RESEARCH ARTICLE OPEN ACCESS

# Identifying Aging and Alzheimer Disease–Associated Somatic Variations in Excitatory Neurons From the Human Frontal Cortex

Meng Zhang, MSc, Gerard A. Bouland, MSc, Henne Holstege, PhD, and Marcel J.T. Reinders, PhD

*Neurol Genet* 2023;9:e200066. doi:10.1212/NXG.0000000000200066

**Correspondence**
Dr. Reinders
m.j.t.reinders@tudelft.nl

## Abstract

### Background and Objectives

With age, somatic mutations accumulated in human brain cells can lead to various neurologic disorders and brain tumors. Because the incidence rate of Alzheimer disease (AD) increases exponentially with age, investigating the association between AD and the accumulation of somatic mutation can help understand the etiology of AD.

### Methods

We designed a somatic mutation detection workflow by contrasting genotypes derived from whole-genome sequencing (WGS) data with genotypes derived from scRNA-seq data and applied this workflow to 76 participants from the Religious Order Study and the Rush Memory and Aging Project (ROSMAP) cohort. We focused only on excitatory neurons, the dominant cell type in the scRNA-seq data.

### Results

We identified 196 sites that harbored at least 1 individual with an excitatory neuron–specific somatic mutation (ENSM), and these 196 sites were mapped to 127 genes. The single base substitution (SBS) pattern of the putative ENSMs was best explained by signature SBS5 from the Catalogue of Somatic Mutations in Cancer (COSMIC) mutational signatures, a clock-like pattern correlating with the age of the individual. The count of ENSMs per individual also showed an increasing trend with age. Among the mutated sites, we found 2 sites tend to have more mutations in older individuals (16:6899517 [*RBFOX1*], $p = 0.04$; 4:21788463 [*KCNIP4*], $p < 0.05$). In addition, 2 sites were found to have a higher odds ratio to detect a somatic mutation in AD samples (6:73374221 [*KCNQ5*], $p = 0.01$ and 13:36667102 [*DCLK1*], $p = 0.02$). Thirty-two genes that harbor somatic mutations unique to AD and the *KCNQ5* and *DCLK1* genes were used for gene ontology (GO)–term enrichment analysis. We found the AD-specific ENSMs enriched in the GO-term "vocalization behavior" and "intraspecies interaction between organisms." Of interest we observed both age-specific and AD-specific ENSMs enriched in the $K^+$ channel–associated genes.

### Discussion

Our results show that combining scRNA-seq and WGS data can successfully detect putative somatic mutations. The putative somatic mutations detected from ROSMAP data set have provided new insights into the association of AD and aging with brain somatic mutagenesis.

From the Delft Bioinformatics Lab (M.Z., G.A.B., H.H., M.J.T.R.), Delft University of Technology; Department of Human Genetics (M.Z., H.H.), Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC; and Department of Human Genetics (G.A.B., M.J.T.R.), Leiden University Medical Center, the Netherlands.

Funding information and disclosures are provided at the end of the article. Full disclosure form information provided by the authors is available with the full text of this article at Neurology.org/NG.

## Glossary

**AD** = Alzheimer disease; **Aβ** = amyloid beta; **COSMIC** = Catalogue of Somatic Mutations in Cancer; **CSTB** = cystatin B; **FDR** = false discovery rate; **GO** = gene ontology; **MAP** = the Rush Memory and Aging Project; **ROS** = the Religious Order Study; **SBS** = single base substitution; **WGS** = whole-genome sequencing.

Somatic mutations are postzygotic genetic variations that can result in genetically different cells within a single organism.[1] Possible reasons for the occurrence and accumulation of somatic mutations in human brains are errors occurring during DNA replication and gradual failing of DNA repair mechanisms caused by extensive oxidative stress.[2,3] Previous studies have shown that brain somatic mutations originating in neuronal stem/progenitor cells can lead to various neurologic disorders and brain tumors.[4-6] While mutations in postmitotic neurons have been found to play an important role in age-related and neurodegenerative diseases,[7] this association remains relatively poorly understood. The link between the accumulation of age-related mutations in neurons and neurodegenerative disease is intuitively worth exploring, considering aging is a major risk factor of many neurodegenerative diseases such as Alzheimer disease (AD).[8]

AD is the most predominant form of dementia and characterized by the extracellular accumulation of amyloid beta (Aβ) plaques and the intracellular aggregation of phosphorylated tau protein into neurofibrillary tangles.[9] A recent study identified several putative pathogenic brain somatic mutations enriched in genes that are involved in hyperphosphorylation of tau.[10] These results indicate that the aggregation of these neuropathologic substrates can be partly explained by the accumulation of brain somatic mutations, which raises a new direction for investigating the pathogenic mechanism of AD.

Most age-related somatic mutations are only present in a small group of postmitotic neurons or even in a single neuron. For this reason, ultra-deep bulk sequencing and matched peripheral tissues are often required.[10] These type of data are often generated for 1 specific research question with a relatively high cost and are not always available from public databases. By contrast, the availability of public single-cell RNA sequencing (scRNA-seq) data sets has exploded because of continuous technological innovations, increasing throughput, and decreasing costs.[11] scRNA-seq data are most often used for expression-based analyses, such as revealing complex and rare cell populations, uncovering regulatory relationships between genes, and tracking the trajectories of distinct cell lineages in development.[12,13] We hypothesized that scRNA-seq data can also be used to detect somatic mutations. We are not the first to realize this; in fact, other studies pioneered on different solutions to call variants in this setting. For example, researchers[14] compare 3 different variant callers (GATK, Strelka2, and Mutect2) and show that a two-fold higher number of single nucleotide variants (SNVs) can be detected from the pooled scRNA-seq when compared with bulk data. As another example, Vu et al.,[15] developed a specific variant caller (SCmut) that can identify specific cells that harbor mutations discovered in bulk cell data by smartly controlling the false positives. Both studies applied their methodology to detect single-cell somatic mutations in cancer.

