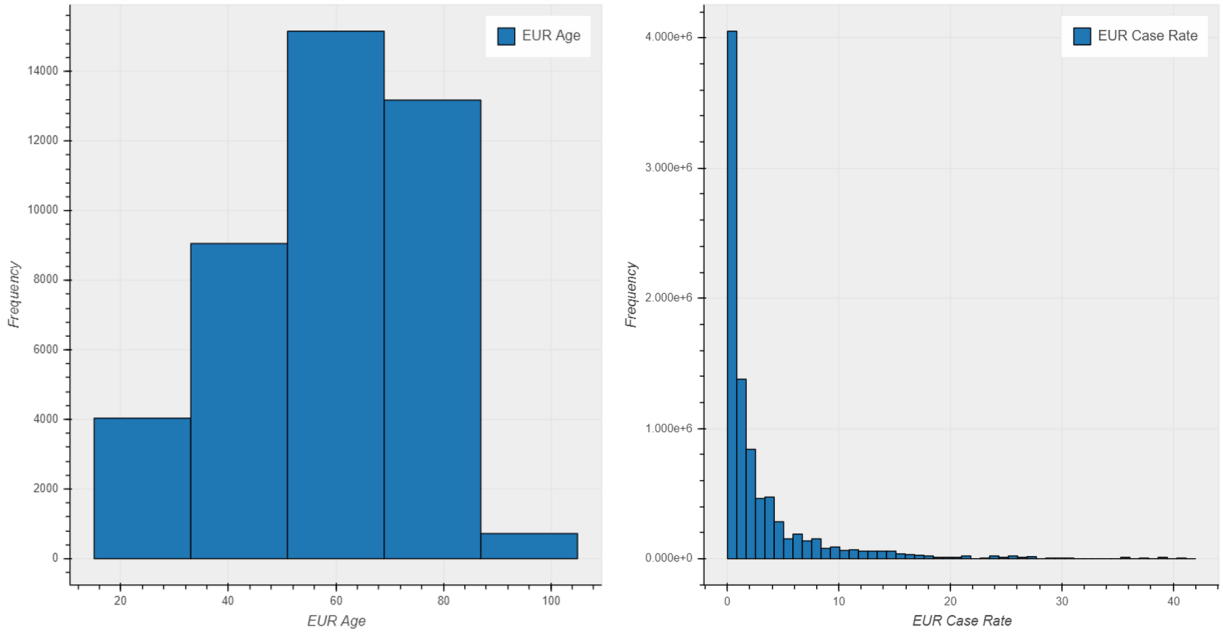

Figure S1. Age and case rate distribution of the European cohort.

AFR population

Of all the 17,396 African-predicted individuals, ages ranged from 19 to 99, with a mean of 51. 17,396 (59.03%) of them were self-reported females. The case rate was calculated for 7,711,555 phenotype-variant pairs within the AFR cohort. The case rates for the phenotype-variant pairs with at least 20 cases were from 0.13% to 40.43%, with a mean of 2.90%. Figure S2 shows the distribution of current age and case rate for the participants.

