# Modeling Community and Genomic Factors of HIV Susceptibility in the *All of Us* Research Program

by

**Dominika E. Oliver**

Bachelors of Science in Psychology, University of Pittsburgh, 2016

Submitted to the Graduate Faculty of the

School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2024

UNIVERSITY OF PITTSBURGH

SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

**Dominika E. Oliver**

It was defended on

April 22, 2024

and approved by

Jenna C. Carlson, PhD, Assistant Professor, Departments of Human Genetics and Biostatistics

George C. Tseng, PhD, Professor, Department of Biostatistics

Reena S. Cecchini, PhD, Research Assistant Professor, Department of Biostatistics

Jeremy J. Martinson, DPhil, Assistant Professor, Departments of Infectious Diseases and Microbiology and Human Genetics

Thesis Advisor: Jenna C. Carlson, PhD, Assistant Professor, Departments of Human Genetics and Biostatistics

**Modeling Community and Genomic Factors of HIV Susceptibility in the *All of Us* Research Program**

Dominika Oliver, M.S.

University of Pittsburgh, 2024

*Objective:* To investigate the association between genes in the region of 46 million base pairs (MBP) and 47MBP on chromosome 3, community factors, and HIV susceptibility using the *All of Us* research program.

*Methods:* 4100 individuals enrolled in the *All of Us* research program, 2050 healthy controls and 2050 HIV patients, were propensity score-matched on age, sex, and race. Community factors from subject resident ZIP codes at time of enrollment were modeled using logistic regression against HIV susceptibility for all 4100 subjects. Separately, 3227 individuals with available short read genomic information had separate logistic regression models run on 64 different genetic variants from the chromosome 3 region of interest to determine their association with HIV susceptibility alone and controlling for community factors found to be significant in the community-factors only model. Relationships were considered statistically significant with a Bonferroni-corrected p-value of $8.8099 \times 10^{-5}$.

*Results:* In the community-only model, race/ethnicity, percentage of individuals on assisted income, percentage of individuals with at least a high school education, and percentage of vacant housing were found to be significantly related to HIV susceptibility. In the genomics-only models, 24 genetic variants were found to be statistically significantly related to HIV susceptibility. After controlling for community factors, no genetic variants were found to be statistically significantly related to HIV susceptibility.

*Conclusion:* This study found no significant genetic relationships to HIV susceptibility within the chromosome 3 region of interest after controlling for community factors. This is one of the first studies to model both community and HIV factors using a racially diverse cohort. Future studies should consider using a larger sample size and a larger genetic region of interest.

## Table of Contents

## List of Tables

## List of Figures

## Preface

I gratefully acknowledge *All of Us* participants for their contributions, without whom this research would not have been possible. I also thank the National Institutes of Health's *All of Us* Research Program for making available the participant data [and/or samples and/or cohort] examined in this study.

I also want to acknowledge Dr. Jenna Carlson, Alexis Cenname, and Kristina Boyd. Dr. Carlson has been my advisor throughout my Masters and has always been kind and supportive of my goals. Alexis Cenname is a peer of mine and her knowledge of the tools available in the *All of Us* Research Program was instrumental to this paper. Kristina Boyd is one of my best friends and her support, both as a statistician and a friend, meant so much to me over the years.

"The All of Us Research Program is supported by the National Institutes of Health, Office of the Director: Regional Medical Centers: 1 OT2 OD026549; 1 OT2 OD026554; 1 OT2 OD026557; 1 OT2 OD026556; 1 OT2 OD026550; 1 OT2 OD 026552; 1 OT2 OD026553; 1 OT2 OD026548; 1 OT2 OD026551; 1 OT2 OD026555; IAA #: AOD 16037; Federally Qualified Health Centers: HHSN 263201600085U; Data and Research Center: 5 U2C OD023196; Biobank: 1 U24 OD023121; The Participant Center: U24 OD023176; Participant Technology Systems Center: 1 U24 OD023163; Communications and Engagement: 3 OT2 OD023205; 3 OT2 OD023206; and Community Partners: 1 OT2 OD025277; 3 OT2 OD025315; 1 OT2 OD025337; 1 OT2 OD025276. In addition, the All of Us Research Program would not be possible without the partnership of its participants."

## 1.0 Introduction: HIV as a Public Health Issue

Human immunodeficiency virus (HIV) is a lentivirus whose acquisition is a major public health issue in the United States. In 2023, approximately 1.2 million people in the United live with this infection (a prevalence of 11.5 per 100,000 people), many of which (approximately 13%) may be unaware of due to lack of testing (Santos et al., 2015). HIV infection takes place in three stages, with the later stage becoming acquired immunodeficiency syndrome (AIDS), resulting in reduction or loss of immune function to fight even minor infections (Filip, 2023). In 2019 US dollars, the estimated average lifetime HIV-related medical cost for a person with HIV is $420,285 (2019 US$) (Bingham, Shrestha, Khurana, Jacobson, & Farnham, 2021).

HIV disproportionately affects minority populations, such as non-white or queer individuals. For example, Rich and colleagues show an exceptionally high prevalence of HIV among queer and minority sex workers compared to White and non-queer sex workers (Rich et al., 2017). Additionally, (Lewis, Herring, Chinnock, & Nelson, 2024). Additionally, innumerable studies highlight the disproportionate burden of HIV/AIDS among Black men (Lewis et al., 2024). Recent efforts have been made to increase the racial and ethnic diversity in curative HIV/AIDS treatment, but recent work by Dube and colleagues shows there is still a persistent gap in HIV research in diverse populations in the US (Dube et al., 2022).

Broadly speaking, there are two types of factors that influence HIV infection susceptibility: community factors and genomic factors. In the context of HIV infection research, community factors refer to the immediate physical and social surroundings of individuals that increase or decrease an individual's chance of contracting HIV (MacQueen et al., 2001). In contrast, genomic factors refer to interactions of an individual's genome with each other and with their environment,

influencing health and HIV disease risk (Lorenzo-Redondo, Ozer, Achenbach, D'Aquila, & Hultquist, 2021). While much research has been done on the factors that influence HIV infection, there have been none that combine both community and genomic factors in models of HIV susceptibility, and fewer that do so in diverse populations. To address this gap, in this study, using data from the *All of Us* Research Program, we combine community data along with genomics to investigate the interplay between environment and genetics in HIV susceptibility.

## 1.1 Factors Influencing Susceptibility of HIV Infection

### 1.1.1 Community Variables

In the context of HIV infection research, community factors refer to the immediate physical and social surroundings of individuals that increase or decrease an individual's chance of contracting HIV (MacQueen et al., 2001). MacQueen's definition of community factors can be further subdivided into demographic and economic factors.

### 1.1.2 Economic Factors

Frew et al. found that financial instability and poverty played a significant role in increasing risk-taking behavior in women, increasing likelihood of HIV infection (Frew et al., 2016). Women who had been diagnosed with HIV were interviewed, with factors in their lives being categorized as exosystem (community), mesosytem (network), microsystem, and individual. Factors such as

poverty, lack of access to education, housing instability, housing discrimination, and perceived discrimination all contributed to an increased likelihood of an HIV infection (Frew et al., 2016).

