SYSTEMS GENETICS OF HUMAN COMPLEX DISEASES - VOLUME II

EDITED BY: Guiyou Liu, Qinghua Jiang and Liangcai Zhang
PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88976-828-8
DOI 10.3389/978-2-88976-828-8

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

SYSTEMS GENETICS OF HUMAN COMPLEX DISEASES - VOLUME II

Topic Editors:
Guiyou Liu, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences (CAS), China
Qinghua Jiang, Harbin Institute of Technology, China
Liangcai Zhang, Janssen Research and Development (United States), United States

Citation: Liu, G., Jiang, Q., Zhang, L., eds. (2022). Systems Genetics of Human Complex Diseases - Volume II. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88976-828-8

Table of Contents

ORIGINAL RESEARCH
published: 09 August 2021
doi: 10.3389/fgene.2021.716364

Evaluating the Causal Association Between Educational Attainment and Asthma Using a Mendelian Randomization Design

*Yunxia Li[1]†, Wenhao Chen[1]†, Shiyao Tian[1], Shuyue Xia[1] and Biao Yang[2]**

[1] Department of Respiratory and Critical Care Medicine, Affiliated Central Hospital, Shenyang Medical College, Shenyang, China, [2] Department of Pathogen Biology, Shenyang Medical College, Shenyang, China

Asthma is a common chronic respiratory disease. In the past 10 years, genome-wide association study (GWAS) has been widely used to identify the common asthma genetic variants. Importantly, these publicly available asthma GWAS datasets provide important data support to investigate the causal association of kinds of risk factors with asthma by a Mendelian randomization (MR) design. It is known that socioeconomic status is associated with asthma. However, it remains unclear about the causal association between socioeconomic status and asthma. Here, we selected 162 independent educational attainment genetic variants as the potential instruments to evaluate the causal association between educational attainment and asthma using large-scale GWAS datasets of educational attainment (n = 405,072) and asthma (n = 30,810). We conducted a pleiotropy analysis using the MR-Egger intercept test and the MR pleiotropy residual sum and outlier (MR-PRESSO) test. We performed an MR analysis using inverse-variance weighted, weighted median, MR-Egger, and MR-PRESSO. The main analysis method inverse-variance weighted indicated that each 1 standard deviation increase in educational attainment (3.6 years) could reduce 35% asthma risk [odds ratio (OR) = 0.65, 95% confidence interval (CI) 0.51–0.85, P = 0.001]. Importantly, evidence from other MR methods further supported this finding, including weighted median (OR = 0.55, 95% CI 0.38–0.80, P = 0.001), MR-Egger (OR = 0.48, 95% CI 0.16–1.46, P = 0.198), and MR-PRESSO (OR = 0.65, 95% CI 0.51–0.85, P = 0.0015). Meanwhile, we provide evidence to support that educational attainment protects against asthma risk dependently on cognitive performance using multivariable MR analysis. In summary, we highlight the protective role of educational attainment against asthma. Our findings may have public health applications and deserve further investigation.

Keywords: asthma, educational attainment, genome-wide association study, Mendelian randomization, inverse-variance weighted

Edited by:
Guiyou Liu,
Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, China

Reviewed by:
Yang Hu,
Harbin Institute of Technology, China
Keshen Li,
The First Affiliated Hospital of Jinan University, China

***Correspondence:**
Biao Yang
yaobiaoyb@126.com

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

***Received:** 28 May 2021*
***Accepted:** 12 July 2021*
***Published:** 09 August 2021*

Citation:
Li Y, Chen W, Tian S, Xia S and Yang B (2021) Evaluating the Causal Association Between Educational Attainment and Asthma Using a Mendelian Randomization Design. Front. Genet. 12:716364. doi: 10.3389/fgene.2021.716364

INTRODUCTION

Asthma is a common chronic respiratory disease (Beasley et al., 2015; Han et al., 2020; von Mutius and Smits, 2020). It is estimated that asthma could affect over 300 million people in the world and result in a substantial burden (Beasley et al., 2015; Han et al., 2020; von Mutius and Smits, 2020). During the past 30 years, asthma death rates have decreased greatly (Beasley et al., 2015;

von Mutius and Smits, 2020). However, there are still no effective therapeutic regimens (Beasley et al., 2015; von Mutius and Smits, 2020). Hence, it is important to identify the risk factors for asthma, especially those with the causation association for asthma (Beasley et al., 2015; von Mutius and Smits, 2020).

In the past 10 years, genome-wide association study (GWAS) has been widely used to identify the common asthma genetic variants (Demenais et al., 2018; Zhu et al., 2018; Shrine et al., 2019; Han et al., 2020). In 2018, the Trans-National Asthma Genetic Consortium (TAGC) conducted a GWAS analysis of asthma using 23,948 cases and 118,538 controls from kinds of populations, including European, African, Japanese, and Latino ancestries (von Mutius and Smits, 2020). They successfully found five new asthma loci (Demenais et al., 2018). Zhu et al. (2018) conducted a genome-wide cross-trait analysis of asthma and allergic diseases using large-scale GWAS datasets from the UK Biobank, including 33,593 cases and 76,768 controls of European ancestry. They found a significant genetic correlation between asthma and allergic diseases and highlighted 38 shared loci (Zhu et al., 2018). Shrine et al. (2019) carried out a GWAS analysis to identify common genetic variants associated with moderate-to-severe asthma by a two-stage design, including 5,135 asthma cases and 25,675 controls in stage 1 and 5,414 asthma cases and 21,471 controls in stage 2. Importantly, all these selected individuals are of European ancestry (Shrine et al., 2019). Interestingly, they reported 24 novel genetic variants to be significantly associated with moderate-to-severe asthma (Shrine et al., 2019). Han et al. (2020) conducted a GWAS analysis of asthma using 64,538 asthma cases and 329,321 controls from the UK Biobank. They further performed an asthma GWAS meta-analysis of the UK Biobank and the TAGC (Demenais et al., 2018; Han et al., 2020). Finally, Han et al. identified 66 novel asthma loci (Demenais et al., 2018; von Mutius and Smits, 2020).

Importantly, these publicly available asthma GWAS datasets provide important data support to investigate the causal association of kinds of risk factors with asthma by Mendelian randomization (MR) design or polygenic score (Granell et al., 2014; Minelli et al., 2018; Skaaby et al., 2018; Rosa et al., 2019; Xu et al., 2019; Zhao and Schooling, 2019; Chen et al., 2020; Mulugeta et al., 2020; Shen et al., 2020; Sun et al., 2020; Au Yeung et al., 2021a; Park et al., 2021; Raita et al., 2021). Some risk factors have been reported to increase the risk of asthma, including soluble interleukin-6 receptor level (Rosa et al., 2019; Raita et al., 2021), childhood body mass index (BMI) (Au Yeung et al., 2021a), adult BMI (Granell et al., 2014; Skaaby et al., 2018; Xu et al., 2019; Sun et al., 2020; Au Yeung et al., 2021a), major depressive disorder (Mulugeta et al., 2020), early pubertal maturation (Chen et al., 2020), and age at puberty (Minelli et al., 2018). Meanwhile, other risk factors are associated with reduced risk of asthma, including estimated glomerular filtration rate (Park et al., 2021), lifetime smoking (Shen et al., 2020), and linoleic acid (Zhao and Schooling, 2019).

In addition to these risk factors discussed earlier, socioeconomic status is also associated with asthma (Eagan et al., 2004; Hancox et al., 2004; Kozyrskyj et al., 2010; Brite et al., 2020). However, it remains unclear about the causal association between socioeconomic status and asthma (Eagan et al., 2004; Hancox et al., 2004; Kozyrskyj et al., 2010; Brite et al., 2020). Here, we selected 162 independent educational attainment genetic variants as the potential instruments to evaluate the causal association between educational attainment and asthma.

MATERIALS AND METHODS

Educational Attainment Genome-Wide Association Study Dataset

We selected 162 independent genetic variants that influence educational attainment to be the potential instrumental variables (Okbay et al., 2016). In brief, these genetic variants are identified by a recent large-scale GWAS dataset of educational attainment in individuals of European descent (n = 405,072) (Okbay et al., 2016). The educational attainment was a continuous variable measuring by the number of years of schooling completed (EduYears) and was assessed at age or older than 30 years (Okbay et al., 2016). This large-scale GWAS dataset is based on the meta-analysis of GWAS results from the discovery stage (Social Science Genetic Association Consortium, including 293,723 individuals) and replication stage (UK Biobank, including 111,349 individuals) (Okbay et al., 2016). The participating cohorts in the discovery stage are provided in **Table 1**. Finally, this meta-analysis identified 162 independent genetic variants with the genome-wide significance (P < 5.00E-08), as provided in **Supplementary Table 1** (Okbay et al., 2016).

Cognitive Performance Genome-Wide Association Study Dataset

We selected a large-scale GWAS dataset of cognitive performance in 257,841 individuals of European descent (Okbay et al., 2016). It is based on the sample-size-weighted meta-analysis of two large-scale GWAS datasets from the COGENT consortium (n = 35,298) and UK Biobank (n = 222,543) (Okbay et al., 2016). In COGENT, the phenotype measure was the first unrotated principal component of performance on at least three neuropsychological tests (or at least two IQ-test subscales) (Okbay et al., 2016). In the UK Biobank, the phenotype measure was a standardized score on a test of verbal–numerical reasoning (Okbay et al., 2016). More detailed information is provided in the original study (Okbay et al., 2016).

Asthma Genome-Wide Association Study Dataset

We selected the large-scale asthma GWAS dataset in 30,810 individuals of European ancestry, including 5,135 moderate–severe asthma cases and 25,675 controls, as described in the original study (Shrine et al., 2019). These selected moderate–severe asthma cases are from the Genetics of Asthma Severity and Phenotypes study (GASP, n = 1,858), the Unbiased Biomarkers in Prediction of respiratory disease outcomes project (U-BIOPRED, n = 281), and the UK Biobank (n = 2,996) (Shrine et al., 2019). The selected controls are from the U-BIOPRED (n = 75) and the UK Biobank (n = 25,600) (Shrine et al., 2019). In GASP and U-BIOPRED, moderate-to-severe asthma patients

TABLE 1 | Participating cohorts in Educational attainment GWAS discovery stage (Okbay et al., 2016).

Study	Country	Sample size	Birth year (mean/range)	Female %
ACPRC	England	1,713	1923 (1903–1948)	0.71
AGES	Iceland	3,212	1927 (1908–1936)	0.58
ALSPAC	England	2,877	1959 (1948–1963)	1
ASPS	Austria	777	1932 (1909–1949)	0.57
BASE—II	Germany	1,619	1948 (1925–1983)	0.52
CoLaus	Switzerland	3,269	1950 (1928–1970)	0.53
COPSAC2000	Germany	318	1966 (1964–1969)	0.47
CROATIA—Korčula	Croatia	842	1950 (1909–1977)	0.64
deCODE	Iceland	46,758	1945 (1894–1983)	0.57
DHS	Germany	953	1949 (1929–1974)	0.53
DIL	England	2,578	1958 (1958–1958)	0.52
EGCUT1	Estonia	5,597	1950 (1905–1980)	0.55
EGCUT2	Estonia	1,328	1957 (1911–1979)	0.53
EGCUT3	Estonia	2,047	1966 (1930–1982)	0.73
ERF	Netherlands	2,433	1952 (1914–1974)	0.55
FamHS	United States	3,483	1941 (1900–1965)	0.53
FINRISK	Finland	1,685	1946 (1923–1977)	0.46
FTC	Finland	2,418	1945 (1910–1972)	0.56
GOYA	Denmark	1,459	1947 (1944–1954)	0
GRAPHIC	England	727	1951 (1942–1965)	0.53
GS	Scotland	8,776	1955 (1909–1981)	0.59
H2000 Cases	Finland	797	1949 (1924–1970)	0.5
H2000 Controls	Finland	819	1949 (1924–1969)	0.52
HBCS	Finland	1,617	1941 (1934–1944)	0.57
HCS	Australia	1,946	1940 (1920–1951)	0.49
HNRS (CorexB)	Germany	1,401	1942 (1926–1955)	0.5
HNRS (Oexpr)	Germany	1,347	1942 (1926–1955)	0.5
HNRS (Omni1)	Germany	778	1942 (1927–1955)	0.52
HRS	United States	9,963	1940 (1900–1979)	0.42
Hypergenes	Italy/United Kingdom/Belgium	815	1945 (1914–1971)	0.46
INGI—CARL	Italy	947	1946 (1910–1975)	0.58
INGI—FVG	Italy	943	1951 (1917–1978)	0.6
KORA S3	Germany	2,655	1945 (1920–1964)	0.51
KORA S4	Germany	2,721	1949 (1926–1970)	0.51
LBC1921	Scotland	515	1921 (1921–1921)	0.58
LBC1936	Scotland	1,003	1936 (1936–1936)	0.49
LifeLines	Netherlands	12,539	1960 (1921–1980)	0.58
MCTFR	United States	3,819	1953 (1926–1974)	0.54
MGS	United States	2313	1951 (1914–1976)	0.5
MoBa	Norway	622	1971 (1966–1976)	1
NBS	Netherlands	1,808	1941 (1923–1972)	0.5
NESDA	Netherlands	1,820	1958 (1939–1977)	0.64
NFBC66	Finland	5,297	1966 (1966–1966)	0.52
NTR	Netherlands	5,246	1958 (1917–1989)	0.64
OGP	Italy	370	1950 (1916–1976)	0
OGP—Talana	Italy	544	1949 (1910–1977)	0.59
ORCADES	Scotland	1,828	1952 (1914–1979)	0.6
PREVEND	Netherlands	3,578	1948 (1923–1968)	0.48
QIMR	Australia	8,006	1956 (1900–1984)	0.59
RS—I	Netherlands	6,108	1922 (1893–1938)	0.6
RS—II	Netherlands	1,667	1935 (1906–1944)	0.52
RS—III	Netherlands	3,040	1950 (1910–1960)	0.56
Rush—MAP	United States	887	1921 (1901–1948)	0.72

(Continued)

TABLE 1 | Continued

Study	Country	Sample size	Birth year (mean/range)	Female %
Rush—ROS	United States	808	1921 (1896–1946)	0.66
SardiNIA	Italy	5,616	1955 (1901–1983)	0.58
SHIP	Germany	3,556	1945 (1918–1971)	0.5
SHIP—TREND	Germany	901	1956 (1928–1980)	0.57
STR—Salty	Sweden	4,832	1951 (1943–1958)	0.52
STR—Twingene	Sweden	9,553	1941 (1916–1958)	0.53
THISEAS	Greece	829	1950 (1909–1979)	0.33
TwinsUK	England	4,012	1949 (1919–1978)	1
WTCCC58C	England	2,804	1958 (1958–1958)	0.48
YFS	Finland	2,029	1969 (1962–1977)	0.55
23andMe	Primarily US	76,155	1961 (1901–1985)	0.52

were evaluated using clinical records based on the British Thoracic Society 2014 guidelines (Shrine et al., 2019). In the UK Biobank, moderate-to-severe asthma cases were diagnosed by a doctor (Shrine et al., 2019). The key demographic characteristics, including age and sex, are provided in **Table 2** or the original study (Shrine et al., 2019).

Pleiotropy Analysis

MR is established based on three key assumptions. Assumption 1: genetic variants (instrumental variables) should be significantly associated with the exposure (educational attainment). Hence, we selected 162 independent genetic variants associated with educational attainment with the genome-wide significance ($P < 5.00E\text{-}08$), as described earlier. Both assumption 2 and assumption 3 are known as no pleiotropy, as described in recent MR studies (Larsson et al., 2020; Au Yeung et al., 2021b; Sun et al., 2021; Yuan et al., 2021; Zhao and Schooling, 2021; Zhuang et al., 2021b). Hence, we conducted a pleiotropy analysis using the MR-Egger intercept test (Bowden et al., 2015; Burgess and Thompson, 2017) and the MR pleiotropy residual sum and outlier (MR-PRESSO) test (Verbanck et al., 2018); both have widely used in recent MR studies (Larsson et al., 2020; Au Yeung et al., 2021b; Sun et al., 2021; Yuan et al., 2021; Zhao and Schooling, 2021; Zhuang et al., 2021b). The significance threshold $P < 0.05$ indicated evidence of pleiotropy.

Mendelian Randomization Analysis

For univariable MR analysis, we selected the inverse-variance weighted (IVW) as the main MR analysis method. Meanwhile, we also selected other additional MR analysis methods, including weighted median, MR-Egger method, and MR-PRESSO, as used in recent MR studies (Bowden et al., 2015; Burgess and Thompson, 2017; Liu et al., 2018; Larsson et al., 2020; Au Yeung et al., 2021b; Sun et al., 2021; Yuan et al., 2021; Zhao and Schooling, 2021; Zhuang et al., 2021b). For multivariable MR analysis, we selected the multivariable IVW method, multivariable median-based method, and multivariable MR-Egger method. The odds ratio (OR) and 95% confidence interval (CI) of asthma correspond to approximately per 3.6 years increase [approximately 1 standard deviation (SD)] in EduYears. R (version x64 4.0.3), R package "MendelianRandomization," and MR-PRESSO were used to perform the MR analysis. The significance threshold $P < 0.05$ indicated evidence of causal association. To test the influence of a single genetic variant, we also conducted a sensitivity analysis using leave-one-out permutation (Liu et al., 2018).

TABLE 2 | Baseline characteristics of asthma cases and controls (Shrine et al., 2019).

Phenotypes	Cases (n = 5,135)	Controls (n = 25,675)
Age, years	55 (12)	56 (8)
Female	3,170 (61.7%)	14,626 (57.0%)
Male	1,965 (38.3%)	11,049 (43.0%)
FEV*1*, % predicted	72.4% (21.4)	91.8% (17.4)
FEV*1*/FVC	0.67 (0.12)	0.76 (0.06)
Smoking status Ever smoker	2,265 (44.1%)	11,913 (46.4%)
Smoking status Never smoker	2,647 (51.6%)	13,487 (52.5%)
Smoking status Unknown	223 (4.3%)	275 (1.1%)
Rhinitis or eczema status Yes	1,897 (36.9%)	8†
Rhinitis or eczema status No	2,062 (40.2%)	25 667†
Rhinitis or eczema status Unknown	1,176 (22.9%)	0†
Rhinitis or eczema status Oral corticosteroid use (prednisolone)	222/3,710 (6.0%)	NA

Data are mean (SD) or n (%), unless otherwise stated.

FEV1, forced expiratory volume in 1 s; FVC, forced vital capacity; NA, not applicable; U-BIOPRED, Unbiased biomarkers in prediction of respiratory disease outcomes.

†Patients in the U-BIOPRED cohort were not screened for rhinitis or eczema before sample selection but were subsequently found to comprise eight patients with rhinitis, eczema, or allergy.

Power Analysis

The variance of educational attainment (R^2) explained by the selected genetic variants was calculated using the effect allele frequency, the effect size beta (β), and the number of the selected genetic variants (k), as described in a previous study (Locke et al., 2015).

$$R^2 = \sum_{i=1}^{k} \beta_i^2 * (1 - EAF_i) * EAF_i * 2$$

Based on the R^2 and other necessary information, including sample size, type-I error rate, proportion of cases in the study, and true OR of the outcome variable per SD of the exposure variable, the statistical power was calculated using mRnd (Power calculations for MR) (Brion et al., 2013).

RESULTS

Educational Attainment Genetic Variants and Asthma

We selected 162 independent genetic variants influencing educational attainment and extracted their corresponding summary statistics in the asthma GWAS dataset. The results showed that 141 unique genetic variants were available in the asthma GWAS dataset. Only five genetic variants are associated with asthma risk with $P < 0.05$, including rs1378214 ($P = 0.000531$), rs76878669 ($P = 0.00607$), rs7772172 ($P = 0.00666$), rs113520408 ($P = 0.0183$), and rs9556958 ($P = 0.0388$). These findings indicated that all these selected genetic variants showed a more significant trend associated with educational attainment. **Table 3** provides the more detailed results about these 141 genetic variants.

Pleiotropy Analysis

Evidence from the MR-Egger intercept test supported that these 141 genetic variants showed no significant pleiotropy with intercept = 0.005, $P = 0.581$. Importantly, evidence from the MR-PRESSO global test further highlighted no significant horizontal pleiotropy $P = 0.375$. Hence, these 141 genetic variants could be selected as the effective instrumental variables.

Univariable Mendelian Randomization Analysis

The main analysis method IVW indicated that each 1 SD increase in educational attainment (3.6 years) could reduce 35% asthma risk (OR = 0.65, 95% CI 0.51–0.85, $P = 0.001$). Importantly, evidence from other MR methods further supported this finding, including weighted median (OR = 0.55, 95% CI 0.38–0.80, $P = 0.001$), MR-Egger (OR = 0.48, 95% CI 0.16–1.46, $P = 0.198$), and MR-PRESSO (OR = 0.65, 95% CI 0.51–0.85, $P = 0.0015$). **Figures 1–3** show the individual causal estimates using the IVW method, weighted median, and MR-Egger, respectively. We further conduct a sensitivity analysis using the leave-one-out permutation. The results suggested no single genetic variant to significantly affect the estimates between educational attainment and the risk of asthma.

Multivariable Mendelian Randomization Analysis

In multivariable MR analysis, we evaluated the effect of cognitive performance on the causal association between educational attainment and the risk of asthma. However, all three multivariable MR analysis methods indicated no significant causal association between educational attainment and the risk of asthma, including multivariable IVW method (OR = 0.64, 95% CI 0.35–1.17, $P = 0.144$), multivariable median-based method (OR = 0.63, 95% CI 0.28–1.42, $P = 0.265$), and multivariable MR-Egger method (OR = 0.32, 95% CI 0.08–1.22, $P = 0.094$). Hence, these findings provide evidence to support that educational attainment protects against asthma risk dependently on cognitive performance.

Power Analysis

One hundred forty-one educational attainment genetic variants finally selected in our MR analysis explain a total of 3.87% of educational attainment variance. Power analysis using mRnd showed that our MR analysis had 80% power to detect OR of 0.79 or lower per SD increase in educational attainment for the risk of asthma. Meanwhile, our MR analysis has 100% power to detect the OR of 0.65 using IVW, the OR of 0.55 using weighted median, the OR of 0.48 using MR-Egger, and the OR of 0.65 using MR-PRESSO.

DISCUSSION

Until recently, multiple large-scale GWAS analyses have been conducted to report novel asthma genetic variants (Demenais et al., 2018; Zhu et al., 2018; Shrine et al., 2019; Han et al., 2020). Importantly, these GWAS datasets are publicly available and promote additional analyses, such as MR analysis, to evaluate the causal association between common risk factors and asthma. These risk factors include soluble interleukin-6 receptor level (Rosa et al., 2019; Raita et al., 2021), childhood BMI (Au Yeung et al., 2021a), adult BMI (Granell et al., 2014; Skaaby et al., 2018; Xu et al., 2019; Sun et al., 2020; Au Yeung et al., 2021a), major depressive disorder (Mulugeta et al., 2020), early pubertal maturation (Chen et al., 2020), age at puberty (Minelli et al., 2018), estimated glomerular filtration rate (Park et al., 2021), lifetime smoking (Shen et al., 2020), and linoleic acid (Zhao and Schooling, 2019).

It is reported that socioeconomic status is also a risk factor for asthma (Eagan et al., 2004; Hancox et al., 2004; Kozyrskyj et al., 2010; Brite et al., 2020). In the World Trade Center Health Registry study, Brite et al. (2020) analyzed the data from 30,452 individuals and found that individuals with lower socioeconomic status had worse asthma outcomes. In the Western Australian Pregnancy Cohort (Raine) Study, Kozyrskyj et al. (2010) analyzed the data from 2,868 children and found that children with lower socioeconomic status tended to develop persistent asthma. However, Hancox et al. (2004) reported inconsistent findings in a prospective cohort study including approximately 1,000 individuals in New Zealand. They found no significant between socioeconomic status during childhood and the prevalence of asthma (Hancox et al., 2004). Hence, the causal association between socioeconomic status and asthma remains unclear, which further promotes us to perform an MR analysis using the large-scale GWAS datasets.

Using 162 independent educational attainment genetic variants, we successfully extracted the summary association results of 141 unique genetic variants from the asthma GWAS dataset. The pleiotropy analysis indicated these genetic variants to be effective instruments. MR analysis showed each 1 SD increase in educational attainment (4.2 years) reduced 35%

TABLE 3 | Association between 141 educational attainment genetic variants and asthma.

SNP	CHR	Position (b37)	EA	NEA	EAF	Beta	SE	*P*-value
rs56044892	1	41830086	T	C	0.198	0.00536	0.0291	0.854
rs12076635	1	44026656	C	G	0.779	−0.0126	0.0265	0.634
rs12410444	1	44188719	G	A	0.297	−0.0303	0.0241	0.207
rs142328051	1	44371441	C	T	0.142	−0.00193	0.0289	0.947
rs2568955	1	72762169	C	T	0.739	−0.0225	0.0262	0.39
rs12142680	1	73615892	A	G	0.0757	0.018	0.0461	0.696
rs12145291	1	74161795	C	T	0.0546	−0.0294	0.053	0.579
rs1008078	1	91189731	T	C	0.397	0.0316	0.0226	0.163
rs12134151	1	96202443	C	G	0.494	0.00288	0.0219	0.895
rs4378243	1	98395881	T	G	0.834	−0.04	0.0299	0.182
rs17372140	1	98572382	A	G	0.295	0.0299	0.0244	0.22
rs648163	1	199315998	T	C	0.27	−0.00598	0.0247	0.809
rs11588857	1	204587047	A	G	0.205	0.0112	0.0272	0.681
rs35771425	1	211609768	C	T	0.221	−0.0371	0.0265	0.162
rs78365243	1	211737950	C	T	0.0499	0.058	0.0515	0.26
rs2992632	1	243503764	T	A	0.287	−0.00229	0.0248	0.927
rs7590368	2	10961474	C	T	0.257	0.0258	0.0251	0.305
rs76076331	2	10977585	T	C	0.123	0.016	0.0338	0.636
rs17504614	2	51080481	C	T	0.186	−0.0113	0.0286	0.692
rs56158183	2	60632924	A	G	0.0795	0.0447	0.0407	0.272
rs7593947	2	60704933	A	T	0.529	−0.0343	0.0226	0.129
rs356992	2	60753593	G	C	0.695	−0.0127	0.024	0.598
rs268134	2	65608363	G	A	0.752	−0.00901	0.0254	0.723
rs6715849	2	100306378	G	A	0.559	0.0105	0.0224	0.638
rs4851251	2	100753490	T	C	0.274	0.0216	0.0248	0.383
rs12987662	2	100821548	A	C	0.4	−0.0294	0.0225	0.191
rs71413877	2	100924822	A	G	0.0419	−0.0846	0.0558	0.13
rs34106693	2	101151830	G	C	0.171	0.0371	0.0301	0.218
rs77702819	2	101328728	T	G	0.0926	−0.0211	0.0388	0.587
rs17824247	2	144152539	C	T	0.41	−0.0298	0.0223	0.182
rs10178115	2	155451738	G	T	0.45	0.0216	0.0223	0.332
rs10930008	2	161854736	A	G	0.738	−0.0478	0.0249	0.0549
rs16845580	2	161920884	C	T	0.37	−0.00443	0.023	0.847
rs4500960	2	162818621	T	C	0.483	−0.0128	0.022	0.56
rs1596747	2	193802478	G	A	0.493	0.000776	0.0219	0.972
rs4675248	2	202880230	G	A	0.57	−0.0423	0.0229	0.0642
rs12694681	2	226609241	G	T	0.308	0.0044	0.0238	0.853
rs11687170	2	237058144	C	T	0.168	−0.0209	0.0293	0.475
rs7429990	3	47901803	A	C	0.275	0.00642	0.0245	0.793
rs140711597	3	48469441	G	C	0.0213	−0.0101	0.0823	0.902
rs34638686	3	48682658	T	C	0.0999	0.0572	0.0371	0.124
rs3172494	3	48731487	T	G	0.108	−0.0447	0.0357	0.211
rs113011189	3	49250007	T	C	0.0896	0.0311	0.039	0.424
rs13090388	3	49391082	T	C	0.303	−0.0368	0.0239	0.123
rs11130222	3	49901060	T	A	0.42	0.0279	0.0223	0.21
rs6800916	3	50052873	A	T	0.0953	−0.0565	0.0425	0.184
rs2624818	3	50056265	A	G	0.103	0.00892	0.0363	0.806
rs112634398	3	50075494	G	A	0.0487	0.0799	0.0527	0.13
rs71326918	3	50174844	A	C	0.116	−0.0201	0.0347	0.561
rs35971989	3	51469248	G	A	0.158	−0.00101	0.0306	0.974
rs7610856	3	71579022	A	C	0.43	−0.0393	0.0224	0.08
rs62263923	3	85674790	G	A	0.362	−0.0225	0.0231	0.33

(Continued)

TABLE 3 | Continued

SNP	CHR	Position (b37)	EA	NEA	EAF	Beta	SE	*P*-value
rs56262138	3	86183716	A	T	0.299	0.00764	0.0245	0.755
rs9755467	3	127143885	T	C	0.159	−0.00189	0.0303	0.95
rs12646808	4	3249828	C	T	0.35	0.0203	0.0235	0.388
rs1967109	4	28720915	G	A	0.837	−0.000441	0.03	0.988
rs4308415	4	67821874	G	C	0.57	0.0253	0.022	0.25
rs6839705	4	106144735	C	A	0.662	0.0272	0.0232	0.241
rs4863692	4	140764124	T	G	0.325	0.011	0.0234	0.64
rs1912528	4	140945966	T	C	0.358	−0.00849	0.0229	0.71
rs12640626	4	176626272	A	G	0.568	−0.00204	0.0222	0.927
rs4493682	5	45188024	C	G	0.18	−0.0144	0.0287	0.615
rs1562242	5	57566494	C	T	0.516	0.0191	0.0221	0.388
rs61160187	5	60111579	G	A	0.403	0.0123	0.022	0.576
rs113474297	5	60554934	T	C	0.14	0.0511	0.0319	0.109
rs10223052	5	60800336	G	A	0.645	0.0181	0.0231	0.434
rs775326	5	62918416	A	C	0.322	0.0117	0.0234	0.617
rs12653396	5	87847273	A	T	0.568	−0.00266	0.0225	0.906
rs6882046	5	87968864	G	A	0.267	0.0359	0.025	0.151
rs700590	5	88106258	C	T	0.401	0.00243	0.0226	0.914
rs152603	5	106774922	G	A	0.361	0.00366	0.0229	0.873
rs660001	5	113866598	A	G	0.212	0.0201	0.0269	0.454
rs62379838	5	120102028	C	T	0.299	0.0269	0.024	0.263
rs7776010	6	14723608	C	T	0.19	−0.0418	0.0283	0.139
rs7772172	6	16662928	G	A	0.599	0.0609	0.0225	0.00666
rs6939294	6	16950631	T	C	0.228	−0.0092	0.0265	0.728
rs56231335	6	98187291	C	T	0.327	−0.0149	0.0235	0.525
rs1338554	6	98346801	G	A	0.504	−0.00034	0.0221	0.988
rs9401593	6	98549801	C	A	0.483	−0.0256	0.0222	0.25
rs56081191	6	98557732	A	G	0.0785	−0.00977	0.0418	0.815
rs11756123	6	152218079	T	A	0.633	−0.0368	0.0228	0.107
rs113779084	7	11871787	A	G	0.302	0.00464	0.0243	0.849
rs12531458	7	39090698	C	A	0.487	0.0113	0.0224	0.614
rs12702087	7	44812607	A	G	0.437	−0.00242	0.0225	0.914
rs756912	7	71741797	T	C	0.525	0.0203	0.022	0.356
rs11976020	7	72247800	A	G	0.228	0.0067	0.0261	0.798
rs12534506	7	92662327	T	A	0.539	0.0131	0.0226	0.564
rs148490894	7	99531755	G	A	0.0275	0.0856	0.0674	0.204
rs11771168	7	113904061	T	C	0.255	0.0428	0.0263	0.104
rs113520408	7	128402782	A	G	0.278	−0.0582	0.0247	0.0183
rs17167170	7	133302345	G	A	0.205	0.00347	0.0276	0.9
rs320700	7	137049477	A	G	0.641	−0.0159	0.0229	0.486
rs1106761	8	142619234	A	G	0.385	0.0206	0.023	0.371
rs11774212	8	145686505	T	C	0.515	0.00315	0.022	0.886
rs4741343	9	14075095	A	G	0.17	−0.0273	0.0293	0.352
rs4741351	9	14222782	G	A	0.705	−0.039	0.0244	0.111
rs7029201	9	23358081	A	G	0.418	0.00472	0.0224	0.833
rs7033137	9	72055158	G	C	0.26	0.0412	0.0253	0.104
rs17425572	9	88006338	G	A	0.53	−0.00799	0.0221	0.718
rs10821136	9	96238731	T	C	0.33	−0.0115	0.0234	0.625
rs10818606	9	124618386	C	T	0.588	−0.0151	0.0223	0.498
rs10761741	10	65066186	T	G	0.415	0.0302	0.0224	0.177
rs7914680	10	67965010	G	T	0.271	0.0302	0.0252	0.23
rs1925576	10	68689083	G	A	0.451	−0.0128	0.0228	0.575
rs149613931	10	103550281	T	G	0.0564	−0.0516	0.0481	0.284

(Continued)

TABLE 3 | Continued

SNP	CHR	Position (b37)	EA	NEA	EAF	Beta	SE	*P*-value
rs73344830	10	103816828	G	A	0.582	–0.0213	0.0223	0.341
rs61874768	10	103880118	T	G	0.18	–0.0165	0.0292	0.572
rs10786662	10	103989812	C	G	0.582	–0.0126	0.0223	0.573
rs12761761	10	133775375	T	C	0.253	–0.0146	0.0252	0.562
rs7945718	11	12748819	G	A	0.412	0.0346	0.0225	0.123
rs76878669	11	66092567	G	C	0.252	0.072	0.0262	0.00607
rs7948975	11	90424638	C	T	0.398	0.0278	0.0224	0.215
rs111321694	11	110950386	T	C	0.175	–0.0242	0.0292	0.408
rs79925071	11	121998253	T	G	0.571	–0.0136	0.0224	0.545
rs10772644	12	13417617	C	G	0.887	–0.00369	0.0364	0.919
rs7964899	12	14595756	A	G	0.431	–0.0359	0.0222	0.105
rs1389473	12	92154270	A	G	0.388	0.0109	0.0223	0.626
rs10773002	12	123746961	T	A	0.755	0.0173	0.0256	0.501
rs8002014	13	58358159	A	G	0.264	0.0374	0.0248	0.132
rs9556958	13	99100046	T	C	0.519	0.0465	0.0225	0.0388
rs34344888	14	23387585	G	A	0.606	–0.0238	0.0225	0.291
rs1115240	14	27090388	C	G	0.742	–0.0239	0.0251	0.341
rs10483349	14	29629456	G	A	0.184	0.0299	0.0287	0.297
rs58694847	14	84916511	C	G	0.266	0.0107	0.025	0.667
rs1378214	15	47579004	C	T	0.628	0.0795	0.0229	0.000531
rs6493271	15	47613593	C	T	0.179	–0.0475	0.0291	0.102
rs281302	15	47686662	A	G	0.546	–0.0416	0.0227	0.0665
rs12900061	15	66009248	A	G	0.172	0.00978	0.029	0.736
rs4076457	15	78007213	T	C	0.257	–0.00426	0.0253	0.866
rs28420834	15	82513121	G	A	0.573	–0.0238	0.0229	0.298
rs9914544	17	18787828	C	A	0.374	0.0426	0.0227	0.0608
rs9964724	18	35159124	T	C	0.682	–0.00238	0.0237	0.92
rs12956009	18	44768024	C	T	0.426	–0.0195	0.0224	0.384
rs62100765	18	50735418	T	C	0.403	–0.00695	0.0224	0.757
rs1382358	19	13171424	C	T	0.13	–0.0209	0.033	0.527
rs111730030	19	13268826	T	G	0.0578	–0.024	0.0476	0.615
rs12462428	19	16694610	C	T	0.193	–0.0149	0.0278	0.592
rs78387210	20	47823441	T	C	0.0904	–0.0153	0.0393	0.698
rs6065080	20	59832791	C	T	0.645	0.0244	0.023	0.289
rs35532491	22	34329603	T	A	0.101	0.0106	0.0364	0.771
rs7286601	22	51121416	G	T	0.458	–0.00681	0.0222	0.759

EA, effect allele; NEA, non-effect allele; EAF, effect allele frequency; SE, standard error; Beta is regression coefficients based on effect allele.

asthma risk (OR = 0.65, 95% CI 0.51–0.85, *P* = 0.001) using IVW. Importantly, other additional analysis methods and sensitivity methods supported this finding. However, multivariable MR analysis showed that educational attainment protected against asthma risk dependently on cognitive performance.

Until now, univariable and multivariable MR studies have evaluated the association of educational attainment and/or cognitive performance on other human complex diseases or phenotypes. Wang et al. (2021) conducted a two-sample univariable and multivariable MR to evaluate the causal effects of educational attainment and cognition on the risk of epilepsy. Using univariable MR analysis, they found that both educational attainment and cognitive performance could reduce the risk of epilepsy (Wang et al., 2021). Using multivariable MR analysis, they found that only educational attainment protected against epilepsy independent of cognitive performance (Wang et al., 2021).

Gill et al. (2019) conducted a two-sample univariable MR to evaluate the effect of education and cognitive performance, respectively, on the risk of coronary heart disease and ischemic stroke. Meanwhile, they performed a multivariable MR to adjust for the effects of cognitive performance and education, respectively (Gill et al., 2019). Using univariable MR analysis, they found a causal association between high education and reduced risk of coronary heart disease and stroke (Gill et al., 2019). Meanwhile, they also found that high cognitive performance could also reduce the risk of coronary heart disease but not stroke (Gill et al., 2019). Using multivariable MR analysis, they found that education could protect against coronary heart disease and stroke independent of cognitive function

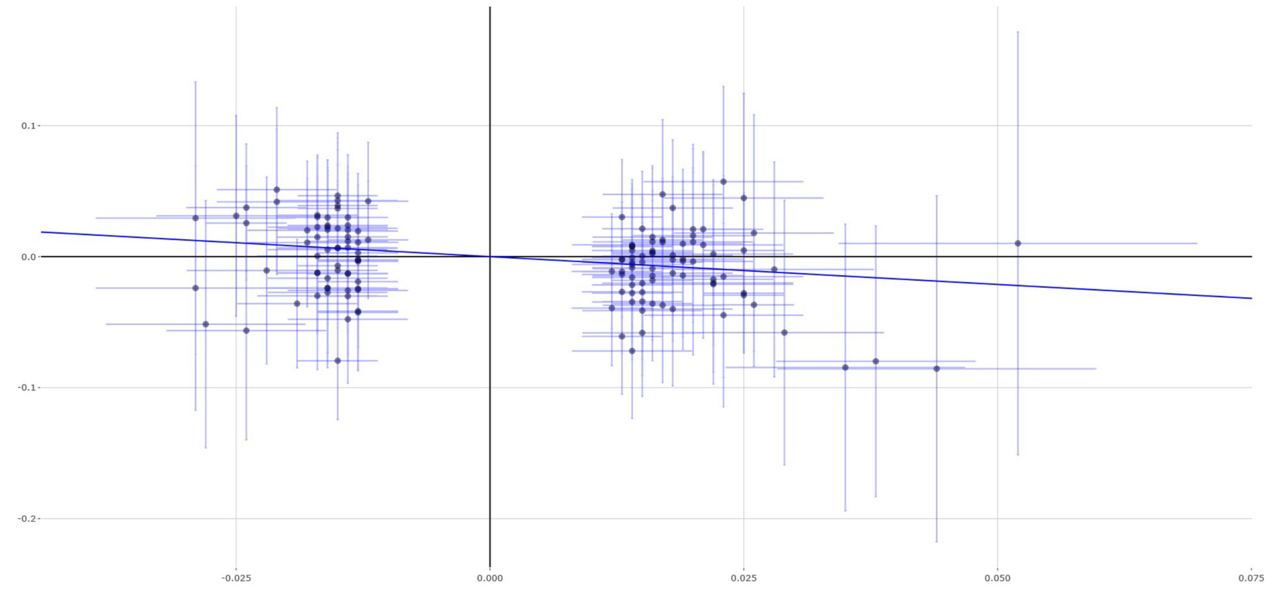


FIGURE 1 | Single estimates about causal association between educational attainment and asthma from MR analysis using IVW method. This scatter plots represent 141 genetic variants associated with educational attainment on *x*-axis and risk of asthma on *y*-axis. Continuous line represents causal effect of educational attainment on risk of asthma. IVW, inverse variance weighted.

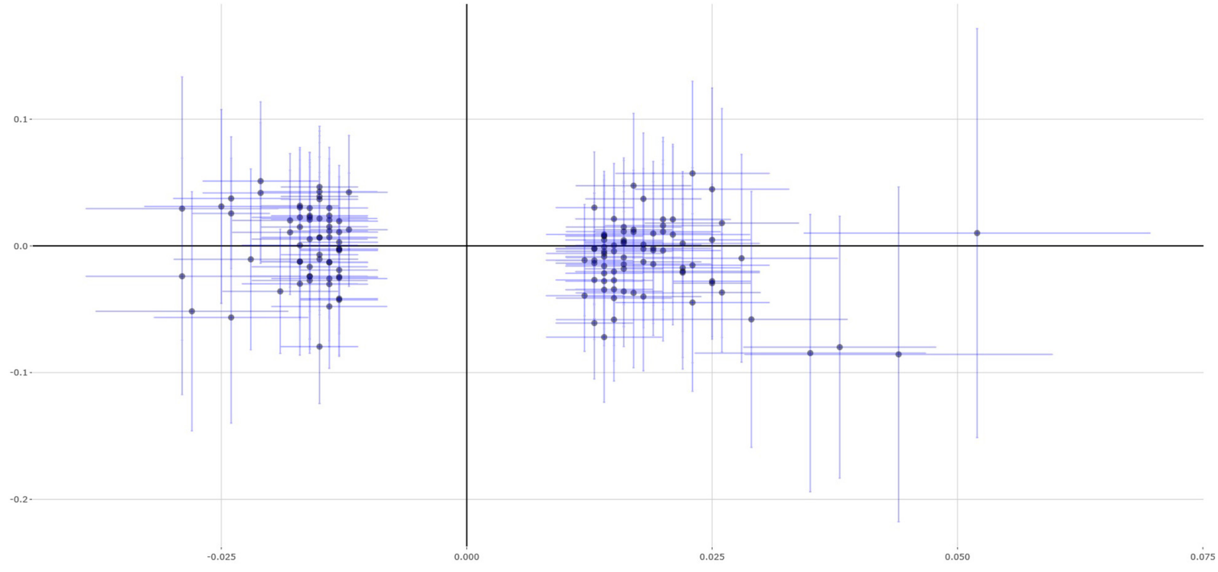


FIGURE 2 | Single estimates about causal association between educational attainment and asthma from MR analysis using weighted median method. This scatter plots represent 141 genetic variants associated with educational attainment on *x*-axis and risk of asthma on *y*-axis. Continuous line represents causal effect of educational attainment on risk of asthma.

(Gill et al., 2019). However, the cognitive performance had no causal association with coronary heart disease or stroke by adjusting for education (Gill et al., 2019). Carter et al. (2019) found that BMI, systolic blood pressure, and smoking behavior could mediate the protective role of education on the risk of cardiovascular outcomes, including coronary heart disease, stroke, myocardial infarction, and cardiovascular disease (all subtypes; all measured in OR).

Liang et al. (2021) identified that educational attainment protected against type 2 diabetes independently of cognitive performance. Rosoff et al. (2020) found that educational attainment could reduce the risk of suicide attempts in individuals with and without psychiatric disorders independent of cognition. Zhang et al. (2020) found that high educational attainment, but not cognitive performance, was causally associated with a reduced risk of amyotrophic lateral sclerosis.

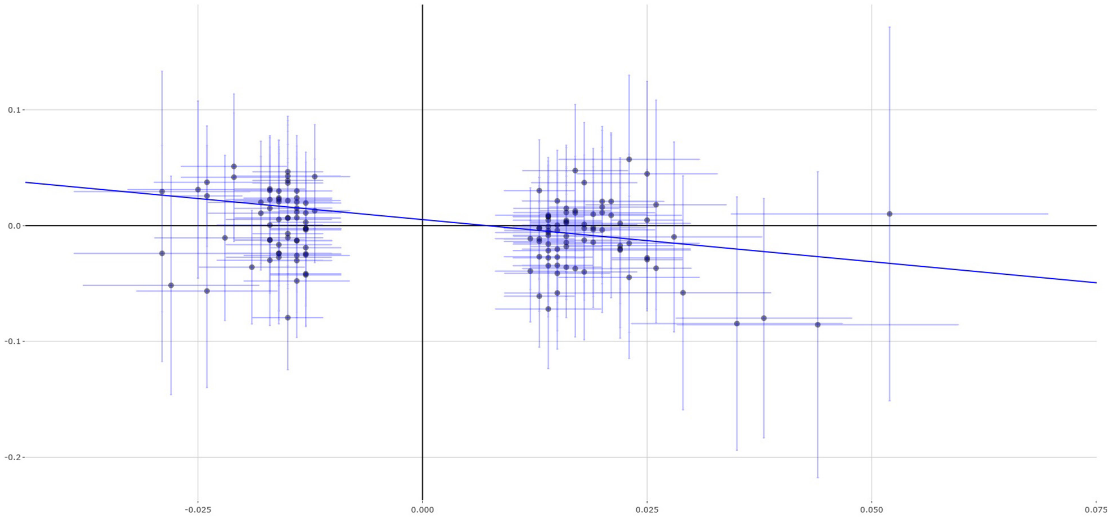


FIGURE 3 | Single estimates about causal association between educational attainment and asthma from MR analysis using MR-Egger method. This scatter plots represent 141 genetic variants associated with educational attainment on *x*-axis and risk of asthma on *y*-axis. Continuous line represents causal effect of educational attainment on risk of asthma.

Meanwhile, MR studies have found that increased education could reduce the risk of ischemic stroke (Gill et al., 2019; Xiuyun et al., 2020; Harshfield et al., 2021) and Alzheimer's disease (Larsson et al., 2017; Anderson et al., 2020; Andrews et al., 2021; Zhuang et al., 2021a). Anderson et al. (2020) recently examined whether educational attainment and cognitive performance had causal effects on the risk of Alzheimer's disease, independently of each other. They found that educational attainment affected the risk of Alzheimer's disease dependently of cognitive performance (Anderson et al., 2020). However, cognitive performance affected the risk of Alzheimer's disease independently of educational attainment (Anderson et al., 2020).

Hence, all these findings discussed earlier indicated that educational attainment had causal effects on the risk of epilepsy (Wang et al., 2021), coronary heart disease (Gill et al., 2019), stroke (Gill et al., 2019), type 2 diabetes (Liang et al., 2021), and suicide attempt (Rosoff et al., 2020), independently of cognitive performance. However, the causal effect of educational attainment on the risk of Alzheimer's disease may be mediated by cognitive performance (Anderson et al., 2020). Our findings are consistent with recent MR findings in other human complex diseases or phenotypes.

Since 2018, multiple large-scale asthma GWAS datasets have been reported, as described in the *Introduction*. Here, we only selected the large-scale asthma GWAS dataset in 30,810 individuals of European ancestry from Shrine et al. (2019). In brief, these GWAS samples are from GASP, U-BIOPRED, and the UK Biobank (Shrine et al., 2019). In 2018, TAGC examined the common asthma variants by a meta-analysis of worldwide asthma GWAS datasets, including 23,948 asthma cases and 118,538 controls (Demenais et al., 2018). However, all these individuals are from ethnically diverse populations, including European ancestry, African ancestry, Japanese ancestry, and Latino ancestry (Demenais et al., 2018). It is known that all these selected educational attainment genetic variants are from the large-scale GWAS dataset in individuals of European descent (n = 405,072) (Okbay et al., 2016). Hence, we did not select the asthma GWAS dataset from TAGC in our MR analysis (Demenais et al., 2018). Zhu et al. (2018) conducted a genome-wide cross-trait analysis to investigate the shared genetic etiology in asthma and allergic diseases by analyzing large-scale GWAS datasets from the UK Biobank, including 25,685 allergic diseases subjects, 14,085 asthma subjects, and 76,768 controls. Hence, both Shrine et al. (2019) and Zhu et al. (2018) have used the UK Biobank samples. Hence, we did not select the asthma GWAS dataset from Zhu et al. (2018) in our MR analysis. Han et al. (2020) conducted a GWAS using 64,538 asthma cases and 329,321 controls from UK Biobank and then performed a meta-analysis using the UK Biobank and the TAGC datasets. However, they did provide the effect size and the corresponding standard error for each variant in the GWAS summary dataset (Han et al., 2020). Importantly, there is a sample overlap in both studies from Shrine et al. (2019) and Han et al. (2020), as both shared the UK Biobank samples. Hence, we did not select the GWAS dataset from the UK Biobank or the GWAS dataset from the meta-analysis of the UK Biobank and TAGC in our MR analysis (Han et al., 2020).

Meanwhile, our MR analysis still has some limitations. First, our findings are based on the educational attainment GWAS dataset and asthma GWAS dataset in individuals of European ancestry (Okbay et al., 2016; Shrine et al., 2019). It remains unclear about the causal association between educational attainment and asthma in other ancestries. Hence, replication MR studies are required to investigate our findings in the future. Second, both the educational attainment GWAS dataset and asthma GWAS dataset include the samples from the UK Biobank (Okbay et al., 2016; Shrine et al., 2019). In brief, the replication stage in the educational attainment GWAS dataset included

111,349 individuals from the UK Biobank (Okbay et al., 2016). The asthma GWAS dataset included 2,996 asthma cases from the UK Biobank and 25,600 controls from the UK Biobank (Shrine et al., 2019). Hence, the educational attainment GWAS dataset and the asthma GWAS dataset may not be independent. Hence, independent GWAS datasets are also required to evaluate our findings further.

In summary, we highlight the protective role of educational attainment against asthma with 100% statistical power using univariable MR analysis. Meanwhile, we provide evidence to support that educational attainment protects against asthma risk dependently on cognitive performance using multivariable MR analysis. Our findings may have public health applications and deserve further investigation.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

YL and BY designed the project and analyzed the data. All authors wrote the first draft of the manuscript, revised, and approved the final manuscript.

ACKNOWLEDGMENTS

We thank the Social Science Genetic Association Consortium and the UK Biobank for providing the educational attainment GWAS datasets. We also thank GASP, U-BIOPRED, and the UK Biobank for the asthma GWAS datasets.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.716364/full#supplementary-material

REFERENCES

Anderson, E. L., Howe, L. D., Wade, K. H., Ben-Shlomo, Y., Hill, W. D., Deary, I. J., et al. (2020). Education, intelligence and Alzheimer's disease: evidence from a multivariable two-sample Mendelian randomization study. *Int. J. Epidemiol.* 49, 1163–1172. doi: 10.1093/ije/dyz280

Andrews, S. J., Fulton-Howard, B., O'reilly, P., Marcora, E., and Goate, A. M. (2021). Causal associations between modifiable risk factors and the Alzheimer's phenome. *Ann. Neurol.* 89, 54–65. doi: 10.1002/ana.25918

Au Yeung, S. L., Li, A. M., and Schooling, C. M. (2021a). A life course approach to elucidate the role of adiposity in asthma risk: evidence from a Mendelian randomisation study. *J. Epidemiol. Commun. Health* 75, 277–281.

Au Yeung, S. L., Zhao, J. V., and Schooling, C. M. (2021b). Evaluation of glycemic traits in susceptibility to COVID-19 risk: a Mendelian randomization study. *BMC Med.* 19:72. doi: 10.1186/s12916-021-01944-3

Beasley, R., Semprini, A., and Mitchell, E. A. (2015). Risk factors for asthma: is prevention possible? *Lancet* 386, 1075–1085. doi: 10.1016/S0140-6736(15)00156-7

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* 44, 512–525. doi: 10.1093/ije/dyv080

Brion, M. J., Shakhbazov, K., and Visscher, P. M. (2013). Calculating statistical power in Mendelian randomization studies. *Int. J. Epidemiol.* 42, 1497–1501. doi: 10.1093/ije/dyt179

Brite, J., Alper, H. E., Friedman, S., Takemoto, E., and Cone, J. (2020). Association between socioeconomic status and asthma-related emergency department visits among world trade center rescue and recovery workers and survivors. *JAMA Netw. Open* 3:e201600. doi: 10.1001/jamanetworkopen.2020.1600

Burgess, S., and Thompson, S. G. (2017). Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur. J. Epidemiol.* 32, 377–389. doi: 10.1007/s10654-017-0255-x

Carter, A. R., Gill, D., Davies, N. M., Taylor, A. E., Tillmann, T., Vaucher, J., et al. (2019). Understanding the consequences of education inequality on cardiovascular disease: Mendelian randomisation study. *BMJ* 365:l1855. doi: 10.1136/bmj.l1855

Chen, Y. C., Fan, H. Y., Yang, C., and Lee, Y. L. (2020). Early pubertal maturation and risk of childhood asthma: a Mendelian randomization and longitudinal study. *Allergy* 75, 892–900. doi: 10.1111/all.14009

Demenais, F., Margaritte-Jeannin, P., Barnes, K. C., Cookson, W. O. C., Altmuller, J., Ang, W., et al. (2018). Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat. Genet.* 50, 42–53. doi: 10.1038/s41588-017-0014-7

Eagan, T. M., Gulsvik, A., Eide, G. E., and Bakke, P. S. (2004). The effect of educational level on the incidence of asthma and respiratory symptoms. *Respir. Med.* 98, 730–736. doi: 10.1016/j.rmed.2004.02.008

Gill, D., Efstathiadou, A., Cawood, K., Tzoulaki, I., and Dehghan, A. (2019). Education protects against coronary heart disease and stroke independently of cognitive function: evidence from Mendelian randomization. *Int. J. Epidemiol.* 48, 1468–1477. doi: 10.1093/ije/dyz200

Granell, R., Henderson, A. J., Evans, D. M., Smith, G. D., Ness, A. R., Lewis, S., et al. (2014). Effects of BMI, fat mass, and lean mass on asthma in childhood: a Mendelian randomization study. *PLoS Med.* 11:e1001669. doi: 10.1371/journal.pmed.1001669

Han, Y., Jia, Q., Jahani, P. S., Hurrell, B. P., Pan, C., Huang, P., et al. (2020). Genome-wide analysis highlights contribution of immune system pathways to the genetic architecture of asthma. *Nat. Commun.* 11:1776. doi: 10.1038/s41467-020-15649-3

Hancox, R. J., Milne, B. J., Taylor, D. R., Greene, J. M., Cowan, J. O., Flannery, E. M., et al. (2004). Relationship between socioeconomic status and asthma: a longitudinal cohort study. *Thorax* 59, 376–380. doi: 10.1136/thx.2003.010363

Harshfield, E. L., Georgakis, M. K., Malik, R., Dichgans, M., and Markus, H. S. (2021). Modifiable lifestyle factors and risk of stroke: a Mendelian randomization analysis. *Stroke* 52, 931–936. doi: 10.1161/STROKEAHA.120.031710

Kozyrskyj, A. L., Kendall, G. E., Jacoby, P., Sly, P. D., and Zubrick, S. R. (2010). Association between socioeconomic status and the development of asthma: analyses of income trajectories. *Am. J. Public Health* 100, 540–546. doi: 10.2105/AJPH.2008.150771

Larsson, S. C., Mason, A. M., Kar, S., Vithayathil, M., Carter, P., Baron, J. A., et al. (2020). Genetically proxied milk consumption and risk of colorectal, bladder, breast, and prostate cancer: a two-sample Mendelian randomization study. *BMC Med.* 18:370. doi: 10.1186/s12916-020-01839-9

Larsson, S. C., Traylor, M., Malik, R., Dichgans, M., Burgess, S., and Markus, H. S. (2017). Modifiable pathways in Alzheimer's disease: Mendelian randomisation analysis. *BMJ* 359:j5375. doi: 10.1136/bmj.j5375

Liang, J., Cai, H., Liang, G., Liu, Z., Fang, L., Zhu, B., et al. (2021). Educational attainment protects against type 2 diabetes independently of cognitive performance: a Mendelian randomization study. *Acta Diabetol.* 58, 567–574. doi: 10.1007/s00592-020-01647-w

Liu, G., Zhao, Y., Jin, S., Hu, Y., Wang, T., Tian, R., et al. (2018). Circulating vitamin E levels and Alzheimer's disease: a Mendelian randomization study. *Neurobiol. Aging* 72, 181.e1–189.e9. doi: 10.1016/j.neurobiolaging.2018.08.008

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature* 518, 197–206. doi: 10.1038/nature14177

Minelli, C., Van Der Plaat, D. A., Leynaert, B., Granell, R., Amaral, A. F. S., Pereira, M., et al. (2018). Age at puberty and risk of asthma: a Mendelian randomisation study. *PLoS Med.* 15:e1002634. doi: 10.1371/journal.pmed.1002634

Mulugeta, A., Zhou, A., King, C., and Hypponen, E. (2020). Association between major depressive disorder and multiple disease outcomes: a phenome-wide Mendelian randomisation study in the UK Biobank. *Mol. Psychiatry* 25, 1469–1476. doi: 10.1038/s41380-019-0486-1

Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., et al. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* 533, 539–542. doi: 10.1038/nature17671

Park, S., Lee, S., Kim, Y., Cho, S., Kim, K., Kim, Y. C., et al. (2021). Kidney function and obstructive lung disease: a bidirectional Mendelian randomisation study. *Eur. Respir. J.* doi: 10.1183/13993003.00848-2021 [Epub ahead of print].

Raita, Y., Zhu, Z., Camargo, C. A. Jr., Freishtat, R. J., Ngo, D., Liang, L., et al. (2021). Relationship of soluble interleukin-6 receptors with asthma: a Mendelian randomization study. *Front. Med.* 8:665057. doi: 10.3389/fmed.2021.665057

Rosa, M., Chignon, A., Li, Z., Boulanger, M. C., Arsenault, B. J., Bosse, Y., et al. (2019). A Mendelian randomization study of IL6 signaling in cardiovascular diseases, immune-related disorders and longevity. *NPJ Genom. Med.* 4:23. doi: 10.1038/s41525-019-0097-4

Rosoff, D. B., Kaminsky, Z. A., Mcintosh, A. M., Davey Smith, G., and Lohoff, F. W. (2020). Educational attainment reduces the risk of suicide attempt among individuals with and without psychiatric disorders independent of cognition: a bidirectional and multivariable Mendelian randomization study with more than 815,000 participants. *Transl. Psychiatry* 10:388. doi: 10.1038/s41398-020-01047-2

Shen, M., Liu, X., Li, G., Li, Z., and Zhou, H. (2020). Lifetime smoking and asthma: a Mendelian randomization study. *Front. Genet.* 11:769. doi: 10.3389/fgene.2020.00769

Shrine, N., Portelli, M. A., John, C., Soler Artigas, M., Bennett, N., Hall, R., et al. (2019). Moderate-to-severe asthma in individuals of European ancestry: a genome-wide association study. *Lancet Respir. Med.* 7, 20–34. doi: 10.1016/S2213-2600(18)30389-8

Skaaby, T., Taylor, A. E., Thuesen, B. H., Jacobsen, R. K., Friedrich, N., Mollehave, L. T., et al. (2018). Estimating the causal effect of body mass index on hay fever, asthma and lung function using Mendelian randomization. *Allergy* 73, 153–164. doi: 10.1111/all.13242

Sun, J. Y., Zhang, H., Zhang, Y., Wang, L., Sun, B. L., Gao, F., et al. (2021). Impact of serum calcium levels on total body bone mineral density: a mendelian randomization study in five age strata. *Clin. Nutr.* 40, 2726–2733. doi: 10.1016/j.clnu.2021.03.012

Sun, Y. Q., Brumpton, B. M., Langhammer, A., Chen, Y., Kvaloy, K., and Mai, X. M. (2020). Adiposity and asthma in adults: a bidirectional Mendelian randomisation analysis of the HUNT study. *Thorax* 75, 202–208. doi: 10.1136/thoraxjnl-2019-213678

Verbanck, M., Chen, C. Y., Neale, B., and Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* 50, 693–698. doi: 10.1038/s41588-018-0099-7

von Mutius, E., and Smits, H. H. (2020). Primary prevention of asthma: from risk and protective factors to targeted strategies for prevention. *Lancet* 396, 854–866. doi: 10.1016/S0140-6736(20)31861-4

Wang, M., Zhang, Z., Liu, D., Xie, W., Ma, Y., Yao, J., et al. (2021). Educational attainment protects against epilepsy independent of cognitive function: a Mendelian randomization study. *Epilepsia* 62, 1362–1368. doi: 10.1111/epi.16894

Xiuyun, W., Qian, W., Minjun, X., Weidong, L., and Lizhen, L. (2020). Education and stroke: evidence from epidemiology and Mendelian randomization study. *Sci. Rep.* 10:21208. doi: 10.1038/s41598-020-78248-8

Xu, S., Gilliland, F. D., and Conti, D. V. (2019). Elucidation of causal direction between asthma and obesity: a bi-directional Mendelian randomization study. *Int. J. Epidemiol.* 48, 899–907.

Yuan, S., Mason, A. M., Carter, P., Burgess, S., and Larsson, S. C. (2021). Homocysteine, B vitamins, and cardiovascular disease: a Mendelian randomization study. *BMC Med.* 19:97. doi: 10.1186/s12916-021-01977-8

Zhang, L., Tang, L., Xia, K., Huang, T., and Fan, D. (2020). Education, intelligence, and amyotrophic lateral sclerosis: a Mendelian randomization study. *Ann. Clin. Transl. Neurol.* 7, 1642–1647. doi: 10.1002/acn3.51156

Zhao, J. V., and Schooling, C. M. (2019). The role of linoleic acid in asthma and inflammatory markers: a Mendelian randomization study. *Am. J. Clin. Nutr.* 110, 685–690. doi: 10.1093/ajcn/nqz130

Zhao, J. V., and Schooling, C. M. (2021). Using Mendelian randomization study to assess the renal effects of antihypertensive drugs. *BMC Med.* 19:79. doi: 10.1186/s12916-021-01951-4

Zhu, Z., Lee, P. H., Chaffin, M. D., Chung, W., Loh, P. R., Lu, Q., et al. (2018). A genome-wide cross-trait analysis from UK Biobank highlights the shared genetic architecture of asthma and allergic diseases. *Nat. Genet.* 50, 857–864. doi: 10.1038/s41588-018-0121-0

Zhuang, Z., Gao, M., Yang, R., Liu, Z., Cao, W., and Huang, T. (2021a). Causal relationships between gut metabolites and Alzheimer's disease: a bidirectional Mendelian randomization study. *Neurobiol. Aging* 100, 119.e15–119.e18. doi: 10.1016/j.neurobiolaging.2020.10.022

Zhuang, Z., Yao, M., Wong, J. Y. Y., Liu, Z., and Huang, T. (2021b). Shared genetic etiology and causality between body fat percentage and cardiovascular diseases: a large-scale genome-wide cross-trait analysis. *BMC Med.* 19:100. doi: 10.1186/s12916-021-01972-z

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 September 2021
doi: 10.3389/fgene.2021.731391

An Updated Mendelian Randomization Analysis of the Association Between Serum Calcium Levels and the Risk of Alzheimer's Disease

Yuchen Shi[1]†, Ruifei Liu[2]†, Ying Guo[3], Qiwei Li[1], Haichun Zhou[2], Shaolei Yu[2], Hua Liang[1] and Zeguang Li[3]**

[1] Heilongjiang University of Chinese Medicine, Harbin, China, [2] Second Affiliated Hospital, Heilongjiang University of Chinese Medicine, Harbin, China, [3] First Affiliated Hospital, Heilongjiang University of Chinese Medicine, Harbin, China

Edited by:
Guiyou Liu,
Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences (CAS), China

Reviewed by:
Hui Lu,
Capital Medical University, China
Yiqiang Zhan,
Helmholtz Association of German Research Centres, Germany

***Correspondence:**
Hua Liang
prolh01@126.com
Zeguang Li
lzgzrys@163.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *27 June 2021*
Accepted: *09 August 2021*
Published: *08 September 2021*

Citation:
Shi Y, Liu R, Guo Y, Li Q, Zhou H, Yu S, Liang H and Li Z (2021) An Updated Mendelian Randomization Analysis of the Association Between Serum Calcium Levels and the Risk of Alzheimer's Disease. Front. Genet. 12:731391. doi: 10.3389/fgene.2021.731391

It has been a long time that the relationship between serum calcium levels and Alzheimer's disease (AD) remains unclear. Until recently, observational studies have evaluated the association between serum calcium levels and the risk of AD, however, reported inconsistent findings. Meanwhile, a Mendelian randomization (MR) study had been conducted to test the causal association between serum calcium levels and AD risk, however, only selected 6 serum calcium SNPs as the instrumental variables. Hence, these findings should be further verified using additional more genetic variants and large-scale genome-wide association study (GWAS) dataset to increase the statistical power. Here, we conduct an updated MR analysis of the causal association between serum calcium levels and the risk of AD using a two-stage design. In discovery stage, we conducted a MR analysis using 14 SNPs from serum calcium GWAS dataset (*N* = 61,079), and AD GWAS dataset (*N* = 63,926, 21,982 cases, 41,944 cognitively normal controls). All four MR methods including IVW, weighted median, MR-Egger, and MR-PRESSO showed a reduced trend of AD risk with the increased serum calcium levels. In the replication stage, we performed a MR analysis using 166 SNPs from serum calcium GWAS dataset (*N* = 305,349), and AD GWAS dataset (*N* = 63,926, 21,982 cases, 41,944 cognitively normal controls). Only the weighted median indicated that genetically increased serum calcium level was associated with the reduced risk of AD. Hence, additional studies are required to investigate these findings.

Keywords: Alzheimer's disease, serum calcium, GWAS, Mendelian randomization, weighted median

INTRODUCTION

It has been a long time that the relationship between serum calcium levels and Alzheimer's disease (AD) remains unclear (3–4), as few studies had investigated the association of serum calcium levels with AD (Deary et al., 1987; Landfield et al., 1991; Conley et al., 2009). Until recently, observational studies have evaluated the association between serum calcium levels and the risk of AD

(Sato et al., 2019; Ma et al., 2021). However, these observational studies have highlighted inconsistent findings about the association of serum calcium levels with the risk of AD. Some observational studies have found the protective role of high serum calcium levels in AD (Deary et al., 1987; Landfield et al., 1991; Conley et al., 2009; Sato et al., 2019). Landfield et al. (1991) and Conley et al. (2009) found that AD cases had lower serum calcium levels compared with normal age-matched controls. Meanwhile, Shore et al. found that the severely demented patients had lower serum calcium levels compared with mildly affected individuals (Deary et al., 1987). Sato et al. (2019) analyzed the neuroimaging data of 234 mild cognitive impairment (MCI) participants from the Japanese Alzheimer's Disease Neuroimaging Initiative (J-ADNI) study cohort. They found that low serum calcium levels could increase the conversion of MCI to early AD (Sato et al., 2019).

However, other observational studies have identified the harmful role of high serum calcium levels in AD. In a longitudinal population-based study, Kern et al. (2016) reported that compared with women without calcium supplementation, women with calcium supplements had increased risk of dementia and stroke-related dementia. Ma et al. (2021) analyzed the neuroimaging data of 1,224 non-demented elders including 413 cognitively normal and 811 MCI from ADNI. Their results indicated that serum calcium levels increased with the disease severity (Ma et al., 2021). High serum calcium could increase the cognitive decline and the conversion from non-demented status (cognitively normal and MCI) to AD (Ma et al., 2021).

In order to test the causal association between serum calcium levels and AD risk, He et al. (2020) conducted a Mendelian randomization (MR) study using genome-wide association study (GWAS) datasets from serum calcium and AD. He et al. (2020) found that genetically increased serum calcium levels could significantly reduce the risk of AD. This MR analysis still has two limitations. First, He et al. (2020) only selected 8 serum calcium related genetic variants as the potential instrumental variables. They further excluded two genetic variants using the pleiotropy analysis, and the remaining six genetic variants could only explain 0.81% of the serum calcium variance (He et al., 2020). Second, He et al. (2020) used four MR analysis methods including inverse-variance weighted (IVW), Weighted median, MR-Egger, and MR-PRESSO. However, the main analysis method IVW only indicated suggestive association (P = 0.031). Hence, these findings should be further verified using additional more genetic variants and large-scale GWAS dataset to increase the statistical power.

Until recently, large-scale GWAS of serum calcium levels (N = 305,349) and AD (N = 63,926, 21,982 cases, 41,944 cognitively normal controls) have been reported (Kunkle et al., 2019; Young et al., 2021). There GWAS included larger sample size than previous GWAS of serum calcium levels (N = 61,079) (O'seaghdha et al., 2013) and AD (N = 54,162, 17,008 AD cases and 37,154 controls) (Lambert et al., 2013), as used by He and colleagues, respectively (He et al., 2020). Importantly, these datasets are publicly available. Hence, we conduct an updated MR analysis of the causal association between serum calcium levels and the risk of AD using serum calcium GWAS datasets (O'seaghdha et al., 2013; Young et al., 2021), and AD GWAS dataset (Kunkle et al., 2019).

MATERIALS AND METHODS

Study Design Overview

This MR analysis is a two-sample MR study. Hence, we used the GWAS datasets from the exposure (serum calcium) and the outcome (AD) to estimate the effect of exposure on outcome (He et al., 2020). MR analysis has three assumptions, which have been widely described (Liu et al., 2018; Anderson et al., 2020; He et al., 2020; Wang L. et al., 2020; Zhang et al., 2020; Ou et al., 2021; Sproviero et al., 2021). Ethical approvals were provided in the original articles (Lambert et al., 2013; Young et al., 2021). Here, our MR analysis only used the GWAS summary datasets from serum calcium and AD (Lambert et al., 2013; Young et al., 2021). Hence, the informed consent is not needed. **Figure 1** provides the framework of MR.

Genetic Instrument Selection (Discovery)

In discovery stage, 14 serum calcium single nucleotide polymorphisms (SNPs) were selected including 8 SNPs at the genome-wide significance threshold (P < 5.00E-08), and 6 SNPs with P < 1.00E-04 (O'seaghdha et al., 2013). The 14 serum calcium SNPs were identified by a GWAS using 61,079 individuals of European descent (O'seaghdha et al., 2013). These 14 serum calcium SNPs, especially the 8 SNPs at the genome-wide significance threshold, have been widely used as the potential instrumental variables to evaluate the association of serum calcium with other human complex diseases or phenotypes (Larsson et al., 2017, 2019; Xu et al., 2017; Meng et al., 2020; Wang Y. et al., 2020; Qu et al., 2021; Sun et al., 2021; Young et al., 2021; Yuan et al., 2021). Detailed information about these 14 SNPs is presented in **Supplementary Table 1**.

Genetic Instrument Selection (Replication)

In replication stage, 208 independent SNPs associated serum calcium levels at the genome-wide significance threshold (P < 5.00E-08) were identified by a recent GWAS using 305,349 individuals from the UK Biobank (Young et al., 2021). Compared with 7 SNPs explaining 0.9% of the total variance of total serum calcium, these 208 SNPs explain 5.8% of the total variance of total serum calcium (Young et al., 2021). Detailed information about these 208 SNPs is presented in **Supplementary Table 2**.

AD GWAS Selection

The discovery GWAS summary statistics of AD were obtained from the International Genomics of Alzheimer's Project (IGAP) stage 1 including 21,982 AD and 41,944 cognitively normal controls of European descent (Kunkle et al., 2019). The IGAP stage 1 is based the meta-analysis of four AD GWAS datasets

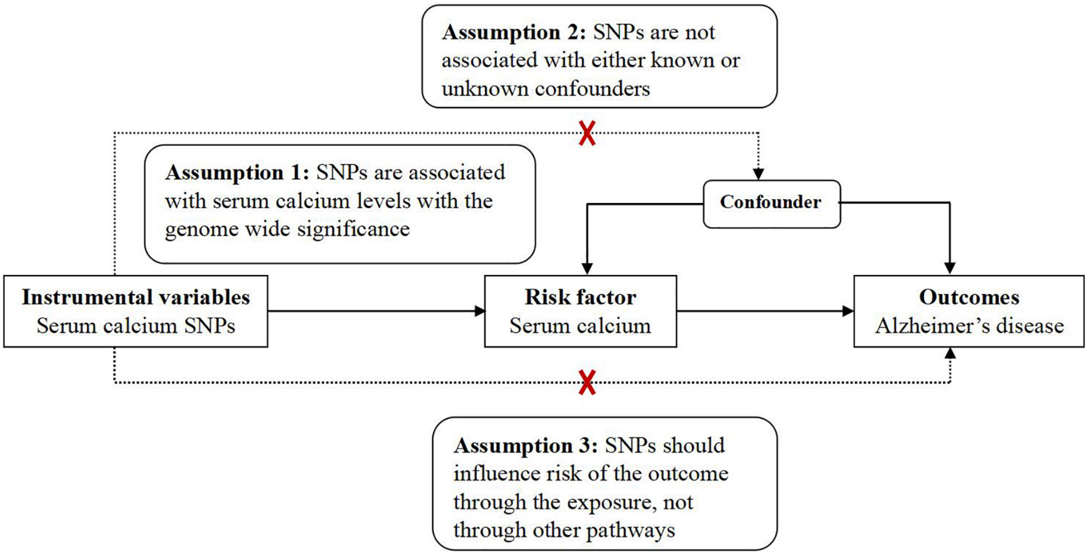

Assumption 2: SNPs are not associated with either known or unknown confounders
Assumption 1: SNPs are associated with serum calcium levels with the genome wide significance
Confounder
Instrumental variables
Serum calcium SNPs
Risk factor
Serum calcium
Outcomes
Alzheimer's disease
Assumption 3: SNPs should influence risk of the outcome through the exposure, not through other pathways

FIGURE 1 | The framework of MR. MR analysis has three assumptions. Assumption 1: SNPs are associated with serum calcium levels with the genome wide significance; Assumption 2: SNPs are not associated with either known or unknown confounders; Assumption 3: SNPs should influence risk of the outcome through the exposure, not through other pathways.

TABLE 1 | Association of 14 serum calcium SNPs with AD risk.

SNP	Serum calcium						AD		
	EA	NEA	EAF	Beta	SE	*P* value	Beta	SE	*P* value
rs10491003	T	C	0.09	0.027	0.005	4.80E-09	−0.0287	0.0246	0.2442
rs11967485	G	A	0.9	0.026	0.005	9.40E-07	−0.0026	0.0248	0.9175
rs12150338	T	C	0.09	0.03	0.006	1.50E-06	0.0491	0.0285	0.08516
rs1550532	C	G	0.31	0.018	0.003	8.20E-11	−0.0027	0.0154	0.8593
rs1570669	G	A	0.34	0.018	0.003	9.10E-12	−0.0015	0.015	0.9188
rs17711722	T	C	0.47	0.015	0.003	8.20E-09	−0.0112	0.0171	0.5115
rs1801725	T	G	0.15	0.071	0.004	8.90E-86	−0.0346	0.0202	0.08741
rs2281558	T	G	0.25	0.015	0.003	5.10E-06	−0.0267	0.0168	0.1128
rs2885836	A	G	0.24	0.012	0.003	5.40E-05	−0.023	0.017	0.1759
rs4074995	A	G	0.28	0.013	0.003	4.60E-06	−0.0153	0.016	0.3385
rs7336933	G	A	0.85	0.022	0.004	9.10E-10	−0.0113	0.0203	0.5779
rs7481584	G	A	0.7	0.018	0.003	1.20E-10	0.0161	0.0157	0.3042
rs780094	T	C	0.42	0.017	0.003	1.30E-10	0.0177	0.0145	0.2216
rs9447004	A	G	0.48	0.012	0.003	3.30E-06	0.0111	0.0143	0.4387

SNP, single-nucleotide polymorphism; EA, effect allele; NEA, non-effect allele; EAF, effect allele frequency; SE, standard error; Beta, regression coefficient based on the effect allele.

including Alzheimer Disease Genetics Consortium, Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium (CHARGE), The European Alzheimer's Disease Initiative (EADI), and Genetic and Environmental Risk in AD/Defining Genetic, Polygenic and Environmental Risk for Alzheimer's Disease Consortium (GERAD/PERADES) (Kunkle et al., 2019). AD cases were autopsy-confirmed or clinically confirmed using the NINCDS-ADRDA criteria or DSM-IV guidelines (Kunkle et al., 2019). The IGAP AD GWAS summary statistics have been widely used in recent MR analysis (Liu et al., 2018; Anderson et al., 2020; He et al., 2020; Wang L. et al., 2020; Zhang et al., 2020; Ou et al., 2021; Sproviero et al., 2021).

MR Method Selection

Four MR methods were selected to evaluate the causal association between serum calcium and the risk of AD including the main analysis method inverse-variance weighted meta-analysis (IVW) (Bowden et al., 2016), and other three additional analysis methods weighted median (Bowden et al., 2016), MR-Egger (Burgess and Thompson, 2017), and Mendelian

TABLE 2 | MR results in discovery and replication stages.

Stage	Method	OR	95% CI	*P* value
Discovery	Weighted median	0.67	0.40–1.12	1.22E-01
Discovery	IVW	0.76	0.51–1.15	1.94E-01
Discovery	MR-Egger	0.66	0.30–1.42	2.87E-01
Discovery	MR-PRESSO Raw	0.76	0.51–1.15	2.17E-01
Discovery	MR-PRESSO Outlier-corrected	NA	NA	NA
Replication	Weighted median	0.15	0.02–0.90	3.80E-02
Replication	IVW	1.14	0.36–3.64	8.18E-01
Replication	MR-Egger	0.27	0.03–2.19	2.22E-01
Replication	MR-PRESSO Raw	1.15	0.36–3.64	8.19E-01
Replication	MR-PRESSO Outlier-corrected	0.85	0.30–2.39	7.62E-01

OR, odds ratio; CI, confidence interval; IVW, Inverse-variance weighted meta-analysis.

randomization pleiotropy residual sum and outlier (MR-PRESSO) (Verbanck et al., 2018). Meanwhile, MR-Egger intercept test and MRPRESSO Global test were used to evaluate the evidence of pleiotropy (Bowden et al., 2016; Burgess and Thompson, 2017; Verbanck et al., 2018; Bowden and Holmes, 2019). The odds ratio (OR) and 95% confidence interval (CI) of AD corresponds to 1 standard deviation (SD) in serum calcium levels. The statistical significance threshold was $P < 0.05$. All analyses were performed using R Version 4.0.3 and R packages ("MendelianRandomization") and ("MRPRESSO") (Yavorska and Burgess, 2017; Verbanck et al., 2018).

RESULTS

MR Analysis in the Discovery Stage

All these 14 serum calcium SNPs are available in the AD GWAS dataset. We then extracted their corresponding summary statistics for MR analysis, as provided in **Table 1**. The main and other additional MR methods indicated no significant association between serum calcium and the risk of AD including weighted median (OR = 0.67, 95% CI: 0.40–1.12, *P* = 1.22E-01), IVW (OR = 0.76, 95% CI: 0.51–1.15, *P* = 1.94E-01), MR-Egger (OR = 0.66, 95% CI: 0.30–1.42, *P* = 2.87E-01), and MR-PRESSO

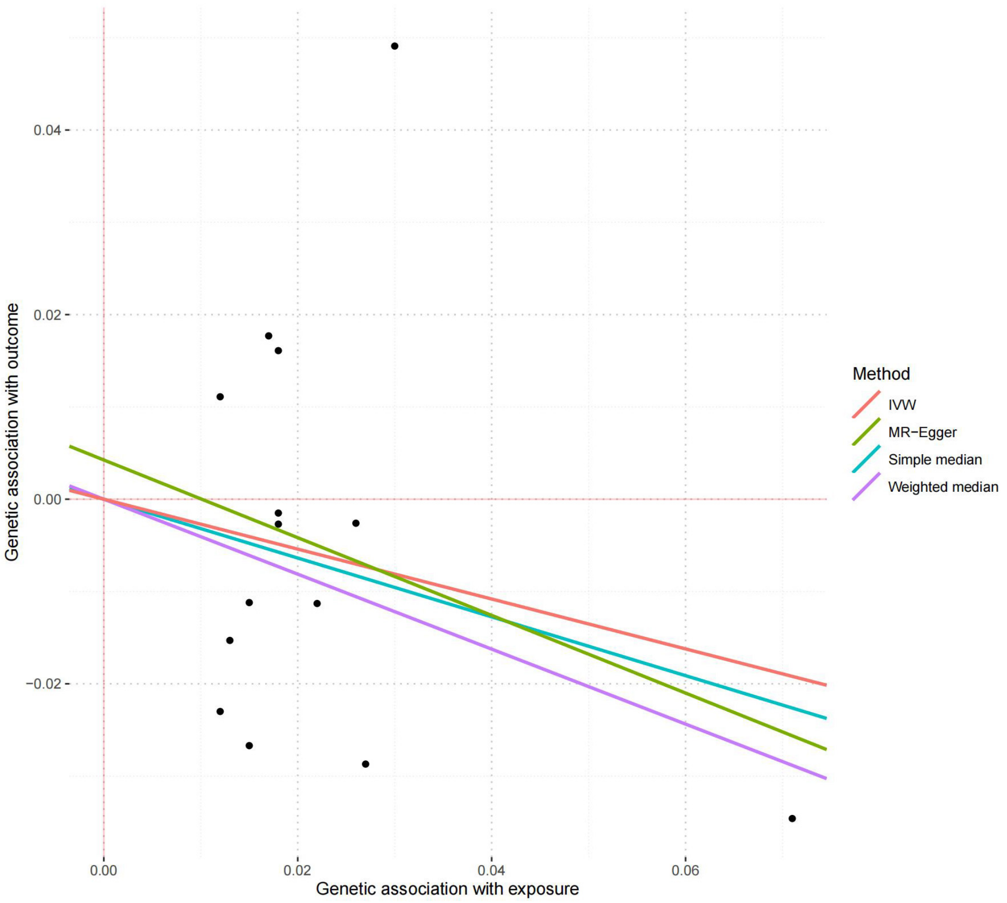


FIGURE 2 | The scatter plot of the MR analysis in discovery stage using different methods. The scatter plot is based on the single causal estimates from 14 serum calcium SNPs using IVW, weighted median, simple median and MR-Egger, respectively. The scatter plot depicts the causal relationship between serum calcium level and the risk of AD. The *X*-axis stands for the effect estimate (beta coefficient) of serum calcium level utilizing a certain SNP; stands for the effect estimate (beta coefficient) of AD risk utilizing a certain IVW, Inverse variance weighting.

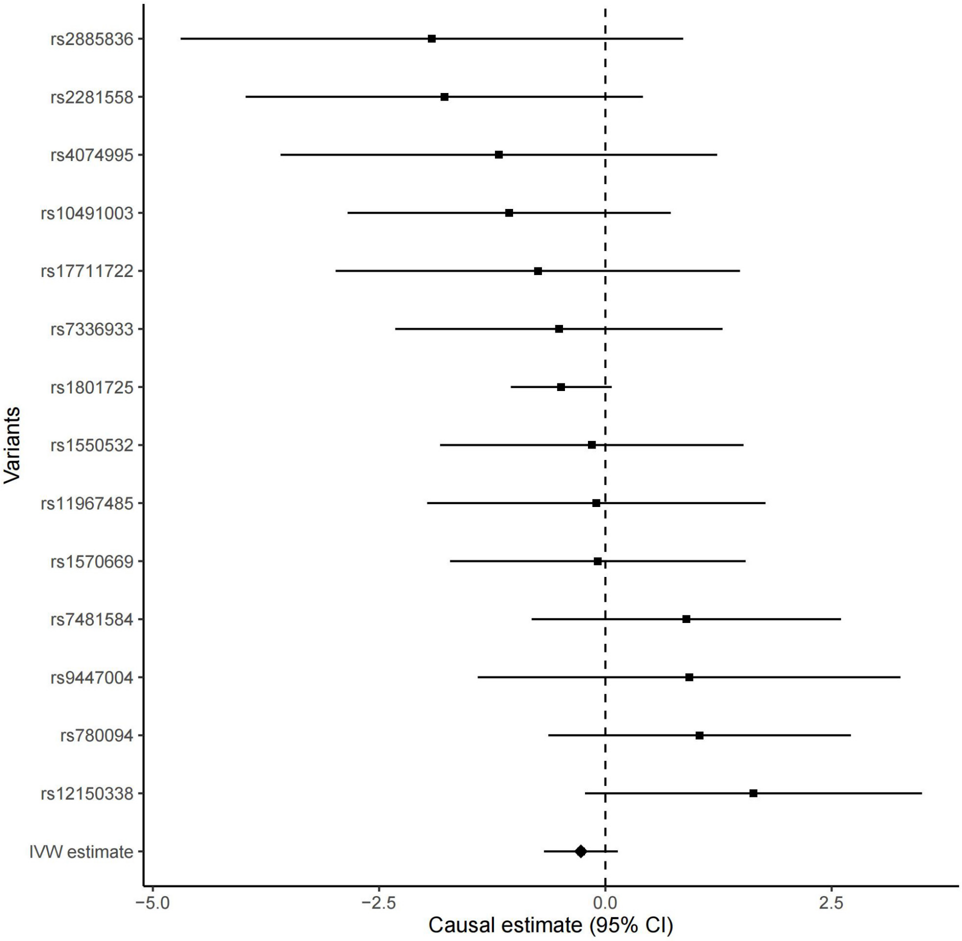


FIGURE 3 | The forest plot of the single Mendelian randomization causal estimates for the association between genetically predicted serum calcium and the risk of AD from 14 serum calcium SNPs using IVW. The black point showed the causal effect estimate (beta coefficient) of serum calcium level on the risk of AD utilizing a certain SNP, and the black line indicated the 95% CI of the estimate. "IVW estimate" reports the effect using all SNPs estimated by the inverse-variance weighted method. CI, confidence interval.

(OR = 0.76, 95% CI: 0.51–1.15, *P* = 2.17E-01), as provided in **Table 2**. However, all these four methods showed a reduced trend of AD risk with the increased serum calcium levels. Meanwhile, the MR-Egger intercept test (with intercept = 0.004, and *P* = 0.650) and MRPRESSO Global Test (*P* = 0.337) did not indicate evidence of pleiotropy. **Figure 2** is the scatter plot of the single causal estimates from these 14 serum calcium SNPs using IVW, weighted median, simple median and MR-Egger. **Figures 3**, **4** are the forest plot, and funnel plot of the single causal estimates from these 14 serum calcium SNPs using IVW, respectively.

MR Analysis in the Replication Stage

166 of the 208 serum calcium SNPs are included in the AD GWAS dataset. We then extracted the summary statistics of these 166 SNPs for the MR analysis, as provided in **Supplementary Table 3**. Using the weighted median, we found that the genetically increased serum calcium level (per 1 SD increase) was associated with the reduced risk of AD (OR = 0.15, 95% CI: 0.02–0.90, *P* = 3.80E-02) (**Table 2**). However, the other MR methods did not reported any significant results including IVW (OR = 1.14, 95% CI: 0.36–3.64, *P* = 8.18E-01) and MR-Egger (OR = 0.27, 95% CI: 0.03–2.19, *P* = 2.22E-01). Meanwhile, MR-Egger intercept test did not indicate evidence of pleiotropy with intercept = 0.005, and *P* = 0.105. Using MRPRESSO, we found evidence of pleiotropy with Global Test *P* = 0.006. The MR-PRESSO Raw estimate is OR = 1.15, 95% CI: 0.36–3.64, *P* = 8.19E-01. The MR-PRESSO Outlier-corrected estimate is OR = 0.85, 95% CI: 0.30–2.39, *P* = 7.62E-01. **Figure 5** is the scatter plot of the single causal estimates from these 166 serum calcium SNPs using IVW, weighted median, simple median and MR-Egger.

DISCUSSION

Calcium signaling is involved in many different intracellular and extracellular processes (Marambaud et al., 2009). It is known that AD is characterized by the extracellular accumulation of

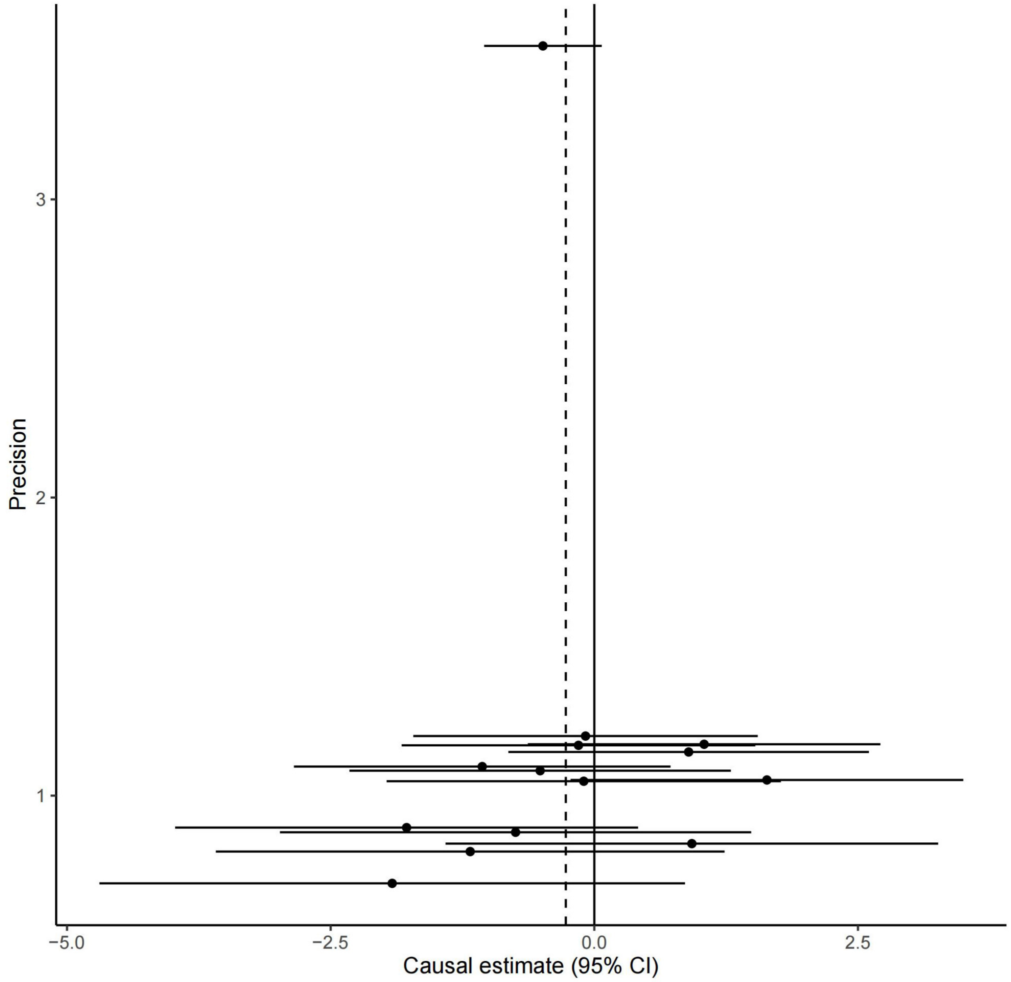


FIGURE 4 | The funnel plot of the single causal estimates from 14 serum calcium SNPs using IVW. The funnel plot shows the potential bias of the selected 14 serum calcium SNPs. The *X*-axis stands for the causal effect estimate (beta coefficient) of serum calcium level on the risk of AD utilizing a certain SNP and the *Y*-axis is the reciprocal of standard error for each causal effect estimate. CI, confidence interval.

amyloid (Aβ) plaques and intracellular neurofibrillary tangles (NFTs) in the brain (Tong et al., 2018). Evidence shows that the calcium dysregulation occurs prior the key AD pathologies including plaques, tangles, and synaptic deficits (Tong et al., 2018). The disrupted calcium could further induce synaptic deficits, and promote the accumulation of Aβ plaques and NFTs (Tong et al., 2018). Hence, deregulated calcium homeostasis may play an important role in the pathogenesis of AD (Marambaud et al., 2009).

Until now, observational studies by analyzing the neuroimaging data have evaluated the association between serum calcium levels and the risk of AD, however, reported inconsistent findings (Sato et al., 2019; Ma et al., 2021). Sato et al. (2019) concluded that low serum calcium levels increased the conversion of MCI to early AD. Ma et al. (2021) found that high serum calcium increased the cognitive decline and the conversion from non-demented status (cognitively normal and MCI) to AD. Two reasons have caused these inconsistent findings. First, Sato et al. (2019) selected a total of 234 MCI individuals, and Ma et al. (2021) selected 413 cognitively normal and 811 MCI. Hence, the sample size may have affected the conclusions from both studies. Second, the samples used in both studies are of different descents including one from Japanese and the other European. Hence, different descents may have also affected the conclusions from both studies (Sato et al., 2019; Ma et al., 2021). Meanwhile, a longitudinal population-based study had tested the association between calcium supplementation and dementia in 700 dementia-free women aged 70–92 years (Kern et al., 2016). The results indicated that women with calcium supplements had higher risk of developing dementia than women without calcium supplementation (Kern et al., 2016). A cross sectional study in 337 subjects in India indicated that increased calcium level could increase the cognitive score (Basheer et al., 2016).

Here, we conduct an updated MR analysis of the causal association between serum calcium levels and the risk of AD using a two-stage design. In discovery stage, we conducted a

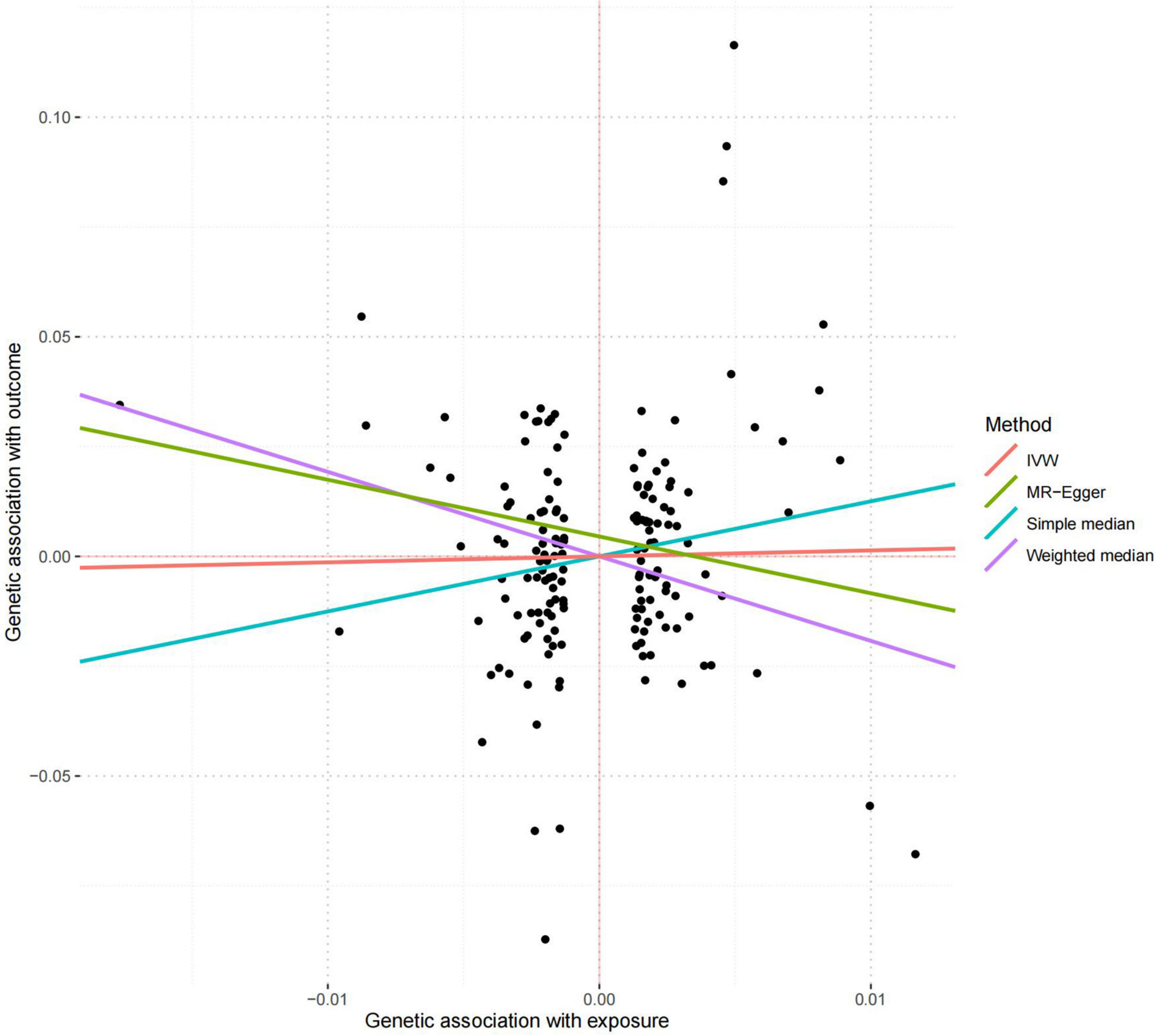


FIGURE 5 | The scatter plot of the MR analysis in replication stage using different methods. The scatter plot is based on the single causal estimates from 166 serum calcium SNPs using IVW, weighted median, simple median and MR-Egger, respectively. The scatter plot depicts the causal relationship between serum calcium level and the risk of AD. The *X*-axis stands for the effect estimate (beta coefficient) of serum calcium level utilizing a certain SNP; stands for the effect estimate (beta coefficient) of AD risk utilizing a certain IVW, Inverse variance weighting.

MR analysis using 14 SNPs from serum calcium GWAS dataset (*N* = 61,079) (O'seaghdha et al., 2013), and AD GWAS dataset (*N* = 63,926, 21,982 cases, 41,944 cognitively normal controls) (Kunkle et al., 2019). All four MR methods including IVW, weighted median, MR-Egger, and MR-PRESSO showed a reduced trend of AD risk with the increased serum calcium levels. In the replication stage, we performed a MR analysis using 166 SNPs from serum calcium GWAS dataset (*N* = 305,349) (Young et al., 2021), and AD GWAS dataset (*N* = 63,926, 21,982 cases, 41,944 cognitively normal controls) (Kunkle et al., 2019). Only the weighted median indicated that genetically increased serum calcium level was associated with the reduced risk of AD, which indicates that 50% of the weight comes from the valid instrumental variables (Bowden et al., 2016; Bowden and Holmes, 2019). Our findings may have clinical application that high serum calcium level by diet or calcium supplementation may contribute to reduce the risk of AD. However, IVW, MR-Egger, and MR-PRESSO indicated no causal association between serum calcium level and the risk of AD. Hence, additional studies including MR studies and especially randomized controlled trials are required to investigate these findings.

Compared with the original MR study from He and colleagues, our MR analysis may have several strengths. First, we selected a large-scale AD GWAS dataset (*N* = 63,926, 21,982 cases, 41,944 cognitively normal controls) (Kunkle et al., 2019), which included more additional samples compared with the original study (*N* = 54,162, 21,982 cases, 41,944 cognitively normal controls), as used by He and colleagues (Lambert et al., 2013). Second, we selected 14 serum calcium SNPs in the discovery stage and 166 serum calcium SNPs in the replication stage. He and colleagues only selected six SNPs as the effective instrumental variables, and only observed suggestive association ($P = 0.031$) (He et al., 2020). Our undated MR analysis significantly increased the number of instrumental variables, which may contribute to the increases statistical power in MR analysis (He et al., 2020). Meanwhile, this two-sage method may contribute to test the replication and robustness of MR estimate. Third, the individuals from both the serum calcium and AD GWAS are of European descent. Hence, our MR analysis may have reduced the population stratification bias. Fourth, multiple MR and pleiotropy analysis methods including IVW, weighted median, MR-Egger, and MR-PRESSO were selected to reduce the pleiotropy.

Meanwhile, our MR study may also have some limitations. First, we selected 14 SNPs in the discovery stage, and 208 SNPs in the replication stage, as the potential instrumental variables. However, they are not completely in linkage disequilibrium. Hence, the linkage disequilibrium may have influenced the MR findings. Second, the 208 serum calcium SNPs are identified using is based UK Biobank samples (Young et al., 2021), and the AD GWAS dataset is based on the 21,982 AD and 41,944 cognitively normal controls of European descent (Kunkle et al., 2019). Hence, we could not ensure that the serum calcium GWAS dataset and AD GWAS dataset are completely independent with each other. Hence, the cryptic relatedness may have influenced the MR findings. Third, our MR findings are based on the individuals of European descent. Considering the genetic heterogeneity across the different descents, the MR findings between serum calcium levels and the risk of AD may be different. Hence, our findings are required to be tested in other populations. Fourth, we have evaluated the pleiotropy using both the MR-Egger intercept test and MRPRESSO test. However, we could not completely exclude all the pleiotropy. Hence, there may be other confounding factors, which may have influenced our MR findings. Hence, future studies are required to verify our findings.

CONCLUSION

Collectively, our updated MR analysis highlighted a reduced trend of AD risk with the increased serum calcium levels in the discovery stage, and reduced risk of AD in the replication stage. Meanwhile, additional studies are required to investigate our findings.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

ZL and HL: conception and design. YS and RL: development of methodology and acquisition of data. YS, RL, YG, QL, HZ, and SY: analysis and interpretation of data, writing, review, and revision of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the National Natural Science Foundation of China (No. 81473563). The i-Select chips was funded by the French National Foundation on AD and related disorders. EADI was supported by the LABEX (laboratory of excellence program investment for the future) DISTALZ grant, Inserm, Institut Pasteur de Lille, Université de Lille 2 and the Lille University Hospital. GERAD was supported by the Medical Research Council (Grant no 503480), Alzheimer's Research UK (Grant no 503176), the Wellcome Trust (Grant no 082604/2/07/Z) and German Federal Ministry of Education and Research (BMBF): Competence Network Dementia (CND) grant nos 01GI0102, 01GI0711, and 01GI0420. CHARGE was partly supported by the NIH/NIA grant R01 AG033193 and the NIA AG081220 and AGES contract N01-AG-12100, the NHLBI grant R01 HL105756, the Icelandic Heart Association, and the Erasmus Medical Center and Erasmus University. ADGC was supported by the NIH/NIA grants: U01 AG032984, U24 AG021886, U01 AG016976, and the Alzheimer's Association grant ADGC-10-196728.

ACKNOWLEDGMENTS

We thank International Genomics of Alzheimer's Project (IGAP) for their AD GWAS summary data. The investigators within IGAP contributed to the design and implementation of IGAP and/or provided data but did not participate in analysis or writing of this report. IGAP was made possible by the generous participation of the control subjects, the patients, and their families.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.731391/full#supplementary-material

REFERENCES

Anderson, E. L., Richmond, R. C., Jones, S. E., Hemani, G., Wade, K. H., and Dashti, H. S. (2020). Is disrupted sleep a risk factor for Alzheimer's disease? Evidence from a two-sample Mendelian randomization analysis. *Int. J. Epidemiol.* 50, 817–828.

Basheer, M. P., Kumar, K. M. P., Sreekumaran, E., and Ramakrishnac, T. (2016). A study of serum magnesium, calcium and phosphorus level, and cognition in the elderly population of South India. *Alex. J. Med.* 52, 303–308. doi: 10.1016/j.ajme.2015.11.001

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* 40, 304–314. doi: 10.1002/gepi.21965

Bowden, J., and Holmes, M. V. (2019). Meta-analysis and Mendelian randomization: a review. *Res. Synth. Methods* 10, 486–496. doi: 10.1002/jrsm.1346

Burgess, S., and Thompson, S. G. (2017). Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur. J. Epidemiol.* 32, 377–389. doi: 10.1007/s10654-017-0255-x

Conley, Y. P., Mukherjee, A., Kammerer, C., Dekosky, S. T., Kamboh, M. I., Finegold, D. N., et al. (2009). Evidence supporting a role for the calcium-sensing receptor in Alzheimer disease. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 150B, 703–709. doi: 10.1002/ajmg.b.30896

Deary, J., Hendrickson, I. E., and Alistairburns, A. (1987). Hair and serum copper, zinc, calcium, and magnesium concentrations in Alzheimer-type dementia. *Pers. Individ. Dif.* 8, 75–80.

He, Y., Zhang, H., Wang, T., Han, Z., Ni, Q. B., Wang, K., et al. (2020). Impact of Serum Calcium Levels on Alzheimer's Disease: a Mendelian Randomization Study. *J. Alzheimers Dis.* 76, 713–724.

Kern, J., Kern, S., Blennow, K., Zetterberg, H., Waern, M., Guo, X., et al. (2016). Calcium supplementation and risk of dementia in women with cerebrovascular disease. *Neurology* 87, 1674–1680. doi: 10.1212/wnl.0000000000003111

Kunkle, B. W., Grenier-Boley, B., Sims, R., Bis, J. C., Damotte, V., Naj, A. C., et al. (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing. *Nat. Genet.* 51, 414–430.

Lambert, J. C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., et al. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* 45, 1452–1458.

Landfield, P. W., Applegate, M. D., Schmitzer-Osborne, S. E., and Naylor, C. E. (1991). Phosphate/calcium alterations in the first stages of Alzheimer's disease: implications for etiology and pathogenesis. *J. Neurol. Sci.* 106, 221–229. doi: 10.1016/0022-510x(91)90261-5

Larsson, S. C., Burgess, S., and Michaelsson, K. (2017). Association of Genetic Variants Related to Serum Calcium Levels With Coronary Artery Disease and Myocardial Infarction. *JAMA* 318, 371–380. doi: 10.1001/jama.2017.8981

Larsson, S. C., Traylor, M., Burgess, S., Boncoraglio, G. B., Jern, C., Michaelsson, K., et al. (2019). Serum magnesium and calcium levels in relation to ischemic stroke: mendelian randomization study. *Neurology* 92, e944–e950.

Liu, G., Zhao, Y., Jin, S., Hu, Y., Wang, T., Tian, R., et al. (2018). Circulating vitamin E levels and Alzheimer's disease: a Mendelian randomization study. *Neurobiol. Aging* 72, e181–e189.

Ma, L. Z., Wang, Z. X., Wang, Z. T., Hou, X. H., Shen, X. N., Ou, Y. N., et al. (2021). Serum Calcium Predicts Cognitive Decline and Clinical Progression of Alzheimer's Disease. *Neurotox. Res.* 39, 609–617. doi: 10.1007/s12640-020-00312-y

Marambaud, P., Dreses-Werringloer, U., and Vingtdeux, V. (2009). Calcium signaling in neurodegeneration. *Mol. Neurodegener.* 4:20. doi: 10.1186/1750-1326-4-20

Meng, Q., Huang, L., Tao, K., Liu, Y., Jing, J., Wang, W., et al. (2020). Integrated Genetics and Micronutrient Data to Inform the Causal Association Between Serum Calcium Levels and Ischemic Stroke. *Front. Cell Dev. Biol.* 8:590903. doi: 10.3389/fcell.2020.590903

O'seaghdha, C. M., Wu, H., Yang, Q., Kapur, K., Guessous, I., Zuber, A. M., et al. (2013). Meta-analysis of genome-wide association studies identifies six new Loci for serum calcium concentrations. *PLoS Genet.* 9:e1003796. doi: 10.1371/journal.pgen.1003796

Ou, Y. N., Yang, Y. X., Shen, X. N., Ma, Y. H., Chen, S. D., Dong, Q., et al. (2021). Genetically determined blood pressure, antihypertensive medications, and risk of Alzheimer's disease: a Mendelian randomization study. *Alzheimers Res. Ther.* 13:41.

Qu, Z., Yang, F., Hong, J., Wang, W., Li, S., Jiang, G., et al. (2021). Causal relationship of serum nutritional factors with osteoarthritis: a Mendelian randomization study. *Rheumatology* 60, 2383–2390. doi: 10.1093/rheumatology/keaa622

Sato, K., Mano, T., Ihara, R., Suzuki, K., Tomita, N., Arai, H., et al. (2019). Lower Serum Calcium as a Potentially Associated Factor for Conversion of Mild Cognitive Impairment to Early Alzheimer's Disease in the Japanese Alzheimer's Disease Neuroimaging Initiative. *J. Alzheimers Dis.* 68, 777–788. doi: 10.3233/jad-181115

Sproviero, W., Winchester, L., Newby, D., Fernandes, M., Shi, L., Goodday, S. M., et al. (2021). High Blood Pressure and Risk of Dementia: a Two-Sample Mendelian Randomization Study in the UK Biobank. *Biol. Psychiatry* 89, 817–824. doi: 10.1016/j.biopsych.2020.12.015

Sun, J. Y., Zhang, H., Zhang, Y., Wang, L., Sun, B. L., Gao, F., et al. (2021). Impact of serum calcium levels on total body bone mineral density: a mendelian randomization study in five age strata. *Clin. Nutr.* 40, 2726–2733. doi: 10.1016/j.clnu.2021.03.012

Tong, B. C., Wu, A. J., Li, M., and Cheung, K. H. (2018). Calcium signaling in Alzheimer's disease & therapies. *Biochim. Biophys. Acta Mol. Cell Res.* 1865, 1745–1760.

Verbanck, M., Chen, C. Y., Neale, B., and Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* 50, 693–698. doi: 10.1038/s41588-018-0099-7

Wang, L., Qiao, Y., Zhang, H., Zhang, Y., Hua, J., Jin, S., et al. (2020). Circulating Vitamin D Levels and Alzheimer's Disease: a Mendelian Randomization Study in the IGAP and UK Biobank. *J. Alzheimers Dis.* 73, 609–618. doi: 10.3233/jad-190713

Wang, Y., Gao, L., Lang, W., Li, H., Cui, P., Zhang, N., et al. (2020). Serum Calcium Levels and Parkinson's Disease: a Mendelian Randomization Study. *Front. Genet.* 11:824. doi: 10.3389/fgene.2020.00824

Xu, L., Lin, S. L., and Schooling, C. M. (2017). A Mendelian randomization study of the effect of calcium on coronary artery disease, myocardial infarction and their risk factors. *Sci. Rep.* 7:42691.

Yavorska, O. O., and Burgess, S. (2017). MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int. J. Epidemiol.* 46, 1734–1739. doi: 10.1093/ije/dyx034

Young, W. J., Warren, H. R., Mook-Kanamori, D. O., Ramirez, J., Van Duijvenboden, S., Orini, M., et al. (2021). Genetically Determined Serum Calcium Levels and Markers of Ventricular Repolarization: a Mendelian Randomization Study in the UK Biobank. *Circ. Genom .Precis. Med.* 14:e003231.

Yuan, S., Giovannucci, E. L., and Larsson, S. C. (2021). Gallstone disease, diabetes, calcium, triglycerides, smoking and alcohol consumption and pancreatitis risk: mendelian randomization study. *NPJ Genom. Med.* 6:27.

Zhang, H., Wang, T., Han, Z., Wang, L., Zhang, Y., and Liu, G. (2020). Impact of Vitamin D Binding Protein Levels on Alzheimer's Disease: a Mendelian Randomization Study. *J. Alzheimers Dis.* 74, 991–998.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

BRIEF RESEARCH REPORT
published: 22 September 2021
doi: 10.3389/fgene.2021.744299

Joint Analysis of Genome-Wide Association Data Reveals No Genetic Correlations Between Low Back Pain and Neurodegenerative Diseases

Pengfei Wu[1,2†], *Bing Du*[1†], *Bing Wang*[3,4*], *Rui Yin*[5], *Xin Lv*[3], *Yuliang Dai*[3,4], *Wan Zhang*[2,6] *and Kun Xia*[1,7,8*]

[1] Center for Medical Genetics & Hunan Key Laboratory of Medical Genetics, School of Life Sciences, Central South University, Changsha, China, [2] Department of Neurology, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA, United States, [3] Department of Spine Surgery, The Second Xiangya Hospital, Central South University, Changsha, China, [4] Center for Digital Spine Surgery, Central South University, Changsha, China, [5] Department of Biomedical Informatics and Harvard Medical School, Boston, MA, United States, [6] Department of Biology, College of Arts & Sciences, Boston University, Boston, MA, United States, [7] CAS Center for Excellence in Brain Science and Intelligence Technology (CEBSIT), Shanghai, China, [8] Hengyang Medical School, University of South China, Hengyang, China

Edited by:
Guiyou Liu,
Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, China

Reviewed by:
Peng-Peng Niu,
First Affiliated Hospital of Zhengzhou University, China
Yi Li,
Air Force General Hospital PLA, China
Huifang Shang,
Sichuan University, China

****Correspondence:***
Bing Wang
wbxyeyy@csu.edu.cn
Kun Xia
xiakun@sklmg.edu.cn

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *20 July 2021*
Accepted: *23 August 2021*
Published: *22 September 2021*

Citation:
Wu P, Du B, Wang B, Yin R, Lv X, Dai Y, Zhang W and Xia K (2021) Joint Analysis of Genome-Wide Association Data Reveals No Genetic Correlations Between Low Back Pain and Neurodegenerative Diseases. Front. Genet. 12:744299. doi: 10.3389/fgene.2021.744299

Background: We aimed to explore the genetic correlation and bidirectional causal relationships between low back pain (LBP) and three neurodegenerative diseases, Alzheimer's disease (AD), Parkinson's disease (PD), and amyotrophic lateral sclerosis (ALS).

Methods: Summary-level statistics were obtained from genome-wide association studies of LBP (n = 177,860), AD (n = 63,926), PD (n = 482,730), and ALS (n = 80,610). We implemented linkage disequilibrium score regression to calculate heritability estimates and genetic correlations. To investigate possible causal associations between LBP and three neurodegenerative diseases, we also conducted a bidirectional two-sample Mendelian randomization (MR) study. Inverse variance-weighted MR was employed as the primary method to generate overall estimates, whereas complementary approaches and sensitivity analyses were conducted to confirm the consistency and robustness of the findings.

Results: There was no evidence of genetic correlations between LBP and AD (Rg = −0.033, p = 0.766). MR analyses did not support the causal effect of LBP on AD (OR = 1.031; 95% CI, 0.924–1.150; p = 0.590) or the effect of AD on LBP (OR = 0.963; 95% CI, 0.923–1.006; p = 0.090). Likewise, this study failed to identify genetic correlations between LBP and two other neurodegenerative diseases. MR results of the associations of LBP with PD and ALS, and the reverse associations, did not reach Bonferroni-corrected significance.

Conclusion: The study did not support genetic correlations or causations between LBP and three common neurodegenerative diseases, AD, PD, and ALS in the European population.

Keywords: low back pain, Alzheimer's disease, Parkinson's disease, amyotrophic lateral sclerosis, Mendelian randomization, linkage disequilibrium score regression

INTRODUCTION

Neurodegenerative diseases have imposed a heavy burden on the global healthcare in line with the accelerated trend of population aging. Alzheimer's disease (AD) is the most common neurodegenerative disorder and the leading cause of dementia characterized by severe decline in cognitive function (Weller and Budson, 2018). Parkinson's disease (PD) is the second most common neurodegenerative disease and the primary movement disorder attributed to neurodegeneration (Balestrino and Schapira, 2020). Amyotrophic lateral sclerosis (ALS), also known as Lou Gehrig's disease, is the most common type of motor neuron disease (Hardiman et al., 2017; Liu et al., 2018). With their etiology and mechanism largely unknown, there are no effective treatments to slow down the progression of neurodegenerative diseases so far (Dorst et al., 2018; Piton et al., 2018; de Bie et al., 2020). Patients get worse gradually and lose basic activities of daily living in the last stage. With enhancement in international collaboration and advancement in genomic sciences, especially large-scale genome-wide association studies (GWAS), genetic underpinnings of neurodegenerative diseases are being elucidated (Nicolas et al., 2018; Kunkle et al., 2019; Nalls et al., 2019; Roberts et al., 2020). Low back pain (LBP) is a common health condition with escalating healthcare utilization. In the last three decades, LBP has been the leading level-3 cause of years lived with disability (YLDs) globally, and particularly in high-income countries (Vos et al., 2012; Hoy et al., 2014). According to the most recent Global Burden of Disease Study (GBD, 2020), LBP was responsible for 780 YLDs per 100,000 population, and among 692 million non-communicable disease YLDs the proportion contributed by LBP was approximately 9.2%. LBP affects all age groups with a lifetime prevalence of about 40% (Manchikanti et al., 2014), which increases with aging and is slightly higher in women (Shmagel et al., 2016). Apart from behavioral and social-economic factors, the genetic basis of LBP has been well recognized in previous studies (Livshits et al., 2011; Junqueira et al., 2014; Suri et al., 2021).

Possible relationships between LBP and neurodegenerative diseases have been previously postulated (Broetz et al., 2007; Aggarwal et al., 2010; Miller et al., 2013; Udeh-Momoh et al., 2019; Silveira Barezani et al., 2020). In a prospective cohort of 690 participants at the preclinical stage of AD (Udeh-Momoh et al., 2019), back pain was among the most frequently occurring (3.0%) safety events, whereas in a recent cross-sectional study of 115 patients with sporadic PD (Silveira Barezani et al., 2020), 58.3% of participants reported to have LBP. A higher prevalence of back pain in PD patients (75/101, 74.3%) when compared with age-matched control patients (35/132, 26.5%) was reported in another prior study (Broetz et al., 2007). With regard to ALS, back pain was also among top safety concerns (8/32, 25%) in prior clinical trials (Aggarwal et al., 2010; Miller et al., 2013). Notably, these studies had limited sample size due to ethical and economic restrictions, and unmeasured confounding and reverse causation would incur biases to the findings as well. Meanwhile, established at parental gamete formation and insusceptible to later-life environmental confounders, genetic variants precede disease onset and hence are ideal epidemiological instruments. The last two decades have witnessed great strides in GWASs (Visscher et al., 2017), particularly increased samples and augmented power, and numerous single-nucleotide polymorphisms (SNPs) have been identified for common disorders, including self-reported back pain (Freidin et al., 2019) and chronic back pain (Suri et al., 2018). From the perspective of human genomics and genetic epidemiology, cutting-edge statistical tools such as linkage disequilibrium score regression (LDSC) (Bulik-Sullivan et al., 2015; Zheng et al., 2017) and Mendelian randomization (MR) (Hemani et al., 2018; Walker et al., 2019), have made it possible to use GWAS summary-level data to explore genetic correlation (Wang et al., 2020; Zhuang et al., 2021) and make causal inference (He et al., 2020; Zhang et al., 2020) within a wide spectrum of complex traits.

In this study, we utilized LDSC to investigate genetic correlations and further conducted two-sample bidirectional MR to explore relationships between LBP and three neurodegenerative diseases.

MATERIALS AND METHODS

Data Sources

This study was based on publicly available GWAS datasets, with informed consent from participants and approval by ethics committees completed in original studies (Nicolas et al., 2018; Kunkle et al., 2019; Nalls et al., 2019; FinnGen, 2021).

Summary association statistics for LBP was retrieved from the FinnGen study (FinnGen, 2021). LBP was defined as back pain localized between the costal margin and the inferior gluteal folds. From the Finnish registries of hospital discharge and cause of death, cases of LBP were ascertained using electronic health records with specific International Classification of Diseases (ICD) code (ICD-10, M54.5; ICD-9, 724.2; ICD-8, 728.7). Patients with symptoms of back pain caused by other specific diseases, such as fracture of lumbar vertebra (ICD-10, S32.0) and ankylosing spondylitis (ICD-10, M45), were excluded. Totally, there were 13,178 cases of LBP and 164,682 controls of the European ancestry (**Supplementary Table 1**). GWAS was performed in SAIGE, version 0.36.3.2 (Zhou et al., 2018), with sex, age, genotyping batches, and first 10 principal components incorporated as covariates. Variant positions which were initially presented in base pairs on build GRCh38 underwent coordinate conversion to GRCh37 using the command line tool *liftOver* and reference chain files from the UCSC Genome Browser Database (Lee et al., 2020). Effect size was reported in the unit of log-transformed odds ratio (OR) per additional copy of the alternative allele (**Supplementary Table 2**).

Summary-level GWAS data of three neurodegenerative diseases were from large-scale meta-analyses of AD (Kunkle et al., 2019), PD (Nalls et al., 2019), and ALS (Nicolas et al., 2018) in the European population. There were 21,982 clinically diagnosed cases and 41,944 controls in the GWAS of AD (Kunkle et al., 2019), 33,674 cases and 449,056 controls in the GWAS of PD (Kunkle et al., 2019), and 20,806 cases and 59,804 controls in the GWAS of ALS (Kunkle et al., 2019). More details of demographic information and case ascertainment were described in **Supplementary Materials** of original studies. GWAS meta-analyses were implemented using PLINK v1.90 (Purcell et al., 2007). Coordinates of SNPs according to the GRCh37 build were adopted; thus, no conversion was required. Likewise, the effect size represented change in log-OR of AD, PD, or ALS in the additive logistic regression (**Supplementary Table 3**).

Linkage Disequilibrium Score Regression

We used the common line tool *ldsc* v1.0.1 (Bulik-Sullivan et al., 2015) to compute heritability estimates and genetic correlations from summary-level statistics. Pre-calculated reference LD scores according to the 1000 Genomes EUR panel were adopted.[1] First, we filtered our data to keep HapMap3 SNPs (International HapMap 3 Consortium, Altshuler et al., 2010), using the recommended SNP list in the LD hub (Zheng et al., 2017). These variants had minor allele frequencies above 1% and were well-imputed in most European-ancestry GWASs, which benefited minimizing biases in the ensuing analyses. Variants at the MHC locus were not considered due to their great potential of pleiotropy and the complexity of local LD structure, which would affect the robustness of LDSC results. Those SNPs with large effect sizes ($\chi^2 > 80$) were filtered, since outliers could disproportionately influence the regression. Totally, 1,160,464 SNPs for LBP, 1,204,767 for AD, 1,120,769 for PD and 1,170,115 for ALS were retained. Heritability (H^2) on the observed scale, genomic inflation factor (λ_{GC}), mean chi-square (χ^2), and intercept statistics were derived from the SNP heritability analysis (command-line, –h2) for LBP and three neurodegenerative diseases. We divided the heritability estimate by its related standard error (SE) to calculate heritability *z*-scores. Suggested criteria (Zheng et al., 2017) to get reliable estimates of the genetic correlation were all met for LBP and three neurodegenerative diseases. The genetic correlation estimate (*Rg*) and its associated SE were computed with the −rg command flag. In the genetic correlation analysis, the *p*-value below the Bonferroni-corrected threshold ($p < 0.05/3 = 0.017$) was considered to be significant.

Mendelian Randomization

We performed bidirectional MR using the TwoSampleMR (version 0.5.6) package (Hemani et al., 2018) in R 3.6.3 (R Foundation for Statistical Computing, Vienna, Austria). First, instrumental SNPs robustly associated with traits of interest were selected. Using the default clumping threshold ($r^2 < 0.001$ within a 10,000 kb distance) in the MR-Base platform (Walker et al., 2019), we obtained 20, 23, and 6 SNPs associated with AD, PD, and ALS, respectively, reaching the significance threshold ($p < 5 \times 10^{-8}$). Regarding LBP, however, there were no genome-wide significant loci identified outside the MHC locus. Therefore, we relaxed the threshold ($p < 5 \times 10^{-6}$), as previous studies did (Schooling and Ng, 2019; Kwok et al., 2020; Ng and Schooling, 2020; Kwok and Schooling, 2021), to select 17 instrumental variants of LBP. For instrumental SNPs which were not present in the outcome datasets, we also searched for available proxies ($r^2 > 0.8$, 1000 Genomes EUR). We aligned effect alleles within each exposure–outcome pair, and the harmonized and merged datasets were utilized for subsequent analyses. As the primary MR analysis, we employed the inverse variance weighted (IVW) model to compute the overall estimate (Burgess et al., 2013). The weighted median approach would provide robust estimates on the assumption that more than 50% of weights came from valid instruments (Bowden et al., 2016). MR-Egger regression was capable of examining unbalanced horizontal pleiotropy *via* the intercept and provided causal estimate with adjustment for pleiotropy *via* the regression slope (Burgess and Thompson, 2017). The weighted mode-based method would obtain a robust overall causal estimate when the majority of similar individual estimates were from valid instrumental SNPs (Hartwig et al., 2017). Nevertheless, the weighted median, MR-Egger, and weighted mode estimates had compromised power (Slob and Burgess, 2020), as indicated by wide confidence intervals (CIs), and hence were performed as complimentary methods. As for MR results, ORs represented the relative odds of the occurrence of the outcome concerned (i.e., AD) given exposure to the trait of interest (i.e., LBP). The power calculation was performed using a web application, mRnd (Brion et al., 2013). We estimated the proportion of variance explained by instrumental SNPs for the exposure using the formula $2 \times \text{EAF} \times (1\text{-EAF}) \times \text{Beta}^2$, where EAF is the effect allele frequency and Beta denotes the effect size. Then, assuming a power of 80% and an alpha of 5%, we calculated the detectable range of OR with sufficient power for the outcome of interest. The significance threshold was set at $p < 0.05/6 = 0.008$ after applying Bonferroni correction for multiple MR tests.

RESULTS

Heritability Estimates and Genetic Correlations

Common SNPs (∼1.1 million, EUR phase 3 HapMap) cumulatively explained 1.86% of the total heritability of LBP, suggesting the small effects of SNPs in the genetic contribution to complex disorders. In the GWAS of LBP, genomic inflation factor ($\lambda_{GC} = 1.096$) demonstrated slight inflation; with the intercept (1.035) being close to 1, the inflation should be attributed to the polygenic genetic architecture. As shown in **Table 1**, the heritability estimate on the observed scale, genomic inflation factor, and LDSC intercept for AD, PD, and ALS in this study were similar to those in the original GWASs. Moreover, all these statistics satisfied the following criteria, heritability $H^2/\text{SE} > 4$, mean $\chi^2 > 1.02$ and intercepts between 0.9 and 1.1, indicating the suitability and reliability for estimating genetic correlations.

There was no evidence for the genetic correlation between LBP and AD ($Rg = -0.033$, $p = 0.766$). As detailed in **Table 2**, correlations between LBP and PD ($Rg = -0.079$, $p = 0.279$) and ALS ($Rg = 0.069$, $p = 0.583$) did not reach nominal significance, either.

Bidirectional MR Analyses

Overall, MR estimates suggested that genetically predicted higher risks of LBP were not associated with the liability to AD, PD, or ALS. By the IVW approach, genetically predicted

[1] https://data.broadinstitute.org/alkesgroup/LDSCORE/

TABLE 1 | Heritability estimates based on single-nucleotide polymorphisms for low back pain and three neurodegenerative diseases.

Traits	H^2(SE)	λ_{GC}	Mean χ^2	Intercept (SE)
Low back pain	1.86% (0.32%)	1.096	1.102	1.035 (0.008)
Alzheimer's disease	7.13% (1.14%)	1.093	1.118	1.030 (0.008)
Parkinson's disease	1.85% (0.18%)	1.090	1.137	0.985 (0.007)
Amyotrophic lateral sclerosis	3.17% (0.70%)	1.044	1.071	1.020 (0.007)

H^2, heritability estimate on the observed scale; SE, standard error; λ_{GC}, genomic inflation factor.

TABLE 2 | Genetic correlations between low back pain and three neurodegenerative diseases.

Phenotypes	R_g (95% CI)	*p*-value
Alzheimer's disease	−0.033 (−0.252, 0.186)	0.766
Parkinson's disease	−0.079 (−0.223, 0.064)	0.279
Amyotrophic lateral sclerosis	0.069 (−0.177, 0.315)	0.583

R_g, genetic correlation estimate; CI, confidence interval.

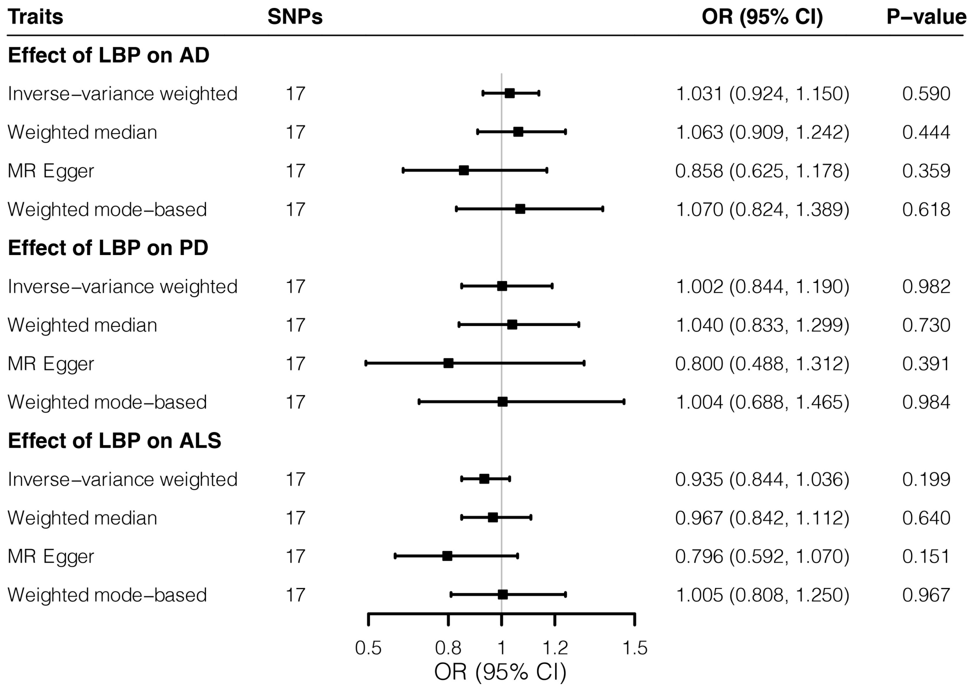
predisposition to LBP was not associated with the risk of AD (*OR* = 1.031; 95% CI, 0.924–1.150; *p* = 0.590). Likewise, causal effects of LBP on PD (*OR* = 1.002; 95% CI, 0.844–1.190; *p* = 0.982) and ALS (*OR* = 0.935; 95% CI, 0.844–1.036; *p* = 0.199) did not reach significance threshold in the main analysis. Complementary MR methods provided consistent results (**Figure 1** and **Supplementary Figure 1**). Notably, our analysis might be underpowered (**Supplementary Table 4**) to detect small causal effects given the small proportion of variance explained by instrumental SNPs. No unbalanced horizontal pleiotropy (all $p > 0.05$) was indicated by MR-Egger regression intercepts (**Supplementary Table 5**). Cochran's *Q* tests provided no evidence for the existence of heterogeneity (**Supplementary Table 6**), whereas leave-one-out plots (**Supplementary Figure 2**) did not identify any outlier variants.

In the reverse direction, MR analyses did not support the effects of neurodegenerative diseases on LBP. A one-unit increase in log-OR of AD was not associated with change in risks of LBP (*OR* = 0.963; 95% CI, 0.923–1.006; *p* = 0.090) by the IVW method, whereas the weighted median estimate reached nominal significance, albeit failing the Bonferroni-corrected threshold ($p = 0.009 > 0.05/6$). Similarly, as shown in **Figure 2**, the relationship between PD and LBP (*OR* = 0.960; 95% CI, 0.922–1.000; *p* = 0.048) reached nominal significance. However, there was no evidence for the association of ALS with LBP (*OR* = 1.030; 95% CI, 0.935–1.135; *p* = 0.545). According to scatter plots (**Supplementary Figure 3**) and leave-one-out plots (**Supplementary Figure 4**), no evident outliers existed, while additional analyses (**Supplementary Tables 5, 6**) demonstrated no horizontal pleiotropy or heterogeneity.

DISCUSSION

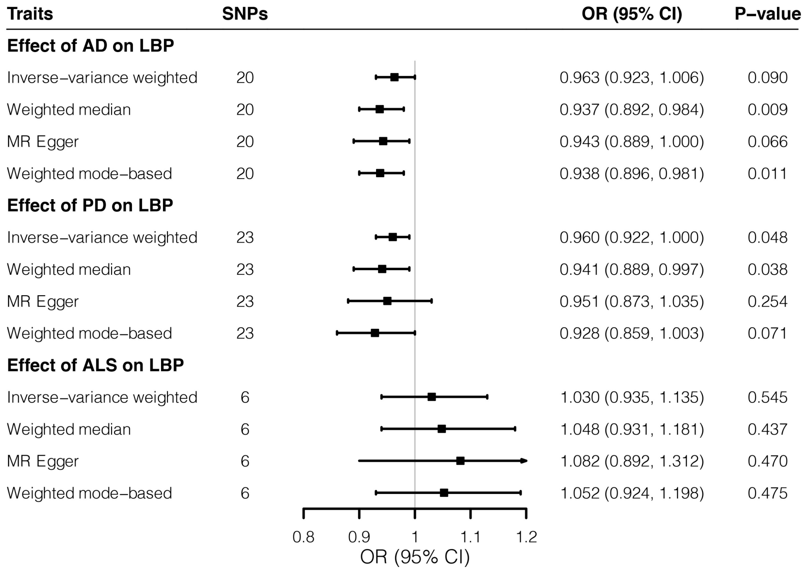
In this study, we did not find evidence supporting genetic correlations or causations between non-specific LBP and three common neurodegenerative diseases. Back pain has been commonly studied as a self-reported symptom (Suri et al., 2018; Freidin et al., 2019) and studied in spine-related diseases like lumbar spinal stenosis (Suri et al., 2021). For example, a previous GWAS (Freidin et al., 2019) of self-reported back pain in 509,000 Europeans identified three significant loci ($p < 5 \times 10^{-8}$), but genetic correlation estimates between back pain and AD ($R_g = 0.115$, $p = 0.147$), PD ($R_g = 0.029$, $p = 0.586$), and ALS ($R_g = 0.166$, $p = 0.030$) all failed Bonferroni-corrected significance. Notably, only a small part of LBP has clear causes and can be classified into specific diseases; however, there exists the majority with unknown mechanisms. Such LBP has been seen as an entity itself in the electronic health record, and as a complex trait, GWAS and related tools are likely to be powerful to disentangle the genetic underpinnings. Here, we employed LDSC and MR to elucidate their relationships based on biobank association data of LBP and the most up-to-date GWASs of AD, PD, and ALS.

Observational studies exploring the relationship between LBP and neurodegenerative diseases have been conducted before (Broetz et al., 2007; Aggarwal et al., 2010; Miller et al., 2013; Udeh-Momoh et al., 2019; Silveira Barezani et al., 2020). Several studies reported a high occurrence of LBP during the non-interventional course of AD (Udeh-Momoh et al., 2019), and the interventional diagnostic and therapeutic procedure of AD (Landen et al., 2013; Alcolea et al., 2014). Similarly, LBP was a common complaint during the treatment of ALS (Aggarwal et al., 2010; Miller et al., 2013). We could not tell whether there are causal mechanisms underlying such findings, given the complexity of insufficiently controlled factors in traditional epidemiology. Regarding the potential role of LBP in PD, in a recent questionnaire-based study (Silveira Barezani et al., 2020), about 40% patients reported the onset of LBP before the diagnosis of PD, and higher pain scores were associated with more advanced stage and rating scales of PD. The interaction of LBP and PD undoubtedly leads to more difficulty and disability in daily activities. Besides, both PD and ALS extensively involved neural and musculoskeletal systems with a variety of manifestations and unbalanced musculoskeletal dynamics due to gait abnormality, posture alteration and chronic joint trauma in the progressive course were likely to result in LBP (Ozturk and Kocer, 2018; Duncan et al., 2019). The vicious cycle of LBP and neurodegenerative diseases should have a severe influence on the life quality of patients. Identifying possible links underlying LBP, AD, PD, and ALS from the perspective of genetic correlations would provide more informative knowledge and ultimately benefit in developing effective interventions. In this study, we found no evidence for the causal effects of LBP on neurodegenerative diseases, neither did the reverse effects reach Bonferroni-corrected threshold ($p < 0.05/6 = 0.008$). The effects of AD and PD on LBP reached nominal significance, and interestingly, the genetic predisposition to AD and PD seemed to be associated with the lower occurrence of LBP in this study. The findings failed to agree with previous observational studies and were against common intuition to a certain extent. Notably, it may as well be common sense that more environmental components (i.e., sedentary behaviors)



Traits	SNPs	OR (95% CI)	P–value
Effect of LBP on AD			
Inverse–variance weighted	17	1.031 (0.924, 1.150)	0.590
Weighted median	17	1.063 (0.909, 1.242)	0.444
MR Egger	17	0.858 (0.625, 1.178)	0.359
Weighted mode–based	17	1.070 (0.824, 1.389)	0.618
Effect of LBP on PD			
Inverse–variance weighted	17	1.002 (0.844, 1.190)	0.982
Weighted median	17	1.040 (0.833, 1.299)	0.730
MR Egger	17	0.800 (0.488, 1.312)	0.391
Weighted mode–based	17	1.004 (0.688, 1.465)	0.984
Effect of LBP on ALS			
Inverse–variance weighted	17	0.935 (0.844, 1.036)	0.199
Weighted median	17	0.967 (0.842, 1.112)	0.640
MR Egger	17	0.796 (0.592, 1.070)	0.151
Weighted mode–based	17	1.005 (0.808, 1.250)	0.967

FIGURE 1 | Effects of low back pain on three neurodegenerative diseases by Mendelian randomization analyses. Relative odds of the occurrence of three neurodegenerative diseases given exposure to low back pain were generated by three Mendelian randomization methods and presented in forest plots. AD, Alzheimer's disease; CI, confidence interval; ALS, amyotrophic lateral sclerosis; LBP, low back pain; OR, odds ratio; PD, Parkinson's disease; SNP, Single-nucleotide polymorphism.



Traits	SNPs	OR (95% CI)	P–value
Effect of AD on LBP			
Inverse–variance weighted	20	0.963 (0.923, 1.006)	0.090
Weighted median	20	0.937 (0.892, 0.984)	0.009
MR Egger	20	0.943 (0.889, 1.000)	0.066
Weighted mode–based	20	0.938 (0.896, 0.981)	0.011
Effect of PD on LBP			
Inverse–variance weighted	23	0.960 (0.922, 1.000)	0.048
Weighted median	23	0.941 (0.889, 0.997)	0.038
MR Egger	23	0.951 (0.873, 1.035)	0.254
Weighted mode–based	23	0.928 (0.859, 1.003)	0.071
Effect of ALS on LBP			
Inverse–variance weighted	6	1.030 (0.935, 1.135)	0.545
Weighted median	6	1.048 (0.931, 1.181)	0.437
MR Egger	6	1.082 (0.892, 1.312)	0.470
Weighted mode–based	6	1.052 (0.924, 1.198)	0.475

FIGURE 2 | Effects of three neurodegenerative diseases on low back pain by Mendelian randomization analyses. Relative odds of the occurrence of low back pain given exposure to three neurodegenerative diseases were generated by three Mendelian randomization methods and presented in forest plots. AD, Alzheimer's disease; CI, confidence interval; ALS, amyotrophic lateral sclerosis; LBP, low back pain; OR, odds ratio; PD, Parkinson's disease; SNP, Single-nucleotide polymorphism.

rather than genetic underpinnings would underlie the liability to LBP when compared with neurodegenerative diseases. In the current statistical model of MR, however, both the exposures and outcomes of interest were genetically predicted "ideal" traits, which were proxied by common variants without taking account of other factors. Undoubtedly, MR estimates alone were not enough. Triangulating evidence across multiple lines of studies is necessary to shed light on relationships between complex traits.

The major strength of this study was the utilization of the state-of-the-art tools, LDSC and MR, to explore the relationships between complex disorders. Using millions of summary-level statistics from hundreds of thousands of participants, LDSC was a powerful tool to estimate the genetic correlation. Based on a subset of instrumental SNPs strongly associated with the exposure-trait of interest, MR was capable of estimating the causal effect on the outcome-trait concerned, while circumventing reverse causation and minimizing biases by confounders. There were also several limitations. Firstly, LBP was defined by electronic health record codes with more reliability and less misclassification, but we could not tell whether the relationship existed between chronic LBP and neurodegenerative diseases. LBP was studied as a whole, without separating the acute and chronic type as generally included in self-reported questionnaires. Neither did we differentiate between subgroups of neurodegenerative diseases like AD subtypes based on neuropathology and neuroimaging, PD subtypes by age at onset (i.e., early-onset and late-onset), and ALS subgroups classified by site of onset (i.e., bulbar and spinal). Secondly, gender differences in the prevalence of LBP and three neurodegenerative diseases have been proposed; however, we could not address the meaningful question since no sex-stratified association data were available. Lastly, this study was based on datasets from European-ancestry GWASs, and great attention should be paid when generalizing the findings to the other populations.

In summary, our results provided no evidence for the genetic correlations between LBP and three common neurodegenerative diseases, AD, PD, and ALS.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

PW, BW, and KX conceptualized the study. PW, BD, RY, and WZ contributed to acquisition and analysis of data and validation and visualization of results. PW, BD, XL, and YD took part in drafting and reviewing the main manuscript. BW and KX played a role in project administration and funding acquisition. All authors contributed to the article and approved the final version of the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (81730036) and Key R&D Program of Hunan Province (2021SK1010) to KX. PW have received a visiting Ph.D. scholarship (201906370155) from the China Scholarship Council.

ACKNOWLEDGMENTS

We acknowledge the participants and investigators of the FinnGen study. We are grateful to them for making summary-level association statistics of low back pain possible and accessible. We also gratefully thank Kunkle B. W., Nicolas A., Nalls M. A., et al., and all investigators and related consortia for conducting studies of Alzheimer's disease, Parkinson's disease, and amyotrophic lateral sclerosis and sharing GWAS summary statistics. We also acknowledge the participants of these studies.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.744299/full#supplementary-material

REFERENCES

Aggarwal, S. P., Zinman, L., Simpson, E., McKinley, J., Jackson, K. E., Pinto, H., et al. (2010). Safety and efficacy of lithium in combination with riluzole for treatment of amyotrophic lateral sclerosis: a randomised, double-blind, placebo-controlled trial. *Lancet Neurol.* 9, 481–488. doi: 10.1016/s1474-4422(10)70068-5

Alcolea, D., Martínez-Lage, P., Izagirre, A., Clerigué, M., Carmona-Iragui, M., Alvarez, R. M., et al. (2014). Feasibility of lumbar puncture in the study of cerebrospinal fluid biomarkers for Alzheimer's disease: a multicenter study in Spain. *J. Alzheimers Dis.* 39, 719–726. doi: 10.3233/jad-131334

Balestrino, R., and Schapira, A. H. V. (2020). Parkinson disease. *Eur. J. Neurol.* 27, 27–42. doi: 10.1111/ene.14108

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* 40, 304–314. doi: 10.1002/gepi.21965

Brion, M. J., Shakhbazov, K., and Visscher, P. M. (2013). Calculating statistical power in Mendelian randomization studies. *Int. J. Epidemiol.* 42, 1497–1501. doi: 10.1093/ije/dyt179

Broetz, D., Eichner, M., Gasser, T., Weller, M., and Steinbach, J. P. (2007). Radicular and nonradicular back pain in Parkinson's disease: a controlled study. *Mov. Disord.* 22, 853–856. doi: 10.1002/mds.21439

Bulik-Sullivan, B. K., Loh, P. R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, et al. (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* 47, 291–295. doi: 10.1038/ng.3211

Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* 37, 658–665. doi: 10.1002/gepi.21758

Burgess, S., and Thompson, S. G. (2017). Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur. J. Epidemiol.* 32, 377–389. doi: 10.1007/s10654-017-0255-x

de Bie, R. M. A., Clarke, C. E., Espay, A. J., Fox, S. H., and Lang, A. E. (2020). Initiation of pharmacological therapy in Parkinson's disease: when, why, and how. *Lancet Neurol.* 19, 452–461. doi: 10.1016/S1474-4422(20)30036-3

Dorst, J., Ludolph, A. C., and Huebers, A. (2018). Disease-modifying and symptomatic treatment of amyotrophic lateral sclerosis. *Ther. Adv. Neurol. Disord.* 11:1756285617734734. doi: 10.1177/1756285617734734

Duncan, R. P., Van Dillen, L. R., Garbutt, J. M., Earhart, G. M., and Perlmutter, J. S. (2019). Low back pain–related disability in Parkinson disease: impact on functional mobility, physical activity, and quality of life. *Phys. Ther.* 99, 1346–1353. doi: 10.1093/ptj/pzz094

FinnGen. (2021). *FinnGen Documentation of R5 Release*. Available online at: https: //www.finngen.fi/en/access_results (accessed May 24, 2021).

