In silico detection of variable number tandem repeats associated with Alzheimer's disease from short-read sequencing data

Francesca C. Lucas

Technische Universiteit Delft

TU Delft

In silico detection of variable number tandem repeats associated with Alzheimer's disease from short-read sequencing data

by

Francesca C. Lucas

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Friday July 16, 2021 at 1:00 PM.

Student number:	4140699	
Project duration:	September 7, 2020 – July 16, 2021	
Thesis committee:	Prof. dr. ir. M. J. T. Reinders,	TU Delft, supervisor
	Dr. ir. C. C. S. Liem,	TU Delft
	Msc. N. Tesi,	TU Delft, supervisor
	Dr. H. Holstege,	VU Medical Center

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Preface

Five years ago, I decided to make the switch from Industrial Design Engineering to the Computer Science master. Finally, a challenge! During the switch I discovered my love for computer- and data science, especially when it can be applied to a subject as fascinating as neurology. For this MSc thesis I got to do exactly that: making use of the possibilities computer science offers to get valuable insights out of large amounts of data, and even contributing to unraveling the genetic basis of Alzheimer's disease. I could not think of a MSc project I would rather have done. I hope you enjoy reading this thesis as much as I did making it.

This work would not have been possible without the help from many people. First of all, I would like to thank Marcel Reinders for giving me the opportunity to make the switch, for which I'm eternally grateful, as well as the opportunity to do this MSc thesis. Your feedback, insights and sharp-witted questions have contributed to my critical thinking. To Niccolò Tesi, thank you for guiding me through the universe of bioinformatics research, the generous amounts of skype meetings and constructive feedback. Thank you Henne Holstege, for welcoming me in your wonderful research group and giving me the opportunity to work on this captivating topic. I would like to give a special thanks to my wonderful friends Maaike, Mila, Rebecca and Hester, for their enthusiasm, (virtual) coffees together and moral support throughout these years. Finally, I would like to thank my family for their endless source of love, support and encouragement, I could not have done this without you.

Francesca C. Lucas
The forests of Vorden, July 2021

1

Glossary

Term	Definition
Allele	Alleles are different forms of the same genetic region.
Anchored IRR	For paired-end reads: when one of the reads is an IRR and the other can be mapped to flanking sequence.
Coverage	The number of unique reads that include a given nucleotide in the reconstructed DNA sequence.
Diploid	Two sets of chromosomes in a cell (corresponding to the number of possible alleles).
Flanking reads	Reads which contain part repeating region and part flanking region.
Flanking region	The sequence on either side of a repeat.
Fragment length	Applicable to paired-end reads, where fragment length corresponds to the size of the two reads plus the insert.
Genotype	The genetic sequence of an organism.
Haploid	Single set of chromosomes / genome.
In repeat read (IRR)	Read which completely consists of repeating material.
In silico	Conducted or produced by means of computer modelling or computer simulation.
Indel	An insertion or deletion of nucleotides in the genome of an organism.
Long read data	Sequencing data for which reads have a size in the range of 10 000 - 100 000bp.
Motif	The sequence that is being repeated.
Paired IRR	For paired-end reads: when both reads are an IRR.
Paired-end read	Pair of two short reads, connected by an unsequenced middle section.
Pathogenic	Causing disease.
Phenotype	An observable physical property of an organism.
Read length	The size of a read, expressed in number of base pairs (bp).
Repeat number	How often a motif is repeated.
Short read data	Sequencing data for which reads have a size in the order of hundreds (100-600bp).
Spanning reads	Reads which fully encompasses a repeat, i.e. this includes the whole repeat as well as a part of the flanking regions.
Tandem repeat	A repeated DNA sequence where the repeats are adjacent.
Variable number tandem repeat	A tandem repeat with a varying number of repeats throughout a population.

2

Brief introduction biological concepts

2.1. Genome

A **genome** is the complete set of genetic material from an organism, which is contained within the **DNA**. A DNA sequence consists of nucleotides (denoted by A, C, G and T). The human's genome has a total length of 6 billion nucleotides, divided over 46 pieces: the **chromosomes**. These chromosomes are grouped into 23 pairs, with each chromosome in a pair coming from one parent. Because the chromosomes are paired our genome is **diploid**.

Each single chromosome consists of two strands (**double-stranded**), linked through chemical bonds, always connecting A's to T's and C's to G's. Two coupled nucleotides make a **base pair**. When looking at a single strand of DNA, the opposite strand is its **reverse complement**. For example, the reverse complement of CAAC would be GTTG. The ends of the chromosomes are called **telomeres**. As cells divide, the telomeres shorten, leading to shortening telomeres as we age.

Each cell contains the complete DNA, but only a small part is used depending on the cell type. DNA is like a cookbook, with many recipes encoding the creation of various **proteins** (the building blocks of our body). One recipe at a time gets transcribed into **RNA**, which is single-stranded and can be translated into proteins. The parts of our DNA that get translated into proteins are what we call **genes**, or **coding regions**. Different versions of a gene are called **alleles**. Genes only form a small part of our DNA, in between are large stretches of **non-coding regions**. Some are known to regulate cellular processes, such as the rate of protein production. For many other non-coding regions, their functionality remains unknown. It is suspected, however, that they help protect coding regions from damage, simply by being so abundant that mutations are more likely to happen in these non-coding parts.

Mutations can happen at all times and for a variety of reasons. There are many types of mutations that can occur, of which we will discuss the three relevant to this paper. The first type is a single nucleotide polymorphism (**SNP**), where one nucleotide gets substituted by another. The second type are **indels**, an abbreviation for insertions and deletions, where nucleotides are inserted or deleted from the sequence. The third type are **tandem repeat** variations, where there is a repeating sequence, and the number of repeats changes.

2.2. Reference genome and sequencing

To study genetic variation, we need a way to get an individual's genome and compare it to that of others. To do this, DNA gets **sequenced**: the process of 'reading' its sequence of nucleotides. However, it's not possible to read all 6 billion nucleotides in one go. Instead, we get overlapping bits and pieces, called **reads**, which have to be reassembled into one complete genome. This is done by mapping them to a **reference genome**. A reference genome is a single-stranded consensus genome, based on the genomes of several individuals. Using this as a template, reads can be allocated to their most probable position.

Genomes are sequenced multiple times, producing different, partially overlapping reads of the same area.

How often each nucleotide of the genome is captured in a read is called **coverage** (or depth). The higher the coverage, the more confidently we can reassemble an individuals genome. The longer the reads, the easier the reassembly. (more confident on variations) **Short reads** are currently the standard, with 100-200 nucleotides per read. The number of nucleotides is denoted by **bp** (base pairs). The field is transitioning towards **long reads** (tens of thousands of bp long), however, they are not as commonly available as short reads yet.

In this paper we use a specific type of short reads: **paired-end reads**. These are pairs of two connected short reads, sequenced from each end towards the middle / each other. There is a piece of DNA between these reads that is not sequenced. The total length of the two reads plus the unsequenced middle part is the **fragment length**.

2.3. Alzheimer's disease

Alzheimer's disease (AD) is a neurodegenerative disease, leading to damaged or destroyed neurons. This initially affects cognitive skills such as memory and problem-solving, but will eventually lead to difficulties with basic bodily functions such as swallowing and walking. Despite the prevalence of AD, its mechanics are not adequately understood.

One of the hallmarks of AD are **amyloid plaques** in the brains of AD patients, which are clumps of a protein called amyloid beta (**A**β). It is believed that the Aβ plaques affect synapses and thereby neuron-to-neuron communication, contributing to neural cell death.

Another hallmark of AD is the formation of **tau tangles**, an accumulation of tau proteins. Tau proteins stabilize microtubules, vital for the structure and functioning of neuronal cells, thus abnormal tau leads to unstable microtubules. Furthermore, tau tangles block the transport of essential molecules, such as nutrients, inside neurons.

Master's thesis

In silico detection of variable number tandem repeats associated with Alzheimer's disease from short-read sequencing data

Francesca Lucas

Department of Pattern Recognition and Bioinformatics, Faculty EEMCS, Delft University of Technology, The Netherlands

Abstract

Motivation: Alzheimer's disease (AD) is a highly prevalent disease whose genetic risk factors remain largely unknown. One potential genetic risk factor is tandem repeat expansions, which have been associated with over 40 diseases, most of which affect the nervous system. Detecting VNTRs from short-read data is a challenging task, leaving many VNTRs unidentified. To date only one variable number tandem repeat (VNTR) expansion (in the ABCA7 gene) has been linked to AD. We hypothesize there are many more VNTR expansions to be discovered that associate with an increased risk of AD.

Results: We created a pipeline with which we overcame the common limitations of VNTR detection (namely, the need for a predefined set of repeats and limited detectable VNTR sizes due to read length). We performed a genome-wide search for VNTRs with a motifsize $\geq$ 7 bp that show repeat size variations associated with AD. We detected 71 VNTR expansions and 1242 contractions, including expansions in genes *ADAMTSL3, ARHGEF10, DIP2C, EVC2, GRM8, MPPED1, PID1* and an expansion in the *SCIMP* gene close to a well-known AD single nucleotide polymorphism (SNP). Our pipeline is, to our knowledge, one of the very few to detect VNTRs exceeding read length without a predefined set of repeats. It is able to detect both previously reported and novel VNTRs, resulting in a promising set of VNTRs showing an association with AD.

1 Introduction

Alzheimer's disease (AD) is a complex neurodegenerative disease, characterised by a steady decline of cognitive capabilities. It is currently the most common form of dementia, the most prevalent cause of death at old age and estimated to affect 50 million people worldwide. This number is predicted to be tripled by 2050, making AD one of the major health challenges of the 21st century [23]. Despite its high prevalence, the biology underlying AD remains largely unknown. Understanding the genetic risk factors for AD is an essential step in developing and refining a cure. Based on twin studies, an estimated 60-80% of the risk of AD has been attributed to genetics [20], with an approximate 30% of the genetic risk being attributed to the ε4 allele of the *APOE* gene [4, 8], making genetics an essential research direction to learn more about AD.

One of the recent efforts to determine the genetic risk factors of AD was a genome wide association study (GWAS), analysing 680K samples, of which 110K were AD patients. They identified 83 single nucleotide polymorphisms (SNPs) associated with an increased risk of AD [5]. Although SNPs are an extensively studied type of genetic variation, they can not always be directly related to biological consequences affecting the disease they have been associated with. A recent example of this is the AD association with the *ABCA7* gene. For African Americans the association could be explained by a SNP, leading to a premature termination codon (PTC). However, this did not hold for Caucasian cohorts, leading the researches to explore other types of genetic variations. They discovered a tandem repeat expansion that leads to isoforms due to alternative splicing, explaining the *ABCA7* association to AD for Caucasian cohorts [9].

Tandem repeats (TRs) are a type of genetic variation where a piece of DNA sequence, the motif, is adjacently repeated at least two times. Based on the size of the motif, we distinguish between a short tandem repeat (STR), with a motif size of 2 - 6bp, and its longer counterpart, a variable number tandem repeat (VNTR) with a motif size $\geq$ 7 bp [49]. Tandem repeats tend to be unstable in their number of repeats: when the repeat number is increased or decreased, we speak of an expansion or contraction, respectively.

The identified VNTR in the *ABCA7* gene has a repeating motif of 25bp, with a pathogenic boundary around 230 repeats (5720bp total TR size) [9]. Furthermore, the VNTR is in linkage disequilibrium with the rs3764650 SNP (D'=0.92 and r^2=0.23) [9], showing that AD-associated SNPs can be an indication for TR variations nearby. To date, the *ABCA7* VNTR expansion is the only TR variation that has been associated with AD specifically. Still, more than 40 known diseases, most of which related to the nervous system, are associated with repeat expansions [56]. For example, a CAG repeat (pathogenic at 35+ repeats) in the *HTT* gene causes Huntington's disease [50], and a GGGGCC expansion (pathogenic at 24+ repeats) in the *C9ORF72* gene causes amyotrophic lateral sclerosis (ALS) [21] and fronto-temporal dementia (FTD) [28].