### 1.1.3 Demographic Factors

Race, ethnicity, and sexual orientation are related to HIV infection disparities. Benbow et al. examined which community, demographic, and economic factors had affected HIV infection rates among Latino populations across the United States. Using a mixed-effects Poisson model they found that counties with fewer Latinos, more rural, and had lower non-Latino-White prevalence tended to have higher disparities in HIV infection rates (Benbow, Aaby, Rosenberg, & Brown, 2020). Non-straight and queer women are more likely to contract HIV than their heterosexual counterparts (Frew et al., 2016).

### 1.1.4 Genomic Factors

Host genetic factors play a crucial role in determining susceptibility to HIV infection and progression of the disease to AIDS. Starting in the asymptomatic period, there are differences in individual responses. Some individuals stay asymptomatic and others experience general immune dysfunction followed by death. The progression phenotype is governed by a complex gamut of environmental and genetic factors (Lama & Planelles, 2007). One of those factors is genetic variation; homozygotic twins display less variation in susceptibility to HIV infection compared to heterozygous twins (Lama & Planelles, 2007). Another active area of research is the identification of host protein genes, since HIV relies on specific cellular proteins to invade host cells and

replicate, including proteins governing viral entry, RNA genome integration, transcription, translation, and virion assembly (Lama & Planelles, 2007).

The *CCR5* gene is one gene that has been shown to play a role in HIV susceptibility. The *CCR5* gene encodes a protein called the CCR5 receptor, which plays a role in the body's inflammatory immune response (Flanagan, 2014; Oppermann, 2004). HIV uses the CCR5 receptor as a co-receptor to enter host cells to begin its replication process (Flanagan, 2014). When HIV binds to the CD4 receptor on the cell surface, it also interacts with the CCR5 receptor, allowing viral entry (Oppermann, 2004). A naturally occurring mutation in the *CCR5* gene known as CCR5-Delta32 is a 32 base pair deletion in the *CCR5* gene, homozygosity for which confers strong protection against HIV-1. Heterozygous CCR5-delta32 carriers also show reduced susceptibility to HIV infection in exposed uninfected individuals (Flanagan, 2014; Lama & Planelles, 2007). This mutation is found in 10% of the population, primarily in individuals who are racially white and have Amish, Finish, and other European ancestry. This mutation becomes less prevalent in other racial groups: Individuals with African decent have an allele frequency of 1.8%, South Asian have a 1.7% frequency, and East Asian have a 0.016% frequency.

## 1.2 Study Objectives and Aims

The objectives of this study are to:

(1) Utilize the *All of Us* research database to construct a cohort of both HIV positive and negative individuals

(2) To conduct a case-control analysis of community and genetic factors influencing HIV susceptibility

The All of Us Research Program is a large-scale research repository that collects individual level demographic information, electronic health record, and genomic data, allowing researchers to easily obtain a statistically powerful dataset and utilize various data types without the cost of lab work. Utilizing this framework, we can match patients with Asymptomatic and Symptomatic HIV infection with be matched with healthy controls using propensity score matching on key demographic factors, as well as import information on CCR5 variants in order to build a logistic regression model that can predict HIV susceptibility.

# 2.0 Methods

## 2.1 Study Population

Data was obtained via the All of Us Research Program, a nationwide cohort study sponsored by the National Institutes of Health (NIH) that aims to further research in underrepresented populations. One of the biggest strengths of the program is the variety of information that it collects, which include survey data, physical measurements, electronic health records, and genomic data for consenting participants. Eligible participants must be 18 years of age and reside in the United States during time of enrollment. As of March 2023, there are over 700,00 participants in the program with over 400,000 participants sharing electronic health records and 250,000 short read genomic samples. Access to samples is divided into three tiers: Public, Registered, and Control.

## 2.2 Preprocessing Steps

Analysis was conducted in the Research Workbench, a cloud-based environment specific to the All of Us Research Program, a cloud-based environment that allows researchers to select participants, build a dataset, and conduct analysis within the All of Us framework. Analysis was conducted using both Python and R software packages.

### 2.2.1 Cohort Building and Covariate Selection

Data for the analytic cohort for this study was obtained in the *All of Us* Researcher Workbench via the Cohort builder, a tool that allows you to select cohorts of participants based on demographics, information in their electronic health record, availability of genomic information, etc. Two cohorts were built: one with asymptomatic HIV patients and one with symptomatic HIV patients. Asymptomatic and symptomatic HIV statuses were derived from *All of Us* database concept IDs which were in turn derived from ICD-10 codes from participant electronic health records. Participants were excluded if they declined to share demographic information with researchers, which included sex assigned at birth, gender, race, and Hispanic or Latino ethnicity origin, or did not have genomic information available. Individual non-genetic covariates extracted from the *All of Us* research database included gender, race, ethnicity, sex assigned at birth.

The *All of Us* research program asks participants for their ZIP code information and includes community variables from the American Communities Survey (ACS), a part of the Census Bureau. Additional variables were extracted for the sample to measure community-level demographic information through the ZIP-code linked ACS data including percent assisted income, percent of individuals in ZIP code with high school education, median income, percent of individuals in ZIP code without health insurance, percentage of individuals in the ZIP code living below the federal poverty limit, percentage of vacant housing, and Area Deprivation Index score.

### 2.2.2 Case-Control Matching

HIV patients were matched on age and race utilizing the R package MatchIt. MatchIT utilizes Propensity Score Matching, a statistical technique that reduces the effect of confounding

variables when estimating the effect of a "treatment" variable. The basic idea is to match participants from a "treatment" group to those with a similar propensity score (i.e., probability of being in the treatment group, which is based on based on possible covariates) from the control group. The goal is to ensure balance on covariates between the control group and the experimental group to enable causal inference of the treatment effect. MatchIt provides several options in how participants are matched, including nearest neighbor, exact, and kernel. In this study, propensity score matching was used to match HIV patients with control counterparts via age, sex at birth, and race, using the six nearest neighbors.

### 2.2.3 Genomic Data

Genomic data in the *All of Us* Research Program includes data derived from three modalities—short read whole genome sequencing (srWGS), long read whole genome sequencing, and microarray genotyping arrays. For this study, we used the v7 data release of the VariantDataSet (VDS), which is a data storage format for single nucleated polymorphisms (SNPs) and indel variant data called from srWGS.

*CCR5* genotypes were extracted using HAIL, an open-source, scalable framework utilizing Python developed by the Broad Institute. HAIL is used in the All of Us Research Program due to its utility in handling large-scale genomic data, and can provide several types of analysis including variant quality control, population stratification analysis, genome-wide association studies (GWAS), and variant annotation. HAIL utilizes a MatrixTable format, with each participant as the column and genes as the rows.