Freidin, M. B., Tsepilov, Y. A., Palmer, M., Karssen, L. C., Suri, P., Aulchenko, Y. S., et al. (2019). Insight into the genetic architecture of back pain and its risk factors from a study of 509,000 individuals. *Pain* 160, 1361–1373. doi: 10.1097/j.pain.0000000000001514

GBD (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet* 396, 1204–1222. doi: 10.1016/S0140-6736(20)30925-9

Hardiman, O., Al-Chalabi, A., Chio, A., Corr, E. M., Logroscino, G., Robberecht, W., et al. (2017). Amyotrophic lateral sclerosis. *Nat. Rev. Dis. Primers* 3:17071. doi: 10.1038/nrdp.2017.71

Hartwig, F. P., Davey Smith, G., and Bowden, J. (2017). Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int. J. Epidemiol.* 46, 1985–1998. doi: 10.1093/ije/dyx102

He, Y., Zhang, H., Wang, T., Han, Z., Ni, Q. B., Wang, K., et al. (2020). Impact of serum calcium levels on alzheimer's disease: a Mendelian randomization study. *J. Alzheimers Dis.* 76, 713–724. doi: 10.3233/JAD-191249

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., et al. (2018). The MR-Base platform supports systematic causal inference across the human phenome. *Elife* 7:e34408. doi: 10.7554/eLife.34408

Hoy, D., March, L., Brooks, P., Blyth, F., Woolf, A., Bain, C., et al. (2014). The global burden of low back pain: estimates from the Global Burden of Disease 2010 study. *Ann. Rheum. Dis.* 73, 968–974. doi: 10.1136/annrheumdis-2013-204428

International HapMap 3 Consortium, Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., et al. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* 467, 52–58. doi: 10.1038/nature09298

Junqueira, D. R., Ferreira, M. L., Refshauge, K., Maher, C. G., Hopper, J. L., Hancock, M., et al. (2014). Heritability and lifestyle factors in chronic low back pain: results of the Australian twin low back pain study (The AUTBACK study). *Eur. J. Pain* 18, 1410–1418. doi: 10.1002/ejp.506

Kunkle, B. W., Grenier-Boley, B., Sims, R., Bis, J. C., Damotte, V., Naj, A. C., et al. (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Abeta, tau, immunity and lipid processing. *Nat. Genet.* 51, 414–430. doi: 10.1038/s41588-019-0358-2

Kwok, M. K., Kawachi, I., Rehkopf, D., and Schooling, C. M. (2020). The role of cortisol in ischemic heart disease, ischemic stroke, type 2 diabetes, and cardiovascular disease risk factors: a bi-directional Mendelian randomization study. *BMC Med.* 18:363. doi: 10.1186/s12916-020-01831-3

Kwok, M. K., and Schooling, C. M. (2021). Herpes simplex virus and Alzheimer's disease: a Mendelian randomization study. *Neurobiol. Aging* 99, 101.e11–101.e13. doi: 10.1016/j.neurobiolaging.2020.09.025

Landen, J. W., Zhao, Q., Cohen, S., Borrie, M., Woodward, M., Billing, C. B. Jr., et al. (2013). Safety and pharmacology of a single intravenous dose of ponezumab in subjects with mild-to-moderate Alzheimer disease: a phase I, randomized, placebo-controlled, double-blind, dose-escalation study. *Clin. Neuropharmacol.* 36, 14–23. doi: 10.1097/WNF.0b013e31827db49b

Lee, C. M., Barber, G. P., Casper, J., Clawson, H., Diekhans, M., Gonzalez, J. N., et al. (2020). UCSC genome browser enters 20th year. *Nucleic Acids Res.* 48, D756–D761. doi: 10.1093/nar/gkz1012

Liu, X., He, J., Gao, F. B., Gitler, A. D., and Fan, D. (2018). The epidemiology and genetics of Amyotrophic lateral sclerosis in China. *Brain Res.* 1693, 121–126. doi: 10.1016/j.brainres.2018.02.035

Livshits, G., Popham, M., Malkin, I., Sambrook, P. N., Macgregor, A. J., Spector, T., et al. (2011). Lumbar disc degeneration and genetic factors are the main risk factors for low back pain in women: the UK twin spine study. *Ann. Rheum. Dis.* 70, 1740–1745. doi: 10.1136/ard.2010.137836

Manchikanti, L., Singh, V., Falco, F. J., Benyamin, R. M., and Hirsch, J. A. (2014). Epidemiology of low back pain in adults. *Neuromodulation* 17(Suppl. 2), 3–10. doi: 10.1111/ner.12018

Miller, T. M., Pestronk, A., David, W., Rothstein, J., Simpson, E., Appel, S. H., et al. (2013). An antisense oligonucleotide against SOD1 delivered intrathecally for patients with SOD1 familial amyotrophic lateral sclerosis: a phase 1, randomised, first-in-man study. *Lancet Neurol.* 12, 435–442. doi: 10.1016/s1474-4422(13)70061-9

Nalls, M. A., Blauwendraat, C., Vallerga, C. L., Heilbron, K., Bandres-Ciga, S., Chang, D., et al. (2019). Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* 18, 1091–1102. doi: 10.1016/S1474-4422(19)30320-5

Ng, J. C. M., and Schooling, C. M. (2020). Effect of Glucagon on ischemic heart disease and its risk factors: a Mendelian randomization study. *J. Clin. Endocrinol. Metab.* 105:dgaa259. doi: 10.1210/clinem/dgaa259

Nicolas, A., Kenna, K. P., Renton, A. E., Ticozzi, N., Faghri, F., Chia, R., et al. (2018). Genome-wide analyses identify KIF5A as a novel ALS gene. *Neuron* 97, 1268–1283e1266. doi: 10.1016/j.neuron.2018.02.027 1268-1283 e1266

Ozturk, E. A., and Kocer, B. G. (2018). Predictive risk factors for chronic low back pain in Parkinson's disease. *Clin. Neurol. Neurosurg.* 164, 190–195. doi: 10.1016/j.clineuro.2017.12.011

Piton, M., Hirtz, C., Desmetz, C., Milhau, J., Lajoix, A. D., Bennys, K., et al. (2018). Alzheimer's disease: advances in drug development. *J. Alzheimers Dis.* 65, 3–13. doi: 10.3233/JAD-180145

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Roberts, J. S., Patterson, A. K., and Uhlmann, W. R. (2020). Genetic testing for neurodegenerative diseases: ethical and health communication challenges. *Neurobiol. Dis.* 141:104871. doi: 10.1016/j.nbd.2020.104871

Schooling, C. M., and Ng, J. C. M. (2019). Reproduction and longevity: a Mendelian randomization study of gonadotropin-releasing hormone and ischemic heart disease. *SSM Popul. Health* 8:100411. doi: 10.1016/j.ssmph.2019.100411

Shmagel, A., Foley, R., and Ibrahim, H. (2016). Epidemiology of chronic low back pain in US adults: data from the 2009-2010 national health and nutrition examination survey. *Arthritis Care Res. (Hoboken)* 68, 1688–1694. doi: 10.1002/acr.22890

Silveira Barezani, A. L., de Figueiredo Feital, A. M. B., Gonçalves, B. M., Christo, P. P., and Scalzo, P. L. (2020). Low back pain in Parkinson's disease: a cross-sectional study of its prevalence, and implications on functional capacity and quality of life. *Clin. Neurol. Neurosurg.* 194:105787. doi: 10.1016/j.clineuro.2020.105787

Slob, E. A. W., and Burgess, S. (2020). A comparison of robust Mendelian randomization methods using summary data. *Genet. Epidemiol.* 44, 313–329. doi: 10.1002/gepi.22295

Suri, P., Palmer, M. R., Tsepilov, Y. A., Freidin, M. B., Boer, C. G., Yau, M. S., et al. (2018). Genome-wide meta-analysis of 158,000 individuals of European ancestry identifies three loci associated with chronic back pain. *PLoS Genet.* 14:e1007601. doi: 10.1371/journal.pgen.1007601

Suri, P., Stanaway, I. B., Zhang, Y., Freidin, M. B., Tsepilov, Y. A., Carrell, D. S., et al. (2021). Genome-wide association studies of low back pain and lumbar spinal disorders using electronic health record data identify a locus associated with lumbar spinal stenosis. *Pain* 162, 2263–2272. doi: 10.1097/j.pain.0000000000002221

Udeh-Momoh, C., Price, G., Ropacki, M. T., Ketter, N., Andrews, T., Arrighi, H. M., et al. (2019). Prospective evaluation of cognitive health and related factors in elderly at risk for developing Alzheimer's dementia: a longitudinal cohort study. *J. Prev. Alzheimers Dis.* 6, 256–266. doi: 10.14283/jpad.2019.31

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., et al. (2017). 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* 101, 5–22. doi: 10.1016/j.ajhg.2017.06.005

Vos, T., Flaxman, A. D., Naghavi, M., Lozano, R., Michaud, C., Ezzati, M., et al. (2012). Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 380, 2163–2196. doi: 10.1016/S0140-6736(12)61729-2

Walker, V. M., Davies, N. M., Hemani, G., Zheng, J., Haycock, P. C., Gaunt, T. R., et al. (2019). Using the MR-Base platform to investigate risk factors and drug targets for thousands of phenotypes. *Wellcome Open Res.* 4:113. doi: 10.12688/wellcomeopenres.15334.2

Wang, X., Jia, J., and Huang, T. (2020). Shared genetic architecture and casual relationship between leptin levels and type 2 diabetes: large-scale cross-trait meta-analysis and Mendelian randomization analysis. *BMJ Open Diabetes Res. Care* 8:e001140. doi: 10.1136/bmjdrc-2019-001140

Weller, J., and Budson, A. (2018). Current understanding of Alzheimer's disease diagnosis and treatment. *F1000Res* 7:F1000FacultyRev–1161. doi: 10.12688/f1000research.14506.1

Zhang, H., Wang, T., Han, Z., and Liu, G. (2020). Mendelian randomization study to evaluate the effects of interleukin-6 signaling on four neurodegenerative diseases. *Neurol. Sci.* 41, 2875–2882. doi: 10.1007/s10072-020-04381-x

Zheng, J., Erzurumluoglu, A. M., Elsworth, B. L., Kemp, J. P., Howe, L., Haycock, P. C., et al. (2017). LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* 33, 272–279. doi: 10.1093/bioinformatics/btw613

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* 50, 1335–1341. doi: 10.1038/s41588-018-0184-y

Zhuang, Z., Yao, M., Wong, J. Y. Y., Liu, Z., and Huang, T. (2021). Shared genetic etiology and causality between body fat percentage and cardiovascular diseases: a large-scale genome-wide cross-trait analysis. *BMC Med.* 19:100. doi: 10.1186/s12916-021-01972-z

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

SYSTEMATIC REVIEW
published: 27 September 2021
doi: 10.3389/fgene.2021.757601

The 10-Repeat 3′-UTR VNTR Polymorphism in the *SLC6A3* Gene May Confer Protection Against Parkinson's Disease: A Meta-analysis

Qiaoli Zeng[1,2†], Fan Ning[1,3†], Shanshan Gu[1,3†], Qiaodi Zeng[4], Riling Chen[1,5,2], Liuquan Peng[5], Dehua Zou[1,2], Guoda Ma[1]* and Yajun Wang[6]**

[1]Maternal and Children's Health Research Institute, Shunde Women and Children's Hospital, Guangdong Medical University, Foshan, China, [2]Key Laboratory of Research in Maternal and Child Medicine and Birth Defects, Guangdong Medical University, Foshan, China, [3]Institute of Neurology, Affiliated Hospital of Guangdong Medical University, Zhanjiang, China, [4]Department of Clinical Laboratory, People's Hospital of Haiyuan County, Zhongwei, China, [5]Department of Pediatrics, Shunde Women and Children's Hospital, Guangdong Medical University, Foshan, China, [6]Institute of Respiratory, Shunde Women and Children's Hospital, Guangdong Medical University, Foshan, China

OPEN ACCESS

Edited by:
Guiyou Liu,
Tianjin Institute of Industrial Biotechnology (CAS), China

Reviewed by:
Nicholas B. Larson,
Mayo Clinic, United States
Zhijie Han,
Chongqing Medical University, China

****Correspondence:***
Dehua Zou
757384895@qq.com
Guoda Ma
sihan1107@126.com
Yajun Wang
wangyajuny1977@aliyun.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *12 August 2021*
Accepted: *14 September 2021*
Published: *27 September 2021*

Citation:
Zeng Q, Ning F, Gu S, Zeng Q, Chen R, Peng L, Zou D, Ma G and Wang Y (2021) The 10-Repeat 3′-UTR VNTR Polymorphism in the SLC6A3 Gene May Confer Protection Against Parkinson's Disease: A Meta-analysis. Front. Genet. 12:757601. doi: 10.3389/fgene.2021.757601

The dopamine transporter (DAT) is encoded by the SLC6A3 gene and plays an important role in the regulation of the neurotransmitter dopamine. The SLC6A3 gene contains several repetition alleles (3–11 repeats) of a 40-base pair variable number of tandem repeats (VNTR) in the 3′-untranslated region (3′-UTR), which may affect DAT expression levels. The 10-repeat (10R) allele could play a protective role against PD. However, inconsistent findings have been reported.

Methods: A comprehensive meta-analysis was performed to accurately estimate the association between the 10R allele of the 3′-UTR VNTR in SLC6A3 and PD among four different genetic models.

Results: This meta-analysis included a total of 3,142 patients and 3,496 controls. We observed a significant difference between patients and controls for the allele model (10R vs. all others: OR = 0.860, 95% CI: 0.771–0.958, P = 0.006), pseudodominant model (10R/10R + 10R/9R vs. all others: OR = 0.781, 95% CI: 0.641–0.952, P = 0.014) and pseudorecessive model (10R/10R vs. all others: OR = 0.858, 95% CI: 0.760–0.969, P = 0.013) using a fixed effects model. No significant differences were observed under the pseudocodominant model (10R/9R vs. all others: OR = 1.079, 95% CI: 0.945–1.233, P = 0.262). By subgroup analysis, the 10R, 10R/10R and 10R/9R genotypes were found to be significantly different from PD in Asian populations.

Conclusion: Our findings suggest that the *SLC6A3* 10R may be a protective factor in susceptibility to PD.

Keywords: Parkinson's disease, Slc6a3, dopamine transporter, variable number of tandem repeats, meta-analysis

INTRODUCTION

Parkinson's disease (PD) is a very common neurodegenerative disorder. One of the critical neuropathologies of PD is the degeneration of dopamine-producing neurons in the substantia nigra, resulting in impairment of the dopaminergic pathway and the subsequent depletion of dopamine levels (Balestrino and Schapira 2020). Another pathologic hallmark is the presence of ubiquitinated protein deposits named Lewy bodies, which cause dopaminergic cell death (Balestrino and Schapira, 2020; Singleton et al., 2003). These pathologic changes result in depletion of dopamine levels that underlie the etiology of PD. Therefore, dopaminergic transmission and metabolism pathway genes have been investigated and are considered to be candidate genes for PD.

The dopamine transporter (DAT) plays an important role in dopaminergic neurotransmission. It is mainly present on the terminals of neurons in the substantia nigra and is responsible for controlling the duration and intensity of neurotransmission by rapid dopamine uptake into the presynaptic terminals; thus, DAT is critical in the temporal and spatial buffering of released dopamine and its recycling (Cheng and Bahar 2015; Uhl 2003; Bannon et al., 2001). In addition, DAT is considered a gateway for neurotoxicants because the nigrostriatal toxicant 1-methyl-4-phenylpyridinium (Mpp$^+$) is taken up selectively by presynaptic DAT, and access to dopaminergic neurons leads to dopaminergic cell toxicity (Krontiris 1995; Uhl et al., 1994); DAT has also been shown to interact with alpha-synuclein (a kind of Lewy body) (Lee et al., 2001; Wersinger et al., 2003; Thomas and Beal 2007). These findings provide evidence for a role of DAT in PD and seem to explain why the density of DAT correlates with the extent of dopaminergic cell loss in PD brains.

DAT is coded by the *SLC6A3* gene. The 3′-UTR of the *SLC6A3* gene includes a 40 bp variable number tandem repeat (VNTR) polymorphism. Between 3 and 11 copies of the 40 bp VNTR have been identified in normal populations (Uhl 2003), and the 9 and 10 repeat alleles are most frequent in both PD patients and several populations (Uhl 2003; Bannon et al., 2001). The *SLC6A3* VNTR itself seems to be a functional polymorphism. A recent study found that the seed region of miR-491 is located in the VNTR fragment of the DAT mRNA e 3′-UTR, and the effect of miR-491 on DAT expression is dependent on the VNTR copynumber (Jia et al., 2016). Thus, *SLC6A3* polymorphic VNTRs may directly influence DAT expression (Fuke et al., 2001; Heinz et al., 2002; Mill et al., 2002; Lynch et al., 2003). Lin reported that the 10R alleles conferred protection against PD compared to other alleles (Lin et al., 2003), but other noteworthy studies showed different results. Given these controversial conclusions, we performed a meta-analysis to systematically, quantitatively, and objectively summarize the association between the 10R of the 3′-UTR VNTR in *SLC6A3* and PD susceptibility.

MATERIALS AND METHODS

Literature Search

The PubMed, Google Scholar, and Chinese National Knowledge Infrastructure databases were systematically searched for potentially qualified studies using a combination of the keywords "dopamine transporter," "DAT," "DAT1," "*SLC6A3*," "VNTR," "3′-UTR," "polymorphism," "rs28363170" and "Parkinson." with no language or date restrictions. All studies were evaluated on the basis of the title and abstract and we excluded studies that were clearly irrelevant. Then, potentially eligible studies were reviewed in full to determine the inclusion in the meta-analysis.

Inclusion and Exclusion Criteria

Studies included in the meta-analysis had to meet all the following inclusion criteria: 1) case-control or cohort studies

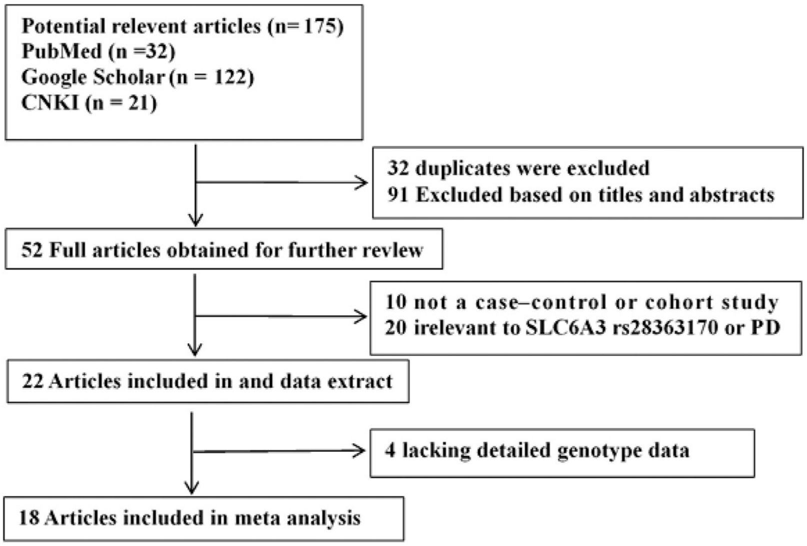

Potential relevent articles (n= 175)
PubMed (n =32)
Google Scholar (n = 122)
CNKI (n = 21)
32 duplicates were excluded
91 Excluded based on titles and abstracts
52 Full articles obtained for further revlew
10 not a case–control or cohort study
20 irelevant to SLC6A3 rs28363170 or PD
22 Articles included in and data extract
4 lacking detailed genotype data
18 Articles included in meta analysis

FIGURE 1 | Flow diagram of the literature search and selection.

TABLE 1 | Characteristics of each study included in this meta-analysis.

Author	Year	Ethnic	Case/Control	Allele distribution						Genotype distribution							
				Cases, n			Control, n			Cases, n				Control, n			
				9R	10R	Other	9R	10R	Other	9R/9R	9R/10R	10R/10R	Other	9R/9R	9R/10R	10R/10R	Other
Chang et al.	2018	Chinese	52/60	19	85	0	23	97	0	5	9	38	0	6	11	43	0
Lu et al.	2016	Chinese	521/502	76	966	0	66	938	0	6	64	451	0	6	54	442	0
BENITEZ et al.	2010	South American	99/131	37	161	0	59	203	0	3	31	65	0	5	49	77	0
Ritz et al.	2009	Latino, Asian, and Native American	324/334	—	—	—	—	—	—	28	113	179	4	16	109	200	9
Kelada et al.	2005	non-Hispanic Caucasian	251/355	147	346	9	179	525	6	23	101	119	8	28	120	202	5
Zhao et al.	2004	Chinese	138/184	10	249	17	16	341	11	0	10	113	15	1	13	160	10
Lin et al.	2003	Chinese	193/254	32	342	12	30	465	13	1	29	151	12	1	26	214	13
Lynch et al.	2003	African-American, and Other	100/63	—	—	—	—	—	—	10	44	42	4	4	24	32	3
Goudreau et al.	2002	Caucasian	183/146	114	249	3	76	211	5	—	—	—	—	—	—	—	—
Kimura et al.	2001	Japanese	204/300	17	371	20	25	551	24	—	—	—	—	—	—	—	—
Kim et al.	2000	Korean	116/128	32	179	21	37	209	10	12	7	84	13	15	6	101	6
Zhang et al.	2000	Chinese	128/85	13	231	12	2	156	12	0	13	104	11	0	2	73	10
Wang et al.	2000	Chinese	171/180	20	300	22	13	333	14	0	20	130	21	0	13	153	14
Mercier et al.	1999	French	75/78	48	99	3	52	102	2	10	26	36	3	8	36	32	2
Nicholl et al.	1999	Caucasian	206/206	—	—	—	—	—	—	15	73	113	5	17	86	100	3
Leighton et al.	1997	Chinese	203/230	28	366	11	35	415	10	0	27	164	12	2	31	187	10
Le Couteur et al.	1997	Caucasian	100/200	51	144	5	112	286	2	7	36	52	5	15	81	102	2
Plante-Bordeneuve et al.	1997	British,French	78/60	42	108	6	34	85	1	—	—	—	—	—	—	—	—

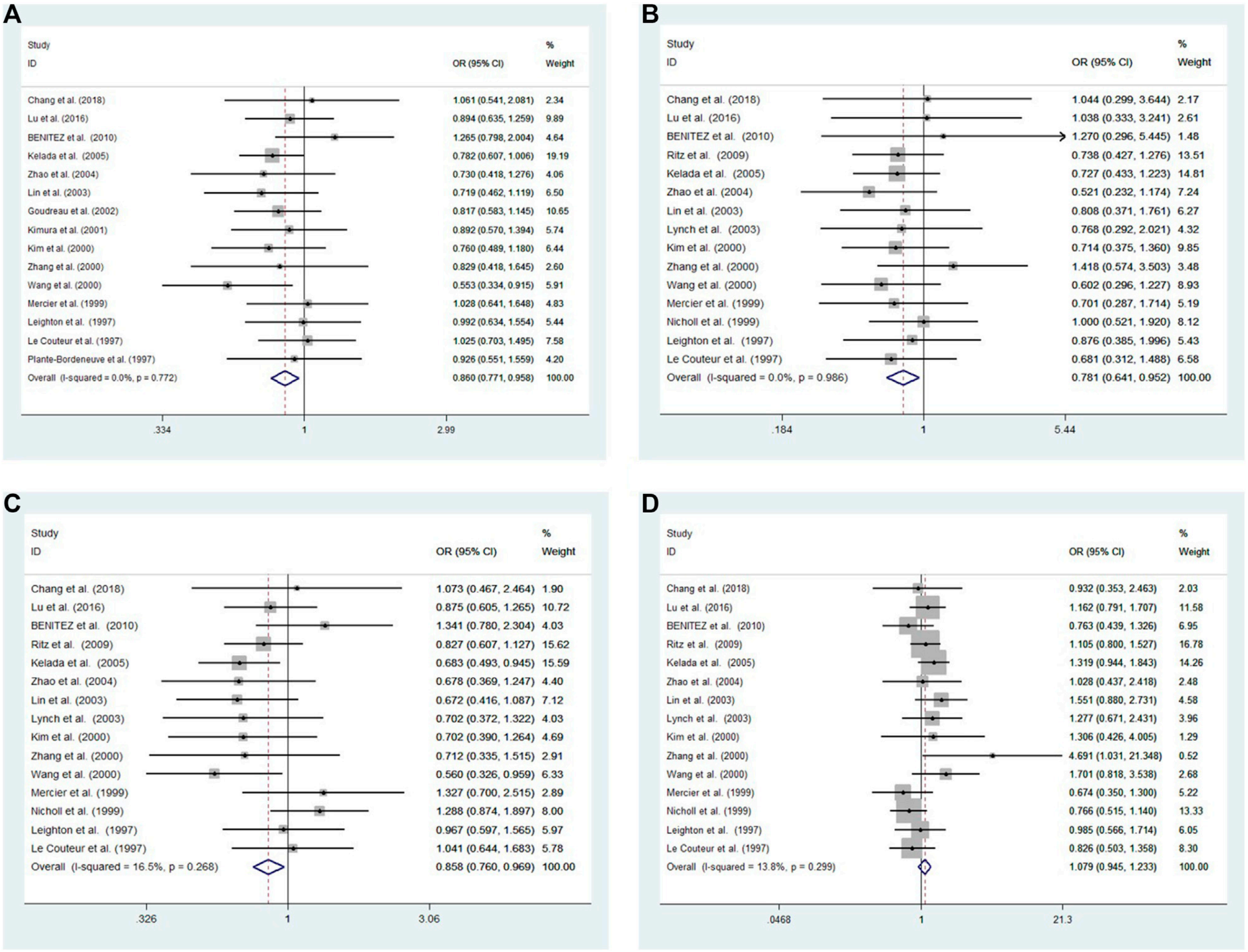


FIGURE 2 | Meta-analysis with a fixed effects model for the association between the 3′-UTR VNTR in SLC6A3 and PD susceptibility. **(A)** Allele model, 10R vs. all others **(B)** Pseudodominant model, 10R/10R + 10R/9R vs. all others **(C)** Pseudorecessive model, 10R/10R vs. all others **(D)** Pseudocodominant model, 10R/9R vs. all others OR: odds ratio, CI: confidence interval, I-squared: measured to quantify the degree of heterogeneity in meta-analyses.

that evaluated the associations between the *SLC6A3* 3′-UTR VNTR polymorphism and the risk of PD; 2) available data for estimating odds ratios (ORs) with corresponding 95% confidence intervals (CIs); and 3) studies in which all PD patients had been diagnosed according to the common diagnostic criteria (Jankovic 2008).

Studies were excluded from the current analysis with the following criteria: 1) not a case-control or cohort study; 2) irrelevant to PD or *SLC6A3* 3′-UTR VNTR; 3) genotype distribution of the control subjects is not in Hardy-Weinberg equilibrium (HWE); and 4) reports lacking detailed genotype data.

Data Extraction

The following data were independently extracted from the included studies and entered into a database to ensure the validity of the data: first author's name, year of publication, ethnicity, number of patients and controls, allele distribution, and genotype distribution. Studies were excluded if they did not provide the above information.

Statistical Analysis

Four genetic models were used in the meta-analysis: the allele model (10R vs. all others), the pseudodominant model (10R/10R + 10R/9R vs. all others), the pseudorecessive model (10R/10R vs. all others), and the pseudocodominant model (10R/9R vs. all others). Genetic heterogeneity was evaluated using the Q-test and I^2 test. I^2 ranged from 0 to 100%. Significant heterogeneity was defined with $p < 0.01$ and $I^2 > 50\%$ (He et al., 2015; Shen et al., 2015; Zhang et al., 2016). If there was no significant heterogeneity among the total of studies, ORs with corresponding 95% CIs were calculated by the fixed effect model (Mantel–Haenszel); otherwise, ORs were calculated by a random-effect model. Z test was used to determine the significance of OR. Additionally, publication bias was investigated with Egger's test and Begg's test (Li et al., 2016; Liu et al., 2017; Han et al., 2019). Statistical

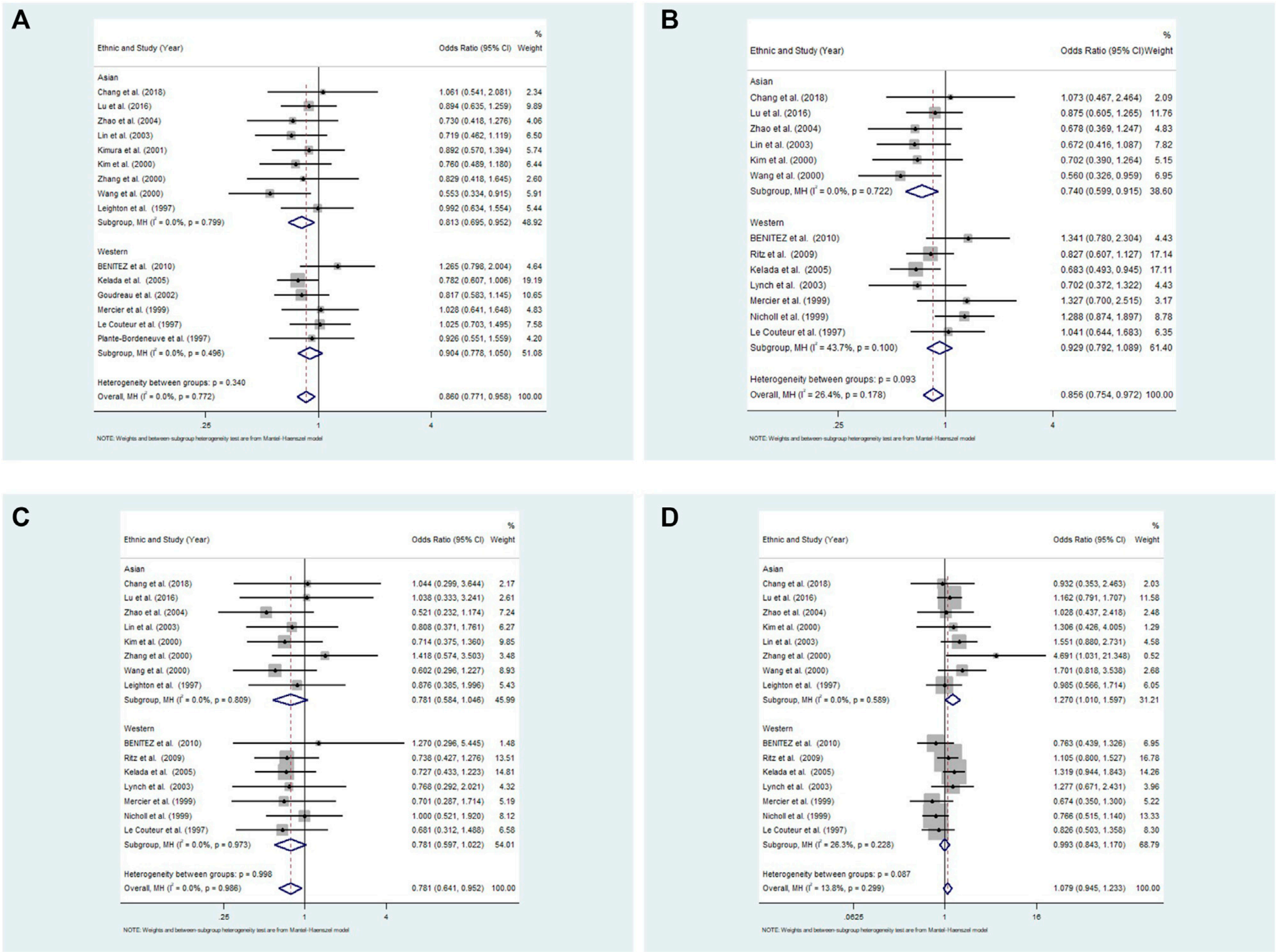


FIGURE 3 | Meta-analysis with a fixed effects model for the association between the 3′-UTR VNTR in SLC6A3 and PD susceptibility in Asian and Western populations. **(A)** Allele model, 10R vs. all others; **(B)** Pseudodominant model, 10R/10R + 10R/9R vs. all others; **(C)** Pseudorecessive model, 10R/10R vs. all others; **(D)** Pseudocodominant model, 10R/9R vs. all others OR: odds ratio, CI: confidence interval, I-squared: measure to quantify the degree of heterogeneity in meta-analyses.

analyses were performed using STATA v.16.0 software (Stata Corporation, Texas, United States).

RESULTS

Study Inclusion and Characteristics

A total of 175 potential studies were retrieved through the initial search. Thirty-two duplicates were excluded. Then, 143 studies were screened on title and abstract, 91 of which were excluded. The remaining 52 articles were evaluated by full-text reading, 34 of which were excluded because 10 were not case-control or cohort studies, 20 were not related to the *SLC6A3* 3′-UTR VNTR or PD, and 4 did not provide sufficient data. A flow chart of study selection in the meta-analysis is shown in **Figure 1**. There were 18 potentially relevant papers, including 14 in English and 4 in Chinese; among them, 15 studies provided allele model data (Chang et al., 2018; Lu et al., 2016; Benitez et al., 2010; Kelada et al., 2005; Zhao et al., 2004; Lin et al., 2003; Goudreau et al., 2002; Kimura et al., 2001; Kim et al., 2000; Zhang et al., 2000; Wang et al., 2000; Mercier et al., 1999; Leighton et al., 1997; Le Couteur et al., 1997; plante-Bordeneuve et al., 1997), and 15 studies had pseudodominant, pseudorecessive and pseudoadditive model data (Chang et al., 2018; Lu et al., 2016; Benitez et al., 2010; Ritz et al., 2009; Kelada et al., 2005; Zhao et al., 2004; Lin et al., 2003; Lynch et al., 2003; Kim et al., 2000; Zhang et al., 2000; Wang et al., 2000; Mercier et al., 1999; Nicholl et al., 1999; Leighton et al., 1997; Le Couteur et al., 1997). The characteristics of each study are shown in **Table 1**.

Heterogeneity Analysis

Cochran's Q and I^2 test results revealed low heterogeneity among studies in four models (10R vs. all others p = 0.772 I^2 = 0.0%; 10R/10R + 10R/9R vs. all others: p = 0.986 I^2 = 0.0%; 10R/10R vs. all others: p = 0.268 I^2 = 16.5%; 10R/9R vs. all others p = 0.299 I^2 = 13.8%, respectively) (**Figure 2**).

In the subgroup analysis by ethnicity, the results also revealed low heterogeneity among studies in four models in

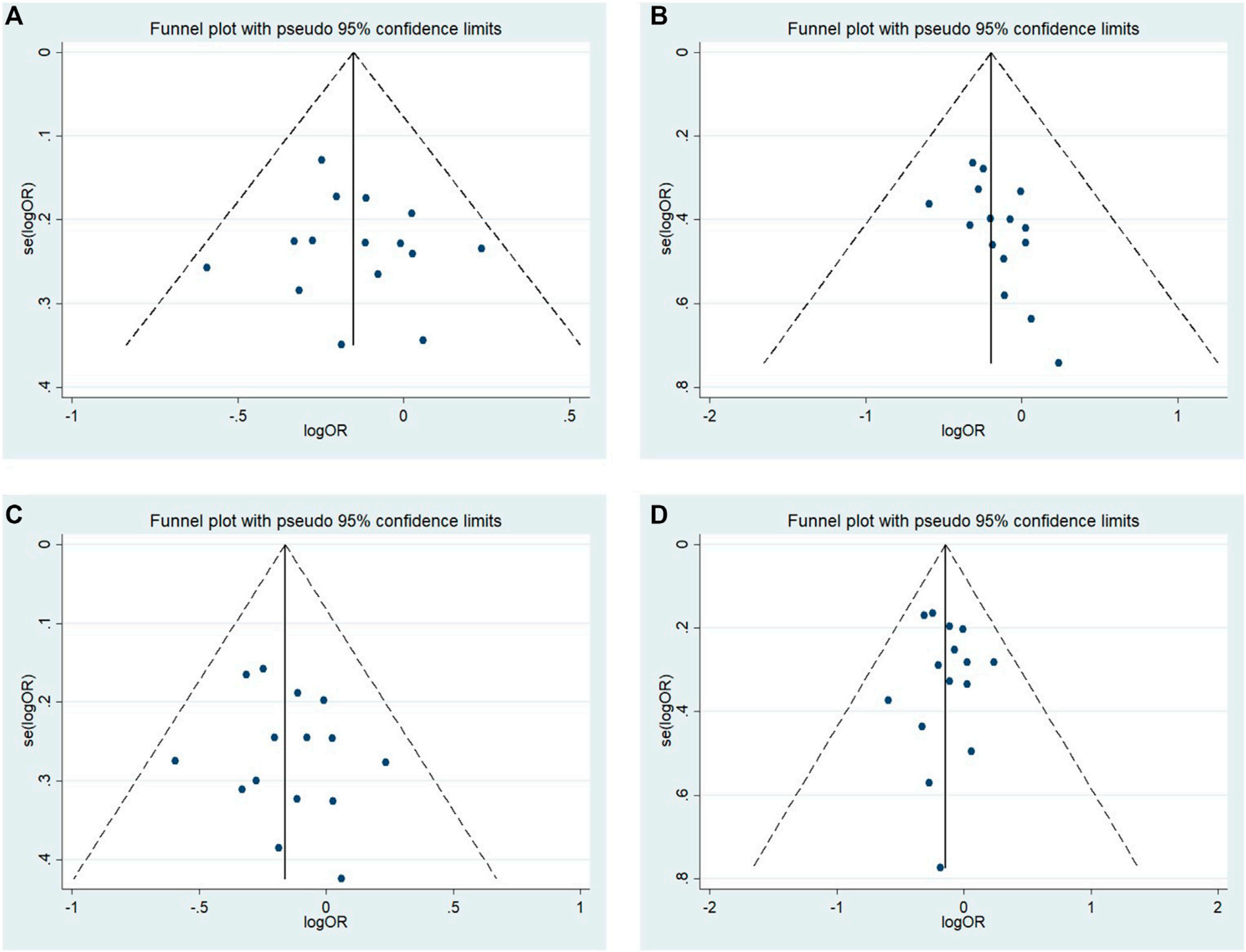


FIGURE 4 | Funnel plot of the odds ratios in the meta-analysis. **(A)** Allele model, 10R vs. all others **(B)** Pseudodominant model, 10R/10R + 10R/9R vs. all others **(C)** Pseudorecessive model, 10R/10R vs. all others **(D)** Pseudocodominant model, 10R/9R vs. all others OR: odds ratio, CI: confidence interval, I-squared: measured to quantify the degree of heterogeneity in meta-analyses.

the Asian populations (10R vs. all others p = 0.799 I^2 = 0.0%; 10R/10R + 10R/9R vs. all others: p = 0.809 I^2 = 0.0%; 10R/10R vs. all others: p = 0.792 I^2 = 0.0%; 10R/9R vs. all others p = 0.589 I^2 = 0.0%, respectively) and in Western populations (10R vs. all others p = 0.496 I^2 = 0.0%; 10R/10R + 10R/9R vs. all others: p = 0.973 I^2 = 0.0%; 10R/10R vs. all others: p = 0.100 I^2 = 43.7%; 10R/9R vs. all others p = 0.228 I^2 = 26.3%, respectively) (**Figure 3**).

The Association Between the 10-Repeat of the 3′-UTR VNTR in *SLC6A3* and PD

A fixed-effect model was used to analyze four models. The results showed a significant difference between patients and controls for the allele model (10R vs. all others: OR = 0.860, 95% CI: 0.771–0.958, p = 0.006), pseudodominant model (10R/10R + 10R/9R vs. all others: OR = 0.781, 95% CI: 0.641–0.952, p = 0.014) and pseudorecessive model (10R/10R vs. all others: OR = 0.858, 95% CI: 0.760–0.969, p = 0.013). No significant differences were observed under the pseudocodominant model (10R/9R vs. all others: OR = 1.079, 95% CI: 0.945–1.233, p = 0.262) (**Figure 2**).

In the subgroup analysis by ethnicity, the results showed a significant difference between patients and controls for the allele model (10R vs. all others: OR = 0.813, 95% CI: 0.695–0.952, p = 0.010), pseudorecessive model (10R/10R vs. all others: OR = 0.769, 95% CI: 0.637–0.928, p = 0.006) and pseudocodominant model (10R/9R vs. all others: OR = 1.270, 95% CI: 1.010–1.597, p = 0.041), but no significant differences were observed under the pseudodominant model (10R/10R + 10R/9R vs. all others: OR = 0.781, 95% CI: 0.584–1.046, p = 0.097) in Asian populations with a fixed-effect model. There was no significant difference between patients and controls for the four models in the Western populations (10R vs. all others: OR = 0.904, 95% CI: 0.778–1.050, p = 0.187; 10R/10R + 10R/9R vs. all others: OR = 0.781, 95% CI: 0.597–1.022, p = 0.071; 10R/10R vs. all others: OR = 0.929, 95% CI: 0.792–1.089, p = 0.361; 10R/9R vs. all others: OR = 0.993, 95% CI: 0.843–1.170, p = 0.930) (**Figure 3**).

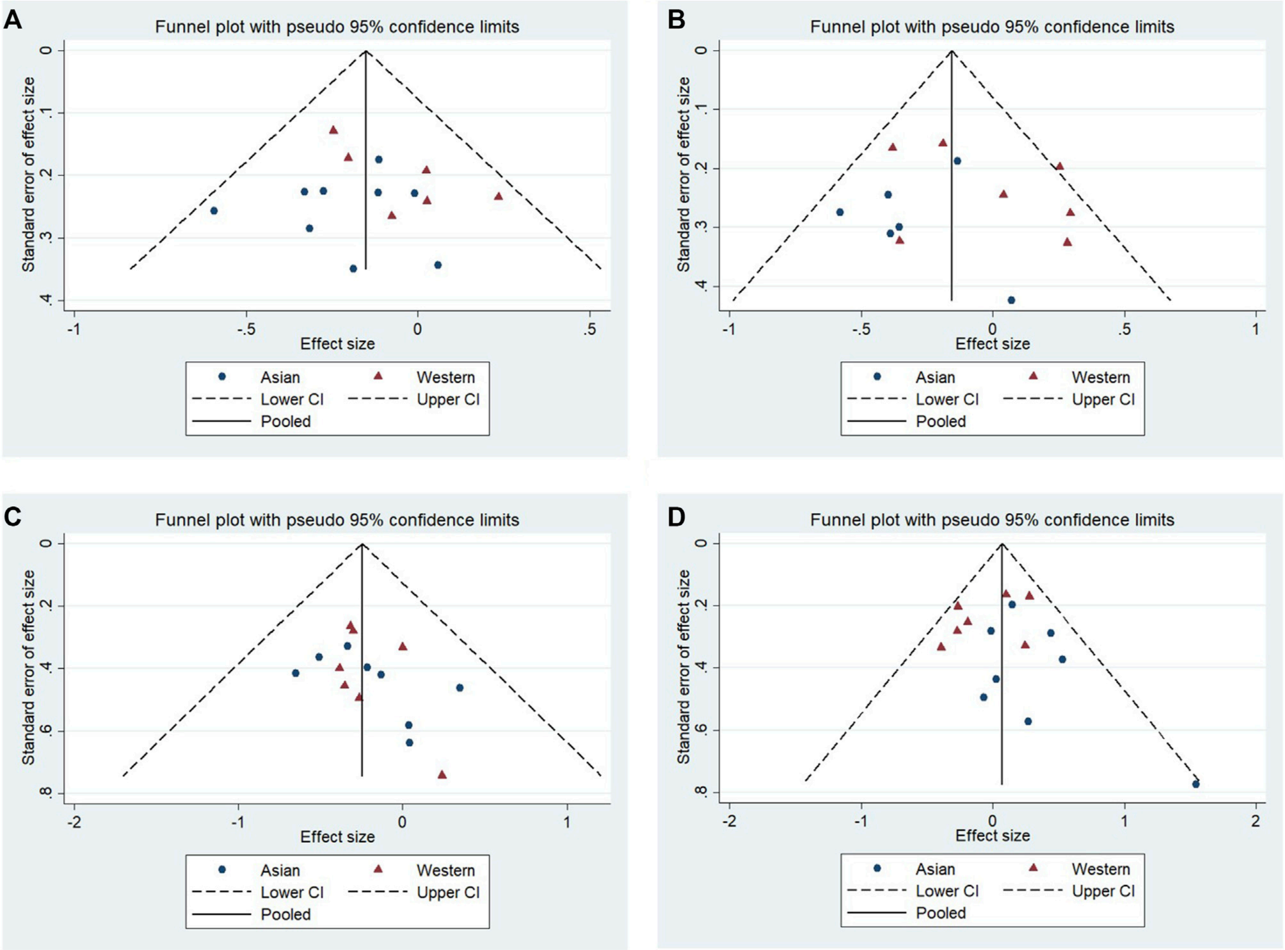


FIGURE 5 | Funnel plot of the odds ratios in the subgroup: Asian populations and Western populations. **(A)** Allele model, 10R vs. all others. **(B)** Pseudodominant model, 10R/10R + 10R/9R vs. all others. **(C)** Pseudorecessive model, 10R/10R vs. all others. **(D)** Pseudocodominant model, 10R/9R vs. all others.

Publication Bias

No significant publication bias was observed in any of the above genetic models via Begg's funnel plot and Egger's test (all $p > 0.05$, data not shown), and the funnel plot was symmetrical, with studies not coagulating into one quadrant of the funnel (**Figures 4**, **5**).

DISCUSSION

This meta-analysis assessed the association between the 10R allele of the 3′-UTR VNTR in the *SLC6A3* gene and PD, and it included a total of 18 published studies. In general, our findings suggested that the 10R alleles and 10R/10R and 10R/10R + 10R/9R genotypes of the VNTR polymorphism in the *SLC6A3* gene confer protection against PD. The 10R alleles and 10R/10R genotype results were replicated in Asian populations, and the 10R/9R genotype was associated with an increased risk of PD in Asian populations. The current meta-analysis confirmed most of the previous findings showing that the 10R allele of the 3′-UTR VNTR in the *SLC6A3* gene may be a protective factor in susceptibility to PD.

Previous studies have shown that the prevalence of PD in Asia is low, approximately half that of Caucasians (Zhang and Román 1993; Leighton et al., 1997). This may be related to the discrepancies in genetic polymorphisms among populations of different racial and ethnic groups. There was a difference in allelic frequency in the *SLC6A3* VNTR polymorphism (Vandenbergh et al., 1992; Sano et al., 1993; Le Couteur et al., 1997; Leighton et al., 1997; Mercier et al., 1999; Kim et al., 2000) and the distribution was similar among the different Asian ethnic populations (Chinese, Korean and Japanese), but it was different from the Western populations. There are research findings that the frequencies of the 10 and 11 repeats of *SLC6A3* in Asian populations were higher than those in Caucasians, but the 9R of *SLC6A3* was much lower in normal Asian populations. The results of our meta-analysis indicate that the 10R may be a protective factor against susceptibility to PD in Asian populations, which may be one of the reasons for the low prevalence of PD in Asia.

Several studies indicate that the 9R allele demonstrates more enhanced transcription activity than the 10R allele of the *SLC6A3*

VNTR polymorphism (Purcaro et al., 2019; Miller and Madras, 2002; Michelhaugh et al., 2001). From a clinical point of view, increased DAT expression due to the 9R allele might exacerbate striatal neuronal damage over time by increasing the presynaptic uptake of potentially neurotoxic endogenous or exogenous substrates via DAT, such as 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) (Contin et al., 2004; Tipton and Singer, 1993). However, the 10R/10R genotype of the *SLC6A3* gene may result in the most stable expression, which may confer nerve terminal protection against Mpp^+-like compounds and prevent the toxicity of dopaminergic neurons (Lin et al., 2003). This may effectively reduce the incidence of PD. Moreover, interindividual genetic differences in DAT might also play a role in the therapeutic outcome of levodopa-treated PD patients (Contin et al., 2004). The DAT 9R allele has been suggested to be a predictor of dyskinesias or psychosis in PD patients (Kaiser et al., 2003). In general, research has shown that changes in the number of VNTR copies are closely related to PD, and our meta-analysis suggests that the 10R allele may be a protective factor in susceptibility to PD. We also conducted heterogeneity analysis, and we found low heterogeneity in our meta-analysis. In addition, our meta-analysis showed no publication bias.

There are potential limitations to the current study. First, PD is a complex disorder that develops as a result of age, environmental, and genetic factors, but age and exposure to environmental agents were often not discussed in our included studies. Moreover, interactions between multiple genes might affect the risk of PD. Additionally, since some are a bit ambiguous from the current Ethnic column (e.g., Ritz et al., Lynch et al.). Ritz's study inclued 13 Asian populations, and Lynch‘s sudy inclued 9 other populations. Although these quantities account for a relatively small proportion of the total, these were difficult to conduct more accurate analyses. Therefore, the findings should be interpreted with caution. Further studies are necessary to establish larger sample sizes and consider SNP-SNP, gene–gene and gene–environmental interactions before reaching robust conclusions.

CONCLUSION

Our findings suggest that the 10R of the 3′-UTR VNTR in *SLC6A3* may be a protective factor in susceptibility to PD. This result was also confirmed in Asian populations.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

QZ, FN and SG were responsible for the statistical analysis, study design, and manuscript preparation. Qd Z, RC, and LP managed the literature searches and analyses. The study was supervised by DZ, GM, and YW.

FUNDING

Support for this work includes funding from the National Natural Science Foundation of China (81670252 and 81770034), Guangdong Basic and Applied Basic Foundation (2019A1515011306 and 2020A1515010240), Guangdong Medical University Research Foundation (GDMUM2020008 and GDMUM2020012); Medical Research Project of Foshan Health Bureau (20210188 and 20210289).

REFERENCES

Balestrino, R., and Schapira, A. H. V. (2020). Parkinson Disease. *Eur. J. Neurol.* 27 (1), 27–42. doi:10.1111/ene.14108

Bannon, M. J., Michelhaugh, S. K., Wang, J., and Sacchetti, P. (2001). The Human Dopamine Transporter Gene: Gene Organization, Transcriptional Regulation, and Potential Involvement in Neuropsychiatric Disorders. *Eur. Neuropsychopharmacol.* 11, 449–455. doi:10.1016/s0924-977x(01)00122-5

Benitez, B. A., Forero, D. A., Arboleda, G. H., Granados, L. A., Yunis, J. J., Fernandez, W., et al. (2010). Exploration of Genetic Susceptibility Factors for Parkinson's Disease in a South American Sample. *J. Genet.* 89, 229–232. doi:10.1007/s12041-010-0030-1

Chang, P., Zhao, P., Wang, X., Wang, C., and Bao, B. (2018). Relationship between Dopamine Transporter Gene Polymorphism and Parkinson's Disease Susceptibility and PET- CT Imagine. *China Chin. J. Geriatr.* 21, 5244–5246. [in Chinese].

Cheng, M. H., and Bahar, I. (2015). Molecular Mechanism of Dopamine Transport by Human Dopamine Transporter. *Structure* 23 (11), 2171–2181. doi:10.1016/j.str.2015.09.001

Contin, M., Martinelli, P., Mochi, M., Albani, F., Riva, R., Scaglione, C., et al. (2004). Dopamine Transporter Gene Polymorphism, Spect Imaging, and Levodopa Response in Patients with Parkinson Disease. *Clin. Neuropharmacology* 27 (3), 111–115. doi:10.1097/00002826-200405000-00004

Fuke, S., Suo, S., Takahashi, N., Koike, H., Sasagawa, N., and Ishiura, S. (2001). The VNTR Polymorphism of the Human Dopamine Transporter (DAT1) Gene Affects Gene Expression. *Pharmacogenomics J.* 1, 152–156. doi:10.1038/sj.tpj.6500026

Goudreau, J. L., Maraganore, D. M., Farrer, M. J., Lesnick, T. G., Singleton, A. B., Bower, J. H., et al. (2002). Case-control Study of Dopamine Transporter-1, Monoamine Oxidase-B, and Catechol-O-Methyl Transferase Polymorphisms in Parkinson's Disease. *Mov Disord.* 17, 1305–1311. doi:10.1002/mds.10268

Han, Z., Wang, T., Tian, R., Zhou, W., Wang, P., Ren, P., et al. (2019). BIN1 Rs744373 Variant Shows Different Association with Alzheimer's Disease in Caucasian and Asian Populations. *BMC bioinformatics* 20, 691. doi:10.1186/s12859-019-3264-9

He, D., Ma, L., Feng, R., Zhang, L., Jiang, Y., Zhang, Y., et al. (2015). Analyzing Large-Scale Samples Highlights Significant Association between Rs10411210 Polymorphism and Colorectal Cancer. *Biomed. Pharmacother.* 74, 164–168. doi:10.1016/j.biopha.2015.08.023

Heinz, A., Goldman, D., Jones, D. W., Palmour, R., Hommer, D., Gorey, J. G., et al. (2002). Genotype Influences *In Vivo* Dopamine Transporter Availability in Human Striatum. *Neuropsychopharmacology* 22, 133–139. doi:10.1016/S0893-133X(99)00099-8

Jankovic, J. (2008). Parkinson's Disease: Clinical Features and Diagnosis. *J. Neurol. Neurosurg. Psychiatry* 79, 368–376. doi:10.1136/jnnp.2007.131045

Jia, X., Wang, F., Han, Y., Geng, X., Li, M., Shi, Y., et al. (2016). miR-137 and miR-491 Negatively Regulate Dopamine Transporter Expression and Function in Neural Cells. *Neurosci. Bull.* 32 (6), 512–522. doi:10.1007/s12264-016-0061-6

Kaiser, R., Hofer, A., Grapengiesser, A., Gasser, T., Kupsch, A., Roots, I., et al. (2003). L -Dopa-Induced Adverse Effects in PD and Dopamine Transporter Gene Polymorphism. *Neurology* 60 (11), 1750–1755. doi:10.1212/01.wnl.0000068009.32067.a1

Kelada, S. N., Costa-Mallen, P., Checkoway, H., Carlson, C. S., Weller, T.-S., Swanson, P. D., et al. (2005). Dopamine Transporter (SLC6A3) 5′ Region Haplotypes Significantly Affect Transcriptional Activity *In Vitro* but Are Not Associated with Parkinson's Disease. *Pharmacogenetics and genomics* 15, 659–668. doi:10.1097/01.fpc.0000170917.04275.d6

Kim, J. W., Kim, D. H., Kim, S. H., and Cha, J. K. (2000). Association of the Dopamine Transporter Gene with Parkinson's Disease in Korean Patients. *J. Korean Med. Sci.* 15, 449–451. doi:10.3346/jkms.2000.15.4.449

Kimura, M., Matsushita, S., Arai, H., Takeda, A., and Higuchi, S. (2001). No Evidence of Association between a Dopamine Transporter Gene Polymorphism (1215A/G) and Parkinson's Disease. *Ann. Neurol.* 49, 276–277. doi:10.1002/1531-8249(20010201)49:2<276:aid-ana54>3.0.co;2-2

Krontiris, T. (1995). Minisatellites and Human Disease. *Science* 269, 1682–1683. doi:10.1126/science.7569893

Le Couteur, D. G., Leighton, P. W., McCann, S. J., and Pond, S. M. (1997). Association of a Polymorphism in the Dopamine-Transporter Gene with Parkinson's Disease. *Mov Disord.* 12, 760–763. doi:10.1002/mds.870120523

Lee, F. J. S., Liu, F., Pristupa, Z. B., and Niznik, H. B. (2001). Direct Binding and Functional Coupling of α-synuclein to the Dopamine Transporters Accelerate Dopamine-induced Apoptosis. *FASEB j.* 15, 916–926. doi:10.1096/fj.00-0334com10.1096/fsb2fj000334com

Leighton, P. W., Le Couteur, D. G., Pang, C. C. P., McCann, S. J., Chan, D., Law, L. K., et al. (1997). The Dopamine Transporter Gene and Parkinson's Disease in a Chinese Population. *Neurology* 49, 1577–1579. doi:10.1212/wnl.49.6.1577

Li, Y., Song, D., Jiang, Y., Wang, J., Feng, R., Zhang, L., et al. (2016). CR1 Rs3818361 Polymorphism Contributes to Alzheimer's Disease Susceptibility in Chinese Population. *Mol. Neurobiol.* 53, 4054–4059. doi:10.1007/s12035-015-9343-7

Lin, J. J., Yueh, K. C., Chang, D. C., Chang, C. Y., Yeh, Y. H., and Lin, S. Z. (2003). The Homozygote 10-Copy Genotype of Variable Number Tandem Repeat Dopamine Transporter Gene May Confer Protection Against Parkinson's Disease for Male, But Not to Female Patients. *J. Neurol. Sci.* 209, 87–92. doi:10.1016/s0022-510x(03)00002-9

Liu, G., Xu, Y., Jiang, Y., Zhang, L., Feng, R., and Jiang, Q. (2017). PICALM Rs3851179 Variant Confers Susceptibility to Alzheimer's Disease in Chinese Population. *Mol. Neurobiol.* 54, 3131–3136. doi:10.1007/s12035-016-9886-2

Lu, Q., Song, Z., Deng, X., Xiong, W., Xu, H., Zhang, Z., et al. (2016). SLC6A3 Rs28363170 and Rs3836790 Variants in Han Chinese Patients with Sporadic Parkinson's Disease. *Neurosci. Lett.* 629, 48–51. doi:10.1016/j.neulet.2016.06.053

Lynch, D. R., Mozley, P. D., Sokol, S., Maas, N. M. C., Balcer, L. J., and Siderowf, A. D. (2003). Lack of Effect of Polymorphisms in Dopamine Metabolism Related Genes on Imaging of TRODAT-1 in Striatum of Asymptomatic Volunteers and Patients with Parkinson's Disease. *Mov Disord.* 18, 804–812. doi:10.1002/mds.10430

Mercier, G., Turpin, J. C., and Lucotte, G. (1999). Variable Number Tandem Repeat Dopamine Transporter Gene Polymorphism and Parkinson's Disease: No Association Found. *J. Neurol.* 246, 45–47. doi:10.1007/s004150050304

Michelhaugh, S. K., Fiskerstrand, C., Lovejoy, E., Bannon, M. J., and Quinn, J. P. (2001). The Dopamine Transporter Gene (SLC6A3) Variable Number of Tandem Repeats Domain Enhances Transcription in Dopamine Neurons. *J. Neurochem.* 79 (5), 1033–1038. doi:10.1046/j.1471-4159.2001.00647.x

Mill, J., Asherson, P., Browes, C., D'Souza, U., and Craig, I. (2002). Expression of the Dopamine Transporter Gene Is Regulated by the 3? UTR VNTR: Evidence from Brain and Lymphocytes Using Quantitative RT-PCR. *Am. J. Med. Genet.* 114, 975–979. doi:10.1002/ajmg.b.10948

Miller, G. M., and Madras, B. K. (2002). Polymorphisms in the 3′-untranslated Region of Human and Monkey Dopamine Transporter Genes Affect Reporter Gene Expression. *Mol. Psychiatry* 7 (1), 44–55. doi:10.1038/sj.mp.4000921

Nicholl, D. J., Bennett, P., Hiller, L., Bonifati, V., Vanacore, N., Fabbrini, G., et al. (1999). A Study of Five Candidate Genes in Parkinson's Disease and Related Neurodegenerative Disorders. *Neurology* 53, 1415. doi:10.1212/wnl.53.7.1415

Planté-Bordeneuve, V., Taussig, D., Thomas, F., Said, G., Wood, N. W., Marsden, C. D., et al. (1997). Evaluation of Four Candidate Genes Encoding Proteins of the Dopamine Pathway in Familial and Sporadic Parkinson's Disease. *Neurology* 48, 1589–1593. doi:10.1212/wnl.48.6.1589

Purcaro, C., Vanacore, N., Moret, F., Di Battista, M. E., Rubino, A., Pierandrei, S., et al. (2019). DAT Gene Polymorphisms (Rs28363170, Rs393795) and Levodopa-Induced Dyskinesias in Parkinson's Disease. *Neurosci. Lett.* 690, 83–88. doi:10.1016/j.neulet.2018.10.021

Ritz, B. R., Manthripragada, A. D., Costello, S., Lincoln, S. J., Farrer, M. J., Cockburn, M., et al. (2009). Dopamine Transporter Genetic Variants and Pesticides in Parkinson's Disease. *Environ. Health Perspect.* 117, 964–969. doi:10.1289/ehp.0800277

Sano, A., Kondoh, K., Kakimoto, Y., and Kondo, I. (1993). A 40-nucleotide Repeat Polymorphism in the Human Dopamine Transporter Gene. *Hum. Genet.* 91, 405–406. doi:10.1007/BF00217369

Shen, N., Chen, B., Jiang, Y., Feng, R., Liao, M., Zhang, L., et al. (2015). An Updated Analysis with 85,939 Samples Confirms the Association between CR1 Rs6656401 Polymorphism and Alzheimer's Disease. *Mol. Neurobiol.* 51, 1017–1023. doi:10.1007/s12035-014-8761-2

Singleton, A. B., Farrer, M., Johnson, J., Singleton, A., Hague, S., Kachergus, J., et al. (2003). -Synuclein Locus Triplication Causes Parkinson's Disease. *Science* 302, 841. doi:10.1126/science.1090278

Thomas, B., and Beal, M. F. (2007). Parkinson's Disease. *Hum. Mol. Genet.* 16, R183–R194. doi:10.1093/hmg/ddm159

Tipton, K. F., and Singer, T. P. (1993). Advances in Our Understanding of the Mechanisms of the Neurotoxicity of MPTP and Related Compounds. *J. Neurochem.* 61 (4), 1191–1206. doi:10.1111/j.1471-4159.1993.tb13610.x

Uhl, G. R. (2003). Dopamine Transporter: Basic Science and Human Variation of a Key Molecule for Dopaminergic Function, Locomotion, and Parkinsonism. *Mov Disord.* 18, S71–S80. doi:10.1002/mds.10578

Uhl, G. R., Walther, D., Mash, D., Faucheux, B., and Javoy-Agid, F. (1994). Dopamine Transporter Messenger RNA in Parkinson's Disease and Control Substantia Nigra Neurons. *Ann. Neurol.* 35, 494–498. doi:10.1002/ana.410350421

Vandenbergh, D. J., Persico, A. M., Hawkins, A. L., Griffin, C. A., Li, X., Jabs, E. W., et al. (1992). Human Dopamine Transporter Gene (DAT1) Maps to Chromosome 5p15.3 and Displays a VNTR. *Genomics* 14, 1104–1106. doi:10.1016/s0888-7543(05)80138-7

Wang, J., Liu, Z., Chen, B., Li, J., and Chen, L. (2000). Association between Genetic Polymorphism of Dopamine Transporter Gene and Susceptibility to Parkinson's Disease. *Zhonghua Yi Xue Za Zhi* 80, 346–348. [in Chinese].

Wersinger, C., Prou, D., Vernier, P., and Sidhu., A. (2003). Modulation of Dopamine Transporter Function by α-synuclein Is Altered by Impairment of Cell Adhesion and by Induction of Oxidative Stress. *FASEB j.* 17, 1–30. doi:10.1096/fj.03-0152fje

Zhang, L., Shao, M., Xu, Q., Dong, X., Yang, J., and Li, Y. (2000). Association between Dopamine Transporter Gene Polymorphism and Parkinson's Disease. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi* 80, 431–434. [in Chinese].

Zhang, S., Li, X., Ma, G., Jiang, Y., Liao, M., Feng, R., et al. (2016). CLU Rs9331888 Polymorphism Contributes to Alzheimer's Disease Susceptibility in Caucasian but Not East Asian Populations. *Mol. Neurobiol.* 53, 1446–1451. doi:10.1007/s12035-015-9098-1

Zhang, Z.-X., and Román, G. C. (1993). Worldwide Occurrence of Parkinson's Disease: an Updated Review. *Neuroepidemiology* 12, 195–208. doi:10.1159/000110318

Zhao, X., Zhao, W., Xie, H., Su, J., Hao, Y., Han, H., et al. (2004). Parkinson ' S Disease Sensitivity and the 40-bp VNTR Polymorphism of DAT Gene. *China Chin. J. Geriatr.* 7, 457–459. [in Chinese].

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

CORRECTION
published: 21 October 2021
doi: 10.3389/fgene.2021.789112

Corrigendum: The 10-Repeat 3′-UTR VNTR Polymorphism in the *SLC6A3* Gene May Confer Protection Against Parkinson's Disease: A Meta-Analysis

Qiaoli Zeng [1,2†], Fan Ning [1,3†], Shanshan Gu [1,3†], Qiaodi Zeng [4], Riling Chen [1,5,2], Liuquan Peng [5], Dehua Zou [1,2], Guoda Ma [1*] and Yajun Wang [6*]*

[1]Maternal and Children's Health Research Institute, Shunde Women and Children's Hospital, Guangdong Medical University, Foshan, China, [2]Key Laboratory of Research in Maternal and Child Medicine and Birth Defects, Guangdong Medical University, Foshan, China, [3]Institute of Neurology, Affiliated Hospital of Guangdong Medical University, Zhanjiang, China, [4]Department of Clinical Laboratory, People's Hospital of Haiyuan County, Zhongwei, China, [5]Department of Pediatrics, Shunde Women and Children's Hospital, Guangdong Medical University, Foshan, China, [6]Institute of Respiratory, Shunde Women and Children's Hospital, Guangdong Medical University, Foshan, China

Keywords: Parkinson's disease, Slc6a3, dopamine transporter, variable number of tandem repeats, metaanalysis

OPEN ACCESS

Approved by:
Frontiers Editorial Office,
Frontiers Media SA, Switzerland

Correspondence:
Dehua Zou
757384895@qq.com
Guoda Ma
sihan1107@126.com
Yajun Wang
wangyajuny1977@aliyun.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *04 October 2021*
Accepted: *04 October 2021*
Published: *21 October 2021*

Citation:
Zeng Q, Ning F, Gu S, Zeng Q, Chen R, Peng L, Zou D, Ma G and Wang Y (2021) Corrigendum: The 10-Repeat 3′-UTR VNTR Polymorphism in the SLC6A3 Gene May Confer Protection Against Parkinson's Disease: A Meta-Analysis. Front. Genet. 12:789112. doi: 10.3389/fgene.2021.789112

A Corrigendum on

The 10-Repeat 3′-UTR VNTR Polymorphism in the SLC6A3 Gene May Confer Protection Against Parkinson's Disease: A Meta-Analysis
by Zeng, Q., Ning, F., Gu, S., Zeng Q., Chen, R., Peng, L., Zou, D., Ma, G., and Wang Y. (2021). Front. Genet. 12:757601. doi: 10.3389/fgene.2021.757601

In the original article, there were some mistake in the **Legends** for FIGURE 2 | Meta-analysis with a fixed effects model for the association between the 3′-UTR VNTR in SLC6A3 and COPD susceptibility and FIGURE 3 | Meta-analysis with a fixed effects model for the association between the 3′-UTR VNTR in SLC6A3 and COPD susceptibility in Asian and Western populations as published. The "COPD" in the legends of Figures 2 and 3 should be "PD." The correct legend appears below.

FIGURE 2 | Meta-analysis with a fixed effects model for the association between the 3′-UTR VNTR in SLC6A3 and PD susceptibility.

FIGURE 3 | Meta-analysis with a fixed effects model for the association between the 3′-UTR VNTR in SLC6A3 and PD susceptibility in Asian and Western populations.

Additionally, there were some minor formatting errors in **References**. Following references: Chang et al., 2018; Wang et al., 2000; Zhang et al., 2000; Zhao et al., 2004 as "(chinese)," should be "in Chinese" And the for reference: Lin et al., 2003, "Lin, J.-J., Yueh, K.-C., Chang, D.-C., Chang, C.-Y., Yeh, Y.-H., and Lin, S.-Z. (2003). The Homozygote 10-copy Genotype of Variable Number Tandem Repeat Dopamine Transporter Gene May Confer protection against Parkinson's Disease "for Male, but "Not to Female Patients. J. Neurol. Sci. 209, 87–92. doi: 10.1016/s0022-510x(03)00002-9, it should be Lin, J. J., Yueh, K. C., Chang, D. C., Chang, C. Y., Yeh, Y. H., and Lin, S. Z. (2003). The homozygote 10-copy genotype of variable number tandem repeat dopamine transporter gene may confer protection against Parkinson's disease for male, but not to female patients. J. Neurol. Sci. 209, 87–92. doi: 10.1016/s0022-510x(03)00002-9."

Finally, **Figure 5** was incorrectly cited in the Discussion section. A correction has been made to Section: Discussion, Paragraph 1:

"This meta-analysis assessed the association between the 10R allele of the 3′-UTR VNTR in the SLC6A3 gene and PD, and it included a total of 18 published studies. In general, our findings suggested that the 10R alleles and 10R/10R and 10R/10R + 10R/9R genotypes of the VNTR polymorphism in theSLC6A3 gene confer protection against PD. The 10R alleles and 10R/10R genotype results were replicated in Asian populations, and the 10R/9R genotype was associated with an increased risk of PD in Asian populations. The current meta-analysis confirmed most of the previous findings showing that the 10R allele of the 3′-UTR VNTR in the SLC6A3 gene may be a protective factor in susceptibility to PD."

The authors apologize for this error and state that this does not change the scientific conclusions of the article in any way. The original article has been updated.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 21 October 2021
doi: 10.3389/fgene.2021.745224

Rheumatoid Arthritis and Cardio-Cerebrovascular Disease: A Mendelian Randomization Study

*Shizheng Qiu[1]†, Meijie Li[2]†, Shunshan Jin[3]†, Haoyu Lu[1] and Yang Hu[1]**

[1] School of Life Sciences and Technology, Harbin Institute of Technology, Harbin, China, [2] Department of Neurology, Xuanwu Hospital, Capital Medical University, Beijing, China, [3] General Hospital of Heilongjiang Province Land Reclamation Bureau, Harbin, China

Significant genetic association exists between rheumatoid arthritis (RA) and cardiovascular disease. The associated mechanisms include common inflammatory mediators, changes in lipoprotein composition and function, immune responses, etc. However, the causality of RA and vascular/heart problems remains unknown. Herein, we performed Mendelian randomization (MR) analysis using a large-scale RA genome-wide association study (GWAS) dataset (462,933 cases and 457,732 controls) and six cardio-cerebrovascular disease GWAS datasets, including age angina (461,880 cases and 447,052 controls), hypertension (461,880 cases and 337,653 controls), age heart attack (10,693 cases and 451,187 controls), abnormalities of heartbeat (461,880 cases and 361,194 controls), stroke (7,055 cases and 454,825 controls), and coronary heart disease (361,194 cases and 351,037 controls) from United Kingdom biobank. We further carried out heterogeneity and sensitivity analyses. We confirmed the causality of RA with age angina (OR = 1.17, 95% CI: 1.04–1.33, p = 1.07E−02), hypertension (OR = 1.45, 95% CI: 1.20–1.75, p = 9.64E−05), age heart attack (OR = 1.15, 95% CI: 1.05–1.26, p = 3.56E−03), abnormalities of heartbeat (OR = 1.07, 95% CI: 1.01–1.12, p = 1.49E−02), stroke (OR = 1.06, 95% CI: 1.01–1.12, p = 2.79E−02), and coronary heart disease (OR = 1.19, 95% CI: 1.01–1.39, p = 3.33E−02), contributing to the understanding of the overlapping genetic mechanisms and therapeutic approaches between RA and cardiovascular disease.

Keywords: Mendelian randomization, genome-wide association studies, rheumatoid arthritis, cardiovascular disease, inverse-variance weighted

OPEN ACCESS

Edited by:
Liangcai Zhang,
Janssen Research and Development,
United States

Reviewed by:
Quan Zou,
University of Electronic Science and Technology of China, China
Balachandran Manavalan,
Ajou University, South Korea

Correspondence:
Yang Hu
huyang@hit.edu.cn

† These authors share first authorship

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *21 July 2021*
Accepted: *20 August 2021*
Published: *21 October 2021*

Citation:
Qiu S, Li M, Jin S, Lu H and Hu Y (2021) Rheumatoid Arthritis and Cardio-Cerebrovascular Disease: A Mendelian Randomization Study. Front. Genet. 12:745224. doi: 10.3389/fgene.2021.745224

INTRODUCTION

Cardiovascular disease remains the leading cause of human death, with an estimated 17.3 million people worldwide dying of cardiovascular disease each year, which is expected to increase to 23.6 million by 2030 (Laslett et al., 2012; Smith et al., 2012; Leong et al., 2017). Epidemiological studies have shown that the occurrence of cardiovascular disease is caused by various factors, with obesity, diabetes, smoking, hyperlipidemia, atherosclerosis, hypertension, and blood viscosity being its potential risk factor (Leong et al., 2017; Xu et al., 2019). Importantly, traditional cardiovascular disease risk factors account for a large proportion of rheumatoid arthritis (RA) (An et al., 2016). RA patients were 48% more likely to have cardiovascular disease than normal people and a 50% higher incidence of cardiovascular disease-related mortality (Avina-Zubieta et al., 2008, 2012; Sokka et al., 2008; England et al., 2018). However, most of the previous studies have examined the association between RA and atherosclerosis and congestive heart failure, ignoring other phenotypes of heart disease and vascular problems (England et al., 2018). Moreover, the exact causality is still unknown.

Mendelian randomization (MR) could estimate causality without bias, which has been used in previous studies to explore the association between phenotypes (Jansen et al., 2014; Smith et al., 2017; Hemani et al., 2018; Cheng et al., 2019b, 2021; Zhuang et al., 2019; Qiu et al., 2021). Causality between multiple metabolic characteristics, nutrient elements, and common diseases with cardiovascular disease have been demonstrated (Ference et al., 2017; Larsson et al., 2017, 2020; Yeung et al., 2018; Rosoff et al., 2020; Arvanitis et al., 2021). However, strong evidence linking RA to cardiovascular disease is still lacking. Herein, we mainly selected inverse-variance weighted (IVW), weighted median, and MR-Egger methods for MR analysis. We provided strong evidence that RA contributed to six vascular-/heart problem-related phenotypes, which could be of great significance for clinical disease prevention and treatment.

MATERIALS AND METHODS

Genome-Wide Association Study Dataset Sources

We obtained large-scale genome-wide association study (GWAS) summary datasets from the United Kingdom Biobank on RA and six cardiovascular disease phenotypes, and all of the participants were of European ancestry. From 2006 to 2010, the United Kingdom Biobank Assessment Center recruited 386,005 participants from the United Kingdom to participate in self-reporting of non-cancer illness (Li et al., 2015; Sun et al., 2019). RA GWAS (462,933 cases and 457,732 controls) was derived from the non-cancer illness study. At the same time, the United Kingdom Biobank Assessment Center carried out the study of vascular/heart problems diagnosed by doctors, covering 501,555 participants. These heart or vascular problems included age angina (461,880 cases and 447,052 controls), age high blood pressure (461,880 cases and 337,653 controls), age stroke (7,055 cases and 454,825 controls), and age heart attack (10,693 cases and 451,187 controls). In addition, we supplemented two other United Kingdom Biobank studies on coronary heart disease (CHD) (361,194 cases and 351,037 controls) and abnormalities of heartbeat (461,880 cases and 361,194 controls).

Quality Control and Identifying Genetic Instruments

In order to enhance the statistical power of genetic variants, we deleted single-nucleotide polymorphisms (SNPs) with a minor allele frequency (MAF) < 1%. Moreover, we removed variants with physical distance less than 10,000 kb and $R^2 < 0.001$ to avoid linkage disequilibrium (LD). For preprocessed exposure (RA) data, we selected genetic variants that passed genome-wide association threshold ($p < 5\text{E}-08$) as instrumental variables (IVs) to satisfy IV assumption 1 of MR analysis: variants should be strongly associated with exposure (RA) (Hemani et al., 2018).

Two-Sample Mendelian Randomization Analysis

In the absence of individual-level data, we used MR, a powerful statistical method, to infer the causality between two phenotypes. MR analysis eliminates the need to consider confounders and reverse causality. Two-sample MR requires that the samples of exposure and outcome be independent, which greatly expands the application range of MR (Hemani et al., 2018). Details of MR analysis have been described in previous reports (Davey Smith and Hemani, 2014; Bowden et al., 2015, 2016; Yavorska and Burgess, 2017; Cheng et al., 2018, 2019b; Hemani et al., 2018; Hu et al., 2020, 2021; Qiu et al., 2021). Herein, we first aligned alleles on the forward strand and harmonized SNP effects of exposure and outcome (Ong et al., 2021). If the variant in IVs was lacking in outcome, we allowed the proxy SNP with a strong LD to replace it (Hemani et al., 2018). Subsequently, we performed the inverse-variance weighted (IVW) estimator to estimate the association between RA and vascular/heart disease (Burgess and Thompson, 2017). In the case of certain invalid instruments or directional pleiotropy bias, the weighted median and MR-Egger estimators could help to make further judgment (Bowden et al., 2017; Burgess and Thompson, 2017; Hartwig et al., 2017; Cheng, 2019; Cheng et al., 2019a). Finally, we carried out reverse MR analysis to evaluate the evidence for reverse causal association.

Sensitivity Analysis

We performed a series of sensitivity tests to ensure that our results were robust, including heterogeneity tests to assess heterogeneity between IVs, leave-one-out analysis (Wei et al., 2019, 2021; Govindaraj et al., 2020; Hasan et al., 2021) to assess whether a single SNP over-drove outcome, and funnel plots and MR-Egger to assess potential horizontal pleiotropy (Bowden et al., 2017; Hartwig et al., 2017; Rosoff et al., 2020). The statistical tests for MR analysis were undertaken using the R package of meta and TwoSampleMR (Hemani et al., 2018). The statistically significant association is defined as $p < 0.05$.

RESULTS

Association of Rheumatoid Arthritis With Cardiovascular Disease

Eight genetic variants were used as IVs to evaluate the association between RA and cardiovascular disease (**Table 1**). Two SNPs needed to be proxied, but we lacked evidence for the presence of wrong effect alleles, strand issues, palindromic SNPs, and incompatible alleles. Due to the heterogeneity in some studies, we preferred to use the random-effects model (**Table 2**). By performing IVW analysis, we confirmed the causality of RA with age angina (OR = 1.17, 95% CI: 1.04–1.33, p = 1.07E−02), hypertension (OR = 1.45, 95% CI: 1.20–1.75, p = 9.64E−05), age heart attack (OR = 1.15, 95% CI: 1.05–1.26, p = 3.56E−03), abnormalities of heartbeat (OR = 1.07, 95% CI: 1.01–1.12, p = 1.49E−02), stroke (OR = 1.06, 95% CI: 1.01–1.12, p = 2.79E−02), and CHD (OR = 1.19, 95% CI: 1.01–1.39, p = 3.33E−02) (**Figure 1**). Detailed MR results are shown in

TABLE 1 | Characteristics of eight genetic variants as instrumental variables (IVs).

SNP	Chr	Pos	Effect allele	Other allele	Beta	EAF	SE	p
rs185320691	6	32,490,292	C	G	0.0076	0.10	0.00040	2.30E−82
rs28559870	6	31,377,974	T	C	0.0017	0.18	0.00029	4.10E−09
rs35175534	6	32,531,108	C	A	0.0076	0.14	0.00035	3.50E−105
rs460568	6	33,232,025	T	C	0.0019	0.17	0.00029	4.10E−11
rs6679677	1	114,303,808	A	C	0.0031	0.10	0.00036	4.90E−18
rs7731626	5	55,444,683	A	G	−0.0015	0.38	0.00023	2.00E−11
rs7760841	6	32,574,868	T	C	0.0083	0.17	0.00030	7.29E−172
rs9265076	6	31,287,765	T	C	0.0021	0.38	0.00026	1.10E−15

Beta is the estimated effect size for the effect allele; Beta > 0 and Beta < 0 means that this effect allele could increase and reduce RA risk, respectively. EAF, effect allele frequency.

TABLE 2 | The results of Mendelian randomization (MR) sensitivity analysis.

Phenotypes Methods	Angina	Hypertension	Heart attack	Abnormalities of heartbeat	Stroke	Coronary heart disease
MR-Egger	0.494	0.090	0.373	0.481	0.448	0.306
Heterogeneity tests	0.010	0.472	0.038	0.293	0.566	0.003
Leave-one-out analysis	0.0107	9.64E−05	0.00356	0.0149	0.0279	0.0333

The values in the table are all p-values.
MR-Egger: p > 0.05 means that the original hypothesis is rejected, and horizontal pleiotropy (non-zero intercept) has no significant effect on the results.
Heterogeneity tests: p > 0.05 means no heterogeneity.
Leave-one-out analysis: p < 0.05 means that no single single-nucleotide polymorphism (SNP) over-drives the overall results.

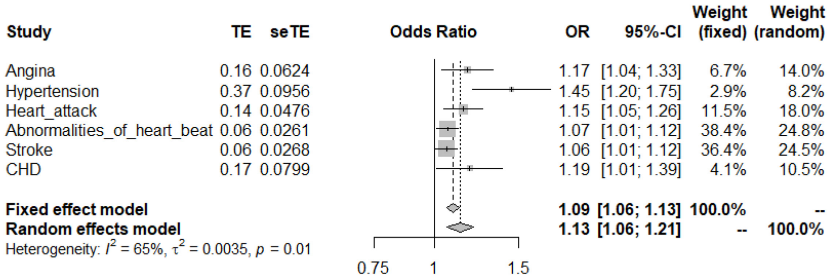


FIGURE 1 | Mendelian randomization (MR) analysis between rheumatoid arthritis (RA) and six cardiovascular diseases. TE, treatment effects (β); se TE, standard error of treatment effect (se).

the **Supplementary Material**. No evidence of reverse causality existed in any of the studies. Thus, RA made a significant contribution to common phenotypes associated with vascular or cardiac problems.

Sensitivity Analysis

Unlike IVW, MR-Egger allows horizontal pleiotropy between IVs and exposure and outcome, and the weighted median allows a more powerful variant to have a greater impact on the overall result (Bowden et al., 2017; Hartwig et al., 2017; Qiu et al., 2021). Other MR calculation methods are a powerful supplement to IVW, especially when the IV assumptions of MR framework is not satisfied perfectly. In all methods, our results were robust, with small intercepts and high *p*-values in MR-Egger, which meant that horizontal pleiotropy made almost no effect on the results (**Figure 2** and **Table 2**). According to funnel plots, rs7731626 and rs460568 in angina, and rs6679677 and rs9265076 in CHD, there existed a certain horizontal pleiotropy; however, little influence affected the overall results (**Figure 3**). Moreover, no single SNP over-drove the overall results.

DISCUSSION

In this study, we carried out MR analysis to demonstrate that RA was positively associated with six heart and vascular diseases. Observational studies evaluated that the patients with RA had a significantly increased risk of cardiovascular disease (An et al., 2016; Crowson et al., 2018). Crowson et al. (2018) followed up 5,638 RA patients for 5.8 years and found that about 30% of them eventually developed cardiovascular disease. RA might increase the risk of the six heart and vascular diseases we mentioned at the same time

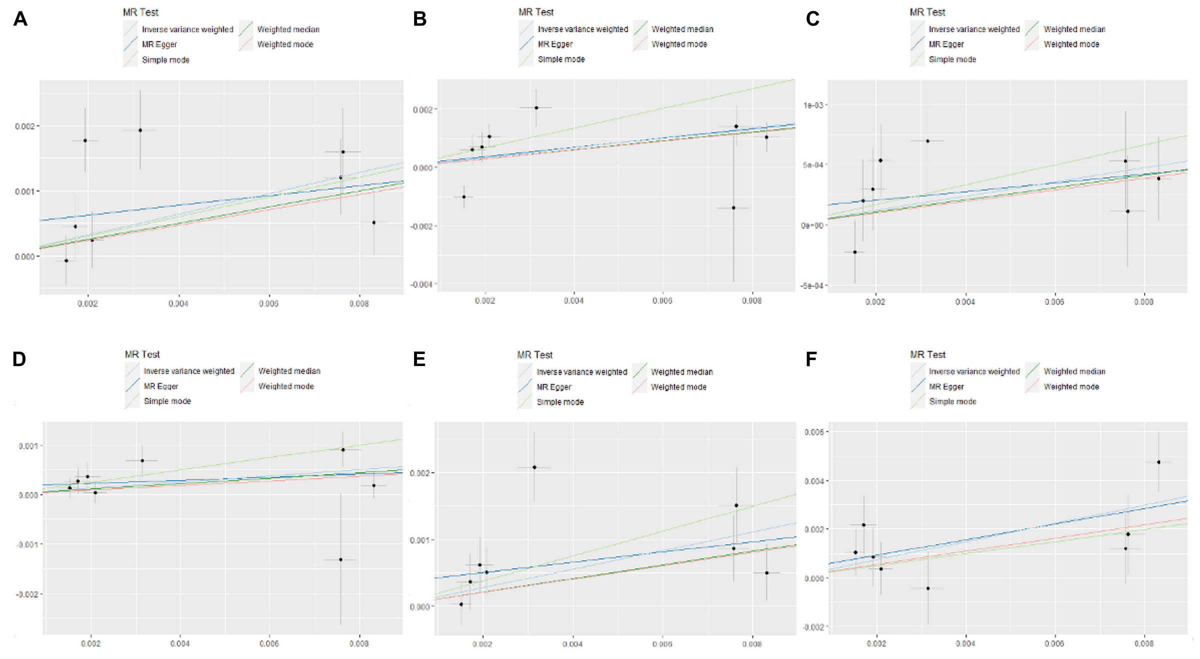


FIGURE 2 | MR tests of RA with angina, hypertension, heart attack, abnormalities of heartbeat, stroke, and coronary heart disease. The estimate of intercept can be interpreted as an estimate of the average pleiotropy of all single-nucleotide polymorphisms (SNPs), and the slope coefficient provides an estimate of the bias of the causal effect. **(A)** Angina. **(B)** Hypertension. **(C)** Heart attack. **(D)** Abnormalities of heartbeat. **(E)** Stroke. **(F)** Coronary heart disease.

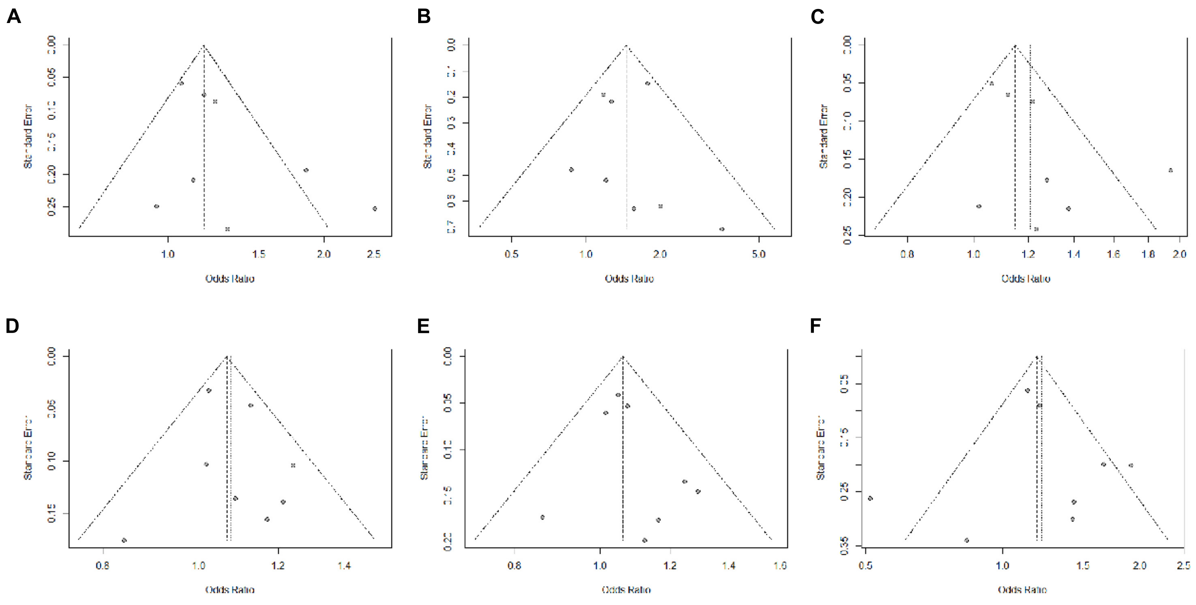


FIGURE 3 | Funnel plots of RA with angina, hypertension, heart attack, abnormalities of heartbeat, stroke, and coronary heart disease. The x-axis represents odds ratio (OR), and the y-axis represents standard error (se). **(A)** Angina. **(B)** Hypertension. **(C)** Heart attack. **(D)** Abnormalities of heartbeat. **(E)** Stroke. **(F)** Coronary heart disease.

(Kitas and Gabriel, 2011; Dougados et al., 2014; Hadwen et al., 2021). Thus, we subdivided the phenotypes of cardiac and vascular diseases to provide different risk values, which might be a significant help for the clinical cotreatment of RA and cardiovascular disease.

Our study may have many advantages over previous observational studies. First, we used seven large-scale GWAS datasets. RA GWAS alone involved more than 900,000 participants. The seven studies were all from European descent, avoiding potential population stratification. Second, MR greatly avoids the influence of confounding factors and reverse causality because the alleles of the SNP site are randomly assigned much earlier than the occurrence of any potential confounding factors. Third, we applied a variety of MR methods to jointly verify the robustness of the results. When some instruments were unavailable or horizontal pleiotropy, an unbiased causal estimate could still be given.

However, certain limitations existed in our study. First, there may be overlap of some samples in RA GWAS and cardiovascular disease GWAS, which leads to weak instruments bias (Pierce and Burgess, 2013). In addition, some of the genetic variants have a certain degree of heterogeneity or horizontal pleiotropy, such as rs7731626 and rs9265076, in the analysis of RA and angina. For population stratification from gender, age, and ancestry, inappropriate proxy SNP may be the potential reasons for them.

In conclusion, we explored the causality between RA and six cardio-cerebrovascular diseases for the first time, and all the results showed the risk effect of RA. Eight variants as IVs may be the link between RA and cardiovascular disease. We expect to find out the genetic association between chronic diseases more deeply in the future.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/ **Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

SQ and LH designed the study, analyzed the data, and wrote the manuscript. ML and SJ revised the manuscript. YH supervised the research and revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Key R&D Program of China (2017YFC1201201, 2018YFC0910504, and 2017YFC0907503), the Natural Science Foundation of China (61801147 and 82003553), and Heilongjiang Postdoctoral Science Foundation (LBH-Z6064), all to YH.

ACKNOWLEDGMENTS

We would like to thank the United Kingdom Biobank for the GWAS summary statistics (http://www.nealelab.is/uk-biobank/).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.745224/full#supplementary-material

REFERENCES

An, J., Alemao, E., Reynolds, K., Kawabata, H., Solomon, D. H., Liao, K. P., et al. (2016). Cardiovascular outcomes associated with lowering low-density lipoprotein cholesterol in rheumatoid arthritis and matched nonrheumatoid arthritis. *J. Rheumatol.* 43, 1989–1996. doi: 10.3899/jrheum.160110

Arvanitis, M., Qi, G. H., Bhatt, D. L., Post, W. S., Chatterjee, N., Battle, A., et al. (2021). Linear and nonlinear mendelian randomization analyses of the association between diastolic blood pressure and cardiovascular events the J-curve revisited. *Circulation* 143, 895–906. doi: 10.1161/circulationaha.120.049819

Avina-Zubieta, J. A., Choi, H. K., Sadatsafavi, M., Etminan, M., Esdaile, J. M., and Lacaille, D. (2008). Risk of cardiovascular mortality in patients with rheumatoid arthritis: a meta-analysis of observational studies. *Arthritis Rheum.* 59, 1690–1697.

Avina-Zubieta, J. A., Thomas, J., Sadatsafavi, M., Lehman, A. J., and Lacaille, D. (2012). Risk of incident cardiovascular events in patients with rheumatoid arthritis: a meta-analysis of observational studies. *Ann. Rheum. Dis.* 71, 1524–1529. doi: 10.1136/annrheumdis-2011-200726

Bowden, J., Del Greco, M. F., Minelli, C., Davey Smith, G., Sheehan, N., and Thompson, J. (2017). A framework for the investigation of pleiotropy in two-sample summary data mendelian randomization. *Stat. Med.* 36, 1783–1802. doi: 10.1002/sim.7221

Bowden, J., Smith, G. D., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* 44, 512–525. doi: 10.1093/ije/dyv080

Bowden, J., Smith, G. D., Haycock, P. C., and Burgess, S. (2016). Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* 40, 304–314. doi: 10.1002/gepi.21965

Burgess, S., and Thompson, S. G. (2017). Interpreting findings from mendelian randomization using the MR-Egger method. *Eur. J. Epidemiol.* 32, 377–389. doi: 10.1007/s10654-017-0255-x

Cheng, L. (2019). Computational and biological methods for gene therapy. *Curr. Gene Ther.* 19, 210–210. doi: 10.2174/156652321904191022113307

Cheng, L., Zhao, H., Wang, P., Zhou, W., Luo, M., Li, T., et al. (2019a). Computational methods for identifying similar diseases. *Mol. Ther. Nucleic Acids* 18, 590–604. doi: 10.1016/j.omtn.2019.09.019

Cheng, L., Zhu, Z., Wang, C., Wang, P., He, Y. O., and Zhang, X. (2021). COVID-19 induces lower levels of IL-8, IL-10, and MCP-1 than other acute CRS-inducing diseases. *Proc. Natl. Acad. Sci. U. S. A.* 118:e2102960118. doi: 10.1073/pnas.2102960118

Cheng, L., Zhuang, H., Ju, H., Yang, S., Han, J., Tan, R., et al. (2019b). Exposing the causal effect of body mass index on the risk of type 2 diabetes mellitus: a mendelian randomization study. *Front. Genet.* 10:94. doi: 10.3389/fgene.2019.00094

Cheng, L., Zhuang, H., Yang, S., Jiang, H., Wang, S., and Zhang, J. (2018). Exposing the causal effect of C-reactive protein on the risk of type 2 Diabetes mellitus: a mendelian randomization study. *Front. Genet.* 9:657. doi: 10.3389/fgene.2018.00657

Crowson, C. S., Rollefstad, S., Ikdahl, E., Kitas, G. D., Van Riel, P., Gabriel, S. E., et al. (2018). Impact of risk factors associated with cardiovascular outcomes in patients with rheumatoid arthritis. *Ann. Rheum. Dis.* 77, 48–54.

Davey Smith, G., and Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* 23, R89–R98.

Dougados, M., Soubrier, M., Antunez, A., Balint, P., Balsa, A., Buch, M. H., et al. (2014). Prevalence of comorbidities in rheumatoid arthritis and evaluation of their monitoring: results of an international, cross-sectional study (COMORA). *Ann. Rheum. Dis.* 73, 62–68.

England, B. R., Thiele, G. M., Anderson, D. R., and Mikuls, T. R. (2018). Increased cardiovascular risk in rheumatoid arthritis: mechanisms and implications. *BMJ* 361, k1036. doi: 10.1136/bmj.k1036

Ference, B. A., Ginsberg, H. N., Graham, I., Ray, K. K., Packard, C. J., Bruckert, E., et al. (2017). Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the European Atherosclerosis Society Consensus Panel. *Eur. Heart J.* 38, 2459–2472. doi: 10.1093/eurheartj/ehx144

Govindaraj, R. G., Subramaniyam, S., and Manavalan, B. (2020). Extremely-randomized-tree-based Prediction of N(6)-methyladenosine Sites in *Saccharomyces cerevisiae*. *Curr. Genomics* 21, 26–33. doi: 10.2174/1389202921666200219125625

Hadwen, B., Stranges, S., and Barra, L. (2021). Risk factors for hypertension in rheumatoid arthritis patients-a systematic review. *Autoimmun. Rev.* 20:102786. doi: 10.1016/j.autrev.2021.102786

Hartwig, F. P., Davey Smith, G., and Bowden, J. (2017). Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption. *Int. J. Epidemiol.* 46, 1985–1998. doi: 10.1093/ije/dyx102

Hasan, M. M., Shoombuatong, W., Kurata, H., and Manavalan, B. (2021). Critical evaluation of web-based DNA N6-methyladenine site prediction tools. *Brief. Funct. Genomics* 20, 258–272. doi: 10.1093/bfgp/elaa028

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., et al. (2018). The MR-Base platform supports systematic causal inference across the human phenome. *Elife* 7:e34408.

Hu, Y., Sun, J. Y., Zhang, Y., Zhang, H., Gao, S., Wang, T., et al. (2021). rs1990622 variant associates with Alzheimer's disease and regulates TMEM106B expression in human brain tissues. *BMC Med.* 19:11. doi: 10.1186/s12916-020-01883-5

Hu, Y., Zhang, H., Liu, B., Gao, S., Wang, T., Han, Z., et al. (2020). rs34331204 regulates TSPAN13 expression and contributes to Alzheimer's disease with sex differences. *Brain* 143:e95. doi: 10.1093/brain/awaa302

Jansen, H., Samani, N. J., and Schunkert, H. (2014). Mendelian randomization studies in coronary artery disease. *Eur. Heart J.* 35, 1917–1924. doi: 10.1093/eurheartj/ehu208

Kitas, G. D., and Gabriel, S. E. (2011). Cardiovascular disease in rheumatoid arthritis: state of the art and future perspectives. *Ann. Rheum. Dis.* 70, 8–14. doi: 10.1136/ard.2010.142133

Larsson, S. C., Back, M., Rees, J. M. B., Mason, A. M., and Burgess, S. (2020). Body mass index and body composition in relation to 14 cardiovascular conditions in UK Biobank: a mendelian randomization study. *Eur. Heart J.* 41, 221–226. doi: 10.1093/eurheartj/ehz388

Larsson, S. C., Burgess, S., and Michaelsson, K. (2017). Association of genetic variants related to serum calcium levels with coronary artery disease and myocardial infarction. *JAMA* 318, 371–380. doi: 10.1001/jama.2017.8981

Laslett, L. J., Alagona, P., Clark, B. A., Drozda, J. P., Saldivar, F., Wilson, S. R., et al. (2012). The worldwide environment of cardiovascular disease: prevalence, diagnosis, therapy, and policy issues a report from the American College of Cardiology. *J. Am. Coll. Cardiol.* 60, S1–S49.

Leong, D. P., Joseph, P. G., Mckee, M., Anand, S. S., Teo, K. K., Schwalm, J. D., et al. (2017). Reducing the global burden of cardiovascular disease, part 2 Prevention and treatment of cardiovascular disease. *Circ. Res.* 121, 695–710. doi: 10.1161/circresaha.117.311849

Li, P., Guo, M., Wang, C., Liu, X., and Zou, Q. (2015). An overview of SNP interactions in genome-wide association studies. *Brief. Funct. Genomics* 14, 143–155. doi: 10.1093/bfgp/elu036

Ong, J. S., Dixon-Suen, S. C., Han, X., An, J., Esophageal Cancer Consortium, 23 and Me Research Team, et al. (2021). A comprehensive re-assessment of the association between vitamin D and cancer susceptibility using mendelian randomization. *Nat. Commun.* 12:246.

Pierce, B. L., and Burgess, S. (2013). Efficient design for mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *Am. J. Epidemiol.* 178, 1177–1184. doi: 10.1093/aje/kwt084

Qiu, S., Cao, P., Guo, Y., Lu, H., and Hu, Y. (2021). Exploring the causality between hypothyroidism and non-alcoholic fatty liver: a mendelian randomization study. *Front. Cell Dev. Biol.* 9:643582. doi: 10.3389/fcell.2021.643582

Rosoff, D. B., Smith, G. D., Mehta, N., Clarke, T. K., and Lohoff, F. W. (2020). Evaluating the relationship between alcohol consumption, tobacco use, and cardiovascular disease: a multivariable mendelian randomization study. *PLoS Med.* 17:e1003410. doi: 10.1371/journal.pmed.1003410

Smith, G. D., Paternoster, L., and Relton, C. (2017). When will mendelian randomization become relevant for clinical practice and public health? *JAMA* 317, 589–591. doi: 10.1001/jama.2016.21189

Smith, S. C., Collins, A., Ferrari, R., Holmes, D. R., Logstrup, S., Mcghie, D. V., et al. (2012). Our time: a call to save preventable death from cardiovascular disease (heart disease and stroke). *Circulation* 126, 2769–2775. doi: 10.1161/cir.0b013e318267e99f

Sokka, T., Abelson, B., and Pincus, T. (2008). Mortality in rheumatoid arthritis: 2008 update. *Clin. Exp. Rheumatol.* 26, S35–S61.

Sun, L., Liu, G., Su, L., and Wang, R. (2019). HS-MMGKG: a fast multi-objective harmony search algorithm for two-locus model detection in GWAS. *Curr. Bioinform.* 14, 749–761. doi: 10.2174/1574893614666190409110843

Wei, L., He, W., Malik, A., Su, R., Cui, L., and Manavalan, B. (2021). Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework. *Brief. Bioinform.* 22:bbaa275.

Wei, L., Su, R., Luan, S., Liao, Z., Manavalan, B., Zou, Q., et al. (2019). Iterative feature representations improve N4-methylcytosine site prediction. *Bioinformatics* 35, 4930–4937. doi: 10.1093/bioinformatics/btz408

Xu, L., Huang, J., Zhang, Z., Qiu, J., Guo, Y., Zhao, H., et al. (2019). Bioinformatics study on serum triglyceride levels for analysis of a potential risk factor affecting blood pressure variability. *Curr. Bioinform.* 14, 376–385. doi: 10.2174/1574893614666190109152809

Yavorska, O. O., and Burgess, S. (2017). MendelianRandomization: an R package for performing mendelian randomization analyses using summarized data. *Int. J. Epidemiol.* 46, 1734–1739. doi: 10.1093/ije/dyx034

Yeung, S. L. A., Luo, S., and Schooling, C. M. (2018). The impact of glycated hemoglobin (HbA(1c)) on Cardiovascular disease risk: a mendelian randomization study using UK Biobank. *Diabetes Care* 41, 1991–1997. doi: 10.2337/dc18-0289

Zhuang, H., Zhang, Y., Yang, S., Cheng, L., and Liu, S. L. (2019). A Mendelian randomization study on infant length and type 2 diabetes mellitus risk. *Curr. Gene Ther.* 19, 224–231. doi: 10.2174/1566523219666190925115535

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Genome-wide Identification and Analysis of Splicing QTLs in Multiple Sclerosis by RNA-Seq Data

Yijie He[†]*, Lin Huang*[†]*, Yaqin Tang, Zeyuan Yang and Zhijie Han**

Department of Bioinformatics, School of Basic Medicine, Chongqing Medical University, Chongqing, China

Multiple sclerosis (MS) is an autoimmune disease characterized by inflammatory demyelinating lesions in the central nervous system. Recently, the dysregulation of alternative splicing (AS) in the brain has been found to significantly influence the progression of MS. Moreover, previous studies demonstrate that many MS-related variants in the genome act as the important regulation factors of AS events and contribute to the pathogenesis of MS. However, by far, no genome-wide research about the effect of genomic variants on AS events in MS has been reported. Here, we first implemented a strategy to obtain genomic variant genotype and AS isoform average percentage spliced-in values from RNA-seq data of 142 individuals (51 MS patients and 91 controls). Then, combing the two sets of data, we performed a *cis*-splicing quantitative trait loci (sQTLs) analysis to identify the *cis*-acting loci and the affected differential AS events in MS and further explored the characteristics of these *cis*-sQTLs. Finally, the weighted gene coexpression network and gene set enrichment analyses were used to investigate gene interaction pattern and functions of the affected AS events in MS. In total, we identified 5835 variants affecting 672 differential AS events. The *cis*-sQTLs tend to be distributed in proximity of the gene transcription initiation site, and the intronic variants of them are more capable of regulating AS events. The retained intron AS events are more susceptible to influence of genome variants, and their functions are involved in protein kinase and phosphorylation modification. In summary, these findings provide an insight into the mechanism of MS.

Keywords: multiple sclerosis, alternative splicing, RNA-seq, splicing quantitative trait loci, function analysis

OPEN ACCESS

Edited by:
Liangcai Zhang,
Janssen Research and Development,
United States

Reviewed by:
Qingxia Yang,
Nanjing University of Posts and
Telecommunications, China
Zhaoxia Yu,
University of California, Irvine,
United States

****Correspondence:***
Zhijie Han
zhijiehan@cqmu.edu.cn

[†]These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to
Statistical Genetics and Methodology,
a section of the journal
Frontiers in Genetics

Received: *02 September 2021*
Accepted: *18 October 2021*
Published: *12 November 2021*

Citation:
He Y, Huang L, Tang Y, Yang Z and Han Z (2021) Genome-wide Identification and Analysis of Splicing QTLs in Multiple Sclerosis by RNA-Seq Data.
Front. Genet. 12:769804.
doi: 10.3389/fgene.2021.769804

INTRODUCTION

Multiple sclerosis (MS) is a serious autoimmune disease of central nervous system (CNS) and is characterized by inflammatory demyelinating lesions in the white matter (Compston and Coles, 2008). According to the most recent survey in 2020 (the *Atlas of MS* investigation), the estimated number of the people affected by MS has reached approximately 2.8 million worldwide (Walton et al., 2020). Similar to most of the complex diseases, genetic factors are the major contributors to the individual differences in MS susceptibility, and the role of genetic variants and transcriptional regulation in MS may be the key to understanding its pathogenesis (Fugger et al., 2009; Olsson et al., 2017; Yang et al., 2019).

Recently, alternative splicing (AS), a process that enables a gene to generate different transcript isoforms, has been found to have the characteristic of high complexity and play an important role in primates and human CNS (Barbosa-Morais et al., 2012; Merkin et al., 2012; GTEx Consortium, 2015;

GTEx Consortium, 2020). Further, previous studies demonstrate that dysregulation of AS events in genes significantly influences the progression of many nervous system diseases, including MS. For example, the RNA helicase DDX39B, a repressor of AS of IL7R exon 6, is downregulated in MS peripheral blood mononuclear cells, and consequently, the overexpression of the soluble form of the interleukin-7 receptor alpha chain gene (sIL7R) increases MS risk (Galarza-Munoz et al., 2017). Inclusion of AS4 exon in Nrxn 1-3 is significantly increased in the prefrontal cortex of a murine MS model, and the abnormal splicing promotes the expression of IL-1β, which is an important mediator of inflammation and leading to cognitive dysfunction in MS (Marchese et al., 2021). The dysregulated AS of the A1β transcript results in a significantly diminished adenosine A1 receptor protein, which is an important therapeutic target in the treatment of MS in peripheral blood mononuclear cells and brain tissue of MS patients (Johnston et al., 2001).

Moreover, previous studies demonstrate that genetic variants can control the regulation of AS events by directly altering nucleotide sequences in the splice site or as splicing quantitative trait loci (sQTLs) in a genome-wide manner (Battle et al., 2014; GTEx Consortium, 2015; Takata et al., 2017; GTEx Consortium, 2020). For MS, numerous disease-related risk single nucleotide polymorphisms (SNPs) have been identified by genome-wide association studies (GWAS) (International Multiple Sclerosis Genetics Consortium et al., 2013; Sawcer et al., 2014; Patsopoulos, 2018), and a part of them as the regulation factors of AS events can contribute to the pathogenesis of MS. For instance, MS risk variants rs35476409 and rs61762387 can affect the splicing of exon 3 of the PRKCA gene, which is considered to be a functional contributor to MS predisposition (Paraboschi et al., 2014). Another MS risk SNP rs6897932 locates in the functional AS exon of IL7R. Through disrupting the exonic splicing silencer, it can increase skipping of IL7R exon 6 to produce more soluble and membrane-bound isoforms of IL7R protein (IL7Ra), which is a key factor in the immune response pathway of MS (Gregory et al., 2007). The SNP rs3130253, located within the MOG gene, has a proven genetic susceptibility to MS. The minor allele (A) of rs3130253 is associated with the increased splicing of MOG exon 2 to 3 in the oligodendrocyte cell (1.7-fold) and influences the extracellular and transmembrane domains of MOG to induce the development of MS (Jensen et al., 2010). Although these findings provide valuable insights into the direct influence of SNPs on AS events in MS, the profile and function of sQTLs throughout the genome remain poorly understood.

Our previous studies systematically describe the influence of genomic variants on gene expression in a genome-wide manner and find that this impact is more significant among the regions of long intergenic noncoding RNA for MS (Han et al., 2018; Han et al., 2020). However, by far, the genome-wide research about the effect of these genomic variants on AS events in MS has been not yet reported. To solve this problem, in this study, we used the blood RNA-seq data from 51 MS patients and 91 controls of European descent that have been previously successfully used for our expression quantitative trait loci (eQTLs) analysis (Han et al., 2020). Particularly, we first comprehensively detected the AS events on a whole-genome scale and performed a differential splicing analysis between the MS patients and healthy individuals by using the RNA-seq data. Then, based on the same data, we genotyped the large-scale genomic variants (mainly the SNPs) in the entire human genome. According to the previous studies, genotyping using RNA-seq can be effectively performed in a lower sample scale (typically tens to hundreds of individuals) and higher genetic heterogeneity and is more conducive to the discovery of functional SNPs than the traditional approaches (e.g., SNP arrays) (Wang et al., 2009; Davey et al., 2011). Next, combining the data of AS isoform average percentage spliced-in (PSI) and genomic variant genotype, we performed a sQTL analysis to identify the *cis*-acting loci and the affected AS events in MS. Further, we explored the distribution characteristics and disease specificity of these *cis*-sQTL loci. Finally, we conducted the weighted gene coexpression network analysis (WGCNA) and gene set enrichment analysis (GSEA) to investigate the interaction pattern of the AS affected genes and the functions of these genes to the pathogenesis of MS. The flow chart is shown in **Figure 1**.

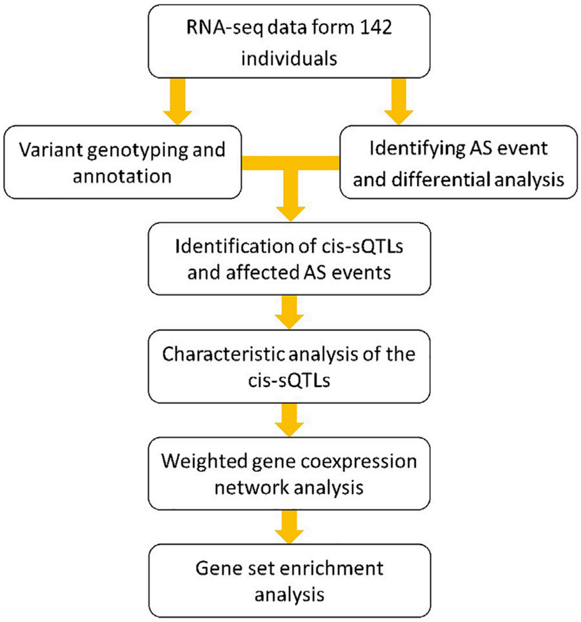

RNA-seq data form 142 individuals
Variant genotyping and annotation
Identifying AS event and differential analysis
Identification of cis-sQTLs and affected AS events
Characteristic analysis of the cis-sQTLs
Weighted gene coexpression network analysis
Gene set enrichment analysis

FIGURE 1 | The flow chart of the study design for exploring the influence of genome variants on gene AS and their functions to pathogenesis of MS.

MATERIALS AND METHODS

Sample Collection and Preprocessing

A total of 142 individuals, including 51 MS patients and 91 age- and gender-matched healthy controls, were selected from the Utrecht Medical Center (UMCU) and VU University Medical Center (VUMC) of Netherlands. The RNA-seq data of blood samples from these individuals were used for this study (**Table 1**). The details are described in previous studies (Best et al., 2017;

TABLE 1 | Summary of the 142 individuals studied in this work.

Individuals	Institution	Ethnicity	Sample size	Mean age (s.d.)	Male/female (%)
MS patients	VUMC	European	51	46.14 (7.54)	25.5/74.5
Healthy controls	VUMC and UMCU	European	91	46.92 (8.50)	34.1/65.9
Total			142	46.64 (8.18)	31.0/69.0

VUMC, VU University Medical Center; Amsterdam, Netherlands; UMCU, Utrecht Medical Center, Utrecht, Netherlands. This information is also described in our previous study (Han et al., 2020).

Han et al., 2020). Briefly, the mirVana miRNA isolation kit was used to extract the total RNA of these samples. The Truseq Nano DNA Sample Preparation Kit and Illumina Hiseq 2500 platform were used for library preparation and sequencing, respectively. After the RNA read quality control, these sequence data were stored in the NCBI Sequence Read Archive (SRA) database (SRP093349). We used the SRA Toolkit software to download these sequence data and converted them into FASTQ files.

Variant Genotyping and Annotation

The procedure of variant genotyping and annotation on a whole-genome scale using FASTQ files has been described in our previous study (Han et al., 2020). Briefly, the BWA software was first used to align the sequenced reads to the human reference genome (hg19) with its default parameter settings and generated the sequence alignment/map (SAM) files (Li and Durbin, 2009). Then, the SAMtools and BCFtools software were used with their default parameter settings to perform the format conversion of these SAM files and variant calling, respectively (Li, 2011; Li et al., 2009). The genotyped variants were stored in the VCF file. Further, based on the annotation databases, refGene (about the functional information of variants) (Pruitt et al., 2007) and snp138 of dbSNP (about the genomic position and ID of variants) (Day, 2010), we used the ANNOVAR software to annotate these genotyped variants (Yang and Wang, 2015). Finally, we preformed quality control, which is based on the sequencing quality and variant annotation. We conducted a Hardy-Weinberg equilibrium (HWE) test using the R package 'Genetics' (https://cran.r-project.org/web/packages/genetics/index.html).

According to the findings of previous studies (Greif et al., 2011; Quinn et al., 2013), we filtered the low-quality genotyped variants if their HWE p value $<5 \times 10^{-5}$ or root mean square (RMS) mapping quality <10 or read depth (DP) < 10 or minor allele frequency (MAF) < 1%. Moreover, other studies suggest that only the results catalogued in dbSNP should be retained to reduce the false positives when performing the SNP calling (Chepelev et al., 2009; Cirulli et al., 2010; Liu et al., 2012; Xu et al., 2013). Therefore, we further removed genotyped variants that are not catalogued in dbSNP according to the annotation results.

Identification and Differential Analysis of AS Events

Based on the RNA-seq data of the same samples, we used the vast-tools software to detect the AS events and calculate their PSI values on a whole-genome scale (Irimia et al., 2014). In particular, we first aligned the sequenced fragments to human reference genome (hg19) using the align tool module of vast-tools software with its default parameters to identify AS events and calculate their PSI values in each sample. Then, the results (five subfiles for each AS event) were merged using the combine tool module of vast-tools software to generate a file containing PSI of each AS event and quality control content for all samples. The quality control threshold is according to quality scores in the merged file, i.e., the mapped reads >10. Next, we used the multiple imputation method with the generalized linear model to impute missing PSI values of each AS event by the R package "mice" (Van Buuren and Groothuis-Oudshoorn, 2011) and counted the number of each type of AS events. Finally, based on the PSI values, we used the diff tool module of vast-tools software with its default parameters to perform a Bayesian inference-based differential AS analysis. The threshold of significance was set at the minimum value for absolute value of differential PSI between MS cases and controls (MV|ΔPSI|) at 0.95 confidence level greater than 10% according to the previous studies (Fagg et al., 2020; Ha et al., 2021; Hekman et al., 2021).

Identification of *cis*-s Quantitative Trait Loci and Characteristic Analysis

Combining the PSI values of AS events and the data of the genomic variant genotype from the same samples, we performed an sQTL analysis to identify the *cis*-acting loci and the affected AS events. Particularly, according to previous studies (GTEx Consortium, 2015; GTEx Consortium, 2020), we first considered it as the cis region where the distance between variants and transcription initiation site (TSS) of AS event corresponding genes less than 1 M, and selected all the suitable variant and AS event pairs for the *cis*-sQTL analysis. The genomic locations of the variants and the TSS of AS event corresponding genes are based on the annotation files of the dbSNP (snp138) and Ensembl databases (release 75), respectively. Then, we used the genotype data of the variants in combination with the PSI values of AS events to perform the sQTL analysis by the R package "Matrix eQTL" with a linear regression model (Shabalin, 2012). The parameters age and gender were used as the covariates. The threshold of significance level was set at a false discovery rate (FDR) q value <0.05. The p values are corrected for multiple testing by the Benjamini–Hochberg method. Finally, we calculated the percentage of various types of the *cis*-sQTL variants and the affected AS events, respectively, and compared them with the original proportion using a two-tailed Fisher exact test (the threshold of p < .05). Moreover, we further explored the

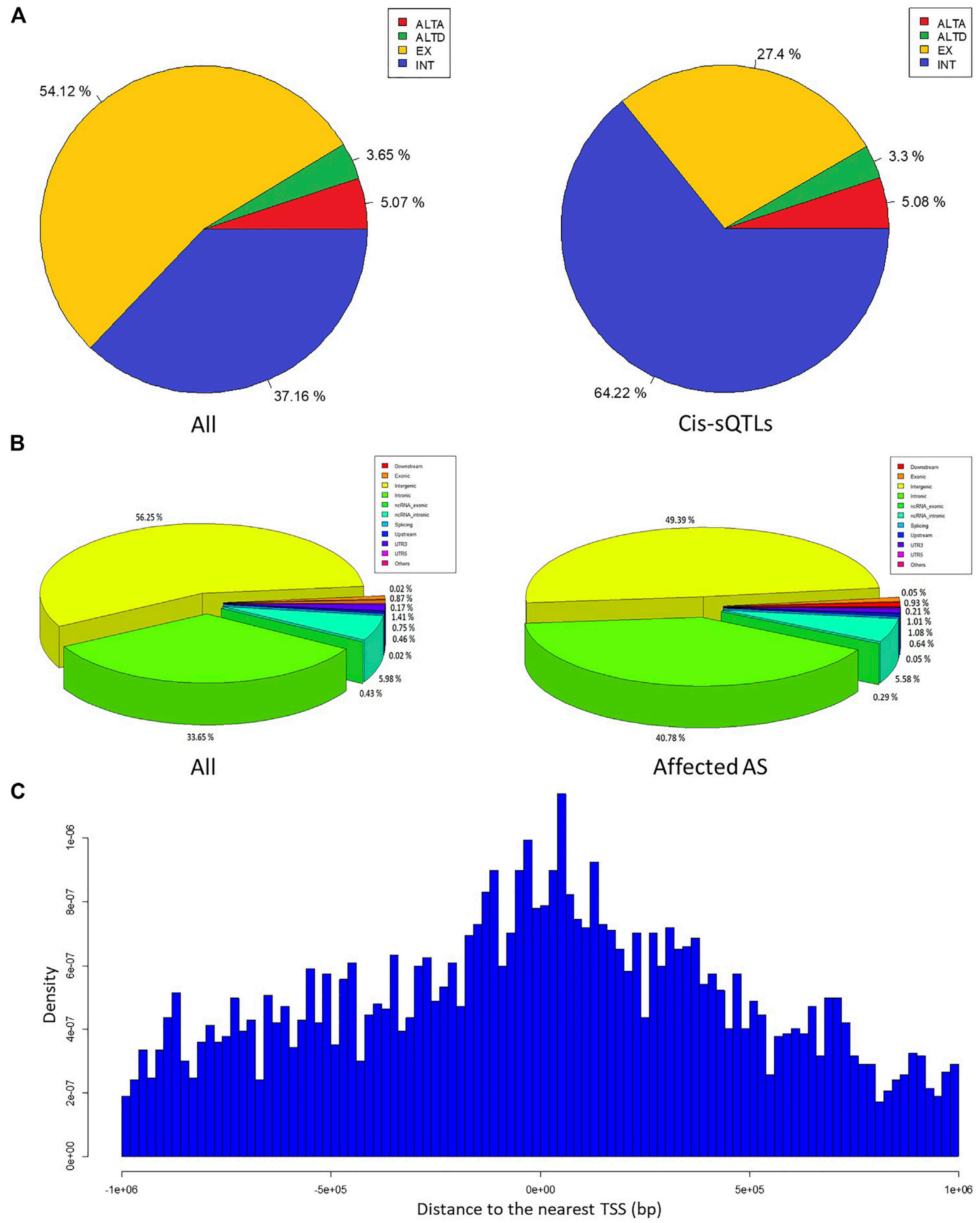


FIGURE 2 | The characteristic of the *cis*-sQTL variants and the affected AS events. **(A)** The pie charts show the percentage of all variants (left) and *cis*-sQTL variants (right) annotated with each class (intergenic, intronic, exonic, ncRNA intronic, ncRNA exonic, 5′/3′-UTR, upstream/downstream, splicing site, and others), respectively. **(B)** The pie charts show the proportion in all AS events (left) and affected AS events (right) annotated with each class (EX, INT, ALTA, and ALTD), respectively. **(C)** The bar graph indicates the relationship between the abundance of the *cis*-sQTL variants and the distance of them to the nearest TSS of AS events corresponding genes.

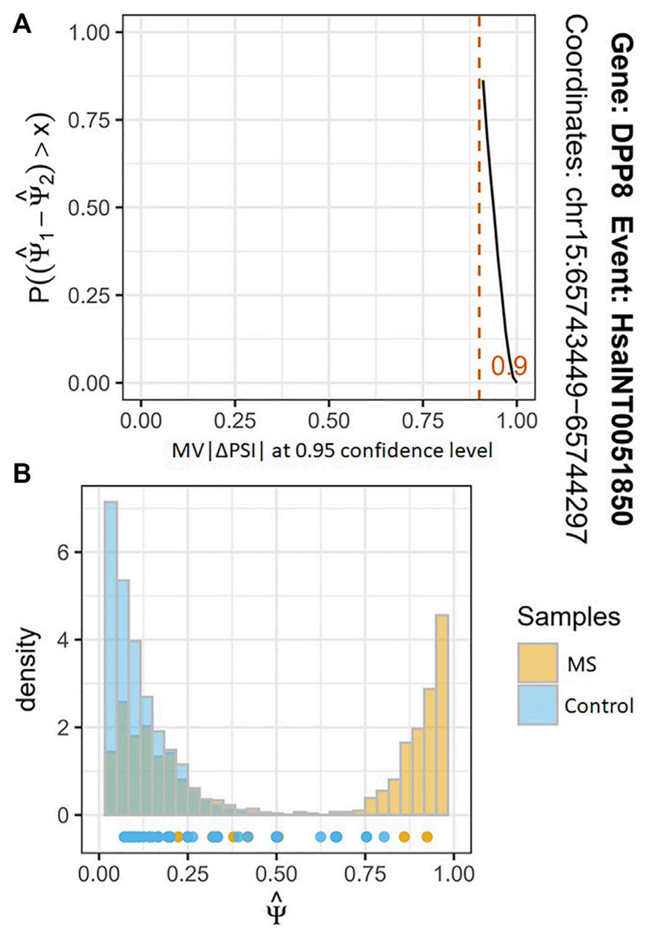


FIGURE 3 | The results of differential analysis of AS event HsaINT0051850. (**A**) The *x*-axis represents MV|ΔPSI | at a 95% confidence level. The *y*-axis represents the probability of ΔPSI being greater than some magnitude value of x. The red line indicates that the maximum probability of ΔPSI of AS event HsaINT0051850 between MS cases and controls is greater than 0.90. (**B**) The histogram shows the two joint posterior distributions over PSI and the points below the histograms estimate for each replicate.

relationship between the abundance of the *cis*-sQTL variants and the distance of them to the nearest TSS.

Weighted Gene Coexpression Network Analysis and Gene Set Enrichment Analysis

To explore the interaction pattern of the AS affected genes and their functions to the pathogenesis of MS, we performed the WGCNA and GSEA in turn. Particularly, we first downloaded the gene expression count data of the 51 MS patients and 91 healthy individuals from Gene Expression Omnibus (GEO) data set GSE89843 (Best et al., 2017) and carried out a standardized processing of these data using the "preprocess" function of R package "caret" (https://cran.r-project.org/web/packages/ caret/). Then, we conducted quality control to identify the outlier samples using the "hclust" function of R package "WGCNA" (Langfelder and Horvath, 2008). Further, to ensure the scale-free topology criterion of the coexpression network, we used the "pickSoftThreshold" function of R package "WGCNA" to choose the satisfactory soft threshold power β. Next, based on the satisfactory soft threshold power β, we used Pearson's method to calculate the weighted correlation of gene pairs in an adjacency matrix and used the dynamic cut-tree algorithm to construct the hierarchical clustering dendrogram by the R package "WGCNA." Finally, we calculated the correlation between the module membership and the importance of genes in this module to clinical traits to assess the relationship between the coexpression module and the clinical traits (including gender, age, and disease status) by the R package "WGCNA."

We further use the genes in the modules that are significantly associated with MS disease status to perform GSEA by DAVID software (Jiao et al., 2012). The default background of DAVID, i.e., three pathway data sets (BBID, BIOCARTA, and KEGG_PATHWAY), three gene ontology data sets (GOTERM_BP_DIRECT, GOTERM_CC_DIRECT, and GOTERM_MF_DIRECT), three functional categories (COG_ONTOLOGY, UP_KEYWORDS, and UP_SEQ_FEATURE), three protein domains (INTERPRO, PIR_SUPERFAMILY, and SMART), and one disease data set (OMIM_DISEASE) for the GSEA. The threshold of significance was set at FDR q < 0.05. The other parameters were set according to the default values of the DAVID software.

RESULTS AND DISCUSSION

Variant Genotyping by RNA-Seq Data

We obtained a total of about 3.2 billion sequenced reads from the blood RNA-seq data of 51 MS patients and 91 healthy controls. Based on these RNA-seq data, we aligned the sequenced reads to human reference genome (hg19) using BWA software and used these aligned reads to call the variant genotypes by SAMtools and BCFtools software. After quality control based on DP, RMS mapping quality, MAF, HWE, and dbSNP catalog, we obtained 620,339 genotyped variants. Finally, the results of annotation using ANNOVAR software showed that a total of 600,872 genotyped SNPs and 19,467 indels are included in these genotyped variants, and approximately 56.25%, 33.65%, 0.87%, 5.98%, 0.43%, 1.58%, 1.21%, and 0.02% of them are categorized into the intergenic, intronic, exonic, ncRNA intronic, ncRNA exonic, 5′/3′-UTR, upstream/downstream, and splicing site classes, respectively. These findings reveal an uneven distribution of these variants in the genome (**Figure 2A**).

Identification and Differential Analysis of Alternative Splicing Events

Based on the FASTQ files from the same samples, we used the corresponding tool modules of vast-tools software to identify the AS event with their PSI values and performed the differential analysis of them. After the quality control, we found a total of 2272 significant differential AS events between the MS cases and healthy individuals (MV|ΔPSI| at 0.95 confidence level ≥10%) from the more than seven million identified AS events. These differential AS events are involved

TABLE 2 | The top 30 significant results of the sQTL variants and the differential AS events affected by them.

SNP ID	Position	Gene	Ensembl ID	AS event	TSS	Beta	*p* Value	FDR q value
rs1950969	94236929	GOLGA5	ENSG00000066455	HsaEX0027985	93260576	34.2500	0.00E + 00	1.05E-303
rs1950970	94236975	GOLGA5	ENSG00000066455	HsaEX0027985	93260576	34.2500	0.00E + 00	1.05E-303
rs8017818	93651054	GOLGA5	ENSG00000066455	HsaEX0027985	93260576	–34.2500	0.00E + 00	1.05E-303
rs12226058	43190629	ACCSL	ENSG00000205126	HsaEX6001613	44069531	–12.7000	0.00E + 00	1.05E-303
rs12795809	43190576	ACCSL	ENSG00000205126	HsaEX6001613	44069531	12.7000	0.00E + 00	1.05E-303
rs61690000	43523415	ACCSL	ENSG00000205126	HsaEX6001613	44069531	–12.7000	0.00E + 00	1.05E-303
rs72898940	43315617	ACCSL	ENSG00000205126	HsaEX6001613	44069531	–12.7000	0.00E + 00	1.05E-303
rs74545163	43424312	ACCSL	ENSG00000205126	HsaEX6001613	44069531	–12.7000	0.00E + 00	1.05E-303
rs7931142	43189976	ACCSL	ENSG00000205126	HsaEX6001613	44069531	–12.7000	0.00E + 00	1.05E-303
rs890245	43201830	ACCSL	ENSG00000205126	HsaEX6001613	44069531	12.7000	0.00E + 00	1.05E-303
rs113384165	26738788	NSMCE1	ENSG00000169189	HsaEX6042948	27280115	–40.0000	0.00E + 00	1.05E-303
rs6498005	27270200	NSMCE1	ENSG00000169189	HsaEX6042948	27280115	–40.0000	0.00E + 00	1.05E-303
rs7187853	27267403	NSMCE1	ENSG00000169189	HsaEX6042948	27280115	–40.0000	0.00E + 00	1.05E-303
rs2976708	125398800	SNX4	ENSG00000114520	HsaEX6058167	125239041	5.1100	0.00E + 00	1.05E-303
rs543453	3139759	PIAS4	ENSG00000105229	HsaEX6091950	4007748	1.2500	0.00E + 00	1.05E-303
rs644193	3139715	PIAS4	ENSG00000105229	HsaEX6091950	4007748	1.2500	0.00E + 00	1.05E-303
rs16949296	45984949	SCRN2	ENSG00000141295	HsaEX6023334	45918699	–66.6580	3.86E-238	1.75E-233
rs11643492	2791938	SRRM2	ENSG00000167978	HsaEX6041902	2802330	52.7000	1.46E-172	6.04E-168
rs2858609	49620817	PIM3	ENSG00000198355	HsaEX6027387	50354161	22.1493	2.47E-124	9.47E-120
rs73179160	50082716	PIM3	ENSG00000198355	HsaEX6027387	50354161	–11.0746	2.47E-124	9.47E-120
rs11671147	8227499	ELAVL1	ENSG00000066044	HsaEX0022092	8070529	49.6873	2.55E-56	7.47E-52
rs62638003	7908051	ELAVL1	ENSG00000066044	HsaEX0022092	8070529	49.6873	2.55E-56	7.47E-52
rs216272	3013971	PIAS4	ENSG00000105229	HsaEX6091950	4007748	–0.5013	5.37E-50	1.53E-45
rs57414916	141780202	EIF2C2	ENSG00000123908	HsaEX6082596	141645718	–5.0980	7.11E-50	1.97E-45
rs2020857	15030752	USP9Y	ENSG00000114374	HsaEX0070061	14813160	–8.2372	1.40E-49	3.66E-45
rs138123250	105087582	CALHM2	ENSG00000138172	HsaEX6090238	105212660	–8.8666	1.27E-45	3.16E-41
rs12610435	8021331	ELAVL1	ENSG00000066044	HsaEX0022092	8070529	42.5593	5.34E-45	1.30E-40
rs192519226	48400006	XYLT2	ENSG00000015532	HsaEX6023498	48423453	–22.2205	2.90E-44	6.55E-40

in 1542 genes (**Supplementary Table S1**). **Figure 3** shows the most significant differential AS event HsaINT0051850 of DPP8 gene (MV|ΔPSI| at 0.95 confidence level = 0.90). According to the classification criteria of vast-tools, the types of AS events contain alternative exon skipping (EX), retained intron (INT), alternative splice site acceptor choice (ALTA), and alternative splice site donor choice (ALTD). We found that approximately 54.12%, 37.16%, 5.07%, and 3.65% of these identified AS events are categorized into EX, INT, ALTA, and ALTD classes, respectively, which also revealed an uneven distribution of them (**Figure 2B**).

Identification of *Cis*-s Quantitative Trait Loci and Characteristic Analysis

Combining the PSI values of AS events with the genotype data of the genomic variant in the cis region from the same samples, we used a linear regression model to perform the *cis*-sQTL analysis by R package "Matrix eQTL" with the parameters age and gender serving as covariates. In total, we identified 5835 variants affecting 672 AS events (involving 482 genes) of all these 2272 significant differential AS events with a significance level of q < 0.05. The top 30 significant results are shown in **Table 2** (the full information is presented in **Supplementary Table S2**). Further, we found that approximately 49.39%, 40.78%, 0.93%, 5.58%, 0.29%, 1.22%, 1.72%, and 0.05% of the *cis*-sQTL variants are categorized into the intergenic, intronic, exonic, ncRNA intronic, ncRNA exonic, 5’/3′-UTR, upstream/downstream, and splicing site classes, respectively (**Figure 2A**), and approximately 27.40%, 64.22%, 5.08%, and 3.30% of the affected AS events are categorized into EX, INT, ALTA, and ALTD classes, respectively (**Figure 2B**). By the two-tailed Fisher exact test, we found that the percentage of main types both in the *cis*-sQTL variants and the affected AS events show a significant difference compared with the original proportion. Particularly, the percentage of the *cis*-sQTL intergenic variants is 49.39%, but its original proportion in all of the variants is 56.25% (odds ratio (OR) = 0.76, $p = 1.84 \times 10^{-45}$); the percentage of the *cis*-sQTL intronic variants is 40.78%, but its original proportion in all of the variants is only 33.65% (OR = 1.36, $p = 1.09 \times 10^{-52}$); the percentage of the affected EX events is 27.40%, but its original proportion in all AS events is 54.12% (OR = 0.32, $p = 9.65 \times 10^{-69}$); the percentage of the affected INT events is 64.22%, but its original proportion in all AS events is only 37.16% (OR = 3.03, $p = 5.12 \times 10^{-70}$). This reveals a specific regulation of the AS events by variants in MS. Moreover, we also found that these *cis*-sQTL variants tend to be distributed in the proximity of the TSS of AS events corresponding genes (**Figure 2C**).

Weighted Gene Coexpression Network Analysis for Affected Alternative Splicing Events Corresponding Genes

We performed WGCNA to explore the characteristics of the affected AS event corresponding genes in MS. According to the sample clustering results for quality control, we removed eight outlier

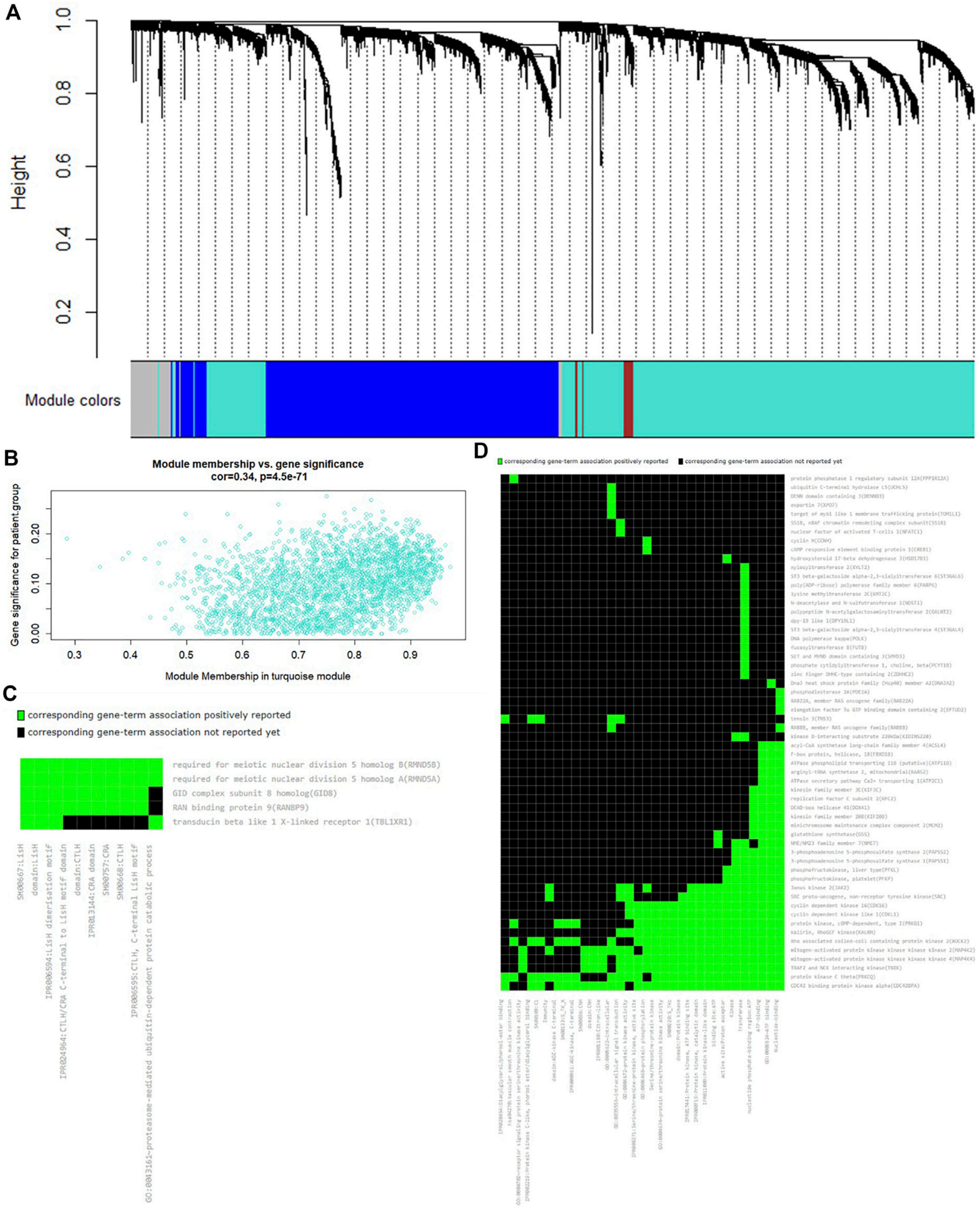


FIGURE 4 | The results of WGCNA and GSEA. **(A)** The expression clustering dendrogram of all 4722 genes in the GSE89843 data set. There are four clustered modules in the hierarchical clustering dendrogram, which contain 360 of 482 affected AS events corresponding genes. These clustered modules are marked as four different colors, respectively, i.e., turquoise, blue, brown, and grey. **(B)** The correlation between the module membership and the gene significance in the turquoise module, which reveals a relatively strong correlation with the disease status (cor = 0.34 and $p = 4.5 \times 10^{-71}$). The gene significance is defined as the correlation between a single gene expression and sample trait (e.g., gender, age, and disease status) **(C)** The annotation cluster 1 contains 10 functionally highly similar enriched terms involved in the protein–protein interaction domain motif. (**D**) The annotation cluster 2 contains 32 functionally highly similar enriched terms involved in protein kinase and phosphorylation modification. This figure can be viewed more clearly by enlarging in the electronic version.

samples (**Supplementary Figure S1**). Then, we found that the model fitting index R-squared reaches 0.85 for the first time, and the mean connectivity approaches zero simultaneously when the soft threshold power β equals 12 (**Supplementary Figure S2**). Therefore, we calculated the weighted correlation of gene pairs and constructed the coexpression network using the R package "WGCNA" with the parameter β = 12. The results show that there is a total of four modules (i.e., MEturquoise, MEblue, MEbrown, and MEgrey) in the coexpression network. The modules are defined as clusters in which the densely interconnected genes are coexpressed with each other. The unsupervised clustering analysis with a topological overlap index was used to measure the network interconnectedness. They contain a total of 4722 clustered genes according to their interconnectedness, and 360 of them belong to the affected AS event corresponding genes (**Figure 4A**). These AS affected genes are generally evenly distributed in the four modules according to their scale. The results of correlation analysis reveal some association of all the modules with individual gender or age ($p < .05$). Among them, however, only the turquoise module shows a relatively strong correlation with the disease status (cor = 0.34 and $p = 4.5 \times 10^{-71}$) (**Figure 4B**), which means that the interaction of the genes in the turquoise module is relevant to pathogenesis of MS. In the grey module, for example, the cor and p value are −.027 and .65, respectively.

Gene Set Enrichment Analysis of Alternative Splicing Affected Genes in Multiple Sclerosis–Related Module

Based on the results of WGCNA, we used the 198 AS affected genes in the MS-related turquoise module to perform the GSEA. According to the significance threshold FDR q < 0.05, we identified a total of 30 enriched terms. The most significant of them contain the AS-related terms, e.g., alternative splicing ($q = 2.0 \times 10^{-8}$) and splicing variant ($q = 1.0 \times 10^{-3}$), which are consistent with the findings of sQTL analysis. Most of the other significant enriched terms are involved in epigenetic modification, which is the common biological process associated with the pathogenesis of MS (**Supplementary Table S3**). Further, we performed a functional annotation clustering analysis of the enriched terms. We identified two annotation clusters with enrichment score more than 2, which contain 10 and 32 functionally highly similar terms, respectively. Particularly, annotation cluster 1 (enrichment score = 3.78) contains the protein–protein interaction domain (e.g., LisH, CTLH, and CRA) motif-related terms, which are the basic biological properties for eukaryotes (**Figure 4C**). The annotation cluster 2 (enrichment score = 2.12) contains protein kinase and phosphorylation modification terms, which are significantly associated with the pathogenesis of MS (**Figure 4D**). For example, Feng *et al.* found that the type I interferons and the p38 MAP kinase can induce tyrosine and serine phosphorylation of STAT1 in MS patients, respectively, and the excessive phosphorylation of STAT1 can induce inflammatory cytokines and demyelination to aggravate the development of MS (Feng et al., 2002). Trinschek *et al.* found that phosphorylation of protein kinase B/c-Akt in MS autoaggressive T effector cells (Teff) is able to induce the unresponsiveness of the $CD4^+$ and $CD8^+$ course independent MS-Teff by stimulation of the active regulatory T cells and thereby lead to the ineffective treatment of MS (Trinschek et al., 2013). Delgado-Roche *et al.* found that ozone therapy can promote the phosphorylation of the transcriptional factor NF-E2-related factor 2 through upregulating the expression of MAP kinase CK2, which can reduce oxidative stress and pro-inflammatory cytokines in MS (Delgado-Roche et al., 2017).

CONCLUSIONS

In this study, based on the MS RNA-seq data, we genotyped 620,339 variants and identified 2272 significant differential AS events in the same samples. Then, combing the two sets of data, we performed a *cis*-sQTL analysis and identified 5835 variants affecting 672 differential AS events in MS. Further, the results of characteristic analysis showed that the intronic variants are more capable of regulating AS events, and INT AS events are more susceptible to the influence of genome variants. Moreover, the *cis*-sQTL variants tend to be distributed in the proximity of the TSS of AS events corresponding genes. Finally, the results of WGCNA and GSEA demonstrate that the regulation of AS by genome variants are important to MS and their potential function may be involved in protein–protein interaction domain motif protein phosphorylation modification. All in all, we performed a strategy to explore the regulation of AS by genome variants in MS by RNA-seq data, and these findings will benefit the improvement of understanding MS pathogenesis.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

ZH designed the research. ZH, YH, LH, YT and ZY collected the data. YH, LH and ZH performed the research, analyzed data. YH, LH and ZH wrote the paper. ZH and YT reviewed and modified the article. All authors discussed the results and contributed to the final article. All authors read and approved the final article

FUNDING

This research is financially supported by the Start-up fund of Chongqing Medical University (R1017), and the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJQN202100402).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.769804/full#supplementary-material

REFERENCES

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., et al. (2012). The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. *Science* 338, 1587–1593. doi:10.1126/science.1230612

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., et al. (2014). Characterizing the Genetic Basis of Transcriptome Diversity through RNA-Sequencing of 922 Individuals. *Genome Res.* 24, 14–24. doi:10.1101/gr.155192.113

Best, M. G., Sol, N., In 't Veld, S. G. J. G., Vancura, A., Muller, M., Niemeijer, A.-L. N., et al. (2017). Swarm Intelligence-Enhanced Detection of Non-small-Cell Lung Cancer Using Tumor-Educated Platelets. *Cancer Cell* 32, 238–252. doi:10.1016/j.ccell.2017.07.004

Chepelev, I., Wei, G., Tang, Q., and Zhao, K. (2009). Detection of Single Nucleotide Variations in Expressed Exons of the Human Genome Using RNA-Seq. *Nucleic Acids Res.* 37, e106. doi:10.1093/nar/gkp507

Cirulli, E. T., Singh, A., Shianna, K. V., Ge, D., Smith, J. P., Maia, J. M., et al. (2010). Screening the Human Exome: a Comparison of Whole Genome and Whole Transcriptome Sequencing. *Genome Biol.* 11, R57. doi:10.1186/gb-2010-11-5-r57

Compston, A., and Coles, A. (2008). Multiple Sclerosis. *Lancet* 372, 1502–1517. doi:10.1016/S0140-6736(08)61620-7

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide Genetic Marker Discovery and Genotyping Using Next-Generation Sequencing. *Nat. Rev. Genet.* 12, 499–510. doi:10.1038/nrg3012

Day, I. N. M. (2010). dbSNP in the Detail and Copy Number Complexities. *Hum. Mutat.* 31, 2–4. doi:10.1002/humu.21149

Delgado-Roche, L., Riera-Romo, M., Mesta, F., Hernández-Matos, Y., Barrios, J. M., Martínez-Sánchez, G., et al. (2017). Medical Ozone Promotes Nrf2 Phosphorylation Reducing Oxidative Stress and Pro-inflammatory Cytokines in Multiple Sclerosis Patients. *Eur. J. Pharmacol.* 811, 148–154. doi:10.1016/j.ejphar.2017.06.017

Fagg, W. S., Liu, N., Braunschweig, U., Chen, X., Widen, S. G., Donohue, J. P., et al. (2020). Definition of Germ Cell Lineage Alternative Splicing Programs Reveals a Critical Role for Quaking in Specifying Cardiac Cell Fate. *bioRxiv*. doi:10.1101/2020.12.22.423880

Feng, X., Petraglia, A. L., Chen, M., Byskosh, P. V., Boos, M. D., and Reder, A. T. (2002). Low Expression of Interferon-Stimulated Genes in Active Multiple Sclerosis Is Linked to Subnormal Phosphorylation of STAT1. *J. Neuroimmunol.* 129, 205–215. doi:10.1016/s0165-5728(02)00182-0

Fugger, L., Friese, M. A., and Bell, J. I. (2009). From Genes to Function: the Next challenge to Understanding Multiple Sclerosis. *Nat. Rev. Immunol.* 9, 408–417. doi:10.1038/nri2554

Galarza-Muñoz, G., Briggs, F. B. S., Evsyukova, I., Schott-Lerner, G., Kennedy, E. M., Nyanhete, T., et al. (2017). Human Epistatic Interaction Controls IL7R Splicing and Increases Multiple Sclerosis Risk. *Cell* 169, 72–84. doi:10.1016/j.cell.2017.03.007

Gregory, S. G., Schmidt, S., Schmidt, S., Seth, P., Oksenberg, J. R., Hart, J., et al. (2007). Interleukin 7 Receptor α Chain (IL7R) Shows Allelic and Functional Association with Multiple Sclerosis. *Nat. Genet.* 39, 1083–1091. doi:10.1038/ng2103

Greif, P. A., Eck, S. H., Konstandin, N. P., Benet-Pagès, A., Ksienzyk, B., Dufour, A., et al. (2011). Identification of Recurring Tumor-specific Somatic Mutations in Acute Myeloid Leukemia by Transcriptome Sequencing. *Leukemia* 25, 821–827. doi:10.1038/leu.2011.19

GTEx Consortium (2015). Human Genomics. The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans. *Science* 348, 648–660. doi:10.1126/science.1262110

GTEx Consortium (2020). The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues. *Science* 369, 1318–1330. doi:10.1126/science.aaz1776

Ha, K. C. H., Sterne-Weiler, T., Morris, Q., Weatheritt, R. J., and Blencowe, B. J. (2021). Differential Contribution of Transcriptomic Regulatory Layers in the Definition of Neuronal Identity. *Nat. Commun.* 12, 335. doi:10.1038/s41467-020-20483-8

Han, Z., Qu, J., Zhao, J., and Zou, X. (2018). Genetic Variant Rs755622 Regulates Expression of the Multiple Sclerosis Severity Modifier D-Dopachrome Tautomerase in a Sex-specific Way. *Biomed. Res. Int.* 2018, 1–7. doi:10.1155/2018/8285653

Han, Z., Xue, W., Tao, L., Lou, Y., Qiu, Y., and Zhu, F. (2020). Genome-wide Identification and Analysis of the eQTL lncRNAs in Multiple Sclerosis Based on RNA-Seq Data. *Brief Bioinform.* 21, 1023–1037. doi:10.1093/bib/bbz036

Hekman, R. M., Hume, A. J., Goel, R. K., Abo, K. M., Huang, J., Blum, B. C., et al. (2021). Actionable Cytopathogenic Host Responses of Human Alveolar Type 2 Cells to SARS-CoV-2. *Mol. Cel* 81, 212. doi:10.1016/j.molcel.2020.12.028

International Multiple Sclerosis Genetics ConsortiumBeecham, A. H., Patsopoulos, N. A., Xifara, D. K., Davis, M. F., Kemppinen, A., et al. (2013). Analysis of Immune-Related Loci Identifies 48 New Susceptibility Variants for Multiple Sclerosis. *Nat. Genet.* 45, 1353–1360. doi:10.1038/ng.2770

Irimia, M., Weatheritt, R. J., Ellis, J. D., Parikshak, N. N., Gonatopoulos-Pournatzis, T., Babor, M., et al. (2014). A Highly Conserved Program of Neuronal Microexons Is Misregulated in Autistic Brains. *Cell* 159, 1511–1523. doi:10.1016/j.cell.2014.11.035

Jensen, C. J., Stankovich, J., Butzkueven, H., Oldfield, B. J., and Rubio, J. P. (2010). Common Variation in the MOG Gene Influences Transcript Splicing in Humans. *J. Neuroimmunol.* 229, 225–231. doi:10.1016/j.jneuroim.2010.07.027

Jiao, X., Sherman, B. T., Huang, D. W., Stephens, R., Baseler, M. W., Lane, H. C., et al. (2012). DAVID-WS: a Stateful Web Service to Facilitate Gene/protein List Analysis. *Bioinformatics* 28, 1805–1806. doi:10.1093/bioinformatics/bts251

Johnston, J. B., Silva, C., Gonzalez, G., Holden, J., Warren, K. G., Metz, L. M., et al. (2001). Diminished Adenosine A1 Receptor Expression on Macrophages in Brain and Blood of Patients with Multiple Sclerosis. *Ann. Neurol.* 49, 650–658. doi:10.1002/ana.1007

Langfelder, P., and Horvath, S. (2008). WGCNA: an R Package for Weighted Correlation Network Analysis. *BMC Bioinformatics* 9, 559. doi:10.1186/1471-2105-9-559

Li, H. (2011). A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data. *Bioinformatics* 27, 2987–2993. doi:10.1093/bioinformatics/btr509

Li, H., and Durbin, R. (2009). Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Liu, Q., Guo, Y., Li, J., Long, J., Zhang, B., and Shyr, Y. (2012). Steps to Ensure Accuracy in Genotype and SNP Calling from Illumina Sequencing Data. *BMC Genomics* 13 (Suppl. 8), S8. doi:10.1186/1471-2164-13-S8-S8

Marchese, E., Valentini, M., Di Sante, G., Cesari, E., Adinolfi, A., Corvino, V., et al. (2021). Alternative Splicing of Neurexins 1-3 Is Modulated by Neuroinflammation in the Prefrontal Cortex of a Murine Model of Multiple Sclerosis. *Exp. Neurol.* 335, 113497. doi:10.1016/j.expneurol.2020.113497

Merkin, J., Russell, C., Chen, P., and Burge, C. B. (2012). Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. *Science* 338, 1593–1599. doi:10.1126/science.1228186

Olsson, T., Barcellos, L. F., and Alfredsson, L. (2017). Interactions between Genetic, Lifestyle and Environmental Risk Factors for Multiple Sclerosis. *Nat. Rev. Neurol.* 13, 25–36. doi:10.1038/nrneurol.2016.187

Paraboschi, E. M., Rimoldi, V., Solda, G., Tabaglio, T., Dall'Osso, C., Saba, E., et al. (2014). Functional Variations Modulating PRKCA Expression and Alternative Splicing Predispose to Multiple Sclerosis. *Hum. Mol. Genet.* 23, 6746–6761. doi:10.1093/hmg/ddu392

Patsopoulos, N. A. (2018). Genetics of Multiple Sclerosis: An Overview and New Directions. *Cold Spring Harb Perspect. Med.* 8, a028951. doi:10.1101/cshperspect.a028951

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007). NCBI Reference Sequences (RefSeq): a Curated Non-redundant Sequence Database of Genomes, Transcripts and Proteins. *Nucleic Acids Res.* 35, D61–D65. doi:10.1093/nar/gkl842

Quinn, E. M., Cormican, P., Kenny, E. M., Hill, M., Anney, R., Gill, M., et al. (2013). Development of Strategies for SNP Detection in RNA-Seq Data: Application to

Lymphoblastoid Cell Lines and Evaluation Using 1000 Genomes Data. *PLoS One* 8, e58815. doi:10.1371/journal.pone.0058815

Sawcer, S., Franklin, R. J. M., and Ban, M. (2014). Multiple Sclerosis Genetics. *Lancet Neurol.* 13, 700–709. doi:10.1016/S1474-4422(14)70041-9

Shabalin, A. A. (2012). Matrix eQTL: Ultra Fast eQTL Analysis via Large Matrix Operations. *Bioinformatics* 28, 1353–1358. doi:10.1093/bioinformatics/bts163

Takata, A., Matsumoto, N., and Kato, T. (2017). Genome-wide Identification of Splicing QTLs in the Human Brain and Their Enrichment Among Schizophrenia-Associated Loci. *Nat. Commun.* 8, 14519. doi:10.1038/ncomms14519

Trinschek, B., Lüssi, F., Haas, J., Wildemann, B., Zipp, F., Wiendl, H., et al. (2013). Kinetics of IL-6 Production Defines T Effector Cell Responsiveness to Regulatory T Cells in Multiple Sclerosis. *PLoS One* 8, e77634. doi:10.1371/journal.pone.0077634

Van Buuren, S., and Groothuis-Oudshoorn, C. G. (2011). Mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* 45, 1–67. doi:10.18637/jss.v045.i03

Walton, C., King, R., Rechtman, L., Kaye, W., Leray, E., Marrie, R. A., et al. (2020). Rising Prevalence of Multiple Sclerosis Worldwide: Insights from the Atlas of MS, Third Edition. *Mult. Scler.* 26, 1816–1821. doi:10.1177/1352458520970841

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-seq: a Revolutionary Tool for Transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi:10.1038/nrg2484

Xu, X., Zhu, K., Liu, F., Wang, Y., Shen, J., Jin, J., et al. (2013). Identification of Somatic Mutations in Human Prostate Cancer by RNA-Seq. *Gene* 519, 343–347. doi:10.1016/j.gene.2013.01.046

Yang, C., Wu, Q., Huang, K., Wang, X., Yu, T., Liao, X., et al. (2019). Genome-Wide Profiling Reveals the Landscape of Prognostic Alternative Splicing Signatures in Pancreatic Ductal Adenocarcinoma. *Front. Oncol.* 9, 511. doi:10.3389/fonc.2019.00511

Yang, H., and Wang, K. (2015). Genomic Variant Annotation and Prioritization with ANNOVAR and wANNOVAR. *Nat. Protoc.* 10, 1556–1566. doi:10.1038/nprot.2015.105

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Genetics

ORIGINAL RESEARCH
published: 18 November 2021
doi: 10.3389/fgene.2021.754795

Matrix Metalloproteinases in Relation to Bone Mineral Density: A Two-Sample Mendelian Randomization Study

Xin Lv[1†], *Pengfei Wu*[2,3†], *Shipeng Xiao*[1†], *Wan Zhang*[4], *Yawei Li*[1], *Bolin Ren*[5,6], *Zhihong Li*[5,6], *Kun Xia*[2,3,7] *and Bing Wang*[1]*

[1] *Department of Spine Surgery, The Second Xiangya Hospital, Central South University, Changsha, China,* [2] *Center for Medical Genetics and Hunan Key Laboratory of Medical Genetics, School of Life Sciences, Central South University, Changsha, China,* [3] *Hunan Key Laboratory of Animal Models for Human Diseases, Central South University, Changsha, China,* [4] *Department of Biology, Boston University, Boston, MA, United States,* [5] *Department of Orthopedics, The Second Xiangya Hospital, Central South University, Changsha, China,* [6] *Hunan Key Laboratory of Tumor Models and Individualized Medicine, The Second Xiangya Hospital, Central South University, Changsha, China,* [7] *Hengyang Medical School, University of South China, Hengyang, China*

Background: We aimed at investigating causal associations between matrix metalloproteinases (MMPs) and bone mineral density (BMD) by the Mendelian randomization (MR) analysis.

