# Computational Approaches to Deciphering the Molecular and Cellular Heterogeneity of Alzheimer's Disease

Bouland, G.A.

**DOI**
10.4233/uuid:871ba6f5-63bc-4423-a44b-53a03af7ddec

**Publication date**
2025

**Document Version**
Final published version

**Citation (APA)**
Bouland, G. A. (2025). *Computational Approaches to Deciphering the Molecular and Cellular Heterogeneity of Alzheimer's Disease.* [Dissertation (TU Delft), Delft University of Technology].
https://doi.org/10.4233/uuid:871ba6f5-63bc-4423-a44b-53a03af7ddec

# Computational Approaches to Deciphering the Molecular and Cellular Heterogeneity of Alzheimer's Disease

Gerard Albert Bouland

**Computational Approaches to Deciphering the Molecular and Cellular Heterogeneity of Alzheimer's Disease**

**Proefschrift**

ter verkrijging van de graad van doctor aan de Technische

Universiteit Delft,

op gezag van de Rector Magnificus, Prof.dr.ir. T.H.J.J. van der Hagen,

voorzitter van het College voor Promoties,

in het openbaar te verdedigen op

Maandag, 10 maart 2025 om 12:30 uur

Door

**Gerard Albert BOULAND**

Master of Science in Computer Science, Universiteit Leiden, Nederland

geboren te Haarlem, Nederland

Dit proefschrift is goedgekeurd door de promotoren.

Samenstelling promotiecommissie bestaat uit:

| | |
|---|---|
| Rector magnificus, | voorzitter |
| Prof. dr. Ir. M.J.T. Reinders, | Technische Universiteit Delft, promotor |
| Dr. A. Mahfouz | Leids Universitair Medisch Centrum, copromotor |

Onafhankelijke leden:

| | |
|---|---|
| Prof.dr. L.F.A. Wessels | Technische Universiteit Delft, Nederlands Kanker Instituut |
| Prof.dr. M. van der Wiel | Amsterdam Universitair Medische Centra |
| Prof.dr. B.J.L. Eggen | Universitair Medisch Centrum Groningen |
| Dr. M.G.P. van der Wijst | Universitair Medisch Centrum Groningen |
| Dr. S. van der Sluis | Vrije Universiteit Amsterdam |

Reserve lid:

| | |
|---|---|
| Prof.dr.Ir. B.P.F. Lelieveldt | Technische Universiteit Delft, Leids Universitair Medische Centrum |

Chapter 6 of this dissertation was prepared with significant contributions from and in close collaboration with Prof. Dr. Manolis Kellis (Massachusetts Institute of Technology).

Printed by: Proefschrift Specialist

Front & Back cover: Generated with Dall-E & Adobe Firefly

ISBN: 978-94-93431-10-2

# Contents

# Summary

Alzheimer's disease is a neurodegenerative disease that progressively impairs cognitive functions, ultimately leading to death. As the global population continues to age, the prevalence of Alzheimer is on the rise, making it a significant public health concern. Given its impact, researchers are extensively studying the disease with the aim to slow down its progression, preventing its onset, and eventually discovering a cure.

Genetics significantly influences the risk of developing Alzheimer's disease, with an estimated 60-80% of this risk being inherited. To uncover the genetic basis of Alzheimer's, researchers use genome-wide association studies (GWAS) to identify specific genetic variants, known as single nucleotide polymorphisms (SNPs), that are more prevalent in individuals with Alzheimer compared to healthy individuals. Many Alzheimer associated SNPs are common, exhibit small effect sizes, and are primarily located in non-coding and intergenic regions of the genome. These SNPs act as markers for haplotypes (genetic regions, typically around 300 kilobases in length) that are passed down through generations. Such haplotypes often include at least one genetic factor that contributes to disease risk, potentially by modulating the expression of one or more genes. Although GWAS-identified SNPs are unlikely to be directly causative, they are frequently in partial linkage with the true causal variants. Given that many SNPs reside in non-coding regions, which do not directly change protein-coding sequences, researchers aim to understand their functional significance by correlating these genetic variants with changes in mRNA expression and other molecular data such as protein levels. Investigating the biological processes involving these genes and proteins may provide insights into the mechanisms underlying Alzheimer's disease and ultimately lead to novel therapeutic interventions.

Recent advances in single-cell RNA sequencing (scRNAseq) have provided important insights into Alzheimer's disease by enabling scientists to analyze gene expression at the level of individual cells, offering a detailed view of cellular activity in both healthy and diseased brains. Alzheimer's disease is characterized by considerable heterogeneity, with individuals displaying diverse phenotypic manifestations. Moreover, Alzheimer does not affect all brain regions or cell types uniformly; it progresses in stages, impacting different regions and cell types at varying times. The heterogeneity and progressive nature of Alzheimer's disease pose significant challenges for research, as varying levels of pathology can be present across different patients and even within different regions of the same patient's brain. Despite the availability of extensive phenotypic and scRNAseq data, methods to effectively utilize this data to study variations within and between individuals have yet to be developed.

In this thesis, we introduce analytical approaches to investigate the downstream biological effects GWAS-identified SNPs. We present an approach that utilizes the principal axis of variance among a set of genes linked by a common biological component to create a single variable that is subsequently correlated with GWAS-identified SNPs. Additionally, we designed a gene prioritization strategy that emphasizes genes located near GWAS-identified SNPs, ranking them based on the extent of significant changes in gene-gene correlation patterns. Furthermore,

we identified differential correlation quantitative trait loci (QTLs), defined as GWAS-identified SNPs associated with alterations in protein-protein correlation networks. These approaches provide additional context to the potential downstream consequences of AD-SNPs and offer deeper insights into the mechanisms underlying Alzheimer's disease.

Further, we introduce a novel method that accounts for the heterogeneity of Alzheimer, both between different brain cell types within the same individual and between individuals, providing a more comprehensive understanding of the disease's complexity.

In addition to the challenges associated with using scRNAseq to study Alzheimer, there are technical challenges inherent to the technique itself, particularly the issue of data sparsity. The amount of messenger RNA present in a single cell is much lower than in bulk tissue samples, leading to many genes appearing as if they are not expressed at all. This sparsity can be misleading, and researchers have adopted two primary strategies to address it: imputing the zero measurements or interpreting the absence of gene expression as biologically meaningful. This thesis advocates for the latter approach, emphasizing the significance of these zero measurements in understanding cellular biology. We demonstrate that by representing gene expression in single-cell data as binary (expressed or not expressed), various downstream analytical tasks can be performed with the same effectiveness as when using count-based data.

In summary, the contributions within this thesis advance Alzheimer's research by introducing new computational tools and methods to better understand the genetics of the disease and cellular mechanisms. Additionally, showing that single-cell gene expression can be effectively analyzed in a binary format (expressed or not) simplifies genomic data analysis, making it more accessible, efficient, and applicable to a range of diseases and conditions.

# Samenvatting

Ziekte van Alzheimer is een neurodegeneratieve aandoening die geleidelijk de cognitieve functies aantast en uiteindelijk tot de dood leidt. Naarmate de wereldbevolking vergrijst, neemt de prevalentie van Alzheimer toe, wat het een belangrijk probleem voor de volksgezondheid maakt. Vanwege deze gevolgen bestuderen onderzoekers de ziekte intensief met als doel de progressie ervan te vertragen, het ontstaan ervan te voorkomen en een geneesmiddel te vinden.

Genetische factoren spelen een grote rol bij het risico op het ontwikkelen van de ziekte van Alzheimer, waarbij naar schatting 60-80% van dit risico erfelijk is. Om de genetische basis van Alzheimer te achterhalen, maken onderzoekers gebruik van genome-wide associatiestudies (GWAS) om specifieke genetische varianten, bekend als single nucleotide polymorfisme (SNP's), te identificeren die vaker voorkomen bij mensen met Alzheimer dan bij gezonde personen. Veel van deze Alzheimer-geassocieerde SNP's zijn algemeen, hebben een kleine effectgrootte en bevinden zich hoofdzakelijk in niet-coderende gebieden van het genoom en in gebieden tussen de genen. Deze SNP's fungeren als indicatoren voor haplotypen (genetische segmenten, doorgaans ongeveer 300 kilobasen lang) die van generatie op generatie worden doorgegeven. Haplotypen bevatten vaak ten minste één genetische factor die bijdraagt aan het ziekterisico, mogelijk door de expressie van een of meer genen te beïnvloeden. Hoewel door GWAS geïdentificeerde SNP's waarschijnlijk niet direct de oorzaak zijn, blijken ze vaak in gedeeltelijke koppeling te staan met de echte causale varianten. Omdat veel SNP's in niet-coderende gebieden liggen, die geen directe veranderingen in eiwit coderende sequenties veroorzaken, proberen onderzoekers de functionele betekenis te achterhalen door deze genetische varianten te koppelen aan veranderingen in mRNA-expressie en andere moleculaire data, zoals eiwit data. Het bestuderen van de biologische processen die bij deze genen en eiwitten betrokken zijn, kan inzicht bieden in de onderliggende mechanismen van de ziekte van Alzheimer en uiteindelijk leiden tot nieuwe therapeutische interventies.

Recente ontwikkelingen in single-cell RNA-sequencing (scRNAseq) hebben belangrijke inzichten opgeleverd in de ziekte van Alzheimer, doordat wetenschappers nu de genexpressie op het niveau van individuele cellen kunnen onderzoeken. Dit biedt een gedetailleerd beeld van de cellulaire activiteit in zowel gezonde als zieke hersenen. De ziekte van Alzheimer wordt gekenmerkt door aanzienlijke heterogeniteit: patiënten kunnen sterk uiteenlopende fenotypische kenmerken vertonen. Bovendien tast Alzheimer niet alle hersengebieden of celtypen op gelijke wijze aan; de ziekte vordert in fasen en beïnvloedt verschillende gebieden en celtypen op verschillende tijdstippen. Deze heterogeniteit en het progressieve karakter van de ziekte maken onderzoek lastig, omdat verschillende voortgang niveaus van pathologie kunnen voorkomen bij verschillende patiënten en zelfs binnen verschillende hersengebieden van één patiënt. Ondanks de beschikbaarheid van uitgebreide fenotypische en scRNAseq-data zijn er nog geen methoden ontwikkeld om deze data effectief te benutten voor het bestuderen van variaties binnen en tussen individuen.

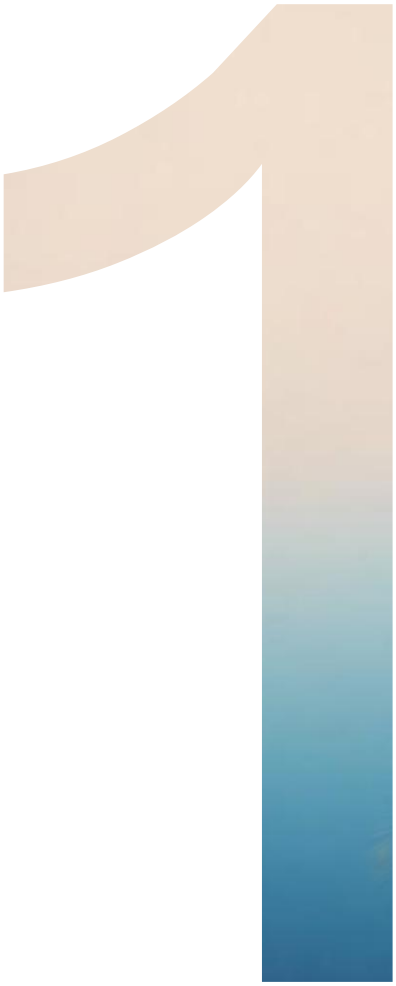
In dit proefschrift introduceren we verschillende analytische methodes om de downstream biologische effecten van door GWAS geïdentificeerde SNP's te

onderzoeken. We presenteren een methode die gebruikmaakt van de as van de meeste variatie binnen een reeks genen, welke verbonden zijn via een gemeenschappelijke biologische component. Hiermee creëren we een enkele variabele die vervolgens wordt gecorreleerd met deze door GWAS geïdentificeerde SNP's. Daarnaast hebben we een strategie ontwikkeld om genen te prioriteren die zich in de buurt van zulke SNP's bevinden; we rangschikken deze genen op basis van de mate waarin er significante veranderingen optreden in hun correlatiepatronen met andere genen. Verder hebben we differentiële correlatie-quantitative trait loci (dc-QTL's) gedefinieerd, oftewel door GWAS geïdentificeerde SNP's die gerelateerd zijn aan veranderingen in eiwit-eiwitcorrelatienetwerken. Deze strategieën bieden extra context voor de mogelijke downstream gevolgen van Alzheimer-SNP's en geven dieper inzicht in de mechanismen achter de ziekte van Alzheimer.

Bovendien introduceren we een nieuwe methode die rekening houdt met de heterogeniteit van Alzheimer, zowel tussen verschillende hersenceltypen binnen één individu als tussen meerdere individuen. Deze aanpak biedt een omvattender beeld van de complexiteit van de ziekte.

Naast de uitdagingen die gepaard gaan met het gebruik van scRNAseq om Alzheimer te onderzoeken, zijn er ook technische obstakels die inherent zijn aan deze technologie, met name het probleem van de grote hoeveelheden nul observaties. De hoeveelheid messenger-RNA (mRNA) in een enkele cel is aanzienlijk lag  er dan in een stukje weefsel, waardoor veel genen ten onrechte lijken te ontbreken in een enkele cel. Dit verschijnsel kan misleidend zijn, en onderzoekers hanteren twee hoofdstrategieën om hiermee om te gaan: het imputeren van de nulmetingen of het interpreteren van de afwezigheid van genexpressie als biologisch betekenisvol. In dit proefschrift pleiten we voor de laatstgenoemde benadering en benadrukken we het belang van deze nulmetingen voor het begrijpen van cellulaire processen. We laten zien dat, door genexpressie in single-cel data binair (wel of niet uitgedrukt) weer te geven, verschillende downstream analysetaken met dezelfde effectiviteit kunnen worden uitgevoerd als bij gebruik van op count gebaseerde data.

Samenvattend leveren de in dit proefschrift een bijdrage aan het onderzoek naar de ziekte van Alzheimer door het introduceren van nieuwe computationele methoden om de genetica van de ziekte en de onderliggende cellulaire mechanismen beter te begrijpen. Bovendien laat we zien dat single-cell genexpressie effectief binair kan worden geanalyseerd, wat de data-analyse vereenvoudigt.



# Introduction

## 1.1. Alzheimer's Disease

Alzheimer's disease (AD) is a neurodegenerative disorder characterized by the decline of cognitive abilities and eventually death[1]. One of the main risk factor for AD is aging[2]. As the global population is increasingly reaching older age, the incidence of AD also continues to rise. Given the devastating nature of AD, extensive research efforts are underway to deepen our understanding of the disease, with the aim of slowing its progression, preventing it, and ultimately finding a cure.

A major portion of AD research focuses on the etiological processes that contribute to AD. Researchers aim to identify the biological processes that are disrupted and could lead to the development and progression of AD. Several hypotheses have emerged in this area, each proposing different mechanisms that may underlie the disease (**Fig. 1**).

One leading theory is the amyloid-beta (Aβ) cascade hypothesis[3], which suggests that the accumulation of amyloid-beta peptides in the brain initiates a cascade of events leading to neurodegeneration. According to this hypothesis, these peptides aggregate to form plaques that disrupt cell function and trigger a series of pathological processes, including inflammation and oxidative stress.

The tau hypothesis[4] centers on the role of tau proteins, which are normally involved in stabilizing microtubules in neurons. In AD, tau proteins become abnormally phosphorylated, leading to the formation of neurofibrillary tangles inside neurons. These tangles disrupt the normal functioning of neurons and contribute to cell death and cognitive decline.

The inflammation hypothesis[5] posits that chronic inflammation in the brain plays a critical role in AD. Microglia, the brain's resident immune cells, become activated in response to amyloid plaques and other factors, leading to a sustained inflammatory response. This inflammation can exacerbate neuronal damage and further drive the progression of the disease.

The oxidative stress hypothesis[6] suggests that an imbalance between the production of reactive oxygen species (ROS) and the brain's ability to detoxify these harmful molecules leads to oxidative damage. Neurons are particularly vulnerable to oxidative stress[7], which can damage cellular components such as DNA, proteins, and lipids, contributing to neurodegeneration. It is increasingly recognized that AD likely results from a combination of these pathological processes rather than a single cause.

Heritability studies suggest that approximately 60-80% of the risk for developing AD can be attributed to genetic factors. Hence understanding genetic variants associated with these pathological processes requires a integrative approach to analyzing genetic and molecular data, which might result in a better understanding of the disease's mechanisms and point towards potential therapeutic targets.

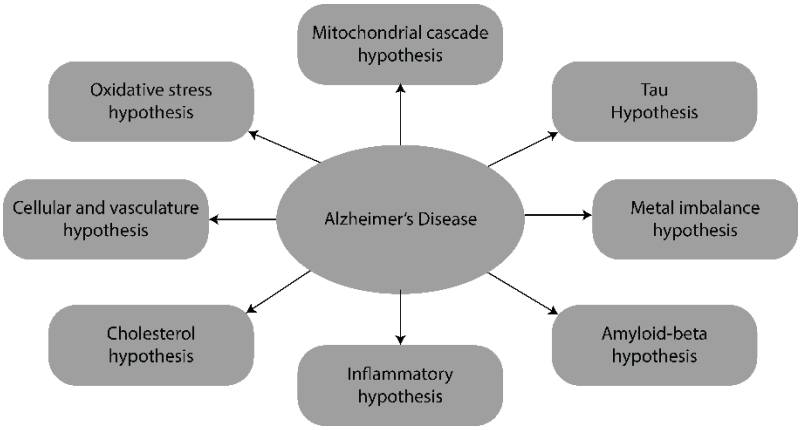



**Figure 1:** Alzheimer's disease and eight leading hypotheses (Figure adapted from Hroudová et al[8])

# 1.2. Genetics of Alzheimer's Disease

Although individual human genomes are approximately 99.9% similar, variation exists at about 88 million genomic positions where variations are observed with a frequency of more than 1% among individuals[9]. Such variations, called single nucleotide polymorphism (SNP), can alter the codon sequence within the coding regions of genes (**Fig. 2a**), potentially changing the polypeptide sequence of proteins and thereby affecting protein function, which might lead to diseases. However, most genetic variations are found in non-coding regions[10], making the task of understanding how these variations influence biological processes and contribute to diseases a difficult task. Moreover, due to recombination, many nearby variants tend to co-occur within individuals[11], complicating the identification of which variant or groups of variants (haplotypes) are critical for disease pathogenesis. Additionally, the interaction of distant SNPs, known as epistasis[12], may also play a role in disease susceptibility. Altogether, outlining the complexity of understanding genetics of human diseases.

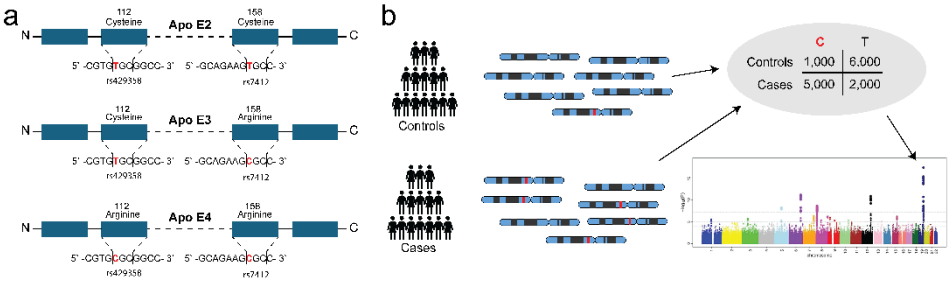



**Figure 2: a)** Schematic representation of the APOE alleles. When individuals have a T allele for both the rs429358 and rs7412 SNPs, a cysteine peptide is encoded into the APOE protein at positions 112 and 158, known as the APOE2 allele. If an individual has a C allele for SNP rs7412, the codon at position 158 changes to encode arginine, resulting in the APOE3 allele. When both rs429358 and rs7412 SNPs have a C allele, both cysteines are replaced by arginine, forming the APOE4 allele. **b)** Schematic representation of a genome wide association study (GWAS). A GWAS begins with the collection of genetic data from two large groups of individuals: controls and cases. For each genetic position where variation exists in the population, it is tested whether variants are enriched or depleted

in the case group compared to the control group as shown in the contingency table where C is the risk variant. This results in an association measure and a p-value, which is shown in the Manhattan plot, where every dot is a variant, the x-axis is the location of that variant on the genome and the y-axis is the -log10 p-value.

## 1.2.2. Genome-wide association studies (GWASs)

SNPs associated with diseases, such as AD, are identified through Genome-wide association studies (GWASs, **Fig. 2b**)[13]. These studies start by collecting and sequencing genomes from a large number of individuals, including both individuals affected by a disease and unaffected controls. Researchers then examine whether SNPs that have variation in the study population are more frequent in those with the disease than in those without. A variety of statistical tests[14] can be used, such as Fisher's exact test, chi-squared tests, and linear and logistic regression analyses, are used to explore the relationships between SNPs and disease.

When tests are performed, it is important to account for the fact that humans are diploid, meaning they carry two copies of each chromosome. As such, the analyses must consider the effects of having one versus two copies of a genetic risk variant, with the assumption that being homozygous (having two identical nucleotides) might influence the disease more than being heterozygous (having two different nucleotides).

Additionally, it is important to realize that not all individuals carrying a genetic risk variant will develop the associated disease. For example, not even every carrier of the APOE4 allele, the largest genetic risk factor for AD, is guaranteed to develop AD. Which is even more true for genetic risk factors with less pronounced effects. Thus, to detect statistically significant associations with relatively subtle effects, larger sample sizes are required. As such, over the years we have seen a large increase in study sizes of GWASs. For instance, one of the earliest GWASs for AD performed, in 2009 included 16,000 participants identified only three statistically significant genetic independent risk variants[15]. While the latest GWAS for AD included almost 800,000 participants and identified 75 statistically significant genetic independent risk variants[16].

In studying the genetics of AD, it is essential to make a distinction between Familial Alzheimer's Disease (FAD) and Late-Onset Alzheimer's Disease (LOAD). FAD is a rare autosomal dominant disorder typically caused by mutations in one of three genes: *APP*, *PSEN1*, or *PSEN2*, leading to the development of Alzheimer's before the age of 65[17]. In this thesis, we focus on LOAD, which is associated with over 75 independent genetic risk variants[16], with the APOE4 allele being the primary genetic risk factor[18].

One of the most well-studied genetic risk factors associated with AD is the APOE4 allele (**Fig. 1a**), which has been shown to significantly increase the risk of developing the disease[18,19]. Individuals who carry one copy of the APOE4 allele have a higher risk compared to those with the more common APOE3 allele, and those with two copies of APOE4 have an even greater risk. This allele is believed to play a role in numerous processes (e.g., lipid metabolism[20], inflammation[5],

oxidative stress[5]) that increase the risk of AD and is linked to pathological hallmarks (tau pathology[21] and amyloid beta accumulation[22]) associated with AD.

In addition to APOE4, there are currently approximately 75 independent genetic variants known to be associated with an increased risk of developing AD[16]. Some of these genetic variants have been linked to genes involved in various biological pathways, including immune response, cell signaling, and lipid transport, highlighting the complex nature of AD. Despite advances in identifying these genetic variants, understanding their role in the pathogenesis of AD remains difficult. The interactions between these genetic factors and environmental influences, as well as their combined effect on the molecular mechanisms leading to AD, are still not understood. Ongoing research aims to unravel these complexities to better understand the disease and eventually develop effective therapeutic strategies.

### 1.2.3. Brain somatic mutations

The genetic variations discussed so far are inherited from either parent. However, some genetic variants, known as somatic mutations, are acquired during an individual's lifetime. These mutations can arise from various causes, including errors during DNA replication in cell division[23], exposure to environmental mutagens[24], and defects in DNA repair mechanisms[25]. Additionally, the accumulation of somatic mutations increases with age, which is noteworthy since aging is a major risk factor for AD. This connection has been explored in previous studies and indeed showed significant associations of somatic mutations with AD and AD related pathologies[26,27]. As such this should be taken into account in genetic research related to AD. While the specific role of these somatic mutations in the disease progression is not well understood, recent studies suggest that they tend to occur in genes involved in the PI3K-AKT, MAPK, and AMPK pathways, which are known to contribute to tau hyperphosphorylation[27].

### 1.2.4. Quantitative trait loci (QTLs)

One approach to understanding how SNPs contribute to diseases involves identifying quantitative trait loci (QTLs[28]). A QTL is a SNP that is significantly associated with a quantitative trait, such as gene expression (eQTL[29]), protein levels (pQTL[30]), metabolite levels (mQTL[31]), lipid levels (lQTL[32]), and other molecular biotypes. Essentially, QTLs are SNPs linked to variations in molecular measurements.

Unlike GWASs that focus on disease status as the outcome variable, QTL analyses use molecular measurements as the outcome. While GWASs test each SNP's association with disease status, resulting in a number of tests equal to the number of SNPs (for example, 6 million SNPs equals 6 million tests), QTL analyses multiply this number by assessing multiple SNPs against multiple biomolecules. For example, analyzing 10,000 genes in an eQTL study would require 60 billion tests. Luckily, it is generally observed that SNPs affecting gene or protein levels are located near each other (<1 megabase pairs)[33], reducing the need to test every gene or protein against all SNPs across the genome and focusing instead on nearby SNPs, also known as cis-regulation, in contrast to

trans-regulation (>1 megabase pair). This localized testing is supported by known biological pathways showing how SNPs can influence gene expression[34,35].

Once QTLs are identified, overlaps with disease-associated SNPs can be examined to determine if a disease-related SNP also affects, for example, the expression of a gene crucial for lipid metabolism. Understanding such links could point to disrupted lipid metabolism as a potential disease mechanism, which would necessitate further functional studies.

Recent advancements in QTL research include single-cell QTLs (sc-QTLs[36]), which analyze SNP associations with gene expression specific to cell types, and co-expression QTLs (co-QTLs[37]), which explore how SNPs relate to the interaction between two genes. This includes investigating whether a specific pair of genes shows a co-expression pattern in individuals with certain genotypes, and how this co-expression might change in a different genetic context.

### 1.2.5. Challenges with QTL analyses

Investigating whether a risk locus functions as an expression quantitative trait locus (eQTL), indicating a link to abnormal messenger RNA (mRNA) expression, is a reasonable approach. However, despite mRNAs encoding proteins, their expression levels often do not correlate with protein levels[38–40]. Consequently, eQTLs frequently do not correspond to protein-QTLs (pQTLs)[41]. Since proteins are the cell's functional units, understanding genetic regulation or deregulation related to risk alleles through their impact on protein expression might be more insightful.

Focusing solely on eQTLs and pQTLs might overlook the downstream effects of causative variants. For instance, a causative missense variant might not affect mRNA levels but could alter the amino acid sequence of a protein, significantly impacting its function and interactions. This can modify various biological pathways, disrupt protein-protein interactions, enzyme activities, and signal transduction pathways, leading to cellular dysfunction and different functional associations between proteins.

Assuming that co-expressed proteins are functionally related, allele-specific protein correlation patterns might indicate unique regulatory states of biological pathways in response to an allele. From this perspective, disease-associated alleles could have downstream functional consequences, detectable by comparing changes in protein abundance co-expression between carriers and non-carriers of risk alleles. This approach could provide deeper insights into the genetic regulation and its impact on cellular functions, aiding in the understanding of disease mechanisms.

## 1.3. Using Single-cell RNAseq Data to Understand Alzheimer's Disease

The introduction of single-cell RNA sequencing (scRNAseq, **Fig. 3a**) has significantly enriched AD research[43,44]. scRNAseq quantifies the abundance of messenger RNA (mRNA) molecules, which are the molecules responsible for

encoding proteins, the functional units of cells. scRNAseq allows to analyze gene expression at the level of individual cells, providing a detailed view of cellular functions and interactions. By comparing RNA expression profiles between healthy individuals and those diagnosed with AD, genes have been identified that may play critical roles in the pathology of the disease[45]. Previously it was only possible to investigate transcriptional differences at a bulk tissue level, which measure average gene expression across all cells in a sample. Bulk RNA sequencing obscures the contributions of specific cell types, making it difficult to discern the roles of different cells in the disease process (**Fig. 3b**). In contrast, scRNAseq enables the identification of cell-type-specific gene expression changes, offering insights into how distinct cell populations, such as neurons and microglia (the brain's resident immune cells), are uniquely affected by AD. This detailed cellular and molecular information is important for understanding the complex interactions that might drive the disease. Despite its promise, scRNAseq is still a relatively new technology and presents several open challenges[46].

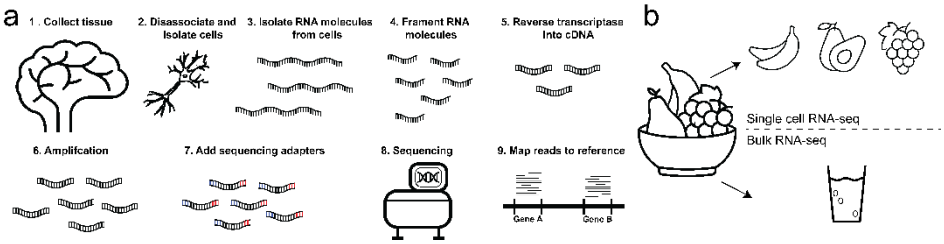



**Figure 3: a)** Workflow of scRNA sequencing (scRNA-seq). **(1)**Tissue samples are collected from the subject, and **(2)** cells are separated from the collected tissue. **(4)** RNA molecules are then extracted from the isolated cells and **(4)** subsequently fragmented into smaller pieces. **(5)** These RNA fragments are reverse transcribed to create complementary DNA (cDNA), which is **(6)** then amplified to produce sufficient quantities for sequencing. **(7)** Sequencing adapters are attached to the amplified cDNA fragments, **(8)** which are then sequenced using a sequencing machine. **(9)** Finally, the sequenced reads are mapped to a reference genome for analysis and identification of gene expression levels. **b)** An illustration of the differences between scRNA-seq and bulk RNA sequencing. The fruit bowl represents a sample containing various distinct cell types, symbolized by different fruits (banana, avocado, grapes). scRNAseq, depicted by the individual fruits, allows for the examination of gene expression at the single-cell level, identifying the specific expression profiles of different cell types within a mixed sample. In contrast, bulk RNAseq, represented by the juice glass (mix of fruits), homogenizes the sample, providing an averaged gene expression profile that masks the unique profiles of individual cell types.

## 1.3.1. Challenges of using scRNAseq data to understand Alzheimer's disease

One major development regarding scRNAseq data is the increased efficiency of sequencing protocols coupled with a reduction in costs. This progress now enables the measurement of cellular profiles at a population scale across multiple conditions. Historically, scRNAseq analyses often focused on measuring many cells from a single individual. Because of this, the primary goal was often to identify various cell types within that individual. With the capability to include multiple individuals in studies, the research paradigm has shifted. Now, researchers can explore differences between individuals, such as comparing groups diagnosed with AD to healthy control groups. However, current analysis methods tend to assume that cells from a given individual uniformly represent the

phenotype of their respective donors, overlooking the variation within an individual's cells. Consequently, the current single-cell analysis paradigm first constructs reference maps of discrete cell types and subtypes. It then identifies differentially expressed genes between “cases” and “controls” for each cell group or subgroup, thereby neglecting the cellular variation within individual donors.

However, as an example, AD pathology does not affect the brain uniformly. For instance, considering Braak staging[47], pathology starts in the lower brainstem and progresses in the end to the neocortex, impacting regions sequentially rather than all at once. Thus, at any time-point, various brain regions of a single individual show differing levels of AD pathology. And this variability in disease manifestations is also present at the cell type level[48].

Some methods[49,50] have been developed to utilize the multi-individual nature of these datasets, allowing for the identification of cell states associated with disease by exploiting phenotypic variation among individuals without requiring predefined cell types. While identifying phenotype-specific cell states provides important insights into the cellular consequences of phenotypes, it overlooks the heterogeneous nature of disease. To address this, it is essential to determine which donors have cells in the respective cell states and which do not, as well as to identify other phenotypic characteristics of donors whose cells exhibit these states. Current methods fall short in investigating the relationship between differences in phenotypic traits and variations in cellular phenotypic manifestations.

## 1.4. Sparsity of single-cell RNAseq data

The amount of mRNA in a single cell is significantly lower than in bulk data[51]. Moreover, transcription is a bursty process; in the same cell mRNA expression can be high at one moment and absent at a later moment[52]. Furthermore, the sequencing process involves sampling[53], which means not all mRNA molecules that are present in the cells are captured during the sequencing process. Next to that, technological advancements now allow for population-scale scRNAseq studies, encompassing data from more than 100 individuals, over 1 million cells, and upwards of 20,000 genes[44]. With this increased capacity, it has been debated whether it is better to measure a larger number of mRNA molecules in fewer cells, or to distribute resources across a larger number of cells with fewer mRNA molecules per cell. It has been found that the latter approach tends to provide a more comprehensive representation of cell biology within the same budget, albeit at the cost of increased data sparsity[54,55].

In summary, scRNAseq data has become sparser, i.e. for more genes no transcripts are read or only a few. There are two main approaches to manage the sparsity observed in scRNAseq data. The first approach considers the zero measurements as missing data, attempting to impute these values[56]. The second approach, views these zero measurements as meaningful biological information[57,58]. However, this raises concerns about potentially discarding valuable information. In this thesis, we advocate for the second approach, emphasizing the biological significance of the zero measurements.

### 1.4.1. Challenges regarding the sparsity of single-cell RNAseq data

The challenge of viewing the zero measurements as biological meaningful information lies in the inadequacy of standard count distribution models, such as Poisson, which worked well for bulk RNA-seq but fail to account for the number of observed zero measurements in scRNAseq data, necessitating the exploration of alternative handling and modeling techniques.

Sarkar and Stephens[53] proposed a perspective on this by proposing to model observed counts through a combination of an expression model and a measurement model. The expression model represents the true biological distribution of a gene, which varies per gene. For example, housekeeping genes might be uniformly distributed across a cell population, and for other genes, due to transcriptional dynamics, mRNA levels might follow a Gaussian distribution within a specific cell populations. The measurement model sees the sequencing process as a series of Bernoulli experiments, resulting in a Poisson distribution. Combining the expression and measurement models yields the count distributions as observed in practice. However, this approach is complex because there is no one-size-fits-all solution, as the distribution differs for each gene and depends heavily on the cell population.

## 1.5. Thesis contributions

The research presented in this thesis addresses several challenges and introduces new methodologies and approaches in the study of scRNAseq, proteomics and genetics data and its application in understanding AD. Firstly, it addresses a major issue with scRNAseq data—its sparsity—by developing a new analytical method for differential gene expression analyses and by outlining the opportunities this increased sparsity presents. Secondly, the thesis improves understanding of how disease-associated genetic risk variants impact mRNAs and proteins, as well as the extent to which mRNAs interact with other mRNAs and proteins with other proteins in relation to these risk variants, and consequently, which biological processes might be disrupted. Moreover, it introduces a new approach to analyzing scRNAseq data by accounting for cellular heterogeneity both within and between individuals who share similar phenotypic characteristics, leading to a better understanding of the heterogeneity of AD. This approach has allowed for the identification of nine distinct cellular components of AD progression, thereby advancing our understanding of the disease's heterogeneity and its underlying mechanisms.

**In Chapter 2**, we present a gene prioritization approach for identifying the genes that are most likely associated with a genetic risk variant and the respective disease. First, we examined differences in correlation patterns of scRNAseq gene expression data between individuals diagnosed with AD and undiagnosed controls. Through this analysis, we identified key "hub" genes that exhibit changes in their correlation with numerous other genes in individuals with AD compared to those without a diagnosis. Given that many of these hub genes are known to regulate gene expression, we hypothesized that they might play a crucial role in the development of AD. This hypothesis was supported by the fact

that many of these hubs are also implicated in the genetic risk factors for AD. Consequently, we identified systematic differences in gene expression associated with AD, which are likely coordinated by these hub genes.

In **Chapter 3** we introduced gene-set-QTLs (GS-QTLs). With GS-QTLs we developed an approach that considers the shared maximum variability of the gene set as a whole, rather than looking at the sum of individual gene associations, thereby association risk variants directly whole sets of genes. Using scRNAseq and genetics data from 666 healthy and Alzheimer's Disease individuals we show that GS-QTLs can identify cell type specific associations between genetic risk variants, biomolecules, and pathways that are missed by conventional methods.

In **Chapter 4,** we performed a large-scale pQTL analysis, identifying many variants associated with the differential abundance of proteins expressed in the Medial Temporal Gyrus. Additionally, we found pairs of proteins that show differential correlation in relation to AD risk variants. These findings offer insights into potential biological mechanisms that are altered by genetic risk variants, thereby enhancing our understanding of the genetic predisposition to AD.

In **Chapter 5** we present a cell type specific somatic mutation identification pipeline using scRNAseq data and whole genome sequencing (WGS) data. Using this pipeline, we identified sites of somatic mutations in the excitatory neuron that were more often mutated in older individuals compared to younger individuals. And we found sites that were associated with AD. Meaning that individuals with AD were more likely to have a somatic mutation at these sites.

In **Chapter 6,** we introduce cell projected phenotypes (CPP). With CPP, we introduce a new approach for analyzing population-scale multi-condition scRNAseq data. Our method uses transcriptional similarities between cells from different individuals, with different phenotypic characteristics, to provide a better understanding of cellular variations between individuals and within a single individual, thereby addressing the complexity of cellular heterogeneity. Our analyses underscore that a) not every cell from an individual exhibits the transcriptional signature of that individual's phenotypic characteristics, b) phenotypes are expressed to varying degrees across different cell types, and c) across individuals with the same phenotypic traits different cell types may be implicated. Although such variability is well-recognized in biological studies, this study is the first to systematically consider these aspects using single-cell RNAseq data, offering new insights into the cellular basis of phenotypic variations.

In **Chapter 7** we present Binary Differential Analysis (BDA), a differential expression analysis method for scRNAseq data relying on binarized scRNAseq data, where every zero remains a zero, and every non-zero value is assigned a one. Through extensive experiments we show that BDA performs similarly as count based differential expression analysis methods, and in some cases even outperforms count based methods. With this research we concluded that the binary patterns of gene expression; measured (1) – not measured(0) is actually biologically meaningful. In **Chapter 8**, we extend the findings from Chapter 7,

finding that counts add very little information on top of the binary pattern of expression and that for various scRNAseq analysis methods the binary pattern is sufficient. As such, we recommended the development of specialized tools for bit-aware implementations of downstream analytical tasks, reducing the amount of required computational resources, thereby enabling the analysis of bigger datasets and a more fine-grained resolution of biological heterogeneity.

Finally, we conclude the thesis in **Chapter 9** with a discussion of our contributions, a discussion on the future of how sparsity could be treated in scRNAseq studies, a discussion on how to analyze genetic risk variants and their downstream consequences and finally a discussion on the implications of CPP on scRNAseq analyses and genetics research.

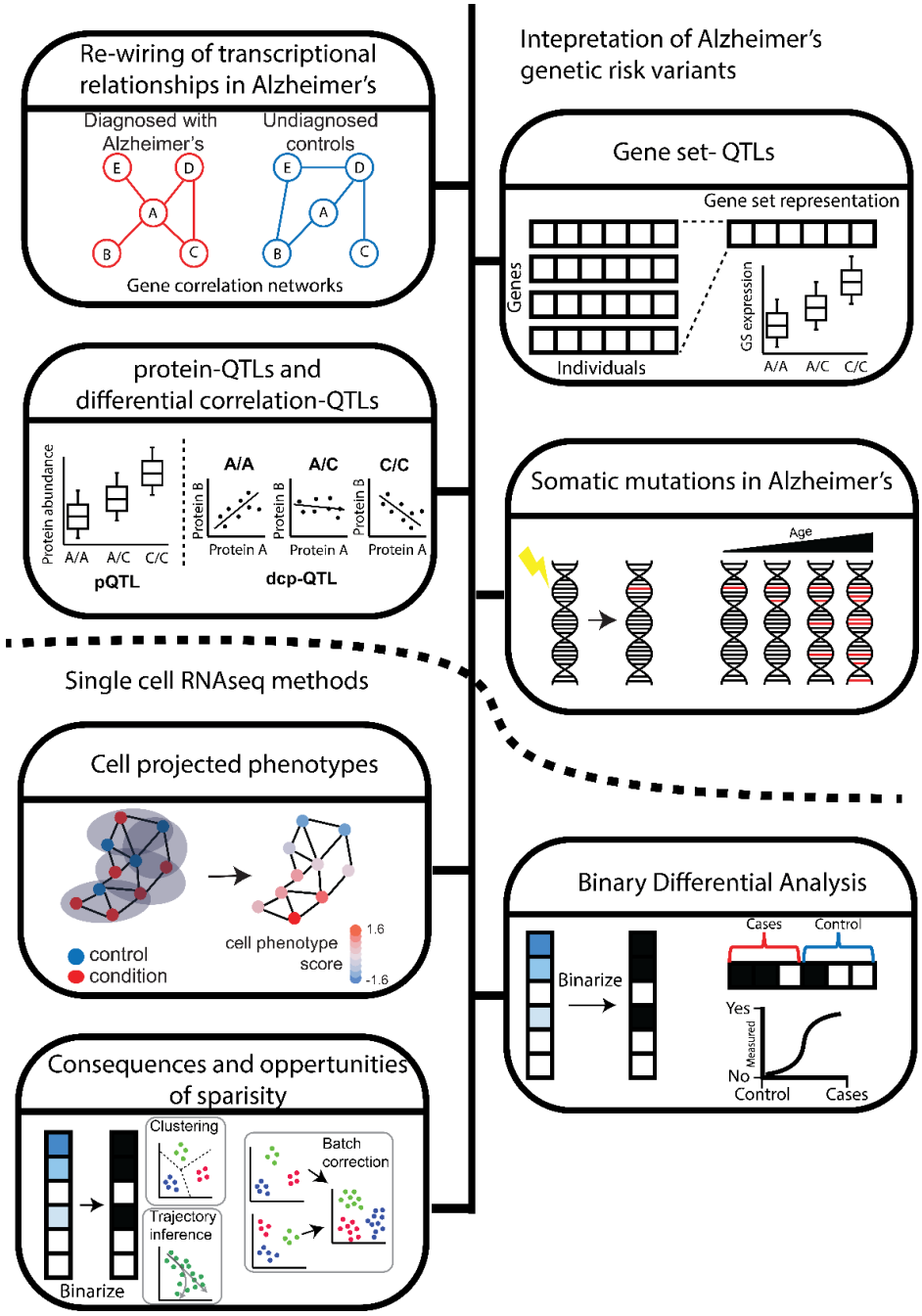



**Figure 4:** Overview of chapter 2-8.

# References


1. Thies, W. & Bleiler, L. 2012 Alzheimer's disease facts and figures. *Alzheimer's Dement.* **8**, 131–168 (2012).

2. Guerreiro, R. & Bras, J. The age factor in Alzheimer's disease. *Genome Med.* **7**, 1–3 (2015).

3. Ricciarelli, R. & Fedele, E. The Amyloid Cascade Hypothesis in Alzheimer's Disease: It's Time to Change Our Mind. *Curr. Neuropharmacol.* **15**, 926 (2017).

4. Mohandas, E., Rajmohan, V. & Raghunath, B. Neurobiology of Alzheimer's disease. *Indian J. Psychiatry* **51**, 55 (2009).

5. Caberlotto, L., Marchetti, L., Lauria, M., Scotti, M. & Parolo, S. Integration of transcriptomic and genomic data suggests candidate mechanisms for APOE4-mediated pathogenic action in Alzheimer's disease OPEN. *Nat. Publ. Gr.* (2016) doi:10.1038/srep32583.

6. Markesbery, W. R. Oxidative stress hypothesis in Alzheimer's disease. *Free Radic. Biol. Med.* **23**, 134–147 (1997).

7. Chen, X., Guo, C. & Kong, J. Oxidative stress in neurodegenerative diseases. *Neural Regen. Res.* **7**, 376 (2012).

8. Hroudová, J., Singh, N., Fišar, Z. & Ghosh, K. K. Progress in drug development for Alzheimer's disease: An overview in relation to mitochondrial energy metabolism. *Eur. J. Med. Chem.* **121**, 774–784 (2016).

9. Auton, A. *et al.* A global reference for human genetic variation. *Nature* vol. 526 68–74 at https://doi.org/10.1038/nature15393 (2015).

10. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**, R102 (2015).

11. Slatkin, M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet. 2008 96* **9**, 477–485 (2008).

12. Wei, W. H., Hemani, G. & Haley, C. S. Detecting epistasis in human complex traits. *Nat. Rev. Genet. 2014 1511* **15**, 722–733 (2014).

13. Uffelmann, E. *et al.* Genome-wide association studies. *Nat. Rev. Methods Prim. 2021 11* **1**, 1–21 (2021).

14. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559 (2007).

15. Harold, D. *et al.* Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat. Genet.* **41**, 1088–1093 (2009).

16. Bellenguez, C. *et al.* New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat. Genet. 2022 544* **54**, 412–436 (2022).

17. Shea, Y. F. *et al.* A systematic review of familial Alzheimer's disease: Differences in presentation of clinical features among three mutated genes and potential ethnic differences. *J. Formos. Med. Assoc.* **115**, 67–75 (2016).

18. Raulin, A. C. *et al.* ApoE in Alzheimer's disease: pathophysiology and therapeutic strategies. *Mol. Neurodegener. 2022 171* **17**, 1–26 (2022).

19. Sando, S. B. *et al.* APOE ε4 lowers age at onset and is a high risk factor for Alzheimer's disease; A case control study from central Norway. (2008) doi:10.1186/1471-2377-8-9.

20. Serrano-Pozo, A. *et al.* Effect of APOE alleles on the glial transcriptome in normal aging and Alzheimer's disease. *Nat. Aging 2021 110* **1**, 919–931 (2021).

21. Strittmatter, W. J. *et al.* Hypothesis: Microtubule Instability and Paired Helical Filament Formation in the Alzheimer Disease Brain Are Related to Apolipoprotein E Genotype. *Exp. Neurol.* **125**, 163–171 (1994).

22. Aleshkov, S., Abraham, C. R. & Zannis, V. I. Interaction of nascent ApoE2, ApoE3, and ApoE4 isoforms expressed in mammalian cells with amyloid peptide beta (1-40). Relevance to Alzheimer's disease. *Biochemistry* **36 34**, 10571–80 (1997).

23. Manders, F., van Boxtel, R. & Middelkamp, S. The Dynamics of Somatic Mutagenesis During Life in Humans. *Front. Aging* **2**, (2021).

24. Wermuth, C. G., Ganellin, C. R., Lindberg, P. & Mitscher, L. A. Chapter 36. Glossary of Terms Used in Medicinal Chemistry (IUPAC Recommendations 1997. *Annu. Rep. Med. Chem.* **33**, 385–395 (1998).

25. Cheong, A. & Nagel, Z. D. Human Variation in DNA Repair, Immune Function, and Cancer Risk. *Front. Immunol.* **13**, (2022).

26. Miller, M. B., Reed, H. C. & Walsh, C. A. Brain Somatic Mutation in Aging and Alzheimer's Disease. *Annu. Rev. Genomics Hum. Genet.* **22**, 239 (2021).

27. Park, J. S. *et al.* Brain somatic mutations observed in Alzheimer's disease associated with aging and dysregulation of tau phosphorylation. *Nat. Commun. 2019 101* **10**, 1–12 (2019).

28. Molecular quantitative trait loci. *Nat. Rev. Methods Prim. 2023 31* **3**, 1–1 (2023).

29. Iacono, G., Massoni-Badosa, R. & Heyn, H. Single-cell transcriptomics unveils gene regulatory network plasticity. *Genome Biol.* **20**, 110 (2019).

30. Yao, C. *et al.* Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat. Commun. 2018 91* **9**, 1–11 (2018).

31. Nicholson, G. *et al.* A genome-wide metabolic QTL analysis in Europeans implicates two loci shaped by recent positive selection. *PLoS Genet.* **7**, (2011).

32. Linke, V. *et al.* A large-scale genome-lipid association map guides lipid identification. *Nat. Metab.* **2**, 1149 (2020).

33. Veyrieras, J. B. *et al.* High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. *PLoS Genet.* **4**, (2008).

34. Corradin, O. & Scacheri, P. C. Enhancer variants: Evaluating functions in common disease. *Genome Med.* **6**, 1–14 (2014).

35. Cortini, R. & Filion, G. J. Theoretical principles of transcription factor traffic on folded chromatin. *Nat. Commun. 2018 91* **9**, 1–10 (2018).

36. Fujita, M. *et al.* Cell subtype-specific effects of genetic variation in the Alzheimer's disease brain. *Nat. Genet. 2024 564* **56**, 605–614 (2024).

37. Van Der Wijst, M. G. P. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493–497 (2018).

38. Guo, Y. *et al.* How is mRNA expression predictive for protein expression? A correlation study on human circulating monocytes. *Acta Biochim. Biophys. Sin. (Shanghai).* **40**, 426–436 (2008).

39. De Sousa Abreu, R., Penalva, L. O., Marcotte, E. M. & Vogel, C. Global signatures of protein and mRNA expression levels. *Molecular BioSystems* vol. 5 1512–1526 at https://doi.org/10.1039/b908315d (2009).

40. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13**, 227–232 (2012).

41. Robins, C. *et al.* Genetic control of the human brain proteome. *Am. J. Hum. Genet.* **108**, 400–410 (2021).

42. Gorbunova, V., Seluanov, A., Mao, Z. & Hine, C. Changes in DNA repair during aging. *Nucleic Acids Res.* **35**, 7466–7474 (2007).

43. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).

44. Mathys, H. *et al.* Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to Alzheimer's disease pathology. *Cell* **186**, 4365-4385.e27 (2023).

45. Sun, N. *et al.* Single-nucleus multiregion transcriptomic analysis of brain vasculature in Alzheimer's disease. *Nat. Neurosci.* **26**, 970–982 (2023).

46. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol. 2020 211* **21**, 1–35 (2020).

47. Braak, H. & Braak, E. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol.* **82**, 239–259 (1991).

48. Tijms, B. M. *et al.* Cerebrospinal fluid proteomics in patients with Alzheimer's disease reveals five molecular subtypes with distinct genetic risk profiles. *Nat. Aging 2024* 1–15 (2024) doi:10.1038/s43587-023-00550-7.

49. Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat. Biotechnol. 2021* 1–9 (2021) doi:10.1038/s41587-021-01033-z.

50. Reshef, Y. A. *et al.* Co-varying neighborhood analysis identifies cell populations associated with phenotypes of interest from single-cell transcriptomics. *Nat. Biotechnol.* **40**, 355–363 (2022).

51. Dzamba, D., Valihrach, L., Kubista, M. & Anderova, M. The correlation between expression profiles measured in single cells and in traditional bulk samples. *Sci. Reports 2016 61* **6**, 1–9 (2016).

52. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science (80-. ).* **297**, 1183–1186 (2002).

53. Sarkar, A. & Stephens, M. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nat. Genet. 2021 536* **53**, 770–777 (2021).

54. Zhang, M. J., Ntranos, V. & Tse, D. Determining sequencing depth in a single-cell RNA-seq experiment. *Nat. Commun. 2020 111* **11**, 1–11 (2020).

55. Schmid, K. T. *et al.* scPower accelerates and optimizes the design of multi-sample single cell transcriptomic studies. *Nat. Commun. 2021 121* **12**, 1–18 (2021).

56. Cheng, Y., Ma, X., Yuan, L., Sun, Z. & Wang, P. Evaluating imputation methods for single-cell RNA-seq data. *BMC Bioinformatics* **24**, 1–24 (2023).

57. Qiu, P. Embracing the dropouts in single-cell RNA-seq analysis. *Nat. Commun.* **11**, 1–9 (2020).

58. Li, R. & Quon, G. ScBFA: Modeling detection patterns to mitigate technical noise in large-scale single-cell genomics data. *Genome Biol.* **20**, 193 (2019).

# Single-cell RNA sequencing data reveals rewiring of transcriptional relationships in Alzheimer's Disease associated with risk variants

Gerard A. Bouland, Kevin I. Marinus, Ronald E. van Kesteren, August B. Smit, Ahmed Mahfouz, and Marcel J.T. Reinders

# Abstract

Understanding how genetic risk variants contribute to Alzheimer's Disease etiology remains a challenge. Single-cell RNA sequencing (scRNAseq) allows for the investigation of cell type specific effects of genomic risk loci on gene expression. Using seven scRNAseq datasets totalling >1.3 million cells, we investigated differential correlation of genes between healthy individuals and individuals diagnosed with Alzheimer's Disease. Using the number of differential correlations of a gene to estimate its involvement and potential impact, we present a prioritization scheme for identifying probable causal genes near genomic risk loci. Besides prioritizing genes, our approach pin-points specific cell types and provides insight into the rewiring of gene-gene relationships associated with Alzheimer's.


# 2.1 Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disease characterized by loss of cognitive functions and autonomy, eventually leading to death[1]. Many hypotheses about the etiology of AD exists, e.g. the amyloid-beta (Aβ) cascade hypothesis, the tau hypothesis, the inflammation hypothesis, the oxidative stress hypothesis and more[2,3], highlighting the complexity of AD. Genome-wide association studies (GWASs) have provided a compendium of genomic loci that are associated with the risk for AD[4–7]. However, understanding how these risk variants contribute to AD etiology remains a challenge. As the number of GWASs is still rising steadily and are increasingly becoming larger in sample size, new genomic risk loci are regularly identified, while studies that generate mechanistic understanding lag behind[8]. Methods such as mendelian randomization[9] and colocalization[10] provide insight in causality but fail to provide insight in downstream molecular consequences. Single-cell genomics has made it possible to investigate genetic regulation in distinct cell types and paves the way to new approaches that will provide a more detailed understanding of cell type specific dysregulation in AD, genetics and downstream consequences.

Additionally, scRNAseq has provided insight into cellular heterogeneity and is increasingly used to understand transcriptional differences at a single-cell level[11,12,13]. For AD, several scRNAseq studies have been performed[14,15,16,17,18] that have generated new insights into AD pathophysiology. Many scRNAseq studies focus on cell type abundance[18], cell type specific differential expression[14,18], identifying novel cell types[16] or exploring cellular differentiation trajectories[19] – where each cell is kept as an independent entity, while being categorized into distinct cell types. An alternative approach utilizing scRNAseq data involves aggregating multiple measurements of genes within pre-defined cell populations, often delineated by cell type and individual, generating pseudo bulk datasets[20,21]. This approach has successfully been used to identify differential cellular states across conditions[21], exploring cell type specific responses[22], and identifying cell type specific gene regulation under genetic control[23].

While scRNAseq data are well suited for e.g. differential expression analysis (DEA), determining expression correlations in scRNAseq data remains

challenging[24]. scRNAseq data is characterized by large numbers of zero counts; the lower the expression, the more abundant the zeros[25]. Consequently, low-expressed genes can appear highly correlated due to just a few paired measurements, while the remaining measurements are zero[26,27]. The pseudo bulk approach provides a solution, as each gene would be represented by the aggregated value within a cell population, delineated by cell type and individual. As such, even low-expressed genes are represented by a robust aggregated value and the correlation is determined by the collinearity between genes across individuals instead of single cells. However, even though most scRNAseq datasets contain large numbers of cells, these are often derived from a small number of individuals, making it challenging to identify meaningful correlations in pseudo bulk data.

To overcome these challenges, we here combined seven previously published AD scRNAseq datasets[14,15,16,17,18] and generated seven cell type specific pseudo bulk datasets (excitatory neurons, inhibitory neurons, astrocytes, oligodendrocytes, oligodendrocyte progenitor cells (OPCs), microglia and endothelial cells), ranging from 132 to 192 individuals. We used this data to investigate differential correlation[28,29] of genes between healthy individuals (control, CT) and patients diagnosed with AD. In contrast to DEA, differential correlation analysis (DCA) provides insight in whether transcriptional changes are independent or coordinated and provides insight into dynamic associations of key regulators subject to AD. For each cell type we explored gene-gene correlations that are significantly different in AD compared to CT. Using a network representation of differential correlations, we identified distinct sets of regulatory hubs for each cell type. Using the number of differential correlations to rank genes located near AD genetic risk variants, we prioritized known causal genes and identified potential novel ones. In addition, this approach revealed altered states of biological processes in AD associated with the prioritized genes. Finally, taking advantage of the characteristics of pseudo bulk data, we performed co-expression analysis between genes expressed in excitatory neurons and four other cell types (inhibitory neurons, astrocytes, oligodendrocytes and microglia) to identify pairs of co-expressed genes that are expressed in different cell types.

# 2.2 Results

## 2.2.1. Analysis workflow

The analysis workflow consists of six major components. The first component describes the demographics of the cell type specific pseudo bulk datasets that were composed of seven separate AD scRNAseq datasets (**Fig. 1a**). In the second component, a general overview of differential correlation results between CT and AD is presented (**Fig. 1b**). Then we continue to investigate hubs; genes that have the majority of differential correlations with other genes within each respective cell type **(Fig. 1c)**. In the fourth component, we compare hubs between cell types; are they cell type spcific of shared? Here we test whether shared hubs also share neigbourhoods. In the next component, we use the number of differential correlations of genes genomically located near AD risk variants in a prioritization scheme for identifying putative causal genes and cell types (**Fig.**

**1d**). In the sixth and final analysis, we perform a co-expression analysis in healthy individuals between genes expressed in excitatory neurons and inhibitory neurons, excitatory neurons and astrocytes, excitatory neurons and oligodendrocytes and finally excitatory neurons and microglia (**Fig. 1e**), thus asking the question whether there are gene-pairs co-expressed across different cell types.

