UNIVERSITY OF LIVERPOOL

Healthy cognitive ageing: focussing on the role of repetitive DNA elements

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor of Philosophy by

Ana Illera López

March 2020

Acknowledgements

I am extremely grateful to both of my supervisors, Professor John P. Quinn and Dr Vivien J. Bubb, for their continuous support and belief; and, for being understanding and compassionate of not only my professional, but also my personal aspirations. I also extend my thanks to Dr Gerald S. Schumann at the Paul-Ehrlich-Institut (PEI); and, Dr Fahad Ali and Dr Mohammed Uddin at Mohammed Bin Rashid University (MBRU) for welcoming me into their laboratories, providing the invaluable opportunity to develop as an international scientist. Sincere thanks go to Professors Antony Payton and Neil Pendleton at the University of Manchester for giving us access to the Dyne Steele cohort, a priceless resource that has allowed the fruition of this thesis.

I would like to sincerely thank Veridiana Pessoa and Dr Maurizio Manca, who welcomed me into the lab and were extremely helpful on the exciting, nonetheless unsettling process of starting a PhD. A wholeheartedly thank you goes to Dr Abigail L. Savage for all of her support throughout some of the most difficult days during my PhD and for the transmission of her invaluable expertise and knowledge, but more importantly for being a great friend and an awesome human being. I would like to thank Ben Middlehurst, Ashley Hall, Emma Price, Jack Marshall and Olympia Gianfrancesco, my colleagues in White Block for always being supportive and working as a team. A special mention goes to Olivia Grech for being a brilliant MRes student and, a great source of inspiration and reassurance.

To Hazel, my best friend, who understands me without the need for words and has wiped the tears away when I could not see pass them. To Alberto Vera Muñoz, my partner, for his blind belief in me and for his continuous love and care. To María Esther López Parrilla and Rubén Illera López, my mother and brother, who have been an enormous support, not only during my PhD, but for as long as I can remember. I would most definitely not be standing strong today if it were not for you four.

Finally, I am enormously thankful to the Wellcome Trust for funding my research.

Abbreviations

A

AD Alzheimer's disease

ALS Amyotrophic lateral sclerosis

B

BED Browser extensible data

BLAST Basic local alignment search tool

BLAT BLAST-like alignment too

bp Base pair

BR Broad range

C

CAA Cerebral amyloid angiopathy

cDNA Complementary DNA

CGI CpG island

CNS Central nervous system

CNV Copy number variation

CpG CG dinucleotide

D

DLB Dementia with Lewy bodies

DNA Deoxyribonuleic acid

DNAJC5 DnaJ heat shock protein family (Hsp40) member C5

DNMT DNA methyltransferases

dNTP Deoxynucleotide triphosphate

dsDNA Double stranded DNA

E

EDTA Ethylenediamine tetraacetic acid

G

GxE Gene x environment

GTEx Genotype-tissue expression project

GWAS Genome wide association studies

H

HA Healthy ageing

HERV Human endogenous retrovirus

Hg Human genome

HS High sensitivity

HSP40 Heat shock protein 40 family

HSP70 Heat shock protein 70 family

K

KAP1 KRAB-associated protein 1

kb Kilobase

KEGG Kyoto encyclopaedia of genes and genomes

KRAB-ZFPs Krüppel-associated box domain zinc finger proteins

L

LARII Luciferase assay reagent II

LB Luria broth

LD Linkage disequilibrium

LINE-1 Long interspersed element class 1

LINE-2 Long interspersed element class 2

LOAD Late onset Alzheimer's disease

LTR Long terminal repeat

M

$MgCl_2$ Magnesium chloride

MIR941 MicroRNA-941

miRNA MicroRNA

mRNA Messenger RNA

Myrs Million years

N

NCL Neuronal ceroid lipofuscinosis

ncRNA Non-coding RNA

NPCs Neuronal precursor cells

NSCs Neural stem cells

O

ORF Open reading frame

P

PBS Phosphate buffered saline

PCR Polymerase chain reaction

PD Parkinson's disease

PLB Passive lysis buffer

PRS Polygenic risk score

R

RC Retrotransposition competent

RE Restriction enzyme

REs Repetitive elements

RIP Retrotransposon insertion polymorphism

RNA Pol II RNA Polymerase II

RNA Ribonucleic acid

Rpm Revolutions per minute

S

SINE Short interspersed element

SNP Single nucleotide polymorphism

SV40 Simian virus 40

SVA SINE-VNTR-*Alu*

SVD Small vessel disease

T

TAE Tris-acetate-EDTA

TBE Tris-borate-EDTA

TDP43 TAR DNA-binding protein 43

TE Transposable element

TPRT Target primed reverse transcription

TR Tandem repeat

TSD Target-site duplication

TSS Transcriptional start site

U

UCSC University of California, Santa Cruz

UTR Untranslated region

UV Ultraviolet

V

VNTR Variable number tandem repeat

X

XCI X-chromosome inactivation

Abstract

The maintenance of healthy cognitive ageing is a complex process that can be modulated through the interaction between genes and environment in a GxE mechanism. The modulation of this complex process is highly regulated by the action of non-coding regulatory elements.

The epigenetic landscape of non-coding regulatory elements is the focus of the first half of this thesis. As the understanding of epigenetic mechanisms such as DNA methylation increases, it becomes clearer that the comprehensive study of epigenetic modifications of the non-coding repetitive regions of the genome is required to expand our knowledge on their regulation. The majority of new long interspersed element class 1 (LINE-1) insertions are the offspring of a set of specific LINE-1s named hot retrotransposition competent LINE-1s (hot RC-L1s) as they give rise to the majority of insertions. In order to prompt insertions, hot RC-L1s need to be full length and actively transcribed. Therefore, the research in this part of the thesis assessed the methylation status of LINE-1s globally, and of hot RC-L1s specifically in order to evaluate their regulatory potential. Further, variable number tandem repeats (VNTRs) are non-coding regulatory elements. The focus was on a VNTR located at the chromosome 20q13.3 locus, which is polymorphic in copy number across the population. A *microRNA* termed *MIR941*, which is dysregulated in ageing, resides within it. *MIR941* plays a role in the regulation of its host gene *DNAJC5*, which is demonstrated to have neuroprotective properties. Therefore, the work in this section of the thesis addressed both the association of the polymorphic nature of the VNTR with health status in the elderly and the methylation status of this element.

The identification of non-reference genome retrotransposon insertion polymorphisms (RIPs) is the focus of the second half of this thesis. As well as playing regulatory roles, retrotransposons, which are repetitive elements, contribute towards genome plasticity and diversity. In particular, LINE-1s have been suggested to have a role in adult neurogenesis and may contribute positively towards memory. With this in mind, the research on this half of the thesis is identifying non-reference genome *Alu*, LINE-1 and SVA insertions in healthy aged (HA) compared to Alzheimer's disease (AD) people using sequencing data generated by two different protocols and bioinformatic analysis. These elements can also be polymorphic, and thus, the work in this part of the thesis studied not only the number, but also the presence/absence of these elements and the potential association with protective effects in healthy cognitive function in the elderly. This work identified that the number of *Alu*, LINE-1 and SVA RIPs does not vary drastically between HA and AD.

The research in this thesis suggests that there is no association between the number of presence/absence LINE-1 RIPs and health status. Our data rather suggests a role for methylation of the active LINE-1s in ageing. The aim of this work is primarily to extend our understanding of the role of non-coding repetitive elements and the epigenetic mechanisms regulating these elements during ageing. Despite the limitation on the small number of samples due to the difficulty of acquiring matched tissue, the data presented here is of high importance to understand the landscape of TEs because of the extensive analysis carried out in each of the individuals studied.

Thesis layout overview

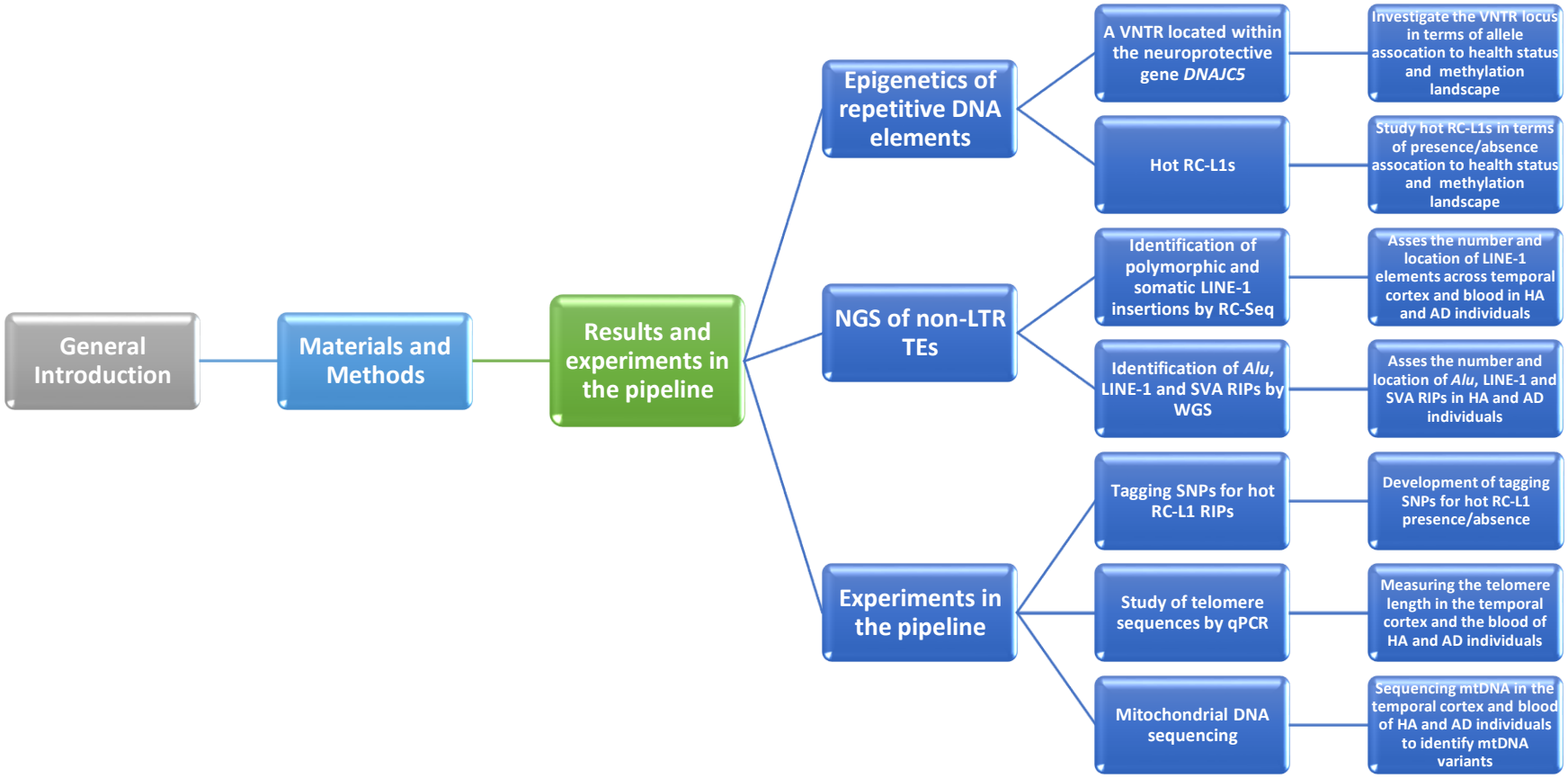


Table of contents

Chapter 1

Introduction

Chapter 1 Introduction

1.1. General introduction

It has been demonstrated that there is a genetic risk for age-related neurodegenerative conditions [1]. What is less clear is if there is a genetic component for healthy cognitive ageing in humans [2]. Maintaining cognitive functioning is more strongly associated with wellbeing in the oldest individuals than presence or absence of disease [3], as good cognition makes an essential contribution to the quality of life rather than simply living for a long time. Unravelling the genetics of heritable and polygenic traits such as cognitive function during ageing has been consistently proven difficult. It was not until genome-wide association studies (GWAS) that the first genetic associations with cognition and ageing were established [4]. The large sample size and the ability of GWAS to look at the small aggregative effect of several genetic variants towards a single phenotype [polygenic risk score (PRS)] allowed establishing such associations. The influence of the environment on the target genetic variants (GxE) in the brain and during ageing has also contributed towards gaining a deeper understanding of healthy cognitive ageing, which is highly impacted by lifestyle choices [3]. These environmental signals function in the majority of cases via non-coding DNA (ncDNA) to alter gene expression or genome structure by acting as transcriptional regulators [5-11]. Therefore, the emerging role of ncDNA elements such as non-long terminal repeat (non-LTR) retrotransposons in ageing is of growing importance as they can both alter transcription/post-transcription [7, 12-15] and modify genome structure [10, 16-18].

In this thesis, we discuss the genetic component of healthy ageing in relation to good cognitive function. In particular, we consider the potential role of non-coding elements such as non-LTR retrotransposons and their intricate relationship as putative regulatory mechanisms altering transcriptional, post-transcriptional and epigenetic parameters with the functioning of the central nervous system (CNS).

1.2. Healthy cognitive ageing: is there such a thing?

There is a degree of cognitive decline associated with ageing (Figure 1.1), and this varies widely between individuals [3, 19]. Data from 14 longitudinal population-based studies of cognitive ageing covering 12 countries and 5 continents [a total of 42,170 individuals aged 54-105 (42% male)] demonstrated that cognitive function in late-life decreases with age and at more rapid rates with increasing age [20]. The low incidence of diseases across healthy long-lived individuals suggests that they either delay or escape most age-related burdens [21]. Furthermore, the complexity of long healthy living implies that either disease reduction or absence are not sufficient, and that maintaining cognitive function is crucial. In addition, the environment and lifestyle (diet, physical and social activity) play a major role to differentiate healthy cognitive ageing from longevity alone [3]. Previous studies in model organisms have shown that manipulation of specific protein coding genes can extend longevity in such species, but these findings do not always translate to human longevity [22]. For instance, when looking at GenAge (https://genomics.senescence.info/genes/), a database of age-related genes, >1000 genes have been associated with ageing and/or longevity in model organisms, >100 of which are in mice, with 51 of these showing

life-extending effects. However, this association could not be established in Deelen *et al.* one of the largest GWAS for human longevity [23].

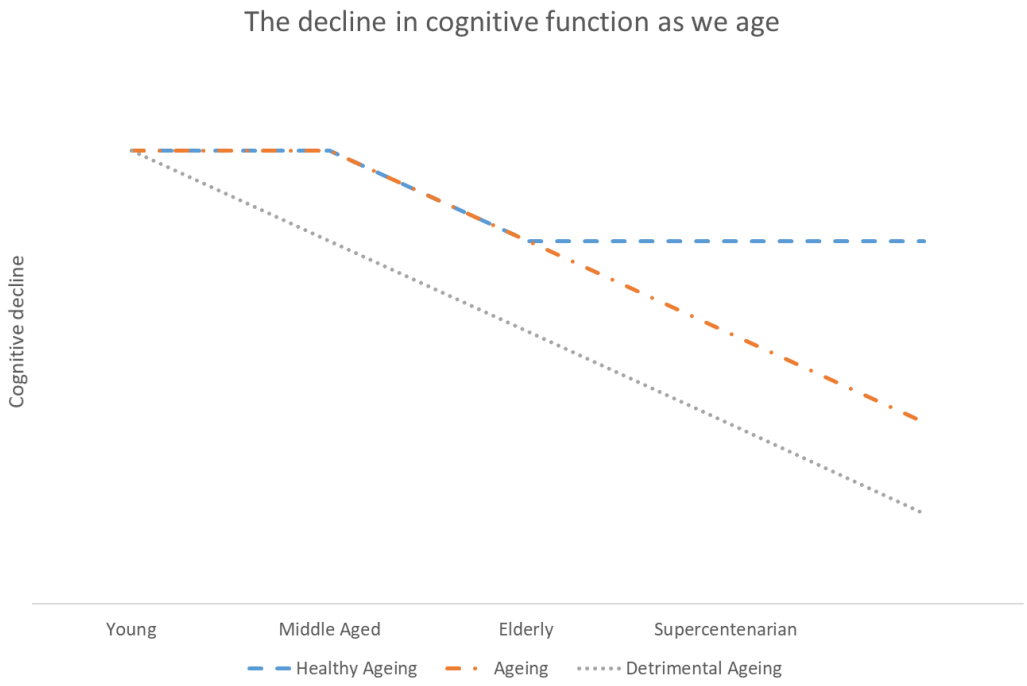


Fig. 1.1. An schematic view of the decline in cognitive function as we age. The maintenance of good cognitive function is essential for healthy ageing. During normal ageing (orange), there will inevitably be a progressive degree of cognitive decline, particularly from midlife onwards. The cognitive decline experienced in detrimental ageing (grey) is a lot more accentuated. However, it is hypothesised that during the course of healthy ageing (blue), the degree of cognitive decline is minimum throughout.

1.2.1 An example of pathological cognitive ageing: Alzheimer's disease

Alzheimer's disease (AD) is the foremost trigger of dementia, which is the leading cause of death in England and Wales with an estimated 44 million people living with dementia worldwide [24]. As the major risk for developing AD is age [19], this number is predicted to increase over three times by 2050 as a result of an increasing ageing population [24].

On the one hand, most AD cases are sporadic and driven by a complex interaction between genetic and environmental factors (GxE) [19]. GWAS studies have found the *APOE* gene to be a major risk variant for sporadic AD [25]. As with healthy cognitive ageing, physical exercise and education are considered protective factors against AD, whereas hypertension and diabetes increase risk of disease onset [19, 24]. On the other hand, about 0.5% of AD is familial and caused by mutations in amyloid beta precursor protein (*APP*), presenilin 1 (*PSEN1*) and presenilin 2 (*PSEN2*) (Figure 1.2) [24]. Early clinical features in the elderly often include progressive problems on episodic memory, difficulties with multi-tasking and loss of confidence up until the person becomes reliant on others. In general, there is a life expectancy of 8.5 years from disease onset [24].

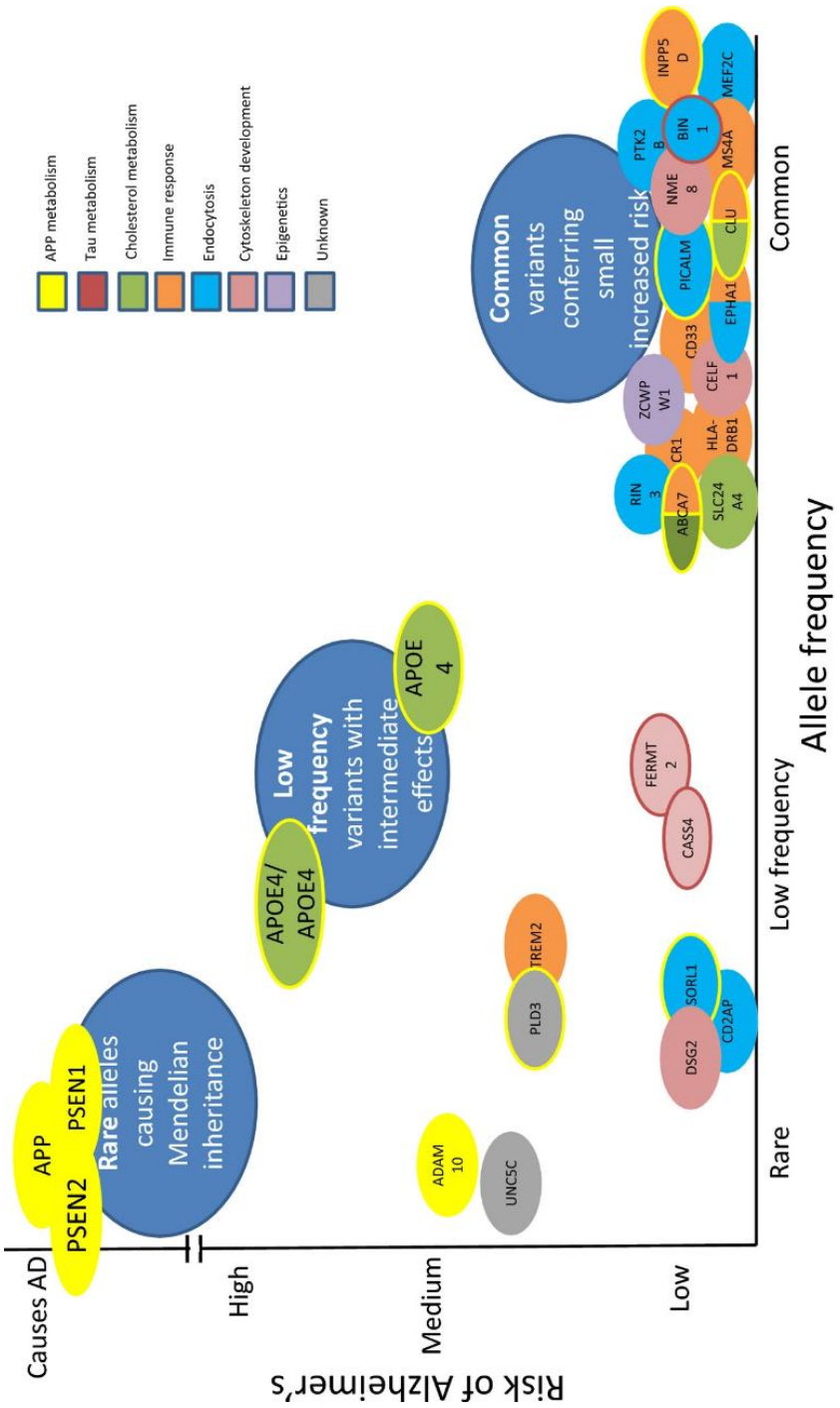


Fig. 1.2. An overview of genes implicated in Alzheimer's disease to date [24]. The function of the gene is determined by the internal colour, and when there are two internal colours, the gene has been implicated in two pathways. Yellow-circled genes are thought to influence amyloid beta precursor protein (*APP*) metabolism. Red-circled genes are thought to influence tau metabolism. Taken from Lane, C.A., J.

Hardy, and J.M. Schott, *Alzheimer's disease.* European Journal of Neurology, 2018. **25**(1): p. 59-70.

The defining lines between AD and normal ageing are ambiguous, as three major symptoms of AD namely episodic memory function disruption, brain atrophy and formation of amyloid plaques are also found in the healthy elderly [3, 19, 26]. However, the organised pattern of brain atrophy in AD can be a distinctive factor between healthy ageing (HA) and AD [27]. In normal ageing, specific cognitive abilities such as executive function and episodic memory often suffer a reduction, whereas verbal abilities and world knowledge remain throughout [19]; however, there is a more global loss of function in cognitive and verbal abilities for AD dementia [19, 24]. In fact, the age-specific changes in cognitive function are measured in terms of fluid intelligence and vocabulary ability; and, are often associated to changes in brain structure [19]. Such changes may include reduction in gross brain volume commonly observed in the cerebral cortex, especially in the frontal and temporal lobes, and in the hippocampus (Figure 1.3) [19]. In high performing elderly, cross-sectional studies do not show reproducibly such a reduction [28].

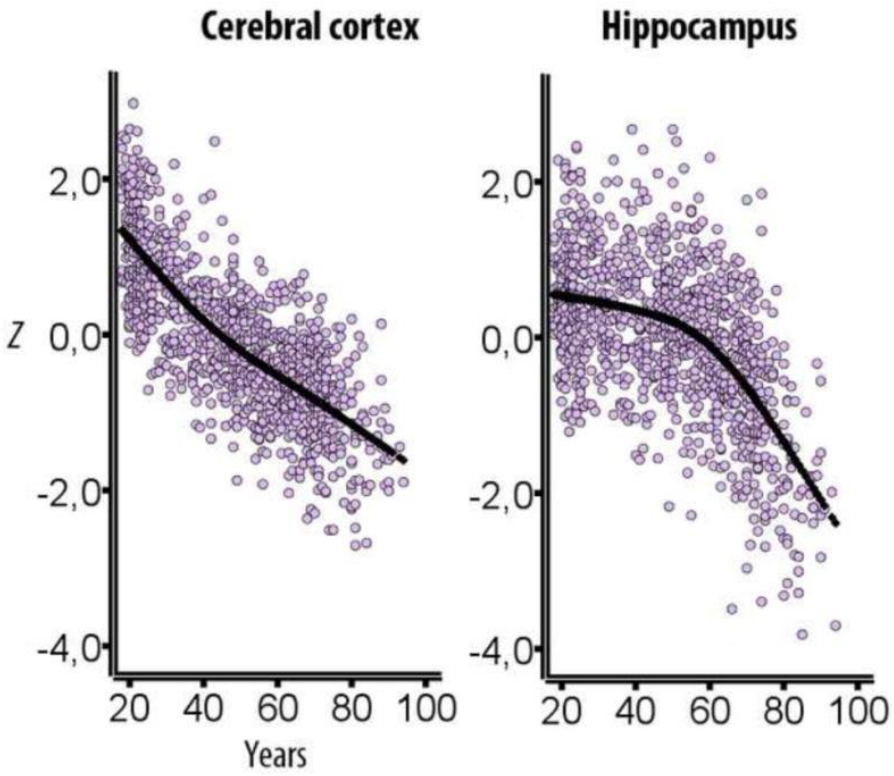


Fig. 1.3. Cross-sectional estimates of adult life-span trajectories of total cerebral cortex volume and total hippocampal volume [19]. Volume is expressed in units of standard deviations. Taken from Fjell, A.M., *et al.*, *What is normal in normal aging? Effects of aging, amyloid and Alzheimer's disease on the cerebral cortex and the hippocampus.* Progress in neurobiology, 2014. 117: p. 20-40.

1.2.2 Regions of the brain important for cognitive function during ageing

The brain is topographically organised into distinct regions based on their biological properties; however, the function of each individual region is yet to be elucidated [29]. Despite the functional heterogeneity associated to each biologically different brain region, neuroimaging provides a comprehensive description of functional data in relation to brain regions [30]. The complex nature of the brain makes it difficult to be definitive about regions of the brain that are important in cognitive function. However, in this section we have attempted to briefly elucidate the role in cognitive function of the cerebral cortex and the hippocampus known to undergo structural changes during ageing [19]:

- The cerebral cortex is the largest and most recently evolved region of the human brain [31]. It is divided into four areas:
    - The frontal lobe lies beneath our forehead and has a wide variety of attributed functions such as our ability to reason, to organise ideas, to plan and to inhibit inappropriate behaviour [31]. As it is connected via nerve fibres to the amygdala and thalamus, it is also part of the emotional brain [29, 31].
    - The parietal lobe is located in the upper rear of the human brain and is responsible for our senses and integrating sensory information [31].
    - The temporal lobe is located near the ears and it is the biggest contributor to long-term memory and language [31].
    - The occipital lobe is located at the back end of the brain and controls visual functions. Visual memory is located in the nearby area to the temporal lobe [31].

- The hippocampus is a major structure of the limbic system. It is located at the base of the temporal lobe, and exchanges signals with the entire cerebral cortex via nerve fibres that connect the two regions together. The hippocampus is essential to short-term memory [32, 33].

Amnesic and AD individuals often present damage to the temporal and frontal lobes [34]. Degeneration of the hippocampus and other limbic structures is also associated with the loss in short-term memory that occurs in AD [31]. Therefore, temporal and frontal lobes as well as the hippocampus are proven essential to maintain cognitive abilities and in turn for healthy cognitive ageing. The temporal lobe has been the brain region of choice throughout the analyses carried out in this thesis. Not only it plays a major role in maintaining good cognitive function during ageing, but it was available to us matched with blood from the same individual, an invaluable resource difficult to acquire that allowed us to address CNS individual- and tissue-specific genomic differences/mutations.

1.3. Genetic variation within the human genome

The human genome is composed of 3 billion nucleotide pairs [35]. Only about 1.5 % is protein coding, whereas the other 98.5 % is non-coding DNA (ncDNA) (Figure 1.4). NcDNA is known to act as regulatory, and exhibits genetic variation, which can be associated with a specific disease risk and alter gene expression [11, 36-41]. In an attempt to explain individual genetic differences in healthy cognitive ageing, we have addressed the most common sources of genetic variation within the human genome in this section.

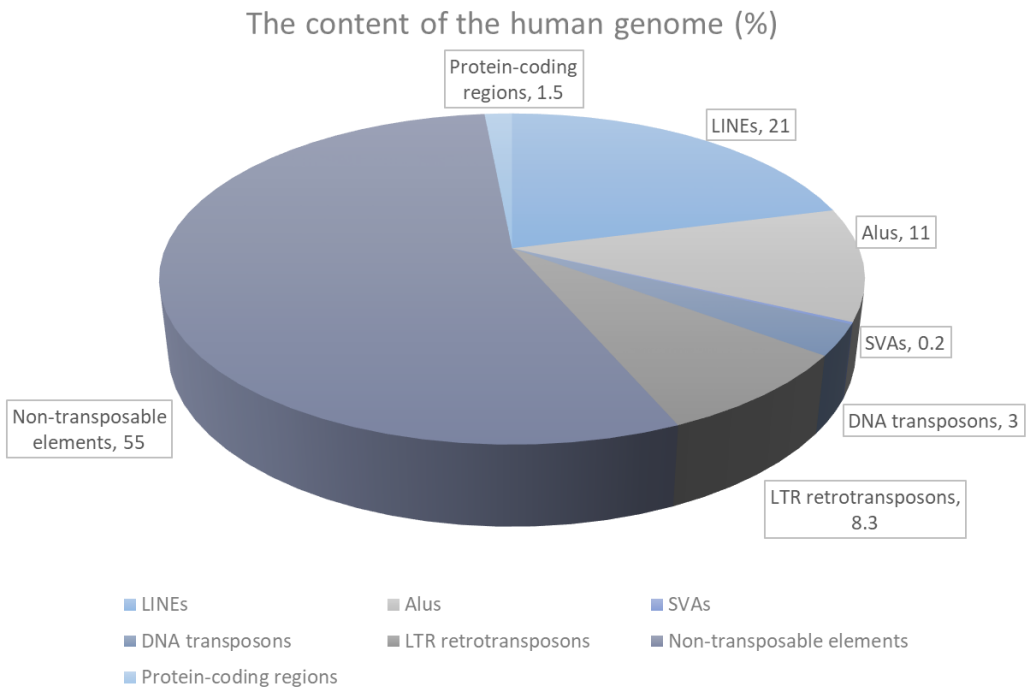


Fig. 1.4. The content of the human genome (%). About 45 % of the human genome is composed of transposable elements (TEs), the vast majority (42 %) of which are non-LTR retrotransposons such as *Alu*, LINE and SVA elements.

1.3.1 Single nucleotide polymorphisms (SNPs)

Single nucleotide polymorphisms (SNPs) are the simplest, most common and most extensively studied class of genetic variation [42]. If a non-synonymous change where a single base variation within a coding exon resulting in an amino acid change and in turn a protein change occurs, it can be mechanistically and functionally relevant. Furthermore, if occurring within ncDNA, it can alter the affinity or specificity between a transcription factor and the specific DNA sequence or prompt alternative splicing, ultimately modifying the function of the DNA element as a result. Despite SNPs being relevant both mechanistically and functionally, a SNP in isolation is rarely linked to a specific condition; and thus, the ability of addressing SNPs not in isolation but in

terms of PRS in GWAS in order to find the genetic overlap between different conditions has become crucial in a growing technological era [11].

1.3.2 Repetitive DNA

It is estimated that about two-thirds of the human genome is composed of repetitive elements (REs) [12, 43, 44]. Five categories of repetitive DNA sequences have been identified so far. These include four minor categories accounting for about 10 % of genomic DNA (simple sequence repeats, segmental duplications, processed pseudogenes, and tandem repeats or satellite DNA sequences), and a more abundant category accounting for about 45 % of the human genome named transposable elements (TEs) (Figure 1.4) [12, 18].

1.3.2.1 Tandem repeats (TRs) and satellite DNA sequences

A widely studied minor category of repetitive DNA comprises TRs and satellite sequences. Satellite DNA consists of arrangements of nucleotide TRs variable in sequence length [11]. Depending on their sequence length they are classified as micro- (1-8 bp), mini- (16-64 bp) and macro- (>64 bp) satellites or variable number tandem repeats (VNTRs). VNTRs occur when the repeat unit copy number is variable in the population [45]. In the human genome, there are 600,000 candidate VNTRs, many of which are located in intron-exon splicing junctions or promoters, functional regions where the repetitive nature of the element can provide multiple transcription binding sites [46, 47]. VNTRs have been repeatedly linked to disease predisposition and have a role in gene regulation [39, 48-50]. In fact, VNTRs regulate gene-expression in a tissue and allele specific manner both *in vitro* and *in vivo* [39, 51] and,

our group and others have previously demonstrated VNTRs act as transcriptional regulators [39, 48, 49].

1.3.2.2 Transposable elements (TEs)

The vast bulk, yet much overlooked class of repetitive DNA sequences are TEs. Interestingly, TEs are able to mobilise throughout the genome giving rise to genome diversity [18]. Scientist Barbara McClintock first discovered transposable elements in maize in the 1940s, but as the observation was originally dismissed as ambiguous, the findings were not published up until the 1950s in one of her seminal papers [52]. McClintock's work was ground-breaking, replacing the idea of a static heritable genome with the concept of a dynamic genome. While TEs were initially thought of as redundant within the genome providing no apparent contribution to its host [18], current research has proven that TEs are major contributors to genome variability, playing key roles in genome evolution and gene function [53]. TEs can be divided in two classes (Figure 1.5), DNA transposons, which comprise about 3 % of the human genome and move via a DNA intermediate using a cut and paste mechanism, and retrotransposons, which mobilise via an RNA intermediate [18]. While there is no evidence of active DNA transposons in the human genome, many retrotransposons are still active today [18]. In fact, DNA transposons domestication resulted in the catalytic domain of Recombination Activating Gene protein 1 (RAG1) [54], Recombination Activating Gene protein 2 (RAG2) [55] and possibly the recombination signal sequences (RSS) [54], which are the central components of V(D)J recombination [56]. V(D)J recombination is one of the main processes to generate the wide range of antibodies necessary for the recognition of an infinity array of

antigens and hence, it is crucial to adaptive immunity [56, 57]. For the purpose of this thesis, our focus is on the active non-LTR subclass of retrotransposons, further explained below.

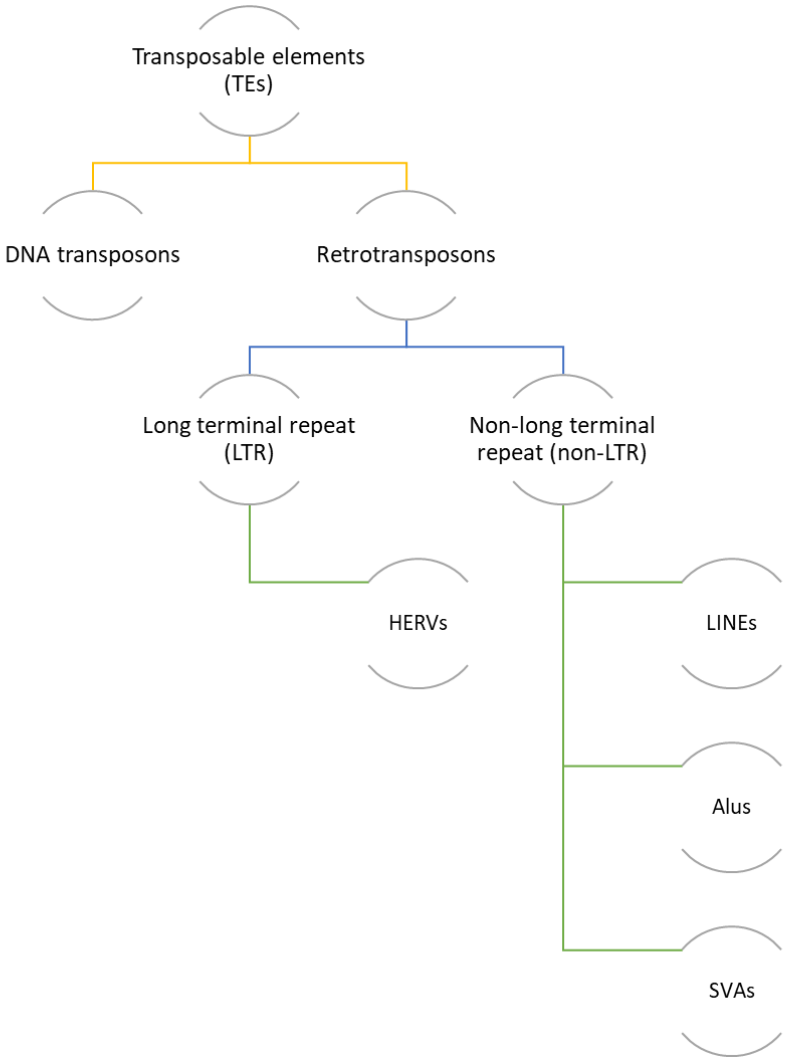


Fig. 1.5. The itemisation of human transposable elements by class. Human transposable elements (TEs) are divided into two main subclasses named DNA transposons and retrotransposons. DNA transposons comprise about 3% of the human genome, whereas retrotransposons represent about 42% of the human genome. Retrotransposons are further classified into long terminal repeats also known as HERVs, and non-long terminal repeats such as *Alu*, LINEs and SVAs.

1.3.2.2.1 Retrotransposons

Retrotransposons are the most abundant class of TEs and can be further divided into those that contain long terminal repeats (LTRs) and those that do not (non-LTRs) [58].

LTR retrotransposons

LTR retrotransposons comprise about 8 % of the human genome and receive their name because of the long terminal repeats that flank their sequences. [53]. The main class of LTR retrotransposon termed human endogenous retrovirus (HERVs) arose from the repeated infection of germ cells by exogenous retroviruses [59]. Subsequently, mutations that accumulated in their proviral DNA rendered them unable to re-infect [60]. Therefore, endogenous retrovirus proviral DNA expression does not result in infectious particles [59, 60]. A full-length HERV element is about 9 kb in length comprising ancient retroviral sequences and two LTRs flanking three open reading frames (ORFs) coding for *gag*, *pro*, *pol* and *env* viral proteins [18]. Despite the hypothesis that HERVs have been rendered inactive due to mutations in their proviral DNA sequence, HERVs polymorphic for their presence/absence in the human genome have been recently identified, suggesting putative mobilisation capabilities of HERVs in the past 6 Myrs in humans [61]. Neurological diseases such as multiple sclerosis, schizophrenia and amyotrophic lateral sclerosis (ALS) have been associated with HERVs [62, 63], though there have been other studies that suggest further research is required to establish the HERVs/ALS correlation [64].

Non-LTR retrotransposons

Non-LTR retrotransposons are the only class of TEs known to retain their ability to mobilise [58]; and, thus they continue to have a major impact on the evolution of

human genome, both structurally and functionally [6, 58, 65]. This class of TEs mobilise via a copy and paste mechanism named target-primed reverse transcription (TPRT) [66] through which an RNA intermediate is reverse transcribed, and the cDNA copy of the element is inserted at a new genomic locus of the host genome [58]. The integration of active TEs results in target site duplication (TSD) sites, which are short sequences of genomic DNA carried over upon insertion and vary in size depending on the element (Figure 1.6) [18]. Non-LTR retrotransposons account for about one third of the human genome and comprise the following active classes: (1) short interspersed elements (SINEs), (2) long interspersed elements class 1 (LINE-1), and (3) a class of composite element consisting of a SINE, a VNTR and *Alu*-like element, known as SVAs.

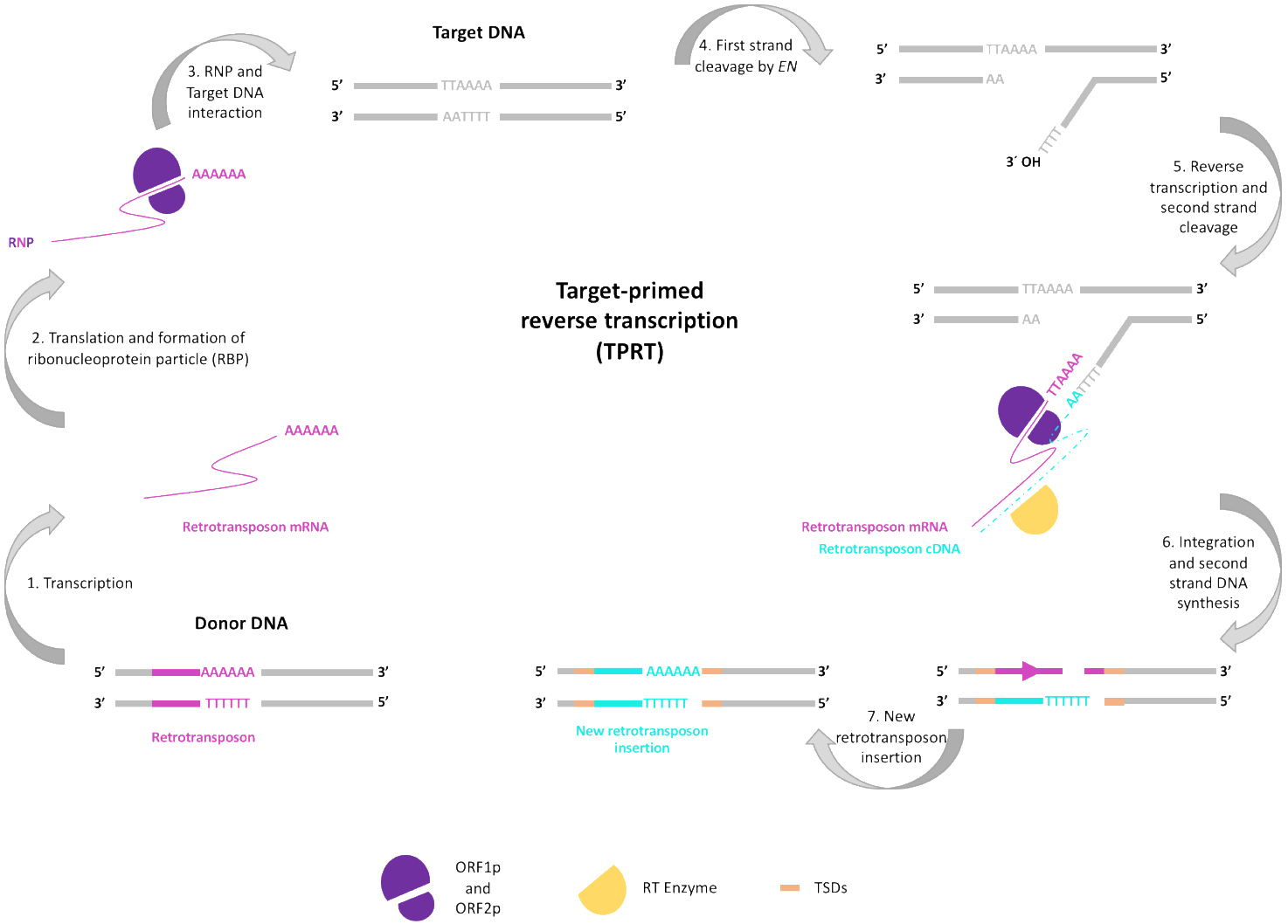

Target-primed reverse transcription (TPRT)
Target DNA
5′ TTAAAA 3′
3′ AATTTT 5′
3. RNP and Target DNA interaction
AAAAAA
RNP
2. Translation and formation of ribonucleoprotein particle (RBP)
AAAAAA
Retrotransposon mRNA
1. Transcription
Donor DNA
5′ AAAAAA 3′
3′ TTTTTT 5′
Retrotransposon
4. First strand cleavage by EN
5′ TTAAAA 3′
3′ AA 5′
3′ OH
TTTT
5. Reverse transcription and second strand cleavage
5′ TTAAAA 3′
3′ AA 5′
TTAAAA
AATTTT
Retrotransposon mRNA
Retrotransposon cDNA
6. Integration and second strand DNA synthesis
5′ AAAAAA 3′
3′ TTTTTT 5′
New retrotransposon insertion
7. New retrotransposon insertion
5′ 3′
3′ TTTTTT 5′
ORF1p and ORF2p
RT Enzyme
TSDs

Fig 1.6. The mechanism of target-primed reverse transcription (TPRT). 1. A retrotransposon, consisting of 5'- and 3'-UTR regions and two open reading frames [ORF1 and ORF2 (encoding a reverse transcriptase (RT) and an endonuclease (EN)], is transcribed by an RNA polymerase. The mRNA transcript of element produced is translocated to the cytoplasm. **2.** Upon translation, ORF1 and ORF2 exhibit cis-preference and bind their encoding mRNA to form a ribonucleoprotein particle (RNP). **3.** The RNP translocates into the nucleus to act on the target DNA. **4.** The EN enzyme makes a single-strand cleavage to the target DNA at a loosely defined consensus site (5'-TTTT/A-3'). **5.** The free 3'-OH generated is used by the RT enzyme to prime first-strand cDNA synthesis of the element at the site of the nick. The opposite DNA strand of the target DNA is also cleaved, but the exact mechanism is unknown. **6**. The gap in the target DNA is repaired, by either a host DNA polymerase or the RT enzyme and ligase. **7.** The end result of TPRT is the integration of an element at a new genomic location, which in most cases is flanked by variable length target-site duplications (TSDs).

Alu elements are active and mobile members of the SINE family in humans. Being the most abundant family of repetitive elements, *Alu* make up about 11% of the human genome [58]. *Alu* elements arose 65 million years ago by fusion of the 5' and 3' end of the 7SL RNA gene giving rise to the first *Alu* monomers (FAMs), which were about 160 bp in length [41]. The fusion of two distinct FAMs resulted in modern *Alu* elements, which are about 300 bp in length [41, 58] and have a dimeric structure composed of two monomers separated by an A-rich linker region (Figure 1.7A). The 5' region contains an RNA polymerase III promoter (A and B boxes) and the 3' region

a poly-A tail. Modern *Alu* elements are classified into subfamilies, which share common nucleotide substitutions [67], according to their evolutionary ages [41]. *Alu*-Y is the youngest *Alu* subfamily present in humans and primates [67]. *Alu* elements have no coding activity and rely on L1 retrotransposition machinery in order to mobilize [5, 68-71]. Therefore, *Alu* are non-autonomous TEs [58].

LINE-1s make up about 17 % of the total LINE (21 %) content of the human genome comprising over 500,000 copies [44, 72]. L1s are the only autonomous non-LTR TE remaining mobile in humans [16, 18]. However, due to mutations, truncations and internal rearrangements, only about 80-100 L1 copies remain intact and retrotransposition competent (RC) in an average human genome [58, 72, 73]. A functional, full length LINE-1 (Figure 1.7B) is about 6 kb in length [16]. Simplistically, it consists of a 5' untranslated region (5' UTR) which encloses a sense and antisense RNA polymerase II promoter, two open reading frames (ORF1 and ORF2) and a 3' untranslated region (3' UTR) followed by a polyadenylation signal and a poly-A tail [16, 62, 74]. ORF1 and ORF2 encode for a protein with RNA-binding capabilities and nucleic acid chaperone activity, and endonuclease (EN) and reverse transcriptase (RT) activities, respectively, conferring LINE-1 elements with the unique ability to autonomously retrotranspose in the human genome [58, 62].

SVAs are the most recently evolved hominid-specific class of TEs [58] and represent only 0.13 % of the human genome comprising about 3,000 elements [14, 37]. A full-length SVA (Figure 1.7C) is a composite unit of about 2 kb in length including a hexamer repeat region $(CCCTCT)_n$, an *Alu*-like region, a VNTR region, a SINE region and a polyadenylation signal followed by a poly-A tail. According to their evolutionary

age, SVA elements are divided into seven subfamilies named A – F1, with SVA A being the oldest (13.6 Myrs) subfamily, B and D being the most abundant (15 and 40 %, respectively) subfamilies and SVA E and F (3.5 and 3.2 Myrs, respectively) and F1 being the youngest subfamilies [14, 37]. Along with SVA D (9.6 Myrs), the youngest SVA subfamilies, which are human specific, are polymorphic within the human population [37, 44]. While SVA elements are active in the human genome, these, like *Alu*, are non-autonomous TEs and rely on L1 retrotransposition molecular machinery for mobilisation [5, 58, 68-70, 75, 76].

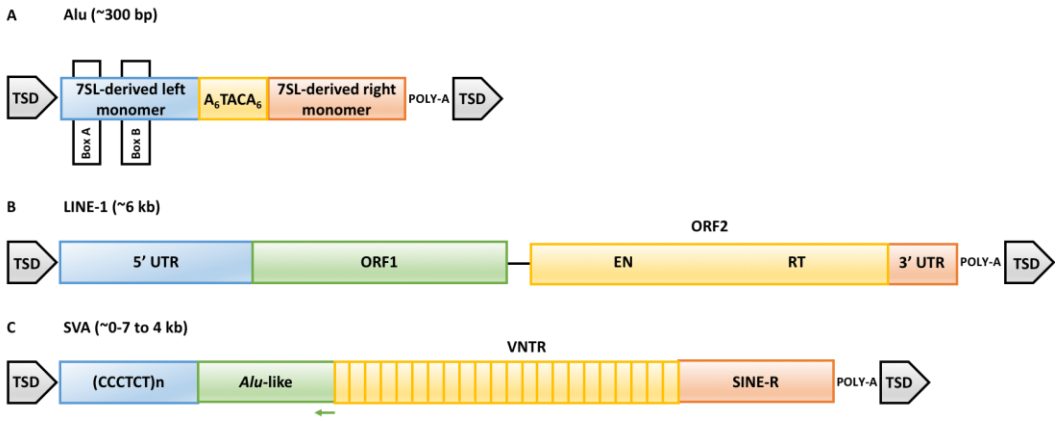


Fig. 1.7. The structure of non-LTR retrotransposons. A. A full-length *Alu* element, which consists of two monomers linked together by an A-rich region. **B.** A full-length LINE-1 element, which comprises two open reading frames (ORF1 and ORF2) flanked either side by 5' and 3' untranslated regions (UTRs). **C.** A full length SVA element, which has a composite structure that starts with a hexamer repeat region, followed by an *Alu*-like element, a VNTR region and a SINE region. All three non-LTR elements are often found framed by target site duplications either end of the element generated upon insertion.

1.3.2.2.1.1 Retrotransposon insertion polymorphisms (RIPs)

Non-LTR elements have been shown to play an important role in evolution and continue to shape our genomes today [68], especially those that are known to be

polymorphic as their burden across/within an individual varies and so do the genes and pathways affected by them [77]. In this sense, non-LTR retrotransposons can be polymorphic in two ways (A) presence/absence in the genome, named retrotransposons insertion polymorphisms (RIPs) (B) polymorphic within each retrotransposon itself (Figure 1.8) [5, 58, 65, 77-84]. In fact, the first publication of the assembled human genome reference sequence in 2001 [44] revealed a genome-wide view of the TE content in humans. However, non-LTR elements are known to be actively retrotransposing at rates of around 2 to 5 new insertions per 10-100 live births [85, 86] and hence, the human genome consists of a combination of vastly overlooked transposable element insertions that are not present in the reference genome in addition to reference insertions [76, 85]. Therefore, presence/absence polymorphisms (RIPs) are a major source of genetic variation between individuals, and it is estimated that within the human population there are 392 million private insertions unique to those harbouring them [68]. Recent development in sequencing technologies and software tools for mobile element insertion detection has allowed detection of *de novo* retrotransposition events [85, 87, 88]. Some of these sequencing technologies include retrotransposon capture sequencing (RC-Seq) [88], long interspersed element sequencing (L1-Seq) [89] and amplification typing of L1 active subfamilies (ATLAS) sequencing [90]. Further, software tools commonly used comprise TEBreak [85], mobile element location tool (MELT) [87], transposable element analyzer (Tea) [85], McClintock [91] and transposon insertion finder (TIF) [92]. *De novo* retrotransposition events are estimated to occur in roughly 1 of 20 live births for *Alu* elements, 1 of 150 for LINE-1 and 1 of 1000 for SVA elements [84, 93, 94]. However, and supporting the idea of the constantly changing retrotransposon

component, a recent paper suggested these numbers to half to 1 in 40 live births for *Alu* elements, and to increase to 1 in 63 for LINE-1 and 1 in 63 for SVA elements [86]. A review on the detection of TEs from WGS data collates those RIPs that have already been identified [85] and there is also RIP information in databases such as the European database of human specific LINE-1 RIPs (euL1db.unice.fr). However, because of the polymorphic nature of RIPs, as further genomes are analysed, we expect the number of *Alu*, L1 and SVA RIPs to keep varying [62]. Many disease-causing events such as in hereditary cancer, haemophilia, X-linked dystonia parkinsonism and neurofibromatosis type 1 [95], are rare insertions, but there are also common RIPs that affect the function of host genes [96]. As mentioned, both reference non-LTR retrotransposons, but more so RIPs have the potential to alter the function or expression of host genes and as a result being a predisposing factor for certain genetic conditions. To explore this hypothesis, RIPs need to be investigated in distinct matched populations to determine if there are insertions that occur more frequently in one compared to the other. In addition, the functional consequences of any insertions identified that are associated with a specific genetic risk also need to be addressed. As an example of a successful approach analysing the role and impact of RIPs on complex disorders, Payer *et al.* in two of their recent seminal studies suggest that *Alu* RIPs are correlated with trait-associated SNPs and alter mRNA splicing mainly promoting exon skipping, where the variant can be implicated in disease risk [97, 98].

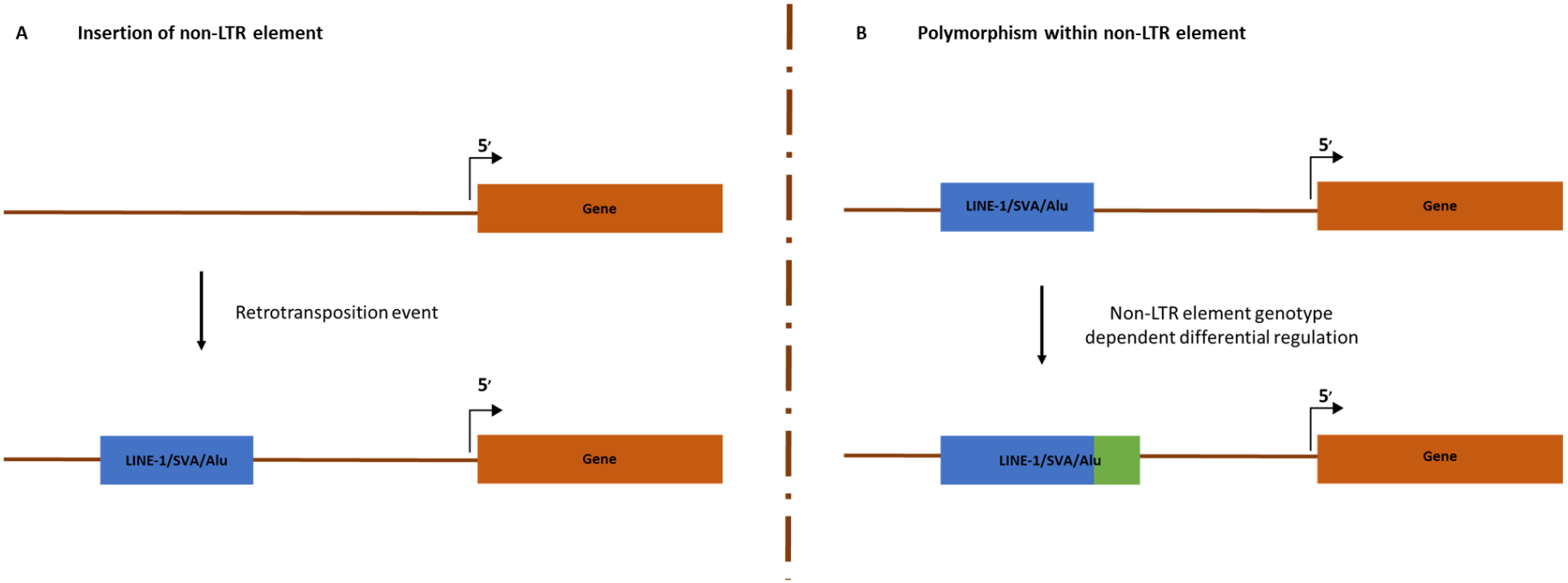


Fig. 1.8. Schematic representation of the two major non-LTR elements polymorphisms. A. Presence/absence of non-LTR element. Insertion of a non-LTR element causes disruption of normal gene expression after a retrotransposition event. These retrotransposition events are the drivers of genomic diversity and somatic mutation **B.** A polymorphism within the non-LTR element differentially alters gene expression. Note that non-LTR elements can also be polymorphic within genes.

1.4. The impact of genetic variation on the human genome

The first studies addressing the impact of genetic variation on the human genome focused on SNPs in exons encoding proteins [42]. In fact, the significance of SNPs was addressed when these resulted in the alteration of proteins such as non-synonymous amino acid changes or protein truncation [11]. However, with the appearance of GWAS, it was demonstrated that ncDNA also contained relevant risk SNPs [11]. In this sense, the study of the impact of SNPs in ncDNA regions of the genome has proven more difficult. For instance, in scenarios where a SNP is located in a ncDNA regulatory region, it could affect transcription factor binding efficiency to such regulatory domain [36, 40, 99, 100].

