Utilising SNP-SV Correlations to find SVs Associated with Alzheimer's Disease

A Novel Approach to Identifying and Analysing Alzheimer's-Related Structural Variants

Joris Belder[1]
Supervisors: Marcel Reinders[1]**, Niccoló Tesi**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Joris Belder
Final project course: CSE3000 Research Project
Thesis committee: Marcel Reinders, Niccoló Tesi, Andy Zaidman

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

Nucleic Acids Research, 2024, 1–10

Utilising SNP-SV Correlations to find SVs Associated with Alzheimer's Disease

Joris Belder,[1,*] Niccoló Tesi[2] and Marcel Reinders[2]

[1]BSc Computer Science and Engineering, EEMCS, Delft University of Technology, Netherlands and [2]Pattern Recognition and Bioinformatics, EEMCS, Delft University of Technology, Netherlands

*Corresponding author

Abstract

Alzheimer's disease (AD) is a neurodegenerative disease affecting r oughly 4 0 m illion p eople. 7 0% o f the heritability of AD is expected to be explained by Structural Variants (SVs), however these have been scarcely studied in the context of AD. This study aims to find S Vs a ssociated w ith A D a nd t o i nvestigate the properties of these correlated SVs. To this end, a dataset created using Third Generation Sequencing of Single Nucleotide Polymorphisms (SNPs) and their correlation with SVs was utilised and combined with the full summary statistics and fine-mapped r esults o f a l arge A D G enome W ide A ssociation S tudy. This resulted in 85 unique SVs with significant c orrelations t o k nown A D a ssociated S NPs, o f w hich 5 were also discovered in previous research, and 80 were novel. SVs were then associated with their nearest genes, however this resulted in a relatively low overlap with known AD genes and few to no results when applied to Gene Set Enrichment Analysis. Additionally, the data was tested for enrichment of Tandem Repeats (TRs), Transposable Elements, regulatory elements and mechanisms of gene expression, which found enrichment of TRs and heterochromatin areas and a depletion of deletions, weak enhancers, and transcription elongation areas.

Key words: Alzheimer's disease (AD), structural variant (SV), single nucleotide polymorphism (SNP), third generation sequencing (TGS)

Introduction

Alzheimer's Disease (AD) is a neurodegenerative disease which affects about 40 million people, most of whom are over 65 years old [1]. People with AD exhibit symptoms such as disorientation, cognitive impairment and memory loss. More than 20 factors involved in the progression of AD have already been identified. These include the formation of Amyloid Beta (Aβ) plaques, which are deposits of the Aβ protein — a fragment of the Amyloid-beta Precursor Protein (APP) — primarily in the brain, and the formation of neurofibrillary tangles, which are an aggregate of tau proteins that deteriorate the structural support of neurons [2].

Studying the genetic factors influencing a disease, may provide better insight into the workings of a disease and its possible treatments. Genome Wide Association Studies (GWAS) aim to identify associations between the genotype and the expression of traits by testing for differences in the frequency of genetic components between individuals exhibiting a trait and a control group [3].

As of writing, the NHGRI-EBI GWAS Catalog reports over 150 different GWAS done into AD, some of which include more than a million individuals [4]. However, all of these studies consider only one particular type of genetic component called a Single Nucleotide Polymorphism (SNP). SNPs are a single base pair change and are therefore the simplest form of variance in the DNA [5]. SNPs occur all throughout the genome and most of them are expected to have no functional consequences [6]. Depending on their location however, SNPs may influence gene expression, although their functional consequences may be non-obvious [7].

Another form of genetic variance are Structural Variants (SVs). These are variants of more than fifty base pairs in size, including Transposable Elements (TEs), which come in the form of insertions and deletions of sequences of base pairs; Tandem Repeats (TRs), which are sequences of base pairs that are repeated multiple times in a row; and more [8]. Due to their size, SVs result in larger changes to the genome than SNPs and can therefore have a greater functional impact [6]. Moreover, the functional impact of SVs usually increases with their size [9]. SVs are also able to affect genes up to hundreds of kilobases away [9] and they often affect multiple genes at once [10].

Variants can influence gene expression in many different ways. Variants which appear in the parts of genes from which proteins are transcribed, alter the proteins themselves and may thereby influence their functionality [11]. Intergenic variants can also

affect gene expression through mechanisms such as regulatory elements [10, 11], which are regions outside of genes which regulate the expression of nearby genes [12].

The gene or genes affected by a genomic variant are usually the genes closest in the genome to the variants themselves [3]. The location of genes and variants within the genome are specified in relation to a reference genome, which is a digital and representative version of the entire genome of a species assembled from a large number of donors [13]. A different method of identifying the genes affected by SNPs is to make use of Quantitative Trait Loci (QTL). QTLs are positions in the genome, called loci, that are associated with the variation of a cell- or tissue-specific quantitative trait [14]. Many large-scale QTL studies have publicly available results using which, the genes influenced by a SNP can be determined.

Given a set of affected genes, it is possible to determine the affected biological functions through Gene Set Enrichment Analysis (GSEA). GSEA determines which functions associated with a gene set used as input are statistically over represented. To this end, GSEA makes use of databases such as gene ontology databases, which maintain the biological functions of genes and their products [15]. The statistically over represented genes are then returned as output.

One factor complicating the interpretation of the functional implications of genetic variants is the concept of Linkage Disequilibrium (LD). LD refers to the non-random association of genetic variants at different positions in the genome. This implies that certain genetic variants, also those of different types such as SNPs and SVs, tend to appear together in the genome [16]. To account for LD in GWAS, fine-mapping is employed. Fine-mapping aims to cluster all SNPs found, around the few most statistically significant SNPs, which can then be used as a representative set of the results for reporting [3].

All AD-related SNPs found thus far, have been unable to explain the estimated heritability of AD of 70% [8]. Given that many GWAS with large sample sizes have already been performed, it seems likely that other genetic components besides SNPs are influencing AD. SVs are a likely candidate and are expected to account for most of the missing heritability of AD [8].

SVs have been studied scarcely in the past, due to the difficulty in detecting them accurately [10]. In spite of this, some studies have made use of such technologies to find SVs associated with AD.

Wang et al. [8] summarises the landscape of SV analysis in AD and the AD associated SVs that have been found. The publication outlines the SVs found in more than 10 papers. These 10 papers were studied, however they all rely on older sequencing technologies that are unable to accurately capture SVs. Moreover some of the studies make use of relatively small sample sizes as low 60 individuals.

In a subsequent work, Wang et al. [17] found a significant burden of deletions and duplications in AD cases with a sample size of 16,905 subjects collected by the Alzheimer's Disease Sequencing Project (ADSP). Moreover, the study reports 21 SVs in LD with AD associated SNPs and 45 ultra-rare SVs on AD genes. However, this work also relies on sequencing technologies that are unable to capture SVs accurately.

Recent advances in sequencing technologies, referred to as Third Generation Sequencing (TGS), have improved the detection of SVs significantly compared to previous methods [18]. TGS technologies provide new opportunities for more comprehensive studies of SVs than ever. Using these new technologies, a dataset was created from 214 individuals of which 93 are patients with AD and the rest are cognitively healthy centenarians. The dataset details the correlation between the length of two types of SVs — TEs and TRs — and SNPs within 500 kilobases up- and downstream of the SV. The correlation is given as a two-sided p-value calculated using a linear regression model and adjusted for population stratification. As of this writing, the dataset has not been made publicly available, and only one study has utilized the data. However, this study focuses on a single SV, thereby using only a portion of the full dataset.

This study aims to employ the SNP-SV dataset to identify new correlations between SVs and AD, and to investigate the properties of these correlated SVs, the associated genes, and their functions. To the best of our knowledge, this is the first paper to make full use of such a dataset, detailing the correlations between SNPs and the length of SVs, with the aim of finding SVs correlated with AD.

Methodology

SNP-SV Dataset Description

A total of 214 individuals, of which 93 are patients diagnosed with Alzheimer's disease from Amsterdam University Medical Center were included [21]. The remaining 121 individuals are Dutch or Dutch-speaking, cognitively healthy centenarians from the 100-plus Study cohort [22]. All participants included in these cohorts consented to the studies performed and provided written informed consent for participation in the genetic studies. Additional information regarding the study cohort is available elsewhere [23].