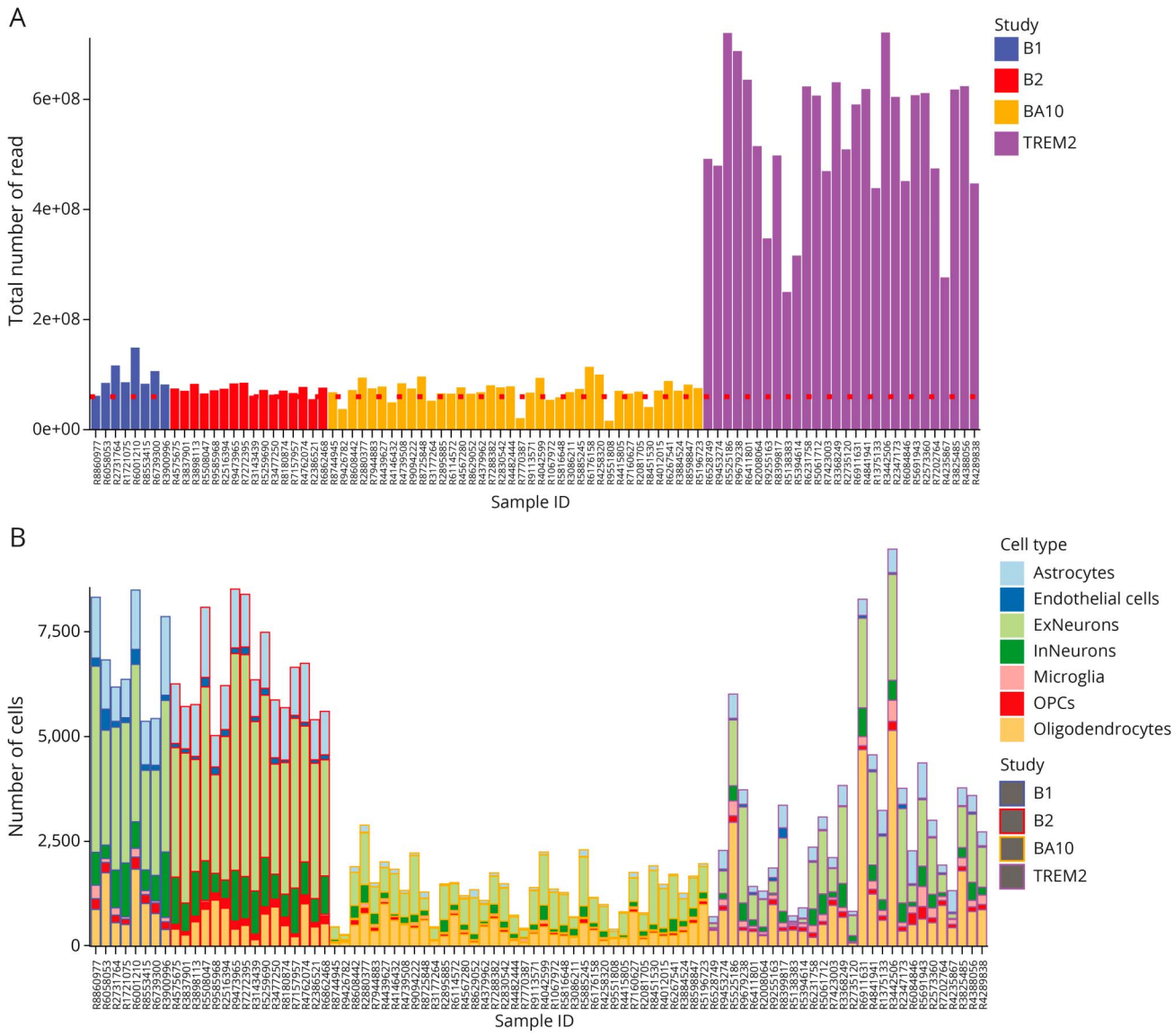
In this study, we designed a workflow to detect brain-specific somatic mutation by contrasting genotypes identified with whole-genome sequencing (WGS) data with genotypes identified with scRNA-seq data. To call variants in single-cell data, we exploit the VarTrix caller from 10× Genomics[16] and apply various filters to ensure their quality. For each putative somatic mutation, we investigated associated genes and their respective relationship with AD and age. In addition, we investigated whether AD and age coincide with an increasing number of somatic mutations.

## Methods

### Case Selection

The scRNA-seq data and WGS data were obtained from the Religious Order Study (ROS) and the Rush Memory and Aging Project (MAP), 2 longitudinal cohort studies of aging and dementia.[17] Information collected as part of these studies, collectively known as ROSMAP, includes clinical data, detailed postmortem pathologic evaluations, and tissue omics profiling. The scRNA-seq data used in this project were from 3 sources: (1) snRNAseqMFC study (n = 24), (2) snRNAseqAD_TREM2 study (n = 32), and (3) snRNAseqPFC_BA10 study (n = 48); specifically, these 3 studies used single-nuclei RNA sequencing data. All specimens for these 3 scRNA-seq data sources were collected postmortem from the frontal cortex; subregions might slightly differ between studies. The scRNA-seq data from the 3 studies were all sequenced according to the 10× Genomics manufacturer's protocol. Detailed information for cell partitioning, reverse transcription, library construction, and sequencing run configuration for the 3 studies is available on Synapse (snRNAseqMFC: syn16780177, snRNAseqAD_TREM2: syn21682120, snRNAseqPFC_BA10: syn21261143). WGS data were from a subset of the ROSMAP participants with DNA obtained from brain tissue, whole blood, or lymphocytes transformed with the EBV. The details for WGS library preparation and sequencing and WGS Germline variants calling were previously described.[18] Individuals (n = 90) with both scRNA-seq data and WGS data (27 from brain tissue and 63 from whole blood) available were selected for this study. Individuals annotated with no cognitive impairment or mild cognitive impairment were defined as nondemented (ND) controls; patients with AD with or without other cause of cognitive impairment were defined as AD samples.

**Figure 1** scRNA Reads and Cell Count Across Selected Samples




Participants (n = 90) from the ROSMAP project with both scRNA-seq data and whole genome sequencing data available were selected for this study. (A) The distribution of the number of scRNA reads across individuals. The dashed red line indicates the cutoff of $<6 \times 10^7$ for the minimal read coverage, i.e., individuals below this line were excluded from the study (n = 9). The colors indicated the study that included an individual. Individuals who colored either blue or red were from the 2 batches (B1 and B2) of the snRNAseqMFC study. Individuals colored orange were from the snRNAseqAD_BA10 study, and individuals colored purple were from the snRNAseqPFC_TREM2 study. (B) The number of cells per cell type per individual. The cell types were distinguished with 7 different colors (see legend). The colors of the edges indicated different studies, as in (A). Abbreviations: ExNeurons = excitatory neurons; InNeurons = inhibitory neurons; OPCs = oligodendrocyte progenitor cells; scRNA-seq = single-cell RNA sequencing; scRNA = single-cell RNA.

## Standard Protocol Approvals, Registrations, and Patient Consents

The ROS/MAP studies and substudies were all approved by an Institutional Review Board of Rush University Medical Center, and all participants signed an informed consent, Anatomical Gift Act, and a repository consent to share data and biospecimens.

## Cell Type Annotation

Each scRNA-seq data set was separately processed for clustering and cell type annotation, which was performed as follows. The processed count matrix was loaded in Seurat (version 3.2.2). Data were log-normalized and scaled before analysis. Next, with the 2,000 most variable genes (default with Seurat), principal components analysis was performed. The number of principal components used for clustering was determined using the elbow method. Furthermore, Seurat's FindNeighbours and FindCluster functions were used, which use Louvain clustering; the resolution was set at 0.5. An Uniform Manifold Approximation and Projection plot (eFigure 1, links.lww.com/NXG/A593) was made to visualize and inspect the clusters. The following cell types were identified using known and previously used markers: excitatory neurons (*SLC17A7*, *CAMK2A*, and *NRGN*), inhibitory neurons (*GAD1* and *GAD2*), astrocytes (*AQP4* and *GFAP*),

**Table** Summary Characteristics of Selected Sample From the ROSMAP Study

| Group | Cogdx[a] | n | Sex | Age, mean ± SD (range) |
|---|---|---|---|---|
| **Nondemented** | 1 | 33 | 23 F; 19 M | 85.7 ± 4.2 (76–90) |
| | 2 | 8 | | |
| | 3 | 1 | | |
| **Alzheimer disease** | 4 | 32 | 19 F; 14 M | 87.1 ± 3.9 (74–90) |
| | 5 | 1 | | |
| **Other dementia** | 6 | 1 | 1F | 83 |

[a] Cognitive diagnosis (cogdx) is defined under 6 categories: 1, NCI: no cognitive impairment (no impaired domains); 2, MCI: mild cognitive impairment (1impaired domain) and NO other cause of CI; 3, MCI: mild cognitive impairment (1 impaired domain) AND another cause of CI; 4, AD: Alzheimer dementia and NO other cause of CI (NINCDS PROB AD); 5, AD: Alzheimer dementia AND another cause of CI (NINCDS POSS AD); and 6, Other dementia: other primary cause of dementia.

oligodendrocytes (*MBP*, *MOBP*, and *PLP1*), oligodendrocyte progenitor cell (*PDGFRA*, *VCAN*, and *CSPG4*), microglia (*CSF1R*, *CD74*, and *C3*), and endothelial cells (*FLT1* and *CLDN5*).[19] Based on the markers' expression patterns across clusters determined by Seurat's FindMarkers function, cell types were assigned to cells (eAppendix, 1, links.lww.com/NXG/A591). When clusters were characterized by markers of multiple cell types, they were assigned "unknown."