Another widely studied example of genetic variation, which has an impact on the host genome, is repetitive DNA [80]. This includes VNTRs, which can act as transcriptional regulators, such as the serotonin transporter [39], the dopamine transporter [49] and monoamine oxidase A (*MAOA*) VNTRs. In these three examples, the copy number of the repeat unit within the VNTR specifically modulates its function and as such, it is the risk factor [11, 48]. The copy number of the VNTR repeat unit has also been shown to act as a regulator of gene expression, such as a VNTR in the internal promoter of *MIR137*, which supports differential gene expression *in vitro* [11, 101].

Despite often being overlooked, another form of repetitive DNA that alters the human genome are TEs [11]. The impact of TEs on the human genome may occur via either mobilisation or post-insertion. A common example of the local impact of TEs via mobilisation is insertional mutagenesis (Figure 1.8A). The effects of non-LTR retrotransposons at a local genomic scale via insertional mutagenesis may include

disruption of protein-coding regions and the creation and repair of DNA breaks [28]. The effects of TEs post-insertion on the genome are more global and can influence genome structure, gene function and chromosome dynamics. These may include functional polymorphisms that alter gene expression or alteration of the epigenetic landscape as a consequence of TE presence [18]. In addition, non-LTR retrotransposons can also influence genome structure by generating genomic rearrangements such as deletions, duplications and inversions [102]. The occurrence of 5'- and 3' transductions, whereby non-LTRs transport flanking sequences to new genomic locations [103, 104] and non-allelic homologous recombination [81], are also common genomic rearrangements triggered by non-LTR TEs. Furthermore, non-LTR retrotransposons can have an impact not only at a DNA level, but also at an RNA level such as RNA editing [62]. All ultimately regulating host gene expression. In fact, the effect of non-LTR retrotransposon insertions on gene expression can occur by numerous mechanisms [62] including:

- The use of non-LTR elements sense and/or antisense promoters to initiate host gene transcription [10].
- The function of non-LTRs as transcription factor binding sites ultimately acting as enhancers or repressors of gene expression [105].
- The insertion into exons, which can result in loss of function mutation [70].
- RNA editing leading to an amino acid substitution when it occurs in the coding sequence, alternative splicing or modification of microRNA [41].

- The introduction of premature polyadenylation [106] and/or RNA polymerase II transcriptional pause sites into genes resulting in termination of transcription within the retrotransposons sequence or reducing gene expression [13, 107].
- Epigenetic regulation such as DNA methylation and heterochromatin formation at the integration site of a new retrotransposon insertion can restrict retrotransposon expression [53].

Accumulation of these events results in genome instability and may be associated with human genetic disorders, but also with evolutionary novelty [18]. While playing a major role in human genome integrity and diversity [108-110] and in non-pathogenic ageing [111, 112] non-LTR retrotransposons have often been identified as risk elements for human health [53]. Their association with genetic diseases via insertion mutagenesis has been demonstrated in instances such as X-Linked Dystonia Parkinsonism, where the presence of an SVA is directly linked to the disease [113]. However, it is important to note that although the presence of an active non-LTR element is thought of as causing a mutation the consequence is not obligatorily deleterious; for example, variation in the poly-T tract of an *Alu* sequence within intron 6 of the *TOMM40* gene is associated with non-pathogenic cognitive ageing [111]. Furthermore, it has been demonstrated that non-LTR retrotransposons may play a role in long-term synaptic plasticity, which may occur as a result of the epigenetic changes that occur during memory formation and learning affecting their function [114]. Non-LTR insertions have functional consequences which are found in healthy populations, and thus, they cannot possibly be identified as deleterious only, but may

also incur regulatory changes and gene expression modifications that are selected as advantageous during human genome evolution [65].

1.4.1 TEs expression and mobilisation in the brain: a mosaic of genomes

In contrast to genes, TEs density in the genome is directly proportional to the biological complexity of the organism [10]. In fact, TEs are present in the genome of most organisms examined today, and the proportion of the human genome that is occupied by TEs is high (˜70 %) [43]. These data imply that most TEs-caused genetic changes are not harmful to the genome, but significantly impact on development and genome architecture and regulation [10]. Not only endogenous TEs affect genome structure, gene regulatory networks and protein function during embryogenesis, but mobilisation of active TEs has been shown to trigger genetic variation both during development and in fully differentiated tissues [10]. In particular, LINE-1 elements, which are the only active autonomous retrotransposons in humans are thought to mobilise at least twice during development: the early embryo and the developing/adult brain [95].

LINE-1 mRNAs are expressed in neuronal precursor cells (NPCs) in the brain, and new L1 insertions can accumulate in cultured human NPCs [115, 116]. The development of next generation DNA sequencing and single cell analysis has demonstrated that the human brain is a mosaic of genomes [9, 69, 117, 118]. Though the rate of retrotransposition in the brain is unclear, it has been proposed to be ubiquitous in the hippocampus of the human brain [118], but there is little information about the frequency of retrotransposition in other brain regions such as the temporal cortex. Nonetheless, LINE-1 activity in the brain could be the mechanism to guarantee that

any two human brains are genetically different. Although additional research is required, several studies suggest that the somatic activity of TEs might be restricted to the human brain [9, 69, 117-121].

1.4.2 The epigenetic regulation of TEs: Bridging the GxE response in ageing

Epigenetic cues can also be affected by genetic variation such as TEs, both at a local, gene-specific, and global, multigene, level [11]. A common epigenetic alteration, DNA methylation, involves the addition of a methyl group to the 5' cytosine of CG pairs named CpG dinucleotides. 70 to 80 % of the total CpG content of the human genome is methylated, and most of the unmethylated portion of CpGs is grouped at CpG islands (CGIs) to the 5' end of genes to aid gene expression [122].

DNA methylation occurs in three phases: establishment or *de novo* DNA methylation, maintenance and demethylation. There are two major *de novo* DNA methylation enzymes (DNMTs) in mammals named DNMT3A and DNMT3B [123]. A third catalytically inactive DNMT named DNMT3L stimulates the activity of DNMT3A and DNMT3B in the germline [124]. *De novo* DNA methylation is not sequence specific, but upon DNA replication only symmetrical CpGs methylation is sustained by the methylation maintenance enzyme DNMT1 [125]. Genome-wide DNA demethylation occurs both during early embryogenesis and in germline reprogramming. During post-fertilization reprogramming, gamete-specific DNA demethylation takes place instead [125]. Throughout genome-wide demethylation, *de novo* DNA methylation has been observed in the human embryonic genome mainly targeting TEs [126]. In addition, maintenance of TEs DNA methylation during early embryogenesis and germline development occurs mainly at evolutionary young and potentially active

TEs, in particular SVAs [127, 128]. This retention of DNA methylation is aided via sequence-specific recruitment of KRAB-associated protein 1 (KAP1) mediated by Krüppel-associated box domain zinc finger proteins (KRAB-ZFPs) [129] signifying the importance of DNA methylation in the regulation of TEs. In fact, the main targets of DNA methylation in mammalian genomes are not genes but TEs, precisely retrotransposons [125]. In general, TEs transcription is repressed when regulatory regions are methylated due to the inability of transcription factors to access the DNA sequence, whereas hypomethylation often suggests high levels of expression [11, 130-132].

As we have mentioned, TEs have been often identified as deleterious to the human genome [53]. In this sense, when active TEs increase in copy number and mobilise, the host genome diminishes or eliminates this activity acquiring then a selective advantage [133]. If this was to occur consistently, eventually we would expect complete silencing and TE loss. However, as this is not the case, it is considered that TEs and host genome experience antagonistic coevolution, which gives rise to genome defence mechanisms against TE replication and mobility, as well as instances where TEs themselves provide a selective advantage to the host [134]. Epigenetic cues such as DNA methylation and/or chromatin modification are one of such defence mechanisms that occur at a genomic level, and whereby silencing of TEs involves altering protein accessibility to DNA required for transcription and hence, the ability of TEs to mobilise [135]. Most TEs tend to be methylated and thus, rendered immobile. As epigenetic changes are heritable, but not permanent, if these were to alter in due course, the element may be reactivated [18]. The DNA

methylation level is known to vary as we age; and, age-related hypomethylation has been observed in humans [122]. In this sense, if there is a tendency towards DNA hypomethylation as we get older [136, 137], we may expect higher TE activity to occur in part as a result of this epigenetic change. In fact, when analysing LINE-1 expression in fetal NPCs versus other somatic cells (skin) from the same donor, Coufal *et al.* demonstrated that there was a subtle change in LINE-1 promoter DNA methylation level, which may explain why LINE-1 mRNAs are expressed selectively in NPCS [115]. Furthermore, it would appear that such a mechanism of TE regulation goes awry in pathological state, with accumulation of retrotransposon's transcripts reported in neurodegenerative diseases such as AD [138], suggesting DNA methylation as an important TE regulatory mechanism during healthy cognitive ageing.

1.5. Genome-wide association studies (GWAS) on cognitive function

GWAS have greatly contributed towards research on the genetic association with cognition and neurodegeneration [4]. Maintaining cognitive function throughout life ultimately translates into living longer and in more desirable conditions [26] and as such, it is important for HA [139]. The environmental factors that contribute towards HA such as diet, physical and social activity are well-established. However, up to the appearance of GWAS, the genetic contributions remained poorly understood [3, 4]. Before GWAS, pedigree-based and candidate gene studies, as well as genetic linkage analysis were used to investigate the association between genotype and phenotype. The large sample number used for GWAS, allowed the genetic association with traits such as health and cognition, where generally – with some exceptions – there is a

small contribution of a large number of genetic variants towards phenotypic variation rather than a direct genotype/phenotype association [140]. The first GWAS on human cognitive function was performed in 2011 [141] and established that human intelligence was highly heritable and polygenic. Subsequently, a GWAS which assessed SNPs obtained from 5,000 individuals from the Cognitive Genomics (COGENT) consortium in four schizophrenia case-control cohorts, established an association between lower cognitive ability and increased risk for schizophrenia, which suggested a genetic overlap between cognitive function and schizophrenia [142]. In 2014, the first GWAS on childhood intelligence (age 6-18 years) reported that childhood intelligence is like adult intelligence, highly heritable and polygenic [141, 143]. This study also concluded that formin binding protein 1-like (*FNBP1L*), which plays a role on reorganization of the actin cytoskeleton during endocytosis, and was previously reported as the most significant gene associated with adult intelligence [141], was also significantly associated with childhood intelligence [143]. A multigenerational GWAS from the same year comprised 7,100 Caucasian individuals addressing 2.5 million SNPs as well as gene-based analysis and found no single-SNPs that reached genome-wide significance but confirmed once again that general cognitive ability is heritable and polygenic [144]. In 2015, a second GWAS on cognitive function by Davies *et al.* conducted a meta-analysis of GWAS across 53,949 people from 31 cohorts. The study reported 13 genome-wide significant SNPs in three different genetic loci 6q16.1, 14q12 and 19q13.32. A significant association with high mobility group nucleosome-binding domain 1 (*HMGN1*) gene on chromosome 21, previously associated with neuropsychiatric phenotypes, was also reported [145], supporting the genetic overlap between cognition and psychiatric conditions [142].

A subsequent study, also by Davies *et al.* used data from the UK Biobank to assess cognitive function based on verbal-numerical reasoning, memory, reaction time and educational attainment [146]. This study provided novel genome-wide significant genetic variants associated to cognitive function and neurodegeneration [146]. Trampush *et al.* assessed the association of about 8 M SNPs to general cognitive function by GWAS meta-analysis in 35,298 healthy individuals of European ancestry across 24 cohorts in the Cognitive Genomics Consortium (COGENT) and identified two novel SNPs in two distinct loci. Whereas the major allele at SNP rs1523041 was strongly associated with better cognitive function and lower expression of cAMP-regulated phosphoprotein 21 (ARPP21), the minor allele at rs2568955 was associated with poorer cognitive ability and higher expression of Ribosomal Protein L31 Pseudogene 12 (*RPL31P12*) in brain tissue, although this gene is annotated as a pseudogene [4]. In the same year, Sniekers *et al.* GWAS meta-analysis of 78,308 individuals reported 15 novel genomic loci and 11 novel genes associated to human intelligence. The identified genes by Sniekers *et al.* are predominantly expressed in brain tissue [147]. The strongest association with intelligence was rs2490272 in an intronic region of forkhead box O3 (*FOXO3*), which has been previously associated with longevity [147, 148]. Savage *et al.*, in 2018, presented a large-scale genetic association study of human intelligence of 269,867 individuals identifying 190 novel genomic loci and 1939 novel genes. This study supported that associated genes are predominantly expressed in the brain [149]. In 2018, yet another study by Davies *et al.* of 300,486 individual aged 16-102, combined cognitive and genetic data from the CHARGE and COGENT consortia, and the UK Biobank and found 148 genome-wide

significant loci associated with cognitive function [139]. The results from the above described GWAS are summarised on Table 1.1.

Table 1.1. Details of breakthrough GWAS of cognitive function to date

Author	Year	n	GWAS SNP hits		GWAS gene hits	
Davis *et al.* **[141]**	2011	3511	0		1 gene	
Lencz *et al.* **[142]**	2014	5000	0		NA	
Benyamin *et al.* **[143]**	2014	17,989	0		0	
Kirkpatrick *et al.* **[144]**	2014	7100	0		0	
Davies *et al.* **[145]**	2015	53,949	3 loci	13 SNPs	1 gene	
Davies *et al.* **[146]**	2016	112,151	20 loci	1,300 SNPs	46 loci	137 genes
Trampush *et al.* **[4]**	2017	35,298	2 loci	7 SNPs	3 loci	7 genes
Sniekers *et al.* **[147]**	2017	78,308	18 loci	336 SNPs	47 genes	
Savage *et al.* **[149]**	2018	269,867	205 loci	NA	1,016 genes	
Davies *et al.* **[139]**	2018	300,486	148 loci	11,600 SNPs	709 genes	

It can be observed that the increase in the number of SNPs and sample size in GWAS is a crucial step towards gaining a deeper understanding of human genetic variation. However, even if GWAS is helping discover genetic variants associated to specific phenotypes, unravelling their functional consequences is challenging due to their small aggregative effect [150]. In fact, a tagging SNP is representative of a large section of DNA that is inherited as one; and thus, the SNP itself may not be the causative agent but rather highlights a target region of DNA [11]. Despite the

polygenic nature of HA and cognitive function, and hence the difficulties associated with the study of such traits, GWAS have repeatedly found genome wide significant genetic associations between the *APOE/TOMM40* locus and cognition [151-153].

1.5.1 A well characterised genetic factor in human cognitive ageing: the *APOE* locus

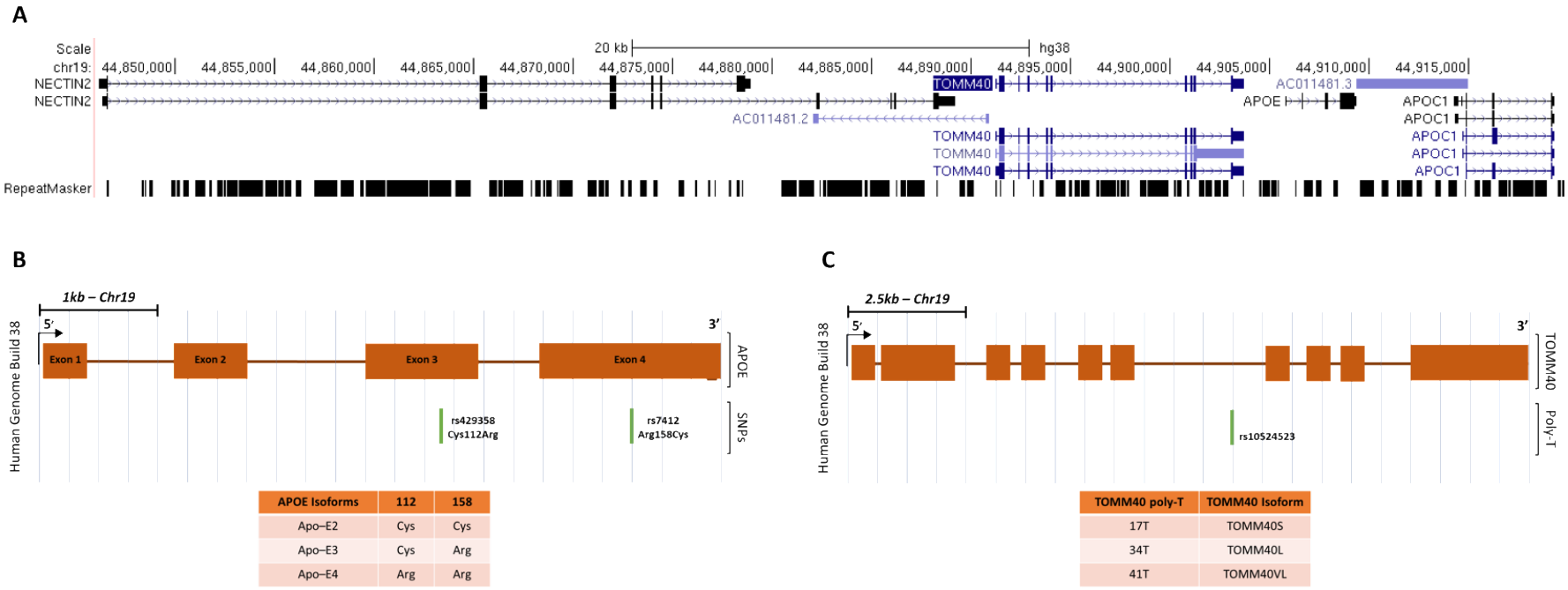
APOE is located in a gene-dense region of chromosome 19 where genes are in moderate to strong linkage disequilibrium (LD) (Figure 1.9A). This region includes Poliovirus Receptor-Related 2 (*PVRL2*/*NECTIN2*), encoding for a membrane glycoprotein component of adherent junctions, which serves as an entry for certain viral strains; *TOMM40*, which codes for TOM protein involved in the transport and sorting of proteins across the mitochondrial membrane [111]; and, Apolipoprotein E (*APOE*), and Apolipoprotein C (*APOC*), which have a role in lipid metabolism [154].

Two *APOE* SNPs are in strong LD giving rise to three *APOE* isoforms named Apo-E2, -E3 and -E4 encoded by alleles Ɛ2, Ɛ3 and Ɛ4 of the *APOE* gene, respectively. The three isoforms differ from each other at amino acid residues 112 and/or 158 (Figure 1.9B). Whereas the *APOE* Ɛ4 allele has been found to be associated with major risk of sporadic AD [24] and lower performance on cognitive tests [154, 155], the *APOE* Ɛ2 allele has been established to be positively associated with survival and longevity in older adults, particularly in older females [112]. Therefore, the *APOE* Ɛ2 allele appears to have a buffering effect of Late Onset Alzheimer's disease (LOAD) [112], which could contribute to healthy cognitive ageing. In addition to the extensive literature on *APOE* driving an observed genetic association to cognition, *TOMM40*, which is a gene

adjacent to *APOE*, is also associated with AD [156-158], cognitive phenotypes in the elderly [153, 155, 159-161] and exceptional longevity [111, 162].

TOMM40 intron 6, a non-coding region of the gene, presents a human specific block of 5 non-LTR *Alu* and a FLAM_A element. The latter comprises a poly-T [111], which presents three variants named short, long and very long differing in the number of T residues (Figure 1.9C). *TOMM40* shorter variations of the poly-T were demonstrated to be associated with reduced vocabulary ability and a slower rate of vocabulary decline with age when compared to the very long poly-T variants [111]. The very long poly T variants are linked to a higher risk of developing LOAD.

TOMM40 short and very long variants are in strong LD with the *APOE* ε3 [163] whereas the *TOMM40* long variant is in strong LD with *APOE* ε4 [156]. Though not specifically studied in the thesis, this data exemplifies that not only SNPs at *APOE* exonic regions, but also the poly-T variants of non-LTR insertions in *TOMM40* intron 6 may distinguish longevity from healthy cognitive ageing. Furthermore, since *TOMM40* and *APOE* are adjacent to one another, then their differential expression could not only be coordinated but aberrant expression of either could have a profound effect on all the genes at the locus.



APOE Isoforms	112	158
Apo–E2	Cys	Cys
Apo–E3	Cys	Arg
Apo–E4	Arg	Arg

TOMM40 poly-T	TOMM40 Isoform
17T	TOMM40S
34T	TOMM40L
41T	TOMM40VL

Figure 1.9. A. Hg38 USCS image of the *APOE* locus showing the genes in LD with *APOE*, named *PVRL2* (*NECTIN2*), *TOMM40* and *APOC* and the repeat masker track illustrating the repetitive DNA elements of the region. **B.** Schematic representation of *APOE*. Two *APOE* SNPs (rs429358 and rs7412) represented by two green vertical lines are in strong LD. They give rise to three *APOE* isoforms named Apo-E2, -E3 and –E4. The three isoforms differ from each other at amino acid residues112 and/or 158. **C.** Schematic representation of *TOMM40* poly-T variants. *TOMM40* FLAM_A region poly-T variants (17T, 34T and 41T), differ in the number of T residues; and, these give rise to the three *TOMM40* isoforms named, short, long and very long.

As we have explained, the appearance of GWAS engendered the first genetic associations with cognition and ageing [3, 4]. However, GWAS have failed to reproducibly establish genetic associations to healthy cognitive ageing other than the *APOE* locus, which has been consistently correlated with cognitive changes [11]. In an attempt to explain the missing genetic component of healthy cognitive ageing, the research on this thesis addresses repetitive DNA elements, including a VNTR and non-LTR TEs, in HA and AD temporal cortex and blood tissues.

As such, the remaining of the thesis is divided upon four sections: chapter 2 explains the materials and methods used throughout the research carried out for the fruition of this thesis, chapter 3 addresses the epigenetic landscape of non-coding regulatory elements, chapter 4 undertakes the identification of non-reference genome RIPs and chapter 5 summarises the research from the thesis and highlights future work. Chapters 3 and 4 are further divided into two subchapters (3.1 & 3.2, and 4.1 & 4.2, respectively), and therefore at the introduction of each chapter there is a small section on the information that encloses the two subchapters. In chapter 3.1, a VNTR located within *DNAJC5*, which functions in neuroprotection [164-166] is addressed. Chapter 3.2 considers hot RC-L1s, which are responsible for the majority of retrotransposition activity in the human genome [73]. Chapters 4.1 and 4.2 assess *Alu*, LINE-1 and SVA elements, which are non-LTR elements active in the human genome [58]. Taken together, we present individual- and tissue- specific genetic variation considering repetitive DNA.

Chapter 2

Materials and Methods

Chapter 2 Materials and Methods

2.1. Materials

2.1.1 Commonly used buffers and reagents

TBE buffer (5x) – 108 g Tris base (Sigma), 55 g Boric acid (Sigma), 5.84 g EDTA (Sigma), made up to 2 L with distilled water.

TAE buffer (10x) – 48.4 g Tris base (Sigma), 3.7 g EDTA (Sigma), 11.4 ml Glacial acetic acid (Sigma) made up to 1 L with distilled water.

LB Broth – 25 g/L in distilled water; autoclaved (Fluka Analytical).

LB Agar – 40 g/L in distilled water; autoclaved (Fluka Analytical).

All other solutions used throughout the thesis were provided with the assay kit used for the specific experiment unless otherwise stated.

2.1.2 Human DNA samples used throughout the project

2.1.2.1 Human blood DNA samples for genotype analysis

2.1.2.1.1 Dyne Steele cohort

The Dyne Steele cohort is a longitudinal study that has been ongoing for about 15 to 20 years screening for cognitive and psychiatric changes. All individuals' background is Caucasian (with equal numbers of volunteers from Manchester and Newcastle). At the time of recruitment, the study partakers were 50 years old and over (mean age at the time of recruitment of 63 years) and showed no signs of cognitive decline. The following information on the samples dates from 2016 when the research carried out on this thesis started. There are 1850 DNA samples with GWAS data available. Plasma, lymphocytes and urine are also available for approximately 400 individuals

from the cohort. Approximately 3 –400 volunteers have agreed to donate their brains after death. There are approximately 100 brains in the bank so far, and it is predicted that 200-250 brains will be collected by time the study is completed. There is pathology data for the approximately 100 brains: 20 samples are healthy aged (age changes only), 11 developed late onset Alzheimer's disease (AD) by the time of death and the rest (71) experience some sort of cognitive decline [incipient, mild, moderate or possible AD, limbic dementia with Lewy bodies (DLB), mild small vessel disease (SVD), etc].

Blood genomic DNA from 281 individuals from the Dyne Steele cohort was extracted from 20 ml of blood using the DNAce MaxiBlood Purification System (Bioline) according to manufacturer's instructions and the concentration determined using the Genemeter (Abgene, UK) that measures absorbance at 260nm. It was then diluted to 100 mg/ml and stored at -40°C. Blood genomic DNA was provided by Prof. Antony Payton and Prof. Neil Pendleton from the Institute for Collaborative Research on Ageing at the University of Manchester, United Kingdom. All blood DNA samples were delivered with information from the longitudinal study detailed on Appendix 1.

2.1.2.1.2 Georgia cohort

Blood genomic DNA from 200 individuals from the Georgia cohort was provided by the Coriell Institute for Medical Research. All individuals' background was Caucasian. A panel of 100 nonagenarians/centenarians (age range 98-108) enclosed 17 males and 83 females. A panel of 100 controls (age range 20-59) included 43 males and 57 females. All blood DNA samples were delivered with information detailed on Appendix 1.

2.1.2.2 Human temporal cortex and blood DNA samples for epigenetic analysis and next generation sequencing

Out of the approximately 100 brains, our laboratory obtained the temporal cortex from 28 people, which included the 18 healthiest samples that showed no signs of cognitive decline and 10 that developed late onset AD by the time of death. 16 out of the 28 temporal cortex samples had matched blood available.

Temporal cortex genomic DNA from a subset of 16 elderly (mean age at the time of death of 89 years; range 78-104 years) individuals from the Dyne Steele cohort was extracted from post-mortem temporal cortex brain tissue using the Gentra Puregene Tissue Kit (QIAgen). 11 individuals (mean age 88 years, range 78-94 years) showed no signs of cognitive decline and died of causes other than neurodegeneration and were thus classified as healthy cognitively aged. 5 individuals (mean age 94 years, range 88-104 years) died having developed late onset AD. Matched blood DNA from each individual was extracted from 20 ml of blood using the DNAce MaxiBlood Purification System (Bioline) according to manufacturer's instructions and provided by Prof. Antony Payton and Prof. Neil Pendleton from the Institute for Collaborative Research on Ageing at the University of Manchester, United Kingdom. All temporal cortex tissue and blood DNA samples were delivered with detailed phenotypic and pathology information from the longitudinal study detailed on Appendix 1.

2.1.3 Human cell line and media

Human cell line – SH-SY5Y (CRL-2266) is a human neuroblastoma cell line from the American Type Culture Collection (ATCC).

Complete media for SH-SY5Y cell line – Minimal Essential Medium Eagle (Sigma) with Nutrient Mixture F-12 Ham (Sigma), supplemented with 10 % (v/v) foetal bovine serum (Sigma), 1 % penicillin/streptomycin (100 U/ml, 100 mg/ml; (Sigma)), 1 % (v/v) 200 mM L-glutamine (Sigma), and 1 % (v/v) 100 mM sodium pyruvate (Sigma).

Freezing media for long-term storage of SH-SY5Y cell line in liquid nitrogen – Freezing media for long-term storage of SH-SY5Y cell line in liquid nitrogen – 90 % foetal bovine serum (Sigma), 10 % DMSO (Sigma).

2.2. Methods

2.2.1 Bioinformatic analysis of transposable elements

The UCSC genome browser and other software freely available on the internet were used for the analysis of the location, structure and tissue expression of transposable elements. These are listed below with the version used:

- UCSC genome browser Hg19, Hg38 (https://genome.ucsc.edu)
- GTEX (https://gtexportal.org/home/)
- Galaxy (http://galaxyproject.org/)

For detailed explanation of these, see sections where they were used for specific purposes.

2.2.2 Primer design for PCR

Primers for PCR amplification were designed by downloading the sequence for the target region with flanking sequence in order to allow room for primer design from the UCSC genome browser. Primer sequences of 17-23 bp were preferred. The potential formation of hairpin structures, as well as the likelihood of homo- and hetero-dimers, and thus the suitability of the primers for PCR was assessed using an Integrated DNA Technologies (IDT) tool named OligoAnalyzer (https://eu.idtdna.com/calc/analyzer) according to manufacturer's instructions. Primers with a melting temperature between 55-65 °C and a GC content of 40 -60 % were chosen; and, primer specificity was verified using the UCSC genome browser In Silico PCR and Blat tools. Primers were purchased from Sigma Aldrich.

2.2.3 Human DNA purification from temporal cortex brain tissue using the Gentra Puregene Tissue Kit

Genomic DNA from temporal cortex tissue was purified using the Gentra Puregene Tissue Kit (QIAgen) according to manufacturer's instructions unless otherwise stated below. This protocol was used for purification of genomic DNA from 100 mg of frozen solid tissue. 100 mg of temporal cortex frozen tissue was quickly thawed and fresh tissue ground on ice with a mortar and pestle. The ground tissue was incubated at 55 °C overnight with 3 ml of cell lysis solution and 15 µl of puregene proteinase K until tissue was completely lysed for maximum yields. Once the resulting pellet was homogenised with DNA hydration solution, the resulting mix was incubated at room temperature (15-25 °C) overnight with gentle shaking. The concentration of purified temporal cortex tissue DNA was measured both by NanoDrop 8000 (Thermofisher), which is a spectrophotometric method to determine the concentration of DNA; and, by Qubit double stranded DNA (dsDNA) broad range (BR) assay (Thermofisher), which is a fluorescent method to determine the concentration of DNA; according to manufacturer's instructions.

2.2.4 Cell culture

2.2.4.1 Culturing of SH-SY5Y cell line

Adherent SH-SY5Y cells were grown in culture media outlined in section 2.1.3 in T75 flasks. When cells reached 70-80 % confluency, these were passaged into new T75 flasks. In order to passage cells, the media was aspirated from the flask and the cells were washed with 10 ml sterile PBS (Gibco). Following the washing of cells, 2 ml of 1x trypsin (Sigma) were added to the flask and spread across the internal surface of the flask by gentle rocking. The flask was placed in the incubator at 37 °C for 3 minutes

until the cells began to detach from the surface. Trypsin was neutralised using 10 ml of culture media. The resulting mix was transferred to a 15 ml tube and the cells were centrifuged at 130 g for 5 minutes. The supernatant was removed and the resulting pellet of cells was resuspended in 10 ml of culture media. 1-2 ml of cell suspension was subsequently transferred to a new T75 flask with 18 ml of culture media.

2.2.4.2 Cell counts with a haemocytometer

To determine the number of cells per ml of media a cell count was carried out using a haemocytometer. A T75 flask of cells at approximately 70 % confluency was passaged as in methods 2.2.4.1 up to when the cell pellet was resuspended in 10 ml of culture media. The haemocytometer and glass coverslip were disinfected with ethanol prior to and after use. The centre of the haemocytometer's counting surface is composed of 25 squares (5x5) limited by three parallel lines each containing 25 smaller squares (5x5). The glass coverslip was placed on top of the counting surface of the haemocytometer and 10 µl of cell suspension were pipetted under either end of the coverslip. The counting surface was visualised under the 10x objective of a light microscope, and the number of cells within the 25 larger squares were counted. Any cells on the top or left hand borders of the 25 squares were included in the count. However, any cells on the bottom or right hand borders were excluded from the count. The counting area of the haemocytometer corresponds to 0.1 mm^3. Therefore, the average number of cells was multiplied by 10,000 to determine the number of cells in 1 cm^3, which is the equivalent of 1 ml. This gave the number of cells per ml of media.

2.2.4.3 Freezing cells for long-term storage in liquid nitrogen

The cells were frozen in freezing media outlined in 2.1.3 for long-term storage in liquid nitrogen. A T75 flask of cells at approximately 70-80 % confluency was passaged as in methods 2.2.4.1 with the cell pellet being resuspended in 10 ml of freezing media instead of culture media. 1.5 ml of cell suspension were transferred into each cryovial. The resulting cryovials were placed into a Mr Frosty freezing container (Thermofisher) with isopropanol and kept at -80 °C for 24 hrs for slow freezing. The cryovials were then transferred to liquid nitrogen for long-term storage.

2.2.5 Agarose gel electrophoresis

PCR amplicons were analysed by applying an electric field (i.e. electrophoresis) to an agarose gel, which allows the separation of DNA fragments by mass as they migrate through the gel. The appropriate amount (0.5 to 2.5 %) of agarose (Bioline) was dissolved in 0.5x TBE buffer (2.1.1) by heating up to a boiling point. Ethidium bromide (EtBr, Sigma, conc. 500 µg/ml), an intercalating nucleic acid stain, was added at 0.8 µl per 10 ml of TBE to allow visualisation of DNA fragments. The percentage of the gel used was inversely proportional to the size of the fragments run on the gel. The agarose gel suspension was poured into the appropriate sized casting tray and a comb inserted at the top of the tray. This was allowed to set for at least 20 minutes at room temperature. The gels were then placed into horizontal gel tanks filled with 0.5x TBE and EtBr (5 µl EtBr/ 1 L of TBE). Each PCR sample was loaded into a single well of the gel. If no loading dye was already present in the PCR reaction, 6x loading dye was added. For sizing of the fragments on the gel, either a 100 bp (Promega) or 1kb (Promega) DNA ladder was loaded depending on the expected size of the fragments.

The voltage (standard running conditions 5 V/cm) and the time for which the gel was run were dependant on the size and percentage of the gel, and the expected fragment size. The DNA was then visualised using a UV transillumintor (BioDoc-It imaging system) and an image was captured.

2.2.6 QIAxcel capillary electrophoresis

The QIAxcel (QIAgen) system allowed high-throughput automated separation of DNA fragments according to their molecular weight by capillary electrophoresis. The QIAxcel was set up and run according to QIAxcel advanced user manual. A gel cartridge, running buffer, and wash buffer, were required for QIAxcel set up. Either a high-resolution cartridge or a screening cartridge were chosen based on the expected bp difference between fragments. The QIAxcel high resolution and screening cartridge allow analysis of DNA fragments between 15 bp and 5 kb in size; a resolution of 3 – 5 bp can be obtained for fragments smaller than 500 bp with the former, whereas the latter allows fragments smaller than 1000 bp to be separated with a resolution of 20 – 50 bp. The appropriate QIAxcel method was selected according to the quantity of sample expected in the product (low (L), medium (M) or high (H)). A range of size and alignment markers are available from QIAgen and were selected based on the expected fragment size and following QIAxcel DNA handbook guidelines. A minimum sample volume of 10 µl was required for analysis. Less than 0.1 µl of the sample was injected into the QIAxcel gel cartridge from a 96-well plate or 12-tube strips placed in the sample plate holder. Desired data collection settings were selected and the samples were run through the QIAxcel gel cartridge. Data was analysed using the BioCalculator software (QIAgen).

2.2.7 Methods for cloning

The genomic sequence of the locus to be cloned was retrieved from UCSC genome browser with flanking sequence, and primers for the target PCR product were designed as outlined in section 2.2.2. A standard 65-55 °C touchdown protocol, in which cycles begin at an initial annealing temperature of 65 °C, and decrease each cycle to a final temperature of 55 °C, as described in Nature Protocols [167]; and, GoTaq Hot Start (Promega) DNA polymerase were used for PCR amplification. This allowed the generation of PCR products with 3´ A-overhangs for high efficiency in TA (thymine/adenine) cloning, which was the method of choice for cloning in this thesis (Figure 2.1).

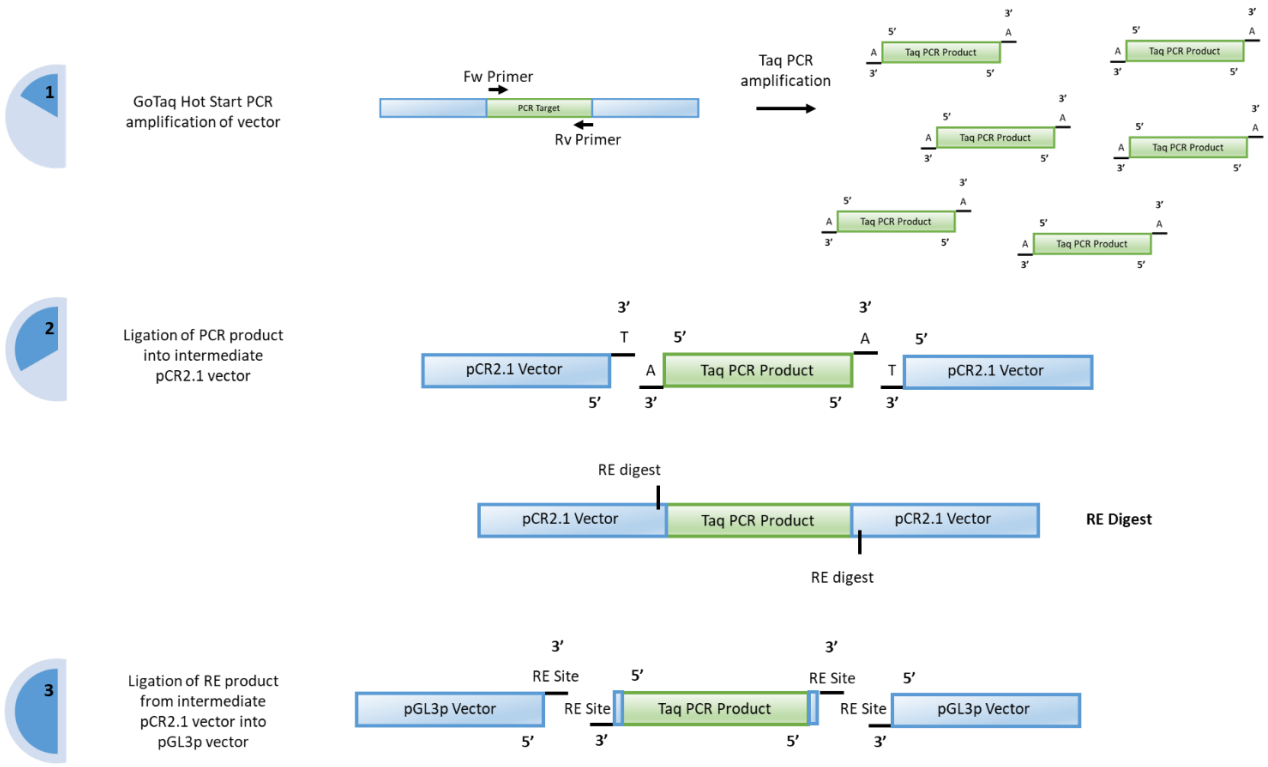


Fig. 2.1. Flow diagram of TA Cloning from PCR amplification of target sequence to pGL3p. 1. GoTaq Hot Start PCR amplification of vector generates a PCR product with 3'-A overhangs. **2.** Ligation of PCR product into intermediate pCR2.1 vector with 5'-T overhangs complementary to 3'-A overhangs of the PCR product generated by GoTaq Hot Start PCR amplification. A restriction enzyme digest (RE digest) is carried out to generate complimentary ends both in the intermediate pCR2.1 vector and the gene expression reporter vector pGL3p. **3.** Ligation of RE product from intermediate pCR2.1 vector into RE digested pGL3p vector to generate gene expression reporter vector of the target sequence.

PCR amplicons were analysed by gel electrophoresis as outlined in section 2.2.5. Following gel electrophoresis, the target bands were carefully excised from the gel and purified using the Wizard SV Gel and PCR Clean-Up System (Promega) according to manufacturer's instructions. DNA was eluted in 30 µl of nuclease free water to keep the concentration of DNA high (10-20 ng/µl). Restriction enzyme digests were used either to create specific nucleic acid overhangs for ligation or as a diagnostic tool for determining the presence and/or orientation of inserts. For restriction enzymes (Promega/NEB) the reaction components shown on Table 2.1 were used. Each enzyme required a specific buffer for optimum activity according to manufacturer's instructions. The digestion reaction was incubated at the appropriate temperature for the enzyme of use, typically for 1-2 hours. Following digestion, 25 % of the reaction mix was run on an agarose gel as outlined in section 2.2.5 to confirm restriction and/or the presence and orientation of the insert.

Table 2.1. Restriction enzyme digest reaction components

Component	Volume (µl)
Buffer 10x	2
BSA (10 mg/ml)	0.2
Enzyme (10 U/µl)	0.5
Template DNA (500 ng)	x
Nuclease free water	y
Total volume	20

2.2.7.1 Ligation of DNA fragments into pCR2.1 intermediate vector

For cloning of PCR fragments into pCR2.1 intermediate vector the TA Cloning Kit (Thermofisher) was used according to manunfacturer's instructions. The TA Cloning Kit uses the pCR2.1 cloning vector and ExpressLink T4 DNA ligase to generate a ligation product. The pCR2.1 vector provides 3'-T overhangs for direct ligation of Taq-amplified PCR product. The amount of insert required for ligation was calculated using the following equation:

Insert (ng) = pCR2.1 vector (ng) x size of insert (kb)/ size of vector (kb)

Three ratios from 10:1 to 100:1 of insert:vector were used in the following ligation reaction (Table 2.2). The ligation reaction into pCR2.1 intermediate vector was incubated at 14 ˚C overnight, and then used in the transformation of competent DH5α E. coli cells as outlined in section 2.2.7.3 or stored at -20°C.

Table 2.2. Ligation into pCR2.1 intermediate vector reaction components

Component	Volume (µl)
ExpressLink T4 DNA Ligase Buffer (5x)	2
pCR2.1vector (25 ng/µl)	2
Insert (~10 ng)	x
ExpressLink T4 DNA Ligase (5 units)	1
Nuclease free water	up to 5
Total volume	10

2.2.7.2 Ligation of DNA fragments into reporter gene pGL3p vector

The cloning of inserts into pGL3p vector was completed excising a fragment from the intermediate cloning vector pCR2.1 and ligating it to the pGL3p vector with complementary overhangs generated after restriction enzyme digest. The ligation (Table 2.3) was generally carried out at a molar ratio of insert to vector of 3:1. The amount of insert required was calculated using the following equation:

Insert (ng) = pGL3p vector (ng) x size of insert (kb)/ size of vector (kb)

The ligation reaction into reporter gene pGL3 vectors was incubated at room temperature for 3 hours and then used in the transformation of competent DH5α E. coli cells as outlined in section 2.2.7.3 or stored at -20°C.

Table 2.3. Ligation into reporter gene pGL3p vector reaction components

Component	Volume (µl)
Ligation Buffer (10x)	1
pGL3 vector (x ng/µl)	x
Insert (~500 ng)	y
T4 DNA Ligase (5 units)	1
Nuclease free water	z
Total volume	10

2.2.7.3 Transformation of the ligation reaction into competent DH5α E. coli cells

The ligation mix was transformed into competent DH5α E. coli cells (Invitrogen) according to manufacturer's instructions. In summary, 5 µl of ligation mix were added to 50 µl aliquots of previously thawed competent DH5α cells, tapped gently to mix and then incubated on ice for 30mins. The transformation reaction was then heat shocked at 42°C for 20-25 seconds and placed on ice for 2 minutes. 950 µl of pre-warmed LB broth (2.1.1) was added to the transformation mixture and placed in a 225 rpm shaking incubator at 37°C for 1 hour. 200 µl of the transformation mixture was spread onto pre-warmed LB agar (2.1.1) plates. The plates had been prepared with the appropriate amount of specific antibiotic (100 µg/ml ampicillin). The plates were incubated at 37 °C overnight. The remaining transformation reaction was stored at 4 °C for future use if required.

2.2.7.4 Purification of plasmid DNA from transformed E. coli

2.2.7.4.1 Mini-prep of plasmid DNA

Individual colonies were picked from an overnight incubation of transformed E. coli culture on LB agar plates (100 µg/ml ampicillin) prepared as outlined in section 2.2.7.3, and grown overnight in 5 ml of LB broth with 100 µg/ml ampicillin at 37 ºC, shaking at 225 rpm. Plasmid DNA was extracted and purified from 1 ml of the 5 ml overnight culture using either the QIAprep Spin Miniprep Kit (QIAgen) or the Wizard Plus SV Miniprep DNA purification system (Promega) according to manufacturer's instructions. DNA was eluted in 30 µl nuclease free water and either stored at -20 ºC or directly used for restriction enzyme digest.

2.2.7.4.2 Maxi-prep of plasmid DNA

For successful transfection of eukaryote cells as on section 2.2.7.5.1, plasmids with great purity and yield are required. 100 µl from 5 ml mini-prep bacterial culture (section 2.2.7.4.1) were added to 100 ml LB broth with 100 µg/ml ampicillin and grown overnight at 37 °C shaking at 225 rpm. Plasmid DNA was extracted and purified using the QIAgen Plasmid Maxi Kit (QIAgen) according to manufacturer's instructions. DNA pellets were resuspended in 150 µl nuclease free water and the concentration determined using a NanoDrop 8000 before storing for later use at -20 ºC.

2.2.7.4.3 Sequencing of PCR products and plasmids

Sequencing of PCR products and plasmids with successfully cloned inserts was carried out externally by Source Bioscience. Sequencing required 5 µl of sample at 100 ng/µl, and 5 µl of each sequencing primer at 3.2 pmol/µl.

2.2.7.5 Analysis of reporter gene expression

2.2.7.5.1 Transient transfection of reporter gene constructs into SH-SY5Y cell line

Turbofect (Thermo Scientific) was used to transfect SH-SY5Y cell line according to manufacturer's guidelines. The cells were counted as outlined in section 2.2.4.2. The cells were plated 48 hours prior to transfection into 24-well plates at a concentration of 100,000 cells per well in 1 ml of cell culture media. For each transfection, 1 µg of the reporter gene construct, 10 ng of the internal control and 2 µl of Turbofect were combined in a total volume of 100 µl of serum free media, vortexed and then incubated for 30 minutes at room temperature and the 100 µl mix added to 1 ml of cells in culture media in each well. Following transfection, cells were maintained in serum free media for 4 hours to prevent growth, then washed with PBS to remove

the transfection mixture, thus reducing cell death and finally cultured in fresh media for 48 hours.

48 hours after transfection, the cells were lysed using passive lysis buffer (PLB) in preparation for the Dual Luciferase Reporter Assay (Promega). The cells were washed with PBS prior to the addition of PLB. 100 µl of 1x PLB were added to each well. 24-well plates were then placed on a rocking platform for 15 minutes. 20 µl of the cell lysate were transferred to an opaque 96 well plate for the dual luciferase assay.

2.2.7.5.2 Measuring levels of reporter gene activity using the Dual Luciferase Reporter Assay

The appropriate amount of luciferase assay reagent II (LARII) and Stop and Glo reagent was prepared for the number of measurements required and allowed to reach room temperature. The opaque 96 well plate containing the cell lysate was placed into a Glomax 96 Microplate Luminometer (Promega). Both the LARII and the Stop and Glo reagents were automatically dispensed after one another at automated intervals via two injectors. Prior to starting the assay, the injectors were flushed with distilled water, 70% ethanol, a second round of distilled water and air in order to clean them. Injector 1 was then primed with LARII whereas Stop and Glo primed injector 2. Subsequently, the Promega dual luciferase program was run, which measured the bioluminescence from the reaction catalysed by the firefly and renilla luciferase enzymes. The LARII was added first to measure the bioluminescence produced by the reaction catalysed by the firefly luciferase protein and then, the Stop and Glo quenched this reaction and was used to measure the bioluminescence from the reaction catalysed by the renilla luciferase protein. Using the measurements from

the action of the two reporter gene constructs that were co-transfected, the activity of the constructs across the different wells could be accurately compared as the internal control reduces/allows one to compensate and adjust for experimental variability caused by differences in transfection efficiencies. A two-tailed student's t-test was carried out for statistical analysis.

2.2.8 Genotyping of human DNA samples

The VNTR at the *DNAJC5*/*MIR941* locus and selected hot RC-L1 elements were PCR amplified for genotyping.

2.2.8.1 Genotyping of *MIR941*/VNTR region

The VNTR at the *DNAJC5*/*MIR941* locus was genotyped in 281 individuals from the Dyne Steele cohort (section 2.1.2.1) by PCR. Primers were designed as outlined in section 2.2.2. PCR amplification was carried out using 20 ng of blood genomic DNA and using primers (Sigma) designed to target *MIR941*/VNTR including Fw: 5'-ACGTGTCCGGGGAGAGGACG-3' and, Rv: 5'-CCCGGTCCGACGCAGGAC-3'. The PCR reaction was performed using GoTaq Hot Start (Promega) DNA polymerase (Table 2.4). A thermocycler set to Standard Touchdown 65-55° as described in Nature Protocols [167] and 4 ˚C infinite hold was used for PCR amplification. PCR products were run both on an agarose gel as outlined in section 2.2.5; and, using QIAxcel capillary electrophoresis as outlined in section 2.2.6. IBM SPSS Statistics v23.0 software was utilised for statistical analysis by Prof. Neil Pendleton from the Institute for Collaborative Research on Ageing at the University of Manchester, United Kingdom. A series of statistical tests such as t-test, chi-square test and cross

tabulation were used depending on the phenotypic data to be scrutinised. The statistical analysis is detailed on Appendix 2.

Table 2.4. GoTaq Hot Start DNA polymerase PCR reaction components

Component	Volume (µl)
PCR buffer (5x)	4.0
MgCl2 (25 mM)	2.5
dNTPs (10mM each)	0.4
Forward primer (20 mM)	0.1
Reverse primer (20 mM)	0.1
DNA polymerase (5 u/µl)	0.1
Nuclease free water	10.8
DNA template (5 ng/µl)	4.0
Final volume	22

2.2.8.2 Hot RC-L1s genotyping

Five hot RC-L1s were selected for genotyping. The elements were shortlisted based on their high level of activity in a cellular retrotransposition assay, high number of germline offspring elements from 3' transductions analysis or high number of somatic insertions in cancer from 3' transductions analysis. Hot RC-L1s elements were genotyped in 12 individuals from the Dyne Steele cohort (section 2.1.2.2) by PCR. Primers were designed as outlined in section 2.2.2. PCR amplification was carried out using 5 ng of DNA and using primers (Sigma) designed to target both the empty site (ES) and the 5' end of the active L1 element. The PCR reaction was performed using

GoTaq Hot Start (Promega) DNA polymerase (Table 2.5 and table 2.6). PCR was performed either single, for those L1 elements where the empty site could not be multiplexed together with the 5' end; or, as a multiplex reaction to simultaneously visualise both the empty site and the 5' end of the insertion. For empty site PCR amplification, forward and reverse primers were used. For 5' end PCR amplification forward or reverse (depending on whether the L1 was on the positive or negative strand), and 5' L1 primers were used. For multiplex PCR amplification, a combination of forward, reverse and 5' L1 primers was used. For schematic representation and description of PCR amplification of hot RC-L1s see figure 2.2. A thermocycler set to 95 ˚C for 2 mins; 95 ˚C for 30 secs, 60/64 ˚C for 30 secs and 72 ˚C for 1 min per kb for 35 cycles; 72 ˚C for 5 mins and 4 ˚C infinite hold was used for PCR amplification. PCR products were run both on an agarose gel as outlined in section 2.2.5; and using QIAxcel outlined in section 2.2.6 for high throughput. Plink, on python programming language, was utilised for statistical analysis. A series of statistical tests such as t-test and chi-square test were used. The statistical analysis is detailed on Appendix 3.

Table 2.5. GoTaq Hot Start DNA polymerase single PCR reaction components

Component	Volume (µl)
PCR buffer (5x)	5.0
MgCl2 (25 mM)	4.0
dNTPs (10mM each)	0.5
Forward primer (20 mM)	0.5
Reverse primer (20 mM)	0.5
DNA polymerase (5 u/µl)	0.125
Nuclease free water	13.375
DNA template (5 ng/µl)	1
Final volume	25 µl

Table 2.6. GoTaq Hot Start DNA polymerase multiplex PCR reaction components

Component	Volume (µl)
PCR buffer (5x)	5.0
MgCl2 (25 mM)	4.0
dNTPs (10 mM each)	0.5
Forward primer (20 mM)	0.5
Reverse primer (20 mM)	0.25
5' L1 primer (20 mM)	0.25
DNA polymerase (5u/µl)	0.125
Nuclease free water	13.375
DNA template (5 ng/µl)	1
Final volume	25 µl

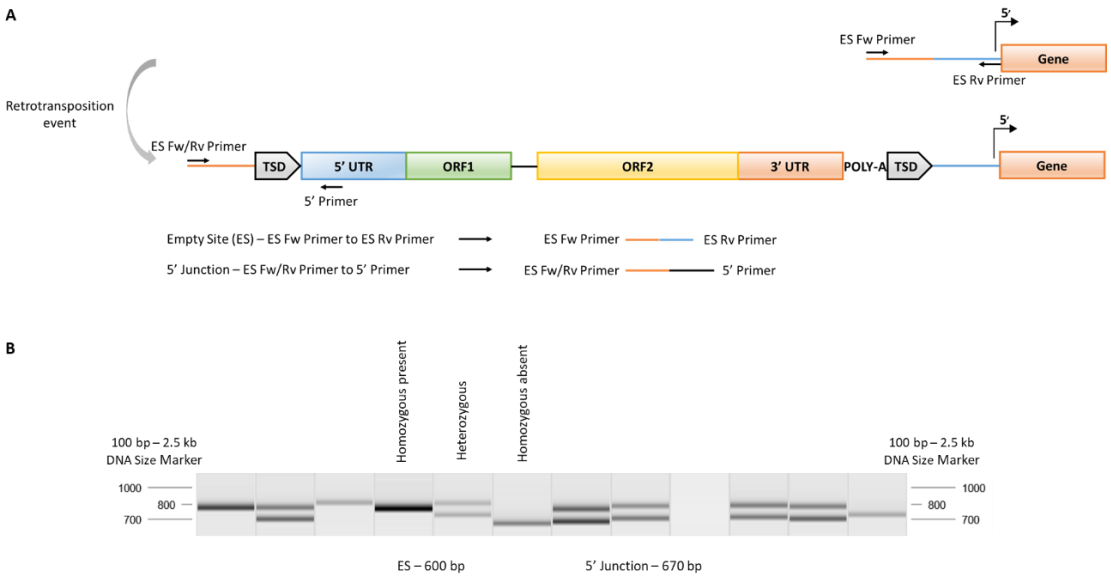


Fig. 2.2. A. Schematic representation of a LINE-1 Empty Site (ES)/5' Junction PCR. The ES PCR involves amplification by primers annealing in the 5' and 3' genomic flanks of the putative insertion. The expected ES is usually favoured during amplification and competes with the amplification of the filled site if there was to be one. 5' junction amplification requires a primer annealing either in the 5' or 3' genomic flanks of the putative insertion, and a primer annealing in the consensus L1 sequence in reverse sense. **B. Example of QIAxcel gel images of an Empty ES/5' Junction PCR product of Ref_chr8_q24.22 L1.** The ES PCR product expected size is 600 bp, whereas the 5' junction of the L1 insertion PCR product expected size is 670 bp. QIAxcel capillary electrophoresis, AL420 method at an injection time of 20 seconds was used to analyse ES/5' Junction PCR. A 15 bp to 3 kb alignment marker and a 100 bp to 2.5 kb size marker were chosen to measure DNA fragments. The inconsistency between the expected and the apparent band sizes (i.e. heterozygous versus homozygous) is possibly due to the difficulty to size repetitive LINE-1 elements.

2.2.9 Bisulphite DNA modification and pyrosequencing

2.2.9.1 Bisulphite treatment

Sodium bisulphite treatment of genomic DNA to convert the unmethylated cytosines into uracil was carried out using the EZ-96 DNA methylation-gold kit (Zymo Research, Cambridge Bioscience). All solutions used were provided in the kit unless otherwise stated. 250 ng of temporal cortex and blood DNA were diluted up to a volume of 20 µl in a 0.2 ml PCR tube. 130 µl of the CT conversion reagent prepared according to manufacturer's instructions were then added. A thermocycler was set to 98 °C for 10 mins, 64 °C for 2.5 hrs for bisulphite modification of genomic DNA.

600 µl of M-binding buffer were then added to a 1.5 ml tube and mixed thoroughly with previously prepared 150 µl of bisulphite conversion reaction. The resulting mix was then loaded into the Zymo-spin IC binding plate for isolation of bisulphite converted DNA. All centrifugation steps were performed at >10,000 g for 1 minute unless otherwise indicated and flow-through was discarded. 200 µl of M-wash buffer were used to wash the columns followed by a centrifugation step. 200 µl of M-desulphonation buffer were added to each column and incubated at room temperature for 20 minutes before a 30 secs centrifugation step. A final wash step was performed by adding 300 µl of M-wash buffer to each column followed by a 4 minutes centrifugation step. Finally, bisulphite treated DNA was eluted by adding 20 µl of M-elution buffer pre-warmed to 65 ˚C to each well of the plate; and, DNA was recovered by centrifugation after 5 minutes incubation

at room temperature. Bisulphite modified DNA was stored at -20°C for later use. When long-term storage was required, bisulphite modified DNA was stored at -80°C.