Only the SVs observed in at least eleven genomes (2.5%) were included in the dataset. Every SV was aligned to the reference genome GRCh38 patch 14 and was classified as being either a TR or TE. For the association between SVs and SNPs, a QTL approach was taken, where the size of the SV was used as a quantitative trait. This approach was chosen in favour of an LD approach due to the high variability of SVs. For each SV the association between SV size and SNPs within 500 kilobases up/downstream of the SV location was determined using a linear regression model adjusted for population stratification. It is important to note, that the AD status of the individuals was not used during the analysis and that the data represents SNP-SV correlations without any relation made to AD. To ensure privacy and confidentiality, the obtained results were processed into summary statistics which cannot be traced back to any specific person. This dataset will henceforth be referred to as the SNP-SV dataset.

SNP-AD Dataset Description

To link the SNP-SV correlation data to AD, a dataset containing SNP-AD associations was needed. To this end, results from the study titled "New insights into the genetic etiology of Alzheimer's disease and related dementias" were used [20]. From this study, a dataset containing all collected SNP-AD correlation values, henceforth referred to as the full summary statistics; and the reported results after fine-mapping and gene annotation, henceforth referred to as the fine-mapped results, were used. The study was selected due to its high citation count, large sample size of more than 780,000 participants of which over 111,000 are AD cases, and the availability of the full summary statistics. The full summary statistics were used in conjunction with the SNP-SV

dataset to find SVs associated with AD through SNPs associated with both. Both the full summary statistics and fine-mapped results were obtained from GWAS catalog [4].

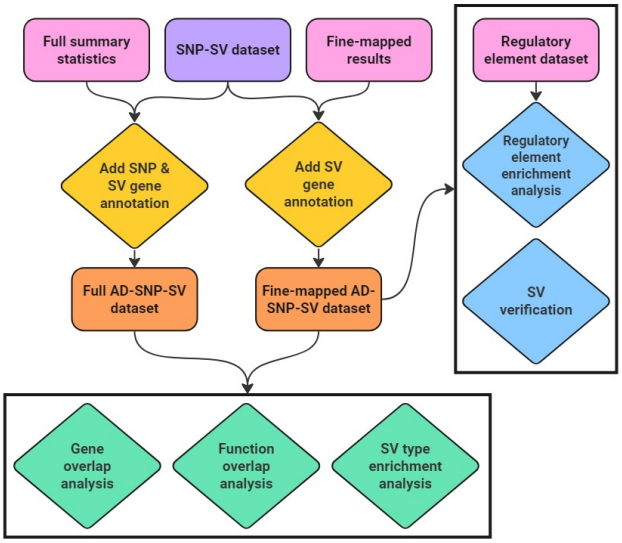

Fig. 1: **Flowchart of the performed analysis.** Rounded boxes are used to represent the different datasets used. The diamonds represent different actions or analyses performed on the datasets. An arrow that points towards a box indicates that all the actions in the box are performed for the originating dataset.

Data Analysis

Figure 1 shows an outline of the analytical steps taken and may be used as a visual aid for understanding the ensuing process description. The full summary statistics, SNP-SV dataset and fine-mapped results reference the datasets described in the preceding sections.

First, all associations with a p-value greater than the suggestive significance level of $1 \cdot 10^{-5}$ in the full summary statistics were filtered out, because higher p-values were deemed insignificant, as per the study from which the data originated [20]. All SNP-SV associations in the SNP-SV dataset with a p-value greater than $1 \cdot 10^{-5}$ were also filtered out.

Next, the full summary statistics and the fine-mapped results were merged with the SNP-SV dataset. The datasets were merged on the location, chromosome, locus, and the allele of the SNP to study the effect of SVs on AD-risk SNPs. When combining the datasets, the alleles corresponding to a positive association (β-value > 0) with AD were used. This required matching on the effect allele when the β-value was greater than 0, and flipping the sign of the β-value as well as matching on the alternative allele otherwise.

Add (SNP &) SV gene annotation. For both the SNPs and SVs, the closest upstream (U/S) and downstream (D/S) genes were calculated based on the positions of genes in reference genome GRCh38 patch 14. Gene symbols with the "LOC" prefix were removed, because many are largely uncharacterised. When a variant was located entirely within a gene, the encompassing gene was used as both the nearest upstream and downstream gene. From the nearest upstream and downstream gene, the one closest to the variant was chosen as the final nearest gene. This approach is subsequently referred to as the nearest-gene approach. This approach was chosen in favour of other QTL-based approaches due to a lack of available QTL data for SVs. The nearest-gene approach was not used for the SNPs in the dataset derived from the fine-mapped results, because all SNPs had already been mapped to genes using more rigorous analysis [20]. The nearest-gene approach applied to the SNPs in the fine-mapped results showed a 91% overlap between the genes, demonstrating the efficacy of the nearest-gene approach.

The gene-annotated dataset derived by combining the full summary statistics with the SNP-SV dataset will be referred to as the full AD-SNP-SV dataset. Similarly, the gene-annotated dataset obtained by merging the fine-mapped results with the SNP-SV dataset will be referred to as the fine-mapped AD-SNP-SV dataset. The gene and function overlap analysis as well as the SV type enrichment analysis were performed for both the aforementioned datasets, whereas the the regulatory element enrichment analysis and SV verification steps were only conducted for the fine-mapped AD-SNP-SV dataset.

Gene overlap analysis. To gain better insight into the identified genes, the overlap between the genes found and known AD associated genes was calculated. The known genes were extracted by collecting the gene annotations of the SNPs in the fine-mapped results. For the SNPs mapped to multiple genes, all the genes were included. The genes from the full and fine-mapped AD-SNP-SV datasets were extracted using three different methods. *Method 1.* Select all unique genes closest to the SVs. *Method 2.* Select all unique nearest upstream and downstream genes to the SVs. *Method 3.* Select the unique nearest upstream and downstream genes to the SVs, which are also the nearest gene to the corresponding SNP. For the fine-mapped AD-SNP-SV dataset, the mapped gene of the corresponding SNP is used instead of the nearest gene. The second and third methods were added as two alternative methods with the aim of achieving a more significant overlap with the known AD genes. The second method was chosen, because SVs on average affect 1.82 nearby genes [10], thus the nearest upstream and downstream genes are most likely to be affected by a SV. The third method was chosen such that only those genes which are close to both the SV and the corresponding SNP are selected. A gene with two nearby variants has a higher chance of being influenced. A Venn diagram of the overlap between the gene sets retrieved using the three methods and the set of known genes was calculated after stripping all genes of any additional gene suffixes such as "-AS1" or "-DT". The gene suffixes were removed to focus solely on the genes, excluding any specific transcripts.

Function overlap analysis. All sets of genes obtained using the three methods previously described were further studied using GSEA with the online g:Profiler functional profiling tool [24]. The databases used were, Gene Ontology biological processes, KEGG, Reactome and HP. When prompted, the ensemble ID with the most Gene Ontology annotations was chosen. Lastly, the overlap between the outputted functional implications of the gene sets with the GSEA results of the fine-mapped result genes was calculated and studied to understand whether the biological pathways match up with the known pathways.

SV type enrichment analysis. To attain greater insight into the SVs found, calculations were done to determine whether TRs, insertions or deletions were enriched or depleted relative to the SNP-SV dataset. These calculations were done using a Fisher exact test to obtain an OR-value and p-value, which were then adjusted for multiple testing using Bonferroni correction.

Regulatory element enrichment analysis. To better understand the mechanisms by which the identified SVs generally influence gene expression, the SVs were mapped to a regulatory element or other mechanism of gene expression. This mapping utilised a dataset that partitions the genome into regions and their corresponding regulatory element or other mechanism of gene expression, as described in the study titled "Mapping and analysis of chromatin state dynamics in nine human cell types" [25]. This dataset will henceforth be referred to as the regulatory element dataset. This dataset was selected due to its high citation count and its partitioning of the entire genome. This full partitioning is important given the limited number of SVs, as a sparse dataset would lead to few results.

The regulatory element dataset is based on the NCBI36 reference genome, which is an older version than the GRCh38 reference genome used in the fine-mapped AD-SNP-SV dataset. Therefore, this dataset had to be realigned to reference genome GRCh38. The realignment was performed using the pyliftover python package [26] in conjunction with the UCSC Chain file for converting from NCBI36 to GRCh38 [27]. Once realigned, every SV was assigned the regulatory elements or other mechanisms of gene expression of the region or regions the SV overlaps. The mapped element of all SVs were then tested for enrichment and depletion relative to the original dataset, using another Fischer exact test with Bonferroni correction for multiple testing.