We hypothesize that there may be more TR variations to be discovered associated with an increased risk of AD. So far, mainly STRs have been

studied due to technical constraints imposed by the length of short reads (100-200bp). As VNTR sizes often exceed short read length, it becomes challenging to detect and estimate their sizes. Long read (10Kb-100Kb) data would be able to capture and genotype these VNTRs, however, long reads are not as accessible yet, making larger samples sizes unfeasible at the moment and leaving most of the VNTRs understudied.

We set out to push the boundaries of detecting VNTRs using paired-end short read data. In order to do so, we utilise ExpansionHunter (EH) [12, 13] and ExpansionHunter Denovo (EHdn) [11], which are computational tools able to detect VNTRs exceeding short read length. EHdn has been used to characterize genome-wide repeat expansion variations [18] and to identify TR expansions associated with autism [67], proving the potential of EHdn. However, our pipeline of chaining EHdn and EH together does not only overcome the common limitations (namely, the need for a predefined set of repeats and limited detectable VNTR sizes due to read length), but also provides accurate size estimates. This kind of pipeline has, to our knowledge, not been reported yet. We utilised this pipeline to detect common (occurring in >1% of cases [60]) VNTRs (motifsize $\geq$7bp), whose contractions or expansions are associated with an increased risk of AD.

2 Methods

2.1 Study population & data preprocessing

In this study, we use a sample of individuals from the Alzheimer's Disease Sequencing Project (ADSP). The ADSP project aims to identify gene variants that are a risk factor for, or protect against, Alzheimer's disease (AD) by sequencing and analysing genomes of a large number of well phenotyped individuals. Additional information about the ADSP study design, AD diagnosis assessment as well as ethical committee approvals are publicly reported on the ADSP website[1].

We 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) families, retaining 1 sample per family, (iv) phenotypes, retaining those with label 'no dementia' (controls), and 'definite AD' or 'probable AD' (cases). Except for filtering the samples, we directly use the paired-end sequencing CRAM files as supplied by ADSP. The CRAM files are aligned to the GRCh38 build (*GRCh38_full_analysis_set_plus_decoy_hla*[2]). We estimated coverage for each sample using mosdepth [57].
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)[3].

2.2 Summary of the pipeline

An overview of the pipeline is given in Figure 1. The first step in the pipeline is the discovery of candidate VNTRs. To do so, we used ExpansionHunter Denovo (EHdn), which identifies In-Repeat-Reads (IRRs) (i.e reads consisting of a repeating motif) for each sample. 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 cases compared to controls are reported by EHdn: these are the candidate VNTRs. Because EHdn considers a subset of all aligned reads (only IRRs) and its size estimates are essentialy a proxy for coverage, the output should be interpreted as an indication that a VNTR may be present at a specific location.

To obtain more precise repeat size estimates, we use ExpansionHunter (EH) to further analyse the candidate VNTRs identified by EHdn. However, prior to running EH we need to convert the reported candidate VNTRs into a variant catalog. 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 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 VNTR based on the reference genome during the variant catalog creation. EH then analyses each of the repeats from the 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. To do so, we first test for cohort-wide differences by comparing the distributions of repeat size between AD cases and healthy controls using the Wilcoxon Rank Sum Test. This gives us a set of VNTRs that show cohort-wide differences. Since expanded/contracted VNTRs are likely to occur in a subset of individuals, additionally we implemented an outlier analysis to specifically compare outlier counts between AD cases and controls using Fisher's exact test and the odds ratio (OR). We take the union of the two Fisher's-detected and OR-detected sets and end up with a set of VNTRs that are expanded and a set of VNTRs that are contracted in AD cases.

The identified sets of AD-associated VNTRs were put in genomic context, annotating whether they fall within or near either a gene or a functional non-genic element (e.g. enhancer), and we check for gene-set enrichment. Furthermore, we check the genomic context for GWAS identified SNPs associated with AD using snpXplorer. Finally, we report the most promising VNTRs that show an association with AD.

2.3 Candidate VNTR discovery

2.3.1 Identifying candidate VNTRs using ExpansionHuner Denovo

To identify candidate VNTRs without a predefined set of repeats, we run ExpansionHunter Denovo (EHdn) (v0.9.0) with each of our samples. We first generate an 'STR profile' (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 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.

We run the '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 outputted motif is the lexicographically smallest repeat unit under circular permutations and reverse complement operations.

Additionally, we run the 'outlier.py' executable with the 'locus' option in a similar fashion. Compared to the 'casecontrol' 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

[1] `niagads.org/adsp`

[2] `ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/GRCh38_reference_genome`

[3] `dss.niagads.org`

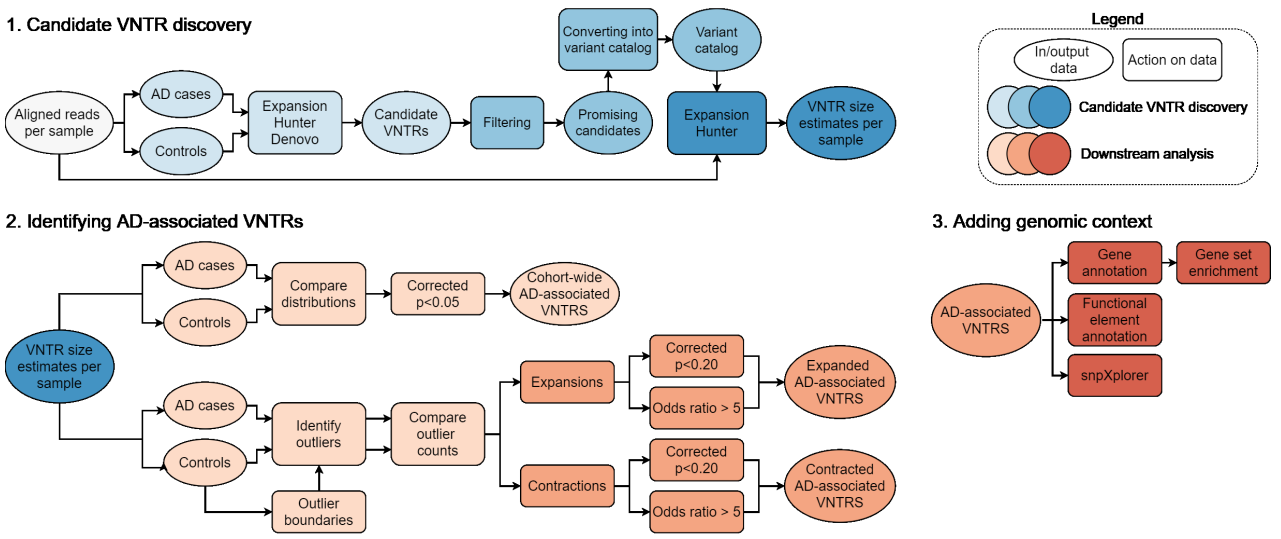


Fig. 1: An overview of our pipeline, distinguishing between the three main steps: (i) Candidate VNTR discovery, (ii) Identifying the VNTRs associated with AD and (iii) Adding genomic context tot the AD-associated VNTRs.

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. The outputted repeats are our candidate VNTRs with which we will proceed.

2.3.2 Filtering candidate VNTRs

To focus on the most interesting repeats, we filter the candidate VNTRs outputted by EHdn. Since we run EHdn in two settings ('casecontrol' and 'outlier'), we have two outputs to select from. For the case-control outputs, we select those with an uncorrected p-value < 0.05 to select the repeats showing differences between cases and controls. For the outlier outputs, we select the repeats where the top z-score for a case is ≥ 10 to select the most variable VNTRs and where a minimum of 15 (3%) cases has a z-score > 1 to select the common repeat expansions. For both outputs, we select only the repeats with a motif size > 6 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.

2.3.3 Converting candidate VNTRs into a variant catalog

In order to run ExpansionHunter (EH) on the candidate VNTRs identified by ExpansionHunter Denovo (EHdn), the repeats need to be encoded in a 'variant catalog'. Such a 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 basepairs, but the approximate locations ($\pm$500bp) as outputted by EHdn are too imprecise to be used directly. Therefore we must identify more precise start and end coordinates.

To determine the precise start and end positions, for each candidate VNTR we extract the corresponding sequence from the reference genome using the pysam package[4] between the coordinates outputted by EHdn, extended by 500bp on each side . We then scan the obtained reference sequence for occurrences of the repeat motif, keeping track of perfect stretches of copies. Each perfect stretch of copies is extended if an adjacent imperfect copy is followed by another perfect copy of the motif. We assume that the longest stretch in the reference is most likely to be the source of the repeat, thus we select the longest stretch of copies. If there are multiple positions with the maximum number of copies, we can not establish which is the most likely source of the repeat. In such cases, we decided to be conservative, and we retain all positions. Only the real source of the repeat will come up as significant in our downstream analysis, so the only downside of this is a larger variant catalog, meaning longer running times.

In addition to defining more accurate start and end position for the candidate VNTRs, an additional check on the repeated motif is needed. In fact, 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, we should decide between the outputted motif and its reverse complement. To do so, for each VNTR, we scan the extracted reference sequence for the reverse complement of the motif as well and again select the longest stretch of copies. We compare the results from both the outputted motif and its reverse complement and select the positions with the longest stretch of copies. The re-established positions plus the selected motif are formatted and outputted. The detailed algorithm can be found in Appendix A and the source code is available through GitHub[5].

2.3.4 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 ExpansionHunter (EH) for each sample individually. We run EH (v4.0.3) with the variant catalog, reference genome and the sample's CRAM files as inputs. 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.

[4] `github.com/pysam-developers/pysam`

[5] `github.com/francesca-lucas/ehdn-to-eh`

2.4 Identifying VNTRs associated with AD

After running EH, a diploid repeat size estimate for each sample and each VNTR is available. We subsequently divide the diploid repeat size 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 the shortest allele, the longest allele and the sum of these two (summed allele), separately. We then take the union of the VNTRs detected based on these separate allele values.

2.4.1 Cohort-wide detection of AD-associated VNTRs

We define cohort-wide VNTRs as those that show a significantly different distribution for cases compared to controls. To detect these, we compare the repeat size distributions of cases and controls using the Wilcoxon Rank Sum test and select the entries with a false discovery rate (FDR) corrected p-value of < 0.05.

2.4.2 Identifying outliers

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

$$\begin{aligned} x &< Q1 - c \cdot IQR \quad \text{or} \\ x &> Q3 + c \cdot IQR \end{aligned} \tag{1}$$

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.

2.4.3 Outlier-based detection of AD-associated VNTRs

We define common pathogenic VNTRs as those that show an expansion or contraction in $\geq$1% of cases, which is $\geq$ 6 cases in our study. To detect expanded VNTRs, we count the number of datapoints above the upper boundary (Equation 1) 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 (OR). To detect contracted VNTRs, we start with the counts of datapoints below the lower boundary (Equation 1), 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 (OR) > 5 and that occur in $\geq$1% of cases (in our study, $\geq$6 cases).

2.4.4 Visualisation of AD-associated VNTRs using scatterplots

We visualised the identified VNTRs using scatterplots. As we are dealing with discrete data, we added some random noise to the datapoints to be able to see the point clouds. We use the scatterplots to visually check the identified pathogenic VNTRs.

2.5 Adding genomic context

2.5.1 Gene annotation

To get an idea of where a VNTR might exert influence we provide genomic context. We annotate genes based on the RefSeq genes (v98) from NCBI[6].

