

https://dspace.org/

Genetic and environmental factors associated with cognition in an ageing South African population

Soo, Cassandra Claire

2021
https://hdl.handle.net/10539/33511

Downloaded from a DSpace repository. https://dspace.org/

Genetic and environmental factors associated with cognition in an ageing South African population

Cassandra Claire Soo



An integrated Thesis submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of PhD in Human Genetics

Johannesburg, 2021

Student declaration

I, Cassandra Claire Soo, (student number 0713996A) declare that this Thesis is my own work and that analyses performed by others are acknowledged and specified. Due to my involvement with the planning of the HAALSI and AWI-Gen studies, I proposed the use of the cognition data that were being collected for HAALSI for genomic studies using genomic data planned for the AWI-Gen study and formulated the initial research proposal for this degree. I was also responsible for DNA extraction and sample preparation for the AWI-Gen study as the SBIMB Biobank Manager. I confirm that I performed a detailed literature review on cognitive genetics. I supervised and interpreted the SNP genotyping for the *APOE* locus and interpreted and compared TOMM40 length polymorphism between platforms. I performed relevant analyses (statistics and GWAS) and data interpretation, wrote the first drafts of the papers and included amendments and elaborations on the advice of my supervisors and co-authors for:

1. The publication: Soo CC, Farrell MT, Tollman S, Berkman L, Nebel A, Ramsay M. Apolipoprotein E Genetic Variation and Its Association With Cognitive Function in Rural-Dwelling Older South Africans. Vol. 12, *Frontiers in Genetics*. 2021. p.1683. https://doi.org/10.3389/fgene.2021.689756 submitted here as part of my thesis (Chapter 2)
2. The manuscript submitted for peer-review and presented as part of this thesis (Chapter 3)
3. The work presented in block format in Chapter 4

This Thesis is being submitted for the Degree of PhD in Human Genetics at the University of the Witwatersrand, Johannesburg. I confirm that it has not been submitted before for any degree or examination at any other University.



Signature

Signed on the 29th day of October 2021 in Crowthorne, United Kingdom

Co-author agreement

Agreement by co-authors: By signing this declaration, the co-authors listed below agree to the use of these articles by Cassandra Claire Soo as part of her Thesis.

Article 1: Title: *APOE* genetic variation and its association with cognitive function in rural-dwelling older South Africans

Accepted for publication by *Frontiers in Genetics: Behavioural and Psychiatric Genetics* on 18 August 2021.

Authors	Name	Signature	Date
Primary	Cassandra C. Soo		26 Sep 2021
Co-author	Meagan T. Farrell		25 Oct 2021
Co-author	Lisa Berkman		october 25,2021
Co-author	Stephen Tollman		26 Oct 2021
Senior	Almut Nebel		27 Oct 2021
Senior	Michèle Ramsay		7 Oct 2021

Comments by primary supervisor: The PhD candidate was responsible for generating the genetic data and analysing the data. She wrote the first draft of the paper and was responsible for revising the manuscript and submitting it to the journal. She was also responsible for all revisions and interactions with the journal.

Article 2: Title: Genome-Wide Association Study of Population Standardised Cognitive Performance Phenotypes in a rural South African Community

Submitted for peer-review.

Authors	Name	Signature	Date
Primary	Cassandra C. Soo		26 Sep 2021
Co-author	Jean-Tristan Brandenburg		29 Sep 2021
Co-author	Almut Nebel		27 Oct 2021
Co-author	Stephen Tollman		26 Oct 2021
Co-author	Lisa Berkman	Lisa Berkman	oct 25,2021
Senior	Ananyo Choudhury	Ananyo Choudhury	7th October 2021
Senior	Michèle Ramsay		7 October 2021

Comments by primary supervisor: The PhD candidate was responsible for analysing the data. She wrote the first draft of the paper and was responsible for revising the manuscript and submitting it to the journal.

Dedication

This is dedicated to my father, Kwong Peter Michael Soo (24 Sep 1950 - 4 Oct 2020). By completing this degree, I will hold you forever in name, as I hold you in my heart, and on my body.

For my lil girl, my guiding star, Ichigo John Eirene Ah Ling – don't be afraid to dream.

Co-authored publications related to work published during the PhD research period

Ali SA, Soo C, Agongo G, Alberts M, Amenga-Etego L, Boua RP, Choudhury A, et al. Genomic and environmental risk factors for cardiometabolic diseases in Africa: methods used for Phase 1 of the AWI-Gen population cross-sectional study. Glob Health Action. 2018;11(sup2):1507133. doi: 10.1080/16549716.2018.1507133. PMID: 30259792; PMCID: PMC6161608.

Boua, P.R., Soo, C.C., Debpuur, C. *et al*. Prevalence and socio-demographic correlates of tobacco and alcohol use in four sub-Saharan African countries: a cross-sectional study of middle-aged adults. *BMC Public Health* **21,** 1126 (2021). https://doi.org/10.1186/s12889-021-11084-1

Soo CC, Farrell MT, Tollman S, Berkman L, Nebel A, Ramsay M. Apolipoprotein E Genetic Variation and Its Association With Cognitive Function in Rural-Dwelling Older South Africans. Vol. 12, *Frontiers in Genetics*. 2021. p.1683. https://doi.org/10.3389/fgene.2021.689756

Abstract

This research was the first to look at genetic associations with latent cognitive function using socially appropriate screening tools in a rural-dwelling African population. By leveraging on existing infrastructure and the added resources of the collaborative studies; The Health and Aging in Africa: A Longitudinal Study of an INDEPTH Community in South Africa (HAALSI) and the Africa Wits-INDEPTH Partnership for Genomic Studies (AWI-Gen), we were able to pioneer the first cross-sectional population study of genomics and cognitive function in a sub-Saharan African population.

Three separate studies, each with their own objectives, explored aspects of genetic associations with five cognitive traits namely, total cognition score, verbal episodic memory, executive function, language, and visuospatial cognition

Apolipoprotein E (*APOE*) and its effect on cognitive function in this cohort was examined due to documented associations of *APOE* ε4 with poorer cognitive performance in studies assessing cognitive function at the domain level. Marginal associations were found where *ε*2 homozygotes exhibited higher levels of executive function and *ε*3/*ε*4 heterozygotes had poorer visuospatial cognition.

Our genome-wide association study detected novel genome-wide significant signals for verbal episodic memory and language performance. Common and rare African variants had modest effects on cognitive function, which showed the importance of using the Illumina Human Health and Heredity in Africa custom array with a population-specific reference panel for imputation.

Finally, an *in-silico* method for determining the length of a homopolymer (T) repeat in Translocase of Outer Mitochondrial Membrane 40 (*TOMM40)* was developed and validated. Allele frequency distributions and the co-segregation patterns of *TOMM4*0 and *APOE* in our cohort versus that of African and European 1000 Genomes populations revealed differences in linkage disequilibrium structure between populations. Regression models showed marginal significance for the association between individuals homozygous for the long allele of *TOMM40* and lower episodic memory scores.

This work has set the stage for future genetic studies of cognitive function in Africa. AWI-Gen Phase 2 presents an opportunity to increase sample size and replicate observed associations across the whole AWI-Gen cohort. Furthermore, longitudinal data from later HAALSI waves may enable assessment of the prevalence and genetic correlates influencing cognitive decline and dementia.

Acknowledgements

Thank you to the members of the Agincourt community who volunteered to participate in this study, and without whom we would not have been able to pioneer this research.

I would like to acknowledge the MRC/Wits Rural Agincourt team under the leadership of Prof Stephen Tollman, for their incredible efforts at maintaining the Agincourt HDSS and continuing to foster the interests of the community through numerous collaborative research projects.

I would like to thank Prof Michèle Ramsay and the central AWI-Gen team for their work in establishing a truly valuable genomic resource, thank you for the opportunities that were provided to me throughout my career at the Sydney Brenner Institute for Molecular Bioscience (SBIMB), University of the Witwatersrand, Johannesburg.

To Prof Almut Nebel, thank you for offering your expertise and supervision during this seemingly endless journey with me.

A special thank you to Dr Jean-Tristan Brandenburg for his tireless efforts, aid, and interest in my research. His role in the development of the bioinformatics pipelines used in this thesis, along with the work of Dr Ananyo Choudhury, Dr Dhriti Sengupta, and Shaun Aron under the guidance of Prof Scott Hazelhurst and H3ABionet were integral to this research.

I also thank Dr Meagan Farrell and Prof Lisa Berkman from the Harvard T.H Chan School of Public Health for allowing my research to follow on from theirs as part of the HAALSI collaboration. And Dr Innocent Maposa from the Wits School of Public Health for his biostatistics aid.

To my colleagues at the SBIMB (Dr Palwende Romuald Boua, Dr Andrew May, Dr Phelelani Mpangase, Wenlong Chen, Jorge de Rocha, Natalie Smyth, Busisiwe Mthembu, Jocelyne Gayenga, Mahtaab Hayat, and Mmatshepho Tauyane) thank you for your support and sharing your knowledge in the areas of statistics and bioinformatics, and for lightening my administrative load.

Table of Contents

Student declaration ..... *i*

Co-author agreement ..... *ii*

Dedication ..... *iv*

Co-authored publications related to work published during the PhD research period ..... *v*

Abstract ..... *vi*

Acknowledgements ..... *viii*

List of Figures ..... *xii*

Chapter 1 ..... **xii**

Chapter 2 ..... **xii**

Chapter 3 ..... **xii**

Chapter 4 ..... **xii**

Appendix 2 ..... **xiii**

Appendix 3 ..... **xiii**

Appendix 4 ..... **xiv**

List of Tables ..... *xv*

Chapter 2 ..... **xv**

Chapter 3 ..... **xv**

Chapter 4 ..... **xv**

Appendix 2 ..... **xv**

Appendix 4 ..... **xvi**

List of Abbreviations ..... *xvii*

Chapter 1 ..... *1*

1 Background **1**

1.1 Epidemiology and implications for healthy ageing in Africa 1

1.2 Sample population 1

1.3 Cognitive function 4

1.4 Genetic association studies 9

1.5 Rationale, aims, and objectives 15

1.6 Ethical considerations 20

Chapter 2 ..... *21*

2 *APOE* genetic variation and its association with cognitive function in rural-dwelling older South Africans **21**

2.1 Abstract 21

2.2 Introduction 22

2.3 Materials and Methods 24

2.4 Results 27

2.5 Discussion 34

2.6 Data availability statement 38

2.7 Author Contributions 39

2.8 Funding 39
2.9 Acknowledgments 40

Chapter 3 ..... *41*

3 Genome-wide association study of population standardised cognitive performance phenotypes in a rural South African community **41**
3.1 Abstract 41
3.2 Introduction 41
3.3 Methods 43
3.4 Results 46
3.5 Discussion 58
3.6 Acknowledgements and funding 61
3.7 Data availability statement 62

Chapter 4 ..... *63*

4 Estimating the length of the poly-T tract in *TOMM40* and its association with cognitive function in Southern Africans **63**
4.1 Introduction 63
4.2 Methods 66
4.3 Results 72
4.4 Discussion 82

Chapter 5 ..... *87*

5 Discussion and conclusions **87**
5.1 Overview 87
5.2 Strengths and limitations 93
5.3 Future considerations 97

References *101*

Appendix 1 ..... *119*

A1.1 Ethics approval certificate **119**

A1.2 Plagiarism declaration **120**

Appendix 2 ..... *121*

A2.1 *APOE* Sanger sequencing validation **121**

A2.2 Diagnostics of Regression **124**
A2.2.1 Total cognition score 124
A2.2.2 Executive Function 127
A2.2.3 Language 129
A2.2.4 Verbal Episodic Memory 131
A2.2.5 Visuospatial ability 133

Appendix 3 ..... *135*

A3.1 Unreported GWAS results for total cognition score **135**

A3.2 Unreported GWAS results for verbal episodic memory **137**

A3.3 Unreported GWAS results for executive function **137**

A3.4 Unreported GWAS results for language **139**

A3.5 Unreported GWAS results for visuospatial cognition **142**

Appendix 4 ..... *147*

A4.1 Homopolymer length detection 147
A4.1.1 Poly-T length fragment standards 147
A4.1.2 Poly-T length estimation methods 152
A4.2 Phylogenetic analysis 158
A4.3 Diagnostics of regression 159
A4.3.1 Total cognition score 159
A4.3.2 Executive function 161
A4.3.3 Language 163
A4.3.4 Verbal episodic memory 165
A4.3.5 Visuospatial ability 167

List of Figures

Chapter 1

Figure 1.1: Agincourt HDSS Study Site…… **2**

Figure 1.2.: Age-specific survivorship curves for Agincourt HDSS…… **4**

Figure 1.3: Details of intercorrelations and factor loadings for the final CFA domain model of cognitive variables from the OCS-Plus (from Farrell et al. (2020)) …… **8**

Figure 1.4: Stepwise GWAS QC process for automated pipeline……... **14**

Figure 1.5: Research summary……..……... **18**

Figure 1.6: Power estimation for GWAS of total cognition score……... **19**

Chapter 2

Figure 2.1: Comparison of APOE allele distribution in our dataset (Study data) compared to 1000 Genomes super population datasets AFR, EUR, EAS, SAS, and AMR…… **29**

Figure 2.2: Comparison of *APOE* allele frequency distributions from four African studies, showing sample size for each study…… **31**

Chapter 3

Figure 3.1: Population structure and affinities of the HAALSI/AWI-Gen participants… **47**

Figure 3.2: Genome-wide and suggestive associations for verbal episodic memory…… **50**

Figure 3.3: Genome-wide and suggestive associations for language**….**……... **52**

Supplementary Figure 3.1: Box-plots showing effects of variants on cognitive function **55**

Chapter 4

Figure 4.1: Workflow diagram detailing each step in the homopolymer detection pipeline developed inhouse by Dr Brandenburg…… **68**

Figure 4.2: Workflow diagram detailing each step in the imputation process developed by Dr Brandenburg…… **70**

Figure 4.3: Comparison of imputation quality results for BEAGLE and SHAPEIT.. **74**

Figure 4.4: Allele distribution of *APOE* and *TOMM40* poly-T length polymorphism in Africans vs Europeans…….. **75**

Figure 4.5: Principal component analysis showing the distribution of *TOMM40-APOE* haplotypes based on LD patterns…….. **76**

Appendix 2

Figure A.2.1: Electrophoretograms for APOE for 27 samples that were successfully sequenced…… **122-124**
Figure A.2.2: Diagnostic regression plots for the residuals of total cognition score **126**
Figure A.2.3: Diagnostic regression plots for the residuals of executive function **128**
Figure A.2.4: Diagnostic regression plots for the residuals of language…… **130**
Figure A.2.5: Diagnostic regression plots for the residuals of verbal episodic memory **132**
Figure A.2.6: Diagnostic regression plots for the residuals of visuospatial ability **134**

Appendix 3

Figure A3.1: rs9893796 on chromosome 17 associated with total cognition score **135**
Figure A3.2: rs80026667 on chromosome 7 associated with total cognition score **135**
Figure A3.3: rs6043030 on chromosome 20 associated with total cognition score **136**
Figure A3.4: rs12059793 on chromosome 1 associated with total cognition score **136**
Figure A3.5: rs9652190 on chromosome 13 associated with total cognition score **136**
Figure A3.6: rs140372794 on chromosome 8 associated with episodic memory… **137**
Figure A3.7: rs10929007 on chromosome 2 associated with executive function… **138**
Figure A3.8: rs2990655 on chromosome 1 associated with executive function….. **138**
Figure A3.9: rs2463700 on chromosome 6 associated with executive function….. **138**
Figure A3.10: rs146628683 on chromosome 20 associated with executive function **139**
Figure A3.11: rs76885705 on chromosome 21 associated with executive function **139**
Figure A3.12: rs10405447 on chromosome 19 associated with language…… **140**
Figure A3.13: rs140234736 on chromosome 1 associated with language…… **140**
Figure A3.14: rs1163379 on chromosome 3 associated with language…… **140**
Figure A3.15: rs4736068 on chromosome 8 associated with language…… **141**
Figure A3.16: rs2289416 on chromosome 15 associated with language……... **141**
Figure A3.17: rs114096889 on chromosome 2 associated with language…… **141**
Figure A3.18: rs189202094 on chromosome 6 associated with language…… **142**
Figure A3.19: rs311477 on chromosome 1 associated with language……... **142**
Figure A3.20: rs189202094 on chromosome 6 associated with visuospatial ability **143**
Figure A3.21: rs140578927 on chromosome 6 associated with visuospatial ability **143**
Figure A3.22: rs112938427 on chromosome 4 associated with visuospatial ability **143**
Figure A3.23: rs1035815 on chromosome 1 associated with visuospatial ability **144**

Figure A3.24: rs114096889 on chromosome 6 associated with visuospatial ability **144**
Figure A3.25: rs147982326 on chromosome 11 associated with visuospatial ability **144**
Figure A3.26: rs191611493 on chromosome 5 associated with visuospatial ability **145**
Figure A3.27: rs149151348 on chromosome 6 associated with visuospatial ability **145**
Figure A3.28: rs6451188 on chromosome 5 associated with visuospatial ability **146**
Figure A3.29: rs78190320 on chromosome 17 associated with visuospatial ability **146**

Appendix 4

Figure A4.1: Gene synthesis report for clone for fragment length T=35 and T=16 **148**
Figure A4.2: Alignment files for fragment standards T=35 and T=16…… **149**
Figure A4.3: Gene synthesis report for clone for fragment length T=18 and T=22 **150**
Figure A4.4: Alignment files for fragment standards T=18 and T=22…… **151**
Figure A4.5 Allele report from fragment analysis from PCR fragment analysis for AB0009…… **156**
Figure A4.6: Inhouse STR-caller results for AB0009 (homozygous T=16, S/S)… **157**
Figure A4.7: Unrooted evolutionary tree diagrams…… **158**
Figure A.4.8: Diagnostic regression plots for the residuals of total cognition score **160**
Figure A.4.9: Diagnostic regression plots for the residuals of executive function **162**
Figure A.4.10: Diagnostic regression plots for the residuals of language…… **164**
Figure A.4.11: Diagnostic regression plots for the residuals of episodic memory... **166**
Figure A.4.12: Diagnostic regression plots for the residuals of visuospatial ability **168**

List of Tables

Chapter 2

Table 2.1: Summary characteristics of study population…… **27**

Table 2.2: *APOE* allele and genotype frequencies…… **28**

Table 2.3: *APOE* allele association with HIV status …… **30**

Table 2.4: Linear regression model 1 using all genotypes to estimate the effects of *APOE* and covariates on cognitive function ……..... **32**

Supplementary Table 2.1: Linear regression models estimating the effects of *APOE* and various covariates on cognitive function……... **33**

Chapter 3

Table 3.1: Descriptive statistics showing cognitive trait and covariate distribution **48**

Table 3.2: Minor allele frequency distribution of genome-wide significant signals in this sample compared to 1000 Genomes Super Populations …… **49**

Supplementary Table 3.1: Summary of independent association signals (3e-06) detected in GWASs for the five cognitive phenotypes…….. **54**

Supplementary Table 3.2: Replication of previous associations for language and verbal episodic memory …… **56-57**

Chapter 4

Table 4.1: Haplotype distribution for African panel vs CEU from 1000 Genomes **73**

Table 4.2: HAALSI-AWI-Gen participant summary and phenotypic description.. **77**

Table 4.3: Allele and genotype distribution of *TOMM40*……... **78**

Table 4.4: *APOE_TOMM40* haplotype distribution for the HAALSI-AWI-Gen participants (N=2234 individuals)…… **78**

Table 4.5: Linear regression models estimating the effects of *TOMM40* and various covariates on cognitive function…… **80-81**

Appendix 2

Table A.2.1: TaqMan genotyping results compared to sequencing validation.. **121-122**

Appendix 4

Table A4.1: Comparison of results of three methods for poly-T length determination**152-155**

List of Abbreviations

AA - African American

ACB - African Caribbeans in Barbados

AD - Alzheimer's disease

AFR - 1000 Genomes African super population group

AGI - AWI-Gen Agincourt (Bushbuckridge) sample population

AMR - 1000 Genomes Admixed American super population group

APOE - Apolipoprotein E gene

ASW - Americans of African Ancestry in SW USA

AWI-Gen - Africa Wits-INDEPTH Partnership for Genomic Studies

CEU - Utah Residents (CEPH) with Northern and Western European Ancestry

CFA - Confirmatory factor analysis

CHARGE - Cohorts for Heart and Aging Research in Genomic Epidemiology

COGENT - Cognitive Genomics Consortium

CVDs - Cardiovascular diseases

DIK - AWI-Gen Dikgale (DIMAMO) sample population

EAS - 1000 Genomes East Asian super population group

EFA - Exploratory factor analysis

ELSA - English Longitudinal Study of Ageing

ESN - Esan in Nigeria

EUR - 1000 Genomes European super population group

FIN - Finnish in Finland

GBR - British in England and Scotland

GEMMA - Genome-Wide Efficient Mixed Model Association

GRCh37 - Genome Reference Consortium Human genome build 37

GWAS - Genome-wide association study

GWD - Gambian in Western Divisions in the Gambia

HAALSI - The Health and Aging in Africa: A Longitudinal Study of an INDEPTH Community in South Africa

HAND - HIV-associated neurocognitive disorder

HDSS - Health and Demographic Surveillance system

HIV - Human Immunodeficiency Virus

HMM - Hidden Markov model

HREC (Medical) - Human Research Ethics Committee

HWE - Hardy Weinberg equilibrium

IBS - Iberian Population in Spain

INDEL - Insertion-deletion

IQ - Intelligence quotient

KGP - 1000 Genomes Project

LMICs - Low and middle-income countries

LASI - Longitudinal Ageing study in India

LD - Linkage disequilibrium

LOAD - Late-onset Alzheimer's disease

LWK - Luhya in Webuye, Kenya

QC - Quality control

MAF - Minor allele frequency

MCI - Mild cognitive impairment

MRC - Medical Research Council

MSL - Mende in Sierra Leone

MTAG - Multi-trait analysis of GWAS

NCDs - Non-communicable diseases

OCS-Plus - Oxford Cognition Screen Plus

PCA - Principal component analysis

PCoA - Principal coordinate analysis

PWBT - Positional Burrows Wheeler transform

SA - South African

SAS - 1000 Genomes South Asian super population group

SBIMB - Sydney Brenner Institute for Molecular Bioscience

SD - Standard deviation

SE - Standard error

SNPs - Single nucleotide polymorphisms

SOW - AWI-Gen Soweto (DPHRU) population sample

SSA - Sub-Saharan Africa

TOMM40 - Translocase of outer mitochondrial membrane 40 gene

TSI - Toscani in Italia

UKBB - United Kingdom Biobank

US HRS - United States Health and Retirement Study

WAIS - Wechlsler Adult Intelligence Scale

WGS - Whole genome sequence data

Wits - University of the Witwatersrand, Johannesburg

YRI - Yoruba in Ibadan, Nigeria

Chapter 1

1 Background

1.1 Epidemiology and implications for healthy ageing in Africa

The global population is ageing and as the proportion of people aged 60 years and older increases, so does the burden of age-related disorders and disability (1). This is of considerable concern as low- and middle-income countries (LMICs) especially in sub-Saharan Africa (SSA) are expected to shoulder this increased burden on top of the health burdens related to infectious and non-communicable disease (2–4). In South Africa (SA), which is undergoing an epidemiological transition fueled by lifestyle changes, the increased incidence of chronic and age-related conditions is observed to be more prevalent in those infected with human immunodeficiency virus (HIV) (2–4). Furthermore, a demographic shift has occurred where the proportion of adults over the age of 65 is greater than the proportion of the population under age five (2,4,5).

Normal cognitive function is essential for daily tasks requiring cognitive demand, and impairment is associated with poorer life outcomes and decreased quality of life especially in old age (1,6,7). Cognitive function declines naturally with age in some individuals, and much of this decline in the cognitive domains becomes noticeable over the age of 55 years (8,9). Cognitive decline, specifically in the domain of memory, is important as memory loss coupled with functional domain decline are characteristic features of age-related neurodegenerative disorders such as late-onset Alzheimer's disease (LOAD) and dementia (10,11). Little data on cognitive function and its decline in some individuals in the general populace is documented in SSA populations (12). Thus, assessing cognitive function and its decline in older adults is becoming vital as the elderly population in Africa expands.

1.2 Sample population

The Health and Ageing in Africa: A Longitudinal Study of an INDEPTH Community in South Africa (HAALSI) was established to assess the social and biological determinants of trends in mortality and morbidity associated with the demographic and epidemiological transition occurring in SA (13). This study, the first in SSA, was modelled on large

longitudinal ageing cohorts such as the United States Health and Retirement Study (US HRS) in the United States of America, the English Longitudinal Study of Ageing (ELSA) in the United Kingdom, and the Longitudinal Ageing Study in India (LASI) (13). This study was nested within the already existing health and socio-demographic surveillance system (HDSS) infrastructure of the MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt) (F***igure 1.1***) (13,14). The Agincourt HDSS site extends across thirty-one villages and some 120 000 people living in Bushbuckridge, rural Mpumalanga, SA (13,14). Five thousand and fifty-nine randomly selected individuals (men and women) aged 40 years and over - representative of the underlying Agincourt population - participated in the HAALSI study (13,14). This population-based study was initiated in 2014 and designed to have three-year intervals between each round of longitudinal follow-up in order to better understand factors influencing the health and wellbeing of this ageing rural population (13).

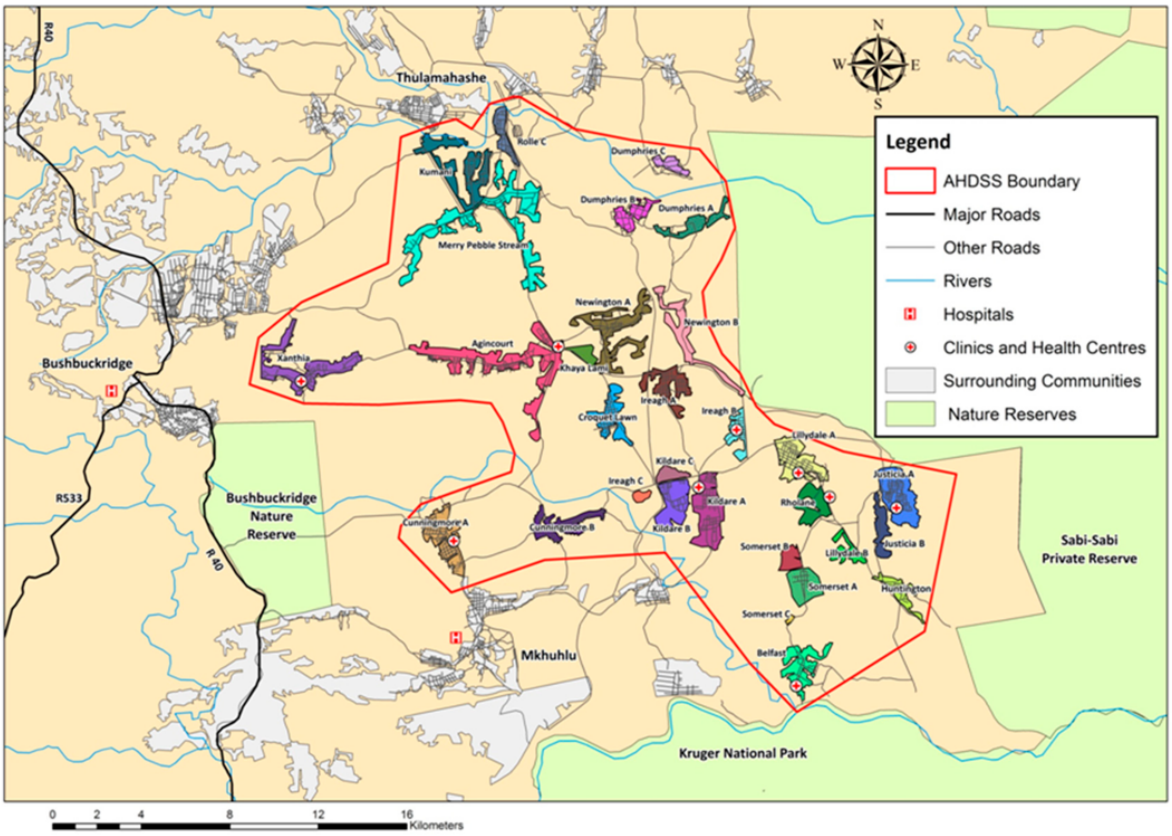


Figure 1.1: Agincourt HDSS Study Site

Mapped villages and communities (shown in different colours) fall within the boundaries of the Agincourt health and socio-demographic surveillance system site (red line). Image (http://www.agincourt.co.za/index.php/about/maps/).

Data collection was extensive, and the administered questionnaire featured detailed sections on demography, employment, social and psychological well-being, cognition, self-reported health indices and behaviours relating to cardiometabolic conditions and HIV/AIDS (substance use, physical activity, diet, sexual behaviour, adherence to medication, utilisation of available healthcare facilities), and household measures informative of socio-economic status (SES) (13). The study also included a number of anthropometric measures, and tests assessing physical frailty (13). Point of care tests for lipids, glucose, haemoglobin, and HIV (with feedback and counselling) were administered during the interviews (13). Blood spots were collected for HbA1c and C-reactive protein testing (13). If individuals tested positive for HIV, blood samples were taken to for confirmation testing and thereafter for assessment of viral load and traces of antiretroviral therapy (ART) (emtricitabine and lamivudine) (13). This cohort of rural-dwelling older adults is characterised by low levels of education (reflecting educational practices under *apartheid*) and high unemployment rate (13). Prevalence of HIV infection was 23%, and 69% of seropositive individuals confirmed being treated through ART (13,15).

A sub-group of the HAALSI sample contributed to the Africa Wits-INDEPTH Partnership for Genomic Studies (AWI-Gen) which falls under the Human Health and Heredity in Africa (H3A) Consortium. AWI-Gen was established with the intention to explore genetic and environmental risk factors for cardiometabolic diseases across six study sites in four different SSA countries (16). This study leveraged existing HDSS centres/sites which had been collecting longitudinal data for extended periods on large urban and rural communities in Burkina Faso, Ghana, Kenya, and South Africa (16). Detailed questionnaires were administered to consented research participants capturing data on demography, general health and infectious disease, NCD indices, SES variables, family history, and anthropometry (including ultrasound images) (17). Blood and urine samples were collected for various biochemical assays related to cardiometabolic health, as well as for DNA extraction and downstream analysis (17). AWI-Gen was designed to include men and women aged 40 to 60 years old, but older individuals were also included from the HAALSI study (17). Data on survivorship for 2014 in this community suggested that this dataset did not include the oldest members of the Bushbuckridge community (***Figure 1.2***) (5). Although this was not a focus of the research presented here, it is worth noting that the SA sites within the AWI-Gen study

had high prevalence of hypertension (~40%) which may have implications for vascular dementia in this study population (18,19).

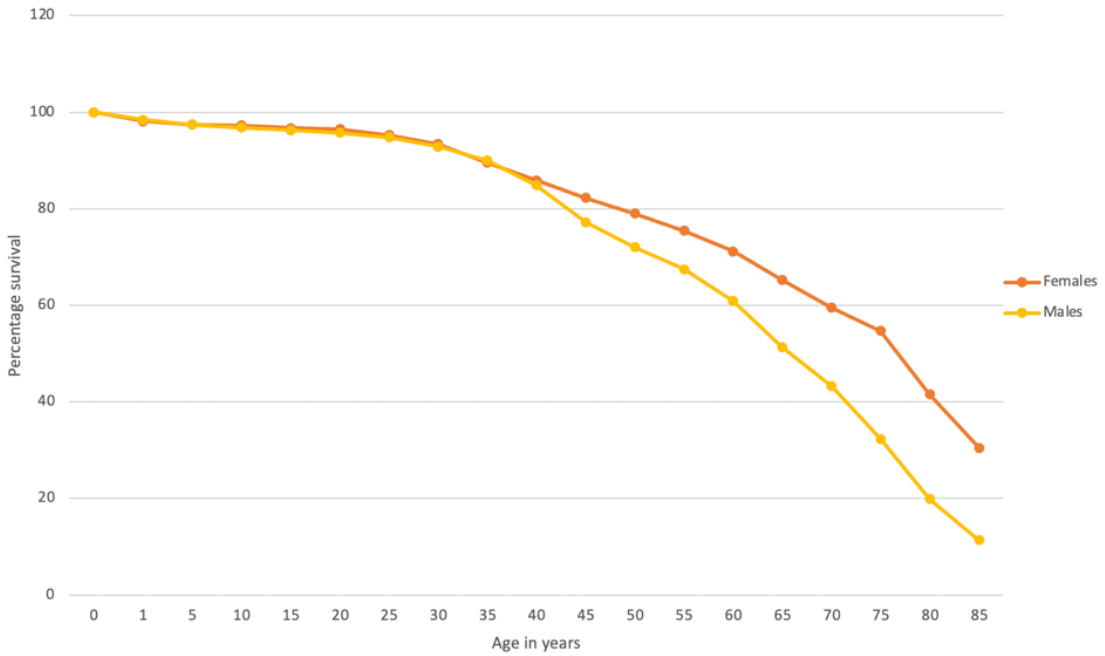


Figure 1.2: Age-specific survivorship curves for Agincourt HDSS

Data plotted from raw data supplied by Kabudula et. al. (2014)

1.3 Cognitive function

1.3.1 Definition

Human cognition had already interested scholars in the times of Aristotle, and although behavioural research in the field of psychology sought to understand the basis of intelligence in the 19th century, we are still met with significant challenges in the field of neuroscience today (20). Cognition is defined as the mental action, or process of acquiring knowledge and understanding through thought, experience, and perception (http://www.lexico.com [accessed 08/03/2021]). It is traditionally defined in terms of cognitive function based on the specific screening tools used to capture cognition at the individual level (20). Part of the difficulty in understanding cognition emerges from the biological complexity of the brain, which is further complicated by the complex nature of cognition itself. Individual cognitive performance, although highly heritable, is a product of the dynamic effects of genetic, molecular, structural, and environmental factors (21). As such, it is important to distinguish

between neurocognitive traits and have detailed and well-defined phenotypes especially when exploring possible correlations and associations in neuroscience.

Generally, individual cognition is determined by tests assessing intelligence quotient (IQ) which screen for degrees of mismatch within a set of predetermined criteria which are undeniably contingent on the individual's educational background and SES (22). Intelligence has further been classified into either fluid or crystallised intelligence (23). Crystallised intelligence is the ability to apply acquired knowledge and learned skills and otherwise access stored information (23,24). Hence tests of crystallised intelligence include those testing inter-individual vocabulary, numeracy, and general knowledge (24). Fluid intelligence relies on information processing and reasoning applied to novel stimuli (23,24). Tests of fluid intelligence measure those cognitive domains of memory, reasoning, processing speed, visuospatial ability, and executive function (24). Not only do these types of intelligence differ by definition and how they are measured, but there is evidence which suggests that crystallised intelligence is more stable across lifespan, and that differences in fluid intelligence contribute more to the variation of cognitive function in middle to older age (24).

It is well-known that education is significantly correlated with cognitive function, and most screening tools are biased toward educational attainment (22). This means that cognition screening tools are often inadequate when it comes to assessing levels of cognitive function in low-income settings where people have less access to formal education, and literacy levels are low (25). Another problem is that often cognition tests do not take into account language or cultural differences in study populations, which further adds to their inadequacy and bias in terms of accurate reporting (25).

It is also worth noting that many large meta-analyses use a derived phenotype, "general cognitive ability", or Spearman's "*g*," to assess differences in cognitive ability between individuals (26). Spearman's *g* is derived from the Wechlsler Adult Intelligence Scale (WAIS) which consists of tests for assessing verbal comprehension, perceptual reasoning, working memory, and processing speed (26,27). General cognitive ability utilises the correlation of different cognitive function tests in order to derive a factor from principal component analysis (PCA) of results in different studies to produce a single factor score representing cognitive function (1,26,28).

1.3.2 Description and data collection

For the purpose of this thesis five specific cognitive phenotypes (total cognition score, episodic memory, executive function, language, and visuospatial ability) are described in detail, along with the screening mechanisms employed to score inter-individual cognition in this cohort.

A questionnaire-based cognitive screen adapted from the US HRS to reflect cultural differences from its original setting was administered to participants of the HAALSI study (13). This screening tool was translated into the local dialect, Shangaan (or Xitsonga) spoken by the people in Bushbuckridge, and captured data assessing orientation, immediate and delayed recall, and numeracy (13). The original metric of this cognitive battery was a score from 0 to 26, where points were assigned per correct answer per section (2). Four points were assigned to participants if they could state the correct day, month, year, and name of the South African president at that time (2). A list of ten words were read aloud and participants were asked to recall as many words as possible, then scored accordingly for immediate recall (2). They were asked to repeat these words after the administration of a different test, and delayed recall out of 10 was scored after one minute (2). This made up the memory segment of the test scored out of 20. One point was given to participants who could count to twenty, and then another point was given if they could complete the numeric sequence: "2, 4, 6..." with "8" (2). Analysis of this metric revealed that it was still subject to educational bias, so the numeracy segment was dropped to remove the aforementioned bias (2). Therefore, total cognition score ranged from 0 to 24 for the downstream analyses of this cohort (2).

To overcome non-transferability of traditional testing methods, the tablet-based Oxford Cognition Screen Plus (OCS-Plus) was validated in this study population for use in low literacy settings (25). All testing stimuli were adapted to fit into the region's cultural context and were presented in the local dialect Shangaan (25). Oral responses were audio-recorded (25). Language tasks consisted of picture naming (participants were asked to name a series of four objects) and semantics (two sets of naming specific known objects, and two sets of identifying the object which did not fit into the semantic category presented) (25). Memory tasks included those for orientation (where participants were asked the same four questions related to orientation as per the questionnaire-based test), word encoding (five words were displayed and read out loud for participants to recall immediately afterwards, and then

repeated again for the second test of immediate recall), delayed recall and word recognition (respondents were asked to repeat the five words from memory after a five minute interval during which they performed another task for delayed recall, and then for each incorrectly remembered word presented with five semantically related words and tasked to recognise the word from the recall test), and incidental memory (participants were presented with three sets of 4 pictures and asked to select the object seen previously in the picture naming task, and then the last set consisting of four words of which one was from the previous recall tasks) (25). Tasks of executive function consisted of trails (connecting circles from smallest to largest, connecting squares from largest to smallest, alternating connecting circles and squares also alternating in size - largest square to smallest circle, etc.) and rule finding tests (participants were asked to predict the next movement of a red dot as it moved along a grid of randomly arranged shapes following either rules related to position or colour) (25).

Participants were given a practice round prior to performing tasks of executive function to ensure that they understood how to complete the tests (25). The OCS-Plus also measured auditory attention by playing three target words and then asking participants to tap the blue button when they heard the target word, or the yellow button upon hearing another word (25). Up to three practice rounds of this task were allowed, and participants only proceeded to testing if they had correctly identified target words in the practice rounds (25). Testing of visuospatial ability was performed through figure copy (a complex figure consisting of seven elements was presented at the top of the screen and participants were asked to copy it below) and figure recall (participants were asked to draw the figure from memory) where each element of the object was scored as either present or absent (29).

Methods used by Seidlecki, Honig, and Stern (2008) were employed to explore the underlying structure of the OCS-Plus screening tool (29). Exploratory factor analysis (EFA) of each of the tests in the OCS-Plus was performed in order to find a best fit model to determine the underlying structure of this data for classification of each domain (29,30). The final measurement model after EFA consisted of four factors namely language, executive function, verbal episodic memory, and visuospatial ability (29). The rest of the OCS-Plus variables (orientation, incidental memory, and auditory attention) were dropped from the model as they only had weak loadings, but in order to increase meaningfulness of executive function the rule finding task formed part of the loadings for this domain (29). Confirmatory

factor analysis (CFA) was performed on the resulting four factors to ensure that each variable only loaded onto the factor to which it had the highest loading fit (29,30). Intercorrelation between domain factors were determined, and only those correlations and factor loadings which had significant p values less than 0.001 are displayed in ***Figure 1.3*** which is a reproduction of Figure 2 in the publication by Farrell, *et al*. (2020). This explains how tasks contributed to loadings onto the four domains used for final analysis.

Standardised z-scores were derived through principal component analysis (PCA) for each individual using the indices from the resulting domain factor scores for each of the cognitive domains representing executive function, language, verbal episodic memory, and visuospatial ability for association analysis (29). This is important, as cognitive function in this cohort can only be representative of cognition within this sample and the underlying population. Thus, by not comparing the cognition in this sample to that of other population samples, the stigma attached to measures of IQ is avoided.

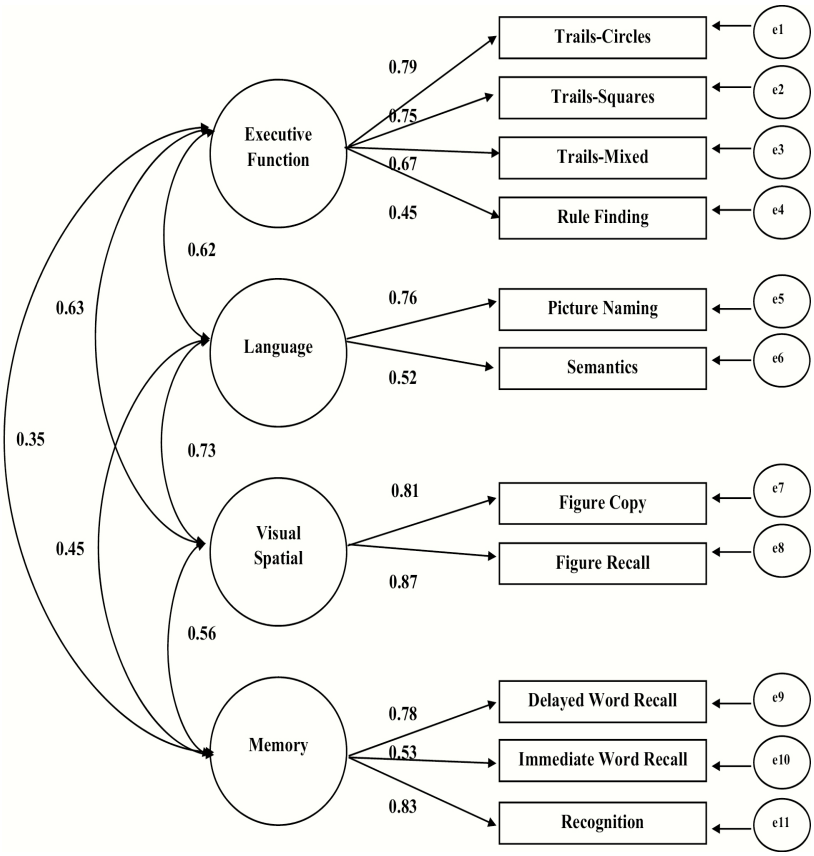


Figure 1.3: Details of intercorrelations and factor loadings for the final CFA domain model of cognitive variables from the OCS-Plus (from Farrell et al. (2020))

Standardised coefficients are shown for factor loadings of the EFA of each task loaded onto each of the four main factor domains. CFA shows intercorrelation between each of the domains where $p < 0.001$.

1.4 Genetic association studies

Genetic studies of cognitive function are often presented in relation to either cognitive decline or an associated pathological condition like Alzheimer's disease (AD), or as studies of intelligence through derived or proxy phenotypes (1,22). These studies are plagued with limitations arising from differences in ethnicity, study sample conditions, sample size, methodology, and screening instruments. Genetic variation due to population substructure is the basis for performing studies where research participants are of the same ethnicity, or ethnicity is adjusted for in the analyses. Much of the published genomic data for cognitive studies is based on either European or East Asian populations where large sample sizes and multiple time points of longitudinal data is available (1). Studies performed in African American (AA) populations may not be representative of, or transferable to other African populations (31).

1.4.1 African genetic variation

The African continent is the region of the world with populations that are most genetically diverse. Differences in admixture exhibited between populations are indicative of an ethnolinguistic cline; from the southern regions (where admixture with Khoe-San hunter-gatherer populations is observed) to northern regions (where substantial Eurasian admixture is observed) (32). Variation in genetic substructure is also observed between east, west, central, and southern Africa implying that marked variability driven by different selective pressures may not only be observed at a global level but within the continent as well (32,33). This is important as population differences in linkage disequilibrium (LD) patterns and differences in observed allele frequencies may impact study replication. The increased variation and smaller LD blocks in African population data may provide higher resolution fine mapping and improved detection of causal loci (32). This is of course, provided that there is sufficient coverage of common African variation in order to increase power of association through more accurate imputation (32).

A recent publication by Choudhury, *et al*. (2020) assessing the genetic variation across Africa (using high-depth whole genome sequence (WGS) data) confirmed that a large amount of rare and novel variation at multi-allelic sites observed in African populations had not been previously reported (33). Further annotation of these variants is required to provide better insight into the genomics of traits and disease states in Africa. They also confirmed that

although genetic data from Nigeria (west Africa) is the most well-characterised in light of their study, utilising these reference populations (e.g. Yoruba) provide poor proxies for other SSA populations (33). Hence, an Illumina genotyping array was developed by the H3Africa Consortium and the Pan African Bioinformatics Network for H3Africa (H3ABionet) (https://www.h3abionet.org/h3africa-chip) to capture common variation in African populations (33). A reference panel was also developed for imputation to aid African genomic research, using data generated by Choudhury, *et al*. (2020) (33).

1.4.2 Twin studies and heritability

In order to determine the heritability (genetic component) of cognition, as with other traits, studies of mono- and dizygotic (DZ) twins have been performed. Monozygotic (MZ) twins are regarded as genetically identical hence allowing researchers to isolate genetic versus environmental effects of complex traits. These studies involved recruiting large numbers of twin pairs to assess their IQ using various psychometric tests (differing according to country and study) over different time points (34). Twin studies compare the similarity between MZ and DZ twin pairs (DZ twin pairs only share fifty percent of their genetic makeup) to determine whether traits are more influenced by genetics (MZ twins more similar than DZ twins) or by shared and unique environmental factors which may have affected only one twin. The method for comparison is based on a model integrating additive genetics, common environment, and unique environment using maximum likelihood (34).

Twin studies of cognitive function have stated high levels of heritability especially for measures of IQ, although shared environment would also contribute to this similarity (1,34,35). A collaborative study consisting of twin cohorts in the Netherlands, Australia, and Japan, observed the heritability for IQ to range from 71-87%, but when examining distinct domain-related (processing speed and working memory) tasks heritability was lower (33-64%) (34). These results are similar to those observed in twin studies for AD, where heritability is estimated between 58% and 78% depending on whether it is early (familial) or late-onset AD (36). The heritability estimates are significantly lower in genome-wide association studies (GWAS) and candidate gene studies where only 24-33% of heritability has been attributed to genetic effects (36,37). This observed difference is largely due to differences in environmental factors which may influence the phenotype of complex traits (34,37). In twin studies, the variance decomposition is lower due to shared environment. Whereas in population studies gene-environment and gene-gene interactions may not

necessarily be controlled for, resulting in much lower estimates of the genetic contribution to the phenotype of interest (36).

1.4.3 Linkage and candidate gene studies

Linkage studies of cognitive function relied on using families with proband individuals diagnosed with neurocognitive disorders such as AD, dyslexia, and alcohol dependence, to identify chromosomal regions influencing multiple correlated cognitive measures (38–40).

The apolipoprotein E (*APOE)* locus on chromosome 19 was discovered through linkage analysis to be associated with susceptibility to AD (39,41). Evidence in support of linkage within a region on chromosome 19 for late onset familial AD was confirmed and later localised to 19q13 which contains *APOE* (39,42). A study examining microsatellite markers on chromosome 19 showed that no other genes within this region contributed to the linkage peak observed for AD supporting previous evidence that *APOE* was associated with AD risk (41). In the latest Genome Reference Consortium human build (GRCh38) *APOE* lies between 44905791-44909393bp on chromosome 19 (http://www.ensembl.org/) (43). It encodes a protein essential for cholesterol metabolism and transport within the central nervous system (44). *APOE* has three major allelic variants ε2, ε3, and ε4 (44). The resulting protein isoforms are characterised by non-synonymous amino acid changes at positions 112 and 158 (cysteine to arginine) where ε2 has cysteine at both positions, ε3 has an arginine at 158, and ε4 has arginine at both positions (44).

APOE allele frequencies are known to differ across ethnicities and will be discussed in detail in Chapter 2. Aside from its association with increased susceptibility to AD and earlier age of onset, ε4 has also been associated with normal brain structure and function and cognitive ageing without dementia (45–47). Clear associations have been observed in European populations, and lesser so in African American (AA) studies, but there has been little evidence from Africa supporting associations between ε4, cognitive function, or AD (48–50). Extant continental African studies often have small sample sizes and are underpowered to detect small effects on complex traits (12).

Another gene on chromosome 19, Translocase of Outer Mitochondrial Membrane 40 (*TOMM40*) has been associated with AD risk in European, Asian, and AA cohorts (8,51–53). This association has been linked with the length of a thymine (T) homopolymer tract found in

intron 6 of *TOMM40* (54–60). This variable length polymorphism is also suggested to affect cognitive function during normal ageing, cognitive performance in type 2 diabetes patients, and gray matter volume (55,57,60). The marker most used to represent this locus in the literature is rs10524523, which represents the insertion-deletion (INDEL) where the length is T = 20 (56,57,61). Alleles may be classified as short (S) where T ≤ 20, long (L) where 20 < T < 30, and very long (VL) where T ≥ 30 (54–60). The gene-gene interaction between *APOE* and *TOMM40* is of great interest as the available literature, mostly from European populations, describes near complete LD between the L allele and ε4 (54,56,58,62).

1.4.4 Genome-wide association studies

Advancements in technology over the last few decades have made it possible to analyse large amounts of genetic data, spurring the advent of GWAS and large meta-analyses (with sample sizes in the order of magnitude of six digits) increasing power to detect polygenic effects. Genotyping arrays tagging millions of SNPs across the human genome coupled with population-specific imputation panels inferred using WGS data have given researchers another method for detecting variation that contributes to complex traits and diseases (63).

GWAS of cognition have been performed mostly in large European Consortia such as Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE), the Cognitive Genomics Consortium (COGENT), and recently UK Biobank (UKBB) studies (1). These consortia have described variants associated with neurocognitive phenotypes ranging from general cognitive ability to cognitive decline and dementias, proxy phenotypes such as educational attainment, and contributed to studies of psychiatric disorders such as schizophrenia (21,24,64–66).

Despite significant findings, the importance of functional analysis and integration of neuroimaging is integral to understanding the biology and physiological effects of genetic and epigenetic interactions contributing to inter-individual cognitive performance (1). GWAS analysis and meta-analysis is a largely computational process once phenotypic and genomic data have been acquired. It requires computational and bioinformatics capacity and is often an automated process of various statistical analyses coupled with graphical representations of results for interpretation.

1.4.5 APOE SNP genotyping

Two single nucleotide polymorphisms (SNPs), rs429358 (19:44908684 T/C) and rs7412 (19:44908822 C/T), make up the haplotypes that determine *APOE* ε alleles where ε4 (where both alleles from the SNPs above are C) is the ancestral haplotype (referred to as an allele in this context) (67). TaqMan genotyping for *APOE* was performed in the Sydney Brenner Institute for Molecular Bioscience (SBIMB) Biobank and Research Laboratory for rs429358 and rs7412. These *APOE* markers are often absent from arrays due to failure to accurately genotype variants at rs429358 (68).

Although it is possible to impute these genotypes using reference panels from the 1000 Genomes Project, we chose to use SNP-genotyping as a cost-effective and efficient way to determine *APOE* genotypes for our study sample. Genotyping was validated by Sanger sequencing (by Inqaba biotech ™ (Gauteng, SA)) of the target region of thirty genotyped samples and chromatogram peaks were aligned with those results (**Appendix 2**).

1.4.6 GWAS analysis

DNA samples were assayed on the H3Africa genotyping array by Illumina (San Diego, CA, USA) and data quality control (QC) performed using standard metrics. The data was available from the AWI-Gen study and stored at the Sydney Brenner Institute for Molecular Bioscience (SBIMB), University of the Witwatersrand, Johannesburg. Along with the development of the array, H3ABioNet developed an automated GWAS pipeline using Nextflow (https://github.com/h3abionet/h3agwas) for the subsequent QC and analysis of array data from the H3Africa Consortium studies (69).

The standard QC process is a stepwise removal of SNP markers and individuals (***Figure 1.4)***. During sample QC, individuals with discordant sex data (genotypic versus recorded) and SNP missingness greater than 0.02 were excluded. Heterozygosity cut-offs were applied to exclude individuals with excessive heterozygosity (> 0.343) which is an indicator of genotyping errors, and low heterozygosity (< 0.15) which is an effect of inbreeding. Relatedness cut-offs (π^ - pi hat) were between 11 and 70% to account for cryptic relatedness observed in the sample and to retain as many individuals as possible. The bias in estimates of effects that this could create was mitigated by using Genome-Wide Efficient Mixed Model

Association (GEMMA) (70) (https://github.com/genetics-statistics/GEMMA) which has a built in Bayesian kinship matrix which adjusts for relatedness between individuals.