SV verification. To verify the correctness of the SVs found using the methods described, the 21 SVs presented in the study titled "Structural Variation Detection and Association Analysis of Whole-Genome-Sequence Data from 16,905 Alzheimer's Diseases Sequencing Project Subjects" were used [8]. This study was chosen, because it presented the largest number of AD associated SVs and provides these SVs with their exact location based on reference genome GRCh38. All SVs in the fine-mapped AD-SNP-SV dataset were compared with the 21 SVs in LD with AD from the study. SVs were said to match if they are of the same SV type and are on the same chromosome. Furthermore the SVs must have start loci within 100 base pairs of one another and in the case of TRs, they must also have end loci within 100 base pairs of one another to account for small differences in length, measuring errors and differences in the reference genomes used.

Results

Detailed overview of the datasets

Table 1 shows statistics regarding the size and the number of elements in the SNP-SV dataset, the full summary statistics, the full AD-SNP-SV dataset, the fine-mapped results, and the fine-mapped AD-SNP-SV dataset. Table 1 shows that 5,248 of the 12,634 SNPs (41.54%) present in the full summary statistics appear in the SNP-SV dataset. A similar percentage appears for the fine-mapped results, where 39 SNPs out of the 89 (43.82%) in the fine-mapped results are included in the fine-mapped AD-SNP-SV dataset.

Figures 2 to 4 provide a detailed overview of the full AD-SNP-SV dataset in detail. Figure 2 displays the association p-values between AD, SNPs, and SVs using a Miami plot and highlights some of the genes that appear in both the results as well as the fine-mapped AD-SNP-SV dataset. The SNP-SV pairs exhibit much stronger correlations on average than the SNP-AD pairs. This can be attributed to the larger variance of SVs, given the wide range of their sizes compared to SNPs.

Figure 3 presents a volcano plot of the β-values against the corresponding p-values for the SNP-AD correlations, while Figure 4 shows a similar plot, depicting the SNP-SV correlations instead. Figure 3 shows relatively small β-values compared to Figure 4, which can be explained by the small size of SNPs relative to SVs. Additionally, the larger sample size used to collect the full summary statistics, compared to the SNP-SV dataset, allows for the detection of SNPs with very small effect sizes, further lowering the observed effect sizes [28]. The marks in Figure 4 are placed more sparsely than those in Figure 3. This observation may also be explained by the larger variance of SVs compared to SNPs.

Lastly, the 85 AD-SNP-SV associations in the fine-mapped AD-SNP-SV dataset are displayed in Table 2. The 25 associations which are highlighted in Table 2, share the property that either the upstream or the downstream gene is also the gene mapped to the corresponding SNP. Or in other terms: the rows from which the genes selected by *method 3* when applied to the fine-mapped AD-SNP-SV dataset originate are highlighted. The common gene between the SNP and SV of these associations are expected to be affected by both the SNP and the SV, therefore the expression of this gene may be more strongly impacted.

As an example, the SNP rs6733839 located on chromosome 2 at locus 127135234, was found to be significantly associated with AD ($P = 6.48 \cdot 10^{-90}$). A T nucleotide at the location of the SNP is positively associated with AD ($\beta = 0.17$), meaning that a T nucleotide increases the likelihood of developing AD. This SNP was mapped to the genes *NIFKP9* and *BIN1* in the fine-mapped results using a more rigorous method than the nearest-gene approach [20]. These mapped genes are the genes which the SNP is expected to affect. The SNP was found to be significantly associated ($P = 2.67 \cdot 10^{-7}$) with the SV chr2:127119795:INS which is an insertion on chromosome 2 at locus 127119795. The SNP is positive associated with the length of this SV ($\beta = 35.87$), indicating that the presence of the T nucleotide at the location of the SNP is linked to a longer insertion. The nearest upstream and downstream gene of the SV are *CYP27C1* and *BIN1* respectively, with *BIN1* being the closest. The SNP and the SV are thus both expected to affect the *BIN1* gene, suggesting that the expression of this gene is strongly influenced in individuals with AD.

Table 1. Summary statistics of the datasets. For each dataset, the total size is given as the number of rows. The number of unique SNPs and SVs are also indicated. Additionally, the counts of TRs, insertions, and deletions are provided, regardless of uniqueness.

Dataset	Total Size	#Unique SNPs	#Unique SVs	#TRs	#Insertions	#Deletions
SNP-SV dataset	7,209,765	3,535,362	22,590	3,323,889	1,993,380	1,892,496
Full summary statistics	12,634	12,634	N/A	N/A	N/A	N/A
Full AD-SNP-SV dataset	11,841	5,248	307	6,329	3,379	2,133
Fine-mapped results	89	89	N/A	N/A	N/A	N/A
Fine-mapped AD-SNP-SV dataset	85	39	85	52	21	12

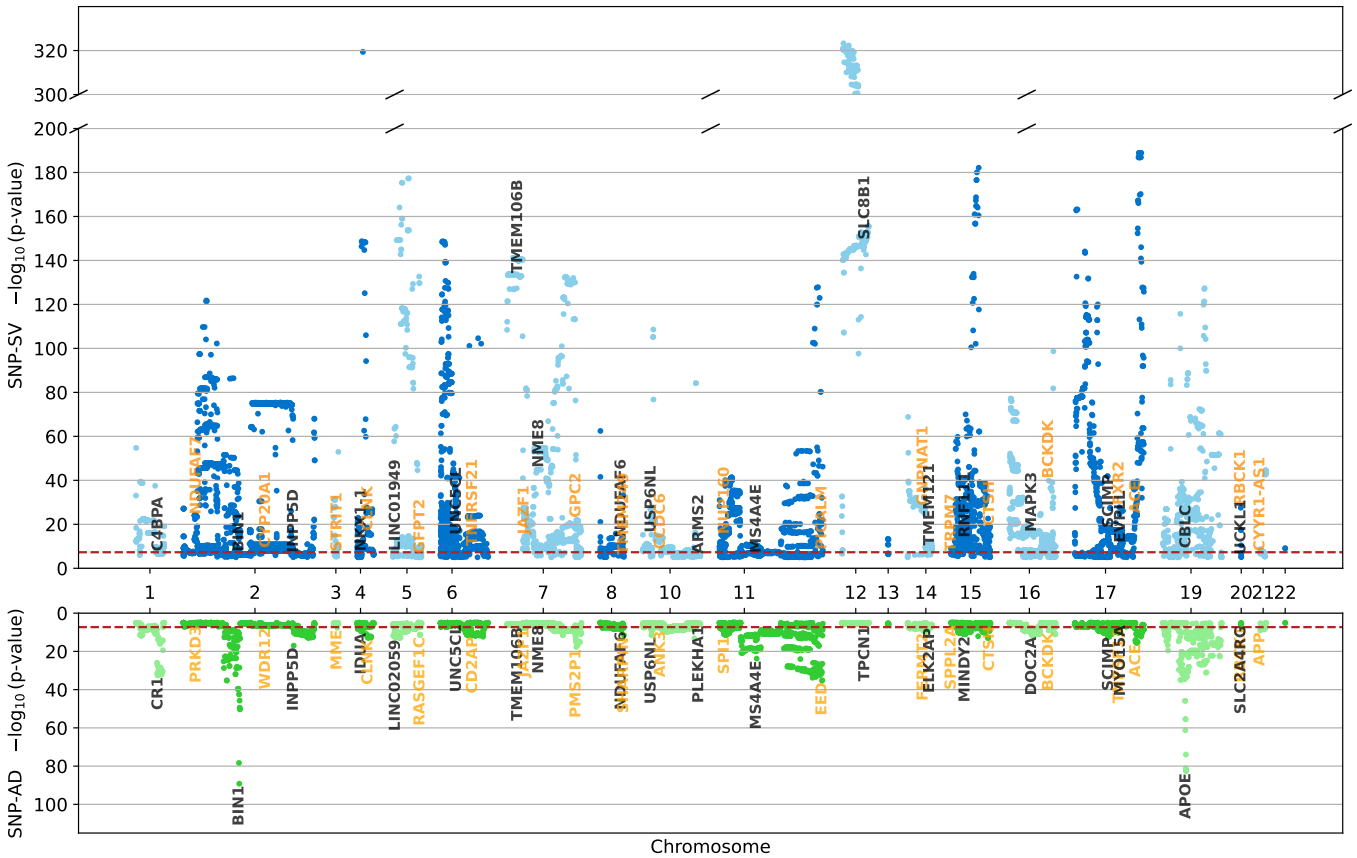

Fig. 2: **Miami plot of the full AD-SNP-SV dataset.** The $-\log_{10}$(p-value) of SNPs correlated with AD on the bottom and of SNPs correlated with SVs on the top. When multiple SVs are associated with a SNP, multiple markers are placed in the same vertical line corresponding to the different p-values of the SNP-SV pairs. The dashed line represents the genome-wide significance level ($5 \cdot 10^{-8}$) and only dots are shown that are more significant than the suggestive significance level ($1 \cdot 10^{-5}$). The loci which are also present in the fine-mapped results have their nearest gene annotated in the bottom plot and the SVs corresponding to these loci have their nearest gene annotated. Only one SV per SNP is gene-annotated for readability purposes. A break in the y-axis of the top graph is added for display purposes and lastly, the chosen colours do not hold meaning and are merely chosen to aid in viewing.