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) coordinate
- 'coding' from coding start (cdsStart) to end (cdsEnd) coordinate
- 'exon' from each exon start (exonStarts[i]) and its corresponding end (exonEnds[i]) coordinate
- 'telomere' within 5Mb of the ends of chromosome arms
- 'promotor' as 1Kb before transcription start (txStart)

A VNTR gets an annotation when it overlaps the defined regions. We define a VNTR to be 'genic' if it overlaps one of these regions: transcription, coding or exon. If a VNTR has no direct genic or functional element annotation, we search for the nearest gene within 50Kb and annotate this as its 'nearest gene'.

2.5.2 Functional element annotation

There are many non-genic functional elements that influence biological processes (e.g. promoters or enhancers). In order to annotate these, we use the NCBI RefSeq Functional Elements table[7]. This table contains various functional elements that have been experimentally validated, such as regulatory elements, protein binding sites, mobile elements, recombination features and sequence features. We included all entries from the table in this annotation step. We use the start (chromStart) and end (chromEnd) positions to define each region and the type of element (name). If a VNTR has no direct annotation to a functional element, we search for the nearest functional element within 50Kb and annotate this as its 'nearest functional element'.

2.5.3 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[8]. 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. We selected the 'gene' annotations and combined them into our set of input genes. We used the default g:GOSt settings except for FDR correction as multiple testing correction. The default settings include annotations from all data sources made available by g:GOSt, which are the following: (i) Gene Ontology sources (GO molecular function, GO cellular component, GO biological process), (ii) biological pathway sources (KEGG, Reactome, WikiPathways), (iii) regulatory motifs in DNA (TRANSFAC, miRTarBase), (iv) protein databases (Human Protein Atlas, CORUM) and (v) Human phenotype ontology (HP).

---

[6] `genome.ucsc.edu/cgi-bin/hgTables/ncbiRefSeqCurated`

[7] `genome.ucsc.edu/cgi-bin/hgTables/refSeqFuncElems`

[8] `biit.cs.ut.ee/gprofiler/gost`

Table 1. An overview of the samples included in our study. EOAD indicates the percentage of early-onset Alzheimer's disease (onset age < 65) vs late-onset AD (onset age $\geq$ 65) within our AD cases. Ethnicity values correspond to: 0 = Non-Hispanic, 1 = Hispanic. The reported ages are those at AD onset for cases and those at inclusion into the study for controls.

	AD Cases	Controls
Total	513	616
EOAD	13%	n/a
Female	51%	63%
Ethnicity (0-1-NA)	502 - 8 - 3	492 - 124 - 0
Age	76 $\pm$9	80 $\pm$6
Age range	47 - 89	60 - 90
Coverage	38x $\pm$ 4	38x $\pm$ 5
Read length	148 $\pm$10	151 $\pm$0.5
Fragment length	369 $\pm$16	373 $\pm$16

2.5.4 Link to GWAS data using snpXplorer

In addition to visualising the promising VNTRs using scatterplots, it is interesting to put VNTRs in the context of known genome-wide association studies summary statistics. To do so, we use the web application[9], which has the possibility to superimpose association densities from multiple studies, displaying regional information such as SNP associations and structural variations [66].

3 Results

3.1 Study population

In this study, we used a sample of Alzheimer's disease (AD) cases (N=513, mean age at onset 76 $\pm$ 9 years) and non-demented controls (N=616, mean age at inclusion 80 $\pm$6 years) from the ADSP project. The AD cases were a mix of early-onset AD (EOAD, age at onset < 65, N=66), and late-onset AD (LOAD, age at onset $\geq$ 65, N=447). We analysed their paired-end short-read whole-genome sequencing data, with 98% of individuals having a read length of 150bp, while 2% (N=24) had a read length of 100bp. A summary of metadata on the study population can be found in Table 1.

3.2 Candidate VNTRs

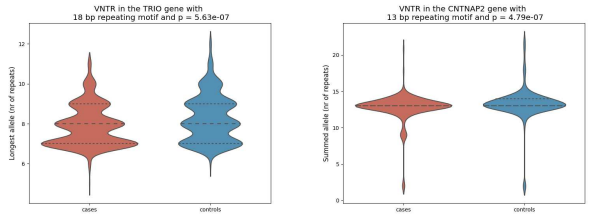
3.2.1 Candidate VNTRs identified by ExpansionHunter Denovo

The initial set of candidate VNTRs were generated with ExpansionHunter Denovo (EHdn). We ran both the case-control and the outlier analysis, detecting VNTRs occurring in, respectively, a high and low fraction of the cases. The case-control analysis returned 11 794 VNTRs, with 1319 (11%) having an uncorrected p-value < 0.05 and 41 (0.3%) having a Bonferroni corrected p-value < 0.05. The outlier analysis reported 319 091 VNTRs, with 3824 having at least one case with a top z-score $\geq$10. The full output for both analyses can be found in Supplementary Table 1. Of the candidate VNTRs returned by the case-control analysis, 97% also occured in the output of the outlier analysis.

3.2.2 Filtering candidate VNTRs and creating the variant catalog

We filtered the outputs to focus on the most promising candidate VNTRs. See Methods Section 2.3.2 for a detailed explanation of the applied filters. We retained 923 from the case-control analysis after filtering on motif size (motif $\geq$ 7bp) and p-value (p < 0.05). From the outlier analysis we retained 1771 candidate VNTRs after filtering on motif size (motif $\geq$ 7bp), z-score ($\geq$ 10), and support ($\geq$ 3%). Taking the union of these sets gives us a total of 2129 candidate VNTRs, which were transformed into a variant catalog of 3818 entries.

[9] snpxplorer.net



(a) VNTR in the *TRIO* gene with its FDR corrected p-value.

(b) VNTR in the *CTNAP* gene with its FDR corrected p-value.

Fig. 2: Violin plots for the two most significant VNTRs having a 'gene' annotation, detected with our cohort-wide analysis.

3.3 Identified VNTRs associated with AD

3.3.1 Cohort-wide AD-associated VNTRs

For each candidate VNTR, we tested for cohort-wide differences in the repeat numbers estimated by EH (see Methods 2.4.1). Our analysis returned 1064 regions with an FDR corrected p-value < 0.05 (Supplementary Table 2). Despite a significant difference according to the test, upon visual inspection the distributions did not differ much and it seems the differences were in the tails. The violin plots of the 10 most significant VNTRs with a 'gene' annotation can be found in Supplementary Figure 1, of which we show the two most significant in Figure 2.

3.3.2 Outlier-based AD-associated VNTRs

In addition to the cohort-wide analysis, we performed an outlier analysis on each candidate VNTR to identify differences in tails (see Methods 2.4.3). We detected VNTRs that show either an expansion or contraction for cases. The number of detected VNTRs are summarised in Table 2.

Expansions

We detected significant (FDR corrected p < 0.2) VNTRs based on each allele value, giving us 29 based on the shortest allele, 22 based on the longest allele and 24 based on the summed allele. There is overlap between these results; taking the union gives 52 unique VNTRs.

We detected 52 expansions with an odds ratio (OR) > 5 and occurring in more than 1% ($\geq$6) of cases, with 25 based on the shortest, 25 on the longest and 27 on the summed allele. Taking the union of the 52 Fisher's-detected expansions and the 52 OR-detected expansions gives us a total of 71 VNTRs showing an expansion in AD cases compared to non-demented controls.

Contractions

We detected 1237 VNTRs that show a contraction in cases compared to controls and have an FDR corrected p < 0.2. From those, 45 VNTRs came up as significant based on the shortest allele, 1138 based on the longest allele and 1056 based on the summed allele.

We detected 610 contractions with an odds ratio (OR) > 5 and occurring in more than 1% ($\geq$6) of cases, with 49 based on the shortest, 518 on the longest and 485 on the summed allele. Taking the union of the 1237 Fisher's-detected contractions and the 610 OR-detected contractions gives us a total of 1242 VNTRs showing a contraction.

Table 2. The number of detected AD-associated VNTRs showing either an expansion or a contraction in cases and based on which allele they were detected.

	Method	Shortest allele	Longest allele	Summed allele	Total
Expansions	Fisher's	29	22	24	52
	Odds ratio	25	25	27	52
				Total	71
Contractions	Fisher's	45	1138	1056	1237
	Odds ratio	49	518	485	610
				Total	1242

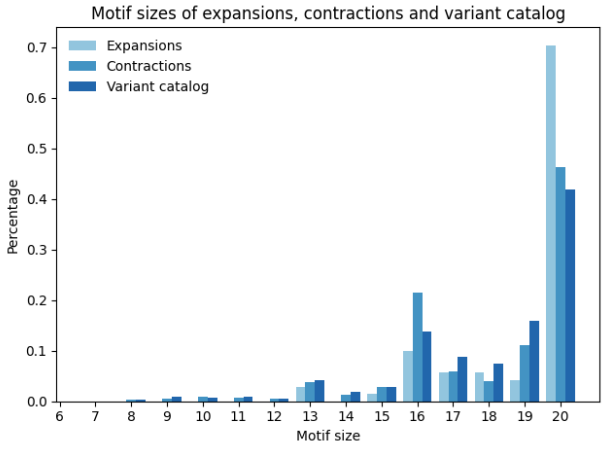


Fig. 3: The distribution of motif sizes in bp for the expanded (N=71) and contracted (N=1242) AD-associated VNTRs as well as for the whole variant catalog.

The distribution of motif sizes for the detected 71 expansions and 1242 contractions can be found in Figure 3. The full list of detected expansions and contractions can be found in Supplementary Table 3 and Supplementary Table 4, respectively.

3.4 Genomic context

3.4.1 Gene and functional element annotation

We annotated the whole variant catalog as well as the expansions and contractions that came up as significant. We noticed a telomeric enrichment in the regions in the variant catalog ($p < 1e\text{-}16$) and the detected contractions ($p < 1e\text{-}16$). For the significant expansions, the telomeric enrichment is less but still present ($p = 0.038$). None of the regions from the variant catalog occurs in promotor regions. The annotation frequencies are summarised in Table 3.

3.4.2 Gene set enrichment

We selected the 'gene' annotations from the expanded and contracted VNTRs and checked the two sets of genes for enrichment. As some VNTRs overlap multiple genes, this gave us 19 genes for the expanded and 218 genes for the contracted VNTRs.

For the expansions, 3 of the 19 genes were excluded by g:GOSt as they exclude uncharacterized LOC genes. For the remaining 16 genes, annotation terms that came up as significant were from the categories GO cellular component (GO:CC), biological pathway WikiPathways (WP) and

Table 3. Annotation frequencies of the VNTRs

	Variant catalog		expanded VNTRs		contracted VNTRs	
total	3818	100%	71	100%	1242	100%
genic	1611	42%	15	21%	371	30%
telomeric	1441	38%	10	14%	352	28%
promotor	0	0%	0	0%	0	0%
exonic	57	1.5%	0	0%	5	0.4%
func elem	25	0.7%	0	0%	11	0.9%
near gene (50kb)	964	25%	26	37%	400	32%
near func elem (50kb)	55	1.4%	7	8%	9	0.7%

protein database CORUM. An overview of the gene set enrichment can be found in Figure 4a and Table 4b, for which the full results can be found in Supplementary Table 6.

For the contractions, 33 of the 218 genes were excluded by g:GOSt as these were uncharacterized LOC genes. For the remaining 185 genes, annotation terms that came up as significant belong to the categories: GO molecular function (GO:MF), biological pathway sources KEGG and Reactome (REAC), protein databse CORUM, regulatory motifs in DNA from TRANSFAC (TF) and Human phenotype ontology (HP). An overview of the gene set enrichment can be found in Figure 4c and Table 4d, for which the full results can be found in Supplementary Table 6.