## 2.3 Logistic Regression

### 2.3.1 Modeling Infection Susceptibility

Logistic regressions are one of the most widely used statistical models for categorical outcomes. If there are only two events associated with the outcome, like the presence or absence of a disease state such as HIV, it can be described as a binary logistic regression. Let $p_i$ represent the outcome, HIV infection, occurring ($Y_i = 1$) with predictor ($X_i$) and adjustment covariates ($\beta Z_i$). The logistic regression equation is given by equation (1):

$$(1)\ logit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i + \boldsymbol{\beta Z_i}$$

$\beta_1$ is the fixed effect coefficient for the predictor. The null hypothesis for this model is given by:

$$H_0: \beta_1 = 0$$

This is equivalent to assuming an odds ratio equal to 1, or equivalently, that there is no statistically significant relationship between outcome $Y_i$ and predictor variable $X_i$ after adjusting for covariates $\boldsymbol{\beta Z_i}$. The odds ratio (OR) can be calculated as follows using equation (2):

$$(2)\ \widehat{OR} = \frac{\left(\frac{p_i}{1-p_i}\middle|(x_i = 1)\right)}{\left(\frac{p_i}{1-p_i}\middle|(x_i = 0)\right)} = \frac{\exp(\beta_0 + \beta_1 + \boldsymbol{\beta Z_i})}{\exp(\beta_0 + \boldsymbol{\beta Z_i})} = \exp(\beta_1)$$

## 3.0 Results

## 4.0 Summary Statistics

A total of 4,100 individuals, 2,050 in the HIV cases and 2,050 healthy controls were used in this analysis. Table 1 shows the results of the propensity score matching on age, race, and sex assigned at birth. Demographic covariates include Age, Gender, Sex at Birth, Race, and Hispanic or Latino Ethnicity. Community Variables are associated with the ZIP code that the participant is located in and include Percentage Poverty, Percent High School Graduation, Percent on Assisted Income, Percentage of Vacant Housing, Percentage of No Health Insurance, and Depravation Index. The model pseudo r-squared is 0.1175.

**Table 1. Descriptive Statistics between the Control Group and HIV Patient Group following propensity score matching**

| Characteristics | N | Overall N = 4100 | Control N = 2050 | HIV Patient N = 2050 | p-value |
|---|---|---|---|---|---|
| **Age** | 4100 | 57 (43, 60) | 55 (38, 68) | 58 (48, 64) | < 0.001 |
| **Sex at Birth** | 4100 | | | | 0.025 |
| Female | | 1,495 (36%) | 782 (38%) | 713 (35%) | |
| Male | | 2,605 (64%) | 1,268 (62%) | 1,337 (65%) | |
| **Gender** | 4100 | | | | 0.041 |
| Female | | 1499 (37%) | 781 (38%) | 713 (35%) | |
| Male | | 2,601 (63%) | 1,269 (62%) | 1,332 (65%) | |
| **Race** | 4100 | | | | <0.001 |
| Black or African American | | 2058 (50%) | 819 (40%) | 1,239 (60%) | |
| Other | | 694 (17%) | 634 (31%) | 60 (2.9%) | |
| Unknown | | 312 (7.6%) | 0 (0%) | 312 (15%) | |
| White | | 1,036 (25%) | 597 (29%) | 439 (21%) | |
| **Ethnicity** | 4100 | | | | <0.001 |
| Hispanic or Latino | | 409 (10%) | 51 (2.5%) | 358 (17%) | |

| | | | | | |
|---|---|---|---|---|---|
| Not Hispanic or Latino | | 3,619 (90%) | 1,999 (98%) | 1,692 (83%) | |
| **Assisted Income** | 4100 | 16 (13, 22) | 15 (11, 19) | 17 (15, 22) | <0.001 |
| **High School Education** | 4100 | 87 (83, 90) | 87 (83, 91) | 86 (83, 89) | <0.001 |
| **Median Income** | 4100 | 59,191 (55,344,71,783) | 61,024 (55,324, 74,084) | 57,307 (56,249,65,575) | <0.001 |
| **No Health Insurance** | 4100 | 10.6 (6.9, 12.9) | 10.3 (6.9, 12.9) | 11.0 (6.9, 12.9) | <0.001 |
| **Vacant Housing** | 4100 | 11.6 (6.7, 12.4) | 10.4 (6.6, 12.4) | 12.2 (7.7,12.4) | <0.001 |
| **Poverty** | 4100 | 18 (14,21) | 16 (13, 21) | 18 (16, 21) | <0.001 |
| **Depravation Index** | 4100 | 0.34 (0.30, 0.39) | 0.33 (0.29, 0.39) | 0.35 (0.31, 0.39) | <0.001 |

## 4.1 Logistic Regression Models

### 4.1.1 Community Model

In order to assess what covariates would be included in the full genomic model, a model with only the community variables was created. A logistic model was created using age, race, ethnicity, gender, sex at birth, assisted income, high school education, median income, poverty, no health insurance, vacant housing, and deprivation index. In order to assess collinearity, variance inflation factor (VIF) was calculated. If a variable's VIF is larger than 10, it is likely that they are colinear with other variables in the model and redundant. Sex at Birth, Gender, and Deprivation Index all had VIF's over 40, so Gender and Depravation index were removed. This makes sense since Gender and Sex at Birth are very similar values, and Deprivation Index is an index made of the other community factors.

The final community model found that for demographic factors, Race (OR = .71) and Ethnicity (OR = 0.06) greatly contributed to HIV susceptibility, while Age and Sex at Birth did not contribute. Community Factors that were highly statistically significant were Assisted Income (OR = 1.13), High School Education (OR = 1.10), Vacant Housing (OR = 0.96), and No Health Insurance (OR = 1.06).

**Table 2. Summary of community variable logistic regression model**

| Characteristic | Odds Ratio | 95% CI | p-value |
|---|---|---|---|
| **Age** | 1.01 | 1.00, 1.01 | 0.009 |
| **Race** | 0.71 | 0.63, 0.79 | <0.001 |

| | | | |
|---|---|---|---|
| **Ethnicity** | 0.06 | 0.04, 0.09 | <0.001 |
| **Sex at Birth** | 0.90 | 0.77, 1.04 | 0.14 |
| **Assisted Income** | 1.13 | 1.11, 1.16 | <0.001 |
| **High School Education** | 1.10 | 1.08, 1.13 | <0.001 |
| **Poverty** | 1.01 | 0.98, 1.03 | 0.7 |
| **Vacant Housing** | 0.96 | 0.93, 0.98 | <0.001 |
| **No Health Insurance** | 1.06 | 1.04, 1.08 | <0.001 |

### 4.1.2 Genomic Models

Of our initial population of 4100, 3447 samples had genomic data available and after accounting for sample relatedness, the final model contains 3227 participants and 563 genetic variants. Sample relatedness was accounted for using the srWGS Related Kinship Score, with kinship scores above 0.1 removed from the final sample ("How the All of Us Genomic Data are Organized," 2024).

Logistic regression was performed separately for each genetic variant with community variables selected as covariates. When building each model, Gender, assisted income and ZIP code deprivation index covariates needed to be removed due to their similarity to other variables in the model. The model examined the region on Chromosome 3 at the loci between 46M and 47M as this is where the CCR5 gene and similar genes are located. HAIL does not utilize a p-value correction method, so a Bonferroni correction was calculated by dividing the value used for statistical significance (0.05) by the 563 variants, meaning that a variant needed to have a p-value below $8.8099 \times 10^{-5}$ to be statistically significant.

For each genetic variant, two models were created: one with only genomic information and one with community and demographic covariates. The purely genomic model creates a table and applies a logistic regression to each genetic variant (equation (3)). The second model adds to the first one by including the list of covariates (equation (4)). Here $p_i$is the probability of being an HIV patient, $X_i$ is the genotype being modeled, and, $\boldsymbol{\beta_i Z_i}$ are the community variables.

$$(3)\ logit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i$$

$$(4)\ logit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_i + \boldsymbol{\beta_i Z_i}$$

The model with only genomic information found that there were 24 genetic variants that were statistically significant utilizing a Bonferroni-corrected p-value. Allele frequencies from the *All of Us* Research Program were obtained via the *All of Us* Data Browser. General population allele frequencies were obtained via the Genome Aggregation Database (gnomAD).