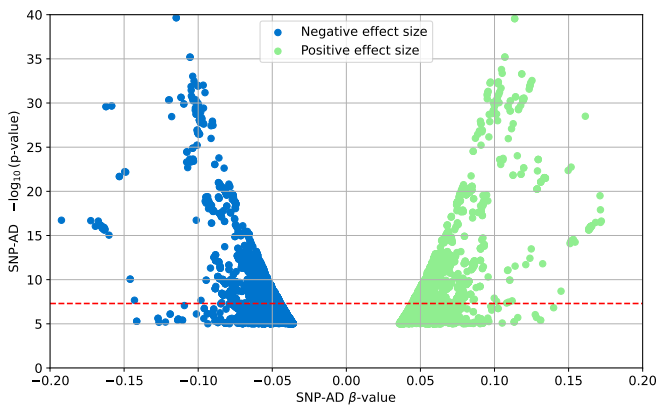

Fig. 3: **Volcano plot of SNP-AD associations in the full AD-SNP-SV dataset.** The $-\log_{10}$(p-value) of SNPs correlated with AD on the y-axis and the β value on the x-axis. Only correlations with a p-value less than the suggestive significance level ($1 \cdot 10^{-5}$). The x-axis and y-axis are limited to the ranges $[-0.20, 0.20]$ and $[0, 40]$ respectively, for visibility purposes.

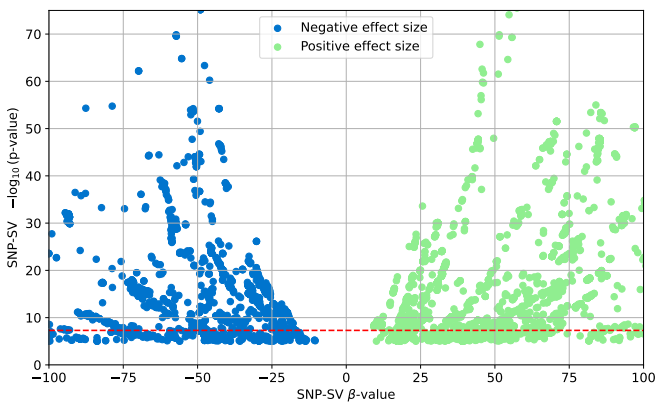

Fig. 4: **Volcano plot of SNP-SV associations in the full AD-SNP-SV dataset.** The $-\log_{10}$(p-value) of SNPs correlated with SV on the y-axis and the β value on the x-axis. Only correlations with a p-value less than the suggestive significance level ($1 \cdot 10^{-5}$). The x-axis and y-axis are limited to the ranges $[-100, 100]$ and $[0, 70]$ respectively, for visibility purposes.

Table 2. Fine-mapped AD-SNP-SV dataset summary. For every SNP the corresponding rsID, location in the form of chromosome and position, correlation with AD in the form of a p-value and β-value, and the gene(s) it was mapped to in the fine-mapped results are presented. The nucleotide the SNP and SV were matched on is given in the *Allele* column. For every SV the corresponding identifier indicating its location and type, correlation with the SNP in the form of a p-value and β-value, and the nearest up- and downstream genes are presented. The actual nearest gene to the SV is indicated by a dagger†. Rows where either the nearest upstream gene or downstream gene of the SV is included in the SNP gene(s) are highlighted.