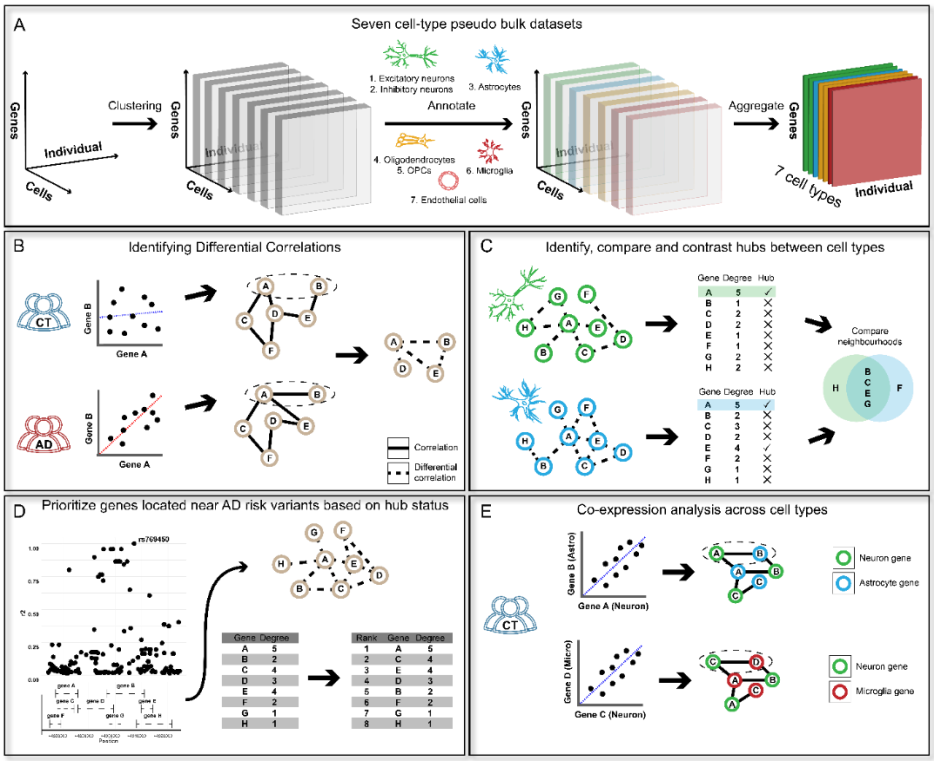



**Figure 1** Schematic overview of the analysis workflow. **A** Overview of how the seven cell type specific pseudo bulk datasets were combined. Starting with three dimensions (genes, cells and individuals), the data was clustered along the cells axis. Next, the clusters were annotated, after which the annotated data was aggregated on cell type and individual. This resulted in seven cell type specific datasets with genes defining the rows and the individuals defining the columns. **B** Schematic overview of differential correlations and differential correlation network (DCN). The degree of correlation is first calculated between pairs of genes in both healthy controls (CT) and Alzheimer's Disease (AD) patients separately, resulting in different co-expression networks. The differential correlation network is defined by the difference between both co-expression networks. **C** Schematic representation of comparing DCNs and the corresponding hubs between cell types. Different cell types have different DCNs and these similarities and differences are identified. Additionally, the network neighborhood genes of the shared hubs is compared. **D** Schematic representation of the gene prioritization scheme. Genes located near AD risk variants are ranked based on their hub status (i.e., based on the number of genes that have an altered association in AD compared to CT). **E** Schematic overview of differential correlation between genes expressed in neurons and genes expressed in astrocytes.

## 2.2.2. Demographics of cell type specific datasets

We collected seven scRNAseq Alzheimer's Disease (AD) datasets, comprised of 1,341,953 cells, and generated seven cell type specific pseudo bulk datasets (*See*

*methods: Aggregation, integration and batch correction*). The excitatory and inhibitory neuron datasets comprise of five datasets (**Table 1**) and consist of 180 individuals, of which 81 were diagnosed with AD and 84 had no cognitive impairment (CT). A total of 15 individuals with mild cognitive impairment (MCI) and/or having other causes for MCI were characterized as other (O) and were excluded from any analyses. The astrocyte, oligodendrocyte and oligodendrocyte progenitor cell (OPC) datasets comprise six datasets and consist of 180 individuals ($N_{AD}$ = 87, $N_{ct}$ = 90, $N_O$ = 15 ). The microglia dataset comprises five datasets and consists of 168 individuals ($N_{AD}$ = 73, $N_{ct}$ = 73, $N_O$ = 22). The endothelial cell dataset comprises four datasets and consists of 132 individuals ($N_{AD}$ = 60, $N_{ct}$ = 70, $N_O$ = 2).

**Table 1** Dataset characteristics and demographics

| Cell type | BA 9 | BA 10 | BA 9-46 | LAU | ENT | SEA-AD | BA 9-46-Micro | Original No. cells | No. genes | No. subjects | %female | AD | CT | O |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Excitatory Neurons** | X | X | X | X | | X | | 663,175 | 9,501 | 180 | 51% | 81 | 84 | 15 |
| **Inhibitory neurons** | X | X | X | X | | X | | 268,060 | 3,556 | 180 | 51% | 81 | 84 | 15 |
| **Astrocytes** | X | X | X | X | X | X | | 109,713 | 2,615 | 192 | 50% | 87 | 90 | 15 |
| **Oligodendrocytes** | X | X | X | X | X | X | | 185,175 | 2,018 | 192 | 50% | 87 | 90 | 15 |
| **OPCs** | X | X | X | X | X | X | | 41,611 | 243 | 192 | 50% | 87 | 90 | 15 |
| **Microglia** | | X | X | X | | X | X | 58,443 | 1,356 | 168 | 53% | 73 | 73 | 22 |
| **Endothelial cells** | | X | X | X | | X | | 15,776 | 368 | 132 | 52% | 60 | 70 | 2 |

## 2.2.3. Alzheimer's Disease is characterized by altered correlations between gene pairs across cell types

To identify altered gene-gene relationships between CT and AD individuals, we performed differential correlation analysis within each cell type. Across all cell types, a total of 374,243 pairs of genes (~0.65% of all tested pairs, **Fig. 2a**) had altered transcriptional relationships in AD ($P_{adj} \leq 0.01$, $|\Delta r| \geq 0.5$). For 253,135 pairs, an increase in correlation coefficient ($\Delta r \geq 0.5$) was observed in AD and for 121,108 pairs a decrease ($\Delta r \leq -0.5$, **Fig.2b**). Most altered relationships were identified in excitatory neurons (n = 313,756, 0.70%), followed by inhibitory neurons (n = 44,974, 0.72%), astrocytes (n = 7,669, 0.22%), microglia (n = 4,061, 0.44%), oligodendrocytes (n = 3,219, 0.61%), endothelial cells (n = 515, 0.77%, **Fig. 2c**) and OPCs (n = 49, 0.17%). We next identified genes that are differentially correlated with age or Braak stage in AD individuals compared to CT (e.g., no correlation between gene expression and age in CT but a positive correlation in AD). Across all cell types, 169 genes were significantly differentially correlated with age (**Fig. 2d-j**, e.g. upregulated with age in CT while downregulated with age in AD) and 215 genes were significantly differentially correlated with Braak stage (**SFig. 1**) . *PTPN3* (**Fig. 2k**, $r_{CT}$ = -0.29, $r_{AD}$ = 0.44, $P_{adj} = 1.35 \times 10^{-4}$) showed the

most extreme changes in association with age in excitatory neurons from AD patients.

Next, we investigated whether the observed differential correlations are explained by the differential expression of the genes between CT and AD individuals. Most genes that were differentially correlated were not significantly differentially expressed. Of all genes that were differentially correlated with at least one other gene (n = 18,321), 3,187(~17%) genes were also significantly differently expressed ($P_{FDR} \leq 0.01$, **STable 1**). Interestingly, most pairs of genes that showed an increase in correlation coefficient in AD had the same directional effects in gene expression; both up- or both downregulated. Vice versa, most pairs of genes that showed a decrease in correlation coefficient showed opposite directional effects in gene expression; one up- and one downregulated. Testing the association between correlation (increase/decrease) coefficient and directionality of effects (same/opposite) on gene expression resulted in a log odds ratio of 2.78 (95%CI = 2.76, 2.80, **STable 2**).

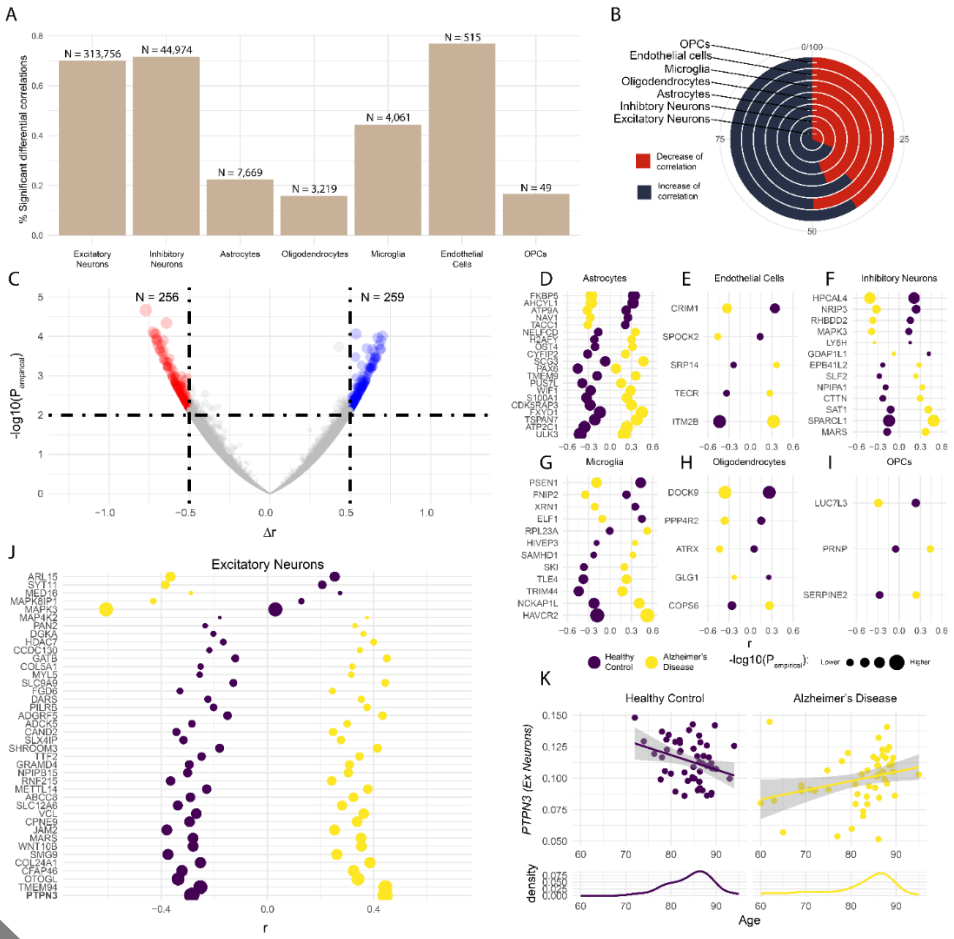



**Figure 2 A** Percentage and number of significant differential correlations for each cell type. **B** For each cell type the percentage of pairs of genes that have an increased or decreased correlation. Increasing meaning a higher correlation coefficient between a pair of genes in individuals with

Alzheimer's Disease (AD) compared to healthy controls (CT), and vice versa. **C** Volcano plot of the differential correlations in endothelial cells. Each dot represents a gene-pair, the x-axis represents the difference in correlation coefficient for the respective pair and the y-axis represents the –log10 empirical P-value. Blue dots are significant differential correlations where an increased correlation coefficient was found in AD. Conversely, red dots are significant differential correlations where a decreased correlation coefficient was found in AD. **D, E, F, G, H, I, J** Plots of the genes that are differentially correlated with age within the respective cell types. The x-axes represent the Spearman's rank correlation coefficient. Yellow dots represent the correlation coefficient in AD for the respective gene with age and purple dots represent the correlation coefficient in CT for the respective gene with age. The size of the dots correspond to the –log10 p-value. K Dot plot for PTPN3 (excitatory neurons). Each dot is an individual, the x-axes represent the age of the individuals in years and the y-axes the expression of the respective genes in the respective cell types. Yellow dots are individuals with Alzheimer's and purple dots are healthy controls. Beneath the dot plot a density plot of age is shown.

## 2.2.4. Regulatory hubs are primarily cell type specific

Next, we constructed gene differential correlation networks for each cell type, where vertices represent genes and edges the significant differential correlations between genes (**Fig. 1b**). Degree distributions of these networks followed a power law (**SFig. 2a-g, STable 3**), showing that these networks have scale-free topology and that per network only a few central genes (hubs) are involved in the majority of altered relationships. Comparing hubs between cell types (**Fig. 1c**) showed that 824 (95%) hubs were cell type specific, and 42 (5%) hubs were shared between at least two cell types (**Fig. 3a**). Of all identified hubs (N= 866), 261 (30%) hubs had known regulatory functions; 62 (7%) were known transcription factors (TFs)[30], 70 (8%) were known cofactors[30] and 154 (18%) hubs were regulators of molecular functions (GO:0065009, **Fig. 3b**). Interestingly, when pairwise comparing the neighbourhoods of excitatory neuron TF hubs ($N_{pairs}$ = 930), we found 214 TF pairs with opposite differential correlations with the same genes. For example, the TF-hub *ZNF579* was negatively correlated with *CDH10* in CT (r = -0.28) and positively correlated in AD (r = 0.31, $P_{adj} = 4.38 \times 10^{-4}$). Conversely, TF-hub *ZNF33A* was positively correlated with *CDH10* in CT (r = 0.38) and negatively in AD (r = -0.23, $P_{adj} = 3.29 \times 10^{-4}$). This suggests that there are genes that are under control by different TFs in AD compared to CT. Furthermore, within the respective cell types, the neighbourhoods of 132 hubs were significantly enriched ($P_{FDR} \leq 0.01$) for the KEGG AD pathway (**SFig. 3**, **STable 4**), namely in excitatory neurons (N= 97), inhibitory neurons (N = 33) and astrocytes (N = 2). We identified hub genes that were differentially correlated with age, including sixteen excitatory neuron hubs, two inhibitory neuron hubs (MARS and SLF2), two astrocyte hubs (ARHGEF9 and CYFIP2), and one hub from endothelial cells (SPOCK2). Additionally, we identified hub genes that were differentially correlated with Braak stages, which included one astrocyte hub (ZNF302) and three microglia hubs (ALCAM, RAB11A and RASA3).Of the cell type specific regulatory hubs, only excitatory and inhibitory neuron hubs showed functional enrichment (GO terms), albeit for distinct processes. For example, hubs of excitatory neurons were enriched for regulation of transferase activity (N = 27, $P_{FDR} = 6.20 \times 10^{-6}$, **Fig. 3c**) and negative regulation of protein phosphorylation (N = 15, $P_{FDR} = 2.42 \times 10^{-4}$), and hubs of inhibitory neurons were enriched for positive regulation of RNA biosynthetic process (N = 21, $P_{FDR} = 5.65 \times 10^{-5}$, **Fig. 3d**) and regulation of transcription by RNA polymerase II (N = 24, $P_{FDR} = 9.99 \times 10^{-5}$).

Next, we examined hubs that were shared between cell types (N = 42). *SARAF* was the only hub in four cell types and *DGKZ*, *EMC7*, *HNRNPH1*, *MOK, NDUFV3* and *S100A6* were shared hubs in three cell types. To investigate whether hubs shared between cell types also share neighbourhoods we performed a Fisher's exact test between the neighbourhoods of the respective cell types for each hub that was shared between at least two cell types. Considering all cell types, we found 21 hubs with significantly overlapping neighbourhoods between cell types (*fisher exact test*, $P_{FDR} \leq 0.01$, **SFig. 4**, **STable 5**). Of the 21 hubs shared between excitatory and inhibitory neurons, 19 had significantly overlapping neighbourhoods. These results show that when hub genes are shared between cell types the putative gene expression regulatory disruptions are also shared.

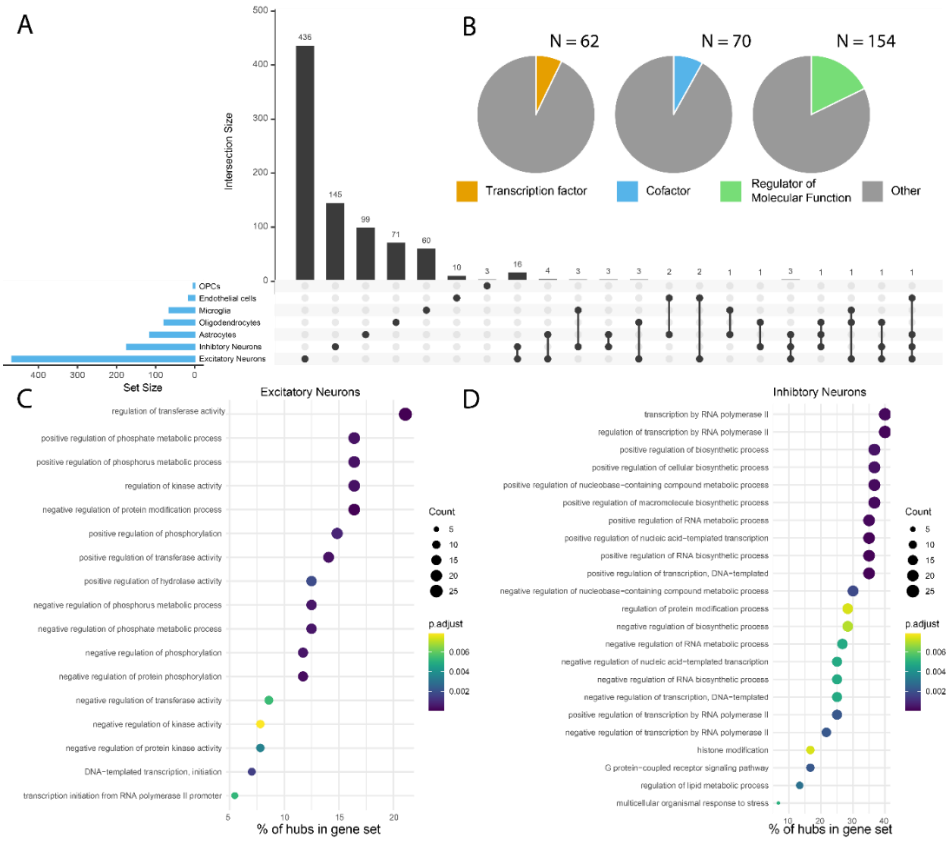



**Figure 3 A)** An UpSet plot of the hubs identified in each cell type, showing the degree of overlap of hubs between the cell types. **B** Pie charts showing the number of hubs that have known regulatory functions; transcription factor (orange), cofactor (blue) and regulator of molecular function (green). **C-D** Dot plot of the hub enrichment results for excitatory neurons (**C)** and inhibitory neurons (**D**), The x-axes represent the % of how many hubs belong to the GO term relative to all genes comprising the GO term. The y-axes represent the GO terms. The color of the dots represents the adjusted P-value (FDR) for the term and the size of the dots represent the number of hubs belonging respective GO term.

## 2.2.5. Differential correlation-based gene prioritization for Alzheimer's Disease risk variants

Considering that many hubs are annotated to have known regulatory functions, we argue that the number of differential correlations can be seen as a measure of importance and could potentially hint at causal involvement of the respective hub in AD. As such, we hypothesized that the number of differential correlations can be used to prioritize genes located near AD risk variants. We focused on 79 AD risk variants identified by Wightman et al[4] and Bellenguez et al[5]. Using CONQUER[31], we identified 2,528 genes near these 79 variants (see methods). Of these genes, 975 were present in our collection of single cell datasets (**SFig. 5**). In total, for 32 variants at least one hub was located nearby (**Fig. 4a**, $N_{excitatory\ neurons}$= 19, $N_{inhibitory\ neurons}$= 9, $N_{astrocytes}$= 6, $N_{oligodendrocytes}$= 3, $N_{microglia}$= 5). rs61732533 had four nearby located hubs; *MAF1* and *BOP1* in excitatory neurons and *PLEC* and *CPSF1* in inhibitory neurons. Interestingly, a dysfunction of *PLEC* in neurons is known to be associated with tau accumulation[32]. rs769450 also had four nearby located hubs; *OPA3*, *APOE* and *FBXO46* in excitatory neurons and *APOE* in inhibitory neurons. For five variants the annotated gene was also a hub (rs769450 - excitatory and inhibitory neurons - *APOE*, rs1065712 - excitatory neurons - *CTSB*, rs141749679 – astrocytes - *SORT1,* rs72777026 - astrocytes - *ADAM17* and rs450674 microglia - *MAF).*

Next, we calculated a cell type specific normalized rank based on the number of significant differential correlations (higher number of differential correlations = higher priority). Using this rank we prioritized genes within each cell type and for each risk variant. In total, we prioritized 230 genes in all cell types (**Fig. 4b, STable 6**). For 29 variants the previously annotated gene corresponded to the highest prioritized gene in different cell types: e.g. rs769450-*APOE*, rs602602-*ADAM10*, rs4663105-*BIN1*, rs11218343-*SORL1*, rs1532278-*CLU* and *rs141749679-SORT1.* For 67 variants, another potential new risk gene ranked highest in the various cell types among which nineteen transcription factors were prioritized (e.g. rs1140239 - *ZNF785*, rs7384878 - *CUX1*). Another prioritized gene was *MAF1* for rs61732533 in excitatory neurons. *MAF1* is a stress responsive transcription factor and mTOR effector[33]. Aberrant mTOR signaling is suggested to strongly contribute to AD, mainly through oxidative stress[34]. *KCNC3* was prioritized for rs9304690 in inhibitory neurons. *KCNC3* is a potassium channel. A dysfunction of potassium channels has been associated with AD and many other neurological disorders[35]. To evaluate our prioritization approach, we compared it to prioritization using DEA. In short, using DEA we prioritized genes located nearby risk variants that were the most differentially expressed. The two sets of prioritized genes (DCA- and DEA-based) were evaluated using DisGeNET[36]. DisGeNET scores disease-gene associations (GDA, 0.01 – 0.9) based on their level of evidence in literature and from curated sources (higher GDA = more evidence). DCA- and DEA-based gene prioritization resulted in 70 and 24 genes respectively that were previously associated with AD (**Fig. 4c**). The highest scoring genes using DCA were *APP* (GDA = 0.9) and *APOE* (GDA = 0.70). The highest scoring genes using DEA were TOMM40 (GDA = 0.50) and BIN1 (GDA = 0.50). Thus, we identified more disease-associated genes and with higher levels of evidence (i.e., higher GDA score) using DCA compared to DEA,

offering an internal validation that DCA can be used as a prioritization method to identify putative risk genes. Altogether, with this prioritization approach, well-known AD genes were prioritized as well as genes that have not yet been associated with AD previously. Our method hints at involvement of these lesser-known genes in AD, and as such they might be of interest for future studies.

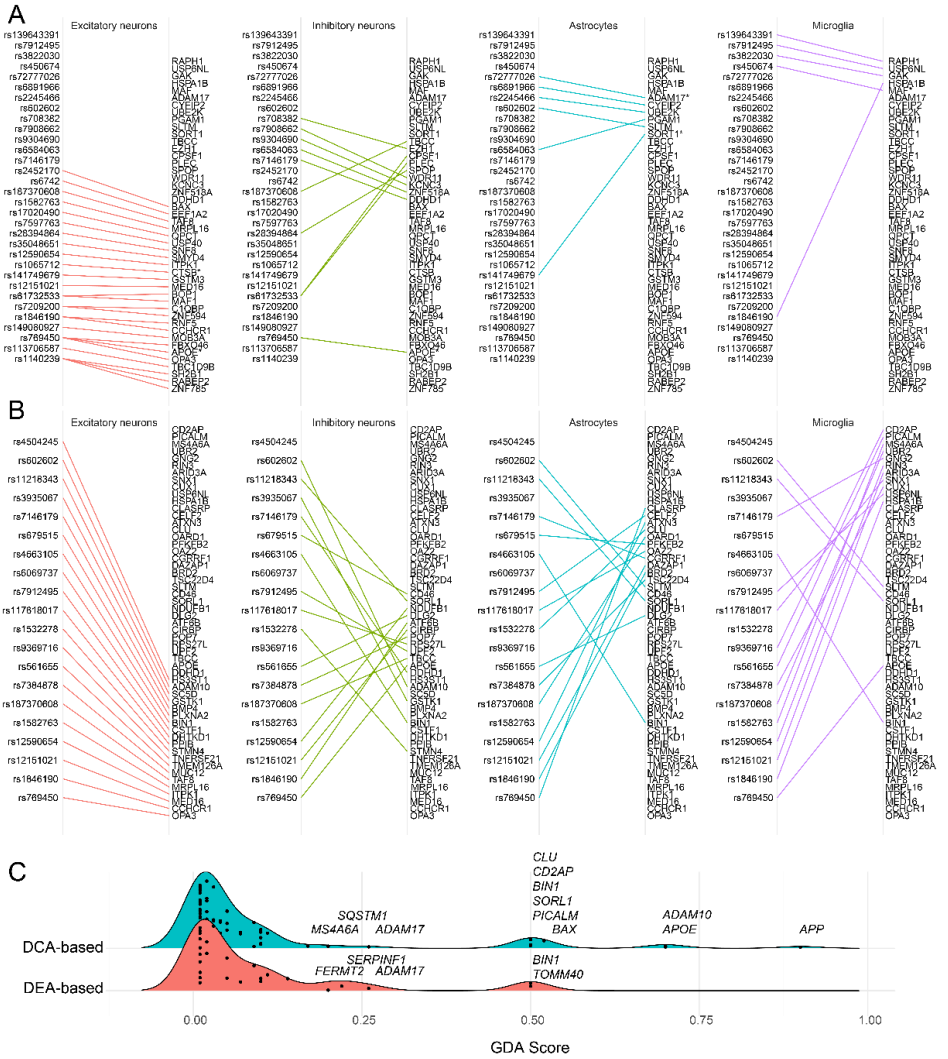



**Figure 4** Prioritization plots of excitatory neurons, inhibitory neurons, astrocytes and microglia. **A)** hubs located nearby risk variants. **B)** The prioritized genes for the top 20 risk variants from Wightman et al (sorted on significance). **C)** Distribution of GDA scores for the DCA-based (blue) and DEA-based (red) prioritized genes.

## 2.2.6. Co-expression analysis across cell types suggests glia-to-neuron intercellular directionality of gene expression regulation

Finally, taking advantage of the characteristics of pseudo bulk data, we explored whether co-expression of genes exceeds cell type boundaries and whether it can also be indicative of transcriptional regulation across cell types. We focused on interactions between excitatory neurons and inhibitory neurons, excitatory neurons and astrocytes, excitatory neurons and oligodendrocytes, and finally excitatory neurons and microglia. Using only CT data, co-expression between pairs of genes was assumed at a spearman rank correlation coefficient of $|r| \geq 0.6$. Excitatory neurons and inhibitory neurons had the highest co-expression rate as 2.69% (n = 909,061, **Fig. 5a**) of all tested gene pairs were co-expressed. For astrocytes, oligodendrocytes and microglia this was 0.06% (n = 16,123), 0.08% (n = 16,196) and 0.02% (n= 2,078) respectively. Interestingly, when constructing co-expression graphs of each tested cell type pair we found that genes expressed in astrocytes (**Fig. 5b**), oligodendrocytes (**Fig. 5c**) and microglia (**Fig. 5d**) were more densely connected to genes in excitatory neurons, than the other way around (**Fig. 5e-h**). For example, the most densely connected gene in astrocytes, *HINT1,* was co-expressed with 728 genes in excitatory neurons, which is 7.66% of all genes measured in the excitatory neurons. The most densely connected excitatory neuron gene, *FAU (***Fig. 5b***),* was co-expressed with only 65 (2.49%) genes in astrocytes. Of note, *HINT1* was only co-expressed with 238 excitatory neuron genes in the AD population, meaning it lost co-expression with 490 genes. Given that *HINT1* is implicated in $Ca^{2+}$ signalling[37] and that an increase of astrocytic $Ca^{2+}$ signalling is associated with AD[38] and thought to have downstream effects on neuronal metabolism[39], astrocytic *HINT1* might be involved in this dysregulation. In oligodendrocytes, the most densely connected gene was *YWHAH* (n = 1,158, 12.2%), which is implicated in the regulation of many signaling pathways[40]. In microglia *UBC* (n = 361, 3.80%) was most densely connected. *UBC* is involved in ubiquitination, which is a post-translational modification process involved in the regulation of many processes[41]. The function of these densely connected genes hints at transcriptional regulation of excitatory neuron by genes expressed in inhibitory neurons, astrocytes, oligodendrocytes and microglia.

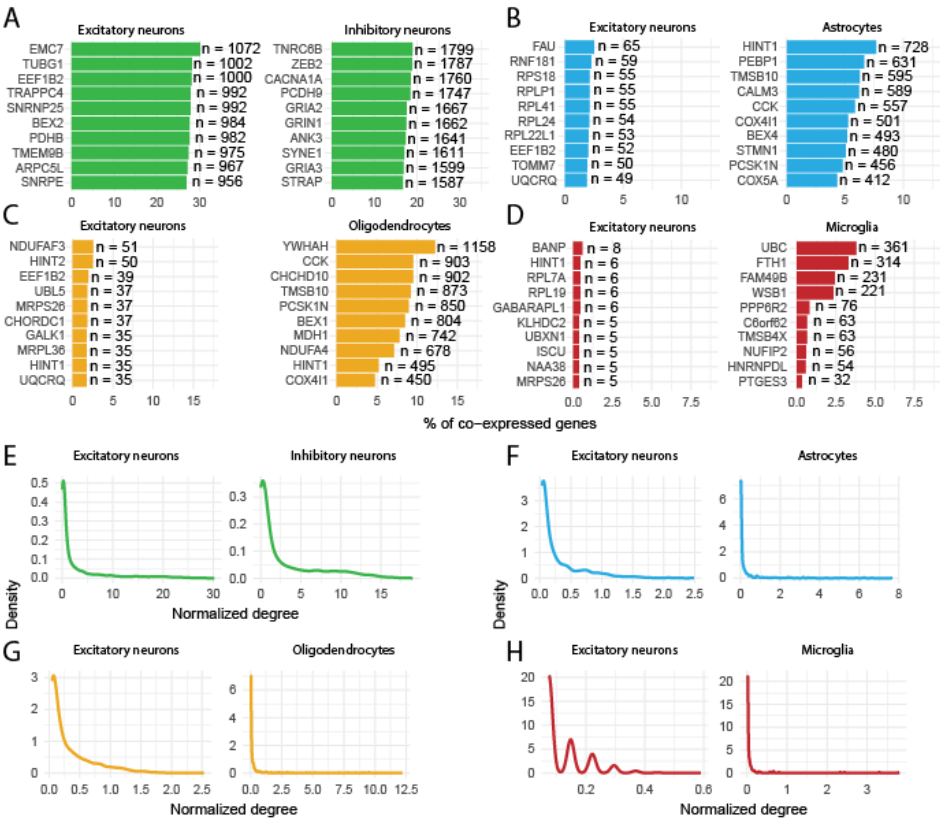



**Figure 5:** Co-expression between cell types. Top ten (defined by no. of co-expression) excitatory neuron genes co-expressed with **A)** inhibitory neuron genes, **B)** astrocyte genes, **C)** oligodendrocytes and **D)** microglia, and vice versa. E.g. EMC7 expressed in excitatory neurons is co-expressed with 1,072 genes expressed in inhibitory neurons, TNRC6B expressed in inhibitory neurons is co-expressed with 1,799 genes expressed in excitatory neurons. **E-H)** Distribution of normalized degree of **E)** genes expressed in excitatory neurons co-expressed with inhibitory neuron genes (left), and vice versa (right). **F)** Of genes expressed in excitatory neurons co-expressed with astrocyte genes, and vice versa. **G)** Of genes expressed in excitatory neurons co-expressed with oligodendrocyte genes, and vice versa. **H)** Of genes expressed in excitatory neurons co-expressed with oligodendrocyte genes, and vice versa. Degrees are normalized for total number of genes of the "other" cell type.

## 2.3. Discussion

In this study, we have provided insight into cell type specific and coordinated transcriptional changes in AD and pin-pointed putative key transcriptional regulators of the observed changes. Most importantly, we have provided a prioritization scheme that identifies probable causal genes and important cell types by superimposing the set of most differentially correlated genes onto genes located near AD risk variants. Finally, we have shown that transcriptional relationships and differences in these relationships do exceed the confines of cell types, hinting at altered intercellular communication AD. Altogether, performing differential correlation analysis (DCA) on scRNAseq data provided a

comprehensive insight into transcriptional changes and consequences that are associated with AD.

Our results show that the number of altered associations a gene has with respect to other genes in the trait of interest opposed to healthy controls can be used as a measure of involvement and severity of consequence for that trait, with the respective hub as a probable key actor. In the case of AD, examples of this are *APOE*, *SORL1* and *ADAM10,* three well known AD genes[42,43,44,45], which were identified as high-ranking hub genes. The ε4 allele of *APOE* is the largest contributor to genetic risk for AD[46]. Of note, *APOE* ranked especially high in neurons (both excitatory and inhibitory), while *APOE* expression is generally low in neurons. This suggest that DCA analysis of pseudo bulk data is particularly capable of also identifying novel cell type specific interactions of risk genes. Interestingly, a recent study confirmed that under stress some neurons indeed express *APOE*[47], which might be reflected in our results.

A strength of our prioritization scheme is that it does not require expression quantitative trait loci (eQTL) or colocalization[10] analyses for variant-gene mapping, which is generally done in GWASs. The effect of a variant on a gene is not always reflected in differential expression of the respective gene. For instance, a variant might alter the amino acid sequence of a protein, without changing the extent to which the transcript is expressed. As the function of the protein is possibly changed, it can also alter functional relationships with other proteins and their respective transcripts. Support for this is shown with the association between rs769450 (part of the ε4 allele) and *APOE.* In brain eQTL and pQTL studies[48], rs769450 has been shown not to influence the abundance of the *APOE* protein or transcript. Additionally, in differential expression analyses of AD, *APOE* is often not among the most differentially expressed genes[18,14]. However, in our analysis, *APOE* is ranked among the most differentially corelated hub genes, highlighting the importance of looking beyond changes in expression of only one single gene at a time. Alternatively, variants might alter the enhancer or promotor regions of a gene, and as a result the respective gene might be under control of different TFs in AD compared to healthy controls. Our results indeed suggest that when comparing AD with healthy controls, there might be genes that are under transcriptional control by different TFs compared to healthy controls. Whether this is due to changes in enhancer or promotor regions remains to be elucidated.

We mainly focused on hub genes in differential correlation networks, which were defined as genes with the most (top 5%) differential correlations. As hubs are involved in the majority of altered associations, we expected these to have a regulatory function. However, a previous study evaluating differential co-expressions showed that differentially regulated targets are more likely to be identified as hubs, instead of the regulators (TFs)[28]. With this in mind, two additional layers of evidence were collected to strengthen the regulatory status of the identified hubs. First, annotations concerning a regulatory function were collected; are the hubs TFs[30], cofactors[30] or regulators of molecular function? The second layer of evidence was disease association; do the hubs have altered transcriptional relationships with known AD genes[49,50]? Additionally, the gene

prioritization scheme provides a third layer of evidence, as it is generally assumed that the causal gene is located near the identified risk variant.

Our results showed that when a pair of genes has increased correlation in the tested group, and both genes are significantly differentially expressed, most often the genes of that pair have similar directional effects in the tested group in terms of mean expression (either up- or downregulated in both groups), whereas a decrease of the correlation coincides with opposite directional effects. However, most genes that are differentially correlated are not differentially expressed. This shows that genes that are subjected to an increase of their correlation are more likely to respond in a similar direction subjected to the perturbation, suggesting shared and altered TF control. Loss of shared control may result from regulation by other TFs. In contrast to differential expression analysis, DCA has an added value in providing a more detailed view of transcriptional changes, and whether the changes are coordinated or not.

One limitation of this study is that different brain regions were confounded by the batches, and therefore were corrected for. It is known that different brain regions have different gene co-expression networks and different cell types and cell-to-cell connectivity, and that different brain regions respond differently to AD[51]. As such, our combination of different brain regions likely favoured transcriptional changes that are shared between brain regions.

Overall, we performed DCA in single-cell data and have shown that AD is associated with coordinated transcriptional changes. Our analysis highlights the complexity and heterogeneity of cell type specific responses to AD. And lastly, with our bottom-up approach towards gene prioritization we provide a compendium of genes that could serve as guidance for functional follow-up studies.

# 2.4. Methods

## Single-cell RNAseq data

Four 10x single-cell RNAseq (scRNAseq) datasets were acquired from AMP-AD knowledge portal of which the subjects were participants of the Religious Orders Study and the Memory and Aging Project (ROS/MAP). Two 10x datasets[17,18] were acquired from GEO (GSE157827 and GSE138852). The Seattle Alzheimer's Disease Brain Cell Atlas (**SEA-AD**) was obtained from (https://registry.opendata.aws/allen-sea-ad-atlas/). The first dataset (***BA9***, ID: syn16780177) consisted of 24 subjects and originated from the dorsolateral prefrontal cortex (DLPFC), specifically Brodmann area 9 (BA9). Raw fastq files were obtained of this dataset. The second dataset (***BA10***, ID: syn18485175) consisted of 48 subjects and originated from the prefrontal cortex (PFC), specifically BA10. A count matrix was obtained of this dataset as it was already processed with CellRanger aligning reads to the hg38 genome[14]. The third dataset (***BA9-46***, ID: syn21670836) consisted of 32 subjects and originated from the DLPFC, BA9 and BA46. Of this dataset a count matrix was obtained as it was also processed with CellRanger aligning reads to the hg38 genome[52]. The fourth dataset (***BA9-46-Micro***, ID: syn12514624) was a microglia only dataset,

consisted of 12 subjects and originated from the DLPFC, BA9 and BA46[15,16]. Of this dataset, raw fastq files were obtained. The fifth dataset (**LAU**[17], GSE157827) consisted of 21 PFC samples and originated from 12 individuals diagnosed with AD and 9 healthy controls. Of this dataset a raw count matrix was acquired as it was also processed with CellRanger aligning reads to the hg38. The sixth dataset (**ENT**[18], GSE138852) consisted from 12 entorhinal cortex samples and originated from 6 individuals diagnosed with AD and 6 healthy controls. Of this dataset a raw count matrix was acquired. The seventh dataset (**SEA-AD**) consisted of 89 middle temporal gyrus samples, 23 of which were diagnosed with AD and 32 were specified as CT. Of this dataset the raw count matrix was acquired.

## Clinical data

Clinical data were acquired from the AMP-AD knowledge portal (ID: syn3157322). The variable cogdx was used to characterize controls (CT), Alzheimer's disease (AD) and other (O). Cogdx represents the clinical consensus diagnosis of cognitive status at time of death and is indicated with a value ranging from one to six. A value of one represents no cognitive impairment (CI), as such, individuals with a cogdx of one were characterized as CT. A value of four represents Alzheimer's dementia and no other cause of CI, as such, these individuals were characterized as AD. The remaining values represent mild CI and/or other causes for dementia and these individuals were characterized as O. Besides clinical diagnosis, *APOE* genotype, Braak stage, sex, and age at time of death was also available. However, age at time of death is censored above the age 90 years. Of the **LAU**, **ENT** and **SEA-AD** datasets the clinical data were acquired from the corresponding sources. For both datasets; age, sex, clinical diagnosis and Braak stage were available.

## Single-cell RNA-seq data alignment and pre-processing

The two datasets (***BA9*** and ***BA9-46-Micro***) of which fastq files were acquired were processed with CellRanger (4.0.0) aligning reads to the hg38 genome, default parameters were used. Next, all datasets were jointly pre-processed. Cells with ≤20% mitochondrial counts, ≥300 total counts, ≤20,000 total counts and ≥200 measured genes, were kept for downstream analyses.

## Clustering and cell type annotation

Each dataset was separately processed for clustering and cell type annotation which was done as follows. The processed count matrix was loaded in Seurat 3.2.2[53]. The data was log-normalized, such that: $y_{ij} = \log\left(\frac{x_{ij}}{\sum_j x_{ij}} \times 10^4\right)$, where $x_{ij}$ and $y_{ij}$ are the raw and normalized values for every gene $i$ in every cell $j$, respectively. Next, with the 2,000 most variable genes (default with Seurat), principal components analysis (PCA) was performed. The number of principal components (PCs) used for clustering was determined using the elbow method (***BA9***:12 PCs, ***BA10***:11 PCs, ***LAU*:**11 PCs*,* ***BA9-46***:10 PCs, ***BA9-46-Micro:*** 7PCs, ***ENT***: 6 PCs).Next, Seurat's `FindNeighbours` and `FindCluster` functions were used, which utilizes Louvain clustering, the resolution was set at 0.5. A UMAP plot was made to visualise and inspect the clusters. The following

cell types were identified using known and previously used markers[14]: excitatory neurons (*SLC17A7*, *CAMK2A*, *NRGN*), inhibitory neurons (*GAD1*, *GAD2*), astrocytes (*AQP4*, *GFAP*), oligodendrocytes (*MBP*, *MOBP*, *PLP1*), oligodendrocyte progenitor cell (*PDGFRA*, *VCAN*, *CSPG4*), microglia (*CSF1R*, *CD74*, *C3*) and endothelial cells (*FLT1*, *CLDN5*). Based on differential expression of these markers between clusters, determined with Seurat's `FindMarkers` function, cell types were assigned (SFiles. 1). When clusters were characterized by markers of multiple cell types, they were assigned as: "Unknown". Of the **LAU** and **SEA-AD** the accompanying cell type labels were used.

## Aggregation, integration and batch correction

Per dataset, for each cell type, pseudo bulk data was generated. For instance, for each subject, cells annotated as astrocytes were aggregated in a single vector. As such, we generated cell type specific datasets. Aggregation was done based on the binary expression pattern, since the percentage of zeros for a gene in a cell population is highly associated with its mean expression[25,54]. The aggregated value of a gene for an individual was defined by the percentage of non-zero measurements in a specific cell population. Genes were kept for aggregation if they were expressed in ≥ 1% of the respective cell population in all datasets. Per cell type, the datasets were combined. Each new cell type specific dataset was batch corrected with respect to a reference dataset. First by performing a median ratio normalization[55] and then, batch correction was performed with the `ComBat` function from the R-package sva (3.36.0)[56]. For the excitatory neurons, inhibitory neurons, astrocytes, oligodendrocytes, OPCs and endothelial cells, the ***BA9*** dataset was used as reference and for the microglia the ***BA9-46-Micro*** dataset was used as reference. Integration was confirmed with a PCA and visually inspecting the first four principal components.

## Differential correlation analysis (DCA)

Differential correlation or differential co-expression was investigated between controls and AD individuals. Differential correlation we calculated similarly as described by McKenzie et al.[29] First, Spearman's rank correlation coefficient between a pair of genes was calculated separately for the groups of interest based on the aggregated detection rates. This results in a correlation coefficient for each group. Next, the correlation coefficients are transformed to z-scores by means of the Fisher z-transformation[57]. Then, the difference between z-scores can be calculated with equation 1:

$$\Delta z = \frac{(z_x - z_y)}{\sqrt{var(r_x) + var(r_y)}} \quad (1)$$

where var(r) is calculated by 1.06/(n-3) , with n being the sample size of the respective groups. As $\Delta z$ is normally distributed, a two-sided P-value for the differential correlation between each pair of proteins can be calculated. Besides the P-value resulting from the Z-test, empirical P-values were also calculated. The empirical null distribution was generated by permuting the group labels a 1,000 times and performing the Z-test on each pairwise combination of genes.

The resulting P-values contributed to the empirical null distribution ($x_0$). Next, the empirical P-value was calculated as:

$$P_{emp} = \frac{\sum_{n=1}^{N} I(P \geq x_0)}{N}$$

Where $I()$ is an indicator function, $N$ is the total number of P-values that make up the empirical null distribution and $P$ is the actual P-value for which we want to determine the empirical P-value. Significance was assumed at $P_{emp} \leq 0.01$.

## Classification of differential correlations

The directional change of correlation between two genes from one group to another does not reveal whether an association is lost or gained. As illustration, a change from r = -0.9 to r = -0.05 and r = 0.05 to r= 0.9 both have an increase of the correlation coefficient. However, in the first example a very strong association is lost, while in the second example a very strong association is gained. To evaluate differential correlations in terms of loss and gain of association we classified each differential correlation. First the state $\boldsymbol{f}(\boldsymbol{r})$ in both group is determined as follows:

$$f(r) = \begin{cases} 0, -0.25 \leq r \leq 0.25 \\ +, \quad r > 0.25 \\ -, \quad r < -0.25 \end{cases}$$

Where $\boldsymbol{r}$ is the Spearman's rank correlation coefficient of a pair of genes in the respective group. A "**0**" represents a state of no correlation, "+" represents a state of positive correlation and "−" a state of negative correlation. When the state of a pair of genes is "−" or "+" in CT and 0 in AD, then we classify it as a *loss of association* (**-/0, +/0**). Vice versa is defined as a *gain of association* (**0/-, 0/+**) and a change from "−" to "+" or from "+" to "−", is defined as a *flip of association* (**-/+, +/-**). Differential correlations can also remain in the same state between groups (e.g. r = 0.30 to r = 0.95), these are defined as no *change of association*.

## Differential Expression

As the aggregated detection rates were normally distributed across the individuals, differential expression analysis (DEA) was performed with a linear model, where the gene's expression was specified as the outcome variable and the group assignment (CT = 0 and AD =1) was used as predictor variable. The function `lm` from the *stats* package from R (4.0.5) was used. Within each cell type, each gene was tested on differential expression. P-values were corrected for multiple tests, per cell type, with the Benjamini-Hochberg procedure and significance was assumed at an adjusted P-value of $P_{FDR} \leq 0.01$.

## Network analysis

For each cell type, the results from the DCA were used to construct networks using igraph (1.2.6). In these networks, genes were defined as nodes and an edge between two nodes was drawn when they were significantly differentially correlated. The centrality of each gene was determined by the degree (no. of

differential correlations) within the respective networks. To test whether these networks followed the power law the `fit_power_law` function from igraph was used. Hubs were defined as the top 5% of genes having the highest degree. Transcription factors and cofactors were downloaded from AnimalTFDB[30]. Genes involved in regulation of molecular functions (GO:0065009) were identified using the R-packages GOfuncR (1.10.0) and org.Hs.eg.db (3.12.0).

## GO term enrichment analysis

For each cell type, a GO term enrichment analysis was performed with the hubs that have known regulatory functions (TFs, cofactors, GO:0065009). The GO term enrichment analysis was executed with the R-package clusterProfiler (v3.18.1)[58]. Gene symbols were translated to entrez IDs making use of org.Hs.eg.db (3.12.0). For each cell type the background was defined by all genes that were present in the respective cell type specific dataset. P-values were corrected for multiple tests with the Benjamini-Hochberg procedure and significance was assumed at an adjusted P-value of $P_{FDR} \leq 0.01$. The GO term regulation of molecular functions (GO:0065009) was excluded as the hubs were partly pre-filtered with this GO term.

## KEGG Alzheimer's Disease pathway enrichment

To investigate to what degree a hub was associated with Alzheimer's Disease, each hub was subjected to a KEGG[49] AD pathway enrichment analysis. For each hub we performed a gene set enrichment analysis on the genes that were significantly differently correlated with the respective hub. Genes belonging to the KEGG AD pathway (ID: 05010) were defined with org.Hs.eg.db (3.12.0). Enrichment was calculated with the fisher exact test from *stats* package from R (4.0.5). P-values were corrected for multiple tests, per cell type, with the Benjamini-Hochberg procedure. For each cell type the background was defined by all genes that were present in the respective cell type specific dataset.

## Hub overlap

When hubs were identified in multiple cell types, we investigated to what degree the hubs overlap between the respective cell types, in terms of genes the hubs were differentially correlated with (e.g. are they differentially correlated with the same genes in the different cell types). This was done with the fisher exact test from *stats* package from R (4.0.5) and the background was defined by the genes that were present in both cell type specific datasets. P-values were corrected for multiple tests with the Benjamini-Hochberg procedure. A significant overlap was assumed at P-value of $P_{FDR} \leq 0.01$.

## Gene prioritization

For the prioritization we started with the hubs that were identified as previously described. Next, we extracted the RS IDs from two GWASs[4,5]. In total 79 AD risk variants were extracted. For each variant, CONQUER[31] was used to identify genes genomically located near the respective risk variants. Besides defining a fixed window of 1 Mb around the respective variant, CONQUER uses chromatin interaction to dynamically expand the search space. For each variant and cell

type, genes were prioritized that were located near the respective variant and ordered based on the number of differential correlations within the respective cell type. In other words, more differential correlations, higher priority. For each of these genes, the regulatory status was evaluated (see *Network analysis*). Finally, the highest-ranking hubs were compared to the previously annotated genes, provided that the previously annotated gene was present in the respective datasets. For the variants extracted from Wightman et al[4] the genes were assigned based on colocalization, fine-mapping and previous literature. For the variants extracted from Bellenguez et al[5] these genes were the nearest protein coding.

# References


1. Thies, W. & Bleiler, L. 2012 Alzheimer's disease facts and figures. *Alzheimer's Dement.* **8**, 131–168 (2012).

2. Du, X., Wang, X. & Geng, M. Alzheimer's disease hypothesis and related therapies. *Transl. Neurodegener. 2018 71* **7**, 1–7 (2018).

3. Mohandas, E., Rajmohan, V. & Raghunath, B. Neurobiology of Alzheimer's disease. *Indian J. Psychiatry* **51**, 55 (2009).

4. Wightman, D. P. *et al.* A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat. Genet. 2021 539* **53**, 1276–1282 (2021).

5. Bellenguez, C. *et al.* New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat. Genet. 2022 544* **54**, 412–436 (2022).

6. Holstege, H. *et al.* Exome sequencing identifies rare damaging variants in ATP8B4 and ABCA1 as risk factors for Alzheimer's disease. *Nat. Genet. 2022 5412* **54**, 1786–1794 (2022).

7. Prokopenko, D. *et al.* Whole-genome sequencing reveals new Alzheimer's disease–associated rare variants in loci related to synaptic function and neuronal development. *Alzheimer's Dement.* **17**, 1509–1527 (2021).

8. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From Association to Function. *Am. J. Hum. Genet.* **102**, 717–730 (2018).

9. S, B. & SG, T. Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur. J. Epidemiol.* **32**, 377–389 (2017).

10. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, e1004383 (2014).

11. Nagy, C. *et al.* Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. *Nat. Neurosci.* **23**, 771–781 (2020).

12. Wilk, A. J. *et al.* A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nat. Med.* **26**, 1070–1076 (2020).

13. Segerstolpe, Å. *et al.* Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab.* **24**, 593–607 (2016).

14. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).

15. Olah, M. *et al.* A transcriptomic atlas of aged human microglia. *Nat. Commun.* **9**, 1–8 (2018).

16. Olah, M. *et al.* Single cell RNA sequencing of human microglia uncovers a subset associated with Alzheimer's disease. *Nat. Commun.* **11**, 1–18 (2020).

17. Lau, S. F., Cao, H., Fu, A. K. Y. & Ip, N. Y. Single-nucleus transcriptome analysis reveals dysregulation of angiogenic endothelial cells and neuroprotective glia in Alzheimer's disease. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 25800–25809 (2020).

18. Grubman, A. *et al.* A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat. Neurosci.* **22**, 2087–2097 (2019).

19. Eze, U. C., Bhaduri, A., Haeussler, M., Nowakowski, T. J. & Kriegstein, A. R. Single-cell atlas of early human brain development highlights heterogeneity of human neuroepithelial cells and early radial glia. *Nat. Neurosci.* **24**, 584–594 (2021).

20. Squair, J. W. *et al.* Confronting false discoveries in single-cell differential expression. *Nat. Commun. 2021 121* **12**, 1–15 (2021).

21. Crowell, H. L. *et al.* muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat. Commun.* **11**, 1–12 (2020).

22. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).

23. Van Der Wijst, M. G. P. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493–497 (2018).

24. Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. & Murali, T. M. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods 2020 172* **17**, 147–154 (2020).

25. Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology* vol. 38 147–150 at https://doi.org/10.1038/s41587-019-0379-5 (2020).

26. Skinnider, M. A., Squair, J. W. & Foster, L. J. Evaluating measures of association for single-cell transcriptomics. *Nat. Methods* **16**, 381–386 (2019).

27. Sanchez-Taltavull, D. *et al.* Bayesian correlation is a robust gene similarity measure for single-cell RNA-seq data. *NAR Genomics Bioinforma.* **2**, (2020).

28. Bhuva, D. D., Cursons, J., Smyth, G. K. & Davis, M. J. Differential co-expression-based detection of conditional relationships in transcriptional data: Comparative analysis and application to breast cancer. *Genome Biol.* **20**, 236 (2019).

29. McKenzie, A. T., Katsyv, I., Song, W. M., Wang, M. & Zhang, B. DGCA: A comprehensive R package for Differential Gene Correlation Analysis. *BMC Syst. Biol.* **10**, (2016).

30. Hu, H. *et al.* AnimalTFDB 3.0: A comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Res.* **47**, D33–D38 (2019).

31. Bouland, G. A. *et al.* CONQUER: an interactive toolbox to understand functional consequences of GWAS hits. *NAR Genomics Bioinforma.* **2**, (2020).

32. Valencia, R. G. *et al.* Plectin dysfunction in neurons leads to tau accumulation on microtubules affecting neuritogenesis, organelle trafficking, pain sensitivity and memory. *Neuropathol. Appl. Neurobiol.* **47**, 73–95 (2021).

33. Cai, Y. & Wei, Y. H. Stress resistance and lifespan are increased in C. elegans but decreased in S. cerevisiae by mafr-1/maf1 deletion. *Oncotarget* **7**, 10812 (2016).

34. Perluigi, M., Di Domenico, F., Barone, E. & Butterfield, D. A. mTOR in Alzheimer disease and its earlier stages: Links to oxidative damage in the progression of this dementing disorder. *Free Radic. Biol. Med.* **169**, 382–396 (2021).

35. Villa, C., Suphesiz, H., Combi, R. & Akyuz, E. Potassium channels in the neuronal homeostasis and neurodegenerative pathways underlying Alzheimer's disease: An update. *Mech. Ageing Dev.* **185**, 111197 (2020).

36. Piñero, J., Saüch, J., Sanz, F. & Furlong, L. I. The DisGeNET cytoscape app: Exploring and visualizing disease genomics data. *Comput. Struct. Biotechnol. J.* **19**, 2960–2967 (2021).

37. Linde, C. I., Feng, B., Wang, J. B. & Golovina, V. A. Histidine triad nucleotide-binding protein 1 (HINT1) regulates Ca2+ signaling in mouse fibroblasts and neuronal cells via store-operated Ca2+ entry pathway. *Am. J. Physiol. - Cell Physiol.* **304**, C1098 (2013).

38. Kuchibhotla, K. V., Lattarulo, C. R., Hyman, B. T. & Bacskai, B. J. Synchronous hyperactivity and intercellular calcium waves in astrocytes in Alzheimer mice. *Science (80-. ).* **323**, 1211–1215 (2009).

39. Åbjørsbråten, K. S. *et al.* Impaired astrocytic Ca2+ signaling in awake-behaving Alzheimer's disease transgenic mice. *Elife* **11**, (2022).

40. Sato, S., Fujita, N. & Tsuruo, T. Regulation of kinase activity of 3-phosphoinositide-dependent protein kinase-1 by binding to 14-3-3. *J. Biol. Chem.* **277**, 39360–39367 (2002).

41. Popovic, D., Vucic, D. & Dikic, I. Ubiquitination in disease pathogenesis and treatment. *Nat. Med. 2014 2011* **20**, 1242–1253 (2014).

42. Kim, J., Basak, J. M. & Holtzman, D. M. The Role of Apolipoprotein E in Alzheimer's Disease. *Neuron* vol. 63 287–303 at https://doi.org/10.1016/j.neuron.2009.06.026 (2009).

43. Kanekiyo, T., Xu, H. & Bu, G. ApoE and Aβ in Alzheimer's disease: Accidental encounters or partners? *Neuron* vol. 81 740–754 at https://doi.org/10.1016/j.neuron.2014.01.045 (2014).

44. Yin, R. H., Yu, J. T. & Tan, L. The Role of SORL1 in Alzheimer's Disease. *Molecular Neurobiology* vol. 51 909–918 at https://doi.org/10.1007/s12035-014-8742-5 (2015).

45. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).

46. Gatz, M. *et al.* Role of genes and environments for explaining Alzheimer disease. *Arch. Gen. Psychiatry* **63**, 168–174 (2006).

47. Zalocusky, K. A. *et al.* Neuronal ApoE upregulates MHC-I expression to drive selective neurodegeneration in Alzheimer's disease. *Nat. Neurosci.* **24**, 786–798 (2021).

48. Robins, C. *et al.* Genetic control of the human brain proteome. *Am. J. Hum. Genet.* **108**, 400–410 (2021).

49. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* vol. 28 27–30 at https://doi.org/10.1093/nar/28.1.27 (2000).

50. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: Integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**, D545–D551 (2021).

51. Lancour, D. *et al.* Analysis of brain region-specific co-expression networks reveals clustering of established and novel genes associated with Alzheimer disease. *Alzheimer's Res. Ther. 2020 121* **12**, 1–11 (2020).

52. Zhou, Y. *et al.* Human and mouse single-nucleus transcriptomics reveal TREM2-dependent and TREM2-independent cellular responses in Alzheimer's disease. *Nat. Med.* **26**, 131–142 (2020).

53. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

54. Bouland, G. A., Mahfouz, A. & Reinders, M. J. T. Consequences and opportunities arising due to sparser single-cell RNA-seq datasets. *Genome Biol. 2023 241* **24**, 1–10 (2023).

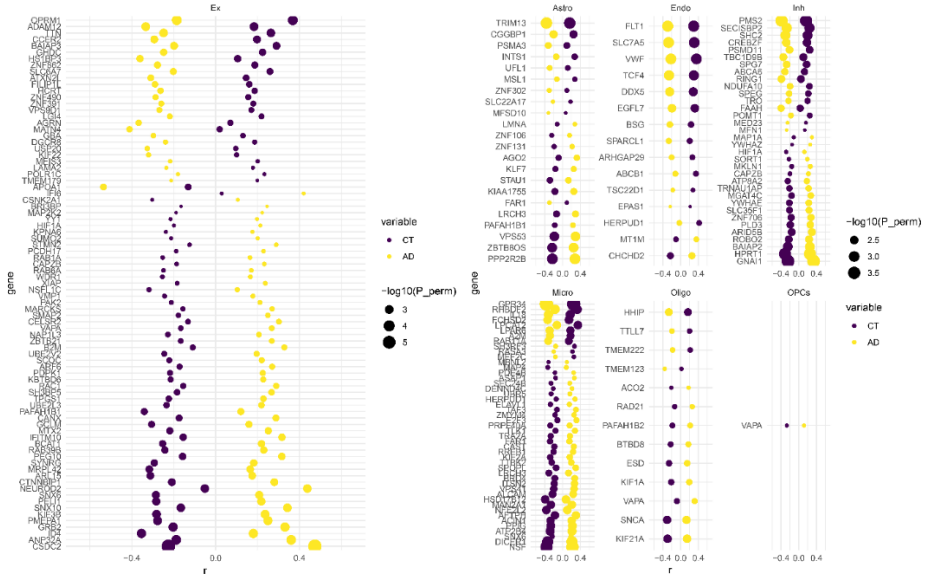
55. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

56. Leek, J. T. *et al.* The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinforma. Appl. NOTE* **28**, 882–883 (2012).