**Table 3. Summary of 10 genetic variants with lowest p-value in the logistic regression model with only genetics**

| Locus | Alleles | Beta | Standard Error | Z-stat | p-value | N | AoU Allele Frequency | Population Allele Frequency |
|---|---|---|---|---|---|---|---|---|
| chr3:46897696 | [C, A] | 0.467691 | 0.055594 | 8.412683 | <1e10-6 | 4 | 0.110965 | 0.1234 |
| chr3:46859905 | [C, T] | 0.475939 | 0.057126 | 8.33138 | <1e10-6 | 4 | 0.104430 | 0.1144 |
| chr3:46860639 | [T, G] | 0.321823 | 0.044 | 7.31417 | <1e10-6 | 4 | 0.321705 | 0.1698 |
| chr3:46370771 | [C, T] | 0.530833 | 0.07894 | 6.724489 | <1e10-6 | 4 | 0.052021 | 0.05801 |
| chr3:46863218 | [C, A] | 0.286968 | 0.045018 | 6.374494 | <1e10-6 | 4 | 0.274584 | 0.1567 |
| chr3:46710663 | [G, A] | -0.725203 | 0.122021 | -5.94328 | <1e10-6 | 4 | 0.076727 | 0.07895 |
| chr3:46902878 | [T, C] | 0.423149 | 0.071801 | 5.893344 | <1e10-6 | 4 | 0.074302 | 0.02651 |
| chr3:46708966 | [T, C] | -0.930894 | 0.165003 | -5.641664 | <1e10-6 | 5 | 0.041635 | 0.04863 |
| chr3:46902784 | [T, C] | 0.307122 | 0.054849 | 5.59942 | <1e10-6 | 4 | 0.655902 | 0.6624 |
| chr3:46701258 | [C, A] | -0.914864 | 0.165359 | -5.532597 | <1e10-6 | 5 | 0.041793 | 0.1142 |

When accounting for covariates, there were no genetic variants that passed the p-value threshold for statistical significance. Listed below are the four genetic variants with the lowest p-values when covariates were added.

**Table 4. Summary of genetic variants with lowest p-value in the logistic regression model accounting for covariates**

| Locus | Alleles | Beta | Standard Error | Z stat | p-value | N | AoU Allele Frequency | Population Allele Frequency |
|---|---|---|---|---|---|---|---|---|
| chr3:46701257 | ["G","A"] | -5.40e01 | 1.83e-01 | -2.95e+00 | 3.23e-03 | 4 | 0.0417 | 0.04837 |
| chr3:46701258 | ["C","A"] | -5.44e-01 | 1.84e-01 | -2.96e+00 | 3.06e-03 | 4 | 0.041793 | 0.04841 |
| chr3:46708966 | ["T","C"] | -5.48e-01 | 1.83e-01 | -2.99e+00 | 2.79e-03 | 4 | 0.041635 | 0.04863 |
| chr3:46863767 | ["C","T"] | -7.19e-01 | 2.51e-01 | -2.86e+00 | 4.21e-03 | 4 | 0.021570 | 0.02207 |