SNP ID	Chrom	Position	SNP-AD P	SNP β	SNP Gene(s)	Allele	SV ID	SNP-SV P	SV β	SV U/S Gene	SV D/S Gene
rs679515	1	207577223	5.15e-33	0.12	CR1	T	chr1:207213754-207214039:TR	1.93e-07	130.58	CD55	C4BPA†
rs17020490	2	37304796	3.29e-06	0.05	PRKD3	C	chr2:37249519-37249678:TR	3.23e-24	29.69	NDUFAF7†	NDUFAF7†
rs17020490	2	37304796	3.29e-06	0.05	PRKD3	C	chr2:37166424-37166601:TR	7.71e-17	905.1	SULT6B1†	EIF2AK2
rs6733839	2	127135234	6.48e-90	0.17	NIFKP9,BIN1	T	chr2:127119795:INS	2.67e-07	35.87	CYP27C1	BIN1†
rs139643391	2	202878716	2.56e-07	0.06	WDR12	TC	chr2:203251778-203251817:TR	5.13e-09	162.91	CYP20A1†	CYP20A1†
rs139643391	2	202878716	2.56e-07	0.06	WDR12	TC	chr2:202986867-202986908:TR	1.89e-11	36.86	CARF†	CARF†
rs139643391	2	202878716	2.56e-07	0.06	WDR12	TC	chr2:203034349:DEL	1.08e-75	-5438.84	NBEAL1†	NBEAL1†
rs139643391	2	202878716	2.56e-07	0.06	WDR12	TC	chr2:203257940:INS	8.02e-09	55.5	CYP20A1†	CYP20A1†
rs10933431	2	233117202	1.04e-17	0.09	INPP5D	C	chr2:233143712-233144825:TR	7.00e-07	66.52	INPP5D†	INPP5D†
rs61762319	3	155084189	2.14e-08	0.14	MME	G	chr3:155307274-155307534:TR	2.36e-07	40.45	PLCH1	STRIT1†
rs3822030	4	993555	5.04e-10	0.05	IDUA	T	chr4:1397436-1398660:TR	1.83e-07	-901.13	NKX1-1†	UVSSA
rs3822030	4	993555	5.04e-10	0.05	IDUA	T	chr4:1391665-1391893:TR	5.36e-10	33.45	NKX1-1	UVSSA†
rs3822030	4	993555	5.04e-10	0.05	IDUA	T	chr4:1372668-1372904:TR	3.95e-08	131.82	UVSSA†	UVSSA†
rs3822030	4	993555	5.04e-10	0.05	IDUA	T	chr4:1356977-1357314:TR	1.74e-09	-159.8	UVSSA†	UVSSA†
rs3822030	4	993555	5.04e-10	0.05	IDUA	T	chr4:1282657-1283811:TR	7.18e-07	-99.52	MAEA†	CTBP1-DT
rs3822030	4	993555	5.04e-10	0.05	IDUA	T	chr4:1278815-1279291:TR	2.62e-07	74.32	MAEA†	CTBP1-DT
rs3822030	4	993555	5.04e-10	0.05	IDUA	T	chr4:1025958-1026310:TR	1.12e-06	-54.15	FGFRL1†	FGFRL1†
rs3822030	4	993555	5.04e-10	0.05	IDUA	T	chr4:926812-928005:TR	1.64e-06	170.86	GAK†	GAK†
rs3822030	4	993555	5.04e-10	0.05	IDUA	T	chr4:872915-873059:TR	3.74e-06	17.02	GAK†	GAK†
rs3822030	4	993555	5.04e-10	0.05	IDUA	T	chr4:867472-867675:TR	1.87e-07	-23.42	GAK†	GAK†
rs3822030	4	993555	5.04e-10	0.05	IDUA	T	chr4:1308029:INS	3.13e-09	-139.95	MAEA†	MAEA†
rs3822030	4	993555	5.04e-10	0.05	IDUA	T	chr4:1391670:DEL	9.81e-10	-33.9	NKX1-1	UVSSA†
rs6846529	4	11023507	1.25e-13	0.07	LINC02498,MIR572	C	chr4:10963590-10964348:TR	1.07e-16	16.12	HS3ST1	CLNK†
rs62374257	5	86927378	1.41e-13	0.07	LINC02059,MIR4280HG	C	chr5:87076866:INS	1.24e-07	124.3	LINC01949†	LINC02059
rs113706587	5	180201150	3.38e-12	0.09	RASGEF1C	A	chr5:180296732-180296983:TR	1.73e-06	48.09	GFPT2†	MAPK9
rs10947943	6	41036354	6.21e-06	0.05	OARD1,UNC5CL	G	chr6:40959080:INS	5.16e-15	-176.4	UNC5CL†	LRFN2
rs7767350	6	47517390	5.05e-12	0.06	CD2AP	T	chr6:47345040:INS	7.29e-10	140.27	CD2AP	TNFRSF21†
rs7767350	6	47517390	5.05e-12	0.06	CD2AP	T	chr6:47505628:INS	1.73e-10	360.46	CD2AP†	CD2AP†
rs7767350	6	47517390	5.05e-12	0.06	CD2AP	T	chr6:47596994:DEL	1.02e-16	767.72	CD2AP†	CD2AP†
rs13237518	7	12229967	5.12e-07	0.04	TMEM106B	C	chr7:12242078:DEL	6.30e-134	-323.98	TMEM106B†	TMEM106B†
rs1160871	7	28129126	1.12e-07	0.05	JAZF1	GTCTT	chr7:28174682-28175046:TR	3.62e-15	-215.76	JAZF1†	JAZF1†
rs1160871	7	28129126	1.12e-07	0.05	JAZF1	GTCTT	chr7:28110447-28110666:TR	3.06e-24	-100.08	JAZF1†	JAZF1†
rs6966331	7	37844191	4.81e-06	0.04	NME8,GPR141	C	chr7:37847104-37848371:TR	1.63e-45	-205.38	NME8†	GPR141
rs6966331	7	37844191	4.81e-06	0.04	NME8,GPR141	C	chr7:37646264-37646416:TR	1.38e-06	18.36	GPR141†	ELMO1-AS1
rs6966331	7	37844191	4.81e-06	0.04	NME8,GPR141	C	chr7:37797035:INS	3.41e-11	-140.75	NME8†	GPR141
rs7384878	7	100334426	2.13e-18	0.08	PMS2P1	T	chr7:100172634-100172834:TR	1.02e-22	-32.96	GPC2†	GPC2†
rs13276936	8	94983473	2.83e-09	0.05	NDUFAF6	T	chr8:95060003:INS	1.06e-13	-121.05	NDUFAF6†	NDUFAF6†
rs34173062	8	144103704	2.93e-12	0.11	SHARPIN	A	chr8:144347174:INS	6.76e-06	161.01	TMEM249†	SCRT1
rs7912495	10	11676714	2.87e-12	0.06	ECHDC3,USP6NL-AS1	G	chr10:11514162-11514319:TR	6.01e-16	-45.06	USP6NL†	USP6NL†
rs7068231	10	60025170	6.79e-09	0.05	ANK3,LINC01553	G	chr10:59882002-59882627:TR	2.50e-08	779.91	CCDC6†	CCDC6†
rs7908662	10	122413396	3.30e-06	0.04	PLEKHA1	A	chr10:122457364:DEL	2.90e-06	-89.53	HTRA1	ARMS2†
rs10437655	11	47370397	8.21e-12	0.06	SPI1	A	chr11:47865987-47866642:TR	3.11e-15	-270.25	PTPRJ	NUP160†
rs10437655	11	47370397	8.21e-12	0.06	SPI1	A	chr11:47785089:INS	1.32e-15	153.82	NUP160†	NUP160†
rs10437655	11	47370397	8.21e-12	0.06	SPI1	A	chr11:47866116:DEL	6.98e-15	273.65	PTPRJ	NUP160†
rs1582763	11	60254475	1.65e-24	0.09	MS4A4A	G	chr11:60206462-60206529:TR	1.03e-06	64.14	MS4A4E†	MS4A4E†
rs3851179	11	86157598	6.50e-36	0.11	LINC02695,RNU6-560P	C	chr11:86130846-86130897:TR	3.48e-08	30.98	EED	PICALM†
rs6489896	12	113281983	2.54e-06	0.07	TPCN1	C	chr12:113300042-113300190:TR	3.52e-149	-67.23	SLC8B1†	SLC8B1†
rs6489896	12	113281983	2.54e-06	0.07	TPCN1	C	chr12:113245307:DEL	6.25e-267	326.59	TPCN1†	TPCN1†
rs17125924	14	52924962	5.82e-10	0.09	FERMT2	G	chr14:52807759-52807996:TR	3.38e-28	145.58	FERMT2	GNPNAT1†
rs7157106	14	105761758	1.46e-07	0.06	IGHG3,IGHG1	A	chr14:105520681-105521705:TR	3.85e-10	-1037.73	TMEM121†	TEDC1
rs7157106	14	105761758	1.46e-07	0.06	IGHG3,IGHG1	A	chr14:105502093-105502283:TR	2.90e-11	-149.24	TMEM121	TEDC1†
rs7157106	14	105761758	1.46e-07	0.06	IGHG3,IGHG1	A	chr14:105502115:INS	2.72e-11	-155.87	TMEM121	TEDC1†
rs8025980	15	50701814	6.09e-06	0.04	RN7SL354P,SPPL2A	A	chr15:50564247-50564314:TR	4.25e-07	-72.74	TRPM7†	TRPM7†
rs8025980	15	50701814	6.09e-06	0.04	RN7SL354P,SPPL2A	A	chr15:50897104:INS	1.39e-09	-362.59	AP4E1†	SPPL2A
rs602602	15	58764824	9.65e-12	0.06	MINDY2-DT,SNORD3P1	T	chr15:58982256-58982361:TR	1.05e-13	20.18	RNF111†	SLTM
rs602602	15	58764824	9.65e-12	0.06	MINDY2-DT,SNORD3P1	T	chr15:58620650:DEL	1.93e-29	-233.85	ADAM10†	ADAM10†
rs602602	15	58764824	9.65e-12	0.06	MINDY2-DT,SNORD3P1	T	chr15:58877172:INS	1.50e-20	-176.76	SLTM†	MINDY2
rs12592898	15	78936857	1.28e-06	0.06	CTSH	G	chr15:78945368-78946282:TR	4.28e-19	53.76	RASGRF1	CTSH†
rs12592898	15	78936857	1.28e-06	0.06	CTSH	G	chr15:78768022-78768123:TR	5.11e-09	-72.26	ADAMTS7†	ADAMTS7†
rs12592898	15	78936857	1.28e-06	0.06	CTSH	G	chr15:78644314:INS	6.58e-06	-170.22	ADAMTS7	CHRNB4†
rs12592898	15	78936857	1.28e-06	0.06	CTSH	G	chr15:78785999:INS	9.61e-06	178.7	ADAMTS7†	ADAMTS7†
rs1140239	16	30010081	4.61e-12	0.06	DOC2A	C	chr16:30153076-30153473:TR	2.90e-16	-66.91	CORO1A	MAPK3†
rs1140239	16	30010081	4.61e-12	0.06	DOC2A	C	chr16:29843499-29843605:TR	1.47e-06	-107.33	MVP†	MVP†
rs1140239	16	30010081	4.61e-12	0.06	DOC2A	C	chr16:30124001:INS	2.96e-06	-458.39	CORO1A	MAPK3†
rs1140239	16	30010081	4.61e-12	0.06	DOC2A	C	chr16:30153303:DEL	6.37e-16	66.48	CORO1A	MAPK3†
rs889555	16	31111250	1.03e-09	0.06	BCKDK	C	chr16:31104046-31105481:TR	1.37e-40	-433.62	BCKDK†	VKORC1
rs889555	16	31111250	1.03e-09	0.06	BCKDK	C	chr16:31104491:DEL	3.34e-11	64.02	BCKDK†	VKORC1
rs7225151	17	5233752	2.60e-12	0.09	ZNF594-DT,SCIMP	A	chr17:5240001-5241339:TR	3.71e-18	452.03	RABEP1	SCIMP†
rs7225151	17	5233752	2.60e-12	0.09	ZNF594-DT,SCIMP	A	chr17:4878941-4879282:TR	2.50e-11	22.85	MINK1†	MINK1†
rs7225151	17	5233752	2.60e-12	0.09	ZNF594-DT,SCIMP	A	chr17:4910154:INS	2.07e-11	59.87	GP1BA	CHRNE†
rs58840546	17	7560327	2.12e-07	0.09	TNFSF12-TNFSF13,TNFSF13	TCAA	chr17:7609863-7610066:TR	1.84e-28	-99.02	FXR2†	FXR2†
rs58840546	17	7560327	2.12e-07	0.09	TNFSF12-TNFSF13,TNFSF13	TCAA	chr17:7621096:DEL	4.20e-14	-439.94	SAT2†	FXR2
rs2242595	17	18156140	4.55e-06	0.06	MYO15A	G	chr17:18388026-18388064:TR	2.36e-11	-21.14	EVPLL†	EVPLL†
rs2242595	17	18156140	4.55e-06	0.06	MYO15A	G	chr17:17619233-17619302:TR	3.79e-11	-769.97	SMCR2	PEMT†
rs2242595	17	18156140	4.55e-06	0.06	MYO15A	G	chr17:17619244:INS	1.01e-09	-1220.55	SMCR2	PEMT†
rs4277405	17	63471557	7.24e-16	0.07	PPIAP55,CYB561	T	chr17:63500872-63501666:TR	1.22e-24	-121.64	KCNH6	ACE†
rs4277405	17	63471557	7.24e-16	0.07	PPIAP55,CYB561	T	chr17:63488530:INS	3.65e-29	211.95	ACE†	ACE†
rs1358782	20	413334	1.86e-06	0.05	RBCK1	G	chr20:407888-408064:TR	3.12e-28	54.12	RBCK1†	TRIB3
rs6742	20	63743088	2.49e-06	0.05	SLC2A4RG	C	chr20:63941895-63942378:TR	6.16e-06	-74.7	UCKL1†	UCKL1†
rs6742	20	63743088	2.49e-06	0.05	SLC2A4RG	C	chr20:63849687-63849979:TR	7.04e-06	-26.83	ABHD16B†	ZBTB46-AS1
rs6742	20	63743088	2.49e-06	0.05	SLC2A4RG	C	chr20:63846949-63847337:TR	6.94e-06	38.69	ABHD16B†	ZBTB46-AS1
rs6742	20	63743088	2.49e-06	0.05	SLC2A4RG	C	chr20:63731268-63731761:TR	3.40e-07	-37.8	ZGPAT†	ZGPAT†
rs6742	20	63743088	2.49e-06	0.05	SLC2A4RG	C	chr20:63682195-63682514:TR	1.46e-07	-15.61	RTEL1-TNFRSF6B†	RTEL1-TNFRSF6B†
rs6742	20	63743088	2.49e-06	0.05	SLC2A4RG	C	chr20:63760716:DEL	3.74e-11	-144.37	ZBTB46†	ZBTB46†
rs2154481	21	26101558	1.02e-09	0.05	APP	T	chr21:26338643:INS	1.89e-07	-16.63	CYYR1-AS1†	APP