4 µl (about 60 ng) of the bisulphite converted DNA were used for PCR amplification to check the DNA quality. PCR amplification was performed using the Pyromark PCR kit (QIAgen, Table 2.7) and a primer (MWG Eurofins) mix designed to target LINE-1 (Table 2.7.1), one of the primers which was biotinylated. A thermocycler set to 95 ˚C for 15 mins; 94 ˚C for 30 secs, 58 ˚C for 45 secs and 72 ˚C for 45 secs for 40 cycles; 72 ˚C for 10 mins and 4 ˚C infinite hold was used for PCR amplification. 3 µl of biotinylated PCR product were run on a 2% agarose gel and separated by electrophoresis at 80 mV for 40 mins. It was only necessary to run 10% of the samples for confirmation of the quality of PCR amplification.

Table 2.7. Pyromark PCR kit PCR reaction components

Component	Volume (µl)
Pyromark master mix (2x)	15.0
CoralLoad Concentrate (10x)	3.0
Primer mix	1.2
Nuclease free water	6.8
Treated DNA template (60 ng)	4.0
Final volume	30

Table 2.7.1. LINE-1 primer mix

Component	Volume (µl)
Biotinylated Primer (20 mM)	7.5
Non-biotinylated Primer (20 mM)	15
Nuclease free water	177.5
Final Volume	200

2.2.9.2 Clean-up of biotinylated PCR product for pyrosequencing

Biotinylated PCR products were cleaned-up using a Q96 vacuum prep workstation (QIAgen). The trays of the vacuum workstation were prepared with PyroMark Gold Q96 reagents and PyroMark buffer solutions according to manufacturer's instructions. Biotinylated PCR products were bound to streptavidin-coated sepharose beads (GE Healthcare). The beads with the bound biotinylated PCR product were captured with the vacuum tool on the vacuum workstation; thoroughly washed with 70% ethanol for 10 seconds; and, subsequently washed with denaturation buffer (sodium hydroxide) and rewashed with wash buffer (sodium acetate).

Following the clean-up reaction, single-stranded DNA was obtained and ready for pyrosequencing. 25 µl of template DNA were added to the PyroMark Q96 plate containing the sequencing primer (Table 2.8) targeting LINE-1 consensus sequence diluted to 0.3 µM in annealing buffer. A shaking hot plate was set at 80 °C to perform primer annealing for 2 minutes, and subsequently the plate was incubated at room temperature for 5 minutes to cool down.

The plate was placed into the PyroMark instrument, and the PyroMark Q96 cartridge loaded with the PyroMark Gold reagents, including the enzymes, nucleotides, and substrate for the pyrosequencing reaction. These were pipetted into the dispensing PyroMark Q96 reagent cartridges, according to the volumes provided by the software, and placed into the PyroMark Q96 ID to perform the pyrosequencing run following

manufacturer's instructions (QIAgen). The methylation status of each CpG site on the LINE-1 target region was then analysed individually.

PyroMark Q96 ID Software 2.5 (QIAgen). was used for the methylation analysis of LINE-1 according to manufacturer's instructions. The global methylation status of the 5' promoter CGI of LINE-1 was calculated including 6 CpG sites present in the sequenced target region. A student's t-test was used to determine statistical significance.

Table 2.8. LINE-1 primer sequences [168]

Primer name	Primer sequence
LINE-1 Fw	5'-BIO-TAG GGA GTG TTA GAT AGT GG-3'
LINE-1 Rv	5'-AAC TCC CTA ACC CCT TAC-3'
LINE-1 Seq	5'-CAA ATA AAA CAA TAC CTC-3'

2.2.10 Isolation of unmethylated and methylated DNA from human temporal cortex and blood DNA

Unmethylated and methylated DNA was isolated using the CpG MethylQuest kit (Millipore) according to manufacturer's instructions unless otherwise stated below. Briefly, 500 ng of temporal cortex and blood DNA were used as starting material. Genomic DNA shearing was performed using S220 focused-ultrasonicator (Covaris). To obtain a 500 bp fragment size the duty cycle was set to 5%, intensity was set to 3 and cycles per burst were 200, during 40 seconds x4. A 1:1.1 Agencourt AMPure XP Beads (Beckman Coulter) ratio was used for concentration of fragmented DNA. DNA was eluted

in 25 µl of nuclease free water for the following step. 1 µl of eluted DNA was used in the Qubit dsDNA BR Assay Kit (Fischer Scientific) to check the DNA concentration after sonication. The QIAxcel (QIAgen) AL420 method (20 seconds injection time) was used to check the DNA size distribution using 2 µl of eluted DNA as template. After CpG pulldown, 100 µl of unmethylated and methylated DNA were isolated for downstream applications as per manufacturer's instructions using 22 µl of fragmented DNA as starting material.

2.2.10.1 Positive and negative control of isolated unmethylated and methylated DNA by PCR

Analysis of isolated DNA was performed according to manufacturer's instructions unless otherwise stated below. PCR amplification was performed using template DNA and positive and negative control primers provided with the CpG MethylQuest kit. The positive control primers target *SNRPN* CGI, an imprinted gene which produces a 230 bp PCR amplicon in both the unmethylated and methylated DNA fractions and would be used in all samples (HeLa control DNA and temporal cortex and blood DNA). The negative control primers target *COX2* CGI, which produces a 442 bp PCR amplicon in the unmethylated portion of HeLa control DNA only. The PCR reaction was performed using AllTaq (QIAgen) DNA Polymerase (Table 2.9). A thermocycler set to 95 ˚C for 2 mins; 95 ˚C for 5 secs, 60 ˚C for 15 secs and 72 ˚C for 10 secs for 35 cycles; and, 4 ˚C infinite hold was used for PCR amplification. 16 µl of PCR product were run on a 1.8% agarose gel and separated by electrophoresis at 100 mV for 90 mins.

Table 2.9. AllTaq DNA polymerase PCR reaction components

Component	Volume (µl)
PCR buffer (5x)	4.0
dNTPs (10 mM each)	0.4
+ve/-ve control primers (0.5 µM each/ 1.0 µM each)	1.0
Mastermix tracer (125x)	0.2
DNA polymerase (5 u/µl)	0.4
Nuclease free water	12.0-13.5
Treated DNA template	0.5-2.0
Final volume	20 µl

2.2.11 qPCR for telomere length quantification

qPCR for telomere length quantification was performed using the absolute human telomere length quantification qPCR assay kit (Science Cell) according to manufacturer's instructions unless otherwise stated below. The qPCR reaction was performed using FastStart Essential DNA Green Master Mix (Roche, Table 2.10) in an Applied Biosystems 7900HT Fast Real-Time PCR System (Table 2.11).

Table 2.10. FastStart Essential DNA Green Master Mix qPCR reaction

Component	Volume (µl)
Reference genomic DNA sample/Genomic DNA template (5 ng/µl)	1.0
Primer stock solution (SCR/telomere)	2.0
qPCR master mix (2x)	10
ROX Solution	1.0
Nuclear free water	6
Total volume	20

Table 2.11. qPCR protocol for Applied Biosystems 7900HT Fast Real-Time PCR System

Step	Temperature	Time	Number of Cycles
Initial denaturation	95 °C	10 min	1x
Denaturation	95 °C	20 sec	
Annealing	52 °C	20 sec	32x
Extension	72 °C	45 sec	
Data acquisition	Plate Read		
Optional	Melting curve analysis		1x
Hold	20 °C	Indefinite	1x

2.2.12 Retrotransposon Capture Sequencing (RC-Seq)

Sanchez-Luque F. J. *et al.* (2016) Retrotransposon Capture Sequencing (RC-Seq): A Targeted High-Throughput Approach to Resolve Somatic L1 Retrotranposition in Humans [88] was used as stated below while learning the protocol at the Paul-Ehlich-Insitut (PEI) in Langen (Germany). The adaptation to perform RC-Seq at the University of Liverpool is described as alternatively below and, in full on Appendix 4.

2.2.12.1 DNA shearing for library preparation

5 µg of temporal cortex and matched blood DNA were used as starting material for shearing. DNA shearing was performed using M220 focused-ultrasonicator (Covaris). To obtain a 250 bp fragment size the peak power was set to 50, duty factor was set to 20 and cycles per burst were 200, during 120 seconds. Alternatively, DNA shearing was performed using S220 focused-ultrasonicator (Covaris). To obtain a 250 bp fragment size the duty cycle was set to 10%, intensity was set to 5 and cycles per burst were 200, during 60 seconds x3.

2.2.12.2 LINE-1 library preparation

DNA was concentrated by using Agencourt AMPure XP Beads (Beckman Coulter). 1 µl of DNA was used in the Qubit dsDNA BR Assay Kit (Fischer Scientific) to check the DNA concentration after sonication and following DNA concentration. 2 µl of DNA were used in the Fragment Analyser (Advanced Analytical) to check the DNA size distribution. Alternatively, the QIAxcel (QIAgen) AL320 method (10 seconds injection time) was used to check the DNA size distribution as outlined in section 2.2.6.

1 µg of DNA was used as starting material for library preparation. A thermocycler set to 30 ˚C block, 40 ˚C lid was used for end repair of sonicated DNA during 30 mins. DNA was purified by using Agencourt AMPure XP Beads. A thermocycler set to 37 ˚C for 30 mins, 70 ˚C for 5 mins and 4 ˚C infinite hold, 80 ˚C lid was used for A-tailing. A thermocycler set to 30 ˚C block, 40 ˚C lid was used for ligation to adapters during 10 mins. The ligation reaction was stopped using stop ligase mix. DNA was purified by two subsequent rounds of Agencourt AMPure XP Beads.

2.2.12.3 Agarose gel-size selection

A 2% high-resolution agarose (Sigma) gel was prepared in TAE buffer (section 2.2.1). DNA samples were loaded in the gel. Gel electrophoresis was run at 120 V. 290-310 bp, 310-350 bp, 350-380 bp and 380-410 bp bands per sample were cut out and subsequently purified from the gel using the MiniElute Gel Extration Kit (QIAgen) according to manufacturer's instructions. Alternatively, an extra round of Agencourt AMPure XP Beads was used for size selection whereby 0.65:1 volumes of beads followed by 0.2:1.65 volumes of beads were used for size selection. 30 µl of sample were transferred to a 0.2 ml tube for the following step.

A thermocycler set to 98 ˚C for 45 secs; 98 ˚C for 15 secs, 60 ˚C for 30 secs and 72 ˚C for 30 secs for six cycles; 72 ˚C for 5 mins and 4 ˚C infinite hold was used for ligation mediated (LM) PCR. DNA was purified by using Agencourt AMPure XP Beads. 1 µl of DNA was used in the Qubit dsDNA BR Assay Kit to check the DNA concentration after size selection. 2 µl of DNA, were used in either the Fragment Analyser or the QIAxcel to check the DNA

size distribution. A preferable fragment size distribution between 340 and 410 bp, with a median fragment peak of 370bp was selected to proceed to the next step. Alternatively, an extra round of Agencourt AMPure XP Beads was used for size selection.

2.2.12.4 Hybridization of LINE-1 libraries to sequencing probes

1 µg of total DNA was used as starting material for hybridization. 500 ng of DNA from both temporal cortex and blood DNA library were used as starting material for hybridization. Temporal cortex and blood DNA libraries were pooled on a 1:1 ratio. 1µl of sequence capture developer reagent, 1 µl of universal blocking oligo, which bind to library adapter sequences to reduce off-target capture during library enrichment, and 1 µl of index-specific blocker oligo, which allowed multiplexed sequencing of libraries, per 1µg of total DNA were used. The whole sample was dried using a speed vac (Eppendorf) at 70 ˚C for 60 mins. A thermoblock set to 95 ˚C for 5 minutes was used to incubate the sample for hybridization. A thermocycler set to 95 ˚C block, 105 ˚C lid for 3 minutes was used to ligate locked nucleic acid (LNA) probes to sample. A thermocycler set to 47 ˚C block, 57 ˚C lid for three days was used to complete hybridisation.

2.2.12.5 Capture recovery and amplification of LINE-1 libraries

The captured library was pulled-down by using previously washed DYNAL Dynabeads M-270 Streptavidin (Fischer Scientific). A thermocycler set to 47 ˚C block, 57 ˚C lid for 45 mins was used to pull-down captured libraries. A thermocycler set to 98 ˚C for 45 secs; 98 ˚C for 15 secs, 60 ˚C for 30 secs and 72 ˚C for 30 secs for eight cycles; 72 ˚C for 5 mins and 4 ˚C infinite hold was used for LM-PCR. DNA was purified by using the MiniElute Gel

Extration Kit according to manufacturer's instructions. 1 µl of DNA was used in the Qubit dsDNA high sensitivity (HS) Assay Kit (Thermofisher) to check the DNA concentration after capture recovery and amplification. 5' and 3' captured libraries were pooled in a 3:7 ratio. If there was not enough captured library DNA, three cycles of LM-PCR and subsequent DNA purification by using Agencourt AMPure XP Beads was performed. 2 µl of DNA, were used in either the Fragment Analyser or the QIAxcel to check the DNA size distribution.

2.2.12.6 Sequencing of LINE-1 libraries

12 LINE-1 enriched libraries were pooled together to send for sequencing according to Sanchez-Luque *et al.* protocol. Sequencing was done on an Illumina NextSeq 500 system by the Centro Pfizer – Universidad de Granada – Junta de Andalucía de Genómica e Investigación Oncológica (GENYO) which resulted in 8 fastq files per samples [4 forward (R1)) and four reverse (R2)].

2.2.12.7 Calculating sequencing coverage of LINE-1 libraries

In order to estimate the sequence coverage of LINE-1 libraries, the number of uniquely mapping reads at the 5´and 3´ ends of full-length reference L1 elements was calculated. Our former post-doc researcher, Abigail L. Savage extracted the coordinates of human specific full-length L1s from UCSC and removed those L1s with evidence of being RIPs. Two bed files covering 100 bp at the 5´and 3´ ends of full-length reference L1s from the above list were generated. Subsequently, using bedtools multicov and the bam files for

each of the samples, the number of uniquely mapping reads over these regions was measured.

The average number of reads over the 5′ and 3′ends of human specific full-length L1 elements, respectively was considered as the sequence coverage at 5′and 3′ ends of LINE-1 libraries (Supplementary figure 4.1.1).

2.2.13 Whole Genome Sequencing (WGS)

Whole genome sequencing of temporal cortex and blood DNA at 40x depth was carried out externally by the Australian Genome Research Facility (AGRF) by providing 1 µg of genomic DNA per sample in a 96-well plate.

2.2.14 Bioinformatic analysis of next generation sequencing

2.2.14.1 Setting up bioinformatic analysis for RC-Seq and WGS

To analyse sequencing data, resulting fastq files were downloaded. The University of Liverpool host server name is chadwick.liv.ac.uk. To log into the server, FileZilla and login in credentials were used:

- Host: chadwick.liv.ac.uk
- Username: username
- Password: xxxxxxxx
- Port: 22

Using FileZilla, a folder was created for each sample on the scratch directory (/scratch/username). Fastq files were then uploaded to chadwick server via FileZilla. In order to access the server, PuTTy, which is a terminal emulation software that runs on Microsoft Windows, was used. PuTTY was selected as terminal of choice to run commands using Python, which is a programming language that allows working quickly and integrating systems more effectively. PuTTy was launched; and, username and password were typed into the black window that appears when opening PuTTY. When logging into putty, the default location is the home directory (/home/username). All our commands were run from the scratch directory. To change directory the command cd (cd /scratch/username) was used. Figure 2.3 illustrates the outlook of FileZilla, PuTTy and FastQC, which is described below.

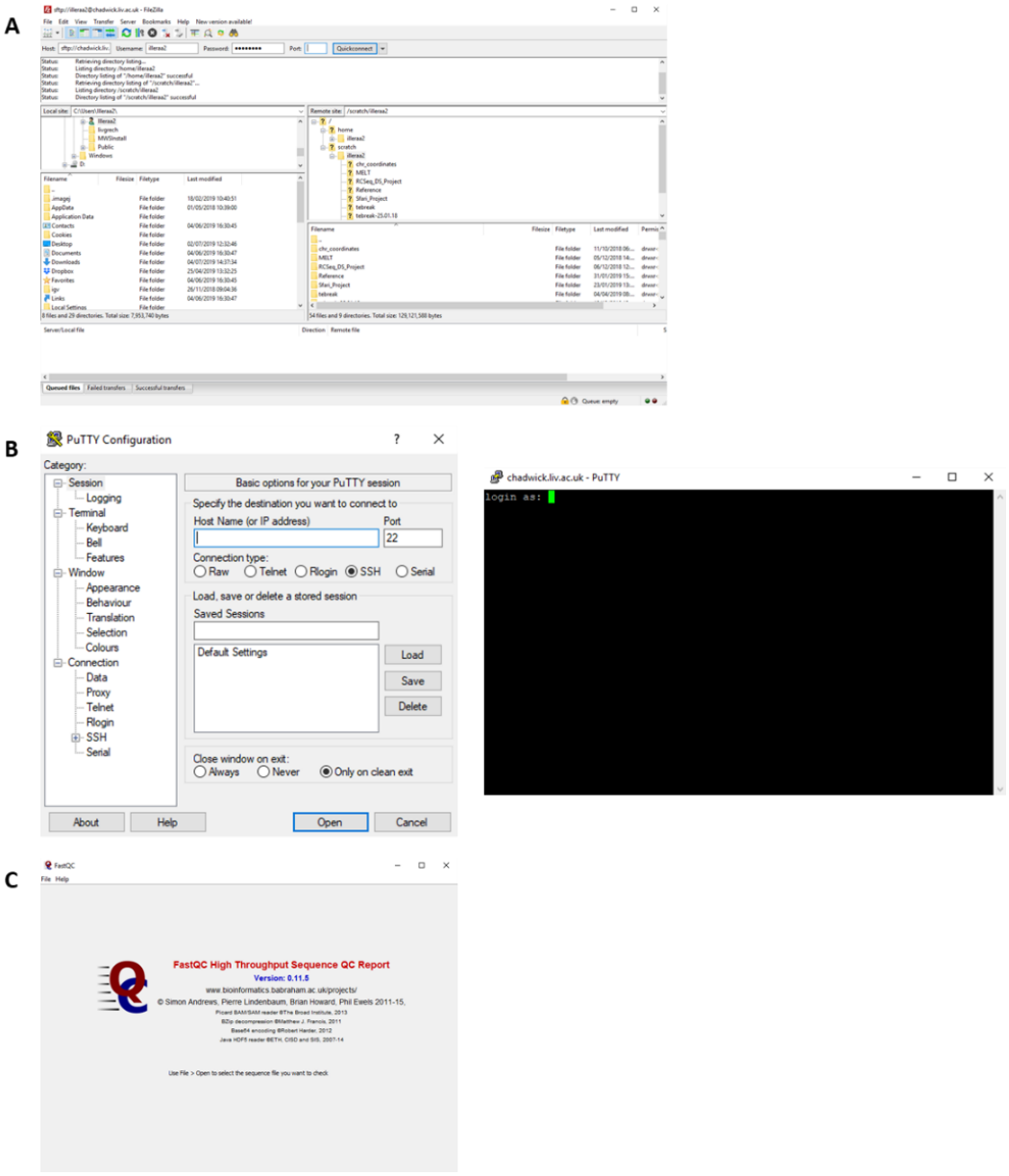


Fig. 2.3. Bioinformatics set up. A. FileZilla, which is a freely available software for transferring files over the Internet. **B.** PuTTY, which terminal emulation software that runs on Microsoft Windows. **C.** FastQC, which is a quality control tool aiming to provide a way to do quality control checks on raw sequence data from high throughput sequencing pipelines.

RC-Seq libraries were sequenced on Nextseq, which resulted in 8 fastq files per sample, including 4 forward (R1) and 4 reverse (R2) reads. WGS was carried out on IlluminaSeq, which resulted in 4 fastq files per sample, including 2 forward (R1) and 2 reverse (R2) reads. To assemble overlapping read pairs into a single fastq file in a process named concatenation, the files containing forward reads were combined together and so were the reverse reads files. To concatenate the resulting fastq forward and reverse files, the command was run from the sample folder, and thus, changing directory from scratch to the sample folder was required. The command below was used for fastq file concatenation:

- cat *R1.fastq.gz > sample.fastq.gz
- cat *R2.fastq.gz > sample.fastq.gz

R1 is the forward read and R2 is the reverse reads. The * is a wildcard symbol so anything ending in .fastq.gz will be concatenated.

A schematic representation of the workflow for insertion detection bioinformatic analysis of data from RC-Seq and WGS is presented on figure 2.4.

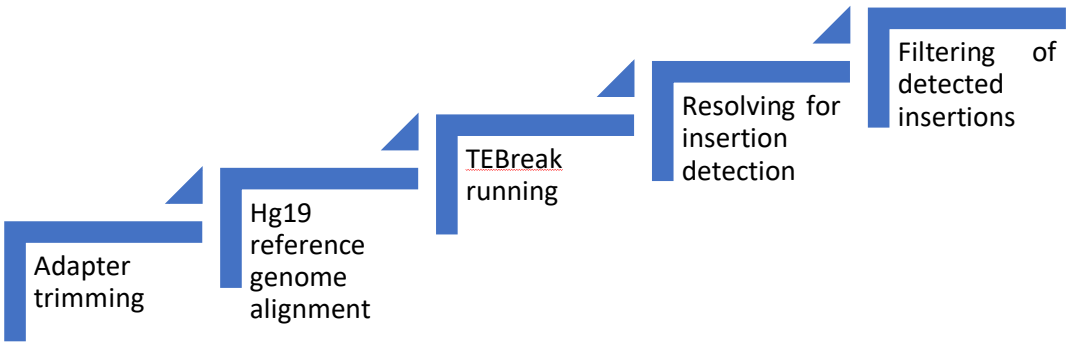


Fig. 2.4 RC/WGS Seq Flow Diagram. The step by step bioinformatic analysis used for insertion detection in both RC-Seq and WGS.

2.2.14.2 Methods for RC-Seq bioinformatic analysis

Fastqc quality control was carried out to check the number of sequencing reads and their quality. Fastqc was installed on the computer. The concatenated fastq files were downloaded from the server to the computer and run through fastqc.

Trimming of fastq files was then carried out to trim adapter sequences. To submit any script to the server, the command qsub was used. For fastq file trimming:

- qsub trimmomatic_rcseq.sh

A second course of Fastqc quality control was carried out to check for the presence of adapter sequences after trimming. The trimmed fastq files were downloaded from the server to the computer and run through fastqc. Concatenated and trimmed files were then aligned to hg19 reference genome. For fastq file alignment:

- qsub alignment_rcseq.sh

TEBreak, which is a bioinformatic tool for generalised insertion detection (*Alu*, LINE-1 and SVAs), was used for identification of L1 insertions. TEBreak was loaded as an environment into the server. For loading TEBreak as an environment:

- module load TEBreak

For TEBreak detection of L1 insertions:

- qsub rcseq_tebreaknew_2samples.sh

The file resulting from TEBreak analysis was a pickle file. Pickling is powerful algorithm for serializing a Python object structure; and thus, the process whereby a Python object hierarchy is converted into a byte stream. For screening the pickle file:

- qsub rcseq_picklescreen.sh

In certain occasions, when resolving the output from TEBreak, the process was aborted by signal kill due to resource starvation. This normally occurred when the pickle file was too big to process. The size limit for resolving the pickle file varied. So, if the pickle file was too big and thus, signal killed occurred, then it was split:

- qsub splitpickle.sh

The pickle file is then resolved for insertion detection. Here, a LINE-1 library only was used for insertion detection. Two different commands were used depending on whether the pickle file was kept as it was output from TEBreak:

- qsub rcseq_resolve_new.sh

Alternatively, if the pickle file was split;

- qsub rcseq_resolvesplit_new.sh

The output from resolving is a table of insertions that requires filtering,

for polymorphic L1 insertions filtering:

- qsub genfilter_rcseq8_150_0.9_0.95_new.sh

or, for somatic L1 insertions filtering:

- qsub genfilter_rcseq4_150_0.9_0.95_mv2_new.sh

For RC-Seq bioinformatic analysis' detailed scripts see Appendix 4.

2.2.14.3 Methods for WGS bioinformatic analysis

Fastqc quality control, trimming of fastq files, and a second course of Fastqc quality control were carried out as in section 2.2.14.2, but this time, the scrip for trimming of fastq files was:

- qsub trimmomatic38_WGS_DS.sh

Concatenated and trimmed files were then aligned to hg19 reference genome. For fastq file alignment:

- qsub alignment_trimmed_WGS_DS.sh

TEBreak was used for identification of generalised insertions (*Alu*, LINE-1 and SVAs),

for loading TEBreak as an environment:

- module load TEBreak

for TEBreak analysis:

- qsub TEBreak_trimmed_WGS_DS.sh

The file resulting from TEBreak analysis was a pickle file. The size limit of the pickle file for subsequent resolving varied. So, if the pickle file was too big, then it was split:

- qsub splitpickle_trimmed_WGS_DS.sh

The pickle file is then screened and resolved for generalised insertion detection. Here, a TE library including *Alu*, LINE-1 and SVA was used for insertion detection. Two different commands were used depending on whether the pickle file was kept as it was output from TEBreak:

- qsub resolve_trimmed_WGS.sh

Alternatively, if the pickle file was split;

- qsub resolvesplit_trimmed_WGS_DS.sh

The output from resolving is a table of insertions that requires filtering,

for *Alu*, L1 and SVA polymorphic insertions detection:

- qsub generalfilter_trimmed_WGS_DS.sh

For WGS bioinformatic analysis' detailed scripts see Appendix 5.

Chapter 3

Epigenetics of repetitive DNA elements

Chapter 3 Epigenetics of repetitive DNA elements

Chapter 3 addresses the epigenetics of repetitive DNA elements. It is divided upon two subchapters, 3.1 and 3.2. Each subchapter has its own introduction, aim or aims, materials and methods, results and discussion. This brief introductory section covers the communalities between the two subchapters to avoid reiteration.

Positive and negative control of isolated unmethylated and methylated DNA by PCR

The temporal cortex and blood DNA of 9 HA and 3 AD individuals was separated into unmethylated and methylated fractions using the CpG MethylQuest kit (Millipore) and used for DNA methylation analysis in the following chapter sections (3.1 and 3.2). Table 3.1 contains phenotypic information on the samples used for methylation analysis.

Table 3.1. Information on the Dyne Steele cohort samples used for methylation analysis. Those where age changes only for path diagnosis 1 were considered healthily cognitively aged and those where path diagnosis 1 is Alzheimer's disease were considered as cases.

Case	Code	Cohort ID	Age at death	Sex	Path diagnosis 1	Path diagnosis 2
09/24	HAs1	11427	78	M	Age changes only	mild SVD
09/26	HAs2	22110	84	M	Age changes only	mild SVD
09/31	HAs3	11508	94	F	Age changes only	mild to moderate SVD
11/06	HAs4	20935	91	F	Age changes only	mild SVD
11/22	HAs6	20088	89	F	Age changes only	
14/04	HAs8	21092	89	F	Age changes only	
14/46	HAs9	11052	94	F	Age changes only	mild SVD
15/01	HAs10	12504	90	M	Age changes only	
15/28	HAs11	22708	91	F	Age changes only	Cerebral infarction
Case	**Code**	**Cohort ID**	**Age at death**	**Sex**	**Path diagnosis 1**	**Path diagnosis 2**
15/11	ADs2	10640	104	F	Alzheimer's disease	Secondary TDP43, Limited SVD
16/03	ADs3	11233	91	M	Alzheimer's disease	Severe CAA
16/13	ADs5	21596	91	M	Alzheimer's disease	

As described in section 2.2.10, a positive and negative control of isolated DNA for subsequent PCR amplification were required. *SNRPN* is an imprinted gene and thus, when targeting the *SNRPN* CGI, a PCR amplicon with the same intensity is expected in both the unmethylated and methylated DNA fractions as one allele is unmethylated and the other is methylated (Figure 3.1A). Similarly, *COX2* CGI is unmethylated in HeLa, and hence, a PCR amplicon is only expected in the unmethylated fraction (Figure 3.1B). In order to analyse the results from the control reactions, we calculated the percentage of PCR product in the methylated fraction of DNA to the unmethylated fraction of DNA for *SNRPN* PCR amplicons, which is estimated to give a value of 0.5 because of the expected 1:1 ratio (Figure 3.1C). Our ratio for temporal cortex is 0.36 and for blood is 0.42. A reason for the observed deviating *SNRPN* ratio could be the polyploidy genotype of Hela cells. In addition, the method of choice for analysis of the data is not entitled to be quantitative, and thus, this could be an additional reason for the observed deviation.

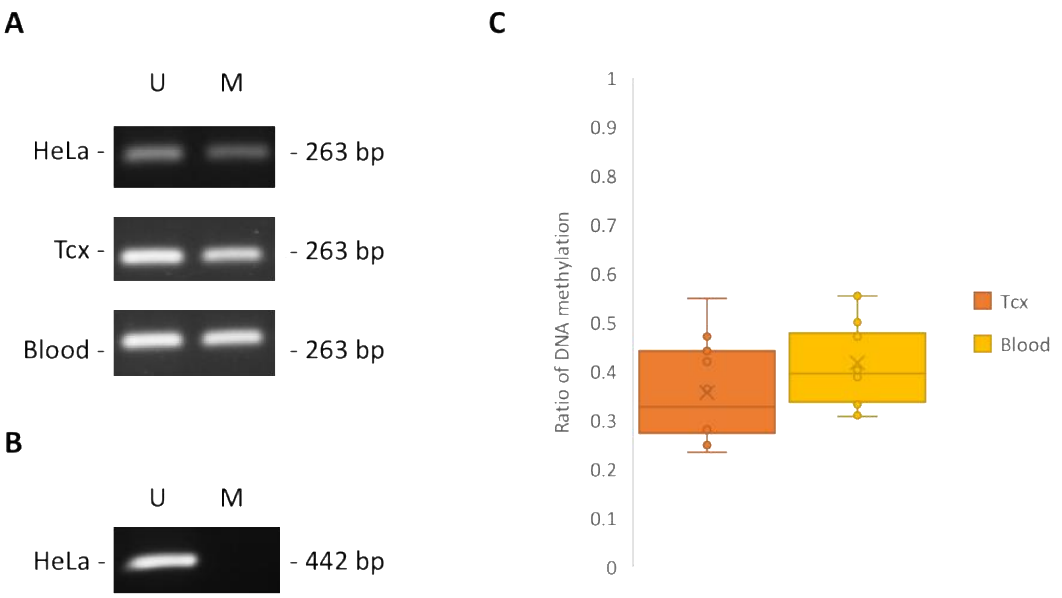


Fig. 3.1. Analysis of methylation status of *SNRPN* CGI and COX2 CGI in unmethylated and methylated fractions of DNA samples. A. Example gel images of PCR product of *SNRPN* CGI as a positive control in unmethylated and methylated fractions of HeLa control DNA provided with the kit and DNA from temporal cortex and blood. **B.** Example gel image of PCR product of *COX2* CGI as a negative control (*COX2* CGI is always unmethylated in HeLa DNA) in unmethylated and methylated fractions of HeLa control DNA. **C.** Box plot representation of the methylation status of the CGI of the imprinted *SNRPN* locus by comparing the band intensity of the PCR product in the unmethylated and methylated fractions of DNA using Image J. The *SNRPN* is an imprinted locus; therefore, one allele is methylated whereas the other is unmethylated. Hence, the ratio is an estimated value of 0.5. Average methylation status of *SNRPN* locus – Tcx – 0.36, blood – 0.42. Tcx – n=12 (HA=9, AD=3), blood – n=12 (HA=9, AD=3). Tcx – temporal cortex, U – unmethylated fraction of DNA, M – methylated fraction of DNA.

Chapter 3.1

The implications for the

MIR941/VNTR locus in ageing

3.1. The implications for the *MIR941*/VNTR locus in ageing

3.1.1 Introduction

Differences in gene expression are thought to be one of the major causes of phenotypic variation across species, including human-specific features such as language, tool-making and lifespan [169]. Several regulatory elements affect gene expression. In this chapter, VNTRs and epigenetic modifications as regulators of gene expression are addressed.

An element shown to be a key regulator of gene expression is the polymorphic domain termed VNTRs, which vary in sequence and copy number of repeats. Polymorphisms are associated with genetic risk for a particular disease and often correlated with differential gene regulation [39, 48, 170]. As an example, the VNTR in the internal promoter of the schizophrenia candidate gene *MIR137*, has been shown to have regulatory properties *in vitro*, where the copy number of the VNTR supports differential reporter gene expression [11]. In addition, monoamine oxidase A (*MAOA*) VNTR genotype has also been identified as a major regulator of the levels of MAOA in the brain *in vitro* [48] and copy number is risk for numerous mental health conditions. MicroRNAs (miRNAs) are single-stranded short RNA transcripts (20 to 24 nucleotides in length) responsible for post-transcriptional regulation of messenger RNA (mRNA) [169, 171, 172]. These play a major role in gene expression regulation by means of targeting multiple mRNAs and translationally repressing or initiating mRNA degradation [171]. *MIR941* is a miRNA, which is embedded within a VNTR, with regulatory effects in the brain affecting genes involved in neurotransmitter signalling, including its host gene *DNAJC5* [169]. *DNAJC5* itself encodes Cysteine String Protein

Alpha (CSPα), a chaperone protein that facilitates correct folding of client proteins and is involved in neuroprotection, neurotransmitter release and synaptic survival [173]. *DNAJC5* dysregulation is linked to a variety of age-related neurodegenerative disorders such as Parkinson's and Alzheimer's disease [174, 175]; and, mutations within *DNAJC5* cause adult-onset neuronal ceroid lipofuscinosis, (ANCL), a neurodegenerative disorder involving accumulation of lipopigments within multiple organs including the brain [166]. In ageing, expression of *DNAJC5* is downregulated [176]. Previous studies have shown that intronic miRNAs are usually transcribed along with their host genes [177], and hence *MIR941* may also be downregulated in ageing.

An extra layer imposed on primary DNA sequence to regulating gene expression are epigenetic modifications of the genome. As an example, DNA methylation within genes is important for gene expression regulation in a tissue- and cell-specific manner [178]. In this context, Shumay *et al.* found that the *MAOA* promoter exhibits a variable DNA methylation pattern in white blood cells, which was correlated with MAOA enzymatic activity in the brain [178].

Using UCSC genome browser, we observe that *MIR941* resides within a polymorphic expanse of tandem repeats, which is also a CGI, located within the first intron of the neuroprotective gene *DNAJC5* [179]. In the *MIR941*/VNTR locus, the VNTR is variable in copy number. Gianfrancesco 2018 previously identified 4 alleles of the VNTR, which differ in copy number of a 56 bp repeat element and vary in frequency in an schizophrenia cohort and matched controls [180]. In this chapter, the VNTR polymorphism in an elderly cohort was further addressed. Furthermore, the

methylation status of the VNTR to address potential regulatory properties was studied. Finally, in order to investigate the potential functional impacts of the copy number of the VNTR on gene expression reporter gene assays in a neuroblastoma cell line were used. It is therefore clear that addressing the intricate relationship of a polymorphic repeat embedding a miRNA that targets a neuroprotective gene with the DNA methylation pattern and expression of the repeat itself or with phenotypic characteristics is important to deepen our understanding of regulatory elements of gene expression in ageing.

3.1.2 Aim

To address whether any genetic or epigenetic variation at *MIR941*/VNTR locus correlated with age using:

- The UCSC Genome Browser to analyse CpG content, SNPs data and conservation across species at the *MIR941*/VNTR locus
- PCR amplification:
  - To analyse the genetic distribution of the polymorphic VNTR in an elderly cohort
  - To compare the genetic distribution from the elderly cohort with:
    - previous data on schizophrenia and matched controls from Gianfrancesco 2018 [180]
    - and, with both a young (age range 20-59) and a supercentenarian (age range 98-108) group from the Georgia cohort
- CpG sites methylation pull-down to analyse the DNA methylation status of the VNTR in temporal cortex and blood DNA
- Reporter gene constructs to assess the regulatory effect of VNTR copy number on expression of the VNTR

3.1.3 Methods

3.1.3.1 Bioinformatic analysis of the *MIR941*/VNTR locus

Bioinformatic analysis of the *MIR941*/VNTR locus was performed using the UCSC Genome Browser (https://genome.ucsc.edu). Analysis was carried out both on the 2009 GRCh37/hg19 and the 2013 GRCh38/hg38 genome build which contained distinct data sets.

3.1.3.2 Genotyping of the VNTR at the *MIR941*/VNTR locus

The VNTR at the *MIR941*/VNTR locus was genotyped in 281 people from the Dyne Steele cohort described in section 2.1.2.1 using 20 ng of blood DNA as template. GoTaq Hot Start Polymerase (Promega) was used for amplification with reagents and PCR cycling conditions outlined in section 2.2.8.1. PCR products were run on a 1.2% agarose gel as outlined in section 2.2.5; these were ran at 100 V for 90 minutes. Images were taken using a trans-illuminator, while running a 100 bp ladder in parallel to the DNA fragments to determine their size. Furthermore, QIAxcel was set up and run as outlined in section 2.2.6. AM320 method at an injection time of 20 seconds was selected to determine the size of the expected fragments. A 15 bp to 3 kb alignment marker and a 100 bp to 2.5 kb size marker were chosen to measure the DNA fragments.

3.1.3.3 Analysis of the methylation status of the VNTR at the *MIR941*/VNTR locus by PCR amplification

As the expected size of the VNTR at the *MIR941*/VNTR target region was too large (800 bp to 1 kb) for reproducible PCR amplification on sonicated DNA (<500 bp), the flanking regions [flank 1 (5') and flank 2 (3')] of the *MIR941*/VNTR locus (Supplementary figure 3.1.1) were analysed. PCR amplification was carried out in 12

people from the Dyne Steele cohort (9 HA and 3 AD) as a proxy for analysis of the methylation status of the locus. PCR amplification was performed using the unmethylated and methylated fractions of temporal cortex and matched blood DNA from the CpG sites pull-down as a template. CpG sites pull-down was performed as described in section 2.2.10. GoTaq Hot Start Polymerase (Promega) was used for amplification with reagents outlined in Table 3.1.1. Primers (Sigma) designed to target each of the flanking sites selected for analysis (Table 3.1.2) were used for PCR. A thermocycler set to 95 ˚C for 2 mins; 95 ˚C for 30 secs, 60 ˚C for 30 secs and 72 ˚C for 30 secs for 35 cycles; 72 ˚C for 5 mins and 4 ˚C infinite hold was used for PCR amplification. 8 µl of PCR product were run on a 1.2% agarose gel using EtBr and separated by electrophoresis at 100 mV for 90 mins. PCR product was visualised using a UV transilluminator. ImageJ software [181] was used to quantify the intensity of the PCR product, and the percentage of methylated DNA calculated by comparing the band intensity of the PCR product in the unmethylated and methylated fractions of DNA in the temporal cortex and blood. It is important to note that this is endpoint PCR (plateau stage), and whilst band quantification via ImageJ may not be a precise reflection of the level methylated versus unmethylated DNA, the size of the flanking regions amplicons (309 bp and 181 bp) is considered too big for qPCR, which products need to be ˜120-150 bp, so endpoint PCR and ImageJ were used as the methods of choice for the preliminary study. qPCR offers a more quantitative approach to measuring the level of DNA methylation, and hence primers that would allow for qPCR amplification of smaller amplicons while still being specific for the VNTR flanking regions need to be designed. This proxy also applies for the analysis of hot RC-L1s. A student's t-test was used to determine statistical significance. The

information on the samples used for analysis of the methylation status of the VNTR at the *MIR941*/VNTR locus is on Table 3.1.

Table 3.1.1. GoTaq Hot Start DNA polymerase PCR reaction components for the flanking regions of the VNTR at the MIR941/VNTR locus

Component	Volume (µl)
PCR buffer (5x)	5.0
MgCl2 (25 mM)	4.0
dNTPs (10 mM each)	0.5
Forward primer (20 mM)	0.5
Reverse primer (20 mM)	0.5
DNA polymerase (5 u/µl)	0.125
Nuclease free water	12.875-14.375
DNA template (5 ng/µl)	0.5-2.0
Final volume	25 µl

Table 3.1.2. VNTR at the MIR941/VNTR locus flanking regions [flank 1 (5´) and 2 (3´)] primer sequences

Primer name	Primer sequence
VNTR_5'End_Fw1	5'-ACACTGAGATTGCACCTGGA-3'
VNTR _5'End_Rv1	5'-CGTCCTCTCCCCGGACACGT-3'
VNTR _3'End_Fw2	5'-CACTCTGTGCTCTGTGTCTG-3'
VNTR _3'End_Rv2	5'-TATGACCTCGGCTCCTTCAC-3'

3.1.3.4 RT-PCR to analyse *DNAJC5* and *MIR941*/VNTR expression in cell lines

RT-PCR amplification was carried out to analyse *DNAJC5* and *MIR941*/VNTR expression using SH-SY5Y, SK-N-AS and HAP-1 cell line cDNA. GoTaq Hot Start Polymerase (Promega) was used for amplification with reagents and PCR cycling conditions outlined below (Table 3.1.4). Primers (Sigma) designed to target exon 1 of *DNAJC5* (*DNAJC5*_L_cDNA Fw) and *MIR941*/VNTR (*DNAJC5*_S_cDNA Fw), respectively; and, exon 2 (*DNAJC5*_cDNA Rv), common to both (Table 3.1.5) were used to determine the presence/absence of *DNAJC5* and VNTR transcripts. ACTB and GAPDH primers [182] were used as a control for endogenous gene expression. A thermocycler set to 95 ˚C for 2 mins; 95 ˚C for 30 secs, 64 ˚C for 30 secs and 72 ˚C for 30 secs for 30 cycles; 72 ˚C for 5 mins and 4 ˚C infinite hold was used for PCR amplification. The PCR products were run on an agarose gel as outlined in section 2.2.5.

Table 3.1.4. GoTaq Hot Start DNA polymerase reaction components for *DNAJC5* and *MIR941*/VNTR PCR

Component	Volume (µl)
PCR buffer (5x)	5.0
MgCl2 (25 mM)	4.0
dNTPs (10mM each)	0.5
Forward primer (20 mM)	0.5
Reverse primer (20 mM)	0.5
DNA polymerase (5 u/µl)	0.125
Nuclease free water	13.375
DNA template (5 ng/µl)	4.0
Final volume	22

Table 3.1.5. *DNAJC5* and *MIR941*/VNTR expression analysis primer sequences

Primer name	Primer sequence
***DNAJC5*_S_cDNA Fw**	5'-AAGAGCAGGCTCAGGGCTCT-3'
***DNAJC5*_L_cDNA Fw**	5'-AGGCTGAGGAGTGCGTCGGC-3'
***DNAJC5*_cDNA Rv**	5'-GGACGTGGTACAATGACTCCC-3'
***GAPDH* Fw [182]**	5'-TCTCCTCTGACTTCAACAGCGAC-3'
***GAPDH* Rv [182]**	5'-CCCTGTTGCTGTAGCCAAATTC-3'
***ACTB* Fw [182]**	5'-GCCCTGAGGCACTCTTCCA-3'
***ACTB* Rv [182]**	5'-CGGATGTCCACGTCACACTTC-3'

3.1.3.5 Generation of reporter gene constructs for use in the Dual Luciferase Reporter Assay by TA intermediate vector cloning

VNTR alleles were amplified using GoTaq Hot Start DNA polymerase as outlined in section 2.2.7 and ligated into the pCR2.1 intermediate vector as described in section 2.2.7.1. EcoRI was used in the restriction enzyme (RE) digest to determine the orientation of the insert in the pCR2.1 vector. Table 3.1.6 summaries the information for the intermediate vector generated.

The pCR2.1 intermediate vector *MIR941*/VNTR alleles were then used to clone the target fragments into pGL3p in the endogenous orientation using RE sites as described in section 2.2.7.2. The ligation reactions were transformed into competent DH5α E. coli as described in section 2.2.7.3. Table 3.1.7 summaries the information for the construct generated.

For downstream applications, the reporter gene constructs were both mini- and maxi- prepped as outlined in section 2.2.7.4. To analyse reporter gene expression, transient transfection of reporter gene constructs into SH-SY5Y cell line was carried out as specified in section 2.2.7.5.1. The levels of reporter gene were then measured by using the Dual Luciferase Reporter assay as described in section 2.2.7.5.2. A two-tailed student's t-test was carried out to determine whether the fold change of the VNTR constructs in comparison to the control pGL3p constructs was significant.

Table 3.1.6. Plasmids generated using fragments cloned into the pCR2.1 intermediate vector. This table contains information on the intermediate vectors generated for use in downstream cloning of VNTR alleles.

Insert	Primers for amplification of insert	Template DNA for PCR	*RE
VNTR alleles	5'-ACGTGTCCGGGGAGAGGACG-3' 5'-CCCGGTCCGACGCAGGAC-3'	Human genomic DNA	EcoRI

* Restriction enzyme used to determine orientation of insert in the PCR2.1 intermediate vector

Table 3.1.7. The pGL3p reporter gene constructs generated for use in *in vitro* luciferase assays. This table shows information regarding the vector used and the nature of the inserts for generating reporter gene constructs containing different VNTR alleles.

Name	Vector	Orientation	*RE	**RE	***RE	****RE
VNTR alleles	pGL3p	Endogenous	Nhel-HF	Spel-HF and Xbal	KpnI-HF and Sfil	NotI-HF

* Restriction enzyme used to generate sites of insertion in the pGL3p vector

** Restriction enzymes used to generate RE complementary sites in the pCR2.1 intermediate vector

*** Restriction enzyme used to detect presence of insert once ligated into the pGL3p vector

**** Restriction enzyme used to determine orientation of insert once ligated into the pGL3p vector

3.1.4 Results

3.1.4.1 Bioinformatic analysis of the *MIR941*/VNTR locus

The UCSC genome browser was used to analyse the *MIR941*/VNTR locus. Using UCSC genome browser, we can analyse the genome for potential regulatory elements, as well as a wide variety of structural information about these, including, but not limited to SNPs data, the location of CGIs, the repetitiveness of the elements and the conservation across species. Because of the variability between UCSC genome browser hg19 and hg 38 at this particular locus, we used both versions for analysis.

Analysis of *DNAJC5* and the *MIR941*/VNTR locus using the UCSC genome browser (GRCh37/hg19) identified *DNAJC5*, a large transcript of 40.9 kb (Figure 3.1.1A) and, *AK128776*, a shorter transcript of 16.6 kb (Figure 3.1.1B). Although named the shorter transcript, *AK128776* is composed of 6 exons in total, whereas the larger *DNAJC5* transcript has 5 exons. The *MIR941*/VNTR locus is shown to be located within the 5' UTR of the *AK128776* transcript, which overlaps a CGI and the VNTR. Two pre-*MIR941* transcripts (*MIR941*-1 and -3) were identified within the VNTR region. Use of Multiz Alignments tool demonstrated the *MIR941*/VNTR locus is conserved back to chimps, gibbons and rhesus monkeys, whereas no conservation was observed in gorillas, orangutans, mice and zebrafish, which suggests lack of sequencing data over the region for gorillas and orangutans as this is evolutionary not possible. Numerous SNPs, small insertions and deletions from common SNP build 151 were found within the *MIR941*/VNTR locus, with particular clustering in the 5' promoter region of the *AK128776*, confirming this region as highly polymorphic.

In comparison, analysis of the *MIR941*/VNTR locus using the UCSC genome browser (GRCh38/hg38) identified a single *DNAJC5* transcript of 40.9 kb (Figure 3.1.2). *MIR941* is shown to be located within the first intron of *DNAJC5* transcript, still overlapping a CGI and the VNTR. Five instead of two pre-*MIR941* transcripts (*MIR941*-1 to -5) were identified within the VNTR region. Use of Multiz Alignments tool demonstrated the *MIR941*/VNTR locus is conserved back to chimps, orangutans, gibbons and rhesus monkeys, whereas no conservation was observed in gorillas, mice and zebrafish, which suggests still a lack of sequencing data over the region for gorillas as again this is evolutionary not possible. Use of common SNP build 151 demonstrated the same variability pattern across the *MIR941*/VNTR locus.

Not only differences in conservation between hg19 and hg38, but also phylogenetics suggest a lack of sequencing data over the region for gorillas and orangutans for the oldest version of UCSC genome browser (hg19), which appears to have been further attempted in both cases in the most up to date version of UCSC (hg38). In order to test this, a PCR for the target region on gorillas and orangutans gDNA could be performed, as sequencing of repetitive regions of DNA is rather difficult, and hence conservation data from UCSC genome browser may not always be accurate.

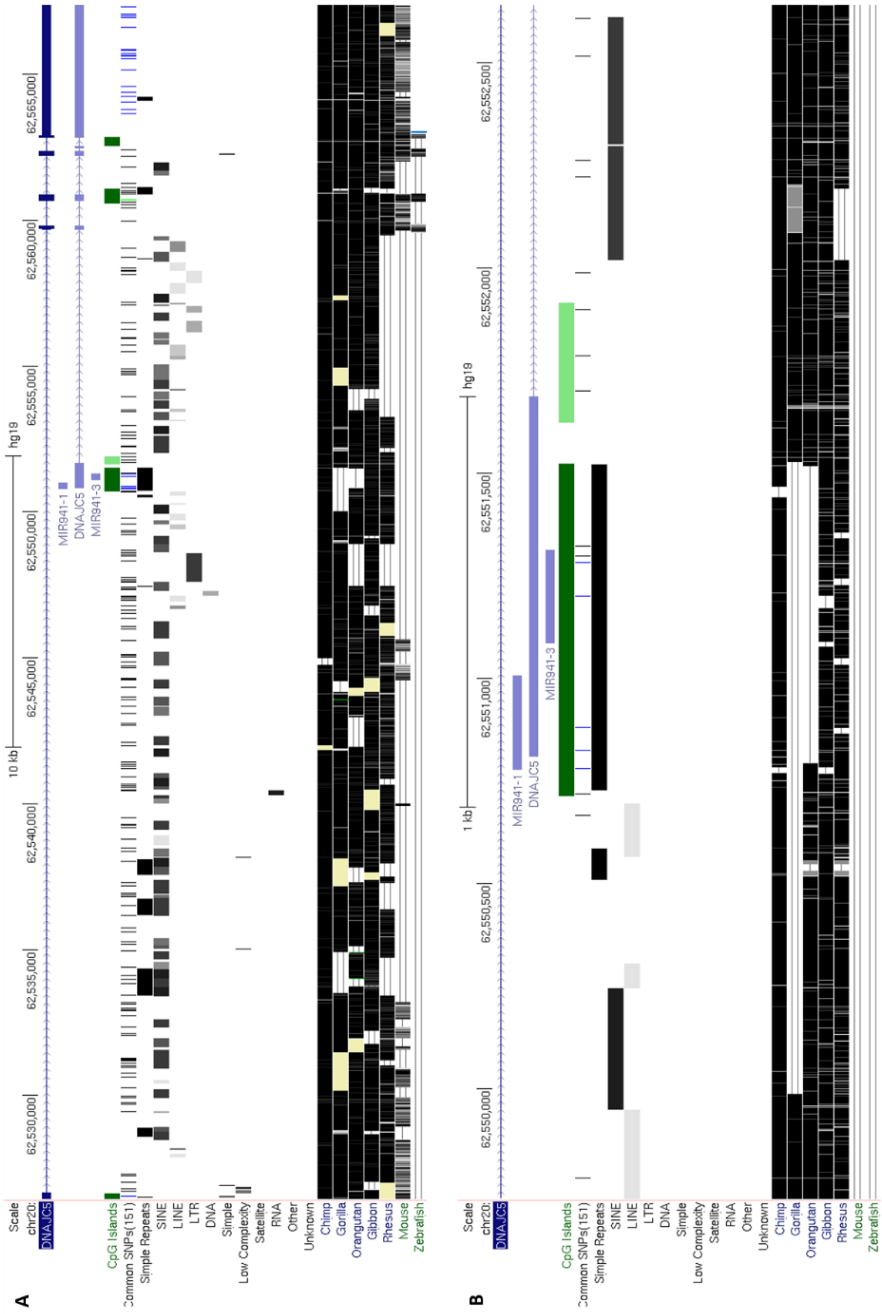


Fig. 3.1.1. *DNAJC5* and *MIR941*/VNTR locus showing *DNAJC5* and *AK128776* transcripts from hg19 UCSC genome browser (chr20:62,526,455-62,567,384). A. Two full-length transcripts, *DNAJC5* and *AK128776*, are shown at the locus, in which both pre-*MIR941*-1 and pre-*MIR941*-3 appear within the 5' UTR of the short *AK128776* transcript. **B.** Zoomed image of the short transcript, *AK128776*, shows that a VNTR and CGI are found to span the 5' UTR. There are numerous SNPs shown within these sequences.

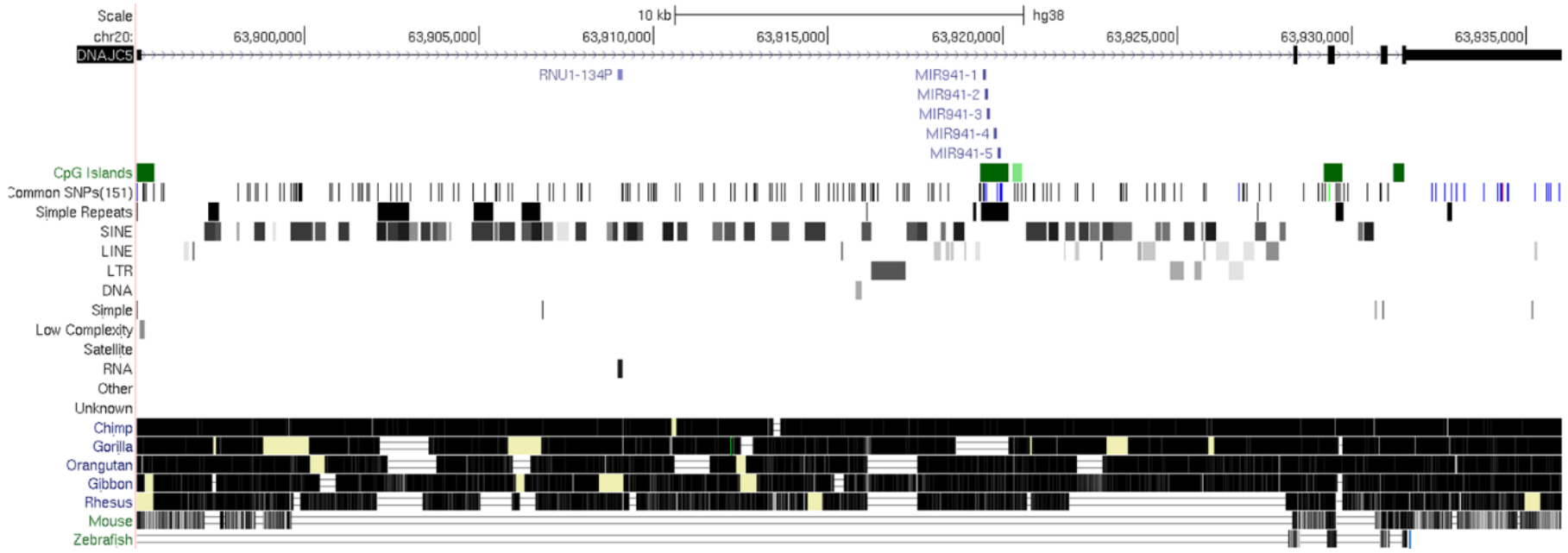


Fig. 3.1.2. *DNAJC5* locus from hg38 UCSC genome browser (chr20:62,526,455-62,567,384). A single *DNAJC5* transcript is shown at the locus, in which pre-*MIR941*-1 to 5 appear. Tandem repeats are found within the first intron spanning a CGI. There are numerous SNPs shown to be within these sequences.

3.1.4.2 There is no association of the VNTR polymorphism to ageing

In order to gain insight into the potential association of the genetic variability at the *MIR941*/VNTR locus to ageing, we analysed the VNTR polymorphism at the locus, [169].

The genetic variation of the VNTR was analysed in the blood DNA of 281 individuals from the Dyne Steele cohort (section 2.1.2.1) to determine if there was an association of repeat unit copy number with ageing. These were 281 elderly individuals, 205 of which were females, and 76 who were males. Primers described in section 2.2.8.1 were used for genotyping. In order to determine whether there was an association with ageing, the genetic variation of the VNTR in the elderly cohort was further compared to: previous data from Gianfrancesco 2018 [180] on controls and matched schizophrenics (control n=340, 152 males and 188 females; schizophrenics n=342, 212 males and 13 females); and, to data for the Georgia supercentenarian (n=100) and young (n=100) cohorts (section 2.1.2.1).