Methods: From genome-wide association studies of European ancestry, we selected instrumental variables for MMP-1, MMP-3, MMP-7, MMP-8, MMP-10, and MMP-12. Accordingly, we retrieved summary statistics of three site-specific BMD, namely, forearm, femoral neck, and lumbar spine. We conducted an inverse variance weighted MR as the primary method to compute overall effects from multiple instruments, while additional MR approaches and sensitivity analyses were implemented. Bonferroni-adjusted significance threshold was set at $p < 0.05/18 = 0.003$.

Results: Totally, there was no evidence for causal effects of genetically-predicted levels of MMPs on BMD measurement at three common sites. MR results indicated that there were no causal associations of circulating MMPs with forearm BMD (all $p \geq 0.023$) by the inverse variance weighted method. Similarly, there were no causal effects of MMPs on femoral neck BMD (all $p \geq 0.120$) and MR results did not support causal relationships between MMPs and lumbar spine BMD (all $p \geq 0.017$). Multiple sensitivity analyses suggested the robustness of MR results, which were less likely to be biased by unbalanced pleiotropy or evident heterogeneity.

Conclusion: We found no evidence for the causal relationship between MMPs and BMD in the European population.

Keywords: matrix metalloproteinase, bone mineral density, mendelian randomization, genome-wide association study, summary statistics, causal inference

OPEN ACCESS

Edited by:
*Guiyou Liu,
Tianjin Institute of Industrial Biotechnology (CAS), China*

Reviewed by:
*Yingfang Ao,
Peking University Third Hospital, China
Biao Zeng,
Icahn School of Medicine at Mount Sinai, United States
Jianru Xiao,
Shanghai Changzheng Hospital, China*

***Correspondence:**
*Bing Wang
wbxyeyy@csu.edu.cn*

[†]*These authors have contributed equally to this work and share first authorship*

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *07 August 2021*
Accepted: *02 November 2021*
Published: *18 November 2021*

Citation:
Lv X, Wu P, Xiao S, Zhang W, Li Y, Ren B, Li Z, Xia K and Wang B (2021) Matrix Metalloproteinases in Relation to Bone Mineral Density: A Two-Sample Mendelian Randomization Study. Front. Genet. 12:754795. doi: 10.3389/fgene.2021.754795

INTRODUCTION

Bone mineral density (BMD) is a key measurement of bone mass and an essential indicator of osteoporosis, which is prevalent in the aging society. In 1994, the World Health Organization gave the diagnosis standard of osteoporosis as 2.5 SD or more below the young adult average value (Kanis, 1994). The main characteristics of osteoporosis include loss of bone mass, deterioration of the bone microarchitecture, decrement of bone strength and increased risk of fractures, which lead to a systemic skeletal disorder with negative consequences on general health and quality of life in post menopause and in old age (Lane, 2006; Vidal et al., 2019; Capozzi et al., 2020). Fractures due to osteoporosis more likely occur on the hip, vertebral body and distal forearm, therefore, the BMD measurements of forearm (FA), lumbar spine (LS) and femoral neck (FN) are always taken by dual-energy X-ray absorptiometry (DXA) in patients to estimate the general risk of osteoporosis. With the continued ageing of the population worldwide, osteoporotic fractures could present an increasing prevalence and thus lead to higher rates of chronic pain, disability and even death in patients, as well as impose a major economic burden on healthcare systems (Sambrook and Cooper, 2006; Catalano et al., 2017). Current studies have found several risk factors that may decrease BMD (Kenny and Prestwood, 2000; Raisz, 2005; Li and Wang, 2018), but overall, the cause of osteoporosis still remains unclear, which brings difficulty in seeking for effective therapy for this disease.

Matrix metalloproteinases (MMPs) are a family of zinc-dependent neutral endopeptidases capable of degrading extracellular matrix components (Johansson et al., 2000). Previous studies have found that MMPs are expressed in bone tissue as key players in the digestion of bone matrix by osteoblasts, and are involved in bone-destructive lesions (Wahlgren et al., 2001; Azevedo et al., 2018; Fatemi et al., 2020), which indicates that MMPs may play a role in the pathogenesis of osteoporosis. It has been reported that the gene polymorphism of MMP-1 was associated with osteoporosis (Liang et al., 2019), and MMP-3 was negatively related to the osteoblast function markers of serum bone-specific alkaline phosphatase and osteocalcin while positively related to the resorptive function marker of serum cross-linked N-telopeptides of type I collagen (Momohara et al., 2005). Increased levels of MMP-7 and 9 in osteoclasts were reported to be associated with rheumatic osteoporosis (Yang et al., 2013), while MMP-8 participated in the healing process as well as embryonic bone development, and may play an important role in the remodeling of extracellular matrix molecules during bone and cartilage formation (Sasano et al., 2002). MMP-10 was found strongly expressed in osteoclasts and most mononuclear cells within the marrow and produced in an active form with associated degradation (Bord et al., 1998). Meanwhile, recombinant MMP-12 cleaved the putative functional domains of osteopontin and bone sialoprotein, two bone matrix proteins that strongly influence osteoclast activities, such as attachment, spreading and resorption (Hou et al., 2004). These studies strongly suggested the possibility that MMPs are related to osteoporosis. However, due to current randomized controlled trials which were based on either small samples or observational epidemiological studies, whether changes in MMP levels are correlated with BMD remains controversial.

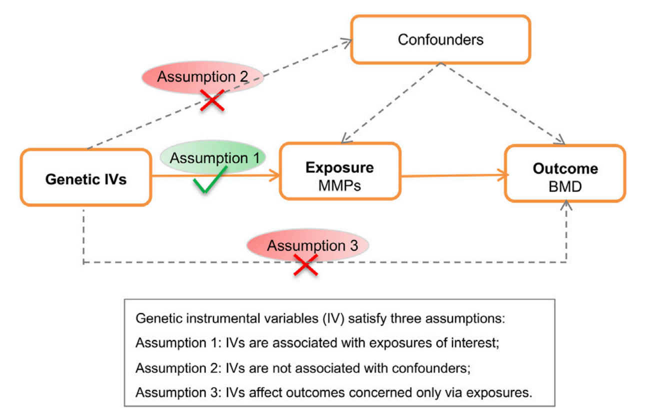

Confounders
Assumption 2
Genetic IVs
Assumption 1
Exposure
MMPs
Outcome
BMD
Assumption 3
Genetic instrumental variables (IV) satisfy three assumptions:
Assumption 1: IVs are associated with exposures of interest;
Assumption 2: IVs are not associated with confounders;
Assumption 3: IVs affect outcomes concerned only via exposures.

FIGURE 1 | Schematic of the Mendelian randomization analysis. BMD, bone mineral density; MMP; matrix metalloproteinase; SNP, single nucleotide polymorphism.

Genome-wide association studies (GWAS) provide a new perspective for understanding genetic determinants that underlie complex disease. The technique of Mendelian randomization (MR), which employs single nucleotide polymorphism (SNPs) as instrumental variables, has been developed to identify causations between a wide range of risk factors and complex diseases. Unlike traditional observational studies, this analytical tool was less susceptible to confounding and reverse causation (Davey Smith and Hemani, 2014). MR has also been widely used these years to explore the causes of osteoporosis (Larsson et al., 2019; Zheng et al., 2019). Given that MMPs were hypothesized to participate in the development of osteoporosis, here we carried out an MR study to identify whether there existed causal associations between MMPs and BMD.

MATERIALS AND METHODS

The MR schematic was shown in **Figure 1**. There were three underlying assumptions: 1) relevance assumption, genetic instrumental variables are associated with the risk factor of interest; 2) independence assumption, genetic variants are not associated with confounders; and 3) exclusion-restriction assumption, instrumental SNPs influence the outcome concerned only through the risk factor (Burgess et al., 2019). This study utilized publicly accessible datasets from published studies wherein formal consent from participants and ethical approval by committees had been obtained.

Data Sources

Summary-level association data for MMPs were obtained from GWASs of European ancestry (Salminen et al., 2017; Folkersen et al., 2020). Folkersen et al. (Folkersen et al., 2020) recently conducted a large-scale mapping of protein quantitative trait loci.

Circulating levels of MMPs, including MMP-1 (n = 16,889), MMP-3 (n = 20,791), MMP-7 (n = 18,245), MMP-10 (n = 16,933), and MMP-12 (n = 19,178) were measured among a panel of 90 candidate biomarkers related to cardiovascular risk. Summary statistics were released by the SCALLOP consortium (http://www.scallop-consortium.com/scallop_downloads/).
Genetic variants associated with MMPs at genome-wide significant significance (p = 5 × 10^{-8}) and clumped at the threshold (r^2 = 0.001 within ±1 Mb, EUR 1000 Genomes phase 3) were selected as instrumental variables (**Supplementary Table S1**). Salminen et al. (Salminen et al., 2017) conducted a GWAS of MMP-8 concentrations in 6,049 Europeans and strongest associations were identified at locus 1q31.3. Two independent SNP associated with MMP-8 meeting the above criteria were utilized as instrumental variables in the ensuing MR analysis. Effect size was given in the unit of SD change in circulating concentration per additional effect allele (**Supplementary Table S2–7**).

Summary statistics for BMD used in this study were gained from the GWAS datasets released by the GEnetic Factors for OSteoporosis Consortium. Zheng et al. (Zheng et al., 2015) performed a large-scale meta-analysis in 2015 to identify genetic variants associated with BMD including FA-BMD (n = 8,143), FN-BMD (n = 32,735) and LS-BMD (n = 28,498) in individuals of European ancestry from the general population. It is the largest GWAS on DXA-measured BMD so far. The associations for BMD were derived from whole-genome sequencing, whole-exome sequencing, deep imputation, and *de novo* replication genotyping. The association of each SNP with BMD was tested and adjusted for sex, age, square of age and weight. When instrumental SNPs were not present in the BMD datasets, proxies (r^2 > 0.8) were searched and utilized if available. Effect size was given in SDs of BMD in association tests with the additive model (**Supplementary Table S2–7**). Summary statistics of MMPs and BMD were harmonized in terms of effect allele, and subsequent analyses were based on the merged exposure-outcome dataset.

Mendelian Randomization

The MR analysis was conducted using the TwoSampleMR (version 0.5.4) package (Hemani et al., 2018) in R 3.6.3 (R Foundation for Statistical Computing, Vienna, Austria). First, individual estimate of the causal effect MMPs on site-specific BMD mediated by each instrumental SNP was computed as the Wald ratio (Walker et al., 2019). Then, the primary method, the inverse variance weighted (IVW) MR was employed to generate overall estimates (Burgess et al., 2013). Two complementary approaches were implemented, considering that IVW estimates would be biased in the presence of invalid instruments or horizontal pleiotropy. Weighted median approach would give robust effect estimates when less than half instruments were invalid (Bowden et al., 2016). MR-Egger regression would serve as a tool to detect unbalanced horizontal pleiotropy, and generate estimates adjusted for pleiotropy (Burgess and Thompson, 2017). IVW estimates were generally more precise, whereas effect estimates given by weighted median and MR-Egger were accompanied by wide confidence intervals (CIs) in the forest plots. Causal effects on BMD were presented in SD units per 1-SD increase in circulating levels of MMPs. The Bonferroni-corrected significance level at p < 0.05/18 = 0.003 was adopted in the scenario of multiple tests.

RESULTS

Mendelian Randomization Analyses of Matrix Metalloproteinases on FA-Bone Mineral Density

MR results demonstrated that genetically-predicted levels of MMPs were not associated with changes in FA-BMD (**Figure 2**). By the primary method, causal effects on FA-BMD were 0.024 SD (−0.018–0.402, p = 0.402) per 1-SD increase in MMP-1 levels, −0.005 SD (−0.074–0.065; p = 0.896) per 1-SD increase in MMP-3 levels, −0.218 SD (−0.461–0.025; p = 0.079) per 1-SD increase in MMP-7 levels, −0.252 SD (−0.535–0.032; p = 0.082) per 1-SD increase in MMP-8 levels, -0.271 SD (−0.504–−0.038; p = 0.023) per 1-SD increase in MMP-10 levels, and −0.016 SD (−0.070–0.039; p = 0.575) per 1-SD increase in MMP-10 levels. MR results were generally consistent among causal estimates given by IVW methods and two additional approaches (**Supplementary Table S8**). In MR analyses with three or more instrumental variables (except for MMP-8), no horizontal pleiotropy was detected according to MR-Egger intercepts and no evident heterogeneity was identified (**Supplementary Table S8**).

Mendelian Randomization Analyses of Matrix Metalloproteinases on FN-Bone Mineral Density

Overall, MR estimates suggested that circulating concentrations of MMPs were not associated with FN-BMD. As shown in **Figure 2**, there was no evidence for causal effects of MMP-1 (−0.018 SD; −0.059–0.024; p = 0.402), MMP-3 (0.006 SD; −0.027–0.040; p = 0.708), MMP-7 (0.017 SD; −0.070 to 0.104; p = 0.697), MMP-8 (−0.073 SD; −0.168–0.023; p = 0.135), MMP-10 (−0.145 SD; −0.327–0.038; p = 0.120) and MMP-12 (−0.016 SD; −0.042–0.010; p = 0.238) by the IVW approach. Complimentary methods further verified the robustness of MR results by the primary method, and there was no evidence for the existence of unbalanced horizontal pleiotropy or heterogeneity (**Supplementary Table S9**).

Mendelian Randomization Analyses of Matrix Metalloproteinases on LS-Bone Mineral Density

MR analyses showed that genetically-predicted MMPs were not in relation to LS-BMD (**Figure 2**). Causal relationships between circulating levels of MMP-1 (−0.007 SD; −0.046–0.032; p = 0.718), MMP-3 (0.013 SD; −0.020–0.047; p = 0.430), MMP-7 (0.028 SD; −0.077–0.134; p = 0.599), MMP-8 (−0.107 SD; −0.194–−0.019; p = 0.017), MMP-10 (−0.099 SD; −0.223–0.025; p = 0.118) and MMP-12 (−0.006 SD; −0.036–0.025; p = 0.721) and measurement in LS-BMD were not significant by the IVW method.

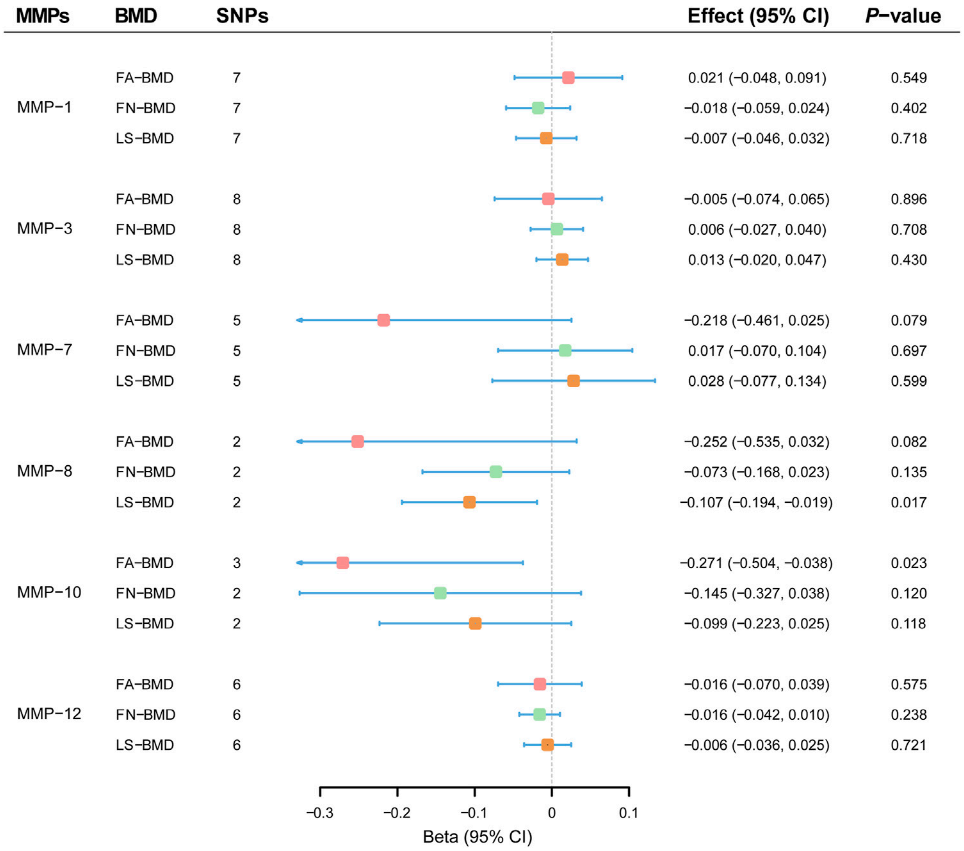


FIGURE 2 | Effect estimates of matrix metalloproteinases on bone mineral density in the Mendelian randomization study. BMD, bone mineral density; CI; confidence interval; FA, forearm; FN, femoral neck; LS, lumbar spine; MMP; matrix metalloproteinase; SNP, single nucleotide polymorphism.

According to sensitivity analyses (**Supplementary Table S10**), MR results by different methods were consistent; besides, unbalanced horizontal pleiotropy or obvious heterogeneity was not present.

DISCUSSION

Osteoporosis is a common cause of morbidity and mortality worldwide especially in people aged over 60 years. Studies have shown that for decrease per 10 percent in bone mineral density, the risk of fracture increases 2–3 folds (Nguyen et al., 1993), and the mortality rate of patients caused by hip and spine fractures increases to 10–20% (Ioannidis et al., 2009). The causes of decrease in BMD have always been discussed in order to benefit for seeking effective therapy, and more and more risk factors are being identified to better predict the occurrence of osteoporosis and therefore avoid the severe complications of fracture.

The family of matrix metalloproteinases have been considered involved in basic pathological processes of osteoporosis for acting as key roles in the digestion of bone matrix by osteoblasts (Azevedo et al., 2018). However, different studies showed conflict results. For example, Zuo et al. (Zuo et al., 2020) found that MMP-8 was involved in the 17β-Estradiol replacement therapy for postmenopausal osteoporosis, while Viljakainen et al. (Viljakainen et al., 2017) found that there was no significant correlation between MMP-8 levels and low BMD. MR is an effective tool for identifying the causal association between certain exposure and disease while circumventing confounders, which might be the main cause of these inconsistent results. In the recent 3 years, a lot of factors that had been reported related to osteoporosis before have been re-evaluated by MR. Some were further confirmed to be associated with BMD, for instance, serum calcium (Sun et al., 2021), sex hormone-binding globulin (Qu et al., 2021) and age at menarche (Magnus et al., 2020), while others such as vascular endothelial growth factor, uric acid and serum vitamin D got no evidence for their correlations with osteoporosis (Lee and Song, 2019; Sun et al., 2019; Keller-Baruch et al., 2020). In a previous MR study of heel-ultrasound estimated BMD (Folkersen et al., 2020), there were no causal effects of MMP-1, 3, 7, 10, 12 in the European population. In this study, we found no evidence for the causal relationship between MMPs and DXA measured BMD at three common sites.

There are some limitations in present study. First, we could not identify the non-linear relationship between MMPs and BMD. Second, we only evaluated the effect of a small set of MMPs on BMD, but missed such types as MMP-2, -9 and -13, which might be in relation to osteoporosis according to previous studies (Bolton et al., 2009; Zheng et al., 2018). Further MR studies were warranted when relevant datasets are available. Third, it is noteworthy that our study was limited to the effect of circulating MMP levels on BMD, but the intracellular function of MMPs cannot be denied. Forth, both association data of MMPs and BMD were obtained from Europeans in this study. We should be cautious when generalizing the conclusion to other populations.

In this study, we found no evidence for causal relationships between MMPs (MMP-1, 3, 7, 8, 10, 12) and BMD in the European population.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

XL, PW, KX, and BW contributed to the conceptualization of the study. XL, PW, SX, and WZ played a part in the acquisition and analysis of data, and validation and visualization of results. SX, YL, BR, and ZL participated in drafting and reviewing the main manuscript. ZL, KX, and BW contributed to the project administration, and funding acquisition. All authors contributed to the article and approved the final version of the manuscript.

ACKNOWLEDGMENTS

The authors are grateful to Dr. Folkersen, Dr. Salminen, Dr. Zheng, et al. for making summary-level association statistics of MMPs and BMD possible and accessible. We also want to acknowledge all participants and investigators in these studies. PW has received a visiting PhD scholarship from the China Scholarship Council.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.754795/full#supplementary-material

REFERENCES

Azevedo, A., Prado, A. F., Feldman, S., de Figueiredo, F. A. T., Dos Santos, M. C. G., and Issa, J. P. M. (2018). MMPs Are Involved in Osteoporosis and Are Correlated with Cardiovascular Diseases. *Cpd* 24, 1801–1810. doi:10.2174/1381612824666180604112925

Bolton, C., Stone, M., Edwards, P., Duckers, J., Evans, W., and Shale, D. (2009). Circulating Matrix Metalloproteinase-9 and Osteoporosis in Patients with Chronic Obstructive Pulmonary Disease. *Chron. Respir. Dis.* 6, 81–87. doi:10.1177/1479972309103131

Bord, S., Horner, A., Hembry, R. M., and Compston, J. E. (1998). Stromelysin-1 (MMP-3) and Stromelysin-2 (MMP-10) Expression in Developing Human Bone: Potential Roles in Skeletal Development. *Bone* 23, 7–12. doi:10.1016/s8756-3282(98)00064-7

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* 40, 304–314. doi:10.1002/gepi.21965

Burgess, S., and Thompson, S. G. (2017). Interpreting Findings from Mendelian Randomization Using the MR-Egger Method. *Eur. J. Epidemiol.* 32, 377–389. doi:10.1007/s10654-017-0255-x

Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian Randomization Analysis with Multiple Genetic Variants Using Summarized Data. *Genet. Epidemiol.* 37, 658–665. doi:10.1002/gepi.21758

Burgess, S., Davey Smith, G., Davies, N. M., Dudbridge, F., Gill, D., Glymour, M. M., et al. (2019). Guidelines for Performing Mendelian Randomization Investigations. *Wellcome Open Res.* 4, 186. doi:10.12688/wellcomeopenres.15555.1

Capozzi, A., Scambia, G., and Lello, S. (2020). Calcium, Vitamin D, Vitamin K2, and Magnesium Supplementation and Skeletal Health. *Maturitas* 140, 55–63. doi:10.1016/j.maturitas.2020.05.020

Catalano, A., Martino, G., Morabito, N., Scarcella, C., Gaudio, A., Basile, G., et al. (2017). Pain in Osteoporosis: From Pathophysiology to Therapeutic Approach. *Drugs Aging* 34, 755–765. doi:10.1007/s40266-017-0492-4

Davey Smith, G., and Hemani, G. (2014). Mendelian Randomization: Genetic Anchors for Causal Inference in Epidemiological Studies. *Hum. Mol. Genet.* 23, R89–R98. doi:10.1093/hmg/ddu328

Fatemi, K., Rezaee, S. A., Banihashem, S. A., Keyvanfar, S., and Eslami, M. (2020). Importance of MMP-8 in Salivary and Gingival Crevicular Fluids of Periodontitis Patients. *Iran J. Immunol.* 17, 236–243. doi:10.22034/iji.2020.81170.1512

Folkersen, L., Gustafsson, S., Wang, Q., Hansen, D. H., Hedman, Å. K., Schork, A., et al. (2020). Genomic and Drug Target Evaluation of 90 Cardiovascular Proteins in 30,931 Individuals. *Nat. Metab.* 2, 1135–1148. doi:10.1038/s42255-020-00287-2

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., et al. (2018). The MR-Base Platform Supports Systematic Causal Inference across the Human Phenome. *Elife* 7, e34408. doi:10.7554/eLife.34408

Hou, P., Troen, T., Ovejero, M. C., Kirkegaard, T., Andersen, T. L., Byrjalsen, I., et al. (2004). Matrix Metalloproteinase-12 (MMP-12) in Osteoclasts: New Lesson on the Involvement of MMPs in Bone Resorption. *Bone* 34, 37–47. doi:10.1016/j.bone.2003.08.011

Ioannidis, G., Papaioannou, A., Hopman, W. M., Akhtar-Danesh, N., Anastassiades, T., Pickard, L., et al. (2009). Relation between Fractures and Mortality: Results from the Canadian Multicentre Osteoporosis Study. *Can. Med. Assoc. J.* 181, 265–271. doi:10.1503/cmaj.081720

Johansson, N., Ahonen, M., and Kähäri, V. M. (2000). Matrix Metalloproteinases in Tumor Invasion. *Cell Mol. Life Sci. (Cmls)* 57, 5–15. doi:10.1007/s000180050495

Kanis, J. A., and Kanis, J. A. (1994). Assessment of Fracture Risk and its Application to Screening for Postmenopausal Osteoporosis: Synopsis of a WHO Report. *Osteoporos. Int.* 4, 368–381. doi:10.1007/bf01622200

Keller-Baruch, J., Forgetta, V., Manousaki, D., Zhou, S., and Richards, J. B. (2020). Genetically Decreased Circulating Vascular Endothelial Growth Factor and Osteoporosis Outcomes: A Mendelian Randomization Study. *J. Bone Miner Res.* 35, 649–656. doi:10.1002/jbmr.3937

Kenny, A. M., and Prestwood, K. M. (2000). Osteoporosis. *Rheum. Dis. Clin. North Am.* 26, 569–591. doi:10.1016/s0889-857x(05)70157-5

Lane, N. E. (2006). Epidemiology, Etiology, and Diagnosis of Osteoporosis. *Am. J. Obstet. Gynecol.* 194, S3–S11. doi:10.1016/j.ajog.2005.08.047

Larsson, S. C., Michaëlsson, K., and Burgess, S. (2019). Mendelian Randomization in the Bone Field. *Bone* 126, 51–58. doi:10.1016/j.bone.2018.10.011

Lee, Y. H., and Song, G. G. (2019). Uric Acid Level, Gout and Bone mineral Density: A Mendelian Randomization Study. *Eur. J. Clin. Invest.* 49, e13156. doi:10.1111/eci.13156

Li, L., and Wang, Z. (2018). Ovarian Aging and Osteoporosis. *Adv. Exp. Med. Biol.* 1086, 199–215. doi:10.1007/978-981-13-1117-8_13

Liang, L., Zhu, D. P., Guo, S. S., Zhang, D., and Zhang, T. (2019). MMP-1 Gene Polymorphism in Osteoporosis. *Eur. Rev. Med. Pharmacol. Sci.* 23, 67–72. doi:10.26355/eurrev_201908_18631

Magnus, M. C., Guyatt, A. L., Lawn, R. B., Wyss, A. B., Trajanoska, K., Küpers, L. K., et al. (2020). Identifying Potential Causal Effects of Age at Menarche: a Mendelian Randomization Phenome-wide Association Study. *BMC Med.* 18, 71. doi:10.1186/s12916-020-01515-y

Momohara, S., Okamoto, H., Yago, T., Furuya, T., Nanke, Y., Kotake, S., et al. (2005). The Study of Bone mineral Density and Bone Turnover Markers in Postmenopausal Women with Active Rheumatoid Arthritis. *Mod. Rheumatol.* 15, 410–414. doi:10.1007/s10165-005-0435-510.3109/s10165-005-0435-5

Nguyen, T., Sambrook, P., Kelly, P., Jones, G., Lord, S., Freund, J., et al. (1993). Prediction of Osteoporotic Fractures by Postural Instability and Bone Density. *Bmj* 307, 1111–1115. doi:10.1136/bmj.307.6912.1111

Qu, Z., Jiang, J., Yang, F., Huang, J., Zhao, J., and Yan, S. (2021). Genetically Predicted Sex Hormone-Binding Globulin and Bone Mineral Density: A Mendelian Randomization Study. *Calcif Tissue Int.* 108, 281–287. doi:10.1007/s00223-020-00770-8

Raisz, L. G. (2005). Pathogenesis of Osteoporosis: Concepts, Conflicts, and Prospects. *J. Clin. Invest.* 115, 3318–3325. doi:10.1172/jci27071

Salminen, A., Vlachopoulou, E., Havulinna, A. S., Tervahartiala, T., Sattler, W., Lokki, M.-L., et al. (2017). Genetic Variants Contributing to Circulating Matrix Metalloproteinase 8 Levels and Their Association with Cardiovascular Diseases. *Circ. Cardiovasc. Genet.* 10. doi:10.1161/circgenetics.117.001731

Sambrook, P., and Cooper, C. (2006). Osteoporosis. *Lancet* 367, 2010–2018. doi:10.1016/s0140-6736(06)68891-0

Sasano, Y., Zhu, J.-X., Tsubota, M., Takahashi, I., Onodera, K., Mizoguchi, I., et al. (2002). Gene Expression of MMP8 and MMP13 during Embryonic Development of Bone and Cartilage in the Rat Mandible and Hind Limb. *J. Histochem. Cytochem.* 50, 325–332. doi:10.1177/002215540205000304

Sun, J.-y., Zhao, M., Hou, Y., Zhang, C., Oh, J., Sun, Z., et al. (2019). Circulating Serum Vitamin D Levels and Total Body Bone mineral Density: A Mendelian Randomization Study. *J. Cel Mol Med.* 23, 2268–2271. doi:10.1111/jcmm.14153

Sun, J.-y., Zhang, H., Zhang, Y., Wang, L., Sun, B.-l., Gao, F., et al. (2021). Impact of Serum Calcium Levels on Total Body Bone mineral Density: A Mendelian Randomization Study in Five Age Strata. *Clin. Nutr.* 40, 2726–2733. doi:10.1016/j.clnu.2021.03.012

Vidal, M., Thibodaux, R. J., Neira, L. F. V., and Messina, O. D. (2019). Osteoporosis: a Clinical and Pharmacological Update. *Clin. Rheumatol.* 38, 385–395. doi:10.1007/s10067-018-4370-1

Viljakainen, H. T., Koistinen, H. A., Tervahartiala, T., Sorsa, T., Andersson, S., and Mäkitie, O. (2017). Metabolic Milieu Associates with Impaired Skeletal Characteristics in Obesity. *PLoS One* 12, e0179660. doi:10.1371/journal.pone.0179660

Wahlgren, J., Maisi, P. i., Sorsa, T., Sutinen, M., Tervahartiala, T., Pirilä, E., et al. (2001). Expression and Induction of Collagenases (MMP-8 and -13) in Plasma Cells Associated with Bone-Destructive Lesions. *J. Pathol.* 194, 217–224. doi:10.1002/path.854

Walker, V. M., Davies, N. M., Hemani, G., Zheng, J., Haycock, P. C., Gaunt, T. R., et al. (2019). Using the MR-Base Platform to Investigate Risk Factors and Drug Targets for Thousands of Phenotypes. *Wellcome Open Res.* 4, 113. doi:10.12688/wellcomeopenres.15334.2

Yang, P. T., Meng, X. H., Yang, Y., and Xiao, W. G. (2013). Inhibition of Osteoclast Differentiation and Matrix Metalloproteinase Production by CD4+CD25+ T Cells in Mice. *Osteoporos. Int.* 24, 1113–1114. doi:10.1007/s00198-012-2014-x

Zheng, H. F., Forgetta, V., Hsu, Y. H., Estrada, K., Rosello-Diez, A., Leo, P. J., et al. (2015). Whole-genome Sequencing Identifies EN1 as a Determinant of Bone Density and Fracture. *Nature* 526, 112–117. doi:10.1038/nature14878

Zheng, X., Zhang, Y., Guo, S., Zhang, W., Wang, J., and Lin, Y. (2018). Dynamic Expression of Matrix Metalloproteinases 2, 9 and 13 in Ovariectomy-Induced Osteoporosis Rats. *Exp. Ther. Med.* 16, 1807–1813. doi:10.3892/etm.2018.6356

Zheng, J., Frysz, M., Kemp, J. P., Evans, D. M., Davey Smith, G., and Tobias, J. H. (2019). Use of Mendelian Randomization to Examine Causal Inference in Osteoporosis. *Front. Endocrinol.* 10, 807. doi:10.3389/fendo.2019.00807

Zuo, H.-L., Xin, H., Yan, X.-N., Huang, J., Zhang, Y.-P., and Du, H. (2020). 17β-Estradiol Improves Osteoblastic Cell Function through the Sirt1/NF-κB/MMP-8 Pathway. *Climacteric* 23, 404–409. doi:10.1080/13697137.2020.1758057

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 18 November 2021
doi: 10.3389/fgene.2021.777942

Association of *GAK* rs1564282 With Susceptibility to Parkinson's Disease in Chinese Populations

He Li[1†], *Chen Zhang*[2†] *and Yong Ji*[1*]

[1] *Tianjin Key Laboratory of Cerebrovascular and of Neurodegenerative Diseases, Department of Neurology, Tianjin Huanhu Hospital, Tianjin, China,* [2] *Tianjin Key Laboratory of Cerebrovascular and of Neurodegenerative Diseases, Department of Neurosurgery, Tianjin Huanhu Hospital, Tianjin, China*

The susceptibility of the *GAK* rs1564282 variant in Parkinson's disease (PD) in Europeans was identified using a series of published genome-wide association studies. Recently, some studies focused on the association between rs1564282 and PD risk in Chinese populations but with inconsistent results. Thus, we conducted an updated meta-analysis with a total of 7,881 samples (4,055 PD cases and 3,826 controls) from eligible studies. After excluding significant heterogeneity, we showed that the rs1564282 variant was significantly associated with PD in Chinese populations (p = 1.00E-04, odds ratio = 1.28 and 95% confidence interval = 1.16–1.42). The sensitivity analysis showed that the association between rs1564282 and PD was not greatly influenced, and there was no significant publication bias among the included studies. Consequently, this meta-analysis indicates that the *GAK* rs1564282 variant is significantly associated with susceptibility to PD in Chinese populations.

Keywords: Parkinson's disease, genome-wide association study, GAK, rs1564282, Chinese population

OPEN ACCESS

Edited by:
Liangcai Zhang,
Janssen Research and Development,
United States

Reviewed by:
Jinchen Li,
Central South University, China
Judong Shen,
Merck and Co., Inc., United States

***Correspondence:**
Yong Ji
jiyongusa@126.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *16 September 2021*
Accepted: *01 November 2021*
Published: *18 November 2021*

Citation:
Li H, Zhang C and Ji Y (2021) Association of GAK rs1564282 With Susceptibility to Parkinson's Disease in Chinese Populations.
Front. Genet. 12:777942.
doi: 10.3389/fgene.2021.777942

INTRODUCTION

Parkinson's disease (PD) is the second-most common neurodegenerative disease after Alzheimer's disease (Ascherio and Schwarzschild, 2016). With the widespread use of genome-wide association studies (GWAS), more genetic components of PD have been identified, and potential mechanisms of PD have been uncovered (Nalls et al., 2014; Nalls et al., 2019). In 2009, Pankratz et al. designated *GAK/DGKQ* as a new PD risk region in a Caucasian population (Pankratz et al., 2009). The following GWAS showed that the *GAK* rs1564282 variant was associated with the increasing risk of PD (Spencer et al., 2011). Subsequently, the underlying associations between rs1564282 and PD were investigated in Chinese populations.

Li et al. selected 812 PD patients and 763 control individuals from west China and first corroborated that rs1564282 was associated with PD in a Chinese population (p = 0.017) (Li et al., 2012). Then a meta-analysis using a European population reached a similar conclusion (Li et al., 2012).

In 2013, Chen et al. recruited 376 PD patients and 277 healthy controls from west China and identified an association between rs1564282 and PD (Chen et al., 2013). The presence of rs1564282 was reported to significantly increase the risk of PD progression (Chen et al., 2013). However, Lin

Abbreviations: CI, confidence interval; GWAS, genome-wide association studies; HWE, Hardy–Weinberg equilibrium; LRRK2, leucine-rich repeat kinase 2; OR, odds radio; PD, Parkinson's disease.

team and Tseng team evaluated Chinese populations from Taiwan and Singapore respectively and demonstrated no association between rs1564282 and PD (Lin et al., 2013; Tseng et al., 2013).

In 2015, Yu et al. analyzed 534 PD patients and 435 neurologically healthy controls from west China and found that rs156428 was significantly associated with PD (Yu et al., 2015).

The inconsistent association results from the previous studies may be due to at least two reasons: genetic heterogeneity and small sample sizes. Firstly, genetic heterogeneity among the previous studies may lead to the inconsistency. Although all the previous studies included Chinese populations, their population compositions (or structures) and geographical environment at largely varied. In other words, the associations between the risk variants and PD may be different among different populations. Secondly, the smaller sample sizes of the previous studies may also contribute to the inconsistency. In these previous studies, Tseng et al. recruited 978 and 777 samples from Taiwan and Singapore population respectively while the Tian et al. used 2049 individuals for analysis. The results of these studies showed that larger sample sizes provided greater power in discovering significant genetic associations. Because of the inconsistent results, the association between rs1564282 and PD in Chinese populations needs further research. Thus, we conducted a new meta-analysis to investigate the association between rs1564282 and PD via combining previous case–control cohort data.

MATERIALS AND METHODS

Systemic Literature Search

A systemic literature search was performed in four databases: PubMed (http://www.ncbi.nlm.nih.gov/pubmed), Google Scholar (https://scholar.google.com/), China National Knowledge Infrastructure (CNKI, http://www.cnki.net/) and Wanfang Medicine database (http://www.wanfangdata.com.cn/). We screened all the relevant studies using the following terms: "Parkinson's disease", "GAK" and "Chinese or China". Literature published before July 31, 2021 was selected. The detailed content of the inclusion and exclusion criteria is given in the Study Selection section.

Study Selection

The eligible studies satisfied the inclusion criteria: 1) case–control designed studies in humans, 2) studies calculating the association between rs1564282 variant and PD and 3) studies providing the number of rs1564282 genotypes or adequate data for the calculation of the odds radio (OR) and a 95% confidence interval (CI). Studies that did not satisfy the inclusion criteria were excluded.

Data Extraction

Two investigators independently extracted the following available data from studies: 1) name of the first author; 2) year of publication; 3) population of study; 4) numbers of PD cases and controls; 5) genotype distribution of rs1564282 in cases and controls; and 6) OR with 95% CI or data for calculating OR and 95% CI.

Genetic Model

The additive genetic model was used to estimate the association between rs1564282 and PD: the T allele versus the C allele.

Statistical Analysis

The Hardy–Weinberg equilibrium (HWE) of rs1564282 in the control for each study was calculated respectively with a chi-squared test at $p < 0.001$. We conducted the heterogeneity test using Cochran's Q test and I^2 statistic (Liu et al., 2013). The Q statistic follows a χ^2 distribution with k−1 degrees of freedom (k means the number of researches selected in calculation). The p-value of Cochran's Q test <0.1 means a significant heterogeneity exists among the studies. The statistic I^2 ($I^2 = \frac{Q-(k-1)}{Q} \times 100\%$) reflects the percentage of variation across studies caused by heterogeneity. I^2 ranges between 0 and 100% (I^2 = 0–25%, 25–50%, 50–75% and 75–100%), with a higher percentage indicating a greater degree of heterogeneity (Liu et al., 2013; Liu et al., 2017). If there was a large amount of heterogeneity ($p < 0.1$ of Q statistic and $I^2 > 50\%$), a random-effect model was used for meta-analysis, otherwise a fixed-effect model was chosen. The statistical significance of OR was calculated utilizing a Z-test, with $p < 0.05$ considered significant. We completed the sensitivity analysis through removing any study from the included studies in turn to evaluate the influence of each study on pooled OR and related p-value (Liu et al., 2017). Publication bias was estimated by a funnel plot. The regression method propounded by Egger was utilized to test publication bias of the selected studies (Egger et al., 1997). The significant threshold was 0.01. All statistical computations were performed utilizing R (http://www.r-project.org/).

RESULTS

Systematic Literature Search

Utilizing our literature search methods, we obtained 20 articles from four databases (**Figure 1**). Firstly, three articles were removed due to duplication or being a review. Subsequently, 11 articles were excluded because they did not estimate the association between rs1564282 variant and PD or have sufficient data to compute OR. Finally, six articles that included seven studies, with a total of 4,055 PD patients and 3,826 controls, were selected for the meta-analysis. Detailed characteristics of the eligible studies are listed in **Table 1**.

HWE and Heterogeneity Test

We evaluated the HWE of rs1564282 in controls for each study respectively. We did not find significant deviation from HWE at $p < 0.001$. Neither Cochran's Q test nor I^2 statistic identified significant heterogeneity of rs1564282 polymorphism among the seven studies in Chinese populations ($p = 0.44$ and $I^2 = 0\%$).

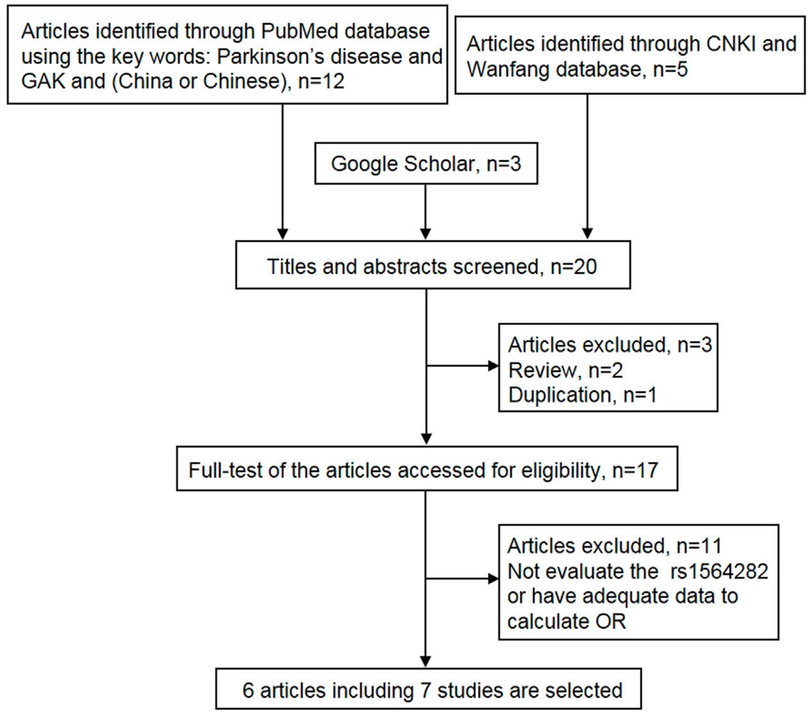

Articles identified through PubMed database using the key words: Parkinson's disease and GAK and (China or Chinese), n=12
Articles identified through CNKI and Wanfang database, n=5
Google Scholar, n=3
Titles and abstracts screened, n=20
Articles excluded, n=3
Review, n=2
Duplication, n=1
Full-test of the articles accessed for eligibility, n=17
Articles excluded, n=11
Not evaluate the rs1564282 or have adequate data to calculate OR
6 articles including 7 studies are selected

FIGURE 1 | Flow chart of the literature search process.

TABLE 1 | Characteristics of seven eligible studies on the association between rs1564282 and PD.

Study	Population	Case	Control	HWE in control	Case genotypes			Control genotypes		
					CC	CT	TT	CC	CT	TT
[a] Tian. (2012)	South China	1,019	1,030	0.11	814	186	19	866	152	12
[b] Li et al. (2012)	West China	812	762	0.22	616	183	13	616	142	4
[a] Chen et al. (2013)	West China	376	277	1	285	81	10	227	48	2
[a] Lin et al. (2013)	Taiwan	448	452	0.60	341	97	10	363	85	4
[b] Tseng et al. (2013)	Taiwan	483	495	0.37	381	97	5	387	104	4
[b] Tseng et al. (2013)	Singapore	388	389	0.095	306	77	5	311	77	1
[a] Yu et al. (2015)	West China	529	421	0.046	385	132	12	331	89	1

HWE: Hardy–Weinberg equilibrium.

[a] Study tested multiple SNPs including the GAK rs1564282 with PD in Chinese populations.

[b] Study only tested the association of one SNP (GAK rs1564282) with PD in Chinese populations.

Meta-Analysis

Because there was no significant heterogeneity of rs1564282 polymorphism, we computed the general OR with 95% CI using a fixed-effect model. The meta-analysis results demonstrated significant association between *GAK* rs1564282 and PD with p = 1.00E-04, OR = 1.28 and 95% CI = 1.16–1.42. More information on the meta-analysis results are shown in **Figure 2**.

Sensitivity Analysis and Publication Bias Analysis

The sensitivity analysis was conducted by removing each study at a time. We found that omitting any eligible study did not substantially influence the overall association between rs1564282 and PD (**Figure 3**).

The funnel plot was used to estimate the publication bias of the included studies. The shape of the funnel plot was symmetrical and inverted (**Figure 4**). The regression test showed no significant publication bias among the seven included studies in this meta-analysis (p = 0.77).

DISCUSSION

The genetic association between the *GAK* rs1564282 and PD was first reported in a familial PD GWAS and replicated by following studies in European populations (Pankratz et al., 2009; Hamza

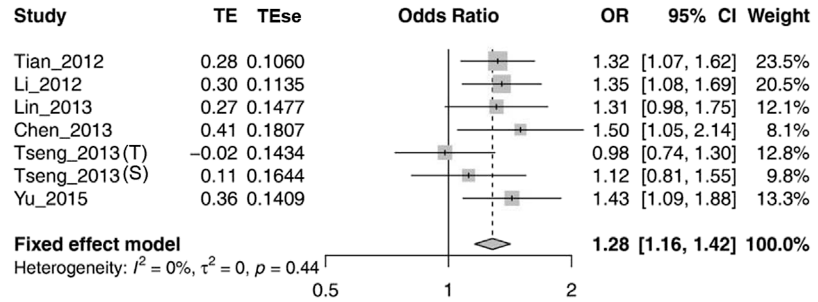


FIGURE 2 | Forest plot for meta-analysis of the association between rs1564282 polymorphism and the risk of PD. Tseng_2013 (T) stands for the Taiwan population in the Tseng_2013 study. Tseng_2013 (S) represents the Singapore population in the Tseng_2013 study. TE = Treatment Effect; TEse = Treatment Effect standard error; CI = Confidence Interval.

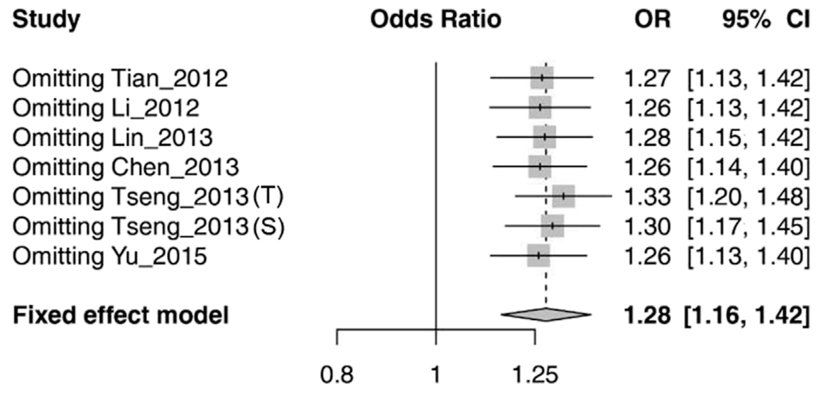


FIGURE 3 | Sensitivity analysis of meta-analysis by omitting every study in turn. Tseng_2013 (T) stands for the Taiwan population in the Tseng_2013 study. Tseng_2013 (S) represents the Singapore population in the Tseng_2013 study. CI = Confidence Interval.

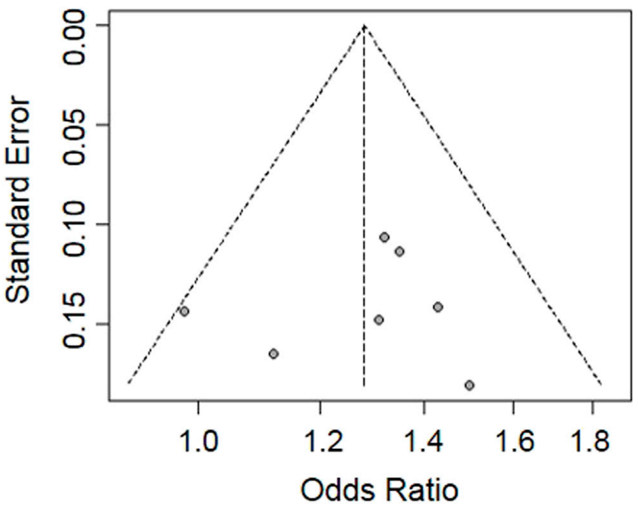


FIGURE 4 | Funnel plot for publication bias analysis of the eligible studies evaluating the relationship between rs1564282 polymorphism and the risk of PD. The *x*-axis and *y*-axis represent the ORs and standard errors for every eligible study, respectively.

et al., 2010; Lill et al., 2012; Nalls et al., 2014). Furthermore, the underlying mechanism of *GAK* in PD has been explored. *GAK* is ubiquitously expressed and participates in various biological processes such as clathrin-mediated membrane traffic and hepatitis C virus entry (Olszewski et al., 2014; Neveu et al., 2015).

In the pathogenesis of PD, rs1564282 was significantly associated with a higher expression level of α-synuclein expression (encoded by *SNCA* gene) in the cortex of PD cases than controls using microarray data (Dumitriu et al., 2011). Dumitriu et al. further investigated the interaction between *GAK* expression and SNCA (Dumitriu et al., 2011). They executed small interfering RNA knockdown of *GAK* in HEK293 cells that overexpressed the SNCA protein and reported that lack of *GAK* expression increased the cytotoxicity based on the overexpression of a-synuclein (Dumitriu et al., 2011).

In addition to the synergistic action with SNCA, evidence also showed that *GAK* impacted the leucine-rich repeat kinase 2 (LRRK2) by forming a complex (Beilina et al., 2014). The gene for LRRK2 has been identified as risk both for monogenic and sporadic PD (Gasser, 2009; Sharma et al., 2012). Beilina et al. utilized the protein–protein arrays to explore the potential interaction mechanisms of LRRK2 in PD pathogenesis (Beilina et al., 2014). The results indicated that *GAK* was a part of a LRRK2-related complex that helped the autophagy–lysosome system to clean vesicles from the Golgi (Beilina et al., 2014).

Nagle's team conducted deep RNA sequencing in human brain tissue from dead PD patients (Nagle et al., 2016). Compared with controls, *GAK* was a unique gene in the 4p16.3 region which had significantly increased expression in PD after adjustment (q value = 4.80E-09) (Nagle et al., 2016). Song et al. studied the function of *auxilin*, the *Drosophila* GAK homolog, via an *in vivo* model (Song et al., 2017). Through systematic experimentation, auxilin was identified as playing a vital role in PD pathogenesis (Song et al., 2017). Researchers proved that reduced auxilin expression resulted in the progressive loss of dopaminergic neurons (Song et al., 2017). Furthermore, the concurrence of reduced auxilin expression and increased SNCA expression accelerated the early death of dopaminergic neurons (Song et al., 2017). Recent evidence showed that *GAK* was one candidate PD gene that had association with N^6-methyladenosine modification (Qiu et al., 2020).

So far, PD GWASs and relevant large-scale meta-analyses have identified tens of risk loci in European population (Hamza et al., 2010; Nalls et al., 2011; Nalls et al., 2014; Nalls et al., 2019). As a vital part of the world population, Chinese population accounts for a certain proportion of global PD patients. Strong evidence provided by Foo team identified that *SNCA*, *LRRK2* and *MCCC1* genes had genome-wide significant associations with PD susceptibility in both Chinese and European population (Foo et al., 2017). In the analysis of risk loci, they inferred that *MAPT* and *GBA* genes might be "European-specific variant loci" (Foo et al., 2017). In subsequent studies, some PD risk loci with genome-wide significance identified in European population had been confirmed to have association in Chinese population, for example *GALC*, *IL1R2*, *SATB1*, *BIN3* and *COQ7* genes (Li et al., 2018; Chen et al., 2019; Hu et al., 2020).

Several studies estimated the underlying association between rs1564282 and PD risk in Chinese populations in China and Singapore (Tseng et al., 2013; Yu et al., 2015). However, the results of these studies were not consistent. We integrated the pooled data of previous studies and conducted a new meta-analysis with 4,055 PD patients and 3,826 controls in all. Firstly, we identified that there was no significant genetic heterogeneity of rs1564282 in the included Chinese populations. Subsequently, the meta-analysis using a fixed-effect model showed a significant association between rs1564282 and PD in Chinese populations. Finally, we performed sensitivity and publication bias analysis. Results showed that the association between rs1564282 and PD was not greatly influenced substantially and that there was no significant publication bias among the eligible studies. In conclusion, our meta-analysis provides good evidence on the risk of the *GAK* rs1564282 variant on PD in Chinese populations.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

HL and YJ conceived and designed the study. HL analyzed data and wrote the manuscript. YJ was responsible for research supervision and manuscript revision. CZ provided technical support. All listed authors approved the final version for submission.

FUNDING

This work was supported by the National Natural Science Foundation (Grant Number 82172282), Science and Technology Project of Tianjin Municipal Health and Health Committee (Grant Nos. ZC20121 and KJ20048).

REFERENCES

Ascherio, A., and Schwarzschild, M. A. (2016). The Epidemiology of Parkinson's Disease: Risk Factors and Prevention. *Lancet Neurol.* 15 (12), 1257–1272. doi:10.1016/s1474-4422(16)30230-7

Beilina, A., Rudenko, I. N., Kaganovich, A., Civiero, L., Chau, H., Kalia, S. K., et al. (2014). Unbiased Screen for Interactors of Leucine-Rich Repeat Kinase 2 Supports a Common Pathway for Sporadic and Familial Parkinson Disease. *Proc. Natl. Acad. Sci. U S A.* 111 (7), 2626–2631. doi:10.1073/pnas.1318306111

Chen, X., Xiao, Y., Guo, W., Zhou, M., Huang, S., Mo, M., et al. (2019). Relationship between Variants of 17 Newly Loci and Parkinson's Disease in a Chinese Population. *Neurobiol. Aging* 73, e1–e231230. doi:10.1016/j.neurobiolaging.2018.08.017

Chen, Y. P., Song, W., Huang, R., Chen, K., Zhao, B., Li, J., et al. (2013). GAK Rs1564282 and DGKQ Rs11248060 Increase the Risk for Parkinson's Disease in a Chinese Population. *J. Clin. Neurosci.* 20 (6), 880–883. doi:10.1016/j.jocn.2012.07.011

Dumitriu, A., Pacheco, C. D., Wilk, J. B., Strathearn, K. E., Latourelle, J. C., Goldwurm, S., et al. (2011). Cyclin-G-associated Kinase Modifies -synuclein Expression Levels and Toxicity in Parkinson's Disease: Results from the GenePD Study. *Hum. Mol. Genet.* 20 (8), 1478–1487. doi:10.1093/hmg/ddr026

Egger, M., Smith, G. D., Schneider, M., and Minder, C. (1997). Bias in Meta-Analysis Detected by a Simple, Graphical Test. *Bmj* 315 (7109), 629–634. doi:10.1136/bmj.315.7109.629

Foo, J. N., Tan, L. C., Irwan, I. D., Au, W.-L., Low, H. Q., Prakash, K.-M., et al. (2017). Genome-wide Association Study of Parkinson's Disease in East Asians. *Hum. Mol. Genet.* 26 (1), ddw379–232. doi:10.1093/hmg/ddw379

Gasser, T. (2009). Molecular Pathogenesis of Parkinson Disease: Insights from Genetic Studies. *Expert Rev. Mol. Med.* 11, e22. doi:10.1017/s1462399409001148

Hamza, T. H., Zabetian, C. P., Tenesa, A., Laederach, A., Montimurro, J., Yearout, D., et al. (2010). Common Genetic Variation in the HLA Region Is Associated with Late-Onset Sporadic Parkinson's Disease. *Nat. Genet.* 42 (9), 781–785. doi:10.1038/ng.642

Hu, X., Mao, C., Hu, Z., Zhang, Z., Zhang, S., Yang, Z., et al. (2020). Association Analysis of 15 GWAS-Linked Loci with Parkinson's Disease in Chinese Han Population. *Neurosci. Lett.* 725, 134867. doi:10.1016/j.neulet.2020.134867

Li, G., Cui, S., Du, J., Liu, J., Zhang, P., Fu, Y., et al. (2018). Association of GALC, ZNF184, IL1R2 and ELOVL7 with Parkinson's Disease in Southern Chinese. *Front. Aging Neurosci.* 10, 402. doi:10.3389/fnagi.2018.00402

Li, N.-N., Chang, X.-L., Mao, X.-Y., Zhang, J.-H., Zhao, D.-M., Tan, E.-K., et al. (2012). GWAS-Linked GAK Locus in Parkinson's Disease in Han Chinese and Meta-Analysis. *Hum. Genet.* 131 (7), 1089–1093. doi:10.1007/s00439-011-1133-3

Lill, C. M., Roehr, J. T., McQueen, M. B., Kavvoura, F. K., Bagade, S., Schjeide, B.-M. M., et al. (2012). Comprehensive Research Synopsis and Systematic Meta-Analyses in Parkinson's Disease Genetics: The PDGene Database. *Plos Genet.* 8 (3), e1002548. doi:10.1371/journal.pgen.1002548

Lin, C.-H., Chen, M.-L., Tai, Y.-C., Yu, C.-Y., and Wu, R.-M. (2013). Reaffirmation of GAK, but Not HLA-DRA, as a Parkinson's Disease Susceptibility Gene in a Taiwanese Population. *Am. J. Med. Genet.* 162 (8), 841–846. doi:10.1002/ajmg.b.32188

Liu, G., Xu, Y., Jiang, Y., Zhang, L., Feng, R., and Jiang, Q. (2017). PICALM Rs3851179 Variant Confers Susceptibility to Alzheimer's Disease in Chinese Population. *Mol. Neurobiol.* 54 (5), 3131–3136. doi:10.1007/s12035-016-9886-2

Liu, G., Zhang, S., Cai, Z., Ma, G., Zhang, L., Jiang, Y., et al. (2013). PICALM Gene Rs3851179 Polymorphism Contributes to Alzheimer's Disease in an Asian Population. *Neuromol Med.* 15 (2), 384–388. doi:10.1007/s12017-013-8225-2

Nagle, M. W., Latourelle, J. C., Labadorf, A., Dumitriu, A., Hadzi, T. C., Beach, T. G., et al. (2016). The 4p16.3 Parkinson Disease Risk Locus Is Associated with GAK Expression and Genes Involved with the Synaptic Vesicle Membrane. *PLoS One* 11 (8), e0160925. doi:10.1371/journal.pone.0160925

Nalls, M. A., Blauwendraat, C., Vallerga, C. L., Heilbron, K., Bandres-Ciga, S., Chang, D., et al. (2019). Identification of Novel Risk Loci, Causal Insights, and Heritable Risk for Parkinson's Disease: a Meta-Analysis of Genome-wide Association Studies. *Lancet Neurol.* 18 (12), 1091–1102. doi:10.1016/s1474-4422(19)30320-5

Nalls, M. A., Nalls, M. A., Plagnol, V., Hernandez, D. G., Sharma, M., Sheerin, U. M., et al. (2011). Imputation of Sequence Variants for Identification of Genetic Risks for Parkinson's Disease: a Meta-Analysis of Genome-wide Association Studies. *Lancet* 377 (9766), 641–649. doi:10.1016/s0140-6736(10)62345-8

Nalls, M. A., Pankratz, N., Pankratz, N., Lill, C. M., Do, C. B., Hernandez, D. G., et al. (2014). Large-scale Meta-Analysis of Genome-wide Association Data Identifies Six New Risk Loci for Parkinson's Disease. *Nat. Genet.* 46 (9), 989–993. doi:10.1038/ng.3043

Neveu, G., Ziv-Av, A., Barouch-Bentov, R., Berkerman, E., Mulholland, J., and Einav, S. (2015). AP-2-associated Protein Kinase 1 and Cyclin G-Associated Kinase Regulate Hepatitis C Virus Entry and Are Potential Drug Targets. *J. Virol.* 89 (8), 4387–4404. doi:10.1128/jvi.02705-14

Olszewski, M. B., Chandris, P., Park, B.-C., Eisenberg, E., and Greene, L. E. (2014). Disruption of Clathrin-Mediated Trafficking Causes Centrosome Overduplication and Senescence. *Traffic* 15 (1), 60–77. doi:10.1111/tra.12132

Pankratz, N., Wilk, J. B., Latourelle, J. C., DeStefano, A. L., Halter, C., Pugh, E. W., et al. (2009). Genomewide Association Study for Susceptibility Genes Contributing to Familial Parkinson Disease. *Hum. Genet.* 124 (6), 593–605. doi:10.1007/s00439-008-0582-9

Qiu, X., He, H., Huang, Y., Wang, J., and Xiao, Y. (2020). Genome-wide Identification of m6A-Associated Single-Nucleotide Polymorphisms in Parkinson's Disease. *Neurosci. Lett.* 737, 135315. doi:10.1016/j.neulet.2020.135315

Sharma, M., Ioannidis, J. P. A., Aasly, J. O., Annesi, G., Brice, A., Van Broeckhoven, C., et al. (2012). Large-scale Replication and Heterogeneity in Parkinson Disease Genetic Loci. *Neurology* 79 (7), 659–667. doi:10.1212/WNL.0b013e318264e353

Song, L., He, Y., Ou, J., Zhao, Y., Li, R., Cheng, J., et al. (2017). Auxilin Underlies Progressive Locomotor Deficits and Dopaminergic Neuron Loss in a Drosophila Model of Parkinson's Disease. *Cel Rep.* 18 (5), 1132–1143. doi:10.1016/j.celrep.2017.01.005

Spencer, C. C., Spencer, C. C. A., Plagnol, V., Strange, A., Gardner, M., Paisan-Ruiz, C., et al. (2011). Dissection of the Genetics of Parkinson's Disease Identifies an Additional Association 5' of SNCA and Multiple Associated Haplotypes at 17q21. *Hum. Mol. Genet.* 20 (2), 345–353. doi:10.1093/hmg/ddq469

Tian, J. (2012). Polymorphism Analysis and Poly Gene Association Analysis of PARK16, PARK17, PARK18 and BST1 Genes in Parkinson's Disease. China National Knowledge Infrastructure Available at: http://cdmd.cnki.com.cn/Article/CDMD-10533-1012475946.htm.

Tseng, W.-E. J., Chen, C.-M., Chen, Y.-C., Yi, Z., Tan, E.-K., and Wu, Y.-R. (2013). Genetic Variations of GAK in Two Chinese Parkinson's Disease Populations: a Case-Control Study. *PLoS One* 8 (6), e67506. doi:10.1371/journal.pone.0067506

Yu, W.-J., Cheng, L., Li, N.-N., Wang, L., Tan, E.-K., and Peng, R. (2015). Interaction between SNCA, LRRK2 and GAK Increases Susceptibility to Parkinson's Disease in a Chinese Population. *eNeurologicalSci* 1 (1), 3–6. doi:10.1016/j.ensci.2015.08.001

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

SYSTEMATIC REVIEW
published: 22 November 2021
doi: 10.3389/fgene.2021.774489

Association Between Insulin-like Growth Factor-1 rs35767 Polymorphism and Type 2 Diabetes Mellitus Susceptibility: A Meta-Analysis

Qiaoli Zeng[1,2,3†], Dehua Zou[2,3,4†], Qiaodi Zeng[5†], Xiaoming Chen[6]*, Yue Wei[7]* and Runmin Guo[1,2,3,6]*

[1]Department of Internal Medicine, Shunde Women and Children's Hospital (Maternity and Child Healthcare Hospital of Shunde Foshan), Guangdong Medical University, Foshan, China, [2]Key Laboratory of Research in Maternal and Child Medicine and Birth Defects, Guangdong Medical University, Foshan, China, [3]Shunde Women and Children's Hospital (Maternity and Child Healthcare Hospital of Shunde Foshan), Matenal and Child Research Institute, Guangdong Medical University, Foshan, China, [4]State Key Laboratory for Quality Research of Chinese Medicines, Macau University of Science and Technology, Taipa, Macau (SAR) China, [5]Department of Clinical Laboratory, People's Hospital of Haiyuan County, Zhongwei, China, [6]Department of Endocrinology, Affiliated Hospital of Guangdong Medical University, Zhanjiang, China, [7]Department of Ultrasound, Shunde Women and Children's Hospital (Maternity and Child Healthcare Hospital of Shunde Foshan), Guangdong Medical University, Foshan, China

OPEN ACCESS

Edited by:
Liangcai Zhang,
Janssen Research and Development,
United States

Reviewed by:
Tonia Carter,
Marshfield Clinic, United States
Yanfang Chen,
Wright State University, United States

***Correspondence:**
Xiaoming Chen
chenxiaoming@gdmu.edu.cn
Yue Wei
weiyue138@163.com
Runmin Guo
runmin.guo@gdmu.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: 12 September 2021
Accepted: 25 October 2021
Published: 22 November 2021

Citation:
Zeng Q, Zou D, Zeng Q, Chen X, Wei Y and Guo R (2021) Association Between Insulin-like Growth Factor-1 rs35767 Polymorphism and Type 2 Diabetes Mellitus Susceptibility: A Meta-Analysis. Front. Genet. 12:774489. doi: 10.3389/fgene.2021.774489

Background: Insulin-like growth factor-1 (IGF-1) has been demonstrated to increase fatty acid β oxidation during fasting, and play an important role in regulating lipid metabolism and type 2 diabetes mellitus (T2DM). The rs35767 (T > C) polymorphism, a functional SNP was found in *IGF-1* promoter, which may directly affect *IGF-1* expression. However, the inconsistent findings showed on the *IGF-1* rs35767 polymorphism and T2DM risk.

Methods: We performed a comprehensive meta-analysis to estimate the association between the *IGF-1* rs35767 and T2DM risk among four genetic models (the allele, additive, recessive and dominant models).

Results: A total 49,587 T2DM cases and 97,906 NDM controls were included in the allele model, a total 2256 T2DM cases and 2228 NDM controls were included in the other three genetic models (the additive; recessive and dominant models). In overall analysis, the *IGF-1* rs35767 was shown to be significantly associated with increased T2DM risk for the allele model (T vs. C: OR = 1.251, 95% CI: 1.082–1.447, *p* = 0.002), additive model (homozygote comparisons: TT vs. CC: OR = 2.433, 95% CI: 1.095–5.405, *p* = 0.029; heterozygote comparisons: TC vs. CC: OR = 1.623, 95% CI: 1.055–2.495, *p* = 0.027) and dominant model (TT + CT vs. CC: OR = 1.934, 95% CI: 1.148–3.257, *p* = 0.013) with random effects model. After omitting Gouda's study could reduce the heterogeneity, especially in the recessive model (TT vs. CC + CT: I^2 = 38.7%, *p* = 0.163), the fixed effects model for recessive effect of the T allele (TT vs. CC + CT) produce results that were of borderline statistical significance (OR = 1.206, 95% CI: 1.004–1.448, *p* = 0.045). And increasing the risk of T2DM in Uyghur population of subgroup for the allele model.

Conclusion: The initial analyses that included all studies showed statistically significant associations between the rs35767 SNP and type 2 diabetes, but after removing the Gouda et al. study produced results that were mostly not statistically significant. Therefore, there is not enough evidence from the results of the meta-analysis to indicate that the rs35767 SNP has a statistically significant association with type 2 diabetes.

Keywords: type 2 diabete mellitus, insulin-like growth factor-1, rs35767, susceptibility, meta-analysis

1 INTRODUCTION

Diabetes is one of the most common chronic metabolic disorder diseases in the worldwide, over 90% of the diabetes patients are type 2 diabetes mellitus (T2DM), which is characterized by insulin resistance in peripheral tissues and dysregulated insulin secretion by pancreatic beta (β) cells (Banerjee and Vats, 2014; Song et al., 2015). Substantial evidence suggests that insulin resistance, an inherited genetic defect, is the basis and major feature of T2DM (DeFronzo and Tripathy, 2009; Cai et al., 2019; Li et al., 2021). Insulin resistance is attributable to excess fatty acids and proinflammatory cytokines, which leads to impaired glucose transport and increases fat breakdown. Since there is an inadequate response or production of insulin, the body responds by inappropriately increasing glucagon, thus further contributing to hyperglycemia. Accumulated data have revealed that lipid abnormalities are associated with insulin resistance and contribute to T2DM (Johnson and Olefsky, 2013; Perry et al., 2014; Li et al., 2021). Studies have also revealed lipid metabolism-related genes and their single-nucleotide polymorphisms (SNPs) associated with insulin resistance and the development of T2DM (Ruchat et al., 2009; Dupuis et al., 2010; Chistiakov et al., 2012; Langberg et al., 2012; Mannino et al., 2013; Li et al., 2014; Thankamony et al., 2014; Yuan et al., 2015; Li et al., 2021).

Insulin-like growth factor-1 (*IGF-1*) is a circulating growth factor which structure is highly homologous with pro-insulin. *IGF-1* is expresses in insulin-resistant tissue, it downregulates free fatty acid and increases fatty acid β oxidation during fasting (Thankamony et al., 2014; Li et al., 2021). It plays an important role in regulating lipid metabolism and insulin sensitivity (Seppä et al., 2015; Gouda et al., 2019), since it effects on glucose homeostasis and associated with insulin resistance (Li et al., 2011; Li et al., 2014; Dai et al., 2015; Ming et al., 2015; Yuan et al., 2015; Wei et al., 2018; Liao et al., 2019; Regué et al., 2019; Li et al., 2021). Previous studies have been reported that people with a low *IGF-1* level are prone to have diabetes mellitus (Chen et al., 2013; Colao et al., 2013; Shankar and Li, 2013). Polymorphisms in the *IGF-1* gene can directly affect *IGF-1* expression. The rs35767 (T > C) polymorphism, a functional SNP was found in *IGF-1* promoter, in which the promoter with C allele showed a higher transcriptional activity than promoter with T allele (Telgmann et al., 2009). Therefore, rs35767 may contribute to insulin resistance involving lipid metabolism in T2DM.

A significant association of *IGF-1* rs35767 with T2DM has been reported in several case-control studies (Gouda et al., 2019; Gulixiati·Maimaitituersun. 2020; Wang. 2019; Zhang et al., 2017; Song et al., 2015). However, seven studies failed to replicate the results (Dupuis et al., 2010; Hu et al., 2010; Fujita et al., 2012; Liu et al., 2012; Zhao et al., 2016; Li et al., 2019; Li et al., 2021). In veiw of the inconsistent results, whether *IGF-1* rs35767 is associated with T2DM remains to be determined. In this meta-analysis, we estimate the association of *IGF-1* rs35767 with T2DM among four different genetic models.

2 MATERIALS AND METHODS

2.1 Literature Search

The Google Scholar, PubMed and Chinese National Knowledge Infrastructure were comprehensively searched for related studies published before July 31, 2021, using the key terms: “insulin-like growth factor 1 or *IGF-1* or *IGF1*,” “rs35767 or rs35767 (T > C) or rs35767 (A > G),” “polymorphism or SNP or mutation or variant” and “diabetes or type 2 diabetes or T2DM.” All searches had no language limitations. Eligible studies were estimated by reading full texts, and excluded substandard studies.

2.2 Inclusion and Exclusion Criteria

The following inclusion criteria: 1) case-control or cohort studies that relate to the *IGF-1* rs35767 and T2DM risk; 2) sufficient raw data or adequate data for assessing odds ratios (ORs) with corresponding 95% confidence intervals (CIs); and 3) The diagnostic standard of T2DM conformed to the World Health Organization.

The following exclusion criteria: 1) not a case-control study; 2) irrelevant to *IGF-1* rs35767 and T2DM risk; 3) lacking detailed data; and 4) control subjects is not in Hardy-Weinberg equilibrium (HWE).

2.3 Data Extraction

Data were independently extracted by two authors from the eligible studies and collected the following data: first author, year of publication, origin, the numbers of T2DM cases and NDM controls, gender and age, BMI (kg/m^2), the distributions number of genotype and alleles, ORs with 95% CI, or ability to calculate the OR and 95% CI. *p*-value for the HWE of NDM controls.

2.4 Statistical Analysis

Statistical analyses using the STATA v.14.0 software (Stata Corporation, TX, United States). Four genetic models were evaluated in this meta-analysis: the allele model (T vs. C); the additive model (homozygote comparisons: TT vs. CC; heterozygote comparisons: TC vs. CC); the recessive model

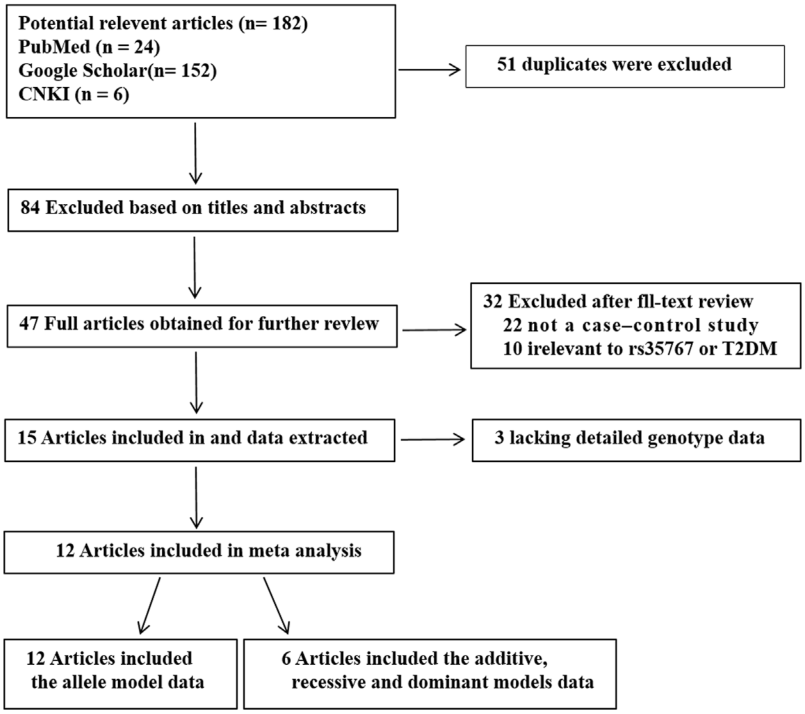

Potential relevent articles (n= 182)
PubMed (n = 24)
Google Scholar(n= 152)
CNKI (n = 6)
51 duplicates were excluded
84 Excluded based on titles and abstracts
47 Full articles obtained for further revlew
32 Excluded after fll-text review
22 not a case–control study
10 irelevant to rs35767 or T2DM
15 Articles included in and data extracted
3 lacking detailed genotype data
12 Articles included in meta analysis
12 Articles included
the allele model data
6 Articles included the additive,
recessive and dominant models data

FIGURE 1 | Flow chart of researches selection in the meta-analysis.

(TT vs. CC + CT) and the dominant model (TT + CT vs. CC). Using Q-test and I^2 test to estimate the genetic heterogeneity. OR with corresponding 95% CIs were calculated by the random effectss model when $p < 0.01$ and $I^2 > 50\%$ (He et al., 2015; Zhang et al., 2016). Otherwise, the fixed effectss model were used. Sensitivity analyses were implemented to evaluate the stability of the overall effect by excluding a study at a time. The Hardy-Weinberg equilibrium for the NDM controls was assessed using Pearson's Chi-squared test. Using Bgger's test to evaluate publication bias (Shen et al., 2015; Li et al., 2016; Liu et al., 2017; Han et al., 2019).

3 RESULTS

3.1 Study Inclusion and Characteristics

A total of 182 potential articles obtained through initial search. 51 duplicates were excluded. Then 131 studies were screened on title and abstract, 84 of them were excluded. The left 47 articles were evaluated by full-text reading, 35 of them were excluded cause that 22 were not case-control researchs, 10 were not related to rs35767 or T2DM, three did not provided sufficient data. 12 articles were included that there are six articles including five in English and one in Chinese just of the allele model data, and other six articles including four in English and two in Chinese of four genetic models data (the allele, additive, recessive and dominant models). Flow chart of researches selection in the meta-analysis was shown in **Figure 1**. A total 49,587 T2DM cases and 97,906 NDM controls were included in the allele model, a total 2256 T2DM cases and 2228 NDM controls were included in the other three genetic models (the additive; recessive and dominant models). The characteristics of each study are shown in **Table 1** and **Supplementary Table S1**.

3.2 Meta-Analysis

The association between the *IGF-1* rs35767 polymorphism and T2DM were evaluated using ORs and 95% CI in the allele model (12 studies, 49587 T2DM cases and 97906 NDM controls) and the additive; recessive and dominant models (6 studies, 2256 T2DM cases and 2228 NDM controls).

In overall analysis, A random effects model were used to analyze the allele, additive, recessive and dominant models. The *IGF-1* rs35767 was shown to be significantly associated with increased T2DM risk for the allele model (T vs. C: OR = 1.251, 95% CI: 1.082–1.447,p = 0.002), additive model (homozygote comparisons: TT vs. CC: OR = 2.433, 95% CI: 1.095–5.405, p = 0.029; heterozygote comparisons: TC vs. CC: OR = 1.623, 95% CI: 1.055–2.495, p = 0.027) and dominant model (TT + CT vs. CC: OR = 1.934, 95% CI: 1.148–3.257, p = 0.013). The results showed no significant difference for the recessive model (TT vs. CC + CT: OR = 1.876, 95% CI: 0.989–3.559, p = 0.054) (**Figure 2**).

3.3 Sensitivity Analysis

Aiming to estimate the influence of each study on the overall OR below four genetic models and analysis the sources of high heterogeneity, sensitivity meta-analysis was performed with

TABLE 1 | Characteristics of each study included in this meta-analysis.

					Allele distribution				Genotype distribution					
					T2DM, n		NDM, n		T2DM, n			NDM, n		
Authors	**Origin**	**Gender**	**T2DM/ NDM, n**	**ORs with 95% CI (T vs. C)**	**C**	**T**	**C**	**T**	**CC**	**CT**	**TT**	**CC**	**CT**	**TT**
Li et al. (2021)	Chinese (Yunnan)	M/F	1194/ 1274	0.928 (0.826–1.042)	1538	850	1597	951	513	512	169	500	597	177
Gulixiati et al. (2020)	Chinese (Xinjiang)	M/F	220/229	1.452 (1.092–1.931)	287	153	335	123	93	101	26	120	65	14
Gouda et al. (2019)	Egyptian	F	180/165	5.103 (3.641–7.153)	72	288	185	145	12	48	120	60	65	40
Wang et al. (2019)	Chinese (Tianjin)	M/F	367/367	1.322 (1.065–1.641)	460	274	506	228	146	168	53	176	154	37
Zhang et al. (2017)	Chinese (Hebei)	M/F	244/142	1.388 (1.025–1.879)	280	208	185	99	77	126	41	56	73	13
Song et al. (2015)	Chinese (Xinjiang)	M/F	51/51	1.900 (1.006–3.587)	67	33	81	21	21	25	4	34	13	4
Li et al. (2019)	Chinese (Tianjin)	F	80/1160	1.043 (0.727–1.494)	–	–	–	–	–	–	–	–	–	–
Zhao et al. (2016)	Chinese (Xinjiang)	M/F	130/135	1.480 (0.980-2.230)	–	–	–	–	–	–	–	–	–	–
Fujita et al. (2012)	Japanese	M/F	2632/ 2050	0.990 (0.912-1.075)	–	–	–	–	–	–	–	–	–	–
Liu et al. (2011)	Chinese (Beijing, Shanghai)	M/F	424/ 1899	0.935 (0.795-1.099)	–	–	–	–	–	–	–	–	–	–
Hu et al. (2010)	Chinese (Shanghai)	M/F	3410/ 3412	1.027 (0.956-1.103)	–	–	–	–	–	–	–	–	–	–
Dupuis et al. (2010)	European	M/F	40655/ 87022	0.962 (0.890-1.038)	–	–	–	–	–	–	–	–	–	–

n, Number; M, Male; F, Female; T2DM, type 2 diabetes mellitus; NDM, Non-diabetic subject;OR, odds ratio; CI, confidence interval; (-), not applicable.

random effects model. The results were showed in **Figure3**, omitting Gouda's study could reduce the heterogeneity, especially in the recessive model (TT vs. CC + CT: I^2 = 38.7%, p = 0.163), the fixed effects model for recessive effect of the T allele (TT vs. CC + CT) produce results that were of borderline statistical significance (OR = 1.206, 95% CI: 1.004–1.448, p = 0.045); In addtion, other three modal show moderate degree of heterogeneity (T vs. C: I^2 = 64.9%, p = 0.002; TT vs. CC: I^2 = 70.3%, p = 0.009; TC vs. CC: I^2 = 84.0%, p = 0.000; TT + CT vs. CC: I^2 = 85.8%, p = 0.000, respectively), the result showed no significant association between *IGF-1* rs35767 and T2DM risk with random effects model (T vs. C: OR = 1.065, 95% CI: 0.983–1.153, p = 0.126; TT vs. CC: OR = 1.603, 95% CI: 0.996–2.578, p = 0.052; TC vs. CC: OR = 1.407, 95% CI: 0.937–2.112, p = 0.099; TT + CT vs. CC: OR = 1.469, 95% CI: 0.978–2.207, p = 0.064, respectively) (**Figure 4**). Since there are not sufficient evidence to draw the conclusion that the rs35767 SNP is associated with T2DM.

3.4 Subgroup-Analyses

As high heterogeneity was observed, we performed subgroup-analysis according to origin to evaluate the association between rs35767 and T2DM susceptibility in the allele model. The results suggested that rs35767 was significantly related to the risk of T2DM in *XinJiang*, China subgroup (T vs. C: OR = 1.508, 95% CI: 1.210–1.878, p = 0.000) with fixed effects model; a random effects model were used to analyze the other provinces, China, rs35767 was shown no significant association with T2DM risk (T vs. C: OR = 1.051, 95% CI: 0.943–1.173, p = 0.369); and not a significantly associatied in the other countries subgroup (excluding Gouda et al. Literature) (T vs. C: OR = 0.975, 95% CI: 0.922–1.031, p = 0.376) with fixed effects model (**Figure 5**).

3.5 Publication Bias

The funnel plot was showed to be visually symmetrical (**Supplementary Figure S1**, **Supplementary Figure S2**, **Supplementary Figure S3**). Begg's and Egger's tests were performed to detect publication bias. There was no significant publication bias appeared in all genetic models in overall analysis *via* Begg's test (all p > 0.05, **Supplementary Table S2**), but for Egger's test, there was publication bias in the additive model (heterozygote comparisons) (Egger's test, p = 0.006, **Supplementary Table S2**). We did not determine publication bias for Begg's test after omitting Gouda's study and subgroup analysis in all genetic models (all p > 0.05, **Supplementary Table S3**, **Supplementary Table S4**). However, for Egger's test, there were publication bias in the allele model (Egger's test, p = 0.039, **Supplementary Table S3**) and additive model (heterozygote comparisons) (Egger's test, p = 0.031, **Supplementary Table S4**).

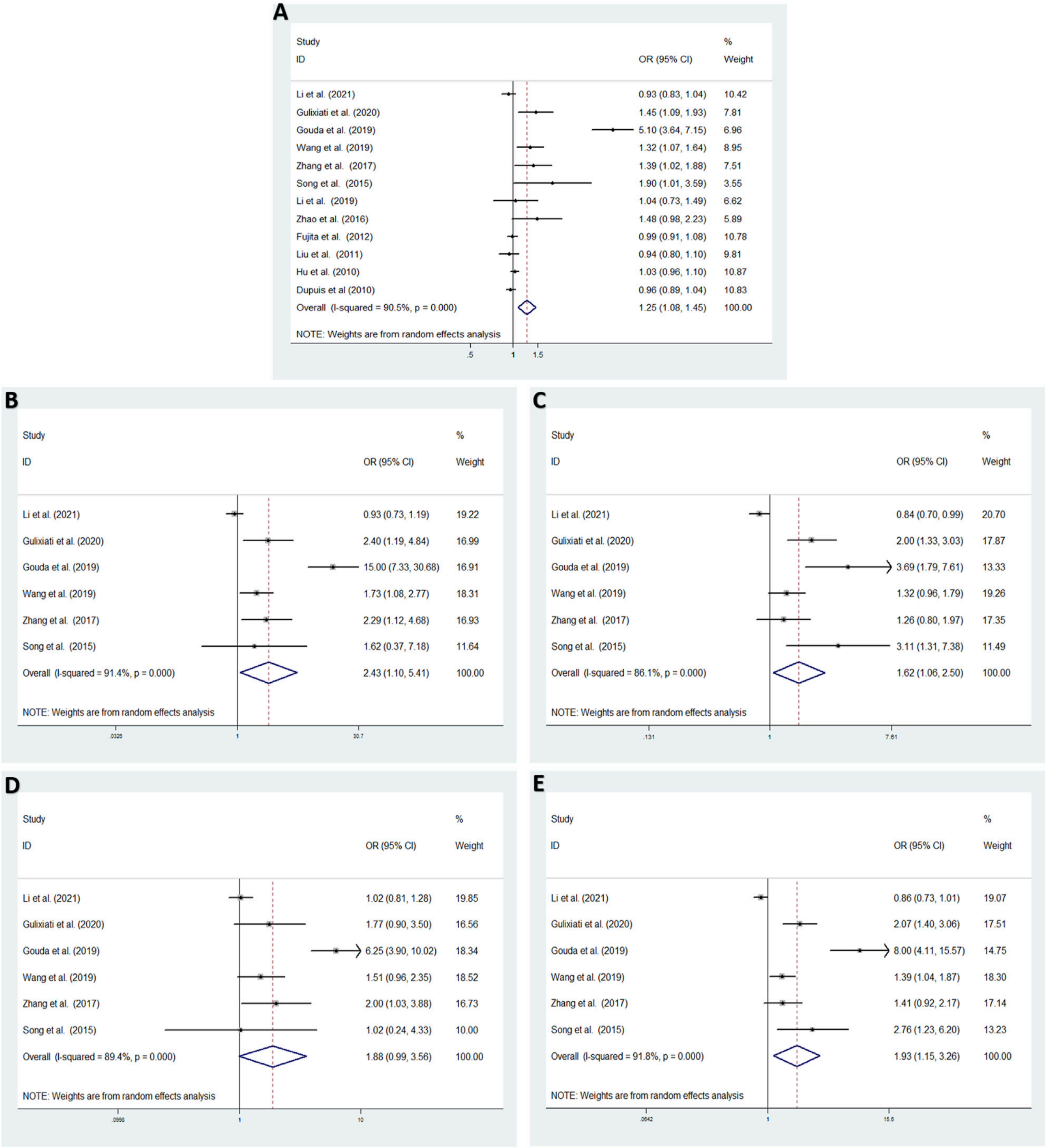


FIGURE 2 | Meta-analysis with a random effects model for the association between the *IGF-1* rs35767 and T2DM susceptibility. **(A)** Allele model, T vs. C. **(B)** Additive model (homozygote comparisons): TT vs. CC.**(C)** Additive model (heterozygote comparisons): TC vs. CC. **(D)** Recessive model, TT vs. CC + CT.**(E)** Dominant model, TT + CT vs. CC. OR: odds ratio, CI: confidence interval, I-squared: measure to quantify the degree of heterogeneity in meta-analyses.

4 DISCUSSION

Previous studies have showed the inconsistent findings of association between *IGF-1* rs35767 and the risk of T2DM. Gouda et al. revealed that the TT, TT + CT genotypes of rs35767 were associated with an increased risk of T2DM in pregnant Egyptian women respectively (Gouda et al., 2019). This results were successfully replicated in a Chinese Han

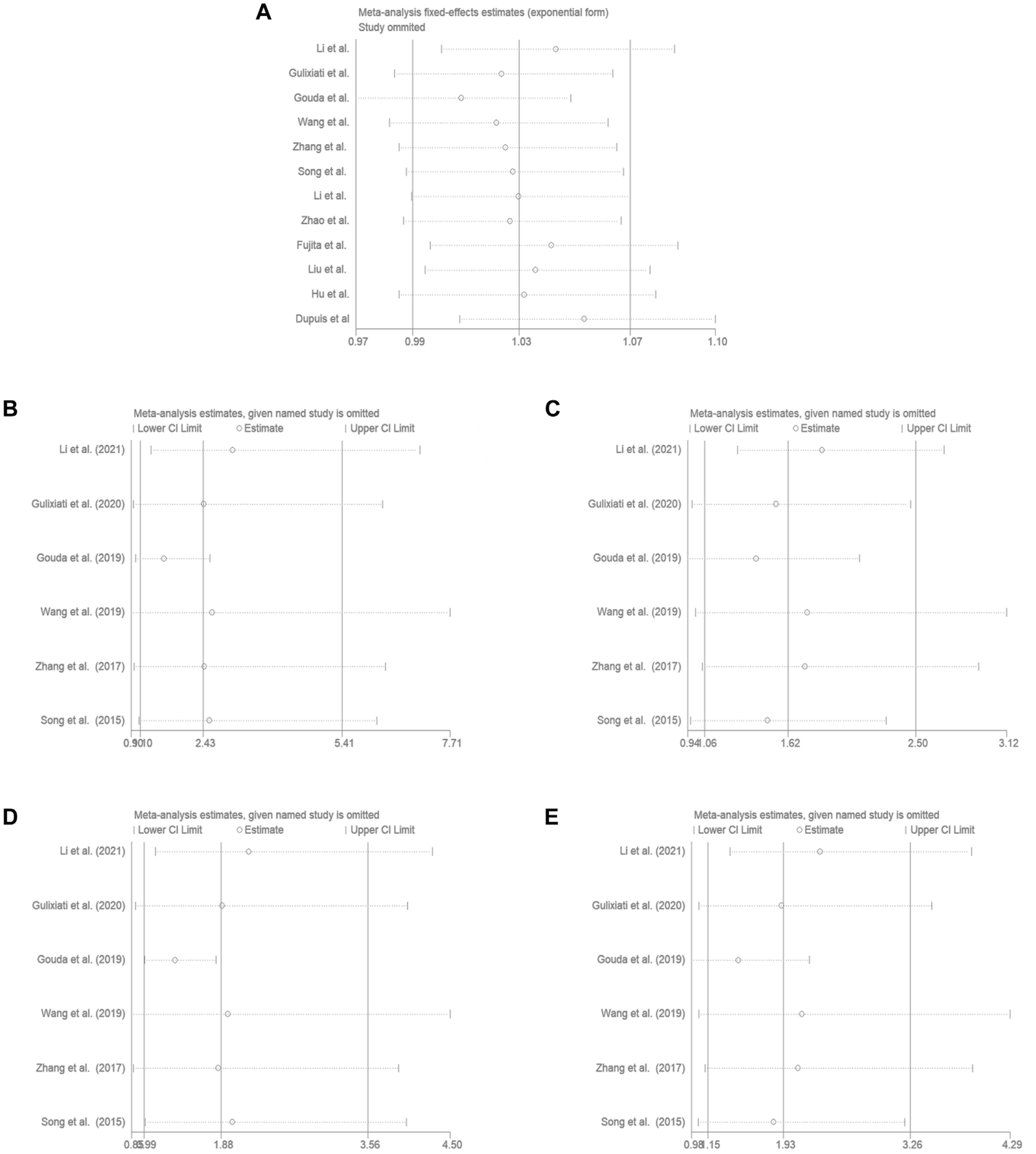


FIGURE 3 | Sensitivity analysis by iteratively removing one study at a time. **(A)** Allele model, T vs. C. **(B)** Additive model (homozygote comparisons): TT vs. CC. **(C)** Additive model (heterozygote comparisons): TC vs. CC. **(D)** Recessive model, TT vs. CC + CT. **(E)** Dominant model, TT + CT vs. CC.

population (Wang. 2019). Zhang et al. found that the A allele of rs35767 contributed to the risk of developing T2DM in a Chinese Han population (Zhang.et al., 2017). More recently, two studies documented that the association of the rs35767 in *IGF-1* was associated with T2DM in a Uyghur population in China (GulixiatiMaimaitituersun. 2020; Song et al., 2015). However, some studies did not find evidence of an association between rs35767 and T2DM (Dupuis et al.,

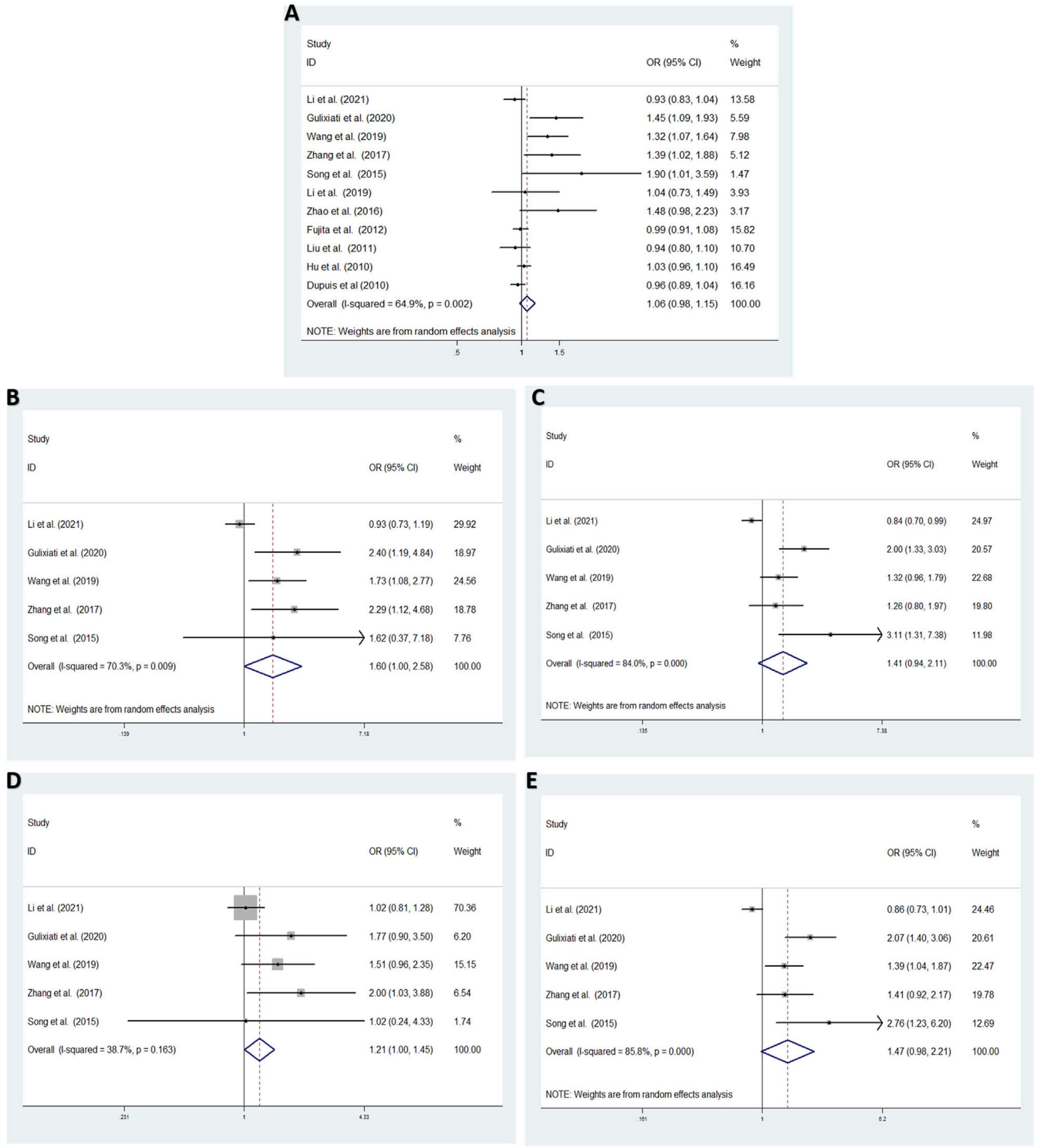


FIGURE 4 | Meta-analysis for the association between the *IGF-1* rs35767 and T2DM susceptibility after omitting Gouda's study. **(A)** Allele model, T vs. C (random effects model). **(B)** Additive model (homozygote comparisons): TT vs. CC. (random effects model). **(C)** Additive model (heterozygote comparisons): TC vs. CC (random effects model) **(D)** Recessive model, TT vs. CC + CT (fixed effects model) **(E)** Dominant model, TT + CT vs. CC (random effects model). OR: odds ratio, CI: confidence interval, I-squared: measure to quantify the degree of heterogeneity in meta-analyses.

2010; Hu et al., 2010; Fujita et al., 2012; Liu et al., 2012; Zhao et al., 2016; Li et al., 2019; Li et al., 2021). It is worth noting that the four largest studies with the most statistical power (Dupuis et al., Hu et al., Fujita et al., and Li et al., 2021) did not report statistically significant associations between the rs35767 SNP and T2DM, whereas five small studies (Song et al., Zhang et al.,

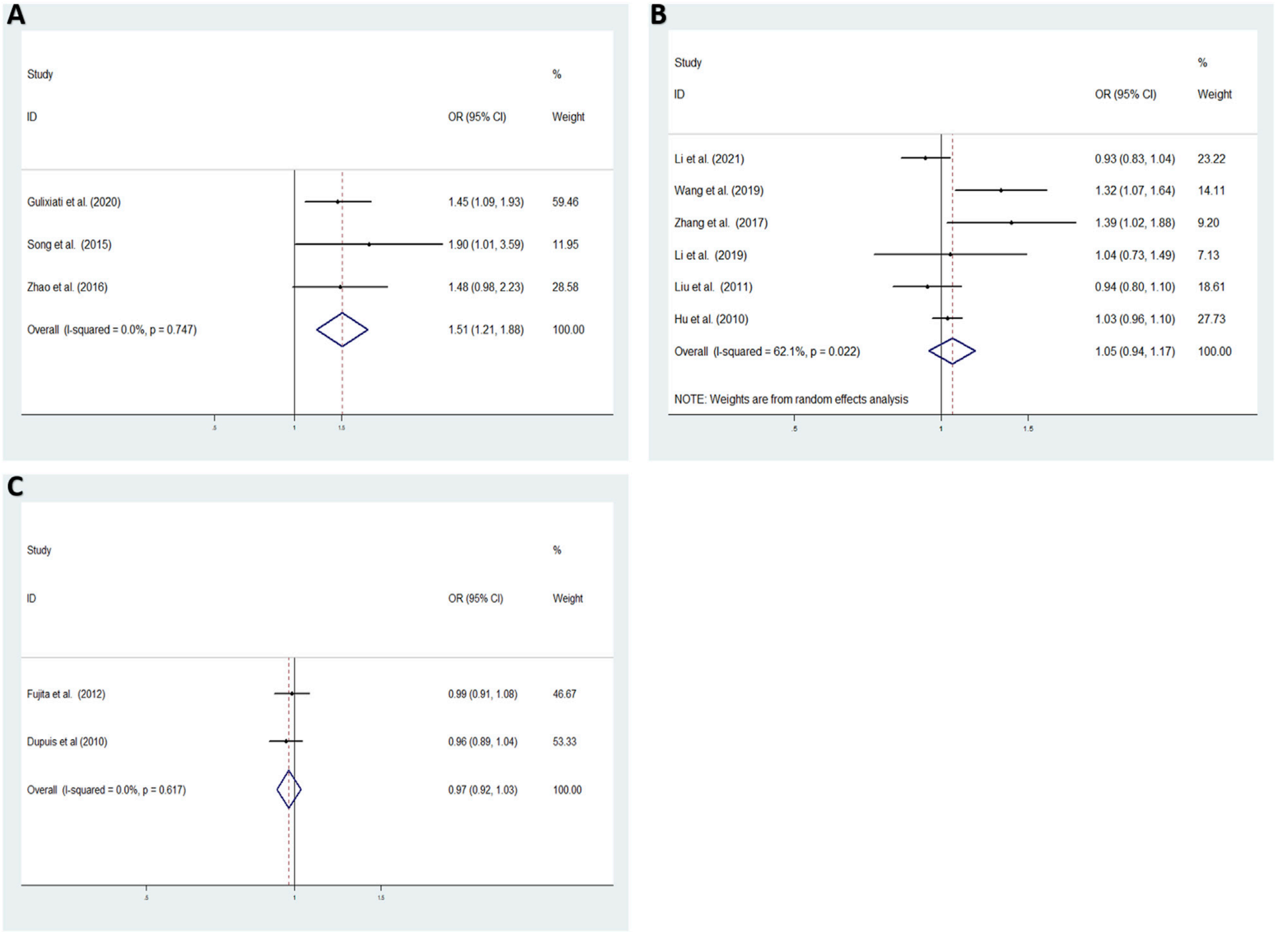


FIGURE 5 | The association between the *IGF-1* rs35767 and T2DM susceptibility in the subgroup for the allele model (T vs. C) **(A)** *XinJiang*, China (fixed effects model) **(B)** *Other provinces*, China (random effects model) **(C)** Other countries (fixed effects model). OR: odds ratio, CI: confidence interval, I-squared: measure to quantify the degree of heterogeneity in meta-analyses.

Wang et al., Gouda et al., and Gulixiati et al.) all reported statistically significant associations. Compared the subjects selected for the studies between the larger and smaller studies, we found that in the Gouda et al. study, the mean body mass index (BMI) of subjects with T2DM was 34.26 ± 5.7, which is very different from a mean BMI of 26.96 ± 4.57 for control subjects without T2DM. There were a significant difference concerning BMI between T2DM and control groups. This study included only pregnant women, was observed to be very influential on the initial analyses.

Previous studies have been reported that people with a low *IGF-1* level are prone to have diabetes mellitus (Colao et al., 2013; Shankar and Li, 2013). The functional SNP rs35767 (T > C) in *IGF-1* promoter, with C allele showed a higher transcriptional activity than promoter with T allele (Telgmann et al., 2009). In terms of mechanism, the higher transcriptional activity of the C allele *IGF-1* promoter was contributed by the C/EBPD transcription activator, which bound exclusively to the C allele, but not to the T allele (Chen et al., 2013; Telgmann et al., 2009). Therefore, there may have low *IGF-1* expression level when promoter with rs35767 T allele, which contribute to the development of T2DM. In our meta-analysis, we found that T allele, TT genotype, TT + CT genotype of rs35767 increased T2DM risk in overall analysis, as well as increasing the risk of T2DM in Uyghur population. After omitting Gouda's study, the result showed TT genotype were of borderline statistical significance.

In overall analysis, high heterogeneity among studies were detected in four genetic models, which might be a result of the difference in ethnicity, country, genetic background and environmental factors (e.g., dietary, life style, climates) (Qin et al., 2010). Then omittied Gouda's study, the heterogeneity was reduced. We found that the subjects of Gouda's study were pregnant Egyptian women, but participants of other studies were both male and female. Thus gender ratio may also had a certain impact on heterogeneity. The subgroup-analyses were detected by origin in allele model, the subgroup of Uyghur in Xinjiang, China have no heterogeneity, but other subgroups still had high heterogeneity. It is noteworthy that a previous study in Xinjiang found that 19.6% of Uyghur had diabetes, exceptionally higher than that in Kazakh (7.3%) and Han Chinese (9.1%) (Li et al., 2012; Song et al., 2015). The marriage pattern and unique life style might be responsible for the observation. One hand, the practice of endogamy in Uyghur population might also be a reason (Wang et al., 2003; Mamet et al., 2005). On the other hand, The most Uyghurs have different dietary habits from Han Chinese. They have more meat, high carbohydrate diets with a higher salt (more than 20 g per day) and less unsaturated fatty acids compared with Han Chinese (Zhai et al., 2007).

There still have several limitations in our meta-analysis. Firstly, there were limited studies which estimated *IGF-1* rs35767 and T2DM risk, only six articles had four gene models data, and the other six articles had only one allele model data. Secondly, the results did not adjustment the potential risk factors, including gender, body mass index, age, drinking and smoking status, and environmental factors. Thirdly, some results showed relatively obvious heterogeneity, but research the source of heterogeneity needs to more larger sample. Finally, some groups results existed potential publication bias in Egger's test. Therefore, the results of the article should be interpreted carefully.

5 CONCLUSION

This is the first time to perform a meta-analysis to systematically summarize the association between the *IGF-1* rs35767 and T2DM susceptibility. Overall, there is not enough evidence from the results of the meta-analysis to indicate that the rs35767 SNP has a statistically significant association with T2DM. Further more studies are necessary to verify the results.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material** further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

QZ, DZ, and QdD were suitable for the study design, literature searches, statistical analysis, and manuscript preparation. The study was supervised by XC, YW, and RG.

FUNDING

Support for this work includes funding from the National Natural Science Foundation of China (81873649); Doctoral scientific research Initiate funding project of Shunde Women and Children's Hospital of Guangdong Medical University (Maternity and Child Healthcare Hospital of Shunde Foshan) (2020BSQD007); Guangdong Medical University Research Foundation (GDMUM2020008 and GDMUM2020012); Medical Research Project of Foshan Health Bureau (20210188 and 20210289).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.774489/full#supplementary-material

REFERENCES

Banerjee, M., and Vats, P. (2014). Reactive Metabolites and Antioxidant Gene Polymorphisms in Type 2 Diabetes Mellitus. *Indian J. Hum. Genet.* 20, 10–19. doi:10.4103/0971-6866.132747

Cai, X., Xia, L., Pan, Y., He, D., Zhu, H., Wei, T., et al. (2019). Differential Role of Insulin Resistance and β-cell Function in the Development of Prediabetes and Diabetes in Middle-Aged and Elderly Chinese Population. *Diabetol. Metab. Syndr.* 11, 24. doi:10.1186/s13098-019-0418-x

Chen, H. Y., Huang, W., Leung, V. H. K., Fung, S. L. M., Ma, S. L., Jiang, H., et al. (2013). Functional Interaction between SNPs and Microsatellite in the Transcriptional Regulation of Insulin-like Growth Factor 1. *Hum. Mutat.* 34, 1289–1297. doi:10.1002/humu.22363

Chistiakov, D. A., Nikitin, A. G., Smetanina, S. A., Bel'chikova, L. N., Suplotova, L. A., Shestakova, M. V., et al. (2012). The Rs11705701 G>A Polymorphism of IGF2BP2 Is Associated with IGF2BP2 mRNA and Protein Levels in the Visceral Adipose Tissue - A Link to Type 2 Diabetes Susceptibility. *Rev. Diabet Stud.* 9, 112–122. doi:10.1900/RDS.2012.9.112

Colao, A., Di Somma, C., Cascella, T., Pivonello, R., Vitale, G., Grasso, L. F. S., et al. (2008). Relationships between Serum IGF1 Levels, Blood Pressure, and Glucose Tolerance: an Observational, Exploratory Study in 404 Subjects. *Eur. J. Endocrinol.* 159, 389–397. doi:10.1530/EJE-08-0201

Dai, N., Zhao, L., Wrighting, D., Krämer, D., Majithia, A., Wang, Y., et al. (2015). IGF2BP2/IMP2-Deficient Mice Resist Obesity through Enhanced Translation of Ucp1 mRNA and Other mRNAs Encoding Mitochondrial Proteins. *Cel Metab.* 21, 609–621. doi:10.1016/j.cmet.2015.03.006

DeFronzo, R. A., and Tripathy, D. (2009). Skeletal Muscle Insulin Resistance Is the Primary Defect in Type 2 Diabetes. *Diabetes Care* 32 (Suppl. 2), S157–S163. doi:10.2337/dc09-S302

Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A. U., et al. (2010). New Genetic Loci Implicated in Fasting Glucose Homeostasis and Their Impact on Type 2 Diabetes Risk. *Nat. Genet.* 42, 105–116. doi:10.1038/ng.520

Fujita, H., Hara, K., Shojima, N., Horikoshi, M., Iwata, M., Hirota, Y., et al. (2012). Variations with Modest Effects Have an Important Role in the Genetic Background of Type 2 Diabetes and Diabetes-Related Traits. *J. Hum. Genet.* 57, 776–779. doi:10.1038/jhg.2012.110

Gouda, W., Mageed, L., AzmyOkasha, O. A., Okasha, A., Shaker, Y., and Ashour, E. (2019). Association of Genetic Variants in *IGF-1* Gene with Susceptibility to Gestational and Type 2 Diabetes Mellitus. *Meta Gene* 21, 100588. doi:10.1016/j.mgene.2019.100588

Gulixiati·Maimaitituersun (2020). "IGF-1Gene Polymorphism and Hyperuricemia in Xinjiang Uyghur Population and Type 2 Diabetes," in *A Dissertation Submitted to Xinjiang Medical University in Partial Fullfillment of the Requirements for the Degree of Master of Public Health*. Xinjiang, China: Xinjiang Medical University. (in Chinese).

Han, Z., Wang, T., Tian, R., Zhou, W., Wang, P., Ren, P., et al. (2019). BIN1 Rs744373 Variant Shows Different Association with Alzheimer's Disease in Caucasian and Asian Populations. *BMC bioinformatics* 20, 691. doi:10.1186/s12859-019-3264-9

He, D., Ma, L., Feng, R., Zhang, L., Jiang, Y., Zhang, Y., et al. (2015). Analyzing Large-Scale Samples Highlights Significant Association between Rs10411210 Polymorphism and Colorectal Cancer. *Biomed. Pharmacother.* 74, 164–168. doi:10.1016/j.biopha.2015.08.023

Hu, C., Zhang, R., Wang, C., Wang, J., Ma, X., Hou, X., et al. (2010). Variants from GIPR, TCF7L2, DGKB, MADD, CRY2, GLIS3, PROX1, SLC30A8 and IGF1 Are Associated with Glucose Metabolism in the Chinese. *PLoS One* 5, e15542. doi:10.1371/journal.pone.0015542

Johnson, A. M. F., and Olefsky, J. M. (2013). The Origins and Drivers of Insulin Resistance. *Cell* 152, 673–684. doi:10.1016/j.cell.2013.01.041

Langberg, K. A., Ma, L., Ma, L., Sharma, N. K., Hanis, C. L., Elbein, S. C., et al. (2012). Single Nucleotide Polymorphisms in JAZF1 and BCL11A Gene Are Nominally Associated with Type 2 Diabetes in African-American Families from the GENNID Study. *J. Hum. Genet.* 57, 57–61. doi:10.1038/jhg.2011.133

Li, L., Yang, Y., Yang, G., Lu, C., Yang, M., Liu, H., et al. (2011). The Role of JAZF1 on Lipid Metabolism and Related Genes *In Vitro*. *Metabolism* 60, 523–530. doi:10.1016/j.metabol.2010.04.021

Li, N., Wang, H., Yan, Z., Yao, X., Hong, J., and Zhou, L. (2012). Ethnic Disparities in the Clustering of Risk Factors for Cardiovascular Disease Among the Kazakh, Uygur, Mongolian and Han Populations of Xinjiang: a Cross-Sectional Study. *BMC Public Health* 12, 499. doi:10.1186/1471-2458-12-499

Li, W., Huang, Y., Liu, H., Zhang, S., Wang, L., Li, N., et al. (2019). Significance of SNPs from Previous Genome-wide Association Study in Prediction of Postpartum Diabetes Among Pregnant Women with Gestational Diabetes Mellitus. *Chin. J. Public Health* 35, 6, 2019. (in Chinese). doi:10.11847/zgggws1117960

Li, X., Yang, M., Wang, H., Jia, Y., Yan, P., Boden, G., et al. (2014). Overexpression of JAZF1 Protected ApoE-Deficient Mice from Atherosclerosis by Inhibiting Hepatic Cholesterol Synthesis via CREB-dependent Mechanisms. *Int. J. Cardiol.* 177, 100–110. doi:10.1016/j.ijcard.2014.09.007

Li, Y., He, S., Li, C., Shen, K., Yang, M., Tao, W., et al. (2021). Evidence of Association between Single-Nucleotide Polymorphisms in Lipid Metabolism-Related Genes and Type 2 Diabetes Mellitus in a Chinese Population. *Int. J. Med. Sci.* 18, 356–363. doi:10.7150/ijms.53004

Li, Y., Song, D., Jiang, Y., Wang, J., Feng, R., Zhang, L., et al. (2016). CR1 Rs3818361 Polymorphism Contributes to Alzheimer's Disease Susceptibility in Chinese Population. *Mol. Neurobiol.* 53, 4054–4059. doi:10.1007/s12035-015-9343-7

Liao, Z. Z., Wang, Y. D., Qi, X. Y., and Xiao, X. H. (2019). JAZF1, a Relevant Metabolic Regulator in Type 2 Diabetes. *Diabetes Metab. Res. Rev.* 35, e3148. doi:10.1002/dmrr.3148

Liu, C., Li, H., Qi, L., Loos, R. J. F., Qi, Q., Lu, L., et al. (2012). Variants in GLIS3 and CRY2 Are Associated with Type 2 Diabetes and Impaired Fasting Glucose in Chinese Hans. *PLoS One* 6, e21464. doi:10.1371/journal.pone.0021464

Liu, G., Xu, Y., Jiang, Y., Zhang, L., Feng, R., and Jiang, Q. (2017). PICALM Rs3851179 Variant Confers Susceptibility to Alzheimer's Disease in Chinese Population. *Mol. Neurobiol.* 54, 3131–3136. doi:10.1007/s12035-016-9886-2

Mamet, R., Jacobson, C. K., and Heaton, T. B. (2005). Ethnic Intermarriage in Beijing and Xinjiang, China, 1990. *J. Comp. Fam. Stud.* 36, 187–204. doi:10.3138/jcfs.36.2.187

Mannino, G. C., Greco, A., De Lorenzo, C., Andreozzi, F., Marini, M. A., Perticone, F., et al. (2013). A Fasting Insulin-Raising Allele at IGF1 Locus Is Associated with Circulating Levels of *IGF-1* and Insulin Sensitivity. *PLoS One* 8, e85483. doi:10.1371/journal.pone.0085483

Ming, G.-f., Xiao, D., Gong, W.-j., Liu, H.-x., Liu, J., Zhou, H.-h., et al. (2014). JAZF1 Can Regulate the Expression of Lipid Metabolic Genes and Inhibit Lipid Accumulation in Adipocytes. *Biochem. Biophysical Res. Commun.* 445, 673–680. doi:10.1016/j.bbrc.2014.02.088

Perry, R. J., Samuel, V. T., Petersen, K. F., and Shulman, G. I. (2014). The Role of Hepatic Lipids in Hepatic Insulin Resistance and Type 2 Diabetes. *Nature* 510, 84–91. doi:10.1038/nature13478

Qin, L., Zhao, J., Wu, Y., Zhao, Y., Chen, C., Xu, M., et al. (2010). Association between Insulin-like Growth Factor 1 Gene Rs35767 Polymorphisms and Cancer Risk. *Medicine (Baltimore)* 98, e18017. doi:10.1097/MD.0000000000018017

Regué, L., Minichiello, L., Avruch, J., and Dai, N. (2019). Liver-specific Deletion of IGF2 mRNA Binding protein-2/IMP2 Reduces Hepatic Fatty Acid Oxidation and Increases Hepatic Triglyceride Accumulation. *J. Biol. Chem.* 294, 11944–11951. doi:10.1074/jbc.RA119.008778

Ruchat, S.-M., Elks, C. E., Loos, R. J. F., Vohl, M.-C., Weisnagel, S. J., Rankinen, T., et al. (2009). Association between Insulin Secretion, Insulin Sensitivity and Type 2 Diabetes Susceptibility Variants Identified in Genome-wide Association Studies. *Acta Diabetol.* 46, 217–226. doi:10.1007/s00592-008-0080-5

Seppä, S., Voutilainen, R., and Tenhola, S. (2015). Markers of Insulin Sensitivity in 12-Year-Old Children Born from Preeclamptic Pregnancies. *J. Pediatr.* 167, 125–130. doi:10.1016/j.jpeds.2015.04.015

Shankar, A., and Li, J. (2013). Positive Association between High-Sensitivity C-Reactive Protein Level and Diabetes Mellitus Among US Non-hispanic Black Adults. *Exp. Clin. Endocrinol. Diabetes* 116, 455–460. doi:10.1055/s-2007-1004563

Shen, N., Chen, B., Jiang, Y., Feng, R., Liao, M., Zhang, L., et al. (2015). An Updated Analysis with 85,939 Samples Confirms the Association between CR1 Rs6656401 Polymorphism and Alzheimer's Disease. *Mol. Neurobiol.* 51, 1017–1023. doi:10.1007/s12035-014-8761-2

Song, M., Zhao, F., Ran, L., Dolikun, M., Wu, L., Ge, S., et al. (2015). The Uyghur Population and Genetic Susceptibility to Type 2 Diabetes: Potential Role for Variants inCDKAL1,JAZF1, andIGF1Genes. *OMICS: A J. Integr. Biol.* 19, 230–237. doi:10.1089/omi.2014.0162

Telgmann, R., Dördelmann, C., Brand, E., Nicaud, V., Hagedorn, C., Pavenstädt, H., et al. (2009). Molecular Genetic Analysis of a Human Insulin-like Growth Factor 1 Promoter P1 Variation. *FASEB j.* 23, 1303–1313. doi:10.1096/fj.08-116863

Thankamony, A., Capalbo, D., Marcovecchio, M. L., Sleigh, A., Jørgensen, S. W., Hill, N. R., et al. (2014). Low Circulating Levels of IGF-1 in Healthy Adults Are Associated with Reduced β-Cell Function, Increased Intramyocellular Lipid, and Enhanced Fat Utilization during Fasting. *J. Clin. Endocrinol. Metab.* 99, 2198–2207. doi:10.1210/jc.2013-4542

Wang, l. (2019). *Relation between SNPs in IGF1 Promoter Region and Colorectal Cancer Risk in Type 2 Diabetes Patients*. MS Dissertation. Tianjin: Tianjin Medical University. (in Chinese).

Wang, W., Wise, C., Baric, T., Black, M. L., and Bittles, A. H. (2003). The Origins and Genetic Structure of Three Co-resident Chinese Muslim Populations: the Salar, Bo'an and Dongxiang. *Hum. Genet.* 113, 244–252. doi:10.1007/s00439-003-0948-y

Wei, Q., Zhou, B., Yang, G., Hu, W., Zhang, L., Liu, R., et al. (2018). JAZF1 Ameliorates Age and Diet-Associated Hepatic Steatosis through SREBP-1c -dependent Mechanism. *Cell Death Dis* 9 (9), 859. doi:10.1038/s41419-018-0923-0

Yuan, L., Luo, X., Zeng, M., Zhang, Y., Yang, M., Zhang, L., et al. (2015). Transcription Factor TIP27 Regulates Glucose Homeostasis and Insulin Sensitivity in a PI3-kinase/Akt-dependent Manner in Mice. *Int. J. Obes.* 39, 949–958. doi:10.1038/ijo.2015.5

Zhai, F., He, Y., Wang, Z., and Hu, Y. (2007). [Status and Characteristic of Dietary Intake of 12 minority Nationalities in China]. *Wei Sheng Yan Jiu* 36, 539–541. doi:10.3969/j.issn.1000-8020.2007.05.004 (in Chinese).

Zhang, J., Chen, X., Zhang, L., and Peng, Y. (2017). *IGF1* Gene Polymorphisms Associated with Diabetic Retinopathy Risk in Chinese Han Population. *Oncotarget* 8, 88034–88042. doi:10.18632/oncotarget.21366

Zhang, S., Li, X., Ma, G., Jiang, Y., Liao, M., Feng, R., et al. (2016). CLU Rs9331888 Polymorphism Contributes to Alzheimer's Disease Susceptibility in Caucasian but Not East Asian Populations. *Mol. Neurobiol.* 53, 1446–1451. doi:10.1007/s00702-014-1260

Zhao, F., Mamatyusupu, D., Wang, Y., Fang, H., Wang, H., Gao, Q., et al. (2016). The Uyghur Population and Genetic Susceptibility to Type 2 Diabetes: Potential Role for Variants in CAPN 10 , APM 1 and FUT 6 Genes. *J. Cel. Mol. Med.* 20, 2138–2147. doi:10.1111/jcmm.12911

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 01 December 2021
doi: 10.3389/fgene.2021.756908

Genetic Liability to Insomnia and Lung Cancer Risk: A Mendelian Randomization Analysis

Jiayi Shen[1,2,3,4†], Huaqiang Zhou[1,2,3†], Jiaqing Liu[1,2,3†], Yaxiong Zhang[1,2,3], Ting Zhou[1,2,3], Gang Chen[1,2,3], Wenfeng Fang[1,2,3], Yunpeng Yang[1,2,3], Yan Huang[1,2,3] and Li Zhang[1,2,3]**

[1] Department of Medical Oncology, Sun Yat-sen University Cancer Center, Guangzhou, China, [2] State Key Laboratory of Oncology in South China, Guangzhou, China, [3] Collaborative Innovation Center for Cancer Medicine, Guangzhou, China, [4] Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, China

Lung cancer is the second most frequently diagnosed cancer and the leading cause of cancer death worldwide, making its prevention an urgent issue. Meanwhile, the estimated prevalence of insomnia was as high as 30% globally. Research on the causal effect of insomnia on lung cancer incidence is still lacking. In this study, we aimed to assess the causality between the genetic liability to insomnia and lung cancer. We performed a two-sample Mendelian randomization analysis (inverse variance weighted) to determine the causality between the genetic liability to insomnia and lung cancer. Subgroup analysis was conducted, which included lung adenocarcinoma and lung squamous cell carcinoma. In the sensitivity analysis, we conducted heterogeneity test, MR Egger, single SNP analysis, leave-one-out analysis, and MR PRESSO. There were causalities between the genetic susceptibility to insomnia and increased incidence of lung cancer [odds ratio (95% confidence interval), 1.35 (1.14–1.59); *P*, < 0.001], lung adenocarcinoma [odds ratio (95% confidence interval), 1.35 (1.07–1.70); *P*, 0.01], and lung squamous cell carcinoma [odds ratio (95% confidence interval), 1.35 (1.06–1.72), *P*, 0.02]. No violation of Mendelian randomization assumptions was observed in the sensitivity analysis. There was a causal relationship between the genetic susceptibility to insomnia and the lung cancer, which was also observed in lung adenocarcinoma and lung squamous cell carcinoma. The underlying mechanism remains unknown. Effective intervention and management for insomnia were recommended to improve the sleep quality and to prevent lung cancer. Moreover, regular screening for lung cancer may be beneficial for patients with insomnia.

Keywords: insomnia, lung cancer, mendelian randomization analysis, causality, prevention

OPEN ACCESS

Edited by:
Guiyou Liu,
Tianjin Institute of Industrial Biotechnology (CAS), China

Reviewed by:
Hongmei Jiang,
Northwestern University, United States
Meng Zhu,
Nanjing Medical University, China

****Correspondence:***
Yan Huang
huangyan@sysucc.org.cn
Li Zhang
zhangli6@mail.sysu.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *11 August 2021*
Accepted: *11 November 2021*
Published: *01 December 2021*

Citation:
Shen J, Zhou H, Liu J, Zhang Y, Zhou T, Chen G, Fang W, Yang Y, Huang Y and Zhang L (2021) Genetic Liability to Insomnia and Lung Cancer Risk: A Mendelian Randomization Analysis. Front. Genet. 12:756908. doi: 10.3389/fgene.2021.756908

Abbreviations: BMI, body mass index; CI, confidence interval; GWAS, Genome-wide Association Study; IARC, International Agency for Research on Cancer; ILCCO, International Lung Cancer Consortium; IVW, inverse variance weighted; LUAD, lung adenocarcinoma; LUSQ, lung squamous cell carcinoma; MR, Mendelian randomization; OR, odds ratio; RCT, randomized controlled trial; SNPs, single-nucleotide polymorphisms; UKB, United Kingdom Biobank.

INTRODUCTION

Lung cancer is the second most frequently diagnosed cancer for both male and female in the world, of which the estimated number of new cases was 228,150 in 2019 (Siegel et al., 2019). It is also the leading cause of cancer death worldwide, with the estimated number of new deaths as 142,670 in 2019. Regarding its high incidence and mortality, lung cancer has long been a heavy burden in public health, making lung cancer prevention an urgent issue. For this reason, it is meaningful to investigate whether there are causalities between potential risk factors and lung cancer, to provide guidance in lung cancer prevention.

Insomnia has become a common sleep disorder worldwide, with the estimated prevalence as 30% (Roth, 2007). Previous studies mainly revealed the association between poor sleep habits like prolonged or shortened sleep duration and cancer incidence (Kakizaki et al., 2008; Chen et al., 2019). The incidence of lung cancer increased when sleep duration was ≤6.5 h or ≥8 h(Luojus et al., 2014). However, insomnia disorder is not simply characterized by reduced sleep duration but more by difficulties falling asleep and sleep disturbance (Morin et al., 2015). Research focusing on the causal effect of insomnia on lung cancer incidence is still lacking. We think it necessary to analyze the causality between insomnia and lung cancer, considering the urgency of lung cancer prevention, high prevalence and the potential carcinogenicity of insomnia.

Mendelian randomization (MR) analysis is a novel epidemiological approach for the estimation of causality between exposure and outcome (Smith and Ebrahim, 2003). In MR analysis, single-nucleotide polymorphisms (SNPs), which have been identified to be robustly correlated with the exposure, are used as proxies of exposure. SNPs of exposure should be correlated with the risk of outcome to the extent predicted by their influence on exposure, if the causality between exposure and outcome exists (Smith et al., 2008). MR analysis can be a potential mimic of randomized controlled trial (RCT) by utilizing SNPs, especially when RCT is too costly or infeasible (Smith and Ebrahim, 2004). SNPs, instrument variables in MR analysis, are randomly allocated during gamete formation and fertilization in the population, which is similar to the randomization in RCT. In this way, biases from confounders and inverse causality can also be avoided in MR analysis, which are common in observational studies (Davey Smith and Hemani, 2014).

In this study, we aimed to assess the causality between genetic liability to insomnia and lung cancer, utilizing two-sample MR analysis. We present the following article in accordance with the STROBE reporting checklist.

MATERIAL AND METHODS

Summary Data From Genome-wide Association Study on Insomnia and Lung Cancer

In a meta-analysis by Jansen et al., 248 SNPs were identified to be robustly correlated with insomnia (Jansen et al., 2019). Data from United Kingdom Biobank (UKB) version 2 (n = 386,533) and 23andMe (n = 944,477) were pooled. Sample size in total was 1,331 010. (**Table 1**). Information about insomnia was collected utilizing a self-report sleep questionnaire. Prevalence of insomnia in the combined sample was 29.9%. The questionnaires used by UKB and 23andMe were with high accuracy (sensitivity/specificity of UKB = 98/96%; sensitivity/specificity of 23andMe = 84/80%). The 248 SNPs were genome-wide significant ($P < 5 \times 10^{-8}$). These 248 SNPs were in linkage equilibrium with each other at $r^2 < 0.1$, and they could explain 2.6% of the variance in insomnia. Conclusively, the 248 SNPs can serve as the genetic instrumental variables for insomnia with enough statistical power. These 248 SNPs were utilized as the proxies of insomnia in this MR analysis.

We used summary data from a Genome-wide Association Study (GWAS) by International Lung Cancer Consortium (ILCCO) on lung cancer (11,348 cases and 15,861 controls), lung adenocarcinoma (LUAD) (3,442 cases and 14,894 controls), and lung squamous cell carcinoma (LUSQ) (3,275 cases and 15,038 controls). (**Table 1**) (Wang et al., 2014). The effects of the SNPs of insomnia on lung cancer, LUAD and LUSQ, the effect size and standard error, were extracted from the GWAS by ILCCO in the form of summary data through MR-base (Hemani et al., 2018). SNPs Summary data of insomnia and the outcomes were harmonized, where effect of each SNP on insomnia and outcomes were estimated and 12 SNPs were removed for being palindromic with intermediate allele frequencies (rs11126082, rs12454003, rs12991815, rs1731951, rs2030672, rs2221119, rs4858708, rs6545798, rs7044885, rs8180817, rs9373590, rs9540729). (**Supplementary Table S1**) Effect allele and frequency of effect allele were also provided.

To assess the potential existence of weak instrumental bias in this MR study, we also calculated the statistical power and F statistic of this study, with four presumed and fixed range of odds ratio (OR) (Brion et al., 2013) (**Table 2**). Power and F statistic were dependent on the strength of association between the SNPs and insomnia and the sample size of the outcome GWAS studies. The larger the power and F statistics were, the smaller the possibility of weak instrumental bias would be (cut off value for judgement, power, 80%; F statistic, 10) Powers were larger than 80%, only when OR was set to be “0.75 or 1.33” and “0.67 or 1.50” for lung cancer and when ORs were set to be “0.67 or 1.50” for LUAD and LUSQ. F statistics were far greater than 10 for lung cancer, LUAD and LUSQ.

No patients were involved in the study design. Recruitment or conduct and the need for ethical approval was waived.

Two-Sample MR Analysis

Two-sample MR was utilized to investigate the causality between the genetic liability to insomnia and lung cancer incidence. Two-sample MR can improve statistical power, with the utilization of summary data of SNPs from large scale GWAS (Burgess et al., 2015; Lawlor, 2016). In two-sample MR, the effects of SNPs of exposure on exposure and outcome are derived from GWAS of exposure and GWAS of outcome respectively. Specifically, inverse variance weighted (IVW) was used (Hemani et al., 2018). Subgroup analyses of LUAD and LUSQ were conducted to assess whether there was a difference between the MR estimate

TABLE 1 | Genome-wide Association Study Utilized in this MR Analysis.

Consortium	Phenotype	Participants	Web source
CTGlab of CNCR	Insomnia	133,1010	https://ctg.cncr.nl/software/summary_statistics
ILCCO	Lung cancer, LUAD, LUSQ	27,209	ilcco.iarc.fr

CTGlab, complex trait genetics lab; CNCR, center for neurogenomics and cognitive research; ILCCO, international lung cancer consortium; LUAD, lung adenocarcinoma; LUSQ, lung squamous cell carcinoma.

TABLE 2 | Power and F statistic for Conventional Mendelian Randomization Analysis (two-sided α = 0.05).

Outcome	Sample size of GWAS on outcome	Proportion of cases	Power to identify OR of 0.91 or 1.10	Power to identify OR of 0.83 or 1.20	Power to identify OR of 0.75 or 1.33	Power to identify OR of 0.67 or 1.50	F Statistic
Lung cancer	27,209	0.4171	0.24	0.68	0.96	1.00	727.32
LUAD	18,336	0.1877	0.13	0.38	0.61	0.98	490.46
LUSQ	18,313	0.1788	0.13	0.37	0.59	0.97	489.85

GWAS, genome-wide association study; OR, odds ratio; LUAD, lung adenocarcinoma; LUSQ, lung squamous cell carcinoma.

of lung cancer and those of LUAD and LUSQ, considering the reported difference in the etiologies between LUAD and LUSQ (Herbst et al., 2018).

Genetic instrumental variable used in MR analysis must fulfill three assumptions: 1) the instrumental variable is associated with the exposure; 2) the instrumental variable is associated with the outcome through the studied exposure merely; and 3) the instrumental variable is independent of other factors which affect the outcome (Boef et al., 2015). In terms of sensitivity analysis, we produced a MR regression slopes chart to display the difference between the result of IVW and those of MR Egger and weighted median. Additionally, the heterogeneity test was also conducted by performing Cochran's Q test on the IVW and the MR-Egger estimate. If there is heterogeneity (Cochrane's Q *p*-value < 0.05) and a random effect model was employed to it. To assess whether the assumptions of MR were violated, MR-Egger analysis was performed to detect directional horizontal pleiotropy, and a funnel plot was also generated (Bowden et al., 2015; Burgess and Thompson, 2017). The existence of pleiotropy means the instrument variable can be associated with the observed outcome through other mechanisms than insomnia. Furthermore, the detection of directional horizontal pleiotropy suggests that the sum of pleiotropy does not equal to zero, which means the violation of the 2) assumption (exclusion restriction assumption). If the intercept is close to 0 and *P* is close to one in MR-Egger analysis, the MR study will be free of directional horizontal pleiotropy. Single SNP analysis and leave-one-out analysis was performed to assess whether the result was driven by a single SNP. MR PRESSO was also conducted for the estimation of horizontal pleiotropy, which included global test, outlier test, and distortion test (Verbanck et al., 2018). If a horizontal pleiotropy was detected by the global test, the outlier test would be performed, figuring out the outlying SNPs. Subsequently, an outlier-corrected causal estimate would be assessed and compared with the original MR estimate, which was the distortion test, providing a *p*-value for the comparison.

TABLE 3 | Mendelian randomization estimates of the causality between insomnia and lung cancer.

Exposure	Outcome	Inverse variance weighted	
		Or (95%CI)	*p*-value
Insomnia	Lung Cancer	1.35 (1.14–1.59)	<0.001
Insomnia	LUAD	1.35 (1.07–1.70)	0.01
Insomnia	LUSQ	1.35 (1.06–1.72)	0.02

OR, odds ratio; CI, confidence interval; LUAD, lung adenocarcinoma; LUSQ, lung squamous cell carcinoma.

The statistical analysis was performed utilizing the package TwoSampleMR (version 0.4.25) in R (version 3.6.1). The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

RESULTS

Causality From Insomnia to Lung Cancer

The genetic susceptibility to insomnia was causally associated with increased lung cancer incidence based on the results of IVW method {OR [95% confidence interval (CI)], 1.35 (1.14–1.59); *p* < 0.001; Cochrane's Q *p*-value = 0.00078}. (**Table 3**). MR regression slopes showed positive correlation between the effect of SNP on insomnia and that on lung cancer. (**Supplementary Figure S1**) The three slopes according to the three different MR analyses were close to each other. Single SNP analysis indicated that the MR estimate of single SNP varied from each other. (**Supplementary Table S2**; **Supplementary Figure S2**) However, results in the leave-one-out analysis were similar to each other, indicating that there was no driving SNP in this MR analysis (**Supplementary Table S3**; **Supplementary Figure S3**) According to the result of MR-Egger analysis, directional horizontal pleiotropy was not detected, which meant the SNPs

TABLE 4 | Results of MR Egger for the estimation of directional horizontal pleiotropy.

Outcome	Intercept	Standard error	*p*-value
Lung cancer	−0.001	0.006	0.91
Lung adenocarcinoma	0.004	0.009	0.64
Lung squamous cell carcinoma	0.005	0.009	0.62

of insomnia did not affect the incidence of lung cancer through other traits than insomnia. (**Table 4**). The funnel plot was symmetrical, with the indication of no directional horizontal pleiotropy (**Supplementary Figure S4**) Horizontal pleiotropy and outlying SNPs were identified in the MR PRESSO global test, while the distortion test did not show a statistically significant difference between the original MR estimate and the outlier-corrected MR estimate. (**Table 5**).

Causal Effect From Insomnia to LUAD and LUSQ

According to the result of IVW, insomnia was positively correlated with the incidence of LUAD [OR (95% CI), 1.35 (1.07–1.70); *P*, 0.01; Cochrane's Q *p*-value = 0.056] and LUSQ [OR (95% CI), 1.35 (1.06–1.72); *P*, 0.02; Cochrane's Q *p*-value = 0.0050]. (**Table 3**) The MR regression slopes of both LUAD and LUSQ showed positive associations between the SNP effect on insomnia and the SNP effect on outcomes (**Supplementary Figures S5**; **Supplementary Figure S6**) However, the three regression curves did not overlap with each other, as were displayed in the two **Supplementary Figures**. Like the results of lung cancer, single SNP analysis of LUAD and LUSQ indicated varied MR effect size of single SNP (**Supplementary Table S2**; **Supplementary Figure S7**; **Supplementary Figure S8**), while leave-one-out analysis did not show signs of driving SNPs (**Supplementary Table S3**; **Supplementary Figure S9**; **Supplementary Figure S10**) MR Egger detected no directional horizontal pleiotropy for LUAD and LUSQ. (**Table 4**). Funnel plots of LUAD and LUSQ were symmetric, in support of the intercept and *p*-value mentioned (**Supplementary Figure S11**; **Supplementary Figure S12**) Horizontal pleiotropy was indicated in the MR analysis of LUAD and LUSQ, according to the result of the global test in MR PRESSO. (**Table 5**). Rs76145129 was identified as the outlying SNP in LUSQ. However, the removal of this SNP brought no significant difference in the MR estimate.

DISCUSSION

In this study, we estimated the causality between insomnia and lung cancer. The genetic liability to insomnia was causally correlated with lung cancer incidence. The positive associations were also observed in LUAD and LUSQ.

Insomnia usually leads to circadian disruption which has been classified as probably carcinogenic to humans (Group 2A) by the IARC (Humans and International Agency for Research on, 2010). Moreover, circadian disruption can alter the secretion patterns of melatonin (Kim et al., 2015). Melatonin was reported to have multiple anti-tumor effect, by modulating cell cycle, stimulating cell differentiation, inducing apoptosis, inhibiting metastasis and angiogenesis, and activating immune system, which was also found among lung cancer patients (Mediavilla et al., 2010) (Du-Quiton et al., 2010; Bhattacharya et al., 2019; Gurunathan et al., 2021). However, we should note that the disruption of the circadian rhythm of melatonin secretion was observed mainly in chronic primary insomnia patients (Hajak et al., 1995). Additionally, lower sleep quality was associated with decreased level of Klotho, an aging-suppressing protein, which also inhibits lung cancer cell growth and promotes lung cancer cell apoptosis (Chen et al., 2010; Chen et al., 2012; Mochón-Benguigui et al., 2020).

Some intermediate phenotypes can mediate the association between insomnia and lung cancer. A bidirectional causal relationship has been found between insomnia and smoking (Gibson et al., 2019). Specifically, smoking initiation and cigarettes smoked per day were positively correlated with insomnia. And insomnia was also found to be a promoter of smoking heaviness and an obstructor of smoking cessation. Considering the carcinogenicity of smoking in lung cancer, tobacco consumption may be an important mediator between insomnia and lung cancer (Hecht, 2002). Insomnia patients whose sleep duration was short or long were at higher risk of obesity and central obesity (Cai et al., 2018). In another MR analysis by Gao et al., body mass index (BMI) was identified as a risk factor of lung cancer (Gao et al., 2016). The above two previous studies indicated the potential intermediate effect of obesity in the positive relationship between insomnia and lung cancer, as was found in this study. The potential mediation of tobacco consumption and high BMI still needs verification in further research. Yet, the currently uncertain mediation status should not

TABLE 5 | Results of MR PRESSO for the estimation of horizontal pleiotropy.

Outcome	*p*-value of global test	Outlying SNP	Outlier-corrected causal estimate	*p*-value of outlier-corrected causal estimate	*p*-value of distortion test
Lung cancer	<0.001	rs73079014, rs76145129	0.269	<0.001	0.61
LUAD	0.03	No significant outliers	NA	NA	NA
LUSQ	0.01	rs76145129	0.284	0.02	0.83

LUAD, lung adenocarcinoma; LUSQ, lung squamous cell carcinoma.

detract the importance of management for insomnia as one way to decrease the lung cancer risk.

This is the first large-scale MR study to illustrate the causal relationship between the genetic liability to insomnia and lung cancer. First, this study demonstrated its great clinical significance. It is an important mission to identify and intervene modifiable risk factors of lung cancer, and finally reduce the incidence of lung cancer. With the utilization of MR analysis, the genetic susceptibility to insomnia was found to be a risk factor of lung cancer. We advocate medical intervention on insomnia, along with the advocacy of other healthy lifestyles, like cigarette cessation (Siegel et al., 2019). Second, several issues may confuse the result of observational study on the association between insomnia and lung cancer. In this MR analysis, we used genetic liability to insomnia as a proxy of exposure. The effect of confounders and inverse causality, which are common in observational study, were avoided in MR analysis (Fewell et al., 2007; Davey Smith and Hemani, 2014). With the utilization of SNPs as proxies of exposure, confounders can be avoided, as SNPs are randomly allocated in the population during gamete formation, serving as a mimic of the randomization in RCT. To our concern, the problem of inverse causality should be paid extra attention to in the study of the relationship between insomnia and lung cancer, because insomnia is a common mental disorder in lung cancer patients (Savard and Morin, 2001). Third, RCT has been regarded as a steady approach for the estimation of causality. However, in the assessment of the causality between insomnia and lung cancer, RCT is infeasible and unethical. MR analysis was used in this study instead. Last but not least, no violation of the assumptions of MR analysis was observed in this study. The sample size was also large enough to support our findings.