## scRNA-seq Short Variants Calling

Single-nuclei RNA reads were mapped to the reference human genome GRCh37 using STAR aligner (STAR v2.7.9a). After alignment, duplicate reads were identified using MarkDuplicates (Picard v2.25.0), and reads with unannotated cell barcodes were removed using samtools (smatools v1.11). Reads containing Ns in their cigar string were split into multiple supplementary alignments using SplitNCigarReads (GATK v4.2.0.0) to match the conventions of DNA aligner. Base Quality Recalibration was performed per sample to detect and correct for patterns of systematic errors in the base quality scores using BaseRecalibrator and ApplyBQSR (GATK v4.2.0.0). Short variant discovery was performed on chromosome 1-22 with a 2-step process. HaplotypeCaller was run on each sample separately in genomic variant call format mode (GATK v4.2.0.0) producing an intermediate file format called gVCF (for genomic VCF). gVCFs from each individual were combined together and run through a joint genotyping step (GATK v4.2.0.0) to produce a multisample VCF file. eFigure 2 (links.lww.com/NXG/A593) indicates the steps of scRNA-seq short variants calling in a flow chart. Variant filtration was then performed using bcftools (bcftools v1.11). A basic hard filtering referring to GATK technical documentation[20] was performed using cutoffs of (1) the total read depth DP <50,000; (2) the quality of calling QUAL >100; (3) the quality by depth >2; (4) the strand odds ratio <2; and (5) the strand bias Fisher exact test FS <10.

## Identical Individual Check Using IBD Estimation

To make sure the sequences of scRNA-seq and WGS are matching and from the same individual, we performed a pairwise identical by descent (IBD) estimation using filtered variants from scRNA-seq and WGS in a combined VCF file. The estimation was calculated using PLINK v1.9. The proportion IBD value PI_HAT from the output of PLINK was used as the estimator; when the profiles are from the same individual, the PI_HAT value will be close to 1; otherwise it will be close to 0.

## Somatic Mutation Detection Using VarTrix

VarTrix, a software tool for extracting single-cell variant information from 10× Genomics single-cell data, was used to detect somatic mutations. For single-nuclei gene expression data, VarTrix requires a precalled variant set in VCF format, an associated set of alignments in BAM or compressed alignment file format, a genome FASTA file, and a cell barcodes file produced by Cell Ranger as input. After an exploratory phase, we observed that only cells annotated as excitatory neuron had enough read coverage for somatic mutation detection. Therefore, for each individual, a subset of the BAM file including only reads from cells annotated as excitatory neuron was used as the input of VarTrix. Correspondingly, the precalled variant set was also detected from the subset of the BAM file, which included only barcodes from cells annotated as excitatory neuron.

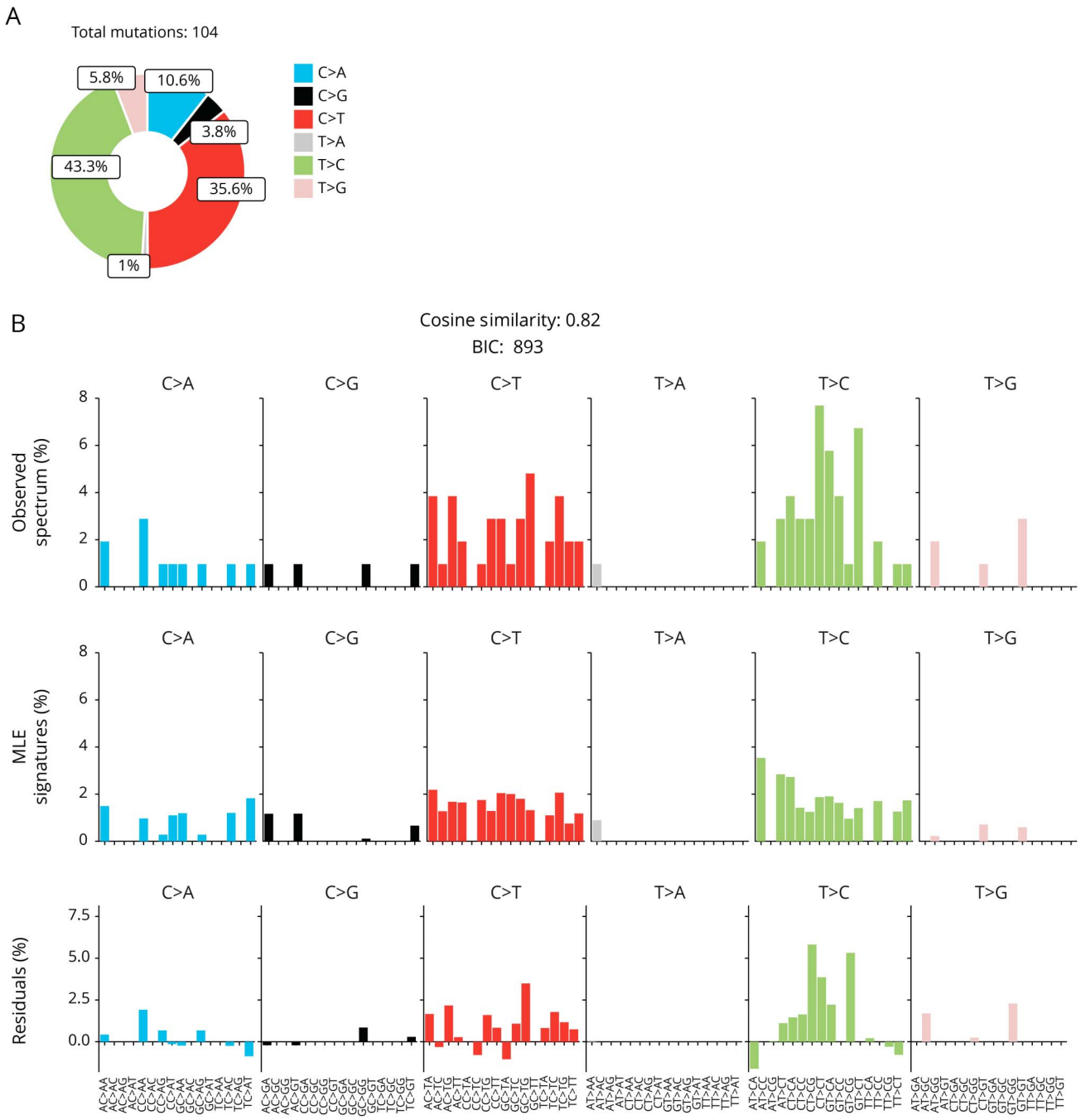
Human reference genome GRCh37 was used as the genome FASTA file. In this study, VarTrix was run in coverage mode generating a reference coverage matrix and an alternate coverage matrix indicating the number of reads that support the reference allele and the alternate allele. These matrices were later used for filtering variant sites and detecting somatic mutations in the excitatory neurons.