Analysis of the VNTR by PCR amplification identified four common alleles, which differed in copy number of the repeat unit. An example gel image from gel electrophoresis of the four alleles identified in shown in Figure 3.1.3A. Gianfrancesco 2018 [180], previously identified by sequencing that the main repeating unit of the VNTR was 56 bp, with each repeat embedding a 22 bp copy of *MIR941* (Supplementary figure 3.1.2). A schematic representation of the *MIR941*/VNTR region illustrating the number of VNTR repeat units, each of which embeds a copy of *MIR941*, is shown in Figure 3.1.3B. The shortest and rarest allele was found to contain 9 copies of both the VNTR and *MIR941* (9R), and had an allele frequency of 4.1 % in

the Dyne Steele cohort, similarly to 5 % in the control population from Gianfrancesco 2018 [180] and 4 % in the Georgia supercentenarian population (Supplementary figure 3.1.3). The second most common allele, was found to contain 10 repeats of the VNTR and *MIR941* (10R), and had an allele frequency of 16.6 % in the Dyne Steele cohort, as opposed to 6 % in the control population from Gianfrancesco 2018 [180] and 7 % in the Georgia supercentenarian population (Supplementary figure 3.1.3). The most common allele was identified as 13 copies of the VNTR and *MIR941* (13R), and had an allele frequency of 70.7 % in the Dyne Steele cohort, similarly to 78.8 % in the control population from Gianfrancesco 2018 [180] and 78 % in the Georgia supercentenarian population (Supplementary figure 3.1.3). The second rarest allele, with an allele frequency of 8.8 % in the Dyne Steele cohort, similarly to 10.2 % in the control population from Gianfrancesco 2018 [180] and 11 % in the Georgia supercentenarian population (Supplementary figure 3.1.3) had 15 predicted copies of both the VNTR and *MIR941* (15R) (Supplementary figure 3.1.4).

In summary, allele distribution was from most to least common 13R, 10R, 15R and 9R for the Dyne Steele cohort. However, both for the controls from Gianfrancesco 2018 [180] and the Georgia supercentenarian population, it was 13R, 15R, 10R and 9R instead. Our data suggesting the enrichment of the 10R allele as a distinctive feature and thus of interest for the Dyne Steele cohort.

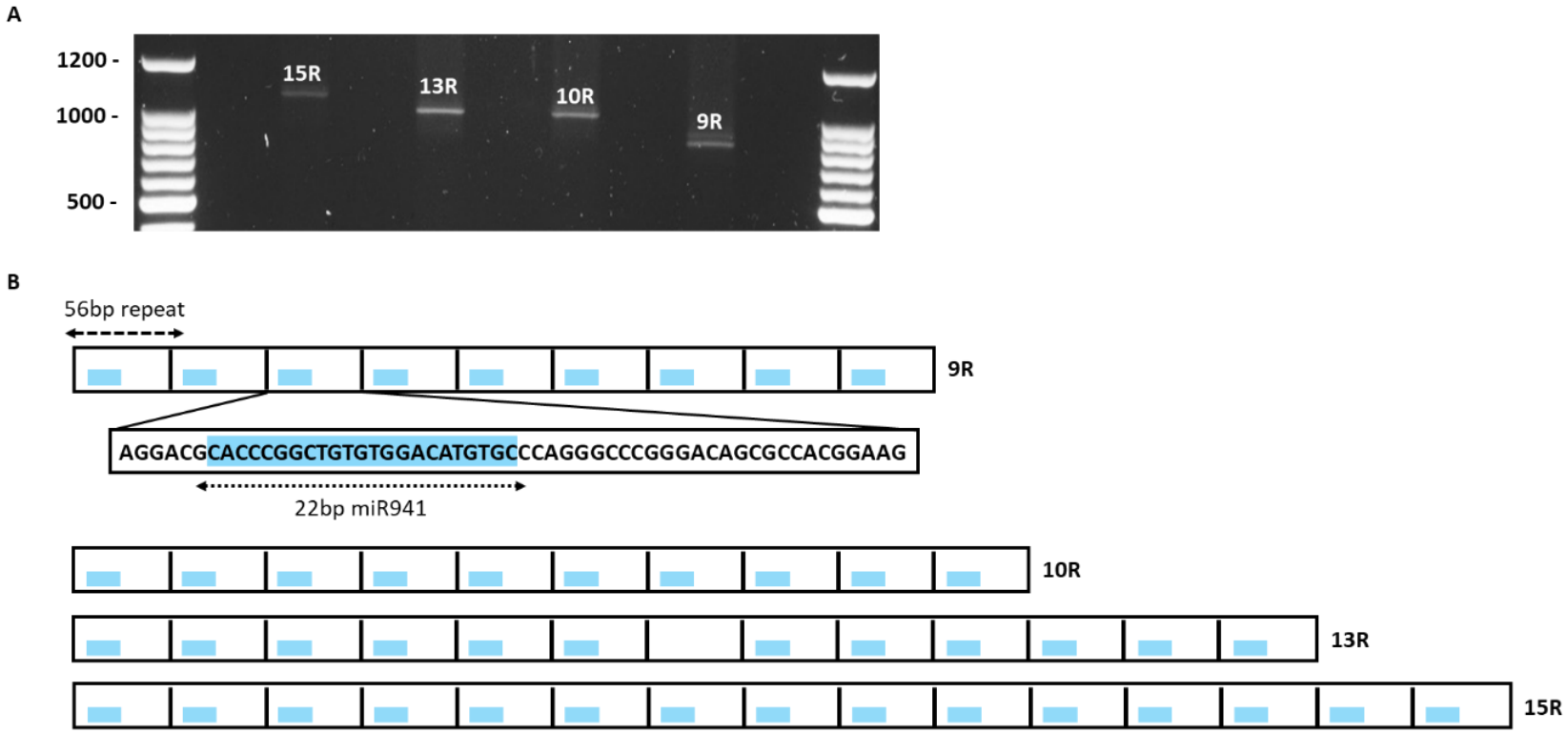


Fig. 3.1.3. Identification of *MIR941*/VNTR alleles in the Dyne Steele cohort. A. PCR amplification across the *MIR941*/VNTR region identified four main alleles of the VNTR, with PCR fragments ranging between approximately 800 bp to over 1 kb in size. **B.** Schematic representation of *MIR941*/VNTR region demonstrates the number of VNTR repeat units, each of which embeds a copy of *MIR941*, of the four different alleles identified by PCR amplification.

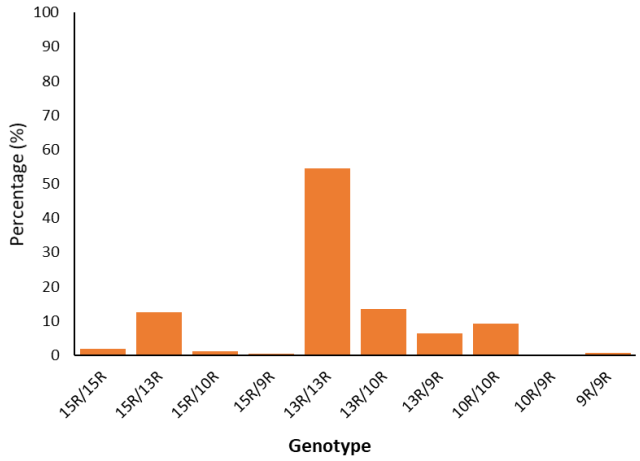
In combination, the four alleles made up 10 possible different genotypes; 9 of which were present in the elderly Dyne Steele cohort. Genotype analysis of the *MIR941*/VNTR across the Dyne Steele cohort (Figure 3.1.4) compared to data previously reported for schizophrenia and controls in Gianfrancesco 2018 [180], and to data from the Georgia cohort presented in the supplementary material of this thesis demonstrated the 10R/10R genotype to be enriched in the Dyne Steele cohort. 26 individuals in the Dyne Steele cohort were found to be homozygous for 10R/10R genotype (9 %), as opposed to 1.2 % in the control population from Gianfrancesco 2018 [180] and 0 % in the Georgia supercentenarian population. Furthermore, one of the two genotypes that were previously reported as present exclusively in the schizophrenia population, 15R/10R (1.46%) was found at 1 % frequency in the Dyne Steele cohort. The 10R/10R genotype, which is found to be enriched in the Dyne Steele cohort compared to the schizophrenia/control and Georgia cohorts, was further analysed. The 26 people from the Dyne Steele cohort that presented the 10R/10R genotype were compared by SPSS v23.0 with the 255 people from the same cohort with all other possible genotypes. In order to decipher whether there is something specific about this subset of the population that presents an enriched genotype, phenotypic information available including ethnic background, mental/physical health, lifestyle, survival and gender associations was addressed. Interesting findings were an association to a more negative mental health as well as lower survival rates of the group that were homozygous 10R, suggesting the 10R allele as potentially deleterious for ageing. Mental health was assessed by three different tests (CMI test (M-R) – psychological and psychiatric, depression test and Cross tabulation: depression (beck 1 and yesavege 1). Data from mental health

analyses showed consistently that people with the 10R/10R genotype presented a tendency towards more negative mental health symptoms. Survival was also assessed by the Kaplan-Meier test, which considers how long people have been part of the study and whether they are dead or alive to date. Survival analysis showed that people without the 10R/10R genotype have better survival rates. Although none of these trends reached significance, this may be due to the small n number used for the analysis. Statistical analysis is included on Appendix 2.

Genotype analysis was further stratified by sex (Figure 3.1.5), which revealed no significant differences in genotype frequency in the Dyne Steele cohort. However, it is important to note that the number of females (n=205) is higher than the number of males (n=76).

Genotype data was stratified by health status (Figure 3.1.6) in a non-overlapping subset of the Dyne Steele cohort where there was pathology data available (28 individuals, 18 HA and 10 AD). This analysis demonstrated observed differences in 15R/13R genotype frequency between healthy aged people and AD patients in the Dyne Steele cohort. However, no risk genotypes were identified. Furthermore, one of the two genotypes that were previously reported as present exclusively in the schizophrenia population, 10R/9R genotype (1.17%) was found at 6 % frequency in the Dyne Steele healthy aged people.

A

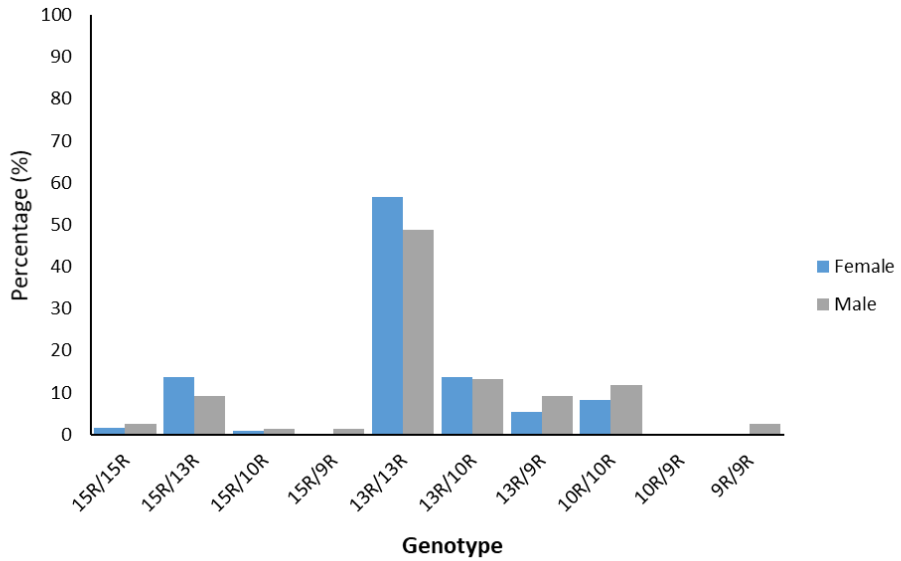


B

Genotype	Dyne Steele	
	Frequency	%
15R/15R	5	2
15R/13R	35	12
15R/10R	3	1
15R/9R	1	0
13R/13R	153	54
13R/10R	38	14
13R/9R	18	6
10R/10R	26	9
10R/9R	0	0
9R/9R	2	1

Fig. 3.1.4. Genotyping data across *MIR941*/VNTR region showed enrichment of a specific genotype (10R/10R) in the Dyne Steele cohort when compared to other groups. A. Graphic representation of the genotypic distribution of the ten possible genotypes arising from combining the four alleles identified by PCR amplification of *MIR941*/VNTR region in blood DNA in the Dyne Steele cohort. 281 people who enrolled with no identification of cognitive decline at the age of 55 of which 205 are females and 76 are males were used for the study. **B.** Table of the genotype frequency across the Dyne Steele cohort.

A

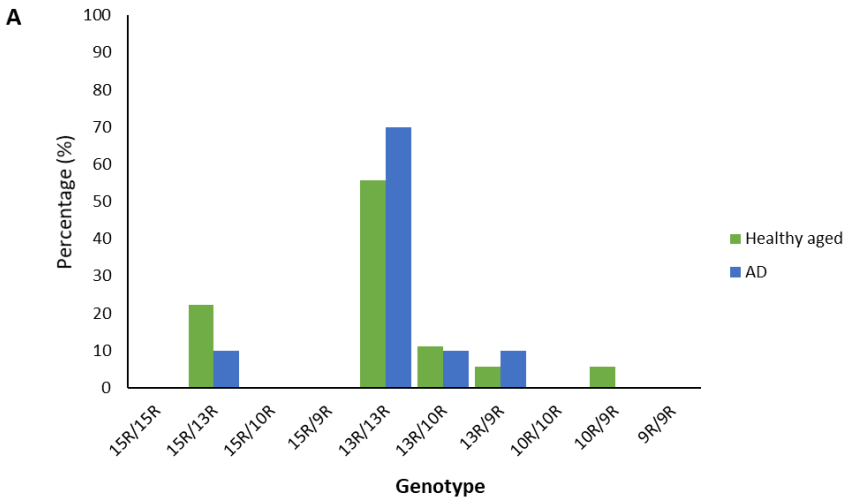


B

Genotype	Dyne Steele female		Dyne Steele male	
	Frequency	%	Frequency	%
15R/15R	3	1	2	3
15R/13R	28	14	7	9
15R/10R	2	1	1	1
15R/9R	0	0	1	1
13R/13R	116	57	37	49
13R/10R	28	14	10	13
13R/9R	11	5	7	9
10R/10R	17	8	9	12
10R/9R	0	0	0	0
9R/9R	0	0	2	3

Fig. 3.1.5. Genotyping data across *MIR941*/VNTR region showed no significant difference between males and females in the Dyne Steele cohort. A. Graphic representation of the genotypic distribution of the ten possible genotypes arising from combining the four alleles identified by PCR amplification of *MIR941*/VNTR region in blood DNA in the Dyne Steele cohort stratified by sex. 281 people who enrolled with no identification of cognitive decline at the age of 55 of which 205 are females and 76 are males were used for the study. **B.** Table of the genotype frequency across the Dyne Steele cohort.

A



B

Genotype	Dyne Steele healthy aged		Dyne Steele Alzheimer's Disease	
	Frequency	%	Frequency	%
15R/15R	0	0	0	0
15R/13R	4	22	1	10
15R/10R	0	0	0	0
15R/9R	0	0	0	0
13R/13R	10	56	7	70
13R/10R	2	11	1	10
13R/9R	1	6	1	10
10R/10R	0	0	0	0
10R/9R	1	6	0	0
9R/9R	0	0	0	0

Fig. 3.1.6. Genotyping data across *MIR941*/VNTR region showed no risk disease-associated genotype when looking at healthy aged people and Alzheimer's disease patients within the Dyne Steele cohort. A. Graphic representation of the genotypic distribution of the ten possible genotypes arising from combining the four alleles identified by PCR amplification of *MIR941*/VNTR region in temporal cortex DNA in the Dyne Steele cohort. The temporal cortex of 28 people, 18 of which were healthy aged and 10 of which developed Alzheimer's disease after enrolment, was used for the study. **B.** Table of the genotype frequency across the Dyne Steele cohort.

In order to test the statistical significance of the observed differences in genotype, a student's t-test was used. This analysis showed no statistically significant differences between any of the groupings. While statistical analysis is useful, it is likely that the small n number would mask any potential statistical significance, so this analysis would need to be extended to a larger cohort before potential statistical significance could be identified.

Therefore, observed differences in the 10R/10R genotype for Dyne Steele cohort; as well as the 15R/10R and 10R/9R genotypes previously identified as being specific to individuals with schizophrenia and now found also in elderly individuals, remain of interest in understanding the potential correlation of the copy number of the VNTR with ageing.

3.1.4.3 *MIR941*/VNTR methylation pattern is tissue-specific and variable across the elderly

The epigenetic landscape of the VNTR at the *MIR941*/VNTR locus in matched (same individual) temporal cortex and blood DNA was addressed to gain a deeper understanding of the potential tissue-specific roles of *MIR941*/VNTR.

To assess the epigenetic landscape of the VNTR, we analysed by PCR amplification the unmethylated and methylated fractions of temporal cortex and matched blood DNA isolated by CpG sites pull-down in 9 HA people and 3 AD patients from the Dyne Steele cohort (Table 3.1). As mentioned above, the analysis of the methylation status of the *MIR941*/VNTR locus was carried out by targeting the flanking regions either side of *MIR941*/VNTR (Supplementary figure 3.1.1) due to the size of the sonicated DNA being too small (<500 bp) to PCR across the entire region consistently.

The analysis of the methylation status of the 5' and 3' ends of *MIR941*/VNTR in all the individuals demonstrated that when comparing the temporal cortex and the blood, the methylation of the locus is significantly higher in the blood than in the temporal cortex (Figure 3.1.7). Furthermore, when looking at the temporal cortex alone, we found that the 3' end is significantly more methylated than the 5' end (Figure 3.1.8), whereas there were no significant differences in the blood. Given the differences between the methylation of *MIR941*/VNTR in the temporal cortex and the blood, we further stratified the methylation data by health status into those that were healthy cognitively aged and AD patients (Supplementary figure 3.1.5). There were no observed differences in the 5' end region neither for temporal cortex (average methylation status of *MIR941* VNTR, HA – 0.12, AD – 0.11), nor for blood (average

methylation status of *MIR941* VNTR, HA – 0.82, AD – 0.84). However, when looking at the methylation status of the 3' end of *MIR941* VNTR, AD patients showed a higher level of methylation in the temporal cortex (average methylation status of *MIR941* VNTR, HA – 0.38, AD – 0.49), but a lower level of methylation in the blood (average methylation status of *DNAJC5* VNTR – HA – 0.94, AD – 0.82). None of the observed differences reached statistical significance using a student´s t-test, but the n number is too small for it to be so; therefore, this analysis should be extended to a larger cohort. In order to further interrogate the data, we attempted correlating the methylation status of the VNTR at the *MIR941*/VNTR locus with genotyping data (Figure 3.1.7A). The methylation pattern was variable across individuals and not correlated with their genotype.

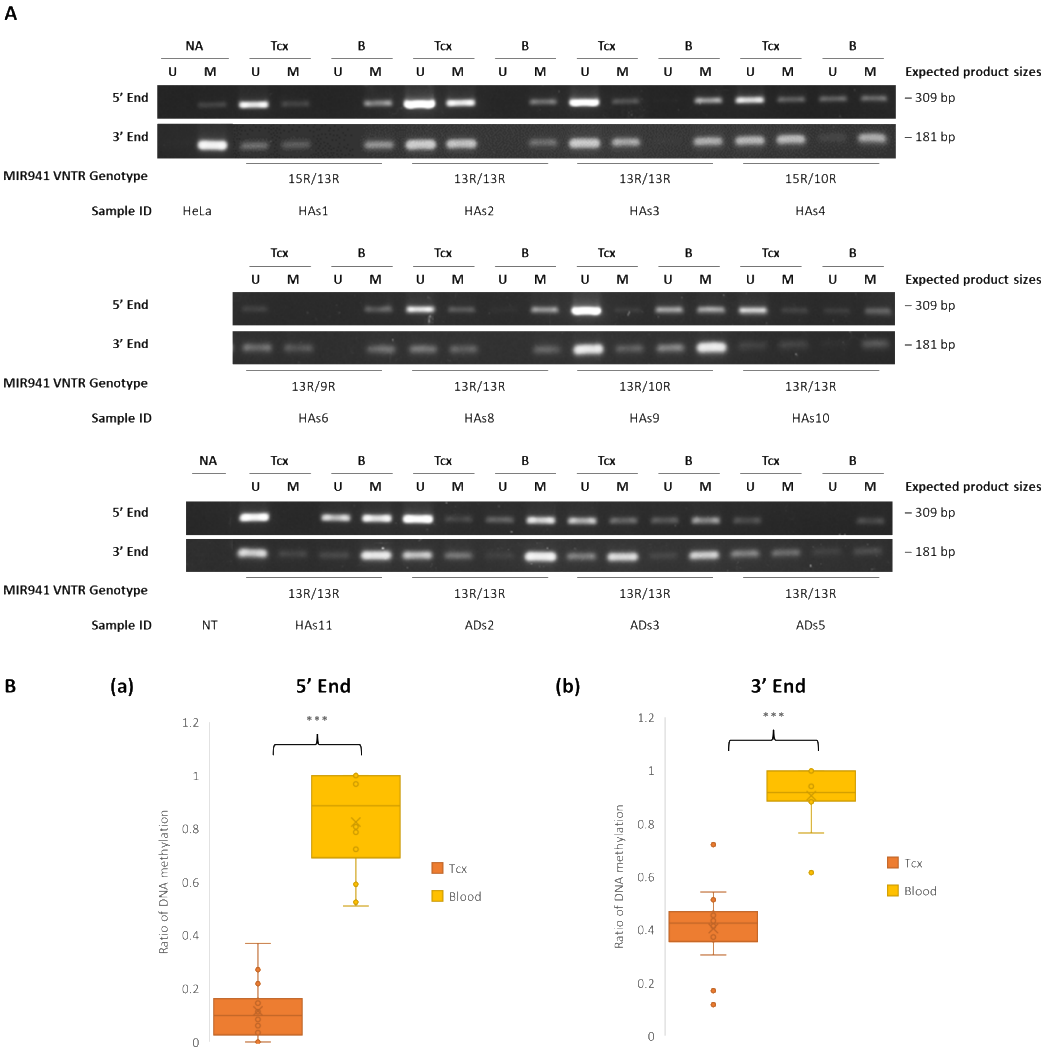


Fig. 3.1.7. Analysis of methylation status of *MIR941*/VNTR region. Methylation status of 5' and 3' end flanking regions of *MIR941*/VNTR locus. **A.** Agarose gel images of PCR product of the 5' and 3' end flanking regions of *MIR941*/VNTR in unmethylated and methylated fractions of HeLa control DNA and DNA from temporal cortex and blood from HA and AD. **B.** (a) The percentage of methylation over the 5' end flanking region of *MIR941*/VNTR is significantly higher in the blood than in the temporal cortex (p-value=2.8E-7). Average methylation status, Tcx – 0.12, blood – 0.83. (b) The percentage of methylation over the 3' end flanking region of *MIR941*/VNTR is significantly higher in the blood than in the temporal cortex (p-value=1.7E-6). Average methylation status, Tcx – 0.40, blood – 0.91 (n=12; HA=9, AD=3). Tcx – temporal cortex, U – unmethylated fraction of DNA, M – methylated fraction of DNA, HA – Healthy aged, AD – Alzheimer's disease, NT – Non-Template.

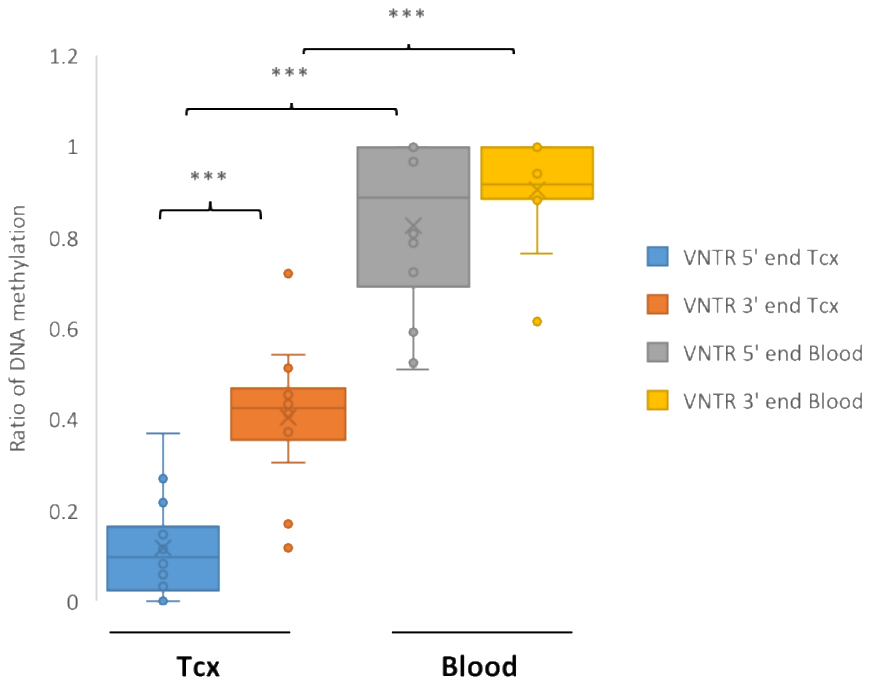


Fig. 3.1.8. Analysis of methylation status of *MIR941*/VNTR region summary. Methylation status of 5' and 3' end flanking regions of *MIR941*/VNTR region. The percentage of methylation over the 5' end flanking region of *MIR941*/VNTR region is significantly higher in the blood than in the temporal cortex (p-value=2.8E-7). Average methylation status, Tcx – 0.12, blood – 0.83. The percentage of methylation over the 3' end flanking region of *MIR941*/VNTR is significantly higher in the blood than in the temporal cortex (p-value=1.7E-6). Average methylation status, Tcx – 0.40, blood – 0.91. The percentage of methylation in the temporal cortex over the 3' end flanking region of *MIR941*/VNTR is significantly higher than in the 5' end flanking region of *MIR941*/VNTR (p-value=8.5E-6). Temporal cortex and blood n=12 (HA=9, AD=3). Tcx – temporal cortex.

3.1.4.4 *MIR941*/VNTR acts as a regulatory domain *in vitro*

Expression and functional analysis of the VNTR component of *MIR941*/VNTR was carried out as a proxy to better understand the role of this element.

Chromatin state and histone modification data from Gianfrancesco 2018 [180] around the transcriptional start sites of the major *DNAJC5* transcript, and the *AK128776* transcript using the HaploReg v4.1 tool suggested promoter activity in all tissues and cell lines tested for *DNAJC5*, and limited regulatory activity in the adult brain for *AK128776*. Hg19 illustrates *MIR941*/VNTR to be located within the 5' UTR region of the shorter transcript of *DNAJC5*, *AK128776* (Figure 3.1.1B). However, when looking at the same region on hg38, the shorter transcript of *DNAJC5* no longer appears (Figure 3.1.2) locating *MIR941*/VNTR simply within the first intron of the long transcript of *DNAJC5*. In order to demonstrate either the presence or absence of the shorter transcript and thus, the potential regulatory effect suggested by Gianfrancesco 2018 [180] data, we performed PCR analysis of SH-SY5Y, HAP-1 and S-K-NAS cell lines cDNA. We used hg19 data to design primers targeting exon 1 of *DNAJC5* transcript and *AK128776*, respectively and exon 2 common to both transcripts. An expected 156 bp band for the long *DNAJC5* transcript as well as a 210 bp band from the *AK128776* transcript confirmed the expression of both transcripts in SK-N-AS and HAP-1 cells. However, no expression of either transcript was observed in SH-SY5Y cell line, but potentially a shorter isoform instead (Figure 3.1.9A). Furthermore, whereas consistent expression was observed across *ACTB*/*GAPDH* control genes, there was an observed change in the level of expression across different cell lines both for the *DNAJC5* and *AK128776* transcripts. This change in

expression seemed to correlate with the VNTR genotype, whereby in SH-SY5Y having two copies of the polymorphic variant (13R and 10R) correlated with no expression of *DNAJC5*, whereas in HAP-1 having a single copy of *MIR941*/VNTR (15R) correlated with an increased level of expression compared to S-K-NAS. Here, our purpose was not to address the specific level of *DNAJC5* and *AK128776* expression, but a presence/absence scenario; however, because of the potential correlation observed, qPCR instead of RT-PCR would be the right approach to quantitatively measure *DNAJC5* and *AK128776* expression. GTEX data confirms ubiquitous expression of *DNAJC5* in the brain (Figure 3.1.10), but there was no GTEX information for *AK128776*.

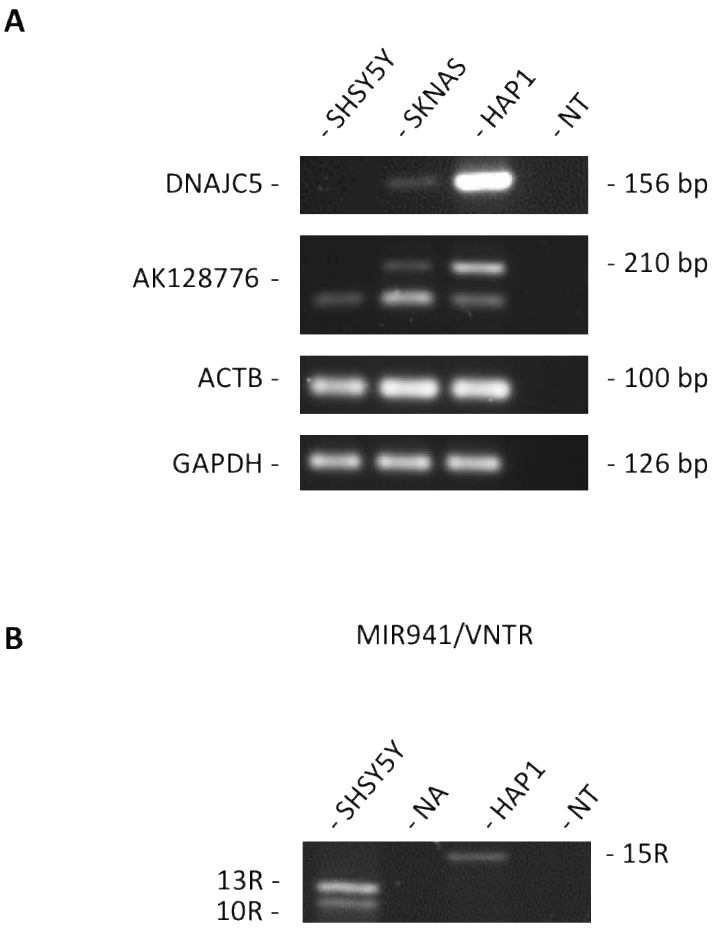


Fig. 3.1.9. Both *DNAJC5* and *AK128776* transcripts are expressed in in S-K-NAS and HAP-1 cell lines A. Expression of *DNAJC5* (156 bp expected band size) and *AK128776* (210 bp expected band size) in cDNA from SH-SY5Y, S-K-NAS and HAP-1. *ACTB* (100 bp expected band size) and *GAPDH* (126 bp expected band size) RT-PCRs are used as control for the level of gene expression. **B.** *MIR941*/VNTR genotype in SH-SY5Y and HAP-1 cell lines. NT – Non-Template.

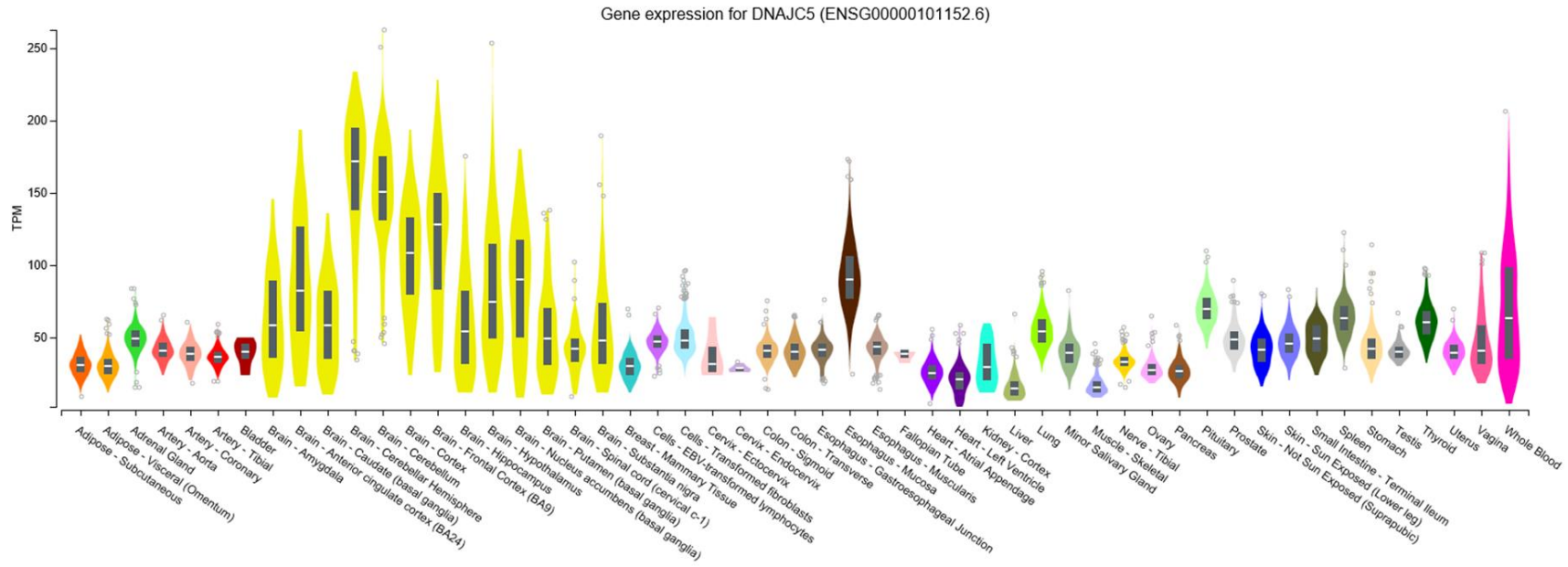


Fig. 3.1.10. *DNAJC5* gene expression data for GTEX demonstrates that *DNAJC5* is ubiquitously expressed in the brain. There is no GTEX data for expression of *AK128776*.

Functional analysis of the VNTR component of *MIR941*/VNTR was then carried out to test its ability to act as a transcriptional regulator *in vitro* using reporter gene constructs. Two variants of the VNTR (10R and 13R) in the endogenous orientation were cloned into pGL3p reporter gene constructs. We were not successful at cloning the other two VNTR variants (9R and 15R) perhaps due to the repetitive nature of the region and the high GC content. However, as we managed to clone both the most common variant (13R) and the variant which was observed as being enriched in the elderly population (10R), this data is still of interest for potential differential regulatory properties of the VNTR locus. The activity of the constructs was measured in the human neuroblastoma cell line SH-SY5Y (Figure 3.1.11). A significant difference in the level of reporter gene expression for both alleles of the VNTR (10R and 13R) was observed compared to the minimal SV40 promoter of the pGL3p vector alone. Both alleles of the VNTR repressed reporter gene expression in SH-SY5Y, suggesting the *MIR941*/VNTR variants as negative regulators in this model. There was not a significant difference in reporter gene activity when the two alleles were compared to each other; suggesting regulation is not dependent on the copy number of the VNTR in this particular model. Though similar experiments were performed using the human neuroblastoma cell line SK-N-AS (data not shown), the data obtained was different. Therefore, this effect is likely cell line specific; and, as such, this region is expected to behave differently in other cell lines and *in vivo*. This is exemplified by the 5-HTT VNTR *in vitro* and *in vivo* [51].

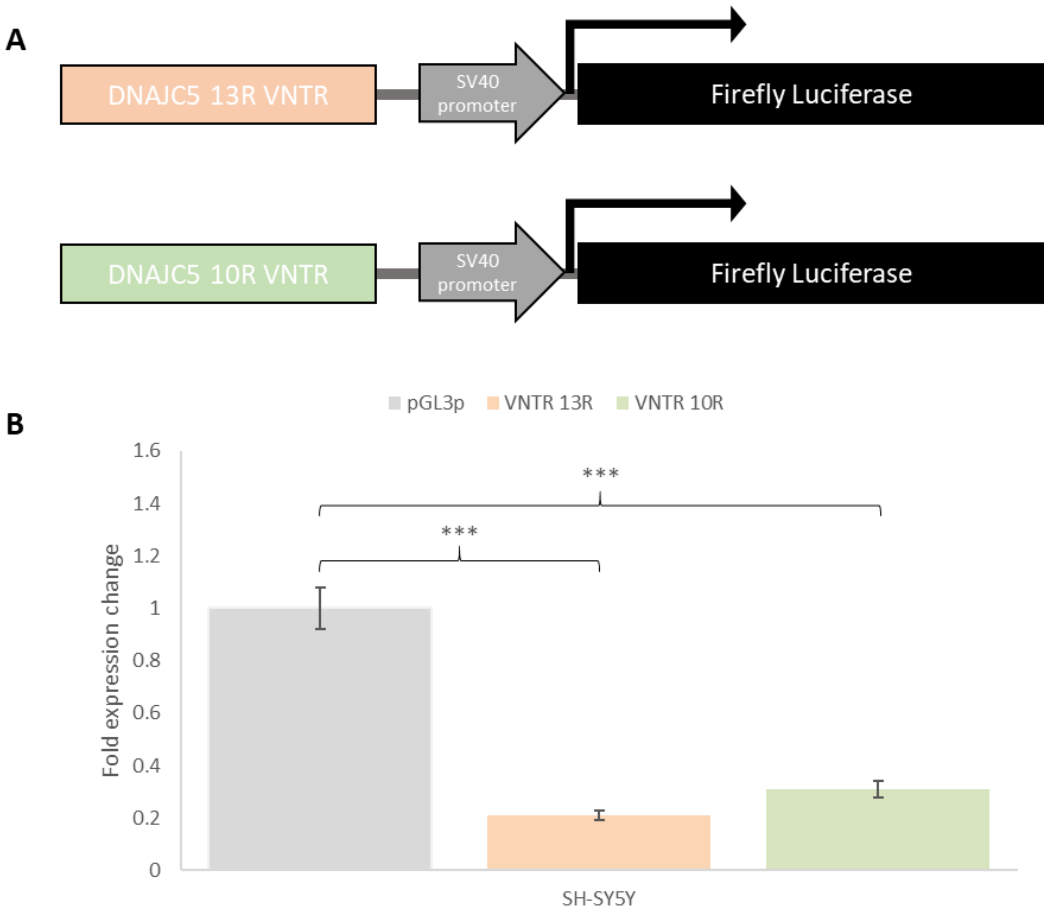


Figure 3.1.11. The VNTR at the *MIR941*/VNTR acts as a functional regulatory domain *in vitro*. A. Schematic representation of pGL3p constructs including 13R and 10R VNTR in endogenous orientation, upstream of a SV40 minimal promoter and firefly luciferase gene. **B.** There is no significant difference between 13R and 10R VNTR in SH-SY5Y cells, both of which are significantly repressive with respect to pGL3p ($p<0.0005$, $n=4$).

3.1.5 Discussion

DNAJC5 at chromosome 20q13.3 is known to be essential for neuroprotection and thus, for healthy functioning of the CNS as it is a major player in neurotransmitter release and synaptic maintenance [164-166]. Differential regulation of gene expression plays a major role in phenotypic variation across individuals [169]. *MIR941* is a hominoid-specific brain-expressed miRNA located within the first intron of *DNAJC5*. In addition, the region embedding *MIR941*, which is a VNTR, has been reported as being variable in copy number in the population [169].

In this project, we have expanded on a previous study on *MIR941*/VNTR in the context of schizophrenia [180]. Our group had previously shown data suggesting *MIR941*/VNTR as a transcriptional regulator in the brain [180], and VNTRs have been shown to play a role in regulating gene expression based on genotype [170, 183]. Gianfrancesco 2018 [180] findings suggested two *MIR941*/VNTR genotypes, 15R/10R and 10R/9R, as a risk factor for schizophrenia. In this context, we found evidence to support that *MIR941*/VNTR genotype is variable across the population (Figure 3.1.4). However, the two genotypes previously found only in schizophrenics were also found in the elderly population in this study (Figures 3.1.5 and 3.1.6). Interestingly, we found only one healthy aged person (6 % genotype frequency, n=28) with the previously schizophrenia associated genotype 10R/9R (1.17 % genotype frequency, n=342), whereas three elderly people (1 % genotype frequency, n=281) presented the 15R/10R genotype previously reported in schizophrenics only. Furthermore, the elderly population from the Dyne Steele cohort compared to previously reported data on schizophrenics and control from Gianfrancesco 2018 [180], and data from a

supercentenarian and a young cohorts (Supplementary figure 3.1.3), was enriched for the 10R/10R genotype. Statistical analysis suggested that people with the 10R/10R genotype presented a tendency towards more negative mental health symptoms and worse survival rates. These two findings correlate with one another as mental health problems are often linked to shorter lifespan [184]. In addition, data from Gianfrancesco 2018 [180] is consistent with the potential correlation of the 10R allele with worse mental health.

Differential DNA methylation in the brain has often been associated with neurodegeneration and psychiatric disorders [32, 185]. The level of DNA methylation in the brain is often inferred from blood or other peripheral tissues DNA methylation levels due to post-mortem brain tissue being the main source of access to measuring DNA methylation levels in the brain; and the effects of post-mortem itself not being clear on DNA methylation changes [185]. With this in mind, nevertheless we found that the methylation of *MIR941*/VNTR is significantly higher in the blood than in the temporal cortex regardless of health status (Figure 3.1.7). This is consistent with the fact that *DNAJC5* is expressed in the brain [169], and suggests shared potential transcription regulation of *DNAJC5* and the VNTR. Data also showed that there were no significant differences in the level of methylation in HA people compared to AD patients when looking at the temporal cortex or the blood alone (Supplementary figure 3.1.5); suggesting the DNA methylation level in the *MIR941*/VNTR locus was not correlated with disease. Given the small n number (n=12) and the variability observed across the population in the methylation pattern, it was difficult to establish a correlation amongst *MIR941*/VNTR genotype, methylation pattern and health

status. Further, PCR serves as a preliminary approach to determine the DNA methylation landscape at the *MIR941*/VNTR locus, but a more quantitative approach such as bisulphite sequencing should be used in future to overcome the limitations of PCR.

In order to take the analysis further, we validated the regulatory activity of *MIR941*/VNTR *in vitro* via reporter gene assays in the SH-SY5Y neuroblastoma cell line. This data demonstrated that two of the four *MIR941*/VNTR alleles acted as repressors of expression in this model (Figure 3.1.11). This is consistent with other studies, which have also demonstrated VNTRs can act as regulatory elements [39, 48].

Chapter 3.2

The methylation of hot RC-L1 elements as a potential biomarker in ageing

3.2. The methylation of hot RC-L1 elements as a potential biomarker in ageing

3.2.1 Introduction

The non-LTR TE subclass L1 are the only autonomous elements that remain mobile in the human genome [73, 87, 186]. Despite mainly being thought of as having deleterious effects [53], L1s play a major role in human genome structure and function [16]. An average human genome comprises over 500,000 (17 %) L1 elements, most of which are unable to mobilise due to 5' truncations, internal rearrangements and mutations of their ORFs regions [186]. Approximately 80 to 100 L1s are thought of as being able to retrotranspose in the average human genome [73]. However, testing of these elements in a cellular retrotransposition assay pointed towards a subset responsible for the majority of the retrotransposition activity to date termed hot RC-L1s [73]. Furthermore, 3' transductions whereby the polyadenylation signal of the L1 element is evaded resulting in carry over of flanking regions [87], have allowed identification of source hot RC-L1 elements that trigger new germline and/or somatic insertions [186, 187]. About 15% of new L1 insertions in the human genome result in 3' transductions [186]. In addition, many of these hot RC-L1s are polymorphic for their presence or absence [73] and their sequence [188, 189]. Therefore, each individual presents a distinct complement of hot RC-L1s.

The genomic instability prompted by TEs is often counteracted by epigenetic control mechanisms such as chromatin-remodelling and CpG dinucleotides DNA methylation [135]. Human retrotransposons are often associated with a CGI [179], most of them subjected to DNA methylation [135]. Epigenetic alterations during ageing such as chromatin-remodelling and a decrease in the level of DNA methylation have often

been correlated with transcriptional activation of retrotransposons and, thus, the age-associated observed increase in TE expression and copy number [135]. Therefore, an increase in the number of hot RC-L1s as well as a decrease in their methylation index may be a causing mechanism of the cellular senescence burden associated with ageing [190].

In this study, the global level of L1 methylation in the temporal cortex and blood genomic DNA of HA and AD individuals was measured using bisulphite pyrosequencing. As there were no biological significant differences in the level of methylation of the L1 family, seven hot RC-L1 elements were shortlisted from the literature based on three different criteria. Firstly, their level of activity in a cellular retrotransposition assay was more than 89% of a known hot RC-L1 named $L1_{RP}$ [73, 186]. Secondly, the number of germline offspring elements (>12) [87]; and, thirdly, the number of somatic insertions (>60) in cancer detected from 3' transductions analysis were both high [187]. Five (2 non-reference and 3 reference) out of the seven L1s were polymorphic for their presence or absence and were genotyped in a small number of HA and AD. Although for each individual L1 tested there was not an association with healthy cognitive ageing we did see an increase in the total number of hot RC-L1s present in AD males. In addition, the methylation status of four (1 non-reference and 3 reference) hot RC-L1s was assessed in the temporal cortex and blood genomic DNA of a small number of HA and AD individuals. Our data suggested a decreased level of methylation of the hot RC-L1s in the temporal cortex of elderly people irrespective of health status. This highlights the need to address the frequency and methylation of this subset of L1s in a much larger cohort to understand the role

they may be playing in healthy cognitive ageing, and also in neurodegenerative diseases such as but not limited to AD.

3.2.2 Aims

- To analyse the global L1 methylation status in temporal cortex and matched blood genomic DNA in HA and AD people in order to determine any potential associations of methylation status of L1 elements with disease and/or tissue
- To assess the location of hot RC-L1 elements using the UCSC genome browser
- To define the presence/absence polymorphism of hot RC-L1 elements in HA and AD people in order to determine any potential associations of RIP frequency with disease
- To analyse the methylation status of hot RC-L1 elements in temporal cortex and matched blood genomic DNA in HA and AD people in order to determine any potential associations of hot RC-L1 elements methylation index with disease and/or tissue

3.2.3 Methods

3.2.3.1 Global L1 methylation

Global L1 methylation analysis was carried out in 16 people from the Dyne Steele cohort previously described in section 2.1.2.2 using 250 ng of temporal cortex and matched blood genomic DNA as starting material for bisulphite treatment and pyrosequencing as described in section 2.2.9. Briefly, temporal cortex and matched blood genomic DNA are bisulphite modified, whereby unmethylated Cs are converted to Us, whereas methylated Cs remain stable (Figure 3.2.1A). A PCR amplification reaction of the 5' promoter GGI L1 consensus sequence where one primer is biotinylated in order track the PCR product for pyrosequencing is then carried out (Figure 3.2.1B). Figure 3.2.1C is an example of an agarose gel of a LINE-1 PCR amplification of bisulphite treated DNA. Finally, purification and isolation of the single strand of DNA that is biotinylated by incubation of the PCR product with streptavidin-coated sepharose beads and subsequent strand separation with NaOH is carried out as exemplified in Figure 3.2.2A. The product is then sequenced by using a pyrosequencing primer (Table 2.10). Figure 3.2.2B is a representative pyrogram of the L1 consensus sequence analysed in our study.

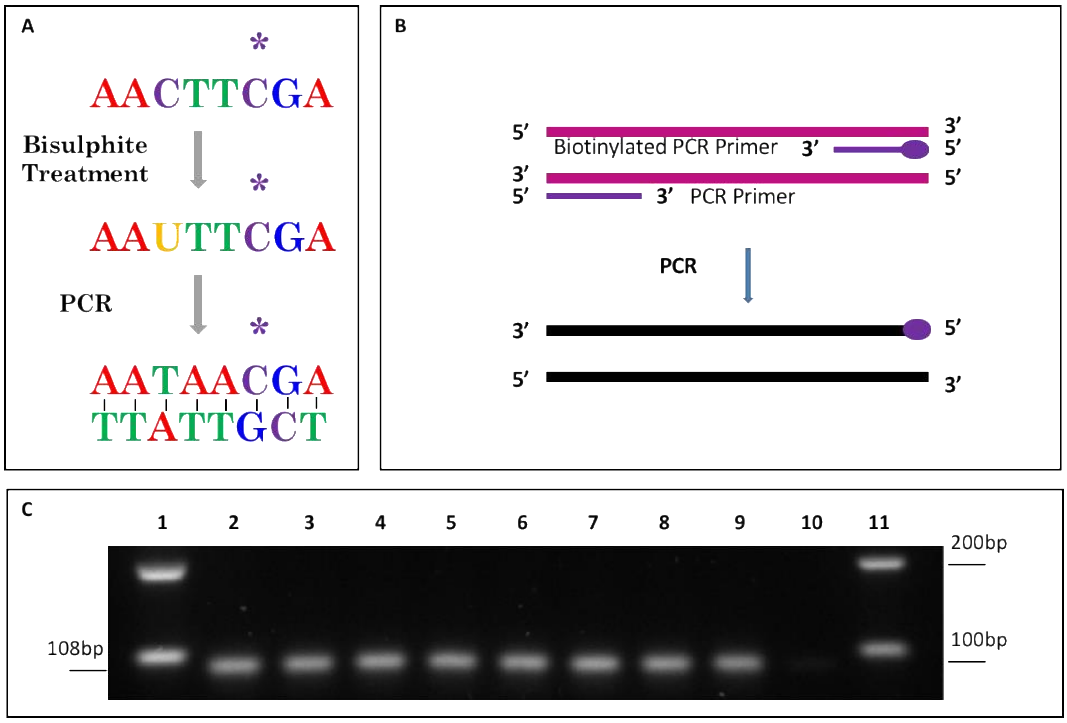


Fig. 3.2.1. Schematic representation of DNA methylation analysis of genomic DNA by bisulphite treatment and pyrosequencing. A. Graphic representation of the effect of bisulphite treatment on genomic DNA whereby unmethylated Cs are converted into Us whereas methylated Cs remain as Cs. **B.** PCR amplification of target region where one primer is biotinylated in order track the PCR product for pyrosequencing. **C.** LINE-1 PCR amplification of bisulphite treated DNA. Lanes 1 and 11 contain 2.0 µl of 100bp ladder. Lanes 2 through to 9 contain 10 % of the bisulphite modified genomic DNA that were amplified by PCR where the expected band was 108 bp. Lane 10 is the negative control.

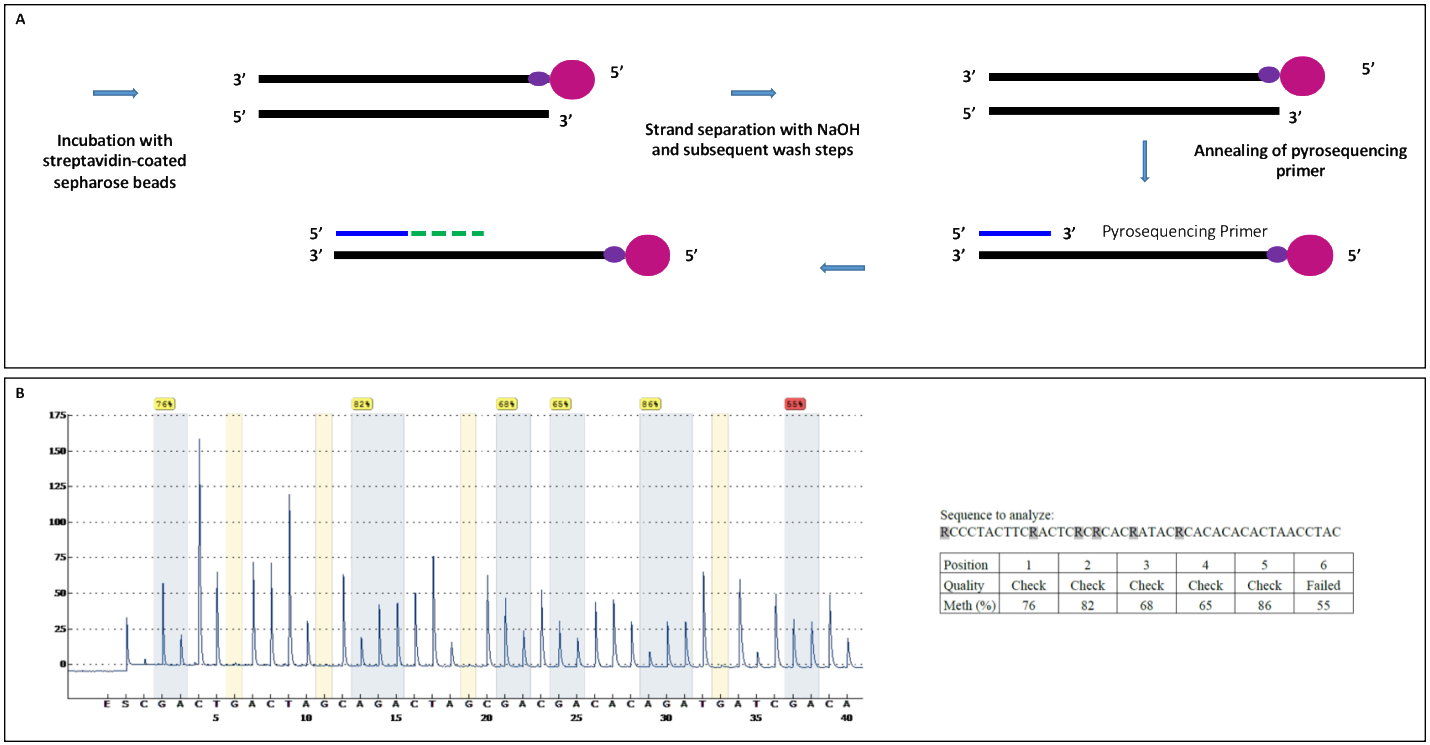


Fig. 3.2.2. Schematic representation of pyrosequencing of PCR amplified bisulphite treated DNA. A. Template preparation to purify and isolate a single strand of DNA by incubation of the PCR product with streptavidin-coated sepharose beads and subsequent strand separation with NaOH. Target PCR product is then sequenced by using a pyrosequencing primer. **B.** Representative pyrogram of the LINE-1 sequence examined in our study. The percentages in boxes indicate the individual CpGs methylation values. The percentage of DNA Methylation (%) is calculated as the average values of the examined CpG dinucleotides.

3.2.3.2 Bioinformatic analysis of hot RC-L1 elements

Hot RC-L1 elements selected for genotyping and for targeted methylation analysis were located using the UCSC genome browser (http://genome.ucsc.edu/index.html). Primers were designed for genotyping and PCR amplification of unmethylated and methylated fractions of DNA as outlined in section 2.2.2.

3.2.3.3 Genotyping of hot RC-L1 elements

Five out of the seven shortlisted hot RC-L1 elements were genotyped in 16 people from the Dyne Steele cohort described in section 2.1.2.2 using 5 ng of temporal cortex DNA as template. GoTaq Hot Start Polymerase (Promega) was used for amplification with reagents and PCR cycling conditions outlined in section 2.2.8.2. Primer sequences used for genotyping are outlined on table 3.2.1. Two hot RC-L1 elements (Ref_chr1_p12 and Ref_chr22_q12.1) were not considered polymorphic in the population and thus, were not genotyped. In order to genotype hot RC-L1 elements, an Empty Site (ES)/5' Junction PCR was carried out (Figure 2.2). The empty site refers to the PCR product when the putative L1 insertion is not present, and thus, involves the flanking regions of the putative insertion was this present. The filled site refers to the PCR product when the putative L1 insertion is present, and thus, involves the L1 element, which when full length and including the flanks can be up to 6/7 kb. Therefore, it is difficult to amplify the filled site, as the empty site is often favoured mainly due to size constrains, though kits exist to overcome this limitation. In our case, 5' junction PCR amplification was used to confirm the presence of the L1 insertion, whereby a primer in the flanking region (the same as for the empty site PCR) and a primer in the consensus sequence of L1 elements is used to amplify the 5'

end of the putative insertion. PCR products were run on a 1.4% agarose gel as outlined in section 2.2.5; these were run at 100 V for 90 minutes. Images were taken using a trans-illuminator, while running a 100 bp ladder in parallel to the DNA fragments. For high throughput analysis of PCR products, QIAxcel was set up and run as outlined in section 2.2.6. AL420 method at an injection time of 20 seconds was selected to determine L1s polymorphism. A 15 bp to 3 kb alignment marker and a 100 bp to 2.5 kb size marker were chosen to measure DNA fragments. Results were statistically analysed as described in section 2.2.8.2.