However, there are still some limitations in this study. First, the two GWASs utilized were based on the United Kingdom population. The application of our conclusion in other populations may cause some unknown biases. Second, with the application of GWAS summary data, we couldn't make stratification of the sample, because we didn't have access to the individual characteristics of the studied population, like age, smoking status and so on. Third, the genetic liability to insomnia was used in this study as a proxy of insomnia, but it did not mean that every individual with those SNPs would necessarily suffer from insomnia. While insomnia is a disorder closely correlated with physical illness, behavioral factors, environment and medications (Kamel and Gammack, 2006). Researchers should be cautious when interpreting our result. Fourth, bidirectional MR analysis between insomnia and lung cancer and multivariable MR study were also infeasible, because of the limited data accessed. However, the result of this study was still rational because directional horizontal pleiotropy was not found in MR Egger and the distortion test did not deny our MR estimates even though the global test identified horizontal pleiotropy. MR analysis is an effective method in terms of causality estimation, a mimic of RCT (Smith and Ebrahim, 2003, 2004). Finally, the direct underlying mechanism is still unknown and needs further exploration or verification.

CONCLUSION

In conclusion, the genetic susceptibility to insomnia was causally correlated with higher incidence of lung cancer, along with its histological subtype, LUAD and LUSQ. Effective intervention and management for insomnia were recommended to improve the sleep quality itself and to prevent lung cancer. Moreover, regular screening for lung cancer may be beneficial for patients with insomnia. However, further prospective studies are warranted to confirm the results and clarify the underlying mechanism.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

Conception and design: YH, LZ, JS, and HZ Administrative support: YH and LZ Provision of study materials or patients: JS, HZ, and JL Collection and assembly of data: JS, HZ, and JL Data analysis and interpretation: JS, HZ, JL, and YZ Manuscript writing: All authors Final approval of manuscript: All authors.

FUNDING

This work was supported by the National Key R and D Program of China (grant numbers 2016YFC0905500, 2016YFC0905503) and Science and Technology Program of Guangzhou (grant number 201704020072).

ACKNOWLEDGMENTS

The authors acknowledge the efforts of the Complex Trait Genetics lab (CTGlab) of Center for Neurogenomics and Cognitive Research (CNCR), International Lung Cancer Consortium (ILCCO), Tobacco and Genetics consortium (TAG), Global Lipids Genetics Consortium (GLGC), and Genetic Investigation of ANthropometric Traits consortium (GIANT) in providing high quality GWAS data for researchers.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.756908/full#supplementary-material

REFERENCES

Bhattacharya, S., Patel, K. K., Dehari, D., Agrawal, A. K., and Singh, S. (2019). Melatonin and its Ubiquitous Anticancer Effects. *Mol. Cel Biochem* 462 (1-2), 133–155. doi:10.1007/s11010-019-03617-5

Boef, A. G. C., Dekkers, O. M., and le Cessie, S. (2015). Mendelian Randomization Studies: a Review of the Approaches Used and the Quality of Reporting. *Int. J. Epidemiol.* 44 (2), 496–511. doi:10.1093/ije/dyv071

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian Randomization with Invalid Instruments: Effect Estimation and Bias Detection through Egger Regression. *Int. J. Epidemiol.* 44 (2), 512–525. doi:10.1093/ije/dyv080

Brion, M.-J. A., Shakhbazov, K., and Visscher, P. M. (2013). Calculating Statistical Power in Mendelian Randomization Studies. *Int. J. Epidemiol.* 42 (5), 1497–1501. doi:10.1093/ije/dyt179

Burgess, S., Scott, R. A., Scott, R. A., Timpson, N. J., Davey Smith, G., and Thompson, S. G. (2015). Using Published Data in Mendelian Randomization: a Blueprint for Efficient Identification of Causal Risk Factors. *Eur. J. Epidemiol.* 30 (7), 543–552. doi:10.1007/s10654-015-0011-z

Burgess, S., and Thompson, S. G. (2017). Interpreting Findings from Mendelian Randomization Using the MR-Egger Method. *Eur. J. Epidemiol.* 32 (5), 377–389. doi:10.1007/s10654-017-0255-x

Cai, G.-H., Theorell-Haglöw, J., Janson, C., Svartengren, M., Elmståhl, S., Lind, L., et al. (2018). Insomnia Symptoms and Sleep Duration and Their Combined Effects in Relation to Associations with Obesity and central Obesity. *Sleep Med.* 46, 81–87. doi:10.1016/j.sleep.2018.03.009

Chen, B., Ma, X., Liu, S., Zhao, W., and Wu, J. (2012). Inhibition of Lung Cancer Cells Growth, Motility and Induction of Apoptosis by Klotho, a Novel Secreted Wnt Antagonist, in a Dose-dependent Manner. *Cancer Biol. Ther.* 13 (12), 1221–1228. doi:10.4161/cbt.21420

Chen, B., Wang, X., Zhao, W., and Wu, J. (2010). Klotho Inhibits Growth and Promotes Apoptosis in Human Lung Cancer Cell Line A549. *J. Exp. Clin. Cancer Res.* 29, 99. doi:10.1186/1756-9966-29-99

Chen, P., Wang, C., Song, Q., Chen, T., Jiang, J., Zhang, X., et al. (2019). Impacts of Sleep Duration and Snoring on the Risk of Esophageal Squamous Cell Carcinoma. *J. Cancer* 10 (9), 1968–1974. doi:10.7150/jca.30172

Davey Smith, G., and Ebrahim, S. (2003). 'Mendelian Randomization': Can Genetic Epidemiology Contribute to Understanding Environmental Determinants of Disease? *Int. J. Epidemiol.* 32 (1), 1–22. doi:10.1093/ije/dyg070

Davey Smith, G., and Hemani, G. (2014). Mendelian Randomization: Genetic Anchors for Causal Inference in Epidemiological Studies. *Hum. Mol. Genet.* 23 (R1), R89–R98. doi:10.1093/hmg/ddu328

Du-Quiton, J., Wood, P. A., Burch, J. B., Grutsch, J. F., Gupta, D., Tyer, K., et al. (2010). Actigraphic Assessment of Daily Sleep-Activity Pattern Abnormalities Reflects Self-Assessed Depression and Anxiety in Outpatients with Advanced Non-small Cell Lung Cancer. *Psycho-oncology* 19 (2), 180–189. doi:10.1002/pon.1539

Fewell, Z., Davey Smith, G., and Sterne, J. A. C. (2007). The Impact of Residual and Unmeasured Confounding in Epidemiologic Studies: a Simulation Study. *Am. J. Epidemiol.* 166 (6), 646–655. doi:10.1093/aje/kwm165

Gao, C., Patel, C. J., Michailidou, K., Peters, U., Gong, J., Schildkraut, J., et al. (2016). Mendelian Randomization Study of Adiposity-Related Traits and Risk of Breast, Ovarian, Prostate, Lung and Colorectal Cancer. *Int. J. Epidemiol.* 45 (3), 896–908. doi:10.1093/ije/dyw129

Gibson, M., Munafò, M. R., Taylor, A. E., and Treur, J. L. (2019). Evidence for Genetic Correlations and Bidirectional, Causal Effects between Smoking and Sleep Behaviors. *Nicotine Tob. Res.* 21 (6), 731–738. doi:10.1093/ntr/nty230

Gurunathan, S., Qasim, M., Kang, M.-H., and Kim, J.-H. (2021). Role and Therapeutic Potential of Melatonin in Various Type of Cancers. *Onco Targets Ther.* 14, 2019–2052. doi:10.2147/OTT.S298512

Hajak, G., Rodenbeck, A., Staedt, J., Bandelow, B., Huether, G., and Rüther, E. (1995). Nocturnal Plasma Melatonin Levels in Patients Suffering from Chronic Primary Insomnia. *J. Pineal Res.* 19 (3), 116–122. doi:10.1111/j.1600-079x.1995.tb00179.x

Hecht, S. S. (2002). Cigarette Smoking and Lung Cancer: Chemical Mechanisms and Approaches to Prevention. *Lancet Oncol.* 3 (8), 461–469. doi:10.1016/s1470-2045(02)00815-x

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., et al. (2018). The MR-Base Platform Supports Systematic Causal Inference across the Human Phenome. *Elife* 7, e34408. doi:10.7554/eLife.34408

Herbst, R. S., Morgensztern, D., and Boshoff, C. (2018). The Biology and Management of Non-small Cell Lung Cancer. *Nature* 553 (7689), 446–454. doi:10.1038/nature25183

Humans, I. W. G. o. t. E. o. C. R. t., and International Agency for Research on, C. (2010). *Painting, Firefighting, and Shiftwork*. Geneva, Switzerland: World Health Organization.

Jansen, P. R., Watanabe, K., Watanabe, K., Stringer, S., Skene, N., Bryois, J., et al. (2019). Genome-wide Analysis of Insomnia in 1,331,010 Individuals Identifies New Risk Loci and Functional Pathways. *Nat. Genet.* 51 (3), 394–403. doi:10.1038/s41588-018-0333-3

Kakizaki, M., Kuriyama, S., Sone, T., Ohmori-Matsuda, K., Hozawa, A., Nakaya, N., et al. (2008). Sleep Duration and the Risk of Breast Cancer: the Ohsaki Cohort Study. *Br. J. Cancer* 99 (9), 1502–1505. doi:10.1038/sj.bjc.6604684

Kamel, N. S., and Gammack, J. K. (2006). Insomnia in the Elderly: Cause, Approach, and Treatment. *Am. J. Med.* 119 (6), 463–469. doi:10.1016/j.amjmed.2005.10.051

Kim, T. W., Jeong, J.-H., and Hong, S.-C. (2015). The Impact of Sleep and Circadian Disturbance on Hormones and Metabolism. *Int. J. Endocrinol.* 2015, 1–9. doi:10.1155/2015/591729

Lawlor, D. A. (2016). Commentary: Two-Sample Mendelian Randomization: Opportunities and Challenges. *Int. J. Epidemiol.* 45 (3), 908–915. doi:10.1093/ije/dyw127

Luojus, M. K., Lehto, S. M., Tolmunen, T., Erkkilä, A. T., and Kauhanen, J. (2014). Sleep Duration and Incidence of Lung Cancer in Ageing Men. *BMC Public Health* 14, 295. doi:10.1186/1471-2458-14-295

Mediavilla, M. D., Sanchez-Barcelo, E. J., Tan, D. X., Manchester, L., and Reiter, R. J. (2010). Basic Mechanisms Involved in the Anti-cancer Effects of Melatonin. *Curr. Med. Chem.* 17 (36), 4462–4481. doi:10.2174/092986710794183015

Mochón-Benguigui, S., Carneiro-Barrera, A., Castillo, M. J., and Amaro-Gahete, F. J. (2020). Is Sleep Associated with the S-Klotho Anti-aging Protein in Sedentary Middle-Aged Adults? the FIT-AGEING Study. *Antioxidants* 9 (8), 738. doi:10.3390/antiox9080738

Morin, C. M., Drake, C. L., Harvey, A. G., Krystal, A. D., Manber, R., Riemann, D., et al. (2015). Insomnia Disorder. *Nat. Rev. Dis. Primers* 1, 15026. doi:10.1038/nrdp.2015.26

Roth, T. (2007). Insomnia: Definition, Prevalence, Etiology, and Consequences. *J. Clin. Sleep Med.* 3 (5 Suppl. l), S7–S10. doi:10.5664/jcsm.26929

Savard, J., and Morin, C. M. (2001). Insomnia in the Context of Cancer: A Review of a Neglected Problem. *J. Clin. Oncol.* 19 (3), 895–908. doi:10.1200/jco.2001.19.3.895

Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer Statistics, 2019. *CA A. Cancer J. Clin.* 69 (1), 7–34. doi:10.3322/caac.21551

Smith, G. D., and Ebrahim, S. (2004). Mendelian Randomization: Prospects, Potentials, and Limitations. *Int. J. Epidemiol.* 33 (1), 30–42. doi:10.1093/ije/dyh132

Smith, G. D., Timpson, N., and Ebrahim, S. (2008). Strengthening Causal Inference in Cardiovascular Epidemiology through Mendelian Randomization. *Ann. Med.* 40 (7), 524–541. doi:10.1080/07853890802010709

Verbanck, M., Chen, C.-Y., Neale, B., and Do, R. (2018). Detection of Widespread Horizontal Pleiotropy in Causal Relationships Inferred from Mendelian Randomization between Complex Traits and Diseases. *Nat. Genet.* 50 (5), 693–698. doi:10.1038/s41588-018-0099-7

Wang, Y., McKay, J. D., Rafnar, T., Wang, Z., Timofeeva, M. N., Broderick, P., et al. (2014). Rare Variants of Large Effect in BRCA2 and CHEK2 Affect Risk of Lung Cancer. *Nat. Genet.* 46 (7), 736–741. doi:10.1038/ng.3002

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Genetics

ORIGINAL RESEARCH
published: 20 December 2021
doi: 10.3389/fgene.2021.758639

Investigating the Causal Relationship Between Physical Activity and Chronic Back Pain: A Bidirectional Two-Sample Mendelian Randomization Study

Shaowei Gao[1], *Huaqiang Zhou*[2], *Siyu Luo*[1], *Xiaoying Cai*[1], *Fang Ye*[1], *Qiulan He*[1], *Chanyan Huang*[1], *Xiaoyang Zheng*[1], *Ying Li*[1], *Zhanxin Du*[1], *Yaqing Wang*[1], *Zhihui Qi*[1] and *Zhongxing Wang*[1]*

[1]*Department of Anesthesia, Sun Yat-sen University First Affiliated Hospital, Guangzhou, China,* [2]*Department of Medical Oncology, Sun Yat-sen University Cancer Center, Guangzhou, China*

Background: Recent observational studies have reported a negative association between physical activity and chronic back pain (CBP), but the causality of the association remains unknown. We introduce bidirectional Mendelian randomization (MR) to assess potential causal inference between physical activity and CBP.

Materials and Methods: This two-sample MR used independent genetic variants associated with physical activity and CBP as genetic instruments from large genome-wide association studies (GWASs). The effects of both directions (physical activity to CBP and CBP to physical activity) were examined. Inverse variance-weighted meta-analysis and alternate methods (weighted median and MR-Egger) were used to combine the MR estimates of the genetic instruments. Multiple sensitivity analyses were conducted to examine the robustness of the results.

Results: The MR set parallel GWAS cohorts, among which, those involved in the primary analysis were comprised of 337,234 participants for physical activity and 158,025 participants (29,531 cases) for CBP. No evidence of a causal relationship was found in the direction of physical activity to CBP [odds ratio (OR), 0.98; 95% CI, 0.85–1.13; p = 0.81]. In contrast, a negative causal relationship in the direction of CBP to physical activity was detected (β = –0.07; 95% CI, –0.12 to –0.01; p = 0.02), implying a reduction in moderate-vigorous physical activity (approximately 146 MET-minutes/week) for participants with CBP relative to controls.

Conclusion: The negative relationship between physical activity and CBP is probably derived from the reduced physical activity of patients experiencing CBP rather than the protective effect of physical activity on CBP.

Keywords: mendelian randomization, physical activity, chronic back pain, causal inference, instrumental variable

OPEN ACCESS

Edited by:
Guiyou Liu,
Tianjin Institute of Industrial Biotechnology (CAS), China

Reviewed by:
Huaizhen Qin,
University of Florida, United States
Nicholas B. Larson,
Mayo Clinic, United States

Correspondence:
Zhongxing Wang
wzhxing@mail.sysu.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *14 August 2021*
Accepted: *25 November 2021*
Published: *20 December 2021*

Citation:
Gao S, Zhou H, Luo S, Cai X, Ye F, He Q, Huang C, Zheng X, Li Y, Du Z, Wang Y, Qi Z and Wang Z (2021) Investigating the Causal Relationship Between Physical Activity and Chronic Back Pain: A Bidirectional Two-Sample Mendelian Randomization Study. Front. Genet. 12:758639. doi: 10.3389/fgene.2021.758639

Abbreviations: CBP, chronic back pain; GWAS, genome-wide association studies; IVW, inverse variance-weighted; MR, Mendelian randomization; MR-PRESSO, MR-pleiotropy residual sum and outlier; SNP, single-nucleotide polymorphism.

INTRODUCTION

Back pain, especially low back pain, has become a large burden worldwide, as it is estimated to affect more than 510 million people and cause over 57 million "years lived with disability" in 2016 (Disease, 2018; Wu et al., 2020). At least one-third of patients with back pain report persistent pain after an acute episode and eventually develop chronic back pain (CBP) (Qaseem et al., 2017), which is generally defined as back pain lasting ≥3 months (Deyo et al., 2014). A key step in preventing CBP is the identification of possible risk factors, especially intervenable risk factors. To date, well-known risk factors for CBP have included smoking (Shiri et al., 2010), obesity (Zhang et al., 2018), previous episodes of back pain (Taylor et al., 2014), other chronic conditions (e.g., diabetes, headache) (Ferreira et al., 2013), and poor mental health (Hartvigsen et al., 2018; Power et al., 2001). However, the role of physical activity on CBP is inconclusive (**Table 1**).

Physical activity is defined as musculoskeletal movement that results in energy consumption (Caspersen et al., 1985). As shown in **Table 1**, recent meta-analyses reviewed tens of observational studies and found a negative relationship between physical activity and CBP (Alzahrani et al., 2019a; Shiri and Falah-Hassani, 2017). A similar conclusion was also reported by other cross-sectional studies (Alzahrani et al., 2019b; B. Amorim et al., 2019). However, studies with high-level evidence (such as randomized control studies), which can address the problem of causal inference, are lacking. Consequently, whether the negative relationship between physical activity and CBP is due to the protective effect of physical activity on CBP or the tendency of patients with CBP to reduce physical activity remains unknown.

Randomized control studies on physical activity are difficult to conduct, as it is unethical to constrain participants' physical activity. Mendelian randomization (MR) is an alternative method to achieve randomization for this situation by treating genetic variation as a natural experiment in which individuals are randomly assigned to different levels of nongenetic exposure during their lifetime (Davey Smith and Ebrahim, 2003). In addition, MR can strengthen causal inferences by importing a bidirectional design.

In this study, we first applied bidirectional MR to determine the causal association between physical activity and CBP (**Figure 1**). We aim to clarify the causal relationship behind this observed negative association between physical activity and CBP. We hypothesize that CBP resulted in reduced physical activity whereas physical activity per se did not have protective effect on CBP.

MATERIALS AND METHODS

This is a Mendelian randomization study with a bidirectional and two-sample design, as illustrated in **Figure 1**. All the data

TABLE 1 | Representative studies for the association between physical activity and chronic back pain.

Study	Type	Design	Region	Time	Sample size	Results	Note	References number
Alzahrani (2019a)	Meta-analysis	Observational studies (cohort or cross-sectional)	Nonspecific	Earliest-March 2017	35 studies, 106,776 participants	Medium physical activity was significantly associated with a lower prevalence of low back pain	This meta-analysis did not specify acute or chronic low back pain	11
Alzahrani (2019b)	Clinical study	Cross-sectional study	Participants form the United Kingdom	1994–2008	60,134 participants	Total PA volume was inversely associated with the prevalence of chronic back conditions	The outcome was chronic back conditions, among which low back pain is one of the most common	13
Shiri (2017)	Meta-analysis	Observational studies (prospective, cohort)	Nonspecific	Earliest-July 2017	36 studies, 158,475 participants	Leisure time physical activity may reduce the risk of chronic low back pain by 11–16%	The exposure was leisure time physical activity	12
Heneweer (2009)	Clinical study	Cross-sectional study	Dutch	1998	3,364 participants	There is some evidence that the relation between physical activity and chronic low back pain is U-shaped	Type of activity (daily routine, leisure time and sport activity), intensity of and time spent on these activities, and back exertion during sport activities were taken into account	18
Kamada (2014)	Clinical study	Cross-sectional study	Japan	2009	4,559 participants	There were no significant linear or quadratic relationships between self-reported physical activity and chronic low back pain	The population were aged 40–79 years	16

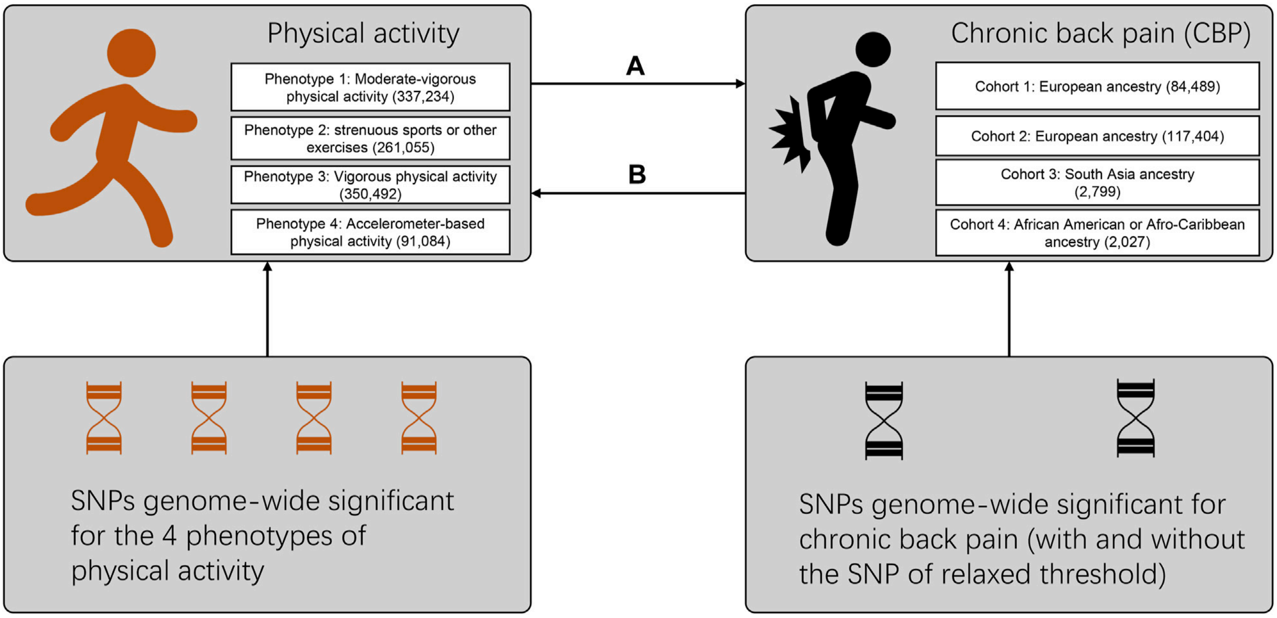

Physical activity
Phenotype 1: Moderate-vigorous physical activity (337,234)
Phenotype 2: strenuous sports or other exercises (261,055)
Phenotype 3: Vigorous physical activity (350,492)
Phenotype 4: Accelerometer-based physical activity (91,084)
A
B
Chronic back pain (CBP)
Cohort 1: European ancestry (84,489)
Cohort 2: European ancestry (117,404)
Cohort 3: South Asia ancestry (2,799)
Cohort 4: African American or Afro-Caribbean ancestry (2,027)
SNPs genome-wide significant for the 4 phenotypes of physical activity
SNPs genome-wide significant for chronic back pain (with and without the SNP of relaxed threshold)

FIGURE 1 | Flow diagram for the design of the bidirectional, two-sample Mendelian randomization study. Multiple phenotypes and cohorts were cross-validated to maintain the robustness of our results. The direction marked **(A)** refers to the effect of physical activity on chronic back pain, while that marked **(B)** refers to the reverse effect. Details on the SNPs used as trait instruments are summarized in **Supplementary Tables S2–S4**. The numbers of participants for different phenotypes or cohorts are labeled in the brackets. SNP: single-nucleotide polymorphism.

used are summary-level and derived from public genome-wide association studies (GWAS), which had obtained ethical permissions from their respective institutional review boards and written informed consent from their respective participants. Neither patients nor the public were involved in this MR study. The study was conducted under Burgess's guidelines and reported according to the STROBE-MR statement (Supplementary checklist 2) (Burgess et al., 2019; Davey Smith et al., 2019). We analyzed these data from April 20, 2021 to June 20, 2021.

Selection of Instruments and Outcome Data

Physical Activity

The physical activity instruments were based on Klimentidis's GWAS conducted with participants of the United Kingdom Biobank cohort (19). This GWAS, using a population of predominantly European ancestry, examined the following four physical activity phenotypes: (Disease, 2018) self-reported moderate-vigorous physical activity [continuous phenotype, 337,234 participants, in standardized units of inverse normalized metabolic equivalent minutes per week (MET-minutes/week)]; (Wu et al., 2020) self-reported vigorous physical activity (binary phenotype, 261,055 participants with 98,060 cases, ≥ 3 vs. 0 for days per week), (Qaseem et al., 2017) self-reported strenuous sports or other exercises (binary phenotype, 350,492 participants with 124,842 cases, ≥ 2–3 vs. 0 for days per week), and (Deyo et al., 2014) seven-day average acceleration from a wrist-worn accelerometer (continuous phenotype, 91,084 participants, in milligravities). The characteristics for each phenotype are summarized in **Supplementary Table S1**. We chose SNPs from the first phenotype (self-reported moderate-vigorous physical activity) for the primary analysis, as this phenotype yielded the largest number of significant SNPs. To ensure robustness, the SNPs from the other three phenotypes were used in a sensitivity analysis (**Supplementary Table S2**). In addition, as the GWAS of the accelerometer-based activity identified only two SNPs but had higher heritability than that of the self-reported activity (~14 vs. ~5%), the top SNPs meeting a relaxed threshold ($p < 1 \times 10^{-7}$) were also imported to our study (**Supplementary Table S3**) in a sensitivity analysis; the method of using SNPs with relaxed thresholds has been used for other MR studies when insufficient SNPs are available (Gage et al., 2017; Hartwig et al., 2017; Choi et al., 2019). We retained only the top independent SNPs by selecting one representative SNP among highly correlated SNPs ($r^2 > 0.001$), a process known as "clumping". If an instrument SNP was not present in the outcome GWAS, then a proxy SNP that was in linkage disequilibrium with the instrument SNPs was searched for instead. Clumping and proxy SNPs are both based on reference data from the 1,000 Genomes Project (Genomes Project et al., 2015).

For the other direction, in which physical activity is regarded as the outcome trait, we again applied Klimentidis's GWAS (Klimentidis et al., 2018). The completed summary data can be accessed from the OpenGWAS database through the MR-base platform (Elsworth et al., 2020; Hemani et al., 2018a). Similarly, data for all four phenotypes above are available, while moderate-vigorous physical activity was used for the primary analysis.

Chronic Back Pain

Genetic instruments for CBP were derived from a genome-wide meta-analysis comprising adults of European ancestry from 16 cohorts (26), in which positive cases were obtained by examining the questionnaires from the participants. These cohorts did not have a consistent definition of CBP: two cohorts used "≥ 1 month of back pain in consecutive years"; nine cohorts used "≥ 6 months of back pain"; six cohorts used "≥ 3 months of back pain". The control group enrolled participants who reported not having back pain or reported back pain of insufficient duration as cases. Most of the included cohorts did not include question items regarding localization of the pain to the low back or lumbar region specifically. Therefore, a general definition examining chronic "back pain" rather than a more specific chronic "low back pain" definition was applied. This meta-analysis identified four SNPs associated with chronic back pain, one of which met a relaxed threshold ($p = 3.9 \times 10^{-7}$), while the others met strict criteria ($p <$ 5 × 10–8) (**Supplementary Table S4**). Similarly, we introduced a sensitivity analysis by eliminating the SNP with a relaxed threshold.

For the outcome data, we searched the OpenGWAS database and found four GWAS cohorts with completed summary data (**Supplementary Table S5**). Two out of the four cohorts are of European ancestry, while the other two contain South Asian populations and African American or Afro-Caribbean populations. Because the MR results may be uninformative for the magnitude (rather than the direction) of the effect when the exposure and outcome studies are derived from different populations (Hemani et al., 2018a), we selected one European ancestry cohort with the maximum sample size (117,404 participants and 80,588 cases) for the primary analysis and the other three for the sensitivity analyses.

Statistical Analysis

The R package "TwoSampleMR" developed by researchers in the MR-base platform was used for this Mendelian randomization study (Hemani et al., 2018a). Briefly, the algorithm in this package combines the effect sizes of the instruments on exposure traits with those of the instruments on outcome traits using the principle of meta-analysis. In addition to the effect size, the effect allele and its frequency for each instrument—whether for exposure or outcome—must be extracted to determine the direction of the strand.

As the primary method for combining MR estimates, we used the multiplicative random-effect IVW method, which translates to a weight regression of instrument-outcome effects on instrument-exposure effects where the intercept is restricted to zero (Burgess et al., 2013). In this way, bias may occur if horizontal pleiotropy (in which the instruments influence the outcome through causal pathways other than the exposure) is present. We therefore introduced two other MR methods: the weighted median method and MR-Egger regression. The weighted median method chooses the median MR estimate of the instruments as the result, while MR-Egger regression allows the intercept to be a value other than zero (Bowden et al., 2015; Bowden et al., 2016). Both methods are more robust for horizontal pleiotropy, although at the cost of reduced statistical power (Hemani et al., 2018b). Generally, the effect size for the binary outcome should be represented as odds ratio (OR) (i.e., exponentiated β). However, in Klimentidis's GWAS, a mixed model-model linear regression was used even for binary phenotypes (vigorous PA and strenuous sports or other exercises), leading to unreliable estimates of effect sizes (but not influencing the direction and statistical power) (Klimentidis et al., 2018). We therefore reported the effect estimates in the β value for PA as an outcome trait (we avoided translating the meaning of β for the binary phenotypes) and in the OR for CBP as an outcome trait.

TABLE 2 | MR results for the effect of self-reported moderate-vigorous physical activity on chronic back pain (CBP).

Method	OR (95% CI)[b]	*p* Value	No. of SNPs
With outlier[a]			
IVW	0.98 (0.85–1.13)	0.81	9
Weighted median	0.96 (0.84–1.11)	0.59	9
MR-Egger	0.91 (0.48–1.73)	0.77	9
Without outlier[a]			
IVW	0.94 (0.84–1.05)	0.26	8
Weighted median	0.96 (0.84–1.09)	0.51	8
MR-Egger	1.00 (0.61–1.63)	1.00	8

[a]*The outlier was rs1043595, which was detected with the MR-pleiotropy residual sum and outlier method.*

[b]*Indicates odds for CBP per 1-SD increase in moderate-vigorous physical activity (1-SD equals 2084 MET-minutes/week in Klimentidis's GWAS).*

Abbreviations: IVW: inverse variance weighted; CBP, chronic back pain; MR, Mendelian randomization; OR, odds ratio; CI, confidence interval; SNP, single-nucleotide polymorphism.

A series of methods were applied for the sensitivity analyses: in addition to setting multiple comparisons among different phenotypes and different cohorts, the funnel plot, Cochran's Q statistic, leave-one-out analyses, MR-PRESSO, and the MR-Egger intercept test of deviation from the null were used to detect heterogeneity and horizontal pleiotropy (Burgess and Thompson, 2017). By implementing a homonymous R package, MR-PRESSO also detects and corrects outlier SNPs reflecting pleiotropic biases (Verbanck et al., 2018). Finally, to determine potential pleiotropy, we searched each instrument used for the primary analysis in the PhenoScanner GWAS database (version 2; http://phenoscanner.medschl.cam.ac.uk) to find any existing associations with potential confounding traits; then, we removed these SNPs to control the pleiotropic effects and to see if the primary results could be reversed.

RESULTS

The cohorts used for extracting instruments in the primary analysis were comprised of 337,234 participants for physical activity and 158,025 participants (29,531 cases) for CBP. Details for all parallel cohorts were summarized on the section of **Section 2** and **Supplementary Tables S1, S5**.

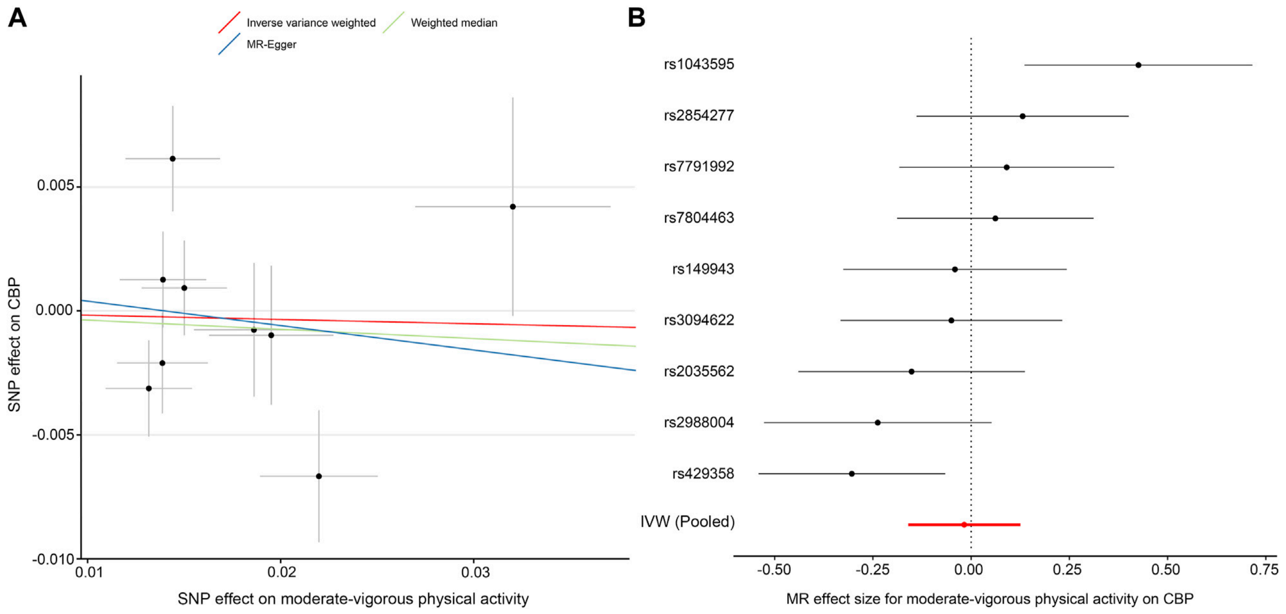


FIGURE 2 | MR plots for the effect of moderate-vigorous physical activity on chronic back pain (CBP). **(A)** Scatter plot of the SNP effect on moderate-vigorous physical activity vs. that on CBP. The slope of each fitted line represents the pooled MR effect calculated by each method. **(B)** Forest plot of individual and pooled MR effect sizes for moderate-vigorous physical activity on CBP. Each point and its corresponding line represent the β value with its 95% CI, respectively. Abbreviations: SNP, single-nucleotide polymorphism; CBP, chronic back pain; MR, Mendelian randomization; IVW, inverse variance weighted.

TABLE 3 | MR results for the effect of chronic back pain (CBP) on self-reported moderate-vigorous physical activity.

Method	β (95% CI)[a]	*p* Value	No. of SNPs[b]
IVW	−0.07 (−0.12 to −0.01)	0.02	4
Weighted median	−0.07 (−0.13 to −0.01)	0.03	4
MR-Egger	−0.08 (−0.25 to 0.09)	0.47	4

[a] *Indicates a change in multiple of SD of moderate-vigorous physical activity (1-SD equals 2084 MET-minutes/week in Klimentidis's GWAS) for participants with CBP vs control status.*

[b] *No outlier was detected with MR-pleiotropy residual sum and outlier method*
Abbreviations: IVW, inverse variance weighted; CBP, chronic back pain; MR, Mendelian randomization; SNP single-nucleotide polymorphism.

Effect of Physical Activity on CBP

In this direction, we found no evidence of a discernible causal effect of physical activity on CBP. In our primary analysis—the effect of self-reported moderate-vigorous physical activity on the largest CBP cohort with European ancestry—the combined inverse variance-weighted (IVW) OR was close to 1 (IVW OR, 0.98; 95% CI, 0.85–1.13; $p = 0.81$) (**Table 2**; **Figure 2**), which indicated that there is no effect of physical activity on CBP. The results were almost consistent for different exposure phenotypes and different outcome cohorts (**Supplementary Table S6**). The funnel plot did not detect obvious asymmetry, and the leave-one-out analysis did not change the pattern of the result (**Supplementary Figure S1**). The MR-Egger intercept test suggested no directional horizontal pleiotropy (intercept, 0.001; standard error, 0.005; $p = 0.81$), even though Cochran's Q test indicated moderate heterogeneity ($Q = 19.8$; $p = 0.011$). The method of MR-pleiotropy residual sum and outlier (MR-PRESSO) detected one outlier (rs1043595), but the result remained negative when this outlier was removed (**Table 2**).

Effect of CBP on Physical Activity

In contrast to the previous analysis, we found a robust negative causal relationship between CBP and physical activity. In our primary analysis—the effect of CBP represented by all four single-nucleotide polymorphisms (SNPs) on self-reported moderate-vigorous physical activity—the MR estimate with the IVW method was significantly less than zero (IVW β, −0.07; 95% CI, −0.12 to −0.01; $p = 0.02$) (**Table 3**; **Figure 3**), implying that participants with CBP tended to reduce their physical activity by approximately 146 MET-minutes/week with respect to those without CBP. The weighted median and MR-Egger tests yielded similar patterns of effects (**Table 3**). The results were consistent not only with analyses with different outcome traits, such as self-reported strenuous sports and accelerometer-based physical activity, but also with analyses where the SNP with the relaxed threshold was removed for CBP (**Supplementary Table S7**). The leave-one-out analysis showed that no single SNP was strong for reversely driving the overall effect of CBP on physical activity but detected one SNP (rs12310519) that played a relatively predominant role (**Supplementary Figure S2A**). Furthermore, the funnel plot presents with a symmetric pattern (**Supplementary Figure S2B**), and Cochran's Q test suggested no heterogeneity ($Q = 0.3$; $p = 0.96$). In addition, MR-PRESSO found no outliers, and the MR-Egger intercept test indicated no consistent pleiotropy (intercept, 0.001; standard error, 0.004; $p = 0.91$).

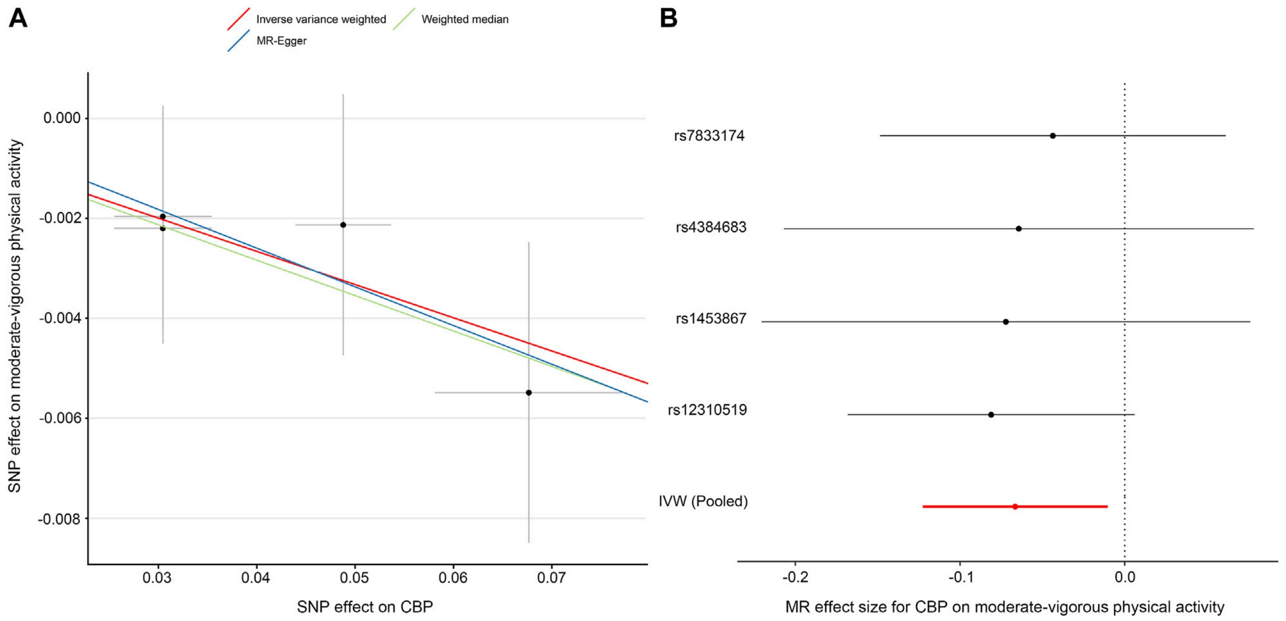


FIGURE 3 | MR plots for the effect of chronic back pain (CBP) on moderate-vigorous physical activity. **(A)** Scatter plot of the SNP effect on CBP vs. that on moderate-vigorous physical activity. The slope of each fitted line represents the pooled MR effect calculated by each method. **(B)** Forest plot of individual and pooled MR effect sizes for CBP on moderate-vigorous physical activity. Each point and its corresponding line represent the β value and its 95% CI, respectively. Abbreviations: SNP, single-nucleotide polymorphism; CBP, chronic back pain; MR, Mendelian randomization; IVW, inverse variance weighted.

Potential Pleiotropy Searched in PhenoScanner

In total, thirteen SNPs were included in our primary analyses (9 for physical activity to CBP, four for CBP to physical activity). We searched PhenoScanner database for these SNPs and found that the most potential pleiotropy was "trunk fat/fat-free mass", which was involved in 7/13 of all SNPs (**Supplementary Table S8**). After removing these involved SNPs, the pattern of the primary results did not change (physical activity to CBP: IVW OR, 1.09; 95% CI 0.85–1.40; $p = 0.52$; CBP to physical activity: IVW β, −0.077; 95% CI, −0.15 to −0.003; $p = 0.04$) (**Supplementary Figure S3**).

DISCUSSION

To the best of our knowledge, this is the first MR study to explore the causal relationship between physical activity and CBP. We examined the effects in both directions and found that engaging in more physical activity was not associated with a reduced risk of CBP, but having CBP was associated with reduced physical activity (including both self-reported and accelerometer-based physical activity). The result supports the more intuitive view that the negative association between physical activity and CBP arises from the fact that patients with CBP tend to reduce their physical activity.

Heritability and Genetics of Selected Variables

The heritability of physical activity varies in terms of different measurements: objective measurement (i.e., accelerometry-based method) has higher heritability than self-reported one (14 vs 5%) (Klimentidis et al., 2018). The study (Klimentidis et al., 2018) we used for extracting instrument SNPs of physical activity applied multi-variable models to adjust covariates such as age, sex, genotyping chip, BMI. This dataset has been involved in several powerful MR studies (Choi et al., 2019; Choi et al., 2020; Papadimitriou et al., 2020), most of which selectively analyzed a few of measurements. To make our results robust, we used all measurements for sensitivity analysis and obtained consistent results, which deeply strengthen our conclusions.

In contrast, the heritability of back pain ranges from 0 to 67%, and is always higher for chronic than acute conditions (Ferreira et al., 2013). The mechanisms of CBP are not only due to anatomic disorders, such as intervertebral disc degeneration, but also to psychological factors. Some previous studies have discovered possible susceptibility genes involved in CBP including SPOCK2, DCC, SLC10A7. (Suri et al., 2018; Freidin et al., 2021). SPOCK2 encodes a protein binding to glycosaminoglycans to form part of the extracellular matrix (Ren et al., 2020), while DCC encodes a transmembrane receptor for netrin-1, an axonal guidance molecule involved in the development of commissural neurons (Finci et al., 2015). SLC10A7, Solute Carrier Family 10 Member 7, is involved in teeth and skeletal development. The evidences above imply that CBP is a complex syndrome, and to some extent related to genetics. The study from which we extracted instrument SNPs of CBP is a meta-analysis including 15 different cohorts, each adjusted for covariates like age, sex, study-specific covariates, and population substructure (Suri et al., 2018). The nature of meta-analysis made the instrument SNPs more robust.

Comparisons With Previous Traditional Studies

Previous studies reported conflicting results regarding the effect of physical activity on CBP. Some studies showed no association between physical activity and CBP (Kamada et al., 2014; Picavet and Schuit, 2003) or a U-shaped relationship, in which very low and very high levels of physical activity increased the risk of CBP (Heneweer et al., 2009). However, a recent observational study with a large population and two meta-analyses supported a negative relationship between physical activity and CBP (Shiri and Falah-Hassani, 2017; Alzahrani et al., 2019a; Alzahrani et al., 2019b). The observational study involved a population 60,134 adults, but its cross-sectional design was insufficient for identifying the causal inference between physical activity and CBP (Alzahrani et al., 2019b). Although the meta-analysis recruited prospective studies (Shiri and Falah-Hassani, 2017), the observational design was "apt in generating hypotheses and suggesting causality but can never prove it" (De Rango, 2016). In contrast, MR can mimic the design of randomized controlled trials (Hemani et al., 2018a). Given that a SNP is known to be related to a trait (the so-called "instrument variable"), according to Mendel's law, the alleles at the SNP are causally upstream of the corresponding trait and expected to be random with respect to potential confounders. In an MR study, participants are randomly assigned to the treatment group or control group according to the genotype at the instrument SNP of exposure. Then, the effect size of the causal inference can be calculated as the ratio between the SNP effect on the outcome and the SNP effect on the exposure. Our study extends the current literature from the level of association to the level of causal inference.

Robustness

Our results were robust to different pairs of exposure and outcome cohorts (**Supplementary Tables S6, S7**). In the direction of physical activity to CBP, engaging in more physical activity did not significantly change the risk of CBP except in the "ukb-e-3571_AFR" cohort (**Supplementary Table S6**). The small sample size (approximately 2000 participants) of the "ukb-e-3571_AFR" cohort and the wide range of the OR indicate that the exception probably derives from a random error. In addition, the generalization of our results to different races (e.g., Chinese and African) is limited due to the fact that the exposure and outcome datasets were mostly from European population. Future studies on this issue will require analyses of other races.

In the other direction, from CBP to physical activity, reporting CBP was always associated with reporting reduced physical activity (**Supplementary Table S7**). However, in the leave-one-out analysis, we found one predominant SNP, rs12310519, without which the OR of reporting CBP on reporting moderate-vigorous physical activity was no longer statistically significant (the 95% CI for the OR included 1) (**Supplementary Figure S2A**). To examine the extent of the influence of this SNP, we repeated the leave-one-out analysis on the other three phenotypes of physical activity; interestingly, however, this SNP (rs12310519) was not the predominant SNP for self-reported vigorous physical activity and self-reported strenuous sports or other exercises (**Supplementary Figure S4**). This result may imply different mechanisms by which genetic variance influences different levels of one phenotype.

After looking up the SNPs used for the primary analysis in Phenoscanner database, a potential pleiotropy, "trunk fat/fat-free mass", was detected (**Supplementary Table S8**). This trait has been reported as a common predictive factor for both physical activity and CBP (Ness et al., 2007; Urquhart et al., 2011; Brady et al., 2019) and served as an exposure-outcome confounder for the current study. Nevertheless, the pattern of the primary results did not change after controlling this pleiotropy, possibly due to the balance of the multiple SNPs that have effects of different directions on this confounder.

Limitations

This study has several limitations. First, although different levels of physical activity were included in this study, the CBP was an all-or-none variable. Thus, it was impossible to compare the effect between different levels of CBP. It will be interesting to determine in future studies if the effects of physical activity are similar on different levels of CBP. Second, there were overlapping samples in both the exposure and outcome studies because the physical activity source study and the CBP outcome data both involved participants from the United Kingdom Biobank project. Results from MRs with overlapping samples may be biased due to the winner's curse phenomenon (Bowden and Dudbridge, 2009). However, we used a sensitivity analysis in which weaker instruments were excluded, which can minimize the bias from sample overlap (Pierce and Burgess, 2013). Finally, the CBP phenotype we used represents a symptom rather than a disease or a biomarker. Compared with other more detailed phenotypes, such as osteoarthritis, additional mechanisms may be involved in CBP, such as muscle injury, nerve root compression, or intervertebral disc degeneration. Thus, a single genome-wide association study is insufficient for finding all SNPs as instruments for CBP. Although the genome-wide meta-analysis we selected for this MR included 16 CBP cohorts, it detected only three to four SNPs, which might partially cover all the mechanisms.

Another point we should clarify is that we used chronic back pain instead of chronic low back pain, a more commonly used phenotype, as the exposure phenotype. The primary reason for this is that the questionnaires used for the included cohorts did not specifically isolate the low back region (Suri et al., 2018)). Given the high agreement between general back pain and low back pain-specific questions (Denard et al., 2010) and since upper/mid back pain without concurrent low back pain is uncommon (Hartvigsen et al., 2009), we believe that our results with CBP can well represent those with chronic low back pain, as exemplified in other studies using similar substitutions (Suri et al., 2018; Suri et al., 2017).

Importance

Despite these limitations, the MR study performed here provides a novel insight into genetic variants as instruments for assessing the causal inference between physical activity and CBP and obviates typical challenges in observational research while providing an internal explanation for such studies (Koes et al., 2010; Shiri and

Falah-Hassani, 2017; Alzahrani et al., 2019a; Alzahrani et al., 2019b). If the negative relationship between physical activity and CBP is truly a reverse causality, the concept that patients with CBP should be engaging in activity, which is recommended by current guidelines (Koes et al., 2010), may need to be reconsidered.

CONCLUSION

This study applied MR to examine the causal inference between physical activity and CBP. The negative relationship between these two traits is probably derived from the fact that patients experiencing CBP tend to reduce their physical activities.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://gwas.mrcieu.ac.uk/.

AUTHOR CONTRIBUTIONS

ZW takes responsibility for the content of the manuscript. SG and ZW conceived the idea and designed the study. SL and XC searched and reviewed relevant articles. FY, QH, and CH helped collect and assemble data from relevant studies. SG and HZ did the major work of data extraction and analysis. Data interpretation, writing and edition were provided by all of the authors. The authors meet the criteria for authorship as recommended by the International Committee of Medical Journal Editors (ICMJE). All of the author approved the final manuscript.

FUNDING

The work was supported by the Key Training Program for Young Teachers of Sun Yat-sen University (No. 19ykzd12) and Guangzhou Science and Technology Foundation (No. 201904010418).

ACKNOWLEDGMENTS

We appreciate the researchers in the MR-base platform, as well as those who have being collecting and preserving massive GWAS data. We also thank the American Journal Experts (AJE) corporation for their professional native English-Language editing. The article has been shared in a preprint (Gao et al., 2021).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.758639/full#supplementary-material

REFERENCES

Alzahrani, H., Mackey, M., Stamatakis, E., Zadro, J. R., and Shirley, D. (2019). The Association between Physical Activity and Low Back Pain: a Systematic Review and Meta-Analysis of Observational Studies. *Sci. Rep.* 9 (1), 8244. doi:10.1038/s41598-019-44664-8

Alzahrani, H., Shirley, D., Cheng, S. W. M., Mackey, M., and Stamatakis, E. (2019). Physical Activity and Chronic Back Conditions: A Population-Based Pooled Study of 60,134 Adults. *J. Sport Health Sci.* 8 (4), 386–393. doi:10.1016/j.jshs.2019.01.003

B. Amorim, A., Simic, M., Pappas, E., Zadro, J. R., Carrillo, E., Ordoñana, J. R., et al. (2019). Is Occupational or Leisure Physical Activity Associated with Low Back Pain? Insights from a Cross-Sectional Study of 1059 Participants. *Braz. J. Phys. Ther.* 23 (3), 257–265. doi:10.1016/j.bjpt.2018.06.004

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian Randomization with Invalid Instruments: Effect Estimation and Bias Detection through Egger Regression. *Int. J. Epidemiol.* 44 (2), 512–525. doi:10.1093/ije/dyv080

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* 40 (4), 304–314. doi:10.1002/gepi.21965

Bowden, J., and Dudbridge, F. (2009). Unbiased Estimation of Odds Ratios: Combining Genomewide Association Scans with Replication Studies. *Genet. Epidemiol.* 33 (5), 406–418. doi:10.1002/gepi.20394

Brady, S. R. E., Urquhart, D. M., Hussain, S. M., Teichtahl, A., Wang, Y., Wluka, A. E., et al. (2019). High Baseline Fat Mass, but Not Lean Tissue Mass, Is Associated with High Intensity Low Back Pain and Disability in Community-Based Adults. *Arthritis Res. Ther.* 21 (1), 165. doi:10.1186/s13075-019-1953-4

Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian Randomization Analysis with Multiple Genetic Variants Using Summarized Data. *Genet. Epidemiol.* 37 (7), 658–665. doi:10.1002/gepi.21758

Burgess, S., Davey Smith, G., Davies, N. M., Dudbridge, F., Gill, D., Glymour, M. M., et al. (2019). Guidelines for Performing Mendelian Randomization Investigations. *Wellcome Open Res.* 4, 186. doi:10.12688/wellcomeopenres.15555.2

Burgess, S., and Thompson, S. G. (2017). Interpreting Findings from Mendelian Randomization Using the MR-Egger Method. *Eur. J. Epidemiol.* 32 (5), 377–389. doi:10.1007/s10654-017-0255-x

Caspersen, C. J., Powell, K. E., and Christenson, G. M. (1985). Physical Activity, Exercise, and Physical Fitness: Definitions and Distinctions for Health-Related Research. *Public Health Rep.* 100 (2), 126–131.

Choi, K. W., Chen, C.-Y., Stein, M. B., Klimentidis, Y. C., Wang, M.-J., Koenen, K. C., et al. (2019). Assessment of Bidirectional Relationships between Physical Activity and Depression Among Adults. *JAMA Psychiatry* 76 (4), 399–408. doi:10.1001/jamapsychiatry.2018.4175

Choi, K. W., Stein, M. B., Nishimi, K. M., Ge, T., Coleman, J. R. I., Chen, C.-Y., et al. (2020). An Exposure-wide and Mendelian Randomization Approach to Identifying Modifiable Factors for the Prevention of Depression. *Ajp* 177 (10), 944–954. doi:10.1176/appi.ajp.2020.19111158

Davey Smith, G., Davies, N. M., Dimou, N., Egger, M., Gallo, V., Golub, R., et al. (2019). *Guidelines for Strengthening the Reporting of Mendelian Randomization Studies*. Peer. doi:10.7287/peerj.preprints.27857v1

Davey Smith, G., and Ebrahim, S. (2003). 'Mendelian Randomization': Can Genetic Epidemiology Contribute to Understanding Environmental Determinants of Disease? *Int. J. Epidemiol.* 32 (1), 1–22. doi:10.1093/ije/dyg070

De Rango, P. (2016). Prospective Cohort Studies. *Eur. J. Vasc. Endovascular Surg.* 51 (1), 151. doi:10.1016/j.ejvs.2015.09.021

Denard, P. J., Holton, K. F., Miller, J., Fink, H. A., Kado, D. M., Marshall, L. M., et al. (2010). Back Pain, Neurogenic Symptoms, and Physical Function in Relation to Spondylolisthesis Among Elderly Men. *Spine J.* 10 (10), 865–873. doi:10.1016/j.spinee.2010.07.004

Deyo, R. A., Dworkin, S. F., Amtmann, D., Andersson, G., Borenstein, D., Carragee, E., et al. (2014). Focus Article: Report of the NIH Task Force on Research

Standards for Chronic Low Back Pain. *Eur. Spine J.* 23 (10), 2028–2045. doi:10.1007/s00586-014-3540-3

Disease, G. B. D., Global, Regional, and National Incidence, Prevalence, and Years Lived with Disability for 354 Diseases and Injuries for 195 Countries and Territories, 1990-2017: a Systematic Analysis for the Global Burden of Disease Study 2017. *Lancet* (2018) 392(10159):1789–1858. doi:10.1016/S0140-6736(18)32279-7

Elsworth, B., Lyon, M., Alexander, T., Liu, Y., Matthews, P., Hallett, J., et al. (2020). *The MRC IEU OpenGWAS Data Infrastructure*. bioRxiv. doi:10.1101/2020.08.10.244293

Ferreira, P. H., Beckenkamp, P., Maher, C. G., Hopper, J. L., and Ferreira, M. L. (2013). Nature or Nurture in Low Back Pain? Results of a Systematic Review of Studies Based on Twin Samples. *Ejp* 17 (7), 957–971. doi:10.1002/j.1532-2149.2012.00277.x

Finci, L., Zhang, Y., Meijers, R., and Wang, J.-H. (2015). Signaling Mechanism of the Netrin-1 Receptor DCC in Axon Guidance. *Prog. Biophys. Mol. Biol.* 118 (3), 153–160. doi:10.1016/j.pbiomolbio.2015.04.001

Freidin, M. B., Tsepilov, Y. A., Stanaway, I. B., Meng, W., Hayward, C., Smith, B. H., et al. (2021). Sex- and Age-specific Genetic Analysis of Chronic Back Pain. *Pain* 162 (4), 1176–1187. doi:10.1097/j.pain.0000000000002100

Gage, S. H., Jones, H. J., Burgess, S., Bowden, J., Davey Smith, G., Zammit, S., et al. (2017). Assessing Causality in Associations between Cannabis Use and Schizophrenia Risk: a Two-Sample Mendelian Randomization Study. *Psychol. Med.* 47 (5), 971–980. doi:10.1017/S0033291716003172

Gao, S., Zhou, H., Luo, S., Cai, X., Ye, F., He, Q., et al. (2021). *Investigating the Causal Relationship between Physical Activity and Chronic Back Pain: A Bidirectional Two-Sample Mendelian Randomization Study*. MedRxiv. doi:10.1101/2021.07.20.21260847

Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., et al. (2015). A Global Reference for Human Genetic Variation. *Nature* 526 (7571), 68–74. doi:10.1038/nature15393

Hartvigsen, J., Hancock, M. J., Kongsted, A., Louw, Q., Ferreira, M. L., Genevay, S., et al. (2018). What Low Back Pain Is and Why We Need to Pay Attention. *The Lancet* 391 (10137), 2356–2367. doi:10.1016/S0140-6736(18)30480-X

Hartvigsen, J., Nielsen, J., Kyvik, K. O., Fejer, R., Vach, W., Iachine, I., et al. (2009). Heritability of Spinal Pain and Consequences of Spinal Pain: A Comprehensive Genetic Epidemiologic Analysis Using a Population-Based Sample of 15,328 Twins Ages 20-71 Years. *Arthritis Rheum.* 61 (10), 1343–1351. doi:10.1002/art.24607

Hartwig, F. P., Borges, M. C., Horta, B. L., Bowden, J., and Davey Smith, G. (2017). Inflammatory Biomarkers and Risk of Schizophrenia. *JAMA Psychiatry* 74 (12), 1226–1233. doi:10.1001/jamapsychiatry.2017.3191

Hemani, G., Bowden, J., and Davey Smith, G. (2018). Evaluating the Potential Role of Pleiotropy in Mendelian Randomization Studies. *Hum. Mol. Genet.* 27 (R2), R195–R208. doi:10.1093/hmg/ddy163

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., et al. (2018). The MR-Base Platform Supports Systematic Causal Inference across the Human Phenome. *Elife* 7, 7. doi:10.7554/eLife.34408

Heneweer, H., Vanhees, L., and Picavet, S. J. H. (2009). Physical Activity and Low Back Pain: a U-Shaped Relation? *Pain* 143 (1-2), 21–25. doi:10.1016/j.pain.2008.12.033

Kamada, M., Kitayuguchi, J., Lee, I.-M., Hamano, T., Imamura, F., Inoue, S., et al. (2014). Relationship between Physical Activity and Chronic Musculoskeletal Pain Among Community-Dwelling Japanese Adults. *J. Epidemiol.* 24 (6), 474–483. doi:10.2188/jea.je20140025

Klimentidis, Y. C., Raichlen, D. A., Bea, J., Garcia, D. O., Wineinger, N. E., Mandarino, L. J., et al. (2018). Genome-wide Association Study of Habitual Physical Activity in over 377,000 UK Biobank Participants Identifies Multiple Variants Including CADM2 and APOE. *Int. J. Obes.* 42 (6), 1161–1176. doi:10.1038/s41366-018-0120-3

Koes, B. W., van Tulder, M., Lin, C.-W. C., Macedo, L. G., McAuley, J., and Maher, C. (2010). An Updated Overview of Clinical Guidelines for the Management of Non-specific Low Back Pain in Primary Care. *Eur. Spine J.* 19 (12), 2075–2094. doi:10.1007/s00586-010-1502-y

Ness, A. R., Leary, S. D., Mattocks, C., Blair, S. N., Reilly, J. J., Wells, J., et al. (2007). Objectively Measured Physical Activity and Fat Mass in a Large Cohort of Children. *Plos Med.* 4 (3), e97. doi:10.1371/journal.pmed.0040097

Papadimitriou, N., Dimou, N., Tsilidis, K. K., Banbury, B., Martin, R. M., Lewis, S. J., et al. (2020). Physical Activity and Risks of Breast and Colorectal Cancer: a Mendelian Randomisation Analysis. *Nat. Commun.* 11 (1), 597. doi:10.1038/s41467-020-14389-8

Picavet, H. S. J., and Schuit, A. J. (2003). Physical Inactivity: a Risk Factor for Low Back Pain in the General Population? *J. Epidemiol. Community Health* 57 (7), 517–518. doi:10.1136/jech.57.7.517

Pierce, B. L., and Burgess, S. (2013). Efficient Design for Mendelian Randomization Studies: Subsample and 2-sample Instrumental Variable Estimators. *Am. J. Epidemiol.* 178 (7), 1177–1184. doi:10.1093/aje/kwt084

Power, C., Frank, J., Hertzman, C., Schierhout, G., and Li, L. (2001). Predictors of Low Back Pain Onset in a Prospective British Study. *Am. J. Public Health* 91 (10), 1671–1678. doi:10.2105/ajph.91.10.1671

Qaseem, A., Wilt, T. J., McLean, R. M., and Forciea, M. A. (2017). Noninvasive Treatments for Acute, Subacute, and Chronic Low Back Pain: A Clinical Practice Guideline from the American College of Physicians. *Ann. Intern. Med.* 166 (7), 514–530. doi:10.7326/M16-2367

Ren, F., Wang, D., Wang, Y., Chen, P., and Guo, C. (2020). SPOCK2 Affects the Biological Behavior of Endometrial Cancer Cells by Regulation of MT1-MMP and MMP2. *Reprod. Sci.* 27 (7), 1391–1399. doi:10.1007/s43032-020-00197-4

Shiri, R., and Falah-Hassani, K. (2017). Does Leisure Time Physical Activity Protect against Low Back Pain? Systematic Review and Meta-Analysis of 36 Prospective Cohort Studies. *Br. J. Sports Med.* 51 (19), 1410–1418. doi:10.1136/bjsports-2016-097352

Shiri, R., Karppinen, J., Leino-Arjas, P., Solovieva, S., and Viikari-Juntura, E. (2010). The Association between Smoking and Low Back Pain: a Meta-Analysis. *Am. J. Med.* 123 (1), 87e7–87. doi:10.1016/j.amjmed.2009.05.028

Suri, P., Boyko, E. J., Smith, N. L., Jarvik, J. G., Williams, F. M. K., Jarvik, G. P., et al. (2017). Modifiable Risk Factors for Chronic Back Pain: Insights Using the Co-twin Control Design. *Spine J.* 17 (1), 4–14. doi:10.1016/j.spinee.2016.07.533

Suri, P., Palmer, M. R., Tsepilov, Y. A., Freidin, M. B., Boer, C. G., Yau, M. S., et al. (2018). Genome-wide Meta-Analysis of 158,000 Individuals of European Ancestry Identifies Three Loci Associated with Chronic Back Pain. *Plos Genet.* 14 (9), e1007601. doi:10.1371/journal.pgen.1007601

Taylor, J. B., Goode, A. P., George, S. Z., and Cook, C. E. (2014). Incidence and Risk Factors for First-Time Incident Low Back Pain: a Systematic Review and Meta-Analysis. *Spine J.* 14 (10), 2299–2319. doi:10.1016/j.spinee.2014.01.026

Urquhart, D. M., Berry, P., Wluka, A. E., Strauss, B. J., Wang, Y., Proietto, J., et al. (2011). 2011 Young Investigator Award Winner. *Spine* 36 (16), 1320–1325. doi:10.1097/BRS.0b013e3181f9fb66

Verbanck, M., Chen, C.-Y., Neale, B., and Do, R. (2018). Detection of Widespread Horizontal Pleiotropy in Causal Relationships Inferred from Mendelian Randomization between Complex Traits and Diseases. *Nat. Genet.* 50 (5), 693–698. doi:10.1038/s41588-018-0099-7

Wu, A., March, L., Zheng, X., Huang, J., Wang, X., Zhao, J., et al. (2020). Global Low Back Pain Prevalence and Years Lived with Disability from 1990 to 2017: Estimates from the Global Burden of Disease Study 2017. *Ann. Transl Med.* 8(6):299. doi:10.21037/atm.2020.02.175

Zhang, T.-T., Liu, Z., Liu, Y.-L., Zhao, J.-J., Liu, D.-W., and Tian, Q.-B. (2018). Obesity as a Risk Factor for Low Back Pain. *Clin. Spine Surg.* 31 (1), 22–27. doi:10.1097/BSD.0000000000000468

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

SYSTEMATIC REVIEW
published: 05 January 2022
doi: 10.3389/fgene.2021.783078

Different Associations Between *CDKAL1* Variants and Type 2 Diabetes Mellitus Susceptibility: A Meta-analysis

Qiaoli Zeng[1,2,3†], *Dehua Zou*[2,3,4†], *Shanshan Gu*[3,5†], *Fengqiong Han*[6], *Shilin Cao*[7*], *Yue Wei*[8*] *and Runmin Guo*[1,2,3,9*]

[1]Department of Internal Medicine, Shunde Women and Children's Hospital (Maternity and Child Healthcare Hospital of Shunde Foshan), Guangdong Medical University, Foshan, China, [2]Key Laboratory of Research in Maternal and Child Medicine and Birth Defects, Guangdong Medical University, Foshan, China, [3]Matenal and Child Research Institute, Shunde Women and Children's Hospital (Maternity and Child Healthcare Hospital of Shunde Foshan), Guangdong Medical University, Foshan, China, [4]State Key Laboratory for Quality Research of Chinese Medicines, Macau University of Science and Technology, Taipa, Macau SAR, China, [5]Institute of Neurology, Affiliated Hospital of Guangdong Medical University, Zhanjiang, China, [6]Department of Obstetric, Shunde Women and Children's Hospital (Maternity and Child Healthcare Hospital of Shunde Foshan), Guangdong Medical University, Foshan, China, [7]Department of Medical, Shunde Women and Children's Hospital (Maternity and Child Healthcare Hospital of Shunde Foshan), Guangdong Medical University, Foshan, China, [8]Department of Ultrasound, Shunde Women and Children's Hospital (Maternity and Child Healthcare Hospital of Shunde Foshan), Guangdong Medical University, Foshan, China, [9]Department of Endocrinology, Affiliated Hospital of Guangdong Medical University, Zhanjiang, China

OPEN ACCESS

Edited by:
Liangcai Zhang,
Janssen Research and Development,
United States

Reviewed by:
Dalin Li,
Cedars Sinai Medical Center,
United States
Huaizhen Qin,
University of Florida, United States

****Correspondence:***
Shilin Cao
33033950@qq.com
Yue Wei
weiyue138@163.com
Runmin Guo
runmin.guo@gdmu.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *25 September 2021*
Accepted: *13 December 2021*
Published: *05 January 2022*

Citation:
Zeng Q, Zou D, Gu S, Han F, Cao S, Wei Y and Guo R (2022) Different Associations Between CDKAL1 *Variants and Type 2 Diabetes Mellitus Susceptibility: A Meta-analysis. Front. Genet. 12:783078. doi: 10.3389/fgene.2021.783078*

Background: *CDK5 regulatory subunit associated protein 1 like 1* (*CDKAL1*) is a major pathogenesis-related protein for type 2 diabetes mellitus (T2DM). Recently, some studies have investigated the association of *CDKAL1* susceptibility variants, including rs4712523, rs4712524, and rs9460546 with T2DM. However, the results were inconsistent. This study aimed to evaluate the association of *CDKAL1* variants and T2DM patients.

Methods: A comprehensive meta-analysis was performed to assess the association between *CDKAL1* SNPs and T2DM among dominant, recessive, additive, and allele models.

Results: We investigated these three *CDKAL1* variants to identify T2DM risk. Our findings were as follows: rs4712523 was associated with an increased risk of T2DM for the allele model (G vs A: OR = 1.172; 95% CI: 1.103–1.244; $p < 0.001$) and dominant model (GG + AG vs AA: OR = 1.464; 95% CI: 1.073–1.996; $p = 0.016$); rs4712524 was significantly associated with an increased risk of T2DM for the allele model (G vs A: OR = 1.146; 95% CI: 1.056–1.245; $p = 0.001$), additive model (GG vs AA: OR = 1.455; 95% CI: 1.265–1.673; $p < 0.001$) recessive model (GG vs AA + AG: OR = 1.343; 95% CI: 1.187–1.518; $p < 0.001$) and dominant model (GG + AG vs AA: OR = 1.221; 95% CI: 1.155–1.292; $p < 0.001$); and rs9460546 was associated with an increased risk of T2DM for the allele model (G vs T: OR = 1.215; 95% CI: 1.167–1.264; $p = 0.023$). The same results were found in the East Asian subgroup for the allele model.

Conclusions: Our findings suggest that *CDKAL1* polymorphisms (rs4712523, rs4712524, and rs9460546) are significantly associated with T2DM.

Keywords: type 2 diabetes mellitus, *CDKAL1*, polymorphisms, susceptibility, meta-analysis

1 INTRODUCTION

Type 2 diabetes mellitus (T2DM) is a complex disease characterized by insulin resistance in peripheral tissues and dysregulated insulin secretion by pancreatic β-cells (Li et al., 2020). The incidence of T2DM in adults has been increasing over recent decades (Yang et al., 2010; Tian et al., 2019) and is estimated to increase to over 700 million by 2045 (Saeedi et al., 2019; Li et al., 2020). T2DM is caused by genetic and environmental factors (Tian et al., 2019; Wu et al., 2014). Genetic variants are thought to be involved in the development of T2DM. Genome-wide association studies have indicated that some single nucleotide polymorphisms (SNPs) are critical risk factors for T2DM (Tian et al., 2019).

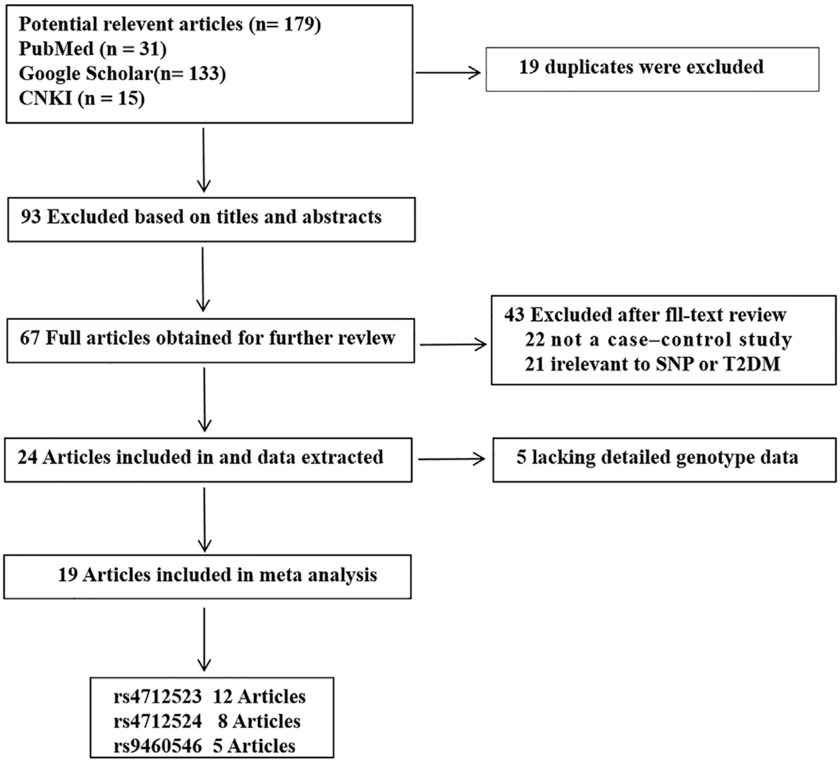

Potential relevent articles (n= 179)
PubMed (n = 31)
Google Scholar(n= 133)
CNKI (n = 15)
19 duplicates were excluded
93 Excluded based on titles and abstracts
67 Full articles obtained for further revlew
43 Excluded after fll-text review
22 not a case–control study
21 irelevant to SNP or T2DM
24 Articles included in and data extracted
5 lacking detailed genotype data
19 Articles included in meta analysis
rs4712523 12 Articles
rs4712524 8 Articles
rs9460546 5 Articles

FIGURE 1 | Flow diagram of the literature search and selection.

TABLE 1 | Characteristics of each study included in rs4712523 of meta-analysis.

Author	Year	Ethnic	T2DM/NDM	ORs with 95% CI (G vs A)	Allele distribution				Genotype distribution					
					T2DM, n		NDM, n		T2DM, n			NDM, n		
					A	G	A	G	AA	AG	GG	AA	AG	GG
Liju et al.	2020	India	1183/1188	1.077 (0.893–1.300)	1640	726	1684	692	—	—	—	—	—	—
Tian et al.	2019	Chinese	510/503	1.420 (1.190–1.690)	508	512	588	418	131	246	133	175	238	90
Qian et al.	2019	Chinese	526/526	1.027 (0.956–1.103)	590	462	556	496	164	262	100	149	258	119
Rao et al.	2016	Chinese	458/429	0.924 (0.766–1.114)	525	391	475	383	154	217	87	138	199	92
Ren et al.	2013	Chinese	98/97	1.521 (1.018–2.273)	99	97	118	76	9	81	8	26	66	5
Li et al.	2013	Chinese	192/190	1.654 (1.237–2.212)	202	182	246	134	22	158	12	62	122	6
Lu et al.	2012	Chinese	2897/3259	1.223 (1.139–1.314)	3105	2689	3816	2702	848	1409	640	1120	1576	563
Gong et al.	2016	Chinese	91/186	1.380 (1.250–1.520)	—	—	—	—	—	—	—	—	—	—
Long et al.	2012	African Americans	1549/2722	0.960 (0.870–1.070)	—	—	—	—	—	—	—	—	—	—
Takeuchi et al.	2009	Japanese	5629/6406	1.270 (1.210–1.330)	—	—	—	—	—	—	—	—	—	—
Takeuchi et al.	2009	Europeans	14586/17968	1.120 (1.080–1.160)	—	—	—	—	—	—	—	—	—	—
Rung et al.	2009	Caucasian	180/165	1.200 (1.140–1.260)	—	—	—	—	—	—	—	—	—	—
Scott et al.	2007	Finnish	1161/1174	1.123 (1.032–1.222)	—	—	—	—	—	—	—	—	—	—

n, Number; T2DM, type 2 diabetes mellitus; NDM, Non-diabetic subject; OR, odds ratio; CI, confidence interval.

TABLE 2 | Characteristics of each study included in rs4712524 of meta-analysis.

Author	Year	Ethnic	T2DM/NDM	Allele distribution				Genotype distribution					
				T2DM, n		NDM, n		T2DM, n			NDM, n		
				A	G	A	G	AA	AG	GG	AA	AG	GG
Liju et al.	2020	India	1183/1188	658	1708	624	1752	—	—	—	—	—	—
Li et al.	2020	Chinese	1169/1277	1324	1014	1551	1003	375	574	220	470	611	196
Azarova et al.	2020	Russian	1579/1627	1988	1170	2204	1050	636	716	227	721	762	144
Tian et al.	2019	Chinese	508/493	506	510	570	416	130	246	132	171	228	94
Li et al.	2018	Chinese	123/311	128	118	327	295	34	60	29	94	139	78
Rao et al.	2016	Chinese	456/417	521	391	457	377	150	221	85	125	207	85
Unoki et al.	2008	Japanese	4795/3441	5119	4471	4019	2863	1431	2257	1107	1176	1667	598
Lu et al.	2012	Chinese	2899/3260	3157	2641	3868	2652	880	1397	622	1156	1556	548

n, Number; T2DM, type 2 diabetes mellitus; NDM, Non-diabetic subject (-), not applicable.

CDK5 regulatory subunit associated protein 1 like 1 *(CDKAL1)* is a crucial pathogenesis-related protein for T2DM. The *CDKAL1* gene encodes cyclin-dependent kinase 5 regulatory subunit-associated protein 1 (CDK5RAP1)-like 1. Cyclin-dependent kinase 5 (CDK5) is a serine/threonine protein kinase that contributes to the glucose-dependent regulation of insulin secretion (Li et al., 2020); therefore, it plays a critical role in the pathophysiology of β-cell dysfunction and predisposition to T2DM (Li et al., 2020; Wei et al., 2005; Ubeda et al., 2006). The associations of many SNPs in *CDKAL1* with T2DM have been examined in some meta-analyses, but no published meta-analysis has evaluated the role of *CDKAL1* rs4712523, rs4712524 and rs9460546 variants in the susceptibility to T2DM. Several studies have examined the association between *CDKAL1* polymorphisms (rs4712523, rs4712524 and rs9460546) and T2DM risk, but some findings were failed to replicate. Therefore, performing a meta-analysis is needed to evaluate the association between *CDKAL1* polymorphisms (rs4712523, rs4712524, and rs9460546) and T2DM.

2 MATERIALS AND METHODS

This meta-analysis was conducted according to Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) guidelines.

TABLE 3 | Characteristics of each study included in rs9460546 of meta-analysis.

Author	Year	Ethnic	T2DM/NDM	ORs with 95% CI (G vs T)
Li et al.	2020	Chinese	1169/1277	1.133 (1.011–1.270)
Hu et al.	2009	Chinese	1849/1785	1.145 (1.041–1.260)
Herder et al.	2008	German	433/1438	1.410 (1.190–1.680)
Unoki et al.	2008	Japanese	4775/3442	1.226 (1.152–1.305)
Maller et al.	2012	European	632/677	1.250 (1.150–1.350)

T2DM, type 2 diabetes mellitus; NDM, Non-diabetic subject; OR, odds ratio; CI, confidence interval.

2.1 Literature Search

The Google Scholar, PubMed and Chinese National Knowledge Infrastructure databases were systematically searched for relevant studies using the following terms:

1 "CDKAL1" or "rs4712523" or "polymorphism" and "T2DM";
2 "CDKAL1" or "rs4712524" or "polymorphism" and "T2DM";
3 "CDKAL1", or "rs9460546" or "polymorphism" and "T2DM", respectively.

The search was performed with no date or language restrictions. All the studies were evaluated by reading the title and abstract and excluding irrelevant studies. The full texts of eligible studies were then assessed by reading the full text to confirm inclusion in the study.

2.2 Inclusion and Exclusion Criteria

The inclusion criteria of the studies were as follows: 1) case-control/cohort studies; 2) studies that evaluated the association between *CDKAL1* SNPs (rs4712523, rs4712524, and rs9460546) and T2DM; 3) adequate raw data or sufficient data to calculate odds ratios (ORs) with corresponding 95% confidence intervals (CIs); 4) a T2DM diagnosis based on the clinical criteria of the World Health Organization.

The exclusion criteria were as follows: 1) not a case-control/ cohort study; 2) not related to *CDKAL1* SNPs (rs4712523, rs4712524, and rs9460546) and T2DM; 3) insufficient data; 4) NDM data not in Hardy-Weinberg equilibrium (HWE).

2.3 Data Extraction

Two authors independently extracted the following data from the included studies: first author, ethnicity, year of publication, numbers of T2DM patients and NDM controls, distribution of alleles and genotypes, and ORs with 95% CIs of the allele distribution**.**

2.4 Statistical Analysis

Four genetic models were evaluated in rs4712523 and rs4712524: the dominant model (GG + AG vs AA), recessive model (GG vs AA + AG), additive model (GG vs AA) and allele model (G vs A). Additionally, the allele model (G vs T) was evaluated in rs9460546. Genetic heterogeneity was estimated using Q-test

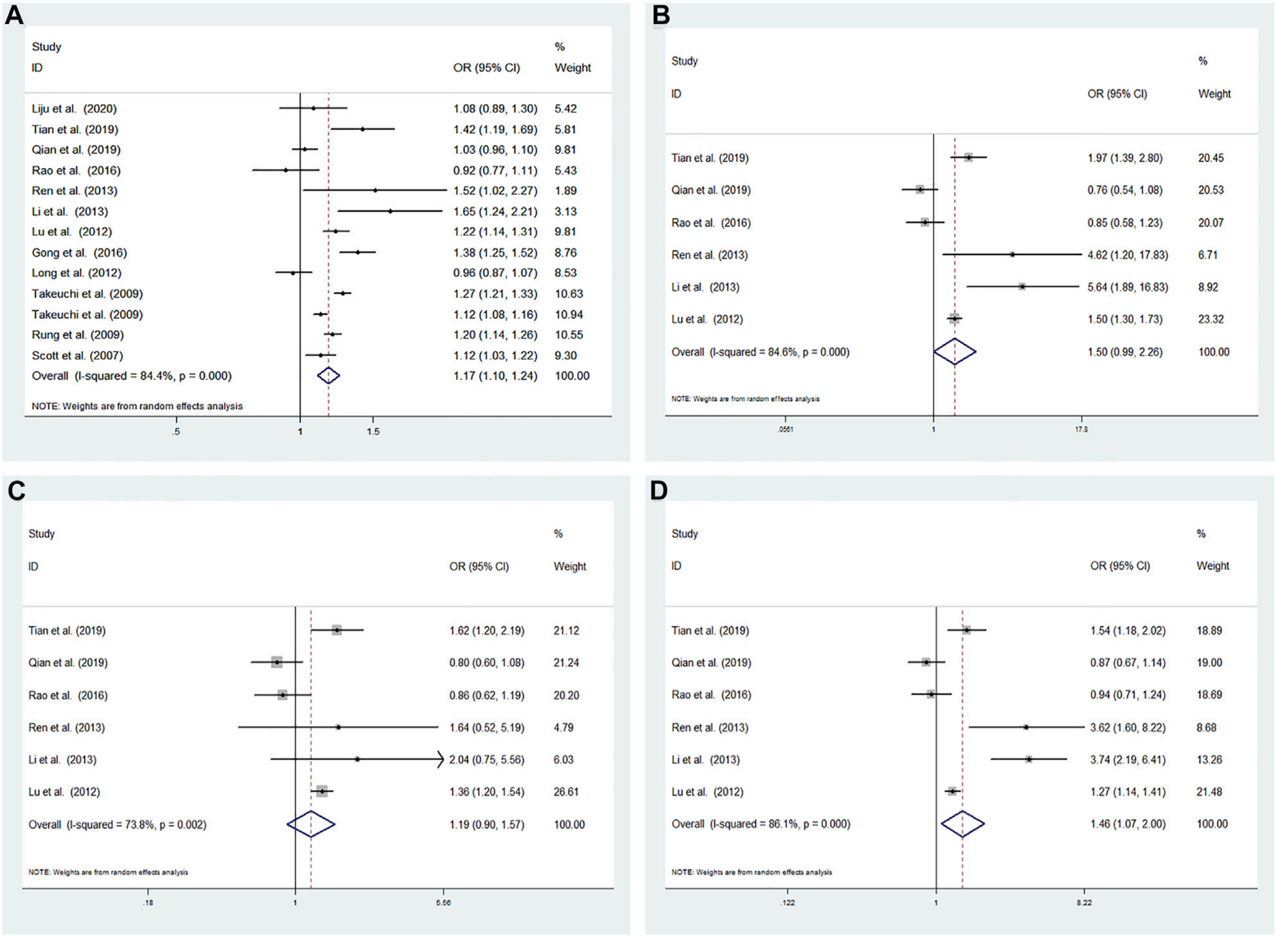


FIGURE 2 | Meta-analysis using a random effects model for the association between the CDKALI rs4712523 polymorphism and T2DM susceptibility **(A)** Allele model, G vs A **(B)** Additive model, GG vs AA **(C)** Recessive model, GG vs AA + AG **(D)** Dominant model, GG + AG vs AA. OR: odds ratio, CI: confidence interval, I-squared: measure to quantify the degree of heterogeneity in meta-analyses.

and I^2 test. Lower heterogeneity was defined as I^2 <50% and p > 0.01, using the fixed effects model (Mantel–Haenszel) to calculate ORs with corresponding 95% CIs. Otherwise, the random effects model (Mantel–Haenszel) was used. The significance of the ORs was evaluated using the Z test. Begg's and Egger's tests were used to determine publication bias. STATA v.14.0 software (Stata Corporation, Texas, United States) was used to perform all statistical analyses.

3 RESULTS

3.1 Study Inclusion and Characteristics

A total of 179 potential studies were searched using the inclusion and exclusion criteria. **Figure 1** shows a flow chart of the study selection process. Twelve articles, including 7 in English and 5 in Chinese, had rs4712523 data. Eight articles, including 5 in English, 2 in Chinese and 1 in Russian, had rs4712524 data. Five articles, including 5 in English, had rs9460546 data. The characteristics of each included study are shown in **Tables 1–3**.

3.2 Heterogeneity Analysis

3.2.1 rs4712523

High heterogeneity among studies (Scott et al., 2007; Rung et al., 2009; Takeuchi et al., 2009; Long et al., 2012; Lu et al., 2012; Gong, 2016; Li et al., 2013; Ren et al., 2013; Rao et al., 2016; Qian, 2019; Tian et al., 2019; Liju et al., 2020) was detected in the allele model (G vs A: I^2 = 84.4%; $p < 0.001$), additive model (GG vs AA: I^2 = 84.6%; $p < 0.001$), recessive model (GG vs AA + AG: I^2 = 73.8%; $p = 0.002$), and dominant model (GG + AG vs AA: I^2 = 86.1%; $p <$ 0.001) (**Figure 2**).

3.2.2 rs4712524

High heterogeneity among studies (Unoki et al., 2008; Lu et al., 2012; Rao et al., 2016; Li, 2018; Tian et al., 2019; Azarova, 2020; Li et al., 2020; Liju et al., 2020) was detected in the allele model (G vs A: I^2 = 75.1%; $p < 0.001$). A moderate degree of heterogeneity among studies was detected under the additive model (GG vs AA: I^2 = 58.7%; p = 0.024) and recessive model (GG vs AA + AG: I^2 = 57.8%; p = 0.027). Low heterogeneity among studies was detected under the

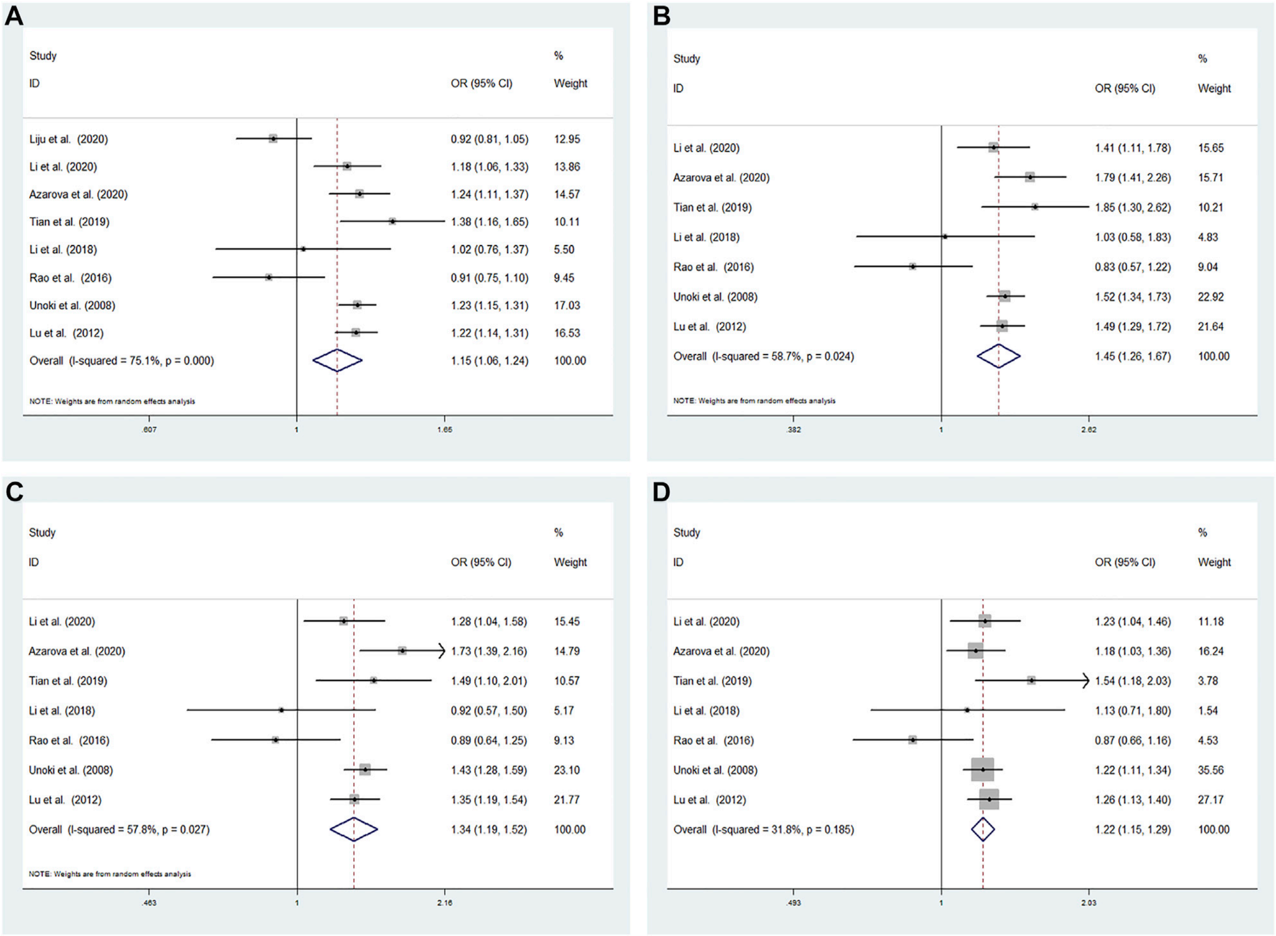


FIGURE 3 | Meta-analysis for the association between the CDKALI rs4712524 polymorphism and T2DM susceptibility **(A)** Allele model, G vs A (random effects model) **(B)** Additive model, GG vs AA (random effects model) **(C)** Recessive model, GG vs AA + AG (random effects model) **(D)** Dominant model, GG + AG vs AA (fixed effects model). OR: odds ratio, CI: confidence interval, I-squared: measure to quantify the degree of heterogeneity in meta-analyses.

dominant model (GG + AG vs AA: I^2 = 31.8%; p = 0.185) (**Figure 3**).

3.2.3 rs9460546

Low heterogeneity among studies (Herder et al., 2008; Unoki et al., 2008; Hu et al., 2009; Maller et al., 2012; Li et al., 2020) was detected in the allele model (G vs T: I^2 = 37.0%; p = 0.174) (**Figure 4**).

3.3 Meta-Analysis Results

3.3.1 rs4712523

A significant difference was found between T2DM patients and NDM controls for the allele model (G vs A: OR = 1.172; 95% CI: 1.103–1.245; $p < 0.001$) and dominant model (GG + AG vs AA: OR = 1.464; 95% CI: 1.073–1.996; p = 0.016). No significant associations were found under the additive model (GG vs AA: OR = 1.495; 95% CI: 0.990–2.257; p = 0.056) and recessive model (GG vs AA + AG: OR = 1.188; 95% CI: 0.900–1.568; p = 0.223) using a random effects model (**Figure 2**).

3.3.2 rs4712524

A random effects model was used to analyze the allele, additive and recessive models, and the dominant model was analyzed using a fixed effects model. A significant difference was found between T2DM patients and NDM controls for the allele model (G vs A: OR = 1.146; 95% CI: 1.056–1.245; p = 0.001), additive model (GG vs AA: OR = 1.455; 95% CI: 1.265–1.673; $p < 0.001$) recessive model (GG vs AA + AG: OR = 1.343; 95% CI: 1.187–1.518; $p < 0.001$) and dominant model (GG + AG vs AA: OR = 1.221; 95% CI: 1.155–1.292; $p < 0.001$) (**Figure 3**).

3.3.3 rs9460546

A significant difference was found between T2DM patients and NDM controls for the allele model (G vs T: OR = 1.215; 95% CI: 1.167–1.264; p = 0.023) using a fixed effects model (**Figure 4**).

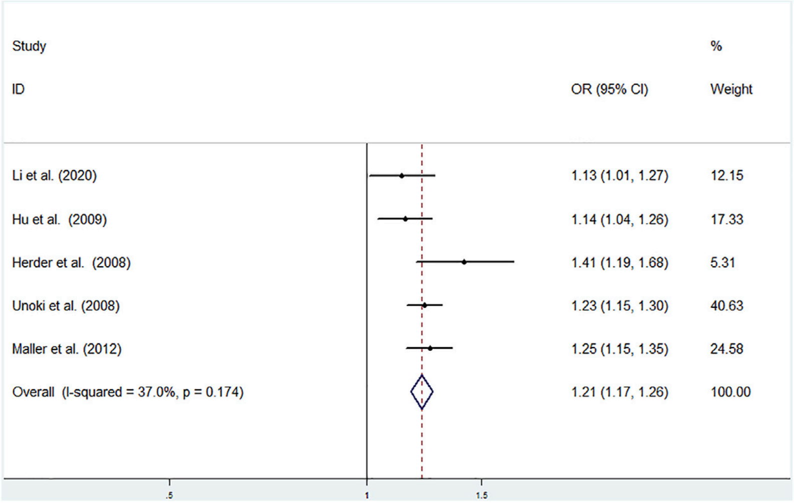


FIGURE 4 | Meta-analysis using a fixed effects model for the association between the *CDKAL1* rs9460546 polymorphism and T2DM susceptibility (Allele model, G vs T). OR: odds ratio, CI: confidence interval, I-squared: measure to quantify the degree of heterogeneity in meta-analyses.

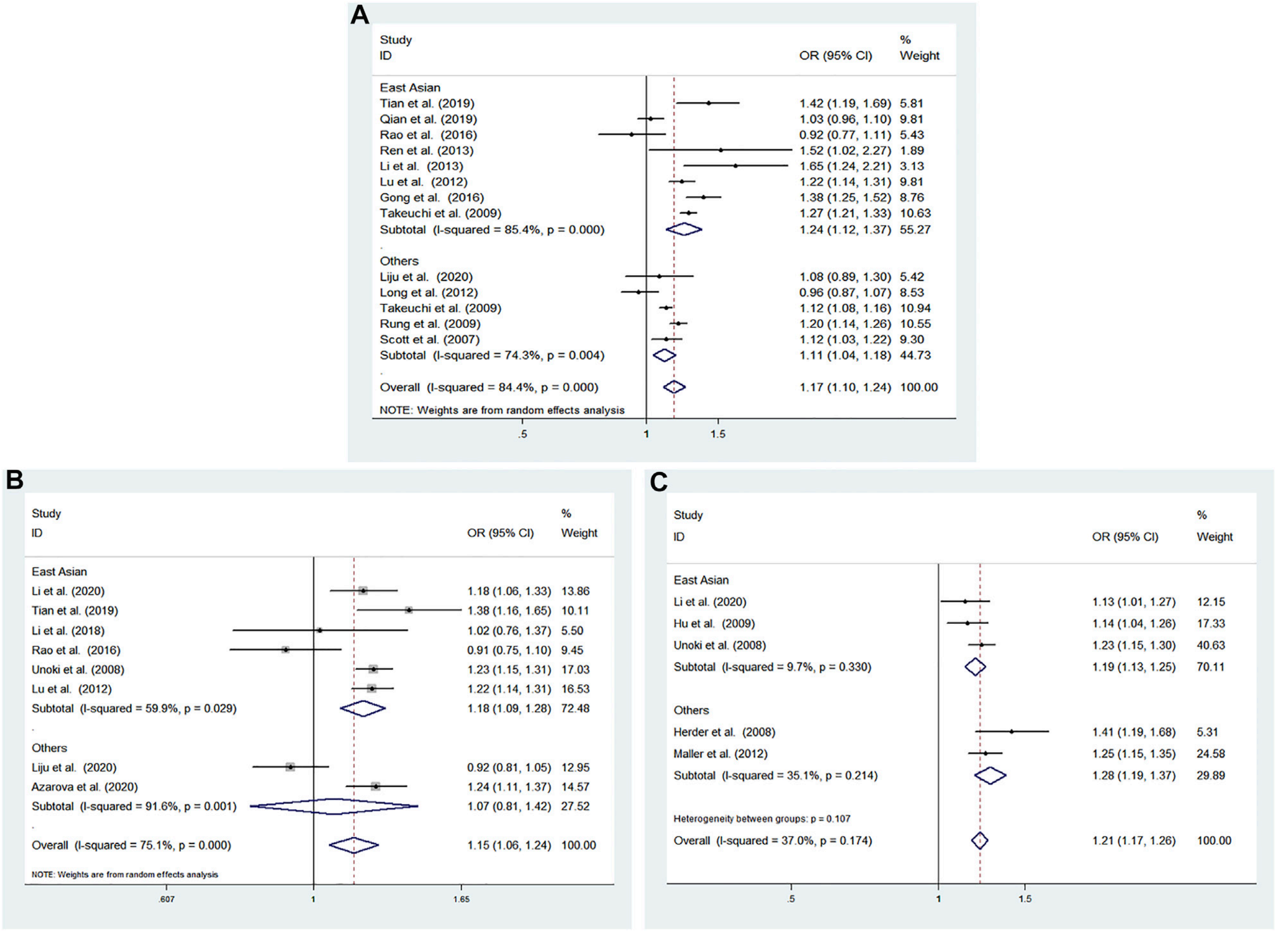


FIGURE 5 | Association between the CDKALI variants and T2DM susceptibility in the subgroup for the allele model **(A)** rs4712523: G vs A (random effects model) **(B)** rs4712524: G vs A (random effects model) **(C)** rs9460546: G vs T (fixed effects model). OR: odds ratio, CI: confidence interval, I-squared: measure to quantify the degree of heterogeneity in meta-analyses.

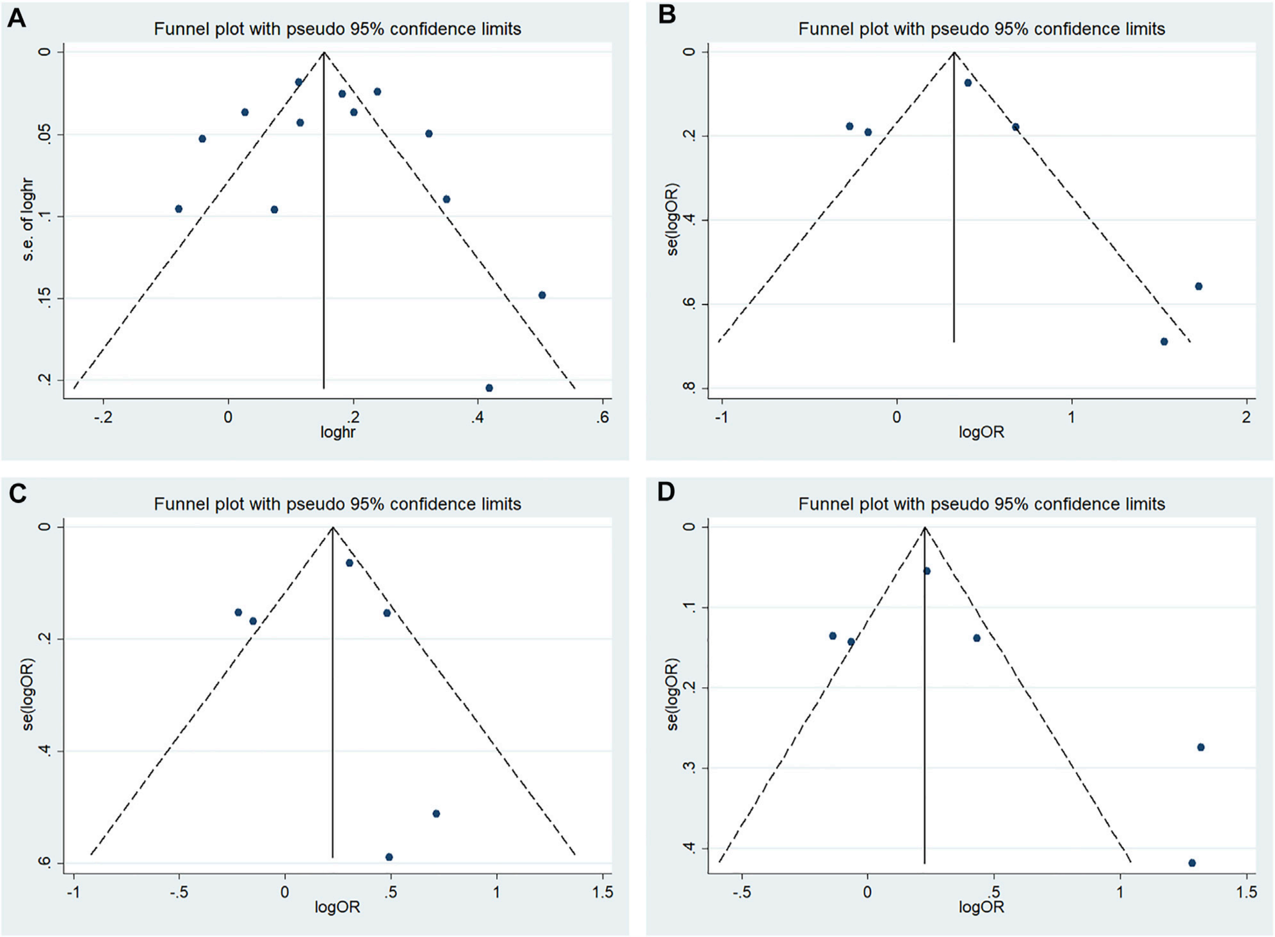


FIGURE 6 | Funnel plot of the odds ratios in the CDKALI rs4712523 meta-analysis **(A)** Allele model, G vs A **(B)** Additive model, GG vs AA **(C)** Recessive model, GG vs AA + AG **(D)** Dominant model, GG + AG vs AA.

3.4 Subgroup Analyses

3.4.1 rs4712523

We performed subgroup analysis according to ethnicity to evaluate the association between rs4712523 and T2DM susceptibility in the allele model. Rs35767 was significantly related to the risk of T2DM in the East Asian (G vs A: OR = 1.241; 95% CI: 1.123–1.371; $p < 0.001$) and others subgroup (G vs A: OR = 1.108; 95% CI: 1.039–1.180; $p = 0.002$) using a random effects model (**Figure 5A**).

3.4.2 rs4712524

We performed subgroup analysis according to ethnicity to evaluate the association between rs4712524 and T2DM susceptibility in the allele model. Rs4712524 was significantly related to the risk of T2DM in the East Asian (G vs A: OR = 1.182; 95% CI: 1.095–1.277; $p < 0.001$), but no significant associations were found in others subgroup (G vs A: OR = 1.071; 95% CI: 0.807–1.423; $p = 0.634$) using a random effects model (**Figure 5B**).

3.4.3 rs9460546

We performed subgroup analysis according to ethnicity to evaluate the association between rs9460546 and T2DM susceptibility in the allele model. Rs9460546 was significantly related to the risk of T2DM in the East Asian (G vs T: OR = 1.189; 95% CI: 1.134–1.247; $p < 0.001$) and others subgroup (G vs T: OR = 1.277; 95% CI: 1.188–1.373; $p < 0.001$) using a fixed effects model (**Figure 5C**).

3.5 Publication Bias

According to Begg's and Egger's tests, no significant publication bias was found in each of the genetic models (all $p > 0.05$, data not shown), and the funnel plots are shown in **Figures 6–9**.

4 DISCUSSION

CDKAL1 is a key pathogenesis-related protein for T2DM (Tian et al., 2019). Genetic variants may play an essential role in T2DM

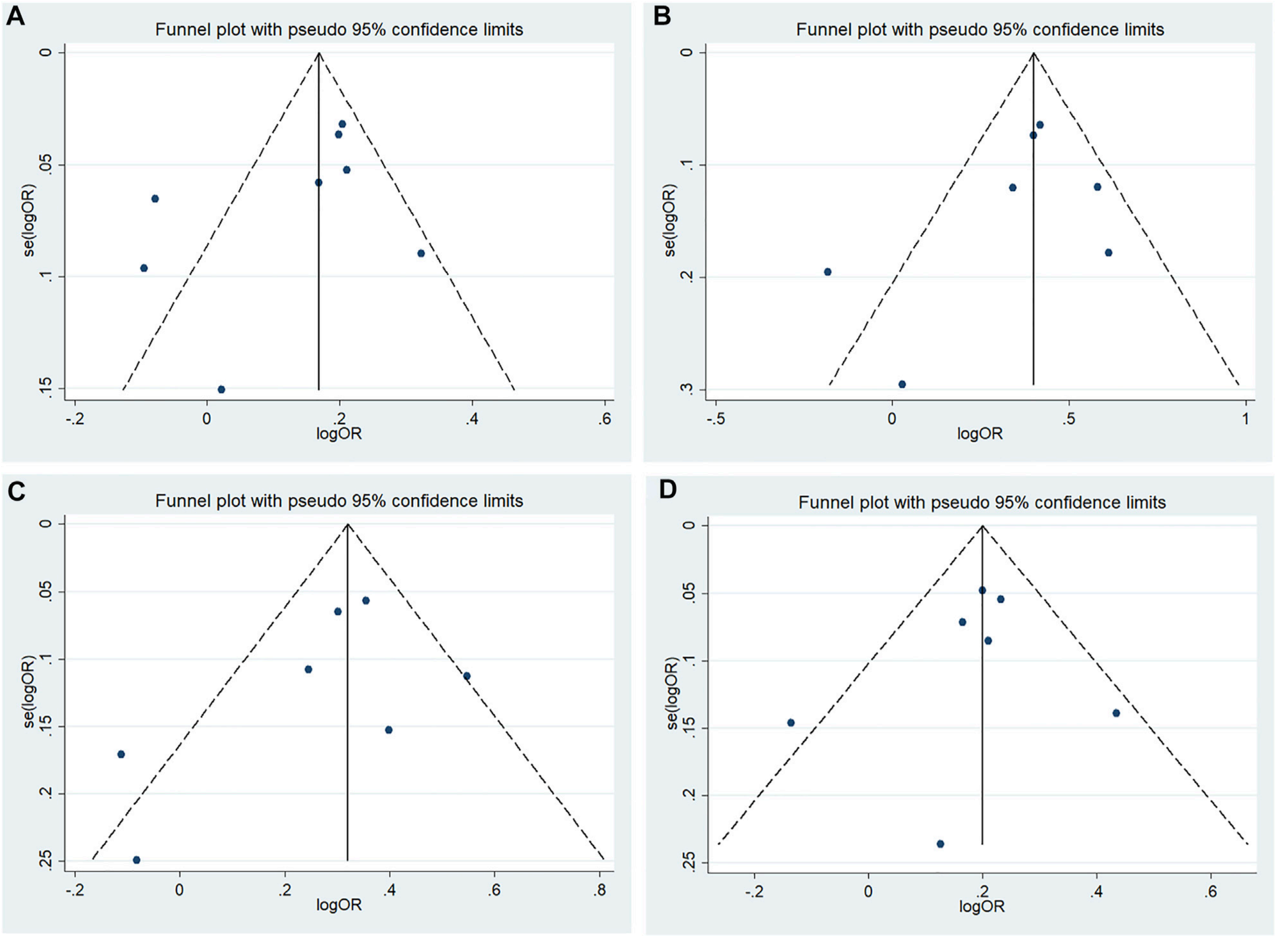


FIGURE 7 | Funnel plot of the odds ratios in the CDKALI rs4712524 meta-analysis **(A)** Allele model, G vs A **(B)** Additive model, GG vs AA **(C)** Recessive model, GG vs AA + AG **(D)** Dominant model, GG + AG vs A.

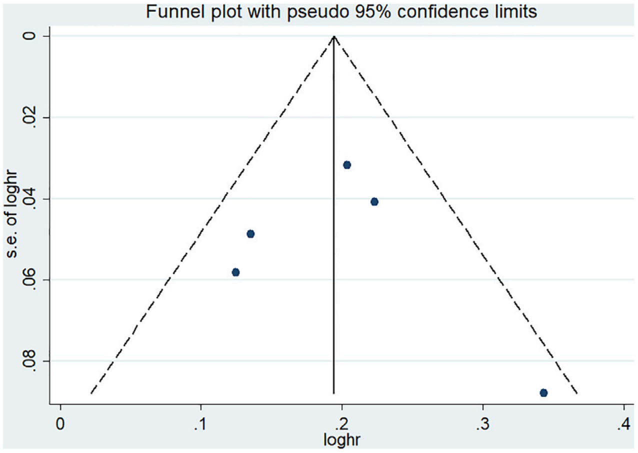


FIGURE 8 | Funnel plot of the odds ratios in the CDKALI rs9460546 meta-analysis for the allele model (G vs T).

susceptibility. In this meta-analysis, three SNPs (rs4712523, rs4712524, and rs9460546) from previous studies were evaluated to determine the association of *CDKAL1* polymorphisms with T2DM. *CDKAL1* polymorphisms (rs4712523, rs4712524, and rs9460546) showed a significant association with T2DM. Our results were consistent with some previous study findings.

The results revealed that the G allele and GG + AG genotypes of rs4712523 were associated with an increased risk of T2DM. Nine of the thirteen previous studies investigated rs4712523 showed an association between the G allele and T2DM (Scott et al., 2007; Rung et al., 2009; Takeuchi et al., 2009; Long et al., 2012; Lu et al., 2012; Gong, 2016; Li et al., 2013; Ren et al., 2013; Tian et al., 2019), and four studies found an association between the GG + AG genotypes and T2DM (Lu et al., 2012; Li et al., 2013; Ren et al., 2013; Tian et al., 2019). In addition, the rs4712524 G allele, GG and GG + AG genotypes were associated with an increased risk of T2DM susceptibility. That have been confirmed previous observations (Unoki et al., 2008; Lu et al., 2012; Tian et al., 2019; Azarova, 2020; Li et al., 2020). Additionally, the

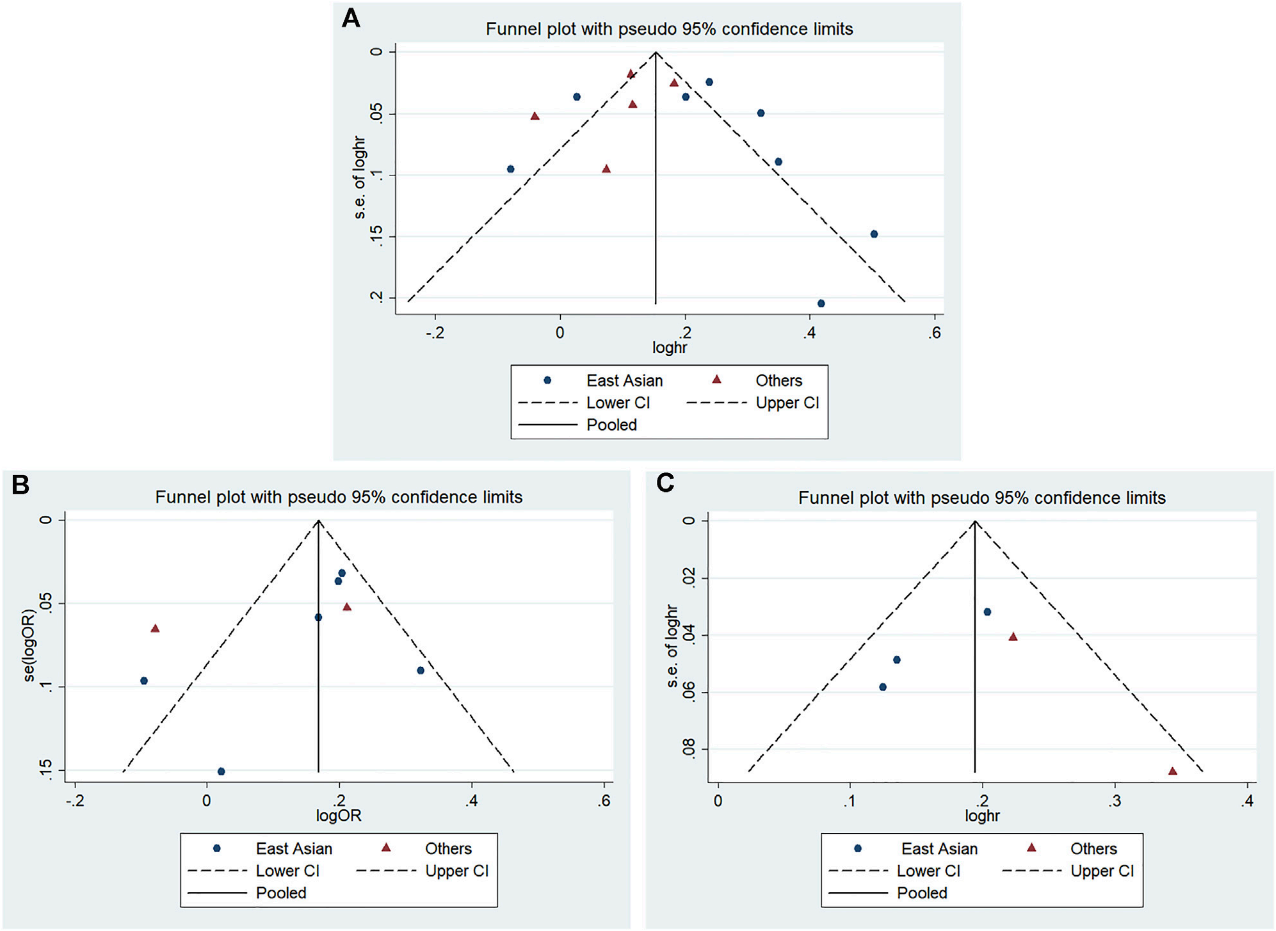


FIGURE 9 | Funnel plot of the odds ratios in the CDKALI variants in the subgroup meta-analysis for the allele model **(A)** rs4712523: G vs A **(B)** rs4712524: G vs A **(C)** rs9460546: G vs T.

results showed that rs9460546 G allele was associated with T2DM susceptibility. Markedly, all five studies found that the rs9460546 G allele was associated with T2DM in various populations (Herder et al., 2008; Unoki et al., 2008; Hu et al., 2009; Maller et al., 2012; Li et al., 2020). Moreovr, rs4712523, rs4712524, and rs9460546 showed a significant association with T2DM in the East Asian subgroup for the allele model. In general, Our results have confirmed previous observations suggesting that *CDKAL1* may play a role in T2DM. But it is worth noting that high heterogeneity among studies was detected in rs4712523 and rs4712524 likely because of the difference in country, ethnicity, genetic background and environmental factors. Subgroup analyses were performed by ethnicity in the allele model, and the subgroup still had high heterogeneity. Importantly, the high heterogeneity among studies might have affected our data.

CDKAL1 expression in human pancreatic β-cells increases insulin secretion by inhibiting CDK5 (Li et al., 2020; Wei et al., 2005; Ubeda et al., 2006; Ching et al., 2002). Subsequently, several studies have shown the association of genetic variants in *CDKAL1* with defects in proinsulin conversion and the insulin response following glucose stimulation (Pascoe et al., 2007; Steinthorsdottir et al., 2007; Tian et al., 2019). Thus, *CDKAL1* is involved in the development of T2DM. Genome-wide association studies have identified several SNPs in the *CDKAL1* gene associated with T2D (Saxena et al., 2007; Scott et al., 2007; Tian et al., 2019). Our results confirmed the significant association between *CDKAL1* SNPs and T2DM susceptibility. However, the mechanisms must be verified in functional studies. Our association results provide reference data to identify new biomarkers of T2DM that could contribute to the diagnosis of T2DM.

This meta-analysis has a few limitations. First, because of the limited examination of *CDKAL1* variants in T2DM, the included studies had comparatively small sample sizes, which might affect the results of the meta-analysis because of insufficient statistical power. Thus, studies must be performed across different geographical and ethnic groups. Additionally, the factors of T2DM might be complex, with the contribution of genetic, environmental and dietary habits. Therefore, further study is required to evaluate whether other risk factors together with the *CDKAL1* gene influence T2DM susceptibility.

5 CONCLUSION

To our knowledge, this study is the first to assess the role of *CDKAL1* polymorphisms (rs4712523, rs4712524, and rs9460546) in T2DM. Significant associations were found between the *CDKAL1* rs4712523, rs4712524, and rs9460546 polymorphisms and susceptibility to T2DM.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

QZ, DZ and SG were responsible for the study design, statistical analysis, and manuscript preparation. QZ and FH managed the literature searches and analyses. The study was supervised by SC, YW and RG.

FUNDING

Support for this work includes funding from the National Natural Science Foundation of China (81873649); Doctoral scientific research Initiate funding project of Shunde Women and Children's Hospital of Guangdong Medical University (Maternity and Child Healthcare Hospital of Shunde Foshan) (2020BSQD007); Guangdong Medical University Research Foundation (GDMUM2020008 and GDMUM2020012); Medical Research Project of Foshan Health Bureau (20210188 and 20210289).

REFERENCES

Azarova, L. (2020). POLYMORPHIC VARIANTS rs4712524 AND rs6931514 IN THE CDKAL1 GENE: ASSOCIATION WITH THE RISK OF DEVELOPMENT OF TYPE 2 DIABETES MELLITUS IN RUSSIAN POPULATIONE. Естественные и технические науки (in Russian) 9. doi:10.25633/ETN.2020.09.02

Ching, Y.-P., Pang, A. S. H., Lam, W.-H., Qi, R. Z., and Wang, J. H. (2002). Identification of a Neuronal Cdk5 Activator-Binding Protein as Cdk5 Inhibitor. *J. Biol. Chem.* 277 (18), 15237–15240. doi:10.1074/jbc.C200032200

Gong, X. (2016). "The Genetic Diversity Analysis of Candidate Genes Associated with Type 2 Diabetes in Xinjiang Ethnic Minority Populations," in *A Dissertation Submitted to Xinjiang Medical University* (Xinjiang: Xinjiang Medical University) (in Chinese).

Herder, C., Rathmann, W., Strassburger, K., Finner, H., Grallert, H., Huth, C., et al. (2008). Variants of thePPARG,IGF2BP2,CDKAL1,HHEX, andTCF7L2Genes Confer Risk of Type 2 Diabetes Independently of BMI in the German KORA Studies. *Horm. Metab. Res.* 40 (10), 722–726. doi:10.1055/s-2008-1078730

Hu, C., Zhang, R., Wang, C., Wang, J., Ma, X., Lu, J., et al. (2009). PPARG, KCNJ11, CDKAL1, CDKN2A-Cdkn2b, IDE-KIF11-HHEX, IGF2BP2 and SLC30A8 Are Associated with Type 2 Diabetes in a Chinese Population. *PLoS One* 4 (10), e7643. doi:10.1371/journal.pone.0007643

Li, C., Shen, K., Yang, M., Yang, Y., Tao, W., He, S., et al. (2020). Association between Single Nucleotide Polymorphisms in *CDKAL1* and *HHEX* and Type 2 Diabetes in Chinese Population. *Dmso* 13, 5113–5123. doi:10.2147/DMSO.S288587

Li, X., Su, Y., Yan, Z., Gu, L., Li, C., and Li, A. (2013). CDKALl Rs4712523 Polymorphism Is Associatedwith Type 2 Diabetes in Han Population in Inner Mongolia. *Basic Clin. Med.* 33, 3, 2013 (in Chinese). doi:10.16352/j.issn.1001-6325.2013.03.017

Li, Y. (2018). "Association between Genetic Polymorphisms and Comorbidity of Coronary Heart Disease and Type 2 Diabetes in the Elderly," in A *Dissertation Submitted to Peking Union Medical College* (Beijing: Peking Union Medical College) (in Chinese).

Liju, S., Chidambaram, M., Mohan, V., and Radha, V. (2020). Impact of Type 2 Diabetes Variants Identified through Genome-wide Association Studies in Early-Onset Type 2 Diabetes from South Indian Population. *Genomics Inform.* 18 (3), e27. doi:10.5808/GI.2020.18.3.e27

Long, J., Edwards, T., Signorello, L. B., Cai, Q., Zheng, W., Shu, X.-O., et al. (2012). Evaluation of Genome-wide Association Study-Identified Type 2 Diabetes Loci in African Americans. *Am. J. Epidemiol.* 176 (11), 995–1001. doi:10.1093/aje/kws176

Lu, F., Qian, Y., Li, H., Dong, M., Lin, Y., Du, J., et al. (2012). Genetic Variants on Chromosome 6p21.1 and 6p22.3 Are Associated with Type 2 Diabetes Risk: a Case-Control Study in Han Chinese. *J. Hum. Genet.* 57 (5), 320–325. doi:10.1038/jhg.2012.25

Maller, J. B., McVean, G., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., et al. (2012). Bayesian Refinement of Association Signals for 14 Loci in 3 Common Diseases. *Nat. Genet.* 44 (12), 1294–1301. doi:10.1038/ng.2435

Pascoe, L., Tura, A., Patel, S. K., Ibrahim, I. M., Ferrannini, E., Zeggini, E., et al. (2007). Common Variants of the Novel Type 2 Diabetes Genes CDKAL1 and HHEX/IDE Are Associated with Decreased Pancreatic -Cell Function. *Diabetes* 56 (12), 3101–3104. doi:10.2337/db07-0634

Qian, X. (2019). "Associations of Polymorphisms of CDKALI Gene with Type 2 Diabetes Mellitus and Diabetes-Related Traits," in *A Thesis Submitted to Zhengzhou University for the Degree of Master* (Zhengzhou: Zhengzhou University) (in Chinese).

Rao, P., Yu, X., Gai, S., Wang, H., Fang, H., Wang, Y., et al. (2016). Association of IGF2BP2, CDKAL1 Gene Polymorphism and Gene Environment Interaction with Type 2 Diabetes Mellitus. *Mod. instrument Med. Treat.* 22, 2, 2016 (in Chinese). doi:10.11876/mimt201602008

Ren, X., Yan, Z., Li, X., Su, Y., and Zhang, S. (2013). Association of CDKAL1 Rs4712523 Polymorphism with Susceptibility to the Blood Lipid of Type 2 Diabetes in Han Population of Inner Mongolia. *Chin. J. Clinicians* 7, 17, 2013 (in Chinese). doi:10.3877/cma.j.issn.1674-0785.2013.17.013

Rung, J., Cauchi, S., Albrechtsen, A., Shen, L., Rocheleau, G., Cavalcanti-Proença, C., et al. (2009). Genetic Variant Near IRS1 Is Associated with Type 2 Diabetes, Insulin Resistance and Hyperinsulinemia. *Nat. Genet.* 41 (10), 1110–1115. doi:10.1038/ng.443

Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., et al. (2019). Global and Regional Diabetes Prevalence Estimates for 2019 and Projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th Edition. *Diabetes Res. Clin. Pract.* 157, 107843. doi:10.1016/j.diabres.2019.107843

Saxena, R., Voight, B. F., Lyssenko, V., Burtt, N. P., de Bakker, P. I. W., Chen, H., et al. (2007). Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels. *Science* 316 (5829), 1331–1336. doi:10.1126/science.1142358

Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L., et al. (2007). A Genome-wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants. *Science* 316 (5829), 1341–1345. doi:10.1126/science.1142382

Steinthorsdottir, V., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Jonsdottir, T., Walters, G. B., et al. (2007). A Variant in CDKAL1 Influences Insulin

Response and Risk of Type 2 Diabetes. *Nat. Genet.* 39 (6), 770–775. doi:10.1038/ng2043

Takeuchi, F., Serizawa, M., Yamamoto, K., Fujisawa, T., Nakashima, E., Ohnaka, K., et al. (2009). Confirmation of Multiple Risk Loci and Genetic Impacts by a Genome-wide Association Study of Type 2 Diabetes in the Japanese Population. *Diabetes* 58 (7), 1690–1699. doi:10.2337/db08-1494

Tian, Y., Xu, J., Huang, T., Cui, J., Zhang, W., Song, W., et al. (2019). A Novel Polymorphism (Rs35612982) in CDKAL1 Is a Risk Factor of Type 2 Diabetes: A Case-Control Study. *Kidney Blood Press. Res.* 44 (6), 1313–1326. doi:10.1159/000503175

Ubeda, M., Rukstalis, J. M., and Habener, J. F. (2006). Inhibition of Cyclin-dependent Kinase 5 Activity Protects Pancreatic Beta Cells from Glucotoxicity. *J. Biol. Chem.* 281 (39), 28858–28864. doi:10.1074/jbc.M604690200

Unoki, H., Takahashi, A., Kawaguchi, T., Hara, K., Horikoshi, M., Andersen, G., et al. (2008). SNPs in KCNQ1 Are Associated with Susceptibility to Type 2 Diabetes in East Asian and European Populations. *Nat. Genet.* 40 (9), 1098–1102. doi:10.1038/ng.208

Wei, F.-Y., Nagashima, K., Ohshima, T., Saheki, Y., Lu, Y.-F., Matsushita, M., et al. (2005). Cdk5-dependent Regulation of Glucose-Stimulated Insulin Secretion. *Nat. Med.* 11 (10), 1104–1108. doi:10.1038/nm1299

Wu, Y., Ding, Y., Tanaka, Y., and Zhang, W. (2014). Risk Factors Contributing to Type 2 Diabetes and Recent Advances in the Treatment and Prevention. *Int. J. Med. Sci.* 11 (11), 1185–1200. doi:10.7150/ijms.10001

Yang, W., Lu, J., Weng, J., Jia, W., Ji, L., Xiao, J., et al. (2010). Prevalence of Diabetes Among Men and Women in China. *N. Engl. J. Med.* 362 (12), 1090–1101. doi:10.1056/NEJMoa0908292

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 07 February 2022
doi: 10.3389/fgene.2022.756901

Low Intelligence Predicts Higher Risks of Coronary Artery Disease and Myocardial Infarction: Evidence From Mendelian Randomization Study

Fangkun Yang[1†], *Teng Hu*[2†], *Songzan Chen*[3], *Kai Wang*[3], *Zihao Qu*[3] *and Hanbin Cui*[4]*

[1]*Department of Cardiology, Ningbo Hospital of Zhejiang University (Ningbo First Hospital), School of Medicine, Zhejiang University, Ningbo, China,* [2]*School of Medicine, Ningbo University, Ningbo, China,* [3]*School of Medicine, Zhejiang University, Hangzhou, China,* [4]*Cardiology Center, Ningbo First Hospital, Ningbo University, Ningbo, China*

OPEN ACCESS

Edited by:
Liangcai Zhang,
Janssen Research and Development,
United States

Reviewed by:
Tong Wang,
Chinese Academy of Sciences (CAS),
China
Alessandro Gialluisi,
Mediterranean Neurological Institute
Neuromed (IRCCS), Italy

****Correspondence:***
Hanbin Cui
hbcui_nbdyyy@outlook.com

†*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *11 August 2021*
Accepted: *10 January 2022*
Published: *07 February 2022*

Citation:
Yang F, Hu T, Chen S, Wang K, Qu Z and Cui H (2022) Low Intelligence Predicts Higher Risks of Coronary Artery Disease and Myocardial Infarction: Evidence From Mendelian Randomization Study.
Front. Genet. 13:756901.
doi: 10.3389/fgene.2022.756901

Background: Low intelligence has been shown to be associated with a high risk of cardiovascular disease in observational studies. It remains unclear whether the association is causal. This study aimed to explore the causal association of intelligence with coronary artery disease (CAD) and myocardial infarction (MI).

Methods: A two-sample Mendelian randomization study was designed to infer the causality. A total of 121 single nucleotide polymorphisms were selected as a genetic instrumental variable for intelligence. Summary data on CAD (n = 184,305) and MI (n = 171,875) were obtained from the Coronary ARtery DIsease Genome-wide Replication and Meta-analysis (CARDIoGRAM) plus The Coronary Artery Disease (C4D) Genetics (CARDIoGRAMplusC4D) consortium and the FinnGen study. Inverse variance weighting method was used to calculate the effect estimates. Sensitivity analyses including other statistical models and leave-one-out analysis were conducted to verify the robustness of results. MR-Egger test was performed to assess the pleiotropy.

Results: Genetically predicted higher intelligence was significantly associated with lower risk of CAD (OR, .76; 95%CI, .69–.85; $p = 1.5 \times 10^{-7}$) and MI (OR, .78; 95%CI, .70–.87; $p = 7.9 \times 10^{-6}$). The results remained consistent in the majority of the sensitivity analyses and were repeated in the FinnGen datasets. MR-Egger test suggested no evidence of directional pleiotropy for the association with coronary artery disease (intercept = −.01, p = .19) and myocardial infarction (intercept = −.01, p = .06).

Conclusion: This Mendelian randomization analysis provided genetic evidence for the causal association between low intelligence and increased risks of CAD and MI.

Keywords: intelligence, coronary artery disease, myocardial infarction, Mendelian randomization, causal association

INTRODUCTION

Cardiovascular diseases (CVD) have represented a major cause of death and disability in the past few decades (Joseph et al., 2017). The global number of deaths associated with CVD has increased by 12.5% during the past 10 years (GBD 2015 Maternal Mortality Collaborators, 2016). In Europe, CVD cause more than 4 million deaths each year, accounting for 45% of all deaths (Townsend et al., 2016). The burden of CVD remains a great challenge, though great efforts have been made to manage this disease (GBD 2013 Mortality and Causes of Death Collaborators, 2015).

In addition to diagnosis and treatment, prevention strategies for CVD are also indispensable (Goff et al., 2014). Many risk factors have been found independently associated with CVD, such as age, sex, hyperlipidemia, hypertension, diabetes, smoking, and family history (Chiu et al., 2018; Madhavan et al., 2018). The risk of incidence of CVD could be effectively reduced by intervening on several modifiable factors among them. There are also other newly discovered risk factors. Several observational studies demonstrated that low intelligence is associated with a high risk of CVD (Roberts et al., 2013; Dobson et al., 2017). However, it is unclear whether this association is causal or spurious.

Randomized controlled trials (RCTs) are the most reliable methods to explore the direct association between the exposures and the outcomes. However, these trials are difficult to carry out due to ethics or others. In recent years, Mendelian randomization (MR) studies, considered an analogy of RCTs, have been increasingly used to ascertain the cause of diseases (Bennett and Holmes, 2017). MR studies use genetic variations as instrumental variables for the exposures, randomly allocated at conception (Sekula et al., 2016; Emdin et al., 2017). Therefore, MR studies are less prone to environmental confounders. Moreover, reverse causality is avoided considering that alleles were always allocated before the onset of the diseases (Sekula et al., 2016; Emdin et al., 2017). A recent regression analysis and MR study investigated the association between intelligence and coronary artery disease (CAD) risk (Li et al., 2021). However, the MR part was simple and not rigorous enough in the selection of the instrumental variable, outcome dataset, and statistical methods. Moreover, the role of physical activity, alcohol use, sleep traits, and psychological factor needs to be further investigated.

This study aims to resort to the MR study to provide consistent evidence for the causal association of genetically determined intelligence with the risk of CAD and myocardial infarction (MI).

MATERIALS AND METHODS

Study Design

A two-sample MR study was designed to estimate the causal association between intelligence and the risk of CAD and MI. Three core assumptions for identifying the genetic instrumental variables are the basis of the MR analyses (Sekula et al., 2016; Emdin et al., 2017). First, the genetic instruments should be strongly associated with intelligence, generally at the genome-wide significant level ($p < 5 \times 10^{-8}$). Second, the instruments should be independent of the confounders. Third, the instruments should be only associated with the CAD and MI *via* intelligence.

Construction of the Genetic Instrumental Variable

The exposure was genetically predicted intelligence. Genetic associations with intelligence were taken from the largest meta-analysis of the genome-wide association study (GWAS) of intelligence to date (n = 269,867) (Savage et al., 2018). That meta-analysis included 14 independent cohorts of European ancestry, adjusted for age, sex, and ancestry principal components. Although intelligence was assessed using different neurocognitive tests in each cohort, the cognitive test scores remained robust in multiple populations (Savage et al., 2018). In that study, 242 lead single-nucleotide polymorphisms (SNPs) were identified as significantly associated with intelligence at a genome-wide significant level ($p < 5 \times 10^{-8}$). The SNPs were further quality-controlled based on a minor allele frequency >1%. For palindromic SNPs, if the minor allele frequency is smaller than .42, then this SNP was regarded as inferrable. Any palindromic SNPs with minor allele frequency larger than .42 were regarded as not inferrable and would be removed. The pairwise-linkage disequilibrium of SNPs was tested using LD-Link (https://ldlink.nci.nih.gov/) based on the European 1,000 Genomes Project reference panel ($r^2 < .001$ and clump distance >10,000 kb). If SNPs were in linkage disequilibrium, the SNP with a greater p-value would be removed (Machiela and Chanock, 2015; Myers et al., 2020). Then, those 157 SNPs were looked up in PhenoScanner 2.0 (a database of human genotype-phenotype associations) manually (Staley et al., 2016). The SNPs associated with other traits that may influence the results at a genome-wide significance level ($p < 5 \times 10^{-8}$) were further removed. We found that 25 SNPs were associated with body mass index, height, weight, or waist circumference and 11 SNPs were associated with cholesterol level, blood pressure, diabetes, alcohol intake, or smoking (**Supplementary Table 1**). After excluding these 36 SNPs, the remaining 121 SNPs were finally selected as the instrumental variable of intelligence.

Data Sources

The summary statistics for genetic associations with CAD and MI were acquired from Coronary ARtery DIsease Genome-wide Replication and Meta-analysis (CARDIoGRAM) plus The Coronary Artery Disease (C4D) Genetics (CARDIoGRAMplusC4D) consortium [n = 184,305, the majority (77%) were of European ancestry] (Nikpay et al., 2015). That study involved 60,801 CAD cases (~70% were MI sub-phenotype) and 123,504 controls. The participants were phenotyped based on clinical diagnosis and medical records. The data was publicly available in CARDIoGRAMplusC4D consortium (http://cardiogramplusc4d.org/). The replication datasets were from the FinnGen study, which was launched in Finland in 2017, including genome and health data from about 500,000 Finnish participants (https://www.finngen.fi/en). We used the fifth release of the results of genome-wide association analysis on CAD, including 21,012 cases and

197,780 controls. The genetic associations for MI included 12,801 cases and 187,840 controls. Genetic associations with smoking and alcohol use were obtained from the GWAS and Sequencing Consortium of Alcohol and Nicotine (GSCAN) use (Liu et al., 2019). Genetic associations with physical activity were acquired from a GWAS including about 90,000 individuals of European ancestry (Doherty et al., 2018). Genetic associations with sleep duration and insomnia were from the Sleep Disorder Knowledge Portal (Dashti et al., 2019; Jansen et al., 2019). Genetic associations with depression were obtained from Psychiatric Genomics Consortium (PGC) (Howard et al., 2019). Studies contributing data to the outcome datasets had already received ethical approval from relevant institutional review boards. In the present study, we only made use of the summarized data from these studies. Hence, no additional ethics approval was required.

Statistical Analyses

Two-sample MR analyses were used to estimate the causal associations of intelligence with the risk of CAD and MI. Specifically, we calculated the Wald ratio (quotient of the genetic association with outcome and the genetic association with the intelligence) and standard error for each SNP and then meta-analyzed them using the inverse variance weighting (IVW) method with fixed effect as our main MR effect estimates. In the sensitivity analyses, in order to test the robustness of the main results, the MR analyses with various statistical models, such as maximum likelihood, the IVW with multiplicative random effect (Bowden et al., 2017), penalized IVW, penalized robust IVW, simple median, weighted median (Bowden et al., 2016), and Mendelian Randomization Pleiotropy Residual Sum and Outlier (MR-PRESSO) (Verbanck et al., 2018), were conducted. The MR-Egger intercept test was used to assess the violation of the "no directional pleiotropy" assumption (Bowden et al., 2015). The visual inspection of scatter plots, funnel plots (Sterne et al., 2011), and leave-one-out plots were also performed to detect the potential horizontal pleiotropy (Bowden et al., 2017). Multivariable MR analysis was performed to investigate whether the association between intelligence and CAD/MI would be affected by potential confounders, including lifestyle factors [smoking (Liu et al., 2019), drinking (Liu et al., 2019), physical activity (Doherty et al., 2018), sleep duration (Dashti et al., 2019), insomnia (Jansen et al., 2019)], and psychological factor [depression (Howard et al., 2019)] (Burgess and Thompson, 2015; Sanderson et al., 2019). Specifically, we obtained summary-level data of the intelligence-related SNPs with confounding factors from corresponding genetic consortia. Then, the data was combined with the genetic associations between intelligence and outcomes for each SNP. The multivariable MR analysis allowed the genetic variants to be associated with all the risk factors in the statistical model (Burgess and Thompson, 2015). Causal estimates reflecting direct causal effects of the primary risk factor were provided, adjusted for the influence of a secondary risk factor or mediator. For power calculation, we used an online tool named mRnd (https://shiny.cnsgenomics.com/mRnd/) based on sample size, type-I error rate, proportion of cases, odds ratio of outcome per standard deviation of exposure, and proportion of variance explained by the included SNPs (Freeman et al., 2013). All the analyses needed to achieve the statistical power of at least 80%. A two-sided *p*-value of <.025 (=.05/2 outcomes) was defined as statistically significant. All the statistical analyses in the current study were implemented by the R software (version 3.6.3) together with the R package "MendelianRandomization" (https://github.com/cran/MendelianRandomization) and "MR-PRESSO" (https://github.com/rondolab/MR-PRESSO) (Yavorska and Burgess, 2017; Verbanck et al., 2018). In the MR analyses using the IVW method, we chose the fixed-effect, random-effect, penalized, or robust model. And for other analyses, default settings were used.

RESULTS

After excluding SNPs that might violate the three core assumptions, 121 SNPs were identified as a genetic instrument in our main analysis. The characteristics of these SNPs and their genetic associations with the intelligence and the outcome are shown in **Supplementary Table 2**.

The scatter plots of the associations between the genetically predicted intelligence and CAD/MI are displayed in **Figure 1**. The associations between the genetically predicted intelligence and CAD/MI are shown in **Figure 2**. The fixed-effect IVW method showed that higher genetically predicted intelligence was significantly associated with lower risks of CAD (OR .76 per SD increase; 95%CI, .69–.85; $p = 1.5 \times 10^{-7}$) and MI (OR .78 per SD increase; 95%CI, .70–.87; $p = 7.9 \times 10^{-6}$). Similar results were observed using the maximum likelihood, the multiplicative random effect IVW, penalized IVW, penalized robust IVW, simple median, weighted median (for CAD), and MR-PRESSO method. However, the associations were not evident using the weighted median (for MI) and MR-Egger test. The main results were repeated based on genetic data for CAD and MI from the FinnGen study, indicating the robustness and consistency of the main results (**Table 1**; **Supplementary Table 3**).

The MR-Egger intercept test is shown in **Table 2**, which did not provide strong evidence of potential directional pleiotropy for the associations between the genetically predicted intelligence and CAD (intercept = −.01, p = .19) and MI (intercept = −.01, p = .06). Funnel plots were symmetric distribution, indicating no obvious potential pleiotropic effects (**Supplementary Figure 1**). The results of the leave-one-out analysis suggested that the associations between the genetically predicted intelligence and CAD/MI were stable and not drastically driven by individual SNP (**Figure 3**; **Supplementary Figure 2**). The association pattern remained after adjusting for most of the potential confounding traits (**Table 3**). The MR estimates were slightly attenuated after adjusting for smoking and sleep duration. However, limited evidence was found for the mediating effect of smoking and sleep duration between intelligence and CAD/MI. The MR analyses have 98% and 90% statistical power at the type I error rate of .05 for association with CAD and MI, respectively (**Supplementary Table 4**).

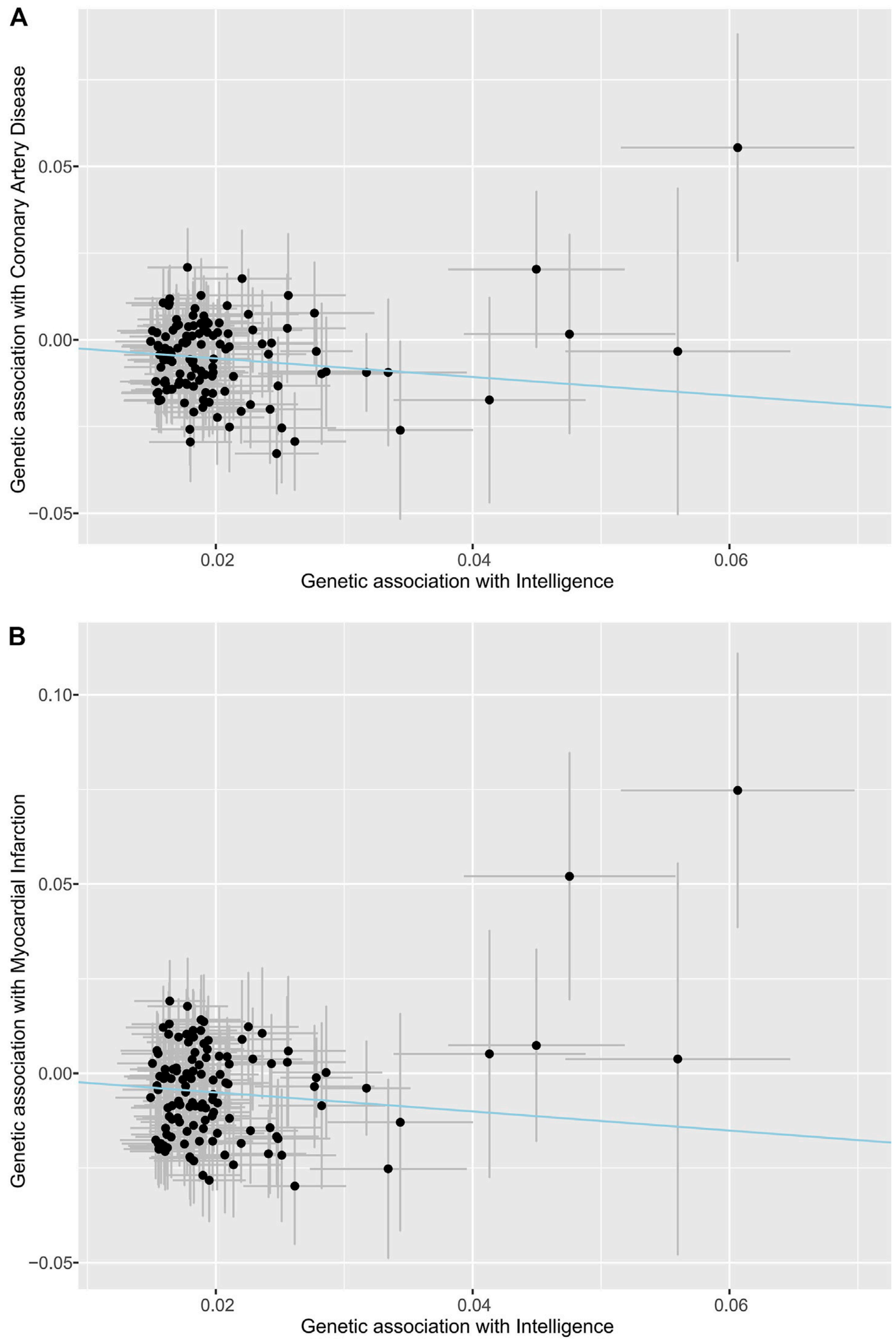


FIGURE 1 | Associations of intelligence-related variants with outcomes. **(A)** Coronary artery disease. **(B)** Myocardial infarction. The dots indicate the causal effect of each SNP. The bars indicate the 95% confidence intervals. The blue line indicates the estimate of effect using the inverse-variance weighted method.

DISCUSSION

We conducted a two-sample MR study to explore the causal effects of intelligence on CAD and MI. We found that the higher genetically predicted intelligence was significantly associated with the lower risk of CAD and MI. In addition, the results remained consistent in the majority of the sensitivity analyses with different statistical models and leave-one-out analyses.

CVD represents a leading cause of illness and disability associated with high morbidity and mortality (Benjamin et al., 2018). Except for the well-established risk factors, such as age, sex, and hypertension, intelligence has been newly discovered as an intriguing risk factor

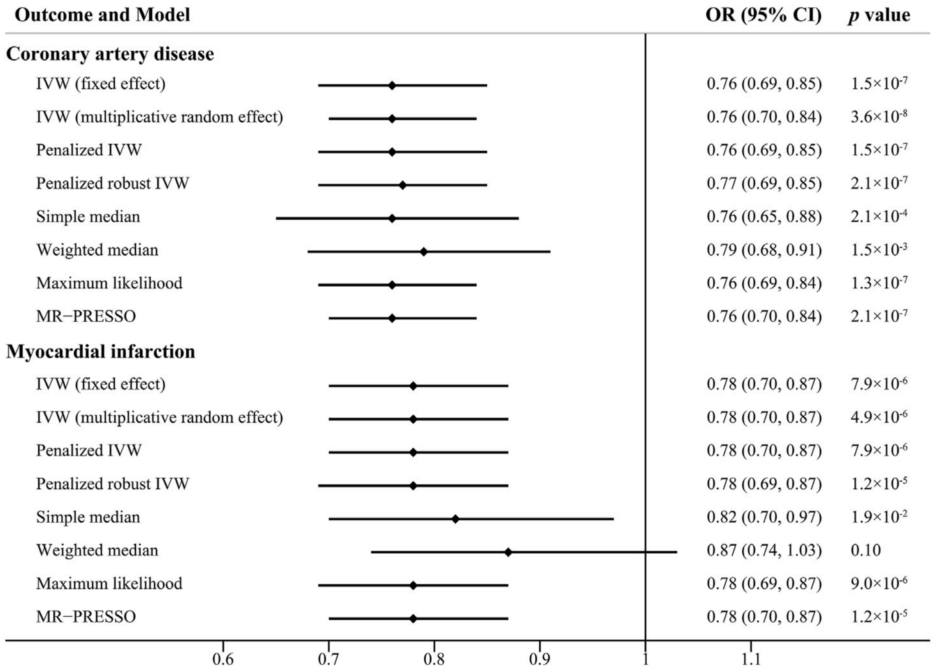


FIGURE 2 | Causal effect estimates of genetically predicted intelligence on coronary artery disease and myocardial infarction using different statistical models. OR, odds ratio; CI, confidence interval.

TABLE 1 | The associations between intelligence and coronary artery disease/myocardial infarction using genetic data from the FinnGen study.

Outcome	Statistical model	OR	95% CI	*p*-value
CAD	IVW (random effects)	.82	(.70, .95)	6.7E-03
	IVW (fixed effects)	.82	(.71, .94)	4.6E-03
	Weighted median	.91	(.74, 1.12)	.36
	MR-Egger	1.47	(.71, 3.03)	.30
	Maximum likelihood	.81	(.70, .94)	4.5E-03
	MR-PRESSO	.82	(.70, .95)	7.7E-03
MI	IVW (random effects)	.76	(.64, .89)	9.2E-04
	IVW (fixed effects)	.76	(.64, .90)	1.2E-03
	Weighted median	.80	(.63, 1.01)	6.3E-02
	MR-Egger	1.06	(.46, 2.46)	.89
	Maximum likelihood	.75	(.63, .89)	1.2E-03
	MR-PRESSO	.76	(.64, .89)	1.2E-03

CAD, coronary artery disease; MI, myocardial infarction; IVW, inverse-variance weighted; MR-PRESSO, mendelian randomization pleiotropy residual sum and outlier; OR, odds ratio; CI, confidence interval.