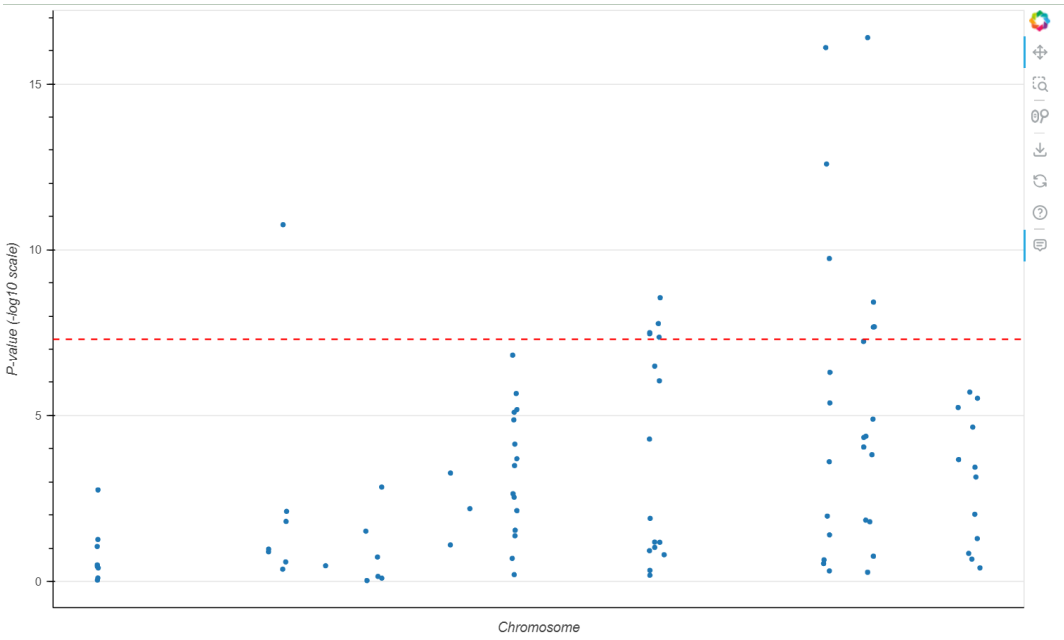


**Figure 1. Manhattan plot of unadjusted logistic regression of only genetic variables**

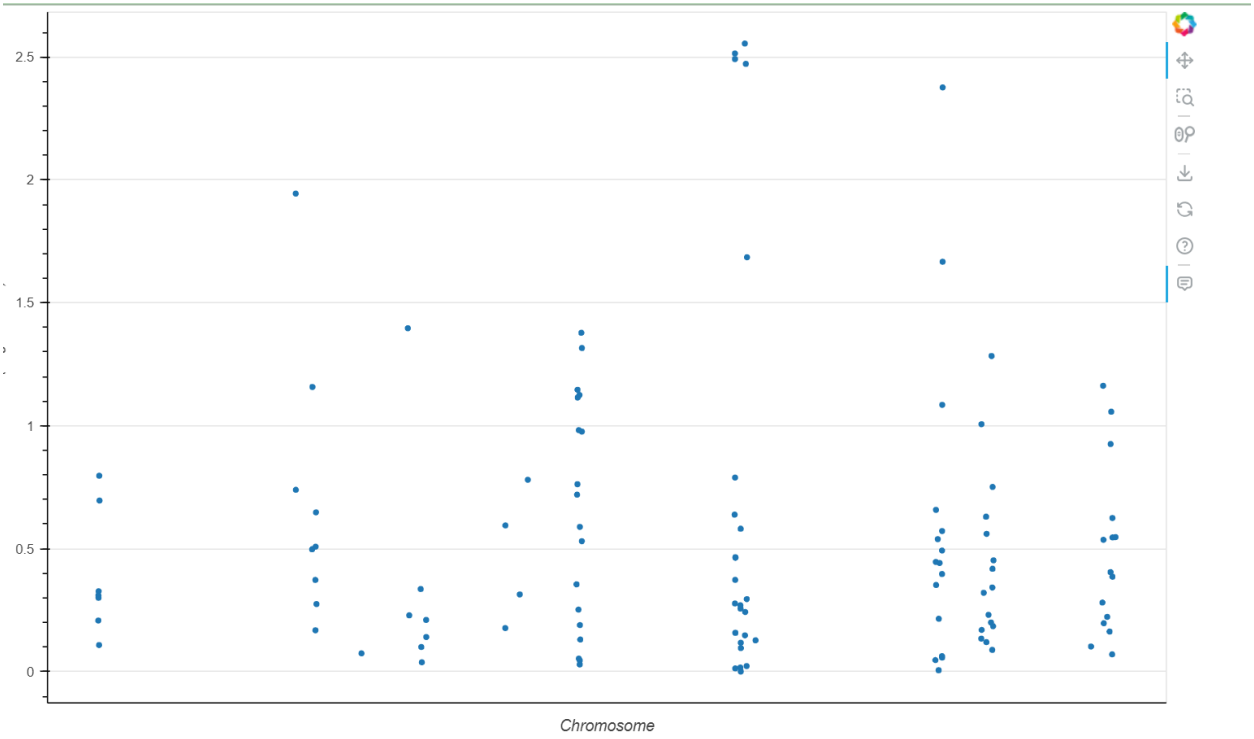


**Figure 2. Manhattan plot of adjusted logistic regression results of variants and community and demographic covariates**

## 5.0 Discussion

In the United States, approximately 0.36% of the population is diagnosed with HIV. In the *All of Us* Research program, approximately 2.04% of the study population is an HIV patient, which is higher than the general population.

When examining HIV susceptibility, individuals with Hispanic or Latino ethnicity are more susceptible to HIV, and Black and African American individuals are more susceptible to HIV than their white counterparts. This was touched upon by Frew et al. and Benbow when examining community factors, finding increased susceptibility due to race and ethnicity factors. HIV susceptibility also increases in zip codes as the percentage of individuals receiving assisted income increases, percentage of high school education increases, percent of individuals without health insurance increases, and the amount of vacant housing decreases. The effect sizes are small (OR of 1.13, 1.10, 1.06, and 0.96) and these metrics do not paint the full picture of a community nor the individual factors that may drive risk-taking behavior.

When examining the genetic variants, it is interesting to note the differences between the genetic variants that are statistically significant in the genetic-only model compared to the model with genetics and covariates. Several of the variants with the lowest p-values in the community model are located around the 46.70M loci, which are in proximity to the ALS2CL gene, which is primarily expressed in the esophagus but also expressed in the skin and possibly affecting the skin barrier in a way leading to decreased susceptibility. The variants in the pure genomic model are further down the region (46.85-46.89M), encompassing genes such as PTH1R and MYL3. Interestingly, variants such as the 3-46897696-C-A and 3-46859905-C-T variants are primarily found in individuals with an African decent, while many of the other significant variants are found

in higher proportions individuals with a European ancestry. Due to the focus on recruiting underserved populations, there are some gene variants with differing frequencies between the All of Us study population and the general population that has been studied thus far.

## 5.1 Strengths and Limitations

This analysis has several strengths. One of the strengths are the diversity of the population being examined. Many genomic studies focus on patients with a Eurocentric ancestry, but the cohort built from this analysis had significant non-White representation. A second strength of this analysis was the inclusion of both genomic and community variables, which made this study relatively novel.

This analysis has several limitations. First, there is participation bias. Individuals who participate in the *All of Us* study likely differ from those who did not participate in the *All of Us* study. Second, there is information bias originating from self-selection bias, since only three-quarters of the identified cohort had genomic information available. Participation in the genomic portion of *All of Us* was voluntary, and those that elected to participate likely differed from those that did not elect to participate. Finally, the community variables were from the patient's ZIP code, and information on how long a patient was a resident of that community was not available, and are not as informative as individual variables. The overall allele frequency of the *CCR5* Delta 32 mutation in the *All of Us* Research Program is 7.026%, which is lower than the general population. Together these limitations limit the external validity and overall generalizability of this analysis.

## 5.2 Future Considerations

This analysis only examined a small portion of the genome, a section of about 1 million base pairs on Chromosome 3. In order to examine the full scope of the role of genomics in susceptibility, future analysis can utilize a GWAS to examine all relevant genes in the genome.

A more thorough genetic association study may be beneficial in examining the differences in the severity of HIV symptoms between individuals. The All of Us Research Program contains both electronic health record data as well as testing done on samples, making it possible to identify individuals in different stages of the pathology.

The *CCR5* Delta 32 mutation and the genes surrounding it are also found to be connected to other infectious disease, including COVID infection. It has been found that the inhibition of the CCR5 receptor results in the relief of symptoms for those who are infected with Covid-19. Additionally, this region has been associated with other infectious diseases such as *Toxoplasma gondii* and *Staphylococcus aureus*, and should be included in further research using the *All of Us* Research Program.

## 5.3 Public Health Implications

Modeling susceptibility is how we can determine what populations are most at risk and therefore can allocate more resources towards prevention. This is particularly important for state health departments to identify counties that may have higher rates of individuals without health insurance or other factors that have been associated with increased susceptibility.

Exploring which genomic factors may be associated increased or decreased susceptibility to HIV infection is important to both identify populations in need of intervention and explore new mechanisms for pharmacological intervention. Variants such as the *CCR5* Delta 32 mutation show us the mechanisms of how an individual can be infected with HIV, and by examining these differences in a diverse population, we may find better ways to treat and prevent infection.

## Appendix A Tables

**Appendix Table 1. Descriptive Statistics between the Control Group and HIV Patient Group following propensity score matching**

| Characteristics | N | Overall N = 4100 | Control N = 2050 | HIV Patient N = 2050 | p-value |
|---|---|---|---|---|---|
| **Age** | 4100 | 57 (43, 60) | 55 (38, 68) | 58 (48, 64) | < 0.001 |
| **Sex at Birth** | 4100 | | | | 0.025 |
| Female | | 1,495 (36%) | 782 (38%) | 713 (35%) | |
| Male | | 2,605 (64%) | 1,268 (62%) | 1,337 (65%) | |
| **Gender** | 4100 | | | | 0.041 |
| Female | | 1499 (37%) | 781 (38%) | 713 (35%) | |
| Male | | 2,601 (63%) | 1,269 (62%) | 1,332 (65%) | |
| **Race** | 4100 | | | | <0.001 |
| Black or African American | | 2058 (50%) | 819 (40%) | 1,239 (60%) | |
| Other | | 694 (17%) | 634 (31%) | 60 (2.9%) | |
| Unknown | | 312 (7.6%) | 0 (0%) | 312 (15%) | |
| White | | 1,036 (25%) | 597 (29%) | 439 (21%) | |
| **Ethnicity** | 4100 | | | | <0.001 |
| Hispanic or Latino | | 409 (10%) | 51 (2.5%) | 358 (17%) | |
| Not Hispanic or Latino | | 3,619 (90%) | 1,999 (98%) | 1,692 (83%) | |
| **Assisted Income** | 4100 | 16 (13, 22) | 15 (11, 19) | 17 (15, 22) | <0.001 |
| **High School Education** | 4100 | 87 (83, 90) | 87 (83, 91) | 86 (83, 89) | <0.001 |

| | | | | | |
|---|---|---|---|---|---|
| **Median Income** | 4100 | 59,191 (55,344,71,783) | 61,024 (55,324, 74,084) | 57,307 (56,249,65,575) | <0.001 |
| **No Health Insurance** | 4100 | 10.6 (6.9, 12.9) | 10.3 (6.9, 12.9) | 11.0 (6.9, 12.9) | <0.001 |
| **Vacant Housing** | 4100 | 11.6 (6.7, 12.4) | 10.4 (6.6, 12.4) | 12.2 (7.7,12.4) | <0.001 |
| **Poverty** | 4100 | 18 (14,21) | 16 (13, 21) | 18 (16, 21) | <0.001 |
| **Depravation Index** | 4100 | 0.34 (0.30, 0.39) | 0.33 (0.29, 0.39) | 0.35 (0.31, 0.39) | <0.001 |