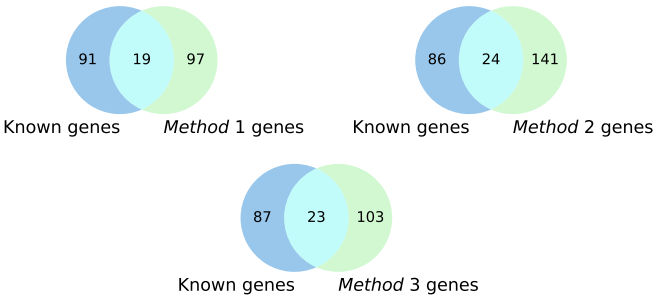

Fig. 5: **Venn diagram of gene sets from the full AD-SNP-SV dataset overlapping with known AD genes.** The known genes were obtained by extracting all unique genes from the fine-mapped results. The other gene sets were derived by applying the three methods to the full AD-SNP-SV dataset.

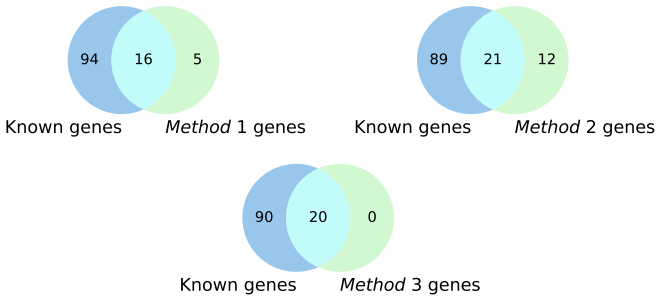

Fig. 6: **Venn diagram of gene sets from the fine-mapped AD-SNP-SV dataset overlapping with known AD genes.** The known genes were obtained by extracting all unique genes from the fine-mapped results. The other gene sets were derived by applying the three methods to the fine-mapped AD-SNP-SV dataset.

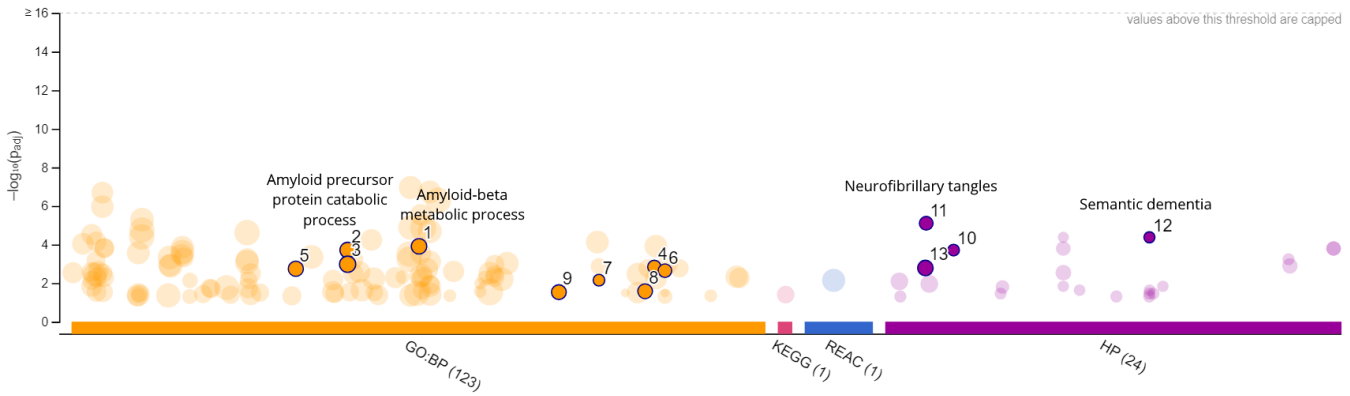

Fig. 7: **Scatter plot of g:Profiler GSEA tool applied to known genes.** The known genes were extracted from the fine-mapped results. The scatter plot, created by the g:Profiler functional profiling tool, displays Gene Ontology biological processes, KEGG, Reactome and HP database functions as marks, with the corresponding enrichment p-values on the y-axis [24]. Function names containing the terms "amyloid-beta", "amyloid precursor", "neurofbrillary tangle" and "dementia" have their marks highlighted. Additionally, some marks corresponding to functions mentioned in the text have been annotated with the function's name.

Gene overlap analysis

Figure 5 shows the overlap between known AD genes obtained from the fine-mapped results and the genes extracted from the full AD-SNP-SV dataset using three different methods: *Method 1.* Select all unique genes closest to the SVs. *Method 2.* Select all unique nearest upstream and downstream genes to the SVs. *Method 3.* Select the unique nearest upstream and downstream genes to the SVs, which are also the nearest gene to the corresponding SNP. Figure 6 shows a similar figure, but instead the genes are extracted from the fine-mapped AD-SNP-SV dataset and in *method 3*, the mapped gene from the fine-mapped results is used instead of nearest SNP gene.

Figure 5 shows a relatively low overlap for each method compared to the number of genes in each method's set. This suggests that the genes extracted using the three different methods are either previously unknown or that the methods do not accurately capture the genes affected by the SVs. Figure 6shows a relatively high overlap when compared to the number of genes obtained using the three methods. However, the overlap compared to the number of known genes remains low, which is partly due to the number of unique genes in the fine-mapped AD-SNP-SV dataset being lower than the number of known genes. Figures 5 and 6 both show a similar overlap for each method indicating that the methods perform equally. The small differences in the overlap of the methods can be attributed to the variations in the number of genes extracted by each method.

Function overlap analysis

Figure 7 shows the 149 functions significantly enriched by the known AD genes, obtained by performing GSEA using g:Profiler [24] on the known AD genes. Many of the resulting functions relate to processes known to be associated with AD, including "Amyloid precursor protein catabolic process" (related to APP), "Amyloid-beta matabolic process" (related to Aβ), "Neurofibrillary tangles"; "Semantic dementia"; and more.

Figure 8 depicts the overlap between the functions significantly enriched by the known AD genes and the functions derived by performing GSEA on the method's gene sets. Fewer significant functions were found when using the gene sets extracted by the three different methods compared to the known functions. This

Table 3. Regulatory element counts in the fine-mapped AD-SNP-SV dataset compared to the genome. This table shows the number of each regulatory element in the genome compared to the count of each regulatory element mapped to all SVs in the fine-mapped AD-SNP-SV dataset using the regulatory element dataset. The table has been split into two major rows for visibility purposes.