Because the scRNA-seq data were collected from 3 studies, the average coverage varied between different sources. To minimize the batch effect from different studies, we filtered the variant site based on the read number of each individual. Specifically, we calculated a cutoff $C_i$ for each individual $i$ as below:

$$C_i = \frac{n_i}{\sum_{n=1}^{N} n_i / N} C$$

where $n_i$ is the number of reads for individual $i$, and $N$ is the total number of individuals. The constant value $C$ is set as 25 to guarantee that a sufficient amount of reads (>5) can support a variant site for every sample. A variant site would be used for somatic mutation detection when for all individuals the read depth at this site is higher than the cutoff $C_i$ for that individual. Next, a somatic mutation was identified as present in an individual when: (1) the genotype of this individual at the site in WGS was ref/ref, and the ratio of reads that support the alternate allele in scRNA-seq is larger than 0.1 at the same site or (2) the

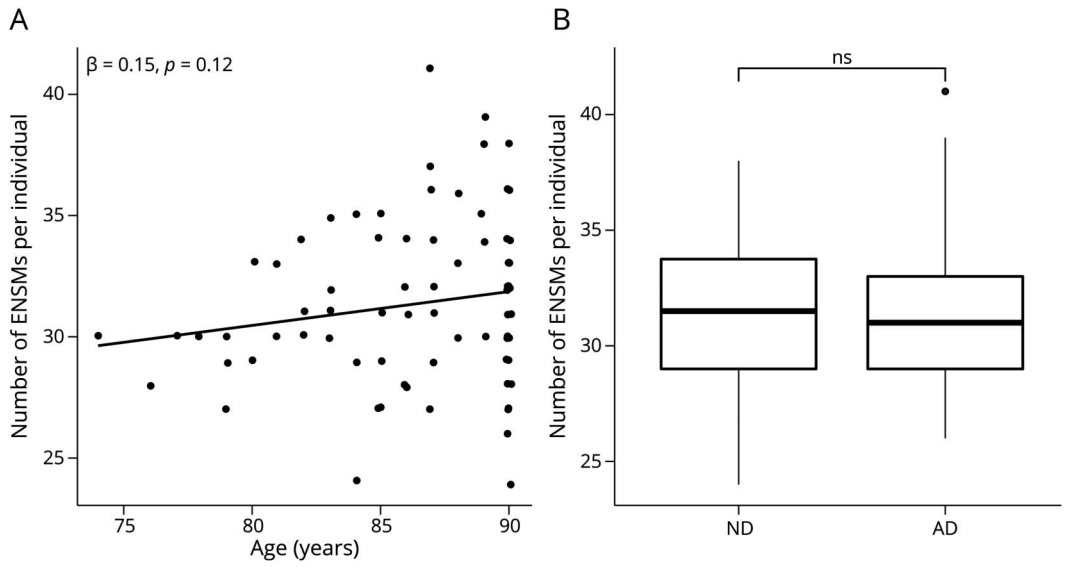
**Figure 2** Mutation Signature of 104 Putative Excitatory Neuron–Specific Single-Nucleotide Variations in the Brain




Among the 30 COSMIC SBS signatures, SBS5 was identified as the model that best explains the observed pattern of putative somatic SNVs by Mutalisk. The cosine similarity with the 104 putative excitatory neuron–specific SNVs and the corresponding BIC for each COSMIC SBS signature are shown in eFigure 5 (links.lww.com/NXG/A593). (A) The percentage of each substitution subtype in the 104 putative excitatory neuron–specific SNVs. Subtype T > C and C > T are the dominate subtypes and account for 43.3% and 35.6% of the fraction separately. (B) The top panel shows the observed distribution of 104 putative excitatory neuron–specific SNVs across the 96 possible mutation types; the middle panel shows the distribution of the identified signature (SBS5); the bottom panel shows the difference of each base substitution subtype between the top and middle panels. The same plots of the other top 5 mutational signatures in largest cosine similarity (i.e., signatures 25, 12, 26, and 9, except for signature 5) are shown in eFigure 5 (links.lww.com/NXG/A593). Abbreviations: BIC = Bayesian information criterion; SBS = single base substitution; SNV = single-nucleotide variation.

genotype of this individual at the site in WGS was alt/alt, and the ratio of reads that support the reference allele in scRNA-seq is larger than 0.1 at the same site. When the genotype of an individual at a certain site was heterozygote in WGS, we ignored the site for that individual, regardless of the allele ratio in scRNA-seq, because we cannot distinguish an observed homozygous variant at a site in scRNA-seq is due to somatic mutagenesis or reads missing when there is a heterozygous variant in WGS at the same site.

**Figure 3** Quantitative Comparison of the Number of ENSMs Regarding AD and Aging




(A) The number of ENSMs per individual against the age of the individual. The line shows how this number regresses with age. The significance of the coefficient ($\beta \neq 0$) was tested using a *t* test. The same analysis for AD and non-AD samples separately is shown in eFigure 6 (links.lww.com/NXG/A593). (B) Boxplot of the number of ENSMs in ND controls and patients with AD. The Wilcoxon rank sum test does not show a significance difference (ns). Abbreviations: AD = Alzheimer disease; ENSMs = excitatory neuron–specific somatic mutations; ND = nondemented.

### Mutation Signature Analysis

To characterize the contribution of mutation signatures, we pooled all putative somatic SNVs for signature analysis. We formatted the pooled SNVs in a VCF file and used it as input for running Mutalisk[21] with the following configurations: maximum likelihood estimation method; linear regression. The input file was compared with 30 single base substitution (SBS) signatures from the COSMIC mutational signature database. The best model of signature combination was suggested from the tool by considering the Bayesian information criterion.

### Variants Annotation and Effect Prediction

The gene annotation and functional effect prediction for all putative variants were performed using SnpEff (SnpEff v5.0).[22] The human genome GRCh37 was used as reference genome. If there were multiple genes mapping to 1 variant site, the gene having higher putative effect was used for the disease and age association analyses.

### GO-Term Enrichment Analysis

The gene ontology (GO)–term enrichment analysis was performed using topGo package (version 2.38.1) in R and compressed by reduce and visualize gene ontology (REVIGO) with semantic similarity score "Lin." The genes that were annotated to the variant sites with read depths higher than the cutoffs for all samples were used as background. The *p* values from the uneliminated GO-terms were corrected using "Benjamini&Hochberg" method, and significant results were reported with false discovery rate (FDR) <0.05.

### Statistical Analysis

All calculations were performed using R (version 3.6.3). The R-scripts for statistical analysis are available on GitHub: github.com/mzhang0215/ENSM_project. The Wilcoxon rank sum test, linear regression, Fisher exact test, and logistic regression were performed using the "stats" R package. By categorizing the "presence" of a somatic mutation as 1 and the "absence" of a somatic mutation as 0, the logistic function was defined as follows: $p = 1/(1+\exp(-(\beta_0 + \beta_1 age + \beta_2 group)))$, where *age* is the age of the sample at death, *group* is the assigned group for the individual based on the cogdx category, and $\beta_{0..2}$ are the coefficients of the intercept and the explanatory variables. For this analysis, only individuals from the AD and ND groups were used.

### Data Availability

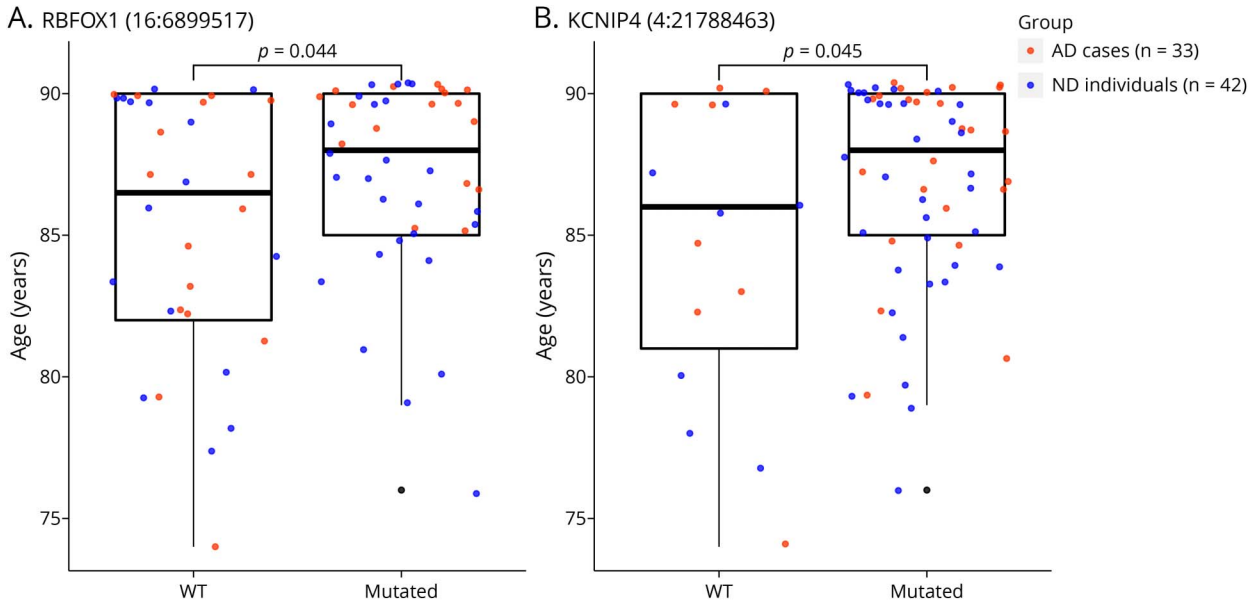
Data are available from the AD Knowledge Portal (contact via adknowledgeportal.org) for researchers who meet the criteria for access to confidential data.