Parameters for the exclusion of SNPs included the removal of duplicate markers and non-autosomal SNPs from the dataset. Markers with a minor allele frequency (MAF) below 0.01 and greater than 0.99 (1-0.01) were removed, as were SNPs with missingness greater than 0.01. Markers that did not conform to genotype distributions in Hardy-Weinberg equilibrium (HWE) were removed as the assumption is that genetic variation within this sample has remained constant.

Specific QC parameters were applied to the dataset used in Chapter 3 which included modified cut-offs for the removal of both SNPs and individuals from the analysis. In order to increase the number of SNPs in the final analysis, imputation was performed as described and referenced in Chapter 3. Essentially this consisted of (1) pre-imputation QC, (2) imputation, and (3) post-imputation QC. Association analysis was also performed using the automated pipeline for GEMMA with the full dataset including imputed genotypes. This is further explained in Chapter 3.

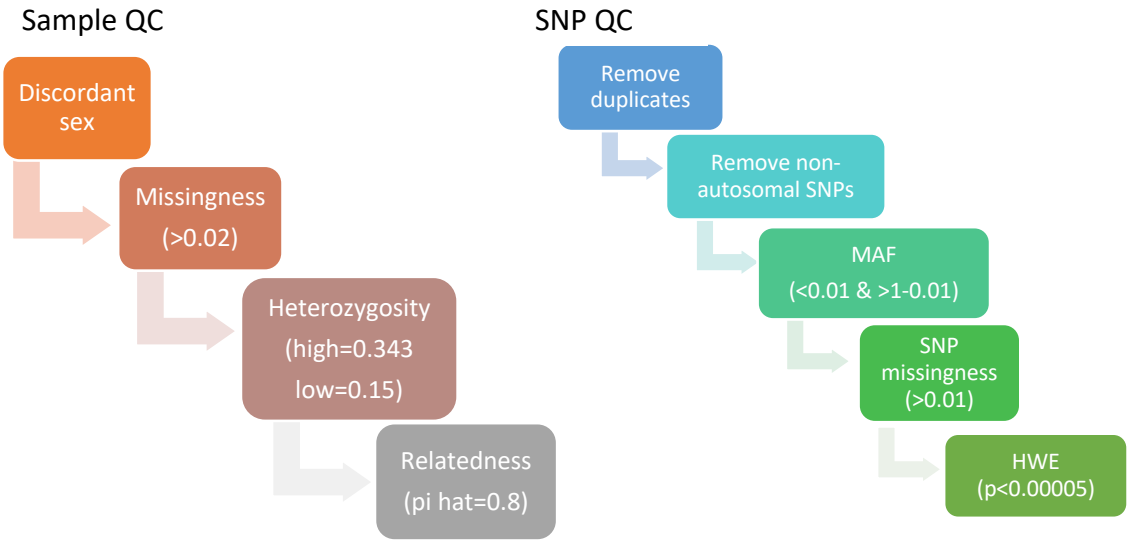


Figure 1.4: Stepwise GWAS QC process for automated pipeline

Sample QC parameters are shown on the left with cut-off values displayed in parentheses for the exclusion of individuals. SNP QC parameters are displayed on the right with cut-offs in parentheses for the removal of markers for subsequent analysis.

1.4.7 TOMM40 *allele length determination*

To determine the lengths of the *TOMM40* poly-T repeat, an in-house detection algorithm was developed by Dr Jean-Tristan Brandenburg specifically for this project. It was tested on a small subset of samples for which WGS data was available. Further details are provided in Chapter 4 of this thesis, where the algorithm was applied to existing 1000 Genomes Project data and then to the HAALSI/AWI-Gen sample subset.

1.5 Rationale, aims, and objectives

There is a distinct lack of African data on cognitive function. As the population demographic in Africa shifts to one where people are living to older age and the population over the age of 65 years grows, it is becoming increasingly important to include cognitive assessments in primary care regimes. Normal cognitive function declines with age, and in the case of dementias like AD is accelerated. In SA which has a high prevalence of people living with HIV who are susceptible to HIV-associated neurocognitive impairment, as well as those suffering from hypertension, regular testing of cognitive function would enable a better understanding of trends in cognitive ageing in the context of the current health and epidemiological transition.

The validation of a domain-specific cognitive function screening instrument; designed to be administered in lower income populations with less access to formal education, in the HAALSI cohort was integral in the provision of a large baseline sample with well-defined cognitive variables in a rural SA setting. Overlap with the AWI-Gen study provided an opportunity to add genomic data to this extensive phenotypic data. The availability of DNA samples for targeted genotyping and WGS, on top of the genotyped and imputed data from the H3A array set a stage to explore the influence of genetic and environmental factors on latent cognitive function.

This body of research was encompassed by the overarching hypothesis that the utilisation of well-defined and very specific phenotypes for assessing genetic variation influencing inter-individual cognitive function in an African population would reveal novel associations and replicate findings observed in other populations. The aim of this study was to explore the effects of genetic variation on five population-standardised cognitive phenotypes in a sample

of rural-dwelling South Africans aged 40 years and over of African ancestry from rural Mpumalanga Province, northeast South Africa. This aim was actioned through three objectives:

1. Determine *APOE* allele frequencies for study participants by SNP-genotyping methods and assess the effects of genotype on five cognitive performance metrics. We hypothesised that ε4 would have a detrimental effect on cognitive function resulting in ε4-carriers having lower performance scores than those that do not carry ε4 (**Chapter 2**).
2. Perform GWAS for each of the five cognitive function phenotypes (total cognition score, verbal episodic memory, language, executive function, and visuospatial cognition) in the AWI-Gen/HAALSI subset adjusting for the covariates age, sex, highest level of education, and HIV status. This was purely exploratory in nature so there was no previously formulated hypothesis (**Chapter 3**).
3. Determine *APOE/TOMM40* poly-T length polymorphism haplotypes and assess the effects of this haplotype on cognitive function, based on the hypothesis that the segregation of the long (L) allele with ε4 would confer negative effects on cognitive performance (**Chapter 4**).

The cognitive phenotypes used for analysis are described in detail in an earlier section, (1) a score for total cognition calculated by using an adapted cognitive battery resembling the one used for the US HRS, and standardised z-scores for each of the cognitive domains (2) executive function, (3) language, (4) verbal episodic memory, and (5) visuospatial ability. The phenotype data was collected by the HAALSI field team, and the cognition phenotypes and z-scores determined by M. Farrell and described in Farrell *et al*., Disparity in Educational Attainment Partially Explains Cognitive Gender Differences in Older Rural South Africans, *The Journals of Gerontology: Series B*, Volume 75, Issue 7, September 2020, Pages e161–e173, https://doi.org/10.1093/geronb/gbaa035. The *APOE* SNP genotyping was performed under my guidance, and I calculated the haplotypes to determine the epsilon alleles and genotypes for each participant. The bioinformatics algorithm for the imputation of the *TOMM40* alleles and genotypes was developed by Jean-Tristan Brandenburg. I interpreted the PCR data and assessed concordance between the STR-callers and the PCR

fragment analysis. I performed the statistical analysis and bioinformatics analyses for each of the three sub-studies, interpreted the results, and wrote the draft manuscripts.

The data collection and pipeline development used in the three studies presented herein involved collaborative efforts from different institutions, and cross-disciplinary expertise from contributing researchers. This is outlined in a schematic indicating the steps taken to complete this research (***Figure 1.5***).

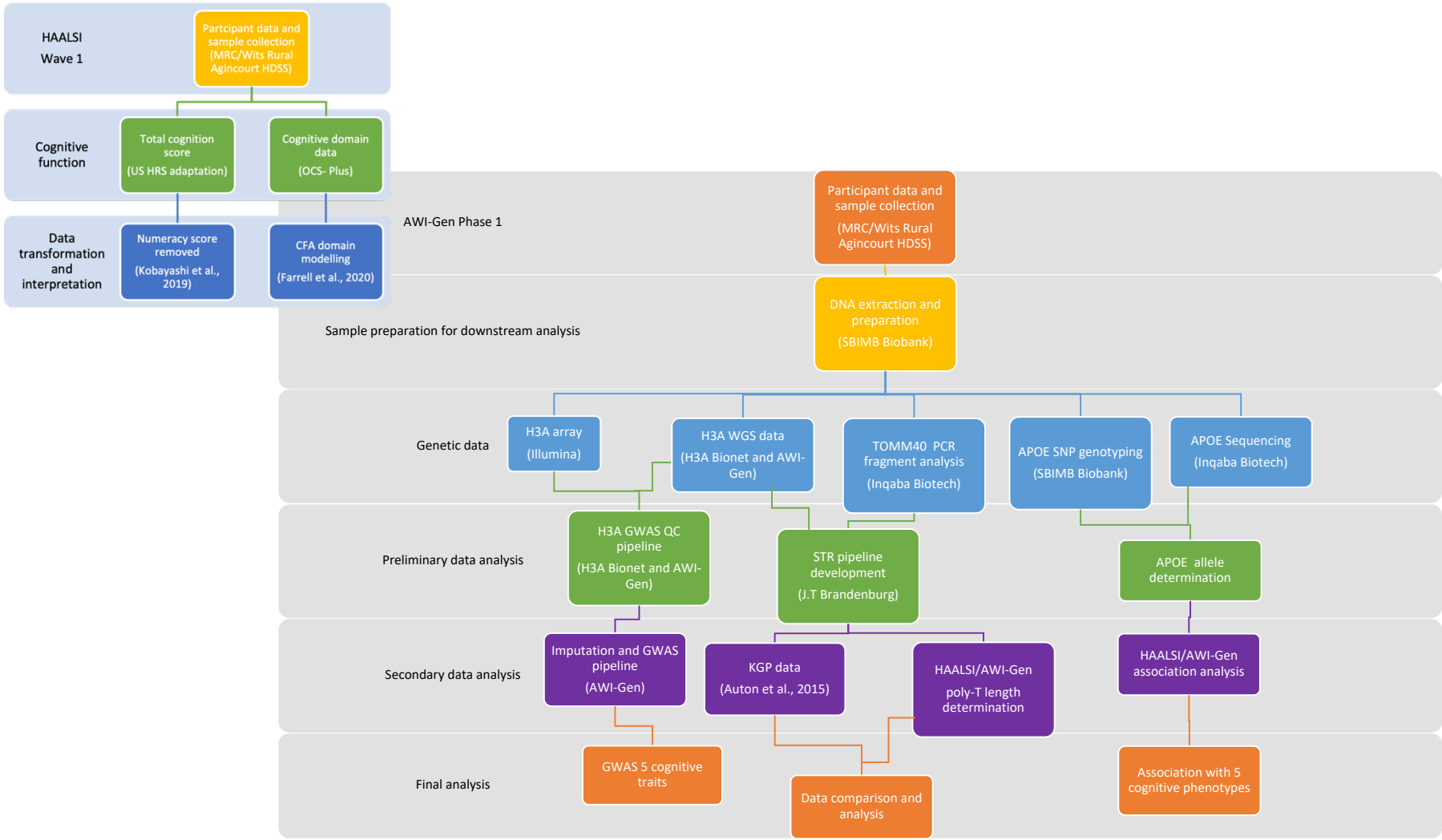


Figure 1.5: Research summary

Schematic diagram showing workflow from data collection to analysis.

Each of the three objectives makes up a chapter in this thesis. The first objective is presented as a published article and taking into consideration the comments made by the examiners. The second objective is presented in manuscript format and has been submitted to a peer-reviewed journal. Additional changes were made to the content to reflect suggestions made by the examiners for improvement. The third objective is presented as a stand-alone chapter detailing the work performed and a manuscript is still in draft stage.

GWAS catalog (https://www.ebi.ac.uk/gwas/efotraits/) results for the traits cognition, cognitive function measurement, verbal and visual memory, and memory performance informed GWAS power calculations using Quanto (71) for total cognition score as the representative continuous trait prior to analysis (***Figure 1.6***). A summary of effect sizes from various studies with a minimum value of β=0.0075 and maximum effect size of β=1.0 were used. Power analysis indicates that with a sample size of 2000, the power is insufficient to detect the effects of rare variants with a minor allele frequency of 0.01 or less, but this study should be adequately powered at 80% to detect common variation (MAF between 0.1 and 0.5) with effect sizes of β >0.2.

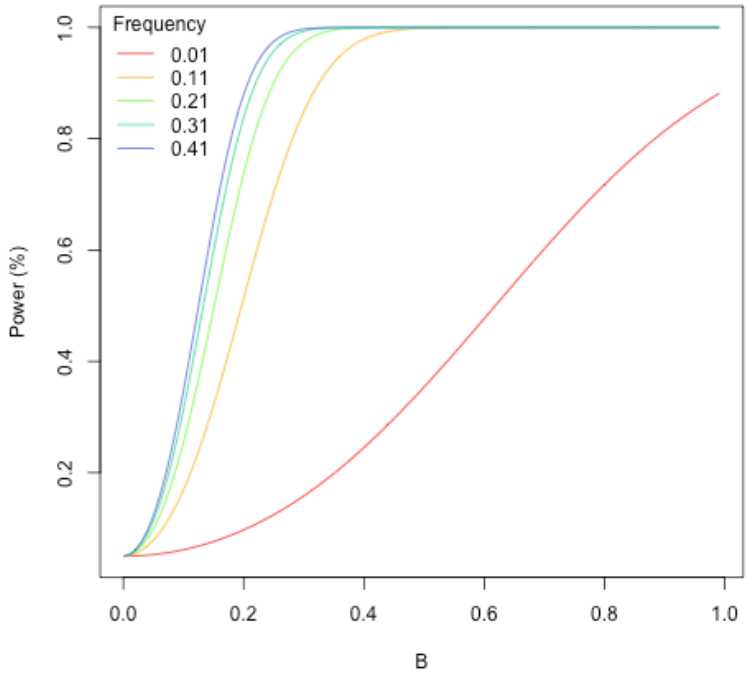


Figure 1.6: Power estimation for GWAS of total cognition score

Power estimates based on a fixed sample size of 2000 individuals for effect size β=0.0075 to 1.0 showing curves representing minor allele frequencies from 0.01 to 0.41.

1.6 Ethical considerations

Ethical approval at a provincial level for this study was granted by the Mpumalanga Provincial Research and Ethics Committee (approved: 2014/10/22) as the study included participants recruited from the Agincourt HDSS. Informed consent for the collection of data and biospecimens from the participants recruited for HAALSI was granted by the Harvard T.H. Chan School of Public Health, Office of Human Research Administration (C13-1608-02) and M141159 by the University of the Witwatersrand, Johannesburg, Human Research Ethics Committee (Wits HREC (Medical)). The AWI-Gen study was approved by the Wits HREC (Medical)) (M121029, and renewal M170880) for these same participants, as was independent ethical approval for this doctoral research (**M170916**) (**Appendix 1**).

Chapter 2

2 *APOE* genetic variation and its association with cognitive function in rural-dwelling older South Africans

2.1 Abstract

Apolipoprotein E (*APOE*) ε4 allele carrier status is well-known for its association with an increased likelihood of developing Alzheimer's disease, but its independent role in cognitive function is unclear. *APOE* genetic variation is understudied in African populations; hence, this cross-sectional study in a rural South African community examined allele and genotype frequencies, and their associations with cognitive function.

Cognitive function was assessed using two different screening methods to produce a total cognition score and four domain-specific cognition scores for verbal episodic memory, executive function, language, and visuospatial ability. Cognitive phenotype and *APOE* genotype data were used to determine whether *APOE* variation was significantly associated with cognitive function in this population.

Observed allele frequencies for 1 776 participants from the HAALSI study [age 40–80years (mean=56.19, SD=10.36); 58.2% female] were 58.1% (ε3), 25.4% (ε4) and 16.5% (ε2). Allele distributions were similar to the African super population, but different from all non-African super populations from the 1000 Genomes Project. The ε3 homozygous genotype was most common (34.9%) and used as the base genotype for comparison in regression models.

Four models were tested for each of the five cognitive phenotypes to explore association of *APOE* variation with cognitive function. In the first model assessing association with all genotypes for all individuals, marginally significant associations were observed for ε2 homozygotes where executive function scored higher by ~0.5 standard deviations (p=0.037, β=0.489, SE=0.23), and for ε3/ε4 heterozygotes where visuospatial ability scores were lower (p=0.046, β=-0.286, SE=0.14). These did not survive correction for multiple testing.

Regional African population differences were observed at the *APOE* locus. Marginally, significant associations between *APOE* genotype, and executive function and visuospatial ability indicate the need for larger studies to better examine these associations in African populations. Furthermore, longitudinal data could shed light on *APOE* genetic association with rate of change, or decline, in cognitive function with age.

2.2 Introduction

Studies conducted in different populations (mostly of European ancestry) have reported associations between apolipoprotein E (*APOE*) ε4 and various detrimental outcomes. These include susceptibility to, and severity of cardiovascular diseases, HIV infection and its associated comorbidities, late onset Alzheimer's disease, and recently COVID-19 (44,45,72–77). Aside from its well-known involvement in disease pathogenesis, there is interest in understanding the role of *APOE* variation in normal cognitive function throughout lifespan. This is especially true as studies begin to identify associations across the globe in heterogeneous populations (76).

Furthermore, cognition is a complex multifactorial trait and studies examining normal cognitive ageing and brain morphology have observed associations with *APOE* (45–47,78–80). These results have however, been inconsistent, and larger meta-analyses have observed small, significant differences in cognitive performance and brain morphology in European populations between ε4-carriers and non-ε4-carriers (45,78,79). Studies focused on Alzheimer's disease (AD) have shown that the presence of ε4 is associated not only with increased susceptibility to AD, from 2-fold to 33-fold depending on the number of risk alleles and ethnicity, but also with an earlier age of onset of the disease (46,79,81,82). It has also been found to be associated with other neurocognitive phenotypes, including mild cognitive impairment (MCI), HIV-associated neurocognitive disorder (HAND), brain atrophy and structural pathologies, and the pace of age-related cognitive decline (45,50,73,80,82–85). However, the results of these studies were mixed, as were the strengths of associations and effects for different ethnic groups (75,79,80,84). This may be due to differences in study design, sample size, characterisation and collection of cognition data, and genetic diversity due to differences in ethnicity.

The *APOE* alleles ε2, ε3, and ε4 encode structurally different protein isoforms with variable molecule-binding affinity (86). APOE is involved in lipid metabolism, specifically of cholesterol, and is integral to cholesterol transport and clearance within the central nervous system (46). Global variation in *APOE* allele frequencies have been reported between ethnic groups (67,86,87). The ε3 allele is the most common allele across all populations, and ε4 frequencies are higher in African and aboriginal populations than in Europeans and Asians (67,81,86,88). Despite this, evidence for its association with AD and other neurocognitive phenotypes in populations of African ancestry is either lacking, or very weak (48–50,81,83,89–91).

Genomic studies of cognitive function among Africans are limited by small sample sizes and low transferability of cognitive performance tests in regions where access to formal education and lower levels of literacy may affect reproducibility of results observed in largely European cohorts (2,25,29,44). The Health and Aging in Africa: A Longitudinal Study of an INDEPTH Community in South Africa (HAALSI) was established to assess trends in socio-demographic determinants, morbidity, and mortality associated with population greying and the current epidemiological and behavioural transition underway in South Africa (13,19). Cognitive performance data were captured for over 5 000 individuals aged 40 years and older from rural communities in Bushbuckridge, Mpumalanga, SA, a region known to have a high prevalence of HIV (13,15,92).

Two screening tools were used to capture cognitive function: (1) a culturally adapted version of the United States Health and Retirement Study (US HRS) screening tool translated into the local language Shangaan and (2) the tablet-based Oxford Cognition Screen (OCS-Plus), which was adapted for and validated in this study population (2,13,25,29). It provides domain- specific scores for cognitive function in episodic memory, executive function, language, and visuospatial ability without bias to levels of literacy or numeracy (2,25,29).

Here, a subset (n = 1 776) of the larger HAALSI cohort was genotyped for the *APOE* locus. Our aim was to examine the associations of *APOE* with total cognition score and specific cognitive domains adjusting for known moderators, such as age, level of education and sex. HIV status was also included as a covariate, and we sought to explore a possible relationship between HIV status and *APOE* in this cohort.

2.3 Materials and Methods

2.3.1 Study design and participants

Demographic, socioeconomic, health and cognitive function data were collected from 5 059 adults (men and women, aged 40 years and older) recruited from Bushbuckridge, rural Mpumalanga, South Africa, from November 2014 to November 2015, as the baseline for the HAALSI longitudinal study (13). African participants who had been living in the area for at least 12 months preceding data collection were selected from the Health and Demographic Surveillance System database and consented (14). Cognitive function variables, HIV status, age, sex, and educational attainment expressed as highest level of education (none, primary, secondary, and tertiary) were used in this population-based cross-sectional community study. Sample size differed per outcome due to missing data.

Approximately 2 000 of these same participants overlapped with the Africa Wits-INDEPTH Partnership for Genomic Studies (AWI-Gen) for which blood samples were collected for DNA extraction and genetic analyses (16,17). Data collection and informed consent templates for the HAALSI and AWI-Gen studies were approved by the University of the Witwatersrand, Johannesburg, Human Research Ethics Committee [Wits HREC (Medical); M141159, M121029], the Harvard T.H. Chan School of Public Health, Office of Human Research Administration (C13-1608-02), and the Mpumalanga Provincial Research and Ethics Committee (approved: 2014/10/22). Ethics approval for this study was independently granted by Wits HREC (Medical) (M170916). All research participants from the HAALSI/AWI-Gen study sample provided written consent for their data and samples to be used.

2.3.2 Cognition screening tools and data

Two different tools were utilised to collect data on cognitive function (2,13,29). The first tool was adapted from the US HRS cognition screening tool. It was translated into the local language, Shangaan, and then back-translated to ensure cultural appropriateness and transferability (2,29). It included questions assessing the following measures: orientation, immediate and delayed word recall, numeracy, self-rated memory, and literacy (2,13,29). A score from 0 to 24, defined here as, ‘total cognition score’, representing global cognition, was used as one outcome variable for this study. The OCS-Plus was developed and then validated in a subset of HAALSI participants to capture cognitive domain- specific data without

literacy or numeracy bias (25). Episodic memory, language, executive function, and visuospatial cognition data were collected and transformed into z-scores for each individual per domain thus providing standardised domain-specific scores for this population using methods described by Seidlecki et al (2008) (29). The sample size was larger for total cognition score than for the domain outcomes.

2.3.3 DNA extraction and APOE genotyping

DNA was extracted from buffy coat samples using the QIAsymphony SP (QIAGEN GmbH, QIAGEN Strasse 1, 40, 724 Hilden, Germany) at the Sydney Brenner Institute for Molecular Bioscience (SBIMB). TaqMan™ SNP genotyping was performed using DNA from 1 952 individuals for rs429358 and rs7412, of which 1 776 were successfully genotyped at both loci. Genotyping was performed on the QuantStudio™ 5 Real-Time PCR System (Applied Biosystems, Thermo Fisher Scientific, Carlsbad, United States), and results were analysed using the QuantStudio™ Design & Analysis Software. *APOE* alleles were determined by using the genotypes for the SNPs rs429358 (T, C) and rs7412 (C, T). When the combined haplotype was TT, this represented the $\varepsilon 2$ allele. When the haplotype was TC $\varepsilon 3$, and when the haplotype was CC, this was interpreted as $\varepsilon 4$.

SNP data for 2 504 individuals from the 1000 Genomes Project (Phase 3) were downloaded from the Ensembl genome browser for the Genome Reference Consortium Human Genome build 37 (GRCh37; ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/ phase3/data; Accessed 11 Feb, 2021; Auton et al., 2015; 1000 Genomes Project and AWS, RRID:SCR_008801). Samples were divided according to super population codes as representative of European (EUR), African (AFR), East Asian (EAS), South Asian (SAS), and Admixed American (AMR) ancestries (Auton et al., 2015; 1000 Genomes Project and AWS, RRID:SCR_008801). *APOE* alleles were determined as above, and frequencies for each of these super populations were compared to our data.

2.3.4 Statistical analysis

Allele and genotype frequencies were calculated, and Hardy- Weinberg equilibrium (HWE) was tested using the Pearson's Chi-squared test. Chi-squared tests were used to test for differences in *APOE* allele distribution and categorical measures of age, sex, level of education, HIV status, each of the 1000 Genomes super populations, and data from four previous African studies (Willis et al., 2003; Gureje et al., 2006; Chen et al., 2010 and Joska

et al., 2010). The allele frequency data are represented as pie diagrams on the backdrop of a map of Africa created in MapChart[1] (2021).

To assess correlations between *APOE* distribution and age, age was categorised firstly as 10-year intervals, and then as younger (40–59 years) vs older (60 years and older). Descriptive statistics was used to summarise continuous (age, total cognition score, and domain-specific z-scores) and categorical variables (HIV status and level of education). Two-sample hypothesis testing was used to further draw out the observed association between *APOE* and HIV. This was followed by logistic regression to determine the effects of each allele on HIV status while adjusting for the presence or absence of one of the other alleles.

Cognitive function, either as total cognition score or as one of the four above-mentioned domains, were used as outcome variables to test for associations. Linear regression models for each cognitive outcome were tested. The four models were as follows: (1) regression using six known genotypes (2) regression comparing $\varepsilon 3$ homozygotes to $\varepsilon 4$-carriers (3) regression comparing $\varepsilon 4$ homozygotes to the rest of the genotypes and (4) regression comparing $\varepsilon 2$ homozygotes to $\varepsilon 4$ homozygotes.

Chi-squared tests, two-sample proportion testing and logistic regression analysis reporting McFadden's R^2 were performed using STATA 15.0.585 (StataCorp, 2017 Stata Statistical RRID:SCR_012763), and summary statistics, regression analyses and correction for multiple testing were performed using R (R Core Team, 2020. R: A language and environment for statistical computing; R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/; R Project for Statistical Computing, RRID:SCR_001905). Bonferroni correction was applied for multiple testing.

[1] https://mapchart.net/africa.html

2.4 Results

The phenotypic characteristics and covariates of the study population along with the number of participants (n) with data for each variable are shown in **Table 2.1**. The mean age was 56.19 years (SD=10.36), and there were more female than male participants (58.22%). Education levels in the community were low, with the majority having no formal education or having only attended primary school (74.66%). HIV infection was high with 28.48% of the study population living with HIV. Total cognition score was normally distributed with a mean of 11.98 (SD=4.4). The four standardised domain scores showed considerable variability. The range was particularly wide for visuospatial cognition (–10.627 to 4.771) and executive function (–4.687 to 4.328).

Table 2.1: Summary characteristics of study population

Measure	***n***	***Range***	***Mean or percentage***	***SD***
Age (years)	1776	40 to 80	56.19	10.36
Sex	1776			
Female			58.22%	
Male			41.78%	
Level of education	1772			
No formal education			36.00%	
Primary			38.66%	
High			21.11%	
Tertiary			4.24%	
HIV status	1773			
Negative			71.52%	
Positive			28.48%	
Cognitive measures				
Total cognition score	1776	0 to 24	11.98	4.44
Cognitive domains	1502			
Executive function		-4.687 to 4.328		
Episodic memory		-3.824 to 1.456		
Language		-2.533 to 1.279		
Visuospatial cognition		-10.627 to 4.771		

Sample number (n) differs per covariate/measure due to data missingness. Z-scores for cognitive domain data are represented only as ranges. Means and standard deviation (SD) are shown for continuous variables and percentages for categorical ones.

In total, 1 776 individuals were successfully genotyped, and all six genotypes were observed with ε2 homozygotes having the lowest frequency (2.9%) and ε3/ε3 the highest (34.9%) (**Table 2.2**). The genotype distribution was in Hardy Weinberg equilibrium (p=0.28). Our sample had a similar allele frequency distribution to that of the AFR super population group of Phase 3 of the 1000 Genomes Project ($\chi 2$=1.0983, df=2, p=0.35) (**Figure 2.1**). The frequencies of ε2 and ε4 were higher in our sample than in the other populations (EUR, EAS, SAS and AMR representing South American ethnicities). When compared to other African studies (**Figure 2.2**), a significant difference in allele frequency distribution was observed between our data and that of Chen et al., 2010 in a study from Kenya ($\chi 2$=11.058, df=2, p=0.004), but due to the small sample size, this difference may not be a true representation of allele frequency differences between east and southern Africans.

Table 2.2: *APOE* allele and genotype frequencies

Allele/Genotype	***Number (Frequency)***
ε2	587 (0.165)
ε3	2064 (0.581)
ε4	901 (0.254)
Total	**3552 alleles**
ε2/ε2	51 (0.029)
ε2/ε3	330 (0.186)
ε2/ε4	155 (0.087)
ε3/ε3	619 (0.349)
ε3/ε4	496 (0.279)
ε4/ε4	125 (0.070)
Total	**1776 individuals**

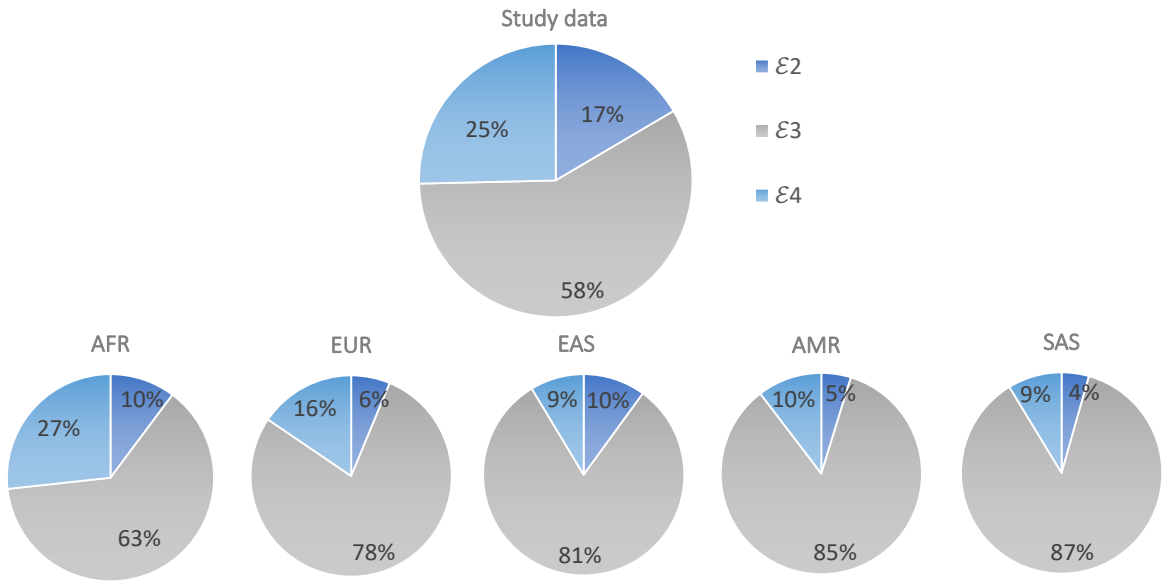


Figure 2.1: Comparison of *APOE* allele distribution in our dataset (Study data) compared to 1000 Genomes super population datasets AFR, EUR, EAS, SAS, and AMR

Pearson's Chi-squared test was used to assess differences in *APOE* allele frequency between sample populations from each of the 1000 Genomes super populations and our study sample.

APOE allele distribution showed no correlation with sex ($\chi2=1.3329$, df=2, $p=0.514$) or level of education ($\chi2=5.7393$, df=6, $p=0.453$). No significant difference in allele frequency was observed between age categories per decade nor for younger vs older individuals ($\chi2=2.0264$, df=6, $p=0.917$, and $\chi2=0.4175$, df=2, $p=0.812$, respectively). There was evidence of a relationship between HIV status and *APOE* allele frequency ($\chi2=6.9266$, df=2, $p=0.031$). Results of hypothesis testing indicated that the proportion of ε3 alleles was higher in HIV positive individuals than in those without HIV ($p=0.001$, OR=1.149 [95% CI, 0.82-1.60]) and the opposite was observed for ε4 ($p=0.007$, OR=0.843 [95% CI=0.66-1.08]). Logistic regression results showed the size and direction of the effects of each allele on HIV status, but these associations were not significant as our model had little predictive value (McFadden's pseudo $R^2=0.0023$; **Table 2.3**).

Table 2.3: *APOE* allele association with HIV status

Allele	*HIV + (n)*	*HIV - (n)*	*OR [95% CI]*	*z*	*p*
ε2	0.163 (165)	0.166 (421)	0.997 [0.77-1.29]	-0.02	0.980
ε3	0.612 (618)	0.569 (1443)	1.149 [0.82-1.60]	0.82	0.412
ε4	0.225 (227)	0.265 (672)	0.843 [0.66-1.08]	-1.37	0.172

Logistic regression was performed to determine size and direction of effects of *APOE* on HIV status.

Four multiple regression models were applied, to allow for comparative analyses with other studies, to each of the five cognitive phenotypes. Each model was representative of the comparison between *APOE* genotypes (Model 1: all genotypes; Model 2: ε4-carriers vs ε3 homozygotes; Model 3: the rest of the genotypes vs ε4 homozygotes; and Model 4: ε4 homozygotes vs ε2 homozygotes) and their effect on cognition while adjusting for covariates (age, sex, level of education, and HIV status). Age and level of education were significantly associated with cognitive function in all the models. Sex was significant for executive function, language, and visuospatial cognition for models 1–3 where being male increased the standard estimates for each domain (**Table 2.4** and **Supplementary table S2.1**). Significant associations were observed in Model 1 (**Table 2.4**) for *APOE* genotypes, where ε2/ε2 was associated with higher levels of executive function by approximately half a standard deviation (β=0.489, SE=0.57, p=0.037) and lower levels (but not significantly) when participants were heterozygous ε2-carriers, and ε3/ε4 was associated with lower levels of visuospatial cognition (β=–0.286, SE=0.14, p=0.046). These effects (f^2=0.395 and f^2=0.597, respectively) were not significant after Bonferroni correction for multiple testing (p=0.337 and 0.416, respectively).

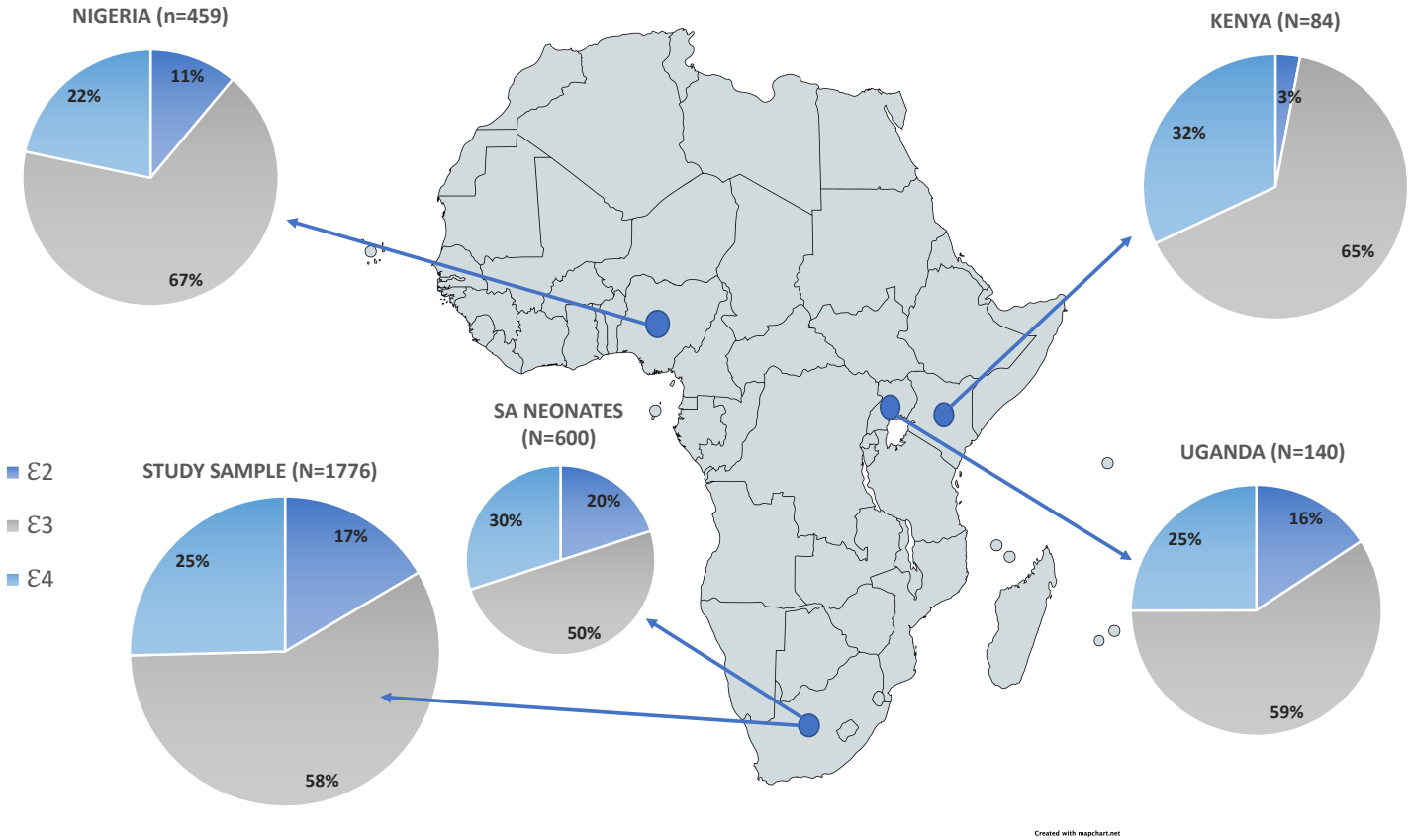


Figure 2.2: Comparison of *APOE* allele frequency distributions from four African studies, showing sample size for each study

Pearson's Chi-squared test was used to assess differences in *APOE* allele frequency between sample populations from each of the African studies referenced by the country of origin/description: Uganda: Willis et al., 2003; Nigeria: Gureje et al., 2006; Kenya: Chen et al., 2010; SA neonates: Joska et al., 2010.

Table 2.4: Linear regression model 1 using all genotypes to estimate the effects of *APOE* and covariates on cognitive function

	Total cognition		**Executive function**		**Language**		**Memory**		**Visuospatial cognition**	
Model	*Coefficent*	*SE*	*Coefficent*	*SE*	*Coefficent*	*SE*	*Coefficent*	*SE*	*Coefficent*	*SE*
1 - All genotypes										
Ɛ3/Ɛ4	-0.306	0.24	-0.042	0.10	-0.048	0.04	-0.062	0.06	-0.286*	0.14
Ɛ3/Ɛ2	-0.134	0.27	0.024	0.11	0.013	0.04	0.021	0.07	-0.063	0.16
Ɛ2/Ɛ4	-0.107	0.35	-0.025	0.15	-0.015	0.05	0.066	0.09	-0.060	0.22
Ɛ4/Ɛ4	0.114	0.39	-0.066	0.16	-0.058	0.06	-0.036	0.09	-0.164	0.24
Ɛ2/Ɛ2	0.434	0.57	0.489*	0.23	0.155	0.09	0.099	0.14	0.470	0.35
Age	-0.093***	0.01	-0.037***	<0.01	-0.013***	<0.01	-0.029***	<0.01	-0.081***	0.01
Male	-0.018	0.19	0.264*	0.08	0.059*	0.03	-0.05	0.05	0.383*	0.12
Education	1.758***	0.12	0.823***	0.05	0.355***	0.02	0.349***	0.03	1.394***	0.07
HIV positive	0.017	0.21	-0.004	0.09	-0.025	0.03	-0.041	0.05	-0.175	0.13
Adjusted R-squared	0.221	3.91	0.283	1.48	0.325	0.54	0.245	0.87	0.374	2.20

*Standardised coefficients represent estimated changes in cognitive function. Domain estimates (executive function, language, memory, visuospatial cognition) change in standard deviation units. Adjusted R-squared values per model with residual standard error (SE) is shown. *p < 0.05, **p < 0.00001, ***p < 2e-16.*

Supplementary Table 2.1: Linear regression models estimating the effects of APOE and various covariates on cognitive function

	Total cognition		**Executive function**		**Language**		**Memory**		**Visuospatial cognition**	
	Coefficent	*SE*	*Coefficent*	*SE*	*Coefficent*	*SE*	*Coefficent*	*SE*	*Coefficent*	*SE*
2 - Ɛ3/Ɛ3										
Ɛ4-carriers	-0.195	0.21	-0.039	0.09	-0.042	0.03	-0.033	0.05	-0.221	0.13
Age	-0.097***	0.01	-0.037***	<0.01	-0.013**	<0.01	-0.030***	<0.01	-0.080***	0.01
Male	-0.157	0.22	0.238*	0.09	0.051	0.03	-0.054	0.05	0.385*	0.13
Education	1.701***	0.14	0.840***	0.06	0.358***	0.02	0.364***	0.03	1.399***	0.08
HIV positive	0.021	0.24	0.02	0.10	-0.015	0.04	-0.028	0.06	-0.095	0.15
Adjusted R-squared	0.219	3.94	0.291	1.48	0.338	0.53	0.267	0.85	0.379	2.19
3 - Ɛ4/Ɛ4										
Rest of cohort	-0.230	0.37	0.070	0.15	0.049	0.06	0.029	0.09	0.070	0.23
Age	-0.093***	0.01	-0.037***	<0.01	-0.013***	<0.01	-0.029***	<0.01	-0.080***	0.01
Male	-0.037	0.19	0.261*	0.08	0.056	0.03	-0.053	0.05	0.367*	0.12
Education	1.763***	0.12	0.822***	0.05	0.355***	0.02	0.350***	0.03	1.397***	0.07
HIV positive	0.026	0.21	-0.002	0.09	-0.023	0.03	-0.037	0.05	-0.161	0.13
Adjusted R-squared	0.222	3.91	0.282	1.48	0.324	0.54	0.246	0.87	0.372	2.20
4 - Ɛ2/Ɛ2										
Ɛ4/Ɛ4	-0.464	0.61	-0.532	0.28	-0.214	0.11	-0.149	0.17	-0.648	0.46
Age	-0.105*	0.03	-0.025	0.01	-0.011	0.01	-0.029*	0.01	-0.071*	0.02
Male	0.767	0.58	0.344	0.26	0.163	0.11	-0.046	0.17	0.845	0.44
Education	1.916**	0.35	0.870**	0.16	0.362**	0.06	0.418*	0.10	1.489**	0.26
HIV positive	-1.286	0.69	0.117	0.32	-0.052	0.13	-0.148	0.20	-0.358	0.53
Adjusted R-squared	0.262	3.63	0.257	1.51	0.263	0.62	0.221	0.95	0.315	2.50

*Standardised coefficients represent estimated changes in cognitive function. Domain estimates (executive function, language, memory, visuospatial cognition) change in standard deviation units. Adjusted R-squared values per model with residual standard error (SE) is shown. *p < 0.05, **p < 0.00001, ***p < 2e-16.*

2.5 Discussion

This cross-sectional study assessed the potential effect of *APOE* genetic variation on five measures of cognition in a rural- dwelling SA population. Participants were adults of African ethnicity aged 40 years and older. We aimed to address the lack of African data for large cohorts that may be used to examine the genetics of cognition and cognitive decline, specifically in rural areas where vulnerability to cognitive impairment and dementia is becoming a major public health concern (2).

Association studies have observed genetic variation in neurocognitive traits between different ethnicities (1,44). Here, we determined *APOE* allele frequency distribution for our dataset and then compared this to existing data from other population groups. The genetic contribution to cognitive function is complex, and the tests used to define levels of cognition have made equitable comparisons between studies difficult (1,26,44). This is especially challenging in a low literacy setting despite the development of tests to assess cognition regardless of levels of literacy. Nonetheless, both literacy and level of education are still associated with cognition in this study cohort (2,29).

Traditionally, intelligence has been measured using the intelligence quotient (IQ) or Spearman's *g* (used in meta- analyses), which are biased towards access to formal education (as they focus mostly on aspects related to literacy and numeracy) (1,26). Here, we focused on latent cognitive ability which encompasses domain-specific and global cognitive function outside of educational attainment (1,26). Collecting cognition data is further complicated by using screening tools which have not been validated for transferability across populations; therefore, the OCS-Plus tool was developed to provide an alternative method for capturing cognition data in low-income populations that was neither reliant on literacy nor numeracy (25). The HAALSI baseline study provided an appropriate setting in which to validate this tablet-based tool, supplying population- specific cognitive function data for an understudied and socially homogeneous population (13,25). Leveraging these data and further capitalising upon the overlap of AWI-Gen participants for which DNA was available, we sought to explore the genetic effects of *APOE* on cognitive function.

APOE genotyping revealed that ε2 and ε4 are generally more common in African populations, including in our own, with SA and Uganda having the highest frequencies of ε2 among African populations (**Figure 2.2)**. Although these allele frequencies differed significantly from the Chen et al., (2010) cross-sectional study of a population in Kenya, the small sample size of this Kenyan study was insufficient for accurate comparison of intra-continental variation of *APOE*. The higher ε4 frequency in Africans has been attributed to selective advantage as protection against infectious disease (86,93–95). *APOE* ε4 has been suggested to be protective against hepatitis C infection as well as carriers being observed to have a better prognosis post-infection (95–97). There is also evidence that ε4 inhibits growth of *Plasmodium falciparum* conveying a protective effect against malaria (93–95). Higher observed ε4 frequency in Africans has also been attributed to the hypothesis that it is beneficial for fertility and early infant survival (94,95,98). Despite these proposed early life advantages, it has also been associated with increased morbidity and mortality later in life (44,45,72–77,94,95). Smaller African studies mainly assessing *APOE* ε4 association with AD showed no significant associations (48,89,90,98). The genetic associations between ε4 and neurocognitive phenotypes (mostly AD due to its highly replicated association with increased susceptibility) have observed weaker effects in those of admixed African ancestry (AA and Brazilians) and lack of association in Africans (45,49,50,78,79,89,99).

Previous epidemiological studies on the HAALSI cohort have reported the effects of age, sex, level of education, and HIV status on these same cognitive phenotypes (2,15,29). Higher educational attainment and younger age are known to be associated with better cognitive performance, and we observed these same education-related effects across all cognitive phenotypes in our study and in our regression models. In this study population, we observed better performance in the domains of executive function and language in male participants, a finding that had already been published by Farrell et al. (2020) who attributed this sex difference to socio-cultural differences and access to education (2). Other studies have suggested that older women generally have better cognitive performance than their male counterparts possibly due to the effects of oestrogen or differences in cognitive reserve (84,100–103). Sex effects appear to be domain-specific, and this has been observed in studies where women performed better in tests assessing memory than men irrespective of educational background (102,103).

For this study, we used HIV status as a covariate because of the high prevalence of HIV and its known contribution to cognitive impairment through HAND (46,75,83,104–106). Upon assessment of the *APOE* distributions for each of the covariates, we observed statistical differences between allele distribution according to HIV status. When modelling the effects of each allele on HIV status, logistic regression indicated that our model consisting of HIV status conditional on *APOE* allele was inadequate, which could be due to small sample size. Although the proportion of ε3 alleles was significantly higher in HIV positive individuals, the relative effect of this allele on HIV status compared to the effects of the other alleles was not significant. This is contrary to another SA study reporting lower ε3 allele frequency in HIV positive adults vs newborn controls (98). The opposite relationship for ε4, with a higher proportion observed in those without HIV, was also not significant, but our results suggested that ε3 is the risk allele for HIV infection. *APOE* ε4 has been linked to AIDS severity and increased mortality especially in ε4 homozygotes (73,82,83,107). It has been associated with faster disease progression and higher viral load in seropositive individuals due to enhanced entry of the virus into T cells (73,96). This may explain why we see a higher proportion of ε3 in our HIV positive sample, as faster disease progression and poorer prognosis may have resulted in earlier HIV-related mortality in ε4-carriers. The mortality profile of the community from which this cohort was recruited was subject to high levels of HIV-associated death, and increased all cause death in age groups from 15 to 64 years in the years preceding data collection for this study (5). HIV progression is accompanied by domain-specific cognitive deficits characterised as HAND, where ε4 is associated with reduced performance in these domains (74,75,82,83,105). Some studies have failed to replicate these findings and have reported either no association between ε4 and HIV, or lack of association between ε4 and HAND (91,98,108,109). HIV status had no significant effect on cognitive function in any of our models. This result may be attributed to widespread use of antiretroviral therapy in this community, and also an indication that people living with HIV may have better access to healthcare and thus are better able to manage both age-related and HIV-associated morbidity (5,15,92). It appears that this sample population may have a high prevalence of HAND within the HIV positive group, but further evidence for trends in the development of accurately diagnosed HAND and seropositivity in the context of *APOE* have not been explored (15,92).

Although we observed possible associations of *APOE* genetic variants with executive function and visuospatial cognition, these associations did not survive correction for multiple testing. *APOE* ε2 homozygosity was linked to increased executive function scores and ε3/ε4 heterozygosity to lower visuospatial performance. Other studies comparing performance on cognition tests between ε3 homozygotes and ε4-carriers observed impaired executive function, episodic memory, global cognition, and processing speed in ε4-carriers which was exacerbated in ε4 homozygotes (45,47,110). Replication of these findings has proven challenging (1,44,45). We observed an increase of ~0.5 standard deviations in the domain of executive function for ε2/ε2 individuals and ~ 0.3 standard deviations reduction in visuospatial ability of ε3/ε4 individuals, larger effects than those of a large meta-analysis by Small et al. (2004). They observed reduced performance of <0.1 standard deviations for global cognition, episodic memory, and executive function in ε4-carriers, with significantly poorer cognitive performance associated with ε4 homozygosity (78). A later meta-analysis of 77 studies totaling 40 942 individuals observed similar results when assessing the effects of *APOE* on normal cognitive function (79). The resulting confidence intervals for global cognition, episodic memory, and executive function for ε4-carriers vs non- ε4-carriers in our study vs Wisdom *et al*. (2011) overlap, indicating that there is no significant difference between our study population and that of the European population meta-analysis. The wideness of the confidence intervals for our study (total cognition [-0.61-0.22], executive function [-0.22-0.14] and episodic memory [-0.13-0.07]) are an artifact of small sample size, but the effects of ε4 in the European study (global cognition [-0.08 to -0.04], executive function [-0.13 to -0.05] and episodic memory [-0.06 to-0.01]) although captured in the results of the HAALSI models indicates our study was insufficiently powered to detect these effects. Contrary to the findings of Wisdom *et al*. (2011), we did not observe significantly higher memory scores in ε2 homozygotes compared to ε4 homozygotes.

We report on baseline genetic and domain-specific cognitive function data for a large African study cohort. However, sample size was a limitation for observing small genetic effects for associations between *APOE* genotypes and cognitive function. Meta-analyses have estimated the effect sizes of *APOE* on different cognitive domains to vary between approximately 0.07 and 0.002 (78,79). But to detect the effect size of 0.01 estimated by Wisdom et al. (2011) for visuospatial ability, we would need to double our sample size. Our models explained modest variation of cognitive function mostly due to highly significant associations observed with

age (cognition score decreasing with increasing age) and level of education (cognition score increasing as higher levels of education are attained), which are known to influence normal cognitive function. It has been suggested that cognition changes throughout the lifespan and that ε4-carriers perform better at cognition tests when they are younger due to antagonistic pleiotropy and the compensation of other pathways which improve cognition and cognitive reserve at younger ages, but become detrimental over the age of 60 years (45,46,74,75). We were unable to confirm this in our study, but longitudinal data from the second wave of HAALSI data collection will enable the assessment of the effects of *APOE* on age-related cognitive decline.

There is little data on cognitive function and domain-specific genetic associations in African populations. Research in European and Asian populations have shown that *APOE* genetic variation is associated with cognition and differences in susceptibility to several diseases. Despite previous associations between ε4 and cognition, we were unable to replicate these findings. Instead, we observed a trend towards higher executive function in ε2 homozygotes and lower visuospatial cognition in ε3/ε4 heterozygotes. These associations were strongly moderated by education and age, and there was minimal indication for the independent influence of *APOE* on latent cognitive function.

Cognition screening tools which are culturally adapted and unbiased in terms of literacy and numeracy are essential for the accurate interpretation of cognition in populations with low levels of education. The inconclusive results of this study may further catalyse African exploratory studies of the complex genetics behind cognition, and the role of known environmental confounders that are becoming increasingly important in the prediction, prevention, and treatment of dementia in low- to middle-income countries.

2.6 Data availability statement

The HAALSI baseline data are publicly available at the Harvard Center for Population and Development Studies (HCPDS) programme website [www.haalsi.org]. Data are also accessible through the MRC/Wits Agincourt Research Unit's data repository [https://data.agincourt.co.za/index.php/catalog/18], the Inter- university Consortium for Political and Social Research (ICPSR) at the University of Michigan [www.icpsr.umich.edu],

and the INDEPTH Data Repository [https://www.indepth-ishare.org/ index.php/catalog/113]. The linked genetic and phenotypic data is available through the Harvard Center for Population and Development Studies (HCPDS) programme website [https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QWSQXR]. Requests for these data through the website are made to the principal investigators of the HAALSI study.