**Appendix Table 2. Summary of community variable logistic regression model**

| Characteristic | Odds Ratio | 95% CI | p-value |
|---|---|---|---|
| **Age** | 1.01 | 1.00, 1.01 | 0.009 |
| **Race** | 0.71 | 0.63, 0.79 | <0.001 |
| **Ethnicity** | 0.06 | 0.04, 0.09 | <0.001 |
| **Sex at Birth** | 0.90 | 0.77, 1.04 | 0.14 |
| **Assisted Income** | 1.13 | 1.11, 1.16 | <0.001 |
| **High School Education** | 1.10 | 1.08, 1.13 | <0.001 |
| **Poverty** | 1.01 | 0.98, 1.03 | 0.7 |
| **Vacant Housing** | 0.96 | 0.93, 0.98 | <0.001 |
| **No Health Insurance** | 1.06 | 1.04, 1.08 | <0.001 |

**Appendix Table 3. Summary of 10 genetic variants with lowest p-value in the logistic regression model with only genetics**

| Locus | Alleles | Beta | Standard Error | Z-stat | p-value | N | AoU Allele Frequency | Population Allele Frequency |
|---|---|---|---|---|---|---|---|---|
| chr3:46897696 | [C, A] | 0.467691 | 0.055594 | 8.412683 | <1e10-6 | 4 | 0.110965 | 0.1234 |
| chr3:46859905 | [C, T] | 0.475939 | 0.057126 | 8.33138 | <1e10-6 | 4 | 0.104430 | 0.1144 |
| chr3:46860639 | [T, G] | 0.321823 | 0.044 | 7.31417 | <1e10-6 | 4 | 0.321705 | 0.1698 |
| chr3:46370771 | [C, T] | 0.530833 | 0.07894 | 6.724489 | <1e10-6 | 4 | 0.052021 | 0.05801 |
| chr3:46863218 | [C, A] | 0.286968 | 0.045018 | 6.374494 | <1e10-6 | 4 | 0.274584 | 0.1567 |
| chr3:46710663 | [G, A] | -0.725203 | 0.122021 | -5.94328 | <1e10-6 | 4 | 0.076727 | 0.07895 |
| chr3:46902878 | [T, C] | 0.423149 | 0.071801 | 5.893344 | <1e10-6 | 4 | 0.074302 | 0.02651 |
| chr3:46708966 | [T, C] | -0.930894 | 0.165003 | -5.641664 | <1e10-6 | 5 | 0.041635 | 0.04863 |
| chr3:46902784 | [T, C] | 0.307122 | 0.054849 | 5.59942 | <1e10-6 | 4 | 0.655902 | 0.6624 |
| chr3:46701258 | [C, A] | -0.914864 | 0.165359 | -5.532597 | <1e10-6 | 5 | 0.041793 | 0.1142 |

**Appendix Table 4. Summary of genetic variants with lowest p-value in the logistic regression model accounting for covariates**

| Locus | Alleles | Beta | Standard Error | Z stat | p-value | N | AoU Allele Frequency | Population Allele Frequency |
|---|---|---|---|---|---|---|---|---|
| chr3:46701257 | ["G","A"] | -5.40e01 | 1.83e-01 | -2.95e+00 | 3.23e-03 | 4 | 0.0417 | 0.04837 |
| chr3:46701258 | ["C","A"] | -5.44e-01 | 1.84e-01 | -2.96e+00 | 3.06e-03 | 4 | 0.041793 | 0.04841 |
| chr3:46708966 | ["T","C"] | -5.48e-01 | 1.83e-01 | -2.99e+00 | 2.79e-03 | 4 | 0.041635 | 0.04863 |
| chr3:46863767 | ["C","T"] | -7.19e-01 | 2.51e-01 | -2.86e+00 | 4.21e-03 | 4 | 0.021570 | 0.02207 |

## Appendix B R Code

Note: Due to privacy concerns, the code below is just a framework.

```
library(tidyverse)
library(bigrquery)
library(lubridate)
```

Data for each cohort was called using concept ID's from a database corresponding to demographic and zip code data.

```
listDF <- list(demographics, zip)

Asymptomatic <- listDF %>% reduce(inner_join, by='person_id')
Lastdate <- "2024-03-01"

calc_age <- function(birthDate, refDate = Sys.Date(), unit = "year") {

    require(lubridate)

    if(grepl(x = unit, pattern = "year")) {
        as.period(interval(birthDate, refDate), unit = 'year')$year
    } else if(grepl(x = unit, pattern = "month")) {
        as.period(interval(birthDate, refDate), unit = 'month')$month
    } else if(grepl(x = unit, pattern = "week")) {
        floor(as.period(interval(birthDate, refDate), unit = 'day')$day / 7)
    } else if(grepl(x = unit, pattern = "day")) {
        as.period(interval(birthDate, refDate), unit = 'day')$day
    } else {
        print("Argument 'unit' must be one of 'year', 'month', 'week', or 'da
y'")
        NA
    }
```

```
}

Asymptomatic <- Asymptomatic %>% mutate(ageyear = calc_age(date_of_birth, Las
tdate))


str(Asymptomatic)

gender <- ggplot(data = Asymptomatic) +

            stat_count(mapping = aes(x = gender))


gender

Remove <- list("Gender Identity: Additional Options", "I prefer not to answer
", "Not man only, not woman only, prefer not to answer, or skipped", "PMI: Sk
ip", "Gender Identity: Non Binary", "Gender Identity: Transgender")


Asymptomatic <- Asymptomatic[ ! Asymptomatic$gender %in% Remove, ]

Asymptomatic$genderfact=ifelse(Asymptomatic$gender=='Female', 1 ,

                       ifelse(Asymptomatic$gender=='Male', 2,

                              ifelse(Asymptomatic$gender=='Gender Identity: T
ransgender', 3, 4)))

table(Asymptomatic$race)


RemoveRace <- list("PMI: Skip")

Asymptomatic <- Asymptomatic[ ! Asymptomatic$race %in% RemoveRace, ]

notindicated <- list("I prefer not to answer", "None Indicated", "None of the
se")

replacenone <- "unknown"


otherrace <- list("Asian", "Middle Eastern or North African", "Native Hawaiia
n or Other Pacific Islander", "More than one population")

replaceother <- "other"


Asymptomatic <- Asymptomatic %>% mutate(race = ifelse(race %in% notindicated,
replacenone, race))


Asymptomatic <- Asymptomatic %>% mutate(race = ifelse(race %in% otherrace, re
placeother, race))

Asymptomatic$racefact=ifelse(Asymptomatic$race=="White", 1, ifelse(Asymptomat
ic$race=="Black or African American", 2, ifelse(Asymptomatic$race=="other", 3
, 4)))
```

```
table(Asymptomatic$racefact)

#Ethnicity

table(Asymptomatic$ethnicity)


Removeeth <- list("PMI: Prefer Not To Answer", "What Race Ethnicity: Race Eth
nicity None Of These")


Asymptomatic <- Asymptomatic[ ! Asymptomatic$ethnicity %in% Removeeth, ]

Asymptomatic$ethnicityfact=ifelse(Asymptomatic$ethnicity=="Hispanic or Latino
", 1, 2)

table(Asymptomatic$sex_at_birth)


Removesex <- list("I prefer not to answer", "Intersex", "No matching concept"
, "None", "PMI: Skip")


Asymptomatic <- Asymptomatic[ ! Asymptomatic$sex_at_birth %in% Removesex, ]


Asymptomatic$sex_at_birthfact=ifelse(Asymptomatic$sex_at_birth=="Male", 1, 2)


table(Asymptomatic$sex_at_birthfact)

Asymptomatic$hivcase <- c(1)


Asymptomatic$hivseverity <- c(1)


Asymptomatic <- Asymptomatic %>% sample_n(2000, replace = FALSE)

Symptomatic$hivcase <- c(1)


Symptomatic$hivseverity <- c(2)

#Added variables for HIV suseptibility and severity


control$hivcase <- c(0)


control$hivseverity <- c(0)

dim(Asymptomatic)

dim(Symptomatic)
```

```
dim(control)


master <- rbind(Asymptomatic, Symptomatic, control)


master$a
m.out <- matchit(hivcase ~ ageyear + sex_at_birthfact + racefact, data = mast
er,

                 method = "nearest",


                 distance = "glm", ratio = 1)


summary(m.out)


#plotting the balance between smokers and non-smokers
plot(m.out, type = "jitter", interactive = FALSE)
plot(summary(m.out), abs = FALSE)


#put matched pairs into own dataset
mastermatched <- match.data(m.out)
```