## Results

### Excitatory Neuron–Specific Somatic Mutations

To study somatic mutations acquired over age and between demented (AD) and nondemented (ND) persons, we retrieved data from 90 participants from the ROSMAP study for which WGS data in blood or brain, as well as scRNA-seq data of the frontal cortex was present (Methods). Because the scRNA-seq data (n = 90) were collected within 3 different studies, the read coverage for samples varied between the studies (Figure 1A). To reduce the bias generated from the unbalanced read coverage, we excluded individuals (n = 9) with a total read count smaller than $6 \times 10^7$ and applied a sample-specific cutoff for the required read coverage to detect a somatic mutation based on the total read count per sample (Methods). Cells from the scRNA-seq data were annotated

**Figure 4** Occurrence of Somatic Mutation With Age in (A) *RBFOX1* and (B) *KCNIP4* Genes




Red dots: cases with AD; blue dots: ND individuals. Logistic regression was used to test the prevalence of somatic mutations with increasing age. Abbreviations: AD = Alzheimer disease; ND = nondemented; WT = wild type.

according to 7 major cell types (Methods). Because the amount of cells varied for different cell types (Figure 1B), we first explored the feasibility of detecting somatic mutations for each cell type. This exploratory analysis showed that somatic mutations could only be detected from the excitatory neurons (when requiring a minimum number of reads (≥5) per sample for a putative variant site, Methods), the dominant cell type in our scRNA-seq data. This underpins that a sufficient amount of cells is needed for scRNA-seq–based somatic mutation detection. As a consequence, we focus our analysis on excitatory neurons only. To further ensure data quality, we excluded individuals (n = 5) who had less than 200 excitatory neurons. After filtering, 76 participants (23 from the snRNAseqMFC study, 30 from the snRNAseqPFC_BA10 study, and 23 from the snRNAseqAD_TREM2 study) had an adequate read coverage and sufficient number of excitatory neurons. Demographic data (sex, age at death, and cognitive diagnosis [cogdx] categories[23]) of these participants are summarized in Table. More than 72% of them were 85 years of age or older at death; 56% were women. Individuals were grouped based on their cognitive diagnosis in either being nondemented (n = 42) or being an AD sample (n = 33).

## Summary of Detected ENSMs

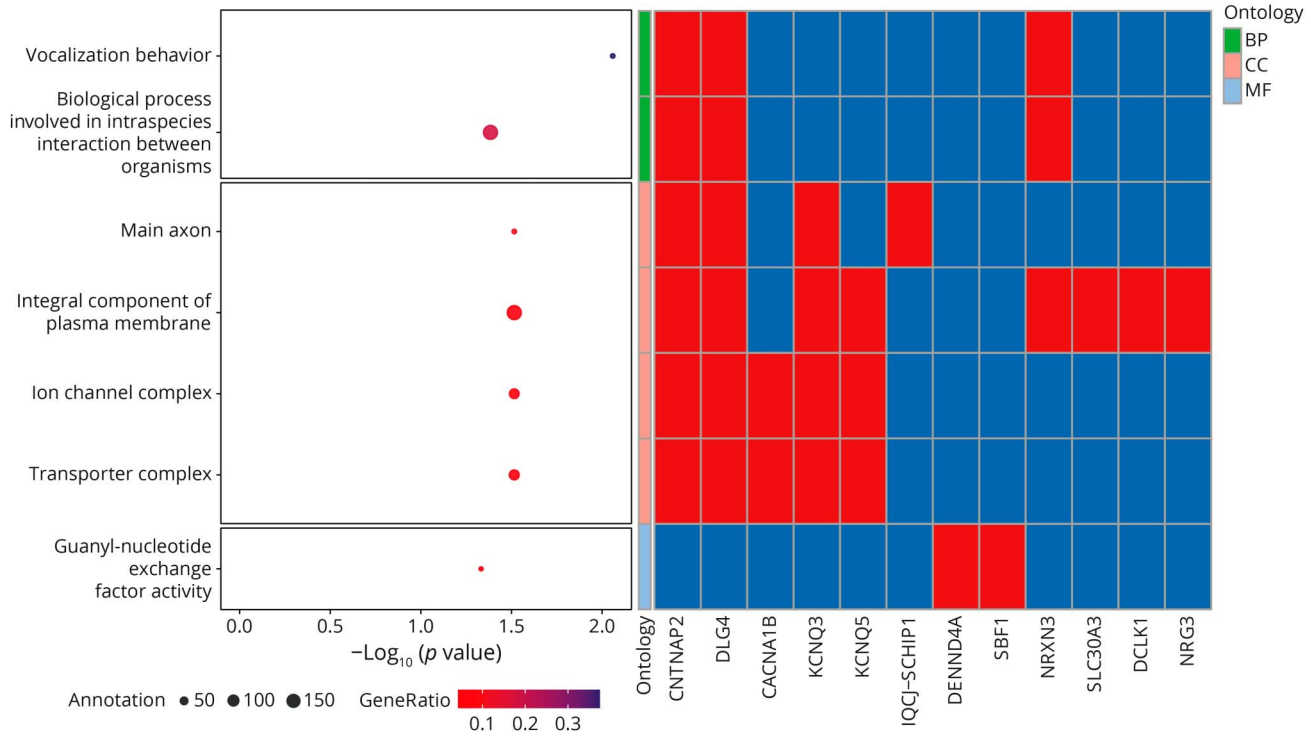
Somatic mutations in the 76 participants were detected using the workflow described in the Methods. For that, the scRNA-seq data of the excitatory neurons are compared with WGS data of the blood (n = 23) or brain (n = 53). IBD estimation using shared variant sites confirmed the matching between the scRNA-seq and WGS samples (pair-wised PI_HAT >0.85, eFigure 3, links.lww.com/NXG/A593, Methods). From the 9,751,193 short variants called from the scRNA-seq data, we identified 196 sites that harbored excitatory neuron–specific somatic mutations (ENSMs). These genetic sites map to 127 genes (Methods), and 104 sites among them were SNVs. From these 196 sites, 98 were shared between multiple individuals (n > 2) and thus are recurrent somatic mutations (eFigure 4, links.lww.com/NXG/A593). A few sites have mutations present in almost all individual genomes, which are likely to be either RNA editing events[24]; transcription errors, which can occur in a wide variety of genetic contexts with several different patterns[25,26]; or technical errors.[27] 53 sites have mutations uniquely present in the brains of the AD samples (eTable 1, links.lww.com/NXG/A594).

Per individual genome, the number of ENSMs ranged from 24 to 41. This does not seem to contradict the other observations that found an average of approximately 12 somatic SNVs in hippocampal formation tissue using deep bulk exome sequencing[10] and an average amount of approximately 1,700 somatic mutations (substitutions approximately 1,500; indels approximately 200) in neurons using a whole-genome duplex single-cell sequencing protocol.[28] However, this comparison might be complicated by the differences in sequencing and somatic mutation detection methods, as well as brain regions.