2.7 Author Contributions

The study was developed by CS together with MR, AN, ST and LB. MF transformed and analysed the raw cognition data and provided information on cognitive phenotypes. CS performed the analyses and drafted the manuscript, and all other authors critically revised and approved the manuscript.

2.8 Funding

The HAALSI study was funded by the National Institute on Aging (P01 AG041710) and is nested within the Agincourt Health and Demographic Surveillance System site, supported by the University of the Witwatersrand and Medical Research Council, South Africa, and the Wellcome Trust, United Kingdom (grants no. 058893/Z/99/A; 069683/Z/02/Z; 085477/Z/08/Z; and 085477/B/08/Z). It has been carried out through a collaboration between the Harvard Center for Population and Development Studies from Harvard T.H. Chan School of Public Health, MRC/Wits Rural Public Health and Health Transitions Research Unit from School of Public Health at the University of the Witwatersrand in South Africa, and the INDEPTH Network in Accra, Ghana. The content is solely the responsibility of the authors and does not necessarily represent the official views of National Institute on Aging or the Wellcome Trust.

The AWI-Gen Study was funded by the National Human Genome Research Institute (NHGRI), Office of the Director (OD), Eunice Kennedy Shriver National Institute Of Child Health & Human Development (NICHD), the National Institute of Environmental Health Sciences (NIEHS), the Office of AIDS research (OAR) and the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), of the National Institutes of Health (NIH)

under award number U54HG006938 and its supplements, as part of the H3Africa Consortium. The study design, conclusions, and opinions expressed in this paper do not necessarily represent the official views of the National Institutes of Health.

CS was supported by the NIH grant U54HG006938. CS was funded by the National Research Foundation (NRF) Thuthuka funding instrument grant no. TTK160602167377, and the NIH Fogarty International Centre SEED funding (D43TW008330) under the umbrella of the Wits Non-Communicable Disease Research Leadership Program. The views expressed herein are those of the authors and do not necessarily reflect those of the NIH, the NRF, or the NIH Fogarty International Centre.

2.9 Acknowledgments

Busisiwe Mthembu performed the *APOE* genotyping under the supervision of CS. Andrew May and Innocent Maposa provided guidance and advice on the statistical methods and using R.

Chapter 3

3 Genome-wide association study of population standardised cognitive performance phenotypes in a rural South African community

3.1 Abstract

Cognitive function is an indicator for global physical and mental health, and cognitive impairment has been associated with poorer life outcomes and earlier mortality. An adapted cognition battery along with the Oxford Cognition Screen-Plus was used to capture cognitive performance as five traits (total cognition score, verbal episodic memory, executive function, language, and visuospatial ability) for 2 246 adults in a rural-dwelling population of South Africans. Analysis of ~14 million markers imputed from the H3Africa genotyping array resulted in the detection of three genome-wide significant associations. For episodic memory a novel common variant rs73485231 (p=8.99e-09, β=0.229) and a rare African-specific variant (rs140372794, p=8.59e-08, β=0.324) were observed. Another genome-wide significant signal (rs139259554, p=4.84e-08, β=0.695) was identified for language score.

This pioneer African genome-wide association study identified novel genetic associations with general cognition and domain-specific cognitive pathways. Larger African studies with longitudinal data are essential to better understand the mechanisms behind cognitive ageing, and cognitive decline associated with neurocognitive disorders in Africa.

3.2 Introduction

Normal cognitive function is an essential determinant for health and quality of life indicators. Deficits have been associated with poorer socio-economic status (SES), physical, neurodegenerative and neuropsychiatric disease, and earlier mortality making it increasingly important to decipher relevant genetic markers and biological pathways for intervention and treatment (1,21,64,111–113). Genome-wide association studies (GWAS) for cognitive function have been challenging despite twin studies suggesting heritability scores up to ~80% for different cognitive phenotypes (1,21,28,64,114–116). The question of heritability is further complicated by evidence that heritability appears to vary across lifespan and has

different trajectories throughout the life course, with relative stability observed from middle to old age (1,35,117,118).

Despite the complex and polygenic nature of neurocognitive phenotypes, meta-analyses with large sample sizes (> 50,000) were able to detect many single nucleotide polymorphisms (SNPs) associations and have successfully replicated association findings with genome-wide significance ($p<5x10^{-8}$) (21,24,64,66,112,119). In order to perform these meta-analyses, general cognitive ability (or Spearman's *g*) was derived from diverse positively but not perfectly correlated cognitive performance tests (capturing ~40% of phenotypic variance), or proxy phenotypes such as educational attainment (21,22,24,64,66,112,119). Studies have used different metrics, measures, and tests to describe traits such as intelligence (fluid or crystallised), cognitive function, non-pathological cognitive ageing, and domain-specific cognitive outcomes hence the adoption *g* to account for testing heterogeneity (1,21,22,25,99,114).

Functional studies have shown that each of the cognitive domains have an impact on gene expression in different regions of the brain, making latent cognitive ability an amalgamation of activity within the brain acting through different biological pathways (1,6,7,23,120–124). A further limitation of these studies is that they suffer from sample heterogeneity in terms of age of participants, SES, and participant's access to education (21,24,25,118). As noted, cognitive trajectory changes throughout lifespan require participants to be within similar age ranges to accurately capture cognitive ability in comparative studies (1,35,117,118). Education is also a major moderating factor for assessing cognitive ability, and cognitive performance tests usually rely on literacy and numeracy (1,22,25,29,35,114,125). In some settings, SES is a major determinant influencing access to education so cognitive batteries may be measuring educational exposure rather than innate cognitive function (1,22,25,29,35,114,125).

There is little research on the genetics of cognitive function in African populations, or in those of African ancestry (1,12,115,126,127). The lack of diverse ethnic representation in studies to-date limit the discovery of associated and potentially causal variants. The differences in linkage disequilibrium (LD) with smaller LD blocks in Africans, compared to Europeans, could enhance discovery of causal variants in Africans (31). The Health and

Aging in Africa: A Longitudinal Study of an INDEPTH Community in South Africa (HAALSI) collected baseline cognition data for over 5 000 older adults in Bushbuckridge, rural Mpumalanga, South Africa (SA) (13). A sub-set of 2 246 participants from this study were also recruited as part of the Africa Wits-INDEPTH Partnership for Genomic Studies (AWI-Gen) for whom genotype data were available from the Illumina Human Heredity and Health in Africa (H3Africa) array (16,17). The combined dataset with phenotype and genotype data was used to explore genetic associations with latent cognitive ability based on multiple quantitative traits (total cognition score, episodic memory, executive function, language, and visuospatial ability) for ~2 000 individuals in five independent GWAS. This is the first large study to explore genome-wide contributions to cognition in an African setting.

3.3 Methods

3.3.1 Participants

Participants were enrolled for both the AWI-Gen and HAALSI studies. Ethical approval was granted through the University of the Witwatersrand, Johannesburg, Human Research Ethics Committee under the following certificate numbers: AWI-Gen M121029 and M170880; HAALSI M141159; and the current study M170916. Socio-demographic data, infection history, and cognitive performance data were collected from 5 059 consented participants (male and female) aged 40 years and older recruited from Bushbuckridge, Mpumalanga (November 2014 to November 2015) and a sub-set of 2 246 of these participants had genotype data. All participants provided written informed consent. Descriptive statistics was performed using R (R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna. Austria. https://www.R-project.org/).

3.3.2 Questionnaire-based cognitive assessment

The US HRS cognition screening tool was culturally adapted and translated into the local vernacular Shangaan (also referred to as Xitsonga). This tool consisted of questions representing the domains of memory and orientation, and was scored from 0-24 (2,13,29). This tool is described in detail in publications by Kobayashi *et al*. (2019) and Farrell *et al*. (2020).

3.3.3 Tablet-based cognitive assessment

The OCS-Plus developed by Humphreys *et al*. (2017) is an electronic cognitive assessment administered using a tablet and was validated for use in this cohort (25). It consists of nine domain-specific cognitive tests which assess language, episodic memory, executive function, attention, and pattern recognition (25). A factor score was derived for each cognitive domain (episodic memory, executive function, language, and visuospatial ability) (29). This method is based on work by Seidlecki, Honig, and Stern (2008), and provides population-standardised domain z-scores for each participant (29).

3.3.4 Genotyping and imputation

Genotyping of the full AWI-Gen dataset (10 900 participants) was performed using the H3Africa array by Illumina (San Diego, CA, USA). This custom chip of ~2.3 million SNPs was developed to be enriched for common African variants (http://chipinfo.h3abionet.org) (33). Data from AWI-Gen were processed through the H3A GWAS pipeline (https://github.com/h3abionet/h3agwas) where individuals with SNP missingness greater than 0.05 were removed from the dataset. SNPs were removed if they had genotype missingness above 0.05, minor allele frequency (MAF) below 0.01 and were not in Hardy-Weinberg equilibrium (HWE) $p < 1x10^{-6}$. SNPs were matched to Genome Reference Consortium Human Genome build 37 (GRCh37) (https://www.ncbi.nlm.nih.gov/grc) (128) and ambiguous SNPs were removed. The 1.71 million SNP dataset was then imputed using the African Genome Resources reference panel at the Sanger Imputation Server (https://imputation.sanger.ac.uk/) (129). EAGLE2 (130) was selected for pre-phasing and positional Burrows-Wheeler transformation (PBWT) (131) algorithm for imputation. Poorly imputed SNPs scoring less than 0.6, with MAF below 0.01 and HWE p-value cut-off $<10^{-6}$ were excluded, and the final dataset included ~14 million SNPs. The HAALSI/AWI-Gen samples were extracted from this dataset.

Principal component analysis (PCA) using EIGENSTRAT (132) was performed to assess population stratification within the samples as well as to find the genetic affinities of our cohort to other African ancestry populations from the 1000 Genome Project (KGP) dataset (133). Population outliers were removed based on visual inspection of the PCA plots using Genesis version 0.2.6 (134).

3.3.5 Genome-wide association analysis

GWAS was performed for each of the five cognitive phenotypes. Total cognition was used as a continuous trait with scores ranging from 0 to 24. Cognitive domain scores for 1 903 individuals from the OCS-Plus were rank normalized using R (https://www.R-project.org/) as the standardized z-scores were not normally distributed. Association was performed on the full imputed dataset using Genome-wide Efficient Mixed-Model Association (GEMMA) (70), and adjusting for age as a continuous covariate, and sex, highest level of education attained (primary, secondary, tertiary), and HIV status as categorical covariates. LD scores from the KGP African reference panel and a reference panel specific to AWI-Gen's SA data were used to adjust for LD structure. Analyses were run on an automated H3Africa workflow for GWAS (http://github.com/h3abionet/h3agwas/) (69).

3.3.6 Visualisation and post-GWAS analysis

Association output files from GEMMA were analysed using Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) (https://fuma.ctglab.nl/) for partitioning of signals based on linkage disequilibrium (LD), visualization and functional annotation (135). Manhattan plots and QQ-plots for SNP-based association were generated using FUMA. Genomic inflation factors were calculated using a local R script. Locus zoom plots (136) were created for selected association signals based on the summary statistics from GEMMA and SA-specific LD panel. Box-plots were constructed from Kruskal-Wallis tests for comparison of cognitive function between individuals by genotype at each relevant SNP. The GWAS Catalog (137) (http://ebi.ac.uk/gwas/) [accessed 14/10/2021], and Phenoscanner v2 (138,139) (http://www.phenoscanner.medschl.cam.ac.uk/) [accessed 14/10/2021] were used to infer previous associations of the lead SNPs. Previous associations in 100 kb genomic regions on either side of each lead SNP were also examined. Ensembl (140) and literature mining were used to assess the functional interpretation.

3.4 Results

3.4.1 Population structure and affinities

Principal component analysis (PCA) showed little population heterogeneity (***Figure 3.1a***) and 19 outlying individuals were removed, resulting in a final sample size of 2 227. The PCA with other African Ancestry populations from the KGP Project (133) datasets showed a distinct clustering from East, West and Central-West African populations, and African Americans (***Figure 3.1b***).

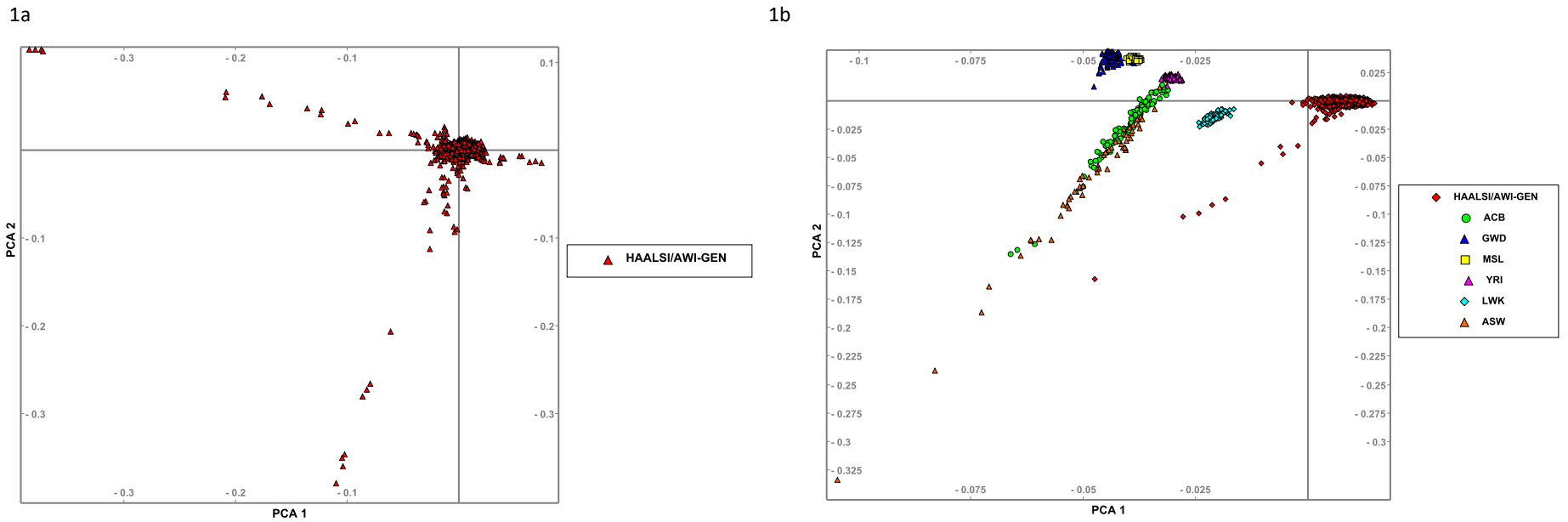


Figure 3.1: Population structure and affinities of the HAALSI/AWI-Gen participants

Principal component analysis (PCA) of individuals from the HAALSI/AWI-Gen sample showing PC1 and 2. 1a shows the absence of any major population structure. 1b shows a PCA comparison of our study participants with African population datasets (African Caribbeans in Barbados (ACB) and Americans of African Ancestry in SW USA (ASW)), East (Luhya in Webuye, Kenya (LWK)), and West Africans (Yoruba in Ibadan, Nigeria (YRI), Gambian in Western Divisions in the Gambia (GWD), and Mende in Sierra Leone (MSL)) from the 1000 Genomes Project.

3.4.2 Genome-wide association study results

We performed GWAS for five cognition-related traits (***Table 3.1***). This sample has more women (~58%) than men. There was little access to education where ~77% of the sample population had not progressed beyond primary school. After rank normalization, cognitive domain data were available for 1 903 genotyped participants. The ranges displayed in ***Table 3.1*** are for the population standardized z-scores prior to transformation and show a particularly wide range of performance for visuospatial cognition (-10.627 to 4.771). Total cognition score data was available for all 2 227 genotyped participants. The imputed dataset included 13 972 012 SNPs.

Table 3.1: Descriptive statistics showing cognitive trait and covariate distribution

Measure	***n***	***Range***	***Mean or percentage***	***SD***
Age (years)	2246	40 to 80	57.57	10.84
Sex	2246			
Female			58.37%	
Male			41.63%	
Level of education	2246			
No formal education			39.40%	
Primary			37.49%	
High			19.19%	
Tertiary			3.92%	
HIV status	2246			
Negative			73.02%	
Positive			26.98%	
Cognitive measures				
Total cognition score	2246	0 to 24	11.72	4.38
Cognitive domains*	1903			
Executive function		-4.687 to 4.328		
Episodic memory		-3.824 to 1.456		
Language		-2.533 to 1.279		
Visuospatial cognition		-10.627 to 4.771		

Sample number (n) differs per covariate/measure due to data missingness. Z-scores for cognitive domain data are represented only as ranges. Means and standard deviation (SD) are shown for continuous variables and percentages for categorical ones.

3.4.2.1 Cognitive domain traits

The GWAS for verbal episodic memory, identified a genome-wide significant signal for rs73485231 (p=8.99e-09, β=0.229, SE=0.040) on chromosome 13 (***Figure 3.2a*** shows the Manhattan plot, ***3.2b*** the QQ plot with **λ**=0.9927, and ***2c*** the Locus zoom plot). The mean episodic memory score was significantly lower in G homozygotes (p=4.4e-04) (***Supplementary Figure 3s1a***). This intergenic SNP between *GNG5P5* and *HTR2A* had a notably higher minor allele frequency (MAF=0.132) in Africans (AFR), compared to Europeans (EUR), Americans (AMR) and Asians (EAS and SAS) KGP groups (***Table 3.2***). Although, no previous associations with cognition had been reported, the GWAS Catalog reported this SNP to be associated with adolescent idiopathic scoliosis (***Figure 3.2c***) ((137), GCST006287 (141), accessed 14/10/2021)).

A novel signal for verbal episodic memory, rs140372794 (p= 8.59e-08, β=0.324, SE=0.061), was observed on chromosome 8 (***Supplementary Table 3.1***). This SNP falls within 100kb of *LINC02055*, a long intergenic non-protein coding RNA gene harbouring several SNPs previously associated with AD disease pathology ((137), GCST010341 (142), accessed 14/10/2021)), mathematical ability ((137), GCST006573 (143), accessed 14/10/2021)), and general cognitive function ((137), GCST90011293 (144), GCST008595 (122), GCST006269 (64), accessed 14/10/2021)).

Table 3.2: Minor allele frequency distribution of genome-wide significant signals in this sample compared to 1000 Genomes Super Populations

Trait	***Variant ID***	***Minor allele***	***MAF sample***	***MAF AFR***	***MAF AMR***	***MAF EAS***	***MAF EUR***	***MAF SAS***
Episodic memory	rs73485231	A	0.141	0.132	0.020	0.024	0.015	0.069
Episodic memory	rs140372794	A	0.053	0.041	0.007	0	0	0
Language	rs139259554	G	0.011	0.009	0	0	0	0
Language	rs34595826	A	0.240	0.231	0.140	0.134	0.158	0.094
Total cognition score	rs138832740	C	0.025	0.039	0	0	0	0

1000 Genomes Project Super Populations groups AFR (African), AMR (Admixed American), EAS (East Asian), EUR (European), and SAS (South Asian).

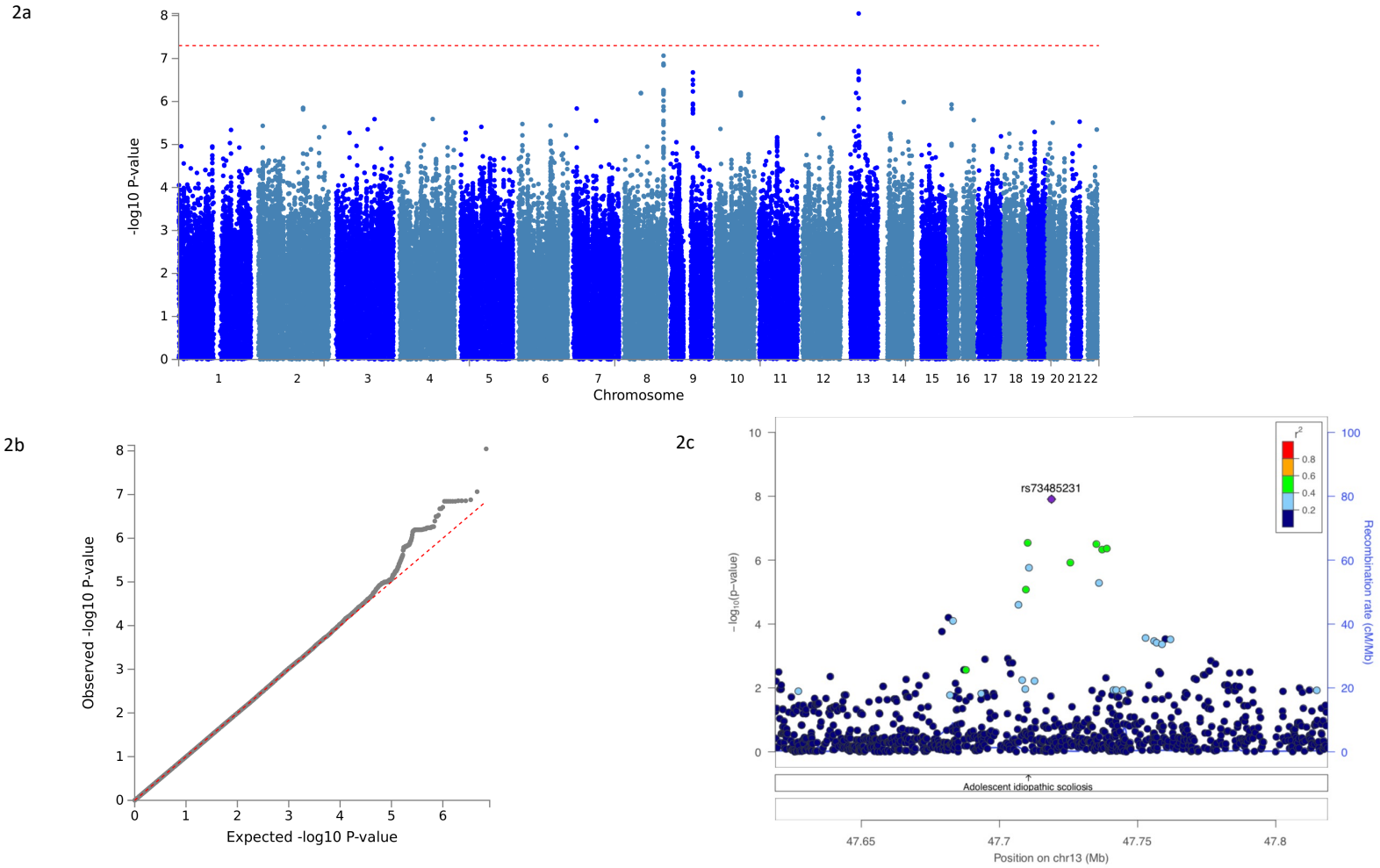


Figure 3.2: Genome-wide and suggestive associations for verbal episodic memory

2a. Manhattan plot. Genome wide significance cut-off 5x10^{-8} shown by dashed red line. 2b. QQ-plot for association summary statistics 2c. Locus zoom plot for rs73485231. LD based on a South African LD panel.

The GWAS for language detected a genome-wide significant association on chromosome 6 (rs139259554, p=4.84e-08, β=0.695, SE=0.127) (***Figure 3.3a*** and ***3.3b***). The MAF (C allele) for rs139259554 was 0.011 in our cohort and was not reported in population groups other than the African super group in the KGP (***Table 3.2***).

A suggestive signal on the chromosome 7 (rs34595826; p=1.21e-07) was detected for language score. The SNP mapped to an intronic region within the *LHFPL3* gene (***Figure 3.3c***; ***Supplementary Table 3.1***). The MAF of this high-effect variant (β=-0.166, SE=0.031) was 0.240 and it was supported by many suggestive and near suggestive signals (***Figure 3.3c***). Accordingly, there was a significant difference between language scores among the genotype classes corresponding to rs34595826 (***Supplementary Figure 3s1b***), with the homozygous A individuals showing significantly lower language scores (p=8.4e-06). The minor allele A was observed at a higher frequency in Africans compared to other populations (***Table 3.2***). Phenoscanner showed association of rs34595826 with schizophrenia (p=1.33e-04, β=0.053) and "risk taking" (p=4.90e-03, β=0.004) (138,139,145,146). Regional lookup (7:104367152-104567152) revealed previous associations with educational attainment ((137), GCST006442 (143), accessed 14/10/2021)), schizophrenia ((137), GCST003048 (147), accessed 14/10/2021)), and Tau cytotoxicity in Alzheimer's disease ((137), GCST010340 (142), accessed 14/10/2021)). At the gene level *LHFPL3* has been previously associated with several neurocognitive, neurodegenerative and neuropsychiatric phenotypes, and proxy phenotypes related to educational attainment ((137), GCST90011293 (144), GCST002367 (148), GCST002045 (149), GCST003048 (147), GCST006573 and GCST006574 and GCST006442 (143), GCST010340 (142), GCST006269 (64), accessed 14/10/2021)).

In addition to the genome-wide significant signals observed for verbal episodic memory and language, several suggestive signals (shown in ***Supplementary Table 3.1***) were observed across the domain phenotypes and include suggestive signals for executive function and visuospatial ability.

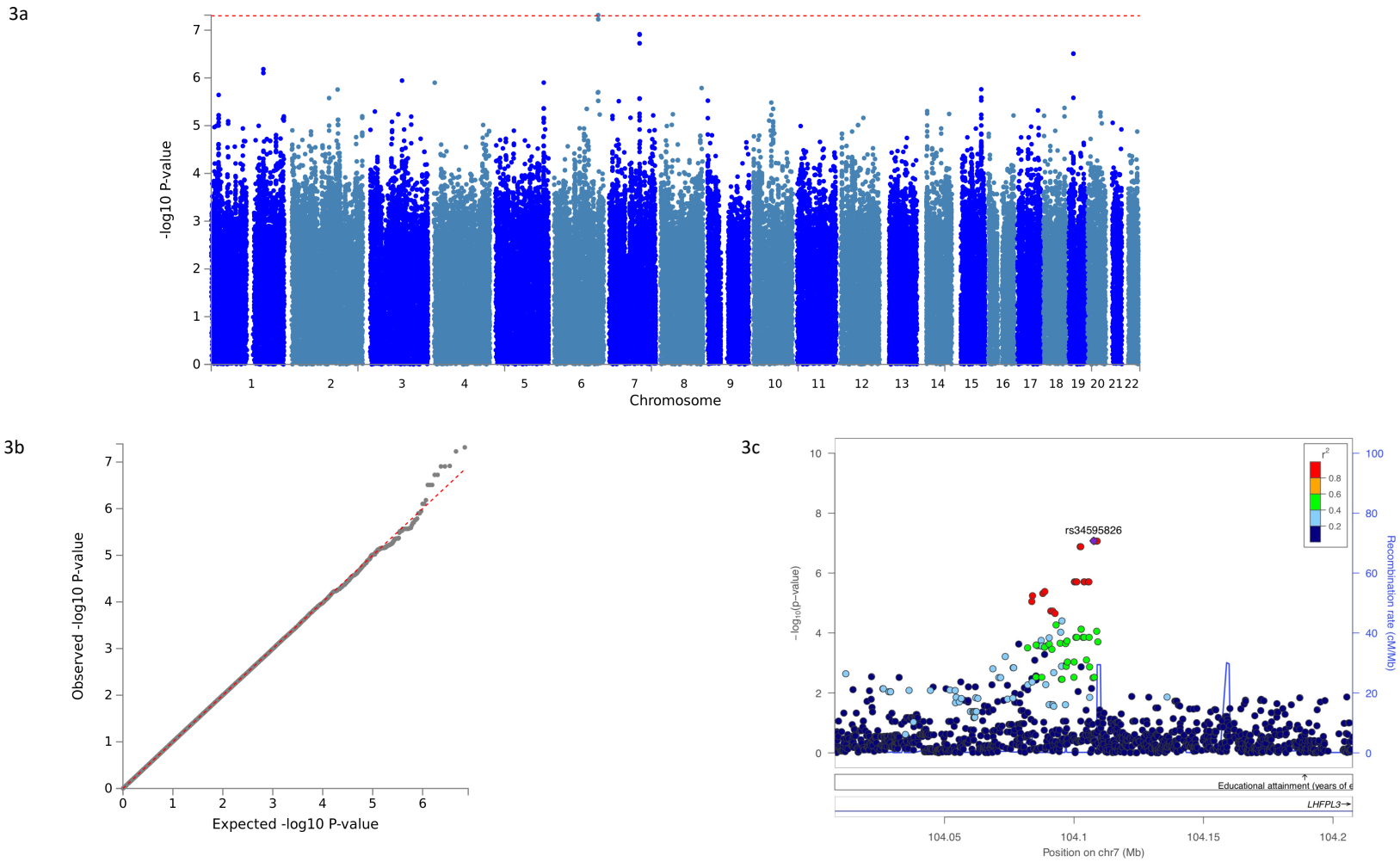


Figure 3.3: Genome-wide and suggestive associations for language

3a. Manhattan plot. Genome wide significance cut-off 5x10^{-8} by dashed red line. 3b. QQ-plot for association summary statistics. 3c. Locus zoom plot for rs34595826. LD values are based on a South African LD panel.

3.4.2.2 Total cognition score

Although no SNPs reached genome-wide significance for association with total cognition score (***Supplementary Table 3.1***), the lead SNP rs138832740 (p=1.10e-07, β=-2.004, SE=0.376) was supported by several suggestive signals in the same region. The SNP was detected to be African ancestry-specific in the KGP dataset (***Table 3.2***). No previous associations had been reported for rs138832740 likely due to its low frequency and apparent continental specificity. Several variants in the surrounding region on chromosome 1 had been previously associated with cognitive traits (intelligence, general cognitive ability, cognitive function measurement) and correlated phenotypes (educational attainment and schizophrenia). These include associations to numerous psychiatric ((137), GCST90016595 (150), GCST010649 (151), GCST007543 (152), GCST009371 (153), GCST007556 (154), accessed 14/10/2021), behavioral ((137), GCST007576 (155), GCST007983 (156), GCST004497 (157), GCST004497 (158), accessed 14/10/2021), and anthropometric ((137), GCST000296 (159), GCST009871 (160), accessed 14/10/2021) traits that map to the nearby *RN7SL831P* gene.

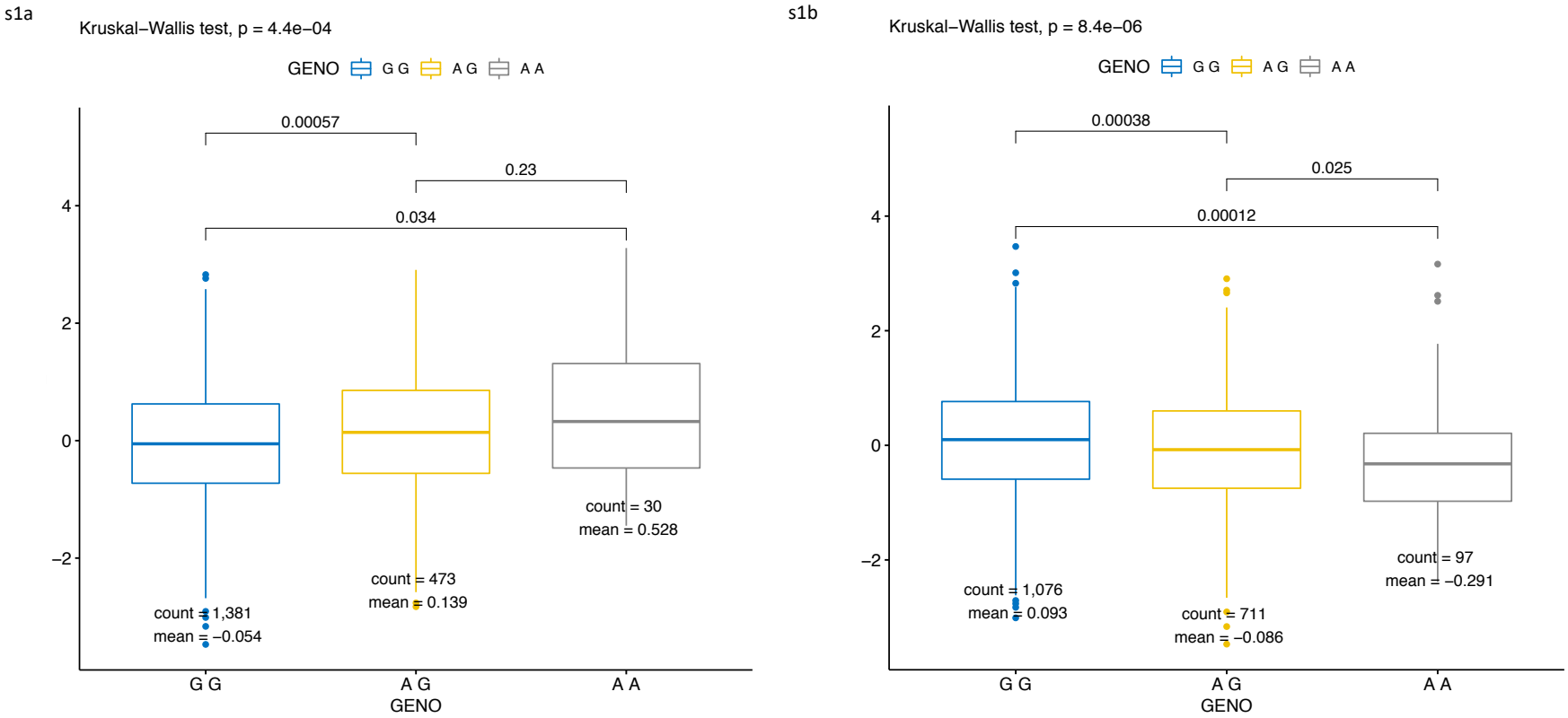
3.4.3 GWAS replication

We tested whether any of the known genome-wide significant signals for language and episodic memory from other populations ((137), downloaded 14/10/2021) showed replication in our cohort (***Supplementary Table 3.2***). Of the three genome-wide significant associations for language (language performance in older adults (161), language impairment (162), receptive language perception (163)) previously reported in European studies, one (rs3801203) was not present in our dataset and the other two failed to reach our replication P-value threshold of $5x10^{-3}$. Of the eighteen SNPs reported with genome-wide significant associations for memory (whether verbal, logical, or related specifically to decline in memory performance), six were not in our dataset and the rest failed to reach even a relaxed P-value threshold of $5x10^{-3}$. Previous associations with the traits executive function and visuospatial ability were not investigated as our study did not reveal any genome-wide significant SNPs for those phenotypes.

The SNP, rs138832740, significantly associated with total cognition score in our study could not be replicated as it is monomorphic in European populations, in whom most cognitive function GWAS have been performed.

Supplementary Table 3.1: Summary of independent association signals (3e-06) detected in GWASs for the five cognitive phenotypes

Study phenotype	chromosome	position (GRCh38)	rsID	effect allele	effect allele frequency	p	beta	se	annotation
Episodic memory	13	47144696	rs73485231	A	0.141	8.99e-09	0.229	0.040	intergenic
Episodic memory	8	136009868	rs140372794	A	0.053	8.59e-08	0.324	0.061	intergenic
Language	6	150737600	rs139259554	G	0.011	4.84e-08	0.695	0.127	intronic
Language	7	104467152	rs34595826	A	0.240	1.21e-07	-0.166	0.031	intronic
Language	19	16348801	rs8101774	G	0.626	3.10e-07	0.141	0.027	intergenic
Total cognition score	1	96531287	rs138832740	C	0.025	1.10e-07	-2.004	0.376	intergenic
Total cognition score	17	79246119	rs9893796	C	0.055	1.39e-07	-1.325	0.259	intronic
Total cognition score	9	79144373	rs12336525	A	0.061	7.87e-07	-1.205	0.243	intronic
Total cognition score	14	20993593	rs61073164	G	0.047	9.05e-07	-2.004	0.275	3'UTR
Total cognition score	8	131523064	rs73343547	G	0.059	1.06e-06	-1.226	0.251	intergenic
Total cognition score	20	14967004	rs6043030	A	0.076	1.15e-06	1.102	0.226	intronic
Executive function	2	127118084	rs10929007	G	0.766	2.78e-07	-0.164	0.032	intergenic
Executive function	1	170277624	rs2990655	G	0.845	5.49e-07	-0.191	0.038	intronic
Executive function	22	25233246	rs79191028	T	0.040	1.12e-06	-0.336	0.007	intergenic
Executive function	20	13692248	rs146628683	T	0.027	1.20e-06	0.413	0.047	intergenic
Executive function	12	72148685	rs10507203	A	0.169	1.40e-06	-0.176	0.036	intergenic
Executive function	21	31952049	rs76885705	G	0.018	1.47e-06	0.499	0.103	intronic
Visuospatial ability	4	24632114	rs112938427	G	0.016	1.70e-06	0.476	0.099	regulatory region
Visuospatial ability	8	142085116	rs567232080	C	0.020	1.93e-06	-0.424	0.089	intergenic
Visuospatial ability	11	117975655	rs147982326	C	0.033	2.23e-06	-0.327	0.069	intergenic
Visuospatial ability	1	54839010	rs1035815	G	0.532	2.26e-06	-0.115	0.024	intronic



Supplementary Figure 3.1: Box-plots showing effects of variants on cognitive function

s1a. *rs73485231 showing difference in the median z-scores for episodic memory between observed genotypes.* ***s1b.*** *rs34595826 the medians z-scores for language between observed genotypes. The means and number of individuals observed for each genotype are included in the diagrams.*

Supplementary Table 3.2: Replication of previous associations for language and verbal episodic memory

	GWAS catalog						*Study sample*		
SNP	**effect allele**	**P-value**	**reported effect allele frequency**	**beta**	**Reported trait**	**Location (GRCh38)**	**Phenotype**	**effect allele frequency**	**p**	**beta**
rs7594645	G	4.0e-09	0.072	0.072 unit decrease	Episodic memory	2:206781950	Verbal episodic memory	0.164	7.27e-01	-1.29e-02
rs429358	C	2.0e-18	0.25	0.42/0.36 unit decrease	Logical memory (delayed/immediate recall)	19:44908684	Verbal episodic memory	0.242	8.78e-01	4.90e-03
rs4420638	G	1.0e-16	0.18	NA	Verbal declarative memory	19:44919689	Verbal episodic memory	0.278	5.92e-01	-1.67e-02
rs6857	T	4.0e-13	0.16	NA	Verbal declarative memory	19:44888997	Verbal episodic memory	0.040	5.97e-01	-3.66e-02
rs769449	A	3.0e-12	Not reported	NA	Verbal memory performance (residualized delayed recall level)	19:44906745	Verbal episodic memory	NA	NA	NA
rs157582	A	8.0e-10	Not reported	NA	Verbal memory performance (immediate recall change)	19:44892962	Verbal episodic memory	0.592	1.74e-01	3.87e-02
rs73705514	C	3.0e-09	0.02	0.84 unit decrease	Logical memory (delayed recall) in mild cognitive impairment	7:88777238	Verbal episodic memory	0.423	1.89e-01	3.72e-02
rs3801203	A	3.0e-09	0.19	0.261 unit decrease	Language performance in older adults (adjusted for episodic memory)	7:42148801	Verbal episodic memory	NA	NA	NA
rs12268753	C	2.0e-08	0.21	0.3 unit increase	Logical memory (delayed recall) in normal cognition	10:52058389	Verbal episodic memory	NA	NA	NA
rs118130881	G	2.0e-08	0.04	0.61 unit decrease	Logical memory (immediate recall) in mild cognitive impairment	8:31371254	Verbal episodic memory	NA	NA	NA
rs111882035	G	3.0e-08	0.02	0.93 unit decrease	Logical memory (delayed recall) in mild cognitive impairment	4:85495401	Verbal episodic memory	NA	NA	NA

	GWAS catalog						Study sample		
SNP	**effect allele**	**P-value**	**reported effect allele frequency**	**beta**	**Reported trait**	**Location (GRCh38)**	**Phenotype**	**effect allele frequency**	**p**	**beta**
rs11074779	T	3.0e-08	0.19	0.4033 unit increase	Verbal declarative memory	16:26440122	Verbal episodic memory	0.060	5.15e-01	3.88e-02
rs6813517	T	3.0e-08	0.21	0.3686 unit increase	Verbal declarative memory	4:167601600	Verbal episodic memory	0.129	2.07e-01	-5.25e-02
rs13360092	A	3.0e-08	0.07	0.7655 unit decrease	Verbal declarative memory	5:51233341	Verbal episodic memory	0.049	8.72e-01	1.02e-02
rs74834332	A	4.0e-08	0.02	0.79 unit increase	Logical memory (delayed recall) in normal cognition	1:57273491	Verbal episodic memory	0.011	1.06e-01	2.21e-01
rs13374761	A	4.0e-08	Not reported	0.01 unit increase	Memory decline	1:160643161	Verbal episodic memory	0.292	7.95e-01	7.86e-03
rs73076581	T	4.0e-08	Not reported	0.7224384 unit decrease	Face memory	3:44824490	Verbal episodic memory	NA	NA	NA
rs67017972	A	4.0e-08	Not reported	0.6 unit decrease	Verbal memory	5:53775458	Verbal episodic memory	0.351	7.79e-03	7.80e-02
rs3801203	A	3.0e-09	0.19	0.261 unit decrease	Language performance in older adults	7:42148801	Language	NA	NA	NA
rs4280164	G	4.0e-08	0.767	3.872 unit decrease	Language impairment	14:24302079	Language	0.111	6.29e-01	1.98e-02
rs11787922	T	1.0e-08	Not reported	23.05 unit decrease	Receptive language perception	9:80173921	Language	0.170	1.90e-01	-4.54e-02

3.5 Discussion

Few genetic association studies for cognitive traits have been performed in continental Africans, and meta-analyses suffer from the limitations of grouping different cognitive phenotypes together of which data was collected using different screening tools (1,22,44). The latter are often ill-adapted to screening populations with lower literacy and numeracy levels, confounding comparative analyses (1,44). This pioneer African GWAS uncovered genome-wide significant and suggestive association signals from population standardised cognitive domain scores and total cognition. We used baseline cognitive function data from a well-characterised rural South African cohort (13), genetic data enriched for common African variants and imputed using an African-enriched reference panel, and the OCS-Plus (25) cognitive assessment tool specifically developed for low-income settings where access to formal education is limited, and language may present barriers. Although of modest size, compared to many recent meta-analyses of cognitive traits, our study detected two genome-wide significant associations of which one was novel, and the other was African-specific.

The first was an association for verbal episodic memory with a SNP (rs73485231) on chromosome 13 in a region previously associated with adolescent idiopathic scoliosis (141). This SNP is localised to an intergenic region between G protein subunit gamma 5 (*GNG5P5*) and 5-hydroxytryptamine receptor 2A (*HTR2A*), both of which have also been associated with substance use behaviors and intellectual disability (164–166). In our sample, the A allele of rs73485231 was associated with improved memory test performance where AA homozygotes exhibiting a higher mean memory score than GG homozygotes (p=0.035) was observed (***Supplementary Figure 3s1a***). The near genome-wide significant signal associated with verbal episodic memory, rs140372794 on chromosome 8 lies in close proximity to long intergenic non-protein coding RNA gene, *LINC02055*. Several SNPs in this 200kb region have been associated with cognitive phenotypes including mathematical ability and AD pathology (142–144,167).

The third was for language, with a genome-wide significant association for rs139259554, located within an intron of Pleckstrin homology and Rho guanine nucleotide exchange factor (RhoGEF) domain (*PLEKHG1*). Although, previous associations for this African-specific SNP had not been reported for language or any other cognitive performance phenotypes,

other variants within *PLEKHG1* have been associated with cerebral white matter intensities (an indication of susceptibility to vascular dementia) and systolic blood pressure in sickle cell populations (168–170).

Among the suggestive associations, rs34595826, a common intronic variant in the Lipoma high-mobility group protein isoform C (HMGIC) fusion partner-like (LHFPL) tetraspan subfamily member 3 gene (*LHFPL3*) was associated with language. Individuals homozygous for the minor allele (A) had significantly lower language performance scores than those homozygous for the major allele (G) (p=9.8e-05) (***Supplementary Figure 3s1b***). The effect allele occurs at higher frequencies in African populations and is an example of common African variation contributing to a complex trait. This SNP was previously associated with organ-limited amyloidosis and schizophrenia (145,146). The genomic region around the SNP (+/- 100kb) harbours SNPs previously associated with hyperphosphorylated Tau in Alzheimer's disease pathology and educational attainment which is often used as a proxy phenotype for cognitive ability (142,143). Other *LHFPL3* variants have been associated with general cognitive ability and social interaction at genome-wide significant levels (64,143,148,171).

The adaptation of the US HRS cognition battery proved to be adequately robust as an assessment of total cognition based on memory and orientation (2). Although this test originally included questions on numeracy, these were excluded as they were shown to be biased toward participants with higher levels of education (2). For total cognition score, using this adapted battery of assessments, we observed a suggestive signal for rs138832740 on chromosome 1. This variant has only been observed to be polymorphic in African populations (133) but SNPs within the 200kb region flanking this SNP have previously been associated with neurocognitive and neuropsychiatric traits, as well as anthropometric traits related to body mass (150,151,153,154,158–160). Genome-wide signals in this region at rs635754 and rs137858393 were observed in a sample size of over 1 million individuals for educational attainment, and another SNP rs1073090 was significantly associated with mathematical ability (n=430 439) (143). The closest gene according to Phenoscanner was RNA 7SL cytoplasmic 831 pseudogene (*RN7SL831P*) for which Wang *et al*. (2020) noted co-expression along with *PLXNC1* (which codes for a transmembrane receptor which regulates axon guidance and immune response) in amyloid-β pathology in AD (139,142).

The same promoter targets *RN7SL831P* and *PTBP2* which are primarily expressed in brain tissue and have been associated with reaction time in cognitive performance, schizophrenia, social interaction, and autism spectrum disorder (151–154,158).

Larger longitudinal cohorts need to be established in low- and middle-income countries where data on the prevalence and progression of cognitive impairment due to ageing, non-communicable diseases, infectious disease, or dementias is scarce. This will promote research into the risk factors and causes of cognitive decline in populations experiencing disproportionate adversity. In our study, we used a screening tool which captured intra-population domain-specific cognition, rather than exploring the genetic basis of educational attainment as a proxy for cognitive function, as many other studies have done. Education attainment is known to be a biased and inadequate metric in communities such as that targeted in our this research, where low levels of education observed likely correspond to extreme educational inequality in rural communities in SA during the apartheid era, when these individuals were young (2,14,25,29). Having a set of well-defined traits that are population standardised provides more accurate phenotype distributions for isolating variants associated with cognitive traits, as well as mitigating stigma attached to traits labelled inappropriately as “intelligence”.

The OCS-Plus was a valuable tool in our study community which is known to have long-standing poor access and quality of education further limited by low employment rates (14). This mature sample population aged 40-80 years was itself a strength as the literature states that the heritability of cognitive function appears to change across the lifespan (1,117). The environmental factors moderating cognitive function may be more, or less potent during the life course and trends between domains differ progressively with age, but stabilise at middle-age (1,117). We were limited by sample size, but our inability to replicate previous genome-wide significant signals may also by related to differences in the genetic architecture of these traits in Africans, as most previous studies have been performed in populations with European ancestry. Moreover, the phenotype data and analysis approaches also differed between our study and previous genetic association studies.

The AWI-Gen/HAALSI collaboration is a catalyst and foundation for genetic studies on neurocognitive traits in South and sub-Saharan Africa with evidence of novel associations and some replication of previous associations. The next steps would include genetic association studies on longitudinal follow-up data to explore the role of genetic variation in the rate of cognitive decline during the course of normal ageing. It would also be interesting to see whether the effects of signals associated with cognitive decline differ in the direction or size of effect in this population versus in other population groups.

Expanding the use of the OCS-Plus in future African studies may produce replication cohorts and contribute to the greater understanding of cognitive function from a genetic perspective especially at this stage of rapid health and social transition in South Africa.

3.6 Acknowledgements and funding

CCS and AC were supported by the NIH grant U54HG006938. Bioinformatics training and support was provided by members of H3ABioNet team and the Sydney Brenner Institute for Molecular Bioscience. MR is a South African Research Chair in Genomics and Bioinformatics of African populations hosted by the University of the Witwatersrand, funded by the Department of Science and Technology, and administered by the National Research Foundation. We are grateful to the patience of the study participants without whom this research would not have been possible.

AWI-Gen is funded by the National Human Genome Research Institute (NHGRI), Office of the Director (OD), Eunice Kennedy Shriver National Institute Of Child Health & Human Development (NICHD), the National Institute of Environmental Health Sciences (NIEHS), the Office of AIDS research (OAR) and the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), of the National Institutes of Health (NIH) under award number U54HG006938 and its supplements, as part of the H3Africa Consortium. The study design, conclusions, and opinions expressed in this paper do not necessarily represent the official views of the National Institutes of Health.

HAALSI was funded by the National Institute on Aging (P01 AG041710) nested within the Agincourt Health and Demographic Surveillance System supported by the University of the

Witwatersrand and Medical Research Council, South Africa, and the Wellcome Trust, UK (grants 058893/Z/99/A; 069683/Z/02/Z; 085477/Z/08/Z; 085477/B/08/Z). This collaboration lies between the Harvard Center for Population and Development Studies Harvard from T.H. Chan School of Public Health, MRC/Wits Rural Public Health and Health Transitions Research Unit, School of Public Health, University of the Witwatersrand in South Africa, and the INDEPTH Network in Accra, Ghana.

CCS was funded by the National Research Foundation (NRF) Thuthuka funding instrument grant number TTK160602167377, and the NIH Fogarty International Centre SEED funding (D43TW008330) under the umbrella of the Wits Non-Communicable Disease Research Leadership Program.

The views expressed herein are those of the authors and do not necessarily represent the official views of the funders.

3.7 Data availability statement

The HAALSI baseline data are publicly available at the Harvard Center for Population and Development Studies (HCPDS) programme website [www.haalsi.org]. Data are also accessible through the MRC/Wits-Agincourt Research Unit's data repository [https://data.agincourt.co.za/index.php/catalog/18], the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan [www.icpsr.umich.edu], and the INDEPTH Data Repository [http://www.indepth-ishare.org/index.php/cata log/113]. Genome-wide genomic data from the AWI-Gen study are in the European Genome-phenome Archive (EGA; https://ega-archive.org/) with accession number: EGAD00010001996. The phenotype dataset is available at study number EGA00001002482 [https://ega459archive.org/datasets/EGAD00001006425]. The H3A-African GWAS pipeline, QC, association testing and fine-mapping approaches are available at (https://github.com/h3abionet/h3agwas).

Chapter 4

4 Estimating the length of the poly-T tract in *TOMM40* and its association with cognitive function in Southern Africans

4.1 Introduction

Apolipoprotein E (*APOE*) ε4 is well-known for its association with increased susceptibility to, and earlier age of onset of, Alzheimer's disease (AD) (59). The three major alleles ε2, ε3, and ε4 encode different protein isoforms which differ in their efficacy in cholesterol transport within the brain (46). The strong association of *APOE* ε4 with neurodegenerative decline was discovered through linkage studies, fine mapping, and later genome-wide association studies (GWAS). *APOE* lies within an ~20kb region on chromosome 19q encompassing the three genes Translocase of Outer Mitochondrial Membrane 40 *(TOMM40)*, *APOE*, and Apolipoprotein C1 (*APOC1*) (54,172).

The variable length thymine (T) homopolymer tract found in intron 6 of *TOMM40* has also been associated with AD risk and the age of onset of LOAD in European, African American, and Asian cohorts (54–60). The length of this poly-T tract has been reported to be associated with levels of cognitive function during normal ageing, cognitive performance in type 2 diabetes patients, and gray matter volume (55,57,60). In the literature the poly-T tract is denoted as the rs10524523 variant (chr 19:44899792) which represents the insertion-deletion (INDEL) where the length of the common T-tract allele is presented as T=20. The T-tract has alleles of variable length; categorised as short (S), long (L), and very long (VL) (as described below), and is polymorphic (140). Poly-T alleles are classified as S where T ≤ 19, L where T equals 20 to 29, and VL where T ≥ 30 (54). The reason for the T=20 allele being the reference allele is likely because the L allele appears at the highest frequency in European populations (56).

Allele frequencies and segregation patterns of the *TOMM40* alleles with ε4 appear to differ between population groups, with a well-documented near-complete LD between the L allele of *TOMM40* and *APOE* ε4 in European populations (54,56,59,60,173). According to the literature, African American (AA) populations and West African (Ghana and Yoruba)

populations have a higher frequency of the S allele when compared to Europeans, and ε4 segregates with all three allele lengths rather than just the L allele as in Europeans (56,58–60,174,175). Furthermore, the range of allele lengths (from T=11 to T=54) differs according to ethnicity, where African populations have much longer VL alleles than those observed in East Asian and European populations (56,173).

