The Association of *VMAT2* Gene Polymorphisms With the Development of Schizophrenia

by

Alexis Cenname

B.S., University of Pittsburgh, 2020

Submitted to the Graduate Faculty of

the School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Sciences

University of Pittsburgh

2023

UNIVERSITY OF PITTSBURGH

SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Alexis Cenname

It was defended on

April 24th 2023

and approved by

Jenna Carlson, PhD, Department of Biostatistics

Jeanine Buchanich, PhD, Department of Biostatistics

John Shaffer, PhD, Department of Human Genetics

Thesis Advisor: Jenna Carlson, PhD, Department of Biostatistics

Copyright © by Alexis Cenname
2023

The Association of *VMAT2* Gene Polymorphisms With the Development of Schizophrenia

Alexis Cenname, M.S.

University of Pittsburgh, 2023

Objective: To investigate the association between common single nucleotide polymorphisms (SNPs) in the vesicular monoamine transporter type 2 (*VMAT2*) gene and a diagnosis of schizophrenia in the US population.

Methods: 968 individuals with a diagnosis of schizophrenia were ancestry-matched to healthy controls to create a final cohort of 1,936 individuals. Diagnosis criteria was determined from Electronic Health Records provided by the AllofUs Research Program. Additive and dominant logistic regression analyses were done for the promoter SNPs rs363324 and rs363371. Additional stratified analyses were performed using sex and ancestry variables.

Results: No significant results were observed for rs363324 and rs363371 using the additive model. The dominant model suggested a protective effect for rs363324 for the entire cohort ($p = 0.007$; $OR = 0.75[0.61, 0.92]$) and within the African ancestral group ($p = 0.028$; $OR = 0.76[0.56, 0.97]$). Stratification by sex did not give significant results for either genotype model, nor did stratification by ancestry+sex.

Conclusion: Premature mortality rates and the need for targeted treatments in schizophrenia make it an important disease to study for public health. This study found that 'GA' and 'AA' genotypes have an equal protective effect from the development of schizophrenia for rs363324. Unlike previous studies, the results for rs363371 were not significant. Future studies should use a larger sample size and include variables concerning environmental factors.

Table of Contents

List of Tables

List of Figures

Preface

I would like to thank my thesis advisor, Dr. Jenna Carlson, for supporting me throughout my graduate career. Dr. Carlson has been an outstanding advisor and a great help throughout this process.

I would also like to thank my additional committee members, Dr. Jeanine Buchanich and Dr. John Shaffer. I appreciate you taking the time to give me suggestions and listen to my defense.

"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

Schizophrenia (SCZ) is a highly complex mental disorder that affects a relatively low percentage of the population, yet it remains one of the top 15 causes of disability in the world [1]. A wide range of psychological and behavioral symptoms are associated with the disorder, like hallucinations, loss of motivation, and disorganized thought processes. If left untreated, these symptoms usually worsen. High suicide rates in the population, along with under-diagnosis and under-treatment of comorbidities, account for an estimated 28.5 years of life lost to the disease [2]. To improve premature mortality rates and overall burden on public health, it is critical to identify key risk factors involved in its development for early diagnosis and targeted treatment. As genetic data is becoming more accessible, many genome-wide association studies (GWAS) have been conducted to understand the biological processes behind SCZ. This data suggests it to be highly heritable and polygenic, warranting further analysis of its genetic underpinnings [3].

A recent review cross-examined significant single nucleotide polymorphisms (SNPs) from previous GWAS analyses and differentially expressed (DE) genes in schizophrenic patients to determine gene-function correlation. *SLC18A2* (or *VMAT2*) was one of the nine genes with significant loci and gene expression, indicating risk alleles within this gene may influence regulation factors [4]. The *VMAT2* gene encodes for a protein contained inside the synaptic membrane of neurons, which transports amine neurotransmitters into synaptic vesicles for eventual release into the body [5]. A visual representation of this process is in Figure 1. Since VMAT2 is the only transporter that delivers cytoplasmic dopamine to central nervous system (CNS) vesicles, it is thought to play an important role in dopamine regulation [7]. There are several functions that dopamine performs in the brain, including memory, movement, motivation, and pleasure. Disruption of the *VMAT2* gene in this biological structure can result in either overproduction or underproduction of dopamine, interfering with the signaling of dopaminergic neurons [8]. An imbalance in the dopaminergic system may lead to a variety of mental health disorders, including clinical symptoms of schizophrenia [9]. To treat these symptoms, VMAT2 inhibitors have been used to maintain therapeutic efficacy and reduce