3.5 A selection of VNTRs

3.5.1 A selection of expanded VNTRs in genes

We made a selection of the most interesting expanded VNTRs based on their annotation. As there are no expanded VNTRs falling within functional element annotations, we select only those with a 'gene' annotation giving us a shortlist of 15 VNTRs, summarised in Supplementary Table 5. Upon visual inspection of their scatterplots, we noticed that the VNTRs in *GALNT17* and *DLGAP2* did not show a clear division between cases and controls and showed many controls with a large expansion (see Supplementary Figure 2). The *GRM8* gene was detected based on the shortest allele, and although there was not a clear separation visible on that axis, it did show a clear separation for the longest allele, therefore we decided to retain this VNTR in our selection. The remaining shortlist consists of 13 expanded VNTRs, which are summarised in Table 4.

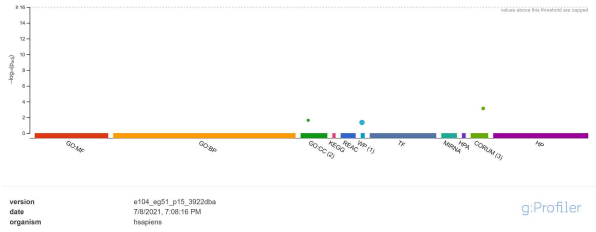
The corresponding scatterplots and snpXplorer images can be found in Supplementary Figure 3. In Figure 5 we highlight the three VNTRs in genes *ADAMTSL3*, *MPPED1* and *SCIMP*, as these showed a clear expansion for cases and AD-associated SNPs in the snpXplorer.

3.5.2 A selection of contracted VNTRs in genes

We curated five contracted VNTRs based on three criteria: (i) the VNTR has a 'gene' annotation, (ii) the scatterplot shows a clear separation for the contracted AD cases, (iii) the VNTR does not have neighbouring variant catalog entries with a similar motif (to avoid reporting technical effects) and (iv) snpXplorer shows activity for the concerned gene. Here we show the plots for the contracted VNTR in *RASA3* and *DNM2*. An overview of these five VNTRs are captured in Table 5, and its scatterplots and snpXplorer images in Supplementary Figure 4.

3.5.3 Which VNTRs get detected is sensitive to the outlier constant

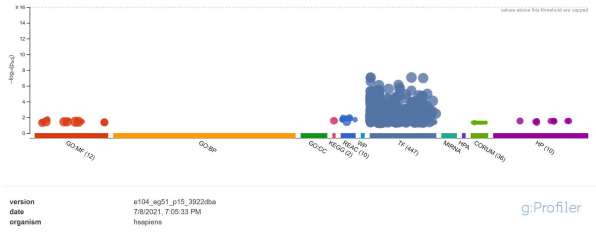
We noticed that the outlier constant greatly affects which VNTRs could be detected. Both smaller and larger constants result in the detection of different VNTRs that show a clear expansion for cases. We show a detected VNTR for each of the outliers constants 0.5, 1.5, 2.0 and 2.5 in Supplementary Figure 5. These four VNTRs show that a different constant



(a) A visual overview of the gene set enrichment analysis for expanded VNTRs.

Source	Term name	Adjusted p
GO:CC	cell trailing edge membrane	0.023202
GO:CC	uropod membrane	0.023202
WP	GPCRs, Class C Metabotropic glutamate, pheromone	0.044345
CORUM	Glutamate receptor 8 complex, metabotropic	0.000735
CORUM	mGluR2-mGluR8 complex	0.000735
CORUM	mGluR3-mGluR8 complex	0.000735

(b) The significant results from the gene set enrichment analysis for expanded VNTRs.



(c) A visual overview of the gene set enrichment analysis for contracted VNTRs.

Source	Term name	Adjusted p
GO:MF	acetylgalactosaminyltransferase activity	0.018199
KEGG	Calcium signaling pathway	0.026744
REAC	O-linked glycosylation	0.011054
REAC	DAG and IP3 signaling	0.011054
CORUM	PKC-alpha-PLD1-PLC-gamma-2 signaling complex	0.046943
TF	Factor: AP-2; motif: MKCCCSCNGGCG	0.000735
HP	Bilateral generalized polymicrogyria	0.028588

(d) The most significant result per source from the gene set enrichment analysis for contracted VNTRs.

Fig. 4: The gene set enrichment plots for the expanded and contracted VNTRs. The full results for this analysis can be found in Supplementary Table 6.

Table 4. An overview of the expanded VNTRs within genes, listing for each VNTRs the gene, their location on the reference genome (Location GRCh38), repeated motif & motif size, Fisher's p-value, its FDR corrected counterpart, the odds ratio (OR), number of cases above the outlier boundary (Cases), number of controls (Controls) above the outlier boundary and if the VNTR has been reported before by Linthorst et al. [44].

Gene	Location GRCh38	Repeated motif	Fisher's p	Corrected	OR	Cases	Controls	Status
ADAMTSL3	15: 83674672 - 83674772	ACACACATATATACATATAT (20)	3.3e-5	0.01	inf	13	0	novel
ARHGEF10	8: 1919435 - 1919475	CCATGGGTGATGGAGCTGTT (20)	1.9e-2	1.00	8.5	7	1	reported
CLDN14 & LOC107984737	21: 36496794 - 36496826	AAGGAAGGGAGGGAGG (16)	1.2e-3	0.16	8.0	13	2	novel
DIP2C & DIP2C-AS1	10: 658264 - 658288	ACCTGCCCCTGG (12)	2.6e-3	0.06	4.3	24	7	novel
EVC2	4: 5706427 - 5706447	ACATAGATAGATAGATAGAT (20)	8.6e-3	1.00	inf	6	0	reported
GRM8	7: 126903616 - 126903756	ATATATATATGTATATGTGT (20)	3.9e-4	0.09	1.6	166	143	novel
LINC02050	3: 80770373 - 80770613	AACGTACGTGCGCTCCTCTC (20)	1.4e-4	0.20	6.2	20	4	reported
LOC101928269	21: 35993052 - 35993084	AACTCACACACACCCC (16)	1.9e-2	1.00	8.5	7	1	novel
LOC101928764	13: 21290679 - 21290759	AATACAGATATGACACCCGC (20)	1.1e-3	0.16	2.8	31	14	reported
MPPED1	22: 43432765 - 43432841	AAAGGGAGGAGAGAGAGAG (19)	6.6e-6	0.002	22.3	18	1	reported
PID1	2: 229184495 - 229184512	ATATATATATATCCCGT (17)	2.3e-3	1.00	12.2	10	1	reported
SCIMP & ZNF594-DT	17: 5232351 - 5232365	AAACAGTGCAGTGT (14)	9.5e-3	1.00	9.7	8	1	reported
TRANK1	3: 36880047 - 36880065	AAACATACAAATATATGT (18)	4.1e-4	0.09	3.3	29	11	reported

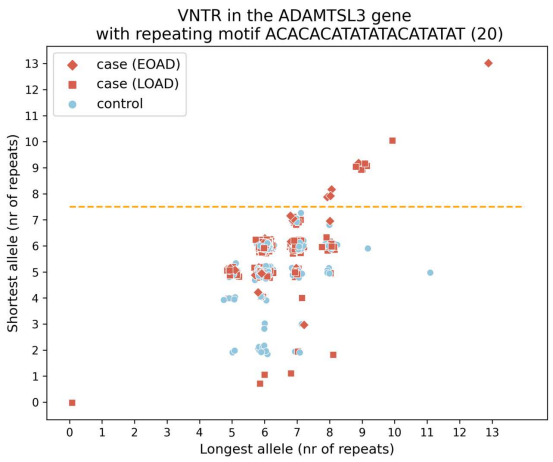
can lead to the detection of other expanded VNTRs, indicating a sensitivity to this constant. Here we highlight a VNTR in the *ALK* gene, detected with an outlier constant of c = 1.5, showing a massive expansion for the longest allele (Figure 7). This VNTRs is located at 2:29831155-29831164 on the reference, with an 'AAGAAGGAG' (9 bp) motif, 9 cases and 1 control above the outlier boundary, an uncorrected Fisher's p = 4.7e-3, FDR corrected p = 0.6 and OR = 11.

4 Discussion

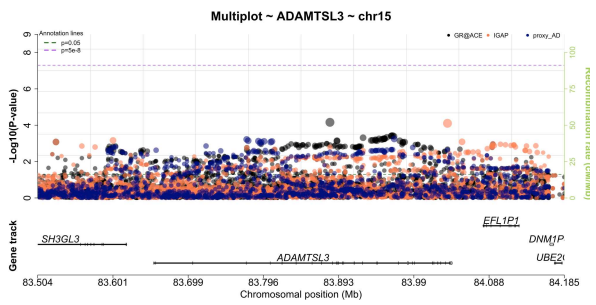
We present our findings of a genome-wide search for variable number tandem repeats (VNTRs) related to Alzheimer's disease (AD). We focussed on VNTRs with a motifsize $\geq$ 7 bp and whose total size exceed read length. We report 71 VNTR expansions and 1242 contractions, of which 15 expansions and 371 contractions occur in genes. The gene set analysis shows an enrichment for neurological elements, with some explicitly linked to AD. For example, within the enriched terms for the expanded VNTRs (Figure 4b) is the 'mGluR2-mGlurR8 complex', with mGluR2 activation triggering the production of Amyloid β [38] and leading to neuronal degeneration [41].

4.1 Expanded VNTRs

The largest expansion was detected using a different outlier constant (see Section 3.5.3). This concerns the VNTR in the *ALK* gene (Figure 7), showing an expansion on the longest allele, on which the largest expansion has an estimated size of 288 repeats. The VNTR has a 9bp

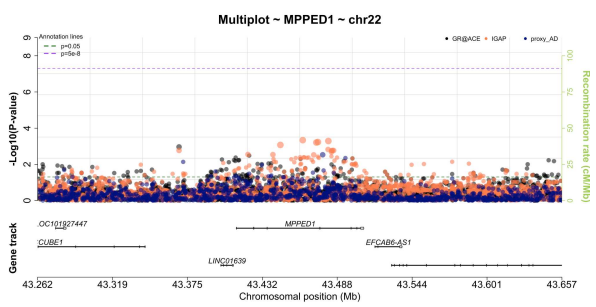


(a) The number of repeats on each allele for the expanded VNTR in *ADAMTSL3*, showing the outlier boundary as a dashed line.

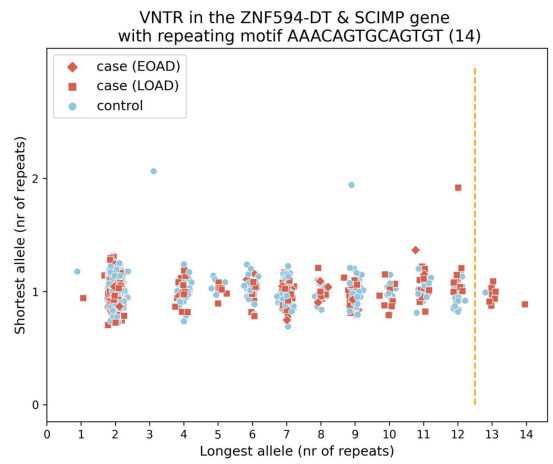


(b) The SNPs in and around *ADAMTSL3*, obtained from snpXplorer.

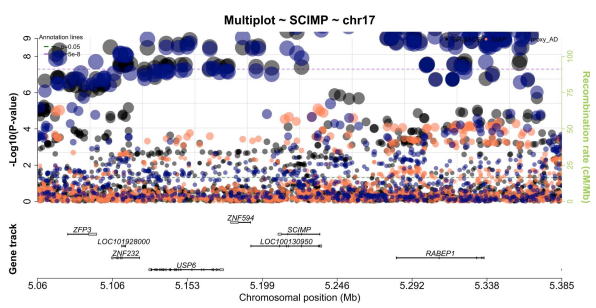
(c) The number of repeats on each allele for the expanded VNTR in *MPPED1*, showing the outlier boundary as a dashed line.