Table 3.2.1. Hot RC-L1 primer sequences for genotyping

Primer name	Primer sequence	PCR type	Annealing temp (˚C)	PCR (bp)
5’ L1	5’-AACTCCCTGACCCCTTGC-3’	NA	NA	NA
Non-ref_chr2_q24.1_Fw	5’-GGAAGGTTGTAGGGGTCAC-3’	Two PCRs	60	ES: 1099
Non-ref_chr2_q24.1_Rv	5’-CCCCAAAAGCACAGACAACA-3’			5’ End: 985
Non-ref_chr6_q24.1_Fw	5’-AGTCAGGAGCAGGGGTAAAC-3’	Multiplex	60	ES: 246
Non-ref_chr6_q24.1_Rv	5’-CAGGCAAAGTTGTAGTAGCGA-3’			5’ End: 392
Ref_chrX_p22.2_Fw	5’-GGCTAACTGTGGGAGAGGAA-3’	Multiplex	64	ES: 194
Ref_chrX_p22.2_RV	5’-TTGGCCTTTTGTGACACTGG-3’			5' End: 398
Ref_chr6_p22.3_Fw	5’-ACCTGGCCTTCTCTCATTCT-3’	Two PCRs	64	ES: 230
Ref_chr6_p22.3Rv	5’-TTCCCCGCAAGCTCTCTCTTTA-3’		60	5’ End: 341
Ref_chr8_q24.22_Fw	5’-ATCCTTGCACCGATTCCTCA-3’	Multiplex	60	ES: 600
Ref_chr8_q24.22_Rv	5’-GCTCACAATCCCAAACTGCA-3’			5' End: 670

* ES – Empty site PCR. 5’ End – 5’ End of L1 element PCR

3.2.3.4 Analysis of the methylation status of active L1 elements by PCR amplification

Four out of the seven hot RC-L1s were chosen for analysis of their methylation status. The size of the resulting PCR product was required to be <500 bp as the template DNA from CpG sites pull-down was sheared to the above-mentioned size; and therefore, Non-ref_chr2_q24.1, Ref_chrX_p22.2 and Ref_chr8_q24.22 were not analysed for their methylation status as the PCR product was too big for reproducible PCR amplification across the elements. Furthermore, two hot RC-L1 elements analysed for methylation status (Ref_chr1_p12 and Ref_chr22_q12.1) were not genotyped as these were reference L1s and there was no evidence that they were polymorphic. PCR amplification was performed using unmethylated and methylated DNA isolated with the CpG MethylQuest kit. CpG sites pull-down was performed as described in section 2.2.10. The LINE-1 PCR reaction was performed using GoTaq Hot Start DNA polymerase (Table 3.2.2 and table 3.2.3) and using primers (Sigma) designed to target each of the hot RC-L1 elements selected for analysis (Table 3.2.4) as described in section 2.2.8.2. A thermocycler set to 95 ˚C for 2 mins; 95 ˚C for 30 secs, 60/64 ˚C for 30 secs and 72 ˚C for 1 min per kb for 35 cycles; 72 ˚C for 5 mins and 4 ˚C infinite hold was used for PCR amplification. 16 µl of PCR product were run on a 1.2% agarose gel using EtBr and separated by electrophoresis at 100 mV for 90 mins. PCR product was visualised using a UV transilluminator while running a 100 bp ladder. ImageJ software was used to quantify the intensity of the PCR product, and the percentage of methylated DNA calculated by comparing the band intensity of the PCR product in the unmethylated and methylated fractions of DNA in the temporal cortex and blood. A student's t-test was used to determine statistical significance.

The complete list and information of the samples used for analysis of the methylation status of hot RC-L1 elements is on table 3.1.

Table 3.2.2. GoTaq Hot Start DNA polymerase single PCR reaction components

Component	Volume (µl)
PCR buffer (5x)	5.0
MgCl2 (25 mM)	4.0
dNTPs (10 mM each)	0.5
Forward primer (20 mM)	0.5
Reverse primer (20 mM)	0.5
DNA polymerase (5 u/µl)	0.125
Nuclease free water	12.875-14.375
DNA template (5 ng/µl)	0.5-2.0
Final volume	25 µl

Table 3.2.3. GoTaq Hot Start DNA polymerase multiplex PCR reaction components

Component	Volume (µl)
PCR buffer (5x)	5.0
MgCl2 (25 mM)	4.0
dNTPs (10 mM each)	0.5
Forward primer (20 mM)	0.5
Reverse primer (20 mM)	0.25
5' L1 primer (20 mM)	0.25
DNA polymerase (5 u/µl)	0.125
Nuclease free water	12.875-14.375
DNA template (5 ng/µl)	0.5-2.0
Final volume	**25 µl**

Table 3.2.4. Hot RC-L1s primer sequences for measuring the methylation level

Primer name	Primer sequence	PCR type	Annealing temp (˚C)	PCR (bp)
5’ L1	5’-AACTCCCTGACCCCTTGC-3’	NA	NA	NA
Non-ref_chr6_q24.1_Fw	5’-AGTCAGGAGCAGGGGTAAAC-3’	Multiplex	60	ES: 246
Non-ref_chr6_q24.1_Rv	5’-CAGGCAAAGTTGTAGTAGCGA-3’			5’ End: 392
Ref_chr1_p12_Fw	5’-TGTGACAGTTGAGGACGTGA-3’	Single	60	5’ End: 440
Ref_chr1_p12_RV	5’-GCTTTGGAATAGGGCTGTCC-3’			
Ref_chr22_q12.1_Fw	5’-GATAAGAAACTTCCACCGGGGC-3’	Single	60	5’ End: 453
Ref_chr22_q12.1_Rv	5’-AACTGGTCACACTTCTGGGA-3’			
Ref_chr6_p22.3_Fw	5’-ACCTGGCCTTCTCTCATTCT-3’	Two PCRs	64	ES: 230
Ref_chr6_p22.3Rv	5’-TTCCCCGCAAGCTCTCTCTTTA-3’		60	5’ End: 341

* ES – Empty site PCR. 5’ End – 5’ End of L1 element PCR.

3.2.4 Results

3.2.4.1 There is no biological association of the level of global L1 methylation and healthy cognitive ageing

Previous studies have identified global L1 DNA hypomethylation as an epigenetic biomarker correlated with genomic instability in ageing [135, 137, 191] and diseases such as non-small cell lung cancer [168] amongst others [192-195] when comparing healthy to diseased tissue. In this chapter, using bisulphite modification of DNA and pyrosequencing assay, we measured the methylation levels of 6 CpG dinucleotides located to the 5' promoter CGI of L1s in the temporal cortex and matched blood genomic DNA of 11 HA and AD individuals from the Dyne Steele cohort as a potential biomarker in healthy cognitive ageing.

The methylation index in the temporal cortex is significantly higher (<5 % difference) than in the blood (paired *t*-test, p-value = 3.831E-05, Figure 3.2.3) when HA and AD people are analysed together, suggesting a tissue specific level of global L1 methylation independently of health status. When stratified by health status, the same trend stands whereby methylation index in the temporal cortex is significantly higher than in the blood both in HA (paired *t*-test, p-value = 0.004, Figure 3.2.4) and AD (paired *t*-test, p-value = 0.002, Figure 3.2.4) individuals. However, no significant differences were found in the level of global L1 methylation when comparing the temporal cortex DNA from HA people versus AD patients (HA – 74.00 %, AD – 74.15 %, 2 sample *t*-test assuming equal variances, p-value = 0.520, Figure 3.2.4). Similarly, no significant differences were found in the level of global L1 methylation when the matched blood DNA from HA versus AD individuals was compared (HA – 73.49 %, AD – 72.85 %, 2 sample *t*-test assuming equal variances, p-value = 0.490, Figure 3.2.4).

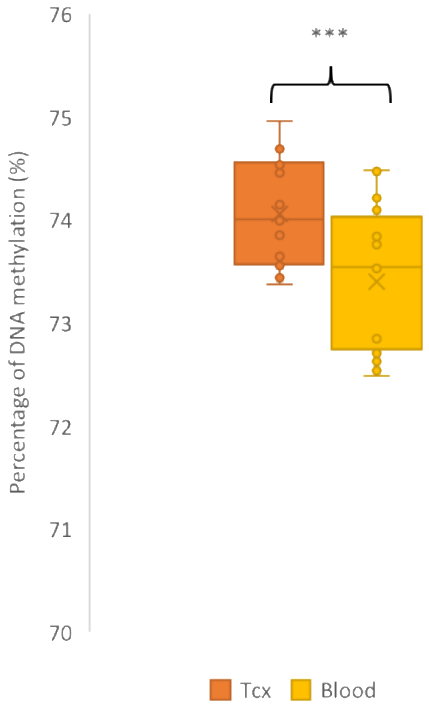


Fig. 3.2.3. There is differential methylation of global LINE-1 in the temporal cortex compared to the blood. Box plot representation of global LINE-1 methylation status of bisulphite treated DNA by pyrosequencing in the temporal cortex versus matched blood DNA of HA people and AD patients together (n=16, p-value = 3.831E-05). *p* value is calculated by 2-tailed paired *t*-test. Tcx – temporal cortex.

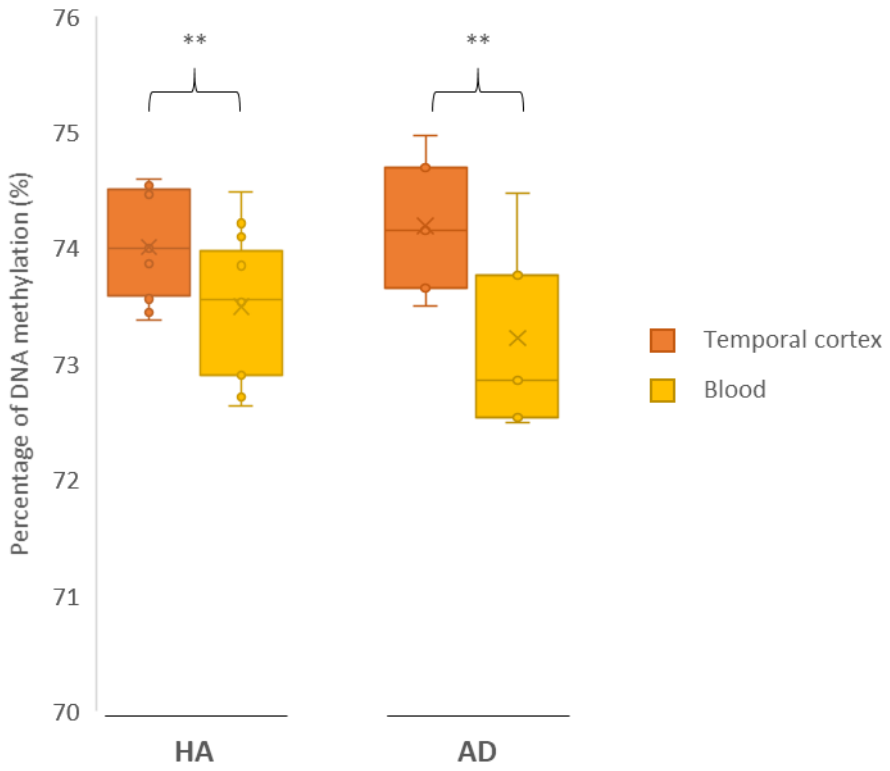


Fig. 3.2.4. There is no differential methylation of global LINE-1 in healthy aged people versus Alzheimer's disease patients. Box plot representation of global LINE-1 methylation status in HA and AD individuals. No significant differences were found in the level of global L1 methylation neither when we compared the temporal cortex DNA from HA people versus AD patients (p-value = 0.520); nor when the matched blood DNA from HA people versus AD patients was compared (p-value = 0.490). *p* values are calculated by two sample *t*-test assuming equal variances. A significant difference was found when we compared the methylation status of global L1 of the temporal cortex and matched blood DNA of HA people (p-value = 0.004); and also, when the methylation status of global L1 of the temporal cortex and matched blood DNA of AD people (p-value = 0.002) were compared. *p* values are calculated by 2-tailed paired *t*-test. HA – Healthy aged (n=11); AD- Alzheimer's disease (n=5).

3.2.4.2 Seven hot RC-L1 elements and their location in the genome

As we could not determine a biological association of the global L1 methylation index with healthy cognitive ageing neither in the temporal cortex nor in the blood, our analysis going forward focussed on seven hot RC-L1 elements. The UCSC genome browser was used to analyse the seven hot RC-L1 elements shortlisted from the literature for either genotyping or methylation analysis. Using UCSC genome browser, we can analyse the genome for potential regulatory elements, as well as a wide variety of information about these, including, but not limited to the location of L1 insertions with regards to genes, such as L1 elements being located within genes or in their vicinity, and the presence of CGIs.

There were three classes of hot RC-L1s. Non-ref L1 RIPs (Non-ref_chr2_q24.1 and Non-ref_chr6_q24.1) are not in the human reference genome and are polymorphic for their presence/absence. Ref L1 RIPs (Ref_chrX_p22.2, Ref_chr6_p22.3 and Ref_chr8_q24.22) are present in the human reference genome and are polymorphic for their presence/absence. Ref L1s (Ref_chr1_p12 and Ref_chr22_q12.1) are present in the human reference genome and there is currently no evidence that they are polymorphic for their presence/absence. As not all of the hot RC-L1 elements were in the reference genome their location was assessed using either the Blat or PCR search tool on the hg19 version of the UCSC genome.

Figure 3.2.5 illustrates the location of the hot RC-L1 elements chosen for analysis and their distance from a CGI if located near any. Below we briefly describe each individual element. Non-ref_chr2_q24.1 (Figure 3.2.5A, Table 3.2.5) is a Non-ref RIP L1 located to the long arm of chromosome 2. It is intergenic, and it is not located

within 100+ kb of a CGI. A 150 % retrotransposition activity of $L1_{RP}$ was attributed to this element [73, 186]; and it gave rise to the highest number of germline offspring elements (41/121) [87]. Ref_chr22_q12.1 (Figure 3.2.5B, Table 3.2.5) is a Ref L1 located to the long arm of chromosome 22. It is intragenic, and it is located within the first intron of Tetratricopeptide Repeat Domain 28 (*TTC28*) main isoform, which is a protein coding gene that plays a role in the formation of the midbody during mitosis and is associated with cleft soft palate [196]. Located 7.8 kb upstream in the same orientation than *TTC28* is the Checkpoint Kinase 2 (*CHEK2*) gene. *CHEK2* encodes a protein which acts as a tumour suppressor by regulating cell division and is associated with prostate cancer and Li-Fraumeni Syndrome 2 [196]. There is a CGI in the promoter region of *TTC28*, 7.8 kb upstream the Ref L1. A 13.8 % retrotransposition activity of $L1_{RP}$ was attributed to this element [73, 186]; and despite the retrotransposition activity of $L1_{RP}$ not reaching >89 %, it gave rise to a high number of somatic insertions in cancer (61) and thus, was shortlisted as hot RC-L1 [187]. Ref_chr1_p12 (Figure 3.2.5C, Table 3.2.5) is a Ref L1 located to the short arm of chromosome 1. It is intergenic, and it is located at the 3' end of T-Box Transcription Factor 15 (*TBX15*), which belongs to the T-box DNA-binding domain family of genes encoding transcription factors that regulate developmental processes and is associated with Cousin syndrome [196]. It is not located within the 100+ kb of a CGI. Non-ref_chr6_q24.1 (Figure 3.2.5D, Table 3.2.5) is a Non-ref RIP L1 located to the long arm of chromosome 6. It is intragenic, and it is located within the third intron of the main isoform of Phosphatase and Actin Regulator 1 (*PHACTR1*), which encodes a protein involved in reorganisation of the actin cytoskeleton [196]. Located 17.8 kb upstream in the opposite orientation to *PHACTR1* is the Cell Migration-Inducing

Protein 23 (*TBC1D7*) gene, which plays a role in cell growth and differentiation and is associated with autosomal recessive megalencephaly [196]. It is not located within 100+ kb a CGI. A 141 % retrotransposition activity of $L1_{RP}$ was attributed to this element [73, 186]; and it gave rise to the highest number of somatic insertions in cancer (75) [187]. Ref_chrX_p22.2 (Figure 3.2.5E, Table 3.2.5) is a Ref RIP L1 located to the short arm of chromosome X. It is intergenic, and it is not located within 100+ kb a CGI. A 132 % retrotransposition activity of $L1_{RP}$ was attributed to this element [73, 186]. Ref_chr6_p22.3 (Figure 3.2.5F, Table 3.2.5) is a Ref RIP L1 located to the short arm of chromosome 6. It is intragenic, and it is located within RHO Family Interacting Cell Polarization Regulator 2 (*FAM65B*/*C6ORF62*) main isoform, which encodes a protein involved in T cell and neutrophil polarization and is associated with deafness [196]. There are three CGIs within 100 kb of the element, located to the promoter regions of *FAM65B*/*C6ORF62* and Geminin DNA Replication Inhibitor (*GMNN*), crucial in cell cycle regulation [196]. A 112.7 % retrotransposition activity of $L1_{RP}$ was attributed to this element [73, 186]. Ref_chr8_q24.22 (Figure 3.2.5G, Table 3.2.5) is a Ref RIP L1 located to the long arm of chromosome 8. It is intergenic, and it is not located within 100+ kb of a CGI. An 89.4 % retrotransposition activity of $L1_{RP}$ was attributed to this element [73, 186]. A summary of the information on these seven hot RC-L1 elements can be found on table 3.2.5.

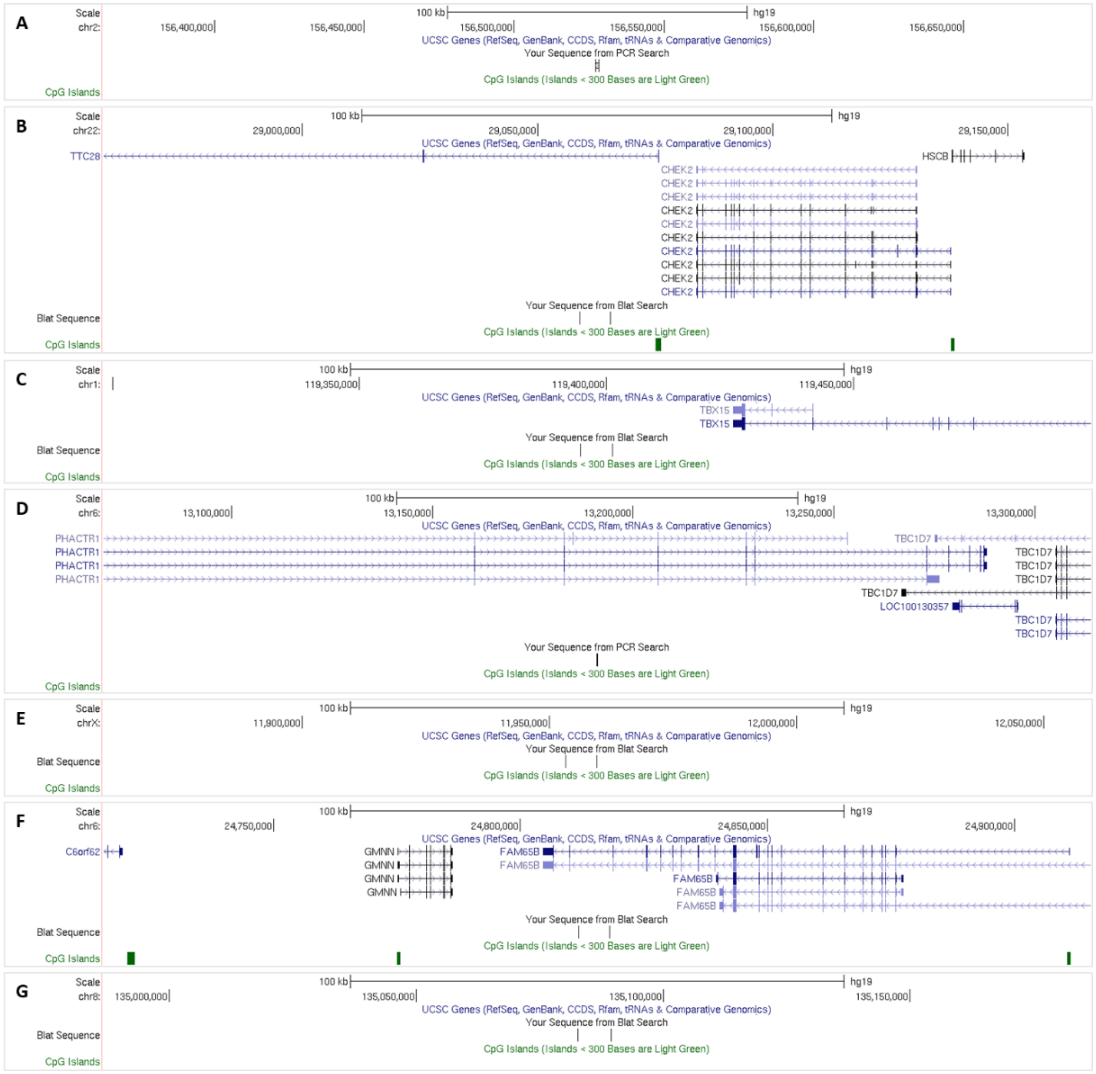


Fig. 3.2.5. UCSC image of the location of hot RC-L1 elements. A. Non-ref_chr2_q24.1 is a Non-ref RIP L1 located to the long arm of chromosome 2. **B.** Ref_chr22_q12.1 s a Ref L1 located to the long arm of chromosome 22. **C.** Ref_chr1_p12 is a Ref L1 located to the short arm of chromosome 1. **D.** Non-ref_chr6_q24.1 is a Non-ref RIP L1 located to the long arm of chromosome 6. **E.** Ref_chrX_p22.2 is a Ref RIP L1 located to the short arm of chromosome X. **F.** Ref_chr6_p22.3 is a Ref RIP L1 located to the short arm of chromosome 6. **G.** Ref_chr8_q24.22 is a Ref RIP L1 located to the long arm of chromosome 8.

Table 3.2.5. Hot RC-L1 elements chosen for genotyping analysis in the Dyne Steele cohort. The L1s were shortlisted based on their high level of activity in a cellular retrotransposition assay*,** [73], high number of germline offspring elements from 3' transductions analysis*** or high number of somatic insertions in cancer from 3' transductions analysis**** [186, 187].

Name	% of retrotransposition activity of L1RP*,**	Ref/non-ref	Insertion Allele Frequency**,***	Number of germline offspring elements from 3' transduction analysis***	Number of 3' transduction somatic insertions in cancer****	Chromosomal loci
Non-ref_chr2_q24.1	150	Non-ref RIP	na,0.16	41/121	16	chr2:156527848 intergenic
Non-ref_chr6_q24.1	a141	Non-ref RIP	na,0.18	14/121	75	chr6:13191033 Intron *PHACTR1*
Ref_chrX_p22.2	132	Ref RIP	0.34,0.74	1/121	11	chrX:11953208 intergenic
Ref_chr6_p22.3	112.7	Ref RIP	0.30,0.61	2/121	0	chr6:24811907 Intron *FAM65B*
Ref_chr8_q24.22	89.4	Ref RIP	0.44,1	2/121	1	chr8:135082987 intergenic
Name	**% of retrotransposition activity of L1RP*,****	**Ref/non-ref**	**Insertion Allele Frequency**,*****	**Number of germline offspring elements from 3' transduction analysis*****	**Number of 3' transduction somatic insertions in cancer******	**Chromosomal loci**
Ref_chr22_q12.1	13.8	Ref	1,1	2/121	61	Chr22:29059272 Intron *TTC28*
Ref_chr1_p12	-	Ref	1,1	13/121	5	chr1:119394974 intergenic

Non-ref RIP – L1s that are not in the human reference genome and are polymorphic for their presence/absence. Ref RIP – L1s that are present in the human reference genome and are polymorphic for their presence/absence. Ref – L1s present in the human reference genome and there is currently no evidence that they are polymorphic for their presence/absence.

3.2.4.3 The frequency of hot RC-L1 elements does not correlate with healthy cognitive ageing

Out of the seven hot RC-L1 elements shortlisted from the literature, five (Non-ref_chr2_q24.1, Non-ref_chr6_q24.1, Ref_chr6_p22.3, Ref_chr8_q24.22 and Ref_chrX_p22.2) are RIPs. We therefore genotyped the L1 RIP variants in 16 individuals (11 HA and 5 AD) from the Dyne Steele cohort (section 2.1.2.2) to determine if there was an association of presence/absence or frequency of L1 RIPs with healthy cognitive ageing. Primers outlined on table 3.2.1 were used for genotyping and, a description and an example of a gel image from QIAxcel capillary electrophoresis of the ES/5' junction PCR fragments of Ref_chr8_q24.22 L1 in HA and AD samples in shown in figure 2.2. The genotype frequencies observed for each element are summarised on table 3.2.6.

Individuals can be either homozygous present, homozygous absent or heterozygous for the hot RC-L1 RIPs (Figure 2.2B). We assessed the frequency of the five L1 RIPs together. The percentage of individuals with 0 to 10 alleles present (two alleles possible for each element) of the five L1 RIPs was calculated in HA and AD individuals in order to assess whether there was a correlation between the frequency of L1 RIPs and healthy cognitive ageing (Figure 3.2.7A). We did not observe any correlation in the number of present alleles and healthy cognitive ageing. Further, we observed no differences between HA and AD individuals in the number of present alleles when the data is stratified by gender (Figure 3.2.7B&C). Several studies correlate allele sequence variation with mobilisation capabilities instead [188, 189] and hence with associated disease risk, which was not assessed in the present study and hence, sequencing and reassessment is required to further identify any correlation between

hot RC-L1 sequence variation and healthy cognitive ageing. Further, because of the small n number, one would only really expect to see differences if there was dramatic and pronounced effect. Therefore, it is noteworthy for future larger studies that there was a tendency in AD males towards a higher frequency of hot RC-L1 elements.

As there was a trend regarding L1 RIPs frequency when the data was stratified by gender, we further interrogated the data excluding Ref_chrX_p22.2, a Ref RIP located to the short arm of chromosome X. The percentage of individuals with 0 to 8 alleles present of four out of the five L1 RIPs was calculated in HA and AD individuals (Figure 3.2.8). The trend did not vary when Ref_chrX_p22.2 was excluded from the analysis.

Table 3.2.6. There is no association of hot RC-L1 RIPs frequency with either Alzheimer's disease or healthy cognitive ageing. Each individual L1 was genotyped in AD and HA samples. **A.** Allele and genotype frequencies of Non-ref_chr2_q24.1, Non-ref_chr6_q24.1, Ref_chr6_p22.3 and Ref_chr8_q24.22. **B.** Allele and genotype frequencies of Ref_chrX_p22.2 broken down by gender as it is on the X chromosome. IAF – insertion allele frequency, AA – homozygous absent for L1, PA – heterozygous for the presence/absence of L1, PP – homozygous present for L1, A – hemizygous absent for L1, P – hemizygous for presence of L1.

A

			Genotype (%)		
		IAF	AA	PA	PP
Non-ref_chr2_q24.1	Healthy aged (11)	0.14	72.7	27.3	0
	Alzheimer's Disease (5)	0.10	80	20	0
Non-ref_chr6_q24.1	Healthy aged (11)	0.09	81.8	18.2	0
	Alzheimer's Disease (5)	0.20	60	40	0
Ref_chr6_p22.3	He aalthy aged (11)	0.18	63.6	36.4	0
	Alzheimer's Disease (5)	0.20	60	40	0
Ref_chr8_q24.22	Healthy aged (11)	0.23	9.1	54.5	36.4
	Alzheimer's Disease (5)	0.45	20	20	60

B

	Female		Genotype (%)		
		IAF	AA	PA	PP
Ref_chrX_p22.2	Healthy aged (7)	0.29	14.3	57.1	28.6
	Alzheimer's Disease (2)	0.25	0	50	50
	Male	**IAF**	**A**	**P**	**PP**
Ref_chrX_p22.2	Healthy aged (4)	0.13	75	25	NA
	Alzheimer's Disease (3)	0.33	33.3	66.7	NA

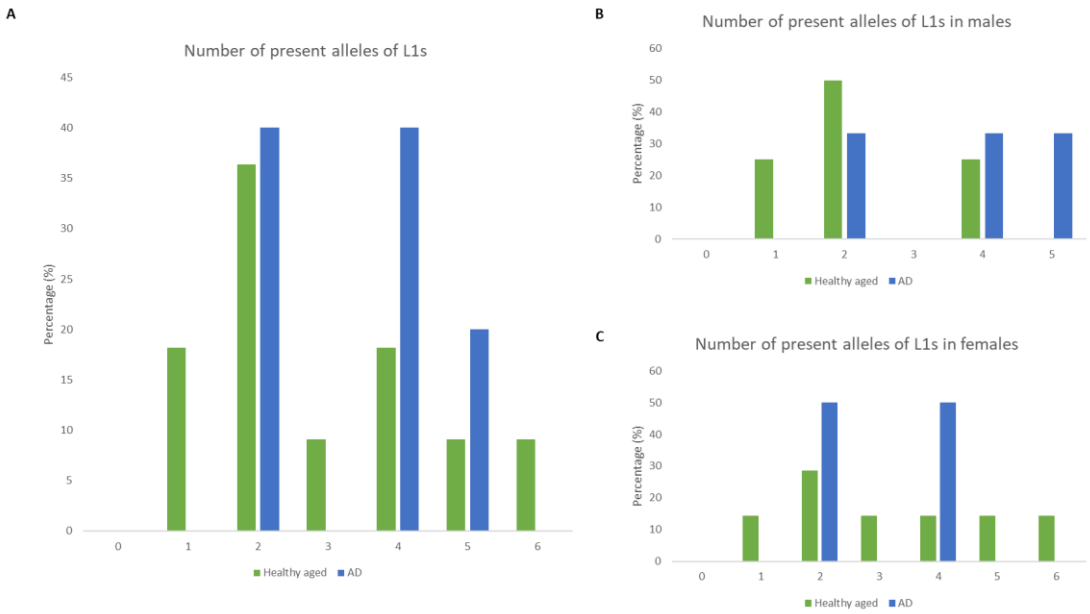


Fig. 3.2.7. There is no significant difference in the combined burden of present alleles of the five hot RC-L1 RIPs genotyped in healthy aged samples compared to Alzheimer's disease. A. The percentage of individuals with 0-10 present alleles of the five L1 RIPs genotyped from table 3.2.5 (Non-ref_chr2_q24.1, Non-ref_chr6_q24.1, Ref_chrX_p22.2, Ref_chr6_p22.3 and Ref_chr8_q24.22) in HA and AD. There is no significant difference in the distribution of the number of present alleles when comparing HA and AD using t-test (p value=0.24). HA – n=11, AD n=5 **B.** The percentage of male individuals with 0-10 present alleles of the 5 L1s RIPs genotyped from table 3.2.5 in HA and AD. There is no significant difference in the distribution of the number of present alleles when comparing HA and AD using t-test (p value=0.7). HA – n=4, AD n=3. **C.** The percentage of female individuals with 0-10 present alleles of the 5 L1 RIPs from table 3.2.5 in HA and AD. There is no significant difference in the distribution of the number of present alleles when comparing HA and AD using t-test (p value=0.07). HA – n=7, AD n=2.

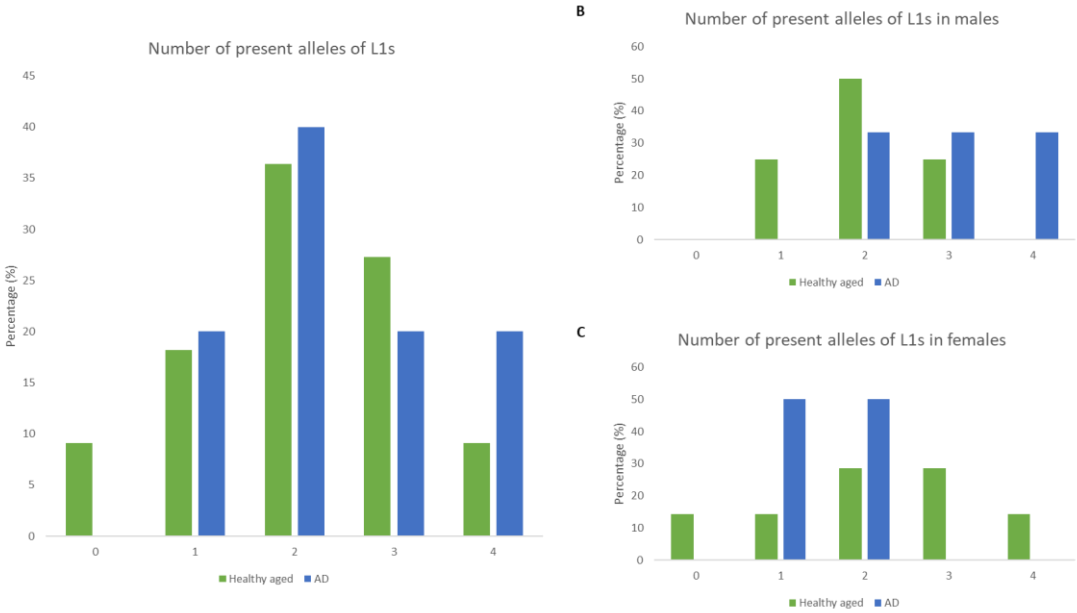


Fig. 3.2.8. There is no significant difference in the combined burden of present alleles of four out of the five hot RCL1 RIPs genotyped in healthy aged samples compared to Alzheimer's dise-ase when excluding Ref_chrX_p22.2. A. The percentage of individuals with 0-8 present alleles of the four L1 RIPs genotyped from table 3.2.5 (Non-ref_chr2_q24.1, Non-ref_chr6_q24.1, Ref_chr6_p22.3 and Ref_chr8_q24.22) in HA and AD. There is no significant difference in the distribution of the number of present alleles when comparing HA and AD using t-test (p value=0.24). HA – n=11, AD n=5 **B.** The percentage of male individuals with 0-8 present alleles of the four L1 RIPs genotyped from table 3.2.5 in HA and AD. There is no significant difference in the distribution of the number of present alleles when comparing healthy aged samples and AD using t-test (p value =0.7). HA – n=4, AD n=3. **C.** The percentage of female individuals with 0-8 present alleles of the four L1 RIPs genotyped from table 3.2.5 in HA and AD. There is no significant difference in the distribution of the number of present alleles when comparing healthy aged samples and AD using t-test (p value =0.09). HA – n=7, AD n=2.

3.2.4.4 There is a potential association of the methylation status of hot RC-L1s and ageing

In ageing and neurodegenerative conditions, there is an expected increase in L1 copy number and expression [135]. As neither analysis of the global L1 methylation index nor the frequency of L1 RIPs stand as a clear factor associated with such variable expression, we focused on the methylation status of four hot RC-L1 elements. As mentioned, DNA methylation analyses of the other three elements shortlisted as active from the literature was not possible due to PCR amplification not being reproducible on sheared DNA.

In this study, we assessed the methylation status of four hot RC-L1s in the temporal cortex and matched blood DNA of 9 HA people and 3 AD patients (table 3.1).

The methylation index of two L1 insertions (Non-ref_chr6_q24.1, n=3 (2 HA & 1 AD); and, Ref_chr6_p22.3, n=5 (3 HA & 2 AD)) was addressed in heterozygous carriers of the L1 element. The other two elements shortlisted for methylation analysis (Ref_chr22_q12.1 and Ref_chr1_p12) were reference L1s with no evidence of being polymorphic and thus, we did not perform genotyping analysis and had no data on heterozygotes. By analysing heterozygous L1 carriers, we determined that the allele harbouring the L1 element had a significantly higher level of methylation than the allele that did not have an L1 insertion both in the temporal cortex and in the blood (2-tailed paired *t*-test, temporal cortex p-value = 0.01; blood p-value = 0.0002, Figure 3.2.9B). When we looked at the temporal cortex and blood together, the allele harbouring the L1 element presented an even more significant higher level of

methylation than the allele that did not have an L1 insertion (2-tailed paired *t*-test, p-value = 8.849E-06, Supplementary figure 3.2.1).

Further, we addressed the methylation index differences of the four hot RC-L1 elements between the temporal cortex and matched blood DNA. Our data demonstrated that the level of methylation of these L1 elements is significantly higher in blood than in temporal cortex (average methylation status of L1s – temporal cortex – 0.48, blood – 0.74; 2-tailed paired *t*-test, p-value = 0.002, Figure 3.2.10B). We further stratified the data by health status comparing HA with AD individuals and found that there were no significant differences between HA and AD individuals neither in the temporal cortex nor in the blood (2-tailed unpaired *t*-test assuming unequal variances, temporal cortex p-value = 0.49, blood p-value = 0.99, Figure 3.2.10C).

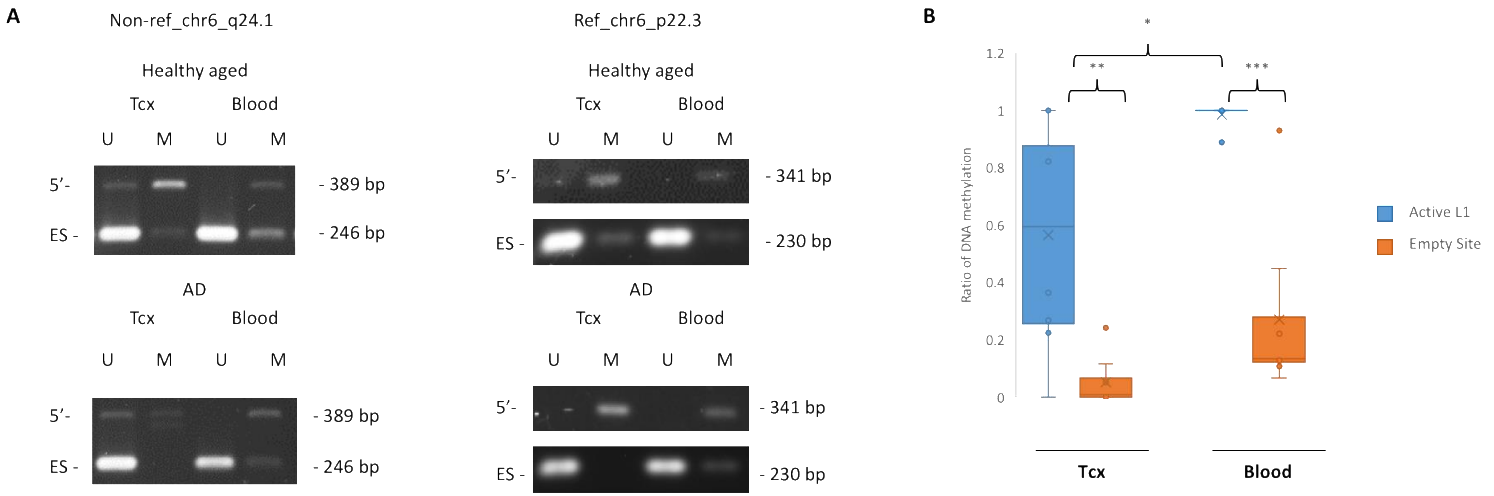


Fig. 3.2.9. The allele harbouring the hot RC-L1 insertions shows a significantly higher level of methylation compared to the allele lacking the insertion both in the temporal cortex and in the blood. A. Example gel images of PCR products of the 5' end of the L1 insertions and their empty site [Non-ref_chr6_q24.1, n=3 (2 HA & 1 AD); Ref_chr6_p22.3, n=5 (3 HA & 2 AD)] of heterozygous carriers in the unmethylated and methylated fractions of DNA. **B**. Methylation status of the 5' end of the L1 insertion and the empty site of the allele lacking the L1 insertion by comparing the band intensity of the PCR product in the unmethylated and methylated fractions of DNA in the temporal cortex and blood of heterozygous carriers. Tcx average methylation status – 5' end of active L1 insertion =0.56 and empty site=0.53 (p-value = 0.01). Blood average methylation status – 5' end of active L1 insertion =0.99 and empty site=0.27 (p-value = 0.0002). The methylation level of the temporal cortex 5' end of active L1 insertion is significantly different from the blood (p-value=0.02). *p* values are calculated by 2-tailed paired *t*-test. Tcx – temporal cortex, U – unmethylated fraction of DNA, M – methylated fraction of DNA.

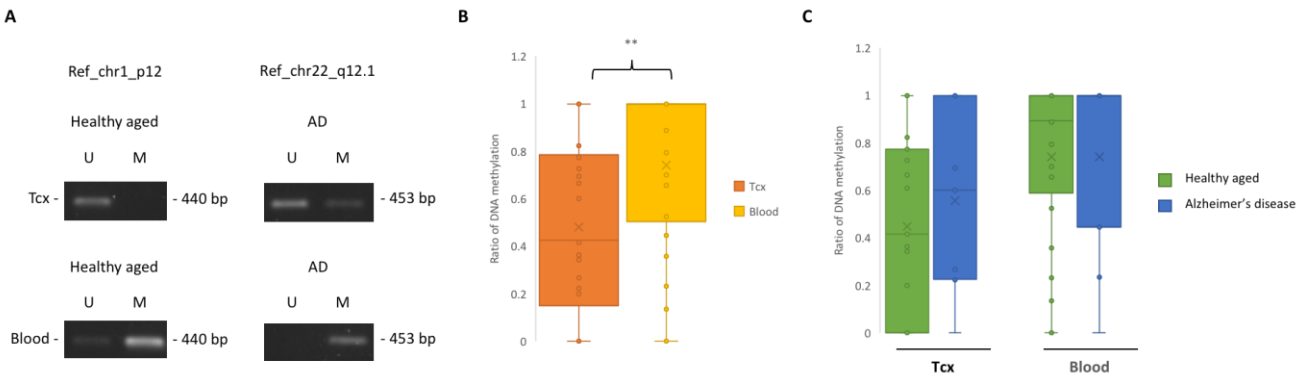


Fig. 3.2.10. Hot RC-L1s from table 3.2.5 show a statistically significant higher level of methylation in the blood compared to the temporal cortex, but no statistically significant difference in the level of methylation in the blood and the temporal cortex when Alzheimer's disease samples are compared to healthy aged individuals. A. Example gel images of PCR products of the 5' end of the L1 insertions (Ref_chr1_p12, Ref_chr22_q12.1) in the unmethylated and methylated fractions of DNA. **B.** Methylation status of the 5' end of L1 insertions from table 3.2.5 by comparing the band intensity of the PCR product in the unmethylated and methylated fractions of DNA in the temporal cortex and blood. Average methylation status of active L1 – temporal cortex – 0.48, blood – 0.74 (p-value = 0.002). *p* values are calculated by 2-tailed paired *t*-test. **C.** Methylation status of the 5' end of L1 insertions from table 3.2.5 by comparing the band intensity of the PCR product in the unmethylated and methylated fractions of DNA in the temporal cortex and blood of healthy aged people and AD individuals. Temporal cortex average methylation status of active L1 – healthy aged– 0.45, AD – 0.56 (p-value=0.49). Blood average methylation status of active L1 – HA– 0.74, AD – 0.74 (p-value=0.99). *p* values are calculated by 2-tailed unpaired *t*-test assuming unequal variances. Tcx – temporal cortex, U – unmethylated fraction of DNA, M – methylated fraction of DNA.

3.2.5 Discussion

In the previous chapter, we assessed the methylation of the *MIR941*/VNTR locus in relation to healthy cognitive ageing due to the role of the region as a regulatory element located in a neuroprotective gene named *DNAJC5*. In this chapter, a similar approach was taken to study the properties of a subclass of non-LTR retrotransposons named LINE-1. L1s are the only autonomous element that remains mobile in the human genome. An average human genome comprises over 500,000 L1 copies, 80-100 of which are retrotransposition competent. Despite the ability of these to retrotranspose, the majority of retrotransposition activity has been attributed to a group of L1 named hot RC-L1s. Because of their polymorphic nature, each individual presents a different L1 make-up [73, 87, 186, 187]. In line with this observation, an age-associated increase in L1 expression and copy number is observed [135]. In fact, epigenetic changes such as DNA methylation widely vary amongst individuals and, this epigenetic drift becomes more widespread as we age [137]. In the human genome, while CGIs are generally unmethylated, the majority of CGIs associated to a retrotransposon are often methylated [179]. Retrotransposons have often being identified as having a deleterious effect to the host, and thus, their high level of methylation may be counteracting their impact to the human genome [53]. The increase in the copy number of L1s associated with age may be attributed to transcriptional activation of these elements, which in turn could be resulting from global (CGIs) or more localised (non-LTR TEs themselves) methylation level changes observed during ageing [135].

In this chapter, using bisulphite pyrosequencing we assessed the global methylation index of LINE-1 elements. We found that there was no biological association of the level of global L1 methylation with either health status or tissue. Although we did see a significant difference in the level of global L1 methylation in the temporal cortex compared to the blood, differences of less than 5 % are associated with assay to assay variability [168]; and thus, despite there being a statistically significant difference, we conclude that there is no biological tissue-specific association of the global L1 methylation index. In this context, it may be helpful to remark that this approach targets the 5' promoter GGI L1 consensus sequence. Coufal *et al.* suggests that the main DNA methylation changes linked to L1 expression occur at the 3' end of the CGI [115], which is not addressed in our analyses. Further, assessing the DNA methylation level looking at a consensus sequence results in the average methylation of thousands of LINE-1 elements. Therefore, changes in the methylation level, whether tissue-specific or disease-associated, could be the result from variation occurring at specific L1s instead [135].

We hypothesise that hot RC-L1 methylation specifically may be crucial to regulating global L1 retrotransposition due to most mobilisation events being attributed to this subset of L1 elements particularly. In order to assess the methylation changes that result potentially from specific elements of the L1 family, we measured the methylation index of the group of L1 elements that are responsible for the majority of retrotransposition

events named hot RC-L1s as we may expect more drastic methylation changes on those more active elements.

Firstly, as these elements are polymorphic for their presence/absence, we genotyped them in order to weigh whether there was a correlation between an individual's hot RC-L1s complement and healthy cognitive ageing. The small n number made the correlation difficult, but it was observed that AD males showed an increased number of hot RC-L1 elements. Hot RC-L1 RIPs are not only polymorphic for their presence/absence, but also for their allelic sequence [188, 189], which may be further correlated with healthy cognitive ageing. Therefore, hot RC-L1s allele sequencing should be addressed in future to study a putative correlation.

By further analysis of the methylation index of these elements, our data suggests a higher level of methylation associated with L1 presence (Figure 3.2.9B) irrespective of tissue and health status. Previous studies have reported that it is not clear whether it is the genomic environment that incurs methylation changes on an element or *de novo* methylation of adjacent regions occurs after insertion [179]. Our findings suggest that methylation changes in the genomic environment result from the L1 insertion itself and may be additionally correlated to allele sequence variation [188, 189]. Further interrogation of the data suggests that the level of methylation of hot RC-L1 elements is higher in blood than in temporal cortex tissue (Figure 3.2.10B) regardless of health status, suggesting that L1 elements are more active in the temporal cortex of elderly people and supporting previous literature. It is important to note that each hot RC-L1

has its own genomic environment (Figure 3.2.6) [135, 179]. Therefore, by grouping highly active elements for analysis, their discrete genomic surroundings are not considered. Because of the small n number of people assessed for methylation changes, hot RC-L1 elements needed to be considered as a group. It is therefore clear, that the study of hot RC-L1 elements in a much larger cohort and taking into account sequence-specific variation is required to understand the role they may be playing in healthy cognitive ageing, and also in neurodegenerative diseases. The development of single nucleotide polymorphisms (SNPs) where the genotype of the SNP can be associated with L1 presence/absence would be useful to expand not only the genotyping analysis of hot RC-L1s, but also the methylation changes that occur within each element via bioinformatic approaches.

Chapter 4

Next generation sequencing (NGS)

Chapter 4 Next generation sequencing (NGS)

Chapter 4 addresses next generation sequencing (NGS) of non-LTR TEs. It is divided upon two subchapters, 4.1 and 4.2. Each subchapter has its own introduction, aim or aims, materials and methods, results and discussion. This brief introductory section covers the communalities between the two subchapters to avoid reiteration.

NGS sample information

The temporal cortex and blood DNA of HA and AD individuals was used to prepare: LINE-1 enriched libraries for retrotransposon capture sequencing (RC-Seq), and whole genome sequencing (WGS) libraries; and, resulting sequencing data analysed throughout the following two chapters (4.1 and 4.2). Table 4.1 contains phenotypic information on the samples used for RC-Seq and WGS.

Table 4.1. Information on the Dyne Steele cohort samples used for RC-Seq/WGS. Those where age changes only for path diagnosis 1 were considered healthily cognitively aged and those where path diagnosis 1 is Alzheimer's disease were considered as cases. Highlighted in green are the samples we did the analyses for, and in red, those where there was not sufficient DNA to perform analyses.

Case	Code	Cohort ID	Age at death	Sex	Path diagnosis 1	Path diagnosis 2	RC-Seq	WGS
09/24	HAs1	11427	78	M	Age changes only	mild SVD		
09/26	HAs2	22110	84	M	Age changes only	mild SVD		
09/31	HAs3	11508	94	F	Age changes only	mild to moderate SVD		
11/06	HAs4	20935	91	F	Age changes only	mild SVD		
11/07	HAs5	10132	80	F	Age changes only			
11/22	HAs6	20088	89	F	Age changes only			
11/29	HAs7	23350	89	M	Age changes only	mild SVD		
14/04	HAs8	21092	89	F	Age changes only			
14/46	HAs9	11052	94	F	Age changes only	mild SVD		
15/01	HAs10	12504	90	M	Age changes only			
15/28	HAs11	22708	91	F	Age changes only	Cerebral infarction		

Case	Code	Cohort ID	Age at death	Sex	Path diagnosis 1	Path diagnosis 2	
10/07	ADs1	11426	88	F	Alzheimer's disease	moderate SVD, mild TDP-43	
15/11	ADs2	10640	104	F	Alzheimer's disease	Secondary TDP43, Limited SVD	
16/03	ADs3	11233	91	M	Alzheimer's disease	Severe CAA	
16/09	ADs4	22683	96	M	Alzheimer's disease	Limbic DLB	
16/13	ADs5	21596	91	M	Alzheimer's disease		

TEBreak, the pipeline used to analyse sequencing data from RC-Seq and WGS

Even though some of the mobile element insertion detection software tools available have unique features, in general, they follow a common approach and the method of choice depends on the particular use [85]. TEBreak has been demonstrated as a more reliable pipeline than others available and it is suitable for analysing sequencing data from both RC-Seq and WGS, and hence it was our pipeline of choice for bioinformatic analyses [16, 85]. The output from TEBreak is a tab-delimited file in the form of a table (Figure 4.1), which we open as an excel file. A header defines each column and a full description of what each header stands for can be found on TEBreak's manual (https://github.com/adamewing/tebreak/blob/master/doc/manual.pdf).

The insertions identified using TEBreak's output can be classified into polymorphic and tissue-specific (Figure 4.2). New tissue-specific insertions can be mapped. In fact, these insertions are present in one tissue alone (either temporal cortex or matched blood) and not previously reported in the literature. Throughout the course of this thesis, we refer to these as somatic insertions. In addition, computational analysis can map polymorphic insertions, which can be private or already described in the population. Polymorphic private insertions are present in both tissues, the temporal cortex and blood, whereas polymorphic insertions described in the population are also present in the annotated genome. For the purpose of this thesis, we refer to both groups as polymorphic insertions.

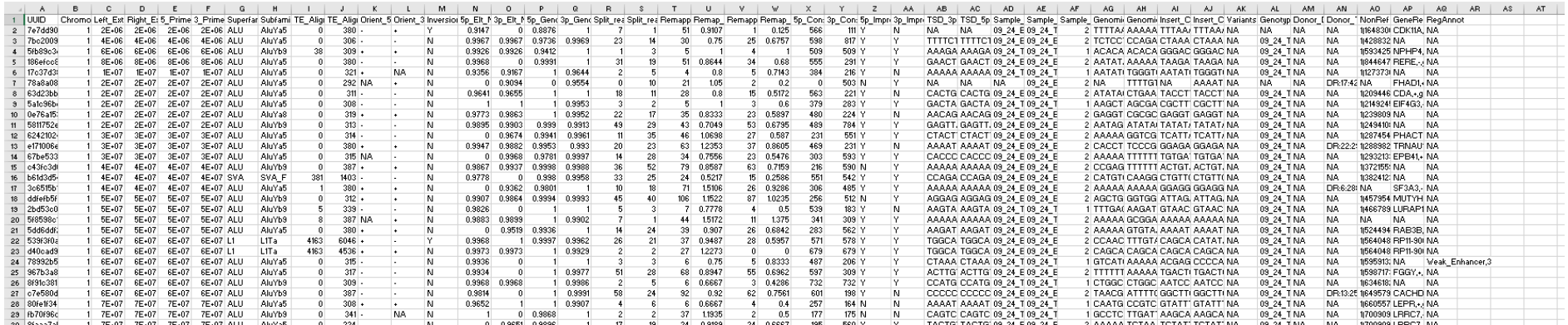


Fig. 4.1. An example of the output from TEBreak analysis of a WGS library in excel. The output from TEBreak is in the form of a table, which we open as an excel file. Each column is defined by a header and a full description of TEBreak's output can be found on TEBreak's manual (https://github.com/adamewing/tebreak/blob/master/doc/manual.pdf). Using the custom filter options from excel, we can select the data to address our specific interests.

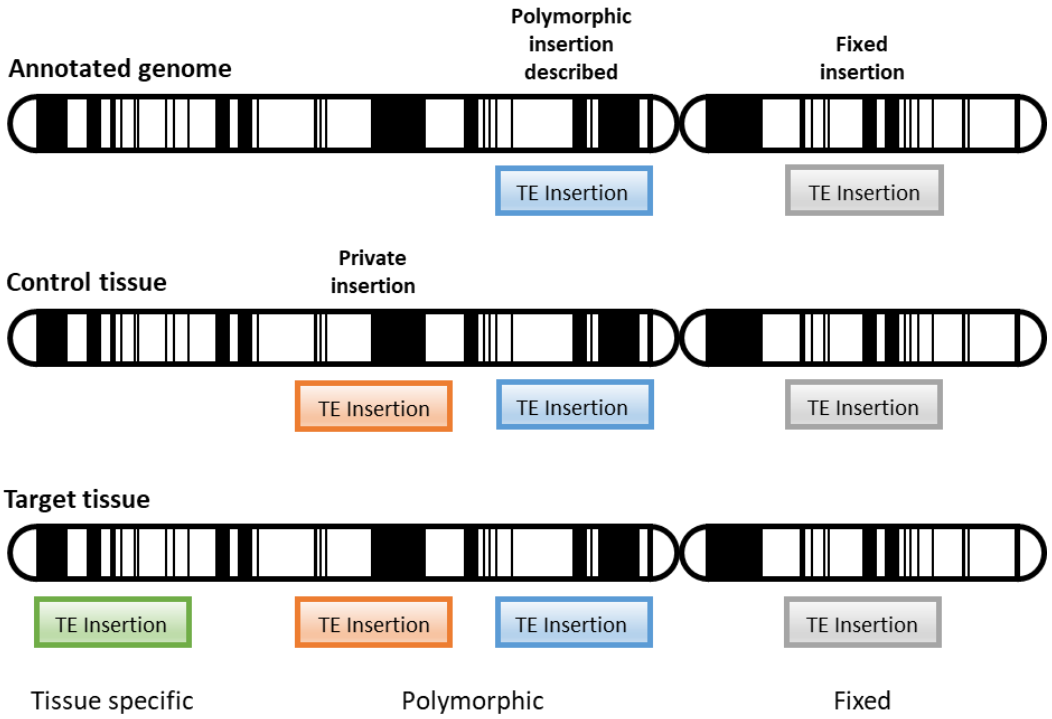


Fig. 4.2. A schematic representation of the types of TE insertion. The annotated genome is the genome we are using as reference. Whereas our target tissue is temporal cortex, our matched control tissue is blood. A tissue specific insertion (green) is defined as that present only in either tissue and not previously described (i.e. somatic). There are two types of polymorphic insertions (non-reference RIPs): a private insertion not previously described (orange), present in control and target tissue, but not in the annotated genome; and, a polymorphic insertion previously described (blue), also present in the annotated genome. A fixed insertion (grey) is present in annotated, control and target tissues, and it is not polymorphic. Adapted from Sanchez-Luque *et al* [88].

Chapter 4.1

The analysis of LINE-1 insertion polymorphisms and somatic variation in the context of ageing

4.1. The analysis of LINE-1 insertion polymorphisms and somatic variation in the context of ageing

4.1.1 Introduction

As we have mentioned throughout the course of this thesis despite the deleterious nature often associated to non-LTR retrotransposons, they have played a fundamental role in shaping human genome structure and function [53, 58, 68]. As an example, retrotransposition in embryonic stem cells can be explained by the selfish nature of the elements trying to reach the germline [197]; however, insertions in neuronal progenitor cells may be attributed to putative functional roles of the elements in the brain such as altering the expression of neuronal genes [116, 198]. Therefore, not only resident elements, but also polymorphic (i.e. RIPs) and somatic insertions have the potential to alter an individual's phenotype depending on when and where they insert on the genome and their specific functional impact [62]. Because of their mutagenic nature, transposition-competent retrotransposons are often heavily methylated and transcriptionally inactivated [199, 200]. Despite these efforts from the host genome to supress retrotransposition, a considerable amount of somatic L1 retrotransposition has been detected in neuronal cell lineages, which could be of major significance to human neurobiology [116, 197, 198]. Whereas fixed insertions are rarely found in regions where they generate a harmful phenotype because of strong selection against these during evolution, polymorphic and somatic events may be more likely to positively or negatively affect protein-coding loci and key regulatory domains in a specific environmental context [69]. However, the identification of RIPs and, specifically somatic insertions amongst inherent elements in the human genome is technically very challenging [201, 202] as it depends vastly

on the number of cells carrying the insertion in a heterogeneous cell population and requires costly high sequencing depth [53].