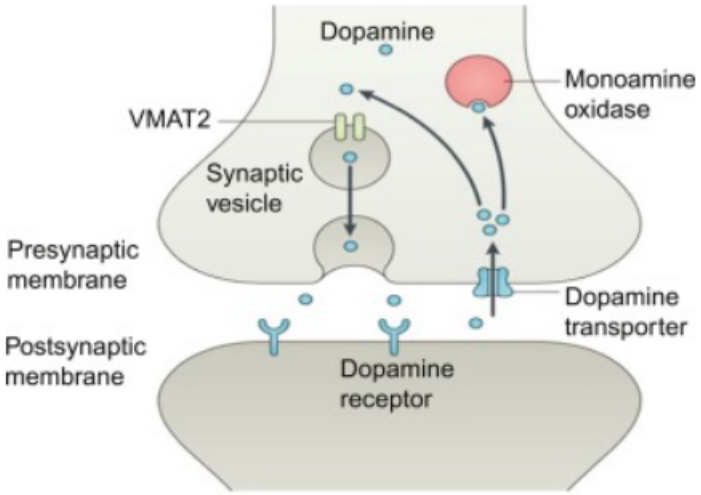

Figure 1: A diagram representing the function of the VMAT2 protein as a transporter for dopamine within the neuron. Adapted from Jankovic et al. 2017 [6].

side effects when combined with a reduced dose of anti-psychotic drugs [10].

Despite extensive research on the *VMAT2* gene, little SNP analysis has been done with it in regards to schizophrenia. Researchers from an Italian journal examined eight variants in *VMAT2* possibly relating to Parkinson's disease (PD); only 2 SNPs in the promoter region were significant [11]. Since schizophrenia and PD share genetic risk factors, rs363324 and rs263371 were subsequently studied in the Chinese Han population to determine their relationship to schizophrenia. A χ^2 test was conducted on the age-sex matched case and control groups, along with an additional stratification by sex. The only significant finding was a protective AA genotype effect for rs363371 against SCZ in the males [12]. Other association studies conducted on the *VMAT2* gene returned mixed results, and did not focus on the promoter region, which regulates transcription factors for protein production [13,14].

The purpose of this paper was to extend upon the Chinese Han analysis using a more diverse population and a larger sample size. To do this, information from the AllofUs database was used, a program that aims to collect health data on diverse populations to help

researchers better diagnose and treat diseases. To find an association between schizophrenia and regulatory SNPs, a logistic regression using the additive and dominant genetic models was performed. Ancestry was stratified in addition to sex. Unlike the previous analysis, the cases were matched to controls based on their ancestry, rather than age and sex. Section 2 gives an overview of these statistical methods and describes the data extraction and cleaning process. Section 3 provides an explanation of the results and any significant findings. The final section is a discussion of limitations and ways to extend this research.

2.0 Methods

2.1 Data Source

The All of Us Research Program is a longitudinal cohort study sponsored by the National Institutes of Health (NIH) which aims to collect data on many types of health outcomes for a broad range of participants in the United States. Currently, the AllofUs database contains electronic health records (EHRs), genomic data, physical measurements, and survey data on 372,380 individuals. Access to the data is divided into three tiers: Public, Registered, and Controlled. A training program was completed to facilitate access to the Controlled tier, which contains the individual-level genomic data necessary for completing this study. Eligibility for enrollment in the cohort requires participants to be 18 years or older and living in the United States (or U.S. territory) at time of enrollment. Anyone who is imprisoned or incapable of consent cannot enroll [15]. All statistical analyses for these data were done using Python in the Researcher Workbench—a cloud platform that stores health data collected by AllofUs.

2.2 Preprocessing Steps

2.2.1 Phenotype Data

Because of the vast amount of phenotypes available, the Researcher Workbench provides a tool called the CohortBuilder. This tool was used to exclude participants who did not match the intended criteria for the case and control groups. For the entire cohort, participants must have provided their Whole Genome Sequence (WGS) and sex at birth assignment. This left about 99,000 participants. An additional criterion was set to establish the case group, which only included the 1,008 individuals diagnosed with schizophrenia (SNOMED-58214004). The control group had slightly more restrictions, with an age cutoff

of 30 years at enrollment and exclusion of all participants with a mental disorder (SNOMED-74732009), including schizophrenia diagnoses. Following these adjustments, approximately 55,000 controls remained for selection.

Having selected the inclusion criteria, phenotypic data from the case and control groups were loaded into Python for further analysis. Additional derived data elements provided by AllofUs were also imported, as they contained relevant information about relatedness and ancestry for participants. Since heritability violates the independence assumption of the logistic regression, 2,306 people with a kinship score greater than 0.1 were removed from the sample pool.