(Dobson et al., 2017). In the past few decades, observational epidemiological studies have accumulated evidence for an inverse association between intelligence and the risk of CVD. A prospective cohort study in Scotland with 938 participants and a 25-year follow-up showed that childhood intelligence quotient (IQ) was significantly inversely related to CVD events in individuals aged up to 65 (Hart et al., 2004). The Newcastle Thousand Families study with 412 members and a 40-year follow-up suggested that individuals with higher childhood intelligence had a lower risk of atherosclerosis in middle age (Roberts et al., 2013). In a meta-analysis of five longitudinal studies with 17,256 participants, each standard deviation decrease in childhood IQ was associated with an increase of 16% in the risk of CVD (Dobson et al., 2017). Moreover, a prospective cohort study of 49,321 Swedish males and another cohort study of 4,316 Vietnam males demonstrated that lower IQ scores in early adulthood were associated with an increased risk of coronary heart disease (CHD) and acute myocardial infarction (AMI) (Hemmingsson et al., 2007; Batty et al., 2008). However, these studies fail to distinguish between the causal and spurious associations because of the unmeasured confounding and reverse causality. The present study can largely overcome these shortcomings and provide a reliable causal inference. Our study, together with previous evidence, suggested that intelligence was causally associated with the risks of CAD and MI.

TABLE 2 | MR-Egger tests of intelligence with CAD and MI.

Outcome	MR-Egger	Estimate	LCI	UCI	*p*-value
CAD	Slope	.06	−.45	.56	.82
	Intercept	−.01	−.02	.0033	.20
MI	Slope	.27	−.29	.83	.34
	Intercept	−.01	−.02	.0005	.06

LCI, lower confidence interval; UCI, upper confidence interval; CAD, coronary artery disease; MI, myocardial infarction.

Though the associations between the low premorbid intelligence and the increased risk of CVD and the high rate of later mortality have been explored, the exact mechanism remains unclear. Several plausible hypotheses have been proposed. First, socioeconomic factor was put forward to explain the association between intelligence and

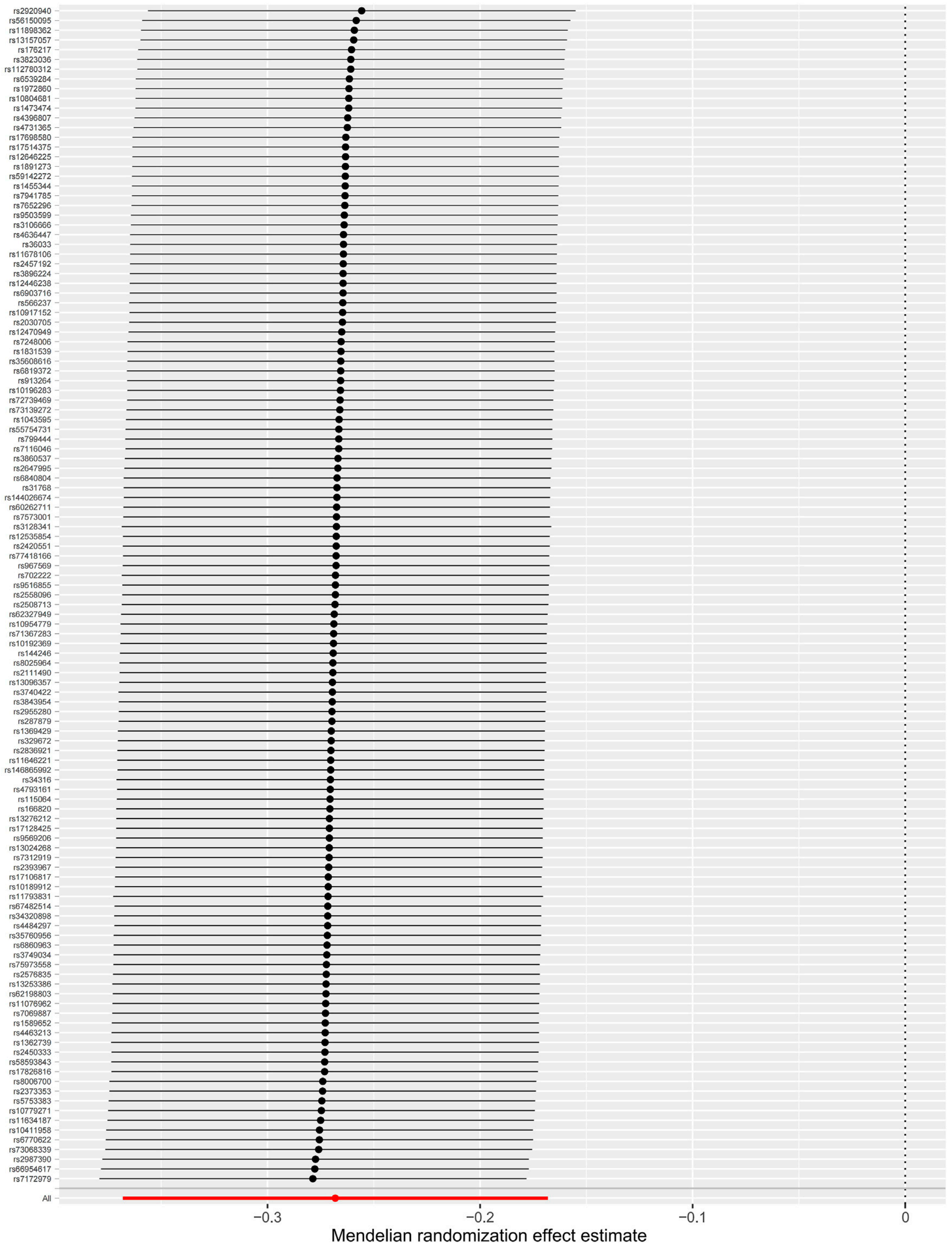


FIGURE 3 | Leave-one-out analyses of the associations between intelligence and coronary artery disease. The dots indicate the causal effect using the inverse-variance weighted method when the SNP is removed. The bars indicate a 95% confidence interval.

CAD risk. Individuals with low intelligence were less prone to educational success and well-remunerated employment, which provided protection against CAD (Batty et al., 2009). Second, the effect of intelligence could be mediated *via* health literacy. Individuals with low intelligence were less likely aware of their health conditions and even had more difficulties understanding health messages (Dobson et al., 2017). They rarely knew how to prevent the diseases or take medicine properly. It was demonstrated that

TABLE 3 | Multivariable Mendelian randomization analyses of genetically determined intelligence and the risk of coronary artery disease and myocardial infarction adjusting for potential confounding traits.

Outcome	Adjusting for	OR	95% CI	*p*-value
CAD	Smoking	.76	(.69, .85)	5.5E-07
	Alcohol use	.76	(.69, .84)	1.3E-07
	Physical activity	.76	(.68, .84)	9.4E-08
	Insomnia	.76	(.69, .84)	1.3E-07
	Sleep duration	.77	(.70, .85)	4.9E-07
	Depression	.77	(.69, .85)	2.4E-07
MI	Smoking	.79	(.70, .88)	5.4E-05
	Alcohol use	.77	(.69, .87)	7.6E-06
	Physical activity	.77	(.69, .86)	7.6E-06
	Insomnia	.77	(.69, .87)	7.2E-06
	Sleep duration	.78	(.70, .87)	1.4E-05
	Depression	.78	(.70, .87)	1.4E-05

CAD, coronary artery disease; MI, myocardial infarction; OR, odds ratio; CI, confidence interval.

higher intelligence was associated with improved disease prevention or better health behaviors, including quitting smoking, having a prudent diet, and persisting with moderate physical activity (Sörberg Wallin et al., 2015). However, the present MR study found limited evidence for the mediating effect of smoking, physical activity, and sleep duration between intelligence and CAD/MI. Third, these associations could also be partly explained by the congenital or childhood health damage in both intelligence and physiological functions, which eventually increased CAD risk in later life (Dobson et al., 2017). Future investigations were warranted to elucidate the exact mechanism by which the low intelligence was causally associated with the increased CAD risk.

The strength of this study is the design of the two-sample MR study. MR analysis is a novel technique that uses genetic variants as instrumental variables to estimate the causal effect of exposure on the outcomes (Sekula et al., 2016). The genetic variants are not associated with other confounding factors because of the random allocation at the conception, greatly reducing the potential bias (Sekula et al., 2016; Emdin et al., 2017). MR analysis can also avoid reverse causation because genotyping is always earlier than phenotyping (Sekula et al., 2016; Emdin et al., 2017). MR analysis represents a reliable method to infer the causal associations, even described as the best alternative to RCTs (Nitsch et al., 2006). In addition, we investigated the causal association between intelligence and CAD and MI based on a large-scale cohort, which could improve the effectiveness of the statistical test.

There are several limitations to our study. First, the potential pleiotropy cannot be completely ruled out, which may lead to biased causal estimates. However, the MR-Egger intercept test suggested no potential directional pleiotropy, and MR-PRESSO found no evidence of horizontal pleiotropic outliers. The result was almost consistent in the sensitivity analysis except for the MR-Egger method. On the one hand, it could be explained by a potential violation of the Instrument Strength Independent of Direct Effect (InSIDE) assumption. This assumption was unlikely completely satisfied, to which the MR-Egger method was sensitive. On the other hand, MR-Egger regression was very conservative compared to other methods, especially when no violation of the horizontal pleiotropy assumption was evident. Second, we did not explore the associations between the genetic instrumental variables and the observed confounders, such as body mass index and cholesterol level. Nevertheless, we had excluded the SNPs related to potential confounders by looking up the SNPs in PhenoScanner. Third, the phenotype of the intelligence varied among the 14 independent cohorts included in the GWAS, but the test scores remained robust in multiple populations. This result needed to be verified when a uniform and precise phenotype of the intelligence was available in GWAS studies. Moreover, the results of the current study were based on samples from individuals of European ancestry, and the effect of low intelligence on the risk of CAD/MI needed to be further investigated in other racial and ethnic groups. Finally, we only revealed the causal association between intelligence and CAD and MI from a genetic perspective without involving other environmental factors.

CONCLUSION

Our two-sample MR study provides genetic evidence for the causal association between low intelligence and the increased risk of CAD and MI. Early recognition coupled with appropriate care of individuals with low intelligence may have significant clinical and public health implications. Further studies are warranted to verify our findings and reveal the potential mechanism.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

Ethical review and approval were not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

FY, SC, and HC designed the study and wrote the analysis plan. FY and ZQ undertook analyses. FY wrote the first draft of the manuscript with critical revisions from TH, KW, and HC. All authors interpreted the study results and gave final approval of the version to be published.

FUNDING

This work was supported by grants from the Zhejiang Provincial Key Research Project (Grant no. 2021C03096 to HC) and Major

Project of Science and Technology Innovation 2025 in Ningbo, China (Grant no. 2021Z134).

ACKNOWLEDGMENTS

The authors thank all the investigators of the FinnGen study, GSCAN Consortium, Sleep Disorder Knowledge Portal, and Psychiatric Genomics Consortium for providing the data publicly.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.756901/full#supplementary-material

REFERENCES

Batty, G. D., Shipley, M. J., Mortensen, L. H., Gale, C. R., and Deary, I. J. (2008). IQ in Late Adolescence/early Adulthood, Risk Factors in Middle-Age and Later Coronary Heart Disease Mortality in Men: the Vietnam Experience Study. *Eur. J. Cardiovasc. Prev. Rehabil.* 15, 359–361. doi:10.1097/HJR.0b013e3282f738a6

Batty, G. D., Wennerstad, K. M., Smith, G. D., Gunnell, D., Deary, I. J., Tynelius, P., et al. (2009). IQ in Early Adulthood and Mortality by Middle Age: Cohort Study of 1 Million Swedish Men. *Epidemiology* 20, 100–109. doi:10.1097/EDE.0b013e31818ba076

Benjamin, E. J., Virani, S. S., Callaway, C. W., Chamberlain, A. M., Chang, A. R., Cheng, S., et al. (2018). Heart Disease and Stroke Statistics-2018 Update: A Report from the American Heart Association. *Circulation* 137, e67–e492. doi:10.1161/CIR.0000000000000558

Bennett, D. A., and Holmes, M. V. (2017). Mendelian Randomisation in Cardiovascular Research: an Introduction for Clinicians. *Heart* 103, 1400–1407. doi:10.1136/heartjnl-2016-310605

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian Randomization with Invalid Instruments: Effect Estimation and Bias Detection through Egger Regression. *Int. J. Epidemiol.* 44, 512–525. doi:10.1093/ije/dyv080

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* 40, 304–314. doi:10.1002/gepi.21965

Bowden, J., Del Greco M, F., Minelli, C., Davey Smith, G., Sheehan, N., and Thompson, J. (2017). A Framework for the Investigation of Pleiotropy in Two-Sample Summary Data Mendelian Randomization. *Statist. Med.* 36, 1783–1802. doi:10.1002/sim.7221

Burgess, S., and Thompson, S. G. (2015). Multivariable Mendelian Randomization: the Use of Pleiotropic Genetic Variants to Estimate Causal Effects. *Am. J. Epidemiol.* 181, 251–260. doi:10.1093/aje/kwu283

Chiu, M. H., Heydari, B., Batulan, Z., Maarouf, N., Subramanya, V., Schenck-Gustafsson, K., et al. (2018). Coronary Artery Disease in post-menopausal Women: Are There Appropriate Means of Assessment? *Clin. Sci. (Lond)* 132, 1937–1952. doi:10.1042/CS20180067

Dashti, H. S., Jones, S. E., Wood, A. R., Lane, J. M., van Hees, V. T., Wang, H., et al. (2019). Genome-wide Association Study Identifies Genetic Loci for Self-Reported Habitual Sleep Duration Supported by Accelerometer-Derived Estimates. *Nat. Commun.* 10, 1100. doi:10.1038/s41467-019-08917-4

Dobson, K. G., Chow, C. H. T., Morrison, K. M., and Van Lieshout, R. J. (2017). Associations between Childhood Cognition and Cardiovascular Events in Adulthood: A Systematic Review and Meta-Analysis. *Can. J. Cardiol.* 33, 232–242. doi:10.1016/j.cjca.2016.08.014

Doherty, A., Smith-Byrne, K., Ferreira, T., Holmes, M. V., Holmes, C., Pulit, S. L., et al. (2018). GWAS Identifies 14 Loci for Device-Measured Physical Activity and Sleep Duration. *Nat. Commun.* 9, 5257. doi:10.1038/s41467-018-07743-4

Emdin, C. A., Khera, A. V., and Kathiresan, S. (2017). Mendelian Randomization. *JAMA* 318, 1925–1926. doi:10.1001/jama.2017.17219

Freeman, G., Cowling, B. J., and Schooling, C. M. (2013). Power and Sample Size Calculations for Mendelian Randomization Studies Using One Genetic Instrument. *Int. J. Epidemiol.* 42, 1157–1163. doi:10.1093/ije/dyt110

GBD 2013 Mortality and Causes of Death Collaborators (2015). Global, Regional, and National Age-Sex Specific All-Cause and Cause-specific Mortality for 240 Causes of Death, 1990-2013: a Systematic Analysis for the Global Burden of Disease Study 2013. *Lancet* 385, 117–171. doi:10.1016/S0140-6736(14)61682-2

GBD 2015 Maternal Mortality Collaborators (2016). Global, Regional, and National Levels of Maternal Mortality, 1990-2015: a Systematic Analysis for the Global Burden of Disease Study 2015. *Lancet* 388, 1775–1812. doi:10.1016/S0140-6736(16)31470-2

Goff, D. C., Lloyd-Jones, D. M., Bennett, G., Coady, S., D'Agostino, R. B., Gibbons, R., et al. (2014). 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk. *Circulation* 129, S49–S73. doi:10.1161/01.cir.0000437741.48606.98

Hart, C. L., Taylor, M. D., Smith, G. D., Whalley, L. J., Starr, J. M., Hole, D. J., et al. (2004). Childhood IQ and Cardiovascular Disease in Adulthood: Prospective Observational Study Linking the Scottish Mental Survey 1932 and the Midspan Studies. *Soc. Sci. Med.* 59, 2131–2138. doi:10.1016/j.socscimed.2004.03.016

Hemmingsson, T., Essen, J. v., Melin, B., Allebeck, P., and Lundberg, I. (2007). The Association between Cognitive Ability Measured at Ages 18-20 and Coronary Heart Disease in Middle Age Among Men: a Prospective Study Using the Swedish 1969 Conscription Cohort. *Soc. Sci. Med.* 65, 1410–1419. doi:10.1016/j.socscimed.2007.05.006

Howard, D. M., Adams, M. J., Adams, M. J., Clarke, T.-K., Hafferty, J. D., Gibson, J., et al. (2019). Genome-wide Meta-Analysis of Depression Identifies 102 Independent Variants and Highlights the Importance of the Prefrontal Brain Regions. *Nat. Neurosci.* 22, 343–352. doi:10.1038/s41593-018-0326-7

Jansen, P. R., Watanabe, K., Watanabe, K., Stringer, S., Skene, N., Bryois, J., et al. (2019). Genome-wide Analysis of Insomnia in 1,331,010 Individuals Identifies New Risk Loci and Functional Pathways. *Nat. Genet.* 51, 394–403. doi:10.1038/s41588-018-0333-3

Joseph, P., Leong, D., McKee, M., Anand, S. S., Schwalm, J.-D., Teo, K., et al. (2017). Reducing the Global Burden of Cardiovascular Disease, Part 1: The Epidemiology and Risk Factors. *Circ. Res.* 121, 677–694. doi:10.1161/CIRCRESAHA.117.308903

Li, L., Pang, S., Zeng, L., Güldener, U., and Schunkert, H. (2021). Genetically Determined Intelligence and Coronary Artery Disease Risk. *Clin. Res. Cardiol.* 110, 211–219. doi:10.1007/s00392-020-01721-x

Liu, M., Jiang, Y., Jiang, Y., Wedow, R., Li, Y., Brazel, D. M., et al. (2019). Association Studies of up to 1.2 Million Individuals Yield New Insights into the Genetic Etiology of Tobacco and Alcohol Use. *Nat. Genet.* 51, 237–244. doi:10.1038/s41588-018-0307-5

Machiela, M. J., and Chanock, S. J. (2015). LDlink: a Web-Based Application for Exploring Population-specific Haplotype Structure and Linking Correlated Alleles of Possible Functional Variants. *Bioinformatics* 31, 3555–3557. doi:10.1093/bioinformatics/btv402

Madhavan, M. V., Gersh, B. J., Alexander, K. P., Granger, C. B., and Stone, G. W. (2018). Coronary Artery Disease in Patients ≥80 Years of Age. *J. Am. Coll. Cardiol.* 71, 2015–2040. doi:10.1016/j.jacc.2017.12.068

Myers, T. A., Chanock, S. J., and Machiela, M. J. (2020). LDlinkR: An R Package for Rapidly Calculating Linkage Disequilibrium Statistics in Diverse Populations. *Front. Genet.* 11, 157. doi:10.3389/fgene.2020.00157

Nikpay, M., Goel, A., Won, H-H., Hall, L. M., Willenborg, C., Kanoni, S., et al. (2015). A Comprehensive 1000 Genomes-Based Genome-wide Association Meta-Analysis of Coronary Artery Disease. *Nat. Genet.* 47, 1121–1130. doi:10.1038/ng.3396

Nitsch, D., Molokhia, M., Smeeth, L., DeStavola, B. L., Whittaker, J. C., and Leon, D. A. (2006). Limits to Causal Inference Based on Mendelian Randomization: a Comparison with Randomized Controlled Trials. *Am. J. Epidemiol.* 163, 397–403. doi:10.1093/aje/kwj062

Roberts, B. A., Batty, G. D., Gale, C. R., Deary, I. J., Parker, L., and Pearce, M. S. (2013). IQ in Childhood and Atherosclerosis in Middle-Age: 40 Year Follow-

Up of the Newcastle Thousand Families Cohort Study. *Atherosclerosis* 231, 234–237. doi:10.1016/j.atherosclerosis.2013.09.018

Sanderson, E., Davey Smith, G., Windmeijer, F., and Bowden, J. (2019). An Examination of Multivariable Mendelian Randomization in the Single-Sample and Two-Sample Summary Data Settings. *Int. J. Epidemiol.* 48, 713–727. doi:10.1093/ije/dyy262

Savage, J. E., Jansen, P. R., Stringer, S., Watanabe, K., Bryois, J., de Leeuw, C. A., et al. (2018). Genome-wide Association Meta-Analysis in 269,867 Individuals Identifies New Genetic and Functional Links to Intelligence. *Nat. Genet.* 50, 912–919. doi:10.1038/s41588-018-0152-6

Sekula, P., Del Greco M, F., Pattaro, C., and Köttgen, A. (2016). Mendelian Randomization as an Approach to Assess Causality Using Observational Data. *Jasn* 27, 3253–3265. doi:10.1681/ASN.2016010098

Sörberg Wallin, A., Falkstedt, D., Allebeck, P., Melin, B., Janszky, I., and Hemmingsson, T. (2015). Does High Intelligence Improve Prognosis? the Association of Intelligence with Recurrence and Mortality Among Swedish Men with Coronary Heart Disease. *J. Epidemiol. Community Health* 69, 347–353. doi:10.1136/jech-2014-204958

Staley, J. R., Blackshaw, J., Kamat, M. A., Ellis, S., Surendran, P., Sun, B. B., et al. (2016). PhenoScanner: a Database of Human Genotype-Phenotype Associations. *Bioinformatics* 32, 3207–3209. doi:10.1093/bioinformatics/btw373

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., et al. (2011). Recommendations for Examining and Interpreting Funnel Plot Asymmetry in Meta-Analyses of Randomised Controlled Trials. *BMJ* 343, d4002. doi:10.1136/bmj.d4002

Townsend, N., Wilson, L., Bhatnagar, P., Wickramasinghe, K., Rayner, M., and Nichols, M. (2016). Cardiovascular Disease in Europe: Epidemiological Update 2016. *Eur. Heart J.* 37, 3232–3245. doi:10.1093/eurheartj/ehw334

Verbanck, M., Chen, C.-Y., Neale, B., and Do, R. (2018). Detection of Widespread Horizontal Pleiotropy in Causal Relationships Inferred from Mendelian Randomization between Complex Traits and Diseases. *Nat. Genet.* 50, 693–698. doi:10.1038/s41588-018-0099-7

Yavorska, O. O., and Burgess, S. (2017). MendelianRandomization: an R Package for Performing Mendelian Randomization Analyses Using Summarized Data. *Int. J. Epidemiol.* 46, 1734–1739. doi:10.1093/ije/dyx034

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 07 February 2022
doi: 10.3389/fgene.2021.782953

Genetic Influence Underlying Brain Connectivity Phenotype: A Study on Two Age-Specific Cohorts

Shan Cong[1†], *Xiaohui Yao*[1†], *Linhui Xie*[2], *Jingwen Yan*[3] *and Li Shen*[1*]
and the Alzheimer's Disease Neuroimaging Initiative

[1] *Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States,* [2] *Department of Electrical and Computer Engineering, School of Engineering, Indiana University Purdue University Indianapolis, Indianapolis, IN, United States,* [3] *Department of BioHealth Informatics, School of Informatics and Computing, Indiana University Purdue University Indianapolis, Indianapolis, IN, United States*

Edited by:
Guiyou Liu,
Tianjin Institute of Industrial Biotechnology (CAS), China

Reviewed by:
Zhaoxia Yu,
University of California, Irvine, United States
Tomas Drgon,
United States Food and Drug Administration, United States

Correspondence:
Li Shen
li.shen@pennmedicine.upenn.edu

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *25 September 2021*
Accepted: *16 November 2021*
Published: *07 February 2022*

Citation:
Cong S, Yao X, Xie L, Yan J and Shen L and the Alzheimer's Disease Neuroimaging Initiative (2022) Genetic Influence Underlying Brain Connectivity Phenotype: A Study on Two Age-Specific Cohorts. Front. Genet. 12:782953. doi: 10.3389/fgene.2021.782953

Background: Human brain structural connectivity is an important imaging quantitative trait for brain development and aging. Mapping the network connectivity to the phenotypic variation provides fundamental insights in understanding the relationship between detailed brain topological architecture, function, and dysfunction. However, the underlying neurobiological mechanism from gene to brain connectome, and to phenotypic outcomes, and whether this mechanism changes over time, remain unclear.

Methods: This study analyzes diffusion-weighted imaging data from two age-specific neuroimaging cohorts, extracts structural connectome topological network measures, performs genome-wide association studies of the measures, and examines the causality of genetic influences on phenotypic outcomes mediated via connectivity measures.

Results: Our empirical study has yielded several significant findings: 1) It identified genetic makeup underlying structural connectivity changes in the human brain connectome for both age groups. Specifically, it revealed a novel association between the minor allele (G) of rs7937515 and the decreased network segregation measures of the left middle temporal gyrus across young and elderly adults, indicating a consistent genetic effect on brain connectivity across the lifespan. 2) It revealed rs7937515 as a genetic marker for body mass index in young adults but not in elderly adults. 3) It discovered brain network segregation alterations as a potential neuroimaging biomarker for obesity. 4) It demonstrated the hemispheric asymmetry of structural network organization in genetic association analyses and outcome-relevant studies.

Discussion: These imaging genetic findings underlying brain connectome warrant further investigation for exploring their potential influences on brain-related complex diseases, given the significant involvement of altered connectivity in neurological, psychiatric and physical disorders.

Keywords: causal inference, body mass index, genome-wide association study, human connectomics, network segregation

1 INTRODUCTION

Brain structural connectivity is a major organizing principle of the nervous system. Estimating interregional neural connectivity, reconstructing geometric structure of fiber pathways, and mapping the network connectivity to corresponding inter-individual variabilities provide fundamental insights in understanding detailed brain topological architecture, function and dysfunction. A large body of research has been devoted to extracting and investigating macro-scale brain networks from diffusion-weighted imaging (DWI) data (Xie et al., 2018; Jiang et al., 2019; van den Heuvel et al., 2019; Bertolero et al., 2019; Elsheikh et al., 2020), and various behavioral, neurological and neuropsychiatric disorders have been linked to the disrupted brain connectivity (Jiang et al., 2019; van den Heuvel et al., 2019). As structural changes of brain connectivity are phenotypically associated with massive complex traits across different categories, the brain-wide connectome has been extensively studied.

It is worth noting that human brain connectome re-configures its network structure dynamically and adaptively in response to genetic, lifestyle, environmental factors (Cohen and D'Esposito, 2016; Cauda et al., 2018), brain development and aging (Sala-Llonch et al., 2015; Alloza et al., 2018; Varangis et al., 2019). However, the underlying neurobiological mechanism from gene to brain connectome, and to cognitive and behavioral outcomes, and whether this mechanism changes over time, remain unclear. To bridge this gap, we perform a genetic study of brain connectome phenotypes on two different age-specific cohorts: one contains healthy young adults (age: 28.7 ± 3.6), and the other contains elderly participants (age: 73.8 ± 7.0). Our goal is to identify genetic factors affecting brain connectivity and examine their consistency and discrepancy between these two age-specific groups.

Emerging advances in multimodal brain imaging, high throughput genotyping and sequencing techniques provide exciting new opportunities to ultimately improve our understanding of brain structure and neural dynamics, their genetic architecture and their influences on cognition and behavior (Shen and Thompson, 2020). Present studies investigating direct associations among human connectomics, genomics and clinical phenotyping are primarily focused on four aspects: 1) estimating genetic heritability of basic connectome measures such as number of fibers, length of fibers and fractional anisotropy (FA) (Jahanshad et al., 2013; Thompson et al., 2013; Elliott et al., 2018); 2) discovering pairwise univariate associations between single nucleotide polymorphisms (SNPs) and imaging phenotypic traits such as above mentioned basic connectome measures at each edge (Jahanshad et al., 2013; Karwowski et al., 2019) and white matter properties at each voxel (Kochunov et al., 2010; Alloza et al., 2018; Guo et al., 2020); 3) discovering pairwise univariate associations between SNPs and clinical phenotypes such as cognitive or behavioral outcomes (Jahanshad et al., 2013; Elsheikh et al., 2020); and 4) discovering pairwise univariate associations between basic connectome measures and clinical phenotypes (Jiang et al., 2019; van den Heuvel et al., 2019).

Among the studies mentioned above, there exist two major limitations. First, these studies were conducted based on basic connectome measures such as number of fibers, length of fibers and FA, but the complex-network attributes were overlooked, which included network segregation, integration, centrality and resilience and important network components such as hubs, communities, and rich clubs (Sporns, 2013). These attributes were extensively adopted to detect network integration and segregation, quantitatively measure the centrality of network regions and pathways, characterize patterns of local anatomical circuitry, and test resilience of networks to insult (Rubinov and Sporns, 2010). Second, these studies performed analyses by examining the association between an independent variable (e.g., SNP) and a dependent variable (e.g., cognitive or behavioral outcome), without taking into consideration the mediator(s) linking these variables (Baron and Kenny, 1986). Mediation analysis can help identify the underlying mechanism of outcome-relevant genetic effects implicitly mediated by neuroimaging phenotypes (e.g., connectome measures). Of note, mediation analysis requires the independent variable to be significantly associated with both the dependent variable and the mediator. This makes applying it in brain neuroimaging studies a challenge due to the modest effect size of an individual genetic variant on both behavioral and imaging phenotypes (Saykin et al., 2015; Cong et al., 2018), as well as limited size of the sample with all diagnostic, imaging and genetic data available.

With the demand of measuring complex-network attributes, a few recent genome-wide association studies (GWAS) (Bertolero et al., 2019; Elsheikh et al., 2020) recognized the first problem mentioned above and adopted quantitative measurement approaches for complex-network attributes, and treated the attributes as neuroimaging traits for the explorations of complex imaging genomic associations. They successfully identified a number of loci susceptible for Alzheimer's disease (Elsheikh et al., 2020), and demonstrated the associations between loci and segregated network patterns, which may be involved in brain development, evolution, and disease (Bertolero et al., 2019). However, a notable limitation is that these studies only focus on the brain networks of either young or elderly participants, as a result, their study outcomes are lack of validations in multiple data sets. Since there is an age-related discrepancy for genetic effects on human connectome alterations across lifespan (Varangis et al., 2019), it remains an under-explored topic to examine genetic consistency and discrepancy for complex-network attributes among cohorts different in age. Another factor that may cause discrepancy in the network architecture is the hemispheric asymmetry (Jiang et al., 2019), and the hemispheric asymmetry of network organization has been linked to development processes (Zhong et al., 2017) and neuropsychiatric disorders (Sun et al., 2017). It remains a challenge to understand the genetic basis for the network attributes of two hemispheres as they may be distinctively correlated to cognition level, physical and psychological development.

Among a large number of complex-network attributes, it has been well documented in recent literatures (Cohen and D'Esposito, 2016; Xie et al., 2018) that segregation of neural information such as modularity, transitivity, clustering

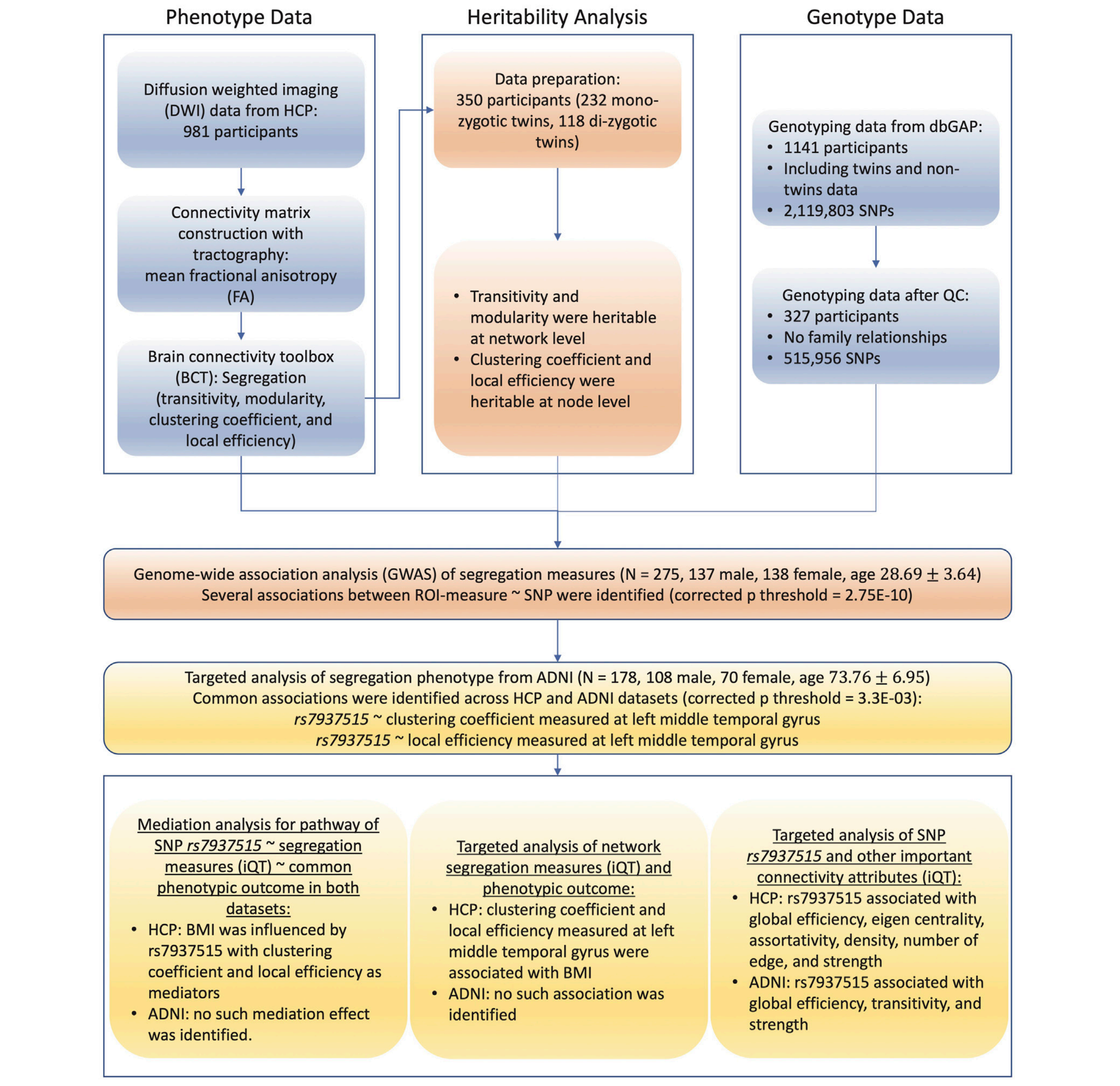

Phenotype Data
Heritability Analysis
Genotype Data
Diffusion weighted imaging (DWI) data from HCP: 981 participants
Connectivity matrix construction with tractography: mean fractional anisotropy (FA)
Brain connectivity toolbox (BCT): Segregation (transitivity, modularity, clustering coefficient, and local efficiency)
Data preparation: 350 participants (232 mono-zygotic twins, 118 di-zygotic twins)
• Transitivity and modularity were heritable at network level
• Clustering coefficient and local efficiency were heritable at node level
Genotyping data from dbGAP:
• 1141 participants
• Including twins and non-twins data
• 2,119,803 SNPs
Genotyping data after QC:
• 327 participants
• No family relationships
• 515,956 SNPs
Genome-wide association analysis (GWAS) of segregation measures (N = 275, 137 male, 138 female, age 28.69 ± 3.64)
Several associations between ROI-measure ~ SNP were identified (corrected p threshold = 2.75E-10)
Targeted analysis of segregation phenotype from ADNI (N = 178, 108 male, 70 female, age 73.76 ± 6.95)
Common associations were identified across HCP and ADNI datasets (corrected p threshold = 3.3E-03):
rs7937515 ~ clustering coefficient measured at left middle temporal gyrus
rs7937515 ~ local efficiency measured at left middle temporal gyrus
Mediation analysis for pathway of SNP rs7937515 ~ segregation measures (iQT) ~ common phenotypic outcome in both datasets:
• HCP: BMI was influenced by rs7937515 with clustering coefficient and local efficiency as mediators
• ADNI: no such mediation effect was identified.
Targeted analysis of network segregation measures (iQT) and phenotypic outcome:
• HCP: clustering coefficient and local efficiency measured at left middle temporal gyrus were associated with BMI
• ADNI: no such association was identified
Targeted analysis of SNP rs7937515 and other important connectivity attributes (iQT):
• HCP: rs7937515 associated with global efficiency, eigen centrality, assortativity, density, number of edge, and strength
• ADNI: rs7937515 associated with global efficiency, transitivity, and strength

FIGURE 1 | Flowchart of brain connectome GWAS design. Abbreviations: SNPs, single nucleotide polymorphisms; ADNI, Alzheimer's disease neuroimaging initiative; HCP, human connectome project; dbGaP, database of genotypes and phenotypes; QC, quality control; ROI, region of interest; iQT: imaging quantitative trait; BMI, body mass index.

coefficients and local efficiency represent the connectivity of local network communities that are intrinsically densely connected and strongly coupled. A converging evidence (Cohen and D'Esposito, 2016; Karwowski et al., 2019) is shown that local, within-network communication is critical for motor execution, whereas integrative, between-network communication is critical for measuring connectome (Bertolero et al., 2019). Thus, network segregation is thought to be essential for describing and understanding of complex neural connectome systems (Sporns, 2013). In addition, segregation measures are highly reliable and heritable network attributes (Xie et al., 2018), and these measures have been linked to the disruption of neural network connectivity in brain development, evolution, disease (Cohen and D'Esposito, 2016; Mak et al., 2016; Bertolero et al., 2019), and immunodeficiency (Bell et al., 2018). Given the importance of network segregation, in this study, we first focus on quantifying measures of network segregation, analyzing heritability of segregation measures and performing genetic association analyses by treating them as neuroimaging traits. Then, our next priority is to explore the genetic basis for the rest of the complex-network attributes (e.g. integration, centrality and resilience).

To overcome the challenges mentioned above, this study aims to develop and implement computational and statistical strategies for a systematic characterization of structural connectome optimized for imaging genetic studies, and to determine genetic basis of structural connectome. Specifically, the framework is organized and described in **Figure 1**, and the primary goals are to address the following six critical issues: 1) construction of basic network connectivity with diffusion tractography, 2) systematic extraction of complex-network attributes, 3) heritability analysis of complex-network attributes, 4) genome-wide association studies of quantitative endophenotypes, 5) examination of mediation effect that intermediately bridges genes and outcomes, and 6) identification of outcome-relevant neuroimaging biomarkers. Given the enormously broad scope of brain connectome, our focus is on studying 1) static tractography-based structural connectome and complex-network attributes characterizing segregation, integration, centrality and resilience; 2) genetic consistency and discrepancy for complex-network attributes among cohorts different in age; and 3) mediation effects of network attributes on outcome-relevant genetics.

The major contributions of this study are fivefold:

- **New challenges in human connectome:** we elucidate the neurobiological pathway from SNPs to brain connectome, and to phenotypic outcomes. By integrating connectomics and genetics, this study provides new genetic mechanism insights into understanding detailed brain topological architecture, and encoding (or mapping) inter-regional connectivity in the genome.
- **New genetic insights for brain phenotype:** we validate the study outcomes by examining genetic consistency and discrepancy for complex-network attributes between young adult cohort and elderly adult cohort, which illustrates the genetic basis for human connectome in different life stages.
- **Biological findings:** we treat network segregation measures as imaging quantitative traits (iQT), and demonstrate that body mass index [BMI, which is related to multiple complex diseases (Emmerzaal et al., 2015; Stenholm et al., 2017)] is influenced by a locus rs7937515 with network segregation attributes (e.g., clustering coefficient and local efficiency) measured at the left middle temporal gyrus as mediators, which reveals the intermediate effects of brain connectivity in the pathway of outcome-relevant genetics.
- **Biological findings:** we discover network segregation as an important neuroimaging biomarker for BMI and weight-related disorders, and illustrate the importance of the left middle temporal gyrus for BMI.
- **Biological findings:** we demonstrate the hemispheric asymmetry of structural network organization in genetic association analyses and outcome-relevant studies.

TABLE 1 | Participant characteristics in HCP and ADNI genetic association analyses.

Cohort	HCP	ADNI	*p*
Number	275	178	—
Gender (M/F)	137/138	108/70	**3.02E-02**
Age	28.69 ± 3.64	73.76 ± 6.95	**5.56E-175**
Education	15.14 ± 1.64	16.03 ± 2.78	**1.41E-04**
MMSE	29.09 ± 1.04	27.37 ± 2.54	**2.28E-15**
Weight	77.70 ± 17.06	77.71 ± 15.92	1.00
BMI	25.99 ± 4.73	27.28 ± 5.24	**8.87E-03**
clus coef ROI 087	0.51 ± 0.05	0.29 ± 0.13	**1.41E-55**
loc effi ROI 087	0.52 ± 0.05	0.39 ± 0.17	**1.20E-18**

p-values were assessed because of significant differences among diagnosis groups, and were computed using one-way ANOVA (except for gender using χ^2 test). The p *values < 0.05 are shown in bold. HC = healthy control; EMCI = early mild cognitive complaint; LMCI = late mild cognitive complaint; AD = Alzheimer's disease.*

2 MATERIALS AND METHODS

2.1 Study Datasets

With the purpose of examining genetic consistency and discrepancy for complex-network attributes between young and elderly adults, and illustrating genetic basis for human connectome in different life stages, our analysis was respectively conducted on Human Connectome Project (HCP) database for young adults and Alzheimer's disease Neuroimaging Initiative (ADNI) database for elderly adults.

2.1.1 HCP Young Adult Dataset

HCP (Van Essen et al., 2013) is a major endeavor to map macroscopic human brain circuits and their relationship to behavior in a large population. It aims to reveal the neural pathways that underlie brain function and behavior, by acquiring and analyzing human brain connectivity from high-quality neuroimaging data in healthy young adults. The HCP datasets serve as a key resource for the neuroscience research community, as it provides valuable resources for characterizing human brain connectivity and function, their relationship to behavior, and their heritability and genetic underpinnings, which enables discoveries of how the brain is wired and how it functions in different individuals.

2.1.2 ADNI Elderly Adult Dataset

Alzheimer's disease Neuroimaging Initiative (ADNI) database was initially launched in 2004 as a public-private partnership, and led by the Principal Investigator Michael W. Weiner, MD. One primary aim of ADNI has been to examine whether serial imaging biomarkers extracted from MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. For up-to-date information, see www.adni-info.org.

2.2 Demographics

We initially downloaded 981 subjects from HCP database, including a part of twin subjects, then one individual from each family was randomly selected and excluded. As a result, 275 unrelated participants were selected for further population-based genetic analyses. ADNI data were collected by selecting the participants who had both genotype data and baseline DWI data at their first visit, family relationship was also removed in the same way as described above for HCP data filtration. Detailed characteristic information and the number of subjects in each data cohort are shown in **Table 1**. In this study, we analyzed a total of 275 participants (age: 28.7 ± 3.6; gender: 137 male, 138 female; education: 15.1 ± 1.6) from the HCP database, and a total of 178 participants (age: 73.8 ± 7.0; gender: 108 male, 70 female; education: 16.0 ± 2.8) from the ADNI database. This study was approved by institutional review boards of all participating institutions, and written informed consent was obtained from all participants or authorized representatives.

2.3 Genotyping Data Acquisition and Processing

2.3.1 HCP Young Adults Dataset

HCP samples were genotyped using MEGA array with PsychChip and ImmunoChip content. 1,141 genotype data was downloaded from dbGAP. Quality control was performed in PLINK v1.90 (Purcell et al., 2007) using the following criteria: 1) call rate per marker $\geq 98\%$, 2) minor allele frequency (MAF) $\geq 5\%$, 3) Hardy Weinberg Equilibrium (HWE) test $p \leq 1.0E\text{-}6$, and 4) call rate per participant $\geq 98\%$. Variants with no "rs" number, and samples with evidence of identity-by-descent (IBD) ≥ 0.25 or heterozygosity rate ±3 standard deviations from the mean were further excluded. Following quality control process, the number of samples with genotype data reduced to 327, we then checked the missing data by matching subjects information between phenotype and genotype data. As a result, this study comprised a total of 327 unrelated subjects and 515,956 SNPs.

2.3.2 ADNI Elderly Adults Dataset

Genotyping data were obtained from the ADNI database (adni.loni.usc.edu). They were quality-controlled as described in (Cong et al., 2020; Yao et al., 2020). We then performed imputation to maximize the number of overlaps between HCP GWAS findings and ADNI SNPs, see (Yao et al., 2019) for details. Briefly, genotyping was performed on all ADNI participants following the manufacturer's protocol using blood genomic DNA samples and Illumina GWAS arrays (610-Quad, OmniExpress, or HumanOmni2.5-4v1) (Saykin et al., 2010). Quality control was performed in PLINK v1.90 (Purcell et al., 2007) using the following criteria: 1) call rate per marker $\geq 95\%$, 2) minor allele frequency (MAF) $\geq 5\%$, 3) Hardy Weinberg Equilibrium (HWE) test $p \leq 1.0E\text{-}6$, and 4) call rate per participant $\geq 95\%$. In total, 5,574,300 SNPs were included for further targeted genetic association analysis.

2.4 Tractography and Network Construction

2.4.1 Tractography

We downloaded high spatial resolution DWI data and genotype data from both HCP and ADNI databases. DWI data from HCP was processed following the MRtrix3 guidelines (Tournier et al., 2012), detailed procedures have been previously reported (Xie et al., 2018) and are briefly described below: 1) generating a tissue-segmented image; 2) estimating the multi-shell multi-tissue response function and performing the multi-shell multi-tissue constrained spherical deconvolution; 3) generating the initial tractogram and applying the successor of Spherical-deconvolution Informed Filtering of Tractograms (SIFT2) methodology (Smith et al., 2015); and 4) mapping the SIFT2 output streamlines onto the MarsBaR automated anatomical labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002) with 90 ROIs to produce the structural connectome with edge value equal to the mean fractional anisotropy (FA).

DWI data from ADNI was acquired following the scanning protocols described in (Elsheikh et al., 2020), and processed following the procedures discussed in (Yan et al., 2018). Tractography was performed in Camino (Cook et al., 2006) based on white matter fiber orientation distribution function (ODF). As Camino adopted a deterministic approach, streamlines were modeled with a multi-tensor modeling approach (voxels fitted up to three fiber orientations, this way accounting for most of the fiber-crossings) of the ODF data. To generate a comparable tractography, the streamlines were also mapped onto AAL atlas with 90 ROIs to produce the structural connectome with edge value equal to the mean FA.

2.4.2 Network Construction

Network was created and defined by connectivity matrix M where M_{ij} stores the connectivity measure between ROIs i and j. As described in the previous section, we considered FA for defining M_{ij}. Since the diffusion tensor is a symmetric 3 × 3 matrix, it can be described by its eigenvalues (λ_1, λ_2 and λ_3) and eigenvectors (v_1, v_2 and v_3) for tractography analysis. FA at edge-level is an index for the amount of diffusion asymmetry within a voxel, defined in terms of its eigenvalues:

$$FA = \sqrt{\frac{(\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_1 - \lambda_3)^2}{2\left(\lambda_1^2 + \lambda_2^2 + \lambda_3^2\right)}}. \tag{1}$$

Thus, mean FA is a normalized measure of the fraction of the tensor's magnitude due to anisotropic diffusion, corresponding to the degree of anisotropic diffusion or directionality.

2.5 Complex-Network Attributes

With an undirected and weighted connectivity matrix M (defined in **Section 2.4.2**), we assessed a comprehensive set of network features such as segregation (e.g., transitivity, clustering coefficients, local efficiency and modularity), integration (e.g., global efficiency), centrality (e.g., eigen centrality) and resilience (e.g., assortativity) of the structural connectome using Brain Connectivity Toolbox (BCT) (Rubinov and Sporns, 2010). Given the importance and priority of segregation measures in this study, we only introduced the definitions of segregation measures, and the definitions of the rest complex-network attributes were explained in (Rubinov and Sporns, 2010).

For the following sub-sections, we define N as the set of all nodes in the network, n as the number of nodes, t_i as geometric mean of triangles around node i $\left(t_i = \frac{1}{2}\sum_{j,h\in N}(M_{ij}M_{ih}M_{jh})^{1/3}\right)$,

k_i as weighted degree of i ($k_i = \sum_{j\in N} M_{ij}$), a_{ij} as the connection status between i and $j\, a_{ij} = 1$ when link (i, j) exists, $a_{ij} = 0$ otherwise), d_{ij} as shortest weighted path length between i and $j\,(d_{ij} = \sum_{a_{uv}\in g_{i\hookrightarrow j}} f(M_{uv})$, where f is a map from weight to length and $g_{i\hookrightarrow j}$ is the shortest weighted path between i and j).

2.5.1 Transitivity

Transitivity measures the ratio of triangles to triplets in the network. By following the definition in (Newman, 2003):

$$T = \frac{\sum_{i\in N} 2t_i}{\sum_{i\in N} k_i (k_i - 1)}, \tag{2}$$

where T is the transitivity measured at network level.

2.5.2 Clustering Coefficient

Clustering coefficient measures the degree to which nodes in a network tend to cluster together by evaluating the fraction of triangles around a node and is equivalent to the fraction of node's neighbors that are neighbors of each other. By following the definition in (Onnela et al., 2005):

$$C = \frac{1}{n}\sum_{i\in N} C_i = \frac{1}{n}\sum_{i\in N}\frac{2t_i}{k_i (k_i - 1)}, \tag{3}$$

where C_i is the clustering coefficient of node i and C is the clustering coefficient measured at network level.

2.5.3 Local Efficiency

Local efficiency measures the efficiency of information transfer limited to neighboring nodes by evaluating the global efficiency computed on node neighborhoods. By following the definition in (Latora and Marchiori, 2001):

$$E_{\text{loc}} = \frac{1}{n}\sum_{i\in N}\frac{\sum_{j,h\in N, j\neq i}\left(M_{ij}M_{ih}\left[d_{jh}(N_i)\right]^{-1}\right)^{1/3}}{k_i (k_i - 1)}, \tag{4}$$

where E_{loc} is the local efficiency of node i, and $d_{jh}(N_i)$ is the length of the shortest path between j and h, that contains only neighbors of i.

2.5.4 Modularity

Modularity measures network segregation into distinct networks, and it is a statistic that quantifies the degree to which the network may be subdivided into such clearly delineated groups (Newman, 2006):

$$Q = \frac{1}{l}\sum_{i,j\in N}\left[M_{ij} - \frac{k_i k_j}{l}\right]\delta_{m_i,m_j}, \tag{5}$$

where Q is the modularity measured at network level, m_i is the module containing node i, and $\delta_{m_i,m_j} = 1$ if $m_i = m_j$, and 0 otherwise.

2.6 Heritability Analysis

Heritability analysis focused on identifying highly heritable measures of structural brain networks, and it was a commonly adopted and critical measurement to describe properties of the inheritance of iQT. An iQT such as network attributes must be heritable, which was defined as the proportion of phenotypic variance due to genetic differences between individuals (Jørstad and Næevdal, 1996). In this study, we estimated heritability of four segregation measures from twin subjects in the HCP young adult cohort (N = 350, 232 mono-zygotic twins, 118 di-zygotic twins) and SOLAR-Eclipse software (Kochunov et al., 2015) was employed for this task. The inputs to this software included phenotype traits, covariates measures and a kinship matrix indicating the pairwise relationship between twin individuals. A maximum likelihood variance decomposition method was applied to estimate the variance explained by additive genetic factors and environmental factors respectively. The output from SOLAR-Eclipse included heritability (h2), standard error and the corresponding significance p-value for each feature. We estimated the heritability of connectomic features, including transitivity, clustering coefficients, local efficiency and modularity. Since many previous studies had reported the effect of age (linear nonlinear), gender and their interactions on structural brain connectivity (Burzynska et al., 2010; Gong et al., 2011; Lopez-Larson et al., 2011; Zhao et al., 2015), all heritability analyses were performed with age, age^2, sex, age×sex and age^2×sex as covariates. In addition, we extracted the total variance explained by all covariate variables.

2.7 Brain Connectome Genetic Association Analysis

2.7.1 HCP Cohort

GWAS on the brain network segregation measures of the 90 ROIs were performed using linear regression under an additive genetic model in PLINK v1.90 (Purcell et al., 2007). Age, gender and education were included as covariates. Our GWAS was focused on analyzing the following network segregation measures: 1) modularity and transitivity, which were network-level measures; and 2) clustering coefficient and local efficiency, which were node-level measures. As a result, in total, we have 2 + 90 × 2 = 182 measures. Our post-hoc analysis used Bonferroni correction for correcting the genome-wide significance (GWS) for the number of quantitative traits (i.e., 5E-8/182 = 2.75E-10).

2.7.2 ADNI Cohort

Genetic findings of the segregation measures from HCP young adult dataset were treated as genotypic candidates and segregation measures at specific ROIs as phenotypic candidates, we further examined in ADNI elderly adult dataset regarding their associations. Apart from including age, gender and education as covariates, we also added diagnostic status into the linear regression model, as a large part of ADNI participants suffered from cognitive disorders. By validating the genetic findings from HCP data using ADNI participants, we examined genetic consistency and discrepancy for network segregation attributes between young and elderly adults, which illustrated the consistency and discrepancy of genetic basis for human connectome in different life stages.

In addition, the validated genetic findings were used to further explore connectivity variances with all important complex-network attributes excepting segregation measures such as integration (e.g., global efficiency and network density), centrality (e.g., eigen centrality) and resilience (e.g., assortativity), and we examined the targeted genetic basis on certain brain ROIs (e.g., middle temporal gyrus). As previously stated, linear regression models were used. In particular, we applied additive genetic models implemented in PLINK v1.90, with age, gender, education as covariates.

2.8 Mediation Analysis

To examine the causal assumption, we followed the Baron-Kenny procedure (Baron and Kenny, 1986) to perform standard mediation analysis to identify the mediated effect, and we treated iQTs (e.g., network segregation measures) as mediating variables, which intermediately linked the pathological path from genetic findings to clinical phenotypes. Specifically, we constructed a set of candidate SNPs which were found significantly associated to segregation measures in both young and elderly participants, and we constructed a set of candidate clinical phenotyping information by extracting overlapped clinical outcomes collected in both HCP and ADNI databases. We then employed the mediation model to detect the indirect effect of loci on clinical outcomes via iQT.

Specifically, mediation analysis was performed using the non-parametric bootstrap approximation with the R "mediation" package developed by Imai et al. (2010). Let y be the dependent variable which was a clinical outcome in our study, x be the independent variable which was a candidate SNP, z be the covariates (age, gender and education), and M be the mediating variable which was brain iQT. The mediation analysis was conducted in 3 steps:

1) fit a linear model to regress the mediating variable M against SNP x while controlling for z;
2) fit a linear model to regress the clinical outcome y against SNP x while controlling for z;
3) adopted the non-parametric bootstrap approximation to estimate the direct effect, mediation effect, proportion of total effect via mediation, their 95% confidence intervals (CI) and p values, by conducting 1,000 simulations.

2.9 Outcome-Relavent Brain Connectome Association Analysis

To discover the outcome-relevant biomarkers which mapped brain connectivity alterations to cognitive or behavioral outcomes, we performed pairwise univariate association analysis between network segregation attributes and outcome data. We selected BMI and Mini-Mental State Examination (MMSE) as outcomes as they were not only measures available in both HCP and ADNI cohorts but also important quantitative traits related to complex diseases such as weight-related disorders as well as neurological and psychiatric disorders. We used linear regression to regress the phenotypic outcomes against network segregation measures for both HCP and ADNI datasets, and explored outcome-relevant brain neuroimaging biomarkers. By comparing young and elderly participants, we illustrated the consistency and discrepancy of human brain connectome in different ages regarding on BMI and MMSE variations.

3 RESULTS

3.1 Heritability of Network Segregation

As illustrated in **Figure 1**, we examined segregation measures estimated at both network-level and node-level prior to GWAS. All of the segregation measures such as clustering coefficients (node-level), local efficiencies (node-level), transitivity (network-level) and modularity (network-level) showed significantly high heritability after Bonferroni correction ($p < 0.05/182 = 2.75E - 04$). The mean (±std) heritability of 182 segregation measures (h2 score) was 0.81 (±0.05), and more detailed results of heritability analysis were listed in Supplementary Table. We included all 182 segregation measures in the subsequent GWAS analysis.

3.2 GWAS of Network Segregation in HCP Young Adults

In the HCP cohort, genome-wide associations between 515, 956 SNPs and 182 structural network segregation measures were assessed under the additive genetic model and controlled for age, gender and education. GWAS identified 20 significant associations between 10 SNPs and 7 segregation measures (**Table 2**), after correcting the genome-wide significance (GWS) for the number of phenotypes using Bonferroni method (i.e., $p < 5E - 08/182 = 2.75E - 10$). Respectively shown in **Figure 2** were Manhattan plots of GWAS results of clustering coefficient and local efficiency measured in left middle temporal gyrus. GWAS of HCP data showed high consistency for clustering coefficient and local efficiency, where nine SNP-ROI associations were discovered for these two segregation measures. After Bonferroni correction, there was no significant finding for the network level segregation measures (i.e., transitivity and modularity).

3.3 Targeted Genetic Association of Segregation in ADNI Elderly Adults

Given the list of significant findings from the aforementioned GWAS of HCP segregation measures, we further examined their statistical significance in the ADNI cohort to identify brain network relevant genetic variants which were consistent for brain aging. We assessed the associations of 15 out of 20 HCP GWAS findings in ADNI cohort, as three SNPs (rs4841664, rs1461192 and rs147446959 are corresponding to 5 associations in **Table 2**) were not included in ADNI genotyping data. Associations of rs7937515 with clustering coefficient and local efficiency measured in left middle temporal gyrus were duplicated and validated in ADNI cohort with p values of 1.63E-03 and 1.34E-03, respectively, where the Bonferroni corrected significant level $p < 0.05/15 = 3.33E - 03$ was applied (**Table 2**).

TABLE 2 | Significant associations between SNPs and segregation measures: statistics in the HCP and ADNI cohorts.

Segregation measure	ROI	CHR	SNP	BP^{a}	Closest gene^{b}	HCP		ADNI	
						Beta	*p*	Beta	P^{c}
Clustering coefficient	FMidO_R	6	rs6930337	148788006	—	−0.36	**1.25E-10**	−0.10	1.65E-01
		18	rs1940608	5927441	TMEM200C	−0.39	**8.00E-12**	0.09	2.15E-01
		18	rs4798416	5930979	TMEM200C	−0.39	**7.52E-12**	0.09	2.15E-01
	FMedO_R	10	rs2104994	5273767	AKR1C4	−0.37	**7.71E-11**	−0.10	1.98E-01
	TPMid_L	4	rs9994092	66 436 114	—	−0.37	**1.51E-10**	−0.05	4.71E-01
		4	rs10032124	66 485 112	—	−0.39	**8.26E-12**	−0.04	5.68E-01
		8	rs4841664	11 859 985	DEFB134/DEFB135/ DEFB136	−0.38	**7.16E-11**	—	—
		11	rs7937515	71 841 325	ANAPC15/LRTOMT/ FOLR3/LAMTOR1	−0.37	**1.09E-10**	−3.20	**1.63E-03**
		11	rs1461192	130043580	ST14	−0.40	**5.54E-12**	—	—
Local efficiency	FMidO_R	6	rs6930337	148788006	—	−0.38	**4.68E-11**	−0.07	3.51E-01
		18	rs1940608	5927441	TMEM200C	−0.39	**7.75E-12**	0.10	2.02E-01
		18	rs4798416	5930979	TMEM200C	−0.39	**7.36E-12**	0.10	2.02E-01
	FMedO_L	10	rs2104994	5273767	AKR1C4	−0.37	**1.56E-10**	−0.11	1.33E-01
	FMedO_R	10	rs2104994	5273767	AKR1C4	−0.38	**1.86E-11**	−0.11	1.62E-01
	TPMid_L	4	rs9994092	66 436 114	—	−0.38	**6.84E-11**	−0.05	4.82E-01
		4	rs10032124	66 485 112	—	−0.40	**3.71E-12**	−0.04	5.71E-01
		8	rs4841664	11 859 985	DEFB134/DEFB135/ DEFB136	−0.38	**5.98E-11**	—	—
		11	rs7937515	71 841 325	ANAPC15/LRTOMT/ FOLR3/LAMTOR1	−0.38	**4.22E-11**	−0.24	**1.34E-03**
		11	rs1461192	130043 580	ST14	−0.39	**1.74E-11**	—	—
		21	rs147446959	29 291 173	—	−0.37	**2.72E-10**	-	—

^{a}*Build 37, assembly hg19.*

^{b}*Genes located ±100 kb of the top SNP.*

^{c}p *value reaching the Bonferroni corrected threshold (0.05/20 = 2.25E-03) is shown in bold.*

Abbreviations: F = frontal, TP = temporal pole, Mid = middle, Med = medial, O = orbital, L = left, R = right.

The minor allele G of rs7937515 (rs7937515-G) was associated with lower level of both clustering coefficient and local efficiency, compared to its major allele A (**Figure 3**). We will discuss the risk effect of rs7937515-G on brain function and dysfunction in the discussion section.

3.4 Mediation Analysis

According to the genetic association results from the HCP and ADNI subjects, we identified a common genetic finding SNP rs7937515, which was associated with two segregation measures in left middle temporal gyrus (e.g., clustering coefficient and local efficiency). In addition, we extracted two common behavioral and cognitive outcome measures (e.g., BMI and MMSE) by comparing the outcome evaluation methods across the HCP and ADNI databases. Thus, in this section, we had two major focuses: 1) exploring the genetic effect of SNP rs7937515 on outcomes including BMI and MMSE, and gaining deeper insights to the molecular mechanisms of the identified genetic variant, and 2) examining the mediated effect of iQTs (e.g., segregation measures) and illustrating their implicit effects in **Eq. 1**.

To achieve those two goals, mediation analysis of outcome was performed for evaluating both the direct and implicit effects of rs7937515 on outcomes (i.e., BMI and MMSE) through segregation measurements in left middle temporal gyrus. Mediation analysis required the independent variable (i.e., rs7937515) to be significantly associated with both the dependent variable (i.e., BMI or MMSE) and the mediator (i.e., segregation measurements). Below we respectively reported the mediation results analyzed from both HCP and ADNI data.

For the first focus, the minor allele G of rs7937515 was significantly associated with the increased BMI in HCP cohort (p = 1.62E-03; **Figure 4A**). The same increasing trend was also observed from the ADNI data, although the association between rs7937515 and BMI was not significant (p = 0.22; **Figure 4B**). For the second focus, **Figure 5** illustrated the results of mediation analysis with BMI as an outcome measure, from which both clustering coefficient and local efficiency of the left middle temporal gyrus demonstrated the significant intermediate roles between rs7937515 and BMI, with mediation effects of 0.98 (95% CI = [0.06, 2.29], p = 3.60E-02) and 0.99 (95%CI = [0.02, 2.11], p = 4.60E-02), respectively. There was no significant association between rs7937515 with MMSE in the HCP young adult dataset, so no mediation analysis regarding MMSE was performed. In the ADNI elderly adult dataset, there were no significant associations observed between rs7937515 with BMI nor MMSE; therefore it was not necessary to further examine mediation effects.

Since the brain can be viewed as a predictor, a mediator, or outcome of a health condition (e.g., obesity) (Lowe et al., 2019), it is unclear whether the brain regulates the condition (e.g., structural connectome alteration considered as a mediator for a physical condition such as BMI), or, conversely, brain is affected by the condition. For completeness, we also explored the potential reciprocal relationship from the other direction. The above experiment was repeated with BMI as a mediator and

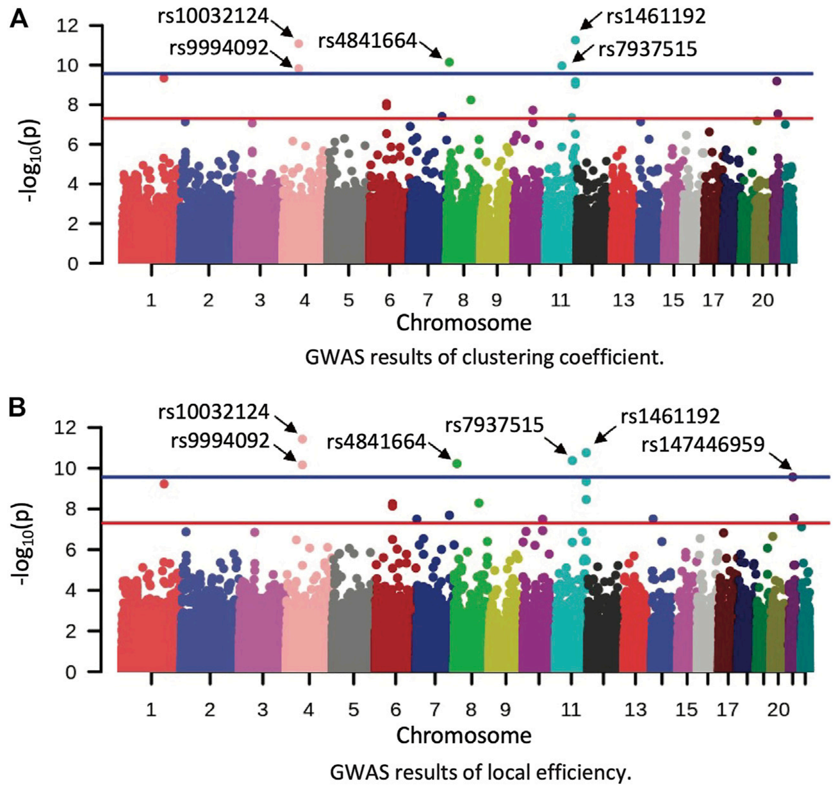


FIGURE 2 | Manhattan plot of GWAS results in the HCP dataset. **(A,B)** show the GWAS results of clustering coefficient and local efficiency on left middle temporal gyrus, respectively. Red and blue lines correspond to the p-value of 5E-08 and 2.75E-10, respectively.

connectivity measures as outcomes. No significant findings were identified, and thus no evidence was observed for BMI as a significant mediator between gene and brain connectivity.

3.5 Outcome-Relevant Neuroimaging Biomaker Discoveries

On one hand, for the HCP cohort, we respectively identified significantly negative associations (p <0.05/4 = 1.25E − 02) between BMI with clustering coefficient (p = 3.92E-05) and local efficiency (p = 4.57E-05) measured in left middle temporal gyrus. On the other hand, for the ADNI cohort, we examined the associations between BMI and the above mentioned segregation measures in a pair-wise manner, but there was no significant findings satisfying the corrected p threshold. Regarding the relationship between cognitive score (e.g., MMSE) and network segregation measures, there was no significant associations identified for both HCP and ADNI cohorts.

3.6 Targeted Genetic Association of Other Important Network Attributes in the Left Middle Temporal Gyrus

To review the genetic effect of SNP rs7937515 from different aspects of network connectivity attributes of the left middle temporal gyrus, we assessed the relationship between rs7937515 and additional node-level measures on reported brain ROI (i.e., left middle temporal gyrus) as well as network-level measures in both HCP and ADNI datasets. **Table 3** showed association statistics of rs7937515 with segregation, integration, centrality and resilience measures. After correcting for the number of examined network measures (i.e., $p < 0.05/9 = 5.56e - 03$), both HCP and ADNI identified significant associations between the targeted SNP with global efficiency (integration) and transitivity (resilience), together with our previous finding that rs7937515 was associated with segregation measures such as clustering coefficient and local efficiency, our results showed the consistent genetic effect of rs7937515 on brain structural network segregation, integration and resilience across aging. Besides the common findings between young and elderly adults, rs7937515 was associated with several other node-level and network-level attributes including network density (integration) and eigenvector centrality (centrality) in HCP data, but not in ADNI. Our results suggested the possible genetic discrepancy for certain brain connectivities in different life stages.

3.7 Hemispheric Asymmetry of Brain Connectome

In this study, we noticed a hemispheric asymmetry of outcome-relevant brain connectivity alterations in the left

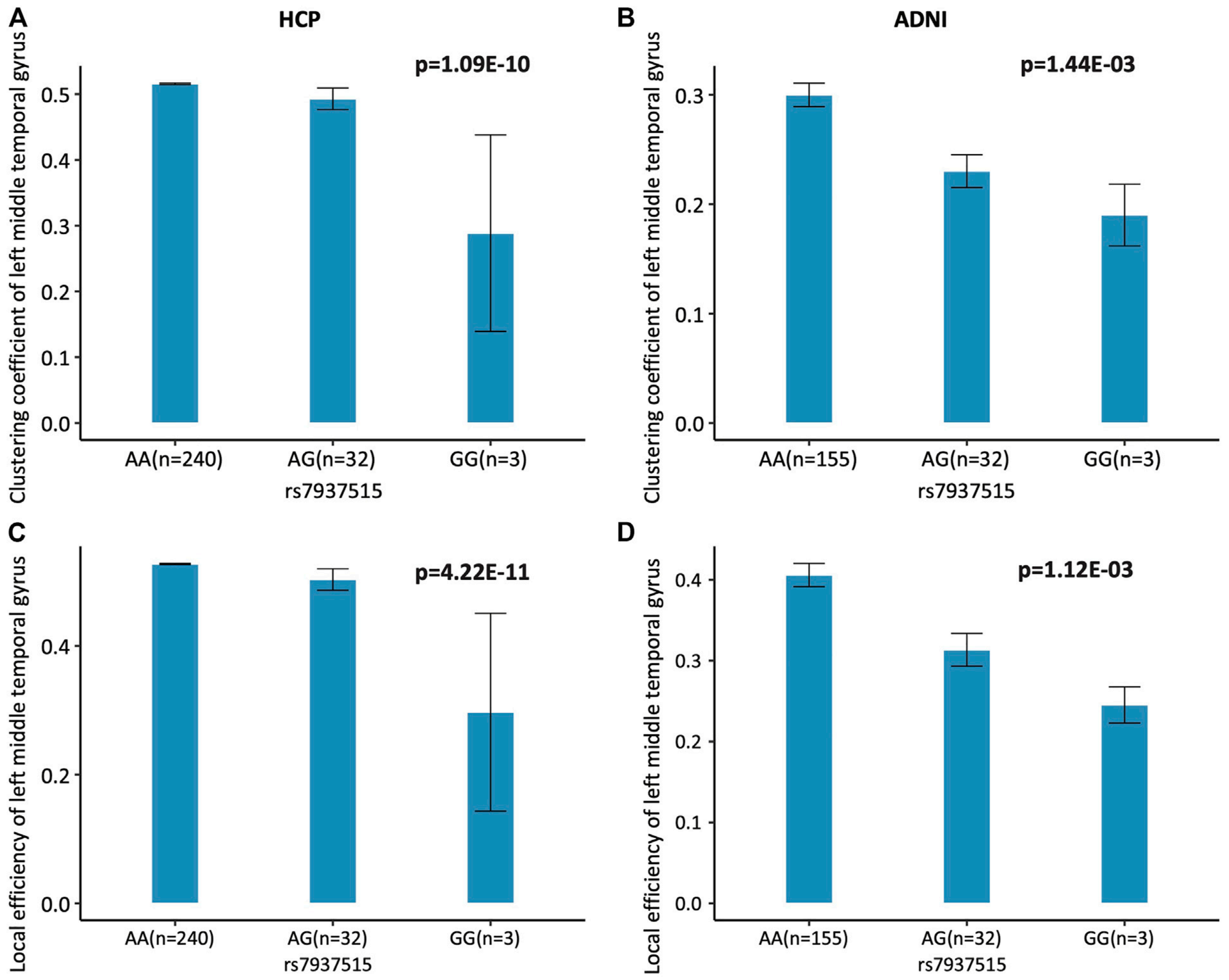


FIGURE 3 | Association of rs7937515 on clustering coefficient and local efficiency of the left middle temporal gyrus. **(A,B)** Mean clustering coefficient with standard errors are plotted against the rs7937515 genotype groups (i.e., AA, AG and GG) for the HCP and ADNI cohorts, respectively. **(C,D)** Mean local efficiency with standard errors are plotted against the rs7937515 genotype groups (i.e., AA, AG and GG) for the HCP and ADNI cohorts, respectively. *p* values are calculated from GWAS (HCP) and targeted analysis (ADNI), and significant *p* values are marked in bold.

and right middle temporal gyrus (**Table 3**). Due to two brain regions (e.g., left and right MTG), two segregation measures (e.g., clustering coefficient and local efficiency) and one outcome measure (e.g., BMI), we applied a Bonferroni corrected p threshold in this section ($p < 5E-02/8 = 6.25E-03$). In the HCP young adult cohort, for the left MTG, we respectively identified significant associations of BMI with clustering coefficient (*p* = 3.92E-05), and with local efficiency (*p* = 4.57E-05); for the right MTG, even though there were no significant associations of BMI with clustering coefficient (*p* = 2.24E-02), and with local efficiency (*p* = 2.90E-02), both clustering coefficient and local efficiency in left and right MTG showed negative associations with BMI. In the ADNI cohort, as reported in the previous section, network segregation was not associated with BMI, so it was not necessary and proper to conduct analyses regarding ADNI data in this section.

4 DISCUSSION

As summarized in **Figure 1**, prior to GWAS, we first performed heritability analysis for network attributes screening, and only heritable measures of network segregation were treated as iQT for GWAS. Based on experimental outcomes, all of the segregation measures were highly heritable: transitivity and modularity were heritable at network level, clustering coefficient and local efficiency were heritable at all nodes, which suggested segregation measures were suitable for genetic analyses. Then, we performed GWAS of segregation attributes in 275 HCP subjects, and identified several pairwise associations between SNPs and iQTs as listed in **Table 2**. These GWAS findings were validated in 178 ADNI subjects. As a validation result, we identified several genetic consistency and discrepancy patterns for human connectome in different life stages (as shown in **Table 2**). As common findings in both HCP young adult and

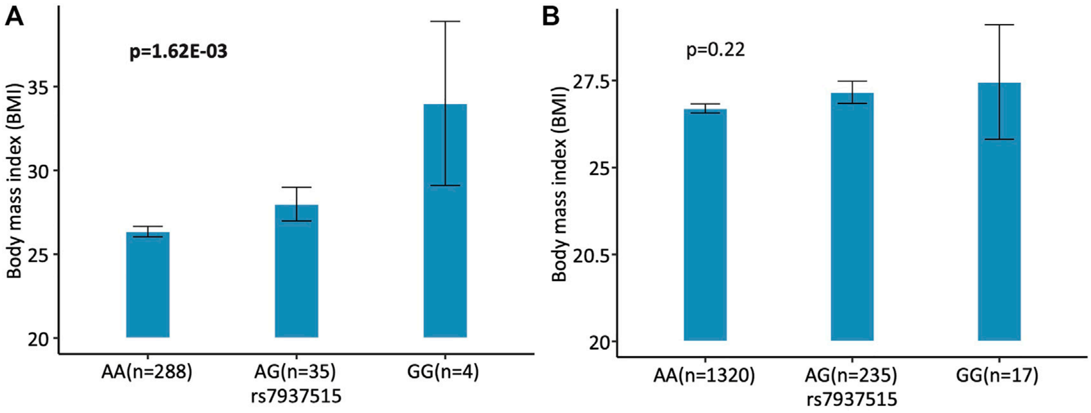


FIGURE 4 | Association of rs7937515 on BMI in the HCP and ADNI cohorts. **(A)** Mean BMI with standard errors are plotted against the rs7937515 genotype groups (i.e., AA, AG and GG) for the HCP cohort. **(B)** Mean BMI with standard errors are plotted against the rs7937515 genotype groups (i.e., AA, AG and GG) for the ADNI cohort. *p* values are calculated from mediation analysis, and significant *p* values are marked in bold.

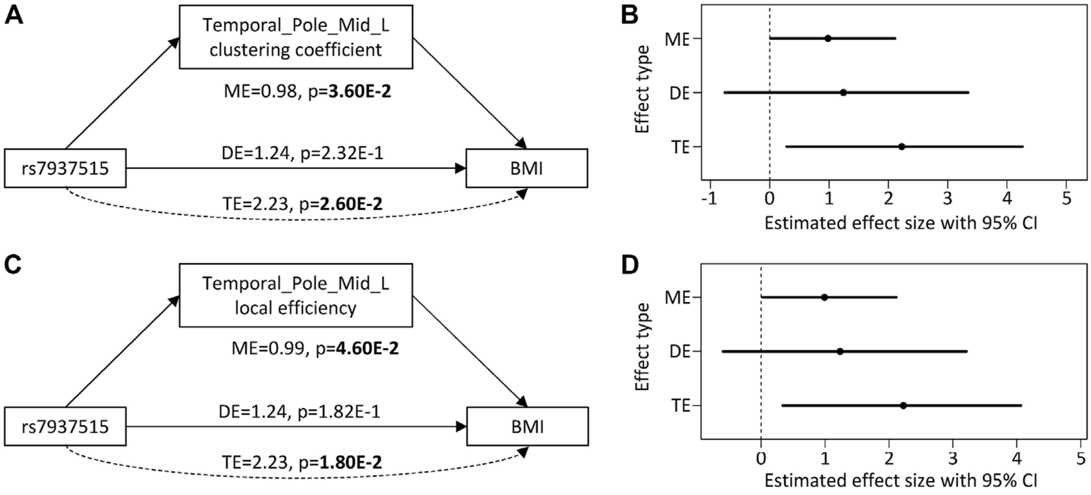


FIGURE 5 | Direct and mediation effect of rs7937515 on BMI through left middle temporal gyrus. **(A,B)** illustrate the effect size, 95% CI and *p* value from rs7937515 mediation analysis of BMI via left middle temporal clustering coefficient. **(C,D)** illustrate the effect size, 95% confidence interval and *p* value from rs7937515 mediation analysis of BMI via left middle temporal local efficient. TE = total effect; DE = direct effect; ME = mediation effect; CI = confidence interval.

ADNI elderly adult cohorts, rs7937515 was negatively associated with clustering coefficient and local efficiency respectively measured at left middle temporal gyrus. To the best of our knowledge, this was among the first GWAS of human brain high-level network measures across both young and elderly participants. As shown in **Figures 3A,C**, the minor allele G of rs7937515 was associated with decreased clustering coefficient and local efficiency of the left middle temporal gyrus in both young and elderly participants. As concluded in (Rudie et al., 2013; Keown et al., 2017; Karwowski et al., 2019; Varangis et al., 2019), the weakness of segregated network connectivity (e.g., modularity, clustering coefficient, and local efficiency) was linked to multiple brain disorders such as age-related cognitive declines and autism spectrum disorder. Thus, our GWAS findings for HCP young adults demonstrated that rs7937515 played a risk effect on human network segregation. This neurorisk effect was also confirmed in a targeted genetic association analysis for ADNI elderly participants (as shown in **Figures 3B,D**), these common discoveries between HCP and ADNI datasets suggested a consistent genetic risk effect across young and old life stages.

This study was further conducted by performing several post-hoc analyses in the following three aspects (shown as bottom sections in **Figure 1**): 1) examining genetic mechanisms for common outcome measures in the HCP and ADNI studies, and elucidating the mediated effect of iQTs for such outcome-relevant genetic association, 2) discovering outcome-relevant imaging biomarkers, and 3) exploring the genetic mechanisms of other important complex-network attributes.

For the first aspect, our goal was to elucidate the neurobiological pathway from SNPs to brain connectome, and

TABLE 3 | Associations between rs7937515 and brain network measures.

Class	QT	ROI or Global	HCP		ADNI	
			Beta	*p*	Beta	*p*
Segregation	Clustering coefficient	TPMid_L	−0.37	**1.09E-10**	−0.24	**1.63E-03**
	Clustering coefficient	TPMid_R	−0.20	**3.53E-04**	−0.22	**3.94E-03**
	Local efficiency	TPMid_L	−0.38	**4.22E-11**	−0.24	**1.34E-03**
	Local efficiency	TPMid_R	−0.22	**7.05E-05**	−0.22	**2.92E-03**
	Transitivity	Global	−0.23	**3.65E-05**	−0.24	**1.17E-03**
	Modularity	Global	0.20	**5.32E-04**	−0.12	9.32E-02
Integration	Global efficiency	Global	−0.29	**1.63E-07**	−0.24	**1.48E-03**
	Density	Global	−0.26	**2.64E-06**	0.03	7.11E-01
Centrality	Betweenness centrality	TPMid_L	−0.09	1.28E-01	−0.05	5.28E-01
	Betweenness centrality	TPMid_R	−0.06	3.24E-01	−0.03	6.75E-01
	Eigenvector centrality	TPMid_L	−0.32	**9.58E-08**	−0.13	7.85E-02
	Eigenvector centrality	TPMid_R	−0.20	**6.11E-04**	−0.03	6.98E-01
Resilience	Assortativity coefficient	Global	0.10	1.14E-01	0.06	3.95E-01

Abbreviations: TP = temporal pole, Mid = middle, L = left, R = right, QT, quantitative trait. p values reaching the Bonferroni corrected threshold (0.05/9 = 5.56E-03) are shown in bold.

to phenotypic outcome. In addition, we aimed to discover the role of iQTs in the outcome-relevant genetic associations by performing mediation analyses in both HCP and ADNI datasets. For the HCP young participants, we identified that BMI was positively associated with rs7937515 in the first step of mediation analysis, demonstrating a risk effect. rs7937515 located in the regions of *FAM86C1*/FOLR3 was previously discussed in literatures (Gao et al., 2015; Gao, 2017) and positively linked to waist circumference in the meta-analysis based on the Insulin Resistance Atherosclerosis Family Study (IRASFS) (Palmer et al., 2015), which was designed to investigate the genetic and environmental basis of insulin resistance and adiposity. *FAM86C1* (Family With Sequence Similarity 86 Member C1) and *FOLR3* (Folate Receptor Gamma) had been reported for their associations with various weight-related phenotypes such as bone mineral density (Li et al., 2019) and BMI (Hair, 2014; Mrozikiewicz et al., 2019), which closely related to osteoporosis (Li et al., 2019; Mrozikiewicz et al., 2019) and obesity (Gómez-Ambrosi et al., 2004). In the second and third steps of mediation analysis, we illustrated that BMI was indirectly influenced by rs7937515 (**Figures 4**, **5**), and iQTs such as clustering coefficient and local efficiency measured at the left middle temporal gyrus respectively played a mediating role. We also examined the genetic association with MMSE, but no evidence indicated any genetic associations to MMSE. In contrary, for the ADNI elderly participants, neither significant associations between rs7937515 and BMI nor MMSE were identified in the first step of mediation analysis, so there was not a necessary to examine mediated effect in this dataset. Our results demonstrated a disappearance of outcome-relevant genetic effect in the elderly participants, this discrepancy from young to elderly participants might due to the dominated influences from life style, environment or other non-genetic factors.

For the second aspect, recent studies (Lowe et al., 2019; Azevedo et al., 2019) showed that structural changes in brain tissues could affect food consumption behaviors and mediate BMI, which implied connectome alteration could be a causal agent and a promising imaging biomarker in this study. Thus, our goal was set to reveal the mapping between connectivity alterations and phenotypic outcome, and discover outcome-relevant imaging biomarkers. For young adult participants, segregation measures (e.g., clustering coefficient or local efficiency measured at left middle temporal gyrus) previously demonstrated their potential to play a mediating role in genetic association discoveries, in this step, we focused on examining their direct associations to the outcomes. Thus, we performed a targeted association analysis between the mentioned segregation measures and the common outcomes (e.g., BMI or MMSE) evaluated in both HCP and ADNI studies (**Table 2**) by employing linear regression models. For the young participants, clustering coefficient and local efficiency measured at left middle temporal gyrus were negatively associated with BMI. Similar observation was obtained in (Chen et al., 2018) which linked lower structural network segregation to obesity (higher BMI). Our findings suggested that there was a mapping between brain network segregation attributes and human physical conditions, and segregation features of the left middle temporal gyrus showed their potential as neuroimaging biomarkers to detect BMI-associated complex diseases such as dementias (Emmerzaal et al., 2015), cardiovascular disease, cancer, respiratory disease and diabetes (Stenholm et al., 2017). For elderly adult participants, no significant associations were identified between segregation measures and any outcomes, which suggested an interesting topic for further explorations.

Multiple regression analyses demonstrated that middle temporal gyrus was linked to weight-related issues. For example, Veit et al. (2014) and Gómez-Apo et al. (2018) revealed that BMI, visceral fat and age were negatively associated with cortical thickness of the left middle temporal gyrus, Ou et al. (2015) indicated that greater adiposity was associated with lower gray matter (GM) volumes in the middle temporal gyrus, Yokum et al. (2012) found positive correlation between BMI and white matter (WM) volume in the middle temporal gyrus, Rapuano et al. (2016) illustrated left middle temporal gyrus was detected significantly greater activation in response to food commercials than to non-food commercials, Salzwedel et al. (2019) concluded that maternal adiposity influenced neonatal brain functional connectivity in middle temporal gyrus, and Peven et al. (2019) identified that cardiorespiratory fitness was negatively associated with functional connectivity in the right middle temporal gyrus. To the best of our knowledge, our investigations for the association between structural connectivity in the middle temporal gyrus and BMI was among the first weight-related studies with high-level imaging features measured from structural network connectivity, and our results confirmed several previous findings analyzed from thickness data, T1-weighted MRI data, and fMRI data.

For the third aspect, since there was an emerging interest in understanding the segregation and the integration of brain networks (Cohen and D'Esposito, 2016; Mohr et al., 2016) as well as other important network attributes such as centrality (Zuo et al., 2012) and resilience (Karwowski et al., 2019), our goal was to expand our focus on comprehensively discussed segregation attributes to a more complete set of network attributes including segregation, integration, centrality and resilience. For both node level network attributes measured at left and right middle temporal gyrus and global network attributes, we applied targeted genetic association analyses on global efficiency and density (integration, network level), betweeness and eigenvector centrality (centrality, node level) and assortativity coefficient (resilience, network level) of the structural connectivity. We identified several pairwise associations between rs7937515 and these network attributes in both HCP and ADNI datasets (**Table 3**), and noticed a significant association between rs7937515 and global efficiency in both datasets, which suggested that rs7937515 was involved into the dynamic fluctuations of segregation and integration of neural information. This finding partially answered an elusive question of revealing genetic basis for brain mechanisms of balancing network segregation and integration. Another worth noting finding was that rs7937515 was associationed density and eigenvector centrality respectively in our targeted analyses, while such associations were vanished in elderly participants, which suggested inconsistent genetic influences across different life stages.

With the awareness of the hemispheric asymmetry of network organization, a genetic basis to explain the differences in connectome between two hemispheres were under discovered. In this work, we identified an obvious inconsistency of genetic influences on human connectome in different brain hemispheres (**Table 3**). As reported in several recent studies (Tian et al., 2011; Shu et al., 2015; Jiang et al., 2019), the topological organizations of structural networks were not uniformly affected across brain hemispheres, which lead to a non-uniformly distributed destruction on neural network of the left and right hemispheres. Our finding gave an explanation from the point-view of genetics, with the potential for further investigations as many of the destruction on neural network (as iQT) were linked to cognitive and behavioral functions and dysfunctions, and their genetic mechanisms were still under discovered.

5 CONCLUSION

In this work, we constructed the structural network connectivity, extracted complex-network attributes and examined the heritability of network segregation measures. Then, we revealed a novel association between the minor allele (G) of rs7937515 and decreased network segregation measures of the left middle temporal gyrus across HCP young participants and ADNI elderly participants, which demonstrated a consistent genetic risk effect on brain network connectivity across lifespan. We elucidated the neurobiological pathway from SNP rs7937515 and genes *FAM86C1*/*FOLR3* to brain network segregation, and to BMI. In such pathway, we concluded a genetic risk effect on BMI due to their positive association, examined the mediated effect of network segregation measures, and discovered network segregation of the left middle temporal gyrus as BMI-related neuroimaging biomarkers by identifying a negative association between them. We also examined genetic associations of a more complete set of important network attributes including integration, centrality and resilience, and demonstrated the consistency and discrepancy in genetic associations in brain aging. At last, we illustrated hemispheric asymmetry of network organization in both datasets in the aspect of genetic effect.

In sum, with the awareness that BMI is directly and indirectly associated to multiple complex diseases, this study performed a systematic analysis that integrated genetics, connectomics and phenotypic outcome data, with the goal of revealing biological mechanisms from the genetic determinant to intermediate brain connectomic traits and to the BMI phenotype at two different life stages. We identified the genetic effect of rs7937515 on human brain network segregation in both young and elderly participants and on BMI in young adult cohort. Our findings confirmed several previous genetic and imaging biomarker discoveries. Moreover, we provided outcome-relevant genetic insights in the aspect of complex-network attributes of human brain connectome. Similar analytical strategies can be applied to future integrative studies linking genomics with connectomics, including the analyses of structural and functional connectivity measures, additional network attributes, longitudinal or dynamic connectomic features, as well as other types of brain imaging genomic data.

6 THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data, but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: the ADNI website (http://adni.loni.usc.edu/) and the HCP website (https://www.humanconnectome.org/).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Boards (IRB) at University of Pennsylvania. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SC: Conceptualization, Methodology, Software, Formal analysis, Validation, Writing—Original Draft. XY: Formal analysis, Validation, Writing—Review and Editing. JY: Data Curation, Resources Writing—Review and Editing. LX: Data Curation, Software, Writing—Review and Editing. LS: Supervision, Conceptualization, Methodology, Writing—Review and Editing.

FUNDING

This work was supported in part by NIH R01 EB022574, R01 LM013463 and R21 AG066135, and by NSF 1837 964, 1942 394, and 1755 836.

ACKNOWLEDGMENTS

ADNI Acknowledgements: Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. **HCP Acknowledgements:** Data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH09-1657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.782953/full#supplementary-material

REFERENCES

Alloza, C., Cox, S. R., Blesa Cábez, M., Redmond, P., Whalley, H. C., Ritchie, S. J., et al. (2018). Polygenic Risk Score for Schizophrenia and Structural Brain Connectivity in Older Age: A Longitudinal Connectome and Tractography Study. *Neuroimage* 183, 884–896. doi:10.1016/j.neuroimage.2018.08.075

Azevedo, E. P., Pomeranz, L., Cheng, J., Schneeberger, M., Vaughan, R., Stern, S. A., et al. (2019). A Role of Drd2 Hippocampal Neurons in Context-Dependent Food Intake. *Neuron* 102, 873–886. doi:10.1016/j.neuron.2019.03.011

Baron, R. M., and Kenny, D. A. (1986). The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations. *J. Personal. Soc. Psychol.* 51, 1173–1182. doi:10.1037/0022-3514.51.6.1173

Bell, R. P., Barnes, L. L., Towe, S. L., Chen, N.-k., Song, A. W., and Meade, C. S. (2018). Structural Connectome Differences in Hiv Infection: Brain Network Segregation Associated with Nadir Cd4 Cell Count. *J. Neurovirol.* 24, 454–463. doi:10.1007/s13365-018-0634-4

Bertolero, M. A., Blevins, A. S., Baum, G. L., Gur, R. C., Gur, R. E., Roalf, D. R., et al. (2019). The Network Architecture of the Human Brain Is Modularly Encoded in the Genome. arXiv preprint arXiv:1905.07606.

Burzynska, A. Z., Preuschhof, C., Bäckman, L., Nyberg, L., Li, S.-C., Lindenberger, U., et al. (2010). Age-Related Differences in white Matter Microstructure: Region-Specific Patterns of Diffusivity. *Neuroimage* 49, 2104–2112. doi:10.1016/j.neuroimage.2009.09.041

Cauda, F., Nani, A., Manuello, J., Premi, E., Palermo, S., Tatu, K., et al. (2018). Brain Structural Alterations Are Distributed Following Functional, Anatomic and Genetic Connectivity. *Brain* 141, 3211–3232. doi:10.1093/brain/awy252

Chen, V. C.-H., Liu, Y.-C., Chao, S.-H., McIntyre, R. S., Cha, D. S., Lee, Y., et al. (2018). Brain Structural Networks and Connectomes: the Brain–Obesity Interface and its Impact on Mental Health. *Neuropsychiatr. Dis. Treat.* 14, 3199–3208. doi:10.2147/ndt.s180569

Cohen, J. R., and D'Esposito, M. (2016). The Segregation and Integration of Distinct Brain Networks and Their Relationship to Cognition. *J. Neurosci.* 36, 12083–12094. doi:10.1523/jneurosci.2965-15.2016

Cong, S., Risacher, S. L., West, J. D., Wu, Y.-C., Apostolova, L. G., Tallman, E., et al. (2018). Volumetric Comparison of Hippocampal Subfields Extracted from 4-minute Accelerated vs. 8-minute High-Resolution T2-Weighted 3t Mri Scans. *Brain Imaging Behav.* 12, 1583–1595. doi:10.1007/s11682-017-9819-3

Cong, S., Yao, X., Huang, Z., Risacher, S. L., Nho, K., Saykin, A. J., et al. (2020). Volumetric Gwas of Medial Temporal Lobe Structures Identifies an Erc1 Locus Using Adni High-Resolution T2-Weighted Mri Data. *Neurobiol. Aging* 95, 81–93. doi:10.1016/j.neurobiolaging.2020.07.005

Cook, P., Bai, Y., Nedjati-Gilani, S., Seunarine, K., Hall, M., Parker, G., et al. (2006). "Camino: Open-Source Diffusion-Mri Reconstruction and Processing," in 14th scientific meeting of the international society for magnetic resonance in medicine, Seattle WA, USA, May 6–12, 2006, 2759.

Elliott, L. T., Sharp, K., Alfaro-Almagro, F., Shi, S., Miller, K. L., Douaud, G., et al. (2018). Genome-Wide Association Studies of Brain Imaging Phenotypes in uk Biobank. *Nature* 562, 210–216. doi:10.1038/s41586-018-0571-7

Elsheikh, S. S. M., Chimusa, E. R., Mulder, N. J., and Crimi, A. (2020). Genome-Wide Association Study of Brain Connectivity Changes for Alzheimer's Disease. *Sci. Rep.* 10, 1433. doi:10.1038/s41598-020-58291-1

Emmerzaal, T. L., Kiliaan, A. J., and Gustafson, D. R. (2015). 2003-2013: A Decade of Body Mass Index, Alzheimer's Disease, and Dementia. *J. Alzheimers Dis.* 43, 739–755. doi:10.3233/JAD-141086

Farahani, F. V., Karwowski, W., and Lighthall, N. R. (2019). Application of Graph Theory for Identifying Connectivity Patterns in Human Brain Networks: a Systematic Review. *Front. Neurosci.* 13, 585. doi:10.3389/fnins.2019.00585

Gao, C. (2017). "Investigation of the Genetic Architecture of Cardiometabolic Disease. Ph.D. thesis. 1834 Wake Forest Rd, Winston-Salem, NC 27109: Wake Forest University.

Gao, C., Wang, N., Guo, X., Ziegler, J. T., Taylor, K. D., Xiang, A. H., et al. (2015). A Comprehensive Analysis of Common and Rare Variants to Identify Adiposity Loci in Hispanic Americans: the Iras Family Study (Irasfs). *PloS one* 10 (11), e0134649. doi:10.1371/journal.pone.0134649

Gómez-Ambrosi, J., Catalán, V., Diez-Caballero, A., Martínez-Cruz, L. A., Gil, M. J., García-Foncillas, J., et al. (2004). Gene Expression Profile of Omental Adipose Tissue in Human Obesity. *FASEB J.* 18, 215–217. doi:10.1096/fj.03-0591fje

Gómez-Apo, E., García-Sierra, A., Silva-Pereyra, J., Soto-Abraham, V., Mondragón-Maya, A., Velasco-Vales, V., et al. (2018). A Postmortem Study of Frontal and Temporal Gyri Thickness and Cell Number in Human Obesity. *Obesity* 26, 94–102. doi:10.1002/oby.22036

Gong, G., He, Y., and Evans, A. C. (2011). Brain Connectivity. *Neuroscientist* 17, 575–591. doi:10.1177/1073858410386492

Guo, Y., Shen, X.-N., Hou, X.-H., Ou, Y.-N., Huang, Y.-Y., Dong, Q., et al. (2020). Genome-Wide Association Study of white Matter Hyperintensity Volume in Elderly Persons without Dementia. *NeuroImage: Clin.* 26, 102209. doi:10.1016/j.nicl.2020.102209

Hair, B. (2014). Body Mass Index, Breast Tissue, and the Epigenome. Ph.D. thesis. Chapel Hill, NC: University of North Carolina at Chapel Hill.

Imai, K., Keele, L., and Tingley, D. (2010). A General Approach to Causal Mediation Analysis. *Psychol. Methods* 15, 309–334. doi:10.1037/a0020761

Jahanshad, N., Rajagopalan, P., Hua, X., Hibar, D. P., Nir, T. M., Toga, A. W., et al. (2013). Genome-Wide Scan of Healthy Human Connectome Discovers Spon1 Gene Variant Influencing Dementia Severity. *Proc. Natl. Acad. Sci. U S A.* 110, 4768–4773. doi:10.1073/pnas.1216206110

Jiang, X., Shen, Y., Yao, J., Zhang, L., Xu, L., Feng, R., et al. (2019). Connectome Analysis of Functional and Structural Hemispheric Brain Networks in Major Depressive Disorder. *Transl Psychiatry* 9, 136. doi:10.1038/s41398-019-0467-9

Jørstad, K. E., and Næevdal, G. (1996). Breeding and Genetics. *Dev. Aquacult. Fish. Sci.* 29, 655–725.

Keown, C. L., Datko, M. C., Chen, C. P., Maximo, J. O., Jahedi, A., and Müller, R.-A. (2017). Network Organization Is Globally Atypical in Autism: a Graph Theory Study of Intrinsic Functional Connectivity. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* 2, 66–75. doi:10.1016/j.bpsc.2016.07.008

Kochunov, P., Glahn, D. C., Lancaster, J. L., Winkler, A. M., Smith, S., Thompson, P. M., et al. (2010). Genetics of Microstructure of Cerebral white Matter Using Diffusion Tensor Imaging. *Neuroimage* 53, 1109–1116. doi:10.1016/j.neuroimage.2010.01.078

Kochunov, P., Jahanshad, N., Marcus, D., Winkler, A., Sprooten, E., Nichols, T. E., et al. (2015). Heritability of Fractional Anisotropy in Human white Matter: A Comparison of Human Connectome Project and enigma-dti Data. *Neuroimage* 111, 300–311. doi:10.1016/j.neuroimage.2015.02.050

Latora, V., and Marchiori, M. (2001). Efficient Behavior of Small-World Networks. *Phys. Rev. Lett.* 87, 198701. doi:10.1103/physrevlett.87.198701

Li, L., Wang, X., Wang, X., Liu, X., Guo, R., and Zhang, R. (2019). Integrative Analysis Reveals Key Mrnas and Lncrnas in Monocytes of Osteoporotic Patients. *Math. biosciences Eng. MBE* 16, 5947–5970. doi:10.3934/mbe.2019298

Lopez-Larson, M. P., Anderson, J. S., Ferguson, M. A., and Yurgelun-Todd, D. (2011). Local Brain Connectivity and Associations with Gender and Age. *Develop. Cogn. Neurosci.* 1, 187–197. doi:10.1016/j.dcn.2010.10.001

Lowe, C. J., Reichelt, A. C., and Hall, P. A. (2019). The Prefrontal Cortex and Obesity: a Health Neuroscience Perspective. *Trends Cognitive Sciences* 23, 349–361. doi:10.1016/j.tics.2019.01.005

Mak, E., Colloby, S. J., Thomas, A., and O'Brien, J. T. (2016). The Segregated Connectome of Late-Life Depression: A Combined Cortical Thickness and Structural Covariance Analysis. *Neurobiol. Aging* 48, 212–221. doi:10.1016/j.neurobiolaging.2016.08.013

Mohr, H., Wolfensteller, U., Betzel, R. F., Mišić, B., Sporns, O., Richiardi, J., et al. (2016). Integration and Segregation of Large-Scale Brain Networks during Short-Term Task Automatization. *Nat. Commun.* 7, 13217. doi:10.1038/ncomms13217

Mrozikiewicz, A. E., Bogacz, A., Barlik, M., Górska, A., Wolek, M., and Kalak, M. (2019). The Role of Folate Receptor and Reduced Folate Carrier Polymorphisms in Osteoporosis Development. *Herba Pol.* 65, 30–36. doi:10.2478/hepo-2019-0011

Newman, M. E. J. (2006). Modularity and Community Structure in Networks. *Proc. Natl. Acad. Sci.* 103, 8577–8582. doi:10.1073/pnas.0601602103

Newman, M. E. J. (2003). The Structure and Function of Complex Networks. *SIAM Rev.* 45, 167–256. doi:10.1137/s003614450342480

Onnela, J. P., Saramäki, J., Kertész, J., and Kaski, K. (2005). Intensity and Coherence of Motifs in Weighted Complex Networks. *Phys. Rev. E Stat. Nonlin Soft Matter Phys.* 71, 065103. doi:10.1103/PhysRevE.71.065103

Ou, X., Andres, A., Pivik, R. T., Cleves, M. A., and Badger, T. M. (2015). Brain gray and white Matter Differences in Healthy normal Weight and Obese Children. *J. Magn. Reson. Imaging* 42, 1205–1213. doi:10.1002/jmri.24912

Palmer, N. D., Goodarzi, M. O., Langefeld, C. D., Wang, N., Guo, X., Taylor, K. D., et al. (2015). Genetic Variants Associated with Quantitative Glucose Homeostasis Traits Translate to Type 2 Diabetes in Mexican Americans: The Guardian (Genetics Underlying Diabetes in Hispanics) Consortium. *Diabetes* 64, 1853–1866. doi:10.2337/db14-0732

Peven, J. C., Litz, G. A., Brown, B., Xie, X., Grove, G. A., Watt, J. C., et al. (2019). Higher Cardiorespiratory Fitness Is Associated with Reduced Functional Brain Connectivity during Performance of the Stroop Task. *Brain Plast.* 5, 57–67. doi:10.3233/BPL-190085

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). Plink: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* 81, 559–575. doi:10.1086/519795

Rapuano, K. M., Huckins, J. F., Sargent, J. D., Heatherton, T. F., and Kelley, W. M. (2016). Individual Differences in Reward and Somatosensory-Motor Brain Regions Correlate with Adiposity in Adolescents. *Cereb. Cortex* 26, 2602–2611. doi:10.1093/cercor/bhv097

Rubinov, M., and Sporns, O. (2010). Complex Network Measures of Brain Connectivity: Uses and Interpretations. *Neuroimage* 52, 1059–1069. doi:10.1016/j.neuroimage.2009.10.003

Rudie, J. D., Brown, J. A., Beck-Pancer, D., Hernandez, L. M., Dennis, E. L., Thompson, P. M., et al. (2013). Altered Functional and Structural Brain Network Organization in Autism. *NeuroImage: Clin.* 2, 79–94. doi:10.1016/j.nicl.2012.11.006

Sala-Llonch, R., Bartrés-Faz, D., and Junqué, C. (2015). Reorganization of Brain Networks in Aging: A Review of Functional Connectivity Studies. *Front. Psychol.* 6, 663. doi:10.3389/fpsyg.2015.00663

Salzwedel, A. P., Gao, W., Andres, A., Badger, T. M., Glasier, C. M., Ramakrishnaiah, R. H., et al. (2019). Maternal Adiposity Influences Neonatal Brain Functional Connectivity. *Front. Hum. Neurosci.* 12, 514. doi:10.3389/fnhum.2018.00514

Saykin, A. J., Shen, L., Foroud, T. M., Potkin, S. G., Swaminathan, S., Kim, S., et al. (2010). Alzheimer's Disease Neuroimaging Initiative Biomarkers as

Quantitative Phenotypes: Genetics Core Aims, Progress, and Plans. *Alzheimer's Demen.* 6, 265–273. doi:10.1016/j.jalz.2010.03.013

Saykin, A. J., Shen, L., Yao, X., Kim, S., Nho, K., Risacher, S. L., et al. (2015). Genetic Studies of Quantitative Mci and Ad Phenotypes in Adni: Progress, Opportunities, and Plans. *Alzheimer's Demen.* 11, 792–814. doi:10.1016/j.jalz.2015.05.009

Shen, L., and Thompson, P. M. (2020). Brain Imaging Genomics: Integrated Analysis and Machine Learning. *Proc. IEEE* 108, 125–162. doi:10.1109/jproc.2019.2947272

Shu, N., Liu, Y., Duan, Y., and Li, K. (2015). Hemispheric Asymmetry of Human Brain Anatomical Network Revealed by Diffusion Tensor Tractography. *Biomed. Research International* 2015, 908917. doi:10.1155/2015/908917

Smith, R. E., Tournier, J.-D., Calamante, F., and Connelly, A. (2015). Sift2: Enabling Dense Quantitative Assessment of Brain white Matter Connectivity Using Streamlines Tractography. *Neuroimage* 119, 338–351. doi:10.1016/j.neuroimage.2015.06.092

Sporns, O. (2013). Network Attributes for Segregation and Integration in the Human Brain. *Curr. Opin. Neurobiol.* 23, 162–171. doi:10.1016/j.conb.2012.11.015

Stenholm, S., Head, J., Aalto, V., Kivimäki, M., Kawachi, I., Zins, M., et al. (2017). Body Mass index as a Predictor of Healthy and Disease-free Life Expectancy between Ages 50 and 75: A Multicohort Study. *Int. J. Obes.* 41, 769–775. doi:10.1038/ijo.2017.29

Sun, Y., Chen, Y., Collinson, S. L., Bezerianos, A., and Sim, K. (2017). Reduced Hemispheric Asymmetry of Brain Anatomical Networks Is Linked to Schizophrenia: A Connectome Study. *Cereb. Cortex* 27, 602–615. doi:10.1093/cercor/bhv255

Thompson, P. M., Ge, T., Glahn, D. C., Jahanshad, N., and Nichols, T. E. (2013). Genetics of the Connectome. *Neuroimage* 80, 475–488. doi:10.1016/j.neuroimage.2013.05.013

Tian, L., Wang, J., Yan, C., and He, Y. (2011). Hemisphere- and Gender-Related Differences in Small-World Brain Networks: A Resting-State Functional MRI Study. *Neuroimage* 54, 191–202. doi:10.1016/j.neuroimage.2010.07.066

Tournier, J.-D., Calamante, F., and Connelly, A. (2012). Mrtrix: Diffusion Tractography in Crossing Fiber Regions. *Int. J. Imaging Syst. Technol.* 22, 53–66. doi:10.1002/ima.22005

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated Anatomical Labeling of Activations in Spm Using a Macroscopic Anatomical Parcellation of the Mni Mri Single-Subject Brain. *Neuroimage* 15, 273–289. doi:10.1006/nimg.2001.0978

van den Heuvel, M. P., Scholtens, L. H., de Lange, S. C., Pijnenburg, R., Cahn, W., van Haren, N. E. M., et al. (2019). Evolutionary Modifications in Human Brain Connectivity Associated with Schizophrenia. *Brain* 142, 3991–4002. doi:10.1093/brain/awz330

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., Ugurbil, K., et al. (2013). The Wu-Minn Human Connectome Project: An Overview. *Neuroimage* 80, 62–79. doi:10.1016/j.neuroimage.2013.05.041

Varangis, E., Habeck, C. G., Razlighi, Q. R., and Stern, Y. (2019). The Effect of Aging on Resting State Connectivity of Predefined Networks in the Brain. *Front. Aging Neurosci.* 11, 234. doi:10.3389/fnagi.2019.00234

Veit, R., Kullmann, S., Heni, M., Machann, J., Häring, H.-U., Fritsche, A., et al. (2014). Reduced Cortical Thickness Associated with Visceral Fat and Bmi. *NeuroImage: Clin.* 6, 307–311. doi:10.1016/j.nicl.2014.09.013

Xie, L., Amico, E., Salama, P., Wu, Y.-c., Fang, S., Sporns, O., et al. (2018). "Heritability Estimation of Reliable Connectomic Features," in *International Workshop on Connectomics in Neuroimaging*. Editors G. Wu, I. Rekik, M. Schirmer, A. Chung, and B. Munsell (Cham: Springer), 58–66. doi:10.1007/978-3-030-00755-3_7

Yan, J., Liu, K., Lv, H., Amico, E., Risacher, S. L., Wu, Y.-C., et al. (2018). "Joint Exploration and Mining of Memory-Relevant Brain Anatomic and Connectomic Patterns via a Three-Way Association Model," in 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, United States, April 4–7, 2018 (IEEE), 6–9. doi:10.1109/isbi.2018.8363511

Yao, X., Cong, S., Yan, J., Risacher, S. L., Saykin, A. J., Moore, J. H., et al. (2020). Regional Imaging Genetic Enrichment Analysis. *Bioinformatics* 36, 2554–2560. doi:10.1093/bioinformatics/btz948

Yao, X., Risacher, S. L., Nho, K., Saykin, A. J., Wang, Z., Shen, L., et al. (2019). Targeted Genetic Analysis of Cerebral Blood Flow Imaging Phenotypes Implicates the Inpp5d Gene. *Neurobiol. Aging* 81, 213–221. doi:10.1016/j.neurobiolaging.2019.06.003

Yokum, S., Ng, J., and Stice, E. (2012). Relation of Regional gray and white Matter Volumes to Current Bmi and Future Increases in Bmi: a Prospective Mri Study. *Int. J. Obes.* 36, 656–664. doi:10.1038/ijo.2011.175

Zhao, T., Cao, M., Niu, H., Zuo, X.-N., Evans, A., He, Y., et al. (2015). Age-Related Changes in the Topological Organization of the white Matter Structural Connectome across the Human Lifespan. *Hum. Brain Mapp.* 36, 3777–3792. doi:10.1002/hbm.22877

Zhong, S., He, Y., Shu, H., and Gong, G. (2017). Developmental Changes in Topological Asymmetry between Hemispheric Brain white Matter Networks from Adolescence to Young Adulthood. *Cereb. Cortex* 27, 2560–2570. doi:10.1093/cercor/bhw109

Zuo, X.-N., Ehmke, R., Mennes, M., Imperati, D., Castellanos, F. X., Sporns, O., et al. (2012). Network Centrality in the Human Functional Connectome. *Cereb. Cortex* 22, 1862–1875. doi:10.1093/cercor/bhr269

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 11 February 2022
doi: 10.3389/fgene.2022.822261

A Pan-Cancer Analysis Reveals the Prognostic and Immunotherapeutic Value of ALKBH7

Kaijie Chen[1,2†], Dongjie Shen[3†], Lin Tan[4], Donglin Lai[1,2], Yuru Han[1,2], Yonggang Gu[5], Changlian Lu[1*] and Xuefeng Gu[1,2,6*]*

[1]Shanghai Key Laboratory of Molecular Imaging, Zhoupu Hospital, Shanghai University of Medicine and Health Sciences, Shanghai, China, [2]School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, China, [3]Department of General Surgery, Ruijin Hospital Lu Wan Branch, Shanghai Jiaotong University School of Medicine, Shanghai, China, [4]Xiangya School of Medicine, The Affiliated Zhuzhou Hospital Xiangya Medical College CSU, Central South University, Changsha, China, [5]Department of TCM, Shanghai Pudong Hospital, Shanghai, China, [6]School of Pharmacy, Shanghai University of Medicine and Health Sciences, Shanghai, China

OPEN ACCESS

Edited by:
*Guiyou Liu,
Tianjin Institute of Industrial Biotechnology (CAS), China*

Reviewed by:
*Li Hongdong,
Gannan Medical University, China
Hailin Zhao,
Imperial College London, United Kingdom*

***Correspondence:**
*Xuefeng Gu
guxf@sumhs.edu.cn
Changlian Lu
lvcl@sumhs.edu.cn
Yonggang Gu
gyg0217@qq.com*

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

***Received:** 02 November 2021*
***Accepted:** 14 January 2022*
***Published:** 11 February 2022*

Citation:
Chen K, Shen D, Tan L, Lai D, Han Y, Gu Y, Lu C and Gu X (2022) A Pan-Cancer Analysis Reveals the Prognostic and Immunotherapeutic Value of ALKBH7. Front. Genet. 13:822261. doi: 10.3389/fgene.2022.822261

Recent studies have identified a role for ALKBH7 in the occurrence and progression of cancer, and this protein is related to cellular immunity and immune cell infiltration. However, the prognostic and immunotherapeutic value of ALKBH7 in different cancers have not been explored. In this study, we observed high ALKBH7 expression in 17 cancers and low expression in 5 cancers compared to paired normal tissues. Although ALKBH7 expression did not correlate relatively significantly with the clinical parameters of age (6/33), sex (3/33) and stage (3/27) in the cancers studied, the results of the survival analysis reflect the pan-cancer prognostic value of ALKBH7. In addition, ALKBH7 expression was significantly correlated with the TMB (7/33), MSI (13/33), mDNAsi (12/33) and mRNAsi (13/33) in human cancers. Moreover, ALKBH7 expression was associated and predominantly negatively correlated with the expression of immune checkpoint (ICP) genes in many cancers. Furthermore, ALKBH7 correlated with infiltrating immune cells and ESTIMATE scores, especially in PAAD, PRAD and THCA. Finally, the ALKBH7 gene coexpression network is involved in the regulation of cellular immune, oxidative, phosphorylation, and metabolic pathways. In conclusion, ALKBH7 may serve as a potential prognostic pan-cancer biomarker and is involved in the immune response. Our pan-cancer analysis provides insight into the role of ALKBH7 in different cancers.

Keywords: ALKBH7, pan-cancer, prognosis, immunotherapy, immune infiltration

INTRODUCTION

The AlkB family consists of Fe (II) and α-ketoglutarate-dependent dioxygenases. Nine AlkB homologues have been identified, including ALKBH1-8 and FTO. Previous experimental studies have found that these proteins are involved in biological processes such as RNA modification and fatty acid metabolism and the DNA damage response (Wu et al., 2016; Bian et al., 2019; Rajecka et al., 2019). In addition, recent studies have also discovered the potential of the AlkB family to participate in immune responses. ALKBH5 regulates the immune response by controlling CD4$^+$ T cells (Zhou et al., 2021), regulating lactic acid levels (Li et al., 2020), and regulating HMGB1 expression (Chen et al., 2021). Moreover, many studies have reported a role for the AlkB family in the development of BLCA, HNSC, LUAD and OV, and this family is involved in regulating the immune response (Fujii et al., 2013; Pilžys et al., 2019; Cai et al., 2021; Wu et al., 2021), suggesting that AlkB homologues may be promising therapeutic targets.

ALKBH7 is a member of the AlkB family. Multiple studies have shown that ALKBH7 participates in biological processes such as lipid metabolism and programmed necrosis (Wang et al., 2014). ALKBH7 deficiency increases body weight and body fat in mice (Solberg et al., 2013) and protects mouse hearts from ischaemia-reperfusion (IR) injury (Kulkarni et al., 2020). ALKBH7 plays a key role in the process of alkylation and oxidation-induced programmed necrosis (Fu et al., 2013) and drives tissue- and sex-specific necrotic cell death responses (Jordan et al., 2017). In addition, recent studies have identified a role for ALKBH7 in the progression of several cancers and its relationship with immune cell infiltration. ALKBH7 expression is significantly elevated in hepatocellular carcinoma and negatively correlates with CD4$^+$ cells, macrophages and neutrophils (Peng et al., 2021). ALKBH7 is associated with overall survival in individuals with lung adenocarcinoma and negatively correlates with CD8$^+$ T cells and macrophages (Wu et al., 2021). ALKBH7 correlates with the pathological stage of ovarian serous carcinoma and positively correlates with the infiltration of CD8$^+$ T cells, dendritic cells and neutrophils (Cai et al., 2021). ALKBH7 is involved in cellular immunity and the proliferation of HeLa cervical cancer cell lines (Meng et al., 2019).

However, the prognostic value and immunological role of ALKBH7 in cancer have not been systematically investigated. In the present study, we explored changes in the expression and prognostic value of ALKBH7 in 33 cancers. Then, we investigated ALKBH7 expression in different cancer immune and molecular subtypes. In addition, we performed an in-depth study of the immune mechanism of ALKBH7 in different cancers to explore its potential immunotherapeutic value. Overall, this work provides evidence to elucidate the immunotherapeutic role of ALKBH7 in cancer, which may be helpful for further functional experiments.

MATERIALS AND METHODS

Data Acquisition and Software Availability

The genomic and clinicopathological information, somatic mutation and stemness score data of 33 cancers were obtained from TCGA (https://cancergenome.nih.gov/) and UCSC Xena (https://xena.ucsc.edu/) database (Goldman et al., 2020). In addition, to obtain more normal tissue genomic data, we downloaded tumor and normal tissue gene expression data combined with TCGA and GTEx database on the UCSCXenaShiny (https://shiny.hiplot.com.cn/ucsc-xena-shiny/) website (Wang S. et al., 2021). R 4.1.0 was used to integrate, analyse the original data and visualize the results.

Differential ALKBH7 Expression Analysis in the Normal, Tumor, Various Age, Gender, and Pathological Stage Tissues

The discrepancy of the gene expression between various types of cancer and paired normal tissues was investigated to explore whether ALKBH7 is associated with cancer development. Differential expression analysis of ALKBH7 has been investigated in a variety of cancers with patients' age, gender and pathological stage by using wilcox test.

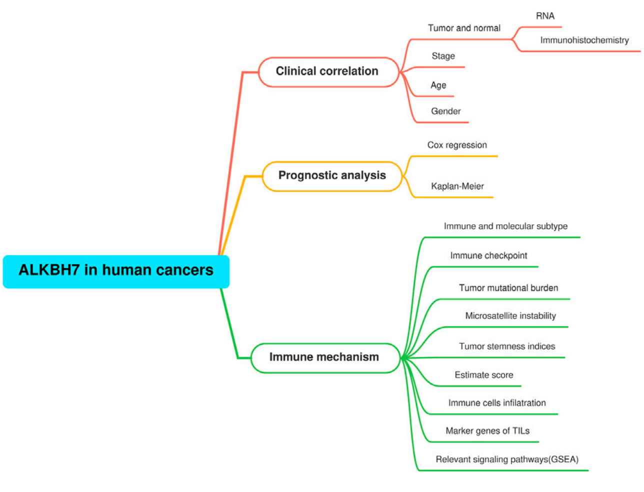


FIGURE 1 | The analysis and indicators employed in our research. In clinical correlation section, differential ALKBH7 expression analyses were performed between different tissues (tumor versus normal), ages (≤60 *versus* >60), genders (male versus female), stages (stage I + II versus stage III + IV). Prognostic analysis was based on univariate Cox regression and Kaplan-Meier survival curve. In immune mechanism section, relevant signaling pathways were explored by GSEA based on the ALKBH7 expression.

Immunohistochemical Staining

The Human Protein Atlas (https://www.proteinatlas.org/) contains over 25,000 antibodies and a collection of over 10 million immunohistochemical (IHC) images (Thul and Lindskog, 2018). To further compare the expression of ALKBH7 gene in tumors and corresponding normal tissues, antibody-based ALKBH7 protein profiles using immunohistochemistry were obtained from the HPA database.

Prognostic Analysis of ALKBH7 in Human Cancers

Univariate Cox regression analysis and Kaplan-Meier curve were used to analyse the relationship between ALKBH7 expression and clinical survival data including overall survival (OS), disease-specific survival (DSS), disease-free interval (DFI) and progression-free interval (PFI) in 33 cancers.

Analysis of ALKBH7 Expression in Different Subtypes of Human Cancers

The TISIDB database (http://cis.hku.hk/TISIDB/) is an online integrated repository portal integrating multiple types of data resources in oncoimmunology (Ru et al., 2019). The relationship between ALKBH7 expression and immune or molecular subtypes of different cancer types was explored through the TISIDB database.

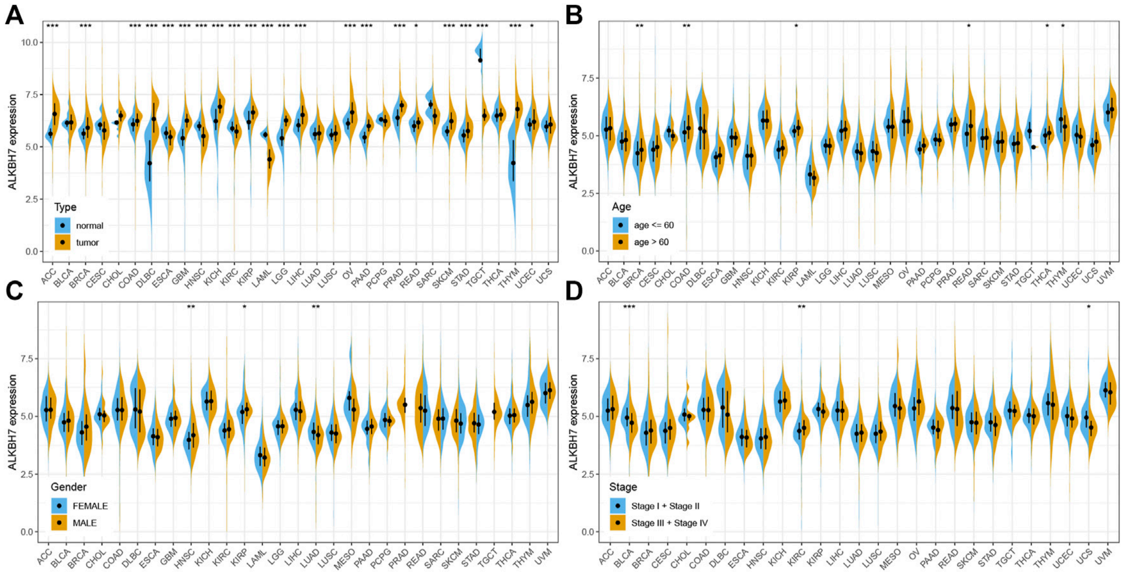


FIGURE 2 | The clinical correlation of ALKBH7 expression. **(A)** Differential expression of ALKBH7 in normal and tumor samples from patients with 33 cancers; the correlations of ALKBH7 with age **(B)**, sex **(C)** and stage **(D)** in 33 cancers. “*” indicates $p < 0.05$, “**” indicates $p < 0.01$ and “***” indicates $p < 0.001$.

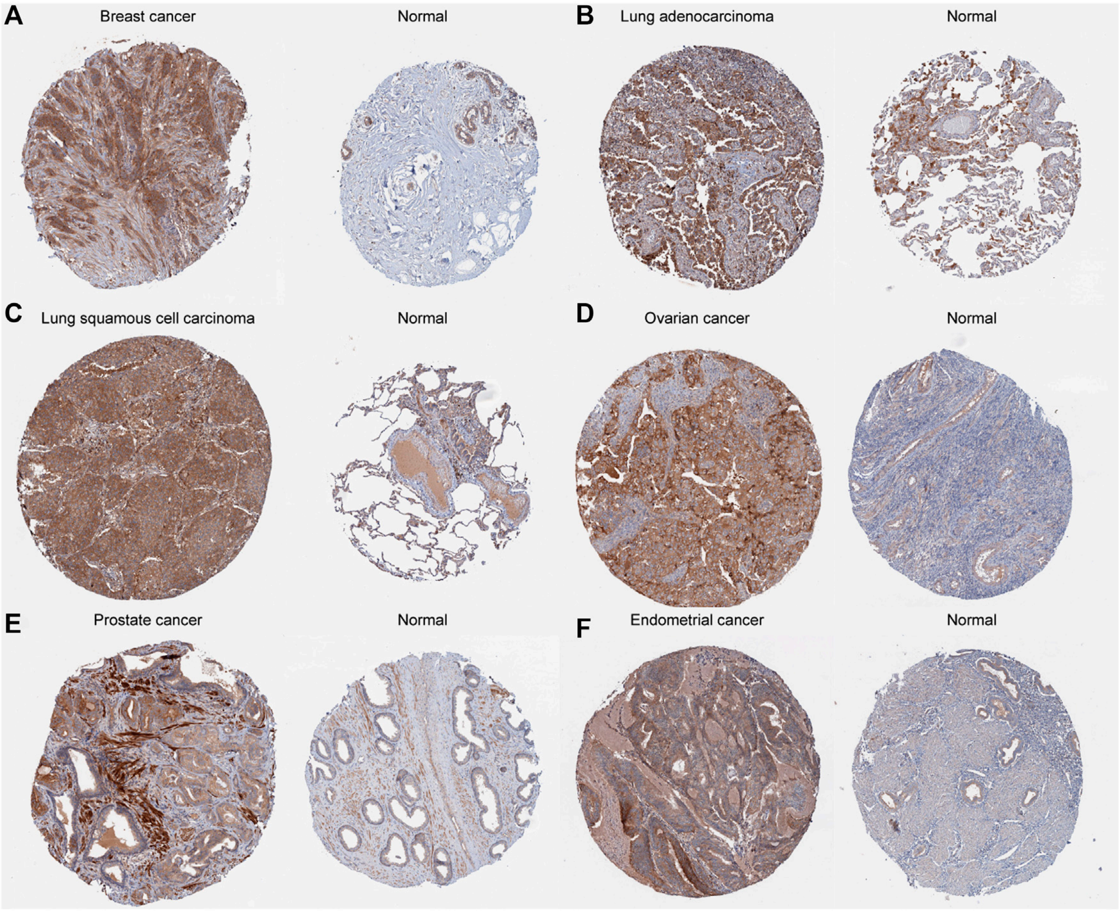
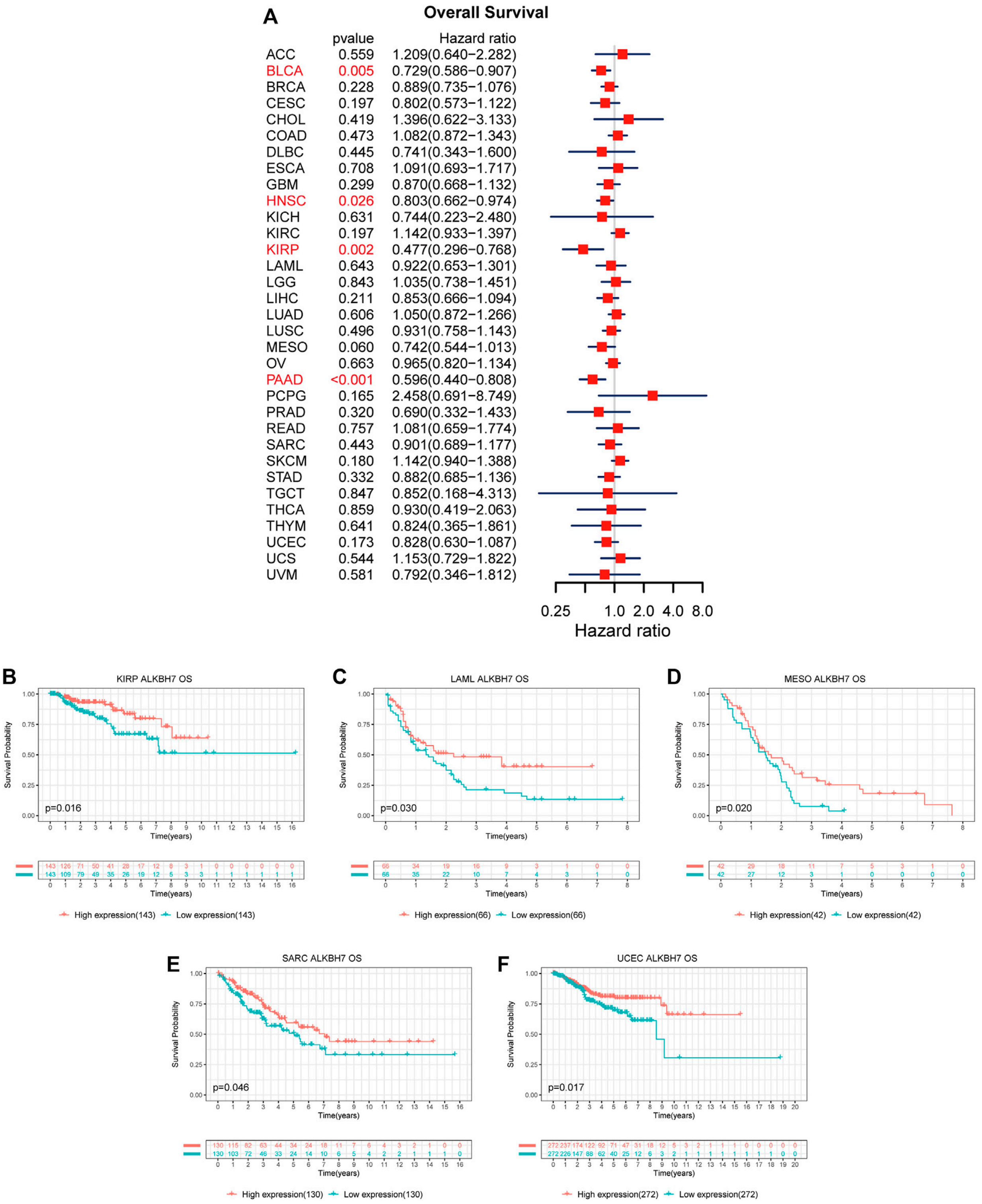


FIGURE 3 | Representative ALKBH7 immunohistochemical staining in tumor and normal tissues. The expression of ALKBH7 gene in BRCA **(A)**, LUAD **(B)**, LUSC **(C)**, OV **(D)**, PRAD **(E)**, and UCEC **(F)** is significantly higher than that in the corresponding normal tissues.



	pvalue	Hazard ratio
ACC	0.559	1.209(0.640−2.282)
BLCA	0.005	0.729(0.586−0.907)
BRCA	0.228	0.889(0.735−1.076)
CESC	0.197	0.802(0.573−1.122)
CHOL	0.419	1.396(0.622−3.133)
COAD	0.473	1.082(0.872−1.343)
DLBC	0.445	0.741(0.343−1.600)
ESCA	0.708	1.091(0.693−1.717)
GBM	0.299	0.870(0.668−1.132)
HNSC	0.026	0.803(0.662−0.974)
KICH	0.631	0.744(0.223−2.480)
KIRC	0.197	1.142(0.933−1.397)
KIRP	0.002	0.477(0.296−0.768)
LAML	0.643	0.922(0.653−1.301)
LGG	0.843	1.035(0.738−1.451)
LIHC	0.211	0.853(0.666−1.094)
LUAD	0.606	1.050(0.872−1.266)
LUSC	0.496	0.931(0.758−1.143)
MESO	0.060	0.742(0.544−1.013)
OV	0.663	0.965(0.820−1.134)
PAAD	<0.001	0.596(0.440−0.808)
PCPG	0.165	2.458(0.691−8.749)
PRAD	0.320	0.690(0.332−1.433)
READ	0.757	1.081(0.659−1.774)
SARC	0.443	0.901(0.689−1.177)
SKCM	0.180	1.142(0.940−1.388)
STAD	0.332	0.882(0.685−1.136)
TGCT	0.847	0.852(0.168−4.313)
THCA	0.859	0.930(0.419−2.063)
THYM	0.641	0.824(0.365−1.861)
UCEC	0.173	0.828(0.630−1.087)
UCS	0.544	1.153(0.729−1.822)
UVM	0.581	0.792(0.346−1.812)

FIGURE 4 | Associations between ALKBH7 expression and OS of patients with cancer. **(A)** Forest plot showing the hazard ratios of ALKBH7 in 33 cancers; Kaplan-Meier survival curves of OS forpatients stratified according to different ALKBH7 expression profiles in KIRP **(B)**, LAML **(C)**, MESO **(D)**, SARC **(E)** and UCEC **(F)**.

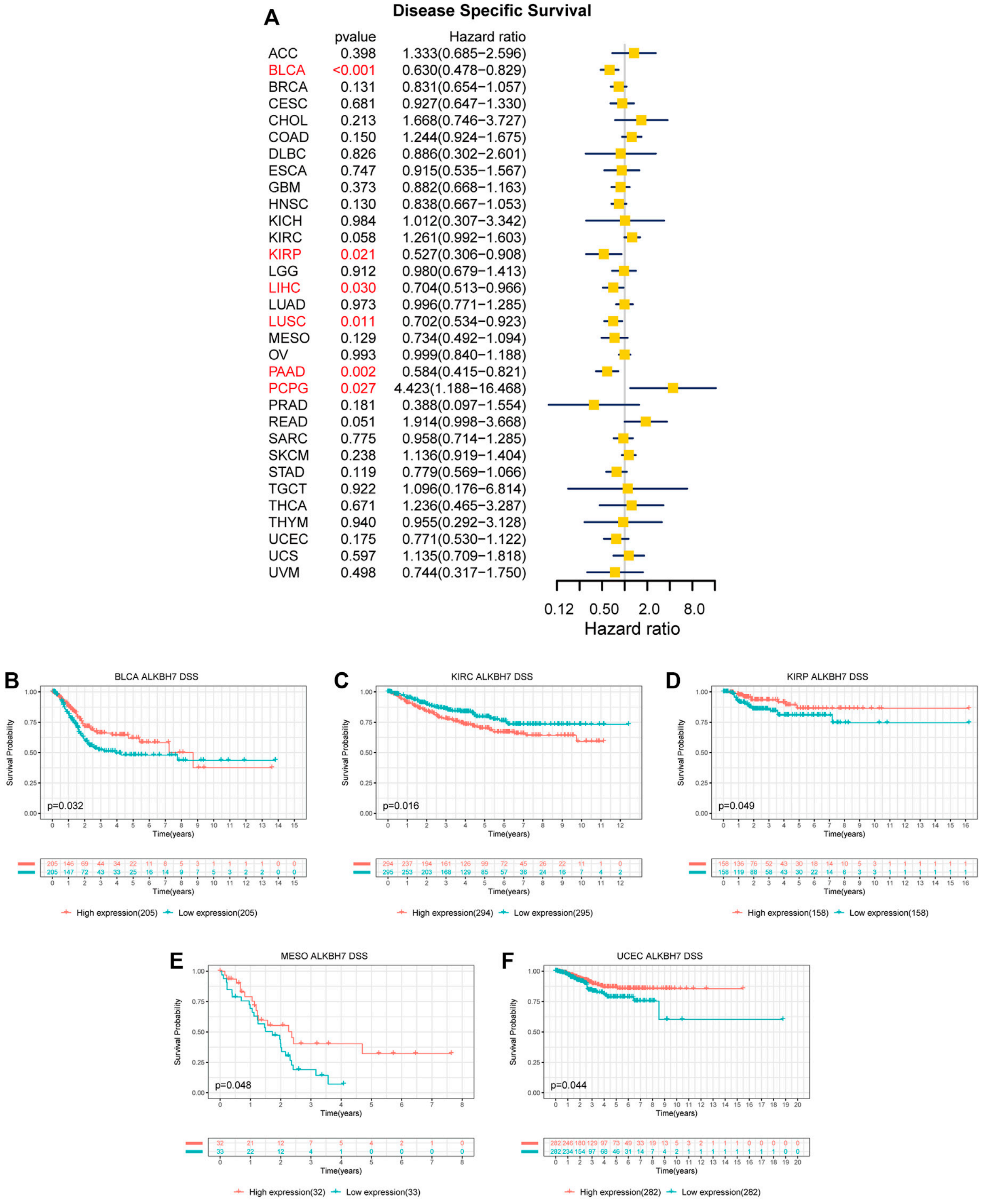


Figure 5

FIGURE 5 | Associations between ALKBH7 expression and DSS of patients with cancer. **(A)** Forest plot showing hazard ratios of ALKBH7 in 32 cancers; Kaplan-Meier survival curves of DSS forpatients stratified according to different ALKBH7 expression profiles in BLCA **(B)**, KIRC **(C)**, KIRP **(D)**, MESO **(E)** and UCEC **(F)**.

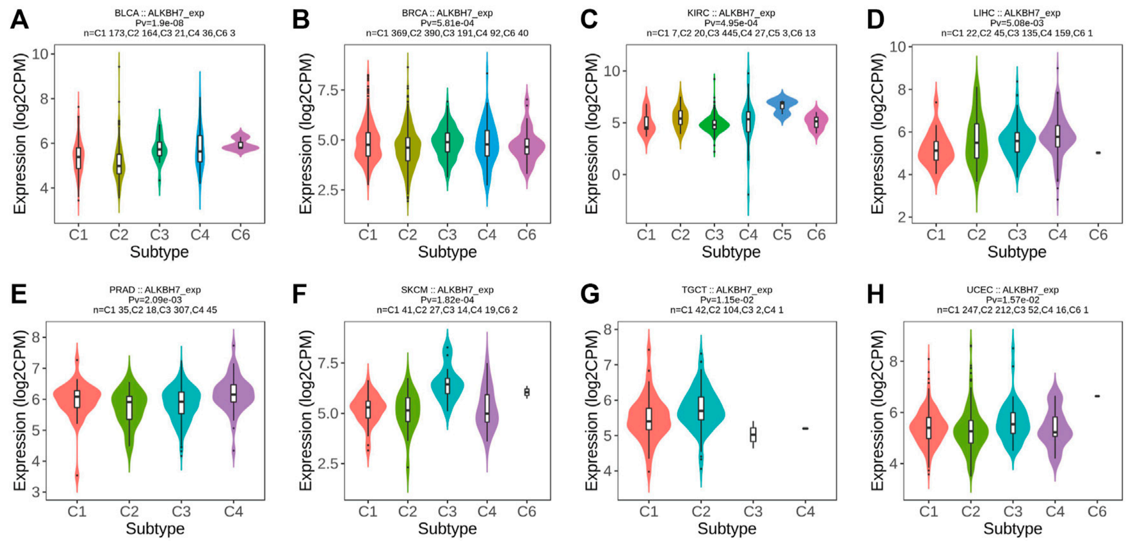


FIGURE 6 | The relationship between ALKBH7 expression and immune subtypes in BLCA **(A)**, BRCA **(B)**, KIRC **(C)**, LIHC **(D)**, PRAD **(E)**, SKCM **(F)**, TGCT **(G)** and UCEC **(H)**. [C1 (wound healing); C2 (IFN-gamma dominant); C3 (inflammatory); C4 (lymphocyte depleted); C5 (immunologically quiet); C6 (TGF-beta dominant)].

Correlation Analysis of ALKBH7 Expression With Immune Checkpoint Genes, Tumor Mutational Burden, Microsatellite Instability and Tumor Stemness Index in Human Cancers

The correlation between ALKBH7 expression and the expression of immune checkpoint (ICP) genes, was explored *via* the SangerBox website (http://sangerbox.com/). The tumor mutational burden (TMB), microsatellite instability (MSI) score and tumor stemness index of each TCGA tumor case were obtained from somatic mutation data and previously published studies respectively (Topalian et al., 2015; Bonneville et al., 2017). Tumor stemness indices are indicators for assessing the degree of oncogenic dedifferentiation. Among them, mRNAsi is a gene expression-based stemness index while mDNAsi is a DNA methylation-based stemness index. Correlations between ALKBH7 expression and TMB, MSI, mRNAsi and mDNAsi were analyzed using Spearman's method.

Analysis of Immune Infiltration-Related Factors and Pathways

The TIMER database (https://cistrome.shinyapps.io/timer/), which collected 10,897 samples across 32 cancer types from TCGA, was created to analyze the level of tumor-associated immune cell infiltration in the TME (Li et al., 2016; Li et al., 2017). The correlation between ALKBH7 expression and six immune cells (B cells, $CD4^+$ T cells, $CD8^+$ T cells, neutrophils, macrophages and dendritic cells) and tumor-infiltrating lymphocyte (TIL) marker genes in human cancers was investigated using the TIMER database. ESTIMATE is an algorithm that predicts the presence of immune and stromal cells in tumor tissue which based on gene expression profiles (Yoshihara et al., 2013). We calculated the stromal score, immune score and estimate socre of each case by using the ESTIMATE package. xCell is a powerful web tool for inferring the proportion of immune cell subtypes in tumor tissue (Aran et al., 2017). A spearman correlation heat map of ALKBH7 expression with 36 immunoinfiltrating subtypes of cells in human cancers was established. Finally, to further investigate the relevant signalling pathways, gene set enrichment analysis (GSEA) was performed to explore pathways of ALKBH7 coexpression gene network.

RESULTS

Clinical Landscape of ALKBH7 Expression in 33 Cancers

The details of the analysis are summarized and presented in **Figure 1** for a more comprehensive perspective. As illustrated in **Figure 2A**, significantly higher ALKBH7 expression was detected in most human cancers than in adjacent normal tissues, such as ACC, BRCA, COAD, DLBC, GBM, KICH, KIRP, LGG, LIHC, OV, PAAD, PRAD, READ, SKCM, STAD, THYM and UCEC. In contrast, significantly lower ALKBH7 expression was observed in a few human cancers (ESCA, HNSC, KIRC, LAML and TGCT). ALKBH7 was highly differentially expressed among elderly patients in the THCA, BRCA, KIRP, READ and COAD groups, whereas it was weakly expressed in patients with THYM (**Figure 2B**). Meanwhile, the results indicated significant sex-based differences in ALKBH7 expression in HNSC, KIRP and LUAD (**Figure 2C**). In addition, ALKBH7 expression was significantly correlated with the pathological stage of some cancers, including BLCA, KIRC and UCS (**Figure 2D**). Finally, we used immunohistochemistry to validate ALKBH7 expression. Compared with normal tissues, ALKBH7 was

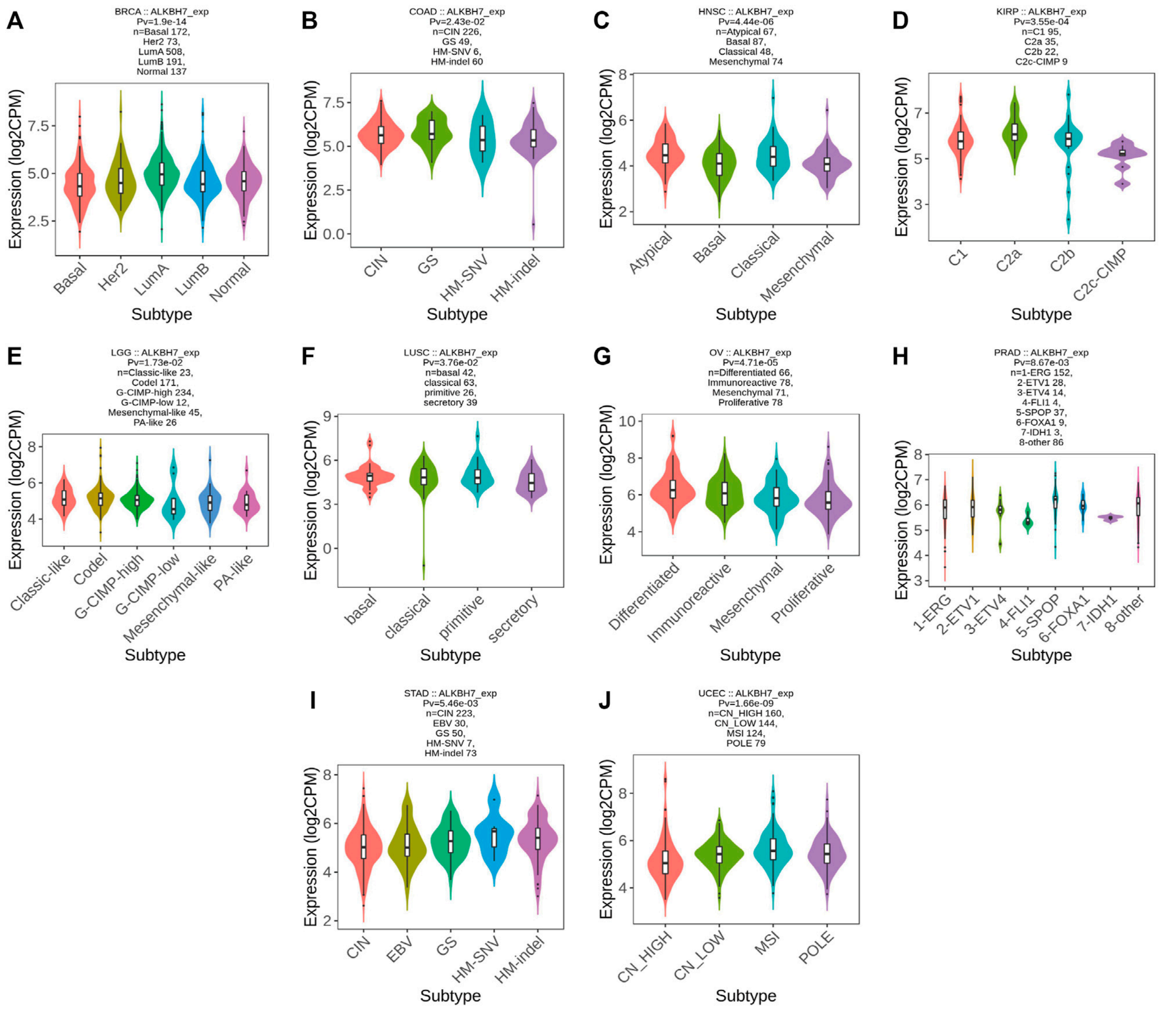


FIGURE 7 | The relationship between ALKBH7 expression and molecular subtypes in BRCA **(A)**, COAD **(B)**, HNSC **(C)**, KIRP **(D)**, LGG **(E)**, LUSC **(F)**, OV **(G)**, PRAD **(H)**, STAD **(I)** and UCEC **(J)**.

highly expressed in BRCA, LUAD, LUSC, OV, PRAD and UCEC (**Figure 3**).