In 1988, the discovery of a *de novo* LINE-1 insertion as the cause of haemophilia A in a male patient was the first unequivocal proof of active L1 retrotransposition activity in the human genome [203]. To date, LINE-1 elements are the only active autonomous non-LTR retrotransposon able to mobilise in the human genome; and, they are responsible for mobilisation of other non-LTRs such as *Alu* or SVAs, which are active, but non-autonomous elements [72, 73]. In this context, L1 mobilisation and L1-mediated retrotransposition events lead to a significant amount of polymorphic and somatic insertional events in the brain [204]. Therefore, mobilisation of non-LTR retrotransposons clearly contributes to mutational events and evolution [58, 69]; and the accumulation of harmful mutations that lead to loss of function, cell death or uncontrolled growth is a hallmark of disease and ageing [58, 69]. In particular, ageing is characterised by an increase in transcription and copy number of LINE-1 elements amongst other TEs. However, it remains to be demonstrated if such an increase occurs through *de novo* insertion of the elements leading to somatic mutation [135].

A technique named retrotransposon capture sequencing (RC-Seq) was developed to elucidate polymorphic and somatic L1 insertions. Prior to the development of RC-Seq in 2011 [69], proof of endogenous L1 retrotransposition by mapping and sequence characterization of L1 insertions in the brain (i.e. ATLAS-Seq [90]) was scarce [88]. Indeed, such evidence is essential to gain a deeper understanding of the impact of L1 activity due to insertional patterns such as copy number or location of L1 insertions.

RC-Seq allowed identification and genome localisation of L1 insertions, in particular those present in a subset of cells within a tissue, demonstrating the extent of somatic L1 mosaicism in the brain [69, 205]. RC-Seq attempts to surpass the technical challenge of resolving uncharacterised polymorphic and somatic insertions by generating targeted sequencing libraries, which are enriched for the 5' and 3' ends of LINE-1 elements. In RC-Seq, genomic DNA from target and control tissue from the same individual is used as starting material to distinguish RIPs from somatic L1 insertions. Because of the nature of L1 insertions, often truncated at the 5' end, but intact at the 3' end; and, the location of the probes used to capture L1 sequences in RC-Seq (5' and 3' termini), many of the characterised insertions are so by their 3' end only. Validation by PCR amplification and capillary sequencing of the matched 5' L1-genome junction of the 5' truncated L1 insertions is the most common approach to corroborate legitimate insertions [88].

Here, using temporal cortex (target tissue) and blood (matched tissue) genomic DNA as starting material for RC-Seq, we were able to identify polymorphic and somatic insertions in 11 HA people and 5 AD patients (Table 4.1). Following computationally intensive bioinformatic analysis of the resulting sequencing files, we interrogated the data in order to determine whether a specific genetic signature of polymorphic or somatic L1 insertions was present in HA people and AD patients. L1 RIPs were defined as those present in both target and matched tissue whereas somatic L1 insertions were those that either appeared in the target or matched tissue alone and not previously described (Figure 4.1). Furthermore, as a proof of principle, we used PCR amplification to validate two polymorphic 5' truncated L1 insertions in a subset of HA

people. Lastly, we examined the data to assess the location of polymorphic and somatic insertions in reference to genes. Our results demonstrate that the number of polymorphic L1 insertions is not vastly different in HA versus AD individuals and support previous findings that L1 insertions locate to genes predominantly expressed in the brain [69] independently of health status.

4.1.2 Aim

To compare the temporal cortex and matched blood DNA of healthy aged and Alzheimer's disease individuals in order to establish whether the number, nature and/or location of polymorphic or somatic LINE-1 insertions is correlated with healthy cognitive ageing using:

- Retrotransposon capture sequencing (RC-Seq)
    - To determine the number of polymorphic LINE-1 insertions (RIPs)
    - To determine the number of putative somatic LINE-1 insertions
- Bioinformatics to assess the location of identified LINE-1 insertions
    - with regards to genes
    - with regards to AD, fluid intelligence and vocabulary ability haploblocks
- DAVID to carry out pathway analysis in order to analyse if LINE-1 insertions occur in genes that trigger significant enrichment of specific pathways

4.1.3 Methods

4.1.3.1 Preparation of next generation sequencing libraries enriched for LINE-1 5' and 3' termini

The RC-Seq protocol used and adapted in this chapter was described by Sanchez-Luque *et al.* 2016 [88] for preparation of LINE-1s termini enriched libraries. To learn the procedure, we spent two months at the Paul-Ehrlich-Institut (PEI) in Langen (Germany), where the protocol is routinely used. RC-Seq to identify polymorphic and somatic L1 insertion in human tissue was performed as described in Sanchez-Luque *et al.* 2016 and outlined in section 2.2.12. Briefly, temporal cortex genomic DNA was extracted from ~100 mg of post-mortem brain tissue using the Gentra Puregene Tissue Kit (QIAgen) following standardised protocol unless otherwise stated in this thesis. Blood genomic DNA was extracted from 20 ml of blood using the DNAce MaxiBlood Purification System (Bioline) according to manufacturer's instructions. 5 µg of genomic DNA from either tissue were used as starting material for library preparation. The Fragment Analyser (Advanced Analytical) was used to check the DNA size distribution and Qubit dsDNA HS/BR Assay was used to check the DNA concentration whenever needed throughout the protocol. During hybridization to locked nucleic acid (LNA) probes, 1µl of Sequence Capture Developer Reagent, 1 µl of Universal Blocking Oligo and 1 µl of Index-specific Blocker Oligo per 1 µg of total DNA were used. In order to adapt RC-Seq to the Quinn's Laboratory at the University of Liverpool (UoL), a few changes were made to the protocol described in Sanchez-Luque *et al.* 2016. Please, note that most of the protocol was carried out as described in Sanchez-Luque *et al.* 2016. However, due to the need to adapt to the technology available to us at the UoL, and to improve the yield of DNA obtained after size

selection, the main changes were the usage of the QIAxcel instead of the Fragment Analyser and the use of AMPure ® XP beads instead of agarose gel, respectively. A systematic protocol of RC-Seq adapted for performing at the UoL can be found on Appendix 4.

4.1.3.2 Non-reference LINE-1 polymorphic insertions validation by PCR

Two 5' truncated polymorphic LINE-1 insertions were selected for PCR validation (chr4:47007784 and chr18:1851762729). Empty site/Filled site (ES/FS) PCR was used to validate the two putative polymorphic insertions in 6 individuals from the Dyne Steele cohort (Figure 2.2). Primers were designed as outlined in section 2.2.2 to target the putative polymorphic L1 insertions, including chr4_L1_Fw: 5'-GCCTCCTGAAGATCCAAGGA-3', chr4_L1_Rv: 5'-CACAGTTTTGCTAAGCCCCA-3', chr18_L1_Fw: 5'-ACGCGTCGGTTTCTTCCTAT-3' and chr18_L1_Rv: 5'-GATCCTGGCTGCCCATATTA-3'. The L1 PCR reaction was performed using 5 ng of genomic DNA as starting material and GoTaq Hot Start (Promega) DNA polymerase with reagents outlined in table 4.1.1. A thermocycler set to 94 ˚C for 2 mins; 94 ˚C for 30 secs, 60 ˚C for 30 secs and 72 ˚C for 30 seconds for 30 cycles; 72 ˚C for 5 mins and 4 ˚C infinite hold was used for PCR amplification. PCR products were run on a 1 % agarose gel using EtBr and separated by electrophoresis at 100 mV for 90 mins. PCR product was visualised using a UV transilluminator while running a 100 bp ladder in parallel to the DNA fragments to determine their size.

Table 4.1.1. GoTaq Hot Start DNA polymerase ES/FS PCR reaction components

Component	Volume (µl)
PCR buffer (5x)	4.0
MgCl2 (25 mM)	2.5
dNTPs (10mM each)	0.4
Forward primer (20 mM)	0.1
Reverse primer (20 mM)	0.1
DNA polymerase (5 u/µl)	0.1
Nuclease free water	11.8
DNA template (5 ng/µl)	1
Final volume	20 µl

4.1.3.3 Bioinformatic analysis of LINE-1 libraries to detect non-reference polymorphic and somatic insertions

A step by step guide to bioinformatic analysis of LINE-1 libraries from RC-Seq is described in section 2.2.14. Bioinformatic scripts are detailed on Appendix 4. TEBreak (https://github.com/adamewing/tebreak) was the software of choice for detection of non-reference polymorphic and somatic L1 insertions. A full description of TEBreak can be found on its manual (https://github.com/adamewing/tebreak/blob/master/doc/manual.pdf). The output from TEBreak is in the form of a table, which we open as an excel file and custom filter in order to assess different parameters including, but not limited to, the number of non-reference L1 insertions, the location of insertions such as genes, and the specific tissue the insertion is present on. An example is shown in figure 4.1. Following bioinformatic analysis to resolve the sequencing data, haplotype block analysis and pathway analysis were used to further interrogate the data.

4.1.3.4 Haplotype block analysis to characterise non-reference L1 insertions distribution

In order to identify whether non-reference polymorphic and putative somatic L1 insertions were enriched in specific regions of the human genome such as AD risk loci or regions associated with cognitive function such as fluid intelligence or vocabulary ability, we split the genome into haplotype blocks. A haplotype block bed file for the human genome, GRch37/hg19 previously defined by Berisa *et al.* was intersected with GWAS hits for AD from the GWAS catalog and the top 150 GWAS for fluid intelligence/vocabulary ability, in order to create AD and healthy cognition haplotype

blocks. In total, there are 495 AD haploblocks, 98 fluid intelligence haploblocks and 104 vocabulary ability haploblocks.

Using Galaxy for assessing the haplotype block enrichment of non-reference polymorphic and putative somatic LINE-1 insertions from HA and AD individuals

The bedtools Intersect intervals from Galaxy (https://usegalaxy.org/) was used to find overlapping intervals between non-reference polymorphic and putative somatic L1 HA/AD insertions bed files and AD, fluid intelligence and vocabulary ability haplotype block bed files.

4.1.3.5 Using DAVID for pathway analysis

Pathway analysis was completed using DAVID (https://david.ncifcrf.gov/). A list of genes with non-reference polymorphic or putative somatic L1 RIPs from HA/AD was generated from TEBreak's output and used as DAVID's input. In order to understand the major biological pathways associated to our input gene list, we used default parameters for DAVID pathway analysis, and interrogated KEGG pathway, gene ontology (GO) annotations including biological processes, molecular function and cellular component, and UP Tissue categories, briefly explained below [206-208]:

- KEGG pathway – a collection of pathway maps including current knowledge on the molecular interaction, reaction and relation networks for metabolism, processing of genetic and environmental information, cellular processes, organismal systems, human diseases and drug development
- GO biological processes – the involvement of a gene on a biological change
- GO molecular function – the potential of a gene to perform actions

- GO cellular component – the location of a gene within a cell or its environment
- UP Tissue – tissue-specific expression of a gene

4.1.4 Results

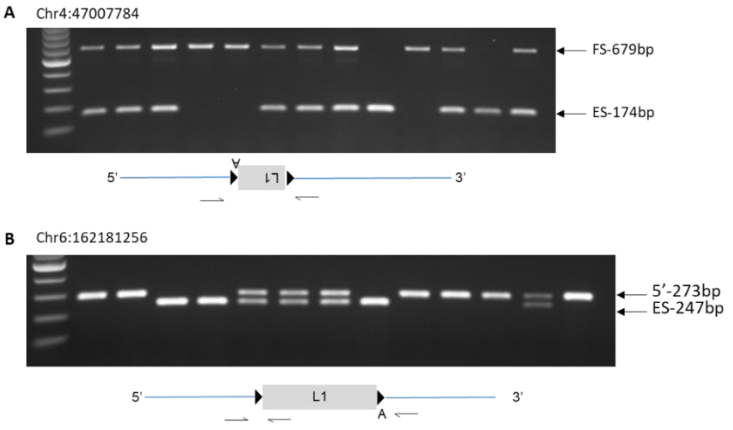
4.1.4.1 Validation of non-reference polymorphic LINE-1 insertions to corroborate accuracy, sensitivity and specificity of the parameters used for bioinformatic analysis

Previous to bioinformatic analysis of sequencing data from RC-Seq libraries, our former post-doc researcher, Abigail L. Savage, performed an extensive amount of PCR validation (Figure 4.1.1) on polymorphic LINE-1 insertions detected using TEBreak in order to validate the accuracy, sensitivity and specificity of TEBreak to detect non-reference L1 RIPs with custom parameters. The filtering step in TEBreak's bioinformatic analysis is crucial to reduce the rate of false positive and negative L1 insertions detection. The definition of the custom parameters used for the filtering step of bioinformatic analysis of L1 libraries and the custom values for these are below:

- Split read is the minimum number of supporting split read mappings and, it was set to 8 and 4 for polymorphic and somatic L1 insertions, respectively.
- Disc read is the minimum number of supporting discordant read mappings and it was set to 4 both for polymorphic and somatic L1 insertions
- Conslen is the minimum total consensus length (bp) and it was set to 150 bp both for polymorphic and somatic L1 insertions
- Eltmatch is the minimum element match (%) and it was set to 0.90 both for polymorphic and somatic L1 insertions
- Refmatch is the minimum reference match (%) and it was set to 0.95 both for polymorphic and somatic L1 insertions

- Maxvar is the maximum number of variants (SNPs) and it was set to 2 for somatic L1 insertions

PCR validation of non-reference L1 RIPs from TEBreak analysis with the above-mentioned filtering parameters showed that the false positive rate of insertions was minimum (0.03), meaning that we can trust most polymorphic insertions to be hypothetically true based on band length obtained by PCR validation compared to TEBreak bioinformatic analysis. The false negative rate was a bit higher (0.14) meaning that TEBreak bioinformatic analysis misses true insertions from some individuals. PCR validation of somatic L1 insertions from TEBreak analysis was not as successful, so for the purpose of our analysis, somatic insertions are always classed as putative. Even though it is not likely to occur, it is important to note that if the somatic insertion was to be in a single cell, validation by PCR is actually impossible, as the insertion served its purpose for library preparation and validation was attempted on genomic DNA to avoid PCR artifacts.



C

	L1
False positive rate	0.03
False negative rate	0.14
Accuracy	0.93
Sensitivity	0.86
Specificity	0.97

Fig. 4.1.1. PCR validation of non-reference L1 retrotransposon insertion polymorphisms (RIPs) identified in RC-Seq data using TEBreak. A. Empty/filled site PCR validation of a 5' truncated L1 polymorphic insertion into chr4:47007784. **B.** Multiplex PCR amplifying the empty site and 5'end of the L1 insertion of a polymorphic L1 at chr6:162181256. **C.** 13 different polymorphic L1 insertions (3 of which had not been previously reported in the literature) were PCR validated in 24 individuals for which RC-Seq libraries were generated. ES – empty site, FS – filled site. The ability of RC-seq to detect the L1 insertions in these individuals was assessed by using the following calculations:

True positive (TP) = present in TEBreak and PCR analysis

True negative (TN) = absent in TEBreak and PCR analysis

False positive (FP) = present in TEBreak and absent in PCR analysis

False negative (FN) = absent in TEBreak and present in PCR analysis

False positive rate (FPR) = FP/(FP+TN); False negative rate (FNR) = FN/(FN+TP)

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Sensitivity = TP/(TP+FN)

Specificity = (TN/(TN+FP)

In this project, LINE-1 libraries were generated for 16 individuals, 11 HA and 5 AD. We run TEBreak in all sequenced libraries. All scripts run for bioinformatic analysis of RC-Seq libraries were the same for detecting either polymorphic or putative somatic insertions except for the last filtering script, where two separate filters differing in the number of split reads were applied. The output was two tab-delimited files, one for polymorphic L1 insertions and one for putative somatic insertions.

4.1.4.2 The number of L1 insertion polymorphisms is on average higher in healthy aged people than in Alzheimer's disease patients

Using TEBreak's output from running genfilter_rcseq8_150_0.9_0.95_new.sh (8 split reads) script, polymorphic L1 insertions were considered those that TEBreak reported as being present both in the temporal cortex and blood genomic DNA (Figure 4.2). This was done by selecting the value 2 from the *Sample_count* column, where the value can be set to 1 or 2 depending on whether bioinformatic analysis found the insertion in 1 or 2 of the input tissues. LINE-1 where the *TE_Align_Start* (5') and *TE_Align_End* (3') columns were outside range, 0-54 bp and 5,921 bp onwards, respectively were excluded from the analysis. This is because the probe for the 5' end finishes at 54 bp into the L1, and the one from the 3' end starts at 5921 in the L1, so for it to be a true insertion, the putative L1 insertion needs to contain at least one of those regions. Following data curation, the number of non-reference L1 RIPs per individual stratified by health status was counted and graphed in Figure 4.1.2A. The number of L1 polymorphisms is higher in average in HA people (average no. of insertions: 124.3) than in AD patients (average no. of insertions: 90.6). However, due to the variability across individuals, our data suggests that the number of polymorphic L1 insertions was not vastly different in HA versus AD individuals.

In order to validate TEBreak's polymorphic L1 insertion calling, two putative L1 RIPs were selected and validated by PCR (Figure 4.1.2B&C) in the temporal cortex and blood of 6 individuals. At the time of PCR validation, sequencing data for only 6 out of the 16 individuals was available, and thus, we only validated the putative polymorphic L1 insertions in those individuals. TEBreak's output suggested both insertions (chr4 L1 and chr18 L1) were present in 5 out of the 6 individuals. There were no discordances between TEBreak L1 insertion calling and PCR data (TP rate = 1.00).

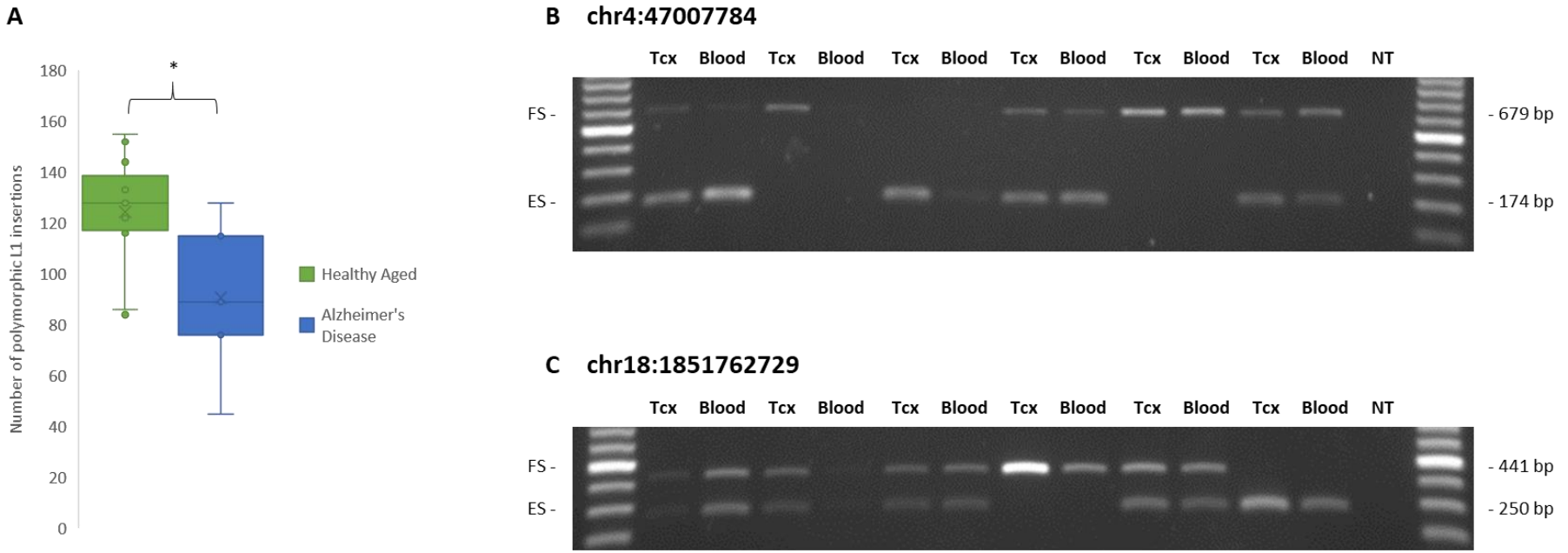


Fig. 4.1.2. Non-reference LINE-1 retrotransposon insertion polymorphisms (RIPs) identified by retrotransposon capture sequencing (RC-Seq) and validated by PCR. A. The number of polymorphic LINE-1 insertions identified in HA individuals and AD samples using TEBreak. The number of polymorphic LINE-1 insertions in HA people (average no. of insertions: 124.3) is higher than in AD patients (average no. of insertions: 90.6). *p-value* (p-value = 0.0326) was calculated by two sample t-test assuming equal variances (HA n=11, AD n=5). **PCR validation of two non-reference L1 RIPs identified in RC-Seq data using TEBreak. B.** Empty/filled site PCR validation of a 5' truncated L1 RIP into chr4:47007784 (n=6). **C.** Empty/filled site PCR validation of a 5' truncated L1 RIP into chr18:1851762729 (n=6).

4.1.4.3 There is a trend towards the number of putative somatic L1 insertions being on average higher in the temporal cortex of Alzheimer's disease patients than in healthy aged people

Using TEBreak's output from running genfilter_rcseq4_150_0.9_0.95_mv2_new.sh (4 split reads) script, putative somatic L1 insertions were considered those that TEBreak reported as being present either in the temporal cortex or in blood genomic DNA alone and not previously reported in the literature. This was done by selecting the value 1 from the *Sample_count* column. In addition to excluding LINE-1 where the *TE_Align_Start* (5') and *TE_Align_End* (3') columns were out of range, L1PA2 elements were also omitted from the analysis, as they are older L1s and do not retrotranspose anymore, so they are not able to mobilise to make somatic insertions. Following data curation, the number of putative somatic L1 insertions per individual stratified by health status was counted and graphed in Figure 4.1.3. The number of putative somatic L1 insertions is higher in AD patients than in HA people; however, this difference does not reach statistical significance. In fact, the trend whereby there is more putative somatic L1 insertions in AD patients than in HA people is not consistent. Three out of the five AD people present a higher number of putative somatic insertions; but two out of the eleven HA individuals do so too. Therefore, if we were to increase the n number of AD patients, this difference may dissipate.

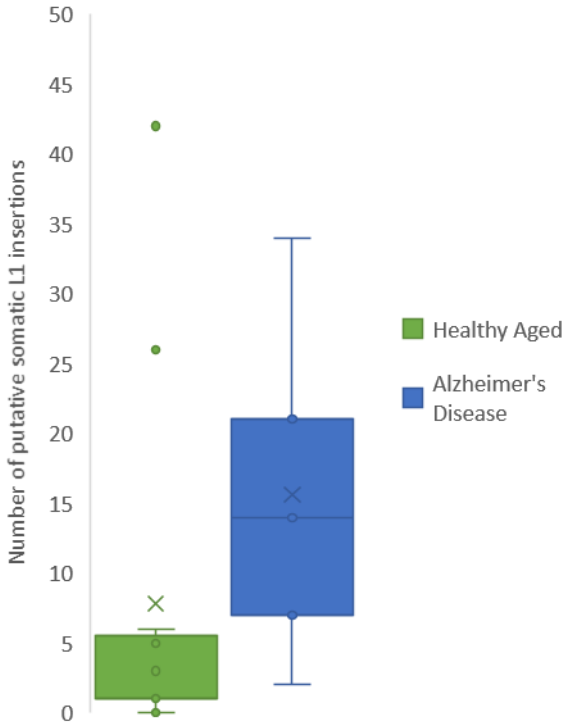


Fig. 4.1.3. The number of putative somatic LINE-1 insertions identified by retrotransposon capture sequencing (RC-Seq). The number of putative somatic LINE-1 insertions identified in each individual using TEBreak. The number of putative somatic L1 insertions is higher in AD patients (average no. 15.6) than in HA people (average no. 7.8); however, this difference does not reach statistical significance. *p-value* (p-value = 0.296) was calculated by two sample t-test assuming equal variances (healthy aged n=11, AD n=5).

We further interrogated the data to establish whether the likelihood of putative somatic L1 insertions was tissue-specific. The number of putative somatic LINE-1 insertions was higher in the blood than in the temporal cortex (Figure 4.1.4A) when we looked at HA and AD individuals together. However, this trend is driven by two outliers with an outstanding number of putative somatic LINE-1 insertions in the blood and hence, this difference is likely not true. The data was further stratified by health status in order to assess whether this difference, despite not significant, was consistent in either group. Whereas the number of putative somatic LINE-1 insertions was higher in the blood than in the temporal cortex (Figure 4.1.4B) of HA individuals, the number of putative somatic LINE-1 insertions was higher in the temporal cortex than in the blood (Figure 4.1.4B) of AD patients. It is important to note that the tendency towards a higher number of putative somatic L1 insertions in the blood of HA individuals is driven by the two outliers mentioned above, and hence, this trend is likely not true.

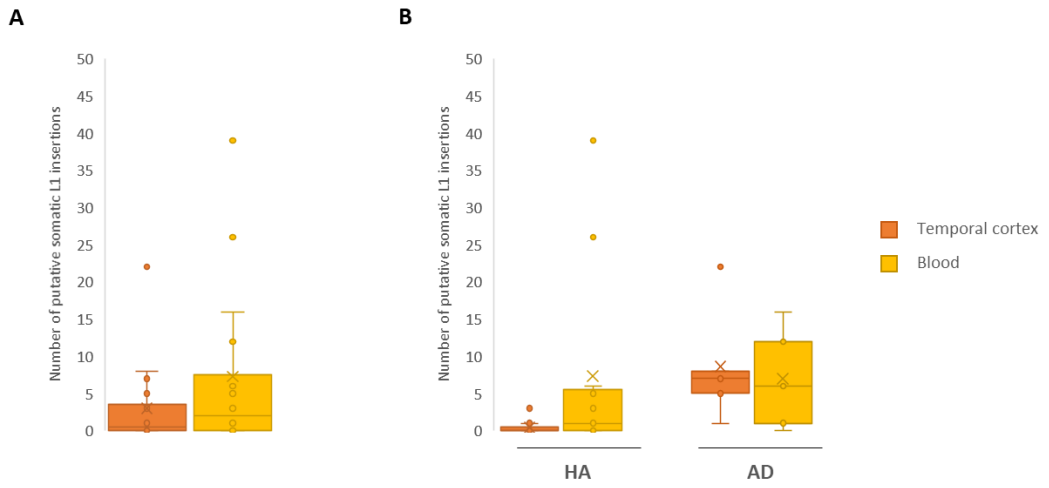


Fig. 4.1.4. The number of putative somatic LINE-1 insertions identified by retrotransposon capture sequencing (RC-Seq). A. The number of putative somatic LINE-1 insertions identified in the temporal cortex and blood of HA and AD individuals together using TEBreak and there is no tissue-specific significant difference in the number (average no. temporal cortex 3; average no. blood 7.3). *p-value* (p-value = 0.189) was calculated by 2 sample t-test assuming equal variances. **B.** The number of putative somatic LINE-1 insertions identified in the temporal cortex and blood DNA of HA individuals and AD patients separately using TEBreak. The number of putative somatic LINE-1 insertions was higher in the blood than in the temporal cortex of HA people (HA, average no. temporal cortex 0.5; average no. blood 7.4, p-value = 0.09). The number of putative somatic LINE-1 insertions was higher in the temporal cortex than in the blood of AD patients (AD, average no. temporal cortex 8.6; average no. blood 7.0, p-value = 0.68). *p-values* were calculated by 2 sample paired t-test (healthy aged n=11, AD n=5).

4.1.4.4 Haplotype block analysis reveals enrichment of putative somatic L1 insertions from healthy aged individuals in cognitive function associated haploblocks

In order to define AD and healthy cognition (as a measure of fluid intelligence and vocabulary ability) associated haploblocks, we used the bedtools Intersect in Galaxy to find overlapping regions between Berisa *et al.* haplotype block division of the human genome bed file, and AD GWAS and fluid intelligence/vocabulary ability GWAS bed files, respectively. Subsequent analysis involved using the bedtools Intersect in Galaxy to elucidate whether polymorphic and/or putative somatic L1 insertions from HA and AD individuals are enriched in AD or cognitive function haploblocks. The original number of polymorphic and putative somatic L1 insertions is different in HA (polymorphic n=429; somatic n=85) and AD (polymorphic n=275; somatic n=80) individuals. When looking at the percentage of polymorphic and putative somatic L1 insertions in AD/cognitive function associated haploblocks (Table 4.1.1), our data demonstrates that putative somatic L1 HA insertions are enriched in fluid intelligence (4.71 %) and vocabulary ability haploblocks (9.41 %) compared to putative somatic L1 AD insertions (2.5 % and 1.25 %, respectively). This suggests that a specific putative somatic L1 genetic signature is present in healthy cognitive ageing. However, because validation of putative somatic L1 insertions was not possible using nested PCR, this can only be a hypothesis subject to corroboration of legitimate insertions.

Table 4.1.1. Non-reference LINE-1 RIPs in AD, fluid intelligence and vocabulary ability associated haplotype blocks.

Health variable	No. of Haploblocks	Intersect with	Polymorphic L1 insertions in haploblocks			Putative somatic L1 insertions in haploblocks		
			No. of haploblocks	No. of insertions	%	No. of haploblocks	No. of insertions	%
AD	495	AD insertions	80	275	29.09	20	80	25.00
		HA insertions	134	429	31.24	25	85	29.41
Fluid Intelligence	98	AD insertions	16	275	5.82	2	80	**2.5**
		HA insertions	20	429	4.66	4	85	**4.71**
Vocabulary Ability	104	AD insertions	22	275	8.00	1	80	**1.25**
		HA insertions	26	429	6.06	8	85	**9.41**

4.1.4.5 The likelihood of non-reference L1 insertion polymorphisms being intragenic is the same in healthy aged people and Alzheimer's disease patients

In order to assess whether non-reference L1 insertion polymorphisms were intragenic, using TEBreak's output, we selected all values other than NA from the *GeneRegion* column, which included all insertions that were located within genes. The likelihood of non-reference L1 RIPs being intragenic is very similar in HA people and AD patients (Figure 4.1.5).

Analysis of putative somatic L1 insertions in reference to genes was also carried out (data not shown). As with L1 RIPs, the likelihood of them being intragenic is very similar in HA people and AD patients.

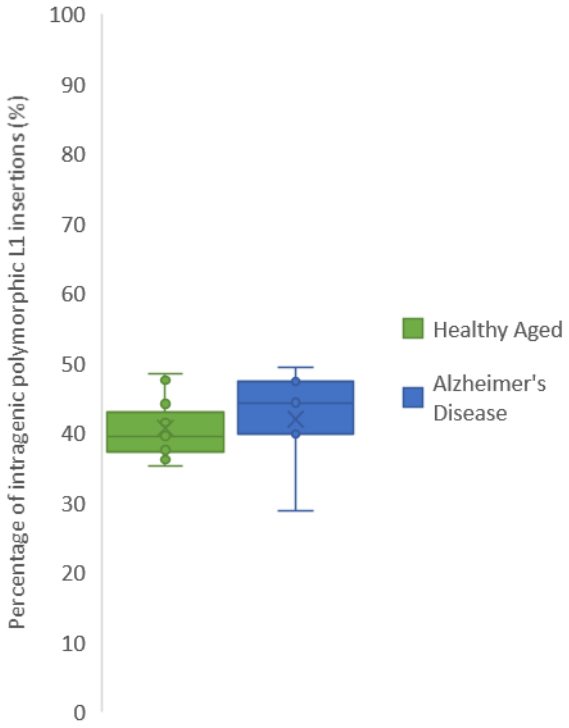


Fig. 4.1.5. The distribution of non-reference LINE-1 RIPs identified by retrotransposon capture sequencing (RC-Seq). The percentage of non-reference L1 RIPs in each individual that are intragenic. There is no significant difference of the frequency of intragenic L1 RIPs in those with AD compared to healthy aged people (average %, HA – 40.7, AD – 42.0). 2 sample t-test assuming equal variances, p-value = 0.69; healthy aged n=11, AD n=5).

4.1.4.6 Polymorphic L1 insertions from RC-Seq occur in genes expressed in the brain

As the likelihood of polymorphic insertions being intragenic was similar in HA people and AD patients, we generated two separate gene lists of polymorphic and putative somatic L1 insertions from either HA people or AD patients. Using either of the four gene lists as input for DAVID v 6.8 and the following categories – KEGG pathway, GO Biological processes, GO cellular component, GO molecular function and UP tissue, we assessed whether our genes of interest associated with differential major biological pathways.

A bias of the data is that gene lengths are not considered in DAVID analyses. As LINE-1 retrotransposition occurs randomly in the genome [209], gene size can influence gene enrichment analyses. In any case, due to either preferential targeting of insertions or gene length, our data suggests that genes harbouring polymorphic L1 insertions from HA people are predominantly located to cell junctions and are significantly expressed in the brain (Table 4.1.2). Analysis was also carried out for putative somatic L1 insertions from HA people (data not shown), but there were no pathways that were significantly enriched. Our data suggests that genes harbouring polymorphic L1 insertions from AD patients are predominantly located to the cytoskeleton and are significantly expressed in the brain (Table 4.1.3). Analysis was also carried out for putative somatic L1 insertions from AD people (data not shown), suggesting that genes harbouring putative somatic L1 insertions from AD patients are predominantly located to synapse.

Table 4.1.2. Data with statistical significance from DAVID pathway analysis of genes harbouring polymorphic L1 insertions from HA people

Category	Term	Count	%	Fold Enrichment	Bonferroni
GOTERM_CC_DIRECT	**Cell junction**	**12**	**9.0**	**3.9**	**0.038**
GOTERM_CC_DIRECT	Adherens junction	5	3.8	15.2	0.057
GOTERM_CC_DIRECT	Postsynaptic membrane	8	6.0	5.8	0.084
GOTERM_CC_DIRECT	Synapse	7	5.3	5.9	0.202

Category	Term	Count	%	Fold Enrichment	Bonferroni
UP_TISSUE	**Brain**	**77**	**57.9**	**1.4**	**0.006**
UP_TISSUE	Trachea	10	7.5	3.9	0.097
UP_TISSUE	Fetal brain	13	9.8	2.5	0.457
UP_TISSUE	Skeletal muscle	9	6.8	2.3	0.984

Table 4.1.3. Data with statistical significance from DAVID pathway analysis of genes harbouring polymorphic L1 insertions from AD people

Category	Term	Count	%	Fold Enrichment	Bonferroni
GOTERM_CC_DIRECT	**Cytoskeleton**	**9**	**9.7**	**5.5**	**0.034**
GOTERM_CC_DIRECT	Adherens junction	4	4.3	17.9	0.196
GOTERM_CC_DIRECT	Dendrite	7	7.5	4.7	0.429
GOTERM_CC_DIRECT	Cell junction	8	8.6	3.9	0.470

Category	Term	Count	%	Fold Enrichment	Bonferroni
UP_TISSUE	**Brain**	**53**	**56.9**	**1.4**	**0.048**
UP_TISSUE	Trachea	6	6.5	3.5	0.917
UP_TISSUE	Epithelium	20	21.5	1.6	0.966

4.1.5 Discussion

To date, LINE-1 elements are the only autonomous non-LTR retrotransposons that remain active in the human genome [72, 73, 186, 210]. These elements are a great source of human genome diversity and, there is evidence of insertion occurrence in neural cell lines giving rise to polymorphic and somatic mutations [198, 211]. Detection of the latter is technically very challenging because of the higher abundance of fixed insertions to the human genome [88]. RC-Seq, a targeted sequencing approach, was developed to overcome this practical challenge [69].

In this chapter, using temporal cortex and blood tissue as starting material for RC-Seq, we assessed the polymorphic and somatic L1 insertion scenery in 11 HA individuals and 5 AD patients in reference to number and location of insertions. Research on the field has often attributed a deleterious nature to increased copy number of LINE-1 [53, 212, 213], which is a hallmark of ageing [135]. Here, using a comprehensive approach, we interrogated the output from TEBreak bioinformatic analysis to compare the number of polymorphic and putative somatic insertions between HA individuals and AD patients. In this context, we found that the number of polymorphic L1 insertions was on average higher in HA than in AD individuals (Figure 4.1.2), but not vastly different. In addition, further interrogation of the data to assess the putative somatic L1 landscape in HA and AD individuals demonstrated a tendency towards a higher number of putative somatic L1 insertions in the temporal cortex of AD patients compared to HA individuals (Figure 4.1.4B). Our data suggests that despite previous literature supporting the idea that a tsunami of insertions occurs in malignancy [133], the number of L1 insertions it is not vastly

different between HA and AD and hence, there is no such an increase of insertions. Previous studies using RC-Seq have identified tissue-specific L1 insertions directly associated with malignancies including but not limited to ovarian tumour [214-216]. These have been more focussed on identifying specific *de novo* L1 elements that may play a major role in the development of malignancies [214-216]. In future, using a more targeted approach and with the curated data already available, we could assess full length *de novo* insertions in more detail to elucidate which, if any, AD specific insertions might be the offspring of hot RC-L1 source elements. Haplotype block analysis (Table 4.1.1) revealed enrichment of putative somatic L1 HA insertions in fluid intelligence and vocabulary ability haploblocks suggesting that the location of putative somatic L1 HA insertions is contributing to healthy cognitive ageing. In order to investigate further the genomic location of polymorphic and putative somatic L1 insertions, using TEBreak's output data, we calculated the percentage of polymorphic and putative somatic L1 insertions that are intragenic (Figure 4.1.5). In both cases (data for putative somatic L1 insertions not shown), less than 50 % L1 insertions were intragenic. Literature confirms that even if L1 elements can be found throughout the genome, preferential sites of insertion include AT-rich, low-recombining and gene deprived regions of the genome [217-219]. L1s are also enriched in sex chromosomes [220] and in low- and monoallelic-expression genes [221, 222]. Furthermore, the age and size of L1 elements plays a role in L1 genome location. Whereas older L1 elements are located on average further from genes than younger elements, full-length elements are more abundant on sex chromosomes [217]. Our data supported the fact that L1 elements are preferentially located in gene deprived regions of the genome. A more in depth bioinformatic analysis of the data will also allow either

supporting or refuting whether younger L1 insertions from this study are in fact located closer to genes, and if full-length L1 are enriched in sex chromosomes. Those polymorphic and putative somatic L1 insertions from the study that were intragenic, were located mainly in brain-expressed genes. In fact, Gianfrancesco 2018 [180] demonstrated that LINE-1 insertions are highly overrepresented at Gamma-Aminobutyric Acid (*GABA*) and glutamate family genes. Our data demonstrated that at least one polymorphic and somatic L1 insertion in HA and AD people was present within GABA Type B Receptor Subunit 1 (*GABR1*) and 3 (*GABR3*), and Glutamate Ionotropic Receptor Kainate Type Subunit 2 (*GRIK2*) genes. However, this data did not support the idea that L1 RIPs are enriched at *GABA* and glutamate family genes. This may be because Gianfrancesco's 2018 [180] analysis included L1PA2 and 3, older L1s which were not predominant in this analysis, and which were the main ones being overrepresented at *GABA* and glutamate genes in Gianfrancesco's 2018 analysis [180].

Our analysis is restricted by TEBreak limitations; therefore, with the already available sequencing data, and in order to compile a more informative approach, future studies using other software tools for mobile element insertion detection such as MELT [87] should also be carried out. It is important to note that even using RC-Seq, which has been designed to minimise artifacts by keeping the PCR cycles to a minimum, could give rise to illegitimate insertions. In particular, the false positive rate of somatic L1 insertions might be high because of the technical difficulty to identify these. Therefore, validation of more polymorphic and particularly putative somatic L1 insertions by nested/heminested PCR (i.e. increasing extension cycles from ˜30x to

~80x) and capillary sequencing is required before definite conclusions could be drawn from this analysis.

Mobilisation of non-LTR retrotransposons clearly contributes to mutational events as 400 million retrotransposon-derived structural variants are present in humans and over 70 diseases involve heritable and *de novo* retrotransposition events [58, 69]. However, the potential beneficial impact of L1 retrotransposition is yet to be explained. Our data supports the idea that the polymorphic and somatic signature of L1 insertions in the brain is an important factor to healthy cognitive ageing, demonstrating that further analysis of which specific insertions can present a beneficial impact in cognitive function during ageing is critical.

Chapter 4.2

The analysis of *Alu*, LINE-1 and SVA insertion polymorphisms in the context of ageing

4.2. The analysis of *Alu*, LINE-1 and SVA insertion polymorphisms in the context of ageing

4.2.1 Introduction

The mutagenic potential of TEs renders these as major contributors to the genetic diversity of the human genome and, they have the potential to function as regulatory elements, be beneficial or cause disease [5, 53]. About 2 to 5 new insertions per 10-100 live births have been attributed to active human TEs [76, 85], but this rate is constantly increasing as research on the field expands. As a result, the human genome consists of a combination of recent insertions along with fixed insertions [76]. Yet, detection of recently mobilised insertions is often problematic because of the large amount of endogenous elements in the genome [75]. Most recent insertions arise from endogenous human retrotransposons that remain active; hence, an approach for identifying recently mobilised transposons in the human genome has been to identify species-specific insertions by comparing the human and chimpanzee genomes [75]. Most species-specific elements and thus mobilised during the past 6 million years (the last common ancestor of humans and chimps) belong to *Alu*, L1 and SVA elements. Since the discovery of a *de novo* LINE-1 insertion as the cause of haemophilia A in 1988 as proof of active L1 retrotransposition in the human genome and to date, L1 elements are the only recognised active autonomous non-LTR retrotransposons and they are responsible for *Alu* and SVA retrotransposition [72, 73]. Therefore, not only LINE-1, but also *Alu* and SVA elements, which rely upon L1-encoded proteins for mobilisation [76], are of great interest to human evolution as these continue to trigger genetic diversity in the human genome and also have the potential to alter gene function [76].

De novo retrotransposition events, which are estimated to occur in roughly 1 of 20 live births for *Alu* elements, 1 of 150 for LINE-1 and 1 of 1000 for SVA elements, are an important source of genetic variation [68]. A human genome has on average 1283 *Alu*, 180 LINE-1 and 56 SVA presence/absence polymorphic insertions [5], genetic variants termed RIPs. In this context, L1 mobilisation and L1-mediated retrotransposition events lead to a significant amount of polymorphic and somatic insertional events in the brain [204]. To date, 124 different LINE-1-mediated insertions have been identified as the genetic cause of diseases. A few examples of these include a *de novo Alu* insertion which results in neurofibromatosis type 1 [223], the above mentioned *de novo* LINE-1 insertion in haemophilia A [203] and the expansion of a hexameric repeat within an SVA retrotransposon insertion in TAF1 which correlates with X-linked dystonia-parkinsonism' onset [224].

High-throughput sequencing has been applied to the study of RIPs [225]. Because of the nature of most sequencing data available to date, which is often generated on Illumina platforms and consists of 100-150 bp reads in pairs, detection of structural variants by comparing to the reference genome must be inferred from these short sequences that generally do not span the whole region affected [85]. To overcome the associated limitations, a three way approach whereby discordant read-pair mappings, clustering of split reads sharing common alignment junctions and sequence assembly and re-alignment of assembled contigs is generally used for structural variant detection from short pair-end read data [85]. Furthermore, the repetitive nature of TEs and the thousands of elements scattered across the human genome confounds detection of insertions [85]. Therefore, the identification of

legitimate non-LTR insertions from sequencing data is further supported by TSDs sequences which are the resulting signature of TPRT [110] or transduced sequences that arise from carrying over of flanking regions during retrotransposition [186]. While both polymorphic and somatic insertions have the potential to alter an individual's phenotype [69], the latter specifically have been proven difficult to identify due to the often small number of cells carrying the insertion in a heterogeneous cell population and the requirement for costly high-depth sequencing [201, 202]. Despite the promising nature of RC-Seq for sequence identification and genome localisation of LINE-1 insertions [88], because of the technical challenge encountered for validation of putative somatic insertions and the requirement to assess *Alu,* L1 and SVA RIPS insertions in a way that minimises cost and time, we decided to use WGS as an alternative/comparative approach to RC-Seq. Previous literature suggests that the detection of insertions by different approaches, often yields a substantial overlap, but an equally if not more substantial expanse of non-overlap [83]. Therefore, by taking a different approach to detection of polymorphic L1 insertions, and further identification of *Alu* and SVA insertions, we can encompass a broader landscape of putative RIPs. In addition, a comparative approach between methods whereby we can check the amount of L1 RIPs overlap between the two is necessary to establish which, if either is more appropriate for insertion detection.

In this study, using temporal cortex (target tissue) and blood (matched tissue) genomic DNA as starting material for WGS library preparation, we were able to identify *Alu*, LINE-1 and SVA RIPs in 9 HA people and 5 AD patients. Following a very similar bioinformatic analysis to that used to detect L1 insertions alone from RC-Seq

libraries, we interrogated the sequencing data in order to determine whether a specific genetic signature of *Alu*, LINE-1 and SVA RIPs is present in HA people and AD patients. Furthermore, we examined the data to assess the location of RIPs in reference to genes. Our results demonstrate that the number of polymorphic *Alu*, L1 and SVA insertions is not vastly different in HA versus AD individuals and support previous findings that intragenic RIPs locate to genes predominantly expressed in the brain [69] independently of health status. Furthermore, an approach to comparing RC-Seq versus WGS data demonstrates that trends for polymorphic L1 insertions are the same across methods and that most of the L1 insertions identified by RC-Seq are also detected by WGS.

4.2.2 Aims

To use temporal cortex and matched blood DNA of healthy aged and Alzheimer's disease individuals in order to establish whether the number and/or location of polymorphic *Alu*, LINE-1 and SVA insertions correlated with healthy cognitive ageing using:

- TEBreak insertion detection software to analyse WGS data to determine the number, genome location and sequence of polymorphic *Alu*, LINE-1 and SVA insertions
- Bioinformatics to assess the location of identified *Alu*, LINE-1 and SVA insertions
    - with regards to genes
    - with regards to AD, fluid intelligence and vocabulary ability haploblocks
- DAVID to carry out pathway analysis in order to analyse if *Alu*, LINE-1 and SVA insertions occur in genes that trigger significant enrichment of specific pathways

To compare LINE-1 insertions detected by WGS with those from RC-Seq in order to assess differences of bioinformatic methods for TEs analysis

4.2.3 Methods

Whole Genome Sequencing (WGS) of temporal cortex and blood DNA at 40x depth was carried out externally at the Australian Genome Research Facility (AGRF) by providing 1 µg of genomic DNA per sample in a 96-well plate.

4.2.3.1 Bioinformatic analysis of WGS libraries to detect non-reference *Alu*, LINE-1 and SVA insertion polymorphisms

A step by step guide to bioinformatic analysis of WGS libraries is described in section 2.2.14. Bioinformatic scripts are detailed on Appendix 5. As for LINE-1 libraries from RC-Seq, TEBreak (https://github.com/adamewing/tebreak) was the software of choice for detection of non-reference polymorphic *Alu*, LINE-1 and SVA insertions. The output from TEBreak is in the form of a table, which we open as an excel file and custom filter in order to assess different parameters. Such parameters include, but are not limited to, the number of non-reference *Alu*, LINE-1 and SVA RIPs, the location of insertions such as within genes or regulatory regions, and the specific tissue the insertion is present on. An example is shown in figure 4.2. Following bioinformatic analysis to resolve the sequencing data, haplotype block analysis and pathway analysis were used to further interrogate the data.

4.2.3.2 Haplotype block analysis to characterise non-reference *Alu*, LINE-1 and SVA RIPs distribution

In order to identify whether non-reference polymorphic *Alu*, LINE-1 and SVA insertions were enriched in specific regions of the human genome such as AD risk loci or regions associated with cognitive function such as fluid intelligence or vocabulary ability, we assessed the genome using the same haplotype block bed files previously generated for RC-Seq haplotype block analysis.

Using Galaxy for assessing the haplotype block enrichment of polymorphic Alu, LINE-1 and SVA insertions from HA and AD individuals

The bedtools Intersect intervals from Galaxy (https://usegalaxy.org/) was used to find overlapping intervals between non-reference polymorphic *Alu*, LINE-1 and SVA HA/AD insertions bed files and AD, fluid intelligence and vocabulary ability haplotype block bed files.

4.2.3.3 Using DAVID for pathway analysis

Pathway analysis was completed using DAVID (https://david.ncifcrf.gov/). A list of genes that contained non-reference *Alu*, LINE-1 and SVA RIPs either from HA or AD individuals was generated from TEBreak's output and used as DAVID's input. In order to understand the major biological pathways associated to our input gene list, we used default parameters for DAVID pathway analysis, and interrogated KEGG pathway, gene ontology (GO) annotations including biological processes, molecular function and cellular component and UP Tissue categories.

4.2.4 Results

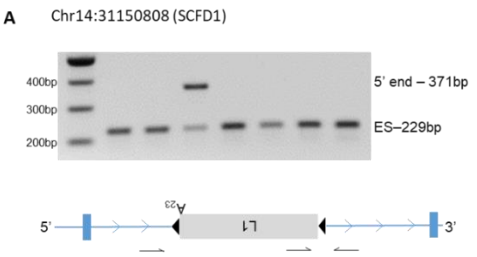
4.2.4.1 Validation of non-reference *Alu*, LINE-1 and SVA RIPs to corroborate accuracy, sensitivity and specificity of the parameters used for bioinformatic analysis

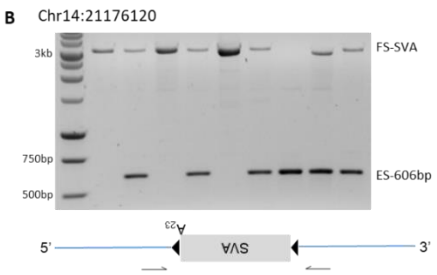
Previous to bioinformatic analysis of sequencing data from WGS libraries, our former post-doc researcher, Abigail L. Savage, performed an extensive amount of PCR validation (Figure 4.2.1) on non-reference *Alu*, LINE-1 and SVA RIPs detected using TEBreak on WGS in order to validate the accuracy, sensitivity and specificity of TEBreak to detect these with custom parameters. The custom values for the parameters defined on the previous chapter used for the filtering step of bioinformatic analysis of WGS libraries are below:

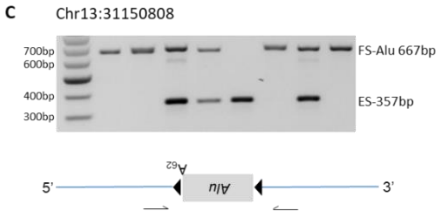
- Split read was set to 4
- Disc read was set to 4
- Conslen was set to 150 bp
- Eltmatch was set to 0.90
- Refmatch was set to 0.95

PCR validation of RIPs detected using TEBreak with the above-mentioned filtering parameters showed that the false positive rate of insertions was zero, meaning that we can trust polymorphic insertions from TEBreak bioinformatic analysis of WGS data to be hypothetically true. The false negative rate was a bit higher (0.04 for *Alu*, 0.21 for LINE-1 and 0.35 for SVA) meaning that TEBreak bioinformatic analysis of WGS data misses true insertions for some individuals. Whereas PCR validation of RIPs demonstrates a high level of specificity of TEBreak for *Alu*, LINE-1 and SVA insertion polymorphisms from WGS data, it shows a lower sensitivity for detecting LINE-1 and

SVA insertion polymorphisms as these are larger elements and present a longer string of repetitive DNA becoming more difficult to identify than *Alu* elements which are much shorter.







D

	Alu	LINE-1	SVA
False positive rate	0.00	0.00	0.00
False negative rate	0.04	0.21	0.35
Accuracy	0.98	0.91	0.87
Sensitivity	0.96	0.79	0.65
Specificity	1.00	1.00	1.00

Fig. 4.2.1. PCR validation of non-reference RIPs identified in WGS data using TEBreak. A. Multiplex PCR validation of full length L1 insertion in an intron of the SCFD1 gene using primers flanking the insertion site to generate an empty site product (229 bp) and a primer in the 5'end end of the L1 sequence to generate a PCR product (371 bp) if L1 is present. **B.** PCR validation of an SVA insertion into chr14 using one primer pair to amplify the empty and filled sites. Empty site PCR product was 606 bp; however, the filled site product's size cannot be predicted by TEBreak due to their large size and variability in length between elements due to their VNTR. **C.** PCR validation of an *Alu* insertion into chr13 using one primer pair to amplify the empty site (357 bp) and the filled site (667 bp). **D.** A total of 147 *Alu*, 117 LINE-1 and 71 SVA RIP validation PCRs were carried out for 12 different *Alu*, 11 different L1 and 6 different SVA RIPs in a minimum of 9 individuals each whose genomes had been analysed using TEBreak. ES – empty site, FS – filled site. The ability of WGS to detect the RIPs in these individuals was assessed by using the following calculations:

True positive (TP) = present in TEBreak and PCR analysis

True negative (TN) = absent in TEBreak and PCR analysis

False positive (FP) = present in TEBreak and absent in PCR analysis

False negative (FN) = absent in TEBreak and present in PCR analysis

False positive rate (FPR) = FP/(FP+TN); False negative rate (FNR) = FN/(FN+TP)

Accuracy = (TP+TN)/(TP+TN+FP+FN); Sensitivity = (TP/(TP+FN); Specificity = (TN/(TN+FP)

WGS libraries were generated for 14 individuals, 9 HA and 5 AD patients. We run TEBreak in all sequenced libraries. All scripts run for bioinformatic analysis of WGS libraries are detailed on Appendix 5. The output from TEBreak was a tab-delimited file of non-reference *Alu*, LINE-1 and SVA RIPs.

4.2.4.2 There is a trend towards the number of non-reference *Alu* insertion polymorphisms being on average higher in healthy aged people than in Alzheimer's disease patients

Using TEBreak's output, polymorphic *Alu*, LINE-1 and SVA insertions were those that TEBreak reported as being present both in the temporal cortex and blood genomic DNA. This was done by selecting the value 2 from the *Sample_count* column, where the value can be set to 1 or 2 depending on whether bioinformatic analysis found the insertion in 1 or 2 of the input tissues. Those insertions that were present in one tissue only, but previously reported in the literature were also established as polymorphic. There were no insertions detected in one tissue only and previously not reported, so WGS data was not a good source for somatic insertion detection. Following data curation, the number of *Alu*, LINE-1 and SVA RIPs per individual stratified by health status was counted and graphed in figure 4.2.2. Despite no significant differences in the number of *Alu*, LINE-1 and SVA polymorphisms between HA people and AD patients; our data demonstrates a trend towards the number of *Alu* insertion polymorphisms being on average higher in HA people than in AD patients.

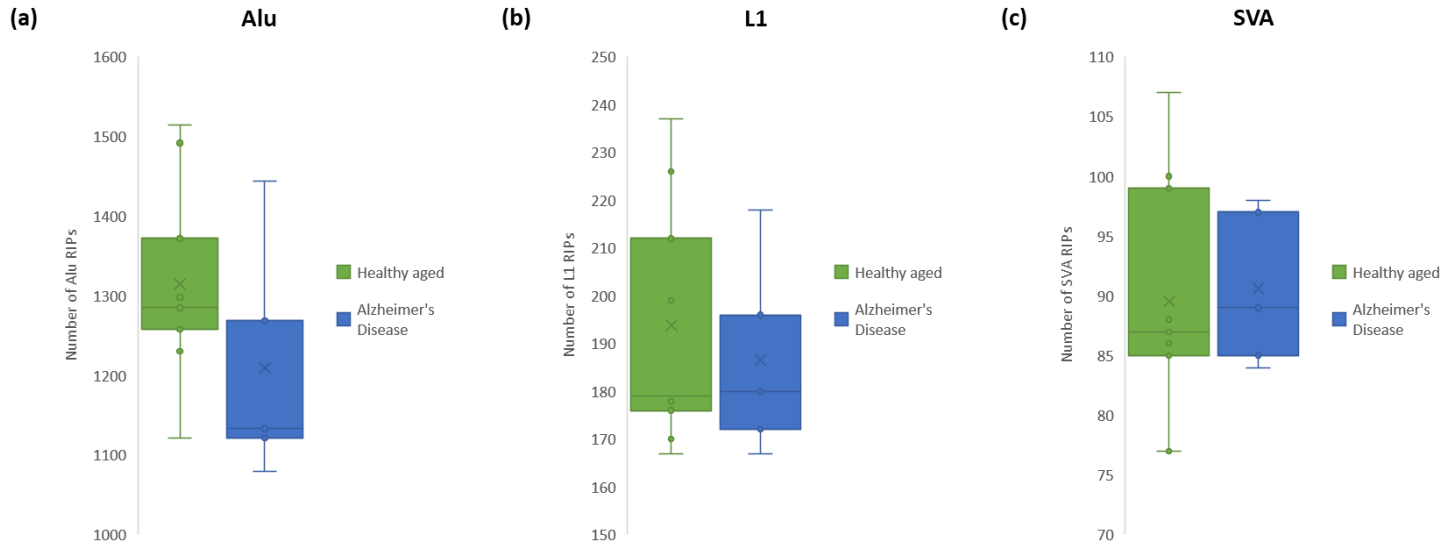


Fig. 4.2.2. The number of non-reference retrotransposon insertion polymorphisms identified in whole genome sequencing data using TEBreak. The number of *Alu* (a), LINE-1 (b) and SVA (c) RIPs identified in the genome of each individual using TEBreak. There is no significant difference in the number of RIPs in those with AD compared to healthy aged controls. The number of *Alu* RIPs in HA people (average no. of insertions: 1314.6) is higher than in AD patients (average no. of insertions: 1209.4), p-value = 0.185. The number of LINE-1 RIPs in HA people (average no. of insertions: 193.8) is ever so slightly higher than in AD patients (average no. of insertions: 186.6), p-value = 0.605. The number of SVA RIPs in HA people (average no. of insertions: 89.6) is very similar to the number in AD patients (average no. of insertions: 90.6), p-value = 0.843. *p-values* were calculated by two sample t-test assuming equal variances. In total, the number of different non-reference RIPs across the 14 individuals were as follows 3642 *Alu*, 587 L1 and 363 SVAs (HA n=9, AD n=5).