2.2.2 Principal Components of Ancestry

The Principal Components Analysis (PCA) for genetic ancestry was done prior to this study by AllOfUs. PCA is a dimensionality reduction method that reduces the number of variables in a large matrix, $\boldsymbol{C}$. When applied to genetics, the matrix dimensions nxm represent n individual samples and m high-quality variants. Samples can be represented as points in an m-dimensional space, where each SNP is its own axis. The principal component measure is a linear combination of these axes, with the first principal component representing the largest possible variance, the second principal component representing the next largest possible variance orthogonal to the first, and so on [16]. The $\boldsymbol{C}$ matrix can be standardized using the Hardy-Weinberg equilibrium model, which normalizes genotype variances to $1/m$. This method was employed by for this data by using the *hwe_normalized_pca* package in Hail [17]. This creates a new matrix, $\boldsymbol{M}$, whose entries can be calculated by:

$$M_{ij} = \frac{C_{ij} - 2p_j}{\sqrt{2p_j(1-p_j)(m)}} \tag{1}$$

The value C_{ij} is the number of alternate alleles (0, 1, 2) per variant j carried by sample i. Half of the mean alternate allele frequency for each variant is represented by p_j. Once these entries are calculated, the matrix $\boldsymbol{M}$ is used to compute principal component measures using singular-value decomposition (SVD):

$$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T \tag{2}$$

$$\boldsymbol{P} = \boldsymbol{X}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{S} \tag{3}$$

Equation (2) is a breakdown of the standardized matrix into three separate matrices. The columns of the nxk matrix $\boldsymbol{U}$ represent the eigenvectors of $\boldsymbol{M}\boldsymbol{M}^{\boldsymbol{T}}$. Similarly, $\boldsymbol{V}$ is a mxk matrix of eigenvectors for $\boldsymbol{M}^{\boldsymbol{T}}\boldsymbol{M}$. The $\boldsymbol{S}$ diagonal matrix with dimensions kxk represents the square root of the eigenvalues for $\boldsymbol{M}\boldsymbol{M}^{\boldsymbol{T}}$ and $\boldsymbol{M}^{\boldsymbol{T}}\boldsymbol{M}$.[18]. k is defined by the rank of the matrix; $k = 16$ for the AllofUs analysis. Once the $\boldsymbol{P}$ matrix is calculated from the SVD, the PC measures are projected into a two-dimensional space. For classification into ancestry groups, AllofUs used a random forest classifier [19].

2.2.2.1 Case-Control Matching

It is common in ancestry matching to find the control group members whose principal component measures are closest to the cases. Once this is figured out, a one-to-one match is performed. An R package called PCAmatchR is used to match controls to cases by converting PCs into Mahalanobis distance metrics, a standardized distance between samples [20,21]. The greater the distance, the less similarity and vice versa. Due to the lack of a package for PCA matching in Python, a simpler method was used for this analysis. Among the ancestral categories available to each participant were European, African, Admixed American/Latino, Middle Eastern, South Asian, and East Asian. The aggregate counts for each category in the case cohort were calculated. Controls were separated by ancestry, and the corresponding counts determined the number of individuals randomly selected from each group. A final control group was formed by combining these selections. It is important to note that any subgroups with aggregate counts less than 20 are not displayed in this paper for privacy purposes.

2.2.3 Genotype Data

Genotype data for the participants are available in an auxiliary file provided by AllofUs. The interval regions for rs363324 and rs363371 were extracted from the WGS of every sample. The Hg38 genome assembly was used as the reference sequence. In both variants, the reference allele was coded as 'G' and the alternate allele as 'A'. The control group was subjected to data quality checks for both loci due to variant inconsistencies anticipated in the schizophrenia diagnosis group. Minor Allele Frequency (MAF) and Allele Balance (AB) tests were performed using thresholds of 0.05 and 0.2, respectively. Additional Hardy-Weinberg tests were done in each ancestry category ($p < 0.01$). Once the genomic data was cleaned, it was combined with case and control demographic data to develop the logistic regression models.

2.2.4 Covariates

The covariates used in the logistic regression analysis are sex, age, and the first 3 PC measures of ancestry. Differences in sex distribution between schizophrenia and healthy groups were analyzed using a χ^2 test ($p < 0.05$). A two-tailed t-test with a significance level of 0.05 was used to determine whether the mean age distribution was similar between the groups.

2.3 Binary Logistic Regression

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, it can be described as a binary logistic regression. These models can have one or multiple predictor variables, with additional covariates for adjustment. To estimate the p-value, odds ratio, and 95% confidence interval, the Wald test statistic, z, is used.