## Number of ENSMs Increase With age

To characterize the ENSMs, a mutation signature analysis was performed on the 104 detected putative somatic SNVs (Methods). The results show that, from the 30 COSMIC mutational signatures, SBS5 best explains the observed pattern of putative somatic SNVs by Mutalisk (Figure 2, eFigure 5, links.lww.com/NXG/A593). SBS5 is a clock-like signature, i.e., the number of mutations correlates with the age of the individual. This suggests that the underlying mutational

**Figure 5** GO-Terms Enriched With Genes Having AD-Specific ENSMs




32 genes that have ENSMs seen only in AD samples, and the *KCNQ5* and *DCLK1* genes that have a higher occurrence in AD samples are used in the GO-term enrichment analysis. The left panel of the figure shows the enriched terms, their corrected *p* value, the number of genes annotated with that term (size of circle), and the fraction of overlapping genes that harbor an AD-specific ENSM (color of circle). The false discovery rate–corrected significant GO-terms are grouped into 3 categories: BP, CC, and MF. The right panel shows the subset of genes having an AD-specific ENSM that are annotated with the enriched GO terms, red squares, while a blue square indicates that the gene does not have that annotation. Those genes that are not annotated with any of these GO-terms are not included in this panel. Abbreviations: AD = Alzheimer disease; BP = biological process; CC = cellular component; ENSMs = excitatory neuron-specific somatic mutations; GO = gene ontology; MF = molecular function.

processes of the found ENSMs might be part of the normal aging process in excitatory neurons.[29] A previous study using bulk exome sequencing also found an abundance of the SBS5 signature in aged brain tissues.[10]

When studying the count of somatic mutation in our analyses, we found only a slight increase with age (β = 0.15, Figure 3A) that was not statistically significant ($p = 0.12$). Similar results were observed when performing the same analysis in AD samples and ND individuals separately (eFigure 6, links.lww.com/NXG/A593). We should note that the number of samples is relatively low and represent a relatively narrow age range (from 74 to 90 year of age). Moreover, participants with an age older than 90 years were all censored by age 90 years, which could also influence the significance of the age trend. A significant trend is observed when we exclude individuals at age 90 years from the regression (β = 0.37, $p = 0.005$; eFigure 7, links.lww.com/NXG/A593).

## *RBFOX1* and *KCNIP4* Harbor Age-Associating ENSMs

Because several detected ENSMs are being detected in multiple individual genomes (eFigure 4, links.lww.com/NXG/A593), we next tested the association of age with somatic mutation prevalence for each site *individually* using a logistic regression (Methods). We added AD status as an explanatory term and excluded the sample with other primary cause of dementia (Methods) from this analysis. Two sites (16:6,899,517 (*RBFOX1*), $p = 0.04$; 4:21,788,463 (*KCNIP4*), $p < 0.05$) are found to have significantly more mutations in older individuals. The age distributions in mutated and unmutated samples for these 2 sites are shown in Figure 4. Some caution should be treated when interpreting this plot for individuals older than 90 years because these are all mapped to 90-year-olds. To assess the effect due to censoring on age, we performed a sensitivity analysis by removing all samples with an age 90 years or older. The results indicated stronger signals for these 2 sites (16:6,899,517 [*RBFOX1*], $p = 0.02$; 4:21,788,463 [*KCNIP4*], $p = 0.03$; eFigure 8, links.lww.com/NXG/A593).

## ENSM Sites in *KCNQ5* and *DCLK1* Associate With AD Status

Genes that were enriched with somatic mutations in AD samples might have a higher possibility to be associated with AD. We found 53 ENSM sites that were detected only in AD samples. This prompted the question whether the number of ENSMs associate with AD status. A Wilcoxon rank sum test indicated that there was no significant difference ($p = 0.71$) in the average count of ENSMs between AD samples and non-demented controls (Figure 3B). This finding is in line with a

previous report[10,28,30] that indicated that somatic mutations are associated with AD in certain patterns, but not by amount.

Next, we examined whether the occurrence of an ENSM is overrepresented within AD samples. A Fisher exact test that identifies sites that have a higher odds ratio to detect a somatic mutation in AD samples (Methods) yielded 2 sites with significant odds ratios. These sites are mapped to 2 genes (6:73,374,221 [*KCNQ5*], $p$ = 0.01 and 13:36,667,102 [*DCLK1*], $p$ = 0.02).

### Genes Harboring AD-Specific ENSMs Do Relate to Alzheimer or Processes Involved in Alzheimer

The 53 AD-specific ENSM sites map to 42 genes. When we exclude genes for which also an ENSM occurs in an ND individual (n = 10), we end up with 32 genes that have ENSMs only seen in AD samples (eAppendix 2, links.lww.com/NXG/A592). Among these 32 genes, there are several well-known AD-associated genes, such as *SLC30A3*, *TTL*, and *CTSB*, which thus harbor somatic mutations unique for AD.

Together with the 2 genes for which AD samples had a higher occurrence of ENSMs (*KCNQ5* and *DCLK1*), we conducted a GO-term analysis to investigate the biological pathways that may be involved (Methods). The most enriched biological process is "vocalization behavior" (FDR <0.001). In addition, "intraspecies interaction between organisms" is found to be significant (FDR <0.04). Detected genes with these functions are *DLG4*, *CNTNAP2*, and *NRXN3* (Figure 5). Our results also identified a group of genes (*CACNA1B*, *CNTNAP2*, *DLG4*, *KCNQ3*, and *KCNQ5*) enriched with the GO-term "ion channel complex" (FDR <0.03). *KCNQ* genes encode 5 members of the $K_v7$ family of $K^+$ channel subunits ($K_v7.1$–$7.5$). Four of these ($K_v7.2$–$7.5$) are expressed in the nervous system.[31] Concerning AD-related neuropathology, a link between Aβ accumulation and $K_v7$ channels has been reported by some studies.[32,33]

## Discussion

Late-onset Alzheimer disease, whose incidence increases with age, is often referred to as an age-related disease. Although the accumulation of Aβ peptides and phosphorylated tau proteins are the main neuropathologic characteristics of AD, they fail to fully explain the molecular pathogenesis. As such, a cell-level investigation might be necessary to study the underlying pathogenic mechanism. In this study, we identified somatic mutations using public data collected from 76 ROSMAP donors and investigated their associations with AD and aging.

Although scRNA-seq data are normally used for expression-based analyses, our results have shown that scRNAseq data can be used for the detection of somatic mutations at a cell-type specific level. As long as RNA sequences align correctly to a reference genome, the pipeline that was used for variant calling can be used for both bulk RNA-seq and scRNAseq data.[34] However, calling variants for each cell separately is not efficient, experiences low coverage, and each cell is likely to have a unique set of identified variants. For this reason, we aggregated cells per individual and per cell type, generating cell type–specific pseudobulk data. An exploratory run of this workflow revealed that we were able to confidently detect somatic mutations only for excitatory neuron because this was the most abundant cell type in the scRNA-seq data and thus resulting in sufficient read coverage. Hence, it is imperative to have a sufficient amount of cells or relatively deep sequencing to reliably detect somatic mutations from scRNA-seq data.

Our analysis showed that the prevalence of somatic mutations in the *KCNIP4* and *RBFOX1* genes are associated with increasing age (when corrected for AD status). *KCNIP4* encodes a member of the family of voltage-gated potassium ($K^+$) channel–interacting proteins (KCNIPs), which suggests altered ion transports/channels may be associated with the aging process.[35] *RBFOX1* is a neuron-specific splicing factor predicted to regulate neuronal splicing networks clinically implicated in neurodevelopmental disorders.[36,37] The increased somatic mutations in *RBFOX1* with age indicates neurodevelopmental disorders may also associate with human brain aging.