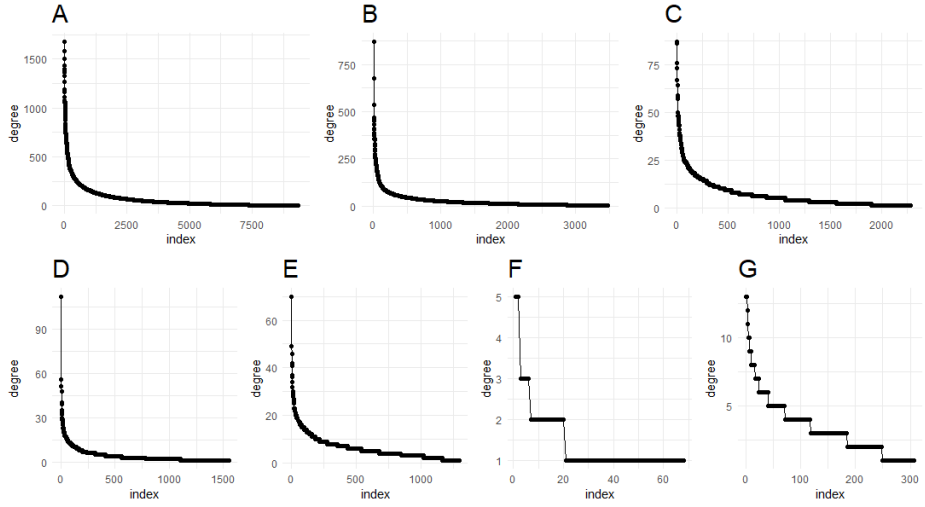
57. Fisher, R. A. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika* **10**, 507 (1915).

58. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *Omi. A J. Integr. Biol.* **16**, 284–287 (2012).

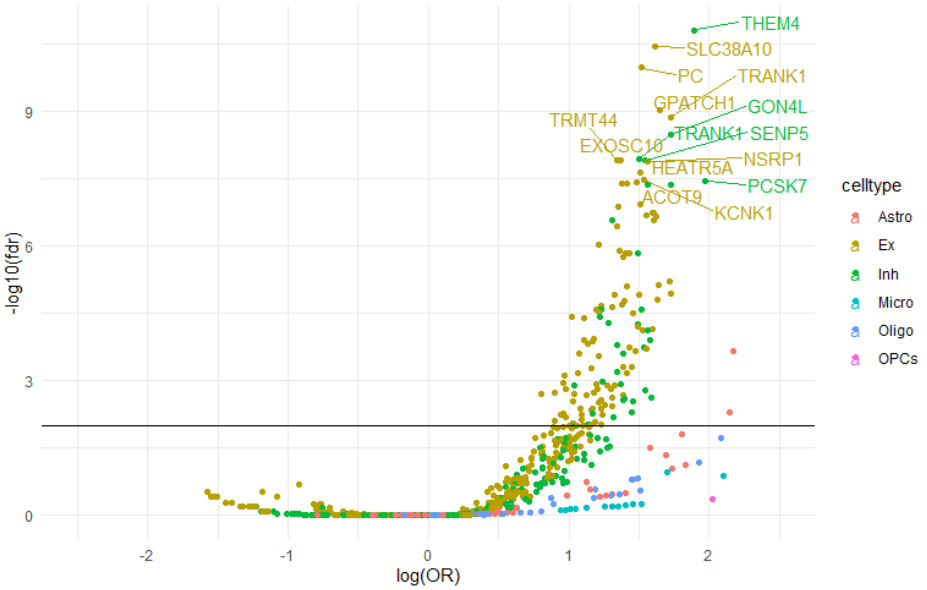
# Supplements



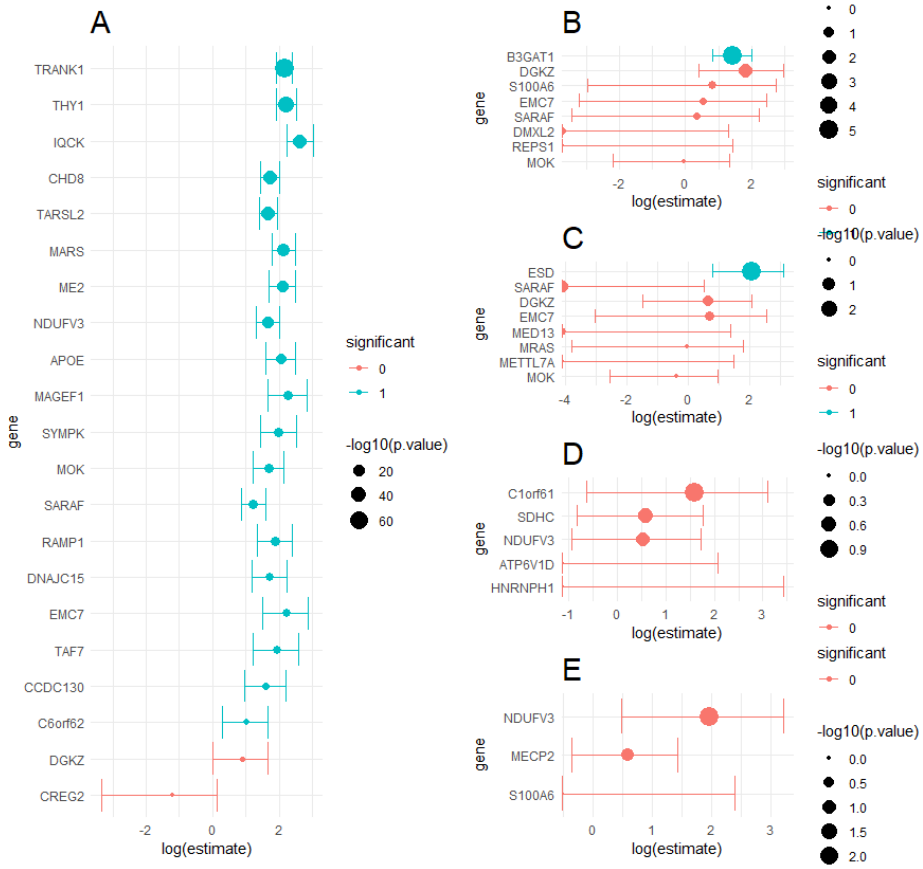

**Supplementary Figure 1**: Differential correlation of genes with the braak stage between CT and AD individuals. X-axis represents the correlation coefficient of the gene with braak stage in the respective group (yellow = AD, purple = CT). Y-axis are the genes.



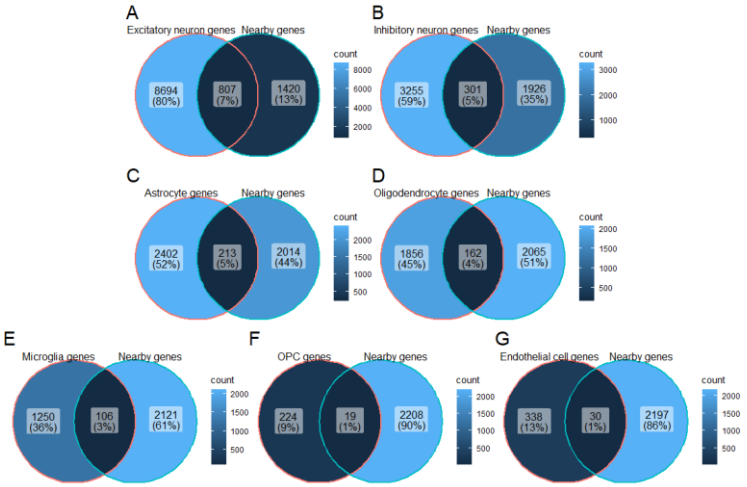

**Supplementary Figure 2:** Degree distribution of differential correlation networks of **A)** excitatory neurons, **B)** inhibitory neurons, **C)** astrocytes, **D)** oligodendrocytes, **E)** microglia, **F)** OPCs **G)** endothelial cells. Every dot is a gene, the x-axis the index of the gene sorted on degree and the y-axis is the degree




**Supplementary Figure 3:** Volcano plot of neighborhood AD enrichment. Every dot represents a hub. The x-axis represents the log odds ratio of the overlap between the hub neighborhood and the KEGG AD pathway. The y-axis represents the –log10 $P_{FDR}$ p-value of the fisher exact test. Colors are the different cell types.




Supplementary Figure 4: Pairwise neighborhood enrichment of shared hubs between A) excitatory neurons and inhibitory neurons, B) inhibitory neurons and astrocytes, C) excitatory neurons and astrocytes, D) excitatory neurons and oligodendrocytes, E) inhibitory neurons and oligodendrocytes. X-axis is the log odds ratio of the overlap between neighborhoods of the respective hub.




**Supplementary figure 5:** Venn diagrams of gene overlap between genes identified nearby AD risk variants and genes expressed in **A)** excitatory neurons, **B)** inhibitory neurons, **C)** astrocytes, **D)** oligodendrocytes, **E)** microglia, **F)** OPCs and **G)** endothelial cells.

# gsQTL: Associating genetic risk variants with gene sets by exploiting their shared variability

Gerard A. Bouland, Niccolò Tesi, Ahmed Mahfouz, and Marcel J.T. Reinders

# Abstract

To investigate the functional significance of genetic risk loci identified through genome-wide association studies (GWASs), genetic loci are linked to genes based on their capacity to account for variation in gene expression, resulting in expression quantitative trait loci (eQTL). Following this, gene set analyses are commonly used to gain insights into functionality. However, the efficacy of this approach is hampered by small effect sizes and the burden of multiple testing. We propose an alternative approach: instead of examining the cumulative associations of individual genes within a gene set, we consider the collective variation of the entire gene set. We introduce the concept of gene set QTL (gsQTL), and show it to be more adept at identifying links between genetic risk variants and specific gene sets. Notably, gsQTL experiences less susceptibility to inflation or deflation of significant enrichments compared with conventional methods. Furthermore, we demonstrate the broader applicability of shared variability within gene sets. This is evident in scenarios such as the coordinated regulation of genes by a transcription factor or coordinated differential expression.


# 3.1. Background

Genome-wide association studies (GWASs) identify genomic loci linked to specific traits, but their impact is hard to understand as the majority of associated loci fall in non-coding and intergenic regions of the genome[1,2]. To determine the functional significance of genetic variants, they are often linked to changes in mRNA expression in bulk RNAseq data (eQTLs)[3] or single cell RNAseq data (sc-eQTLs)[4–7], as well as other molecular data, such as lipids (lipid-QTLs[8]), metabolites (mQTLs[9,10]), and microRNA expression (miQTLs[11]). Once these QTLs are identified, gene set analysis is commonly used to identify affected pathways[12–16]. In the context of complex diseases, this approach typically begins with multiple SNPs, as these are associated with multiple genes, enabling overrepresentation analysis (ORA). However, when focusing on a single variant, this becomes challenging due to the limited number of significantly associated genes, which arises from the burden of multiple testing and the small effect sizes observed in QTL analyses[3,17], especially when associations between a single variant and the whole transcriptome is considered (trans-eQTLs). Another approach is gene set enrichment analysis (GSEA), which in contrast to ORA considers the entire list of genes ranked by their change in expression levels or other relevant metrics (e.g., p-value, effect size) without requiring a predefined threshold for differential expression. It assesses whether a predefined gene set shows significant, consistent differences in expression across the entire ranked list. However, GSEA approaches often yield higher numbers of significantly enriched gene sets compared to ORA approaches, but also have been found to have elevated false positive rates[18]. To address the limitations of functional enrichment analysis of GWAS loci, we propose directly assessing the impact of SNPs on the overall variability of expression of a gene set, contrasting with traditional post-hoc aggregation approaches (e.g. ORA or GSEA). Our gene set QTL (gsQTL) approach starts by combining genes into gene sets based on known interactions (gene sets). Previous studies have demonstrated the

superiority of principal component analysis (PCA) over other popular methods for hidden variable inference for QTL related analyses[19]. However, principal components frequently capture gene expression variance that arises from a mix of various biological and technical sources within a single component[20]. Our approach guides the identification of principal components by calculating them within a highly controlled environment, specifically focusing on predefined gene sets. This approach effectively isolates the biological variance directly associated with the specific biological factor represented by the gene set. Furthermore, unlike other methodologies[21], our components are immediately interpretable due to the inclusion of the biological factor (gene set) that directly links the corresponding genes.

The advantages of this approach for QTL analyses include the ability to capture collective variation within a gene set using PCA, which can lead to stronger associations with risk variants, even when the effect size of the individual genes are modest, as demonstrated in **Fig. 1.** Moreover, the number of gene sets is lower than the number of genes, which reduces the burden of multiple testing correction, and hence increases statistical power. In addition, multiple genes are tested at once for each SNP and these genes do not need to be located near the SNP, allowing for an integration of cis and trans regulatory effects.

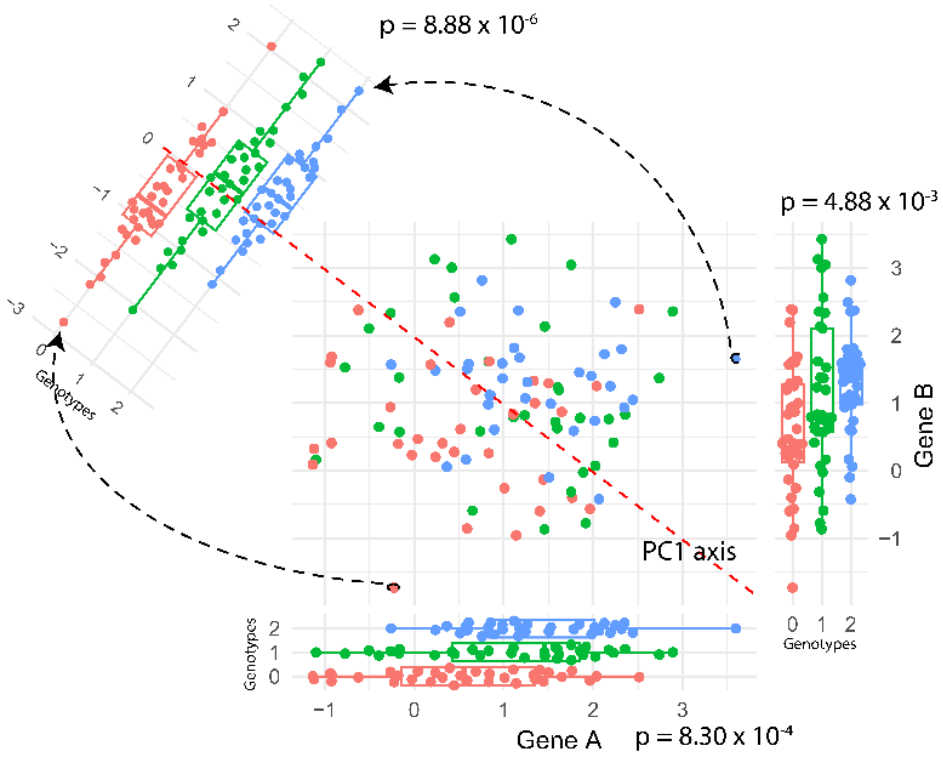



**Figure 1:** *Graphical explanation with simulated data of the proposed gene set QTL (gsQTL) approach*. Illustrated are the effects of three genotypes of a SNP on the expression of Gene A (x-axis) and Gene B (y-axis) that are together forming one gene set, across a set of measured samples. When we inspect the expression of an individual gene for both variations (box plots at bottom and right side of

the figure), there is a significant difference in expression, however, by representing the shared variation of the two genes within the gene set, using the first component of a principle component analysis (PCA), we can observe that the shared variation in the gene set shows a much stronger significant difference in expression when the SNP varies (box plot at diagonal of figure).

## 3.2. Results

Using six jointly pre-processed Alzheimer's Disease (AD) scRNAseq datasets[23–28], including genetics data from 666 individuals (N=249 AD cases, N=242 healthy controls, and N = 175 with mild cognitive impairment with other causes than AD), we demonstrate that gsQTL detects novel functional implications of AD-related SNPs and indeed has more power to detect them than current post-hoc approaches. gsQTL analyses were performed on previously identified AD risk variants[29,30]. Following pre-processing, 44 variants were retained for testing. For seven major brain cell types (excitatory neurons, inhibitory neurons, astrocytes, oligodendrocytes, microglia, OPCs and endothelial cells), we report 66 significant gsQTLs comprised of 30 AD risk variants and 59 unique gene sets, including microRNA targeted genes[31], metabolite interacting genes[32], and KEGG pathway-related gene sets[33] (**Fig. 2a , Sup. Tables 1-3**).

For 53 of the 66 significant gsQTLs, the association of the gene set with the variant had a smaller nominal p-value than the association between the variant and any of the individual genes within the respective gene set (**Supp. Fig 1, Supp Table 4**), showing empirically that the shared variation among genes within a gene set can indeed yield stronger associations with a variant compared to those observed when considering only individual genes of that gene set. Among the significant microRNA gene sets identified, hsa-let-7d-3p ($p=8.03\times10^{-6}$, **Fig. 2b**) and hsa-let-7i-3p ($p=2.36\times10^{-4}$) showed significant associations with rs13237518 (intronic of *TMEM106B*) in excitatory neurons, both of which have been previously implicated in AD pathology[34,35]. Additionally, we identified hsa-miR-6761-3p ($5.90\times10^{-4}$) associated with rs2526377 (intronic of *TSPOAP1*) in OPCs, also a microRNA previously implicated in AD pathology[36]. Further, through the gsQTL analysis considering the gene sets interacting with metabolites, we uncover a potential role of rs1358782 (intronic of *RBCK1*) in pentose phosphate pathway, through the gsQTL in astrocytes with Ribose 5-Phosphate[37] ($p=7.29\times10^{-5}$, **Fig. 2c**). And, among the KEGG pathways, we identified a significant association between taurine and hypotaurine metabolism ($p=5.60\times10^{-4}$, **Fig. 2d**) and rs10933431 (intronic of *INPP5D*) in astrocytes. This finding is particularly interesting given the growing recognition of astrocytes' crucial role in cognitive functioning[38,39] and the neuroprotective properties of taurine, which is believed to enhance cognitive performance[40]. These results suggest that the AD-risk allele of rs10933431 might be involved in a dysregulation of taurine and hypotaurine metabolism, specifically in astrocytes. Collectively, these comparisons and results demonstrate the capacity of gsQTLs to reveal associations with gene sets that may be overlooked when focusing solely on individual genes initially.

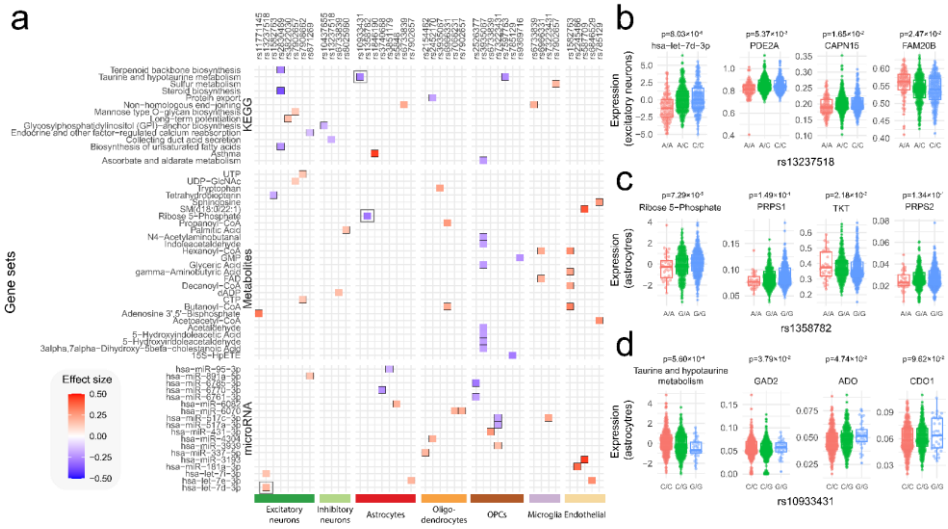



**Figure 2: a,** Significant association between gene sets (rows) and genetic variant (columns) grouped by major cell type (bottom colors) and with the effect size (β, red-blue gradient) of the association between gene set and risk variant. **b,c,d,** Boxplots of a selected set of gsQTL: **b,** hsa-let-7d-3p-rs13237518, **c** Ribose 5-Phosphate-rs1358782 and **d,** taurine and hypotaurine metabolism-rs10933431, with the genotypes (x-axis) of the respective SNPs and the expression levels (y-axis) of the whole gene set as well as three individual genes that belong to the respective gene sets.

To compare gsQTLs with more traditional post-hoc approaches, we considered: 1) overrepresentation analysis (ORA) using the fisher exact test, and 2) gene set enrichment analysis (GSEA) using the R-package fgsea[41]. With ORA, we identified only two gene sets significantly associated with a variant (**Supp. Table5-7**), which is consistent with the limitations of this approach, which is generally more effective when multiple SNPs are analyzed simultaneously. With GSEA, we detected 246 gene sets significantly associated with a variant, including 188 KEGG pathways, 46 metabolite interacting gene sets and 30 microRNA targeted gene sets (**Supp. Table8-10**), comprising 43 unique variants and 81 unique gene sets. Only 5 gene sets-variants-cell type combination overlapped with those identified by our gsQTL method (**Fig. 3a**). Upon closer examination, we observed that the ribosome pathway was linked to 28 distinct variants specifically in excitatory neurons. Despite this, the ribosome pathway did not emerge as significant in our gsQTL analysis, even though 111 out of 127 ribosomal genes showed moderate associations ($p \leq 0.05$) with at least one variant. These ribosomal genes exhibit a high degree of collinearity amongst each other, resulting in the shared variability being captured predominantly by the first principal component in PCA (**Fig. 3b**). But this signal is not strong enough to find an association with any of the variants. As GSEA has been reported to have elevated false positive rates[18], this might be the case here too, suggesting that our gsQTL analysis effectively accounts for the inflation of association signals caused by collinearity among genes.

In our approach, we assume that shared variability among genes reflects a common underlying regulatory mechanism, such as a microRNA targeting a gene set or participation in a KEGG pathway. By analyzing the association between

expression changes in TFs and the coordinated expression shifts in their downstream targets, we aimed to test how broadly our shared variation approach could be applied across different regulatory contexts. More specifically, we computed the correlation between the expression of the TF and the shared representation of the set of genes targeted by that TF. Using data from all individuals (N = 706) in the excitatory neuron dataset and regressing out potential confounding variables (age, sex, dataset and braak stage), we conducted the gsQTL analysis for 256 TFs obtained from the TRUST transcription factor database[42] that overlapped with the measured expression data. We observed that for 42 TFs (16%), their expression patterns correlated with the shared representation of their target genes ($P \leq 4.98 \times 10^{-46}$, $r \geq 0.50$). Interestingly, for 43 of the TFs, the shared representation showed a stronger correlation with the TF than any of the individual TF target genes (**Fig. 3c-d**), underscoring the considerable value of shared maximum variability within a gene set in association analyses.

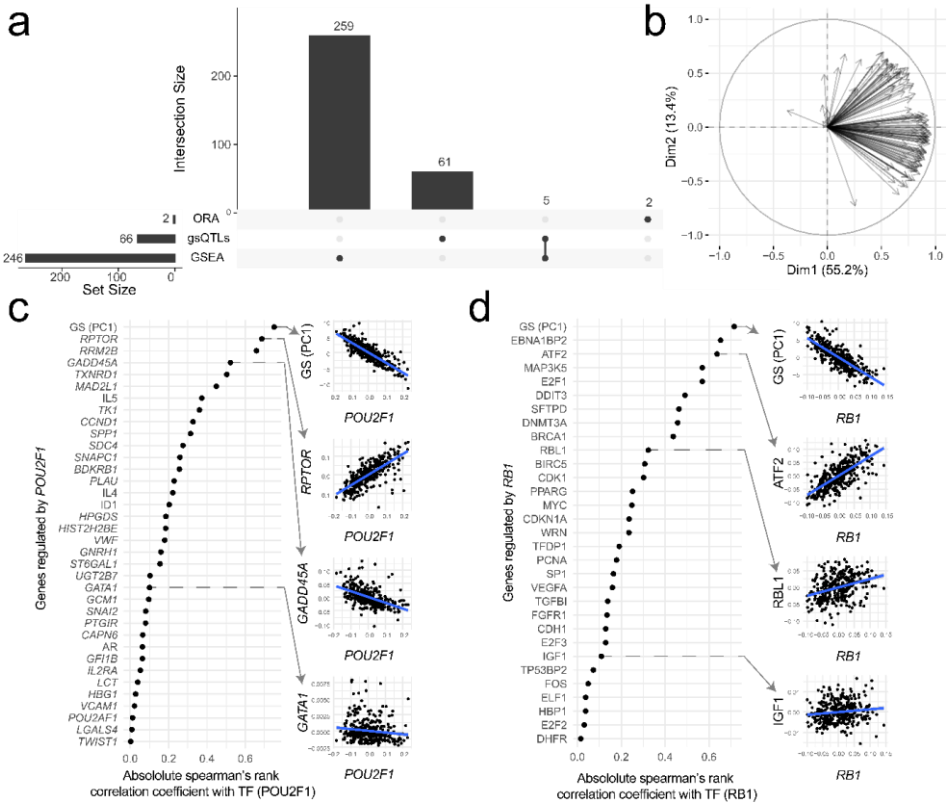



**Figure 3: a,** Overlap between the gene sets identified with ORA, gsQTLs and GSEA. **b,** PCA correlation circle plot of all genes comprising the ribosome KEGG pathway and expressed in our excitatory neuron dataset. The vectors represent the individual ribosomal genes. The closer the vector to each other, the more similar the correlations of the respective genes with respect to the PCs are. **c,d** Absolute correlations between the **c,** TF POU2F1 and **d,** TF RB1 and their targets (x-axis) and the gene set as a whole (GS PC1), with each of the individual targets and the gene set as a whole represented on the y-axis. On the right of the respective plots scatter plots showing the TF

expression(x-axis) and target expressions (y-axis) in excitatory neurons where every dot represents a donor.

To further explore the utility of our shared variation approach, we investigated whether a shared representation of a gene set is more effective in detecting differential expression across conditions compared to traditional methods involving differential expression per gene followed by ORA or GSEA for functional interpretation. For this purpose, we utilized an acute myeloid leukaemia (AML) bulk RNAseq dataset comprising 3,383 individuals[43]. Our aim was to identify metabolic gene sets exhibiting differential expression between AML-subtypes characterized by genetic abnormalities: KMT2A.1 ($N_{individuals}$ = 540), RUNX1_RUNX1T1 ($N_{individuals}$ = 300), FLT3_ITD ($N_{individuals}$ = 661), CBFB_MYH11 ($N_{individuals}$ = 310) and NPM1 ($N_{individuals}$ = 508), TET2 ($N_{individuals}$ = 194). We evaluated 233 metabolome-related gene sets and tested whether their shared representation was differentially expressed between individuals with and without the respective genetic abnormalities. Metabolic gene sets were significantly ($P_{FDR}$ ≤ 0.05) differentially expressed in FLT3_ITD ($N_{gene\ sets}$ = 40), followed by KMT2A.1 ($N_{gene\ sets}$ = 39), CBFB_MYH11 ($N_{gene\ sets}$ = 32), RUNX1_RUNX1T1 ($N_{gene\ sets}$ = 31) and NPM1 ($N_{gene\ sets}$ = 27). Moreover, most metabolic gene sets were differentially expressed in more than two subtypes (**Supp. Fig. 3a**). For example, the most pronounced differential expression between blood and bone marrow was observed for the metabolic gene set related to *phosphatidylcholine* in the KMT2A.1 subtype ($P_{FDR} = 3.77 \times 10^{-53}$, **Supp. Fig. 3b**), which was not detected using the post-hoc analysis approach. Notably, phosphatidylcholine is the product of a reaction catalyzed by genes of the LPC acyltransferase family. *LPCAT1,* a member of this family, is suggested as biomarker to guide treatment choice in AML patients.[44] When applying the ORA approach, only five metabolic gene sets are significantly associated with an AML-subtype, whereas with GSEA, no metabolic gene sets are significantly associated with all AML-subtypes, showing that our shared variability approach identifies associations of genetic AML abnormalities with gene sets that would otherwise go unnoticed.

## 3.3. Conclusion

We present an approach for exploring the functional implications of genetic risk variants on gene expression, biomolecules, and pathways. Rather than focusing on individual gene associations, our method identifies and examines the shared variability of a gene set in relation to genetic risk variants. By defining gene sets based on interactions with various biomolecules such as metabolites or microRNAs, but also based on biological pathways, our approach provides additional context to cell type specific changes associated with genetic risk variants. Our findings underscore that a priori identification of shared variability within gene sets, via PCA, facilitates the discovery of putative coordinated expression changes. These findings can be broadly applied, as demonstrated in our study, for associating with genetic risk variants, coordinating the regulation of genes targeted by a transcription factor, or coordinated differential expression.

## 3.4. Methods

### Single-cell RNAseq data

Five 10x single-cell RNAseq (scRNAseq) datasets were acquired from AMP-AD knowledge portal of which the subjects were participants of the Religious Orders Study and the Memory and Aging Project (ROS/MAP)[45]. The Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD) was obtained from (https://registry.opendata.aws/allen-sea-ad-atlas/). The first dataset (**BA9**, ID: syn16780177) consisted of 24 subjects and originated from the dorsolateral prefrontal cortex (DLPFC), specifically Brodmann area 9 (BA9). Raw fastq files were obtained of this dataset. The second dataset (***BA10***, ID: syn18485175) consisted of 48 subjects and originated from the prefrontal cortex (PFC), specifically BA10. A count matrix was obtained of this dataset as it was already processed with CellRanger aligning reads to the hg38 genome[25]. The third dataset (**TREM**, ID: syn18485175) consisted of 32 subjects and originated from the DLPFC, BA9 and BA46. Of this dataset a count matrix was obtained as it was also processed with CellRanger aligning reads to the hg38 genome[27]. The fourth dataset (**DLPFC2**: ID: syn31512863) consisted of 424 individuals. Of this dataset also a count matrix was obtained as it was also processed with CellRanger aligning reads to the hg38 genome[24]. The fifth dataset (**MIT:** ID: syn52293433) consisted of 427 individuals. Of this dataset also a count matrix was obtained as it was also processed with CellRanger aligning reads to the hg38 genome[28]. The sixth dataset (**SEA-AD**) consisted of 89 middle temporal gyrus samples, 23 of which were diagnosed with AD and 32 were specified as CT[23]. Of this dataset the raw count matrix was acquired.

### Clinical data

Clinical data were acquired from the AMP-AD knowledge portal (ID: syn3157322). The variable cogdx was used to characterize controls (CT), Alzheimer's disease (AD) and other (O). Cogdx represents the clinical consensus diagnosis of cognitive status at time of death and is indicated with a value ranging from one to six. A value of one represents no cognitive impairment (CI), as such, individuals with a cogdx of one were characterized as CT. A value of four represents Alzheimer's dementia and no other cause of CI, as such, these individuals were characterized as AD. The remaining values represent mild CI and/or other causes for dementia and these individuals were characterized as O. Besides clinical diagnosis, *APOE* genotype, Braak stage, sex, and age at time of death was also available. However, age at time of death is censored above the age 90 years. Of the **SEA-AD** datasets the clinical data were acquired from the corresponding source. For both datasets; age, sex, clinical diagnosis and Braak stage were available.

### Genetics data

Genotyping data were sourced from the Synapse AD portal and consisted of 3 batches. Batch 1 and batch 2 (SynID: syn17008939) included 1709 and 382 individuals, respectively, from the ROSMAP study[45]. Batch 3 (SynID: syn28257618) included 95 samples from SEA-AD study[23]. Batch 2 and batch 3

data were aligned to GRCh37 (hg19), while batch 1 data was aligned to GRCh36 (hg18) and lifted over to GRCh37 (hg19). Standard quality control was applied to each batch independently (variant call rate >98%, individual call rate >98%, and deviation from Hardy-Weinberg was considered significant at $p<1e^{-6}$). Variant ID, strand, and allele frequencies were compared, for each batch, to the Haplotype Reference Consortium (HRC, HRC-1000G-check-bim-v4.2.7.pl)[46]. Genotyping data were combined and high-quality genotyping was ensured (variant call rate >98%, individual call rate >98%). All autosomal variants were submitted to the TOPMED imputation server (https://imputation.biodatacatalyst.nhlbi.nih.gov). The server uses Eagle (v2.4) to phase data and imputation to the reference panel (TOPMED R2 v1.0) was performed with Minimac4[47–49]. A total of 2,115 individuals passed quality control. Prior to analysis, we extracted individuals for which scRNA data was also available, leaving 527 individuals (N=171 AD cases and N=184 healthy controls and N=172 specified as O (other)) for analyses. For these individuals, we further selected variants known to associate with AD from previous GWAS[29,30]. Only variants for which all three genotypes were present in at least 5% of the total population were tested. Quality control of genotype data was performed with PLINK (v1.90b4.6 and v2.00a2.3)[50,51]. Liftover of the genetics data was performed with liftOver R-package (v1.10).

## Bulk AML RNAseq data

The bulk AML RNAseq datast was obtained from Severens, et al[43]. This dataset consisted of 3,656 individuals and 60,660 transcripts. First, individuals having ≤ 20,000 or ≥35,000 zero measurements were removed. Next, genes that were measured in less than 90% of the individuals were removed. The resulting matrix (23,418 genes x 3,383 individuals) was normalized using median ratio normalization[52]. The dataset was comprised of five different source datasets, as such, batch correction was done using Combat from the R-package sva(v 3.46.0)[53]. BiomaRt[54] was used to translate the ensembl gene IDs to HGNC gene symbols.

## Cell type annotation

The DLPFC2 dataset was used to identify marker genes as it was already annotated. For excitatory neurons (n = 3.154), inhibitory neurons (n = 457), astrocytes (n = 456), oligodendrocytes(n = 283), microglia (646), OPCs (n = 274) and endothelial cells (n = 517) markers were identified. First, pseudo bulk data was generated for each cell type and concatenated, resulting in a gene by individual matrix, in which each individual is present seven times (one for each cell type). Then, for each cell type a differential expression analyses, using Wilcoxon-rank sum test, was performed where the groups were defined by whether the measurement was from the respective cell type (group 1) or not (group 2). A gene was considered a marker gene when $P \leq 5 \times 10^{-10}$ and $\log_2$ fold-change ≥ 3. Next, with these markers the cells from the TREM, BA10 and BA9 datasets were annotated. This was done using the AddModuleScore function from Seurat[55], which assigns a score to every cell for each cell type based on the expression of cell type markers. Each cell was annotated as the cell type for which the score was the highest. However, if the second highest score was within

25% of the highest score, it was annotated as hybrid and removed for subsequent analyses.

## Generating pseudo bulk data

For each dataset, for each cell type, pseudo bulk was generated. Aggregation was done based on the binary expression pattern, since the percentage of zeros for a gene in a cell population is highly associated with its mean expression[56] and aggregating based on the percentage of zero results in less false positive in downstream analyses opposed to aggregating based on the mean[57]. Next, for each cell type, all the datasets were combined, and genes expressed in less than 10% of the individuals were removed. Then, the expression was normalized using median ratio normalization[58] and batch correction was performed using ComBat from the sva R-package (v3.46.0)[53]. The DLPFC2 dataset consisted of 60 batches and the other datasets were each considered a batch, as such, in total there were 64 batches. Batch correction was confirmed with a PCA and visually inspecting the principal components and visually inspecting boxplots of the individuals' gene expressions. For the endothelial cells only the DLPFC2 dataset was used.

## Shared variability representation of gene sets

Starting with a gene-by-sample expression matrix we first subset the genes that belong to a specific gene set. Using the subsetted gene-by-sample matrix we first scale each gene such that the mean = 0 and the standard deviation = 1, then we perform a principal component analysis using the prcomp function from the stats R-package. If the first principal component explains ≥10% of the total variance of the gene set, and the gene set is comprised of at least 5 genes, then we store the principal component in the new gene set-by-sample matrix (**Supp. Fig 2**).

## Gene sets

We used four different gene sets. Three of these gene set databases (KEGG[33], TRRUST Transcription Factors[42] and Metabolomics Workbench[32]) we downloaded from the webserver of enrichR[59]: https://maayanlab.cloud/Enrichr/#libraries. The microRNA database was downloaded from miRTarBase[31], where we only considered functionally validated microRNA targets.

## QTL analyses

QTL identification was performed using a linear model which was defined as follows:

$$E_{\boldsymbol{gs},k} = \beta_0 G_k + \sum_i \beta_i cov_{i,k} + \varepsilon_k \qquad (1)$$

where $E_{gs}$ is the shared variable representation of gene set *gs* (when performing gsQTL), or, alternatively, the gene expression of a single gene (when performing eQTL), for an individual *k*. $G_k$ is the genotype dosage of an individuals of the SNP of study (0-2) . $Cov_{i,k}$ are the covariates for each of the individuals. $\varepsilon_k$ is an error term that gets minimized. In both, the GSQTL and eQTL analyses, we adjusted for age, sex, diagnosis, dataset, and the first five gene expression PCs. For the

bulk AML analyses, *G* represents subtype assignment (0-1) and we adjusted for sex and tissue of origin (bone marrow and peripheral blood).

## Transcription factor validation experiment

Using the healthy individuals and the excitatory neuron dataset, we evaluated the association between the expression of transcription factors (TFs) and shared representation derived from the genes targeted by the respective TF. Targeted genes were derived from the trust transcription factor database[42]. We only evaluated TFs that were present in both the trust database as well as in our expression data. Potential covariance caused by effects of age, sex, dataset and Braak stage on the gene expression were regressed out. We only considered TF targeted gene sets, when there were minimally five target genes, and the shared variability had to explain at least 5% of the total variance of the gene set. The association between the expression of the TF and the expression in the shared variable representation was evaluated using Spearman's rank correlation coefficient. The resulting p-values were adjusted for testing multiple TF target gene tests using the Benjamini-Hochberg Procedure, and significance of the association was assumed at $P_{FDR} \leq 0.05$.

## ORA: Over representation analysis

Overrepresentation analysis (ORA) was performed using the fisher exact test for each variant separately including the genes that are nominally significant, i.e. eQTLs for which $P \leq 0.05$. The Benjamini-Hochberg Procedure was used to corrects for testing multiple gene sets, and a gene set was assumed significant at $P_{FDR} \leq 0.05$.

## GSEA: Gene set enrichment analysis

Gene set enrichment analysis (GSEA) was performed using the R-package fgsea[41] (v 1.24.0). For each variant, the βs of the associations with the genes were used to rank the genes, that served as input for the gene set enrichment analysis. The resulting p-values were adjusted for testing multiple tests per variant using the Benjamini-Hochberg Procedure, and significance of the association was assumed at $P_{FDR} \leq 0.05$.

# References

1. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 9362–9367 (2009).

2. Smigielski, E. M., Sirotkin, K., Ward, M. & Sherry, S. T. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.* **28**, 352–355 (2000).

3. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* vol. 45 580–585 at https://doi.org/10.1038/ng.2653 (2013).

4. Neavin, D. *et al.* Single cell eQTL analysis identifies cell type-specific genetic control of gene expression in fibroblasts and reprogrammed induced pluripotent stem cells. *Genome Biol.* **22**, 1–19 (2021).

5. Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**, (2022).

6. Van Der Wijst, M. G. P. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493–497 (2018).

7. van der Wijst, M. G. P. *et al.* The single-cell eQTLGen consortium. *Elife* **9**, (2020).

8. Tabassum, R. *et al.* Genetic architecture of human plasma lipidome and its link to cardiovascular disease. *Nat. Commun. 2019 101* **10**, 1–14 (2019).

9. Rhee, E. P. *et al.* A Genome-wide Association Study of the Human Metabolome in a Community-Based Cohort. *Cell Metab.* **18**, 130–143 (2013).

10. Gallois, A. *et al.* A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context. *Nat. Commun. 2019 101* **10**, 1–13 (2019).

11. Huan, T. *et al.* Genome-wide identification of microRNA expression quantitative trait loci. *Nat. Commun. 2015 61* **6**, 1–9 (2015).

12. Bouland, G. A. *et al.* Diabetes risk loci-associated pathways are shared across metabolic tissues. *BMC Genomics 2022 231* **23**, 1–9 (2022).

13. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun. 2015 61* **6**, 1–9 (2015).

14. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun. 2017 81* **8**, 1–11 (2017).

15. Xu, T., Jin, P. & Qin, Z. S. Regulatory annotation of genomic intervals based on tissue-specific expression QTLs. *Bioinformatics* **36**, 690–697 (2020).

16. Wang, B., Yang, J., Qiu, S., Bai, Y. & Qin, Z. S. Systematic Exploration in Tissue-Pathway Associations of Complex Traits Using Comprehensive eQTLs Catalog. *Front. Big Data* **4**, 719737 (2021).

17. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nat. Genet. 2023 5511* **55**, 1866–1875 (2023).

18. Geistlinger, L. *et al.* Toward a gold standard for benchmarking gene set enrichment analysis. *Brief. Bioinform.* **22**, 545–556 (2021).

19. Zhou, H. J., Li, L., Li, Y., Li, W. & Li, J. J. PCA outperforms popular hidden variable inference methods for molecular QTL mapping. *Genome Biol.* **23**, 1–17 (2022).

20. Zhernakova, D. V. *et al.* Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet. 2016 491* **49**, 139–145 (2016).

21. Vochteloo, M. *et al.* PICALO: principal interaction component analysis for the identification of discrete technical, cell-type, and environmental factors that mediate eQTLs. *Genome Biol.* **25**, 1–26 (2024).

22. Tomfohr, J., Lu, J. & Kepler, T. B. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics* **6**, 1–11 (2005).

23. Gabitto, M. *et al.* Integrated multimodal cell atlas of Alzheimer's disease. *Res. Sq.* (2023) doi:10.21203/RS.3.RS-2921860/V1.

24. Fujita, M. *et al.* Cell-subtype specific effects of genetic variation in the aging and Alzheimer cortex. *bioRxiv* 2022.11.07.515446 (2022) doi:10.1101/2022.11.07.515446.

25. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).

26. Olah, M. *et al.* Single cell RNA sequencing of human microglia uncovers a subset associated with Alzheimer's disease. *Nat. Commun.* **11**, 1–18 (2020).

27. Zhou, Y. *et al.* Human and mouse single-nucleus transcriptomics reveal TREM2-dependent and TREM2-independent cellular responses in Alzheimer's disease. *Nat. Med.* **26**, 131–142 (2020).

28. Mathys, H. *et al.* Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to Alzheimer's disease pathology. *Cell* **186**, 4365-4385.e27 (2023).

29. Wightman, D. P. *et al.* A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat. Genet. 2021 539* **53**, 1276–1282 (2021).

30. Bellenguez, C. *et al.* New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat. Genet. 2022 544* **54**, 412–436 (2022).

31. Huang, H. Y. *et al.* miRTarBase update 2022: an informative resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res.* **50**, D222–D230 (2022).

32. Sud, M. *et al.* Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* **44**, D463–D470 (2016).

33. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: Integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**, D545–D551 (2021).

34. Dakterzada, F. *et al.* Identification and validation of endogenous control miRNAs in plasma samples for normalization of qPCR data for Alzheimer's disease. *Alzheimer's Res. Ther.* **12**, 1–8 (2020).

35. Leidinger, P. *et al.* A blood based 12-miRNA signature of Alzheimer disease patients. *Genome Biol.* **14**, R78 (2013).

36. Lu, L., Dai, W. Z., Zhu, X. C. & Ma, T. Analysis of Serum miRNAs in Alzheimer's Disease. *Am. J. Alzheimers. Dis. Other Demen.* **36**, (2021).

37. Palmer, A. M. The activity of the pentose phosphate pathway is increased in response to oxidative stress in Alzheimer's disease. *J. Neural Transm.* **106**, 317–328 (1999).

38. Santello, M., Toni, N. & Volterra, A. Astrocyte function from information processing to cognition and cognitive impairment. *Nat. Neurosci.* **22**, 154–166 (2019).

39. Mathys, H. *et al.* Single-cell multiregion dissection of Alzheimer's disease. *Nat. 2024* 1–11 (2024) doi:10.1038/s41586-024-07606-7.

40. Rafiee, Z., García-Serrano, A. M. & Duarte, J. M. N. Taurine Supplementation as a Neuroprotective Strategy upon Brain Dysfunction in Metabolic Syndrome and Diabetes. *Nutrients* **14**, (2022).

41. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv* 060012 (2021) doi:10.1101/060012.

42. Han, H. *et al.* TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* **46**, D380–D386 (2018).

43. Severens, J. F. *et al.* Mapping AML heterogeneity - multi-cohort transcriptomic analysis identifies novel clusters and divergent ex-vivo drug responses. *Leukemia* **38**, 751–761 (2024).

44. Wang, K. *et al.* Identification of LPCAT1 expression as a potential prognostic biomarker guiding treatment choice in acute myeloid leukemia. *Oncol. Lett.* **21**, 1–1 (2021).

45. Bennett, D. A. *et al.* Religious Orders Study and Rush Memory and Aging Project. *J. Alzheimers. Dis.* **64**, S161–S189 (2018).

46. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

47. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

48. Loh, P. R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet. 2016 4811* **48**, 1443–1448 (2016).

49. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation. *Bioinformatics* **31**, 782–784 (2015).

50. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559 (2007).

51. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

52. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol. 2014 1512* **15**, 1–21 (2014).

53. Leek, J. T. *et al.* The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinforma. Appl. NOTE* **28**, 882–883 (2012).

54. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinforma. Appl. NOTE* **21**, 3439–3440 (2005).

55. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

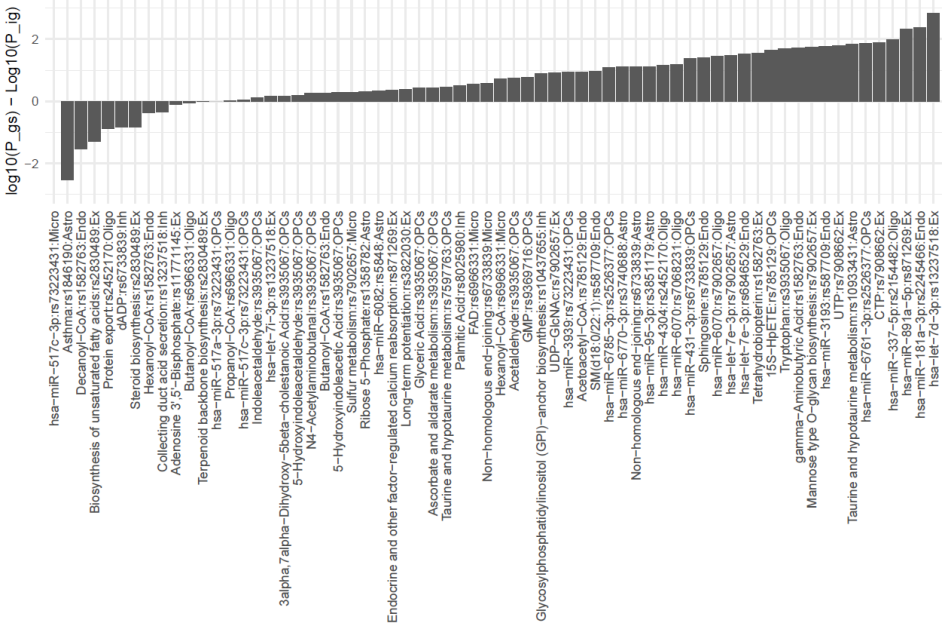
56. Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology* vol. 38 147–150 at https://doi.org/10.1038/s41587-019-0379-5 (2020).

57. Bouland, G. A., Mahfouz, A. & Reinders, M. J. T. Consequences and opportunities arising due to sparser single-cell RNA-seq datasets. *Genome Biol. 2023 241* **24**, 1–10 (2023).

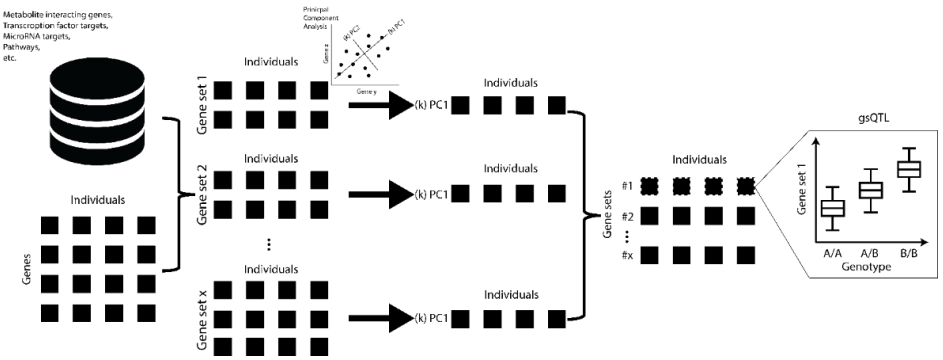
58. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

59. Xie, Z. *et al.* Gene Set Knowledge Discovery with Enrichr. *Curr. Protoc.* **1**, e90 (2021).

# Supplements

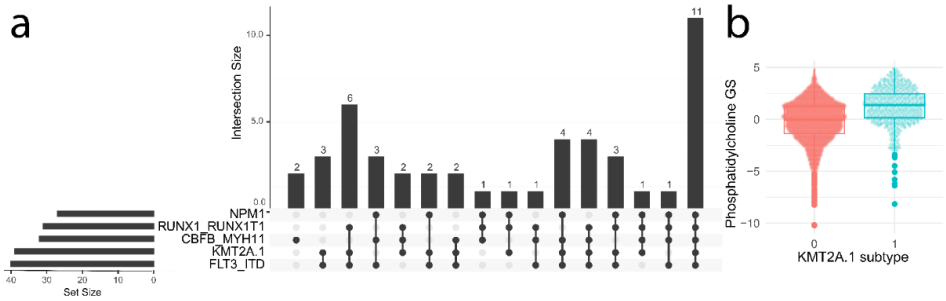



**Supplementary Figure 1:** A bar plot where the x-axis represent significantly identified gsQTLs and the y-axis represents the difference between the -log10(p-value) of the gsQTL and the -log10(p-value) of the most significant individual gene that was part of the respective gene set used to calculate the proxy values.




**Supplementary Figure 2:** Visual representation of the approach. First, we take as input a gene expression matrix where the rows represent the genes and the columns represent the individuals or samples. Then, for the selected gene sets (e.g. pathways, metabolite interacting genes, transcription factor or microRNA targets), the respective genes are subsetted from the input dataset. For each gene set PCA is performed, either a linear PCA or a kernel-PCA. Next, PCs are filtered based on the

number of input genes and based on percentage variance explained. Finally, QTL analyses are performed on the PCs.




**Supplementary Figure 3: A)** Upset plot of the overlap of significantly identified proxy metabolites between the AML subtypes. **B)** Boxplot of the phosphatidylcholine GS-metabolite. X-axis represents KMT2A.1 status where 0 is negative and 1 is positive. Y-axis represents the expression value.

# 4

# Genetic Variants that Modulate Alzheimer's Disease Risk Deregulate Protein-Protein Correlations in the Gyrus Temporalis Medius

Gerard A. Bouland, Niccolò Tesi, Meng Zhang, Andrea B. Ganz, Marc Hulsman, Sven van der Lee, Marieke Graat, Annemieke Rozemuller, Martijn Huisman, Natasja van Schoor, Wiesje van der Flier, Jeroen Hoozemans, August B Smit, Marcel J.T. Reinders and Henne Holstege

# Abstract

Through a comprehensive protein quantitative trait loci (pQTL) analysis, we identified 8,081 genetic variants linked to the abundance of 227 proteins in the Gyrus Temporalis Medius (GTM). This includes novel associations between variants and proteins, not found in a previous pQTL study in the prefrontal cortex and from expression quantitative loci (eQTL) analyses across 12 brain areas. We observed a link between the rs429358-T variant, known for encoding the APOE4 allele, and increased APOE levels in the GTM, pointing to a potential explanation for GTM's greater vulnerability to Alzheimer's Disease (AD). We show that AD risk variants deregulate protein-protein correlations, providing a genetic basis for coordinated modulation of protein associations. Specifically, significant effects on protein interactions in the GTM are found for three SNPs: rs9381040 in TREML2, rs34173062 in SHARPIN, and rs11218343 near SORL1. Notably, DDX17 may exert a protective role in individuals with the rs9381040-T/T genotype by tightly regulating synuclein levels.


# 4.1. Introduction

Alzheimer's Disease (AD) is a progressive disease marked by the loss of cognitive functions and autonomy, eventually leading to death[1]. Numerous genome wide association studies (GWASs) have been conducted to identify genetic modifiers of AD risk, including attempts to understand their role in AD etiology[2–4]. However, elucidating the mechanistic pathways through which these genetic loci influence AD-related processes is difficult. Many risk variants are common, have a small effect size on AD, and are located in non-coding and intergenic regions. These variants merely act as 'markers' of haplotypes, genetic regions averaging 300 kb that are inherited across generations[5,6]. These haplotypes typically include at least one genetic factor that increases disease risk, possibly by affecting the expression of one or more genes in the risk locus. Therefore, while GWAS 'risk' variants are unlikely to be directly causative, they are likely in (partial) linkage with the causal variant(s).

It is thus reasonable to investigate whether a risk locus is an expression quantitative trait locus (eQTL), meaning it is associated with changes in the expression of messenger RNAs (mRNAs)[7]. However, while mRNAs encode proteins, their expression levels are not always correlated with protein expression levels[8–10]. Consequently, eQTLs are often not protein-QTLs (pQTLs)[11], while proteins are the functional units within the cell. Understanding how AD risk alleles associate with protein expression could therefore better pursued to reveal the genetic (de)regulation associated with the risk alleles. Additionally, causative variants might have downstream effects that are missed when only investigating eQTLs/pQTLs. For instance, while a causative missense variant might not alter mRNA levels, the change in the amino acid sequence of a protein can lead to a significant change in protein function and interactions with other proteins, potentially altering various biological pathways. Assuming that co-expressed proteins are functionally related, allele-specific correlation patterns of protein abundance might indicate a unique regulatory state of biological pathways in response to an allele.

Based on this, we hypothesize that AD-associated alleles may have downstream functional consequences, identifiable by comparing changes in co-expression of protein abundance between carriers and non-carriers of risk alleles. In this study, we investigated genetic control of protein abundance using a unique collection of brains from 88 AD patients, 53 non-demented individuals, and 49 cognitively healthy centenarians. The Gyrus Temporalis Medius (GTM) region, known for its vulnerability to AD-related neuropathological changes[12,13], was specifically analyzed, enabling us to conduct a comprehensive pQTL analysis and compare changes in co-expression between pairs of proteins between individuals who carry an AD-risk allele compared to those the protective risk allele.

# 4.2. Results

## 4.2.1. Analysis workflow

We performed an extensive pQTL analysis to assess whether the abundance of measured proteins in the GTM correlated with the occurrence of genetic variants (Fig. 1a). These identified pQTLs were then compared with previously identified brain pQTLs[11] and eQTLs across twelve brain areas in the GTEx database[7]. Furthermore, all genetic variants identified as pQTL were evaluated for their association with AD risk. We then explored whether AD risk variants identified in a previous GWAS[2] were linked to distinctive correlation-structures in protein abundance, by conducting a differential correlation QTL analysis (dcQTL, Fig. 1b).

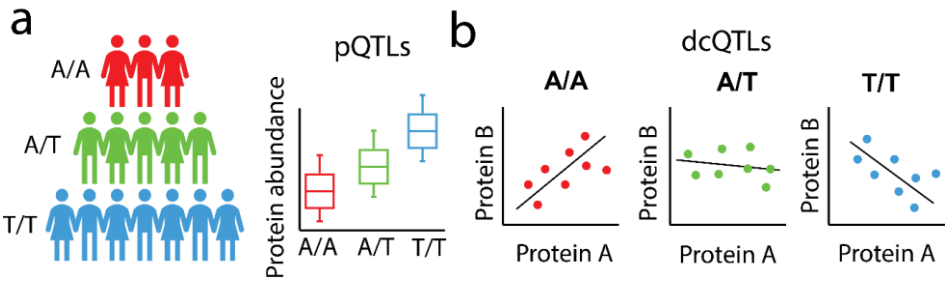



**Figure 2:** Overview of the pQTL and dcQTL analyses within this study. a) Schematic representation of a pQTL analysis. Individuals are grouped according to the genotypes of a genetic variant. A pQTL is identified when the expression of a protein is significantly (linearly) associated with the genotype. b) Schematic overview of the dcQTL analysis, identifying changes in co-expression relative to genotype. Individuals are grouped based on the genotypes of a genetic variant and pairs of proteins are identified whose correlation within each genotype differs between genotypes ($P_{FDR} \leq 0.05$).

## 4.2.2. Demographics

After quality control of the genetic data, 6,607 individuals were included in the analyses (mean age of 68.4±15.8, 53.7% females). For the protein expression data in the GTM, 190 individuals were available (mean age of 86.8±13.8, 73.7% females). Both genetics and proteomics data were available for 140 individuals (mean age of 91.0±14.2, 74.3% females, Table 1).

**Table 1:** Population characteristics of individuals for whom genotyping and protein data were available, including the intersection. AD = Alzheimer's Disease cases, ND = Non-demented controls, CEN = centenarians.

| | genotyping data | | | Protein data | | | intersection | | |
|---|---|---|---|---|---|---|---|---|---|
| Number of individuals | 6,607 | | | 190 | | | 140 | | |
| Females (%) | 3549 (53.7) | | | 140 (73.7) | | | 104 (74.3) | | |
| Diagnosis | AD | ND | CEN | AD | ND | CEN | AD | ND | CEN |
| N | 2,416 | 3,848 | 343 | 88 | 53 | 49 | 67 | 27 | 46 |
| Age (σ) | 70.2 (10.5) | 61.1 (14.8) | 101.0 (2.5) | 81.2 (11.2) | 81.1 (12.0) | 103.0 (2.3) | 79.6 (12.2) | 83.4 (8.7) | 103.1 (2.2) |

## 4.2.3. Genetic modulation of protein abundance is largely independent from modulation of its corresponding mRNA in the Gyrus Temporalis Medialis

A pQTL analysis (Fig. 1a) including 140 individuals for which both genetic and proteomics data was available, was conducted (Table 1). pQTLs were identified using linear regression models, adjusting for estimated cell type composition (neurons, microglia/macrophages, and oligodendrocytes) and population substructure using the first five principal components (Supplement '*pQTL linear regression model'*). A total of 3,427 proteins were tested for association with genetic cis-variants lying within 250 Kbp of the TSS. We identified 8,081 variants significantly associated with the abundance of 227 proteins in the GTM ($P_{FDR} \leq 0.05$, Fig. 2a). Of these, 5,331 (~66%) are new associations involving 150 proteins, while 2,750 variants (34%), where associated with 77 proteins, overlapping with pQTLs previously identified in a QTL study of the dorsolateral prefrontal cortex[11].