Let p_i represent the probability of the outcome occuring ($Y_i = 1$) with predictor (X_i) and adjusting for covariates ($\boldsymbol{\beta} Z_i$). The logistic regression equation is as follows:

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

The variable β_1 is the fit effect coefficient for the predictor. This model tests the null hypothesis that $\beta_1 = 0$. This is equivalent to assuming the odds ratio is equal to 1 when comparing the event and non-event groups.

The predictor can have multiple categories (1,...,n), but here is a case of a simple binary predictor. The OR can be calculated from the logistic regression equation as follows:

$$\hat{OR} = \frac{\frac{\hat{p_i}}{1-\hat{p_i}}|(x_i = 1)}{\frac{\hat{p_i}}{1-\hat{p_i}}|(x_i = 0)} = \frac{exp(\beta_0 + \beta_1 * 1 + \boldsymbol{\beta} Z_i)}{exp(\beta_0 + \beta_1 * 0 + \boldsymbol{\beta} Z_i)} = exp(\beta_1) \tag{5}$$

Here, β_1 represents the log-odds ratio of the outcome group comparing a predictor of x=0 to x=1 after adjusting for covariates.

2.3.1 Genotype Coding

When testing for genetic associations, there are three models that can be used to make the statistical power better. These are the additive, recessive, and dominant models. The additive model assumes that the risk of disease increases linearly with the number of alternate alleles. In this case, I assume the risk of developing schizophrenia increases linearly with the number of 'A' alleles. The common way to code this model is (0,1,2). The dominant model is slightly different, and assumes the risk of developing disease in the same whether you have one or two alternate alleles. Therefore, it is coded as (0,1,1). Finally, the recessive model assumes only those with the homozygous alternate genotype are at increased risk of developing the trait/disease. This is coded as (0,0,1). In this paper, I will only look at the additive and dominant models. This follows the structure of the Brighina et al. analysis.

2.3.2 Models for rs363324 and rs363371

To test my hypothesis, a similar structural equation was used. The presence (=1) and absence (=0) of schizophrenia was the identified outcome. The number of alternate alleles for each SNP was used as the predictor. The additive genotype model was used as the predictor variable ($g_i = 0, 1, 2$) for the first set of analyses. The same equation is used for the dominant model, but with a differently coded predictor variable ($g_i = 0, 1, 1$). The adjusted covariates in the model are sex, ancestry, and age.

$$logit(p_i) = \beta_0 + \beta_1 g_i + \boldsymbol{\beta} Z_i \tag{6}$$

2.4 Stratification

To eliminate potential confounding in the analysis, a cohort can be split into groups based on phenotype. Since the cohort is diverse, it is a way to investigate patterns in subgroups. In basic terms, a logistic regression is run within each group to find significant associations with the predictor and disease. The data was split into groups depending on sex and ancestry. There was a stratification of only sex, only race, and sex plus race. The covariates being used in the stratification were omitted from the $\boldsymbol{Z}$ vector in their respective logistic regression equations.

3.0 Results

3.1 Summary Statistics

A total of 1936 participants–968 cases and controls–were used in the final analysis. The case cohort included 566 males and 402 females with an average age of 50.44 $\pm$ 13.42 years (range 19-84 years). All ancestral groups were present in the cohort with the highest counts being in the European, African, and American Admixed/Latino categories. Controls were ancestry-matched and included 421 males and 547 females. The mean age of this cohort was 53.77 $\pm$ 13.24 years (range 30-103 years). Those without a mental health diagnosis were an average of 3 years older than those diagnosed with schizophrenia; a t-test supported this as significant ($p = 4.16e-08$). This difference is likely due to the 30-year cutoff introduced to the larger AllofUs control cohort. A χ^2 test of independence was done to compare sex distribution between groups. The results were significant, indicating that male subjects are 1.83 times more likely to be in the case group than the control group ($p = 5.43e-11$). Both SNPs passed QC checks described in Methods section.

3.2 Logistic Regressions

Individual logistic regressions were performed for each SNP using the additive and dominant genotype models. These were adjusted for sex, age, and the first three principal components of ancestry. The results of these regressions are in Table 2. For the additive model, no associations were found for rs363324 ($p = 0.056$; $OR = 0.87[0.76, 1.00]$) and rs363371 ($p = 0.988$; $OR = 1.00[0.75, 1.19]$). The dominant model showed significant associations for rs363324 ($p = 0.007$; $OR = 0.75[0.61, 0.92]$), but not rs363371 ($p = 0.707$; $OR = 0.96[0.78, 1.18]$). This suggests the odds of schizophrenia development in the heterozygous and homozygous 'A' allele groups is 0.75 times lower than that in the homozygous 'G' allele group.