(d) The SNPs in and around *MPPED1*, obtained from snpXplorer.



(e) The number of repeats on each allele for the expanded VNTR in *SCIMP*, showing the outlier boundary as a dashed line.

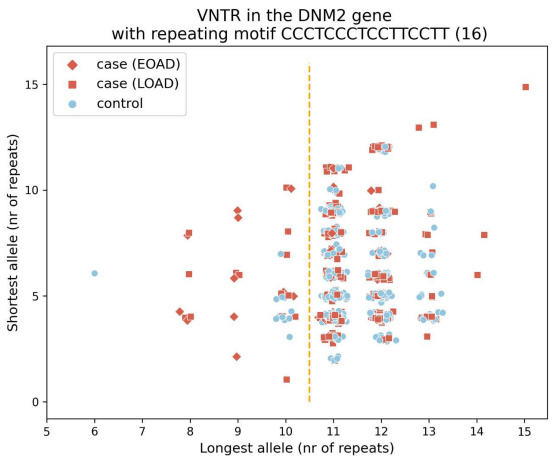


(f) The SNPs in and around *SCIMP*, obtained from snpXplorer.

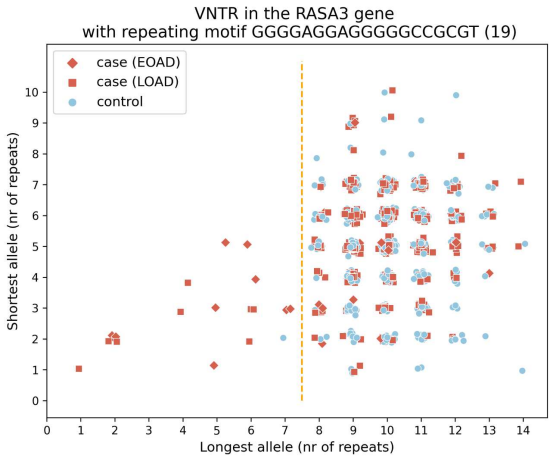
Fig. 5: For the three expanded VNTRs in *ADAMTSL3*, *MPPED1* and *SCIMP* from Table 4 we show the number of repeats on each allele and the regional SNPs as provided by snpXplorer. Similar plots for all entries from Table 4 can be found in Supplementary Figure 3.

Table 5. An overview of the contracted VNTRs near genes, listing for each VNTRs the gene, their location on the reference genome (Location GRCh38), repeated motif & motif size, the FDR corrected Wilcoxon rank sum test p-value, the odds ratio (OR), number of cases above the outlier boundary (Cases), number of controls (Controls) below the outlier boundary and if the VNTR has been reported before by Linthorst et al. [44].

Gene	Location GRCh38	Repeated motif	Corrected	OR	Cases	Controls	Status
C2CD3	11: 74117086 - 74117240	AATATATATATATG (14)	0.04	2.0	35	22	known
DNM2	19: 10758432 - 10758496	CCCTCCCTCCTTCCTT (16)	9.7e-4	17.3	14	1	known
PCCA	13: 100451910 - 100451985	CCTCTCCCTCTCTCT (15)	1.7e-4	11.8	19	2	known
RASA3	13: 114010765 - 114010784	CCCTCCCTCCTTCCTT (19)	3.1e-5	inf	16	0	novel
SEMA4D	9: 89462958 - 89462976	AGCGAGCGAGGGGAGGGG (18)	1.8e-3	10.0	19	4	known



(a) Contracted VNTR in the *DNM2* gene.



(b) Contracted VNTR in the *RASA3* gene.

Fig. 6: Two of the contracted VNTRs.

motif, with 9 cases and 1 control above the outlier boundary (uncorrected Fisher's p=0.0047. FDR corrected p=0.6, OR=11). ALK has been reported as a crucial protein in the tau-dependent neural degradation, as ALK causes abnormal accumulation of phosphorylated tau in neurons. The brains of AD patients showed significantly elevated ALK levels. Even more so, the pharmocological inhibition of ALK activity reversed the tau accumulation and memory impairment in transgenic mice [54]. Alk

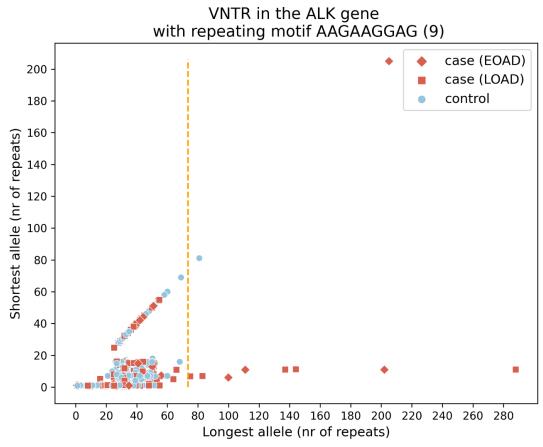


Fig. 7: Detected VNTR in the *ALK* gene showing a massive expansion, with c = 1.5 as the outlier boundary constant.

inhibition in *Drosophila* has been shown to extend lifespan and described as a target for longevity [72].

Furthermore, we selected the most promising expanded VNTRs based on their annotation (selecting those within genes) and a visual check of their scatterplots. These thirteen VNTRs are summarised in Table 4. Eight of the thirteen VNTRs (those in *ARHGEF10*, *EVC2*, *LINC02050*, *LOC101928764*, *MPPED1*, *PID1*, *SCIMP* and *TRANK1*) have previously been reported by Linthorst et al. (2020) [44], who used long-read sequencing to detect genome-wide structural variations (SVs), including tandem repeats and VNTRs. We consider the five VNTRs that have not been reported by [44] as novel: this concerns the VNTRs in *ADAMTSL3*, *CLDN14*, *DIP2C*, *GRM8* and *LOC101928269*. This suggests both that our pipeline works well at detecting VNTRs and that we are able to detect novel ones.

The majority of the genes in which we detected VNTRs have either been linked to AD, linked to a disease sharing a genetic similarity with AD or are active in the central nervous system.

ADAMTSL3 has been described as a candidate locus for schizophrenia [14], which shows a link to AD [39], and colorectal cancers [40]. The protein encoded by *ADAMTSL3* is a glycoprotein and is suspected to play a role in cell-cell interactions [6]. The overarching ADAM and ADAMTS proteins are believed to play a role in various pathologies, including AD, and have been mentioned as promising drug targets [6]. The VNTR in *ADAMTSL3* has a 20bp motif and shows an expansion with the same number of repeats on both alleles (Figure 5a), where 13 AD cases and 0 controls fall above the outlier boundary (FDR-corrected p = 0.01, OR = inf). snpXplorer shows suggestive association signals with AD for this

genomic region (Figure 5b), with rs34321903 being the most significant (p = 6.9e-5). Furthermore, the leading SNP in the region (rs34321903) is an intergenic SNP for which consequences at gene and protein level are unknown, thus making this VNTR a very promising one.

MPPED1 is listed as one of the regionally enriched genes in the brain cortex [15]. Furthermore, it has been included in the expression analysis of the transcriptional signatures of tau and amyloid neuropathologies [7]. A SNP in *MPPED1* has been associated to autism in GWAS [45], and autism has been shown to share genetic similarities with AD [30, 63, 47]. The VNTR in *MPPED1* has a 19bp motif and shows various expansions on both alleles (Figure 5c), with 18 cases and 1 control above the outlier boundary (FDR corrected p=0.002, OR=22.3). For this region, snpXplorer shows various SNPs (Figure 5d) with the most significant SNP having a significance of p = 4.6e-4, making this VNTR another very promising one.

SCIMP has been implicated as a risk gene for AD, showing significant association [31]. The VNTR in *SCIMP* has a repeated motif of 14bp and shows an expansion on the longest allele (Figure 5e), where 8 AD cases and 1 control fall above the outlier boundary at 12 repeats (uncorrected p=9.5e-3, OR=9.7). This VNTR is relatively close to (distance of 1.4Kb) the leading SNP in the region, rs7225151 (p = 9.3e-13), which has been found to be genome-wide significant in a GWAS [51].

The VNTR in *TRANK1* has an 18bp motif, with 29 cases and 11 controls above the outlier boundary (FDR corrected p=0.09, OR=3.3). A decreased expression of the *TRANK1* gene influences many genes related to neural development and differentiation [33]. Thus far *TRANK1* has Gene Ontology (GO) annotations to protein binding and ATP binding [1, 3] and has been linked to bipolar disorder and schizophrenia [33]. Schizophrenia, in turn, is associated with an elevated risk for developing AD [39]. Even more so, *TRANK1* has been mentioned as a novel genomic region of interest for rare variants of AD [52]. In our dataset this genetic variation is above the common threshold of 1%, instead of rare.

The VNTR in *GRM8* has an 20bp motif, with 166 cases and 143 controls above the outlier boundary (FDR corrected p=0.09, OR=1.6). *GRM8* has been linked to autism [24, 43] and multiple sclerosis [71]. Recently, an upregulation of *GRM8* has been shown to protect against neural inflammation [71]. Neuroinflammation contributes greatly to the pathogenesis of AD [25]. Furthermore, Aβ plaques, another hallmark of AD, have been associated with the downregulation of *Grm8* and a drug has been reported to prevent this Aβ-associated downregulation of *Grm8* in AD mouse models [55].

The protein encoded by the *DIP2C* gene (reported VNTR: 12 bp motif, FDR corrected p=0.06, OR=4.3) has an influence on transcription factor binding and is expressed in the nervous system [1, 3]. A SNP in *DIP2C* has been linked to overall cognitive ability and to psychosis in AD [26]. SNPs in *ARHGEF10* (VNTR: 20 bp motif, uncorrected p=1.9e-2, OR=8.5) has been associated with neurofibrillary tangles[62], and came up in a study showing genetic overlap between Amyotrophic lateral sclerosis (ALS) and fronto-temporal dementia (FTD) [36]. There is evidence to link ALS and FTD to AD [70]. Finally, *PID1* (VNTR: 17 bp motif, uncorrected p=1.9e-2, OR=8.5) mRNA is lower in brains of AD patients [34, 17, 19], *EVC2* (VNTR: 20 bp motif, uncorrected p=8.6e-3, OR=inf) has been reported as a rare variant for early-onset familial AD [27].

Three of the thirteen expanded VNTRs occur in non-coding RNA (ncRNA) genes (*LINC02050*, *LOC101928269*, *LOC107984737*). There is substantial evidence implicating ncRNA in the regulation of the main hallmarks of AD pathology, such as Aβ, tau and inflammation [42, 48, 64]. Although these specific ncRNA expansions/variations have not been linked to AD yet, further investigation could be valuable.

4.2 Contracted VNTRs

We detected approximately 15x more contractions than expansions. This could be due to a technical effect, for example that one VNTR is divided over multiple variant catalog entries. This could lead to some entries showing an expansion and some a contraction, which could cancel each other out upon merging these entries. Alternatively, there is a biological implication that the number of contractions increases exponentially as the repeat length increases [61]. Interestingly, contractions can occur, just like expansions, due to repair slippage [59].

