Dissertation

GENETIC DETERMINANTS OF BRAIN AGEING

submitted by

Piyush Gajananrao GAMPAWAR

for the Academic Degree of

Doctor of Philosophy (PhD)

at the

Medical University of Graz

Research Unit-Genetic Epidemiology
Gottfried Schatz Research Centre for Cell Signalling, Metabolism and Aging
Molecular Biology and Biochemistry

under the Supervision of

Univ. Prof. Dr.med.univ. Dr.phil. Helena Schmidt

2020

Declaration

I hereby declare that this thesis is my own original work and that I have fully acknowledged by name all of those individuals and organisations that have contributed to the research for this thesis. Due acknowledgement has been made in the text to all other material used. Throughout this thesis and in all related publications I followed the "Guidelines of the Medical University of Graz on Good Scientific Practice".

Piyush Gajananrao Gampawar

Date: 9th September 2020
Graz, Austria

Disclosures

The part of this dissertation has been published in following two articles

Gampawar P, Schmidt R and Schmidt H (2020) Leukocyte Telomere Length Is Related to Brain Parenchymal Fraction and Attention/Speed in the Elderly: Results of the Austrian Stroke Prevention Study. *Front. Psychiatry* 11:100. doi: 10.3389/fpsyt.2020.0010 [1]

Gampawar P, Saba Y, Werner U, Schmidt R, Müller-Myhsok B and Schmidt H (2019) Evaluation of the Performance of AmpliSeq and SureSelect Exome Sequencing Libraries for Ion Proton. *Front. Genet.* 10:856. doi: 10.3389/fgene.2019.00856 [2]

The following co-authors contributed to my first-author publications:

Yasaman Saba[1], Ulrike Werner[1], Reinhold Schmidt[2], Bertram Müller-Myhsok[3], Helena Schmidt[1]

[1]Research Unit-Genetic Epidemiology, Gottfried Schatz Research Centre for Cell Signaling, Metabolism and Aging, Molecular Biology and Biochemistry, Medical University Graz, Graz Austria

[2]Department of Neurology, Clinical Division of Neurogeriatrics, Medical University Graz, Graz, Austria

[3]Max Planck Institute of Psychiatry, Munich, Germany

I confirm that all co-authors have agreed to use their data in my thesis. I have permission from the publisher to reproduce figures. Both papers (Gampawar et al (2020) and Gampawar et al (2019)) are open-access articles distributed under the terms of the **Creative Commons Attribution License (CC BY)**. The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice.

I contributed to following publication during my PhD studies

Tripolszki K, **Gampawar P**, Schmidt H, Nagy Z.F, Nagy D, Klivenyi P, Engelhardt JI, and Széll, M. (2019) Comprehensive genetic analysis of a Hungarian amyotrophic lateral sclerosis cohort. *Front. Genet*, 10:732

Acknowledgements

As a PhD student I was funded by the Medical University of Graz through the PhD Program "Molecular Medicine"

The research reported in this dissertation was funded by

- The Austrian Science Fund grant number P13180 and P20545-B05
- The Austrian National Bank Anniversary Fund, P15435
- The Austrian Ministry of Science under the aegis of the EU Joint Programme—Neurodegenerative Disease Research—www.jpnd.eu. The project is supported through the following funding organisations under the aegis of EU Joint Programme—Neurodegenerative Disease Research—www.jpnd.eu: Australia, National Health and Medical Research Council, Austria, Federal Ministry of Science, Research and Economy; Canada, Canadian Institutes of Health Research; France, French National Research Agency; Germany, Federal Ministry of Education and Research; Netherlands, The Netherlands Organisation for Health Research and Development; United Kingdom, Medical Research Council. This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 643417.
- The Franz Lanyar Stiftung from the Medical University of Graz, Austria
- The Austrian Atherosclerosis Society.

Foreword

I am grateful to all people who helped and supervised me during this journey of my PhD.
First and foremost, I would like to express my gratitude towards my supervisor Prof. Helena Schmidt for bestowing her trust in me for this excellent and exciting research opportunity and guiding me through all the hurdles along the path. Without her continuous supervision and patience, I would not have been able to produce quality results.
I want to convey my regards to my thesis committee members Dr Reinhold Schmidt and Dr Bertram Mueller, for all guidance and support. A special thanks to Reinhold for providing the neurological data to work with, guidance to understand and analyse it.
I am thankful to all my past and current lab colleagues with whom working was a pleasant experience, especially Hans and Irmi, who taught me to work with new machines and protocols. Not being a German speaker at least in the initial days is quite stressful, but both Hans and Irmi were there to help me with all official works in German.
I am grateful to all my endless amazing friends here in Graz, back home in India and quite a few places around the world, considering the space writing your names will be impossible for me. You guys are amazing, and my life will not be so much fun without you.

Finally, words can never justify how indebted I am with my family, who always stood firmly behind me throughout my life. A farfetched dream that I had as a child to study genetics could not have been possible without my parent's constant support and motivation.

Table of Contents

Abbreviations

AD	Alzheimer's Disease
ApoE	Apolipoprotein
ASPS	Austrian Stroke Prevention Study
ASPS-Fam	Austrian Stroke Prevention Family Study
AxD	Axial Diffusivity
BAM	Binary Alignment
BED	Browser Extensible Data
BMI	Body Mass Index
BPF	Brain Parenchymal Fraction
BWA	Burrows-Wheeler Aligner
Chr	Chromosome
CVD	Cardiovascular Diseases
DHS	Dallas Heart Study
DM	Diabetes Mellitus
DNA	Deoxy Ribonucleic Acid
DTI	Diffusion Tensor Imaging
ECG	Electrocardiogram
ETR	Effective target region
EU	European Union
FA	Fractional Anisotropy
FLAIR	Fluid-Attenuated Inversion Recovery
FNs	False Negatives
FPs	False Positives
GATK	Genome Analysis Toolkit
GSHA	Graz Study Health and Ageing
GWAS	Genome With Association Studies
HCR	High Confidence Region
HDL	High-Density Lipoprotein
IGV	Integrated Genomic Viewer
LDL	Low-Density Lipoprotein
LTL	Leukocyte Telomere Length
MB	Million Bases
MD	Mean Diffusivity
MRI	Magnetic Resonance Imaging
NGS	Next Generation Sequencing
OTR	Overlapping Target Region
PCR	Polymerase Chain Reaction
pM	Pico mole

PPV	Positive Predictive Value
PSMD	Peak Width of Skeletonized Mean Diffusivity
qPCR	Quantitative Polymerase Chain Reaction
RD	Radial Diffusivity
RFLP	Restriction Fragment Length Polymorphism
RT-PCR	Real-Time Polymerase Chain Reaction
S	Single copy gene
SNP	Single Nucleotide Polymorphism
T	Telomere
TMAP	Torrent Mapping Alignment Program
TPs	True Positives
TTR	Total Target Region
TVC	Torrent Variant Caller
UCSC	University of California—Santa Cruz
UK	United Kingdom
VCF	Variant Calling Format
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
WMH	White Matter Hyperintensities
µl	Micro litre

Abstract

The general aim of this thesis was to contribute to a better understanding of brain ageing using genetic epidemiological approaches. We specifically focused on 2 aims 1) exploring the role of leukocyte telomere length (LTL) a marker of cellular senescence in structural and functional alteration of the brain during ageing and 2) establishing robust high validity protocols for whole exome/genome sequencing (WES/WGS) and genomic databases based on WES/WGS data generated in our cohorts. These accomplishments form the basis for further studies dissecting the pathomechanism of brain ageing. The two aims are presented in Part I and Part II. In detail, in Part I we first explored the association between LTL and global MRI markers of brain ageing such as brain parenchymal fraction (BPF) and white matter hyperintensities (WMH) as well as cognitive functions within the Austrian Stroke Prevention Study (ASPS) (N=909; age=65.9±8.0; female=57.3%). MRIs was done on a 1.5T scanner and LTL measurements using RT-PCR. Longer LTL was significantly associated with larger BPF (ß=0.43, $p < 0.001$), larger WMH (ß=0.03, p=0.04) and better performance in the attention/speed domain. The effect was confined to those overweight (BMI ≥25, ß=0.04, p=0.05) and with lower education (≤10 years, ß=0.04, p=0.05). Importantly, the beneficial effect of longer LTL on attention/speed was partly mediated by BPF in both subgroups (ß= 0.02, 95%CI=0.01-0.03). Our results support a strong protective role of longer LTL on the brain, especially in the presence of risk factors. In order to follow up on these findings, we next initiated a collaboration within the JPND BRIDGET consortium. In the frame of this initiative, we conducted a pilot study on 90 participants of the Graz Study Health & Aging (GSHA) (age=67.6±9.0; female=63.3%). Our pilot study provided a proof of principle that LTL extracted from WGS data using the TelSeq software can be used to generate and combine LTL data from the various cohorts within BRIDGET. The use of the combined data provides several advantages 1) increases the sample size to >2000, 2) assessing the effect of LTL at the microstructural level of the brain, and 3) investigating the effect of LTL on the brain over a much wider age range (20->90ys). In Part II of this thesis, we present the evaluation of 2 library preparation methods for WES, namely AmpliSeq and SureSelect and an improved variant calling protocol for Ion Proton. We used 12 in-house DNAs and the NA12878 reference DNA. We found comparable sensitivity (93%) but a higher positive predictive value (PPV) for AmpliSeq than SureSelect (84%vs80%). Our improved protocol substantially reduced false positives by 90% and increased PPV to 97%. As WGS became available as part of the BRIDGET collaboration, we selected in total 150 ASPS

family and GSHA participants to be sequenced on Illumina HiSeq at the McGill Genomic Centre. The joint variant calling and quality control have been completed, and a study is ongoing to identify rare genetic variants associated with structural markers of brain ageing.

In summary, we show that longer LTL is associated with larger brain size, which in turn transfers to better cognitive performance in the attention/speed domain. Importantly, longer LTL is especially beneficial in the presence of risk factors such as overweight and low education. In order to follow-up on these findings, we initiated a collaborative study using LTL extracted from WGS in the frame of the BRIDGET consortium. Further on we established several resources including high validity NGS protocol on Ion Proton and comprehensive databases of genomic variants derived from WES, WGS in our cohorts.

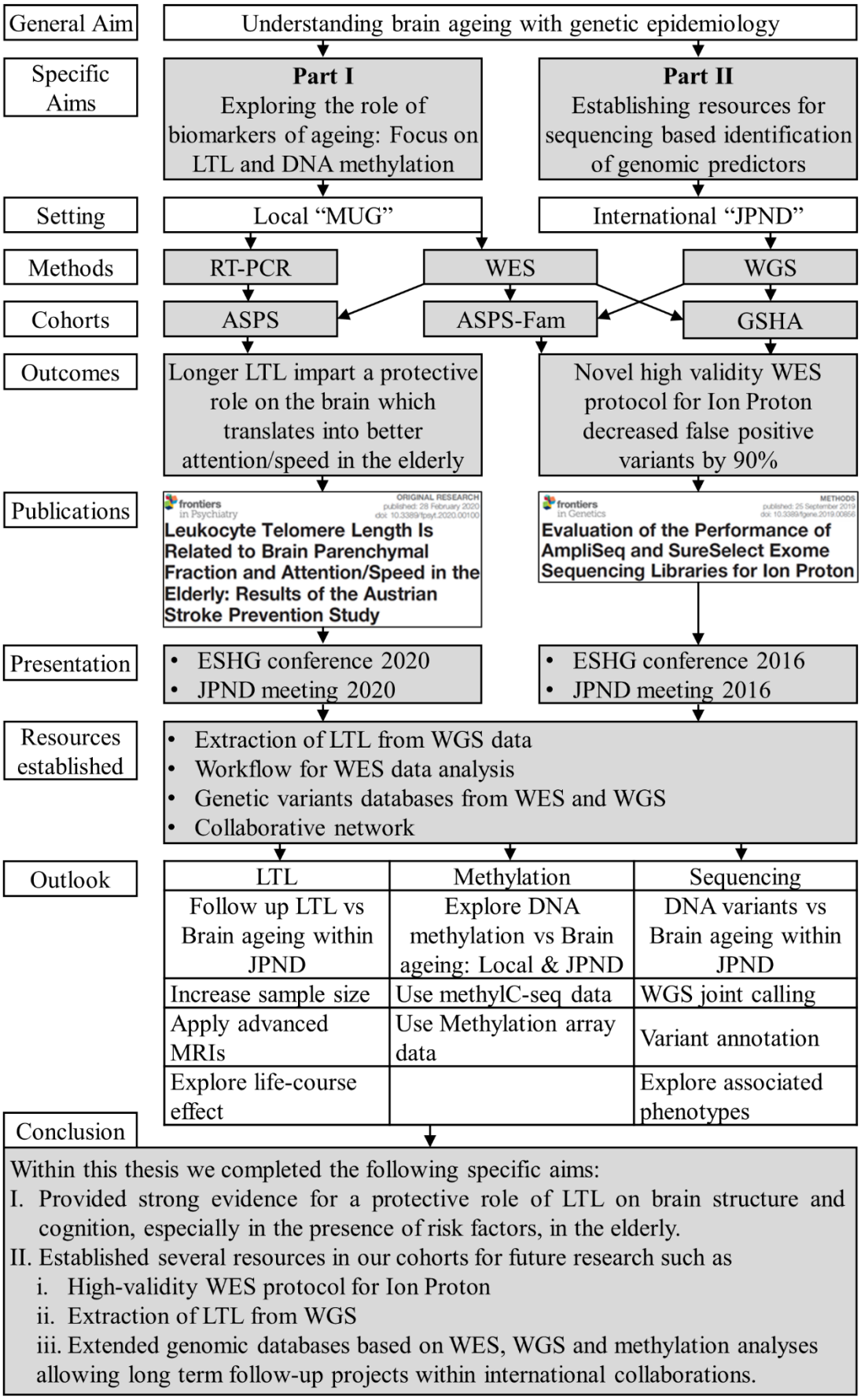
Zusammenfassung

Das allgemeine Ziel dieser Doktorarbeit war, zu einem besseren Verständnis der Alterung des Gehirns mittels genetischer epidemiologischer Studien beizutragen. Wir haben uns dabei auf 2 Themen fokussiert 1) die Erforschung der Rolle von Leukozyten Telomerlänge (LTL), ein Marker der zellulären Alterung bei altersbedingten strukturellen und funktionellen Veränderungen des Gehirns und 2) die Entwicklung eines zuverlässigen Protokolls mit hoher Validität zur whole exome/genome Sequenzierung (WES/WGS) und das Erstellen von auf WES/GWS basierten genomischen Datenbanken in unseren Kohorten. Die Resultate aus diesen Projekten bilden die Basis für zukünftige Studien um den Pathomechanismus der Gehirnalterung weiter zu erforschen. Die 2 Schwerpunkte sind in Teil I und Teil II präsentiert. Im Detail untersuchen wir in Teil I die Assoziation zwischen LTL und globalen MRI Markern der Gehirnalterung so wie Gehirn Parenchyma Fraktion (BPF), Marklager Hyperintensitäten (WMH) und kognitive Funktionen im Rahmen der Austrian Stroke Prevention Study (ASPS). MRI wurde auf 1,5T Scanner durchgeführt, LTL mittels rtPCR gemessen. Längere LTL war signifikant assoziiert mit grösserer BPF, und WMH und mit besseren kognitiven Leistungen in der Aufmerksamkeit/ Schnelligkeit Domäne. Der Effekt war besonders stark in der Gruppe mit Übergewicht (BMI>=25) bzw. niedrigeren Schulbildung (<10J). Der günstige Effekt auf Aufmerksamkeit/ Schnelligkeit wurde in beiden Subgruppen zum Teil durch den Effekt auf BPF meditiert. Unsere Resultate weisen auf einen starken Schutzeffekt von langen Telomären auf das Gehirn, insbesondere bei Vorhandensein von Risikofaktoren, hin. Zwecks Nachverfolgung dieser Befunde haben wir eine Zusammenarbeit im Rahmen des JPND BRIDGET Konsortiums initiiert. Im Rahmen dieser Initiative haben wir eine Pilotstudie mit 90 TeilnehmerInnen der Graz Study on Health & Aging durchgeführt. Unsere Pilotstudie hatte das „proof of principles“ dafür geliefert, dass LTL, welche aus den WGS Daten mittels TELSEQ Software extrahiert werden, geeignet sind, LTL Daten in den an BRIDGET teilnehmenden Kohorten zu generieren und zusammenzufügen. Die Verwendung der so entstandenen zentralen Datenbank bietet zahlreiche Vorteile für Follow-up Studien unserer Resultate: 1) Vergrösserung der Stichprobe >2000 TeilnehmerInnen, 2) Beurteilung des LTL Effektes auf Mikrostrukturen des Gehirns und 3) Untersuchung des LTL Effektes über einen breiten Altersbereich von 20J bis zu >90J.

Im Teil II der Doktorarbeit präsentieren wir die umfassende Evaluation der 2 WES Methoden, AmpliSeq und SureSelekt, und die Entwicklung eines neuen Variant Calling Protokolls auf dem Ion Proton Sequenzierer. Wir haben 12 hausinterne bzw eine kommerzielle, NA12878, Referenz DNA als Proben verwendet. Wir haben eine vergleichbare Sensitivität beider Methoden aber einen signifikant höheren positiven prädiktiven Wert (PPV) für AmpliSeq, verglichen mit SureSelekt, gefunden. Unser verbessertes Protokoll hat zudem die Anzahl der falsch positiven Varianten signifikant um 90% reduziert und gleichzeitig den PPV auf 97% erhöht. Nachdem WGS innerhalb des BRIDGET Konsortium angeboten worden ist, haben wir 150 TeilnehmerInnen der ASPS-Fam und GSHA Studie für Sequenzierung an Illumina HiSeq, durchgeführt am McGill Genomic Centre, ausgewählt. Der gemeinsame Variant Calling und die zentrale Qualitätskontrolle der Daten wurde bereits abgeschlossen, und die ersten Studien zur Identifizierung von seltenen genetischen Varianten in Bezug auf MRI Marker der Gehirnalterung wurden gestartet.

Zusammenfassend zeigen wir, dass die LTL-Länge mit der Gehirngröße und diese wiederum mit besseren kognitiven Leistungen (Aufmerksamkeit/ Schnelligkeit) positiv assoziiert ist. Wichtigerweise ist längere LTL besonders schützend bei Vorhandensein von Risikofaktoren wie Übergewicht und niedrigerer Schulbildung. Zwecks Follow-up dieser Befunde haben wir eine kooperative Studie im Rahmen des BRIDGET Konsortiums unter Verwendung von aus WGS extrahierten LTL initiiert. Darüber hinaus, haben wir zahlreiche wichtige Ressourcen für zukünftige Projekte etabliert, darunter ein verbessertes NGS Protokoll mit hoher Validität entwickelt und eine umfangreiche auf WES/WGS basierte genomische Datenbank in unseren Kohorten angesetzt.

Thesis outline



Grey boxes show the work completed under this thesis.

1 Introduction

1.1 Brain ageing

Ageing, in essence, construed as a time-dependent progressive deterioration of the physiological and functional integrity of an organism, which increases the risk of age-related diseases such as neurodegenerative diseases, chronic kidney diseases, coronary heart diseases, stroke, type II diabetes, osteoarthritis and common cancers and ultimately mortality. Ageing at the molecular and cellular levels [3,4] is manifested by genomic instability, telomere attrition, epigenetic alterations, loss of proteostasis, deregulated nutrient sensing, mitochondrial dysfunction, cellular senescence, stem cell exhaustion, and altered intercellular communications. These processes were also shown to accelerate ageing in experimental models[3].

1.1.1 Epidemiology and clinical relevance

The last century observed an unprecedented increase in human life expectancy from approximately 50 years in early 1900 to over 80 years presently due to better medical care and living conditions [5]. In the age group of 60 years and more, there is an exponential increase in the incidence and prevalence of cardiovascular diseases, cancers, dementia and neurodegenerative diseases [5]. The rate of ageing and the vulnerability of the different organ systems to ageing are highly variable at the population level resulting in substantial differences in the age at onset and the clinical manifestations of age-related diseases [6].

According to the 2019 Alzheimer Europe Yearbook "Estimating the prevalence of dementia in Europe" [7] the prevalence of dementia increases from 0.6% in the age range of 60-64 years to 40.8% in age >90 years. In total, 1.73% of the population in EU countries is suffering from dementia with higher prevalence in women (0.9-44.8%) than in men (0.2 -29.7%). Worldwide, 47 million people live with dementia, and the expected number by 2050 is 131 million[8]. Dementia takes years to decades to develop, and we poorly understand the processes taking place in the preclinical phase. Age is the strongest risk factor for dementia, and clinically silent pathology in the brain is often seen in the ageing population. The most common form of dementia is Alzheimer’s dementia (60-80%), followed by vascular dementia (10-25%) and Lewy body dementia (7-25%). Mixed forms with vascular and degenerative components are common [9]. It is estimated that 25-30% of people who die by the age of 75 years, already have a substantial amount of cerebral lesions typical of Alzheimer's disease (AD), though clinically

they had normal cognitive function [10]. Moreover, Bennet et al., (2006) have reported that 45% of people without dementia has AD pathology as detected by brain autopsy [11]. The pathological processes causing dementia and AD begin 10-20 years before the actual manifestation of the disease and clinical diagnosis, causing hurdles in efficient preventive strategies [10]. Therefore, it is essential to understand the normal ageing processes taking place in the brain and when and how these processes change its trajectories towards pathological degeneration and ultimately, clinical disease [12].

1.1.2 Assessing brain ageing in epidemiological studies

1.1.2.1 MRI markers of brain ageing

During ageing, the brain undergoes typical functional changes manifesting as cognitive decline, particularly affecting domains of processing speed, memory, reasoning and executive functions [13]. At the macrostructural level, a decrease in grey and white matter volume, cortical thinning, ventricular enlargement and a decrease in weight of the brain is evident in ageing [13,14]. In epidemiological studies, MRI is used to detect and assess the extent of age-related macro- and microstructural changes as well as vascular lesions in the brain. These MRI correlates of brain ageing are referred as MRI markers of brain ageing. Most frequently used such markers are intracranial volume, brain parenchymal fraction (BPF), cortical thickness/surface area/volume indicating neurodegenerative processes as well as white matter hyperintensities (WMH), covert brain infarcts, cerebral microbleeds and extended perivascular spaces related to vascular processes [15]. Recently, also automatically generated regional measures of the brain such as lobular volumes, hippocampal volume, and even voxel-based signal intensities became available, allowing for a better spatial resolution. These MRI markers are powerful predictors of clinically manifest dementia and stroke in the general population [16,17].

A large number of epidemiological studies have been exploring trajectories of these MRI markers over the years in order to identify their predictors, clinical consequences, the underlying mechanisms and pathways. Recently, the longitudinal analysis of more than 12,000 MRI scans from the Rotterdam Study found that different global, cortical, subcortical and lobar MRI markers show a nonlinear course which accelerates with advancing age and seen earlier in men than in women [12]. A meta-analysis of 94 studies reported that MRI markers of covert vascular brain injury such as WMH burden, brain infarcts and cerebral microbleeds were associated with a higher risk of stroke, intracerebral haemorrhage, dementia, Alzheimer's disease and death [15] in the elderly. Higher global, as well as regional brain volumes, were shown

to be associated with decreased risk of mortality independent of the presence of vascular risk factors and co-morbidities in the Icelandic population [18]. By using more advanced MRI technologies such as diffusion tensor imaging (DTI) investigating the integrity of the brain at the microstructure level became possible. Such recently emerging markers such as lower fractional anisotropy (FA), higher mean and radial diffusivity (MD & RD) throughout the white matter, as well as the tracts passing through temporal and posterior brain regions, were able to distinguish between AD patients and cognitively intact elderly controls [19].

1.1.2.2 Cognitive ageing

There is an age-related gradual decline in overall cognitive abilities generally termed as "cognitive ageing". Cognitive ageing is observed as a global cognitive decline or a decline in specific cognitive domains such as attention/speed, conceptualisation, memory and visuospatial reasoning. Evaluation of cognitive functions is done by using different cognitive test batteries. Cognitive performance over the different domains is highly intercorrelated. Global cognition, often termed as g-factor, is the first unrotated principal component that represents the positive correlation between different cognitive tests. The g-factor accounts for approximately 40%-50% of the inter-person variations for a given cognitive test [20].

A gradual age-associated decline in general cognition, fine motor skills, processing speed and visuospatial ability was reported in the cross-sectional setting of Rotterdam study [21]. In middle-age women, a 4.9% and 2% decline was observed in the cognitive domains of processing speed and memory respectively from the mean base score over a 10-year of observation time [22]. The cognitive decline during late life (6-15 years before death) is gradual and domain-specific. Nonetheless, it accelerates rapidly and gets global during the final 2-3 years before death [23]. A multi-ethnic meta-analysis of 14 longitudinal studies concluded that there is an age-related gradual decline in memory, processing speed, language, and executive functioning test scores in the elderly and a further accentuation in final stages of life [24]. Importantly, the rate of age-related cognitive decline varies with ethnicity, sex and presence of risk factors [24].

1.1.2.3 Association of MRI markers with cognitive ageing

The association between MRI markers of brain ageing and cognition has been extensively studied. Progression of WMH was associated with a decline in memory, conceptualisation and visuopractical skills [25]. A higher load of small vessel disease on MRI (WMH, microbleeds, lacunes, and perivascular spaces) were associated with lower general cognition (g-factor) [26]. Longitudinal studies on older adults showed that reduction of total cerebral, hippocampal and

grey matter volume was associated with faster decline in global cognition; an increase of cerebral and white matter atrophy with a decrease in verbal memory; while hippocampal atrophy and ventricular expansion were associated with a decline in verbal memory and executive functions [27,28]. On DTI measures, better hippocampal mean diffusivity, white matter integrity and brain peak width of skeletonised mean diffusivity (PSMD) were related to better verbal and working memory [29], reasoning, cognitive flexibility and processing speed [30,31]. Understanding of these MRI markers, cognitive decline as well as their relationship with each other during normal brain ageing is vital in order to perceive the path when normal brain ageing diverts to pathological changes [12].

1.1.3 Heritability of MRI markers and cognitive traits

Heritability is defined as the proportion of phenotypic variance explained by genetic variance. In population genetics, the observed phenotypic variation of a trait is due to unobserved genetic and environmental factors. Therefore, heritability estimates of a trait in a population give an idea about variation seen among individuals that are attributable to differences in their genetic make-up [32]. Heritability estimates of a trait can differ across ages, calendar times and populations due to change in the genetic variance, environmental variance and correlation between genetic and environmental variance [32].

1.1.3.1 Family-based heritability

Classically heritability is studied in twins and family studies. Twin studies provided a design where genetic and environmental effects could be estimated. These studies carried out based on the assumption that monozygotic twins share both additive (due to alleles) and non-additive (due to allelic interactions) genetic factors. In contrast, dizygotic twins share only on an average 50% of their genes. Considering the effect of the environment is constant, genetic factors explain the phenotypic difference between monozygotic and dizygotic twins. In the family-based study design, heritability is estimated by shared genetic markers or estimates of expected genetic relatedness from the family structure. It is referred as the broad-sense heritability [33].

Based on twin studies, heritability estimates of MRI measures of brain volume, grey matter and white matter vary from >70% in the neonate, 57-91% in children aged 5-19 years, 75-94% in adults (19-59 years) and 72-85% in elderly (>60 years) (Jansen et al., 2015). The brain white matter heritability estimates of DTI measure of FA were 88% (whole brain), 53-90% (regional tracts) [34], and that of MD, RD and axial diffusivity (AxD) was 73%, 74% and 72% respectively [35].

Reports on the heritability estimates on general cognition showed a trend of increasing genetic influence with age, from 23% in childhood to 80% in older twins [36–38]. Though different studies showed different heritability estimates, general cognition had generally higher heritability than domain-specific test batteries (processing speed: ~50%, memory: ~40%, verbal reasoning: ~70%) [39–41]. These estimates differ across different ages and studies, yet there is a strong element of genetic influence playing a role in various phases of brain development, morphology and function [42].

1.1.3.2 Genotype based heritability

Though twin or family-based studies provide an optimal setting to study the relative effect of genes and environment on brain development and functioning, it often gives inflated estimates due to shared environmental effects, non-additive genetic variations and epigenetic factors. With genome-wide association studies (GWAS) using single nucleotide polymorphism (SNP) arrays, estimates of genetic variance due to SNPs in unrelated individuals can be calculated, known as SNP-based heritability. SNP-based heritability is defined as the proportion of phenotypic variance explained by any set of SNPs either obtained from microarray, imputed from reference or sequencing data [43].

SNP-based heritability from UK biobank for total, grey matter and white matter volumes and DTI measures of white matter tract were 44-66%, 41-67%,43-68% and 49% [44,45]. Heritability for general cognition was 20-30%, verbal-numerical reasoning 31%, reaction time 11% and memory 5% [46,47]. GWAS derived SNP-based heritability could thus only explain a minor proportion of the heritability of these traits as compared to that of twin or family-based studies. The lower SNP-based heritability as compared to twin studies might be explained by an overestimation of the heritability by twin studies or by an underestimation by SNP-based studies. The latter does not capture rare variants [44], structural variants or epigenetic differences, that all may contribute to heritability assessed by twin studies. Importantly, under the polygenic model of GWAS, heritability is a function of sample size therefore with an increase in sample size and an increase in the number of SNPs or variants analysed, the gap between family-based and SNP based heritability is reducing [43].

1.2 Predictors of brain ageing

Humans do not experience ageing at a similar rate, which is reflected in the high variability of developing age-related diseases and of age at onset of these disease at the population level. The same is true for ageing of the brain and the associated neurological diseases. Ageing or brain

ageing is conceptualised as a complex or multifactorial trait depending on the interaction of genetic, environmental and lifestyle factors. Identifying these factors, therefore, is of major interest as their knowledge facilitate the prediction of age-related diseases as well our understanding of their aetiology and pathomechanism [13]. Heritability indexes derived from family-based studies can estimate the proportion of variance in ageing explained by genetic factors.

Epidemiological approaches using GWAS have identified many SNPs that predicts the risk of developing a disease. For instance, individuals inheriting an E4 allele in the apolipoprotein (*ApoE*) gene have up to 20 times higher risk of developing AD as compared to people with E2 and E3 alleles. Many GWAS studies have not only identified genome-wide SNPs increasing the risk of developing a disease but also genes and pathways leading to the pathogenesis of these diseases [48]. Importantly, due to the composite role of genes and environment in the process of ageing, a variant or polymorphism itself may not be deterministic, causal or involved in the aetiology of a disease but may be related to the susceptibility or could be a powerful antecedent at any stage of disease pathway [49].

Lifestyle behaviours play a central role in ageing that can significantly modify the ageing trajectories and health status in the elderly. Healthy lifestyle, high diet quality and exercise predict the disease-free life, reduce risk of mortality, and poor cognitive functions [50–52]. In addition, moderate caloric restrictions reduce metabolic and hormonal factors implicated in diabetes, cardiovascular diseases, cancers and vascular dementia, indicating likely beneficial effect on ageing [53]. In contrast, stress and exposure to toxins, including drugs of abuse, have a substantial effect on health, longevity and neurodegenerative diseases [54]. A longitudinal study determining ageing trajectories showed that people with physical inactivity, smoking, and alcohol abstinence were associated with consistent poor health and accelerated decline during ageing [55,56]. Physical activity and non-smoking were two strong predictors of better health outcomes in later life as well as associated with decreased risk of cognitive decline and disability [56–58]. Exercise has a wholesome effect on a range of body functions such as strength, balance, flexibility and endurance. In the brain, exercise improves cerebral perfusion, impart positive change in brain volume and connectivity, synaptic plasticity, neurogenesis, and regulation of trophic factors [59], which ultimately translated into better global cognitive functions and with better performance in the cognitive domains of memory, executive function,

and motor skills in elderly. Therefore, genetic as well as lifestyle predictors of brain ageing are useful in population disease risk stratifications [60].

1.3 Biological markers of ageing in epidemiological settings

A biological marker can be defined as a trait that can assess and evaluate normal biological processes, pathological processes or pharmacological response to a therapeutic intervention. In general, a better understanding of biological processes and scientific-technological advances aided the scientific community to identify different biological markers. Biological markers bolster our understanding of the prediction, cause, diagnosis, progression, regression, or outcome of treatment of a disease [49]. With ageing, the body's susceptibility to age-related diseases increases exponentially. A marker for ageing could monitor age-related processes and delineate the events between normal progression and a disease state.

An ageing marker for epidemiology shall be correlated with age, age-related changes, risk factors of age-related diseases as well as associated with age-related diseases and mortality. They shall be non-invasive, without harm for the participants and preferably easily and cheaply measurable. A variety of markers monitoring changes in blood, brain, cerebrospinal fluid, muscle, nerve, skin, and urine have been developed in order to gain insight into the changes within the nervous system during normal as well as pathological ageing [49]. Ageing markers can be classified as global and organ-specific markers. A global marker can assess the wholesome effect of ageing at the organismal level while organ-specific marker delivers information on ageing at the level of specific organs. A well-established global ageing maker is the length of telomeres at the end of the chromosomes [61] and a recently developed novel ageing marker, the so-called epigenetic clock [62].

In the field of brain ageing research, neuroimaging and cognitive tests served as brain-specific markers providing in-depth information on brain structural and functional manifestations of ageing. MRI scans are well established surrogate markers to study ageing-related changes in the brain. Recently, the use of artificial intelligence with MRI data and advanced MRI methods such as DTI facilitated the development of novel brain-specific markers such as brain–predicted age [63] and brain PSMD [31].