Demographics of Cohort

Factor	Case (n=968)	Controls (n=968)	Outcome
Sex, n (%)			
Male	566 (58.47)	421 (45.45)	5.43e-11; 1.83 [1.52, 2.20] **
Female	402 (41.53)	547 (54.55)	
Age, yrs			
Mean±SD	50.44±13.42	53.77±13.24	4.16e-08; -3.34 [-4.52, -2.15] **
rs363371, n (%)			
GG	694 (71.69)	689 (71.18)	
GA	236 (24.38)	250 (25.83)	
AA	38 (3.93)	29 (2.99)	
rs363324, n (%)			
GG	342 (35.33)	292 (30.17)	
GA	387 (39.98)	429 (44.32)	
AA	239 (24.69)	247 (25.51)	

Table 1: This table gives the count and frequency of sex and genotypes for case/control groups. Mean and standard deviation were calculated for age in cohorts. Comparisons were performed using the χ^2 test and t-test for sex and age, respectively. χ^2: p-value and OR [CI]. T-test: p-value and mean difference [CI]. ** represents an outcome where $p \leq 0.05$.

3.3 Stratified Logistic Regressions

For both SNPs, stratifications on sex, ancestry, and sex+ancestry were done. The covariates being stratified were not adjusted in the corresponding regression models. Using males and females as subgroups, the additive and dominant models returned insignificant results for rs363324 and rs363371. When stratifying by ancestry, only groups with counts greater than 20 were analyzed. Therefore, only European, African, and Latino samples were analyzed in the logistic models. The additive model did not return any significant results for any ancestral category. The dominant model returned significant results for those with African ancestry ($p = 0.028$; $OR = 0.76[0.56, 0.97]$). This suggests that the 'A' allele may

Logistic Regression Results

Factor	Additive	Dominant
rs363324	p = 0.056 ; OR = 0.87 [0.76, 1.00]	p = 0.007 ; OR = 0.75 [0.61, 0.92]**
EUR	p = 0.892 ; OR = 0.98 [0.76, 1.27]	p = 0.141 ; OR = 0.66 [0.38, 1.15]
AFR	p = 0.100 ; OR = 0.85 [0.70, 1.03]	p = 0.028 ; OR = 0.76 [0.56, 0.97]**
AMR	p = 0.365 ; OR = 0.87 [0.64, 1.18]	p = 0.621 ; OR = 0.87 [0.50, 1.51]
rs363371	p = 0.988 ; OR = 1.00 [0.75, 1.19]	p = 0.707 ; OR = 0.96 [0.78, 1.18]
EUR	p = 0.613 ; OR = 1.08 [0.81, 1.44]	p = 0.388 ; OR = 0.85 [0.59, 1.23]
AFR	p = 0.900 ; OR = 0.98 [0.75, 1.28]	p = 0.854 ; OR = 1.03 [0.76, 1.39]
AMR	p = 0.773 ; OR = 0.94 [0.61, 1.45]	p = 0.840 ; OR = 1.05 [0.64, 1.73]

Table 2: This table shows the results of the individual logistic regressions and ancestry-stratified analyses for rs363324 and rs363371. The models were adjusted for age, sex, and ancestry (first 3 PC measures). The p-value, ORs, and corresponding 95% CIs are provided. ** represents an outcome where $p \leq 0.05$.

have a protective effect on schizophrenia development for this group. Stratification analyses for sex and ancestry were not significant. These results are given in Appendix A.

4.0 Discussion

When observing the potential effect of *VMAT2* SNPs on schizophrenia diagnosis, this analysis only saw meaningful results for rs363324 in the unstratified dominant model and the dominant model for the African cohort. Stratification by sex proved to be insignificant for both variants. Both significant effects were protective, indicating lower odds for those in the 'GA' and 'AA' genotype groups. The Brighina et al. analysis regarding PD performed similar logistic regression analyses but did not find significant p-values for rs363324. They did, however, observe protective effects for the 'A' allele in the rs363371 dominant model [11]. The recessive χ^2 analysis performed on the Chinese Han population also suggested that the homozygous 'AA' genotype in rs363371 reduced the odds of schizophrenia in males [12]. Before that study, limited association analyses relating to schizophrenia development were performed in this region and returned varied results [13,14]. Though the significant variants differ, nearly all results suggest a protective effect of SNPs in this gene, which indicates that genetic variability in *VMAT2* may play a role in neurodegeneration. Therefore, further studies should examine how the gene can be a potential therapeutic target for mental health diseases, especially in certain ancestral groups.