One of the detected contracted VNTRs falls within the *DNM2* gene, where 14 cases and 1 control fall below the outlier boundary (16 bp motif, FDR corrected p = 9.7e-4, OR = 17.3). We carefully checked the variant catalog for neighbouring entries with a similar motif, which were not present, to avoid reporting a technical effect. Mutations in *DNM2* cause Charcot-Marie-Tooth disease and a rare form of myopathy [65, 16, 22]. In the context of AD, *DNM2* has been described as a susceptibility gene for late-onset AD (LOAD) in non-carriers of the APOE4 allele [2] and at least one SNP has been associated with LOAD [35]. Aβ seems to decrease the expression of DNM2 [37]. The products of *DNM2* have an effect on myelination and microtubules [16, 22, 29, 65]. Both demyelination and unstable microtubules are characteristics of AD [32, 53]. Although this gene came up as significant based on its contraction, in Figure 6a we also noticed a few cases showing an expansion. These are indeed all LOAD cases, which is in line with *DNM2* being described as a susceptibility gene for LOAD.

Until now only one confirmed pathogenic contraction has been reported, related to facioscapulohumeral muscular dystrophy; this contraction occurs in a subtelomeric region resulting in relative demethylation, which is thought to activate an (unknown) gene with negative effects on muscular development that might otherwise be silent [10, 68, 69]. In the context of the *cinetobacter baumannii* bacterium, contractions have been described to have an impact on the expression, structure and activity of cellular proteins [58]. Furthermore, an association was found between individuals who have fewer repeats in the *FMR1* CGG repeat and various deleterious effects, including decreased cognitive functioning [46]. This supports the idea that there is a tight range of repeats which allow for optimal cognitive functioning, and that a low number of repeats as well as a high number of repeats could have pathological effects. Therefore, we speculate that our detected contractions could have a similar deregulatory effect as the expansions.

4.3 Limitations

VNTRs can be quite complex: the motif is often variable, there can be insertions between copies and even repeats within repeats. In this study, we only detect the 'clean' repeats, i.e. repeats that have adjacent copies of the same motif. This limitation is mainly due to the way EHdn and EH work, as they barely allow for differences in motif, and even less so for indels between copies. EH does allow for the encoding of more complex repeats, however the structures need to be precisely captured with a regular expression and provided in the variant catalog. The challenging part is determining the structure of a VNTR, as they are often unstable, leading to many possible variations. Furthermore, in the output of EHdn we noticed repeats with the same or a very similar pattern (differing by 1 or 2 bp) in nearby or even overlapping regions. These are probably part of the same repeat, yet are outputted as separate repeats. In this study, we treat these as separate VNTRs, even though merging similar nearby repeats might be more biologically appropriate.

Both EHdn and EH detect motifs up until 20bp, imposing a limitation on the size of VNTRs that can be detected. For example, we were not be able to detect the *ABCA7* VNTR with a motif of 25bp [9]. This limitation of detectable motif size can be attributed to the constraints imposed by

read length. For example, with a motif of 20bp and a read length of 100, a maximum of 5 repeats could fit within a read. In other words, the longer a motif, the less confident any estimate can be, simply because only a few copies can fit within a short read.

In our current pipeline, we exclude the VNTRs that are absent from the reference genome. These VNTRs have the potential to be pathogenic too. However, we need a basis in the reference in order to use EH, meaning that these VNTRs (detected by EHdn) would require a whole different pipeline to estimate their repeat sizes.

When filtering candidate VNTRs we aim to select the most promising candidates for the rest of our analysis. We can not be certain that this did not exclude candidate VNTRs which might turn out to have an association with AD

Finally, the sensitivity to the outlier constant is an indication that this method of VNTR detection is suboptimal. Furthermore it shows there are many more VNTRs to be discovered using this pipeline and even this dataset.

5 Future work

First and foremost, the detected VNTRs should be validated using either long-read sequencing or in vitro methods. Furthermore, they should be validated using a larger cohort.

One of the major biases in the estimates of repeat size is due to VNTRs being spread out over multiple entries. The current pipeline and utilised tools (EHdn, EH) allow for little variation within and between repeats. One way to reduce these effects would be to add a merging step in the downstream analysis, merging nearby entries with similar repeated motifs as they are likely part of the same VNTR.

During downstream analysis, and specifically in the outlier analysis, we noticed that the choice of the outlier boundary constant determined which VNTRs were detected, leaving many VNTRs within the current dataset, undetected. This indicates that detecting outliers with a boundary derived from the interquartile range is suboptimal and the pipeline would benefit from a more flexible nonparametric method for outlier detection.

The current pipeline is not suitable for detecting VNTRs without a basis in the reference genome due to the need for a variant catalog specifying coordinates on the reference. To analyse these VNTRs, a completely different method would be needed to estimate the number of repeats. There have been pathogenic VNTRs reported without a basis in the reference genome [11], making this direction one worth investigating.

Extending the downstream analysis by linking phenotypes to genotypes could be a valuable addition. For example, a regression analysis with e.g. the number of repeats within a VNTR and age at onset of AD could differentiate between variants specific to either early-onset AD or late-onset AD. As diseases are often caused by combinations of variants [60], another idea would be to encode the set of VNTRs as a binary vector for each sample, annotating whether that VNTR is expanded within that specific sample. These vectors could be clustered to find subgroups of VNTRs that co-occur, defining subgroups of AD patients.

6 Conclusion

We set out to tackle the challenging task of detecting VNTRs using paired-end short read data. We created a pipeline that performs a genome-wide search (i.e. no predefined set of repeats needed) for VNTRs exceeding read length. We utilised this pipeline to identify various promising VNTRs, proving the potential of this pipeline and illustrating the abundance of VNTRs associated with AD.

7 Acknowledgements

This work was carried out at The Delft Bioinformatics Lab, TU Delft, where Marcel J.T. Reinders actively contributed. Furthermore, this project was in close collaboration with the 100-plus Study at the Alzheimer Center Amsterdam, where Niccolò Tesi, Marc Hulsman and Henne Holstege contributed greatly. Finally, Egor Dolzhenko from Illumina assisted to this project.

References

[1]The gene ontology resource: enriching a gold mine. *Nucleic Acids Research*, 49(D1):D325–D334, 2021.

[2]Nuripa Jenishbekovna Aidaralieva, Kouzin Kamino, Ryo Kimura, Mitsuko Yamamoto, Takeshi Morihara, Hiroaki Kazui, Ryota Hashimoto, Toshihisa Tanaka, Takashi Kudo, Tomoyuki Kida, et al. Dynamin 2 gene is a novel susceptibility gene for late-onset alzheimer disease in non-apoe-ε4 carriers. *Journal of human genetics*, 53(4):296–302, 2008.

[3]Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[4]J Wesson Ashford and James A Mortimer. Non-familial alzheimer's disease is mainly due to genetic factors. *Journal of Alzheimer's disease*, 4(3):169–177, 2002.

[5]Celine Bellenguez, Fahri Küçükali, Iris Jansen, Victor Andrade, Sonia Moreno-Grau, Najaf Amin, Adam C Naj, Benjamin Grenier-Boley, Rafael Campos-Martin, Peter A Holmans, et al. New insights on the genetic etiology of alzheimer's and related dementia. *MedRxiv*, 2020.

[6]Chad N Brocker, Vasilis Vasiliou, and Daniel W Nebert. Evolutionary divergence and functions of the adam and adamts gene families. *Human genomics*, 4(1):1–13, 2009.

[7]Isabel Castanho, Tracey K Murray, Eilis Hannon, Aaron Jeffries, Emma Walker, Emma Laing, Hedley Baulf, Joshua Harvey, Lauren Bradshaw, Andrew Randall, et al. Transcriptional signatures of tau and amyloid neuropathology. *Cell reports*, 30(6):2040–2054, 2020.

[8]EH Corder, Al M Saunders, NJ Risch, WJ Strittmatter, DE Schmechel, PC Gaskell, JB Rimmler, PA Locke, PM Conneally, KE Schmader, et al. Protective effect of apolipoprotein e type 2 allele for late onset alzheimer disease. *Nature genetics*, 7(2):180–184, 1994.

[9]Arne De Roeck, Lena Duchateau, Jasper Van Dongen, Rita Cacace, Maria Bjerke, Tobi Van den Bossche, Patrick Cras, Rik Vandenberghe, Peter P De Deyn, Sebastiaan Engelborghs, et al. An intronic vntr affects splicing of abca7 and increases risk of alzheimer's disease. *Acta neuropathologica*, 135(6):827–837, 2018.

[10]Christel Depienne and Jean-Louis Mandel. 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? *The American Journal of Human Genetics*, 2021.

[11]Egor Dolzhenko, Mark F Bennett, Phillip A Richmond, Brett Trost, Sai Chen, Joke JFA van Vugt, Charlotte Nguyen, Giuseppe Narzisi, Vladimir G Gainullin, Andrew M Gross, et al. Expansionhunter denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. *Genome biology*, 21(1):1–14, 2020.

[12]Egor Dolzhenko, Viraj Deshpande, Felix Schlesinger, Peter Krusche, Roman Petrovski, Sai Chen, Dorothea Emig-Agius, Andrew Gross, Giuseppe Narzisi, Brett Bowman, et al. Expansionhunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics*, 35(22):4754–4756, 2019.

[13]Egor Dolzhenko, Joke JFA van Vugt, Richard J Shaw, Mitchell A Bekritsky, Marka van Blitterswijk, Giuseppe Narzisi, Subramanian S

Ajay, Vani Rajan, Bryan R Lajoie, Nathan H Johnson, et al. Detection of long repeat expansions from pcr-free whole-genome sequence data. *Genome research*, 27(11):1895–1903, 2017.

[14]David J Dow, Julie Huxley-Jones, Jamie M Hall, Clyde Francks, Peter R Maycox, James NC Kew, Israel S Gloger, Nalini AL Mehta, Fiona M Kelly, Pierandrea Muglia, et al. Adamtsl3 as a candidate gene for schizophrenia: gene sequencing and ultra-high density association analysis by imputation. *Schizophrenia research*, 127(1-3):28–34, 2011.

[15]Cletus A D'Souza, Vikramjit Chopra, Richard Varhol, Yuan-Yun Xie, Slavita Bohacec, Yongjun Zhao, Lisa LC Lee, Mikhail Bilenky, Elodie Portales-Casamar, An He, et al. Identification of a set of genes showing regionally enriched expression in the mouse brain. *BMC neuroscience*, 9(1):1–14, 2008.

[16]Anne-Cécile Durieux, Bernard Prudhon, Pascale Guicheney, and Marc Bitoun. Dynamin 2 and human diseases. *Journal of molecular medicine*, 88(4):339–350, 2010.

[17]Anat Erdreich-Epstein, Nathan Robison, Xiuhai Ren, Hong Zhou, Jingying Xu, Tom B Davidson, Mathew Schur, Floyd H Gilles, Lingyun Ji, Jemily Malvar, et al. Pid1 (nyggf4), a new growth-inhibitory gene in embryonal brain tumors and gliomas. *Clinical Cancer Research*, 20(4):827–836, 2014.

[18]Sarah Fazal, Matt C Danzi, Vivian P Cintra, Dana M Bis-Brewer, Egor Dolzhenko, Michael A Eberle, and Stephan Zuchner. Large scale in silico characterization of repeat expansion variation in human genomes. *Scientific data*, 7(1):1–14, 2020.

[19]Alexander W Fischer, Kirstin Albers, Christian Schlein, Frederike Sass, Lucia M Krott, Hartwig Schmale, Philip LSM Gordts, Ludger Scheja, and Joerg Heeren. Pid1 regulates insulin-dependent glucose uptake by controlling intracellular sorting of glut4-storage vesicles. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1865(6):1592–1603, 2019.

[20]Margaret Gatz, Chandra A Reynolds, Laura Fratiglioni, Boo Johansson, James A Mortimer, Stig Berg, Amy Fiske, and Nancy L Pedersen. Role of genes and environments for explaining alzheimer disease. *Archives of general psychiatry*, 63(2):168–174, 2006.