For simplicity, in this thesis, we classified markers of ageing into two categories 'Established markers' and 'Recent developments' and further subclassified them as 'Global' and 'Brain-specific' markers.

1.3.1 Established markers

1.3.1.1 Global markers

1.3.1.1.1 Telomere length

Telomeres are nucleoprotein protective caps at the end of the chromosomes containing repeating hexamer, TTAGGG, sequences. The length of telomere in human is 9-15 kilobases with 50-300 single stretch of guanine overhang on 3' end and is evolutionary well conserved across species [61,64]. Telomeres protect the ends of linear chromosomal DNA from being recognised as broken ends; which would be recognised and repaired by several mechanisms such as DNA end joining, DNA replication, or DNA recombination processes. During mitosis, DNA polymerase is unable to complete replication of linear DNA at the end of the chromosomes hence telomeres shortens with each cell cycle. Independent of mitosis, telomere attrition is also possible through oxidative damage of the guanine residues within the telomere repeats resulting in double-stranded breaks [65–67]. Shortening of the telomeres destabilises the genome that in turn leads to senescence of the affected cells [3,61,68]. Telomere shortening is considered as part of the cellular tumour suppressor mechanisms preventing of uncontrolled cell division of senescent cells, a process involved in cancer development. A specialised DNA polymerase, known as telomerase, is required to replicate telomeric repeats. The enzyme is expressed in germline, stem and cancer cells and is lacking in most human somatic cells. Over-expression or reactivation of telomerase reverted and delayed ageing in animal models [3,67]. The telomeres are bound to the sheltering protein complex, which can protect the telomeres from deleterious DNA damages and regulates the exposure of telomeres by the telomerase [61,68].

In humans, telomerase deficiency is involved in diseases such as pulmonary fibrosis, dyskeratosis congenita, probably through a reduction of the regenerative capacity of different tissues. Affected patients have shortened telomeres, and the severity of the disease has been shown to increase with shorter telomere length. Sheltering mutations have also been detected in some cases of these diseases [3,69]. In epidemiological studies, leukocyte telomere length (LTL) is frequently used as a biological marker of ageing, especially of ageing related to cellular senescence. Telomere attrition occurs during the normal ageing; however, increased rate of attrition leads to accelerated ageing, and its amelioration delays normal ageing processes, fulfilling the criterion for a biological ageing marker [3,66].

Telomere length in different human cells varies depending upon their replicative activity, and yet there is a high correlation between telomere length across somatic tissues. This strong

correlation is referred as 'synchrony' and is a conserved trait across the species [70]. Hence, it is postulated that when a person having longer telomeres in a particular tissue, will have longer telomeres in other tissues as well. Therefore, the genetic regulation of telomere length is probably tissue independent [71].

LTL, which can be easily measured in peripheral blood, is used as a surrogate parameter for telomere length elsewhere in the body. LTL is inversely correlated with age with an estimated rate of shortening in the range of 20-30 base pairs per year [72]. The heritability estimates of LTL are ranging from 31% to as high as 88% [73] with substantial inter-individual variability due to the environmental effects as well [74]. In human epidemiological studies, LTL is associated with age-related chronic vascular and degenerative diseases, especially with AD and stroke [61,75]. Shorter LTL is also related to risk for earlier onset of dementia, mortality [76] and all-cause mortality in the elderly [77].

1.3.1.2 Brain-specific markers

1.3.1.2.1 Magnetic resonance imaging

MRI facilitated *in vivo* tracking of various structural and functional changes occurring in the brain with advancing age, enabling to identify various patterns within the broad spectrum of brain ageing [78]. Accentuated contrast difference on T1-weighted sequences of brain MRI enabled to distinguish between grey matter, white matter and cerebrospinal fluid. These images are used to measure the sizes of various cortical and subcortical structures. On T2- weighted MRI, ischemic brain lesions such as WMH are seen as enhanced contrast regions due to oedema and tissue loss and can be quantified [79]. Fluid-attenuated inversion recovery (FLAIR) imaging in which cerebrospinal fluid signals are suppressed can also be used for visualising WMH.

Over the years many population-based studies used structural brain MRIs in order to study atrophy related changes such as total and regional brain volumes, grey and white matter volumes, cortical thickness as well as markers of white matter integrity such as WMH. These changes are used as surrogates to observe brain ageing and to stratify the risk of developing age-related neurological diseases in elderly [80].

1.3.1.2.2 Cognitive tests

With advancing age, cognitive abilities decline. In neuropsychology, evaluation and characterisation of cognitive performances are done by using cognitive test batteries. Within each cognitive domain, there are subdomains, and different neuropsychological tests evaluate the performance within these. The domains are classified based on simplicity as

basic/underlying processes (memory, attention, language, basic sensory and perceptual) or executive functions which involve the execution of more than one basic functioning (reasoning, problem-solving). Domains most frequently investigated in epidemiological studies are attention, processing speed, memory, reasoning, and executive functions [20,25]. Domains are not independent of each other as more complex cognitive processes involve utilisation of more than one basic function and hence, showing a high level of positive correlation in the performance of different cognitive test batteries. The first principle component from the results of these tests is known as 'g-factor' and used as a marker of global cognition. In the epidemiological setting, these markers of cognitive ageing are shown to have a strong association with age and used for risk stratification of developing neurodegenerative diseases and dementia [13].

1.3.2 Recent developments

1.3.2.1 Global markers

1.3.2.1.1 Epigenetic clocks

The epigenetic clock represents an age estimate in years based on DNA methylation patterns in specific CpGs within the genome calculated using a supervised machine learning algorithm. There are few epigenetic clocks available, such as Hannum's clock (71 CpGs), Horvath's clock (353 CpGs), DNAmPhenoAge (513 CpGs) but all are based on the same principle. The correlation of these clocks with chronological age exceeds 0.8 across all ages. They are used to determine a differential ageing pattern in different organs and showed that most tissues and organs from the same individual exhibit similar age as indicated by 'synchronicity' of DNAmAge across tissues. Acceleration of these clocks is associated with age-related conditions such as neuropathology in elderly, physical and cognitive fitness, Parkinson's disease, and centenarian status. Moreover, they predict all-cause mortality and the risk of developing certain cancers. Acceleration of these clocks in the blood is also associated with reduced white matter integrity, cognitive functioning, and dementia status in elderly [62].

1.3.2.2 Brain-specific markers

1.3.2.2.1 Diffusion tensor imaging

DTI estimates patterns of white matter connectivity in the brain using white matter tractography. DTI maps and characterises three-dimensional diffusion of water as a function of spatial location in white matter. It is highly sensitive to changes at the cellular and microstructural level. Water diffusion in white matter is anisotropic, and directionality of flow is along the fibres. DTI is highly sensitive to changes in local tissue microstructure due to injury,

pathological processes, or normal physiological changes occurring during ageing and hence frequently used as a marker for white matter integrity [81].

DTI measures of microstructural integrity such as FA, MD, AxD, and RD are used as markers of neurodegeneration. FA measures the directionality of water molecules along the axons. Higher FA value indicates more intact axons or a higher degree of myelination, whereas lower FA value indicates loss of white matter integrity and injury. AxD and RD are representing the diffusion of water along and perpendicular to axonal fibres respectively, while MD is the mean diffusion in all directions. Increase in MD, AxD and RD suggest damage to the myelin sheath, axonal injury and demyelination respectively [19]. Reduction in FA is associated with age as well as with neurodegenerative and cerebrovascular diseases [80]. PSMD is another DTI based marker based on skeletonisation and histogram analysis. PSMD was associated with processing speed and is a more sensitive imaging marker of ageing than brain volume, volume of white matter hyperintensities, and volume of lacunes [31].

1.3.2.2.2 Brain-age

By virtue of artificial intelligence, especially machine learning, thousands of brain structural MRI scans from healthy adults as a function of age are trained to build a predictive statistical model for ageing. These algorithms take into consideration all the nocent effect piled up over the years in an individual causing changes in brain structure, function and disease risk. The age predicted by these models is known as 'brain age'. Smaller brain age as compared to chronological age indicates biologically younger brain and vice versa. The brain age as well as the difference between brain and chronological age are used as a marker of brain ageing (positive difference: brain appears older than their chronological age, negative: brain appears younger than their chronological age). The older appearing brain is associated with poor performance in cognitive and physiological measures of ageing [13].

1.4 Molecular mechanisms and pathological processes in brain ageing

Age is the primary driver and risk factor for late-onset neurodegenerative diseases [5]. Ageing is a complex process involving genetic and environmental factors as well as their interaction. The working hypothesis is that normal brain ageing forms a continuum with neurodegenerative diseases and neurodegenerative diseases are expressions of accelerated ageing [54]. Several molecular mechanisms have been proposed driving the processes of ageing based on the results of post-mortem brain examinations in humans as well in model organisms, MRI studies *and in vitro* as well as *in vivo* disease models. The described molecular mechanisms/pathological

processes are not necessarily causal but may represent accompanying processes or consequences or of ageing.

1.4.1 Post-mortem examination findings

Brain autopsies of cognitively intact older individuals show the aggregation of amyloid plaques, neurofibrillary tangles, Lewy bodies, inclusions of TAR DNA-binding protein 43, synaptic dystrophy, the loss of neurons and the loss of brain volume [23,54,82]. Brain volume and weight decreases with advancing age by approximately 10% in humans from their fifties to their nineties corresponding to a 150 gm loss in brain tissue or of 5% loss every decade after the age of 40 [14,54,82]. This brain shrinkage might be due to neuronal or glial cell death, decrease in neuronal volume, reduction in dendritic synapses, white matter, myelination, or loss of fluid.

1.4.2 Neuroinflammation and cellular senescence

Low level of chronic inflammation is often present in the ageing brain and linked with age-related neurodegenerative diseases. Reduction in neuronal population, dendritic and axonal arborisation, post-synaptic densities, dendritic spines, presynaptic markers, synapse and cortical volume, are all associated with the presence of chronic inflammation. Senescent glial cells are thought to be a cause of inflammation as an increase in a load of senescent astrocytes and their markers such as p16INK4a and matrix metalloproteinase 3 increases significantly with age. Importantly, an even more prominent increase in inflammatory markers has been observed at autopsy in the brains of Parkinson's disease, AD patients when compared to elderly controls. This suggests a common role of cellular senescence and neuroinflammation as a molecular mechanism during both brain ageing and neurodegeneration [83].

1.4.3 Small vessel disease

Brain MRI studies reported the presence of age-associated ischemic changes including WMH, micro-bleeds, micro-infarcts representing the spectrum of cerebral small vessel disease as well as cortical and subcortical macro-infarcts representing macro-vascular changes in the non-demented elderly [15,84,85]. Small vessel disease is a disorder of small brain arterioles, capillaries and probably venules, which is detected as WMH, lacunes, increase in perivascular spaces, micro-bleeds, superficial siderosis and micro-infarcts on MRI. Mechanisms causing small vessel disease are dysfunction of vascular endothelium, blood-brain barrier, cerebral blood flow, and perivascular flow. Dysfunctional vascular endothelium and blood-brain barrier lead to an increase in interstitial fluids and proteins damaging astrocytes. The presence of damaged astrocytes, in turn, results in the disruption of interstitial nutrient exchange, neuronal energy

supply, leading to a blockage of the maturation of oligodendrocyte precursors finally causing defective myelination, repair and energy support to the axons. The dysfunctional perivascular flow might impede the clearance of interstitial drainage and metabolites from the tissue [86]. These changes in the brain do not necessarily precipitate clinically or may take years to manifest as cognitive, behavioural and functional impairments. At various stages of dementia, the brain shows abnormal protein deposits and damaged neurovasculature [87]. Cerebral small vessel disease evolves as a silent and complex subclinical disease in the elderly finally resulting in an increased risk of stroke, dementia and death [15].

1.4.4 Studies in model organisms

At the cellular and molecular level, the ageing process can be experimentally studied *in vitro* or *in vivo* using model organisms. The processes observed have been described as the 'hallmark of brain ageing by Mattson & Arumugam [88].

1.4.4.1 Mitochondrial dysfunction

Mitochondria from animal brain tissue show age-related altercations such as enlargement, fragmentation, oxidative damage to DNA, and impaired electron transport system. Dysfunctional mitochondria accumulate in ageing neuron and astrocytes as seen in cell culture studies. Dysfunctional mitochondria are supposed to compromise activities necessary for neuronal function and viability [88].

1.4.4.2 Oxidative damage

As a result of an oxidative imbalance in ageing, neurons tend to accumulate damaged proteins and mitochondria. Nitric oxide-mediated oxidative damage is involved in the pathology of vascular dysfunction in the ageing cerebral cortex. Accumulation of oxidatively damaged molecules is seen as accentuators of ageing in mice and drosophila [88].

1.4.4.3 Impaired lysosome and proteasome function

Neurons being post-mitotic cells, their ability to remove damaged and dysfunctional molecules and organelles is central for their functioning. During ageing, autophagic and proteasomal degradation is impaired in neurons exposing them to adversities and death. Stimulators of autophagy increase lifespan, decrease memory and learning deficits, as well as decrease neurodegeneration in model organisms. Hence, lysosome and autophagy play an essential role in protecting neurons from ageing related adversities [88].

1.4.4.4 Dysregulation of neuronal calcium homeostasis

Neuronal abilities to control calcium regulation is hampered during ageing. Dysregulation of calcium homeostasis is involved in age-related cognitive deficit. In aged rats, experimental restoration of dysregulated calcium homeostasis ameliorates cognitive deficits [88].

1.4.4.5 Compromised cellular stress response

Neurons are continuously exposed to ionic, metabolic and oxidative stress in their normal physiological functioning. Chronic stress impairs neuronal plasticity and predisposes neurons to degeneration [88].

1.4.4.6 Aberrant neuronal network activity

Communication within and between different brain regions occur through interlinked neuronal networks. During brain ageing, imbalance in the activity in these networks takes place disrupting the communication within the brain. Moreover, damage to white matter, that contains the axons of neurons is seen in ageing and is linked to cognitive decline [88].

1.4.4.7 Impaired DNA repair

DNA repair systems maintain DNA integrity and remove damaged DNA. Brain tissues analysis showed an increase in damaged DNA and a decrease in the expression of some DNA repair proteins during ageing. People with a mutation in DNA repair genes show a rapid increase in age-related phenotypes already at a young age [88].

1.5 Rationale

1. LTL is an established biomarker for ageing. However, studies are inconsistent regarding its association with structural and functional changes of the brain taking place during ageing. With the observed increase in longevity, also the expected number of individuals affected by neurodegenerative diseases increased exponentially during the last decades. Possible biological markers for brain ageing, such as LTL are therefore of major interest as they may identify individuals from the population who are prone to develop cognitive decline in the future and support to monitor process of brain ageing as well as the effect of preventive and therapeutic measures on it. In addition, being a biomarker, telomere shortening may also be causally involved in ageing of the brain, meaning that protecting telomeres throughout life will also protect the brain when it comes to ageing.
2. Brain ageing related phenotypes show heritability of 30-70%. GWAS on MRI correlates of brain ageing were highly successful and identified numerous common variants explaining a part of the heritability. However, as with other common complex traits, a substantial proportion of the heritability remains unexplained by these studies. We expect that identifying rare variants by next generation sequencing (NGS) at the whole exome/genome level will bolster our understanding of the genetic architecture of brain ageing and the processes driving it. Indeed, whole exome (WES) and whole genome sequencing (WGS) became an extension of microarray-based GWAS. Establishing robust and valid protocols in order to perform NGS on diverse platforms were major achievements and the prerequisite for WES/WGS in large cohorts.

1.6 Working hypothesis

1. Longer LTL is related to
    a. better structural preservation of the brain as detected by MRI markers such as larger BPF and less WMH
    b. better global cognitive performance detected by g-factor and in the domains of attention/speed, conceptualisation, memory and visuopractical skills.
2. MRI markers of brain ageing can be used as powerful intermediate markers to better understand the genetic determinants, and thereby molecular mechanisms underlying the occurrence of dementia.

1.7 Aims

Our proposed aims to test the hypotheses mentioned above are

1. To investigate the association between
    a. global correlates of brain ageing such as BPF and WMH load and Fazekas score with LTL
    b. global (g-factor) and domain-specific cognition such as attention/speed, conceptualisation, memory, and visuopractical skills
2. To calculate telomere length from WGS data and to reproduce results obtained from qPCR-based LTL measurements with brain ageing markers.
3. To identify rare genetic variants associated with structural markers of brain ageing in older community people
    a. establishing the WES sequencing method.
    b. identify rare genetic variants in WES and WGS data associated with structural markers of brain ageing.

2 Methods

2.1 Study setting

In this thesis, we utilised extensive phenotypic and genotypic data from three well-established, prospective community-dwelling cohort and family studies in the elderly population in the city of Graz, Austria, namely Austrian Stroke Prevention Study (ASPS)[89], Austrian Stroke Prevention Family Study (ASPS-Fam) and Graz Study Health and Ageing (GSHA).

2.1.1 ASPS

The ASPS is a single centre prospective follow up study on the effects of vascular risk factors on brain structure and function in the normal elderly population of the city of Graz[89]. The study was established in 1991 and ended in 2003. A total of 2007 participants were randomly selected from the official community register stratified by gender and 5-year age groups[89]. Individuals were excluded from the study if they had a history of neuropsychiatric disease, including previous stroke, transient ischemic attack, and dementia, or an abnormal neurologic examination determined based on a structured clinical interview and a physical and neurologic examination[89]. Participants were undergone extensive workup, including brain MRI, neuropsychological test batteries, complete blood cell counts, blood chemistry, three blood pressure measurements, ECG, and echocardiography. Follow-up examinations were conducted after 3 and 6 years [89]. We studied the association of LTL with MRI correlates of brain ageing and cognition in participants of this study.

2.1.2 ASPS-Fam

ASPS-Fam represents an extension of the ASPS[90]. Between 2006 and 2013, ASPS participants with at least one first grade relative willing to participate were invited to participate in ASPS-Fam[90]. Inclusion criteria were no history of previous stroke or dementia and a normal neurological examination[90]. A total of 410 individuals from 174 families were included in the study[90]. The ethics committee of the Medical University of Graz, Austria, approved the study protocol and written informed consent was obtained from all subjects. Each individual underwent diagnostic workup including clinical history, blood tests, cognitive testing and vascular risk factor assessment according to the ASPS protocol. They were all European Caucasians [90]. We used GWAS chip data from these participants in establishing the WES pipeline and performed WGS on 55 participants.

2.1.3 GSHA

The GSHA is an ongoing community-based longitudinal cohort study. It is an interdisciplinary investigation of ageing for the identification of genetic, environmental and lifestyle risk factors in ageing-dependent preclinical functional and structural lesions in the population of Graz, Austria. Inclusion criteria are residents of the city of Graz with age ≥ 45 years without any malignant tumour and chemotherapeutic treatment currently or in the past two years and pregnancy. The pilot phase of recruiting 100 participants was finished in 2016. Ageing-related phenotypes and clinical correlates collected from different clinical aspects of neurology, dermatology, ophthalmology, cardiology, genetic epidemiology, medical psychology, sports sciences, music sciences and nutrition amounting to a total of over 5700 defined variables. We performed WES and WGS using DNA from these participants

2.2 Diagnostic workup

2.2.1 Brain MRI

2.2.1.1 ASPS

In detail procedure for MRI acquisition was published previously [58,89]. MRI scans were performed on a 1.5-Tesla scanner (Philips Medical Systems, Eindhoven, Netherlands) using proton density- and T2-weighted sequences in a transverse orientation. T1-weighted sequences were generated in the sagittal plane[89]. Baseline and follow-up scans were performed with identical protocols. Scans from each study participants were read independently by three experienced investigators blinded to the clinical phenotype of the participant[58,89]. WMH, lacunes, and brain volume were determined in each study participant. For WMH measurements, the scans were analysed by investigators. They marked and outlined each WMH on transparency that was overlaid on the proton density scans[58,89]. Independent from this visual analysis, WMH load measurements were done on proton density-weighted images on an UltraSPARC workstation (Sun Microsystems, Santa Clara, CA) by a trained operator using DISPImage[89]. The operators used a hardcopy overlaid by the transparency, with all lesions outlined by the experienced readers as a reference and segmented all lesions on the computer image[89]. Lesion areas were then provided by the semi-automated thresholding algorithm implemented in DISPImage[58,89]. The total lesion volume (mm3) was calculated by multiplying the total lesion area by slice thickness[89]. The Fazekas scale on deep white matter changes was used to calculate WMH score where a score of 0 is the absence of any white matter change, 1 is punctate foci, 2 is beginning confluence of foci, and 3 is large confluent areas [91]. Brain

volume was calculated from the T2-weighted spin-echo sequence using the fully automated structural image evaluation of atrophy (Sienax, part of the FMRIB Software Library; http://www.fmrib.ox.ac.uk/fsl)[89]. BPF was estimated from the ratio of parenchymal volume to the total volume given by the outer surface of the brain[58].

2.2.1.2 ASPS-Fam and GSHA

This protocol for MRI acquisition was published previously in Seiler *et al.* [90]. MRI scans were performed on 3 T whole-body scanner (Tim Trio; Siemens, Erlangen, Germany). The imaging protocol included an axial FLAIR sequence (TR = 10000 ms, TE = 69 ms, inversion time = 2500 ms, number of slices = 40, slice thickness = 3 mm, in-plane resolution = 0.86 mm × 0.86 mm) and a high resolution T1 weighted 3D sequences with magnetisation preparation (MPRAGE) and whole-brain coverage (TR = 1900 ms, TE = 2.19 ms, inversion time = 900 ms, flip angle = 9°, isotropic resolution of 1 mm) [90].

WMH, silent non-lacunar and lacunar infarcts were recorded on FLAIR images as in ASPS. All lesions were outlined using a custom written IDL program (Exelis Visual Information Solutions, USA). The total lesion volume (cubic millimetre) was calculated using the program FSLMATHS (FSL, Oxford) by multiplying the lesion area with the slice thickness and normalised by head size. Cortex volume, normalised for the subject head size, was calculated from the T1 weighted MPRAGE images using the fully automated structural image evaluation of atrophy (FSL, Oxford) [90].

2.2.2 Cognition

A set of test batteries has been used in order to evaluate cognitive function within the domains of memory and learning abilities, conceptional reasoning, attention and speed as well as visuopractical skills in each participant [89]. The tests applied were widely used in German-speaking countries and always administered in the same order and under the same conditions[89]. Memory and learning capacity was assessed by a test called Bäumler's Lern- und Gedächtnistest[89]. It is a paper-pencil procedure consisting of six subsets. Three subsets (word, digit association task and story recall) screen for verbal memory, two subsets (trail and design recall) screen for visuospatial memory and one subset included image recognition paradigm[89]. The sum of weighted scores from these tests gave score in memory and total learning. The Wisconsin card sorting test was used to measure conceptional reasoning[89]. Attention and speed were assessed with the Alters Konzentrationstest of Gaterer, the part B of the trail making test, and the digit span test, which is part of the Wechsler Adult Intelligence Scale[89]. The reaction

time was assessed using a computerised complex reaction time task (Wiener Reaktionsgerät) which is part of the Schuhfried psychological test battery[89]. The computer records the number of correct responses and reaction times. Visuopractical skills were evaluated by Purdue's pegboard test. Individual test scores were converted to z-scores by normalising to the mean of test grand, and the average score within each cognitive domain was computed to get summary measures of domain-specific cognitive function[89]. The first unrotated principal component of all test on principal component analysis was used to assess general cognition given as the g-factor.

2.3 Classical predictors

2.3.1 Vascular risk factors

Hypertension was defined as a history of hypertension with repeated blood pressure measurements higher than 140/90 mm Hg, if he/she was treated using antihypertensive treatments or if two reading at the examination were more than 140/90 mm Hg[1]. We coded a participant as diabetic if the participant was treated for diabetes mellitus at the time of examination or if the fasting blood glucose level was exceeded 140 mg/dL[1]. Presence of cardiovascular disease (CVD) was coded as present if there was evidence of cardiac abnormalities knows to be a source for cerebral embolism, evidence of coronary heart disease according to the Rose questionnaire or relevant ECG findings, or if an individual presented with signs of ventricular hypertrophy on echocardiogram or ECG[1]. Total cholesterol, high-density lipoprotein (HDL), and triglycerides were measured from the blood, and low-density lipoprotein (LDL) was calculated based on the Friedewald formula[1]. ApoE genotyping was done using PCR-RFLP resulting in 6 genotypes which were further categorised into heterozygous, homozygous carriers of the E4 allele as well as non-carriers [1,92].

2.3.2 Lifestyle factors

Smoking status was assigned depending upon the history of smoking and coded into 3 categories such as 1) current smokers smoking at the time of the interview, 2) former smokers: smoked previously and 3) non-smokers: never smoked [1,89].

2.4 Genetic predictors

2.4.1 Leukocyte telomere length

2.4.1.1 ASPS

A previously established real-time PCR (RT-PCR) method was used to measure LTL [93,94]. The telomere repeats (T) and single-copy gene (S) (36b4 acidic ribosomal phosphoprotein) were amplified using Power SYBR® Green PCT Mastermix (Applied Biosystems, USA) with sequence-specific primers. Two separate assays were performed for telomere repeats and single-copy gene in two separate MicroAmpTM Fast Optical 384-well reaction plates on a 7900 HT fast Real-Time PCR System (Applied Biosciences, USA). Each sample was measured in triplicate. Inter-plate measurement variation was accounted for by adding three serial dilutions of standard DNA (150, 50, 16.7. 5.55 and 1.86 ng/µl) on each plate, one each for telomere signals and single-copy gene signal. From amplification signals for telomeres (T) and single-copy gene (S), the relative LTL (T/S ratio) was calculated according to the Cawthon's modified method [93] and normalised to the reference DNA pooled from 24 subjects including 50% females (mean age =51.5 years) loaded on each plate[94].

In LTL and brain ageing study, we included 909 ASPS participants who underwent LTL measurements along with MRI and cognitive tests[1,89]. The mean age of the participants was 65.9±8 years (range:46-90), 42.7% were males, 69.4% hypertensives, 10.9 % diabetics, and 40% had CVD[1]. The mean years of education were 11.3±2.6 (range: 9-18 years)[1]. There were 28.3% former smokers and 11% current smokers in the study sample. In total, 898 participants had the *ApoE4* genotypes, with 0.8% being homozygous and 19.3% heterozygous carriers[1]. The median of relative LTL was 0.61 (range: 0.05-2.60). The median of LTL was 0.61 (IQR: 0.47-0.82)[1].

2.4.1.2 ASPS-Fam and GSHA

Telomere length from WGS data was estimated using Computel [95] and TelSeq [96] software. Computel calculates mean LTL in base pairs by realigning the raw fastq files obtained from the WGS to the telomeric reference generated by the application. It counts telomeric reads and then normalises it to the mean coverage of the sample and output mean telomeric length [95].TelSeq looks for sequencing reads containing a fixed number of the telomeric pattern "TTAGGG" in BAM files and then normalises to the number of reads in that sequencing library having GC contents between 48-50%. Reads length of WGS libraries was calculated by samtools [97]. In our

WGS data, most of the reads had a length of 151 base pairs; therefore, a repeat number of 12 was chosen to detect telomeric reads, which were also advised in the literature [98].

Due to the use of the realigning step, Computel is time-consuming and requires much more computational power than TelSeq. As correlations of telomere lengths estimated from the 2 software were highly significant (r^2>99%) ([98], own data), we decided to analyse the samples available in our cohorts using TelSeq.

2.4.2 Single Nucleotide Variants

2.4.2.1 Whole exome sequencing

WES was performed using Ion Proton™ sequencer (Life Technologies, USA) in combination with the Ion AmpliSeq™ library kit plus (Life Technologies, USA). We used in-house DNAs and NA12878 reference DNA to establish WES on Ion Proton[2].

2.4.2.1.1 In-house DNAs

We randomly selected 12 in-house DNA samples which were previously extracted from whole peripheral EDTA blood using the phenol-chloroform method. Out of 12 DNAs, six samples were withdrawn from the ASPS consisting of one female and five males with a mean age of 55.7 years [2,99]. The remaining six samples, containing two females and four males with a mean age of 74.3 years, were part of the Prospective Registry on Dementia in Austria [2] and represented patients clinically diagnosed with probable Alzheimer's dementia. These DNAs were stored at -80 degree Celsius. The DNAs were selected based on the presence of microarray genotyping data either from genome-wide or exome chip or both[2]. Altogether, 11 DNAs were genotyped by Affymetrix Genome-Wide Human SNP Array 6.0, one with Human 610-Quad BeadChip (Thermo Fisher Scientific, USA) for genome-wide microarray genotypes, and by Exome chip Illumina Infinium Exome-24 v1.1 BeadChip array (n=11) (Illumina Inc., USA). The quality of DNAs was checked on 1.5% agarose gel and quantified using NanoDrop 3300 fluorospectrometer (Thermo Fisher Scientific, USA) before sequencing[2]. We used the raw data from microarray without any filtration such as cut off for minor allele frequency to avoid loss of any genotypes for further comparison with WES data[2].

2.4.2.1.2 Reference DNA NA12878

We used reference DNA NA12878 which is genomic DNA derived from a massive growth of the human lymphoblastoid cell lines GM12878, (Reference Material 8398) provided by the "National Institute of Standards and Technology". This DNA is one of the reference materials provided by the Genome in a Bottle consortium and used for assessing the performances of

sequencing technologies. Reference values for single nucleotide variants, small insertions and deletions and homozygous reference genotypes from this DNA is provided to obtain an estimate of true positives (TP), false positives (FP) and false negative (FN) variant calls from the sequencing [100]. Henceforth, the reference values of variants in NA12878 is addressed as the truthset[2].

2.4.2.1.3 Library preparation

2.4.2.1.3.1 Ion AmpliSeq Exome RDY Library

The Ion AmpliSeq™ library kit plus (Life Technologies, USA) used for the preparation of libraries using 100 ng of DNA (100 ng/µl) as per the manufacturer's protocol[2]. Target regions of genomic DNA were amplified using Ion AmpliSeq primer pool, distributed along 12 wells in a row of the Ion AmpliSeq™ Exome RDY 96 well plate[2]. The primer pool consists of 294,000 primer pairs, and each pool consists of 24,000 primer pairs. Amplification of target regions followed by partially digesting the primer sequences using FuPa reagent, and ligation of adapters (Ion P1 adapter) and barcodes (Ion Xpress™) to the amplified amplicons[2]. In total, 50 µl of purified unamplified library retrieved at the end using AMPure XP reagent purification system (Beckman Coulter Life Sciences, USA). In the Ion AmpliSeq Exome RDY plate, a 96 well plate, eight different genomic DNAs are amplified and barcoded at the same time [2].

2.4.2.1.3.2 SureSelect All Human Exome V6

We prepared the library using the SureSelect Target Enrichment System (Agilent Technologies, USA) following the manufacturer's protocol[2]. In total, 1µg of gDNA (100 ng/µl) was fragmented using Ion Shear Plus enzyme mix to obtained approximately 130 base pairs fragments[2]. After purifying and size selecting the sample using AMPure XP beads (Beckman Coulter Life Sciences, USA), library quality was assessed on 2100 Bioanalyser (Agilent Technologies, USA) (distribution between 50 to 250 nucleotides) [2]. Next, ligation of Ion Xpress barcodes and P1 adapters to the end of DNA fragments was followed by amplification of the library[2]. The amplified DNA fragments were hybridised to biotinylated RNA library baits and captured using streptavidin-coated magnetic beads. Finally, captured library fragments were amplified and quality assessed on 2100 Bioanalyser (Agilent Technologies, USA)[2].

2.4.2.1.4 Library quantitation using TaqMan

Ion library TaqMan quantitation kit (Life Technologies, USA) on the 7900 real-time PCR system (Applied Biosystems, USA) for quantitation of both unamplified libraries was used[2]. It is a qPCR-based kit which determines the amount of amplifiable template in an unamplified

Ion library. *E. coli* DH10B 68pM control library from the kit is used in 5 serial dilutions of 6.8 pM, 0.68pM, 0.068pM, 0.0068pM and 0.00068pM to generate a standard curve. Unamplified sample libraries were diluted in 1:2000, and 1:20000 dilutions and three technical replicates of each dilution were prepared[2]. Each reaction was set up with 10µl of the qPCR mix, one µl of 20X quantitation assay, 4 µl of nuclease-free water and five µl of the diluted control or sample library or nuclease-free water depending upon standard, sample library or negative control respectively. The concentration of each replicate was calculated by multiplying dilution factor with quantity obtained from the qPCR output. The average of six concentrations was calculated to get the final concentration of unamplified library[2].

2.4.2.1.5 Template preparation using Ion Chef System

We used Ion Chef Instrument (Life Technologies, USA) and Ion PI™ Hi-Q™ chemistry (Life Technologies, USA) for preparation of template-positive Ion Sphere particles[2]. After quantification, unamplified libraries were diluted to 50 pM, and 25µl of each diluted library was added to the library sample tube to load on a single Ion PI chip v3 (NA12878 DNA). In the case of in-house DNAs, we loaded two samples per chip; therefore, instead of 25µl of a diluted library, we mixed 12.5 µl of two libraries and added to the library sample tube. Each kit has two library sample tubes; hence, template enrichment of four libraries was achieved at the same time. We performed quality control to assess templating efficiency of Ion spheres using Qubit™ 2.0 Fluorometer (Thermo Fischer Scientific, USA). When Ion sphere particle efficiency was between 10-30 %, chip loading was performed[2].