Regulatory element	Weak Enhancer	Weak Txn[1]	Heterochromatin	Strong Enhancer	Weak Promoter	Insulator
Genome	178,463	82,187	74,863	64,078	35,021	33,214
Dataset	2	16	54	2	4	1
Regulatory element	Txn[1] Elongation	Polycomb-repressed	Txn[1] Transition	Active Promoter	Repetitive/CNV[2]	Poised Promoter
Genome	26,479	2,435	16,215	15,256	14,087	5,253
Dataset	9	5	4	0	0	1

[1]Txn = Transcription, [2]CNV = Copy Number Variation

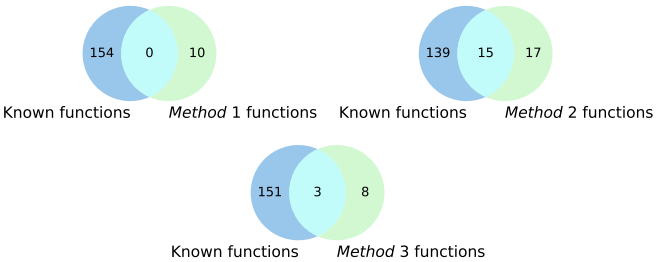

Fig. 8: **Venn diagram of function sets from the full AD-SNP-SV dataset overlapping with known AD functions.** The known function set and the method's sets were created by performing GSEA on the known AD genes and the gene sets obtained through the three methods applied to the fine-mapped AD-SNP-SV dataset respectively.

indicates that the functions of the genes in the method-derived gene sets are not as strongly correlated as those of the known genes. Additionally, the functions obtained from the method-derived gene sets did not include any AD-related terms such as "amyloid-beta", "amyloid precursor", "neurofibrillary tangle", and "dementia".

GSEA was also performed on the gene sets obtained from the three methods applied to the fine-mapped AD-SNP-SV dataset. This resulted in no functions being significantly enriched, likely due to the small size of the gene sets and the lack of significant correlation in the functions of the genes.

SV type enrichment analysis

The full and fine-mapped AD-SNP-SV datasets were tested for enrichment of TRs, deletions and insertions compared to the SNP-SV dataset. The number of SVs of each SV type can be viewed in Table 1. For the full AD-SNP-SV dataset, a significant enrichment was found in the number of TRs ($OR = 1.34,\ P = 4.6 \cdot 10^{-57}$) and a significant depletion in the number of deletions ($OR = 0.62,\ P = 6.0 \cdot 10^{-99}$). For the fine-mapped AD-SNP-SV dataset, the TRs were also significantly enriched ($OR = 1.84,\ P = 0.02$) and the deletions significantly depleted ($OR = 0.46,\ P = 0.03$). These results contrasts previous findings where a burden of both TRs and deletions was found in AD cases [17].

Regulatory element enrichment analysis

The regulatory element dataset was incorporated into the fine-mapped AD-SNP-SV dataset. Table 3 shows the number of each regulatory element mapped to a SV in the fine-mapped AD-SNP-SV dataset compared to the count of each regulatory element in the genome. The regulatory elements mapped to the SVs were tested for enrichment relative to all regulatory elements in the genome. This test found a significant enrichment of heterochromatin areas ($OR = 8.13,\ P = 1.0 \cdot 10^{-21}$) and a significant depletion of both weak enhancers ($OR = 0.05,\ P = 2.0 \cdot 10^{-12}$) and transcription elongation areas ($OR = 0.16,\ P = 0.01$). These results indicate that heterochromatin areas are more frequently the regulatory elements affecting gene expression for the SVs found, while weak enhancers are less frequently involved.

SV verification

All SVs from the fine-mapped AD-SNP-SV dataset, as shown in Table 2, were compared with 21 AD-associated SVs identified in a previous study [8]. This procedure found that 5 out of the 85 unique SVs from the fine-mapped AD-SNP-SV dataset were also present in the previous study. These 5 SVs, along with their corresponding known SVs, are listed in Table 4.

Table 4. Verified SVs. The verified SVs shown together with the SVs they were mapped to from a previous study [8].

Verified SV	Known SV
chr2:203034349:DEL	chr2:203034369-203039560:DEL
chr6:40959080:INS	chr6:40959079-40959079:INS
chr7:12242078:DEL	chr7:12242077-12242399:DEL
chr10:122457364:DEL	chr10:122457302-122457747:DEL
chr12:113245307:DEL	chr12:113245316-113245625:DEL

Discussion

Summary and conclusion

This study aimed to employ the SNP-SV dataset to identify new correlations between SVs and AD and to investigate the properties of these correlated SVs, the associated genes, and their functions. To this end, the SNP-SV dataset was used in combination with the full summary statistics of a large GWAS, to find 307 unique SVs correlated with AD through their corresponding SNPs. Additionally, a significant enrichment was found in the number of TRs and a significant depletion in the number of deletions for both the full AD-SNP-SV dataset and the fine-mapped AD-SNP-SV dataset. This suggests that TRs more often affect AD associated genes and deletions less. All methods used for extracting genes from the full AD-SNP-SV dataset produced relatively low overlap with known AD genes compared to the number of genes extracted. These results were corroborated by the GSEA done using the different genes sets, which also resulted in low overlap with the known AD functions and the results having no obvious relation to AD. The fine-mapped results combined with the SNP-SV dataset produced 85 AD-SNP-SV associations. The SNP

gene from the fine-mapped results and either the upstream or downstream gene of the SV are the same in 25 of these associations. Incorporating the regulatory elements and mechanisms of gene expression into the fine-mapped AD-SNP-SV dataset revealed a significant enrichment of heterochromatin areas and a significant depletion of both weak enhancers and transcription elongation areas. Calculating the overlap between the known AD genes and the genes extracted from the fine-mapped AD-SNP-SV dataset using three different methods, revealed a higher overlap relative to number of genes extracted compared to the results obtained from the full AD-SNP-SV dataset. Applying the genes in GSEA however, resulted in no significant functions for all methods. 5 out of the 85 SVs in the fine-mapped AD-SNP-SV dataset were also present in a set of 21 SVs obtained from previous work [8].

307 unique SVs were found in the full AD-SNP-SV dataset. However many of these SVs relate to SNPs which are in LD with the SNPs actually responsible for AD, adding another level of indirection, making it hard to study whether these SVs actually influence AD. 85 out of the 307 unique SVs were also present in the fine-mapped AD-SNP-SV dataset. These SVs directly associate with the SNPs in the fine-mapped results and thus relate to AD with only one level of indirection. Nevertheless, for many of these SVs it is still unclear and hard to study, through which mechanics they affect the genes associated with AD, due to, for example, the large distances between the SVs and the genes. The significant enrichment of heterochromatin areas points to these areas having a large impact on AD gene expression. Nevertheless, how AD gene expression is affected is more obvious for the 25 associations for which the SNP and SV have a gene in common, as presented in the highlighted rows of Table 2. The effects that variants have on nearby genes are more obvious as variants more commonly affect their closest genes than genes at large distances. Thus when the closest gene to a SV is a gene associated with AD, the SV is more likely to affect the gene than genes at a distance. Additionally, the gene is expected to be affected by both the SNP as well as the SV, increasing the likelihood that the expression of the gene is indeed influenced. Lastly, 5 SVs have also been observed in a previous study [8] and have therefore been found by two independent studies, thereby strengthening the proof that these SVs are related to AD.

Limitations and future work

SNP-SV dataset. This study extensively utilises the SNP-SV dataset, which, to the best of our knowledge, is the first dataset to detail the correlation between SNPs and the lengths of SVs. The dataset was generated from a relatively small sample size of 214 individuals, all of whom are Dutch, were treated in a Dutch hospital, and/or are Dutch-speaking. Extending the dataset by incorporating additional sequencing data from a more ethnically diverse population, generated through TGS, is recommended. Including more sequencing data is likely to reveal additional SNP-SV associations and enhance the reliability of the SVs and findings.

SNP-AD dataset. The SNP-SV dataset used in the analysis was selected due to the high citation count of the corresponding study and the extensive sample size of over 780,000 participants. However, there are many more AD GWAS, which could provide additional insights when combined with the SNP-SV dataset in the manner described. Additionally, the SNP-SV dataset may be combined with GWAS of traits besides AD. Such studies would likely uncover many previously unknown SVs associated with many various traits.

SV gene annotation. SVs were gene annotated using the simple nearest-gene approach. It is unclear whether this method resulted in the selection of appropriate genes, as knowledge regarding which genes are affected by which SVs and the manner in which the SVs affect genes is still limited. QTL studies mapping SVs to quantitative traits, or GWAS with large cohorts which study the association between SVs and a trait, would significantly improve our understanding of the effects of SVs. This knowledge would, in turn, improve the process of identifying the genes associated with specific SVs. Additionally, GWAS would provide more options for verifying the SVs identified. Both the QTL studies and GWAS, would require more SVs sequencing results obtained through TGS to become available.