Determining the length of homopolymer repeats is a much-explored problem in genomics as copy number variants, short tandem repeats (STR), and homopolymer repeats have been previously associated with neurological disease susceptibility, age of onset, and disease progression, and are believed to be involved in gene regulation of complex traits (176,177). Sequencing stretches of repetitive sequences ubiquitous in eukaryotic genomes is challenging on many levels (176–178). Not only is next generation sequencing limited in this regard, but further problems arise during genome assembly, and automated annotation (177,178). Many existing STR-Callers are sensitive to PCR stutter and noise, and are unable to accurately call and phase alleles (176). Generally, microsatellite analysis involves PCR amplification and fragment or amplicon size determination followed by Sanger sequencing (176,177). Most studies have therefore analysed the poly-T tract in *TOMM40* using PCR-based methods along with large fragment cloning, fluorescence, fragment length analysis, and deep sequencing (52–59,61,173–175,179,180).

A collaboration between two major African studies (mentioned in Chapter 1), the Africa Wits-INDEPTH Partnership for Genomic Studies (AWI-Gen) and Health and Aging in Africa: A Longitudinal Study of an INDEPTH Community in South Africa (HAALSI) provided the opportunity to pilot genetic studies on cognitive function in an African population (13,16). HAALSI baseline data for cognitive function was tested for association with *APOE* (see Chapter 2) and only marginal associations were observed for ε2/ε2 individuals for higher executive function scores and ε3/ε4 for poorer visuospatial performance.

Assessing the LD structure for the *APOE-TOMM40* region in our South African (SA) study population was important as we could not assume that the patterns of segregation and even associations would be similar to those seen in other populations. Utilising the methods by Roses *et al*. (2014) and Nishimura, *et al*. (2017), and manufacturing gene synthesis reference

points along with whole genome sequence (WGS) data, we sought to develop a computational algorithm using both local and global alignment of paired ends for the estimation of the length of the *TOMM40* poly-T tract, and then validating the results by comparing them to those of the *in vitro* PCR-based method. We further compared our results to those of a recently developed STR-Caller known for its benchmarked accuracy against the gold standard for STR genotyping, Haplotype Inference and Phasing for STRS (HipSTR) (176). As we already had *APOE* genotype data (see Chapter 2) for many of the individuals in this cohort, we imputed and phased *APOE* and *TOMM40* poly-T data for our sample combining the 1000 Genomes Project (KGP) African (AFR) samples and WGS data for 100 AWI-Gen samples as a reference panel and using PCR and genotyping results to test the accuracy of imputation. We then applied both the developed algorithm and HipSTR to KGP data for AFR and Northern and Western European (EUR) individuals in order to assess allele length and segregation patterns with *APOE*. We determined allele frequency and allele length distributions for *APOE* and *TOMM40* in our sample and compared them to estimates from previous studies.

A previous study reported that longer *TOMM40* poly-T alleles were associated with AD, where individuals with longer alleles had an earlier age of onset of AD (54). Other studies assessing cognitive performance among non-demented individuals observed higher performance levels in S homozygotes when compared to VL homozygotes (57,60). As such we sought to examine whether or not *TOMM40* conferred any effect on cognitive function under the hypothesis that SS individuals would have higher levels of cognitive performance compared to carriers of longer alleles. Due to the weak associations observed with *APOE* in Chapter 2, regression modelling was based on only two models: (a) *TOMM40* genotypes adjusted for age, sex, highest attained level of education, and HIV status across the entire cohort, (b) comparing cognitive performance of individuals with *APOE* $\varepsilon 3/\varepsilon 3$ (i.e., a sub-set of individuals) across all *TOMM40* length genotypes adjusted for the same covariates as the first model. The second model is based on a similar study by Yu *et al*. (2017) assessing cognitive decline in older adults, specifically $\varepsilon 3$ homozygotes. *APOE* $\varepsilon 3$ had been observed to co-segregate with either S or VL *TOMM40* alleles, and $\varepsilon 3$ homozygotes constituted a larger proportion (~60%) of the population they were sampling from (61).

4.2 Methods

4.2.1 Participants and reference samples

Demographic, phenotypic, and genomic data from individuals recruited through AWI-Gen as part of the Human Heredity and Health in Africa (H3Africa) Consortium, and HAALSI, was used to generate the results for this study (13,16). Data and biospecimens were collected from male and female participants of African descent, aged 40 years and above between 2014 and 2015 (13,17).

4.2.2 APOE genotyping

APOE genotype data was available for 1 776 individuals from our sample. TaqMan™ SNP genotyping was performed at the Sydney Brenner Institute for Molecular Bioscience (SBIMB), University of the Witwatersrand, Johannesburg. Genotype was determined for the two SNPs rs429358 and rs7412 (as explained in previous chapters) using the QuantStudio™ 5 Real-Time PCR System (Applied Biosystems, Thermo Fisher Scientific, Carlsbad, USA) and results were analysed using the QuantStudio™ Design & Analysis Software. When the software was unable to call results; amplification plots, multicomponent plots, and allelic discrimination plots were scrutinised as well.

4.2.3 TOMM40 poly-T homopolymer allele length determination

4.2.3.1 PCR and fragment analysis

DNA samples for 100 SA AWI-Gen participants for which WGS data was available were assessed by custom PCR and capillary-based fragment analysis of the poly-T region of *TOMM40* based on methods previously described by Roses *et al*. (2014) and Nishimura *et al*. (2017) (assays performed at Inqaba Biotec™). Custom gene synthesis for cloned and transformed plasmids for allele lengths T=16, T=18, T=22, and T=35 was performed by Biomatik©. Quality assessment information is supplied in Appendix 4, along with the primer sequences for PCR. The poly-T region was amplified using Q5® Hot Start High-Fidelity 2X Master Mix NEB M0494 (New England Biolabs, MA01938, USA). Fluorescently labelled amplicons and the allelic ladder were added to a LIZ500 sizing standard with Hi-Di™ Formamide 4311320 (Thermo Fisher Scientific, Carlsbad, USA) prior to denaturation. Resulting fragments were run on the ABI PRISM™ 3500xl Genetic Analyzer (Applied Biosystems, Thermo Fisher Scientific, Carlsbad, USA) 50cm capillary array, POP-7™ . Data

was analysed using GeneScan™ and interpreted with GeneMarker HID v2.9.5 (SoftGenetics LLC) and by direct scrutiny of the data.

4.2.3.2 Algorithm for the determination of homopolymer length

The algorithm developed inhouse by Dr Jean-Tristan Brandenburg can be found at https://github.com/jeantristanb/h3adetectpoly (**Figure 4.1**). This algorithm uses both local and global alignment of paired ends to generate estimates of the number of T residues at the region of interest (here 19:45403083 marking the S allele). It uses different software to call the lengths of each allele, and a summary of combined methods is provided to give allele length estimates based on the number of observations. **Figure 4.1** describes the workflow of this homopolymer detection pipeline. In order to estimate the unphased poly-T allele lengths for 100 SA individuals from the AWI-Gen Study, the algorithm was applied to WGS data for each individual and compared to the results of the PCR method above, and to the results from HipSTR.

4.2.3.3 HipSTR

HipSTR was specifically developed for the detection and genotyping of STRs in sequence data generated on Illumina sequencing platforms (https://github.com/HipSTR-Tool/HipSTR) (176). It works by capturing the STR stutter profile of each sample, and then using this along with genomic location and Hidden Markov Models (HMM) to realign reads to produce phased haplotypes (176). Array data for the SA cohort of the AWI-Gen study was incorporated into the analysis to improve length estimation. Genotyping and phasing of homopolymers was shown to be accurate according to benchmarking against gold standard methods employing WGS capillary electrophoresis for accurate STR-typing (176). WGS data for the same 100 individuals in the method above was analysed through HipSTR, and then compared to both the PCR and bespoke detection pipeline.

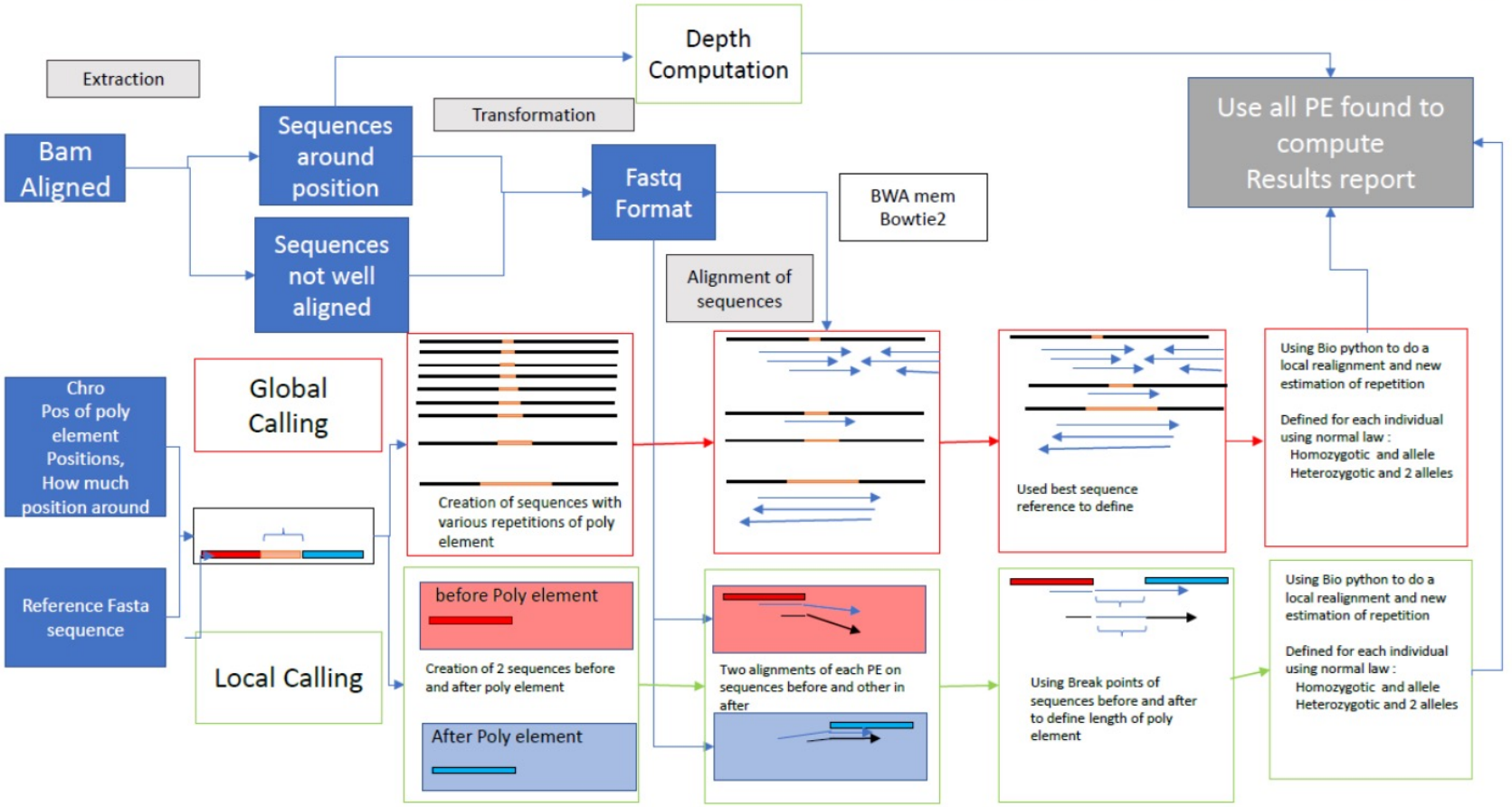


Figure 4.1: Workflow diagram detailing each step in the homopolymer detection pipeline developed inhouse by Dr Brandenburg

4.2.4 Reference panel, imputation, and phasing

4.2.4.1 Reference panel

WGS data for 100 SA individuals from the AWI-Gen study along with WGS data from KGP was incorporated to produce an African reference panel specific to our study. The imputation panel was built using KGP African data (n=579) plus the 100 AWI-Gen WGS samples for which *TOMM40* allele length had been determined. The full reference panel consisted of data for one million positions around *TOMM40* for 679 African individuals. For determination of the *TOMM40* poly-T alleles of each sample, a consensus between sequencing data, HipSTR, PCR and fragment analysis, and our detection algorithm was employed. During allele determination, some difficulty arose due to this locus being a tri-allelic STR where allele length was designated A (S), T (L), G (VL) in order to process imputation.

4.2.4.2 Sequence analysis

Sequence data from KGP for AFR and CEU were analysed to compare allele frequency and haplotype segregation between these two population groups and our sample.

4.2.4.3 Approach to imputation

Imputation was performed employing two different methods using the AWI-Gen SA and AFR KGP reference panel mentioned above:

1. BEAGLE 5.2 (181) (https://faculty.washington.edu/browning/beagle/beagle.html) to call genotypes, phase, and impute genotype data
2. A combination of SHAPEIT4 (182) (https://github.com/odelaneau/shapeit4) to phase alleles and then IMPUTE2 (183) (http://mathgen.stats.ox.ac.uk/impute/impute.html) to call genotypes

BEAGLE 5.2 incorporates both a pre-imputation step which reduces computation time, and utilisation of a reference panel to infer calls (181). The imputation pipeline was documented by Dr Jean-Tristan Brandenburg and may be accessed at: https://github.com/jeantristanb/h3abionetimp_spezone. Pre-imputation of the full dataset was performed by using SHAPEIT4 and IMPUTE2 (on the African reference panel from the Sanger Institute) which employ a Positional Burrow Wheeler Transformation (PBWT) approach to selecting a set of conditional haplotypes to inform phasing (182,184). Thereafter HMM is implemented for more accurate phasing (182,184).

4.2.4.4 Estimation of imputation quality

Estimation of imputation quality for the *TOMM40* T-length polymorphism (S, L, VL) was performed in a four-step process repeated fifty times:

1. Using a partial reference panel minus two individuals
2. Using two genotyped individuals for whom the base pair positions have been extracted from either the H3A array data or that of the pre-imputation panel
3. Imputation using the partial reference panel and the two imputation approaches above
4. Imputation results of the excluded individuals compared to that of the initial calls for *TOMM40*

The workflow is shown in Figure 4.2.

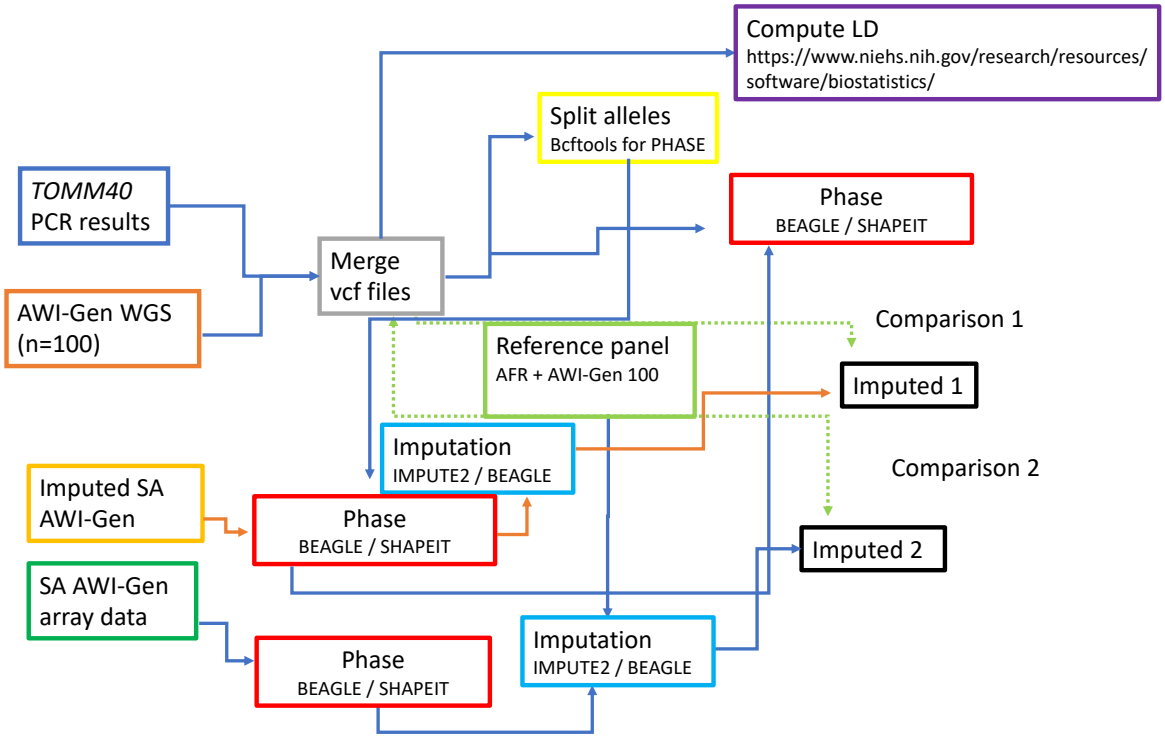


Figure 4.2: Workflow diagram detailing each step in the imputation process developed by Dr Brandenburg

4.2.5 Genetic context of TOMM40 and APOE segregation patterns

Allele distribution for *APOE* and *TOMM40* poly-T was determined for our sample and compared to that of African and European data from KGP. Principal component analysis (PCA) using SNPs within a 0.5Mb region on either side of the *TOMM40 locus* was performed for our sample along with AFR and European KGP data. Segregation of *TOMM40* alleles was compared using pairwise genetic distances between individuals in distance matrices and building evolutionary trees or by principal coordinate analysis (PCoA) in R (https://www.R-project.org/) (R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria) which is a function of the "ape" package (Paradis, E. (2012) Analysis of Phylogenetics and Evolution with R (Second Edition). New York: Springer.)

4.2.6 Statistical analysis

Allele and genotype frequencies were determined for *TOMM40* and Hardy-Weinberg equilibrium (HWE) was tested using the Pearson $\chi 2$ test in STATA 15.0.585 (StataCorp. 2017. Stata Statistical Software: Release 15. College Station. TX: StataCorp LLC). Haplotype frequency was also determined for *APOE/TOMM40*.

Using the results from the imputation, linear regression models for each cognitive outcome were tested for association with *TOMM40* poly-T length alleles. As we had observed only marginal associations of *APOE* in Chapter 2, we focused only on two models for assessing the association of *TOMM40* with cognitive function:

1. *TOMM40* genotypes adjusted for the covariates age, sex, highest level of education, and HIV status in the full dataset
2. *TOMM40* allele lengths adjusted for the covariates age, sex, highest level of education, and HIV status in a sub-set of individuals homozygous for *APOE* $\varepsilon 3$

Summary statistics, regression analyses, and correction for multiple testing were performed in R (R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna. Austria. https://www.R-project.org/). Bonferroni correction was applied for multiple testing.

4.3 Results

4.3.1 Comparison of STR-callers vs PCR methods

PCR results were available for 98 of the original 100 AWI-Gen samples with WGS data. When compared to the PCR method, genotype calls using our bespoke STR-caller for poly-T detection were 81.6% (80 individuals) concordant with the PCR results. Although HipSTR was unable to call genotypes for five samples, the remaining results were 88.2% concordant with the genotypes called by the PCR method. HipSTR was therefore used to detect poly-T length in the KGP data moving forward. Worth noting is that when the results from two STR-callers were compared, excluding the five samples HipSTR was unable to call, only five samples had discordant genotype results between the two STR-callers (94% concordance). Result files for each of the samples are available in Appendix 4.

4.3.2 Imputation quality, calling, and genetic context

Imputation quality results indicated that BEAGLE had a lower error rate (0.08) than SHAPEIT (0.18) when calling *TOMM40* allele length genotypes (**Figure 4.3**).

APOE and *TOMM40* poly-T allele distribution frequencies were determined for the 100 AWI-Gen samples (split into groups by site Agincourt, Dikgale, and Soweto) and for KGP populations of African (Esan in Nigeria - ESN, Gambian in Western Divisions in the Gambia - GWD, Luhya in Webuye, Kenya - LWK, Mende in Sierra Leone - MSL, and Yoruba in Ibadan, Nigeria -YRI) and European (Utah Residents (CEPH) with Northern and Western European Ancestry - CEU, Finnish in Finland - FIN, British in England and Scotland - GBR, Iberian Population in Spain - IBS, and Toscani in Italia - TSI) descent using both STR-callers (**Figure 4.4**). Consolidated *APOE* frequency data were as expected where Africans had higher frequencies of ε2 and ε4 (0.12 and 0.27) than in Europeans (0.06 and 0.15 respectively). Allele length distribution for *TOMM40* showed higher frequencies for the S (0.69 vs 0.43) and VL (0.22 vs 0.10) alleles in Africans compared to Europeans when data for all populations was considered, and the long allele had the highest frequency in Europeans (0.47).

Imputation and phasing of the reference panel and European KGP data showed that segregation patterns between *APOE* and *TOMM40* differ between populations (**Table 4.1**). In

the African dataset ε2 segregated almost entirely with the S allele, but was observed with all *TOMM40* alleles in Europeans. In Africans, just under half of ε4 alleles (43.3%) segregated with the S allele of *TOMM40*, whereas in Europeans 99.3% of ε4 alleles segregated with longer allele lengths (L and VL).

Table 4.1: Haplotype distribution for African panel vs CEU from 1000 Genomes

Alleles		***African data***		***European data***	
APOE	***TOMM40***	**Count**	**Frequency**	**Count**	**Frequency**
ε2	**S**	132	0.115	25	0.025
ε2	**L**	1	0.001	33	0.033
ε2	**VL**	1	0.001	5	0.005
ε3	**S**	525	0.459	410	0.408
ε3	**L**	17	0.150	364	0.362
ε3	**VL**	161	0.141	14	0.014
ε4	**S**	133	0.116	1	0.001
ε4	**L**	84	0.073	72	0.072
ε4	**VL**	90	0.079	82	0.082
	Total	**1144**		**1006**	

PCA results (**Figure 4.5**) show differences in segregation patterns between Africans (a) and Europeans (b) where it appears that the VL allele segregating with ε3 and ε4, and the L allele segregating with ε4 haplotype groups separate in different directions (VL_3 - upper part of the PCA in Africans, lower in Europeans, and the opposite for VL_4, and L_4).

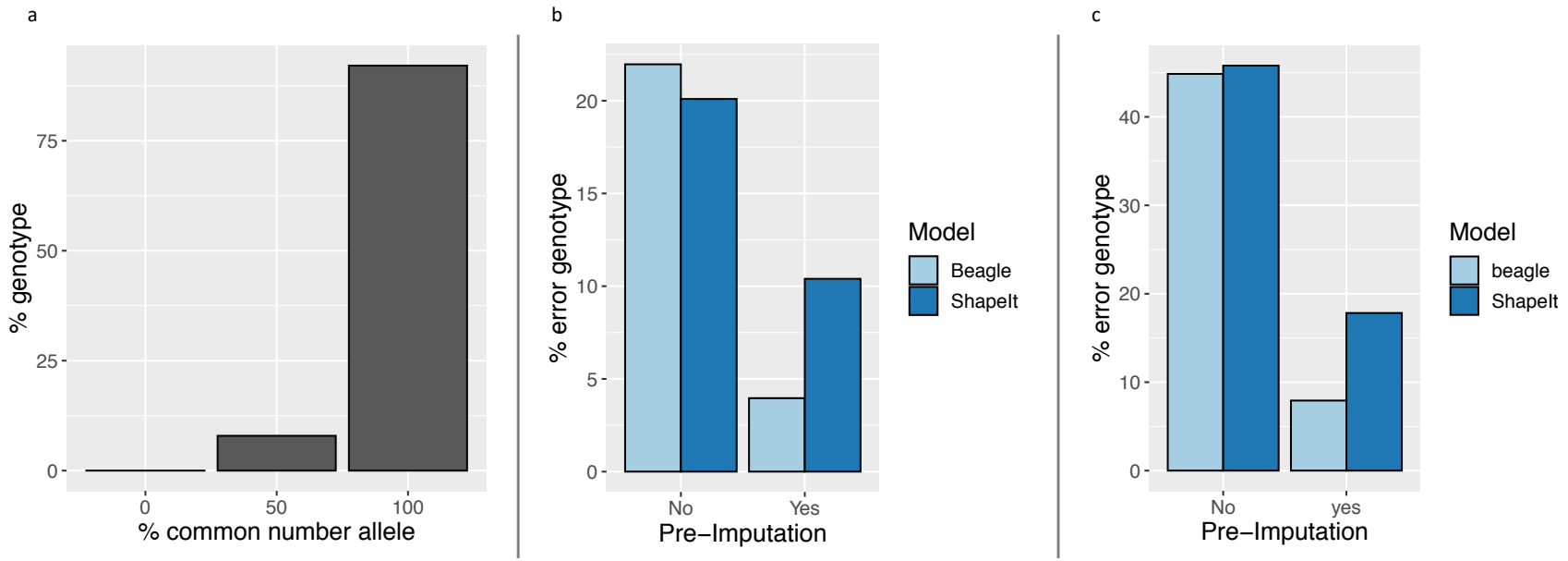


Figure 4.3: Comparison of imputation quality results for BEAGLE and SHAPEIT

(a) shows percentage of concordant alleles for the poly-T tract imputation using BEAGLE vs PCR where 100% and 50% indicate concordance and discordance respectively. Panel (b) shows error estimation for allele calls with pre-imputation applied to BEAGLE and SHAPEIT and (c) shows error estimation for genotype calls with pre-imputation applied to BEAGLE and SHAPEIT.

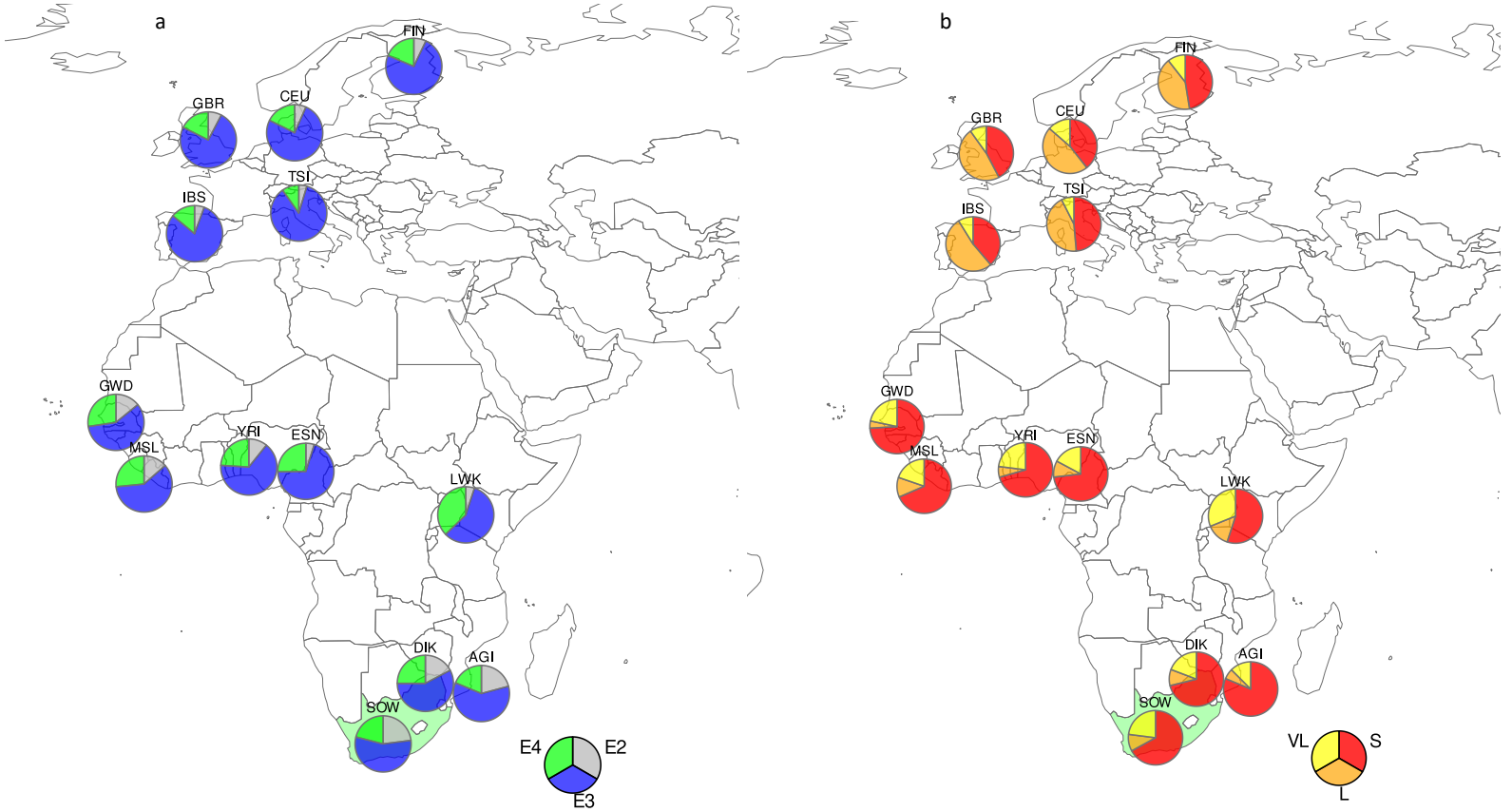


Figure 4.4: Allele distribution of *APOE* and *TOMM40* poly-T length polymorphism in Africans vs Europeans

(a) Distribution map of APOE alleles showing ε2, ε3, and ε4 allele frequencies in each population

(b) TOMM40 poly-T length allele frequency distribution for S, L, and VL in each population

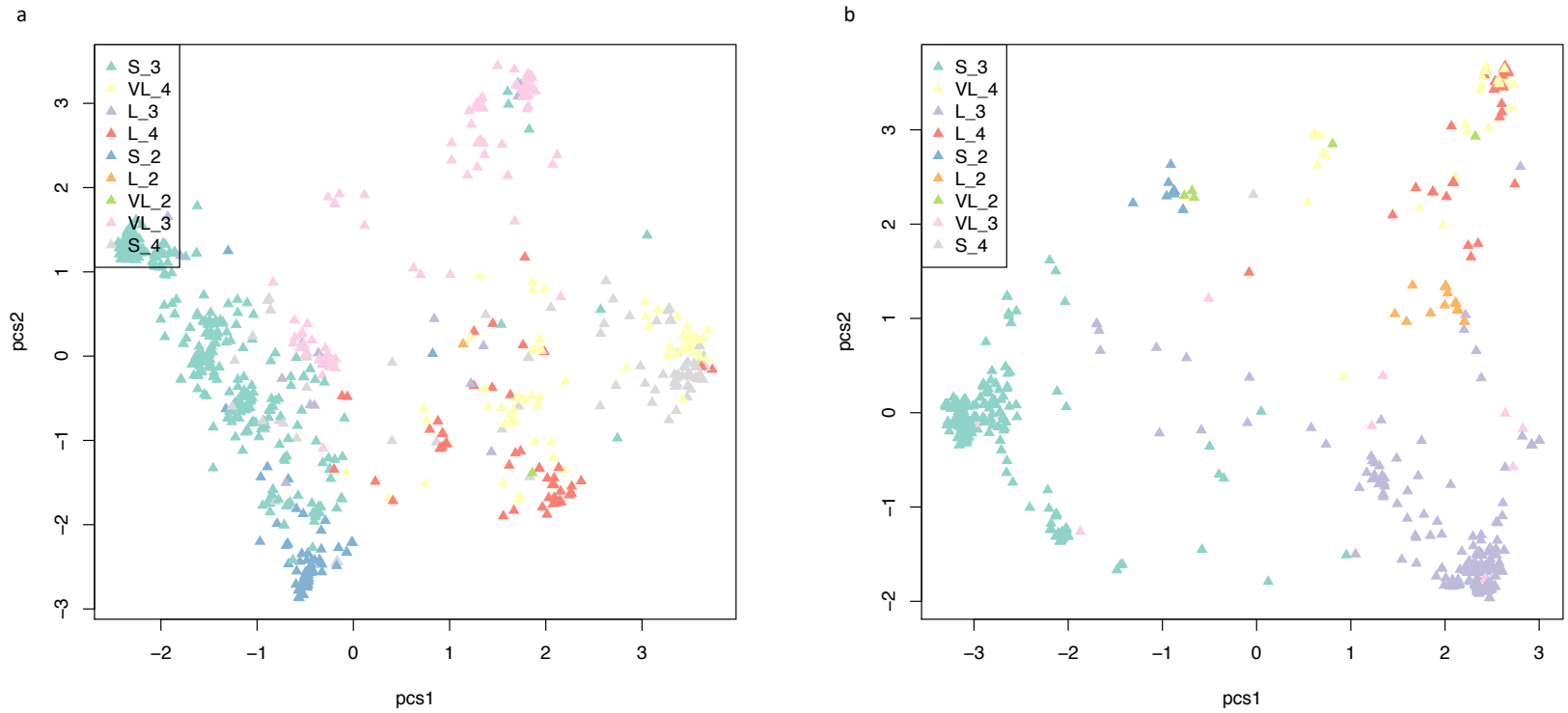


Figure 4.5: Principal component analysis showing the distribution of *TOMM40-APOE* haplotypes based on LD patterns

Haplotypes are coded as per the legend for each PCA plot. (a) shows haplotype distribution in Africans and (b) in Europeans.

4.3.3 Sample description and haplotype distribution

The participants included are those individuals for whom we had *TOMM40* allele length calls after the imputation process (n=2234) (**Table 4.2**). Most of the sample was female (57.63%), and more than three quarters had little to no formal education (77.07%). The full dataset had available data for total cognition score, but cognitive domain data was only available for 1 909 participants. Summary statistics are expressed as means or percentages with standard deviations where applicable. Cognitive domain data are expressed as ranges as they are z-score values per individual.

Table 4.2: HAALSI-AWI-Gen participant summary and phenotypic description

Measure	***n***	***Range***	***Mean or percentage***	***SD***
Age (years)	2234	40-80	57.63	10.86
Sex	2234			
Female			58.46%	
Male			41.54%	
Level of education	2229			
No formal education			39.43%	
Primary			37.64%	
High			19.07%	
Tertiary			3.86%	
HIV status	2232			
Negative			73.25%	
Positive			26.75%	
Cognitive measures				
Total cognition score	2234	0-24	11.69	4.38
Cognitive domains	1909			
Executive function		-4.687-4.328		
Episodic memory		-3.912-1.456		
Language		-2.533-1.279		
Visuospatial cognition		-10.627-4.771		

Allele and genotype frequency distribution for the *TOMM40* poly-T showed that the S allele was most common in our sample (0.732) with S homozygotes making up more than half of the sample (0.537) (**Table 4.3**). Genotypes were in HWE (χ^2=2.7786, df=3, p=0.427).

Table 4.3: Allele and genotype distribution of *TOMM40*

Allele/Genotype	***Number (Frequency)***
S	3269 (0.732)
L	217 (0.049)
VL	982 (0.220)
Total	**4468 alleles**
S/S	1199 (0.537)
S/L	148 (0.066)
S/VL	723 (0.324)
L/L	7 (0.003)
L/VL	55 (0.025)
VL/VL	102 (0.046)
Total	**2234 individuals**

Haplotype segregation patterns showed that the S allele mostly segregated with ε3 in our sample and that each of the poly-T length alleles (S, L and VL) were observed in linkage with ε4 (**Table 4.4**).

Table 4.4: *APOE TOMM40* haplotype distribution for the HAALSI-AWI-Gen participants (N=2 234 individuals)

APOE	***TOMM40***	***Count***	***Frequency***
ε2	**S**	709	0.159
ε2	**L**	10	0.002
ε2	**VL**	5	0.001
ε3	**S**	2002	0.448
ε3	**L**	12	0.003
ε3	**VL**	647	0.145
ε4	**S**	558	0.125
ε4	**L**	195	0.044
ε4	**VL**	330	0.074
	Total	**4468**	

4.3.4 Association analysis

For regression analysis, the intercept was the S/S genotype for *TOMM40* as it is the most commonly observed genotype in this population (**Table 4.5**). In the first model where all individuals were included so as to compare the effects of *TOMM40* genotype on cognitive function, there were three main outcomes: 1. age and highest level of education were significantly associated with cognitive function across all traits where cognitive performance levels were lower with increased age, and higher with higher levels of education; 2. significant sex effects were observed for executive function, language, and visuospatial ability; and 3. in terms of genotype, a marginally statistically significant association was observed where L/L individuals had 0.7 standard deviations lower memory performance score than S/S individuals ($p=0.052$, $\beta=-0.716$, $SE=0.37$). Another possible association was observed where VL/VL individuals had 0.7 standard deviations lower total cognition scores than the intercept ($p=0.082$, $\beta-=0.703$, $SE=0.40$).

In the second model assessing the effects of the combined *APOE/TOMM40* haplotypes in only *ε*3 homozygotes the adjusted R^2 values ranged from 0.221 to 0.398 although the sample size consisted of only 807 participants for total cognition and under 700 participants for each of the cognitive domains. Again, the strong significant effects of age and education were observed but this time an interesting association was seen where being male lowered episodic memory performance by 0.2 standard deviations ($p=0.005$, $SE=0.07$). In terms of genotype, marginal p-values (~0.07) were observed for VL_ε3/ VL_ε3 having more than 1 standard deviation lower total cognition score than S_ε3/ S_ε3 homozygotes, and S_ε3/ L_ε3 heterozygotes performing almost 1.5 standard deviations better at executive function than S_ε3/ S_ε3 individuals.

Table 4.5: Linear regression models estimating the effects of *TOMM40* and various covariates on cognitive function

	Total cognition			**Executive function**			**Language**			**Memory**			**Visuospatial cognition**		
Model	*Coeff.*	*SE*	*p*	*Coeff.*	*SE*	*p*	*Coeff.*	*SE*	*p*	*Coeff.*	*SE*	*p*	*Coeff.*	*SE*	*p*
1 - All genotypes															
S/VL	0.073	0.18	0.693	-0.014	0.07	0.852	-0.027	0.03	0.336	-0.003	0.05	0.956	-0.046	0.11	0.678
S/L	0.090	0.34	0.792	0.039	0.14	0.784	-0.064	0.05	0.220	-0.033	0.09	0.699	0.048	0.21	0.816
VL/VL	-0.703	0.40	**0.082**	-0.165	0.16	0.301	-0.011	0.06	0.850	0.082	0.10	0.401	-0.022	0.23	0.924
L/VL	0.069	0.54	0.899	0.031	0.22	0.890	-0.015	0.08	0.853	0.126	0.14	0.358	-0.020	0.33	0.952
L/L	-0.451	1.48	0.761	-0.017	0.60	0.978	-0.160	0.22	0.473	-0.716	0.37	**0.052**	-1.337	0.88	0.130
Age	-0.087	0.01	<2e-16***	-0.036	<0.01	<2e-16***	-0.013	<0.01	<2e-16***	-0.030	<0.01	<2e-16***	-0.076	0.01	<2e-16***
Male	-0.119	0.17	0.484	0.215	0.07	0.002**	0.055	0.03	0.033*	-0.056	0.04	0.183	0.290	0.10	0.004**
Education	1.607	0.11	<2e-16***	0.844	0.04	<2e-16***	0.363	0.02	<2e-16***	0.351	0.03	<2e-16***	1.398	0.07	<2e-16***
HIV positive	-0.017	0.19	0.930	0.009	0.08	0.909	-0.008	0.03	0.788	-0.046	0.05	0.339	-0.080	0.12	0.492
Adjusted R-squared	0.200	3.91	<2.2e-16***	0.290	1.47	<2.2e-16***	0.331	0.54	<2.2e-16***	0.249	0.90	<2.2e-16***	0.378	2.16	<2.2e-16***

	Total cognition			Executive function			Language			Memory			Visuospatial cognition		
Model	*Coeff.*	*SE*	*p*	*Coeff.*	*SE*	*p*	*Coeff.*	*SE*	*p*	*Coeff.*	*SE*	*p*	*Coeff.*	*SE*	*p*
2 - Ɛ3/Ɛ3 and TOMM40 genotypes															
S/VL	0.099	0.29	0.736	0.073	0.12	0.532	0.016	0.04	0.706	0.037	0.07	0.611	0.088	0.17	0.603
S/L	-3.283	1.97	0.096	1.466	0.83	**0.078**	0.239	0.31	0.439	-0.409	0.52	0.432	1.297	1.22	0.287
VL/VL	-1.036	0.58	**0.074**	-0.261	0.22	0.240	-0.005	0.08	0.954	0.071	0.14	0.607	-0.208	0.32	0.523
L/VL	4.704	2.78	0.091	1.153	1.44	0.423	0.751	0.53	0.160	0.539	0.90	0.550	2.571	2.11	0.223
Age	-0.100	0.01	1.37e-11***	-0.036	<0.01	6.41e-10	-0.012	<0.01	2.77e-08***	-0.028	<0.01	2.36e-14***	-0.070	<0.01	7.69e-16***
Male	-0.200	0.28	0.484	0.249	0.11	0.029*	0.071	0.04	0.093	-0.202	0.07	**0.005****	0.129	0.17	0.441
Education	1.583	0.18	<2e-16***	0.867	0.07	<2e-16***	0.366	0.03	<2e-16***	0.368	0.05	2.86e-15***	1.469	0.11	<2e-16***
HIV positive	-0.035	0.31	0.911	-0.016	0.12	0.899	-0.021	0.05	0.641	-0.049	0.08	0.531	-0.022	0.18	0.906
Adjusted R-squared	0.221	3.91	<2.2e-16***	0.318	1.43	<2.2e-16***	0.343	0.53	<2.2e-16***	0.251	0.90	<2.2e-16***	0.398	2.10	<2.2e-16***

*Standardised coefficients represent estimated changes in cognitive function for each model. Models are denoted in bold to highlight the effect of interest, with predictors in italics. Domain estimates (executive function, language, memory, visuospatial cognition) change in standard deviation units. Adjusted R-squared values per model with residual standard error (SE) is shown. *p < 0.05, **p < 0.00001, ***p < 2e-16. Bold p-values indicate marginal significance or associations of interest mentioned.*

4.4 Discussion

In studies in European populations, the high linkage disequilibrium between *APOE* and *TOMM40* led to the conclusion that the causal allele for association with AD was *APOE* ε4 and that *TOMM40* L alleles were largely "hitchhikers" on the same haplotype block (57,58,60,185). The study described in this chapter was initially conceptualised to extend the genetic association between *APOE* and cognitive function in Africans by including *TOMM40* into the association analyses. We sought to understand the independent association of *TOMM40* genotypes with cognition in an African population. To do so, we first needed to determine the allele lengths of the poly-T region for our sample, which proved to be a considerable challenge. To maximise sample size while being mindful of the cost of laboratory assays, it was decided that we would develop and validate an *in silico* STR-caller capable of accurately calling homopolymer length alleles.

An inhouse algorithm for calling *TOMM40* poly-T alleles was developed by Dr Brandenburg and successfully validated by sequencing and comparing the estimates with another STR-caller, HipSTR. Utilising these tools, we further tested different imputation and phasing methods for this tri-allelic locus (S, L and VL). The aim was to determine linkage and segregation patterns between *APOE* and *TOMM40* for African vs European populations using KGP data combined with a sub-set of our AWI-Gen data. And finally, we imputed and phased *TOMM40* and *APOE* for a dataset of over 2 000 participants with genotype and cognition data for regression analysis. The model that included analysing all the samples together in terms of *TOMM40* genotype proved more informative than that same model for *APOE* for the cognitive domain traits, and the second model that included analysing the effect of *TOMM40* haplotypes in a subset of only ε3 homozygotes in association with cognition marginally improved upon the first.

The detection and estimation of homopolymer repeat lengths is highly problematic in any PCR-based assay or sequencing strategy, including NGS (176–178). Knowing that it was integral to this research to be able to determine the poly-T length alleles of *TOMM40*, we avoided the costly laboratory-based capillary electrophoresis by developing an inhouse computational method which we compared to the results of a recently developed STR-caller which had been validated against the gold standard for STR estimation. For a small number

of samples for which we already had WGS data, we performed the laboratory-based PCR assay and fragment length analysis methods followed by Nishimura *et al*. (2017) to ensure that thorough consideration was given to the accuracy of the approach prior to implementing these workflows into this research.

Rather than using the position of rs10524523 (representing the L allele) to mark the position of the poly-T in our imputation, we instead shifted the position of the marker to represent the S allele in our sample as previous literature had already indicated that L alleles were not as frequently observed in African populations (56). Our algorithm was able to accurately determine the genotype for 81.6% of the calls compared to our validation sample according to fragment analysis results. This was comparable to the findings using HipSTR which called 88.2% of the genotypes correctly. Results from the inhouse caller were more easily interpreted, and the addition of sequencing depth analysis assisted with genotype estimation when calls were uncertain. The benefit of using HipSTR was that it also phases the STR and allows for the incorporation of a reference panel (176). By utilising both these methods and comparing them to *in vitro* results for the same samples, we are confident that the bespoke algorithm may be used to estimate the length of homopolymers or other STRs in future studies either as an alternative to, or in conjunction with HipSTR as was done here.

Imputation and phasing of *APOE* and *TOMM40* was performed using BEAGLE 5.2 and then SHAPEIT4 and IMPUTE2 and results were compared as proof of concept using the KGP data and the 100 AWI-Gen samples with WGS data for which we had already determined poly-T lengths using the PCR-based electrophoresis method. Pre-imputation was applied to both methods in order to reduce computation time. However, allele length polymorphisms had to be split and then recombined as the software could not impute tri-allelic markers (181). Imputation was performed for 1 million SNPs centred around our target position for *TOMM40* (19:45403083) to produce enough information to accurately assess co-segregation between markers.

Imputed results comparing African to European data showed that African populations have a higher frequency of the S allele (0.691 vs 0.436) which was present in the African sample as three haplotypes with linkage to each of the *APOE* alleles. This was previously reported by Roses et al. (2014) who observed that African Americans and Yorubans from Nigeria had a

trimodal distribution when assessing linkage of *TOMM40* alleles with ε4, whereas those of European ancestry showed apparently complete linkage of ε4 with the L allele. Furthermore, all *TOMM40* alleles showed some linkage with ε3 in those with African ancestry but only with S and VL in Europeans (51,174). Our data indicate that the S and L alleles occur at similar frequencies in Europeans, but the previous claim for complete linkage between ε4 and the L allele is inaccurate. Rather one may say that ε4 segregates almost entirely with longer (L and VL) alleles in European populations, whereas this is not so for Africans, nor is it true for East Asians (173,174). These differences in segregation patterns between ethnicities may be indicative of different evolutionary events occurring in different ethnic groups (54).

The final imputed dataset in over 2 000 individuals of African ancestry with comprehensive cognitive function data had *TOMM40* allele frequencies similar to what had been observed in other African populations. Our sample had a higher frequency of the S allele, and the L allele was observed at the lowest frequency. Furthermore, in the second regression model there were no L/L individuals amongst the ε3 homozygotes. We observed higher frequencies of ε2-carriers than in European studies allowing us to observe the linkage between ε2 and *TOMM40* where it is in almost complete LD with the S allele (~98%). This may have implications for susceptibility to AD, as S homozygotes have been reported to display better cognitive performance than VL homozygotes and it has been repeatedly suggested that ε2 may have a protective effect with regard to AD risk (57,60). A possible buffering effect may be at play where the S allele somehow mitigates the harm of ε4. Earlier studies have reported that age of onset of AD is later in S_ε4 haplotype-carriers than in ε4-carriers who have longer poly-T length alleles (51,54,174).

To assess whether or not poly-T length was associated with cognitive function, our first model looked at it in terms of *TOMM40* genotype as per the first model in the *APOE* associations as that was the most informative model from that study (Chapter 2). The strong significant associations between highest level of education and age were observed throughout the regression analyses, as was some improved performance (at least in visuospatial ability and executive function) in men. This effect is assumed to be the result of educational disparity between the sexes in this community, where women have less access to formal education either through cultural or financial limitations (14,29). Although no significant associations were observed for any of the genotypes, modelling for association with

TOMM40 accounted for marginally more of the variation of the cognitive domain phenotypes than *APOE* in this population. A marginal association was observed between L homozygosity and verbal episodic memory, where performance levels were more than half a standard deviation lower when compared to S homozygotes. The same drop in total cognition score was observed in VL homozygotes when compared to S/S individuals albeit not significantly. This aligns with studies indicating that S homozygotes perform better at memory tests than VL homozygotes or carriers of longer alleles (55,57,58,60).

In order to be aligned with previous studies assessing the association between the *TOMM40* poly-T and neurocognitive traits, only ε3 homozygotes were included in the second model. A recent study on non-Hispanic African Americans suggested that S_ε3 homozygotes experience faster cognitive decline than non-S-carrying ε3 homozygotes (186). This recessive effect of the S allele on cognitive decline had also been reported by Yu et al. (2017). Different rates of decline were observed between African vs European S_ε4-carriers where the rate of decline was slower in Africans (186). An interesting significant association emerged from the second model, where men had lower memory performance scores than women which is unexpected and goes against what we have seen in our results thus far. Also, a full standard deviation reduction in total cognition score between S_ε3 homozygotes and VL_ε3 homozygotes was observed but was not statistically significant. Post-hoc power calculations of the reduced sample size (only ε3 homozygotes) suggest that for these associations, the study was less than 40% powered to detect significant associations ($p=0.001$) with an effect size 0.02, and therefore requires further investigation in a larger sample size.

This study complimented what was performed in Chapter 2 for *APOE*, by determining allele and genotype frequency distribution of *TOMM40* for the same study sample, and then testing associations between the poly-T tract and cognitive function. A new method for the estimation of homopolymer length was developed to detect the length of the poly-T tract in *TOMM40* from WGS data and was validated against PCR fragment analysis results. This method was assessed alongside an already benchmarked STR-caller and may confidently be applied to future studies. The rigorous testing and comparison of phasing and imputation methods added to the strength of the results. Attempting to delve deeper into the genetic context of *TOMM40* and its co-segregation with *APOE* added an extra dimension to assessing

the differences in haplotype frequency distribution between Africans and Europeans. In future studies we may perform association analyses with the haplotypes, rather than *APOE* (Chapter 2) and *TOMM40* (Chapter 4) separately. This may reveal associations with the region and could help pinpoint potentially causative loci. Exploring the epistatic interaction between *TOMM40* and *APOE* may further provide insight into the relationship between these two markers and cognitive decline.

The evolutionary aspects of this research are of great interest but have yet to be teased out. Evolutionary tree diagrams comparing the African reference panel used in this study to the KGP Europeans may be found in Appendix 4. We have added to extant data by improving the resolution of LD patterns between African and European populations. The sample size was small, and power was insufficient to detect significant associations observed in studies of the combined effects of *TOMM40* and *ε*3 homozygosity, but it appears that *TOMM40* was marginally more informative for cognitive function than *APOE* for our sample.

Chapter 5

5 Discussion and conclusions

5.1 Overview

This is the first population-based continental African study of this size exploring the genetic aetiology of normal cognitive function. It is embedded within two high profile African collaborative studies, Health and Aging in Africa: A Longitudinal Study of an INDEPTH Community in South Africa (HAALSI), and Africa Wits-INDEPTH partnership for Genomic Studies (AWI-Gen) (13,16). Overlapping sampling from these studies, which capitalised on the longitudinal research infrastructure of the Agincourt health and socio-demographic surveillance system (HDSS), provided the complementary genomic data integral to this analysis.

Baseline cognition data was captured with two different tools; a culturally-adapted and translated mini-mental state examination, and the OCS-Plus specifically developed for low-income low literacy settings and validated in this cohort (2,25,29). Together, these provided thorough phenotypic assessments of five well-defined cognitive traits (total cognition score, executive function, language, verbal episodic memory, and visuospatial ability) to support genetic association studies in a well-characterised cohort of ~2 000 individuals aged 40 years and older. The research presented here consisted of three objectives, each using cognitive performance scores derived for each of the traits mentioned above, as well as socio-demographic (age, sex, educational attainment) and health-related (HIV-status) data for the participants described above and extensively documented in the previous chapters of this thesis.

5.1.1 APOE genetic variation and its association with cognitive function in rural-dwelling older South Africans

The purpose of the first part (Chapter 2) of this research was two-fold; firstly to determine the allele frequency distribution at the *APOE* locus in our participants using a TaqMan genotyping method (validated with sequence information for a small number of genotyped individuals), and then to test the association between *APOE* genetic variation and each of the five examined traits representing cognitive function. The approach was based on similar study designs and previous findings in other populations, some of which indicated that ε4-

carrier status is associated with cognitive deficits and lower cognitive performance scores in "normal" cognition (47,78,80,82). The allele frequency distribution in our participants was similar to that observed in other African populations including the 1000 Genomes (KGP) African Super Population and smaller African studies in eastern, western, and southern Africa where allele frequencies of ε4 and ε2 are higher than those observed in the other KGP population groups (Europeans, East and South Asians, and South Americans) (48,67,89,90,98,133). Additionally, the value of studying *APOE* variation in Africans is that with the higher ε2 allele frequencies it is possible to perform better powered comparative studies on the effects of ε2 as well as ε4, as this has been a drawback in elucidating the effects of ε2 in various diseases in other populations (187).