2.4.2.1.6 Sequencing

Ion Proton™ sequencer was cleaned using chloride solution (once a week) and 18 MΩ water (every day) obtained from the Milli-Q® integral water treatment system (Merck KGaA, Germany). After calibrating the sequencer, an Ion PI chip loaded with the enriched library was inserted into the sequencer and sequenced using PI™ Hi-Q™ sequencing 200 chemistry (Life Technologies, USA)[2].

2.4.2.2 Whole genome sequencing

WGS was performed at McGill University, Genome Quebec Innovation Centre in Canada on Illumina Hiseq X platform (Illumina, San Diego, CA, USA) using pair-end reads at 35X average read depth. Burrows Wheeler Aligner [101] was used to align sequence reads to GRCh 38 assembly of the human genome and genome analysis toolkit [102] haplotype caller version 3.7 was used to call the variant from the whole genome.

2.5 Data analysis

2.5.1 LTL and brain ageing

2.5.1.1 Statistical analyses

The statistical analysis was performed using IBM SPSS statistics version 25. The normal distribution of the variables was tested using the Kolmogorov-Smirnov test and by visual inspection of histograms[1]. LTL measurement was transformed into z-scores relative to the mean of the whole group. We converted the skewed distribution of WMH load into normal distribution by log transformation after adding 1 to the volume[1]. BPF was converted into percentage by multiplying the value of fraction by 100 to facilitate the interpretation of the results from linear regression[1]. Cognitive tests scores were normally distributed. One outlier with a value of -5.95 within the attention/speed domain was removed[1]. Co-variates for multiple linear regression models were selected based on their correlation with outcome variables and LTL (Pearson's correlation $p<0.1$) and/or based on previous reports on their association with the phenotypes[1].

Linear regression models were used to test the effect of LTL on brain morphological measures and cognition in the presence of age, sex, risk factors such as hypertension, diabetes, cardiovascular disease, BMI, HDL, years of education, smoking, and *ApoE*4 genotypes[1]. The model I was adjusted for age and sex, Model II, additionally for hypertension, diabetes, CVD, BMI, education, HDL, and smoking status, and Model III for *ApoE*4 carrier status[1].

2.5.1.2 Subgroup analyses

We divided our cohort into subgroups based on sex (men/women), age (≤ 65y/>65y), hypertension (normotensives/hypertensives), BMI (normal weight: ≤ 25 Kg/m^2 / overweight: BMI<25), education (basic education≤ 10 years / >10 years mid to high education), diabetes (No/Yes), CVD (No/Yes) and *ApoE*4 carrier status (No/ Yes)[1]. We did not use education as a covariate in the linear regression model in education subgroup and *ApoE*4 status in ApoE4 subgroup, respectively[1].

We formally tested if the risk factors modulate the effect of LTL by using the interaction terms namely (gender × zLTL), (age × zLTL), (hypertension × zLTL), (BMI × zLTL), (education × zLTL), (DM × zLTL), (CVD × zLTL) and (*ApoE*4 status × zLTL) in the model III of regression[1].

2.5.1.3 Mediation analysis

We used bootstrapping (PROCESS macro version 3.4) [103] to test if the effect of the independent variable, LTL on the dependent variable, attention/speed is mediated through BPF or WMH in the presence of confounders (Model III)_[1]. Bootstrapping gives the estimates of direct and indirect effects. The extent of the effect which was mediated through BPF or WMH was calculated by repeating the sampling procedure 5000 times[1]. Effect sizes, as well as 95% confidence intervals (95%CI), are given. Significant mediation is present when 95%CI does not include the value of zero[1].

2.5.2 Establishing methods: WES

2.5.2.1 Data Analysis

The data analysis was performed using Ion Torrent Suite version 5.4 (Life Technologies, USA). The Torrent Mapping Alignment Program (TMAP) version 5.2 and the Torrent Variant Caller (TVC) version 5.4 were used for alignment against human hg19 assembly and for variant calling under default low stringency settings respectively[2].The default low stringency settings included 1) minimum allele frequency of 0.1, 2) minimum coverage of 5, 3) minimum coverage on each strand of zero and 4) maximum strand bias of 0.98. We analysed the variants in the library-specific total target region (TTR)_[2].

In case of AmpliSeq, we analysed variants also in the effective target region (ETR). The ETR was introduced by the manufacturer into the default Ion Torrent pipeline to exclude poor performing regions enriched for FPs or having low coverage[2]. After evaluating AmpliSeq's performance, they identified the set of regions where AmpliSeq performs poorly[2]. These regions are at the edge of amplicons, where sequence error interact with alignment to cause FPs. This approach by the manufacturer included[2] "

1. Selectively trimming of 9.4 MB of intronic regions from the amplicon with the largest number of filtered candidates within introns (targeting false positives in introns)
2. Trimming of 0.4 MB of the non-Medical Exome Project exonic regions from amplicons with consistently very low read coverage (targeting false negative)
3. Trimming of 1.4MB of the non-medical Exome Project exonic regions from amplicons with the largest number of filtered candidates within an intron (targeting false positives in exons)"

We downloaded RefSeq, Ensembl, and UCSC defined coding regions from UCSC genome browser in the form of BED files (20/04/2017)_[2].

We validated our data against the NA12878 v3.3.2 high confidence reference calls provided by the Genome in the Bottle project [100,104] as downloaded from their ftp server. For optimising our pipeline, we used the high confidence region (HCR), provided as BED file [100,104]. HCR specifies those regions in the genome where genotypes can be called confidently. These regions were generated after arbitrating between 11 whole genome and three exome data sets from 5 sequencing platforms and seven mappers and carefully filtering uncertain sites[100]. They identified possible SNP sites in the genome that differ between platforms and arbitrate between different datasets. After arbitration, they filtered out regions "1) with simple repeats which are not entirely covered by reads from any data set 2) with known tandem duplications and not in the GRCh37 assembly 3) paralogous to the 1000 Genomes Project "decoy reference" which contains sequence not in GRCh37 reference genome 4) in the RepeatSeq dataset 5) located inside structural variants for NA12878 that have been submitted to dbVar and 6) sites where they could not determine the reason for discordant genotypes" [100]. After doing this, they provided a BED region in which they can confidently call the variants. We intersected the target regions from AmpliSeq and SureSelect with the provided HCR to get the HCR regions in the respective target design.

We used bedtools [105] to manipulate BED files and vcf files and bcftools to calculate the true positives (TPs), FNs, and FPs. We used vcflib vcfallelicprimitives [106] module to generate phased genotypes and vt to regularise the variants [107]. The vcfallelicprimitive module splits the multiple representations of a single record in a vcf file into multiple lines[2]. This is necessary as indels and complex variants are frequently called differently depending upon the aligner used to create BAM files[2]. It results in the representation of multi-nucleotide polymorphisms as two SNVs[2]. The vt performs normalisation by left alignment and presents a variant in as few nucleotides as possible. The normalisation helps to compare the variants called by the different variant caller to minimise errors[2].

For fine-tuning, we applied different combinations of parameters for variant calling on TVC. The parameters which we considered for parameter tuning were 1) minimum allele frequency, 2) minimum coverage, 3) minimum coverage on either strand, and 4) maximum strand bias[2]. We stepwise changed minimum allele frequency from 0.1 to 0.2, minimum coverage from 5 to 10, minimum coverage on either strand from 0 to 5 and maximum strand bias from 0.98 to 0.8 to get a balance between FNs and FPs[2].

In order to explore the effect of duplicates on the performance of the libraries, we performed duplicate removal using three different tools, namely Picard, samtools [101] and a java-based tool "MarkDupbyStartEnd" [108]. Picard and samtools were designed for Illumina data and considered only start position of a read to mark it as a duplicate[2]. This raises serious problems in AmpliSeq libraries as it is PCR based and uses primers to capture the exome[2]. Therefore, in AmpliSeq, the start position of the reads is the same[2]. Next, we used a java-based tool "MarkDupbyStartEnd" that marks a read as a duplicate only if both its start and end positions are the same[2]. For *in silico* downsampling, we used samtools view -s option which selects the desired number of reads from a big BAM file [101]. Finally, we visualised all FNs and FPs on Chr 1, 7, 16, 19, and X using IGV [109]. We used Rstudio for statistical computation and graphics.

2.5.2.2 Categorisation of FNs and FPs

We performed manual inspection for all FNs and FPs on Chr 1, 7, 16, 19 and X[2]. Chr 1 was chosen as it is the largest chromosome and has the highest number of FNs, Chr 7 and 16 as they have a high density of exonic monomer repeats, Chr 19 as it has the highest density of sequence repeats and Chr X as a representation of a sex chromosome [110]. We classified FNs due to possible causes related to 1) library derived issues such as coverage, genotype and read quality or a combination of these, and 2) sequencer derived issues such as location in a homopolymer region and signal shifts or both[2]. When we cannot identify the reason behind an FN, we categorised it as unknown. We classified FPs into six categories by inspecting each position on IGV[2]. 1) Strand bias: ≤ 2% of reads of alternate alleles are from one strand, 2) Read end: a variant present within 5 nucleotides at the end of a read, 3) Low quality: the quality of the variant call was less than 20, 4) Homopolymer: a variant inside or next to a repeat stretch of 4 or more nucleotides, 5) Mixed allele: more than one alternate allele was present at that particular position, 6) Unknown: variants failed to be categorised under the above 5 categories[2].

2.5.2.3 Sensitivity and positive predictive value

We reported the sensitivity and PPV to evaluate the performances of libraries. Sensitivity is defined as the proportion of actual true positive variants (truthset variants) that are correctly detected by the library (TP), and PPV is defined as the proportion of total variants detected by the library are true variants (TP)[2].

We did not calculate the specificity of the methods as we do not know the number of true negatives. Only if a variant detected by any of the libraries is negative, but the variant is present in NA12878 truthset, we can see the discrepancy and can call these variants as FNs[2]. The same

applies to those variants which are called by libraries but are not present if truthset (FNs). Therefore, we could not calculate specificity out of the available data but rather provide beside sensitivity the PPV[2].

Formulas for sensitivity and PPV are as follows:

$$Sensitivity = \frac{TruePositives}{TotalPositive}$$

$$PositivePredictiveValue = \frac{TruePositives}{TotalVariants}$$

2.5.2.4 Z score for coverage comparison and evenness of coverage

We divided read depth into 45 categories. For the lower end of read distribution (0-10X), we used an increment of 5X, through 10X - 400X an increment of 10X and above 400X that of 200X[2]. This allowed us a high-resolution comparison of the distribution of reads in the low read depth region (<10X) and in the callable region[2]. The callable range is defined between 5X to 400X and was set by the manufacturer in order to reduce the computation time[2]. Next, we calculated the difference in the coverage between the AmpliSeq and SureSelect for each category. We computed the normalised difference in coverage as follows:

$$Diffcoverage = AmpliSeqcoverage - \text{SureSelectcoverage}$$

$$Z = Diffcoverage - \frac{Mean(Diffcoverage)}{SD(Diffcoverage)}$$

Where SD(Diffcoverage) is the standard deviation of the difference in coverage.

Z > 0 means higher AmpliSeq coverage than SureSelect and vice versa.

We calculated the evenness of coverage for both libraries to compare the target enrichment by dividing the per-base coverage by the average depth.

3 Results

Most of the results presented here are already published as parts of my first author publications[1,2]. In the first part of this section, we presented results from association testing between LTL, and brain ageing phenotypes and in the second part, WES and WGS performed on our study cohorts.

3.1 Part I: LTL and brain ageing

3.1.1 Association of LTL and brain ageing phenotypes within ASPS

LTL, as well as LTL attrition, has been associated with structural brain phenotypes on MRI such as regional and global brain volumes as well as with WMH [111,112]. Results are, however, largely inconsistent regarding the effect of LTL on cognitive functions and decline. The Dallas Heart Study (DHS) investigating 2606 individuals, found no relationship between LTL and cognition while recently 2 large meta-analyses reported that longer LTL is associated with better cognition affecting general cognition [77], memory, speed, and executive function as well as general cognition (g-factor) [113]. Recently, LTL attrition was also shown to be associated with cognitive decline in a longitudinal setting [114]. Importantly, studies studying the effect of LTL on brain structure and functions concurrently at the population level, are so far largely missing.

In this part, we tested our hypothesis that longer LTL is related to 1) better structural preservation of the brain such as larger BPF and less WMH, 2) better cognitive performances including g-factor and composite scores for attention/speed, conceptualisation, memory, visuopractical skills in the ASPS cohort[1]. We used subgroup analyses to test whether the presence of risk factors such as sex, age, hypertension, BMI, education, diabetes, CVD, and *ApoE*4 carrier status modifies the effect of LTL on brain structure and function. We also explored if the observed significant effects of LTL on cognition are mediated by structural brain changes (**Figure 1**).

3.1.1.1 Description of the participants

In total, 909 ASPS participants with LTL (909), MRI (852), and cognitive tests (891) (mean age=65.9 years, men 388(42.7%) were selected [1]. In the study cohort, there were 631(69.4%) hypertensives, 99(10.9%) diabetics 364(40%) with CVD, and 180(20.1%) were *ApoE4* carriers[1].

LTL was negatively (r=-0.092, p=0.006) correlated with age and was significantly lower in the subgroup of participants with CVD (β= -0.136, p=0.046) (**Figure 2**)[1], while we observed no significant difference by sex, hypertension, BMI, diabetes, education and ApoE4 carrier status[1]. Smaller BPF was observed in individuals with age >65years (p=2.92×10^{-42}), hypertension (p=9.19×10^{-7}), DM (p= 7.8 ×10^{-4}) and CVD (p= 0.03)[1]. WMH load was higher in women (p=0.002), >65years (p= 2.7×10^{-40}), hypertensives (p=7.2 ×10^{-14}), in those with ≤10years of education (p=0.02), diabetics (p=0.008), CVD (p=0.001) and in *ApoE*4 carriers (p=0.04)[1]. Similarly, WMH score were significantly higher in the subgroups of women (p=0.03), >65 years (p=1.5×10^{-32}), hypertensives (p=2.7×10^{-9}), diabetics (p=0.01) and CVD (p=0.04) but not in education ≤10years (p=0.3) and *ApoE*4 carriers (p=0.6)[1]. Domain-specific cognitive scores and g-factor were significantly lower in all high-risk groups. There was no significant difference in conceptualisation by sex, for attention/speed and conceptualisation by BMI and visuopractical skills by education (**Table 1**)[1].

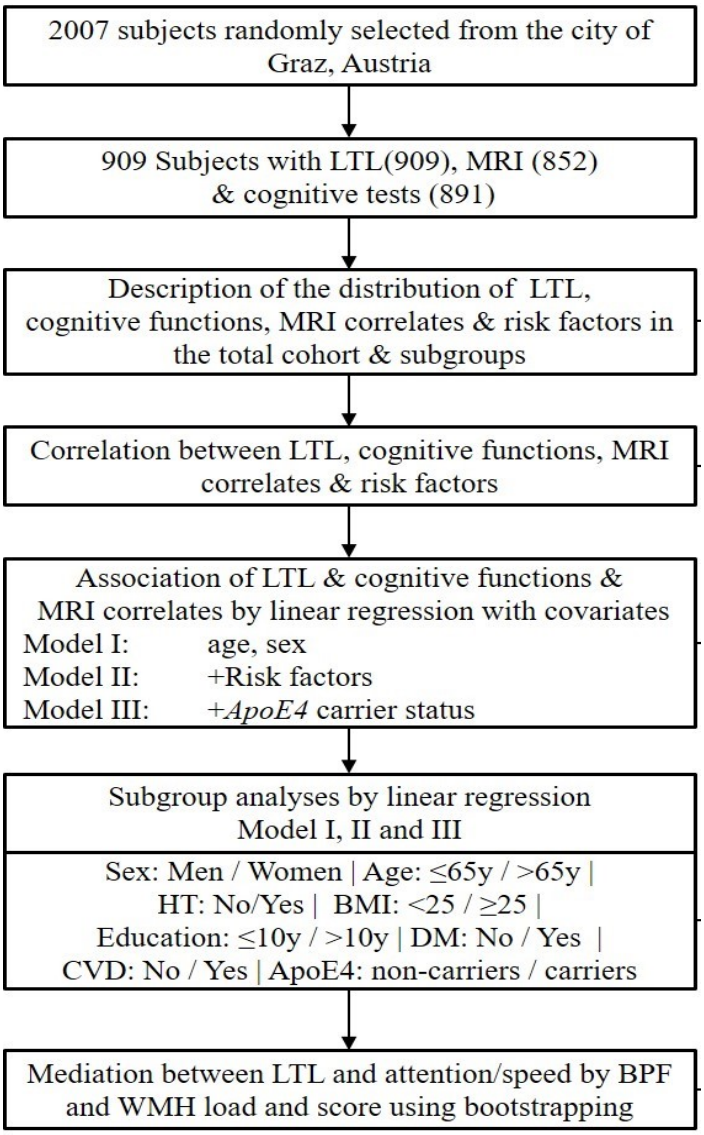


Figure 1: Workflow of the study design. *The current study cohort is a subsample of the Austrian Stroke Prevention Study. Risk factors include hypertension, diabetes, cardiovascular diseases, body mass index, education, high-density lipoprotein, and smoking status. LTL: Leukocyte Telomere Length, HT: Hypertension, DM: Diabetes Mellitus, CVD: Cardiovascular Disease, BPF: Brain Parenchymal Fraction, WMH: White Matter Hyperintensities [Figure reproduced from Gampawar et al., Frontiers in Psychiatry 2020* [1]*].*

	g factor	Visuo practical Skills	Memory	Conceptu alization	Attention/ Speed	LWMH	BPF	HDL	Total Cholester ol	Education	BMI	DBP	SBP	Age	zLTL
zLTL	0.062	0.065	0.024	0.047	0.081*	0.005	0.169***	-0.089**	-0.015	-0.012	.067*	-0.032	-0.022	-0.092**	
Age	-0.593***	-0.626***	-0.428***	-0.217***	-0.448***	0.513***	-0.577***	0.190***	-0.049	-0.107**	0.075*	0.129***	0.347***		
SBP	-0.237***	-0.205***	-0.183***	-0.118***	-0.162***	0.261***	-0.189***	0.035	0.081*	-0.116***	0.227***	0.716***			
DBP	-0.108**	-0.067*	-0.086*	-0.035	-0.108**	0.145***	-0.136***	-0.019	0.123***	-0.032	0.250***				
BMI	-0.113**	-0.117***	-0.097**	-0.067*	-0.070*	0.025	-0.053	-0.258***	-0.027	-0.126***					
Education	0.335**	0.091**	0.373***	0.269***	0.224***	-0.098**	0.023	-0.062	-0.010						
Total Cholesterol	0.032	0.073*	0.025	0.034	0.005	-0.002	0.021	0.223***							
HDL	-0.114**	-0.076*	-0.065	-0.002	-0.122***	0.151***	0-.211***								
BPF	0.370***	0.439***	0.202***	0.073	0.331***	-0.287***									
LWMH	-0.349***	-0.349***	-0.243***	-0.132***	-0.263***										
Attention/Speed	0.640***	0.447***	0.378***	0.253***											
Conceptualization	0.549***	0.269***	0.446***												
Memory	0.806***	0.436***													
Visuopractical Skills	0.703***														
g factor															

Pearson's correlation

-1 -0.9 -0.7 -0.5 -0.3 0 0.3 0.5 0.7 0.9 1

* 0.01<P< 0.05 ** 0.001<P< 0.01 *** < 0.001

Figure 2:Pearson's correlation between two continuous variables*. zLTL: z Score leukocyte telomere length, SBP: Systolic Blood Pressure, DBP: Diastolic Blood Pressure, BMI: Body Mass Index, HDL: High Density Lipoprotein, BPF: Brain Parenchymal Fraction, LWMH: log-transformed White Matter Hyperintensities[Figure reproduced from Gampawar et al., Frontiers in Psychiatry 2020 [1]].*

3.1.1.2 LTL versus MRI correlates and cognitive functions

3.1.1.2.1 Total Cohort

We observed a significant correlation between LTL and BPF (r=0.169, p=3.71×10^{-6}) as well as attention/speed (r=0.0081, p=0.018) (**Figure 2**)[1]. We used 3 linear regression models for studying the association between LTL and MRI as well as cognitive functions. The Model I was adjusted for age and sex (Model I), Model II additionally adjusted for risk factors such as hypertension, DM, CVD, BMI, education, HDL and smoking status and Model III for *ApoE*4 carrier status[1]. LTL was significantly associated with BPF in Model I (β=0.44, p=1.3×10^{-4}), II (β =0.43, p=2.1×10^{-4}) and III (β=0.43, p=2.1×10^{-4})[1]. We observed a significant association between LTL and WMH load as well as the WMH score only in Model II (β=0.02, p=0.04, β=0.05, p=0.04 respectively) and III (β=0.02, p=0.04, β=0.05, p=0.04 respectively)[1]. LTL explained 1.2% of the variation in BPF and 0.4% of the variation in WMH load and score[1]. In none of the multivariable models, LTL had a significant association with any of the cognitive measures (**Table 2**)[1].

3.1.1.2.2 Subgroups

We further studied association by repeating the linear regression models in subgroups divided by sex, age, hypertension, BMI, education, DM, CVD, and by *ApoE4* carrier status[1]. We observed a highly significant effect of LTL on BPF in the subgroups of women (β=0.51, p=0.001), individuals with age >65years (β=0.58, p=0.002), hypertension (β=0.51, p=7.0×10^{-4}), BMI≥25(β=0.40, p=0.004), education≤10years (β=0.42, p=0.002), CVD (β=0.33, p=0.02), in non-diabetics (β=0.42, p=4.0×10^{-4}) and *ApoE4* non-carriers (β=0.49, p=1.1×10^{-4})[1]. The

significant effect of LTL was present on both WMH load and score in subgroups of education ≤10years, non-diabetics (β=0.03, *p*=0.04, β=0.03, *p*=0.008; β=0.06, *p*=0.02, β=0.06, *p*=0.02) and *ApoE4* non-carriers (β=0.03, *p*=0.02, β=0.07, *p*=0.01)[1]. On WMH load, the effect was in addition significant in hypertensives (β=0.04, *p*=0.02) and on WMH score in subjects without CVD (β=0.06, *p*=0.05)[1]. LTL was significantly associated only in the attention/speed cognitive domain within the subgroup of those with BMI≥25 (β=0.04, *p*=0.05) and ≤10years of education (β=0.04, *p*=0.05)[1]. We formally tested the interaction between LTL and the risk factors on MRI and cognitive phenotypes using interaction terms. However, none of the interaction terms reached statistical significance (*p*>0.05) (**Table 3**)[1].

We tested whether BPF mediates the significant effect of LTL on attention/speed in model III in the subgroup of BMI≥25 and education ≤10years or WMH load/score in the subgroup of education ≤10years using bootstrapping. We found significant mediation by BPF in both subgroups (β=0.02, 95%CI 0.01-0.03 for both subgroups) while the mediation effect by WMH load/score was not significant (load: β= -0.01, 95%CI -0.02-0.001; score: β= -0.03, 95%CI -0.01-0.001) (**Figure 3** and **Table 4**)[1].

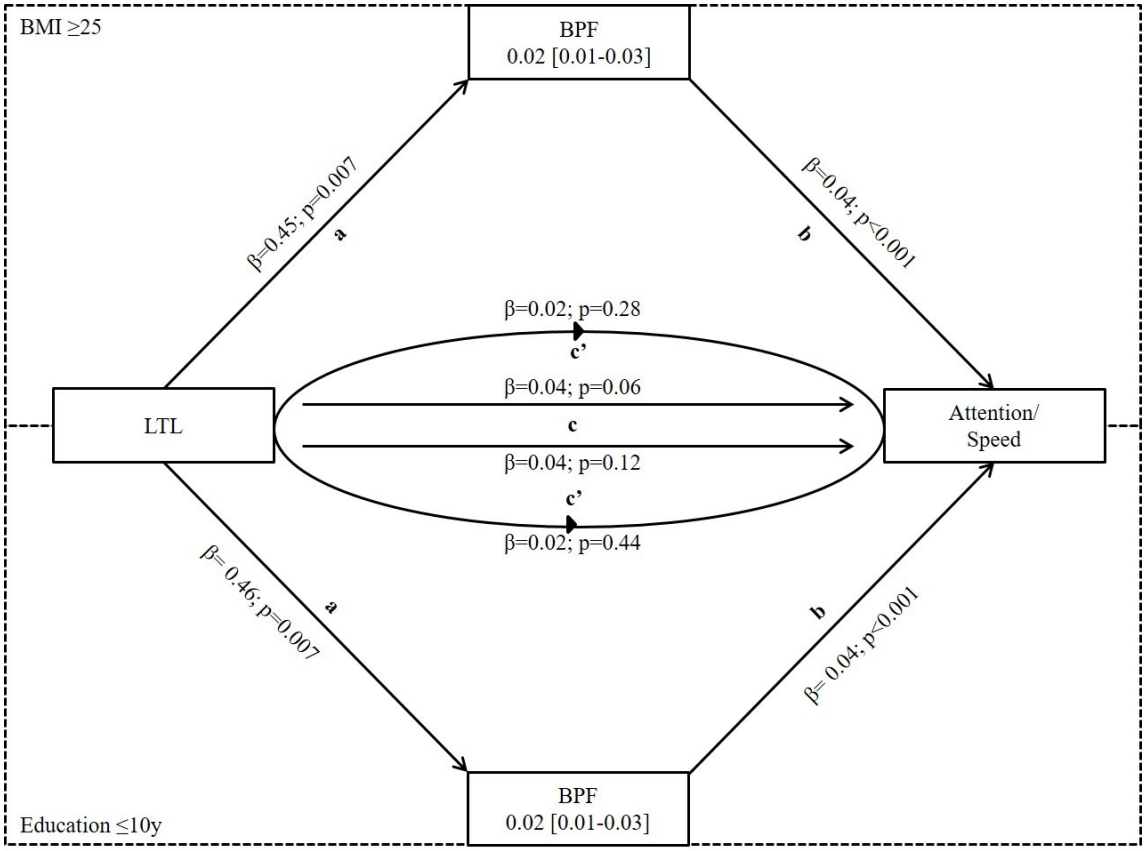


Figure 3:Mediation analysis of LTL on attention/speed via BPF in subgroups of BMI≥ 25 and Education ≤ 10y where LTL was significantly associated with both BPF and attention/speed. *Mediation is calculated after adjusting for covariates from Model III of linear regression. a- effect of LTL on BPF, b-effect of BPF on attention/speed, c-total effect of LTL on attention/speed, and c'-direct effect of LTL on attention/speed when controlled for BPF. It shows partial mediation because total effect (c) of LTL on attention/speed is reduced in the presence of BPF (c'). LTL: Leukocyte Telomere Length, BPF: Brain Parenchymal Fraction. [Figure reproduced from Gampawar et al., Frontiers in Psychiatry 2020 [1]]*

3.1.2 WGS derived telomere length and brain ageing

Use of NGS in research settings has increased tremendously due to the decrease in its cost, which in turn led to the development of many bioinformatic tools that could estimate telomere lengths from WGS data[115]. Many studies have used WGS based telomere length in the study of depression, cancers[115], and GWAS of telomere length from 75,000 WGS samples[98]. In this part, we compared Computel[95] and TelSeq[96] two telomeric reads extraction tools from WGS data and estimated LTL from GSHA participants.

3.1.2.1 Comparison of Computel and TelSeq

TelSeq required 2-3 hours, whereas Computel 8-10 hours to extract LTL from a single sample. We used both the software on 5 GSHA samples, and the results are shown in **Table 5**. TelSeq estimated higher LTL than Computel however, the estimates are highly correlated (r =0.996).

3.1.2.2 Analysis of GSHA

In GSHA, there were 90 participants with WGS data available. The mean age of the participants was 67.6±9.0 years, and 33 (36.7%) were men. The correlation between age and LTL was negative (-0.272) and significant (p=0.010). The correlation was stronger in women and significant (r= -0.294, p = 0.026) as compared to nonsignificant correlation in men (r= -0.220, p = 0.218) (**Figure 4**).

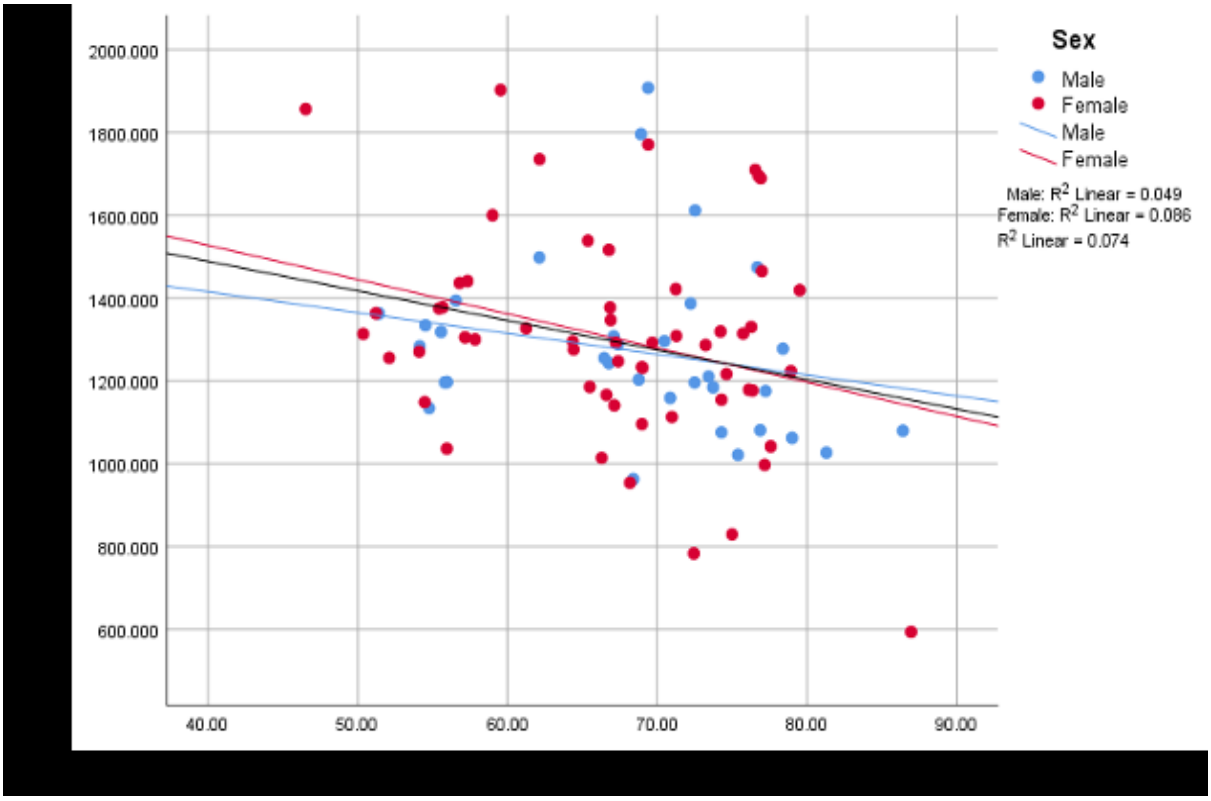


Figure 4:Correlation of LTL with age. Correlation of LTL with age. *LTL shows a decrease with increase in age*

3.2 Part II: Establishing methods: WES and WGS

These results are already published as a part of my first author publication[2].

3.2.1 Establishing WES for Ion Proton

Library preparation is the critical step in NGS workflow as target regions from the genomic DNA are selected either by hybridisation using probes or amplification by primers. Most of the library preparation methods are available for Illumina platforms as Illumina has the largest NGS market[116]. For the Ion Proton platform, only two WES library preparation methods are available; however, studies evaluating their performances are so far missing in the literature. Previously, studies compared the performance of AmpliSeq on Ion platforms with various WES library preparation methods available for Illumina platforms[117–119] and reported that AmpliSeq on Ion platforms is a faster method with high throughput but faces problem in complex genomic regions. However, to our knowledge, no studies have compared AmpliSeq and SureSelect for Ion Proton. This part is aimed 1) comparing the performance of AmpliSeq and SureSelect and 2) developing an optimised protocol for variant calling for WES on Ion Proton platform[2].

We utilised 12 in-house DNAs with genome-wide and exome chip genotyping data to calculate concordance rates between variants detected by genotyping and sequencing and NA12878, a well characterised reference DNA. We validated our sequencing protocol against NA12878 v3.3.2 reference calls by documenting sensitivity and PPV at each optimisation step [100,104] (**Figure 5**)[2]. We also extend previous findings of AmpliSeq WES library on Ion Proton by evaluating different target regions, coverage ranges (44x to 270x) using wet lab sequencing and by manually inspecting all FNs and FPs variants Chr 1,7,16,19 and X and categorising them based on their possible causes[2].