Furthermore, we compared the trend for LINE-1 elements from WGS libraries to that from RC-Seq libraries. In both cases, there is a trend whereby the number of polymorphic L1 insertions is on average higher in HA than in AD individuals, but not drastically different (Figure 4.2.3). Moreover, the number of identified LINE-1 insertions is slightly higher by analysis of WGS data than RC-Seq data (Figure 4.2.4). In fact, up to 88.6 % of L1 RIPs detected by RC-Seq overlap with those detected by WGS. As aforementioned, different approaches to TEs detection, often yield different results. Because the false positive rate from WGS for LINE-1 insertions was zero, our data suggests that RC-Seq misses true polymorphic insertions, though it is more stringent for avoiding false negative insertions.

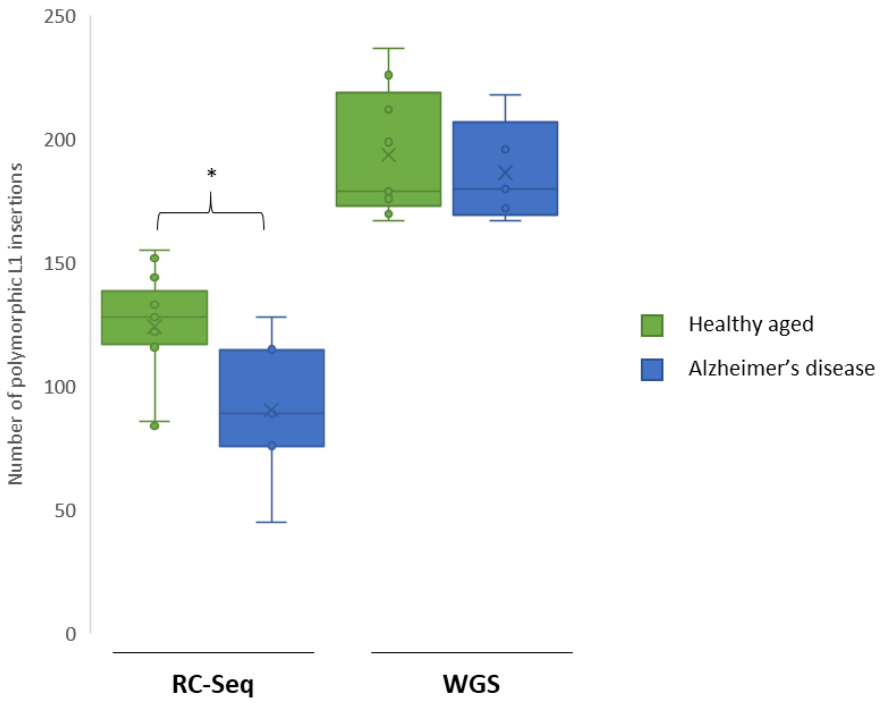


Fig. 4.2.3. The number of non-reference LINE-1 retrotransposon insertion polymorphisms identified in retrotransposon capture sequencing compared to whole genome sequencing data using TEBreak. The number of LINE-1 RIPs in HA people is on average higher than in AD patients when looking at LINE-1 insertions from RC-Seq and WGS, though not drastically different (RC-Seq – HA n=11, AD n=5; WGS – HA n=9, AD n=5).

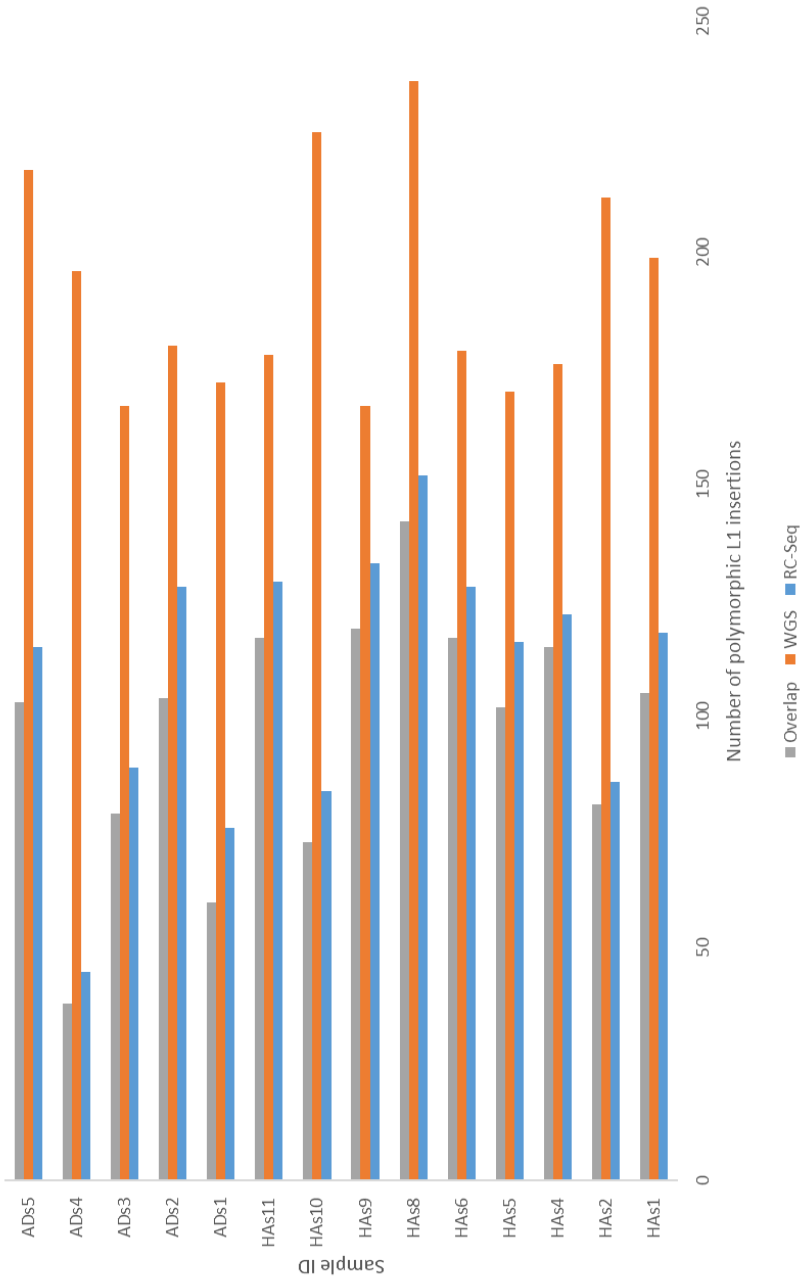


Fig. 4.2.4. The overlap of the number of LINE-1 non-reference retrotransposon insertion polymorphisms between retrotransposon capture sequencing and whole genome sequencing data using TEBreak. The number of LINE-1 non-reference retrotransposon insertion polymorphisms is graphed per sample. In blue, the number of L1 insertions from RC-Seq data. In orange, the number of L1 insertions from WGS data. In grey, the number of L1 insertions overlapping from the two data sets. An average of 88.6 % of L1 insertions from RC-Seq data overlap with insertions from WGS data (RC-Seq – HA n=11, AD n=5; WGS – HA n=9, AD n=5).

4.2.4.3 Haplotype block analysis reveals enrichment of non-reference SVA insertion polymorphisms from Alzheimer's disease individuals in AD associated haploblocks

AD and healthy cognition associated haploblocks used on the previous chapter for analysis of LINE-1 insertions detected by TEBreak from RC-Seq were also used for haplotype block analysis of *Alu*, L1 and SVA insertions from WGS libraries. Analysis was carried out using the bedtools Intersect in Galaxy to elucidate whether non-reference *Alu*, L1 and SVA RIPs from HA and AD individuals are enriched in AD or cognition haploblocks. The number of insertions detected is different in HA and AD individuals for *Alu*, L1 and SVA insertions. When looking at the percentage of polymorphic insertions in AD/cognitive function associated haploblocks (Table 4.2.1), our data suggests the number of AD SVA insertions has a tendency of being enriched in AD haploblocks (>2 % difference). Because of the small difference (>2 %) and the small n number (14 individuals), this is only preliminary data that suggests SVAs as putatively important for AD.

Table 4.2.1. The number of polymorphic *Alu*, LINE-1 and SVA RIPs in AD, fluid intelligence and vocabulary ability associated haplotype blocks

Case ID	No. of Haploblocks	Intersect with	*Alu* RIPs in haploblocks			L1 RIPs in haploblocks			SVA RIPs in haploblocks		
			No. of haploblocks	No. of insertions	%	No. of haploblocks	No. of insertions	%	No. of haploblocks	No. of insertions	%
AD	495	AD insertions	808	2593	31.2	130	421	30.9	71	233	**30.5**
		HA insertions	964	3178	30.3	147	490	30.0	83	292	**28.4**
Fluid Intelligence	98	AD insertions	175	2593	6.7	25	421	5.9	19	233	8.2
		HA insertions	220	3178	6.9	28	490	5.7	27	292	9.2
Vocabulary Ability	104	AD insertions	173	2593	6.7	29	421	6.9	28	233	12.0
		HA insertions	226	3178	7.1	28	490	5.7	30	292	10.3

4.2.4.4 Non-reference SVA insertion polymorphisms are more frequently intragenic and in regulatory domains compared to LINE-1 and *Alu* as are *Alu* compared to LINE-1

Using TEBreak's output, we assessed whether non-reference *Alu*, L1 and SVA RIPs were intragenic and/or in regulatory domains (weak/strong enhancer, active/poised/weak promoters, insulator) independently of health status. In order to do this, we selected all values other than NA from the *GeneRegion* column, which included all insertions that were located within genes, and all values other than NA from the *RegAnnot* column, which included all insertions that were located in regulatory domains. The likelihood of polymorphic SVA insertions being intragenic and in regulatory domains is higher compared to L1 and *Alu* as is that of polymorphic *Alu* insertions compared to L1 when HA and AD are analysed together (Figure 4.2.5).

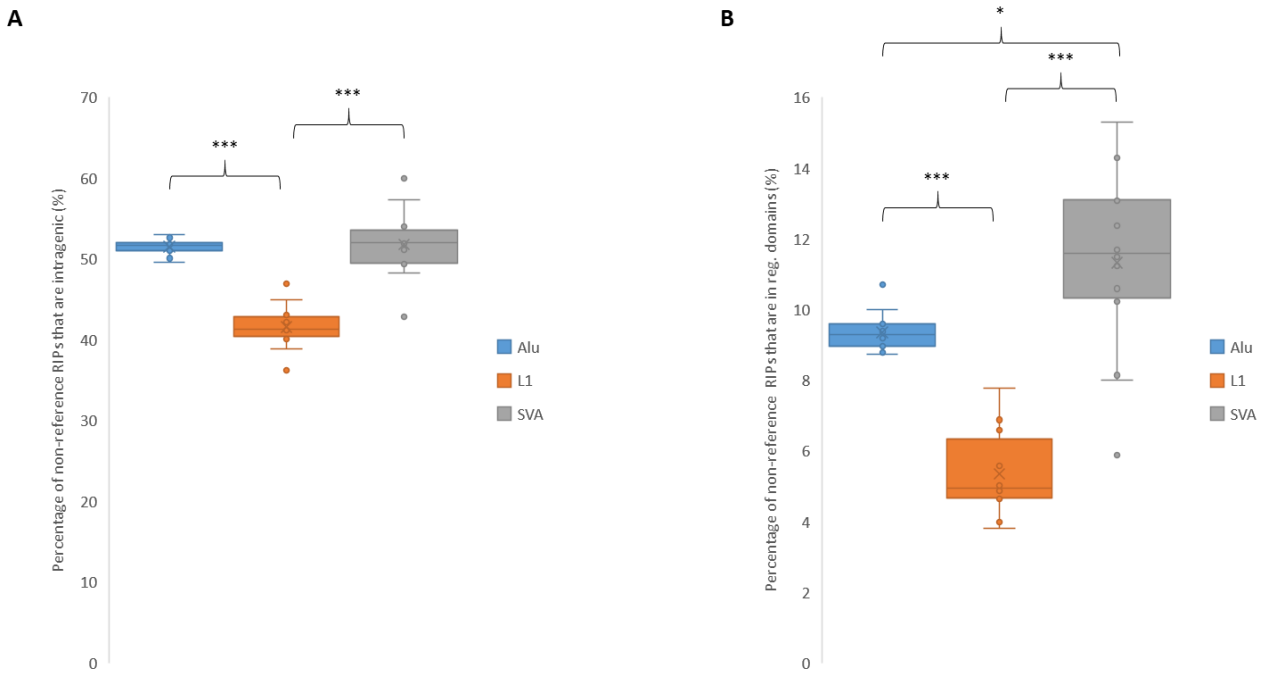


Fig. 4.2.5. The distribution of non-reference retrotransposon insertion polymorphisms identified in whole genome sequencing data using TEBreak. A. The percentage of *Alu*, L1 and SVA RIPs in each individual that are intragenic. SVA RIPs are more frequently intragenic compared to L1 (p-value= 4.5E-06) and *Alu* (p-value= 0.8) as are *Alu* RIPs compared to L1 (p-value=1.2E-09). **B.** The percentage of *Alu*, L1 and SVA RIPs in each individual that are located in regulatory domains. SVA RIPs are more frequently found in regulatory domains compared to *Alu* (p-value=0.02) and L1 (p-value=2.7E-06) RIPs as are *Alu* RIPs compared to L1 (p-value=7.0E-09).

In order to interrogate the data further, we stratified it by health status (Figure 4.2.6). The likelihood of *Alu* and SVA RIPs being intragenic is significantly higher in AD patients than in healthy aged people, whereas the likelihood of L1 insertions being intragenic is the same in HA people and AD patients. When comparing RC-Seq and WGS data in terms of likelihood of LINE-1 insertions being intragenic (Figure 4.2.7), the trend is the same. Finally, the likelihood of *Alu*, L1 and SVA RIPs being in regulatory domains is the same in HA people and AD patients (Figure 4.2.8).

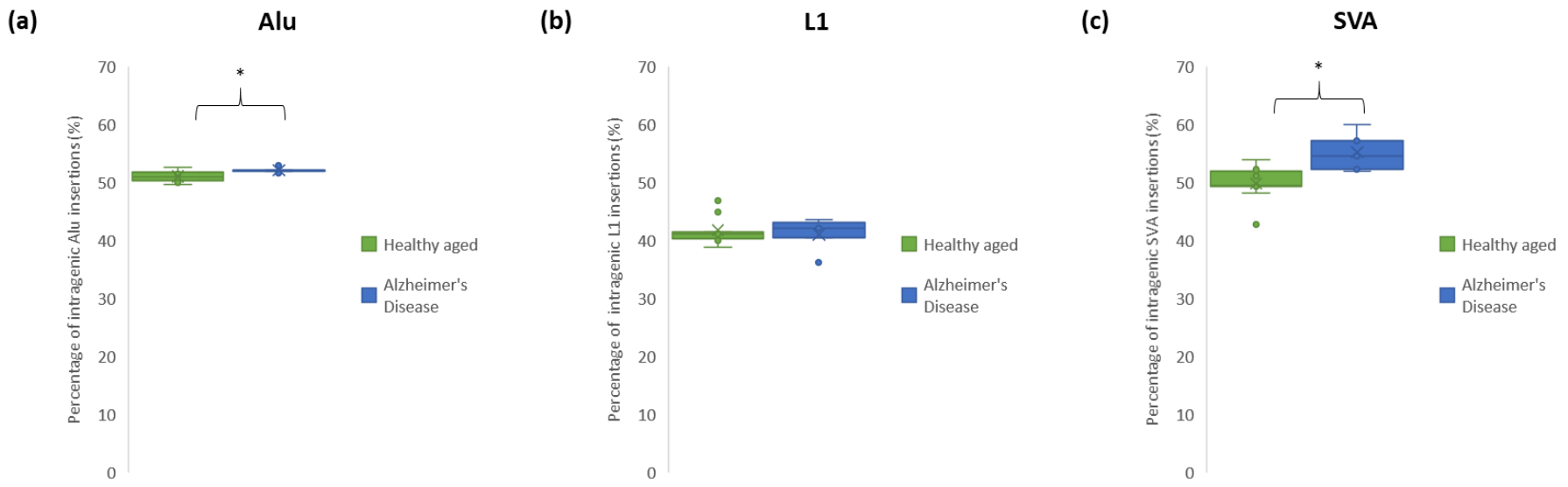


Fig. 4.2.6. The percentage of non-reference retrotransposon insertion polymorphisms identified in intragenic regions stratified by health status. The % of intragenic *Alu* (a), LINE-1 (b) and SVA (c) RIPs identified in whole genome sequencing data using TEBreak There is a significant difference of the percentage of intragenic RIPs of *Alu* (p-value= 0.04) and SVA (p-value=0.01) in those with AD compared to HA controls (HA n=9, AD n=5).

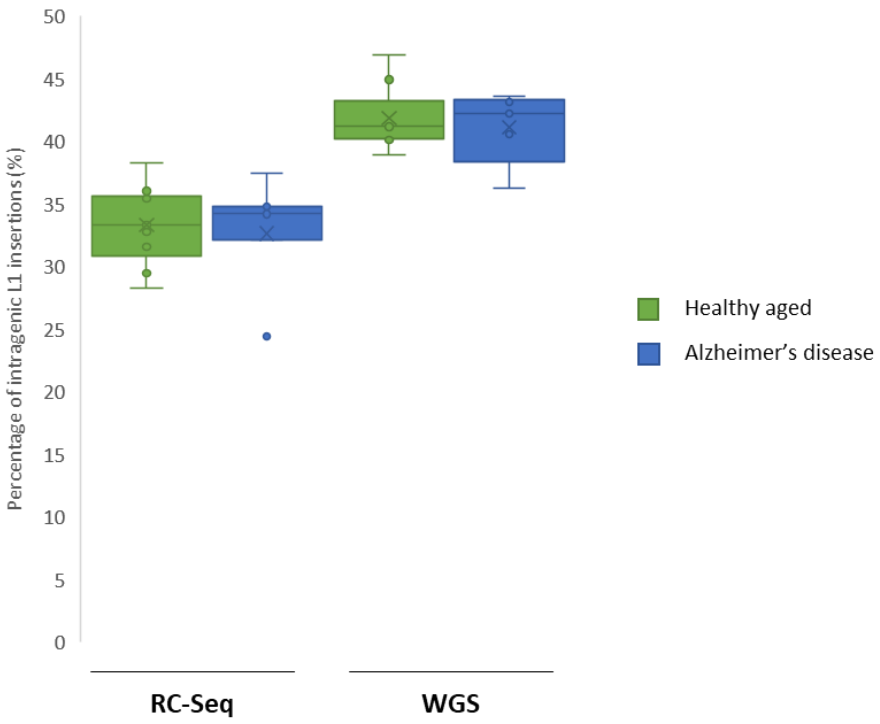


Fig. 4.2.7. The distribution of LINE-1 non-reference retrotransposon insertion polymorphisms by retrotransposon capture sequencing compared to whole genome sequencing. The percentage of L1 RIPs in HA and AD individuals that are intragenic. There is no significant difference of the frequency of intragenic L1 RIPs in those with AD compared to HA people (RC-Seq – HA n=11, AD n=5; WGS – HA n=9, AD n=5).

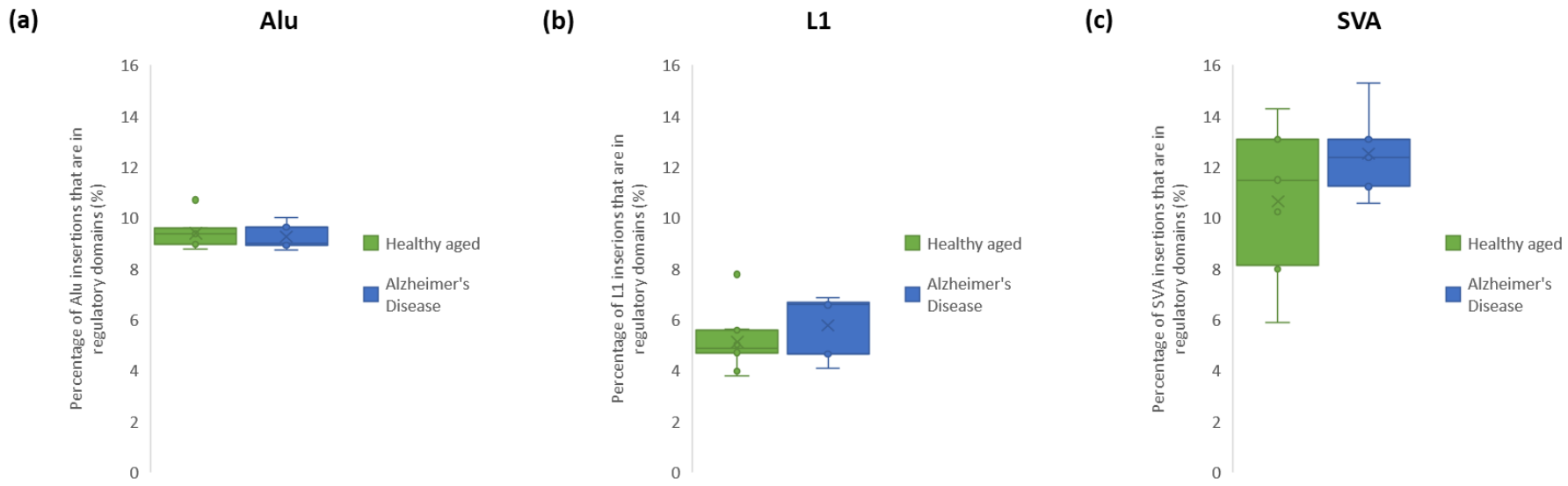


Fig 4.2.8. The percentage of non-reference retrotransposon insertion polymorphisms identified in regulatory domains stratified by health status. The % of *Alu* (a), LINE-1 (b) and SVA (c) RIPs in regulatory regions identified in whole genome sequencing data using TEBreak There is no significant difference of the percentage of RIPs located within regulatory domains of all three families (*Alu*, LINE-1 and SVA) in those with AD compared to HA controls (HA n=9, AD n=5).

4.2.4.5 Intragenic non-reference RIPs identified from WGS data occur in genes expressed in the brain

We generated two separate gene lists of non-reference *Alu*, LINE-1 and SVA RIPs from HA people and AD patients. Using either of the two gene lists as input for DAVID v 6.8 and the following categories – KEGG pathway, GO Biological processes, GO cellular component, GO molecular function, we assessed whether our genes of interest associated with differential major biological pathways. Note that all RIPs detected were considered together as the number of genes with SVA insertions alone is not sufficient for pathway analysis.

Again, bearing in mind that the data might be biased as gene lengths are not considered in DAVID analyses; our data from tissue expression suggests that genes harbouring polymorphic insertions from HA people are significantly expressed in the brain, hippocampus, epithelium and amygdala (Table 4.1.2). Those genes harbouring polymorphic insertions from AD patients are significantly expressed in the brain, hippocampus and amygdala (Table 4.1.3). In line with the tissue expression category, the rest of categories analysed (data not shown) presented a very similar pattern between HA and AD genes with RIPs, suggesting that when analysed all together, there is no biological pathways differentially enriched by genes harbouring polymorphic insertions neither in AD nor in HA individuals.

Table 4.2.2. Tissue expression data with statistical significance from DAVID pathway analysis of genes harbouring polymorphic insertions from HA people

Category	Term	Count	%	Fold Enrichment	Bonferroni
UP_TISSUE	**Brain**	**668**	**47.6**	**1.237014**	**1.00E-11**
UP_TISSUE	Hippocampus	56	3.9	1.833323	0.00483
UP_TISSUE	Epithelium	235	16.7	1.27586	0.010574
UP_TISSUE	Amygdala	67	4.77	1.665587	0.014297

Table 4.2.3. Tissue expression data with statistical significance from DAVID pathway analysis of genes harbouring polymorphic insertions from AD patients

Category	Term	Count	%	Fold Enrichment	Bonferroni
UP_TISSUE	**Brain**	**589**	**49.7**	**1.255089**	**1.09E-11**
UP_TISSUE	Amygdala	67	5.6	1.916588	1.37E-04
UP_TISSUE	Hippocampus	51	4.3	1.921243	0.003591

4.2.5 Discussion

Despite LINE-1 elements being the only recognised autonomous non-LTR retrotransposons that remain active in the human genome to date [72, 73, 186, 210], *Alu* and SVA elements, which rely on LINE-1 machinery for retrotransposition, [76] also remain mobile and are all great sources of human genome diversity. The human genome is a combination of endogenous elements and recent insertions that are polymorphic amongst individuals, the latter are named RIPs [76]. Because of their major contribution towards human evolution and disease, the detection of RIPs is crucial, but often difficult due to the large number of inactive fixed insertions in the genome [75]. Furthermore, the repetitive nature of non-LTR elements also complicates their identification using short-reads sequencing data. In particular, detection of somatic mutations is technically very challenging [88]. RC-Seq was developed to overcome this practical challenge [69]. In the previous chapter, we used LINE-1 enriched libraries generated using RC-Seq (average coverage at the 5´end – 602.8; at the 3´end – 239.8) to identify non-reference L1 insertions, both polymorphic and somatic. However, due the unsuccessful validation of putative somatic L1 insertions using nested PCR and the importance of *Alu* and SVA elements in human evolution [58], WGS at 40x depth was used for comparison as a more global approach for identifying non-reference *Alu*, L1 and SVA RIPs.

In this chapter, using temporal cortex and blood tissue as starting material for WGS library preparation, we assessed the number and location of non-reference *Alu*, LINE-1 and SVA insertion polymorphisms in 9 healthy aged individuals and 5 AD patients. Research on the field has often attributed a deleterious nature to *de novo* non-LTR

insertions [203, 223, 224]. Here, using a comprehensive approach, we interrogated the output from TEBreak's bioinformatic analysis to compare the number of RIPs between HA individuals and AD patients. In this context, we found that the number of *Alu* and L1 insertions was on average higher in HA than in AD individuals (Figure 4.2.2). A recent genome-wide scan found that a number of adaptive human polymorphic TE insertions are positively selected for and implicated in gene regulation. In fact, this study identified 3 *Alu*, 3 L1 and 1 SVA RIPs that have increased in frequency in specific human populations due to positive selection; five which are located in tissue-specific enhancers [226], suggesting *de novo* retrotransposition is not always deleterious. Our data supports the finding that polymorphic *Alu* and L1 elements can have a role in the context of HA; and, correlates with previous findings from RC-Seq when comparing across the two methods. However, it is important to notice that there is not a vast difference in the number RIPs between HA and AD as previously suggested in the literature [133]. An average human genome is estimated to have 1283 *Alu*, 180 LINE-1 and 56 SVA RIPs [5]. Using TEBreak to curate WGS data, we find the number of *Alu* insertions (average no. of insertions: HA – 1314.6, AD – 1209.4) as well as the number of LINE-1 insertions (average no. of insertions: HA – 193.8, AD – 186.6) is consistent with these expectations. However, the number of SVA insertions (average no. of insertions: HA – 89.6, AD – 90.6) is slightly higher than those expected. In fact, data from Abigail L. Savage (Figure 4.2.1) demonstrated that the false negative rate for SVA RIPs (0.35) was higher than that for *Alu* (0.04) and LINE-1 (0.21), respectively, meaning that TEBreak fails to detect true SVA RIPs detected otherwise by PCR. Therefore, our data suggests that both previous literature and our analysis using TEBreak may be underestimating the amount of non-

reference SVA RIPs in the genome. A recent paper on the field, suggests an increase in the SVA retrotransposition rate to 1 in 63 birds [86] likely due to the use of multiple mobile element insertion calling tools that allow more accurate insertion detection, and supporting the idea of the vast amount of missing *de novo* SVA insertions to date [86]. When addressing LINE-1s alone and comparing WGS to RC-Seq data, the number of LINE-1 insertions (average no. of insertions: HA – 193.8, AD – 186.6) from WGS libraries is more within the bulk of expected insertions than that from RC-Seq libraries (average no. of insertions: HA – 124.3, AD – 90.6), which is slightly lower than expected. In fact, an average of 88.6 % of L1 insertions from RC-Seq data overlap with insertions from WGS data (Figure 4.2.4). This would suggest the overlapping insertions to be legitimate. It is important to aim for a balance between missing true insertions (high false negative rate) and including illegitimate insertions (high false positive rate). Not only the method of choice, but also the parameters used for bioinformatic analysis play a major role on finding this equilibrium.

Further interrogation of the data by haplotype block analysis (Table 4.2.1) revealed enrichment of SVA AD insertions (30.5 %) when compared to SVA HA insertions (28.4 %) in AD associated haploblocks suggesting that the location of SVA insertions might be contributing to AD. In order to investigate further the genomic location of RIPs, using TEBreak's output data, we calculated the percentage of *Alu*, LINE-1 and SVA RIPs that are either intragenic or in regulatory domains (Figure 4.2.5). Whereas 50-60 % of SVA RIPs are intragenic, only 40-50 %. *Alu* and L1 RIPs are within genes. In addition, 10-15 % of SVA RIPs are located in regulatory domains, but 5-10 % *Alu* and L1 RIPs are in these loci. Our data suggests that previous literature focussing on L1s

could be misleading as not only there seems to be a missing SVA component, but SVA RIPs are more often within genes and regulatory domains; and hence, they may be more likely modulators of gene expression and regulatory elements [37]. Furthermore, pathway analysis using DAVID indicated that those RIPs identified from the study that were intragenic, were located mainly in brain-expressed genes, supporting the idea that TEs play a crucial role in the brain [69].

Despite the deleterious nature often associated to LINE-1-mediated insertions as the genetic cause of diseases [203, 223, 224], the potential beneficial impact of retrotransposition is yet to be elucidated. Previous studies have been focussed on endogenous/fixed retrotransposons as potential regulatory sequences. However, as explained and by definition, these are not the source of most human genetic diversity today [226]. Our data supports the idea that non-reference RIPs are an important factor to healthy cognitive ageing and AD, demonstrating that the study of RIPs as major contributors to genetic variation across individuals might help unravel the adaptive effects of active TEs [97, 98].

Chapter 5

Discussion

Chapter 5 Discussion

5.1. Thesis summary

The ultimate goal of the research presented in this thesis was to characterise the landscape of repetitive elements and present a genome-wide analysis of TE number and location in healthy cognitive ageing and AD. Due to historical technical limitations, and more recently because of the exclusion of repetitive DNA from analysis of sequencing data, TEs have not often been included in genetic studies and hence they are a largely overlooked source of genetic variation in the genome. In fact, TEs are often associated with a deleterious nature and so their potential beneficial role is often disregarded. More recently, by deepening our knowledge and through technological advances in the field, we have a better understanding of TEs modus operandi and there are emerging more accurate methods for TE insertion detection.

Cognitive function varies widely between individuals. Furthermore, higher cognitive ability during childhood and adolescence is generally associated with a tendency to stay longer in education, which in turn results in higher qualifications and better-paid jobs, and ultimately translates into living longer and in more desirable conditions [26]. Maintaining cognitive function throughout life is therefore important for HA, as the maintenance of good cognitive function is more important for wellbeing in the elderly than any onset of disease [139]. Both genetic and environmental factors contribute towards the aetiology of such a medically and socially important trait. The influence of environmental factors such as a balanced diet, moderate physical activity, managing stress and recurrent social activity are widely studied and well established. However, the genetic contributions remain poorly understood [3, 227].

It was not until the advent of large sample size GWAS that the first genetic associations with cognition and ageing were established [3, 4]. To date, GWAS have consistently associated the *APOE* locus significantly with cognitive changes [11]. *APOE* isoforms relate either positively [3, 111, 112, 151, 161], by contributing towards non-pathological ageing, or detrimentally [153-155, 157, 158, 160], by increasing the risk of neurodegenerative disorders such as AD [11]. In addition, genetic variation in a non-coding region of *TOMM40*, which is a gene adjacent to *APOE*, is also associated with AD [156-158], cognitive phenotypes in the elderly [153, 155, 159-161] and exceptional longevity [111, 162]. However, GWAS have failed to reproducibly establish further genetic associations to healthy cognitive ageing.

Despite TE's known role as regulatory elements, they are often disregarded in genetic studies due to their repetitive nature. In fact, the ability of the non-LTR retrotransposons to mobilize is critical for genome diversity and evolution [228]. Furthermore, this mobilisation, as well as during earlier stages of life, can occur in the adult and could increase in the elderly [116, 197, 198]. This increase in the elderly could correlate with the age of onset of neurodegenerative conditions. Ultimately, awry regulation of non-LTRs retrotransposition may be a causative effect or the result of the detrimental effects associated with normal ageing or simply support degeneration as a secondary mechanism [229]. Due to the deleterious nature often associated to these elements [53], the human genome has evolved ways to limit the amount of retrotransposition. Epigenetic mechanisms such as DNA methylation or histone modifications are the main limiters of retrotransposition by restraining the ability of non-LTR retrotransposons to mobilize, ultimately leading to transcriptional

silencing of retrotransposition events, though not necessarily blocking their genomic function [230]. As such, the inhibition of retrotransposition drivers such as hot RC-L1s may play a major role on preserving the integrity of the human genome [108]. This implies that a balance in the amount of retrotransposition is crucial to maintaining a healthy cognitive phenotype as the retrotransposon regulating mechanism seems to be awry during ageing and neurodegeneration.

In 2013 a study defined the nine hallmarks of ageing, including - but not limited to - genomic instability, telomere attrition, mitochondrial dysfunction, cellular senescence and epigenetic alterations [231]. The work in this thesis has attempted to deepen our knowledge on the contribution of repetitive DNA to these. Taken together, the data presented here suggests that TE variation could be involved in healthy cognitive ageing. Further, it suggests that not only could endogenous TE be a valuable source for deepening our understanding of the genetics of the phenotypically diverse trait that is ageing, but that active and newly integrating TE variants are potentially an even more valuable source of genetic variation as they greatly contribute to individual to individual differences. Whereas the environmental factors that influence healthy cognitive ageing have been extensively studied, at present, there is only a limited knowledge of the genetic basis of HA. The analysis of TEs, widely unexplored and overlooked in genetic analysis, is therefore likely to be crucial in unravelling the missing genetic basis of not only HA but also disease.

5.2. Concluding remarks

Throughout the work presented in this thesis, an overview was presented on the putative regulatory role, the epigenetic regulation and a genome-wide analysis of repetitive DNA in healthy cognitive ageing compared to AD. In order to do so, we have covered a wide variety of laboratory techniques from molecular biology such as PCR, to epigenetics such as bisulphite pyrosequencing to next generation sequencing including RC-Seq and WGS, and bioinformatics. For all these analyses, we used an immensely valuable resource named the Dyne Steele cohort, and the majority of the data has been generated around 16 elderly individuals, 11 HA and 5 AD. Table 5.1 presents a summary on the data available for each individual. In the pipeline, tagging SNPs, measuring telomere length and mitochondrial DNA (mtDNA) sequencing are briefly described as a proxy for future experiments to deepen our understanding of the genetics of healthy cognitive ageing.

Table 5.1. Data generated throughout the course of this project on the Dyne Steele cohort individuals. Highlighted in green is data available, in red non-available and in blue data that will be available in the future. Note that wherever data is not available for an individual is due to a lack of sufficient DNA at the time of analysis.

Case	Code	DNA available	DNAMet	PyroSeq	RC-Seq	WGS	mtDNA
09/24	HAs1	Temporal cortex and blood DNA					
09/26	HAs2						
09/31	HAs3						
11/06	HAs4						
11/07	HAs5						
11/22	HAs6						
11/29	HAs7						
14/04	HAs8						
14/46	HAs9						
15/01	HAs10						
15/28	HAs11						
10/07	ADs1						
15/11	ADs2						
16/03	ADs3						
16/09	ADs4						
16/13	ADs5						
11/25	HAs12	Temporal cortex DNA					
12/23	HAs13						
14/16	HAs14						
14/20	HAs15						
15/26	HAs16						
15/30	HAs17						
15/31	HAs18						
09/15	ADs6						
10/08	ADs7						
10/40	ADs8						
13/10	ADs9						
15/16	ADs10						

DNAMet – DNA methylation from CpG pulldown, which was used in VNTR and hot RC-L1s analysis in chapters 3.1 and 3.2
PyroSeq – Bisulphite and pyrosequencing, which was used in global L1 methylation analysis in chapter 3.2
RC-Seq – Retrotransposon capture sequencing, which was used in L1 analysis in chapter 4.1
WGS – Whole genome sequencing, which was used in *Alu*, L1 and SVA analysis in chapter 4.2
mtDNA – Mitochondrial DNA sequencing, which will be used to analyse mtDNA variants
HA – Healthy ageing; AD – Alzheimer's disease; s – Sample

Taken together, we have assessed the individual- and tissue- specific genetic variation considering repetitive DNA.

In chapter 3.1, a variable number tandem repeat (VNTR) located on the long arm of chromosome 20 was studied. VNTRs are an important source of genetic variation with previous research in the field consistently demonstrating the regulatory activity of VNTRs, which is often allele-specific and dependent on the number of tandem repeats in the element. The VNTR at the 20q13.3 locus is of particular interest because it comprises a hominoid-specific brain expressed miRNA named *MIR941*, and it is located within *DNAJC5*, which functions in neuroprotection [164-166]. In this chapter, we found evidence to support that *MIR941*/VNTR genotype is variable across an elderly population (Figure 3.1.4) and we found four common alleles. A genotype where each allele presented 10 copies of the VNTR was enriched in the elderly who presented a tendency towards more negative mental health symptoms and shorter survival rates. These two findings support the idea that mental health problems are often linked to shorter lifespan [184]. In addition, we demonstrated no genotypic association of the VNTR to healthy cognitive ageing or AD in this particular study. As the VNTR is a CGI, we also assessed the DNA methylation at the VNTR locus and found that the methylation of *MIR941*/VNTR is significantly higher in the blood than in the temporal cortex irrespective of health status (Figure 3.1.7). This is consistent with the fact that *DNAJC5* is expressed in the brain [169], and suggests shared transcription regulation of *DNAJC5* and the VNTR, but again demonstrates no evidence of association of the VNTR to disease risk. Finally, we found that that two

of the four *MIR941*/VNTR alleles acted as repressors of expression in this model (Figure 3.1.11) suggesting a regulatory role of the VNTR in this particular model.

In chapter 3.2, we considered hot RC-L1s, which are responsible for the majority of retrotransposition activity in the human genome. In fact, hot RC-L1s are vastly polymorphic and as such, each individual presents a different L1 make-up [73, 87, 186, 187]. In order to limit their mobile activity, host genomes have evolved a defence mechanism named DNA methylation [179]. Furthermore, ageing is characterised by methylation level changes, which may in turn be responsible for the increased copy number of L1 observed in the elderly [135]. In this chapter, using a similar approach to that of chapter 3.1, our data suggests that harbouring a greater number of L1 RIPs correlated to risk of AD in males. In addition, L1 RIP's presence was linked to a higher level of methylation both in the temporal cortex and in the blood regardless of health status (Figure 3.2.9B), suggesting that methylation changes in the genomic environment result from the L1 insertion itself. Further interrogation of the data suggested that the level of methylation of L1 elements is higher in blood than in temporal cortex (Figure 3.2.10B) irrespective of health status, suggesting that L1 elements are more active in the temporal cortex of elderly people.

In chapters 4.1 and 4.2, we assessed *Alu*, LINE-1 and SVA elements, which are non-LTR elements active in the human genome [58]. In fact, even if constantly changing, *de novo* insertions are estimated to occur in roughly 1 of 20 live births for *Alu* elements, 1 of 150 for LINE-1 and 1 of 1000 for SVA elements and are an important source of genetic variation [68]. The reference genome has on average 140,000 *Alu-Y*, 500,000 LINE-1 and 3,600 SVA annotated elements [83]. However, because of the

rate of insertion of these elements, there is a missing component of non-annotated elements named non- reference RIPs [62]. In chapter 4.1, using targeted sequencing (RC-Seq) for LINE-1s, we attempted to characterise polymorphic and somatic LINE-1 insertions, the latter technically very challenging to detect [88]. Our data demonstrated that the number of polymorphic L1 insertions was not vastly different in HA versus AD individuals (Figure 4.1.2). Further, the data presented in this chapter supports the notion that LINE-1 preferentially insert in gene-deprived regions of the genome (Figure 4.1.5) [217]. In fact, those LINE-1 insertions that occurred in genes in this study were located mainly in brain-expressed genes. In chapter 4.2, we used WGS to study the whole complement of non-LTR elements named *Alu*, LINE-1 and SVA. Our data from WGS on LINE-1 correlated with that from RC-Seq. In the context of *Alu*, LINE-1 and SVA elements, we found that SVA RIPs are more frequently intragenic and in regulatory domains compared to LINE-1 and *Alu* (Figure 4.2.5), suggesting that despite these elements relying on LINE-1 machinery for mobilisation, they also have a pronounced impact on human genome regulation. In addition, pathway analysis demonstrated that those RIPs identified from the study that were intragenic, were located mainly in brain-expressed genes, supporting the idea that TEs play a crucial role in the brain [69]. Taken together, our data supports the idea that non-reference RIPs in the brain are an important factor to consider in healthy cognitive ageing, demonstrating that further analysis of which specific insertions can present a beneficial impact in cognitive function during ageing is critical.

5.3. In the pipeline

5.3.1 Development of tagging SNPs for hot RC-L1s genotyping

The analysis of the hot RC-L1s genotype by PCR analysis did not show a significant association with HA or AD. However, using PCR and due to the limited number of samples available, this analysis could only be addressed in a very small population. Therefore, setting out to pinpoint tagging SNPs is a more useful high-throughput method to identify hot RC-L1s presence/absence in an individual. Not only could we address hot RC-L1s genotype in a larger cohort of HA and AD individuals, but also in other neurological conditions where these may play a role such as ALS or PD.

Our former postdoc Abigail L. Savage identified tagging SNPs (rs1150602, rs6932875, rs7844570, rs7594648, rs6640825) for the five hot RC-L1s analysed by PCR in chapter 4.2 using genotype data for ALS and control individuals, and the SNP data for those elements. Subsequently, using WG or other sources where there is SNP data available and a program such as Haploview, we could infer hot RC-L1s presence/absence in much larger cohorts to establish an association with HA, AD or other neurodegenerative conditions.

5.3.2 Measuring telomere length in ageing

The role of non-coding DNA elements such as telomeres has been long established in ageing [232]. Telomere maintenance is essential for cell cycle division and thus, to prevent early cellular senescence, a hallmark of ageing. Despite neurons being found in a post mitotic state in the adult nervous system, there is still ongoing debate as to whether telomere attrition could be relevant in ageing neurons in addition to playing a role in the supporting non-neuronal cells which undergo cell cycle [233].

The role of telomeres in ageing

Telomeres serve as a protective cap located at the end of chromosomes and shorten with every cell cycle until eventually the telomere becomes too short for the cell to divide, at which point the cell undergoes senescence [232]. Telomerase buffers this effect to some extent by either maintaining or lengthening telomeres in dividing cells. In non-dividing cells such as neurons, telomerase has been demonstrated to have other roles in processes such as neuronal differentiation, neuronal survival and neuritogenesis. In supporting self-renewing neural stem cells (NSCs) and NPCs, there seems to be high telomerase in both the developing and adult brains of humans [234]. It is well established that telomeres shorten with age in dividing cells. The extent to which DNA damage occurs in post-mitotic neurons is debatable, and thus, the question as to the role of telomere shortening may have in neurons remains unanswered [235]. Whereas it is well established that telomere shortening occurs in the brain in a similar manner to that of peripheral tissues such as blood, there is ongoing debate as to whether telomere erosion in the brain arises from neurons or from supporting non-neuronal cells such as NSCs and NPCs [233]. In fact, using quantitative fluorescence *in situ* hybridization, telomere length changes associated to age in neurons and glial cells have been analysed and they demonstrated that telomeres are significantly longer in neurons than in glial cells in adults, but there is no difference between these cell types when assessing infants [236]. Furthermore, older individuals´ glial cells had shorter telomeres than those of younger individuals, showing evidence for telomere attrition in glial cells, but not in neurons [236].

Using qPCR as described in section 2.2.11, we measured the telomere length in both the temporal cortex and blood of 11 healthy aged individuals and 5 AD patients. Preliminary results (Figure 5.1) demonstrated that the average length on each chromosome end is higher in AD than in HA individuals both in the temporal cortex and in the blood. Yet, shorter telomere lengths have been repeatedly associated with increased likelihood of developing age-related diseases such as cancer, cardiovascular disease, obesity, diabetes, chronic pain and some neurodegenerative disorders [232-234, 237-241]. Our findings are only preliminary and in a small population. Furthermore, qPCR is a single point measurement of telomere length and does not account for the rate of shortening. As we do not know the starting length of the telomeres for either group, we cannot conclude differences being due to disease pathways. Expanding the study towards a larger cohort where we can still compare tissue-specific differences and different time points will be the best way to overcome the limitations of this preliminary, yet interesting study.

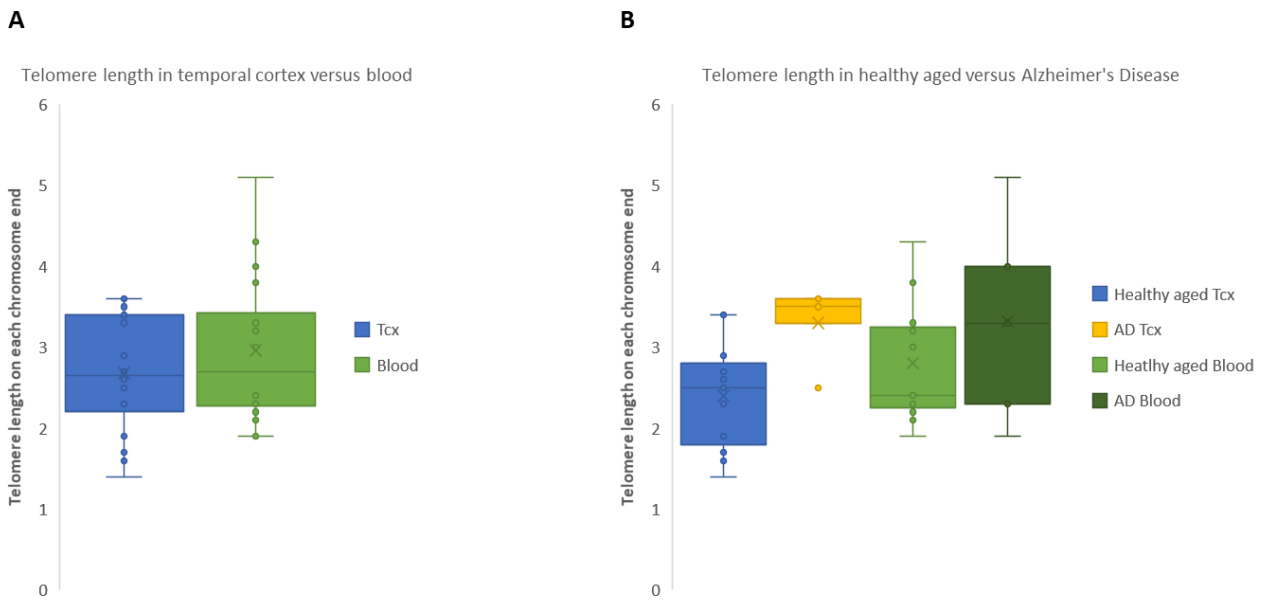


Fig. 5.1. Average telomere length on each chromosome end. A. Average telomere length on each chromosome end on temporal cortex compared to blood DNA of aged individuals. There is no significant difference on the telomere length on the temporal cortex and blood DNA of aged individuals (p-value = 0.302*). p-value* was calculated using a two-tail paired t-test **B.** Average telomere length on each chromosome end on temporal cortex compared to blood DNA stratified by healthy aged and Alzheimer's disease. The average telomere length on each chromosome end is significantly higher in AD (average=3.3) than in HA (average=-2.4) individuals when assessing the temporal cortex alone (p-value=0.0198). The trend is the same when looking at the blood alone (p-value=0.328). *p-value* was calculated using a two-tail t-test assuming equal variances (n=16, 11 HA and 5 AD). Tcx – temporal cortex, AD – Alzheimer's disease.

5.3.3 Sequencing of mitochondrial DNA

Mitochondria are the energy providers of the cell [242]. As such, antagonistic functions of mitochondria have an intimate relationship with both the energy of the youth and the decline of the old [243]. In fact, mitochondrial dysfunction is a hallmark of normal ageing and correlated to age-related diseases [231, 243]. At present, mitochondria are known not only as bioenergetics factories, but also platforms for intracellular signalling and regulators of immunity [243]. As such, the decline in mitochondrial activity causing reduced mitochondrial respiratory complex activity and increased oxidative stress leads to DNA damage, cell senescence and inflammation, where mutations to mitochondrial DNA (mtDNA) are the driving force [242]. mtDNA is inherited maternally [244], and the mitochondrial function of an individual is largely determined by nuclear and mitochondrial genetics as well as by lifestyle choices [245]. Accordingly, research on the field is of major interest to deepen our understanding on the GxE component of our research. Furthermore, in line with our interest on genetic variation is the study of mtDNA polymorphisms, which are crucial to understanding the mitochondrial genotype. In fact, subsets of mtDNA polymorphisms grouped together are named mitochondrial haploblocks and these can cause functional differences in mitochondrial activity and energy metabolism.

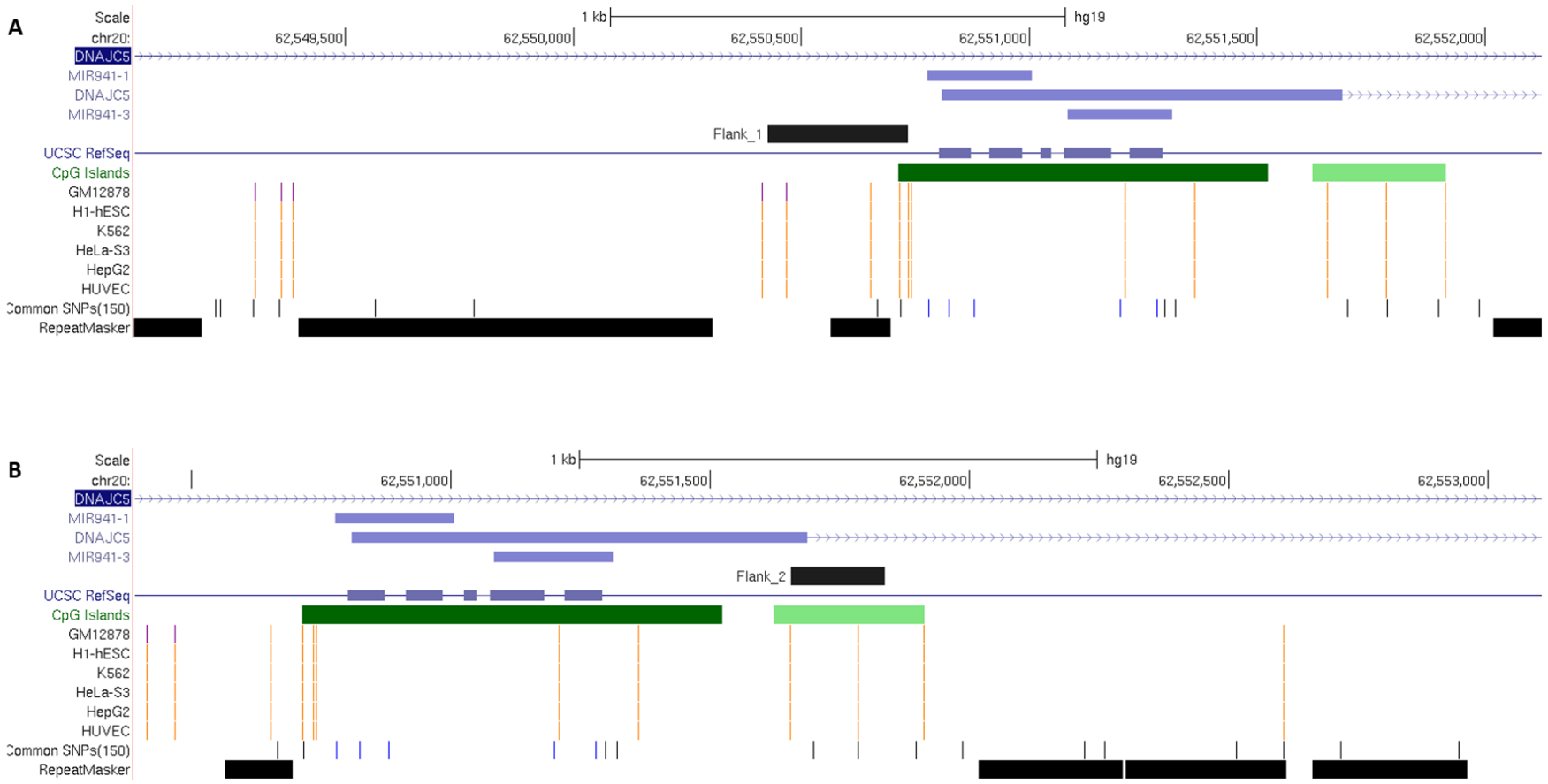
Professor Ana Alfirevic from the Institute of Translational Medicine at the University of Liverpool developed an assay to sequence mtDNA. In an attempt to characterise the mtDNA scenario in HA and AD, we used Prof Alfirevic's approach to sequence the mtDNA

of 18 HA and 10 AD individuals. Bioinformatic analysis using MToolBox, Haplogrep and Phy-mer to characterise mtDNA variants across temporal cortex and blood DNA from the same individual and comparing HA and AD individuals will be carried out. In addition, MToolBox, Haplogrep and Phy-mer will be used for predicting haplotype groups' classification of mtDNA variants.

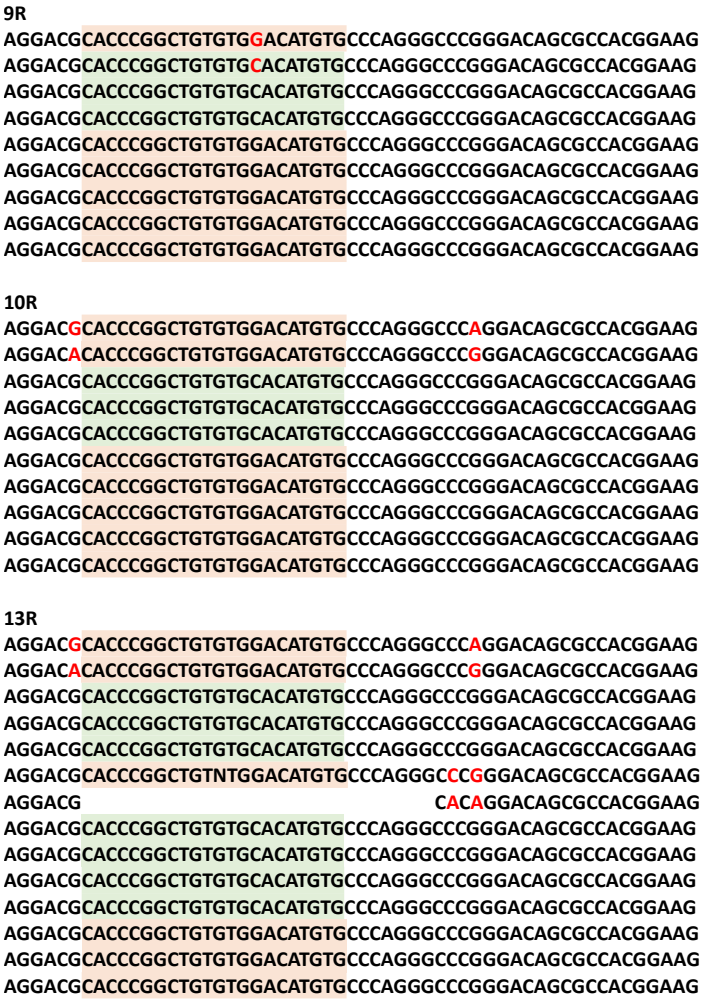
Supplementary material

Supplementary material

Chapter 3.1

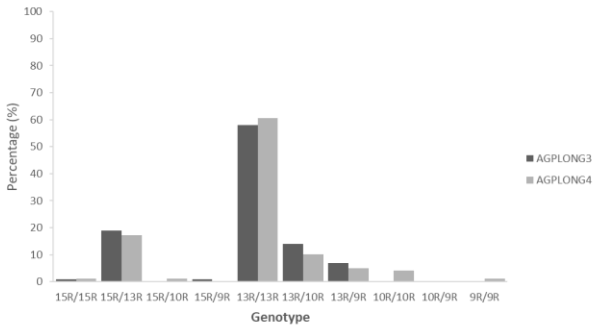


Supplementary fig. 3.1.1. hg19 UCSC image of the 5' and 3' flanking regions of DNAJC5 VNTR used for the analysis of the methylation status of the VNTR/MIR941 locus. A. hg19 UCSC image of the 5' flanking region of *DNAJC5* VNTR (Flank_1) analysed by PCR amplification. **B.** hg19 UCSC image of the 3' flanking region of *DNAJC5* VNTR (Flank_2) analysed by PCR amplification.

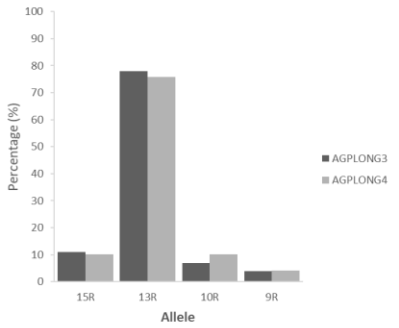


Supplementary fig. 3.1.2. Sequencing across VNTR/MIR941 alleles demonstrated that each contained a different number of VNTR repeats which resulted in a different number of MIR941 copies ranging from nine repeats in the smallest allele (9R) to a predicted 15 repeats in the largest allele (15R). The 9R allele has nine repeats of the VNTR and miRNA, with three *MIR941* copies (green) and six *MIR941** copies (orange), which are defined by a C/G SNP at base 15 in the mature miRNA sequence. The 10R allele has 10 repeats of the VNTR with three copies of *MIR941* and seven copies of *MIR941**. The 13R allele has 13 repeats of the VNTR with seven copies of *MIR941* and six copies of *MIR941**. The high GC content and repetitive nature of this region meant that we were unable to reliably sequence the 15R allele, however, due to its approximate 100 bp size difference compared to the 13R allele as visualised on an agarose gel, we predict that this allele may have an additional two copies of the 56 bp repeat.

A



C



B

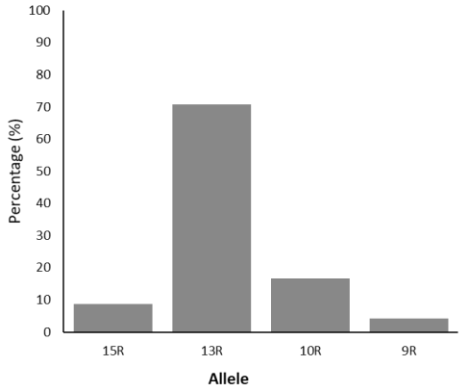
	AGPLONG3		AGPLONG4	
Genotype	Frequency	Percentage	Frequency	Percentage
15R/15R	1	1	1	1.01010101
15R/13R	19	19	17	17.17171717
15R/10R	0	0	1	1.01010101
15R/9R	1	1	0	0
13R/13R	58	58	60	60.60606061
13R/10R	14	14	10	10.1010101
13R/9R	7	7	5	5.050505051
10R/10R	0	0	4	4.04040404
10R/9R	0	0	0	0
9R/9R	0	0	1	1.01010101

D

	AGPLONG3		AGPLONG4	
Allele	Frequency	Percentage	Frequency	Percentage
15R	22	11	20	10.1010101
13R	156	78	150	75.75757576
10R	14	7	20	10.1010101
9R	8	4	8	4.04040404

Supplementary fig. 3.1.3. Genotyping data across DNAJC5 VNTR in the Georgia cohort showed similar distribution to the Dyne Steele cohort. A. Graphic representation of the genotypic distribution of the ten possible genotypes arising from combining the four alleles identified by PCR amplification of *DNAJC5* VNTR in blood DNA in the Georgia cohort. **B.** Table of the genotype frequency across the Georgia cohort. **C.** Graphic representation of the allelic distribution of the four possible alleles identified by PCR amplification of *DNAJC5* VNTR in blood DNA in the Georgia cohort. **D.** Table of the allele frequency across the Georgia cohort. AGPLONG3 cohort comprises 100 people (100.4 av. years), 83 of which are females from 98 to 108 yrs (100.5 av. yrs) and 17 of which are males from 98 to 103 yrs (99.9 av. yrs). AGPLONG4 cohort comprises 99 people (46.8 av. years) – 57 of which are females from 20 to 59 yrs (44.3 av. yrs) and 42 of which are males from 20 to 59 yrs (42.7 av. yrs).

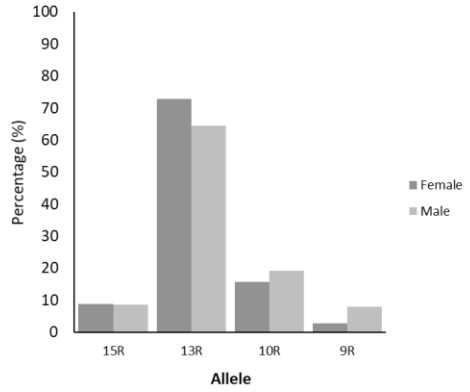
A (a)



(b)

Allele	Dyne Steele	
	Frequency	Percentage
15R	49	8.72
13R	397	70.64
10R	93	16.55
9R	23	4.09

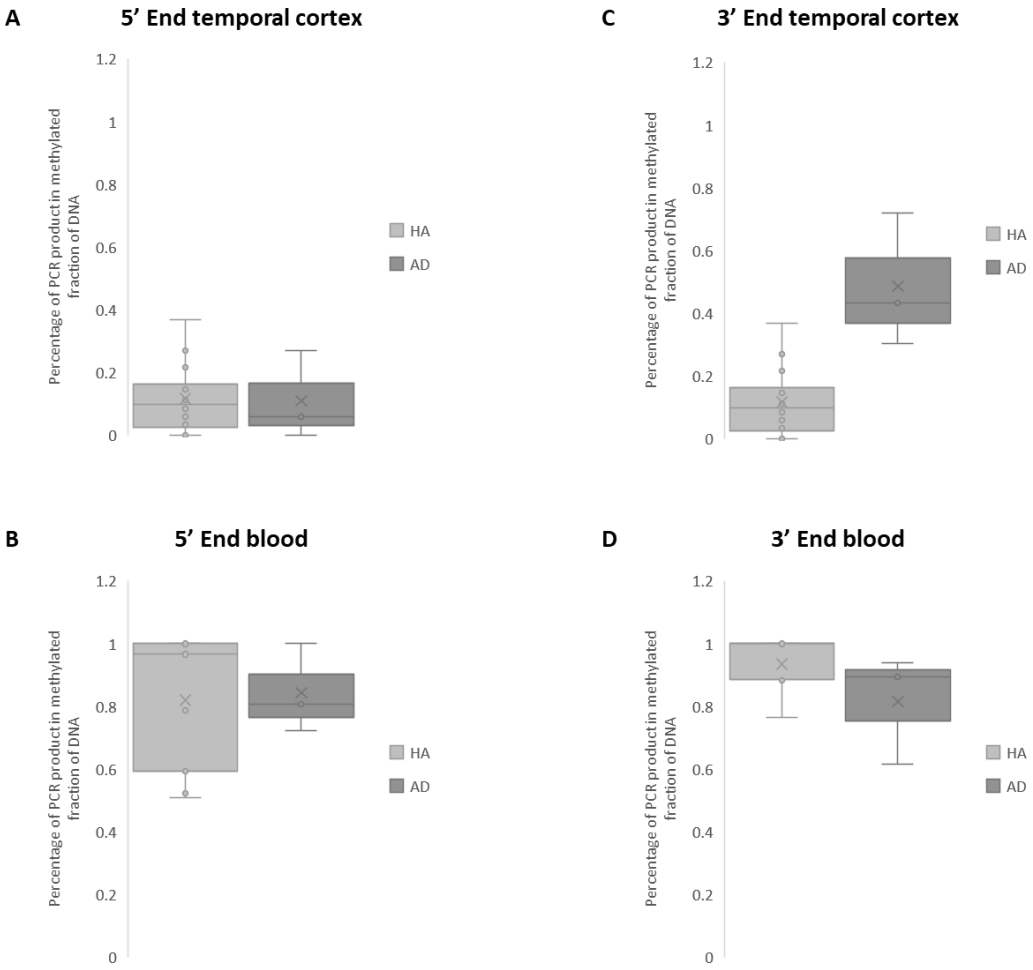
B (a)



(b)

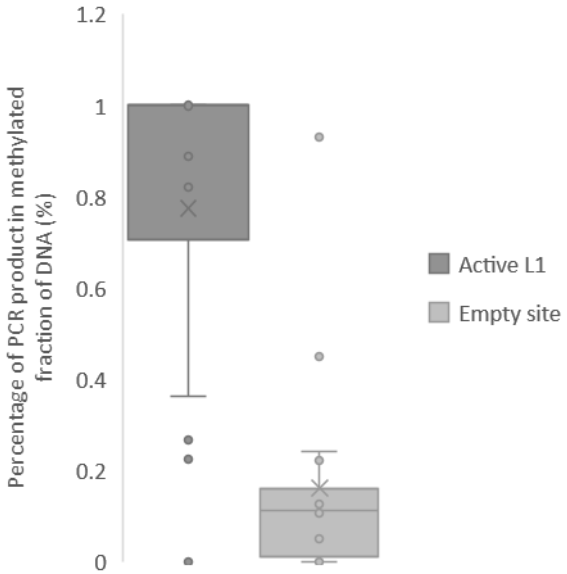
Allele	Dyne Steele Female		Dyne Steele Male	
	Frequency	Percentage	Frequency	Percentage
15R	36	9	13	9
13R	299	73	98	64
10R	64	16	29	19
9R	11	3	12	8

Supplementary fig. 3.1.4. A. Allele frequency of DNAJC5 VNTR region in the Dyne Steele cohort. (a) Graphic representation of the allelic distribution of the four possible alleles identified by PCR amplification of *DNAJC5* VNTR region in blood DNA in the Dyne Steele cohort. **(b)** Table of the allele frequency across the Dyne Steele cohort. **B. Allelic data across DNAJC5 VNTR region showed no significant difference between males and females in the Dyne Steele cohort. a)** Graphic representation of the allelic distribution of the four possible alleles identified by PCR amplification of *DNAJC5* VNTR region in blood DNA in the Dyne Steele cohort stratified by gender. **(b)** Table of the allele frequency across the Dyne Steele cohort stratified by gender. 281 people who enrolled with no identification of cognitive decline at the age of 55 of which 205 are females and 76 are males were used for the study.



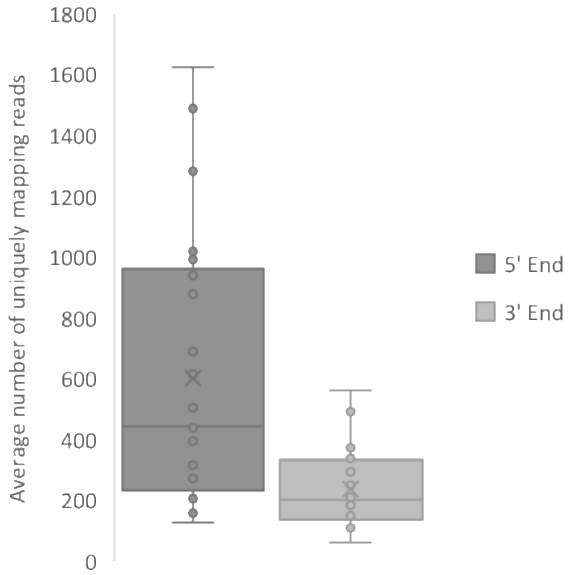
Supplementary fig. 3.1.5. Analysis of methylation status of MIR941/VNTR region stratified by tissue and health status. A&B refer to the upstream flanking region (5' End) of VNTR/MIR941. A. Average methylation status in the temporal cortex, HA– 0.12, AD – 0.11. **B.** Average methylation status in the blood, HA – 0.82, AD – 0.84. **C&D refer to the downstream flanking region (3' End) of MIR941 VNTR. C.** Average methylation status in the temporal cortex, HA – 0.38, AD – 0.49. **D.** Average methylation status in the blood, HA – 0.94, AD – 0.82. Temporal cortex n=12 (HA=9, AD=3). HA – healthy aged, AD – Alzheimer's disease.

Chapter 3.2



Supplementary fig. 3.2.1. The allele harbouring the L1 insertions shows a significantly higher level of methylation compared to the allele lacking the active L1 insertion both in the temporal cortex and in the blood. Methylation status of the 5' end of the active L1 insertion and the empty site of the allele lacking the L1 insertion by comparing the band intensity of the PCR product in the methylated and unmethylated fractions of DNA of heterozygous carriers. Average methylation status – 5' end of active L1 insertion =0.76 and empty site=0.16 (p-value = 8.849E-06). *p* values are calculated by 2-tailed paired *t*-test.

Chapter 4.1



Supplementary fig. 4.1.1. Average number of uniquely mapping reads over the 5' and 3'ends of human specific full-length reference L1s. The average number of reads over the 5´ and 3´ends of human specific full-length L1 elements, respectively was considered as the sequence coverage at 5´and 3´ ends of LINE-1 libraries. The average number of reads at the 5´end was 602.8. The average number of reads at the 3´end was 239.8.

Appendices

Appendices

Documents available upon request.

Contact Professor John Quinn: jquinn@liverpool.ac.uk

Appendix 1 – Human DNA

- Dyne Steele information
- Georgia Collection information

Appendix 2 – MIR941/VNTR locus (Chapter 3.1)

- MIR941/VNTR genotype statistical analysis

Appendix 3 – LINE-1 methylation (Chapter 3.2)

- Bisulphite and Pyroseq reports
- Hot RC-L1s RIPs genotype statistical analysis

Appendix 4 – RC-Seq (Chapter 4.1)

- RC-Seq scripts
- RC-Seq protocol

Appendix 5 – WGS (Chapter 4.2)

- WGS scripts

Reference list

Reference list

1. Pihlstrom, L., S. Wiethoff, and H. Houlden, *Genetics of neurodegenerative diseases: an overview.* Handb Clin Neurol, 2017. **145**: p. 309-323.
2. Stern, Y., *Cognitive reserve in ageing and Alzheimer's disease.* Lancet Neurol, 2012. **11**(11): p. 1006-12.
3. Brooks-Wilson, A.R., *Genetics of healthy aging and longevity.* Hum Genet, 2013. **132**(12): p. 1323-38.
4. Trampush, J.W., et al., *GWAS meta-analysis reveals novel loci and genetic correlates for general cognitive function: a report from the COGENT consortium.* Mol Psychiatry, 2017. **22**(3): p. 336-345.
5. Bennett, E.A., et al., *Natural genetic variation caused by transposable elements in humans.* Genetics, 2004. **168**(2): p. 933-51.
6. Chuong, E.B., N.C. Elde, and C. Feschotte, *Regulatory activities of transposable elements: from conflicts to benefits.* Nat Rev Genet, 2017. **18**(2): p. 71-86.
7. Chureau, C., et al., *Ftx is a non-coding RNA which affects Xist expression and chromatin structure within the X-inactivation center region.* Human Molecular Genetics, 2011. **20**(4): p. 705-718.
8. Conley, A.B., J. Piriyapongsa, and I.K. Jordan, *Retroviral promoters in the human genome.* Bioinformatics, 2008. **24**(14): p. 1563-7.
9. Erwin, J.A., M.C. Marchetto, and F.H. Gage, *Mobile DNA elements in the generation of diversity and complexity in the brain.* Nat Rev Neurosci, 2014. **15**(8): p. 497-506.
10. Garcia-Perez, J.L., T.J. Widmann, and I.R. Adams, *The impact of transposable elements on mammalian development.* Development, 2016. **143**(22): p. 4101-4114.
11. Quinn, J.P., A.L. Savage, and V.J. Bubb, *Non-coding genetic variation shaping mental health.* Curr Opin Psychol, 2019. **27**: p. 18-24.
12. Criscione, S.W., et al., *Transcriptional landscape of repetitive elements in normal and cancer human cells.* BMC Genomics, 2014. **15**: p. 583.
13. Han, J.S., S.T. Szak, and J.D. Boeke, *Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes.* Nature, 2004. **429**(6989): p. 268-74.
14. Quinn, J.P. and V.J. Bubb, *SVA retrotransposons as modulators of gene expression.* Mob Genet Elements, 2014. **4**: p. e32102.
15. Savage, A.L., et al., *Characterisation of the potential function of SVA retrotransposons to modulate gene expression patterns.* BMC Evol Biol, 2013. **13**: p. 101.
16. Beck, C.R., et al., *LINE-1 elements in structural variation and disease.* Annu Rev Genomics Hum Genet, 2011. **12**: p. 187-215.
17. Bodea, G.O., E.G.Z. McKelvey, and G.J. Faulkner, *Retrotransposon-induced mosaicism in the neural genome.* Open Biol, 2018. **8**(7): p. 180074.

18. Klein, S.J. and R.J. O'Neill, *Transposable elements: genome innovation, chromosome diversity, and centromere conflict.* Chromosome Res, 2018. **26**(1-2): p. 5-23.
19. Fjell, A.M., et al., *What is normal in normal aging? Effects of aging, amyloid and Alzheimer's disease on the cerebral cortex and the hippocampus.* Prog Neurobiol, 2014. **117**: p. 20-40.
20. Lipnicki, D.M., et al., *Age-related cognitive decline and associations with sex, education and apolipoprotein E genotype across ethnocultural groups and geographic regions: a collaborative cohort study.* PLoS Med, 2017. **14**(3): p. e1002261.
21. Gierman, H.J., et al., *Whole-genome sequencing of the world's oldest people.* PLoS One, 2014. **9**(11): p. e112430.
22. de Magalhães, J.P., *Why genes extending lifespan in model organisms have not been consistently associated with human longevity and what it means to translation research.* Cell Cycle, 2014. **13**(17): p. 2671-2673.
23. Deelen, J., et al., *Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age.* Hum Mol Genet, 2014. **23**(16): p. 4420-32.
24. Lane, C.A., J. Hardy, and J.M. Schott, *Alzheimer's disease.* Eur J Neurol, 2018. **25**(1): p. 59-70.
25. Rawle, M.J., et al., *Apolipoprotein-E (Apoe) epsilon4 and cognitive decline over the adult life course.* Transl Psychiatry, 2018. **8**(1): p. 18.
26. Christensen, K. and M. McGue, *Genetics: Healthy ageing, the genome and the environment.* Nat Rev Endocrinol, 2016. **12**(7): p. 378-80.
27. Alzheimer's, A., *2016 Alzheimer's disease facts and figures.* Alzheimers Dement, 2016. **12**(4): p. 459-509.
28. Fjell, A.M., et al., *Accelerating cortical thinning: unique to dementia or universal in aging?* Cereb Cortex, 2014. **24**(4): p. 919-34.
29. Duncan, J., *The structure of cognition: attentional episodes in mind and brain.* Neuron, 2013. **80**(1): p. 35-50.
30. Genon, S., et al., *How to Characterize the Function of a Brain Region.* Trends Cogn Sci, 2018. **22**(4): p. 350-364.
31. Ackerman, S., *Discovering the Brain*. Major Structures and Functions of the Brain. 1992, Washington (DC): National Academies Press (US).
32. Altuna, M., et al., *DNA methylation signature of human hippocampus in Alzheimer's disease is linked to neurogenesis.* Clin Epigenetics, 2019. **11**(1): p. 91.
33. Rehfeld, K., et al., *Dancing or Fitness Sport? The Effects of Two Training Programs on Hippocampal Plasticity and Balance Abilities in Healthy Seniors.* Front Hum Neurosci, 2017. **11**: p. 305.
34. Rapp, B. and R.W. Wiley, *Re-learning and remembering in the lesioned brain.* Neuropsychologia, 2019. **132**: p. 107126.

35. Committee, N.R.C.U., *Mapping and Sequencing the Human Genome. Mapping and Sequencing the Human Genome*. Introduction. 1988, Washington (DC): : National Academies Press (US).
36. Davidson, S., et al., *Analysis of the effects of depression associated polymorphisms on the activity of the BICC1 promoter in amygdala neurones.* The pharmacogenomics journal, 2016. **16**(4): p. 366-374.
37. Gianfrancesco, O., V.J. Bubb, and J.P. Quinn, *SVA retrotransposons as potential modulators of neuropeptide gene expression.* Neuropeptides, 2017. **64**: p. 3-7.
38. Kines, K.J. and V.P. Belancio, *Expressing genes do not forget their LINEs: transposable elements and gene expression.* Frontiers in bioscience (Landmark edition), 2012. **17**: p. 1329-1344.
39. Vasiliou, S.A., et al., *The SLC6A4 VNTR genotype determines transcription factor binding and epigenetic variation of this gene in response to cocaine in vitro.* Addict Biol, 2012. **17**(1): p. 156-70.
40. Gianfrancesco, O., et al., *Identification and Potential Regulatory Properties of Evolutionary Conserved Regions (ECRs) at the Schizophrenia-Associated MIR137 Locus.* J Mol Neurosci, 2016. **60**(2): p. 239-47.
41. Hasler, J. and K. Strub, *Alu elements as regulators of gene expression.* Nucleic Acids Res, 2006. **34**(19): p. 5491-7.
42. Bell, J.I., *Single nucleotide polymorphisms and disease gene mapping.* Arthritis research, 2002. **4 Suppl 3**(Suppl 3): p. S273-S278.
43. de Koning, A.P.J., et al., *Repetitive elements may comprise over two-thirds of the human genome.* PLoS genetics, 2011. **7**(12): p. e1002384-e1002384.
44. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
45. Pelley, J.W., *Organization, Synthesis, and Repair of DNA*, in *Elsevier's Integrated Biochemistry*. 2007, Mosby: Philadelphia. p. 123-133.
46. Sawaya, S., et al., *Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements.* PLoS One, 2013. **8**(2): p. e54710.
47. Breen, G., et al., *Variable number tandem repeats as agents of functional regulation in the genome.* IEEE Eng Med Biol Mag, 2008. **27**(2): p. 103-4, 108.
48. Manca, M., et al., *The Regulation of Monoamine Oxidase A Gene Expression by Distinct Variable Number Tandem Repeats.* J Mol Neurosci, 2018. **64**(3): p. 459-470.
49. Michelhaugh, S.K., et al., *The dopamine transporter gene (SLC6A3) variable number of tandem repeats domain enhances transcription in dopamine neurons.* J Neurochem, 2001. **79**(5): p. 1033-8.
50. Lupan, I., et al., *Lineage specific evolution of the VNTR composite retrotransposon central domain and its role in retrotransposition of gibbon LAVA elements.* BMC Genomics, 2015. **16**(1): p. 389.

51. MacKenzie, A. and J. Quinn, *A serotonin transporter gene intron 2 polymorphic region, correlated with affective disorders, has allele-dependent differential enhancer-like properties in the mouse embryo.* Proceedings of the National Academy of Sciences of the United States of America, 1999. **96**(26): p. 15251-15255.
52. Mc, C.B., *The origin and behavior of mutable loci in maize.* Proc Natl Acad Sci U S A, 1950. **36**(6): p. 344-55.
53. Goodier, J.L., *Restricting retrotransposons: a review.* Mob DNA, 2016. **7**: p. 16.
54. Kapitonov, V.V. and J. Jurka, *RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons.* PLoS biology, 2005. **3**(6): p. e181-e181.
55. Kapitonov, V.V. and E.V. Koonin, *Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon.* Biology direct, 2015. **10**: p. 20-20.
56. Carmona, L.M. and D.G. Schatz, *New insights into the evolutionary origins of the recombination-activating gene proteins and V(D)J recombination.* The FEBS journal, 2017. **284**(11): p. 1590-1605.
57. Zhang, Y., et al., *Transposon molecular domestication and the evolution of the RAG recombinase.* Nature, 2019. **569**(7754): p. 79-84.
58. Cordaux, R. and M.A. Batzer, *The impact of retrotransposons on human genome evolution.* Nat Rev Genet, 2009. **10**(10): p. 691-703.
59. Flockerzi, A., et al., *Human endogenous retrovirus HERV-K14 families: status, variants, evolution, and mobilization of other cellular sequences.* Journal of virology, 2005. **79**(5): p. 2941-2949.
60. Mager, D.L. and J.P. Stoye, *Mammalian Endogenous Retroviruses.* Microbiol Spectr, 2015. **3**(1).
61. Wildschutte, J.H., et al., *Discovery of unfixed endogenous retrovirus insertions in diverse human populations.* Proceedings of the National Academy of Sciences of the United States of America, 2016. **113**(16): p. E2326-E2334.
62. Savage, A.L., et al., *Retrotransposons in the development and progression of amyotrophic lateral sclerosis.* J Neurol Neurosurg Psychiatry, 2019. **90**(3): p. 284-293.
63. Küry, P., et al., *Human Endogenous Retroviruses in Neurological Diseases.* Trends in Molecular Medicine, 2018. **24**(4): p. 379-394.
64. Mayer, J., et al., *Transcriptional profiling of HERV-K(HML-2) in amyotrophic lateral sclerosis and potential implications for expression of HML-2 proteins.* Molecular neurodegeneration, 2018. **13**(1): p. 39-39.
65. Wang, L., E.T. Norris, and I.K. Jordan, *Human Retrotransposon Insertion Polymorphisms Are Associated with Health and Disease via Gene Regulatory Phenotypes.* Front Microbiol, 2017. **8**: p. 1418.
66. Luan, D.D., et al., *Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition.* Cell, 1993. **72**(4): p. 595-605.
67. Batzer, M.A. and P.L. Deininger, *Alu repeats and human genomic diversity.* Nat Rev Genet, 2002. **3**(5): p. 370-9.

68. Faulkner, G.J., *Retrotransposons: mobile and mutagenic from conception to death.* FEBS Lett, 2011. **585**(11): p. 1589-94.
69. Baillie, J.K., et al., *Somatic retrotransposition alters the genetic landscape of the human brain.* Nature, 2011. **479**(7374): p. 534-7.
70. Hancks, D.C. and H.H. Kazazian, Jr., *Active human retrotransposons: variation and disease.* Curr Opin Genet Dev, 2012. **22**(3): p. 191-203.
71. Klawitter, S., et al., *Reprogramming triggers endogenous L1 and Alu retrotransposition in human induced pluripotent stem cells.* Nat Commun, 2016. **7**: p. 10286.
72. Beck, C.R., et al., *LINE-1 retrotransposition activity in human genomes.* Cell, 2010. **141**(7): p. 1159-70.
73. Brouha, B., et al., *Hot L1s account for the bulk of retrotransposition in the human population.* Proc Natl Acad Sci U S A, 2003. **100**(9): p. 5280-5.
74. Singer, T., et al., *LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes?* Trends in neurosciences, 2010. **33**(8): p. 345-354.
75. Mills, R.E., et al., *Recently mobilized transposons in the human and chimpanzee genomes.* Am J Hum Genet, 2006. **78**(4): p. 671-9.
76. Mills, R.E., et al., *Which transposable elements are active in the human genome?* Trends Genet, 2007. **23**(4): p. 183-91.
77. Yu, Q., et al., *Population-wide sampling of retrotransposon insertion polymorphisms using deep sequencing and efficient detection.* Gigascience, 2017. **6**(9): p. 1-11.
78. Quinn, J.P., A.L. Savage, and V.J. Bubb, *Non-coding genetic variation shaping mental health.* Current Opinion in Psychology, 2019. **27**: p. 18-24.
79. Savage, A.L., et al., *Characterisation of the potential function of SVA retrotransposons to modulate gene expression patterns.* BMC Evolutionary Biology, 2013. **13**: p. 101-101.
80. Quinn, J.P., et al., *Polymorphic variation as a driver of differential neuropeptide gene expression.* Neuropeptides, 2013. **47**(6): p. 395-400.
81. Robberecht, C., et al., *Nonallelic homologous recombination between retrotransposable elements is a driver of de novo unbalanced translocations.* Genome Res, 2013. **23**(3): p. 411-8.
82. Schrider, D.R., et al., *Gene copy-number polymorphism caused by retrotransposition in humans.* PLoS Genet, 2013. **9**(1): p. e1003242.
83. Stewart, C., et al., *A comprehensive map of mobile element insertion polymorphisms in humans.* PLoS Genet, 2011. **7**(8): p. e1002236.
84. Xing, J., et al., *Mobile elements create structural variation: analysis of a complete human genome.* Genome research, 2009. **19**(9): p. 1516-1526.
85. Ewing, A.D., *Transposable element detection from whole genome sequence data.* Mob DNA, 2015. **6**: p. 24.
86. Feusier, J., et al., *Pedigree-based estimation of human mobile element retrotransposition rates.* Genome Research, 2019. **29**(10): p. 1567-1577.

87. Gardner, E.J., et al., *The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology.* Genome Res, 2017. **27**(11): p. 1916-1929.
88. Sanchez-Luque, F.J., S.R. Richardson, and G.J. Faulkner, *Retrotransposon Capture Sequencing (RC-Seq): A Targeted, High-Throughput Approach to Resolve Somatic L1 Retrotransposition in Humans*, in *Transposons and Retrotransposons: Methods and Protocols*. 2016, Springer New York: New York, NY. p. 47-77.
89. Doucet, T.T. and H.H. Kazazian, Jr., *Long Interspersed Element Sequencing (L1-Seq): A Method to Identify Somatic LINE-1 Insertions in the Human Genome.* Methods in molecular biology (Clifton, N.J.), 2016. **1400**: p. 79-93.
90. Badge, R.M., R.S. Alisch, and J.V. Moran, *ATLAS: a system to selectively identify human-specific L1 insertions.* American journal of human genetics, 2003. **72**(4): p. 823-838.
91. Nelson, M.G., R.S. Linheiro, and C.M. Bergman, *McClintock: An Integrated Pipeline for Detecting Transposable Element Insertions in Whole-Genome Shotgun Sequencing Data.* G3 (Bethesda, Md.), 2017. **7**(8): p. 2763-2778.
92. Nakagome, M., et al., *Transposon Insertion Finder (TIF): a novel program for detection of de novo transpositions of transposable elements.* BMC bioinformatics, 2014. **15**: p. 71-71.
93. Ewing, A.D. and H.H. Kazazian, Jr., *High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes.* Genome Res, 2010. **20**(9): p. 1262-70.
94. Cordaux, R., et al., *Estimating the retrotransposition rate of human Alu elements.* Gene, 2006. **373**: p. 134-137.
95. Hancks, D.C. and H.H. Kazazian, Jr., *Roles for retrotransposon insertions in human disease.* Mob DNA, 2016. **7**: p. 9.
96. Steranka, J.P., et al., *Transposon insertion profiling by sequencing (TIPseq) for mapping LINE-1 insertions in the human genome.* Mobile DNA, 2019. **10**: p. 8-8.
97. Payer, L.M., et al., *Structural variants caused by Alu insertions are associated with risks for many human diseases.* Proceedings of the National Academy of Sciences of the United States of America, 2017. **114**(20): p. E3984-E3992.
98. Payer, L.M., et al., *Alu insertion variants alter mRNA splicing.* Nucleic acids research, 2019. **47**(1): p. 421-431.
99. Hing, B., et al., *A polymorphism associated with depressive disorders differentially regulates brain derived neurotrophic factor promoter IV activity.* Biological psychiatry, 2012. **71**(7): p. 618-626.
100. Davidson, S., et al., *Differential activity by polymorphic variants of a remote enhancer that supports galanin expression in the hypothalamus and amygdala: implications for obesity, depression and alcoholism.* Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology, 2011. **36**(11): p. 2211-2221.
101. Gianfrancesco, O., et al., *Novel brain expressed RNA identified at the MIR137 schizophrenia-associated locus.* Schizophr Res, 2017. **184**: p. 109-115.

102. Gilbert, N., S. Lutz-Prigge, and J.V. Moran, *Genomic Deletions Created upon LINE-1 Retrotransposition.* Cell, 2002. **110**(3): p. 315-325.
103. Goodier, J.L., E.M. Ostertag, and H.H. Kazazian, Jr., *Transduction of 3'-flanking sequences is common in L1 retrotransposition.* Hum Mol Genet, 2000. **9**(4): p. 653-7.
104. Pickeral, O.K., et al., *Frequent human genomic DNA transduction driven by LINE-1 retrotransposition.* Genome research, 2000. **10**(4): p. 411-415.
105. Sun, X., et al., *Transcription factor profiling reveals molecular choreography and key regulators of human retrotransposon expression.* Proc Natl Acad Sci U S A, 2018. **115**(24): p. E5526-E5535.
106. Perepelitsa-Belancio, V. and P. Deininger, *RNA truncation by premature polyadenylation attenuates human mobile element activity.* Nature Genetics, 2003. **35**(4): p. 363-366.
107. Kim, D.S. and Y. Hahn, *Identification of human-specific transcript variants induced by DNA insertions in the human genome.* Bioinformatics, 2011. **27**(1): p. 14-21.
108. Aschacher, T., et al., *LINE-1 induces hTERT and ensures telomere maintenance in tumour cell lines.* Oncogene, 2016. **35**(1): p. 94-104.
109. Morrish, T.A., et al., *Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres.* Nature, 2007. **446**(7132): p. 208-212.
110. Sen, S.K., et al., *Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome.* Nucleic Acids Res, 2007. **35**(11): p. 3741-51.
111. Payton, A., et al., *A TOMM40 poly-T variant modulates gene expression and is associated with vocabulary ability and decline in nonpathologic aging.* Neurobiology of Aging, 2016. **39**.
112. Davies, G., et al., *A genome-wide association study implicates the APOE locus in nonpathological cognitive ageing.* Mol Psychiatry, 2014. **19**(1): p. 76-87.
113. Aneichyk, T., et al., *Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly.* Cell, 2018. **172**(5): p. 897-909 e21.
114. Maag, J.L.V., et al., *Dynamic expression of long noncoding RNAs and repeat elements in synaptic plasticity.* Frontiers in Neuroscience, 2015. **9**: p. 351.
115. Coufal, N.G., et al., *L1 retrotransposition in human neural progenitor cells.* Nature, 2009. **460**(7259): p. 1127-31.
116. Muotri, A.R., et al., *Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition.* Nature, 2005. **435**(7044): p. 903-10.
117. Evrony, G.D., et al., *Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain.* Cell, 2012. **151**(3): p. 483-96.
118. Upton, K.R., et al., *Ubiquitous L1 mosaicism in hippocampal neurons.* Cell, 2015. **161**(2): p. 228-39.
119. Belancio, V.P., et al., *Somatic expression of LINE-1 elements in human tissues.* Nucleic acids research, 2010. **38**(12): p. 3909-3922.

120. Macia, A., et al., *Engineered LINE-1 retrotransposition in nondividing human neurons.* Genome Res, 2017. **27**(3): p. 335-348.
121. Erwin, J.A., et al., *L1-associated genomic regions are deleted in somatic cells of the healthy human brain.* Nature neuroscience, 2016. **19**(12): p. 1583-1591.
122. Jung, S.E., K.J. Shin, and H.Y. Lee, *DNA methylation-based age prediction from various tissues and body fluids.* BMB Rep, 2017. **50**(11): p. 546-553.
123. Okano, M., et al., *DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development.* Cell, 1999. **99**(3): p. 247-257.
124. Ooi, S.K.T., et al., *DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA.* Nature, 2007. **448**(7154): p. 714-717.
125. Greenberg, M.V.C. and D. Bourc'his, *The diverse roles of DNA methylation in mammalian development and disease.* Nature Reviews. Molecular Cell Biology, 2019. **20**: p. 590-607.
126. Zhu, P., et al., *Single-cell DNA methylome sequencing of human preimplantation embryos.* Nature Genetics, 2018. **50**(12-19).
127. Guo, H., et al., *The DNA methylation landscape of human early embryos.* Nature, 2014. **511**(7511): p. 606-610.
128. Tang, W.W.C., et al., *A Unique Gene Regulatory Network Resets the Human Germline Epigenome for Development.* Cell, 2015. **161**(6): p. 1453-1467.
129. Rowe, H.M., et al., *KAP1 controls endogenous retroviruses in embryonic stem cells.* Nature, 2010. **463**(7278): p. 237-240.
130. Gnanakkan, V.P., et al., *TE-array--a high throughput tool to study transposon transcription.* BMC genomics, 2013. **14**: p. 869-869.
131. MacLennan, M., et al., *Mobilization of LINE-1 retrotransposons is restricted by Tex19.1 in mouse embryonic stem cells.* eLife, 2017. **6**: p. e26152.
132. Hackett, J.A., et al., *Promoter DNA methylation couples genome-defence mechanisms to epigenetic reprogramming in the mouse germline.* Development (Cambridge, England), 2012. **139**(19): p. 3623-3632.
133. Dubnau, J., *The Retrotransposon storm and the dangers of a Collyer's genome.* Curr Opin Genet Dev, 2018. **49**: p. 95-105.
134. McLaughlin, R.N., Jr. and H.S. Malik, *Genetic conflicts: the usual suspects and beyond.* The Journal of experimental biology, 2017. **220**(Pt 1): p. 6-17.
135. Cardelli, M., *The epigenetic alterations of endogenous retroelements in aging.* Mech Ageing Dev, 2018. **174**: p. 30-46.
136. Baccarelli, A., et al., *Repetitive element DNA methylation and circulating endothelial and inflammation markers in the VA normative aging study.* Epigenetics, 2010. **5**(3): p. 222-8.
137. Bollati, V., et al., *Decline in genomic DNA methylation through aging in a cohort of elderly subjects.* Mech Ageing Dev, 2009. **130**(4): p. 234-9.
138. Guo, C., et al., *Tau Activates Transposable Elements in Alzheimer's Disease.* Cell Rep, 2018. **23**(10): p. 2874-2880.

139. Davies, G., et al., *Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function.* Nat Commun, 2018. **9**(1): p. 2098.
140. Deary, I.J., S.E. Harris, and W.D. Hill, *What genome-wide association studies reveal about the association between intelligence and physical health, illness, and mortality.* Current Opinion in Psychology, 2019. **27**: p. 6-12.
141. Davies, G., et al., *Genome-wide association studies establish that human intelligence is highly heritable and polygenic.* Molecular psychiatry, 2011. **16**(10): p. 996-1005.
142. Lencz, T., et al., *Molecular genetic evidence for overlap between general cognitive ability and risk for schizophrenia: a report from the Cognitive Genomics consorTium (COGENT).* Molecular Psychiatry, 2014. **19**(2): p. 168-174.
143. Benyamin, B., et al., *Childhood intelligence is heritable, highly polygenic and associated with FNBP1L.* Mol Psychiatry, 2014. **19**(2): p. 253-8.
144. Kirkpatrick, R.M., et al., *Results of a "GWAS plus:" general cognitive ability is substantially heritable and massively polygenic.* PLoS One, 2014. **9**(11): p. e112390.
145. Davies, G., et al., *Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N=53949).* Mol Psychiatry, 2015. **20**(2): p. 183-92.
146. Davies, G., et al., *Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N=112 151).* Mol Psychiatry, 2016. **21**(6): p. 758-67.
147. Sniekers, S., et al., *Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence.* Nat Genet, 2017. **49**(7): p. 1107-1112.
148. Willcox, B.J., et al., *FOXO3A genotype is strongly associated with human longevity.* Proc Natl Acad Sci U S A, 2008. **105**(37): p. 13987-92.
149. Savage, J.E., et al., *Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence.* Nat Genet, 2018. **50**(7): p. 912-919.
150. Zhou, L. and P. Verstreken, *Reprogramming neurodegeneration in the big data era.* Curr Opin Neurobiol, 2018. **48**(1873-6882 (Electronic)): p. 167-173.
151. Beekman, M., et al., *Genome-wide linkage analysis for human longevity: Genetics of Healthy Aging Study.* Aging Cell, 2013. **12**(2): p. 184-93.
152. Harold, D., et al., *Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease.* Nature Genetics, 2009. **41**(10): p. 1088-1093.
153. De Jager, P.L., et al., *A genome-wide scan for common variants affecting the rate of age-related cognitive decline.* Neurobiol Aging, 2012. **33**(5): p. 1017 e1-15.
154. Kim, J., J.M. Basak, and D.M. Holtzman, *The role of apolipoprotein E in Alzheimer's disease.* Neuron, 2009. **63**(3): p. 287-303.
155. Deary, I.J., et al., *Cognitive change and the APOE epsilon 4 allele.* Nature, 2002. **418**(6901): p. 932.

156. Roses, A.D., et al., *A TOMM40 variable-length polymorphism predicts the age of late-onset Alzheimer's disease.* Pharmacogenomics J, 2010. **10**(5): p. 375-84.
157. Corder, E.H., et al., *Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families.* Science, 1993. **261**(5123): p. 921-3.
158. Strittmatter, W.J., et al., *Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease.* Proc Natl Acad Sci U S A, 1993. **90**(5): p. 1977-81.
159. Reynolds, C.A., et al., *Longitudinal memory performance during normal aging: twin association models of APOE and other Alzheimer candidate genes.* Behav Genet, 2006. **36**(2): p. 185-94.
160. Schiepers, O.J., et al., *APOE E4 status predicts age-related cognitive decline in the ninth decade: longitudinal follow-up of the Lothian Birth Cohort 1921.* Mol Psychiatry, 2012. **17**(3): p. 315-24.
161. Small, B.J., et al., *Apolipoprotein E and cognitive performance: a meta-analysis.* Psychol Aging, 2004. **19**(4): p. 592-600.
162. Sebastiani, P., et al., *Genetic signatures of exceptional longevity in humans.* PLoS One, 2012. **7**(1): p. e29848.
163. Lutz, M.W., et al., *Genetic variation at a single locus and age of onset for Alzheimer's disease.* Alzheimers & Dementia, 2010. **6**(2): p. 125-131.
164. Fontaine, S.N., et al., *DnaJ/Hsc70 chaperone complexes control the extracellular release of neurodegenerative-associated proteins.* EMBO J, 2016. **35**(14): p. 1537-49.
165. Burgoyne, R.D. and A. Morgan, *Cysteine string protein (CSP) and its role in preventing neurodegeneration.* Semin Cell Dev Biol, 2015. **40**: p. 153-9.
166. Lopez-Ortega, E., R. Ruiz, and L. Tabares, *CSPalpha, a Molecular Co-chaperone Essential for Short and Long-Term Synaptic Maintenance.* Front Neurosci, 2017. **11**: p. 39.
167. Korbie, D.J. and J.S. Mattick, *Touchdown PCR for increased specificity and sensitivity in PCR amplification.* Nat Protoc, 2008. **3**(9): p. 1452-6.
168. Daskalos, A., et al., *Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer.* Int J Cancer, 2009. **124**(1): p. 81-7.
169. Hu, H.Y., et al., *Evolution of the human-specific microRNA miR-941.* Nat Commun, 2012. **3**: p. 1145.
170. Warburton, A., et al., *Characterization of a REST-Regulated Internal Promoter in the Schizophrenia Genome-Wide Associated Gene MIR137.* Schizophr Bull, 2015. **41**(3): p. 698-707.
171. Lu, J. and A.G. Clark, *Impact of microRNA regulation on variation in human gene expression.* Genome Res, 2012. **22**(7): p. 1243-54.
172. Cao, D.D., L. Li, and W.Y. Chan, *MicroRNAs: Key Regulators in the Central Nervous System and Their Implication in Neurological Diseases.* Int J Mol Sci, 2016. **17**(6): p. 842.

173. Tiwari, S.S., et al., *Evidence that the presynaptic vesicle protein CSPalpha is a key player in synaptic degeneration and protection in Alzheimer's disease.* Mol Brain, 2015. **8**: p. 6.
174. Gorenberg, E.L. and S.S. Chandra, *The Role of Co-chaperones in Synaptic Proteostasis and Neurodegenerative Disease.* Front Neurosci, 2017. **11**: p. 248.
175. Roosen, D.A., et al., *DNAJC proteins and pathways to parkinsonism.* FEBS J, 2019. **286**(16): p. 3080-3094.
176. Brehme, M., et al., *A chaperome subnetwork safeguards proteostasis in aging and neurodegenerative disease.* Cell Rep, 2014. **9**(3): p. 1135-50.
177. Najafi-Shoushtari, S.H., et al., *MicroRNA-33 and the SREBP Host Genes Cooperate to Control Cholesterol Homeostasis.* Science, 2010. **328**(5985): p. 1566.
178. Shumay, E., et al., *Evidence that the methylation state of the monoamine oxidase A (MAOA) gene predicts brain activity of MAO A enzyme in healthy men.* Epigenetics, 2012. **7**(10): p. 1151-60.
179. Grothaus, K., et al., *Genome-wide methylation analysis of retrocopy-associated CpG islands and their genomic environment.* Epigenetics, 2016. **11**(3): p. 216-26.
180. Gianfrancesco, O., *Regulation at the schizophrenia-associated MIR137 locus and repetitive DNA in the regulation and evolution of brain-related pathways.*, in *Molecular and Clinical Pharmacology*. 2018, University of Liverpool. p. 324.
181. Schneider, C.A., W.S. Rasband, and K.W. Eliceiri, *NIH Image to ImageJ: 25 years of image analysis*. 2012: Nature methods p. 671-675
182. Zhang, C., et al., *Selection of reference genes for gene expression studies in human bladder cancer using SYBR-Green quantitative polymerase chain reaction.* Oncol Lett, 2017. **14**(5): p. 6001-6011.
183. Paredes, U.M., J.P. Quinn, and U.M. D'Souza, *Allele-specific transcriptional activity of the variable number of tandem repeats in 5' region of the DRD4 gene is stimulus specific in human neuronal cells.* Genes Brain Behav, 2013. **12**(2): p. 282-7.
184. Prince, M., et al., *Global mental health 1 - No health without mental health.* Lancet, 2007. **370**(9590): p. 859-877.
185. Braun, P.R., et al., *Genome-wide DNA methylation comparison between live human brain and peripheral tissues within individuals.* Transl Psychiatry, 2019. **9**(1): p. 47.
186. Brouha, B., et al., *Evidence consistent with human L1 retrotransposition in maternal meiosis I.* Am J Hum Genet, 2002. **71**(2): p. 327-36.
187. Tubio, J.M.C., et al., *Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes.* Science, 2014. **345**(6196): p. 1251343.
188. Seleme, M.d.C., et al., *Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity.* Proceedings of the National Academy of Sciences, 2006. **103**(17): p. 6611.
189. Lutz, S.M., et al., *Allelic heterogeneity in LINE-1 retrotransposition activity.* American journal of human genetics, 2003. **73**(6): p. 1431-1437.

190. De Cecco, M., et al., *L1 drives IFN in senescent cells and promotes age-associated inflammation.* Nature, 2019. **566**(7742): p. 73-78.
191. Lange, N.E., et al., *Alu and LINE-1 methylation and lung function in the normative ageing study.* BMJ Open, 2012. **2**(5): p. e001231.
192. Zhu, Z.Z., et al., *Repetitive element hypomethylation in blood leukocyte DNA and cancer incidence, prevalence, and mortality in elderly individuals: the Normative Aging Study.* Cancer Causes Control, 2011. **22**(3): p. 437-47.
193. Cho, N.Y., et al., *Hypermethylation of CpG island loci and hypomethylation of LINE-1 and Alu repeats in prostate adenocarcinoma and their relationship to clinicopathological features.* J Pathol, 2007. **211**(3): p. 269-77.
194. Wilson, A.S., B.E. Power, and P.L. Molloy, *DNA hypomethylation and human diseases.* Biochim Biophys Acta, 2007. **1775**(1): p. 138-62.
195. Schulz, W.A., *L1 retrotransposons in human cancers.* J Biomed Biotechnol, 2006. **2006**(1): p. 83672.
196. Stelzer G, R.R., Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Iny Stein T, Nudel R, Lieder I, Mazor Y, Kaplan S, Dahary D, Warshawsky D, Guan - Golan Y, Kohn A, Rappaport N, Safran M, and Lancet D., *The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analysis*. 2016: Current Protocols in Bioinformatics.
197. Kano, H., et al., *L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism.* Genes Dev, 2009. **23**(11): p. 1303-12.
198. McConnell, M.J., et al., *Mosaic copy number variation in human neurons.* Science, 2013. **342**(6158): p. 632-7.
199. Garcia-Perez, J.L., et al., *Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells.* Nature, 2010. **466**(7307): p. 769-73.
200. Yang, N. and H.H. Kazazian, Jr., *L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells.* Nat Struct Mol Biol, 2006. **13**(9): p. 763-71.
201. Tang, Z., et al., *Human transposon insertion profiling: Analysis, visualization and identification of somatic LINE-1 insertions in ovarian cancer.* Proc Natl Acad Sci U S A, 2017. **114**(5): p. E733-E740.
202. Komkov, A.Y., et al., *An advanced enrichment method for rare somatic retroelement insertions sequencing.* Mob DNA, 2018. **9**: p. 31.
203. Kazazian, H.H., Jr., et al., *Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man.* Nature, 1988. **332**(6160): p. 164-6.
204. Scott, E.C. and S.E. Devine, *The Role of Somatic L1 Retrotransposition in Human Cancers.* Viruses, 2017. **9**(6): p. 131.
205. Richardson, S.R., S. Morell, and G.J. Faulkner, *L1 retrotransposons and somatic mosaicism in the brain.* Annu Rev Genet, 2014. **48**(1): p. 1-27.
206. Hill, D.P., et al., *Gene Ontology annotations: what they mean and where they come from.* BMC Bioinformatics, 2008. **9 Suppl 5**(Suppl 5): p. S2.

207. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nature Protocols, 2008. **4**: p. 44.
208. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Res, 2009. **37**(1): p. 1-13.
209. Flasch, D.A., et al., *Genome-wide de novo L1 Retrotransposition Connects Endonuclease Activity with Replication.* Cell, 2019. **177**(4): p. 837-851.e28.
210. Faulkner, G.J. and V. Billon, *L1 retrotransposition in the soma: a field jumping ahead.* Mob DNA, 2018. **9**: p. 22.
211. Kurnosov, A.A., et al., *The evidence for increased L1 activity in the site of human adult brain neurogenesis.* PLoS One, 2015. **10**(2): p. e0117854.
212. Bundo, M., et al., *Increased l1 retrotransposition in the neuronal genome in schizophrenia.* Neuron, 2014. **81**(2): p. 306-13.
213. Gualtieri, A., et al., *Increased expression and copy number amplification of LINE-1 and SINE B1 retrotransposable elements in murine mammary carcinoma progression.* Oncotarget, 2013. **4**(11): p. 1882-93.
214. Carreira, P.E., et al., *Evidence for L1-associated DNA rearrangements and negligible L1 retrotransposition in glioblastoma multiforme.* Mob DNA, 2016. **7**: p. 21.
215. Nguyen, T.H.M., et al., *L1 Retrotransposon Heterogeneity in Ovarian Tumor Cell Evolution.* Cell Rep, 2018. **23**(13): p. 3730-3740.
216. Schauer, S.N., et al., *L1 retrotransposition is a common feature of mammalian hepatocarcinogenesis.* Genome Res, 2018. **28**(5): p. 639-653.
217. Graham, T. and S. Boissinot, *The genomic distribution of L1 elements: the role of insertion bias and natural selection.* J Biomed Biotechnol, 2006. **2006**(1): p. 75327.
218. Deininger, P.L., et al., *Mobile elements and mammalian genome evolution.* Curr Opin Genet Dev, 2003. **13**(6): p. 651-8.
219. Medstrand, P., L.N. van de Lagemaat, and D.L. Mager, *Retroelement distributions in the human genome: variations associated with age and proximity to genes.* Genome Res, 2002. **12**(10): p. 1483-95.
220. Bailey, J.A., et al., *Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: The Lyon repeat hypothesis.* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(12): p. 6634-6639.
221. Han, J.S., S.T. Szak, and J.D. Boeke, *Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes.* Nature, 2004. **429**(6989): p. 268-274.
222. Allen, E., et al., *High concentrations of long interspersed nuclear element sequence distinguish monoallelically expressed genes.* Proc Natl Acad Sci U S A, 2003. **100**(17): p. 9940-5.

223. Wallace, M.R., et al., *A de novo Alu insertion results in neurofibromatosis type 1.* Nature, 1991. **353**(6347): p. 864-6.
224. Bragg, D.C., et al., *Disease onset in X-linked dystonia-parkinsonism correlates with expansion of a hexameric repeat within an SVA retrotransposon in TAF1.* Proceedings of the National Academy of Sciences of the United States of America, 2017. **114**(51): p. E11020-E11028.
225. Ewing, A.D. and H.H. Kazazian, Jr., *Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans.* Genome Res, 2011. **21**(6): p. 985-90.
226. Rishishwar, L., et al., *Evidence for positive selection on recent human transposable element insertions.* Gene, 2018. **675**: p. 69-79.
227. Atallah, N., et al., *How Healthy Lifestyle Factors at Midlife Relate to Healthy Aging.* Nutrients, 2018. **10**(7).
228. Linker, S.B., et al., *Examining non-LTR retrotransposons in the context of the evolving primate brain.* BMC Biol, 2017. **15**(1): p. 68.
229. Dubnau, J., *The Retrotransposon storm and the dangers of a Collyer's genome.* Curr Opin Genet Dev, 2018. **49**(1879-0380 (Electronic)): p. 95-105.
230. Kazazian, H.H., Jr. and J.V. Moran, *Mobile DNA in Health and Disease.* N Engl J Med, 2017. **377**(4): p. 361-370.
231. Lopez-Otin, C., et al., *The hallmarks of aging.* Cell, 2013. **153**(6): p. 1194-217.
232. Aubert, G. and P.M. Lansdorp, *Telomeres and aging.* Physiol Rev, 2008. **88**(2): p. 557-79.
233. Eitan, E., E.R. Hutchison, and M.P. Mattson, *Telomere shortening in neurological disorders: an abundance of unanswered questions.* Trends Neurosci, 2014. **37**(5): p. 256-63.
234. Liu, M.Y., A. Nemes, and Q.G. Zhou, *The Emerging Roles for Telomerase in the Central Nervous System.* Front Mol Neurosci, 2018. **11**: p. 160.
235. Saretzki, G.C., *Does telomerase protein protect our neurons?* 2016: Institute for Cell and Molecular Biosciences, Newcastle Institute for Ageing, Campus for Ageing and Vitality, Newcastle University, UK.
236. Tomita, K.I., et al., *Changes in telomere length with aging in human neurons and glial cells revealed by quantitative fluorescence in situ hybridization analysis.* Geriatr Gerontol Int, 2018. **18**(10):1507-1512.
237. Arsenis, N.C., et al., *Physical activity and telomere length: Impact of aging and potential mechanisms of action.* Oncotarget, 2017. **8**(27): p. 45008-45019.
238. Frenck, R.W., Jr., E.H. Blackburn, and K.M. Shannon, *The rate of telomere sequence loss in human leukocytes varies with age.* Proc Natl Acad Sci U S A, 1998. **95**(10): p. 5607-10.
239. Starnino, L., et al., *Psychological Profiles in the Prediction of Leukocyte Telomere Length in Healthy Individuals.* PLoS One, 2016. **11**(10): p. e0165482.
240. Starr, J.M., et al., *Oxidative stress, telomere length and biomarkers of physical aging in a cohort aged 79 years from the 1932 Scottish Mental Survey.* Mech Ageing Dev, 2008. **129**(12): p. 745-51.

241. Zhang, J., et al., *Ageing and the telomere connection: An intimate relationship with inflammation.* Ageing Res Rev, 2016. **25**: p. 55-69.
242. Bonora, M. and P. Pinton, *Mitochondrial DNA keeps you young.* Cell Death Dis, 2018. **9**(10): p. 992.
243. Sun, N., R.J. Youle, and T. Finkel, *The Mitochondrial Basis of Aging.* Mol Cell, 2016. **61**(5): p. 654-666.
244. Zhang, R., et al., *Independent impacts of aging on mitochondrial DNA quantity and quality in humans.* BMC Genomics, 2017. **18**(1): p. 890.
245. Shokolenko, I.N., G.L. Wilson, and M.F. Alexeyev, *Aging: A mitochondrial DNA perspective, critical analysis and an update.* World J Exp Med, 2014. **4**(4): p. 46-57.