[21]Tania F Gendron and Leonard Petrucelli. Disease mechanisms of c9orf72 repeat expansions. *Cold Spring Harbor perspectives in medicine*, 8(4):a024224, 2018.

[22]Arlek M González-Jamett, Valentina Haro-Acuña, Fanny Momboisse, Pablo Caviedes, Jorge A Bevilacqua, and Ana M Cárdenas. Dynamin-2 in nervous system disorders. *Journal of neurochemistry*, 128(2):210–223, 2014.

[23]Ma"elenn Guerchet, Martin Prince, et al. Numbers of people with dementia worldwide: An update to the estimates in the world alzheimer report 2015. 2020.

[24]Dexter Hadley, Zhi-liang Wu, Charlly Kao, Akshata Kini, Alisha Mohamed-Hadley, Kelly Thomas, Lyam Vazquez, Haijun Qiu, Frank Mentch, Renata Pellegrino, et al. The impact of the metabotropic glutamate receptor and other gene family interaction networks on autism. *Nature communications*, 5(1):1–10, 2014.

[25]Michael T Heneka, Monica J Carson, Joseph El Khoury, Gary E Landreth, Frederic Brosseron, Douglas L Feinstein, Andreas H Jacobs, Tony Wyss-Coray, Javier Vitorica, Richard M Ransohoff, et al. Neuroinflammation in alzheimer's disease. *The Lancet Neurology*, 14(4):388–405, 2015.

[26]Paul Hollingworth, Robert Sweet, Rebecca Sims, Denise Harold, Giancarlo Russo, Richard Abraham, Alexandra Stretton, Nicola Jones, Amy Gerrish, Jade Chapman, et al. Genome-wide association study of alzheimer's disease with psychotic symptoms. *Molecular psychiatry*, 17(12):1316–1327, 2012.

[27]BV Hooli, Zs M Kovacs-Vajna, K Mullin, MA Blumenthal, Manuel Mattheisen, C Zhang, C Lange, G Mohapatra, Lars Bertram, and RE Tanzi. Rare autosomal copy number variations in early-onset familial alzheimer's disease. *Molecular psychiatry*, 19(6):676–681, 2014.

[28]Alfredo Iacoangeli, Ahmad Al Khleifat, Ashley R Jones, William Sproviero, Aleksey Shatunov, Sarah Opie-Martin, Karen E Morrison, Pamela J Shaw, Christopher E Shaw, Isabella Fogh, et al. C9orf72 intermediate expansions of 24–30 repeats are associated with als. *Acta neuropathologica communications*, 7(1):1–7, 2019.

[29]Nobuhisa Ishida, Yuichi Nakamura, Kenji Tanabe, Shun-Ai Li, and Kohji Takei. Dynamin 2 associates with microtubules at mitosis and regulates cell cycle progression. *Cell structure and function*, pp. 1011260065–1011260065, 2010.

[30]Yanina Ivashko-Pachima, Adva Hadar, Iris Grigg, Vlasta Korenková, Oxana Kapitansky, Gidon Karmon, Michael Gershovits, C Laura Sayas, R Frank Kooy, Johannes Attems, et al. Discovery of autism/intellectual disability somatic mutations in alzheimer's brains: mutated adnp cytoskeletal impairments and repair as a case study. *Molecular psychiatry*, 26(5):1619–1633, 2021.

[31]Iris E Jansen, Jeanne E Savage, Kyoko Watanabe, Julien Bryois, Dylan M Williams, Stacy Steinberg, Julia Sealock, Ida K Karlsson, Sara Hägg, Lavinia Athanasiu, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimer's disease risk. *Nature genetics*, 51(3):404–413, 2019.

[32]Daphney C Jean and Peter W Baas. It cuts two ways: microtubule loss during alzheimer disease. *The EMBO journal*, 32(22):2900–2902, 2013.

[33]Xueying Jiang, Sevilla D Detera-Wadleigh, Nirmala Akula, Barbara S Mallon, Liping Hou, Tiaojiang Xiao, Gary Felsenfeld, Xinglong Gu, and Francis J McMahon. Sodium valproate rescues expression of trank1 in ipsc-derived neural cells that carry a genetic variant associated with serious mental illness. *Molecular psychiatry*, 24(4):613–624, 2019.

[34]Yuji Kajiwara, Sonia Franciosi, Nagahide Takahashi, Lisa Krug, James Schmeidler, Kevin Taddei, Vahram Haroutunian, Ulrik Fried, Michelle Ehrlich, Ralph N Martins, et al. Extensive proteomic screening identifies the obesity-related nyggf4 protein as a novel lrp1-interactor, showing reduced expression in early alzheimer's disease. *Molecular neurodegeneration*, 5(1):1–11, 2010.

[35]Eiichiro Kamagata, Takashi Kudo, Ryo Kimura, Hitoshi Tanimukai, Takashi Morihara, Md Golam Sadik, Kouzin Kamino, and Masatoshi Takeda. Decrease of dynamin 2 levels in late-onset alzheimer's disease alters aβ metabolism. *Biochemical and biophysical research communications*, 379(3):691–695, 2009.

[36]Celeste M Karch, Natalie Wen, Chun C Fan, Jennifer S Yokoyama, Naomi Kouri, Owen A Ross, Gunter Höglinger, Ulrich Müller, Raffaele Ferrari, John Hardy, et al. Selective genetic overlap between amyotrophic lateral sclerosis and diseases of the frontotemporal dementia spectrum. *JAMA neurology*, 75(7):860–875, 2018.

[37]Brent L Kelly, Robert Vassar, and Adriana Ferreira. β-amyloid-induced dynamin 1 depletion in hippocampal neurons: a potential mechanism for early cognitive decline in alzheimer disease. *Journal of Biological Chemistry*, 280(36):31746–31753, 2005.

[38]Soong Ho Kim, Paul E Fraser, David Westaway, Peter H St George-Hyslop, Michelle E Ehrlich, and Sam Gandy. Group ii metabotropic glutamate receptor stimulation triggers production and release of alzheimer's amyloid β42 from isolated intact nerve terminals. *Journal of Neuroscience*, 30(11):3870–3875, 2010.

[39]Peter Kochunov, Artemis Zavaliangos-Petropulu, Neda Jahanshad, Paul M Thompson, Meghann C Ryan, Joshua Chiappelli, Shuo Chen, Xiaoming Du, Kathryn Hatch, Bhim Adhikari, et al. A white matter

connection of schizophrenia and alzheimer's disease. *Schizophrenia bulletin*, 47(1):197–206, 2021.

[40]Bon-Hun Koo, Tiina Hurskainen, Katrina Mielke, Phyu Phyu Aung, Graham Casey, Helena Autio-Harmainen, and Suneel S Apte. Adamtsl3/punctin-2, a gene frequently mutated in colorectal tumors, is widely expressed in normal and malignant epithelial cells, vascular endothelial cells and other cell types, and its mrna is reduced in colon cancer. *International journal of cancer*, 121(8):1710–1716, 2007.

[41]Hyoung-gon Lee, Xiongwei Zhu, Gemma Casadesus, Mercé Pallàs, Antoni Camins, Michael J O'Neill, Shigetada Nakanishi, George Perry, and Mark A Smith. The effect of mglur2 activation on signal transduction pathways and neuronal cell survival. *Brain research*, 1249:244–250, 2009.

[42]Dingfeng Li, Juan Zhang, Xiaohui Li, Yuhua Chen, Feng Yu, and Qiang Liu. Insights into lncrnas in alzheimer's disease mechanisms. *RNA biology*, pp. 1–11, 2020.

[43]Hui Li, Yun Li, Jie Shao, Rong Li, Yufeng Qin, Chunhong Xie, and Zhengyan Zhao. The association analysis of reln and grm8 genes with autistic spectrum disorder in chinese han population. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 147(2):194–200, 2008.

[44]Jasper Linthorst, Wim Meert, Matthew S Hestand, Jonas Korlach, Joris Robert Vermeesch, Marcel JT Reinders, and Henne Holstege. 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, 2020.

[45]Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, et al. The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic acids research*, 45(D1):D896–D901, 2017.

[46]Marsha R Mailick, Jinkuk Hong, Paul Rathouz, Mei W Baker, Jan S Greenberg, Leann Smith, and Matthew Maenner. Low-normal fmr1 cgg repeat length: phenotypic associations. *Frontiers in genetics*, 5:309, 2014.

[47]A Malishkevich, N Amram, G Hacohen-Kleiman, I Magen, E Giladi, and I Gozes. Activity-dependent neuroprotective protein (adnp) exhibits striking sexual dichotomy impacting on autistic and alzheimer's pathologies. *Translational psychiatry*, 5(2):e501–e501, 2015.

[48]Rotem Maoz, Benjamin P Garfinkel, and Hermona Soreq. Alzheimer's disease and ncrnas. *Neuroepigenomics in aging and disease*, pp. 337–361, 2017.

[49]Avinash Marwal, Anurag Kumar Sahu, and R.K. Gaur. Chapter 16 - molecular markers: Tool for genetic analysis. In Ashish S. Verma and Anchal Singh, editors, *Animal Biotechnology*, pp. 289–305. Academic Press, San Diego, 2014.

[50]Elisabeth Möncke-Buchner, Stefanie Reich, Merlind Mücke, Monika Reuter, Walter Messer, Erich E Wanker, and Detlev H Krüger. Counting cag repeats in the huntington's disease gene by restriction endonuclease eco p15i cleavage. *Nucleic acids research*, 30(16):e83–e83, 2002.

[51]Sonia Moreno-Grau, Itziar de Rojas, Isabel Hernández, Inés Quintela, Laura Montrreal, Montserrat Alegret, Begoña Hernández-Olasagarre, Laura Madrid, Antonio González-Perez, Olalla Maroñas, et al. Genome-wide association analysis of dementia and its clinical endophenotypes reveal novel loci associated with alzheimer's disease and three causality networks: the gr@ ace project. *Alzheimer's & Dementia*, 15(10):1333–1347, 2019.

[52]Adam C Naj, Ganna Leonenko, Xueqiu Jian, Benjamin Grenier-Boley, Maria Carolina Dalmasso, Céline Bellenguez, Jin Sha, Yi Zhao, Sven J van der Lee, Rebecca Sims, et al. Genome-wide meta-analysis of late-onset alzheimer's disease using rare variant imputation in 65,602 subjects identifies novel rare variant locus nck2: The international genomics of alzheimer's project (igap). *medRxiv*, 2021.

[53]Ewa Papuć and Konrad Rejdak. The role of myelin damage in alzheimer's disease pathology. *Archives of medical science: AMS*, 16(2):345, 2020.

[54]Jisu Park, Hyunwoo Choi, Young Doo Kim, Seo-Hyun Kim, Youbin Kim, Youngdae Gwon, Dong Young Lee, Sung-Hye Park, Won Do Heo, and Yong-Keun Jung. Aberrant role of alk in tau proteinopathy through autophagosomal dysregulation. *Molecular Psychiatry*, pp. 1–15, 2021.

[55]Eduardo Pauls, Sergi Bayod, Lídia Mateo, Victor Alcalde, Teresa Juan-Blanco, Takaomi C Saido, Takashi Saito, Antoni Berenguer-Llergo, Camille Stephan-Otto Attolini, Marina Gay, et al. Identification and drug-induced reversion of molecular signatures of alzheimers disease onset and progression in appnl-gf, appnl-f and 3xtg-ad mouse models. *bioRxiv*, 2021.

[56]Henry Paulson. Chapter 9 - repeat expansion diseases. In Daniel H. Geschwind, Henry L. Paulson, and Christine Klein, editors, *Neurogenetics, Part I*, volume 147 of *Handbook of Clinical Neurology*, pp. 105–123. Elsevier, 2018.

[57]Brent S Pedersen and Aaron R Quinlan. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*, 34(5):867–868, 2018.