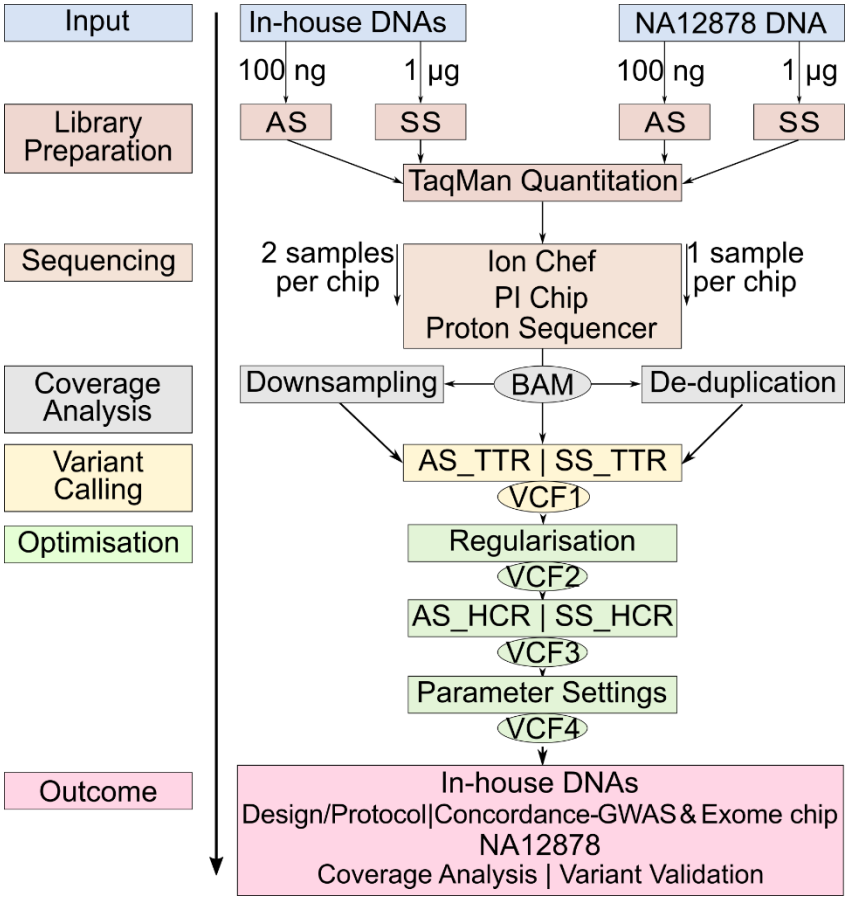


Figure 5:Workflow of the study design. *The same colour represents the steps at the same level. Identical steps are used to analyses both methods. AS-AmpliSeq, SS- SureSelect, TTR-Total Target Region, ETR-Effective Target Region, OTR- Overlapping Target Region, TPs- True Positives, FNs- False Negatives, FPs- False Positives, PPV- Positive Predictive Value. [Figure reproduced from Gampawar et al., 2019 Frontiers in Genetics [2].]*

3.2.1.1 Comparison of AmpliSeq and SureSelect laboratory protocol and design

An in-detail analysis is presented in **Table 6.** AmpliSeq is a PCR based library preparation method, amplifies the target regions from genomic DNA using primers[2]. It requires 100 ng of DNA to start with and finishes library preparation in around 6 hours[2]. On the other hand, SureSelect is a hybridisation-based library preparation method, uses RNA library baits to select the targets from the enzymatically fragmented 1 µg of genomic DNA[2]. The library preparation takes around 2.5 days.

By design AmpliSeq TTR targets 57,742,646 bps and SureSelect TTR 60,456,963 bps[2]. The overlap between AmpliSeq TTR and SureSelect TTR, henceforth referred as overlapping target region (OTR), is 43,173,762 bps[2]. AmpliSeq TTR covers 91.1%, 88.6% and 87.9% while SureSelect TTR 87.9%, 87.8%, and 87.4%, of RefSeq, Ensembl and UCSC defined coding regions, respectively[2]. The size of AmpliSeq ETR is 46,347,343bps and covers 86.3%, 83.7%

and 83.1% of the RefSeq, Ensemble and UCSC coding regions[2]. From RefSeq coding region, AmpliSeq design missed 3,016,767 while SureSelect missed 4,227,905 bases[2].

3.2.1.2 Analysis of in-house DNA samples

3.2.1.2.1 Target enrichment efficiency and variant detection

On average AmpliSeq produced had 34.2 million and SureSelect 39.8 million reads[2]. AmpliSeq had the mean read depth of 92X as compared to 69X for SureSelect[2]. In total, 94% of the total reads produced by AmpliSeq and 86% by SureSelect were mapped to their respective target regions, a statistically significant difference (p<0.0001)[2]. The percentage of bases covered >50X and >5X were 63% and 97.3% with AmpliSeq and 50.3% and 97.9% with SureSelect (Supplementary table 1)[2]. The average number of variants over all samples were 51,413 in AmpliSeq and 51,783 in SureSelect TTR under default low stringency settings of TVC (Supplementary table 2). Averagely in AmpliSeq ETR, we found 37,268 variants, 14,145 variants less than that of AmpliSeq TTR[2].

3.2.1.2.2 Microarray genotypes

3.2.1.2.2.1 Exome chip

The microarray genotype concordance analysis was done in TTR of respective libraries[2]. In AmpliSeq out of ~7,004 variants detected by both sequencing and exome chip, ~6,862 variants were concordant[2]. Similarly, in SureSelect, out of ~7,343, ~7,166 variants were concordant[2]. On an average, approximately 300 more variants were both, detected and concordant, between SureSelect and exome chip than that of AmpliSeq and exome chip[2]. However, the mean concordance rate was around 98% with exome chip for both libraries. For this comparison, we used nine samples as data from two exome chip samples (801058 & 801020) was not used due to its low quality and sample 13 did not have exome chip data (Supplementary table 3)[2].

We categorised the variants located within the exomic regions identified by both exome chip and sequencing using the genome aggregation database into variants with frequency of 1) > 5%, 2) 5% - 1%, 3) 1% - 0.05% and 4) < 0.05%[2]. Concordance rates were similar over the first three categories (>98%) but substantially lower among variants with a frequency of <0.05% (AmpliSeq: 95%, SureSelect: 92%)[2].

3.2.1.2.2.2 GWAS chip

Similar to exome chip genotypes, in AmpliSeq, fewer variants were detected (~3,664) and concordant (~3,646) between sequencing and GWAS chip[2]. In SureSelect, ~4,910 variants were detected by both sequencing and GWAS chip out of which ~4,883 were concordant[2]. There

were approximately 1,237 more concordant variants between sequencing and GWAS chip detected by SureSelect than AmpliSeq. Both the methods had an excellent concordance rate of >99% across all 12 samples (Supplementary table 3)[2]. For these comparisons, we used only homozygous variant or heterozygous variant calls, and we disregarded homozygous reference calls[2].

3.2.1.3 Analysis of NA12878 reference DNA

3.2.1.3.1 Coverage

The average depth of coverage was 270X for AmpliSeq and 115X for SureSelect[2]. From 57.74 million bases (MB) targeted by AmpliSeq TTR, 0.40 MB were covered with > 5X, 45.85 MB between 5X to 400X, and 11.94 MB with > 400X coverage amounting to 0.69%, 79.40% and 19.91% respectively[2]. Similarly, out of 60.46 MB targeted by SureSelect, 0.37 MB covered with < 5X, 59.54 MB between 5X to 400X and 0.90 MB with >400X coverage corresponding to 0.61%, 98.49% and 0.90% respectively (Supplementary table 4)[2]. Both libraries covered approximately 40% of the total targeted bases with more than the average depth (**Figure 6**)[2]. However, in AmpliSeq percentage of bases covered less than 10X and >150X were fairly higher than that in SureSelect (**Figure 6** B & C)[2].

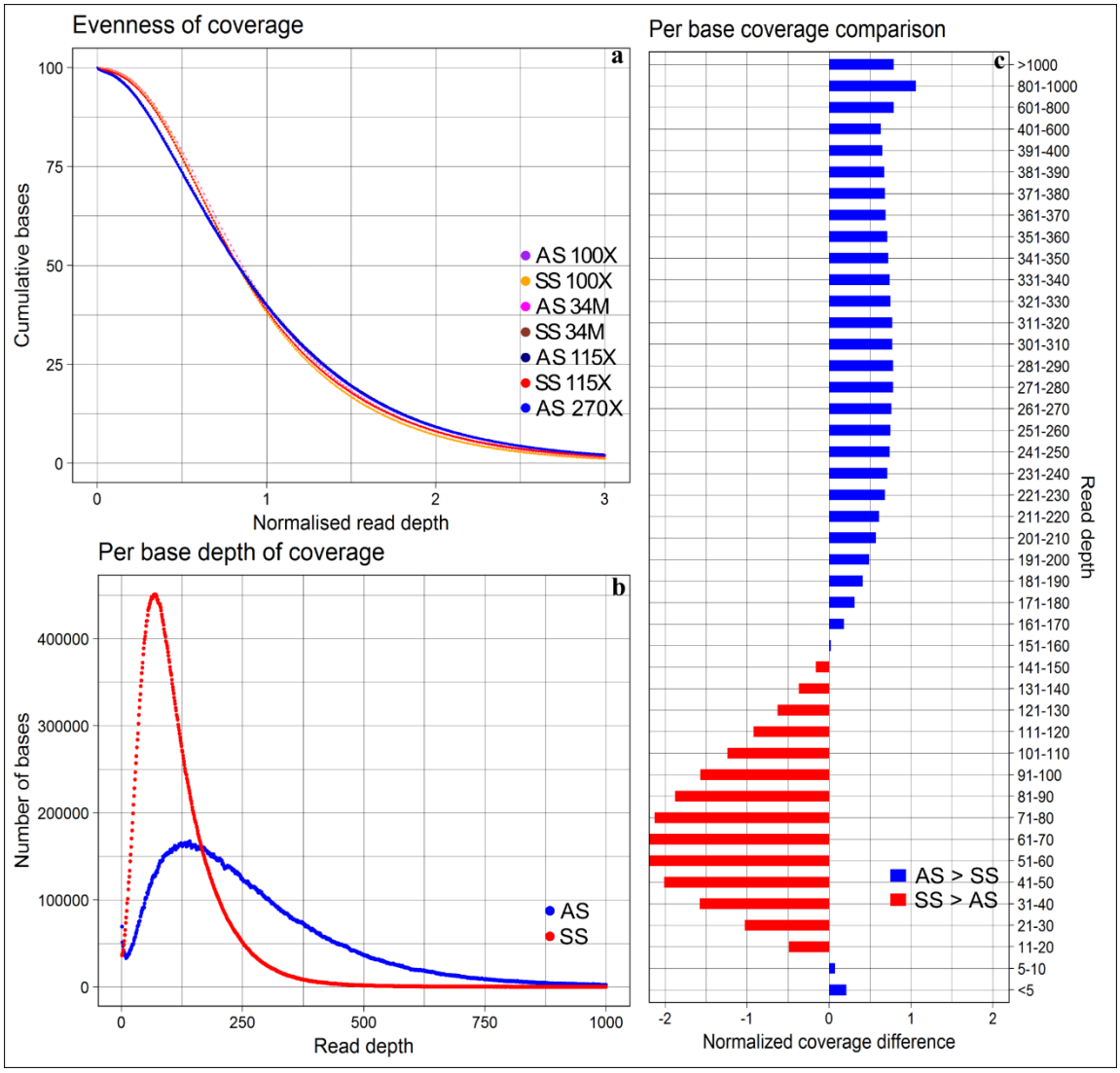


Figure 6: Evenness of coverage, per base depth of coverage and its comparison between AmpliSeq and SureSelect methods. *(a) Evenness of coverage plotted for original and downsampled BAM files (b) Scatter plot showing the distribution of per base coverage of AmpliSeq and SureSelect till 1000X read depth. (c) A bar chart is showing the difference in coverage after dividing the depth of coverage into 45 groups and normalization. SureSelect covers more bases in the coverage range of 11X to 150X than AmpliSeq. AS-AmpliSeq, SS-SureSelect [Figure reproduced from Gampawar et al., 2019 Frontiers in Genetics [2]].*

3.2.1.3.2 Variant detection

By using AmpliSeq, we identified 54,351 variants (VCF1) whereas truthset had 49,340 variants[2]. Out of the 54,351 variants detected by AmpliSeq, 45,946 were TP, and 8,405 FPs[2]. The overall sensitivity and PPV of AmpliSeq were 93.1% and 84.5%[2]. In case of SNVs and exonic SNVs, AmpliSeq detected 45,092 and 16,964 variants out of which 43,840 and 16,588 were TP, respectively[2]. Nevertheless, only 2,106 total indels were detected correctly from 4,248

and 231 exonic indel from 329 by AmpliSeq. The overall sensitivity and PPV for detecting SNVs were higher than for detecting indels[2].

By using SureSelect, we identified 54,934 variants out of which 43,929 were TP, and 11,005 FP while truthset had 46,982 variants[2]. The sensitivity and PPV for SureSelect were 93.5% and 80% for SureSelect. The number of true positive total SNVs, exonic SNVs, total indels and exonic indels were 42,230, 15,846, 1,699 and 195 respectively[2]. The sensitivity and PPV for detecting SNVs were higher indels same as that of in AmpliSeq (**Table 7**)[2].

3.2.1.4 Optimisation-variant calling pipeline

The optimisation was consisting of three steps, namely regularisation, use of HCR and fine-tuning by parameter settings[2]. Upon regularisation, as recommended by Zook et al.,[100] in AmpliSeq, the number of TPs increased by 714 to 46,660 and in SureSelect by 622 to 44,551 (VCF2)[2]. At the same time, FPs increased by 176 to 8,581 and by 275 to 11,280 respectively[2]. Use of HCR BED file excluded difficult-to-sequence target regions leading to a significant reduction in FPs from 8,581 to 1,218 in AmpliSeq and from 11,005 to 947 in SureSelect while the number of TPs decreased minimally by 0.7% (VCF3) (**Table 8**)[2]. The sensitivity and PPV for detection of SNVs were 98.7% and 98.3% for AmpliSeq and 98.8% and 98.6% for SureSelect respectively[2]. Corresponding values for indels were 52.7% and 82.7% for AmpliSeq and 49.1% and 84.4% for SureSelect (**Figure 7** and Supplementary table 5)[2].

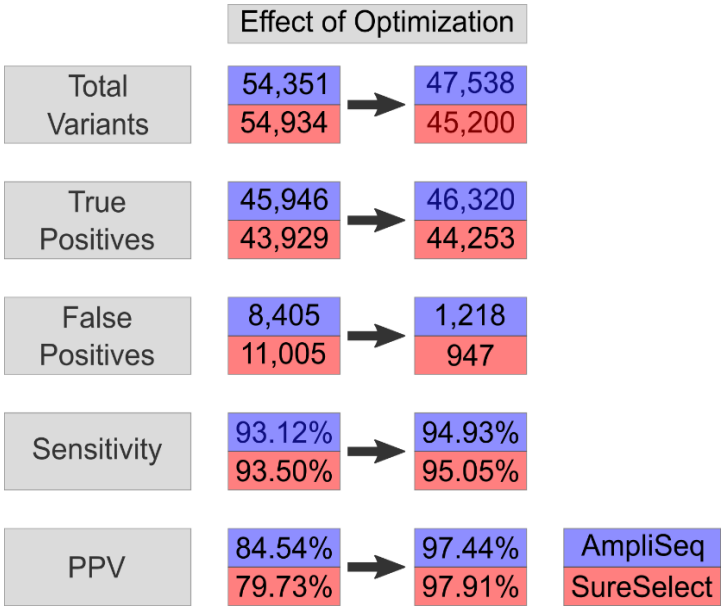


Figure 7:Effect on optimisation -variant calling pipeline *.Effect of optimisation steps shown on total variants, true positives, false positives, sensitivity and PPV in AmpliSeq and SureSelect. Blue represents AmpliSeq and red SureSelect. PPV: positive predictive value [Figure reproduced from Gampawar et al., 2019 Frontiers in Genetics [2]].*

The total number of TP, FN, and FP indels were 2,121, 1,904 and 445 by AmpliSeq and 1,710, 1,771 and 335 in SureSelect respectively[2]. Out of these, 870, 642, and 31 TP, FN, and FP indels were missed by both libraries respectively[2]. Out of 1,261 indels missed by AmpliSeq, SureSelect detected 132, while out of 1,128 indels missed by SureSelect, AmpliSeq detected 247(Supplementary table 6)[2]. Finally, we stepwise changed default parameters settings and repeated variant calling only using HCR as a target region. AmpliSeq had the best balance between TP, FN and FPs when the parameter "minimum allele frequency" was changed to 0.2 (step 2) while SureSelect had the best performance with the default parameter settings[2].In AmpliSeq, PPV improved from 97.4% to 98.1% with a reduction of sensitivity by 0.3% (Supplementary table 7 and **Figure 8**)[2].

3.2.1.5 Downsampling

We equalised read depth to 34 million reads and to an average depth of 100X by performing downsampling to reduce the coverage bias. Repeating variant calling on 34 million randomly selected reads from both libraries resulted in 53,068 and 52,918 variants within AmpliSeq and SureSelect TTRs[2]. AmpliSeq had a sensitivity of 91.8% and PPV of 85.3% whereas SureSelect had a sensitivity of 91% and PPV of 80.8%. When variant calling was performed on 100X average depth downsampled libraries, AmpliSeq detected 53,121 and SureSelect 54,612 variants[2]. The proportion of the increase in detection of total variants was also seen into TP detected by both libraries. The sensitivity of AmpliSeq was 91.8% while that of SureSelect increase to 93.1%. PPV of AmpliSeq was 85.3% and of SureSelect was 80.1% (Supplementary table 8)[2]. Application of our optimisation pipeline for variant calling on downsampled libraries, we observed minimum increase in sensitivity and a significant increase in PPV to 98% similar to a similar increase observed in original libraries (AmpliSeq 270X and SureSelect 115X) (Supplementary table 8)[2].

3.2.1.6 AmpliSeq ETR

AmpliSeq detected a total number of 38,651 variants in ETR out of which 33,119 were TPs, 1,251 FNs, and 5,532 FPs2. Performing regularisation and restricting variant calling to HCR resulted in a substantial reduction of FPs by 91% and increased PPV to 98.5% while maintaining sensitivity at (Supplementary table 9)[2]

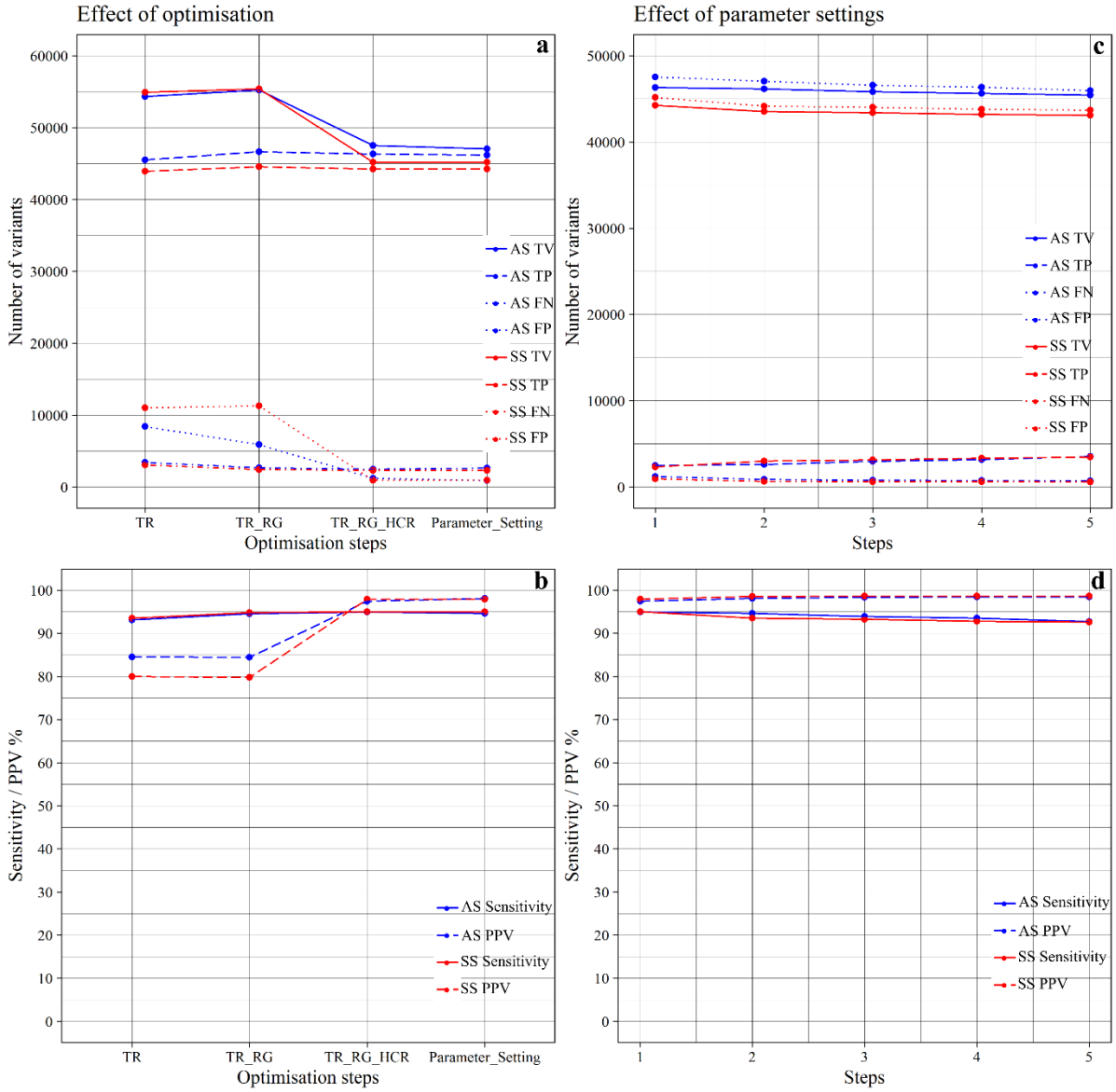


Figure 8:Effect of stepwise optimisation strategies on AmpliSeq (blue) and SureSelect (red) performance*. Effect of 3 step optimization on a) total variants (TV), true positives (TP), false positives (FP) and false negatives (FN) b) on sensitivity and positive predictive value (PPV). Effect of low to high stringency 5 step parameter setting on the variants is shown in c and d. AS-AmpliSeq SS-SureSelect [Figure reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]].*

3.2.1.7 Variant detection in RefSeq coding region and overlapping target region

To compare the efficiency of two methods for sequencing of the human protein-coding region of the genome, we used transcripts of NCBI RefSeq datasets. RefSeq coding region had a total of 19,270 variants in NA12878 truthset, from which AmpliSeq detected 17,836 and SureSelect 17,312[2] . Regularisation and use of HCR had a similar effect here too, reducing the number of FPs in AmpliSeq from 3,748 to 322 and in SureSelect from 4,337 to 304[2]. Due to optimisation, sensitivity increased slightly from 92.6% to 93.4% for AmpliSeq and from 89.8% to 90.5% for

SureSelect whereas PPV increased considerably from 82.6% to 98.1% and 80% to 98.1%, respectively (**Table 9**)[2].

We compared the performance of both the libraries at 115X average depth in 43.2MB OTR and saw a similar improvement like that of TTR or RefSeq coding region[2]. The sensitivity of both methods was around 95%, and PPV was improved from 85% to 98% (**Table 9**)[2]. In OTR, out of the total TPs called by each method, 30,266 were shared, leaving 1-2% of variants specific to each library (**Figure 9**)[2].

3.2.1.8 Duplicate removal

Removal of duplicate sequencing using Picard and samtools resulted in an 88% loss of reads in AmpliSeq and 30% loss in SureSelect[2]. Therefore, we did not perform variant calling in AmpliSeq. Using the tool "MarkDupbyStartEnd", the loss was 13% in AmpliSeq and 0.1% in SureSelect[2]. Although in AmpliSeq duplicate removal reduced the number of FPs to 1,111 from 1,218 (TTR), TP variants were also reduced by 711 to 45,609 leading to a slight decrease in sensitivity by 1.5% with negligible improvement in PPV[2]. In the case of SureSelect, there was no markable change in the performance by removing duplicates by any of the strategies and sensitivity remained around 94% and PPV 98% (Supplementary table 10)[2].

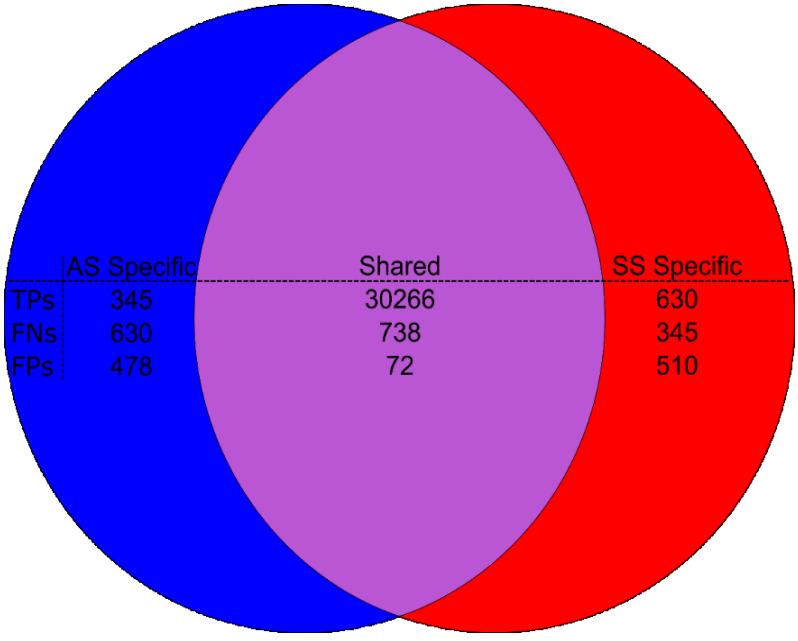


Figure 9:Number of variants shared between two libraries and specific to each library in the overlapping target region *In the overlapping target region of 43.17 million bases, both libraries shared 98% of variants detected. We performed this analysis on after equalizing depth of coverage to 115X for both libraries [Figure reproduced from Gampawar et al., Frontiers in Genetics 2019[2]].*

3.2.1.9 Exploration of FNs and FPs

On manual inspection of all FNs on Chr 1,7,16,19 and X (**Figure 10**) we found that the library derived issues (74-95%) were mainly responsible for FN SNVs due to whereas library derived, sequencer derived or both issues (14-58%) were responsible for FN indels in both libraries (Supplementary table 11-17)[2]. We scrutinised FN positions shared by both methods to validate our classification of FNs (Chr 1: 35 and Chr X: 18) and found that, except for two positions, the classification was concordant. Strand bias was the primary cause for classifiable FPs SNVs in AmpliSeq (51-61%), (**Figure 11**)[2] while in SureSelect homopolymers played a prominent role (18-50)[2]. Homopolymer related issues explained most FP indels (44-79%) in both methods (Supplementary table 18-24). There were no significant differences between the causes of FNs or FPs in the respective TTRs or library-specific regions in either library[2].

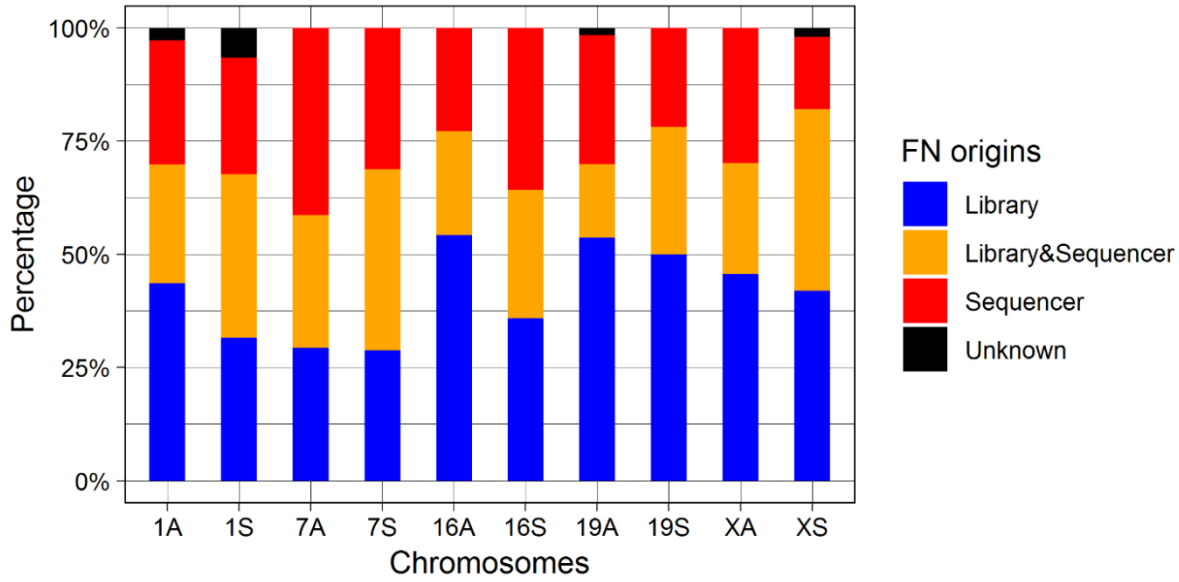


Figure 10:Comparison of the percentage of false negatives in a library derived, sequencer derived, both, and unknown categories. *Each bar represents a chromosome and A or S represent AmpliSeq or SureSelect. On each bar, 4 colours represent the percentage of false negatives in 4 categories [Figure reproduced from Gampawar et al., Frontiers in Genetics 2019[2]].*

3.2.1.10 Effect of increasing average read depth on AmpliSeq performance

Increasing average depth from 44X to 270X led to decrease in bases covered < 5X (3,188,163 to 400,116) but on the same time also to a disproportional increase in the number of bases covered > 400X (19,602 to 11,494,834)(**Figure 12**)[2]. The increase in read depth resulted in a relative decrease in the number of bases in the callable range (5X-400X from 55,687,667 at

44X to 45,880,764 at 270X)[2]. Sensitivity increased significantly from 86.2% to 94.9% while the change in PPV was negligible (96.9% to 97.4%) (Supplementary table 25)[2].

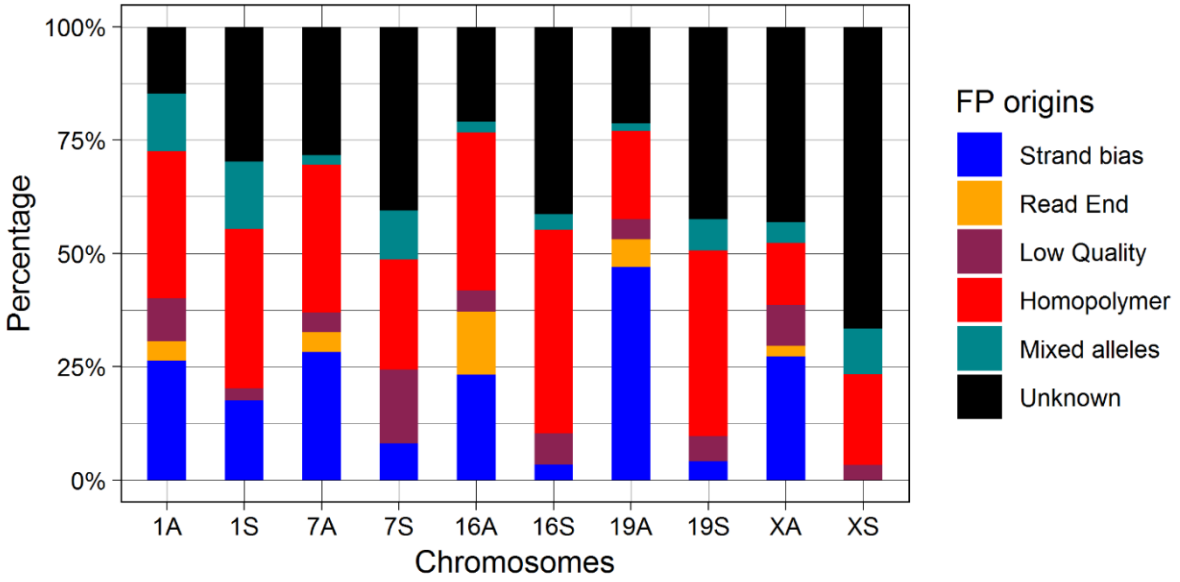


Figure 11:Comparison of the percentage of false positives in strand bias, read end, low quality, homopolymer, mixed alleles and unknown categories*. Each bar represents a chromosome and A or S represent AmpliSeq or SureSelect. On each bar, 6 colours represent the percentage of false negatives in 6 categories [Figure reproduced from Gampawar et al., Frontiers in Genetics 2019[2]].*

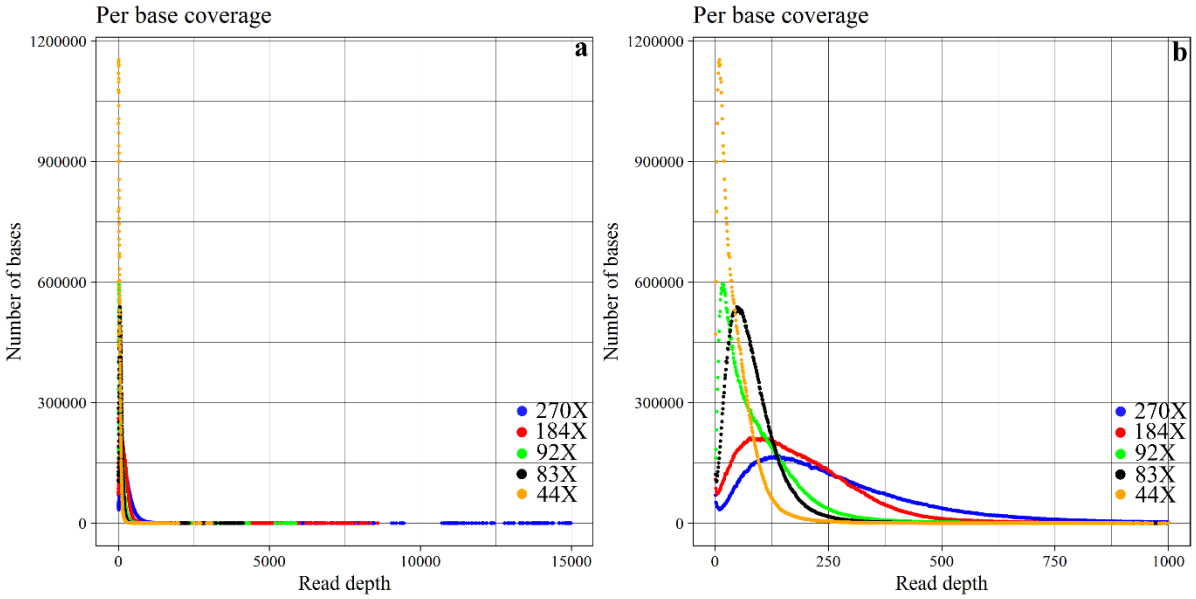


Figure 12:Comparison between AmpliSeq runs at a different average depth of coverage. . *(a)Scatter plot showing the distribution of per base coverage of 5 sequencing run of AS over total read depth. The average depth of coverage of 270x (blue), 184X (red), 92X (green), 83X (black) and 44X (yellow) are seen. (b) Scatter plot showing the distribution of per base coverage of AmpliSeq and SureSelect till 1000X read depth. With the increase in average depth of coverage more than 100X, the proportion of a relative number of bases covered with more than 400X increases more than between 5X to 400X, leading to increase in average depth of coverage [Figure reproduced from Gampawar et al., Frontiers in Genetics 2019[2]].*

3.2.2 Whole Genome sequencing

WGS was performed on 147 samples at the McGill Genomic Centre as decided in consent by the participating cohorts in the BRIDGET Consortium. The average read depth of WGS was 37.1±2.9X. Due to unforeseen technical problems at the McGill Genomic Centre, WGS BAM and VCF files were transferred to the participating groups with substantial delay. This led to a prolongation of the project running time by >1year. The QC and the joint variant calling within the consortium was finalised by the end of 2019. Therefore, our previously laid aim to identify rare genetic variants associated with MRI markers/cognition using generated WES and WGS data could not be completed so and is not part of this thesis. However, data analyses have been initiated, and first results with the contribution of the Graz samples are expected by the end of 2020.