Substance abuse is more prevalent among those with mental illnesses, so several studies have been conducted to examine how *VMAT2* affects these conditions. In mice, the chronic use of nicotine and early withdrawal from the drug showed up-regulation of the VMAT2 protein [22]. Additional studies have found an association between variants in *VMAT2* with nicotine, alcohol, and opioid dependence [23,24,25]. Unfortunately, an attempt to include survey responses on substance abuse in this cohort resulted in a substantially decreased sample size. Ideally, future research will use covariates in their analysis relating to substance abuse. A variable including the history of medication may also be helpful since up to 25% of patients receiving long-term first-generation anti-psychotic treatment are affected by Tardive Dyskinesia (TD) [26]. Similar to this study, the 'AA' genotype of rs363324 suggested a protective effect against the movement disease. Another limitation of this study is the introduction of participation bias through AllofUs data submitted voluntarily. The require-

ment for schizophrenia diagnosis in the case group may exacerbate this issue. If the disease is severe or someone is untreated, the willingness to join a research program may decrease.

The considerable strengths of this study were the use of a diverse population and a large sample size (n=1,936). Most studies on the *VMAT2* gene analyzed small samples of homogeneous European or East Asian populations. To current knowledge, this is the only study examining rs363324 and rs363371 for schizophrenia development in a cohort with diverse genetic ancestry. Even so, expanding the sample size and diversity of future investigations would be advantageous. In addition, including a recessive model and more SNPs in the area will also be beneficial.

In conclusion, the log-additive models did not suggest that rs363324 and rs363371 were associated with SCZ. However, the dominant models suggested protective effects for the 'A' allele in rs363324. These results further suggest that targeting the VMAT2 protein may be a therapeutic target for certain neurological disorders. Future studies should include environmental and additional genetic factors to understand how this genetic structure affects SCZ development.

Appendix A Additional Stratified Analyses Results

Additional Logistic Regression Results

Factor	Additive	Dominant
rs363324		
MALE	p = 0.400 ; OR = 0.92 [0.76, 1.12]	p = 0.086 ; OR = 0.77 [0.58, 1.04]
M-EUR	p = 0.783 ; OR = 1.05 [0.74, 1.49]	p = 0.314 ; OR = 0.68 [0.32, 1.44]
M-AFR	p = 0.516 ; OR = 0.91 [0.69, 1.20]	p = 0.236 ; OR = 0.81 [0.57, 1.15]
M-AMR	p = 0.392 ; OR = 0.81 [0.50, 1.31]	p = 0.466 ; OR = 0.73 [0.31, 1.70]
FEMALE	p = 0.111 ; OR = 0.85 [0.70, 1.04]	p = 0.053 ; OR = 0.73 [0.54, 1.00]
F-EUR	p = 0.619 ; OR = 0.91 [0.62, 1.33]	p = 0.259 ; OR = 0.62 [0.27, 1.43]
F-AFR	p = 0.101 ; OR = 0.80 [0.61, 1.05]	p = 0.052 ; OR = 0.70 [0.49, 1.00]
F-AMR	p = 0.628 ; OR = 0.90 [0.60, 1.36]	p = 0.966 ; OR = 0.98 [0.46, 2.10]
rs363371		
MALE	p = 0.135 ; OR = 0.83 [0.65, 1.06]	p = 0.061 ; OR = 0.76 [0.57, 1.01]
M-EUR	p = 0.553 ; OR = 0.89 [0.60, 1.31]	p = 0.160 ; OR = 0.70 [0.43, 1.15]
M-AFR	p = 0.165 ; OR = 0.77 [0.53, 1.12]	p = 0.185 ; OR = 0.75 [0.49, 1.15]
M-AMR	p = 0.487 ; OR = 0.80 [0.42, 1.51]	p = 0.774 ; OR = 0.90 [0.43, 1.89]
FEMALE	p = 0.140 ; OR = 1.21 [0.94, 1.56]	p = 0.216 ; OR = 1.21 [0.90, 1.62]
F-EUR	p = 0.154 ; OR = 1.37 [0.89, 2.11]	p = 0.807 ; OR = 1.01 [0.63, 1.80]
F-AFR	p = 0.232 ; OR = 1.26 [0.86, 1.84]	p = 0.116 ; OR = 1.41 [0.92, 2.15]
F-AMR	p = 0.763 ; OR = 1.10 [0.60, 2.00]	p = 0.594 ; OR = 1.20 [0.61, 2.38]

Table 3: This table shows the results of the logistic regressions when grouping by sex and sex+ancestry for rs363324 and rs363371. The models were adjusted for age and/or ancestry (first 3 PC measures). The p-value, ORs, and corresponding 95% CIs are provided.