Next, we identified a set of 222 independent pQTL variants by prioritizing the most significant protein-variant pair within a linkage disequilibrium (LD) block (500 Kbp, R2>0.001). 64 of these 222 pQTL variants (29%) were also eQTL variants in at least one of the twelve brain areas available in the GTEx Portal[7](Supp Table 1), and sixteen pQTLs variants being eQTL variants in all twelve brain areas (Fig. 2b). For the matching pQTLs and eQTLs, the direction of change in protein levels corresponded with the direction of change in gene transcript levels (Supp Fig. 1). Additionally, the effect sizes of matching pQTLs and eQTLs were significantly correlated across all brain regions ($P \leq 6.39 \times 10^{-3}$, $r \geq 0.46$).

This indicates that while genetic modulation of protein abundance is largely independent from genetic modulation of mRNA levels, when shared modulation occurs, mRNA levels do correlate with protein abundances.

## 4.2.4. rs429358 and rs6857 associate with increased APOE abundance and increased Alzheimer's risk

Next, to elucidate the potential connection between AD-genetic risk factors and dysregulation of protein abundance, we investigated whether the identified pQTLs are associated with AD risk. Using an independent dataset of individuals with genetics data[14–18] (N = 6,479, N cases = 2,361 , N controls = 4,118 see

Methods), we checked whether the identified 8,081 pQTL variants were associated with AD risk. We found a significant association with AD for rs6857 and rs429358 (Fig. 2c, Supp. Fig. 3). Rs6857 had an odds-ratio (OR) of 3.22 ($P_{FDR} = 6.0 \times 10^{-145}$) and rs429358 had an OR of 3.56 ($P_{FDR} = 1.5 \times 10^{-167}$) for AD risk. Both variants were pQTLs associated with *APOE* abundance. Rs6857 had a β of 0.18 (SE = 0.04, $P_{FDR} = 2.57 \times 10^{-5}$, Fig. 2d) and rs429358 had a β of 0.18 (SE = 0.04, $P_{FDR} = 2.80 \times 10^{-5}$, Fig. 2e) associated with APOE abundances.

A Bayesian test of colocalization (see Methods *Colocalization analysis*) revealed a posterior probability (PP) of 77% that *APOE* abundance and AD risk share the same causal variant, with rs429358 being the most likely variant (PP = 1). Notably, the rs429358-T variant encodes for arginine, resulting in the APOE4 allele[19], which is known to increase AD risk. Our findings show that the rs429358-T variant is linked to higher APOE levels in the GTM, a result not observed in prior pQTL studies of the prefrontal cortex. This may help explain the greater vulnerability of the GTM to AD, as elevated APOE4 levels could potentially lead to increased amyloid-beta accumulation specifically in this brain area.

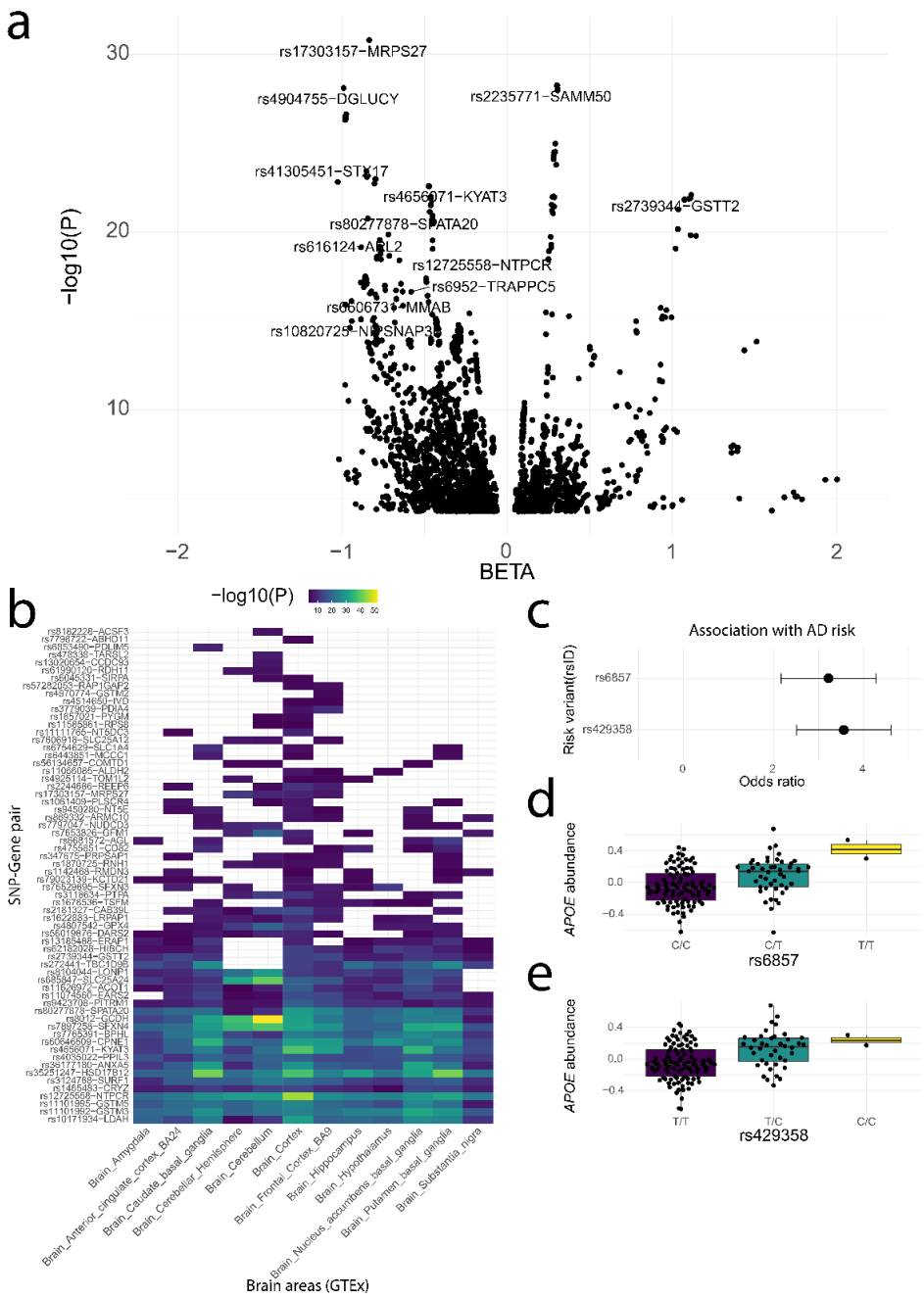



**Figure 3** Overview of found significant pQTLs. a) Volcano plot of significant pQTLs ($P_{FDR} \leq 0.05$). X-axis represents the effect size of the pQTL association, and the y-axis represents -$\log_{10}$ P-value of the pQTL association. Labels are shown for pQTLs with P-value ≤ $1 \times 10^{-15}$. b) Heatmap of all significant pQTL variants that are also an eQTL variant. Colored squares indicate significant eQTLs based on the GTEx database, with color representing the -$\log_{10}$ P-value. The x-axis shows the tested brain tissues, and the y-axis the protein/gene names. c) Association of variants with AD risk. The x-

axis represents the odd ratio (with error bars), and the y-axis shows the two significant variants. d,e) Boxplots of APOE protein residuals grouped by genotypes of d) rs6857, and e) rs429358. The x-axis represents the genotypes of rs6857/rs429358, and the y-axis represents the residuals of APOE protein residuals.

## 4.2.5. Genetic risk variants for Alzheimer's Disease associate with changing associations between proteins

In addition to examining whether genetic risk variants affect the abundance of individual proteins, we explored whether AD risk variants influence the correlation between proteins, potentially indicating a genetic basis for coordinated changes in their functional associations. To investigate this, we tested for differential correlation between pairs of proteins with respect to the genotypes of 33 known AD risk variants[2]. These variants were selected such that each genotype was represented by at least 10 individuals to mitigate population size discrepancies and reduce the likelihood of false positives (Supp Table 2). Differential correlation QTL (dcQTL) analysis (see Methods) was performed on the 140 individuals for whom both genetic and proteomics data were available. We identified 238 pairs of proteins that were significantly differentially correlated with respect to one of the 33 AD risk variants ($P_{FDR} \leq 0.05$, Supp Table 3).

To assess whether the identified protein pairs could have a functional association, we checked if the protein pairs were expressed in the same cell type (FPKM > 0.1, See Methods *Human brain cell type transcriptome profile*). All protein pairs shared cell types in which they were expressed, except for two pairs (*DDX17–PKLR* and *HNRNPL- ICAM5*), which could not be validated as *PKLR* and *ICAM5* did not exceed the expression threshold in any cell type. These findings reveal that although most AD risk variants do not significantly impact the abundance of individual proteins in the GTM, they are associated with modifications in protein-protein correlation patterns.

## 4.2.6. Potential role for DDX17 in mediating the protective effect of rs9381040-T through tight regulation of synuclein abundance

Next, we aimed to characterize the risk variants and the proteins whose associations with other proteins were altered in relation to the respective AD risk variants. We found that three SNPs (rs9381040 in TREML2, rs34173062 in SHARPIN and rs11218343 near SORL1) had the strongest regulatory effects on protein-protein associations. Specifically, most differentially correlated proteins (23%) were found between the two homozygous genotypes of variant rs9381040 (closest gene = TREML2, 54 differentially correlated pairs, including 90 different proteins, Fig. 3a). These 90 proteins were enriched for neuron cell type markers (N = 17, OR = 3.55, 95%CI = 1.47-9.17, $P = 2.71 \times 10^{-3}$), suggesting that rs9381040 may predominantly influence neuronal protein networks and potentially impact neuronal function in AD.

Among the 54 differentially correlated protein pairs, DDX17 had the most altered associations, being differentially correlated with 15 proteins (Fig 3b-d). Functional enrichment analysis of the proteins differentially correlated with DDX17 revealed

enrichment for 'PFAM Protein Domains: Synuclein' ($N = 3$, $1.38 \times 10^{-7}$). All members of synuclein family (SNCA, SNCB and SNCG, $\Delta r \geq 1.15$, $P_{FDR} \leq 2.98 \times 10^{-2}$) were differentially correlated with DDX17 with respect to the genotypes of rs9381040. In individuals with the T/T genotype (protective), DDX17 was highly correlated with SNCA, SNCB and SNCG ($r \geq 0.94$). These results suggest that DDX17 may play a protective role in individuals with the rs9381040-T/T genotype by regulating synuclein levels, potentially preventing its aggregation and the subsequent neurotoxicity seen in AD[20,21].

33 pairs of proteins (14%, including 60 different proteins, Fig. 3c, Fig. 3d) showed differential correlation between homozygous rs34173062-G and heterozygous rs34173062-G/A. Rs34173062 is a missense variant in SHARPIN. Among these 33 differentially correlated protein pairs, SPTBN2 exhibited the most altered associations with other proteins (N=5). SPTBN2, a brain spectrin, has been implicated in several neurodegenerative diseases[22,23], including AD[24,25].

These findings suggest a potential association between SHARPIN and SPTBN2 concerning rs34173062. However, the functional association of this association to AD risk requires further investigation. The set of 60 proteins was not enriched for specific cell type markers. 32 of these proteins were involved in GO biological process of 'transport' ($P_{FDR} = 8.54 \times 10^{-5}$), 50 proteins were associated with the GO cellular component 'cytoplasmic part' ($P_{FDR} = 5.89 \times 10^{-6}$), and 16 proteins were linked to the mitochondrial part ($P_{FDR} = 7.03 \times 10^{-6}$).

We observed that 15 protein pairs were dcQTL with variant rs11218343 (closest gene = *SORL1,* involving 29 proteins*,* Fig. 3e, Fig. 3f). These proteins showed distinct correlations between homozygous (T/T) and heterozygous (T/C) individuals. However, the group of 29 proteins did not show enrichment for specific cell type markers. 18 of these proteins were involved in the *GO* biological process of 'localization' ($P_{FDR} = 4.90 \times 10^{-3}$), 8 proteins were associated with the *GO* cellular component 'dendrite' ($P_{FDR} = 1.00 \times 10^{-4}$), and 9 proteins were linked to 'neuron projection' ($P_{FDR} = 6.80 \times 10^{-4}$).

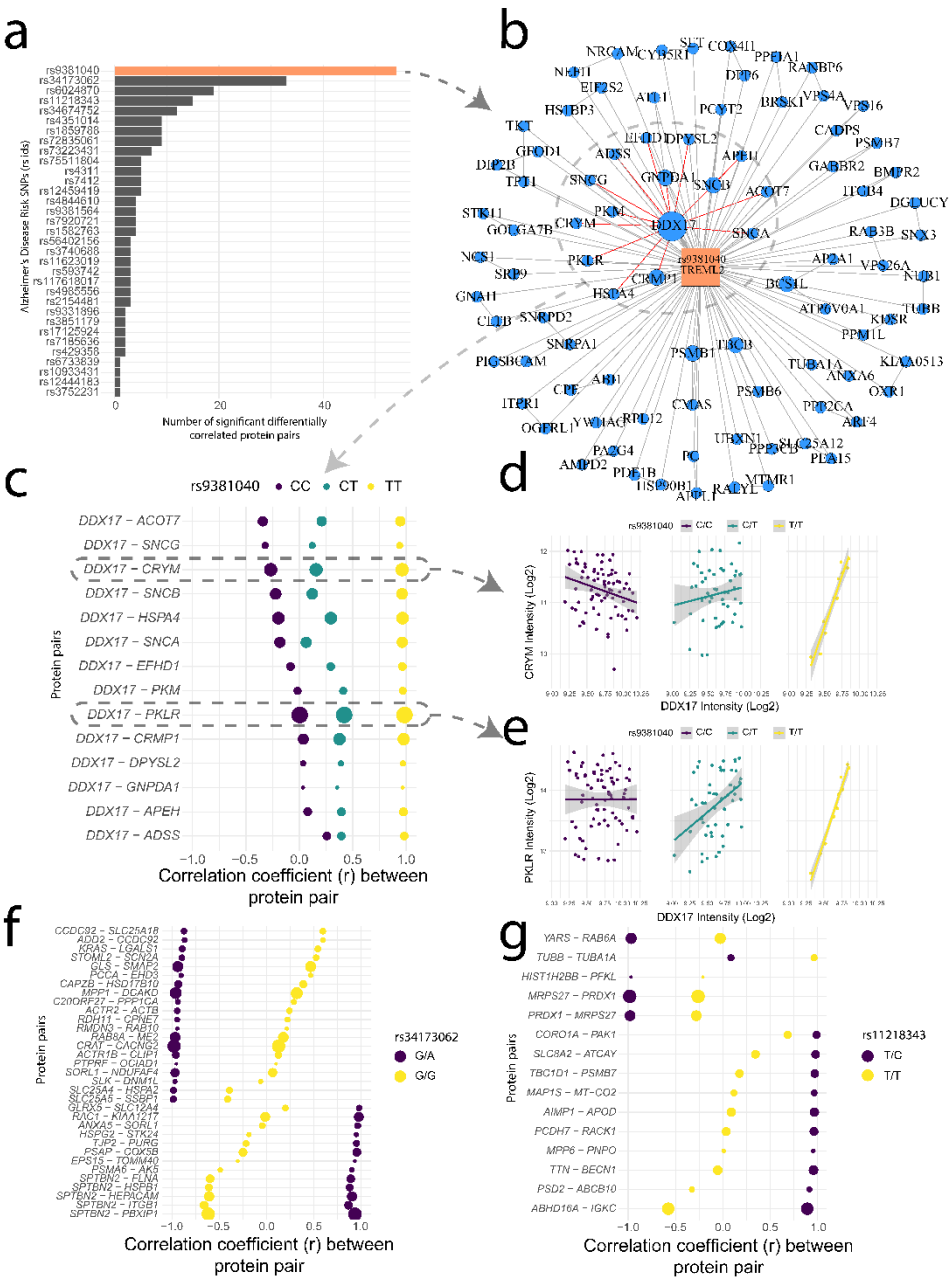



**Figure 3** Overview of dcQTL results with respect to AD variants. a) Network graph illustrating differentially correlated proteins associated with rs9381040. Blue nodes are proteins, an edge indicates that two proteins are differentially correlated with respect to rs9381040. Node size reflects the degree of connectivity. b) DDX17's differential correlation with other proteins relative to the rs9381040 genotype. The x-axis shows Pearson's correlation coefficients between protein pairs across individuals within the respective genotypes. The y-axis lists each protein pair, and dot color indicates different genotypes. Dot size corresponds to -$\log_{10}$ p-value c, d) Scatter plots depicting two proteins (c, PKLR ; d, CRYM) differentially correlated with DDX17 (x-axis) concerning rs9381040

genotypes. Each dot represents an individual, colored by genotype (purple for C/C, green for C/T, yellow for and T/T). e, f) All proteins differentially correlated with e) rs34173062, and f) rs11218343. The x-axis represents Pearson's correlation coefficients between the protein pairs across individuals withing the respective genotypes . The y-axis lists each protein pair, with dot colors indicating different genotypes. Dot size corresponds to $-\log_{10}$ p-value.

## 4.3. Discussion

We conducted a comprehensive pQTL analysis in the GTM, identifying associations between the expression of 227 proteins and 8,081 genetic variants. Our findings not only align with previous studies in proteomics and transcriptomics, but also revealed novel variant-protein associations that may indicate specific vulnerabilities of brain regions to AD.

Our comparison of AD genetics with brain proteomics suggests that individuals carrying AD-associated variants in/near *TREML2*, *SHARPIN* and *SORL1* exhibit distinct protein-protein correlation structures compared to non-carriers. This implies that individual AD associated risk alleles may exert significant control over protein-protein correlation patterns, highlighting potential mechanisms underlying AD pathogenesis.

The identified dcQTLs often involved central *hub* proteins. For example, variant rs9381040 (near *TREML2* gene) was associated with differential correlations among 54 pairs of proteins. In this network, *DDX17* (DEAD-Box Helicase 17) played a pivotal role by exhibiting differential correlation with 15 proteins. *DDX17* functions as a transcriptional co-regulator for various target genes[26] and plays a significant role in the androgen signaling pathway[26], which is implicated in protective mechanisms against neurodegenerative diseases by potentially reducing β-amyloid accumulation[27].

Moreover, *DDX17* is involved in amyloidogenesis, a crucial process linked to AD pathogenesis, and a possible mediator contributing to AD[28]. Although *TREML2* (Triggering Receptor Expressed On Myeloid Cells Like 2) was not measured in our proteomics dataset, its expression is known to increase in neutrophils and macrophages during immune responses to inflammatory factors[29]. We speculate that the protective effects observed with rs9381040 may involve altered functions of DDX17 and potentially be initiated through immune-related responses mediated by TREML2. However, further research is needed to elucidate the functional relationship between *TREML2* and *DDX17* regarding their protective action against AD.

*SPTBN2* (Spectrin Beta, Non-Erythrocytic 2) was identified as *hub protein* linked to the effects of rs34173062, a missense variant within *SHARPIN*. Brain spectrins have garnered attention in various neurodegenerative diseases[22,23], including AD[24,25]. Our findings suggest a connection between *SHARPIN* and *SPTBN2* concerning genetic risk for AD. The functional implications of this association in relation to AD risk require further investigation.

QTL studies traditionally aim to uncover how genetic variations regulate transcription or protein levels. However, our findings underscore the importance of moving beyond straightforward associations between genetic variants and biomolecules. A variant may not directly influence the expression levels of a

transcript or protein; instead, it could trigger changes in biological states where the interactions between proteins or transcripts are redefined. While studies using single-cell RNA sequencing data increasingly adopt this perspective[30,31], our research highlights the continued relevance of bulk data, particularly in proteomics, for unraveling these intricate genetic interactions.

This study benefits from the inclusion of single cell data to complement bulk data-derived results. By considering cell type specificities, we verified that nearly all differentially correlated proteins were expressed in the same cells, thereby affirming the hypothesis of functional association through co-expression in our findings. Another strength lies in the inclusion of individuals with extreme phenotypes, particularly cognitively healthy centenarians, who were found to be depleted with genetic variants associated with increased AD risk. This inclusion enhances the power and effect size for AD specific variants[32].

However, it's important to note that applying our differential correlation approach on a proteome- and genome-wide scale is impractical due to the vast number of pair-wise tests between proteins and variants involved. Therefore, a hypothesis driven approach, as employed in this study, becomes essential.

In summary, our study bridged genetic variants with proteomics in the GTM. Moreover, through differential correlation analysis based on genotypes of AD risk variants, we demonstrated that this approach holds promise as a valuable addition to GWAS- and QTL-studies. It effectively identifies proteins potentially involved in downstream effects of disease-associated risk variants. Additionally, our findings present promising results and suggest new research opportunities for exploring genetic implications of AD risk variants.

# 4.4. Methods

## Population of the study

Individuals classified as AD patients were derived from two sources: 1) clinically diagnosed with probable AD patients from the Amsterdam Dementia Cohort[14] (N=2,668), and 2) pathologically confirmed AD patients from the Netherlands Brain Bank[16] (N=436). The non-demented controls included individuals from various cohorts: 1) 1,779 individuals aged 55-58 years from the Longitudinal Aging Study Amsterdam[33] (LASA), 2) 1,206 individuals with subjective cognitive decline assessed at the memory clinic of the Alzheimer center Amsterdam confirmed as cognitively normal after thorough examination, 3) 40 healthy individuals from the Netherlands Brain Bank, 4) 201 individuals from the twin study[17], and 5) 444 individuals from the 100-plus Study cohort[18]. The 100-plus Study cohort comprises of Dutch-speaking individuals aged 100 years and older, who self-reported to be cognitively healthy confirmed by their family members and partners. For this study, a total of N=358 cognitive health centenarians and N=86 partners of centenarian's children were considered. Genetic data was available for all individuals. For 140 of the 190 individuals, proteomics data was also available. These samples with both genetics and proteomics data were used for the pQTL analysis in this study. The Medical Ethics Committee of the Amsterdam UMC (METC) approved all studies. All participants and/or their legal

representatives provided written informed consent for participation in clinical and genetic studies.

## Genetic data processing

Genetic variants were identified using standard genotyping and imputation methods, followed by established quality control procedures. Genotyping was performed on individuals using Illumina Global Screening Array (GSAsharedCUSTOM_20018389_A2). High-quality genotyping was retained (individual call rate > 98%, variant call rate > 98%), with exclusions for sex mismatches and significant departures from Hardy–Weinberg equilibrium ($P<1\times10^{-6}$). Genotypes were prepared for imputation using provided scripts (HRC-1000G-check-bim.pl)[34] to compare variant ID, strand and allele frequencies to the Haplotype Reference Panel (HRC v1.1, April 2016)[35]. All autosomal variants were submitted to the Sanger imputation server (https://imputation.sanger.ac.uk), which uses MACH for phasing and PBWT for imputation against the HRC v1.1, April 2016 reference panel.

In total, 3,670 population subjects and 3,106 AD cases passed quality control. Before analysis, individuals of non-European ancestry were excluded based on 1000Genomes clustering, those with a family relation (identity-by-descent > 0.2) were also excluded. This resulted in the exclusion of 205 population controls and 152 AD cases with non-European ancestry, and 217 population controls and 100 AD cases with family relations. Consequently, 4,191 control subjects and 2,416 AD cases remained for the analyses, yielding a total sample size of 6,607. Of these, 140 individuals also had proteomics data.

## Summary statistics pQTL study

The pQTL summary statistics of Robins et al.[11] were obtained from http://brainqtl.org. The pQTLs were identified in 144 healthy individuals originally part of the ROSMAP study, a population composition of 63.1% females and a median age of 86.5 (range: 67.4 to 102.7). Protein expression data were sourced from the dorsolateral prefrontal cortex. In contrast to our pQTL analysis, variants within 50 Kbp up- and downstream of the transcription start site (TSS) of the respective proteins were tested. The dataset includes pQTL summary statistics for 7,901 proteins and 2,599,383 variants, totaling, 4,199,577 pQTLs. Proteins were identified using their Uniprot accession IDs and variants by their GRCh37/hg19 genomic coordinate. P-values were corrected for multiple tests using both Bonferroni and FDR methods. Using Bonferroni correction, 2,955 significant pQTLs ($P_{BONF} \leq 0.05$) were identified, while FDR correction identified 28,211 significant pQTLs ($P_{FDR} \leq 0.05$).

## eQTLs from GTEx

The eQTL data was accessed through the GTEx Application programming interface (API)[7]. Twelve brain regions were analyzed, each brain with varying numbers of individuals who had genotype and RNA-seq data available (Table S1). The total population comprised 395 individuals, 72% of whom were male. While GTEx does not report the exact ages of these individuals, the specified age

ranges are as follows: 20-29 years (N = 8), 30-39 years (N = 10), 40-49 years (N = 36), 50-59 years (N = 119), 60-69 years (N = 200), and 70-79 years (N = 22).

The eQTL statistics obtained from GTEx include a normalized effect size (NES), which is the slope of the linear regression, indicating the effect of the alternative allele (ALT) relative to the reference allele (REF) according to human genome reference GRCh38/hg38. The data also contains nominal p-values for the eQTL association and a p-value threshold, determined by $P_{FDR} \leq 0.05$, but is translated to a nominal p-value. Variants are defined by their reference SNP identification number (rs IDs), while transcripts are defined by their gene symbol and an Ensembl transcript ID.

## Human brain cell type transcriptome profile (HBCT)

The cell type specific transcript expression data of Zhang et al.[36] were obtained from https://www.brainrnaseq.org/. A total of 21.661 transcripts were measured across five cell types: astrocytes, neurons, oligodendrocytes, microglia/macrophages, and endothelial cells, sampled from individuals ranging from 8 years to 63 years old. Specifically, transcripts measurements in astrocytes were derived from 26 samples, with fourteen from male individuals and twelve from females, including four samples from the tumor core or their vicinity. Neuron-specific expression came from one male individual, while oligodendrocyte-specific expression was derived from 5 individuals (4 makes and 1 female). Microglia/macrophages expression was available from three individuals (2 males and 1 female), and the endothelial cells from two females. The transcript expression levels were normalized using Fragments Per Kilobase Million (FPKM), providing a standardized representation of transcript abundance.

## Cell type markers, composition, and enrichment

Cell type markers were estimated using the HBCT transcriptome dataset. First, cells that originated from a tumor or its surroundings were excluded (N = 4). The average FPKM of each gene, in each of the five cell types (astrocytes, neurons, oligodendrocytes, microglia/macrophages, and endothelial cells) was calculated. For genes with multiple measurements, an average FPKM value was calculated. A gene was annotated as unique cell type marker for a particular cell type when the fold change of the average FPKM was ≥ 3 compared to all other cell types.

Subsequently, the cell type composition for individuals with available protein data (N = 190, See Table 1: *Protein data*) was estimated. This was achieved by averaging the protein intensities of the unique cell type makers present in the GTM protein dataset for each cell type, thus providing an estimation of the cell type abundance for each individual based on their respective protein abundance. Finally, cell type enrichments were performed with the Fisher's exact test[37], based on the unique cell type markers of the five cell types, to calculate the enrichment of a set of proteins for unique cell type markers.

## Gyrus Temporalis Medialis Proteomics data

Proteomics data from the GTM was measured for a total of 237 individuals from Netherlands Brain Bank[38], from whom measurements were taken for 4,829

proteins. Among them, 102 individuals were diagnosed with Alzheimer's Disease (AD), 62 were cognitively health centenarians (CHC), and 73 were non-demented (ND) controls. Proteomics data was generated using the Sequential Window Acquisition of All Theoretical Mass Spectra (SWATH- MS) method employing a Data-Independent Acquisition (DIA) approach. Spectrum annotation and relative protein quantification were performed using MaxQuant software[39], with the Uniprot human reference proteome[40] used as reference.

## Gyrus Temporalis Medialis Proteomics Quality Control and Pre-processing

Quality control was performed separately on both a sample basis and protein basis. Initially, samples with more than 34% of low-quality peptides ($Q \geq 0.01$) were excluded from the analyses (N = 35). After removing low-quality samples, a reference peptide intensity distribution was calculated by averaging the peptide intensity distributions of the remaining samples. The distance between each individual peptide intensity distribution and the reference distribution was then calculated using the Kolmogorov–Smirnov test. Samples with a distribution distance (*D*) greater than 0.04 from the reference distribution were excluded from the analyses (N = 1). For replicate samples, lower quality samples were determined using a paired t-test on the quality measures, resulting in the exclusion of eleven replicates.

The proteomics data was generated in bottom-up fashion, where peptides were measured and used to estimate the expression of their respective protein. If the peptides comprising a single protein were of low-quality in more than 10% of the samples, the respective protein was excluded. Proteins were represented by the sum of intensities of their respective peptides. Finally, protein intensities were $\log_2$ transformed to ensure the normality of the protein intensity distributions. The final proteomics dataset consists of 3,556 proteins and 190 individuals.

Next, batch effects, which have no biological meaning, were removed during the pre-processing step. Initially, the association between variations in the proteomics data and the variables age, sex, Braak stage I-VI, post-mortem delay (PMD), *APOE* genotype ($\log_2$ Polygenic Risk Score), and batch were tested using the R-package variancePartition[41,42]. VariancePartition employs a mixed linear model to determine the percentage of variation attributable to each variable. Among the tested variables, substantial proportions of the variation were explained by age, Braak stage and batch. To remove the variation associated with batch from the protein intensity data, the combat function from the R-package sva[43] was used.

## pQTL identification

pQTL analysis was performed on the subset of 140 individuals for which both genetics and proteomics data was available after quality control. Genetic variants associated with protein expression were identified with Plink (v2.00a2LM)[44]. Linear models were employed for the association analysis, with genotype dosages as predictors for protein expression, assuming additive genetic effects. The analyses were corrected for estimated cell type composition (neurons,

microglia/macrophages and oligodendrocytes, See Materials and methods *Cell type markers, composition, and enrichment*) and population substructure using the first five principal components (See methods *pQTL linear regression model*). An additional analysis was performed correcting for phenotype status (AD, ND controls and CHC, See Table 1: *intersection*).

Resulting effect-sizes (β) were calculated with the minor allele relative to the major allele in our population. Association P-values were corrected for multiple tests with False Discovery Rate (FDR), with assumed at $P_{FDR} \leq 0.05$. The analyses were restricted to variants with a MAF higher than 5% and variants located 250 Kbp down- and upstream of the TSS of the respective proteins. Four window sizes were tested (50 Kbp, 250 Kbp, 500 Kbp and 1 Mb, Supp. Fig. 4). Using the 250 Kbp window, we reduced to total number of tests while still capturing most of pQTLs.

Genomic locations of the TSSs were acquired with biomaRt (v2.42.0)[45,46]. The retrieved genomic locations of the TSSs were for genomic build GRCh38/hg38. The liftOver R-package (v1.10.0)[47,48] was used to lift over the genomic coordinates to build GRCh37/hg19, as the genotype files were based on this genomic build.

## pQTL linear regression model

For identifying significant pQTLs, the generalized linear model (GLM) from Plink (v2.00a2LM)[44] was used, which is the primary association analysis method in Plink for quantitative phenotypes. The model applied to our data was as follows:

$$P_y = \beta_0 G + \beta_1 \overline{N} + \beta_2 \overline{MG} + \beta_3 \overline{O} + \beta_4 PC_1 + \cdots + \beta_8 PC_5 + \varepsilon \quad (3)$$

where: $P_y$ is the $\log_2$ intensity for the individuals of respective protein (quantitative phenotype); G are the dosages for the individuals of the respective variant that is tested; $\overline{N}$ are the mean intensity of all Neuron cell type markers; $\overline{MG}$ are the mean intensity of all Microglia/Macrophage cell type markers; $\overline{O}$ are the mean intensity of all Oligodendrocytes cell type markers; $PC_i$ are the principal components for the individuals of the population substructure; $\varepsilon$ is an error term that gets minimized with least squares minimization.

## Testing pQTL variants on association with AD risk

All significant pQTL variants were tested on association with AD risk using all individuals for which genetics data was available. This, which comprised 6,479 individuals (2,361 AD cases and 4,118 ND controls). To ensure an independent population, individuals used in the pQTL identification (N=128 of 140) were excluded, ensuring no overlap of individuals between the differential expression analysis population and the genetic association test population. The association of pQTL variants with AD status was tested using a logistic regression model in R (v3.6.3), with AD status as discrete outcome variable (ND = 0, AD = 1) and the pQTL variant's dosages as predictor variable. The model was adjusted for population substructure using the first five principal components. P-values were adjusted for multiple tests using FDR, with significant association assumed at $P_{FDR} \leq 0.05$.

## Colocalization analysis

For the colocalization analysis, we initially identified proteins associated with pQTL variant that also showed associated with AD status (See methods *Testing pQTL variants on association with AD risk*). Subsequently, we extracted all variants within 250 Kbp up- and downstream of the TSS of these proteins. These variants were then evaluated for association with AD status using the same cohort of individuals described previously (N = 6,479, See methods *Testing pQTL variants on association with AD risk*).

A logistic regression model was performed with AD status (ND = 0, AD = 1) as the discrete outcome variable and genotypes of the aforementioned variants as predictor variable. The model was adjusted for population substructure using the first five principal components.

To estimate the probability that the specified genomic region contains a pQTL variant influencing both protein abundance and AD risk, an approximate Bayes Factor colocalization analysis was performed using the coloc R-package (v3.2.1)[49]. This analysis utilized summary statistics including p-values, sample size and MAF. The colocalization analysis tests five hypotheses. $H_0$: There is no association of the genomic region with protein abundance and AD risk. $H_1$: There is only an association with protein abundance. $H_2$, there is only an association with AD risk. $H_3$: The genomic region is associated with both protein abundance and AD risk, but through two different variants. $H_4$: The genomic region is associated with both protein abundance and AD risk, through a single variant. For each hypothesis, a posterior probability is computed to assess the likelihood of colocalization between protein abundance and AD risk in the specified genomic region.

## pQTL and eQTL comparison

We examined whether the significant pQTLs were also an eQTL variant using the eQTL data from twelve brain tissues (Table S1) from GTEx (v8)[7]. An independent set of clumped pQTLs (See methods *Clumping)* was used to minimize the number of requests that needed to be sent to the API of GTEx. We investigated associations of pQTLs from the GTM with eQTLs both across the twelve brain tissues and within specific brain regions. Each gene-variant pair was queried across twelve brain tissues using the get_eQTL_bulk function from the R-package CONQUER (v1.0)[50], which requires tissue ID, gene symbol and RS ID to be supplied. Significance of the tested eQTLs was determined using the P-value thresholds provided by GTEx. To compare the directional effects of pQTLs with their synonymous eQTLs, we calculated Pearson's correlation coefficient between effect sizes using the cor.test function.

## Clumping

We created an independent set of pQTLs using LD-based clumping. Variants that are located close to each other often exhibit linkage disequilibrium, meaning that they are correlated and show similar associations with the same protein. The clumping procedure allows to retain only the most strongly associated variant within a specified window. We performed clumped individually for each protein

using Plink (v1.90b4.6)[44], with criteria set at $R^2 \geq 0.001$ and MAF ≥ 0.05. Linkage disequilibrium between variants was calculated using European individuals from the 1,000 Genomes Project reference panel[51].

## Differential correlation

We examined whether protein correlation changed when comparing two distinct groups based on AD variant genotypes. Initially, Pearson's correlation between a pair of proteins was calculated separately for each group of interest. Let's denote these coefficients as $r_x$ for group *x* and $r_y$ for group *y*. Subsequently, these correlation coefficients were transformed into z-scores using Fisher's z-transformation[52] (Eq. 4).

$$z = atanh(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \quad (4)$$

The difference between z-scores $z_x$ and $z_y$ was then calculated using equation 5:

$$\Delta z = \frac{(z_x - z_y)}{\sqrt{var(r_x) + var(r_y)}} \quad (5)$$

where $var(r)$ is calculated by $\frac{1}{n-3}$, with $n$ being the sample size of the respective groups. Since $\Delta z$ follows a normal distribution, a two-sided P-value for the differential correlation between each pair of proteins can be determined.

## Differential correlation with respect to AD variants genotype

In this analysis, we used individuals with both genetics and proteomics data: 67 AD individuals, 27 ND controls, and 46 CHC. We conducted a differential correlation analysis of proteins based on the genotypes of established AD variants. Initially, we considered 41 variants[2] known to influence AD risk. We selected variants where each genotype was represented by at least 10 individuals to mitigate population size discrepancies and reduce the likelihood of false positives, resulting in 33 remaining variants.

For variants where all three genotypes were present, we calculated the differential correlation between the two homozygous genotypes, reporting the correlation involving the heterozygous genotype separately. When only two genotypes were present, we assessed the differential correlation between the homozygous genotype and the heterozygous genotype.

The differential correlation method for these variants was implemented in R (v3.6.3)[53]. P-values were FDR corrected based on the total number of tests conducted. Significance was established at $P_{FDR} \leq 0.05$.

# References

1. Thies, W. & Bleiler, L. 2012 Alzheimer's disease facts and figures. *Alzheimer's Dement.* 8, 131–168 (2012).

2. de Rojas, I. *et al.* Common variants in Alzheimer's disease and risk stratification by polygenic risk scores. *Nat. Commun. 2021 121* 12, 1–16 (2021).

3. Wightman, D. P. *et al.* A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat. Genet. 2021 539* 53, 1276–1282 (2021).

4. Bellenguez, C. *et al.* New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat. Genet. 2022 544* 54, 412–436 (2022).

5. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science (80-. ).* 337, 1190–1195 (2012).

6. Odell, S. G. *et al.* Modeling allelic diversity of multiparent mapping populations affects detection of quantitative trait loci. *G3 Genes|Genomes|Genetics* 12, (2022).

7. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* vol. 45 580–585 at https://doi.org/10.1038/ng.2653 (2013).

8. Guo, Y. *et al.* How is mRNA expression predictive for protein expression? A correlation study on human circulating monocytes. *Acta Biochim. Biophys. Sin. (Shanghai).* 40, 426–436 (2008).

9. De Sousa Abreu, R., Penalva, L. O., Marcotte, E. M. & Vogel, C. Global signatures of protein and mRNA expression levels. *Molecular BioSystems* vol. 5 1512–1526 at https://doi.org/10.1039/b908315d (2009).

10. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* 13, 227–232 (2012).

11. Robins, C. *et al.* Genetic control of the human brain proteome. *bioRxiv* 816652 (2019) doi:10.1101/816652.

12. de Flores, R. *et al.* Medial Temporal Lobe Networks in Alzheimer's Disease: Structural and Molecular Vulnerabilities. *J. Neurosci.* 42, 2131–2141 (2022).

13. Dalton, M. A., Tu, S., Hornberger, M., Hodges, J. R. & Piguet, O. Medial temporal lobe contributions to intra-item associative recognition memory in the aging brain. *Front. Behav. Neurosci.* 7, 57029 (2014).

14. Van Der Flier, W. M. & Scheltens, P. Amsterdam dementia cohort: Performing research to optimize care. *Journal of Alzheimer's Disease* vol. 62 1091–1111 at https://doi.org/10.3233/JAD-170850 (2018).

15. Hoogendijk, E. O. *et al.* The Longitudinal Aging Study Amsterdam: cohort update 2016 and major findings. *Eur. J. Epidemiol.* 31, 927–945 (2016).

16. Rademaker, M. C., de Lange, G. M. & Palmen, S. J. M. C. The Netherlands Brain Bank for Psychiatry. in *Handbook of Clinical Neurology* vol. 150 3–16 (Elsevier B.V., 2018).

17. Willemsen, G. *et al.* The Netherlands twin register biobank: A resource for genetic epidemiological studies. *Twin Res. Hum. Genet.* 13, 231–245 (2010).

18. Holstege, H. *et al.* The 100-plus Study of cognitively healthy centenarians: rationale, design and cohort description. *Eur. J. Epidemiol.* 33, 1229–1249 (2018).

19. Raulin, A. C. *et al.* ApoE in Alzheimer's disease: pathophysiology and therapeutic strategies. *Mol. Neurodegener. 2022 171* 17, 1–26 (2022).

20. Crews, L., Tsigelny, I., Hashimoto, M. & Masliah, E. Role of Synucleins in Alzheimer's Disease. *Neurotox. Res.* 16, 306–317 (2009).

21. van der Gaag, B. L. *et al.* Distinct tau and alpha-synuclein molecular signatures in Alzheimer's disease with and without Lewy bodies and Parkinson's disease with dementia. *Acta Neuropathol.* 147, 1–22 (2024).

22. Nicita, F. *et al.* Heterozygous missense variants of SPTBN2 are a frequent cause of congenital cerebellar ataxia. *Clin. Genet.* 96, 169–175 (2019).

23. Ordonez, D. G., Lee, M. K. & Feany, M. B. α-synuclein Induces Mitochondrial Dysfunction through Spectrin and the Actin Cytoskeleton. *Neuron* 97, 108-124.e6 (2018).

24. Sihag, R. K. & Cataldo, A. M. Brain β-spectrin is a component of senile plaques in Alzheimer's disease. *Brain Res.* 743, 249–257 (1996).

25. Yan, X.-X. & Jeromin, A. Spectrin Breakdown Products (SBDPs) as Potential Biomarkers for Neurodegenerative Diseases. *Curr. Transl. Geriatr. Exp. Gerontol. Rep.* 1, 85–93 (2012).

26. Samaan, S. *et al.* The Ddx5 and Ddx17 RNA helicases are cornerstones in the complex regulatory array of steroid hormone-signaling pathways. doi:10.1093/nar/gkt1216.

27. Pike, C. J. *et al*. Androgen cell signaling pathways involved in neuroprotective actions. *Hormones and Behavior* vol. 53 693–705 at https://doi.org/10.1016/j.yhbeh.2007.11.006 (2008).

28. Liu, Y. *et al.* DEAD-Box Helicase 17 Promotes Amyloidogenesis by Regulating BACE1 Translation. *Brain Sci. 2023, Vol. 13, Page 745* 13, 745 (2023).

29. Klesney-Tait, J., Turnbull, I. R. & Colonna, M. The TREM receptor family and signal integration. *Nature Immunology* vol. 7 1266–1273 at https://doi.org/10.1038/ni1411 (2006).

30. Van Der Wijst, M. G. P. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* 50, 493–497 (2018).

31. Li, S. *et al.* Identification of genetic variants that impact gene co-expression relationships using large-scale single-cell data. *Genome Biol.* 24, 1–37 (2023).

32. Tesi, N. *et al.* Cognitively healthy centenarians are genetically protected against Alzheimer's disease. *Alzheimers. Dement.* 20, 3864–3875 (2024).

33. Huisman, M. *et al.* Cohort Profile: The Longitudinal Aging Study Amsterdam How did the study come about? *Int. J. Epidemiol.* 40, 868–876 (2011).

34. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* 48, 1284–1287 (2016).

35. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* 48, 1279–1283 (2016).

36. Zhang, Y. *et al.* Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. *Neuron* 89, 37–53 (2016).

37. Fisher, R. A. Statistical Methods for Research Workers. in 66–70 (Springer, New York, NY, 1992). doi:10.1007/978-1-4612-4380-9_6.

38. Ganz, A. B. *et al.* Proteomic profiling of aging brains identifies key proteins by which cognitively healthy centenarians defy their age by decades. *medRxiv* 2023.11.30.23299224 (2023) doi:10.1101/2023.11.30.23299224.

39. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* 26, 1367–1372 (2008).

40. D506-D515. UniProt: a worldwide hub of protein knowledge The UniProt Consortium. *Nucleic Acids Res.* 47, (2019).

41. Hoffman, G. E. & Schadt, E. E. variancePartition: Interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* 17, 483 (2016).

42. Hoffman, G. E. & Roussos, P. dream: Powerful differential expression analysis for repeated measures designs. *bioRxiv* 432567 (2018) doi:10.1101/432567.

43. Leek, J. T. *et al.* The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinforma. Appl. NOTE* 28, 882–883 (2012).

44. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 7 (2015).

45. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. *Nat. Protoc.* 4, 1184–1191 (2009).

46. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinforma. Appl. NOTE* 21, 3439–3440 (2005).

47. Bioconductor - liftOver. https://www.bioconductor.org/packages/release/workflows/html/liftOver.html.

48. Hinrichs, A. S. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* 34, D590–D598 (2006).

49. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* 10, e1004383 (2014).

50. Bouland, G. A. *et al.* Understanding functional consequences of type 2 diabetes risk loci using the universal data integration and visualization R package CONQUER. *bioRxiv* 2020.03.27.011627 (2020) doi:10.1101/2020.03.27.011627.

51. Auton, A. *et al.* A global reference for human genetic variation. *Nature* vol. 526 68–74 at https://doi.org/10.1038/nature15393 (2015).

52. Fisher, R. A. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika* 10, 507 (1915).

53. R Core Team. R: A Language and Environment for Statistical Computing. at https://www.r-project.org/ (2020).

# Supplements

**pQTL linear regression model**

For identifying significant pQTLs, the generalized linear model (GLM) from Plink (v2.00a2LM)[60] is used. Which is the primary association analysis method in Plink for quantitative phenotypes. The model applied on our data was as follows:

$$P_y = \beta_0 G + \beta_1 \bar{N} + \beta_2 \overline{MG} + \beta_3 \bar{O} + \beta_4 PC_1 + \cdots + \beta_8 PC_5 + e$$

Where:

$P_y$ is the $\log_2$ intensity for the individuals of respective protein (quantitative phenotype).

G are the dosages for the individuals of the respective variant that is tested.

$\bar{N}$ are the mean intensity of all Neuron cell type markers

$\overline{MG}$ are the mean intensity of all Microglia/Macrophage cell type markers

$\bar{O}$ are the mean intensity of all Oligodendrocytes cell type markers

$PC_i$ are the principal components for the individuals of the population substructure.

e is an error term that gets minimized with least squares minimization.

**Protein residuals**

For each individual, for each protein measured in the GTM protein dataset we calculated the residual after correcting for the abundance of three cell types (neurons, microglia/macrophages, and oligodendrocytes). The pQTL analysis was also corrected for the abundance of these cell types. The residuals were calculated in order to truthfully visualize the pQTL associations.

First for each protein we fitted a linear model:

$$Protein_{int} = \beta_0 + \beta_1 \bar{N} + \beta_2 \overline{MG} + \beta_3 \bar{O} + \varepsilon$$

Where:

$\bar{N}$ = the mean intensity of all neuron cell type markers

$\overline{MG}$ = the mean intensity of all microglia/macrophage cell type markers

$\bar{O}$ = the mean intensity of all oligodendrocytes cell type markers

Next, with the fitted model, we predicted the protein expression:

$$Protein_{pred} = \beta_0 + \beta_1 \bar{N} + \beta_2 \overline{MG} + \beta_3 \bar{O} + \varepsilon$$

And finally, we subtracted $Protein_{pred}$ from $Protein_{int}$ to get the protein residuals.

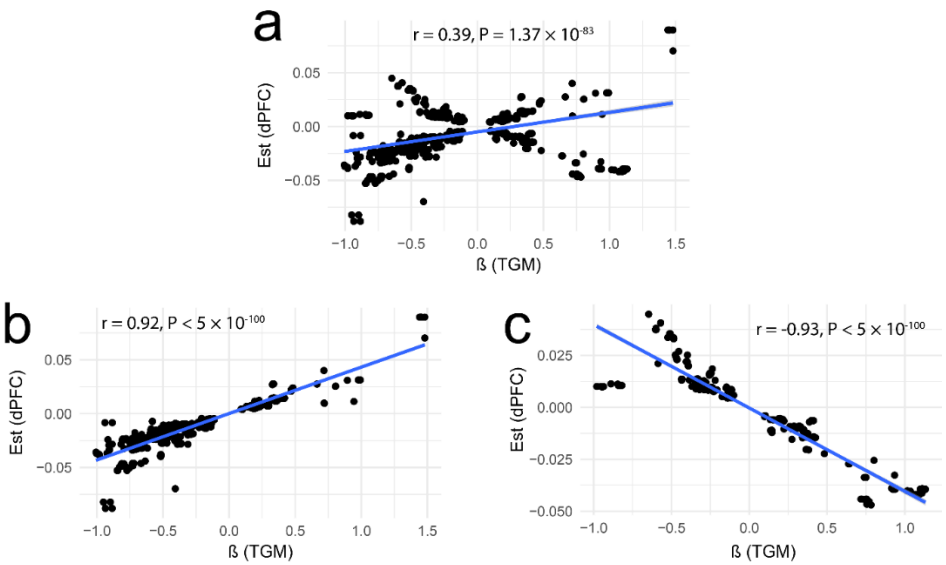
**Supplementary Table 1:** Genotyped and RNAseq sample sizes from GTEx for all twelve investigated brain regions

| Tissue | # RNASeq and Genotyped samples | # RNASeq Samples |
|---|---|---|
| Cerebellum | 209 | 241 |
| Cortex | 205 | 255 |
| Nucleus accumbens (basal ganglia) | 202 | 246 |
| Caudate (basal ganglia) | 194 | 246 |
| Cerebellar Hemisphere | 175 | 215 |
| Frontal Cortex (BA9) | 175 | 209 |
| Hypothalamus | 170 | 202 |
| Putamen (basal ganglia) | 170 | 205 |
| Hippocampus | 165 | 197 |
| Anterior cingulate cortex (BA24) | 147 | 176 |
| Amygdala | 129 | 152 |
| Substantia nigra | 114 | 139 |

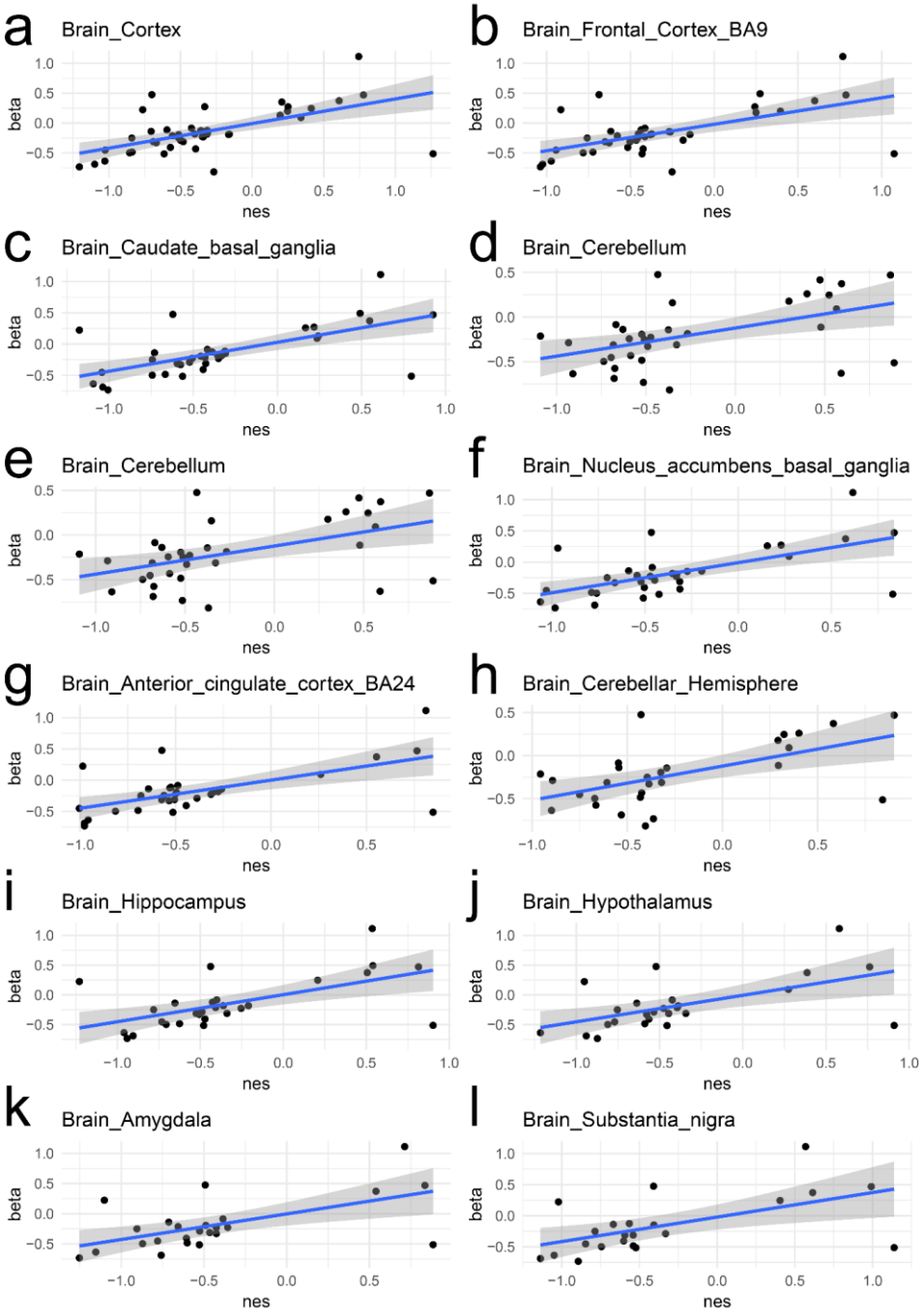
**Supplementary Table 2**: AD risk variants subject in differential correlation analysis

| RS ID | Chromosome | Genomic location | Closest Gene | Genotypes | | |
|---|---|---|---|---|---|---|
| rs6733839 | 2 | 127892810 | BIN1 | C/C 43 | C/T 71 | T/T 26 |
| rs9381040 | 6 | 41154650 | TREML2 | C/C 78 | C/T 51 | T/T 11 |
| rs1859788 | 7 | 99971834 | PILRA | A/A 13 | A/G 58 | G/G 69 |
| rs73223431 | 8 | 27219987 | PTK2B | C/C 43 | C/T 81 | T/T 16 |
| rs9331896 | 8 | 27467686 | CLU | C/C 18 | C/T 64 | 58 T/T |
| rs34674752* | 8 | 145154222 | SHARPIN | A/A 0 | G/A 11 | G/G 129 |
| rs7920721 | 10 | 11720308 | ECHDC3 | A/A 60 | A/G 56 | G/G 24 |
| rs3740688 | 11 | 47380340 | SPI1 | G/G 22 | G/T 68 | T/T 50 |
| rs1582763 | 11 | 60021948 | MS4A4A | A/A 18 | G/A 77 | G/G 45 |
| rs3851179 | 11 | 85868640 | PICALM | C/C 53 | T/C 66 | T/T 21 |
| rs11218343* | 11 | 121435587 | SORL1 | C/C 0 | T/C 12 | T/T 128 |
| rs12444183 | 16 | 81773209 | PLCG2 | A/A 20 | A/G 65 | G/G 55 |
| rs4311 | 17 | 61560763 | ACE | C/C 34 | T/C 75 | T/T 31 |
| rs12459419 | 19 | 51728477 | CD33 | C/C 68 | C/T 58 | T/T 14 |
| rs2154481 | 21 | 27473875 | APP | C/C 35 | C/T 64 | T/T 41 |

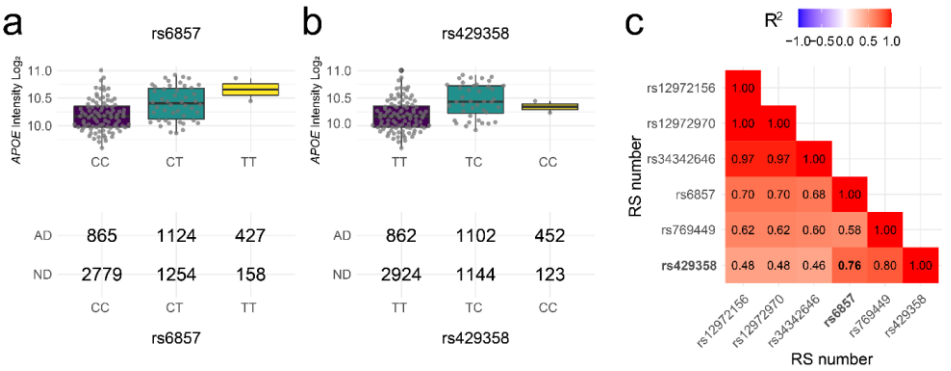
* = variant of which two genotypes were present in population



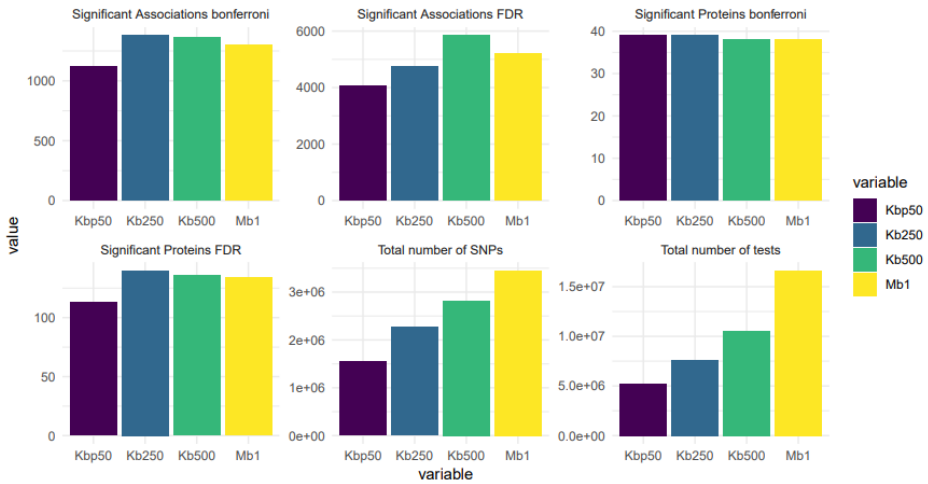

**Supplementary figure 1**: Estimate comparison of pQTL studies, in all sub-figures the x-axis represents the estimates of this current study and y-axis represents the estimates from [16]. a) Estimates of all matching pQTLs. b) Estimates of all matching pQTLs where the directional effects were the same. c) Estimates of all matching pQTLs where the directional effects opposite.