4 Discussion

The general aim of this thesis was to contribute to a better understanding of the genetic architecture of brain ageing and thus shed light on possible molecular mechanisms underlining the structural and functional decline of the ageing brain ultimately resulting in reduced quality of life, disabilities and severe clinical manifestations such as stroke and dementia. This understanding is the fundament of developing a novel preventive and therapeutic measures to slow down the process and reduce the risk of stroke and dementia. The thesis followed up on two specific aims namely 1) exploring the role of biomarkers with the focus on LTL and DNA methylation in brain ageing and 2) establishing resources for the identification of genomic predictors by using WES and WGS at the local level as well as in the frame of international collaborations.

4.1 The setting

Our studies were conducted within the ASPS, ASPS-Fam and the GSHA, three extensively phenotyped (>5000 variables/proband) community-based local cohorts on brain ageing and ageing in general with a partly longitudinal design over a follow-up time of 3-20 years. The three local cohorts are embedded in several international collaborative networks. Especially the Cohorts for Heart and Ageing Research in Genomic Epidemiology (CHARGE) and the Joint Programme in Neurodegenerative Diseases funded BRain Imaging, cognition, Dementia and next generation GEnomics: a Transdisciplinary approach (BRIDGET) consortia are to mention, as here the interactions are especially intensive and fruitful. CHARGE and BRIDGET are collecting the leading cohorts in the field such as Framingham Heart Study, Rotterdam Study, Cardiovascular Health Study, Study of Health in Pomerania. They not only allow for high impact publications by increasing the sample size but most importantly to an efficient transfer of knowledge among the groups and a prompt reaction to the fast-evolving fields of genomics, bioinformatics, MRI technologies and statistical approaches, to mention only a few. A clear strength of the thesis is its foundation on the combination of the local resources with international collaborations, which promoted the development of novel sequencing protocols, and opens up the perspectives to follow-up on the findings on LTL and brain ageing in a much larger sample and using advanced MRI technologies.

4.2 The role of LTL in brain ageing

4.2.1 Association of LTL and brain ageing phenotypes within ASPS

Our results showed a highly significant association between LTL and BPF independent of age, sex, the presence of vascular risk factors, and *ApoE4* allele in the elderly[1]. The observed effect was significantly stronger in women and individuals with a higher risk for cerebrovascular disease and dementia such as those older than 65 years, hypertensives, overweight (BMI≥25), having education less than 10 years, and CVD. The effect of LTL on BPF was significantly stronger in *ApoE4* non-carriers[1]. We did not observe any significant effect of LTL on cognition except for the domain of attention/speed in individuals having education less than 10 years and overweight. The association between LTL and attention/speed in these subgroups was significantly mediated by in these subgroups[1].

4.2.1.1 LTL and MRI correlates

There are contradictory results from studies which investigated brain size in relation to LTL. The largest multi-ethnic study, the DHS, reported a highly significant association with total cerebral volume and volume of the cerebral white and grey matter[111]. In their cohort, LTL explained 1.3% of the cerebral volume, which is very close to our finding of LTL explaining 1.2% of BPF (partial R^2=1.2%) (**Table 2**)[1]. This corresponds to around 1/20 of the proportion explained by age (partial R^2=20 in DHS and R^2=24 in ASPS) in both studies. This pinpoints to a robust and reproducible effect of LTL on brain size, even in diverse populations[1]. The effect size of LTL on BPF was approximately twice as large as in their counterparts in women, individuals older than 65 years, hypertensives, and those with CVD[1]. The especially strong effect of LTL on BPF in the high-risk groups in our cohort suggests a crucial protective role for longer telomeres when the brain is exposed to damaging factors[1]. The Swedish subsample of CASCADE cohort reported similar subgroup specific results on subcortical atrophy where the effect of LTL was stronger in those above the median age of 69.6 years [120]. Importantly, in the case of *ApoE*, those who did not carry the E4 allele, the effect of LTL on BPF was more than three times as high (p=0.0001) as in carriers (p=0.6). This was a surprising finding and might indicate a different mechanism linking LTL to BPF as in the traditional high-risk groups[1].

We noticed an unexpected positive association between LTL and WMH load as well as score in the total cohort as well in some subgroups suggesting longer telomere length was associated with more lesions[1]. The significance and the effect sizes were moderate but became more significant and larger after adjusting for vascular risk factors and *ApoE4* carrier status in addition to age and sex[1]. These findings are contrary to reports from studies on American

Indians[112] and the Swedish subsample of CASCADE[120], which reported a significant association between shorter LTL and increased WMH load. We did not see the differential effect by age; instead, the association was significant within hypertensives, non-diabetics and those with education less than 10 years[1]. Most likely, this difference was at least partly due to differences in sample sizes. We observed different effect sizes for WMH load in hypertensives vs normotensives and *ApoE4* non-carriers vs carriers[1]. Presently we cannot give a biological explanation for the detrimental effect of longer telomeres on WMH, however, it might be a chance finding as this significance diminishes after correcting for multiple testing.

4.2.1.2 LTL and cognition

In our cohort, we observed a very specific effect of LTL on cognition with longer telomeres were associated with better performances only in attention/speed domain in those with overweight and education less than 10 years[1]. Previously, population-based cohort studies reported divergent findings on the association between LTL and cognition, varying from no association in 2606 participants of DHS[121] to a significant association in meta-analyses of four prospective cohorts (N=5955) [122]. However, a significant association of longer LTL to better general cognition in this meta-analyses became insignificant adjusting for risk factors. The meta-analyses so far performed in European ancestry cohorts (N=17052) reported longer telomeres are associated with better cognitive performance, including memory, executive function, and importantly similarly to our study with speed[113].

A previous longitudinal study of the ASPS cohort reported that BPF loss was a predictor of worse performance in memory, visuopractical skills and attention/speed domain. The study also showed that the association between WMH progression and cognitive decline diminishes when controlling for brain volume, a finding suggesting that cognitive decline relates directly to the loss of brain substance rather than to the progression of lesion burden[25]. Therefore, in the future, we intend to test the hypothesis that the observed protective role of longer LTL on the brain parenchyma also translates to better performance in other cognitive domains. Clearly, larger and preferentially longitudinal studies are needed to investigate the connections between LTL, brain size, cognition and the interplay with risk factors.

4.2.1.3 Effect mediation by BPF on attention/speed

The effect of LTL was significant on both BPF and attention/speed in subgroups of BMI≥25and education<10 years. Within these subgroups, BPF mediated about half of the effect of LTL on cognition. This finding for the first time support the hypothesis that longer telomeres protect

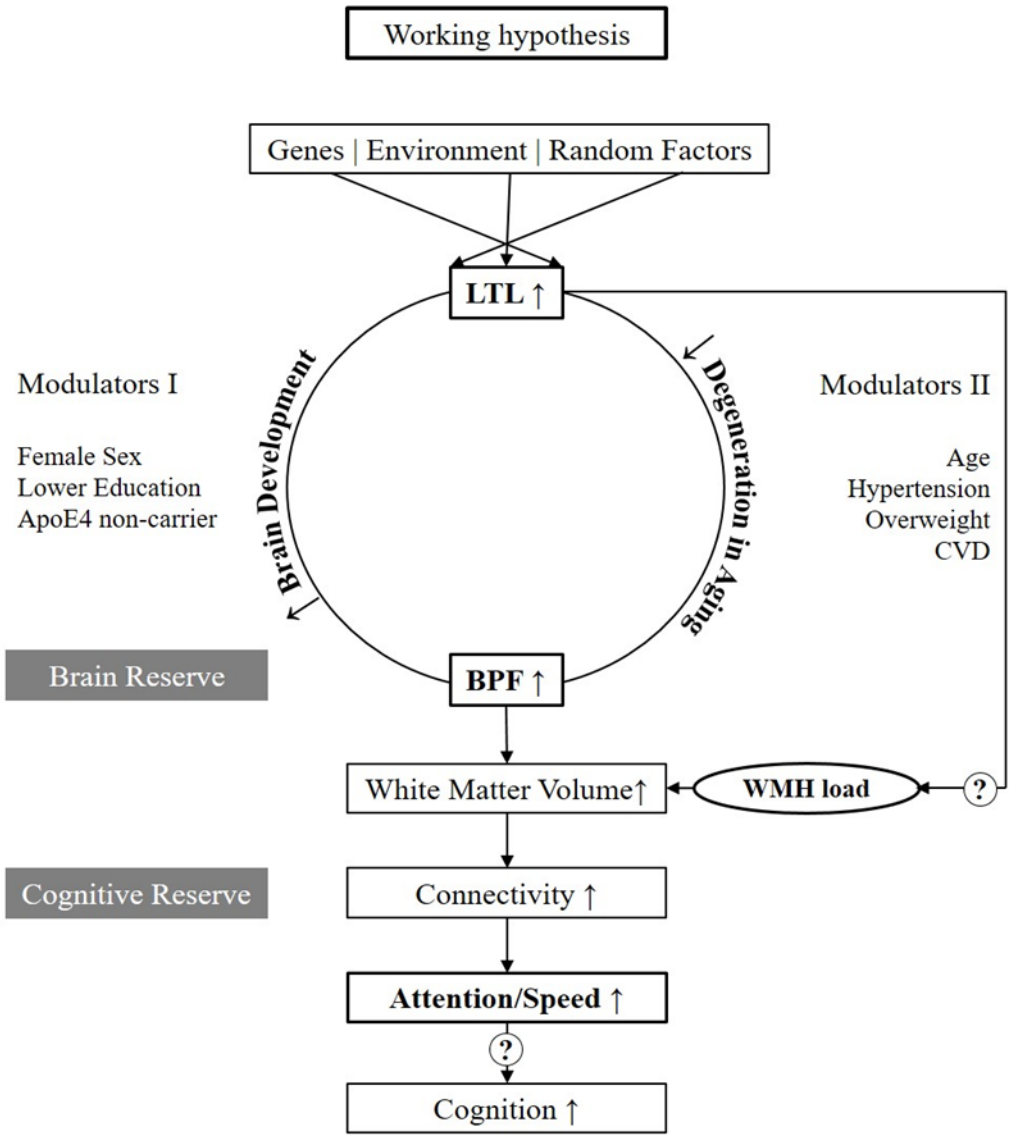
the brain, and this eventually transforms to better cognitive performance, especially within the attention/speed domain[1]. Attention/speed is considered as a basic cognitive function, which declines with age linearly, and its decline also affects other cognitive domains [123,124]. In our cohort, BPF was only significantly related to attention/speed but not to other cognitive domains or g-factor when tested by linear regression using model III.

4.2.1.4 Working hypothesis

Figure 13 presents our working hypothesis on the relation between LTL and ageing related structural and functional changes in the brain. In this model, we hypothesise that longer LTL plays a causal and protective role in the process of brain ageing. A causal role for telomeres for AD was suggested by Mendelian randomisation studies[125,126]. The effect of LTL on BPF might be mediated by two major pathways, one linked to the development of the brain up to puberty and the second related to ageing. Telomere attrition is accelerated in these two phases that led to major inter-individual differences in telomere length [66]. In the first pathway, we hypothesise that those inheriting longer telomeres or loosing less of their telomere during growth develop a larger brain translating to higher brain reserve. Indeed, in younger life, faster LTL attrition was associated with poorer global cognitive function as well as worse performance in domains that include processing speed in midlife [127]. We further hypothesise that the effect modulators acting already early in life, such as female sex, lower education, and ApoE4 non-carrier status, mainly act on this developmental pathway (Modulators I). In the second pathway, we hypothesise that in those with longer LTL, the process of brain ageing is slowed down, and brain parenchyma is longer preserved. Besides, during ageing, the brain is challenged by risk factors such as old age, hypertension, overweight and CVD (Modulators II). Those, however, who enter the process of ageing with a larger brain reserve are better protected against atrophy and cognitive decline. This is in line with the findings of Brickman et al. (2011) [128], showing that for any given level of cognitive function those with higher reserve -derived by latent variable analyses have higher WMH load [128]. A difference, which especially gets measurable when risk factors are present. At present, our data only supports the connection between longer LTL, larger BPF and better performances in attention/speed (**Figure 13**)[1]. However, whether better attention/speed translates further to better cognitive performances in other domains as well, need to be investigated in larger studies. In addition, the use of DTI markers to study microstructural substrate of the hypothesised larger brain reserve needs to be followed up [129].

The strong association of PSMD and attention/speed as reported in later life[31]. Undoubtedly, the paradox association between LTL and WMH mandates further investigation.

Figure 13:Working hypothesis. *Variables in bold are those which are tested in the present study. LTL: Leukocyte Telomere Length, BPF: Brain Parenchymal Fraction, WMH: White Matter Hyperintensities, CVD: Cardiovascular disease. [Figure reproduced from Gampawar et al., Frontiers in Psychiatry 2020 [1].]*



4.2.1.5 Strengths and weaknesses

Apart from being a well-established population-based cohort, the clear strength of studying the role of LTL in brain ageing within ASPS is a large sample size with LTL measurements, MRI correlates and comprehensive measures on cognitive functions[1]. The large sample size allowed us stratification of our cohort into subgroups based on the presence or the absence of relevant risk factors and define significant effect modulators. Nonetheless, the sample size in the subgroup of participants with DM was too small to reach statistical power for meaningful

testing[1]. Our results from mediation analyses provided the first evidence of the beneficial effect of LTL on BPF that eventually transforms into a positive effect on attention/speed by mediation analyses[1]. Due to the explorative nature of our study, we used the 3 linear regression models within 16 subgroups for all MRI and cognitive phenotypes[1]. Indeed, correction for multiple testing using FDR, only the significant association between LTL and BPF (p<0.05) persisted but not for WMH (p=0.1) or attention/speed[1].

A weakness of the study is its cross-sectional setting, which makes causal interference difficult as well as the possibility for residual confounding in spite to test for a wide range of possible confounders. The ASPS participants were scanned on a 1.5 Tesla scanner with large voxels (3x3x3 mm), and advanced MRI phenotypes such as regional volumes, subcortical structural volumes, brain tissue integrity markers from DTI were not available. In addition, LTL estimation was done using RT-PCR, which itself poses some limitations. The LTL measured by RT-PCR is a relative measurement with respect to the study-specific reference DNA and is prone to variability due to differences in the protocol as well as calibration methods used by the laboratories. This hampers the combined analyses of LTL data originating from different laboratories and studies [72]. Considering these limitations, we will turn to the ASPS-Fam and GSHA cohorts in which advanced MRI phenotypes are available, and LTL measurement derived from WGS data presently is under evaluation.

4.2.2 WGS derived telomere length and brain ageing

We compared Computel and TelSeq and found that both tools work well. Though TelSeq overestimates LTL as compared to Computel, the results are highly correlated. One large study (n=3362) compared LTL derived from these two tools also found the similar results (r = 0.98) [98]. Moreover, TelSeq is time-efficient as compared to Computel as it does not involve read alignment step. Due to these reasons, we used TelSeq for a pilot study. Our preliminary results from GSHA samples show a statistically significant inverse correlation between age and LTL and sex-specific effect. But this is in a small sample size of 90 participants and need further investigation.

4.3 Establishing resources

During the last five years, the field of genetic epidemiology moved from studying common genetic variations identified by microarray analyses to the more comprehensive detection of variants using NGS. As prices were falling and throughput was increasing, NGS at the exomic and later at the genomic scale became available in even large cohorts. Know-how performing WES and WGS from running wet-lab protocols to performing data-analyses needed to be established in our research unit in order to set up strategic resources for future research based on WES and WGS databases. Establishing these resources includes careful selection of wet lab methodology and improvement of the available protocols in order to gain the highest possible validity of genomic data for future analyses.

4.3.1 Establishing WES on Ion Proton

Here, we established WES on Ion Proton by comparing currently available only two WES library preparation methods namely, AmpliSeq and SureSelect, as well as by developing a novel variant calling pipeline. AmpliSeq design covers a slightly larger proportion of the RefSeq, Ensembl and UCSC coding regions than SureSelect design[2]. Concordance rate with microarray genotype data was excellent for both methods (<97%)[2]. When validated against NA12878 truthset both methods had comparable sensitivity of 93% but AmpliSeq had a higher PPV (84.5%) than SureSelect (80%)[2] Application of our novel 3 step variant calling pipeline reduced FPs by 90% and improved final sensitivity to 95% and PPV to 97% in both methods[2].

4.3.1.1 Protocol and design

AmpliSeq protocol, due to its PCR based design, is considerably faster (6h), comprises of fewer preparation steps and needs less hands-on time than SureSelect, allowing identification of exomic variants DNA from within 48h. The requirement of low amount of input DNA in is a further advantage over SureSelect. Therefore, when the amount of starting material is low and time is a constraint, AmpliSeq is the method of choice[2].

4.3.1.2 Microarray concordance

We observed excellent concordance rates of >97% against exome chip and >99% against GWAS chip genotype data in both methods[2]. A more favourable distribution of per base coverage in SureSelect despite having the lower average read depth probably a cause of the comparable concordance rates[2]. As we describe in the NA12878 sequencing results, SureSelect had a significantly higher proportion of bases in the callable range (5-400X) as compared to AmpliSeq (98.49% vs 79.49%)[2]. The callable range was set up by the manufacturer to reduce

the computation time. Sequencing had an added advantage over the exome chip genotyping in the detection of very rare variants (MAF <0.05%) as it detected on average 90% more variants than the exome chip[2].

4.3.1.3 Variant detection

Validating sequencing performance by microarray genotypes, however, have several drawbacks. Importantly, microarrays do not provide genotype data for all SNVs called by sequencing, they may contain errors, and comparison of sequencing data with microarrays leads to an overestimation of sensitivity. It is therefore recommended to validate sequencing methods by comparing their results with high confidence genotypes obtained from reference DNAs [100,104]. Therefore, as recommended, we utilised the NA12878 high confidence variant calls to validate the two libraries and found that sensitivity was comparable for both methods (93.1% vs 93.5%) while PPV was relevantly higher for AmpliSeq (84.5% vs 80%).

4.3.1.4 Optimisation-variant calling pipeline

To improve the validity of methods, we used three steps namely, 1) regularisation, 2) restriction to HCR and 3) changing default parameters for variant calling. Regularisation breaks down complex variants into simplest form generating phased genotypes. In both methods, regularisation had the most substantial effect on reduction of FNs by 20% [2].

Use of HCR as a target region leads to the exclusion of the complex, difficult to sequence areas such as simple repeats, tandem duplications, regions inside structural variants [100]. This led to the most remarkable reduction of FPs by 85.6% in AmpliSeq and 91.6% in SureSelect[2]. This is in line with a recent report which stated that PCR based methods are more prone to errors in complex genomic regions[119]. However, restricting analyses to HCR reduces the target coding region by 5.5 MB in AmpliSeq and 6.8MB in SureSelect and could miss the variant in this region[2]. This is an important consequence for the identification of pathogenic variants in rare diseases in the coding region. If pathogenic variants are expected in these excluded regions, we recommend variant calling using the TTR with a hotspot list of all known pathogenic variants. This strategy delivers genotypes at all positions given on the hotspot list even when the individual is homozygous for the reference allele[2].

Fine-tuning by stepwise changing the default parameters for variant calling had an effect only in the AmpliSeq method where FPs were reduced by 27% (Supplementary table 26) [2]. This is due to the fact that the AmpliSeq method suffers PCR based errors as seen by the high number of FPs due to parameters such as strand bias and minimal allele frequency [2].

Both AmpliSeq and SureSelect had high sensitivity and PPV for the detection of SNVs but lower values for detection of Indels. We found similar estimates for AmpliSeq to previously reported in the literature[130]; however, no studies are available for comparing SureSelect performance on Ion Proton,.

4.3.1.5 AmpliSeq ETR

When we used our optimisation pipeline on AmpliSeq ETR, we observed a similar reduction in the number of FPs and FNs, resulting in a relevant improvement in PPV from 85.7% to 98.5%[2]. Though the use of ETR is advised by manufacturer, it significantly reduces TPs by 28%. We, however, strongly recommend using our pipeline together with TTR instead of ETR.

4.3.1.6 RefSeq coding region and overlapping target region

A more direct and valid comparison of two methods can be achieved by focussing their performances in the RefSeq coding region and OTR. AmpliSeq detected 3% more variants in the RefSeq coding region as compared SureSelect[2]. This was probably due to the fact that AmpliSeq by design covers 3.2% more of the RefSeq coding region than SureSelect does. In OTR both the methods performed equally well.

Importantly, we showed that it is possible to combine data from both methods for meta-analysis, as within OTR, ~98% TP variants were shared by both libraries and the number of library-specific variants were reduced to a minimum.

4.3.1.7 Duplicate removal

Duplicate removal is meant for removal of PCR duplicates from the library. As AmpliSeq is a PCR based method, most of the reads generated are PCR duplicates. Therefore, duplicate removal using standard practices led to loss of most reads (88%)[2]. When duplicates were removed using tool "MarkDupbyStartEnd", only 18% of the reads were removed. Nonetheless, it did not improve the outcome and inversely, the sensitivity of the assay reduced. The manufacturer also does not suggest removing the duplicates from the AmpliSeq libraries. Although SureSelect is a hybridisation-based method, and removing duplicates did not reduce the reads as much as in AmpliSeq, it did not improve the outcome. Furthermore, studies have shown that duplicate removal had a minimal effect on downstream variant calling[131]. Therefore, we do not suggest duplicate removal on either of the libraries on Ion Proton.

4.3.1.8 Exploration of FNs and FPs

Library-related issues dominated FN SNVs in both libraries whereas library, as well as sequencer related problems were equally responsible for FN indels[2]. A PCR related issue of

strand bias was predominantly responsible for FP SNVs in AmpliSeq therefore, increasing the stringency of parameter "minimum allele frequency" led to a significant reduction of FPs by 27%[2]. Homopolymer and complex regions were responsible caused for most of the FP indels in both methods. FPs categorised as "unknown" were located mainly in complex regions[2]. This is similar to previous reports showing a high rate of errors in complex genomic regions using Torrent technology [119]. Therefore, regardless of the chemistry used for library preparation, FN and FP indels on Ion Proton were mainly due to the sequencing technology.

4.3.1.9 Effect of coverage on AmpliSeq performance

Contrary to our expectations, increasing the mean read depth from 44X to 270X did not result in a reduction of FNs and FPs; however, AmpliSeq performance reached a plateau between 80-90X coverage. Increasing the mean read depth, exponentially increased the number of bases covered >400X but not those covered <5X (Supplementary table 25)[2]. which could be the result of PCR based biases during the target selection and library enrichment. These results echoed the previous finding using *in silico* downsampling[130]. SureSelect design targeted 62% of the regions which are either missed by or covered with less than 5X by AmpliSeq. Out of these 62% targeted regions, 72%were covered by ≥5X by SureSelect **(Figure 14).**

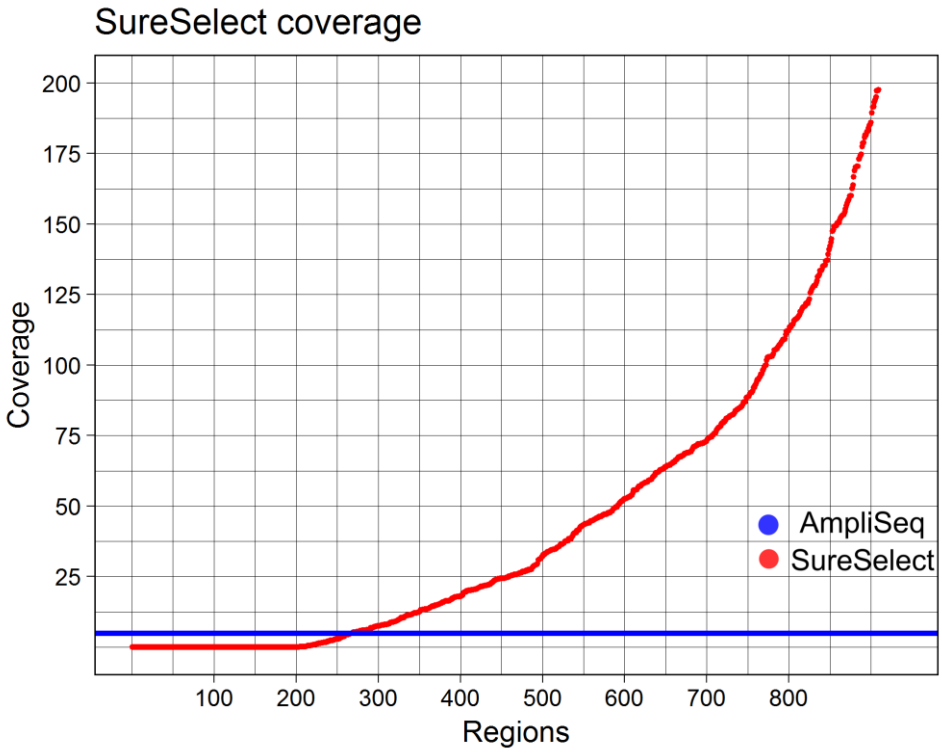


Figure 14:Coverage of SureSelect on low covered AmpliSeq regions. *SureSelect design covered 122,709 bases (spanned over 935 regions) out of 196,820 bases which are covered with < 5X by AmpliSeq. Out of these 935 regions in SureSelect, 268 regions had coverage of <5X and remaining 667 regions of ≥5X [Figure reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]]*

4.3.1.10 Mapping and variant calling pipelines

The Torrent Suite contains user friendly TMAP+TVC pipeline for alignment and variant calling and is provided by the manufacturer. This pipeline is optimised for analysis of data produced from Ion Torrent platforms. We did not test alternative variant calling pipelines because previous reports clearly showed that for the analysis of data from Ion Torrent platforms, Torrent Suite is the most sensitive and specific [132–134]. Especially, Yeo et al. [135]compared the TMAP+TVC pipeline versus alternative pipelines such as frequently used BWA + GATK pipeline. Although they found high sensitivity (97%) and specificity (100%) and a low false discovery rate (0%) for SNV calling in the training set, in the validation set the values were somewhat less convincing (93%, 100% and 3% respectively). Recent studies also showed that the GATK pipeline on Ion Torrent data has comparable sensitivity detecting SNVs, but it is inferior for the detection of indels [136]. Similarly, Vanni *et al.*, (2015) [133] also reported that BWA+GATK pipeline on Ion Torrent data worked equally well to TMAP+TVC for detection of SNVs; however, it had a high error rate in indel calling. They concluded that though for SNVs, both pipelines are comparable, for indels TMAP+TVC performs better.

4.3.1.11 Strengths and weaknesses

A thorough and stepwise comparison of the AmpliSeq and SureSelect methods including, their design, protocols, sequencing performances, and variant calling add to the strength of this study[2]. We validated both methods using microarray genotyping data as well as by NA12878 reference calls and reported the number of TPs, FNs, FPs, sensitivity, and PPV[2]. Our novel variant calling pipeline reduced FPs in both methods and improved sensitivity and PPV substantially. Manual inspection and categorisation of all FNs and FPs on Chr 1, 7, 16, 19, and X gave details about the most likely origin of these errors. We explored the potential of increasing average read depth using AmpliSeq as a measure to improve variant calling and provide evidence that an average read depth of 80-100X, which is achievable by loading two samples/PI chip, represents an optimal setting[2].

We primarily addressed scientist using Ion Proton, therefore, we did not compare these methods with WES library preparation methods available for Illumina platforms. Due to limited resources, we were only able to explore the effect of coverage in AmpliSeq but not in SureSelect. We also did not explore the regions failed to be sequenced by either library. It was, however, already done for AmpliSeq by a recent study [130]. We devised our pipeline on GRCh 37 reference assembly. When we performed uplifting of GRCh37 data to GRCh38, target

regions of both the libraries reduced significantly. AmpliSeq HCR decreased from 53.6MB to 49.7MB and SureSelect from 54.2MB to 50.19MB. Due to reduction of the target region, the variants in the truthset also decreased from 48,796 to 45,749 in AmpliSeq HCR and from 46,557 to 43,453 in SureSelect HCR. Therefore, a re-optimisation of our protocol is necessary in case data aligned to GRCh 38 assembly is used.

4.3.2 Comprehensive detection of genomic variants by NGS

In contrast to GWAS where CHARGE and BRIDGET identified several novel common variants associated with manifestations of brain ageing, we expect that in the future data derived from WES/WGS will allow for the discovery of rare variants with a large effect on brain ageing phenotypes with new information on pathways driving the process.

The utilisation of the established multidimensional database in the ASPS, ASPS-Fam and GSHA cohorts represents the fundament for future studies dissecting the complex process of brain ageing that will ultimately lead to a more precise assessment of the individual predisposition for stroke and dementia. By pinpointing to specific molecular pathways, these studies represent an essential step in finding new preventive and therapeutic targets against these diseases. The availability of comprehensive genetic epidemiological studies together with the established resources, will constitute the bases of precision medicine

5 Outlook

The work conducted within the frame of this thesis, including the two publications as well as the established databases, know-how and international networks will facilitate follow-up studies in the future. On the short term, we aim to focus on the role of 1) LTL using advanced MRI markers, 2) variation in DNA methylation pattern and 3) rare DNA variants derived from WES/WGS data in the process of brain ageing.

5.1 LTL using advanced MRI markers

To follow up on our findings published in 'Frontiers in Psychology', we initiated a collaborative approach utilising LTL derived from WGS data within the BRIDGET consortium. This collaboration offers three advantages 1) an increase in sample size and statistical power to investigate classical MRI markers such as total intracranial volume, total brain volume, grey matter volume, hippocampal volume, and WMH as well as cognition, 2) advanced MRI markers such as MD, FA, RD and PSMD on DTI or BrainAge and AD score [137] constructed by supervised machine learning, allowing to investigate the effect at the microstructural level, and 3) widening the age range of the participants from 20 in i-Share to >90 years in most of the cohorts in order to study the life course effect of LTL. The total number of participants with harmonised phenotypic data on MRI and cognitive tests eligible for this analysis will be approximately 2300.

The pilot study in GSHA already provided a proof of concept. Our results and project proposal for this collaborative effort has been presented and discussed at the BRIDGET meeting in 2020. According to the agreement, a central database containing phenotypic and LTL data will be created in the consortium, and the final analysis will be performed at the Research Unit Genetic Epidemiology in Graz.

5.2 Variation in DNA methylation pattern

Since the beginning, one of the aims of the BRIDGET consortium was to examine the association between epigenetic alterations and brain ageing-related phenotypes across the lifespan. The focus was laid on DNA methylation, and the participating BRIDGET cohorts went through a methylC-seq capture assay for a comprehensive exploration of DNA methylation at the McGill Genomic Centre. In total, 100 participants of GSHA are part of this project. The total sample size over all cohorts is approximately 1000. The data is currently processed centrally and is at the quality control stage. In order to further increase the sample

size, a methylation profiling of 900 ASPS participants, selected based on the availability of clinical data, using Infinium Methylation EPIC Bead Chip Kit from Illumina that interrogates over 850,000 methylation sites across the genome has been initiated. The results of this analysis will be available by the end of 2020.