We detected the occurrence of somatic mutations within some well-known AD-associated genes, such as *SLC30A3*, *TTL*, and *CTSB*. *SLC30A3* is known to be downregulated in the prefrontal cortex of patients with AD.[38] *SLC30A3* is assumed to play a protective role against endoplasmic reticulum stress, which has been thought to be involved to neurodegenerative diseases such as AD.[39] *TTL* is a cytosolic enzyme involved in the post-translational modification of alpha-tubulin.[40] A previous study found that levels of *TTL* were decreased in lysates from AD brains compared with age-matched controls and that, by contrast, D2 tubulin was significantly higher in the AD brains, indicating that loss of *TTL* and accompanying accumulation of D2 tubulin are hallmarks of both sporadic and familial AD.[41] Gene *CSTB* encodes cystatin B (CSTB), an endogenous inhibitor of cystine proteases.[42] Human CSTB has been proposed to be a partner of Aβ and colocalizes with intracellular inclusions of Aβ in cultured cells.[43] Protein levels of CSTB have been also reported to increase in the brains of patients with AD.[44] Apart from these well-known AD-associated genes, we also identified that the *DCLK1* gene harbored more somatic mutations in patients with AD. A study reported that *DCLK1*, which has both microtubule-polymerizing activity and protein kinase activity, phosphorylates *MAP7D1* on Ser 315 to facilitate the axon elongation of cortical neurons.[45] These observations suggest that somatic mutations may initiate or are involved in the AD process in many ways.

Advance AD-related dementia is often accompanied with language problems, behavioral issues, and cognitive decline.[8] Our results identified AD-associated somatic mutations in the genes *CNTNAP2*, *DLG4*, and *NRXN3*, which are involved in, among other processes, vocalization behavior and intraspecies interaction between organisms. These results may indicate that AD-related speech or language problems and withdrawal from social activities

might be associated with somatic mutations in excitatory neurons. In addition, we identified AD-associated somatic mutations in *CACNA1B*, *CNTNAP2*, *DLG4*, *KCNQ3*, and *KCNQ5*, which are all ion channels or involved with ion channels. Previous studies have reported on the possible role of altered neuronal excitability, controlled by different ion channels and their associated proteins, occurring early during AD pathogenesis.[46,47] Specifically, $K^+$ channels, which are the most numerous and diverse channels present in the mammalian brain, may partly explain this alteration in neuronal excitability.[48] In addition, a dysfunction of $K^+$ channels has been observed in fibroblasts[49] and platelets[44] of patients with AD. In addition, Aβ has been demonstrated to not only be involved in the AD pathogenesis but also modulate $K^+$ channel activities[50] and may have a physiologic role in controlling neuronal excitability.[51] Somatic mutations involved in $K^+$ channels were detected to associate with both AD and age, indicating the existence of common processes behind neurodegenerative disease and aging. It also seems that $K^+$ channels are naturally subjected to oxidation by reactive oxygen species (ROS) in both aging and neurodegenerative disease, which are characterized by high levels of ROS.[52]

Calling variants and detecting somatic mutations from public scRNA-seq data expand the use and scope of scRNA-seq data and may provide new insight into postzygotic genetic change at a cell type–specific level. The use of a single cell type (excitatory neurons) and the minimal read coverage requirement minimized biases driven by gene-specific expression. However, some limitations can also not be ignored. First, the workflow is relatively complex, and results are sensitive to the chosen settings of the parameters. Consequently, quality control was highly critical for this study. Nevertheless, we would like to stress the value of further validation of the proposed workflow, e.g., by validating candidate ENSMs using targeted amplicon sequencing in excitatory neurons. Besides these technical aspects, RNA editing events and transcription errors that happen in RNA sequences might also be identified as somatic mutations using this workflow, which may explain the recurrent mutations that we identified. However, the association between this type of mutation and AD or aging could also be interesting.[53] Another limitation of this study is the relative narrow age range of the included individuals. Moreover, ages older than 90 years were censored to be 90 years. These 2 factors may explain that we found only a relative weak association between age and the accumulation of somatic mutations. On the contrary, the significant trend after removing individuals with an age older than 90 years might also suggest that nonagenarians and centenarians generally have a healthier individual genome. Another limitation of our work is that heterozygous variants from the WGS data were ignored in this study (due to potential ambiguity because of differences in gene expression). Therefore, many potential somatic mutations were excluded from the start. In addition, to reduce the effect of technical noise, we need more than 10% of the reads to support a mutational base, which may exclude the mutations present in just 1 or a few neurons. Finally, because 10× scRNA-seq data were used to detect somatic mutations, only variants located on the DNA that gets transcribed into mRNA were detected.

Our study has explored the feasibility of using scRNA-seq data to generate potential new insights into the association of AD and aging with brain somatic mutagenesis. It should be noted that follow-up studies with larger cohorts are required to validate our findings.


## Acknowledgment

The authors thank and acknowledge all participants and their family members of the ROSMAP study. The results published in this study are in whole or in part based on data obtained from the AD Knowledge Portal (adknowledgeportal.org).

## Study Funding

This research was supported by an NWO Gravitation project: BRAINSCAPES: A Roadmap from Neurogenetics to Neurobiology (NWO: 024.004.012).


## Disclosure

The authors report no disclosures relevant to the manuscript. Full disclosure form information provided by the authors is available with the full text of this article at Neurology.org/NG.


## Publication History

Previously published in medRxiv (doi: https://doi.org/10.1101/2022.05.25.22275538). Received by *Neurology: Genetics* October 24, 2022. Accepted in final form February 3, 2023. Submitted and externally peer reviewed. The handling editor was Associate Editor Suman Jayadev, MD.


## Appendix Authors

| Name | Location | Contribution |
|---|---|---|
| **Meng Zhang, MSc** | Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands; Department of Human Genetics, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands | Drafting/revision of the article for content, including medical writing for content; major role in the acquisition of data; study concept or design; and analysis or interpretation of data |
| **Gerard A. Bouland, MSc** | Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands; Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands | Drafting/revision of the article for content, including medical writing for content; study concept or design; and analysis or interpretation of data |
| **Henne Holstege, PhD** | Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands; Department of Human Genetics, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands | Drafting/revision of the article for content, including medical writing for content |
| **Marcel J.T. Reinders, PhD** | Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands; Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands | Drafting/revision of the article for content, including medical writing for content; study concept or design; and analysis or interpretation of data |