**Supplementary figure 2:** Estimate comparison of pQTLs versus the eQTL NESs from GTEx for all investigated brain regions. X-axes represent the NESs from GTEx for a particular eQTL – eGene pair. The y-axes represent the betas of the pQTL – protein pair synonymous for the eQTL – eGene pair.




**Supplementary figure 3:** Overview of APOE associated pQTL variants. a) Boxplot of rs6857 genotypes versus APOE intensity, x-axis represent the genotypes, y-axis represents the $log_2$ normalized intensity of APOE. b) Boxplot of rs429358 genotypes versus APOE intensity, x-axis represent the genotypes, y-axis represents the $log_2$ normalized intensity of APOE. c) LD correlation between the six pQTL variants associated with APOE.




**Supplementary figure** 4: Overview of pQTL variant mapping window. X-axes represent the mapping windows of 50 Kbp, 250 Kbp, 500 Kbp and 1 Mb. The y-axes represent the count of the respective statistic that is shown. The title above each plot is the respective statistic.

# 5

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

Meng Zhang, Gerard Bouland, Henne Holstege and Marcel Reinders

# Abstract

**Background and Objectives:** With age, somatic mutations accumulated in human brain cells can lead to various neurological disorders and brain tumors. Since the incidence rate of Alzheimer's 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 WGS data with genotypes derived from scRNA-seq data, and applied this workflow to 76 participants from the 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 one 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 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 two sites to have significantly more mutations in older individuals (16:6899517 (*RBFOX1*), p=0.04; 4:21788463 (*KCNIP4*), p<0.05). Also, two 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). 32 genes that harbor somatic mutations unique to AD and the *KCNQ5* and *DCLK1* genes were used for GO-term enrichment analysis. We found the AD-specific ENSMs enriched in the GO-term “vocalization behavior” and “intraspecies interaction between organisms”. Interestingly, we observed both age- and AD-specific ENSMs enriched in the $K^+$ channels-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 dataset have provided new insights into the association of AD and aging with brain somatic mutagenesis.


# 5.1. Introduction

Somatic mutations are post-zygotic 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 neurological disorders and brain tumors.[4–6] While mutations in post-mitotic 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 for many neurodegenerative diseases, like Alzheimer's 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 (NFTs).[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 neuropathological 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 post-mitotic neurons or even in a single neuron. For this reason, ultra-deep bulk sequencing and matched peripheral tissues are often required.[10] This type of data is often generated for one specific research question with relatively high cost and are not always available from public databases. In contrast, the availability of public single cell RNA sequencing (scRNA-seq) datasets has exploded due to continuous technological innovations, increasing throughput, and decreasing costs.[11] scRNA-seq data is 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, Prashant et al.,[14] compares three different variant callers (GATK, Strelka2, Mutect2) and show that a two-fold higher number of SNVs can be detected from the pooled scRNA-seq as compared to 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 10x 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. Additionally, we investigated whether AD and age coincide with an increasing number of somatic mutations.

# 5.2. Results

## 5.2.1. Excitatory neuron-specific somatic mutations (ENSMs)

To study somatic mutations acquired over age and between demented (AD) and non-demented (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). Since the scRNA-seq data (n=90) were collected within three 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 $6\times10^7$, and applied a sample-specific cut-off 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 according to seven major cell types (Methods). As 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 dominate 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) which 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. The demographic data (sex, age-at-death, and cognitive diagnosis (cogdx) categories[23]) of these participants are given in Table 1. 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 non-demented (n=42) or being an AD sample (n=33).

Table 1. Summary characteristics of selected sample from the ROSMAP study

| Group | Cogdx* | n | Sex | Age, mean ±SD (range) |
|---|---|---|---|---|
| Non-demented | 1<br>2<br>3 | 33<br>8<br>1 | 23 F; 19 M | 85.7±4.2 (76-90) |
| Alzheimer's disease | 4<br>5 | 32<br>1 | 19 F; 14 M | 87.1±3.9 (74-90) |
| Other dementia | 6 | 1 | 1F | 83 |

*Cognitive diagnosis (cogdx) is defined as six categories: 1, NCI: No cognitive impairment (No impaired domains); 2, MCI: Mild cognitive impairment (One impaired domain) and NO other cause of CI; 3, MCI: Mild cognitive impairment (One impaired domain) AND another cause of CI; 4, AD: Alzheimer's dementia and NO other cause of CI (NINCDS PROB AD); 5, AD: Alzheimer's dementia AND another cause of CI (NINCDS POSS AD); 6, Other dementia: Other primary cause of dementia.

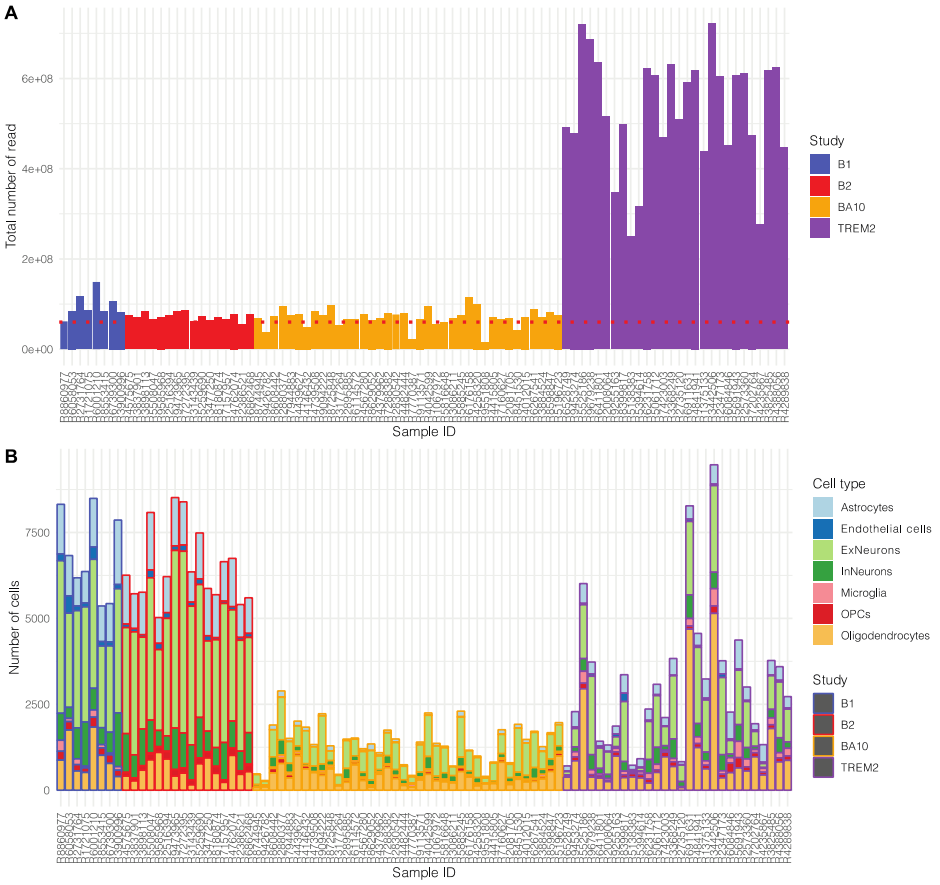



**Figure 1. Single nuclei RNA (snRNA) reads and cell count across selected samples.** Participants (n=90) from the ROSMAP project with both single cell RNA sequencing (scRNA-seq) data and whole genome sequencing (WGS) data available were selected for this study. **A.** The distribution of the number of snRNA reads across individuals. The dashed red line indicates the cutoff of <6×10^7 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 two batches (B1 and B2) of the snRNAseqMFC study. Individuals colored orange were from the snRNAseqAD_BA10 study, and individuals colored purple from the snRNAseqPFC_TREM2 study. **B.** The number of cells per cell type per individual. The cell types were distinguished with seven different colors (see legend). The colors of the edges indicated different studies, as in **A**. Abbreviation: ExNeurons, excitatory neurons; InNeurons, inhibitory neurons; OPCs, oligodendrocyte progenitor cells.

## 5.2.2. 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 to WGS data of 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, 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 single-nucleotide variants (SNVs). From these 196 sites, 98 were shared between multiple individuals n>2, and thus are recurrent somatic mutations (eFigure 4). 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).

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 ~12 somatic SNVs in hippocampal formation tissue using deep bulk exome sequencing[10], and an average amount of ~1700 somatic mutations (substitutions ~1500; indels ~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.

## 5.2.3. 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). 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 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). We should note that the number of samples is relatively low and represent a relatively narrow age range (from 74 to 90 years old). Moreover, participants with an age older than 90 years were all censored by age 90, which could also influence the significance of the age trend. A significant trend is observed when we exclude individuals at age 90 from the regression (β=0.37, P=0.005; eFigure 7).

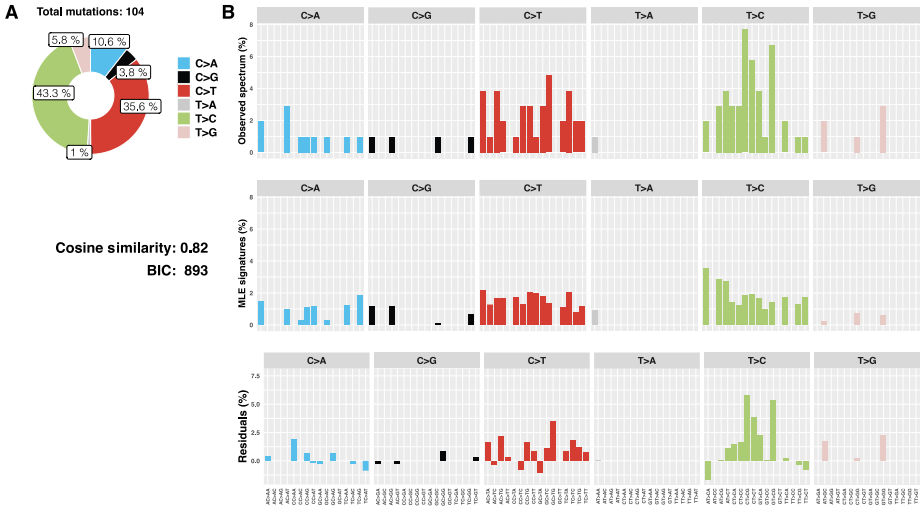



**Figure 2. The mutation signature of 104 putative excitatory neuron-specific single nucleotide variations (SNVs) in the brain.** Among the 30 COSMIC single base substitution (SBS) signatures, SBS5 was identified as the model that best explains the observed pattern of putative somatic SNVs by Mutalisk. The cosine similarity with the 104 putative excitatory neuron-specific SNVs and the corresponding Bayesian information criterion (BIC) for each COSMIC SBS signature are shown in eFigure 5. **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 panel. 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.

## 5.2.4. *RBFOX1* and *KCNIP4* harbor age-associating ENSMs

As several detected ENSMs are being detected in multiple individual genomes (eFigure 4), 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 (16:6899517 (*RBFOX1*), p=0.04; 4:21788463 (*KCNIP4*), p<0.05) are found to have significantly more mutations in older individuals. The age distributions in mutated and un-mutated samples for these two sites are shown in Figure 4. Some caution should be treated when interpreting this plot for individuals older than 90 years as these are all mapped to 90 years old. To assess the effect due to censoring on age, we performed a sensitivity analysis by removing all samples with an age ≥90. The results indicated stronger signals for these two sites (16:6899517 (*RBFOX1*), p=0.02; 4:21788463 (*KCNIP4*), p=0.03; eFigure 8).

## 5.2.5. 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 only detected 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's exact test that identifies sites that have a higher odds ratio to detect a somatic mutation in AD samples (Methods), yielded two sites with significant odds ratios. These sites are mapped to two genes (6:73374221 (*KCNQ5*), p=0.01 and 13:36667102 (*DCLK1*), p=0.02).

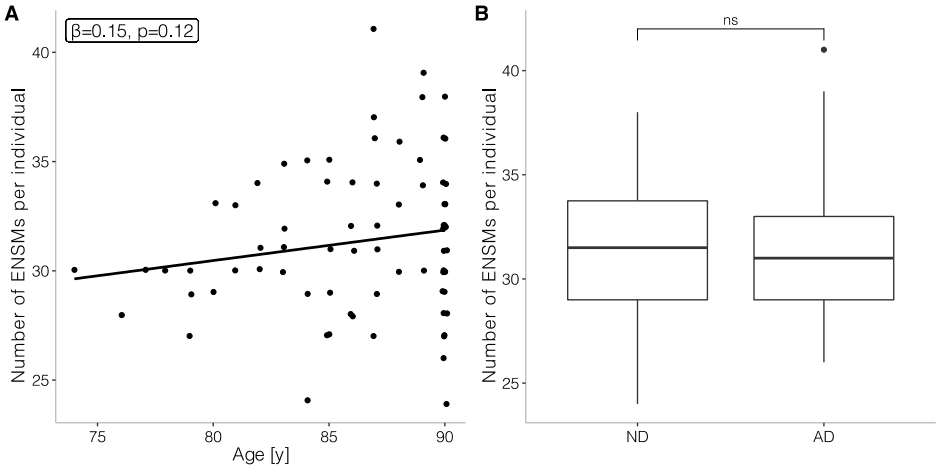



**Figure 3.** Quantitative comparison of the number of excitatory neuron-specific somatic mutations (ENSMs) in terms of AD and aging. A. The number of ENSMs per individual against the age of the individual. The line shows how this number regresses with age. The significance of the coefficient (β≠0) was tested using a t-test. The same analysis for AD and non-AD samples separately is shown in eFigure 6. B. Boxplot of the number of ENSMs in non-demented controls (ND) and AD patients (AD). The Wilcoxon rank sum test does not show a significance difference (ns).

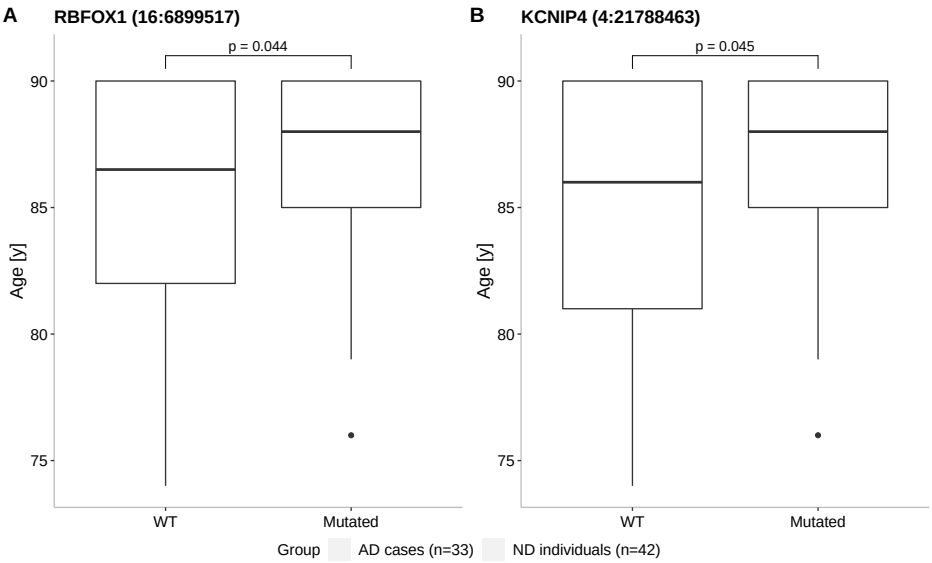



**Figure 4. The occurrence of somatic mutation with age in (A) *RBFOX1* and (B) *KCNIP4* genes.** Red dots: AD cases; blue dots: non-demented (ND) individuals. Logistic regression was used to test the prevalence of somatic mutations with increasing age.

## 5.2.6. 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). Among these 32 genes, there are several well-known AD-associated genes, like *SLC30A3*, *TTL*, and *CTSB*, which thus harbor somatic mutations unique for AD.

Together with the two 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). Also, “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 five 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]

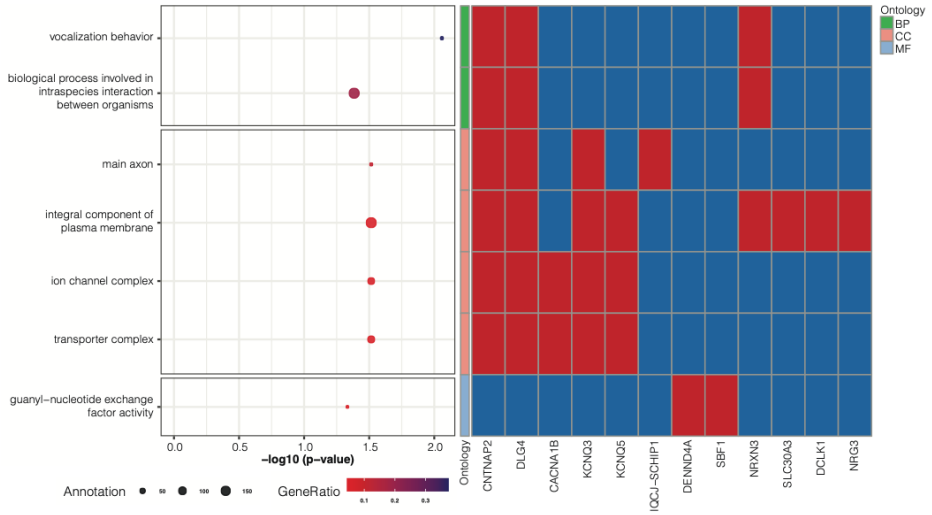



**Figure 5. GO-terms enriched with genes having AD-specific ENSMs.** 32 genes that have ENSMs only seen in AD samples, and the *KCNQ5* and *DCLK1* genes that have a higher occurrence in AD samples are used in the GO-term enrichment analysis. The left panel of the figure shows the enriched terms, their 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 FDR corrected significant GO-terms are grouped into three categories: Biological Process (BP), Cellular Component (CC), and Molecular Function (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.

## 5.3. Discussion

Late-onset Alzheimer's 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 neuropathological 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. Here, 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, suffers from 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 pseudo-bulk data. An exploratory run of this workflow revealed that we were only able to confidently detect somatic mutations for excitatory neuron as 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, like *SLC30A3*, *TTL*, and *CTSB*. *SLC30A3* is known to be down-regulated in the prefrontal cortex of AD patients.[38] *SLC30A3* is assumed to play a protective role against ER 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 to age-matched controls and that, in 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 AD patients.[44] Apart from these well-known AD-associated genes, we also identified that the *DCLK1* gene harbored more somatic mutations in AD patients. 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] Also, a dysfunction of $K^+$ channels has been observed in fibroblasts[49] and platelets[44] of AD patients. Additionally, Aβ has been demonstrated to not only be involve in the AD pathogenesis, but also modulate $K^+$ channel activities[50] and may have a physiological 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 expands the use and scope of scRNA-seq data, and may provide new insight into post-zygotic 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 above 90 were censored to be 90. These two factors may explain that we only found a relative weak association between age and the accumulation of somatic mutations. On the other hand, the significant trend after removing individuals with an age higher than 90 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 as a result of differences in gene expression). Therefore, many potential somatic mutations were excluded from the start. Also, 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 one or a few neurons. Finally, as 10x scRNA-seq data was 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.

# 5.4. 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), two longitudinal cohort studies of aging and dementia.[17] Information collected as part of these studies, collectively known as ROSMAP, includes clinical data, detailed post-mortem pathological evaluations and tissue omics profiling. The scRNA-seq data used in this project were from three sources: 1) snRNAseqMFC study (n=24), 2) snRNAseqAD_TREM2 study (n=32), and 3) snRNAseqPFC_BA10 study (n=48);

specifically, these three studies used single-nuclei RNA sequencing data. All specimens for these three scRNA-seq data sources were collected post-mortem from the frontal cortex, sub-regions might slightly differ between studies. The scRNA-seq data from the three studies were all sequenced according to the 10x Genomics manufacturer's protocol. Detailed information for cell partitioning, reverse transcription, library construction, and sequencing run configuration for the three studies is available on Synapse (snRNAseqMFC: syn16780177, snRNAseqAD_TREM2: syn21682120, snRNAseqPFC_BA10: syn21261143). WGS data was from a subset of the ROSMAP participants with DNA obtained from brain tissue, whole blood or lymphocytes transformed with the EBV virus. The details for WGS library preparation and sequencing, and WGS Germline variants calling were described previously.[18] The individuals (n=90) that have 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 non-demented (ND) controls; AD patients with or without other cause of cognitive impairment were defined as AD samples.

## Standard Protocol Approvals, Registrations, and Patient Consents

The ROS/MAP studies and sub-studies 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 dataset was separately processed for clustering and cell type annotation which was done as follows. The processed count matrix was loaded in Seurat (version 3.2.2). The data was log-normalized and scaled before analysis. Next, with the 2,000 most variable genes (default with Seurat), principal components analysis (PCA) was performed. The number of principal components used for clustering was determined using the elbow method. Further, Seurat's FindNeighbours and FindCluster functions were used, which utilizes Louvain clustering, the resolution was set at 0.5. A UMAP plot (eFigure 1) 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*), 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). 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 splitted 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 two-step process. HaplotypeCaller was run on each sample separately in GVCF 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 multi-sample VCF file. eFigure 2 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 <50000; 2) the quality of calling QUAL >100; 3) the quality by depth QD >2; 4) the strand odds ratio SOR <2; and 5) the strand bias Fisher's 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 10x Genomics single cell data, was used to detect somatic mutations. For single nuclei gene expression data, VarTrix requires a pre-called variant set in VCF format, an associated set of alignments in BAM or CRAM 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 pre-called variant set was also detected from the subset of the BAM file which only including 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.

Since the scRNA-seq data were collected from three 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 samples. A variant site would be used for somatic mutation detection when for all individuals the read depth at this site is higher than the cut-off $C_i$ for that individual. Next, a somatic mutation was identified as present in one 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 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.

## Mutation signature analysis

To characterize the contribution of mutation signatures, we pooled all putative somatic single nucleotide variations (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 (MLE) method; linear regression. The input file was compared with 30 single base substitution (SBS) signatures from the COSMIC mutational signatures database. The best model of signature combination was suggested from the tool by considering the Bayesian information criterion (BIC).

## 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 one 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 REVIGO with semantic similarity score “Lin”. The genes that were annotated to the variant sites with read depths higher than the cut-offs for all samples were used as background. The p-values from the uneliminated GO-terms were corrected using “Benjamini&Hochberg” method, 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: https://github.com/mzhang0215/ENSM_project. Wilcoxon rank sum test, linear regression, Fisher's 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: $p = 1/(1 + \exp(-(\beta_0 + \beta_1 age + \beta_2 group)))$, 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 group were used.

# References


1. Biesecker LG, Spinner NB. A genomic view of mosaicism and human disease. *Nat Rev Genet*. 2013;14(5):307-320. doi:10.1038/nrg3424

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). doi:10.1101/cshperspect.a025130

3. Wang X, Wang W, Li L, Perry G, Lee H gon, Zhu X. Oxidative stress and mitochondrial dysfunction in Alzheimer's disease. *Biochim Biophys Acta - Mol Basis Dis*. 2014;1842(8):1240-1247. doi:10.1016/j.bbadis.2013.10.015

4. Poduri A, Evrony GD, Cai X, Walsh CA. Somatic mutation, genomic variation, and neurological disease. *Science (80- )*. 2013;341(6141). doi:10.1126/science.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. doi:10.1016/j.coisb.2016.12.004

6. McConnell MJ, Moran J V., Abyzov A, et al. Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. *Science (80- )*. 2017;356(6336). doi:10.1126/science.aal1641

7. Kennedy SR, Loeb LA, Herr AJ. Somatic mutations in aging, cancer and neurodegeneration. *Mech Ageing Dev*. 2012;133(4):118-126. doi:10.1016/j.mad.2011.10.009

8. Burns A, Iliffe S. Alzheimer's disease. *BMJ*. 2009;338(7692):467-471. doi:10.1136/bmj.b158

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. doi:10.1016/j.jalz.2011.10.007

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). doi:10.1038/s41467-019-11000-7

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. doi:10.1016/j.coisb.2017.07.004

12. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med*. 2018;50(8):1-14. doi:10.1038/s12276-018-0071-8

13. Chen G, Ning B, Shi T. Single-cell RNA-seq technologies and related computational data analysis. *Front Genet*. 2019;10(APR):317. doi:10.3389/fgene.2019.00317

14. N M P, Liu H, Dillard C, et al. Improved SNV Discovery in Barcode-Stratified scRNA-seq Alignments. *Genes (Basel)*. 2021;12(10):1558. doi:10.3390/genes12101558

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. doi:10.1093/bioinformatics/btz288

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 101*. 2019;10(1):1-16. doi:10.1038/s41467-019-11591-1

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. doi:10.3233/JAD-179939

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. doi:10.1038/SDATA.2018.142

19. Mathys H, Davila-Velderrain J, Peng Z, et al. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature*. 2019;570(7761):332-337. doi:10.1038/s41586-019-1195-2

20. Van der Auwera G, O'Connor B. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. O'Reilly Media; 2020. Accessed December 14, 2022. https://www.oreilly.com/library/view/genomics-in-the/9781491975183/

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. Accessed January 27, 2022. https://academic.oup.com/nar/article/46/W1/W102/5001159

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. doi:10.4161/fly.19695

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. doi:10.1212/01.WNL.0000271090.28148.24

24. Gott JM, Emeson RB. FUNCTIONS AND MECHANISMS OF RNA EDITING. *https://doi.org/101146/annurev.genet341499*. 2003;34:499-531. doi:10.1146/ANNUREV.GENET.34.1.499

25. Gout JF, Li W, Fritsch C, et al. The landscape of transcription errors in eukaryotic cells. *Sci Adv*. 2017;3(10). doi:10.1126/sciadv.1701484

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. doi:10.1073/pnas.1525329113

27. Navin NE. Cancer genomics: one cell at a time. *Genome Biol*. 2014;15(8):452. doi:10.1186/s13059-014-0452-9

28. Abascal F, Harvey LMR, Mitchell E, et al. Somatic mutation landscapes at single-molecule resolution. *Nature*. Published online April 28, 2021:1-6. doi:10.1038/s41586-021-03477-4

29. Lodato MA, Rodin RE, Bohrson CL, et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science (80- )*. 2018;359(6375):555-559. doi:10.1126/science.aao4426

30. Miller MB, Huang AY, Kim J, et al. Somatic genomic changes in single Alzheimer's disease neurons. *Nature*. 2022;604(7907):714-722. doi:10.1038/s41586-022-04640-1

31. Brown DA, Passmore GM. Neural KCNQ (Kv7) channels. *Br J Pharmacol*. 2009;156(8):1185-1195. doi:10.1111/j.1476-5381.2009.00111.x

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):134385. doi:10.1371/journal.pone.0134385

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). doi:10.1016/j.neurobiolaging.2013.02.005

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 Biol*. 2019;20(1):242. doi:10.1186/s13059-019-1863-4

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. doi:10.1523/JNEUROSCI.4599-08.2009

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. doi:10.1093/hmg/dds240

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. doi:10.3389/fnmol.2020.00066

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. doi:10.1016/j.neurobiolaging.2014.06.015

39. Kurita H, Okuda R, Yokoo K, Inden M, Hozumi I. Protective roles of SLC30A3 against endoplasmic reticulum stress via ERK1/2 activation. *Biochem Biophys Res Commun*. 2016;479(4):853-859. doi:10.1016/j.bbrc.2016.09.119

40. Lewis SA, Cowan NJ. Tubulin Genes: Structure, Expression, and Regulation. In: *Microtubule Proteins*. CRC Press; 2018:37-66. doi:10.1201/9781351074643-2

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:e056351. doi:10.1002/alz.056351

42. Turk V, Bode W. The cystatins: Protein inhibitors of cysteine proteinases. *FEBS Lett*. 1991;285(2):213-219. doi:10.1016/0014-5793(91)80804-C

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. doi:10.1074/jbc.M109.024620

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. doi:10.1016/S0140-6736(98)03200-0

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. doi:10.1002/dneu.22428

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. doi:10.1016/j.neuron.2007.07.025

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-e2100. doi:10.1038/cddis.2016.18

48. Zaydman MA, Silva JR, Cui J. Ion Channel Associated Diseases: Overview of Molecular Mechanisms. *Chem Rev*. 2012;112(12):6319-6333. doi:10.1021/CR300360K

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. doi:10.1073/PNAS.90.17.8209

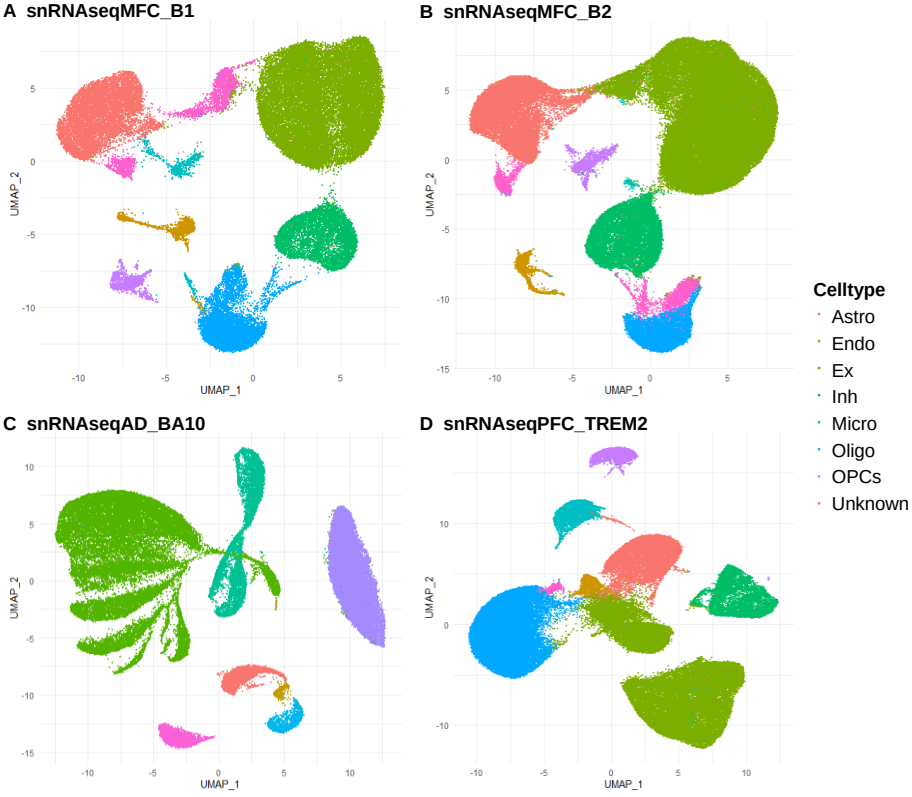
50. Plant LD, Webster NJ, Boyle JP, et al. Amyloid β peptide as a physiological modulator of neuronal 'A'-type K+ current. *Neurobiol Aging*. 2006;27(11):1673-1683. doi:10.1016/j.neurobiolaging.2005.09.038

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. doi:10.1016/S0006-8993(02)03547-3

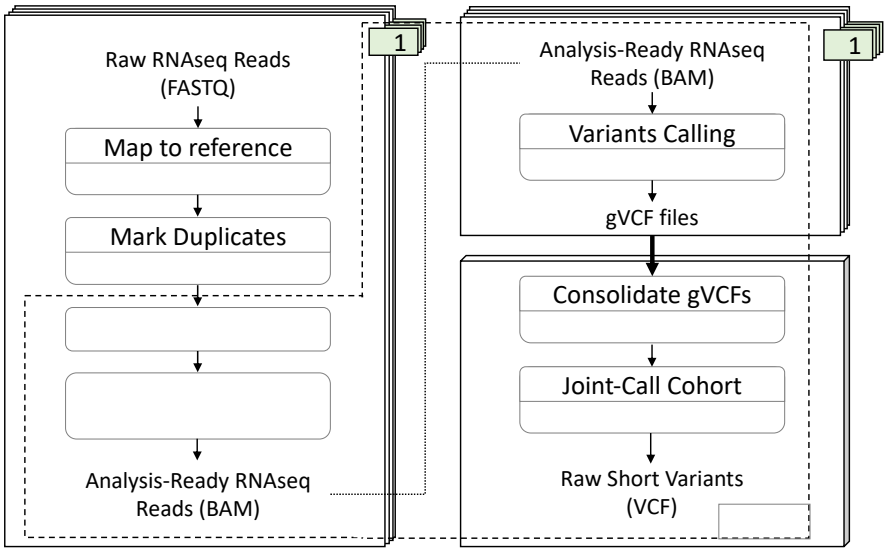
52. Sesti F. Oxidation of K+ channels in aging and neurodegeneration. *Aging Dis*. 2016;7(2):130-135. doi:10.14336/AD.2015.0901

53. Anagnostou ME, Chung C, McGann E, et al. Transcription errors in aging and disease. *Transl Med Aging*. 2021;5:31-38. doi:10.1016/j.tma.2021.05.002

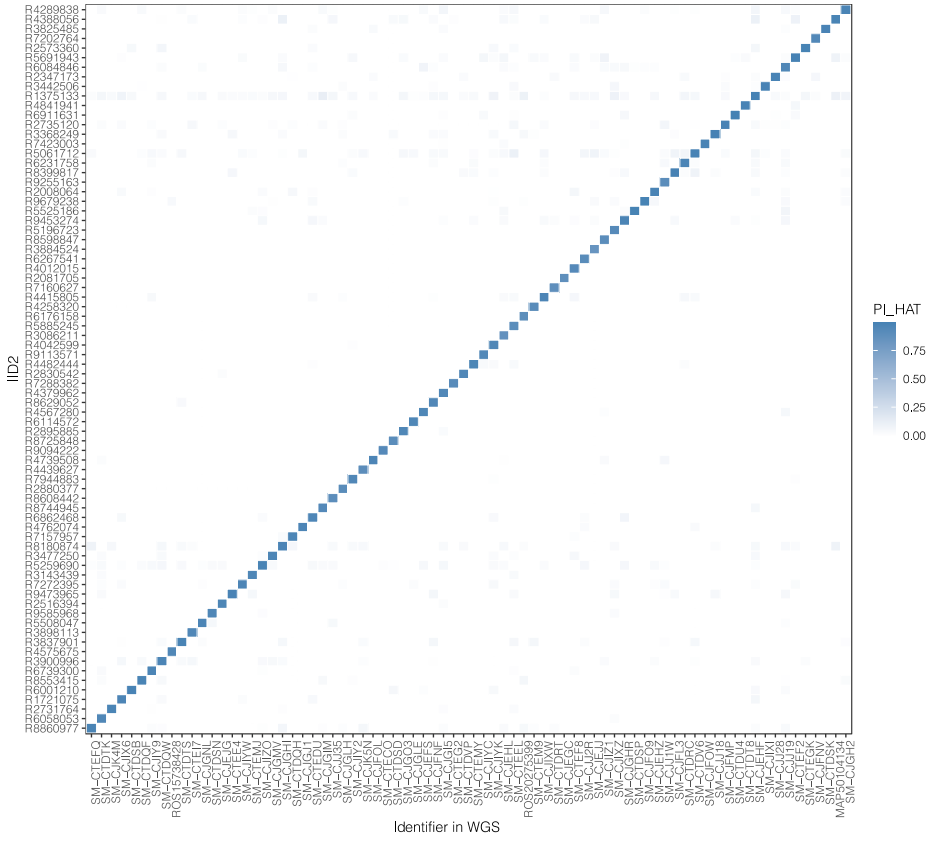
# Supplements



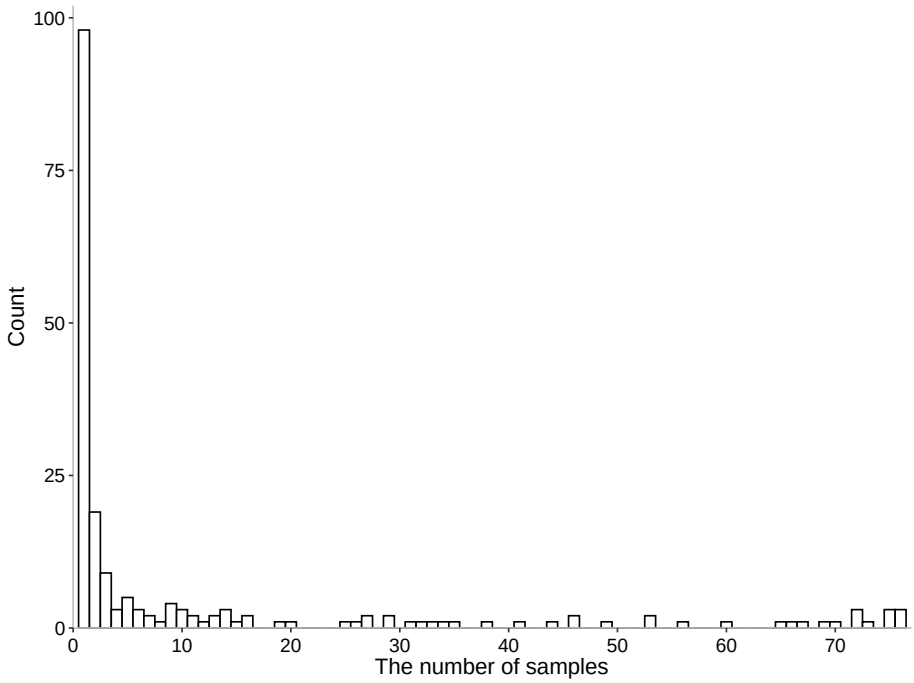
**Supplementary Figure 1**. The UMAP plot for cell type clustering in each study. The details for cell type clustering see Methods. Abbreviation: Astro, astrocytes; Endo, endothelial cells; Ex, excitatory neurons; Inh, inhibitory neurons; Micro, microglia; Oligo, oligodendrocytes; OPCs, Oligodendrocyte precursor cells.



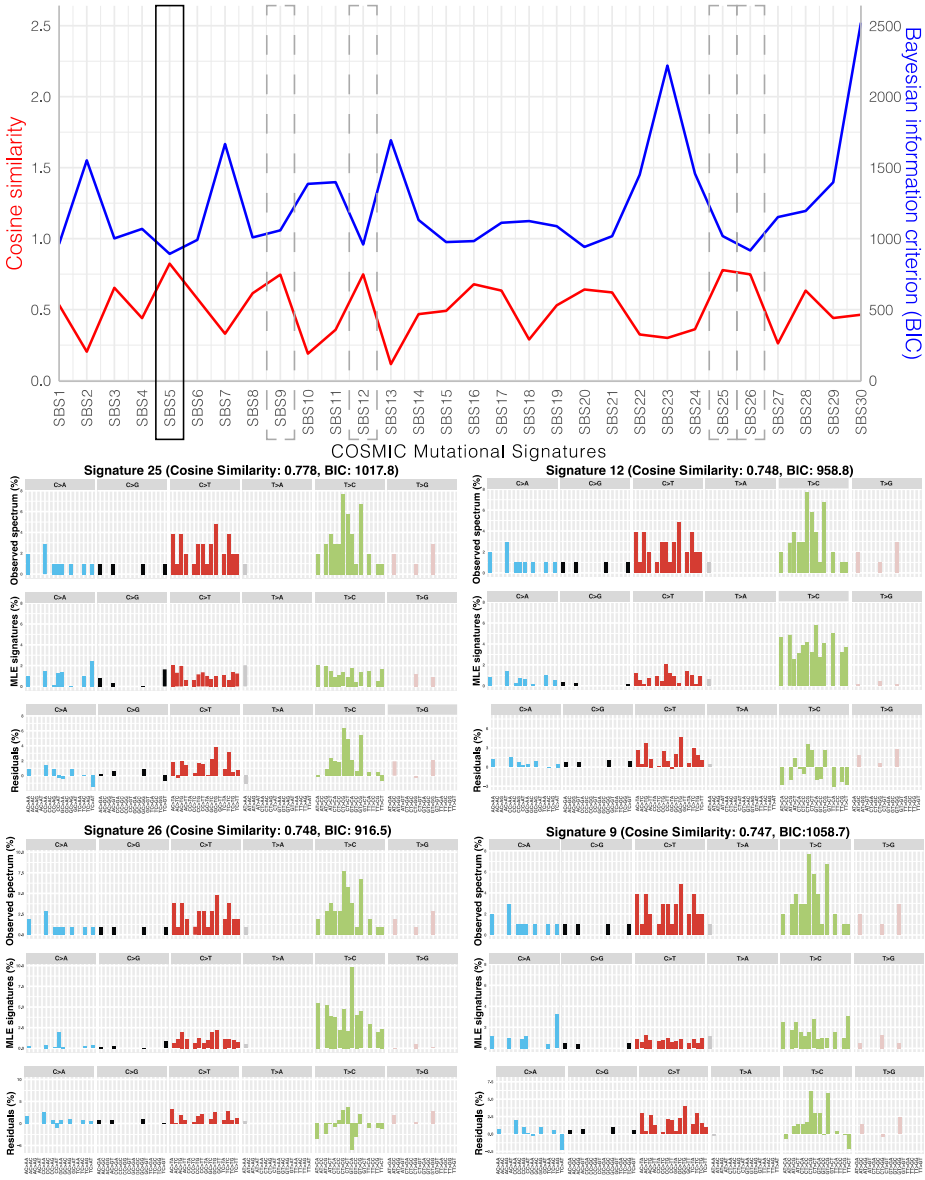

**Supplementary Figure 2.** The pipeline of short snRNA-seq variants calling. This pipeline follows the best practices workflows from GATK.



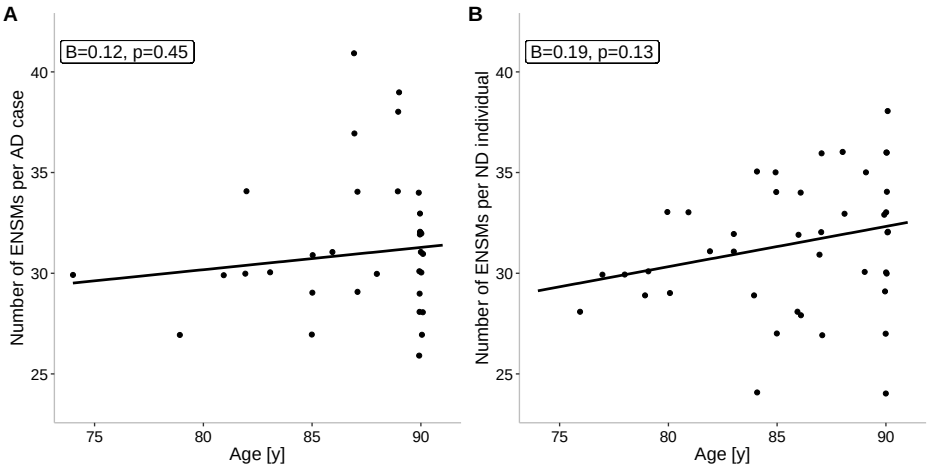

**Supplementary Figure 3.** IBD estimation between paired genetic profiles from WGS and snRNA-seq. The IBD estimation was performed to ensure the WGS and snRNA-seq profiles which share the same identifier were from the same individual. PI_HAT is a measure of overall IBD alleles. If the genetic profiles are from different persons, the value PI_HAT will be close to 0. On the contrary, if the profiles are from the same person, the value PI_HAT will be close to 1.



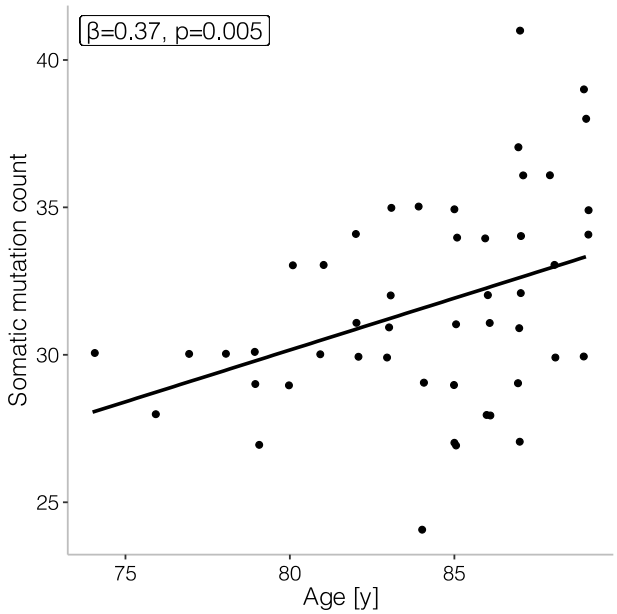

**Supplementary Figure 4.** Distribution of sample counts with excitatory neuron-specific somatic mutations at the same site.



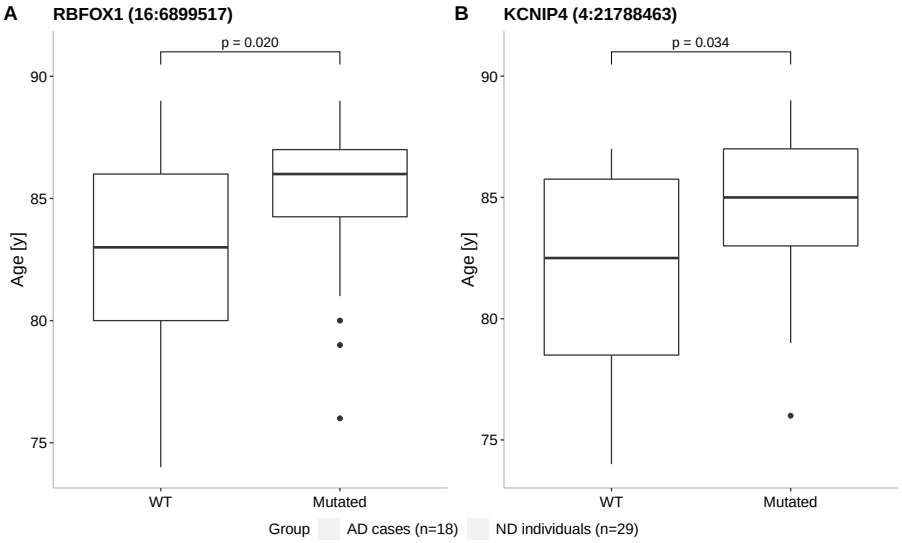

**Supplementary Figure 5**. The results of mutation signature analysis. In the top panel, we calculated the cosine similarity (red line) and the corresponding Bayesian information criterion (BIC, blue line) between the 104 putative excitatory neuron-specific SNVs and each of the 30 COSMIC SBS signatures using Mutalisk. Signature 5, highlighted in the top panel, showed the highest similarity and lowest BIC value. The other top 5 mutational signatures in cosine similarity (i.e., signatures 25, 12, 26, and 9, except for signature 5) were also highlighted in the top panel, and their mutation patterns were shown in the bottom panel.




**Supplementary Figure 6.** The count of mutations regressed with age in AD cases (A) and non-demented individuals (B).




**Supplementary Figure 7.** The count of mutations regressed with age. Individuals at age 90 were removed from this figure. The significance of the coefficient (β≠0) was tested using t-test.




**Supplementary Figure 8.** The occurrence of somatic mutation with age in (A) RBFOX1 and (B) KCNIP4 genes. As a sensitive analysis, samples with age ≥90 were excluded from these figures. Red dots: AD cases; blue dots: non-demented (ND) individuals. Logistic regression was used to test the prevalence of somatic mutations with increasing age.

# Cell-projected phenotypes link transcriptional and phenotypic heterogeneity in Alzheimer's disease

Gerard Bouland, Ahmed Mahfouz, Marcel Reinders and Manolis Kellis

# Abstract

Single-cell transcriptomics can identify disease driver genes and biomarkers with high resolution, revealing heterogeneity both between and within individual donors. However, current analysis methods often focus on creating reference maps of cell types, overlooking phenotypic variation. Recent advances in profiling large disease cohorts enable a new approach: directly associating each cell's transcriptional state with phenotypic variation between individuals. We achieve this by linking each cell to multiple transcriptional 'neighborhoods' and their corresponding phenotypic enrichments based on donor variables. This method connects molecular and disease-level variations within and between individuals. By using data from 3.4 million cellular transcriptomes across 599 donors, we discovered that not all cells from an individual reflect their phenotypic traits, with manifestations varying across cell types. Our findings include the identification of potential compensatory mechanisms in Alzheimer's disease, such as neurotransmitter switching, and elevated glucose and cortisol levels in excitatory neurons, possibly linked to neuronal hyperexcitability, as well as a neuroinflammation-regulating process.


# 6.1. Introduction

The majority of human diseases are multifactorial, affecting multiple tissues, cell types, and biological processes throughout the body. Even in monogenic disorders individual genes often participate in multiple pathways and biological processes, acting differently in various cell types and leading to diverse manifestations across individuals, tissues and cells. This complexity is further amplified in complex traits that are influenced by diverse genetic and environmental factors. Capturing this heterogeneity of dysregulation across cells is crucial for understanding the wide range of phenotypic manifestations across individuals. Such understanding can significantly impact personalized therapeutics, by guiding interventions based on each individuals' unique biological dysregulation. It can also help identify different subclasses of disease and their corresponding molecular manifestations, by linking phenotypic variation with potential therapeutic targets in a cell type-specific, pathway-specific, and tissue-specific way.

To enable this understanding, single-cell RNA sequencing (scRNAseq) experiments have the potential to enable detailed exploration of cellular heterogeneity at the transcriptomic level. However, current analysis paradigms implicitly assume that the cells from a given individual uniformly represent the phenotype of their respective donors, thereby ignoring within-individual cell variation. This results in the current single-cell analysis paradigm that first constructs reference maps of discrete cell types and subtypes, and then identifies differentially-expressed genes between "cases" vs. "controls" for each group or subgroup of cells, thereafter neglecting the cellular variation within individual donors. As recent single-cell cohorts include up to hundreds of individuals with diverse phenotypes[1–4], the opportunity arises to develop new analysis methods that exploit between-individual variation. Some methods[5,6] have been developed that utilize the multi-individual nature of these datasets, allowing the identification of cell states associated with disease by exploiting the phenotypic variation

among individuals without requiring predefined cell types. Although identifying phenotype-specific cell states offers valuable insights into the cellular impacts of phenotypes, it tends to overlook the complexity of disease heterogeneity as current approaches fail to explore the connection between variations in phenotypic traits and differences in cellular phenotypic manifestations. To address this, it is essential to determine which donors have cells in the respective cell states, as well as to identify other phenotypic characteristics of donors that associate with these cell states.

Alzheimer's Disease (AD) is particularly amenable to such an approach due to its urgency and major health impact, its highly heterogeneous nature, and the recent availability of two large-scale scRNAseq studies by our group[1] and the Jager study[2] involving 599 unique donors together. AD is characterized by progressive neurodegeneration, loss of cognitive functions, and ultimately death[7], but its complexity includes multiple phenotypic signatures implicated in AD progression, including the amyloid-beta (Aβ) cascade, tau, inflammation, and oxidative stress[8]. These diverse hypotheses suggest heterogeneity among individuals, and that the targeting of any one pathway in therapeutic development may lead to the currently-observed low success rates in clinical trials[9], as different pathways may be dysregulated in different individuals. In addition, AD pathology does not uniformly affect the brain, instead beginning in the lower brainstem, progressing next to subcortical regions, and ultimately to the neocortex. This sequential progression, reflected in Braak staging[10], has the implication that at any given time point, different brain regions in an individual, and likely different cells within them, will exhibit very different levels of AD pathology.

Extending these observations to the cellular level, we introduce an analysis framework to calculate cell-projected phenotypes, a new approach for analyzing population-scale, multi-condition scRNAseq datasets. Departing from the conventional practice of assuming cells uniformly represent the phenotypic characteristics of the respective individual, we assign phenotype scores to each cell, conceptualizing disease involvement as a spectrum rather than a fixed classification. This approach provides a more nuanced understanding of AD, reflecting its complex and heterogeneous nature.

Using cell-projected phenotypes and a total of 3.4 million scRNAseq cellular transcriptomes across 599 unique AD and age-matched control individuals, we investigated the associations of cell phenotypic manifestations with donor-level phenotypes as well as gene regulation circuitries. We discovered that individuals diagnosed with AD have subsets of cells that are transcriptionally similar to healthy cells from non-affected individuals. We observed distinct phenotypic manifestations varying in intensity across different cell types. Additionally, we identified nine distinct AD components, each linked to different cell types and clinical manifestations. Characterizing these distinct components, we identified unique sets of genes and pathways associated. This integration of single cell gene expression with population-level phenotype information and molecular data opens up new possibilities for better understanding the complex heterogeneity of human disease.

# 6.2. Results

## 6.2.1. Cell-level Phenotypic Projections

Here, we present a new analysis framework to get cell-projected phenotypes for analyzing population-scale, multi-condition scRNAseq data (Fig. 1d). We use a 'guilt-by-association' strategy to systematically assign phenotype association scores to individual cells. Unlike traditional approaches that categorize cells based on the phenotypic values or categorical disease states of the donor—thereby oversimplifying the complexity of biological systems—this framework calculates phenotype scores for each cell, allowing for a spectrum of phenotype or disease involvement rather than a fixed classification.

Briefly, our approach consists of the following steps:

First, we define partially overlapping cellular neighborhoods based on transcriptional similarities (using miloR[5]). To ensure cellular neighborhoods consist of multiple individuals, neighborhoods with cells coming from fewer than three individuals were removed. Additionally, to prevent any single individual's phenotypic trait from dominating the results, we downsample individuals with an excessively high number of cells, as this could disproportionately increase the representation of their cells within any given neighborhood. These neighborhoods are derived from the KNN graph and typically results in a distribution centered at roughly 50 cells per neighborhood (Supplementary Fig. 1).

The second step involves calculating the association between these neighborhoods and the phenotypic traits of the individuals, while correcting for confounding variables(Supplementary Table 1). For example, if 40 out of 50 cells in a neighborhood originate from individuals diagnosed with AD, the neighborhood is assigned a high positive AD association score. Conversely, if 40 out of 50 cells originate from non-AD individuals, the neighborhood receives a negative AD association score.

And third, the phenotype associations of the neighborhoods are propagated to the individual cells. On average, cells are part of approximately five neighborhoods(Supplementary Fig. 2). We reason that the closer a cell is to the center of a neighborhood, the more accurately the neighborhood's phenotypic enrichment applies to the cell. Therefore, for each cell, we calculate the distance to the centers of its neighborhoods and apply a fading-membership function to these distances that allows to assign higher levels of importance to the nearest neighborhoods. Finally, the cell’s phenotypic score is determined as the weighted average of its associated neighborhoods. By calculating cell-projected phenotypes we assume that most cells align with the characteristics of the donors, which means that, on average, individuals diagnosed with AD have more cells with elevated AD scores than individuals with no AD diagnosis.

Given current technologies and available datasets, we used transcriptional definitions of cellular states. However, this approach is generalizable and can be applied to single-cell proteomics and future single-cell profiling of metabolic, lipidomic, or other cellular states.

## 6.2.2. Cell-projected phenotypes reveal intra individual cellular heterogeneity of phenotype manifestation

In this study, we used a total of 3.4 million scRNAseq cellular transcriptomes across 599 unique AD and age-matched control individuals. The cellular transcriptomes were independently profiled in two cohorts[1,2] and included 228 individuals that were independently profiled by both cohorts. To prevent double-dipping and to provide a measure of generalizability and robustness to our results, the cellular transcriptome that were profiled by our group[1] were used for defining the cellular transcriptional neighborhoods, and the Jager data[2] was used to derive additional gene expression-informed measures to associate with the cell-projected phenotypes. Furthermore, the other dataset was used to validate the projected cellular phenotypes.

We defined 171,470 cellular transcriptional neighborhoods using 1.8 million scRNAseq cellular transcriptomes across 397 AD and age-matched control individuals(Fig. 1a-c). We projected eight donor-level characteristics onto individual cells: AD diagnosis, amyloid, tangle density, neuritic plaque load, cognitive impairment (CI), cognitive resilience (CR) APOE e4 status, and sex (Fig. 1c, Supplementary Fig. 3). For each cell, we computed a score indicating how much it is related to each of these donor-level characteristics. With cell-projected phenotypes we are able to distinguish cells that are ‘representative of’ and ‘not representative of’ the phenotypic trait of the respective donor, thereby offering a more nuanced perspective of cells, phenotypes and their association within and between individuals.

To illustrate that we can distinguish cells ‘representative of AD’ from those ‘not displaying AD characteristics’, we analyzed cell-projected AD diagnosis across 102,991 astrocytes derived from 322 donors, including 182 diagnosed with AD and 140 with no AD diagnosis. Our analysis revealed 34k cells exhibiting AD characteristics (AD score≥ 0.25), 36k healthy cells (AD score ≤ -0.25), and 33k neutral cells (|AD score| ≤ 0.25). As expected, the vast majority of AD-like astrocytes (79%) came from diagnosed individuals, and the large majority of healthy cells (66%) were from donors with no AD diagnosis. However, 34% of healthy-like astrocytes were from AD donors, and 21% of AD-like astrocytes were from non-AD donors (Supplementary Fig. 4).These findings suggest that, although the transcriptional states of cells generally align with the clinical diagnoses of the donors, individuals diagnosed with the disease still possess cells that exhibit a healthy-like state. Conversely, individuals without an AD diagnosis do have some cells displaying transcriptional profiles characteristic of AD, potentially indicating early, preclinical signs of pathology.

To validate the transcriptional disease states of these cells, we compared gene expressions between AD astrocytes from diagnosed donors and healthy astrocytes from non-AD donors, and vice versa. We find 1,124 genes significantly differentially expressed ($P_{bonf}$≤0.01,Supplementary Table 2) when comparing healthy donor-healthy cell vs. AD donor-AD cell. A ~35-fold decrease in the number of differentially expressed genes (Supplementary Table 3) was observed when comparing healthy cells from diagnosed and non-AD donors, supporting that healthy-like cells in diagnosed individuals resemble those in non-AD donors

and underscoring that an individual's phenotypic trait is not uniformly reflected across all their cells.