Pan-Cancer Analysis of the Multifaceted Prognostic Value of ALKBH7

The association between ALKBH7 expression and patient prognosis was estimated in the pan-cancer dataset. The survival metrics included OS, DSS, DFI, and PFI. Univariate Cox regression analysis of the results from 33 types of cancer suggested that ALKBH7 expression significantly correlated with OS in 4 types of cancer, including BLCA, HNSC, KIRP, and PAAD. Kaplan–Meier survival curves indicated that downregulated ALKBH7 expression was remarkably associated with shorter OS of patients with KIRP, LAML, MESO, SARC, and UCEC (**Figure 4**). The relationship between ALKBH7 expression and DSS in patients with cancer was examined. ALKBH7 expression affected DSS in six types of cancer, including BLCA, KIRP, LIHC, LUSC, PAAD, and PCPG. The Kaplan–Meier analysis indicated that decreased ALKBH7 expression indicated shorter DSS of patients with BLCA, KIRP, MESO, and UCEC, while increased ALKBH7 expression corresponded with shorter DSS of patients with KIRC (**Figure 5**). Cox regression analysis of the DFI revealed that ALKBH7 expression significantly correlated with DFI in 4 types of cancer, including LUSC, OV, PAAD, and THCA. The results from the Kaplan–Meier analysis suggested that increased ALKBH7 expression was

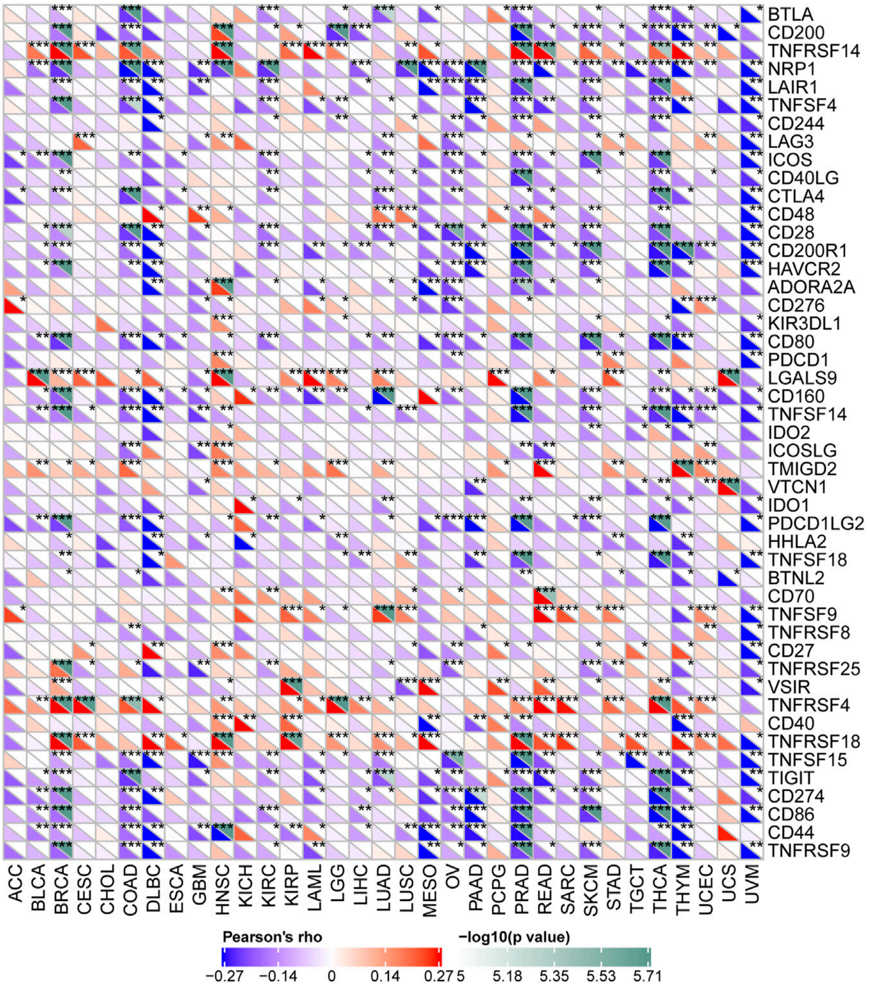


FIGURE 8 | Correlations between the expression of ALKBH7 and immune checkpoint genes in 33 types of cancer. "*" indicates $p < 0.05$, "**" indicates $p < 0.01$ and "***" indicates $p < 0.001$.

associated with a poor prognosis for patients with PRAD, while decreased ALKBH7 expression was associated with a poor prognosis for patients with THCA (**Supplementary Figure S1**). We also assessed the association between ALKBH7 expression and PFI and identified that ALKBH7 expression influenced PFI in patients with BLCA, KIRC, LUSC and PAAD. Kaplan–Meier PFI curves revealed that decreased ALKBH7 mRNA expression correlated with an unfavourable PFI in patients with BLCA and PAAD. In contrast, increased ALKBH7 mRNA expression correlated with an unfavourable PFI in patients with KIRC (**Supplementary Figure S2**).

ALKBH7 Expression Is Related to Immune and Molecular Subtypes in Human Cancers

Based on accumulating evidence, immunophenotyping reflects the comprehensive immune status of a tumor, which is closely related to immunotherapy and the tumor microenvironment (Ma et al., 2021). Different molecular subtypes correspond to the unique molecular biology of cancer and may facilitate the selection of molecular targeted therapies and immunotherapy strategies (Kim et al., 2019; Bai et al., 2021). Next, ALKBH7 expression in immune and molecular subtypes of human cancer was explored using the TISIDB website. ALKBH7 expression was significantly different in different immune subtypes of BLCA, BRCA, KIRC, LIHC, PRAD, SKCM, TGCT, and UCEC (**Figure 6**). In addition, the trends for the up- and downregulation of ALKBH7 expression were also different in different immune subtypes of a specific cancer type. Taking SKCM as an example, low ALKBH7 expression was detected in C2 and C4 types and high expression was observed in the C3 type. Regarding different molecular subtypes of cancers, a significant correlation with ALKBH7 expression was observed in BRCA, COAD, HNSC, KIRP, LGG, LUSC, OV, PRAD, STAD and UCEC (**Figure 7**). Based on the results described above, we suggest that ALKBH7 may play an important role in the tumor immune microenvironment and modulate the effect of immunotherapy.

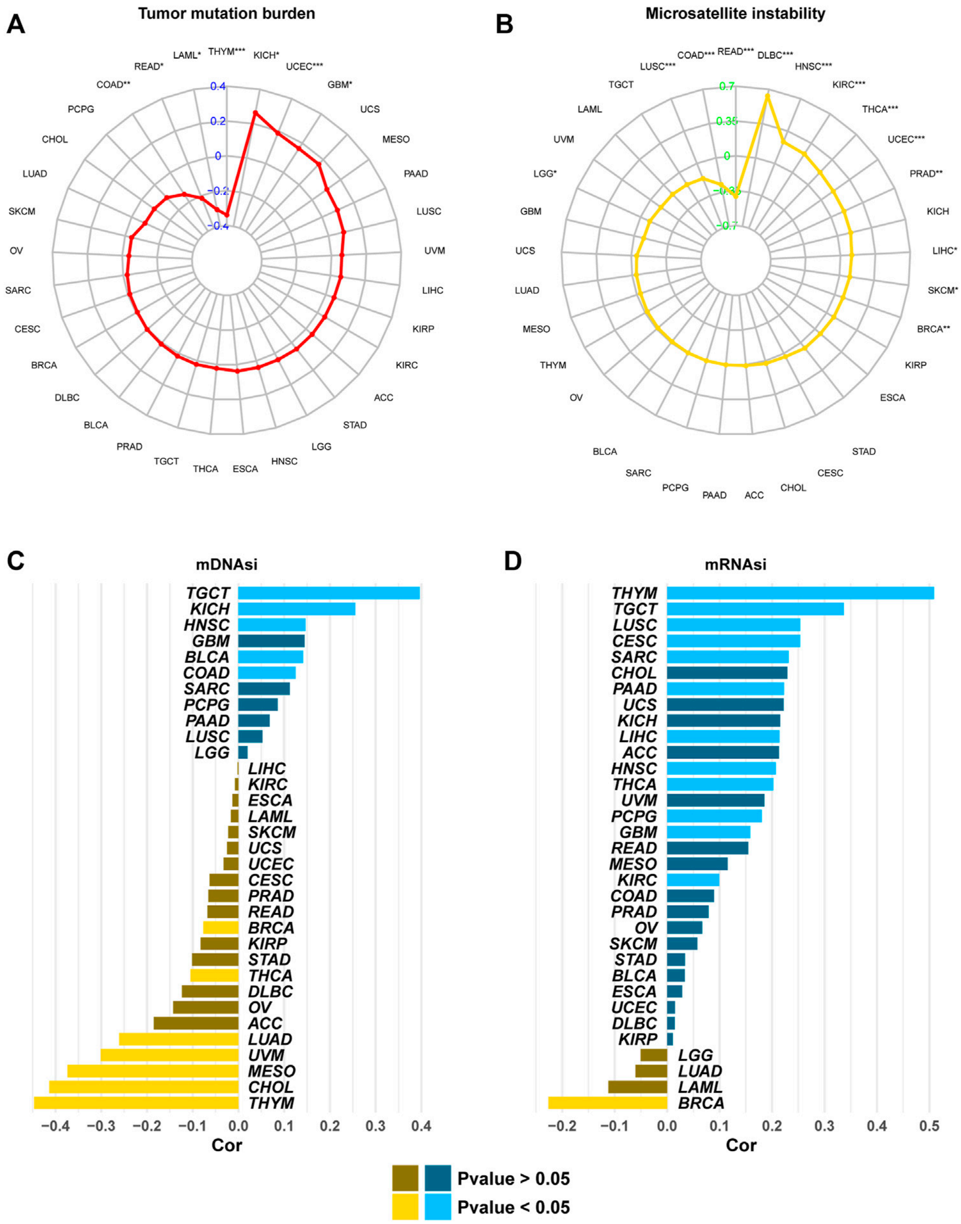


FIGURE 9 | The correlation between ALKBH7 expression and the TMB **(A)**, MSI **(B)**, mDNAsi **(C)**, and mRNAsi **(D).** "*" indicates $p < 0.05$, "**" indicates $p < 0.01$ and "***" indicates $p < 0.001$.

ALKBH7 Expression Is Related to Immune Checkpoint Genes in Human Cancers

Studies have shown that immune checkpoint genes have important implications for immunotherapy in many cancers (Topalian et al., 2015; Li et al., 2019). Here, we collected expression patterns of 47 common immune checkpoint genes and analysed the relationship between ALKBH7 expression and immune checkpoint gene expression to explore the potential role of ALKBH7 in immunotherapy. As shown in **Figure 8**, ALKBH7 expression significantly correlated with the expression of most ICP genes in many cancers, such as BRCA COAD, HNSC, KIRC, LUAD, OV, PAAD, PRAD, READ, SKCM, THCA, THYM, and UVM. Among them, a negative correlation was the main trend; for example, in PRAD, ALKBH7 expression was negatively correlated with the expression of 30 ICP genes and positively correlated with the expression of 5 ICP genes. Thus, high levels of ALKBH7 expression may predict unsatisfactory immunotherapy outcomes when targeting ICP genes. On the other hand, ALKBH7 inhibitors may be potential alternative therapeutic approaches. Therefore, we hypothesized that ALKBH7, a potential pan-cancer biomarker or a novel immunotherapeutic target, may predict the response to immunotherapy or achieve promising therapeutic outcomes, respectively.

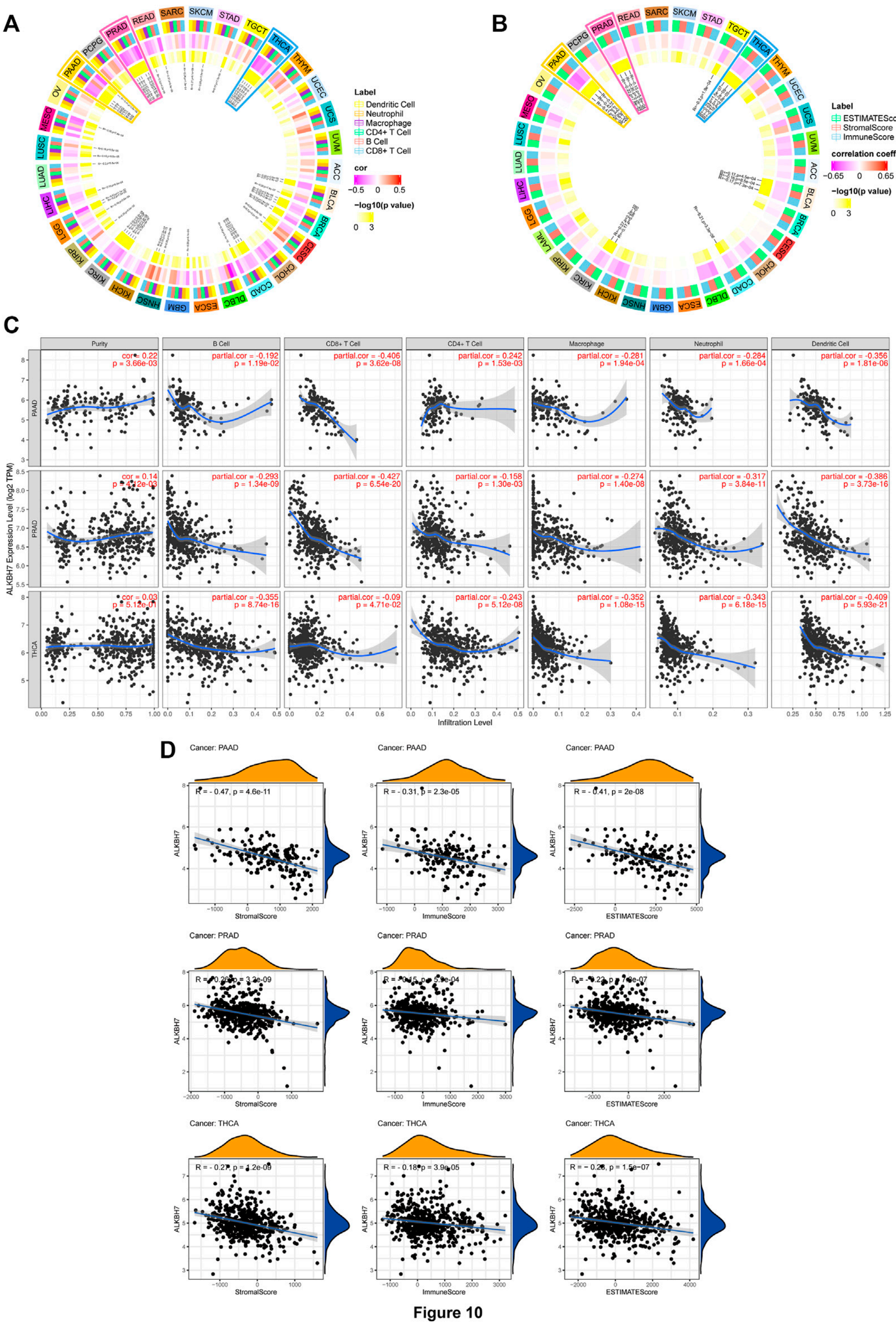


FIGURE 10 | Correlations between ALKBH7 expression and both immune cell infiltration and ESTIMATE score. **(A)** The relationship between the ALKBH7 expression level and numbers of infiltrating B cells, CD4^{+} T cells, CD8^{+} T cells, macrophages, neutrophils, dendritic cell in human cancers. **(B)** The relationship between ALKBH7 expression and the ESTIMATE score in human cancers. **(C)** Correlation of ALKBH7 expression with immune cell infiltration levels in PAAD, PRAD, and THCA. **(D)** Correlation of ALKBH7 expression with ESTIMATE scores in PAAD, PRAD, and THCA.

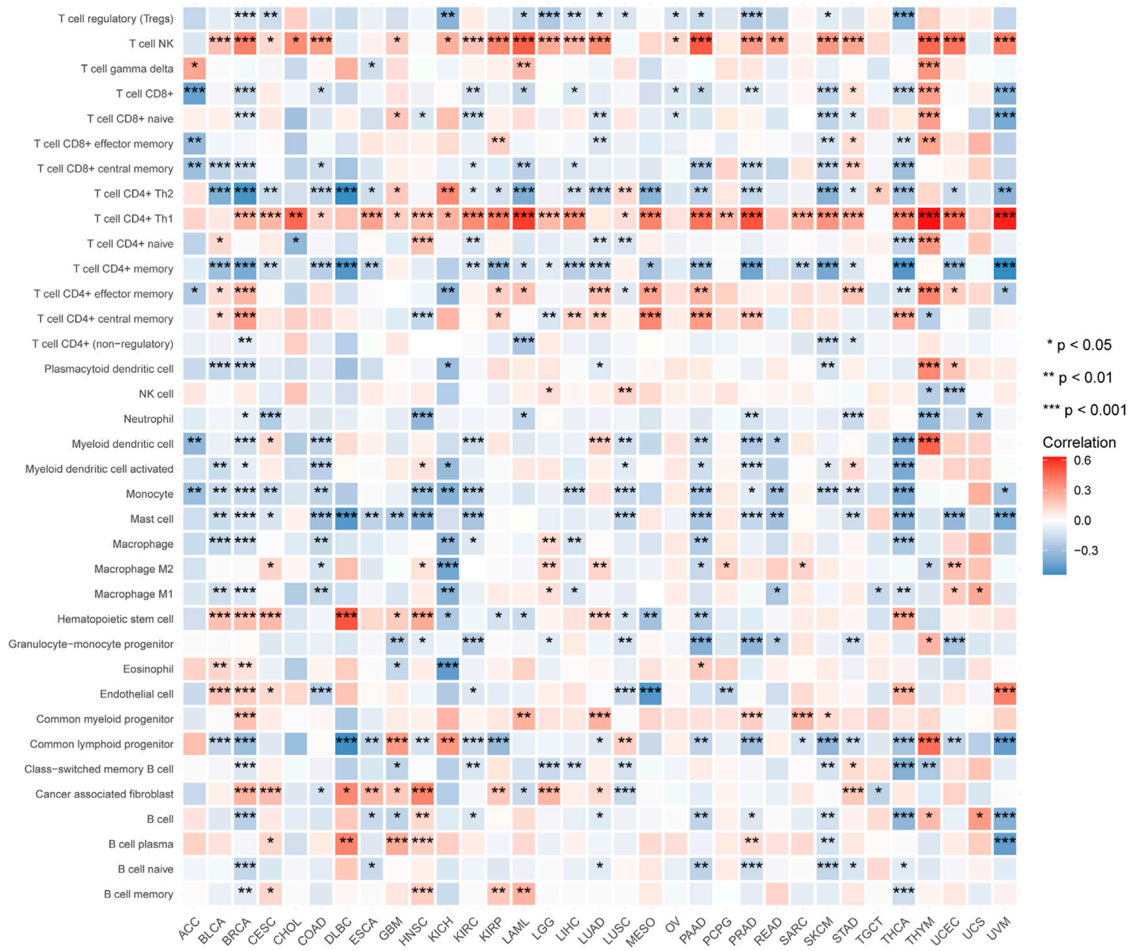


FIGURE 11 | Heat map of the correlations between ALKBH7 expression and immune cell subtypes in 33 types of cancer. "*" indicates $p < 0.05$, "**" indicates $p < 0.01$ and "***" indicates $p < 0.001$.

ALKBH7 Expression Is Related to the Tumor Mutational Burden, Microsatellite Instability, and Tumor Stemness Index

We analysed the correlations between ALKBH7 expression and the TMB, MSI, and tumor stemness index to explore the role of ALKBH7 in the immune mechanism and immune response of the tumor microenvironment (TME). The TMB, MSI, and tumor stemness index in the tumor microenvironment are related to antitumor immunity and might predict the therapeutic efficacy of tumor immunotherapy (Lee et al., 2016; Yarchoan et al., 2017; Malta et al., 2018). As presented in **Figure 9**, ALKBH7 was associated with the TMB in 7 cancers and MSI in 13 cancers. In addition, ALKBH7 was related to mDNAsi in 12 cancers and mRNAsi in 13 cancers. Among them, ALKBH7 expression was negatively correlated with the TMB and MSI in COAD and READ, while it was positively correlated with the TMB and MSI in UCEC. Based on this finding, ALKBH7 might exert an indirect effect on the immunotherapeutic response of COAD, READ and UCEC. ALKBH7 was positively correlated with mRNAsi and mDNAsi in TGCT and HNSC, but negatively correlated with mRNAsi and mDNAsi in BRCA. High ALKBH7 expression in TGCT and HNSC may be related to the low sensitivity to immune checkpoint blockade therapy; in contrast, high ALKBH7 expression in BRCA may be related to the high sensitivity to immune checkpoint blockade therapy. Interestingly, ALKBH7 was positively correlated with mRNAsi but negatively correlated with mDNAsi in THCA and THYM. This result might arise from the discrepancies between mRNAsi and mDNAsi caused by DNA hypermethylation (Malta et al., 2018).

Correlation Analysis Between ALKBH7 Expression and Infiltrating Immune Cells and the ESTIMATE Score

The tumor microenvironment contains immune cells and fibroblasts, which affect the effect of immunotherapy (Aran et al., 2015). We analysed the correlation between ALKBH7 expression and six types of infiltrating immune cells, including

TABLE 1 | Correlation analysis between ALKBH7 and gene markers of immune cells in TIMER.

Description	Gene markers	THCA				PRAD				PAAD			
		None		Purity		None		Purity		None		Purity	
		Cor	Pvalue	Cor	Pvalue	Cor	Pvalue	Cor	Pvalue	Cor	Pvalue	Cor	Pvalue
$CD8^+$ T cell	CD8A	−0.202	[a]	−0.192	[a]	−0.314	[a]	−0.254	[a]	−0.292	[a]	−0.233	[b]
	CD8B	−0.152	[a]	−0.143	[b]	−0.157	[a]	−0.124	[c]	−0.279	[a]	−0.217	[b]
T cell (general)	CD3D	−0.26	[a]	−0.25	[a]	−0.153	[a]	−0.071	0.146	−0.19	[c]	−0.122	0.112
	CD3E	−0.305	[a]	−0.295	[a]	−0.259	[a]	−0.2	[a]	−0.238	[b]	−0.17	[c]
	CD2	−0.316	[a]	−0.303	[a]	−0.253	[a]	−0.16	[b]	−0.292	[a]	−0.224	[b]
B cell	CD19	−0.183	[a]	−0.167	[a]	−0.059	0.187	−0.015	0.759	−0.186	[c]	−0.118	0.123
	CD79A	−0.209	[a]	−0.194	[a]	−0.153	[a]	−0.089	0.068	−0.194	[b]	−0.123	0.109
Monocyte	CD86	−0.404	[a]	−0.393	[a]	−0.414	[a]	−0.336	[a]	−0.399	[a]	−0.332	[a]
	CSF1R	−0.346	[a]	−0.343	[a]	−0.37	[a]	−0.299	[a]	−0.347	[a]	−0.289	[a]
TAM	CCL2	−0.26	[a]	−0.244	[a]	−0.223	[a]	−0.165	[a]	−0.282	[a]	−0.253	[a]
	CD68	−0.358	[a]	−0.34	[a]	−0.35	[a]	−0.289	[a]	−0.228	[b]	−0.157	[c]
	IL10	−0.292	[a]	−0.279	[a]	−0.338	[a]	−0.251	[a]	−0.258	[a]	−0.208	[b]
M1 Macrophage	IRF5	−0.324	[a]	−0.317	[a]	−0.234	[a]	−0.263	[a]	0.027	0.719	0.058	0.454
	PTGS2	−0.319	[a]	−0.31	[a]	−0.268	[a]	−0.179	[a]	−0.237	[b]	−0.253	[a]
M2 Macrophage	CD163	−0.335	[a]	−0.32	[a]	−0.414	[a]	−0.343	[a]	−0.438	[a]	−0.377	[a]
	VSIG4	−0.353	[a]	−0.345	[a]	−0.379	[a]	−0.303	[a]	−0.347	[a]	−0.278	[a]
	MS4A4A	−0.334	[a]	−0.321	[a]	−0.363	[a]	−0.297	[a]	−0.379	[a]	−0.309	[a]
Neutrophils	CEACAM8	−0.196	[a]	−0.197	[a]	0.004	0.922	0.017	0.724	−0.059	0.433	−0.005	0.95
	ITGAM	−0.411	[a]	−0.401	[a]	−0.361	[a]	−0.309	[a]	−0.285	[a]	−0.191	[c]
	CCR7	−0.293	[a]	−0.276	[a]	−0.274	[a]	−0.217	[a]	−0.178	[c]	−0.124	0.106
Dendritic cell	HLA-DPB1	−0.305	[a]	−0.295	[a]	−0.119	[b]	−0.057	0.246	−0.232	[b]	−0.159	[c]
	HLA-DQB1	−0.291	[a]	−0.295	[a]	−0.198	[a]	−0.147	[b]	−0.243	[b]	−0.189	[c]
	HLA-DRA	−0.377	[a]	−0.365	[a]	−0.351	[a]	−0.295	[a]	−0.35	[a]	−0.289	[a]
	HLA-DPA1	−0.364	[a]	−0.351	[a]	−0.339	[a]	−0.257	[a]	−0.356	[a]	−0.3	[a]
	CD1C	−0.336	[a]	−0.319	[a]	−0.319	[a]	−0.242	[a]	−0.199	[b]	−0.14	0.067
	NRP1	−0.258	[a]	−0.245	[a]	−0.29	[a]	−0.271	[a]	−0.471	[a]	−0.439	[a]
	ITGAX	−0.363	[a]	−0.348	[a]	−0.299	[a]	−0.277	[a]	−0.185	[c]	−0.091	0.236

[c] *p < 0.05.*

[b] *p < 0.01.*

[a] *p < 0.001.*

B cells, $CD4^+$ T cells, $CD8^+$ T cells, neutrophils, macrophages, and dendritic cells. The results revealed a significant correlation in 31 cancer types. ALKBH7 expression displayed a strong relationship with dendritic cells in 8 cancer types, macrophages in 9 cancer types, neutrophils in 11 cancer types, $CD8^+$ T cells in 14 cancer types, B cells in 8 cancer types and $CD4^+$ T cells in 3 cancer types (**Figure 10A**). Results from the TIMER database included these results, and all details are shown in **Supplementary Table S1**. Subsequently, the correlation between ALKBH7 expression and stromal, immune and ESTIMATE scores was analysed (**Figure 10B**). All results are presented in **Supplementary Table S2**. Interestingly, the most significant correlation between ALKBH7 expression and the two parameters described above was observed in PAAD, PRAD and THCA. As shown in **Figures 10C,D**, ALKBH7 expression was negatively correlated with TILs and stromal, immune, and ESTIMATE scores. Therefore, ALKBH7 may be involved in inhibiting immune cell infiltration in PAAD, PRAD and THCA.

We also used the xCell web tool to explore the association between ALKBH7 gene expression and the infiltration of various subtypes of immune cells. The Spearman correlation heat map is shown in **Figure 11**. NK T cells and $CD4^+$ Th1 T cells were positively correlated with ALKBH7 gene expression in most cancers. In contrast, $CD4^+$ Th2 T cells, memory $CD4^+$ T cells, monocytes and mast cells were negatively correlated with ALKBH7 gene expression in most cancers. In PAAD, PRAD and THCA, ALKBH7 expression was associated with most subtypes of immune cells and generally exhibited negative correlations.

Correlation Between ALKBH7 Expression and Various Immune Markers

We validated the correlations between ALKBH7 expression and diverse immune signatures in PAAD, PRAD and THCA using the TIMER database to obtain a better understanding of ALKBH7 crosstalk with the immune response. The genes listed in **Table 1** were used to characterize immune cells, including $CD8^+$ T cells, T cells, B cells, monocytes, tumor-associated macrophages (TAMs), M1 macrophages, M2 macrophages, neutrophils and dendritic cells. Tumor purity is an important aspect affecting the number of infiltrating immune cells in clinical cancer biopsies. After adjusting for tumor purity, ALKBH7 expression was significantly negatively correlated with most of the immune markers of divergent types of immune cells in PAAD, PRAD and THCA (**Table 1**).

We also examined the correlations between ALKBH7 expression and various functional T cells, including Th1, Th1-like, Th2, Th17, Tfh, Treg, resting Tregs, effector Tregs, effector

TABLE 2 | Correlation analysis between ALKBH7 and gene markers of different types of T cells in TIMER.

Description	Gene markers	THCA				PRAD				PAAD			
		None		Purity		None		Purity		None		Purity	
		Cor	P	Cor	P	Cor	P	Cor	P	Cor	P	Cor	P
Th1	TBX21	−0.199	[a]	−0.192	[a]	−0.2	[a]	−0.17	[a]	−0.194	[b]	−0.143	0.062
	STAT4	−0.317	[a]	−0.317	[a]	−0.281	[a]	−0.216	[a]	−0.207	[b]	−0.179	[c]
	STAT1	−0.451	[a]	−0.436	[a]	−0.421	[a]	−0.339	[a]	−0.355	[a]	−0.302	[a]
	IFNG	−0.231	[a]	−0.221	[a]	−0.206	[a]	−0.144	[b]	−0.277	[a]	−0.238	[b]
	TNF	−0.236	[a]	−0.229	[a]	−0.261	[a]	−0.168	[a]	−0.125	0.096	−0.076	0.32
	IL12A	−0.064	0.152	−0.057	0.206	−0.21	[a]	−0.155	[b]	−0.127	0.09	−0.108	0.159
	IL12B	−0.254	[a]	−0.245	[a]	−0.211	[a]	−0.145	[b]	−0.134	0.075	−0.101	0.187
Th1-like	HAVCR2	−0.395	[a]	−0.382	[a]	−0.364	[a]	−0.3	[a]	−0.352	[a]	−0.28	[a]
	IFNG	−0.231	[a]	−0.221	[a]	−0.206	[a]	−0.144	[b]	−0.277	[a]	−0.238	[b]
	CXCR3	−0.15	[a]	−0.136	[a]	−0.226	[a]	−0.185	[a]	−0.017	0.825	0.046	0.549
	BHLHE40	−0.375	[a]	−0.367	[a]	−0.345	[a]	−0.311	[a]	−0.097	0.196	−0.091	0.238
	CD4	−0.394	[a]	−0.384	[a]	−0.4	[a]	−0.322	[a]	−0.337	[a]	−0.268	[a]
Th2	GATA3	−0.077	0.082	−0.058	0.199	−0.127	[b]	−0.035	0.479	−0.157	[c]	−0.119	0.122
	STAT6	−0.301	[a]	−0.285	[a]	−0.308	[a]	−0.269	[a]	−0.034	0.65	−0.019	0.806
	STAT5A	−0.273	[a]	−0.266	[a]	−0.24	[a]	−0.164	[a]	−0.072	0.339	−0.006	0.942
	IL13	−0.074	0.094	−0.071	0.116	−0.061	0.177	−0.088	0.073	−0.058	0.441	−0.049	0.521
Th17	STAT3	−0.356	[a]	−0.336	[a]	−0.449	[a]	−0.385	[a]	−0.338	[a]	−0.298	[a]
	IL17A	−0.135	[a]	−0.13	[b]	−0.139	[b]	−0.028	0.565	−0.235	[b]	−0.228	[b]
Tfh	BCL6	−0.231	[a]	−0.202	[a]	−0.316	[a]	−0.306	[a]	−0.26	[a]	−0.238	[b]
	IL21	−0.144	[b]	−0.133	[b]	−0.139	[b]	−0.116	[c]	−0.089	0.234	−0.048	0.532
Treg	FOXP3	−0.377	[a]	−0.363	[a]	−0.332	[a]	−0.337	[a]	−0.264	[a]	−0.199	[b]
	CCR8	−0.421	[a]	−0.402	[a]	−0.451	[a]	−0.396	[a]	−0.392	[a]	−0.342	[a]
	STAT5B	−0.209	[a]	−0.189	[a]	−0.465	[a]	−0.409	[a]	−0.126	0.093	−0.127	0.098
	TGFB1	−0.05	0.264	−0.042	0.359	−0.179	[a]	−0.193	[a]	0.07	0.353	0.133	0.083
Resting Treg	FOXP3	−0.377	[a]	−0.363	[a]	−0.332	[a]	−0.337	[a]	−0.264	[a]	−0.199	[b]
	IL2RA	−0.414	[a]	−0.404	[a]	−0.414	[a]	−0.356	[a]	−0.397	[a]	−0.334	[a]
Effector Treg T-cell	FOXP3	−0.377	[a]	−0.363	[a]	−0.332	[a]	−0.337	[a]	−0.264	[a]	−0.199	[b]
	CCR8	−0.421	[a]	−0.402	[a]	−0.451	[a]	−0.396	[a]	−0.392	[a]	−0.342	[a]
	TNFRSF9	−0.38	[a]	−0.361	[a]	−0.454	[a]	−0.377	[a]	−0.362	[a]	−0.309	[a]
Effector T-cell	CX3CR1	−0.17	[a]	−0.162	[a]	−0.383	[a]	−0.262	[a]	−0.262	[a]	−0.236	[b]
	FGFBP2	0.055	0.217	0.051	0.259	−0.118	[b]	−0.094	0.056	−0.22	[b]	−0.196	[c]
	FCGR3A	−0.336	[a]	−0.329	[a]	−0.38	[a]	−0.317	[a]	−0.388	[a]	−0.326	[a]
Naive T-cell	CCR7	−0.293	[a]	−0.276	[a]	−0.274	[a]	−0.217	[a]	−0.178	[c]	−0.124	0.106
	SELL	−0.355	[a]	−0.352	[a]	−0.4	[a]	−0.331	[a]	−0.249	[a]	−0.183	[c]
Effector memory T-cell	DUSP4	−0.29	[a]	−0.278	[a]	−0.162	[a]	−0.153	[a]	−0.111	0.137	0.09	0.242
	GZMK	−0.253	[a]	−0.242	[a]	−0.258	[a]	−0.19	[a]	−0.216	[b]	−0.149	0.052
	GZMA	−0.242	[a]	−0.235	[a]	−0.228	[a]	−0.151	[b]	−0.222	[b]	−0.173	[c]
Resident memory T-cell	CD69	−0.324	[a]	−0.311	[a]	−0.389	[a]	−0.306	[a]	−0.335	[a]	−0.297	[a]
	CXCR6	−0.282	[a]	−0.269	[a]	−0.292	[a]	−0.188	[a]	−0.339	[a]	−0.283	[a]
	MYADM	−0.144	[a]	−0.123	[b]	−0.335	[a]	−0.306	[a]	−0.186	[c]	−0.159	[c]
General memory T-cell	CCR7	−0.293	[a]	−0.276	[a]	−0.274	[a]	−0.217	[a]	−0.178	[c]	−0.124	0.106
	SELL	−0.355	[a]	−0.352	[a]	−0.4	[a]	−0.331	[a]	−0.249	[a]	−0.183	[c]
	IL7R	−0.415	[a]	−0.4	[a]	−0.458	[a]	−0.394	[a]	−0.425	[a]	−0.38	[a]
Exhausted T cell	PDCD1	−0.152	[a]	−0.154	[a]	−0.088	[c]	−0.072	0.142	−0.117	0.119	−0.044	0.568
	CTLA4	−0.317	[a]	−0.304	[a]	−0.151	[a]	−0.115	[c]	−0.227	[b]	−0.159	[c]
	LAG3	−0.211	[a]	−0.207	[a]	−0.031	0.486	0.001	0.98	−0.074	0.323	−0.047	0.543
	HAVCR2	−0.395	[a]	−0.382	[a]	−0.364	[a]	−0.3	[a]	−0.352	[a]	−0.28	[a]
	GZMB	−0.231	[a]	−0.23	[a]	−0.162	[a]	−0.114	[c]	−0.33	[a]	−0.273	[a]
	CXCL13	−0.264	[a]	−0.262	[a]	−0.192	[a]	−0.147	[b]	−0.218	[b]	−0.161	[c]
	LAYN	−0.073	0.098	−0.074	0.101	−0.237	[a]	−0.189	[a]	−0.217	[b]	−0.162	[c]

[c] p < *0.05.*

[b] p < *0.01.*

[a] p < *0.001.*

T cells, naïve T cells, effector memory T cells, resistant memory T cells, and exhausted T cells (**Table 2**). Using the TIMER database, the ALKBH7 expression level was also significantly negatively correlated with 44 of 50 T cell markers in PRAD and THCA and with 31 of 50 T cell markers in PAAD after adjusting for tumor purity (**Table 2**). These findings further support the hypothesis that ALKBH7 may be involved in inhibiting immune cell infiltration in PAAD, PRAD and THCA.

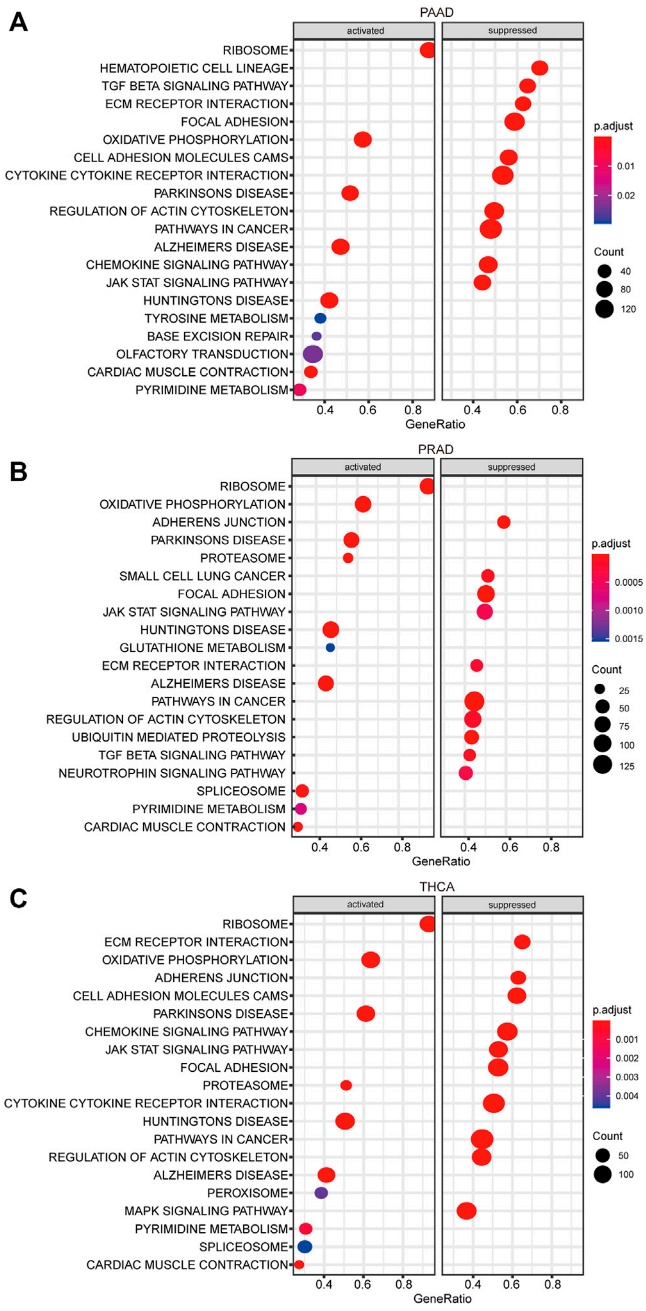


FIGURE 12 | KEGG enrichment analysis of ALKBH7. **(A)** Top 20 enriched KEGG pathways in PAAD. **(B)** Top 20 enriched KEGG pathways in PRAD. **(C)** Top 20 enriched KEGG pathways in THCA.

The ALKBH7 Coexpression Network Relevant Signalling Pathways

The aforementioned results identified significant associations between ALKBH7 expression and the prognosis and immunity of cancers. Considering the robust correlation between ALKBH7 expression and PAAD, PRAD and THCA, GSEA was performed to investigate the potential signalling pathways of ALKBH7 in these cancers. The results presented in **Figure 12** indicate that genes coexpressed with ALKBH7 are enriched in the regulation of immune and inflammatory responses and are negatively associated with these pathways, such as the JAK/STAT signalling pathway and TGF-β signalling pathway. These results suggested that ALKBH7 expression might play an essential role in human cancers by suppressing the immune response of the TME.

DISCUSSION

ALKBH7 is a mitochondrial protein involved in programmed necrosis, fatty acid metabolism and obesity development. In this paper, a comprehensive pan-cancer study of ALKBH7 revealed the potential prognostic and immunotherapeutic value of ALKBH7 in human cancers. First, significantly higher ALKBH7 expression was detected in most cancers compared to paired normal tissues, consistent with previous studies. Cai et al. observed high ALKBH7 expression in ovarian plasmacytoma (Cai et al., 2021), and Peng et al. found high ALKBH7 expression in hepatocellular carcinoma (Peng et al., 2021). However, ALKBH7 expression correlates with clinical parameters (age, sex and pathological stage) in only a few patients with cancer. For example, in BLCA, ALKBH7 expression correlated with the pathological stage of the tumor. Interestingly, ALKBH7 expression has some prognostic value for some cancers. For example, in a univariate survival analysis, ALKBH7 expression was significantly associated with four clinical survival datasets (OS, DSS, DFI and PFI) in patients with PAAD; in Kaplan–Meier survival estimates, downregulated ALKBH7 expression was significantly associated with shorter OS and DSS of patients with UCEC. These results suggest that ALKBH7 is a potential prognostic biomarker.

Next, ALKBH7 expression in different immune subtypes and molecular subtypes of human cancers was explored to determine its potential mechanism of action. ALKBH7 expression was significantly different in different immune subtypes and molecular subtypes in many cancer types, suggesting that ALKBH7 is a promising diagnostic pan-cancer biomarker and participates in immune regulation. Moreover, we documented significant differences in ALKBH7 expression in different immune and molecular subtypes of BRCA, PRAD and UCEC. In fact, differential ALKBH7 expression was detected in the cancers listed above and their normal tissue, indicating that ALKBH7 might play a role in the growth and progression of cancers.

Tumor cells use the immune checkpoint pathway to suppress immune cells and achieve immune escape (Topalian et al., 2015). Based on this principle, immune checkpoint inhibitors (ICIs) have emerged as new therapeutic approaches for cancer treatment (Muenst et al., 2016) and have been successfully applied in the clinic. The most commonly used ICI predictive biomarkers are programmed cell death ligand-1 (PD-L1), microsatellite instability (MSI) and tumor mutational burden (TMB) (Wang Y. et al., 2021). In addition, a study by Malta et al. found that a high tumor stemness index was associated with reduced PD-L1 expression in most cancers (Malta et al., 2018). In the present study, immunotherapy biomarkers (TMB and MSI) and the tumor stemness index showed significant associations with ALKBH7 in some cancers. Moreover, a strong relationship between the expression of ALKBH7 and ICP genes was identified. These results indicate that ALKBH7 has a strong association with ICIs.

Based on accumulating evidence, the tumor microenvironment (TME) is involved in tumor progression and significantly affects the treatment response and clinical outcome (Wu and Dai, 2017; Hinshaw and Shevde, 2019). Tumor-infiltrating lymphocytes (TILs) in the TME have been proven to be an independent predictor of the prognosis of patients with cancer and immunotherapeutic efficacy (Azimi et al., 2012). Our study found that ALKBH7 was related to the immune, stromal, and ESTIMATE scores and immune cell infiltration in the TME of most human cancer types, especially in PAAD, PRAD and THCA. Then, we explored the function of ALKBH7 in PAAD, PRAD and THCA by performing a KEGG analysis. ALKBH7 and its coexpression network were indeed involved in the regulation of the immune response and inflammatory response. In summary, these results strongly indicated the potential of ALKBH7 as a target of anticancer immunotherapy.

Overall, our pan-cancer analysis of ALKBH7 is the first to explore the relationship between ALKBH7 expression in human cancers and clinical prognostic factors, immune subtypes, molecular subtypes, immune checkpoints (ICPs), tumor mutational burden (TMB), microsatellite instability (MSI), tumor stemness index, tumor microenvironment (TME) and tumor-infiltrating lymphocytes (TILs). This information contributes to the understanding of the function of ALKBH7 in cancer development and its role in immunology. However, more experimental studies are required to explore the specific mechanisms of ALKBH7 action in cancer.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repositories and accession numbers can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

XG, CL, and YG conceived and designed the study. KC and DS were responsible for the collection and assembly of data, data analysis, and interpretation. KC, DL, and YH were involved in writing of the manuscript. LT provided help in revising the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (81772829 and 81830052), the Special Program for Collaborative Innovation, the Construction project of Shanghai Key Laboratory of Molecular Imaging (18DZ2260400) and Hunan Province National Science Foundation of China (2018JJ3895), "Top-100 Talent Cultivation Plan" of Shanghai University of Medicine and Health Sciences.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.822261/full#supplementary-material

REFERENCES

Aran, D., Hu, Z., and Butte, A. J. (2017). xCell: Digitally Portraying the Tissue Cellular Heterogeneity Landscape. *Genome Biol.* 18, 220. doi:10.1186/s13059-017-1349-1

Aran, D., Sirota, M., and Butte, A. J. (2015). Systematic Pan-Cancer Analysis of Tumour Purity. *Nat. Commun.* 6, 8971. doi:10.1038/ncomms9971

Azimi, F., Scolyer, R. A., Rumcheva, P., Moncrieff, M., Murali, R., McCarthy, S. W., et al. (2012). Tumor-infiltrating Lymphocyte Grade Is an Independent Predictor of sentinel Lymph Node Status and Survival in Patients with Cutaneous Melanoma. *Jco* 30, 2678–2683. doi:10.1200/JCO.2011.37.8539

Bai, R., Li, L., Li, L., Chen, X., Zhao, Y., Song, W., et al. (2021). Advances in Novel Molecular Typing and Precise Treatment Strategies for Small Cell Lung Cancer. *Chin. J. Cancer Res.* 33, 522–534. doi:10.21147/j.issn.1000-9604.2021.04.09

Bian, K., Lenz, S. A. P., Tang, Q., Chen, F., Qi, R., Jost, M., et al. (2019). DNA Repair Enzymes ALKBH2, ALKBH3, and AlkB Oxidize 5-methylcytosine to 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxylcytosine *In Vitro*. *Nucleic Acids Res.* 47, 5522–5529. doi:10.1093/nar/gkz395

Bonneville, R., Krook, M. A., Kautto, E. A., Miya, J., Wing, M. R., Chen, H.-Z., et al. (2017). Landscape of Microsatellite Instability across 39 Cancer Types. *JCO Precision Oncol.* PO.17, 1–15. doi:10.1200/PO.17.00073

Cai, Y., Wu, G., Peng, B., Li, J., Zeng, S., Yan, Y., et al. (2021). Expression and Molecular Profiles of the AlkB Family in Ovarian Serous Carcinoma. *Aging* 13, 9679–9692. doi:10.18632/aging.202716

Chen, G., Zhao, Q., Yuan, B., Wang, B., Zhang, Y., Li, Z., et al. (2021). ALKBH5-Modified HMGB1-STING Activation Contributes to Radiation Induced Liver Disease via Innate Immune Response. *Int. J. Radiat. Oncology*Biology*Physics* 111, 491–501. doi:10.1016/j.ijrobp.2021.05.115

Fu, D., Jordan, J. J., and Samson, L. D. (2013). Human ALKBH7 Is Required for Alkylation and Oxidation-Induced Programmed Necrosis. *Genes Dev.* 27, 1089–1100. doi:10.1101/gad.215533.113

Fujii, T., Shimada, K., Anai, S., Fujimoto, K., and Konishi, N. (2013). ALKBH2, a Novel AlkB Homologue, Contributes to Human Bladder Cancer Progression by Regulating MUC1 Expression. *Cancer Sci.* 104, 321–327. doi:10.1111/cas.12089

Goldman, M. J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., et al. (2020). Visualizing and Interpreting Cancer Genomics Data via the Xena Platform. *Nat. Biotechnol.* 38, 675–678. doi:10.1038/s41587-020-0546-8

Hinshaw, D. C., and Shevde, L. A. (2019). The Tumor Microenvironment Innately Modulates Cancer Progression. *Cancer Res.* 79, 4557–4566. doi:10.1158/0008-5472.CAN-18-3962

Jordan, J. J., Chhim, S., Margulies, C. M., Allocca, M., Bronson, R. T., Klungland, A., et al. (2017). ALKBH7 Drives a Tissue and Sex-specific Necrotic Cell Death Response Following Alkylation-Induced Damage. *Cell Death Dis* 8, e2947. doi:10.1038/cddis.2017.343

Kim, T. S., da Silva, E., Coit, D. G., and Tang, L. H. (2019). Intratumoral Immune Response to Gastric Cancer Varies by Molecular and Histologic Subtype. *Am. J. Surg. Pathol.* 43, 851–860. doi:10.1097/PAS.0000000000001253

Kulkarni, C. A., Nadtochiy, S. M., Kennedy, L., Zhang, J., Chhim, S., Alwaseem, H., et al. (2020). ALKBH7 Mediates Necrosis via Rewiring of Glyoxal Metabolism. *Elife* 9, e58573. doi:10.7554/eLife.58573

Lee, V., Murphy, A., Le, D. T., and Diaz, L. A., Jr (2016). Mismatch Repair Deficiency and Response to Immune Checkpoint Blockade. *Oncologist* 21, 1200–1211. doi:10.1634/theoncologist.2016-0046

Li, B., Chan, H. L., and Chen, P. (2019). Immune Checkpoint Inhibitors: Basics and Challenges. *Cmc* 26, 3009–3025. doi:10.2174/0929867324666170804143706

Li, B., Severson, E., Pignon, J.-C., Zhao, H., Li, T., Novak, J., et al. (2016). Comprehensive Analyses of Tumor Immunity: Implications for Cancer Immunotherapy. *Genome Biol.* 17, 174. doi:10.1186/s13059-016-1028-7

Li, N., Kang, Y., Wang, L., Huff, S., Tang, R., Hui, H., et al. (2020). ALKBH5 Regulates Anti-PD-1 Therapy Response by Modulating Lactate and Suppressive Immune Cell Accumulation in Tumor Microenvironment. *Proc. Natl. Acad. Sci. USA* 117, 20159–20170. doi:10.1073/pnas.1918986117

Li, T., Fan, J., Wang, B., Traugh, N., Chen, Q., Liu, J. S., et al. (2017). TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells. *Cancer Res.* 77, e108–e110. doi:10.1158/0008-5472.CAN-17-0307

Ma, S., Ba, Y., Ji, H., Wang, F., Du, J., and Hu, S. (2021). Recognition of Tumor-Associated Antigens and Immune Subtypes in Glioma for mRNA Vaccine Development. *Front. Immunol.* 12, 738435. doi:10.3389/fimmu.2021.738435

Malta, T. M., Sokolov, A., Gentles, A. J., Burzykowski, T., Poisson, L., Weinstein, J. N., et al. (2018). Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell* 173, 338–e15. e15. doi:10.1016/j.cell.2018.03.034

Meng, S., Zhan, S., Dou, W., and Ge, W. (2019). The Interactome and Proteomic Responses of ALKBH7 in Cell Lines by In-Depth Proteomics Analysis. *Proteome Sci.* 17, 8. doi:10.1186/s12953-019-0156-x

Muenst, S., Läubli, H., Soysal, S. D., Zippelius, A., Tzankov, A., and Hoeller, S. (2016). The Immune System and Cancer Evasion Strategies: Therapeutic Concepts. *J. Intern. Med.* 279, 541–562. doi:10.1111/joim.12470

Peng, B., Yan, Y., and Xu, Z. (2021). The Bioinformatics and Experimental Analysis of AlkB Family for Prognosis and Immune Cell Infiltration in Hepatocellular Carcinoma. *PeerJ* 9, e12123. doi:10.7717/peerj.12123

Pilžys, T., Marcinkowski, M., Kukwa, W., Garbicz, D., Dylewska, M., Ferenc, K., et al. (2019). ALKBH Overexpression in Head and Neck Cancer: Potential Target for Novel Anticancer Therapy. *Sci. Rep.* 9, 13249. doi:10.1038/s41598-019-49550-x

Rajecka, V., Skalicky, T., and Vanacova, S. (2019). The Role of RNA Adenosine Demethylases in the Control of Gene Expression. *Biochim. Biophys. Acta (Bba) - Gene Regul. Mech.* 1862, 343–355. doi:10.1016/j.bbagrm.2018.12.001

Ru, B., Wong, C. N., Tong, Y., Zhong, J. Y., Zhong, S. S. W., Wu, W. C., et al. (2019). TISIDB: an Integrated Repository portal for Tumor-Immune System Interactions. *Bioinformatics* 35, 4200–4202. doi:10.1093/bioinformatics/btz210

Solberg, A., Robertson, A. B., Aronsen, J. M., Rognmo, O., Sjaastad, I., Wisloff, U., et al. (2013). Deletion of Mouse Alkbh7 Leads to Obesity. *J. Mol. Cel Biol.* 5, 194–203. doi:10.1093/jmcb/mjt012

Thul, P. J., and Lindskog, C. (2018). The Human Protein Atlas: A Spatial Map of the Human Proteome. *Protein Sci.* 27, 233–244. doi:10.1002/pro.3307

Topalian, S. L., Drake, C. G., and Pardoll, D. M. (2015). Immune Checkpoint Blockade: a Common Denominator Approach to Cancer Therapy. *Cancer Cell* 27, 450–461. doi:10.1016/j.ccell.2015.03.001

Wang, G., He, Q., Feng, C., Liu, Y., Deng, Z., Qi, X., et al. (2014). The Atomic Resolution Structure of Human AlkB Homolog 7 (ALKBH7), a Key Protein for Programmed Necrosis and Fat Metabolism. *J. Biol. Chem.* 289, 27924–27936. doi:10.1074/jbc.M114.590505

Wang, S., Xiong, Y., Zhao, L., Gu, K., Li, Y., Zhao, F., et al. (2021). UCSCXenaShiny: An R/CRAN Package for Interactive Analysis of UCSC Xena Data. *Bioinformatics* 38, 527–529. doi:10.1093/bioinformatics/btab561

Wang, Y., Tong, Z., Zhang, W., Zhang, W., Buzdin, A., Mu, X., et al. (2021). FDA-approved and Emerging Next Generation Predictive Biomarkers for Immune Checkpoint Inhibitors in Cancer Patients. *Front. Oncol.* 11, 683419. doi:10.3389/fonc.2021.683419

Wu, G., Yan, Y., Cai, Y., Peng, B., Li, J., Huang, J., et al. (2021). ALKBH1-8 and FTO: Potential Therapeutic Targets and Prognostic Biomarkers in Lung Adenocarcinoma Pathogenesis. *Front. Cel Dev. Biol.* 9, 633927. doi:10.3389/fcell.2021.633927

Wu, T., and Dai, Y. (2017). Tumor Microenvironment and Therapeutic Response. *Cancer Lett.* 387, 61–68. doi:10.1016/j.canlet.2016.01.043

Wu, T. P., Wang, T., Seetin, M. G., Lai, Y., Zhu, S., Lin, K., et al. (2016). DNA Methylation on N6-Adenine in Mammalian Embryonic Stem Cells. *Nature* 532, 329–333. doi:10.1038/nature17640

Yarchoan, M., Hopkins, A., and Jaffee, E. M. (2017). Tumor Mutational Burden and Response Rate to PD-1 Inhibition. *N. Engl. J. Med.* 377, 2500–2501. doi:10.1056/NEJMc1713444

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., et al. (2013). Inferring Tumour Purity and Stromal and Immune Cell Admixture from Expression Data. *Nat. Commun.* 4, 2612. doi:10.1038/ncomms3612

Zhou, J., Zhang, X., Hu, J., Qu, R., Yu, Z., Xu, H., et al. (2021). m 6 A Demethylase ALKBH5 Controls CD4 + T Cell Pathogenicity and Promotes Autoimmunity. *Sci. Adv.* 7, eabg0470. doi:10.1126/sciadv.abg0470

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, orclaim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 11 February 2022
doi: 10.3389/fgene.2021.794820

Educational Attainment and Ischemic Stroke: A Mendelian Randomization Study

Luyan Gao[1], Kun Wang[2], Qing-Bin Ni[2], Hongguang Fan[1], Lan Zhao[1], Lei Huang[1], Mingfeng Yang[3]* and Huanming Li[4]*

[1]Department of Neurology, Tianjin Fourth Central Hospital, The Fourth Central Hospital Affilicated to Nankai University, The Fourth Central Clinical College of Tianjin Medical University, Tianjin, China, [2]Taishan Academy of Medical Sciences, Taian City Central Hospital, Taian, China, [3]Second Affiliated Hospital, Brain Science Institute, Key Laboratory of Cerebral Microcirculation in Universities of Shandong, Shandong First Medical University and Shandong Academy of Medical Sciences, Taian, China, [4]Department of Cardiovascular, Tianjin Fourth Central Hospital, The Fourth Central Hospital Affilicated to Nankai University, The Fourth Central Clinical College of Tianjin Medical University, Tianjin, China

OPEN ACCESS

Edited by:
Liangcai Zhang,
Janssen Research and Development,
United States

Reviewed by:
Yang Hu,
Harbin Institute of Technology, China
Keshen Li,
The First Affiliated Hospital of Jinan
University, China

***Correspondence:**
Mingfeng Yang
mfyang@163.com
Huanming Li
13820788663@163.com

Specialty section:
This article was submitted to
Statistical Genetics and Methodology,
a section of the journal
Frontiers in Genetics

Received: *14 October 2021*
Accepted: *24 December 2021*
Published: *11 February 2022*

Citation:
Gao L, Wang K, Ni Q-B, Fan H, Zhao L,
Huang L, Yang M and Li H (2022)
Educational Attainment and Ischemic
Stroke: A Mendelian
Randomization Study.
Front. Genet. 12:794820.
doi: 10.3389/fgene.2021.794820

Observational studies have evaluated the potential association of socioeconomic factors such as higher education with the risk of stroke but reported controversial findings. The objective of our study was to evaluate the potential causal association between higher education and the risk of stroke. Here, we performed a Mendelian randomization analysis to evaluate the potential association of educational attainment with ischemic stroke (IS) using large-scale GWAS datasets from the Social Science Genetic Association Consortium (SSGAC, 293,723 individuals), UK Biobank (111,349 individuals), and METASTROKE consortium (74,393 individuals). We selected three Mendelian randomization methods including inverse-variance-weighted meta-analysis (IVW), weighted median regression, and MR–Egger regression. IVW showed that each additional 3.6-year increase in years of schooling was significantly associated with a reduced IS risk (OR = 0.54, 95% CI: 0.41–0.71, and $p = 1.16 \times 10^{-5}$). Importantly, the estimates from weighted median (OR = 0.49, 95% CI: 0.33–0.73, and $p = 1.00 \times 10^{-3}$) and MR–Egger estimate (OR = 0.18, 95% CI: 0.06–0.60, and $p = 5.00 \times 10^{-3}$) were consistent with the IVW estimate in terms of direction and magnitude. In summary, we provide genetic evidence that high education could reduce IS risk.

Keywords: stroke, educational attainment, Mendelian randomization, genome-wide association studies, ischaemic stroke

Abbreviations: BMI, body mass index; CI, confidence interval; CARDIoGRAMplusC4D, Coronary ARtery DIsease Genome wide Replication and Meta-analysis (CARDIoGRAM) plus The Coronary Artery Disease (C4D) Genetics) consortium; DIAGRAM, DIAbetes Genetics Replication And Meta-analysis consortium; IS, ischemic stroke; GWAS, genome-wide association study; IVW, inverse-variance-weighted meta-analysis; OR, odds ratio; SD, standard deviation; SNP, single-nucleotide polymorphism; SSGAC, Social Science Genetic Association Consortium.

INTRODUCTION

Stroke is one of the leading causes of serious long-term disability in the world and is the fifth leading cause of death in the United States (Mozaffarian et al., 2016a; Mozaffarian et al., 2016b). Every year, there are more than 795,000 people having a stroke and more than 130,000 deaths from stroke, and the estimated stroke cost is $33 billion in the United States (Mozaffarian et al., 2016b). In recent years, there has been an increased interest for observational studies exploring the impact of socioeconomic factors such as higher education on stroke risk. In fact, a number of studies have reported that high education could reduce the risk of stroke (Nordahl et al., 2014; Ferrario et al., 2017; Kubota et al., 2017; Mchutchison et al., 2017). However, there are still some inconsistent findings. In 2002, Chang et al. found that stroke risk was reduced among less educated women in Africa, compared to highly educated women (Chang et al., 2002). It is well known that persons with cognitive impairment are at a higher risk of stroke (Sajjad et al., 2015). In 2015, Sajjad et al. conducted an observational study of 9,152 participants from the Rotterdam (Sajjad et al., 2015). They identified that education could modify the association between subjective memory complaints and risk of stroke (Sajjad et al., 2015). Higher education is significantly associated with a higher risk of stroke (hazard ratio = 1.39; 95% CI: 1.07–1.81) (Sajjad et al., 2015).

In recent years, large-scale genome-wide association studies (GWAS) promptly identified some common genetic variants and provided insight into the genetics of educational attainment (Okbay et al., 2016) and stroke (Malik et al., 2016). The existing large-scale GWAS datasets provide strong support for investigating the potential causal association of educational attainment with stroke risk by a Mendelian randomization analysis (Mokry et al., 2015; Nelson et al., 2015; Ference et al., 2017; Larsson et al., 2017a; Manousaki et al., 2017; Tillmann et al., 2017). This method could avoid some limitations of observational studies and is widely used to determine the causal inferences (Mokry et al., 2015; Ference et al., 2017; Larsson et al., 2017a; Manousaki et al., 2017; Tillmann et al., 2017; Wang et al., 2020).

It is reported that about 87% of all strokes are ischemic stroke (IS), in which blood flow to the brain is blocked (Mozaffarian et al., 2016a; Mozaffarian et al., 2016b). Intracerebral hemorrhage is the second most common cause of stroke (about 15%–30% of strokes) (An et al., 2017). Here, we performed a Mendelian randomization (MR) study to investigate the association of increased educational attainment with IS risk using the genetic variants from the large-scale educational attainment GWAS dataset (*N* = 405,072 individuals of European descent) and the large-scale IS GWAS dataset (*N* = 29,633, including 10,307 IS cases and 19,326 controls of European descent).

MATERIALS AND METHODS

Study Mesign

MR is based on three principal assumptions (Emdin et al., 2017; Larsson et al., 2017a). First, the genetic variants selected to be instrumental variables should be associated with the exposure (educational attainment) (Emdin et al., 2017; Larsson et al., 2017a). Second, the genetic variants should not be associated with confounders (assumption 2) (Emdin et al., 2017; Larsson et al., 2017a). Third, genetic variants should affect the risk of the outcome (IS) only through the exposure (educational attainment) (assumption 3) (Emdin et al., 2017; Larsson et al., 2017a). Recent studies have provided the more detailed information about the three principal assumptions (Liu et al., 2018; Liu et al., 2019; Zhang et al., 2020; Liu et al., 2021a; Liu et al., 2021b; Sun et al., 2021). This study is based on the publicly available, large-scale GWAS summary datasets. All participants gave informed consent in all these corresponding original studies. All relevant data are within the paper and the **Supplementary Tables S1**. The authors confirm that all data underlying the findings are either fully available without restriction through consortia websites or may be made available from consortia upon request.

Educational Attainment GWAS Dataset

We selected a large-scale GWAS dataset of educational attainment in individuals of European descent whose educational attainment was assessed at or above age 30 (Okbay et al., 2016). The examined phenotype is a continuous variable measuring the number of years of schooling completed (EduYears) (Okbay et al., 2016). This GWAS dataset consisted of 293,723 individuals in the discovery stage [Social Science Genetic Association Consortium (SSGAC), EduYears mean = 14.3, standard deviation (SD) = 3.6] and 111,349 individuals in the independent replication stage (UK Biobank, EduYears mean = 13.7, SD = 5.1) (a total of 405,072 individuals of European descent) (Okbay et al., 2016). In brief, the discovery stage GWAS from SSGAC was performed at the cohort level in individuals of European descent (Okbay et al., 2016). The replication stage GWAS from UK Biobank was conducted using conventionally population-based unrelated individuals with "White British" ancestry in the United Kingdom (Okbay et al., 2016). The meta-analysis of the discovery and replication stages of GWAS identified 162 independent genetic variants with the genome-wide significance ($p < 5.00 \times 10^{-8}$) (Okbay et al., 2016). Here, we selected these 162 independent genetic variants as the potential instrumental variables, as provided in **Table 1** and **Supplementary Table S1**, which could explain 1.6%–1.8% of the variance in education (Tillmann et al., 2017). Meanwhile, Li and others also selected these 162 independent genetic variants in their MR analysis to evaluate the causal association between educational attainment and asthma (Li et al., 2021).

IS GWAS Dataset

The IS GWAS dataset is from the METASTROKE consortium (Malik et al., 2016). The METASTROKE consortium performed a meta-analysis of 12 IS cohorts with a total of 10,307 IS individuals and 19,326 controls of European ancestry (*N* = 29,633 individuals) (Malik et al., 2016). More detailed information is described in the original study (Malik et al., 2016). The significance threshold for the association of these 162 educational attainment genetic variants with IS is $p < 0.05/162 = 3.09 \times 10^{-4}$.

TABLE 1 | 162 independent genetic variants as the potential instrumental variables.

SNP	Effect allele	Non-effect allele	Effect allele frequency	Effect size	Standard error	*p*-value
rs11130222	A	T	0.59	0.025	0.0023	3.68E-28
rs13090388	T	C	0.31	0.026	0.0024	2.58E-26
rs7029201	A	G	0.41	0.025	0.0023	7.16E-27
rs9401593	A	C	0.52	−0.024	0.0022	3.83E-28
rs12987662	A	C	0.39	0.025	0.0023	8.52E-28
rs8002014	A	G	0.27	−0.024	0.0025	3.80E-21
rs34305371	A	G	0.1	0.036	0.0039	1.52E-20
rs10773002	A	T	0.25	0.022	0.0026	7.74E-18
rs6882046	A	G	0.74	−0.019	0.0025	8.12E-14
rs17824247	T	C	0.59	−0.016	0.0023	2.41E-12
rs61160187	A	G	0.61	−0.017	0.0023	2.71E-14
rs11588857	A	G	0.21	0.02	0.0027	3.27E-13
rs2456973	A	C	0.67	−0.019	0.0024	5.83E-16
rs10786662	C	G	0.55	−0.017	0.0022	4.63E-14
rs4863692	T	G	0.32	0.017	0.0024	4.61E-12
rs10223052	A	G	0.36	0.016	0.0023	3.56E-12
rs11998763	A	G	0.54	0.017	0.0022	4.61E-14
rs9964724	T	C	0.68	0.018	0.0024	2.39E-14
rs6839705	A	C	0.36	0.015	0.0023	1.19E-10
rs7964899	A	G	0.44	0.016	0.0022	4.37E-13
rs12410444	A	G	0.7	−0.017	0.0024	6.01E-13
rs112634398	A	G	0.95	0.038	0.0055	2.74E-12
rs1106761	A	G	0.38	−0.016	0.0023	1.37E-11
rs3172494	T	G	0.12	0.023	0.0036	8.98E-11
rs58694847	C	G	0.26	−0.018	0.0025	4.98E-12
rs1008078	T	C	0.4	−0.017	0.0023	3.10E-14
rs34344888	A	G	0.39	−0.016	0.0023	8.87E-13
rs1378214	T	C	0.37	−0.015	0.0023	1.85E-11
rs16845580	T	C	0.63	0.016	0.0023	4.14E-12
rs12900061	A	G	0.18	0.019	0.0029	5.04E-11
rs35771425	T	C	0.79	0.018	0.0027	2.71E-11
rs7776010	T	C	0.82	−0.021	0.003	2.61E-12
rs1338554	A	G	0.5	0.015	0.0022	1.52E-11
rs356992	C	G	0.3	0.017	0.0024	4.03E-12
rs7593947	A	T	0.51	0.015	0.0022	2.39E-11
rs1912528	T	C	0.36	0.014	0.0023	1.53E-09
rs2992632	A	T	0.72	0.016	0.0025	3.25E-11
rs4741351	A	G	0.3	−0.015	0.0024	2.98E-10
rs6715849	A	G	0.44	−0.015	0.0022	1.65E-11
rs660001	A	G	0.21	−0.018	0.0027	1.34E-10
rs320700	A	G	0.65	0.014	0.0023	3.91E-09
rs113474297	T	C	0.13	−0.021	0.0034	8.34E-10
rs28420834	A	G	0.45	−0.014	0.0023	2.67E-10
rs56231335	T	C	0.67	−0.017	0.0024	7.17E-13
rs62263923	A	G	0.64	−0.017	0.0023	1.11E-13
rs12076635	C	G	0.79	0.018	0.0027	3.11E-11
rs9556958	T	C	0.53	−0.015	0.0022	1.21E-11
rs8049439	T	C	0.59	0.015	0.0023	6.99E-11
rs11774212	T	C	0.52	0.016	0.0023	1.51E-12
rs10483349	A	G	0.81	−0.017	0.0028	7.11E-10
rs71326918	A	C	0.1	0.022	0.0039	1.02E-08
rs11687170	T	C	0.83	0.021	0.0035	1.39E-09
rs7286601	T	G	0.54	−0.014	0.0023	1.99E-09
rs73344830	A	G	0.42	0.015	0.0023	9.93E-12
rs12143094	C	G	0.06	0.029	0.0049	2.73E-09
rs34638686	T	C	0.1	0.023	0.0038	1.51E-09
rs10761741	T	G	0.42	0.013	0.0023	7.05E-09
rs75090987	A	C	0.52	0.014	0.0022	1.14E-09
rs4500960	T	C	0.46	−0.014	0.0022	2.56E-10
rs1562242	T	C	0.48	−0.013	0.0022	5.95E-09
rs192818565	T	G	0.8	0.02	0.0029	2.02E-12
rs12534506	A	T	0.47	−0.014	0.0023	3.17E-10
rs10178115	T	G	0.54	0.014	0.0022	5.84E-10

(Continued on following page)

TABLE 1 | (*Continued*) 162 independent genetic variants as the potential instrumental variables.

SNP	Effect allele	Non-effect allele	Effect allele frequency	Effect size	Standard error	*p*-value
rs62100765	T	C	0.42	−0.015	0.0023	1.08E-10
rs12142680	A	G	0.09	0.026	0.0043	8.97E-10
rs71413877	A	G	0.04	0.035	0.0058	1.91E-09
rs149613931	T	G	0.06	−0.028	0.0048	5.54E-09
rs17167170	A	G	0.8	0.019	0.0028	1.79E-12
rs12956009	T	C	0.57	−0.013	0.0022	3.75E-09
rs2179152	T	C	0.37	−0.013	0.0023	9.30E-09
rs7033137	C	G	0.76	0.015	0.0026	1.77E-08
rs4378243	T	G	0.83	0.018	0.0029	1.04E-09
rs4493682	C	G	0.17	0.019	0.003	1.54E-10
rs9755467	T	C	0.16	0.019	0.0031	5.11E-10
rs4851251	T	C	0.27	−0.015	0.0025	1.36E-09
rs7945718	A	G	0.62	0.014	0.0023	1.26E-09
rs1382358	T	C	0.87	0.02	0.0035	1.66E-08
rs148490894	A	G	0.98	0.044	0.0078	1.84E-08
rs12761761	T	C	0.24	0.016	0.0027	1.04E-08
rs142328051	T	C	0.91	0.022	0.0039	3.60E-08
rs55786114	T	C	0.07	−0.03	0.0045	4.11E-11
rs7948975	T	C	0.64	0.014	0.0023	1.14E-09
rs1606974	A	G	0.12	0.022	0.0034	1.82E-10
rs10772644	C	G	0.88	0.02	0.0035	1.65E-08
rs111321694	T	C	0.17	−0.016	0.003	4.33E-08
rs17425572	A	G	0.46	0.014	0.0022	1.38E-09
rs111730030	T	G	0.06	−0.029	0.005	7.51E-09
rs1550973	A	G	0.35	−0.014	0.0023	2.00E-09
rs2406253	A	G	0.81	0.016	0.0028	4.64E-08
rs7772172	A	G	0.4	0.013	0.0023	9.83E-09
rs281302	A	G	0.56	−0.013	0.0022	2.88E-09
rs17372140	A	G	0.3	−0.014	0.0024	9.19E-09
rs12640626	A	G	0.58	0.013	0.0023	1.66E-08
rs113011189	T	C	0.09	−0.025	0.0045	2.91E-08
rs56081191	A	G	0.07	0.028	0.0047	3.67E-09
rs12694681	T	G	0.69	0.014	0.0024	1.81E-08
rs12134151	C	G	0.5	−0.013	0.0022	1.14E-08
rs7914680	T	G	0.71	−0.014	0.0025	1.60E-08
rs6493271	T	C	0.83	0.017	0.0029	4.21E-09
rs152603	A	G	0.65	−0.013	0.0023	2.01E-08
rs7791133	A	C	0.38	−0.014	0.0023	2.33E-09
rs1389473	A	G	0.38	−0.013	0.0023	4.52E-09
rs61874768	T	G	0.18	−0.016	0.0029	3.85E-08
rs10818606	T	C	0.4	−0.014	0.0023	5.67E-10
rs2568955	T	C	0.25	−0.016	0.0026	5.77E-10
rs268134	A	G	0.25	0.014	0.0026	3.53E-08
rs6939294	T	C	0.23	0.016	0.0026	2.90E-09
rs12653396	A	T	0.56	−0.013	0.0022	7.65E-09
rs648163	T	C	0.26	0.014	0.0025	1.38E-08
rs140711597	C	G	0.98	0.052	0.0091	1.66E-08
rs301800	T	C	0.18	0.016	0.0029	2.85E-08
rs12462428	T	C	0.81	0.016	0.0028	3.31E-08
rs11756123	A	T	0.35	−0.015	0.0023	6.43E-11
rs7429990	A	C	0.27	−0.015	0.0026	8.44E-09
rs12702087	A	G	0.46	0.013	0.0022	1.74E-09
rs4076457	T	C	0.25	0.015	0.0026	8.85E-09
rs78387210	T	C	0.09	0.023	0.004	8.41E-09
rs7610856	A	C	0.43	0.012	0.0023	3.02E-08
rs78365243	T	C	0.95	0.029	0.0052	2.22E-08
rs1115240	C	G	0.75	−0.016	0.0026	7.05E-10
rs7605827	A	T	0.29	0.016	0.0029	4.86E-08
rs76076331	T	C	0.14	0.02	0.0032	2.38E-10
rs1596747	A	G	0.51	0.014	0.0022	1.14E-09
rs77702819	T	G	0.09	0.022	0.004	2.93E-08
rs12646808	T	C	0.66	0.015	0.0024	3.79E-10
rs2624818	A	G	0.11	0.021	0.0037	8.63E-09

(Continued on following page)

TABLE 1 | (*Continued*) 162 independent genetic variants as the potential instrumental variables.

SNP	Effect allele	Non-effect allele	Effect allele frequency	Effect size	Standard error	*p*-value
rs7633857	C	G	0.52	–0.014	0.0026	4.74E-08
rs11976020	A	G	0.23	–0.015	0.0027	4.43E-08
rs4308415	C	G	0.44	–0.013	0.0022	2.52E-09
rs700590	T	C	0.59	–0.013	0.0023	2.84E-08
rs756912	T	C	0.52	–0.014	0.0022	1.14E-09
rs7241530	T	C	0.36	–0.013	0.0023	2.28E-08
rs35971989	A	G	0.84	0.018	0.0032	2.95E-08
rs11771168	T	C	0.24	–0.015	0.0027	2.56E-08
rs17504614	T	C	0.8	0.016	0.0028	1.56E-08
rs9914544	A	C	0.62	–0.013	0.0023	4.66E-08
rs4675248	A	G	0.4	–0.012	0.0023	4.39E-08
rs6800916	A	T	0.08	–0.024	0.0043	1.70E-08
rs35532491	A	T	0.9	-0.022	0.0038	7.15E-09
rs56044892	T	C	0.2	–0.016	0.0028	5.37E-09
rs79925071	T	G	0.56	0.013	0.0022	1.52E-08
rs12145291	T	C	0.94	–0.029	0.0051	2.21E-08
rs34106693	C	G	0.83	0.017	0.0031	1.80E-08
rs12754946	T	C	0.57	0.013	0.0023	1.48E-08
rs4741343	A	G	0.18	–0.016	0.0029	2.32E-08
rs76878669	C	G	0.76	0.014	0.0026	4.12E-08
rs775326	A	C	0.32	–0.014	0.0024	1.22E-08
rs10821136	T	C	0.34	0.013	0.0024	3.58E-08
rs1925576	A	G	0.54	–0.012	0.0022	2.23E-08
rs6065080	T	C	0.36	–0.013	0.0023	1.16E-08
rs56158183	A	G	0.07	0.025	0.0043	1.42E-08
rs12531458	A	C	0.52	0.012	0.0022	3.81E-08
rs62379838	T	C	0.69	0.013	0.0024	4.06E-08
rs7590368	T	C	0.73	–0.014	0.0025	2.72E-08
rs113520408	A	G	0.27	0.015	0.0025	7.15E-09
rs62263033	T	C	0.96	0.037	0.0063	5.60E-09
rs11643654	A	C	0.6	0.013	0.0023	2.00E-08
rs10930008	A	G	0.73	–0.014	0.0025	4.14E-08
rs56262138	A	T	0.3	0.014	0.0025	2.29E-08
rs113779084	A	G	0.31	0.014	0.0024	2.70E-08
rs62262721	T	C	0.96	0.042	0.0072	3.41E-09
rs1967109	A	G	0.15	–0.017	0.0031	4.40E-08

Pleiotropy Analysis

We performed a comprehensive pleiotropy analysis to assure that the selected genetic variants do not exert effects on IS through biological pathways independent of education levels. The American Heart Association and American Stroke Association have reported the leading risk factors for stroke, including high blood pressure, high cholesterol, heart disease (coronary artery disease), diabetes, current smoking, obesity, and excessive alcohol drinking (Meschia et al., 2014). In stage 1, we manually evaluated the potential pleiotropy using the GWAS datasets about the known confounders including high blood pressure, high cholesterol, body mass index (BMI), smoking behavior, and alcohol drinking from the UK Biobank (Sudlow et al., 2015); coronary artery disease from the CARDIoGRAMplusC4D [Coronary ARtery DIsease Genome wide Replication and Meta-analysis (CARDIoGRAM) plus The Coronary artery disease (C4D) Genetics] consortium (Nikpay et al., 2015); and type 2 diabetes from the DIAGRAM (DIAbetes Genetics Replication And Meta-analysis) consortium (Zhao et al., 2017). The significance threshold for the association of these 162 genetic variants with the potential confounders is a Bonferroni correction $p < 0.05/162 = 3.09E\text{-}04$.

In stage 2, we selected three statistical methods to perform the pleiotropy analysis. The first statistical method is based on the heterogeneity test (Greco et al., 2015; Hartwig et al., 2017; Liu et al., 2017). The potential heterogeneity in these genetic variants could be evaluated using Cochran's Q test (together with the I^2 index), which is a useful tool to explore the presence of heterogeneity due to pleiotropy or other causes, especially in MR studies with large sample sizes based on summary data (Greco et al., 2015). The second statistical method is the MR–Egger intercept test that provides an assessment of the validity of the instrumental variable assumptions and provides a statistical test of the presence of potential pleiotropy (Dale et al., 2017). The third statistical method is a newly developed method named Mendelian Randomization Pleiotropy RESidual Sum and Outlier (MR-PRESSO) test (Verbanck et al., 2018). In all these three statistical methods, the threshold of statistical significance for evidence of pleiotropy is $p < 0.05$.

Mendelian Randomization Analysis

We selected three MR methods including inverse-variance-weighted meta-analysis (IVW), weighted median regression, and MR–Egger regression, as in recent studies (Dale et al., 2017; Larsson et al., 2017a; Tillmann et al., 2017; Liu et al., 2018; Liu et al., 2019). If there is no clear evidence of pleiotropy, these three methods should give consistent estimates. The odds ratio (OR) as well as 95% confidence interval (CI) of IS corresponds to a per 3.6 increase [about 1 standard deviation (SD)] in educational attainment levels. All analyses were conducted using R (version 3.2.4) and R package "MendelianRandomization" (Yavorska and Burgess, 2017). The statistical significance was $p < 0.05$.

Power Analysis

The proportion of education variance explained by the instrumental variables can be estimated using R^2.

$$R^2 = \sum_{i=1}^{k} \frac{\beta_i^2 * 2 * MAF_{SNP_i}\left(1 - MAF_{SNP_i}\right)}{\mathrm{var}\left(X\right)}$$

where β_i is the effect size (beta coefficient) associated with the education for SNP_i, MAF_{SNPi} is the minor allele frequency for SNP_i, K is the number of genetic variants, and $\mathrm{var}(X)$ is the variance of the education [$\mathrm{var}(X) = 1$ for education, since the beta estimates refer to change in 1 standard deviation (SD)] (Pattaro et al., 2016; Mack et al., 2017). The strength of the instrumental variables was evaluated by the first-stage F-statistic (Noyce et al., 2017; Xu et al., 2017). A common threshold is F > 10 which avoids bias in MR studies (Burgess and Thompson, 2011). Here, we calculated statistical power to estimate the minimum detectable magnitudes of association for IS using the web-based tool mRnd and a two-sided type-I error rate α of 0.05 (Brion et al., 2013).

RESULTS

Association of Educational Attainment Variants With IS

Of the 162 genetic variants associated with educational attainment, we extracted the summary statistics for all these 162 variants in the IS GWAS dataset. The characteristics of 162 genetic variants used as instrumental variables in IS are described in **Supplementary Table S2**. We noticed that none of these 162 genetic variants was significantly associated with IS risk at the Bonferroni-corrected significance threshold ($p < 0.05/162 = 3.09 \times 10^{-3}$) (**Supplementary Table S2**).

Pleiotropy Analysis

In stage 1, 51 of these 162 educational attainment genetic variants are significantly associated with known confounders at the Bonferroni-corrected significance threshold ($p < 0.05/162 = 3.09 \times 10^{-3}$), as described in **Supplementary Tables S3–S9**. In brief, seven genetic variants were significantly associated with smoking. Two genetic variants were significantly associated with coronary artery disease. Six genetic variants were significantly associated with high blood pressure. 43 genetic variants were significantly associated with BMI. To meet the MR assumptions, we excluded these 51 genetic variants in the following analysis. In stage 2, using the remaining 111 genetic variants, the heterogeneity test showed no significant heterogeneity [$I^2 = 0\%$, 95% CI (0%; 16.8%), and $p = 0.7093$]. The MR–Egger intercept test showed no significant pleiotropy (MR–Egger intercept $\beta = 0.018$; $p = 0.064$). The MR-PRESSO test did not identify any horizontal pleiotropic outliers.

TABLE 2 | MR analysis results between educational attainment and IS.

Method	OR	95% CI	*p* value
Inverse-variance weighted	0.54	0.41–0.71	1.16×10^{-5}
Weighted median	0.49	0.33–0.73	1.00×10^{-3}
MR–Egger	0.18	0.06–0.60	5.00×10^{-3}

OR, odds ratio; CI, confidence interval; the significance was at p *< 0.05.*

Association of Educational Attainment Levels With IS

Using the remaining 111 genetic variants, IVW showed that each SD increase in years of schooling (3.6 years) was significantly associated with a reduced IS risk (OR = 0.54, 95% CI: 0.41–0.71, and $p = 1.16 \times 10^{-5}$). Interestingly, the estimates from weighted median (OR = 0.49, 95% CI: 0.33–0.73, and $p = 1.00 \times 10^{-3}$), and MR–Egger estimate (OR = 0.18, 95% CI: 0.06–0.60, and $p = 5.00 \times 10^{-3}$), were consistent with the IVW estimate in terms of direction and magnitude, as provided in **Table 2**. **Figure 1** shows individual causal estimates from each of the 111 genetic variants using different methods.

Power Analysis

Here, all these 111 genetic variants could explain about 1.09% of the educational attainment variance ($R^2 = 1.09\%$). The first-stage F-statistic for the instrument including these 111 genetic variants was 327.56 > 10, so a weak instrument bias is unlikely. The actual N for IS GWAS is 29,633, and the proportion of cases is 0.347822. Our MR study had 80% power to detect effect sizes of moderate magnitude with ORs as low as 0.71 and as high as 1.37 per SD increase in educational attainment levels for IS. Importantly, the power to detect the causal association (OR = 0.54, 95% CI: 0.41–0.71, and $p = 1.16 \times 10^{-5}$) is 100% by selecting these 111 genetic variants as the instrumental variables. Hence, our analysis has enough statistical power to detect robust causal effect estimates.

DISCUSSION

It has been a long time since the relation between the educational attainment and risk of stroke was evaluated. Until November 2015, there have been 79 observational studies including approximately 164,683 strokes (Mchutchison et al., 2017). However, these observational studies have reported both positive and negative associations between higher educational attainment and stroke (Mchutchison et al., 2017). Meanwhile,

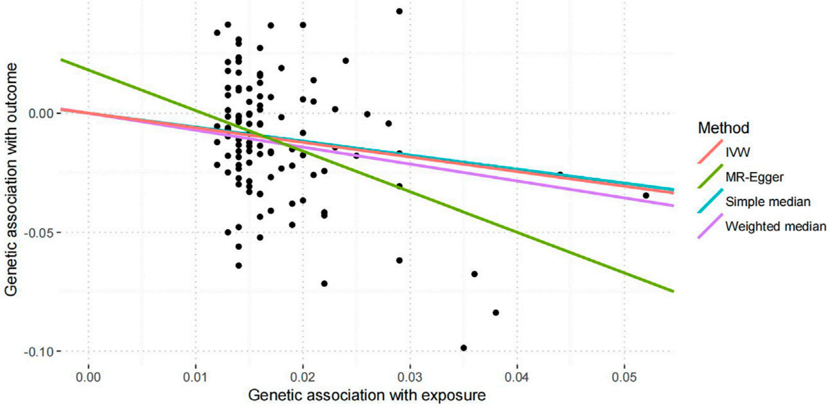


FIGURE 1 | Individual causal estimates from each of the 111 genetic variants. This scatter plot shows individual causal estimates from each of 111 genetic variants associated with educational attainment on the *x*-axis and IS risk on the *y*-axis. The continuous line represents the causal estimate of educational attainment on IS risk.

there was clear between-study heterogeneity in all comparisons, ranging from 76% to 96% (Mchutchison et al., 2017). Until now, it has been difficult to establish causality because of methodological limitations of traditional observational studies.

Here, we performed an MR analysis to evaluate the potential association of educational attainment with IS risk using large-scale GWAS datasets. MR is based on the premise that the human genetic variants are randomly distributed in the population (Emdin et al., 2017). These genetic variants are largely not associated with confounders and can be used as instrumental variables to estimate the causal association of an exposure with an outcome (Emdin et al., 2017), which could avoid the methodological limitations of the traditional observational studies.

Our results indicated that a genetically increased educational attainment was significantly associated with reduced IS risk. IVW showed that each additional 3.6-year increase in years of schooling was significantly associated with a reduced IS risk (OR = 0.54, 95% CI: 0.41–0.71, and $p = 1.16 \times 10^{-5}$). Importantly, other sensitivity analyses further supported this estimate. All these findings show that the causal association between genetically increased educational attainment and reduced IS risk is robust. Hence, our results do seem to hint at what lifestyle choices may help protect against IS. The life experiences that engage the brain, such as higher educational attainment, may protect against IS risk.

Our findings are comparable to findings from traditional observational studies with OR = 0.74 (Mchutchison et al., 2017), 0.65 (men) (Ferrario et al., 2017), and 0.71 (women) (Ferrario et al., 2017). Meanwhile, our findings are also consistent with the results from a recent MR study, which found that one SD increase in years of schooling (3.6 years) was associated with a reduced risk of coronary heart disease (OR = 0.67, 95% CI 0.59–0.77; $p = 3.00 \times 10^{-8}$) (Tillmann et al., 2017). It has been established that coronary artery disease is one of the leading risk factors for stroke (Meschia et al., 2014).

Until now, 3 MR studies have also investigated the causal association between educational attainment and IS. Harshfield et al. assessed the causal effect of 12 lifestyle factors on risk of stroke (Harshfield et al., 2021). They found that genetically increased educational attainment was associated with reduced risk of IS, large artery stroke, and small vessel stroke, and intracerebral hemorrhage using 305 educational attainment genetic variants (Harshfield et al., 2021). Wen et al. selected 58 educational attainment genetic variants and identified a suggestive causal association between education and IS (p = 0.048) (Xiuyun et al., 2020). Gill et al. selected 625 instrument SNPs for educational attainment and found that education was causally associated with stroke risk (Gill et al., 2019). A main difference between our and previous MR studies is the manual pleiotropy analysis. These above 3 MR studies only used the statistical methods to perform the pleiotropy analysis (Gill et al., 2019; Xiuyun et al., 2020; Harshfield et al., 2021).

This MR study has several strengths. First, this study may benefit from the large-scale educational attainment GWAS dataset (N = 405,072 individuals of European descent individuals) and IS GWAS dataset (N = 29,633 individuals of European descent). Importantly, power analysis further provides ample power to detect the association of educational attainment with IS risk. Second, both the educational attainment and IS GWAS datasets are from the European descent, which may reduce the influence on the potential association caused by the population stratification. Third, multiple independent genetic variants are taken as instruments, which may reduce the influence on the potential association caused by the linkage disequilibrium; Fourth, we selected multiple methods to perform MR analysis, as in previous studies (Mokry et al., 2015; Nelson et al., 2015; Emdin et al., 2017; Larsson et al., 2017a; Manousaki et al., 2017; Noyce et al., 2017). Fifth, we performed a comprehensive pleiotropy analysis to evaluate the potential association of these educational attainment genetic variants with known IS risk factors. We excluded 51 genetic variants associated with potential confounders, which meets the MR assumptions.

Despite these interesting results, we recognize some limitations in our study. First, we could not completely rule out that there may be additional confounders, although some other available software or tools may be helpful to identify the pleiotropy, such as GSMR (Zhu et al., 2018) and CAUSE (Morrison et al., 2020). Until now, it is almost impossible to fully rule out pleiotropy present in any MR study (Emdin et al., 2017; Larsson et al., 2017a; Larsson et al., 2017b). Second, it could not be completely ruled out that population stratification may have had some influence on the estimate. Third, the genetic association between education and IS may be different in different ancestries. Hence, this causal association should be further evaluated in other ancestries. In some individuals, the association between a genetic variant and one specific outcome may have been confounded by the hidden population structure (Davies et al., 2018). Thus, MR studies using these individuals could have been biased by population stratification or different ancestries (Davies et al., 2019). In fact, Zheng et al. found that hypertension could play different causal roles on chronic kidney disease across ancestries (Zheng et al., 2022). Fourth, the underlying mechanisms about the causal association between educational attainment and IS remain unclear.

In summary, we provide genetic evidence that high education could reduce IS risk. Our findings could have public health implication to raise awareness of the extent to which educational inequalities are associated with risk of IS. Meanwhile, population-based solutions may contribute to ameliorate the deleterious effects of low educational attainment on health outcomes.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

LG, MY, and HL designed the study; LG, analyzed the data; and all authors contributed to the interpretation of the results and critical revision of the manuscript for important intellectual content and approved the final version of the manuscript. HF, LZ, and LH contribute to revision of the manuscript.

FUNDING

This work was supported by the funding from the National Natural Science Foundation of China (Grant No. 81871855) and the Natural Science Foundation of Shandong Province (No. ZR2020MH141), and Exceptional Young Talents Fostering Foundation 2021 of the Tianjin Fourth Central Hospital (Grant No. tjdszxyy20210003).

ACKNOWLEDGMENTS

This research has been conducted using the IS GWAS datasets from the METASTROKE consortium. We thank the METASTROKE consortium, DIAbetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium, CardiogramplusC4D consortium, Social Science Genetic Association Consortium (SSGAC), and UK Biobank for other GWAS summary statistics.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.794820/full#supplementary-material

REFERENCES

An, S. J., Kim, T. J., and Yoon, B. W. (2017). Epidemiology, Risk Factors, and Clinical Features of Intracerebral Hemorrhage: An Update. *J. Stroke* 19, 3–10. doi:10.5853/jos.2016.00864

Brion, M. J., Shakhbazov, K., and Visscher, P. M. (2013). Calculating Statistical Power in Mendelian Randomization Studies. *Int. J. Epidemiol.* 42, 1497–1501. doi:10.1093/ije/dyt179

Burgess, S., and Thompson, S. G. (2011). Avoiding Bias from Weak Instruments in Mendelian Randomization Studies. *Int. J. Epidemiol.* 40, 755–764. doi:10.1093/ije/dyr036

Chang, C. L., Marmot, M. G., Farley, T. M., and Poulter, N. R. (2002). The Influence of Economic Development on the Association between Education and the Risk of Acute Myocardial Infarction and Stroke. *J. Clin. Epidemiol.* 55, 741–747. doi:10.1016/s0895-4356(02)00413-4

Dale, C. E., Fatemifar, G., Palmer, T. M., White, J., Prieto-Merino, D., Zabaneh, D., et al. (2017). Causal Associations of Adiposity and Body Fat Distribution with Coronary Heart Disease, Stroke Subtypes, and Type 2 Diabetes Mellitus: A Mendelian Randomization Analysis. *Circulation* 135, 2373–2388. doi:10.1161/CIRCULATIONAHA.116.026560

Davies, N. M., Hill, W. D., Anderson, E. L., Sanderson, E., Deary, I. J., and Davey Smith, G. (2019). Multivariable Two-Sample Mendelian Randomization Estimates of the Effects of Intelligence and Education on Health. *Elife* 8, e43990. doi:10.7554/eLife.43990

Davies, N. M., Holmes, M. V., and Davey Smith, G. (2018). Reading Mendelian Randomisation Studies: a Guide, Glossary, and Checklist for Clinicians. *BMJ* 362, k601. doi:10.1136/bmj.k601

Emdin, C. A., Khera, A. V., Natarajan, P., Klarin, D., Zekavat, S. M., Hsiao, A. J., et al. (2017). Genetic Association of Waist-To-Hip Ratio with Cardiometabolic Traits, Type 2 Diabetes, and Coronary Heart Disease. *JAMA* 317, 626–634. doi:10.1001/jama.2016.21042

Ference, B. A., Kastelein, J. J. P., Ginsberg, H. N., Chapman, M. J., Nicholls, S. J., Ray, K. K., et al. (2017). Association of Genetic Variants Related to CETP Inhibitors and Statins with Lipoprotein Levels and Cardiovascular Risk. *JAMA* 318, 947–956. doi:10.1001/jama.2017.11467

Ferrario, M. M., Veronesi, G., Kee, F., Chambless, L. E., Kuulasmaa, K., Jorgensen, T., et al. (2017). Determinants of Social Inequalities in Stroke Incidence across Europe: a Collaborative Analysis of 126 635 Individuals from 48 Cohort Studies. *J. Epidemiol. Community Health* 71, 1210–1216. doi:10.1136/jech-2017-209728

Gill, D., Efstathiadou, A., Cawood, K., Tzoulaki, I., and Dehghan, A. (2019). Education Protects against Coronary Heart Disease and Stroke Independently of Cognitive Function: Evidence from Mendelian Randomization. *Int. J. Epidemiol.* 48, 1468–1477. doi:10.1093/ije/dyz200

Greco, M. F., Minelli, C., Sheehan, N. A., and Thompson, J. R. (2015). Detecting Pleiotropy in Mendelian Randomisation Studies with Summary Data and a Continuous Outcome. *Stat. Med.* 34, 2926–2940. doi:10.1002/sim.6522

Harshfield, E. L., Georgakis, M. K., Malik, R., Dichgans, M., and Markus, H. S. (2021). Modifiable Lifestyle Factors and Risk of Stroke: A Mendelian

Randomization Analysis. *Stroke* 52, 931–936. doi:10.1161/STROKEAHA.120.031710

Hartwig, F. P., Borges, M. C., Horta, B. L., Bowden, J., and Davey Smith, G. (2017). Inflammatory Biomarkers and Risk of Schizophrenia: A 2-Sample Mendelian Randomization Study. *JAMA Psychiatry* 74 (12), 1226–1233. doi:10.1001/jamapsychiatry.2017.3191

Kubota, Y., Heiss, G., Maclehose, R. F., Roetker, N. S., and Folsom, A. R. (2017). Association of Educational Attainment with Lifetime Risk of Cardiovascular Disease: The Atherosclerosis Risk in Communities Study. *JAMA Intern. Med.* 177, 1165–1172. doi:10.1001/jamainternmed.2017.1877

Larsson, S. C., Burgess, S., and Michaelsson, K. (2017a). Association of Genetic Variants Related to Serum Calcium Levels with Coronary Artery Disease and Myocardial Infarction. *JAMA* 318, 371–380. doi:10.1001/jama.2017.8981

Larsson, S. C., Traylor, M., Malik, R., Dichgans, M., Burgess, S., and Markus, H. S. (2017b). Modifiable Pathways in Alzheimer's Disease: Mendelian Randomisation Analysis. *BMJ* 359, j5375. doi:10.1136/bmj.j5375

Li, Y., Chen, W., Tian, S., Xia, S., and Yang, B. (2021). Evaluating the Causal Association between Educational Attainment and Asthma Using a Mendelian Randomization Design. *Front. Genet.* 12, 716364. doi:10.3389/fgene.2021.716364

Liu, G., Jin, S., and Jiang, Q. (2019). Interleukin-6 Receptor and Inflammatory Bowel Disease: A Mendelian Randomization Study. *Gastroenterology* 156, 823–824. doi:10.1053/j.gastro.2018.09.059

Liu, G., Xu, Y., Jiang, Y., Zhang, L., Feng, R., and Jiang, Q. (2017). PICALM Rs3851179 Variant Confers Susceptibility to Alzheimer's Disease in Chinese Population. *Mol. Neurobiol.* 54, 3131–3136. doi:10.1007/s12035-016-9886-2

Liu, G., Zhao, Y., Jin, S., Hu, Y., Wang, T., Tian, R., et al. (2018). Circulating Vitamin E Levels and Alzheimer's Disease: a Mendelian Randomization Study. *Neurobiol. Aging* 72, 189.e1–189.e9. doi:10.1016/j.neurobiolaging.2018.08.008

Liu, H., Zhang, Y., Hu, Y., Zhang, H., Wang, T., Han, Z., et al. (2021a). Mendelian Randomization to Evaluate the Effect of Plasma Vitamin C Levels on the Risk of Alzheimer's Disease. *Genes Nutr.* 16, 19. doi:10.1186/s12263-021-00700-9

Liu, H., Zhang, Y., Zhang, H., Wang, L., Wang, T., Han, Z., et al. (2021b). Effect of Plasma Vitamin C Levels on Parkinson's Disease and Age at Onset: a Mendelian Randomization Study. *J. Transl Med.* 19, 221. doi:10.1186/s12967-021-02892-5

Mack, S., Coassin, S., Vaucher, J., Kronenberg, F., and Lamina, C. (2017). Evaluating the Causal Relation of ApoA-IV with Disease-Related Traits - A Bidirectional Two-Sample Mendelian Randomization Study. *Sci. Rep.* 7, 8734. doi:10.1038/s41598-017-07213-9

Malik, R., Traylor, M., Pulit, S. L., Bevan, S., Hopewell, J. C., Holliday, E. G., et al. (2016). Low-frequency and Common Genetic Variation in Ischemic Stroke: The METASTROKE Collaboration. *Neurology* 86, 1217–1226. doi:10.1212/WNL.0000000000002528

Manousaki, D., Paternoster, L., Standl, M., Moffatt, M. F., Farrall, M., Bouzigon, E., et al. (2017). Vitamin D Levels and Susceptibility to Asthma, Elevated Immunoglobulin E Levels, and Atopic Dermatitis: A Mendelian Randomization Study. *Plos Med.* 14, e1002294. doi:10.1371/journal.pmed.1002294

Mchutchison, C. A., Backhouse, E. V., Cvoro, V., Shenkin, S. D., and Wardlaw, J. M. (2017). Education, Socioeconomic Status, and Intelligence in Childhood and Stroke Risk in Later Life: A Meta-Analysis. *Epidemiology* 28, 608–618. doi:10.1097/EDE.0000000000000675

Meschia, J. F., Bushnell, C., Boden-Albala, B., Braun, L. T., Bravata, D. M., Chaturvedi, S., et al. (2014). Guidelines for the Primary Prevention of Stroke: a Statement for Healthcare Professionals from the American Heart Association/American Stroke Association. *Stroke* 45, 3754–3832. doi:10.1161/STR.0000000000000046

Mokry, L. E., Ross, S., Ahmad, O. S., Forgetta, V., Smith, G. D., Goltzman, D., et al. (2015). Vitamin D and Risk of Multiple Sclerosis: A Mendelian Randomization Study. *Plos Med.* 12, e1001866. doi:10.1371/journal.pmed.1001866

Morrison, J., Knoblauch, N., Marcus, J. H., Stephens, M., and He, X. (2020). Mendelian Randomization Accounting for Correlated and Uncorrelated Pleiotropic Effects Using Genome-wide Summary Statistics. *Nat. Genet.* 52, 740–747. doi:10.1038/s41588-020-0631-4

Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., et al. (2016a). Executive Summary: Heart Disease and Stroke Statistics--2016 Update: A Report from the American Heart Association. *Circulation* 133, 447–454. doi:10.1161/CIR.0000000000000366

Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., et al. (2016b). Heart Disease and Stroke Statistics-2016 Update: A Report from the American Heart Association. *Circulation* 133, e38–360. doi:10.1161/CIR.0000000000000350

Nelson, C. P., Hamby, S. E., Saleheen, D., Hopewell, J. C., Zeng, L., Assimes, T. L., et al. (2015). Genetically Determined Height and Coronary Artery Disease. *N. Engl. J. Med.* 372, 1608–1618. doi:10.1056/NEJMoa1404881

Nikpay, M., Goel, A., Won, H. H., Hall, L. M., Willenborg, C., Kanoni, S., et al. (2015). A Comprehensive 1,000 Genomes-Based Genome-wide Association Meta-Analysis of Coronary Artery Disease. *Nat. Genet.* 47, 1121–1130. doi:10.1038/ng.3396

Nordahl, H., Osler, M., Frederiksen, B. L., Andersen, I., Prescott, E., Overvad, K., et al. (2014). Combined Effects of Socioeconomic Position, Smoking, and Hypertension on Risk of Ischemic and Hemorrhagic Stroke. *Stroke* 45, 2582–2587. doi:10.1161/STROKEAHA.114.005252

Noyce, A. J., Kia, D. A., Hemani, G., Nicolas, A., Price, T. R., De Pablo-Fernandez, E., et al. (2017). Estimating the Causal Influence of Body Mass index on Risk of Parkinson Disease: A Mendelian Randomisation Study. *Plos Med.* 14, e1002314. doi:10.1371/journal.pmed.1002314

Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., et al. (2016). Genome-wide Association Study Identifies 74 Loci Associated with Educational Attainment. *Nature* 533, 539–542. doi:10.1038/nature17671

Pattaro, C., Teumer, A., Gorski, M., Chu, A. Y., Li, M., Mijatovic, V., et al. (2016). Genetic Associations at 53 Loci Highlight Cell Types and Biological Pathways Relevant for Kidney Function. *Nat. Commun.* 7, 10023. doi:10.1038/ncomms10023

Sajjad, A., Mirza, S. S., Portegies, M. L., Bos, M. J., Hofman, A., Koudstaal, P. J., et al. (2015). Subjective Memory Complaints and the Risk of Stroke. *Stroke* 46, 170–175. doi:10.1161/STROKEAHA.114.006616

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., et al. (2015). UK Biobank: an Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *Plos Med.* 12, e1001779. doi:10.1371/journal.pmed.1001779

Sun, J. Y., Zhang, H., Zhang, Y., Wang, L., Sun, B. L., Gao, F., et al. (2021). Impact of Serum Calcium Levels on Total Body Bone mineral Density: A Mendelian Randomization Study in Five Age Strata. *Clin. Nutr.* 40, 2726–2733. doi:10.1016/j.clnu.2021.03.012

Tillmann, T., Vaucher, J., Okbay, A., Pikhart, H., Peasey, A., Kubinova, R., et al. (2017). Education and Coronary Heart Disease: Mendelian Randomisation Study. *BMJ* 358, j3542. doi:10.1136/bmj.j3542

Verbanck, M., Chen, C. Y., Neale, B., and Do, R. (2018). Detection of Widespread Horizontal Pleiotropy in Causal Relationships Inferred from Mendelian Randomization between Complex Traits and Diseases. *Nat. Genet.* 50, 693–698. doi:10.1038/s41588-018-0099-7

Wang, L., Qiao, Y., Zhang, H., Zhang, Y., Hua, J., Jin, S., et al. (2020). Circulating Vitamin D Levels and Alzheimer's Disease: A Mendelian Randomization Study in the IGAP and UK Biobank. *J. Alzheimers Dis.* 73, 609–618. doi:10.3233/JAD-190713

Xiuyun, W., Qian, W., Minjun, X., Weidong, L., and Lizhen, L. (2020). Education and Stroke: Evidence from Epidemiology and Mendelian Randomization Study. *Sci. Rep.* 10, 21208. doi:10.1038/s41598-020-78248-8

Xu, L., Lin, S. L., and Schooling, C. M. (2017). A Mendelian Randomization Study of the Effect of Calcium on Coronary Artery Disease, Myocardial Infarction and Their Risk Factors. *Sci. Rep.* 7, 42691. doi:10.1038/srep42691

Yavorska, O. O., and Burgess, S. (2017). MendelianRandomization: an R Package for Performing Mendelian Randomization Analyses Using Summarized Data. *Int. J. Epidemiol.* 46, 1734–1739. doi:10.1093/ije/dyx034

Zhang, H., Wang, T., Han, Z., and Liu, G. (2020). Mendelian Randomization Study to Evaluate the Effects of Interleukin-6 Signaling on Four Neurodegenerative Diseases. *Neurol. Sci.* 41, 2875–2882. doi:10.1007/s10072-020-04381-x

Zhao, W., Rasheed, A., Tikkanen, E., Lee, J. J., Butterworth, A. S., Howson, J. M. M., et al. (2017). Identification of New Susceptibility Loci for Type 2 Diabetes and Shared Etiological Pathways with Coronary Heart Disease. *Nat. Genet.* 49, 1450–1457. doi:10.1038/ng.3943

Zheng, J., Zhang, Y., Rasheed, H., Walker, V., Sugawara, Y., Li, J., et al. (2022). Trans-ethnic Mendelian-Randomization Study Reveals Causal Relationships between Cardiometabolic Factors and Chronic Kidney Disease. *Int. J. Epidemiol.* 50, 1995–2010. doi:10.1093/ije/dyab203

Zhu, Z., Zheng, Z., Zhang, F., Wu, Y., Trzaskowski, M., Maier, R., et al. (2018). Causal Associations between Risk Factors and Common Diseases Inferred from GWAS Summary Data. *Nat. Commun.* 9, 224. doi:10.1038/s41467-017-02317-2

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Genetics

ORIGINAL RESEARCH
published: 22 February 2022
doi: 10.3389/fgene.2022.795479

Circulating N-Terminal Probrain Natriuretic Peptide Levels in Relation to Ischemic Stroke and Its Subtypes: A Mendelian Randomization Study

Ming Li[1†], *Yi Xu*[1,2†], *Jiaqi Wu*[3], *Chuanjie Wu*[2], *Ang Li*[4] *and Xunming Ji*[1,2,3]*

[1]China-America Institute of Neurology, Xuanwu Hospital, Capital Medical University, Beijing, China, [2]Department of Neuroscience, Xuanwu Hospital, Capital Medical University, Beijing, China, [3]Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, School of Biological Science and Medical Engineering, Beihang University, Beijing, China, [4]Department of Biomedical Engineering, Columbia University, New York City, NY, United States

Mendelian randomization was used to evaluate the potential causal association between N-terminal probrain natriuretic peptide (NT-proBNP) and ischemic stroke based on summary statistics data from large-scale genome-wide association studies. Three single-nucleotide polymorphisms (SNPs) rs198389, rs13107325, and rs11105306 associated with NT-proBNP levels found in large general populations and in patients with acute heart disease were used as instrumental variables. The results of genetic association analysis of each single SNP show that there is no significant association between NT-proBNP levels and ischemic stroke or its subtypes, whereas rs198389 alone has a suggestive association with large-artery atherosclerosis stroke. The MR analysis of three SNPs shows that NT-proBNP levels may reduce the risk of small-vessel occlusion stroke suggestively. This genetic analysis provides insights into the pathophysiology and treatment of ischemic stroke.

Keywords: mendelian randomization, single nucleotide polymorphisms, N-terminal pro-brain natriuretic peptide, stroke, risk predictor

OPEN ACCESS

Edited by:
Liangcai Zhang,
Janssen Research and Development,
United States

Reviewed by:
Jun Zhang,
Texas Tech University Health Sciences
Center, United States
Natalia V. Rivera,
Karolinska Institutet (KI), Sweden

****Correspondence:***
Xunming Ji
jixm@ccmu.edu.cn

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to
Statistical Genetics and Methodology,
a section of the journal
Frontiers in Genetics

Received: *15 October 2021*
Accepted: *10 January 2022*
Published: *22 February 2022*

Citation:
Li M, Xu Y, Wu J, Wu C, Li A and Ji X (2022) Circulating N-Terminal Probrain Natriuretic Peptide Levels in Relation to Ischemic Stroke and Its Subtypes: A Mendelian Randomization Study.
Front. Genet. 13:795479.
doi: 10.3389/fgene.2022.795479

INTRODUCTION

Stroke is the second major cause of global death with a mortality rate of approximately 5.5 million/year and poses a huge financial burden to family members and public health (Donkor, 2018). The study of INTERSTROKE presents 10 potentially modifiable risk factors that are associated with around 90% of acute strokes (O'Donnell et al., 2016), and according to the data from the INTERHEART study, those factors also account for the great majority of the risk of myocardial infarction (Yusuf et al., 2004). Therefore, it is generally acknowledged that a bidirectional interaction exists between brain damage and heart dysfunction (Chen et al., 2017; Scheitz et al., 2018), which may share overlapping cell death pathways (Gonzales-Portillo et al., 2016).

Abbreviations: AAE, Age at examination; AAO, Age at onset; ACS, Acute coronary syndromes; AIS, Acute ischemic stroke; ANP, Atrial natriuretic peptide; ARIC, Atherosclerosis Risk in Communities; BNP, Brain natriuretic peptide; CAD, Coronary artery disease; CES, Cardioembolism stroke; CI, Confidence interval; GWAS, Genome-wide association study; IVW, Inverse-variance weighted; LAS, Large-artery atherosclerosis stroke; MR, Mendelian randomization; NPPB, Natriuretic peptide precursor B; NT-proBNP, N-terminal pro-brain natriuretic peptide; OR, Odds ratio; SD, Standard deviation; SNP, Single-nucleotide polymorphism; SVS, Small-vessel occlusion stroke.

N-terminal probrain natriuretic peptide (NT-proBNP) is an N-terminal fragment of brain natriuretic peptide (BNP), released from the heart muscle in response to the blood pressure and volume overload (Daniels and Maisel, 2007). This factor is widely used in the clinic as a prognostic biomarker to predict mortality in patients with coronary artery disease (CAD), atrial fibrillation, and heart failure (Johansson et al., 2016). Compared with BNP, NT-proBNP presents a longer circulating half-life, higher plasma concentration, and greater diagnostic sensitivity. Due to the connections between cardiac dysfunction and stroke, NT-proBNP is supposed to be a potential predictor for the risk of ischemic stroke (Zhao et al., 2020a).

The relationship between NT-proBNP and risks of stroke remains a popular research subject. The related research dates back to 1996 (Rubattu et al., 1996). The scientific community has increasing interest in this area from 2010 (García-Berrocoso et al., 2010; Kim et al., 2010; Quan et al., 2010) to 2020 (Bhatia et al., 2020; Harpaz et al., 2020; Hotsuki et al., 2020; Khan and Kamal, 2020; Medranda et al., 2020; Rubattu et al., 2020; Shirotani et al., 2020; Tonomura et al., 2020; Wang et al., 2020; Watson et al., 2020; Yang et al., 2020; Zhao et al., 2020b). Several studies explore and identify variable degrees of correlation in different types of stroke. The data from the population-based Atherosclerosis Risk in Communities (ARIC) study shows that NT-proBNP was associated positively with total stroke, non-lacunar ischemic, as well as cardioembolic stroke, but not with lacunar or hemorrhagic stroke (Folsom et al., 2013). NT-proBNP is a strong predictor of atrial fibrillation, which makes it a contributor to the incidence of cardioembolic stroke (Yang et al., 2014). A recent study indicates that serum levels of NT-proBNP higher than 800 pg/ml obtained within 72 h after a transient ischemic attack were associated with an increased risk of stroke (Rodríguez-Castro et al., 2020). More interestingly, in 2019, based on the Biomarkers for Cardiovascular Risk Assessment in Europe-Consortium, Castelnuovo et al. (2019) analyzed data of 58,173 participants free of stroke from six community-based cohort studies and found that, in the European group, levels of NT-proBNP have positive association with risk of ischemic and hemorrhagic stroke, independent from several other conditions and risk factors. These findings cannot be easily explained by the known physiological function of BNP.

The role of NT-proBNP in the incidence of stroke became an unsolved question. A meta-analysis of 16 studies suggests that NT-proBNP provides minor clinical predictive values for the prediction of stroke mortality (García-Berrocoso et al., 2013). According to the research of George et al. (Giannakoulas et al., 2005), no significant correlation was observed between NT-proBNP levels and stroke severity or infarct volume. Another study also denied this association in terms of functional outcomes (Etgen et al., 2005). Evidence suggests the causal relationships of natriuretic peptides to endothelial permeability, which might predispose people to atherosclerosis and hemorrhages (Lee et al., 2007; Lin et al., 2007; Kuhn, 2012; Cannone et al., 2013). Therefore, some researchers hypothesized that NT-proBNP may be involved in the causal physiological path for stroke incidence or be a causal risk factor of stroke (Cushman et al., 2014; Di Castelnuovo et al., 2019). However, a large number of studies confirms that BNP is a protective factor of CAD and a self-regulator of the body's pathological state. The release of BNP improves myocardial relaxation and response to the acute increase of ventricular volume by opposing sodium retention, vasoconstriction, and antidiuretic effects of the activated renin-angiotensin-aldosterone system (Daniels and Maisel, 2007). All of these findings suggest that BNP may also have a protective role in stroke.

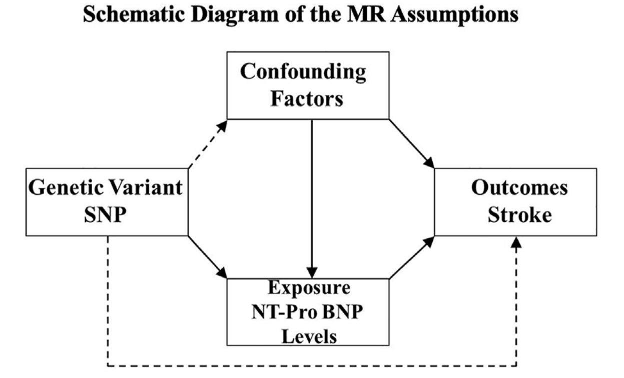

Schematic Diagram of the MR Assumptions
Confounding Factors
Genetic Variant SNP
Outcomes Stroke
Exposure NT-Pro BNP Levels

FIGURE 1 | Schematic diagram of the MR assumptions. The arrows represent possible causal associations between variables. The dashed lines represent possible causal associations between variables that would violate the MR assumptions.

As a result, available clinical observational studies investigating the association between NT-proBNP and risk of stroke show ambiguous results. The confounding factors of the observational studies may cause BNP levels to rise, but this increase is not one of the causes of stroke; and those studies cannot rule out some implicit risk factors of stroke.

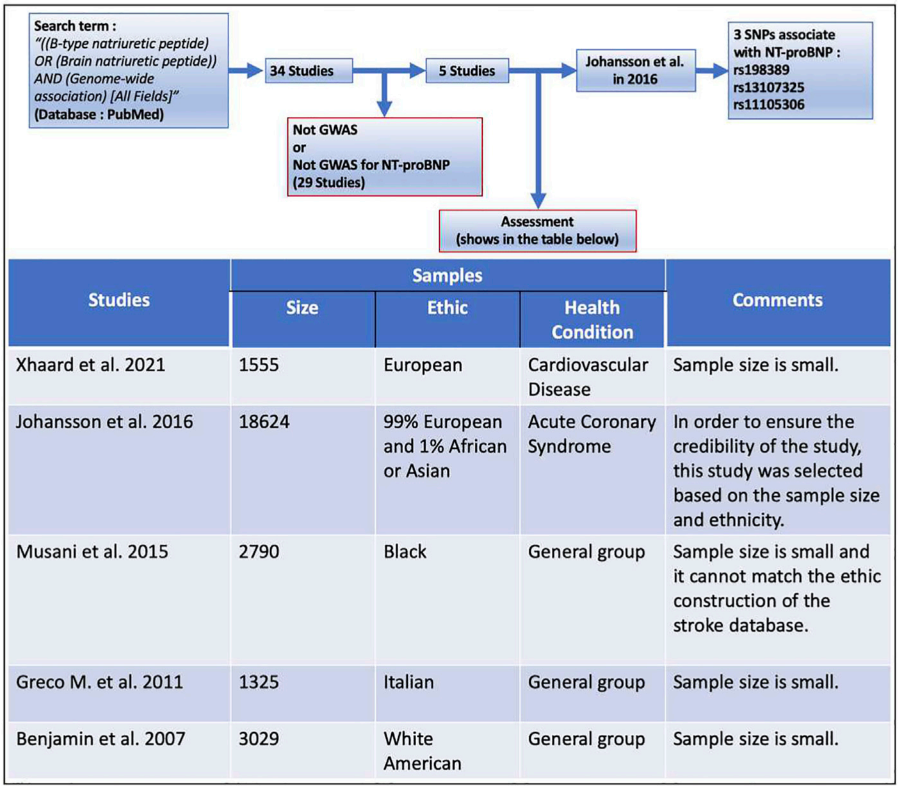
To circumvent the limitations of observational studies, Mendelian randomization (MR) analysis was used to improve causal inference. This technique is based on the premise that human genetic variants are randomly distributed among the population. This method may avoid the potential confounding factors within the exposure–outcome relationship and provide insight into the genetic association between the circulating NT-proBNP levels and ischemic stroke (**Figure 1**). Therefore, we conducted an MR analysis to investigate the causal effect of NT-proBNP on ischemic stroke and its subtypes (cardioembolism stroke, small-vessel occlusion stroke, and large-artery atherosclerosis stroke) by using three single-nucleotide polymorphisms (SNPs) (rs198389, rs13107325, rs11105306) associated with NT-pro-BNP level (Johansson et al., 2016).

MANUSCRIPT FORMATTING

Methods

Selection of Instrumental Variables

To select SNPs associated with NT-proBNP as instrumental variables, the term "[(B-type natriuretic peptide) OR (Brain natriuretic peptide)] AND (Genome-wide association) (All


Search term :
"((B-type natriuretic peptide) OR (Brain natriuretic peptide)) AND (Genome-wide association) [All Fields]"
(Database : PubMed)
34 Studies
5 Studies
Johansson et al. in 2016
3 SNPs associate with NT-proBNP :
rs198389
rs13107325
rs11105306
Not GWAS
or
Not GWAS for NT-proBNP
(29 Studies)
Assessment
(shows in the table below)

Studies	Samples			Comments
	Size	Ethic	Health Condition	
Xhaard et al. 2021	1555	European	Cardiovascular Disease	Sample size is small.
Johansson et al. 2016	18624	99% European and 1% African or Asian	Acute Coronary Syndrome	In order to ensure the credibility of the study, this study was selected based on the sample size and ethnicity.
Musani et al. 2015	2790	Black	General group	Sample size is small and it cannot match the ethic construction of the stroke database.
Greco M. et al. 2011	1325	Italian	General group	Sample size is small.
Benjamin et al. 2007	3029	White American	General group	Sample size is small.

FIGURE 2 | The retrieval process and inclusion/discharge criteria of instrumental variables.

Fields)" was searched in PubMed from 2005 to 2021, and the results showed a total of 34 articles (**Supplementary Appendix**). There are only five studies that found SNPs associated with NT-proBNP, of which the genome-wide association study (GWAS) performed by Johansson et al. (2016) was selected for our study (the retrieval process and inclusion/discharge criteria are shown in **Figure 2**). This GWAS of 18,624 individuals with acute coronary syndrome consisting of 99% European and 1% African or Asian identified two novel SNPs in SCL39A8 (rs13107325, pooled $p = 5.99 \times 10^{-10}$) and POC1B/GALANT4 (rs11105306, pooled $p = 1.02 \times 10^{-16}$) and confirmed one SNP (rs198389, pooled $p = 1.07 \times 10^{-15}$) that were all associated with the serum level of NT-proBNP. Among these three BNPs, rs198389 is proven to be associated with the level of NT-proBNP in several studies. The first study of this SNP was reported in 2007. This GWAS surrounding the natriuretic peptide precursor B (NPPB) gene with plasma BNP levels was performed in 2,970 adults from the general population (Takeishi et al., 2007). NPPB is on chromosome 1, encoding pre-proBNP. rs198389 is located in the NPPB promoter and has previously been found to influence promoter activity by interrupting an E-box consensus motif in the gene promoter (Meirhaeghe et al., 2007; Johansson et al., 2016). The rs13107325 is located in SLC39A8 on chromosome 4. It is a missense variant, which may cause an amino acid change at position 391 of the protein (Johansson et al., 2016). This substitution is predicted to be deleterious to the protein (Johansson et al., 2016). The rs11105306 is located in POC1B/GALANT4 on chromosome 12, which is in an intronic region with no obvious regulatory function (Johansson et al., 2016).

Outcome Data

The statistical data used for MR analysis of genetic associations with stroke was obtained from a multi-ancestry GWAS, including data from 521,612 individuals (67,162 cases and 454,450 controls) (Malik et al., 2018). These participants were selected from 29 investigations, consisting of ancestry groups from European (40,585 cases and 406,111 controls), East Asian (17,369 cases and 28,195 controls), African (5,541 cases and 15,154 controls), Latin American (865 cases and 692 controls), mixed Asian (365 cases and 333 controls), and South Asian (2,437 cases and 6,707 controls) (Malik et al., 2018). To avoid bias produced by a multi-ancestry population, we only used the data from the European group. The MEGASTROKE project was approved by relevant institutional review boards, and informed consent was obtained from each participant. The data set and basic information including sample size, age, and gender composition are presented in **Table 1**.

Statistical Analysis

First, we conducted genetic association analysis to evaluate the association between single NT-proBNP-associated SNPs and

TABLE 1 | The data set and basic information of the stroke GWAS in 2018.

Dataset	Stroke			Control		
	N	% Female	Mean AAO	*N*	% Female	Mean AAE
Metastroke	20,000	44.4%	67.1	19,326	49.9%	61.0
NINDS-SIGN	7,743	46.1%	66.5	17,970	—	—
Charge	4,348	67.0%	75.8	80,613	—	63.7
EPIC-CVD	4,347	48.0%	70.1	7,897	60.2%	64.1
Barcelona	520	41.9%	69.1	315	37.7%	67.5
Biobank Japan	16,256	36.8%	69.9	27,294	60.4%	57.5
CADISP	555	38.9%	43.7	9,259	—	—
Compass	5,541	—	—	15,154	—	—
Decode	5,520	44.2%	78.7	254,000	49.9%	53.3
Glasgow	599	49.7%	69.9	1,775	48.8%	69.6
Finland	501	40.9%	64.0	1,813	—	—
Hisayama	1,113	39.1%	69.7	901	40.5%	69.4
HVH—All	805	65.7%	68.3	1901	50.3%	66.4
Interstroke	2,429	44.3%	64.0	2,128	47.6%	62.5
MDC	202	34.7%	62.9	4,925	59.4%	57.2
RACE1	1,218	47.6%	50.1	1,158	47.0%	51.9
RACE2	1,167	—	—	4,035	—	—
SAHLSIS	298	40.9%	59.3	596	35.6%	56.8
SDS	52	46.2%	55.7	1,514	46.4%	53.0
SIFAP	981	38.9%	41.7	1825	50.7%	55.2
SLESS	546	42.1%	66.2	868	47.9%	58.7
UK young lacunar stroke DNA	1,403	32.8%	60.6	968	47.5%	59.7
ICH	1,545	45.1%	67.0	1,481	40.5%	65.3

AAO, age at onset; AAE, age at examination.

ischemic stroke and its three subtypes (cardioembolism, small-vessel occlusion, and large-artery atherosclerosis strokes). The significance threshold is $p < .005$, considering that many association studies for a single test changed the p value from .05 to .005, and the results with p values between .05 and .005 were considered to be suggestive of significance. Second, we conducted the MR analysis using three MR methods, including inverse-variance weighted (IVW), weighted median, and MR-Egger. IVW is the main MR analysis method, which combines the variant-specific Wald estimators by taking the inverse of their approximate variances as the corresponding weights (Bowden et al., 2016). Weighted median could derive consistent estimates when up to 50% of instruments are not valid (Bowden et al., 2016). MR-Egger could test the presence of potential pleiotropy and account for this potential pleiotropy using the MR-Egger intercept test (Burgess and Thompson, 2017). The odds ratio (OR) as well as 95% confidence interval (CI) of stroke corresponds to about 1 standard deviation (SD) in NT-proBNP level. All the statistical tests were completed using R Packages "Mendelian Randomization" (Yavorska and Burgess, 2017) and a $p < .0042$ (0.05/12 adjusted with Bonferroni method) was considered statistically significant; p between .05 and .0042 were considered suggestive of significance.

Results and Discussion

The genetic association analysis evaluating the association between single NT-proBNP-associated SNPs and ischemic stroke and its three subtypes shows that neither of those SNPs have significant association with ischemic stroke and subtypes, whereas only rs198389 has a suggestive association with LAS (95% CI 0.017~0.116, p = .008686, .05 > p > .005) (**Table 2**). The MR analysis using three MR methods (IVW, weighted median, MR-Egger) shows no significant causal association between BNP levels and the risk of ischemic stroke. However, the weighted median and the IVW present suggestive association in small-vessel occlusion stroke (SVS) (weighted median: OR = −0.268, 95% CI −0.492~−0.044, p = .019; IVW: OR = −0.199, 95% CI −0.389~−0.009, p = .040) with no horizontal pleiotropy, which was identified with the MR-egger method (p = .499) (**Table 3**; **Figure 3**). In conclusion, the genetic association analysis shows that rs198389 has a suggestive association with LAS, and the MR analysis shows that NT-proBNP levels suggestively reduce the risk of SVS.

Analysis of the Negative Results

This MR study overcomes confounding risk factors and shows that there is no significant causal association between BNP levels and the risk of ischemic stroke, which is contrary to the results of most previous prospective studies. The discrepancy therein may be ascribed to the negligence of some hidden risk factors for stroke, which may cause BNP levels to rise without any causal association with stroke. In 2013, in a random community-based sample, Cannone et al. found that rs5065 was associated with increased cardiovascular risk by analyzing the phenotype associated with atrial natriuretic peptide (ANP) genetic variant rs5065. The rs5065 is a genetic variant and its minor allele encodes for an ANP with two additional arginines at the C-terminus, ANP-RR. This research also found that the

TABLE 2 | The genetic association analysis of BNPs and ischemic stroke and its subtypes.

SNP	Stroke types	Allele1	Allele2	Freq1[a]	Effect	StdErr[b]	*p*-value
rs198389	AIS[c]	a	g	0.5846	0.0093	0.0103	0.367
	LAS[d]	a	g	0.5833	0.0667	0.0254	0.008686
	CES[e]	a	g	0.5851	−0.0126	0.0196	0.5212
	SVS[f]	a	g	0.5838	0.0406	0.0236	0.0856
rs13107325	AIS	t	c	0.0748	−0.0065	0.0215	0.7611
	LAS	t	c	0.0802	0.0206	0.0529	0.6965
	CES	t	c	0.0769	−0.0299	0.0435	0.4921
	SVS	t	c	0.0766	0.0234	0.0475	0.6219
rs11105306	AIS	t	c	0.2461	0.0058	0.0123	0.6384
	LAS	t	c	0.2432	−0.0213	0.0294	0.4695
	CES	t	c	0.2439	0.0146	0.0229	0.5232
	SVS	t	c	0.2447	0.0492	0.0271	0.06945

[a] *Frequence.*
[b] *Standard error.*
[c] *Acute ischemic stroke.*
[d] *Large-artery atherosclerosis stroke.*
[e] *Cardioembolism stroke.*
[f] *Small-vessel occlusion stroke.*

TABLE 3 | MR analysis of association between 3 BNPs (rs198389, rs13107325, rs11105306) and ischemic stroke and its subtypes.

Stroke types	Method	Estimate	Std. error	95% CI	*p*-value
IS	Weighted median	−0.04	0.049	−0.136, 0.056	0.415
	IVW	−0.044	0.043	−0.129, 0.041	0.313
	MR-Egger	0.073	0.370	−0.653, 0.799	0.843
	MR-Egger (intercept)	−0.020	0.063	−0.144, 0.104	0.751
LAS	Weighted median	0.100	0.149	−0.191, 0.391	0.501
	IVW	−0.107	0.105	−0.314, 0.099	0.308
	MR-Egger	2.042	0.908	0.263, 3.821	0.024
	MR-Egger (intercept)	−0.371	0.156	−0.676, −0.066	0.017
CES	Weighted median	−0.031	0.095	−0.218, 0.156	0.746
	IVW	−0.023	0.083	−0.185, 0.139	0.779
	MR-Egger	−0.808	0.720	−2.219, 0.604	0.262
	MR-Egger (intercept)	0.135	0.123	−0.106, 0.376	0.273
SVS	Weighted median	−0.268	0.114	−0.492, −0.044	0.019
	IVW	-0.199	0.097	−0.389, −0.009	0.040
	MR-Egger	0.631	0.933	−1.197, 2.459	0.499
	MR-Egger (intercept)	−0.144	0.160	−0.458, 0.170	0.370

endothelial hyperpermeability induced by chronic exposure to ANP-RR may predispose the subject to atherosclerotic disease. Interestingly, the minor allele of rs5065 is associated with higher BNP plasma values. The researchers hypothesized that higher levels of BNP might be originated from the deleterious effects caused by ANP-RR on the heart although it did not reveal any other CAD signs (Cannone et al., 2013). The rs5065 causes both ANP-RR and BNP levels to increase, but only ANP-RR is the causal factor. Pathways such as this may exist in the incidence of stroke and lead to controversial results. Although some evidence suggests causal relationship between natriuretic peptides and endothelial permeability, which might predispose to atherosclerosis and hemorrhages, some research shows that BNP may also have anti-inflammatory endothelial actions (Kuhn, 2012). These two actions are contrary to each other, which may explain the difference between the results mentioned above and our result.

Possible Explanations of the Suggestive Associations

The contradiction stated in the previous paragraph leads us to focus on the suggestive associations found in this study. The genetic association analysis shows that rs198389 alone has a suggestive association with LAS. The MR analysis shows that NT-proBNP levels have a suggestive positive causal effect on LAS in MR-Egger analysis (OR = 2.042, 95% CI 0.263–3.821, $p = .024$), but the MR-Egger intercept (95%CI −0.676~−0.066, $p = .017$) is significantly different from zero, showing a pleiotropic effect on this outcome. The origin of loci may affect the results. In this study, the rs198389 locus came from a large population without special classification, and the other two loci came from people with ACS in GWAS performed by Johansson et al. In our study, genes as instrumental variables need to be absolutely associated with exposure factors. However, the association between NT-proBNP and the loci found in ACS patients is questionable. Therefore, we cannot conclude that NT-proBNP has no causal

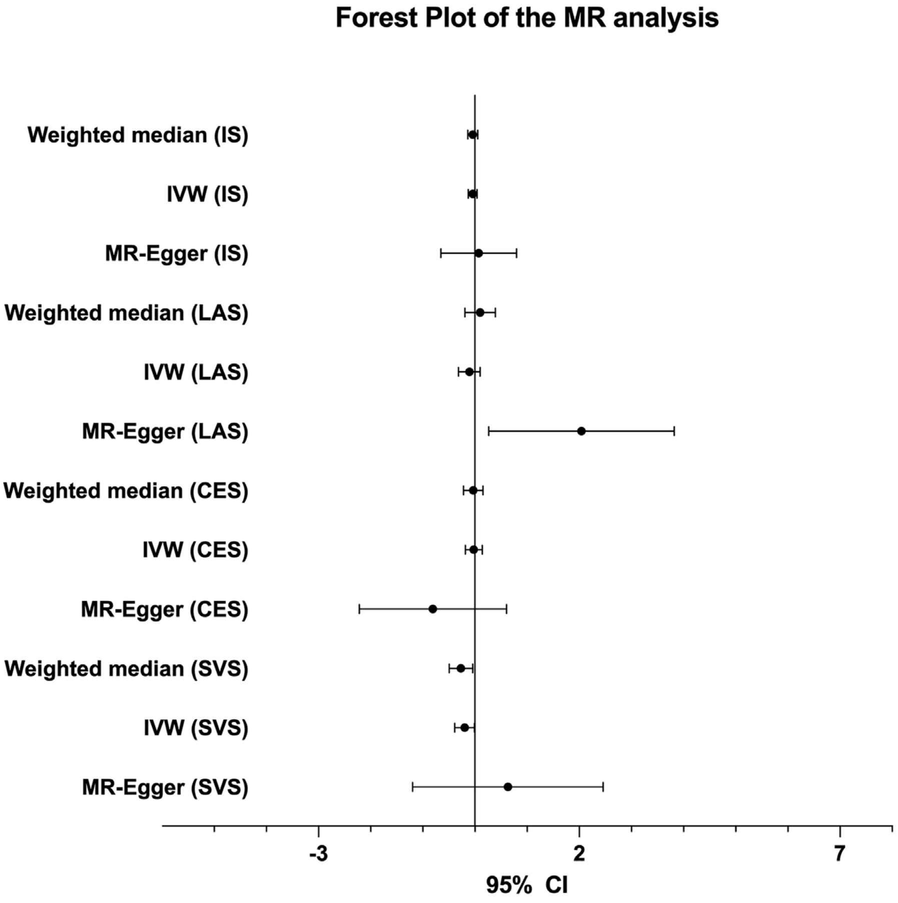


FIGURE 3 | The forest plot of the MR analysis.

relationship with stroke merely based on this study. We chose the GWAS performed by Johansson et al. as the SNPs source because it has the largest sample size among all of the available GWAS of NT-proBNP (**Figure 2**). This is based on the idea that many of the current limitations of GWAS can be overcome to some extent by increasing sample sizes, which makes GWAS with larger sample sizes more reliable (Tam et al., 2019). Therefore, GWAS of NT-proBNP in general populations with a large sample size is anticipated to explore the relationship between SNP and stroke more accurately.

Interestingly, in our study, it is also implied that the serum level of NT-proBNP suggestively reduces the risk of SVS. The role of BNP in lowering blood pressure may be involved in the mechanism behind this phenomenon. BNP is released from the heart muscle in response to blood pressure and volume overload. Its main effects are reducing the preload of the heart by promoting diuresis and capillary permeability, which results in the reduction of the blood pressure (Goetze et al., 2020). In 2013, Wang et al. performed a retrospective study on the association between hypertension and different ischemic stroke subtypes, which involved 11,560 patients with ischemic stroke. The results show that hypertension is significantly related to recurrent stroke in patients with SVS, but not other subtypes of ischemic stroke (Wang et al., 2013). Taken together, we conclude that BNP can reduce the risk of SVS by lowering blood pressure. Whether BNP can reduce the risk of SVS needs to be verified by more accurate and credible studies in the future, which will help us form a better understanding of the pathogenesis and treatment of SVS.

Strengths and Limitations

Our MR study has several strengths. First, stroke is a complex disease with a large number of risk factors and pathophysiological pathways. However, in this study, the relationship between NT-proBNP and stroke was studied at the gene level with a large sample size and directly from the gene, which reduces the possibility of interference from implied risk factors. Second, in this study, the potential confounding factors caused by linkage disequilibrium may be reduced by using three independent genetic variants as instrumental variables. Third, we selected three MR methods to enhance the robustness of estimates. Fourth, three-stage pleiotropy analysis were performed, which may decrease the risk of pleiotropy.

Some limitations still exist in this MR analysis. First, the additional confounders cannot be completely ruled out as well as for the pleiotropy present in any MR study. Second, the obtained analysis results may be influenced by the population stratification, which cannot be fully ruled out. Third, the genetic relationship between NT-proBNP levels and stroke risk may be different in diverse genetic ancestries or ethnicities.

This genetic association should be further evaluated in other ancestries. Fourth, a replication study should be performed to ensure the accuracy and rigor of our original study. However, the GWAS of stroke we used as outcome data had very large sample size. It conducted meta-analyses of 29 studies, which involved every large size of stroke-related database before 2018. As we know, there are no other relative studies that have approximately the same order of magnitude as the previous GWAS. Replication studies should be performed with another large GWAS of ischemic stroke.

CONCLUSION

This research provides evidence that there is no causal relationship between elevated NT-proBNP level and the risk of stroke. It is ineffective to use NT-proBNP as the target for stroke treatment and prevention. NT-proBNP plays an important role in ischemic stroke, but its function is not completely clear, and its association with stroke needs to be further explored.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

ML and YX wrote the first draft of the manuscript; YX performed the data collection and statistical analysis; JW, CW, and AL contributed to manuscript revision; ML and XJ contributed conception and design of the study; XJ takes full responsibility for the data, the analyses and interpretation, and the conduct of the research. All authors read and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (Nos 82102220, 82027802, 61975017, 82071468), Beijing Municipal Natural Science Foundation (721220) and General Projects of Scientific and Technological Plan of Beijing Municipal Education Commission (KM202010025023).

ACKNOWLEDGMENTS

We would like to thank all the people who participated in this study for their cooperation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.795479/full#supplementary-material

REFERENCES

Bhatia, R., Warrier, A. R., Sreenivas, V., Bali, P., Sisodia, P., Gupta, A., et al. (2020). Role of Blood Biomarkers in Differentiating Ischemic Stroke and Intracerebral Hemorrhage. *Neurol. India* 68 (4), 824–829. doi:10.4103/0028-3886.293467

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* 40 (4), 304–314. doi:10.1002/gepi.21965

Burgess, S., and Thompson, S. G. (2017). Interpreting Findings from Mendelian Randomization Using the MR-Egger Method. *Eur. J. Epidemiol.* 32 (5), 377–389. doi:10.1007/s10654-017-0255-x

Cannone, V., Huntley, B. K., Olson, T. M., Heublein, D. M., Scott, C. G., Bailey, K. R., et al. (2013). Atrial Natriuretic Peptide Genetic Variant Rs5065 and Risk for Cardiovascular Disease in the General Community: a 9-year Follow-Up Study. *Hypertension* 62 (5), 860–865. doi:10.1161/hypertensionaha.113.01344

Chen, Z., Venkat, P., Seyfried, D., Chopp, M., Yan, T., and Chen, J. (2017). Brain-Heart Interaction: Cardiac Complications after Stroke. *Circ. Res.* 121 (4), 451–468. doi:10.1161/CIRCRESAHA.117.311170

Cushman, M., Judd, S. E., Howard, V. J., Kissela, B., Gutiérrez, O. M., Jenny, N. S., et al. (2014). N-terminal Pro-B-type Natriuretic Peptide and Stroke Risk: the Reasons for Geographic and Racial Differences in Stroke Cohort. *Stroke* 45 (6), 1646–1650. doi:10.1161/strokeaha.114.004712

Daniels, L. B., and Maisel, A. S. (2007). Natriuretic Peptides. *J. Am. Coll. Cardiol.* 50 (25), 2357–2368. doi:10.1016/j.jacc.2007.09.021

Di Castelnuovo, A., Veronesi, G., Costanzo, S., Zeller, T., Schnabel, R. B., de Curtis, A., et al. (2019). NT-proBNP (N-Terminal Pro-B-type Natriuretic Peptide) and the Risk of Stroke. *Stroke* 50 (3), 610–617. doi:10.1161/strokeaha.118.023218

Donkor, E. S. (2018). Stroke in the21stCentury: A Snapshot of the Burden, Epidemiology, and Quality of Life. *Stroke Res. Treat.* 2018, 1–10. doi:10.1155/2018/3238165

Etgen, T., Baum, H., Sander, K., and Sander, D. (2005). Cardiac Troponins and N-Terminal Pro-brain Natriuretic Peptide in Acute Ischemic Stroke Do Not Relate to Clinical Prognosis. *Stroke* 36 (2), 270–275. doi:10.1161/01.STR.0000151364.19066.a1

Folsom, A. R., Nambi, V., Bell, E. J., Oluleye, O. W., Gottesman, R. F., Lutsey, P. L., et al. (2013). N-terminal Pro-B-type Natriuretic Peptide, and Incidence of Stroke: the Atherosclerosis Risk in Communities Study. *Stroke* 44 (4), 961–967. doi:10.1161/strokeaha.111.000173

García-Berrocoso, T., Fernández-Cadenas, I., Delgado, P., Rosell, A., and Montaner, J. (2010). Blood Biomarkers in Cardioembolic Stroke. *Curr. Cardiol. Rev.* 6 (3), 194–201. doi:10.2174/157340310791658767

García-Berrocoso, T., Giralt, D., Bustamante, A., Etgen, T., Jensen, J. K., Sharma, J. C., et al. (2013). B-type Natriuretic Peptides and Mortality after Stroke: a Systematic Review and Meta-Analysis. *Neurology* 81 (23), 1976–1985. doi:10.1212/01.wnl.0000436937.32410.32

Giannakoulas, G., Hatzitolios, A., Karvounis, H., Koliakos, G., Charitandi, A., Dimitroulas, T., et al. (2005). N-terminal Pro-brain Natriuretic Peptide Levels Are Elevated in Patients with Acute Ischemic Stroke. *Angiology* 56 (6), 723–730. doi:10.1177/000331970505600610

Goetze, J. P., Bruneau, B. G., Ramos, H. R., Ogawa, T., de Bold, M. K., and de Bold, A. J. (2020). Cardiac Natriuretic Peptides. *Nat. Rev. Cardiol.* 17 (11), 698–717. doi:10.1038/s41569-020-0381-0

Gonzales-Portillo, C., Ishikawa, H., Shinozuka, K., Tajiri, N., Kaneko, Y., and Borlongan, C. V. (2016). Stroke and Cardiac Cell Death: Two Peas in a Pod. *Clin. Neurol. Neurosurg.* 142, 145–147. doi:10.1016/j.clineuro.2016.01.001

Harpaz, D., Seet, R. C. S., Marks, R. S., and Tok, A. I. Y. (2020). B-type Natriuretic Peptide as a Significant Brain Biomarker for Stroke Triaging Using a Bedside

Point-of-Care Monitoring Biosensor. *Biosensors (Basel)* 10 (9). doi:10.3390/bios10090107

Hotsuki, Y., Sato, Y., Yoshihisa, A., Watanabe, K., Kimishima, Y., Kiko, T., et al. (2020). B-type Natriuretic Peptide Is Associated with post-discharge Stroke in Hospitalized Patients with Heart Failure. *ESC Heart Fail.* 7 (5), 2508–2515. doi:10.1002/ehf2.12818

Johansson, Å., Eriksson, N., Lindholm, D., Varenhorst, C., James, S., Syvänen, A. C., et al. (2016). Genome-wide Association and Mendelian Randomization Study of NT-proBNP in Patients with Acute Coronary Syndrome. *Hum. Mol. Genet.* 25 (7), 1447–1456. doi:10.1093/hmg/ddw012

Khan, S., and Kamal, M. A. (2020). Cardiac Biomarkers in Stroke, Alzheimer's Disease, and Other Dementia. Are They of Use? A Brief Overview of Data from Recent Investigations. *CNS Neurol. Disord. Drug Targets* 20 (8), 687–693. doi:10.2174/1871527319666201005171003

Kim, M. H., Kang, S. Y., Kim, M. C., and Lee, W. I. (2010). Plasma Biomarkers in the Diagnosis of Acute Ischemic Stroke. *Ann. Clin. Lab. Sci.* 40 (4), 336–341.