#cleaning final dataset

```
str(mastermatched)


finalset <- subset(mastermatched, select = -c(gender_concept_id, date_of_birt
h, race_concept_id, ethnicity_concept_id, sex_at_birth_concept_id))
Communitymodel <- glm(hivcase ~ ageyear + racefact + ethnicityfact + sex_at_b
irthfact + genderfact + assisted_income + high_school_education + poverty  +
vacant_housing + deprivation_index + no_health_insurance, data = finalset, fa
mily = "binomial")
summary(Communitymodel)
install.packages("car")


library(car)
f_calculate_vif <- function(fit) {
  v <- c(v <- car::vif(fit))
```

```
  m <- cbind(v, 1/v)
  colnames(m) <- c("VIF", "1/VIF")
  print(m)
  cat("Mean VIF: ", mean(v))
}


f_calculate_vif(Communitymodel)
Communitymodel2 <- glm(hivcase ~ ageyear + racefact + ethnicityfact + sex_at_
birthfact + assisted_income + high_school_education + poverty + vacant_housin
g + no_health_insurance, data = finalset, family = "binomial")


summary(Communitymodel2)
f_calculate_vif(Communitymodel2)
```

Hosmer-Lemeshow GOF Test

```
#install.packages("performance")
library(performance)


performance::performance_hosmer(Communitymodel2, n_bins = 20)
```

logit summary

```
logit_summary <- function(x){
  stopifnot("glm" %in% class(x)) # input must be of class 'glm'


  preds <- unlist(strsplit(as.character(x$formula[3]), # extract predictors u
sed
                           split = "[[:space:]]\\+[[:space:]]"))
  LL <- stats::logLik(x) # log likelihood
  y <- as.character(x$formula[2]) # outcome variable
  tStat <- with(x, null.deviance - deviance) # chi-square test statistic
  McF <- signif(1 - logLik(x)/logLik(glm(as.formula(paste(y, "1", sep = "~"))
, # McFadden's Pseudo R^2
                              family = binomial(link = "logit"), data = x$data
)), digits = 4)
  AIC <- x$aic # Akaike information criterion
  BIC <- (-2 * LL) + (log(length(x$residuals)) * (length(preds) + 1)) # Bayes
ian Information Criterion
```

```
  pval <- signif(with(x, stats::pchisq(null.deviance - deviance, # p-value of
model
                           df.null - df.residual, lower.tail = FALSE)), digits
= 4)


  mod_stats <- merge(summary(x)$coefficients, exp(confint.default(x)), by = "
row.names") # model stats
  mod_stats$`Odds Ratio` <- exp(mod_stats$Estimate) # add 'Odds Ratio'
  mod_stats <- subset(mod_stats, select = -Estimate) # drop 'Estimate'
  mod_stats <- mod_stats[,c(1, ncol(mod_stats),2:(ncol(mod_stats)-1))] # reor
der columns







  tbl <- data.frame(nrow = length(preds), ncol = 5) # data.frame


  output <- list(LL, y, tStat, McF,AIC, BIC, pval, tbl) # list of diagnostics
  names(output) <- c("log likelihood", "outcome", "LR chi2", "Pseudo R^2",
                     "AIC", "BIC", "Prob > chi2","results") # names for list


  output$results <- mod_stats


  return(output)
}
logit_summary(Communitymodel2)
```