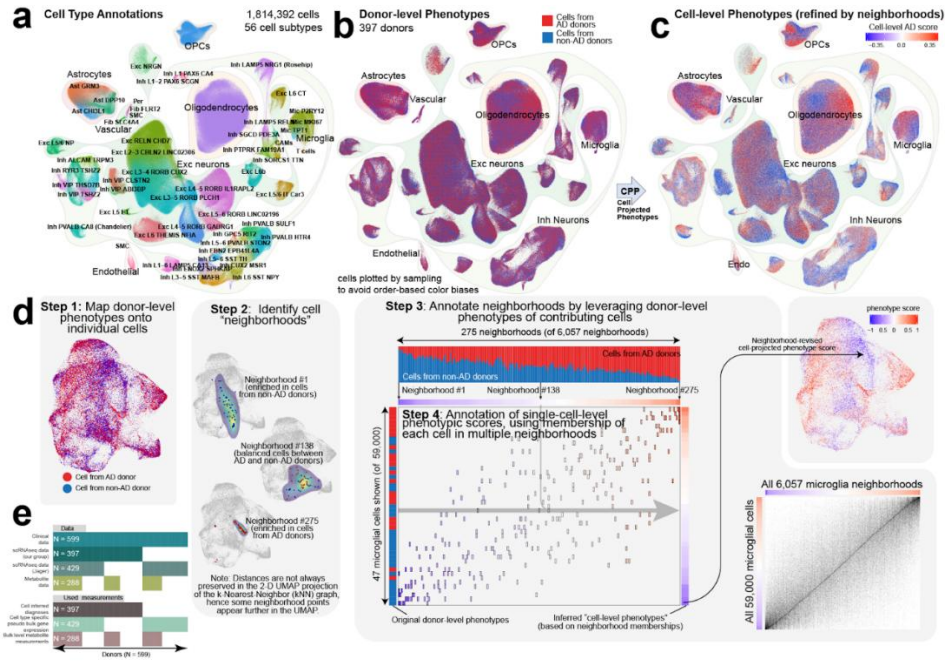



**Figure 1: a, b, c,** Uniform manifold approximation and projection (UMAP) of the main scRNAseq datasets used for calculating the cell-transcriptional neighborhoods where cells are colored by coloured by **(a)** sub cell type **(b)** diagnosis of the donor and **(c)** cell-projected AD-scores where blue represents a non-AD cell state and red represents an AD cell state, capped at AD-scores of -0.35 and 0.35**. d,** The computational steps required to calculate cell-projected phenotypes using microglial cells as an example. **(1)** We start with individual cells, here illustrated with microglia, visualized in a UMAP where blue represents cells from non-AD individuals and red represents cells from AD individuals. **(2)** Next, we identify neighborhoods of transcriptionally similar cells. For microglia, there were 6,057 transcriptional neighborhoods, but for illustration, we show three: Nh#1 (predominantly non-AD cells), Nh#138 (approximately 50/50 non-AD and AD cells), and Nh#275 (mostly AD cells). **(3)** We index the donors contributing cells to these neighborhoods to calculate enrichment scores using their phenotypic characteristics, depicted with a cell(rows)-neighborhood(columns) membership matrix. **(4)** We then calculate a weighted average for each cell based on the phenotype enrichment score of its neighborhoods, assigning higher weights to neighborhoods closely resembling the cells.Finally, we present the complete cell(rows)-neighborhood(columns) membership matrix with 6,057 neighborhoods and 59,000 cells, and display the calculated cell-projected phenotypes in a UMAP. **e** Summary of the included datasets, ROSMAP participants and used measurements **e,** Overview of dataset used.

## 6.2.3. Cell type-specific phenotypes uncover differences in phenotype manifestations across cell types

Given our observation that the phenotypic neighborhoods of individual cells from the same donor can vary greatly, we aggregated these cell-level scores at the cell-type-level across 52 cellular subtypes, to recognize cell-type-level manifestation of phenotypes.(Fig. 2a, Supplementary Tables 4-12). We reasoned that if phenotypic neighborhoods of individual cells of a certain cell type always

correspond to the phenotypic characteristics of the donors, then the phenotype is fully manifested in that cell type.

We observed that AD status (Fig. 2b, r=0.81) and tangle density (Fig. 2b, r=0.84) were most effectively represented by the cellular phenotypic states of microglia P2RY12, underscoring the critical involvement of microglia and neuroinflammation in AD pathology[11,12]. Reflecting recent literature on the pivotal roles of oligodendrocytes and astrocytes in cognitive function, cognitive impairment (Fig. 2b) was best represented in the cellular phenotypic states of oligodendrocytes[13](r = 0.81) and astrocytes[14](r≥0.67). And, in agreement with evidence that OPCs are differentially influenced by sex hormones[15] and show transcriptional distinctions between male and female[16], sex was almost perfectly captured in the cellular phenotypic states of OPCs (Fig. 2b, r=0.91). These findings highlight the utility of cell-projected phenotypes in distilling cell type-phenotype associations, indicating distinct manifestations of phenotypes across different cell types.

Recognizing the circular reasoning of this analysis, we sought to assess the robustness and generalizability of cell type-inferred phenotypes by examining their capability to predict donor-level phenotypes of new, previously unseen donors, using transcriptional profiles of cells from the Jager dataset[2]. We began by training ridge regression models on our dataset, using cell type-inferred phenotypes as predictors and donor-level phenotypes as outcomes(Fig. 3c,d). For each new cell from the unseen donors, we calculated the transcriptional distances to the previously constructed cell neighborhoods and projected the phenotype association scores from the nearest neighborhoods onto these new cells in a weighted fashion. We then computed "predicted" cell-type-inferred phenotypes by averaging the projected cell phenotype scores for each new donor, categorized by cell type. These predicted cell-type-inferred phenotypes (Supplementary Tables 13-12) were subsequently inputted into the pre-trained ridge regression models.

The association between predicted and measured phenotype was highly dependent on the phenotype being predicted. For biological sex (Fig. 3c, $p=1x10^{-172}$) and cognitive impairment (Fig. 3d, $p=5x10^{-15}$) the association between predicted and true phenotype showed a highly significant association. Amyloid-beta load, tangle density, neuritic plaque (Supplementary Fig. 5a, $p=4x10^{-7}$) and AD diagnoses also showed strong significant associations (Fig. 3c,d, $p\leq4x10^{-7}$), while APOE e4 status (Supplementary Fig. 5b, p=0.88) and age (Supplementary Fig. 5c, $p=3x10^{-3}$) were more difficult to predict. We reason that the variability in associations arises from the way different phenotypes manifest in the prefrontal cortex. Consistent with the highly critical function of the prefrontal cortex for cognitive functioning[17–19], the transcriptional signatures of cells in this region correlate with cognitive performance and are therefore predictive. In contrast, the prefrontal cortex is pathologically more heavily affected in the later stages of Alzheimer's[10,20] and early-stage increases in amyloid levels, tangle density, and neuritic plaque load may not be associated with detectable pathological changes in the prefrontal cortex, explaining the reduced strength in association for these measures, which are here averaged across the brain.

## 6.2.4. Cell type-specific phenotypes reveal associations with gene expressions obscured by donor-level phenotypes

We next reasoned that the full extent of phenotype associated transcriptional disturbances might not be captured by donor-level phenotypic assessments, due to the cell-type-specific phenotypic shifts that we are observing across individuals (Fig. 2b).

As an example, for AD-donor #20156469, the transcriptional impact varied significantly across different cell types (Supplementary Fig. 6). The microglia P2RY12 were relatively unaffected, with an AD score at the 51st percentile (AD score = 0.01). In contrast, the oligodendrocytes and astrocyte GRM3 were highly transcriptionally affected, with AD scores at or above the 97th percentile (AD score ≥ 0.74). Conversely, for AD-donor #50106730, we observed the opposite pattern (Supplementary Fig. 7): the oligodendrocytes and astrocyte GRM3 were relatively unaffected with AD scores at the 65th (AD score = 0.08) and 34th (AD score = -0.15) percentiles, respectively. While the microglia P2RY12 showed significant transcriptional impact, scoring in the 98th percentile (AD score = 0.55). These two donors show us that the transcriptional impact varies across different cell types in different individuals. Using cell-projected phenotype we can exploit these differences of phenotype manifestation between individuals for differential expression analyses.

To illustrate this, we performed a differential expression analysis with oligodendrocyte-inferred tangle density (Supplementary Table 22) and compared it with a differential expression analysis performed with donor-level measured tangle density(Fig. 2f, Supplementary Table 23). To prevent double-counting, we associated the oligodendrocyte-inferred tangle density score with pseudo bulk gene expression data generated using oligodendrocytes from different brain samples of overlapping donors (n=178) from the Jager data[2]. Associated with oligodendrocyte-inferred tangle density, we observed more pronounced changes in gene expression and a greater number of significant genes($P_{bonf} \leq 0.01$, N=703) compared to donor-level measured tangle density(N=5, Fig. 2e). The effect sizes between the two analyses were significantly correlated(r=0.65, Supplementary Fig. 8) and all genes we identified as significantly differentially expressed with donor-level measured tangle density were also significantly differentially expressed associated with oligodendrocyte-inferred tangle density. Notably, *METTL7A* (Fig. 2f), a methyltransferase, exhibited the most significant expression change associated with oligodendrocyte-inferred tangle density and was not significant based on the donor-level measured tangle density analysis. Although *METTL7A*'s role in AD and tau aggregation is not well-studied, methylation and methyltransferases are recognized as key regulators of tau aggregation and neuronal health in AD[21].

Altogether, these results demonstrate the ability of cell-projected phenotypes to identify phenotypic shifts and the advantage of accounting for these shifts, thereby enhancing our understanding of the relationship between phenotype, cell type, and gene expression.

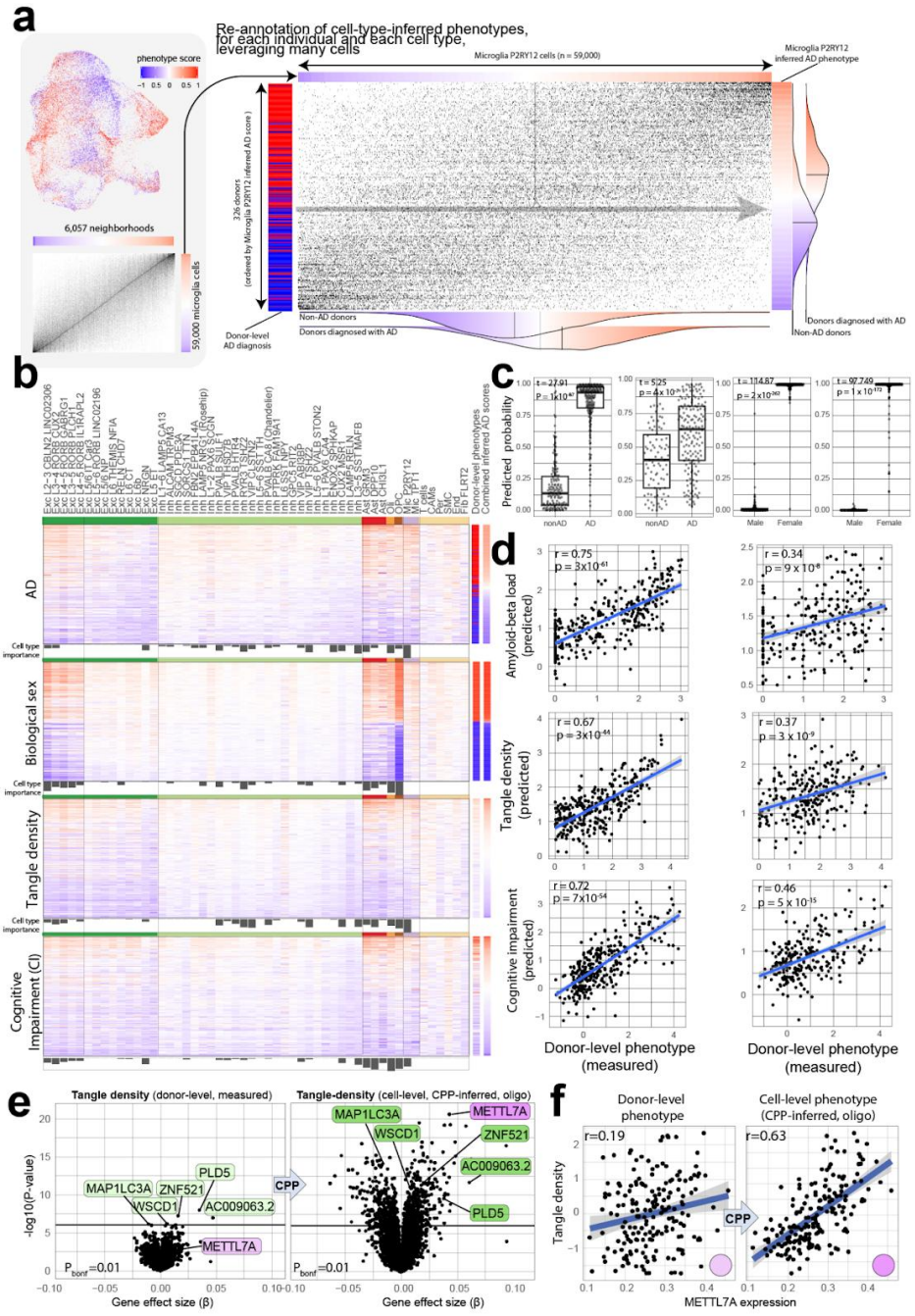



**Figure 2: a,** The computational steps required to calculate cell-type inferred phenotypes. **(1)** Starting with the cell-projected phenotypes from **Fig 1d. (2)** For each donor (rows), we calculate the average using all their cells (columns), displaying the population averages for cell-projected AD scores and microglia-inferred AD scores. Red indicates a high AD-score, and blue indicates a negative AD-score. **(3)** This process is repeated for each cell type,resulting in **(b)**.**b,** Fifty-two cell-type-inferred phenotype (AD, sex, tangle density and cognitive impairment) scores (columns) across 326 donors (rows),

ordered by the cell inferred phenotype score. High scores for each cell type are represented in red, and low scores in blue. The combined cell inferred score, displayed on the right, is derived from predicted donor-level phenotypes using the cell inferred scores as input. Below the plot, cell type importance is illustrated, calculated as the Spearman correlation coefficient between the cell inferred scores and the combined cell inferred score. Values below or equal to the median are set to zero, with the remaining values scaled from 0 to 1. **c,** Associations between reported donor-level phenotypes (x-axis) of 326 donors (left) and 245 previously unseen donors (right) with predicted donor-level phenotypes (y-axis), using the predicted cell inferred phenotype scores of AD diagnoses and sex. **e,** Associations between reported donor-level phenotypes (x-axis) of 326 donors (left) and 245 previously unseen donors (right) with predicted donor-level phenotypes (y-axis), using the predicted cell type inferred phenotype scores of cognitive impairment, amyloid-beta load and tangle density. **e,** Results of the differential expression analysis of reported donor-level tangle density (left) and oligodendrocytes inferred tangle density (right) with genes expressed in oligodendrocytes presented in a volcano plot with gene effect size (β) reported on the x-axis and -log10(p) reported on the y-axis. **g,** Association between *METTL7A* expression on the x-axis and reported donor-level tangle density and oligodendrocytes inferred tangle density on the y-axes.

## 6.2.5 Cell type-centric components of Alzheimer's disease and their association with pathological and cognitive manifestations

Calculating the cell type-specific AD scores, we observed that groups of cell types exhibit similar AD-phenotypic shifts while simultaneously showing different shifts compared to other groups of cell types within the same individuals (Fig. 2b). For instance, in the same AD donors from the previous example, donor #20156469 shows both astrocytes-GRM3 and oligodendrocytes affected by AD, while in donor #50106730, both cell types remain relatively unaffected. In both donors, this does not translate to the microglia-P2RY12 being affected to a similar degree. This observation led us to hypothesize the existence of multiple AD components, with individuals affected to varying degrees in different cell types, where each component may be characterized by specific donor-level phenotypes.

To identify these AD components, we performed a correlation analysis between the cell type-specific AD scores (Supplementary Fig. 9). Indeed, we observed that AD scores in oligodendrocytes were strongly correlated with AD scores in astrocyte-GRM3 cells across all donors ($r=0.85$, $p=8x10^{-91}$), while showing a much weaker correlation with AD scores from e.g., Inh L1-2 PAX6 SCGN cells ($r=0.17$, $p=2x10^{-3}$). Clustering the cell types based on their similarity of AD scores across individuals resulted in the discovery of nine AD components (Supplementary Fig. 10, Fig. 3a).

Next, we investigated the association between these AD components and donor-level phenotypes. Given that the phenotypes are used to diagnose AD and the diagnosis itself informs the calculation of the AD scores, we analyzed only AD donors (N = 177) to prevent circular reasoning and the inflation of associations.

First, we observed that the combined global AD score correlated with all included donor-level phenotypes: global pathology ($r=0.21$, $p=6x10^{-3}$), amyloid-beta load ($r=0.27$, $p=2x10^{-4}$), tangle density ($r=0.29$, $p=9x10^{-5}$), neuritic plaque ($r=0.16$, $p=3x10^{-2}$), and cognitive impairment ($r=0.29$, $p=9x10^{-5}$). However, we also identified outlier donors (Fig. 3b) whose combined global AD scores did not align with their diagnoses, which may suggest the presence of cognitive resilience or potential misdiagnoses.

Further, we observed that the components individually were associated with distinct AD-related pathological characteristics(Fig. 3c). Specifically, tangle density (Fig. 3d) was significantly correlated ($P_{bonf} \leq 0.01$, $r \geq 0.29$) with the Astro-Oligo(x4), Exc(x5), Exc(x8), and Mic(x2) components. Amyloid-beta load (Fig. 3e) was significantly correlated only with the Mic(x2) component, and cognitive impairment (Fig. 3f) was significantly correlated only with the Astro-Oligo(x4) component.

These findings demonstrate that increased AD-related transcriptional disturbances in astrocytes and oligodendrocytes are associated with more severe cognitive impairment and higher tangle density. In contrast, disturbances in microglia correlate with higher amyloid-beta load and tangle density. This indicates that different pathological facets of AD are linked to transcriptional disturbances in distinct cell types. This variation suggests that AD might be best understood as a series of distinct but overlapping components, each connected to specific cellular dysfunctions and specific pathological characteristics. These insights not only redefine our understanding of AD pathology but also highlight the potential significance of targeting these specific cellular changes in developing future therapies.

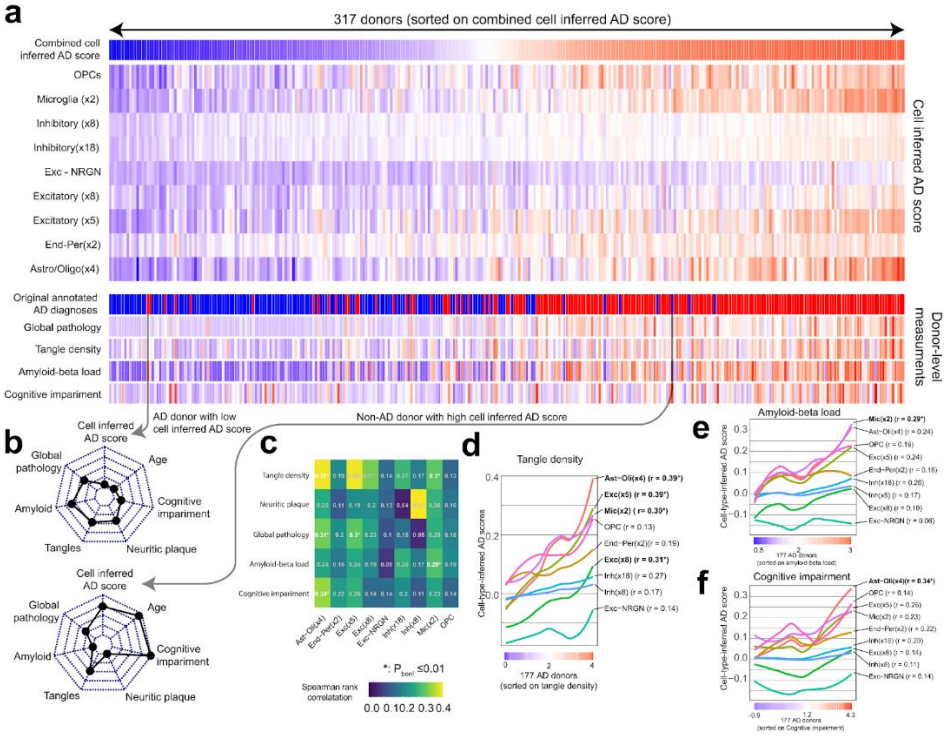



**Figure 3**: **a,** Nine cell-type-inferred AD-scores and five reported donor-level phenotypes (rows) of 317 donors(columns), sorted on the combined AD scores (top row). Red indicates a high value and blue a low value. For the original reported AD diagnoses blue is undiagnosed and red is diagnosed with AD. **b,** Relative values of reported donor-level phenotypes of the two most extreme outlier individuals, where the combined AD-score did not match the reported donor-level diagnosis. **c,** Associations (spearman rank correlation) between reported donor-level characteristics (rows) and the

cell-type-inferred AD-scores (columns). An * indicates significant association at p≤0.01 after Bonferroni correction. **d,e,f,** The increase of the cell-type-inferred AD-scores (y-axis) associated with increasing reported donor-level characteristics (x-axis); tangle density **(d)**, amyloid-beta load **(e)** and cognitive impairment **(f),** of only AD individuals visualized by sorting the individuals on their reported donor-level characteristics and a smoothed (LOESS) line representing each of the cell-type-inferred AD-scores. The spearman's rank correlation coefficients reported in the figures are calculated on the actual underlying (non-smoothened) data.

## 6.2.6. Cell-type-specific Alzheimer components linked to transcriptional alterations in distinct processes

To gain better insight into what might be driving these cell-type-specific AD components, we conducted differential gene expression analyses. For each AD component identified in our dataset, we sought differentially expressed genes in pseudobulk data from the de Jager data, focusing on the contributing cell types and overlapping individuals (Supplementary Table 24). We then performed gene set enrichment analyses for each set of differentially expressed genes (Supplementary Table 25). Our findings reveal that pathways related to neurotransmitters, metabolism, neurodegeneration, the immune system, and metal homeostasis are predominantly disrupted associated with varying cell-type-specific AD components (Fig. 4a, Supplementary Fig. 11). Notably, we found evidence of neurotransmitter switching, neuronal hyperexcitability, and TRP channel-associated inflammatory pathology, all of which may serve as potential therapeutic targets.

## 6.2.7. Neurotransmitter dynamics and potential switching associated with the inhibitory-neuron components

Specifically, we identified six AD components significantly associated with transcriptional alterations in neurotransmitter-related pathways ($P_{bonf} \leq 0.01$, Fig. 4b). Building on this observation, we hypothesized that these components of AD might be associated with alterations in neurotransmitter abundance itself. To test this hypothesis, we identified a metabolite dataset measured from the dorsolateral prefrontal cortex of 521 donors, 160 of whom overlapped with our study. Our analysis revealed significant changes in the abundance of one neurotransmitter (arachidonoyl ethanolamide), four neurotransmitter precursors (tyrosine, choline, glutamine, and tryptophan), and cortisol, which regulates various neurotransmitters (Fig. 4c).

We also detected sings of neurotransmitter switching. Specifically, choline levels were significantly downregulated with increasing pathology in the inhibitory-neuron component of AD (Fig. 4c,d; $r = -0.29$, $P = 2x10^{-4}$). In alignment with a known compensatory mechanism[22], we observed upregulation of a gene encoding a choline transmembrane transporter (Fig. 4e; SLC44A1, $\beta=0.22$, $P = 1x10^{-7}$), however surprisingly, this was in parvalbumin-positive neurons (Inh PVALB HTR4). Parvalbumin-positive neurons typically utilize GABA for inhibitory signaling. However, the increased expression of SLC44A1 suggests that these neurons may switch to using acetylcholine, for which choline is a precursor. This process, termed neurotransmitter switching [23,24], can occur under stress or toxicity conditions and has been documented in Parkinson's disease [25]. The cholinergic system, crucial for cognitive function[26], is notably vulnerable in

Alzheimer's disease[27,28]. Parvalbumin-positive neurons may adopt excitatory roles in an attempt to maintain the critical balance between excitatory and inhibitory signaling in the brain.

Consistent with literature[29–31], together, these findings suggest a widespread dysregulation of neurotransmitter processing and signaling in AD pathology.

## 6.2.8. Elevated brain-glucose and cortisol levels linked to the excitatory neuron component indicate neuronal hyperexcitability

We also observed many metabolic-related pathways significantly associated with seven AD components (Fig. 4f). Oxidative phosphorylation exhibited aberrant gene expression in 16 sub-cell types linked to the Inh (x18) and Exc (x5) components of AD, underscoring the well-known role of oxidative phosphorylation in AD.

Further, in our investigation of the excitatory neuron-centric components of AD, we identified dysregulation in the cAMP signaling pathway associated with neuronal hyperexcitability. Hyperexcitability is driven by reduced GABA levels[32], leading to decreased inhibitory control over excitatory neurons. AD is characterized by reduced GABA levels[33]. Consistent with this, we observed significant downregulation of two genes (*GABBR1* and *GABBR2*) that interact with GABA and are also involved in the cAMP signaling pathway. These genes were markedly downregulated in relation to the excitatory neuron-centric components of AD (Exc (x5): $p_{bonf} = 1x10^{-5}$, r = -0.39; Exc (x8): $p_{bonf} = 4x10^{-5}$, r = -0.38). Additionally, cortisol levels (Fig. 4c), known to reduce GABA levels and enhance neuronal excitability[34], were significantly elevated in association with these excitatory neuron components. Further compounding the hyperexcitability, hyperexcitability increases the metabolic demands of neurons, which may explain the elevated glucose levels observed in relation to the excitatory neuron-centric components (Fig. 4g,h, Exc (x5): $p_{bonf} = 2x10^{-3}$, r = 0.32; Exc (x8): $p_{bonf} = 3x10^{-3}$, r = 0.31).

## 6.2.9. TRP channels and arachidonoyl ethanolamide as promising targets for potential Alzheimer's therapies

We identified multiple disrupted immune-related pathways associated with four AD components (Fig. 4i, Exc (x5), Exc (x8), Inh (x18), and Ast-Oli (x4)). In oligodendrocytes, there was increased expression of genes involved in NF-kappa B signaling($P_{bonf} = 1.7x10^{-4}$). NF-kappa B activation has been suggested as a protective mechanism in oligodendrocytes against inflammation[35]. Additionally, Exc (x8) and Inh (x18) were linked to the regulation of inflammatory mediators of TRP channels($P_{bonf} \leq 3.7x10^{-3}$). TRP channels, including TRPV1 and TRPA1, are activated by inflammatory mediators and contribute to neuroinflammatory responses. The increased activity of these channels can lead to enhanced production of pro-inflammatory cytokines such as TNF-α, IL-1β, and IL-6, exacerbating the inflammatory environment in the AD brain[36]. Furthermore, both Exc(x8) and Inh(x18) were associated with decreased arachidonoyl

ethanolamide levels (Fig. 4c, $P_{bonf} \leq 1.5x10^{-3}$, $r \leq -0.29$). Lower arachidonoyl ethanolamide levels can result in increased expression of TRP channels, leading to greater calcium influx and amplified cellular responses to inflammation[37]. Activation of TRP channels also influences amyloid precursor protein (APP) processing and Aβ production. Specifically, TRPM7 activation has been shown to prevent AD-related Aβ neuropathology, which is pivotal in AD pathology[38]. Arachidonoyl ethanolamide exhibits significant anti-inflammatory and neuroprotective properties by modulating the endocannabinoid system and interacting with cannabinoid receptors (CB1 and CB2), which regulate pain, mood, and inflammation[39]. These findings suggest that TRP channels and anandamide could be promising targets for potential AD therapies.

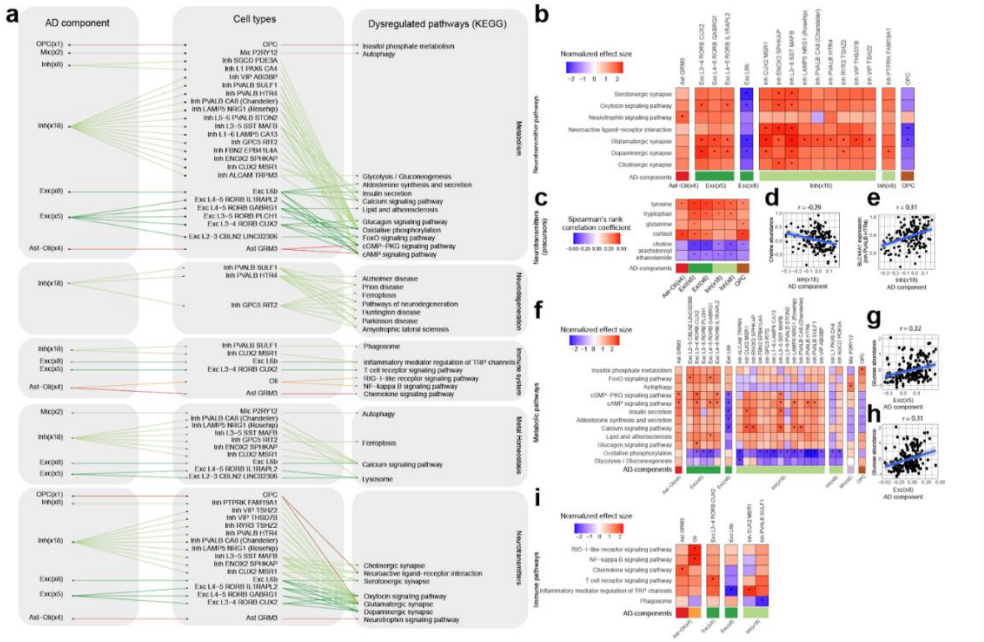



**Figure 4 a,** Significantly as pathways in specific sub-cell-types associated with the cell-type-inferred AD-scores. **b,** Significant dysregulated neurotransmitter-related pathways, colors indicate normalized effect size where red indicates that the genes of the respective pathways (rows) are significantly upregulated in the respective cell types (columns), associated with the respective cell-type-inferred AD-scores (bottom row). An * indicates significant gene set enrichment at p≤0.01 after Bonferroni correction.**c,** Association (spearman rank correlation) between neurotransmitters (precursors and regulator, rows) from the metabolite data and the cell-type-inferred AD-scores (columns). An * indicates significant association at p≤0.01 after Bonferroni correction. **d,** Association between the Inh(x8) inferred AD-scores (x-axis) and brain-choline levels (y-axis). **e,**Association between the Inh(x8) inferred AD-scores (x-axis) and choline transmembrane transporter (SLC44A1) in Inh PVALB HTR4 (y-axis). **f,** Significant dysregulated metabolic-related pathways, colors indicate normalized effect size where red indicates that the genes of the respective pathways (rows) are significantly upregulated in the respective cell types (columns), associated with the respective cell-type-inferred AD-scores (bottom row). An * indicates significant association at p≤0.01 after Bonferroni correction.**g,h,** Association between Exc(x5) **(g)** and Exc(x8) **(h)** inferred AD-scores(x-axes) and brain-glucose.**i,**Significant dysregulated immune-related pathways, colors indicate normalized effect size where red indicates that the genes of the respective pathways (rows) are significantly upregulated in the respective cell types (columns), associated with the respective cell-type-inferred AD-scores (bottom row). An * indicates significant association at p≤0.01 after Bonferroni correction.

## 6.3. Discussion

In this work, we introduced cell-projected phenotypes to address the need for analytical approaches that utilize the rich phenotypic data from recent large-scale single-cell profiling of disease and biobank cohorts, which encompass hundreds of individuals with diverse phenotypes. By computing cell-projected phenotypes, we were able to uncover heterogeneity in cellular phenotype manifestations both within and between individuals. This approach allowed us to redefine phenotypes at the level of individual cell types for each person, leading to a more nuanced understanding of the transcriptional disturbances and dysregulated pathways associated with these phenotypes.

In contrast to existing methods like MiloR[5] and CNA[6], which focus on phenotypic variation among individuals within the context of cellular neighborhoods, our approach goes beyond the neighborhood. These existing methods are limited to investigating aberrant gene expression within localized cellular environments or specific cells linked to a phenotype. However, we believe that the strength of the rich phenotypic data lies in its potential to be mapped back to the donor level. By mapping phenotypic manifestations back to the donor-level, we can explore additional phenotypic characteristics that may be associated with, for example, a more severe manifestation of AD in astrocytes. Moreover, in cohorts such as ROS/MAP[40], where other donor-level molecular measurements (e.g., proteomics, metabolomics) are available, mapping back to donor space enables the investigation of how other aberrant molecular processes correlate with variations in phenotypic manifestations across different cell types in individuals.

This work demonstrates that recognizing the variation in phenotypic manifestations within a cell type across individuals provides a deeper understanding of the transcriptional disturbances associated with these phenotypes. It challenges the existing paradigm, which typically constructs reference maps of discrete cell types and subtypes, followed by identifying differentially expressed genes between “cases” and “controls” within each cell group or subgroup. This traditional approach does not fully capitalize on the rich phenotypic data now available. Through our study using single-cell RNA-seq data, we have introduced a systematic approach that reveals: **(a)** not every cell from an individual exhibits the transcriptional signature of that individual's phenotypic traits; **(b)** phenotypes are expressed to varying degrees across different cell types; and **(c)** even among individuals with the same phenotypic traits, different cell types may be involved.

By moving beyond the existing paradigm, we identified a potential compensatory mechanism in AD, specifically neurotransmitter switching[25]. Our findings include elevated levels of glucose and cortisol, associated specifically with the excitatory neuron components of AD, which may be linked to neuronal hyperexcitability, as well as the identification of a potential neuroinflammation-regulating process in AD.

It is important to acknowledge that our phenotype scores are derived from transcriptional data. However, the extensive ROSMAP cohort[40], coupled with the availability of the Jager data[2] from the same brain region that included different

samples from overlapping individuals, enabled us to calculate cell type-specific phenotype scores from one sample and correlate these with gene expression-based measures from the other sample. Conducting this analysis within a single dataset could have introduced circular reasoning, but the unique composition of the ROSMAP cohort provided a robust framework for these analyses, effectively mitigating such concerns.

Altogether, this study not only advances our understanding of AD at a cellular level but also pioneers a new paradigm for analyzing scRNAseq data across conditions. By using population-scale datasets to investigate intra-individual heterogeneity and cross-individual similarities, our approach challenges conventional scRNAseq analysis techniques that typically ignore the complex spectrum of phenotypic expression across cell types. The insights from our research underscore the need for a more nuanced interpretation of scRNAseq data, which considers the heterogeneity of disease and cellular behavior within and across individuals. Moving forward, it is important that we refine these analytical methodologies to fully make use of the rich, multidimensional information offered by scRNAseq data. This could lead to more precise diagnostic tools and targeted therapies, tailored not only to diseases but also to the individual variabilities within patient populations. Our findings serve as a foundation for rethinking how single-cell genomics can be employed to dissect and understand the biology underlying human health and disease.

# 6.4. Methods

## Single cell RNAseq datasets

Two scRNAseq datasets were acquired from the Religious Orders Study and Rush Memory and Aging Project (ROSMAP)[40]. Both scRNAseq datasets were composed of cells originating from the dorsolateral prefrontal cortex. The first dataset[1] (our study) was acquired from the Synapse Portal (syn52293433). This dataset was already pre-processed. In short, gene counts were obtained by aligning reads to the GRCh38 genome using Cell Ranger. Doublets and poor quality cells were excluded, resulting in the dataset that we obtained that is composed of 2,359,994 cells from 427 individuals. Cell types were annotated using previously published marker genes and single-cell RNA-sequencing data. For a more detailed description of the pre-processing and cell type annotation we refer to the methods section of Hansruedi et al[1].

The second scRNAseq dataset[2] (Jager) was also acquired from the Synapse Portal (syn51123521). This dataset was also already pre-processed. In short, gene counts were obtained by aligning reads to the GRCh38 genome using Cell Ranger. Doublets and poor quality cells were excluded, resulting in the dataset that we obtained that is composed of 1,638,882 cells from 465 individuals. Cells were annotated into eight major cell types (excitatory neurons, inhibitory neurons, astrocytes, microglia, oligodendrocytes, OPCs, endothelial, and pericytes). For a more detailed description of the pre-processing and cell type annotation we refer to the methods section of Green et al[41].

## Metabolite dataset

The metabolite data[42] were acquired from ROSMAP through the Synapse Portal (syn25878459). The metabolite profiles were measured from the dorsolateral prefrontal cortex using the untargeted metabolomics platform from Metabolon Inc. For a more detailed description of the metabolite profiling pipeline we refer to the methods section of [42]. In total, the dataset consisted of 521 individuals and 969 metabolites. First we removed the individuals that were not measured in the Our scRNAseq dataset, leaving us with 160 individuals. Next, we removed metabolites that were measured in less than 80% of the individuals (N=32) and log normalized the data, leaving us with a dataset composed of 160 individuals and 592 metabolites. Remaining missing values were imputed with the R-package eimpute (v0.2.3)[43].

## Clinical and metadata

The clinical and metadata were provided to us by ROSMAP and were available for all individuals included in the two scRNAseq datasets and metabolite dataset. The clinical data that were used in this work were; NIA-Reagan diagnosis of AD (niareagansc), mean of percent area of cortex occupied by amyloid beta of eight brain regions (amyloid), mean tangle density of eight brain regions (tangles), mean neuritic plaque burden of five brain regions (plaq_n), global cognitive function (average of 19 cognitive tests, cogn_global_lv) and global pathology burden (gpath). For a more detailed description of the clinical data we refer to (https://www.radc.rush.edu/docs/var/variables.htm). Beside these clinical variables, we also calculated cognitive resilience to AD (CR), which is defined as unexplained variation in global cognitive function given the global pathological burden of individuals. We calculated this by fitting a linear regression model with global cognitive function as outcome variable (y) and global pathology burden as predictor variable (x) and taking the residuals. Global cognitive function was inverted by multiplying it by -1 and referred to as cognitive impairment. And sex was also inverted where 1 represents female and 0 represents male. Metadata that were used in this work were, age of death, post-mortem interval (pmi), sex (msex), fixation interval (fixation_interval) and which study the individuals were part of; ROS or MAP (study).

## Harmonizing cell type annotations between single cell RNAseq datasets

Our scRNAseq was annotated at a higher resolution (54 sub cell types) than the Jager scRNAseq dataset (eight cell types). As such, we used our dataset as reference to annotate the Jager scRNAseq dataset at the same resolution. First, per major cell type, as annotated in the Mathys dataset (excitatory neurons, inhibitory neurons, vasculature cells, OPCs, oligodendrocytes, astrocytes and immune cell), we constructed reference datasets, such that each reference set contained 1,000 randomly selected cells of each sub cell type. Then, we matched the major cell types from both datasets. Next, per major cell type, in chunks of 20 individuals, we used the FindTransferAnchors, TransferData and AddMetaData functions from Seurat (v4.1.0)[44] to perform label transfer from the reference sets to the cells from the Jager scRNAseq dataset. Quality of the label transfer was

assessed by comparing marker genes from our scRNAseq dataset and marker genes from the newly identified Jager scRNAseq dataset.

## Generating pseudo bulk datasets

Using the Jager scRNAseq dataset, for each sub cell type we generated pseudo bulk gene expression datasets. We did this by binarizing the gene expression values of the single cells, such that every zero remains a zero and every value ≥1 is assigned a 1. Previous work of ours[45] showed that for scRNAseq datasets with large numbers of cells and many individuals, calculating binary-based pseudo bulk performs better than count-based pseudo bulk, resulting in less false positives. Per individual, per gene, we calculated the proportion of measurements (1s), to get the relative expressions of the genes of that individual within the sub cell type. An individual was excluded for a specific sub cell type if that individual had less than 10 cells of the respective sub cell type. And individuals were excluded when too few genes were measured in the respective sub cell type relative to the other individuals. The lower-bound threshold for this was defined as the median of measured genes for all individual minus the IQR*2. For each pseudo bulk dataset median ratio normalization[46] was applied, followed by batch correction using combat from the R-package sva (v3.44.0)[47].

## Cell phenotypic projections

In order to calculate cell projected phenotype a normalized scRNAseq dataset and meta data containing phenotypic information about the individuals within the scRNAseq dataset are required. First, miloR[5] is used to create a k-nearest neighbors (KNN) graph, where cells are connected based on transcriptomic similarities (euclidean distance). Using this KNN graph, overlapping neighborhoods are identified, such that a cell can belong to multiple neighborhoods. This results in neighborhood index matrix M with dimensions c×n, where c is the number of cells and n is the number of neighborhoods, where a 1 indicates membership of a cell to a neighborhood. For each neighborhood, the cell count per individual is determined, resulting in a neighborhood-by-individual count matrix. Then, association between neighborhood and a user defined phenotype is tested using NB GLM from edgeR[48]. During this step TMM normalization is used to account for different numbers of cells across individuals and covariates can be added to account for confounding factors. The resulting log fold-changes, representing neighborhood scores, are stored in vector S. Using the neighborhood index matrix M and the vector of neighborhood scores S we perform a column wise multiplication which replaces the binary indicators with the neighborhood scores, resulting in matrix N.

Next, for each individual cell we calculate the euclidean distance to each of the neighborhoods it belongs to and apply a fading membership function to the distances, also known as a softmax function, to provide the ability to assign larger weights to closer neighborhoods. The neighborhood scores in matrix N are multiplied by the calculated weights. As the weights sum to one, the cell score is calculated by taking the sum of the weighted matrix N. In the context of AD, a cell receives a high score when surrounded in the KNN graph by cells from

individuals diagnosed with AD, indicating transcriptional similarity. Conversely, a low score occurs when surrounded by cells from healthy individuals.

## Pre-processing scRNAseq datasets for cell projected phenotypes

We calculated cell projected phenotypes for each of the major cell types separately. First, individuals with too few cells were removed <Q1 (number of cells of all individuals). And individuals with too many cells (>median + IQR *2) were down-sampled as this could possibly confound the testing of differential abundance per neighborhood. Next, using Seurat, we log normalized the count matrix, such that yij =log((xij/jxij))104 , where xij and yij are the raw and normalized values for every gene i in every cell j, respectively. Next, we identified 2,000 highly variable features sew Seurat's *vst* method, scaled the count matrix and used the 2,000 highly variable genes to run principal component analysis (PCA). Using the PCs of the cells we calculated the lisi score (R-package v1.0)[49] to filter out cells that were on average transcriptionally similar to cells coming from <3 individuals. As such, ensuring that neighborhoods are composed of at least 3 individuals.

## Calculating cell projected phenotypes

With the pre-processed scRNAseq datasets we used miloR[5] (v1.3.6) to calculate the KNN graph (k = 30) per major cell type based on the first twelve PCs and to identify the neighborhoods of transcriptionally similar cells, such that the number of neighborhoods is 10% of the total number of cells. A relatively high k and low number of neighborhoods are required to ensure that neighborhoods are sufficiently overlapping. Next, per neighborhood, per individual the number of cells are counted. Next, the cell counts of these neighborhoods were tested using the NB GLM from edgeR(v3.37.4)[48] on associations as described in **Supplementary Table 1**. The resulting log fold changes of the neighborhoods were weighted with the fading membership function (fading factor = 0.3) and propagated to the individual cells as described in the previous section (see methods: Cell phenotypic projections).

## Calculating cell type specific phenotypes and proportions of AD-like cells

Using the cell specific phenotype associations we calculated cell type specific phenotypes. We did this by taking the mean phenotype score per individual, per sub cell type. Sub cell types were excluded when less than 100 individuals had cell specific phenotype scores for that respective sub cell type. Individuals were excluded when the respective individual had cell specific phenotype scores for less than 20 sub cell types. Remaining missing values were imputed with the R-package eimpute (v0.2.3)[43].

## Clustering cell type specific AD scores to get AD components

To identify the distinct AD components we clustered the sub cell type inferred AD scores. To identify sub cell types with similar scores across individuals we first regressed out the first PC to remove the variation shared between all sub cell types (mainly AD diagnosis). Next, we performed hierarchical clustering with complete linkage, where the distance was defined as 1 – Spearman's correlation coefficient between sub cell types. We cut the tree at h = 0.9, resulting in eight clusters. After visually inspecting the dendrogram we separated the OPCs from the microglia (P2RY12 and TPT1) clusters, due to the relatively large distance of the OPCs to the microglia. To get the AD components, we re-calculated the AD scores, but now within the clusters of sub cell types.

## Predicting cell projected phenotypes of new cells

To predict cell projected phenotypes of new unseen cells, first, using the dataset for which we have cell projected phenotypes we obtain the 200 most correlated genes with the projected phenotypes. Using these 200 genes we calculate the euclidean distance between the new cells and the existing neighborhoods of matching cell types. For each cell, using the five closest neighborhoods we apply the fading membership function (fading factor = 0.1), and sum the weighted association scores (of any phenotype) of the respective neighborhoods that become the predicted cell projected phenotypes.

## Predicting donor-level phenotypes

First, we utilized our dataset to train a ridge regression model for each phenotype, with the cell type-inferred phenotype scores serving as predictors and the reported donor-level phenotype as the outcome variable, using the R package glmnet (v4.1.8)[50]. For binary phenotypes, we used a logistic ridge regression. Cell types were selected based on a Spearman correlation of ≥ 0.5 between the observed and predicted cell type-inferred phenotype scores using the individuals that we measured in both our dataset and the Jager dataset. The lambda parameter was optimized through 10-fold cross-validation, and the optimal lambda was subsequently applied in the final training of the ridge regression model. The trained models were then used to predict donor-level phenotypes of previously unseen individuals. The association between predicted and reported continuous phenotypes was assessed using Spearman's rank correlation, while the association of binary phenotypes was evaluated using a t-test, comparing the reported donor-level phenotypes with the outcome probabilities from the logistic ridge regression.

## Differential expression analysis

Differential expression analyses were performed between the nine AD components and genes expressed in the sub cell types that compose the respective AD component. The differential expression analysis we performed using a simple linear regression with the AD component as predictor variable and the pseudo-bulk genes as dependable variable, while correcting for post mortem

interval, study (ROS or MAP),sex and fixation interval. Per sub cell type Bonferroni correction was used to correct for multiple testing and significance was assumed at $P_{bonf} \leq 0.01$.

## Gene set enrichment analysis

Gene set enrichment analyses were done using the fgsea R-package (v1.24.0)[51]. For each sub-cell-type the genes were sorted on their t value, representative of the association with the AD components calculated during the differential expression analysis. Enrichment was tested for KEGG pathways[52] with at least 5 genes at most 500 genes.The database was downloaded from the web server (https://maayanlab.cloud/Enrichr/#libraries) of EnrichR[53].

# References


1. Mathys, H. *et al.* Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to Alzheimer's disease pathology. *Cell* **186**, 4365–4385.e27 (2023).

2. Fujita, M. *et al.* Cell subtype-specific effects of genetic variation in the Alzheimer's disease brain. *Nat. Genet.* **56**, 605–614 (2024).

3. Crowell, H. L. *et al.* muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat. Commun.* **11**, 1–12 (2020).

4. Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**, (2022).

5. Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat. Biotechnol.* 1–9 (2021).

6. Reshef, Y. A. *et al.* Co-varying neighborhood analysis identifies cell populations associated with phenotypes of interest from single-cell transcriptomics. *Nat. Biotechnol.* **40**, 355–363 (2022).

7. Thies, W. & Bleiler, L. 2012 Alzheimer's disease facts and figures. *Alzheimer's and Dementia* **8**, 131–168 (2012).

8. Hroudová, J., Singh, N., Fišar, Z. & Ghosh, K. K. Progress in drug development for Alzheimer's disease: An overview in relation to mitochondrial energy metabolism. *Eur. J. Med. Chem.* **121**, 774–784 (2016).

9. Kim, C. K. *et al.* Alzheimer's disease: Key insights from two decades of clinical trial failures. *J. Alzheimers. Dis.* **87**, 83–100 (2022).

10. Braak, H. & Braak, E. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol.* **82**, 239–259 (1991).

11. Brelstaff, J. H. *et al.* Microglia become hypofunctional and release metalloproteases and tau seeds when phagocytosing live neurons with P301S tau aggregates. *Sci Adv* **7**, eabg4980 (2021).

12. Leng, F. & Edison, P. Neuroinflammation and microglial activation in Alzheimer disease: where do we go from here? *Nat. Rev. Neurol.* **17**, 157–172 (2021).

13. Shimizu, T. *et al.* Oligodendrocyte dynamics dictate cognitive performance outcomes of working memory training in mice. *Nat. Commun.* **14**, 6499 (2023).

14. Santello, M., Toni, N. & Volterra, A. Astrocyte function from information processing to cognition and cognitive impairment. *Nat. Neurosci.* **22**, 154–166 (2019).

15. Marin-Husstege, M., Muggironi, M., Raban, D., Skoff, R. P. & Casaccia-Bonnefil, P. Oligodendrocyte progenitor proliferation and maturation is differentially regulated by male and female sex steroid hormones. *Dev. Neurosci.* **26**, 245–254 (2004).

16. Yasuda, K. *et al.* Sex-specific differences in transcriptomic profiles and cellular characteristics of oligodendrocyte precursor cells. *Stem Cell Res.* **46**, 101866 (2020).

17. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167–202 (2001).

18. Koechlin, E., Ody, C. & Kouneiher, F. The architecture of cognitive control in the human prefrontal cortex. *Science* **302**, 1181–1185 (2003).

19. Duncan, J. An adaptive coding model of neural function in prefrontal cortex. *Nat. Rev. Neurosci.* **2**, 820–829 (2018).

20. Bossers, K. *et al.* Concerted changes in transcripts in the prefrontal cortex precede neuropathology in Alzheimer's disease. *Brain* **133**, 3699–3723 (2010).

21. Balmik, A. A. & Chinnathambi, S. Methylation as a key regulator of Tau aggregation and neuronal health in Alzheimer's disease. *Cell Commun. Signal.* **19**, 1–13 (2021).

22. Slotkin, T. A., Nemeroff, C. B., Bissette, G. & Seidler, F. J. Overexpression of the high affinity choline transporter in cortical regions affected by Alzheimer's disease. Evidence from rapid autopsy studies. *J. Clin. Invest.* **94**, 696–702 (8 1994).

23. Spitzer, N. C. Neurotransmitter Switching? No Surprise. *Neuron* **86**, 1131–1144 (2015).

24. Li, H.-Q., Pratelli, M., Godavarthi, S., Zambetti, S. & Spitzer, N. C. Decoding Neurotransmitter Switching: The Road Forward. *J. Neurosci.* **40**, 4078–4089 (2020).

25. Steinkellner, T. *et al.* Dopamine neurons exhibit emergent glutamatergic identity in Parkinson's disease. *Brain* **145**, 879–886 (2022).

26. Wallace, T. L. & Bertrand, D. Importance of the nicotinic acetylcholine receptor system in the prefrontal cortex. *Biochem. Pharmacol.* **85**, 1713–1720 (2013).

27. Baker-Nigh, A. *et al.* Neuronal amyloid-β accumulation within cholinergic basal forebrain in ageing and Alzheimer's disease. *Brain* **138**, 1722–1737 (2015).

28. Wurtman, R. J. Choline metabolism as a basis for the selective vulnerability of cholinergic neurons. *Trends Neurosci.* **15**, 117–122 (1992).

29. Babaei, P. NMDA and AMPA receptors dysregulation in Alzheimer's disease. *Eur. J. Pharmacol.* **908**, 174310 (2021).

30. Whitehouse, P. J. Neurotransmitter receptor alterations in Alzheimer disease: a review. *Alzheimer Dis. Assoc. Disord.* **1**, 9–18 (1987).

31. Lombardero, L., Llorente-Ovejero, A., Manuel, I. & Rodríguez-Puertas, R. Chapter 28 - Neurotransmitter receptors in Alzheimer's disease: from glutamatergic to cholinergic receptors. in *Genetics, Neurology, Behavior, and Diet in Dementia* (eds. Martin, C. R. & Preedy, V. R.) 441–456 (Academic Press, 2020).

32. Teleanu, R. I. *et al.* Neurotransmitters-Key Factors in Neurological and Neurodegenerative Disorders of the Central Nervous System. *Int. J. Mol. Sci.* **23**, (2022).

33. Carello-Collar, G. *et al.* The GABAergic system in Alzheimer's disease: a systematic review with meta-analysis. *Mol. Psychiatry* **28**, 5025–5036 (2023).

34. Milani, P. *et al.* Cortisol-induced effects on human cortical excitability. *Brain Stimul.* **3**, 131–139 (2010).

35. Stone, S. *et al.* NF-κB Activation Protects Oligodendrocytes against Inflammation. *J. Neurosci.* **37**, 9332–9344 (2017).

36. Lee, K.-I. *et al.* Role of transient receptor potential ankyrin 1 channels in Alzheimer's disease. *J. Neuroinflammation* **13**, 92 (2016).

37. Hu, H.-Z. *et al.* Potentiation of TRPV3 channel function by unsaturated fatty acids. *J. Cell. Physiol.* **208**, 201–212 (2006).

38. Oh, H. G. & Chung, S. Activation of transient receptor potential melastatin 7 (TRPM7) channel increases basal autophagy and reduces amyloid β-peptide. *Biochem. Biophys. Res. Commun.* **493**, 494–499 (2017).

39. Turcotte, C., Chouinard, F., Lefebvre, J. S. & Flamand, N. Regulation of inflammation by cannabinoids, the endocannabinoids 2-arachidonoyl-glycerol and arachidonoyl-ethanolamide, and their metabolites. *J. Leukoc. Biol.* **97**, 1049–1070 (2015).

40. Bennett, D. A. *et al.* Religious Orders Study and Rush Memory and Aging Project. *J. Alzheimers. Dis.* **64**, S161–S189 (2018).

41. Green, G. S. *et al.* Cellular dynamics across aged human brains uncover a multicellular cascade leading to Alzheimer's disease. *bioRxiv* (2023) doi:10.1101/2023.03.07.531493.

42. Batra, R. *et al.* The landscape of metabolic brain alterations in Alzheimer's disease. *Alzheimers. Dement.* (2022) doi:10.1002/alz.12714.

43. Mazumder, R., Hastie, T. & Tibshirani, R. Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *J. Mach. Learn. Res.* **11**, 2287–2322 (2010).

44. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**, 293–304 (2024).

45. Bouland, G. A., Mahfouz, A. & Reinders, M. J. T. Consequences and opportunities arising due to sparser single-cell RNA-seq datasets. *Genome Biol.* **24**, 86 (2023).

46. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

47. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).

48. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

49. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

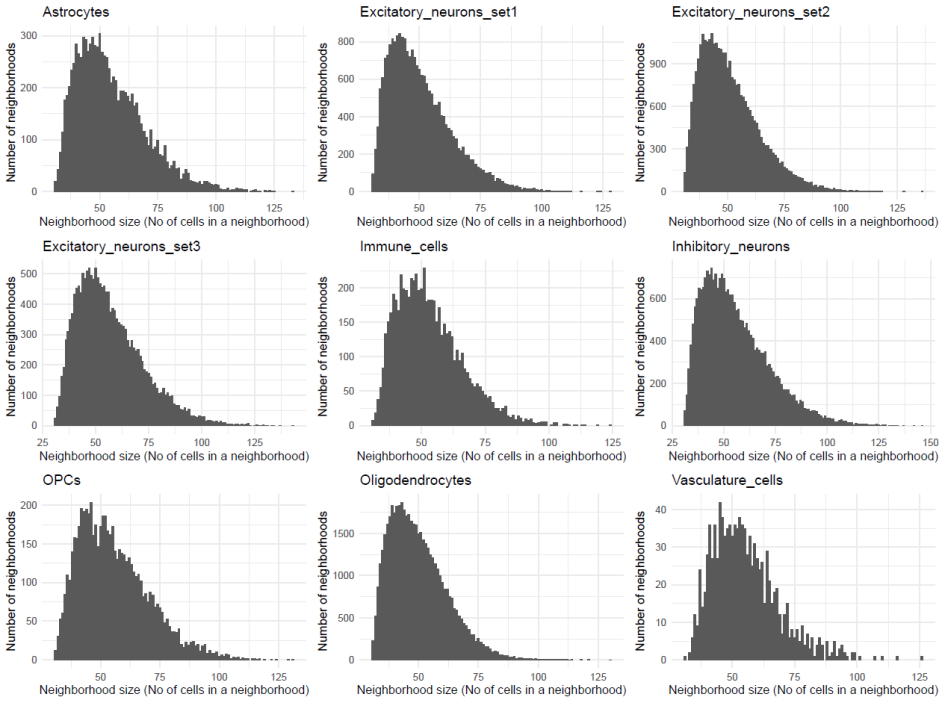
50. Tay, J. K., Narasimhan, B. & Hastie, T. Elastic Net Regularization Paths for All Generalized Linear Models. *J. Stat. Softw.* **106**, (2023).

51. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv* 060012 (2021).

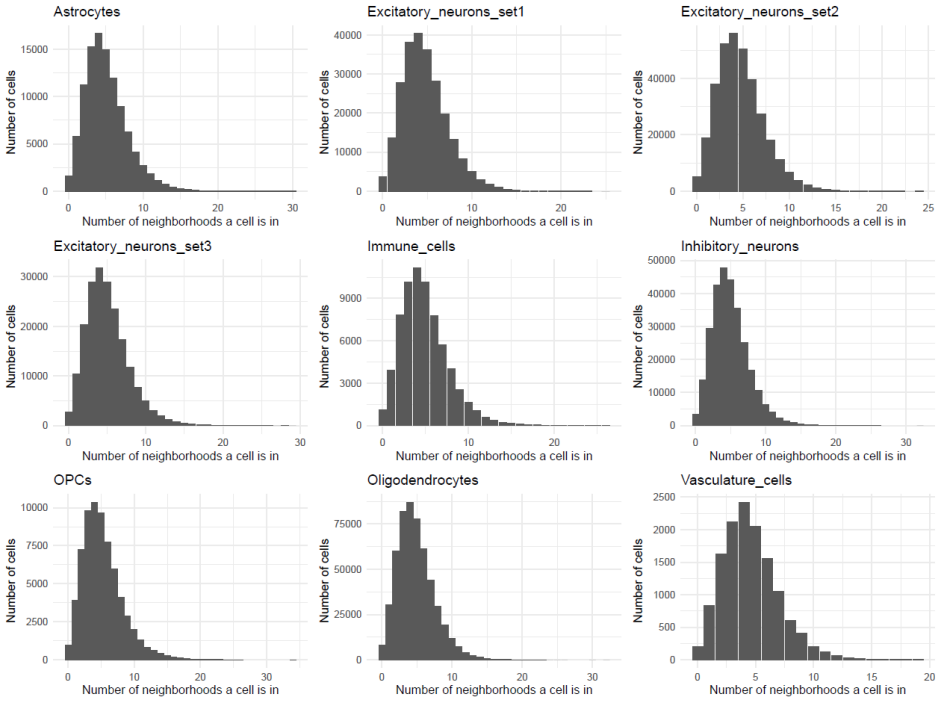
52. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–D592 (2023).