Gene and function overlap analysis. The gene and function overlap analyses show a relatively low overlap between the known AD genes and functions, and those obtained using the three methods applied to the full and fine-mapped AD-SNP-SV datasets. This poor overlap is likely due to the nearest-gene approach used for gene annotation of the SVs, which may have led to inadequate gene selection and, in turn, poor overlap between the functions. As mentioned, SVs commonly affect multiple genes and they are able to affect genes over long distances [10]. Therefore, the nearest-gene approach is likely too simplistic, and the low overlap is not necessarily a result of a poor selection of SVs. Utilising alternative methods for selecting the genes than the ones presented may also help improve the overlap.

SV enrichment analysis. In this study, only the types of SVs were tested for enrichment. Future research could explore other properties of SVs for enrichment, such as their length and the distance between SVs and their corresponding SNPs. The SV type enrichment analysis was not performed for the full AD-SNP-SV dataset because the SVs in this dataset are less likely to be associated with AD compared to those in the fine-mapped AD-SNP-SV dataset. Nevertheless, future studies may consider performing these enrichment tests to uncover further insights.

Regulatory element enrichment analysis. The regulatory element dataset was selected due to the high citation count of the corresponding study and because it partitions the entire genome. The dataset is however based on the outdated reference genome NCBI36, released in 2006. No newer version of the dataset based on reference genome GRCh38 was available, thus the dataset had to be realigned. Some regions could not be realigned, however this did not affect the results as no SVs overlapped with these regions. Future studies using this same method for different SVs may encounter issues however. To resolve these problems, a new regulatory elements dataset based on GRCh38 should be created, or a more robust realignment procedure capable of mapping all regions to GRCh38 should be developed.

SV verification. The verification procedure utilised a dataset containing only 21 SVs. The studies reviewed in a publication by Wang et al. [8] also present SVs associated with AD. However, most describe only one SV, are based on older reference genomes, and do not provide precise genomic locations. Future work should expand the verification set as more SVs associated with AD are identified.

References

1. Querfurth Henry W. and LaFerla Frank M. (2010) Alzheimer's Disease. New England Journal of Medicine, 362, 329–344.
2. Ramachandran,A.K., Das,S., Joseph,A., Gurupur Gautham,S., Alex,A.T. and Mudgal,J. (2021) Neurodegenerative Pathways in Alzheimer's Disease: A Review. Curr Neuropharmacol, 19, 679–692.
3. Uffelmann,E., Huang,Q.Q., Munung,N.S., de Vries,J., Okada,Y., Martin,A.R., Martin,H.C., Lappalainen,T. and Posthuma,D. (2021) Genome-wide association studies. Nat Rev Methods Primers, 1, 1–21.
4. Sollis,E., Mosaku,A., Abid,A., Buniello,A., Cerezo,M., Gil,L., Groza,T., Güneş,O., Hall,P., Hayhurst,J., et al. (2023) The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res, 51, D977–D985.
5. Shastry,B.S. (2009) SNPs: Impact on Gene Function and Phenotype. In Komar,A.A. (ed), Single Nucleotide Polymorphisms: Methods and Protocols. Humana Press, Totowa, NJ, pp. 3–22.
6. Mortazavi,M., Ren,Y., Saini,S., Antaki,D., St. Pierre,C.L., Williams,A., Sohni,A., Wilkinson,M.F., Gymrek,M., Sebat,J., et al. (2022) SNPs, short tandem repeats, and structural variants are responsible for differential gene expression across C57BL/6 and C57BL/10 substrains. Cell Genom, 2, 100102.
7. Robert,F. and Pelletier,J. (2018) Exploring the Impact of Single-Nucleotide Polymorphisms on Translation. Front Genet, 9, 507.
8. Wang,H., Wang,L.-S., Schellenberg,G. and Lee,W.-P. (2023) The role of structural variations in Alzheimer's disease and other neurodegenerative diseases. Front Aging Neurosci, 14, 1073905
9. Weischenfeldt,J., Symmons,O., Spitz,F. and Korbel,J.O. (2013) Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet, 14, 125–138.
10. Scott,A.J., Chiang,C. and Hall,I.M. (2021) Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. Genome Res, 31, 2249–2257.
11. Shastry,B.S. (2009) SNPs: Impact on Gene Function and Phenotype. In Komar,A.A. (ed), Single Nucleotide Polymorphisms: Methods and Protocols. Humana Press, Totowa, NJ, pp. 3–22.
12. Doane,A.S. and Elemento,O. (2017) Regulatory elements in molecular networks. Wiley Interdiscip Rev Syst Biol Med, 9, 10.1002/wsbm.1374.
13. Schneider,V.A., Graves-Lindsay,T., Howe,K., Bouk,N., Chen,H.-C., Kitts,P.A., Murphy,T.D., Pruitt,K.D., Thibaud-Nissen,F., Albracht,D., et al. (2016) Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. 10.1101/072116.
14. Zheng,Z., Huang,D., Wang,J., Zhao,K., Zhou,Y., Guo,Z., Zhai,S., Xu,H., Cui,H., Yao,H., et al. (2020) QTLbase: an integrative resource for quantitative trait loci across multiple human molecular phenotypes. Nucleic Acids Research, 48, D983–D991.
15. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S., et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102, 15545–15550.
16. Slatkin,M. (2008) Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nat Rev Genet, 9, 477–485.
17. Wang,H., Dombroski,B.A., Cheng,P.-L., Tucci,A., Si,Y.-Q., Farrell,J.J., Tzeng,J.-Y., Leung,Y.Y., Malamon,J.S., Wang,L.-S., et al. (2023) Structural Variation Detection and Association Analysis of Whole-Genome-Sequence Data from 16,905 Alzheimer's Diseases Sequencing Project Subjects. medRxiv, 10.1101/2023.09.13.23295505, 13 September 2023, pre-print: not peer-reviewed.
18. Xiao,T. and Zhou,W. (2020) The third generation sequencing: the advanced approach to genetic diseases. Transl Pediatr, 9, 163–173.
19. Wang,H., Wang,L.-S., Schellenberg,G. and Lee,W.-P. (2023) The role of structural variations in Alzheimer's disease and other neurodegenerative diseases. Front Aging Neurosci, 14, 1073905.
20. Bellenguez,C., Küçükali,F., Jansen,I.E., Kleineidam,L., Moreno-Grau,S., Amin,N., Naj,A.C., Campos-Martin,R., Grenier-Boley,B., Andrade,V., et al. (2022) New insights into the genetic etiology of Alzheimer's disease and related dementias. Nat Genet, 54, 412–436.
21. van der Flier,W.M. and Scheltens,P. (2018) Amsterdam Dementia Cohort: Performing Research to Optimize Care. Journal of Alzheimer's Disease, 62, 1091–1111.
22. Holstege,H., Beker,N., Dijkstra,T., Pieterse,K., Wemmenhove,E., Schouten,K., Thiessens,L., Horsten,D., Rechtuijt,S., Sikkes,S., et al. (2018) The 100-plus Study of cognitively healthy centenarians: rationale, design and cohort description. Eur J Epidemiol, 33, 1229–1249.
23. Tesi,N., Van Der Lee,S., Hulsman,M., Van Schoor,N.M., Huisman,M., Pijnenburg,Y., Van Der Flier,W.M., Reinders,M. and Holstege,H. (2024) Cognitively healthy centenarians are genetically protected against Alzheimer's disease. Alzheimer's & Dementia, 10.1002/alz.13810.
24. Reimand,J., Kull,M., Peterson,H., Hansen,J. and Vilo,J. (2007) g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res, 35, W193–W200.
25. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M., et al. (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. Nature, 473, 43–49.
26. Guttman,M. (2017) pyliftover: Python tool to lift-over genome coordinates. Version 0.4. Available at https://pypi.org/project/pyliftover/.
27. Hinrichs,A.S., et al. (2006) The UCSC Genome Browser Database: update 2006. Nucleic Acids Res, 34(Database issue):D590-8. Available at http://hgdownload.cse.ucsc.edu/goldenPath/hg18/liftOver/hg18ToHg38.over.chain.gz.
28. Park,J.-H., Wacholder,S., Gail,M.H., Peters,U., Jacobs,K.B., Chanock,S.J. and Chatterjee,N. (2010) Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet, 42, 570–575.