Kuhn, M. (2012). Endothelial Actions of Atrial and B-type Natriuretic Peptides. *Br. J. Pharmacol.* 166 (2), 522–531. doi:10.1111/j.1476-5381.2012.01827.x

Lee, J. M., Zhai, G., Liu, Q., Gonzales, E. R., Yin, K., Yan, P., et al. (2007). Vascular Permeability Precedes Spontaneous Intracerebral Hemorrhage in Stroke-Prone Spontaneously Hypertensive Rats. *Stroke* 38 (12), 3289–3291. doi:10.1161/strokeaha.107.491621

Lin, K., Kazmi, K. S., Law, M., Babb, J., Peccerelli, N., and Pramanik, B. K. (2007). Measuring Elevated Microvascular Permeability and Predicting Hemorrhagic Transformation in Acute Ischemic Stroke Using First-Pass Dynamic Perfusion CT Imaging. *AJNR Am. J. neuroradiology* 28 (7), 1292–1298. doi:10.3174/ajnr.A0539

Malik, R., Chauhan, G., Traylor, M., Sargurupremraj, M., and Yamaji, T. (2018). Multiancestry Genome-wide Association Study of 520,000 Subjects Identifies 32 Loci Associated with Stroke and Stroke Subtypes. *Nat. Genet.* 50 (4), 524–537. doi:10.1038/s41588-018-0058-3

Medranda, G. A., Salhab, K., Schwartz, R., and Green, S. J. (2020). Prognostic Implications of Baseline B-type Natriuretic Peptide in Patients Undergoing Transcatheter Aortic Valve Implantation. *Am. J. Cardiol.* 130, 94–99. doi:10.1016/j.amjcard.2020.06.017

Meirhaeghe, A., Sandhu, M. S., McCarthy, M. I., de Groote, P., Cottel, D., Arveiler, D., et al. (2007). Association between the T-381C Polymorphism of the Brain Natriuretic Peptide Gene and Risk of Type 2 Diabetes in Human Populations. *Hum. Mol. Genet.* 16 (11), 1343–1350. doi:10.1093/hmg/ddm084

O'Donnell, M. J., Chin, S. L., Rangarajan, S., Xavier, D., Liu, L., Zhang, H., et al. (2016). Global and Regional Effects of Potentially Modifiable Risk Factors Associated with Acute Stroke in 32 Countries (INTERSTROKE): a Case-Control Study. *Lancet (London, England)* 388 (10046), 761–775. doi:10.1016/s0140-6736(16)30506-2

Quan, H. X., Jin, J. Y., Wen, J. F., and Cho, K. W. (2010). beta(1)-Adrenergic Receptor Activation Decreases ANP Release via cAMP-Ca2+ Signaling in Perfused Beating Rabbit Atria. *Life Sci.* 87 (7-8), 246–253. doi:10.1016/j.lfs.2010.06.022

Rodríguez-Castro, E., Hervella, P., López-Dequidt, I., Arias-Rivas, S., Santamaría-Cadavid, M., López-Loureiro, I., et al. (2020). NT-pro-BNP: A Novel Predictor of Stroke Risk after Transient Ischemic Attack. *Int. J. Cardiol.* 298, 93–97. doi:10.1016/j.ijcard.2019.06.056

Rubattu, S., Stanzione, R., Cotugno, M., Bianchi, F., Marchitti, S., and Forte, M. (2020). Epigenetic Control of Natriuretic Peptides: Implications for Health and Disease. *Cell Mol Life Sci* 77 (24), 5121–5130. doi:10.1007/s00018-020-03573-0

Rubattu, S., Volpe, M., Kreutz, R., Ganten, U., Ganten, D., and Lindpaintner, K. (1996). Chromosomal Mapping of Quantitative Trait Loci Contributing to Stroke in a Rat Model of Complex Human Disease. *Nat. Genet.* 13 (4), 429–434. doi:10.1038/ng0896-429

Scheitz, J. F., Nolte, C. H., Doehner, W., Hachinski, V., and Endres, M. (2018). Stroke–heart Syndrome: Clinical Presentation and Underlying Mechanisms. *Lancet Neurol.* 17 (12), 1109–1120. doi:10.1016/s1474-4422(18)30336-3

Shirotani, S., Minami, Y., Saito, C., Haruki, S., and Hagiwara, N. (2020). B-type Natriuretic Peptide and Outcome in Patients with Apical Hypertrophic Cardiomyopathy. *J. Cardiol.* 76 (4), 357–363. doi:10.1016/j.jjcc.2020.03.015

Takeishi, Y., Toriyama, S., Takabatake, N., Shibata, Y., Konta, T., Emi, M., et al. (2007). Linkage Disequilibrium Analyses of Natriuretic Peptide Precursor B Locus Reveal Risk Haplotype Conferring High Plasma BNP Levels. *Biochem. biophysical Res. Commun.* 362 (2), 480–484. doi:10.1016/j.bbrc.2007.08.028

Tam, V., Patel, N., Turcotte, M., Bosse, Y., Pare, G., and Meyre, D. (2019). Benefits and Limitations of Genome-wide Association Studies. *Nat. Rev. Genet.* 20 (8), 467–484. doi:10.1038/s41576-019-0127-1

Tonomura, S., Ihara, M., and Friedland, R. P. (2020). Microbiota in Cerebrovascular Disease: A Key Player and Future Therapeutic Target. *J. Cereb. Blood Flow Metab.* 40 (7), 1368–1380. doi:10.1177/0271678x20918031

Wang, A., Zhang, M., Ding, Y., Mo, X., Zhong, C., Zhu, Z., et al. (2020). Associations of B-type Natriuretic Peptide and its Coding Gene Promoter Methylation with Functional Outcome of Acute Ischemic Stroke: A Mediation Analysis. *J. Am. Heart Assoc.* 9 (18), e017499. doi:10.1161/JAHA.120.017499

Wang, Y., Xu, J., Zhao, X., Wang, D., Wang, C., Liu, L., et al. (2013). Association of Hypertension with Stroke Recurrence Depends on Ischemic Stroke Subtype. *Stroke* 44 (5), 1232–1237. doi:10.1161/STROKEAHA.111.000302

Watson, C. J., Tea, I., O'Connell, E., Glezeva, N., Zhou, S., James, S., et al. (2020). Comparison of Longitudinal Change in sST2 vs BNP to Predict Major Adverse Cardiovascular Events in Asymptomatic Patients in the Community. *J. Cel Mol Med* 24 (11), 6495–6499. doi:10.1111/jcmm.15004

Yang, H. L., Lin, Y. P., Long, Y., Ma, Q. L., and Zhou, C. (2014). Predicting Cardioembolic Stroke with the B-type Natriuretic Peptide Test: A Systematic Review and Meta-Analysis. *J. Stroke Cerebrovasc. Dis.* 23 (7), 1882–1889. doi:10.1016/j.jstrokecerebrovasdis.2014.02.014

Yang, M., Tao, L., An, H., Liu, G., Tu, Q., Zhang, H., et al. (2020). A Novel Nomogram to Predict All-Cause Readmission or Death Risk in Chinese Elderly Patients with Heart Failure. *ESC Heart Fail.* 7 (3), 1015–1024. doi:10.1002/ehf2.12703

Yavorska, O. O., and Burgess, S. (2017). MendelianRandomization: an R Package for Performing Mendelian Randomization Analyses Using Summarized Data. *Int. J. Epidemiol.* 46 (6), 1734–1739. doi:10.1093/ije/dyx034

Yusuf, S., Hawken, S., Ounpuu, S., Dans, T., Avezum, A., Lanas, F., et al. (2004). Effect of Potentially Modifiable Risk Factors Associated with Myocardial Infarction in 52 Countries (The INTERHEART Study): Case-Control Study. *The Lancet* 364 (9438), 937–952. doi:10.1016/s0140-6736(04)17018-9

Zhao, J. J., Zhang, Y., Yuan, F., Song, C. G., Jiang, Y. L., Gao, Q., et al. (2020). Diagnostic Value of N-Terminal Pro B-type Natriuretic Peptide for Nonvalvular Atrial Fibrillation in Acute Ischemic Stroke Patients: A Retrospective Multicenter Case-Control Study. *J. Neurol. Sci.*, 414, 116822. doi:10.1016/j.jns.2020.116822

Zhao, Y. H., Gao, H., Pan, Z. Y., Li, J., Huang, W. H., Wang, Z. F., et al. (2020). Prognostic Value of NT-proBNP after Ischemic Stroke: A Systematic Review and Meta-Analysis of Prospective Cohort Studies. *J. stroke Cerebrovasc. Dis.* 29 (4), 104659. doi:10.1016/j.jstrokecerebrovasdis.2020.104659

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 04 March 2022
doi: 10.3389/fgene.2022.799913

Genome-Wide Analysis for the Regulation of Gene Alternative Splicing by DNA Methylation Level in Glioma and its Prognostic Implications

Zeyuan Yang[1], *Yijie He*[1], *Yongheng Wang*[1,2], *Lin Huang*[1], *Yaqin Tang*[1], *Yue He*[3], *Yihan Chen*[1] *and Zhijie Han*[1]*

[1]*Department of Bioinformatics, School of Basic Medicine, Chongqing Medical University, Chongqing, China,* [2]*International Research Laboratory of Reproduction and Development, Chongqing Medical University, Chongqing, China,* [3]*Group of Mathematics Education Teaching and Research, Chongqing Fudan Secondary School, Chongqing, China*

OPEN ACCESS

Edited by:
Liangcai Zhang,
Janssen Research and Development (United States), United States

Reviewed by:
Xuefeng Gu,
Shanghai University of Medicine and Health Sciences, China
Benoit Lhermitte,
Hôpitaux Universitaires de Strasbourg, France

***Correspondence:**
Zhijie Han
zhijiehan@cqmu.edu.cn

Specialty section:
This article was submitted to "Statistical Genetics and Methodology", a section of the journal Frontiers in Genetics

Received: *22 October 2021*
Accepted: *20 January 2022*
Published: *04 March 2022*

Citation:
Yang Z, He Y, Wang Y, Huang L, Tang Y, He Y, Chen Y and Han Z (2022) Genome-Wide Analysis for the Regulation of Gene Alternative Splicing by DNA Methylation Level in Glioma and its Prognostic Implications. Front. Genet. 13:799913. doi: 10.3389/fgene.2022.799913

Glioma is a primary high malignant intracranial tumor with poorly understood molecular mechanisms. Previous studies found that both DNA methylation modification and gene alternative splicing (AS) play a key role in tumorigenesis of glioma, and there is an obvious regulatory relationship between them. However, to date, no comprehensive study has been performed to analyze the influence of DNA methylation level on gene AS in glioma on a genome-wide scale. Here, we performed this study by integrating DNA methylation, gene expression, AS, disease risk methylation at position, and clinical data from 537 low-grade glioma (LGG) and glioblastoma (GBM) individuals. We first conducted a differential analysis of AS events and DNA methylation positions between LGG and GBM subjects, respectively. Then, we evaluated the influence of differential methylation positions on differential AS events. Further, Fisher's exact test was used to verify our findings and identify potential key genes in glioma. Finally, we performed a series of analyses to investigate influence of these genes on the clinical prognosis of glioma. In total, we identified 130 glioma-related genes whose AS significantly affected by DNA methylation level. Eleven of them play an important role in glioma prognosis. In short, these results will help to better understand the pathogenesis of glioma.

Keywords: glioma, alternative splicing, methylation modification, clinical prognosis, TCGA

INTRODUCTION

Glioma is the most common and highly malignant primary intracranial tumor which is characterized by substantial heterogeneity and extremely poor prognosis in central nervous system (CNS) (Dong and Cui 2020; Pan et al., 2021). The World Health Organization (WHO) defines grade IV glioma as the glioblastoma (GBM). The annual incidence of this disease worldwide is about 5 cases per 100,000 people (Hottinger et al., 2014), and shows a significant mortality and unclarified molecular mechanism of the occurrence and development (Hottinger et al., 2014; Dong and Cui 2020). Although the etiology of glioma has been extensively studied, there are still many challenges and unknowns in the epigenetic mechanism of its pathogenesis and progress (Molinaro et al., 2019).

Recently, the DNA methylation has been demonstrated to extensively participate in the epigenetic mechanisms of CNS (Hwang et al., 2017), and many methyltransferase and demethylase-related genes (e.g., MGMT, CD44, HYAL2, SPP1, MMP2) contribute to the pathogenesis of glioma (Weller et al., 2010; Wiestler et al., 2014; Xiao et al., 2020). A large amount of the evidence showed that DNA methylation is involved in the occurrence and development of glioma tumors (Etcheverry et al., 2010; Chen et al., 2020; Dong and Cui 2020). For example, in GBM patients, the disease-related important signaling pathways (e.g., RB1 and TP53) are affected by CpG island promoter hypermethylation (Etcheverry et al., 2010). The promoter methylation of DNA repair enzymes (O6-methylguanine-DNA methyltransferase) has been identified as a significant prognostic factor for temozolomide resistance in GBM patients (Chen et al., 2020).

Conversely, the previous studies reported that pathogenesis of glioma is significantly associated with the dysregulated alternative splicing (AS) in the brain (Mogilevsky et al., 2018; Pattwell et al., 2020; Zeng et al., 2020). AS is the primary driving force behind generating diverse proteins, which is the basis for the remarkable and complex functional regulation seen in eukaryotic cells (Xie et al., 2019). Genome-wide studies showed that 90–95% of human genes undergo some level of AS, and almost one-third of them were proved to be generated multiple protein isoforms (Kim et al., 2014; Wang et al., 2021). These processes usually show an extreme complexity in brain tissues and can play an important role in the progression of many CNS diseases (Merkin et al., 2012; Galarza-Munoz et al., 2017; Consortium 2020). For glioma, for instance, Mogilevsky et al. discovered that the manipulation of MKNK2 AS significantly suppressed the oncogenic properties of GBM cells and resensitized the cells to chemotherapy (Mogilevsky et al., 2018). Pattwell et al. found that a truncated splice variant, TrkB.T1, increases PDGF-induced Akt and STAT3 signaling and further enhances PDGF-driven GBM *in vivo* (Pattwell et al., 2020). Moreover, many previous studies indicate that there is a strong link between DNA methylation and AS and it generally contributes to the pathogenesis of CNS disorders, including glioma (Feng et al., 2019; Li et al., 2019). For example, transcriptome analysis revealed that PTEN methylation influences mature mRNA transcripts through modulation of pre-mRNA AS, and the methylation-defective PTEN R159K mutant is found most frequently in glioma patients. There was mark dysregulation of splicing factors in the PTEN-deficient GBM samples (Feng et al., 2019). The important oncogene METTL3 is a methyltransferase and it is found to modulate the nonsense-mediated mRNA decay of splicing factors and AS isoform switches in GBM. The methylation modification of serine- and arginine-rich splicing factors by METTL3 promotes GBM tumor growth and progression (Li et al., 2019).

However, so far, there has been no systematic study to explore the relationship between glioma-related DNA methylation and gene AS in the whole genome scale, and the influence of their interaction on the pathogenesis and progress of glioma. Therefore, in this study, we performed a genome-wide analysis by integrating the DNA methylation and AS data of 537 low-grade glioma (LGG) and GBM individuals. First, we downloaded the relevant data from the Cancer Genome Atlas (TCGA), TCGA SpliceSeq and EWASdb database, respectively. Second, we conducted the differential analysis between LGG and GBM samples to identify the glioma-related methylation positions and AS events. Third, based on the results, we performed a splicing quantitative trait methylation loci (defined as me-sQTL (Gutierrez-Arcelus et al., 2015; Han and Lee 2017)) analysis to explore the influence of DNA methylation level on gene AS in glioma. Fourth, we further explored the characteristics of these me-sQTLs and affected AS events. Fifth, combining the data of disease risk methylation positions from EWASdb, we performed the two-tailed Fisher's exact test to investigate the disease specificity of the me-sQTLs and identify the potential key genes related to them in glioma. Finally, based on these potential key genes and clinical data, we conducted the least absolute shrinkage, univariate Cox regression, selection operator (LASSO) regression, clinical correlation and survival analysis to explore the influence of these genes whose AS events affected by DNA methylation on clinical prognosis of glioma. The flow chart is shown in **Figure 1**.

MATERIALS AND METHODS

Data Collection and Processing

Clinical and methylation information of glioma patients was downloaded from the TCGA database (http://cancergenome.nih.gov), a comprehensive resource containing multi-omics data from various cancers. According to the annotation of TCGA, glioma is classified as the LGG and the GBM. TCGA is a global genomic profiling project that utilizes high-throughput microarray technologies to identify molecular subtype classifications of cancers, multigene clinical predictors, new targets for drug therapy, and predictive markers for these therapies (Vigneswaran et al., 2015). The International Classification of Diseases for Oncology has been used for nearly 25 years as a tool for coding diagnoses of neoplasms in tumor and cancer registrars and in pathology laboratories (Warzel et al., 2003). Data analysis was performed with the glioma classification LGG and GBM provided by the TCGA database. Current glioma classifications are based on the 2007 WHO grading scale, which separates gliomas based on cytologic features and degrees of malignancy after hematoxylin and eosin (H&E) staining (Erridge et al., 2011). According to the classification of gliomas in the TCGA database, data analysis is carried out by using the classifications LGG and GBM of gliomas provided by the TCGA database. We accessed these TCGA data using the Genomic Data Commons (GDC) data portal (https://portal. gdc. cancer.gov/). Particularly, based on our previous study (He et al., 2020), we first selected "DNA methylation" for the Data Category, "Illumina human methylation 450" for the Platform, "brain" for the Primary Site and "gliomas" in the Disease Type to screen out the suitable methylation array of patients in the GDC data portal. Then, the"clinical," "brain" and "gliomas" were selected to the Data Category, Primary Site and Disease Type, respectively, to screen out the clinical information of patients in the GDC data portal. Finally, we removed samples that lacked methylation or clinical information.

The AS events of these samples were obtained from the TCGA SpliceSeq database (http://bioinformatics.mdanderson.org/TCGASpliceSeq), which identifies AS events and describes their genome profiles using the RNA-seq data of the TCGA

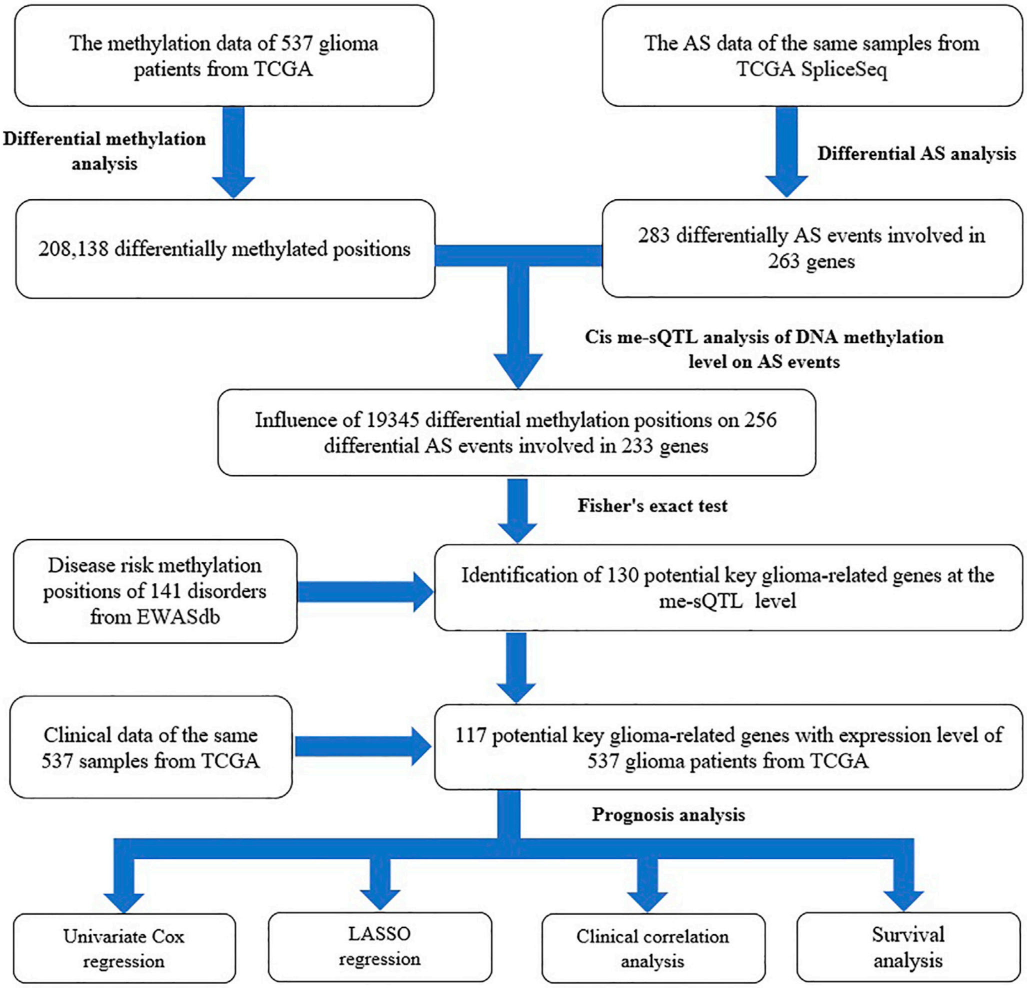

The methylation data of 537 glioma patients from TCGA
The AS data of the same samples from TCGA SpliceSeq
Differential methylation analysis
Differential AS analysis
208,138 differentially methylated positions
283 differentially AS events involved in 263 genes
Cis me-sQTL analysis of DNA methylation level on AS events
Influence of 19345 differential methylation positions on 256 differential AS events involved in 233 genes
Fisher's exact test
Disease risk methylation positions of 141 disorders from EWASdb
Identification of 130 potential key glioma-related genes at the me-sQTL level
Clinical data of the same 537 samples from TCGA
117 potential key glioma-related genes with expression level of 537 glioma patients from TCGA
Prognosis analysis
Univariate Cox regression
LASSO regression
Clinical correlation analysis
Survival analysis

FIGURE 1 | The flow chart of the study design for exploring the influence of DNA methylation level on gene AS in glioma and its impact on disease prognosis.

samples (Ryan et al., 2016). Particularly, we downloaded the AS isoform average percent spliced-in (PSI) values of the LGG and GBM samples, respectively, from TCGA SpliceSeq database with the common parameter settings (i.e., the percentage of samples with PSI value >75%, minimum PSI range >0 and minimum PSI standard deviation >0.1) according to the previous studies (Yang et al., 2019; Rong et al., 2020; Wei et al., 2021). Based on the classification criteria of TCGA SpliceSeq, we classified the types of AS events into Alternate Acceptors (AA), Alternate Donors (AD), Exon Skip (ES), Retained Intron (RI), Alternate Promoters (AP), Alternate Terminators (AT) and Mutually Exclusive Exons (ME). The AS events that are not present in both LGG and GBM samples were removed.

Moreover, the information of disease risk methylation positions was obtained from the EWASdb database (http://www.bioapp.org/ewasdb/index.php/Index/index). EWASdb is a specialized epigenome-wide association database which stores the results of 1,319 epigenome-wide association study (EWAS) studies involved in the 302 diseases/phenotypes with the threshold for significance $p < 1 \times 10^{-7}$ (Liu et al., 2019). We downloaded the EWAS single epi-marker and annotation files (phenotype/disease info) and merged the files by the disease codes.

Differential Analysis of Methylation Positions

To obtain the glioma-related methylation positions, we performed differential methylation analysis between GBM and LGG samples. In particular, we used a Subset-quantile Within Array Normalization method to preprocess the methylation data by the R package “minfi,”, a specialized tool for the analysis of the Illumina methylation 450 array dataset (http://bioconductor.org/packages/release/bioc/html/minfi.html) (Aryee et al., 2014). Then, the quality control of methylation array was conducted “densityBeanPlot” function of this package. The characteristics of the qualified samples show that the methylation levels (beta values) of CpG positions are distributed around 0 and 1, respectively. Finally, based on the qualified methylation array data, we used a bump-hunting algorithm to identify the differentially methylated positions between GBM and LGG subjects by the "dmpFinder" function of this package. The parameter was set by its default value (i.e., type = "categorical") and the significance level was set according to a common threshold for the absolute intercept ≥0.2 (i.e. 20% difference on the beta values) and the p value $<1 \times 10^{-3}$ (Guo et al., 2015).

Differential Analysis of Alternative Splicing Events and Annotation

To identify the glioma-related AS events and corresponding genes, we performed the differential AS events analysis and gene annotation. Particularly, the differential AS events analysis was conducted by the vast-tools software (Irimia et al., 2014). Based on the PSI of each AS event, we performed a Bayesian inference-based differential AS analysis by the "diff" function of vast-tools software with its default parameters. According to the previous studies, we set the threshold for significance at the minimum value for absolute value of differential PSI between GBM and LGG samples (MV|ΔPSI|) at 0.95 confidence level greater than 10% (Ha et al., 2021; Hekman et al., 2021). The gene annotation was conducted by g:Profiler toolset, a web server for conversions between gene identifiers and functional annotation (Raudvere et al., 2019). We used the g:Profiler to identify these AS events corresponding genes, convert their ID and annotate the genome location and type of the genes. The annotation file (hg19) from the database (release 75) were used for these analyses (Aken et al., 2017).

Association Analysis Between DNA Methylation and Alternative Splicing

To explore the effect of methylation on AS events in glioma, we performed a cis me-sQTL analysis by combining the PSI values of differential AS events and the beta values of differentially methylated positions from the same samples. Particularly, we first considered the distance between the differentially methylated positions and the transcription initiation site (TSS) of differential AS events corresponding genes less than 1 M as the cis region, and selected all methylation positions and AS event pairs that met the conditions for the cis me-sQTL analysis. The annotation files of the Illumina methylation 450 array dataset (hg19) and Ensembl database (release 75) were used to locate the genomic locations of the methylated positions and the TSS of AS events corresponding genes, respectively. Then, based on the beta values of the differentially methylated positions in combination with the PSI values of the corresponding differential AS events, we used a linear regression model to perform a cis me-sQTL analysis by the R package "Matrix eQTL" with the parameters, age, and gender as covariates (Shabalin 2012). Finally, we conducted a multiple testing by Benjamini–Hochberg method to correct the p values of the cis me-sQTL analysis and set false discovery rate (FDR) q value less than 0.05 as the threshold for significance level according to the previous studies (Gillies et al., 2018; Drag et al., 2019; Han et al., 2020).

Disease Specificity Analysis of the Cis Me-sQTLs

In order to explore the disease specificity of these cis me-sQTLs and further verify our findings as well as identify the potential key glioma-related genes with affected AS events by methylation level, we performed the two-tailed Fisher's exact test by combining the disease risk methylation positions and the results of cis me-sQTLs analysis. Particularly, we first produced the disease risk methylation position datasets for various disorders including glioma from EWASdb database (Liu et al., 2019). Then, we defined the methylation positions which were unlikely to have an effect on the AS events in cis region ($p > 0.05$) as the non me-sQTLs. Next, by the two-tailed Fisher's exact test, we compared the proportions of all these cis and non me-sQTLs in the disease risk methylation positions for each of the disorders to explore the disease specificity and further verify previous findings. The threshold for significance level was set as the p value <0.05. Finally, to identify the potential key glioma-related genes at the me-sQTL level, we compared the proportions of cis and non me-sQTLs in glioma-related methylation positions for each gene using the two-tailed Fisher's exact test (the threshold of $p < 0.05$). The "fisher.test" function of R was used for these calculations.

Influence of the Me-sQTL Genes on Clinical Prognosis of Glioma

We further analyzed the influence of these potential key genes whose AS events are affected by DNA methylation on clinical prognosis of glioma. First, we calculated the average expression of these genes in each individual and separated the samples into low and high expression groups according to the median of average expression. Then, we used the Kaplan-Meier overall survival curves to compare prognosis between the high expression and low expression individuals. Next, we performed a univariate Cox regression analysis to assess the association between these me-sQTL genes and the prognosis of glioma. The threshold of significance was set at 95% confidence interval (CI) of hazard ratio (HR) $\not\supseteq$ 1 and $p < 0.05$. Then, based on the results of univariate Cox regression analysis, the R package "glmnet" was used to perform the LASSO regression analysis, a fit algorithm based on cyclical coordinate descent and warm start search along a regularization path, to identify the main glioma prognosis-related genes (Simon et al., 2011). According to the common parameter settings, the maxit and alpha were set at 1,000 and 1, respectively, and others were set by their default values. Based on the results, the risk scores were calculated for each subject by the R package "survival" (http://CRAN.R-project.org/package=survival). Further, the receiver operator characteristic (ROC) curve was used to verify the reliability of these risk scores by the R package "survivalROC" (https://CRAN.R-project.org/package=survivalROC). Finally, we used the chi-square test to assess the association between the expression of these glioma prognosis-related genes and other clinical features of the patients, which included the age at initial pathologic diagnosis, the vital status, and the gender. The threshold for significance was set at the p value <0.05.

TABLE 1 | Summary of the 537 individuals studied in this work.

Individuals	Sample Type	Sample Size	Mean Age (Aken et al.)	Male/Female (Han and Lee)	Death Rates (Han and Lee)
GBM subjucts	Primary Tumor	51	61.54 (13.41)	56.00/44.00	66.00
LGG subjucts	Primary Tumor	486	42.91 (13.42)	54.64/45.36	25.15
Total		537	44.66 (14.48)	54.77/45.23	28.97

These samples are from our previous study (He et al., 2020).

RESULTS AND DISCUSSION

The Multi-Omics Data From 537 Glioma Individuals

In total, we obtained the datasets of DNA methylation values, AS events PSI values, gene expression levels and clinical information from 537 glioma samples (including 486 LGG and 51 GBM patients). The summary of these glioma samples was shown in **Table 1**. Particularly, after the missing value filtering and normalization processing, we quantified a total of 369,531 CpGs methylation positions with the normalized values according to the annotation files of Illumina human methylation 450 array. The results of normalization processing were shown in the **Supplementary Figure S1** and our previous study (He et al., 2020). We obtained 7,414 AS events with the PSI values from 537 glioma samples the TCGA SpliceSeq database. These AS events are composed of about 39.0% ES, 27.8% AP, 11.3% AT, 8.4% RI, 6.9% AD, 5.9% AA and 0.5% ME types (**Figure 2A**). The expression data of 20,530 genes of the glioma samples was downloaded from the TCGA database and quantified by RSEM values. The clinical information of these samples contains age, gender, survival time, and vital status. Moreover, after the combination of same disease types and missing value filtering, we obtained a total of 141 disease risk methylation position data sets from the EWASdb database.

Differential Analysis of Methylation Positions and Alternative Splicing Events

We performed a differential methylation analysis between the LGG and GBM subjects to identify the glioma-related DNA methylation positions. All of the methylation array data met quality control metrics. The results showed that the beta values of DNA methylation positions are mainly distributed around 0 and 1, respectively, for each sample. The details are described in the **Supplementary Figure S2** and our previous study (He et al., 2020). By the differential methylation analysis, we identified a total of 208,138 positions with a significantly different methylation level between LGG and GBM subjects. The results are shown in the **Supplementary Table S1** and our previous study (He et al., 2020).

To identify the glioma-related AS events, we further conducted differential AS events between the LGG and GBM subjects. According to the significance threshold MV|ΔPSI| at 0.95 confidence level ≥10%, we identified a total of 287 differential AS events between LGG and GBM subjects. These differential AS events belonged to 263 genes (**Supplementary Table S2**). **Figure 3** shows the most significant differential AS events (SpliceSeq ID: 96726) of LPHN3 gene (MV|ΔPSI| at 0.95 confidence level = 0.25). A recent study reported that LPHN3 was an important paralog of EVA1C which leads to the high infiltration levels of multiple immune cells in glioma (Hu and Qu 2021). Moreover, according to the classification criteria for SpliceSeq database, about 35.5%, 31.0%, 14.0%, 9.1%, 5.2% and 5.2% of these identified AS events are categorized into ES, AP, AT, RI, AD and AA types, respectively (**Figure 2A**). We did not find a significant difference in the proportion of AS event types when compared with the original AS event type proportion by the two-tailed Fisher's exact test (**Figure 2A**). This revealed a typological universality of the differential AS events in glioma.

Association Analysis Between DNA Methylation and Alternative Splicing

Combining the PSI values of differential AS events with the beta values of differentially methylated positions from the same samples, we used a linear regression model to perform the cis me-sQTL analysis by R package "Matrix eQTL" with the parameters, age, and gender serving as covariates. In total, we identified 19,345 methylated positions affecting 256 AS events which are involved in 233 genes (over 88% of the total differential genes) with a significance level of FDR $q < 0.05$. This revealed a general influence of DNA methylation level on gene AS in glioma. The top 25 significant results are shown in **Table 2** (the full information is listed in the **Supplementary Table S3**). Among the 256 affected AS events, we found that about 34.1%, 32.2%, 14.6%, 9.6%, 4.6% and 5.0% of these affected AS events are categorized into ES, AP, AT, RI, AD and AA types, respectively. By the two-tailed Fisher's exact test, we also did not find a significant difference of percentage between the affected and the original AS event types (**Figure 2A**). This revealed a typological non-specific regulation of gene AS by DNA methylation level in glioma. Further, we explored the relationship between the significance of regulation of the cis me-sQTLs and the distance of them to the TSS of the corresponding affected gene, and their distribution characteristics in genome. The results showed that these cis me-sQTLs tended to be distributed in the proximity of the corresponding affected gene TSS, and there were more significant regulatory effects of them in these regions (**Figure 2B**). This was consistent with the findings of previous studies (Pangeni et al., 2018; Chen and Elnitski 2019).

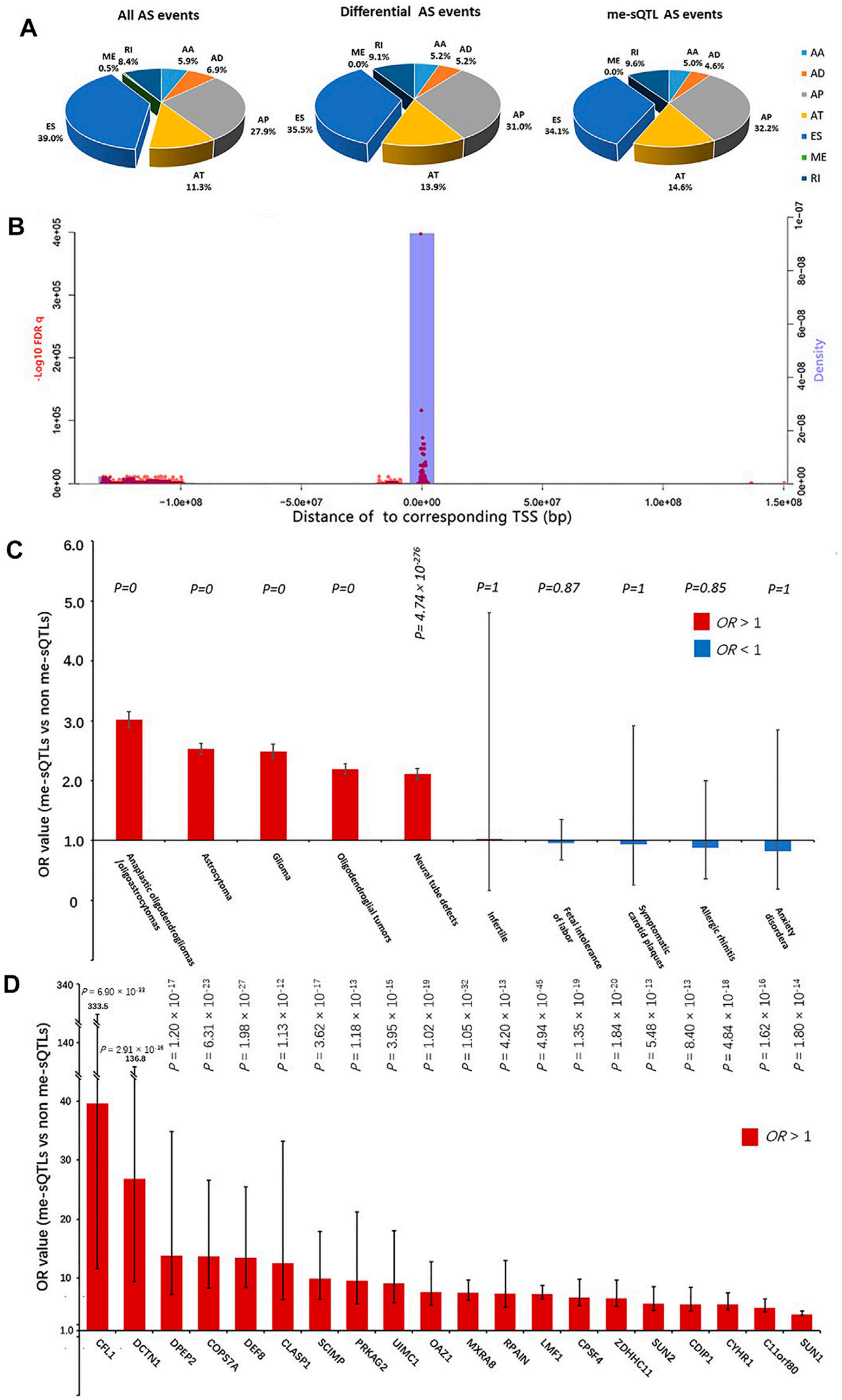


FIGURE 2 | The characteristic of the cis me-sQTLs and the affected AS events. **(A)** The pie charts show the proportion in all (left), differential (middle) and DNA methylation affected AS events (right) annotated with each class (AA, AD, ES, RI, AP, AT and ME), respectively. **(B)** The blue bar graphs indicate the relationship between the abundance of the cis me-sQTLs and the distance of them to TSS of corresponding AS events. The red dots indicate the relationship between the statistical significance of the cis me-sQTLs associated with AS and the distance of them to TSS of corresponding AS events. **(C)** The disease specificity of the cis me-sQTLs by the two-tailed Fisher's exact test. **(D)** The glioma specificity of the cis me-sQTLs in each gene by the two-tailed Fisher's exact test. The black bars in histogram represent 95% confidence intervals.

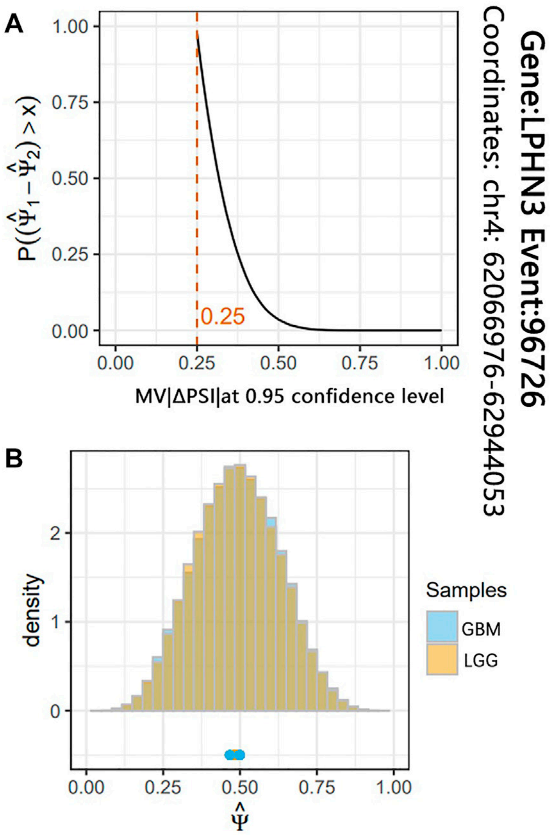


FIGURE 3 | The results of differential analysis for the AS event 96726 of LPHN3 gene. **(A)** The red line indicates that the maximum probability of ΔPSI of AS event 96726 between LGG and GBM subjects is greater than 0.25. **(B)** The histogram shows the two joint posterior distributions over PSI and the point estimates for each replicate.

Disease Specificity Analysis of the Cis Me-sQTLs

To explore the disease specificity of these cis me-sQTLs and verify our findings, as well as further identify the potential key glioma-related genes at the me-sQTL level, we performed the two-tailed Fisher's exact test using the disease risk methylation position data from the EWASdb database. We found that the risk methylation positions of all the 141 diseases are overlapped with the cis me-sQTLs and non me-sQTLs. By comparing the proportions of cis me-sQTLs and non me-sQTLs in each disease risk methylation position dataset (the threshold of Fisher's p value <0.05), we found that the cis me-sQTLs significantly enriched the risk methylation position dataset of 103 diseases, which are mainly composed of CNS disorders and malignant tumor diseases including glioma (odds ratio (OR) = 2.49, p = 0). In contrast, the remaining 38 diseases, whose risk methylation positions are not significantly enriched by the cis me-sQTLs, are mainly composed of the other types of disorders, e.g., the rheumatic heart disease ($p = 5.99 \times 10^{-1}$), septicemia ($p = 6.97 \times 10^{-1}$), and Infertile (p = 1). The top 5 most and least significant results are shown in **Figure 2C** (the full information is listed in the **Supplementary Table S4**). The results revealed the specificity and similarity of neuro-oncological disorders at the me-sQTL level and verified the association of the cis me-sQTLs we identified with glioma. Further, for each type of AS event and each gene, we compared the proportions of their cis and non me-sQTLs in glioma risk methylation position dataset, respectively. The results showed that the cis me-sQTLs of almost all types of AS events are significantly enriched in glioma risk methylation position dataset, i.e., AA (OR = 4.76, $p = 1.18 \times 10^{-36}$), AT (OR = 2.85, $p = 9.63 \times 10^{-66}$), ES (OR = 2.39, $p = 6.26 \times 10^{-112}$), RI (OR = 2.05, $p = 2.85 \times 10^{-30}$), AD (OR = 2.55, $p = 4.78 \times 10^{-21}$), and AP (OR = 2.88, $p = 8.69 \times 10^{-191}$), and there are a total of 130 genes whose cis me-sQTLs are significantly enriched in glioma risk methylation position dataset (p < 0.05). **Figure 2D** shows the top 20 significant results and full information is listed in the **Supplementary Table S5**. We considered that these genes are more correlated with the pathogenesis of glioma at the me-sQTL level and selected them for the following prognosis analysis of glioma.

Influence of the Me-sQTL Genes on Clinical Prognosis of Glioma

We further analyzed the influence of the potential key genes which are associated with glioma in me-sQTL level on the clinical prognosis of glioma. The expression data were obtained from TCGA database and these data are involved in 117 of the 130 potential key genes. We found that the overall survival curve of the subjects with high expression of these genes is significantly longer than the subjects with low expression ($p = 7.56 \times 10^{-1}$ (**Figure 4A**). This revealed that the expression dysregulation of these potential key genes is significantly associated with the bad prognosis of glioma patients. To avoid dependence between the 117 genes and identify the main glioma prognosis-related genes, we performed the univariate Cox regression analysis of the 117 genes. However, the results showed that 61me-sQTL genes identified are high-risk factors for the prognosis of glioma subjects (i.e. 95% CI HR $\not\supseteq$ 1 and p < 0.001) (**Supplementary Figure S3**). We discover that both over-expression of those 30 genes and under-expression of the other 31 genes can lead to a poor prognosis in glioma patients, which is also consistent with common sense. given that patients are in advanced stages of the disease and their survival may be affected by other complications or factors. Then, we further applied the LASSO regression algorithm to conduct the selection and calculate the risk score of each subject to univariate Cox regression results. The results showed that there are 11 genes (i.e., KIF3A, HAUS1, TMCC1, BEND7, B3GNT5, MTMR3, ITGB3, BICD1, EXTL3, SUN1 and MXRA8) identified when the cross-validated partial likelihood deviance reaches its minimum value (**Figures 4B,C**). Among the 11 genes, the coefficients of 7 were positive (i.e., increase risk of disease), and others were negative (i.e., decrease risk of disease). A previous study reported that the low expression of the TMCC1 gene confers poor clinical prognoses of glioma patients which is in accordance with our findings (Pangeni et al., 2018). The area under the curve (AUC) of the ROC is 0.988, which reveals the reliability of the risk score (**Figure 4D**). According to the median

TABLE 2 | The top 25 significant results of the me-sQTLs and the differential AS events affected by the methylated position.

Methylated position	Differential analysis of methylated positions				AS event	Differential analysis of AS			Me-sQTLs			
	Strand	Intercept	f	*p* Value		Gene	E (ΔPSI)	95% MV|ΔPSI|	Statistic	*p* Value	FDR	Beta
cg04928129	1429051−	−2.1144	230.3921	3.34E-44	33029	LMF1	−0.004168	0.11	28.4852	4.34E-109	9.97E-106	0.7302
cg00583426	1209990−	−2.7769	210.9880	4.05E-41	33029	LMF1	−0.004168	0.11	28.1495	1.93E-107	2.22E-104	0.6155
cg08259514	1131634−	−3.8941	291.6219	1.79E-53	33029	LMF1	−0.004168	0.11	27.7357	2.11E-105	1.61E-102	0.5842
cg04603812	1429265−	−4.5946	291.2726	2.01E-53	33029	LMF1	−0.004168	0.11	27.3596	1.51E-103	8.70E-101	0.7149
cg03323597	1131489−	−2.1121	273.3760	8.81E-51	33029	LMF1	−0.004168	0.11	26.8338	6.07E-101	2.79E-98	0.7754
cg09249980	1213919−	1.1469	137.6579	9.86E-29	33029	LMF1	−0.004168	0.11	25.2485	4.74E-93	1.36E-90	1.0011
cg00611495	1120275−	−1.0380	165.5488	1.38E-33	33029	LMF1	−0.004168	0.11	25.1518	1.44E-92	3.31E-90	0.8511
cg20104307	778658+	−1.9372	165.0035	1.71E-33	33029	LMF1	−0.004168	0.11	25.0200	6.57E-92	1.37E-89	0.7347
cg27040104	1384722−	−0.7004	129.0667	3.38E-27	33029	LMF1	−0.004168	0.11	24.8244	6.25E-91	1.20E-88	0.8274
cg00525011	122031+	−1.6582	190.3615	9.36E-38	33029	LMF1	−0.004168	0.11	24.6293	5.92E-90	1.05E-87	0.6617
cg04913730	1121907−	−1.7877	137.0784	1.25E-28	33029	LMF1	−0.004168	0.11	24.5070	2.42E-89	3.98E-87	0.6183
cg00675160	1208531+	−0.7381	141.6593	1.93E-29	33029	LMF1	−0.004168	0.11	24.4083	7.57E-89	1.16E-86	0.7337
cg08438529	1052939−	−1.2133	173.4449	6.27E-35	33029	LMF1	−0.004168	0.11	24.1224	2.05E-87	2.94E-85	0.5957
cg07549278	1204244−	−2.0317	95.9829	4.18E-21	33029	LMF1	−0.004168	0.11	23.9448	1.59E-86	2.15E-84	0.6040
cg16383109	126451−	−1.4219	232.0717	1.82E-44	33029	LMF1	−0.004168	0.11	22.9002	2.77E-81	3.35E-79	0.6947
cg05245533	795877−	−0.9085	145.6310	3.86E-30	33029	LMF1	−0.004168	0.11	22.8553	4.65E-81	5.34E-79	0.6500
cg16443148	776667−	−0.2492	47.2131	1.61E-11	33029	LMF1	−0.004168	0.11	22.6322	6.12E-80	6.12E-78	0.6401
cg09786479	1020419+	−3.2149	77.1741	1.68E-17	33029	LMF1	−0.004168	0.11	22.6067	8.22E-80	7.87E-78	0.5584
cg07336438	1131466−	−0.9484	173.4777	6.19E-35	33029	LMF1	−0.004168	0.11	22.5742	1.20E-79	1.10E-77	0.7216
cg10163825	776685+	−0.4881	18.5564	1.93E-05	33029	LMF1	−0.004168	0.11	22.5302	1.99E-79	1.76E-77	0.9054
cg27127090	1131327+	0.3781	81.8560	2.08E-18	33029	LMF1	−0.004168	0.11	22.1160	2.38E-77	1.95E-75	0.9443
cg07915516	377344−	−1.5503	116.9595	5.29E-25	33029	LMF1	−0.004168	0.11	21.8766	3.77E-76	2.99E-74	0.7060
cg06587435	865125+	1.6381	82.6033	1.49E-18	33029	LMF1	−0.004168	0.11	21.7725	1.25E-75	9.00E-74	1.1040
cg08641445	1080637+	0.4693	58.8931	6.88E-14	33029	LMF1	−0.004168	0.11	21.6790	3.68E-75	2.57E-73	0.9575
cg05272807	1232363+	0.2547	93.9919	9.94E-21	33029	LMF1	−0.004168	0.11	21.6061	8.55E-75	5.78E-73	0.7974

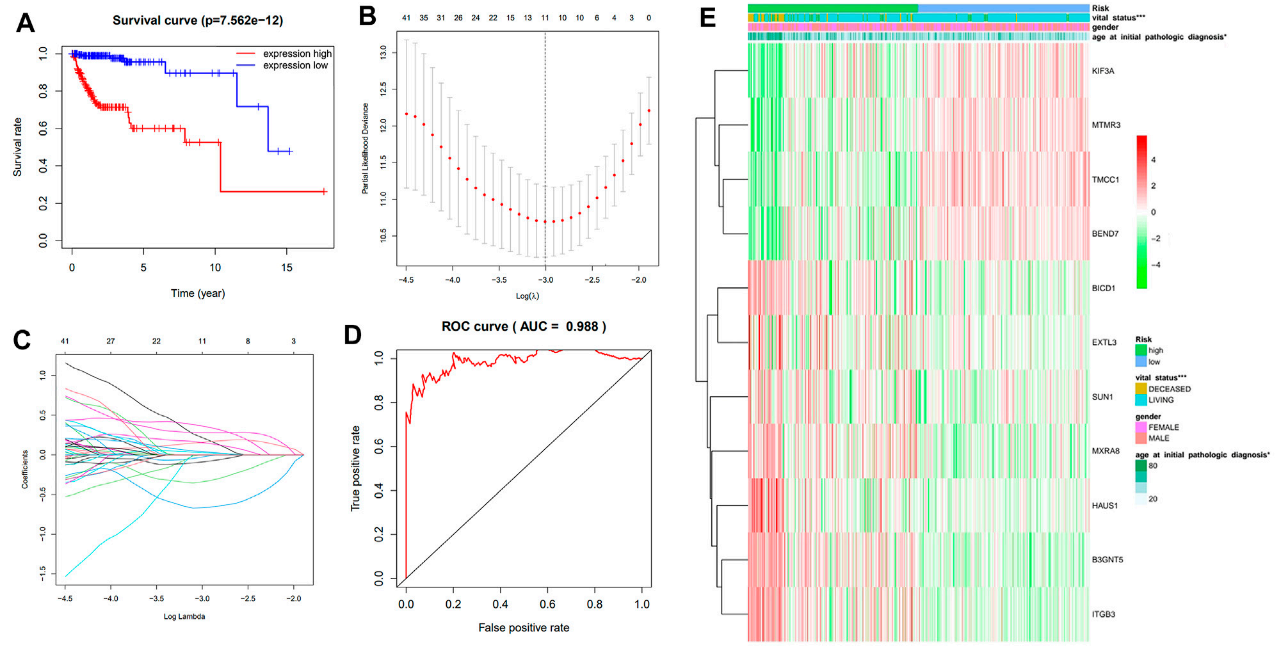


FIGURE 4 | The influence of the glioma-related genes whose AS significantly affected by DNA methylation level on the disease prognosis. **(A)** The Kaplan-Meier overall survival curves of the low (red) and high (blue) expression groups. **(B)** and **(C)** show the results of LASSO regression. There are 11 independent genes with their coefficient when the partial likelihood deviance reaches its minimum value. **(D)** The ROC curve reveals the reliability of the risk score by comparing the true and false positive rate. **(E)** The heatmap shows the association between the risk scores of the prognosis-related me-sQTL genes and the clinical features of glioma patients.

of risk scores, the patients were separated into the low and high-risk groups. We found that the LGG and GBM subjects are mainly distributed in the low and high-risk group, respectively, which reflects the consistency between the risk scores calculated by the prognosis-related me-sQTL genes and the severity of glioma. Moreover, the results of chi-square test showed that the risk scores of the prognosis-related me-sQTL genes are also associated with the age at initial pathologic diagnosis ($p = 4.89 \times 10^{-2}$) and vital status ($p = 1.81 \times 10^{-9}$), but not with the gender of the patients ($p = 4.60 \times 10^{-1}$) (**Figure 4E**). This proves that the 11 key genes we found are meaningful for clinical prognosis of glioma. Among the 11 genes, 7 of them (i.e., B3GNT5, BICD1, KIF3A, HAUS1, MTMR3, ITGB3 and EXTL3) have been confirmed to be associated with the prognosis of glioma (Kim et al., 2011; Sumazin et al., 2011; Huang et al., 2017; Zhou et al., 2018; Jeong et al., 2020; Li et al., 2020; Wang et al., 2020). Our findings imply that the functions of these genes in glioma prognosis may be related to the methylation regulation of their AS events.

CONCLUSION

In this study, we used the TCGA data to explore the role of the me-sQTL process on pathogenesis of glioma and identify the affected genes and further analyze the influence of them on the clinical prognosis of glioma. In total, we identified 130 such genes which have the following three characteristics: 1) they are significantly differentially expressed between the LGG and GBM subjects; 2) their AS events are significantly regulated by DNA methylation level in the cis regions; and 3) the cis me-sQTLs of them are significantly enriched in glioma risk methylation position dataset. Further, the results of clinical data analysis show a significant association between the expression of these genes and the clinical prognosis of glioma, and among them, 11 (i.e., KIF3A, HAUS1, TMCC1, BEND7, B3GNT5, MTMR3, ITGB3, BICD1, EXTL3, SUN1 and MXRA8) are considered the key risk factors for the prognosis and severity of glioma. At the same time, these 130 genes provide new ideas for the study of the interaction between DNA methylation and alternative splicing in gliomas and similar diseases and provide reference for future research on the study of DNA methylation and variable splicing in neurological diseases in the whole genome. In summary, we performed a strategy to explore the influence of DNA methylation level on gene AS in glioma and these findings will help to better understand pathogenesis of glioma.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

ZH and ZY designed the research. ZY, YH, YW, LH, YT, YH, and YC collected the data. ZY and ZH performed the research, analyzed data, and wrote the paper. ZH reviewed and modified the manuscript. All authors discussed the results, and contributed to the final manuscript. All authors read and approved the final manuscript.

FUNDING

This research is financially supported by Start-up fund of Chongqing Medical University (Grant No. R1017) and Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJQN202100402).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.799913/full#supplementary-material

REFERENCES

Aken, B. L., Achuthan, P., Akanni, W., Amode, M. R., Bernsdorff, F., Bhai, J., et al. (2017). Ensembl 2017. *Nucleic Acids Res.* 45, D635–D642. doi:10.1093/nar/gkw1104

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., et al. (2014). Minfi: a Flexible and Comprehensive Bioconductor Package for the Analysis of Infinium DNA Methylation Microarrays. *Bioinformatics* 30, 1363–1369. doi:10.1093/bioinformatics/btu049

Chen, W.-j., Zhang, X., Han, H., Lv, J.-n., Kang, E.-m., Zhang, Y.-l., et al. (2020). The Different Role of YKL-40 in Glioblastoma Is a Function of MGMT Promoter Methylation Status. *Cell Death Dis* 11, 668. doi:10.1038/s41419-020-02909-9

Chen, Y.-C., and Elnitski, L. (2019). Aberrant DNA Methylation Defines Isoform Usage in Cancer, with Functional Implications. *Plos Comput. Biol.* 15, e1007095. doi:10.1371/journal.pcbi.1007095

Consortium, G. T. (2020). The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues. *Science* 369, 1318–1330. doi:10.1126/science.aaz1776

Dong, Z., and Cui, H. (2020). The Emerging Roles of RNA Modifications in Glioblastoma. *Cancers* 12, 736. doi:10.3390/cancers12030736

Drag, M. H., Kogelman, L. J. A., Maribo, H., Meinert, L., Thomsen, P. D., and Kadarmideen, H. N. (2019). Characterization of eQTLs Associated with Androstenone by RNA Sequencing in Porcine Testis. *Physiol. Genomics* 51, 488–499. doi:10.1152/physiolgenomics.00125.2018

Erridge, S. C., Hart, M. G., Kerr, G. R., Smith, C., McNamara, S., Grant, R., et al. (2011). Trends in Classification, Referral and Treatment and the Effect on Outcome of Patients with Glioma: a 20 Year Cohort. *J. Neurooncol.* 104, 789–800. doi:10.1007/s11060-011-0546-0

Etcheverry, A., Aubry, M., de Tayrac, M., Vauleon, E., Boniface, R., Guenot, F., et al. (2010). DNA Methylation in Glioblastoma: Impact on Gene Expression and Clinical Outcome. *BMC Genomics* 11, 701. doi:10.1186/1471-2164-11-701

Feng, J., Dang, Y., Zhang, W., Zhao, X., Zhang, C., Hou, Z., et al. (2019). PTEN Arginine Methylation by PRMT6 Suppresses PI3K-AKT Signaling and Modulates Pre-mRNA Splicing. *Proc. Natl. Acad. Sci. USA* 116, 6868–6877. doi:10.1073/pnas.1811028116

Galarza-Muñoz, G., Briggs, F. B. S., Evsyukova, I., Schott-Lerner, G., Kennedy, E. M., Nyanhete, T., et al. (2017). Human Epistatic Interaction Controls IL7R Splicing and Increases Multiple Sclerosis Risk. *Cell* 169, 72–84. e13. doi:10.1016/j.cell.2017.03.007

Gillies, C. E., Putler, R., Menon, R., Otto, E., Yasutake, K., Nair, V., et al. (2018). An eQTL Landscape of Kidney Tissue in Human Nephrotic Syndrome. *Am. J. Hum. Genet.* 103, 232–244. doi:10.1016/j.ajhg.2018.07.004

Guo, X., Xu, Y., and Zhao, Z. (2015). In-depth Genomic Data Analyses Revealed Complex Transcriptional and Epigenetic Dysregulations of BRAF V600E in Melanoma. *Mol. Cancer* 14, 60. doi:10.1186/s12943-015-0328-y

Gutierrez-Arcelus, M., Ongen, H., Lappalainen, T., Montgomery, S. B., Buil, A., Yurovsky, A., et al. (2015). Tissue-specific Effects of Genetic and Epigenetic Variation on Gene Regulation and Splicing. *Plos Genet.* 11, e1004958. doi:10.1371/journal.pgen.1004958

Ha, K. C. H., Sterne-Weiler, T., Morris, Q., Weatheritt, R. J., and Blencowe, B. J. (2021). Differential Contribution of Transcriptomic Regulatory Layers in the Definition of Neuronal Identity. *Nat. Commun.* 12, 335. doi:10.1038/s41467-020-20483-8

Han, S., and Lee, Y. (2017). *IMAS: Integrative Analysis of Multi-Omics Data for Alternative Splicing*, 1.

Han, Z., Xue, W., Tao, L., Lou, Y., Qiu, Y., and Zhu, F. (2020). Genome-wide Identification and Analysis of the eQTL lncRNAs in Multiple Sclerosis Based on RNA-Seq Data. *Brief Bioinform* 21, 1023–1037. doi:10.1093/bib/bbz036

He, Y., Wang, L., Tang, J., and Han, Z. (2020). Genome-Wide Identification and Analysis of the Methylation of lncRNAs and Prognostic Implications in the Glioma. *Front. Oncol.* 10, 607047. doi:10.3389/fonc.2020.607047

Hekman, R. M., Hume, A. J., Goel, R. K., Abo, K. M., Huang, J., Blum, B. C., et al. (2021). Actionable Cytopathogenic Host Responses of Human Alveolar Type 2 Cells to SARS-CoV-2. *Mol. Cel* 81, 212. doi:10.1016/j.molcel.2020.12.028

Hottinger, A. F., Stupp, R., and Homicsko, K. (2014). Standards of Care and Novel Approaches in the Management of Glioblastoma Multiforme. *Chin. J. Cancer* 33, 32–39. doi:10.5732/cjc.013.10207

Hu, Z., and Qu, S. (2021). EVA1C Is a Potential Prognostic Biomarker and Correlated with Immune Infiltration Levels in WHO Grade II/III Glioma. *Front. Immunol.* 12, 683572. doi:10.3389/fimmu.2021.683572

Huang, S.-P., Chang, Y.-C., Low, Q. H., Wu, A. T. H., Chen, C.-L., Lin, Y.-F., et al. (2017). BICD1 Expression, as a Potential Biomarker for Prognosis and Predicting Response to Therapy in Patients with Glioblastomas. *Oncotarget* 8, 113766–113791. doi:10.18632/oncotarget.22667

Hwang, J.-Y., Aromolaran, K. A., and Zukin, R. S. (2017). The Emerging Field of Epigenetics in Neurodegeneration and Neuroprotection. *Nat. Rev. Neurosci.* 18, 347–361. doi:10.1038/nrn.2017.46

Irimia, M., Weatheritt, R. J., Ellis, J. D., Parikshak, N. N., Gonatopoulos-Pournatzis, T., Babor, M., et al. (2014). A Highly Conserved Program of Neuronal Microexons Is Misregulated in Autistic Brains. *Cell* 159, 1511–1523. doi:10.1016/j.cell.2014.11.035

Jeong, H. Y., Park, S. Y., Kim, H. J., Moon, S., Lee, S., Lee, S. H., et al. (2020). B3GNT5 Is a Novel Marker Correlated with Stem-like Phenotype and Poor Clinical Outcome in Human Gliomas. *CNS Neurosci. Ther.* 26, 1147–1154. doi:10.1111/cns.13439

Kim, J.-H., Zheng, L. T., Lee, W.-H., and Suk, K. (2011). Pro-apoptotic Role of Integrin β3 in Glioma Cells. *J. Neurochem.* 117, 494–503. doi:10.1111/j.1471-4159.2011.07219.x

Kim, M.-S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., et al. (2014). A Draft Map of the Human Proteome. *Nature* 509, 575–581. doi:10.1038/nature13302

Li, F., Yi, Y., Miao, Y., Long, W., Long, T., Chen, S., et al. (2019). N6-Methyladenosine Modulates Nonsense-Mediated mRNA Decay in Human Glioblastoma. *Cancer Res.* 79, 5785–5798. doi:10.1158/0008-5472.CAN-18-2868

Li, S., Wei, Z., Li, G., Zhang, Q., Niu, S., Xu, D., et al. (2020). Silica Perturbs Primary Cilia and Causes Myofibroblast Differentiation during Silicosis by Reduction of the KIF3A-Repressor GLI3 Complex. *Theranostics* 10, 1719–1732. doi:10.7150/thno.37049

Liu, D., Zhao, L., Wang, Z., Zhou, X., Fan, X., Li, Y., et al. (2019). EWASdb: Epigenome-wide Association Study Database. *Nucleic Acids Res.* 47, D989–D993. doi:10.1093/nar/gky942

Merkin, J., Russell, C., Chen, P., and Burge, C. B. (2012). Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. *Science* 338, 1593–1599. doi:10.1126/science.1228186

Mogilevsky, M., Shimshon, O., Kumar, S., Mogilevsky, A., Keshet, E., Yavin, E., et al. (2018). Modulation ofMKNK2alternative Splicing by Splice-Switching Oligonucleotides as a Novel Approach for Glioblastoma Treatment. *Nucleic Acids Res.* 46, 11396–11404. doi:10.1093/nar/gky921

Molinaro, A. M., Taylor, J. W., Wiencke, J. K., and Wrensch, M. R. (2019). Genetic and Molecular Epidemiology of Adult Diffuse Glioma. *Nat. Rev. Neurol.* 15, 405–417. doi:10.1038/s41582-019-0220-2

Pan, T., Wu, F., Li, L., Wu, S., Zhou, F., Zhang, P., et al. (2021). The Role m6A RNA Methylation Is CNS Development and Glioma Pathogenesis. *Mol. Brain* 14, 119. doi:10.1186/s13041-021-00831-5

Pangeni, R. P., Zhang, Z., Alvarez, A. A., Wan, X., Sastry, N., Lu, S., et al. (2018). Genome-wide Methylomic and Transcriptomic Analyses Identify Subtype-specific Epigenetic Signatures Commonly Dysregulated in Glioma Stem Cells and Glioblastoma. *Epigenetics* 13, 432–448. doi:10.1080/15592294.2018.1469892

Pattwell, S. S., Arora, S., Cimino, P. J., Ozawa, T., Szulzewsky, F., Hoellerbauer, P., et al. (2020). A Kinase-Deficient NTRK2 Splice Variant Predominates in Glioma and Amplifies Several Oncogenic Signaling Pathways. *Nat. Commun.* 11, 2977. doi:10.1038/s41467-020-16786-5

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., et al. (2019). g:Profiler: a Web Server for Functional Enrichment Analysis and Conversions of Gene Lists (2019 Update). *Nucleic Acids Res.* 47, W191–W198. doi:10.1093/nar/gkz369

Rong, M.-h., Zhu, Z.-h., Guan, Y., Li, M.-w., Zheng, J.-s., Huang, Y.-q., et al. (2020). Identification of Prognostic Splicing Factors and Exploration of Their Potential Regulatory Mechanisms in Pancreatic Adenocarcinoma. *PeerJ* 8, e8380. doi:10.7717/peerj.8380

Ryan, M., Wong, W. C., Brown, R., Akbani, R., Su, X., Broom, B., et al. (2016). TCGASpliceSeq a Compendium of Alternative mRNA Splicing in Cancer. *Nucleic Acids Res.* 44, D1018–D1022. doi:10.1093/nar/gkv1288

Shabalin, A. A. (2012). Matrix eQTL: Ultra Fast eQTL Analysis via Large Matrix Operations. *Bioinformatics* 28, 1353–1358. doi:10.1093/bioinformatics/bts163

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J. Stat. Soft.* 39, 1–13. doi:10.18637/jss.v039.i05

Sumazin, P., Yang, X., Chiu, H.-S., Chung, W.-J., Iyer, A., Llobet-Navas, D., et al. (2011). An Extensive microRNA-Mediated Network of RNA-RNA Interactions Regulates Established Oncogenic Pathways in Glioblastoma. *Cell* 147, 370–381. doi:10.1016/j.cell.2011.09.041

Vigneswaran, K., Neill, S., and Hadjipanayis, C. G. (2015). Beyond the World Health Organization Grading of Infiltrating Gliomas: Advances in the Molecular Genetics of Glioma Classification. *Ann. Transl Med.* 3, 95. doi:10.3978/j.issn.2305-5839.2015.03.57

Wang, B., Mao, J.-h., Wang, B.-y., Wang, L.-x., Wen, H.-y., Xu, L.-j., et al. (2020). Exosomal miR-1910-3p Promotes Proliferation, Metastasis, and Autophagy of Breast Cancer Cells by Targeting MTMR3 and Activating the NF-Kb Signaling Pathway. *Cancer Lett.* 489, 87–99. doi:10.1016/j.canlet.2020.05.038

Wang, M., Zhou, Z., Zheng, J., Xiao, W., Zhu, J., Zhang, C., et al. (2021). Identification and Validation of a Prognostic Immune-Related Alternative Splicing Events Signature for Glioma. *Front. Oncol.* 11, 650153. doi:10.3389/fonc.2021.650153

Warzel, D. B., Andonaydis, C., McCurry, B., Chilukuri, R., Ishmukhamedov, S., and Covitz, P. (2003). Common Data Element (CDE) Management and Deployment in Clinical Trials. *AMIA Annu. Symp. Proc.* 2003, 1048.

Wei, Y., Zhang, Z., Peng, R., Sun, Y., Zhang, L., and Liu, H. (2021). Systematic Identification of Survival-Associated Alternative Splicing Events in Kidney Renal Clear Cell Carcinoma. *Comput. Math. Methods Med.* 2021, 1–10. doi:10.1155/2021/5576933

Weller, M., Stupp, R., Reifenberger, G., Brandes, A. A., van den Bent, M. J., Wick, W., et al. (2010). MGMT Promoter Methylation in Malignant Gliomas: Ready for Personalized Medicine. *Nat. Rev. Neurol.* 6, 39–51. doi:10.1038/nrneurol.2009.197

Wiestler, B., Capper, D., Sill, M., Jones, D. T. W., Hovestadt, V., Sturm, D., et al. (2014). Integrated DNA Methylation and Copy-Number Profiling Identify Three Clinically and Biologically Relevant Groups of Anaplastic Glioma. *Acta Neuropathol.* 128, 561–571. doi:10.1007/s00401-014-1315-x

Xiao, Y., Cui, G., Ren, X., Hao, J., Zhang, Y., Yang, X., et al. (2020). A Novel Four-Gene Signature Associated with Immune Checkpoint for Predicting Prognosis in Lower-Grade Glioma. *Front. Oncol.* 10, 605737. doi:10.3389/fonc.2020.605737

Xie, Z. c., Wu, H. y., Dang, Y. w., and Chen, G. (2019). Role of Alternative Splicing Signatures in the Prognosis of Glioblastoma. *Cancer Med.* 8, 7623–7636. doi:10.1002/cam4.2666

Yang, C., Wu, Q., Huang, K., Wang, X., Yu, T., Liao, X., et al. (2019). Genome-Wide Profiling Reveals the Landscape of Prognostic Alternative Splicing Signatures in Pancreatic Ductal Adenocarcinoma. *Front. Oncol.* 9, 511. doi:10.3389/fonc.2019.00511

Zeng, Y., Zhang, P., Wang, X., Wang, K., Zhou, M., Long, H., et al. (2020). Identification of Prognostic Signatures of Alternative Splicing in Glioma. *J. Mol. Neurosci.* 70, 1484–1492. doi:10.1007/s12031-020-01581-0

Zhou, C., Wang, Y., Liu, X., Liang, Y., Fan, Z., Jiang, T., et al. (2018). Molecular Profiles for Insular Low-Grade Gliomas with Putamen Involvement. *J. Neurooncol.* 138, 659–666. doi:10.1007/s11060-018-2837-1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

OPINION
published: 10 March 2022
doi: 10.3389/fgene.2022.828827

Lack of Causal Relationships Between Chronic Hepatitis C Virus Infection and Alzheimer's Disease

Lin Huang[1], *Yongheng Wang*[1,2], *Yaqin Tang*[1], *Yijie He*[1] *and Zhijie Han*[1]*

[1]Department of Bioinformatics, School of Basic Medicine, Chongqing Medical University, Chongqing, China, [2]International Research Laboratory of Reproduction and Development, Chongqing Medical University, Chongqing, China

Keywords: chronic hepatitis C (HCV), Alzheimer disease, Mendelian randomization (MR) analysis, GWAS—genome-wide association study, cause

OPEN ACCESS

Edited by:
Liangcai Zhang,
Janssen Research and Development, United States

Reviewed by:
Haoxiang Cheng,
Icahn School of Medicine at Mount Sinai, United States
Jianbo Fu,
Helmholtz Association of German Research Centers (HZ), Germany

****Correspondence:***
Zhijie Han
zhijiehan@cqmu.edu.cn

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *04 December 2021*
Accepted: *20 January 2022*
Published: *10 March 2022*

Citation:
Huang L, Wang Y, Tang Y, He Y and Han Z (2022) Lack of Causal Relationships Between Chronic Hepatitis C Virus Infection and Alzheimer's Disease. Front. Genet. 13:828827. doi: 10.3389/fgene.2022.828827

INTRODUCTION

Inflammation produces hepatic encephalopathy in patients with chronic liver disease by modulating brain functions (O'Beirne et al., 2006). Several studies have reported that patients with chronic hepatitis C virus (HCV) infection tend to exhibit cognitive impairment and may increase the risk for dementia (Chiu et al., 2014; Adinolfi et al., 2015; Choi et al., 2021). Meanwhile, the HCV genome has been detected in the brain tissues of some patients with dementia, which suggests that it may be able to infect the central nervous system (CNS) directly (Forton et al., 2001; Khonsari et al., 2015). A recent study reported that treatment of HCV infection with direct-acting antivirals (e.g., glecaprevir/pibrentasvir, elbasvir/grazoprevir, and ledipasvir/sofosbuvir) significantly reduces mortality risk in patients with Alzheimer's disease (AD) and related dementia (Tran et al., 2021). Furthermore, apolipoprotein E (ApoE) plays a key role in the mechanism of AD by driving amyloid-β (Aβ) peptide accumulation in the brain (Yamazaki et al., 2019). Previous studies demonstrated that the ApoE level affects HCV infection and action in the CNS by regulating the blood–brain barrier permeability and is significantly associated with the neuropsychiatric symptoms in HCV-infected individuals (Gochee et al., 2004; Sheridan et al., 2014; Wozniak et al., 2016). Although evidence has shown that HCV infection is associated with the dysfunctions of the CNS, it is not clear whether any HCV infection influences AD pathogenesis. Observational studies are difficult to interpret because these results may have been influenced by reverse causation and confounding factors. Mendelian randomization (MR) has the potential to evaluate causal relationships between exposure and outcome in the presence of such limitations (Sekula et al., 2016; Davies et al., 2018). In this study, we investigated the causal impact of HCV infection on the risk of late-onset AD by implementing Causal Analysis Using Summary Effect estimates (CAUSE), a novel MR method that can avoid more false positives caused by correlated horizontal pleiotropy (Morrison et al., 2020).

ANALYSIS OF ASSOCIATION BETWEEN HCV INFECTION AND RISK OF LATE-ONSET AD

The summary data for exposure (HCV infection) was downloaded from the National Bioscience Database Center (NBDC) Human Database which includes the complete results of genome-wide association studies (GWAS) based on 5,794 HCV susceptible cases and 206,659 controls (NBDC Research ID: hum0014.v17.CHC.v1) (Ishigaki et al., 2020). The GWAS results of late-onset AD were obtained from the International Genomics of Alzheimer's Project (IGAP) ($n = 17,008$ late-onset AD cases and 37,154 controls) (Lambert et al., 2013). In addition, Manhattan plot of HCV and AD GWAS results are in **Supplementary Figure S1**. According to the manual of CAUSE, there should be as many single-nucleotide polymorphisms (SNPs) as the instrumental variable (IV) to estimate CAUSE posteriors to ensure the

TABLE 1 | The results of causality of HCV infection and AD by comparing causal and sharing model.

Model 1	Model 2	Δ ELPD	s.e. Δ ELPD	Z-Score	*p*-value
Null	Sharing	0.22	0.13	1.70	0.95
Null	Causal	0.96	1.00	0.94	0.83
Sharing	Causal	0.74	0.92	0.80	0.79

Model 1 and Model 2 imply the models being compared. When estimated difference in ELPD (Δ ELPD) = $ELPD_C$ − $ELPD_S$ is negative, model 2 is a better fit.
Key: s.e. Δ ELPD, estimated standard error of Δ ELPD; Z-Score, Δ ELPD/s.e. Δ ELPD.

accuracy of the MR results (https://github.com/jean997/cause). Therefore, we used the more liberal threshold of the GWAS significance and the independence of SNPs (i.e., $p < 0.001$ and $r^2 < 0.01$, respectively) in this study. The "'ld_clump" function of the R package "ieugwasr" was used to calculate pairwise prune linkage disequilibrium (LD) measures between these SNPs based on the 1000 Genomes project phase I and prune the non-independent ones (https://mrcieu.github.io/ieugwasr/). To avoid the pleiotropy effects, CAUSE included as many information from all variants as possible, even weakly associated variants. It computed the test statistic to distinguish variants associated with confounders. At last, we use two-sample MR for further validation (Morrison et al., 2020).

After merging the GWAS results between the NBDC HCV and IGAP AD studies, and further removing the variants with ambiguous and mismatched alleles, we selected 4,289,211 SNPs which match both datasets for the following analyses. The nuisance parameters are estimated by finding the mixing parameters and the maximum a posteriori estimate. According to the significance threshold of the GWAS $p < 0.001$ and the LD analysis $r^2 < 0.01$, there are a total of 606 SNPs as IVs for the CAUSE fitting. As **Table 1** shows, when compared with the null model, the estimated difference in expected log pointwise posterior density (delta EPLD) in both the sharing and causal models is positive, and no significant differences are presented (*p*-value is 0.95 and 0.83, respectively). Further, the fitted delta EPLD of the causal model is still not significantly better than the sharing model (z-score = 0.80 and *p*-value = 0.79). This revealed a similar posterior distributions and proportion of correlated pleiotropic SNPs in the two models. A total of 42 SNPs for any HCV infection were identified (**Supplementary Table S1**). Two-sample MR analysis indicated that genetically predicted HCV infection was not associated with AD (odds ratio (OR) = 0.99, 95% confidence interval (CI) = −0.07to 0.04, $p = 0.69$) (**Supplementary Table S2**). The resulting estimates of the effect of HCV on AD are shown in the scatter plot (**Supplementary Figure S2**). Generalized funnel plot indicated the absence of directional pleiotropy (**Supplementary Figure S3**). Leave-one-out analysis revealed a high stability of our results (**Supplementary Figure S4**). Thus, the present MR study affords no support for causality between HCV and AD, which suggests that HCV infection has no influence on the late-onset AD.

DISCUSSION

The traditional MR methods may tend to lead to false positive results because the assumption about horizontal pleiotropy of instrument is often violated. Therefore, in this study, using the largest available GWAS results on HCV infection and AD, we investigated a potential causal role for HCV infection in late-onset AD using an improved MR analysis. However, the results failed to reveal any causal association between them, which appears to be in conflict with some previous reports about the influence of HCV infection on the human CNS (Weissenborn et al., 2004; Yarlott et al., 2017). Previous studies demonstrate that ApoE and its cell-surface receptor is a key prerequisite for HCV production and infectivity by enriching the virus particles and giving rise to the lipoviral particles hybrid with lipoproteins. Given that ApoE also plays an important role in Aβ peptide accumulation, a potential explanation for these findings could be that the ApoE level may affect HCV infection and also mediate the genetic risk of late-onset AD in patients with HCV and AD (Jiang and Luo, 2009; Hishiki et al., 2010; Yang et al., 2016) and thus leads to a false association between HCV infection and AD. In conclusion, our MR analyses found no evidence for a causal role of HCV for late-onset AD pathogenesis. These findings could further improve the conclusions of previous studies, and further research are needed to elucidate the underlying mechanisms.

CONCLUSION

In this study, we used both CAUSE and two-sample MR study to assess the causal effects of HCV infection on AD. The CAUSE results revealed that HCV infection did not appear to have a causal effect on the risk of AD. Similar trends were observed through two-sample MR. The evidence suggests that previously reported observational associations could have resulted from confounding. Future studies are warranted to clarify the underlying mechanism.

AUTHOR CONTRIBUTIONS

ZH designed the research. LH, YW, YT, YH, and ZH collected the data. LH and ZH performed the research and analyzed the data. LH wrote the paper. ZH reviewed and modified the article. All authors discussed the results and contributed to the final article. All authors read and approved the final article.

FUNDING

This research is financially supported by the Start-up fund of Chongqing Medical University (R1017) and the Science and Technology Research Program of Chongqing Municipal Education Commission (Grant No. KJQN202100402).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.828827/full#supplementary-material

REFERENCES

Adinolfi, L. E., Nevola, R., Lus, G., Restivo, L., Guerrera, B., Romano, C., et al. (2015). Chronic Hepatitis C Virus Infection and Neurological and Psychiatric Disorders: an Overview. *Wjg.* 21 (8), 2269–2280. doi:10.3748/wjg.v21.i8.2269

Chiu, W.-C., Tsan, Y.-T., Tsai, S.-L., Chang, C.-J., Wang, J.-D., and Chen, P.-C. (2014). Health Data Analysis in Taiwan Research, GHepatitis C Viral Infection and the Risk of Dementia. *Eur. J. Neurol.* 21 (8), 1068–e59. doi:10.1111/ene.12317

Choi, H. G., Soh, J. S., Lim, J. S., Sim, S. Y., and Lee, S. W. (2021). Association between Dementia and Hepatitis B and C Virus Infection. *Medicine (Baltimore).* 100 (29), e26476. doi:10.1097/md.0000000000026476

Davies, N. M., Holmes, M. V., and Davey Smith, G. (2018). Reading Mendelian Randomisation Studies: a Guide, Glossary, and Checklist for Clinicians. *BMJ.* 362, k601. doi:10.1136/bmj.k601

Forton, D. M., Allsop, J. M., Main, J., Foster, G. R., Thomas, H. C., and Taylor-Robinson, S. D. (2001). Evidence for a Cerebral Effect of the Hepatitis C Virus. *The Lancet.* 358 (9275), 38–39. doi:10.1016/s0140-6736(00)05270-3

Gochee, P. A., Powell, E. E., Purdie, D. M., Pandeya, N., Kelemen, L., Shorthouse, C., et al. (2004). Association Between Apolipoprotein E ε4 and Neuropsychiatric Symptoms During Interferon α Treatment for Chronic Hepatitis C. *Psychosomatics.* 45 (1), 49–57. doi:10.1176/appi.psy.45.1.49

Hishiki, T., Shimizu, Y., Tobita, R., Sugiyama, K., Ogawa, K., Funami, K., et al. (2010). Infectivity of Hepatitis C Virus Is Influenced by Association with Apolipoprotein E Isoforms. *J. Virol.* 84 (22), 12048–12057. doi:10.1128/jvi.01063-10

Ishigaki, K., Akiyama, M., Kanai, M., Takahashi, A., Kawakami, E., Sugishita, H., et al. (2020). Large-scale Genome-wide Association Study in a Japanese Population Identifies Novel Susceptibility Loci across Different Diseases. *Nat. Genet.* 52 (7), 669–679. doi:10.1038/s41588-020-0640-3

Jiang, J., and Luo, G. (2009). Apolipoprotein E but Not B Is Required for the Formation of Infectious Hepatitis C Virus Particles. *J. Virol.* 83 (24), 12680–12691. doi:10.1128/jvi.01476-09

Khonsari, R. H., Maylin, S., Nicol, P., Martinot-Peignoux, M., Créange, A., Duyckaerts, C., et al. (2015). Sicca Syndrome and Dementia in a Patient with Hepatitis C Infection: a Case Report with Unusual Bifocal Extrahepatic Manifestations. *J. Maxillofac. Oral Surg.* 14 (Suppl. 1), 388–392. doi:10.1007/s12663-014-0632-x

Lambert, J.-C., Ibrahim-Verbaas, C. A., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., et al. (2013). Meta-Analysis of 74,046 Individuals Identifies 11 New Susceptibility Loci for Alzheimer's Disease. *Nat. Genet.* 45 (12), 1452–1458. doi:10.1038/ng.2802

Morrison, J., Knoblauch, N., Marcus, J. H., Stephens, M., and He, X. (2020). Mendelian Randomization Accounting for Correlated and Uncorrelated Pleiotropic Effects Using Genome-wide Summary Statistics. *Nat. Genet.* 52 (7), 740–747. doi:10.1038/s41588-020-0631-4

O'Beirne, J. P., Chouhan, M., and Hughes, R. D. (2006). The Role of Infection and Inflammation in the Pathogenesis of Hepatic Encephalopathy and Cerebral Edema in Acute Liver Failure. *Nat. Rev. Gastroenterol. Hepatol.* 3 (3), 118–119. doi:10.1038/ncpgasthep0417

Sekula, P., Del Greco M, F., Pattaro, C., and Köttgen, A. (2016). Mendelian Randomization as an Approach to Assess Causality Using Observational Data. *Jasn.* 27 (11), 3253–3265. doi:10.1681/asn.2016010098

Sheridan, D. A., Bridge, S. H., Crossey, M. M. E., Felmlee, D. J., Thomas, H. C., Neely, R. D. G., et al. (2014). Depressive Symptoms in Chronic Hepatitis C Are Associated with Plasma Apolipoprotein E Deficiency. *Metab. Brain Dis.* 29 (3), 625–634. doi:10.1007/s11011-014-9520-9

Tran, L., Jung, J., Carlin, C., Lee, S., Zhao, C., and Feldman, R. (2021). Use of Direct-Acting Antiviral Agents and Survival Among Medicare Beneficiaries with Dementia and Chronic Hepatitis C. *Jad.* 79 (1), 71–83. doi:10.3233/jad-200949

Weissenborn, K., Krause, J., Bokemeyer, M., Hecker, H., Schüler, A., Ennen, J. C., et al. (2004). Hepatitis C Virus Infection Affects the Brain-Evidence from Psychometric Studies and Magnetic Resonance Spectroscopy. *J. Hepatol.* 41 (5), 845–851. doi:10.1016/j.jhep.2004.07.022

Wozniak, M. A., Lugo Iparraguirre, L. M., Dirks, M., Deb-Chatterji, M., Pflugrad, H., Goldbecker, A., et al. (2016). Apolipoprotein E-E4 Deficiency and Cognitive Function in Hepatitis C Virus-Infected Patients. *J. Viral Hepat.* 23 (1), 39–46. doi:10.1111/jvh.12443

Yamazaki, Y., Zhao, N., Caulfield, T. R., Liu, C.-C., and Bu, G. (2019). Apolipoprotein E and Alzheimer Disease: Pathobiology and Targeting Strategies. *Nat. Rev. Neurol.* 15 (9), 501–518. doi:10.1038/s41582-019-0228-7

Yang, Z., Wang, X., Chi, X., Zhao, F., Guo, J., Ma, P., et al. (2016). Neglected but Important Role of Apolipoprotein E Exchange in Hepatitis C Virus Infection. *J. Virol.* 90 (21), 9632–9643. doi:10.1128/jvi.01353-16

Yarlott, L., Heald, E., and Forton, D. (2017). Hepatitis C Virus Infection, and Neurological and Psychiatric Disorders - A Review. *J. Adv. Res.* 8 (2), 139–148. doi:10.1016/j.jare.2016.09.005

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 31 March 2022
doi: 10.3389/fgene.2022.782476

Total Brain Volumetric Measures and Schizophrenia Risk: A Two-Sample Mendelian Randomization Study

Dan Zhu[1,2†], Chunyang Wang[3†], Lining Guo[2†], Daojun Si[4], Mengge Liu[2], Mengjing Cai[2], Lin Ma[2], Dianxun Fu[2], Jilian Fu[2], Junping Wang[2*] and Feng Liu[2*]*

[1]Department of Radiology, Tianjin Medical University General Hospital Airport Hospital, Tianjin, China, [2]Department of Radiology and Tianjin Key Laboratory of Functional Imaging, Tianjin Medical University General Hospital, Tianjin, China, [3]Department of Scientific Research, Tianjin Medical University General Hospital, Tianjin, China, [4]National Supercomputer Center in Tianjin, Tianjin, China

OPEN ACCESS

Edited by:
Guiyou Liu,
Tianjin Institute of Industrial Biotechnology (CAS), China

Reviewed by:
Fengmei Lu,
University of Macau, China
Alexandre Bureau,
Laval University, Canada
Minghao Dong,
Xidian University, China

****Correspondence:***
Feng Liu
fengliu@tmu.edu.cn
Junping Wang
wangjunping_tj@163.com
Jilian Fu
fujilian2015@163.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *24 September 2021*
Accepted: *02 March 2022*
Published: *31 March 2022*

Citation:
Zhu D, Wang C, Guo L, Si D, Liu M, Cai M, Ma L, Fu D, Fu J, Wang J and Liu F (2022) Total Brain Volumetric Measures and Schizophrenia Risk: A Two-Sample Mendelian Randomization Study. Front. Genet. 13:782476. doi: 10.3389/fgene.2022.782476

Schizophrenia (SCZ) is an idiopathic psychiatric disorder with a heritable component and a substantial public health impact. Although abnormalities in total brain volumetric measures (TBVMs) have been found in patients with SCZ, it is still unknown whether these abnormalities have a causal effect on the risk of SCZ. Here, we performed a Mendelian randomization (MR) study to investigate the possible causal associations between each TBVM and SCZ risk. Specifically, genome-wide association study (GWAS) summary statistics of total gray matter volume, total white matter volume, total cerebrospinal fluid volume, and total brain volume were obtained from the United Kingdom Biobank database (33,224 individuals), and SCZ GWAS summary statistics were provided by the Psychiatric Genomics Consortium (150,064 individuals). The main MR analysis was conducted using the inverse variance weighted method, and other MR methods, including MR-Egger, weighted median, simple mode, and weighted mode methods, were performed to assess the robustness of our findings. For pleiotropy analysis, we employed three approaches: MR-Egger intercept, MR-PRESSO, and heterogeneity tests. No TBVM was causally associated with SCZ risk according to the MR results, and no significant pleiotropy or heterogeneity was found for instrumental variables. Taken together, this study suggested that alterations in TBVMs were not causally associated with the risk of SCZ.

Keywords: schizophrenia, total brain volumetric measures, genetic, causality, Mendelian randomization

INTRODUCTION

Schizophrenia (SCZ) is one of the most serious mental disorders; it has a high disability rate worldwide and has brought heavy economic burdens and life pressure to families and society (Mueser and McGurk, 2004). SCZ has been shown to have a high rate of heritability (60–80%), much of which is attributable to common risk alleles, suggesting that the genome-wide association study (GWAS) can enhance our understanding of the etiology of SCZ (Kahn et al., 2015). The GWAS has revealed that single-nucleotide polymorphisms (SNPs) at novel loci confer risk for SCZ, and these results have been obtained by enlarging sample sizes and incorporating more ethnicities (Ripke et al., 2013; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014; Lam et al., 2019).

In addition to the genetic basis, substantial efforts have been made in the past decade to elucidate the neural basis of SCZ by using neuroimaging techniques (Kahn et al., 2015). Neuroimaging measures can be considered as endophenotypes, which are quantitative indicators of brain structure or function that index genetic liability for neuropsychiatric disorders (Meyer-Lindenberg and Weinberger, 2006). Compared to neuropsychiatric disorders, endophenotypes are hypothesized to have less polygenicity, have a greater effect size of susceptible SNPs, and require smaller sample sizes to discover the SNPs (Gottesman and Gould, 2003; Meyer-Lindenberg and Weinberger, 2006). A number of studies have reported alterations in total brain volumetric measures (TBVMs), such as total gray matter volume (TGMV), total white matter volume (TWMV), total cerebrospinal fluid volume (TCSFV), and total brain volume (TBV), in patients with SCZ. For example, Haijma et al. (2013) conducted a meta-analysis on TBVMs in more than 18,000 patients and controls, demonstrating a significant reduction in intracranial volume (ICV, sum of TGMV, TWMV, and TCSFV) and TBV (sum of TGMV and TWMV) and an increase in TCSFV in SCZ patients. In addition, progressive decreases in TBV and ventricular expansions (increased in TCSFV) were found in longitudinal studies of SCZ (Kempton et al., 2010; Olabi et al., 2011). However, all these findings were based on observational studies, which may be limited by the possibility of confounding factors and reserve causation; thus, it is still unknown whether TBVM alterations have a causal effect on the risk of SCZ.

Mendelian randomization (MR) is an epidemiological approach that could overcome the limitations in observation studies by using genetic variants associated with exposure as instrumental variables to uncover the causal relationship between an exposure and an outcome (Lawlor et al., 2008). In addition, MR can control the confounding factors and reverse causation that are usually encountered in observation studies. To date, MR has been successfully applied to assess causal relationships in pioneer studies of neuropsychiatric diseases (Hartwig et al., 2017a; Liu et al., 2018; Vaucher et al., 2018; He et al., 2020; Wang et al., 2020; Zhang et al., 2020). For instance, Hartwig et al. found a protective effect of C-reactive protein and a risk-increasing effect of soluble interleukin-6 receptor on SCZ risk (Hartwig et al., 2017a). Vaucher et al. reported that the use of cannabis was causally associated with an increased risk of SCZ (Vaucher et al., 2018). Therefore, in this study, by leveraging data from the largest GWAS summary statistics on both TBVMs and SCZ, we performed a two-sample MR study to estimate the causal effect of TBVMs, including TGMV, TWMV, TCSFV, and TBV, on the risk of SCZ.

MATERIALS AND METHODS

Study Design

MR is an approach that uses genetic variants as instrumental variables to investigate the causal relationship between exposures and outcomes, which should satisfy three principal assumptions: 1) the instrumental variables should be significantly associated with exposure; 2) the instrumental variables should not be associated with any confounders; and 3) the instrumental variables should affect the risk of the outcome only by the exposure. The second and third assumptions are also considered independent of pleiotropy. In this study, MR is based on the publicly available GWAS summary datasets of TBVMs (Smith et al., 2021) and SCZ (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014), and all subjects provided informed consent in the original studies. Specifically, the genetic variants that were significantly associated with TBVMs were used as instrumental variables to examine the causal influence of TBVMs on SCZ risk (**Figure 1**).

Total Brain Volumetric Measure Genome-Wide Association Study Dataset

The GWAS summary data of TBVMs, including TGMV, TWMV, TCSFV, and TBV, were downloaded from an open resource named the Oxford Brain Imaging Genetics (BIG40) web server (https://open.win.ox.ac.uk/ukbiobank/big40/), which included GWAS summary statistics with 33,224 individuals in the United Kingdom Biobank (Smith et al., 2021). The genome-wide significance threshold was set at $p < 5 \times 10^{-8}$ in the discovery cohort (N = 22,138) and $p < 0.05$ in the replication cohort (N = 11,086). Only SNPs that met the significance level in both cohorts were used as instrumental variables in MR analyses, and these SNPs were independent and had no linkage disequilibrium, as described in the original studies (Elliott et al., 2018; Smith et al., 2021). Detailed information about the instrumental variables of TGMV, TWMV, TCSFV, and TBV is shown in **Supplementary Tables S1-S4**.

Genome-Wide Association Study of Schizophrenia

GWAS summary data regarding SCZ were downloaded from a meta-analysis provided by the Schizophrenia Working Group of Psychiatric Genomics Consortium (https://www.med.unc.edu/pgc/pgc-workgroups/schizophrenia/), including 36,989 cases and 113,075 controls of predominantly European ancestry without population stratification. In total, 128 significant associations in 108 genetic loci were identified (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014).

Pleiotropy Analysis

Comprehensive pleiotropy analyses were performed to assure that instrumental variables met the MR assumptions. First, an MR-Egger intercept test was performed to evaluate the potential pleiotropic associations of the instrumental variables with known and unknown confounders (Bowden et al., 2015; Bowden et al., 2016; Burgess and Thompson, 2017). Second, an MR pleiotropy residual sum and outlier (MR-PRESSO) analysis was carried out to detect horizontal pleiotropy (i.e., MR-PRESSO global test) (Verbanck et al., 2018). Heterogeneity across instrumental variables is also an indicator of pleiotropy. Thus, Cochran's Q test and I^2 statistic were calculated to estimate the heterogeneity (Sun et al., 2021). Specifically, Cochran's Q test is a conventional test for heterogeneity and approximately follows a chi-square

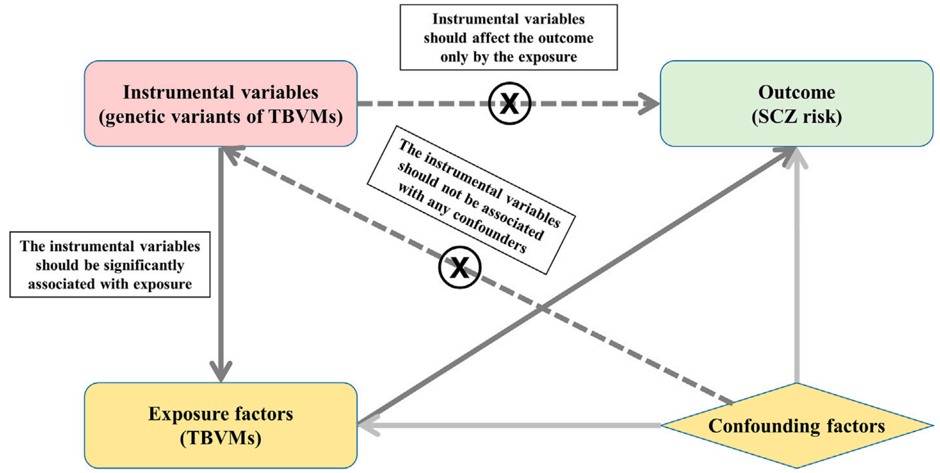

Instrumental variables should affect the outcome only by the exposure
Instrumental variables (genetic variants of TBVMs)
Outcome (SCZ risk)
The instrumental variables should not be associated with any confounders
The instrumental variables should be significantly associated with exposure
Exposure factors (TBVMs)
Confounding factors

FIGURE 1 | Study design based on MR principal assumptions. In this study, MR is based on the publicly available GWAS summary datasets in TBVMs and SCZ. Specifically, the genetic variants that are significantly associated with TBVMs were used as the instrumental variables to examine the causal influence of TBVMs on SCZ risk. Abbreviations: SCZ, Schizophrenia; TBVMs, total brain volumetric measures.

distribution with n-1 degrees of freedom (here, n is the number of instrumental variables). The I^2 index is another measure to quantify heterogeneity, which divides the difference between the Q statistic and its degrees of freedom by the Q statistic itself and then multiplies by 100. The value of the I^2 index ranges from 0 to 100%, with 0%–25%, 25%–50%, 50%–75%, and 75%–100% representing low, moderate, large, and extreme heterogeneity, respectively (Liu et al., 2013; He et al., 2020). The significance threshold of all the MR-Egger intercept, MR-PRESSO, and Cochran's Q tests was set at $p < 0.05$.

Aligning Effect Alleles With Exposure and Outcome

The effect alleles of the instrumental variables were adjusted to be associated with increased TBVMs (i.e., the effect estimates of SNPs were larger than zero). Subsequently, the effect alleles of these genetic variants were aligned to be consistent with the effect alleles in the SCZ GWAS dataset. If the instrumental SNPs were not available in the outcome dataset, a proxy SNP that was in high linkage disequilibrium ($r^2 > 0.8$) with the requested SNP was searched instead with the online tool SNiPA (https://snipa.helmholtz-muenchen.de/snipa3/index.php) (Arnold et al., 2015).

Two-Sample Mendelian Randomization Analysis

The inverse variance weighted (IVW) method was employed to estimate the causal effects of TGMV, TWMV, TCSFV, and TBV on SCZ risk. Specifically, for each TBVM, the effect estimates of each instrumental variable on TBVMs and SCZ were extracted, and Wald estimates and their standard errors were then calculated (Burgess et al., 2017b). The Wald estimates of all the instrumental variables were combined with a weighted mean using inverse variance weights. The significance threshold of the associations between exposures and outcomes was set at $p < 0.05$.

Power Analysis

For each TBVM, the proportion of variance explained by each instrumental variable (R^2) was calculated using the following formula:

$$R^2 = \frac{2 \times MAF \times (1 - MAF) \times \beta^2}{2 \times MAF \times (1 - MAF) \times \beta^2 + 2 \times MAF \times (1 - MAF) \times N \times se(\beta)^2}$$

where MAF represents the minor allele frequency for a given SNP, β represents the effect size associated with the TBVM for a given SNP, $se(\beta)$ represents the standard error of the effect size associated with the exposure for a given SNP, and N represents the sample size of the exposure GWAS data.

Then, the strength of instrument variables can be measured by F-statistics, which were calculated based on the following equation:

$$F = \frac{R^2 \times (N - 2)}{1 - R^2}$$

where R^2 is the proportion of the variance explained by each SNP, and N represents the sample size of the exposure GWAS data. To minimize weak instrument bias, SNPs with F-statistics > 10 were retained for subsequent analyses (Lawlor et al., 2008).

Sensitivity Analysis

A series of sensitivity analyses were conducted to validate the robustness of the results. First, four different MR methods including MR-Egger, weighted median, simple mode, and weighted mode methods were performed to estimate the causal effect of TBVMs on SCZ risk. Specifically, the MR-Egger method allows all variants to have pleiotropic effects and can provide a consistent estimate of the causal effect under a weaker instrument strength independent of direct effects (InSIDE) assumption (Burgess and Thompson, 2017); the weighted median method can provide valid causal estimates even if up to 50% of instruments are not valid (Bowden et al., 2016); and the model-based methods (i.e., simple mode and weighted mode) use the causal effect estimates for individual SNPs to form clusters, and the causal effect is estimated in the largest cluster of SNPs (Hartwig et al., 2017b). Second, a leave-one-out sensitivity analysis was carried out to identify SNPs that could potentially bias the causal relationship. To this aim, by sequentially removing each SNP, we estimated the relationship between the remaining SNPs and the risk of SCZ using the IVW method. Finally, reverse causation bias may occur when the outcome variable is at an earlier time point (i.e., the risk of SCZ causally influences the changes of each TBVM). Therefore, we also tested the possibility of reverse causation by treating the risk of SCZ as an exposure and each TBVM as an outcome. Specifically, the instrumental variables were the significant genetic variants associated with SCZ risk, and the same procedures as the main analyses were used to perform reverse MR causality detection.

All statistical analyses were conducted using *R* version 4.0.4 (R Foundation for Statistical Computing, Vienna, Austria) using the packages of "TwoSampleMR" (Hemani et al., 2018b) and "MR-PRESSO" (Verbanck et al., 2018).

RESULTS

Association of Total Brain Volumetric Measure Variants With Schizophrenia

Only two genetic variants without linkage disequilibrium were found to be associated with TGMV, and their summary statistics were extracted from SCZ GWAS data for MR analyses (**Supplementary Table S1**). Of the five genetic variants associated with TWMV, rs742396 was a palindromic SNP. Thus, we deleted it in the subsequent MR analyses (**Supplementary Table S2**). Seven genetic variants were associated with TCSFV. All seven instrumental SNPs were located on different chromosomes and were not in linkage disequilibrium with each other. However, rs4843550 is a palindromic SNP and was removed from the subsequent MR analyses. The summary statistics for these TCSFV variants are shown in **Supplementary Table S3**. Of the five genetic variants associated with TBV, the summary statistics for the four variants could be extracted from the SCZ GWAS data. The SNP rs2732714 was not available in the SCZ GWAS data, therefore, we used the information of its proxy SNP rs113138968, which was in high linkage disequilibrium ($r^2 > 0.8$), to perform the following analyses. All five instrumental SNPs were not in linkage disequilibrium with each other, and none of them were palindromic SNPs. Detailed information about these five instrumental SNPs is shown in **Supplementary Table S4**.

Pleiotropy Analysis

Both the MR-Egger intercept test and MR-PRESSO test showed no significant pleiotropy for the genetic variants of TBVMs (all *p*s > 0.05). Furthermore, Cochran's *Q* test and I^2 statistic revealed no significant heterogeneity for these SNPs (**Supplementary Table S5**).

Two-Sample Mendelian randomization Analysis

We performed a two-sample MR analysis by using genetic variants from TGMV, TWMV, TCSFV, and TBV as instrumental variables. As shown in **Table 1**, we did not find any causal influence on the risk of SCZ with the IVW method ($p > 0.05$).

Power Analysis

The explained variances (R^2) and *F*-statistics of each instrumental variable are shown in **Supplementary Tables S1-S4**, and the *F*-statistics of each instrumental variable were larger than 10, indicating no weak instrumental bias among these variables.

Sensitivity Analysis

All other MR approaches, including the MR-Egger, weighted median, simple mode, and weighted mode methods, did not identify any significant causal effects of TGMV, TWMV, and TBV on the risk of SCZ (**Table 1**). Although TCSFV was found to be causally associated with the risk of SCZ when using the MR-Egger method (*BETA* = 0.646, *SE* = 0.220, *p value* = 0.042), this result was not validated by other methods. In leave-one-out sensitivity analyses, no genetic variants could significantly affect the MR estimates (**Figure 2**). For the reverse MR causality analysis, 111 leading SNPs associated with SCZ risk were extracted from the GWAS summary data of TBVMs. Among them, ten palindromic SNPs were removed, and the remaining 101 SNPs were retained for subsequent analyses (**Supplementary Table S6**). All the MR methods indicated that there was no causal influence of any TBVM on SCZ risk (**Supplementary Table S7**).

DISCUSSION

SCZ is a chronic, complex mental disorder characterized by an array of symptoms, including delusions, hallucinations, disorganized speech, and impaired cognitive ability, that typically emerges in late adolescence and early adulthood (Mueser and McGurk, 2004; Sheffield and Barch, 2016; Marder and Cannon, 2019; McCutcheon et al., 2020). Several lines of evidence have suggested that structural brain abnormalities play an important role in the pathology of SCZ (Okada et al., 2016; Zhao et al., 2018; Kuo and Pogue-Geile, 2019). Using neuroimaging methods, some researchers found TBVM abnormalities in patients with SCZ relative to age-matched healthy controls (Staal et al., 1998; Haijma et al., 2013), and progressive reductions in TBVMs might be associated with disease progression (Kempton et al., 2010). However, evidence has pointed toward the possibility that antipsychotic drugs might have an effect on TBVM alterations (Olabi et al., 2011; Guo et al., 2015; Emsley et al., 2017). In addition, as a risk factor for SCZ,

TABLE 1 | Results of the causal effect of TBVMs on SCZ risk.

MR methods	TGMV			TWMV			TCSFV			TBV		
	BETA	SE	*p* value	BETA	SE	*p* value	BETA	SE	*p* value	BETA	SE	*p* value
IVW	0.085	0.163	0.601	0.062	0.132	0.636	-0.089	0.089	0.315	0.114	0.090	0.202
MR-Egger	—	—	—	0.004	0.831	0.997	-0.646	0.220	0.042	1.204	0.914	0.279
Weighted median	—	—	—	0.025	0.122	0.841	0.012	0.088	0.089	0.081	0.094	0.384
Simple mode	—	—	—	0.003	0.178	0.986	0.064	0.127	0.635	0.070	0.141	0.649
Weighted mode	—	—	—	0.001	0.159	0.997	0.060	0.118	0.631	0.064	0.145	0.681

Abbreviations: BETA, regression coefficient; IVW, inverse variance weighted; MR, Mendelian randomization; SE, standard error; TBV, total brain volume; TCSFV, total cerebrospinal fluid volume; TGMV, total gray matter volume; TWMV, total white matter volume.
Notably, only the IVW method worked when there were two instrumental variables in TGMV-SCZ MR analysis.

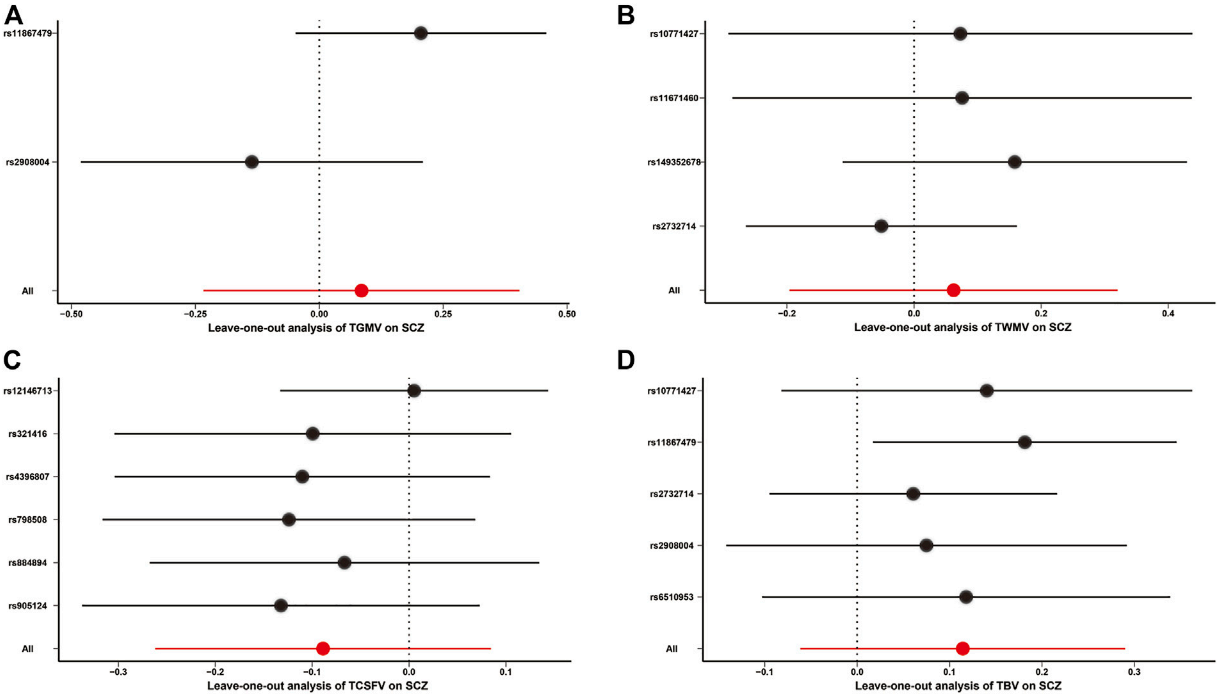


FIGURE 2 | Leave-one-out analysis for MR causality analysis between TBVMs and SCZ risk. **(A)**. Leave-one-out analysis for MR causality analysis between TGMV and SCZ risk. **(B)**. Leave-one-out analysis for MR causality analysis between TWMV and SCZ risk. **(C)**. Leave-one-out analysis for MR causality analysis between TCSFV and SCZ risk. **(D)**. Leave-one-out analysis for MR causality analysis between TBV and SCZ risk. The red points and red lines represent the BETA and 95% confidence interval in MR analyses, while the black points and black lines represent the BETA and 95% confidence interval after removing each SNP sequentially. Of note, only the IVW method was used in the leave-one-out sensitivity analysis. Abbreviations: SCZ, Schizophrenia; TBV, total brain volume; TCSFV, total cerebrospinal fluid volume; TGMV, total gray matter volume; TWMV, total white matter volume.

experience with cannabis use could also lead to brain structural alterations (Kumra et al., 2012; Rapp et al., 2012; Navarri et al., 2022). Hence, the causality between the changes in TBVMs and the risk of SCZ remains largely unclear.

In this study, we aimed to explore whether there is a causal effect of changes in TBVMs on the risk of SCZ by using MR, one of the powerful genetic-epidemiological approaches. Here, we used four reliable TBVMs (TGMV, TWMV, TCSFV, and TBV) derived from structural neuroimaging data. Specifically, genetic variants of TGMV, TWMV, TCSFV, and TBV without any pleiotropy and heterogeneity were selected as the instrumental variables, and five MR methods were used to ensure the reliability of the results. Different from the observational studies, no significant result was found using MR between any TBVMs and SCZ risk. The possible explanations for the difference are as follows: 1) the substantial brain structural heterogeneity exists across the individuals with SCZ (Alnæs et al., 2019). The changes in TBVMs might not be a sensitive risk factor for SCZ, since alterations (increase or decrease) in the volume of some specific brain regions have been reported in patients with SCZ (Kuo and Pogue-Geile, 2019); 2) some observational studies showed that the decrease in TBVMs in SCZ might be the result of antipsychotics, aging, or other unknown confounders (Kumra et al., 2012; Emsley et al., 2017); and 3) SCZ is a cognitive and behavioral dysfunction with complex symptoms (Sheffield and Barch, 2016), the onset of which might be linked to functional abnormalities rather than structural abnormalities of the brain. Hence, more attention should be devoted to the changes in specific brain region volumes by removing the effects of antipsychotics and aging and the functional neural mechanisms of SCZ.

Our study design has many advantages. First, the exposure and outcome datasets were from a large-scale GWAS of TBVMs (N = 33,224) and SCZ (36,989 cases and 113,075 controls). The large sample sizes of GWAS typically led to higher levels of statistical power (van der Sluis et al., 2013). Second, we utilized independent SNPs as the instrumental variables in each MR analysis, which could effectively avoid the influence caused by linkage disequilibrium. Third, a series of pleiotropy and sensitivity analyses based on different principles and assumptions were carried out to detect pleiotropy and heterogeneity to ensure that the instrumental variables we used here were reliable (Burgess et al., 2017a; Hemani et al., 2018a). Finally, to increase the robustness of the MR results, different methods were applied to investigate the causal relationship between the exposures and the outcomes. Assessing the causal relationship by using a variety of methods is more reliable because the different MR methods we used here were based on the different assumptions (Burgess and Thompson, 2017).

Some limitations needed to be addressed in this study. First, the subjects from the outcome GWAS dataset were of transancestral descent (both European and East Asian); however, the subjects from the TBVM GWAS dataset were of pure European descent. Population stratification might have a potential confounding effect on the causal estimate. Second, although a series of statistical methods were used to identify pleiotropy, it is impossible to fully remove all pleiotropy in MR studies. Third, the instrumental variables of TBVMs were obtained from United Kingdom Biobank GWAS summary data. The participants in the United Kingdom Biobank were aged from 45 to 81 years (Smith et al., 2021), which is not the typical age of onset for SCZ (Howard et al., 2000). The genetic variants determining TBVMs in childhood and/or adolescence may differ from those determining TBVMs in adulthood used in this study. Therefore, it would be better to use instrumental variables from large-scale TBVMs GWAS data in childhood and/or adolescence that are not publicly available to date. Fourth, the generalized summary-based MR (GSMR) method is also a popular MR approach to assess the causal association between exposure and outcome (Zhu et al., 2018). The rule of thumb advises that the application of GSMR requires ten or more independent genome-wide significant SNPs, but there were fewer than ten instrumental variables used in each two-sample MR analysis in our study, especially those of TGMV. Thus, we could not use the GSMR method to test the causal associations of TBVMs with SCZ risk. Finally, ICV is also an important TBVM, and alterations in ICV were found in SCZ patients (Haijma et al., 2013). We did not investigate the causal relationship between ICV and SCZ risk in this study because there are no GWAS summary data of ICV in the United Kingdom Biobank database.

CONCLUSION

In conclusion, although the previous neuroimaging studies showed the changes in TBVMs in patients with SCZ, our MR results demonstrated that there was no causal relationship between alterations in TBVMs and the risk of SCZ at the genetic level. Further studies with independent data are warranted to confirm these findings.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: the brain volume GWAS summary data were downloaded from the Oxford Brain Imaging Genetics (BIG40) web server (https://open.win.ox.ac.uk/ukbiobank/big40/). GWAS summary data of SCZ were provided by the Schizophrenia Working Group of Psychiatric Genomics Consortium (https://www.med.unc.edu/pgc/pgc-workgroups/SCZ/).

ETHICS STATEMENT

The MR is based on the publicly available GWAS summary datasets of TBV and SCZ, and all patients have provided informed consent in these corresponding original studies.

AUTHOR CONTRIBUTIONS

FL, JW, and JF conceived and designed the experiments. LM, ML, and DF prepared and managed the data. DZ, DS, CW, LG, and MC conducted the experiment and analyzed the data. DZ, MC, and FL wrote the manuscript. All authors contributed to and have approved the final manuscript.

FUNDING

This work was supported by the Natural Science Foundation of China (grant numbers 82072001, 82171905, and 81871431), Natural Science Foundation of Tianjin Municipal Science and Technology Commission (grant number 18JCYBJC26300), Medical Foundation of Jieping Wu (grant number 320.6750.19089-56), and Medical Foundation of Yicheng Zhao (grant number ZYY2019007).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.782476/full#supplementary-material

REFERENCES

Alnæs, D., Kaufmann, T., van der Meer, D., Córdova-Palomera, A., Rokicki, J., Moberget, T., et al. (2019). Brain Heterogeneity in Schizophrenia and its Association with Polygenic Risk. *JAMA Psychiatry* 76, 739–748. doi:10.1001/jamapsychiatry.2019.0257

Arnold, M., Raffler, J., Pfeufer, A., Suhre, K., and Kastenmüller, G. (2015). SNiPA: an Interactive, Genetic Variant-Centered Annotation Browser. *Bioinformatics* 31, 1334–1336. doi:10.1093/bioinformatics/btu779

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian Randomization with Invalid Instruments: Effect Estimation and Bias Detection through Egger Regression. *Int. J. Epidemiol.* 44, 512–525. doi:10.1093/ije/dyv080

Bowden, J., Del Greco M, F., Minelli, C., Davey Smith, G., Sheehan, N. A., and Thompson, J. R. (2016). Assessing the Suitability of Summary Data for Two-Sample Mendelian Randomization Analyses Using MR-Egger Regression: the Role of the I2 Statistic. *Int. J. Epidemiol.* 45, 1961–1974. doi:10.1093/ije/dyw220

Burgess, S., and Thompson, S. G. (2017). Interpreting Findings from Mendelian Randomization Using the MR-Egger Method. *Eur. J. Epidemiol.* 32, 377–389. doi:10.1007/s10654-017-0255-x

Burgess, S., Bowden, J., Fall, T., Ingelsson, E., and Thompson, S. G. (2017a). Sensitivity Analyses for Robust Causal Inference from Mendelian Randomization Analyses with Multiple Genetic Variants. *Epidemiology* 28, 30–42. doi:10.1097/ede.0000000000000559

Burgess, S., Small, D. S., and Thompson, S. G. (2017b). A Review of Instrumental Variable Estimators for Mendelian Randomization. *Stat. Methods Med. Res.* 26, 2333–2355. doi:10.1177/0962280215597579

Elliott, L. T., Sharp, K., Alfaro-Almagro, F., Shi, S., Miller, K. L., Douaud, G., et al. (2018). Genome-wide Association Studies of Brain Imaging Phenotypes in UK Biobank. *Nature* 562, 210–216. doi:10.1038/s41586-018-0571-7

Emsley, R., Asmal, L., du Plessis, S., Chiliza, B., Phahladira, L., and Kilian, S. (2017). Brain Volume Changes over the First Year of Treatment in Schizophrenia: Relationships to Antipsychotic Treatment. *Psychol. Med.* 47, 2187–2196. doi:10.1017/s0033291717000642

GottesmanII, and Gould, T. D. (2003). The Endophenotype Concept in Psychiatry: Etymology and Strategic Intentions. *Ajp* 160, 636–645. doi:10.1176/appi.ajp.160.4.636

Guo, J. Y., Huhtaniska, S., Miettunen, J., Jääskeläinen, E., Kiviniemi, V., Nikkinen, J., et al. (2015). Longitudinal Regional Brain Volume Loss in Schizophrenia: Relationship to Antipsychotic Medication and Change in Social Function. *Schizophrenia Res.* 168, 297–304. doi:10.1016/j.schres.2015.06.016

Haijma, S. V., Van Haren, N., Cahn, W., Koolschijn, P. C. M. P., Hulshoff Pol, H. E., and Kahn, R. S. (2013). Brain Volumes in Schizophrenia: a Meta-Analysis in over 18 000 Subjects. *Schizophr Bull.* 39, 1129–1138. doi:10.1093/schbul/sbs118

Hartwig, F. P., Borges, M. C., Horta, B. L., Bowden, J., and Davey Smith, G. (2017a). Inflammatory Biomarkers and Risk of Schizophrenia: A 2-Sample Mendelian Randomization Study. *JAMA Psychiatry* 74, 1226–1233. doi:10.1001/jamapsychiatry.2017.3191

Hartwig, F. P., Davey Smith, G., and Bowden, J. (2017b). Robust Inference in Summary Data Mendelian Randomization via the Zero Modal Pleiotropy assumption. *Int. J. Epidemiol.* 46, 1985–1998. doi:10.1093/ije/dyx102

He, Y., Zhang, H., Wang, T., Han, Z., Ni, Q.-b., Wang, K., et al. (2020). Impact of Serum Calcium Levels on Alzheimer's Disease: A Mendelian Randomization Study. *Jad* 76, 713–724. doi:10.3233/jad-191249

Hemani, G., Bowden, J., and Davey Smith, G. (2018a). Evaluating the Potential Role of Pleiotropy in Mendelian Randomization Studies. *Hum. Mol. Genet.* 27, R195–r208. doi:10.1093/hmg/ddy163

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., et al. (2018b). The MR-Base Platform Supports Systematic Causal Inference across the Human Phenome. *Elife* 7, e34408. doi:10.7554/eLife.34408

Howard, R., Rabins, P. V., Seeman, M. V., and Jeste, D. V. (2000). Late-Onset Schizophrenia and Very-Late-Onset Schizophrenia-like Psychosis: An International Consensus. *Am. J. Psychiatry* 157, 172–178. doi:10.1176/appi.ajp.157.2.172

Kahn, R. S., Sommer, I. E., Murray, R. M., Meyer-Lindenberg, A., Weinberger, D. R., Cannon, T. D., et al. (2015). Schizophrenia. *Nat. Rev. Dis. Primers* 1, 15067. doi:10.1038/nrdp.2015.67

Kempton, M. J., Stahl, D., Williams, S. C. R., and DeLisi, L. E. (2010). Progressive Lateral Ventricular Enlargement in Schizophrenia: a Meta-Analysis of Longitudinal MRI Studies. *Schizophrenia Res.* 120, 54–62. doi:10.1016/j.schres.2010.03.036

Kumra, S., Robinson, P., Tambyraja, R., Jensen, D., Schimunek, C., Houri, A., et al. (2012). Parietal Lobe Volume Deficits in Adolescents with Schizophrenia and Adolescents with Cannabis Use Disorders. *J. Am. Acad. Child Adolesc. Psychiatry* 51, 171–180. doi:10.1016/j.jaac.2011.11.001

Kuo, S. S., and Pogue-Geile, M. F. (2019). Variation in Fourteen Brain Structure Volumes in Schizophrenia: A Comprehensive Meta-Analysis of 246 Studies. *Neurosci. Biobehavioral Rev.* 98, 85–94. doi:10.1016/j.neubiorev.2018.12.030

Lam, M., Chen, C.-Y., Chen, C.-Y., Li, Z., Martin, A. R., Bryois, J., et al. (2019). Comparative Genetic Architectures of Schizophrenia in East Asian and European Populations. *Nat. Genet.* 51, 1670–1678. doi:10.1038/s41588-019-0512-x

Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N., and Davey Smith, G. (2008). Mendelian Randomization: Using Genes as Instruments for Making Causal Inferences in Epidemiology. *Stat. Med.* 27, 1133–1163. doi:10.1002/sim.3034

Liu, G., Zhang, S., Cai, Z., Ma, G., Zhang, L., Jiang, Y., et al. (2013). PICALM Gene Rs3851179 Polymorphism Contributes to Alzheimer's Disease in an Asian Population. *Neuromol. Med.* 15, 384–388. doi:10.1007/s12017-013-8225-2

Liu, G., Zhao, Y., Jin, S., Hu, Y., Wang, T., Tian, R., et al. (2018). Circulating Vitamin E Levels and Alzheimer's Disease: a Mendelian Randomization Study. *Neurobiol. Aging* 72, 189.e1–189.e9. doi:10.1016/j.neurobiolaging.2018.08.008

Marder, S. R., and Cannon, T. D. (2019). Schizophrenia. *N. Engl. J. Med.* 381, 1753–1761. doi:10.1056/nejmra1808803

McCutcheon, R. A., Reis Marques, T., and Howes, O. D. (2020). Schizophrenia-An Overview. *JAMA Psychiatry* 77, 201–210. doi:10.1001/jamapsychiatry.2019.3360

Meyer-Lindenberg, A., and Weinberger, D. R. (2006). Intermediate Phenotypes and Genetic Mechanisms of Psychiatric Disorders. *Nat. Rev. Neurosci.* 7, 818–827. doi:10.1038/nrn1993

Mueser, K. T., and McGurk, S. R. (2004). Schizophrenia. *Lancet* 363, 2063–2072. doi:10.1016/s0140-6736(04)16458-1

Navarri, X., Afzali, M. H., Lavoie, J., Sinha, R., Stein, D. J., Momenan, R., et al. (2022). How Do Substance Use Disorders Compare to Other Psychiatric Conditions on Structural Brain Abnormalities? A Cross-Disorder Meta-Analytic Comparison Using the ENIGMA Consortium Findings. *Hum. Brain Mapp.* 43, 399–413. doi:10.1002/hbm.25114

Okada, N., Fukunaga, M., Yamashita, F., Koshiyama, D., Yamamori, H., Ohi, K., et al. (2016). Abnormal Asymmetries in Subcortical Brain Volume in Schizophrenia. *Mol. Psychiatry* 21, 1460–1466. doi:10.1038/mp.2015.209

Olabi, B., Ellison-Wright, I., McIntosh, A. M., Wood, S. J., Bullmore, E., and Lawrie, S. M. (2011). Are There Progressive Brain Changes in Schizophrenia? A Meta-Analysis of Structural Magnetic Resonance Imaging Studies. *Biol. Psychiatry* 70, 88–96. doi:10.1016/j.biopsych.2011.01.032

Rapp, C., Bugra, H., Riecher-Rössler, A., Tamagni, C., and Borgwardt, S. (2012). Effects of Cannabis Use on Human Brain Structure in Psychosis: a Systematic Review Combining *In Vivo* Structural Neuroimaging and post Mortem Studies. *Curr. Pharm. Des.* 18, 5070–5080. doi:10.2174/138161212802884861

Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kähler, A. K., Akterin, S., et al. (2013). Genome-wide Association Analysis Identifies 13 New Risk Loci for Schizophrenia. *Nat. Genet.* 45, 1150–1159. doi:10.1038/ng.2742

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological Insights from 108 Schizophrenia-Associated Genetic Loci. *Nature* 511, 421–427. doi:10.1038/nature13595

Sheffield, J. M., and Barch, D. M. (2016). Cognition and Resting-State Functional Connectivity in Schizophrenia. *Neurosci. Biobehav. Rev.* 61, 108–120. doi:10.1016/j.neubiorev.2015.12.007

Smith, S. M., Douaud, G., Chen, W., Hanayik, T., Alfaro-Almagro, F., Sharp, K., et al. (2021). An Expanded Set of Genome-wide Association Studies of Brain Imaging Phenotypes in UK Biobank. *Nat. Neurosci.* 24, 737–745. doi:10.1038/s41593-021-00826-4

Staal, W. G., Hulshoff Pol, H. E., Schnack, H., van der Schot, A. C., and Kahn, R. S. (1998). Partial Volume Decrease of the Thalamus in Relatives of Patients with Schizophrenia. *Am. J. Psychiatry* 155, 1784–1786. doi:10.1176/ajp.155.12.1784

Sun, J. Y., Zhang, H., Zhang, Y., Wang, L., Sun, B. L., Gao, F., et al. (2021). Impact of Serum Calcium Levels on Total Body Bone mineral Density: A Mendelian Randomization Study in Five Age Strata. *Clin. Nutr.* 40, 2726–2733. doi:10.1016/j.clnu.2021.03.012

van der Sluis, S., Posthuma, D., Nivard, M. G., Verhage, M., and Dolan, C. V. (2013). Power in GWAS: Lifting the Curse of the Clinical Cut-Off. *Mol. Psychiatry* 18, 2–3. doi:10.1038/mp.2012.65

Vaucher, J., Keating, B. J., Lasserre, A. M., Gan, W., Lyall, D. M., Ward, J., et al. (2018). Cannabis Use and Risk of Schizophrenia: a Mendelian Randomization Study. *Mol. Psychiatry* 23, 1287–1292. doi:10.1038/mp.2016.252

Verbanck, M., Chen, C. Y., Neale, B., and Do, R. (2018). Detection of Widespread Horizontal Pleiotropy in Causal Relationships Inferred from Mendelian Randomization between Complex Traits and Diseases. *Nat. Genet.* 50, 693–698. doi:10.1038/s41588-018-0099-7

Wang, L., Qiao, Y., Zhang, H., Zhang, Y., Hua, J., Jin, S., et al. (2020). Circulating Vitamin D Levels and Alzheimer's Disease: A Mendelian Randomization Study in the IGAP and UK Biobank. *J. Alzheimers Dis.* 73, 609–618. doi:10.3233/jad-190713

Zhang, H., Wang, T., Han, Z., Wang, L., Zhang, Y., Wang, L., et al. (2020). Impact of Vitamin D Binding Protein Levels on Alzheimer's Disease: A Mendelian Randomization Study. *J. Alzheimers Dis.* 74, 991–998. doi:10.3233/jad-191051

Zhao, C., Zhu, J., Liu, X., Pu, C., Lai, Y., Chen, L., et al. (2018). Structural and Functional Brain Abnormalities in Schizophrenia: A Cross-Sectional Study at Different Stages of the Disease. *Prog. Neuropsychopharmacol. Biol. Psychiatry* 83, 27–32. doi:10.1016/j.pnpbp.2017.12.017

Zhu, Z., Zheng, Z., Zhang, F., Wu, Y., Trzaskowski, M., Maier, R., et al. (2018). Causal Associations between Risk Factors and Common Diseases Inferred from GWAS Summary Data. *Nat. Commun.* 9, 224. doi:10.1038/s41467-017-02317-2

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 13 April 2022
doi: 10.3389/fgene.2022.782691

A Bidirectional Mendelian Randomization Study of Selenium Levels and Ischemic Stroke

Hui Fang[†], *Weishi Liu*[†], *Luyang Zhang, Lulu Pei, Yuan Gao, Lu Zhao, Rui Zhang, Jing Yang, Bo Song and Yuming Xu**

Department of Neurology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China

Background: Previous observational studies have shown that circulating selenium levels are inversely associated with ischemic stroke (IS). Our aims were to evaluate the causal links between selenium levels and IS, and its subtypes by Mendelian randomization (MR) analysis.

Methods: We used the two-sample Mendelian randomization (MR) method to determine whether the circulating selenium levels are causally associated with the risk of stroke. We extracted the genetic variants (SNPs) associated with blood and toenail selenium levels from a large genome-wide association study (GWAS) meta-analysis. Inverse variance-weighted (IVW) method was used as the determinant of the causal effects of exposures on outcomes.

Results: A total of 4 SNPs (rs921943, rs6859667, rs6586282, and rs1789953) significantly associated with selenium levels were obtained. The results indicated no causal effects of selenium levels on ischemic stroke by MR analysis (OR = 0.968, 95% CI 0.914–1.026, p = 0.269). Meanwhile, there was no evidence of a causal link between circulating selenium levels and subtypes of IS.

Conclusion: The MR study indicated no evidence to support the causal links between genetically predicted selenium levels and IS. Our results also did not support the use of selenium supplementation for IS prevention at the genetic level.

Keywords: selenium, stroke, trace element, cause, Mendelian randomization (MR)

Edited by:
Liangcai Zhang,
Janssen Research and Development,
United States

Reviewed by:
Kei Hang Katie Chan,
City University of Hong Kong, Hong Kong SAR, China
Cheryl D. Cropp,
Samford University, United States

****Correspondence:***
Yuming Xu
xuyuming@zzu.edu.cn

[†]*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *24 September 2021*
Accepted: *07 March 2022*
Published: *13 April 2022*

Citation:
Fang H, Liu W, Zhang L, Pei L, Gao Y, Zhao L, Zhang R, Yang J, Song B and Xu Y (2022) A Bidirectional Mendelian Randomization Study of Selenium Levels and Ischemic Stroke. Front. Genet. 13:782691. doi: 10.3389/fgene.2022.782691

INTRODUCTION

Ischemic stroke (IS) is one of the leading causes of death worldwide and a major cause of serious long-term disability (Campbell et al., 2019). Although IS mortality has been declining globally over the past 2 decades, the number of IS incidents, IS survivors, IS-related deaths, and overall disability-adjusted life years (DALY) lost remains significant and increases year by year (Krishnamurthi et al., 2013). Therefore, early identification of the subjects with a high risk of developing or relapsing IS is of great importance. In addition, the benefit of effective medication for IS (i.e., alteplase) is time-dependent, which limits the wide application of alteplase practice (Phipps and Cronin, 2020). The major challenge of developing new anti-stroke drugs is the presence of the blood–brain barrier and blood circulation gaps, as well as the complexity of signal transduction processes and inflammatory response (Amani et al., 2017;

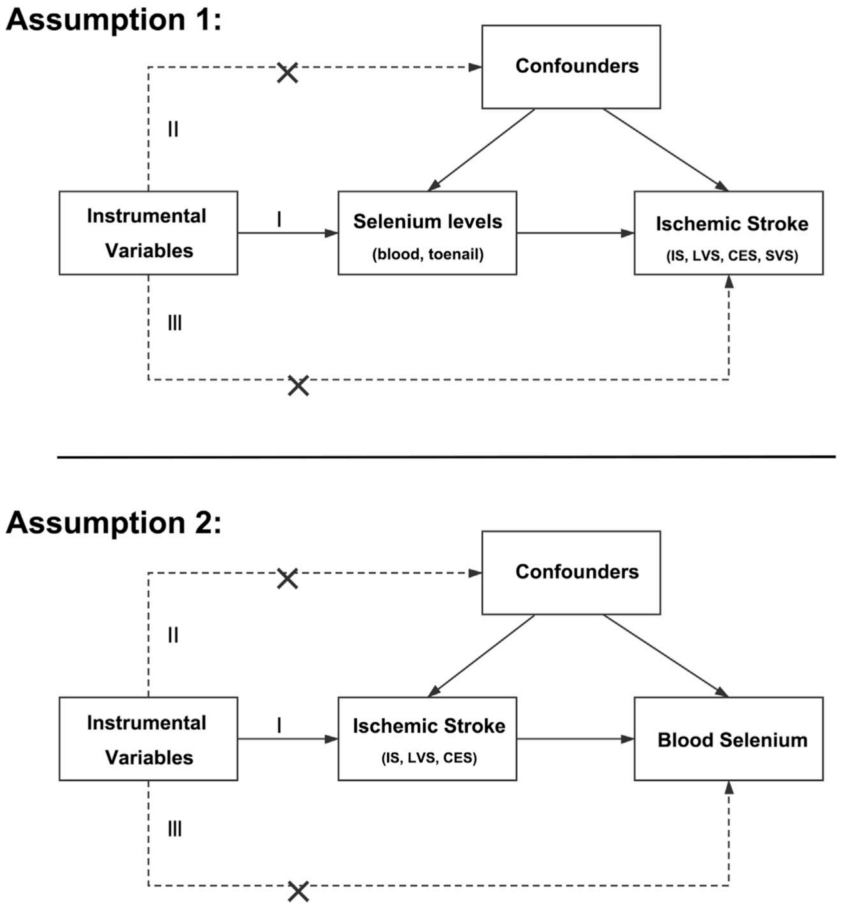
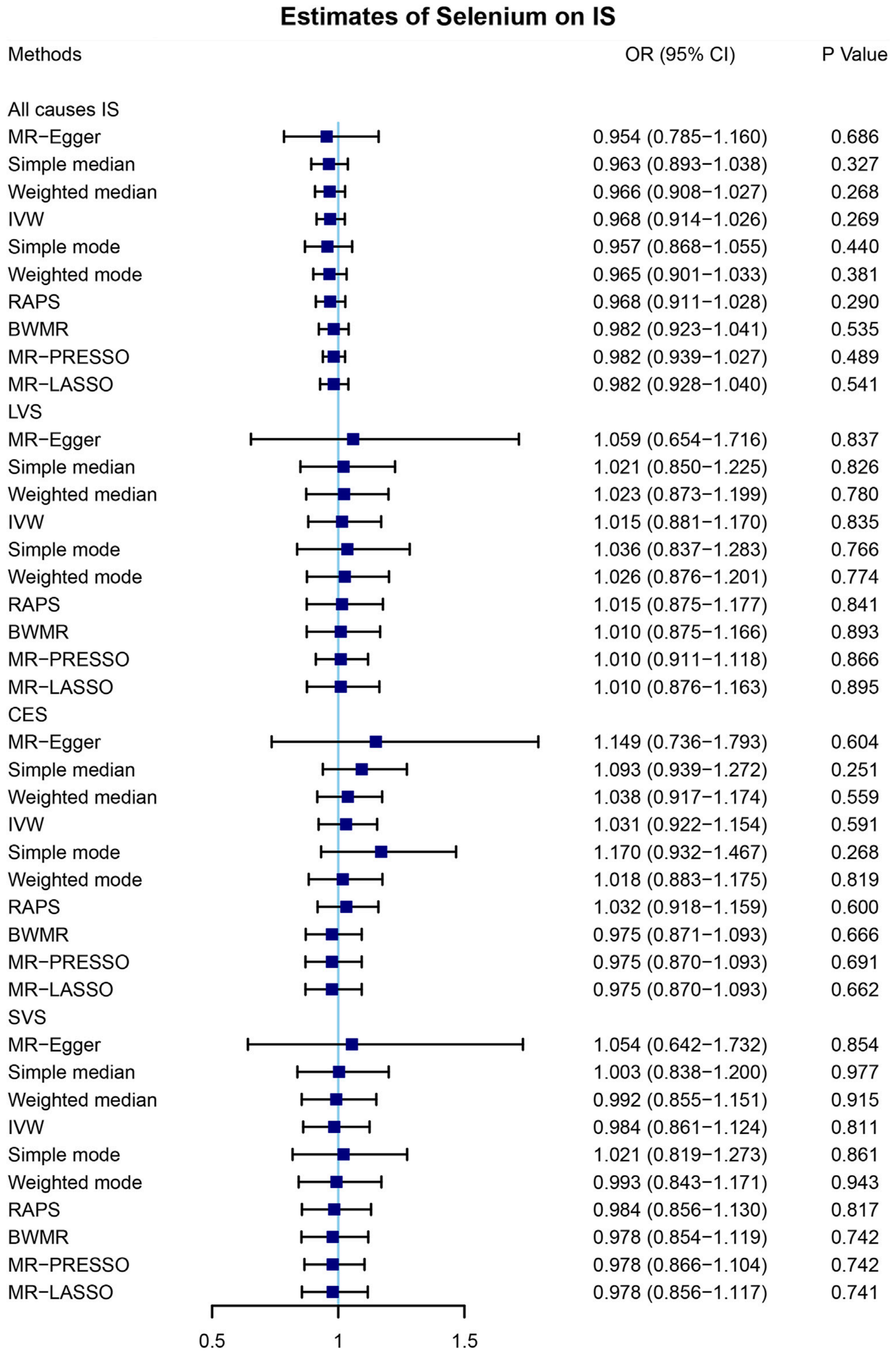


FIGURE 1 | Main assumptions of the Mendelian randomization study of selenium levels and ischemic stroke. IS, ischemic stroke; LVS, large-vessel atherosclerosis stroke; CES, cardio-embolic stroke; SVS, small-vessel occlusion stroke.

Saxton and Sabatini, 2017). Moreover, fast metabolization clearance from blood circulation and poor transport across the blood–brain barrier hinder the efficacy of most central venous system medications (Amani et al., 2017; Amani et al., 2019). All in all, further investigation of risk factors of IS and targeted therapy strategies is warranted.

The major modifiable risk factors of IS include hypertension, diabetes mellitus, hyperlipidemia, and smoking (Go et al., 2014; Feigin et al., 2016). In addition, some trace elements, particularly essential trace elements, have been reported to be associated with IS (Zecca et al., 2004; Scheiber et al., 2014). Selenium is one of the essential trace elements involved in human physiological processes, metabolism, antioxidant defense, immune regulation, and so on (Burk et al., 2014). The main functions of selenoproteins, the main functional form of selenium, in the neural cells are modulation of neurogenesis, regulation of Ca^{2+} channels, and maintenance of the redox balance (Cardoso et al., 2015). Reported *in vitro* studies show that selenium protects mitochondrial functional performance, stimulates mitochondrial biogenesis, and reduces infarct volume after focal cerebral ischemia, through an autophagy-dependent mechanism (Mehta et al., 2012).

Evidence from observational studies indicated that circulating selenium levels were inversely correlated with certain cardiovascular outcomes with a possible U-shaped association, and beneficial effects against IS were found in IS patients as well (Flores-Mateo et al., 2006; Stranges et al., 2010; Rees et al., 2013). However, results from clinical trials were controversial. Specifically, reports of the Selenium and Vitamin E Cancer Prevention Trial (SELECT) and Nutritional Prevention of Cancer Trial (NPC) found no beneficial effects on the incidence and mortality of coronary heart disease and stroke (Stranges et al., 2006; Lippman et al., 2009). In addition, results from a population-based survey revealed that blood selenium concentration might be inversely associated with the prevalence of stroke, and the relationship was non-linear (Hu et al., 2019). However, due to selection bias and reverse causation, the association between selenium levels and the risk of IS may be overestimated. In addition, whether selenium had different impacts on IS subtypes remains unclear. Mendelian randomization (MR), which uses genetic variants as instrumental variables, is a powerful method for inferring causal links between exposures and outcomes. MR analysis uses genetic variants associated with the selenium levels, as the



Estimates of Selenium on IS

Methods	OR (95% CI)	P Value
All causes IS		
MR−Egger	0.954 (0.785−1.160)	0.686
Simple median	0.963 (0.893−1.038)	0.327
Weighted median	0.966 (0.908−1.027)	0.268
IVW	0.968 (0.914−1.026)	0.269
Simple mode	0.957 (0.868−1.055)	0.440
Weighted mode	0.965 (0.901−1.033)	0.381
RAPS	0.968 (0.911−1.028)	0.290
BWMR	0.982 (0.923−1.041)	0.535
MR−PRESSO	0.982 (0.939−1.027)	0.489
MR−LASSO	0.982 (0.928−1.040)	0.541
LVS		
MR−Egger	1.059 (0.654−1.716)	0.837
Simple median	1.021 (0.850−1.225)	0.826
Weighted median	1.023 (0.873−1.199)	0.780
IVW	1.015 (0.881−1.170)	0.835
Simple mode	1.036 (0.837−1.283)	0.766
Weighted mode	1.026 (0.876−1.201)	0.774
RAPS	1.015 (0.875−1.177)	0.841
BWMR	1.010 (0.875−1.166)	0.893
MR−PRESSO	1.010 (0.911−1.118)	0.866
MR−LASSO	1.010 (0.876−1.163)	0.895
CES		
MR−Egger	1.149 (0.736−1.793)	0.604
Simple median	1.093 (0.939−1.272)	0.251
Weighted median	1.038 (0.917−1.174)	0.559
IVW	1.031 (0.922−1.154)	0.591
Simple mode	1.170 (0.932−1.467)	0.268
Weighted mode	1.018 (0.883−1.175)	0.819
RAPS	1.032 (0.918−1.159)	0.600
BWMR	0.975 (0.871−1.093)	0.666
MR−PRESSO	0.975 (0.870−1.093)	0.691
MR−LASSO	0.975 (0.870−1.093)	0.662
SVS		
MR−Egger	1.054 (0.642−1.732)	0.854
Simple median	1.003 (0.838−1.200)	0.977
Weighted median	0.992 (0.855−1.151)	0.915
IVW	0.984 (0.861−1.124)	0.811
Simple mode	1.021 (0.819−1.273)	0.861
Weighted mode	0.993 (0.843−1.171)	0.943
RAPS	0.984 (0.856−1.130)	0.817
BWMR	0.978 (0.854−1.119)	0.742
MR−PRESSO	0.978 (0.866−1.104)	0.742
MR−LASSO	0.978 (0.856−1.117)	0.741

FIGURE 2 | Mendelian randomization analysis of the causal effects of selenium levels on ischemic stroke. A total of 4 SNPs significantly associated with selenium levels were obtained. MR, Mendelian randomization; IS, ischemic stroke; SNP, single-nucleotide polymorphism; OR, odds ratio; CI, confidential interval; IVW, inverse variance-weighted; RAPS, robust adjusted profile score; BWMR, Bayesian weighted Mendelian randomization; MR-PRESSO, Mendelian randomization pleiotropy residual sum and outlier; MR-LASSO, Mendelian randomization least absolute shrinkage and selection operator; LVS, large-vessel atherosclerosis stroke; CES, cardio-embolic stroke; SVS, small-vessel occlusion stroke.

random allocation in randomized controlled trials, to determine the causal effect of the selenium levels on IS, and vice versa (Davies et al., 2018). Since the genes are randomly allocated at conception, genetically predicted selenium levels are not associated with any potential confounders. In addition, random allocation at birth can also avoid the bias caused by reverse causation, as other factors, like disease status cannot affect the genes (Davies et al., 2018). MR analysis was established by three main assumptions (Emdin et al., 2017). First, instrumental variables were significantly associated with the exposure. Next, no links between instrumental variables and confounders were identified. Last, the impact of instrumental variables on outcome was only via exposure (**Figure 1**). Therefore, MR analysis could overcome the limitations of observational studies and provide insights into the association between selenium and IS. And our aims were to evaluate the causal links between selenium levels and IS and their subtypes by MR analysis.

MATERIALS AND METHODS

Data Sources

The genetic variants associated with selenium levels were obtained from a large genome-wide association study (GWAS) meta-analysis of blood selenium (n = 5,477) and toenail selenium (n = 4,162) levels in people of European ancestry (Evans et al., 2013; Cornelis et al., 2015). The genetic variants associated with IS were obtained from a large GWAS by the MEGASTROKE consortium with 34,217 cases and 406,111 controls (Malik et al., 2018). Based on the Trial of ORG 10172 in Acute Stroke Treatment (TOAST) classification, all IS cases could be further divided into large-vessel atherosclerosis stroke (LVS, n = 4,373), cardio-embolic stroke (CES, n = 7,193), and small-vessel occlusion stroke (SVS, n = 5,386) (Adams et al., 1993; Malik et al., 2018). To perform bidirectional MR analysis, the GWAS of the blood selenium level was used as the outcome dataset (Evans et al., 2013).

Sample overlap was calculated in percentages by dividing the number of participants in the GWAS of selenium levels by the number of participants in the respective cohorts in the GWAS of IS and its subtypes (Evans et al., 2013; Cornelis et al., 2015; Malik et al., 2018). An acceptable level of population overlaps between selenium and IS and its subtypes GWAS datasets was 0.22–0.63%.

Selection Criteria of Genetic Variants

We selected genetic variants associated with selenium levels, IS of all causes, LAS, CES, and SVS at genome-wide significance ($p < 5 \times 10^{-8}$) as instrumental variables. Then linkage disequilibrium was tested among the preliminarily selected single-nucleotide polymorphisms (SNPs), and those with $r^2 > 0.01$ in the 1000 Genome Project of Europeans were excluded. The proportion of variance (R^2) in the selenium levels explained by the selected genetic variants was calculated using the following formula: $R^2 = 2 \times \beta^2 \times (1\text{-EAF}) \times \text{EAF}$, where β represents the estimated effect of the genetic variant and EAF represents the effect allele frequency (Palmer et al., 2012). In addition, F-statistic was calculated using the following formula: $F = R^2 \times (N\text{-}k\text{-}1)/k\ (1\text{-}R^2)$, where R^2 represents the proportion of variance explained by the genetic variants, N represents the sample size, and k represents the number of included SNPs (Palmer et al., 2012). The SNPs with an F-statistic <10 were considered weak instruments and were excluded from the MR analysis (Burgess et al., 2011).

Then, the corresponding genetic variants were obtained from the dataset of outcomes (IS or selenium). If selenium-associated SNPs were not available in the outcome datasets, then a proxy SNP in linkage disequilibrium ($r^2 > 0.9$) was searched online (https://ldlink.nci.nih.gov/) as replacement and used in the further analysis.

All genetic variants were searched in the PhenoScanner V2 database to assess whether those variants were significantly associated with the risk factors for IS and its subtypes (Kamat et al., 2019).

Statistical Analysis

All analyses were conducted by R software (version 4.0.3) with R packages TwoSampleMR, MRPRESSO, and MendelianRandomization (Yavorska and Burgess, 2017; Hemani et al., 2018; Verbanck et al., 2018). The estimated effect for blood and toenail selenium levels was presented as Z-score units per effect allele (Evans et al., 2013; Cornelis et al., 2015). Therefore, the Z-score was converted to β and standard error values by the formulas described previously (Kho et al., 2019). The inverse variance-weighted (IVW) method was used as the determinants of the causal effects of exposures on outcomes (Hemani et al., 2018). We also performed MR-Egger, simple median, weighted median, simple mode, weighted mode, robust adjusted profile score (RAPS), Bayesian weighted Mendelian randomization (BWMR), Mendelian randomization pleiotropy residual sum and outlier (MR-PRESSO), and Mendelian randomization least absolute shrinkage and selection operator (MR-LASSO) methods (Bowden et al., 2015; Bowden et al., 2016; Hartwig et al., 2017; Verbanck et al., 2018; Zhao et al., 2020). Sensitivity tests including the heterogeneity test (Cochrane's Q test), pleiotropy test (MR-Egger intercept test), and leave-one-out test were performed (Bowden et al., 2015). Bonferroni correction (corrected $p = 0.05/X/Y$, where X represents the number of exposures and Y represents the number of outcomes) was used for multiple comparisons.

Power Calculation for Bidirectional Mendelian Randomization Analyses

Statistical power for the bidirectional MR analyses was calculated by mRnd (Brion et al., 2013). The minimum effect estimates of selenium levels required to achieve a power of 80% based on the sample size of the outcome datasets and the R2 by the IVs were calculated and is given in **Supplementary Table S1**.

TABLE 1 | SNPs significantly associated with selenium levels and included in the MR study.

SNP	Nearby gene	Ch	E/O allele	EAF	*N*	β	SE	*Z*-score	*p*-value	R^2
rs921943	DMGDH	5	T/C	0.29	9,639	0.295	0.022	13.14	1.90×10^{-39}	0.0358
rs6859667	HOMER1	5	T/C	0.96	9,639	−0.360	0.052	−6.92	4.40×10^{-12}	0.0099
rs6586282	CBS	21	T/C	0.17	9,639	−0.160	0.027	−5.89	3.96×10^{-9}	0.0072
rs1789953	CBS	21	T/C	0.14	9,639	0.162	0.029	5.52	3.40×10^{-8}	0.0063

SNP, single-nucleotide polymorphism; MR, Mendelian randomization; Ch, chromosome; SE, standardized error; E/O, effect or other; EAF, effect allele frequency.

TABLE 2 | MR results of the effect of IS and its subtypes on selenium levels.

SNP	Nearby Gene	Ch.	E/O Allele	EAF	*N*	Exposure			Outcome[a]		
						β	SE	*p*	β	SE	*p*
IS of all causes											
rs2758612[b]	PMF1-BGLAP	1	T/C	0.645	440,328	0.065	0.011	3.68×10^{-9}	NA	NA	NA
rs34311906[b]	ANK2	4	C/T	0.402	440,328	0.065	0.011	1.07×10^{-8}	NA	NA	NA
rs2634074[b]	RP11-119H12.3	4	T/A	0.212	440,328	0.094	0.012	5.90×10^{-15}	0.018	0.037	0.620
rs2066864	FGG	4	A/G	0.245	440,328	0.063	0.012	3.51×10^{-8}	0.036	0.034	0.296
rs11242678	RP11-157J24.2	6	T/C	0.255	440,328	0.072	0.011	2.70×10^{-10}	0.031	0.034	0.358
rs2107595	HDAC9	7	A/G	0.167	440,328	0.088	0.013	2.33×10^{-11}	−0.034	0.041	0.412
rs473238	WTAPP1	11	T/C	0.133	440,328	0.083	0.015	1.65×10^{-8}	0.057	0.046	0.215
rs3184504	SH2B3	12	T/C	0.472	440,328	0.078	0.010	1.23×10^{-14}	−0.002	0.029	0.957
rs4942561	LRCH1	13	T/G	0.759	440,328	0.066	0.012	1.77×10^{-8}	−0.039	0.033	0.247
LVS											
rs7610618[b]	SIAH2	3	T/C	0.013	150,765	0.845	0.149	1.44×10^{-8}	NA	NA	NA
rs2107595	HDAC9	7	A/G	0.168	150,765	0.236	0.032	1.44×10^{-13}	−0.034	0.041	0.412
rs10820405	LINC01492	9	G/A	0.815	150,765	0.181	0.033	4.51×10^{-8}	−0.083	0.038	0.027
rs476762[b]	MMP3	11	A/T	0.133	150,765	0.201	0.035	1.22×10^{-8}	−0.056	0.043	0.189
CES											
rs146390073[b]	RGS7	1	T/C	0.022	211,763	0.669	0.120	2.20×10^{-8}	NA	NA	NA
rs2466455	RP11-119H12.3	4	T/C	0.783	211,763	−0.299	0.022	2.75×10^{-41}	0.018	0.037	0.626
rs6838973	RP11-119H12.3	4	T/C	0.434	211,763	−0.108	0.020	3.58×10^{-8}	−0.014	0.029	0.628
rs12932445	ZFHX3	16	C/T	0.181	211,763	0.176	0.025	6.88×10^{-13}	−0.017	0.044	0.696

CES, cardio-embolic stroke; Ch, chromosome; E/O, effect/other; EAF, effect allele frequency; IS, ischemic stroke; LVS, large vessel atherosclerosis stroke; MR, mendelian randomization; NA, not applicable; SE, standard error; SNP, single nucleotide polymorphism.