53. Xie, Z. *et al.* Gene Set Knowledge Discovery with Enrichr. *Current Protocols* **1**, e90 (2021).

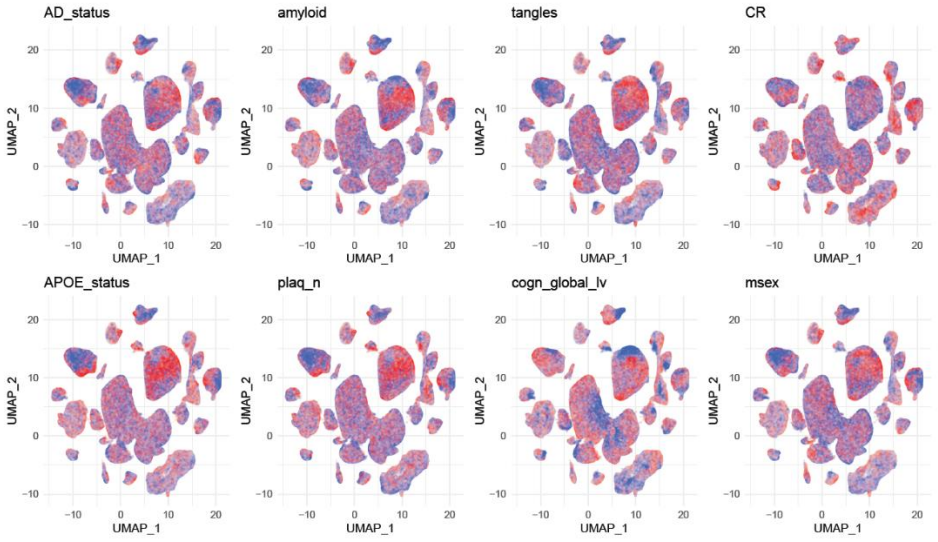
# Supplements




**Supplementary Figure 1:** Distribution of neighborhood sizes, shown per major cell type (panels) with neighborhood size (x-axis) and the number of neighborhoods with the respective neighborhood size (y-axis). Excitatory neurons were run in three sets due to the number of excitatory neurons cells profiled.

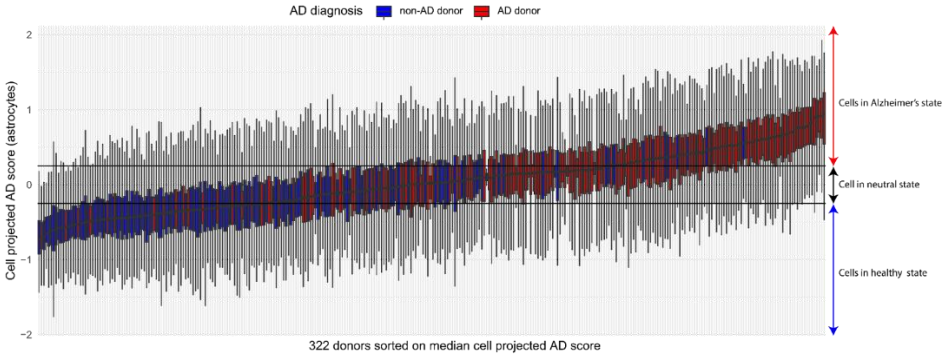



**Supplementary Figure 2:** Distribution of the number of neighborhoods a cell belongs to. shown per major cell type (panels) with the number of neighborhoods per cell (x-axis) and the number of cells with that many neighborhoods (y-axis). Excitatory neurons were run in three sets due to the number of excitatory neurons cells profiled.

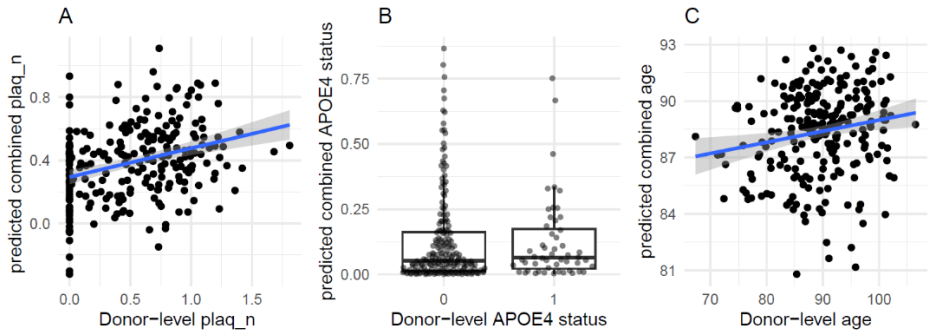



**Supplementary Figure 3,** Uniform manifold approximation and projection (UMAP) of the main scRNAseq datasets used for calculating the cell-transcriptional neighborhoods coloured by cell-

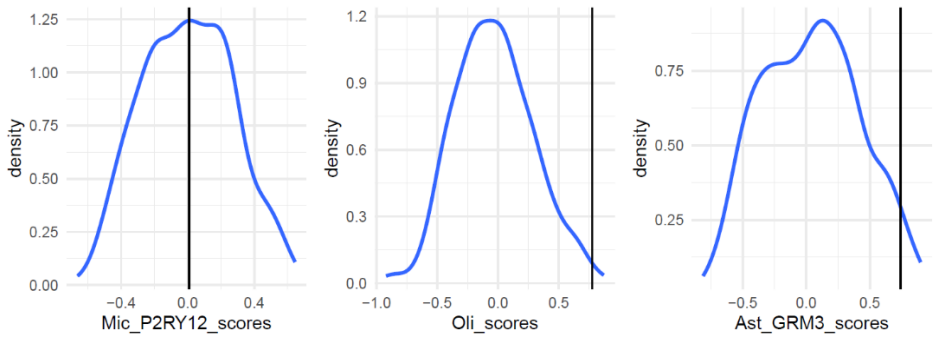
projected phenotype-scores where blue represents a low phenotype scores and red represents high phenotype scores, capped at phenotypes-scores of -1 and 1.



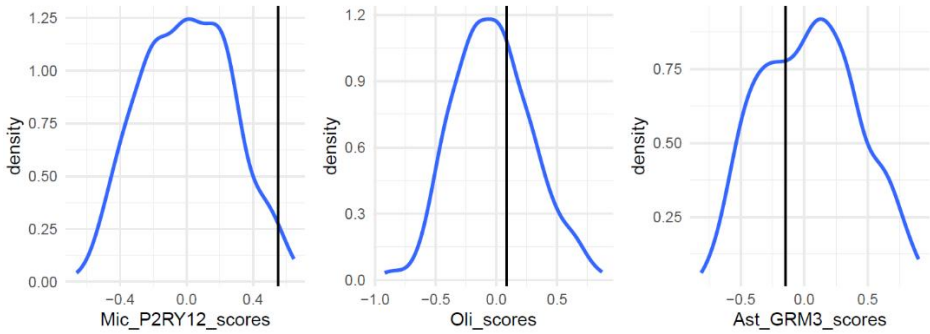

**Supplementary Figure 4,** Boxplots of 322 donors (x-axis) and their cell projected AD scores for astrocytes, showing which individuals primarily have cells in alzheimer state ≥0.25 and which in a healthy state≤-0.25. Bleu boxplots are donors that reported to be non-AD individuals and red boxplots are donors reported to be diagnosed with AD.



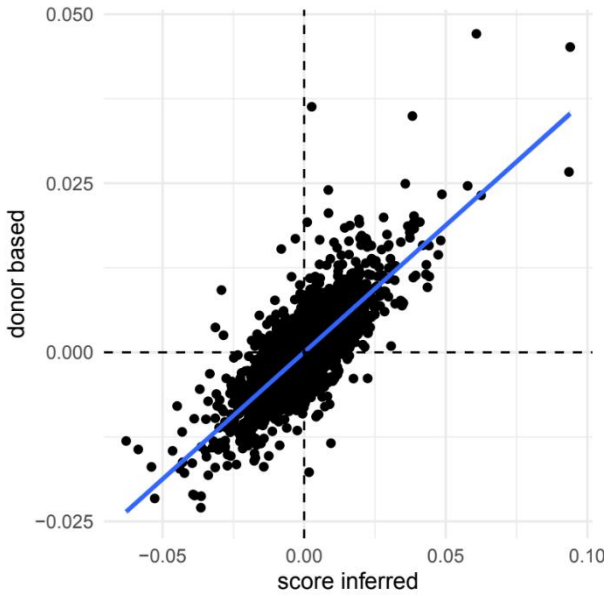

**Supplementary Figure 5:** Associations between reported donor-level phenotypes (x-axis) of 245 previously unseen donors with predicted donor-level phenotypes (y-axis), using the predicted cell type inferred phenotype scores of neuritic plaque**(a)**, APOE4 status **(b)** and age **(c).**



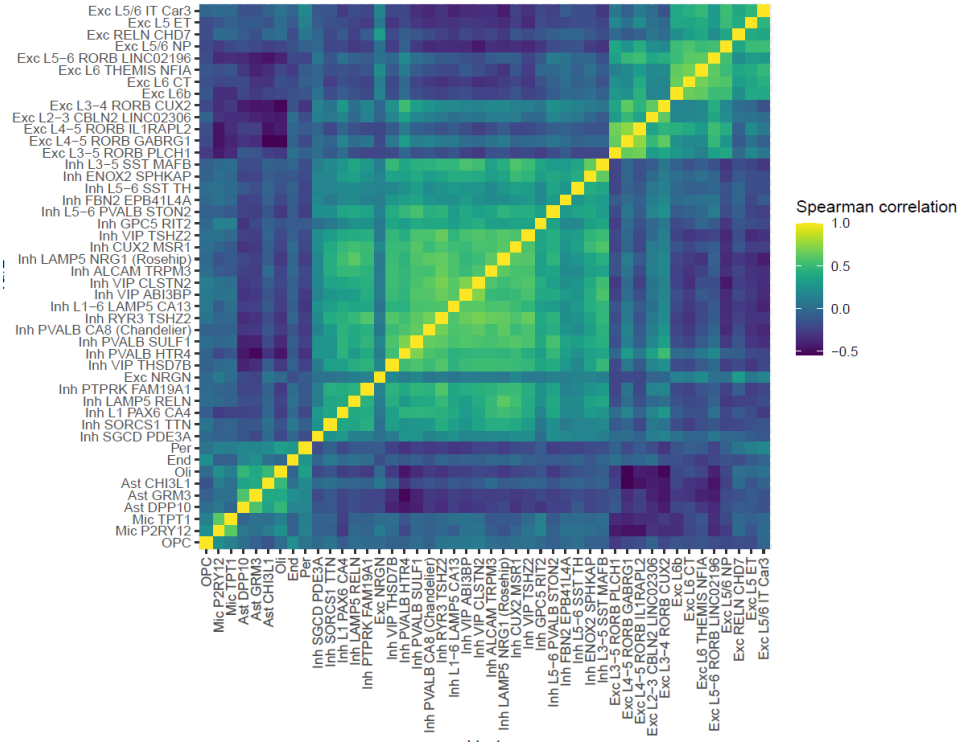
**Supplementary Figure 6** Distribution of Microglia P2RY12, Oligodendrocytes and Astrocytes GRM3 inferred AD scores, and the position of individual #20156469 in that distribution, indicated with a vertical line.



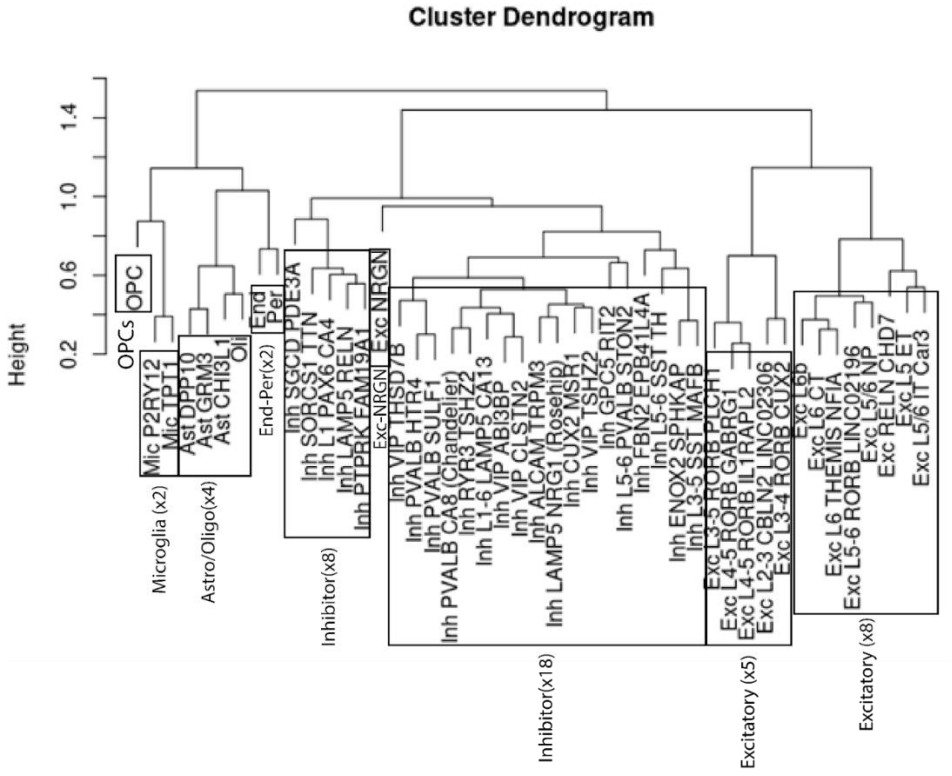

**Supplementary Figure 7:** Distribution of Microglia P2RY12, Oligodendrocytes and Astrocytes GRM3 inferred AD scores, and the position of individual #50106730 in that distribution, indicated with a vertical line.



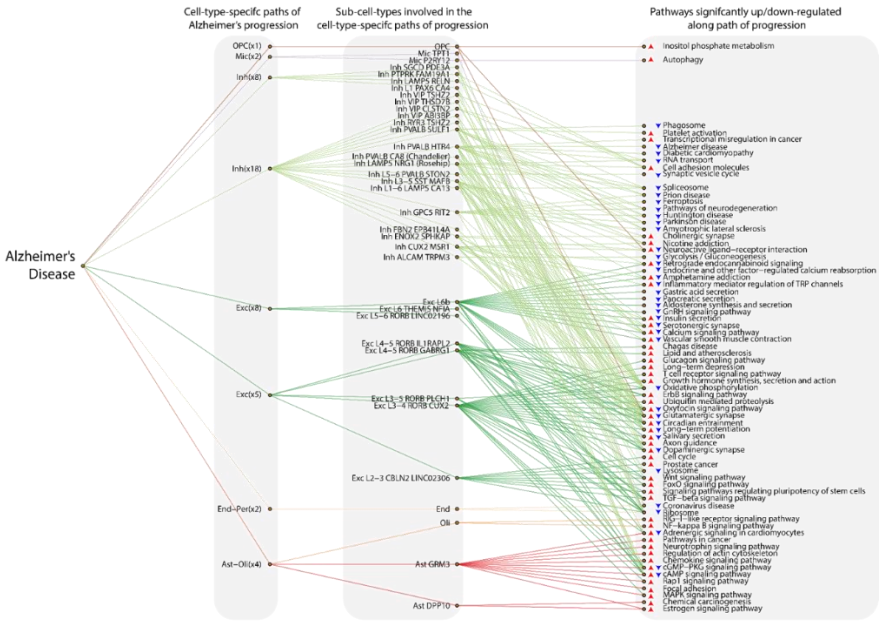

**Supplementary Figure 8:** Association between effect sizes of differential expression analysis performed with oligodendrocyte inferred tangle density score (x-axis) and differential expression analysis performed with reported tangle density(y-axis).




**Supplementary Figure 9:** Correlation (spearman rank correlation coefficient) between 52 cell-type-inferred AD scores(row and columns).




**Supplementary Figure 10:** Dendrogram (complete linkage; default) based on 1-spearman rank correlation coefficient between cell type inferred AD scores.




**Supplementary Figure 11:** Significantly dysregulated pathways in specific sub-cell-types associated with the cell-type-inferred AD-scores.

# Differential analysis of binarized single-cell RNA sequencing data captures biological variation

Gerard Bouland, Ahmed Mahfouz and Marcel Reinders

# Abstract

Single-cell RNA sequencing data is characterized by a large number of zero counts, yet there is growing evidence that these zeros reflect biological variation rather than technical artifacts. We propose to use binarized expression profiles to identify the effects of biological variation in single-cell RNA sequencing data. Using 16 publicly available and simulated datasets, we show that a binarized representation of single-cell expression data accurately represents biological variation and reveals the relative abundance of transcripts more robustly than counts.


# 7.1. Introduction

Single cell RNA sequencing (scRNAseq) data is highly sparse, and the common belief is that the zero values are primarily caused by technical artifacts (often referred to as dropouts). Although more zeros are observed in scRNAseq data than expected, these can largely be explained by biological rather than technical factors(1). Also, the amount of zeros in scRNAseq is in line with distributional models of molecule sampling counts(2, 3). These distributional models show that a zero observation is not simply a missing-value, as a missing-value would provide no information. On the contrary, a zero observation for a gene reveals that the respective gene is unlikely to be highly expressed(3). Methods that utilize zero observations for feature selection(4),(5) and cell type clustering(6) have recently been developed and perform better or comparable with methods relying on the continuous expression values of highly variable genes. For instance, Qiu(6) binarized scRNAseq count data, where each zero remains zero and every non-zero value was assigned a one. With this binary representation, cell type clusters were identified based on co-occurrence of transcripts. Yet, it is not clear whether differences in the number of zeros for a gene also reflect differences across distinct biological cell populations. Therefore, we investigated whether biological differences across cell population can be identified using Binary Differential Analysis (BDA), rather than the commonly used differential expression analysis (DEA). Instead of relying on changes in the expression value of genes across cell populations, which can be sparse and are subject to pre-processing steps, we analyzed the binary expression patterns across biological distinct cell populations, i.e. are there more (or less) zeros for a gene in condition *A* compared to condition *B*. Taken together the main contribution of our work is that we show that the binarization of gene expression is biologically relevant and can be used to test for differences between a wide variety of groupings, and that this holds across different datasets, as well as different single-cell protocols.

# 7.2. Results

## 7.2.1. BDA competitive with Wilcoxon Rank Sum test

As proof of concept, we performed BDA with a simple logistic regression on binarized expression profiles from 16 scRNAseq datasets (662,825 cells in total, Table 1). We compared the results of BDA-LR with those of differently expressed genes (DEGs) detected using the commonly used Wilcoxon Rank Sum test,

which is also top ranked for single cell analyses(9),(24). We tested each gene using both BDA-LR and DEA for differences between conditions (6 datasets), cell types (6 datasets) and normal- versus cancerous tissues (4 datasets, Fig. 1a). Across all datasets, a total of 96,275 significant genes ($P_{FDR} \leq 0.05$) were identified with either BDA-LR (92,381 genes) or DEA (91,521 genes). Of these, 87,627 were identified by both tests, resulting in a Jaccard index of 0.91. This high degree of agreement is also reflected in each individual dataset (median = 0.92, minimum = 0.76, and maximum = 0.99). We did not use a log fold-change (logFC) or log odds-ratio (logOR) threshold, as for each dataset and comparison different thresholds are appropriate. In all datasets, the logFC and logOR were significantly (spearman) correlated (median($\rho$) = 0.90, minimum($\rho$) = 0.49, and maximum($\rho$) = 0.98, $P \leq 5 \times 10^{-100}$). The three datasets with the lowest correlation coefficient between logOR and logFC ($\rho \leq 0.62$) were datasets generated using the Smart-seq protocol (Table 1). Across the datasets, we observed an average increase of 1.80 in logOR (median = 1.70, $Q_1$ = 1.59, $Q_3$ = 2.10), for every increase in logFC (see Fig. 1b for the *cancer atlas (2)* dataset(8)). The high degree of agreement of detected genes shows that BDA-LR performs on par with the Wilcoxon Rank Sum test and the strong correlation of the logFC and logOR across all datasets shows that the results can be interpreted in a similar way.

**Table 1** Single cell datasets included in this study

| Dataset | No. Unique individuals | No. Cells | No. Genes | Contrasting subpopulation defined by: | Description | Protocol | Reference |
|---|---|---|---|---|---|---|---|
| **Alzeheimer's Disease (AD)** | 14 | 13.214 | 10.850 | Control vs AD | Entorhinal cortex | 10x Chromium | (21) |
| **Major Depressive Disorder (MDD)** | 34 | 78.886 | 30.062 | Control vs MDD | Prefrontal cortex | 10x Chromium | (31) |
| **Type 2 Diabetes (T2D)** | 10 | 3.514 | 26.271 | Control vs T2D | Pancreas | Smart-seq2 | (32) |
| **Coronavirus Disease 2019 (COVID19)** | 13 | 44.721 | 26.361 | Control vs COVID19 | PBMCs | Seq-Well | (33) |
| **Lung adenocarcinoma (LUAD, Lung)** | 22 | 88.144 | 29.634 | Normal tissue vs cancerous tissue | Lung | 10x Chromium | (34) |
| **Lung adenocarcinoma (LUAD, Lymph node)** | 17 | 54.577 | 29.634 | Normal tissue vs cancerous tissue | Lymph node | 10x Chromium | (34) |
| **Four cancers (T-cells)** | 14 | 132.549 | 22.815 | Normal tissue vs cancerous tissue | Colon, Endo, Lung, Renal | 10x Chromium | (35) |
| **Aging Mouse Atlas FACS** | 14 | 74.157 | 22.966 | 3m vs 24m | Aging mouse | Smart-seq2 | (7) |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Aging Mouse Atlas Droplet** | 11 | 83.262 | 20.138 | 3m vs 24m | Aging mouse | 10x Chromium | (7) |
| **Allen Brain Atlas (Medialis temporalis Gyrus)** | 8 | 14.689 | 48.304 | Inhibitory neurons vs Excitatory Neurons | Medialis temporalis Gyrus | Smart-Seq v4 | (36) |
| **Colorectal Cancer** | 23 | 63.502 | 27.946 | Normal tissue vs cancerous tissue | Colon CRC cells | 10x Chromium | (37) |
| **Cortex Neurons** | 5 | 9.451 | 28.985 | Inhibitory neurons vs Excitatory Neurons | Cortex | 10x Chromium | (25) |
| **Cortex Oligodendrocytes** | 5 | 307 | 28.985 | OPC vs ODC | Cortex | 10x chromium | (25) |
| **Substantia Nigra** | 7 | 4.711 | 28.985 | OPC vs ODC | Substantia Nigra | 10x chromium | (25) |
| **Cancer atlas (1)** | 171 | 33.346 | 50.705 | Regulatory T cells vs T helper cells | Cancer atlas | Multiple | (8) |
| **Cancer atlas (2)** | 162 | 30.105 | 48.942 | Naive-memory CD4 T cells vs Transitional memory CD4 T cells | Cancer atlas | Multiple | (8) |

## 7.2.2. BDA among the best performing tests on simulated data

To compare the performance of binary methods with methods relying on counts in a controlled manner, we simulated scRNAseq data with muscat(10) using the provided dataset as reference(11). We generated scRNAseq data with varying number of cells and 25% of differentially expressed genes. With 1,000 and 2,000 simulated cells, DEsingle(14) performed the best as the F1-score (Fig. 1c) and positive predictive value (PPV, S Fig. 1) were the highest and the false positive rate (FPR, S Fig. 2) was the lowest. However, this performance comes at a cost in terms of considerably required computational time (S Fig. 3a). For that reason, we excluded DEsingle when simulating 5,000 and 10,000 cells (running time >20min). All binary-based methods performed consistently good, with 1,000 and 2,000 cells ranking tightly together. BDA-fisher and BDA-Phi had decreased relative performance with 5,000 and 10,000 cells, while the performances of BDA-chisq and BDA-LR were also among the best with 5,000 and 10,000 cells. Taken together, this shows that differences in the frequency of zeros between groups can represent biological variation and can most accurately be detected with BDA-chisq and BDA-LR in a time efficient manner.

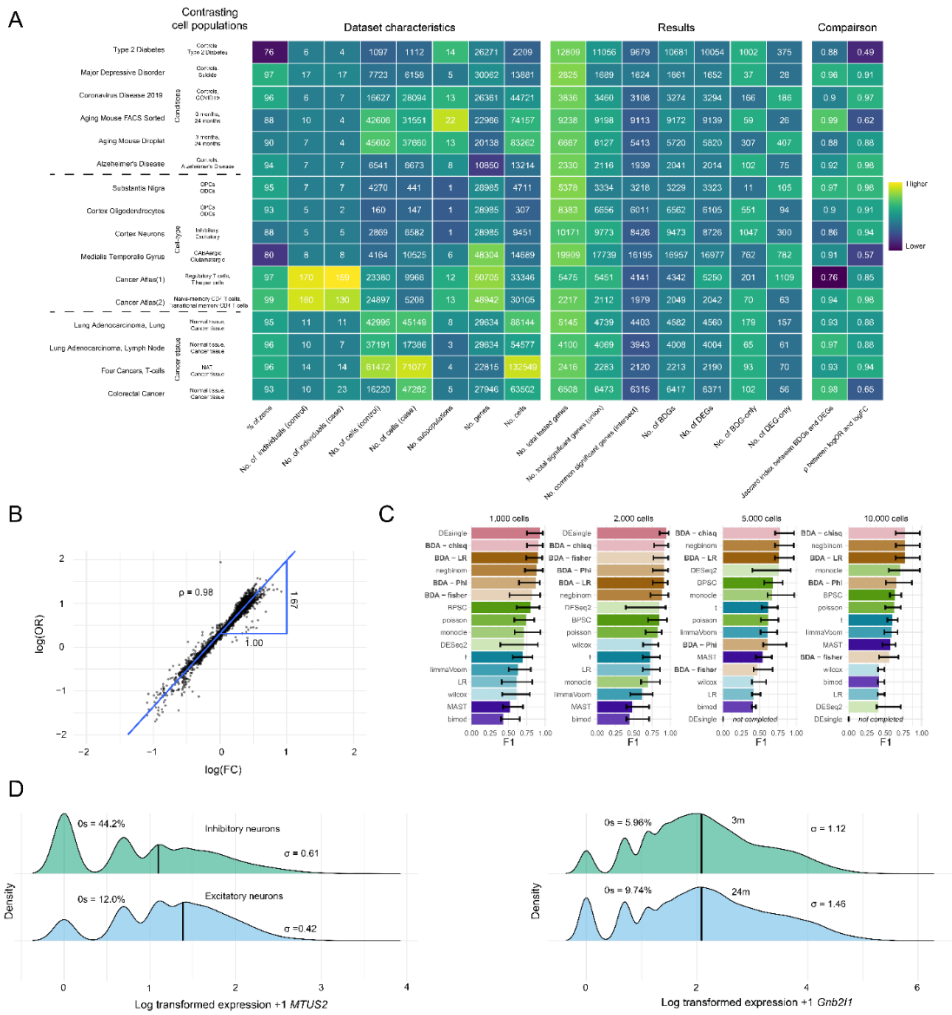



Figure 1 **a,** Heatmap of the dataset characteristics, general overview of the results of both BDA-LR and DEA per dataset, and a comparison of the results. The rows represent the datasets, the first column shows the cell populations that were used as contrast for testing. **b,** Plot of the logOR and logFC of the Cancer Atlas (2) dataset. The x-axis represent the logFCs of each tested gene, and the y-axis represent the logORs for the same genes. The blue lines shows the linear association between the logFC and logOR. The Spearman's rank correlation coefficient (ρ) is also shown in the plot. **c** Barplots of the F-score of four BDA methods and 12 DEA methods on simulated data. Numbers above the barplots show the number of cells that were generated within the simulation. Height of bar defines the median value from 25 simulations, error bars are the first and third quartile. **d** Two density plots of MTUS2 from cortex neuron dataset. The top plot shows the density of MTUS2 in inhibitory neurons and the bottom plot shows the density of MTUS2 in excitatory neurons. **e** Two density plots of Gnb2l1 from the aging mouse atlas droplet dataset. The top plot shows the density of Gnb2l1 in 3-month-old mice and the bottom plot shows the density of Gnb2l1 in 24-month-old mice. Both **d** and **e** are supported with fraction of zeros and variance of each cell population.

## 7.2.3. Differences in test outcomes explained by differences in variance between contrasting cell populations

Despite the observed association between mean expression and number of zeros, which has been previously described(2), and similar performance of the two tests, there were 4,754 and 3,894 genes uniquely identified using BDA and DEA, respectively, across all datasets. To better understand these differences, we highlighted two extreme exemplar cases that were not significant differentially expressed ($P_{FDR} \geq 0.05$), while they were binary differential genes (BDGs, $P_{FDR} \leq 5.27 \times 10^{-115}$). In the *cortex dataset*(25)*, MTUS2* had significantly less zeros in excitatory neurons (logOR = 1.30, $P_{FDR|BDA} = 5.27 \times 10^{-115}$, Fig. 1d) compared to inhibitory neurons, while the median expression levels were not significantly different (logFC = $-1.70 \times 10^{-3}$, $P_{FDR|DEA} = 5.70 \times 10^{-2}$), implying additional high ranked expressions for every additional zero. In the *aging mouse atlas droplet* dataset(7), *Gnb2l1* had significantly less zeros in the 3-month-old mice (logOR = -0.67, $P_{FDR|BDA} = 1.81 \times 10^{-122}$, Fig. 1e) compared to the 24-month-old mice, while again the median expression levels were not significantly different (logFC = $3.15 \times 10^{-3}$, $P_{FDR|DEA} = 6.97 \times 10^{-1}$). These examples show that differences in variance between contrasting cell populations can interfere with the association between observed zeros and mean expression, resulting in disparities between BDA and DEA. Of note, most BDGs-only and DEGs-only had small differences in P-values between the two tests i.e. a borderline significant difference in frequency of zeros while not having a significant difference in median expression (S Fig. 4). The mean $P_{FDR}$ for the 4,754 genes uniquely identified using BDA was $9.57 \times 10^{-3}$, while the mean $P_{FDR}$ of the same genes using DEA was $3.18 \times 10^{-1}$. As for the 3,894 genes uniquely identified using DEA the mean $P_{FDR}$ of BDA was $3.45 \times 10^{-1}$ and mean $P_{FDR}$ of DEA was $8.56 \times 10^{-3}$.

## 7.2.4. Binary differential genes are not driven by technological or biological process

To exclude that the differentially behaving genes between BDA and DEA associate with a specific technological or biological process, we investigated whether there were genes repeatedly detected by a one of the two methods. In most cases, genes that were identified as BDG-only (or DEG-only) were found within a single dataset (S Fig. 5a, S Fig. 5b), suggesting the absence of a driving process for them.

## 7.2.5. Binary differential genes validated with bulk RNA sequencing data

To provide additional insight that differences in zero observations are indeed biologically relevant and represent differential abundance we compared the results of the *Alzheimer's Disease* (AD) dataset(21) (entorhinal cortex) with DEA analysis performed on a bulk RNAseq AD dataset(19), an approach followed by others(26). The bulk RNAseq datatset was comprised of samples from the fusiform gyrus. For genes measured in both, the scRNAseq dataset and the bulk dataset (N = 2,177), the majority of BDG-only (59.4%) were also differentially

expressed in bulk (Fig. 2a). The logOR of the single cell analysis was also significantly correlated with the logFC of the bulk analysis (ρ = 0.39, P = 9.70 × $10^{-79}$, Fig. 2b). Similarly, in a second dataset(22), 65.9% of the BDG-only genes were differentially expressed in bulk samples from the frontal cortex, temporal cortex and hippocampal formation (S Fig 6). Given that the differences in zero observations for genes between the tested groups (expressed in logOR) highly correlates with the differences in median expression in bulk RNAseq data (expressed in logFC), and that the majority of BDG-only were still detected in bulk, further emphasizes that binarized scRNAseq expression data can be used to detect differentially abundant genes.

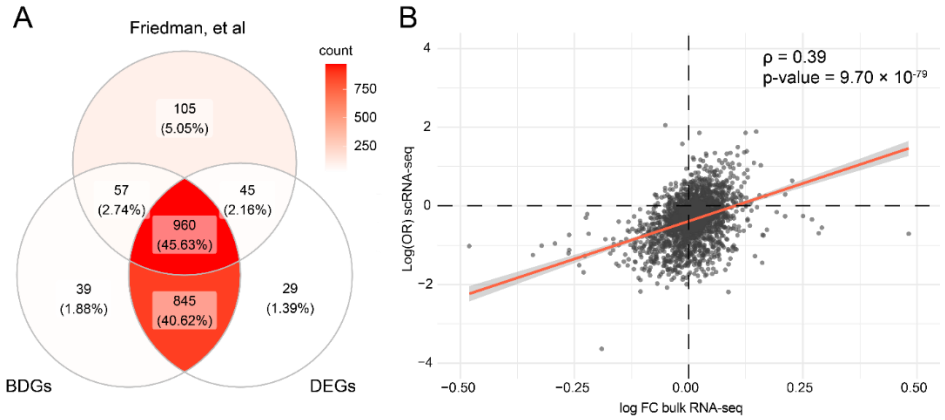



Figure 2 **a,** Venn diagram of genes detected ($P_{FDR}$ ≤ 0.05) in a bulk AD dataset (Friedman, et al), in the single cell AD dataset with BDA-LR (BDGs) and with DEA (DEGs). Each section shows the number and percentage of genes belonging to that section. **b** Plot of the logFC from the AD bulk dataset (x-axis) and the logOR from the single cell AD dataset (y-axis). The red line represents the linear association between the bulk logFC and logOR**.** The Spearman's rank correlation coefficient (ρ) and corresponding association p-value are also shown. Outlier genes (n = 9, **b**) were removed from the plots.

### 7.2.6. Binarization with a threshold of one most appropriate for BDA

To test the binarization scheme, we performed a BDA on the AD dataset for binary profiles generated with different thresholds for binarization (thresholds ranging from one to ten counts). Naturally, for every increase in the threshold, the number of genes with zero measurements across all cells increased, resulting in a decreasing number of tested and significant genes (S Fig 7a, S Fig 7b). With higher thresholds, we found a decrease in correlation of the logORs from BDA with the logFC from DEA (S Fig 7c). These results show that the default binarization scheme where zeros remain zero and every non-zero value is assigned a one, is indeed appropriate.

## 7.3. Discussion

Altogether, our results show that binarized expression patterns across cell populations represent biological variation and can be used as measure of relative abundance of transcripts. Across 16 datasets and a variety of contrasting cell

populations (disease vs healthy, cell types, and cancer status), BDA detected biologically relevant genes that were missed by DEA. While the performance of BDA and DEA on real data is largely comparable, with a known ground truth, BDA performed better than DEA on simulated data. Additionally, BDA benefits reproducibility and is more robust than DEA, since the only pre-processing step required for BDA is the binarization of counts. In contrast, DEA requires normalization and transformation of counts, where an analyst can choose from an excess of equally valid methods(27). Performing BDA on datasets generated using the Smart-seq protocol should be approached with more caution: although the agreement of detected genes between BDA and DEA was high, we observed the lowest correlation between the logOR and logFC for these datasets.

With six of the sixteen datasets we performed the differential analyses between cell types that were based on clusters that were determined with the expression data itself, opposed to a case-control setting. We should note that this is a circular analysis (double dipping) and that the resulting p-values in these comparisons are thus not guaranteed to be controlled for false discoveries. This is, however, still common practice in single-cell differential analyses, as this setup is used to identify cell type markers. For the other ten dataset, the results are not compromised statistically as the case-control definitions are not based on the single-cell data itself.

In our main approach to test for BDA, we have used logistic regression. A logistic regressor for differential expression has been used before(12, 28, 29). These previous applications, however, use continuous expression values of genes as input, while we propose to use the binary expression value. As for MAST(12), a logistic regression on binarized expression values is implemented to take into account the zeros (expressed vs. not expressed) and is combined in a hurdle model with a linear Gaussian model for the continuous values. In contrast to the previously described methods, we show that the frequencies of zeros alone are sufficient to capture biological variation and to identify differential expression of genes between biologically distinct groups in single-cell data.

A commonly used term for observed zeros in single-cell data is dropouts. As zeros in single-cell data can largely be explained by distributional models of molecule sampling counts (2, 3), the use of the term dropout can be misleading, as indicated by Sarkar and Stephens (3). This work contributes to clarifying the origin of zeros in single-cell RNAseq data, by showing that the frequency of zeros can actually be used to identify biological differences.

Performing BDA is normalization-free, time efficient and an accurate alternative for DEA for which we see three potential use cases. First, BDA could be performed in isolation as a fast and accurate alternative to DEA. For different use cases, different BDA tests can be used. For more complex study designs BDA-LR could be used as it allows to adjust for covariates, allowing to take into account biological replicates, which decreases false discoveries(30). More straightforward designs could be performed with BDA-chisq. Second, BDA could be performed in addition to DEA to identify more genes. Finally, BDA could be used to validate pre-processing, normalization and DEA as a big discrepancy between the BDGs and DEGs could indicate an aberration in the DEA results.

# 7.4. Methods

## Single-cell RNA-seq datasets

In total, 16 scRNAseq dataset (14 human and 2 mouse) were used to investigate the utility and biological relevance of binarized expression profiles of genes (Table 1). All datasets had pre-annotated cell types and conditions. From the corresponding references, un-normalized count matrices were acquired, and only annotated cells were kept for further analysis. For each dataset, we extracted the annotated cell type, patient ID, and to which of contrasting cell population the cell belonged from the included meta data. This was slightly different for the *aging mouse atlases* and *cancer atlas*. For the *aging mouse atlases*(7), instead of annotated cell types we retrieved the tissue names. For *the cancer atlas*(8), the contrasting cell populations were defined by cell type, so we retrieved the tissue and the cancer-type for each cell. Each dataset was separately pre-processed. For the Binary Differential Analysis (BDA), the count matrices were transformed to a binary representation, where each zero remain zero and every non-zero value was assigned a one. For the differential expression analysis (DEA), each count matrix was log-normalized using Seurat 3.2.2(9), such that: $y_{ij} = \log\left(\frac{x_{ij}}{\sum_j x_{ij}} \times 10^4\right)$, where $x_{ij}$ and $y_{ij}$ are the raw and normalized values for every gene $i$ in every cell $j$, respectively. This normalizes the feature expression measurements for each cell by the total expression, multiplies this by a scale factor (10,000 by default), and log-transforms the result. The cancer atlas was already normalized, as it was a merger of multiple datasets.

## Statistical analysis

Association p-values were corrected for multiple tests with the Benjamini-Hochberg procedure and significance was assumed at an adjusted P-value of $P_{FDR} \leq 0.05$. Spearman's rank correlation coefficient and the associated p-values were calculated using the `cor.test` function in R v4.0.2.

## Differential expression analysis

DEA was performed using the Wilcoxon Rank Sum test using the FindMarkers function in Seurat 3.2.2(9). Note that the Wilcoxon Rank Sum test from Seurat takes into account zero measurements, and handles them as ties between contrasting cell populations. Genes coding for ribosomal proteins were excluded and we only tested genes that were expressed in at least 10% of the cells in either of the respective groups of interest. This is the default option in the FindMarkers function and speeds up testing by ignoring infrequently expressed genes. Genes coding for ribosomal proteins were excluded. Association p-values were corrected for multiple tests.

## Binary Differential Analysis (BDA)

As the sampling process of biomolecules is the main cause for generating zeros, as illustrated by Svensson(2) and Sarkar and Stephens(3). The probability of measuring a gene is dependent on the relative abundance; more abundant genes

are less likely to result in a zero observation. Extrapolating this to a population of cells, the number of zeros for a gene is representative of the abundance within the respective cell population and differences in the number of zeros between two groups of cells is representative of differential abundance. Lastly, we assume that within a single-cell experiment zeros induced by stochastic processes are not confounded by the groupings. In other words, a stochastically induced zero is equally likely to happen in either cell population, as such, in this setting can be ignored.

## Implementation

In the main analyses, to statistically test for significant differences of zero observations between pre-defined groups in scRNAseq data, we used a logistic regression (BDA-LR). Specifically, the glm(family = “binomial”) function in R v4.0.2, with the binarized expression pattern of the genes as outcome variables and the grouping (i.e. healthy vs diseased) as predictor variable. We have used logistic regression because it allows to add covariates to correct for potential confounding factors. Moreover, predictor variables as well as covariates can be continuous, allowing for complex study designs. All genes that were tested with DEA were also tested with BDA. The resulting association p-values were corrected for multiple tests (see *Statistical analysis*). In addition to logistic regression, we used the Chi-squared test (BDA-chisq), the Fisher's exact test (BDA-fisher) and binary Pearson’s correlation (BDA-Phi) on the simulated data. The Chi-squared test and the Fisher's exact tests were performed with the chisq.test() and fisher.test() R functions, respectively. These tests were performed for each gene on the contingency table representing the binarized gene expression against the pre-defined groupings. The binary Pearson’s correlation was calculated between each binarized gene and the pre-defined groupings and performed with the cor.test() R function, where one group was defined as 0 and the other groups as 1. In a binary setting the outcome statistic of Pearson’s correlation is called Phi (φ).

## BDA – DEA comparison

For the comparison between BDA and DEA, we investigated agreement and disagreement between detected genes and the linear association between the logOR and logFC. Agreement was calculated by the Jaccard index, i.e. number of genes that both tests commonly detected, divided by the total number of genes that were detected. Agreement was calculated on the combination of all datasets and for each individual dataset. The disagreement was investigated by means of inspecting characteristics of BDGs-only and DEGs-only. BDGs-only were defined as genes that were detected ($P_{FDR} \leq 0.05$) by BDA and were not detected ($P_{FDR} > 0.05$) by DEA. Conversely, DEGs-only were defined as genes that were detected ($P_{FDR} \leq 0.05$) by DEA and were not detected ($P_{FDR} > 0.05$) by BDA. The Spearman's rank correlation coefficients between the logOR and logFC were calculated with the estimates of all tested genes of the respective datasets. The scale differences for every dataset, between logOR and logFC, were calculated with a linear model on the estimates of all tested genes of the respective datasets, using the `lm` function in R v4.0.2. The logOR was specified as outcome variable

and the logFC as predictor variable. The resulting slopes were interpreted as scale differences between the logOR and logFC

## Simulation

Data was simulated with muscat 1.2.1(10). The provided PBMC dataset(11) was used as reference. 100 simulated datasets were generated with varying sample sizes (1,000 cells, 2,000 cells, 5,000 cells and 10,000 cells), 25 simulations per sample size. For each simulation 1,000 genes were generated of which 25% were differently expressed between two groups of equal size. For all tests we calculated the False Positive Rate (FPR), Positive Predictive Value (PVV) and accuracy (F1-score) per simulation. Performance was evaluated of 12 DEA methods. 8 methods implemented in Seurat (wilcox, bimod, t, negbinom, poisson, LR, MAST(12), DESeq2(13)), 4 additional methods (DEsingle(14), BPSC(15), monocle(16), limmaVoom(17)) and 4 BDA methods (logistic regression, chi squared test, Fisher's exact test and binary Pearson's correlation). For the runtime benchmark, each run of the simulation of every tests was also timed with `proc.time()` function in R. Tests requiring more than 20 minutes computational time on one simulated dataset were excluded.

## Validation with existing bulk RNA-seq data

The AD bulk RNA-seq datasets were acquired from Gemma(18). The first dataset from Friedman, et al(19) consisted of 33 controls (CT) and 84 samples from individuals diagnosed with Alzheimer's Disease (AD) collected from the fusiform gyrus. This dataset was reprocessed by Gemma and no batch effects were present. For the differential expression analysis in bulk we used the Wilcoxon Rank Sum test from limma v3.44.3(20). In total, 2,228 genes were tested for differential expression, as these genes were also included in the scRNAseq AD dataset(21) analysis. The second dataset from Hokama, et al(22) consisted of 47 controls and 32 AD samples. The samples originated from the frontal cortex ($N_{CT}$ = 18, $N_{AD}$ = 15), temporal cortex ($N_{CT}$ = 19, $N_{AD}$ = 10) and hippocampal formation ($N_{CT}$ = 10, $N_{AD}$ = 7). The data was reprocessed and batch corrected by Gemma. For the differential expression analysis no distinction was made between brain regions. In total, 2,001 genes were tested for differential expression. All resulting association p-values were corrected for multiple tests. For validation, the significant BDGs and DEGs from the scRNAseq AD dataset analyses were compared with the significantly differentially expressed genes from the bulk analyses. Venn diagrams were plotted with ggVennDiagram(v0.3)(23). Correlations were calculated between the logOR and logFC of the single-cell analysis with the bulk logFC.

# References

1. Choi,K., Chen,Y., Skelly,D.A. and Churchill,G.A. (2020) Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics. *Genome Biol.*, **21**, 183.

2. Svensson,V. (2020) Droplet scRNA-seq is not zero-inflated. *Nat. Biotechnol.*, **38**, 147–150.

3. Sarkar,A. and Stephens,M. (2021) Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nat. Genet. 2021 536*, **53**, 770–777.

4. Andrews,T.S., Hemberg,M. and Birol,I. (2019) M3Drop: Dropout-based feature selection for scRNASeq. *Bioinformatics*, **35**, 2865–2867.

5. Li,R. and Quon,G. (2019) ScBFA: Modeling detection patterns to mitigate technical noise in large-scale single-cell genomics data. *Genome Biol.*, **20**, 193.

6. Qiu,P. (2020) Embracing the dropouts in single-cell RNA-seq analysis. *Nat. Commun.*, **11**, 1–9.

7. Almanzar,N., Antony,J., Baghel,A.S., Bakerman,I., Bansal,I., Barres,B.A., Beachy,P.A., Berdnik,D., Bilen,B., Brownfield,D., *et al.* (2020) A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature*, **583**, 590–595.

8. Nieto,P., Elosua-Bayes,M., Trincado,J.L., Marchese,D., Massoni-Badosa,R., Salvany,M., Henriques,A., Mereu,E., Moutinho,C., Ruiz,S., *et al.* (2020) A Single-Cell Tumor Immune Atlas for Precision Oncology. *bioRxiv*, 10.1101/2020.10.26.354829.

9. Stuart,T., Butler,A., Hoffman,P., Hafemeister,C., Papalexi,E., Mauck,W.M., Hao,Y., Stoeckius,M., Smibert,P. and Satija,R. (2019) Comprehensive Integration of Single-Cell Data. *Cell*, **177**, 1888-1902.e21.

10. Crowell,H.L., Soneson,C., Germain,P.L., Calini,D., Collin,L., Raposo,C., Malhotra,D. and Robinson,M.D. (2020) muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat. Commun.*, **11**, 1–12.

11. Kang,H.M., Subramaniam,M., Targ,S., Nguyen,M., Maliskova,L., McCarthy,E., Wan,E., Wong,S., Byrnes,L., Lanata,C.M., *et al.* (2018) Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.*, **36**, 89–94.

12. Finak,G., McDavid,A., Yajima,M., Deng,J., Gersuk,V., Shalek,A.K., Slichter,C.K., Miller,H.W., McElrath,M.J., Prlic,M., *et al.* (2015) MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol. 2015 161*, **16**, 1–13.

13. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol. 2014 1512*, **15**, 1–21.

14. Miao,Z., Deng,K., Wang,X. and Zhang,X. (2018) DEsingle for detecting three types of differential expression in single-cell RNA-seq data. *Bioinformatics*, **34**, 3223–3224.

15. Vu,T.N., Wills,Q.F., Kalari,K.R., Niu,N., Wang,L., Rantalainen,M. and Pawitan,Y. (2016) Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics*, **32**, 2128–2135.

16. Trapnell,C., Cacchiarelli,D., Grimsby,J., Pokharel,P., Li,S., Morse,M., Lennon,N.J., Livak,K.J., Mikkelsen,T.S. and Rinn,J.L. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol. 2014 324*, **32**, 381–386.

17. Law,C.W., Chen,Y., Shi,W. and Smyth,G.K. (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol. 2014 152*, **15**, 1–17.

18. Zoubarev,A., Hamer,K.M., Keshav,K.D., McCarthy,E.L., Santos,J.R.C., Van Rossum,T., McDonald,C., Hall,A., Wan,X., Lim,R., *et al.* (2012) Gemma: a resource for the reuse, sharing and meta-analysis of expression profiling data. *Bioinformatics*, **28**, 2272–2273.

19. Friedman,B.A., Srinivasan,K., Ayalon,G., Meilandt,W.J., Lin,H., Huntley,M.A., Cao,Y., Lee,S.H., Haddick,P.C.G., Ngu,H., *et al.* (2018) Diverse Brain Myeloid Expression Profiles Reveal Distinct Microglial Activation States and Aspects of Alzheimer's Disease Not Evident in Mouse Models. *Cell Rep.*, **22**, 832–847.

20. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**.

21. Grubman,A., Chew,G., Ouyang,J.F., Sun,G., Choo,X.Y., McLean,C., Simmons,R.K., Buckberry,S., Vargas-Landin,D.B., Poppe,D., *et al.* (2019) A single-cell atlas of entorhinal cortex from individuals

with Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat. Neurosci.*, **22**, 2087–2097.

22. Hokama,M., Oka,S., Leon,J., Ninomiya,T., Honda,H., Sasaki,K., Iwaki,T., Ohara,T., Sasaki,T., LaFerla,F.M., *et al.* (2014) Altered expression of diabetes-related genes in Alzheimer's disease brains: The Hisayama study. *Cereb. Cortex*, **24**, 2476–2488.

23. Gao,C.-H. (2019) ggVennDiagram: A 'ggplot2' Implement of Venn Diagram.

24. Soneson,C. and Robinson,M.D. (2018) Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods*, **15**, 255–261.

25. Agarwal,D., Sandor,C., Volpato,V., Caffrey,T.M., Monzón-Sandoval,J., Bowden,R., Alegre-Abarrategui,J., Wade-Martins,R. and Webber,C. (2020) A single-cell atlas of the human substantia nigra reveals cell-specific pathways associated with neurological disorders. *Nat. Commun.*, **11**, 1–11.

26. Mathys,H., Davila-Velderrain,J., Peng,Z., Gao,F., Mohammadi,S., Young,J.Z., Menon,M., He,L., Abdurrob,F., Jiang,X., *et al.* (2019) Single-cell transcriptomic analysis of Alzheimer's disease. *Nature*, **570**, 332–337.

27. Lytal,N., Ran,D. and An,L. (2020) Normalization Methods on Single-Cell RNA-seq Data: An Empirical Survey. *Front. Genet.*, **11**, 41.

28. Satija,R., Farrell,J.A., Gennert,D., Schier,A.F. and Regev,A. (2015) Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol. 2015 335*, **33**, 495–502.

29. Ntranos,V., Yi,L., Melsted,P. and Pachter,L. (2019) A discriminative learning approach to differential expression analysis for single-cell RNA-seq. *Nat. Methods 2019 162*, **16**, 163–166.

30. Squair,J.W., Gautier,M., Kathe,C., Anderson,M.A., James,N.D., Hutson,T.H., Hudelle,R., Qaiser,T., Matson,K.J.E., Barraud,Q., *et al.* (2021) Confronting false discoveries in single-cell differential expression. *Nat. Commun. 2021 121*, **12**, 1–15.

31. Nagy,C., Maitra,M., Tanti,A., Suderman,M., Théroux,J.F., Davoli,M.A., Perlman,K., Yerko,V., Wang,Y.C., Tripathy,S.J., *et al.* (2020) Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. *Nat. Neurosci.*, **23**, 771–781.

32. Segerstolpe,Å., Palasantza,A., Eliasson,P., Andersson,E.M., Andréasson,A.C., Sun,X., Picelli,S., Sabirsh,A., Clausen,M., Bjursell,M.K., *et al.* (2016) Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab.*, **24**, 593–607.

33. Wilk,A.J., Rustagi,A., Zhao,N.Q., Roque,J., Martínez-Colón,G.J., McKechnie,J.L., Ivison,G.T., Ranganath,T., Vergara,R., Hollis,T., *et al.* (2020) A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nat. Med.*, **26**, 1070–1076.

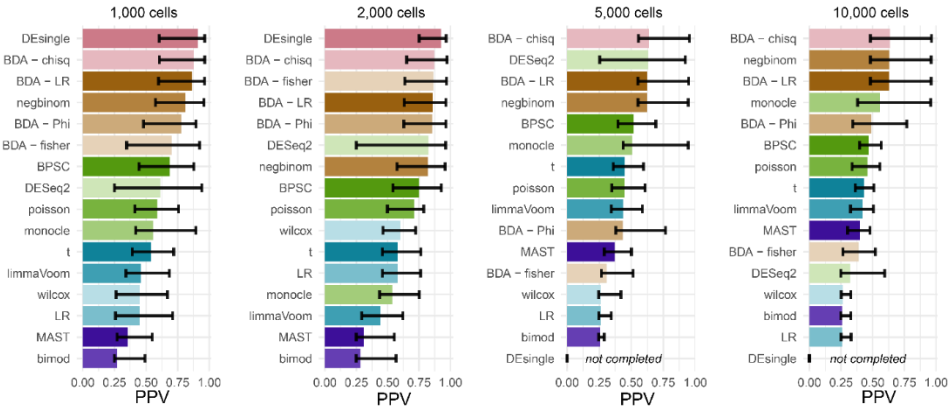
34. Kim,N., Kim,H.K., Lee,K., Hong,Y., Cho,J.H., Choi,J.W., Lee,J. Il, Suh,Y.L., Ku,B.M., Eum,H.H., *et al.* (2020) Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat. Commun.*, **11**, 1–15.

35. Wu,T.D., Madireddi,S., de Almeida,P.E., Banchereau,R., Chen,Y.J.J., Chitre,A.S., Chiang,E.Y., Iftikhar,H., O'Gorman,W.E., Au-Yeung,A., *et al.* (2020) Peripheral T cell expansion predicts tumour infiltration and clinical response. *Nature*, **579**, 274–278.

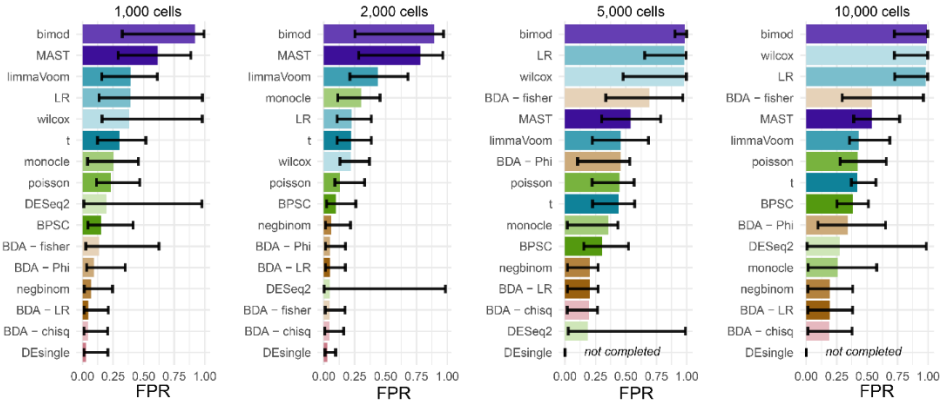
36. Hodge,R.D., Bakken,T.E., Miller,J.A., Smith,K.A., Barkan,E.R., Graybuck,L.T., Close,J.L., Long,B., Johansen,N., Penn,O., *et al.* (2019) Conserved cell types with divergent features in human versus mouse cortex. *Nature*, **573**, 61–68.

37. Lee,H.O., Hong,Y., Etlioglu,H.E., Cho,Y.B., Pomella,V., Van den Bosch,B., Vanhecke,J., Verbandt,S., Hong,H., Min,J.W., *et al.* (2020) Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nat. Genet.*, **52**, 594–603.

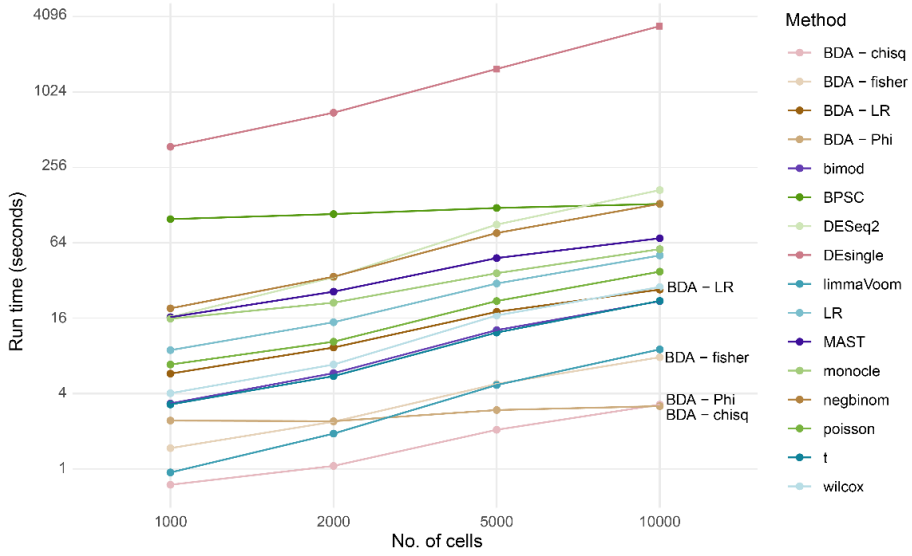
# Supplements



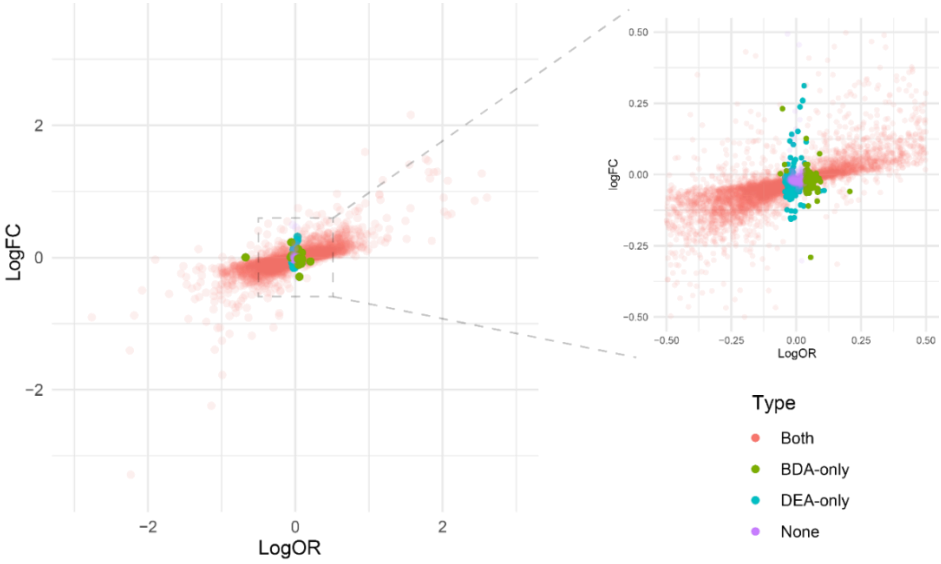

**Supplementary Figure** 1 Barplots of Positive predictive value (PPV) of four BDA and 12 DEA methods on simulated data. Numbers above the barplots shows the number of cells that were generated within the simulation. Height of bar defines the median value from 25 simulations, error bars the first and third quartile.