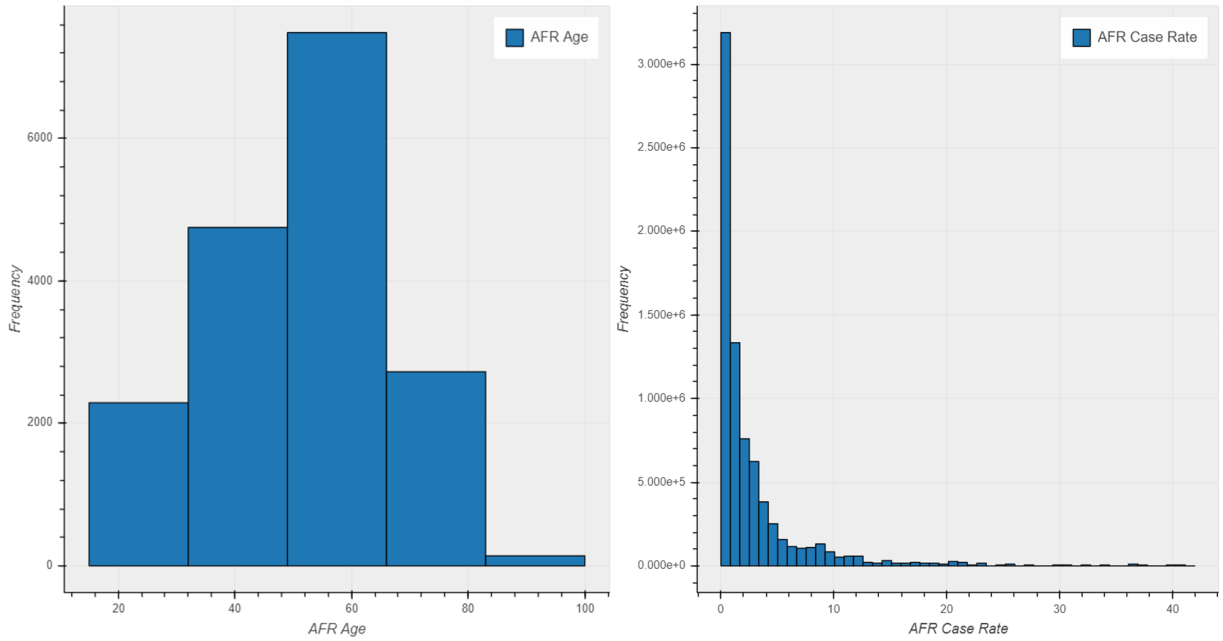

Figure S2. Age and case rate distribution of the African cohort.

EAS population

Of all the 1,656 East Asian-predicted individuals, ages ranged from 19 to 91, with a mean of 47. 1503 (66.79%) of these individuals were self-reported females. The case rate was calculated for 2,998,888 phenotype-variant pairs within the EAS cohort. The case rates for the phenotype-variant pairs with at least 20 cases were from 1.27% to 26.70%, with a mean of 4.40%. Figure S3 shows the distribution of current age and case rate for the participants.

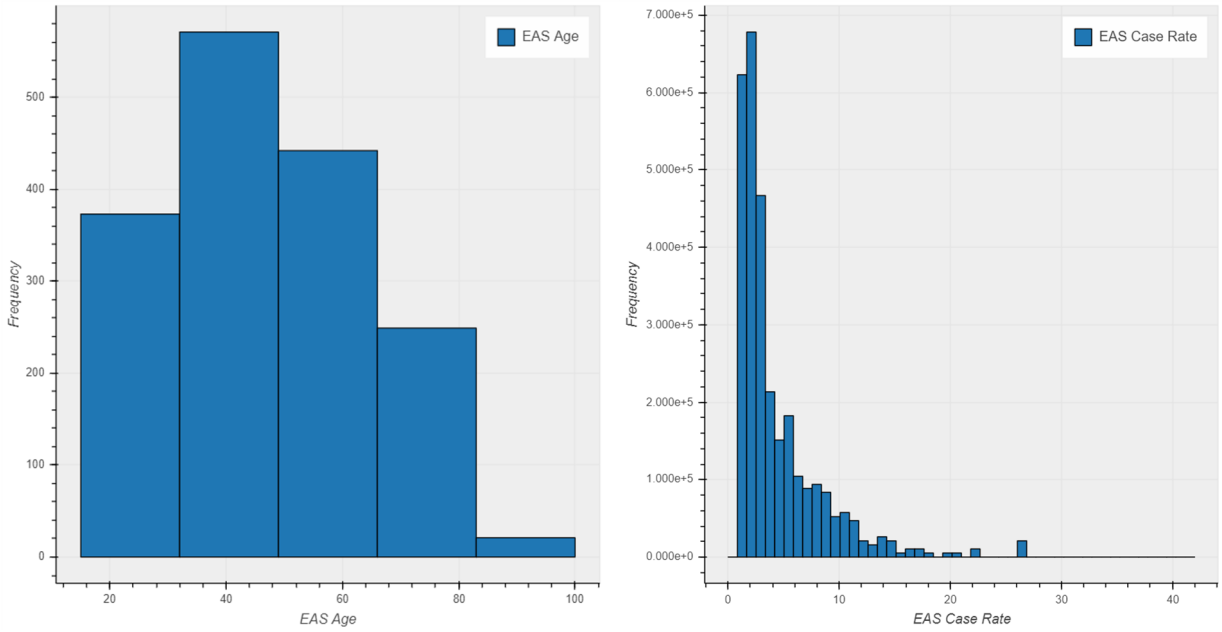

Figure S3. Age and case rate distribution of the East Asian cohort.

Computing environment

Main node: 32 CPUs, 120GB RAM, 100GB Disk
Workers (200/200): 4CPU, 26GB RAM, 150GB Disk
Time & cost: ~2hr / ~$160 (Not including variant filter step)

References:

[1] Denny, J.C. et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat. Biotechnol. 31, 1102–1110 (2013).

[2] 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