Region-specific *APOE* allele frequency variation was also highlighted in our study between participants from different regions of Africa. A small study in Kenya had significantly different allele frequencies to our South African study sample. Although the significance of the comparison is questionable due to the small sample size of the Kenyan study (Chen *et al*. (2010)), it highlighted the misconception that representation from West Africans (Yoruba) in genomics studies is adequate for the whole continent (31–33). Allele frequencies differ significantly across African populations and therefore more African populations need to be investigated. The ages of the participants of the smaller African studies used for comparison may also have affected the observed allele frequencies. The higher ε4 allele frequency in the Joska *et al*. (2010) study is expected as it has been suggested that ε4 is advantageous to infant survival (94,95). The studies from Kenya (n=84) and Nigeria (n=459) were performed on older participants as they were studies assessing association of ε4 with dementias and AD (48,90). The Ugandan study by Willis *et al*. (2003) was performed on a group of participants aged 17-80 years, and the *APOE* allele frequencies from this study were most similar to our study sample. The higher frequencies of ε3 in the Kenyan and Nigerian studies may also be related to the increased morbidity and mortality associated with ε4, especially at older ages (187).

Statistically significant associations between *APOE* genotype and the cognitive domains; executive function and visuospatial cognition, did not survive correction for multiple testing. But when compared to the most commonly occurring genotype (ε3/ε3), ε2 homozygotes scored approximately half a standard deviation higher in executive function, and ε3/ε4

individuals ~0.3 standard deviations lower. Here the covariates of age and education contributed more strongly to the significance of the model. Sex differences were apparent as men performed better in the cognitive domains of language, executive function, and visuospatial ability. This is likely a result of disparities in access to, or attainment of education in this community (29). No significant associations were observed between *APOE* genotype and cognitive function for the models comparing the performance of ε4-carriers to ε3 homozygotes, the rest of the cohort to ε4 homozygotes, or ε2 homozygotes to ε4 homozygotes. In the latter model, the effects of age and level of education were not as significant (age from $p < 2e\text{-}16$ to $p < 0.5$, and education from $p < 2e\text{-}16$ to $p < 0.00001$). It may be that these environmental factors influence cognition to a lesser degree in the presence of ε2 and ε4 homozygosity as these are representative of the extremes *APOE* genotype in terms of cognitive function.

5.1.2 Genome-wide association study of population standardised cognitive performance phenotypes in a rural South African community

The directive behind the study reported in Chapter 3 was to use a genome-wide exploratory approach to identify genetic associations with cognitive traits. This non-hypothesis-based approach assessed allelic associations for SNPs genotyped using the H3Africa array (https://www.h3abionet.org/h3africa-chip) enriched for common variation in African populations. Five separate genome-wide association studies (GWAS) were performed for each of the cognitive phenotypes mentioned above using data imputed from the array and a population-specific imputation reference panel. The top independent signals were documented. Two genome-wide significant signals for association with cognitive phenotypes were observed, along with strongly suggestive signals. The highly significant variants associated with verbal episodic memory - one novel, and an African-specific variant associated with language ability were emphasised here, along with suggestive signals for language performance and global cognition (total cognition score).

The first genome-wide significant signal (rs73485231, $p=8.99e\text{-}09$, $\beta=0.229$) associated with improved verbal episodic memory lies within an intergenic region which has previously been associated with adolescent idiopathic scoliosis (141). Although this is the first time this variant is being associated with cognitive function, it's located between two genes, *GNG5P5* and *HTR2A*, which have previously been implicated in substance use and intellectual

disability, makes this an interesting finding. It implies possible involvement in common biological pathways affecting neurocognitive phenotypes and behaviours. Furthermore, this common African variant has a higher minor allele (A) frequency in African populations (MAF=0.132) than in other KGP population groups. The A allele was more prevalent in this study sample (14.1%) and AG heterozygotes performed significantly better at memory tasks than GG homozygotes. This is an example of how common African variation may assist in the characterisation of more genome-wide markers for association with complex traits.

A novel rare variant, rs140372794, also associated with improved memory performance reached near genome-wide significance (p= 8.59e-08, β=0.324). The effect allele occurred at a frequency of 0.053 in the study sample and was not observed in the Asian or European KGP super populations. Its proximity to a non-protein coding RNA gene with several SNPs associated with general cognitive function, mathematical ability, and AD pathology have been detected lends credibility to the association of this novel variant with episodic memory. This once again supports the strength of using study-specific tools used to capture both genomic and phenotypic data in this study.

The second genome-wide significant signal (rs139259554, p=4.84e-08, β=0.695) was for an African-specific variant found to be polymorphic only in African populations. Although the effect allele (G) is rare (occurring at 1.1% in the population), it was associated with a modest increase in language test performance and overlapped with another near-genome wide significant rare African-specific signal within < 50kb of it. No previous associations had been reported for this SNP as it is monomorphic in other populations, but it is found within a gene that has been associated with susceptibility to vascular dementia (168).

A common variant rs34595826 on chromosome 7, occurring in higher frequencies in African populations compared to other ethnic groups, had a dosage effect where homozygotes for the A allele performed significantly worse in language tests than those who did not carry the A allele. Interestingly, this suggestive (p=1.21e-07, β=-0.166) variant lies within a region previously associated with educational attainment (which is often used as a proxy phenotype for cognition due to high correlation between these phenotypes), schizophrenia, and AD pathology (142,143,147). The Manhattan and Locus zoom plots (***Figure 3.3***) for this variant

were quite convincing, clearly showing a supported peak with many SNPs in high LD with the target.

Another African-specific variant rs138832740, although of only suggestive significance (p=1.10e-07) was found to have a large effect (β=-2.004) on total cognition score. Once again, no previous associations were documented due to these African-specific SNPs being monomorphic in European populations upon whom most GWAS are performed. Despite this, the results are encouraging as several variants in the surrounding region of chromosome 1 have been reported to be associated with cognition and educational attainment, neurocognitive and neuropsychiatric conditions, and behavioural traits including substance use (150–158).

Replication of signals previously associated with episodic memory and language (results in Chapter 3) were challenging as they were either not present in this dataset or failed to reach relaxed significance levels of p below 0.005. The lack of significant associations is likely due to sample size of our study, compared to studies that are orders of magnitudes larger, but may also be indicative of a lack of association on the African continent. Part of the strength of our study lies in the use of the custom H3A array that has common African variants that are either monomorphic or rare in other populations and therefore increases discoverability in an African dataset.

5.1.3 Estimating the length of the poly-T tract in TOMM40 and its association with cognitive function in Southern Africans

The research reported in Chapter 4 involved a study to develop and validate a computational method for estimating the length of the poly-T tract in *TOMM40* (S - short, L - long, VL - very long) using PCR fragment data and WGS data from the AWI-Gen study and the KGP African dataset. The aim was to use this algorithm to determine allele distribution of the TOMM40 length polymorphism, its co-segregation with *APOE* alleles, and to assess genetic associations with *TOMM40* on cognitive function through regression modelling. The hypothesis for association was based on findings by Greenbaum *et al*. (2014) where *TOMM40* S homozygotes performed better in cognitive batteries than VL homozygotes. This was tested in the HAALSI/AWI-Gen sample.

Results from a bespoke algorithm, created inhouse, were validated against and compared to both an *in vitro* PCR and fragment-based analysis and to another *in silico* STR-caller (HipSTR) which had already been benchmarked against the gold standard for STR-typing. The results from Dr Brandenburg's caller algorithm were highly concordant (94%) when compared with the results from HipSTR, of the 100 AWI-Gen samples with whole genome sequence data were used to validate the experiment. When compared to the PCR results, concordance was ~81% compared to HipSTR's 88% concordance. Hence HipSTR was selected for estimating poly-T length in the KGP African and European populations. The development of the inhouse tool further attested to the excellence and bioinformatics abilities of the SBIMB AWI-Gen team, and the importance of nurturing genomics capacity in Africa.

HipSTR was applied to WGS data from the AWI-Gen 100 and African and European KGP (133) populations. Results showed differences in allele and haplotype distributions between the combined African sample and the pooled European sample. Africans had significantly higher frequencies of the S allele, and lower frequencies of the L allele than Europeans. *APOE* allele data was also extracted from KGP sequence data and used to assess linkage between *APOE* and *TOMM40* in the KGP populations. The resulting segregation patterns challenged previous assumptions stating near complete LD between ε4 and the L allele in Europeans (52,54,174). Our work was able to shed light on this misconception as we observed near complete linkage (99.3%) between ε4 and the longer *TOMM40* alleles (L and VL) in the European KGP populations, rather than complete linkage between ε4 and the L allele as previously described. The findings in European populations could have been due to the low frequency of the VL allele in Europeans (in KGP, 0.10), as well as the use of AD cohorts in which the enhanced ε4 frequency means that the likelihood of linked L alleles is higher. The use of *in sillico* tools providing high resolution STR detection may also have contributed to the observed differences as there was no previous evidence that the KGP data had been used for *TOMM40* poly-T studies. The implications of the segregation results present some interesting questions regarding the evolution of this chromosomal region and how it could be linked to biological pathways that involve the way in which *APOE* and *TOMM40* interact.

Post-validation of the algorithm and creation of a reference panel combining the WGS data from KGP Africans and the AWI-Gen 100 for imputation, *TOMM40* genotypes were imputed and phased with *APOE* genotype data. This dataset was then used in linear regression models for association analysis with cognitive function phenotypes. A single near-significant association (p=0.052) was observed for L/L individuals for an ~0.7 standard deviation reduction in memory performance in the model assessing cognitive function across all genotypes. The second model, comparing cognitive performance between *APOE* ε3/ε3 S and VL homozygotes, failed to replicate the findings by Greenbaum et al. (2014). They reported that S/S individuals performed significantly better in tasks of executive function and episodic memory than their VL/VL counterparts (60). Although our results observed that VL homozygotes had 1 standard deviation poorer performance in executive function than S/S individuals, this result was not significant (p=0.074). Furthermore, although the statistical methods used differed between this study (multiple regression analysis) and that of Greenbaum *et al*. (2014) (Multivariate Analysis of Covariance), the results are equivalent despite having different parameters. Our analysis is based on predicting cognitive outcome whereas theirs looked specifically at the differences in variance between the genotypes. Once again, the models confirmed the effects of age, sex, and educational attainment on cognitive function. Despite the modest improvement of the *TOMM40* homopolymer length polymorphism regression models for the cognitive domains compared to that of *APOE*, a larger sample size may be required to detect the genetic contribution to cognitive function of a single locus. Exploring the interaction between *APOE* and *TOMM40* to determine the effects each gene has on the other would provide interesting insights into observed population differences and inform how this interaction influences cognitive decline especially in AD.

5.2 Strengths and limitations

5.2.1 Strengths

This research is a forerunner for larger studies on the African continent focusing on the genetics and epidemiology of cognitive function, cognitive ageing, and neurocognitive decline typical of dementias. Leveraging on the power of international collaboration and funding opportunities coupled with already established African research centres, we were able to generate high quality phenotypic and genotypic data for association studies. The

strengths of this research lie in the use and development of specialised tools for the capture and analysis of genomic and phenotypic data, as well as the cross-disciplinary nature of the collaborating research centres involved.

The translation and cultural adaptation of a cognitive assessment battery used in the US Health and Retirement Study (HRS) proved adequate to measure a total cognition score in this population, but only after removing the questions testing numeracy. Previous analyses by the HAALSI group indicated that the numeracy results showed an educational attainment bias when they were included (2,29). Knowledge of the community and its social, demographic and epidemiological history - along with sustained levels of community engagement - were essential to the success of this study. Previously reported low levels of education and literacy amongst the rural older adult population in Bushbuckridge informed the need for a different method of capturing latent cognitive function within this community (14,25). The OCS-plus which was validated for use in the HAALSI cohort proved to be a better method for accurately assessing cognitive function in this community, as it was not reliant on participants being either literate or numerate (25). Furthermore, the instrument was administered using a tablet so that the direction and timing of all "movements" could be tracked, and verbally-based testing was audio-recorded for quality control purposes (25). Study participants were familiarised with both cognitive screening tools prior to test administration in order to exclude those participants who were likely already suffering from some form of cognitive impairment (2,13,29). Thus this cohort of consented individuals aged 40 years and above represent a population cross-section of "healthy" individuals serving as a baseline for future cognitive research.

For each of the objectives, a relatively cost-effective and efficient alternative method was employed to generate genomic data. TaqMan SNP genotyping was used for *APOE* genotype determination as arrays are notorious for being unable to accurately type this locus (188). Validation by sequencing was performed on a small number of samples to reduce cost at the same time as ensuring the accuracy of *APOE* genotyping. Existing genetic data for the participants in this study from the H3Africa array, customised to represent common African variation, was used. The use of an African-specific array strengthened the resolution by tagging smaller LD blocks and providing data on markers previously unavailable on other arrays (31–33). Underlying genetic substructure and regional variation is a confounder for

GWAS and should be carefully considered when choosing reference panels for imputation. The additional customisation of a South African-specific reference panel for imputation quadrupled the number of SNPs for association. The H3Africa Consortium has greatly contributed to the increase in genomic research and bioinformatics capabilities in Africa, and highlights the importance of addressing the regional variation observed across the continent in genomic studies (12,16,69,189). By choosing to use an inhouse *in silico* method for the estimation of homopolymer lengths for *TOMM40*, we demonstrated that it was not essential to use alternate laboratory tests to confirm the allele lengths. Thereby, costs were greatly reduced as validation using PCR-based fragment analysis was only required for 100 participants for whom we already had WGS data to test the algorithm. Rigorous testing of Dr Brandenburg's STR-caller against benchmarked methods for STR-typing and an existing high resolution STR-caller was performed to ensure that the final dataset was valid. Along with this, the use of 1000 Genomes Project (133) data for characterisation and allele frequency determination of *TOMM40* and *APOE* was novel.

5.2.2 Limitations

Although the strengths of this research outweighed the limitations, it is important to consider how study design, sample selection, and sample size may limit current and future research in this cohort.

More than a quarter of this community is currently living with HIV but at the time of data collection, the assessment and/or diagnosis of HIV-associated neurocognitive disorder was not considered. There was evidence that members of the study sample were likely to be HAND sufferers but this could not be confirmed. The inclusion of these participants may have affected the genotype distribution in terms of resulting cognitive performance, as the effects of ART on cognitive function have not been assessed here. Furthermore, the HAALSI team were not able to accurately diagnose dementia in the cohort although there was interest in AD and dementia when the initial study was conceived. Thus this series of cross-sectional studies focused on normal cognitive function rather than a pathological cognitive state such as Alzheimer's disease. Exclusion of participants with incipient dementia was not possible, nor was the incorporation of a case-control study into the research objectives.

Knowing that the small sample size of this study compared to the meta-analyses portrayed in the literature would be problematic in terms of replicating significant results, sample size was maximised by excluding individuals per trait and covariate per study objective with missing data rather than across the whole cohort. The power calculation performed prior to initiation of this study estimated that the study was 80% powered to detect associations of common variants (MAF>0.10) with beta scores above 0.625 with decreases on either parameter reducing the overall power to detect associations and increasing the chances for type 1 errors. Effect sizes for the power calculation were determined using summary statistics from the GWAS Catalog (137) across relevant cognitive phenotypes. The power of this study was in itself a limitation as the expectation is that each genetic marker assessed for association with cognition (as a complex trait) would have very small effect sizes compared to the 0.625 beta value used in the calculation above. To add to that, the detection of novel or rare variants with smaller effects would be more prone to false positive results. Although the cognitive phenotypes could have been combined using tools such as Multi-trait analysis of GWAS (MTAG) (190) to increase the power of the study, we relied on having more discrete phenotypes rather than a single cognitive metric. Despite this, we were able to detect marginally significant associations (which notably did not survive correction for multiple testing) between *APOE*, *TOMM40* and cognitive function, and confidently identify novel variants associated with cognitive function through GWAS.

The inability to replicate GWAS association signals is correlated with sample size. Compared to larger studies, this cohort (~2 000 participants) was significantly smaller than GWAS and meta-analyses with up to 100 000 individuals. Differences in arrays and imputation panels used to capture and generate population/disease/trait-specific genetic variation also contributed to our inability to replicate previous studies observations. Results in Chapter 3 showed that some previously reported SNPs were not present in the AWI-Gen dataset, and also that African-specific variants - only polymorphic in African populations - had little information regarding previous associations with traits or disease states.

In terms of study replication, currently an appropriate replication cohort for our GWAS is not available. Part of this is due to a lack of African cognitive and genetic data in the comparable participant datasets, and partly because different studies measure and define cognitive traits differently. Meta-analyses often merge different imperfectly correlated phenotypes. This

highlights the importance of having harmonised studies. With future waves of AWI-Gen and HAALSI data collection there may be an opportunity to replicate this research. As this research may be considered a pilot to larger and more comprehensive studies, it was more focused on methodology and proof of principle than it was on discovery and replication outcomes.

5.3 Future considerations

With the second wave of HAALSI cognition data becoming available (and a third wave expected in 2021), this longitudinal data would provide further time points to assess the rates of cognitive change and decline as the cohort ages. Analysis of the longitudinal cognition data may reveal cohort-specific trends in cognitive function, as studies have shown that the different cognitive domains show different trends (stability vs decline and rate of decline) with age (117). At these time points, it would also mean that the older surviving individuals in the study would now be over the age of 80 years, possibly allowing for more research into cognitive decline in the elderly as opposed to in the middle-aged to elderly, as once again different trends and even different genetic factors may come into play. For example ε4 is suggested to provide increased cognitive reserve in younger ages, but also result in faster cognitive decline at much older ages especially in the presence of HIV (75,83).

The older people in this community have an unusual social or household status which appears to shift depending on the structure of each household and its response to epidemiological events (191,192). Rather than being dependent or adding financial strain, older members of families are sometimes the most productive members of the household often contributing directly to food and utilities with government-subsidised pensions and/or being employed (191). The third wave of HAALSI data collection would then provide even more information on the rate of cognitive decline and possibly give an indication of the cognitive health of the older population in this community to better inform public health practice, especially since older women are often primary caregivers for their grandchildren. This is partly due to the high prevalence of HIV and HIV-related mortality between 2000 and 2008, but also an artifact of high unemployment rates (73.5% in the HAALSI cohort according to Gómez-Olivé *et al*. (2018)) (5,13,14,191).

It may be of particular interest in this community to look at HIV and HIV-associated neurocognitive disorder (HAND). The prevalence of HIV is relatively high (~26%) despite the age demographic. A large proportion of people are being treated with ARVs - reflected in steadily rising uptake - which may also have an effect on cognitive function (5,15,193). The effects of *APOE* ε4 on HIV status/infection and HAND is important as it may exacerbate brain atrophy in seropositive individuals over the age of 60 (82). Diagnosis of disorders or dementias in this cohort would enable the possibility of performing case-control studies to look at genetic factors which predispose to neurocognitive or neurodegenerative pathology. In addition, how Covid-19 and infection with SARS-CoV-2 is affecting cognitive function in this community may provide insight into the long-term effects of Covid infection were infection data to become available. As an aside, one could explore SARS-CoV-2 infection in terms of *APOE* allele distribution in Africans and how it affects susceptibility to Covid-19.

Aside from using the longitudinal data for epidemiological studies, GWAS per cognitive phenotype exploring rate of cognitive decline could be performed to find African-specific SNPs associated with cognitive decline. To add to this, AWI-Gen phase 2 has incorporated the total cognition score test battery into its data collection. This will expand the dataset to include East, West, and South Africans of varying ethno-linguistic backgrounds; thus, a larger African cohort will become available for GWAS at least for a total cognition score, where the results from Chapter 3 could then be used as a replication cohort. Unfortunately, due to time and cost constraints, the OCS-Plus could not be administered for AWI-Gen 2 despite its desirability. The HAALSI study, along with the research from this thesis, has shown the strength of using the OCS-Plus to assess cognitive function rather than a standard cognitive test in our African communities.

As this dataset includes extensive demographic, health, biomarker, anthropometric, and lifestyle data - given the strengths of both the underlying HDSS platform coupled with HAALSI and AWI-Gen cohorts - the opportunity to explore causal relationships between cognitive ability, educational attainment, and other correlated factors through Mendelian randomisation is possible. A number of factors; including smoking, sleep, physical activity, iron levels, and disease status, are correlated and associated with cognitive function. Polygenic risk scores (PRS) for each of the cognitive traits studied here can be developed, and then later PRS analysis for subsequent HAALSI datasets could be used to test predictive

value of the PRS created in the first instance. This may provide a better perspective of all the factors which converge to result in a person's cognitive abilities and performance.

Neuroimaging has become a very important tool in establishing the functional biological pathways of cognition. Magnetic resonance imaging is being performed on a small subset of the HAALSI cohort adding great value to African data and research output. Combining imaging data with the well-defined phenotypes and high-quality genomic data from this study may help to uncover the complexity of cognitive function.

The evolutionary trees for *APOE-TOMM40* haplotype analysis alluded to in Chapter 4 (Appendix 4) indicated that these haplotypes evolved independently and at multiple time points between Africans and Europeans. Phylogenetic research looking at the co-segregation of *TOMM40* and *APOE* may pinpoint the different evolutionary events in African versus European populations which may have led to the observed LD patterns.

With growing concern over the ageing global population, there has been a shift in the public health priorities and governments are allocating more budget to the research, intervention, and treatment of dementias. The US National Institutes of Health (NIH) National Institute for Ageing (NIA) has applied for $1.4 billion for Alzheimer's Disease and related dementia research from the US government. Private-public investment into Alzheimer's disease and related-therapeutics in Europe states that the Innovative Medicines Initiative (IMI) has already invested €5.3 billion into collaborative research projects and life science companies. These are only two of many initiatives funding growing research in the fields of neuroscience and cognitive decline. Showcasing the work done in this field in Africa is a good way to motivate for future funding for cognitive research and collaborative studies.

The assessment of cognitive function is essential in older Africans especially in communities where limited access to amenities may have severe impacts on mental and physical well-being. Ideally if the OCS-Plus could be incorporated into primary care services and systems in South Africa for all persons of retirement age, and those with chronic conditions it would enable a better understanding of the cognitive health of our vulnerable populations. This would allow a more accurate picture to help fulfil potential of our communities and each individual, and take a vital step toward active intervention to improve quality of life. Access

to quality education and the availability of skills development programmes may moderate cognitive decline in rural communities. The dissemination of information on lifestyle and other environmental factors which may negatively impact cognitive function may indirectly reduce the rates and severity of cognitive decline related to ageing in these populations.

References

1. Fitzgerald J, Morris DW, Donohoe G. Cognitive Genomics: Recent Advances and Current Challenges. Vol. 22, Current Psychiatry Reports. Springer; 2020.
2. Kobayashi LC, Mateen FJ, Montana L, Wagner RG, Kahn K, Tollman SM, et al. Cognitive Function and Impairment in Older, Rural South African Adults: Evidence from "Health and Aging in Africa: A Longitudinal Study of an INDEPTH Community in Rural South Africa". Neuroepidemiology. 2019;52(1–2):32–40.
3. Guerchet M, Mbelesso P, Ndamba-bandzouzi B, Pilleron S, Desormais I, Lacroix P, et al. Epidemiology of dementia in Central Africa (EPIDEMCA): protocol for a multicentre population-based study in rural and urban areas of the Central African Republic and the Republic of Congo. 2014;1–12.
4. Tollman SM, Norris SA, Berkman LF. Commentary: The value of life course epidemiology in low- and middle-income countries: an ageing perspective. Int J Epidemiol. 2016 Aug;45(4):997–9.
5. Kabudula CW, Tollman S, Mee P, Ngobeni S, Silaule B, Gómez-Olivé FX, et al. Two decades of mortality change in rural northeast South Africa. Glob Health Action. 2014 Oct 29;7:25596.
6. Kamboh MI, Fan KH, Yan Q, Beer JC, Snitz BE, Wang X, et al. Population-based genome-wide association study of cognitive decline in older adults free of dementia: identification of a novel locus for the attention domain. Neurobiol Aging. 2019 Dec 1;84:239.e15-239.e24.
7. Xu C, Zhang D, Wu Y, Tian X, Pang Z, Li S, et al. A genome-wide association study of cognitive function in Chinese adult twins. Biogerontology. 2017 Oct 1;18(5):811–9.
8. Zhang C, Pierce BL. Genetic susceptibility to accelerated cognitive decline in the US Health and Retirement Study. Neurobiol Aging. 2014 Jun;35(6):1512.e11-8.
9. Harris SE, Deary IJ. The genetics of cognitive ability and cognitive ageing in healthy older people. Trends Cogn Sci. 2011 Sep;15(9):388–94.
10. Stacey D, Ciobanu LG, Baune BT. A systematic review on the association between inflammatory genes and cognitive decline in non-demented elderly individuals. Eur Neuropsychopharmacol. 2015 Dec;
11. Achouri-Rassas A, Ali N Ben, Cherif A, Fray S, Siala H, Zakraoui NO, et al. Association between ACE polymorphism, cognitive phenotype and APOE E4 allele in a Tunisian population with Alzheimer disease. J Neural Transm. 2016;123(3):317–21.

12. Akinyemi RO, Owolabi MO, Oyeniyi T, Ovbiagele B, Arnett DK, Tiwari HK, et al. Neurogenomics in Africa: Perspectives, progress, possibilities and priorities. J Neurol Sci. 2016 Jul;366(2):213–23.
13. Xavier Gómez-Olivé F, Montana L, Wagner RG, Kabudula CW, Rohr JK, Kahn K, et al. Cohort profile: Health and ageing in Africa: A longitudinal study of an indepth community in South Africa (HAALSI). Int J Epidemiol. 2018;47(3):689-690J.
14. Kahn K, Collinson MA, Xavier Gómez-olivé F, Mokoena O, Twine R, Mee P, et al. Profile: Agincourt health and socio-demographic surveillance system. Int J Epidemiol. 2012;41(4):988–1001.
15. Asiimwe SB, Farrell M, Kobayashi LC, Manne-Goehler J, Kahn K, Tollman SM, et al. Cognitive differences associated with HIV serostatus and antiretroviral therapy use in a population-based sample of older adults in South Africa. Sci Rep. 2020;10(1):1–14.
16. Ramsay M, Crowther N, Tambo E, Agongo G, Baloyi V, Dikotope S, et al. H3Africa AWI-Gen Collaborative Centre: A resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. Glob Heal Epidemiol Genomics. 2016;1.
17. Ali SA, Soo C, Agongo G, Alberts M, Amenga-Etego L, Boua RP, et al. Genomic and environmental risk factors for cardiometabolic diseases in Africa: methods used for Phase 1 of the AWI-Gen population cross-sectional study. Glob Health Action. 2018;
18. Gómez-Olivé FX, Ali SA, Made F, Kyobutungi C, Nonterah E, Micklesfield L, et al. Regional and Sex Differences in the Prevalence and Awareness of Hypertension: An H3Africa AWI-Gen Study Across 6 Sites in Sub-Saharan Africa. Glob Heart. 2017;12(2):81–90.
19. Houle B, Gaziano T, Farrell M, Gómez-Olivé FX, Kobayashi LC, Crowther NJ, et al. Cognitive function and cardiometabolic disease risk factors in rural South Africa: Baseline evidence from the HAALSI study. BMC Public Health. 2019;19(1):1–11.
20. Gomez-Marin A, Mainen ZF. Expanding perspectives on cognition in humans, animals, and machines. Curr Opin Neurobiol. 2016 Apr;37:85–91.
21. Savage JE, Jansen PR, Stringer S, Watanabe K, Bryois J, De Leeuw CA, et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. Nat Genet. 2018 Jul 1;50(7):912–9.
22. Richardson K. GWAS and cognitive abilities: Why correlations are inevitable and meaningless . EMBO Rep. 2017 Aug;18(8):1279–83.

23. Christoforou A, Espeseth T, Davies G, Fernandes CPD, Giddaluru S, Mattheisen M, et al. GWAS-based pathway analysis differentiates between fluid and crystallized intelligence. Genes Brain Behav. 2014/08/08. 2014 Sep;13(7):663–74.
24. Davies G, Armstrong N, Bis JC, Bressler J, Chouraki V, Giddaluru S, et al. Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N=53949). Mol Psychiatry. 2015/02/03. 2015 Feb;20(2):183–92.
25. Humphreys GW, Duta MD, Montana L, Demeyere N, McCrory C, Rohr J, et al. Cognitive Function in Low-Income and Low-Literacy Settings: Validation of the Tablet-Based Oxford Cognitive Screen in the Health and Aging in Africa: A Longitudinal Study of an INDEPTH Community in South Africa (HAALSI). J Gerontol B Psychol Sci Soc Sci. 2017 Jan;72(1):38–50.
26. Goriounova NA, Mansvelder HD. Genes, cells and brain areas of intelligence. Vol. 13, Frontiers in Human Neuroscience. Frontiers Media S.A.; 2019.
27. Ryan JJ, Schnakenberg-Ott SD. Scoring reliability on the Wechsler Adult Intelligence Scale-Third Edition (WAIS-III). Assessment. 2003 Jun;10(2):151–9.
28. Kirkpatrick RM, McGue M, Iacono WG, Miller MB, Basu S. Results of a 'GWAS plus:' general cognitive ability is substantially heritable and massively polygenic. PLoS One. 2014 Nov 10;9(11):e112390–e112390.
29. Farrell MT, Kobayashi LC, Montana L, Wagner RG, Demeyere N, Berkman L. Disparity in educational attainment partially explains cognitive gender differences in Older Rural South Africans. Journals Gerontol - Ser B Psychol Sci Soc Sci. 2020;75(7):E161–73.
30. Siedlecki KL, Honig LS, Stern Y. Exploring the structure of a neuropsychological battery across healthy elders and those with questionable dementia and Alzheimer's disease. Neuropsychology. 2008 May;22(3):400–11.
31. Pereira L, Mutesa L, Tindana P, Ramsay M. African genetic diversity and adaptation inform a precision medicine agenda. Nat Rev Genet. 2021;
32. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome Variation Project shapes medical genetics in Africa. Nature. 2015;517(7534):327–32.

33. Choudhury A, Aron S, Botigué LR, Sengupta D, Botha G, Bensellak T, et al. High-depth African genomes inform human migration and health. Nature. 2020;586(7831):741–8.
34. Wright M, De Geus E, Ando J, Luciano M, Posthuma D, Ono Y, et al. Genetics of cognition: outline of a collaborative twin study. Twin Res. 2001 Feb;4(1):48–56.
35. Hasan A, Afzal M. Gene and environment interplay in cognition: Evidence from twin and molecular studies, future directions and suggestions for effective candidate gene x environment (cGxE) research. Vol. 33, Multiple Sclerosis and Related Disorders. Elsevier B.V.; 2019. p. 121–30.
36. Cuyvers E, Sleegers K. Genetic variations underlying Alzheimer's disease: evidence from genome-wide association studies and beyond. Lancet Neurol. 2016;15(8):857–68.
37. Ridge PG, Mukherjee S, Crane PK, Kauwe JSK, Consortium ADG. Alzheimer's Disease: Analyzing the Missing Heritability. PLoS One. 2013 Nov 7;8(11):e79771.
38. Marlow AJ, Fisher SE, Francks C, MacPhie IL, Cherny SS, Richardson AJ, et al. Use of multivariate linkage analysis for dissection of a complex cognitive trait. Am J Hum Genet. 2003/02/13. 2003 Mar;72(3):561–70.
39. Pericak-Vance MA, Bebout JL, Gaskell PCJ, Yamaoka LH, Hung WY, Alberts MJ, et al. Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. Am J Hum Genet. 1991 Jun;48(6):1034–50.
40. Dick DM, Aliev F, Bierut L, Goate A, Rice J, Hinrichs A, et al. Linkage analyses of IQ in the Collaborative Study on the Genetics of Alcoholism (COGA) sample. Behav Genet. 2006;36(1):77–86.
41. Blom ES, Holmans P, Arepalli S, Adighibe O, Hamshere ML, Gatz M, et al. Does APOE explain the linkage of Alzheimer's disease to chromosome 19q13? Am J Med Genet B Neuropsychiatr Genet. 2008 Sep 5;147B(6):778–83.
42. Pericak-Vance MA, Grubber J, Bailey LR, Hedges D, West S, Santoro L, et al. Identification of novel genes in late-onset Alzheimer's disease. Exp Gerontol. 2000 Dec;35(9–10):1343–52.
43. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2014. Nucleic Acids Res. 2014 Jan;42(Database issue):D749-55.

44. Savitz J, Solms M, Ramesar R. Apolipoprotein E variants and cognition in healthy individuals: A critical opinion. Brain Res Rev. 2006;51(1):125–35.
45. O'Donoghue MC, Murphy SE, Zamboni G, Nobre AC, Mackay CE. APOE genotype and cognition in healthy individuals at risk of Alzheimer's disease: A review. Cortex. 2018;104:103–23.
46. Rebeck GW. The role of APOE on lipid homeostasis and inflammation in normal brains. J Lipid Res. 2017;58(8):1493–9.
47. Hays CC, Zlatar ZZ, Meloy MJ, Bondi MW, Gilbert PE, Liu TT, et al. APOE modifies the interaction of entorhinal cerebral blood flow and cortical thickness on memory function in cognitively normal older adults. Neuroimage. 2019;202(February):116162.
48. Gureje O, Ogunniyi A, Baiyewu O, Price B, Unverzagt FW, Evans RM, et al. APOE ε4 is not associated with Alzheimer's disease in elderly Nigerians. Ann Neurol. 2006 Jan;59(1):182–5.
49. Osuntokun BO, Sahota A, Ogunniyi AO, Gureje O, Baiyewu O, Adeyinka A, et al. Lack of an association between apolipoprotein E ?4 and Alzheimer's disease in elderly nigerians. Ann Neurol. 1995 Sep 24;38(3):463–5.
50. Hendrie HC, Murrell J, Baiyewu O, Lane KA, Ogunniyi A, Unverzagt FW, et al. decline in African Americans and Yoruba. 2015;26(6):977–85.
51. Lutz MW, Crenshaw DG, Saunders AM, Roses AD. Genetic variation at a single locus and age of onset for Alzheimer's disease. Alzheimers Dement. 2010 Mar;6(2):125–31.
52. Zeitlow K, Charlambous L, Ng I, Gagrani S, Mihovilovic M, Luo S, et al. The biological foundation of the genetic association of TOMM40 with late-onset Alzheimer's disease. Biochim Biophys Acta - Mol Basis Dis. 2017 Nov;1863(11):2973–86.
53. Chu SH, Roeder K, Ferrell RE, Devlin B, DeMichele-Sweet MAA, Kamboh MI, et al. TOMM40 poly-T repeat lengths, age of onset and psychosis risk in Alzheimer disease. Neurobiol Aging. 2011;32(12):2328.e1-2328.e9.
54. Roses AD, Lutz MW, Amrine-Madsen H, Saunders AM, Crenshaw DG, Sundseth SS, et al. A TOMM40 variable-length polymorphism predicts the age of late-onset Alzheimer's disease. Pharmacogenomics J. 2010;10(5):375–84.

55. Johnson SC, La Rue A, Hermann BP, Xu G, Koscik RL, Jonaitis EM, et al. The effect of TOMM40 poly-T length on gray matter volume and cognition in middle-aged persons with APOE ε3/ε3 genotype. Alzheimer's Dement. 2011 Jul;7(4):456–65.
56. Linnertz C, Saunders AM, Lutz MW, Crenshaw DM, Grossman I, Burns DK, et al. Characterization of the poly-t variant in the tomm40 gene in diverse populations. PLoS One. 2012 Feb 16;7(2).
57. Hayden KM, McEvoy JM, Linnertz C, Attix D, Kuchibhatla M, Saunders AM, et al. A homopolymer polymorphism in the TOMM40 gene contributes to cognitive performance in aging. Alzheimer's Dement. 2012;8(5):381–8.
58. Linnertz C, Anderson L, Gottschalk W, Crenshaw D, Lutz MW, Allen J, et al. The cis-regulatory effect of an Alzheimer's disease-associated poly-T locus on expression of TOMM40 and apolipoprotein e genes. Alzheimer's Dement. 2014;10(5):541–51.
59. Yu L, Lutz MW, Wilson RS, Burns DK, Roses AD, Saunders AM, et al. APOE e4-TOMM40 '523 haplotypes and the risk of Alzheimer's disease in older Caucasian and African Americans. PLoS One. 2017;12(7):1–14.
60. Greenbaum L, Springer RR, Lutz MW, Heymann A, Lubitz I, Cooper I, et al. The TOMM40 poly-T rs10524523 variant is associated with cognitive performance among non-demented elderly with type 2 diabetes. Eur Neuropsychopharmacol. 2014;24(9):1492–9.
61. Yu L, Lutz MW, Wilson RS, Burns DK, Roses AD, Saunders AM, et al. TOMM40 ′523 variant and cognitive decline in older persons with APOE ϵ3/3 genotype. Neurology. 2017;88(7):661–8.
62. Roses AD. Genetic and clinical implications of a variable poly-T repeat within the TOMM40 gene for late-onset Alzheimer's disease. Arch Neurol. 2011;67(5):536–41.
63. Jie Huang MMP, Yuan Lin MP. Genotyping Technologies and Applications in the Era of Precision Medicine. North Am J Med Sci. 2017;10(4):176–80.
64. Davies G, Lam M, Harris SE, Trampush JW, Luciano M, Hill WD, et al. Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. Nat Commun. 2018 Dec 1;9(1).
65. Trampush JW, Yang MLZZ, Yu J, Knowles E, Davies G, Liewald DC, et al. GWAS meta-analysis reveals novel loci and genetic correlates for general cognitive function: a report from the COGENT consortium. Mol Psychiatry. 2017 Mar 1;22(3):336–45.

66. Sniekers S, Stringer S, Watanabe K, Jansen PR, Coleman JRI, Krapohl E, et al. Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. Nat Genet. 2017 Jul;49(7):1107–12.
67. Singh PP, Singh M, Mastana SS. APOE distribution in world populations with new data from India and the UK. Ann Hum Biol. 2006 Jan 9;33(3):279–308.
68. Radmanesh F, Devan WJ, Anderson CD, Rosand J, Falcone GJ. Accuracy of imputation to infer unobserved APOE epsilon alleles in genome-wide genotyping data. Eur J Hum Genet. 2014;22(10):1239–42.
69. Baichoo S, Souilmi Y, Panji S, Botha G, Meintjes A, Hazelhurst S, et al. Developing reproducible bioinformatics analysis workflows for heterogeneous computing environments to support African genomics. BMC Bioinformatics. 2018;19(1):457.
70. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44(7):821–4.
71. Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. Am J Epidemiol. 2002 Mar;155(5):478–84.
72. Hendrie HC, Hall KS, Hui S, Unverzagt FW, Yu CE, Lahiri DK, et al. Apolipoprotein E genotypes and Alzheimer's disease in a community study of elderly African Americans. Ann Neurol. 1995;37(1):118–20.
73. Burt TD, Agan BK, Marconi VC, He W, Kulkarni H, Mold JE, et al. Apolipoprotein (apo) E4 enhances HIV-1 cell entry in vitro, and the APOE ε4/ε4 genotype accelerates HIV disease progression. Proc Natl Acad Sci U S A. 2008;105(25):8718–23.
74. Chang L, Andres M, Sadino J, Jiang CS, Nakama H, Miller E, et al. Impact of apolipoprotein E ε4 and HIV on cognition and brain atrophy: Antagonistic pleiotropy and premature brain aging. Neuroimage. 2011 Oct;58(4):1017–27.
75. Panos SE, Hinkin CH, Singer EJ, Thames AD, Patel SM, Sinsheimer JS, et al. Apolipoprotein-E genotype and human immunodeficiency virus-associated neurocognitive disorder: The modulating effects of older age and disease severity. Neurobehav HIV Med. 2013;5(1):11–22.
76. Kotze M, Brand T, Pretorius J, J van Rensburg S, Luckhoff HK. Apolipoprotein E ε-4 as a genetic determinant of Alzheimer’s disease heterogeneity. Degener Neurol Neuromuscul Dis. 2015;9.

77. Kuo CL, Pilling LC, Atkins JL, Masoli JAH, Delgado J, Kuchel GA, et al. ApoE e4e4 Genotype and Mortality with COVID-19 in UK Biobank. Journals Gerontol - Ser A Biol Sci Med Sci. 2020;75(9):1801–3.
78. Small BJ, Rosnick CB, Fratiglioni L, Bäckman L. Apolipoprotein E and cognitive performance: A meta-analysis. Psychol Aging. 2004;19(4):592–600.
79. Wisdom NM, Callahan JL, Hawkins KA. The effects of apolipoprotein E on non-impaired cognitive functioning: A meta-analysis. Neurobiol Aging. 2011;32(1):63–74.
80. Reas ET, Laughlin GA, Bergstrom J, Kritz-Silverstein D, Barrett-Connor E, McEvoy LK. Effects of APOE on cognitive aging in community-dwelling older adults. Neuropsychology. 2019 Mar;33(3):406–16.
81. Liu C-C, Kanekiyo T, Xu H, Bu G. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. Nat Rev Neurol. 2013 Feb 8;9(2):106–18.
82. Wendelken LA, Jahanshad N, Rosen HJ, Busovaca E, Allen I, Coppola G, et al. ApoE ϵ4 is associated with cognition, brain integrity, and atrophy in HIV Over Age 60. J Acquir Immune Defic Syndr. 2016;73(4):426–32.
83. Chang L, Jiang C, Cunningham E, Buchthal S, Douet V, Andres M, et al. Chang et al. 2014 - Effects of Apoe4, age, and HIV on glial metabolits and cog deficits. 2015;
84. Lipnicki DM, Makkar SR, Crawford JD, Thalamuthu A, Kochan NA, Lima-Costa MF, et al. Determinants of cognitive performance and decline in 20 diverse ethno-regional groups: A COSMIC collaboration cohort study. PLoS Med. 2019;16(7):1–27.
85. Espinosa A, Hernández-Olasagarre B, Moreno-Grau S, Kleineidam L, Heilmann-Heimbach S, Hernández I, et al. Exploring Genetic Associations of Alzheimer's Disease Loci With Mild Cognitive Impairment Neurocognitive Endophenotypes. Front Aging Neurosci. 2018 Oct 30;10.
86. Corbo RM, Scacchp R. Apolipoprotein E (APOE) allele distribution in the world. Is APOE*4 a 'thrifty' allele? Ann Hum Genet. 1999;63(4):301–10.
87. Demarchi DA, Salzano FM, Altuna ME, Fiegenbaum M, Hill K, Hurtado AM, et al. APOE polymorphism distribution among Native Americans and related populations. Ann Hum Biol. 2005;32(3):351–65.
88. Sandholzer C, Delport R, Vermaak H, Utermann G. High frequency of the apo ε4 allele in Khoi San from South Africa. Hum Genet. 1995;95(1):46–8.

89. Willis F, Graff-Radford N, Pinto M, Lawson LS, Adamson J, Epstein D, et al. Apolipoprotein ∈4 allele frequency in young Africans of Ugandan descent versus African Americans. J Natl Med Assoc. 2003;95(1):71–6.
90. Chen C-H, Mizuno T, Elston R, Kariuki MM, Hall K, Unverzagt F, et al. A comparative study to screen dementia and APOE genotypes in an ageing East African population. Neurobiol Aging. 2010 May;31(5):732–40.
91. Morgan EE, Woods SP, Letendre SL, Franklin DR, Bloss C, Goate A, et al. Apolipoprotein E4 genotype does not increase risk of HIV-associated neurocognitive disorders. J Neurovirol. 2013;19(2):150–6.
92. Rosenberg M, Gómez-Olivé FX, Wagner RG, Rohr J, Payne CF, Berkman L, et al. The relationships between cognitive function, literacy and HIV status knowledge among older adults in rural South Africa. J Int AIDS Soc. 2020;23(3):1–9.
93. Fujioka H, Phelix CF, Friedland RP, Zhu X, Perry EA, Castellani RJ, et al. Apolipoprotein E4 Prevents Growth of Malaria at the Intraerythrocyte Stage: Implications For Diff erences in Racial Susceptibility to Alzheimer's Disease. J Health Care Poor Underserved. 2014;24(4A):70–8.
94. Van Exel E, Koopman JJE, Van Bodegom D, Meij JJ, De Knijff P, Ziem JB, et al. Effect of APOE ε4 allele on survival and fertility in an adverse environment. PLoS One. 2017;12(7):1–13.
95. Smith CJ, Ashford JW, Perfetti TA. Putative Survival Advantages in Young Apolipoprotein ε4 Carriers are Associated with Increased Neural Stress. J Alzheimer's Dis. 2019;68:885–923.
96. Kuhlmann I, Minihane AM, Huebbe P, Nebel A, Rimbach G. Apolipoprotein E genotype and hepatitis C, HIV and herpes simplex disease risk: a literature review. Lipids Health Dis. 2010 Jan 28;9:8.
97. Mueller T, Fischer J, Gessner R, Rosendahl J, Böhm S, van Bömmel F, et al. Apolipoprotein E allele frequencies in chronic and self-limited hepatitis C suggest a protective effect of APOE4 in the course of hepatitis C virus infection. Liver Int. 2016;
98. Joska JA, Combrinck M, Valcour VG, Hoare J, Leisegang F, Mahne AC, et al. Association between apolipoprotein E4 genotype and human immunodeficiency virus-associated dementia in younger adults starting antiretroviral therapy in South Africa. J Neurovirol. 2010;16(5):377–83.

99. Gouveia MH, Cesar CC, Santolalla ML, Anna HPS, Scliar MO, Leal TP, et al. Genetics of cognitive trajectory in Brazilians: 15 years of follow-up from the Bambuí-Epigen Cohort Study of Aging. Sci Rep. 2019 Dec 1;9(1).
100. Wu Y, Zhang D, Pang Z, Oksuzyan A, Jiang W, Wang S, et al. Gender-specific patterns in age-related decline in general health among Danish and Chinese: a cross-national comparative study. Geriatr Gerontol Int. 2012 Jul;12(3):431–9.
101. Hayat SA, Luben R, Moore S, Dalzell N, Bhaniani A, Anuj S, et al. Cognitive function in a general population of men and women: a cross sectional study in the European Investigation of Cancer-Norfolk cohort (EPIC-Norfolk). BMC Geriatr. 2014 Dec 19;14:142.
102. Sundermann EE, Maki PM, Rubin LH, Lipton RB, Landau S, Biegon A, et al. Female advantage in verbal memory: Evidence of sex-specific cognitive reserve. Neurology. 2016/10/05. 2016 Nov 1;87(18):1916–24.
103. Jack Jr CR, Wiste HJ, Weigand SD, Knopman DS, Vemuri P, Mielke MM, et al. Age, Sex, and APOE ε4 Effects on Memory, Brain Structure, and β-Amyloid Across the Adult Life Span. JAMA Neurol. 2015 May;72(5):511–9.
104. Andres MA. HHS Public Access. J Neuroimmune Pharmacol. 2011;6(3):389–98.
105. Geffin R, McCarthy M. Aging and Apolipoprotein E in HIV Infection. J Neurovirol. 2018;24(5):529–48.
106. Hulgan T, Kallianpur AR, Guo Y, Barnholtz-Sloan JS, Gittleman H, Brown TT, et al. Peripheral Blood Mitochondrial DNA Copy Number Obtained From Genome-Wide Genotype Data Is Associated With Neurocognitive Impairment in Persons With Chronic HIV Infection. 2018.
107. Valcour V, Shikuma C, Shiramizu B, Watters M, Poff P, Selnes OA, et al. Age, apolipoprotein E4, and the risk of HIV dementia: The Hawaii Aging with HIV Cohort. J Neuroimmunol. 2004;157(1-2 SPEC. ISS.):197–202.
108. Becker JT, Martinson JJ, Penugonda S, Kingsley L, Molsberry S, Reynolds S, et al. No association between Apoε4 alleles, HIV infection, age, neuropsychological outcome, or death. J Neurovirol. 2015 Feb 12;21(1):24–31.
109. Cysique LA, Hewitt T, Croitoru-Lamoury J, Taddei K, Martins RN, Chew CSN, et al. APOE ε4 moderates abnormal CSF-abeta-42 levels, while neurocognitive impairment is associated with abnormal CSF tau levels in HIV+ individuals - A cross-sectional observational study. BMC Neurol. 2015;15(1):4–11.

110. Lipnicki DM, Crawford JD, Dutta R, Thalamuthu A, Kochan NA, Andrews G, et al. Age-related cognitive decline and associations with sex, education and apolipoprotein E genotype across ethnocultural groups and geographic regions: a collaborative cohort study. PLoS Med. 2017;14(3):1–21.
111. Hill WD, Davies G, Group CCW, Liewald DC, McIntosh AM, Deary IJ. Age-Dependent Pleiotropy Between General Cognitive Function and Major Psychiatric Disorders. Biol Psychiatry. 2015/09/04. 2016 Aug 15;80(4):266–73.
112. Coleman JRI, Bryois J, Gaspar HA, Jansen PR, Savage JE, Skene N, et al. Biological annotation of genetic loci associated with intelligence in a meta-analysis of 87,740 individuals. Mol Psychiatry. 2019 Feb 1;24(2):182–97.
113. Davies G, Harris SE, Reynolds CA, Payton A, Knight HM, Liewald DC, et al. A genome-wide association study implicates the APOE locus in nonpathological cognitive ageing. Mol Psychiatry. 2012/12/04. 2014 Jan;19(1):76–87.
114. Ohi K, Sumiyoshi C, Fujino H, Yasuda Y, Yamamori H, Fujimoto M, et al. Genetic overlap between general cognitive function and schizophrenia: A review of cognitive GWASs. Vol. 19, International Journal of Molecular Sciences. MDPI AG; 2018.
115. Harvey PD, Sun N, Bigdeli TB, Fanous AH, Aslan M, Malhotra AK, et al. Genome-wide association study of cognitive performance in U.S. veterans with schizophrenia or bipolar disorder. Am J Med Genet Part B Neuropsychiatr Genet. 2019;
116. Ibrahim-Verbaas CA, Bressler J, Debette S, Schuur M, Smith A V, Bis JC, et al. GWAS for executive function and processing speed suggests involvement of the CADM2 gene. Mol Psychiatry. 2015/04/14. 2016 Feb;21(2):189–97.
117. Hansell NK, Halford GS, Andrews G, Shum DHK, Harris SE, Davies G, et al. Genetic basis of a cognitive complexity metric. PLoS One. 2015 Apr 10;10(4):e0123886–e0123886.
118. Reynolds CA, Finkel D. A meta-analysis of heritability of cognitive aging: minding the 'missing heritability' gap. Neuropsychol Rev. 2015/03/03. 2015 Mar;25(1):97–112.
119. Trampush JW, Yang MLZ, Yu J, Knowles E, Davies G, Liewald DC, et al. GWAS meta-analysis reveals novel loci and genetic correlates for general cognitive function: A report from the COGENT consortium. Mol Psychiatry. 2017 Mar 1;22(3):336–45.

120. Ersland KM, Christoforou A, Stansberg C, Espeseth T, Mattheisen M, Mattingsdal M, et al. Gene-based analysis of regionally enriched cortical genes in GWAS data sets of cognitive traits and psychiatric disorders. PLoS One. 2012/02/22. 2012;7(2):e31687–e31687.
121. Stephan Y, Sutin AR, Luchetti M, Caille P, Terracciano A. Polygenic Score for Alzheimer Disease and cognition: The mediating role of personality. J Psychiatr Res. 2018 Dec 1;107:110–3.
122. Lam M, Hill WD, Trampush JW, Yu J, Knowles E, Davies G, et al. Pleiotropic Meta-Analysis of Cognition, Education, and Schizophrenia Differentiates Roles of Early Neurodevelopmental and Adult Synaptic Pathways. Am J Hum Genet. 2019 Aug 1;105(2):334–50.
123. Trzaskowski M, Davis OSP, DeFries JC, Yang J, Visscher PM, Plomin R. DNA evidence for strong genome-wide pleiotropy of cognitive and learning abilities. Behav Genet. 2013/04/23. 2013 Jul;43(4):267–73.
124. Zhao B, Luo T, Li T, Li Y, Zhang J, Shan Y, et al. Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. Nat Genet. 2019/11/01. 2019 Nov;51(11):1637–44.
125. Smith JA, Kho M, Zhao W, Yu M, Mitchell C, Faul JD. Genetic effects and gene-by-education interactions on episodic memory performance and decline in an aging population. Soc Sci Med. 2018;
126. Raj T, Chibnik LB, McCabe C, Wong A, Replogle JM, Yu L, et al. Genetic architecture of age-related cognitive decline in African Americans. Neurol Genet. 2017 Feb;3(1):e125.
127. Yen K, Wan J, Mehta HH, Miller B, Christensen A, Levine ME, et al. Humanin Prevents Age-Related Cognitive Decline in Mice and is Associated with Improved Cognitive Age in Humans. Sci Rep. 2018 Dec 1;8(1).
128. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, et al. Modernizing reference genome assemblies. PLoS Biol. 2011 Jul;9(7):e1001091.
129. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016 Oct;48(10):1279–83.

130. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016/10/03. 2016 Nov;48(11):1443–8.
131. Durbin R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). Bioinformatics. 2014 May;30(9):1266–72.
132. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLOS Genet. 2006 Dec 22;2(12):e190.
133. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74.
134. Buchmann R, Hazelhurst S. The ' Genesis ' Manual. Johannesburg: University of the Witwatersrand; 2015. p. 0–20.
135. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat Commun. 2017;8(1):1826.
136. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010 Sep;26(18):2336–7.
137. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019 Jan;47(D1):D1005–12.
138. Kamat MA, Blackshaw JA, Young R, Surendran P, Burgess S, Danesh J, et al. PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. Bioinformatics. 2019 Nov;35(22):4851–3.
139. Staley JR, Blackshaw J, Kamat MA, Ellis S, Surendran P, Sun BB, et al. PhenoScanner: a database of human genotype-phenotype associations. Bioinformatics. 2016 Oct;32(20):3207–9.
140. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. Nucleic Acids Res. 2020 Jan 8;48(D1):D682–8.
141. Liu J, Zhou Y, Liu S, Song X, Yang X-Z, Fan Y, et al. The coexistence of copy number variations (CNVs) and single nucleotide polymorphisms (SNPs) at a locus can result in distorted calculations of the significance in associating SNPs to disease. Hum Genet. 2018 Jul;137(6–7):553–67.

142. Wang H, Yang J, Schneider JA, De Jager PL, Bennett DA, Zhang H-Y. Genome-wide interaction analysis of pathological hallmarks in Alzheimer's disease. Neurobiol Aging. 2020 Sep;93:61–8.
143. Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat Genet. 2018 Jul;50(8):1112–21.
144. de la Fuente J, Davies G, Grotzinger AD, Tucker-Drob EM, Deary IJ. A general dimension of genetic sharing across diverse cognitive traits inferred from molecular data. Nat Hum Behav. 2021 Jan;5(1):49–58.
145. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014 Jul;511(7510):421–7.
146. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018 Oct;562(7726):203–9.
147. Goes FS, McGrath J, Avramopoulos D, Wolyniec P, Pirooznia M, Ruczinski I, et al. Genome-wide association study of schizophrenia in Ashkenazi Jews. Am J Med Genet Part B, Neuropsychiatr Genet Off Publ Int Soc Psychiatr Genet. 2015 Dec;168(8):649–59.
148. St Pourcain B, Skuse DH, Mandy WP, Wang K, Hakonarson H, Timpson NJ, et al. Variability in the common genetic architecture of social-communication spectrum phenotypes during childhood and adolescence. Mol Autism. 2014 Feb;5(1):18.
149. Rietveld CA, Medland SE, Derringer J, Yang J, Esko T, Martin NW, et al. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. Science. 2013 Jun;340(6139):1467–71.
150. Peyrot WJ, Price AL. Identifying loci with different allele frequencies among cases of eight psychiatric disorders using CC-GWAS. Nat Genet. 2021 Apr;53(4):445–54.
151. Wu Y, Cao H, Baranova A, Huang H, Li S, Cai L, et al. Multi-trait analysis for genome-wide association study of five psychiatric disorders. Transl Psychiatry. 2020 Jun;10(1):209.
152. Demontis D, Walters RK, Martin J, Mattheisen M, Als TD, Agerbo E, et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. Nat Genet. 2019 Jan;51(1):63–75.

153. Pisanu C, Williams MJ, Ciuculete DM, Olivo G, Del Zompo M, Squassina A, et al. Evidence that genes involved in hedgehog signaling are associated with both bipolar disorder and high BMI. Transl Psychiatry. 2019 Nov;9(1):315.
154. Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, et al. Identification of common genetic risk variants for autism spectrum disorder. Nat Genet. 2019/02/25. 2019 Mar;51(3):431–44.
155. Jones SE, Lane JM, Wood AR, van Hees VT, Tyrrell J, Beaumont RN, et al. Genome-wide association analyses of chronotype in 697,828 individuals provides insights into circadian rhythms. Nat Commun. 2019 Jan;10(1):343.
156. Jansen PR, Watanabe K, Stringer S, Skene N, Bryois J, Hammerschlag AR, et al. Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. Nat Genet. 2019 Mar;51(3):394–403.
157. Justice AE, Winkler TW, Feitosa MF, Graff M, Fisher VA, Young K, et al. Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits. Nat Commun. 2017 Apr;8:14977.
158. Soler Artigas M, Sánchez-Mora C, Rovira P, Richarte V, Garcia-Martínez I, Pagerols M, et al. Attention-deficit/hyperactivity disorder and lifetime cannabis use: genetic overlap and causality. Mol Psychiatry. 2020 Oct;25(10):2493–503.
159. Thorleifsson G, Walters GB, Gudbjartsson DF, Steinthorsdottir V, Sulem P, Helgadottir A, et al. Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. Nat Genet. 2009 Jan;41(1):18–24.
160. Zhu Z, Guo Y, Shi H, Liu C-L, Panganiban RA, Chung W, et al. Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK Biobank. J Allergy Clin Immunol. 2020 Feb;145(2):537–49.
161. Deters KD, Nho K, Risacher SL, Kim S, Ramanan VK, Crane PK, et al. Genome-wide association study of language performance in Alzheimer's disease. Brain Lang. 2017/05/31. 2017 Sep;172:22–9.
162. Nudel R, Simpson NH, Baird G, O'Hare A, Conti-Ramsden G, Bolton PF, et al. Genome-wide association analyses of child genotype effects and parent-of-origin effects in specific language impairment. Genes Brain Behav. 2014/03/24. 2014 Apr;13(4):418–29.

163. Nudel R, Christiani CAJ, Ohland J, Uddin MJ, Hemager N, Ellersgaard D, et al. Quantitative genome-wide association analyses of receptive language in the Danish High Risk and Resilience Study. BMC Neurosci. 2020 Jul 7;21(1):30.
164. Gong P, Liu J, Blue PR, Li S, Zhou X. Serotonin receptor gene (HTR2A) T102C polymorphism modulates individuals' perspective taking ability and autistic-like traits . Vol. 9, Frontiers in Human Neuroscience . 2015. p. 575.
165. Norton N, Owen MJ. HTR2A: association and expression studies in neuropsychiatric genetics. Ann Med. 2005;37(2):121–9.
166. Verweij KJH, Vinkhuyzen AAE, Benyamin B, Lynskey MT, Quaye L, Agrawal A, et al. The genetic aetiology of cannabis use initiation: a meta-analysis of genome-wide association studies and a SNP-based heritability estimation. Addict Biol. 2012/07/24. 2013 Sep;18(5):846–50.
167. Lam M, Trampush JW, Yu J, Knowles E, Djurovic S, Melle I, et al. Multi-Trait analysis of gwas and biological insights into cognition: A response to hill (2018). Twin Res Hum Genet. 2018 Oct 1;21(5):394–7.
168. Traylor M, Tozer DJ, Croall ID, Lisiecka-Ford DM, Olorunda AO, Boncoraglio G, et al. Genetic variation in PLEKHG1 is associated with white matter hyperintensities (n = 11,226). Neurology. 2019 Feb;92(8):e749–57.
169. Armstrong NJ, Mather KA, Sargurupremraj M, Knol MJ, Malik R, Satizabal CL, et al. Common Genetic Variation Indicates Separate Causes for Periventricular and Deep White Matter Hyperintensities. Stroke. 2020 Jul;51(7):2111–21.
170. Bhatnagar P, Barron-Casella E, Bean CJ, Milton JN, Baldwin CT, Steinberg MH, et al. Genome-wide meta-analysis of systolic blood pressure in children with sickle cell disease. PLoS One. 2013;8(9):e74193.
171. Day FR, Ong KK, Perry JRB. Elucidating the genetic basis of social interaction and isolation. Nat Commun. 2018 Jul;9(1):2457.
172. Chiba-Falek O, Gottschalk WK, Lutz MW. The effects of the TOMM40 poly-T alleles on Alzheimer's disease phenotypes. Alzheimer's Dement. 2018 May;14(5):692–8.
173. Nishimura A, Nonomura H, Tanaka S, Yoshida M, Maruyama Y, Aritomi Y, et al. Characterization of APOE and TOMM40 allele frequencies in the Japanese population. Alzheimer's Dement Transl Res Clin Interv. 2017;3(4):524–30.

174. Roses AD, Lutz MW, Saunders AM, Goldgaber D, Saul R, Sundseth SS, et al. African-American TOMM40'523-APOE haplotypes are admixture of West African and Caucasian alleles. Alzheimer's Dement. 2014;10(6):592-601.e2.
175. Yu L, Lutz MW, Farfel JM, Wilson RS, Burns DK, Saunders AM, et al. Neuropathologic features of TOMM40 '523 variant on late-life cognitive decline. Alzheimer's Dement. 2017 Dec;13(12):1380–8.
176. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. Nat Methods. 2017;14(6):590–2.
177. Ivády G, Madar L, Dzsudzsák E, Koczok K, Kappelmayer J, Krulisova V, et al. Analytical parameters and validation of homopolymer detection in a pyrosequencing-based next generation sequencing system. BMC Genomics. 2018;19(1):1–8.
178. Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, et al. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res. 2019;47(21):10994–1006.
179. Caselli RJ, Dueck AC, Huentelman MJ, Lutz MW, Saunders AM, Reiman EM, et al. Longitudinal modeling of cognitive aging and the TOMM40 effect. Alzheimer's Dement. 2012 Nov;8(6):490–5.
180. Li G, Bekris LM, Leong L, Steinbart EJ, Shofer JB, Crane PK, et al. TOMM40 intron 6 poly-T length, age at onset, and neuropathology of AD in individuals with APOE ε3/ε3. Alzheimer's Dement. 2013 Sep;9(5):554–61.
181. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet. 2018;103(3):338–48.
182. Delaneau O, Zagury J-F, Robinson MR, Marchini JL, Dermitzakis ET. Accurate, scalable and integrative haplotype estimation. Nat Commun. 2019;10(1):5436.
183. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009 Jun;5(6):e1000529.
184. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 2012 Jul;44(8):955–9.

185. Shao Y, Shaw M, Todd K, Khrestian M, D'Aleo G, Barnard PJ, et al. DNA methylation of TOMM40-APOE-APOC2 in Alzheimer's disease. J Hum Genet. 2018 Apr 25;63(4):459–71.
186. Deters KD, Mormino EC, Yu L, Lutz MW, Bennett DA, Barnes LL. TOMM40-APOE haplotypes are associated with cognitive decline in non-demented Blacks. Alzheimers Dement. 2021 Feb;
187. Wolters FJ, Yang Q, Biggs ML, Jakobsdottir J, Li S, Evans DS, et al. The impact of APOE genotype on survival: Results of 38,537 participants from six population-based cohorts (E2-CHARGE). PLoS One. 2019 Jul 29;14(7):e0219668.
188. Haraksingh RR, Abyzov A, Urban AE. Comprehensive performance comparison of high-resolution array platforms for genome-wide Copy Number Variation (CNV) analysis in humans. BMC Genomics. 2017 Apr;18(1):321.
189. Rotimi C, Abayomi A, Abimiku A, Adabayeri VM, Adebamowo C, Adebiyi E, et al. Research capacity. Enabling the genomic revolution in Africa. Science. 2014 Jun;344(6190):1346–8.
190. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat Genet. 2018/01/01. 2018 Feb;50(2):229–37.
191. Madhavan S, Schatz E, Gómes-Olivé FX, Collinson M. Social Positioning of Older Persons in Rural South Africa: Change or Stability? J South Afr Stud. 2017/10/04. 2017;43(6):1293–307.
192. Schatz E, Ralston M, Madhavan S, Collinson MA, Gómez-Olivé FX. Living Arrangements, Disability and Gender of Older Adults Among Rural South Africa. Journals Gerontol Ser B. 2017;00(00):1–11.
193. Gómez-Olivé FX, Houle B, Rosenberg M, Kabudula C, Mojola S, Rohr JK, et al. Brief Report: HIV Incidence Among Older Adults in a Rural South African Setting: 2010–2015. JAIDS J Acquir Immune Defic Syndr. 2020 Sep 1;85(1):18–22.

Appendix 1

A1.1 Ethics approval certificate

UNIVERSITY OF THE WITWATERSRAND JOHANNESBURG

R14/49 Ms CC Soo et al

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
CLEARANCE CERTIFICATE NO. M170916

NAME: (Principal Investigator)	Ms CC Soo et al
DEPARTMENT:	School of Pathology Sydney Brenner Institute for Molecular Bioscience Medical School
PROJECT TITLE:	Genetic and environmental factors associated with cognition in an ageing South African population
DATE CONSIDERED:	29/09/2017
DECISION:	Approved unconditionally
CONDITIONS:	
SUPERVISOR:	Professor M Ramsay
APPROVED BY:	Professor PE Cleaton-Jones, Chairperson, HREC (Medical)
DATE OF APPROVAL:	15/12/2017

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary on 3rd floor, Phillip V Tobias Building, Parktown, University of the Witwatersrand, Johannesburg.
I/We fully understand the conditions under which I am/we are authorised to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated from the research protocol as approved, I/we undertake to resubmit to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in **September** and will therefore be due in the month of **September** each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

11/01/2018
Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

A1.2 Plagiarism declaration



PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I Cassandra Claire Soo (Student number: 0713996A) am a student registered for the degree of Doctor of Philosphy in Human Genetics in the academic year 2021.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.
- I have included as an appendix a report from "Turnitin" (or other approved plagiarism detection) software indicating the level of plagiarism in my research document.

Signature: _____ Date: 25 October 2021

1

Appendix 2

A2.1 *APOE* Sanger sequencing validation

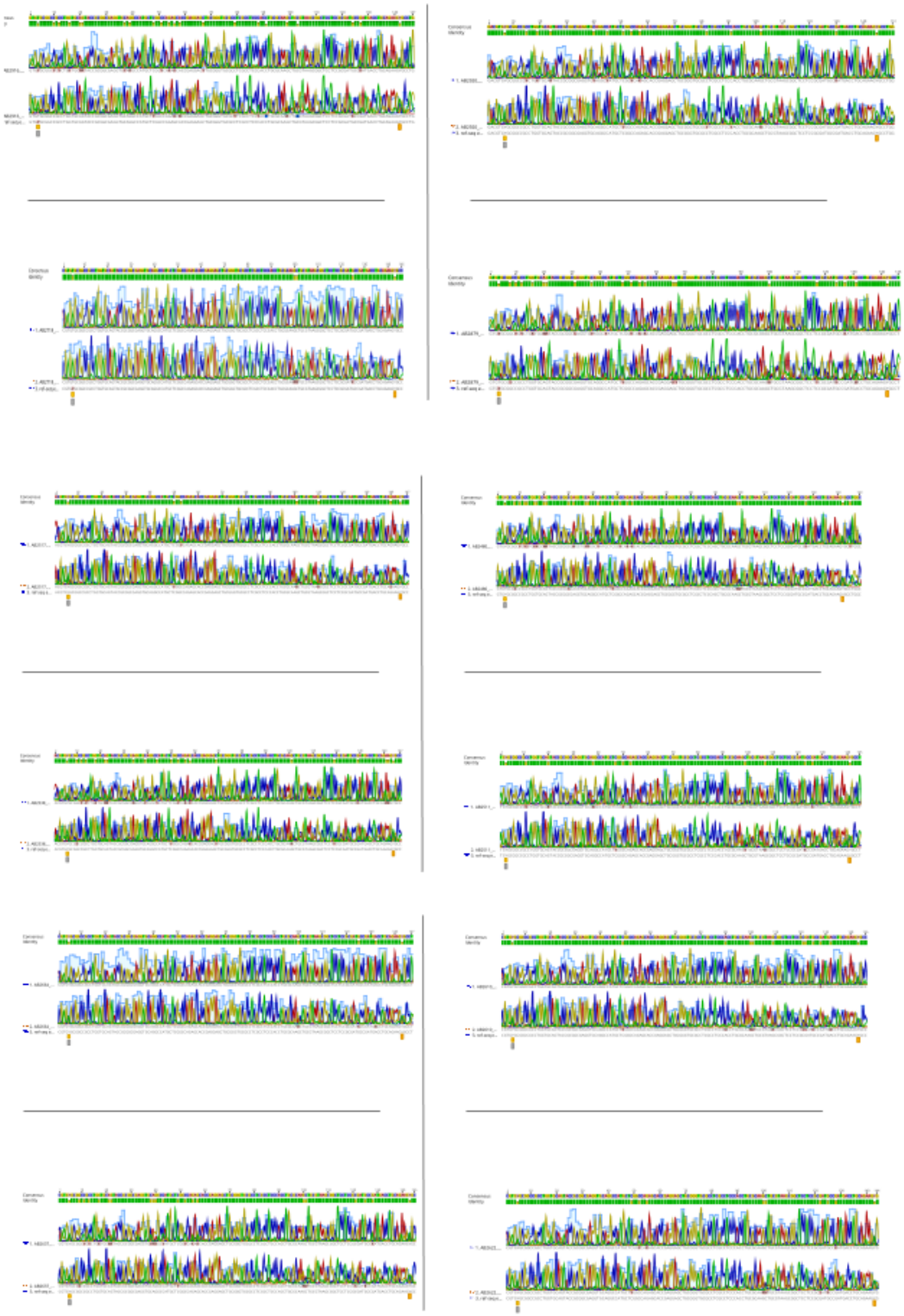
Thirty DNA samples were submitted to Inqaba biotec™ for Sanger sequencing as validation for TaqMan™ SNP genotyping. *APOE* primer sequences were:

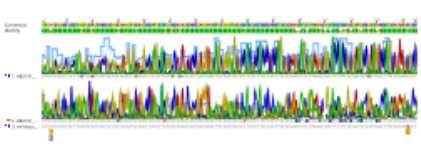
- Forward primer
    - CGGGCACGGCTGTCCAAG
- Reverse primer
    - TCGCGGATGGCGCTGAGG

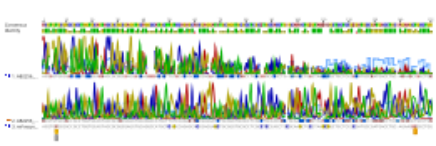
Table A.2.1: TaqMan genotyping results compared to sequencing validation

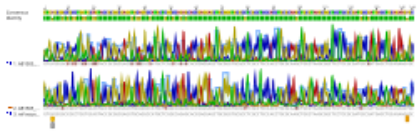
Sample ID	***Sequencing***	***TaqMan_rs429358***	***TaqMan_rs7412***
AB2256	validated	T/T	C/T
AB2191	validated	C/T	C/C
AB0744	validated	C/T	C/T
AB0521	validated	C/T	C/C
AB0352	validated	C/T	C/T
AB0375	validated	C/T	C/T
AB0399	validated	C/T	C/T
AB0423	validated	C/T	C/C
AB0430	validated	C/T	C/C
AB0187	validated	T/T	C/T
AB0198	validated	T/T	C/T
AB0210	validated	C/T	C/T
AB2610	validated	C/T	C/T
AB2637	validated	C/T	C/T
AB2622	validated	C/T	C/T
AB1505	validated	C/T	C/T
AB1492	validated	C/T	C/T
AB2664	validated	C/T	C/T
AB2718	validated	T/T	C/T
AB2739	validated	C/T	C/T
AB2480	validated	C/T	C/T
AB2517	validated	C/T	C/T
AB2511	validated	C/T	C/T

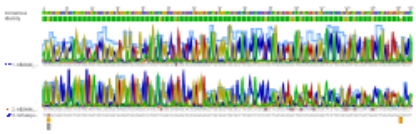
AB2538	validated	C/T	C/T
AB2916	validated	C/T	C/T
AB2920	validated	C/T	C/T
AB2444	validated	T/T	C/T

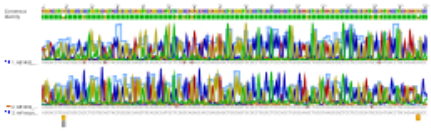


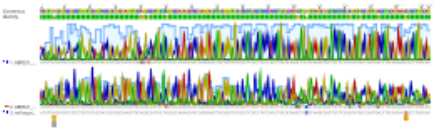


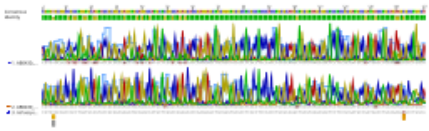


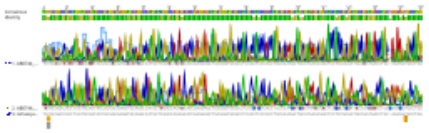


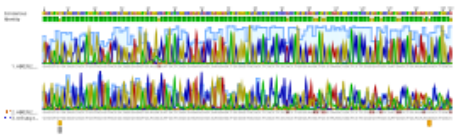


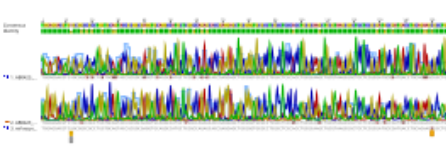


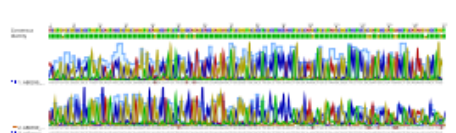


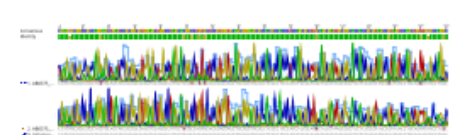












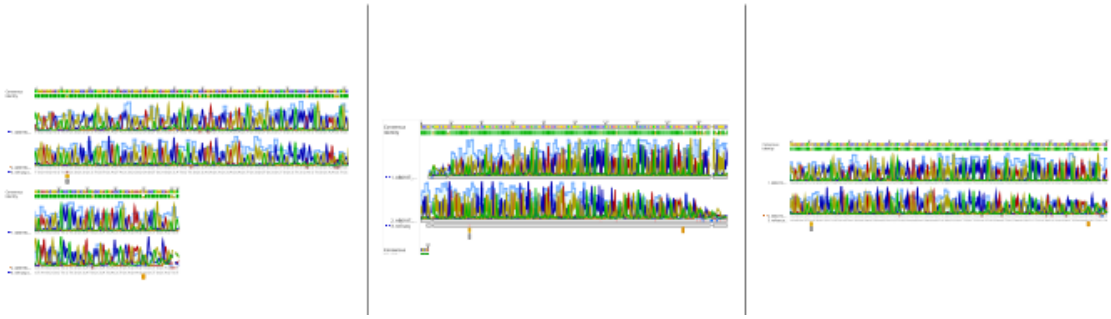


Figure A.2.1: Electrophoretograms for *APOE* for 27 samples that were successfully sequenced

A2.2 Diagnostics of Regression

Diagnostics of regression were performed for the first regression model (Chapter 2) where cognitive phenotype was dependent on *APOE* genotype and adjusted for the covariates age, sex, highest level of education, and HIV status. Diagnostics were performed using the car package in R (Fox J, Weisberg S (2019). *An R Companion to Applied Regression*, Third edition. Sage, Thousand Oaks, CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/) for each of the cognitive function phenotypes below. Tests assessing validity of the regression model included:

1. Assessment of outliers
2. QQ and leverage plots
3. Cook's D and influence plots for influential observations
4. Assessment of normality
5. Test for non-constant error variance
6. Evaluation of multi-collinearity
7. Non-linearity tests
8. The Durbin Watson test for autocorrelated errors

A2.2.1 Total cognition score

There were no outliers for total cognition score as none of the Studentized residuals had Bonferroni p-values < 0.05. The QQ-plot showed little inflation, and leverage plots indicate that age and level of education increase significance of the model as a whole. Cook's distance was <1 signifying no identifiable influencers in the dataset. Residuals were normally distributed, and variance was constant (χ^2=3.88, df=1, p=0.049). The spread level plot

suggests a transformation to the power of 1.29, where a power of 1 means no change. Variance inflation factors were below 5, and there was no autocorrelation of variables (Durbin-Watson p=0.714). And component residual plots indicate how each term influences total cognition score in the model.

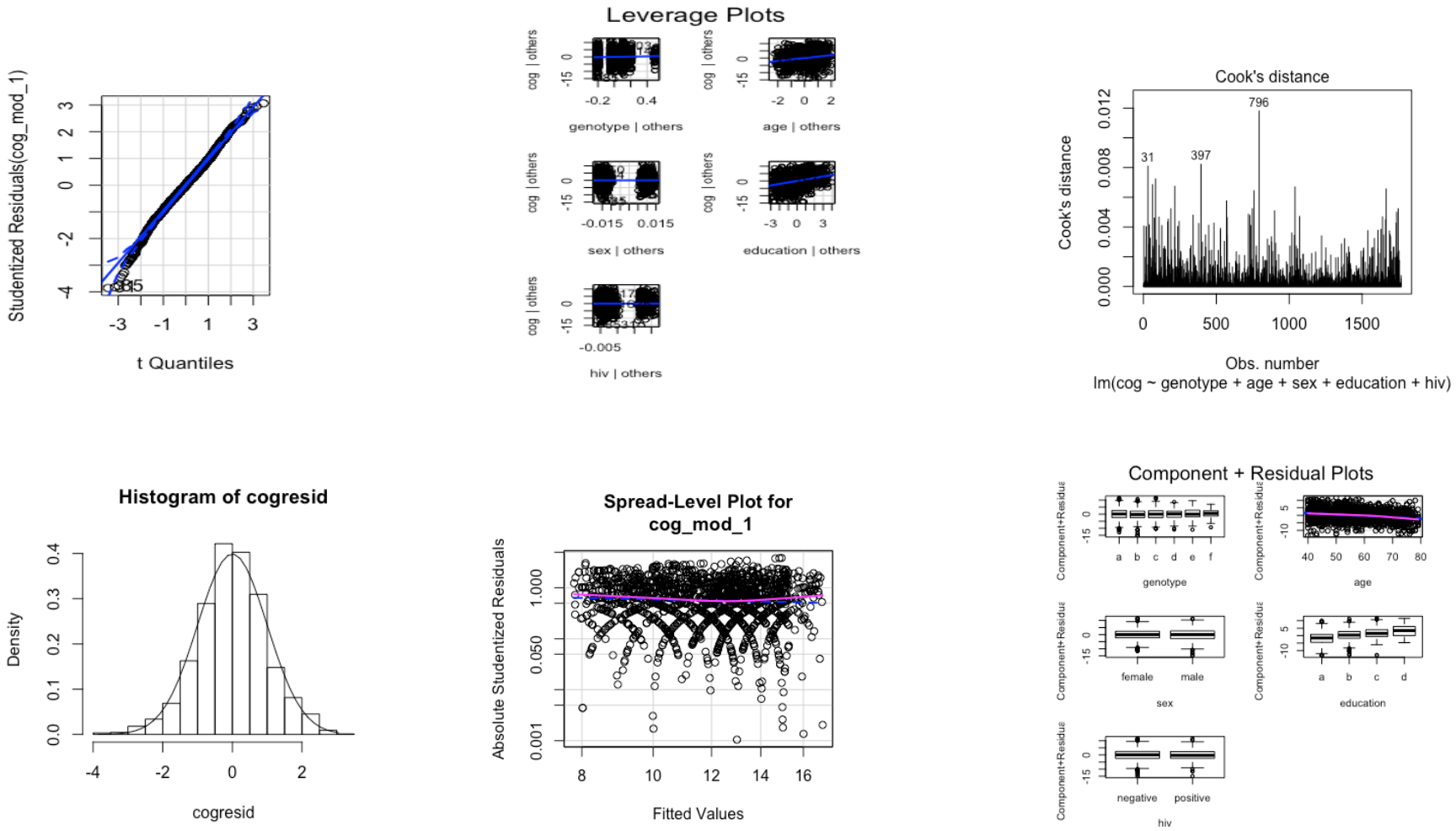


Figure A.2.2: Diagnostic regression plots for the residuals of total cognition score In order: QQ-plot, Leverage plots for each regression term, Cook's plot indicating Cook's distance, histogram showing distribution of residuals, spread level plot, and component-residual plots for each term and its effect on total cognition score

A2.2.2 Executive Function

No outliers were observed for the regression model for executive function and *APOE* genotype. The QQ-plot showed most values of residuals falling along the expected line. Leverage plots indicate strong correlation with age and education level. According to the plot showing Cook's distance, there were no influencing individuals. Residuals were normally distributed about 0 and variance was constant (χ^2=1.37, df=1, p=0.242). Spread level calculations suggested a transformation to the power of 1.03 indicating z-score suitability for linear modelling. Variance inflation levels were appropriate and there was no autocorrelation (p= 0.608).

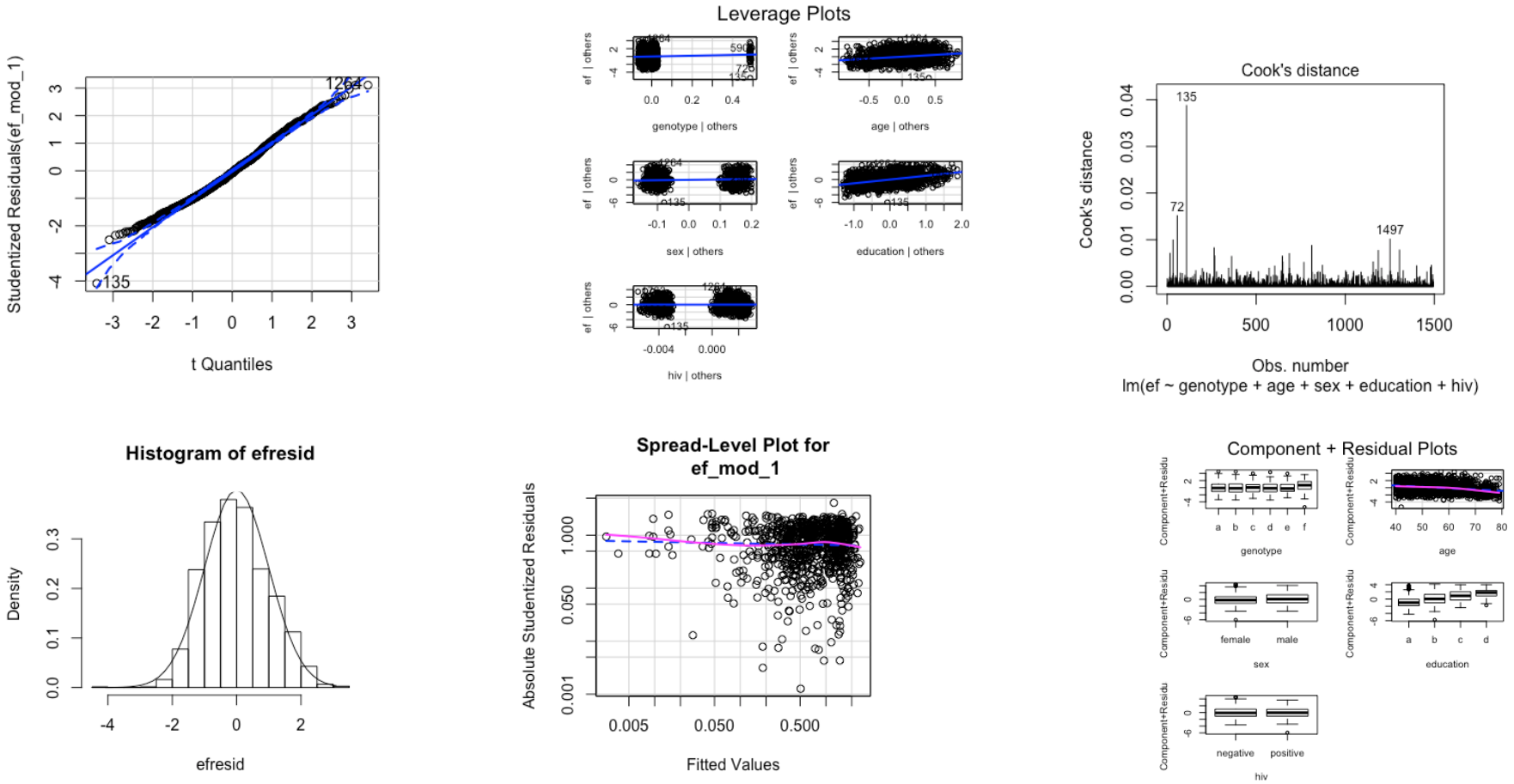


Figure A.2.3: Diagnostic regression plots for the residuals of executive function

In order: QQ-plot, Leverage plots for each regression term, Cook's plot indicating Cook's distance, histogram showing distribution of residuals, spread level plot where negative values were removed, and component-residual plots for each term and its effect on executive function

A2.2.3 Language

There was one outlying residual for language performance (Bonferroni p=1.932e-05) evidenced in the QQ-plot indicated 135. Leverage plots indicate that the education was the term with the most effect on language scores. Cook's distance was below 1 and the outlier (135) had the greatest Cook's distance. Histogram skewness may once again be attributed to the outlying residual value. Heteroskedasticity was observed (χ^2=79.13, df=1, p=< 2.22e-16) which means that the residual variance is not constant across all observations. Transformation was not required (suggested power=1.17). There was little variance inflation and no concern for autocorrelation (p=0.08 for Durbin-Watson test).

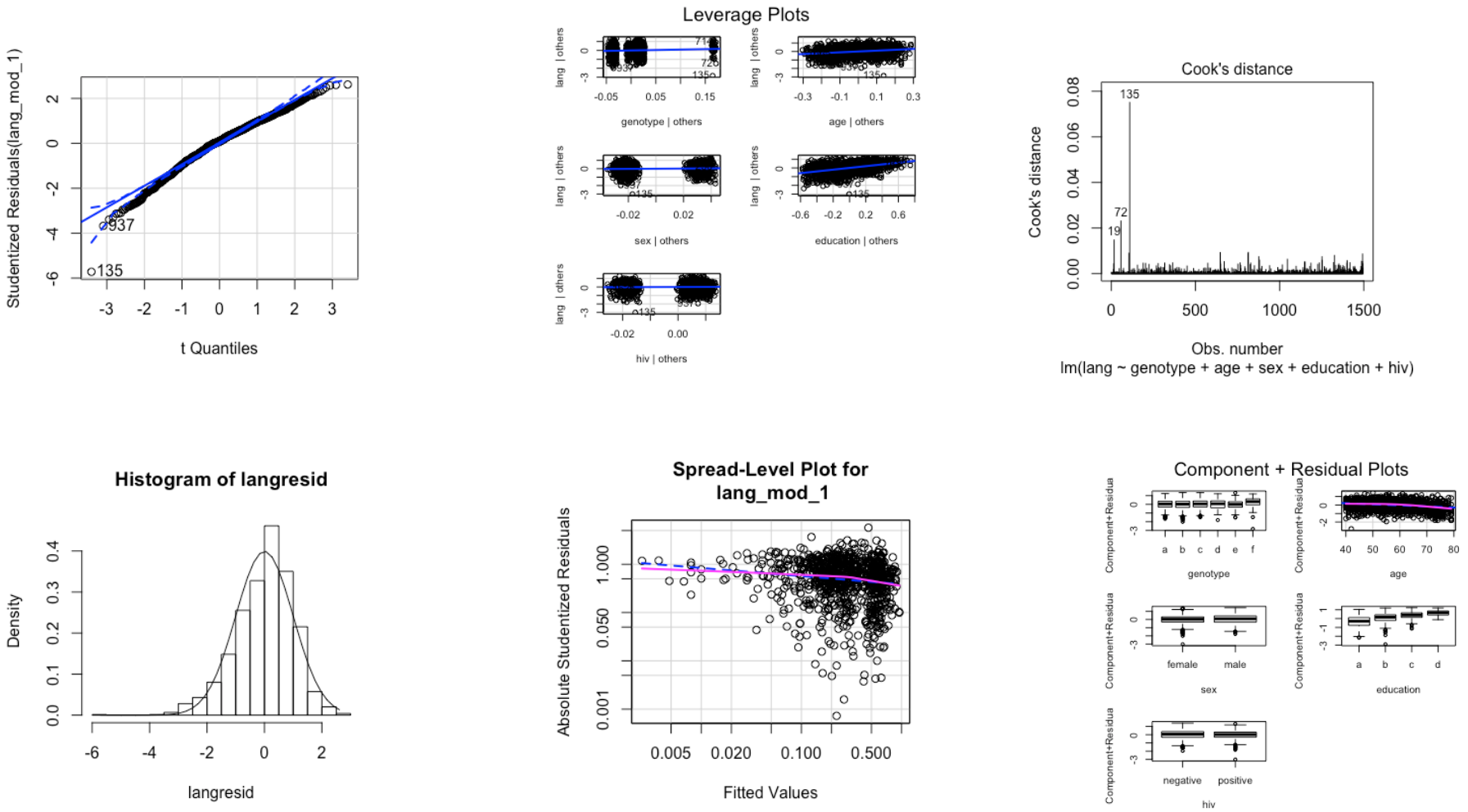


Figure A.2.4: Diagnostic regression plots for the residuals of language

In order: QQ-plot, Leverage plots for each regression term, Cook's plot indicating Cook's distance, histogram showing distribution of residuals, spread level plot where negative values were removed, and component-residual plots for each term and its effect on language.

A2.2.4 Verbal Episodic Memory

There were two outliers for memory residuals individual 135 (p=0.0002) and 1647 (p=0.048). The outlying values caused divergence from the normal in the QQ-plot. Leverage plots show main effects on model as age and education. Highest Cook's distance is for outlier 135. Variance across observations was not constant (χ^2=128.08, df=1, p=< 2.22e-16) and residuals of z-scores were skewed. There was no autocorrelation (p=0.056) and variance inflation factors were below 5.

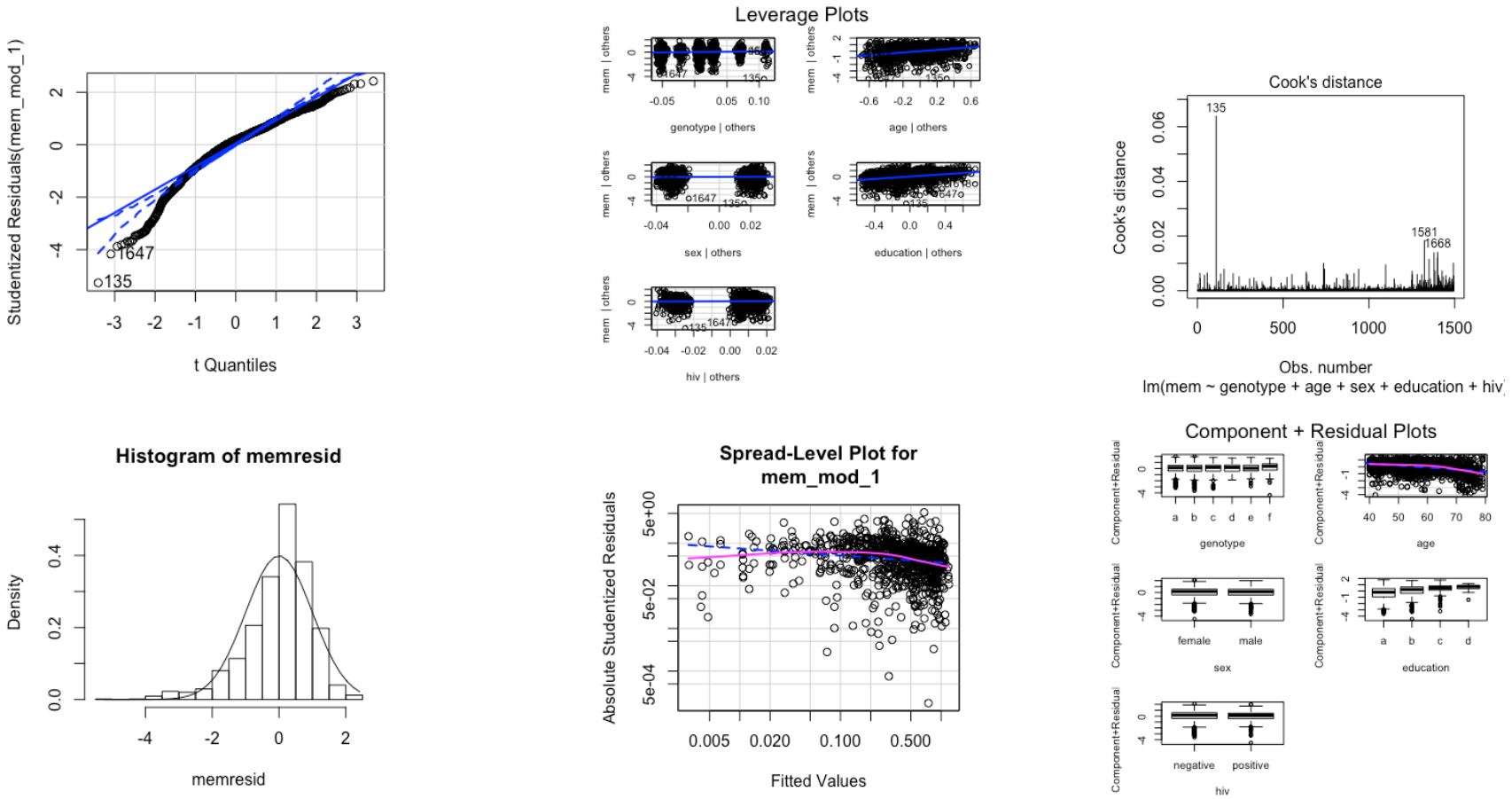


Figure A.2.5: Diagnostic regression plots for the residuals of verbal episodic memory

In order: QQ-plot, Leverage plots for each regression term, Cook's plot indicating Cook's distance, histogram showing distribution of residuals, spread level plot where negative values were removed, and component-residual plots for each term and its effect on memory.

A2.2.5 Visuospatial ability

Two residual outliers were observed during diagnostics of regression analysis, 135 (p=4.519e-06) and 153 (p=1.454e-03). There was minimal effect of the outliers on the distribution of the residuals, and as with the other traits, the most significant effects on the model were age and education. The largest Cook's distance was observed for outlier 135. Non-constant variance was observed for visuospatial ability residual observations (χ^2=74.84, df=1, p=< 2.22e-16) but this does not necessarily imply that linear regression was invalid. Durbin Watson indicated some autocorrelation where p=0.006.

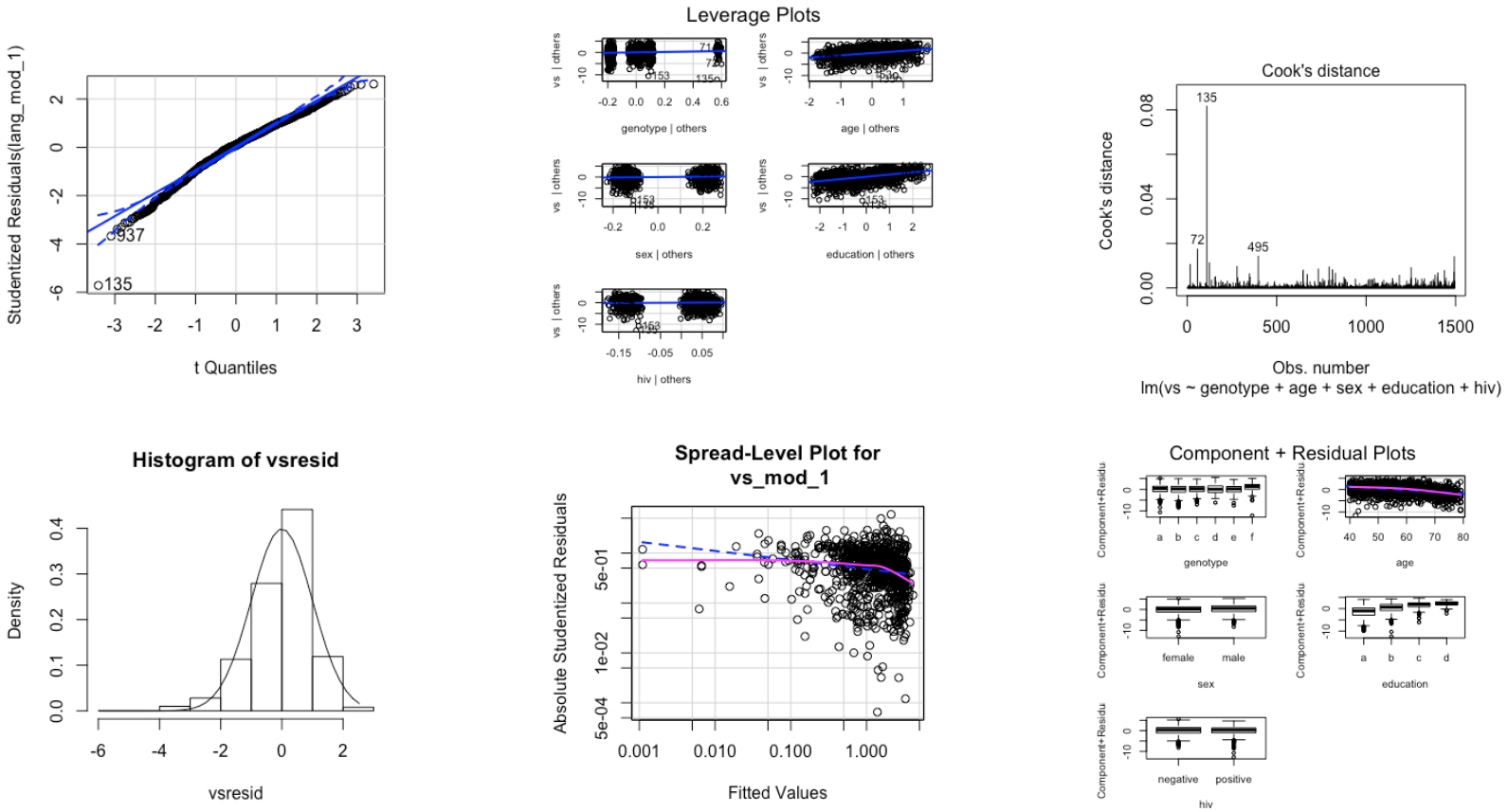


Figure A.2.6: Diagnostic regression plots for the residuals of visuospatial ability

In order: QQ-plot, Leverage plots for each regression term, Cook's plot indicating Cook's distance, histogram showing distribution of residuals, spread level plot where negative values were removed, and component-residual plots for each term and its effect on visuospatial cognition.

Appendix 3

A3.1 Unreported GWAS results for total cognition score

Locus zoom and Kruskal-Wallis plots for the top independent signals associated with total cognition score as listed in Supplementary Table 1 of Chapter 3 with p-values ranging from 1.13e-07 to 1.53e-06 may be found below. There are five SNPs for this trait reported in the table. They were not included in the manuscript as the focus was on genome-wide significance.

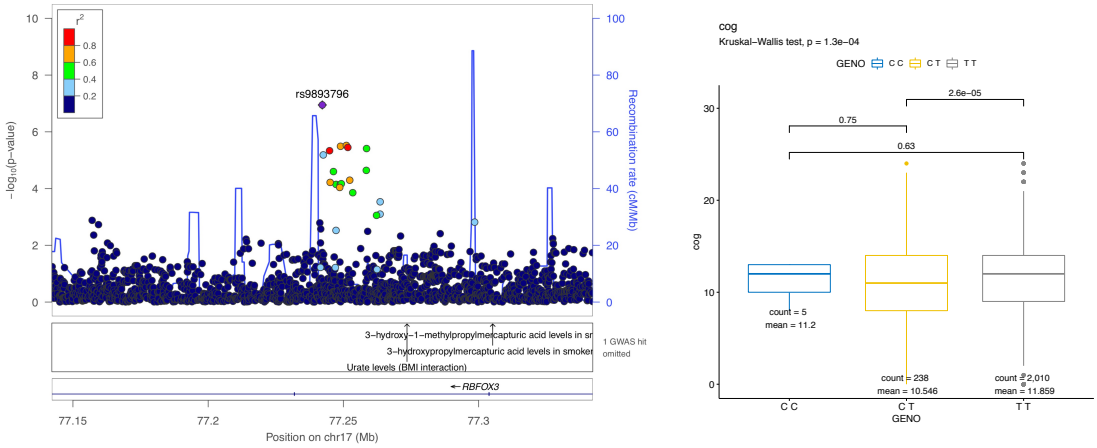


Figure A3.1: rs9893796 on chromosome 17 associated with total cognition score

Locus zoom (left) and Kruskal-Wallis (right) plots.

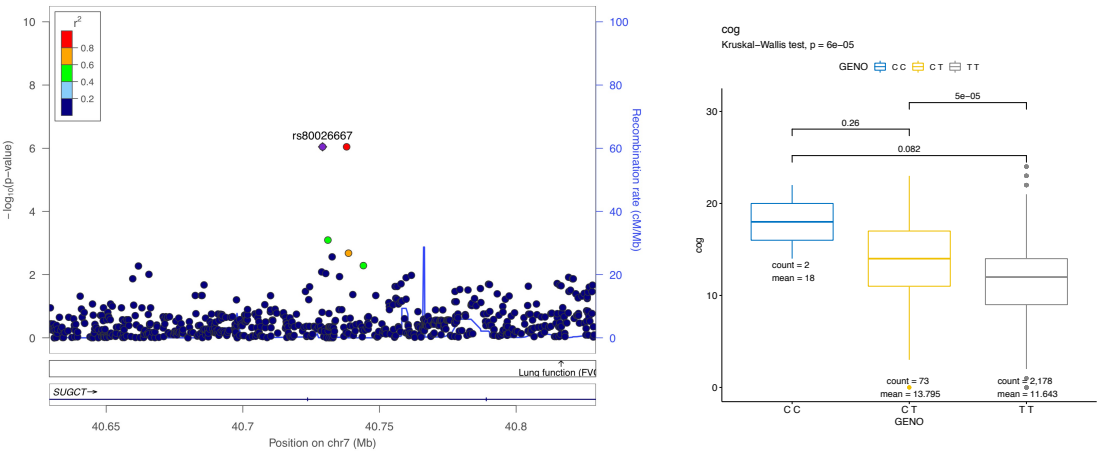


Figure A3.2: rs80026667 on chromosome 7 associated with total cognition score

Locus zoom (left) and Kruskal-Wallis (right) plots.

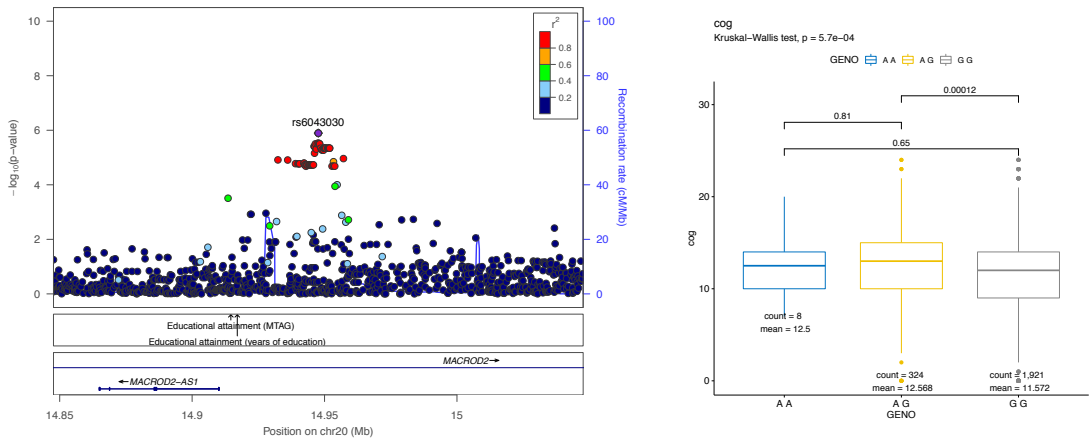


Figure A3.3: rs6043030 on chromosome 20 associated with total cognition score

Locus zoom (left) and Kruskal-Wallis (right) plots.

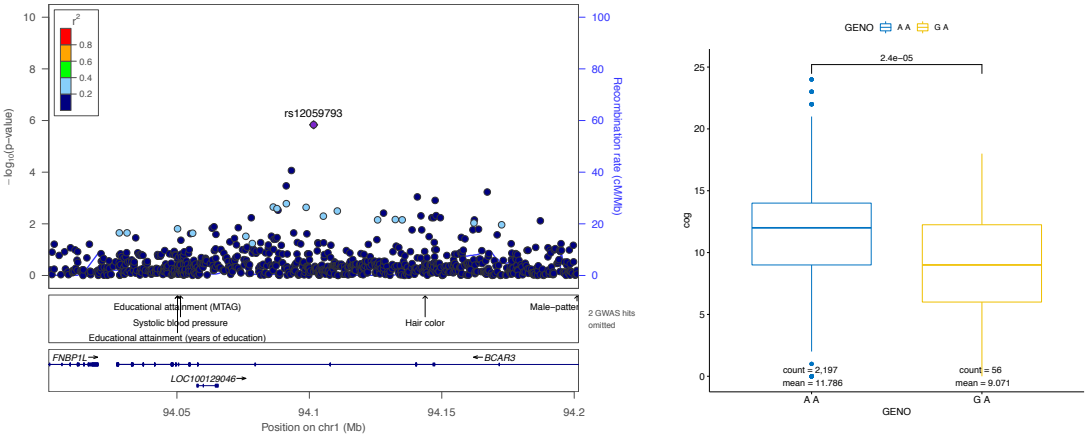


Figure A3.4: rs12059793 on chromosome 1 associated with total cognition score

Locus zoom (left) and Kruskal-Wallis (right) plots.