Like telomeres, DNA methylation is not only a biomarker but also hypothesised to play a causal role in the pathomechanism of age-related diseases by regulating gene expressions. It is also known that environmental exposures and lifestyle factors are able to influence and modify DNA methylation patterns across the lifespan. There are preliminary data suggesting that DNA methylation patterns might be modulated in that way that ageing is slowed down or even halted. These new developments change the concept of ageing from being a resilient to a plastic process, which can be prevented or even treated as any other disease. Therefore, studying the effect of DNA methylation on brain ageing opens a novel scope for understanding and treating age-related diseases as well. Especially the interaction of lifestyle factors with methylation patterns across the lifespan is of major interest, as preventive measures slowing ageing will have high public health impact and major consequences on the society as a whole.

5.3 Rare DNA variants derived from WES/WGS

Use of WGS extends the variant spectrum from common to rare up to individual level SNPs and allows for the detection of structural variants as well. WGS data from the BRIDGET cohorts have already undergone sequencing, joint variant calling, quality control, and variant annotations centrally, and are ready to be analysed in relation to phenotypes of interest focussing on WMH, hippocampal volume, intracranial volume, PSMD, AD score and BrainAge. Analysis plan involves both gene-based and region-based burden analyses as well as single variant analyses after rank-based inverse normal transformation of phenotypes and adjusting for age, sex, and common risk factors. These results will certainly expand our understanding of the biological mechanisms underlying early structural brain alterations that portend an increased dementia risk and identification of early predictors.

We envision that in the future, the use of genetic epidemiological approaches will further expand our understanding of the genomics of complex traits associated with ageing, especially brain ageing. The studies completed and the essential resources established in the frame of this thesis represent the foundation for future studies brain ageing by profiling genomic, epigenomic and environmental markers through both local networks as well as international collaborations.

Bibliography

1 Gampawar P, Schmidt R, Schmidt H. Leukocyte Telomere Length Is Related to Brain Parenchymal Fraction and Attention/Speed in the Elderly: Results of the Austrian Stroke Prevention Study. *Front Psychiatry* 2020; **11**: 1–11.

2 Gampawar P, Saba Y, Werner U, Schmidt R, Müller-Myhsok B, Schmidt H. Evaluation of the Performance of AmpliSeq and SureSelect Exome Sequencing Libraries for Ion Proton. *Front Genet* 2019; **10**: 1–11.

3 López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. Cell. 2013; **153**: 1194.

4 Melzer D, Pilling LC, Ferrucci L. The genetics of human ageing. *Nat Rev Genet* 2019. doi:10.1038/s41576-019-0183-6.

5 Niccoli T, Partridge L. Ageing as a risk factor for disease. Curr. Biol. 2012; **22**: R741–R752.

6 Cole JH, Ritchie SJ, Bastin ME, Valdés Hernández MC, Muñoz Maniega S, Royle N *et al.* Brain age predicts mortality. *Mol Psychiatry* 2018; **23**: 1385–1392.

7 Alzheimer Europe. Dementia in Europe Yearbook 2019 Estimating the prevalence of dementia in Europe. *Alzheimer Eur* 2020; : 74–75.

8 Arvanitakis Z, Shah RC, Bennett DA. Diagnosis and Management of Dementia: Review. *JAMA - J Am Med Assoc* 2019; **322**: 1589–1599.

9 Mayeux R, Stern Y. Epidemiology of Alzheimer disease. *Cold Spring Harb Perspect Med* 2012; **2**. doi:10.1101/cshperspect.a006239.

10 Holtzman DM, Morris JC, Goate AM, John CM, Goate AM. Alzheimer's Disease: The Challenge of the Second Century. *Sci Transl Med* 2011; **3**: 77sr1.

11 Bennett DA, Schneider JA, Arvanitakis Z, Kelly JF, Aggarwal NT, Shah RC *et al.* Neuropathology of older persons without cognitive impairment from two community-based studies. *Neurology* 2006; **66**: 1837–1844.

12 Vinke EJ, de Groot M, Venkatraghavan V, Klein S, Niessen WJ, Ikram MA *et al.* Trajectories of imaging markers in brain aging: the Rotterdam Study. *Neurobiol Aging* 2018; **71**: 32–40.

13 Cole JH, Marioni RE, Harris SE, Deary IJ. Brain age and other bodily 'ages': implications for neuropsychiatry. Mol. Psychiatry. 2019; **24**: 266–281.

14 Peters R. Ageing and the brain. *Postgrad Med J* 2006; **82**: 84–88.

15 Debette S, Schilling S, Duperron MG, Larsson SC, Markus HS. Clinical Significance of Magnetic Resonance Imaging Markers of Vascular Brain Injury: A Systematic Review and Meta-analysis. *JAMA Neurol* 2019; **76**: 81–94.

16 Debette S, Markus HS. The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: systematic review and meta-analysis. *BMJ* 2010; **341**: c3666.

17 Jack CR, Knopman DS, Jagust WJ, Petersen RC, Weiner MW, Aisen PS *et al.* Tracking pathophysiological processes in Alzheimer's disease: An updated hypothetical model of dynamic biomarkers. Lancet Neurol. 2013; **12**: 207–216.

18 Van Elderen SSGC, Zhang Q, Sigurdsson S, Haight TJ, Lopez O, Eiriksdottir G *et al.* Brain Volume as an Integrated Marker for the Risk of Death in a Community-Based Sample: Age Gene/Environment Susceptibility—Reykjavik Study. *J Gerontol A Biol Sci Med Sci* 2016; **71**: 131–137.

19 Nir TM, Jahanshad N, Villalon-Reina JE, Toga AW, Jack CR, Weiner MW *et al.* Effectiveness of regional DTI measures in distinguishing Alzheimer's disease, MCI, and normal aging. *NeuroImage Clin* 2013; **3**: 180–195.

20 Tucker-Drob EM. Global and Domain-Specific Changes in Cognition Throughout Adulthood. *Dev Psychol* 2011; **47**: 331–343.

21 Hoogendam YY, Hofman A, Van Der Geest JN, Van Der Lugt A, Ikram MA. Patterns of cognitive function in aging: The Rotterdam Study. *Eur J Epidemiol* 2014; **29**: 133–140.

22 Karlamangla AS, Lachman ME, Han W, Huang M, Greendale GA. Evidence for Cognitive Aging in Midlife Women: Study of Women's Health Across the Nation. *PLoS One* 2017; **12**: e0169008.

23 Wilson RS, Segawa E, Hizel LP, Boyle PA, Bennett DA. Terminal dedifferentiation of cognitive abilities. *Neurology* 2012; **78**: 1116–1122.

24 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**: e1002261.

25 Schmidt R, Ropele S, Enzinger C, Petrovic K, Smith S, Schmidt H *et al.* White matter lesion progression, brain atrophy, and cognitive decline: The Austrian stroke prevention

study. *Ann Neurol* 2005; **58**: 610–616.

26 Staals J, Booth T, Morris Z, Bastin ME, Gow AJ, Corley J *et al.* Total MRI load of cerebral small vessel disease and cognitive ability in older people. *Neurobiol Aging* 2015; **36**: 2806–2811.

27 Leong RLF, Lo JC, Sim SKY, Zheng H, Tandi J, Zhou J *et al.* Longitudinal brain structure and cognitive changes over 8 years in an East Asian cohort. *Neuroimage* 2017; **147**: 852–860.

28 Gorbach T, Pudas S, Lundquist A, Orädd G, Josefsson M, Salami A *et al.* Longitudinal association between hippocampus atrophy and episodic-memory decline. *Neurobiol Aging* 2017; **51**: 167–176.

29 Anblagan D, Valdés Hernández MC, Ritchie SJ, Aribisala BS, Royle NA, Hamilton IF *et al.* Coupled changes in hippocampal structure and cognitive ability in later life. *Brain Behav* 2018; **8**: e00838.

30 Borghesani PR, Madhyastha TM, Aylward EH, Reiter MA, Swarny BR, Warner Schaie K *et al.* The association between higher order abilities, processing speed, and age are variably mediated by white matter integrity during typical aging. *Neuropsychologia* 2013; **51**: 1435–1444.

31 Deary IJ, Ritchie SJ, Muñoz Maniega S, Cox SR, Valdés Hernández MC, Luciano M *et al.* Brain Peak Width of Skeletonized Mean Diffusivity (PSMD) and Cognitive Function in Later Life. *Front Psychiatry* 2019; **10**. doi:10.3389/fpsyt.2019.00524.

32 Visscher PM, Hill WG, Wray NR. Heritability in the genomics era — concepts and misconceptions. *Nat Rev Genet* 2008; **9**: 255–266.

33 Mayhew AJ, Meyre D. Assessing the Heritability of Complex Traits in Humans: Methodological Challenges and Opportunities. *Curr Genomics* 2017; **18**: 332.

34 Kochunov P, Jahanshad N, Sprooten E, Nichols TE, Mandl RC, Almasy L *et al.* Multi-site study of additive genetic effects on fractional anisotropy of cerebral white matter: Comparing meta and megaanalytical approaches for data pooling. *Neuroimage* 2014; **95**: 136–150.

35 Vuoksimaa E, Panizzon MS, Hagler Jr DJ, Hatton SN, Fennema-Notestine C, Rinker D *et al.* Heritability of white matter microstructure in late middle age: A twin study of tract-based fractional anisotropy and absolute diffusivity indices. *Hum Brain Mapp* 2017; **38**: 2026–2036.

36 Finkel D, Pedersen NL, McGue M, McClearn GE. Heritability of cognitive abilities in adult twins: Comparison of Minnesota and Swedish data. *Behav Genet* 1995; **25**: 421–431.

37 Swagerman SC, de Geus EJC, Kan K-J, van Bergen E, Nieuwboer HA, Koenis MMG *et al.* The Computerized Neurocognitive Battery: Validation, aging effects, and heritability across cognitive domains. *Neuropsychology* 2016; **30**: 53–64.

38 Davis OSP, Haworth CMA, Plomin R. Dramatic increase in heritability of cognitive development from early to middle childhood: an 8-year longitudinal study of 8,700 pairs of twins. *Psychol Sci* 2009; **20**: 1301–8.

39 Plomin R, Pedersen NL, Lichtenstein P, McClearn GE. Variability and stability in cognitive abilities are largely genetic later in life. *Behav Genet* 1994; **24**: 207–215.

40 Benyamin B, Wilson V, Whalley LJ, Visscher PM, Deary IJ. Large, consistent estimates of the heritability of cognitive ability in two entire populations of 11-year-old twins from Scottish Mental Surveys of 1932 and 1947. *Behav Genet* 2005; **35**: 525–534.

41 McGue M, Christensen K. The heritability of cognitive functioning in very old adults: evidence from Danish twins aged 75 years and older. *Psychol Aging* 2001; **16**: 272–80.

42 Jansen AG, Mous SE, White T, Posthuma D, Polderman TJC. What Twin Studies Tell Us About the Heritability of Brain Development, Morphology, and Function: A Review. *Neuropsychol Rev* 2015; **25**: 27–46.

43 Yang J, Zeng J, Goddard ME, Wray NR, Visscher PM. Concepts, estimation and interpretation of SNP-based heritability. Nat. Genet. 2017; **49**: 1304–1310.

44 Elliott LT, Sharp K, Alfaro-Almagro F, Shi S, Miller KL, Douaud G *et al.* Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature* 2018; **562**: 210–216.

45 Zhao B, Ibrahim JG, Li Y, Li T, Wang Y, Shan Y *et al.* Heritability of Regional Brain Volumes in Large-Scale Neuroimaging and Genetic Studies. *Cereb Cortex* 2019; **29**: 2904–2914.

46 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; **9**: 1–16.

47 Davies G, Marioni RE, Liewald DC, Hill WD, Hagenaars SP, Harris SE *et al.* Genome-wide association study of cognitive functions and educational attainment in UK Biobank

(N=112 151). *Mol Psychiatry* 2016; **21**: 758–767.

48 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* 2018; **47**: 1005–1012.

49 Mayeux R. Biomarkers: Potential Uses and Limitations. *NeuroRx* 2004; **1**: 182–188.

50 Atallah N, Adjibade M, Lelong H, Hercberg S, Galan P, Assmann KE *et al.* How healthy lifestyle factors at midlife relate to healthy aging. *Nutrients* 2018; **10**. doi:10.3390/nu10070854.

51 Daskalopoulou C, Koukounari A, Ayuso-Mateos JL, Prince M, Prina AM. Associations of lifestyle behaviour and healthy ageing in five latin American and the Caribbean countries—A 10/66 population-based cohort study. *Nutrients* 2018; **10**. doi:10.3390/nu10111593.

52 Daskalopoulou C, Stubbs B, Kralj C, Koukounari A, Prince M, Prina AM. Associations of smoking and alcohol consumption with healthy ageing: A systematic review and meta-analysis of longitudinal studies. BMJ Open. 2018; **8**: e019540.

53 Most J, Tosti V, Redman LM, Fontana L. Calorie restriction in humans: An update. Ageing Res. Rev. 2017; **39**: 36–45.

54 Wyss-Coray T. Ageing, neurodegeneration and brain rejuvenation. Nature. 2016; **539**: 180–186.

55 Daskalopoulou C, Stubbs B, Kralj C, Koukounari A, Prince M, Prina AM. Physical activity and healthy ageing: A systematic review and meta-analysis of longitudinal cohort studies. Ageing Res. Rev. 2017; **38**: 6–17.

56 Daskalopoulou C, Koukounari A, Wu YT, Terrera GM, Caballero FF, de la Fuente J *et al.* Healthy ageing trajectories and lifestyle behaviour: the Mexican Health and Aging Study. *Sci Rep* 2019; **9**: 1–10.

57 Sen A, Gider P, Cavalieri M, Freudenberger P, Farzi A, Schallert M *et al.* Association of cardiorespiratory fitness and morphological brain changes in the elderly: Results of the Austrian stroke prevention study. *Neurodegener Dis* 2012; **10**: 135–137.

58 Freudenberger P, Petrovic K, Sen A, Töglhofer AM, Fixa A, Hofer E *et al.* Fitness and cognition in the elderly: The Austrian Stroke Prevention Study. *Neurology* 2016; **86**: 418–424.

59 Cabral DF, Rice J, Morris TP, Rundek T, Pascual-Leone A, Gomes-Osman J. Exercise

for Brain Health: An Investigation into the Underlying Mechanisms Guided by Dose. Neurotherapeutics. 2019; **16**: 580–599.

60 Schmidt H, Freudenberger P, Seiler S, Schmidt R. Genetics of subcortical vascular dementia. *Exp Gerontol* 2012; **47**: 873–877.

61 Blackburn EH, Epel ES, Lin J. Human telomere biology: A contributory and interactive factor in aging, disease risks, and protection. *Science (80-)* 2015; **350**: 1193–1198.

62 Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nat. Rev. Genet. 2018; **19**: 371–384.

63 Cole JH, Franke K. Predicting Age Using Neuroimaging: Innovative Brain Ageing Biomarkers. Trends Neurosci. 2017; **40**: 681–690.

64 Thanseem I, Viswambharan V, Poovathinal SA, Anitha A. Is telomere length a biomarker of neurological disorders? *Biomark Med* 2017; **11**: 799–810.

65 Mamdani F, Rollins B, Morgan L, Myers RM, Barchas JD, Schatzberg AF *et al.* Variable telomere length across post-mortem human brain regions and specific reduction in the hippocampus of major depressive disorder. *Transl Psychiatry* 2015; **5**: e636–e636.

66 Sanders JL, Newman AB. Telomere length in epidemiology: A biomarker of aging, age-related disease, both, or neither? *Epidemiol Rev* 2013; **35**: 112–131.

67 Mather KA, Jorm AF, Parslow RA, Christensen H. Is Telomere Length a Biomarker of Aging? A Review. *Journals Gerontol Ser A Biol Sci Med Sci* 2011; **66A**: 202–213.

68 Blackburn EH. Telomeres and telomerase: Their mechanisms of action and the effects of altering their functions. In: *FEBS Letters*. Elsevier, 2005, pp 859–862.

69 Armanios M, Blackburn EH. The telomere syndromes. Nat. Rev. Genet. 2012; **13**: 693–704.

70 Daniali L, Benetos A, Susser E, Kark JD, Labat C, Kimura M *et al.* Telomeres shorten at equivalent rates in somatic tissues of adults. *Nat Commun* 2013; **4**: 1–7.

71 Friedrich U, Griese EU, Schwab M, Fritz P, Thon KP, Klotz U. Telomere length in different tissues of elderly patients. *Mech Ageing Dev* 2000; **119**: 89–99.

72 Müezzinler A, Zaineddin AK, Brenner H. A systematic review of leukocyte telomere length and age in adults. *Ageing Res Rev* 2013; **12**: 509–519.

73 Dugdale HL, Richardson DS. Heritability of telomere variation: It is all about the environment! *Philos Trans R Soc B Biol Sci* 2018; **373**. doi:10.1098/rstb.2016.0450.

74 Broer L, Codd V, Nyholt DR, Deelen J, Mangino M, Willemsen G *et al.* Meta-analysis

of telomere length in 19 713 subjects reveals high heritability, stronger maternal inheritance and a paternal age effect. *Eur J Hum Genet* 2013; **21**: 1163–1168.

75 Haycock PC, Heydon EE, Kaptoge S, Butterworth AS, Thompson A, Willeit P. Leucocyte telomere length and risk of cardiovascular disease: Systematic review and meta- Analysis. *BMJ* 2014; **349**. doi:10.1136/bmj.g4227.

76 Honig LS, Kang MS, Schupf N, Lee JH, Mayeux R. Association of shorter leukocyte telomere repeat length with dementia and mortality. *Arch Neurol* 2012; **69**: 1332–1339.

77 Wang Q, Zhan Y, Pedersen NL, Fang F, Hägg S. Telomere Length and All-Cause Mortality: A Meta-analysis. Ageing Res. Rev. 2018; **48**: 11–20.

78 Vinke EJ, Ikram MA, Vernooij MW. Brain aging: more of the same!? *Aging (Albany NY)* 2019; **11**: 849–850.

79 Koretsky AP. New Developments in Magnetic Resonance Imaging of the Brain. *NeuroRx* 2004; **1**: 155–164.

80 Lockhart SN, Decarli C. Structural Imaging Measures of Brain Aging. *Neuropsychol Rev* 2014; **24**: 271–289.

81 Alexander AL, Lee JE, Lazar M, Field AS. Diffusion Tensor Imaging of the Brain. *Neurotherapeutics* 2007; **4**: 316–329.

82 Elobeid A, Libard S, Leino M, Popova SN, Alafuzoff I. Altered proteins in the aging brain. *J Neuropathol Exp Neurol* 2016; **75**: 316–325.

83 Chinta SJ, Woods G, Rane A, Demaria M, Campisi J, Andersen JK. Cellular senescence and the aging brain. *Exp Gerontol* 2015; **68**: 3–7.

84 Nelson PT, Head E, Schmitt FA, Davis PR, Neltner JH, Jicha GA *et al.* Alzheimer's disease is not 'brain aging': Neuropathological, genetic, and epidemiological human studies. Acta Neuropathol. 2011; **121**: 571–587.

85 Raman MR, Kantarci K, Murray ME, Jack CR, Vemuri P. Imaging markers of cerebrovascular pathologies: Pathophysiology, clinical presentation, and risk factors. Alzheimer's Dement. Diagnosis, Assess. Dis. Monit. 2016; **5**: 5–14.

86 Wardlaw JM, Smith C, Dichgans M. Small vessel disease: mechanisms and clinical implications. Lancet Neurol. 2019; **18**: 684–696.

87 Raz L, Knoefel J, Bhaskar K. The neuropathology and cerebrovascular mechanisms of dementia. J. Cereb. Blood Flow Metab. 2016; **36**: 172–186.

88 Mattson MP, Arumugam T V. Hallmarks of Brain Aging: Adaptive and Pathological

Modification by Metabolic States. Cell Metab. 2018; **27**: 1176–1199.

89 Schmidt R, Fazekas F, Kapeller P, Schmidt H, Hartung H-P. MRI white matter hyperintensities: Three-year follow-up of the Austrian Stroke Prevention Study. *Neurology* 1999; **53**: 132–132.

90 Seiler S, Pirpamer L, Hofer E, Duering M, Jouvent E, Fazekas F *et al.* Magnetization Transfer Ratio Relates to Cognitive Impairment in Normal Elderly. *Front Aging Neurosci* 2014; **6**: 263.

91 Fazekas1 F, Chawluk2 JB, Alavi1 A, Hurtig2 HI, Zimmerma& RA, Fazekas F *et al.* MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. *Am J Neuroradiol* 1987; **8**: 421–426.

92 Schmidt H, Schmidt R, Fazekas F, Semmler J, Kapeller P, Reinhart B *et al.* Apolipoprotein E4allele in the normal elderly: neuropsychologic and brain MRI correlates. *Clin Genet* 1996; **50**: 293–299.

93 Cawthon RM. Telomere measurement by quantitative PCR. *Nucleic Acids Res* 2002; **30**: 47e – 47.

94 Sen A, Marsche G, Freudenberger P, Schallert M, Toeglhofer AM, Nagl C *et al.* Association between higher plasma lutein, zeaxanthin, and vitamin C concentrations and longer telomere length: Results of the Austrian Stroke Prevention Study. *J Am Geriatr Soc* 2014; **62**: 222–229.

95 Nersisyan L, Arakelyan A. Computel: Computation of mean telomere length from whole-genome next-generation sequencing data. *PLoS One* 2015; **10**. doi:10.1371/journal.pone.0125201.

96 Ding Z, Mangino M, Aviv A, Spector T, Durbin R. Estimating telomere length from whole genome sequence data. *Nucleic Acids Res* 2014; **42**. doi:10.1093/nar/gku181.

97 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma Appl NOTE* 2009; **25**: 2078–2079.

98 Taub M, Weinstock J, Iyer K, Yanek L, Conomos M, Brody J *et al.* Novel genetic determinants of telomere length from a multi-ethnic analysis of 75,000 whole genome sequences in TOPMed. *bioRxiv* 2019; : 749010.

99 Schmidt R, Lechner H, Fazekas F, Niederkorn K, Reinhart B, Grieshofer P *et al.* Assessment of Cerebrovascular Risk Profiles in Healthy Persons: Definition of Research Goals and the Austrian Stroke Prevention Study (ASPS). *Neuroepidemiology* 1994; **13**:

308–313.

100 Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* 2014; **32**: 246–51.

101 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; **25**: 1754–1760.

102 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010; **20**: 1297–1303.

103 Hayes A. The PROCESS macro for SPSS and SAS. Processmacro.Org. 2016.http://processmacro.org/index.html (accessed 3 Jun2020).

104 Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* 2016; **3**: 160025.

105 Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010; **26**: 841–842.

106 Garrison E. GitHub - vcflib/vcflib: C++ library and cmdline tools for parsing and manipulating VCF files. Github. 2015.https://github.com/vcflib/vcflib (accessed 3 Jun2020).

107 Tan A, Abecasis GR, Kang HM. Unified representation of genetic variants. *Bioinformatics* 2015; **31**: 2202–2204.

108 Beraldi D. Java-cafe/MarkDupsByStartEnd at master · dariober/Java-cafe · GitHub. Github. 2015.https://github.com/dariober/Java-cafe/tree/master/MarkDupsByStartEnd (accessed 3 Jun2020).

109 Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G *et al.* Integrative genomics viewer. Nat. Biotechnol. 2011; **29**: 24–26.

110 Subramanian S, Mishra RK, Singh L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol* 2003; **4**. doi:10.1186/gb-2003-4-2-r13.

111 King KS, Kozlitina J, Rosenberg RN, Peshock RM, McColl RW, Garcia CK. Effect of Leukocyte Telomere Length on Total and Regional Brain Volumes in a Large Population-Based Cohort. *JAMA Neurol* 2014; **71**: 1247.

112 Suchy-Dicey AM, Muller CJ, Madhyastha TM, Shibata D, Cole SA, Zhao J *et al.* Telomere Length and Magnetic Resonance Imaging Findings of Vascular Brain Injury and Central Brain Atrophy. *Am J Epidemiol* 2018; **187**: 1231–1239.

113 Hägg S, Zhan Y, Karlsson R, Gerritsen L, Ploner A, van der Lee SJ *et al.* Short telomere length is associated with impaired cognitive performance in European ancestry cohorts. *Transl Psychiatry* 2017; **7**: e1100.

114 Devore EE, Prescott J, De Vivo I, Grodstein F. Relative telomere length and cognitive decline in the Nurses' Health Study. *Neurosci Lett* 2011; **492**: 15–18.

115 Lee M, Napier CE, Yang SF, Arthur JW, Reddel RR, Pickett HA. Comparative analysis of whole genome sequencing-based telomere length measurement techniques. *Methods* 2017; **114**: 4–15.

116 Goodwin S, McPherson JD, Richard McCombie W, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016; **17**: 333–351.

117 Loman NJ, Misra R V., Dallman TJ, Constantinidou C, Gharbia SE, Wain J *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* 2012; **30**: 434–439.

118 Boland JF, Chung CC, Roberson D, Mitchell J, Zhang X, Im KM *et al.* The new sequencer on the block: comparison of Life Technology's Proton sequencer to an Illumina HiSeq for whole-exome sequencing. *Hum Genet* 2013; **132**: 1153–63.

119 Samorodnitsky E, Jewell BM, Hagopian R, Miya J, Wing MR, Lyon E *et al.* Evaluation of Hybridization Capture Versus Amplicon-Based Methods for Whole-Exome Sequencing. *Hum Mutat* 2015; **36**: 903–914.

120 Wikgren M, Karlsson T, Söderlund H, Nordin A, Roos G, Nilsson L-G *et al.* Shorter telomere length is linked to brain atrophy and white matter hyperintensities. *Age Ageing* 2014; **43**: 212–217.

121 Kaja R, Reyes SM, Rossetti HC, Brown ES. Association between telomere length and cognitive ability in a community-based sample. *Neurobiol Aging* 2019; **75**: 51–53.

122 Zhan Y, Clements MS, Roberts RO, Vassilaki M, Druliner BR, Boardman LA *et al.* Association of telomere length with general cognitive trajectories: a meta-analysis of four prospective cohort studies. *Neurobiol Aging* 2018; **69**: 111–116.

123 Finkel D, Reynolds CA, McArdle JJ, Pedersen NL. Age Changes in Processing Speed as a Leading Indicator of Cognitive Aging. *Psychol Aging* 2007; **22**: 558–568.

124 Hagenaars SP, Harris SE, Davies G, Hill WD, Liewald DCM, Ritchie SJ *et al.* Shared genetic aetiology between cognitive functions and physical and mental health in UK Biobank (N=112 151) and 24 GWAS consortia. *Mol Psychiatry* 2016; **21**: 1624–1632.

125 Zhan Y, Song C, Karlsson R, Tillander A, Reynolds CA, Pedersen NL *et al.* Telomere length shortening and Alzheimer disease-A mendelian randomization study. JAMA Neurol. 2015; **72**: 1202–1203.

126 Kuźma E, Hannon E, Zhou A, Lourida I, Bethel A, Levine DA *et al.* Which Risk Factors Causally Influence Dementia? A Systematic Review of Mendelian Randomization Studies. *J Alzheimer's Dis* 2018; **64**: 181–193.

127 Cohen-Manheim I, Doniger GM, Sinnreich R, Simon ES, Pinchas R, Aviv A *et al.* Increased attrition of leukocyte telomere length in young adults is associated with poorer cognitive function in midlife. *Eur J Epidemiol* 2016; **31**: 147–157.

128 Brickman AM, Siedlecki KL, Muraskin J, Manly JJ, Luchsinger JA, Yeung LK *et al.* White matter hyperintensities and cognition: Testing the reserve hypothesis. *Neurobiol Aging* 2011; **32**: 1588–1598.

129 Staffaroni AM, Tosun D, Lin J, Elahi FM, Casaletto KB, Wynn MJ *et al.* Telomere attrition is associated with declines in medial temporal lobe volume and white matter microstructure in functionally independent older adults. *Neurobiol Aging* 2018; **69**: 68–75.

130 Damiati E, Borsani G, Giacopuzzi E. Amplicon-based semiconductor sequencing of human exomes: performance evaluation and optimization strategies. *Hum Genet* 2016; **135**: 499–511.

131 Ebbert MTW, Wadsworth ME, Staley LA, Hoyt KL, Pickett B, Miller J *et al.* Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics* 2016; **17 Suppl 7**: 239.

132 Caboche S, Audebert C, Lemoine Y, Hot D. Comparison of mapping algorithms used in high-throughput sequencing: Application to Ion Torrent data. *BMC Genomics* 2014; **15**: 264.

133 Vanni I, Coco S, Truini A, Rusmini M, Dal Bello MG, Alama A *et al.* Next-Generation Sequencing Workflow for NSCLC Critical Samples Using a Targeted Sequencing Approach by Ion Torrent PGM™ Platform. *Int J Mol Sci* 2015; **16**: 28765–82.

134 Guerreiro R, Bras J, Hardy J, Singleton A, Brás J, Hardy J *et al.* Next generation

sequencing techniques in neurological diseases: redefining clinical and molecular associations. *Hum Mol Genet* 2014; **23**: R47-53.

135 Yeo Z, Wong JC, Rozen SG, Lee AS. Evaluation and optimisation of indel detection workflows for ion torrent sequencing of the BRCA1 and BRCA2 genes. *BMC Genomics* 2014; **15**: 516.

136 De Summa S, Malerba G, Pinto R, Mori A, Mijatovic V, Tommasi S. GATK hard filtering: Tunable parameters to improve variant calling for next generation sequencing targeted gene panel data. *BMC Bioinformatics* 2017; **18**: 14–16.

137 Frenzel S, Wittfeld K, Habes M, Klinger-König J, Bülow R, Völzke H *et al.* A Biomarker for Alzheimer's Disease Based on Patterns of Regional Brain Atrophy. *Front Psychiatry* 2020; **10**. doi:10.3389/fpsyt.2019.00953.