Appendix B Python Code

Due to privacy concerns with the AllofUs Research Program, the code presented below is just a framework. The real variables are not used in certain situations.

```
# import packages
import os
import pandas as pd
import matplotlib.pyplot as plt
from IPython.display import HTML, display
import hail as hl
import plotly.graph_objects as go
from scipy.stats import ttest_ind

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

# selecting controls randomly
# this is repeated for each ancestral group
ht.filter(df=="ancestry_group")
ht.sample(anc_prop)

# combining random selections
matched_pheno = afr_dataset.union(amr_dataset,
                                  eas_dataset,
                                  eur_dataset,
```

```
                                    mid_dataset,
                                    sas_dataset)

# combine geno and pheno data for control group
# this is also done for case group and entire cohort
SNP_control = SNP.semi_join_cols(matched_pheno)
SNP_control = SNP_control.annotate_cols(
              pheno = matched_pheno[SNP_control.sample_id])

# QC checks
# AB
SNP_control = SNP_control.filter_rows(
              hl.is_missing(SNP_control.filters))
SNP_control = hl.variant_qc(SNP_control)

# MAF
SNP_control = SNP_control.filter_rows(hl.min(
              SNP_control.variant_qc.AF) > 0.05, keep = True)

# chi square test for sex distribution
hl.eval(hl.fisher_exact_test(c1, c2, c3, c4))

# t test for age distribution
ttest_ind(case_age, control_age)

# HWE
hl.eval(hl.hardy_weinberg_test(hom_ref, het, hom_alt))

# coding for dominant model; the additive model is the default
cohort_df = cohort_df.annotate_entries(
```

```
    alt_dom = cohort_df.GT.is_het()
            + cohort_df.GT.is_hom_var())

# additive logistic regression
covariates = [1.0, cohort_df.pheno.age,
              cohort_df.pheno.sex_male,
              cohort_df.pheno.anc.pca_features[0],
              cohort_df.pheno.anc.pca_features[1],
              cohort_df.pheno.anc.pca_features[2]]

log_reg = hl.logistic_regression_rows(
    test='wald',
    y=cohort_df.pheno.has_schiz,
    x=cohort_df.GT.n_alt_alleles(),
    covariates=covariates
)
# dominant logistic regression
dom_log_reg = hl.logistic_regression_rows(
    test='wald',
    y=cohort_df.pheno.has_schiz,
    x=cohort_df.alt_dom,
    covariates=covariates
)

# stratifications done using the following code
stratified_group = cohort_df.filter_cols()

# the logistic regression framework is similar for stratified analyses,
# but omits the covariate that is being stratified from vector
```

Bibliography

[1] Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet*. 2017 Sep 16;390(10100):1211-1259.

[2] Olfson M, Gerhard T, Huang C, Crystal S, Stroup TS. Premature Mortality Among Adults With Schizophrenia in the United States. *JAMA Psychiatry*. 2015 Dec;72(12):1172-81.

[3] Dennison, C. A., Legge, S. E., Pardinas, A. F., & Walters, J. T. (2020). Genome-wide association studies in schizophrenia: Recent advances, challenges and future perspective. *Schizophrenia Research*, 217, 4-12.

[4] Chu, T. T., & Liu, Y. (2010). An integrated genomic analysis of gene-function correlation on schizophrenia susceptibility genes. *Journal of human genetics*, 55(5), 285-292.

[5] National Center for Biotechnology Information (NCBI). (2023). SLC18A2 solute carrier family 18 member A2 [Homo sapiens (human)]. `https://www.ncbi.nlm.nih.gov/gene/6571`

[6] Jankovic, J. (2017). Progress in Parkinson disease and other movement disorders. *Nature Reviews Neurology*, 13(2), 76-78.

[7] Yelin, R., & Schuldiner, S. (2002). Vesicular neurotransmitter transporters: pharmacology, biochemistry, and molecular analysis. *Neurotransmitter transporters: structure, function, and regulation*, 313-354.

[8] Zucker, M., Valevski, A., Weizman, A., & Rehavi, M. (2002). Increased platelet vesicular monoamine transporter density in adult schizophrenia patients. *European neuropsychopharmacology*, 12(4), 343-347.

[9] da Silva Alves, F., Figee, M., van Amelsvoort, T., Veltman, D., & de Haan, L. (2008). The revised dopamine hypothesis of schizophrenia: evidence from pharmacological MRI studies with atypical antipsychotic medication. *Psychopharmacol Bull*, 41(1), 121-132.

[10] Hoare, S. R., Kudwa, A. E., Luo, R., & Grigoriadis, D. E. (2022). Efficacy of vesicular monoamine transporter 2 inhibition and synergy with antipsychotics in animal models of schizophrenia. *Journal of Pharmacology and Experimental Therapeutics*, 381(2), 79-95.