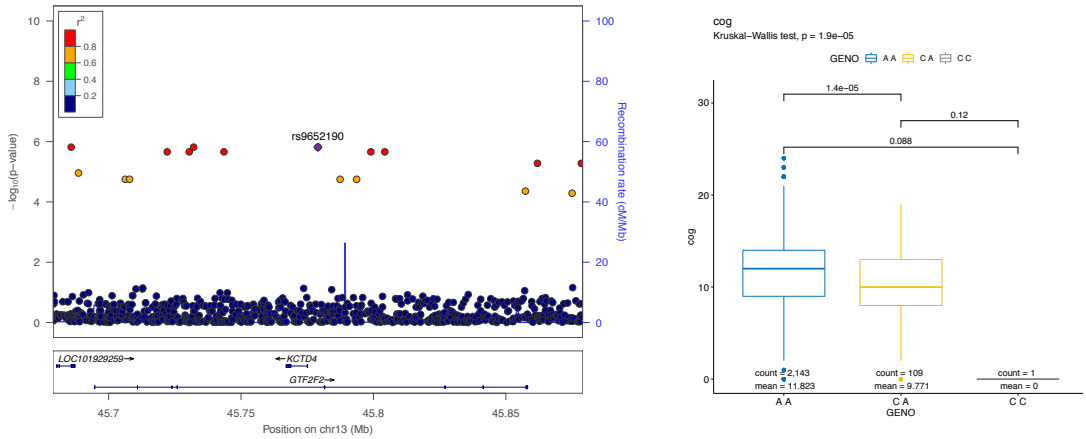


Figure A3.5: rs9652190 on chromosome 13 associated with total cognition score

Locus zoom (left) and Kruskal-Wallis (right) plots.

A3.2 Unreported GWAS results for verbal episodic memory

Locus zoom and Kruskal-Wallis plots for the top signals associated with episodic memory as listed in Supplementary Table 1 of Chapter 3 with p-values ranging from 1.42e-07 to 2.23e-07 may be found below. There were three SNPs for this trait reported in the table but all map to the same region on chromosome 8 so only one figure is included here. They were not included in the manuscript as the focus was on genome-wide significance.

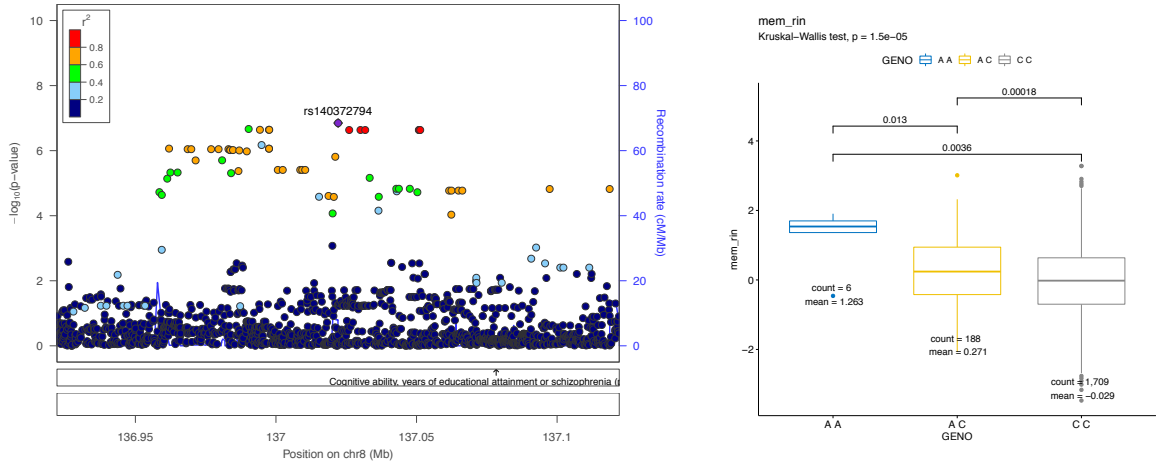


Figure A3.6: rs140372794 on chromosome 8 associated with episodic memory

Locus zoom (left) and Kruskal-Wallis (right) plots.

A3.3 Unreported GWAS results for executive function

Locus zoom and Kruskal-Wallis plots for the top signals associated with executive function as listed in Supplementary Table 1 of Chapter 3 with p-values ranging from 4.68e-07 to 1.41e-06 may be found below. There were seven SNPs for this trait reported in the table as there were no genome-wide significant hits for executive function but two were within 200kb of another signal on their respective chromosome.

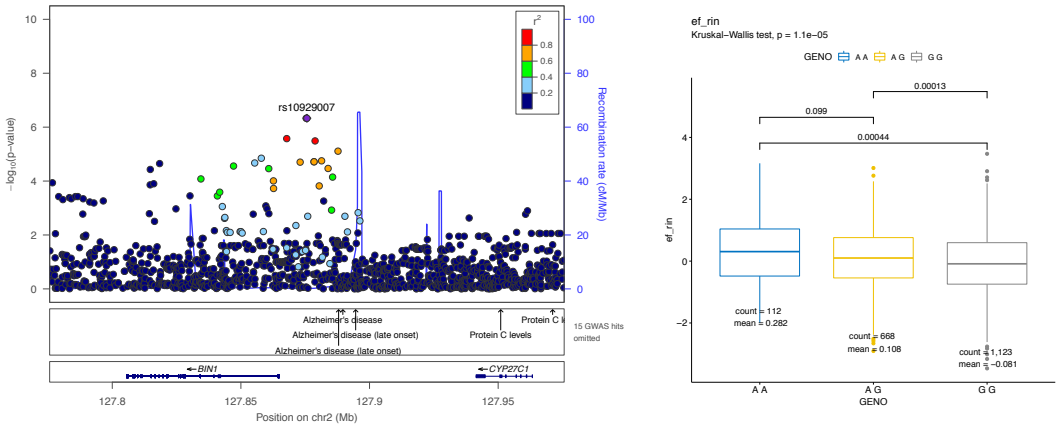


Figure A3.7: rs10929007 on chromosome 2 associated with executive function

Locus zoom (left) and Kruskal-Wallis (right) plots.

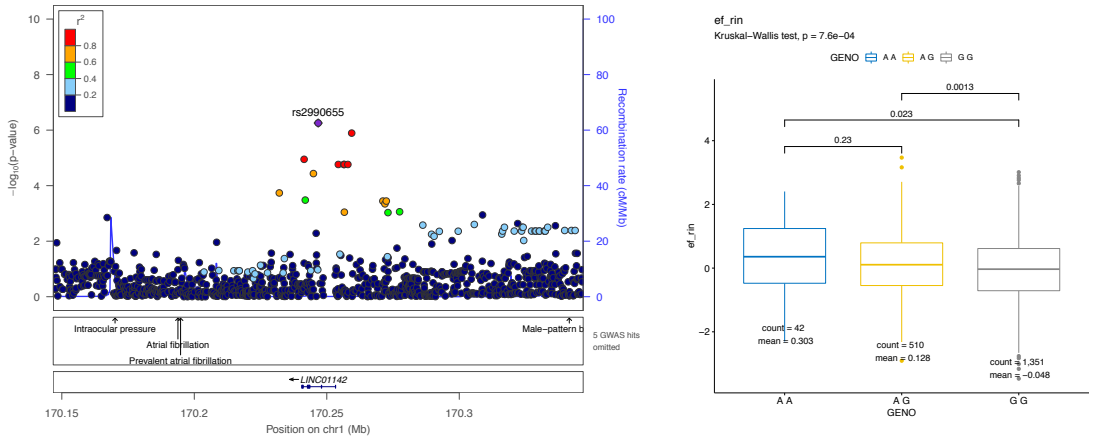


Figure A3.8: rs2990655 on chromosome 1 associated with executive function

Locus zoom (left) and Kruskal-Wallis (right) plots.

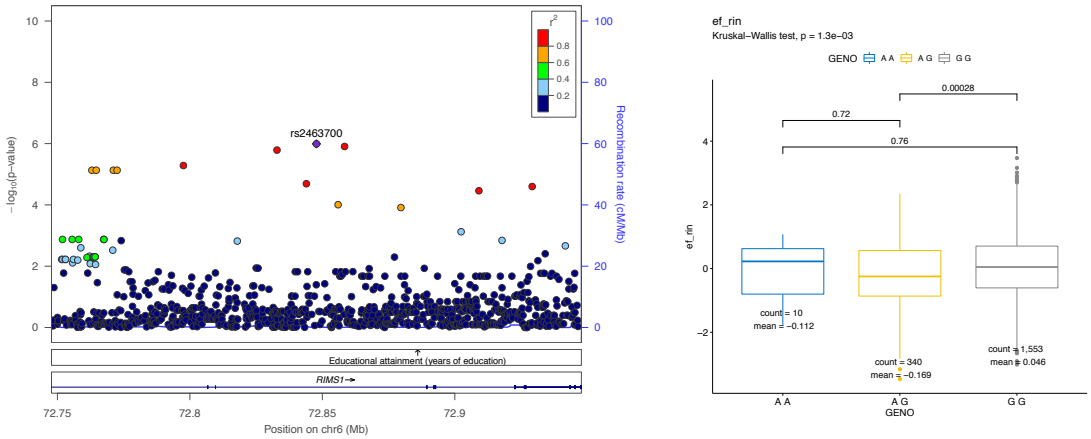


Figure A3.9: rs2463700 on chromosome 6 associated with executive function

Locus zoom (left) and Kruskal-Wallis (right) plots.

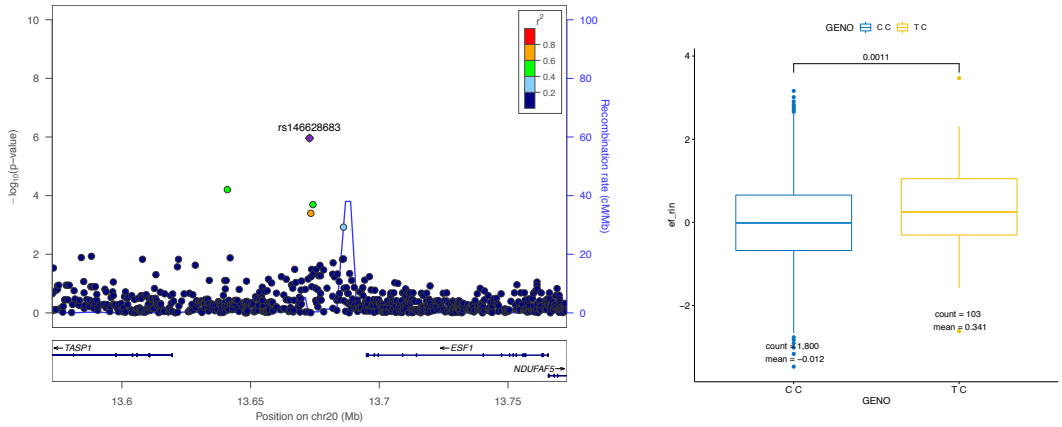


Figure A3.10: rs146628683 on chromosome 20 associated with executive function

Locus zoom (left) and Kruskal-Wallis (right) plots.

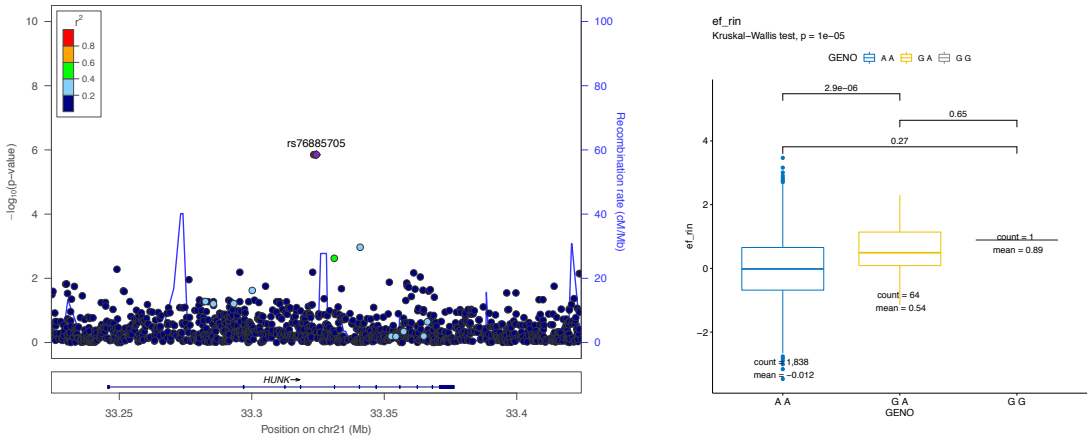


Figure A3.11: rs76885705 on chromosome 21 associated with executive function

Locus zoom (left) and Kruskal-Wallis (right) plots.

A3.4 Unreported GWAS results for language

Locus zoom and Kruskal-Wallis plots for the top signals associated with language as listed in Supplementary Table 1 of Chapter 3 with p-values ranging from 1.94e-07 to 2.03e-06 may be found below.

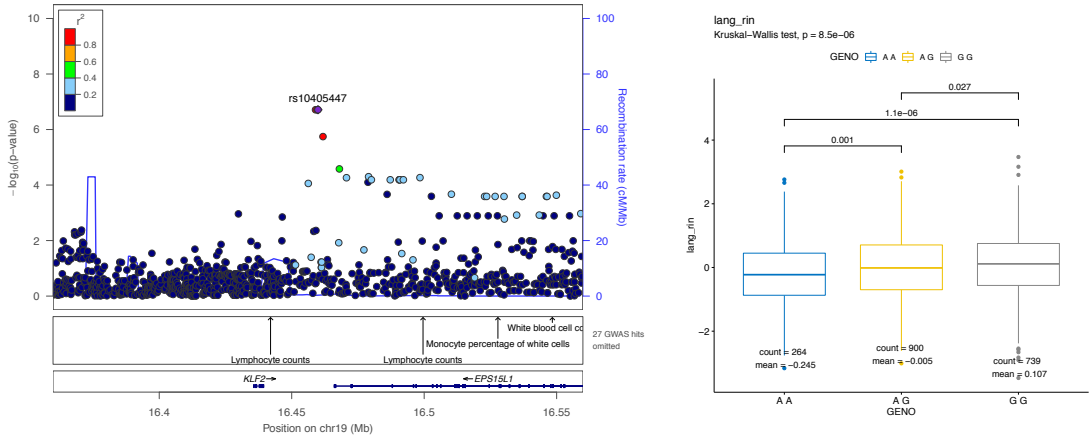


Figure A3.12: rs10405447 on chromosome 19 associated with language

Locus zoom (left) and Kruskal-Wallis (right) plots.

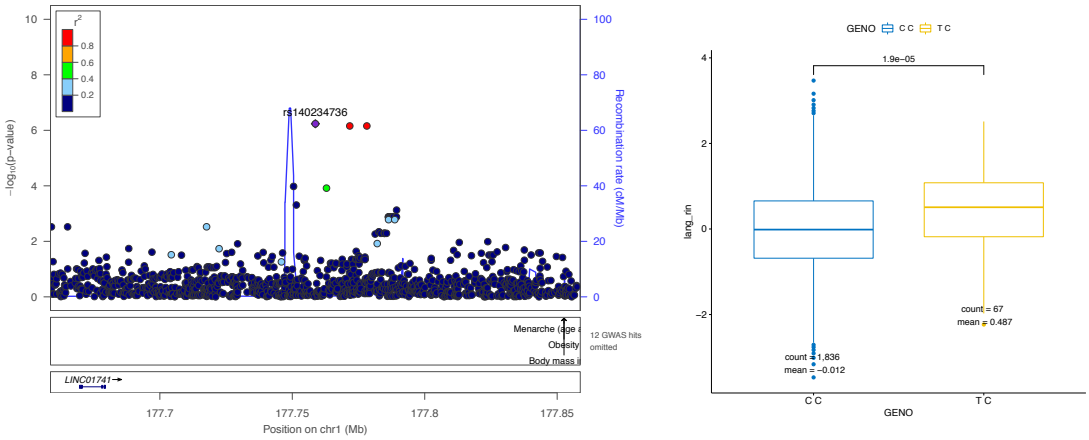


Figure A3.13: rs140234736 on chromosome 1 associated with language

Locus zoom (left) and Kruskal-Wallis (right) plots.

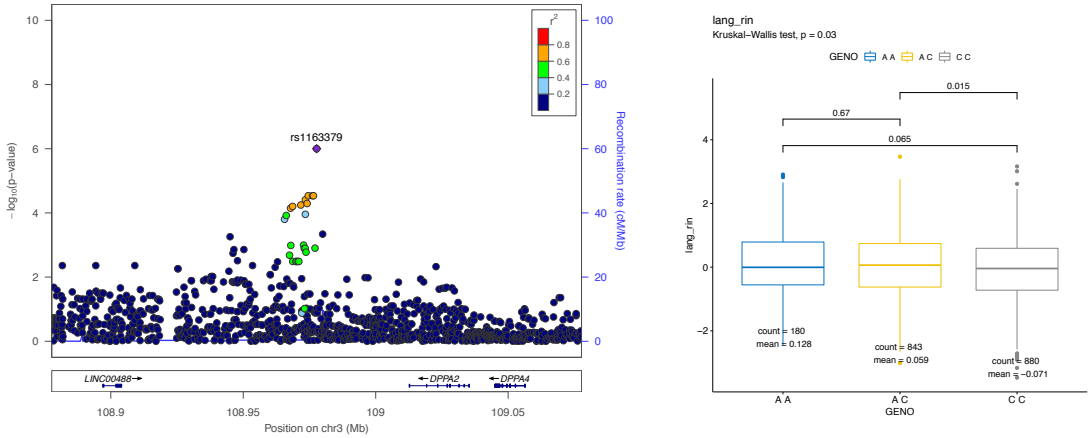


Figure A3.14: rs1163379 on chromosome 3 associated with language

Locus zoom (left) and Kruskal-Wallis (right) plots.

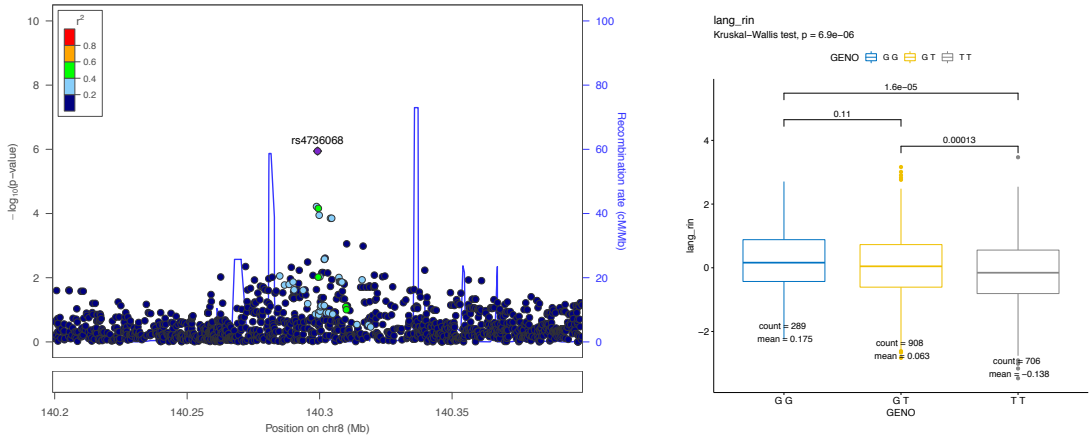


Figure A3.15: rs4736068 on chromosome 8 associated with language

Locus zoom (left) and Kruskal-Wallis (right) plots.

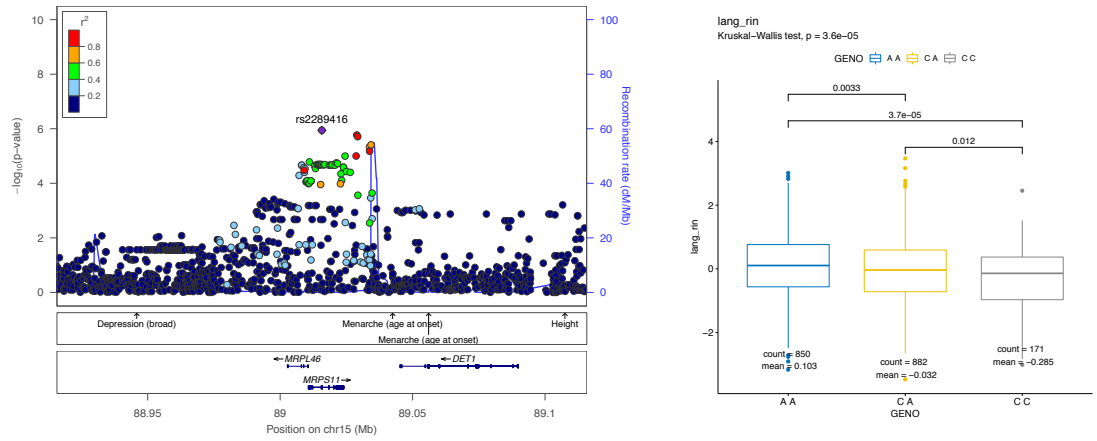


Figure A3.16: rs2289416 on chromosome 15 associated with language

Locus zoom (left) and Kruskal-Wallis (right) plots.

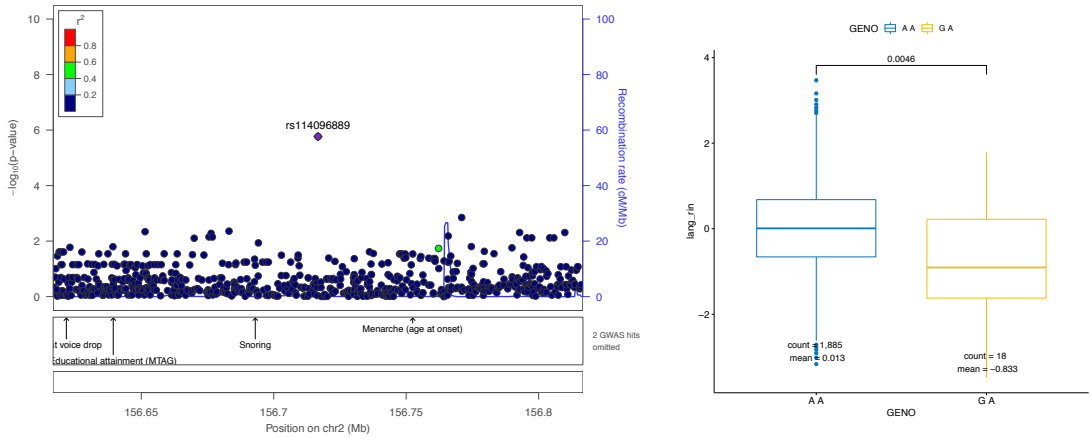


Figure A3.17: rs114096889 on chromosome 2 associated with language

Locus zoom (left) and Kruskal-Wallis (right) plots.

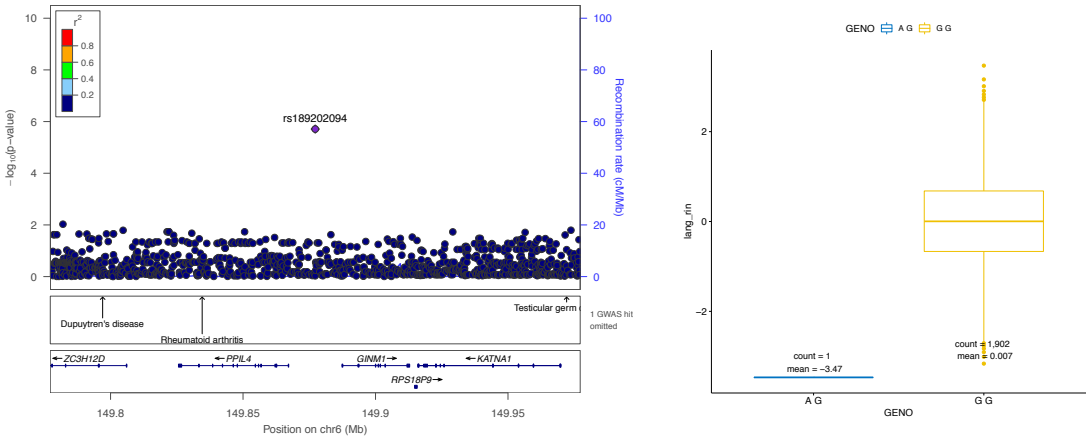


Figure A3.18: rs189202094 on chromosome 6 associated with language

Locus zoom (left) and Kruskal-Wallis (right) plots.

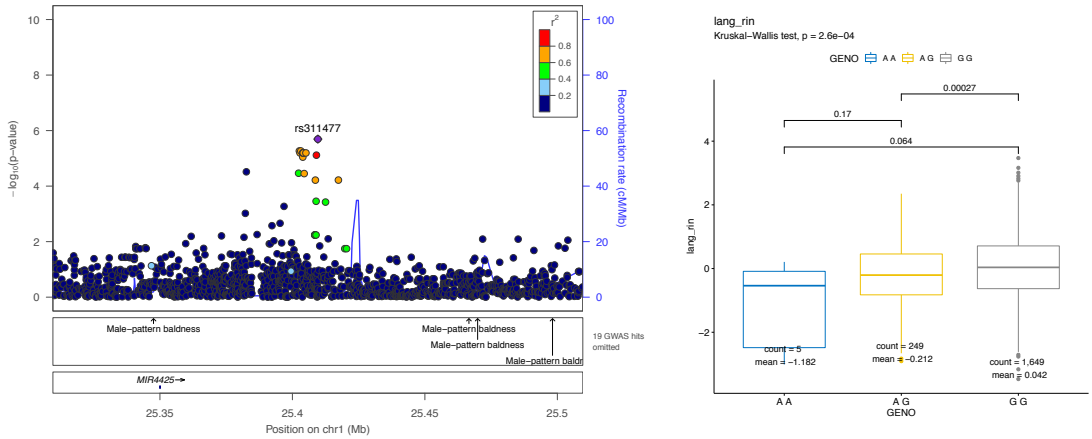


Figure A3.19: rs311477 on chromosome 1 associated with language

Locus zoom (left) and Kruskal-Wallis (right) plots.

A3.5 Unreported GWAS results for visuospatial cognition

Locus zoom and Kruskal-Wallis plots for the top signals associated with visuospatial ability as listed in Supplementary Table 1 of Chapter 3 with p-values ranging from 2.43e-07 to 3.75e-06 may be found below. All ten signals are shown here as none reached genome-wide significance. Three SNPs, rs114096889, rs189202094, and rs140578927 had also been reported above for association with language.

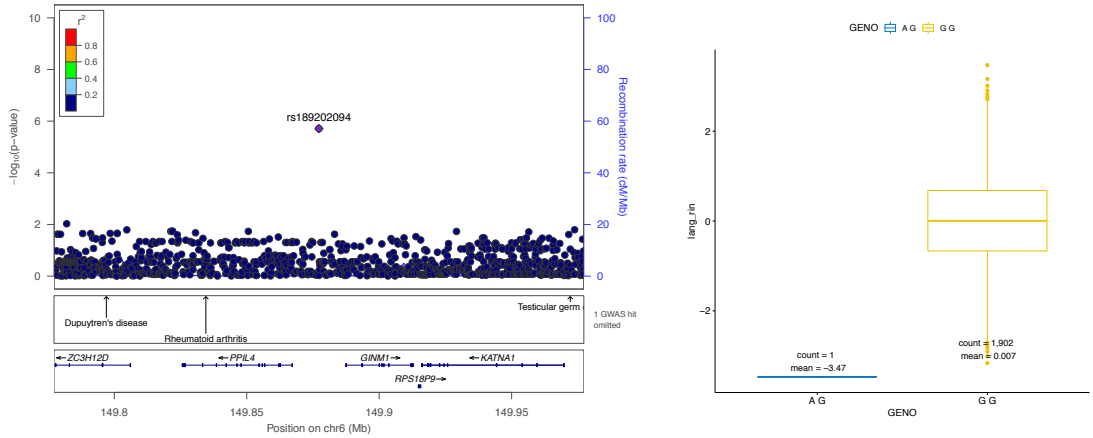


Figure A3.20: rs189202094 on chromosome 6 associated with visuospatial ability

Locus zoom (left) and Kruskal-Wallis (right) plots.

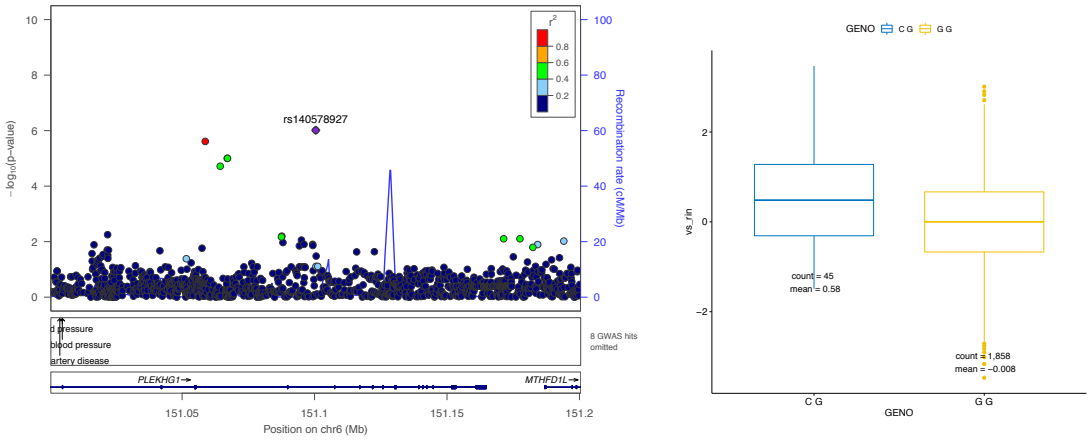


Figure A3.21: rs140578927 on chromosome 6 associated with visuospatial ability

Locus zoom (left) and Kruskal-Wallis (right) plots.

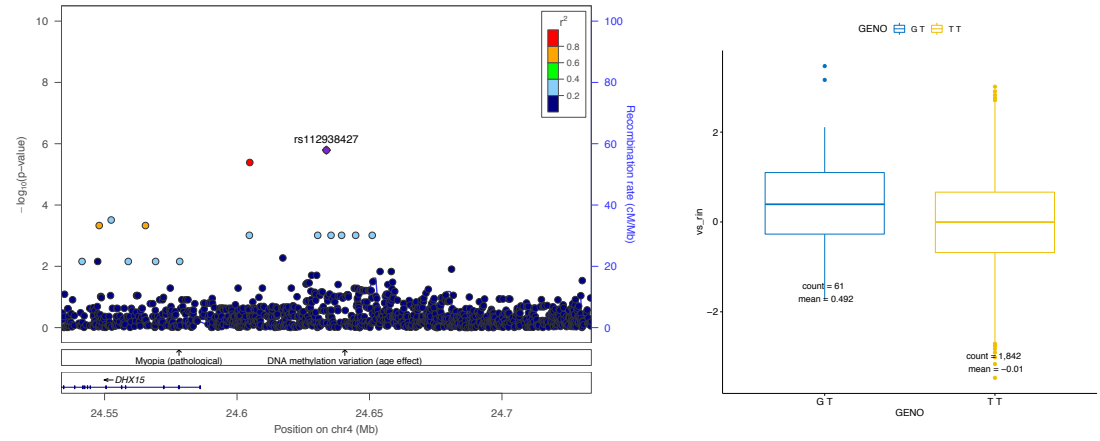


Figure A3.22: rs112938427 on chromosome 4 associated with visuospatial ability

Locus zoom (left) and Kruskal-Wallis (right) plots.

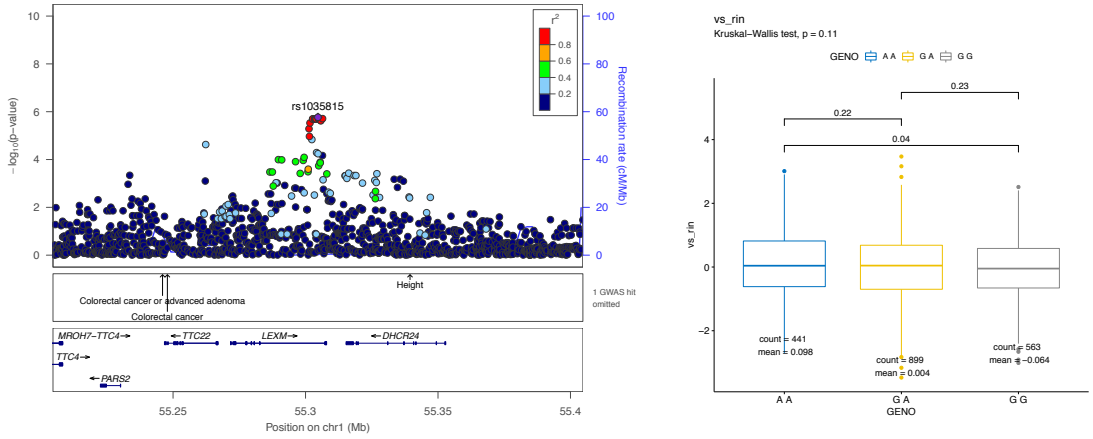


Figure A3.23: rs1035815 on chromosome 1 associated with visuospatial ability

Locus zoom (left) and Kruskal-Wallis (right) plots.

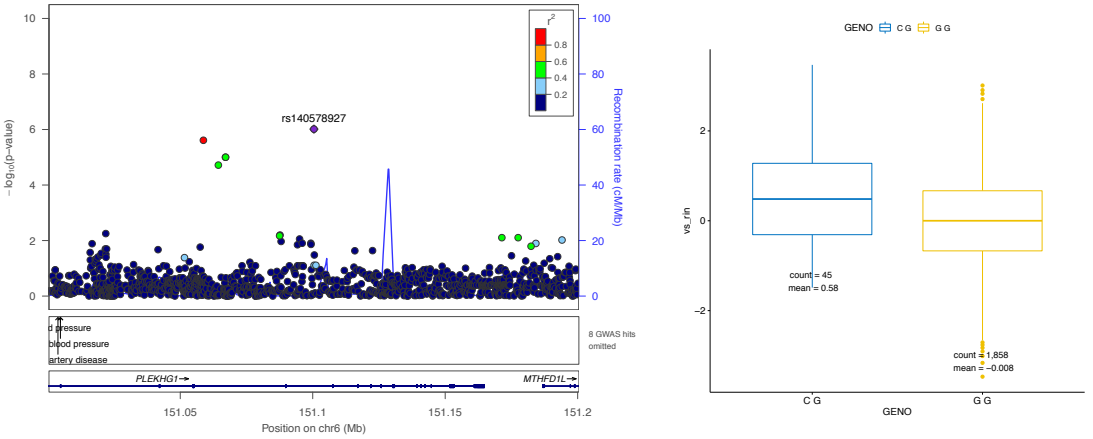


Figure A3.24: rs114096889 on chromosome 6 associated with visuospatial ability

Locus zoom (left) and Kruskal-Wallis (right) plots.

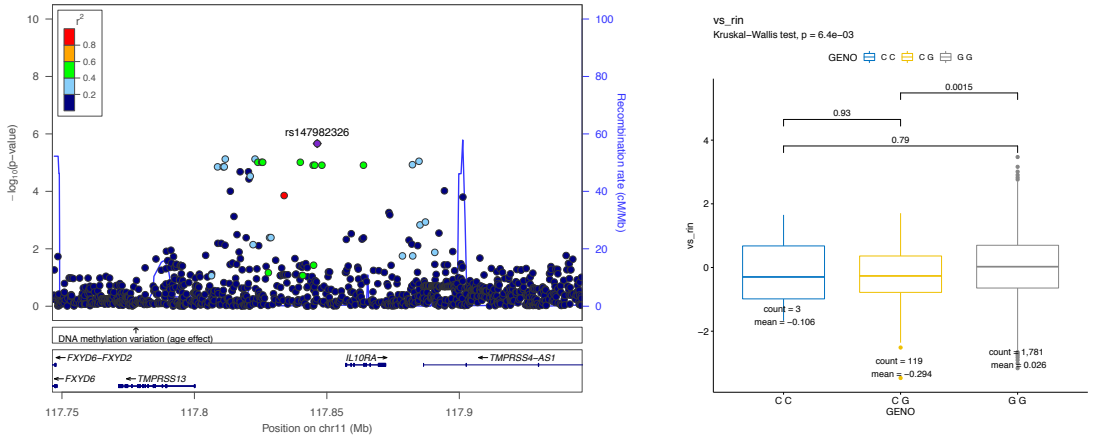


Figure A3.25: rs147982326 on chromosome 11 associated with visuospatial ability

Locus zoom (left) and Kruskal-Wallis (right) plots.

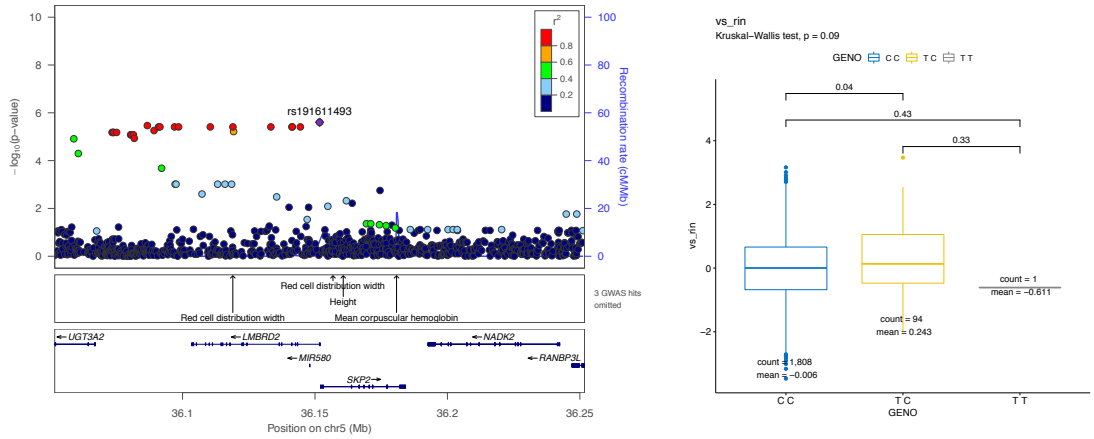


Figure A3.26: rs191611493 on chromosome 5 associated with visuospatial ability

Locus zoom (left) and Kruskal-Wallis (right) plots.

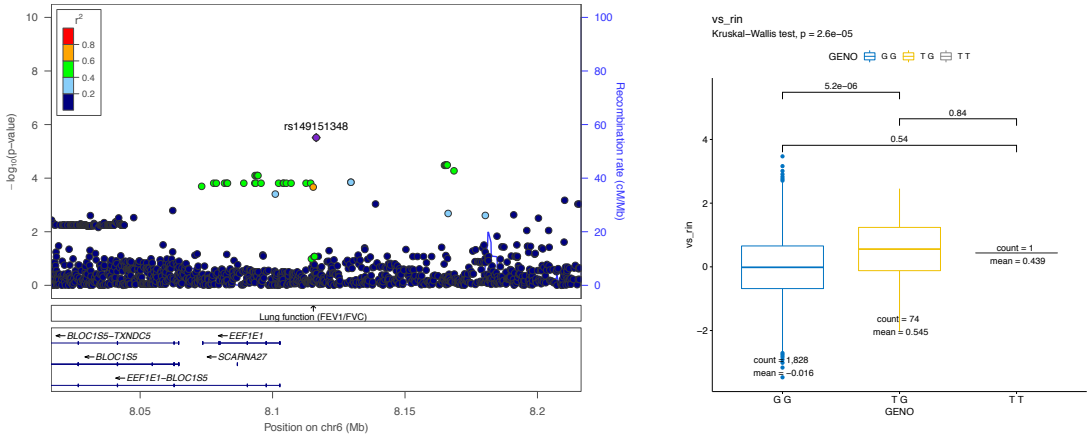


Figure A3.27: rs149151348 on chromosome 6 associated with visuospatial ability

Locus zoom (left) and Kruskal-Wallis (right) plots.

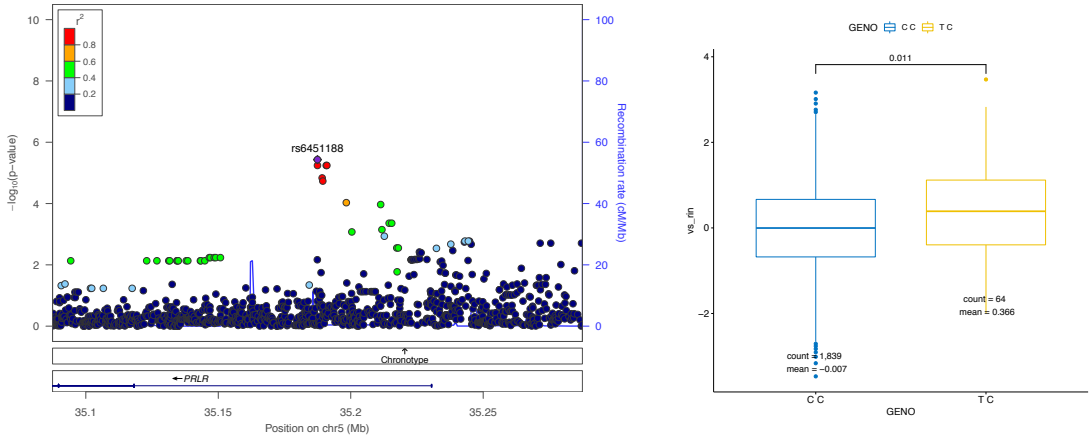


Figure A3.28: rs6451188 on chromosome 5 associated with visuospatial ability

Locus zoom (left) and Kruskal-Wallis (right) plots.

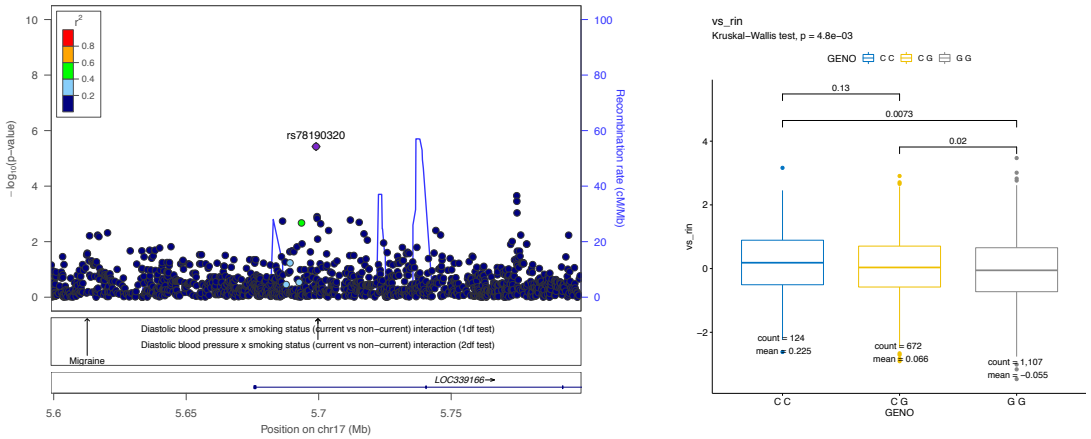


Figure A3.29: rs78190320 on chromosome 17 associated with visuospatial ability

Locus zoom (left) and Kruskal-Wallis (right) plots.

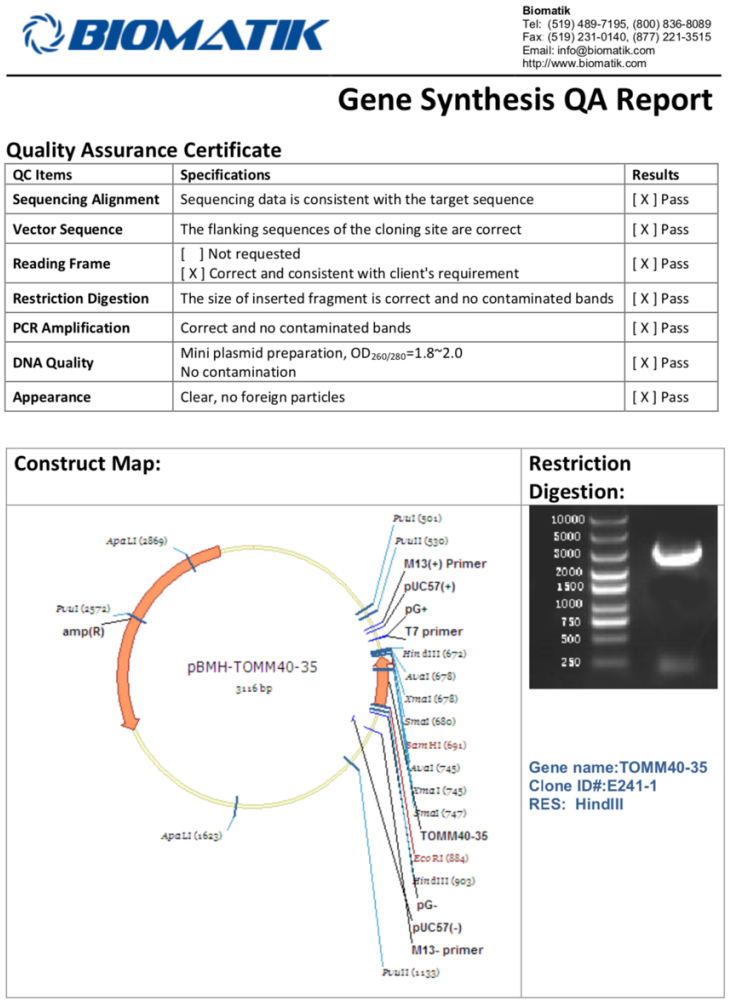
Appendix 4

A4.1 Homopolymer length detection

A4.1.1 Poly-T length fragment standards

The sequence to which alignment is presented for each fragment standard is as follows:

- T=35
  >CTGCTGACCTCAAGCTGTCCTCTTGCCCCAGCCCTCCAAAGCATTGGGAT
  TACTGGCATGAGCCATTGCATCTGGCTTTTTTTTTTTTTTTTTTTTTTTT
  TTTTTTTTGAGATGGGGTCTCACCATGTTGCCCGGGCTGGCTTCGAATGCC
  TAGGCTCAAGCAATCCTCCCTTCTCAGCCTCCC
- T=16
  >CTGCTGACCTCAAGCTGTCCTCTTGCCCCAGCCCTCCAAAGCATTGGGAT
  TACTGGCATGAGCCATTGCATCTGGCTTTTTTTTTTTTTTTTGAGATGGGGT
  CTCACCATGTTGCCCGGGCTGGCTTCGAATGCCTAGGCTCAAGCAATCCTC
  CCTTCTCAGCCTCCC
- T=18
  >CTGCTGACCTCAAGCTGTCCTCTTGCCCCAGCCCTCCAAAGCATTGGGAT
  TACTGGCATGAGCCATTGCATCTGGCTTTTTTTTTTTTTTTTTGAGATGGG
  GTCTCACCATGTTGCCCGGGCTGGCTTCGAATGCCTAGGCTCAAGCAATCC
  TCCCTTCTCAGCCTCCC
- T=22
  >CTGCTGACCTCAAGCTGTCCTCTTGCCCCAGCCCTCCAAAGCATTGGGAT
  TACTGGCATGAGCCATTGCATCTGGCTTTTTTTTTTTTTTTTTTTTGAGAT
  GGGGTCTCACCATGTTGCCCGGGCTGGCTTCGAATGCCTAGGCTCAAGCA
  ATCCTCCCTTCTCAGCCTCCC



Biomatik
Tel: (519) 489-7195, (800) 836-8089
Fax: (519) 231-0140, (877) 221-3515
Email: info@biomatik.com
http://www.biomatik.com

Gene Synthesis QA Report

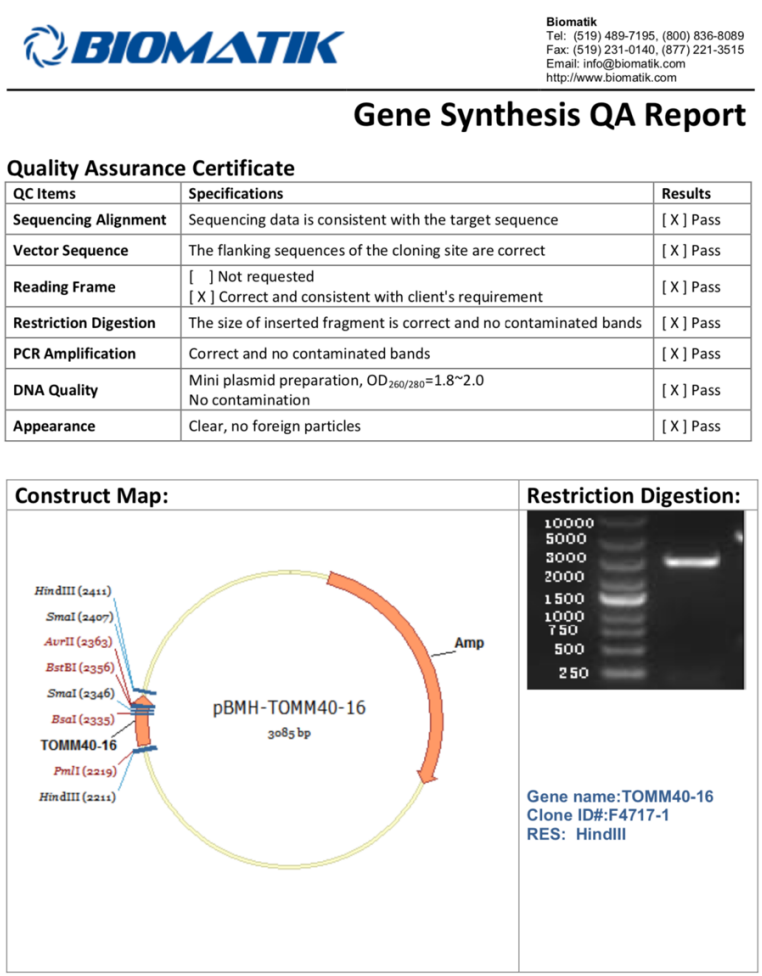
Quality Assurance Certificate

QC Items	Specifications	Results
Sequencing Alignment	Sequencing data is consistent with the target sequence	[X] Pass
Vector Sequence	The flanking sequences of the cloning site are correct	[X] Pass
Reading Frame	[ ] Not requested [X] Correct and consistent with client's requirement	[X] Pass
Restriction Digestion	The size of inserted fragment is correct and no contaminated bands	[X] Pass
PCR Amplification	Correct and no contaminated bands	[X] Pass
DNA Quality	Mini plasmid preparation, $OD_{260/280}$=1.8~2.0 No contamination	[X] Pass
Appearance	Clear, no foreign particles	[X] Pass

Construct Map:

Restriction Digestion:

Gene name:TOMM40-35
Clone ID#:E241-1
RES: HindIII



Biomatik
Tel: (519) 489-7195, (800) 836-8089
Fax: (519) 231-0140, (877) 221-3515
Email: info@biomatik.com
http://www.biomatik.com

Gene Synthesis QA Report

Quality Assurance Certificate

QC Items	Specifications	Results
Sequencing Alignment	Sequencing data is consistent with the target sequence	[X] Pass
Vector Sequence	The flanking sequences of the cloning site are correct	[X] Pass
Reading Frame	[ ] Not requested [X] Correct and consistent with client's requirement	[X] Pass
Restriction Digestion	The size of inserted fragment is correct and no contaminated bands	[X] Pass
PCR Amplification	Correct and no contaminated bands	[X] Pass
DNA Quality	Mini plasmid preparation, $OD_{260/280}$=1.8~2.0 No contamination	[X] Pass
Appearance	Clear, no foreign particles	[X] Pass

Construct Map:

Restriction Digestion:

Gene name:TOMM40-16
Clone ID#:F4717-1
RES: HindIII

Figure A4.1: Gene synthesis report for clone for fragment length T=35 and T=16

Reports provided by BIOMATIK through Inqaba Biotec for cloned fragment standards T=35 (left) and T=16 (right)

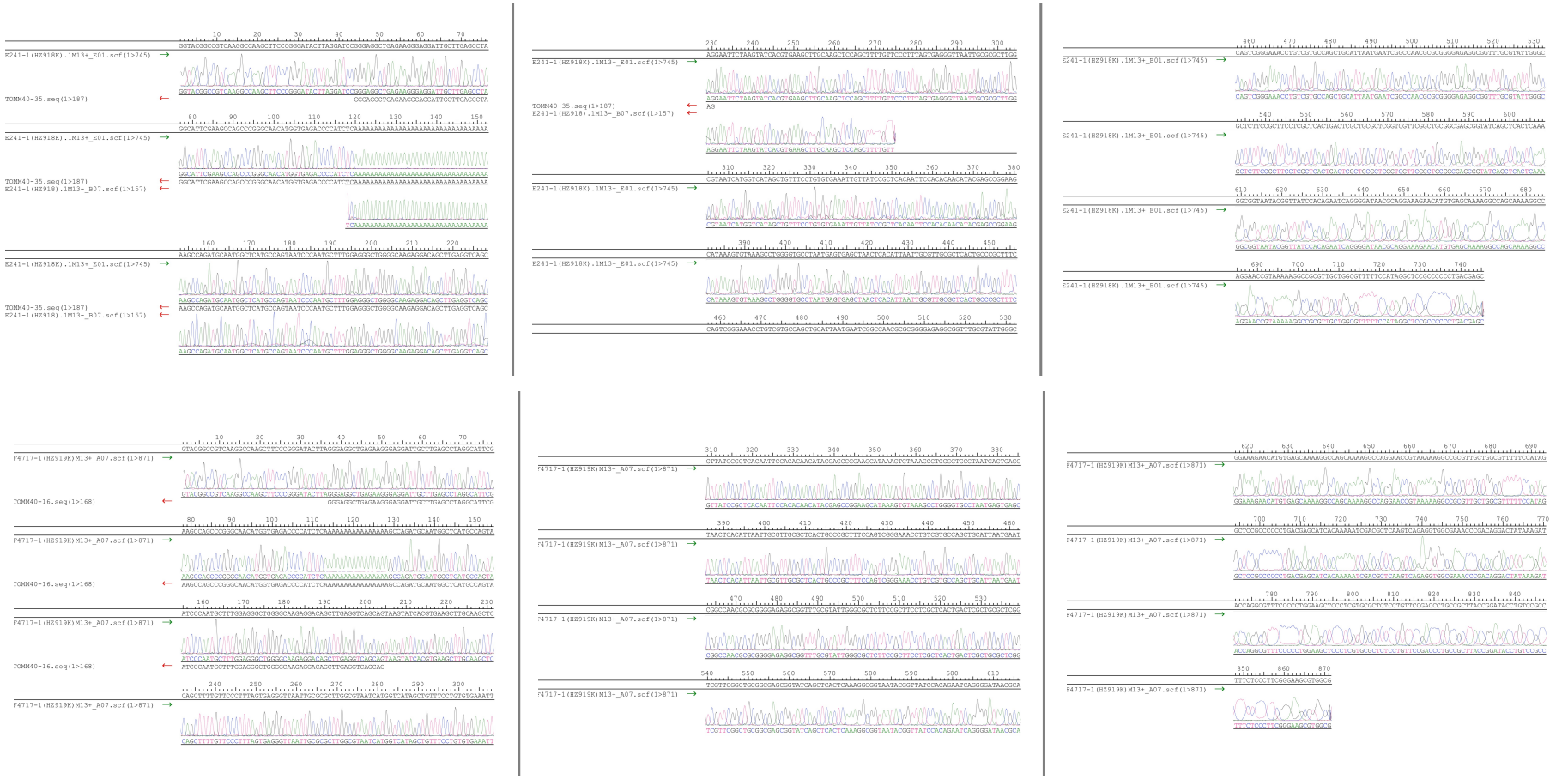
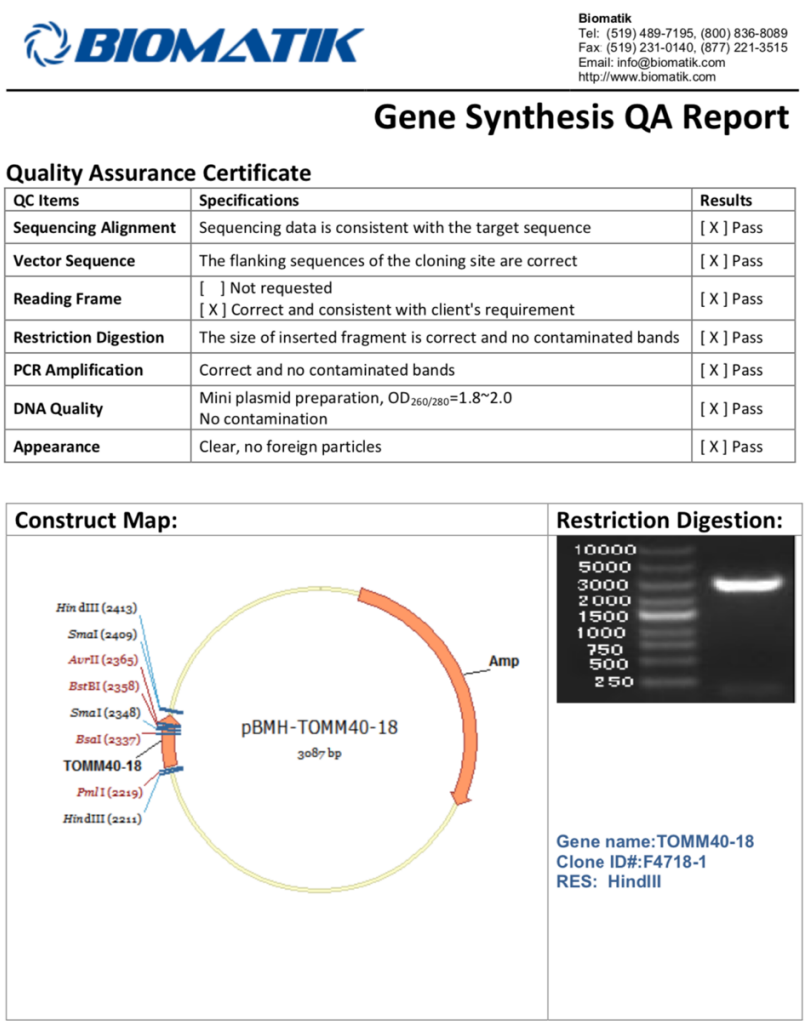


Figure A4.2: Alignment files for fragment standards T=35 and T=16

Alignment results for fragment standard T=35 (top) and T=16 (bottom). Each different coloured peak in the electrophoretogram represents a different base where A is green, T is red, G is black, and C is blue.