[58]Christine Pourcel, Fabrizia Minandri, Yolande Hauck, Silvia d'Arezzo, Francesco Imperi, Gilles Vergnaud, and Paolo Visca. Identification of variable-number tandem-repeat (vntr) sequences in acinetobacter baumannii and interlaboratory validation of an optimized multiple-locus vntr analysis typing scheme. *Journal of clinical microbiology*, 49(2):539–548, 2011.

[59]Calen P Ryan. Tandem repeat disorders. *Evolution, medicine, and public health*, 2019.

[60]Nicholas J Schork, Sarah S Murray, Kelly A Frazer, and Eric J Topol. Common vs. rare allele hypotheses for complex diseases. *Current opinion in genetics & development*, 19(3):212–219, 2009.

[61]Makoto K Shimada, Ryoko Sanbonmatsu, Yumi Yamaguchi-Kabata, Chisato Yamasaki, Yoshiyuki Suzuki, Ranajit Chakraborty, Takashi Gojobori, and Tadashi Imanishi. Selection pressure on human str loci and its relevance in repeat expansion disease. *Molecular Genetics and Genomics*, 291(5):1851–1869, 2016.

[62]Gil Stelzer, Naomi Rosen, Inbar Plaschkes, Shahar Zimmerman, Michal Twik, Simon Fishilevich, Tsippi Iny Stein, Ron Nudel, Iris Lieder, Yaron Mazor, et al. The genecards suite: from gene data mining to disease genome sequence analyses. *Current protocols in bioinformatics*, 54(1):1–30, 2016.

[63]Chao Tai, Che-Wei Chang, Gui-Qiu Yu, Isabel Lopez, Xinxing Yu, Xin Wang, Weikun Guo, and Lennart Mucke. Tau reduction prevents key features of autism in mouse models. *Neuron*, 106(3):421–437, 2020.

[64]Lin Tan, Jin-Tai Yu, Nan Hu, and Lan Tan. Non-coding rnas in alzheimer's disease. *Molecular neurobiology*, 47(1):382–393, 2013.

[65]Kenji Tanabe and Kohji Takei. Dynamic instability of microtubules requires dynamin 2 and is impaired in a charcot-marie-tooth mutant. *Journal of Cell Biology*, 185(6):939–948, 2009.

[66]Niccolo Tesi, Sven J van der Lee, Marc Hulsman, Henne Holstege, and Marcel Reinders. snpxplorer: a web application to explore snp-associations and annotate snp-sets. *bioRxiv*, 2020.

[67]Brett Trost, Worrawat Engchuan, Charlotte M Nguyen, Bhooma Thiruvahindrapuram, Egor Dolzhenko, Ian Backstrom, Mila Mirceta, Bahareh A Mojarad, Yue Yin, Alona Dov, et al. Genome-wide detection of tandem dna repeats that are expanded in autism. *Nature*, 586(7827):80–86, 2020.

[68]Karen Usdin. The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. *Genome research*, 18(7):1011–1019, 2008.

[69]Petra GM van Overveld, Richard JFL Lemmers, Lodewijk A Sandkuijl, Leo Enthoven, Sara T Winokur, Floor Bakels, George W Padberg, Gert-Jan B van Ommen, Rune R Frants, and Silvère M van der Maarel. Hypomethylation of d4z4 in 4q-linked and non-4q-linked facioscapulohumeral muscular dystrophy. *Nature genetics*, 35(4):315–317, 2003.

[70]Agathe Vrillon, Vincent Deramecourt, Florence Pasquier, David Wallon, Pierre Lozeron, Elodie Bouaziz Amar, and Claire Paquet. Association of alzheimer's disease and amyotrophic lateral sclerosis: A series of cases and review of the literature: Neuropsychiatry and behavioral neurology/dementia. *Alzheimer's & Dementia*, 16:e045814, 2020.

[71]Marcel S Woo, Friederike Ufer, Nicola Rothammer, Giovanni Di Liberto, Lars Binkle, Undine Haferkamp, Jana K Sonner, Jan Broder Engler, Sönke Hornig, Simone Bauer, et al. Neuronal metabotropic glutamate receptor 8 protects against neurodegeneration in cns inflammation. *Journal of Experimental Medicine*, 218(5), 2021.

[72]Nathaniel S Woodling, Benjamin Aleyakpo, Miranda Claire Dyson, Lucy J Minkley, Arjunan Rajasingam, Adam J Dobson, Kristie HC Leung, Simona Pomposova, Matías Fuentealba, Nazif Alic, et al. The neuronal receptor tyrosine kinase alk is a target for longevity. *Aging Cell*, 19(5):e13137, 2020.

Appendices

A Variant catalog creation algorithm

We create a variant catalog to run ExpansionHunter (EH) based on the output from ExpansionHunter Denovo (EHdn). EHdn reports for each repeat a position: chr:start-end and a motif. These coordinates can be off by $\pm$500bp, so we use 500 as our margin. For each repeat:

1. Extract sequence from the reference genome from (start - margin, end + margin)
2. For the repeated motif:
   a. Identify positions with stretches of perfect copies of this motif
   b. Extend these stretches when:
      - The next motif differs by 1 base (Levenshtein distance ≤ 1)
      - AND the motif after that is a perfect copy of the repeating motif
3. For the reverse complement of the motif (revmotif), same loop
4. Select the motif or the revmotif, depending on which has the longest stretch of repeats
   a. If they have the same size longest stretch, retain both
5. For the selected motif(s), select the position(s) that have the longest stretch of repeat
   a. If there are multiple longest stretches, retain all of these
6. For each of the retained longest stretches
   a. Calculate the chromosome coordinates
   b. Create a variant catalog entry using the updated coordinates and corresponding motif

EH performs some quality control. One of these quality measures is the number of 'N's in the sequence of the reference genome. An 'N' means that this nucleotide could not be determined precisely. EH extracts repeats from a region around the repeat, by default this is a region of 1Kb on each side of the repeat. If there are more than 5 Ns in this region, EH gives an error. Therefore we exclude repeats from the variant catalog where the region in the reference sequence has more than 5Ns.

Due to the nature of this catalog creation and because EH requires coordinates on the reference genome, repeats that have no basis in the reference get filtered out.

The motif is expressed as a regular expression. In our cases, this would be in the form of (CAG)*, meaning that the tandem repeat consists of zero or more repeats of the CAG motif. We only encode simple repeats, in the sense that we only encode 1 motif that is repeated. EH's sequence graph algorithm allows for the encoding of more complex repeats. For example, (CAG)*CAACAG(CCG)* would encode two tandem repeats with motifs CAG and CCG, separated by a CAACAG sequence.

B ExpansionHunter Denovo

First, it scans each individual genome for tandem repeats that exceed read length. It does so by detecting in-repeat reads (IRRs) and anchored in-repeat reads. The IRRs consist purely of the repeating sequence, the anchored IRRs have one side mapped to the flanking sequence (the anchored part). EHdn considers a minimum mapping quality (MAPQ) for the anchored IRRs and a maximum MAPQ for IRRs as input variables. Default values were used, which are: minimum MAPQ for anchored reads = 40, max MAPQ for IRRs = 50. For each repeating structure, the number of anchored and IRRs and their locations are noted in the STR profile of that genome.

After identifying these IRRs for each individual genome, the normalised IRR counts can be compared between samples. There are two modes of comparison: case control analysis and outlier analysis. The case control analysis checks if the IRR count distributions differ, so this analysis mode is suitable to detect repeats that are expanded in a significant proportion of the cases. This calculates the wilcoxon rank sum score based on the distribution of read counts for cases vs controls. These scores are then reported, along with the copied motif, its approximate region on the reference and the read counts for each sample. Although it is not explicitly stated, there seems to be a filtering on the prevalence amongst cases for each repeat. (This is based on that it reports fewer repeats than the outlier analysis, but does report p-values ranging from 0 - 1.)

The outlier analysis checks for large deviations in the IRR count. Just a single case with a large expansion is sufficient to be reported, so this analysis mode is suitable for detecting rare expansions. It takes 95% of the distribution of read counts (based on controls only? or cases and controls?). Each repeat for which at least one case deviates from the mean by at least 1 standard deviation is reported.

C ExpansionHunter

Duplicated from Literature Survey. The authors of ExpansionHunter distinguish three types of reads that you can encounter: spanning reads (encompassing the whole repeating region), flanking reads (one side of the read is outside the repeating region), and in-repeat reads (IRRs) [**?**].

When estimating an expansion of repeats that are longer than the read length, the reads will consist fully of the repeating motif. In this case we want to identify IRRs to estimate the size of the expansion. To identify IRRs one must first establish if the read fully consists of a repeating motif. To do so, it will be matched to the perfect repeating sequence. To find the closest match, the sequence can be shifted (repeating CAG can also lead to repeats of AGC or GCA) and reversed (the complementary GTC, TCG or CGT). When there is a match, some sort of alignment score is calculated to assess if it is an IRR. This score is called 'Weighted Purity' (WP), which assigns:

1 for a matching basepair,
0.5 for a low-quality mismatch,
-1 for a high- quality mismatch.

Low- and high-quality refer to the read quality. If the read is less certain, a mismatch is less meaningful. These scores are summed for the bases, and then normalized by dividing it by read length. This way, the WP score will range from -1 to 1. Reads with a WP ≥ 0.9 are considered IRRs.

These IRRs can map to different locations in the genome, since certain repeats occur more often. Especially when an STR is relatively short on the reference genome and expanded in the donor genome, the IRR might not map back to the target location. The off-target regions are therefore defined as regions where IRRs might be mistakenly mapped to. It is useful to identify these regions as only those regions have to be taken into account when searching for relevant IRRs, instead of searching the whole genome.

IRR reads are not only selected by having a high WP score (WP ≥ 0.9), but their mapping quality score (MAPQ $= 0$) is taken into account as well. A low value implies it can map to multiple regions, so due to this ambiguity the mapping quality is considered low. In our case, this ambiguous mapping means a higher probability for an IRR, as these reads have a high similarity to the repeating region and can be mapped to multiple regions. This method works best for motifs that are sufficiently long and underrepresented in the rest of the genome, ensuring there will be only a few alternative locations where the reads can be mapped to [**?**].

When IRRs are closer than 500bp to each other, their mapping positions will be merged. The resulting locations, if present in $\geq 50\%$ of the

samples, are considered the off-target regions where IRRs can be mapped to.

We assume that observing a read at a certain base or position follows a Bernoulli distribution, with success probability π.

$$\pi = \frac{read\ depth}{read\ length} \tag{2}$$

Given this assumption, the reads starting at each position in a certain region form a Bernoulli process. Therefore the number of reads starting in this region follows a binomial distribution.

With r being the read length, one of the terminal bases of an IRR has to start at least $N - r$ bases away from flanking regions of the repeat. This leads to the following distribution for observing i of such reads:

$$P(i,\ N-r)\ =\ \binom{N-r}{i}\ \pi^i\ \ (1-\pi)^{N-r-i} \tag{3}$$

with r being the read length and N the repeat size. Since r is known, i can be estimated with the number of IRRs (found in the previous step) and π can be calculated according to Equation 2, then the value of N (repeat size) can be estimated. This estimation is done with a parametric bootstrap.

To identify the spanning reads, first a subselection is made including only those reads that are aligned within 1kb of the target region. These reads are then checked for the presence of the repeating motif. If this is present, the flanking regions from the reads are aligned to the flanking regions on the reference genome. This read also gets a WP score, and should be ≥ 0.9 for both repeating and flanking parts of the sequence to be considered spanning. The more similar a flanking region is to the repeating motif, the more flanking sequence is required to identify the end of repeating and begin of flanking region. This is formalized in requirements on mismatches in flanking regions.