[11] Brighina, L., Riva, C., Bertola, F., Saracchi, E., Fermi, S., Goldwurm, S., & Ferrarese, C. (2013). Analysis of vesicular monoamine transporter 2 polymorphisms in Parkinson's disease. *Neurobiology of aging*, 34(6), 1712.e9–1712.e1.712E13. https://doi.org/10.1016/j.neurobiolaging.2012.12.020

[12] Han, H., Xia, X., Zheng, H., Zhao, C., Xu, Y., Tao, J., & Wang, X. (2020). The Gene Polymorphism of VMAT2 Is Associated with Risk of Schizophrenia in Male Han Chinese. *Psychiatry investigation*, 17(11), 1073–1078. https://doi.org/10.30773/pi.2020.0023

[13] Talkowski, M. E., Kirov, G., Bamne, M., Georgieva, L., Torres, G., Mansour, H., Chowdari, K. V., Milanova, V., Wood, J., McClain, L., Prasad, K., Shirts, B., Zhang, J., O'Donovan, M. C., Owen, M. J., Devlin, B., & Nimgaonkar, V. L. (2008). A network of dopaminergic gene variations implicated as risk factors for schizophrenia. *Human molecular genetics*, 17(5), 747–758. https://doi.org/10.1093/hmg/ddm347

[14] Kunugi, H., Ishida, S., Akahane, A. et al. Exon/intron boundaries, novel polymorphisms, and association analysis with schizophrenia of the human synaptic vesicle monoamine transporter (SVMT) gene. *Mol Psychiatry* 6, 456–460 (2001). https://doi.org/10.1038/sj.mp.4000895

[15] National Institutes of Health (2022). AllofUs Research Program Operational Protocol. https://allofus.nih.gov/sites/default/files/All%20of%20Us%20Research%20Program%20Operational%20Protocol%202022_0.pdf

[16] McVean G. (2009). A genealogical interpretation of principal components analysis. *PLoS genetics*, 5(10), e1000686. https://doi.org/10.1371/journal.pgen.1000686

[17] Hail Team (2023). Genetics - Hail 0.2. https://hail.is/docs/0.2/methods/genetics.html#hail.methods.hwe_normalized_pca

[18] Abraham G, Inouye M (2014) Fast Principal Component Analysis of Large-Scale Genome-Wide Data. *PLOS ONE* 9(4): e93766. https://doi.org/10.1371/journal.pone.0093766

[19] AllofUs Research Program (2022). Genomic Research Data Quality Report. https://www.researchallofus.org/wp-content/themes/research-hubwordpresstheme/media/2022/06/All%20Of%20Us%20Q2%202022%20Release%20Genomic%20Quality%20Report.pdf

[20] Derek W Brown, Timothy A Myers, Mitchell J Machiela, PCAmatchR: a flexible R package for optimal case–control matching using weighted principal components, *Bioinformatics*, Volume 37, Issue 8, 15 April 2021, Pages 1178–1181, https://doi.org/10.1093/bioinformatics/btaa784

[21] Hu H. et al. (2013) Fault diagnosis of analogue circuits with weighted Mahalanobis distance based on entropy theory. *Int. J. Digit. Content Technol. Appl.*, 7, 182.

[22] Duchemin, A. M., Zhang, H., Neff, N. H., & Hadjiconstantinou, M. (2009). Increased expression of VMAT2 in dopaminergic neurons during nicotine withdrawal. *Neuroscience letters*, 467(2), 182-186.

[23] Quik, M., O'Neill, M., & Perez, X. A. (2007). Nicotine neuroprotection against nigrostriatal damage: importance of the animal model. *Trends in pharmacological sciences*, 28(5), 229-235.

[24] Schwab, S. G., Franke, P. E., Hoefgen, B., Guttenthaler, V., Lichtermann, D., Trixler, M., ... & Wildenauer, D. B. (2005). Association of DNA polymorphisms in the synaptic vesicular amine transporter gene (SLC18A2) with alcohol and nicotine dependence. *Neuropsychopharmacology*, 30(12), 2263-2268.

[25] Randesi, M., van den Brink, W., Levran, O., Blanken, P., van Ree, J. M., Ott, J., & Kreek, M. J. (2019). VMAT2 gene (SLC18A2) variants associated with a greater risk for developing opioid dependence. *Pharmacogenomics*, 20(05), 331-341.

[26] Zai CC, Tiwari AK, Mazzoco M, de Luca V, Muller DJ, Shaikh SA, et al. Association study of the vesicular monoamine transporter gene SLC18A2 with tardive dyskinesia. *J Psychiatr Res* 2013;47:1760-1765.