BIOMATIK

Biomatik
Tel: (519) 489-7195, (800) 836-8089
Fax: (519) 231-0140, (877) 221-3515
Email: info@biomatik.com
http://www.biomatik.com

Gene Synthesis QA Report

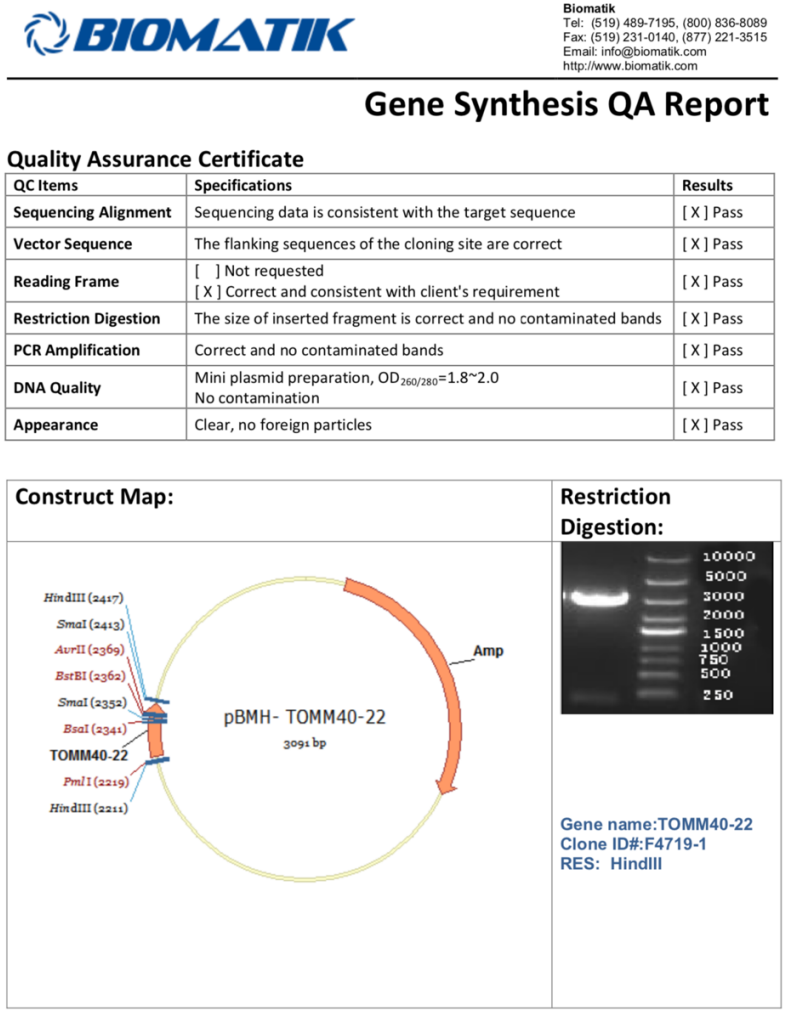
Quality Assurance Certificate

QC Items	Specifications	Results
Sequencing Alignment	Sequencing data is consistent with the target sequence	[X] Pass
Vector Sequence	The flanking sequences of the cloning site are correct	[X] Pass
Reading Frame	[ ] Not requested [X] Correct and consistent with client's requirement	[X] Pass
Restriction Digestion	The size of inserted fragment is correct and no contaminated bands	[X] Pass
PCR Amplification	Correct and no contaminated bands	[X] Pass
DNA Quality	Mini plasmid preparation, $OD_{260/280}$=1.8~2.0 No contamination	[X] Pass
Appearance	Clear, no foreign particles	[X] Pass

Construct Map:

Restriction Digestion:

Gene name:TOMM40-18
Clone ID#:F4718-1
RES: HindIII



BIOMATIK

Biomatik
Tel: (519) 489-7195, (800) 836-8089
Fax: (519) 231-0140, (877) 221-3515
Email: info@biomatik.com
http://www.biomatik.com

Gene Synthesis QA Report

Quality Assurance Certificate

QC Items	Specifications	Results
Sequencing Alignment	Sequencing data is consistent with the target sequence	[X] Pass
Vector Sequence	The flanking sequences of the cloning site are correct	[X] Pass
Reading Frame	[ ] Not requested [X] Correct and consistent with client's requirement	[X] Pass
Restriction Digestion	The size of inserted fragment is correct and no contaminated bands	[X] Pass
PCR Amplification	Correct and no contaminated bands	[X] Pass
DNA Quality	Mini plasmid preparation, $OD_{260/280}$=1.8~2.0 No contamination	[X] Pass
Appearance	Clear, no foreign particles	[X] Pass

Construct Map:

Restriction Digestion:

Gene name:TOMM40-22
Clone ID#:F4719-1
RES: HindIII

Figure A4.3: Gene synthesis report for clone for fragment length T=18 and T=22

Reports provided by BIOMATIK through Inqaba Biotec for cloned fragment standards T=18 (left) and T=22 (right)

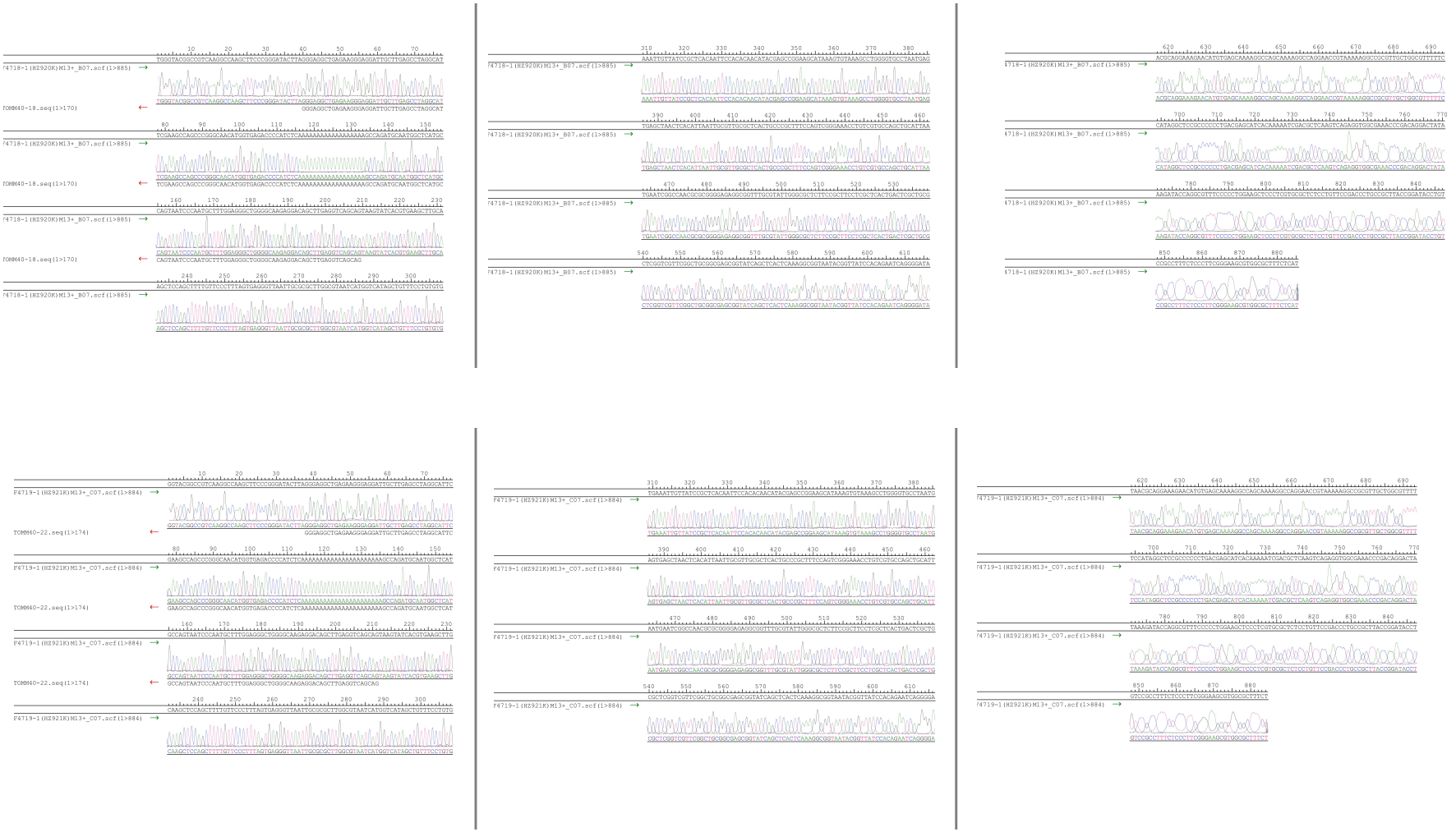


Figure A4.4: Alignment files for fragment standards T=18 and T=22

Alignment results for fragment standard T=18 (top) and T=22 (bottom). Each different coloured peak in the electrophoretogram represents a different base where A is green, T is red, G is black, and C is blue.

A4.1.2 Poly-T length estimation methods

Table A4.1: Comparison of results of three methods for poly-T length determination

ID	PCR analysis	Inhouse caller	HipSTR
AB0009	S/S	S/S	S|S
AB0330	S/S	S/S	S|S
AB0528	S/S	S/S	S|S
AB0867	S/S	S/S	S|S
AB0874	NA	S/S	S|S
AB1089	S/L	S/L	S|L
AB1223	S/VL	S/VL	NA
AB1276	S/VL	S/VL	S|VL
AB1298	S/S	S/S	S|S
AB1313	S/S	S/S	S|S
AB2092	S/S	S/S	S|S
AB2098	S/S	S/S	S|S
AB2163	L/VL	L/VL	L|VL
AB2305	S/VL	S/VL	NA
AB2484	S/S	S/S	S|S
AB2496	S/S	S/S	S|S
AB2502	S/S	S/S	S|S
AB2513	S/VL	S/VL	S|VL
AB2533	S/S	S/S	S|S
AB2543	S/S	S/L	S|L
AB2721	L/VL	L/L	L|VL
AB2867	S/S	S/S	S|S
AB2870	S/S	S/VL	S|VL
AB2967	S/S	S/S	S|S
ACH0L	S/VL	S/VL	S|VL
ACS0X	S/VL	S/VL	S|VL

ADC0I	S/VL	S/VL	S|VL
AFY0J	S/S	S/S	S|S
AGM0B	S/S	VL/VL	NA
AGN0C	NA	S/L	S|L
AGQ0E	S/S	S/S	S|S
AIR0J	S/L	S/L	S|L
AIZ0R	S/VL	S/S	VL|S
AJK0F	S/L	S/L	S|L
AJX0R	S/VL	S/VL	S|VL
AKM0J	S/VL	S/VL	S|VL
	PCR	**Inhouse**	
ID	**analysis**	**caller**	**HipSTR**
ANG0J	S/VL	S/VL	S|VL
ANY0C	S/S	S/S	S|S
AQV0D	L/VL	S/VL	S|VL
ASP0C	L/L	S/L	S|L
AST0G	S/S	S/S	S|S
ATM0C	S/S	L/L	S|L
AVR0I	L/VL	L/VL	VL|L
AXA0V	S/L	S/L	S|L
AZZ0A	S/S	S/S	S|S
BBX0D	S/VL	S/VL	S|VL
BFE0S	S/S	S/S	S|S
BFH0W	S/S	S/S	S|S
BKX0X	S/S	S/S	S|S
BLP0R	S/L	S/L	L|S
BSH0X	S/S	S/S	S|S
BST0J	S/VL	S/VL	S|VL
CBB0J	S/S	S/L	L|S
CBM0W	S/VL	S/S	S|VL
CCL0X	S/VL	S/VL	S|VL
CFF0X	S/L	S/L	S|L
CFJ0C	S/S	S/S	S|S

CFK0D	S/S	S/S	S|S
CGB0V	S/VL	S/VL	S|VL
CGI0D	S/VL	S/VL	S|VL
CMM0V	S/VL	S/VL	S|VL
CPW0I	VL/VL	VL/VL	VL|VL
CVN0L	S/S	S/L	S|L
CWF0F	S/S	S/S	S|S
CXV0X	S/S	S/S	S|S
CYE0I	S/S	S/VL	S|VL
DBB0M	S/VL	S/VL	S|VL
DDB0R	S/S	S/S	S|S
DEC0V	S/S	S/S	S|S
DEP0I	S/S	S/S	S|S
DFX0S	VL/VL	VL/VL	NA
DGN0L	S/VL	S/VL	S|VL
DHY0Y	S/VL	S/VL	S|VL
DLB0J	S/S	L/L	L|L
RCZ0G	S/S	S/S	S|S
RDH0Q	S/S	S/S	S|S
RDJ0S	S/L	S/L	S|L
REN0A	S/L	S/L	S|L
ID	**PCR analysis**	**Inhouse caller**	**HipSTR**
REV0G	S/S	S/S	S|S
REW0H	S/VL	S/S	NA
RFA0M	S/S	S/S	S|S
RFI0W	S/VL	S/VL	S|VL
RFQ0E	S/S	S/S	S|S
RJN0K	S/L	S/L	S|L
RMP0S	S/VL	S/VL	S|VL
RRT0G	S/L	S/L	S|L
RVB0T	S/S	S/S	S|S
RWB0W	S/S	S/S	S|S

RXV0S	S/VL	S/VL	S|VL
RZD0F	S/L	S/S	S|S
TDF0V	S/VL	S/S	VL|S
THT0S	S/S	S/S	S|S
TIH0I	S/VL	S/L	L|S
VBC0Q	S/VL	S/S	S|S
VGD0D	S/VL	S/VL	S|VL
VKS0C	S/VL	S/VL	S|VL
VLD0N	S/VL	S/VL	S|VL
VNB0Q	S/S	S/S	S|S
VSM0L	S/S	S/S	S|S
VVJ0M	S/S	S/S	S|S

Inhouse STR-caller developed by Dr Brandenburg, PCR performed by Inqaba Biotec, HipSTR (Willems, et al., 2017). S where T≤19 , L where T=20 to 29, VL where T≥30. / represents unphased alleles, | represents phased alleles. NA is an undetermined result.

An example of the result files from the PCR and fragment length analysis (**Figure A4.5**) and the output from the homebrew STR-caller developed by Dr Brandenburg (**Figure A4.6**) for the sample AB0009. The peak observed at T~19 was considered to be an artifact of the experiment as every sample had that same observed peak pattern at that spot. HipSTR outputs VCF files counting Ts.

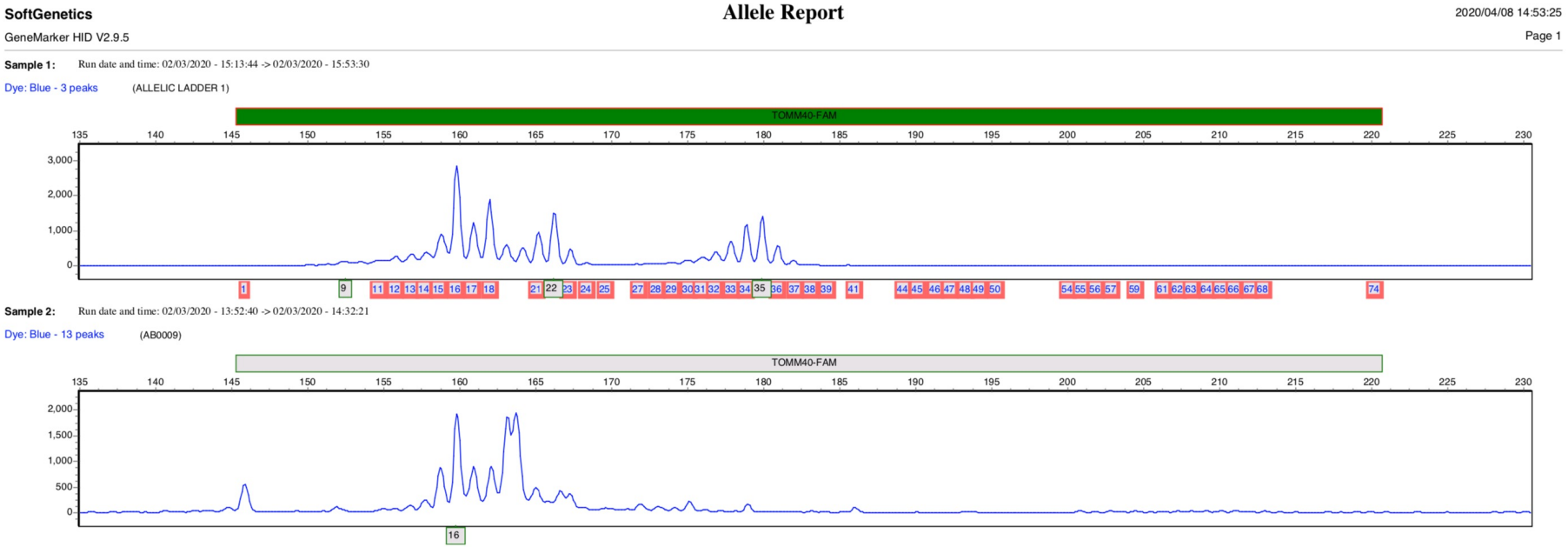


Figure A4.5 Allele report from fragment analysis from PCR fragment analysis for AB0009

The allelelic ladder contained the fragment markers for estimation of poly-T length (top). Sample peaks indicate result is homozygous T=16 therefore S/S (bottom).

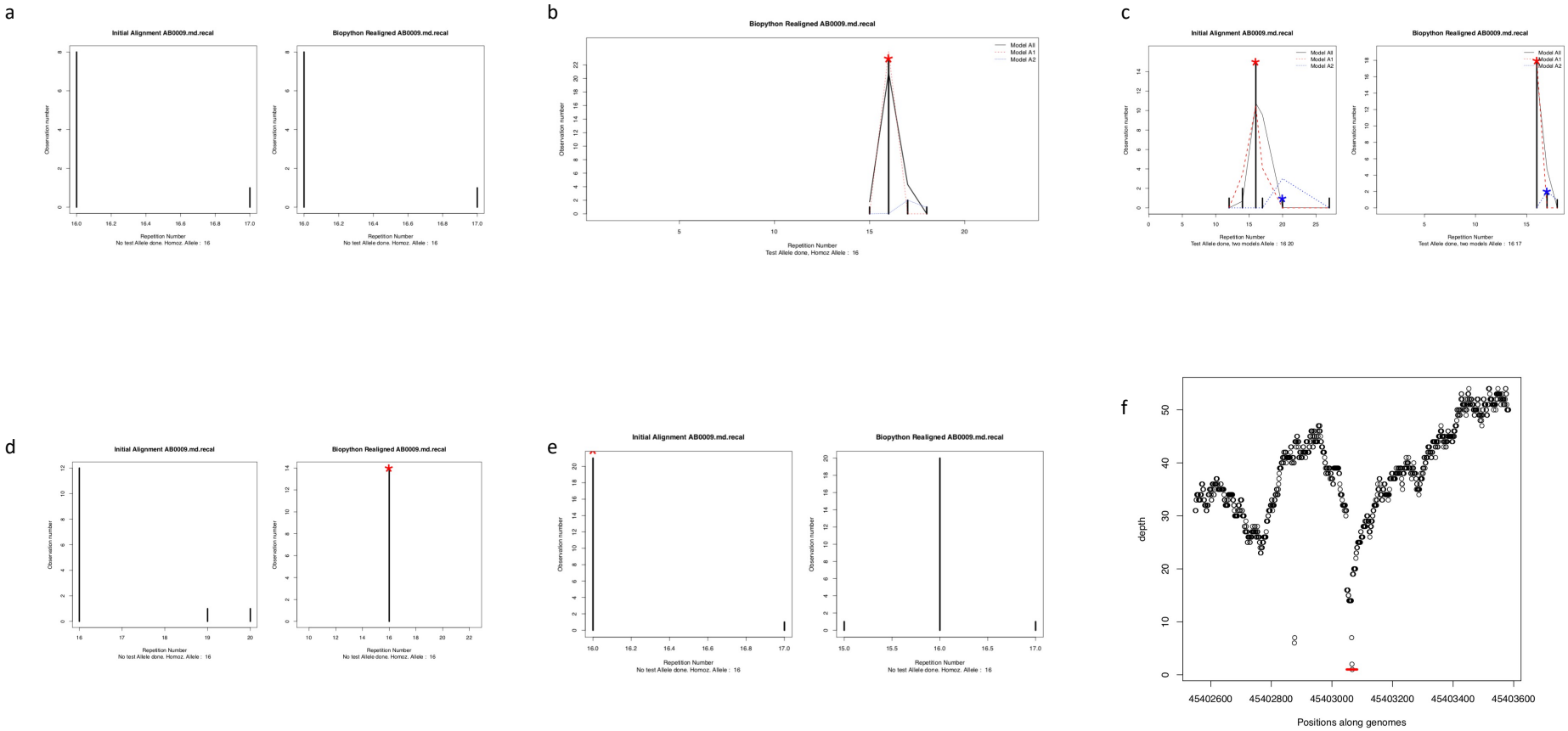


Figure A4.6: Inhouse STR-caller results for AB0009 (homozygous T=16, S/S)

All results religned using Biopython (a) shows results for local alignment using bowtie2 (b) shows combined results for global and local alignments using bowtie2 and bwa (c) shows results for local alignment using bwa (d) shows results for global alignment using bowtie2 (e) shows results for local alignment using bowtie2 (f) shows read depth across the genome where the red line marks the position of TOMM40 poly-T.

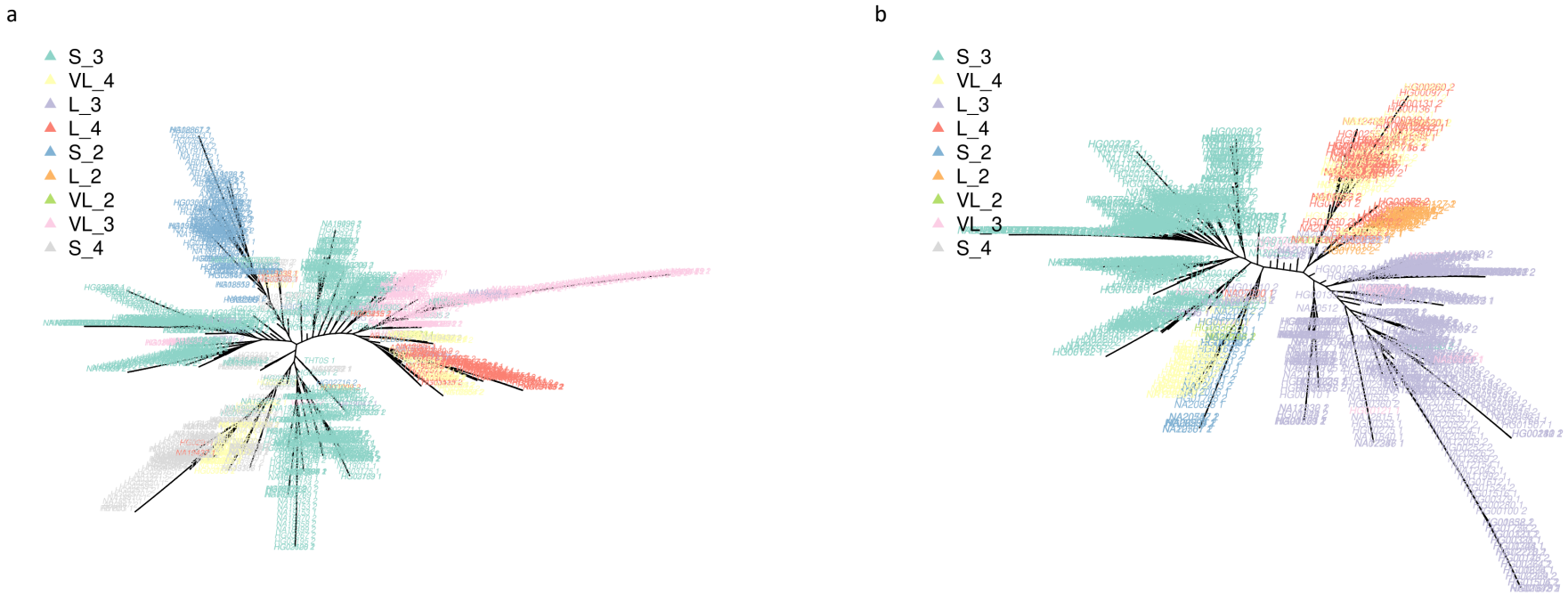


Figure A4.7: Unrooted evolutionary tree diagrams

This data and diagram will be used in a future publication once the evolutionary aspects of TOMM40 have been adequately teased out. Unrooted evolutionary tree diagrams of TOMM40-APOE haplotypes coded as per the legends. (a) represents African data and (b) European data

A4.3 Diagnostics of regression

Diagnostics of regression performed for first regression model assessing the effects of *TOMM40* genotype on cognitive function.

A4.3.1 Total cognition score

No outliers for total cognition score as none of the Studentized residuals had Bonferroni p-values < 0.05. Cook's distance was <1 signifying no identifiable influencers in the dataset. Residuals were normally distributed, and variance was not constant (χ^2= 6.12, df=1, p= 0.013). The spread level plot suggests a transformation to the power of 1.37. Variance inflation factors were below 5, and there was no autocorrelation of variables (Durbin-Watson p= 0.544).

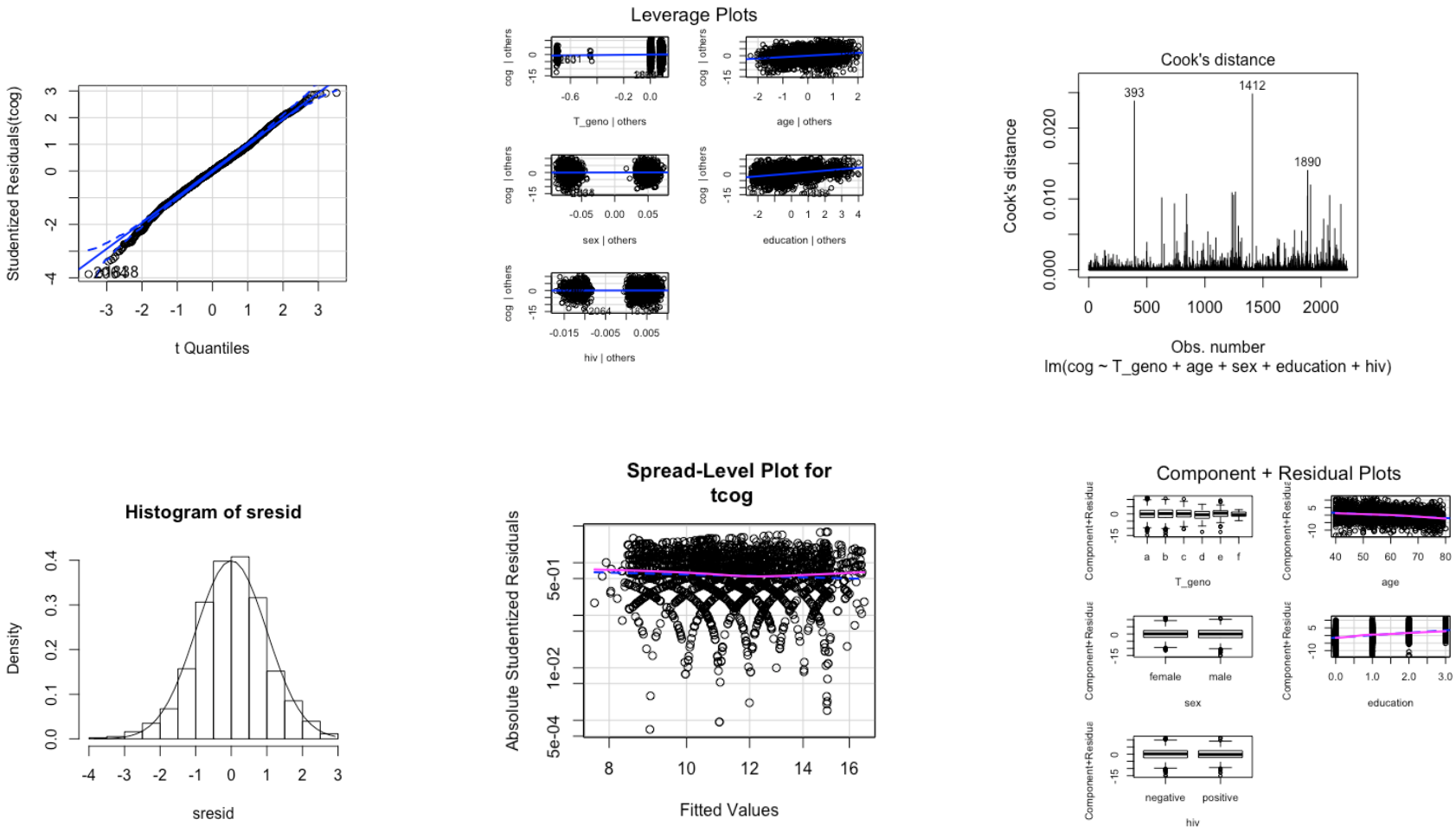


Figure A.4.8: Diagnostic regression plots for the residuals of total cognition score

In order: QQ-plot, Leverage plots for each regression term, Cook's plot indicating Cook's distance, histogram showing distribution of residuals, spread level plot where negative values were removed, and component-residual plots for each term and its effect on total cognition score.

A4.3.2 Executive function

No outliers for executive function as none of the Studentized residuals had Bonferroni p-values < 0.05. Cook's distance was <1 signifying no identifiable influencers in the dataset. Residuals were normally distributed, and variance was constant (χ^2= 2.25, df=1, p= 0.133). The spread level plot suggests a transformation to the power of 1.01 after removing the negative values. Variance inflation factors were below 5, and there was autocorrelation of variables (Durbin-Watson p= 0).

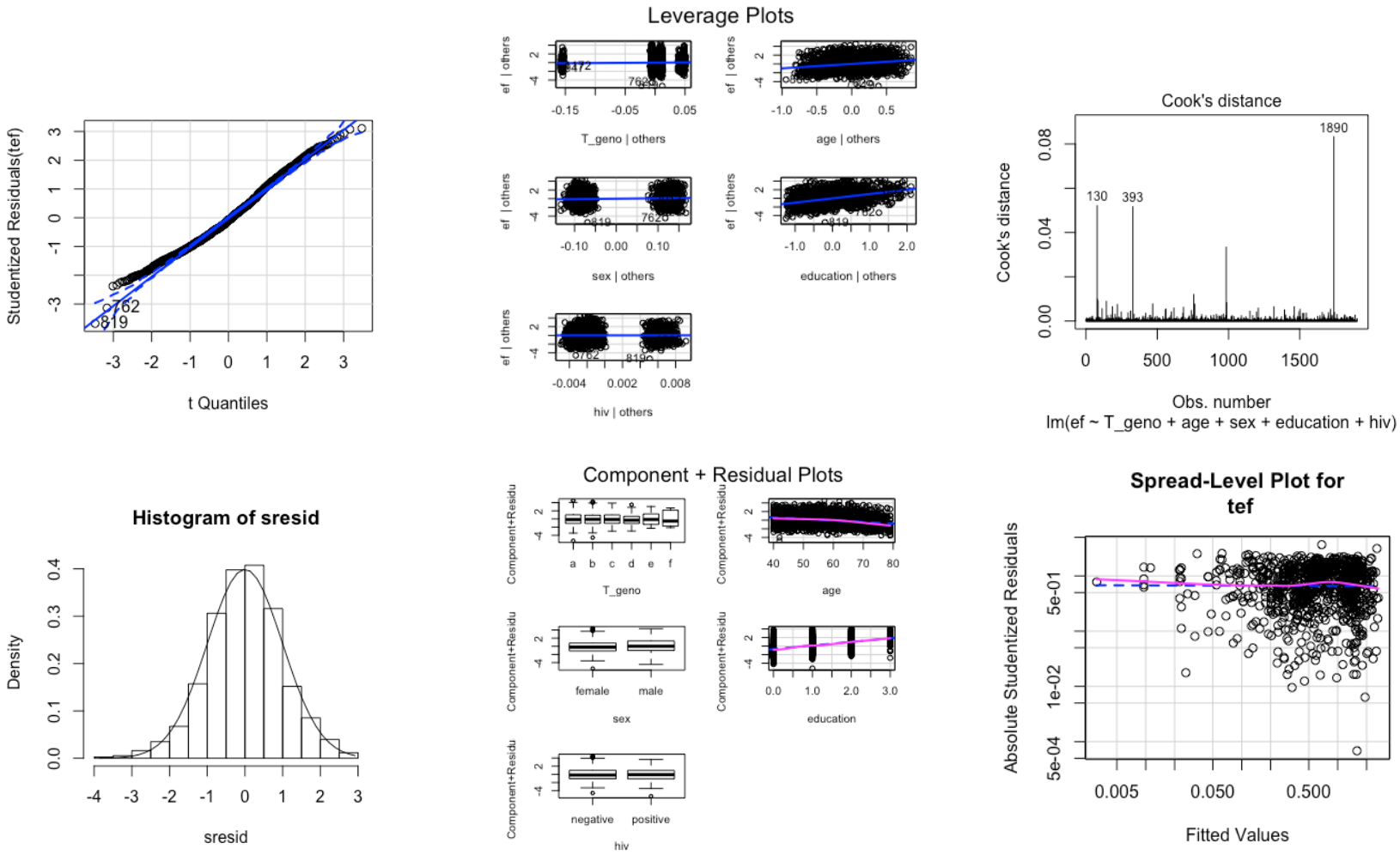


Figure A.4.9: Diagnostic regression plots for the residuals of executive function

In order: QQ-plot, Leverage plots for each regression term, Cook's plot indicating Cook's distance, histogram showing distribution of residuals, spread level plot where negative values were removed, and component-residual plots for each term and its effect on executive function.

A4.3.3 Language

There was one outlier for language (Studentized residuals had Bonferroni p=0.0004). Cook's distance was <1 signifying no identifiable influencers in the dataset. Residuals were normally distributed, and variance was not constant (χ^2=84.8, df=1, p<2.22e-16). The spread level plot suggests a transformation to the power of 1.15 after removing the negative values. Variance inflation factors were below 5, and there was no autocorrelation of variables (Durbin-Watson p= 0.162).

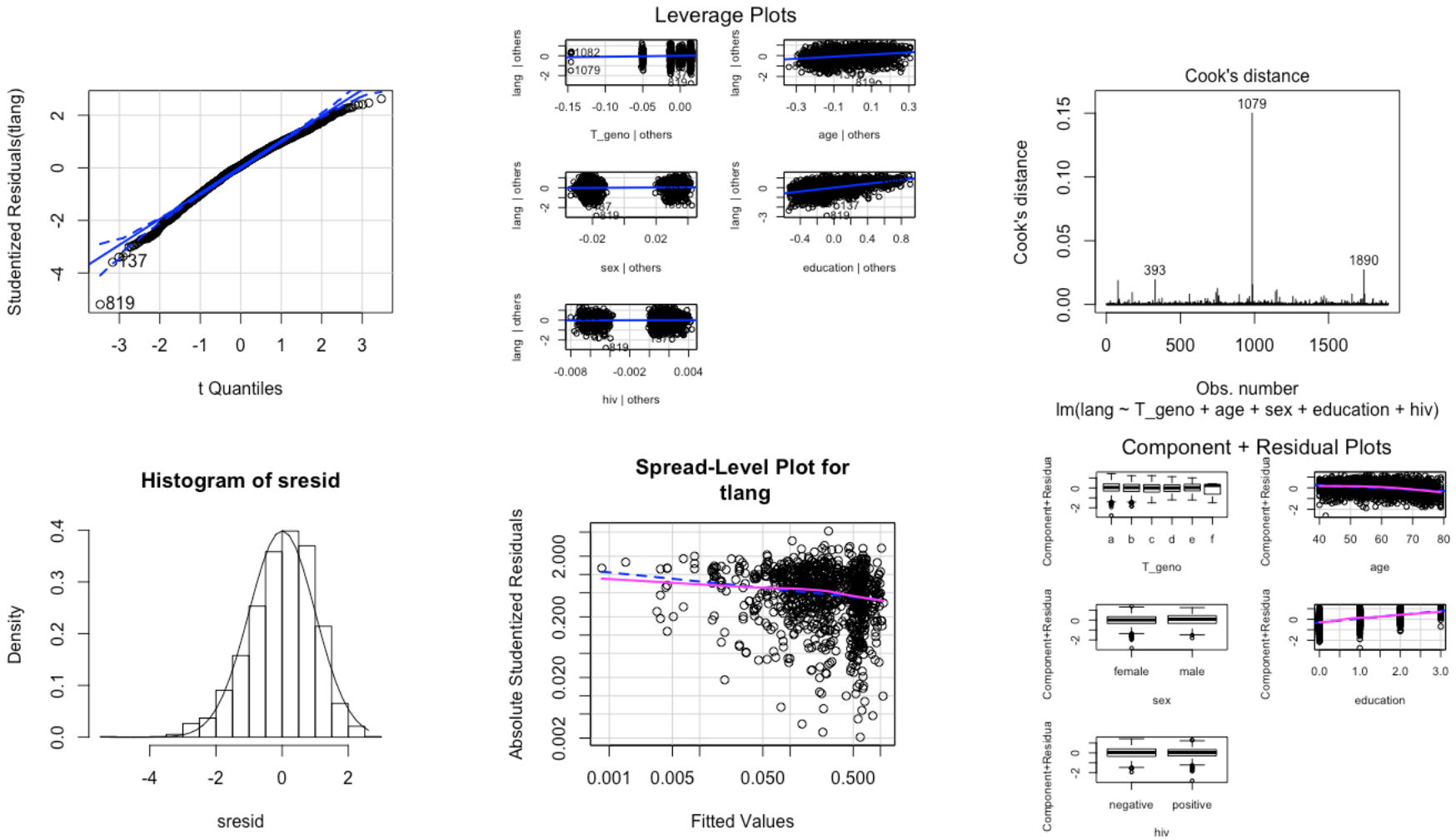


Figure A.4.10: Diagnostic regression plots for the residuals of language

In order: QQ-plot, Leverage plots for each regression term, Cook's plot indicating Cook's distance, histogram showing distribution of residuals, spread level plot where negative values were removed, and component-residual plots for each term and its effect on language.

A4.3.4 Verbal episodic memory

There was one outlier for memory (Studentized residuals had Bonferroni p=0.002). Cook's distance was <1 signifying no identifiable influencers in the dataset. Residuals were normally distributed, and variance was not constant (χ^2= 174.6, df=1, p<2.22e-16). The spread level plot suggests a transformation to the power of 1.18 after removing the negative values. Variance inflation factors were below 5, and there was autocorrelation of variables (Durbin-Watson p=0.002).

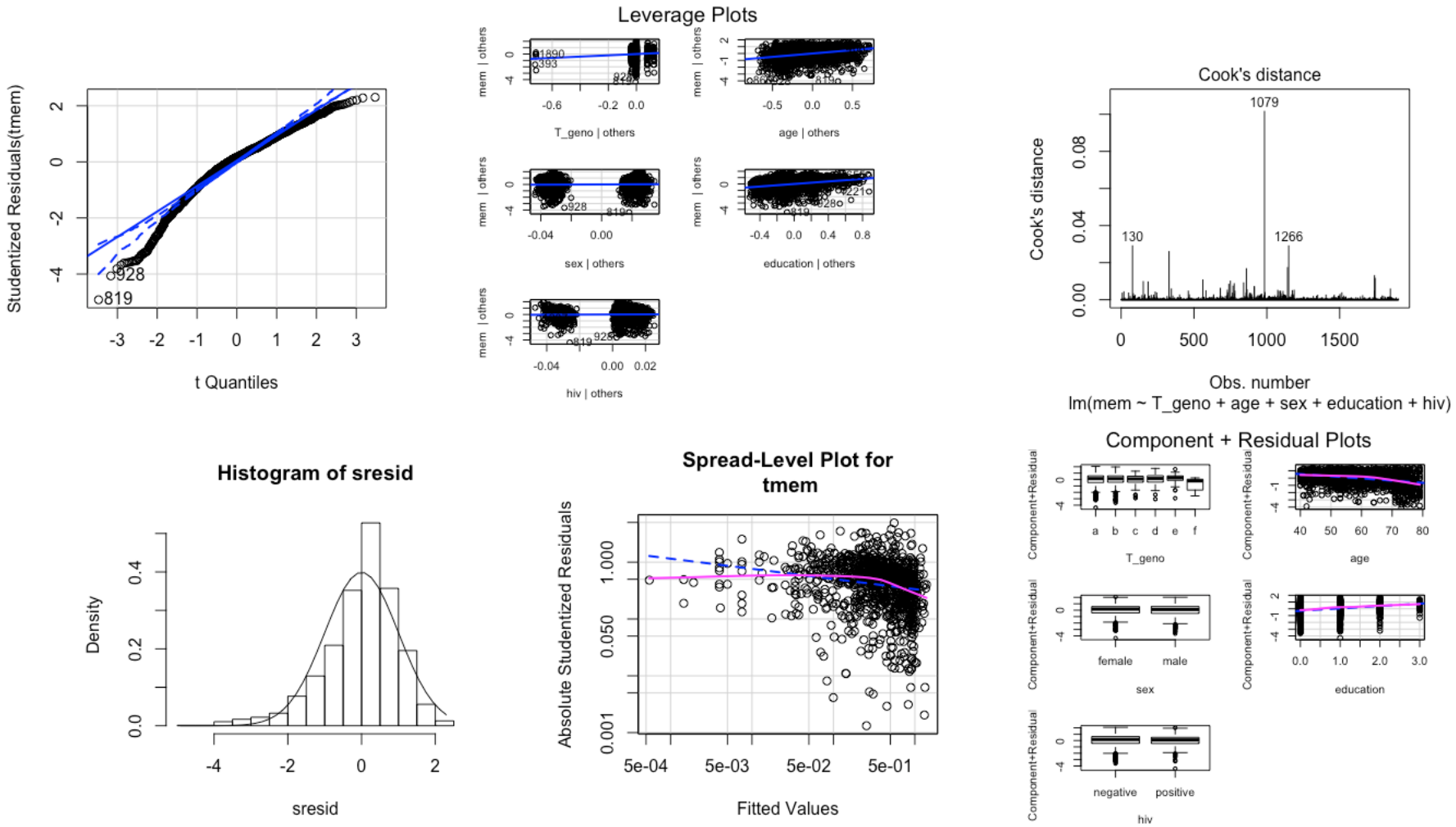


Figure A.4.11: Diagnostic regression plots for the residuals of episodic memory

In order: QQ-plot, Leverage plots for each regression term, Cook's plot indicating Cook's distance, histogram showing distribution of residuals, spread level plot where negative values were removed, and component-residual plots for each term and its effect on episodic memory.

A4.3.5 Visuospatial ability

There were two outliers for visuospatial cognition with Studentized residual p-values of 3.798e-05 and 4.379e-02. Cook's distance was <1 signifying no identifiable influencers in the dataset. Residuals were normally distributed, and variance was not constant (χ^2=66.6, df=1, p<3.311e-16). The spread level plot suggests a transformation to the power of 1.19 after removing the negative values. Variance inflation factors were below 5, and there was some autocorrelation observed (Durbin-Watson p=0.006).

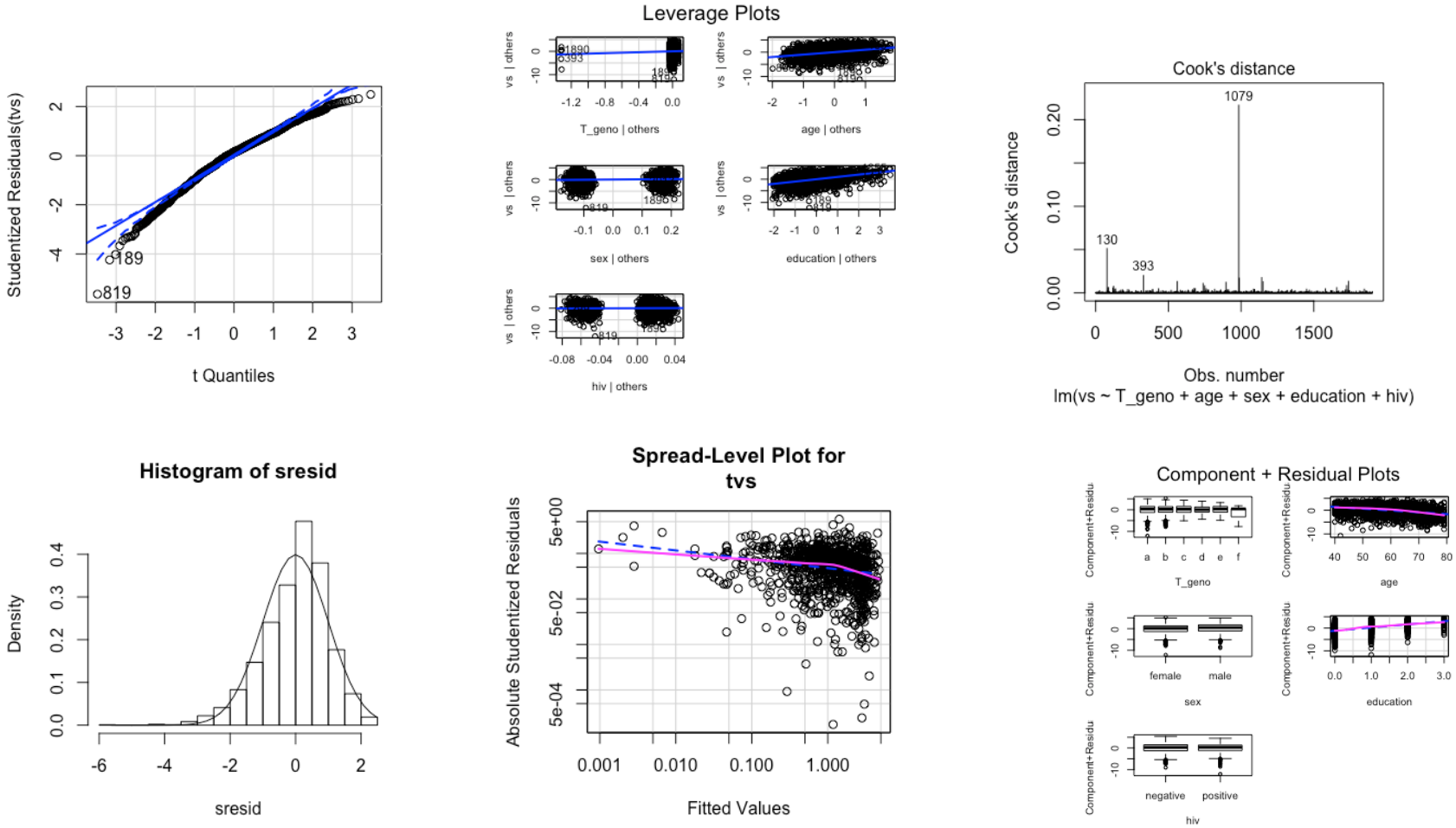


Figure A.4.11: Diagnostic regression plots for the residuals of visuospatial ability

In order: QQ-plot, Leverage plots for each regression term, Cook's plot indicating Cook's distance, histogram showing distribution of residuals, spread level plot where negative values were removed, and component-residual plots for each term and its effect on visuospatial cognition