[a] *represented blood selenium level here.*

[b] *not included in the MR analysis.*

RESULTS

The Causal Effects of Selenium Levels on Ischemic Stroke

A total of 4 SNPs (rs921943, rs6859667, rs6586282, and rs1789953) significantly associated with selenium levels were obtained (**Table 1**). The 4 SNPs explained 5.9% of the variance in the selenium levels, and the corresponding *F*-statistic was about 151.8. Then, we used PhenoScanner V2 to find whether horizontal pleiotropy existed in the 4 SNPs (Kamat et al., 2019). We found that rs6586282 was significantly associated with plasma homocysteine levels, and rs921943 was associated with height. In MR analysis, the IVW method indicated no causal effects of selenium levels on IS of all causes (OR = 0.968, 95% CI 0.914–1.026, p = 0.269), LVS (OR = 1.015, 95% CI 0.881–1.170, p = 0.835), CES (OR = 1.031, 95% CI 0.922–1.154, p = 0.591), and SVS (OR = 0.984, 95% CI 0.861–1.124, p = 0.811) (**Supplementary Table S2** and **Figure 2**). Heterogeneity tests indicated no heterogeneities of the genetic variants for IS of all causes (p = 0.626), LVS (p = 0.472), CES (p = 0.259), and SVS (p = 0.293) (**Supplementary Table S3**), and pleiotropy tests indicated no pleiotropy of the genetic variants for IS of all causes (p = 0.896), LVS (p = 0.874), CES (p = 0.669), and SVS (p = 0.802) (**Supplementary Table S3**). Leave-one-out analysis indicated that the results were still powerful and stable even if they excluded any single SNP (**Supplementary Figure S1**). Likewise, excluding the effect of rs6586282 did not significantly change the results of MR analysis (**Supplementary Figure S1**). Altogether, our results indicated no causal effects of selenium levels on IS and its subtypes by MR analysis.

The Causal Effects of Ischemic Stroke on Blood Selenium

To further explore the association between the blood selenium level and IS and its subtypes, we further performed bidirectional MR analysis to estimate the causal effects of IS and its subtypes on blood selenium level. Overall, 9, 4, and 4 SNPs significantly associated with IS of all causes, LVS, and CES were obtained, respectively (**Supplementary Table S2**). No SNPs significantly

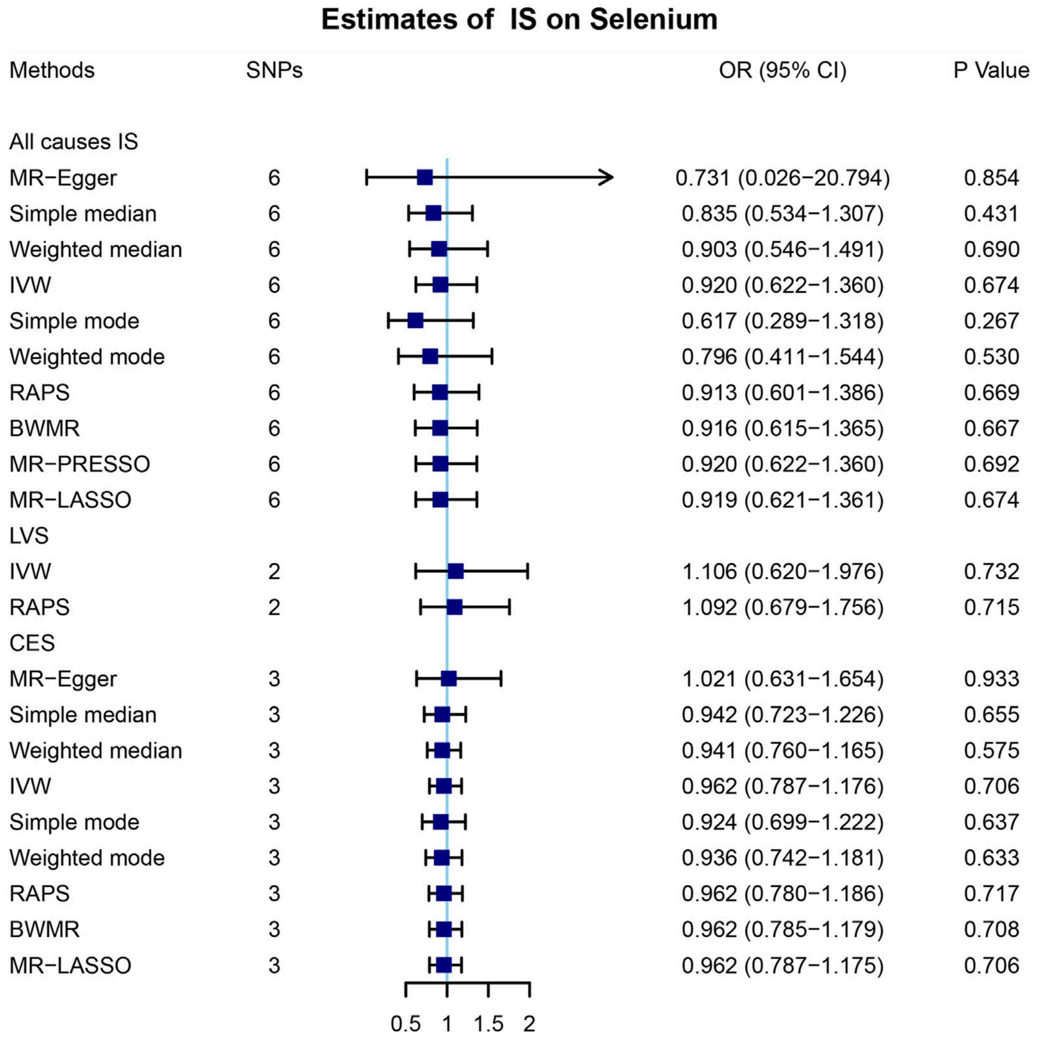


FIGURE 3 | Mendelian randomization analysis of the causal effects of ischemic stroke on blood selenium levels. A total of 6, 2, and 3 SNPs significantly associated with IS of all causes, LVS, and CES were obtained in the reverse Mendelian randomization analysis. MR, Mendelian randomization; IS, ischemic stroke; SNP, single-nucleotide polymorphism; OR, odds ratio; CI, confidential interval; IVW, inverse variance-weighted; RAPS, robust adjusted profile score; BWMR, Bayesian weighted Mendelian randomization; MR-PRESSO, Mendelian randomization pleiotropy residual sum and outlier; MR-LASSO, Mendelian randomization least absolute shrinkage and selection operator; LVS, large-vessel atherosclerosis stroke; CES, cardio-embolic stroke; SVS, small-vessel occlusion stroke.

TABLE 3 | Sensitivity analysis of ischemic stroke and selenium levels.

	Pleiotropy		Heterogeneity	
	Intercept	*p*-value	Q	*p*-value
Exposures				
IS of all causes	0.124	0.404	4.426	0.352
LVS	–	–	4.887[a]	0.027
CES	0.027	0.672	0.157	0.692

IS, ischemic stroke; LVS, large vessel atherosclerosis stroke; CE, cardio-embolic stroke.
[a]by inverse variance weighted method.

associated with SVS were identified. After testing for linkage disequilibrium, 7, 2, and 3 SNPs significantly associated with IS of all causes, LVS, and CES remained, respectively (**Table 2**; **Supplementary Tables S2–S4**). By using the IVW method, our results indicated no causal effects of IS of all causes (OR = 0.920, 95% CI 0.622–1.360, $p = 0.674$), LVS (OR = 1.105, 95% CI 0.620–1.976, $p = 0.732$), and CES (OR = 0.962, 95% CI 0.787–1.176, $p = 0.706$) on the blood selenium level (**Supplementary Tables S2–S4** and **Figure 3**). Sensitivity analysis indicated heterogeneities in the analysis of LVS ($p = 0.027$) and blood selenium level (**Table 3**). No heterogeneities were identified in the analysis of IS of all cause ($p = 0.352$) or CES ($p = 0.692$) (**Table 3**). The pleiotropy test indicated no pleiotropy (IS of all causes: $p = 0.404$; CES: $p = 0.672$) among the genetic variants (**Table 3**). Leave-one-out analysis indicated that the results of our analysis were powerful (**Supplementary Figure S2**). Altogether, our results indicated no causal effects of IS and its subtypes on the blood selenium level by MR analysis.

DISCUSSION

By bidirectional MR analysis based on the summarized data of the GWAS, we found that neither selenium levels were causally associated with IS and its subtypes nor IS and its subtypes were causally associated with selenium levels. The results of our analysis were robust with multiple statistical methods, such as heterogeneity test, pleiotropy test, and leave-one-out analysis.

To our knowledge, the present study is the first study to investigate the causal links between selenium levels and IS and its subtypes by using the bidirectional MR method. Previously, the association between selenium levels and IS was controversial and not well investigated. Prior studies have revealed the potential protective role of selenium in cardiovascular disease. In a case–control study with more than 1,000 Chinese subjects, lower concentrations of selenium were associated with a higher risk of IS (Wen et al., 2019). The inverse association between selenium levels and prevalence of IS was also observed in American subjects (Hu et al., 2019). Nevertheless, Wu et al. (2021) revealed no association between baseline serum selenium levels and stroke in a cohort study (Wei et al., 2004). In a meta-analysis including 12 observational studies, circulating selenium levels were inversely associated with the risk of stroke (Ding and Zhang, 2021). However, in a subgroup analysis, the negative association of selenium levels and stroke was confirmed in the retrospective study group, but not in the prospective study group (Ding and Zhang, 2021). Therefore, the association between selenium levels and IS was controversial and not well investigated. Studies which demonstrated the association between selenium levels and IS with different etiologies were rare. Mironczuk et al. (2021) reported a higher copper-to-selenium ratio in CES patients but a relatively low copper-to-selenium ratio in SVS patients.

The association between selenium levels and stroke is complicated. Selenium is an essential trace element of the human body and shows antioxidant activity by scavenging free radicals (Fang et al., 2002). In the rodent IS model, pretreatment of selenium had significant protective effects on the activity of catalase, superoxide dismutase, and glutathione peroxidase (Ansari et al., 2004). In addition, selenium pretreatment significantly improved hypoxia/ischemia-induced neuron death and reduced infarction volume by alleviating oxidative stress and maintaining mitochondrial function (Mehta et al., 2012). However, the beneficial effect of selenium could be attenuated or even eliminated because of the increasing inflammation and oxidative stress caused by stroke (Ding and Zhang, 2021). Moreover, excess blood selenium concentration (130–150 μg/L) might be associated with minimal mortality (Rayman, 2012).

Gender differences could be a reason for the null finding. Hu et al. (2021) reported a negative association between selenium levels and the first stroke in males but not in females. Different sources (plasma, whole blood, diet, and environment) of selenium used in different studies could be another reason for the null finding and the discrepancy between the present and previous studies (Hu et al., 2017; Merrill et al., 2017; Hu et al., 2019; Wen et al., 2019; Xiao et al., 2019; Hu et al., 2021). Then, regarding the effect of IS on selenium levels, lower selenium levels were observed among acute IS patients in a retrospective study (Angelova et al., 2008). But our analysis provided no evidence of causal effects of IS on selenium levels. Wu et al. (2021) reported genetically predicted selenium levels were negatively causally associated with total cholesterol and low-density lipoprotein cholesterol, which were risk factors for IS (Diener and Hankey, 2020). Furthermore, selenium was reported to be positively correlated with systemic arterial function (Chan et al., 2012). Because previous studies reported non-linear association (including J-shaped and U-shaped) between selenium levels and stroke, the links between selenium levels and IS are rather complicated and still need further investigation (Bleys et al., 2008; Hu et al., 2017; Hu et al., 2019; Hu et al., 2021).

Given the antioxidant activity of selenium and selenoproteins, selenium supplementation was proposed as a potential strategy for the prevention of multiple disorders, like IS, osteoarthritis, rheumatoid arthritis, hypothyroidism, and prostate cancer (Sanmartin et al., 2011). Regarding stroke, selenium supplementation directly into the brain induced the expression of antioxidant glutathione peroxidase 4, which further inhibited the ferroptosis of neurons in a brain hemorrhage model (Alim et al., 2019). In a clinical trial of 29,584 Chinese people, the group receiving selenium supplements for a period of 5 years had a reduction in stroke mortality (9%), but no statistical significance was identified (Mark et al., 1998). Through a secondary analysis of the Nutritional Prevention of Cancer Trial, Stranges et al demonstrated no beneficial effect of selenium supplementation on stroke or cardiovascular disease incidence (Stranges et al., 2006). By bidirectional MR analysis, our results did not support the effectiveness of selenium supplementation in the prevention of IS and its subtypes at the genetic level. Given the impact of selenium levels on blood lipids and arterial function (Chan et al., 2012; Wu et al., 2021), the efficacy of selenium supplementation in subjects with hyperlipidemia or atherosclerotic lesions needed further investigation.

There were some limitations to our study. First, only subjects with European ancestry were included in the MR analysis. The prevalence and incidence of IS vary with ethnicity and so do the proportions of the subtypes of IS (Kim and Kim, 2014). Studies of Western populations indicated CES was the most common subtype of IS, while studies in Asian countries reported a higher prevalence of LVS than CES (Kolominsky-Rabas et al., 2001; Tsai et al., 2013). And the ethnicity differences among the SNPs associated with selenium levels also exist (**Supplementary Table S5**). Therefore, the results of this study needed further validation in Asian or African people. Second, despite including the genetic variants significantly associated with selenium levels from the largest GWAS of selenium levels, only 4 SNPs were finally included in MR analysis. While the 4 SNPs explained approximately 5.9% of the variance of selenium levels and the *F*-statistic of each SNP was more than 10. Therefore, more genetic variants associated with selenium levels, both blood and toenail selenium levels, need to be identified in the future. Third, pleiotropy, which is inevitable in MR analysis, may overestimate the effect of the exposure on the outcome. To eliminate the impact of pleiotropy as much as possible, we sought to identify potential pleiotropic SNPs before the MR analysis. By PhenoScanner, we found one SNP significantly associated with homocysteine. In addition, we performed a pleiotropy test by MR-Egger intercept, and no pleiotropy was found in the present study. Fourth, regarding outcome datasets of selenium levels, only blood selenium levels were used in the MR analysis. So, the causal effects of IS and its subtypes on toenail selenium levels are still unclear. Last, although our analysis suggested no effectiveness of selenium supplementation for patients with IS at the genetic level, large randomized controlled trials are needed to investigate the efficacy and safety of selenium supplementation for IS patients.

CONCLUSION

In conclusion, our bidirectional MR study provides no evidence to support the causal links between genetically predicted selenium levels and IS. Our results also did not support the use of selenium supplementation for IS prevention at the genetic level. Clinical trials with high quality and large sample size are warranted to further elucidate the underlying association between selenium levels and IS and the clinical benefit of selenium supplementation for the prevention of IS.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

YX and HF conceived and designed the study. WL and LZ collected the data. HF, WL and LP analyzed the data. YG, LZ and RZ interpreted the results. JY, BS and YX supervised the study. HF and WL wrote the manuscript. All authors approved the submitted version.

FUNDING

The present study was funded by the Education Department of Henan Province (No. 172102310086) and the Education Department of Henan Province (No. 17A320060).

ACKNOWLEDGMENTS

The authors thank Cornelis et al. and the MEGASTROKE consortium for providing summarized data of GWAS (Evans et al., 2013; Cornelis et al., 2015; Malik et al., 2018).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.782691/full#supplementary-material

Supplementary Figure S1 | Sensitivity analysis of the causal effects of selenium levels on ischemic stroke. **(A)** Leave-one-out analysis of the causal effects of selenium levels on ischemic stroke and its subtypes. **(B)** Scatter plot analysis of the causal effects of selenium levels on ischemic stroke and its subtypes. **(C)** Forest plot analysis of the causal effects of selenium levels on ischemic stroke and its subtypes. **(D)** Funnel plot analysis of the causal effects of selenium levels on ischemic stroke and its subtypes. IS, ischemic stroke; LVS, large-vessel atherosclerosis stroke; CES, cardio-embolic stroke; SVS, small-vessel occlusion stroke; MR, Mendelian randomization; SNP, single-nucleotide polymorphism.

Supplementary Figure S2 | Sensitivity analysis of the causal effects of ischemic stroke on blood selenium levels. **(A)** Leave-one-out analysis of the causal effects of ischemic stroke and its subtypes on blood selenium levels. **(B)** Scatter plot analysis of the causal effects of ischemic stroke and its subtypes on blood selenium levels. **(C)** Forest plot analysis of the causal effects of ischemic stroke on blood selenium levels. **(D)** Funnel plot analysis of the causal effects of ischemic stroke on blood selenium levels. IS, ischemic stroke; LVS, large-vessel atherosclerosis stroke; CES, cardio-embolic stroke; SVS, small-vessel occlusion stroke; MR, Mendelian randomization; SNP, single-nucleotide polymorphism.

REFERENCES

Adams, H. P., Bendixen, B. H., Kappelle, L. J., Biller, J., Love, B. B., Gordon, D. L., et al. (1993). Classification of Subtype of Acute Ischemic Stroke. Definitions for Use in a Multicenter Clinical Trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. *Stroke* 24, 35–41. doi:10.1161/01.str.24.1.35

Alim, I., Caulfield, J. T., Chen, Y., Swarup, V., Geschwind, D. H., Ivanova, E., et al. (2019). Selenium Drives a Transcriptional Adaptive Program to Block Ferroptosis and Treat Stroke. *Cell* 177, 1262–1279. doi:10.1016/j.cell.2019.03.032

Amani, H., Habibey, R., Hajmiresmail, S. J., Latifi, S., Pazoki-Toroudi, H., and Akhavan, O. (2017). Antioxidant Nanomaterials in Advanced Diagnoses and Treatments of Ischemia Reperfusion Injuries. *J. Mater. Chem. B* 5, 9452–9476. doi:10.1039/c7tb01689a

Amani, H., Habibey, R., Shokri, F., Hajmiresmail, S. J., Akhavan, O., Mashaghi, A., et al. (2019). Selenium Nanoparticles for Targeted Stroke Therapy through Modulation of Inflammatory and Metabolic Signaling. *Sci. Rep.* 9, 6044. doi:10.1038/s41598-019-42633-9

Angelova, E. A., Atanassova, P. A., Chalakova, N. T., and Dimitrov, B. D. (2008). Associations between Serum Selenium and Total Plasma Homocysteine during the Acute Phase of Ischaemic Stroke. *Eur. Neurol.* 60, 298–303. doi:10.1159/000157884

Ansari, M. A., Ahmad, A. S., Ahmad, M., Salim, S., Yousuf, S., Ishrat, T., et al. (2004). Selenium Protects Cerebral Ischemia in Rat Brain Mitochondria. *Bter* 101, 73–86. doi:10.1385/BTER:101:1:73

Bleys, J., Navas-Acien, A., and Guallar, E. (2008). Serum Selenium Levels and All-Cause, Cancer, and Cardiovascular Mortality Among US Adults. *Arch. Intern. Med.* 168, 404–410. doi:10.1001/archinternmed.2007.74

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian Randomization with Invalid Instruments: Effect Estimation and Bias Detection through Egger Regression. *Int. J. Epidemiol.* 44, 512–525. doi:10.1093/ije/dyv080

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* 40, 304–314. doi:10.1002/gepi.21965

Brion, M.-J. A., Shakhbazov, K., and Visscher, P. M. (2013). Calculating Statistical Power in Mendelian Randomization Studies. *Int. J. Epidemiol.* 42, 1497–1501. doi:10.1093/ije/dyt179

Burgess, S., Thompson, S. G., and Crp Chd Genetics Collaboration (2011). Avoiding Bias from Weak Instruments in Mendelian Randomization Studies. *Int. J. Epidemiol.* 40, 755–764. doi:10.1093/ije/dyr036

Burk, R. F., Hill, K. E., Motley, A. K., Winfrey, V. P., Kurokawa, S., Mitchell, S. L., et al. (2014). Selenoprotein P and Apolipoprotein E Receptor-2 Interact at the Blood-brain Barrier and Also within the Brain to Maintain an Essential Selenium Pool that Protects against Neurodegeneration. *FASEB j.* 28, 3579–3588. doi:10.1096/fj.14-252874

Campbell, B. C. V., De Silva, D. A., Macleod, M. R., Coutts, S. B., Schwamm, L. H., Davis, S. M., et al. (2019). Ischaemic Stroke. *Nat. Rev. Dis. Primers* 5, 70. doi:10.1038/s41572-019-0118-8

Cardoso, B. R., Roberts, B. R., Bush, A. I., and Hare, D. J. (2015). Selenium, Selenoproteins and Neurodegenerative Diseases. *Metallomics* 7, 1213–1228. doi:10.1039/c5mt00075k

Chan, Y.-H., Siu, C.-W., Yiu, K.-H., Chan, H.-T., Li, S.-W., Tam, S., et al. (2012). Adverse Systemic Arterial Function in Patients with Selenium Deficiency. *J. Nutr. Health Aging* 16, 85–88. doi:10.1007/s12603-011-0086-5

Cornelis, M. C., Fornage, M., Foy, M., Xun, P., Gladyshev, V. N., Morris, S., et al. (2015). Genome-wide Association Study of Selenium Concentrations. *Hum. Mol. Genet.* 24, 1469–1477. doi:10.1093/hmg/ddu546

Davies, N. M., Holmes, M. V., and Davey Smith, G. (2018). Reading Mendelian Randomisation Studies: a Guide, Glossary, and Checklist for Clinicians. *BMJ* 362, k601. doi:10.1136/bmj.k601

Diener, H.-C., and Hankey, G. J. (2020). Primary and Secondary Prevention of Ischemic Stroke and Cerebral Hemorrhage. *J. Am. Coll. Cardiol.* 75, 1804–1818. doi:10.1016/j.jacc.2019.12.072

Ding, J., and Zhang, Y. (2021). Relationship between the Circulating Selenium Level and Stroke: A Meta-Analysis of Observational Studies. *J. Am. Coll. Nutr.*, 1–9. doi:10.1080/07315724.2021.1902880

Emdin, C. A., Khera, A. V., and Kathiresan, S. (2017). Mendelian Randomization. *Jama* 318, 1925–1926. doi:10.1001/jama.2017.17219

Evans, D. M., Zhu, G., Dy, V., Heath, A. C., Madden, P. A. F., Kemp, J. P., et al. (2013). Genome-wide Association Study Identifies Loci Affecting Blood Copper, Selenium and Zinc. *Hum. Mol. Genet.* 22, 3998–4006. doi:10.1093/hmg/ddt239

Fang, Y.-Z., Yang, S., and Wu, G. (2002). Free Radicals, Antioxidants, and Nutrition. *Nutrition* 18, 872–879. doi:10.1016/s0899-9007(02)00916-4

Feigin, V. L., Roth, G. A., Naghavi, M., Parmar, P., Krishnamurthi, R., Chugh, S., et al. (2016). Global burden of Stroke and Risk Factors in 188 Countries, during 1990-2013: a Systematic Analysis for the Global Burden of Disease Study 2013. *Lancet Neurol.* 15, 913–924. doi:10.1016/S1474-4422(16)30073-4

Flores-Mateo, G., Navas-Acien, A., Pastor-Barriuso, R., and Guallar, E. (2006). Selenium and Coronary Heart Disease: a Meta-Analysis. *Am. J. Clin. Nutr.* 84, 762–773. doi:10.1093/ajcn/84.4.762

Go, A. S., Mozaffarian, D., Roger, V. L., Benjamin, E. J., Berry, J. D., Blaha, M. J., et al. (2014). Heart Disease and Stroke Statistics-2014 Update. *Circulation* 129, e28. doi:10.1161/01.cir.0000441139.02102.80

Hartwig, F. P., Davey Smith, G., and Bowden, J. (2017). Robust Inference in Summary Data Mendelian Randomization via the Zero Modal Pleiotropy assumption. *Int. J. Epidemiol.* 46, 1985–1998. doi:10.1093/ije/dyx102

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., et al. (2018). The MR-Base Platform Supports Systematic Causal Inference across the Human Phenome. *Elife* 7, e34408. doi:10.7554/eLife.34408

Hu, H., Bi, C., Lin, T., Liu, L., Song, Y., Wang, B., et al. (2021). Sex Difference in the Association between Plasma Selenium and First Stroke: a Community-Based Nested Case-Control Study. *Biol. Sex. Differ.* 12, 39. doi:10.1186/s13293-021-00383-2

Hu, X. F., Sharin, T., and Chan, H. M. (2017). Dietary and Blood Selenium Are Inversely Associated with the Prevalence of Stroke Among Inuit in Canada. *J. Trace Elem. Med. Biol.* 44, 322–330. doi:10.1016/j.jtemb.2017.09.007

Hu, X. F., Stranges, S., and Chan, L. H. M. (2019). Circulating Selenium Concentration Is Inversely Associated with the Prevalence of Stroke: Results from the Canadian Health Measures Survey and the National Health and Nutrition Examination Survey. *Jaha* 8, e012290. doi:10.1161/JAHA.119.012290

Kamat, M. A., Blackshaw, J. A., Young, R., Surendran, P., Burgess, S., Danesh, J., et al. (2019). PhenoScanner V2: an Expanded Tool for Searching Human Genotype-Phenotype Associations. *Bioinformatics* 35, 4851–4853. doi:10.1093/bioinformatics/btz469

Kho, P. F., Glubb, D. M., Thompson, D. J., Spurdle, A. B., and O'Mara, T. A. (2019). Assessing the Role of Selenium in Endometrial Cancer Risk: A Mendelian Randomization Study. *Front. Oncol.* 9, 182. doi:10.3389/fonc.2019.00182

Kim, B. J., and Kim, J. S. (2014). Ischemic Stroke Subtype Classification: an Asian Viewpoint. *J. Stroke* 16, 8–17. doi:10.5853/jos.2014.16.1.8

Kolominsky-Rabas, P. L., Weber, M., Gefeller, O., Neundoerfer, B., and Heuschmann, P. U. (2001). Epidemiology of Ischemic Stroke Subtypes According to TOAST Criteria. *Stroke* 32, 2735–2740. doi:10.1161/hs1201.100209

Krishnamurthi, R. V., Feigin, V. L., Forouzanfar, M. H., Mensah, G. A., Connor, M., Bennett, D. A., et al. (2013). Global and Regional burden of First-Ever Ischaemic and Haemorrhagic Stroke during 1990-2010: Findings from the Global Burden of Disease Study 2010. *Lancet Glob. Health* 1, e259–e281. doi:10.1016/S2214-109X(13)70089-5

Lippman, S. M., Klein, E. A., Goodman, P. J., Lucia, M. S., Thompson, I. M., Ford, L. G., et al. (2009). Effect of Selenium and Vitamin E on Risk of Prostate Cancer and Other Cancers. *JAMA* 301, 39–51. doi:10.1001/jama.2008.864

Malik, R., Chauhan, G., Chauhan, G., Traylor, M., Sargurupremraj, M., Okada, Y., et al. (2018). Multiancestry Genome-wide Association Study of 520,000 Subjects Identifies 32 Loci Associated with Stroke and Stroke Subtypes. *Nat. Genet.* 50, 524–537. doi:10.1038/s41588-018-0058-3

Mark, S. D., Wang, W., Mark, J. F., Fraumeni, J. F., Li, J.-Y., Taylor, P. R., et al. (1998). Do nutritional Supplements Lower the Risk of Stroke or Hypertension? *Epidemiology* 9, 9–15. doi:10.1097/00001648-199801000-00005

Mehta, S. L., Kumari, S., Mendelev, N., and Li, P. A. (2012). Selenium Preserves Mitochondrial Function, Stimulates Mitochondrial Biogenesis, and Reduces Infarct Volume after Focal Cerebral Ischemia. *BMC Neurosci.* 13, 79. doi:10.1186/1471-2202-13-79

Merrill, P. D., Ampah, S. B., He, K., Rembert, N. J., Brockman, J., Kleindorfer, D., et al. (2017). Association between Trace Elements in the Environment and Stroke Risk: The Reasons for Geographic and Racial Differences in Stroke (REGARDS) Study. *J. Trace Elem. Med. Biol.* 42, 45–49. doi:10.1016/j.jtemb.2017.04.003

Mirończuk, A., Kapica-Topczewska, K., Socha, K., Soroczyńska, J., Jamiołkowski, J., Kułakowska, A., et al. (2021). Selenium, Copper, Zinc Concentrations and Cu/Zn, Cu/Se Molar Ratios in the Serum of Patients with Acute Ischemic Stroke in Northeastern Poland-A New Insight into Stroke Pathophysiology. *Nutrients* 13, 2139. doi:10.3390/nu13072139

Palmer, T. M., Lawlor, D. A., Harbord, R. M., Sheehan, N. A., Tobias, J. H., Timpson, N. J., et al. (2012). Using Multiple Genetic Variants as Instrumental Variables for Modifiable Risk Factors. *Stat. Methods Med. Res.* 21, 223–242. doi:10.1177/0962280210394459

Phipps, M. S., and Cronin, C. A. (2020). Management of Acute Ischemic Stroke. *BMJ* 368, l6983. doi:10.1136/bmj.l6983

Rayman, M. P. (2012). Selenium and Human Health. *The Lancet* 379, 1256–1268. doi:10.1016/S0140-6736(11)61452-9

Rees, K., Hartley, L., Day, C., Flowers, N., Clarke, A., and Stranges, S. (2013). Selenium Supplementation for the Primary Prevention of Cardiovascular Disease. *Cochrane Database Syst. Rev.* 2013, CD009671. doi:10.1002/14651858.CD009671.pub2

Sanmartin, C., Plano, D., Font, M., and Palop, J. A. (2011). Selenium and Clinical Trials: New Therapeutic Evidence for Multiple Diseases. *Cmc* 18, 4635–4650. doi:10.2174/092986711797379249

Saxton, R. A., and Sabatini, D. M. (2017). mTOR Signaling in Growth, Metabolism, and Disease. *Cell* 168, 960–976. doi:10.1016/j.cell.2017.02.004

Scheiber, I. F., Mercer, J. F. B., and Dringen, R. (2014). Metabolism and Functions of Copper in Brain. *Prog. Neurobiol.* 116, 33–57. doi:10.1016/j.pneurobio.2014.01.002

Stranges, S., Marshall, J. R., Trevisan, M., Natarajan, R., Donahue, R. P., Combs, G. F., et al. (2006). Effects of Selenium Supplementation on Cardiovascular Disease Incidence and Mortality: Secondary Analyses in a Randomized Clinical Trial. *Am. J. Epidemiol.* 163, 694–699. doi:10.1093/aje/kwj097

Stranges, S., Navas-Acien, A., Rayman, M. P., and Guallar, E. (2010). Selenium Status and Cardiometabolic Health: State of the Evidence. *Nutr. Metab. Cardiovasc. Dis.* 20, 754–760. doi:10.1016/j.numecd.2010.10.001

Tsai, C.-F., Thomas, B., and Sudlow, C. L. M. (2013). Epidemiology of Stroke and its Subtypes in Chinese vs white Populations: a Systematic Review. *Neurology* 81, 264–272. doi:10.1212/WNL.0b013e31829bfde3

Verbanck, M., Chen, C.-Y., Neale, B., and Do, R. (2018). Detection of Widespread Horizontal Pleiotropy in Causal Relationships Inferred from Mendelian Randomization between Complex Traits and Diseases. *Nat. Genet.* 50, 693–698. doi:10.1038/s41588-018-0099-7

Wei, W.-Q., Abnet, C. C., Qiao, Y.-L., Dawsey, S. M., Dong, Z.-W., Sun, X.-D., et al. (2004). Prospective Study of Serum Selenium Concentrations and Esophageal and Gastric Cardia Cancer, Heart Disease, Stroke, and Total Death. *Am. J. Clin. Nutr.* 79, 80–85. doi:10.1093/ajcn/79.1.80

Wen, Y., Huang, S., Zhang, Y., Zhang, H., Zhou, L., Li, D., et al. (2019). Associations of Multiple Plasma Metals with the Risk of Ischemic Stroke: A Case-Control Study. *Environ. Int.* 125, 125–134. doi:10.1016/j.envint.2018.12.037

Wu, Q., Sun, X., Chen, Q., Zhang, X., and Zhu, Y. (2021). Genetically Predicted Selenium Is Negatively Associated with Serum TC, LDL-C and Positively Associated with HbA1C Levels. *J. Trace Elem. Med. Biol.* 67, 126785. doi:10.1016/j.jtemb.2021.126785

Xiao, Y., Yuan, Y., Liu, Y., Yu, Y., Jia, N., Zhou, L., et al. (2019). Circulating Multiple Metals and Incident Stroke in Chinese Adults. *Stroke* 50, 1661–1668. doi:10.1161/STROKEAHA.119.025060

Yavorska, O. O., and Burgess, S. (2017). MendelianRandomization: an R Package for Performing Mendelian Randomization Analyses Using Summarized Data. *Int. J. Epidemiol.* 46, 1734–1739. doi:10.1093/ije/dyx034

Zecca, L., Youdim, M. B. H., Riederer, P., Connor, J. R., and Crichton, R. R. (2004). Iron, Brain Ageing and Neurodegenerative Disorders. *Nat. Rev. Neurosci.* 5, 863–873. doi:10.1038/nrn1537

Zhao, J., Ming, J., Hu, X., Chen, G., Liu, J., and Yang, C. (2020). Bayesian Weighted Mendelian Randomization for Causal Inference Based on Summary Statistics. *Bioinformatics* 36, 1501–1508. doi:10.1093/bioinformatics/btz749

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 05 May 2022
doi: 10.3389/fgene.2022.868408

The Association Between Vitamin C and Cancer: A Two-Sample Mendelian Randomization Study

Hanxiao Chen[1,2,3]†, Ze Du[4]†, Yaoyao Zhang[2,3]†, Mengling Li[2,3], Rui Gao[2,3], Lang Qin[2,3] and Hongjing Wang[1,2]**

[1]Department of Obstetrics and Gynaecology, West China Second University Hospital, Sichuan University, Chengdu, China, [2]Key Laboratory of Birth Defects and Related Diseases of Women and Children of the Ministry of Education, West China Second University Hospital, Sichuan University, Chengdu, China, [3]Reproductive Center, Department of Obstetrics and Gynaecology, West China Second University Hospital, Sichuan University, Chengdu, China, [4]Department of Orthopedics, Research Institute of Orthopedics, West China Hospital/West China School of Medicine, Sichuan University, Chengdu, China

In recent years, many studies have indicated that vitamin C might be negatively associated with the risk of cancer, but the actual relationship between vitamin C and cancer remains ambivalent. Therefore, we utilized a two-sample Mendelian randomization (MR) study to explore the causal associations of genetically predicted vitamin C with the risk of a variety of cancers. Single-nucleotide polymorphisms (SNPs) associated with vitamin C at a significance level of $p < 5 \times 10–8$ and with a low level of linkage disequilibrium (LD) (r2 < 0.01) were selected from a genome-wide association study (GWAS) meta-analysis of plasmid concentration of vitamin C consisting of 52,018 individuals. The data of the GWAS outcomes were obtained from United Kingdom Biobank, FinnGen Biobank and the datasets of corresponding consortia. In the inverse-variance weight (IVW) method, our results did not support the causal association of genetically predicted vitamin C with the risk of overall cancer and 14 specific types of cancer. Similar results were observed in sensitivity analyses where the weighted median and MR-Egger methods were adopted, and heterogeneity and pleiotropy were not observed in statistical models. Therefore, our study suggested that vitamin C was not causally associated with the risk of cancer. Further studies are warranted to discover the potential protective and therapeutic effects of vitamin C on cancer, and its underlying mechanisms.

Keywords: vitamin c, cancer, GWAS, SNP, Mendelian randomization

Edited by:
Guiyou Liu,
Tianjin Institute of Industrial Biotechnology (CAS), China

Reviewed by:
John Frederick Pearson,
University of Otago, New Zealand
Cheryl D. Cropp,
Samford University, United States

Correspondence:
Lang Qin
cacier@163.com
Hongjing Wang
whjscdx@163.com

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *02 February 2022*
Accepted: *07 April 2022*
Published: *05 May 2022*

Citation:
Chen H, Du Z, Zhang Y, Li M, Gao R, Qin L and Wang H (2022) The Association Between Vitamin C and Cancer: A Two-Sample Mendelian Randomization Study. Front. Genet. 13:868408. doi: 10.3389/fgene.2022.868408

INTRODUCTION

Vitamin C, also called ascorbic acid, is a water-soluble vitamin commonly considered an electron donor with an antioxidant function that can eliminate fatal reactive oxygen species (ROS) (Lane and Richardson, 2014). On the other hand, vitamin C can also be a pro-oxidant at a pharmacological plasma concentration (Padayatty and Levine, 2016). In recent years, many researchers have indicated that vitamin C might be negatively associated with the risk of cancer (Bo et al., 2016; Aune et al., 2018; Jenkins et al., 2021), but the actual relationship and the underlying mechanisms of vitamin C in the pathogenesis or therapeutic effect of cancer remain ambivalent.

Cancer is the second-leading cause of death in the USA and causes approximately 600,000 deaths each year (Islami et al., 2020). Thus, prevention and treatment of cancer are of vital importance. Although cancer is known to be associated with some genetic and environmental

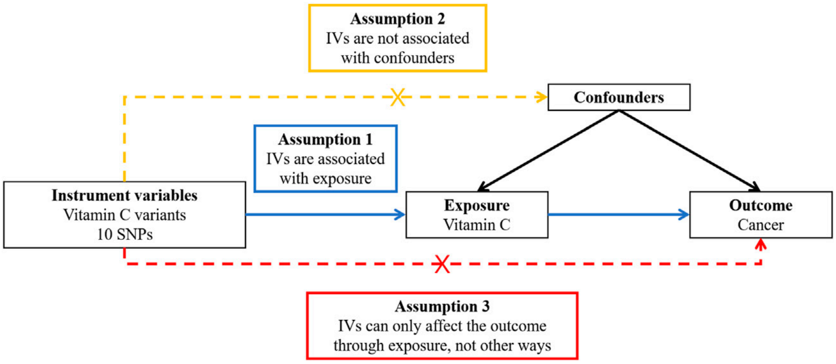

Assumption 2
IVs are not associated with confounders
Confounders
Assumption 1
IVs are associated with exposure
Instrument variables
Vitamin C variants
10 SNPs
Exposure
Vitamin C
Outcome
Cancer
Assumption 3
IVs can only affect the outcome through exposure, not other ways

FIGURE 1 | Overview of the design and three key assumptions of the Mendelian randomization study. IVs, instrument variables; SNPs, single-nucleotide polymorphisms.

factors and different cancers may have different risk factors, some studies suggested that vitamin C may also influence the development of cancer. However, previous studies have yielded inconclusive findings on the potential impact of vitamin C on cancer. One systematic review and dose–response meta-analysis study revealed that when the concentration of vitamin C in blood increased to 50 μmol/L, the relative risk (RR) for total cancer risk was 0.74 (95% confidence interval (CI): 0.66–0.82) (Aune et al., 2018). On the other hand, another systematic review that included 19 trials did not support the positive effect of vitamin C supplementation in patients with cancer on their clinical status and overall survival (van Gorkom et al., 2019). In addition, the relationship between vitamin C and cancer risk may be different in different types of cancer. Vitamin C has been linked to a lower risk of renal cell carcinoma, esophageal cancer, colon cancer, breast cancer, endometrial cancer, and cervical cancer (Bandera et al., 2009; Park et al., 2010; Fulan et al., 2011; Jia et al., 2015a; Bo et al., 2016). However, some studies also suggested that supplementary intake of vitamin C had no relationship with the risk of pancreatic cancer, bladder cancer, prostate cancer, cervical cancer, and ovarian cancer (Jiang et al., 2010; Chen et al., 2015; Cao et al., 2016; Hua et al., 2016; Long et al., 2020). Therefore, the causal role of vitamin C in the development of cancers remains unclear and warrants future studies.

A Mendelian randomization (MR) study uses genetic variation, typically single-nucleotide polymorphisms (SNPs), associated with an exposure to assess its potential causal relationship with an outcome. Compared with traditional observational studies, the MR study provides relatively more convincing evidence for detecting the association between the exposure and the outcome. The MR study can minimize the potential bias generated by potential confounding factors and reverse causality and will not be affected by disease progression because the genetic variants that are used as instrument variables (IVs) in the MR study are strongly and solely related to the exposure (Little, 2018). Using two-sample MR analysis, many studies have found a potential relationship between many risk factors and the risk of cancer (Larsson et al., 2020; Yuan et al., 2020). However, the causal association between vitamin C and the risk of cancer has not yet been fully established using MR analysis. A recent MR study did not support the association between vitamin C and five types of cancer, including lung, breast, prostate, colon, and rectal cancer (Fu et al., 2021), but whether there are causal associations between vitamin C and other types of cancer remains unclear.

Therefore, in this study, we aimed to comprehensively explore the causal associations of genetically predicted vitamin C with the risk of different types of cancer by utilizing a two-sample MR study.

MATERIALS AND METHODS

Study Design

In order to obtain reliable results from a two-sample MR study, the genetic variants used in this study should be in conformity with three principles (**Figure 1**), including the relevance assumption, independence assumption, and exclusion restriction assumption, which means these genetic variants should be strongly related to the exposure (i.e., vitamin C), be not associated with confounding factors of the exposure–outcome relationship, and have an effect on the outcome (i.e., cancer) only through the exposure and not any other pathway (Little, 2018).

Genetic Instrumental Variables for Vitamin C

The SNPs associated with vitamin C were selected from a genome-wide association study (GWAS) meta-analysis of vitamin C (Zheng et al., 2021) consisting of 52,018 individuals from the following studies: 10,771 participants from the Fenland study (Ashor et al., 2017); 16,841 participants from the European Prospective Investigation into Cancer and Nutrition (EPIC)-InterAct study (Consortium, 2011); 16,756 participants from the EPIC Norfolk study (Day et al., 1999) (excluding duplicated samples with EPIC-InterAct); and 7,650

TABLE 1 | Vitamin C SNPs used to construct the instrument variable.

Chr	Position	SNP	Effect allele	Other allele	EAF	Beta	SE	Gene	*p* Value	F Statistics
1	2330190	rs6693447	T	G	0.551	0.039	0.006	RER1	6.25E-10	42.25
2	220031255	rs13028225	T	C	0.857	0.102	0.009	SLC23A3	2.38E-30	128.4444
5	138715502	rs33972313	C	T	0.968	0.36	0.018	SLC23A1	4.61E-90	400
5	176799992	rs10051765	C	T	0.342	0.039	0.007	RGS14	3.64E-09	31.04082
11	61570783	rs174547	C	T	0.328	0.036	0.007	FADS1	3.84E-08	26.44898
12	96249111	rs117885456	A	G	0.087	0.078	0.012	SNRPF	1.70E-11	42.25
12	102093459	rs2559850	A	G	0.598	0.058	0.006	CHPT1	6.30E-20	93.44444
14	105253581	rs10136000	A	G	0.283	0.04	0.007	AKT1	1.33E-08	32.65306
16	79740541	rs56738967	C	G	0.321	0.041	0.007	MAF	7.62E-10	34.30612
17	59456589	rs9895661	T	C	0.817	0.063	0.008	BCAS3	1.05E-14	62.01563

Abbreviations: Chr, chromosome; SNP, single-nucleotide polymorphism; EAF, effect allele frequency; SE, standard error.

TABLE 2 | Characteristics of included studies or consortia of cancer.

Type of Cancer	Source	Year	Sample size	Population
Overall cancer	UKBB	2018	461311	European
	FinnGen Biobank	2020	96499	European
Bronchus and lung	UKBB	2018	361194	European
	FinnGen Biobank	2020	96499	European
	ILCCO	2014	27209	European
Breast	UKBB	2018	462933	European
	FinnGen Biobank	2020	96499	European
	BCAC	2017	228951	European
Pancreas	PanScan1	2009	3,835	European
	FinnGen Biobank	2020	96499	European
Colon	UKBB	2018	462933	European
	FinnGen Biobank	2020	96499	European
Rectum	UKBB	2018	463010	European
	FinnGen Biobank	2020	96499	European
Kidney	UKBB	2018	463010	European
	FinnGen Biobank	2020	96499	European
Bladder	UKBB	2018	462933	European
	FinnGen Biobank	2020	96499	European
Prostate	UKBB	2018	463010	European
	FinnGen Biobank	2020	96499	European
	PRACTICAL	2018	140254	European
Ovary	UKBB	2018	463010	European
	FinnGen Biobank	2020	96499	European
	OCAC	2017	66450	European
Uterus/endometrium	UKBB	2018	462933	European
	FinnGen Biobank	2020	96499	European

Abbreviations: UKBB, United Kingdom, biobank; ILCCO, international lung cancer consortium; BCAC, breast cancer association consortium; PanScan1, Pancreatic Cancer Cohort Consortium GWAS; PRACTICAL, prostate cancer association group to investigate cancer-associated alterations in the genome; OCAC, ovarian cancer association consortium.

participants from the EPIC-CVD study (Danesh et al., 2007) (excluding duplicated samples with EPIC-InterAct or EPIC-Norfolk). A total of 11 independent SNPs were reported to be related to vitamin C at the genome-wide significance level ($p < 5 \times 10^{-8}$). Since rs7740812 was correlated ($r^2 < 0.01$) in linkage disequilibrium (LD) analysis, the remaining 10 SNPs were included to establish the genetic IVs for vitamin C (**Table 1**).

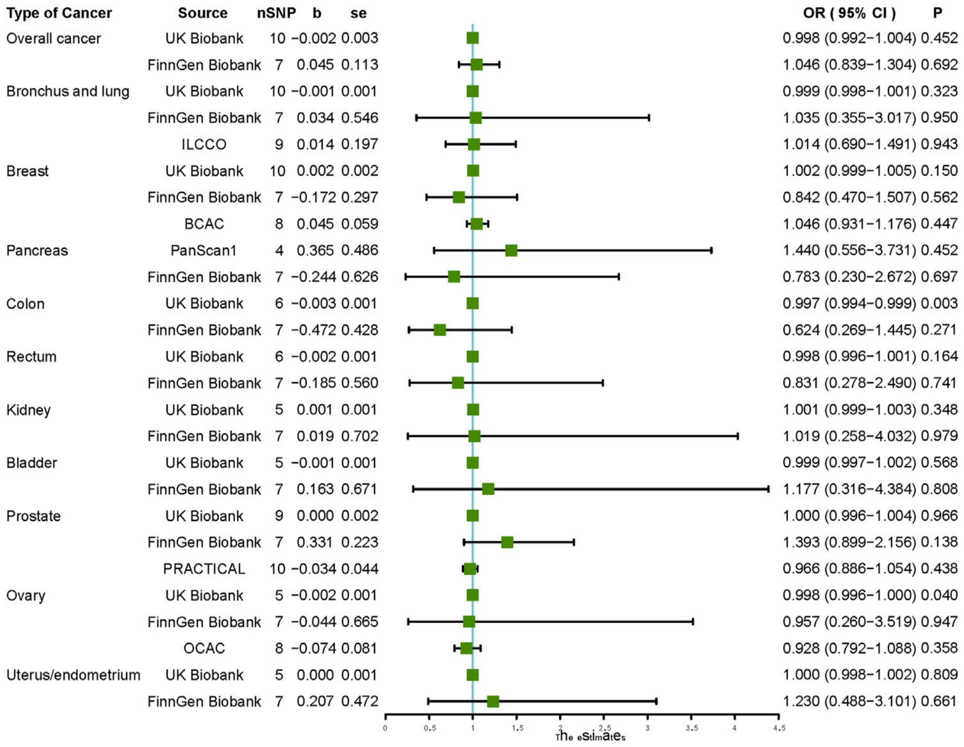
Genetic Association Datasets for Cancer

Overall cancer and ten types of site-specific cancer were included as cancer outcomes in our MR study (**Table 2**). GWAS summary statistics on overall cancer and nine site-specific cancers, including lung, breast, colon, rectum, kidney, bladder, prostate, ovarian, and uterine/endometrial cancer, were obtained from the United Kingdom Biobank dataset. Summary statistics of GWAS on overall cancer and malignant neoplasm of the bronchus and lung, breast, pancreas, colon, rectum, kidney, bladder, prostate, ovary, and corpus uteri were acquired from the FinnGen Biobank database. Summary statistics of GWAS on lung cancer were obtained from the International Lung Cancer Consortium (ILCCO) (Wang et al., 2014). Summary statistics of GWAS on breast cancer were obtained from the Breast Cancer Association Consortium (BCAC) (Michailidou et al., 2017). GWAS summary statistics on pancreatic cancer were obtained from the Pancreatic Cancer Cohort Consortium (PanScan1) (Amundadottir et al., 2009). The GWAS summary of prostate cancer was derived from the Prostate Cancer Association group to Investigate Cancer Associated Alterations in the Genome (PRACTICAL) (Schumacher et al., 2018). Summary statistics of GWAS on ovarian cancer were obtained from the Ovarian Cancer Association Consortium (OCAC) (Phelan et al., 2017). In this study, we extracted the effect estimates and standard errors for each of the 10 vitamin C–related SNPs from the meta-GWAS summary statistics of overall cancer risk and site-specific cancer risk.

TABLE 3 | Associations between genetically predicted vitamin C and risk of cancer.

Type of cancer	Data source	Number of SNPs	Inverse variance weighted		MR-Egger		Simple mode		Weighted median		Weighted mode	
			Or (95%CI)	P	Or (95%CI)	P	Or (95%CI)	P	Or (95%CI)	P	Or (95%CI)	P
Overall cancer	UKBB	10	0.998 (0.992–1.004)	0.452	1.003 (0.994–1.012)	0.562	0.991 (0.977–1.006)	0.28	1.002 (0.994–1.009)	0.663	1.003 (0.995–1.011)	0.492
	FinnGen Biobank	7	1.046 (0.839–1.304)	0.692	1.107 (0.779–1.573)	0.596	1.073 (0.731–1.574)	0.732	1.076 (0.848–1.367)	0.546	1.077 (0.834–1.392)	0.59
Bronchus and lung	UKBB	10	0.999 (0.998–1.001)	0.323	1.000 (0.997–1.003)	0.982	0.996 (0.993–1.000)	0.058	1.000 (0.998–1.002)	0.795	1.000 (0.998–1.002)	0.889
	FinnGen Biobank	7	1.035 (0.355–3.017)	0.95	2.274 (0.522–9.903)	0.323	1.414 (0.388–5.156)	0.618	1.517 (0.582–3.957)	0.394	1.590 (0.561–4.505)	0.416
	ILCCO	9	1.014 (0.690–1.491)	0.943	1.259 (0.678–2.340)	0.49	1.174 (0.757–1.820)	0.494	1.075 (0.830–1.391)	0.586	1.078 (0.829–1.402)	0.592
Breast	UKBB	10	1.002 (0.999–1.005)	0.15	1.002 (0.998–1.007)	0.362	0.999 (0.993–1.006)	0.826	1.002 (0.998–1.006)	0.292	1.002 (0.998–1.007)	0.301
	FinnGen Biobank	7	0.842 (0.470–1.507)	0.562	0.516 (0.246–1.079)	0.139	0.780 (0.381–1.598)	0.523	0.669 (0.430–1.041)	0.075	0.652 (0.420–1.012)	0.105
	BCAC	8	1.046 (0.931–1.176)	0.447	1.039 (0.862–1.252)	0.704	1.002 (0.826–1.215)	0.985	1.042 (0.948–1.146)	0.389	1.053 (0.951–1.166)	0.355
Pancreas	PanScan1	4	1.440 (0.556–3.731)	0.452	0.612 (0.058–6.485)	0.723	1.253 (0.289–5.439)	0.783	1.249 (0.417–3.746)	0.691	1.173 (0.339–4.060)	0.818
	FinnGen Biobank	7	0.783 (0.230–2.672)	0.697	0.873 (0.120–6.358)	0.898	1.043 (0.130–8.376)	0.969	0.721 (0.177–2.925)	0.647	0.897 (0.210–3.830)	0.888
Colon	UKBB	6	0.997 (0.994–0.999)	0.003	1.000 (0.987–1.013)	0.986	0.997 (0.993–1.001)	0.167	0.997 (0.994–1.000)	0.048	0.997 (0.993–1.001)	0.164
	FinnGen Biobank	7	0.624 (0.269–1.445)	0.271	0.590 (0.151–2.297)	0.481	0.633 (0.138–2.889)	0.576	0.616 (0.252–1.503)	0.287	0.619 (0.256–1.497)	0.328
Rectum	UKBB	6	0.998 (0.996–1.001)	0.164	0.993 (0.980–1.006)	0.342	0.998 (0.993–1.002)	0.39	0.998 (0.995–1.001)	0.157	0.997 (0.993–1.001)	0.227
	FinnGen Biobank	7	0.831 (0.278–2.490)	0.741	1.287 (0.252–6.556)	0.774	0.361 (0.064–2.027)	0.291	0.971 (0.273–3.457)	0.964	1.055 (0.252–4.418)	0.944
Kidney	UKBB	5	1.001 (0.999–1.003)	0.348	1.008 (0.996–1.019)	0.296	1.002 (0.998–1.005)	0.405	1.002 (0.999–1.004)	0.168	1.002 (0.999–1.006)	0.319
	FinnGen Biobank	7	1.019 (0.258–4.032)	0.979	3.268 (0.566–18.852)	0.243	1.411 (0.226–8.812)	0.725	1.875 (0.555–6.342)	0.312	1.936 (0.606–6.192)	0.308
Bladder	UKBB	5	0.999 (0.997–1.002)	0.568	1.005 (0.993–1.017)	0.475	1.000 (0.996–1.004)	0.921	1.000 (0.998–1.003)	0.869	1.001 (0.997–1.004)	0.759
	FinnGen Biobank	7	1.177 (0.316–4.384)	0.808	3.023 (0.489–18.67)	0.287	1.916 (0.276–13.28)	0.535	1.694 (0.527–5.442)	0.376	2.039 (0.638–6.515)	0.275
Prostate	UKBB	9	1.000 (0.996–1.004)	0.966	1.000 (0.989–1.011)	0.995	1.002 (0.995–1.009)	0.59	1.000 (0.995–1.005)	0.937	1.001 (0.996–1.007)	0.661
	FinnGen Biobank	7	1.393 (0.899–2.156)	0.138	1.491 (0.779–2.854)	0.282	1.389 (0.612–3.150)	0.462	1.396 (0.826–2.362)	0.213	1.428 (0.831–2.455)	0.245
	PRACTICAL	10	0.966 (0.886–1.054)	0.438	0.974 (0.850–1.116)	0.715	0.984 (0.823–1.177)	0.863	0.980 (0.880–1.091)	0.709	0.986 (0.876–1.108)	0.814
Ovary	UKBB	5	0.998 (0.996–1.000)	0.04	0.996 (0.984–1.007)	0.526	0.997 (0.994–1.001)	0.192	0.998 (0.995–1.000)	0.052	0.997 (0.994–1.000)	0.17
	FinnGen Biobank	7	0.957 (0.260–3.519)	0.947	0.470 (0.068–3.254)	0.479	1.247 (0.124–12.539)	0.857	0.685 (0.148–3.172)	0.628	0.655 (0.117–3.652)	0.646
	OCAC	8	0.928 (0.792–1.088)	0.358	0.801 (0.638–1.006)	0.105	1.093 (0.797–1.498)	0.599	0.891 (0.739–1.074)	0.227	0.857 (0.712–1.033)	0.149
Uterus/ endometrium	UKBB	5	1.000 (0.998–1.002)	0.809	1.004 (0.992–1.016)	0.527	1.000 (0.996–1.003)	0.973	1.000 (0.997–1.003)	0.948	1.000 (0.997–1.003)	0.889
	FinnGen Biobank	7	1.230 (0.488–3.101)	0.661	2.922 (0.743–11.488)	0.185	0.862 (0.134–5.553)	0.881	1.864 (0.607–5.723)	0.276	1.940 (0.661–5.696)	0.273

Abbreviations: SNP, single-nucleotide polymorphism; OR, odds ratio; CI, confidence interval; UKBB, UK biobank; ILCCO, international lung cancer consortium; BCAC, breast cancer association consortium; PanScan1, Pancreatic Cancer Cohort Consortium GWAS; PRACTICAL, prostate cancer association group to investigate cancer-associated alterations in the genome; OCAC, ovarian cancer association consortium.



Type of Cancer	Source	nSNP	b	se	OR (95% CI)	P
Overall cancer	UK Biobank	10	−0.002	0.003	0.998 (0.992−1.004)	0.452
	FinnGen Biobank	7	0.045	0.113	1.046 (0.839−1.304)	0.692
Bronchus and lung	UK Biobank	10	−0.001	0.001	0.999 (0.998−1.001)	0.323
	FinnGen Biobank	7	0.034	0.546	1.035 (0.355−3.017)	0.950
	ILCCO	9	0.014	0.197	1.014 (0.690−1.491)	0.943
Breast	UK Biobank	10	0.002	0.002	1.002 (0.999−1.005)	0.150
	FinnGen Biobank	7	−0.172	0.297	0.842 (0.470−1.507)	0.562
	BCAC	8	0.045	0.059	1.046 (0.931−1.176)	0.447
Pancreas	PanScan1	4	0.365	0.486	1.440 (0.556−3.731)	0.452
	FinnGen Biobank	7	−0.244	0.626	0.783 (0.230−2.672)	0.697
Colon	UK Biobank	6	−0.003	0.001	0.997 (0.994−0.999)	0.003
	FinnGen Biobank	7	−0.472	0.428	0.624 (0.269−1.445)	0.271
Rectum	UK Biobank	6	−0.002	0.001	0.998 (0.996−1.001)	0.164
	FinnGen Biobank	7	−0.185	0.560	0.831 (0.278−2.490)	0.741
Kidney	UK Biobank	5	0.001	0.001	1.001 (0.999−1.003)	0.348
	FinnGen Biobank	7	0.019	0.702	1.019 (0.258−4.032)	0.979
Bladder	UK Biobank	5	−0.001	0.001	0.999 (0.997−1.002)	0.568
	FinnGen Biobank	7	0.163	0.671	1.177 (0.316−4.384)	0.808
Prostate	UK Biobank	9	0.000	0.002	1.000 (0.996−1.004)	0.966
	FinnGen Biobank	7	0.331	0.223	1.393 (0.899−2.156)	0.138
	PRACTICAL	10	−0.034	0.044	0.966 (0.886−1.054)	0.438
Ovary	UK Biobank	5	−0.002	0.001	0.998 (0.996−1.000)	0.040
	FinnGen Biobank	7	−0.044	0.665	0.957 (0.260−3.519)	0.947
	OCAC	8	−0.074	0.081	0.928 (0.792−1.088)	0.358
Uterus/endometrium	UK Biobank	5	0.000	0.001	1.000 (0.998−1.002)	0.809
	FinnGen Biobank	7	0.207	0.472	1.230 (0.488−3.101)	0.661

FIGURE 2 | Causal effect estimates of vitamin C on cancer outcomes. SNP, single-nucleotide polymorphism; OR, odds ratio; CI, confidence interval; UKBB, UKn Biobank; ILCCO, International Lung Cancer Consortium; BCAC, Breast Cancer Association Consortium; PanScan1, Pancreatic Cancer Cohort Consortium GWAS; PRACTICAL, Prostate Cancer Association group To Investigate Cancer-Associated Alterations in the Genome; OCAC, Ovarian Cancer Association Consortium.

Statistical Analysis

An MR analysis was performed utilizing 10 vitamin C–related SNPs as IVs to evaluate the association of vitamin C with overall cancer risk and site-specific cancer risk. We used the inverse-variance weight (IVW) method with random effects to implement the primary MR analysis. The odds ratio (OR) and 95% CI for risk of overall cancer and site-specific cancer were estimated.

We then performed sensitivity analyses, including MR-Egger regression, simple mode, weighted median, and weighted mode methods to determine whether the IVs can influence cancer only through their effect on vitamin C. To test bias from pleiotropic effects, we used MR-Egger regression. In addition, the slope coefficient from an Egger regression provided a reliable estimate of any causal effect (Bowden et al., 2015). The weighted median method could provide a consistent assessment of the finding if more than half of the weight comes from valid IVs (Bowden et al., 2016). When the most common horizontal pleiotropy value was zero regardless of the type of horizontal pleiotropy, we performed the simple mode method to offer a consistent assessment (Bowden et al., 2016). In addition, the weighted mode requires that the largest subset of instruments identifying the same causal effect estimates is contributed by valid IVs (Hartwig et al., 2017). A pleiotropy test was also performed to test whether IVs had horizontal pleiotropy. We also applied the MR-Pleiotropy Residual Sum and Outlier (MR-PRESSO) analysis to determine the horizontal pleiotropy and correct the potential outliers (Verbanck et al., 2018). In addition, we utilized Cochran's Q test on the IVW and MR-Egger estimates to test the heterogeneity of the causal estimates. We also used a leave-one-out sensitivity test to test whether the MR outcome was sensitive to its related IV. MR and sensitivity analyses were performed in R (version 4.0.2) using the Two-Sample MR package (version 0.5.5) and the MRPRESSO package (version 1.0).

RESULTS

Our findings did not support the causal association between vitamin C and the risk of overall cancer in the UK Biobank and FinnGen Biobank (OR: 0.998, 95% CI: 0.992–1.004, p = 0.452, and OR: 1.046, 95% CI: 0.839–1.304, p = 0.692, respectively). The results of MR-Egger, weighted median, simple mode, and weighted mode analyses were similar to those of the IVW (**Table 3**). In sensitivity analysis, heterogeneity was not detected (**Supplementary Table S1**). In addition, we did not detect horizontal pleiotropy *via* pleiotropy tests and MR-PRESSO analysis (**Supplementary Tables S2, S3**). A scatter plot of the association between vitamin C and overall cancer is shown in **Supplementary Figure S1**.

When analyzing the causal relationship between vitamin C and different types of cancer, our IVW results did not support the causal association between vitamin C and the risk of any of the ten types of cancer, including malignant neoplasm of the bronchus and lung, breast, pancreas, colon, rectum, kidney, bladder, prostate, ovary, and endometrium (**Figure 2**). Using MR-Egger, weighted median, simple mode, and weighted mode methods, we obtained similar results to those of IVW, which did not support the causal association between vitamin C and any type of cancer (**Table 3**).

In sensitivity analysis of vitamin C and site-specific cancer, our results did not reveal substantial heterogeneity except for that in lung cancer and breast cancer (**Supplementary Table S1**), and a pleiotropy test using the MR-Egger intercept did not detect any pleiotropy across the studies (**Supplementary Table S2**). In MR-PRESSO analysis, we did not detect horizontal pleiotropy except for the association between vitamin C and lung cancer in the ILCCO dataset (**Supplementary Table S3**). We further found that rs174547 was a potential outlier ($p < 0.01$), and after omitting rs174547, vitamin C was still not associated with the risk of lung cancer (OR: 0.999, 95% CI: 0.998–1.001, $p = 0.481$). Details of the leave-one-out sensitivity test are displayed in **Supplementary Table S4**. A scatter plot of the association between vitamin C and 10 types of site-specific cancer is shown in **Supplementary Figures S2–S11**.

DISCUSSION

The prevention and therapeutic effects of vitamin C on cancer have been debated for decades. In this MR study, we demonstrated that vitamin C was not causally associated with the risk of cancer. In particular, our findings did not support the causal association between vitamin C and the risk of overall cancer or any specific type of cancer, including colon cancer and ovarian cancer, and the risk of malignant neoplasm of the bronchus and lung, breast, pancreas, colon, rectum, kidney, bladder, prostate, ovary, and uterine/endometrium. MR-Egger regression, simple mode, weighted median, and weighted mode methods showed similar findings. In addition, in sensitivity analysis, heterogeneity and horizontal pleiotropy were not detected in most of our studies.

In general, our findings were in line with those of previous studies aimed at investigating the association between vitamin C and cancer. A recent systematic review included 19 clinical trials that did not support the protective effect of vitamin C supplementation in patients with cancer on their clinical status and overall survival (van Gorkom et al., 2019). One meta-analysis included three studies that indicated vitamin C had no significant effect on lung cancer incidence (Cortés-Jofré et al., 2020). A meta-analysis that included 20 observational studies did not support the relationship between vitamin C intake and the risk of pancreatic cancer (Hua et al., 2016). Another meta-analysis of three prospective cohort studies did not observe an association between vitamin C intake and the risk of renal cell carcinoma (Jia et al., 2015b). A meta-analysis involving 16 studies indicated no effect of vitamin C on reducing the risk of ovarian cancer (RR: 0.95, 95% CI: 0.81–1.11) (Long et al., 2020). In addition, for prostate cancer, a meta-analysis that summarized nine RCTs found no relationship between vitamin C intake and the incidence of prostate cancer (RR: 1.45, 95% CI: 0.92–2.29) (Jiang et al., 2010). However, some of our results were inconsistent with those of several observational studies. At the same time, a meta-analysis involving 13 cohort studies suggested that supplementary intake of vitamin C could reduce the risk of colon cancer (RR: 0.81, 95% CI: 0.71–0.92) (Park et al., 2010). Moreover, targeting female-specific tumors, supplementary intake of vitamin C could reduce the risk of cervical neoplasia (OR: 0.58, 95% CI: 0.44–0.75) (Cao et al., 2016). In addition, another meta-analysis included 12 studies suggesting that vitamin C could prevent endometrial cancer (OR: 0.85, 95% CI: 0.73–0.98) (Bandera et al., 2009). But, most of the available clinical studies were cross-sectional, case-control, and cohort studies, the results of which were easily affected by known and unknown confounding factors and reverse causality (Bandera et al., 2009; Bo et al., 2016). Heterogeneity was detected in most of the studies. In addition, case-control studies were also affected by recall and selection biases. The current study used MR analysis, which utilized genetically predicted SNPs as IVs for the exposure, to explore the causal relationship between exposure and outcome that could minimize the effect of the potential confounders and reverse causality. Therefore, the findings of high-quality MR studies could be more convincing than those of the aforementioned observational studies. One previous MR study assessed the relationships between plasma vitamin C levels and five types of cancer, including lung, breast, prostate, colon, and rectal cancer. Similar to our findings, the use of vitamin C supplements was not causally associated with the risk of these types of cancer (Fu et al., 2021).

Previous experiments have well-investigated the therapeutic effects of vitamin C and confirmed that vitamin C is capable of killing cancer cells *in vitro* and shrinking tumor size *in vivo*. Multiple pathways might be involved in the antitumor effect of vitamin C, including targeting redox imbalance, acting as an epigenetic regulator and modifying hypoxia-inducible factor 1 (HIF1) signaling (Cimmino et al., 2017; Ngo et al., 2019). But, there were few experimental studies that supported the prevention effect of vitamin C on the risk of cancer (Reczek and Chandel, 2015). In that case, vitamin C seemed to be unable to reduce cancer incidence but could act as an additional therapeutic agent for cancer treatment. Moreover, even with the usage of supplementary vitamin C, the plasma vitamin C concentration among a healthy population was likely unable to reach the dose of vitamin C utilized in experiments *in vivo* and *in vitro*, which led to the fact that supplementary vitamin C intake failed to reduce the risk of cancer in the general population.

The current study had several advantages and disadvantages. A major strength of this study was the MR study design, which could diminish confounding and reverse causality. Second, in this study, we broadly assessed the causal relationship of plasma vitamin C concentrations with the overall and a wide range of different types of cancer with a large number of cancer cases. Third, for each type of cancer, we validated our results in at least two datasets, which improved the robustness of our findings. However, there were also several limitations to the present study. First, the sample sizes of several types of cancer cases were small, resulting in low precision in the assessment. In that case, we might have ignored some weak associations. To deal with the problem, for those MR results generated from GWASs with small sample sizes, we validated the findings using another GWAS with a larger sample size. It should also be noted that the analyses are limited by the potential of the GWAS studies from which the IVs have been identified. In addition, in our study, the IVs were extracted from the largest GWAS study of vitamin C, and the F-statistics for the IVs were over 10, which could reduce the potential weak instrument bias. Second, our analyses were based on GWAS of European ancestry, and the results may be different in different ancestries; hence, our results might not be generalizable to all populations. Third, our study could only determine the causal relationship between circulating vitamin C levels and cancer risk but did not investigate the therapeutic effect of vitamin C on cancer.

CONCLUSION

This MR study did not support the causal association between vitamin C and the risk of overall or any specific types of cancer. Although previous observational studies and experiments confirmed an anticancer effect of vitamin C, these results might be influenced by confounding factors and were unable to illustrate the actual connection between vitamin C and cancer. Therefore, further studies are warranted to explore the relationship between vitamin C and the risk of cancer.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: Summary statistics of GWAS on overall cancer and nine site-specific cancers, including lung, breast, colon, rectum, kidney, bladder, prostate, ovarian, and uterine/endometrial cancer, were obtained from the United Kingdom Biobank dataset upon application (https://www.ukbiobank.ac.uk/). GWAS summary-level data on overall cancer and malignant neoplasm of the bronchus and lung, breast, pancreas, colon, rectum, kidney, bladder, prostate, ovary, and corpus uteri from the FinnGen consortium are available at https://finngen.gitbook.io/documentation/. GWAS summary-level data on lung cancer, breast cancer, pancreatic cancer, prostate cancer, and ovarian cancer were obtained from ILCCO, BCAC, PanScan1, PRACTICAL, and OCAC *via* https://gwas.mrcieu.ac.uk/datasets/, respectively.

AUTHOR CONTRIBUTIONS

Conceptualization: LQ and HW; data curation: HC, ZD, and YZ; MR analysis: HC; funding acquisition: LQ and HW; software and visualization: YZ, ML and RG; writing—original draft: HC and ZD; writing—review and editing: YZ, LQ, and HW. HC and ZD have verified the underlying data.

FUNDING

This work was supported by the Sichuan Science and Technology Program (2020YFS0127 and 2021YFS0126).

ACKNOWLEDGMENTS

Summary-level data for cancer were obtained from the UK Biobank, FinnGen Biobank, and genetic consortia such as Breast Cancer Association Consortium (BCAC), International Lung Cancer Consortium (ILCCO), Pancreatic Cancer Cohort Consortium (PanScan1), Ovarian Cancer Association Consortium (OCAC), and the Prostate Cancer Association group to Investigate Cancer-Associated Alterations in the Genome (PRACTICAL). The authors thank all investigators for sharing these data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.868408/full#supplementary-material

Supplementary Figure S1 | Scatter plot of vitamin C on overall cancer. **(A)** Scatter plot of vitamin C on overall cancer in UK Biobank. **(B)** Scatter plot of vitamin C on overall cancer in FinnGen Biobank.

Supplementary Figure S2 | Scatter plot of vitamin C on bronchus and lung cancer. **(A)** Scatter plot of vitamin C on bronchus and lung cancer in UK Biobank. **(B)** Scatter plot of vitamin C on bronchus and lung cancer in FinnGen Biobank. **(C)** Scatter plot of vitamin C on lung cancer in ILCCO. ILCCO, International Lung Cancer Consortium.

Supplementary Figure S3 | Scatter plot of vitamin C on breast cancer. **(A)** Scatter plot of vitamin C on breast cancer in UK Biobank. **(B)** Scatter plot of vitamin C on breast cancer in FinnGen Biobank. **(C)** Scatter plot of vitamin C on breast cancer in BCAC. BCAC, Breast Cancer Association Consortium.

Supplementary Figure S4 | Scatter plot of vitamin C on pancreatic cancer. **(A)** Scatter plot of vitamin C on pancreatic cancer in PanScan1. **(B)** Scatter plot of vitamin C on pancreatic cancer in FinnGen Biobank. PanScan1, Pancreatic Cancer Cohort Consortium GWAS.

Supplementary Figure S5 | Scatter plot of vitamin C on colon cancer. **(A)** Scatter plot of vitamin C on colon cancer in UK Biobank. **(B)** Scatter plot of vitamin C on colon cancer in FinnGen Biobank.

Supplementary Figure S6 | Scatter plot of vitamin C on rectal cancer. **(A)** Scatter plot of vitamin C on rectal cancer in UK Biobank. **(B)** Scatter plot of vitamin C on rectal cancer in FinnGen Biobank.

Supplementary Figure 7 | Scatter plot of vitamin C on kidney cancer. **(A)** Scatter plot of vitamin C on kidney cancer in UK Biobank. **(B)** Scatter plot of vitamin C on kidney cancer in FinnGen Biobank.

Supplementary Figure S8 | Scatter plot of vitamin C on bladder cancer. **(A)** Scatter plot of vitamin C on bladder cancer in UK Biobank. **(B)** Scatter plot of vitamin C on bladder cancer in FinnGen Biobank.

Supplementary Figure S9 | Scatter plot of vitamin C on prostate cancer. **(A)** Scatter plot of vitamin C on prostate cancer in UK Biobank. **(B)** Scatter plot of vitamin C on prostate cancer in FinnGen Biobank. **(C)** Scatter plot of vitamin C on prostate cancer in PRACTICAL. PRACTICAL, Prostate Cancer Association group To Investigate Cancer-Associated Alterations in the Genome.

Supplementary Figure S10 | Scatter plot of vitamin C on ovarian cancer. **(A)** Scatter plot of vitamin C on ovarian cancer in UK Biobank. **(B)** Scatter plot of vitamin C on ovarian cancer in FinnGen Biobank. **(C)** Scatter plot of vitamin C on ovarian cancer in OCAC. OCAC, Ovarian Cancer Association Consortium.

Supplementary Figure S11 | Scatter plot of vitamin C on endometrial cancer. **(A)** Scatter plot of vitamin C on endometrial cancer in UK Biobank. **(B)** Scatter plot of vitamin C on endometrial cancer in FinnGen Biobank.

REFERENCES

Amundadottir, L., Kraft, P., Stolzenberg-Solomon, R. Z., Fuchs, C. S., Petersen, G. M., Arslan, A. A., et al. (2009). Genome-Wide Association Study Identifies Variants in the Abo Locus Associated with Susceptibility to Pancreatic Cancer. *Nat. Genet.* 41, 986–990. doi:10.1038/ng.429

Ashor, A., Werner, A., Lara, J., Willis, N., Mathers, J., and Siervo, M. (2017). Effects of Vitamin C Supplementation on Glycaemic Control: A Systematic Review and Meta-Analysis of Randomised Controlled Trials. *Eur. J. Clin. Nutr.* 71, 1371–1380. doi:10.1038/ejcn.2017.24

Aune, D., Keum, N., Giovannucci, E., Fadnes, L. T., Boffetta, P., Greenwood, D. C., et al. (2018). Dietary Intake and Blood Concentrations of Antioxidants and the Risk of Cardiovascular Disease, Total Cancer, and All-Cause Mortality: A Systematic Review and Dose-Response Meta-Analysis of Prospective Studies. *Am. J. Clin. Nutr.* 108, 1069–1091. doi:10.1093/ajcn/nqy097

Bandera, E., Gifkins, D., Moore, D., Mccullough, M., and Kushi, L. (2009). Antioxidant Vitamins and the Risk of Endometrial Cancer: A Dose-Response Meta-Analysis. *Cancer Causes & Control : Ccc* 20, 699–711. doi:10.1007/s10552-008-9283-x

Bo, Y., Lu, Y., Zhao, Y., Zhao, E., Yuan, L., Lu, W., et al. (2016). Association between Dietary Vitamin C Intake and Risk of Esophageal Cancer: A Dose-Response Meta-Analysis. *Int. J. Cancer* 138, 1843–1850. doi:10.1002/ijc.29838

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian Randomization with Invalid Instruments: Effect Estimation and Bias Detection through Egger Regression. *Int. J. Epidemiol.* 44, 512–525. doi:10.1093/ije/dyv080

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using A Weighted Median Estimator. *Genet. Epidemiol.* 40, 304–314. doi:10.1002/gepi.21965

Cao, D., Shen, K., Li, Z., Xu, Y., and Wu, D. (2016). Association between Vitamin C Intake and the Risk of Cervical Neoplasia: A Meta-Analysis. *Nutr. Cancer* 68, 48–57. doi:10.1080/01635581.2016.1115101

Chen, F., Li, Q., Yu, Y., Yang, W., Shi, F., and Qu, Y. (2015). Association of Vitamin C, Vitamin D, Vitamin E and Risk of Bladder Cancer: A Dose-Response Meta-Analysis. *Scientific Rep.* 5, 9599. doi:10.1038/srep09599

Cimmino, L., Dolgalev, I., Wang, Y., Yoshimi, A., Martin, G., Wang, J., et al. (2017). Restoration of Tet2 Function Blocks Aberrant Self-Renewal and Leukemia Progression. *Cell* 170, 1079–1095. E20. doi:10.1016/j.cell.2017.07.032

Consortium, I. A. (2011). Design and Cohort Description of the Interact Project: An Examination of the Interaction of Genetic and Lifestyle Factors on the Incidence of Type 2 Diabetes in the Epic Study. *Diabetologia* 54, 2272.

Cortés-Jofré, M., Rueda, J., Asenjo-Lobos, C., Madrid, E., and Bonfill Cosp, X. (2020). Drugs for Preventing Lung Cancer in Healthy People. *Cochrane Database Syst. Rev.* 3, Cd002141. doi:10.1002/14651858.CD002141.pub3

Danesh, J., Saracci, R., Berglund, G., Feskens, E., Overvad, K., Panico, S., et al. (2007). The Cardiovascular Component of A Prospective Study of Nutritional, Lifestyle and Biological Factors in 520,000 Middle-Aged Participants from 10 European Countries. *Eur. J. Epidemiol.* 22, 129–141. doi:10.1007/s10654-006-9096-8

Day, N. E., Oakes, S. A., Luben, R. N., Khaw, K. T., and Wareham, N. J. (1999). Epic-Norfolk: Study Design and Characteristics of the Cohort. *Br. J. Cancer* 80 (Suppl. 1), 95–103.

Fu, Y., Xu, F., Jiang, L., Miao, Z., Liang, X., Yang, J., et al. (2021). Circulating Vitamin C Concentration and Risk of Cancers: A Mendelian Randomization Study. *Bmc Med.* 19, 171. doi:10.1186/s12916-021-02041-1

Fulan, H., Changxing, J., Baina, W., Wencui, Z., Chunqing, L., Fan, W., et al. (2011). Retinol, Vitamins A, C, and E and Breast Cancer Risk: A Meta-Analysis and Meta-Regression. *Cancer Causes & Control : Ccc* 22, 1383–1396. doi:10.1007/s10552-011-9811-y

Hartwig, F. P., Davey Smith, G., and Bowden, J. (2017). Robust Inference in Summary Data Mendelian Randomization via the Zero Modal Pleiotropy Assumption. *Int. J. Epidemiol.* 46, 1985–1998. doi:10.1093/ije/dyx102

Hua, Y., Wang, G., Jiang, W., Huang, J., Chen, G., and Lu, C. (2016). Vitamin C Intake and Pancreatic Cancer Risk: A Meta-Analysis of Published Case-Control and Cohort Studies. *Plos One* 11, E0148816. doi:10.1371/journal.pone.0148816

Islami, F., Siegel, R. L., and Jemal, A. (2020). The Changing Landscape of Cancer in the Usa - Opportunities for Advancing Prevention and Treatment. *Nat. Rev. Clin. Oncol.* 17, 631–649. doi:10.1038/s41571-020-0378-y

Jenkins, D. J. A., Spence, J. D., Giovannucci, E. L., Kim, Y. I., Josse, R. G., Vieth, R., et al. (2021). Supplemental Vitamins and Minerals for Cardiovascular Disease Prevention and Treatment: Jacc Focus Seminar. *J. Am. Coll. Cardiol.* 77, 423–436. doi:10.1016/j.jacc.2020.09.619

Jia, L., Jia, Q., Shang, Y., Dong, X., and Li, L. (2015a). Vitamin C Intake and Risk of Renal Cell Carcinoma: A Meta-Analysis. *Scientific Rep.* 5, 17921. doi:10.1038/srep17921

Jia, L., Jia, Q., Shang, Y., Dong, X., and Li, L. (2015b). Vitamin C Intake and Risk of Renal Cell Carcinoma: A Meta-Analysis. *Sci. Rep.* 5, 17921. doi:10.1038/srep17921

Jiang, L., Yang, K., Tian, J., Guan, Q., Yao, N., Cao, N., et al. (2010). Efficacy of Antioxidant Vitamins and Selenium Supplement in Prostate Cancer Prevention: A Meta-Analysis of Randomized Controlled Trials. *Nutr. Cancer* 62, 719–727. doi:10.1080/01635581.2010.494335

Lane, D., and Richardson, D. (2014). The Active Role of Vitamin C in Mammalian Iron Metabolism: Much More Than Just Enhanced Iron Absorption. *Free Radic. Biol. Med.* 75, 69–83. doi:10.1016/j.freeradbiomed.2014.07.007

Larsson, S. C., Carter, P., Kar, S., Vithayathil, M., Mason, A. M., Michaëlsson, K., et al. (2020). Smoking, Alcohol Consumption, and Cancer: A Mendelian Randomisation Study in Uk Biobank and International Genetic Consortia Participants. *Plos Med.* 17, E1003178. doi:10.1371/journal.pmed.1003178

Little, M. (2018). Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation. *J. R. Stat. Soc.* 181, 549–550. doi:10.1111/rssa.12343

Long, Y., Fei, H., Xu, S., Wen, J., Ye, L., and Su, Z. (2020). Association about Dietary Vitamin C Intake on the Risk of Ovarian Cancer: A Meta-Analysis. *Biosci. Rep.* 40. doi:10.1042/BSR20192385

Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., et al. (2017). Association Analysis Identifies 65 New Breast Cancer Risk Loci. *Nature* 551, 92–94. doi:10.1038/nature24284

Ngo, B., Van Riper, J. M., Cantley, L. C., and Yun, J. (2019). Targeting Cancer Vulnerabilities with High-Dose Vitamin C. *Nat. Rev. Cancer* 19, 271–282. doi:10.1038/s41568-019-0135-7

Padayatty, S., and Levine, M. (2016). Vitamin C: The Known and the Unknown and Goldilocks. *Oral Dis.* 22, 463–493. doi:10.1111/odi.12446

Park, Y., Spiegelman, D., Hunter, D., Albanes, D., Bergkvist, L., Buring, J., et al. (2010). Intakes of Vitamins A, C, and E and Use of Multiple Vitamin Supplements and Risk of Colon Cancer: A Pooled Analysis of Prospective Cohort Studies. *Cancer Causes & Control : Ccc* 21, 1745–1757. doi:10.1007/s10552-010-9549-y

Phelan, C. M., Kuchenbaecker, K. B., Tyrer, J. P., Kar, S. P., Lawrenson, K., Winham, S. J., et al. (2017). Identification of 12 New Susceptibility Loci for Different Histotypes of Epithelial Ovarian Cancer. *Nat. Genet.* 49, 680–691. doi:10.1038/ng.3826

Reczek, C., and Chandel, N. J. S. (2015). Cancer. *Revisiting Vitamin C And Cancer* 350, 1317–1318. doi:10.1126/science.aad8671

Schumacher, F. R., Al Olama, A. A., Berndt, S. I., Benlloch, S., Ahmed, M., Saunders, E. J., et al. (2018). Association Analyses of More Than 140,000 Men Identify 63 New Prostate Cancer Susceptibility Loci. *Nat. Genet.* 50, 928–936. doi:10.1038/s41588-018-0142-8

van Gorkom, G. N. Y., Lookermans, E. L., Van Elssen, C., and Bos, G. M. J. (2019). The Effect of Vitamin C (Ascorbic Acid) in the Treatment of Patients with Cancer: A Systematic Review. *Nutrients* 11. doi:10.3390/nu11050977

Verbanck, M., Chen, C. Y., Neale, B., and Do, R. (2018). Detection of Widespread Horizontal Pleiotropy in Causal Relationships Inferred from Mendelian Randomization between Complex Traits and Diseases. *Nat. Genet.* 50, 693–698. doi:10.1038/s41588-018-0099-7

Wang, Y., Mckay, J. D., Rafnar, T., Wang, Z., Timofeeva, M. N., Broderick, P., et al. (2014). Rare Variants of Large Effect in Brca2 and Chek2 Affect Risk of Lung Cancer. *Nat. Genet.* 46, 736–741. doi:10.1038/ng.3002

Yuan, S., Kar, S., Carter, P., Vithayathil, M., Mason, A. M., Burgess, S., et al. (2020). Is Type 2 Diabetes Causally Associated with Cancer Risk? Evidence from A Two-Sample Mendelian Randomization Study. *Diabetes* 69, 1588–1596. doi:10.2337/db20-0084

Zheng, J.-S., Luan, J. A., Sofianopoulou, E., Imamura, F., Stewart, I. D., Day, F. R., et al. (2021). Plasma Vitamin C and Type 2 Diabetes: Genome-wide Association Study and Mendelian Randomization Analysis in European Populations. *Diabetes Care* 44, 98–106. doi:10.2337/dc20-1328

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 24 June 2022
doi: 10.3389/fgene.2022.835713

Dissecting Generalizability and Actionability of Disease-Associated Genes From 20 Worldwide Ethnolinguistic Cultural Groups

Emile R. Chimusa[1,2]*, *Shatha Alosaimi*[1] *and Christian D. Bope*[1,2,3,4]

[1]Division of Human Genetics, Department of Pathology, University of Cape Town, Medical School Cape Town, Cape Town, South Africa, [2]Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa, [3]Department of Mathematics and Computer Science, University of Kinshasa, Kinshasa, Congo, [4]Centre for Bioinformatics, Department of Informatics, University of Oslo, Oslo, Norway

OPEN ACCESS

Edited by:
Guiyou Liu,
Chinese Academy of Sciences (CAS), China

Reviewed by:
Segun Fatumo,
University of London, United Kingdom
Juan Carlos Fernandez-Lopez,
Instituto Nacional de Medicina Genómica (INMEGEN), Mexico

****Correspondence:***
Emile R. Chimusa
emile.chimusa@uct.ac.za

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *12 January 2022*
Accepted: *29 April 2022*
Published: *24 June 2022*

Citation:
Chimusa ER, Alosaimi S and Bope CD (2022) Dissecting Generalizability and Actionability of Disease-Associated Genes From 20 Worldwide Ethnolinguistic Cultural Groups. Front. Genet. 13:835713. doi: 10.3389/fgene.2022.835713

Findings resulting from whole-genome sequencing (WGS) have markedly increased due to the massive evolvement of sequencing methods and have led to further investigations such as clinical actionability of genes, as documented by the American College of Medical Genetics and Genomics (ACMG). ACMG's actionable genes (ACGs) may not necessarily be clinically actionable across all populations worldwide. It is critical to examine the actionability of these genes in different populations. Here, we have leveraged a combined WES from the African Genome Variation and 1000 Genomes Project to examine the generalizability of ACG and potential actionable genes from four diseases: high-burden malaria, TB, HIV/AIDS, and sickle cell disease. Our results suggest that ethnolinguistic cultural groups from Africa, particularly Bantu and Khoesan, have high genetic diversity, high proportion of derived alleles at low minor allele frequency (0.0–0.1), and the highest proportion of pathogenic variants within HIV, TB, malaria, and sickle cell diseases. In contrast, ethnolinguistic cultural groups from the non-Africa continent, including Latin American, Afro-related, and European-related groups, have a high proportion of pathogenic variants within ACG than most of the ethnolinguistic cultural groups from Africa. Overall, our results show high genetic diversity in the present actionable and known disease-associated genes of four African high-burden diseases, suggesting the limitation of transferability or generalizability of ACG. This supports the use of personalized medicine as beneficial to the worldwide population as well as actionable gene list recommendation to further foster equitable global healthcare. The results point out the bias in the knowledge about the frequency distribution of these phenotypes and genetic variants associated with some diseases, especially in African and African ancestry populations.

Keywords: actionable gene, incidental finding, whole-genome sequencing, next-generation sequencing, genetic diversity, population genetics actionable gene, population genetics

INTRODUCTION

NGS analysis contributed to the improvement of patient treatment and clinical care. This development has bridged the gap between healthcare and genomics. Furthermore, variant calling is an important aspect of genomics studies as polymorphism information can be used to influence the discovery of actionable pathogenic variants and therefore impact important clinical decisions. Currently, the definition of actionable pathogenic variants varies among scholars (Bope et al., 2019).

The Clinical Genome Resource (ClinGen) presents actionability as clinically prescribed interventions to a genetic disorder that is effective for prevention, lowered clinical burden or delay for a clinical disease, or improved clinical treatments and outcomes in a previously undiagnosed adult (Hunter et al., 2016). On the other hand, the 100,000 Genomes Project protocol presents actionable genes as variants that can significantly prevent (or result in illness or disability that is clinically significant, severely life-threatening, and clinically actionable) disease morbidity and mortality, if identified before symptoms become apparent. However, in any case, the classification of variants to be clinically actionable or not dependent and can only emerge during the process of seeking ethical approval for the study (Hunter et al., 2016).

Overall, in the current literature and most annotation databases, the classification of pathogenicity differs (Sherry et al., 2001; Wang et al., 2010; Landrum et al., 2016; McLaren et al., 2016). Dorschner et al. (1016) leveraged exome data of European and African populations to dissect actionable pathogenic variants, and the result shows that actionable pathogenic variants were disproportionate between European and African populations with an estimated frequency of approximately 3.4 and 1.2%, respectively. This indicates a deficit in the identification or categorization of pathogenic variants in African populations. A similar study conducted by Amendola et al. (2015) also confirmed the findings of Dorschner et al. (1016). One approach to define actionability is to combine many annotation pipelines during filtering and prioritization of mutations, in which casting vote can be applied respectively to allow better prediction of the targeted variant (Lebeko et al., 2017; Bope et al., 2019). Furthermore, on top of ethical approval, the ancestral/derived minor allele frequency of the variants, segregation evidence, and the number of patients affected with the variants and their status as a *de novo* mutation can highly be considered.

In this study, we provide a broad assessment of the possible actionability of variants known to be associated with the top four burden African diseases and a list of actionable genes from the American College of Medical Genetics and Genomics (ACMG) using WGS data of 20 worldwide ethnolinguistic cultural groups. This work aims to 1) perform variant join calling on publicly available data from the African Genome Variation and the 1000 Genomes Project to examine the evolutionary variation of pathogenic mutation; 2) perform disease-gene population structure; and 3) examine the heterozygosity ratio, the proportion of ancestral/derived alleles, and the distribution of minor allele frequencies based on selected known disease-genes from four predominant African burden diseases, HIV/AIDS, malaria, TB, and sickle cell disease, and a set of known actionable genes across 20 worldwide ethnolinguistic cultural groups. These diseases have uniquely shaped ethnolinguistic culturally specific groups and continental-specific genomic variations, and therefore offer unprecedented opportunities to map disease genes.

Our results in support with previous findings indicate higher genetic diversity in ethnolinguistic cultural groups from Africa, based on four African burden diseases and associated actionable genes. The results suggest the limitation of transferability or generalizability and support the use of personalized medicine as beneficial to each worldwide population or ethnolinguistic cultural group. In addition, our results point out the bias in the knowledge about the frequency distribution of these phenotypes and genetic variants associated with some diseases, especially in African and African ancestry populations, suggesting further examination of actionable gene lists to improve equitable global healthcare.

TABLE 1 | Number of SNPs after quality control (QC) in each group of genes associated with HIV, TB, SCD, malaria, and actionable genes.

SNP	Gene	Disease
649,078	114	ACG
2,735,797	460	HIV
265,427	50	Malaria
4,455,648	75	SCD
2,513,341	77	TB

RESULTS

Based on the initial sample description of populations and country labels and leveraging the population culture and ethnolinguistic information (Gudykunst and Schmidt, 1987; Michalopoulos, 2012), we grouped 4,932 samples from their country labels into 20 independent ethnolinguistic cultural groups (**Supplementary Table S1**) and performed an independent joint call (see Materials and Methods), resulting in 90, 641, and 235 curated polymorphisms. We leveraged the dbSNPS database in extracting SNPs associated with 77, 50, 75, 460, and 114 genes known to be associated with tuberculosis, malaria, sickle cell disease, HIV, and ACMG's actionable genes, respectively (**Table 1**), to examine the generalizability and actionability of these disease-associated genes from 20 worldwide ethnolinguistic cultural groups.

Disease and Actionable Gene-Specific Population Structure

To better characterize the genetic relatedness, we first conducted principal component analysis (PCA) on whole-genome SNPs across all these 20 ethnolinguistic cultural groups (**Supplementary Figure S1**). Regardless of ethnolinguistic

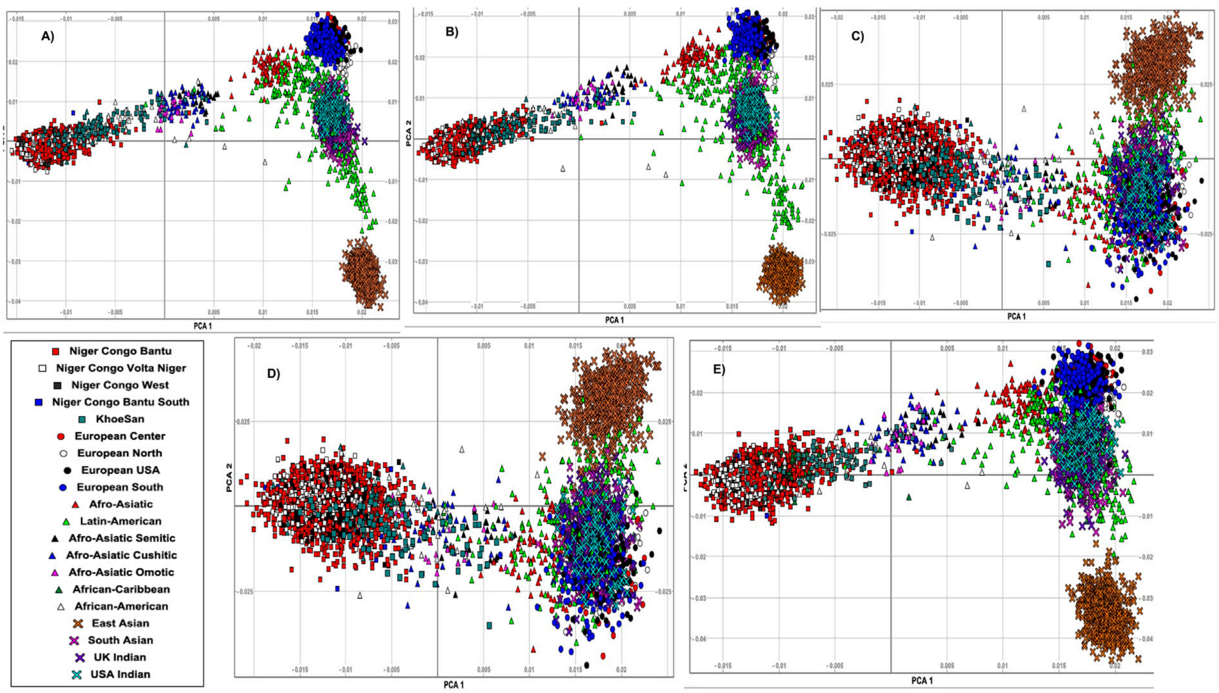


FIGURE 1 | Principal component analysis (PCA) of genes associated with **(A)** HIV-specific, **(B)** TB-specific, **(C)** malaria-specific, **(D)** sickle cell disease-specific, and **(E)** ACG-specific SNPs and plots of the first and second eigenvectors for 20 ethnolinguistic cultural groups.

cultural groups, the results in **Supplementary Figure S1** show a clear separation between African, European, Indian, and Eastern Asian groups. Second, based on the extracted disease-specific SNPs of different diseases, among these 20 different worldwide ethnolinguistic cultural groups (Materials and Methods), we performed principal component analysis (PCA). This PCA produces a set of orthogonal axes for which the remaining variances in the data are maximized by each successive dimension. **Supplementary Table S2** illustrates the genetics distance (Fst) based on disease-specific variants among these 20 ethnolinguistic cultural groups. We present our gene-specific population structure results for HIV (**Figure 1A**), TB (**Figure 1B**), malaria (**Figure 1C**), sickle cell anemia (**Figure 1D**), and ACG (**Figure 1E**). Our results show that HIV variation is observed among Bantu, African–American, Khoesan, and Afro-related ethnolinguistic cultural groups, while the European group is clustering together (**Figure 1A**). Most ethnolinguistic cultural groups from Africa have the highest HIV gene-specific frequency (**Figure 1A**), confirming that HIV infection has a high incidence or prevalence among ethnolinguistic cultural groups from Africa compared to other ethnolinguistic cultural groups. Moreover, a variation in HIV-specific genes shows little overlap between/within ethnolinguistic cultural groups. The first principal component (PC) separates the European-related ethnolinguistic cultural group cluster and the African-related ethnolinguistic cultural cluster from one end to the other with the Afro-Asiatic ethnolinguistic cultural groups, the African–American, and one part of the Latin-Americans in the middle. The second principal component separates the European-related ethnolinguistic cultural cluster and the East Asian ethnolinguistic cultural group from one end to the other with the United Kingdom/United States–Indian group, the South Asian, and one part of the Latin-American ethnolinguistic cultural group in the middle. We also observe a cline between each axis. The dispersion of samples of HIV-specific genes along the lines suggests the existence of an admixture which may have occurred between ethnolinguistic cultural groups located on the same line and added to a strong local adaptation of HIV-specific genes among ethnolinguistic cultural groups located in the middle of each cline. One interesting observation is the intersection of the Latin-American ethnolinguistic cultural group with the Afro-Asiatic ethnolinguistic cultural groups on one side and the United Kingdom/United States–Indian and South Asian ethnolinguistic cultural groups on the other side which may indicate either a possible existence of HIV-specific actionable genes overlapping between these mentioned populations or a differing effect of these genes across these ethnolinguistic cultural groups. As for HIV, a variation in TB-specific genes was observed among Bantu and Khoesan and Afro-related ethnolinguistic cultural groups (**Figure 1B**), while European groups are clustering together, except in North European (explaining the known high incidence of TB in Central and North Europe). As the same observation for TB is similar to HIV, then the same comment applies for TB as well. Malaria-specific worldwide ethnolinguistic cultural groups' genetic structure (**Figure 1C**) shows that ethnolinguistic cultural groups from Africa and African–American ethnolinguistic cultures are still separated from the rest of the

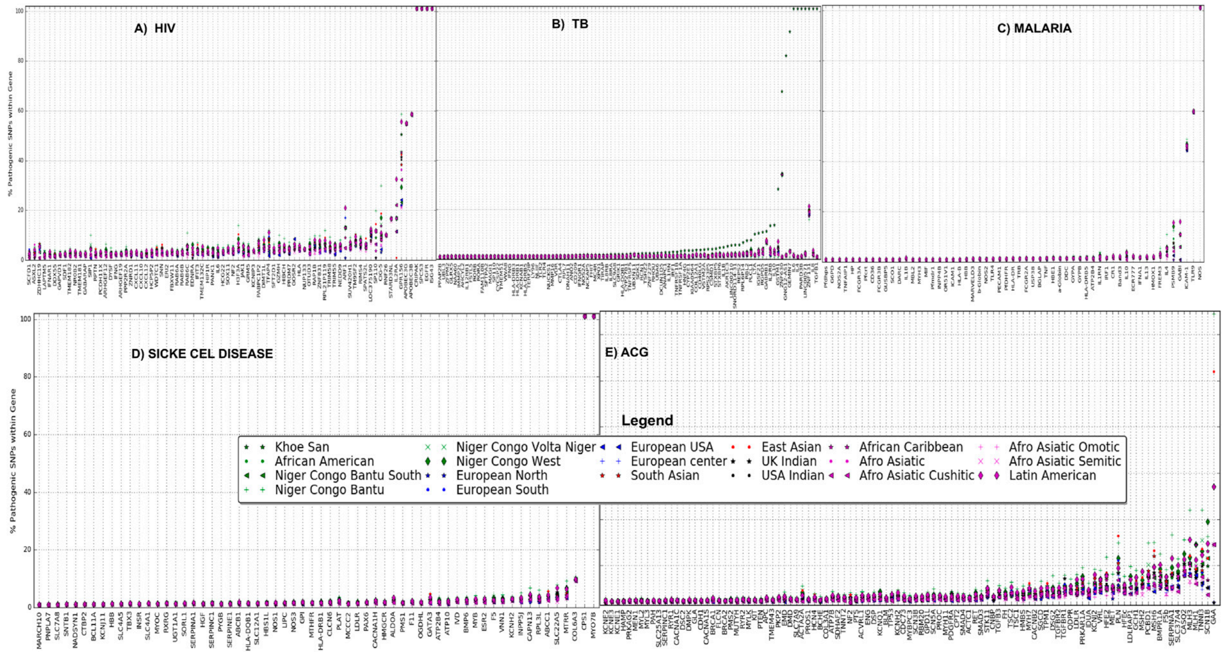


FIGURE 2 | Proportion of pathogenic variants within **(A)** HIV-specific, **(B)** TB-specific, **(C)** malaria-specific, **(D)** sickle cell disease-specific, and **(E)** ACG-specific (actionable genes) genes among all 20 ethnolinguistic cultural groups.

other ethnolinguistic cultural groups. United Kingdom/ United States–Indians and Afro-related, Latin-American, and all Europeans are clustering together based on malaria-specific genes, low prevalence, and/or absence of malaria in their geographic regions, indicating that the malaria-specific genes found in one of these aforementioned populations may not be found in the other population. East/South Asians are clustering apart from ethnolinguistic cultural clusters from Africa and Europe continents. While it is known that malaria has a high prevalence among African and Asian populations, the separate cluster between them may indicate different patterns of linkage disequilibrium, geographic location, and genetic variation in malaria-specific genes. As expected, since malaria and sickle cell disease are known to be genetically correlated, similar results for Malaria are observed with sickle-cell disease-specific genes (**Figure 1D**). The population structure on ACG-specific genes reveals that Africa and European-related ethnolinguistic cultural groups, East-Asian ethnolinguistic cultural groups, and United Kingdom/United States–Indian and South Asian ethnolinguistic cultural groups are separated and clustered in three different clusters (**Figure 1E**). We observed that African–American and Afro-related ethnolinguistic cultural groups are in the convex of these three clusters (**Figure 1E**), justifying that they are the result of the admixture of these ethnolinguistic cultural groups considered geographic ancestral populations. In addition, Latin-America is close to European and South Asian clusters, as seen from the results of the admixture, and they are mainly in the convex between East-Asian, South-Asian, and European groups, and a bit distant to the ethnolinguistic cultural groups from Africa. This result indicates that the transferability or generatability of the actionability of these ACG genes may have differing effects across 20 worldwide ethnolinguistic cultural groups.

Proportion of Pathogenic Polymorphisms Within Disease-Associated Genes

Ethnolinguistic cultural groups from Africa including Bantu and Latin-American and Afro-related groups have a considerable high proportion of pathogenic variants in these HIV-specific genes (**Figure 2A**). We observe that the Khoesan ethnolinguistic cultural group has a high proportion of pathogenic variants within TB-specific genes (**Figure 2B**). Latin-American, Afro-Asiatic, and African ancestry (African diaspora)-related ethnolinguistic cultural groups have a high proportion of pathogenic variants (**Figure 2B**). The low proportion of pathogenic variants is observed across all malaria-specific genes in Bantu, Afro-Asiatic, and Latin-American ethnolinguistic cultural groups (**Figure 2C**); however, except for toll-like receptor 9 (*TLR9*), *FREM3*, *IL4*, *ICAM-1*, and nitric oxide synthase 1 (neuronal), the Bantu-related ethnolinguistic cultural groups and Latin-Americans have a high proportion of pathogenic variants (**Figure 2C**). Bantu, Afro-related ethnolinguistic cultural groups, and Latin America have a similar low proportion of pathogenic variants in most of the sickle cell disease-specific genes, except in *MY O 7B*, *CPS1*, *COL6A3*, *MTRR*, *SLC22A5*, *ABCC1*, and *RPL3L* (**Figure 2D**). We observed a considerable high proportion of pathogenic variants within ACG-specific genes from ethnolinguistic cultural groups out of the African continent

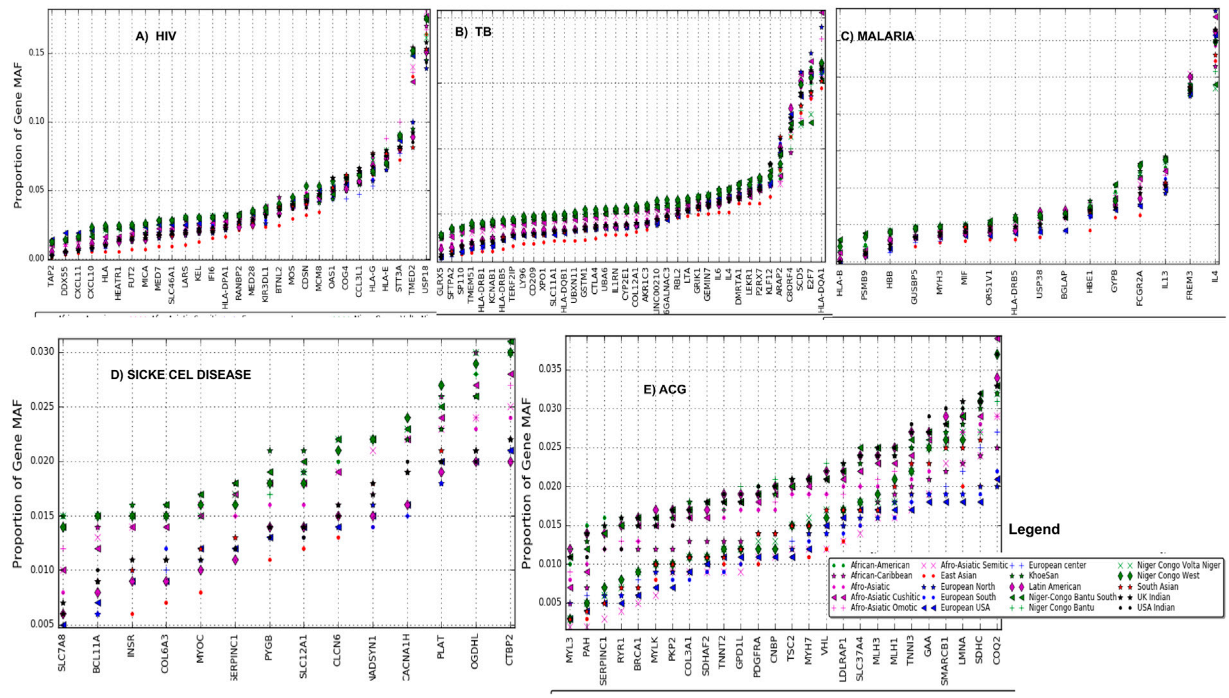


FIGURE 3 | Gene-specificity in SNP minor allele frequency: the distribution of the minor allele frequency at the gene level **(A)** HIV, **(B)** TB, **(C)** malaria, **(D)** sickle cell disease, and **(E)** ACG (actionable genes) among all ethnolinguistic cultural groups.

including Latin America, Afro-Asiatic, and European-related ethnolinguistic cultural groups (**Figure 2E**), while few genes show a high proportion of pathogenic variants in Niger-Bantu and African–American groups (**Figure 2E**).

Distribution of Gene-Specificity in SNP Frequencies

We observed variations in the distribution of MAF at rare variants within MAF bin 0.0–0.05 among these 20 ethnolinguistic cultural groups in four African burden diseases (**Supplementary Figures S2A–D**) and ACMG's actionable genes (**Supplementary Figure S2E**). *BTNL2*, *MOS*, *CDSN*, *USP18*, *MCM8*, *OAS1*, *COG4*, *CCL3L1*, *HLA-G*, *HLA-E*, *STT3A*, *TMED2*, and *USP18* have HIV gene-specificity in SNP frequencies ranging between 5 and 15% (**Figure 3A**) and those ethnolinguistic cultural groups from Africa have the highest. A total of 33 genes have TB gene-specificity in SNP frequencies between 5 and 20% of which all ethnolinguistic cultural groups from Africa have the highest (**Figure 3B**), suggesting that these genes may harbor common effects and contributions to TB among African ethnolinguistic cultural groups. The distribution of malaria gene-specificity in SNP frequencies from **Figure 3C** suggests that four genes include *GYPB*, *FCGR2A*, *IL13*, and *FREM3* with gene-specificity ranging between 4 and 15%, while all sickle cell disease-related genes (**Figure 3D**) show low gene-specificity in SNP frequencies ranging between 0.1 and 0.3% among all 20 ethnolinguistic cultural groups, but all ethnolinguistic cultural groups from Africa have the highest frequencies. The distribution of ACG-gene-specificity in SNP frequencies in **Figure 3E** indicates that all ACG genes have gene-specificity in SNP frequencies lower than 0.4% in all 20 ethnolinguistic cultural groups. However, the gene-specificity in SNP frequencies from most of the ethnolinguistic cultural groups from Africa are higher than those from non-African ethnolinguistic cultural groups, supporting a potential difference effect and contribution of these actionable genes among worldwide ethnolinguistic cultural groups. **Supplementary Table S3** shows the details of gene-specificity in SNP frequencies of these ACG and disease burdens across all these 20 ethnolinguistic cultural groups.

Gene-Specific in Proportion of Derived Alleles and Relationship Between Derived and Ethnolinguistic Cultural-Specific Minor Allele Frequency

Derived alleles are more often minor alleles (<50% allele frequency) and associated with risk than ancestral alleles (32). As for the variation observed in the distribution of MAF at rare variants at low ethnolinguistic and cultural-specific minor allele frequencies (ranging between 0.0 and 0.1, **Supplementary Figure S3**), high variation in the proportion of derived alleles can be observed in HIV (**Supplementary Figure S3A**), TB (**Supplementary Figure S3B**), malaria (**Supplementary Figure S3C**), and sickle cell disease (**Supplementary Figure S3D**), and a set of actionable genes (**Supplementary Figure S3E**) across all ethnolinguistic cultural groups from Africa compared to the rest

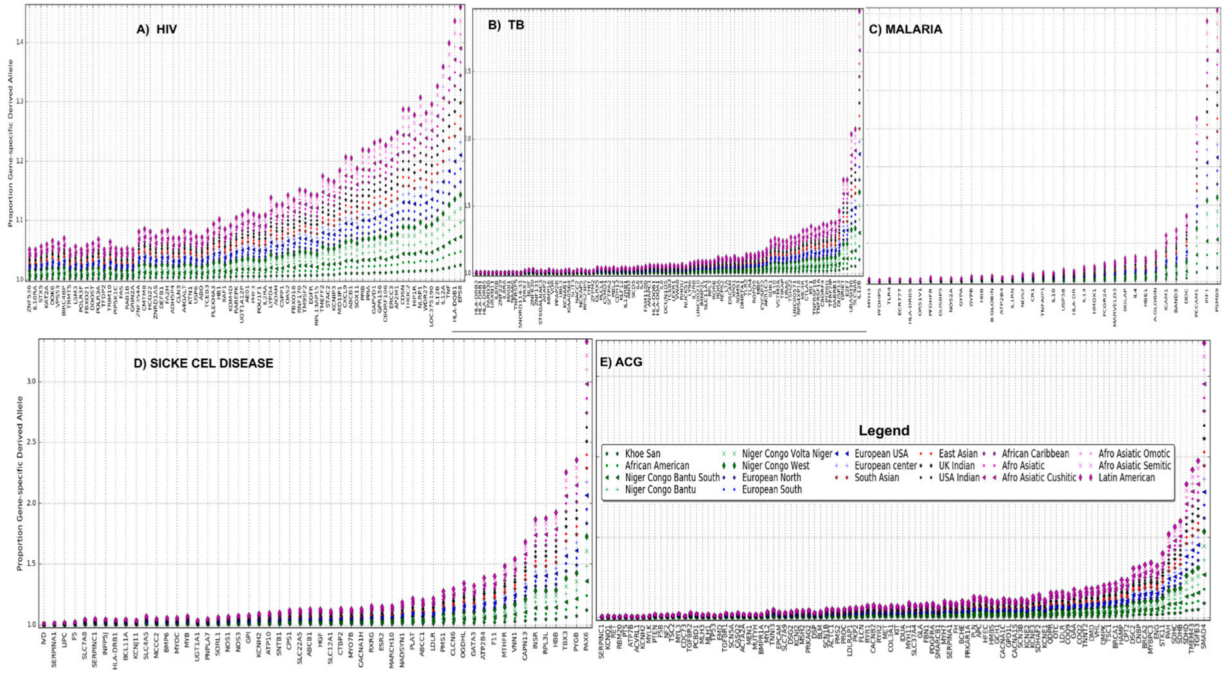


FIGURE 4 | (A) HIV, **(B)** TB, **(C)** malaria, **(D)** sickle cell disease, and **(E)** ACG gene-specific proportion of derived alleles across 20 worldwide ethnolinguistic cultural groups.

of the other ethnolinguistic cultural groups, and that most of the ethnolinguistic cultural groups from Africa have the highest proportion of derived alleles in the range of minor allele frequency bin (0.0–0.1) (**Supplementary Figure S3A**), indicating that different mutations and possible selections occurred in rare variants within genes associated with these four African burden diseases, and ACMG's actionable genes play critical roles and that ethnolinguistic and cultural-specific risk alleles may differentially contribute to the phenotypic variations and clinical outcomes.

To obtain gene-specific proportions of derived alleles, derived allele frequencies were aggregated for all SNPs associated with each of these disease-specific genes (see Materials and Methods). For all African burden diseases including HIV (**Figure 4A**), TB (**Figure 4B**), malaria (**Figure 4C**), and sickle cell diseases (**Figure 4D**), we observe that Latin America and most of Afro-Asiatic, Bantu, and Khoesan ethnolinguistic cultural groups have a considerable and consistently high proportion of gene-specific derived alleles. We observe a consistent high ACG-gene-specific allele in Latin America and most Afro-related ethnolinguistic cultural groups following most of European-related ethnolinguistic cultural groups (**Figure 4E**), while a low ACG-gene-specific allele is observed in most of African ethnolinguistic cultural groups. One can expect actionable genes to have a high proportion of derived alleles; however, this is not the case for most of African ethnolinguistic cultural groups, indicating that the current ACG genes were primarily tailored for non-African ethnolinguistic cultural groups. A full list of the ethnolinguistic and cultural gene-specific proportions of derived alleles based on genes associated with these four African burden diseases and ACMG's actionable genes can be found in **Supplementary Table S4**.

Genetic Diversity: Observed and Expected Heterozygosity

Gene diversity consists of two elements including the abundance (or evenness) of the alleles and the number of alleles. The abundance (or evenness) of the alleles and the number of alleles would increase the expected heterozygosity. If an ethnolinguistic cultural group consists of an excess of homozygotes for different alleles, this leads to low-observed heterozygosity. In **Figure 5**, we observe that ethnolinguistic cultural groups from Africa, particularly Bantus and Khoesan, have the highest gene diversity in HIV, TB, malaria, sickle cell disease, and ACG-associated variants (**Supplementary Table S5**). This result supports the highest genetic diversity found in individuals and communities across the African continent and that the use of personalized medicine will be beneficial to both the continent and world.

DISCUSSION

In this study, we conducted a joint call of 4,932 samples representing 20 worldwide ethnolinguistic cultural groups (**Supplementary Table S1**), to examine the generalizability and actionability of 77, 50, 75, 460, and 114 genes known to be associated with tuberculosis, malaria, sickle cell disease, HIV, and ACG, respectively (**Table 1**). To examine the generalizability

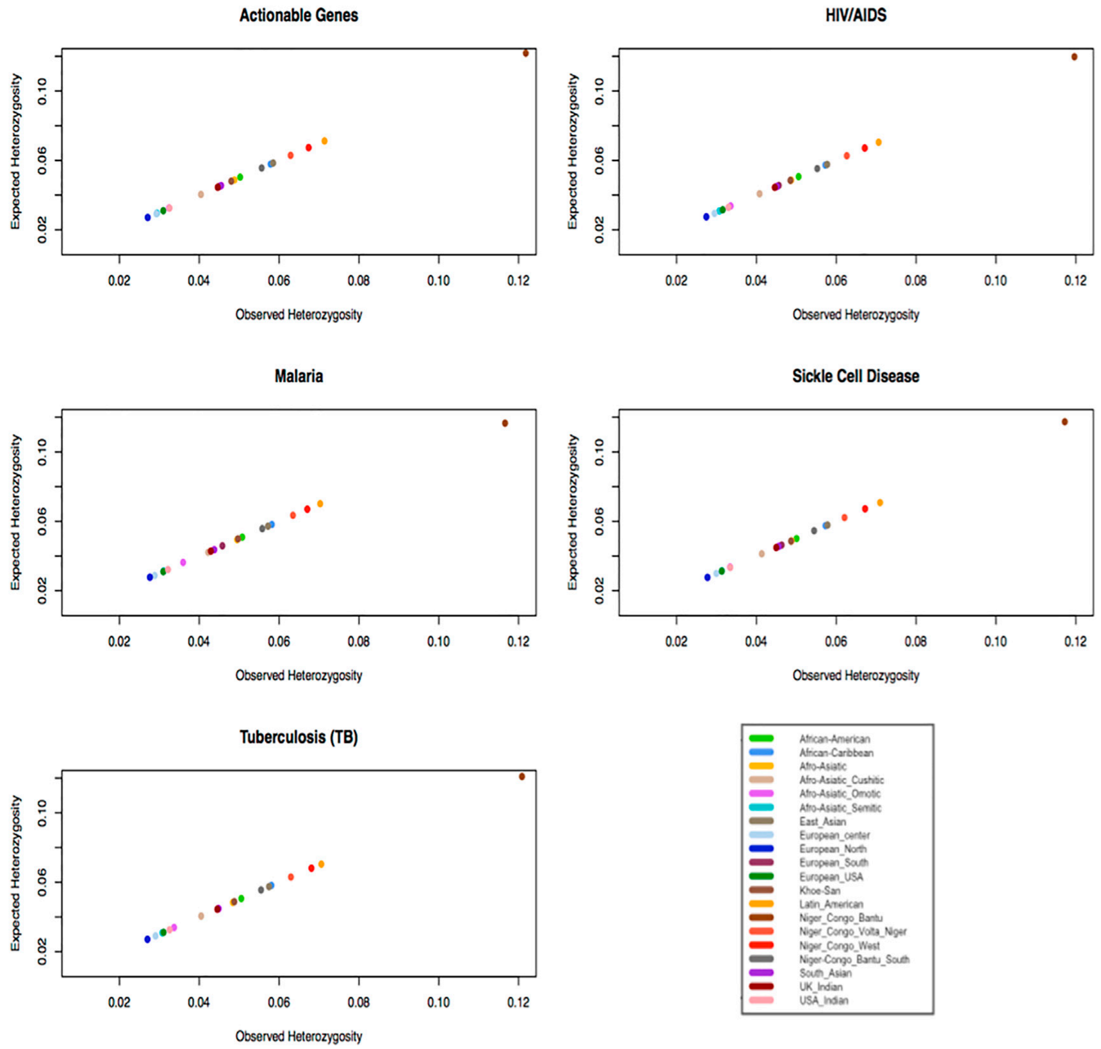


FIGURE 5 | Plot expected heterozygosity as a function of observed heterozygosity per gene of specific diseases within ethnolinguistic cultural groups.

and actionability of genes, we investigated the distribution of (Bope et al., 2019) gene-specificity in SNP frequencies, (Hunter et al., 2016), gene-specificity in the proportion of derived alleles, and (Sherry et al., 2001) gene-specificity in pathogenic mutations. In addition, population-specific genetic structures and expected heterozygosity were observed in all associated SNPs within genes.

The results of HIV/TB indicated that ethnolinguistic cultural groups including Bantu, Latin American, and Afro-Asiatic have the highest proportion of pathogenic variants based on 483 HIV-specific genes. From 77 TB-specific genes, we observed that Latin American and Afro-Asiatic ethnolinguistic cultural groups have the highest proportion of pathogenic variants, important among all African and African diaspora ethnolinguistic cultural groups, and only Khoesan has a high proportion of pathogenic variants within TB-specific genes. Most ethnolinguistic cultural groups from Africa (Bantu and Khoesan) have the highest HIV and TB gene-specific frequency, indicating that HIV disease risk is prevalent among African ethnolinguistic cultural groups compared with other ethnolinguistic cultural groups. Our result identifies *BTNL2*, *MOS*, *CDSN*, *USP18*, *MCM8*, *OAS1*, *COG4*, *CCL3L1*, *HLA-G*, *HLA-E*, *STT3A*, *TMED2*, and *USP18* to have HIV gene-specificity in SNP frequencies ranging between 5 and 15% and those ethnolinguistic cultural groups from Africa have the highest. In addition, 33 genes have TB gene-specificity in SNP frequencies ranging between 5 and 20% of which all African ethnolinguistic cultural groups have the highest frequencies. This suggests that these genes may harbor a common effect and contribution to TB/HIV among African ethnolinguistic cultural groups. Furthermore, HIV/TB gene-specificity has a high proportion of derived alleles at low minor allele frequency (0.0–0.1) from African ethnolinguistic cultural groups and that these proportions of derived alleles vary among African ethnolinguistic cultural groups, suggesting a possible challenge in enabling cross-population actionable gene transferability and possible implementation of precision medicine within different ethnolinguistic cultural groups from Africa.

The results of malaria and sickle cell disease indicate the absence of pathogenic variants in most of the European-related ethnolinguistic cultural groups and a low proportion of pathogenic variants across all malaria-specific genes in Bantu, Afro-Asiatic, and Latin American ethnolinguistic cultural groups, except for toll-like receptor 9 (*TLR9*), *FREM3*, *IL4*, *ICAM-1*, and nitric oxide synthase 1 (neuronal), indicates that Bantu and Latin America ethnolinguistic cultural groups have a high proportion of pathogenic variants. Furthermore, Bantu, Afro-Asiatic, and Latin American ethnolinguistic cultural groups have a similar low proportion of pathogenic variants in most sickle cell disease-specific genes, except in *MY O 7B*, *CPS1*, *COL6A3*, *MTRR*, *SLC22A5*, *ABCC1*, and *RPL3L*. We identify four genes including *GYPB, FCGR2A, IL13,* and *FREM3* with malaria gene-specificity in SNP frequencies ranging between 4 and 15%, while all sickle cell disease-related genes have low gene-specificity in SNP frequencies ranging between 0.1 and 0.3% among all 20 ethnolinguistic cultural groups, but all African and diaspora ethnolinguistic cultural groups have the highest in that range.

The result on ACG showed a considerably high proportion of pathogenic variants within ACG-specific genes from non-African ethnolinguistic cultural groups including Latin American, Afro-Asiatic, and European compared to most of African-related ethnolinguistic cultural groups. This result justifies and indicates that the actionability of these ACG genes may have heterogeneous effects on worldwide ethnolinguistic cultural groups, unraveling cross-ethnic group transferability and generalizability to diverse ethnic groups, particularly African from ACG-specific actionable genes daunting. Our result indicates that all ACG genes have gene-specificity in SNP frequencies lower than 0.4% in all 20 ethnolinguistic cultural groups. However, the gene-specificity in SNP frequencies from most of African ethnolinguistic cultural groups are higher than those from non-African ethnolinguistic cultural groups, supporting the potential common effect and contribution of these actionable genes to non-African ethnolinguistic cultural groups. A high ACG-gene-specific derived allele was observed in Latin-American and most Afro-related ethnolinguistic cultural groups following most of European-related ethnolinguistic cultural groups, while a low ACG--specific derived allele is observed in most of African ethnolinguistic cultural groups.

We leveraged the dbSNP database to extract SNPs associated with these genes per disease. The obtained SNPs per disease were thus extracted from the whole phased data containing 4,932 samples of these 20 ethnolinguistic cultural groups, yielding five disease-specific phased haplotype datasets. From these phased haplotype data, we conducted disease gene-specific population structure, and we examined the distribution and relationship of derived and minor allele frequency and estimated the expected and observed heterozygosity.

The result of this study suggests significant genetic variations among all non-European ethnolinguistic cultural groups, mostly African ethnolinguistic cultural groups, while all European ethnolinguistic cultural groups are genetically and consistently clustering together based on these diseases or actionable-specific variants, suggesting limitations of cross-population transferability of actionable or medically relevant genes, given the exceptional polygenicity of human traits. Furthermore, the result indicates that African and African diaspora ethnolinguistic cultural groups, particularly Bantus and Khoesan ethnolinguistic cultural groups, have the highest gene diversity in HIV, TB, malaria, sickle cell disease, and ACG-associated variants. This supports the highest genetic diversity found in individuals and communities across the African continent. Based on these findings, the use of personalized medicine including African genomics will be beneficial to both the continent and world. One of the limitations of this finding is that although these results depend greatly on laboratory experiments, the distribution of actionable genes across populations may depend on continuous genetic diversity, natural selection, and genetic drift. Such study paves the way for a continuous analysis of disease-specific actionable genes and their genetic mechanism underpinning those diseases.

CONCLUDING REMARKS

In conclusion, our findings suggest the highest genetic diversity in African ethnolinguistic cultural groups in the four African burden diseases and ACMG's actionable genes, and that the distribution of gene-specificity (Bope et al., 2019) in SNP frequencies (Hunter et al., 2016), in the proportion of derived alleles, and (Sherry et al., 2001) in pathogenic mutations based on the obtained 77, 50, 75, 460, and 114 genes was known to associate with tuberculosis, malaria, sickle cell disease, HIV, and ACMG's actionable genes, respectively, indicating significant variation across 20 worldwide ethnolinguistic cultural groups. This suggests (Bope et al., 2019) the limitation of transferability or generalizability; however, the use of personalized medicine will be beneficial to both the African continent and worldwide (Hunter et al., 2016), enabling a recommendation for an African-specific actionable list of genes which will further improve African and diaspora healthcare.

MATERIALS AND METHODS

Data Description and Quality Check

The data Binary Alignment Map (BAM) files were obtained from the 1000 Genomes Project (1KGP) (Siva, 2008) and the African Genome Variation Project (AGVP) (Gurdasani et al., 2015), which has recently characterized the admixture across 18 ethnolinguistic groups from sub-Saharan Africa as shown in **Supplementary Table S1**. A quality control check was conducted on the BAM files using SAMtools (Li et al., 2009). After quality check, a total of 2,504 BAM files from the 1000 Genomes Project and 2,428 BAM files from the AGVP were retained. Based on initial sample description population and country labels, we used the population culture and ethnolinguistic information (Gudykunst and Schmidt, 1987; Michalopoulos, 2012) to group populations from the country label into 20 ethnolinguistic cultural groups (**Supplementary Table S1**). **Supplementary Figure S1** illustrates the genetics relatedness and variation of these 20 ethnolinguistic cultural groups, supporting previous findings (Siva, 2008; Chimusa et al., 2015; Gurdasani et al., 2015; Choudhury et al., 2020), and Supplementary File 1 illustrates the genetics distance (Fst) based on disease-specific variants among the 20 ethnolinguistic cultural groups.

Variants Discovery Analysis and Annotation

LoFreq, a variant calling tool, was used to conduct joint calls across 4,932 samples in 20 worldwide ethnolinguistic cultural groups. The resulting variant sets of all 4,932 samples in the VCF file were filtered using SAMtools, and 4,932 samples remained and were considered for downstream analysis.

The resulting joint call VCF file of 4,932 samples and samples were split into 20 VCF files per ethnolinguistic cultural group as listed in **Supplementary Table S1**. The independent gene-based annotation for each VCF dataset to determine whether SNPs cause protein-coding change and produce a list of amino acids that are affected was conducted using ANNOVAR (Wang et al., 2010). The following setting was used in ANOVA: the population frequency and pathogenicity for each variant were obtained from 1000 Genomes exome, Exome Aggregation Consortium (ExAC), targeted exon datasets, and COSMIC. Gene functions were obtained from RefGene, and different functional predictions were obtained from ANNOVAR's library, which contains up to 21 different functional scores including SIFT (Ng et al., 2006), LRT (Schwarz et al., 2010), MutationTaster (Reva et al., 2011), MutationAssessor (Shihab et al., 2013), FATHMM and FATHMM-MKL (Liu et al., 2011), RadialSVM (Choi and Chan, 2015), LR (Kim et al., 2017), PROVEAN (Kim et al., 2017), MetaSVM (Dong et al., 2015), MetaLR (Rentzsch et al., 2018), CADD (Davydov et al., 2010), GERP++ (Quang et al., 2014), DANN (Jagadeesh et al., 2016), M-CAP (Ionita-Laza et al., 2016), Eigen (Lu et al., 2015), GenoCanyon (Adzhubei et al., 2010), Polyphen2-HVAR and HDIV (Doerks et al., 2002), PhyloP (Garber et al., 2009), and SiPhy (Loh et al., 2016a). In addition, conservative and segmental duplication sites were included, and the dbSNP code and clinical relevance were reported in dbSNP. From each resulting functional annotated dataset, we independently filtered for the predicted functional status, of which each predicted functional status is of "deleterious" (D), "probably damaging" (D), "disease-causing-automatic" (A), or "disease-causing" (D). The selection of mutations was carried out using the following approach: first, the casting vote approach was implemented in our custom Python script, to retain only a variant if it had at least 17 predicted functional status "D" or "A" out of 21 was used and second, the retained variants from each dataset were further filtered for rarity, exonic variants, and nonsynonymous mutations and with a high-quality call as described previously, yielding a final candidate list of predicted mutant variants in each subject group, including the replication group. We report on the aggregated SiPhy score from all identified mutant SNPs within the gene. The following sections provide details on how SNPs were mapped to genes.

Phased and Haplotypes Inference

To increase the accuracy, the resulting VCF file, containing 4,932 samples of 20 ethnolinguistic cultural groups, was used to further conduct quality control in removing all structured, indel, multi-allelic variants and those with a low minor allele frequency (MAF <0.05) prior to phasing. We first phased and inferred the haplotypes using Eagle (Loh et al., 2016b) from the resulting curated data. We further compared site discordances between these haplotype panels and independently with their original VCF file before phasing. The only site with phase switch-errors showed discrepancies in MAF and was removed.

Disease- and Actionable Gene-Specific Population Structure

We obtained the list of genes, known as medically actionable, and Actionable Genome Consortium (ACG) from https://www.coriell.org/1/NIGMS/Collections/ACMG-73-Genes. The list of genes associated with four major African diseases including malaria, TB, HIV, and sickle cell disease was collected from the GWAS Catalog (https://www.ebi.ac.uk/gwas/), and the extraction was based on phenotype classification and from databases such DisGeNET http://www.disgenet.org/and literature. We obtained 50, 77, 460, 75, and 114 genes known to be associated with tuberculosis, malaria, sickle cell anemia, HIV, and ACG, respectively. We leveraged the dbSNP database to extract SNPs associated with these genes per disease, as shown in **Table 1**. The obtained SNPs per disease were extracted from the whole phased data containing 4,932 samples of these 20 ethnolinguistic cultural groups, yielding five disease-specific phased haplotype datasets (**Table 1**).

To evaluate the extent of substructures within disease-specific polymorphism across worldwide ethnolinguistic cultural groups, we leverage each constructed disease-specific phased haplotype dataset, to perform genetic structure analysis based on principal component analysis (PCA) using smartpca, part of the EIGENSOFT 3.0 package (Patterson and Price, 2006). Genesis software http://www.bioinf.wits.ac.za/software/genesis was used to plot PCA.

Proportion of Ancestral/Derived Alleles Among Risk-Conferring Alleles

Each of these four disease-specific phased haplotype datasets was used to analyze the fraction of derived and ancestral alleles and at-risk alleles within each ethnolinguistic cultural group. A previous work showed that derived alleles are more often minor alleles (<50% allele frequency) and associated with risk than ancestral alleles (Gorlova et al., 2012). Therefore, we define risk alleles as follows: if a gene is reported to increase the risk of disease (odd ratio >1) from either the DisGeNET or GWAS Catalog, the risk allele was defined as a minor allele (for all SNPs associated with the gene); otherwise (odd ratio <1), it is defined as a major allele (for all SNPs associated with the gene).

The SNP ancestral alleles were downloaded from the Ensembl, a 59 comparative 32 species alignment (Paten et al., 2008), and we further checked the SNPs for those present in the dbSNP database. Each of these four disease-specific phased haplotype datasets was further annotated using the VCFtools "fillOaa" script (Danecek et al., 2011) with the ancestral allele recorded using the "AA" INFO tag. For each disease-specific dataset, we determined the proportion of risk alleles that were ancestral or derived alleles. We first computed, for each SNP, the fraction of the ancestral allele, which was calculated by dividing the number of times the defined risk allele matched with the ancestral allele by the total number of copies of all alternative alleles across all samples (within each ethnolinguistic cultural group per disease) for a particular SNP. The fraction of the derived allele is equivalent to one minus the fraction of the ancestral allele. As mentioned earlier, derived alleles are more often minor alleles and associated with risk rather than ancestral alleles. Therefore, we investigated the relationship between the fraction of derived alleles, at-risk alleles, and ethnolinguistic cultural group SNP minor allele frequency. To this end, the alternative (minor) alleles were categorized into six bins, (0–0.05, >0.05–0.1, >0.1–0.2, >0.2–0.3, >0.3–0.4, and >0.4–0.5) with respect to each ethnolinguistic cultural dataset frequencies and independently computed the fractions of derived alleles in each bin. Furthermore, we computed the fraction of ancestral/derived alleles for all these known disease-specific genes. To this end, we aggregated the fraction of ancestral/derived alleles at the SNP-based level to gene, considering all SNPs located within the genes' downstream or upstream region (Chimusa et al., 2015).

Distribution of Minor Allele Frequency and Gene-Specificity in SNP Frequencies

To examine the extent of common variants across these 20 ethnolinguistic cultural groups within a specific disease (TB, HIV, sickle cell anemia, and malaria) and known actionable genes from ACG, the distribution of the minor allele frequency was investigated. To this end, the proportion of minor alleles was categorized into six bins (0–0.05, >0.05–0.1, >0.1–0.2, >0.2–0.3, >0.3–0.4, and >0.4–0.5) with respect to each ethnolinguistic cultural group with a disease. The minor allele frequency (MAF) per SNP for each category was computed using Plink software (Purcell et al., 2007). Furthermore, the fraction of gene-specific in SNP frequency for each gene was computed. To this end, the fraction of gene-specific SNP frequency was computed, assuming that SNPs in upstream and downstream within a gene region are close and possibly in linkage disequilibrium (LD). Minor allele frequency per SNP has aggregated a gene level.

Aggregating SNP Summary Statistics at the Gene Level

SNP-specific allele frequencies or the proportion of ancestral/derived alleles from SNPs 40 kb downstream and upstream within a gene region as per the dbSNP database were aggregated (Chimusa et al., 2016). Under the null hypothesis, frequency/proportion P_κ (k = 1,..., L) with a continuous distribution is uniformly distributed at the interval [0,1]. It follows that a parametric cumulative distribution function F can be chosen, and P_κ can be transformed into quantile according to $q_\kappa = F^{-1}(P_\kappa)$. The combined frequency/proportion $C^P = \frac{\sum_{\kappa=1}^{L} P_\kappa}{\sqrt{L}}$ is a sum of independent and identically distributed random variables P_κ. To account for the independence assumption, given the correlation among neighboring genomic markers (Chimusa et al., 2016), we implement the Stouffer–Liptak method accounting for spatial correlations among SNPs within a gene or SNPs within a given sub-network. The overall statistic can be obtained by $P = \phi(C^P)$, in which ϕ is the cumulative distribution function of the standard normal distribution.

Key Points

- Personalized medicine including African genomics will be beneficial both to the continent and worldwide.
- Generalizability and transferability of actionable genes are challenging but will improve clinical population healthcare.
- Investigating the distribution of gene-specificity in SNP frequencies, gene-specificity in proportion of derived alleles, and gene-specificity in burden of pathogenic mutations will reveal population-specific actionable genes.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**; further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

The authors have equally contributed to writing, interpretation of the results and revision of the manuscript. However, SA and EC carried out all the analysis, and EC designed, administrated and supervised the manuscript.

FUNDING

The authors are supported in part by the DAAD, the German Academic Exchange Programme, under reference number 91653117, the National Institutes of Health Common Fund under grant number 1U2RTW012131-01 (COBIP), the National Research Foundation of South Africa for funding (NRF) (Grant #RA171111285157/119,056), and the Sub-Saharan African Network for TB/HIV Research Excellence (SANTHE), a DELTAS African Initiative (grant # DEL-15-006). Responsibility for the information and views expressed in the article lies entirely with the authors.

ACKNOWLEDGMENTS

We are grateful to H3ABioNet consortium members for helpful discussions and the CHPC (https://www.chpc.ac.za/) facility for providing a computing platform. We thank Dr. Denis Awany, Ms Jacquiline Mugo and Mr. Joel Defo to have read our manuscript and for their advises and supports during the preparation of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.835713/full#supplementary-material

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A Method and Server for Predicting Damaging Missense Mutations. *Nat. Methods* 7, 248–249. doi:10.1038/nmeth0410-248

Amendola, L. M., Dorschner, M. O., Robertson, P. D., Salama, J. S., Hart, R., Shirts, B. H., et al. (2015). Actionable Exomic Incidental Findings in 6503 Participants: Challenges of Variant Classification. *Genome Res.* 25, 305–315. doi:10.1101/gr.183483.114

Bope, C. D., Chimusa, E. R., Nembaware, V., Mazandu, G. K., de Vries, J., and Wonkam, A. (2019). Dissecting In Silico Mutation Prediction of Variants in African Genomes: Challenges and Perspectives. *Front. Genet.* 10 (601), 601. doi:10.3389/fgene.2019.00601

Chimusa, E. R., Mbiyavanga, M., Mazandu, G. K., and Mulder, N. J. (2016). ancGWAS: a Post Genome-wide Association Study Method for Interaction, Pathway and Ancestry Analysis in Homogeneous and Admixed Populations. *Bioinformatics* 32 (4), 549–556. doi:10.1093/bioinformatics/btv619

Chimusa, E. R., Meintjies, A., Tchanga, M., Mulder, N., Seoighe, C., Soodyall, H., et al. (2015). A Genomic Portrait of Haplotype Diversity and Signatures of Selection in Indigenous Southern African Populations. *PLoS Genet.* 11, e1005052. doi:10.1371/journal.pgen.1005052

Choi, Y., and Chan, A. P. (2015). PROVEAN Web Server: a Tool to Predict the Functional Effect of Amino Acid Substitutions and Indels. *Bioinformatics* 31 (16), 2745–2747. doi:10.1093/bioinformatics/btv195

Choudhury, A., Aron, S., Botigué, L. R., Sengupta, D., Botha, G., Bensellak, T., et al. (2020). High-depth African Genomes Inform Human Migration and Health. *Nature* 586 (7831), 741–748. doi:10.1038/s41586-020-2859-7

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The Variant Call Format and VCFtools. *Bioinformatics* 27, 2156–2158. doi:10.1093/bioinformatics/btr330

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., and Batzoglou, S. (2010). Identifying a High Fraction of the Human Genome to Be under Selective Constraint Using GERP++. *PLoS Comput. Biol.* 6, e1001025. doi:10.1371/journal.pcbi.1001025

Doerks, T., Copley, R. R., Schultz, J., Ponting, C. P., and Bork, P. (2002). Systematic Identification of Novel Protein Domain Families Associated with Nuclear Functions. *Genome Res.* 12, 47–56. doi:10.1101/gr.203201

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., et al. (2015). Comparison and Integration of Deleteriousness Prediction Methods for Nonsynonymous SNVs in Whole Exome Sequencing Studies. *Hum. Mol. Genet.* 24, 2125–2137. doi:10.1093/hmg/ddu733

Dorschner, M. O., Amendola, L. M., Turner, E. H., Robertson, P. D., Shirts, B. H., and Gallego, C. J. Actionable, Pathogenic Incidental Findings in 1,000 Participants' Exomes. *Am. J. Hum. Genet.* 93, 631–640. doi:10.1016/j.ajhg.2013.08.006

Garber, M., Guttman, M., Clamp, M., Zody, M. C., Friedman, N., and Xie, X. (2009). Identifying Novel Constrained Elements by Exploiting Biased Substitution Patterns. *Bioinformatics* 25, i54–i62. doi:10.1093/bioinformatics/btp190

Gorlova, O. Y., Ying, J., Amos, C. I., Spitz, M. R., Peng, B., and Gorlov, I. P. (2012). Derived SNP Alleles Are Used More Frequently Than Ancestral Alleles as Risk-Associated Variants in Common Human Diseases. *J. Bioinform Comput. Biol.* 10 (2), 1241008. doi:10.1142/S0219720012410089

Gudykunst, W. B., and Schmidt, K. L. (1987). Language and Ethnic Identity: An Overview and Prologue. *J. Lang. Soc. Psychol.* 6 (3-4), 157–170. doi:10.1177/0261927x8763001

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., et al. (2015). The African Genome Variation Project Shapes Medical Genetics in Africa. *Nature* 517, 327–332. doi:10.1038/nature13997

Hunter, J. E., Irving, S. A., Biesecker, L. G., Buchanan, A., Jensen, B., Lee, K., et al. (2016). A Standardized, Evidence-Based Protocol to Assess Clinical Actionability of Genetic Disorders Associated with Genomic Variation. *Genet. Med.* 18, 1258–1268. doi:10.1038/gim.2016.40

Ionita-Laza, I. M. K., McCallum, K., Xu, B., and Buxbaum, J. D. (2016). A Spectral Approach Integrating Functional Genomic Annotations for Coding and Noncoding Variants. *Nat. Genet.* 48 (2), 214–220. doi:10.1038/ng.3477

Jagadeesh, K. A., Wenger, A. M., Berger, M. J., Guturu, H., Stenson, P. D., Cooper, D. N., et al. (2016). M-CAP Eliminates a Majority of Variants of Uncertain Significance in Clinical Exomes at High Sensitivity. *Nat. Genet.* 48 (12), 1581–1586. doi:10.1038/ng.3703

Kim, S. J., Jhong, J. H., Lee, J., and Koo, J. Y. (2017). Erratum to: Meta-Analytic Support Vector Machine for Integrating Multiple Omics Data. *BioData Min.* 10 (2), 8. doi:10.1186/s13040-017-0128-6

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., et al. (2016). ClinVar: Public Archive of Interpretations of Clinically Relevant Variants. *Nucleic Acids Res.* 44, D862–D868. doi:10.1093/nar/gkv1222

Lebeko, K., Manyisa, N., Chimusa, E. R., Mulder, N., Dandara, C., and Wonkam, A. (2017). A Genomic and Protein-Protein Interaction Analyses of Nonsyndromic Hearing Impairment in Cameroon Using Targeted Genomic Enrichment and Massively Parallel Sequencing. *OMICS A J. Integr. Biol.* 21, 90–99. doi:10.1089/omi.2016.0171

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352

Liu, X., Jian, X., and Boerwinkle, E. (2011). dbNSFP: a Lightweight Database of Human Nonsynonymous SNPs and Their Functional Predictions. *Hum. Mutat.* 32, 894–899. doi:10.1002/humu.21517

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., A Reshef, Y., K Finucane, H., et al. (2016). Reference-based Phasing Using the Haplotype Reference Consortium Panel. *Nat. Genet.* 48 (11), 1443–1448. doi:10.1038/ng.3679

Loh, P.-R., Palamara, P. F., and Price, A. L. (2016). Fast and Accurate Long-Range Phasing in a UK Biobank Cohort. *Nat. Genet.* 48, 811–816. doi:10.1038/ng.3571

Lu, Q. H. Y., Hu, Y., Sun, J., Cheng, Y., Cheung, K. H., and Zhao, H. (2015). A Statistical Framework to Predict Functional Non-coding Regions in the Human Genome through Integrated Analysis of Annotation Data. *Sci. Rep.* 5, 10576. doi:10.1038/srep10576

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., et al. (2016). The Ensembl Variant Effect Predictor. *Genome Biol.* 17, 122. doi:10.1186/s13059-016-0974-4

Michalopoulos, S. (2012). The Origins of Ethnolinguistic Diversity. *Am. Econ. Rev.* 102 (4), 1508–1539. doi:10.1257/aer.102.4.1508

Ng, P. C., Henikoff, S., Chun, S., and Fay, J. C. (2006). Predicting the Effects of Amino Acid Substitutions on Protein functionIdentification of Deleterious Mutations within Three Human Genomes. *Annu. Rev. Genom. Hum. Genet.Genome Res.* 719, 611553–801561. doi:10.1146/annurev.genom.7.080505.115630

Paten, B., Herrero, J., Fitzgerald, S., Beal, K., Flicek, P., Holmes, I., et al. (2008). Genome-wide Nucleotide-Level Mammalian Ancestor Reconstruction. *Genome Res.* 18, 1829–1843. doi:10.1101/gr.076521.108

Patterson, N. P. A. L., and Price, D. (2006). Population Structure and Eigenanalysis. *PLoS Genet.* 2 (12), e190. doi:10.1371/journal.pgen.0020190

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* 81, 559–575. doi:10.1086/519795

Quang, D., Chen, Y., Dann, X., and Xie, X. (2014). DANN: a Deep Learning Approach for Annotating the Pathogenicity of Genetic Variants. *Bioinformatics* 31 (5), 761–763. doi:10.1093/bioinformatics/btu703

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., and Kircher, M. (2018). CADD: Predicting the Deleteriousness of Variants throughout the Human Genome. *Nucleic Acids Res.* 47, D886. doi:10.1093/nar/gky1016

Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the Functional Impact of Protein Mutations: Application to Cancer Genomics. *Nucleic Acids Res.* 39, e118. doi:10.1093/nar/gkr407

Schwarz, J. M., Rödelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster Evaluates Disease-Causing Potential of Sequence Alterations. *Nat. Methods* 7, 575–576. doi:10.1038/nmeth0810-575

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP: the NCBI Database of Genetic Variation. *Nucleic Acids Res.* 29 (1), 308–311. doi:10.1093/nar/29.1.308

Shihab, H. A., Gough, J., Cooper, D. N., Day, I. N. M., and Gaunt, T. R. (2013). Predicting the Functional Consequences of Cancer-Associated Amino Acid Substitutions. *Bioinformatics* 29, 1504–1510. doi:10.1093/bioinformatics/btt182

Siva, N. (2008). 1000 Genomes Project. *Nat. Biotechnol.* 26 (3), 256. doi:10.1038/nbt0308-256b

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data. *Nucleic Acids Res.* 38, e164. doi:10.1093/nar/gkq603

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership