The implication of dynamic repetitive DNA in Alzheimer's disease

Rita Guimarães

2026

The implication of dynamic repetitive DNA in Alzheimer's disease

Rita Guimarães

This work was supported by a grant from Alzheimer Nederland (WE.03-2018-07) to Dr. F.M.J. Jacobs, Dr. H. Holstege and Dr. M.J.T. Reinders.

Printing of this thesis was financially supported by F.M.J. Jacobs and Alzheimer Nederland.

ISBN 978-94-6534-611-3

Cover design by Anton Estlund
Printed by proefschriftmaken || www.proefschriftmaken.nl

The implication of dynamic repetitive DNA in Alzheimer's disease

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
ten overstaan van een door
het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op donderdag 29 oktober 2026, te 13.00 uur

door **Rita Leonor Franco Paulo Guimarães**
geboren te Vila Franca de Xira, Lisboa

Promotiecommissie

Promotores:

dr. F.M.J. Jacobs	Universiteit van Amsterdam
prof. dr. ir. M.J.T. Reinders	Technische Universiteit Delft

Copromotor:

dr. H. Holstege	AMC-UvA

Overige leden:

prof. dr. M.P. Smidt	Universiteit van Amsterdam
prof. dr. P.J. Lucassen	Universiteit van Amsterdam
dr. A.A. Dijkstra	Universiteit van Amsterdam
dr. N.N. Kasri	Radboudumc
prof. dr. K. Sleegers	Universiteit Antwerpen
dr. E. de Wit	NKI

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

“Dream is destiny”
- Waking Life

Contents

Introduction 9

Chapter 1: Integration and systematic categorization of hidden structural variation in Alzheimer's disease-associated loci 25

Chapter 2: A hidden layer of structural variation in transposable elements reveals potential genetic modifiers in human disease-risk loci 49

Chapter 3: Functional assessment of human-specific SVAs in AD-associated genes *PICALM* and *PSEN1* 87

Chapter 4: Preliminary evidence for the association of a VNTR in the *RTEL1* gene with Alzheimer's disease risk using short-read sequencing data 109

General discussion 141

Appendix 151
- References 153
- Nederlandse Samenvatting 181
- Publications 183
- Acknowledgements 184
- About the Author 192



Introduction

The human genome as basis for complex diseases

The human genome, containing all the genetic information necessary for creating a human being, serves as a blueprint for an organisms' development, function, and reproduction. It determines individual characteristics, such as physical traits and susceptibility to diseases. Therefore, understanding the genome and its intricacies is imperative for unravelling the molecular mechanisms underlying human health and disease. The landmark Human Genome Project, completed in the early 2000s, provided the first ever comprehensive map of the human genome (Lander et al. 2001). Further advances have built upon this foundation, reaching greater resolution, and identifying critical elements that contribute to genetic diversity (Audano et al. 2019; Chaisson, Wilson, and Eichler 2015). The recent achievement of telomere-to-telomere sequencing represents the most complete view of the human genome (Nurk et al. 2022). It has uncovered previously inaccessible genomic regions, offering new insights into the complexity of genetic regulation and inheritance.

Genetic diseases arise from mutations in an individual's DNA, affecting gene function and leading to a wide spectrum of inherited disorders. These mutations can range from single nucleotide changes to alterations in entire genes or chromosomes. Genetic diseases can be classified into three broad categories: single-gene disorders, chromosomal disorders, and complex disorders. Single-gene disorders, also known as Mendelian diseases, result from mutations in a single gene, which causes a defect within the cell, and follow clear inheritance patterns. Chromosomal disorders, on the other hand, arise from abnormalities in chromosome number or structure, such as Down syndrome, which is caused by the presence of an extra copy of chromosome 21. Finally, complex disorders involve interactions between multiple genetic and environmental factors, making their inheritance patterns more complex. Alzheimer's disease (AD) is a prime example of a complex disorder with a strong genetic component. While ageing remains the greatest risk factor, genetic predisposition plays a crucial role in determining susceptibility. The presence of certain genetic variants significantly influences disease risk, while ongoing research continues to uncover additional genetic regions contributing to its pathology. Despite these insights, the precise etiology of AD remains largely unknown, posing a major challenge for the development of effective treatments. Therefore, understanding the genetic factors at play is critical to unraveling the disease's molecular underpinnings, which could pave the way for early intervention therapies.

Alzheimer's disease clinical manifestation and pathology

In 1906, Alois Alzheimer identified Alzheimer's disease, a neurodegenerative disorder that has since emerged as a major focus of medical research. Alzheimer described the strange case of a 51-year-old woman with unusual brain pathologies (Stelzmann, Norman Schnitzlein, and Reed Murtagh 1995). Initially viewed as a rare, presenile dementia, it was gradually recognized as the leading cause of dementia in the elderly population worldwide. Clinically, AD is characterised by progressive

memory impairment, cognitive decline, and behavioural changes, ultimately affecting the patient's daily life. The loss of independence in advanced stages also burdens family members and healthcare systems around the world.

The pathological hallmarks of AD include the extracellular deposition of amyloid-β (Aβ) plaques, and the presence of intracellular neurofibrillary tangles composed of phosphorylated tau protein (Busche and Hyman 2020) (**Figure 1A**). Discovered in the 80s, Aβ is formed with the cleavage of the transmembrane amyloid precursor protein (APP) (Glenner and Wong 1984; Selkoe 2008) (**Figure 1B**). First, part of the extracellular domain of APP is cleaved into the extracellular space. Then, γ-secretase, of which presenilin-1 (PSEN1) and -2 (PSEN2) are components, mediates the cleavage of APP in its transmembrane domain, producing soluble Aβ peptides. The propensity for Aβ oligomerization is dependent on the cleavage site. It occurs after amino acids 38, 40 or 42, but the Aβ42 peptide has the strongest tendency to oligomerize (L. Liu et al. 2019). It is currently believed that the increased production and/or reduced clearance of Aβ leads to toxic accumulation of Aβ protein aggregates, which in turn triggers a cascade of events leading to neuronal dysfunction and cell death – the amyloid hypothesis (Haass and Selkoe 2007; Selkoe and Hardy 2016). Furthermore, Aβ induces the spread of tau pathology, a microtubule-associated protein (Busche and Hyman 2020) (**Figure 1B**). Tau functions as a stabiliser of microtubules, by binding during the polymerization of tubulin. This process is regulated by the phosphorylation of tau, which in pathological conditions can be abnormally high, leading to microtubules breakdown and overall cytoskeleton disorganisation. The hyperphosphorylated tau will form aggregates in the cytoplasm, contributing to cell toxicity and eventual cell death (Gu and Liu 2020). These two molecules, Aβ and tau, are currently viewed as pathogenic contributors to neuronal dysfunction and synapse loss in AD, causing the characteristic brain atrophy observed in advanced stages.

The third key component of AD pathology is neuroinflammation, triggered by the accumulation of Aβ plaques and neurofibrillary tangles in the brain (Zotova et al. 2010). Initially, microglial cells, the major resident immune cells in the central nervous system, are in a resting phase (**Figure 1A**). Aβ can trigger the transition of microglia to an activated state, which involves morphologic changes, migration, and phagocytosis of the pathogen (Davalos et al. 2005). Clearance of Aβ by microglia is extremely important to prevent pathology, playing an initial neuroprotective role. However, persistent microglial activation can have pathogenic effects as well, shifting the actions of microglia from protective to harmful (Das Sarma 2014). In such cases, the exacerbation of inflammation becomes chronic, contributing to the progression of AD (Kinney et al. 2018).

Age is the primary risk factor for AD. Consequently, as the global population ages, the prevalence of AD is expected to rise, with estimates that the number of cases

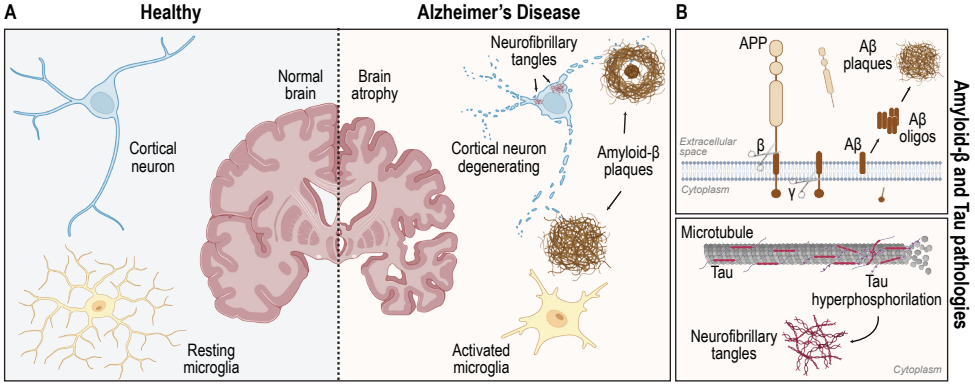


Figure 1: Pathological hallmarks of Alzheimer's disease. A, The brain of an AD patient presents with atrophy and enlarged ventricles. Molecularly, pathology includes extracellular amyloid-β plaques and intracellular neurofibrillary tangles. This causes chronic activation of microglial cells and neuronal cell death. **B,** The amyloid-β pathway involves the processing of APP by β- and γ-secretases. Cleavage by γ-secretase leads to the formation of amyloid-β peptides with a tendency to oligomerize, eventually resulting in amyloid-β plaques. In parallel, tau protein becomes hyperphosphorylated, which induces continuous aggregation of hyperphosphorylated tau, destabilising the neuron cytoskeleton and contributing to cell toxicity. Created with BioRender.com.

will triple by 2050 (https://www.alzint.org/resource/numbers-of-people-with-dementia-worldwide/). Current efforts focus predominantly on disease-modifying therapies that aim to slow the progression of disease by targeting amyloid pathology specifically. Examples include FDA-approved treatments like Aducanumab (Ali, Gupta, and Chawla 2022) and Lecanemab (van Dyck et al. 2023) in the United States, as well as Donanemab, which recently demonstrated promising cognitive benefits and is under regulatory review in Europe (Song et al. 2024). These therapies underscore the importance of early detection and intervention in AD, as their efficacy is most pronounced in the early stages of the disease. Currently, 164 clinical trials are underway, testing 127 drugs, with 48 trials in Phase 3 (Cummings et al. 2024), crucial for regulatory approvals. Despite this progress, significant challenges persist, including the modest efficacy of current treatments and the limited focus on tau and other non-amyloid pathways. The greatest obstacle, however, lies in treating the disease once symptoms are manifested, as neurodegeneration and molecular pathologies are already well underway by this point. This underscores the urgent need for improved methods to predict disease risk at an earlier age, enabling earlier detection of AD which opens up possibilities to start intervening therapies before the onset of clinical symptoms. Since genetic factors are believed to substantially influence the susceptibility and age of onset of AD, understanding these factors offers a pathway toward preventative therapies and personalized approaches to halt disease progression.

The genetics of Alzheimer's disease and missing heritability

AD can be divided into early-onset (EOAD) and late-onset (LOAD) based on the relative age of onset of molecular pathology and symptoms. EOAD, typically presents before the age of 65 and only accounts for 10% of all AD cases, leading to a more rapid disease progression as it directly affects the amyloid pathway (Andrade-Guerrero et al. 2023; Cacace, Sleegers, and Van Broeckhoven 2016). EOAD is almost entirely driven by genetic factors, and a small fraction of cases show an autosomal dominant inheritance due to mutations in *APP*, *PSEN1*, and *PSEN2* (Andrade-Guerrero et al. 2023; Cacace, Sleegers, and Van Broeckhoven 2016; Wingo et al. 2012). In contrast, LOAD, accounting for the majority of cases, manifests after the age of 65. The heritability of LOAD is estimated to be between 60-80% (Gatz et al. 2006) and a great number of genetic variants (i.e. haplotypes) have been identified that are associated with an increased risk for AD (Bellenguez et al. 2022). The most consistent of them is the ε4 risk allele of apolipoprotein E (*APOE*) which has been associated with an estimated 3-4 times increased risk (Jansen et al. 2019). At present, 80 haplotypes are known to have an association with LOAD. Importantly, the presence of any particular AD-associated haplotype is not predicting that an individual will develop AD, but it merely increases the chance of that individual to develop AD in his/her lifetime compared to an individual that does not have the AD-associated haplotype. The different combination of alleles across these loci contribute to the genetic complexity and individual variability observed in LOAD cases. Therefore, the late-onset form is characterised by a more gradual progression and a higher degree of heterogeneity in clinical presentations, complicating both diagnosis and treatment.

Research methodologies to study AD have evolved in parallel with advances in technology and molecular biology. Particularly the field of genetics has revolutionised AD research, beginning with linkage analysis in familial cases of EOAD, leading to the identification of causative mutations (Pericak-Vance et al. 1991). This was followed by genome-wide association studies (GWAS), which have been invaluable for identifying genetic loci that contribute to disease susceptibility in LOAD (Bertram et al. 2007; J.-C. Lambert et al. 2013; Seshadri et al. 2010). These studies rely on high-throughput genotyping technologies, such as microarrays or short-read sequencing, to analyse hundreds of thousands to millions of SNPs across large case-control cohorts (Bush and Moore 2012). The first major discovery came with the identification of the *APOE*-ε4 as a significant risk factor, a finding initially identified through candidate gene studies (Strittmatter et al. 1993) but later robustly confirmed and contextualized by GWAS (Jansen et al. 2019; J.-C. Lambert et al. 2009). GWAS typically identify common variants with small effect sizes which require large sample sizes to achieve statistical significance (Bush and Moore 2012). This need for large-scale data has driven a steady increase in the size of GWAS cohorts over time, with recent studies often including tens of thousands of participants. These efforts have made it possible to identify the genetic loci of *CD33*, *CLU*,

CR1 and *PICALM* as associated to AD (Bertram et al. 2008; Harold et al. 2009; J.-C. Lambert et al. 2009). Soon after, to address the demand for larger datasets, meta-GWAS analyses have aggregated data from multiple studies and biobanks, dramatically increasing sample sizes and identifying loci such as *BIN1* (Seshadri et al. 2010); *ABCA7, MS4A4A, EPHA1, CD33* and *CD2AP* (Hollingworth et al. 2011); *INPP5D, HLA, TREM2, PILRA, PTK2B, FERMT2, RIN3/SLC24A4* and *CASS4* (J.-C. Lambert et al. 2013); *MAPT/KANSL1* (Jun et al. 2016); and *PLCG2*, *ABI3*, and *TREM2* (Sims et al. 2017). Finally, methods like AD-by-proxy have been employed, where individuals without a direct diagnosis of AD but with a family history of the disease are included in the analysis. This approach has not only replicated known associations but also uncovered many novel loci (Bellenguez et al. 2022; Jansen et al. 2019; de Rojas et al. 2021; Schwartzentruber et al. 2021), significantly advancing our understanding of the genetic architecture of AD.

Despite substantial progress in identifying genetic risk factors for AD, a significant fraction of its genetic contribution remains unaccounted for. This challenge is referred to as the problem of missing heritability, which suggests that currently known genetic variants account for only a fraction of the disease's genetic predisposition (Ridge et al. 2013). One contributor to missing heritability is the inability of GWAS to detect rare variants, as it primarily detects common variants. Rare variants often have larger effect sizes than common variants but are not well-captured by standard GWAS platforms due to their low frequency in the population (Bush and Moore 2012). In addition, another potential contributor to the missing heritability problem may not be a single nucleotide change, but rather larger genetic variations of repetitive DNA. Tandem repeats (TRs) and transposable elements (TEs) are particularly relevant, as they are highly polymorphic and are a frequent source of inter-individual genetic variability. However, these elements have been largely overlooked in past GWAS due to technical limitations in short-read sequencing, which struggles to accurately resolve repetitive sequences. Therefore, structural variations of repetitive DNA may be hidden in regions associated with AD where the association to disease is not yet explained, i.e. SNPs are non-coding and thus not predicted to have any functional outcome. Here, the actual causal genetic variant is present within a region of high linkage disequilibrium (LD), a phenomenon where variants in close genomic proximity are inherited together. In these cases, the non-coding SNP serve only a flag for the disorder-associated genetic locus (Edwards et al. 2013). Therefore, investigating the role of repetitive DNA may uncover previously hidden genetic risk factors, helping to explain the missing heritability problem in AD.

Towards identifying causal genetic factors: focussing on repetitive DNA

Of 3 billion base-pairs of DNA that are together forming the human genome, only 1.5% is coding (**Figure 2A**). The remaining 98,5% of non-coding DNA, previously termed as "junk" DNA, is increasingly being recognized as a crucial component of the human genome (Lander et al. 2001). The majority of non-coding DNA is made

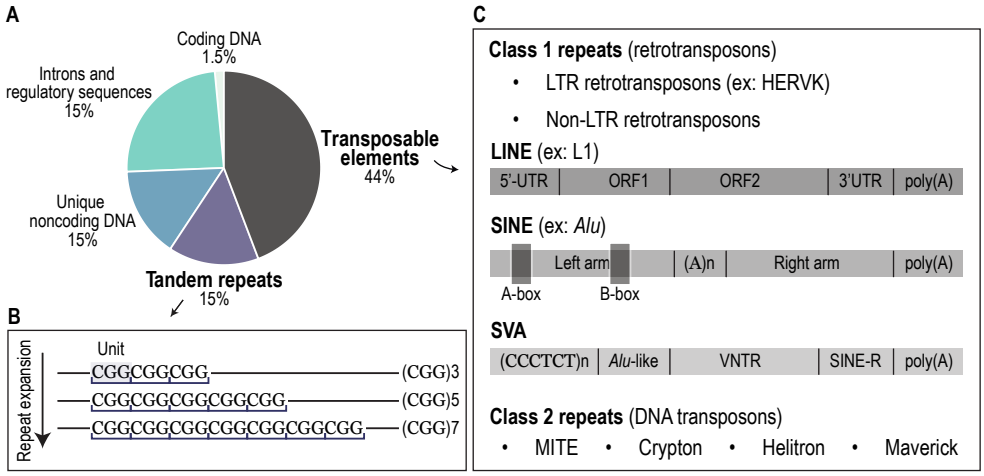


Figure 2: Repetitive DNA in the human genome. A, The main components of the human genome are coding DNA, introns and regulatory sequences, unique noncoding DNA, tandem repeats and transposable elements. **B,** Tandem repeats are composed of units repeated in tandem a number of times that are prone to expand. **C,** Transposable elements originate from viral DNA and are dispersed throughout the genome. These contain class 1 (retrotransposons) and class 2 elements (DNA transposons). Showing the structure of three non-LTR retrotransposons that are currently still active in the human genome: L1, Alu and SVA. Created with BioRender.com.

up of repetitive DNA sequences, which can have significant implications for genetic stability, inter-individual differences in gene expression, and disease susceptibility. Repetitive DNA can be classified into two main types based on their genomic organisation: tandem repeats, which include short sequences repeated in tandem next to each other; and transposable elements, or interspersed repeats, which are dispersed throughout the genome.

TRs are categorised by the length of their repeat unit (or motif). Short tandem repeats (STRs), or microsatellites, with units of 2 to 6 base pairs (bp) and variable number tandem repeats (VNTRs), or minisatellites, with motifs larger than 7 bp (Bakhtiari et al. 2021). TRs tend to be unstable in their number of units because of their repetitive nature (**Figure 2B**), which increases their susceptibility to length variations caused by unequal recombination or slippage events during DNA replication and/or transcription (Richard and Pâques 2000). Therefore, TRs significantly contribute to the genetic diversity in the human population.TEs constitute almost half of the human genome and have played a major role in its evolution (Deininger et al. 2003; Lander et al. 2001). TEs are remnants of ancient retroviruses which have retained the capability to multiply their DNA-sequences throughout the genome. They can be divided into two classes based on their mechanism (**Figure 2C**). Class 1 are retrotransposons, which mobilise via an RNA intermediate in a copy-and-

paste mechanism, whilst class 2 are DNA transposons, which use a cut-and-paste mechanism (Wicker et al. 2007). However, most of the TEs have lost their capability of moving throughout the genome, mostly due to accumulating mutations over evolutionary time. The only three families of TEs that remain actively mobile in the human genome are L1, *Alu* and SVA (Batzer et al. 1991; Kazazian et al. 1988; Ostertag et al. 2003) (**Figure 2C,** *Class 1 repeats*). These elements can integrate anywhere in the genome, resulting in insertional polymorphisms, i.e. are either present or absent in individuals. In parallel to insertional polymorphisms, TEs that are fixed in the human population, i.e. present in all individuals, can also contain structural variations, resulting in a rich source of inter-individual genetic variation.

Both TRs and TEs have had a profound impact on the regulatory landscape of the human genome. TRs play a key role in regulating gene expression. In promoters, TRs can modulate transcription by affecting transcription factor binding and epigenetic modifications, such as DNA methylation (**Figure 3A**). Within genes, intronic TRs can influence alternative splicing, leading to the production of diverse transcript isoforms that results in altered protein functions (**Figure 3B**). Additionally, TRs in untranslated regions (UTRs) contribute to messenger RNA (mRNA) stabilization, localization, and translation, allowing fine-tuned gene regulation. The inherent variable nature of TRs enables them to act as dynamic modulators to fine-tune gene expression, contributing to a rapid species adaptation and evolution (Gemayel et

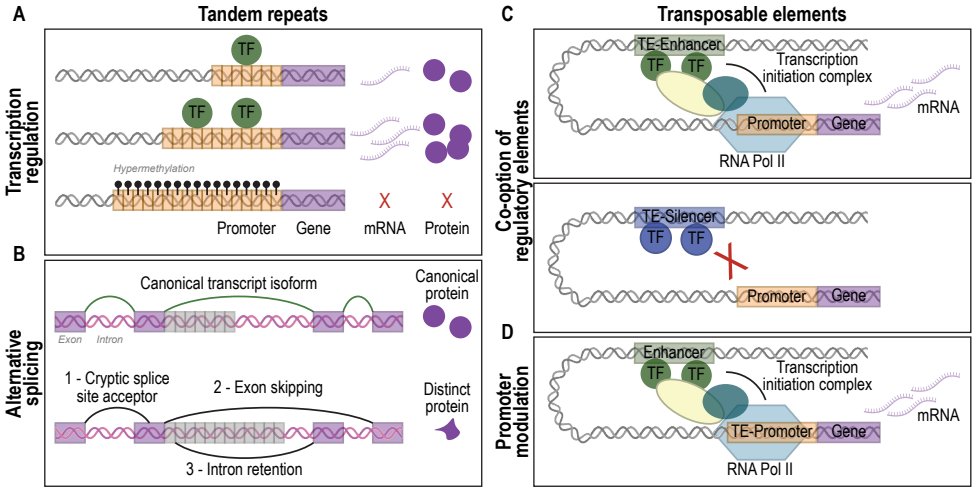


Figure 3: Influence mechanisms of tandem repeats and transposable elements on gene expression. A, TRs at the promoter sites regulate gene transcription through transcription factor binding and epigenetic modifications. Increasing lengths of tandem repeats may induce hypermethylation of the promoter region, silencing gene expression. **B,** TRs located within genes can influence transcription by inducing alternative splicing. **C,D,** Transposable elements have been co-opted into regulatory networks, functioning as enhancers or silencers (**C**), and as novel promoter sequences (**D**). Created with BioRender.com.

al. 2010; Sulovari et al. 2019). Similarly, TEs contribute to gene regulation by acting as mobile regulatory elements. Many TEs harbour transcription factor binding sites and function as enhancers, silencers, or alternative promoters (**Figure 3C,D**). Through their insertional activity, TEs can alter regulatory networks, particularly in a tissue-specific manner, capacity which has been co-opted during primate evolution (Chuong, Elde, and Feschotte 2017; Cordaux and Batzer 2009; Pontis et al. 2019; Sundaram and Wysocka 2020; Trizzino, Kapusta, and Brown 2018). Overall, through structural variations, meaning changes in sequence length, these repetitive DNA elements become a rapidly modifiable platform of the regulatory landscape of the human genome.

Structural variation in TRs and TEs are frequently associated with human brain disorders

Accumulating evidence suggests that both TRs and TEs are genetic modulators which took an instrumental part in shaping human brain function and behaviour (Baillie et al. 2011; Fondon et al. 2008; Gianfrancesco, Bubb, and Quinn 2017; Larsen et al. 2018; Savage et al. 2013, 2014). For example, active retrotransposition of LINE-1 elements is known to occur in neuronal lineages during brain development, causing somatic mosaicism within brain regions (Coufal et al. 2009; Muotri et al. 2005), especially in the hippocampus (Coufal et al. 2009). It was shown that these elements integrate in regions of protein-coding genes active in the brain, associated with neurogenesis and synaptic function (Baillie et al. 2011). VNTRs as well, were found to be predominantly located in genes highly expressed in the brain (Linthorst et al. 2020). However, despite these positive contributions, TRs and TEs also present challenges as they are implicated in various neurological and psychiatric disorders (Ahmadi et al. 2020; Chintalaphani et al. 2021). TRs are the cause of more than 40 neurological diseases, caused by altered repeat lengths that extend beyond a certain threshold, disrupting gene expression (Chintalaphani et al. 2021). An example of these is the “GGGGCC” TR located in the 5’-UTR of *C9orf72* gene, associated with frontotemporal dementia and amyotrophic lateral sclerosis. The TR expansion results in toxic RNA gain-of-function, which forms nuclear RNA foci, and repeat-associated non-AUG (RAN) translation, causing intranuclear neuronal inclusions in the cerebellum, hippocampus and cortex (Gendron and Petrucelli 2018). Moreover, a VNTR in *ABCA7* gene is associated with LOAD. With a repeat unit of 25 bp, it can expand over 10,000 bp in length, causing alternative splicing, leading to dysfunctional proteins (De Roeck et al. 2018).

TEs have also been implicated in various neurodegenerative disorders (Larsen et al. 2017; Wanhe Li et al. 2012; Ravel-Godreuil et al. 2021; Thomas et al. 2020), and specifically with AD (Evering et al. 2023). For instance, the expression of Human Endogenous Retrovirus (HERV), a retrotransposon, is significantly associated with multiple sclerosis (MS) (Morandi et al. 2017). Particularly TEs from the HERV-W retrovirus family, associated with increased risk of MS, expression levels of which

are linked with the severity of the disorder (García-Montojo et al. 2014). TEs from the HERV-K family was shown to be elevated in brains of patients with amyotrophic lateral sclerosis (ALS) (Douville et al. 2011; Wenxue Li et al. 2015), frontotemporal dementia (FTD) (Phan et al. 2021) and AD (Dembny et al. 2020). Furthermore, most elements belonging to the HERV-K family, although not exclusively, were differentially expressed in AD patients and were correlated with the expression of AD-associated genes (Dawson et al. 2023). It is thought that the increased levels of HERV-K leads to neurotoxicity through the production of envelope glycoproteins that trigger inflammatory responses (Wenxue Li et al. 2015). Finally, *Alu* elements, another class of retrotransposons, have been associated with AD by inducing genetic deletions or aberrant transcription of *TOMM40*, which plays a role in mitochondrial function (Larsen et al. 2017). Here, the authors further propose the *Alu* neurodegeneration hypothesis, in which *Alu* elements present in mitochondrial genes may cause an imbalance of mitochondria homeostasis in neuronal cells, ultimately resulting in cell death.

TRs and TEs have long been recognized as major contributors to human brain evolution and function, influencing neuronal development and connectivity. These repetitive DNA elements are particularly enriched in genes closely associated with neuronal processes, meaning that any changes in their size, whether expansions or contractions, can profoundly impact gene regulation. Such alterations may disrupt the finely tuned regulatory mechanisms governing brain function, potentially leading to disease. This delicate balance underscores the importance of studying structural variations in repetitive DNA, not only the element itself but rather the inter-individual genetic variability arising from these elements, as they may also be central to the genetic mechanisms driving AD and other neurodegenerative disorders.

Linking age with increased influence of TRs and TEs on their genomic environment

TEs are typically silenced in healthy cells by mechanisms such as transcriptional repression and RNA degradation. However, as we age the normal epigenetic controls over these elements weaken, unleashing their full regulatory potential that grows more pronounced over time (Wanhe Li et al. 2013; Yushkova and Moskalev 2023). Given that age is the main risk factor for neurodegenerative disorders, the epigenetic state of TRs and TEs are also tightly correlated to age-associated neurodegenerative disorders (Ahmadi et al. 2020; Larsen et al. 2017; Thomas et al. 2020). Not only through age-related epigenetic changes that activate TEs, but also the dysfunction of regulatory proteins like TDP-43 and tau can lead to TE activation. TDP-43 is an RNA-binding protein that plays a significant role in the suppression of TE-derived transcripts, preventing their activation. In certain cases of neurodegenerative disorders, including AD, loss of TDP-43 leads to overexpression of TE-derived transcripts, suggesting that lack of TE control contributes to disease pathology (Wanhe Li et al. 2012). Similarly, tau protein has been implicated in TE activation

through remodelling of chromatin in AD (Guo et al. 2018). The cumulative effect of tau and TDP-43 proteins appear to be critical in maintaining genomic integrity from TE activation. Age-related epigenetic changes, including DNA methylation and histone modifications, further exacerbate this vulnerability by creating a permissive environment for TE activation. Their activity, in turn, causes DNA damage, cellular stress and inflammation, all of which are hallmarks of neurodegenerative disorders (Ravel-Godreuil et al. 2021; Thomas et al. 2020). Ultimately, the progressive exposure of the latent gene-regulatory potential of the repetitive DNA elements in our genome upon ageing may cause an increasingly severe imbalance in brain homeostasis, culminating in cellular toxicity, impairment, and cell death. Given that this gene-regulatory potential may vary according to the structural variation of the repetitive DNA element, understanding the full extent of variability of these elements is, therefore, essential in the context of a rapidly ageing population.

New technical advances allow for more comprehensive identification of AD- associated genetic variants

Only in recent years has the complete picture of structural variation of repetitive DNA come to light. This is due to the highly repetitive nature and the large size of some of these elements, especially VNTRs, as these can reach 10 kb. In the 80s, Southern blotting was the primary method for VNTR genotyping (Jeffreys, Wilson, and Thein 1985), but it was a labour-intensive and time-consuming method, also lacking precision for complex repeat structures. Genotyping then became more efficient with the polymerase chain reaction (PCR) (Saiki et al. 1985), although it still had limitations in accurately characterising extensive or highly repetitive VNTRs. In the early 2000s, the development of next-generation sequencing (NGS) provided a higher throughput, but their short read length of ~150 bp often does not span the full VNTR (Chaisson, Wilson, and Eichler 2015; Mardis 2013). It was only with the development of long-read sequencing technologies in the 2010s and 2020s that generate reads spanning 10-100 Kb and precisely sequence the entire spectrum of genetic variations, including complex TRs and TEs is achievable (Audano et al. 2019; Auton et al. 2015; Chaisson, Wilson, and Eichler 2015; Ebert et al. 2021; Ewing et al. 2020; Linthorst et al. 2020). This breakthrough is invaluable for complex diseases, as it offers an unprecedented view of the genome's most challenging and variable genetic regions.

Genetic association studies, particularly for AD, have continued to rely on NGS technologies, such as whole-genome and whole-exome sequencing. These techniques have provided unprecedented insights into the genetic landscape of AD, revealing rare variants and new risk genes (Bis et al. 2020; Holstege et al. 2022; Prokopenko et al. 2021). However, despite these advances, NGS has limitations in resolving repetitive and structurally complex regions of the genome, which may potentially contribute to the missing heritability of AD. The use of long-read sequencing offers a unique opportunity in this context, as it provides a

comprehensive view of each genetic locus, including variations in repetitive DNA (De Roeck et al. 2018; Salazar et al. 2023). Insights that might hold the key to fully unravelling the genetic architecture of Alzheimer's disease, bridging gaps in our understanding of its heritability and shedding light on novel therapeutic targets.

Summary of experimental chapters

In this thesis we explore the role of repetitive DNA in the genetic architecture of Alzheimer's disease, focusing on how structural variations in TRs and TEs may contribute to disease susceptibility. While GWAS have identified numerous risk loci for AD, a significant portion of its heritability remains unexplained. We aim to uncover hidden genetic variations within AD-associated regions and assess their potential gene-regulatory functions by integrating genetic, epigenetic, and transcriptomic analyses. Through this work, we hope to provide novel insights into the genetic basis of AD.

In **chapter 1** we begin with a comprehensive overview of AD-associated loci, considering structural variants of repetitive DNA present in each locus. Using a published GWAS we performed a fine-mapping analysis and identified 29 genomic risk loci (GRL) associated with AD. Within the boundaries of the GRL, we integrated a dataset of structural variants and found 202 variants in 16 GRL. Notably, we observed that clusters of intergenic GWAS-identified SNPs surround structurally variable SVAs (SV-SVAs), thus we prioritise these loci for further functional analysis.

Chapter 2 explores structural variation in TEs shown to be active in a neuronal environment, and their association with neurological disorders. We used published long-read sequencing and PCR analysis to uncover the high level of structural variation of SVA elements throughout the human genome. Given that a substantial number of SV-SVA reside in loci associated with neurological disorders, we investigated the functional impact of these elements in 3 loci, *BIN1* and *CD2AP*, associated with Alzheimer's, and *BCKDK*, associated with Parkinson's disease. Genetic deletions of the SV-SVAs were engineered in human embryonic stem cells using CRISPR-Cas9 system and generated cortical organoids to provide a neuronal environment. Epigenetic marks were altered upon *BCKDK* SV-SVA deletion, in genes nearby the deletion site, and transcriptomic analysis revealed differential expression of *FUS*, 75 kb way. No epigenetic or transcriptomic changes were seen in *BIN1*-SVA KO. As for the intronic *CD2AP*-SVA, RNA CaptureSeq was performed, revealing reduced expression levels for all *CD2AP* exons downstream of the SVA insertion site. Taking together, our findings unravel a new layer of hidden genetic variation important to consider for risk of complex diseases.

In **chapter 3** we continue the work described in **chapter 1**, by providing a functional assessment of two SV-SVAs present in or upstream the AD-associated genes *PSEN1* and *PICALM*, respectively. Here, we use the same method as described

in **chapter 2**: generated SV-SVA KO cortical organoids and performed RNA CaptureSeq for a high resolution of the loci transcriptome. There was no evidence of a regulatory effect by the SV-SVA in *PICALM* gene in the context of a neuronal environment. However, upon deletion of the intronic SV-SVA in *PSEN1*, preliminary findings revealed elevated expression levels of transcripts associated with nonsense mediated decay.

In **chapter 4** we dive into a genetic association study to identify VNTRs associated with AD risk. We use a short-read sequencing dataset from the ADSP cohort to detect VNTRs whose length is either increased or decreased in most of the AD cases or in a subset, compared to controls. We discovered 52 VNTRs with a suggestive association to disease, of which 5 were replicated in an independent long-read sequencing dataset consisting of AD cases and cognitively healthy centenarians. We highlight the most promising AD-associated VNTR located in a regulatory element within the *RTEL1* gene, in the promoter region of *TNFRSF6B* and in the fusion transcript *RTEL1-TNFRSF6B. RTEL1* is an ageing-related gene involved in cell homeostasis, whilst *TNFRSF6B,* which functions as an immunosuppressive, has been associated with AD. While shorter lengths of the VNTR associates with AD, longer lengths correlate with a protective GWAS SNP rs55765053 in *TNFRSF6B*, which in turn is in modest LD with the main protective SNP rs6742. Further functional analysis should be conducted to validate our preliminary findings.

1

Integration and systematic categorization of hidden structural variation in Alzheimer's disease-associated loci

Rita LFP Guimaraes[1,2], Elena R. Blujdea[1], Marcel JT Reinders[3], Henne Holstege[2,3,4], Frank MJ Jacobs[1,4$]

1 Evolutionary Neurogenomics, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands
2 Section Genomics of Neurodegenerative Diseases and Aging, Department of Clinical Genetics, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands
3 Delft Bioinformatics Lab, Delft University of Technology, The Netherland
4 Amsterdam Neuroscience, Complex Trait Genetics, University of Amsterdam, Amsterdam, The Netherlands

$ corresponding author

Abstract

Predicting a person's risk to develop Alzheimer's disease poses a profound challenge due to its complex genetic nature. Despite efforts to identify genetic causative variants, GWAS mainly identifies non-coding SNPs, leaving many AD-associated loci without a direct causal link. To address this, we explored the role of structural variations in repetitive DNA elements, which have been previously linked to various neurological diseases, including AD. In this study, we performed a fine-mapping analysis with the web-based tool FUMA on summary statistics from de Rojas et al. (2021). Here, in parallel to the annotation of SNPs in the context of genomic position, regulatory function, and 3D chromatin interactions, we integrated positional mapping of structural variants from Audano et al. (2019) and further characterised the type of repetitive DNA that gives rise to the variation. Overall, we identified 202 structural variants in 16 out of 22 highly replicated AD genomic risk loci. *Alu* elements are a major source of variation throughout the loci, but structurally variable SVAs are present in four highly replicated AD-associated loci, *PICALM*, *BIN1*, *CD2AP* and *MS4A4A,* surrounded by high risk GWAS-identified SNPs. Thus, we hypothesise that this previously hidden type of structural variation in AD-associated loci may harbour gene-regulatory variation that may be the actual driver of the reported disease-association for some of the AD-risk loci.

Introduction

Alzheimer's disease (AD) is a genetically complex neurodegenerative disease. It is the most common form of dementia worldwide, with symptoms consisting of progressive decline of memory and cognitive abilities (Soria Lopez, González, and Léger 2019). Age is the primary risk factor for sporadic AD and its heritability is estimated to be between 60-80% (Gatz et al. 2006), though the genetic mechanisms underlying AD remain unknown. In recent years there has been a high influx of genome-wide association studies (GWAS) and meta-analyses that reveal an abundance of AD-associated single nucleotide polymorphisms (SNPs) (Bellenguez et al. 2022; Jansen et al. 2019; Kunkle et al. 2019; de Rojas et al. 2021; Schwartzentruber et al. 2021). Due to these studies, sporadic AD is defined by multiple, common variants with low effect sizes that are typically not the direct culprit for the disease, but are rather a representative for the locus (Edwards et al. 2013; Jansen et al. 2019). This problem of missing heritability may be caused by the complexity of the loci, which is not fully captured by GWAS. Thus, uncovering the true architecture of AD-associated loci will allow for the discovery of genetic causal variants.

Currently there are ~84 genetic loci associated with the risk of AD in the European population (Bellenguez et al. 2022), making the prioritisation of these loci essential. To that end, fine-mapping of AD-associated SNPs through functional annotations is performed as a post-GWAS analysis. Fine-mapping uses associations of a genomic region with a trait and assumes that at least one causal variant exists in that region (Broekema, Bakker, and Jonkers 2020). The areas can be partitioned into loci that have approximately independent effects on traits allowing for the determination of disease-linked effects (Broekema, Bakker, and Jonkers 2020; Hutchinson, Asimit, and Wallace 2020). These techniques of fine-mapping only focus on potential disease-causing candidate SNPs that, through linkage disequilibrium (LD), are in the same haplotype (Hutchinson, Asimit, and Wallace 2020). However, instead of a single nucleotide change, larger structural variations of the genome, such as deletions, insertions, or inversions of more than 50 bp, might be the driver of complex diseases such as AD.

Non-coding regions are rich in transposable elements, and because of their repetitive nature most variations that may exist in transposable elements have remained unexplored and were not taken into account when fine-mapping GWAS analyses. However, it has become clear that the gene-regulatory properties of transposable elements (TEs) have widely integrated in mammalian genomes, and TEs have been associated with both normal gene regulation and gene dysfunction (Gloss and Dinger 2018; Lander et al. 2001). The extent of variation that exists in the non-coding part of our genome has come to light through long-read sequencing methodology, exposing a vast amount of structural variants (SVs) (Audano et al. 2019; Chaisson, Wilson, and Eichler 2015; Linthorst et al. 2020).

SVs can have stable and heritable effects on gene expression. For instance, structural variation of a variable number of tandem repeat (VNTR) was found in the *ABCA7* gene, a highly replicated AD-associated locus. Here, the expansion of the VNTR, in LD with a nearby risk SNP (rs3764650; D'=0.92 and r^2=0.23), causes alternative splicing, leading to a dysfunctional protein or nonsense-mediated decay (De Roeck et al. 2018). To date, the *ABCA7*-VNTR is the only variation of its kind to be associated with AD specifically.

Even when they reside outside of the coding part of a gene, VNTRs can carry over some gene-regulatory properties to genes nearby, often in a length-dependent manner (Bakhtiari et al. 2021). The high propensity of tandem repeats in important gene-regulatory regions to undergo structural changes over time, may have played an important role in accelerating the evolution of species (Gemayel et al. 2010). Also in the human species, a high number of human-specific variations of VNTRs exist in neuronal genes which were shown to be differentially expressed between humans and chimpanzees (Sulovari et al. 2019). One of the sources of human-specific VNTRs is harboured in a retrotransposon called SINE-VNTR-*Alu* (SVA) elements, a hominid-specific family of transposable elements (H. Wang et al. 2005). The VNTR component of SVAs is a major source of structural variation (Audano et al. 2019; van Bree et al. 2022).

SVAs are currently still active in the human genome and give rise to new insertional polymorphisms in the human population (Cordaux and Batzer 2009; H. Wang et al. 2005). New SVA insertions can give rise to genetic variation but can also cause genomic instability and diseases (Hancks et al. 2009; Petrozziello et al. 2020; Pfaff et al. 2023; Pfaff, Singleton, and Kõks 2022). When inserted in a gene, SVAs can cause altered splicing patterns and loss-of-function mutations (Aneichyk et al. 2018; Pfaff, Singleton, and Kõks 2022). When in intergenic regions, SVAs also have the potential to regulate gene expression of nearby genes (Jacobs et al. 2014; Savage et al. 2013, 2014). As an evolutionary consequence of their incorporation in our genome, during primate evolution, SVAs have integrated in gene-regulatory networks (Jacobs et al. 2014; Pontis et al. 2019). Functional properties of VNTRs and VNTR-containing elements, such as the SVA, with high levels of hidden intrinsic structural variation and strong gene regulatory properties, establish VNTRs as prime candidates for SVs causal to gene dysregulation in AD-associated loci.

In this study we used published GWAS meta-analysis summary statistics (de Rojas et al. 2021) to perform a post-GWAS analysis using the web-based tool FUMA (Functional Mapping and Annotation of Genome-Wide Association Studies) (Watanabe et al. 2017). FUMA uses the summary statistics and pre-calculated LD structure to characterise genomic risk loci (GRL) associated with a trait/disease. Furthermore, FUMA annotates candidate SNPs in the context of genomic position, regulatory function, and 3D chromatin interactions. We took advantage of the

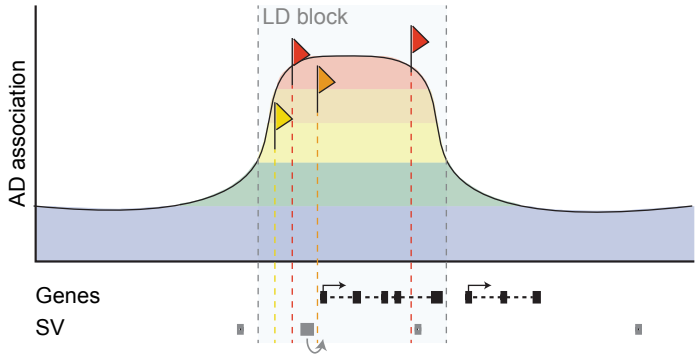


Figure 1: Schematic of hypothesis. In a regional plot of an AD-associated locus, significant risk SNPs (flags) are in a region of high linkage disequilibrium (LD block). Shading areas of cold to warm colours mean increasing association to disease risk. The gene within the LD block is associated with AD, though the risk SNPs are non-coding. Larger structural variations recently uncovered are also located within the LD block and are surrounded by the risk SNPs. We hypothesise that it is the structural variation that drives disease-association (grey arrow) instead of the risk SNP itself. Non-coding risk SNPs only flag the locus.

wide range of analyses available in FUMA to develop a prioritisation method that incorporates positional mapping of structural variation (Audano et al. 2019) in the GRL. We hypothesise that some of the AD-associated SNPs are tightly associated with hidden structural variations in repetitive DNA elements which we argue then to be the actual causative variant involved with gene dysregulation that leads to AD (**Figure 1**).

Results

Alzheimer's disease is a complex disease with new studies reporting new variants and implicating new genes yearly. Thus, to perform a prioritisation of AD-associated loci we used the most up-to-date meta-analysis from de Rojas et al. (2021). The study is composed of individuals from GR@ACE, IGAP and UK Biobank (n = 409,435). We applied FUMA, a web-based tool, to the summary statistics of the referred study to identify potential risk loci with genome-wide significant lead SNPs for AD. A total of 29 genomic risk loci (GRL) were reported (**Table 1**), harbouring 829 independent significant SNPs with a genome-wide significant association with AD status (*P*-value < 5 x 10^{-8}). They were in linkage disequilibrium ($r^2 \geq 0.8$) with other candidate SNPs within each locus (**Supplemental Table S2**). The defined

Table 1: 29 genomic risk loci containing SNPs present in the GWAS catalog of Alzheimer's disease. Table of the full output from FUMA with 29 GRL containing SNPs present in the GWAS catalog of Alzheimer's disease. *P*-values of top lead SNP from de Rojas et al. (2021). Coordinates for the GRL are based on the human genome reference GRCh37. SNPs: number of SNPs within the GRL; GWAS SNPs: number of GWAS-tagged candidate SNPs within the GRL; Independent Sig SNPs: number of independent significant SNPs within the GRL; Lead SNPs: number of lead SNPs within the GRL.

GRL	rsID	p-value	Start	End	SNPs	GWAS SNPs	Independet Sig SNPs	Lead SNPs
1	1:207,802,552	$7.297x10^{-31}$	207,368,589	207,875,537	94	90	12	3
2	2:37,515,958	$1.34x10^{-9}$	37,429,848	37,561,357	47	45	3	1
3	2:127,892,810	$6.43x10^{-77}$	127,772,047	127,895,487	166	149	47	8
4	2:233,981,912	$3.64x10^{-12}$	233,978,212	234,082,577	78	72	6	2
5	4:11,027,619	$9.16x10^{-12}$	11,014,822	11,040,406	19	19	4	1
6	6:41,129,252	$6.95x10^{-14}$	40,706,366	41,162,518	14	14	7	2
7	6:47,443,806	$1.79x10^{-15}$	47,379,843	47,708,879	175	156	19	2
8	7:99,971,834	$1.94x10^{-20}$	99,591,021	100,190,116	160	143	16	3
9	7:143,103,481	$9.63x10^{-14}$	143,092,664	143,132,183	40	37	11	3
10	8:27,467,686	$3.77x10^{-29}$	27,187,520	27,489,959	96	90	28	7
11	8:145,097,720	$4.10x10^{-13}$	144,981,488	145,158,607	10	10	5	2
12	10:11,720,308	$3.09x10^{-11}$	11,707,563	11,723,537	12	12	3	1
13	11:47,380,340	$1.93x10^{-14}$	47,232,038	47,932,666	294	268	24	3
14	11:60,021,948	$8.69x10^{-24}$	59,826,677	60,105,199	460	432	38	3
15	11:85,868,640	$3.87x10^{-38}$	85,624,638	85,881,600	293	275	33	4
16	11:121,435,587	$5.59x10^{-16}$	121,423,640	121,460,949	29	26	8	3
17	14:53,391,680	$6.30x10^{-14}$	53,293,307	53,420,884	109	99	13	2
18	14:92,936,971	$2.41x10^{-13}$	92,926,952	92,938,855	50	50	8	1
19	15:59,045,774	$3.20x10^{-15}$	58,875,398	59,130,927	136	118	16	1
20	15:63,569,902	$7.98x10^{-11}$	63,569,902	63,595,878	3	3	1	1
21	16:70,694,000	$3.91x10^{-10}$	70,694,000	70,696,272	2	2	2	2
22	16:81,773,209	$6.87x10^{-12}$	81,773,003	81,912,580	11	11	4	2
23	17:5,138,304	$6.46x10^{-14}$	4,727,352	5,269,183	248	210	35	2
24	17:61,560,763	$5.51x10^{-10}$	61,499,732	61,574,492	45	45	13	3
25	19:1,043,638	$8.42x10^{-16}$	1,021,637	1,075,979	33	33	16	5
26	19:45,395,266	$2.06x10^{-287}$	44,403,375	46,125,148	1,238	1,107	434	118
27	19:51,728,477	$3.11x10^{-8}$	51,727,962	51,736,383	3	3	1	1
28	20:54,997,568	$2.16x10^{-11}$	54,980,606	55,040,537	26	25	15	1
29	21:27,473,875	$1.39x10^{-11}$	27,473,875	27,563,105	34	32	7	1

GRL also includes 188 lead SNPs that are independent from each other at $r^2 < 0.1$ (**Supplemental Table S3**). See Watanabe et al. (2017) for detailed methodology. A list of the r^2 scores for all pairs of one of the independent lead SNPs and SNPs in LD with $r^2 \geq 0.8$ can be found at **Supplemental Table S4**. For each of the 29 GRLs, only the top lead SNP per GRL is represented in **Table 1** (*P*-value from de Rojas et al. (2021)).

The majority of SNPs in the genomic risk loci are in non-coding regions

Previous attempts to functionally interrogate AD-associated SNPs has largely focused on SNPs that resided in protein-coding or UTRs of genes. In this study, we prioritise GRL based on the positional mapping of their SNPs, to focus on the ones that are not in coding or untranslated regions within a gene, such as exonic, UTRs and splicing sites. As a first step, we analyse the functional consequences of SNPs on genes with ANNOVAR, which annotates candidate SNPs that are in LD with at least one of the independent significant SNPs present in every GRL. We found an enrichment of candidate SNPs in intronic regions ($P = 1.51 \times 10^{-81}$), exonic regions ($P = 1.48 \times 10^{-10}$) and in untranslated regions (3' UTR, $P = 4.41 \times 10$-5; 5' UTR, $P = 3.07 \times 10^{-4}$) (**Figure 2A**, **Supplemental Tables S5, S6**). Furthermore, there is a depletion of intergenic SNPs ($P = 5.91 \times 10^{-124}$), however the total number of intergenic candidate SNPs is 1,079, which still accounts for 27.5% of the total number of candidate SNPs. Taking together, 79.2% of AD-associated SNPs are present in non-coding regions (intergenic = 1,079 and intronic = 2,028) and based

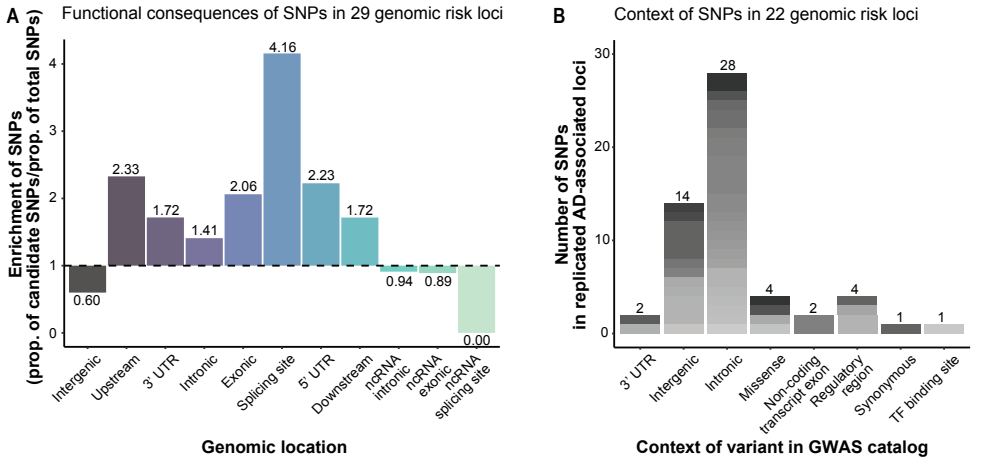


Figure 2: Highly replicated AD-associated genomic risk loci. A, Enrichment analysis of functional consequences of SNPs on genes in 29 GRL, performed by ANNOVAR in FUMA. Up- and downstream refer to SNPs within 1 kb of the transcription start site or transcription end site, respectively. **B,** Annotation of SNPs in context of GWAS catalog of Alzheimer's disease with replication studies. Only include 22 GRL, which are represented in a grey gradient.

on functional annotation of the respective loci, a clear predicted relation between the SNP and AD-associated gene-dysregulation is absent (**Supplemental Table S6**).

To focus on the highly replicated AD-associated loci without a known causative variant, we analysed a list of SNPs reported in the GWAS catalog provided by FUMA. From this list, 56 SNPs located in 22 GRL had previously been reported in an AD GWAS with a replication study (**Supplemental Table S7**). The genomic context of these variants was analysed and most of the variants are either intergenic (25%) or intronic (50%) (**Figure 2B**). Thus, we selected the GRL based on the presence of SNPs in highly replicated AD-associated loci composed of 22 GRL, ordered by the number of GWAS-identified SNPs (**Table 2**). Here we prioritise GRL with intergenic or intronic SNPs, which are 19 out of the 22 (the GRL 23, 9 and 27 are without any non-coding SNPs). Furthermore, GRL 26 of the *APOE* gene has the lowest priority due to its high penetrance pathogenic ε4 allele (Saddiki *et al.*, 2020).

Table 2: 22 highly replicated genomic risk loci associated with AD. Table of a selection of the 22 GRL containing SNPs present in the GWAS catalog of Alzheimer's disease with replication studies. *P*-values of top lead SNP from de Rojas et al. (2021). Coordinates for GRL are based on the human genome reference GRCh37. GRL highlighted with a * indicate that there is no non-coding SNP (intergenic or intronic) from ANNOVAR analysis. SNPs: number of SNPs within the GRL; GWAS SNPs: number of GWAS-tagged candidate SNPs within the GRL; Independent Sig SNPs: number of independent significant SNPs within the GRL; Lead SNPs: number of lead SNPs within the GRL.

GRL	rsID	p-value	Start	End	SNPs	GWAS SNPs	Independet Sig SNPs	Lead SNPs	Genes reported in GWAS
14	11:60,021,948	$8.691x10^{-24}$	59,826,677	60,105,199	460	432	38	3	*MS4A6A, MS4A4E, MS4, MS4A4A, MS4A6E, MS4A*
15	11:85,868,640	$3.873x10^{-38}$	85,624,638	85,881,600	293	275	33	4	*PICALM*
13	11:47,380,340	$1.934x10^{-14}$	47,232,038	47,932,666	294	268	24	3	*CELF1*
7	6:47,443,806	$1.794x10^{-15}$	47,379,843	47,708,879	175	156	19	2	*CD2AP*
3	2:127,892,810	$6.431x10^{-77}$	127,772,047	127,895,487	166	149	47	8	*BIN1*
8	7:99,971,834	$1.944x10^{-20}$	99,591,021	100,190,116	160	143	16	3	*ZCWPW1*
19	15:59,045,774	$3.197x10^{-15}$	58,875,398	59,130,927	136	118	16	1	*ADAM10*
17	14:53,391,680	$6.302x10^{-14}$	53,293,307	53,420,884	109	99	13	2	*FERMT2*
1	1:207,802,552	$7.297x10^{-31}$	207,368,589	207,875,537	94	90	12	3	*CR1*
10	8:27,467,686	$3.772x10^{-29}$	27,187,520	27,489,959	96	90	28	7	*PTK2B, CLU*
4	2:23,3981,912	$3.644x10^{-12}$	233,978,212	234,082,577	78	72	6	2	*INPP5D*
18	14:92,936,971	$2.414x10^{-13}$	92,926,952	92,938,855	50	50	8	1	*SLC24A4, RIN3*
24	17:61,560,763	$5.513x10^{-10}$	61,499,732	61,574,492	45	45	13	3	*ACE*
25	19:1,043,638	$8.417x10^{-16}$	1,021,637	1,075,979	33	33	16	5	*ABCA7*
16	11:121,435,587	$5.591x10^{-16}$	121,423,640	121,460,949	29	26	8	3	*SORL1*
28	20:54,997,568	$2.160x10^{-11}$	54,980,606	55,040,537	26	25	15	1	*CASS4*
6	6:41,129,252	$6.947x10^{-14}$	40,706,366	41,162,518	14	14	7	2	*OARD1, TREM2, TREML2*
12	10:11,720,308	$3.091x10^{-11}$	11,707,563	11,723,537	12	12	3	1	*NR, ECHDC3, USP6NL*
23*	17:5,138,304	$6.459x10^{-14}$	4,727,352	5,269,183	248	210	35	2	*SCIMP*
9*	7:143,103,481	$9.632x10^{-14}$	143,092,664	143,132,183	40	37	11	3	*EPHA1, TAS2R62P*
27*	19:51,728,477	$3.109x10^{-08}$	51,727,962	51,736,383	3	3	1	1	*CD33*
26	19:45,395,266	$2.060x10^{-287}$	44,403,375	46,125,148	1,238	1,107	434	118	*APOE, PVRL2, TOMM40*

High structural variation in AD-associated loci

We hypothesise that previously overlooked structural variation might be playing a role in the highly replicated loci associated with Alzheimer's disease, of which there is no causal genetic variant identified yet. To that end, we analysed SVs present in the 22 GRL, using the SV data reported by Audano et al. (2019). 16 of the 22 GRL contain a total of 202 SVs, with lengths ranging from 50 bp to 4,546 bp for insertions (median = 140 bp), and -6,074 bp to -50 bp for deletions (median = -104 bp) (**Figure 3A**, **Supplemental Table S8**). Most of the structural variants in the GRL are in tandem repeats (TRF), which make up 52.5% of the total variation (**Figure 3B**). Importantly, the Audano list included the SV in the *ABCA7* locus, expansion of which was previously associated with AD (De Roeck *et al.*, 2018) (**Figure 3C**).

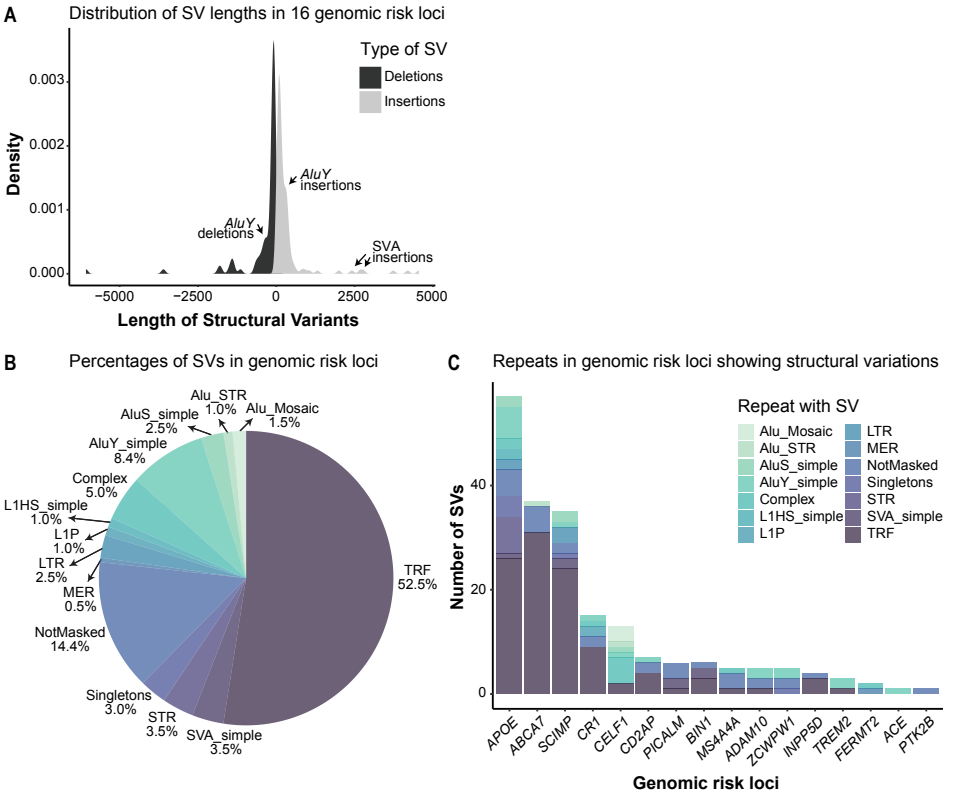


Figure 3: Structural variants in AD-associated genomic risk loci. A, Distribution of lengths of structural variants from Audano et al. (2019) present in 16 GRL. *Alu*Y and SVA elements are highlighted in the density plot as insertional polymorphisms. **B,** Pie chart showing total SV percentages in the 16 GRL. **C,** Number of SVs per GRL. **B-C,** The colouring represents the type of repeat which shows structural variation or full insertion/deletion for *Alu*Y and SVA elements. Nomenclature of repeats used as in Audano et al. (2019).

Of all SVs in the GRL, the *Alu* element family constitutes 13.4% (**Figure 3B)**. The *Alu*Y elements represent 63% of all *Alu* elements and these are present in 9 of the 16 GRL (**Figure 3B-C**). The insertional polymorphisms of *AluY* elements account for the shoulder of the peak around -300 bp (*Alu* insertions absent in the genome assembly) as well as around +300 bp (*Alu* insertions not represented in the genome assembly). (**Figure 3A**). Moreover, there are larger structural variations defined as "Complex" or "NotMasked", which include *Alu* elements, raising the percentage to 18.32% SVs (**Supplemental Table S8**). SVA elements exhibit both new polymorphic SVA insertions (not present in the whole human population) and structural variation in SVA elements that are fixed in the human genome (present in every individual). In total structural variation in SVA elements accounts for 3.5% of the SVs and are present in four GRL: *BIN1*, *PICALM*, *SCIMP* and *APOE* (**Figure 3B-C, Supplemental Table S8**). Three polymorphic SVA insertions contribute to SVs of sizes around ~2,500 bp and ~2,700 bp (**Figure 3A**) and are located in the *SCIMP* and *APOE* loci.

To further investigate the architecture of these loci regarding the structural variations, we intersected the SVs with repetitive elements from the RepeatMasker track on UCSC Genome Browser. Only SVs that fully reside within a repetitive element were considered. SVA elements were represented in more AD-associated GRL than previously observed (**Supplemental Table S9**). This is due to the classification of SVs done by Audano et al. (2019), which called "TRF" (tandem repeats), "singletons" or "NotMasked" to variations occurring within the SVA elements, instead of "SVA_simple" (**Table 3**). A number of SVs were found within SVA sequences that were fixed in the human population, meaning that every human individual carries a particular SVA insertion in its genome, but it has an SV within its VNTR sequence. These SVs

Table 3: Genomic risk loci with structurally variable SVAs. Table of GRL with SV-SVAs, which are highlighted in lilac. Repetitive DNA from RepeatMasker and SVs from Audano et al. (2019).

GRL	Genes	Repeat Name	Repeat Family	Structural Variants (Audano et al. (2019)
3	*BIN1*	(CCT)n	Simple repeat	117 INS TRF
3	*BIN1*	SVA-F	SVA	52 INS NotMasked, 96 INS TRF, 80 DEL SVA_simple, 169 INS SVA_simple, 293 INS TRF
7	*CD2AP*	(TCC)n	Simple repeat	162 INS TRF
7	*CD2AP*	AluYh3	Alu	56 INS NotMasked
7	*CD2AP*	L2	L2	344 INS AluY_simple
7	*CD2AP*	SVA-C	SVA	807 INS TRF, 298 INS TRF
8	*ZCWPW1*	AluSx1	Alu	322 INS AluY_simple
8	*ZCWPW1*	FLAM-A	Alu	343 INS AluY_simple
8	*ZCWPW1*	SVA-D	SVA	50 DEL Singletons
14	*MS4A4A*	(AT)n	Simple repeat	212 INS TRF
14	*MS4A4A*	Charlie1b	hAT−Charlie	70 INS NotMasked
14	*MS4A4A*	Charlie1b	hAT−Charlie	70 INS NotMasked
14	*MS4A4A*	L1MCa	L1	1056 INS NotMasked
14	*MS4A4A*	SVA-D	SVA	50 DEL NotMasked
15	*PICALM*	(AT)n	Simple repeat	96 INS TRF
15	*PICALM*	MIR3	MIR	64 INS NotMasked
15	*PICALM*	SVA-D	SVA	356 INS SVA_simple, 80 INS NotMasked, 63 INS NotMasked, 175 DEL SVA_simple

were found within SVA sequences in the GRL of *BIN1*, *CD2AP*, *ZCWPW1, MS4A4A* and *PICALM* with SV lengths ranging from 50 bp to 807 bp (**Table 3**). One other SV of 4,195 bp lies within a fixed SVA element in the *CELF1* gene. An analysis shows that this is likely caused by a retrotransposon insertion within the fixed SVA element and not an extension of the VNTR as is the case for other SV-SVAs.

Genomic regions of *BIN1* and *PICALM* genes are neuronally active

Next, we focussed on the SVAs in the highly replicated AD-associated GRL of *BIN1*, *CD2AP*, *ZCWPW1, MS4A4A* and *PICALM*, which have fixed but structurally variable SVAs (SV-SVA), i.e. they are present across the human population in different forms. Because issues with mapping of DNA sequences to highly repetitive DNA elements result in the underrepresentation of SNPs in repetitive DNA such as SVAs, we used information about nearby non-coding SNPs as a proxy to the predicted functionality of the genomic locus the SVA resides in. We first established whether any of the nearby non-coding SNPs were predicted to be deleterious (i.e. CADD score higher than 12.37,(Kircher et al. 2014). In the *PICALM* locus, there are three SNPs with high CADD scores of 16.12, 15.63 and 17.79. Also, in the *BIN1*, *CD2AP* and *MS4A4A* loci there are SNPs just around or above the threshold (**Figure 4A**). None of the SNPs in *ZCWPW1* locus have a deleteriousness score close to the threshold, the highest being SNP 7:99843353:C:T with a CADD score of 8.62 (**Supplemental Table S10**), thus the SVA in the *ZCWPW1* locus was not further investigated at this stage.

Since functional annotation relies on chromatin states which is often not accurately captured for highly repetitive DNA elements such as SVAs, we investigated the functional properties of nearby independent significant SNPs as an indirect measure for the possible gene-regulatory potential that is harboured in the four SVA-containing GRL. For each GRL, we then analysed data from the 15-core chromatin states from Roadmap Epigenomics (Kundaje et al. 2015), enhancers from PsychENCODE (Akbarian et al. 2015) and expression quantitative trait loci (eQTL) from GTEx (Lonsdale et al. 2013). Based on Roadmaps epigenetic data, SNPs in the GRL of *BIN1*, *CD2AP* and *PICALM* are present in neuronally active genomic regions, whilst SNPs in *MS4A4A* locus show almost full inactivation in brain tissues (**Figure 4B**). This is also supported by the number of SNPs overlapping enhancers in these loci, especially for *BIN1* and *PICALM* loci (**Figure 4C**). Finally, eQTL data shows that *BIN1* and *CD2AP* gene expression correlates with the presence of the independent SNPs (**Figure 4D**, **Supplemental Table S11**). On the contrary, such association could not be confirmed for *MS4A4A* and *PICALM* genes.

Based on the four datasets together, the GRL of *PICALM* is the most promising locus due to the high activity in neuronal environments and the three independent SNPs that are predicted to be deleterious with a high CADD score (**Figure 4**). The SNP 11:85800279:A:G with the highest deleteriousness score (CADD=17.79) is an intergenic SNP previously reported in GWAS of AD with replication studies. It is

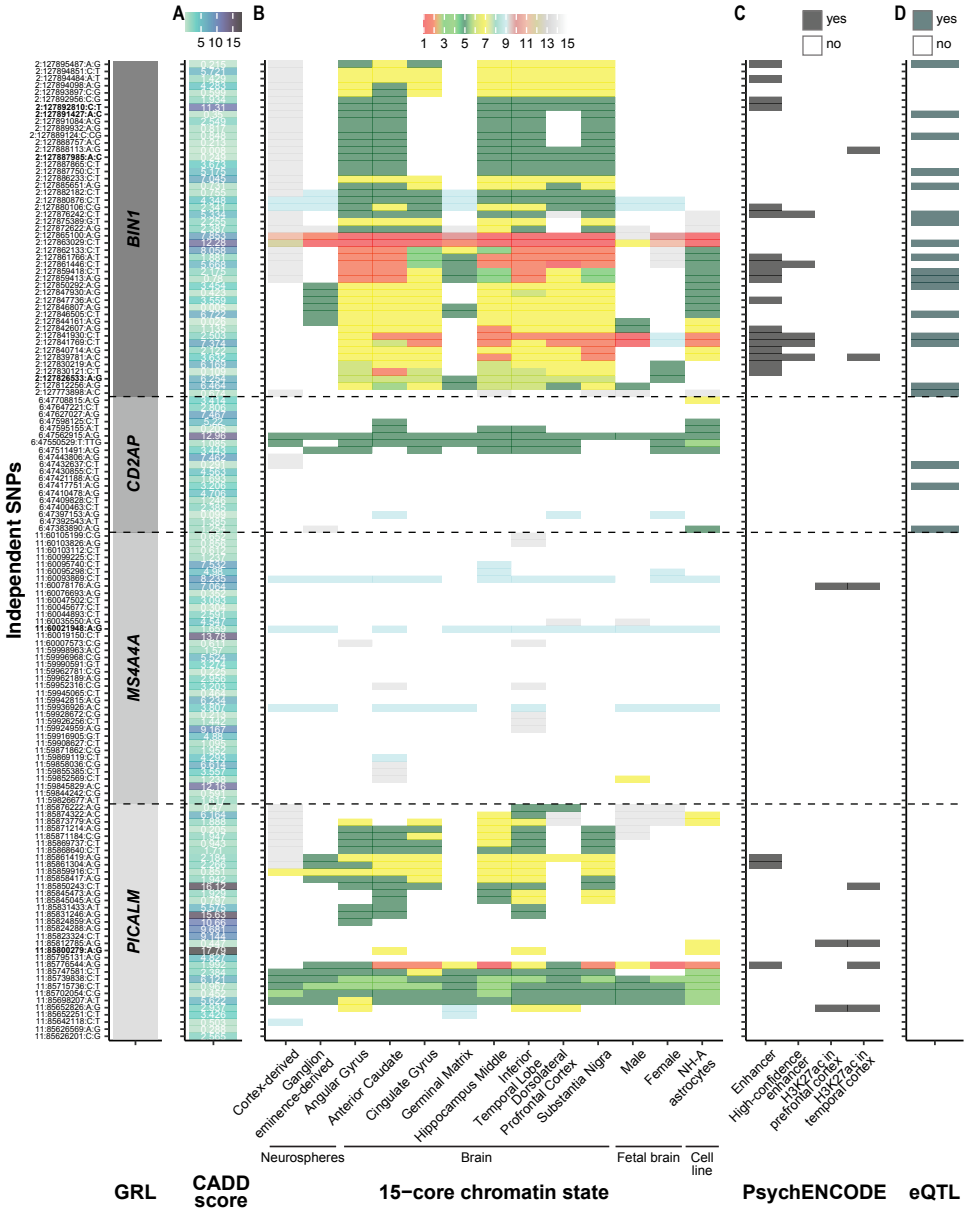


Figure 4: Deleteriousness and functional properties of independent SNPs. A-D, For each independent significant SNP in the 4 GRL we plot: **A,** CADD score for level of deleteriousness: a SNP is considered deleterious with a score higher than 12.37 (Kircher et al., 2014), SNPs in bold are characterised as "intergenic" or "intronic" from ANNOVAR analysis and are present in GWAS catalog for AD with replication study. SNP IDs are based on the human genome reference GRCh37. **B,** 15-core chromatin state, warmer colours from 1 to 7 represent active/ open chromatin whilst colder colours from 8 to 15 represent heterochromatin (see **Methods** for specific chromatin states from Roadmap Epigenomics). **C,** Enhancer dataset from PsychENCODE for brain tissues. **D,** eQTL data for brain tissues from GTEx.

present in an open chromatin region but not overlapping any known enhancers and without eQTL signal. This SNP is located approximately 20 kb downstream of *PICALM* (chr11:85,668,214-85,780,139; GRCh37), and 20 kb upstream of the SV-SVA (chr11:85,820,785-85,823,186; GRCh37). Interestingly, the other independent SNPs with high CADD scores 11:85850243:C:T (CADD=16.12) and 11:85831246:A:G (CADD=15.63) are located downstream of the SV-SVA, thus surrounding this element. Because a disease-associated SNP is usually merely an indicator of a causal variant in its vicinity, the fact that two SNPs with high CADD scores flank the SV-SVA in the upstream region of *PICALM*, may suggest that the SV-SVA contributes to the disease associated dysregulation thereof.

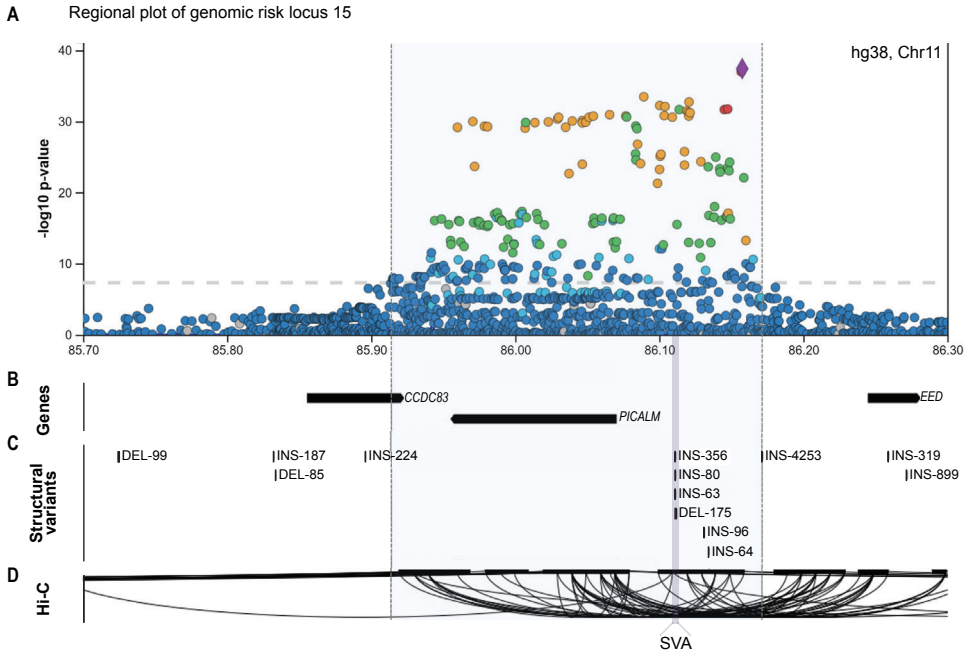


Figure 5: Overview of *PICALM* genomic risk locus. A, Regional plot of AD-associated SNPs from Rojas et al. (2021) plotted with their respective meta-analysis genome-wide significant *P*-values (GWS (Genome-wide significance), $P < 5 \times 10^{-8}$; as $-\log_{10}$ values) and are distinguished by linkage disequilibrium (r^2) of nearby SNPs on a blue to red scale, from $r^2 = 0$ to 1, based on pairwise r^2 values from the 1000 Genomes Phase3 reference panel. **B,** Gene annotations from NCBI RefSeq Select database, **C.** Structural variants from Audano et al. (2019). **D,** Local chromatin interactions from datasets available in FUMA (see **Methods**). SVA highlighted in lilac. Grey shading represents GRL 15. Assembly GRCh38, scale in Mb.

SV-SVAs in four GRL are in a AD-associated SNP-dense region but lack direct chromatin interaction with genes

To investigate whether the SV-SVAs, together with independent SNPs, in the four GRL are involved in regulatory mechanisms through 3D chromatin interactions, we analysed chromatin interaction mapping derived from FUMA (see **Methods**, **Supplemental Table S12**). This dataset consists of regions of 10,000 bp which overlap with at least one SNP in the GRL, in brain tissues. For a visual understanding of each locus, we plotted all 3D chromatin associations, structural variants, and a regional plot of the study from de Rojas et al. (2021). We observe that the *PICALM*-SVA, *BIN1*-SVA and *MS4A4A*-SVA, are interacting with regions either in- or outside of the locus (**Figures 5-7**). Interestingly, in the *PICALM* locus, the region with highest association to AD colocalizes with the SV-SVA and interacts with genes of *PICALM*, *EED* and *CCDC83* (**Figure 5**). Likewise, in the *BIN1* locus, the highly variable SV-SVA is located upstream of the *BIN1* gene on the right side of the GRL, which is

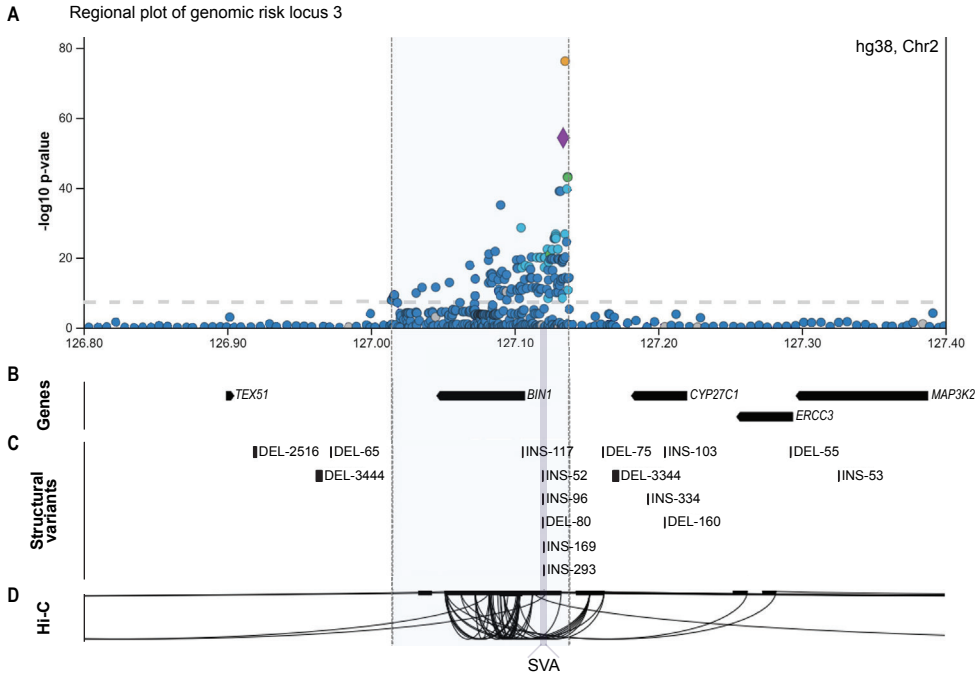


Figure 6: Overview of *BIN1* genomic risk locus. A, Regional plot of AD-associated SNPs from Rojas et al. (2021) plotted with their respective meta-analysis genome-wide significant *P*-values (GWS (Genome-wide significance), $P < 5 \times 10^{-8}$; as $-\log_{10}$ values) and are distinguished by linkage disequilibrium (r^2) of nearby SNPs on a blue to red scale, from $r^2 = 0$ to 1, based on pairwise r^2 values from the 1000 Genomes Phase3 reference panel. **B,** Gene annotations from NCBI RefSeq Select database. **C,** Structural variants from Audano et al. (2019). **D,** Local chromatin interactions from datasets available in FUMA (see **Methods**). SVA highlighted in lilac. Grey shading represents GRL 3. Assembly GRCh38, scale in Mb.

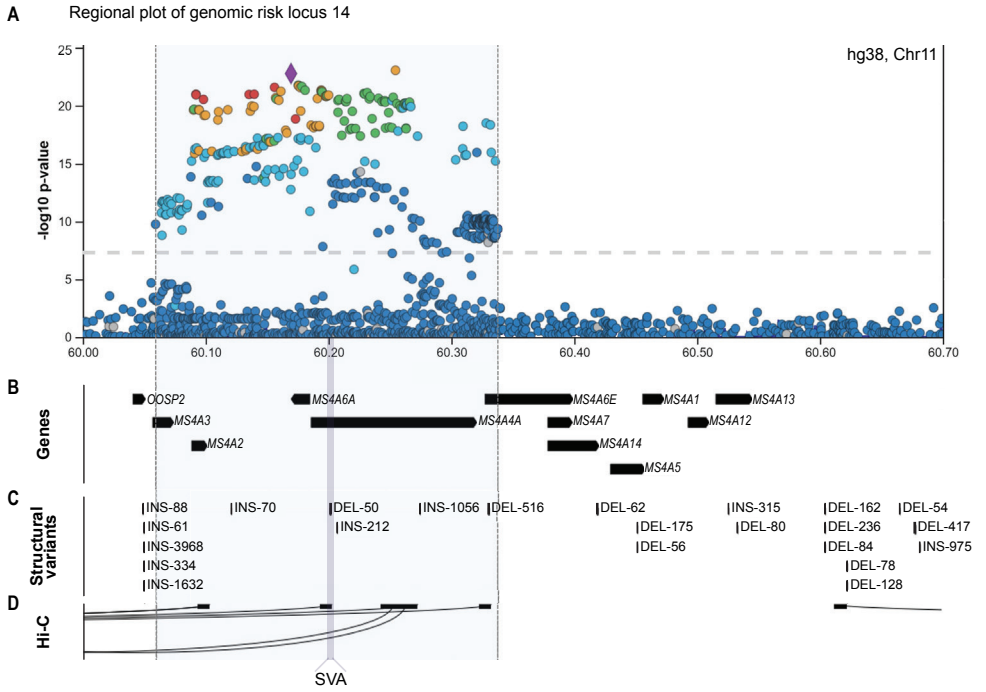


Figure 7: Overview of *MS4A4A* genomic risk locus. A, Regional plot of AD-associated SNPs from Rojas et al. (2021) plotted with their respective meta-analysis genome-wide significant *P*-values (GWS (Genome-wide significance), $P < 5 \times 10^{-8}$; as $-\log_{10}$ values) and are distinguished by linkage disequilibrium (r^2) of nearby SNPs on a blue to red scale, from $r^2 = 0$ to 1, based on pairwise r^2 values from the 1000 Genomes Phase3 reference panel. **B,** Gene annotations from NCBI RefSeq Select database. **C,** Structural variants from Audano et al. (2019). **D,** Local chromatin interactions from datasets available in FUMA (see **Methods**). SVA highlighted in lilac. Grey shading represents GRL 14. Assembly GRCh38, scale in Mb.

the region with highest association with AD (**Figure 6**). In the *MS4A4A* locus, the SV-SVA is located in a region with high association with AD, exhibiting only one SV of 50 bp deletion which is also linked to a region outside of the locus (**Figure 7**). Finally, in the *CD2AP* locus the SV-SVA is also intronic and has two insertions of 298 and 807 bp (**Figures 8**). It is located right in the middle of the region of high AD-association though it does not interact with any other region either in- or outside of the locus. Importantly, all SV-SVAs are surrounded by risk SNPs, indicating a possible mechanism of regulation involved in disease.

For a more detailed resolution of the locus, we intersected each region from the chromatin interaction dataset with structurally variable repetitive DNA elements (**Supplemental Table S13**). We did not find evidence for a direct physical interaction with the SVAs and promoters of nearby genes in the GRL. A possible explanation for

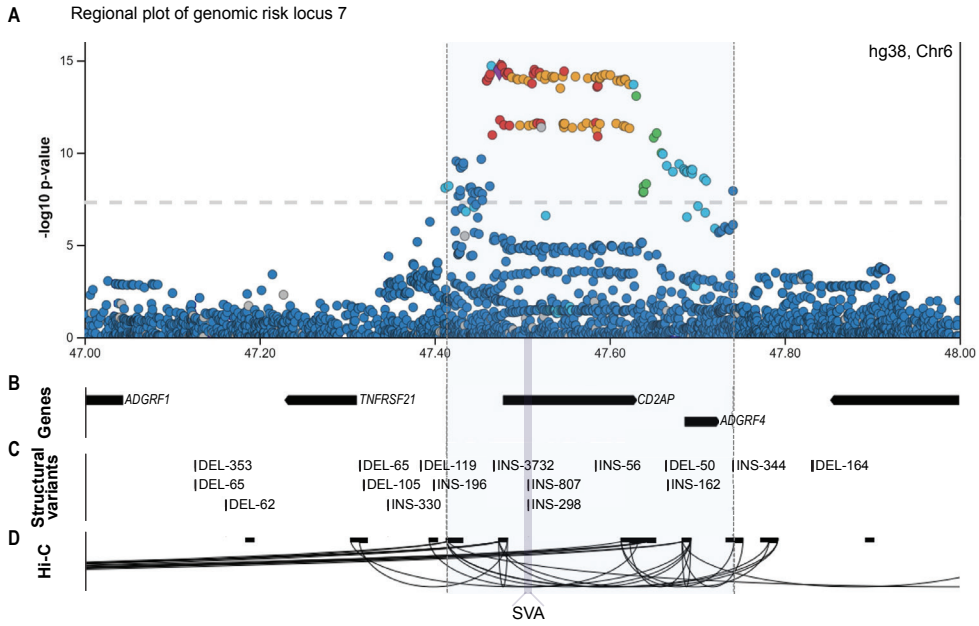


Figure 8: Overview of *CD2AP* genomic risk locus. A, Regional plot of AD-associated SNPs from Rojas et al. (2021) plotted with their respective meta-analysis genome-wide significant *P*-values (GWS (Genome-wide significance), $P < 5 \times 10^{-8}$; as $-\log_{10}$ values) and are distinguished by linkage disequilibrium (r^2) of nearby SNPs on a blue to red scale, from $r^2 = 0$ to 1, based on pairwise r^2 values from the 1000 Genomes Phase3 reference panel. **B,** Gene annotations from NCBI RefSeq Select database. **C,** Structural variants from Audano et al. (2019). **D,** Local chromatin interactions from datasets available in FUMA (see **Methods**). SVA highlighted in lilac. Grey shading represents GRL 7. Assembly GRCh38, scale in Mb.

this is that mapping of interaction sites to highly repetitive DNA is often extremely challenging, potentially leading to the failure to map interaction sites that exist *in vivo*.

Under normal conditions SVAs are heavily covered with repressive histone and DNA methylation marks, which may prevent them from interacting with their genetic environment and exerting their gene regulatory effect. Epigenetic control of SVAs may also cause the temporary lack of direct physical association between the SVA and the nearby genes, which other work of our lab suggests becomes progressively impaired during ageing (Rosenkrantz et al. 2024). Therefore, a lack of a chromatin interaction in the current HiC data does not rule out that this interaction exists in specific cell-types, in pathological conditions or during normal ageing of the human brain.

Discussion

Over the past decades, GWAS have been highly successful in identifying loci and genes associated with Alzheimer's disease. We hypothesise that for some of the AD-associated loci, the pathological dysregulation might be caused by larger structural variations that have remained hidden thus far. Their omission in past GWAS datasets is due to technical limitations in sequence mapping and an overall shortcoming of accurate functional annotation of repetitive DNA. To investigate the involvement of hidden SVs in gene-dysregulation in AD associated loci, this study aimed to identify SVs in AD-associated loci, and we characterise the AD-associated loci based on the integration of structural variants in a GWAS fine-mapping analysis.

Here we identify 16 GRL associated with AD that contain a total of 202 SVs. Overall, structural variations in tandem repeats are present in almost every locus. Tandem repeats encompass VNTRs and STRs (short tandem repeats) and interestingly, unstable tandem repeats are responsible for many human neurological diseases caused by pathological repeat expansions, including in AD (Chintalaphani et al. 2021; De Roeck et al. 2018; Fondon et al. 2008; Pearson, Edamura, and Cleary 2005). Furthermore, a large contribution comes from *Alu* elements, a primate-specific retrotransposon family and they have extensively been linked to various human neurological diseases, including AD (Larsen et al. 2017, 2018; Salazar et al. 2023). There are multiple mechanisms by which *Alu* elements can contribute to the progression of disease, at the genomic or transcriptional levels, including post-transcriptional modifications (Pfaff, Singleton, and Kõks 2022). Thus, exploring the variability in these inherently dynamic regions of our genome might give links to explain the missing heritability of Alzheimer's disease.

We report that amongst the highly replicated AD-associated loci are a few loci which harbour SV-SVA elements surrounded by GWAS-identified risk SNPs with high disease-association values. Four of these loci harbour four different genes, *PICALM*, *BIN1*, *CD2AP* and *MS4A4A*, that play crucial roles in brain lipid metabolism, a process intricately linked to neuroinflammation and microglial functions (Ando et al. 2022; Yang et al. 2022). Microglial cells serve as the brain's immune system, actively maintaining its homeostasis by phagocytosing harmful molecules (Colonna and Butovsky 2017). In the context of Alzheimer's disease, microglia play a pivotal role in the early stages by clearing amyloid-beta (Aβ) aggregates. This process involves endosomal regulator proteins such as *PICALM*, *BIN1*, and *CD2AP* (Ando et al. 2022; Dourlen et al. 2019; Raj et al. 2018). Additionally, the *MS4A4A* gene, belonging to the *MS4A* gene cluster, exhibits high expression in microglia and serves as a modulator of *TREM2* expression, another key player in AD (Deming et al. 2019). *PICALM* and *BIN1*, along with *APOE*, are established as the main susceptibility genes for AD. *CD2AP* and *MS4A4A* have also been associated with AD pathology through multiple independent genetic studies (Bellenguez et al. 2022;

Kunkle et al. 2019; de Rojas et al. 2021; Schwartzentruber et al. 2021). It is believed that *PICALM, BIN1* and *CD2AP* genes affect Aβ clearance through changes on their expression levels (Ando et al. 2022; Dourlen et al. 2019; Gao et al. 2021; E. Lambert et al. 2022; Tao, Chen, and Wu 2019; Xu, Tan, and Yu 2015). Conversely, disruptions in the regulation of the *MS4A4A* gene could impact microglial cells' lipid metabolism and immune response (You et al. 2023). However, the specific genetic mechanism by which dysregulation of *PICALM*, *BIN1, CD2AP* or *MS4A4A* increases the risk of AD remains elusive.

Strikingly, SV-SVAs in these four loci consistently reside in regions strongly correlated with AD, often in close proximity to the risk SNP with highest disease significance. Notably, SV-SVAs in *PICALM*, *BIN1* and *MS4A4A* loci reside in regions of 3D chromatin interactions, suggesting potential functional regulatory links both within and outside of their respective locus. Additionally, the genomic risk loci of *PICALM* and *BIN1* exhibit significant neuronal activity, as indicated by their open chromatin status, presence of enhancers and eQTL signalling in brain tissues. Conversely, *MS4A4A* locus presents silencing in a neuronal context. However, since this gene cluster is predominantly expressed in microglia rather than neurons, advocating further investigation of this *MS4A4A* locus in microglial cells.

Insertional polymorphisms of SVAs are common in the human population, as SVAs are one of the few classes of TEs that remain actively retrotransposing in the human genome (Hancks and Kazazian 2010; Payer and Burns 2019). An insertional SVA polymorphism is present in some but not all human individuals. We find polymorphic SVA insertions in the *APOE* locus, the main susceptibility gene for AD pathology, and *SCIMP*, another gene associated with AD (Bellenguez et al. 2022; Saddiki et al. 2020). Further studies should be conducted to assess the functional consequences of insertional polymorphisms in these AD GRL. Here we did not focus our work on insertional SVA polymorphisms, because their absence in the reference genome precludes the accurate consideration of chromatin states, eQTL and chromatin interactions. Therefore, despite their possible contribution to human disease, our approach was to focus only on fixed SVA insertions, consistently present in the human population, which can show structural variations within the VNTR sequence in the middle part of the SVA.

The high level of structural variation found in SVA elements widens the range of hidden gene-regulatory potential, introducing previously unknown layers of gene regulation that may heavily depend on the changeable epigenetic status of our DNA upon normal ageing (van Bree et al. 2022; Guo et al. 2018; Wanhe Li et al. 2013; Yushkova and Moskalev 2023). One of the most noteworthy changes during ageing is the heterochromatin loss (Villeponteau 1997), resulting in de-repression of elements like SVAs. Concomitantly, the expression level of ZNF91, the main repressor of SVAs (Haring et al. 2021) shows a significant negative correlation to

age in many neuronal tissues (Rosenkrantz et al. 2024). This reduction in ZNF91 expression over time may also lead to the progressive unleashing of the SVA's full gene-regulatory potential, suggesting that their influence on gene regulation is both temporally dynamic and intricately linked to the ageing process. Since Alzheimer's disease is intrinsically related with ageing and transposable elements are active and expressed in the brain (Ahmadi et al. 2020; Baillie et al. 2011; Thomas et al. 2020), SVAs could contribute to the disruption of neuronal function and integrity, influencing disease onset and progression.

In this study, we have gathered the first hints towards a potential functional involvement of hidden SVs and, in particular, SV-SVA elements in loci associated with Alzheimer's disease. In a post-GWAS era, the importance of fine-mapping analysis cannot be overstated when it comes to disentangling the intricate genetic identity of each AD locus previously identified by GWAS. Currently the primary focus of these analyses is on SNPs that are in linkage disequilibrium within a certain genomic region. However, a significant proportion of these SNPs are located in non-coding regions. It is worth noting that the larger variability within these non-coding regions, arising from repetitive DNA elements, are often overlooked in association studies. Therefore, functional characterization of hidden AD genetic factors, such as SV-SVA elements, is imperative to elucidate the mechanisms of AD and may shed light on the observed differential genetic susceptibility to AD in the human population

Methods

Fine-mapping of AD-associated loci using FUMA

Characterization of genomic risk loci

The data analysed in this study consisted of GWAS summary statistics acquired from a meta-GWAS analysis of AD studies performed by de Rojas et al. (2021). The meta-analysis consisted of a Spanish case-control study (GR@ACE/DEGESCO, n = 12,386), a case-control study of International Genomics of Alzheimer project (IGAP, n = 82,771) and the UK Biobank AD-by-proxy case-control study (n = 314,278). We used FUMA (Functional Mapping and Annotation of Genome-Wide Association Studies, version 1.3.6a), which provides comprehensive annotations identifying independent significant SNPs that have a GWS *P*-value < 5 x 10^{-8} and LD of $r^2 \geq 0.8$. A subset of these SNPs are termed lead SNPs when they are in approximate LD with each other at $r^2 < 0.1$. FUMA then assigns independent genomic risk loci by identifying regions in LD with these lead SNPs, within a 500 kb window, using the 1000 Genomes Phase 3 reference panel to compute the LD blocks. The borders of the genomic risk loci are defined by identifying all SNPs in linkage disequilibrium ($r^2 \geq 0.8$) with one of the independent significant SNPs in the locus. The region containing all these candidate SNPs was considered as a single independent genomic risk locus. All genomic annotations are based on the human genome reference GRCh37. Output files regarding SNPs are in **Supplemental Tables S1-S4**, including information on all, independent and lead SNPs, and r^2 scores for all pairs of independent SNPs and SNPs in LD ($r^2 \geq 0.8$).

Annotation of SNPs from FUMA

FUMA annotates functions to independent significant SNPs and of those that are in LD ($r^2 \geq 0.8$) using multiple methods and databases. For functional consequences of SNPs on genes, FUMA uses ANNOVAR with Ensembl genes (build 85) (K. Wang, Li, and Hakonarson 2010) (**Supplemental Tables S5, S6**); their deleteriousness with CADD score (Kircher et al. 2014); regulatory functions with RegulomeDB score (Boyle et al. 2012), PsychENCODE (Akbarian et al. 2015) and 15-core chromatin state in brain regions from Roadmap Epigenomics (Kundaje et al. 2015) predicted by ChromHMM (Ernst and Kellis 2012) (**Supplemental Table S10**); effects on gene expression with eQTL using GTEx v8 (Lonsdale et al. 2013) for brain tissues (**Supplemental Table S11**); chromatin interactions using Hi-C data from PsychENCODE, (Giusti-Rodríguez et al. 2019) and (Schmitt et al. 2016) available in FUMA, using enhancers from PsychENCODE and Roadmap, and promoter regions defined as 1,000 bp up- and 500 bp downstream of transcription start sites (**Supplemental Table S12**). Finally, SNPs are linked to the GWAS catalog (version e96_r2019-09-24) (Welter et al. 2014) to include previously reported associations with traits or diseases (**Supplemental Table S7**).

Filtering of genomic risk loci based on FUMA

We used the outputs from FUMA described above to further filter the SNPs in the GRL. All analysis was done in R (version 4.3.0) with the packages data.table (version 1.14.8), stringr (version 1.5.0), and plots were made with the packages ggplot2 (version 3.4.2), ggpubr (version 0.6.0) and viridis (0.6.3). We used the information in the GWAS catalog to select independent significant SNPs that have been previously reported in Alzheimer's disease association studies (from the "Trait" column) and were replicated (**Supplemental Table S7**). Therefore, from the initial set of 29 GRL from FUMA we selected the ones that contain non-coding independent SNPs previously associated with AD.

Structural variants in genomic risk loci

A list of structural variants was acquired from Audano et al. (2019), generated using long-read sequencing data. In total 99,604 insertions, deletions, and inversions were mapped. To filter the SVs list based on their presence in the GRL, we first used the LiftOver tool on UCSC Genome browser (Kent et al. 2002) to convert the coordinates from the reference genome GRCh37 to GRCh38 (**Table 1**). We selected the SVs present in the genomic risk loci using bedtools intersect (v2.30.0) (Quinlan and Hall 2010) with the options -wa -wb. The SVs list was imported into R, further filtering and analysis was done in the SVs present in the 16 GRL (**Supplemental Table S8**). To highlight in the density plot (**Figure 2A**) only full insertions or deletions from the active classes of mobile elements *AluY* and SVA, we used a cut-off size of 250 bp and 2000 bp, respectively. The percentage of *Alu* elements showing a variation was done by first grepping all "Alu" from the "REPEAT_TYPE" and the "FULL_REPEAT" columns, then binding both tables with unique values. The "FULL_REPEAT" column gives information about the "Complex" or "NotMasked" SVs as well, which are often large insertions/deletions consisting of more than one type of repeat.

To investigate further the type of repetitive DNA that shows structural variation in the 16 GRL, we used the table browser at the UCSC Genome Browser. We intersected the repetitive DNA from RepeatMasker track (A. Smit, Hubley, and Green 1996) within the defined regions (GRCh38) with the SVs list from Audano et al. (2019), previously uploaded onto the browser. Then in R we merged the list of repeats with the SVs, by keeping only the SVs that are fully contained within the repeat (**Supplemental Table S9**). More complex variations partially involving repeats were excluded from the analysis.

Deleteriousness and functional properties of independent SNPs in genomic risk loci

The four GRL we decided to focus on were studied based on the independent significant SNPs present in each locus. We used eQTL data from GTEx (v8) for brain tissues together with the annotation file with CADD score, enhancer data from

PsychENCODE and the 15-core chromatin state in brain regions from Roadmap Epigenomics (**Supplemental Tables S10, S11**). These datasets were used as input for heatmaps done with ggplot2 package in R.

Chromatin interactions in genomic risk loci

3D chromatin interactions were analysed in the four GRL, using Hi-C datasets from PsychENCODE, Giusti-Rodríguez et al. (2019) and Schmitt et al. (2016) available in FUMA (**Supplemental Table S12**). The coordinates from column "region1", which is the end of the chromatin interaction that overlaps with at least one SNP, were converted from the reference genome GRCh37 to GRCh38 with the LiftOver tool on UCSC Genome browser. Then we selected the chromatin interactions that also overlap with a structurally variable repetitive DNA element (**Supplemental Table S9**) using bedtools intersect (v2.30.0) (Quinlan and Hall 2010) with the options -wa -wb (**Supplemental Table S13**).

For visualization purposes, the full chromatin interaction dataset, together with the SVs list, was plotted using pyGenomeTracks (https://pypi.org/project/pyGenomeTracks/) and regional plots were made using LocusZoom (Pruim et al., 2010) (https://my.locuszoom.org/) using data from Rojas et al. (2021). Gene annotations are from NCBI RefSeq Select database, assembly GRCh38.

Code

Full scripts can be found in **Supplemental Code**.

Acknowledgements

This work was supported by a grant from Alzheimer Nederland (WE.03-2018-07) to F.M.J. Jacobs, H. Holstege and M.J.T. Reinders. We thank Niccolò Tesi for his help getting the summary statistics used in this study, and the Evolutionary Neurogenomics for helpful discussions.

Author contributions

F.M.J.J., R.L.F.P.G. and E.R.B. conceptualized the project. E.R.B. performed initial FUMA analysis and created the regional plots. R.L.F.P.G. integrated the premilinary results with structural variants mapping and further deleteriousness and functional properties of SNPs. F.M.J.J., M.J.T.R. and H.H. acquired funding for the study. R.L.F.P.G. and F.M.J.J. wrote the manuscript with contributions of M.J.T.R. and H.H.. F.M.J.J. supervised the study.

2

A hidden layer of structural variation in transposable elements reveals potential genetic modifiers in human disease-risk loci

Elisabeth J. van Bree[1*], Rita L.F.P. Guimarães[1,2,3*], Mischa Lundberg[4], Elena R. Blujdea[1], Jimi L. Rosenkrantz[1], Fred T.G. White[1], Josse Poppinga[1], Paula Ferrer-Raventós[1], Anne-Fleur E. Schneider[1], Isabella Clayton[1], David Haussler[5], Marcel J.T. Reinders[6], Henne Holstege[2,3,6,7], Adam D. Ewing[4], Colette Moses[1], Frank M.J. Jacobs[1,7$]

1 Evolutionary Neurogenomics, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands
2 Genomics of Neurodegenerative Diseases and Aging, Department of Human Genetics, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands
3 Alzheimer Center Amsterdam, Department of Neurology, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands
4 Mater Research Institute - University of Queensland, Woolloongabba, QLD, Australia
5 UC Santa Cruz Genomics Institute, Santa Cruz, CA, USA; Howard Hughes Medical Institute, UC Santa Cruz, Santa Cruz, CA, USA
6 Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands
7 Amsterdam Neuroscience, Complex Trait Genetics, University of Amsterdam, Amsterdam, The Netherlands

* These authors contributed equally to this work
$ corresponding author

This work is published on:
Genome Research 32(4): 656–70.
doi:10.1101/gr.275515.121.

Abstract

Genome-wide association studies (GWAS) have been highly informative in discovering disease-associated loci, but are not designed to capture all structural variations in the human genome. Using long-read sequencing data, we discovered widespread structural variation within SVA (SINE-VNTR-*Alu*) elements, a class of great ape-specific transposable elements with gene-regulatory roles, which represents a major source of structural variability in the human population. We highlight the presence of structurally variable SVAs (SV-SVAs) in neurological disease-associated loci, and further associate SV-SVAs to disease-associated SNPs and differential gene expression using luciferase assays and expression quantitative trait loci data. Finally, we genetically deleted SV-SVAs in the *BIN1* and *CD2AP* Alzheimer's disease-associated risk loci and in the *BCKDK* Parkinson's disease-associated risk locus and assessed multiple aspects of their gene-regulatory influence in a human neuronal context. Together, this study reveals a novel layer of genetic variation in transposable elements that may contribute to identification of the structural variants that are the actual drivers of disease associations of GWAS loci.

Introduction

Discovering the genetic variation underlying human diseases is a common goal in human genetics, and the rapid increase of genome-wide association studies (GWAS) has generated a vast catalog of single nucleotide polymorphisms (SNPs) associated with specific traits and diseases (MacArthur et al. 2017). In most cases, GWAS do not identify the genetic variation that drives the trait, but use SNPs as markers to highlight trait-associated loci through linkage disequilibrium (LD) (Edwards et al. 2013). This calls for elaborate post-GWAS analysis to shed light on the genes and mechanisms involved in specific traits (Backman et al. 2021; Mortezaei and Tavallaei 2021). A comprehensive view of the genetic structural variants that exist within loci containing trait-associated SNPs is an essential first step to assessing how these variants may lead to disease susceptibility on both genetic and functional levels (Eichler 2019).

One source of structural variation that has not been sufficiently considered comes from transposable elements (TEs), which together constitute over 42% of the human genome (Audano et al. 2019; Lander et al. 2001; Linthorst et al. 2020; A. F. Smit 1999). Although the vast majority of TEs do not alter coding regions of our genome, some TE classes harbour strong gene-regulatory potential that can directly affect gene expression levels (Chuong, Elde, and Feschotte 2016; Fuentes, Swigut, and Wysocka 2018; Jacobs et al. 2014; Pontis et al. 2019; J. Wang et al. 2014). The TE-mediated regulatory effect on genes is highly tissue-specific and has been shown to be particularly prominent in a neuronal environment (Jacob-Hirsch et al. 2018; Miao et al. 2020; Pontis et al. 2019; Sundaram and Wysocka 2020; Trizzino, Kapusta, and Brown 2018). TEs are activated during aging, neurodegeneration, and neurological diseases, but whether this is a cause or a consequence of the disease pathology remains unknown in many cases (Frank et al. 2005; Guo et al. 2018; Wanhe Li et al. 2013; Shpyleva et al. 2018; Van Meter et al. 2014).

Only TEs of the *Alu*, L1 and SVA (SINE-VNTR-*Alu*) families can still actively spread through the genome, and new insertions cause variation between individuals in the form of presence/absence TE-insertional polymorphisms (Batzer et al. 1991; Brouha et al. 2003; Kazazian et al. 1988; Ostertag et al. 2003). TEs can alter gene regulation in the locus in which they insert, such that the presence or absence of a TE can lead to inter-individual differences in gene expression. There are approximately 60,475 *Alu*, 10,018 L1 and 6,417 SVA TE-insertional polymorphisms known, with new insertions occurring every 40, 63 and 63 births, respectively (R. L. Collins et al. 2020; Feusier et al. 2019). Some of these new insertions have been linked to diseases (Hancks and Kazazian 2016; Makino et al. 2007; Payer et al. 2017; Payer and Burns 2019; Pfaff et al. 2021; Sekar et al. 2016). Next to presence/absence TE polymorphisms, structural variation within fixed TEs (TE insertions observed in all individuals in the human population) has also been reported (Savage

et al. 2013, 2014), although the prevalence of this type of structural variation has remained elusive. The repetitive nature of TEs increases the propensity for unequal crossover events or DNA polymerase slippage during meiosis, for which variable number of tandem repeats (VNTRs) are especially susceptible (Brookes 2013). SVA elements harbour unusually large VNTRs as their internal segment and have a unique sequence composition compared to other VNTRs in our genome. The structural variation in VNTRs is particularly interesting as they are often associated with gene-regulatory functions, and many genes have accrued VNTRs as essential regulatory elements for their expression (Fondon et al. 2008; Lander et al. 2001).

It is becoming increasingly clear that gene-regulatory properties of TEs were co-opted during evolution, leading to the integration of TEs as novel gene-regulatory elements in pre-existing gene expression networks (Chuong, Elde, and Feschotte 2016; Cordaux and Batzer 2009). As such, TEs have become an integral part of normal human gene regulation. Because our genome has become dependent on TEs for specific aspects of gene regulation, structural variation within fixed TEs could account for inter-individual differences in temporal or spatial aspects of gene expression. Despite the possible roles structurally variable TEs may play in human health or disease, this level of structural variation has remained largely undocumented. This is mainly due to technical limitations associated with the highly repetitive DNA sequences within TEs, which makes identifying structural variations in TEs using short-read sequencing strategies extremely challenging. The development of long-read sequencing techniques provides, for the first time, the opportunity to accurately assess the level of structural variation (Eichler 2019). This allows for the evaluation of possible associations between disease susceptibility and specific structural variations found in fixed TEs in our genome (Audano et al. 2019; Chaisson et al. 2019; Ebert et al. 2021; Ewing et al. 2020; Porubsky et al. 2021; Sulovari et al. 2019).

In this study we discovered that SVA retrotransposons, a great ape-specific class of TEs, constitutes a major source of hidden genetic variation which is not taken into account by conventional genetic case-control studies. We set out to investigate the biological consequences of structural variability in SVAs, focusing on SV-SVAs in Alzheimer's disease (AD)- and Parkinson's disease (PD)-associated GWAS loci. We assessed the gene-regulatory influence of SV-SVAs in a human-neuronal context by genetic deletion of SVAs in 3 disease-associated loci. Our findings highlight the importance of careful mapping of structural variations within fixed TEs in the human population, and argue for their inclusion in complex trait genetics as a layer of genetic variation that may, in some cases, confer the actual disease susceptibility to a GWAS-identified locus.

Results

SVAs are a major contributor to inter-individual structural variation

In our search for structural variations in TEs that may be associated with differential susceptibility to neurological disorders, we first determined which TEs are likely to play a regulatory role in human neuronal gene expression. We specifically focused on TEs active in neuronal cells, because previous research has associated aberrant TE activation with several neurological disorders (Bragg et al. 2017; Frank et al. 2005; Guo et al. 2018; Wanhe Li et al. 2013; Shpyleva et al. 2018; Van Meter et al. 2014). To detect TEs that may have a regulatory influence on nearby genes in neuronal cells, we performed chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) for the enhancer-associated marker EP300 in human embryonic stem cell (hESC)-derived cortical organoids (Eiraku et al. 2008; Visel et al. 2009). Highest neuronal EP300 enrichment on TEs was observed for the active classes of TEs in our genome, including LINEs and all members of the SVA family (**Figure 1A-B**). These data support previously published findings in which histone ChIP-seq showed SVAs become active in a neuronal environment (Pontis et al. 2019).

For the top 20 neuronally active TE classes, we analysed inter-individual structural variations. We used data from the recently published 'patched human reference genome assembly' based on 15 human genomes sequenced by PacBio long read sequencing (Audano et al. 2019). In this assembly, structural variants of 50 base pairs (bp) or greater identified in any of the 15 individuals were included as alternate loci to improve the representation of allelic diversity in the reference assembly. Whereas a low level of structural variation was identified for almost all classes, we found an extraordinarily high level of structural variation in SVA elements (**Figure 1C**). Almost half of full-length SVAs were structurally variable in the human population. The biggest contribution came from SVA-D, SVA-E and SVA-F elements, representing the evolutionarily youngest classes of SVAs which contain the largest VNTR region. This region was highly enriched for structural variations, together with the 5' region which contains a hexamer repeat (**Figure 1D, Supplemental Figure S1A**). Most variations observed in SVAs were between 50-200 bp (**Figure 1E**), but ~20% of the variations were larger than 200 bp in size.

To validate these findings and to assess if any additional structural variation in SVA elements was not correctly captured, we performed an additional BLAST-alignment-based analysis on 5 of the PacBio genome assemblies (**Supplemental Table S1**). We focused on SVA sequences present on Chromosome 1 in the GRCh38 assembly as a proof of principle, and included small structural variations (20-50 bp difference in SVA size) that were not included in previous analyses (Audano et al. 2019). We observed structural variation within 49.1% of full-length SVAs present on Chr1, a slight increase over what was discovered in our first analysis. Structural variation was primarily observed in the centre (VNTR) and 5' (hexamer) region of the SVA

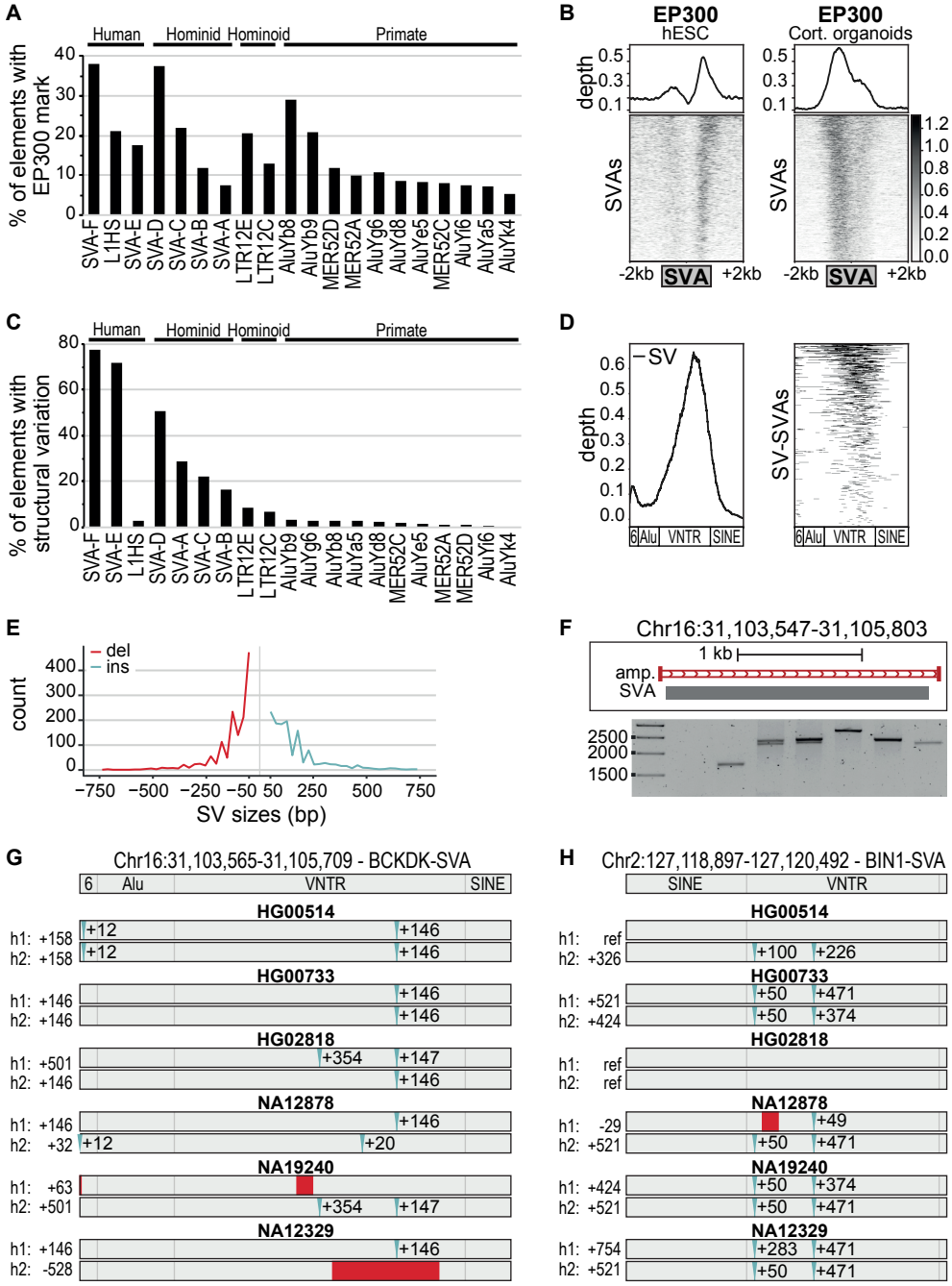


Figure 1: SVAs are a major contributor to inter-individual structural variation. A, Percentage of transposable elements with EP300 enhancer mark in cortical organoids; top 20 enriched elements are shown. **B**, Coverage heatmaps at full length SVAs (GRCh37) in hESCs and cortical organoids for EP300 (hESCs: average of 2 replicates, cortical organoids: average of 2 biological and 2 technical replicates). Bottom grey box: average size SVAs. **C**,

Percentage of 'full-length' TEs per class with structural variation based on Audano et al. (2019), grouped by species they originated in. **D**, Relative abundance of structural variation (left) and corresponding coverage heatmap (right) showing most structural variation resides in the VNTR region of SVAs. Approximate SVA structure is shown below. **E**, Distribution of structural variation (SV) sizes for insertions (ins) and deletions (del) in SVAs. **F**, Example of structural variants for SVA in PCR-amplified region Chr16:31,103,547-31,105,803 (GRCh38 assembly). PCR amplified region shown in red. **G-H**, Schematic overview of SV-SVAs in phased assemblies of Ebert et al. (2021) of listed genomes for specified regions with approximate size shown. Estimated location of insertions (blue) and deletions (red) compared to reference genome.

The distribution of structural variation between SVA classes was comparable to previous analysis, with most SV-SVA elements belonging to the youngest types.

To rule out sequencing or assembly errors leading to the observed variation, we selected five human-specific SVAs in which to validate the structural variations by PCR analysis. We confirmed structural variations within all five studied SVAs in a platform of 236 human genomic DNA samples (**Figure 1F**, **Supplemental Figure S1B-G)** and observed differences in size up to 1000 bp. For the SV-SVAs most extensively analysed (PCR on 81 and 236 genomes) we found additional variations not observed in the initially sequenced human genomes. We therefore used recently published phased genome assemblies (Ebert et al. 2021) and were able to trace all variants observed by PCR analysis in the haplotype-specific sequencing data (**Figure 1F-H, Supplemental Figure S1B**). This suggests the actual variability in SVAs may be even higher in the population than we found based on the 15 unphased PacBio genomes (Audano et al. 2019). Together these data reveal that SVA elements are a major source of structural variation between individuals.

SV-SVAs reside in gene-regulatory regions

We next determined whether SV-SVAs have the potential to influence gene expression. Previous studies have shown that SVAs have strong gene-regulatory potential and are identified as cis-regulatory elements by epigenetic marks (Bragg et al. 2017; Jacobs et al. 2014; Pontis et al. 2019; Savage et al. 2013, 2014; Trizzino et al. 2017). In addition, SVAs are enriched in open and active chromatin regions in numerous tissues and senescent cells, and show active epigenetic marks in the brain (De Cecco et al. 2013; Pontis et al. 2019; Trizzino, Kapusta, and Brown 2018). In further support of the regulatory potential of SVA elements, we showed that an SVA-F element upstream of a minimal promoter induced an 11.5× increase in luciferase activity relative to the empty vector control in mouse ESCs (mESCs) (P = 0.00227, 2-sided t-test with Bonferroni correction; **Figure 2A**). mESCs were used to fully assess the SVA's regulatory potential using an approach described in Jacobs et al. (2014), in which SVA activity is measured in a cellular model system that lacks KRAB zinc finger proteins, a family of endogenous SVA repressors found in

primates. A much lower but still significant enhancement of luciferase activity by the SVA-F element was also observed under conditions of KRAB zinc finger-mediated repression by the recently identified SVA-repressor ZNF91 (P = 0.00441, 2-sided t-test with Bonferroni correction; **Figure 2A**) (Haring et al. 2021; Jacobs et al. 2014).

SVAs have also been shown to influence gene expression through insertion near or within genes (Bragg et al. 2017; Haring et al. 2021; Jacobs et al. 2014; Trizzino, Kapusta, and Brown 2018). We found that 82% of SV-SVAs are located less than 50 kb from a transcription start site (TSS) (**Figure 2B**), 1.33× more than expected by random distribution throughout the genome ($P < 2.2 \times 10^{-16}$, chi-square test). This is in line with previous research showing a higher than expected number of SVAs in genic regions (Savage et al. 2013). SVAs in active chromatin regions are known to preferentially reside in gene bodies, suggesting they may function as intronic regulatory elements (Trizzino, Kapusta, and Brown 2018). We found that a high

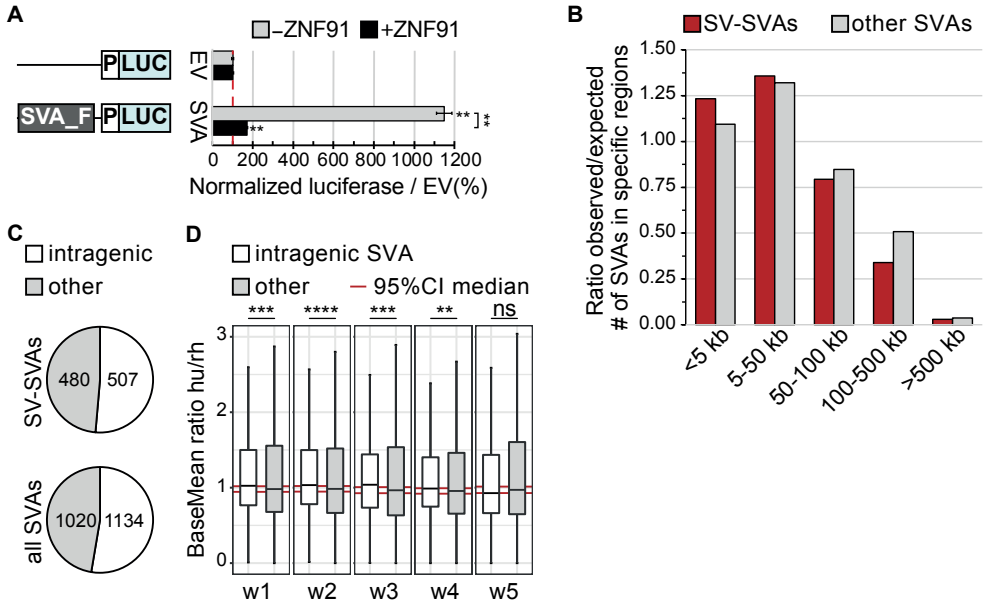


Figure 2: SV-SVAs reside in gene-regulatory regions. A, Luciferase activity of construct without (EV) and with an SVA element (SVA_F) upstream of a minimal promoter (P) in mESCs. N3n9, 2-sided t-test with Bonferroni correction, ** = $P < 0.01$. Error bars: standard error of the mean (SEM). **B**, Distribution of SV-SVAs (red) and non-SV-SVAs (grey) per distance to TSS. Only SVAs > 1 kb shown. **C**, Number of SV-SVAs and non-SV-SVAs that are intragenic is comparable ($\chi 2(1, N = 2{,}154) = 1.10$, $P = 0.29$). Only SVAs > 1,000 bp shown. **D**, Boxplots showing base mean expression ratio (human/rhesus) for transcripts with an intragenic SVA in humans (white; 1,151) and without (grey; 23,296) in ESC-derived cortical organoids of 1 to 5 weeks old. Red line shows 95% CI of 10,000× bootstrapped median of transcripts without an SVA with sample size of 1,151. Wilcoxon rank sum test, **** = $P < 0.0001$, *** = $P < 0.001$, ** = $P < 0.01$, ns = not significant.

percentage (51.4%) of SV-SVAs were intragenic, which was comparable to non-SV-SVAs (P = 0.29, chi-square test; **Figure 2C**). This supports findings from previous work, where additionally a positive correlation between SVA number and gene transcription was found (Gianfrancesco et al. 2019). To further analyse the intronic regulatory effect of SVAs on a genome-wide level, we compared the expression levels of genes with and without an intragenic SVA in cortical organoids derived from human and rhesus macaque embryonic stem cells (Field et al. 2019). Notably, rhesus macaques, like all non-great-ape species, do not contain any SVAs in their genome. Human genes with an intragenic SVA displayed a modest but significant increase in expression levels compared to the non-SVA ortholog in rhesus. This difference was not observed for genes that lack intragenic SVAs in both human and rhesus (**Figure 2D**, **Supplemental Figure S2**). These data provide support for an intronic regulatory role of SVAs. The location of SV-SVAs in close proximity to or within genes is consistent with the hypothesis that these SVAs may have gene-regulatory potential, and that structural variability within these elements could differentially regulate nearby genes.

Disorder-associated loci identified by GWAS are rich in SV-SVAs

Previous research showed that insertional polymorphisms of TEs are in LD with trait-associated SNPs, constituting a potential causative genetic variant for numerous human phenotypes (Payer et al. 2017). We extended this approach beyond TE-insertional polymorphisms, asking whether we could detect SV-SVAs in LD with established neurological disorder-associated SNPs, which could thus have possible involvement in the development of these conditions. Using reported SNPs from the NHGRI-EBI catalog as source, we considered a variety of complex diseases with a presumed major genetic component, and searched for SVAs within disorder-associated LD blocks. We determined the presence of SVAs in disorder-associated LD blocks in AD, PD, bipolar disorder, amyotrophic lateral sclerosis (ALS), autism spectrum disorder (ASD), multiple sclerosis (MS) and schizophrenia. AD-associated LD blocks contained 13 SVAs (in 12 out of 94 LD blocks) and PD-associated LD blocks contained 23 SVAs (in 19 out of 114 LD blocks). A substantial number of SVAs were also found in LD blocks associated with schizophrenia (69 SVAs in 59 out of 470 blocks), ASD (38 SVAs in 28 out of 178 blocks), bipolar disorder (36 SVAs in 21 out of 151 blocks) and MS (19 SVAs in 13 out of 198 blocks). Only one SVA was found in ALS-associated LD blocks, but the number of identified ALS-associated loci was also low (15 loci). The number of SVAs located within LD blocks was higher than the number expected by random distribution of the elements for PD (2.6×, P < .00001, chi-square test), ASD (2.1×, P < .00001), bipolar disorder (2.3×, P < .00001), schizophrenia (1.6×, P = .000106), and MS (1.9×, P = .00555), but not for AD (1.7×, P = .0545), neuroblastoma (1.7×, P = .586), or ALS (1.3×, P = .868).

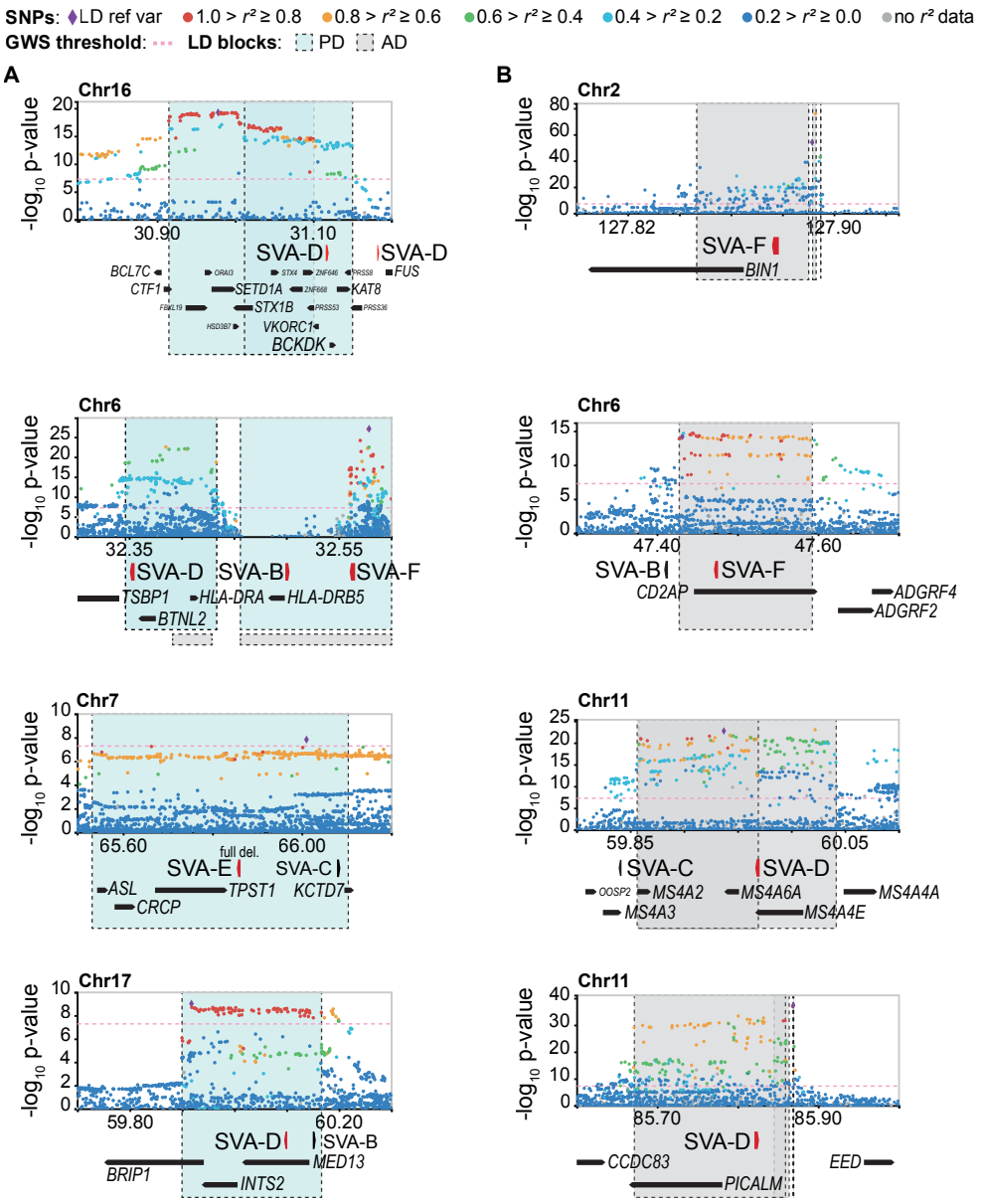


Figure 3: SV-SVAs reside in Parkinson's and Alzheimer's disease-associated LD blocks. **A-B**, Regional SNP association plots with SV-SVAs (red) shown in LD blocks of PD (blue) (**A**) and AD (grey) (**B**). The associated SNPs (AD; de Rojas et al., 2021, PD; Nalls et al., 2019) are plotted with their respective meta-analysis genome-wide significant *P*-values (GWS (Genome-wide significance), $P < 5 \times 10^{-8}$; as $-\log_{10}$ values) and are distinguished by linkage disequilibrium (r^2) of nearby SNPs on a blue to red scale, from $r^2 = 0$ to 1, based on pairwise r^2 values from the 1000 Genomes Phase3 (ALL) reference panel. Gene annotations: NCBI RefSeq Select database. Assembly GRCh37, scale in Mb.

We further focused our investigation on PD- and AD-associated loci, as aging is the greatest risk factor for those diseases, and the epigenetic changes associated with aging can uncover the hidden regulatory potential of TEs (Guo et al. 2018; Wanhe Li et al. 2013). For PD, 23 SVAs were located within 19 LD blocks, of which nine SVAs were structurally variable, and one showed insertional polymorphisms in the population (**Figure 3A, Supplemental Figure S3A**). SV-SVAs were also found intronic or close to the TSS of three other well-studied PD-associated genes: *NURR1*, *SYT11*, and *PARK7* (**Supplemental Figure S4A**) (Jankovic, Chen, and Le 2005; Nalls et al. 2014, 2019; Simón-Sánchez et al. 2009). For AD, the 13 SVAs found within those LD blocks were located near highly validated AD risk genes including *BIN1*, *PICALM* and *CD2AP* (**Figure 3B**) (Hamza et al. 2010; Jansen et al. 2019; Schwartzentruber et al. 2021). Eight of these SVAs were found to be structurally variable (**Figure 3B, Supplemental Figure S3B**). SV-SVAs were also found intronic or close to the TSS of five other well-studied AD-associated genes: *MS4A1*, *BACE1*, *PSEN1*, *DMXL1* and *SPRED2* (**Supplemental Figure S4B**) (Kelleher and Shen 2017; Ma, Yu, and Tan 2015; Schwartzentruber et al. 2021; Sherva et al. 2014; Yu et al. 2016). The GWAS upon which the LD blocks were based were performed without accurate knowledge of the level of structural variation in TEs. Our analysis suggests that like any other structural variation near the tag SNPs, structurally variable SVAs need to be considered as candidate causal factors for the disease risk contained within these loci. This emphasizes the need to assess the gene-regulatory roles of these specific SV-SVAs in more detail.

SVA variants are associated with disease-risk SNPs and have differential gene-regulatory potential

Based on the regulatory potential of SVAs and their presence in neurological disease-associated loci, we hypothesized that structural variation within SVAs could be a causal factor in disease risk for a number of GWAS-identified trait loci. To test this, we first focused on two human-specific SV-SVAs in loci associated with AD and PD. One of these SVAs was chosen as an example of an SVA nearby a gene body, and the other of an SVA in a gene-poor region. Structural variants of these SVAs were identified by PCR, and the presence of the disease-associated *tag* SNP was determined by Sanger sequencing. We focused on individuals homozygous for specific SVA structural variants, to ensure a correct association between the risk SNP and the presence of a certain SVA variant. The first SV-SVA we selected is located 2.7 kb from the TSS of the gene encoding for branched chain ketoacid dehydrogenase kinase (*BCKDK*) (**Figure 3A, Figure 4A-B**). This gene has an exonic SNP, rs14235, for which the minor (risk) allele is associated with a 1.17× to 1.36× increase in mean Lewy body count ($P < 0.041$-0.0026) (Heckman et al. 2017; Nalls et al. 2014). The putative mechanism of involvement of *BCKDK* in PD development is not entirely clear, but the nearby gene *KAT8* which resides in the same LD block may influence PD by modulating autophagic flux (Chang et al. 2017). We identified multiple structural variants for the nearby SVA by PCR analysis (**Figure 1F,**

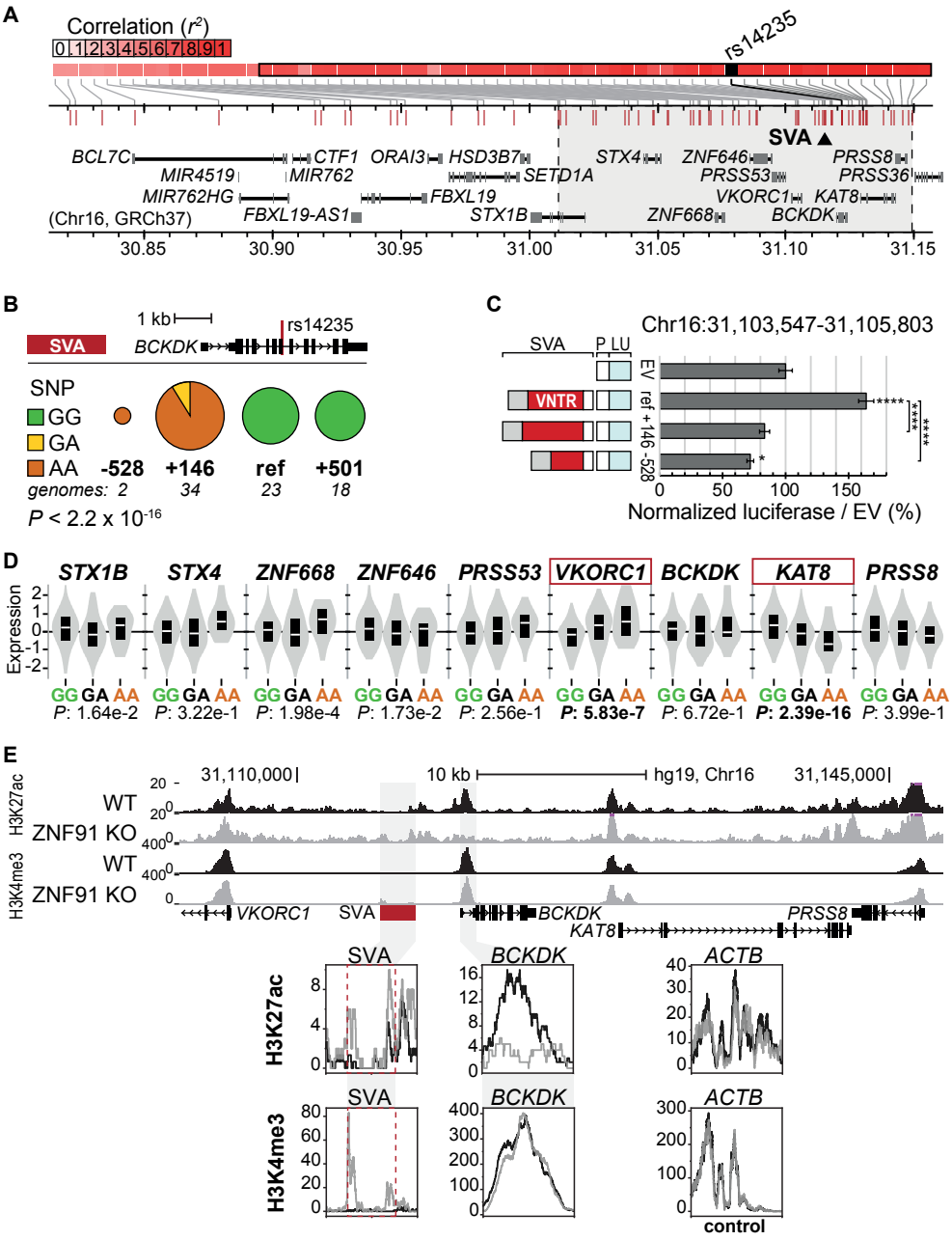


Figure 4: Structurally variable SVA near *BCKDK* links to disease-associated SNP and has the potential to differentially regulate nearby genes. A, Overview of LD block for rs14235, with area $r^2 > 0.8$ highlighted in grey. Approximate location of SVA marked with black triangle. **B**, rs14235 genotyping analysis for individuals homozygotic for *BCKDK*-SVA variants -600 (n = 2), ref (n = 23), +150 (n = 34), and +500 (n = 18). Ancestral allele = G, risk allele = A. Fisher's exact test: $P < 2.2 \times 10^{-16}$. **C**, Schematic overview of luciferase constructs (P = minimal promoter, LU = luciferase gene) with *BCKDK*-SVA variants (Chr16:31,103,547-

31,105,803 GRCh38), with corresponding luciferase activity in transfected mESCs. N3n9, except *BCKDK*-SVA ref (n = 8). One-way ANOVA with Tukey's multiple comparison, **** = $P < 0.0001$, * = $P < 0.05$. Error bars: SEM. **D**, Analysis of eQTL data in cortex for rs14235 for genes within the LD block with $r^2 > 0.8$. Normalized expression shown. Genes considered significant are shown in a red box. **E**, KO of the SVA repressor *ZNF91* lowers H3K27ac at the promoter of *BCKDK* and increases H3K4me3 methylation at the SVA near *BCKDK* in hESCs. *ACTB* shown as control enhancer region. Top: overview of locus, bottom: magnification of regions of interest.

Supplemental Figure S1F), and analysed rs14235 in 76 individuals homozygous for each structural variant of the nwearby SVA. There was a significant relationship between the SNP alleles and the SVA structural variants ($P < 2.2 \times 10^{-16}$, Fisher's exact test). The minor (risk) allele was exclusively observed in individuals containing the SVA variant with either -528 bp or +146 bp structural variations, while all individuals with the reference variant and the +500 bp SVA variant were homozygous for the ancestral (major) allele (**Figure 4B**). This indicates a strong link between specific structural variants of the SVA in this locus and the disease-associated haplotype identified by GWAS. This makes the SVA near *BCKDK* a potential candidate for carrying the actual mechanistic cause for the association of the minor (risk) allele of rs14235 with increased Lewy body count and increased risk of PD.

We next examined whether each of the SVA structural variants associated with the minor (risk) allele exhibited different gene-regulatory potential. Both SVA variants (+146 bp and -528 bp) that consistently occur in conjunction with the minor allele showed significantly reduced gene-regulatory potential compared to the reference variant/ancestral allele ($P < 0.0001$, two-way ANOVA, Tukey's multiple comparison; **Figure 4C**). The difference in reporter gene expression between the +146 bp variant and the reference SVA was still observed following the overexpression of *ZNF91*, a factor mediating strong repression of SVAs in a human-cellular context ($P = 0.0274$, two-way ANOVA, Tukey's multiple comparison; **Supplemental Figure S5A**). This demonstrates that structural variations within the SVA near *BCKDK* harbour highly differential gene-regulatory potential, and each of these SVA variants may therefore have different effects on nearby gene expression.

We next determined whether the presence of different SVA structural variants could also influence gene expression *in vivo.* Based on the knowledge that the rs14235 SNP allele was significantly associated with the structural variant of the nearby SVA, we compared eQTL data for rs14235 in brain tissue. We examined expression data from cortex and substantia nigra, two well-studied brain areas involved in PD, and found rs14235 was significantly associated with *VKORC1* and *KAT8* expression in cortex (**Figure 4D**). *KAT8* expression in the cortex was decreased in the presence of the minor (risk) allele for SNP rs14235. *KAT8* expression in substantia nigra displayed a trend in the same direction (**Supplemental Figure S5C**). These data are

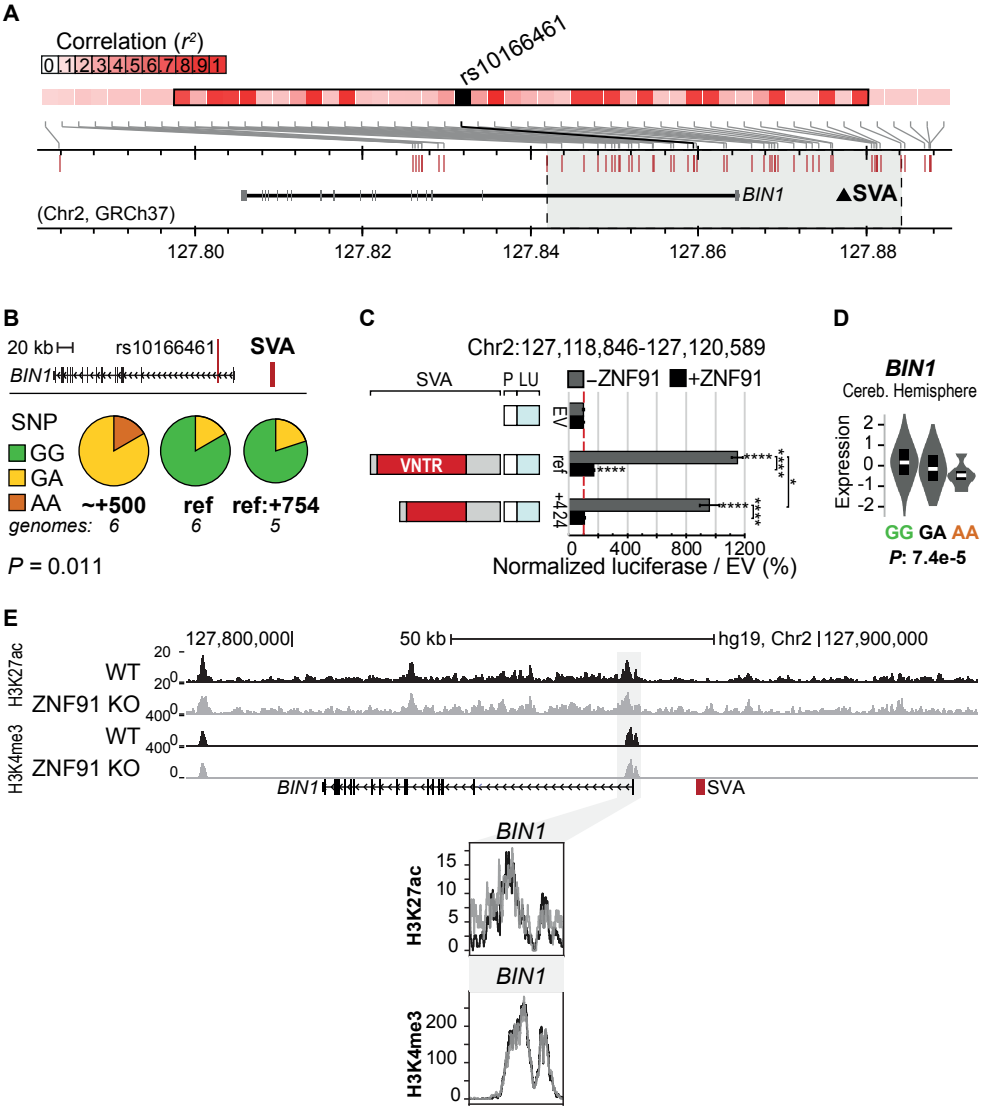


Figure 5: Structurally variable SVA near *BIN1* links to a disease-associated SNP and has the potential to differentially regulate nearby genes. A, Overview of LD block for rs10166461, with area $r^2 > 0.8$ highlighted in grey. Approximate location of SVA marked with black triangle. **B**, rs10166461 genotyping analysis for individuals homozygotic for *BIN1*-SVA variants ref (n = 6) and +424 or +521 (~+500, n = 6) and heterozygotic for ref and +754 (n = 5). Ancestral allele = G, risk allele = A. Fisher's exact test: $P = 0.0108$. **C**, Schematic overview of luciferase constructs (P = minimal promoter, LU = luciferase gene) with *BIN1*-SVA variants, with corresponding luciferase activity in transfected mESCs with and without ZNF91. N3n9. Two-way ANOVA with Tukey's multiple comparison. **** = $P < 0.0001$, * = $P < 0.05$. Error bars: SEM. **D**, Analysis of eQTL data in cerebellar hemisphere for rs10166461 and *BIN1*. **E**, KO of the SVA repressor *ZNF91* does not influence H3K27ac and H3K4me3 at the promoter of *BIN1* in hESCs. Top: overview of locus, bottom: magnification of regions of interest.

consistent with the observed repressive regulatory effect of the two SVA variants that coincide with the minor (risk) allele (**Figure 4B-D**). Individuals with the minor (risk) allele also showed decreased expression of *KAT8*, *ZNF646*, *PRSS36* and *RP11-196G* in other brain areas. No association of rs14235 with *BCKDK* expression levels was found. To further assess the regulatory effect of the SVA in this locus, we asked whether ectopic activation of the SVA would lead to changes in epigenetic marks in the *BCKDK* locus. We previously showed that in human ESCs carrying a *ZNF91* genetic deletion, SVAs become epigenetically activated (Haring et al. 2021). This is also the case for the SVA near *BCKDK*, which showed an increase in H3K4me3 in *ZNF91* knockout (KO) hESCs (**Figure 4E**). We also found epigenetic alterations at the *BCKDK* promoter, indicating that in human ESCs, the ectopic activation of the upstream SVA affected the epigenetic state of the *BCKDK* promoter.

The second SV-SVA we analysed is located 12 kb upstream of the TSS of the AD-associated gene *BIN1*. The upstream region of this gene has been named the second-most important susceptibility locus in late-onset AD (LOAD) (http://www.alzgene.org). BIN1 colocalizes and interacts with tau protein, and SNPs upstream of *BIN1* have been linked to increased expression of the gene and risk for LOAD (Carrasquillo et al. 2011; Chapuis et al. 2013; Hu et al. 2011; J.-C. Lambert et al. 2011, 2013; J. H. Lee et al. 2011; Masoodi et al. 2013; Seshadri et al. 2010; Wijsman et al. 2011). The upstream region of *BIN1* contains a structurally variable SVA (**Figure 3B, Figure 5A-B**) (Chr2:127,118,897-127,120,492, GRCh38), located just 17 kb from the AD-associated risk SNP rs10166461 (allele effect -0.2636, $P = 3.82 \times 10^{-6}$) (Beecham et al. 2014). In a similar PCR validation to determine the association of SVA variants with the AD-associated SNP rs10166461, we found that the risk allele was more often observed in individuals carrying the +424 bp or +521 bp (~+500) SVA variant, and although the number of sequenced individuals was limited, we found a significant association between SNP allele and SVA variant ($P = 0.011$, Fisher's exact test; **Figure 5B**). Structural variants of this SVA associated with the minor (risk) allele showed significant differential regulatory potential in the luciferase reporter assay ($P < 0.05$, two-way ANOVA, Tukey's multiple comparison; **Figure 5C**). In the presence of the SVA repressor ZNF91 this difference was not maintained. Furthermore, eQTL analysis revealed that individuals carrying the risk allele of rs10166461 displayed significantly lower expression of *BIN1,* but this effect was only observed in the cerebellum (**Figure 5D**). Despite the apparent tissue-specific association between rs10166461 and changes in *BIN1* expression, the direction of the gene expression effect corresponded with the results of our luciferase assays, in which the +424 bp variant (most often present in individuals with the risk allele of rs10166461) had a significantly lower regulatory potential than the reference SVA variant (**Figure 5C**). Consistent with the cerebellum-specific association of rs10166461 with differences in *BIN1* expression, in cortical organoids we found no significant alterations for H3K4me3 and H3K27ac histone marks in the *BIN1* locus under *ZNF91* KO conditions (**Figure 5E**). We also examined eQTL data

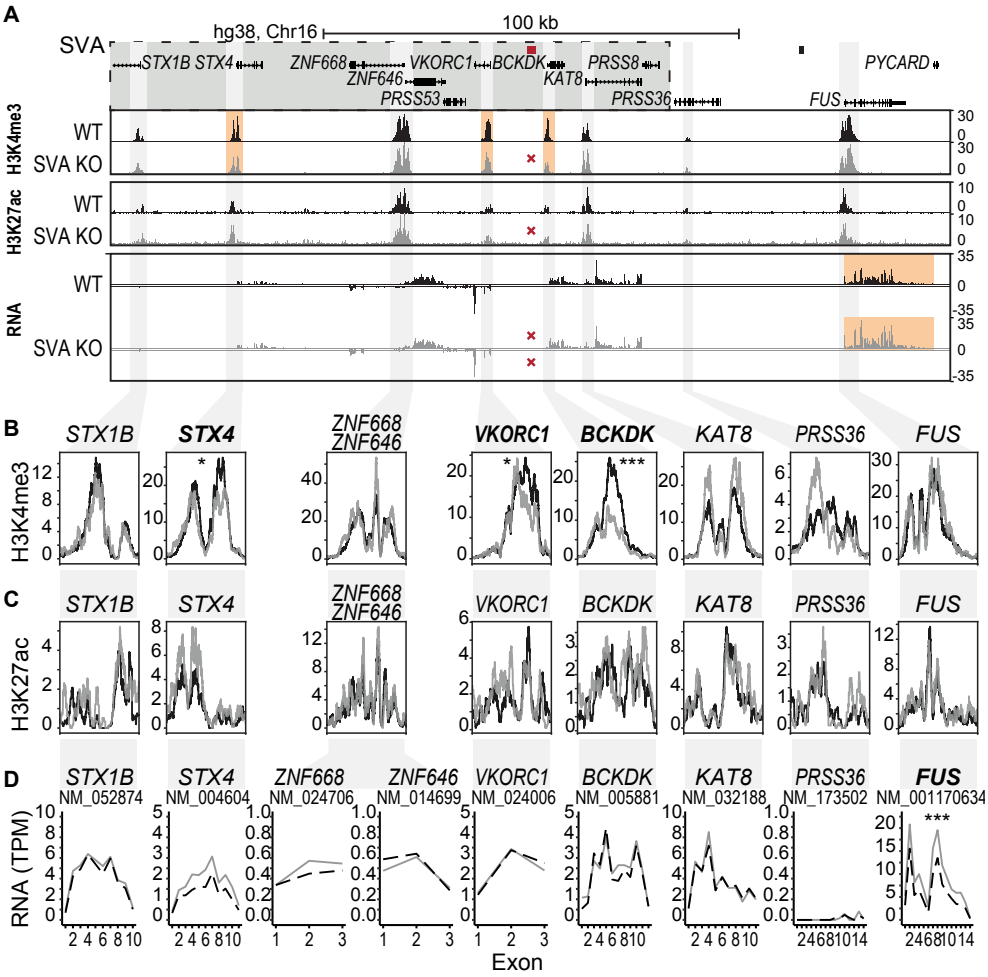


Figure 6: SVA deletion alters epigenome and nearby gene expression. A, Overview of locus. LD block shown in grey box, location of SVA removed by CRISPR-Cas9 KO shown in red. **B-D**, Magnification of genes within a 200 kb region of deleted SVA. **B**, H3K4me3 ChIP-seq, mean of 3 replicates shown. **C**, H3K27ac ChIP-seq, mean of 3 replicates shown. **D**, Mean expression of three replicates shown per exon for transcripts. Adjusted *P*-value of DESeq2 shown for whole transcript and ChIP peaks. *** = $P < 0.001$, * = $P < 0.05$.

for other SV-SVAs in AD and PD loci. Significant differential expression associations for SNPs within the LD blocks overlapping these SVAs were observed for numerous genes, including *CD2AP*, *HLA-DRB1/5/6*, *PICALM*, *MS4A6A*, *MS4A4E*, and *APOE* (see **Supplemental Table S2**). Together, these data show the relationship between SV-SVAs and GWAS-identified risk SNPs, and suggest that these SVA variants could potentially be involved in disease susceptibility through gene regulation in these loci. Further functional research on specific loci remains necessary to fully understand which and how SV-SVAs may contribute to disease development.

Genetic deletion of SVA elements in AD/PD loci alters the epigenome and nearby gene expression

To further investigate the gene-regulatory role of the SVAs in the *BCKDK* and *BIN1* loci, we engineered genetic deletions of the SVAs in hESCs using CRISPR-Cas9 technology. *BIN1*-SVA KO and *BCKDK*-SVA KO hESCs were subsequently directed into a neuronal fate by generating cortical organoids to analyse the regulatory effect of the SVAs in a neuronal context (**Figure 6, Supplemental Figures S6, S7**). Five weeks after the onset of neuronal differentiation, organoids were harvested for RNA-seq and ChIP-seq. We supplemented a transcriptomic profile of the loci in the respective SVA KO cells with ChIP-seq profiles of H3K4me3 and H3K27ac epigenetic marks to identify genes that may be under regulatory influence of the SVA.

Genetic deletion of the SVA located in the *BCKDK* locus resulted in a statistically significant reduction in H3K4me3 at the promoters of *BCKDK*, *VKORC1* and *STX4* (**Figure 6A-B, Supplemental Table S3**). While we observed an increase of H3K4me3 signal in part of the promoter of *KAT8* and *PRSS36*, this result did not reach the threshold of significance. No significant changes were observed for H3K27ac, which was detected at a much lower level overall (**Figure 6C, Supplemental Table S3**). Despite the clear changes in histone marks, the RNA expression levels of *BCKDK, VKORC1*, *STX4* or *KAT8* were not significantly altered in *BCKDK*-SVA KO neurons. Instead, the deletion of the SVA in the *BCKDK* locus resulted in a strong increase in expression of the gene encoding RNA-binding protein FUS (*FUS*; adjusted $P = 1.53 \times 10^{-11}$, DESeq2), located approximately 75 kb from the deletion site (**Figure 6D**). While no direct link has been reported for *FUS* with PD, this gene is linked to other neurodegenerative diseases such as ALS, frontotemporal dementia (FTD) and essential tremor (ET) (Deng, Gao, and Jankovic 2014; Kwiatkowski et al. 2009; Mackenzie, Rademakers, and Neumann 2010; Vance et al. 2009; Wu et al. 2013). The expression of *FUS* was previously shown to be controlled by an SVA directly upstream of the *FUS* gene, which further supports our current observations (Savage et al. 2014). None of the other 61 genes in a window of 1 Mb up- and downstream of the SVA showed differential expression upon SVA deletion, with the exception of *SRCAP*, located 404 kb from the SVA (**Supplemental Figure S8**). Removal of the SVA 12 kb from the TSS of *BIN1* did not result in detectable epigenetic and transcriptional changes in the locus in the context of the cortex model system we used (**Supplemental Figure S6**), but this may reflect the selective cerebellum-specific association of *BIN1* expression with the risk allele and its associated SVA variant (**Figure 5D**). Therefore, use of a model system for cerebellum rather than cortex may be important for investigating the effect of SVA KO in the *BIN1* locus.

In a final approach, we aimed to investigate the gene-regulatory role of an intronic SVA in *CD2AP*, another AD-associated gene (**Figure 3B**). We engineered a genetic deletion of the SVA in hESCs using CRISPR-Cas9 (**Figure 7A**) and generated

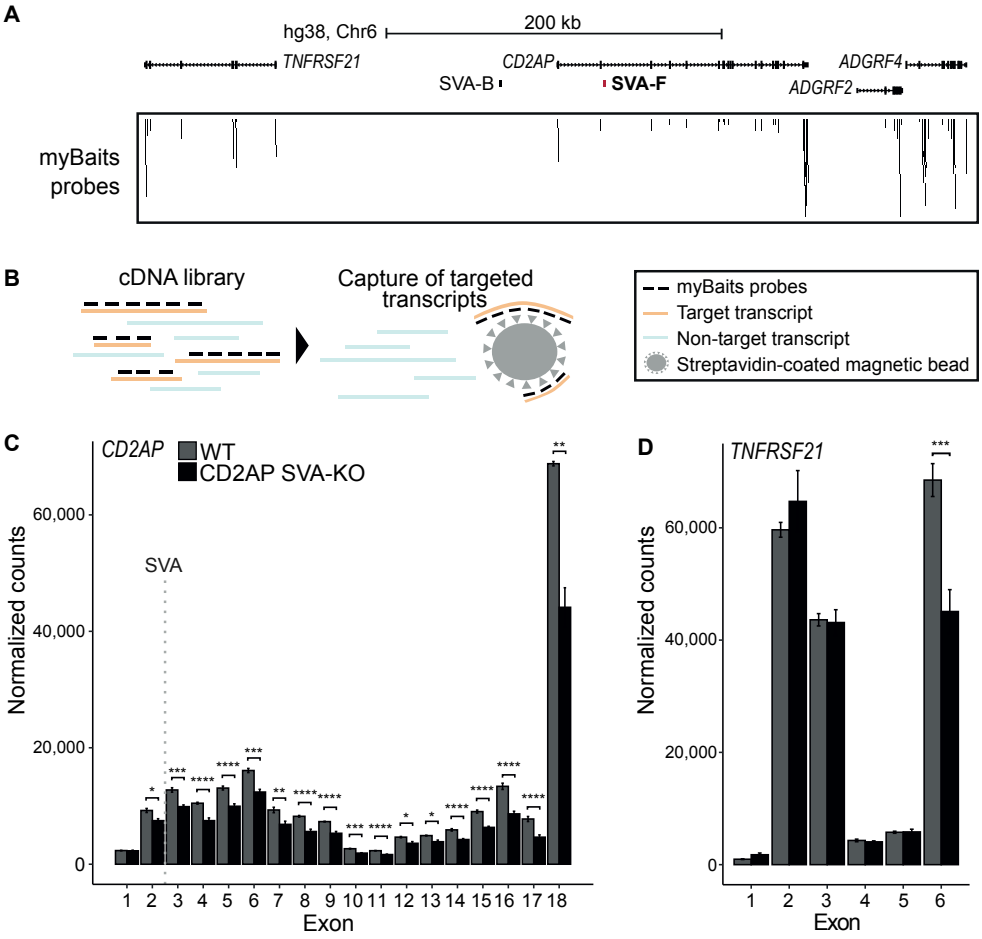


Figure 7: Intronic SVA deletion alters exon expression of *CD2AP* gene. A, Overview of *CD2AP* locus. Location of SVA removed by CRISPR-Cas9 KO shown in red. Probes from myBaits targeting all exons in *CD2AP* locus. **B**, Schematic of capture of targeted transcripts with myBaits probes. **C-D**, Normalized mean expression of three replicates shown per exon of *CD2AP* and the nearby, highly expressed gene *TNFRSF21*. Location of SVA indicated with dashed grey line. Adjusted *P*-value from DESeq2 shown for each exon. **** = $P < 0.0001$, *** = $P < 0.001$, ** = $P < 0.01$, * = $P < 0.05$.

cortical organoids from the *CD2AP*-SVA KO hESCs. We performed targeted mRNA enrichment prior to RNA-seq (RNA CaptureSeq) to specifically enrich for transcripts derived from the *CD2AP* locus and 38 other unrelated control genes, allowing us to track the expression levels of each of the *CD2AP* exons separately. This was important because intronic SVAs may differentially influence the expression level of exons up- and downstream of the SVA, as is observed for a pathogenic intronic SVA that leads to X-linked dystonia-parkinsonism (Bragg et al. 2017). Furthermore,

this RNA CaptureSeq approach increased the overall resolution of the selected gene transcripts, allowing for a more accurate assessment of the consequences of removal of the intronic SVA for *CD2AP* gene expression (**Figure 7B**). DESeq2 analysis comparing gene expression in *CD2AP*-SVA KO cortical organoids to unedited control organoids revealed significantly lower expression levels for all *CD2AP* exons downstream of the SVA insertion site (**Figure 7C**). Of the two exons upstream of the SVA, exon 1 was not significantly affected by SVA removal, and only a modest reduction was observed for exon 2. Although the final exon of the nearby gene *TNRFSF21* also showed a difference in expression between KO and unedited control, a consistent change in expression levels for *TNRFSF21* was not observed (**Figure 7D**). No significant differences in expression levels between *CD2AP*-SVA KO and unedited control organoids were observed for the vast majority of control genes residing in other genetic loci included in the RNA CaptureSeq analysis (**Supplemental Table S4**).

Taken together, these results demonstrate that structurally variable SVAs in important neurological disease-associated loci have the potential to affect the surrounding epigenome and/or transcriptome. While the genetic variants that cause gene-dysregulation remain undiscovered for the majority of the GWAS-identified disease risk loci, our data present structural variation in SVAs and other TEs as novel potential genetic modifiers of gene regulation in these loci.

Discussion

Our study reveals that structural variations in SVAs are previously unknown candidates for causal genetic variants not captured by conventional GWAS. Multiple lines of research indicate that SVAs can influence the expression of nearby genes (Bragg et al. 2017; De Cecco et al. 2013; Haring et al. 2021; Jacobs et al. 2014; Pontis et al. 2019; Savage et al. 2013, 2014; Trizzino et al. 2017; Trizzino, Kapusta, and Brown 2018), and the extensive structural variability that we observed in SVAs genome-wide suggests that this influence could be highly variable within the human population. For two SVAs analysed in detail in this study, we find that specific SVA variants are significantly associated with the presence of nearby risk SNPs identified in GWAS, whereas other variants are not. Although all individuals will carry a certain SVA in a disease-associated locus, the particular structural variation present within this SVA in each individual may have important consequences. In fact, we demonstrated that different SVA structural variants exhibit differential gene-regulatory potential *in vitro*. We propose that inter-individual structural variation in SVA elements may be an important genetic variable that may directly and functionally link to disease risk. Furthermore, in those cases where an SVA influences gene expression in the locus, the structural variation in the SVA may be an even more accurate indicator for disease risk than the risk SNP originally identified by GWAS.

Previous research into structural variation within specific SVAs supports our hypothesis that these variations can be involved in disease. In X-linked dystonia-parkinsonism, the length of a disease-causing SVA insertion in an intron of the *TAF1* gene displayed differential gene-regulatory potential and correlated with disease onset (Bragg et al. 2017). Two other SV-SVAs near ALS- and PD-associated genes were also shown to have differential regulatory potential, although in these cases the causal relationship to the disease remains unproven (Savage et al. 2013, 2014). Our study reveals multiple SV-SVAs in AD- and PD-associated loci that are potential contributors to disease susceptibility. eQTL data of SNPs in the *BCKDK* and *BIN1* loci showed an association of the disease-risk allele to changes in gene expression. Importantly, for both loci the direction of gene expression change associated with the risk alleles was consistent with our luciferase assay data where different SVA variants were explored functionally. The two SVA variants associated with the risk allele rs14235 in the *BCKDK* locus displayed an expansion (+146 bp variant) and deletion (-528 bp variant) relative to the reference, however both variants showed a significant reduction in gene regulatory potential relative to the reference in the luciferase reporter assay. Potential mechanisms underlying differences in regulation of gene expression between SVA variants remain to be explored in future work.

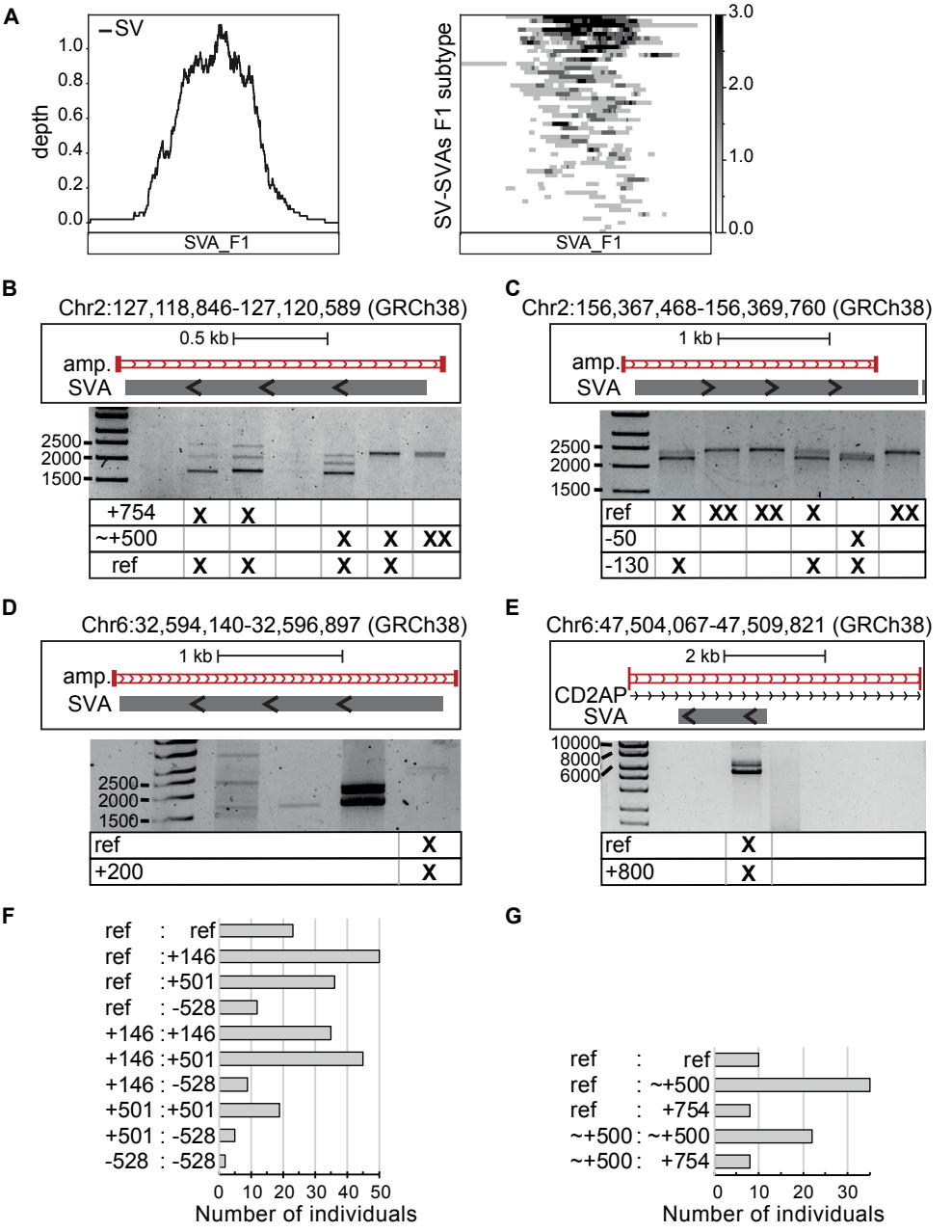
Removal of *BCKDK* and *CD2AP* SV-SVAs produced subtle but significant changes on gene expression and epigenetic landscape, whereas no change was detected upon deletion of the SV-SVA in the *BIN1* locus. This is not completely unexpected,

since SV-SVAs are only one potential layer of gene regulation amongst many others. If SV-SVAs produced large changes in gene expression, we might expect the phenotypic consequences of these variations to be evident in early life. On the contrary, it could be the case that the influence of SV-SVAs is highly tissue- or condition-specific and it is not easily captured in our model systems. Structural variations in SVAs may be particularly important in age-related neurological diseases, as the epigenetic changes associated with aging could uncover the latent regulatory potential of these TEs (Guo et al. 2018; Wanhe Li et al. 2013). In that sense, SV-SVAs may contribute to the progressive character of these diseases, because with increased age the differences in regulatory potential of SVA variants could become increasingly more prominent. For the SVA in the PD-associated *BCKDK* locus, along with an influence on the epigenetic state of nearby gene promoters, we found a possible gene-regulatory effect on *FUS*, a gene located just outside of the PD-associated LD block. These findings are compatible with the concept that the causal variant within a disease-associated locus could be a gene-regulatory element that influences the expression of genes inside and/or outside of the LD block. Even though there is no strong support for a role of *FUS* dysregulation in PD, *FUS* has been associated with other neurodegenerative diseases such as ALS, FTD and ET (Deng, Gao, and Jankovic 2014; Kwiatkowski et al. 2009; Mackenzie, Rademakers, and Neumann 2010; Vance et al. 2009; Wu et al. 2013). In addition, *PINK1* and *PARKIN*, two well-known PD-associated genes, were found to be genetic modifiers of *FUS*-induced neurodegeneration (Chen et al. 2016). Given the role of *FUS* in other neurodegenerative diseases, it is not unlikely that *FUS*-dysregulation, caused by specific SVA variants in the *BCKDK* locus, is in some way a contributor to the neuropathology in PD, but whether *FUS* is indeed regulated by the SVA in the *BCKDK* locus remains to be investigated in detail.

Collectively, this study reveals an extensive level of structural variation in TEs that has escaped detection by SNP arrays and short-read sequencing techniques. Our analysis stresses the importance of accurate mapping and characterization of structural variations in TEs in disease-associated loci. The methods used by GWAS do not identify structural variations due to their repetitive nature and size, so there may be numerous additional risk loci where the disease association is caused by structural variation in TEs. Whole-genome sequencing (WGS) is perfectly suitable for genotyping insertional polymorphisms, however variations in repetitive sequences are still overlooked. Ebert et al. (2021) describe a complementary method to identify structural variations with haplotype-resolved assemblies from long-read sequencing technology, and genotype the identified structural variations on short-read sequencing data for further population analysis. Although this method allows for the discovery of structural variations on short reads, the limiting factor is still highly repetitive DNA regions such as SVAs, thus giving an underrepresentation of the true level of structural variability. Indeed, we found that the poor genotype quality of the specific SVAs of interest in the PanGenie callset

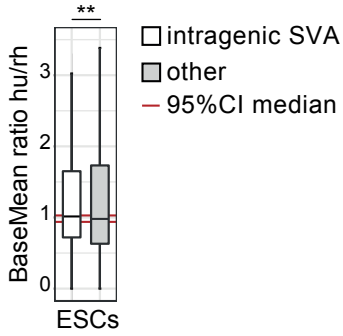
of 1000 Genomes samples (Ebert et al. 2021) made for a non-informative SVA-eQTL analysis. In the future, a targeted genotyping of SVAs using an alternative genotyper may be able to improve the SVA genotyping quality using short reads. Alternatively, long-read sequencing may be a better technology to fully characterize structural variations and their association with disease, because it is estimated that ~83% of insertions are missed with short-read-calling algorithms (Chaisson et al. 2019). Recently haplotype-resolved assemblies have been reported, which allow for more sensitive structural variation discovery by taking into account both alleles for each locus. This strategy outputs a much more accurate representation of the total structural variability in the human genome, indicating that the structural variation dataset used in our study is an underrepresentation of the total structural variation in the human population. Although long-read sequencing is expensive and not yet widely used, it is a matter of time before it becomes the standard way to perform genome sequencing. The reliability of genetic disease susceptibility markers and predictions will be significantly enhanced once GWAS include SNPs alongside all other structural variations, regardless of the size of these variants or whether they are located within TEs. Finally, increased awareness of the hidden gene-regulatory potential of non-coding DNA elements and the availability of human model systems and genetic engineering to study functional non-coding elements will allow us to functionally interrogate these variants to gain a better understanding of the mechanism of disease. Ultimately, this may contribute to the discovery of novel therapeutic targets and strategies for disease treatment.

Supplemental Figures

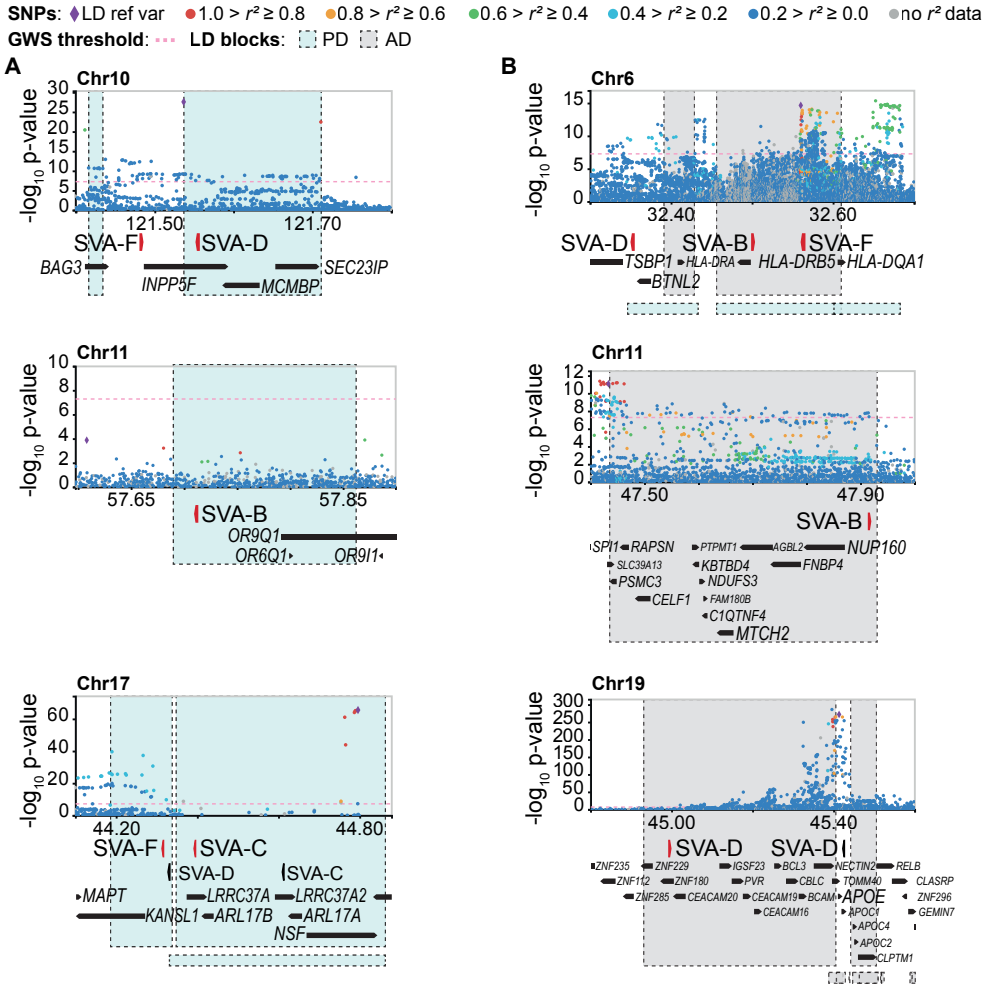


Supplemental Figure S1: Structural variation in SVAs and PCR confirmation of SV-SVA. **A**, Relative abundance of structural variation (left) and corresponding coverage heatmap (right) showing most structural variation resides in the VNTR region of the SVA_F1 subtype. **B-E**, Example of structural variants for SVA in PCR-amplified region Chr2:127,118,846-

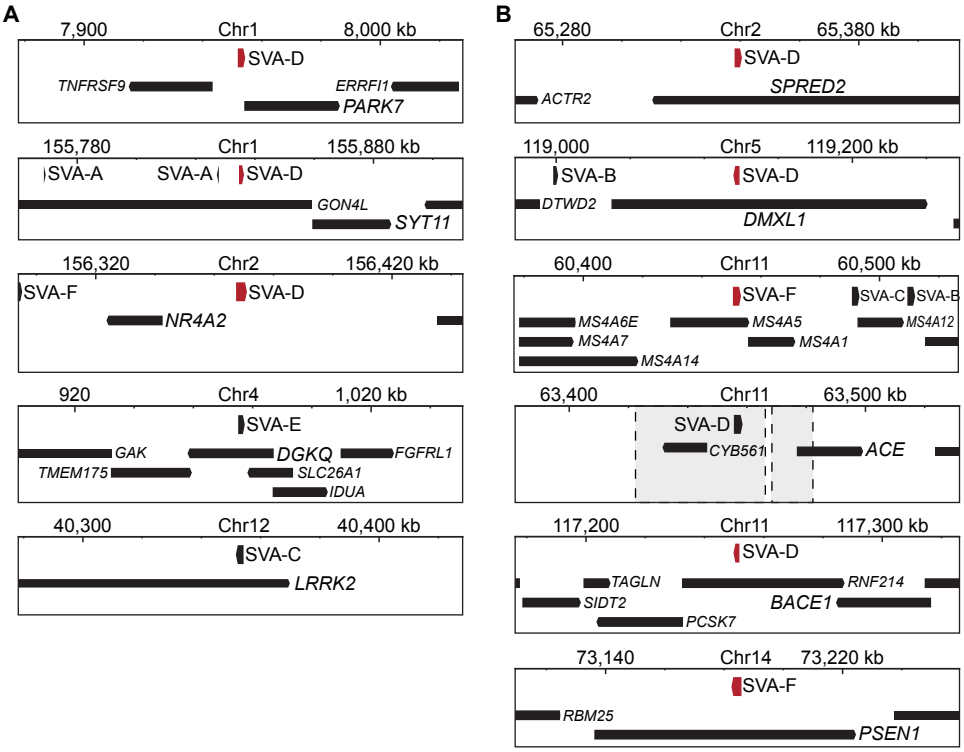
127,120,589 (*BIN1*) (**B**), Chr2:156,367,468-156,369,760 (*NURR1*) (**C**), Chr6:32,594,140-32,596,897 (*HLA-DRB1, last lane, faint bands*) (**D**), and Chr6:47,504,067-47,509,821 (*CD2AP*) (**E**). PCR amplified region shown in red, X indicates allelic variants. **F-G**, Prevalence of most commonly observed SVA variants for SVA in amplified region Chr16:31,103,547-31,105,803 (GRCh38 assembly) (n = 236) (**F**) and Chr2:127,118,846-127,120,589 (n = 81) (**G**).



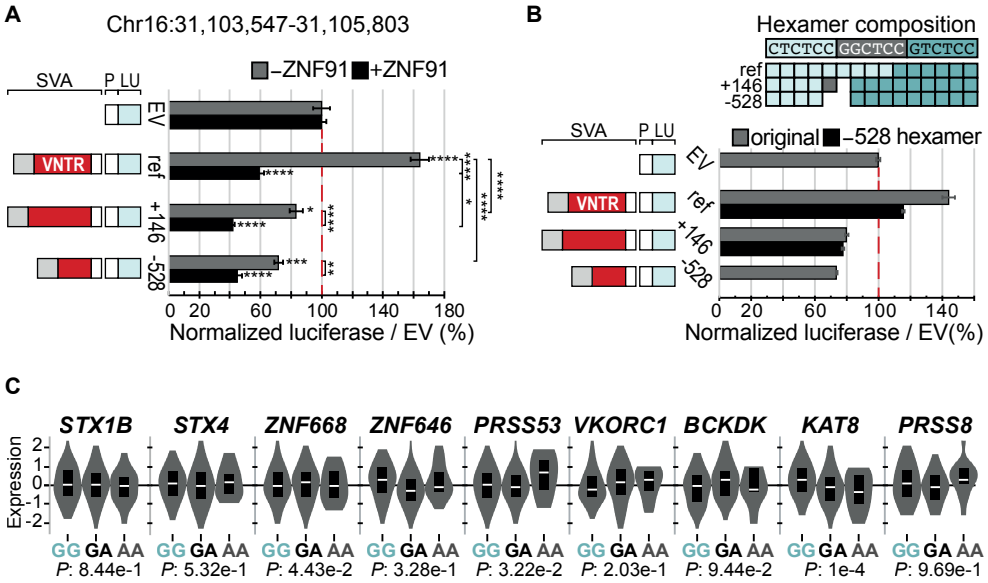
Supplemental Figure S2: Base mean expression ratio (human vs. rhesus) of transcripts with and without an intragenic SVA in humans. Boxplot showing base mean expression ratio (human/rhesus) for transcripts with an intragenic SVA in humans (white; 1,151) and without (grey; 23,296) in embryonic stem cells (ESC). Red line shows 95% CI of 10,000× bootstrapped median of transcripts without an SVA with sample size of 1,151. Wilcoxon rank sum test, **** = $P < 0.0001$, *** = $P < 0.001$, ** = $P < 0.01$, ns = not significant.



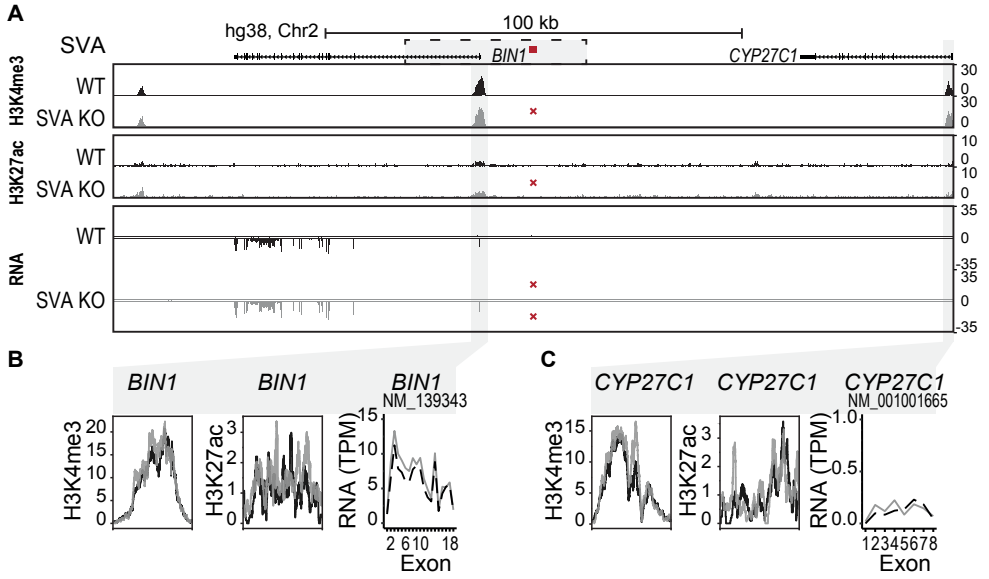
Supplemental Figure S3: SV-SVAs reside in Parkinson's and Alzheimer's disease-associated LD blocks. A-B, Regional SNP association plots with SV-SVAs (red) shown in LD blocks of PD (blue) (**A**) and AD (grey) (**B**). The associated SNPs (AD; Rojas *et al.,* 2021, PD; Nalls *et al*., 2019) are plotted with their respective meta-analysis genome-wide significant *P*-values (GWS, $P < 5 \times 10^{-8}$; as $-\log_{10}$ values) and are distinguished by linkage disequilibrium (r^2) of nearby SNPs on a blue to red scale, from $r^2 = 0$ to 1, based on pairwise r^2 values from the 1000 Genomes Phase3 (ALL) reference panel. Gene annotations: NCBI RefSeq Select database. Assembly GRCh37, scale in Mb.



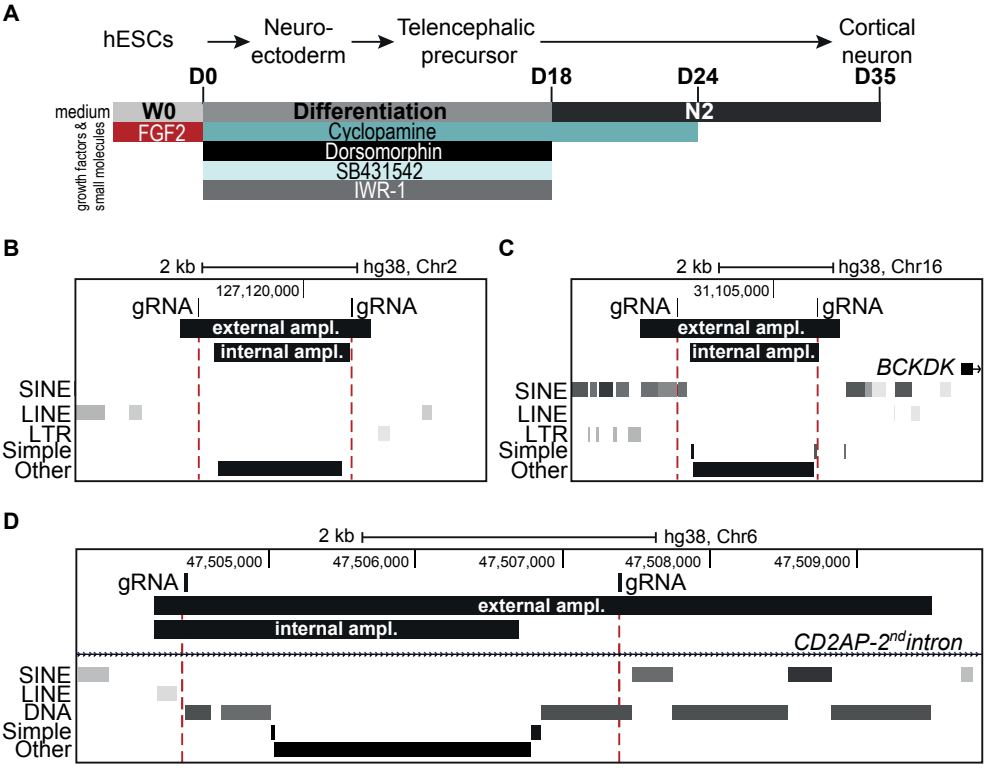
Supplemental Figure S4: SV-SVAs reside near Parkinson's and Alzheimer's disease-associated risk genes. A-B, Overview of PD (**A**) and AD (**B**) associated risk genes with SV-SVA (red) nearby. 150 kb regions shown, except for *DMXL1* (300 kb), with SVA centred. Grey area: LD blocks.



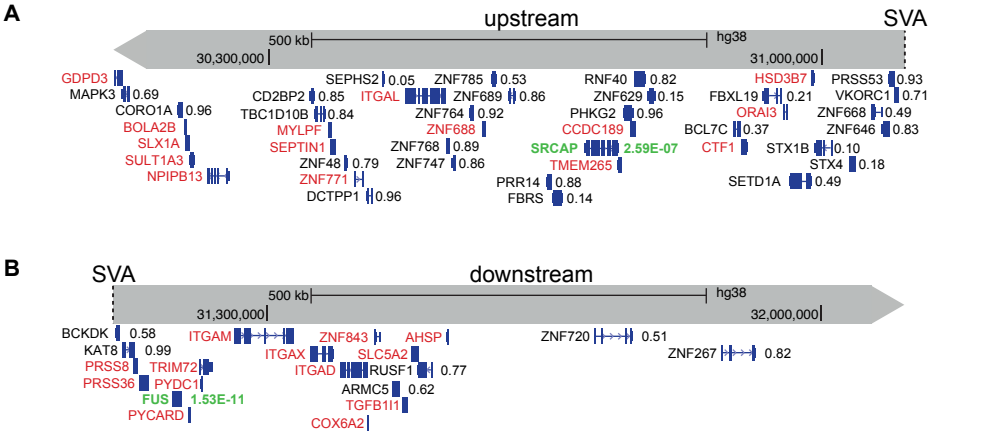
Supplemental Figure S5: SV-SVAs have differential gene-regulatory potential. A, Schematic overview of luciferase constructs with *BCKDK*-SVA variants, with corresponding luciferase activity in transfected mESCs. N3n9, except *BCKDK*-SVA ref without ZNF91 (n = 8). Two-way ANOVA with Tukey's multiple comparison test. **** = $P < 0.0001$, *** = $P < 0.001$, ** = $P < 0.01$, * = $P < 0.05$. **B**, Luciferase activity for *BCKDK*-SVA variants with their original or -528-variant hexamer. Small variations in the composition of the hexamer region between the different SVA variants shown on top. The variation in the composition of the hexamer region of this SVA did not show a major influence on the regulatory potential of the *BCKDK*-SVA variants. N1n3, no statistical testing performed. **C**, Analysis of eQTL data in substantia nigra for rs14235 for genes within the LD block with $r^2 > 0.8$. Normalized expression shown. Error bars (a-b): SE. P = minimal promoter, LU = luciferase gene.



Supplemental Figure S6: SVA deletion does not alter epigenome and nearby gene expression. A, Overview of locus. LD block shown in grey box, location of SVA removed by CRISPR-Cas9 KO shown in red. **B-C**, H3K4me3, H3K27ac and RNA-seq shown for *BIN1* (**B**) and *CYP27C1* (**C**). Mean of three replicates shown.



Supplemental Figure S7: Cortical organoid culture overview and SVA KO confirmation. **A**, Overview of cortical organoid culture protocol of SVA KO hESCs. **B-D**, Location of CRISPR-Cas9 gRNAs and deleted region shown, with PCR amplicons for confirmation of KOs.



Supplemental Figure S8: Representation of genes in *BCKDK* locus following SVA KO. **A-B**, DESeq2 results of EV and *BCKDK*-SVA KO cortical organoids shown for genes within a 1,000 kb window upstream (**A**) and 1,000 kb downstream (**B**) of the deleted SVA. Adjusted *P*-values shown on the right of genes. Green: $P < 0.01$, black: $P > 0.01$, red: no *P*-value.

Methods

Cell culture and SVA KO

hESCs were cultured and treated as described previously (Haring et al. 2021) to generate KOs of the SVAs located at Chr2:127,118,846-127,120,589, Chr16:31,103,565-31,105,709 and Chr6:47,505,039-47,506,780 (GRCh38) using the CRISPR-Cas9 system (**Supplemental Figure S7B-D**). Cortical organoids were grown based on the methods described by (Eiraku et al. 2008). See **Supplemental Material** for details and full methods.

ChIP-seq

ChIP

ChIP on EP300 was performed as described previously (Jacobs et al. 2014), with an excess of EP300 antibody (C-20; sc-585 X; Lot # B0211 and Lot # E2610). For H3K27ac and H3K4me3 data, ChIP was based on (Vermunt et al. 2014), with 5 µg H3K27ac (abcam ab4729, lot #GR3303561-2) and 6 µg H3K4me3 (millipore 07-473, lot #3394198) per sample. ChIP-seq data analysis on EP300 was performed as described previously (Jacobs et al. 2014). For full procedure, see **Supplemental Material**.

Library preparation and sequencing

Paired-end indexed ChIP DNA libraries were prepared using Illumina TruSeq ChIP Sample Preparation Kit according to the guidelines, with minor exceptions (see **Supplemental Material**). For H3K27ac and H3K4me3 2 × 75 bp paired-end sequencing was performed by MAD: Dutch Genomics Service & Support Provider of the University of Amsterdam using a NextSeq 550 Illumina sequencer. EP300 data were sequenced on an Illumina HiSeq 2000 sequencing device.

Enrichment TE classes

EP300 summit and peak files of replicates were merged and only peaks of which the summit overlapped with a repeat were kept to increase specificity. These 19,030 peaks were used as input for the TE-analysis pipeline (https://github.com/4ureliek/TEanalysis) (Kapusta et al. 2013) to calculate repeat classes significantly enriched for EP300. For the top 20 enriched elements ($\log_2$(observed number of elements overlapping with a EP300 peak / expected)), the percentage of elements overlapping with a peak was reported (obs_hits / nb_of_TE_in_genome). See **Supplemental Table S1** for output.

RNA-seq

RNA isolation, library preparation and sequencing

Between 6 and 9 organoids were collected on day 35 after start of differentiation, rinsed in medium, homogenized in 400 µl TRIzol (Invitrogen) by pipetting, and stored at -80°C for later use. RNA was isolated according to manufacturer's protocol

with DNase I treatment (Sigma-Aldrich) and cleaned-up using the RNA Clean & Concentrator-5 kit (Zymo Research). Libraries were generated with the TruSeq Stranded Total RNA Library Prep (Illumina) kit, and 2 × 75 bp paired-end reads were sequenced by MAD: Dutch Genomics Service & Support Provider of the University of Amsterdam using a NextSeq 550 Illumina sequencer.

RNA CaptureSeq

Samples of *CD2AP*-SVA KO and unedited control were used for RNA CaptureSeq. RNA isolation of cortical organoids and library preparation were performed as described previously. Enrichment of targeted transcripts was performed with biotinylated probes from myBaits Custom RNA-seq following the manufacturer's standard protocol v.5.00 (Arbor biosciences, Ref #200320-91). The cDNA libraries were pooled and 100 ng was used for capture. After capture, samples were amplified with 14 PCR cycles and purified with 1× AMPure XP beads. The sequencing was performed as described previously.

Data analysis

Data were analysed on the public Freiburg Galaxy server (Afgan et al. 2018; Goecks et al. 2010) (usegalaxy.eu). Reads were trimmed using Trimmomatic (Bolger, Lohse, and Usadel 2014) version 0.36.5 for paired-end reads, removing adapters (ILLUMINACLIP TruSeq3 paired-end), cutting when average quality per base in a 4-base sliding window was below 20, and dropping reads below a length of 30. Reads were mapped using HISAT2 (Kim et al. 2019) (Galaxy Version 2.1.0+galaxy5) against the built-in hg38. featureCounts (Y. Liao, Smyth, and Shi 2014) (Galaxy Version 1.6.4+galaxy2) was used to assign reads to NCBI RefSeq hg38 features with -p, -d 75 -D 900 -B -C. Output was used for DESeq2 (Love, Huber, and Anders 2014) (Galaxy Version 2.11.40.6+galaxy1) using default settings. bamCoverage (Galaxy Version 3.3.2.0.0) from the deepTools2 package (Ramírez et al. 2016) was used to generate coverage tracks (bin size 1) and scale these with a scaling factor based on the number of uniquely assigned reads from featureCounts. Scaled coverage tracks were merged using wiggletools (Zerbino et al. 2014) mean, and wig files transformed to bigWig files using the wigToBigWig script (http://hgdownload.soe.ucsc.edu/admin/exe/). See **Supplemental Table S3** for DESeq2 output.

Differential exon usage analysis

To visualize differential exon usage (**Figure 6D, Supplemental Figure S6B-C**), raw RNA-seq data were pre-processed and analysed with the RNA-seq analysis pipeline from tijsbliek, Galland, and Chouaref (2018) (https://github.com/KoesGroup/Snakemake_hisat-DESeq), with the following modification of the featureCounts rule using 'exon' as feature type: Snakefile, line 194 featureCounts -p -t exon -g exon_id -T 8 -F GTF -O -M -a {input.gff} -o {output} {input.bams}. Preprocessed reads were mapped to hg38 and reads were assigned to NCBI RefSeq hg38 features. The read

count files were used as input for differential exon usage analysis. From here on analysis was performed in R (version 3.4.1) using the packages dplyr (Wickham et al. 2023) (version 1.0.2), reshape2 (Wickham 2007) (version 1.4.4) and tidyr (Wickham, Vaughan, et al. 2024) (version 1.1.2). The NCBI RefSeq hg38 annotation file was modified to contain only exon information and filtered based on canonical transcript IDs and coding region. Read counts file was merged with prefiltered annotation file through a translation table and exon counts normalized to transcript per million. Data were displayed using ggplot2 (Wickham, Chang, et al. 2024) (version 3.3.3).

RNA CaptureSeq data analysis

RNA CaptureSeq data were pre-processed and analysed with the RNA-seq analysis pipeline from bliek et al. (2021) (https://github.com/BleekerLab/snakemake_rnaseq). Pre-processed reads were mapped to hg38, allowing up to 20 multimappers during the alignment. Reads were assigned to Ensembl V104, counting features for both transcripts and exons. From here on analysis was performed in R (version 4.1.0) using the packages tidyverse (Wickham et al. 2019) (version 1.3.1), reshape2 (Wickham 2007) (version 1.4.4) and plyr (Wickham 2011) (version 1.8.6). The read counts file was used as input for differential exon usage analysis with DESeq2 (Love, Huber, and Anders 2014) (version 1.32.0). To correct for variability in control loci presumably unaffected by SVA KO, size factor was estimated based on all captured genes (**Supplemental Table S1**) excluding genes located in the *CD2AP* locus (**Figure 7A**). See **Supplemental Table S4** for DESeq2 output.

Structural variation analysis

Structural variant data were retrieved from Audano et al. (2019) and alternate loci were removed. For each repeat class of interest, regions with one or both ends outside the repeat that overlapped > 95% with the total repeat length were filtered out, to exclude full insertions from the analysis. Transposable elements that showed EP300 (enhancer) marks in cortical organoids were selected and the coordinates were extracted from RepeatMasker (UCSC Genome browser hg38). The alternate haplotypes were removed from each TE subclass and distribution plots were used to select a minimum length per TE subclass (**Supplemental Table S1**), which was used as criteria to filter the elements. All analyses were performed using GenomicRanges package version 1.36.1 (10.18129/B9.bioc.GenomicRanges) (Lawrence et al. 2013) in R version 3.6.0 (2019-04-26) (R Core Team 2019). For script, see Rcode_SVsinTEs in **Supplemental Code**. Since the annotation of SVAs in RepeatMasker for hg38 includes many SVA fragments near one another, we merged these using the join_sva.py code (see **Supplemental Code**). For **Figure 1D**, structural variation locations were plotted at SVAs using data of Audano et al. (2019) and the deepTools (Ramírez et al. 2016) package at the Galaxy platform (Afgan et al. 2018; Goecks et al. 2010). To include hexamer repeats not annotated in the RepeatMasker (hg38 assembly) as SVA, simple repeats at the 5' side of SVAs with a size over 1,000 bp were merged with the SVA if they were within 20 bp of each other using BEDTools (Quinlan and

Hall 2010) merge. The structural variation data were filtered as described previously. Insertions were transformed into deletions for clarity of overlap (insertions were represented as only 1 bp overlap, which were not clearly visible in the heatmap). computeMatrix (Galaxy Version 3.3.2.0.0) was used to prepare the data for profile plotting with the following settings: regions were scaled to the mean size of SVAs containing structural variation (1,908 bp), --binSize 1, --sortRegions descend, --sortUsing mean, --missingDataAsZero True. Heatmap and coverage plots were generated using plotHeatmap (Galaxy Version 3.3.2.0.1) for only SV-SVAs (987). For **Figure 1E**, sizes of structural variations overlapping with SVAs were plotted in R (R Core Team 2019) using ggplot2 (Wickham, Chang, et al. 2024) geom_freqpoly() with 60 bins. The lines from 0 to 50 and -50 to 0 bp were manually removed, as no structural variations with these sizes were present in the data. For **Figure 1G-H**, BLAST (Altschul et al. 1990) was used to search for the SVAs of interest plus 500 bp flanks in the depicted genomes (Ebert et al. 2021), and MUSCLE (Unipro UGENE) (Okonechnikov et al. 2012) was used for alignment to interpret the results. SVA regions were determined based on the repeat browser (Fernandes et al. 2020). Manually curated results were visualized using Adobe Illustrator CC (Adobe Inc.). See **Supplemental Table S1** for sequences and information about data. Analysis of intragenic SVAs was performed on the public European Galaxy server (usegalaxy.eu). The hg38 NCBI RefSeq GTF file was filtered for transcripts, converted to BED12 using Convert GTF to BED12 (Galaxy Version 357), sorted using BEDTools (Quinlan and Hall 2010) SortBED (Galaxy Version 2.29.2) and overlapped with SV- and non-SV-SVAs using BEDTools (Quinlan and Hall 2010) ClosestBed (Galaxy Version 2.29.2). TSSs were generated from the hg38 NCBI RefSeq transcript GTF file and the GenomicRanges package (version 1.44.0, (10.18129/B9.bioc.GenomicRanges) was used to determine locations relative to TSS in R (version 4.1.0) (R Core Team 2019). See **Supplemental Table S1** for information. Differences in distribution of SV- and non-SV-SVAs were tested using a chi-square test of independence. For distance to TSS: $\chi2(4, N = 2,154) = 7.4468$, $P = 0.1141$. For intragenic vs non-intragenic: $\chi2 (1, N = 2,154) = 1.10$, $P = 0.29$. For number of SV-SVAs within 50kb of TSS: $\chi2 (1, N = 987) = 172.3$, $P < 2.2 \times 10^{-16}$. For number of SVAs in AD/PD LD blocks: PD (23/8.8, $\chi2$ (df: 1) = 23.195, $P < .00001$), ASD (38/17.8, $\chi2$ (df: 1) = 23.057, $P < .00001$), bipolar disorder (36/15.6, $\chi2$ (df: 1) = 26.939, $P < .00001$), schizophrenia (69/43.6, $\chi2$ (df: 1) = 15.018, $P = .000106$), MS (19/10.2, $\chi2$ (df: 1) = 7.691, $P = .00555$), AD (13/7.7, $\chi2$ (df: 1) = 3.696, $P = .0545$), neuroblastoma (1/0.6, $\chi2$ (df: 1) = 0.296, $P = .586$), and ALS (1/0.8, $\chi2$ (df: 1) = 0.0276, $P = .868$).

Genome-wide analysis of structural variation within TEs

Structural variations in SVAs were additionally analysed by aligning reference SVA sequences with SVAs identified in WGS files (GCA_002180035.3_HG00514_prelim_3.0_genomic.fna, HG00733_prelim_1.0, HG01352_prelim_2.1, HG02059_prelim_1.0, NA19240_prelim_3.0), using the te-polymorphisms-analysis.py script (see **Supplemental Code**). As this analysis included extensive manual interpretation

of output, only SVAs on Chromosome 1 were analysed in five individuals to confirm high structural variability in SVAs. In short, reference sequence and 500 bp flanking sequences were retrieved from hg38 using the Bio.SeqIO package from Biopython (Cock et al. 2009). lastdb (Kiełbasa et al. 2011) was used to build an index of the WGS files. SVA-flanking sequences with ≥ 90% similarity to reference were identified in the sequencing contigs using lastal (Kiełbasa et al. 2011), and the sequence in between the flanks with the highest similarity was stored as SVA sequence. Pairwise sequence alignment of reference SVA sequence and SVA sequence identified in contigs was performed using EMBOSS Needle or Stretcher (Myers and Miller 1988) to determine structural variation relative to the reference genome.

LD block analysis

Linkage disequilibrium (LD) blocks (threshold $r^2 \geq 0.8$) were determined with the SNiPA Block Annotation tool ((Arnold et al. 2015), https://snipa.helmholtz-muenchen.de/) using the 1000 Genomes Project (Phase 3/Version 5, European population, GRCh37 assembly, Ensembl 87) data for SNPs from the GWAS Catalog ($P \leq 10^{-8}$) (MacArthur et al. 2017). SNPs for which no LD blocks could be calculated or that were not in LD were filtered out, and duplicate blocks were removed from the list. To minimize bias in calculating the percentage of LD blocks overlapping with an SVA, blocks completely within other blocks were removed. LD blocks were annotated for hg19, and overlapped with the hg19 SVA annotation (see **Supplemental Table S1**) using the UCSC Genome Browser Table Browser tool. See **Supplemental Table S5** for downloaded files and information about LD blocks. Locus overview tracks were plotted using pyGenomeTracks (Lopez-Delisle et al. 2020; Ramírez et al. 2018) (https://pypi.org/project/pyGenomeTracks/) and regional plots were made using LocusZoom (AD) ((Pruim et al. 2010), https://my.locuszoom.org/) or the PD GWAS Locus Browser ((Grenn et al. 2020), https://pdgenetics.shinyapps.io/GWASBrowser/). Data were used from de Rojas et al. (2021) and Nalls et al. (2019). Due to plotting restrictions of LocusZoom, infinite *P*-values ($P = 0$) present in the Rojas et al. (2021) meta-analysis for the *APOE* locus were transformed to 1×10^{-320}. Gene annotations were derived from NCBI RefSeq Select database, assembly GRCh37.

eQTL plots

Violin plots were generated for described regions and SNPs using the GTEx eQTL Dashboard tool (https://gtexportal.org/home/eqtlDashboardPage) on 05/20/2020 for the cortex and substantia nigra.

Pairwise Linkage Disequilibrium

Heatmap matrices were generated using the interactive LDmatrix tool of the LDlink suite ((Machiela and Chanock 2015), https://ldlink.nci.nih.gov/). Data from the European population of the 1000 Genomes Project (Phase 3/Version 5) were

used to match Caucasian gDNA samples. See **Supplemental Table S1** for SNPs included in image.

PCR

SVA and SNP amplification

SVAs were amplified from human gDNA samples of patients diagnosed with neurodegenerative disease, which were a kind gift from Coriell Biorepository (DNA panels NDPT088, NDPT087, NDPT083 from the NINDS Repository), and gDNA from a cohort of cognitively healthy centenarians. See **Supplemental Material** for details and full methods. For primer details, see **Supplemental Table S6**; for details on identified variants for *BCKDK*- and *BIN1*-SVA, see **Supplemental Table S1** and **Supplemental File S1**.

Luciferase assay

Luciferase assays using SVAs cloned upstream of a luciferase reporter were performed in mouse ESCs as described previously (Jacobs et al. 2014). Details of primers used for cloning are provided in **Supplemental Table S6.** A detailed description and full methods are provided in **Supplemental Material.**

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE167409. For scripts, see **Supplemental Codes**. Plasmids and cell lines are available on request to the corresponding author.

Acknowledgements

This work was supported by an HFSP Career Development Award (CDA00030/2016C) to F.M.J Jacobs and an ERC starting grant (ERC-2016-stG-716035) to F.M.J. Jacobs and a grant from Alzheimer Nederland (WE.03-2018-07) to F.M.J. Jacobs, H. Holstege and M.J.T. Reinders. We thank Lindsay Payer and Kathleen Burns for helpful discussions and advice about the project; Evan Eichler, Tobias Marschall, Peter Audano and Marc Bonder for helpful discussion about SV-eQTL analysis; the NINDS Human Genetics Resource Center for gDNA; Gonzalo Congrains Sotomayor for technical assistance with FACS; Cindy Wagemans for technical support during gDNA isolation of SVA KO lines; Sophie Imhof and Elias Brandorff for providing the BIN1-SVA reference plasmid; Sol Katzman for EP300 ChIP-data processing; Wim de Leeuw for technical support with bioinformatics; MAD: Dutch Genomics Service & Support Provider of the University of Amsterdam for sequencing; and the Evolutionary Neurogenomics Group and others at SILS for helpful discussions.

Author contributions

F.M.J.J. and E.J.v.B. conceptualized the project. F.M.J.J., E.J.v.B., R.L.F.P.G., M.L. and A.D.E. conceived or designed elements of the study. E.J.v.B. and R.L.F.P.G. performed SV-TE/SVA analysis. F.M.J.J. and D.H. generated EP300 data. E.J.v.B., M.L., F.M.J.J. and A.D.E. were involved in the BLAST SV-SVA analysis. E.J.v.B., J.P., P.F.R., A.S. and F.M.J.J. performed PCR and SNP genotyping and generated SVA-plasmids. E.J.v.B., R.L.F.P.G., I.C. and E.R.B. were involved in the LD-block analysis and visualization. E.J.v.B., R.L.F.P.G. and I.C. were involved in SVA KO experiments. R.L.F.P.G. performed RNA CaptureSeq experiment with support of J.L.R. and F.T.G.W. for the analysis. E.J.v.B., R.L.F.P.G. and C.M. performed the statistical analyses. F.M.J.J., M.J.T.R. and H.H. acquired funding for the study. E.J.v.B, R.L.F.P.G., F.M.J.J. and C.M. wrote the manuscript with contributions of M.L., A.D.E., M.J.T.R. and H.H.. F.M.J.J. supervised the study.

3

Functional assessment of human-specific SVAs in AD-associated genes *PICALM* and *PSEN1*

Rita LFP Guimaraes[1,2], Marcel JT Reinders[3], Henne Holstege[2,3,4], Frank MJ Jacobs[1,4$]

1 Evolutionary Neurogenomics, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands
2 Section Genomics of Neurodegenerative Diseases and Aging, Department of Clinical Genetics, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands
3 Delft Bioinformatics Lab, Delft University of Technology, The Netherland
4 Amsterdam Neuroscience, Complex Trait Genetics, University of Amsterdam, Amsterdam, The Netherlands

$ corresponding author

Abstract

Alzheimer's disease (AD), a major cause of dementia, presents in early-onset (EOAD) and late-onset (LOAD) forms. While EOAD is associated with highly penetrant mutations in Aβ-related genes (e.g., *APP*, *PSEN1*, *PSEN2*), LOAD involves polygenic factors (~84 genes). This study explores the role of structurally variable SINE-VNTR-*Alu* (SV-SVA) elements in the *PICALM* and *PSEN1* genes, using the CRISPR-Cas9 genome editing system. From genetically modified embryonic stem cells (ESCs) we generated cortical organoids carrying genetic deletions of the SVA elements in the *PICALM* and *PSEN1* loci to address the regulatory properties of the SVA on neuronal gene expression. Gene expression was profiled using RNA CaptureSeq, aimed to acquire high coverage transcriptome data of a pre-specified selection of genes. Our analyses revealed no significant gene expression differences for the *PICALM* gene or nearby genes in the organoids where the *PICALM*-SVA was genetically removed. For *PSEN1*, while overall transcript levels did not show significant differences upon SVA removal, *de novo* transcript assembly of *PSEN1* transcripts, coupled with differential expression analysis revealed elevated transcript isoform levels associated with nonsense mediated decay in the *PSEN1*-SVA KO condition. These preliminary findings underscore the role of some SVA elements as possible structurally variable genetic modifiers of AD-associated genes, providing insights into the complex genetic architecture of AD.

Introduction

Alzheimer's disease (AD) is the most common cause of dementia worldwide. The neurodegenerative disease is characterised by the deposition of amyloid-β (Aβ) peptide in the extracellular space, leading to neurotoxicity and subsequent cell death (Duyckaerts, Delatour, and Potier 2009). Typically, AD is categorised into early-onset (EOAD) and late-onset (LOAD), based on the age of onset, set to 65 years old. EOAD manifests symptoms at or before 65, caused by highly penetrant mutations in genes involved in the production of Aβ, namely *amyloid precursor protein* (*APP*), *presenilin 1* (*PSEN1*), and *presenilin 2* (*PSEN2*) (Andrade-Guerrero et al. 2023; Ayodele et al. 2021; Wingo et al. 2012). Notably, mutations in these genes are not always linked to autosomal dominant transmission, as there are risk-modifiers or non-pathogenic variants as well. Conversely, LOAD is highly polygenic, encompassing approximately ~84 genes with numerous low/intermediate penetrance variants identified through genome-wide association studies (GWAS) (Bellenguez et al. 2022).

One of the main susceptibility genes for LOAD is *PICALM*, encoding for a *phosphatidylinositol binding clathrin assembly protein* (Ando et al. 2022). PICALM is a central player in the Aβ pathway, involved in APP processing and Aβ generation and clearance (Ando et al. 2022; Xu, Tan, and Yu 2015; Zhao et al. 2015). It is known that lower gene expression of *PICALM* reduces Aβ clearance, whilst higher levels of PICALM enhances Aβ clearance (Zhao et al. 2015). While *PICALM* harbours numerous intergenic risk single nucleotide polymorphisms (SNPs) within its locus, no direct causative genetic variant has been identified. Interestingly, these risk SNPs are in close proximity to a structurally variable SINE-VNTR-*Alu* (SV-SVA) element (as detailed in **chapter 1**) (**Figure 1A**). SVA elements, a hominid-specific family of mobile elements still active in the human genome (Cordaux and Batzer 2009; H. Wang et al. 2005) and possess gene-regulatory properties (Jacobs et al. 2014; Savage et al. 2013, 2014). The *PICALM*-SVA lies upstream of the gene, 40kb away from its transcription start site (TSS) (**Figure 1A**), suggesting a potential role as a genetic driver of the disease. Previous work has demonstrated that the SVA in *PICALM*, being surrounded by risk SNPs, is polymorphic in size in the human population (Audano et al. 2019) (**Figure 1B**). The configuration of the SVA in the reference genome (GRCh38) is 2.353 bp and other *PICALM*-SVA polymorphisms exist with insertions of 63, 80 and 356 bp, and a deletion of 175 bp. These variations are located in the central segment of the SVA, in the VNTR region.

PSEN1 gene, encoding *presenilin 1*, is one of the main causative factors for EOAD, together with *PSEN2* and *APP*. Mutations in these genes follow an autosomal dominant inheritance pattern and are responsible for less than 1% of all AD cases (Andrade-Guerrero et al. 2023; Cacace, Sleegers, and Van Broeckhoven 2016; Fernández et al. 2025; Petit et al. 2022). *PSEN1* alone holds 215 mutations and carriers start to develop symptoms between 30 and 50 years old (Cacace, Sleegers,

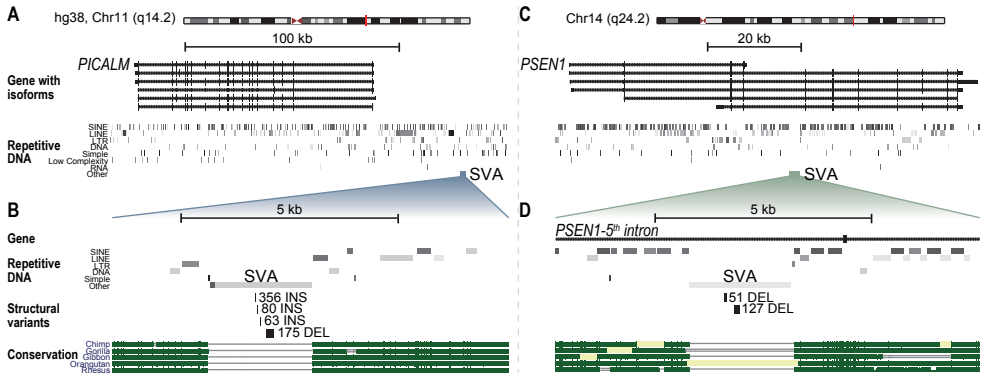


Figure 1: *PICALM*- and *PSEN1*-SVA genomic location. A-D, Architecture of *PICALM* (**A-B**) and *PSEN1* (**C-D**) loci, including repetitive DNA elements from RepeatMasker. SVAs are highlighted in blue (**A**) and green (**C**), respectively. Gene annotation from GENCODE v36. **B,D,** Zoom-in on SVAs in *PICALM* (**B**) and *PSEN1* (**D**) loci, also showing structural variations from Audano et al. (2019) and a primate conservation track, pale yellow colouring reflects lack of sequence. **A-D,** Adapted from the UCSC Genome Browser.

and Van Broeckhoven 2016; De Deyn and Sleegers 2025). PSEN1 serves as the catalytic subunit of γ-secretase complex, involved in APP processing and Aβ generation. Consequently, pathological mutations in *PSEN1* often result in a gain-of-function and over-production of Aβ (De Strooper et al. 1998; Vetrivel et al. 2006). Similarly to the *PICALM* gene, *PSEN1* also contains an SV-SVA element in its locus (**Figure 1C**). Here, however, the SVA is located within the gene, in its fifth intron, with a secondary TSS just downstream of it. Given that intronic SVAs can influence splicing patterns (Aneichyk et al. 2018; Pfaff, Singleton, and Kõks 2022), changes in its length could potentially impact the transcription of this essential gene, thereby contributing to the pathogenesis of the disease. Furthermore, the intronic SVA in *PSEN1* also exists in multiple forms in the human population (Audano et al. 2019). The configuration of the SVA in the reference genome (GRCh38) is 2.343 bp and part of the population has *PSEN1*-SVAs that are slightly shorter, carrying deletions of 51 or 127 bp, present also in the VNTR region (**Figure 1D**).

Both SVAs in *PICALM* and *PSEN1*, two of the most relevant genes associated with Alzheimer's disease, are human-specific, as they are not present in any other primate species (**Figure 1B, D**), meaning that they were inserted in the human genome in the last 6 million years. Since SVAs have been co-opted in gene-regulatory networks, the structural variation in these SVAs, might contribute to inter-individual differences in gene expression and potentially modify disease risk. Indeed, both genes show multiple isoforms arising from alternative splicing, which might suggest a potential role of this element, depending on specific cell type regulation.

In this study we aimed to investigate the functional consequences of SV-SVAs located either upstream or within the key AD-associated genes, specifically *PICALM* and *PSEN1* genes. By employing the CRISPR-Cas9 genome editing system, we engineered deletions of the *PICALM*- and *PSEN1*-SVA in human embryonic stem cells (hESCs) and analysed their impact on gene expression in cortical organoids, representing a human-neuronal context. For *PSEN1*-SVA we perform an exon-level expression analysis to investigate whether specific exons are differentially expressed, as is the case for intronic SVAs in *CD2AP* gene (as detailed in **chapter 2**) or in *TAF1* gene, that leads to X-linked dystonia-parkinsonism (Bragg et al. 2017; Rosenkrantz et al. 2024).

Results

To investigate the functional impact of SVAs in AD-associated loci, we engineered human embryonic stem cell (hESCs) lines carrying genetic deletions of the *PICALM*-SVA and *PSEN1*-SVA using CRISPR-Cas9 genome editing system (Ran et al. 2013) (**Figure 2, 3A-B**). To study the effect of the SVA in a neuronal context, the SVA KO hESC lines were used to generate cortical organoids. We performed targeted mRNA enrichment prior to RNA-seq (RNA CaptureSeq) to specifically enrich for transcripts derived from the *PICALM* and *PSEN1* loci, together with other unrelated control genes. This allowed us to track the expression levels of each exon separately. This is important in particular for *PSEN1*, because intronic SVAs may differentially influence the expression level of exons up- and downstream of the SVA. Furthermore, this

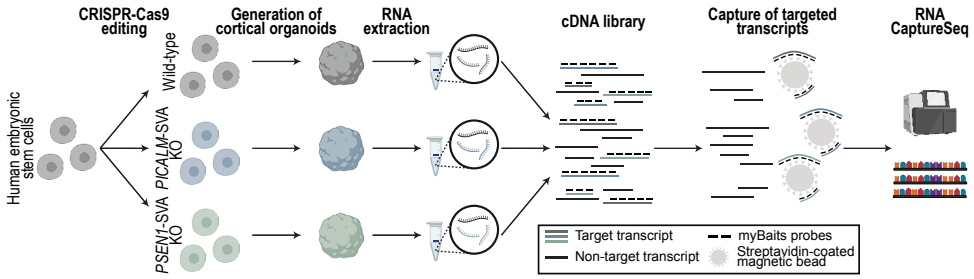
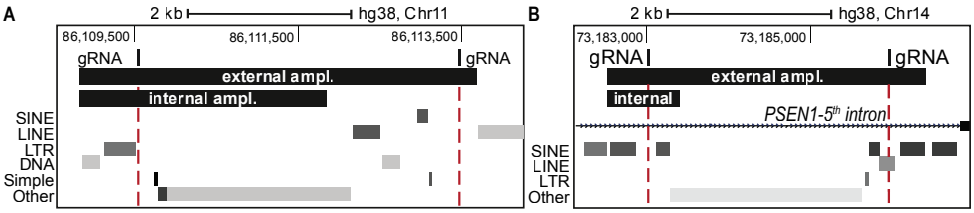


Figure 2: Workflow of functional analysis of SVAs in two AD-associated genes. Genomic deletion of SVAs in human embryonic stem cells with CRISPR-Cas9 gene editing system, generation of cortical organoids with the SVA KO cell lines and RNA CaptureSeq.



C

Targeted transcripts	ΔCt	$2^{\Delta Ct}$	
PSEN1	-10.14	1128.35	Enriched transcripts
CD2AP	-11.16	2290.85	
PAX6	-10.62	1577.40	
SMN	2.73	0.15	Depleted transcripts
GAPDH	3.46	0.09	
DNMT3B	3.26	0.10	

D

Targeted transcripts		Depth	Coverage
WT	rep1	142,270,940	22,609.62
	rep2	144,492,889	22,962.73
	rep3	109,010,010	17,323.81
PICALM-SVA KO	rep1	154,052,761	24,481.98
	rep2	147,324,641	23,412.75
	rep3	105,684,768	16,795.36
PSEN1-SVA KO	rep1	170,759,641	27,137.03
	rep2	210,898,009	33,515.79
	rep3	150,013,755	23,840.10

Figure 3: RNA CaptureSeq of *PICALM*- and *PSEN1*-SVA KO. A-B Location of CRISPR-Cas9 gRNAs and deleted region shown, with PCR amplicons for confirmation of KOs. **C**, Confirmation of enrichment with myBait probes by qPCR. ΔCt calculated with three replicates per sample. **D**, Depth and coverage of targeted transcripts.

RNA CaptureSeq approach increased the overall resolution of the selected gene transcripts, allowing for a more accurate assessment of the consequences of removal of the SVA for *PICALM* and *PSEN1* gene expression.

Prior to sequencing, enrichment of the targeted transcripts was confirmed with qPCR, by amplifying exons in three target genes and three non-target genes in the wild-type (WT) condition. The targeted transcripts showed a high enrichment, with ΔCt values between -10 and -11 (average Ct of non-target – average Ct of target) (**Figure 3C**). RNA CaptureSeq analysis revealed a high coverage of target transcripts between 16,795 and 33,515 across samples and conditions (**Figure 3D**). All samples showed a high percentage of raw read counts in the targeted transcripts, ranging around 88% for WT and *PICALM*-SVA KO, and 90% for *PSEN1*-SVA KO (**Supplemental Figure S1A**). A principal component analysis based on the raw read counts of targeted genes showed each condition (*PICALM*-SVA KO, *PSEN1*-SVA KO and WT) clustered together with some spread between replicates (**Supplemental Figure S1B**). Mapping statistics can be found on **Supplemental Table S6**.

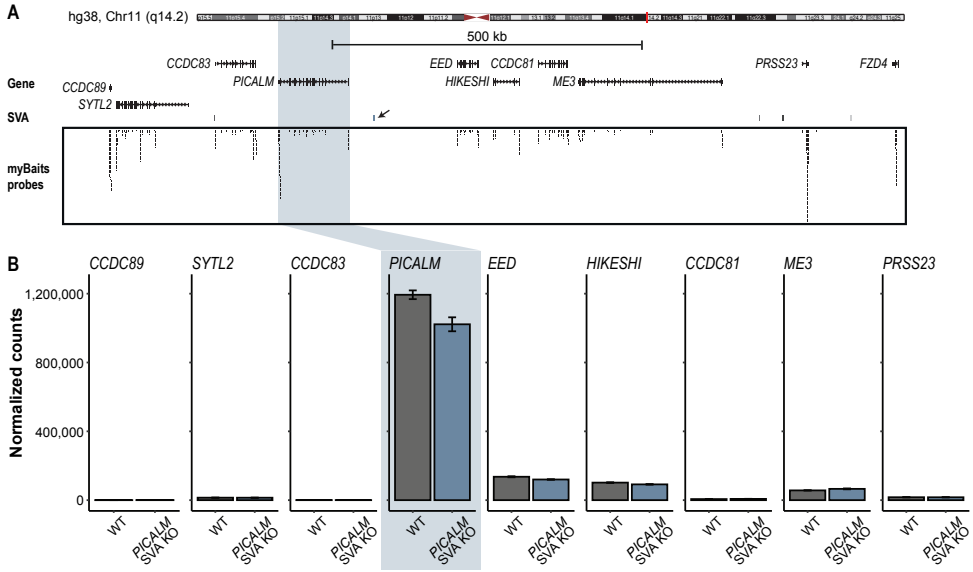


Figure 4: RNA CaptureSeq of *PICALM*-SVA KO. A, Overview of *PICALM* locus, including all genes targeted. Location of SVA removed by CRISPR-Cas9 KO shown in blue, in the promotor region of *PICALM*, indicated with an arrow. Probes from myBaits targeting all exons in the loci. **B,** Normalised mean expression of three replicates of genes *CCDC89*, *SYTL2*, *CCDC83*, *PICALM*, *EED*, *HIKESHI CCDC81*, *ME3* and *PRSS23* in the Wild-type and *PICALM*-SVA KO conditions. Adjusted *P*-value from DESeq2 analysis not significant in all comparisons.

Differential expression analysis comparing *PICALM*-SVA KO cortical organoids to unedited control cortical organoids revealed no statistically significant differential expression levels of genes in the *PICALM* locus (**Figure 4**). Despite both *PICALM* and *EED* genes showing elevated gene expression in the WT condition relatively to the SVA KO (P (*PICALM*) = 7.34 × 10^{-2}; P (*EED*) = 7.20 × 10^{-3}), neither reaches statistical significance after correction. Control genes residing in other genetic loci also showed no overall significant differences, with the exception of *VKORC1* (adjusted P = 2.52 × 10^{-2}), located in a different chromosome (Chr16) (**Supplemental Figure S2, Supplemental Table S8**).

In *PSEN1*-SVA KO cortical organoids, compared to unedited control, differential expression analysis revealed significantly lower expression levels for *RBM25* gene (P = 2.11 × 10^{-6}), located upstream of *PSEN1* (**Figure 5A-B**). *PSEN1* did not exhibit a significant differential gene expression (P = 0.345). However, a differential exon usage analysis revealed two exons, third and fourth exons, upstream of the SVA deletion, with a higher expression in the *PSEN1*-SVA KO condition, compared to unedited controls (**Figure 5C**). We should note that between the *PSEN1*-SVA KO and unedited control conditions, significant differences in expression of unrelated control genes were also observed (**Supplemental Figure S3, Supplemental Table S8**).

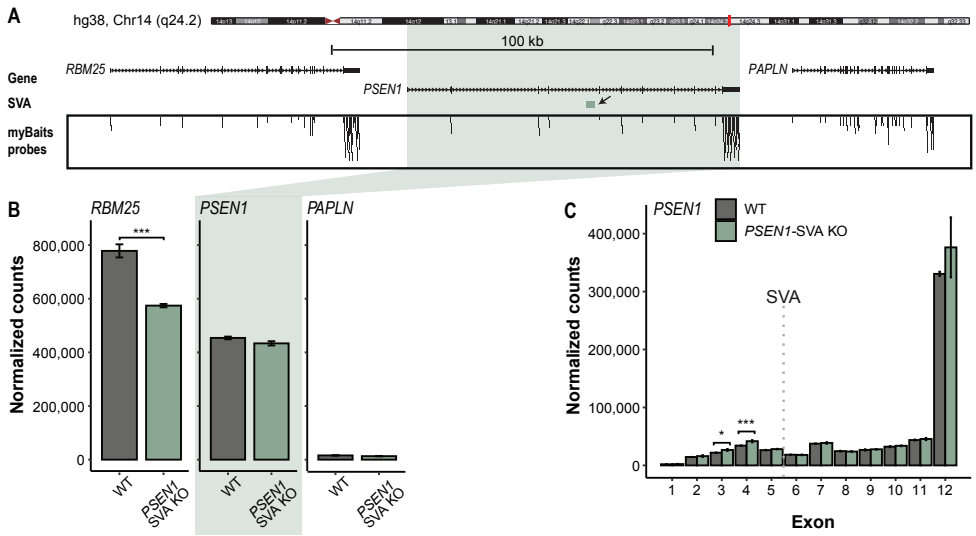


Figure 5: RNA CaptureSeq of *PSEN1*-SVA KO. A, Overview of *PSEN1* locus, including all genes targeted. Location of SVA removed by CRISPR-Cas9 KO shown in green, in the fifth intron of *PSEN1*, indicated with an arrow. Probes from myBaits targeting all exons in the loci. **B,** Normalised mean expression of three replicates of *RBM25*, *PSEN1* and *PAPLN* genes. **C,** Normalised mean expression of three replicates shown per exon of *PSEN1.* Location of SVA indicated with dashed grey line. Adjusted P-value from DESeq2 analysis. *** = P < 0.001, ** = P < 0.01, * = P < 0.05.

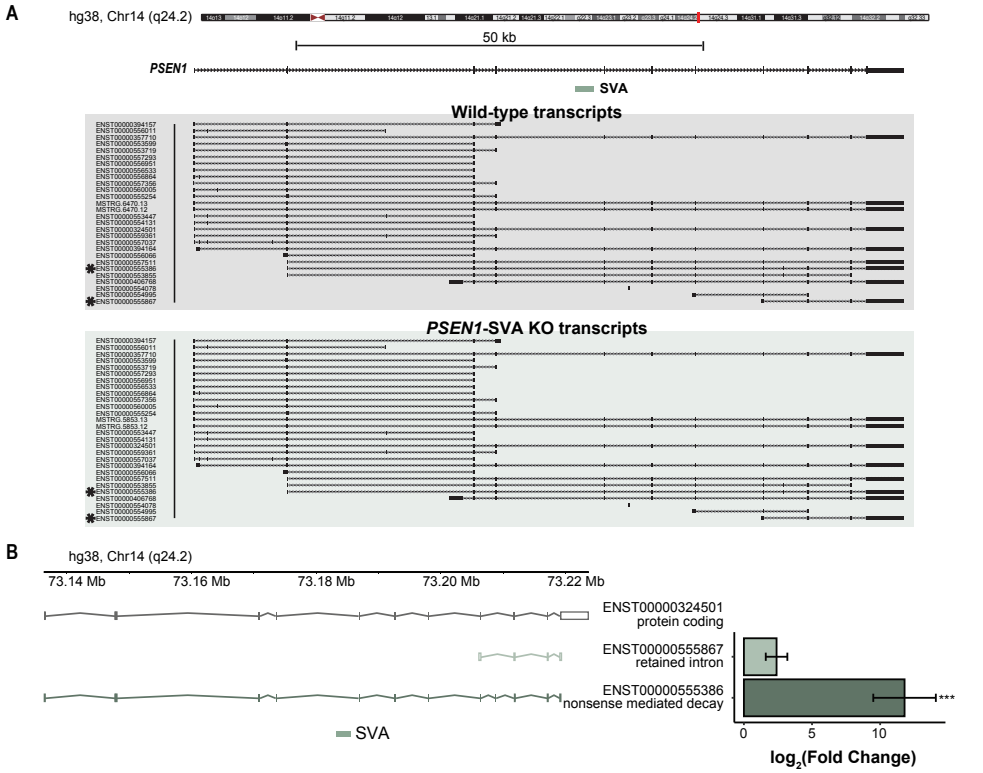


Figure 6: *PSEN1* transcripts from *PSEN1*-SVA KO and wild-type cortical organoids. A, Transcripts assembled *de novo* with StringTie method. * indicates differentially expressed transcripts. **B**, *PSEN1* transcripts: canonical (ENST00000324501) and differentially expressed in *PSEN1*-SVA KO compared to unedited control (ENST00000555867 and ENST00000555386), adapted from the ensembl website (left side), and respective differential expression change, if applicable (right side). Adjusted *P*-value from DESeq2. *** = $P < 0.001$.

To investigate further whether the observed differential exon usage in *PSEN1* gene is caused by alternative transcripts differentially expressed in *PSEN1*-SVA KO compared to unedited controls, we analysed the captured transcripts with StringTie (M. Pertea et al. 2015). StringTie is a computational method that allows for a *de novo* transcript assembly, providing an improved reconstruction of transcripts and expression level estimations. The accuracy of this transcript assembler was evaluated with GFFCompare (G. Pertea and Pertea 2020), which showed that the sensitivity and precision from base level to a locus level is 93.9 and 100%, and 83.4 and 99.7, respectively (**Supplemental Table S11**). StringTie analysis shows no difference in the set of transcript-isoforms that are expressed between the *PSEN1*-SVA KO and unedited control conditions (**Figure 6A**).

DESeq2 analysis was performed to compare transcript counts from StringTie revealed 11.80× higher expression of the ENST00000555386 (Chr14:73,206,153-73,219,396) isoform of *PSEN1* (which includes the third and fourth exons), in the *PSEN1*-SVA KO compared to WT cortical organoids ($P = 1.96 \times 10^{-5}$). Interestingly, this isoform, I-374, has a premature stop codon generating a protein consisting of 374 amino acids (Sahara et al. 1996). Thus, in normal conditions it may be expressed at low levels due to nonsense-mediated decay. A second shorter isoform in *PSEN1*, ENST00000555867 (Chr14:73,136,451-73,219,289), appears to be 2.40× higher expressed in the *PSEN1*-SVA KO condition (**Figure 6B**), but this difference does not reach statistical significance ($P = 8.76 \times 10^{-2}$). This isoform is also not predicted to be translated into any functional protein, as there is intron retention. No other transcripts in the *PSEN1* locus were differentially expressed, including those of *RBM25* gene (**Supplemental Table S11**).

Discussion

Our study aimed to deepen the knowledge about structurally variable SVA elements and their regulatory dynamics in loci associated with both early-onset and late-onset Alzheimer's disease. New SVA insertions within genes are known to create instability, by causing altered splicing through disruptions of splicing sites, adding new cryptic splice sites contained in its sequence in the sense orientation, or the exonization of the SVA element itself (Pfaff, Singleton, and Kõks 2022). Another mechanism is through the formation of highly stable G-quadruplexes within the SVA, which may block the progression of RNA polymerase II, leading to truncated transcripts (Nicoletto et al. 2024; Rosenkrantz et al. 2024). The intronic SVA insertion in the *PSEN1* gene is not a new insertion, but instead it is fixed in the human population. However, inter-individual variability arises in the form of structural variation.

While total *PSEN1* gene expression levels did not change upon the deletion of the intronic SVA, *PSEN1*-SVA KO cells gene displayed elevated expression levels of two *PSEN1* transcript isoforms that were expressed at much lower levels in unedited control organoids. This finding reveals that while the impact of the SVA on the *PSEN1* locus and the expression of *PSEN1* appears to be limited, it may have changed the expression dynamics of some of the isoforms that exist for the *PSEN1* gene. Moreover, expression levels of *RBM25* gene, located upstream of *PSEN1*, were significantly lower in the *PSEN1*-SVA KO cells. *RBM25* (RNA binding motif protein 25) plays a crucial role in the global regulation of alternative splicing (Carlson et al. 2017) and while *RBM25* itself has not been directly linked to AD, other RNA-binding proteins, such as *RBM24* and *RBM8A*, have been implicated in AD pathogenesis. These proteins are involved in maintaining synaptic function and promoting neurogenesis, respectively, highlighting the importance of RNA-binding proteins in AD (Yao et al. 2024; Zhu et al. 2024). Additionally, RBM45 protein is known to aggregate in nuclear stress bodies under chronic stress conditions in neurons and glial cells in neurodegenerative diseases, including AD (M. Collins, Li, and Bowser 2020). The potential regulatory role of the SVA in *PSEN1* and its downstream impact on *RBM25* expression warrant further investigation to gain a deeper understanding of the regulatory mechanism.

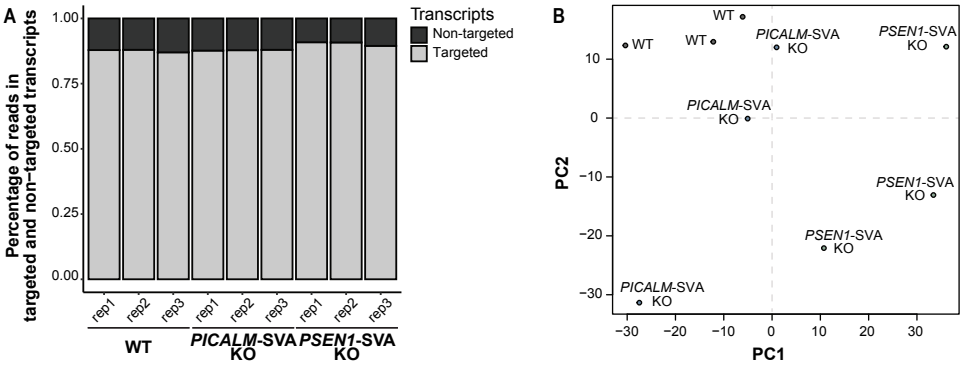
In the *PICALM* gene, genetic deletion of the SVA did not result in significant gene expression differences in cortical organoids. While *PICALM* is prominently expressed in most tissues, it undergoes specific epigenomic regulation in microglial cells (Xiong et al. 2023). Recent research indicates a dynamic regulation of *PICALM*, with down-regulation in early microglial stages of individuals with early AD and up-regulation in later stages (Sun et al. 2023). While five-week-old cortical organoids primarily consist of neural stem cells, cortical neurons, and intermediate progenitors, microglia cells are not yet present. Therefore, transcriptional changes arising from the SVA deletion in the *PICALM* locus may not be captured in our analysis. SVA mediated changes

could potentially be captured in *in vitro* models aimed at generating ESC-derived microglia cells. Future studies should be conducted to further explore the cell-type specific regulatory mechanisms involving SVAs, providing a deeper understanding of the intricate epigenetic dynamics of AD.

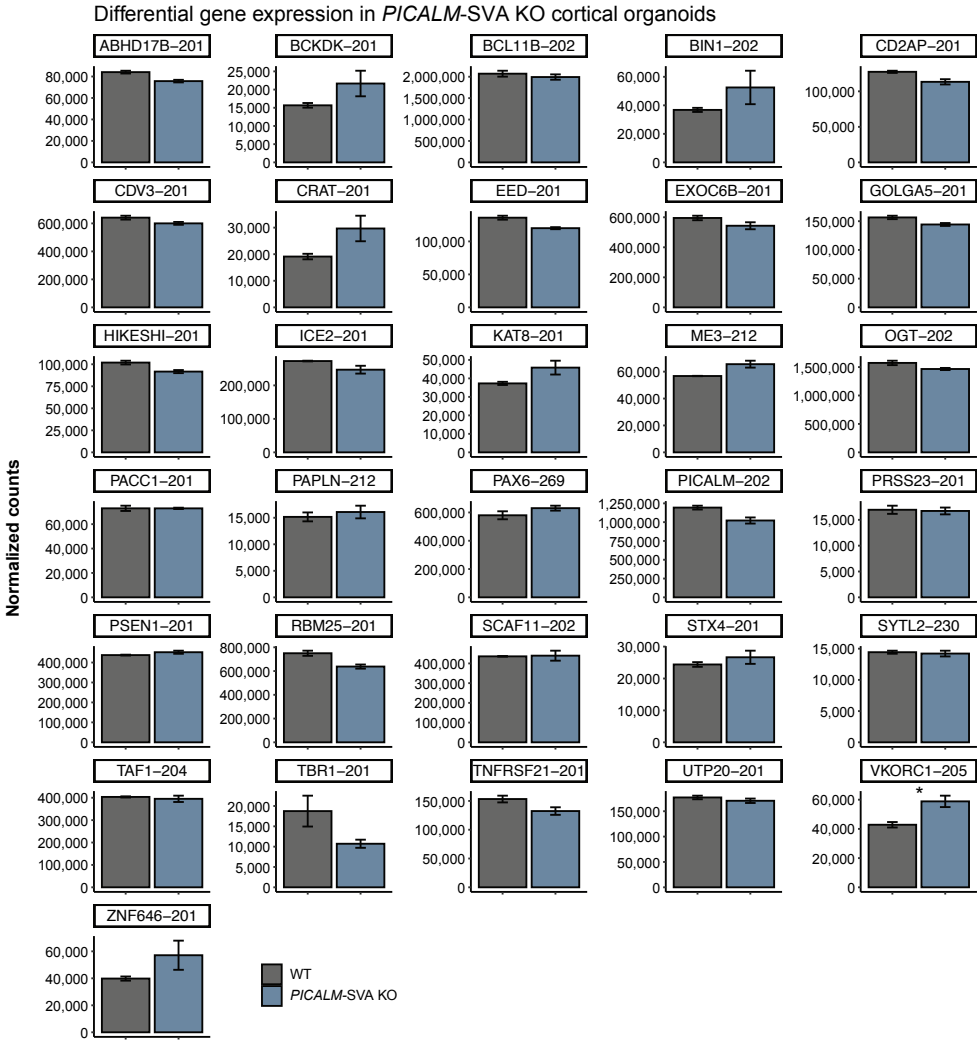
In neurodegenerative conditions like AD, ageing stands out as the primary risk factor, accompanied by a multitude of age-associated epigenetic changes (J.-H. Lee et al. 2020). Mobile elements, among them SVAs, remain repressed in normal conditions but may become activated as part of the ageing process (Guo et al. 2018; Wanhe Li et al. 2013; Salazar et al. 2023). In parallel, gene expression levels of *ZNF91*, a critical repressor of SVA elements belonging to the Krüppel-associated box (KRAB) zinc finger protein family, significantly decreases with age (Rosenkrantz et al. 2024). This observation suggests that the gene-regulatory potential of SVAs in loci associated with neurodegenerative diseases could become more pronounced with advancing age, potentially contributing to the progressive nature of these diseases. Therefore, mimicking the age-related transcriptional changes, for instance for ZNF91, might help uncover the true biological impact of SVAs in AD loci.

In conclusion, this study gathers preliminary evidence of a nuanced involvement of SVAs in the regulation of genes in AD-risk loci. While the differential isoform expression observed in *PSEN1* is notable, the lack of gene expression differences in the *PICALM* gene suggests the need for refining *in vitro* models of AD. Although cortical organoids introduce a neuronal background, it is crucial to consider age-related characteristics and cell-type specifications to enhance the fidelity of these models. In the near future, it will be possible to correlate carriership of a certain SVA with the cell-type specific abundance of entire transcripts identified with single-cell long-read RNA-seq of human tissues, even cell-type specific. Hopefully we can then learn more about the influence of such SVAs on gene transcription/protein regulation.

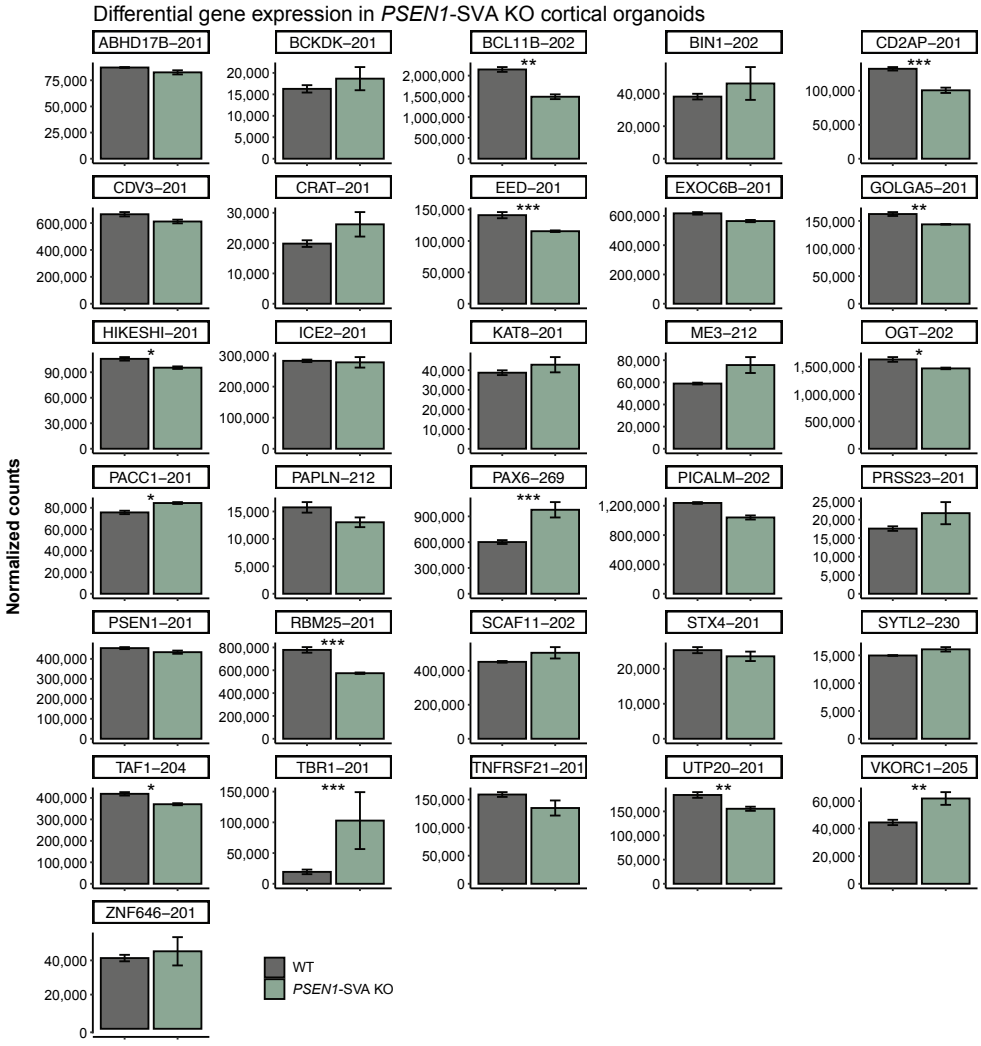
Supplemental Figures



Supplemental Figure S1: Quality control of *PICALM*- and *PSEN1*-SVA KO RNA CaptureSeq. A, Percentage of reads from targeted and non-targeted transcripts across all samples, with three replicates each. **B**, Principal components analysis of raw read counts on targeted transcripts per sample, with three replicates each.



Supplemental Figure S2: Differential gene expression in *PICALM*-SVA KO cortical organoids. Normalised mean expression of three replicates. Adjusted *P*-value from DESeq2 analysis. *** = $P < 0.001$, ** = $P < 0.01$, * = $P < 0.05$.



Supplemental Figure S3: Differential gene expression in *PSEN1*-SVA KO cortical organoids. Normalised mean expression of three replicates. Adjusted *P*-value from DESeq2 analysis. *** = $P < 0.001$, ** = $P < 0.01$, * = $P < 0.05$.

Methods

Cell culture and SVA KO

hESCs were cultured and treated as described previously (Haring et al. 2021)). Plates were coated with 0.1% gelatin (Sigma-Aldrich) for 30 min and washed twice with phosphate buffered saline (PBS). Mitomycin C-treated mouse embryonic fibroblasts (MEFs, GlobalStem) were seeded at a density of approximately 3.17 × 104 cells/cm2 in MEF medium (DMEM-GlutaMAX (Gibco) supplemented with 10% heat inactivated Fetal Bovine Serum (Gibco), 100 U/ml penicillin/100 µg/ml streptomycin (Gibco), 1 mM sodium pyruvate (Gibco)) and incubated at 37°C, 5% CO2. The following day, MEF-covered plates were washed once with PBS before hESC colonies were seeded in hESC medium (DMEM-F12 (Gibco) supplemented with 20% KnockOut Serum Replacement (Gibco), 100 U/ml penicillin/100 µg/ml streptomycin (Gibco), 2 mM GlutaMAX (Gibco), 1× MEM Non-Essential Amino Acids solution (Gibco), 100 µM 2-mercaptoethanol (Gibco)) with 8 ng/ml fresh bFGF (Sigma-Aldrich). Medium was changed daily, until full-grown colonies were passaged manually using a needle (1:4 ratio) or used for cortical organoid differentiation.

CRISPR-Cas9-medidated knock out of SVA elements

Guide RNAs (gRNAs) were designed to KO the SVAs located at Chr11:86,109,539-86,113,492 (*PICALM*-SVA) and Chr14:73,183,017-73,185,961 (*PSEN1*-SVA) (GRCh38) (**Supplemental Figure S1A-B**) using the CRISPR-Cas9 system (Ran *et al.*, 2013). The online tools CRISPOR (Concordet and Haeussler 2018) and CHOPCHOP (Labun et al. 2016) were used to design the gRNA sequences, considering their efficiency and specificity. The oligos containing the gRNA sequences were annealed and cloned into a pX330-U6-Chimeric_BB-CBh-hSpCas9 plasmid (Addgene 42230; hereafter referred to as pX330). For the transfection, hESCs were cultured in matrigel-coated (Corning) dishes in mouse embryonic fibroblasts (MEF)-conditioned medium (MCM), supplemented with 8ng/ml fresh bFGF. One day prior to transfection, cells were plated in a 6-well plate to reach 30-60% confluency on the day of transfection. Cells were incubated 1 hour prior to transfection with 10 ng/ml ROCK inhibitor (Thiazovivin, Sigma-Aldrich). 1.4 µg pX330 plasmid containing the up- and downstream gRNAs (**Supplemental Table S1**) (mixed 1:1), or pX330 plasmid without gRNA sequence inserted, along with 20 ng pCAG-GFP (Addgene #11150), was transfected using Lipofectamine Stem Transfection Reagent (Invitrogen). After 6h, the medium was refreshed with fresh MCM supplemented with bFGF and ROCK inhibitor. 48 hours post-transfection, GFP-positive cells were selected by fluorescence-activated cell sorting (FACS) on a FACS Aria III (BD) with a 100 µm nozzle. 3000 single cells were collected in a MEF-covered 60 mm dish and maintained in MCM + hESC medium (1:1) until colonies were visible, after which they were maintained in hESC medium.

Clonal expansion

Single colonies were manually transferred to matrigel-coated 96-well plates containing 150 µl MCM + 10 ng/ml ROCK inhibitor, 8 ng/ml bFGF, and maintained in MCM + bFGF until ready for passaging. KO and WT clones were identified by gDNA isolation followed by two PCR reactions with internal and external primer sets (**Supplemental Figure S1A-B, Supplemental Table S1).** Clones were considered valid knockouts when a PCR product was observed with primers flanking the deletion region, and no PCR product was observed with primers internal to the deletion region. Expanded cell lines were verified again by repeating gDNA isolation, two separate PCR reactions and sanger sequencing before cryopreservation of hESCs and cortical organoid formation.

Cortical organoids

Cortical organoids were grown based on the methods described by (Eiraku et al. 2008). hESC colonies of 1-2 mm in diameter were lifted with a cell lifter (Corning), manually adjusted to fit the colony to prevent unintentional lifting of MEFs. Before lifting, the medium was changed to the differentiation medium (99% hESC medium supplemented with 1 mM sodium pyruvate (Gibco)). Colonies were transferred with differentiation medium to a 60 mm ultra-low attachment dish (Corning) and embryoid bodies were formed overnight. Medium was replaced the next day for differentiation medium supplemented with 3 µM IWR-1-Endo, 1 µM Dorsomorphin, 10 µM SB-431542 hydrate, and 1 µM Cyclopamine hydrate (day 0 of differentiation). Medium was changed every other day with freshly added inhibitors. On day 3, organoids were placed on a rocker, to prevent fusion of organoids and promote growth. On day 18, medium was changed to Neurobasal/N2 medium (Neurobasal (Gibco) supplemented with 100 U/ml penicillin/100 µg/ml streptomycin (Gibco), 2 mM GlutaMAX (Gibco), 1× N-2 supplement (Gibco)) supplemented with 1 µM Cyclopamine hydrate. From D24 onwards, organoids were cultured in Neurobasal/N2 medium without inhibitors until harvest on D35.

RNA CaptureSeq

Design of capture probes

All probes were designed to target 44 genes (**Supplemental Table S2**), including all known isoforms with 5' and 3' untranslated regions (UTRs). All probes have 80 nucleotides in length, however exon coordinates that were between 65-79 nucleotides were padded with thymines until reaching 80 nucleotides. Probes targeting highly repetitive regions were discarded. All regions have a 3x probe tiling to achieve a high probe coverage which averages the differential performance between probes. To allow identification of new exons or intron retention, the exon-exon junctions were not targeted. A non-transcribed intergenic region was included in the design to identify gDNA contamination (chr6:101280105-101386618, GRCh38). A fasta and bed files can be found in **Supplemental Tables S3, S4**, respectively.

RNA isolation, library preparation and capture

Samples of *PICALM*-SVA KO, *PSEN1*-SVA KO and unedited control were used for RNA CaptureSeq. Between 8 organoids were collected on day 35 after start of differentiation, rinsed in medium, homogenised in 400 µl TRIzol (Invitrogen) by pipetting, and stored at -80°C for later use. RNA was isolated according to the manufacturer's protocol with DNase I treatment (Sigma-Aldrich), and cleaned-up using the RNA Clean & Concentrator-5 kit (Zymo Research). Libraries were generated with the TruSeq Stranded Total RNA Library Prep (Illumina) kit. Enrichment of targeted transcripts was performed with biotinylated probes from myBaits Custom RNA-seq following the manufacturer's standard protocol v.5.00 (Arbor biosciences, Ref #200320-91). The cDNA libraries were pooled and 100 ng was used for capture. After capture, samples were amplified with 14 PCR cycles and purified with 1× AMPure XP beads.

Quantitative PCR and sequencing

To determine enrichment of the targeted transcripts a quantitative PCR (qPCR) was performed using Quantitect SYBR Green (QIAGEN) for three target genes *CD2AP*, *PSEN1* and *PAX6* and three non-targeted transcripts *SMN*, *GAPDH* and *DNMT3B* of the wild-type condition (**Supplemental Table S5, Supplemental Figure S1C**). After confirming the enrichment of targeted transcripts, 2 × 75 bp paired-end reads were sequenced by MAD: Dutch Genomics Service & Support Provider of the University of Amsterdam using a NextSeq 550 Illumina sequencer.

Data analysis

RNA CaptureSeq data was pre-processed and analysed with the RNA-seq analysis pipeline from (bliek et al. 2021) (https://github.com/BleekerLab/snakemake_rnaseq) (see RNA_CaptureSeq in **Supplemental Code**). Pre-processed reads were mapped to hg38, allowing up to 20 multimappers during the alignment. Reads were assigned to Ensembl V104, counting features for both transcripts and exons. Mapping statistics are in **Supplemental Table S6**. From here on analysis was performed in R (version 4.1.0) using the packages tidyverse (version 1.3.1), ggplot2 (version 3.3.5), ggrepel (version 0.9.1), reshape2 (version 1.4.4), plyr (version 1.8.6) and tibble (version 3.1.6). The read counts file was used to calculate the depth, sum of all reads, and coverage of the targeted transcripts. The latter was calculated by multiplying the depth by the average read length (151 bp) and dividing it by the size of the targeted region (**Supplemental Figure S1D**). The percentage of reads from targeted and non-targeted transcripts was also calculated (**Supplemental Figure S1E**). A principal components analysis was performed on the raw read counts of targeted transcripts and the 2 first principal components were plotted (**Supplemental Figure S1F**). Finally, the read counts file, including only the targeted regions, was used as input for differential expression analysis, at both gene and exon levels, with DESeq2 (Love, Huber, and Anders 2014) (version 1.32.0). For the RNA CaptureSeq dataset, to correct for variability in control loci presumably

unaffected by SVA KO, size factor was estimated based on all captured genes excluding genes located in the locus where the SVA was removed from (**Figure 2A, 3A**, **Supplemental Table S2**). See **Supplemental Table S7, S8** for DESeq2 output for differential gene expression in *PICALM*-SVA KO and *PSEN1*-SVA KO, respectively, and **Supplemental Table S9** for DESeq2 output for differential exon expression in *PSEN1*-SVA KO condition. The plots of normalised expression levels are of canonical transcripts only (**Supplemental Table S10**). **Supplemental Figures S2**, **S3** represent an overview of gene expression levels in *PICALM*-SVA KO and *PSEN1*-SVA KO, respectively, whose normalised counts are higher than 10,000.

Alternative transcripts analysis

RNA CaptureSeq data was analysed with StringTie tool (version v2.1.7b) (M. Pertea et al. 2015). Transcripts were assembled from the read alignments obtained with the RNA-seq analysis pipeline from Bliek et al. (2021), with the Ensembl V104 annotation file (-G) and assuming it is a stranded library (-–rf). The resulting GTF files were merged (--merge) to generate a non-redundant set of transcripts present in all the RNA CaptureSeq samples. This file is used as input for the GffCompare tool (version v0.12.6) (G. Pertea and Pertea 2020) to evaluate the accuracy of the transcripts assembled by StringTie (**Supplemental Table S11**). Then the first step is repeated but with this non-redundant set of transcripts as an annotation file (-G) to guide the assembly of transcripts per sample, creating a non-redundant set of transcripts per sample. Finally, each set of transcripts per sample was used as input for the python script prepDE.py3 (https://github.com/gpertea/stringtie/blob/master/prepDE.py3) to create count matrices for genes and transcripts for differential expression analysis with DESeq2. For visualisations purposes, transcripts were merged per condition using bedtools intersect function and uploaded onto the UCSC Genome Browser for a screenshot of *PSEN1* transcripts.

From here on analysis was performed in R (version 4.1.0) using the packages tidyverse (version 1.3.1), ggplot2 (version 3.3.5), ggrepel (version 0.9.1), reshape2 (version 1.4.4), plyr (version 1.8.6), tibble (version 3.1.6), GenomicRanges (version 1.46.1) and DESeq2 (version 1.32.0). Transcript count matrix from StringTie analysis, the annotation file with the non-redundant set of transcripts and a file with the targeted transcripts coordinates were imported into R and transformed to GRanges objects. This was done to filter out any transcripts not included in the capture using the subsetByOverlaps() function from GenomicRanges. This file was used as input for DESeq2 differential expression analysis. Again, the strategy to correct for variability was as described previously, by estimating size factors based on all captured genes excluding genes located in the locus where the SVA was removed from (*PSEN1*, *PAPLN* and *RBM25*). For full scripts see StringTie_analysis in **Supplemental Code**. Finally, only transcripts arising from the *PSEN1* locus were considered (**Supplemental Table S12**). Transcripts not present in the Ensembl V104 annotation file are annotated with the prefix “MSTRG”.

For visualisation purposes, the canonical transcript of *PSEN1* (ENST00000324501), together with the transcripts showing a differential expression (ENST00000555867 and ENST00000555386) were exported from the ensembl website (http://jul2023.archive.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000080815;r=14:73603155-73687109) and adapted in **Figure 6B**.

Acknowledgements

This work was supported by a grant from Alzheimer Nederland (WE.03-2018-07) to F.M.J. Jacobs, H. Holstege and M.J.T. Reinders. We thank Gonzalo Congrains Sotomayor for technical assistance with FACS; Wim de Leeuw for technical support with bioinformatics; Fred White for support with data analysis; MAD: Dutch Genomics Service & Support Provider of the University of Amsterdam for sequencing; and the Evolutionary Neurogenomics Group for helpful discussions and advice about the project, particularly Elise van Bree, Diana Pereira Fernandes, Colette Moses and Jimi Rosenkrantz.

Author contributions

F.M.J.J. and R.L.F.P.G. conceptualized the project. R.L.F.P.G. performed all experiments and analyses. F.M.J.J., M.J.T.R. and H.H. acquired funding for the study. R.L.F.P.G. and F.M.J.J. wrote the manuscript with contributions of M.J.T.R. and H.H.. F.M.J.J. supervised the study.

4

Preliminary evidence for the association of a VNTR in the *RTEL1* gene with Alzheimer's disease risk using short-read sequencing data

Rita LFP Guimaraes[1,2], Niccolò Tesi[1,3], Marc Hulsman[1,3], Francesca Lucas[3], Marcel JT Reinders[3], Frank MJ Jacobs[2,4], Henne Holstege[1,3,4$]

1 Section Genomics of Neurodegenerative Diseases and Aging, Department of Clinical Genetics, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands
2 Evolutionary Neurogenomics, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands
3 Delft Bioinformatics Lab, Delft University of Technology, The Netherland
4 Amsterdam Neuroscience, Complex Trait Genetics, University of Amsterdam, Amsterdam, The Netherlands

$ corresponding author

Abstract

Tandem repeats are dynamic stretches of repetitive DNA with gene-regulatory properties which are the cause of many neurological disorders, both neurodevelopmental and neurodegenerative. In this study we sought to identify *de novo* variable number of tandem repeats (VNTRs) associated with Alzheimer's Disease with short-read sequencing data derived from 748 AD cases and 630 controls from the ADSP cohort. In total, 52 VNTRs were identified using two methods of analysis: (1) a case-control with logistic regression models to test distribution differences between AD cases and controls; and (2) an outlier model, based on outlier detection with interquartile rule and test disease association with Fisher's exact test. In a replication phase using long-read sequencing data we found a VNTR in the 30th intron of the *RTEL1* gene whose contraction is significantly associated with AD. *RTEL1* is an important gene involved in DNA repair, replication, transcription, and telomeric stability. Given that this VNTR is located within a regulatory element of the *RTEL1* gene, VNTR length variations may affect *RTEL1* expression levels. As this VNTR is also in the promoter region of *TNFRSF6B* and in the fusion transcript of *RTEL1-TNFRSF6B*, it may similarly influence their expression levels. While these initial findings hold promise, further exploration is required to elucidate the precise relationship between variations in the length of the VNTR, *RTEL1* and/or *TNFRSF6B* gene expression and the risk of developing AD.

Introduction

Alzheimer's disease (AD) is the most common form of dementia, a complex neurodegenerative disease, characterised by a steady decline of cognitive capabilities. AD is estimated to affect 50 million people worldwide, which is predicted to be tripled by 2050, making AD one of the major health challenges of the 21st century (https://www.alzint.org/resource/numbers-of-people-with-dementia-worldwide/). Despite enormous advances in recent years, the aetiology underlying AD remains elusive. Next to lifestyle factors including education, sleeping habits, etc. (Livingston et al. 2020), a considerable proportion of AD risk is heritable. Based on twin studies, the estimated heritability of sporadic AD ranges 60-80% (Gatz et al. 2006), with an approximate 30% of the genetic risk being attributed to the ε4 allele of the *APOE* gene (Ashford and Mortimer 2002; Corder et al. 1994). Genome-wide association studies (GWAS) have identified 85 genetic loci, marked by single nucleotide polymorphisms (SNPs) that are significantly associated with the risk of AD in the European population (Bellenguez et al. 2022). However, the SNPs genotyped in GWAS studies are mostly benign non-coding 'markers' of genetic regions with which they are in linkage disequilibrium (LD). Thus, most GWAS SNPs only 'tag' the AD-associated loci and do not themselves drive the association with disease (Edwards et al. 2013). To identify actual driving genetic elements within the associated genetic regions it is necessary to explore these risk loci for genetic elements that influence genetic expression. These may concern coding or noncoding variants that influence the expression of nearby genes.

In efforts to fine-map genetic risk loci identified by GWAS, structural variants such as deletions, insertions, or tandem repeats have not been sufficiently taken into account due to technical limitations (Audano et al. 2019; Linthorst et al. 2020). Of all structural variants, tandem repeats (TR) are among the most prevalent together representing ~30% of the human genome. TRs are stretches of DNA in which a specific sequence motif is consecutively repeated in a tandem-like fashion for a variable number of times. Based on the size of the repeated motif, TRs are distinguished between short tandem repeats (STR), with a motif size between 2 and 6 bp, and variable number of tandem repeats (VNTRs), with a motif size larger than 7 bp (Bakhtiari et al. 2021). TRs tend to be unstable in their number of units because of their repetitive nature, which increases their susceptibility to length variations caused by unequal recombination or slippage events during DNA replication and/or transcription (Richard and Pâques 2000). Importantly, TRs may harbour length-dependent gene-regulatory properties, which allows them to influence gene expression (Bakhtiari et al. 2021). Repeats are also specifically prevalent in human neuronal genes, and these may, in fact, have driven the alteration of gene expression levels that underlie the evolutionary changes going from primates to human lineages (Sulovari et al. 2019). However, the presence of polymorphic tandem repeats in neuronal genes also has negative consequences, as their proneness to expand

beyond a critical threshold may give rise to neurological disorders such as Fragile-X Syndrome, Huntington disease, frontotemporal dementia and amyotrophic lateral sclerosis (Chintalaphani et al. 2021), and AD (De Roeck et al. 2018). Together, this makes it especially important to consider the impact of TRs when fine-mapping GWAS regions in the context of neurodegenerative diseases.

VNTR lengths were historically studied with laborious experimental methods such as Southern blotting, the development of high-throughput short-read genetic sequencing approaches allowed for investigation of VNTR-based genetic sequence. In this study, we analysed a short read sequencing dataset from the Alzheimer's disease sequencing project (ADSP) to identify VNTR contractions or expansions that associate with an increased risk of AD, with motif size ≥ 7 bp, that occur in more than 1% of the AD cases (Schork et al. 2009).

To find potentially 'pathogenic' structural variants in VNTRs associated with AD, we compared the size of VNTRs in a sample of paired-end short-read data of 748 AD patients and 630 healthy controls. Since short-read sequencing reads are commonly too short to span the repetitive region of a complete VNTR, especially if it has expanded, we used the overall coverage of sequencing reads encompassing the repetitive sequences as an estimate of the repeat length of both alleles (Dolzhenko et al. 2019). With this approach, in a case control analysis, we observed that nine and 24 VNTRs were respectively longer and shorter in AD cases compared to controls, whilst 11 showed longer or shorter lengths with either the short or long alleles. In parallel, in an outlier analysis we observed three expansions and seven contractions. In total, 52 VNTRs were identified, where two overlapped in both analyses.

Notably, we observed similar differences in the lengths of 5 VNTRs in a novel independent dataset comprising long-read sequencing of 250 AD cases and 251 non-demented controls. Since such long-reads span 10-30 Kb, they can accurately capture and genotype repetitive DNA sequences, which allows for an accurate characterization of SV-length and SV-sequence for each allele separately (Audano et al. 2019; Linthorst et al. 2020). The sample sizes of both the short-read and long-read sequencing datasets are too small to reach genome-wide significance, such that our analysis provides preliminary evidence for the association of a contraction of the VNTR in the *RTEL1* locus with AD. This signal will need to be replicated in a larger independent dataset.

Results

An overview of our analysis pipeline is given in **Figure 1**. To precisely characterise VNTRs between cases and controls, we applied ExpansionHunter (EH) on whole genome-sequencing data (Illumina HiSeqX). However, as EH requires a catalog of VNTR regions, we first discover candidate AD-associated VNTRs using ExpansionHunter Denovo (EHdn), which identifies In-Repeat-Reads (IRRs) (i.e reads consisting of a repeating motif) for each sample (**Figure 1A**). EHdn then analyses if the coverage for each repeating region is significantly higher than the overall coverage as this could indicate an expanded repeat. The repeating regions for which there is an unusually high coverage, for either i) cases compared to controls, or ii) outlier cases, are reported by EHdn: these are the candidate AD-associated VNTRs. Because EHdn considers a subset of all aligned reads (only IRRs) and its size estimates are essentially a proxy for coverage, the output should be interpreted as an indication that a VNTR may be present at a specific location. The candidate AD-associated VNTRs are converted into a variant catalog and put in genomic context, annotating whether they fall within or near either a gene or a functional non-genic element (e.g. enhancer). A variant catalog is a file describing the genomic coordinates and motif information (motif size, motif pattern) of each repeat to be analysed. As EH, in the subsequent step, greatly depends on the accuracy of the provided coordinates and those outputted by EHdn are imprecise, we establish accurate start and end coordinates for each candidate AD-associated VNTR based on the reference genome during the variant catalog creation. Finally, we perform a gene set enrichment analysis for a contextual understanding.

To obtain more precise repeat size estimates, we then use ExpansionHunter (EH) to further analyse the candidate VNTRs identified by EHdn (**Figure 1B**). EH analyses each of the repeats from the established variant catalog, combining the information from multiple types of reads (spanning, flanking and anchored IRRs), and estimates diploid repeat sizes. After estimating repeat sizes for all candidate VNTRs for all samples, we identify the AD-associated VNTRs, again using a case-control setting as well as an outlier approach.

To investigate whether the AD-associated VNTR are in previously identified AD loci, we check the genomic context for GWAS-identified SNPs associated and test the association through a linkage analysis (**Figure 1C**). Finally, we replicate our preliminary findings using long-read sequencing data in a smaller dataset of Alzheimer's disease patients and cognitively healthy centenarians (**Figure 1D**).

Generation of a catalog of candidate AD-associated VNTRs

Using ExpansionHunter Denovo (EHdn), we identified *de novo* VNTRs genome-wide, i.e. without a predefined set of repeats, in a sample of 771 samples consisting of 452 Alzheimer's disease cases and 319 non-demented controls from the ADSP cohort (**Figure 1A**).

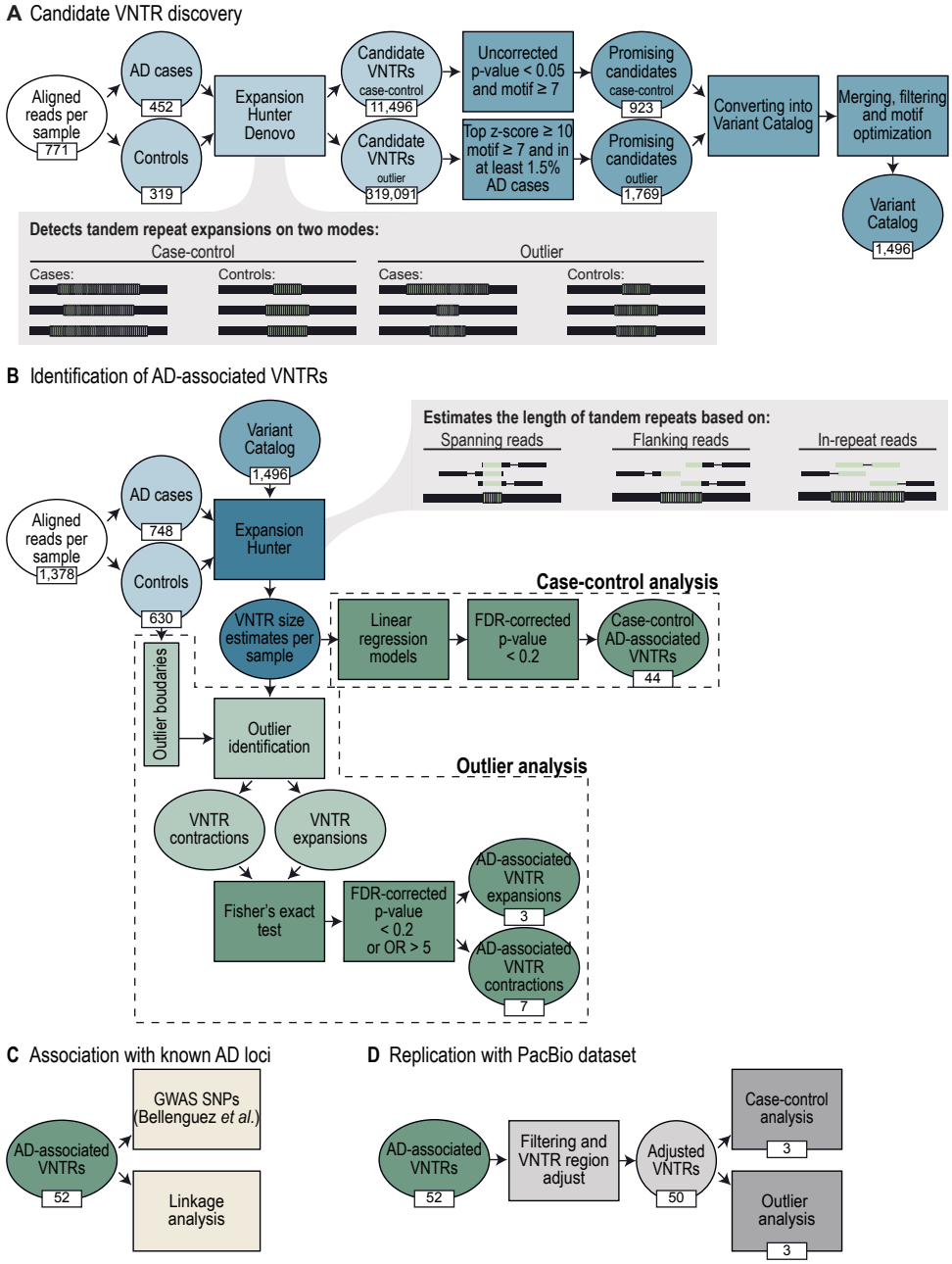


Figure 1: Overview of pipeline. A, Candidate VNTR discovery and catalog creation with ExpansionHunter Denovo. Grey box: schematic representation of analysis modes of ExpansionHunter Denovo. **B,** Identification of VNTRs associated with AD in a case-control and outlier analysis. Grey box: schematic representation of types of reads used in the estimation of lengths of tandem repeats. **C,** Association of VNTRs with known AD loci. **D,** Replication phase with long-read PacBio dataset. Circles represent the in/output data and rectangles represent the action done on data.

We created a set of candidate AD-associated VNTRs using EHdn that operates in (i) a case-control mode, which compares the distributions of IRRs between AD cases and controls; and (ii) an outlier-based mode, which calculates a z-score for cases exceeding the mean of the compared cases + controls IRR count distribution (**Figure 1A**). The EHdn case-control mode detected 11,794 VNTRs in 452 AD cases and 319 controls that had longer or shorter mean lengths in AD cases compared to controls (see **Methods**), of which 923 VNTRs (7.8%) with an uncorrected *P*-value < 0.05 and a motif size ≥ 7 bp. In parallel, the EHdn outlier mode detected 319,091 VNTRs with expansions/contractions that occur only in a subset of the AD cases and may not be detected by the case-control mode due to the low number of samples (see **Methods**). Of these, 1,769 VNTRs (0.6%) had lengths that were at least 10 standard deviations longer or shorter compared to the mean VNTR length of cases and controls combined (a top z-score ≥ 10), with motif size ≥7 bp, and were observed in at least 1% of the AD cases. Of the VNTRs returned by the case-control analysis, 97% are also present in the output of the outlier analysis. The full output for both analyses can be found in **Supplementary Tables S1, S2**.

Candidate AD-associated VNTRs from both the case-control and outlier analysis, were mapped onto the GRCh38 reference genome to derive exact coordinates necessary for downstream analyses (see **Methods**, **Supplemental Data**). After this conversion, both outputs were merged to create a variant catalog of candidate AD-associated VNTRs. To minimise any biases coming from multiple (short) repeat sequences in the same locus, we combined repeats within a window of 5 kb with the same motif size and a motif sequence similarity of at least 80%. In addition, we introduced degeneracy to the motif's DNA sequence by using the International Union of Pure and Applied Chemistry (IUPAC) code (see **Methods**).

Our final variant catalog of candidate AD-associated VNTRs included a total of 1,496 unique VNTRs, with motif sizes ranging from 7 to 20 bp (**Supplemental Table S3**). In total, 58.1% were previously reported as variable in length in Audano et al. (2019), Linthorst et al. (2020), and in a tandem repeat catalog from PacBio (https://github.com/PacificBiosciences/trgt/blob/main/repeats/repeat_catalog.hg38.bed), while the remaining 40.9% may be considered as novel variations (**Supplemental Table S3**). 46.3% of the VNTRs (n = 692) were located within a gene, and 1.7% (n = 26) were exonic. These genic VNTRs might be exceptionally interesting for disease risk as they could exert a direct functional impact. 6.6% of the VNTRs (n = 98) were found in promoter regions, close to the transcription start site (TSS), which often consist of repetitive sequence, containing regulatory CpG islands (Vavouri and Lehner 2012). 13.8% of the VNTRs (n = 207) are located completely or partially within functional elements such as enhancers, regions of CTCF-binding, which allows the interaction between enhancers and promoters, and transcription factor binding sites (Pruitt et al. 2014; Pruitt, Tatusova, and Maglott 2005). Finally, 38.2% of the VNTRs (n = 572) were in telomeric regions, which is in agreement with previous research (Linthorst et al. 2020; Sulovari et al. 2019).

The 692 genic VNTRs mapped in 531 genes. A gene ontology (GO) analysis (see **Methods**) indicated these genes were enriched with GTPase regulator activity (GO:0030695, $P = 2.27 \times 10^{-4}$); regulation of signalling (GO:0023051, $P = 4.79 \times 10^{-7}$) and synaptic signalling (GO:0099536, $P = 4.87 \times 10^{-3}$); and they were enriched to localise to the membrane (GO:0016020, $P = 2.92 \times 10^{-14}$), cytoplasm (GO:0005737, $P = 3.49 \times 10^{-6}$), and (main) axon (GO:0044304, $P = 9.15 \times 10^{-3}$) (**Supplemental Table S4**).

Association of candidate VNTRs to Alzheimer's Disease

After the generation of the variant catalog of 1,469 candidate AD-associated VNTRs, we used EH to get a precise estimate of VNTR lengths for the two separate alleles (which we labelled as short and long). Using a logistic regression model, we then tested whether length distributions are associated with AD status while adjusting for sex, population ancestry, family relations and CRAM/BAM file statistics (**Figure 1B**, *case-control analysis*). To increase statistical power, we accessed the new data release from the ADSP cohort. After quality control, we were able to include a total of 1,378 individuals: 748 AD cases (37% EOAD, early onset, age at onset < 65) and 630 non-demented controls in the analysis. A summary of metadata on the study population can be found in **Table 1**.

Table 1: Study population. Metadata of the EH-analysed samples from the ADSP cohort.

	AD cases	Controls
Total	748	630
Age	73±14	87±5
EOAD	37.17%	NA
Female	55%	59%
Coverage	42±4.8	44±5.9
Read length	151±0.7	151±0.6
Insert size	1760±150	1772±140

Case-control analysis identifies 44 candidate AD-associated VNTRs

Of the 1,496 candidate AD-associated VNTRs, the alleles of 44 VNTRs were significantly longer or shorter in the group of AD patients compared to controls (FDR-corrected *P*-value < 0.2) (**Table 2**). We have found 9 expansions (positive effect size) and 24 contractions (negative effect size) where both short and long alleles follow the same trend. However, 11 VNTRs show either an expansion or contraction with the short allele and the opposite with the long allele. The 44 VNTRs were enriched with 23 VNTRs that localised to subtelomeric regions (Fisher's exact test: $P = 0.039$); 19 were located within genes (intronic), three in promoter regions and six were part of functional elements, suggesting that these VNTRs may have a direct functional impact. Examples of the genes are: *RTEL1*, *ANXA2*, *B4GALNT3*, *DLGAP2* and *SDK1*, which are involved in fundamental cellular processes and/

Table 2: Case-control analysis with logistic regression models comparing VNTR lengths of AD cases and healthy controls in the discovery phase. Showing only significant associations (FDR-corrected *P*-value < 0.2). Columns: Region, coordinates of full VNTR genomic sequence in GRCh38; Motif, motif sequence of VNTR; Motif size, number of base pairs in motif sequence; CN range, range of VNTR copy-number in both alleles for AD cases and controls; Short-allele effect size and *P*-value, statistics of VNTR analysis when tested with the short allele in logistic regression models; Long-allele effect size and *P*-value, statistics of VNTR analysis when tested with the long allele in logistic regression models; Annotation, genomic annotation of VNTR.

Region	Motif	Motif size	CN range	Short-allele		Long-allele		Annotation
				Effect-size	FDR	Effect-size	FDR	
4:49153237-49156525	CCAYTCCATTCCATTSCDTT	20	[15-27]	$1.15x10^{-3}$±$2.40x10^{-4}$	0.002	$-6.17x10^{-4}$±$2.15x10^{-4}$	0.158	Simple repeat (TCCAT)n in functional element
2:207024-207712	AAGGGAGGGAGG	12	[1-17]	$-7.05x10^{-4}$±$1.57x10^{-4}$	0.004	$-5.52x10^{-5}$±$1.66x10^{-4}$	0.962	Telomeric
15:60349890-60350142	AAGGCAGGGAGGSAGGGG	18	[3-37]	$-6.61x10^{-4}$±$1.55x10^{-4}$	0.008	$-5.33x10^{-4}$±$1.53x10^{-4}$	0.069	In *ANXA2* gene
22:45890396-45890846	GMGGGGCGWGTGTGT	15	[9-33]	$-6.29x10^{-4}$±$1.53x10^{-4}$	0.012	$-5.31x10^{-4}$±$1.52x10^{-4}$	0.069	Telomeric in promoter region of *LOC107985535*
8:1470789-1471378	CCCGACCCCTCCAGCYTTT	19	[4-29]	$-6.37x10^{-4}$±$1.57x10^{-4}$	0.012	$-5.93x10^{-4}$±$1.56x10^{-4}$	0.056	Telomeric in *DLGAP2* gene
21:6565394-6566097	ACACACWCC	9	[31-57]	$-6.13x10^{-4}$±$1.58x10^{-4}$	0.020	$-4.71x10^{-4}$±$1.53x10^{-4}$	0.156	Intergenic
4:189379959-189380797	AAGCACACACASCTGC	16	[7-31]	$-5.85x10^{-4}$±$1.56x10^{-4}$	0.029	$1.09x10^{-4}$±$1.54x10^{-4}$	0.879	Telomeric
5:5954850-5955313	SCCCTTCCCTCCTT	14	[2-36]	$-5.52x10^{-4}$±$1.51x10^{-4}$	0.034	$-3.33x10^{-4}$±$1.51x10^{-4}$	0.389	Intergenic
12:545768-546002	GAGAGGAGTGGGGAGGT	17	[0-20]	$-5.50x10^{-4}$±$1.50x10^{-4}$	0.034	$-1.03x10^{-4}$±$1.50x10^{-4}$	0.881	Telomeric in *B4GALNT3* gene
21:41805989-41806413	ACCACCACCACCACCATC	18	[7-23]	$-5.16x10^{-4}$±$1.61x10^{-4}$	0.158	$-5.02x10^{-4}$±$1.61x10^{-4}$	0.156	Telomeric simple repeat (CACCAC)n in *PRDM15* gene
3:71435292-71435681	AMGRAGGGASGGAGGGAGGG	20	[1-32]	$-5.28x10^{-4}$±$1.66x10^{-4}$	0.158	$8.44x10^{-5}$±$1.47x10^{-4}$	0.907	In *FOXP1* gene
9:28189272-28189688	AAGGRAGGRAGGGAGG	16	[5-40]	$-5.08x10^{-4}$±$1.63x10^{-4}$	0.177	$-4.81x10^{-4}$±$1.62x10^{-4}$	0.157	Simple repeat (AGGA)n in *LINGO2* gene
15:52232253-52232333	AAGGAAGGAAGGAGAG	16	[1-9]	$-9.06x10^{-4}$±$2.95x10^{-4}$	0.190	$-5.29x10^{-4}$±$2.12x10^{-4}$	0.263	GA-rich repeat in *MYO5C* gene
12:132290532-132291402	ACAGCACCCACRTCC	15	[3-36]	$-4.47x10^{-4}$±$1.47x10^{-4}$	0.196	$-1.98x10^{-4}$±$1.47x10^{-4}$	0.642	Telomeric in *GALNT9* gene
10:103251021-103251117	ACCACCAYCACCACCWYC	18	[1-33]	$-4.63x10^{-4}$±$1.55x10^{-4}$	0.203	$-4.47x10^{-4}$±$1.51x10^{-4}$	0.157	Intergenic
6:170145995-170146318	AGGWGGGTGAGGTGTGT	17	[2-40]	$-3.93x10^{-4}$±$1.52x10^{-4}$	0.432	$-4.34x10^{-4}$±$1.52x10^{-4}$	0.158	Telomeric
7:144024125-144024220	ACCTCTCATCATCGCCTGT	19	[1-9]	$4.64x10^{-4}$±$2.09x10^{-4}$	0.563	$7.49x10^{-4}$±$2.55x10^{-4}$	0.157	Intergenic
5:14157420-14157546	CTCCCTCTCTCTCCCTGT	18	[3-12]	$-3.16x10^{-4}$±$1.60x10^{-4}$	0.641	$-5.95x10^{-4}$±$1.55x10^{-4}$	0.056	Simple repeat (CTCTT)n in functional element in *TRIO* gene
19:55670466-55670547	CCSTGCACCCTGCTGT	16	[2-20]	$6.53x10^{-4}$±$3.31x10^{-4}$	0.641	$-4.33x10^{-4}$±$1.54x10^{-4}$	0.166	Telomeric in functional element in *U2AF2* gene and promoter region of *EPN1*
22:46566659-46567756	AACCCACTCCCAWGATAACK	20	[3-31]	$-2.90x10^{-4}$±$1.52x10^{-4}$	0.649	$-5.22x10^{-4}$±$1.50x10^{-4}$	0.069	Telomeric complex region with MSTD LTR elements and simple repeat (ACTAACCC)n
6:2944776-2944902	CCCATACATAATTCCCTT	18	[3-10]	$4.32x10^{-4}$±$2.28x10^{-4}$	0.649	$5.88x10^{-4}$±$2.10x10^{-4}$	0.175	Telomeric
11:1670506-1670749	GGGYAGTGT	9	[1-56]	$-2.79x10^{-4}$±$1.48x10^{-4}$	0.649	$-4.43x10^{-4}$±$1.47x10^{-4}$	0.156	Telomeric MLT2D LTR element
6:157774248-157774265	GGGGGAGTGGAGGAGGT	17	[1-23]	$-2.54x10^{-4}$±$1.49x10^{-4}$	0.737	$-5.00x10^{-4}$±$1.56x10^{-4}$	0.128	Simple repeat (GGAGTGGA)n in functional element
20:1877066-1877190	**CCATCCCCCYGTGT**	**14**	**[1-31]**	$\mathbf{-2.31x10^{-4}}$±$\mathbf{1.55x10^{-4}}$	**0.761**	$\mathbf{-4.55x10^{-4}}$±$\mathbf{1.51x10^{-4}}$	**0.156**	**Telomeric**
8:699780-700404	ACACGCGCACAGRCCCGC	18	[2-26]	$2.21x10^{-4}$±$1.52x10^{-4}$	0.770	$-4.31x10^{-4}$±$1.50x10^{-4}$	0.158	Telomeric in *ERICH1* gene
7:3582240-3582285	CTCCCTCAGGTAGGT	15	[0-10]	$2.36x10^{-4}$±$1.70x10^{-4}$	0.785	$1.27x10^{-3}$±$4.20x10^{-4}$	0.156	Telomeric in *SDK1* gene

5:3276135-3276454	AAGGAGGGAGRGGAKG	16	[5-29]	$-1.95\times10^{-4}\pm1.50\times10^{-4}$	0.813	$-4.64\times10^{-4}\pm1.54\times10^{-4}$	0.156	Telomeric
20:6841319-6841370	AGAGAGAGAGGAGGG	15	[1-24]	$-1.97\times10^{-4}\pm1.51\times10^{-4}$	0.813	$-4.32\times10^{-4}\pm1.49\times10^{-4}$	0.158	GA-rich repeat
19:50811607-50812006	ACCATCYCAGGCAGKGAGG	19	[1-13]	$-1.98\times10^{-4}\pm1.53\times10^{-4}$	0.813	$4.36\times10^{-4}\pm1.59\times10^{-4}$	0.189	Intergenic
7:155792048-155792330	AATCCCTCCCTCTCTCCGT	19	[1-18]	$-1.77\times10^{-4}\pm1.49\times10^{-4}$	0.853	$-5.47\times10^{-4}\pm1.61\times10^{-4}$	0.088	Telomeric in functional element
2:426809-427550	CCGCWGGCACGCMGGGGGTT	20	[1-22]	$-1.69\times10^{-4}\pm1.49\times10^{-4}$	0.861	$4.33\times10^{-4}\pm1.49\times10^{-4}$	0.158	Telomeric
3:172644161-172644199	GGGGTGACTCTAGGGCTTT	19	[0-8]	$6.68\times10^{-4}\pm6.33\times10^{-4}$	0.899	$4.23\times10^{-4}\pm1.52\times10^{-4}$	0.181	In *NCEH1* gene
15:42557329-42557401	ATATATAATATATATTTT	18	[1-15]	$-1.36\times10^{-4}\pm1.49\times10^{-4}$	0.928	$4.91\times10^{-4}\pm1.64\times10^{-4}$	0.156	In *HAUS2* gene
16:80196906-80196960	ACACACACATATATAT	18	[1-10]	$1.73\times10^{-4}\pm1.79\times10^{-4}$	0.928	$4.35\times10^{-4}\pm1.52\times10^{-4}$	0.158	Simple repeat (AC)n in *DYNLRB2-AS1* gene
10:132097911-132097940	CCCTTCCCCTTT	12	[1-22]	$1.16\times10^{-4}\pm1.54\times10^{-4}$	0.979	$4.22\times10^{-4}\pm1.48\times10^{-4}$	0.160	Telomeric simple repeat (CTTCCC)n in *JAKMIP3* gene
2:238870127-238871508	ACACACACCACACCC	16	[1-17]	$-6.93\times10^{-5}\pm1.47\times10^{-4}$	0.989	$-5.65\times10^{-4}\pm1.62\times10^{-4}$	0.069	Telomeric simple repeat (AC)n in functional element in *TWIST2* gene
20:63693482-63693758	ACCACCACCTCC	12	[10-23]	$-5.21\times10^{-5}\pm1.54\times10^{-4}$	0.998	$-6.51\times10^{-5}\pm1.62\times10^{-4}$	0.056	Telomeric simple repeat (CCA)n in *RTEL1* gene, in *RTEL1-TNFRSF6B* fusion transcript and in the promoter of *TNFRSF6B*
1:18049728-18049812	AGTGGTGGTGGTGGT	15	[1-12]	$-1.33\times10^{-4}\pm1.27\times10^{-3}$	0.998	$5.67\times10^{-4}\pm1.61\times10^{-4}$	0.069	Simple repeat (TGATGG)n
2:231381278-231381326	ACTCCCTCCTCC	12	[1-18]	$3.52\times10^{-5}\pm1.90\times10^{-4}$	0.998	$5.14\times10^{-4}\pm1.53\times10^{-4}$	0.088	Intergenic
3:20854642-20855529	ACACTGTCACCCCAGAG	17	[1-10]	$-1.97\times10^{-2}\pm4.98\times10^{-1}$	0.998	$7.40\times10^{-4}\pm2.23\times10^{-4}$	0.093	Intergenic
1:3572159-3572294	GCTGGGTCCTCCTGGGTGT	19	[1-9]	$3.43\times10^{-5}\pm3.40\times10^{-4}$	0.998	$4.68\times10^{-4}\pm1.54\times10^{-4}$	0.156	Telomeric in *MEGF6* gene
17:71609424-71609724	ATATCTMTATCTCTATCTCT	20	[0-11]	$-1.51\times10^{-5}\pm2.03\times10^{-4}$	0.998	$5.10\times10^{-4}\pm1.73\times10^{-4}$	0.157	Intergenic
11:95015768-95015876	ATATATATGTGT	12	[1-16]	$3.16\times10^{-5}\pm1.57\times10^{-4}$	0.998	$6.37\times10^{-4}\pm2.22\times10^{-4}$	0.158	Simple repeat (TA)n
18:80258333-80258399	AGGGTGAGGGTT	12	[1-18]	$-3.09\times10^{-5}\pm1.49\times10^{-4}$	0.998	$-4.82\times10^{-4}\pm1.75\times10^{-4}$	0.188	Telomeric G-rich repeat

or have previously been associated with AD (**Figure 2A**, **Supplemental Figure S1**). Despite the significant association of these VNTR alleles to AD, the graphical representation of the results does not clearly portrait the difference between cases and control. Full results are available in **Supplemental Table S5**.

Outlier-based analysis detected 10 candidate AD-associated VNTRs

As contractions or expansions in VNTRs may occur only in a small subset of AD cases (and thus not detectable in a case-control setting due to power issues), we sought to identify outliers using the interquartile rule (IQR) followed by a one-sided test with the alternative hypothesis that cases are more likely to be outliers (**Figure 1B**, *Outlier analysis*). Of our candidates list, seven contractions and three expansions in VNTRs were observed significantly more often in AD cases compared to controls (FDR-corrected Fisher's exact test *P*-value < 0.2 or odds ratio (OR) > 5) (**Table 3**). Four of the seven contracted VNTRs were in genes: *ADAMTSL2, C17orf67, PTPRN2 and LINGO2* (**Supplemental Figure S2A**). Notably, the ADAMTS protein family, to which ADAMTSL2 belongs, has been previously associated with AD in GWAS (*ADAMTS4*, *ADAMTS1*, *ADAMTS20*) (Bellenguez et al. 2022; Jansen et al. 2019; Kunkle et al. 2019; Schwartzentruber et al. 2021). Of the three expanded VNTRs, two were within genes: the long non-coding RNA (lncRNA) *LOC101928269* (**Figure 2C**) and *SDK1* (**Supplemental Figure S2B**). The latter was also observed as one of the 44 VNTRs with a different length distribution in AD cases compared to controls (the case-control analysis). Here, none of the VNTRs associated with AD are part of regulatory elements. Full results are available in **Supplemental Table S6**.

More than half of the VNTRs are in AD-associated loci

After combining the case-control results and the outlier-based results, we had a total of 52 candidate VNTRs, with a suggestive association with AD. Then, we sought to investigate whether these VNTRs were in linkage with SNPs identified by Bellenguez et al. (2022). To that end we performed a correlation analysis between the VNTR length and all independent SNPs within a window of 1 Mb of the detected VNTR, in each locus (**Figure 1C**). 36 of the 52 VNTRs (60%) were in (partial) linkage with at least 1 GWAS-identified SNP (FDR-corrected Spearman's correlation *P*-value < 0.05). Two examples are the *RTEL1*-VNTR, where shorter VNTR lengths associates with AD risk and longer lengths are correlated with the presence of a protective GWAS SNP rs55765053 (**Figure 2A-B**); or the *LOC101928269*-VNTR, whose expansion associates with AD risk and correlates with the presence of a risk GWAS SNP rs45559732 (**Figure 2C-D**).

None of the VNTRs were in strong linkage with AD-haplotypes identified as genome-wide significant by Bellenguez et al. (2022) (**Supplemental Table S7**). However, 8 (15%) and 31 (60%) VNTRs have at least one AD-associated SNP within a 100 kb or 500 kb window with an AD association *P*-value < 5×10^{-4}, respectively (**Supplemental Table S8**).

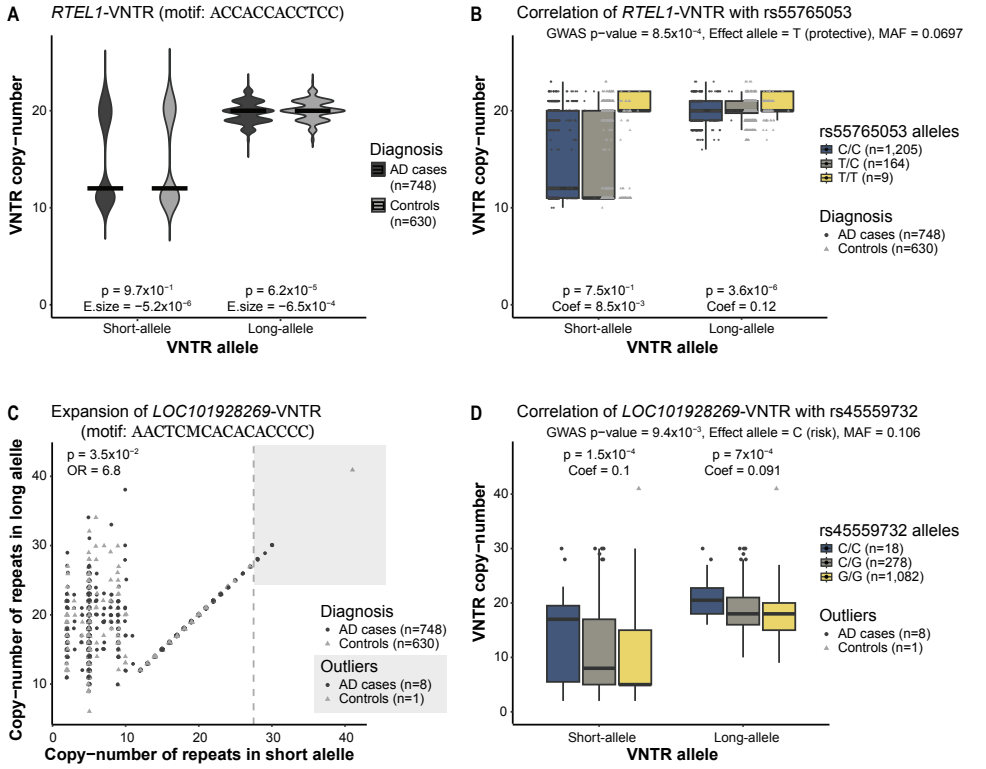


Figure 2: VNTR variations associated with Alzheimer's disease from case-control and outlier analysis. A, Case-control analysis of VNTR copy-numbers in *RTEL1* gene (Chr20:63,693,482-63,693,758) associated to Alzheimer's disease using logistic regression models, for both short and long alleles. **B,** Correlation analysis between copy-numbers of *RTEL1*-VNTR and the status of protective GWAS SNP rs55765053 from Bellenguez et al. (2022), for each short and long allele separately. **C,** Alzheimer's disease association of expanded VNTR in the long non-coding RNA *LOC101928269* (Chr21:35,993,052-35,993,132) present in a subset of cases (outlier analysis). The outlier boundary is shown as a dashed grey line. **D,** Correlation analysis between copy-numbers of *LOC101928269*-VNTR and the status of risk GWAS SNP rs45559732 from Bellenguez et al. (2022), for each short and long allele separately. Showing only the outliers. Showing uncorrected *P*-values of AD-association tests.

Table 3: Outliers analysis using Fisher's exact test for VNTR lengths of AD cases and healthy controls in the discovery phase. Showing only significant associations (FDR-corrected *P*-value < 0.2 or OR > 5). Columns: Region, coordinates of full VNTR genomic sequence in GRCh38; Motif, motif sequence of VNTR; Motif size, number of base pairs in motif sequence; Tested allele, (short or long) allele analysed with Fisher's exact test; Boundary, outlier boundary calculated with IQR based on controls distribution; AD cases and controls [cn range], sample number of AD cases and controls and the corresponding range of VNTR copy-number in both alleles; Odds ratio and FDR-corrected *P*-value, statistics related to the allele analysed with Fisher's exact test; Annotation, genomic annotation of VNTR.

	Region	Motif	Motif size	Tested allele	Boundary	AD cases [cn range]	Controls [cn range]	Odds ratio	FDR-corrected p-value	Annotation
Expansion	21:35993052-35993132	AACTCMCACACACCCC	16	Short	27	8 [28-30]	1 [41]	6.8	$8.80x10^{-1}$	In *LOC101928269* gene
	6:469630-469690	CTCCCTCCTCTCCGT	15	Short	6	10 [7-16]	1 [7-8]	8.51	$8.80x10^{-1}$	Telomeric simple repeat (GTCTCT)n
	7:3582240-3582285	CTCCCTCAGGTAGGT	15	Long	3	51 [3-10]	15 [3-8]	3	$4.93x10^{-2}$	Telomeric in *SDK1* gene
Contraction	21:6565394-6566097	ACACACWCC	9	Short	33	26 [31-49]	3 [31-32]	7.52	$1.97x10^{-2}$	Intergenic
	19:5796199-5796280	CTCCTCCTT	9	Short	3	12 [1-18]	2 [2-18]	5.11	$3.67x10^{-1}$	Simple repeat (CTTCTCCTC)n
	17:56792862-56792880	GATGGTGGTGGTGGTGGT	18	Short	1	8 [0-1]	1 [0-1]	6.79	$4.71x10^{-1}$	Simple repeat (GTG)n in *C17orf67* gene
	21:6565394-6566097	ACACACWCC	9	Long	33	19 [31-32]	3 [31-32]	5.44	$1.66x10^{-1}$	Intergenic
	7:158395655-158395718	AGGGGCG	7	Long	23	23 [3-22]	5 [3-22]	3.96	$1.66x10^{-1}$	Telomeric simple repeat (GAGGGGC)n in *PTPRN2* gene
	9:133564168-133564642	AGAGAGAGAGAGAGGG	16	Long	10	51 [1-9]	18 [1-9]	2.49	$1.66x10^{-1}$	Telomeric in *ADAMTSL2* gene
	11:68644608-68644656	AAAGAAGGAAGGAAGG	16	Long	3	53 [1-2]	20 [1-2]	2.32	$1.66x10^{-1}$	Simple repeat (AGGA)n
	9:28189312-28189624	AAGGAAGGRAGGGAGGGAGG	20	Long	9	79 [1-8]	38 [1-8]	1.84	$1.66x10^{-1}$	Simple repeat (AGGA)n *LINGO2* gene

Contraction of *RTEL1*-VNTR is also observed in an independent long-read sequencing dataset and correlated with a top GWAS SNP

To obtain more support for an AD-association for the expansions/contractions of the identified significant VNTRs, we investigated whether their length differences were also observable in an independent long-read sequencing dataset (**Figure 1D**) of 250 AD cases and 251 cognitively healthy centenarians, as described by Salazar et al. (2023). After QC on the VNTRs and long-read data, 50 VNTRs were genotyped with Treat (Tesi et al. 2024) and their lengths were tested for their association with AD, again with both a case-control and outlier analysis.

The case-control analysis indicated that the lengths of three of the 50 AD-associated VNTRs show contractions, in both short and long alleles, in AD cases compared to controls (**Table 4**, **Supplemental Table S10**): the intronic VNTR in *RTEL1* gene (**Figure 3A**), and two intergenic VNTRs (**Supplemental Figure S3**). Similarly, the outlier analysis indicated that the lengths of three candidate VNTRs were altered in AD cases relative to controls (**Table 5**, **Supplemental Table S11**): one contraction in the long allele of the intronic VNTR in the *RTEL1* gene (uncorrected Fisher's

Table 4: Case-control analysis with logistic regression models comparing VNTR lengths of AD cases and cognitively healthy centenarians in the replication phase. Showing only significant associations (FDR-corrected *P*-value < 0.2). Columns: Locus, coordinates of full VNTR genomic sequence in GRCh38; Motif, motif sequence of VNTR; Short-allele effect size and *P*-value, statistics of VNTR analysis when tested with the short allele in logistic regression models; Long-allele effect size and *P*-value, statistics of VNTR analysis when tested with the long allele in logistic regression models.

Locus	Motif	Short allele		Long allele	
		Effect size	p-value	Effect size	p-value
Chr2:206,996–207,736	AGGGAGGGAGGG	$-1.2x10^{-3}$ ($\pm 8.6x10^{-4}$)	$1.8x10^{-1}$	$-2.3x10^{-3}$ ($\pm 8.7x10^{-4}$)	$7.3x10^{-3}$
Chr6:157,774,214–157,774,281	ACCTCCTCCACTCCCC	$-2.0x10^{-4}$ ($\pm 6.7x10^{-4}$)	$7.7x10^{-1}$	$-1.7x10^{-3}$ ($\pm 6.8x10^{-4}$)	$1.4x10^{-2}$
Chr20:63,693,420–63,693,781	ACCACCACCACCACCTCC	$-1.1x10^{-3}$ ($8.8x10^{-4}$)	$2.2x10^{-1}$	$-1.8x10^{-3}$ ($\pm 8.8x10^{-4}$)	$4.4x10^{-2}$

Table 5: Outliers analysis using Fisher's exact test for VNTR lengths of AD cases and cognitively healthy centenarians in the replication phase. Showing only significant associations (FDR-corrected *P*-value < 0.2 or OR > 5). Columns: Locus, coordinates of full VNTR genomic sequence in GRCh38; Motif, motif sequence of VNTR; Tested allele, (short or long) allele analysed with Fisher's exact test; Boundary, outlier boundary calculated with IQR based on controls distribution; AD cases and CHC, sample number of AD cases and cognitively healthy centenarians; Odds ratio (OR), Fisher's *P*-value and FDR-corrected Fisher's *P*-value (FDR), statistics related to the allele analysed with Fisher's exact test.

	Locus	Motif	Tested allele	boundary	AD cases	CHC	OR	p-value	FDR
Contraction	Chr20:63,693,420–63,693,781	ACCACCACCACCACCTCC	long	2,337.25	9	1	11.30	$4.4x10^{-3}$	0.09
Expansion	Chr12:545,634–546,071	ACTCCTCGCTCCTCCCC	short	441.00	6	1	6.13	$6.1x10^{-2}$	0.73
Expansion	Chr21:35,992,932–35,994,479	AACTCACACACACCCC	long	1,649.00	3	0	Inf	$1.2x10^{-1}$	0.79

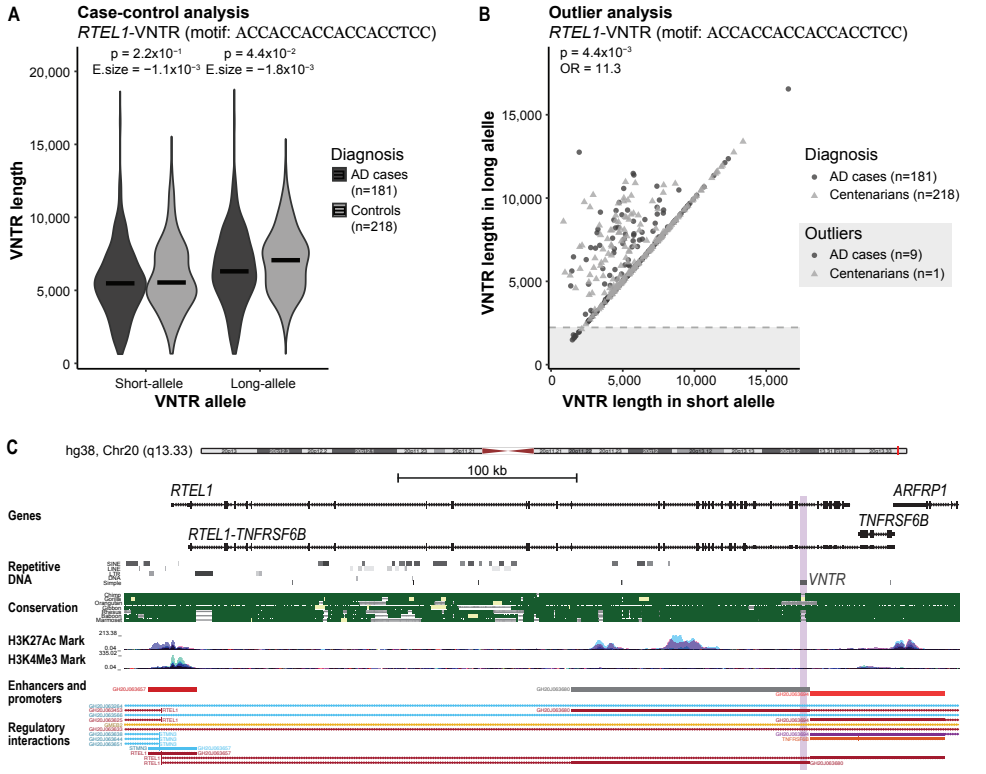


Figure 3: Contraction of *RTEL1*-VNTR. A-B, Alzheimer's disease association of VNTR lengths in *RTEL1* (Chr20:63,693,482-63,693,758) in a case-control (**A**) and outlier detection analyses (**B**). Showing uncorrected *P*-values. **A,** Case-control analysis of intronic VNTRs lengths associated to Alzheimer's disease with logistic regression models, using both short and long alleles. **B,** Alzheimer's disease association of intronic VNTR contractions present in a subset of cases. Outliers are calculated with interquartile rule and are shown in grey shading. The outlier boundary is shown as a dashed grey line. **C,** Architecture of *RTEL1* and surrounding genes, including repetitive DNA, a primate conservation track, histone marks (H3K27ac and H3K4me3), enhancers (grey) and promoters (red) and their interaction with annotated genes from GeneHancer. Lilac shading area highlights the location of the *RTEL1*-VNTR. Adapted from the UCSC Genome Browser.

exact test: *P* = 0.0044, FDR-corrected *P* = 0.09, OR = 11.30) (**Figure 3B**); and two expansions on the short allele of *B4GALNT3*-VNTR (uncorrected Fisher's exact test: *P* = 0.061, FDR-corrected *P* = 0.73, OR = 6.13) and on the long allele of lncRNA *LOC101928269*-VNTR (uncorrected Fisher's exact test *P* = 0.12, FDR-corrected *P* = 0.79) (**Supplemental Figure S4**). Interestingly, the expansion of *B4GALNT3*-VNTR was associated with AD risk, yet in the opposite direction as compared with the discovery phase.

The VNTRs in the *RTEL1* and lncRNA *LOC101928269* are the strongest candidates for an AD association in our study. Thus, we assessed the genetic architecture of these loci regarding activating histone marks, regulatory elements and sequence conservation amongst primate species on UCSC Genome Browser (Kent et al. 2002). The lncRNA *LOC101928269*-VNTR (Chr21:35,992,932-35,994,479, GRCh38) located in its first intron shows high structural variation across primates and a lack of potential regulatory functions as it lays in a region excluded of any regulatory elements (**Supplemental Figure S5**). The VNTR in the 30th intron of the *RTEL1* gene (chr20:63693420-63693781, GRCh38), only present in the human and gibbon species, is also located in the fusion transcript of *RTEL1-TNFRSF6B*, and present in the promoter region of *TNFRSF6B* (**Figure 3C**). Furthermore, this VNTR is part of a regulatory region which interacts with the promoter of *RTEL1*, suggesting the VNTR may play a role in regulating *RTEL1* gene expression levels (**Figure 3C**, *lower panel*).

The *RTEL1*-VNTR has a motif of 18 bp (ACCACCACCACCACCTCC), characterised by Treat. The VNTR comprises 361 bp of the human reference genome GRCh38, corresponding to 20.01 repeating units. Whilst in our long-sequencing cohort the smallest VNTR length is 631 bp and the longest is 16,548 bp, corresponding to 35.01 and 919.33 repeating units, respectively. The variation of this VNTR has been reported as insertions ranging from 86 to 256 bp by Audano et al. (2019) (**Supplemental Table S3**), though its association with AD is new. We also found a correlation with SNP rs55765053 (**Figure 3B**), located in *TNFRSF6B,* 3,346 bp downstream of the VNTR, which in turn is in modest LD with rs6742 ($R^2 = 0.21$, $P < 0.0001$), the main protective SNP in a novel locus associated with AD (OR = 0.95, GWAS $P = 2.6 \times 10^{-9}$) (Bellenguez et al. 2022). However, the correlation between the VNTR lengths and rs6742 did not reach significance with either allele. Further studies are required to validate the association of this VNTR to AD risk.

Discussion

Our analysis of both short and long-read sequencing datasets provides support for an association of a contracted VNTR in *RTEL1* with increased AD risk. The VNTR maps in an enhancer element, located in the 30th intron of the *RTEL1* gene, in the promoter region of *TNFRSF6B* gene and the 27th intron of fusion transcript *RTEL1-TNFRSF6B.*

RTEL1 (Regulator of Telomere Elongation Helicase 1) encodes a DNA helicase crucial for maintaining genomic stability, particularly at telomeres (Uringa et al. 2011, 2012). This protein is involved in DNA repair by resolving secondary DNA structures at promoter regions, such as G-quadruplexes and R-loops, facilitating DNA replication and transcription (Kotsantis et al. 2020). It is also required for transcriptional silencing, affecting epigenetic marks (Olivier et al. 2023). In short, *RTEL1* acts on different nuclear pathways which are implicated in cell homeostasis and ageing (Hourvitz, Awad, and Tzfati 2024). Given that AD is an ageing-related disorder, there is a heightened genomic instability caused by heterochromatin loss (J.-H. Lee et al. 2020; Villeponteau 1997), which may also cause dysregulation of the *RTEL1* gene which, in turn, may contribute to a faster decline with, for instance, the shortening of telomeres and increased DNA damage (Hourvitz, Awad, and Tzfati 2024; Smoom et al. 2023; Uringa et al. 2012). Therefore, the variation of VNTR length may influence the regulatory function of the enhancer, altering *RTEL1* gene expression levels. Gene expression changes may become more prominent upon age-related epigenomic modifications, ultimately contributing to AD pathogenesis. Further replication and functional validation studies are imperative to confirm the association of the *RTEL1*-VNTR variation to increased AD risk.

On the other hand, the *RTEL1*-VNTR is also located on the promoter of *TNFRSF6B* gene and in the fusion transcript *RTEL1-TNFRSF6B*. Even though the fusion transcript *RTEL1-TNFRSF6B* is not predicted to produce a functional protein, being a candidate for nonsense-mediated mRNA decay (NMD), a recent study has reported the transcript is a lncRNA that may modulate *RTEL1* expression. In the study, the authors identified a variant shared between *RTEL1* and *RTEL1-TNFRSF6B* associated with progressive myoclonus epilepsy, a neurodegenerative disorder (Chaudhari et al. 2024). Furthermore, *TNFRSF6B* (tumour necrosis factor receptor superfamily member 6B) is involved in cell cycle processes and immune responses, playing a significant role in cancer progression, particularly pancreatic cancer (Zhang et al. 2021). This gene has also been associated with AD, through its immunosuppressive role it ameliorates neuroinflammation, amyloid-β plaque deposition and cognitive deficits in AD mice models. Specifically, *TNFRSF6B* upregulation has a neuroprotective effect via modulating microglia activation and enhancing its amyloid-β clearance efficiency (Y.-L. Liu et al. 2017). This is in line with our finding that shorter *RTEL1*-VNTR lengths associates with AD, whilst

longer *RTEL1*-VNTR lengths correlates with the presence of SNP rs55765053 in *TNFRSF6B*, which in turn is in modest LD with the main protective SNP rs6742 (Bellenguez et al. 2022). Therefore, investigating the regulatory role of the VNTR in the promoter of *TNFRSF6B* in connection to AD progression warrants further investigation.

Besides the *RTEL1*-VNTR, other VNTRs associated with AD are in the genes *ANXA2* and *DLGAP2*, which have been highlighted in previous research in association with AD. *ANXA2*, member of a calcium-binding protein family involved in membrane trafficking and cell signalling, was shown to be linked to AD through interactions with Tau and Presenilin 1 (Astillero-Lopez et al. 2024; Bustos et al. 2017; Gauthier-Kemper et al. 2018). Whilst *DLGAP2*, a scaffold protein primarily expressed in the brain, plays a role in the regulation of neuronal synapses and has been associated with neuropsychiatric disorders (Hsieh et al. 2023; J.-M. Li et al. 2014; Rasmussen, Rasmussen, and Silahtaroglu 2017). Its association with age-related cognitive decline and AD has also been reported (Ouellette et al. 2020). Furthermore, the VNTR in *ADAMTSL2* gene, which is part of the overarching ADAM and ADAMTS proteins believed to play a role in various pathologies and have been mentioned in recent GWAS meta-analysis of AD (Bellenguez et al. 2022; Jansen et al. 2019; Kunkle et al. 2019; de Rojas et al. 2021; Schwartzentruber et al. 2021). However, none of these VNTRs were replicated in the independent dataset of long-read sequencing data, potentially due to the small sample size available at the time of the study.

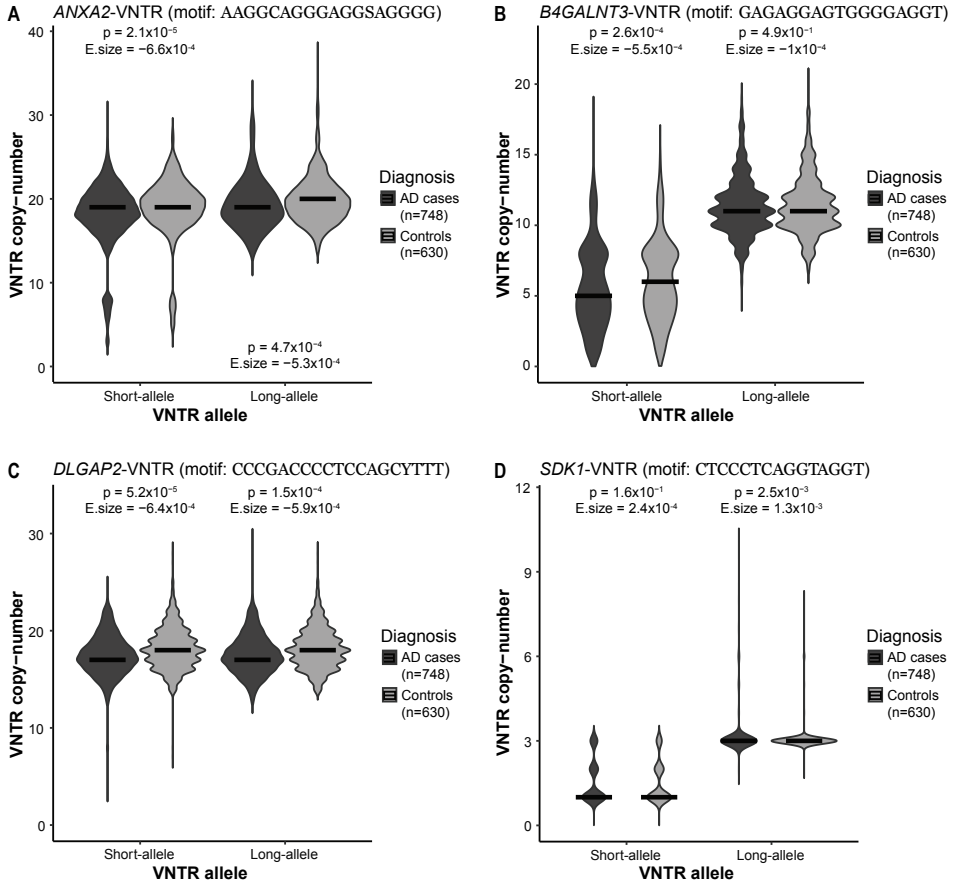
Two VNTR expansions were found to be associated with AD risk in an independent long-read sequencing dataset. The first one is the expanded VNTR in the lncRNA *LOC101928269*. Although this specific lncRNA has not been linked to AD, there is substantial evidence implicating lncRNAs in the Aβ cascade hypothesis (Hao et al. 2022; D. Li et al. 2021). The second gene is *B4GALNT3*, which belongs to a glycosyltransferase family. Glycosylation is an important form of post-translational modification that when disrupted it has been implicated in neurodegeneration, including AD (Moll, Shaw, and Cooper-Knock 2020; Schedin-Weiss, Winblad, and Tjernberg 2014). Despite the indirect association with AD, more research is needed to validate these findings.

The fact that ExpansionHunter Denovo and ExpansionHunter can only detect motifs up to 20 bp poses a limitation on the type of VNTRs that can be identified in our pipeline, leaving many VNTRs undetectable. This includes the *ABCA7*-VNTR, with a motif of 25 bp, that was previously associated with AD (De Roeck et al. 2018). Moreover, because EH needs exact genomic coordinates in the variant catalog, our pipeline cannot consider VNTRs that are absent from the reference genome, which may also have the potential to influence disease risk. Furthermore, ExpansionHunter only detects VNTR structures provided in the variant catalog, meaning that interruptions in the motif sequence and point mutations will affect the

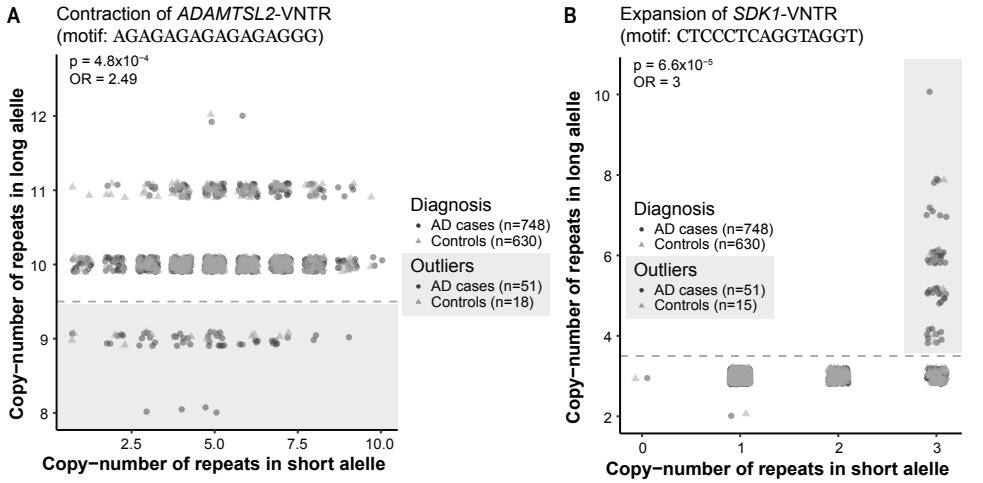
output. Thus, the challenge is to determine the structure and the possible variation of each VNTR. In our analysis, we partly addressed this limitation, i.e. for VNTRs whose motifs have the same length, have 80% sequence similarity, and are located within a 5 kb window, we created a consensus motif by introducing degeneracy with the IUPAC code. This does not fully overcome the issue since we do not account for indels, but it does improve the EH output. On the other hand, we are at risk of merging independent signals from distinct VNTRs with highly similar sequences, which may also impact downstream analysis.

Overall, investigating variation in long tandem repeats using short-read data is complex, independently of the computational tool used. This is mainly due to the intrinsic repetitive nature of tandem repeats and the length that exceeds the one of short reads which causes alignment errors. Therefore, long-read sequencing is an invaluable method to accurately sequence VNTRs and validate preliminary findings. Nevertheless, with our approach, and despite the limitations arising from the short-read sequencing and small sample sizes, we were able to identify various promising VNTRs. We provide preliminary evidence that a VNTR in the 30th intron of *RTEL1* gene and promoter of *TNFRSF6B* is of interest for further investigation as part of a regulatory element in AD risk.

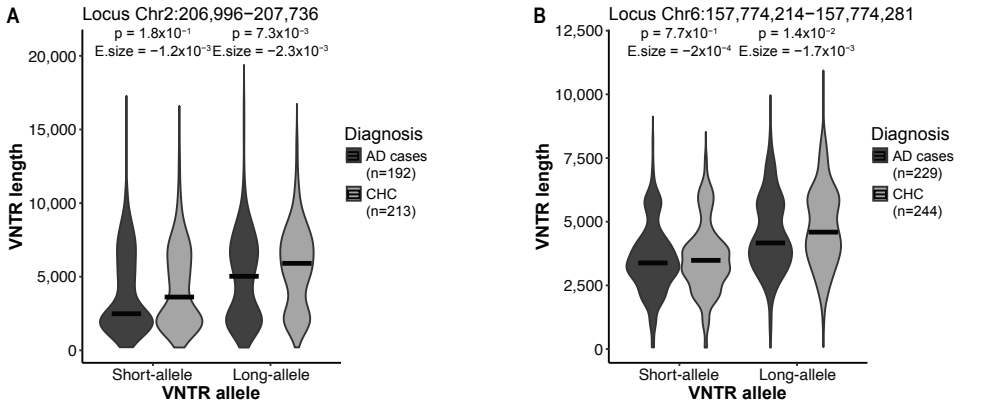
Supplemental Figures



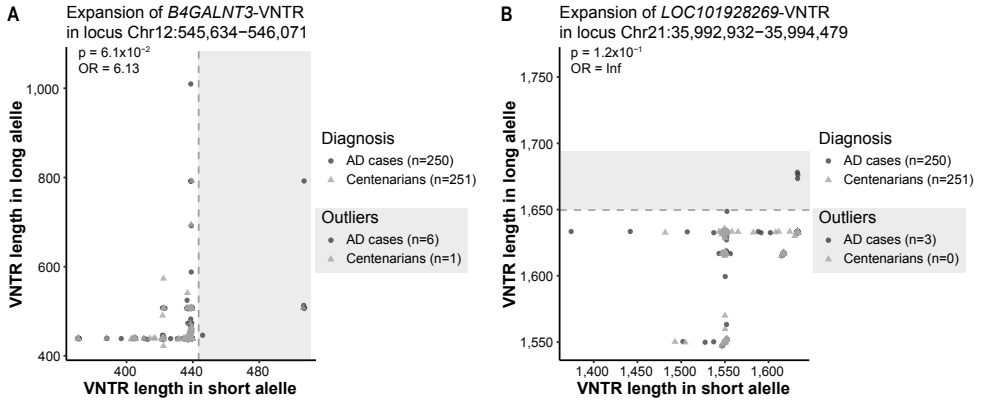
Supplemental figure S1: VNTRs significantly associated with AD from case-control analysis in the discovery phase. A-D, Case-control analysis of VNTR copy-numbers associated to Alzheimer's disease using logistic regression models, for both short and long alleles. Showing uncorrected *P*-values. The VNTRs are: *ANXA2*-VNTR (Chr15:60,349,890-60,350,142) (**A**), *B4GALNT3*-VNTR (Chr12:545,768-546,002) (**B**), *DLGAP2*-VNTR (Chr8:1,470,789-1,471,378) (**C**) and *SDK1*-VNTR (Chr 7:3,582,240-3,582,285) (**D**).



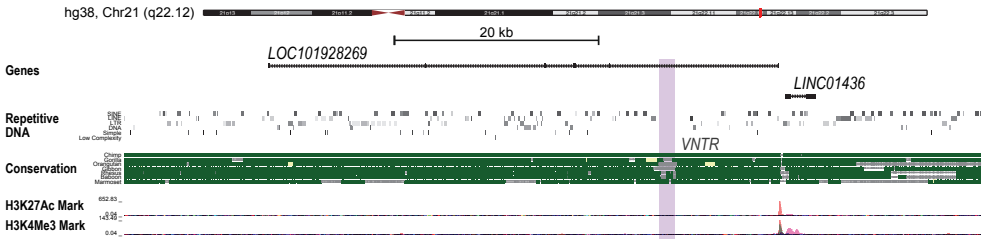
Supplemental figure S2: VNTRs significantly associated with AD from outlier analysis in discovery phase. A-B, Alzheimer's disease association of intronic VNTR contraction in *ADAMTSL2* (Chr9:133,564,168-133,564,642) (**A**) and expansion in *SK1* (Chr7:3,582,240-3,582,285) (**B**) present in a subset of cases with Fisher's exact test. Outliers are calculated with interquartile rule and are shown in gray shading. The outlier boundary is shown as a dashed gray line. Showing uncorrected *P*-values.



Supplemental figure S3: VNTRs significantly associated with AD from case-control analysis in the replication phase. A-B, Intergenic VNTRs in Chr2:206,996-207,736 (**A**) and Chr6:157,774,214-157,774,281 (**B**) significantly associated with Alzheimer's disease when tested with the long allele. Showing uncorrected *P*-values.



Supplemental figure S4: VNTRs significantly associated with AD from outlier analysis in the replication phase. A-B, Intronic *B4GALNT3*-VNTR (Chr12:545,634-546,071) (**A**) and *LOC101928269*-VNTR (Chr21:35,992,932-35,994,479) (**B**) showing expansions significantly associated with Alzheimer's disease. Outliers are calculated with interquartile rule and are shown in gray shading. The outlier boundary is shown as a dashed gray line. Showing uncorrected *P*-values.



Supplemental figure S5: Genetic landscape of lncRNA *LOC101928269*-VNTR. Architecture of *LOC101928269* and upstream gene, including repetitive DNA, a primate conservation track and histone marks (H3K27ac and H3K4me3). Lilac shading area highlights the location of the *LOC101928269*-VNTR. Adapted from the UCSC Genome Browser.

Methods

Study population and preprocessing

We used a sample of individuals from the Alzheimer's Disease Sequencing Project (ADSP). Additional information about the ADSP study design, AD diagnosis assessment as well as ethical committee approvals are publicly reported on the ADSP website (https://niagads.org/adsp). Sequence data from the ADSP is available by application to the NIA Genetics of Alzheimer's Disease Data Storage Site (NIAGADS) Data Sharing Service (DSS) (https://dss.niagads.org).

Initially we used a sample of 948 individuals, filtered the samples based on four aspects: (i) NIH racial category, to all be of the category 'Whites', (ii) origin, excluding samples from ADNI to avoid batch effects, (iii) family relationship, retaining 1 sample per family, (iv) phenotypes, retaining those labelled with "AD case" or "Control" were analysed. At a later stage, to increase statistical power to detect significant VNTRs, we increased the number of samples to 1,378 individuals. Here we filtered based on: (i) European ancestry (based on 1000 Genome clustering (Auton *et al.*, 2015)); (ii) family relationship (IBD > 0.2); (iii) phenotypes (AD case and healthy controls); (iiii) sequencing platform (Illumina HiSeqX) from three different sequencing centres: Broad, NYGC and USUHS; (iv) read length equal or higher than 149 bp. The paired-end sequencing CRAM files supplied by ADSP were converted to BAM format by aligning to the reference genome build GRCh38 using samtools (Li *et al.*, 2009).

Creation of the variant catalog

Identifying candidate VNTRs using ExpansionHunter Denovo

To identify candidate VNTRs without a predefined set of repeats, we run ExpansionHunter Denovo (EHdn) (v0.9.0) (Dolzhenko et al. 2020) on 948 individuals. We first generate an "STRprofile" (short tandem repeat profile) for each sample individually with the "profile" command using the default settings. This procedure takes a sample's aligned CRAM (or BAM) file and the reference genome as input, and outputs a file with all the identified in-repeat reads (IRRs) exceeding read length. Next, we create a manifest.tsv file, listing all samples together with their names (e.g. sample1), labels ("AD case" or "Control"), and the location of the relative STR file. We aggregate all sample's STR profiles using the "merge" command and the manifest.tsv file using the default settings (motif length between 2 and 20 bp). This command outputs a multi-sample STR profile containing the IRR counts of each sample, for each repeat region.

We run the case-control analysis using "casecontrol.py" executable with the "locus" option, taking the generated multi-sample STR profile and the manifest file as inputs. The "locus" option uses the anchored IRRs, identifying the repeats longer than read length but shorter than fragment length. This analysis compares the distribution of IRRs between AD cases and controls and outputs (i) the approximate location of the

VNTR on the reference genome, (ii) the repeated motif, (iii) the normalised IRR counts for each sample, (iv) the *P*-value from Wilcoxon Rank Sum test and (v) the Bonferroni corrected *P*-value. Note that the output motif is the lexicographically smallest repeat unit under circular permutations and reverse complement operations.

We run the outlier-based analysis using "outlier.py" executable with the "locus" option in a similar fashion. Compared to the case-control setting, this analysis does not compare distributions (thus does not report a *P*-value) but rather calculates the z-score for cases exceeding the mean of the compared cases+controls IRR count distribution. Instead of *P*-values, it outputs the highest z-score of a case sample and the IRR counts for cases with a z-score > 1.0.

Filtering candidate VNTRs

To focus on the most interesting repeats, we filtered the candidate VNTRs outputted by EHdn. For the case-control outputs, we selected those with an uncorrected *P*-value < 0.05 to select the repeats showing differences between cases and controls. For the outlier outputs, we selected the repeats where the top z-score for a case was ≥ 10 and a minimum of 7 cases (1%) had a z-score > 1 to select the common repeat expansions. For both outputs, we selected only the repeats with a motif size ≥ 7 as we specifically focus on VNTRs and exclude the repeats on sex- and decoy chromosomes. Only the repeats that remained after these filtering steps were converted into a variant catalog.

Converting candidate VNTRs into a variant catalog

To run ExpansionHunter (EH) (Dolzhenko et al. 2019) on the candidate VNTRs identified by EHdn, the repeats need to be encoded in a variant catalog. This catalog describes the start and end positions on the reference genome and which motif is repeated. The accuracy of the positions in the catalog greatly affects the quality of the estimates by EH. The positions can be off by a few base pairs, however, the approximate locations (± 500 bp) as outputted by EHdn are too imprecise to be used directly. Therefore, we first identified more precise start and end coordinates. To do so, for each candidate VNTR we extracted the corresponding sequence from the reference genome using the pysam package (https://github.com/pysam-developers/pysam) between the coordinates outputted by EHdn, extended by 500bp on each side. We then scan the obtained reference sequence for occurrences of the repeated motif, keeping track of perfect stretches of copies. Each perfect stretch of copies was extended if an adjacent imperfect copy was followed by another perfect copy of the motif. We assumed that the longest stretch in the reference is most likely to be the source of the repeat, thus we selected the longest stretch of copies. If there were multiple positions with the maximum number of copies, we could not establish which was the most likely source of the repeat. In such cases, we decided to be conservative and retain all positions.

In addition to defining more accurate start and end positions for the candidate VNTRs, an additional check on the repeated motif is done. EHdn outputs the lexicographically smallest version of the repeated motif, including reverse complement operations. This means it might be that the reverse complement is the motif that corresponds to the orientation of the reference. Therefore, for each VNTR, we scanned the extracted reference sequence for the reverse complement of the motif as well and again selected the longest stretch of copies. The re-established positions plus the selected motif were finally formatted, and outputted. The detailed algorithm can be found in **Supplemental Data** and the source code is available through GitHub (https://github.com/francesca-lucas/ehdn-to-eh).

The variant catalog was finalised in R (version 4.2.2), using the packages rjson (version 0.2.21), dplyr (version 1.1.2), e1071 (version 1.7.13) and seqinr (version 4.2.30). We took the union of both the outputs of the case-control and outlier analysis to get one variant catalog. Highly similar VNTRs within 5000 bp from each other were merged as likely being part of the same VNTR. VNTR were merged if the motif size was the same and the motif sequence similarity was ≥ 80% (based on hamming distance). Then, for each VNTR, we created a degenerate consensus sequence using the code from International Union of Pure and Applied Chemistry (IUPAC). For full script see Variant_catalog in **Supplemental Code**. Finally, the VNTRs in the variant catalog were classified as "novel" or "reported", depending on whether they had been reported before by: Audano et al. (2019), which reported SVs; Linthorst et al. (2020), which reported mostly VNTRs; or the tandem repeat catalog from PacBio, which reported mostly STRs (https://github.com/PacificBiosciences/trgt/blob/main/repeats/repeat_catalog.hg38.bed) (based on bedtools intersect (v2.30.0) (Quinlan and Hall 2010) with the options -wa -wb).

Identification of VNTRs associated with Alzheimer's Disease

Estimating number of repeats using ExpansionHunter

To get a more accurate and diploid estimate of the number of repeats for each candidate VNTR, we run EH (v5.0.0) on 3,274 individuals. We first extract the regions in the variant catalog with 5 kb padding from each sample, then we run EH with the variant catalog, the reference genome and the sample's CRAM (or BAM) files as input. For each repeat in the variant catalog, the algorithm will (i) collect the reads mapping to that repeat, and (ii) will determine the most probable repeat size for each allele using these reads. The output consists of a report for each sample containing copy number estimates for each repeat. These are diploid estimates, with a consensus value for each of the two alleles as well as the corresponding confidence intervals. For each sample we filtered out the repeats with the label "LowDepth" and kept the ones with "PASS". Then we count the number of repeats removed per sample and only the samples with ≤ 20 missing values were analysed. For these the diploid repeat size was divided into a shortest and longest allele, based on which has the least and the most copies, respectively. We test the association

of the VNTRs with AD risk using both alleles separately, then take the union of the VNTRs detected based on these separate allele values.

Case-control analysis with logistic regression model

We define case-control VNTRs as those that show a significantly different distribution for cases compared to controls. Since the ADSP cohort includes samples from multiple sequencing centres, it introduces variance to the dataset. Thus, we performed a principal components analysis (PCA) to explore sample stratification due to (i) type of sample (blood, brain or unknown), (ii) sequencing centre (Broad, NYGC or USUHS), (iii) DSS study number, (iv) CRAM/BAM file statistics (mean read length, mean and standard deviation insert size and coverage as calculated with samtools stats). We considered the top 7 PCs and added them as covariates to the models. Association models were in total corrected for population stratification (PC1-20), sequencing batch effect (PC1-7), and gender. We compare the repeat size distributions of 748 AD cases and 630 controls with logistic regression models, testing both the short and long alleles separately. Only repeats with < 0.1% missingness rate are analysed, which means only one sample is allowed not to have a repeat copy-number estimation per repeat. VNTRs with a false discovery rate (FDR) corrected *P*-value of < 0.2 are selected. See Discovery_phase in **Supplemental Code** for full script.

Identification of outliers with the interquartile rule

To detect VNTRs that may be expanded/contracted in a small subset of cases, we implement an outlier analysis next to the case-control analysis. To identify the outliers, we use the interquartile rule. This rule is based on the 25th percentile or first quartile (Q1), the 75th percentile or third quartile (Q3) and the interquartile range (IQR = Q3 – Q1). A datapoint x is considered an outlier when:

$$x < Q1 - c \,.\, IQR$$

or

$$x > Q3 + c \,.\, IQR$$

where c is a constant determining the sensitivity of these boundaries. The default value for c is 1.5, however, we use a more inclusive value for the constant: c = 1.0. We establish the outlier boundaries based on the distribution of the controls, as this is our null distribution against which we compare cases. This analysis is done for short and long alleles separately.

Outlier-based detection of AD-associated VNTRs

We define common pathogenic VNTRs as those that show an expansion or contraction in ≥ 1% of cases, which is 7 cases in our study. To detect expanded VNTRs, we count the number of datapoints above the upper boundary for both cases and controls. These outlier counts are captured in a contingency table, with outlier/non-outlier as rows and cases/controls status as columns. We use Fisher's exact

test on this contingency table to test the difference in outlier counts between cases and controls, using both the calculated *P*-values and odds ratio with confidence intervals. To detect contracted VNTRs, we count the number of datapoints below the lower boundary, and process these counts analogous to the expansions. As we are mainly interested in the VNTRs where AD cases show a clear expansion or contraction, we perform a one-sided test with the alternative hypothesis that cases are more likely to be outliers. We finally apply multiple testing correction to the significance values using FDR, and we consider the VNTRs which have an FDR < 0.20 as significant. Due to our cohort size, the repeats which show an expansion or contraction in < 2% of the cases will likely come up as insignificant after FDR correcting Fisher's significance values. This may exclude VNTRs that show a clear separation in repeat size and fall within our scope of common VNTRs (> 1% of cases). To detect these VNTRs (expanded/contracted in 1-2% of cases), we select the VNTRs where the outlier counts show an odds ratio > 5 and that occur in ≥ 1% of cases. See Discovery_phase in **Supplemental Code** for full script.

Genomic context

Gene annotation

To get an idea of where a VNTR might exert influence we provide genomic context. We annotate curated genes based on the RefSeq genes (v98) from NCBI (Pruitt et al. 2014; Pruitt, Tatusova, and Maglott 2005). This file contains multiple gene models for each gene. We first select the entry with the highest number of exons (exonCount). If there are multiple entries left, we select the entry with the widest transcription coordinates. After selecting one entry for each gene, we define the following regions:

- "transcription" from transcription start (txStart) to end (txEnd) coordinates;
- "coding" from coding start (cdsStart) to end (cdsEnd) coordinates;
- "exon" from each exon start (exonStarts) and its corresponding end (exonEnds) coordinates;
- "telomere" within 5Mb of the ends of chromosome arms;
- "promoter" as 5 kb upstream and 1 kb downstream of the transcription start (txStart).

A VNTR gets an annotation when it overlaps the defined regions. We define a VNTR to be "genic" if it overlaps with a region of "transcription". See Genomic_context in **Supplemental Code** for full script.

Functional element annotation

Non-genic functional elements are annotated using the NCBI RefSeq Functional Elements (Pruitt et al. 2014; Pruitt, Tatusova, and Maglott 2005) and the Registry of candidate cis-Regulatory Elements (cCREs) from ENCODE (Dunham et al. 2012; Moore et al. 2020). The NCBI RefSeq Functional Elements table contains various

functional elements that have been experimentally validated, such as regulatory elements, protein binding sites, mobile elements, recombination features and sequence features. The ENCODE cCREs table contains representative DNase hypersensitive sites from ENCODE and Roadmap Epigenomics samples with promoter- or enhancer-like signatures, when supported by H3K4me3 and H3K27ac histone modifications, respectively, or CTCF binding. All entries from both tables are included, the start (chromStart) and end (chromEnd) coordinates are used to define each region and the type of element (name). See Genomic_context in **Supplemental Code** for full script.

Gene set enrichment

To assess whether there is an over-representation of biological pathways in the genic VNTRs, we test for gene set enrichment. Gene set enrichment analysis was performed with the web application g:Profiler, using the g:GOSt analysis (biit.cs.ut.ee/gprofiler/gost). g:GOSt performs gene set enrichment analysis on an input gene list, mapping genes to functional terms and detecting the terms that are statistically significantly enriched in a background set consisting of all annotated genes. We selected the genic VNTRs in the variant catalog and used the g:GOSt settings including annotations from Gene Ontology sources: GO:MF (molecular function), GO:CC (cellular component), GO:BP (biological process).

Linkage analysis

Discovery phase

Linkage between VNTRs and SNPs was assessed with a Spearman rank correlation between SNP allele count and VNTR size. Only independent SNPs (MAF > 0.005) within a window of 1 mb of the VNTR were extracted with PLINK2 (parameters for LD-pruning were --indep 50 5 2) (Purcell et al. 2007) and SNP allele count was then tested with respect to the short and long alleles, separately (see Linkage_analysis in **Supplemental Code**). SNPs with a genome-wide association *P*-value < 0.05 from Bellenguez et al. (2022) were included as suggestive associations to AD.

RTEL1 locus

We analysed whether the SNP rs55765053, which correlates with the *RTEL1*-VNTR in our study, is correlated with the main protective GWAS SNP rs6742 reported by Bellenguez et al. (2022) using the online tool LDlink, specifically LDpair Tool (Machiela and Chanock 2015). The genome build is the GRCh38 High Coverage from the 1000 Genome Project, selecting only European ancestry.

Replication with PacBio dataset

To replicate the findings from the discovery phase, we used a long-read sequencing dataset of 250 AD cases from the Amsterdam Dementia Cohort (van der Flier and Scheltens 2018) and 251 cognitively healthy centenarians from the 100-plus Study

(Holstege et al. 2018). This dataset was generated with PacBio, details about the sequencing are available at Salazar et al. (2023). For genotyping of the VNTRs in the long-read datasets, the coordinates were adjusted to include the complete stretch of the repetitive DNA elements overlapped by our identified VNTRs (**Supplemental Table S10**). This means two repeats (chr9:28189272-28189688 and chr9:28189312-28189624) were merged (chr9:28189212-28189724). Also, one repeat located in a centromeric region (chr4:49153237-49156525), which showed an unusually high coverage, was excluded from the analysis. The genotype of the adjusted VNTRs was done with Treat (Tesi et al. 2024). Briefly, this tool genotypes tandem repeats using the size of the sequences in the regions of interest and a clustering framework. Haplotype-specific VNTR lengths were then extracted, combined, and used for association analysis. The association with AD was performed following the case-control and outlier analysis described previously, including the criteria for VNTRs considered as significantly associated with AD. Short and long alleles were tested separately. See Replication_phase in **Supplemental Code** for full script.

Visualisation of genomic loci

To assess the architecture of the strongest VNTR candidates in our study, we used the UCSC Genome Browser (Kent et al. 2002) with the following tracks: genes from GENCODE (v44) (Frankish et al. 2023) or NCBI RefSeq genes (Pruitt, Tatusova, and Maglott 2005); repetitive DNA from RepeatMasker (A. Smit, Hubley, and Green 1996), a primate conservation track from Multiz Alignments of 100 Vertebrates (Blanchette et al. 2004), histone marks (H3K27ac and H3K4me3) from ENCODE (Dunham et al. 2012), enhancers (grey) and promoters (red) and their interaction with annotated genes from GeneHancer (Fishilevich et al. 2017).

Acknowledgements

This work was supported by a grant from Alzheimer Nederland (WE.03-2018-07) to F.M.J. Jacobs, H. Holstege and M.J.T. Reinders. We thank Alex Salazar for his invaluable insights and the Evolutionary Neurogenomics Group for helpful discussions.

Author contributions

M.J.T.R., N.T., R.L.F.P.G. and F.L. conceptualized the project. M.H. aquired the dataset. N.T., R.L.F.P.G. and F.L developed the methodology. R.L.F.P.G. performed data analysis and vizualisation. F.M.J.J., M.J.T.R. and H.H. acquired funding for the study. R.L.F.P.G., F.L. and N.T wrote the manuscript with contributions of A.S., M.J.T.R., H.H. and F.M.J.J.. H.H. supervised the study.

d

General discussion

Repetitive DNA as a hidden layer of Alzheimer's genetic risk

This thesis addresses a key problem in the study of Alzheimer's disease (AD): the elusive causality of genetic variants. Despite extensive genome-wide association studies (GWAS), a significant portion of heritability remains unexplained, referred to as "missing heritability." One possibility is that causal variants are hidden within the repetitive parts of our DNA, which are traditionally difficult to study. Recent advances in sequencing technologies have allowed for a much more accurate representation of loci associated with disease, shedding light on structural variations of previously overlooked repetitive DNA (Audano et al. 2019; Chaisson, Wilson, and Eichler 2015; Ebert et al. 2021). While the involvement of repetitive DNA in neurological diseases has long been recognized (Chintalaphani et al. 2021; X. Liao et al. 2023), only recently research in complex disorders like AD has begun to focus on these elements (De Roeck et al. 2018; Larsen et al. 2017; Salazar et al. 2023). Compared to non-coding single nucleotide polymorphisms (SNPs), structural variants are increasingly considered as more likely to have a functional implication, thus making them promising candidates to form a connection between genetic variation and disease risk (Roses et al. 2016).

Mapping structural variants in Alzheimer's Disease loci

In **chapter 1**, we integrated structural variation into a fine-mapping analysis of AD risk loci, with a focus on recurrently identified AD-associated loci for which research has failed to identify a causal SNP or genetic variant. Our investigation revealed a high level of variation in both tandem repeats (TRs) and transposable elements (TEs) within these regions.

TRs were especially notable in the *ABCA7*, *APOE*, and *SCIMP* loci. TR variations in *ABCA7* arises from the variable number of tandem repeat (VNTR) associated with an increased risk to develop AD (De Roeck et al. 2018). Beyond the well-established ε2, ε3, and ε4 alleles of *APOE* (Saddiki et al. 2020), other repeat-based variants may act as genetic modifiers. Furthermore, *SCIMP* is associated with LOAD through various GWAS and meta-analysis (Bellenguez et al. 2022; de Rojas et al. 2021; Schwartzentruber et al. 2021). However, this association lacks strong evidence, which may be attributed to variations in TRs in the locus.

TEs, particularly SVA elements, were a shared feature across these highly replicated AD-associated loci. SVA elements are a great ape-specific class of TEs that remain mobile in modern humans and contributed to many insertional polymorphisms in the population (Lander et al. 2001). For AD-associated loci, polymorphic insertions are observed in *APOE* and *SCIMP*. However, even fixed SVAs, being present in every individual in the human population, often contain large structural variations in the VNTR part of the SVA sequence. Notably, structurally variable SVAs (SV-SVAs), fixed SVAs which themselves are a source for structural variants, were located in regions dense with SNPs strongly associated with AD risk. **Chapter 2** expands this analysis to Parkinson's disease (PD) and other neurological disorders, highlighting

the potential role of SVAs in neurodegenerative disorders. For instance, in PD and amyotrophic lateral sclerosis (ALS), SVA elements were identified in the promoter regions of disease-associated genes, suggesting their possible contribution to disease (Savage et al. 2013, 2014).

Another significant contributor to structural variations in the AD-associated loci is the primate-specific class of *Alu* elements. As a primate-specific TE with high retrotransposition activity (Feusier et al. 2019; Lander et al. 2001), numerous *Alu* insertions and deletions were found across AD-associated loci. Prior studies have linked *Alu* activity to deleterious effects in genes such as *ATP5J*, *TOMM40*, and *PSEN1* (Antonell et al. 2012; Boone et al. 2014; Larsen et al. 2017; Le Guennec et al. 2017) and a new *Alu* insertion in *TMEM106b*, a gene with no previously known causal variant, strengthens the case for their role in AD and other neurodegenerative disorders (Salazar et al. 2023).

Our study of structural variation in AD (and PD) aimed to gather evidence and establish the groundwork for functional assessment of the potentially causative repetitive DNA element. However, our characterisation of structural variants in disease-associated loci was limited, as it did not involve a disease-specific cohort for mapping of structural variants. A couple of years ago a new study has been published (not yet peer reviewed), making significant advances in that respect. The authors identified 400,234 structural variants in 16,905 individuals from an AD cohort, detecting ultra-rare structural variants in well-known AD genes, and potential risk structural variants in novel genes (H. Wang et al. 2023). Moreover, in the same year, a genome-wide analysis of structural variants was performed for PD, unravelling 3 novel structural variants associated with PD risk (Billingsley et al. 2023). Both these studies bring invaluable insights into the complex genetic landscape of neurodegenerative disorders.

Functional insights into transposable elements in relation to disease

Building on the observation that SVAs are present in disease-associated loci, we examined their gene-regulatory roles in **chapters 2** and **3**. We found that SV-SVAs are amongst the TE classes that become active in cortical organoids, which is in concordance with previous research (Bragg et al. 2017; Gianfrancesco, Bubb, and Quinn 2017; Savage et al. 2013, 2014). Furthermore, we showed that derepressed SVAs serve as strong enhancers and that this activity depends on the length of their VNTR, where structural variation often occurs. These VNTRs hold a high GC content (H. Wang et al. 2005), which increases the chance of forming G-quadruplexes (G4s), secondary structures known to have a strong potential to influence transcription (Kejnovsky, Tokan, and Lexa 2015; Rhodes and Lipps 2015). As such, G4s in SVAs are a major source of G4-mediated regulatory potential in the human genome (Gianfrancesco, Bubb, and Quinn 2017). Repressive proteins that interact with TEs, such as KRAB zinc-finger 91 (ZNF91), adds yet another layer of regulation, as it

specifically binds to the VNTR region of the SVA, inducing transcriptional repression of the element (Haring et al. 2021; Jacobs et al. 2014). In our experiments, ZNF91 strongly repressed SVA activity in a manner that varied depending on VNTR length. At present it's still unclear whether ZNF91 binds directly to G4s or blocks their formation, but the relationship between VNTR structure and gene regulation is now evident.

Given the differential gene-regulatory potential of SV-SVAs and their presence in AD and PD loci, in **chapter 2** we further investigated their correlation to disease risk. We found that variants with weaker enhancer properties are correlated with GWAS-identified risk SNPs associated with the disorders. Therefore, the susceptibility to disease may be caused by altered gene regulation through variations in length of the SVA elements. To elucidate the functional properties of the SVA elements in AD and PD loci, in **chapters 2** and **3** we used a combination of genetic engineering of human stem cells and cortical organoid technologies to perform a functional assessment of SV-SVAs in a human neuronal environment. We engineered 5 lines of human embryonic stem cells with SVA deletions in disease associated loci and generated cortical organoids. For two of the investigated loci there was no epigenomic or transcriptomic changes induced by the SV-SVA, which may be linked to the *in vitro* model used in the study, or it indicates that these structural variants are neutral and not associated with disease risk at all. For three of the investigated loci, RNA- and ChIP-sequencing revealed significant epigenomic and/or transcriptomic changes upon removal of SV-SVAs located nearby or within genes. These findings are in concordance with studies reporting SVA insertional polymorphisms associated with differential gene expression (Pfaff et al. 2021), or pathogenic SVA insertions within genes that alter the expression levels of downstream exons (Bragg et al. 2017). However, to what extent the structural variation in SVAs influence the expression of the associated genes in an AD or PD patient's brain remains to be elucidated.

Challenges of modelling transposable element regulation in neurodegeneration

Cortical organoids have been increasingly used in biomedical research as they mimic various features of the developing human brain, including cell-to-cell interactions, tissue organization and pathology (Lancaster et al. 2013). Recreating the complex structures and function of specific regions of the human brain is essential in the study of neurodevelopmental disorders, drug screening and disease modelling. However, a major limitation of this model is low reproducibility, due to intrinsic variability between organoids caused by cellular heterogeneity (Di Lullo and Kriegstein 2017; Yoon et al. 2019). Furthermore, cortical organoids are only composed by cells of neural origins, thus lacking vascular or immune cells, such as microglia (Qian, Song, and Ming 2019). Since microglial cells play a major role in AD pathology, there may be a cell-type specific mechanism at play in the regulation exerted by the SVAs that we do not detect in a neuronal environment.

Age is the primary risk factor for neurodegenerative disorders and with ageing several epigenetic mechanisms display progressive alterations (J.-H. Lee et al. 2020). This includes a progressively reduced repression of TEs, which is essential to repress the activity and transcription of TEs. Epigenetic silencing of TEs is maintained through mechanisms like heterochromatin formation, DNA methylation, histone modifications, and the binding of specific transcription factors. As these regulatory systems weaken with age, whether due to loss of heterochromatin or reduced binding of repressors such as ZNF91 for SVAs, TEs can become fully active (Wanhe Li et al. 2013; Rosenkrantz et al. 2024; Salazar et al. 2023). Because ageing-induced TE activation occurs only later in life, cortical organoids derived from healthy donors may not be ideal for studying a disease mechanism that has such a strong age-associated risk factor. A more promising approach may be to use AD-specific cortical organoids developed using human-induced pluripotent stem cells (iPSCs) derived from EOAD patients (Gonzalez et al. 2018; Raja et al. 2016), or by introducing EOAD-associated mutations (Choi et al. 2014). Even though much of the age-associated epigenetic signature is lost during the conversion of blood cells or fibroblasts to iPSCs, these studies have shown that such models replicate key AD pathologies, including amyloid aggregation and tau hyperphosphorylation. Since tau pathology has been shown to contribute to TE activation (Evering et al. 2023; Guo et al. 2018; Singh, Borkar, and Bhatt 2024), using AD patients-derived organoids may better mimic the biological context in which TEs become active, making them a valuable tool for investigating the role of TEs in this age-related disorder.

Preliminary insights into AD-associated VNTRs

In **chapter 4** we conducted a genome-wide detection of VNTRs using a short-read sequencing dataset of AD cases and controls and found 50 VNTRs with a suggestive association to AD risk. Among the VNTRs identified, the most compelling signal was a VNTR located in the 30th intron of *RTEL1*. This repeat is particularly interesting due to its genomic context. It overlaps not only an enhancer element in the *RTEL1* gene, but also the promoter region of *TNFRSF6B* and the fusion transcript *RTEL1-TNFRSF6B*. Shorter alleles of this VNTR were more frequent in AD cases, suggesting a potential regulatory role in disease risk. *RTEL1* (Regulator of Telomere Elongation Helicase 1) encodes a DNA helicase essential for maintaining genomic integrity, specifically at telomeres (Hourvitz, Awad, and Tzfati 2024). It is also involved in DNA repair, replication and transcription by unwinding secondary structures, primarily G4s, at promoter regions (Kotsantis et al. 2020). The location of this VNTR within a regulatory region that interacts with the promoter of *RTEL1* suggests it may influence *RTEL1* expression. This is aligned with previous work showing that TRs can function as transcription enhancers, with their activity modulated by repeat length (Marshall et al. 2021; Usdin 2008).

As AD is closely tied to ageing, there is an increased vulnerability to genomic instability due to the loss of heterochromatin (J.-H. Lee et al. 2020; Villeponteau 1997). This may disrupt the regulation of *RTEL1*, potentially accelerating disease progression via mechanisms such as telomere shortening and increased DNA damage (Hourvitz, Awad, and Tzfati 2024; Smoom et al. 2023; Uringa et al. 2012). While we did not assess the functional impact of VNTR variation in our study, future work should focus on testing how different VNTR lengths affect *RTEL1* gene regulation. For example, luciferase reporter assays could be used to explore how changes in VNTR length may affect gene expression in human-derived neuroblastoma cell lines. The element of ageing could be simulated by inducing cellular senescence through DNA damage (Di Micco et al. 2021).

This VNTR may also impact *TNFRSF6B*, a gene involved in modulating immune responses. *TNFRSF6B* encodes a decoy receptor shown to reduce neuroinflammation, promote amyloid-β clearance, and improve cognition in AD mouse models (Y.-L. Liu et al. 2017). In our data, shorter VNTR alleles were linked to AD, while longer alleles correlated with rs55765053, a SNP located in *TNFRSF6B* that is in modest linkage disequilibrium with the lead protective variant rs6742 (Bellenguez et al. 2022). This suggests a possible mechanism in which VNTR length variation influences *TNFRSF6B* expression and, in turn, AD susceptibility. Thus, to further investigate this, CRISPR-edited iPSCs-derived microglial cells carrying different VNTR alleles could be used in targeted RNA-seq experiments, combined with cytokine secretion assays or immune challenge experiments (e.g., stimulation with amyloid-β peptides) to evaluate whether VNTR-driven changes in *TNFRSF6B* expression modulate microglial immune responses.

An important limitation of short-read data is that it cannot be directly used to assess the size of VNTRs, when the VNTR exceeds the size of the DNA-sequencing reads that were used. However, indirectly, we used the short-read data to infer the size of the VNTR based on the relative coverage of DNA sequencing reads on any typical VNTR in our genome. The results obtained from this analysis were further validated using a small long-read sequencing dataset, which confirmed the AD-associated VNTR in *RTEL1* gene. While our current findings remain preliminary, our analysis proves that short-read data can be used to predict and estimate contractions or expansions of VNTRs in our genomes. This is only an estimation, while long-read sequencing data can accurately genotype the VNTRs. However, currently the sample size of individuals for which long-read sequencing data is available is not yet large enough to allow for an accurate genetic association study. At present, new developments are in place to gather a dataset of AD patients and centenarians of considerable sample size for the study of risk and protective variants.

Concluding remarks

This thesis provides substantial support for the conception that repetitive DNA, particularly VNTRs and SVA elements, should be considered promising candidates for being genetic contributors to the development and progression of neurodegenerative diseases such as Alzheimer's disease. During this study, pioneering advances have been made that enabled the characterization of repetitive DNA elements, either focussing on previous GWAS-identified loci to AD or unravelling completely novel loci. The next step is to pair larger long-read sequencing datasets with functional assays to understand how these genetic elements regulate genes in time- and cell-type-specific ways, and to what extent structural variations causes differential gene regulation between healthy or disease-affected individuals in the human population. Ultimately, the novel insights reflected in this thesis - that the repetitive regions of our DNA should not be ignored, but instead their structure, functionality and polymorphisms should be carefully studied in relation to human disease - could reshape our understanding of the complex genetics underlying neurodegenerative diseases like AD.

a

Appendix

References
Nederlandse Samenvatting
Publications
Acknowledgements
About the Author

References

Afgan, Enis, Dannon Baker, Bérénice Batut, Marius van den Beek, Dave Bouvier, Martin Čech, John Chilton, et al. 2018. 'The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2018 Update'. *Nucleic Acids Research* 46(Web Server issue): W537–44. doi:10.1093/nar/gky379.

Ahmadi, Amirhossein, Ilario De Toma, Natàlia Vilor-Tejedor, Mohammad Reza Eftekhariyan Ghamsari, and Iman Sadeghi. 2020. 'Transposable Elements in Brain Health and Disease'. *Ageing Research Reviews* 64: 101153. doi:10.1016/j.arr.2020.101153.

Akbarian, Schahram, Chunyu Liu, James A Knowles, Flora M Vaccarino, Peggy J Farnham, Gregory E Crawford, Andrew E Jaffe, et al. 2015. 'The PsychENCODE Project'. *Nature neuroscience* 18(12): 1707–12. doi:10.1038/nn.4156.

Ali, Rouchan, Ghanshyam Das Gupta, and Pooja A. Chawla. 2022. 'Aducanumab: A New Hope in Alzheimer's Disease'. *Health Sciences Review* 4: 100039. doi:10.1016/j.hsr.2022.100039.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. 'Basic Local Alignment Search Tool'. *Journal of Molecular Biology* 215(3): 403–10. doi:10.1016/S0022-2836(05)80360-2.

Ando, Kunie, Siranjeevi Nagaraj, Fahri Küçükali, Marie-Ange de Fisenne, Andreea-Claudia Kosa, Emilie Doeraene, Lidia Lopez Gutierrez, Jean-Pierre Brion, and Karelle Leroy. 2022. 'PICALM and Alzheimer's Disease: An Update and Perspectives'. *Cells* 11(24): 3994. doi:10.3390/cells11243994.

Andrade-Guerrero, Jesús, Alberto Santiago-Balmaseda, Paola Jeronimo-Aguilar, Isaac Vargas-Rodríguez, Ana Ruth Cadena-Suárez, Carlos Sánchez-Garibay, Glustein Pozo-Molina, et al. 2023. 'Alzheimer's Disease: An Updated Overview of Its Genetics'. *International Journal of Molecular Sciences* 24(4): 3754. doi:10.3390/ijms24043754.

Aneichyk, Tatsiana, William T. Hendriks, Rachita Yadav, David Shin, Dadi Gao, Christine A. Vaine, Ryan L. Collins, et al. 2018. 'Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly'. *Cell* 172(5): 897-909.e21. doi:10.1016/j.cell.2018.02.011.

Antonell, Anna, Ellen Gelpi, Raquel Sánchez-Valle, Ramiro Martínez, José L. Molinuevo, and Albert Lladó. 2012. 'Breakpoint Sequence Analysis of an AβPP Locus Duplication Associated with Autosomal Dominant Alzheimer's Disease and Severe Cerebral Amyloid Angiopathy'. *Journal of Alzheimer's Disease* 28(2): 303–8. doi:10.3233/JAD-2011-110911.

Arnold, Matthias, Johannes Raffler, Arne Pfeufer, Karsten Suhre, and Gabi Kastenmüller. 2015. 'SNiPA: An Interactive, Genetic Variant-Centered Annotation Browser'. *Bioinformatics* 31(8): 1334–36. doi:10.1093/bioinformatics/btu779.

Ashford, J. Wesson, and James A. Mortimer. 2002. 'Non-Familial Alzheimer's Disease Is Mainly Due to Genetic Factors'. *Journal of Alzheimer's Disease* 4(3): 169–77. doi:10.3233/JAD-2002-4307.

Astillero-Lopez, Veronica, Sandra Villar-Conde, Melania Gonzalez-Rodriguez, Alicia Flores-Cuadrado, Isabel Ubeda-Banon, Daniel Saiz-Sanchez, and Alino Martinez-Marcos. 2024. 'Proteomic Analysis Identifies HSP90AA1, PTK2B, and ANXA2 in the Human Entorhinal Cortex in Alzheimer's Disease: Potential Role in Synaptic Homeostasis and Aβ Pathology through Microglial and Astroglial Cells'. *Brain Pathology (Zurich, Switzerland)*: e13235. doi:10.1111/bpa.13235.

Audano, Peter A., Arvis Sulovari, Tina A. Graves-Lindsay, Stuart Cantsilieris, Melanie Sorensen, AnneMarie E. Welch, Max L. Dougherty, et al. 2019. 'Characterizing the Major Structural Variant Alleles of the Human Genome'. *Cell* 176(3): 663-675.e19. doi:10.1016/j.cell.2018.12.019.

Auton, Adam, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, et al. 2015. 'A Global Reference for Human Genetic Variation'. *Nature* 526(7571): 68–74. doi:10.1038/nature15393.

Ayodele, Temitope, Ekaterina Rogaeva, Jiji T. Kurup, Gary Beecham, and Christiane Reitz. 2021. 'Early-Onset Alzheimer's Disease: What Is Missing in Research?' *Current Neurology and Neuroscience Reports* 21(2): 4. doi:10.1007/s11910-020-01090-y.

Backman, Joshua D., Alexander H. Li, Anthony Marcketta, Dylan Sun, Joelle Mbatchou, Michael D. Kessler, Christian Benner, et al. 2021. 'Exome Sequencing and Analysis of 454,787 UK Biobank Participants'. *Nature* 599(7886): 628–34. doi:10.1038/s41586-021-04103-z.

Baillie, J. Kenneth, Mark W. Barnett, Kyle R. Upton, Daniel J. Gerhardt, Todd A. Richmond, Fioravante De Sapio, Paul M. Brennan, et al. 2011. 'Somatic Retrotransposition Alters the Genetic Landscape of the Human Brain'. *Nature* 479(7374): 534–37. doi:10.1038/nature10531.

Bakhtiari, Mehrdad, Jonghun Park, Yuan-Chun Ding, Sharona Shleizer-Burko, Susan L. Neuhausen, Bjarni V. Halldórsson, Kári Stefánsson, Melissa Gymrek, and Vineet Bafna. 2021. 'Variable Number Tandem Repeats Mediate the Expression of Proximal Genes'. *Nature Communications* 12(1): 2075. doi:10.1038/s41467-021-22206-z.

Batzer, M A, V A Gudi, J C Mena, D W Foltz, R J Herrera, and P L Deininger. 1991. 'Amplification Dynamics of Human-Specific (HS) Alu Family Members.' *Nucleic Acids Research* 19(13): 3619–23.

Beecham, Gary W., Kara Hamilton, Adam C. Naj, Eden R. Martin, Matt Huentelman, Amanda J. Myers, Jason J. Corneveaux, et al. 2014. 'Genome-Wide Association Meta-Analysis of Neuropathologic Features of Alzheimer's Disease and Related Dementias'. *PLoS Genetics* 10(9): e1004606. doi:10.1371/journal.pgen.1004606.

Bellenguez, Céline, Fahri Küçükali, Iris E. Jansen, Luca Kleineidam, Sonia Moreno-Grau, Najaf Amin, Adam C. Naj, et al. 2022. 'New Insights into the Ge-

netic Etiology of Alzheimer's Disease and Related Dementias'. *Nature Genetics* 54(4): 412–36. doi:10.1038/s41588-022-01024-z.

Bertram, Lars, Christoph Lange, Kristina Mullin, Michele Parkinson, Monica Hsiao, Meghan F. Hogan, Brit M.M. Schjeide, et al. 2008. 'Genome-Wide Association Analysis Reveals Putative Alzheimer's Disease Susceptibility Loci in Addition to APOE'. *American Journal of Human Genetics* 83(5): 623–32. doi:10.1016/j.ajhg.2008.10.008.

Bertram, Lars, Matthew B. McQueen, Kristina Mullin, Deborah Blacker, and Rudolph E. Tanzi. 2007. 'Systematic Meta-Analyses of Alzheimer Disease Genetic Association Studies: The AlzGene Database'. *Nature Genetics* 39(1): 17–23. doi:10.1038/ng1934.

Billingsley, Kimberley J., Jinhui Ding, Pilar Alvarez Jerez, Anastasia Illarionova, Kristin Levine, Francis P. Grenn, Mary B. Makarious, et al. 2023. 'Genome-Wide Analysis of Structural Variants in Parkinson Disease'. *Annals of Neurology* 93(5): 1012–22. doi:10.1002/ana.26608.

Bis, Joshua C., Xueqiu Jian, Brian W. Kunkle, Yuning Chen, Kara L. Hamilton-Nelson, William S. Bush, William J. Salerno, et al. 2020. 'Whole Exome Sequencing Study Identifies Novel Rare and Common Alzheimer's-Associated Variants Involved in Immune Response and Transcriptional Regulation'. *Molecular Psychiatry* 25(8): 1859–75. doi:10.1038/s41380-018-0112-7.

Blanchette, Mathieu, W. James Kent, Cathy Riemer, Laura Elnitski, Arian F. A. Smit, Krishna M. Roskin, Robert Baertsch, et al. 2004. 'Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner'. *Genome Research* 14(4): 708–15. doi:10.1101/gr.1933104.

bliek, tijs, Marc Galland, franskloet, and Jihed Chouaref. 2021. 'BleekerLab/Snakemake_rnaseq: Release v0.3.7'. doi:10.5281/zenodo.4707140.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. 'Trimmomatic: A Flexible Trimmer for Illumina Sequence Data'. *Bioinformatics (Oxford, England)* 30(15): 2114–20. doi:10.1093/bioinformatics/btu170.

Boone, Philip M., Bo Yuan, Ian M. Campbell, Jennifer C. Scull, Marjorie A. Withers, Brett C. Baggett, Christine R. Beck, et al. 2014. 'The Alu-Rich Genomic Architecture of SPAST Predisposes to Diverse and Functionally Distinct Disease-Associated CNV Alleles'. *American Journal of Human Genetics* 95(2): 143–61. doi:10.1016/j.ajhg.2014.06.014.

Boyle, Alan P., Eurie L. Hong, Manoj Hariharan, Yong Cheng, Marc A. Schaub, Maya Kasowski, Konrad J. Karczewski, et al. 2012. 'Annotation of Functional Variation in Personal Genomes Using RegulomeDB'. *Genome Research* 22(9): 1790–97. doi:10.1101/gr.137323.112.

Bragg, D. Cristopher, Kotchaphorn Mangkalaphiban, Christine A. Vaine, Nichita J. Kulkarni, David Shin, Rachita Yadav, Jyotsna Dhakal, et al. 2017. '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* 114(51): E11020–28. doi:10.1073/pnas.1712526114.

van Bree, Elisabeth J., Rita L.F.P. Guimarães, Mischa Lundberg, Elena R. Blujdea, Jimi L. Rosenkrantz, Fred T.G. White, Josse Poppinga, et al. 2022. 'A Hidden Layer of Structural Variation in Transposable Elements Reveals Potential Genetic Modifiers in Human Disease-Risk Loci'. *Genome Research* 32(4): 656–70. doi:10.1101/gr.275515.121.

Broekema, R. V., O. B. Bakker, and I. H. Jonkers. 2020. 'A Practical View of Fine-Mapping and Gene Prioritization in the Post-Genome-Wide Association Era'. *Open Biology* 10(1): 190221. doi:10.1098/rsob.190221.

Brookes, K. J. 2013. 'The VNTR in Complex Disorders: The Forgotten Polymorphisms? A Functional Way Forward?' *Genomics* 101(5): 273–81. doi:10.1016/j.ygeno.2013.03.003.

Brouha, Brook, Joshua Schustak, Richard M. Badge, Sheila Lutz-Prigge, Alexander H. Farley, John V. Moran, and Haig H. Kazazian. 2003. 'Hot L1s Account for the Bulk of Retrotransposition in the Human Population'. *Proceedings of the National Academy of Sciences of the United States of America* 100(9): 5280–85. doi:10.1073/pnas.0831042100.

Busche, Marc Aurel, and Bradley T. Hyman. 2020. 'Synergy between Amyloid-β and Tau in Alzheimer's Disease'. *Nature Neuroscience* 23(10): 1183–93. doi:10.1038/s41593-020-0687-6.

Bush, W., and J. Moore. 2012. 'Education Chapter 11 : Genome-Wide Association Studies'. https://www.semanticscholar.org/paper/Education-Chapter-11-%3A-Genome-Wide-Association-Bush-Moore/afe370768da2420ab07adcb69c227337454a3820 (December 12, 2024).

Bustos, Victor, Maria V. Pulina, Ashley Bispo, Alison Lam, Marc Flajolet, Fred S. Gorelick, and Paul Greengard. 2017. 'Phosphorylated Presenilin 1 Decreases β-Amyloid by Facilitating Autophagosome–Lysosome Fusion'. *Proceedings of the National Academy of Sciences of the United States of America* 114(27): 7148–53. doi:10.1073/pnas.1705240114.

Cacace, Rita, Kristel Sleegers, and Christine Van Broeckhoven. 2016. 'Molecular Genetics of Early-Onset Alzheimer's Disease Revisited'. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 12(6): 733–48. doi:10.1016/j.jalz.2016.01.012.

Carlson, Scott M., Cameron M. Soulette, Ze Yang, Joshua E. Elias, Angela N. Brooks, and Or Gozani. 2017. 'RBM25 Is a Global Splicing Factor Promoting Inclusion of Alternatively Spliced Exons and Is Itself Regulated by Lysine Mono-Methylation'. *The Journal of Biological Chemistry* 292(32): 13381–90. doi:10.1074/jbc.M117.784371.

Carrasquillo, Minerva M., Olivia Belbin, Talisha A. Hunter, Li Ma, Gina D. Bisceglio, Fanggeng Zou, Julia E. Crook, et al. 2011. 'Replication of BIN1 Association with Alzheimer's Disease and Evaluation of Genetic Interactions'. *Journal of Alzheimer's disease : JAD* 24(4): 751–58. doi:10.3233/JAD-2011-101932.

Chaisson, Mark J. P., Ashley D. Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J. Gardner, et al. 2019. 'Multi-Platform Discovery of Haplotype-Resolved Structural Variation in Human Genomes'. *Nature Communications* 10(1): 1784. doi:10.1038/s41467-018-08148-z.

Chaisson, Mark J. P., Richard K. Wilson, and Evan E. Eichler. 2015. 'Genetic Variation and the de Novo Assembly of Human Genomes'. *Nature reviews. Genetics* 16(11): 627–40. doi:10.1038/nrg3933.

Chang, Diana, Mike A. Nalls, Ingileif B. Hallgrímsdóttir, Julie Hunkapiller, Marcel van der Brug, Fang Cai, International Parkinson's Disease Genomics Consortium, et al. 2017. 'A Meta-Analysis of Genome-Wide Association Studies Identifies 17 New Parkinson's Disease Risk Loci'. *Nature Genetics* 49(10): 1511–16. doi:10.1038/ng.3955.

Chapuis, J., F. Hansmannel, M. Gistelinck, A. Mounier, C. Van Cauwenberghe, K. V. Kolen, F. Geller, et al. 2013. 'Increased Expression of BIN1 Mediates Alzheimer Genetic Risk by Modulating Tau Pathology'. *Molecular Psychiatry* 18(11): 1225–34. doi:10.1038/mp.2013.1.

Chaudhari, Sima, Lavanya Prakash Acharya, Dushyanth Babu Jasti, Akshay Pramod Ware, Sankar Prasad Gorthi, and Kapaettu Satyamoorthy. 2024. 'Discovery of a Novel Shared Variant Among RTEL1 Gene and RTEL1-TNFRSF6B lncRNA at Chromosome 20q13.33 in Familial Progressive Myoclonus Epilepsy'. *International Journal of Genomics* 2024: 7518528. doi:10.1155/2024/7518528.

Chen, Yanbo, Jianwen Deng, Peng Wang, Mengxue Yang, Xiaoping Chen, Li Zhu, Jianghong Liu, et al. 2016. 'PINK1 and Parkin Are Genetic Modifiers for FUS-Induced Neurodegeneration'. *Human Molecular Genetics* 25(23): 5059–68. doi:10.1093/hmg/ddw310.

Chintalaphani, Sanjog R., Sandy S. Pineda, Ira W. Deveson, and Kishore R. Kumar. 2021. 'An Update on the Neurological Short Tandem Repeat Expansion Disorders and the Emergence of Long-Read Sequencing Diagnostics'. *Acta Neuropathologica Communications* 9(1): 98. doi:10.1186/s40478-021-01201-x.

Choi, Se Hoon, Young Hye Kim, Matthias Hebisch, Christopher Sliwinski, Seungkyu Lee, Carla D'Avanzo, Hechao Chen, et al. 2014. 'A Three-Dimensional Human Neural Cell Culture Model of Alzheimer's Disease'. *Nature* 515(7526): 274–78. doi:10.1038/nature13800.

Chuong, Edward B., Nels C. Elde, and Cédric Feschotte. 2016. 'Regulatory Evolution of Innate Immunity through Co-Option of Endogenous Retroviruses'. *Science (New York, N.Y.)* 351(6277): 1083–87. doi:10.1126/science.aad5497.

Chuong, Edward B., Nels C. Elde, and Cédric Feschotte. 2017. 'Regulatory Activities of Transposable Elements: From Conflicts to Benefits'. *Nature reviews. Genetics* 18(2): 71–86. doi:10.1038/nrg.2016.139.

Cock, Peter J. A., Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, et al. 2009. 'Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics'. *Bioinformatics (Oxford, England)* 25(11): 1422–23. doi:10.1093/bioinformatics/btp163.

Collins, Mahlon, Yang Li, and Robert Bowser. 2020. 'RBM45 Associates with Nuclear Stress Bodies and Forms Nuclear Inclusions during Chronic Cellular

Stress and in Neurodegenerative Diseases'. *Acta Neuropathologica Communications* 8(1): 91. doi:10.1186/s40478-020-00965-y.

Collins, Ryan L., Harrison Brand, Konrad J. Karczewski, Xuefang Zhao, Jessica Alföldi, Laurent C. Francioli, Amit V. Khera, et al. 2020. 'A Structural Variation Reference for Medical and Population Genetics'. *Nature* 581(7809): 444–51. doi:10.1038/s41586-020-2287-8.

Colonna, Marco, and Oleg Butovsky. 2017. 'Microglia Function in the Central Nervous System During Health and Neurodegeneration'. *Annual review of immunology* 35: 441–68. doi:10.1146/annurev-immunol-051116-052358.

Concordet, Jean-Paul, and Maximilian Haeussler. 2018. 'CRISPOR: Intuitive Guide Selection for CRISPR/Cas9 Genome Editing Experiments and Screens'. *Nucleic Acids Research* 46(W1): W242–45. doi:10.1093/nar/gky354.

Cordaux, Richard, and Mark A. Batzer. 2009. 'The Impact of Retrotransposons on Human Genome Evolution'. *Nature reviews. Genetics* 10(10): 691–703. doi:10.1038/nrg2640.

Corder, E. H., A. M. Saunders, N. J. Risch, W. J. Strittmatter, D. E. Schmechel, P. C. Gaskell, J. B. Rimmler, et al. 1994. 'Protective Effect of Apolipoprotein E Type 2 Allele for Late Onset Alzheimer Disease'. *Nature Genetics* 7(2): 180–84. doi:10.1038/ng0694-180.

Coufal, Nicole G, José L Garcia-Perez, Grace E Peng, Gene W Yeo, Yangling Mu, Michael T Lovci, Maria Morell, et al. 2009. 'L1 Retrotransposition in Human Neural Progenitor Cells'. *Nature* 460(7259): 1127–31. doi:10.1038/nature08248.

Cummings, Jeffrey, Yadi Zhou, Garam Lee, Kate Zhong, Jorge Fonseca, and Feixiong Cheng. 2024. 'Alzheimer's Disease Drug Development Pipeline: 2024'. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* 10(2): e12465. doi:10.1002/trc2.12465.

Das Sarma, Jayasri. 2014. 'Microglia-Mediated Neuroinflammation Is an Amplifier of Virus-Induced Neuropathology'. *Journal of NeuroVirology* 20(2): 122–36. doi:10.1007/s13365-013-0188-4.

Davalos, Dimitrios, Jaime Grutzendler, Guang Yang, Jiyun V. Kim, Yi Zuo, Steffen Jung, Dan R. Littman, Michael L. Dustin, and Wen-Biao Gan. 2005. 'ATP Mediates Rapid Microglial Response to Local Brain Injury in Vivo'. *Nature Neuroscience* 8(6): 752–58. doi:10.1038/nn1472.

Dawson, Tyson, Uzma Rentia, Jessie Sanford, Carlos Cruchaga, John S. K. Kauwe, and Keith A. Crandall. 2023. 'Locus Specific Endogenous Retroviral Expression Associated with Alzheimer's Disease'. *Frontiers in Aging Neuroscience* 15: 1186470. doi:10.3389/fnagi.2023.1186470.

De Cecco, Marco, Steven W. Criscione, Edward J. Peckham, Sara Hillenmeyer, Eliza A. Hamm, Jayameenakshi Manivannan, Abigail L. Peterson, et al. 2013. 'Genomes of Replicatively Senescent Cells Undergo Global Epigenetic Changes Leading to Gene Silencing and Activation of Transposable Elements'. *Aging Cell* 12(2): 247–56. doi:10.1111/acel.12047.

De Deyn, Lara, and Kristel Sleegers. 2025. 'The Impact of Rare Genetic Variants on Alzheimer Disease'. *Nature Reviews Neurology* 21(3): 127–39. doi:10.1038/s41582-025-01062-1.

De Roeck, Arne, Lena Duchateau, Jasper Van Dongen, Rita Cacace, Maria Bjerke, Tobi Van den Bossche, Patrick Cras, et al. 2018. 'An Intronic VNTR Affects Splicing of ABCA7 and Increases Risk of Alzheimer's Disease'. *Acta Neuropathologica* 135(6): 827–37. doi:10.1007/s00401-018-1841-z.

De Strooper, Bart, Paul Saftig, Katleen Craessaerts, Hugo Vanderstichele, Gundula Guhde, Wim Annaert, Kurt Von Figura, and Fred Van Leuven. 1998. 'Deficiency of Presenilin-1 Inhibits the Normal Cleavage of Amyloid Precursor Protein'. *Nature* 391(6665): 387–90. doi:10.1038/34910.

Deininger, Prescott L, John V Moran, Mark A Batzer, and Haig H Kazazian. 2003. 'Mobile Elements and Mammalian Genome Evolution'. *Current Opinion in Genetics & Development* 13(6): 651–58. doi:10.1016/j.gde.2003.10.013.

Dembny, Paul, Andrew G. Newman, Manvendra Singh, Michael Hinz, Michal Szczepek, Christina Krüger, Robert Adalbert, et al. 2020. 'Human Endogenous Retrovirus HERV-K(HML-2) RNA Causes Neurodegeneration through Toll-like Receptors'. *JCI Insight* 5(7). doi:10.1172/jci.insight.131093.

Deming, Yuetiva, Fabia Filipello, Francesca Cignarella, Claudia Cantoni, Simon Hsu, Robert Mikesell, Zeran Li, et al. 2019. 'The MS4A Gene Cluster Is a Key Modulator of Soluble TREM2 and Alzheimer Disease Risk'. *Science translational medicine* 11(505): eaau2291. doi:10.1126/scitranslmed.aau2291.

Deng, Hao, Kai Gao, and Joseph Jankovic. 2014. 'The Role of FUS Gene Variants in Neurodegenerative Diseases'. *Nature Reviews. Neurology* 10(6): 337–48. doi:10.1038/nrneurol.2014.78.

Di Lullo, Elizabeth, and Arnold R. Kriegstein. 2017. 'The Use of Brain Organoids to Investigate Neural Development and Disease'. *Nature reviews. Neuroscience* 18(10): 573–84. doi:10.1038/nrn.2017.107.

Di Micco, Raffaella, Valery Krizhanovsky, Darren Baker, and Fabrizio d'Adda di Fagagna. 2021. 'Cellular Senescence in Ageing: From Mechanisms to Therapeutic Opportunities'. *Nature Reviews Molecular Cell Biology* 22(2): 75–95. doi:10.1038/s41580-020-00314-w.

Dolzhenko, Egor, Mark F. Bennett, Phillip A. Richmond, Brett Trost, Sai Chen, Joke J. F. A. van Vugt, Charlotte Nguyen, et al. 2020. 'ExpansionHunter Denovo: A Computational Method for Locating Known and Novel Repeat Expansions in Short-Read Sequencing Data'. *Genome Biology* 21: 102. doi:10.1186/s13059-020-02017-z.

Dolzhenko, Egor, Viraj Deshpande, Felix Schlesinger, Peter Krusche, Roman Petrovski, Sai Chen, Dorothea Emig-Agius, et al. 2019. 'ExpansionHunter: A Sequence-Graph-Based Tool to Analyze Variation in Short Tandem Repeat Regions'. *Bioinformatics* 35(22): 4754–56. doi:10.1093/bioinformatics/btz431.

Dourlen, Pierre, Devrim Kilinc, Nicolas Malmanche, Julien Chapuis, and Jean-Charles Lambert. 2019. 'The New Genetic Landscape of Alzheimer's Disease: From Amyloid Cascade to Genetically Driven Synaptic Failure Hypothesis?' *Acta*

Neuropathologica 138(2): 221–36. doi:10.1007/s00401-019-02004-0.

Douville, Renée, Jiankai Liu, Jeffrey Rothstein, and Avindra Nath. 2011. 'Identification of Active Loci of a Human Endogenous Retrovirus in Neurons of Patients with Amyotrophic Lateral Sclerosis'. *Annals of neurology* 69(1): 141–51. doi:10.1002/ana.22149.

Dunham, Ian, Anshul Kundaje, Shelley F. Aldred, Patrick J. Collins, Carrie A. Davis, Francis Doyle, Charles B. Epstein, et al. 2012. 'An Integrated Encyclopedia of DNA Elements in the Human Genome'. *Nature* 489(7414): 57–74. doi:10.1038/nature11247.

Duyckaerts, Charles, Benoît Delatour, and Marie-Claude Potier. 2009. 'Classification and Basic Pathology of Alzheimer Disease'. *Acta Neuropathologica* 118(1): 5–36. doi:10.1007/s00401-009-0532-1.

van Dyck, Christopher H., Chad J. Swanson, Paul Aisen, Randall J. Bateman, Christopher Chen, Michelle Gee, Michio Kanekiyo, et al. 2023. 'Lecanemab in Early Alzheimer's Disease'. *The New England Journal of Medicine* 388(1): 9–21. doi:10.1056/NEJMoa2212948.

Ebert, Peter, Peter A. Audano, Qihui Zhu, Bernardo Rodriguez-Martin, David Porubsky, Marc Jan Bonder, Arvis Sulovari, et al. 2021. 'Haplotype-Resolved Diverse Human Genomes and Integrated Analysis of Structural Variation'. *Science* 372(6537): eabf7117. doi:10.1126/science.abf7117.

Edwards, Stacey L., Jonathan Beesley, Juliet D. French, and Alison M. Dunning. 2013. 'Beyond GWASs: Illuminating the Dark Road from Association to Function'. *American Journal of Human Genetics* 93(5): 779–97. doi:10.1016/j.ajhg.2013.10.012.

Eichler, Evan E. 2019. 'Genetic Variation, Comparative Genomics, and the Diagnosis of Disease'. *The New England journal of medicine* 381(1): 64–74. doi:10.1056/NEJMra1809315.

Eiraku, Mototsugu, Kiichi Watanabe, Mami Matsuo-Takasaki, Masako Kawada, Shigenobu Yonemura, Michiru Matsumura, Takafumi Wataya, et al. 2008. 'Self-Organized Formation of Polarized Cortical Tissues from ESCs and Its Active Manipulation by Extrinsic Signals'. *Cell Stem Cell* 3(5): 519–32. doi:10.1016/j.stem.2008.09.002.

Ernst, Jason, and Manolis Kellis. 2012. 'ChromHMM: Automating Chromatin State Discovery and Characterization'. *Nature methods* 9(3): 215–16. doi:10.1038/nmeth.1906.

Evering, Teresa H., Jez L. Marston, Li Gan, and Douglas F. Nixon. 2023. 'Transposable Elements and Alzheimer's Disease Pathogenesis'. *Trends in Neurosciences* 46(3): 170–72. doi:10.1016/j.tins.2022.12.003.

Ewing, Adam D., Nathan Smits, Francisco J. Sanchez-Luque, Jamila Faivre, Paul M. Brennan, Sandra R. Richardson, Seth W. Cheetham, and Geoffrey J. Faulkner. 2020. 'Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling'. *Molecular Cell* 80(5): 915-928.e5. doi:10.1016/j.molcel.2020.10.024.

Fernandes, Jason D., Armando Zamudio-Hurtado, Hiram Clawson, W. James Kent, David Haussler, Sofie R. Salama, and Maximilian Haeussler. 2020. 'The UCSC Repeat Browser Allows Discovery and Visualization of Evolutionary Conflict across Repeat Families'. *Mobile DNA* 11: 13. doi:10.1186/s13100-020-00208-w.

Fernández, Sara Gutiérrez, Cristina Gan Oria, Dieter Petit, Wim Annaert, John M. Ringman, Nick C. Fox, Natalie S. Ryan, and Lucía Chávez-Gutiérrez. 2025. 'Spectrum of γ-Secretase Dysfunction as a Unifying Predictor of ADAD Age at Onset across PSEN1, PSEN2 and APP Causal Genes'. *Molecular Neurodegeneration* 20(1): 48. doi:10.1186/s13024-025-00832-1.

Feusier, Julie, W. Scott Watkins, Jainy Thomas, Andrew Farrell, David J. Witherspoon, Lisa Baird, Hongseok Ha, Jinchuan Xing, and Lynn B. Jorde. 2019. 'Pedigree-Based Estimation of Human Mobile Element Retrotransposition Rates'. *Genome Research* 29(10): 1567–77. doi:10.1101/gr.247965.118.

Field, Andrew R., Frank M.J. Jacobs, Ian T. Fiddes, Alex P.R. Phillips, Andrea M. Reyes-Ortiz, Erin LaMontagne, Lila Whitehead, et al. 2019. 'Structurally Conserved Primate LncRNAs Are Transiently Expressed during Human Cortical Differentiation and Influence Cell-Type-Specific Genes'. *Stem Cell Reports* 12(2): 245–57. doi:10.1016/j.stemcr.2018.12.006.

Fishilevich, Simon, Ron Nudel, Noa Rappaport, Rotem Hadar, Inbar Plaschkes, Tsippi Iny Stein, Naomi Rosen, et al. 2017. 'GeneHancer: Genome-Wide Integration of Enhancers and Target Genes in GeneCards'. *Database: The Journal of Biological Databases and Curation* 2017: bax028. doi:10.1093/database/bax028.

van der Flier, Wiesje M., and Philip Scheltens. 2018. 'Amsterdam Dementia Cohort: Performing Research to Optimize Care'. *Journal of Alzheimer's Disease* 62(3): 1091–1111. doi:10.3233/JAD-170850.

Fondon, John W., Elizabeth A. D. Hammock, Anthony J. Hannan, and David G. King. 2008. 'Simple Sequence Repeats: Genetic Modulators of Brain Function and Behavior'. *Trends in Neurosciences* 31(7): 328–34. doi:10.1016/j.tins.2008.03.006.

Frank, Oliver, Michelle Giehl, Chun Zheng, Rüdiger Hehlmann, Christine Leib-Mösch, and Wolfgang Seifarth. 2005. 'Human Endogenous Retrovirus Expression Profiles in Samples from Brains of Patients with Schizophrenia and Bipolar Disorders'. *Journal of Virology* 79(17): 10890–901. doi:10.1128/JVI.79.17.10890-10901.2005.

Frankish, Adam, Sílvia Carbonell-Sala, Mark Diekhans, Irwin Jungreis, Jane E Loveland, Jonathan M Mudge, Cristina Sisu, et al. 2023. 'GENCODE: Reference Annotation for the Human and Mouse Genomes in 2023'. *Nucleic Acids Research* 51(D1): D942–49. doi:10.1093/nar/gkac1071.

Fuentes, Daniel R., Tomek Swigut, and Joanna Wysocka. 2018. 'Systematic Perturbation of Retroviral LTRs Reveals Widespread Long-Range Effects on Human Gene Regulation'. *eLife* 7: e35989. doi:10.7554/eLife.35989.

Gao, Peirong, Lingqi Ye, Hongrong Cheng, and Honglei Li. 2021. 'The Mechanistic Role of Bridging Integrator 1 (BIN1) in Alzheimer's Disease'. *Cellular and*

Molecular Neurobiology 41(7): 1431–40. doi:10.1007/s10571-020-00926-y.

García-Montojo, Marta, Belén de la Hera, Jezabel Varadé, Ana de la Encarnación, Iris Camacho, María Domínguez-Mozo, Ana Arias-Leal, et al. 2014. 'HERV-W Polymorphism in Chromosome X Is Associated with Multiple Sclerosis Risk and with Differential Expression of MSRV'. *Retrovirology* 11(1): 2. doi:10.1186/1742-4690-11-2.

Gatz, Margaret, Chandra A. Reynolds, Laura Fratiglioni, Boo Johansson, James A. Mortimer, Stig Berg, Amy Fiske, and Nancy L. Pedersen. 2006. 'Role of Genes and Environments for Explaining Alzheimer Disease'. *Archives of General Psychiatry* 63(2): 168–74. doi:10.1001/archpsyc.63.2.168.

Gauthier-Kemper, Anne, María Suárez Alonso, Frederik Sündermann, Benedikt Niewidok, Maria-Pilar Fernandez, Lidia Bakota, Jürgen Josef Heinisch, and Roland Brandt. 2018. 'Annexins A2 and A6 Interact with the Extreme N Terminus of Tau and Thereby Contribute to Tau's Axonal Localization'. *The Journal of Biological Chemistry* 293(21): 8065–76. doi:10.1074/jbc.RA117.000490.

Gemayel, Rita, Marcelo D. Vinces, Matthieu Legendre, and Kevin J. Verstrepen. 2010. 'Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences'. *Annual Review of Genetics* 44(1): 445–77. doi:10.1146/annurev-genet-072610-155046.

Gendron, Tania F., and Leonard Petrucelli. 2018. 'Disease Mechanisms of C9ORF72 Repeat Expansions'. *Cold Spring Harbor Perspectives in Medicine* 8(4): a024224. doi:10.1101/cshperspect.a024224.

Gianfrancesco, Olympia, Vivien J. Bubb, and John P. Quinn. 2017. 'SVA Retrotransposons as Potential Modulators of Neuropeptide Gene Expression'. *Neuropeptides* 64: 3–7. doi:10.1016/j.npep.2016.09.006.

Gianfrancesco, Olympia, Bethany Geary, Abigail L. Savage, Kimberley J. Billingsley, Vivien J. Bubb, and John P. Quinn. 2019. 'The Role of SINE-VNTR-Alu (SVA) Retrotransposons in Shaping the Human Genome'. *International Journal of Molecular Sciences* 20(23): 5977. doi:10.3390/ijms20235977.

Giusti-Rodríguez, Paola, Leina Lu, Yuchen Yang, Cheynna A. Crowley, Xiaoxiao Liu, Ivan Juric, Joshua S. Martin, et al. 2019. 'Using Three-Dimensional Regulatory Chromatin Interactions from Adult and Fetal Cortex to Interpret Genetic Results for Psychiatric Disorders and Cognitive Traits'. : 406330. doi:10.1101/406330.

Glenner, George G., and Caine W. Wong. 1984. 'Alzheimer's Disease: Initial Report of the Purification and Characterization of a Novel Cerebrovascular Amyloid Protein'. *Biochemical and Biophysical Research Communications* 120(3): 885–90. doi:10.1016/S0006-291X(84)80190-4.

Gloss, Brian S., and Marcel E. Dinger. 2018. 'Realizing the Significance of Noncoding Functionality in Clinical Genomics'. *Experimental & Molecular Medicine* 50(8): 1–8. doi:10.1038/s12276-018-0087-0.

Goecks, Jeremy, Anton Nekrutenko, James Taylor, and Galaxy Team. 2010. 'Galaxy: A Comprehensive Approach for Supporting Accessible, Reproducible, and

Transparent Computational Research in the Life Sciences'. *Genome Biology* 11(8): R86. doi:10.1186/gb-2010-11-8-r86.

Gonzalez, Cesar, Enrique Armijo, Javiera Bravo-Alegria, Andrea Becerra-Calixto, Charles E. Mays, and Claudio Soto. 2018. 'Modeling Amyloid Beta and Tau Pathology in Human Cerebral Organoids'. *Molecular psychiatry* 23(12): 2363–74. doi:10.1038/s41380-018-0229-8.

Grenn, Francis P., Jonggeol J. Kim, Mary B. Makarious, Hirotaka Iwaki, Anastasia Illarionova, Kajsa Brolin, Jillian H. Kluss, et al. 2020. 'The Parkinson's Disease Genome-Wide Association Study Locus Browser'. *Movement Disorders: Official Journal of the Movement Disorder Society* 35(11): 2056–67. doi:10.1002/mds.28197.

Gu, Jian-lan, and Fei Liu. 2020. 'Tau in Alzheimer's Disease: Pathological Alterations and an Attractive Therapeutic Target'. *Current Medical Science* 40(6): 1009–21. doi:10.1007/s11596-020-2282-1.

Guo, Caiwei, Hyun-Hwan Jeong, Yi-Chen Hsieh, Hans-Ulrich Klein, David A. Bennett, Philip L. De Jager, Zhandong Liu, and Joshua M. Shulman. 2018. 'Tau Activates Transposable Elements in Alzheimer's Disease'. *Cell reports* 23(10): 2874–80. doi:10.1016/j.celrep.2018.05.004.

Haass, Christian, and Dennis J. Selkoe. 2007. 'Soluble Protein Oligomers in Neurodegeneration: Lessons from the Alzheimer's Amyloid β-Peptide'. *Nature Reviews Molecular Cell Biology* 8(2): 101–12. doi:10.1038/nrm2101.

Hamza, Taye H., Cyrus P. Zabetian, Albert Tenesa, Alain Laederach, Jennifer Montimurro, Dora Yearout, Denise M. Kay, et al. 2010. 'Common Genetic Variation in the HLA Region Is Associated with Late-Onset Sporadic Parkinson's Disease'. *Nature Genetics* 42(9): 781–85. doi:10.1038/ng.642.

Hancks, Dustin C., Adam D. Ewing, Jesse E. Chen, Katsushi Tokunaga, and Haig H. Kazazian. 2009. 'Exon-Trapping Mediated by the Human Retrotransposon SVA'. *Genome Research* 19(11): 1983–91. doi:10.1101/gr.093153.109.

Hancks, Dustin C., and Haig Kazazian. 2010. 'SVA Retrotransposons: Evolution and Genetic Instability'. *Seminars in cancer biology* 20(4): 234–45. doi:10.1016/j.semcancer.2010.04.001.

Hancks, Dustin C., and Haig H. Kazazian. 2016. 'Roles for Retrotransposon Insertions in Human Disease'. *Mobile DNA* 7: 9. doi:10.1186/s13100-016-0065-9.

Hao, Yitong, Bo Xie, Xiaoshu Fu, Rong Xu, and Yu Yang. 2022. 'New Insights into lncRNAs in Aβ Cascade Hypothesis of Alzheimer's Disease'. *Biomolecules* 12(12): 1802. doi:10.3390/biom12121802.

Haring, Nina L., Elisabeth J. van Bree, Whitney S. Jordaan, Judith R. E. Roels, Gonzalo Congrains Sotomayor, Tiziana M. Hey, Fred T. G. White, et al. 2021. 'ZNF91 Deletion in Human Embryonic Stem Cells Leads to Ectopic Activation of SVA Retrotransposons and Up-Regulation of KRAB Zinc Finger Gene Clusters'. *Genome Research* 31(4): 551–63. doi:10.1101/gr.265348.120.

Harold, Denise, Richard Abraham, Paul Hollingworth, Rebecca Sims, Amy Gerrish, Marian Hamshere, Jaspreet Singh Pahwa, et al. 2009. 'Genome-Wide As-

sociation Study Identifies Variants at CLU and PICALM Associated with Alzheimer's Disease, and Shows Evidence for Additional Susceptibility Genes'. *Nature genetics* 41(10): 1088–93. doi:10.1038/ng.440.

Heckman, Michael G., Koji Kasanuki, Nancy N. Diehl, Shunsuke Koga, Alexandra Soto, Melissa E. Murray, Dennis W. Dickson, and Owen A. Ross. 2017. 'Parkinson's Disease Susceptibility Variants and Severity of Lewy Body Pathology'. *Parkinsonism & Related Disorders* 44: 79–84. doi:10.1016/j.parkreldis.2017.09.009.

Hollingworth, Paul, Denise Harold, Rebecca Sims, Amy Gerrish, Jean-Charles Lambert, Minerva M Carrasquillo, Richard Abraham, et al. 2011. 'Common Variants in ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP Are Associated with Alzheimer's Disease'. *Nature genetics* 43(5): 429–35. doi:10.1038/ng.803.

Holstege, Henne, Nina Beker, Tjitske Dijkstra, Karlijn Pieterse, Elizabeth Wemmenhove, Kimja Schouten, Linette Thiessens, et al. 2018. 'The 100-plus Study of Cognitively Healthy Centenarians: Rationale, Design and Cohort Description'. *European Journal of Epidemiology* 33(12): 1229–49. doi:10.1007/s10654-018-0451-3.

Holstege, Henne, Marc Hulsman, Camille Charbonnier, Benjamin Grenier-Boley, Olivier Quenez, Detelina Grozeva, Jeroen G. J. van Rooij, et al. 2022. 'Exome Sequencing Identifies Rare Damaging Variants in ATP8B4 and ABCA1 as Risk Factors for Alzheimer's Disease'. *Nature Genetics* 54(12): 1786–94. doi:10.1038/s41588-022-01208-7.

Hourvitz, Noa, Aya Awad, and Yehuda Tzfati. 2024. 'The Many Faces of the Helicase RTEL1 at Telomeres and Beyond'. *Trends in Cell Biology* 34(2): 109–21. doi:10.1016/j.tcb.2023.07.002.

Hsieh, Ming-Yen, Li-Heng Tuan, Ho-Ching Chang, Yu-Chun Wang, Chia-Hsiang Chen, Horng-Tzer Shy, Li-Jen Lee, and Susan Shur-Fen Gau. 2023. 'Altered Synaptic Protein Expression, Aberrant Spine Morphology, and Impaired Spatial Memory in Dlgap2 Mutant Mice, a Genetic Model of Autism Spectrum Disorder'. *Cerebral Cortex* 33(8): 4779–93. doi:10.1093/cercor/bhac379.

Hu, Xiaolan, Eve Pickering, Yingxue Cathy Liu, Stephanie Hall, Helene Fournier, Elyse Katz, Bryan Dechairo, et al. 2011. 'Meta-Analysis for Genome-Wide Association Study Identifies Multiple Variants at the BIN1 Locus Associated with Late-Onset Alzheimer's Disease'. *PloS One* 6(2): e16616. doi:10.1371/journal.pone.0016616.

Hutchinson, Anna, Jennifer Asimit, and Chris Wallace. 2020. 'Fine-Mapping Genetic Associations'. *Human Molecular Genetics* 29(R1): R81–88. doi:10.1093/hmg/ddaa148.

Jacob-Hirsch, Jasmine, Eran Eyal, Binyamin A Knisbacher, Jonathan Roth, Karen Cesarkas, Chen Dor, Sarit Farage-Barhom, et al. 2018. 'Whole-Genome Sequencing Reveals Principles of Brain Retrotransposition in Neurodevelopmental Disorders'. *Cell Research* 28(2): 187–203. doi:10.1038/cr.2018.8.

Jacobs, Frank M. J., David Greenberg, Ngan Nguyen, Maximilian Haeussler, Adam D. Ewing, Sol Katzman, Benedict Paten, Sofie R. Salama, and David Haussler. 2014. 'An Evolutionary Arms Race between KRAB Zinc-Finger Genes ZNF91/93 and SVA/L1 Retrotransposons'. *Nature* 516(7530): 242–45. doi:10.1038/nature13760.

Jankovic, J., S. Chen, and W. D. Le. 2005. 'The Role of Nurr1 in the Development of Dopaminergic Neurons and Parkinson's Disease'. *Progress in Neurobiology* 77(1–2): 128–38. doi:10.1016/j.pneurobio.2005.09.001.

Jansen, Iris E, Jeanne E Savage, Kyoko Watanabe, Julien Bryois, Dylan M Williams, Stacy Steinberg, Julia Sealock, et al. 2019. 'Genome-Wide Meta-Analysis Identifies New Loci and Functional Pathways Influencing Alzheimer's Disease Risk'. *Nature genetics* 51(3): 404–13. doi:10.1038/s41588-018-0311-9.

Jeffreys, Alec J., Victoria Wilson, and Swee Lay Thein. 1985. 'Hypervariable "Minisatellite" Regions in Human DNA'. *Nature* 314(6006): 67–73. doi:10.1038/314067a0.

Jun, Gyungah, Carla A. Ibrahim-Verbaas, Maria Vronskaya, Jean-Charles Lambert, Jaeyoon Chung, Adam C. Naj, Brian W. Kunkle, et al. 2016. 'A NOVEL ALZHEIMER DISEASE LOCUS LOCATED NEAR THE GENE ENCODING TAU PROTEIN'. *Molecular psychiatry* 21(1): 108–17. doi:10.1038/mp.2015.23.

Kapusta, Aurélie, Zev Kronenberg, Vincent J. Lynch, Xiaoyu Zhuo, LeeAnn Ramsay, Guillaume Bourque, Mark Yandell, and Cédric Feschotte. 2013. 'Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs'. *PLoS genetics* 9(4): e1003470. doi:10.1371/journal.pgen.1003470.

Kazazian, Haig H., Corinne Wong, Hagop Youssoufian, Alan F. Scott, Deborah G. Phillips, and Stylianos E. Antonarakis. 1988. 'Haemophilia A Resulting from de Novo Insertion of L1 Sequences Represents a Novel Mechanism for Mutation in Man'. *Nature* 332(6160): 164–66. doi:10.1038/332164a0.

Kejnovsky, Eduard, Viktor Tokan, and Matej Lexa. 2015. 'Transposable Elements and G-Quadruplexes'. *Chromosome Research* 23(3): 615–23. doi:10.1007/s10577-015-9491-7.

Kelleher, Raymond J., and Jie Shen. 2017. 'Presenilin-1 Mutations and Alzheimer's Disease'. *Proceedings of the National Academy of Sciences of the United States of America* 114(4): 629–31. doi:10.1073/pnas.1619574114.

Kent, W. James, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and and David Haussler. 2002. 'The Human Genome Browser at UCSC'. *Genome Research* 12(6): 996–1006. doi:10.1101/gr.229102.

Kiełbasa, Szymon M., Raymond Wan, Kengo Sato, Paul Horton, and Martin C. Frith. 2011. 'Adaptive Seeds Tame Genomic Sequence Comparison'. *Genome Research* 21(3): 487–93. doi:10.1101/gr.113985.110.

Kim, Daehwan, Joseph M. Paggi, Chanhee Park, Christopher Bennett, and Steven L. Salzberg. 2019. 'Graph-Based Genome Alignment and Genotyping

with HISAT2 and HISAT-Genotype'. *Nature Biotechnology* 37(8): 907–15. doi:10.1038/s41587-019-0201-4.

Kinney, Jefferson W., Shane M. Bemiller, Andrew S. Murtishaw, Amanda M. Leisgang, Arnold M. Salazar, and Bruce T. Lamb. 2018. 'Inflammation as a Central Mechanism in Alzheimer's Disease'. *Alzheimer's & Dementia : Translational Research & Clinical Interventions* 4: 575–90. doi:10.1016/j.trci.2018.06.014.

Kircher, Martin, Daniela M. Witten, Preti Jain, Brian J. O'Roak, Gregory M. Cooper, and Jay Shendure. 2014. 'A General Framework for Estimating the Relative Pathogenicity of Human Genetic Variants'. *Nature genetics* 46(3): 310–15. doi:10.1038/ng.2892.

Kotsantis, Panagiotis, Sandra Segura-Bayona, Pol Margalef, Paulina Marzec, Phil Ruis, Graeme Hewitt, Roberto Bellelli, et al. 2020. 'RTEL1 Regulates G4/R-Loops to Avert Replication-Transcription Collisions'. *Cell Reports* 33(12): 108546. doi:10.1016/j.celrep.2020.108546.

Kundaje, Anshul, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Pouya Kheradpour, Zhizhuo Zhang, et al. 2015. 'Integrative Analysis of 111 Reference Human Epigenomes'. *Nature* 518(7539): 317–30. doi:10.1038/nature14248.

Kunkle, Brian W., Benjamin Grenier-Boley, Rebecca Sims, Joshua C. Bis, Vincent Damotte, Adam C. Naj, Anne Boland, et al. 2019. 'Genetic Meta-Analysis of Diagnosed Alzheimer's Disease Identifies New Risk Loci and Implicates Aβ, Tau, Immunity and Lipid Processing'. *Nature genetics* 51(3): 414–30. doi:10.1038/s41588-019-0358-2.

Kwiatkowski, T. J., D. A. Bosco, A. L. Leclerc, E. Tamrazian, C. R. Vanderburg, C. Russ, A. Davis, et al. 2009. 'Mutations in the FUS/TLS Gene on Chromosome 16 Cause Familial Amyotrophic Lateral Sclerosis'. *Science (New York, N.Y.)* 323(5918): 1205–8. doi:10.1126/science.1166066.

Labun, Kornel, Tessa G. Montague, James A. Gagnon, Summer B. Thyme, and Eivind Valen. 2016. 'CHOPCHOP v2: A Web Tool for the next Generation of CRISPR Genome Engineering'. *Nucleic Acids Research* 44(W1): W272–76. doi:10.1093/nar/gkw398.

Lambert, Erwan, Orthis Saha, Bruna Soares Landeira, Ana Raquel Melo de Farias, Xavier Hermant, Arnaud Carrier, Alexandre Pelletier, et al. 2022. 'The Alzheimer Susceptibility Gene BIN1 Induces Isoform-Dependent Neurotoxicity through Early Endosome Defects'. *Acta Neuropathologica Communications* 10(1): 4. doi:10.1186/s40478-021-01285-5.

Lambert, Jean-Charles, Simon Heath, Gael Even, Dominique Campion, Kristel Sleegers, Mikko Hiltunen, Onofre Combarros, et al. 2009. 'Genome-Wide Association Study Identifies Variants at CLU and CR1 Associated with Alzheimer's Disease'. *Nature Genetics* 41(10): 1094–99. doi:10.1038/ng.439.

Lambert, Jean-Charles, Carla A. Ibrahim-Verbaas, Denise Harold, Adam C. Naj, Rebecca Sims, Céline Bellenguez, Gyungah Jun, et al. 2013. 'Meta-Analysis of 74,046 Individuals Identifies 11 New Susceptibility Loci for Alzheimer's Disease'. *Nature Genetics* 45(12): 1452–58. doi:10.1038/ng.2802.

Lambert, Jean-Charles, Diana Zelenika, Mikko Hiltunen, Vincent Chouraki, Onofre Combarros, Maria J. Bullido, Gloria Tognoni, et al. 2011. 'Evidence of the Association of BIN1 and PICALM with the AD Risk in Contrasting European Populations'. *Neurobiology of Aging* 32(4): 756.e11-15. doi:10.1016/j.neurobiolaging.2010.11.022.

Lancaster, Madeline A., Magdalena Renner, Carol-Anne Martin, Daniel Wenzel, Louise S. Bicknell, Matthew E. Hurles, Tessa Homfray, et al. 2013. 'Cerebral Organoids Model Human Brain Development and Microcephaly'. *Nature* 501(7467): 10.1038/nature12517. doi:10.1038/nature12517.

Lander, Eric S., Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, et al. 2001. 'Initial Sequencing and Analysis of the Human Genome'. *Nature* 409(6822): 860–921. doi:10.1038/35057062.

Larsen, Peter A., Kelsie E. Hunnicutt, Roxanne J. Larsen, Anne D. Yoder, and Ann M. Saunders. 2018. 'Warning SINEs: Alu Elements, Evolution of the Human Brain, and the Spectrum of Neurological Disease'. *Chromosome Research* 26(1): 93–111. doi:10.1007/s10577-018-9573-4.

Larsen, Peter A., Michael W. Lutz, Kelsie E. Hunnicutt, Mirta Mihovilovic, Ann M. Saunders, Anne D. Yoder, and Allen D. Roses. 2017. 'The Alu Neurodegeneration Hypothesis: A Primate-Specific Mechanism for Neuronal Transcription Noise, Mitochondrial Dysfunction, and Manifestation of Neurodegenerative Disease'. *Alzheimer's & dementia : the journal of the Alzheimer's Association* 13(7): 828–38. doi:10.1016/j.jalz.2017.01.017.

Lawrence, Michael, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. 2013. 'Software for Computing and Annotating Genomic Ranges'. *PLoS computational biology* 9(8): e1003118. doi:10.1371/journal.pcbi.1003118.

Le Guennec, Kilan, Sarah Veugelen, Olivier Quenez, Maria Szaruga, Stéphane Rousseau, Gaël Nicolas, David Wallon, et al. 2017. 'Deletion of Exons 9 and 10 of the Presenilin 1 Gene in a Patient with Early-Onset Alzheimer Disease Generates Longer Amyloid Seeds'. *Neurobiology of Disease* 104: 97–103. doi:10.1016/j.nbd.2017.04.020.

Lee, Jong-Hyuk, Edward W. Kim, Deborah L. Croteau, and Vilhelm A. Bohr. 2020. 'Heterochromatin: An Epigenetic Point of View in Aging'. *Experimental & Molecular Medicine* 52(9): 1466–74. doi:10.1038/s12276-020-00497-4.

Lee, Joseph H., Rong Cheng, Sandra Barral, Christiane Reitz, Martin Medrano, Rafael Lantigua, Ivonne Z. Jiménez-Velazquez, et al. 2011. 'Identification of Novel Loci for Alzheimer Disease and Replication of CLU, PICALM, and BIN1 in Caribbean Hispanic Individuals'. *Archives of Neurology* 68(3): 320–28. doi:10.1001/archneurol.2010.292.

Li, Dingfeng, Juan Zhang, Xiaohui Li, Yuhua Chen, Feng Yu, and Qiang Liu. 2021. 'Insights into lncRNAs in Alzheimer's Disease Mechanisms'. *RNA Biology* 18(7): 1037–47. doi:10.1080/15476286.2020.1788848.

Li, Jun-Ming, Chao-Lin Lu, Min-Chih Cheng, Sy-Ueng Luu, Shih-Hsin Hsu, Tsung-Ming Hu, Hsin-Yao Tsai, and Chia-Hsiang Chen. 2014. 'Role of the DLGAP2 Gene Encoding the SAP90/PSD-95-Associated Protein 2 in Schizophrenia'.

PLoS ONE 9(1): e85373. doi:10.1371/journal.pone.0085373.

Li, Wanhe, Ying Jin, Lisa Prazak, Molly Hammell, and Josh Dubnau. 2012. 'Transposable Elements in TDP-43-Mediated Neurodegenerative Disorders'. *PLoS ONE* 7(9): e44099. doi:10.1371/journal.pone.0044099.

Li, Wanhe, Lisa Prazak, Nabanita Chatterjee, Servan Grüninger, Lisa Krug, Delphine Theodorou, and Josh Dubnau. 2013. 'Activation of Transposable Elements during Aging and Neuronal Decline in Drosophila'. *Nature Neuroscience* 16(5): 529–31. doi:10.1038/nn.3368.

Li, Wenxue, Myoung-Hwa Lee, Lisa Henderson, Richa Tyagi, Muzna Bachani, Joseph Steiner, Emilie Campanac, et al. 2015. 'Human Endogenous Retrovirus-K Contributes to Motor Neuron Disease'. *Science Translational Medicine* 7(307): 307ra153-307ra153. doi:10.1126/scitranslmed.aac8201.

Liao, Xingyu, Wufei Zhu, Juexiao Zhou, Haoyang Li, Xiaopeng Xu, Bin Zhang, and Xin Gao. 2023. 'Repetitive DNA Sequence Detection and Its Role in the Human Genome'. *Communications Biology* 6(1): 1–21. doi:10.1038/s42003-023-05322-y.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. 'featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features'. *Bioinformatics (Oxford, England)* 30(7): 923–30. doi:10.1093/bioinformatics/btt656.

Linthorst, Jasper, Wim Meert, Matthew S. Hestand, Jonas Korlach, Joris Robert Vermeesch, Marcel J. T. Reinders, and Henne Holstege. 2020. 'Extreme Enrichment of VNTR-Associated Polymorphicity in Human Subtelomeres: Genes with Most VNTRs Are Predominantly Expressed in the Brain'. *Translational Psychiatry* 10(1): 1–13. doi:10.1038/s41398-020-01060-5.

Liu, Lei, Li Ding, Matteo Rovere, Michael S. Wolfe, and Dennis J. Selkoe. 2019. 'A Cellular Complex of BACE1 and γ-Secretase Sequentially Generates Aβ from Its Full-Length Precursor'. *The Journal of Cell Biology* 218(2): 644–63. doi:10.1083/jcb.201806205.

Liu, Yi-Ling, Wei-Ting Chen, Yu-Yi Lin, Po-Hung Lu, Shie-Liang Hsieh, and Irene Han-Juo Cheng. 2017. 'Amelioration of Amyloid-β-Induced Deficits by DcR3 in an Alzheimer's Disease Model'. *Molecular Neurodegeneration* 12: 30. doi:10.1186/s13024-017-0173-0.

Livingston, Gill, Jonathan Huntley, Andrew Sommerlad, David Ames, Clive Ballard, Sube Banerjee, Carol Brayne, et al. 2020. 'Dementia Prevention, Intervention, and Care: 2020 Report of the Lancet Commission'. *The Lancet* 396(10248): 413–46. doi:10.1016/S0140-6736(20)30367-6.

Lonsdale, John, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, et al. 2013. 'The Genotype-Tissue Expression (GTEx) Project'. *Nature Genetics* 45(6): 580–85. doi:10.1038/ng.2653.

Lopez-Delisle, Lucille, Leily Rabbani, Joachim Wolff, Vivek Bhardwaj, Rolf Backofen, Björn Grüning, Fidel Ramírez, and Thomas Manke. 2020. 'pyGenomeTracks: Reproducible Plots for Multivariate Genomic Datasets'. *Bioinformatics* 37(3): 422–23. doi:10.1093/bioinformatics/btaa692.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. 'Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2'. *Genome Biology* 15(12): 550. doi:10.1186/s13059-014-0550-8.

Ma, Jing, Jin-Tai Yu, and Lan Tan. 2015. 'MS4A Cluster in Alzheimer's Disease'. *Molecular Neurobiology* 51(3): 1240–48. doi:10.1007/s12035-014-8800-z.

MacArthur, Jacqueline, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, et al. 2017. 'The New NHGRI-EBI Catalog of Published Genome-Wide Association Studies (GWAS Catalog)'. *Nucleic Acids Research* 45(Database issue): D896–901. doi:10.1093/nar/gkw1133.

Machiela, Mitchell J., and Stephen J. Chanock. 2015. 'LDlink: A Web-Based Application for Exploring Population-Specific Haplotype Structure and Linking Correlated Alleles of Possible Functional Variants'. *Bioinformatics* 31(21): 3555–57. doi:10.1093/bioinformatics/btv402.

Mackenzie, Ian Ra, Rosa Rademakers, and Manuela Neumann. 2010. 'TDP-43 and FUS in Amyotrophic Lateral Sclerosis and Frontotemporal Dementia'. *The Lancet. Neurology* 9(10): 995–1007. doi:10.1016/S1474-4422(10)70195-2.

Makino, Satoshi, Ryuji Kaji, Satoshi Ando, Maiko Tomizawa, Katsuhito Yasuno, Satoshi Goto, Shinnichi Matsumoto, et al. 2007. 'Reduced Neuron-Specific Expression of the TAF1 Gene Is Associated with X-Linked Dystonia-Parkinsonism'. *American Journal of Human Genetics* 80(3): 393–406. doi:10.1086/512129.

Mardis, Elaine R. 2013. 'Next-Generation Sequencing Platforms'. *Annual Review of Analytical Chemistry* 6(1): 287–303. doi:10.1146/annurev-anchem-062012-092628.

Marshall, Jack NG, Ana Illera Lopez, Abigail L Pfaff, Sulev Koks, John P Quinn, and Vivien J Bubb. 2021. 'Variable Number Tandem Repeats – Their Emerging Role in Sickness and Health'. *Experimental Biology and Medicine* 246(12): 1368–76. doi:10.1177/15353702211003511.

Masoodi, Tariq Ahmad, Sulaiman A. Al Shammari, May N. Al-Muammar, Adel A. Alhamdan, and Venkateswar Rao Talluri. 2013. 'Exploration of Deleterious Single Nucleotide Polymorphisms in Late-Onset Alzheimer Disease Susceptibility Genes'. *Gene* 512(2): 429–37. doi:10.1016/j.gene.2012.08.026.

Miao, Benpeng, Shuhua Fu, Cheng Lyu, Paul Gontarz, Ting Wang, and Bo Zhang. 2020. 'Tissue-Specific Usage of Transposable Element-Derived Promoters in Mouse Development'. *Genome Biology* 21(1): 255. doi:10.1186/s13059-020-02164-3.

Moll, Tobias, Pamela J Shaw, and Johnathan Cooper-Knock. 2020. 'Disrupted Glycosylation of Lipids and Proteins Is a Cause of Neurodegeneration'. *Brain* 143(5): 1332–40. doi:10.1093/brain/awz358.

Moore, Jill E., Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, et al. 2020. 'Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes'. *Nature* 583(7818): 699–710. doi:10.1038/s41586-020-2493-4.

Morandi, Elena, Radu Tanasescu, Rachael E. Tarlinton, Cris S. Constantinescu, Weiya Zhang, Christopher Tench, and Bruno Gran. 2017. 'The Association between Human Endogenous Retroviruses and Multiple Sclerosis: A Systematic Review and Meta-Analysis'. *PLOS ONE* 12(2): e0172415. doi:10.1371/journal.pone.0172415.

Mortezaei, Zahra, and Mahmood Tavallaei. 2021. 'Recent Innovations and In-Depth Aspects of Post-Genome Wide Association Study (Post-GWAS) to Understand the Genetic Basis of Complex Phenotypes'. *Heredity* 127(6): 485–97. doi:10.1038/s41437-021-00479-w.

Muotri, Alysson R., Vi T. Chu, Maria C. N. Marchetto, Wei Deng, John V. Moran, and Fred H. Gage. 2005. 'Somatic Mosaicism in Neuronal Precursor Cells Mediated by L1 Retrotransposition'. *Nature* 435(7044): 903–10. doi:10.1038/nature03663.

Myers, E. W., and W. Miller. 1988. 'Optimal Alignments in Linear Space'. *Computer applications in the biosciences: CABIOS* 4(1): 11–17. doi:10.1093/bioinformatics/4.1.11.

Nalls, Mike A., Cornelis Blauwendraat, Costanza L. Vallerga, Karl Heilbron, Sara Bandres-Ciga, Diana Chang, Manuela Tan, et al. 2019. 'Identification of Novel Risk Loci, Causal Insights, and Heritable Risk for Parkinson's Disease: A Meta-Genome Wide Association Study'. *The Lancet. Neurology* 18(12): 1091–1102. doi:10.1016/S1474-4422(19)30320-5.

Nalls, Mike A., Nathan Pankratz, Christina M. Lill, Chuong B. Do, Dena G. Hernandez, Mohamad Saad, Anita L. DeStefano, et al. 2014. 'Large-Scale Meta-Analysis of Genome-Wide Association Data Identifies Six New Risk Loci for Parkinson's Disease'. *Nature Genetics* 46(9): 989–93. doi:10.1038/ng.3043.

Nicoletto, Giulia, Marianna Terreri, Ilaria Maurizio, Emanuela Ruggiero, Filippo M Cernilogar, Christine A Vaine, Maria Vittoria Cottini, et al. 2024. 'G-Quadruplexes in an SVA Retrotransposon Cause Aberrant TAF1 Gene Expression in X-Linked Dystonia Parkinsonism'. *Nucleic Acids Research* 52(19): 11571–86. doi:10.1093/nar/gkae797.

Nurk, Sergey, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, et al. 2022. 'The Complete Sequence of a Human Genome'. *Science* 376(6588): 44–53. doi:10.1126/science.abj6987.

Okonechnikov, Konstantin, Olga Golosova, Mikhail Fursov, and UGENE team. 2012. 'Unipro UGENE: A Unified Bioinformatics Toolkit'. *Bioinformatics (Oxford, England)* 28(8): 1166–67. doi:10.1093/bioinformatics/bts091.

Olivier, Margaux, Amy Hesketh, Marie-Noëlle Pouch-Pélissier, Thierry Pélissier, Ying Huang, David Latrasse, Moussa Benhamed, and Olivier Mathieu. 2023. 'RTEL1 Is Required for Silencing and Epigenome Stability'. *Nucleic Acids Research* 51(16): 8463–79. doi:10.1093/nar/gkad610.

Ostertag, Eric M., John L. Goodier, Yue Zhang, and Haig H. Kazazian Jr. 2003. 'SVA Elements Are Nonautonomous Retrotransposons That Cause Disease in Humans'. *American Journal of Human Genetics* 73(6): 1444–51.

Ouellette, Andrew R., Sarah M. Neuner, Logan Dumitrescu, Laura C. Anderson, Daniel M. Gatti, Emily R. Mahoney, Jason A. Bubier, et al. 2020. 'Cross-Species Analyses Identify Dlgap2 as a Regulator of Age-Related Cognitive Decline and Alzheimer's Dementia'. *Cell reports* 32(9): 108091. doi:10.1016/j.celrep.2020.108091.

Payer, Lindsay M., and Kathleen H. Burns. 2019. 'Transposable Elements in Human Genetic Disease'. *Nature Reviews Genetics* 20(12): 760–72. doi:10.1038/s41576-019-0165-8.

Payer, Lindsay M., Jared P. Steranka, Wan Rou Yang, Maria Kryatova, Sibyl Medabalimi, Daniel Ardeljan, Chunhong Liu, et al. 2017. '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* 114(20): E3984–92. doi:10.1073/pnas.1704117114.

Pearson, Christopher E., Kerrie Nichol Edamura, and John D. Cleary. 2005. 'Repeat Instability: Mechanisms of Dynamic Mutations'. *Nature Reviews Genetics* 6(10): 729–42. doi:10.1038/nrg1689.

Pericak-Vance, M. A., J. L. Bebout, P. C. Gaskell, L. H. Yamaoka, W.-Y. Hung, M. J. Alberts, A. P. Walker, et al. 1991. 'Linkage Studies in Familial Alzheimer Disease: Evidence for Chromosome 19 Linkage'. *American Journal of Human Genetics* 48(6): 1034–50.

Pertea, Geo, and Mihaela Pertea. 2020. 'GFF Utilities: GffRead and GffCompare'. *F1000Research* 9: ISCB Comm J-304. doi:10.12688/f1000research.23297.2.

Pertea, Mihaela, Geo M. Pertea, Corina M. Antonescu, Tsung-Cheng Chang, Joshua T. Mendell, and Steven L. Salzberg. 2015. 'StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads'. *Nature Biotechnology* 33(3): 290–95. doi:10.1038/nbt.3122.

Petit, Dieter, Sara Gutiérrez Fernández, Katarzyna Marta Zoltowska, Thomas Enzlein, Natalie S. Ryan, Antoinette O'Connor, Maria Szaruga, et al. 2022. 'Aβ Profiles Generated by Alzheimer's Disease Causing PSEN1 Variants Determine the Pathogenicity of the Mutation and Predict Age at Disease Onset'. *Molecular Psychiatry* 27(6): 2821–32. doi:10.1038/s41380-022-01518-6.

Petrozziello, Tiziana, Amanda M. Dios, Kaly A. Mueller, Christine A. Vaine, William T. Hendriks, Kelly E. Glajch, Alexandra N. Mills, et al. 2020. 'SVA Insertion in X-Linked Dystonia Parkinsonism Alters Histone H3 Acetylation Associated with TAF1 Gene'. *PLoS ONE* 15(12): e0243655. doi:10.1371/journal.pone.0243655.

Pfaff, Abigail L., Vivien J. Bubb, John P. Quinn, and Sulev Koks. 2021. 'Reference SVA Insertion Polymorphisms Are Associated with Parkinson's Disease Progression and Differential Gene Expression'. *npj Parkinson's Disease* 7(1): 1–9. doi:10.1038/s41531-021-00189-4.

Pfaff, Abigail L., Vivien J. Bubb, John P. Quinn, and Sulev Koks. 2023. 'A Genome-Wide Screen for the Exonisation of Reference SINE-VNTR-Alus and Their Expression in CNS Tissues of Individuals with Amyotrophic Lateral Sclerosis'. *International Journal of Molecular Sciences* 24(14): 11548.

doi:10.3390/ijms241411548.

Pfaff, Abigail L, Lewis M Singleton, and Sulev Kõks. 2022. 'Mechanisms of Disease-Associated SINE-VNTR-Alus'. *Experimental Biology and Medicine* 247(9): 756–64. doi:10.1177/15353702221082612.

Phan, Katherine, Ying He, YuHong Fu, Nicolas Dzamko, Surabhi Bhatia, Julian Gold, Dominic Rowe, et al. 2021. 'Pathological Manifestation of Human Endogenous Retrovirus K in Frontotemporal Dementia'. *Communications Medicine* 1(1): 1–11. doi:10.1038/s43856-021-00060-w.

Pontis, Julien, Evarist Planet, Sandra Offner, Priscilla Turelli, Julien Duc, Alexandre Coudray, Thorold W. Theunissen, Rudolf Jaenisch, and Didier Trono. 2019. 'Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs'. *Cell Stem Cell* 24(5): 724-735.e5. doi:10.1016/j.stem.2019.03.012.

Porubsky, David, Peter Ebert, Peter A. Audano, Mitchell R. Vollger, William T. Harvey, Pierre Marijon, Jana Ebler, et al. 2021. 'Fully Phased Human Genome Assembly without Parental Data Using Single-Cell Strand Sequencing and Long Reads'. *Nature Biotechnology* 39(3): 302–8. doi:10.1038/s41587-020-0719-5.

Prokopenko, Dmitry, Sarah L. Morgan, Kristina Mullin, Oliver Hofmann, Brad Chapman, Rory Kirchner, Sandeep Amberkar, et al. 2021. 'Whole-genome Sequencing Reveals New Alzheimer's Disease–Associated Rare Variants in Loci Related to Synaptic Function and Neuronal Development'. *Alzheimer's & Dementia* 17(9): 1509–27. doi:10.1002/alz.12319.

Pruim, Randall J., Ryan P. Welch, Serena Sanna, Tanya M. Teslovich, Peter S. Chines, Terry P. Gliedt, Michael Boehnke, Gonçalo R. Abecasis, and Cristen J. Willer. 2010. 'LocusZoom: Regional Visualization of Genome-Wide Association Scan Results'. *Bioinformatics* 26(18): 2336–37. doi:10.1093/bioinformatics/btq419.

Pruitt, Kim D., Garth R. Brown, Susan M. Hiatt, Françoise Thibaud-Nissen, Alexander Astashyn, Olga Ermolaeva, Catherine M. Farrell, et al. 2014. 'RefSeq: An Update on Mammalian Reference Sequences'. *Nucleic Acids Research* 42(D1): D756–63. doi:10.1093/nar/gkt1114.

Pruitt, Kim D., Tatiana Tatusova, and Donna R. Maglott. 2005. 'NCBI Reference Sequence (RefSeq): A Curated Non-Redundant Sequence Database of Genomes, Transcripts and Proteins'. *Nucleic Acids Research* 33(suppl_1): D501–4. doi:10.1093/nar/gki025.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, et al. 2007. 'PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses'. *American Journal of Human Genetics* 81(3): 559–75.

Qian, Xuyu, Hongjun Song, and Guo-li Ming. 2019. 'Brain Organoids: Advances, Applications and Challenges'. *Development (Cambridge, England)* 146(8): dev166074. doi:10.1242/dev.166074.

Quinlan, Aaron R., and Ira M. Hall. 2010. 'BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features'. *Bioinformatics* 26(6): 841–42. doi:10.1093/bioinformatics/btq033.

Raj, Towfique, Yang I. Li, Garrett Wong, Jack Humphrey, Minghui Wang, Satesh Ramdhani, Ying-Chih Wang, et al. 2018. 'Integrative Transcriptome Analyses of the Aging Brain Implicate Altered Splicing in Alzheimer's Disease Susceptibility'. *Nature genetics* 50(11): 1584–92. doi:10.1038/s41588-018-0238-1.

Raja, Waseem K., Alison E. Mungenast, Yuan-Ta Lin, Tak Ko, Fatema Abdurrob, Jinsoo Seo, and Li-Huei Tsai. 2016. 'Self-Organizing 3D Human Neural Tissue Derived from Induced Pluripotent Stem Cells Recapitulate Alzheimer's Disease Phenotypes'. *PLOS ONE* 11(9): e0161969. doi:10.1371/journal.pone.0161969.

Ramírez, Fidel, Vivek Bhardwaj, Laura Arrigoni, Kin Chung Lam, Björn A. Grüning, José Villaveces, Bianca Habermann, Asifa Akhtar, and Thomas Manke. 2018. 'High-Resolution TADs Reveal DNA Sequences Underlying Genome Organization in Flies'. *Nature Communications* 9(1): 189. doi:10.1038/s41467-017-02525-w.

Ramírez, Fidel, Devon P. Ryan, Björn Grüning, Vivek Bhardwaj, Fabian Kilpert, Andreas S. Richter, Steffen Heyne, Friederike Dündar, and Thomas Manke. 2016. 'deepTools2: A next Generation Web Server for Deep-Sequencing Data Analysis'. *Nucleic Acids Research* 44(W1): W160-165. doi:10.1093/nar/gkw257.

Ran, F. Ann, Patrick D. Hsu, Jason Wright, Vineeta Agarwala, David A. Scott, and Feng Zhang. 2013. 'Genome Engineering Using the CRISPR-Cas9 System'. *Nature Protocols* 8(11): 2281–2308. doi:10.1038/nprot.2013.143.

Rasmussen, Andreas H., Hanne B. Rasmussen, and Asli Silahtaroglu. 2017. 'The DLGAP Family: Neuronal Expression, Function and Role in Brain Disorders'. *Molecular Brain* 10(1): 43. doi:10.1186/s13041-017-0324-9.

Ravel-Godreuil, Camille, Rania Znaidi, Tom Bonnifet, Rajiv L. Joshi, and Julia Fuchs. 2021. 'Transposable Elements as New Players in Neurodegenerative Diseases'. *FEBS Letters* 595(22): 2733–55. doi:10.1002/1873-3468.14205.

Rhodes, Daniela, and Hans J. Lipps. 2015. 'G-Quadruplexes and Their Regulatory Roles in Biology'. *Nucleic Acids Research* 43(18): 8627–37. doi:10.1093/nar/gkv862.

Richard, Guy-Franck, and Frédéric Pâques. 2000. 'Mini- and Microsatellite Expansions: The Recombination Connection'. *EMBO reports* 1(2): 122–26. doi:10.1093/embo-reports/kvd031.

Ridge, Perry G., Shubhabrata Mukherjee, Paul K. Crane, and John S. K. Kauwe. 2013. 'Alzheimer's Disease: Analyzing the Missing Heritability'. *PLoS ONE* 8(11): e79771. doi:10.1371/journal.pone.0079771.

de Rojas, Itziar, Sonia Moreno-Grau, Niccolo Tesi, Benjamin Grenier-Boley, Victor Andrade, Iris E. Jansen, Nancy L. Pedersen, et al. 2021. 'Common Variants in Alzheimer's Disease and Risk Stratification by Polygenic Risk Scores'.

Nature Communications 12(1): 3417. doi:10.1038/s41467-021-22491-8.

Rosenkrantz, Jimi L., J. Elias Brandorff, Sanaz Raghib, Ashni Kapadia, Christine A. Vaine, D. Cristopher Bragg, Grace Farmiloe, and Frank M. J. Jacobs. 2024. 'ZNF91 Is an Endogenous Repressor of the Molecular Phenotype Associated with X-Linked Dystonia-Parkinsonism (XDP)'. *Proceedings of the National Academy of Sciences of the United States of America* 121(33): e2401217121. doi:10.1073/pnas.2401217121.

Roses, Allen D., P. Anthony Akkari, Ornit Chiba-Falek, Michael W. Lutz, William Kirby Gottschalk, Ann Marie Saunders, Bob Saul, Scott Sundseth, and Daniel Burns. 2016. 'Structural Variants Can Be More Informative for Disease Diagnostics, Prognostics and Translation than Current SNP Mapping and Exon Sequencing'. *Expert Opinion on Drug Metabolism & Toxicology* 12(2): 135–47. doi:10.1517/17425255.2016.1133586.

Saddiki, Hana, Aurore Fayosse, Emmanuel Cognat, Séverine Sabia, Sebastiaan Engelborghs, David Wallon, Panagiotis Alexopoulos, et al. 2020. 'Age and the Association between Apolipoprotein E Genotype and Alzheimer Disease: A Cerebrospinal Fluid Biomarker–Based Case–Control Study'. *PLoS Medicine* 17(8): e1003289. doi:10.1371/journal.pmed.1003289.

Sahara, N., Y. Yahagi, H. Takagi, T. Kondo, M. Okochi, M. Usami, T. Shirasawa, and H. Mori. 1996. 'Identification and Characterization of Presenilin I-467, I-463 and I-374'. *FEBS letters* 381(1–2): 7–11. doi:10.1016/0014-5793(96)00054-3.

Saiki, Randall K., Stephen Scharf, Fred Faloona, Kary B. Mullis, Glenn T. Horn, Henry A. Erlich, and Norman Arnheim. 1985. 'Enzymatic Amplification of β-Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia'. *Science* 230(4732): 1350–54. doi:10.1126/science.2999980.

Salazar, Alex, Niccolò Tesi, Marc Hulsman, Lydian Knoop, Sven van der Lee, Sanduni Wijesekera, Jana Krizova, Marcel Reinders, and Henne Holstege. 2023. 'An AluYb8 Mobile Element Characterises a Risk Haplotype of TMEM106B Associated in Neurodegeneration'. : 2023.07.16.23292721. doi:10.1101/2023.07.16.23292721.

Savage, Abigail L, Vivien J Bubb, Gerome Breen, and John P Quinn. 2013. 'Characterisation of the Potential Function of SVA Retrotransposons to Modulate Gene Expression Patterns'. *BMC Evolutionary Biology* 13: 101. doi:10.1186/1471-2148-13-101.

Savage, Abigail L., Thomas P. Wilm, Kejhal Khursheed, Aleksey Shatunov, Karen E. Morrison, Pamela J. Shaw, Christopher E. Shaw, et al. 2014. 'An Evaluation of a SVA Retrotransposon in the FUS Promoter as a Transcriptional Regulator and Its Association to ALS'. *PLOS ONE* 9(3): e90833. doi:10.1371/journal.pone.0090833.

Schedin-Weiss, Sophia, Bengt Winblad, and Lars O. Tjernberg. 2014. 'The Role of Protein Glycosylation in Alzheimer Disease'. *The FEBS Journal* 281(1): 46–62. doi:10.1111/febs.12590.

Schmitt, Anthony D., Ming Hu, Inkyung Jung, Zheng Xu, Yunjiang Qiu, Catherine L. Tan, Yun Li, et al. 2016. 'A Compendium of Chromatin Contact Maps

Reveal Spatially Active Regions in the Human Genome'. *Cell reports* 17(8): 2042–59. doi:10.1016/j.celrep.2016.10.061.

Schork, Nicholas J., Sarah S. Murray, Kelly A. Frazer, and Eric J. Topol. 2009. 'Common vs. Rare Allele Hypotheses for Complex Diseases'. *Current opinion in genetics & development* 19(3): 212–19. doi:10.1016/j.gde.2009.04.010.

Schwartzentruber, Jeremy, Sarah Cooper, Jimmy Z. Liu, Inigo Barrio-Hernandez, Erica Bello, Natsuhiko Kumasaka, Adam M. H. Young, et al. 2021. 'Genome-Wide Meta-Analysis, Fine-Mapping, and Integrative Prioritization Implicate New Alzheimer's Disease Risk Genes'. *Nature genetics* 53(3): 392–402. doi:10.1038/s41588-020-00776-w.

Sekar, Aswin, Allison R. Bialas, Heather de Rivera, Avery Davis, Timothy R. Hammond, Nolan Kamitaki, Katherine Tooley, et al. 2016. 'Schizophrenia Risk from Complex Variation of Complement Component 4'. *Nature* 530(7589): 177–83. doi:10.1038/nature16549.

Selkoe, Dennis J. 2008. 'Biochemistry and Molecular Biology of Amyloid β-Protein and the Mechanism of Alzheimer's Disease'. In *Handbook of Clinical Neurology*, Dementias, Elsevier, 245–60. doi:10.1016/S0072-9752(07)01223-7.

Selkoe, Dennis J, and John Hardy. 2016. 'The Amyloid Hypothesis of Alzheimer's Disease at 25 Years'. *EMBO Molecular Medicine* 8(6): 595–608. doi:10.15252/emmm.201606210.

Seshadri, Sudha, Annette L. Fitzpatrick, M Arfan Ikram, Anita L. DeStefano, Vilmundur Gudnason, Merce Boada, Joshua C. Bis, et al. 2010. 'Genome-Wide Analysis of Genetic Loci Associated with Alzheimer's Disease'. *JAMA : the journal of the American Medical Association* 303(18): 1832–40. doi:10.1001/jama.2010.574.

Sherva, Richard, Yorghos Tripodis, David A. Bennett, Lori B. Chibnik, Paul K. Crane, Philip L. de Jager, Lindsay A. Farrer, et al. 2014. 'Genome-Wide Association Study of the Rate of Cognitive Decline in Alzheimer's Disease'. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 10(1): 45–52. doi:10.1016/j.jalz.2013.01.008.

Shpyleva, Svitlana, Stepan Melnyk, Oleksandra Pavliv, Igor Pogribny, and S. Jill James. 2018. 'Overexpression of LINE-1 Retrotransposons in Autism Brain'. *Molecular Neurobiology* 55(2): 1740–49. doi:10.1007/s12035-017-0421-x.

Simón-Sánchez, Javier, Claudia Schulte, Jose M. Bras, Manu Sharma, J. Raphael Gibbs, Daniela Berg, Coro Paisan-Ruiz, et al. 2009. 'Genome-Wide Association Study Reveals Genetic Risk Underlying Parkinson's Disease'. *Nature Genetics* 41(12): 1308–12. doi:10.1038/ng.487.

Sims, Rebecca, Sven J. van der Lee, Adam C. Naj, Céline Bellenguez, Nandini Badarinarayan, Johanna Jakobsdottir, Brian W. Kunkle, et al. 2017. 'Rare Coding Variants in PLCG2, ABI3 and TREM2 Implicate Microglial-Mediated Innate Immunity in Alzheimer's Disease'. *Nature genetics* 49(9): 1373–84. doi:10.1038/ng.3916.

Singh, Shrishti, Maheshkumar R. Borkar, and Lokesh Kumar Bhatt. 2024. 'Transposable Elements: Emerging Therapeutic Targets in Neurodegenerative Dis-

eases'. *Neurotoxicity Research* 42(1): 9. doi:10.1007/s12640-024-00688-1.

Smit, A. F. 1999. 'Interspersed Repeats and Other Mementos of Transposable Elements in Mammalian Genomes'. *Current Opinion in Genetics & Development* 9(6): 657–63. doi:10.1016/s0959-437x(99)00031-3.

Smit, AFA, R Hubley, and P Green. 1996. 'RepeatMasker Open-3.0'. http://www.repeatmasker.org.

Smoom, Riham, Catherine Lee May, Vivian Ortiz, Mark Tigue, Hannah M. Kolev, Melissa Rowe, Yitzhak Reizel, et al. 2023. 'Telomouse—a Mouse Model with Human-Length Telomeres Generated by a Single Amino Acid Change in RTEL1'. *Nature Communications* 14(1): 6708. doi:10.1038/s41467-023-42534-6.

Song, Tanya, Yunfei Wang, Bret David Silverglate, and George T. Grossberg. 2024. 'Pharmacokinetic Evaluation of Donanemab for the Treatment of Alzheimer's'. *Expert Opinion on Drug Metabolism & Toxicology* 20(6): 411–17. doi :10.1080/17425255.2024.2357637.

Soria Lopez, Jose A., Hector M. González, and Gabriel C. Léger. 2019. 'Chapter 13 - Alzheimer's Disease'. In *Handbook of Clinical Neurology*, Geriatric Neurology, eds. Steven T. Dekosky and Sanjay Asthana. Elsevier, 231–55. doi:10.1016/B978-0-12-804766-8.00013-3.

Stelzmann, Rainulf A., H. Norman Schnitzlein, and F. Reed Murtagh. 1995. 'An English Translation of Alzheimer's 1907 Paper, "Über Eine Eigenartige Erkankung Der Hirnrinde"'. *Clinical Anatomy* 8(6): 429–31. doi:10.1002/ca.980080612.

Strittmatter, W J, A M Saunders, D Schmechel, M Pericak-Vance, J Enghild, G S Salvesen, and A D Roses. 1993. 'Apolipoprotein E: High-Avidity Binding to Beta-Amyloid and Increased Frequency of Type 4 Allele in Late-Onset Familial Alzheimer Disease.' *Proceedings of the National Academy of Sciences of the United States of America* 90(5): 1977–81.

Sulovari, Arvis, Ruiyang Li, Peter A. Audano, David Porubsky, Mitchell R. Vollger, Glennis A. Logsdon, Wesley C. Warren, et al. 2019. 'Human-Specific Tandem Repeat Expansion and Differential Gene Expression during Primate Evolution'. *Proceedings of the National Academy of Sciences of the United States of America* 116(46): 23243–53. doi:10.1073/pnas.1912175116.

Sun, Na, Matheus B. Victor, Yongjin P. Park, Xushen Xiong, Aine Ni Scannail, Noelle Leary, Shaniah Prosper, et al. 2023. 'Human Microglial State Dynamics in Alzheimer's Disease Progression'. *Cell* 186(20): 4386-4403.e29. doi:10.1016/j.cell.2023.08.037.

Sundaram, Vasavi, and Joanna Wysocka. 2020. 'Transposable Elements as a Potent Source of Diverse Cis-Regulatory Sequences in Mammalian Genomes'. *Philosophical Transactions of the Royal Society B: Biological Sciences* 375(1795): 20190347. doi:10.1098/rstb.2019.0347.

Tao, Qing-Qing, Yu-Chao Chen, and Zhi-Ying Wu. 2019. 'The Role of CD2AP in the Pathogenesis of Alzheimer's Disease'. *Aging and Disease* 10(4): 901–7. doi:10.14336/AD.2018.1025.

Tesi, Niccolo, Alex Salazar, Yaran Zhang, Sven van der Lee, Marc Hulsman, Lydian Knoop, Sanduni Wijesekera, et al. 2024. 'Characterising Tandem Repeat Complexities across Long-Read Sequencing Platforms with TREAT and Otter'. *Genome Research*: gr.279351.124. doi:10.1101/gr.279351.124.

Thomas, Elizabeth Ochoa, Gabbe Zuniga, Wenyan Sun, and Bess Frost. 2020. 'Awakening the Dark Side: Retrotransposon Activation in Neurodegenerative Disorders'. *Current opinion in neurobiology* 61: 65–72. doi:10.1016/j.conb.2020.01.012.

tijsbliek, Marc Galland, and Jihed Chouaref. 2018. 'KoesGroup/Snakemake_hisat-DESeq: Release on Zenodo'. doi:10.5281/zenodo.2393540.

Trizzino, Marco, Aurélie Kapusta, and Christopher D. Brown. 2018. 'Transposable Elements Generate Regulatory Novelty in a Tissue-Specific Fashion'. *BMC Genomics* 19(1): 468. doi:10.1186/s12864-018-4850-3.

Trizzino, Marco, YoSon Park, Marcia Holsbach-Beltrame, Katherine Aracena, Katelyn Mika, Minal Caliskan, George H. Perry, Vincent J. Lynch, and Christopher D. Brown. 2017. 'Transposable Elements Are the Primary Source of Novelty in Primate Gene Regulation'. *Genome Research* 27(10): 1623–33. doi:10.1101/gr.218149.116.

Uringa, Evert-Jan, Kathleen Lisaingo, Hilda A. Pickett, Julie Brind'Amour, Jan-Hendrik Rohde, Alex Zelensky, Jeroen Essers, and Peter M. Lansdorp. 2012. 'RTEL1 Contributes to DNA Replication and Repair and Telomere Maintenance'. *Molecular Biology of the Cell* 23(14): 2782–92. doi:10.1091/mbc.E12-03-0179.

Uringa, Evert-Jan, Jillian L. Youds, Kathleen Lisaingo, Peter M. Lansdorp, and Simon J. Boulton. 2011. 'RTEL1: An Essential Helicase for Telomere Maintenance and the Regulation of Homologous Recombination'. *Nucleic Acids Research* 39(5): 1647–55. doi:10.1093/nar/gkq1045.

Usdin, Karen. 2008. 'The Biological Effects of Simple Tandem Repeats: Lessons from the Repeat Expansion Diseases'. *Genome Research* 18(7): 1011–19. doi:10.1101/gr.070409.107.

Van Meter, Michael, Mehr Kashyap, Sarallah Rezazadeh, Anthony J. Geneva, Timothy D. Morello, Andrei Seluanov, and Vera Gorbunova. 2014. 'SIRT6 Represses LINE1 Retrotransposons by Ribosylating KAP1 but This Repression Fails with Stress and Age'. *Nature Communications* 5: 5011. doi:10.1038/ncomms6011.

Vance, Caroline, Boris Rogelj, Tibor Hortobágyi, Kurt J. De Vos, Agnes Lumi Nishimura, Jemeen Sreedharan, Xun Hu, et al. 2009. 'Mutations in FUS, an RNA Processing Protein, Cause Familial Amyotrophic Lateral Sclerosis Type 6'. *Science (New York, N.Y.)* 323(5918): 1208–11. doi:10.1126/science.1165942.

Vavouri, Tanya, and Ben Lehner. 2012. 'Human Genes with CpG Island Promoters Have a Distinct Transcription-Associated Chromatin Organization'. *Genome Biology* 13(11): R110. doi:10.1186/gb-2012-13-11-r110.

Vermunt, Marit W., Peter Reinink, Jeroen Korving, Ewart de Bruijn, Paul M. Creyghton, Onur Basak, Geert Geeven, et al. 2014. 'Large-Scale Identification of Coregulated Enhancer Networks in the Adult Human Brain'. *Cell Reports* 9(2): 767–79. doi:10.1016/j.celrep.2014.09.023.

Vetrivel, Kulandaivelu S., Yun-wu Zhang, Huaxi Xu, and Gopal Thinakaran. 2006. 'Pathological and Physiological Functions of Presenilins'. *Molecular Neurodegeneration* 1(1): 4. doi:10.1186/1750-1326-1-4.

Villeponteau, Bryant. 1997. 'The Heterochromatin Loss Model of Aging'. *Experimental Gerontology* 32(4): 383–94. doi:10.1016/S0531-5565(96)00155-6.

Visel, Axel, Matthew J. Blow, Zirong Li, Tao Zhang, Jennifer A. Akiyama, Amy Holt, Ingrid Plajzer-Frick, et al. 2009. 'ChIP-Seq Accurately Predicts Tissue-Specific Activity of Enhancers'. *Nature* 457(7231): 854–58. doi:10.1038/nature07730.

Wang, Hui, Beth A. Dombroski, Po-Liang Cheng, Albert Tucci, Ya-Qin Si, John J. Farrell, Jung-Ying Tzeng, et al. 2023. 'Structural Variation Detection and Association Analysis of Whole-Genome-Sequence Data from 16,905 Alzheimer's Diseases Sequencing Project Subjects'. : 2023.09.13.23295505. doi:10.1101/2023.09.13.23295505.

Wang, Hui, Jinchuan Xing, Deepak Grover, Dale J. Hedges, Kyudong Han, Jerilyn A. Walker, and Mark A. Batzer. 2005. 'SVA Elements: A Hominid-Specific Retroposon Family'. *Journal of Molecular Biology* 354(4): 994–1007. doi:10.1016/j.jmb.2005.09.085.

Wang, Jichang, Gangcai Xie, Manvendra Singh, Avazeh T. Ghanbarian, Tamás Raskó, Attila Szvetnik, Huiqiang Cai, et al. 2014. 'Primate-Specific Endogenous Retrovirus-Driven Transcription Defines Naive-like Stem Cells'. *Nature* 516(7531): 405–9. doi:10.1038/nature13804.

Wang, Kai, Mingyao Li, and Hakon Hakonarson. 2010. 'ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data'. *Nucleic Acids Research* 38(16): e164. doi:10.1093/nar/gkq603.

Watanabe, Kyoko, Erdogan Taskesen, Arjen van Bochoven, and Danielle Posthuma. 2017. 'Functional Mapping and Annotation of Genetic Associations with FUMA'. *Nature Communications* 8(1): 1826. doi:10.1038/s41467-017-01261-5.

Welter, Danielle, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, et al. 2014. 'The NHGRI GWAS Catalog, a Curated Resource of SNP-Trait Associations'. *Nucleic Acids Research* 42(Database issue): D1001–6. doi:10.1093/nar/gkt1229.

Wicker, Thomas, François Sabot, Aurélie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, et al. 2007. 'A Unified Classification System for Eukaryotic Transposable Elements'. *Nature Reviews Genetics* 8(12): 973–82. doi:10.1038/nrg2165.

Wickham, Hadley. 2007. 'Reshaping Data with the Reshape Package'. *Journal of Statistical Software* 21: 1–20. doi:10.18637/jss.v021.i12.

Wickham, Hadley. 2011. 'The Split-Apply-Combine Strategy for Data Analysis'. *Journal of Statistical Software* 40: 1–29. doi:10.18637/jss.v040.i01.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. 'Welcome to the Tidyverse'. *Journal of Open Source Software* 4(43): 1686. doi:10.21105/joss.01686.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, et al. 2024. 'Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics'. https://cran.r-project.org/web/packages/ggplot2/index.html (January 28, 2025).

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, Davis Vaughan, Posit Software, and PBC. 2023. 'Dplyr: A Grammar of Data Manipulation'. https://cran.r-project.org/web/packages/dplyr/index.html (January 28, 2025).

Wickham, Hadley, Davis Vaughan, Maximilian Girlich, Kevin Ushey, Posit Software, and PBC. 2024. 'Tidyr: Tidy Messy Data'. https://cran.r-project.org/web/packages/tidyr/index.html (January 28, 2025).

Wijsman, Ellen M., Nathan D. Pankratz, Yoonha Choi, Joseph H. Rothstein, Kelley M. Faber, Rong Cheng, Joseph H. Lee, et al. 2011. 'Genome-Wide Association of Familial Late-Onset Alzheimer's Disease Replicates BIN1 and CLU and Nominates CUGBP2 in Interaction with APOE'. *PLoS genetics* 7(2): e1001308. doi:10.1371/journal.pgen.1001308.

Wingo, Thomas S., James J. Lah, Allan I. Levey, and David J. Cutler. 2012. 'Autosomal Recessive Causes Likely in Early-Onset Alzheimer Disease'. *Archives of Neurology* 69(1): 59–64. doi:10.1001/archneurol.2011.221.

Wu, Yih-Ru, Jia Nee Foo, Louis C. S. Tan, Chiung-Mei Chen, Kumar M. Prakash, Yi-Chun Chen, Jin-Xin Bei, et al. 2013. 'Identification of a Novel Risk Variant in the FUS Gene in Essential Tremor'. *Neurology* 81(6): 541–44. doi:10.1212/WNL.0b013e31829e700c.

Xiong, Xushen, Benjamin T. James, Carles A. Boix, Yongjin P. Park, Kyriaki Galani, Matheus B. Victor, Na Sun, et al. 2023. 'Epigenomic Dissection of Alzheimer's Disease Pinpoints Causal Variants and Reveals Epigenome Erosion'. *Cell* 186(20): 4422-4437.e21. doi:10.1016/j.cell.2023.08.040.

Xu, Wei, Lan Tan, and Jin-Tai Yu. 2015. 'The Role of PICALM in Alzheimer's Disease'. *Molecular Neurobiology* 52(1): 399–413. doi:10.1007/s12035-014-8878-3.

Yang, Danying, Xifeng Wang, Lieliang Zhang, Yang Fang, Qingcui Zheng, Xing Liu, Wen Yu, et al. 2022. 'Lipid Metabolism and Storage in Neuroglia: Role in Brain Development and Neurodegenerative Diseases'. *Cell & Bioscience* 12(1): 106. doi:10.1186/s13578-022-00828-0.

Yao, Dongbo, Xiaoxia Wang, Jing Liu, and Xiu Qin Xu. 2024. 'Rbm24 Modulates Neuronal RNA Splicing to Restrict Cognitive Dysfunction'. *International Journal of Biological Macromolecules* 276(Pt 2): 133853. doi:10.1016/j.ijbiomac.2024.133853.

Yoon, Se-Jin, Lubayna S. Elahi, Anca M. Paşca, Rebecca M. Marton, Aaron Gordon, Omer Revah, Yuki Miura, et al. 2019. 'Reliability of Human 3D Cortical

Organoid Generation'. *Nature methods* 16(1): 75–78. doi:10.1038/s41592-018-0255-0.

You, Shih-Feng, Logan Brase, Fabia Filipello, Abhirami K. Iyer, Jorge Del-Aguila, June He, Ricardo D'Oliveira Albanus, et al. 2023. 'MS4A4A Modifies the Risk of Alzheimer Disease by Regulating Lipid Metabolism and Immune Response in a Unique Microglia State'. *medRxiv*: 2023.02.06.23285545. doi:10.1101/2023.02.06.23285545.

Yu, Minhua, Yue Liu, Jun Shen, Dongwei Lv, and Junjian Zhang. 2016. 'Meta-Analysis of BACE1 Gene Rs638405 Polymorphism and the Risk of Alzheimer's Disease in Caucasion and Asian Population'. *Neuroscience Letters* 616: 189–96. doi:10.1016/j.neulet.2016.01.059.

Yushkova, Elena, and Alexey Moskalev. 2023. 'Transposable Elements and Their Role in Aging'. *Ageing Research Reviews* 86: 101881. doi:10.1016/j.arr.2023.101881.

Zerbino, Daniel R., Nathan Johnson, Thomas Juettemann, Steven P. Wilder, and Paul Flicek. 2014. 'WiggleTools: Parallel Processing of Large Collections of Genome-Wide Datasets for Visualization and Statistical Analysis'. *Bioinformatics (Oxford, England)* 30(7): 1008–9. doi:10.1093/bioinformatics/btt737.

Zhang, Chen, Haoran Li, Yujie Huang, Yuchen Tang, Jie Wang, Yinxiang Cheng, Yijun Wei, et al. 2021. 'Integrative Analysis of TNFRSF6B as a Potential Therapeutic Target for Pancreatic Cancer'. *Journal of Gastrointestinal Oncology* 12(4). doi:10.21037/jgo-21-303.

Zhao, Zhen, Abhay P. Sagare, Qingyi Ma, Matthew R. Halliday, Pan Kong, Kassandra Kisler, Ethan A. Winkler, et al. 2015. 'Central Role for PICALM in Amyloid–β Blood–Brain Barrier Transcytosis and Clearance'. *Nature neuroscience* 18(7): 978–87. doi:10.1038/nn.4025.

Zhu, Chenlu, Xiao Ren, Chen Liu, Yawei Liu, and Yonggang Wang. 2024. 'Rbm8a Regulates Neurogenesis and Reduces Alzheimer's Disease-Associated Pathology in the Dentate Gyrus of 5×FAD Mice'. *Neural Regeneration Research* 19(4): 863–71. doi:10.4103/1673-5374.382254.

Zotova, Elina, James AR Nicoll, Raj Kalaria, Clive Holmes, and Delphine Boche. 2010. 'Inflammation in Alzheimer's Disease: Relevance to Pathogenesis and Therapy'. *Alzheimers Res Ther* 2(1): 1. doi:10.1186/alzrt24.

Nederlandse Samenvatting

In dit proefschrift onderzoeken we de rol van repetitief DNA in de genetische architectuur van de ziekte van Alzheimer, met een focus op hoe structurele variaties in TRs en TEs kunnen bijdragen aan de vatbaarheidvoor de ziekte. Hoewel GWAS-onderzoeken talrijke risicoloci voor AD hebben geïdentificeerd, blijft een aanzienlijk deel van de erfelijkheid onverklaard. Ons doel is om verborgen genetische variaties binnen AD-geassocieerde regio's op te sporen en hun potentiële genregulerende functies te beoordelen door genetische, epigenetische en transcriptomische analyses te integreren. Met dit werk hopen we nieuwe inzichten te bieden in de genetische basis van AD.

In **hoofdstuk 1** beginnen we met een uitgebreid overzicht van AD-geassocieerde loci, waarbij we structurele varianten van repetitief DNA in elk locus in overweging nemen. Met behulp van een gepubliceerde GWAS hebben we een fine-mapping-analyse uitgevoerd en 29 genomische risicoloci (GRL) geïdentificeerd die geassocieerd zijn met AD. Binnen de grenzen van de GRL hebben we een dataset van structurele varianten geïntegreerd en 202 varianten in 16 GRL gevonden. Opvallend is dat clusters van intergene, door GWAS geïdentificeerde SNP's zich rondom structureel variabele SVA's (SV-SVA's) bevinden; daarom geven we prioriteit aan deze loci voor verdere functionele analyse.

Hoofdstuk 2 onderzoekt structurele variatie in TE's die actief zijn gebleken in een neuronale omgeving, en hun associatie met neurologische aandoeningen. We gebruikten gepubliceerde long-read sequencing en PCR-analyse om de hoge mate van structurele variatie van SVA-elementen in het gehele menselijke genoom bloot te leggen. Gezien een aanzienlijk aantal SV-SVA's zich bevindt in loci die geassocieerd zijn met neurologische aandoeningen, hebben we de functionele impact van deze elementen onderzocht in 3 loci: *BIN1* en *CD2AP*, geassocieerd met Alzheimer, en *BCKDK*, geassocieerd met de ziekte van Parkinson. Genetische deleties van de SV-SVA's werden aangebracht in humane embryonale stamcellen met behulp van het CRISPR-Cas9-systeem, waarna corticale organoïden werden gegenereerd als model voor een neuronale omgeving. Epigenetische kenmerken veranderden na deletie van de *BCKDK*-SV-SVA, in genen nabij de deletieplaats, en transcriptomische analyse toonde differentiële expressie van *FUS*, 75 kb verderop. Er werden geen epigenetische of transcriptomische veranderingen waargenomen bij de *BIN1*-SVA-KO. Wat betreft de intronische *CD2AP*-SVA werd RNA CaptureSeq uitgevoerd, waaruit verminderde expressieniveaus bleken voor alle *CD2AP*-exonen stroomafwaarts van de SVA-insertieplaats. Samengevat laten onze bevindingen een nieuwe laag van verborgen genetische variatie zien waarmee rekening moet worden gehouden bij het inschatten van het risico op complexe ziekten.

In **hoofdstuk 3** zetten we het werk uit **hoofdstuk 1** voort door een functionele beoordeling te geven van twee SV-SVA's die zich respectievelijk in of stroomopwaarts van de AD-geassocieerde genen *PSEN1* en *PICALM* bevinden. Hier gebruiken we dezelfde methode als beschreven in **hoofdstuk 2**: we genereerden SV-SVA-KO corticale organoïden en voerden RNA CaptureSeq uit voor een hoge resolutie van het locus-transcriptoom. Er was geen bewijs voor een regulerend effect van de SV-SVA in het *PICALM*-gen in de context van een neuronale omgeving. Bij deletie van de intronische SV-SVA in *PSEN1* toonden voorlopige bevindingen echter verhoogde expressieniveaus aan van transcripten die geassocieerd zijn met nonsense-mediated decay.

In **hoofdstuk 4** duiken we in een genetische associatiestudie om VNTR's te identificeren die geassocieerd zijn met AD-risico. We gebruiken een short-read sequencing dataset van het ADSP-cohort om VNTR's te detecteren waarvan de lengte is toegenomen of afgenomen bij het merendeel van de AD-patiënten, of bij een subgroep daarvan, in vergelijking met controles. We ontdekten 52 VNTR's met een suggestieve associatie met de ziekte, waarvan 5 werden gerepliceerd in een onafhankelijke long-read sequencing dataset bestaande uit AD-patiënten en cognitief gezonde honderdjarigen. We lichten de meest veelbelovende AD-geassocieerde VNTR eruit, gelegen in een regulerend element binnen het *RTEL1*-gen, in het promotorgebied van *TNFRSF6B* en in het fusietranscript *RTEL1-TNFRSF6B*. *RTEL1* is een verouderingsgerelateerd gen dat betrokken is bij celhomeostase, terwijl *TNFRSF6B*, dat een immunosuppressieve functie heeft, geassocieerd is met AD. Waar kortere lengtes van de VNTR geassocieerd zijn met AD, correleren langere lengtes met een beschermende GWAS-SNP rs55765053 in *TNFRSF6B*, die op zijn beurt in matige LD staat met de belangrijkste beschermende SNP rs6742. Verdere functionele analyse is nodig om onze voorlopige bevindingen te valideren.

Publications

Lodewijk, Gerrald A., Matthijs de Geus, **Rita L. F. P. Guimarães**, and Frank M. J. Jacobs. 2024. 'Emergence of the ZNF675 Gene During Primate Evolution–Influenced Human Neurodevelopment Through Changing HES1 Autoregulation'. *Journal of Comparative Neurology* 532(7): e25648. doi:10.1002/cne.25648.

van Bree, Elisabeth J., **Rita L.F.P. Guimarães**, Mischa Lundberg, Elena R. Blujdea, Jimi L. Rosenkrantz, Fred T.G. White, Josse Poppinga, et al. 2022. 'A Hidden Layer of Structural Variation in Transposable Elements Reveals Potential Genetic Modifiers in Human Disease-Risk Loci'. *Genome Research* 32(4): 656–70. doi:10.1101/gr.275515.121.

Acknowledgements

The path to this thesis was rarely straightforward. There were times I was not sure I would make it to this page and it is thanks to the people around me that I did. I am deeply grateful to my mentors, friends and family, who stood by me through the difficult moments and celebrated every small victory along the way.

First of all, **Frank**, you were my gateway to Amsterdam. During my Masters, while searching for internship opportunities I came across your lab which immediately caught my eye – such cool and exciting stuff! I reached out to you about the possibility of joining the team and I am so very grateful you gave me that chance, as it turned out to be a turning point in my life. As a young scientist, seeing your endless excitement and motivation about science was incredibly inspiring so when the possibility to stay as a PhD candidate in your lab came up, I had to take it. It was a rough path, with COVID-19 starting at the end of my first year and all our plans changing. It was a difficult time but at that time and throughout my PhD I always felt you had my best interest at heart. Thank you for all the opportunities you put my way, whether in or outside the lab: from conferences in Chile (memorable to say the least!) and in Canada with the whole team, to paper parties or just borrels on Fridays – it was a fun time! Overall, these years under your supervision were challenging but also very rewarding, both scientifically and personally.

Henne, my co-promotor, it was a pleasure to have joined your lab. When I started, we were only four PhD candidates, but the lab quickly grew into a wonderfully multidisciplinary team, and I believe that diversity is exactly what makes it so successful. In my early years I mainly joined the weekly lab meetings, but during my final year I got to spend much more time with the team, learning from them, and I am very grateful for that opportunity. I have always admired you as a successful woman in science, with such curiosity and eagerness to question everything, and for motivating us to do the same. You always made a point of asking what we thought, which meant we could learn from one another and grow together, and I feel very fortunate to have been part of such a team.

Marcel, my promotor alongside Frank, thank you for your time and input throughout this journey. We may not have spent much time together in person, but you never kept me waiting for a response whenever I sent something your way. Despite your incredibly busy schedule, you somehow always found the time to read my work and share your comments. That responsiveness meant a great deal to me, and I am grateful for it.

I would also like to thank the **members of my committee** for taking the time to read and evaluate this thesis.

Over the course of the PhD one meets a lot of people – some who stay only temporarily, others who are there for the full ride. I am so lucky to have found amazing people in both the Jacobs and Holstege labs.

Diana, you were the person I first talked to about applying for an internship in the Jacobs lab. I wanted your opinion since you had been an intern there yourself. You told me to go ahead and here we are now! I cannot begin to tell you how important it was to have you by my side, sharing a whole liver! We have enjoyed life in the Netherlands, through crazy heatwaves, dancing at a silent disco party in Noorderlicht in the middle of a storm, ice skating on the canals of Amsterdam, or even going for the first swim of the year in January (only you though!). Looking back, it is crazy how much we have changed but that is the whole point. To many more *Amarguinha* sours!

Mattie, we met through a mutual friend at the time, and I am so grateful she connected us the way she did. As my colleague in the Holstege lab as well, we became inseparable, relying on each other when life got too hard. You cannot live life without a tune on, from singing along together in the office, or sharing tunes at whatever time of the day. All the concerts we danced at, or the one dance class we went to! The *gezelligheid* of cooking together, chilling on your balcony, watching *New Girl* and falling asleep. It all made life better. Now together with Chryssa, it warms my heart to know you are happy. Thank you for staying, for trusting and confiding in me. Your friendship means the world.

Angie, my star! You started in the lab during COVID-19 times, which was not the easiest, but you were always determined to make the best out of any situation. With your cheerful spirit you befriended Italians all over Amsterdam-Oost and I honestly thought it would be more useful for me to learn Italian rather than Dutch. You have taught me so much, especially that life is to be lived. And so, we went on living – Taormina, Schull in Ireland (with Grace, Anton and Judith), Tenerife, Thailand (with Diana, Mattie and Juju), and finally, Bali. With you, possibilities feel endless and the world feels small. Writing this I am reminiscing on countless memories but mostly excited for what is to come. Thank you for the love and for standing by me always.

The Jacobs lab has this thing that attracts beautiful people. When I first started as an intern, I had the best supervisor in the world ever in the universe. **Gerrald**, you are one of those people born to do science and it was a privilege to have learned from you. We quickly became friends which made my days a fun time, sharing either cups of coffee or glasses of wine. Never forget the koala hugs either! **Nina**, even though we did not share much time together I was always quite impressed by your work ethic, you looked fearlessly ready to take up whatever SVA came your way! **Elise**, you ended up being a surprise to me, as we only got close after I joined as a PhD and we worked together in the SVA project. It was a tough time

during COVID-19 but I was glad we could support each other! **Elias**, it was always fun being around you and watching your family grow! Also happy you ended up meeting my grandparents in Portugal, when you came with Stephanie and baby Nova to pick up your van, you are always welcome to come back! **Grace**, my voice of reason. You are an absolute gift and I will be forever thankful to have you as a friend. To all the good times in and out of the lab, all the shared beers, dinners and parties, in Amsterdam, Portugal, Ireland and Canada! Hopefully many more to come. **Judith**, it is impossible to step into the stem cell lab without thinking of you. With your high spirits and unmatched perseverance, creating new brands of stem cell medium day after day (together with Grace). All the BBQs and borrels were more fun with you! **Colette**, you were the first postdoc in our lab and we needed your calm and stable presence. Thank you for always being ready to help, for all the talks and support! **Jimi**, we connected since day one. You are an incredible scientist, colleague and friend, always with fun ideas up your sleeve! Thank you for your trust in me, especially with your babies Jaspurr and Tchotchke. I know you have made an awesome life for yourself back in the US and I will come visit sooner or later! **Jade**, we missed each other in the lab but not out of it. We have the same taste in music and shared a few memorable moments on the dancefloor, thank you for your lovely energy! **Gonzalo**, my foodie friend. Thank you for all your support with the FACS and overall experiments in the lab. Very happy we continue to meet up, especially with lovely Aurelia – you are a wonderful father! **Sanaz**, you put the "z" in "snazzy"! You are incredibly dedicated to your work and go above and beyond to make sure you do it well, like bringing a GoPro attached to your head to make sure you do not blackout while culturing stem cells! Thank you for all the laughs. **Fred**, we were both interns in the lab at first, and I cannot begin to tell you how important your support was to me. You can do whatever you put your mind to, and your academic journey is the proof of that, it flows to a rhythm entirely your own.

Molecular neuroscience people, you are the dopamine to our receptors. Life in the lab was better together. **Marten**, thank you for your genuine interest, helpful discussions and motivating words when I needed them most. Your ambition is striking but you love life above all, and it is amazing how you juggle everything. Thank you for bringing me the best beer in the world, made by monks in Belgium! **Lars**, I do not remember how many Duvels we have shared, or the lab outings we have been on. Your dark sense of humour always brightened my days but most importantly, thank you for listening and for encouraging me always. **Marco**, I do not think I have ever seen you on a bad day. You take life with a smile on, and it is always good to have a catch up with you. The conference in Chile was one for the books! **Simone** and **Anke**, we did not hang out too much, but it was always nice to have a chat around the office! **Cindy**, you were the glue of our group, always committed to bringing people together, a very needed and admirable skill! Thank you for your care. **Martijn** and **Monica**, your support in the lab and overall good energy in the office was always something I could count on. **Swip**, thank you for the

deep talks, walks around Science Park, or the long days at work. You have been a steady presence during my PhD, and I appreciate all the time we spent together, including all the boogies you DJ'ed at! **Eddy**, back in the day… you and Gerrald came in a pack. Thank you for all the fun times impersonating the Backstreet Boys! **Jesse**, you and Erik were the protein dream team. The music in your lab never disappointed, and neither did the wooden works you were creating! **Erik**, my buddy and desk neighbour, thank you for your cheerful energy and for constantly rooting for me. It was a pleasure to witness your progress at Macrobian Biotech, and to see you help create Parkinnova Therapeutics from the ground up. So proud of you! **Reinofke**, **Noor**, **Joost**, **Marc**, **Matthijs** and **Luís**, post-COVID times were not easy, but you have all been a joy to work and share a beer with.

When I first started at SILS, I met four fellow interns from molecular neuroscience and other labs at SILS-CNS: **David**, the guitarist with a deep love for peanut butter; **Henok**, the unapologetic flirt; **Sarah**, the football-playing drum n' bass fan; and **Anna**, the badass cyclist with that thick British accent. I am incredibly grateful to have met you and shared the first year with you. Through all the crazy and the chaos, we found connection, we cheered for one another, and after all these years, we still show up. Thank you. I REALLY miss you!

In the Holstege lab, I had the privilege of working alongside many wonderful people. **Marc**, our bioinformatics magician, thank you for being so incredibly kind and helpful. I truly appreciate all the time you put in throughout my PhD, but especially during my final year. **Sven**, a medical doctor and geneticist by training, always with high spirits and available to help everyone, whether that meant solving a problem or setting up new collaborations. **Nina**, you were the first PhD candidate in the lab to graduate. Even in the little time we shared, your kindness did not go unnoticed. **Andrea**, you showed real resilience. Your road was a bumpy one, but you finished strong, uncovering some amazing findings on brain pathology in cognitively healthy centenarians. **Nicco**, you are as laid-back as they come, and proudly Italian through and through. You worked closely with me in my final year, always taking the time to teach me and brainstorm with me, and I honestly could not have finished this thesis without you. Together with Matthijs, Dani and Mayca, we went to Pisa, Italy, to see you and it was the best time. Still waiting for the trip to Rome! **Meng**, you worked mostly from Delft but never missed a lab meeting, sometimes coming in person. You truly are the kindest and it was always nice to see you around! **Alex**, you have just as much good vibes as you have smarts. Thank you for helping me find my footing in the world of bioinformatics, and for all the interesting conversations along the way. **Linda**, you go in head-strong at everything you do. You started as a project manager but soon became a PhD candidate yourself, continuing Nina's work. It has been great watching you take it on! **Susan**, you brought a lot of energy to the lab, always fun to be around! Thank you for being so supportive. **Yaran**, you are usually quiet, but you can take on any challenge when needed and do it well! I was

so glad you agreed to be our chef during the lab retreat that Janna and I planned together, teaching all of us how to prepare delicious Chinese food! **Janna**, you are the goofiest person I know! Planning the lab retreat with you was such a good experience. Thank you for the fun time and cheerful words, always. Even though years later we are still trying to plan that coffee, I am sure it will happen sometime! **Dani**, we met a long time ago, living as roommates together with Fred. It was great to share that time with you, getting to eat the best tortillas, but even better to later have you as a colleague and friend. **Michelle**, **Jenna**, **Georgii**, **Maruelle** and **Willemijn**, even though we did not work together that much, I was always happy to see you around the office or during the weekly lab meetings. **Mayca**, you are the friendliest girl I know! Thank you for all the fun times, for getting in on all the crazy, and the support, always. The practical team members that have left in the meantime, **Myke**, **Dominique**, **Marieke** and **Kimberley**, it was a pleasure to have met you and shared the occasional drink at fun outings and lab retreats!

My support network extends out of the lab, and it is thanks to them that my life in Amsterdam has been beyond words. They remind me that a life well-lived isn't just about the work you do, but also about the people who show up. **Juliette**, life never stops but I always felt that with you it would slow down. Meeting for yoga or pilates and afterwards enjoying a hot drink while feeling the warmth of the sun on our faces was enough to unwind and reset. We connected through music and dancing, and shared countless fun nights with the team! Thank you for your love and unconditional support. **Maria**, you have a grounding, creative way to go about life. From cold plunges to making ceramics together or wandering through weekend markets with Juju. I have appreciated the long chats, travel tips and overall laughs! **Carlos**, *meu amor*. With a loving heart, expanding mind and soulful body, I only wish we had shared more time together! Still, Easter brunches, Tenclub or Pride – it was better because of you. **Kili** & **Kath**, we have grown closer with time and in the last years you have been a constant presence in my life. It has been amazing sharing fun boogies with you, but most importantly seeing you now as parents of beautiful Lex fills my heart! **Mert**, you are such a kind and thoughtful person. Genuinely yourself, a social butterfly dancing away through the world! I really appreciate your reaching out and catching up, no matter how far. **Isa** & **Flavia**, always good energy when you are around, especially on the dancefloor! Thank you for all the joy and the many fun memories. **Antonio**, your traveller spirit knows no borders. Enjoying life above all, keep on *O'brezzing* through the canals of Amsterdam! **Felipe**, **Andrea**, **Agathe**, **Carmen**, **Ben**, **Conrad**, it has been a pleasure to share life in Amsterdam with you!

A special thanks to **Anton** for the most beautiful cover design! But also for the great times we have shared, from amazing themed parties hosted together with Grace, to sunbathing at the swimming spot or the lovely Sunday roasts with the old crew. Your energy shines bright! **Mary Ann** & **Liam**, you are incredibly lovely souls I was lucky enough to share beautiful moments with.

Being a Portuguese in Amsterdam also has its challenges but I was lucky I had my Portuguese friends that brought in those flavours back to my life. My oldest friend **Inês**, we were born in the same hospital almost 12 hours apart. We are night and day, always have been. We spent our lives together until you left Portugal for Canada, but in the summer of 2017 we both announced we were moving to the Netherlands, to Amsterdam and Leiden, how crazy is that? Thank you for your love and care, for your friendship and companionship. Now it is my turn to leave but it warms my heart to know you are well and happy together with Douwe. **Francisco**, your gentle admiration for Portuguese music, poetry, customs, gastronomy, nature connected me to you. Seeing our country through your eyes always made me feel more *saudade* than I would let on. Thank you for taking me back home in endless conversations, dinners or late-night guitar playing sessions. Keep the spontaneity and the warmth. **Gil**, our incredible travel photographer. Thank you for all the good times we shared with Diana and Francisco!

To my *franguitos favoritos*, **Salvador**, your energy fuels my heart, even with a broken foot in Phnom Penh! Your eagerness to see the world matches my own and I cannot wait to meet again somewhere, anywhere. Thank you for your authenticity, for your love and care. **Margarida**, I hold you dearest to my heart. There is no end to your light. All the long talks about life, health, travel or even spirituality – we are aligned. I deeply appreciate the love with which you welcome me always, and for making me feel I am part of the family! **Bruno**, working with you was always a treat! Your chaotic energy to my calm felt like flowing lava meeting the ocean. A natural process that is bound to happen. Thank you for all the laughs that brought tears to my eyes and for always making time for me. **Alexandra**, my artist solo traveller, and **Cunha**, the Rotterdam king, it has been lovely to get to know you!

Always present even from afar. **Sofia**, together since university, with Rita. There was a time when I was going more to your home than my own and I am so grateful you let me in with open arms. Together with the best furry buddies, having lovely dinners and the best parties! **Rita**, your strong will can bring down mountains. It was never meant to be easy but you did show them how the struggle made magic. Thank you for being part of Amsterdam too, and for the unforgettable trip to Iceland, I do hope we go back as promised! **Marisa**, you have been with me for longer than I can remember. Connected through chill vibes, peace & love, we have had amazing days that turned into nights. You mean so much to me, and it makes me so happy seeing your family grow with Fábio, first Lia and now baby Isa! **Inês**, life was always more fun together, thank you for showing me we can be friends for life! **Mike**, throughout all the change our friendship has remained rock steady. You have always shown me you cared and I am very thankful for it. **David** & **Diogo**, *meus putos!* You are like brothers to me and I truly appreciate all the time we have spent together. There are too many memories to point out here and what matters is that we have always celebrated one another. To many more dinners under the stars!

I finished this thesis during my travel in Southeast Asia and I appreciate all the people I met and connected with during that time. Strangers that quickly became friends. During a motorbike road trip in Vietnam, in a small island in Cambodia, in a night market in Laos, or in a pool in Bali, Indonesia. You do not know how much your words mattered.

My family has shown me endless support. When I first moved to the Netherlands I stayed at my cousin's house. **Patricia** & **Pieter**, with **Laurens** and **Ema**, I cannot find the words to tell you how grateful I am for having lived with you, as part of your family. We did not know each other before, but I had grown up hearing about my family in the Netherlands! We were family and so you took a chance on me. Seeing you two dancing, singing, cooking together – building a loving home – is nothing less than what I want for myself. Thank you for the long conversations over dinner with a glass of wine, and your never-ending support! To the rest of the family, **Ema** & **Julio**, **Robbie** & **Tessa**, **Eduardo** & **Monique**, **Stefan** & **Pleunie**, **Daniel** & **Tara**, it was always nice seeing you at family events!

To my aunts and uncles, it feels like a party whenever we are all together! **Hermínia** & **Luís**, thank you for caring for me since I was young. Over the years, every time I came back you always made sure I joined for lunch, feeding me delicious homemade food and sharing precious time together! **Fernando** & **Rosa**, **Graça** & **Ricardo**, **Rute** & **Rubinho**, **Luciano** & **Bela**, **Ana** & **Ricardo**, I am so fortunate for having such a close family and I thank you for all your support. **Sara** & **Hugo**, four years I stayed with you, with **Rodrigo** and **Francisco**, during university. It was home. I am incredibly grateful you took me in, as I understand it could not have been easy. Thank you for being a part of my journey, for cheering me on along the way. Thank you for your love and your kindness. To all my cousins, to **Patito**, I appreciate you all. And my grandparents too, avó **Benedita** & avô **Inácio**, thank you for having created this beautiful family!

My aunt **Iva**, you have always had a genuine interest in my life, no matter the distance. Thank you! Avó **Leonor** and avô **Eduardo**, you have been taking care of me since I was a baby and have shown your support throughout my entire life. I am forever thankful to you both.

Pai, independently of where I wanted to go or what I wanted to do, I could always count on your support. I appreciate all you have done for me and continue to do – embracing my diving adventures! And to my little sister **Lara**, we may be twenty years apart, but watching you grow into your own person is one of my joys in life.

Diogo, I have always been so proud of you and felt so lucky to have grown together, to have you as my brother. You are an incredibly knowledgeable person, and I love to learn from you. So much fun to be around, putting on tunes every chance you get. Thank you for being you and always being there for me, for your support and all the fun times!

Mãe, you are an amazing woman who has given me all the love and care one could hope for. You shaped me into who I am today, and I thank you for absolutely everything: your words of wisdom, your reassurance, and your tight hugs. Thank you, and **Victor**, for joining me on my travels, in Thailand! It was amazing to share that time with you and see your smile shining so bright.

About the Author

Rita Guimarães (1995) was born and raised in Alenquer, Portugal. Following her passion for nature and life sciences, she enrolled in the Bachelor's programme in Cellular and Molecular Biology at NOVA University Lisbon in 2013. During her studies, she was first introduced to experimental biology, which sparked her interest in scientific research.

In 2016, Rita started the Master's programme in Molecular Genetics and Biomedicine at NOVA University Lisbon. A year later, she decided to pursue an Erasmus research internship abroad and joined the laboratory of Dr. Frank Jacobs at the Swammerdam Institute for Life Sciences (SILS), University of Amsterdam, the Netherlands. There, she investigated the evolutionary history of the transcriptional repressor ZNF675 and its role in neuronal gene regulatory networks. During this project, she gained experience in molecular biology techniques while developing a strong interest in comparative genomics and bioinformatics. She obtained her master's degree in 2018.

The following year, Rita returned to Amsterdam to join the laboratory of Dr. Frank Jacobs as a PhD candidate under the supervision of Dr. Frank Jacobs at SILS and Dr. Henne Holstege at Vrije Universiteit Amsterdam. Her doctoral research focused on the role of repetitive DNA in the genetic architecture of Alzheimer's disease, combining computational genomics with experimental approaches to investigate how structural variation within transposable elements and tandem repeats may contribute to disease susceptibility. During her PhD, Rita also contributed to teaching activities as a lecturer for the master's course *Human Genome Biology and Evolution* and supervised master's students in their research projects.

Rita has since left Amsterdam. She first embarked on a 14-month trip in Southeast Asia, where she finalised this thesis. During her travels she discovered scuba diving and spent the final two months volunteering as a diver in Sulawesi, Indonesia. She will soon move to Australia to pursue this newfound passion, with the ambition of combining her scientific background with marine conservation and contributing to the protection of ocean ecosystems.