Tables

Table 1:Characteristics of ASPS cohort used in LTL analysis

		Total (n)	Men (n)	Women (n)	<65 years (n)	>65 years (n)
Age (years)		65.9±8.0 (909)	65.3±7.7 (388)	66.4±8.1 (521)	59.2±4.2 (440)	**72.3±4.7***** (469)
Sex (Men)		388 (42.7%)	NA	NA	202 (45.9%)	186 (39.7%)
Hypertension		631 (69.4%)	274 (70.6%)	357 (68.5%)	244 (55.5%)	**387(82.5%)*****
SBP (mm Hg)		143.8±22.8 (908)	143.3±21.4 (387)	144.2±23.8 (521)	136.7±20.1 (440)	**150.5±23.2***** (468)
DBP (mm Hg)		87.6±10.0 (908)	87.6±9.7 (387)	87.6±10.2 (521)	86.3±9.5 (440)	**88.8±10.3***** (468)
DM		99 (10.9%)	42 (10.8%)	57 (10.9%)	31 (7.0%)	**68 (14.5%)*****
CVD		364 (40.0%)	139 (35.8%)	**225 (43.2%)***	149 (33.9%)	**215(45.8%)*****
BMI (kg/m^2)		26.9±4.1 (909)	27.2±3.3 (388)	**26.7±4.6*** (521)	26.5±4.0 (440)	**27.2±4.1**** (469)
Education		11.1±2.6 (909)	11.8±3.0 (388)	10.6±2.1*** (521)	11.4±2.7 (440)	10.9±2.4** (469)
TC (mg/dl)		227.87±40.4 (909)	219.8±38.2 (388)	**233.9±41.0***** (521)	229.9±41.0 (440)	226.0±39.8 (469)
HDL (mg/dl)		56.5±17.2 (901)	49.4±15.0 (382)	**61.8±16.8***** (519)	53.9±16.7 (438)	**58.9±17.3***** (463)
Smoking						
Never		551 (60.7%)	149 (38.4%)	402 (77.3%)	249 (56.7%)	302 (64.4%)
Former		257 (28.3%)	179 (46.1%)	78 (15.0%)	128 (29.2%)	129 (27.5%)
Current		100 (11.0%)	60 (15.5%)	**40 (7.7%)*****	62 (14.1%)	**38 (8.1%)****
ApoE4						
No allele		718 (80.0%)	303 (78.5%)	415 (81.1%)	343 (78.7%)	375 (81.2%)
Heterozygous		173 (19.3%)	81 (21.0%)	92 (18.0%)	88 (20.2%)	85 (18.4%)
Homozygous		7 (0.8%)	2 (0.5%)	5 (0.9%)	5 (1.1%)	2 (0.4%)
LTL		0.61(0.47-0.82) (909)	0.62(0.47-0.85) (388)	0.60(0.46-0.79) (521)	0.62(0.46-0.87) (440)	0.60(0.47-0.79) (469)
BPF (%)		78.7±3.9 (739)	78.6±3.7 (313)	78.7±4.0 (426)	80.6 ± 3.3 (352)	**76.9 ± 3.6***** (387)
WMH load (mm^3)		0.8(0.2-3.1) (827)	0.6(0.1-2.5) (359)	**1.1(0.2-3.8)**** (468)	0.3(0.0-1.1) (396)	**2.2(0.6-5.5)***** (431)
WMH Score	1	157 (18.4%)	84 (23.0%)	73 (15.0%)	117 (28.5%)	40 (9.1%)
	2	503 (59.0%)	202 (55.2%)	301 (61.9%)	271 (65.9%)	232 (52.6%)
	3	123 (14.4%)	53 (14.5%)	70 (14.4%)	16 (3.9%)	107 (24.3%)
	4	69 (8.1%)	27 (7.4%)	**42 (8.6%)***	7 (1.7%)	**62 (14.1%)*****
Attention/ Speed		0.03±0.53 (853)	0.12±0.50 (365)	**-0.03±0.54***** (488)	0.24±0.41 (423)	**-0.17±0.55***** (430)
Conceptual-isation		0.06±0.59 (858)	0.08±0.61 (361)	0.05±0.58 (497)	0.17±0.61 (429)	**-0.04±0.56***** (429)
Memory		0.01±1.01 (860)	0.12±1.07 (369)	**-0.07±0.96**** (491)	0.37±1.04 (429)	**-0.35±0.84***** (431)
Visuopractical Skills		0.00±0.93 (891)	-0.14±0.88 (375)	**0.10±0.95***** (516)	0.47±0.78 (434)	**-0.45±0.82***** (457)
g factor		0.01±1.01 (823)	0.11±1.04 (350)	**-0.07±0.98*** (473)	0.50±0.84 (413)	**-0.49±0.91***** (410)

Table 1 continue

		Normotensive (n)	Hypertensive (n)	BMI <25 (n)	BMI ≥25 (n)
Age (years)		62.0±7.1 (278)	**67.7±7.7***** (631)	65.3±8.2 (325)	66.3±7.8 (584)
Sex (Men)		114 (41.0%)	274 (43.4%)	109 (33.5%)	**279 (47.8%)*****
Hypertension		NA	NA	192 (59.1%)	**439 (75.2%)*****
SBP (mm Hg)		122.7±10.0 (277)	**153.4±20.1***** (631)	137.9±21.6 (325)	**147.1±22.8***** (583)
DBP (mm Hg)		79.2±4.6 (277)	**91.2±9.5***** (631)	84.8±9.0 (325)	**89.1±10.2***** (583)
DM		20 (7.2%)	**79 (12.5%)***	13 (4.0%)	**86 (14.7%)*****
CVD		92 (33.1%)	**272 (43.1%)****	121 (37.2%)	243 (41.6%)
BMI (kg/m^2)		25.6±3.5 (278)	**27.5±4.2***** (631)	23.0±1.6 (325)	**29.0±3.4**** (584)
Education		11.4±2.6 (278)	11.0±2.6 (631)	11.6±2.8 (325)	10.9±2.5*** (584)
TC (mg/dl)		225.6±39.0 (278)	228.88±41.0 (631)	226.7±38.0 (325)	228.5±41.7 (584)
HDL (mg/dl)		56.3±17.6 (276)	56.6±17.0 (625)	61.7±18.6 (323)	**53.6±15.6***** (578)
Never		163 (58.8%)	388 (61.5%)	231 (71.3%)	320 (54.8%)
Former		77 (27.8%)	180 (28.5%)	66 (20.3%)	191 (32.7%)
Current		37 (13.3%)	63 (10.0%)	27 (8.3%)	**73 (12.5%)*****
ApoE4					
No allele		216 (79.1%)	502 (80.3%)	257 (79.1%)	461 (79.8%)
Heterozygous		52 (19.0%)	121 (19.4%)	59 (18.4%)	114 (19.7%)
Homozygous		5 (1.8%)	2 (0.3%)	4 (1.3%)	3 (0.5%)
LTL		0.62(0.49-0.86) (278)	0.61(0.46-0.82) (631)	0.60(0.48-0.76) (325)	0.62(0.46-0.84) (584)
BPF (%)		79.7 ± 3.67 (231)	**78.2 ± 3.9***** (508)	78.8 ± 4.1 (266)	78.6 ± 3.8 (473)
WMH load (mm^3)		0.4(0.0-1.3) (258)	**1.3(0.3-4.3)***** (569)	0.8(0.2-3.0) (298)	0.9(0.2-3.3) (529)
WMH Score	0	75 (28.4%)	82 (13.9%)	61 (20.0%)	96 (17.6%)
	1	159 (60.2%)	344 (58.5%)	188 (61.6%)	315 (57.6%)
	2	22 (8.3%)	101 (17.2%)	33 (10.8%)	90 (16.5%)
	3	8 (3.0%)	**61 (10.4%)*****	23 (7.5%)	46 (8.4%)
Attention/ Speed		0.15±0.49 (267)	**-0.01±0.54***** (586)	0.07±0.51 (307)	0.02±0.54 (546)
Conceptualisation		0.16±0.59 (268)	**0.02±0.59**** (590)	0.11±0.58 (312)	0.03±0.60 (546)
Memory		0.25±1.04 (265)	**-0.10±0.99***** (595)	0.11±0.97 (311)	**-0.05±1.03*** (549)
Visuopractical Skills		0.29±0.87 (272)	**-0.13±0.92***** (619)	0.13±0.90 (324)	**-0.08±0.93***** (567)
g factor		0.33±0.90 (257)	**-0.14±1.02***** (566)	0.12±0.95 (298)	**-0.06±1.03*** (525)

Table 1 continue

		Education ≤10 (n)	**Education >10 (n)**	**No DM (n)**	**DM (n)**
Age (years)		66.2±7.9 (616)	65.4±8.1 (293)	65.6±7.9 (810)	**69.0±7.4***** (99)
Sex (Men)		233 (37.8%)	**155 (52.9%)*****	346 (42.7%)	**42 (42.4%)**
Hypertension		444 (72.1%)	**187 (63.8%)***	552 (68.2%)	**79 (79.8)***
SBP (mm of Hg)		145.3±23.3 (615)	**140.7±21.4**** (293)	143.0±22.5 (810)	**150.3±24.0**** (98)
DBP (mm of Hg)		87.9±10.1 (615)	87.0±9.7 (293)	87.5±9.7 (810)	88.0±10.2 (98)
DM		80 (13.0%)	**19 (6.5%)****	NA	NA
CVD		261 (42.4%)	**103 (35.2%)***	313 (38.6%)	51 (51.5%)*
BMI (kg/m^2)		27.3±4.2 (616)	**25.9±3.7***** (293)	26.6±4.0 (810)	**29.0±4.1***** (99)
Education		9.6±0.5 (616)	14.4±2.2*** (293)	11.2±2.6 (2.6)	**10.4±2.2**** (99)
TC (mg/dl)		228.9±42.0 (616)	225.8±36.7 (293)	228.7±39.8 (810)	221.4±44.8 (99)
HDL (mg/dl)		56.7±17.3 (609)	56.0±16.9 (292)	57.0±17.3 (8.3)	**52.1±15.8**** (98)
Smoking					
Never		383 (62.3%)	168 (57.3%)	504 (62.3%)	47 (47.5%)
Former		161 (26.2%)	96 (32.8%)	221 (27.3%)	36 (36.4%)
Current		71 (11.5%)	29 (9.9%)	84 (10.4%)	16 (16.2%)*
ApoE4					
No allele		482 (79.3%)	236 (81.4%)	637 (79.5%)	81 (83.5%)
Heterozygous		122 (20.1%)	51 (17.6%)	157 (19.6%)	16 (16.5%)
Homozygous		4 (0.7%)	3 (1.0%)	7 (0.9%)	0 (0.0%)
LTL		0.62(0.47-0.83) (616)	0.60(0.47-0.78) 293)	0.61(0.47-0.82) (810)	0.63(0.45-0.83) (99)
BPF (%)		78.6 ± 3.9 (505)	78.8 ± 3.9 (234)	78.9±3.8 (662)	**77.3±4.3***** (77)
WMH load(mm^3)		1.0(0.2-3.6) (564)	**0.6(0.1-2.5)*** (263)	0.8(0.2-3.0) (737)	1.5(0.3-6.1)** (90)
WMH Score	0	103 (17.8%)	54 (19.7%)	142 (18.7%)	15 (16.3%)
	1	340 (58.8%)	163 (59.5%)	459 (60.4%)	44 (47.8%)
	2	91 (15.7%)	32 (11.7%)	100 (13.2%)	23 (25.0%)
	3	44 (7.6%)	25 (9.1%)	59 (7.8%)	**10 (10.9%)***
Attention/ Speed		-0.04±0.55 (580)	**0.19±0.44***** (273)	0.06±0.5 (764)	**-0.15±0.56***** (89)
Conceptualisation		-0.03±0.60 (574)	**0.25±0.54***** (284)	0.08±0.60 (770)	**-0.80±0.54*** (88)
Memory		-0.22±0.88 (576)	**0.47±1.11***** (284)	0.05±1.02 (773)	**-0.32±0.89***** (87)
Visuopractical Skills		-0.04±0.94 (603)	0.09±0.90 (288)	0.06±0.90 (795)	**-0.49±1.00***** (96)
g factor		-0.20±0.95 (553)	**0.42±0.98***** (270)	0.07±0.99 (741)	**-0.54±1.03***** (82)

Table 1 continue

		No CVD (n)	CVD (n)	Apoe4 non-carriers (n)	Apoe4 carriers (n)
Age (years)		65.2±7.9 (545)	**67.1±8.0*** ** (364)	66.1±8.0 (718)	65.4±7.8 (180)
Sex (Men)		249 (45.7%)	**139 (38.2%)***	303 (42.2%)	83 (46.1)
Hypertension		359 (65.9%)	**272 (74.7%)****	502 (69.9%)	123 (68.3)
SBP (mm Hg)		143.0±23.0 (544)	145.1±22.5 (364)	143.8±22.6 (717)	144.3±23.7 (180)
DBP (mm Hg)		87.3±10.0 (544)	88.0±10.1 (364)	87.5±9.9 (717)	87.8±10.0 (180)
DM		48 (8.8%)	**51 (14.0%)***	81 (11.3%)	16 (8.8%)
CVD		NA	**NA**	285 (39.7%)	76 (42.2)
BMI (kg/m^2)		26.7±4.0 (545)	27.1±4.2 364	26.9±4.1 (718)	26.8±3.9 (180)
Education		11.3±2.7 (545)	**10.8±2.4**** (364)	11.2±2.7 (718)	11.0±2.5 (180)
TC (mg/dl)		227.2±39.8 (545)	228.9±41.3 (364)	227.0±41.3 (718)	232.4±36.8 (180)
HDL (mg/dl)		56.4±16.8 (541)	56.6±17.8 (360)	56.9±17.2 (712)	55.0±17.1 (178)
Smoking					
	Never	345 (65.1%)	206 (56.6%)	429 (59.8%)	116 (64.4%)
	Former	139 (25.6%)	118 (32.4%)	209 (29.1%)	46 (26.6%)
	Current	60 (11.0%)	40 (11.0%)	80 (11.1%)	18 (10.0%)
ApoE4					
	No allele	433 (80.6%)	285 (79.0%)	NA	NA
	Heterozygous	100 (18.6%)	73 (20.2%)		
	Homozygous	4 (0.8%)	3 (0.8%)		
LTL		0.63(0.48-0.86) (545)	**0.60(0.45-0.76)*** (364)	0.61(0.46-0.82) (718)	0.62(0.47-0.86) (180)
BPF (%)		79.0±3.8 (437)	**78.3±4.0*** (302)	78.6±3.9 (589)	79.0±4.0 (140)
WMH load(mm^3)		0.7(0.1-2.6) (494)	**1.2(0.3-4.0)***** (333)	0.80(0.20-2.90) (651)	**1.3(0.20-4.4)*** (165)
WMH Score	0	105 (20.7%)	52 (15.1%)	128 (19.0%)	26 (15.6%)
	1	302 (59.4%)	201 (58.4%)	399 (59.2%)	99 (59.3%)
	2	62 (12.2%)	61 (17.7%)	95 (14.1%)	25 (15.0%)
	3	39 (7.7%)	**30 (8.7%)***	52 (7.7%)	17 (10.2%)
Attention/ Speed		0.09±0.50 (518)	**-0.43±0.56***** (335)	0.04±0.52 (674)	0.02±0.57 (168)
Conceptualisation		0.13±0.54 (523)	**-0.40±0.66***** (337)	0.07±0.60 (683)	0.02±0.57 (164)
Memory		0.13±1.01 (523)	**-0.18±1.00***** (337)	0.01±1.00 (683)	0.05±1.05 (167)
Visuopractical Skills		0.08±0.91 (534)	**-0.13±0.95**** (357)	-0.01±0.93 (705)	0.00±0.88 (175)
g factor		0.14±0.98 (502)	**-0.19±1.02***** (321)	-0.05±1.01 (652)	0.05±0.99 (161)

*0.01<P< 0.05, ** 0.001<P< 0.01, *** < 0.001. *The independent sample t-test was performed for normally distributed continuous variables, Mann-Whitney U test for skewed variables, and the chi-*

square test for categorized variables to compare the groups. For normally distributed variables, mean ±SD, for skewed variables median (IQR), and categorical variables, numbers of observations (%) in the category are given. SBP: Systolic Blood Pressure, DBP: Diastolic Blood Pressure, CVD: Cardiovascular Diseases, BMI: Body Mass Index, TC: Total Cholesterol, HDL: High-Density Lipoprotein, LTL: Leukocyte Telomere Length, BPF: Brain Parenchymal Fraction, WMH: White Matter Hyperintensities, SD: Standard Deviation, IQR: Inter-Quartile Range [Table reproduced from Gampawar et al., Frontiers in Psychiatry 2020 [1]].

Table 2:Multiple regression models used to estimate effect of LTL on different brain phenotypes

	β	**SE**	***p***	**R^2**	**p F change**	**Partial R^2**
BPF (739)						
Model I	0.4430	0.1150	**0.0001**	0.3442	0.0000	0.0135
Model II	0.4285	0.1149	**0.0002**	0.3656	0.0013	0.0124
Model III	0.4290	0.1150	**0.0002**	0.3657	0.6751	0.0124
WMH load (827)						
Model I	0.0198	0.0116	0.0881	0.2628	0.0000	0.0027
Model II	0.0236	0.0116	**0.0424**	0.2745	0.0060	0.0037
Model III	0.0242	0.0116	**0.0367**	0.2832	0.0011	0.0039
WMH Score (852)						
	β	SE	*p*	R^2	p F change	Partial R^2
Model I	0.0469	0.0247	0.0579	0.1972	0.0000	0.0035
Model II	0.0515	0.0249	**0.0390**	0.2074	0.1609	0.0041
Model III	0.0519	0.0249	**0.0372**	0.2112	0.0471	0.0042
Attention/Speed (854)						
Model I	0.0253	0.0172	0.1425	0.2034	0.0000	0.0021
Model II	0.0248	0.0172	0.1482	0.2341	0.0000	0.0019
Model III	0.0248	0.0172	0.1490	0.2349	0.3444	0.0019
Conceptualisation (858)						
	β	SE	*p*	R^2	p F change	Partial R^2
Model I	0.0170	0.0197	0.3884	0.0474	0.0000	0.0008
Model II	0.0171	0.0192	0.3710	0.1293	0.0000	0.0008
Model III	0.0167	0.0191	0.3820	0.1310	0.2001	0.0008
Memory (860)						
Model I	-0.0217	0.0308	0.4826	0.1900	0.0000	0.0005
Model II	-0.0151	0.0290	0.6039	0.3040	0.0000	0.0002
Model III	-0.0148	0.0290	0.6098	0.3040	0.6325	0.0002
Visuopractical Skills (891)						
	β	SE	*p*	R^2	p F change	Partial R^2
Model I	0.0125	0.0238	0.5996	0.4143	0.0000	0.0002
Model II	0.0104	0.0237	0.6621	0.4347	0.0001	0.0001
Model III	0.0102	0.0237	0.6682	0.4350	0.4655	0.0001
g factor (823)						
Model I	0.0077	0.0280	0.7833	0.3520	0.0000	0.0001
Model II	0.0088	0.0263	0.7380	0.4436	0.0000	0.0001
Model III	0.0088	0.0264	0.7381	0.4436	0.9907	0.0001

Bold shows significance level of <0.05. BPF: Brain Parenchymal Fraction, WMH: White Matter Hyperintensities, β-Effect size, SE-Standard Error [Table reproduced from Gampawar et al., Frontiers in Psychiatry 2020[1]].

Table 3: Stratified analyses of the significant association of LTL with MRI correlates and cognitive functions

	β	SE	p	Partial R^2	β	SE	p	Partial R^2
BPF	Men (313)				Women (426)			
	0.330	0.174	0.058	0.008	0.505	0.155	**0.001**	0.016
	≤65 years (352)				>65 years (387)			
	0.350	0.142	**0.014**	0.013	0.583	0.186	**0.002**	0.016
	Normotensive (231)				Hypertensive (508)			
	0.348	0.187	0.064	0.011	0.505	0.148	0.001	0.015
	BMI <25 (266)				BMI ≥25 (473)			
	0.502	0.202	**0.013**	0.014	0.403	0.140	**0.004**	0.012
	Education ≤10 (505)				Education >10 (234)			
	0.424	0.138	**0.002**	0.012	0.418	0.213	0.051	0.011
	No DM (662)				DM (77)			
	0.418	0.117	**0.000**	0.013	0.250	0.575	0.665	0.002
	No CVD (437)				CVD (302)			
	0.328	0.140	**0.020**	0.008	0.584	0.207	**0.005**	0.017
	Apoe4 non-carriers (589)				Apoe4 carriers (140)			
	0.485	0.125	0.000	0.016	0.149	0.299	0.620	0.001
WMH load	Normotensive (258)				Hypertensive (596)			
	-0.003	0.015	0.860	0.000	0.038	0.016	**0.017**	0.008
	Education ≤10 (564)				Education >10 (263)			
	0.029	0.014	**0.039**	0.006	0.018	0.022	0.406	0.002
	No DM (737)				DM (90)			
	0.031	0.012	**0.008**	0.007	-0.076	0.048	0.115	0.021
	Apoe4 non-carriers (589)				Apoe4 carriers (140)			
	0.029	0.013	**0.020**	0.006	-0.001	0.030	0.975	0.000
WMH Score	Education ≤10 (578)				Education >10 (274)			
	0.069	0.030	**0.021**	0.008	0.023	0.047	0.633	0.001
	No DM (760)				DM (92)			
	0.061	0.026	**0.018**	0.006	-0.049	0.102	0.633	0.002
	No CVD (508)				CVD (344)			
	0.060	0.031	**0.049**	0.006	0.042	0.044	0.343	0.002
	Apoe4 non-carriers (674)				Apoe4 carriers (167)			
	0.070	0.027	**0.010**	0.008	-0.019	0.062	0.760	0.000
Attention /Speed	BMI <25 (308)				BMI ≥25 (546)			
	0.001	0.034	0.974	0.000	0.039	0.019	0.048	0.006
	Education ≤10 (581)				Education >10 (273)			
	0.044	0.022	**0.048**	0.005	-0.031	0.025	0.208	0.005

The results are shown from linear regression after adjusting for age, sex, risk factors, and ApoE4 genotypes (Model III) in subgroups. Bold shows significance level of <0.05. For each subgroup, number of participants is given in the brackets. BPF: Brain Parenchymal Fraction, WMH: White Matter Hyperintensities, BMI: Body Mass Index, DM: Diabetes Mellitus, CVD: Cardiovascular Disease, β-Effect size, SE-Standard Error [Table reproduced from Gampawar et al., Frontiers in Psychiatry 2020 [1]*].*

Table 4: Mediation analysis

Subgroups	Effect (β)	SE	CI Lower	CI Upper
BPF				
BMI ≥25	0.019	0.006	0.007	0.033
Education ≤10	0.019	0.006	0.008	0.032
WMH load				
Education ≤ 10	-0.008	0.005	-0.019	0.001
WMH Score				
Education ≤ 10	-0.0029	0.0026	-0.0088	0.0013

The effect size of the mediation calculated using bootstrapping is shown here. We use the 5000 bootstrap samples. When the confidence interval of the effect size does not include zero, it shows a significant effect. BPF: Brain Parenchymal Fraction, WMH: White Matter Hyperintensities, BMI: Body Mass Index, SE: Standard Error, CI: 95% Confidence Interval [Table reproduced from Gampawar et al., Frontiers in Psychiatry 2020 [1]].

Table 5:Comparison between TelSeq and Computel derived LTL

Sample ID	Age	TelSeq	Computel	Difference
200	75	829.721	765.82	63.901
G178202	77	997.649	964.33	33.319
G178599	56	1197.66	1136.6	61.06
G178603	57	1305.4	1216.29	89.11
G179864	75	1216.94	1146.99	69.95

Table 6:Comparison of laboratory protocols and design of AmpliSeq and SureSelect library preparation methods

		AmpliSeq	SureSelect
Protocol	Enrichment approach	PCR	Hybridisation
	DNA input	100 ng	1µg
	Steps in library preparation	3	8
	DNA fragmentation	NA	Enzymatic fragmentation
	Target selection	Amplification using primers	Hybridisation with RNA library baits
	Incubation time	~4 hours	~26 hours
	Library Preparation time	6 hours	2.5 days
Target Region	Total target region	57.74MB	60.45MB
	RefSeq coding	32.30MB (91.13%)	31.15MB (87.88%)
	UCSC coding	32.57MB (88.65%)	32.26MB (87.78%)
	Ensembl coding	32.40MB (87.90%)	32.23MB (87.44%)
	Effective target region	46.35MB	NA
	RefSeq coding	30.58 MB (86.28%)	NA
	UCSC coding	30.76MB (83.72%)	NA
	Ensembl coding	30.62MB (83.07%)	NA

NA - not applicable, MB - million bases [Table reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]]

Table 7:Variant validation of default TVC output (VCF1) against NA12878 truth set

	Total Variants	Truthset	TPs	FNs	FPs	Sensitivity (%)	PPV (%)
AmpliSeq							
Total Variants	54,351	49,340	45,946	3,394	8,405	93.12	84.54
Total SNVs	50,913	45,092	43,840	1,252	7,073	97.22	86.11
Exonic SNVs	19,650	16,964	16,588	376	3,062	97.78	84.42
Total Indels	3,436	4,248	2,106	2,142	1,330	49.58	61.29
Exonic indels	539	329	231	98	308	70.21	42.86
SureSelect							
Total Variants	54,934	46,982	43,929	3,053	11,005	93.50	79.97
Total SNVs	52,013	43,367	42,230	1,137	9,783	97.38	81.19
Exonic SNVs	19,171	16,120	15,846	274	3,325	98.30	82.66
Total Indels	2,921	3,614	1,699	1,915	1,222	47.01	58.17
Exonic indels	312	277	195	82	117	70.40	62.50

Truth set-Variants in v3.3.2 of high confidence calls VCF of NA12878 from Genome in the Bottle project, SNVs- Single Nucleotide Variants, TPs- True Positives, FNs- False Negatives, FPs- False Positives, PPV-Positive Predictive Value, TVC- Torrent Variant Caller, VCF1- This corresponds to figure 7 [Table reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]].

Table 8:Variant validation in various steps of optimization using NA12878 truth set

Steps	Total Variants	Truthset	TPs	FNs	FPs	Sensitivity (%)	PPV (%)
AmpliSeq							
TTR (VCF1)	54,351	49,340	45,946	3,394	8,405	93.12	84.54
RG (VCF2)	55,241	49,340	46,660	2,680	8,581	94.57	84.47
HCR (VCF3)	47,538	48,796	46,320	2,476	1,218	94.93	97.44
SureSelect							
TTR (VCF1)	54,934	46,982	43,929	3,053	11,005	93.50	79.97
RG (VCF2)	55,831	46,982	44,551	2,431	11,280	94.83	79.80
HCR (VCF3)	45,200	46,557	44,253	2,304	947	95.05	97.91

Truth set-Variants in v3.3.2 of high confidence calls VCF of NA12878 from Genome in the Bottle project, SNVs- Single Nucleotide Variants, TPs- True Positives, FNs- False Negatives, FPs- False Positives, PPV-Positive Predictive Value, VCF1-3 - This corresponds to figure 1, TTR - Total Target Region, RG – Regularization, HCR - High Confidence Region [Table reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]].

Table 9:Comparing performance of AmpliSeq vs SureSelect within RefSeq coding region and overlapping target region

	Steps	Total Variants	Truthset	TPs	FNs	FPs	Sensitivity (%)	PPV(%)
RefSeq coding region	**AmpliSeq**							
	TTR (VCF1)	21,584	19,270	17,836	1,434	3,748	92.56	82.64
	RG (VCF2)	21,878	19,270	18,087	1,183	3,791	93.86	82.67
	HCR (VCF3)	18,331	19,270	18,009	1,261	322	93.46	98.24
	SureSelect							
	TTR (VCF1)	21,649	19,270	17,312	1,958	4,337	89.84	79.97
	RG (VCF2)	21,943	19,270	17,523	1,747	4,420	90.93	79.86
	HCR (VCF3)	17,747	19,270	17,443	1,827	304	90.52	98.29
OTR 115X	**AmpliSeq**							
	TTR (VCF1)	35,093	32,213	30,367	1,846	4,723	94.27	86.53
	RG (VCF2)	35,550	32,213	30,788	1,425	4,762	95.58	86.60
	HCR (VCF3)	31,161	31,979	30,611	1,368	550	95.72	98.23
	SureSelect							
	TTR (VCF1)	36,228	32,213	30,651	1,562	5,577	95.15	84.61
	RG (VCF2)	36,744	32,213	31,067	1,146	5,677	96.44	84.55
	HCR (VCF3)	31,478	31,979	30,896	1,083	582	96.61	98.15

Overlapping target region (OTR) - this is the common region covered by both AS and SS design. This region is 43.17Mb, RefSeq coding region -this is the coding region from RefSeq database, downloaded from UCSC table browser on 20/04/2017. NA12878 truth set has 19270 variants, Truth set-Variants in v3.3.2 of high confidence calls VCF of NA12878 from Genome in the Bottle project, SNVs- Single Nucleotide Variants, TPs- True Positives, FNs- False Negatives, FPs- False Positives, PPV-Positive Predictive Value, VCF1-3-This corresponds to figure 1, RG – Regularization, HCR - High Confidence Region [Table reproduced from Gampawar et al., Frontiers in Genetics 2019[2]].

Supplementary tables

Supplementary table 1: Target enrichment efficiency with AmpliSeq and SureSelect methods.

Sample ID	GWAS SNPs passing QC	Exome chip SNPs passing QC	**AmpliSeq**						**SureSelect**					
			Total Reads	Mapped Reads	Mapped Reads %	Reads on Target %	Mean Read Length	Mean Read Depth	Total Reads	Mapped reads	Mapped Reads %	Reads on Target %	Mean Read Length	Mean Read Depth
26500	501,288	242,688	36,844,784	36,514,871	99.1	94.9	173	102	33,403,340	33,267,894	99.6	87.0	115	51
59301	501,288	242,688	29,945,578	29,789,626	99.5	95.3	177	84	41,990,192	41,802,769	99.6	84.9	149	75
90901	501,288	242,688	34,829,164	34,582,922	99.3	95.1	176	98	44,110,131	43,935,592	99.6	84.8	151	80
152301	501,288	242,688	35,414,963	34,943,911	98.7	93.6	157	87	40,376,249	40,200,406	99.6	93.6	149	87
168002	501,288	242,688	20,064,165	19,832,304	98.8	89.1	180	56	53,862,426	53,563,459	99.4	85.8	149	98
801019	491,791	244,688	35,704,021	35,500,942	99.4	95.8	180	104	37,144,327	36,983,305	99.6	87.2	102	51
801020	491,791	NA	50,119,270	49,782,561	99.3	95.9	181	146	40,164,626	39,991,973	99.6	85.8	110	58
801024	491,791	244,688	28,924,948	28,632,133	99.0	95.1	152	69	33,479,566	33,339,769	99.6	82.4	147	58
801042	491,791	244,688	30,488,385	30,178,632	99.0	95.7	153	74	35,129,719	34,927,962	99.4	82.9	145	60
801052	491,791	244,688	27,983,305	27,560,123	98.5	93.4	139	60	45,745,431	45,491,089	99.4	84.3	146	80
801058	491,791	NA	41,756,741	41,288,706	98.9	90.2	184	118	26,070,226	25,979,250	99.7	84.5	150	47
13	501,288	NA	38,495,116	38,051,197	98.8	94.9	171	105	46,329,776	46,148,050	99.6	86.8	138	82
Mean±SD			34,214,203 ±7,617,405	33,889,119 ±7,577,988	99±0.3	94.1±2.2	169±15	92±26	39,817,167 ±7,371,996	39,635,960 ±7,324,890	99.6±0.1	85.8±2.9	133±18	69±17

NA: Not Available [Table reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]]

Supplementary table 2:Variants detected with AmpliSeq and SureSelect libraries using different target regions.

Sample ID	AmpliSeq TTR	SureSelect TTR	AmpliSeq ETR	Difference AmpliSeq TTR vs AmpliSeq ETR (%)
26500	51,704	48,629	37,679	14,025 (27.13)
59301	50,343	53,562	36,576	13,767 (27.35)
90901	52,026	53,492	37,523	14,503 (27.88)
152301	51,875	52,748	37,316	14,559 (28.07)
168002	48,383	52,668	35,625	12,758 (26.37)
809019	52,097	48,087	37,662	14,435 (27.72)
809020	52,325	49,622	37,722	14,603 (27.91)
801024	51,279	52,422	37,198	14,081 (27.46)
801042	51,053	52,416	37,081	13,972 (27.37)
801052	51,018	54,284	37,105	13,913 (27.27)
801058	52,996	51,895	38,186	14,810 (77.95)
13	51,860	51,566	37,547	14,313 (27.60)
Mean ± SD	51,413±1,180	51,783±1,982	37,268±658	14,145±542 (27.50±0.47)

TTR –Total Target Region, ETR –Effective Target Region [Table reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]]

Supplementary table 3:Concordance between sequencing and microarray genotyping.

Exome chip (n=9)						
Sample ID	AmpliSeq			SureSelect		
	Common Positions	Concordant variants	Concordance rate	Common Positions	Concordant variants	Concordance rate
26500	7,142	7,009	98.1%	7,245	7,062	97.5%
59301	6,851	6,709	97.9%	7,329	7,154	97.6%
90901	7,093	6,966	98.2%	7,406	7,229	97.6%
152301	6,950	6,799	97.8%	7,302	7,113	97.4%
168002	6,791	6,669	98.2%	7,377	7,226	98.0%
801019	7,113	6,965	97.9%	7,174	6,974	97.2%
801020	NA					
801024	7,080	6,946	98.1%	7,402	7,236	97.8%
801042	6,995	6,858	98.0%	7,389	7,219	97.7%
801052	7,021	6,836	97.4%	7,465	7,277	97.5%
801058	NA					
13	NA					
Mean ± SD	7,004±121	6,862±102	97.97±0.26%	7,343±91	7,166±99	97.58±0.22%
GWAS chip (n=12)						
Sample ID	AmpliSeq			SureSelect		
	Common Positions	Concordant variants	Concordance rate	Common Positions	Concordant variants	Concordance rate
26500	3,759	3,740	99.5%	4,826	4,790	99.3%
59301	3,523	3,494	99.2%	4,924	4,892	99.4%
90901	3,779	3,768	99.7%	5,034	5,018	99.7%
152301	3,698	3,681	99.5%	5,055	5,024	99.4%
168002	3,541	3,526	99.6%	4,946	4,926	99.6%
801019	3,626	3,612	99.6%	4,657	4,629	99.4%
801020	3,711	3,700	99.7%	4,787	4,767	99.6%
801024	3,628	3,604	99.3%	4,891	4,861	99.4%

801042	3,697	3,673	99.4%	4,991	4,966	99.5%
801052	3,626	3,596	99.2%	4,976	4,958	99.6%
801058	3,718	3,715	99.9%	4,918	4,879	99.2%
13	6,518	6,474	99.3%	8,300	8,257	99.5%
Mean ± SD	3,664±83	3,646±87	99.51±0.23%	4,910±116	4,883±118	99.45±0.16%

Data from two exome chip samples (801058 & 801020) was not used due to its low quality and sample 13 didn't have exome chip data. Genotyped with Affymetrix Genome-Wide Human SNP Array 6.0 Illumina Infinium Exome-24 v1.1 BeadChip and Illumina Human610-Quad BeadChip. The concordance rate was calculated using VCF 1 (figure 7). [Table reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]].

***Supplementary table 4**:Coverage per base level in AmpliSeq and SureSelect libraries with NA12878 DNA.*

Bases covered	AmpliSeq	%	SureSelect	%
<5X	400,116	0.69	371,807	0.61
<10X	587,610	1.02	671,090	1.11
5X-10X	220,562	0.38	378,922	0.63
5X-400X	45,847,696	79.40	59,541,459	98.49
10X-400X	45,660,202	79.08	59,242,176	97.99
>400X	11,494,834	19.91	543,697	0.90
Total target bases	57,742,646		60,456,963	
Average read depth	270X		115X	

[Table reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]].

Supplementary table 5:Variant validation of optimised output (VCF3) against NA12878 truth set.

	Total Variants	Truthset	TPs	FNs	FPs	Sensitivity	PPV
AmpliSeq							
Total Variants	47,535	48,796	46,320	2,476	1,215	94.93%	97.44%
Total SNVs	44,969	44,771	44,199	572	770	98.72%	98.29%
Exonic SNVs	16,953	16,895	16,737	158	216	99.06%	98.73%
Total Indels	2,566	4,025	2,121	1,904	445	52.70%	82.66%
Exonic indels	326	309	230	79	96	74.43%	70.55%
SureSelect							
Total Variants	45,194	46,556	44,253	2,303	941	95.05%	97.92%
Total SNVs	43,169	43,075	42,543	532	626	98.76%	98.55%
Exonic SNVs	16,187	16,054	15,966	88	221	99.45%	98.63%
Total Indels	2,025	3,481	1,710	1,771	315	49.12%	84.44%
Exonic indels	255	263	188	75	67	71.48%	73.73%

Truth set-Variants in v3.3.2 of high confidence calls VCF of NA12878 from Genome in the Bottle project, SNVs- Single Nucleotide Variants, TPs- True Positives, FNs- False Negatives, FPs- False Positives, PPV-Positive Predictive Value [Table reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]].