## References


1. Biesecker LG, Spinner NB. A genomic view of mosaicism and human disease. *Nat Rev Genet.* 2013;14(5):307-320.
2. Maynard S, Fang EF, Scheibye-Knudsen M, Croteau DL, Bohr VA. DNA damage, DNA repair, aging, and neurodegeneration. *Cold Spring Harb Perspect Med.* 2015; 5(10):a025130.
3. Wang X, Wang W, Li L, Perry G, Lee Hg, Zhu X. Oxidative stress and mitochondrial dysfunction in Alzheimer's disease. *Biochim Biophys Acta.* 2014;1842(8):1240-1247.
4. Poduri A, Evrony GD, Cai X, Walsh CA. Somatic mutation, genomic variation, and neurological disease. *Science.* 2013;341(6141):1237758.
5. Paquola ACM, Erwin JA, Gage FH. Insights into the role of somatic mosaicism in the brain. *Curr Opin Syst Biol.* 2017;1:90-94.
6. McConnell MJ, Moran JV, Abyzov A, et al. Intersection of diverse neuronal genomes and neuropsychiatric disease: the Brain Somatic Mosaicism Network. *Science.* 2017; 356(6336):eaal1641.
7. Kennedy SR, Loeb LA, Herr AJ. Somatic mutations in aging, cancer and neurodegeneration. *Mech Ageing Dev.* 2012;133(4):118-126.
8. Burns A, Iliffe S. Alzheimer's disease. *BMJ.* 2009;338(feb05 1):b158-b471.
9. Hyman BT, Phelps CH, Beach TG, et al. National Institute on Aging-Alzheimer's Association guidelines for the Neuropathologic Assessment of Alzheimer's disease. *Alzheimer's Dement.* 2012;8(1):1-13.
10. Park JS, Lee JHJ, Jung ES, et al. Brain somatic mutations observed in Alzheimer's disease associated with aging and dysregulation of tau phosphorylation. *Nat Commun.* 2019;10(1):3090.
11. Angerer P, Simon L, Tritschler S, Wolf FA, Fischer D, Theis FJ. Single cells make big data: new challenges and opportunities in transcriptomics. *Curr Opin Syst Biol.* 2017;4:85-91.
12. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med.* 2018;50(8):1-14.
13. Chen G, Ning B, Shi T. Single-cell RNA-seq technologies and related computational data analysis. *Front Genet.* 2019;10(APR):317.
14. NMP, Liu H, Dillard C, et al. Improved SNV discovery in barcode-stratified scRNA-seq alignments. *Genes (Basel).* 2021;12(10):1558.
15. Vu TN, Nguyen HN, Calza S, Kalari KR, Wang L, Pawitan Y. Cell-level somatic mutation detection from single-cell RNA sequencing. *Bioinformatics.* 2019;35(22):4679-4687.
16. Petti AA, Williams SR, Miller CA, et al. A general approach for detecting expressed mutations in AML cells using single cell RNA-sequencing. *Nat Commun.* 2019;10(1): 3660-3716.
17. Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA. Religious orders study and rush memory and aging project. *J Alzheimer's Dis.* 2018;64(s1): S161-S189.
18. De Jager PL, Ma Y, McCabe C, et al. A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Sci Data.* 2018;5(1):180142.
19. Mathys H, Davila-Velderrain J, Peng Z, et al. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature.* 2019;570(7761):332-337.
20. Van der Auwera G, O'Connor B. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media; 2020.
21. Lee J, Lee AJ, Lee JK, et al. Mutalisk: a web-based somatic MUTation AnaLyIS toolKit for genomic, transcriptional and epigenomic signatures. *Nucleic Acids Res.* 2018; 46(W1):W102-W108.
22. Cingolani P, Platts A, Wang LLL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin).* 2012;6(2):80-92.
23. Schneider JA, Arvanitakis Z, Bang W, Bennett DA. Mixed brain pathologies account for most dementia cases in community-dwelling older persons. *Neurology.* 2007; 69(24):2197-2204.
24. Gott JM, Emeson RB. Functions and mechanisms of RNA editing. *Annu Rev Genet.* 2000;34:499-531.
25. Gout JF, Li W, Fritsch C, et al. The landscape of transcription errors in eukaryotic cells. *Sci Adv.* 2017;3(10):e1701484.
26. Traverse CC, Ochman H. Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. *Proc Natl Acad Sci U S A.* 2016;113(12): 3311-3316.
27. Navin NE. Cancer genomics: one cell at a time. *Genome Biol.* 2014;15(8):452.
28. Abascal F, Harvey LMR, Mitchell E, et al. Somatic mutation landscapes at single-molecule resolution. *Nature.* 2021;593(7859):405-410.
29. Lodato MA, Rodin RE, Bohrson CL, et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science.* 2018;359(6375): 555-559.
30. Miller MB, Huang AY, Kim J, et al. Somatic genomic changes in single Alzheimer's disease neurons. *Nature.* 2022;604(7907):714-722.
31. Brown DA, Passmore GM. Neural KCNQ (Kv7) channels. *Br J Pharmacol.* 2009; 156(8):1185-1195.
32. Mayordomo-Cava J, Yajeya J, Navarro-López JD, Jiménez-Díaz L. Amyloid-β(25-35) modulates the expression of GirK and KCNQ channel genes in the hippocampus. *PLoS One.* 2015;10(7):e0134385.
33. Durán-González J, Michi ED, Elorza B, et al. Amyloid β peptides modify the expression of antioxidant repair enzymes and a potassium channel in the septohippocampal system. *Neurobiol Aging.* 2013;34(8):2071-2076.
34. Liu F, Zhang Y, Zhang L, et al. Systematic comparative analysis of single-nucleotide variant detection methods from single-cell RNA sequencing data. *Genome Bio.l* 2019; 20(1):242.
35. Kadish I, Thibault O, Blalock EM, et al. Hippocampal and cognitive aging across the lifespan: a bioenergetic shift precedes and increased cholesterol trafficking parallels memory impairment. *J Neurosci.* 2009;29(6):1805-1816.
36. Fogel BL, Wexler E, Wahnich A, et al. RBFOX1 regulates both splicing and transcriptional networks in human neuronal development. *Hum Mol Genet.* 2012;21(19): 4171-4186.
37. Casanovas S, Schlichtholz L, Mühlbauer S, et al. Rbfox1 is expressed in the mouse brain in the form of multiple transcript variants and contains functional E boxes in its alternative promoters. *Front Mol Neurosci.* 2020;13:66.
38. Whitfield DR, Vallortigara J, Alghamdi A, et al. Assessment of ZnT3 and PSD95 protein levels in Lewy body dementias and Alzheimer's disease: association with cognitive impairment. *Neurobiol Aging.* 2014;35(12):2836-2844.
39. Kurita H, Okuda R, Yokoo K, Inden M, Hozumi I. Protective roles of SLC30A3 against endoplasmic reticulum stress via ERK1/2 activation. *Biochem Biophysical Res Commun.* 2016;479(4):853-859.
40. Lewis SA, Cowan NJ. Tubulin genes: structure, expression, and regulation. In: *Microtubule Proteins*. Ed. Avila J. CRC Press; 2018:37-66.
41. Parato J, Kumar A, Pero ME, et al. The pathogenic role of tubulin tyrosine ligase and D2 tubulin in Alzheimer's disease. *Alzheimer's Dement J Alzheimer's Assoc.* 2021;17(53):e056351.
42. Turk V, Bode W. The cystatins: protein inhibitors of cysteine proteinases. *FEBS Lett.* 1991;285(2):213-219.
43. Škerget K, Taler-Verčič A, Bavdek A, et al. Interaction between oligomers of stefin B and amyloid-β in vitro and in cells. *J Biol Chem.* 2010;285(5):3201-3210.
44. De Silva HA, Aronson JK, Grahame-Smith DG, Jobst KA, Smith AD. Abnormal function of potassium channels in platelets of patients with Alzheimer's disease. *Lancet.* 1998;352(9140):1590-1593.
45. Koizumi H, Fujioka H, Togashi K, et al. DCLK1 phosphorylates the microtubule-associated protein MAP7D1 to promote axon elongation in cortical neurons. *Dev Neurobiol.* 2017;77(4):493-510.
46. Palop JJ, Chin J, Roberson ED, et al. Aberrant excitatory neuronal activity and compensatory remodeling of inhibitory hippocampal circuits in mouse models of Alzheimer's disease. *Neuron.* 2007;55(5):697-711.
47. Frazzini V, Guarnieri S, Bomba M, et al. Altered Kv2.1 functioning promotes increased excitability in hippocampal neurons of an Alzheimer's disease mouse model. *Cell Death Dis.* 2016;7(2):e2100.
48. Zaydman MA, Silva JR, Cui J. Ion Channel associated diseases: overview of molecular mechanisms. *Chem Rev.* 2012;112(12):6319-6333.
49. Etcheberrigaray R, Ito E, Oka K, Tofel-Grehl B, Gibson GE, Alkon DL. Potassium channel dysfunction in fibroblasts identifies patients with Alzheimer disease. *Proc Natl Acad Sci.* 1993;90(17):8209-8213.
50. Plant LD, Webster NJ, Boyle JP, et al. Amyloid β peptide as a physiological modulator of neuronal 'A'-type K+ current. *Neurobiol Aging.* 2006;27(11):1673-1683.
51. Ramsden M, Henderson Z, Pearson HA. Modulation of Ca2+ channel currents in primary cultures of rat cortical neurones by amyloid β protein (1-40) is dependent on solubility status. *Brain Res.* 2002;956(2):254-261.
52. Sesti F. Oxidation of K+ channels in aging and neurodegeneration. *Aging Dis.* 2016; 7(2):130-135.
53. Anagnostou ME, Chung C, McGann E, et al. Transcription errors in aging and disease. *Translational Med Aging.* 2021;5:31-38.