# Appendix C Python Code

Note: Due to privacy concerns, the code below is just a framework.

```
#import Packaages

import os
import subprocess
import numpy as np
import pandas as pd
import pandas_profiling
import plotnine
from plotnine import *  # Provides a ggplot-like interface to matplotlib.
from IPython.display import display

## Plot setup.
theme_set(theme_bw(base_size = 11)) # Default theme for plots.

def get_boxplot_fun_data(df):
  """Returns a data frame with a y position and a label, for use annotating ggplot boxplots.

  Args:
    d: A data frame.
  Returns:
    A data frame with column y as max and column label as length.
  """
  d = {'y': max(df), 'label': f'N = {len(df)}'}
  return(pd.DataFrame(data=d, index=[0]))

genotype = my_dataframe

#clean and prepare data to join to genomic data.

#Remove Unnessicary Columns (factors, ACS Year, Propensity Score Matching metrics)
genoremoved = genotype.drop(['american_community_survey_year','zip_code', 'observation_datetime',
'genderfact', 'racefact', 'ethnicityfact', 'sex_at_birthfact', 'distance', 'weights', 'subclass', 'outcome'], axis=1)

#Create Dummy Variables
genodummy = pd.get_dummies(genoremoved.set_index(['person_id'])).reset_index()

genodummy['hivcase'].value_counts()

genodummy['hivcase'] = genodummy['hivcase'].astype(int)

genodummy.columns = genodummy.columns.str.replace(' ', '')

genodummy.dtypes

#convert booleans to integers

genodummy.gender_Female = genodummy.gender_Female.replace({True: 1, False: 0})
```

```
genodummy.gender_Male = genodummy.gender_Male.replace({True: 1, False: 0})
genodummy.race_BlackorAfricanAmerican = genodummy.race_BlackorAfricanAmerican.replace({True: 1,
False: 0})
genodummy.race_White = genodummy.race_White.replace({True: 1, False: 0})
genodummy.race_other = genodummy.race_other.replace({True: 1, False: 0})
genodummy.race_unknown = genodummy.race_unknown.replace({True: 1, False: 0})
genodummy.ethnicity_HispanicorLatino = genodummy.ethnicity_HispanicorLatino.replace({True: 1, False:
0})
genodummy.ethnicity_NotHispanicorLatino  =  genodummy.ethnicity_NotHispanicorLatino.replace({True:
1, False: 0})
genodummy.sex_at_birth_Female = genodummy.sex_at_birth_Female.replace({True: 1, False: 0})
genodummy.sex_at_birth_Male = genodummy.sex_at_birth_Male.replace({True: 1, False: 0})

#save as a tsv to be complient with genotype data
genodummy["person_id"] = genodummy["person_id"].astype(str)

genotypes = genodummy
genotypes.dtypes

genotypes.to_csv('hivphenotypes.tsv', index=False, sep='\t')

#import packages for genomic analysis
import hail as hl
from hail.plot import show
from bokeh.plotting import output_file, save
import bokeh.io
from bokeh.io import *
from bokeh.resources import INLINE
#bokeh.io.output_notebook(INLINE)
%matplotlib inline

#Pathways utilized:  GRCh38 reference genome, Clinvar variants

#genomic region that we are examining
test_intervals = ['chr3:46M-47M']

mt = hl.filter_intervals(
   mt,
   [hl.parse_locus_interval(x,)
    for x in test_intervals])

#upload phenotypes
phenotype_filename

phenoandgeno = hl.import_table(phenotype_filename,
                        types={'assisted_income': hl.tfloat64, 'high_school_education': …},
                        key='person_id')

mt = mt.semi_join_cols(phenoandgeno)

mt = mt.annotate_cols(pheno = phenoandgeno[mt.s])

#remove related samples
related_remove = hl.import_table(related_samples_path,
                          types={"sample_id":"tstr"},
                         key="sample_id")
```

```
mt = mt.anti_join_cols(related_remove)

#filter variants that are prevalent in less than 1% of the population
filtered_mt = mt.filter_rows(hl.min(mt.variant_qc.AF) > 0.01, keep = True)

mt = filtered_mt

#List of covariates
c2=[1.0, mt.pheno.ethnicity_HispanicorLatino, mt.pheno.ageyear, mt.pheno.race_BlackorAfricanAmerican,
    mt.pheno.race_White, mt.pheno.sex_at_birth_Male, mt.pheno.poverty,mt.pheno.vacant_housing,
mt.pheno.assisted_income, mt.pheno.high_school_education, mt.pheno.no_health_insurance]

#Create Additive Logistic Regression Model
log_reg = hl.logistic_regression_rows(
   test='wald',
   y=mt.pheno.hivcase,
   x=mt.GT.n_alt_alleles(),
   covariates=[1.0]
)

#Create Additive Logistic Regression Model with Community Covariates
log_reg_community= hl.logistic_regression_rows(
   test='wald',
   y=mt.pheno.hivcase,
   x=mt.GT.n_alt_alleles(),
   covariates=c2
)

#Filter to only genes that pass the Bonferroni Correction
filteredtable = log_reg.filter(log_reg.p_value < 8.8099e-5)

filteredtable.count()

#Create and output Manhattan Plots
p2 = hl.plot.manhattan(log_reg_community.p_value)

output_file("manhattannovariates2.html")
save(p2)
```

## Bibliography


Benbow, N. D., Aaby, D. A., Rosenberg, E. S., & Brown, C. H. (2020). County-level factors affecting Latino HIV disparities in the United States. *PLoS One, 15*(8), e0237269. doi:10.1371/journal.pone.0237269

Bingham, A., Shrestha, R. K., Khurana, N., Jacobson, E. U., & Farnham, P. G. (2021). Estimated Lifetime HIV-Related Medical Costs in the United States. *Sex Transm Dis, 48*(4), 299-304. doi:10.1097/OLQ.0000000000001366

Dube, K., Kanazawa, J., Campbell, C., Boone, C. A., Maragh-Bass, A. C., Campbell, D. M., . . . Sauceda, J. A. (2022). Considerations for Increasing Racial, Ethnic, Gender, and Sexual Diversity in HIV Cure-Related Research with Analytical Treatment Interruptions: A Qualitative Inquiry. *AIDS Res Hum Retroviruses, 38*(1), 50-63. doi:10.1089/AID.2021.0023

Filip, I. (2023). The steep cost of HIV treatment interruptions. *AIDS, 37*(14), N17. doi:10.1097/QAD.0000000000003714

Flanagan, C. A. (2014). Receptor conformation and constitutive activity in CCR5 chemokine receptor function and HIV infection. *Adv Pharmacol, 70*, 215-263. doi:10.1016/b978-0-12-417197-8.00008-0

Frew, P. M., Parker, K., Vo, L., Haley, D., O'Leary, A., Diallo, D. D., . . . Team, H. I. V. P. T. N. S. (2016). Socioecological factors influencing women's HIV risk in the United States: qualitative findings from the women's HIV SeroIncidence study (HPTN 064). *BMC Public Health, 16*(1), 803. doi:10.1186/s12889-016-3364-7

How the All of Us Genomic Data are Organized. (2024). Retrieved from https://support.researchallofus.org/hc/en-us/articles/4614687617556-How-the-All-of-Us-Genomic-data-are-organized#h_01GY7QZR2QYFDKGK89TCHSJSA7

Lama, J., & Planelles, V. (2007). Host factors influencing susceptibility to HIV infection and AIDS progression. *Retrovirology, 4*, 52. doi:10.1186/1742-4690-4-52

Lewis, T. J., Herring, R. P., Chinnock, R. E., & Nelson, A. (2024). Ending the HIV Epidemic in Black America: Qualitative Insights Following COVID-19. *J Racial Ethn Health Disparities*. doi:10.1007/s40615-024-01925-1

Lorenzo-Redondo, R., Ozer, E. A., Achenbach, C. J., D'Aquila, R. T., & Hultquist, J. F. (2021). Molecular epidemiology in the HIV and SARS-CoV-2 pandemics. *Curr Opin HIV AIDS, 16*(1), 11-24. doi:10.1097/COH.0000000000000660

MacQueen, K. M., McLellan, E., Metzger, D. S., Kegeles, S., Strauss, R. P., Scotti, R., . . . Trotter, R. T., 2nd. (2001). What is community? An evidence-based definition for participatory public health. *Am J Public Health, 91*(12), 1929-1938. doi:10.2105/ajph.91.12.1929

Oppermann, M. (2004). Chemokine receptor CCR5: insights into structure, function, and regulation. *Cell Signal, 16*(11), 1201-1210. doi:10.1016/j.cellsig.2004.04.007

Rich, A., Scott, K., Johnston, C., Blackwell, E., Lachowsky, N., Cui, Z., . . . Roth, E. (2017). Sexual HIV risk among gay, bisexual and queer transgender men: findings from interviews in Vancouver, Canada. *Cult Health Sex, 19*(11), 1197-1209. doi:10.1080/13691058.2017.1299882

Santos, I. M., da Rosa, E. A., Graf, T., Ferreira, L. G., Petry, A., Cavalheiro, F., . . . Pinto, A. R. (2015). Analysis of Immunological, Viral, Genetic, and Environmental Factors That Might Be Associated with Decreased Susceptibility to HIV Infection in Serodiscordant Couples in Florianopolis, Southern Brazil. *AIDS Res Hum Retroviruses, 31*(11), 1116-1125. doi:10.1089/aid.2015.0168