Supplementary table 6: Indel detection by AmpliSeq and SureSelect.

Indels	AmpliSeq specific	SureSelect specific	Overlapping indels
True Positives	1,251	841	870
False Negatives	1,261	1,128	642
False Positives	437	310	31

[Table reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]].

Supplementary table 7:Effect of parameter settings on variant detection

Step	Total Variants	TPs	FNs	FPs	Sensitivity	PPV
AmpliSeq						
1	47,538	46,320	2,476	1,218	94.93%	97.44%
2	47,061	46,177	2,619	884	94.63%	98.12%
3	46,594	45,832	2,964	762	93.93%	98.36%
4	46,374	45,639	3,157	735	93.53%	98.42%
5	45,957	45,248	3,548	709	92.73%	98.46%
SureSelect						
Step	Total Variants	TPs	FNs	FPs	Sensitivity	PPV
1	45,191	44,250	2,307	941	95.05%	97.92%
2	44,173	43,542	3,015	631	93.52%	98.57%
3	44,027	43,422	3,135	605	93.27%	98.63%
4	43,806	43,209	3,348	597	92.81%	98.64%
5	43,690	43,103	3,454	587	92.58%	98.66%

Step 1 - MiAF=0.1 MiCo=5 MiCo/str=0 MxStrBi=0.98
Step 2 - MiAF=0.2 MiCo=5 MiCo/str=0 MxStrBi=0.98
Step 3 - MiAF=0.2 MiCo=5 MiCo/str=2 MxStrBi=0.8
Step 4 - MiAF=0.2 MiCo=10 MiCo/str=3 MxStrBi=0.8
Step 5 - MiAF=0.2 MiCo=10 MiCo/str=5 MxStrBi=0.8
MiAF- minimum allele frequency, MiCo-minimum coverage, MiCo/str- minimum coverage on either strand, MxStrBi- maximum strand bias, TPs- True Positives, FNs- False Negatives, FPs- False Positives, PPV-Positive Predictive Value [Table reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]].

Supplementary table 8:Variant validation against NA12878 truth set in downsampled BAM

	Steps	Total Variants	Truthset	TPs	FNs	FPs	Sensitivity	PPV
34 Million reads	AmpliSeq							
	TTR (VCF1)	53,068	49,340	45,267	4,073	7,801	91.75%	85.30%
	RG (VCF2)	53,890	49,340	45,949	3,391	7,941	93.13%	85.26%
	HCR (VCF3)	46,697	48,796	45,613	3,183	1,084	93.48%	97.68%
	SureSelect							
	TTR (VCF1)	52,918	46,982	42,763	4,219	10,155	91.02%	80.81%
	RG (VCF2)	53,705	46,982	43,346	3,636	10,359	92.26%	80.71%
	HCR (VCF3)	43,951	46,557	43,057	3,500	894	92.48%	97.97%
100X Average depth	AmpliSeq							
	TTR (VCF1)	53,121	49,340	45,298	4,042	7,823	91.81%	85.27%
	RG (VCF2)	53,941	49,340	45,981	3,359	7,960	93.19%	85.24%
	HCR (VCF3)	46,745	48,796	45,648	3,148	1,097	93.55%	97.65%
	SureSelect							
	TTR (VCF1)	54,612	46,982	43,734	3,240	10,878	93.09%	80.08%
	RG (VCF2)	55,511	46,982	44,353	2,629	11,158	94.40%	79.90%
	HCR (VCF3)	45,043	46,557	44,054	2,503	989	94.62%	97.80%

Truth set-Variants in v3.3.2 of high confidence calls VCF of NA12878 from Genome in the Bottle project, SNVs- Single Nucleotide Variants, TPs- True Positives, FNs- False Negatives, FPs- False Positives, PPV-Positive Predictive Value, VCF1-3 - This corresponds to figure 1, TTR - Total Target Region, RG – Regularization, HCR - High Confidence Region. [Table reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]].

Supplementary table 9:Variant validation against NA12878 truthset in AmpliSeq ETR

AmpliSeq							
Steps	Total Variants	Truthset	TPs	FNs	FPs	Sensitivity	PPV
ETR (VCF1)	38,651	34,370	33,119	1,251	5,532	96.36%	85.69%
RG (VCF2)	39,112	34,370	33,559	811	5,553	97.64%	85.80%
HCR (VCF3)	33,888	34,118	33,383	735	505	97.85%	98.51%

Truthset- Variants in v3.3.2 of high confidence calls vcf file of NA12878 from Genome in the Bottle project, TPs- True Positives, FNs- False Negatives, FPs- False Positives, PPV-Positive Predictive

Supplementary table 10:Variant validation against NA12878 after duplicate removal

	Total Variants	Truthset	TPs	FNs	FPs	Sensitivity	PPV
AmpliSeq							
PICARD	88% loss of reads						
Samtools	87.8% loss of reads						
Start End	46,720	48,796	45,609	3,187	1,111	93.47%	97.62%
SureSelect							
PICARD	44,662	46,557	43,886	2,691	796	94.26%	98.26%
Samtools	44,662	46,557	43,886	2,691	796	94.26%	98.26%
Start End	45,041	46,557	44,053	2,504	988	94.62%	97.81%

Truthset- Variants in v3.3.2 of high confidence calls vcf file of NA12878 from Genome in the Bottle project, TPs- True Positives, FNs- False Negatives, FPs- False Positives, PPV-Positive Predictive Value [Table reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]]

Supplementary table 11:Distribution of false negatives from chromosome 1 in various categories

AmpliSeq																
	Library derived							Sequencing derived						Both	Unknown	Total
	Min coverage	Min coverage on either strand	Max strand bias	Min quality	Min relative read quality	Mixed	Total	Max homopolymer length	Excess outlier reads	Max common signal shift	Max reference/ variant signal shift	Mixed	Total			
SNVs	2(5%) (67%)	1(2%) (20%)	2(5%) (40%)	2(5%) (29%)	0 (0%) (0%)	34(83%) (60%)	41 (53%)	0 (0%) (0%)	0 (0%) (0%)	9(90%) (32%)	0 (0%) (0%)	1(10%) (9%)	10 (20%)	1 (2%)	3 (60%)	55 (31%)
Indels	1(3%) (33%)	4(11%) (80%)	3(8%) (60%)	5(13%) (71%)	1(3%) (100%)	23(62%) (40%)	37 (47%)	8(21%) (100%)	2(5%) (100%)	19(49%) (68%)	0 (0%) (0%)	10(25%) (91%)	39 (80%)	46 (98%)	2 (40%)	124 (69%)
Total	3(4%)	5(6%)	5(6%)	7(9%)	1(1%)	57(73%)	78	8(16%)	2(4%)	28(57%)	0(0%)	11(22%)	49	47	5	179
Group Total	78 (43.58%)							49 (27.37%)						47 (26.3%)	5 (2.79%)	179
SureSelect																
	Library derived							Sequencing derived						Both	Unknown	Total
	Min coverage	Min coverage on either strand	Max strand bias	Min quality	Min relative read quality	Mixed	Total	Max homopolymer length	Excess outlier reads	Max common signal shift	Max reference / variant signal shift	Mixed	Total			
SNVs	0 (0%) (0%)	0 (0%) (0%)	6(21%) (86%)	1(4%) (14%)	0 (0%) (0%)	21(75%) (81%)	28 (65%)	0 (0%) (0%)	0 (0%) (0%)	0 (0%) (0%)	0 (0%) (0%)	0 (0%) (0%)	0 (0%)	0 (0%)	2 (22%)	30 (22%)
Indels	0 (0%) (0%)	3(20%) (100)	1(7%) (14%)	6(40%) (86%)	0 (0%) (0%)	5(33%) (19%)	15 (35%)	10(29%) (100%)	3(9%) (100%)	11(30%) (100%)	1(3%) (100%)	10(29%) (100%)	35 (100%)	49 (100%)	7 (78%)	106 (78%)
Total	0(0%)	3(7%)	7(16%)	7(16%)	0(0%)	26(61%)	43	10(29%)	3(9%)	11(30%)	1(3%)	10(29%)	35	49	9	136
Group Total	43 (31.62%)							35 (25.74%)						49 (36 %)	9 (6.6%)	136

Max-maximum, Min-Minimum [Table reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]].

Supplementary table 12:Distribution of false negatives from chromosome 7 in various categories

AmpliSeq																
	Library derived							Sequencing derived						Both	Unknown	Total
	Min coverage	Min coverage on either strand	Max strand bias	Min quality	Min relative read quality	Mixed	Total	Max homopolymer length	Excess outlier reads	Max common signal shift	Max reference/variant signal shift	Mixed	Total			
SNVs	0 (0%) (0%)	1(7%) (25%)	2(14%) (100%)	0 (0%) (0%)	0 (0%) (0%)	11(79%) (55%)	14 (50%)	0 (0%) (0%)	0 (0%) (0%)	3(100%) (23%)	0 (0%) (0%)	0 (0%) (0%)	3 (8%)	1 (4%)	0 (0%)	18 (19%)
Indels	0 (0%) (0%)	3(21%) (75%)	0 (0%) (0%)	2(14%) (100%)	0 (0%) (0%)	9(64%) (45%)	14 (50%)	16(43%) (100%)	4(11%) (100%)	10(27%) (77%)	1 (3%) (100%)	6(16%) (100%)	37 (93%)	27 (96%)	0 (0%)	78 (81%)
Total	0(0%)	4 (14%)	2(7%)	2(7%)	0(0%)	20 (71%)	28	16(40%)	4(10%)	13(33%)	1(3%)	6(15%)	40	28	0	96
Group Total	28 (29.2%)							40 (41.2%)						28 (29.2%)	0 (0%)	96
SureSelect																
	Library derived							Sequencing derived						Both	Unknown	Total
	Min coverage	Min coverage on either strand	Max strand bias	Min quality	Min relative read quality	Mixed	Total	Max homopolymer length	Excess outlier reads	Max common signal shift	Max reference / variant signal shift	Mixed	Total			
SNVs	0 (0%) (0%)	0 (0%) (0%)	1(7%) (50%)	4(29%) (50%)	0 (0%) (0%)	9(64%) (69%)	14 (61%)	0 (0%) (0%)	0 (0%) (0%)	1 (100%) (13%)	0 (0%) (0%)	0 (0%) (0%)	1 (4%)	0 (0%)	0 (0%)	15 (39%)
Indels	0 (0%) (0%)	0 (0%) (0%)	1(11%) (50%)	4(44%) (50%)	0 (0%) (0%)	4(44%) (31%)	9 (39%)	13(54%) (100%)	1(4%) (100%)	7(29%) (88%)	1 (4%) (100%)	2(8%) (100%)	24 (96%)	32 (94%)	0 (0%)	65 (61%)
Total	0(0%)	0(0%)	2(9%)	8(35%)	0(0%)	13(57%)	23	13(52%)	1(4%)	8(32%)	1(4%)	2(8%)	25	32	0	80
Group Total	23 (28.8%)							25 (31.3%)						32 (40%)	0 (0%)	80

Max-maximum, Min-Minimum [Table reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]].

Supplementary table 13:Distribution of false negatives from chromosome 16 in various categories

AmpliSeq																
	Library derived							Sequencing derived						Both	Unknown	Total
	Min coverage	Min coverage on either strand	Max strand bias	Min quality	Min relative read quality	Mixed	Total	Max homopolymer length	Excess outlier reads	Max common signal shift	Max reference/ variant signal shift	Mixed	Total			
SNVs	0 (0%) (0%)	0 (0%) (0%)	6(20%) (75%)	1(3%) (100%)	0 (0%) (0%)	23(77%) (68%)	30 (67%)	0 (0%) (0%)	0 (0%) (0%)	1(100%) (10%)	0 (0%) (0%)	0 (0%) (0%)	1 (5%)	1 (5%)	0 (0%)	32 (39%)
Indels	0 (0%) (0%)	2(13%) (100%)	2(13%) (25%)	0 (0%) (0%)	0 (0%) (0%)	11(73%) (32%)	15 (33%)	5(28%) (100%)	0 (0%) (0%)	9(50%) (90%)	0 (0%) (0%)	4(22%) (100%)	18 (95%)	18 (95%)	0 (0%)	51 (61%)
Total	0(0%)	2 (4%)	8 (18%)	1 (2%)	0(0%)	34 (76%)	45	5(26%)	0(0%)	10(53%)	0(3%)	4(21%)	19	19	0	83
Group Total	45 (54.2%)							19 (22.9%)						19 (22.9%)	0 (0%)	83
SureSelect																
	Library derived							Sequencing derived						Both	Unknown	Total
	Min coverage	Min coverage on either strand	Max strand bias	Min quality	Min relative read quality	Mixed	Total	Max homopolymer length	Excess outlier reads	Max common signal shift	Max reference/ variant signal shift	Mixed	Total			
SNVs	1 (100%) (0%)	0 (0%) (0%)	3(75%) (86%)	0 (0%) (0%)	0(0%) (0%)	13(76%) (76%)	17 (71%)	2 (50%) (22%)	1 (25%) (50%)	1 (25%) (10%)	0 (0%) (0%)	0 (0%) (0%)	4 (17%)	0 (0%)	0 (0%)	21 (31%)
Indels	0 (0%) (0%)	0 (0%) (0%)	1(25%) (14%)	0 (0%) (0%)	2 (29%) (100%)	4(57%) (24%)	7 (29%)	7(35%) (78%)	1(5%) (50%)	9(45%) (90%)	0 (0%) (0%)	3(15%) (100%)	20 (83%)	19 (100%)	0 (0%)	46 (69%)
Total	1(4%)	0(0%)	4(17%)	0 (0%)	2(8%)	17(71%)	24	9(38%)	2(8%)	10(42%)	0(0%)	3(13%)	24	19	0	67
Group Total	24 (35.8%)							24 (35.8%)						19 (28.4%)	0 (0%)	67

Max-maximum, Min-Minimum [Table reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]*].*

Supplementary table 14:Distribution of false negatives from chromosome 19 in various categories

AmpliSeq																
	Library derived							Sequencing derived						Both	Unknown	Total
	Min coverage	Min coverage on either strand	Max strand bias	Min quality	Min relative read quality	Mixed	Total	Max homopolymer length	Excess outlier reads	Max common signal shift	Max reference/variant signal shift	Mixed	Total			
SNVs	0 (0%) (0%)	1(2%) (11%)	6(15%) (75%)	3(7%) (75%)	0 (0%) (0%)	31(76%) (69%)	41 (62%)	0 (0%) (0%)	3 (60%) (50%)	2(40%) (32%)	0 (0%) (0%)	0 (0%) (0%)	5 (14%)	0 (0%)	0 (0%)	46 (37%)
Indels	0 (0%) (0%)	8(32%) (89%)	2(8%) (25%)	1(4%) (25%)	0 (0%) (0%)	14(56%) (31%)	25 (38%)	7(23%) (100%)	3(10%) (50%)	17(57%) (68%)	1 (3%) (0%)	2(7%) (91%)	30 (86%)	20 (100%)	2 (100%)	77 (63%)
Total	0(0%)	9 (14%)	8(12%)	4 (6%)	0(0%)	45 (68%)	66	7(20%)	6(17%)	19(54%)	1(3%)	2(6%)	35	20	2	123
Group Total	66 (53.7%)							35 (28.5%)						20 (16.3%)	2 (1.6%)	123
SureSelect																
	Library derived							Sequencing derived						Both	Unknown	Total
	Min coverage	Min coverage on either strand	Max strand bias	Min quality	Min relative read quality	Mixed	Total	Max homopolymer length	Excess outlier reads	Max common signal shift	Max reference/variant signal shift	Mixed	Total			
SNVs	0 (0%) (0%)	1 (2%) (0%)	4(10%) (86%)	2(5%) (14%)	0(0%) (0%)	34(83%) (81%)	41 (65%)	0 (0%) (0%)	0 (0%) (0%)	0 (0%) (0%)	0 (0%) (0%)	0 (0%) (0%)	0 (0%)	2 (6%)	0 (0%)	43 (39%)
Indels	0 (0%) (0%)	2(14%) (100)	1(7%) (14%)	4(29%) (86%)	0 (0%) (0%)	7(50%) (19%)	14 (35%)	10(42%) (100%)	3(13%) (100%)	10(42%) (100%)	0 (0%) (0%)	1(4%) (100%)	24 (100%)	29 (94%)	0 (0%)	67 (61%)
Total	0(0%)	3(5%)	5(9%)	6(11%)	0(0%)	41(75%)	55	10(100%)	3(100%)	10(100%)	0(0%)	1(100%)	24	31	0	110
Group Total	55 (50%)							24 (21.8%)						31 (28.2%)	0 (0%)	110

Max-maximum, Min-Minimum [Table reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]]

Supplementary table 15:Distribution of false negatives from chromosome X in various categories

AmpliSeq																
	Library derived							Sequencing derived						Both	Unknown	Total
	Min coverage	Min coverage on either strand	Max strand bias	Min quality	Min relative read quality	Mixed	Total	Max homopolymer length	Excess outlier reads	Max common signal shift	Max reference/variant signal shift	Mixed	Total			
SNVs	0 (0%) (0%)	0 (0%) (0%)	0 (0%) (0%)	2(17%) (100%)	0 (0%) (0%)	10(83%) (50%)	12 (46%)	1(100%) (20%)	0 (0%) (0%)	0 (0%) (0%)	0 (0%) (0%)	0(0%) (0%)	1 (6%)	1 (7%)	0 (0%)	14 (25%)
Indels	1(8%) (100%)	3 (21%) (100%)	0 (0%) (0%)	0 (0%) (0%)	0 (0%) (0%)	10(71%) (50%)	14 (54%)	4(25%) (80%)	1(6%) (100%)	7(44%) (100%)	0 (0%) (0%)	4(25%) (100%)	16 (94%)	13 (93%)	0 (0%)	43 (75%)
Total	1(4%)	3(11%)	0(0%)	2(8%)	0(0%)	20(77%)	26	5(29%)	1(6%)	7(41%)	0(0%)	4(24%)	17	14	0	57
Group Total	26 (45.6%)							17 (29.8%)						14 (24.6%)	0 (0%)	57
SureSelect																
	Library derived							Sequencing derived						Both	Unknown	Total
	Min coverage	Min coverage on either strand	Max strand bias	Min quality	Min relative read quality	Mixed	Total	Max homopolymer length	Excess outlier reads	Max common signal shift	Max reference/variant signal shift	Mixed	Total			
SNVs	0 (0%) (0%)	0 (0%) (0%)	0 (0%) (0%)	1(6%) (100%)	0 (0%) (0%)	15(94%) (79%)	16 (76%)	0 (0%) (0%)	0 (0%) (0%)	0 (0%) (0%)	0 (0%) (0%)	0 (0%) (0%)	0 (0%) (0%)	1 (5%)	0 (0%)	17 (34%)
Indels	0 (0%) (0%)	0 (0%) (0%)	1(20%) (100%)	0 (0%) (0%)	0 (0%) (0%)	4(80%) (21%)	5 (24%)	0 (0%) (0%)	1(13%) (100%)	3(37%) (100%)	0 (0%) (0%)	4(50%) (100%)	8 (100%)	19 (95%)	1 (100%)	33 (66%)
Total	0(0%)	0(0%)	1(5%)	1(5%)	0(0%)	19(90%)	21	0(0%)	1(13%)	3(37%)	0(0%)	4(50%)	8	20	1	50
Group Total	21 (42%)							8 (16%)						20 (40%)	1 (2%)	50

Max-maximum, Min-Minimum [Table reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]]

Supplementary table 16:False negative SNVs and indels in AmpliSeq and SureSelect

		Region	Library derived (%)	Sequencer derived (%)	Library + Sequencer (%)	Unknown (%)	Total (N)
SNVs	**AmpliSeq**						
	Chr1	TTR	74.5	18.2	1.8	5.5	55
		Specific	81.0	14.3	4.8	0.0	21
	ChrX	TTR	85.7	7.1	7.1	0.0	14
		Specific	75.0	25.0	0.0	0.0	4
	SureSelect						
	Chr1	TTR	93.3	0.0	0.0	6.7	30
		Specific	100.0	0.0	0.0	0.0	14
	ChrX	TTR	94.1	0.0	5.9	0.0	17
		Specific	92.9	0.0	7.1	0.0	14
Indels	**AmpliSeq**						
	Chr1	TTR	29.8	31.5	37.1	1.6	124
		Specific	30.9	33.8	30.9	4.4	68
	ChrX	TTR	32.6	37.2	30.2	0.0	43
		Specific	36.8	36.8	21.1	5.3	19
	SureSelect						
	Chr1	TTR	14.2	33.0	46.2	6.6	106
		Specific	13.0	43.5	41.3	2.2	46
	ChrX	TTR	15.2	24.2	57.6	3.0	33
		Specific	6.7	40.0	46.7	6.7	15

TTR - Total Target Region, Specific - Library specific region (AmpliSeq = 11.4 Mb, SureSelect = 14.1 Mb) SNV- Single Nucleotide Variation [Table reproduced from Gampawar et al., Frontiers in Genetics 2019[2]].

Supplementary table 17:False negative SNVs and indels in AmpliSeq and SureSelect.

		Library derived (%)	Sequencer derived (%)	Library + Sequencer (%)	Unknown (%)	Total (N)
SNVs	**AmpliSeq**					
	Chr1	74.5	18.2	1.8	5.5	55
	Chr7	77.8	16.7	5.6	0.0	18
	Chr16	93.8	3.1	3.1	0.0	32
	Chr19	89.1	10.9	0.0	0.0	46
	ChrX	85.7	7.1	7.1	0	14
	SureSelect					
	Chr1	93.3	0	0	6.7	30
	Chr7	93.3	6.7	0.0	0.0	15
	Chr16	81.0	19.0	0.0	0.0	21
	Chr19	95.3	0.0	4.7	0.0	43
	ChrX	92.9	0	7.1	0	14
Indels	**AmpliSeq**					
	Chr1	29.8	31.5	37.1	1.6	124
	Chr7	17.9	47.4	34.6	0.0	78
	Chr16	29.4	35.3	35.3	0.0	51
	Chr19	32.5	39.0	26.0	2.6	77
	ChrX	32.6	37.2	30.2	0	43
	SureSelect					
	Chr1	14.2	33	46.2	6.6	106
	Chr7	13.8	36.9	49.2	0.0	65
	Chr16	15.2	43.5	41.3	0.0	46
	Chr19	20.9	31.3	43.3	0.0	67
	ChrX	15.2	24.2	57.6	3	33

[Table reproduced from Gampawar et al., Frontiers in Genetics 2019[2]].

Supplementary table 18:Distribution of false positives from chromosome 1.

Categories	Strand bias	Read End	Low Quality	Homopolymer	Mixed alleles	Unknown	Total
AmpliSeq							
Strand bias	12						
Read End	3	2					
Low Quality	4	1	7				
Homopolymer	3	1	2	25			
Mixed alleles	3	0	0	6	12		
Unknown	0	0	0	0	0	14	
Total	25 (26.3%)	4 (4.2%)	9 (9.5%)	31 (32.6%)	12 (12.6%)	14 (14.7%)	95
SureSelect							
Strand bias	6						
Read End	0	0					
Low Quality	5	0	2				
Homopolymer	2	0	0	24			
Mixed alleles	0	0	0	2	11		
Unknown	0	0	0	0	0	22	
Total	13 (17.6%)	0 (0%)	2 (2.7%)	26 (35.1%)	11 (14.9%)	22 (29.7%)	74

In AmpliSeq 1 FP was due to strand bias+homopolymer+read end, 1 FP =strand bias+read end+homopolymer and 1 FP =strand bias+read end+low quality. Total FPs – 98 [Table reproduced from Gampawar et al., Frontiers in Genetics 2019[2]].

Supplementary table 19:Distribution of false positives from chromosome 7.

Categories	Strand bias	Read End	Low Quality	Homopolymer	Mixed alleles	Unknown	Total
AmpliSeq							
Strand bias	6						
Read End	3	2					
Low Quality	1	0	2				
Homopolymer	3	0	0	15			
Mixed alleles	0	0	0	0	1		
Unknown	0	0	0	0	0	13	
Total	13(28.3%)	2 (4.3%)	2 (4.3%)	15 (32.6%)	1 (2.2%)	13 (28.3%)	46
SureSelect							
Strand bias	0						
Read End	0	0					
Low Quality	3	0	4				
Homopolymer	0	0	2	8			
Mixed alleles	0	0	0	1	4		
Unknown	0	0	0	0	0	15	
Total	3 (8.1%)	0 (0%)	6 (16.2%)	9 (24.3%)	4 (10.8%)	15 (40.5%)	37

In AmpliSeq 1 FP = strand bias+low quality+homopolymer. Total FPs – 47 [Table reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]].

Supplementary table 20:Distribution of false positives from chromosome 16.

Categories	Strand bias	Read End	Low Quality	Homopolymer	Mixed alleles	Unknown	Total
AmpliSeq							
Strand bias	8						
Read End	1	3					
Low Quality	0	1	0				
Homopolymer	1	2	2	15			
Mixed alleles	0	0	0	0	1		
Unknown	0	0	0	0	0	9	
Total	10(23.3%)	6 (14%)	2(4.7%)	15 (34.9%)	1 (2.9%)	9 (20.9%)	43
SureSelect							
Strand bias	0						
Read End	0	0					
Low Quality	1	0	1				
Homopolymer	0	0	1	12			
Mixed alleles	0	0	0	1	1		
Unknown	0	0	0	0	0	12	
Total	1 (3.4%)	0 (0%)	2 (6.9%)	13(44.8%)	1 (3.4%)	12 (41.4%)	29

In AmpliSeq 1 FP = read end+low quality+homopolymer. Total FPs – 44. [Table reproduced from Gampawar et al., Frontiers in Genetics 2019[2]].

Supplementary table 21:Distribution of false positives from chromosome 19.

Categories	Strand bias	Read End	Low Quality	Homopolymer	Mixed alleles	Unknown	Total
AmpliSeq							
Strand bias	34						
Read End	3	3					
Low Quality	5	2	4				
Homopolymer	9	2	1	21			
Mixed alleles	2	0	0	1	2		
Unknown	0	0	0	0	0	24	
Total	53 (46.9%)	7 (6.2%)	5(4.4%)	22 (19.5%)	2 (1.8%)	24 (21.2%)	113
SureSelect							
Strand bias	1						
Read End	0	0					
Low Quality	1	0	0				
Homopolymer	1	0	4	30			
Mixed alleles	0	0	0	0	5		
Unknown	0	0	0	0	0	31	
Total	3 (4.4%)	0 (0%)	4 (5.5%)	30(41.1%)	5 (6.8%)	31 (42.5%)	73

In AmpliSeq 2 FP = strand bias +read end+homopolymer, 1 FP = strand bias+ homopolymer+mixed allele. Total FPs – 116 [Table reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]].

Supplementary table 22:Distribution of false positives from chromosome X.

Categories	Strand bias	Read End	Low Quality	Homopolymer	Mixed alleles	Unknown	Total
AmpliSeq							
Strand bias	7						
Read End	0	1					
Low Quality	3	0	4				
Homopolymer	2	0	0	4			
Mixed alleles	0	0	0	2	2		
Unknown	0	0	0	0	0	19	
Total	12 (27.3%)	1 (2.3%)	4 (9.1%)	6 (13.6%)	2 (4.5%)	19 (43.2%)	44
SureSelect							
Strand bias	0						
Read End	0	0					
Low Quality	0	0	1				
Homopolymer	0	0	0	5			
Mixed alleles	0	0	0	1	3		
Unknown	0	0	0	0	0	20	
Total	0 (%)	0 (%)	1 (3.3%)	6 (20%)	3 (10%)	20 (66.7%)	30

[Table reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]].

Supplementary table 23:False positive SNVs and indels in AmpliSeq and SureSelect.

		Region	Strand bias (%)	Read End (%)	Low Quality (%)	Homopolymer (%)	Mixed alleles (%)	Unknown (%)	Total (N)
SNVs	**AmpliSeq**								
	Chr1	TTR	51.1	6.7	20.0	20.0	2.2	13.5	52
		Specific	48.1	11.1	14.8	22.2	3.7	15.6	32
	ChrX	TTR	61.1	5.6	22.2	11.1	0.0	43.8	32
		Specific	50.0	10.0	20.0	20.0	0.0	50.0	20
	SureSelect								
	Chr1	TTR	42.9	0.0	7.1	42.9	7.1	39.1	46
		Specific	18.0	0.0	9.1	54.5	18.2	47.6	21
	ChrX	TTR	0.0	0.0	50.0	50.0	0.0	90.0	20
		Specific	0.0	0.0	50.0	50.0	0.0	60.0	5
Indels	**AmpliSeq**								
	Chr1	TTR	5.6	2.8	0.0	61.1	30.6	16.3	43
		Specific	4.5	0.0	0.0	54.5	40.9	18.5	27
	ChrX	TTR	14.3	0.0	0.0	57.1	28.6	41.7	12
		Specific	0.0	0.0	0.0	66.7	33.3	50.0	6
	SureSelect								
	Chr1	TTR	4.2	0.0	0.0	58.3	37.5	14.3	28
		Specific	10.0	0.0	0.0	50.0	40.0	9.1	11
	ChrX	TTR	0.0	0.0	0.0	62.5	37.5	20.0	10
		Specific	0.0	0.0	0.0	50.0	50.0	33.3	3

TTR - Total Target Region,Specific - Library specific region (AmpliSeq = 11.4 Mb, SureSelect = 14.1 Mb) SNV- Single Nucleotide Variation [Table reproduced from Gampawar et al., Frontiers in Genetics 2019[2]].

Supplementary table 24:False positive SNVs and indels in AmpliSeq and SureSelect.

		Strand bias (%)	Read End (%)	Low Quality (%)	Homopolymer (%)	Mixed alleles (%)	Unknown (%)	Total (N)
SNVs	**AmpliSeq**							
	Chr1	51.1	6.7	20	20	2.2	13.5	52
	Chr7	48.0	8.0	8.0	16.0	0.0	20.0	25
	Chr16	35.7	21.4	7.1	14.3	0.0	21.4	28
	Chr19	59.7	9.1	6.5	6.5	0.0	18.2	77
	ChrX	61.1	5.6	22.2	11.1	0	43.8	32
	SureSelect							
	Chr1	42.9	0	7.1	42.9	7.1	39.1	46
	Chr7	10.7	0.0	17.9	17.9	3.6	50.0	28
	Chr16	5.6	0.0	11.1	33.3	5.6	44.4	18
	Chr19	6.5	0.0	8.7	39.1	2.2	43.5	46
	ChrX	0	0	50	50	0	90	20
Indels	**AmpliSeq**							
	Chr1	5.6	2.8	0	61.1	30.6	16.3	43
	Chr7	5.0	0.0	0.0	50.0	5.0	40.0	20
	Chr16	19.4	0.0	0.0	47.2	5.6	27.8	36
	Chr19	0.0	0.0	0.0	78.6	7.1	14.3	14
	ChrX	14.3	0	0	57.1	28.6	41.7	12
	SureSelect							
	Chr1	4.2	0	0	58.3	37.5	14.3	28
	Chr7	0.0	0.0	11.1	44.4	33.3	11.1	9
	Chr16	0.0	0.0	0.0	63.6	0.0	36.4	11
	Chr19	0.0	0.0	0.0	44.4	14.8	40.7	27
	ChrX	0	0	0	62.5	37.5	20	10

SNV- Single Nucleotide Variation [Table reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]].

Supplementary table 25:Effect of increase in average read on per base coverage in AmpliSeq library.

	270X	**184X**	**92X**	**83X**	**44X**
<5	400,116	605,213	1,102,034	722,020	3,188,163
5 - 10	220,562	459,515	2,543,570	869,884	6,629,198
10 - 400	45,660,202	53,520,615	53,993,427	56,217,098	49,058,469
>400	11,494,834	3,239,836	618,854	108,958	19,602
TPs	46,320	45,781	44,681	45,462	42,041
FNs	2,476	3,015	4,115	3,334	6,755
FPs	1,218	2,983	1,391	1,404	1,341
Sensitivity	94.93%	93.87%	91.57%	93.17%	86.16%
PPV	97.44%	93.88%	96.98%	97.00%	96.91%

TPs- True Positives, FNs- False Negatives, FPs- False Positives, PPV-Positive Predictive Value [Table reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]].

Supplementary table 26:Relative change in the number of TPs, FNs, and FPs with each step of optimisation.

AmpliSeq			
Optimisation Steps	TPs	FNs	FPs
Regularisation	1.6% (714)	-21.0% (-714)	2.1% (176)
High confidence region	-0.7% (-340)	-7.6% (-204)	-85.8 (-7 363)
Parameter settings	-0.3% (-143)	5.8% (143)	-27.4% (-334)
Total effect	0.5% (231)	-22.8% (-775)	-89.5% (-7 521)
SureSelect			
Optimisation Steps	TPs	FNs	FPs
Regularisation	1.4% (622)	-20.4% (-622)	2.5% (275)
High confidence region	-0.7% (-298)	-5.2% (-127)	-91.6% (-10 333)
Parameter settings	0% (-3)	0.1% (3)	-6% (-6)
Total effect	0.7% (321)	-24.4% (-746)	-91.4% (-10 064)

TPs- True Positives, FNs- False Negatives, FPs- False Positives, PPV-Positive Predictive Value [Table reproduced from Gampawar et al., Frontiers in Genetics 2019 [2]].