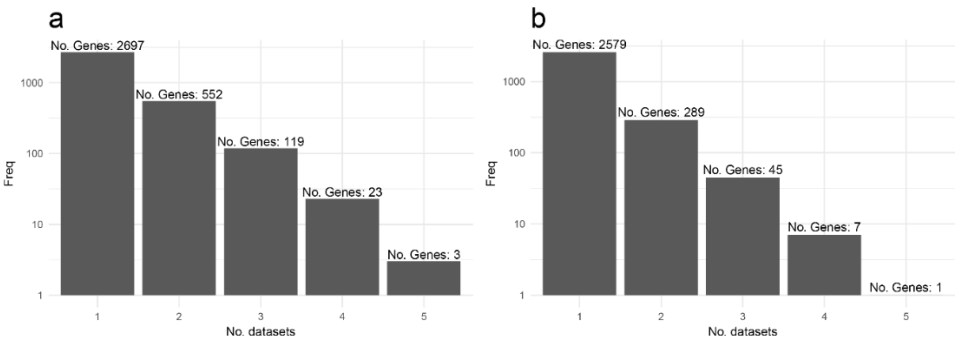

**Supplementary Figure** 2 Barplots of False positive rate (FPR) of four BDA and 12 DEA methods on simulated data. Numbers above the barplots shows the number of cells that were generated within the simulation. Height of bar defines the median value from 25 simulations, error bars the first and third quartile.



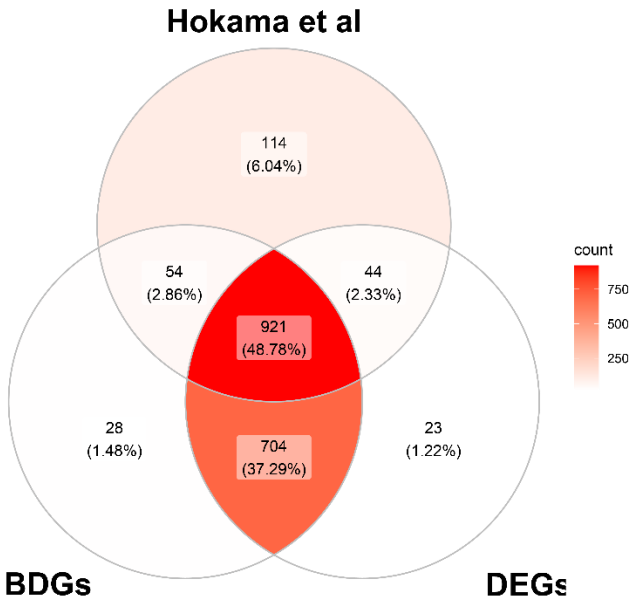

**Supplementary Figure** 3 Results of the timing benchmark. Y-axis represents runtime in seconds on a log2 scale and X-axis the number of cells. Dots are median runtimes based on 25 runs. Squares (DESingle at 5,000 and 10,000 cells) are based on one simulation, as it exceeded the 20 minute threshold only one run was available.



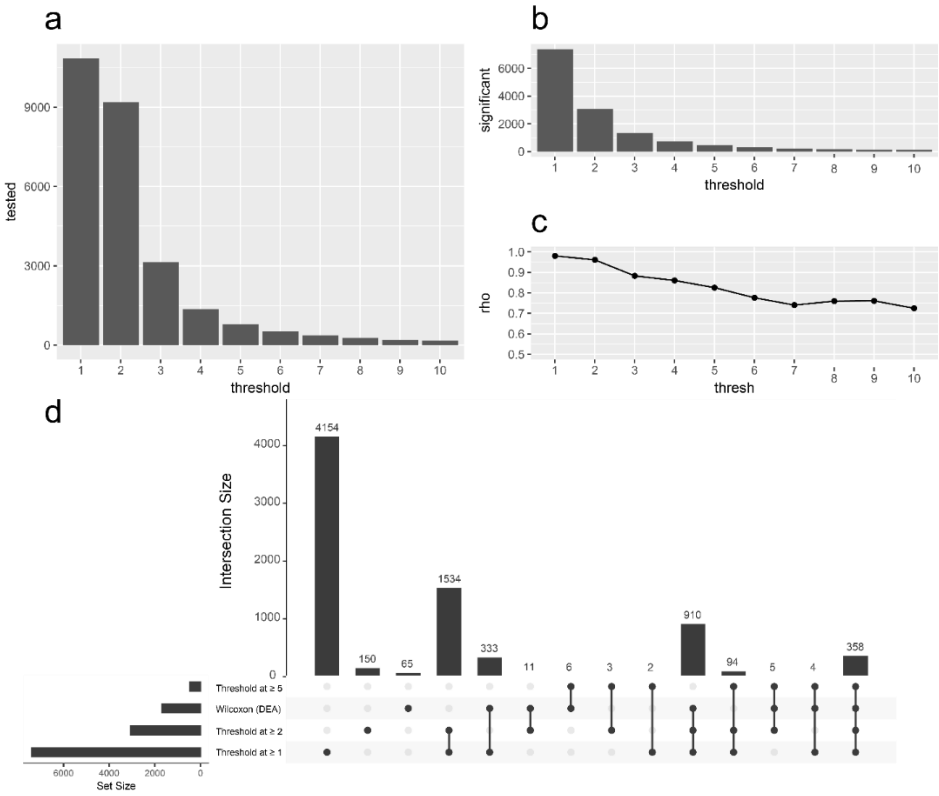

**Supplementary Figure** 4 LogOR – LogFC plot of the Aging mouse atlas droplet dataset. X-axis represents the effect size of differences in frequency of zeros (logOR) and the y-axis represents the effects size of differences of median expression (logFC). Red dots are genes significantly detected by both tests. Purple dots are genes that do not show a significant difference according to both tests. Green dots were only detected with BDA and blue dots were only detected with DEA.




**Supplementary Figure 5a ,** Bar plot of the number of genes that were identified as DDG-only in i datasets. The x-axis represent the number of datasets and the y-axis represent the number of genes that were found with that many datasets. For instance, 2740 genes were detected with a single dataset and 3 genes were detected with 5 datasets. **b,** Bar plot of the number of genes that were repeatedly identified as DEG-only. The x-axis represent the number of datasets and the y-axis represent the number of genes that were found with that many datasets.




**Supplementary Figure 6,** Venn diagram of genes detected ($P_{FDR} \leq 0.05$) in bulk AD dataset (Hokama, et al), in the single cell AD dataset with BDA and with DEA. Each section shows the number and percentage of genes belonging to that section




**Supplementary Figure 7a,** Bar plot of total number of tested genes at each binarization threshold. The x-axis represents the binarization threshold and the y-axis represent the number of genes that were tested with that threshold. **b,** Bar plot of total number of significant genes ($P_{FDR} \leq 0.05$) at each binarization threshold. The x-axis represents the binarization threshold and the y-axis represent the number of genes that were significant with that threshold. **c,** Pearson's correlation coefficient of the logOR of the BDA with different thresholds with the logFC of the DEA. The x-axis represents the binarization threshold and the y-axis represents The Spearman's rank correlation coefficient with that threshold. **d,** Upset plot of detected genes ($P_{FDR} \leq 0.05$) with a threshold of 1, 2, 5 and DEA (Wilcoxon).

# Consequences and opportunities arising due to sparser single-cell RNA-seq datasets

Gerard Bouland, Ahmed Mahfouz and Marcel Reinders

# Abstract

With the number of cells measured in single-cell RNA sequencing (scRNA-seq) datasets increasing exponentially and concurrent increased sparsity due to more zero counts being measured for many genes, we demonstrate here that downstream analyses on binary-based gene expression give similar results as count-based analyses. Moreover, a binary representation scales up to ~50-fold more cells that can be analyzed using the same computational resources. We also highlight the possibilities provided by binarized scRNA-seq data. Development of specialized tools for bit-aware implementations of downstream analytical tasks will enable a more fine-grained resolution of biological heterogeneity.


# 8.1. Background

Since its introduction, single-cell RNA sequencing (scRNA-seq) has been vital in investigating biological questions that were previously impossible to answer[1–4]. Continuous technological innovations are resulting in a consistent increase in the number of cells and molecules being measured in a single experiment. However, at the same time, datasets appear to become sparser, i.e. more zero measurements across the whole dataset. The sparsity has generally been seen as a problem, especially since standard count distribution models (e.g. Poisson) do not account for the excess of zeros. [5–8]. This sparked discussions about whether the excess of zeros can be explained by mainly technological or biological factors[5,8–10]. Jiang et al.[8] discuss the 'zero-inflation controversy', in which a distinction is made between a biological zero, indicating the true absence of a transcript, and a non-biological zero, indicating failure of measuring a transcript that was present in the cell. Similarly, Sarkar and Stephens[11] make a distinction between measurement and expression. They proposed a model that is a combination of an expression model that encodes the true absence of a transcript, i.e. a (biological) zero, with a measurement model, for which they use a Poisson model (which can result in non-biological zeros due to limited sequencing depth). Consequently, even non-biological zeros encode useful biological information as then the gene is unlikely to be highly expressed. Or, in other words: all zeros in scRNA-seq datasets have biological significance. Aligned with this, Qui et al. [12] proposed to 'embrace' all zeros as useful signal and developed a clustering algorithm requiring only binarized scRNA-seq data (a zero representing a zero count and a one for non-zero counts). Using binarized scRNA-seq data, Qui et al. identified clusters similar to clusters identified using a count-based approach. Although this was the first paper explicitly embracing zeros as useful signal, binarization of scRNA-seq was already used to infer gene regulatory networks[13]. Since then, several methods have employed binarized scRNA-seq data. For instance, scBFA[14], a dimensionality reduction method for binarized scRNA-seq data, showed improved visualization and classification of cell identity and trajectory inference when compared to methods that use count data. Likewise, we introduced Binary Differential Analysis (BDA)[15], a differential expression analysis method relying on binarized scRNA-seq data. We showed that differential expression analysis on binary representations of scRNA-seq data faithfully captures biological variation across cell types and conditions.

Provided that a binarized data representation has the potential to reduce required computational resources considerably, and as scRNA-seq datasets are becoming increasingly bigger and sparser, we wondered if binary should be the preferred data representation for other tasks. In this work, we explore the consequences of sparser datasets and the applicability of binarized scRNA-seq data for various single-cell analysis tasks.

## 8.2. Results and Discussion

We downloaded 56 datasets published between 2015 and 2021. Based on these datasets, a clear association between the year of publication and the number of cells can be observed (Pearson's correlation coefficient of r = 0.46, Fig. 1a). For instance, the average dataset in 2015 (n = 7) had 704 cells while the average dataset in 2020 (n = 7) had 58,654 cells. Another clear trend that can be seen is that an increasing number of cells is highly correlated with decreasing detection rates (fraction of non-zero values) (Pearson's correlation coefficient of r = -0.47, Fig. 1b). Note that this trend of measuring more cells per dataset outweighs improved chemistry over time, and thus still results in sparser datasets. It is likely that this trend will continue over the next years as, for many biological questions, shallow sequencing of many cells is more cost effective than deep sequencing of a few cells(16). Moreover, by measuring more cells we can better estimate the probability whether a gene is expressed, and the overall power to detect differentially expressed genes in a given dataset increases [17].This trend will be amplified, as more population scale and multi-condition scRNA-seq datasets are emerging [17,18], for which a low coverage sequencing is sufficient to capture cell type specific gene expression (given enough cells are measured per individual and per cell type)[19]. Altogether, these developments will result in sparser scRNA-seq datasets with larger numbers of cells.

As zeros become more abundant, a binarized expression might be as informative as counts. Using ~1.5 million cells from 56 datasets, we observed on average a strong point-biserial correlation (Pearson correlation coefficient p = 0.93) between the normalized expression counts of a cell and its respective binarized variant, although differences between datasets exist (Additional file 1: Fig. S1). This strong correlation implies that the binarized signal already captures most of the signal present in the normalized count data. This strong correlation is primarily explained by the detection rate (Additional file 1: Fig. S2a) and the variance of the non-zero counts of a cell (Additional file 1: Fig. S2b). In cells where the detection rate is low (many zeros) and the variance of the non-zero counts is small, the correlation between the normalized expression values and their binary representation is high (Fig. 1c). Across all datasets, the detection rate and variance of measured expressions were good predictors for the correlation between the binary representation and the normalized representation, although differences between technologies exist (Fig. 1d). This indicates that as datasets become sparser, counts become less informative with respect to binarized expression.

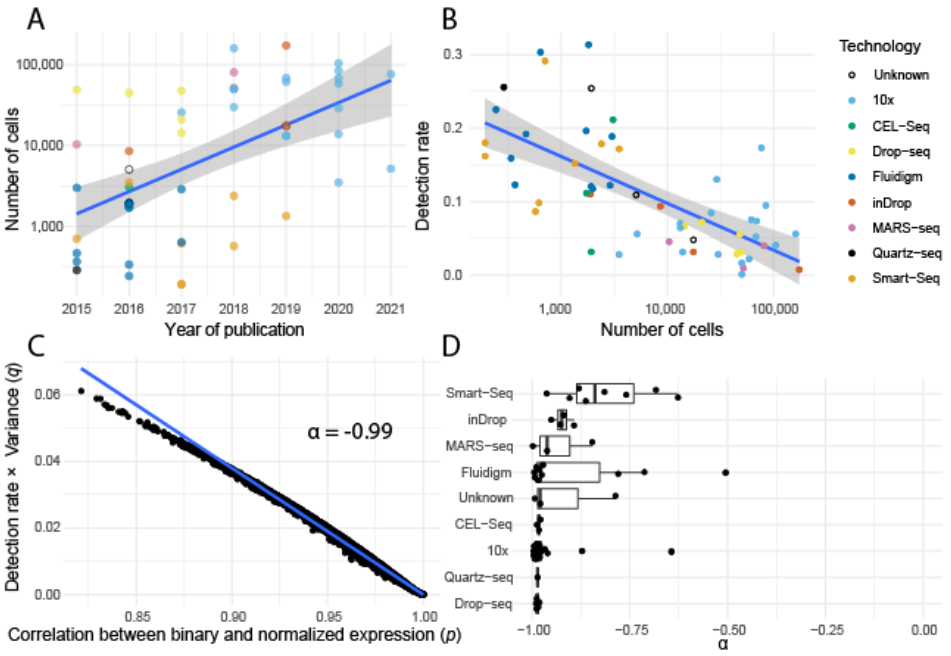



**Figure 1**: More cells, more zeros. Binarized scRNA-seq datasets were generated by binarizing the raw count matrix, where zero remains zero and every non-zero value is assigned a one. A) Association between year of publication, total number of cells. Scatterplot of the number of cells (log scale) against the date of publication. B) Scatterplot of the detection rate (y-axis) against the number of cells (log scale, x-axis). C) On the x-axis the Pearson's correlation coefficient (p) of every cell from the PaulHSC dataset between the binarized and normalized expressions. On the y-axis the product of the detection rate and the variance of the non-zero values (q). α is the Pearson's correlation coefficient between these values p and q across all cells. D) Boxplots of the α-values for all 56 datasets grouped by technology. One dataset (LawlorPancreasData) was excluded as α-value (α = 0.42) for this dataset was a clear outlier.

To assess whether counts can actually be discarded in practice, we assessed whether binarized data can give comparable results to counts in four common single-cell analysis tasks: (1) dimensionality reduction for visualization, (2) data integration, (3) cell type identification, and (4) differential expression analysis using pseudobulk. First, for dimensionality reduction, we used three different dimensionality reduction approaches on binarized scRNA-seq data; (i) scBFA [14], (ii) PCA (Fig. 2a), and (iii) eigenvectors of the Jaccard cell-cell similarity matrix (see Additional file 2). All three approaches were compared to the standard approach of applying PCA to the normalized counts (Fig. 2b, Additional file 1: Fig. S3). Further, for all four methods the first ten components were used to generate a non-linear embedding using UMAP (Additional file 1: Fig. S4). Qualitatively, we observed that the results of binary-based dimensionality reduction are comparable to standard count-based methods. This was confirmed quantitatively, as the pairwise distances between cells based on the binary-based UMAPs were highly correlated with the pairwise distances from the count-based UMAP ($r \geq 0.73$, Additional file 1: Fig. S5). Especially the UMAP generated with the binary-based PCs was visually very similar to the UMAP generated with the count-based PCs (Fig. 2c-d). Calculating the silhouette score (SS) for each cell type

with the reduced dimensions (n = 10) resulted in slightly lower scores for scBFA (SS = 0.32) and binary-based PCA(SS = 0.39) compared to the count-based PCA (SS = 0.44) (Additional file 1: Fig. S6). However, in the UMAP space (2-dimensional), silhouette scores for scBFA (SS = 0.43) and binary-based PCA (SS = 0.42) were higher than count-based PCA (SS = 0.35).

Second, we integrated three scRNA-seq datasets[20–22] with Harmony[23], using count- and binary-based PCA. Both, visually and quantitatively, we observed an improved mixing of cells for the binary representation (LISI = 1.18) as compared to counts (LISI = 1.12) (Additional file 1: Fig. S7-S8). Third, we evaluated the effect of binarization on cell annotation using (i) marker genes and (ii) classification methods. Using a set of known brain cell type markers[24], we annotated the binarized AD dataset[20] based on solely the detection of respective cell type markers (See Additional file 2). The annotations were compared to cell type labels that were originally assigned based on the markers' expression level (i.e., counts). We observed a high level of concordance between annotations as quantified by a median F1-score of 0.93. (Additional file 1: Fig. S9). Additionally, we found that the visualization of the binarized expression of cell type markers to be highly similar to the visualization of their normalized expression in UMAP plots (Fig. 2e-h, Additional file 1: Fig. S10). Next, we compared the performance of automatic cell type identification using scPred and SingleR[25,26] on 22 datasets for which cell type annotations were available. The median F1-scores were highly similar between cell type identifications based on the binarized and the normalized count data, despite large variation of sparseness between these datasets. This finding implies that counts do not add information for cell type identification. This conclusion was further supported by randomly shuffling the non-zero counts, which resulted in a comparable performance (Fig. 2I, Additional file 1: Fig. S11).

Forth, we evaluated whether counts can also be discarded when pseudobulk data is used for differential expression analysis[18]. In a dataset containing scRNA-seq data of the prefrontal cortex of 34 individuals[27], we generated pseudobulk data by either taking the mean expression of each gene across all cells, or the fraction of non-zero values across all cells (detection rate), per individual. The Spearman's rank correlation between the binarized profile and the mean counts (across all genes) was ≥ 0.99 (Additional file 1: Fig. S12) for every individual, implying that pseudobulk aggregation with binarized expression faithfully represents counts. To quantify this further, we generated 960 datasets using muscat [18] with 96 unique simulation settings (see Additional file 2). In each dataset, pseudobulk data for each individual was generated and we identified differential expressed genes using Limma trend[28] for the mean gene expression and a t-test for the detection rate. In general, the F1-scores for the count and binary representations were very similar across the different settings, however, with small sample sizes and fewer cells, analyses based on a count representation performed better, while analyses based on a binarized expression performed better with larger sample sizes and more cells (Additional file 1: Fig. S13). Additionally, count-based analyses resulted in more false positives (Additional file 1: Fig. S14) while binarized-based analyses resulted in more false negatives (Additional file 1: Fig. S15). The false negatives were primarily due to

highly expressed genes that show no differences in the detection rate. At larger sample sizes and with more cells, the false negatives diminished (Additional file 1: Fig. S16). All together, these result show that most of the information is indeed captured in the binary representation, only when genes have a high detection rate (>0.9), or when the number of cells per sample becomes low, then, changes in expression are not reflected in the binary representation and, consequently, information from counts is needed.

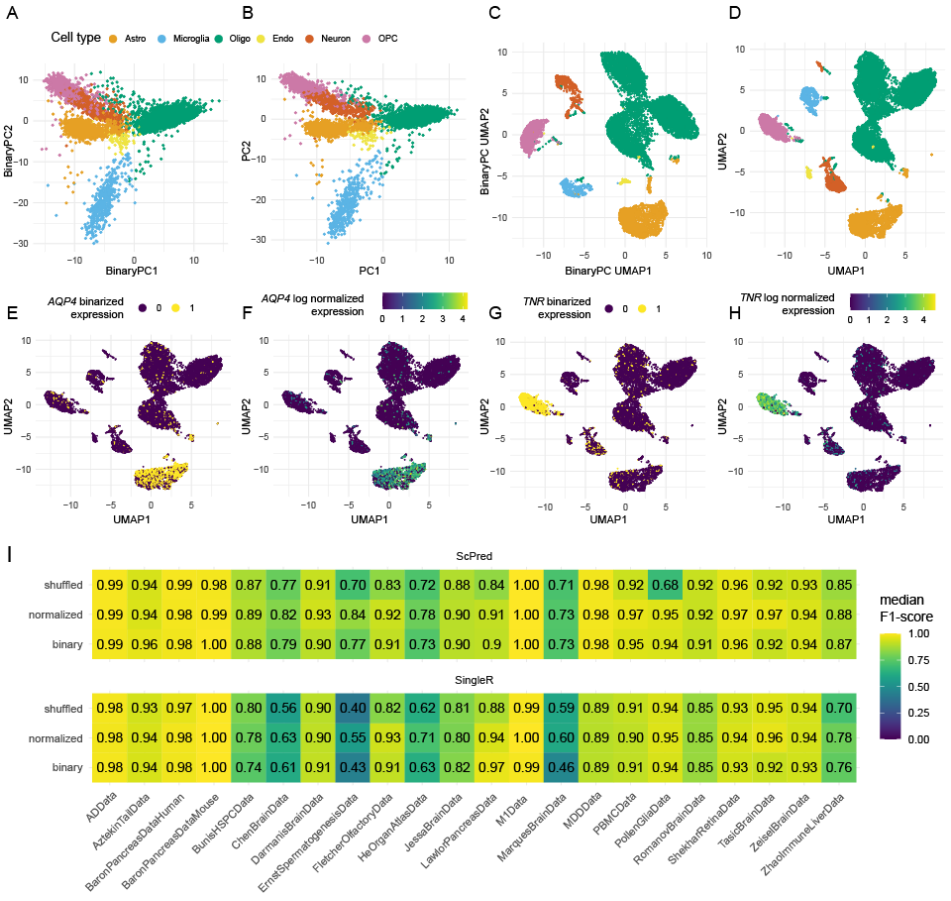



**Figure 2:** A,B) Cells plotted against the first two principle components of the AD dataset[20] (A) PCA based on binary representation, and (B) PCA based on count representation. UMAP generated from data presented with C) the binary-based PCs and D) the count-based PCs. Colors indicate annotated cell type. E,H) UMAP based on the count based PCs, in which cells are colored according to the binary representation of the marker genes *AQP4* (E) and *TNR* (H) which are known markers for astrocytes and OPCs respectively[24] F,G) Similar as E and H but showing the normalized expression of the marker gene I) The performance (median F1-score) of cell type identification by SingleR[25] and scPred[26] when applied to binary (binarized data), normalized (normalized expression) and shuffled (shuffled normalized expression) for 22 datasets.

Whether zero-inflation associates with technical or biological origins is heavily debated[8]. One compelling reason for this debate is the fact that within a single dataset some genes are zero-inflated, while others are not[5,8]. We argue that

this observation is mostly related to whether a gene is only expressed in a subpopulation of cells (e.g. marker genes) or whether a gene has a stable expression (e.g. housekeeping genes). To substantiate our claim, we used BDA[15] to identify the top 100 most differentially expressed genes between two cell populations and the top 100 most stable expressed genes in a 10X dataset[21] as well as a Smart-Seq dataset[29]. Next, we applied scRATE[5] to identify the best distribution model for the observed expression of the identified genes, being either a Poisson, a Negative Binomial or their zero-inflated counterparts. A fisher exact test showed that a zero-inflated model was enriched in the top 100 differentially expressed genes, and a non-zero inflated model was enriched in the top 100 stable expressed genes (Table 1). Hence, like earlier work [5], we conclude biological heterogeneity to be the main driver of zero-inflation.

**Table 1:** Enrichment of zero-inflated distributions for the top100 differential expressed genes and the enrichment of non-zero inflated distributions for the top100 stable genes.

| Platform | Top 100 | Zero-inflated | Not zero-inflated | logOR | 95%CI | p-value |
|---|---|---|---|---|---|---|
| 10x | Differentially expressed genes | 99 | 1 | 5.19 | 3.36, 8.87 | $3.03 \times 10^{-25}$ |
| | Stable genes | 35 | 65 | | | |
| Smart-seq | Differentially expressed genes | 97 | 3 | 3.70 | 2.50, 5.36 | $5.46 \times 10^{-18}$ |
| | Stable genes | 44 | 56 | | | |

Increasingly larger datasets require increasingly more computational resources. The storage required for all 56 datasets used in this study was 764 Gigabytes after normalization using sctransform[30], or 276 Gigabytes when log-normalized and stored as sparse matrices. In contrast, binarizing the same datasets and storing them as bits required only 73 Gigabytes, which is an ~11-fold and ~4-fold reduction in storage requirements, respectively (Additional file 1: Fig. S17). Yet, there are big differences across datasets. For example, a reduction of ~50-fold and ~20-fold, respectively was acquired for the BuettnerESC dataset[31]. The amount of storage that can be saved is highly correlated with the detection rate (Additional file 1: Fig. S18), with the highest gain for datasets with a high detection rate. The considerable storage reduction of the binary representation gives the potential to boost downstream analyses to larger numbers of cells, opening possibilities to get a more fine-grained resolution of biological heterogeneity[32].

We showed that analyses based on a binary representation of scRNA-seq data perform on par with count-based analyses. Working with binarized scRNA-seq data has clear additional advantages. The first is simplicity. For the various tasks that we explored, such as dimensionality reduction, data integration, cell type prediction, differential expression analysis[15] and clustering[12], the binary representations required no normalization. Hence various subjective choices on the normalization could be avoided, which improves reproducibility of these tasks. However, as sequencing depth has an effect on the detection rate of a cell, it is likely this is not the case for all downstream tasks. Second, binarization reduces the amount of required storage significantly and allows the analysis of significantly larger datasets. For example, binary-based data allow for a bit implementation of clustering as has been done before in the field of molecular

dynamics resulting in a significant reduction of run time and peak memory usage compared to existing methods[33]. It has also been suggested that binarization alleviates noise[14] as it is insensitive to count errors. However, binarization remains sensitive to detection errors caused by, e.g., the presence of ambient RNA. Consequently, detection of ambient RNA[34] poses a challenge for binary representations when studying individual cells, and thus might require specialized methods to be developed.

At first glance, binarizing scRNA-seq data seems to remove signal. However, genes that are highly expressed across cells will not have a lot of zeros, whereas genes that are lowly expressed across cells will have many. This implies we might be able to infer the relative expression of a gene within an individual cell by exploiting the detection pattern of similar other cells. Using this reasoning, we indeed were able to reconstruct the expression levels of genes from the detection pattern using neighboring cells ( Additional file 1: Fig. S19, Additional file 2). Hence, we conclude that the detection rate of a gene in a group of cells, such as a cell type, do faithfully represents the (mean) expression levels of that gene in that group of cells, underpinning why binarization for most of the downstream tasks apparently does not have lost signal.

We have shown that sparsity is inversely correlated with the amount of additional signal that is captured with counts. Consequently, binarization will not be useful for all scRNA-seq datasets. Previous work suggested that when the detection rate is >90%, visualizations based on the binary representation do not perform on par with count-based representation [14]. With our simulation experiments, we have shown a similar trend when considering the task of detecting differential expressed genes based on pseudobulk values.

## 8.3. Conclusion

Concluding, our results support existing literature in showing that binarized scRNA-seq data can be used for: dimensionality reduction, data integration, visualization, clustering, trajectory inference, batch correction, differential expression analysis and cell type prediction. We believe scRNA-seq tool developers should be aware of the possibility of using a binary representation of the scRNA-seq data instead of count-based data, as it gives opportunities to develop computational- and time-efficient tools.

## References


1. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).

2. Van Der Wijst, M. G. P. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493–497 (2018).

3. La Manno, G. *et al.* RNA velocity of single cells. *Nat. 2018 5607719* **560**, 494–498 (2018).

4. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat. Methods 2019 168* **16**, 715–721 (2019).

5. Choi, K., Chen, Y., Skelly, D. A. & Churchill, G. A. Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics. *Genome Biol.* **21**, 183 (2020).

6. Pierson, E. & Yau, C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* **16**, 1–10 (2015).

7. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods 2014 117* **11**, 740–742 (2014).

8. Jiang, R., Sun, T., Song, D. & Li, J. J. Statistics or biology: the zero-inflation controversy about scRNA-seq data. *Genome Biol. 2022 231* **23**, 1–24 (2022).

9. Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology* vol. 38 147–150 at https://doi.org/10.1038/s41587-019-0379-5 (2020).

10. Cao, Y., Kitanovski, S., Küppers, R. & Hoffmann, D. UMI or not UMI, that is the question for scRNA-seq zero-inflation. *Nat. Biotechnol. 2021 392* **39**, 158–159 (2021).

11. Sarkar, A. & Stephens, M. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nat. Genet. 2021 536* **53**, 770–777 (2021).

12. Qiu, P. Embracing the dropouts in single-cell RNA-seq analysis. *Nat. Commun.* **11**, 1–9 (2020).

13. Moignard, V. *et al.* Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nat. Biotechnol. 2015 333* **33**, 269–276 (2015).

14. Li, R. & Quon, G. ScBFA: Modeling detection patterns to mitigate technical noise in large-scale single-cell genomics data. *Genome Biol.* **20**, 1–20 (2019).

15. Bouland, G. A., Mahfouz, A. & Reinders, M. J. T. Differential analysis of binarized single-cell RNA sequencing data captures biological variation. *NAR Genomics Bioinforma.* **3**, (2021).

16. Zhang, M. J., Ntranos, V. & Tse, D. Determining sequencing depth in a single-cell RNA-seq experiment. *Nat. Commun. 2020 111* **11**, 1–11 (2020).

17. Schmid, K. T. *et al.* scPower accelerates and optimizes the design of multi-sample single cell transcriptomic studies. *Nat. Commun. 2021 121* **12**, 1–18 (2021).

18. Crowell, H. L. *et al.* muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat. Commun.* **11**, 1–12 (2020).

19. Mandric, I. *et al.* Optimized design of single-cell RNA sequencing experiments for cell-type-specific eQTL analysis. *Nat. Commun. 2020 111* **11**, 1–9 (2020).

20. Grubman, A. *et al.* A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat. Neurosci.* **22**, 2087–2097 (2019).

21. Bakken, T. E. *et al.* Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nat. 2021 5987879* **598**, 111–119 (2021).

22. Nagy, C. *et al.* Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. *Nat. Neurosci. 2020 236* **23**, 771–781 (2020).

23. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods 2019 1612* **16**, 1289–1296 (2019).

24. McKenzie, A. T. *et al.* Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. *Sci. Reports 2018 81* **8**, 1–19 (2018).

25. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol. 2019 202* **20**, 163–172 (2019).

26. Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q. & Powell, J. E. ScPred: Accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol.* **20**, 1–17 (2019).

27. Nagy, C. *et al.* Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. *Nat. Neurosci.* **23**, 771–781 (2020).

28. Ritchie, M. E. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

29. Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).

30. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol. 2019 201* **20**, 1–15 (2019).

31. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155–160 (2015).

32. Sikkema, L. *et al.* An integrated cell atlas of the human lung in health and disease. *bioRxiv* 2022.03.10.483747 (2022) doi:10.1101/2022.03.10.483747.

33. González-Alemán, R. *et al.* BitClust: Fast Geometrical Clustering of Long Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **60**, 444–448 (2020).

34. Yang, S. *et al.* Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol.* **21**, 1–15 (2020).

35. Bouland G.A., Mahfouz A., Reinders M.J.T., Arising_sparsity_scRNAseq. Github; https://github.com/gbouland/Arising_sparsity_scRNAseq (2023).

36. Bouland G.A., Mahfouz A., Reinders M.J.T., Consequences and opportunities arising due to sparser single-cell RNA-seq datasets. Zenodo; https://doi.org/10.5281/zenodo.7732380 (2023).

37. Chew G, Grubman A, Ouyang JF, Rackham O, Polo J, Petretto E. A single-cell atlas of the human cortex reveals drivers of transcriptional changes in Alzheimer's disease in specific cell subpopulations. Gene Expression Omnibus. https://identifiers.org/geo:GSE138852 (2019).

38. Turecki G. Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. Gene Expression Omnibus. https://identifiers.org/geo:GSE144136 (2020).

39. Abdelaal T, Michielsen L, Cats D, Hoogduin D, Me H, Reinders M.J.T, Mahfouz A. A comparison of automatic cell identification methods for single-cell RNA-sequencing data. Zenodo. https://doi.org/10.5281/zenodo.3357167 (2019)

40. Risso D, Cole M (2022). scRNAseq: Collection of Public Single-Cell RNA-Seq Datasets. R package version 2.8.0.

# 8.4. Methods

## Datasets

A total of 56 scRNA-seq datasets were used of which 52 datasets were downloaded using the scRNA-seq R-package (v 2.8.0). Four additional datasets were acquired from the corresponding sources (**Table 2**).

**Table 2:** Overview of datasets

| Name | Description | Number | Tech | Reference |
|---|---|---|---|---|
| AztekinTailData | tail | 13199 | 10x | (1) |

| | | | | |
|---|---|---|---|---|
| BachMammaryData | mammary gland | 25806 | 10x | (2) |
| BacherTCellData | T cells | 104417 | 10x | (3) |
| BaronPancreasData(HumanMouse) | pancreas | 8569 | inDrop | (4) |
| BaronPancreasData(HumanMouse) | pancreas | 1886 | inDrop | (4) |
| BuettnerESCData | embryonic stem cells | 288 | Quartz-seq | (5) |
| BunisHSPCData | haematopoietic stem and progenitor | 5183 | 10x | (6) |
| CampbellBrainData | brain | 21086 | Drop-seq | (7) |
| ChenBrainData | brain | 14437 | Drop-seq | (8) |
| DarmanisBrainData | brain | 466 | Fluidigm | (9) |
| ErnstSpermatogenesisData | testis | 68937 | 10x | (10) |
| FletcherOlfactoryData | olfactory epithelium | 616 | Smart-Seq | (11) |
| GrunHSCData | haematopoietic stem cells | 1915 | CEL-Seq | (12) |
| GrunPancreasData | pancreas | 1728 | CEL-Seq | (12) |
| GiladiHSCData | haematopoietic stem cells | 81024 | MARS-seq | (13) |
| HeOrganAtlasData | various organs | 84363 | 10x | (14) |
| HuCortexData | cortex | 48000 | Drop-seq | (15) |
| KolodziejczykESCData | embryonic stem cells | 704 | Smart-Seq | (16) |
| JessaBrainData | brain | 61595 | 10x | (17) |
| LaMannoBrainData('human-es') | embryonic stem cells | 1715 | Fluidigm | (18) |
| LaMannoBrainData('human-embryo') | embryonic midbrain | 1977 | Fluidigm | (18) |
| LaMannoBrainData('human-ips') | induced pluripotent stem cells | 337 | Fluidigm | (18) |
| LaMannoBrainData('mouse-adult') | adult dopaminergic neurons | 243 | Fluidigm | (18) |
| LaMannoBrainData('mouse-embryo') | embyronic midbrain | 1907 | Fluidigm | (18) |
| LawlorPancreasData | pancreas | 638 | Fluidigm | (19) |
| LedergorMyelomaData | bone marrow plasma cells | 51840 | MARS-seq | (20) |
| LunSpikeInData('416b') | 416B cells | 192 | Smart-Seq | (21) |

| | | | | |
|---|---|---|---|---|
| LunSpikeInData('troph o') | trophoblasts | 192 | Smart-Seq | (21) |
| MacoskoRetinaData | retina | 49300 | Drop-seq | (22) |
| MairPBMCData | peripheral blood mononuclear cells | 29033 | 10x | (23) |
| KotliarovPBMCData | peripheral blood mononuclear cells | 58654 | 10x | (24) |
| MarquesBrainData | brain | 5069 | | (25) |
| MessmerESCData | embryonic stem cells | 1344 | Smart-Seq | (26) |
| MuraroPancreasData | pancreas | 3072 | CEL-Seq | (27) |
| NestorowaHSCData | haematopoietic stem cells | 1920 | | (28) |
| PaulHSCData | haematopoietic stem cells | 10368 | MARS-seq | (29) |
| PollenGliaData | outer radial glia | 367 | Fluidigm | (30) |
| RichardTCellData | CD8+ T cells | 572 | Smart-Seq | (31) |
| RomanovBrainData | brain | 2881 | Fluidigm | (32) |
| SegerstolpePancreasData | pancreas | 3514 | Smart-Seq | (33) |
| ShekharRetinaData | retina | 44994 | Drop-seq | (34) |
| StoeckiusHashingData (mode='mouse') | peripheral blood mononuclear cells | 50000 | 10x | (35) |
| StoeckiusHashingData (mode='human') | peripheral blood mononuclear cells | 50000 | 10x | (35) |
| StoeckiusHashingData (type='mixed') | HEK, THP1, K562, KG1 cells | 30000 | 10x | (35) |
| TasicBrainData | brain | 1809 | Fluidigm | (36) |
| WuKidneyData | kidney | 17542 | | (37) |
| ZeiselBrainData | brain | 3005 | Fluidigm | (38) |
| ZeiselNervousData | nervous system | 160796 | 10x | (39) |
| ZhaoImmuneLiverData | liver immune cells | 68100 | 10x | (40) |
| ZhongPrefrontalData | prefrontal cortex | 2394 | Smart-Seq | (41) |
| ZilionisLungData | lung | 173954 | inDrop | (42) |
| ZilionisLungData('mouse') | lung | 17549 | inDrop | (42) |

| | | | | |
|---|---|---|---|---|
| ADData | brain | 13214 | 10x | (43) |
| M1Data | brain | 76533 | 10x | (44) |
| MDDData | brain | 13881 | 10x | (45) |
| PBMCData | peripheral blood mononuclear cells | 3500 | 10x | (46) |

## Binarization and the detection rate

Binarized scRNA-seq datasets were generated by transforming the raw count matrix such that a zero remains a zero and every non-zero value is assigned a one. The detection rate refers to the fraction of non-zero values. More formally, binarized scRNA-seq data is generated as follows:

$$y_{ij} = \begin{cases} 1 & x_{ij} \geq 1 \\ 0 & \text{otherwise} \end{cases}, \quad \text{for} \quad i \in [1, g], j \in [1, n]$$

where $x_{ij}$ is the expression of gene i in cell j; g is number of genes and n the number of cells in the dataset.

The detection rate $DR_j$ for cell j is then defined as:

$$DR_j = \frac{\sum_{i=1}^{g} y_{ij}}{g}, \qquad \forall j \in [1, n]$$

Similarly, we define the detection rate for a gene $DR_i$ and the detection rate across the whole dataset $DR_d$:

$$DR_i = \frac{\sum_{j=1}^{n} y_{ij}}{n}, \qquad \forall i \in [1, g]$$

$$DR_d = \frac{\sum_{i=1}^{g} \sum_{j=1}^{n} y_{ij}}{ng}$$

Note, the detection rate of a gene can be determined either within a specific cell population or within the dataset.

## Log normalization

Log normalization on scRNA-seq datasets was performed as follows: given a count matrix ($X$) where $x_{ij}$ is the expression of gene i in cell j, the log-normalized version was generated, such that $y_{ij} = \log\left(\frac{x_{ij}}{\sum_j x_{ij}} \times 10^4\right)$, where $y_{ij}$ normalized values for every gene i in every cell j, respectively.

## Dimensionality reduction

Dimensionality reduction was performed on the Alzheimer's Disease dataset from Grubman et al(43). All dimensionality reductions (count- and binary-based) were performed with the same set of highly variable genes. These genes were identified using M3drop(47). The count-based dimensionality reduction (Principal Component Analysis, PCA) was performed using the default Seurat(48) pipeline on data that was log-normalized and scaled. The default Seurat pipeline was also applied on binary data (binary-PCA), however, without the normalization step. Additionally, for the binary-based dimensionality reduction scBFA(49) was used and eigen vectors of the jaccard cell-cell similarity matrix were calculated(Jaccard Eigen Vectors, JEVs). For the comparison of UMAP plots, the first 10 components of all four dimensionality reductions were used to calculate pair-wise Euclidian distances between cells and subsequently obtain the non-linear UMAP embeddings. The resulting plots were visually inspected on whether the cells clustered together according to previous annotations. Cell type annotations were obtained from the original study.

## Batch correction

Three brain datasets (ADData(43), M1Data(44) and MDDData(45))were used for batch correction. From the three datasets, astrocytes, endothelial cells, microglia and oligodendrocytes were extracted, and cell type labels were harmonized. Then, the three datasets were combined and count- and binary-based PCs were calculated using Seurat. These PCs (n = 10) were used as input for Harmony(50). The uncorrected and batch corrected PCs were then used as input for the UMAP and to evaluate count- and binary-based batch corrected data.

## Use of marker genes with binary data

To evaluate the use of marker genes with binarized scRNA-seq data, we annotated the cells from the ADData(43) dataset using markers from the BRETIGEA R-package(51). From the list of marker genes, we construct a one-hot-encoded marker matrix ($M$), where the columns represent cell types and the rows genes. When, gene $g$ is a marker for cell type $c$, then, $m_{gc} = 1$, otherwise $m_{gc} = 0$. Then, we subset the dataset, such that only known marker genes remain. Next, we calculate the Pearson's correlation between the cell type vector ($m_c$) and the binarized expression of the cell ($y_j$). For every cell $j$ we get a measure of association ($\varphi$) with every cell type $c$. Higher values indicate higher association. As such, we annotated every cell as the cell type for which $\varphi$ is the highest. The approach was evaluated by comparing the annotations with the

annotation from the original study. F1-scores for every cell type were calculated and the median F1-score was reported.

## Automatic cell-type identification

Automatic cell-type identification was performed using two existing automatic cell-type identification methods, scPred and SingleR(52, 53). Both methods were applied to all datasets for which cell type labels were available (n = 22 out of 56). Three versions were made of every dataset, (i) a log-normalized version, (ii) a binarized version, and (iii) a shuffled version. The shuffled version was made by randomly shuffling all non-zero values of the log-normalized version. Note, that all zeros remained zero. For all three data representations, 10 reference / target splits were randomly made of 75% (reference) and 25% (target) of the total number of cells. For scPred, Seurat(48) was used to scale the data (zero-mean and standard variance) and calculate the principal components (PCs), which was done on all three data representations. Of note, the binarized data representations were not normalized. SingleR requires no specific pre-processing. The predicted labels were compared with the true labels by calculating the F1-score for every cell type and taking the median of F1-scores across all cell types, using the evaluation function of caret(54). The median F1-score of all 10 runs were used to evaluate the predictions.

## scRNA-seq data simulation and differential expression analysis

scRNA-seq data was simulated with muscat(55) using the provided dataset(56) as reference. In total, 96 settings were generated to evaluate the performance of binarized scRNA-seq data when performing differential expression analysis (DEA) on pseudo bulk data. For the percentage of differentially expressed genes we evaluated 1%, 10%, 20%, 30%, 40%, and 50%. We evaluated datasets with 1,000, 5,000, 10,000, and 50,000 cells. And, for the number of individuals per group we evaluated 5, 10, 20, and 50 individuals. Each combination of the aforementioned settings was evaluated, resulting in the 96 settings. For each setting we generated ten datasets of 500 genes, resulting in 960 simulated datasets. Pseudo bulk data was generated with the mean as aggregation function using the aggregateData function from muscat. Here, for each individual the mean expression of each gene was calculated based on all cells belonging to the respective individual. For the binarized data, the detection rate per gene was calculated as the number of cells per individual in which the gene is observed divided by the total number of cells belonging to the respective individual. The mean pseudo bulk data was normalized using the calcNormFactors function from edgeR(57) and DEA was performed using Limma Trend(58). The t-test was used for binarized data, without normalization. P-values were corrected for multiple testing using the Benjamini-Hochberg procedure. Genes were considered significantly detected at $P_{adj} \leq 0.05$.

## Identification of best count distribution model

To test whether zero-inflation in scRNA-seq data can be explained by biological heterogeneity, we reasoned that a marker gene is a prime example of biological

heterogeneity: it being highly expressed in a specific cell population while virtually absent in other cells. As such, we hypothesized that if zero-inflation is primarily explained by biological heterogeneity, marker genes should be zero-inflated. Using the same reasoning, a stably expressed gene should not be zero-inflated. We tested both hypotheses on two brain datasets(10x(44)and a Smart-Seq v2(59)). We selected two cell types and performed differential expression analyses between the cell types, using BDA(60). Next, we selected the top 100 most differentially expressed genes (sorted on $P_{FDR}$), as well as the top 100 most stably expressed genes (sorted on smallest fold changes). Using scRATE(61), we fitted four different count distribution models (a Poisson, Negative-binomial and their zero-inflated counter parts) to all 200 genes individually. Using a leave-one-out cross validation test, we selected the best count distribution model for each gene, based on the best predictive accuracy. For every gene, we then know whether it is a marker gene or stably expressed, and whether it is zero-inflated or not. With a fisher exact test, we finally evaluate the association between zero-inflated/not zero inflated with marker/stable.

## Comparison of bit-stored and normalized datasets

For the bit-stored datasets, the binary-based datasets were stored as Boolean vectors using the bit R-package(v4.0.4). All count-based datasets were log-normalized using Seurat(48) or normalized using scTransform(62). Before the comparison of the required storage, the normalized matrices were stored as sparse matrices.

## Magnitude recovery

To recover the magnitude of expression from binary-based data, first pairwise cell similarities were calculated using the Jaccard index (JI). Next, for every cell, the neighbourhood is determined by the closest neighbour according to the JI, and the respective cell itself. Then, for every gene in a cell, a weighted average of the binary profile is calculated based on the neighbourhood. The weight is determined by the JI and is proportional to the sum of JIs. After this, the dataset of weighted averages is log normalized, such that $y_{ij} = \log\left(\frac{x_{ij}}{\sum_j x_{ij}} \times 10^4\right)$, where $x_{ij}$ and $y_{ij}$ are the weighted averages and normalized values for every gene i in every cell j, respectively. Finally, all non-zero values that were originally zero in the binary-based data are set to zero.

## References (methods)


1. Aztekin, C. *et al.* Identification of a regeneration organizing cell in the Xenopus tail. *Science* **364**, 653 (2019).

2. Bach, K. *et al.* Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nat. Commun. 2017 81* **8**, 1–11 (2017).

3. Bacher, P. *et al.* Low-Avidity CD4 + T Cell Responses to SARS-CoV-2 in Unexposed Individuals and Humans with Severe COVID-19. *Immunity* **53**, 1258-1271.e5 (2020).

4. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.* **3**, 346-360.e4 (2016).

5. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155–160 (2015).

6. Bunis, D. G. *et al.* Single-Cell Mapping of Progressive Fetal-to-Adult Transition in Human Naive T Cells. *Cell Rep.* **34**, (2021).

7. Campbell, J. N. *et al.* A molecular census of arcuate hypothalamus and median eminence cell types. *Nat. Neurosci. 2017 203* **20**, 484–496 (2017).

8. Chen, R., Wu, X., Jiang, L. & Zhang, Y. Single-Cell RNA-Seq Reveals Hypothalamic Cell Diversity. *Cell Rep.* **18**, 3227–3241 (2017).

9. Darmanis, S. *et al.* A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 7285–7290 (2015).

10. Ernst, C., Eling, N., Martinez-Jimenez, C. P., Marioni, J. C. & Odom, D. T. Staged developmental mapping and X chromosome transcriptional dynamics during mouse spermatogenesis. *Nat. Commun.* **10**, (2019).

11. Fletcher, R. B. *et al.* Deconstructing Olfactory Stem Cell Trajectories at Single-Cell Resolution. *Cell Stem Cell* **20**, 817-830.e8 (2017).

12. Grün, D. *et al.* De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell Stem Cell* **19**, 266–277 (2016).

13. Giladi, A. *et al.* Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. *Nat. Cell Biol.* **20**, 836–846 (2018).

14. He, S. *et al.* Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs. *Genome Biol.* **21**, (2020).

15. Hu, P. *et al.* Dissecting Cell-Type Composition and Activity-Dependent Transcriptional State in Mammalian Brains by Massively Parallel Single-Nucleus RNA-Seq. *Mol. Cell* **68**, 1006-1015.e7 (2017).

16. Kolodziejczyk, A. A. *et al.* Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. *Cell Stem Cell* **17**, 471–485 (2015).

17. Jessa, S. *et al.* Stalled developmental programs at the root of pediatric brain tumors. *Nat. Genet.* **51**, 1702–1713 (2019).

18. La Manno, G. *et al.* Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell* **167**, 566-580.e19 (2016).

19. Lawlor, N. *et al.* Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res.* **27**, 208–222 (2017).

20. Ledergor, G. *et al.* Single cell dissection of plasma cell heterogeneity in symptomatic and asymptomatic myeloma. *Nat. Med.* **24**, 1867–1876 (2018).

21. Lun, A. T. L., Calero-Nieto, F. J., Haim-Vilmovsky, L., Göttgens, B. & Marioni, J. C. Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. *Genome Res.* **27**, 1795–1806 (2017).

22. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).

23. Mair, F. *et al.* A Targeted Multi-omic Analysis Approach Measures Protein Expression and Low-Abundance Transcripts on the Single-Cell Level. *Cell Rep.* **31**, (2020).

24. Kotliarov, Y. *et al.* Broad immune activation underlies shared set point signatures for vaccine responsiveness in healthy individuals and disease activity in patients with lupus. *Nat. Med.* **26**, 618–629 (2020).

25. Marques, S. *et al.* Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science (80-. ).* **352**, 1326–1329 (2016).

26. Messmer, T. *et al.* Transcriptional Heterogeneity in Naive and Primed Human Pluripotent Stem Cells at Single-Cell Resolution. *Cell Rep.* **26**, 815-824.e4 (2019).

27. Muraro, M. J. *et al.* A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst.* **3**, 385-394.e3 (2016).

28. Nestorowa, S. *et al.* A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood* **128**, e20–e31 (2016).

29. Paul, F. *et al.* Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell* **163**, 1663–1677 (2015).

30. Pollen, A. A. *et al.* Molecular identity of human outer radial glia during cortical development. *Cell* **163**, 55–67 (2015).

31. Richard, A. C. *et al.* T cell cytolytic capacity is independent of initial stimulation strength. *Nat. Immunol.* **19**, 849–858 (2018).

32. Romanov, R. A. *et al.* Molecular interrogation of hypothalamic organization reveals distinct dopamine neuronal subtypes. *Nat. Neurosci.* **20**, 176–188 (2017).

33. Segerstolpe, Å. *et al.* Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab.* **24**, 593–607 (2016).

34. Shekhar, K. *et al.* Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell* **166**, 1308-1323.e30 (2016).

35. Stoeckius, M. *et al.* Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, (2018).

36. Tasic, B. *et al.* Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* **19**, 335–346 (2016).

37. Wu, H., Kirita, Y., Donnelly, E. L. & Humphreys, B. D. Advantages of single-nucleus over single-cell RNA sequencing of adult kidney: Rare cell types and novel cell states revealed in fibrosis. *J. Am. Soc. Nephrol.* **30**, 23–32 (2019).

38. Zeisel, A. *et al.* Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).

39. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174**, 999-1014.e22 (2018).

40. Zhao, J. *et al.* Single-cell RNA sequencing reveals the heterogeneity of liver-resident immune cells in human. *Cell Discov.* **6**, (2020).

41. Zhong, S. *et al.* A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* **555**, 524–528 (2018).

42. Zilionis, R. *et al.* Single-Cell Transcriptomics of Human and Mouse Lung Cancers Reveals Conserved Myeloid Populations across Individuals and Species. *Immunity* **50**, 1317-1334.e10 (2019).

43. Grubman, A. *et al.* A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat. Neurosci.* **22**, 2087–2097 (2019).

44. Bakken, T. E. *et al.* Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nat. 2021 5987879* **598**, 111–119 (2021).

45. Nagy, C. *et al.* Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. *Nat. Neurosci.* **23**, 771–781 (2020).

46. Senabouth, A. *et al.* Comparative performance of the BGI and Illumina sequencing technology for single-cell RNA-sequencing. *NAR Genomics Bioinforma.* **2**, (2020).

47. Andrews, T. S., Hemberg, M. & Birol, I. M3Drop: Dropout-based feature selection for scRNASeq. *Bioinformatics* **35**, 2865–2867 (2019).

48. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

49. Li, R. & Quon, G. ScBFA: Modeling detection patterns to mitigate technical noise in large-scale single-cell genomics data. *Genome Biol.* **20**, 1–20 (2019).

50. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods 2019 1612* **16**, 1289–1296 (2019).

51. McKenzie, A. T. *et al.* Brain Cell Type Specific Gene Expression and Co-expression Network Architectures. *Sci. Reports 2018 81* **8**, 1–19 (2018).

52. Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q. & Powell, J. E. ScPred: Accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol.* **20**, 1–17 (2019).

53. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol. 2019 202* **20**, 163–172 (2019).

54. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **28**, 1–26 (2008).

55. Crowell, H. L. *et al.* muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat. Commun.* **11**, 1–12 (2020).

56. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).

57. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

58. Ritchie, M. E. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

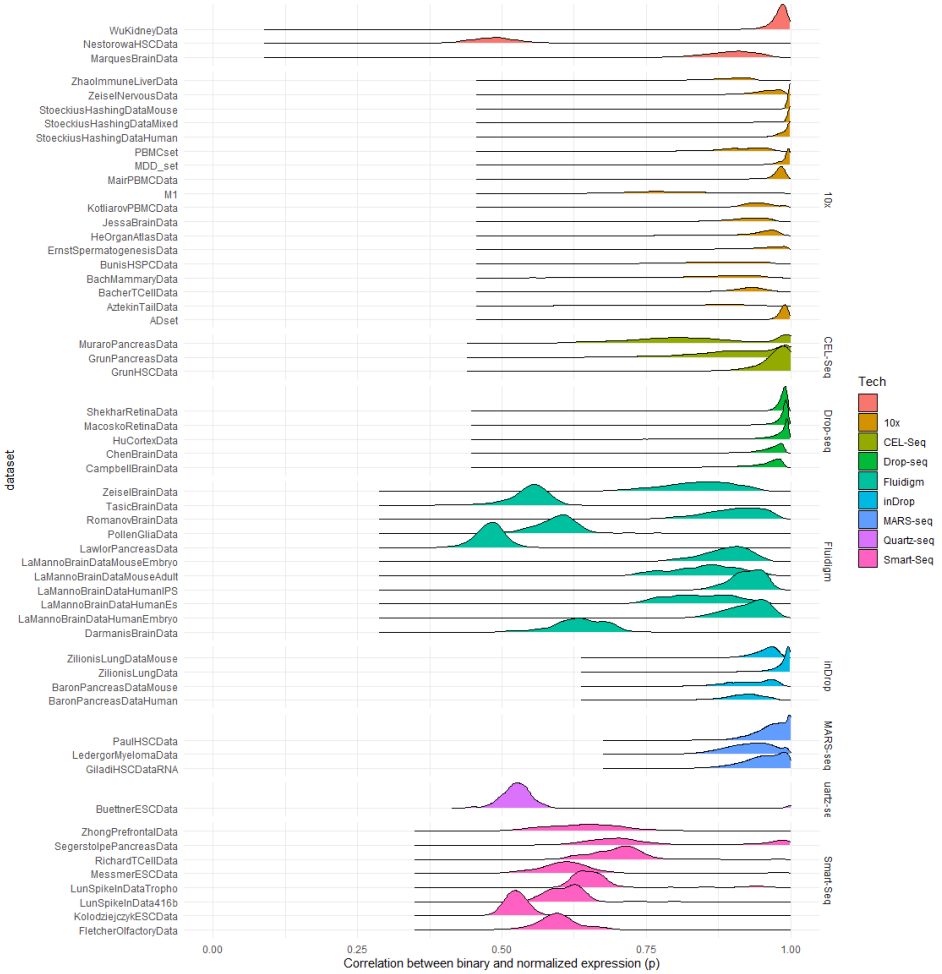
59. Hodge, R. D. *et al.* Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).

60. Bouland, G. A., Mahfouz, A. & Reinders, M. J. T. Differential analysis of binarized single-cell RNA sequencing data captures biological variation. *NAR Genomics Bioinforma.* **3**, (2021).

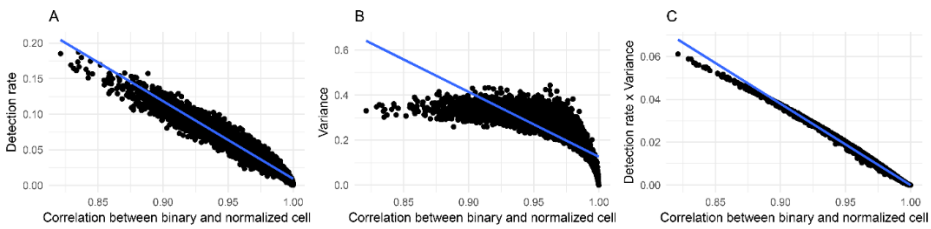
61. Choi, K., Chen, Y., Skelly, D. A. & Churchill, G. A. Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics. *Genome Biol.* **21**, 183 (2020).

62. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol. 2019 201* **20**, 1–15 (2019).

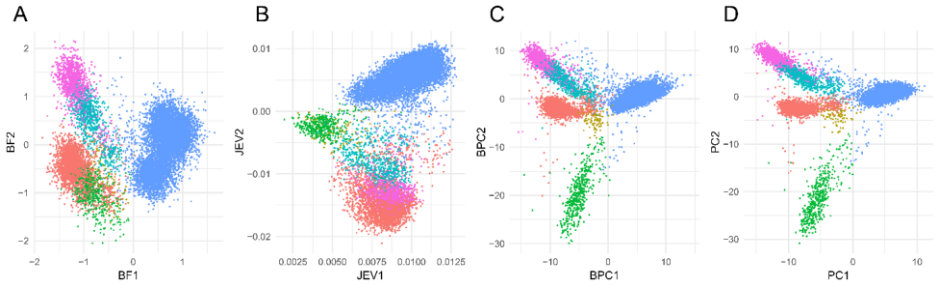
# Supplements



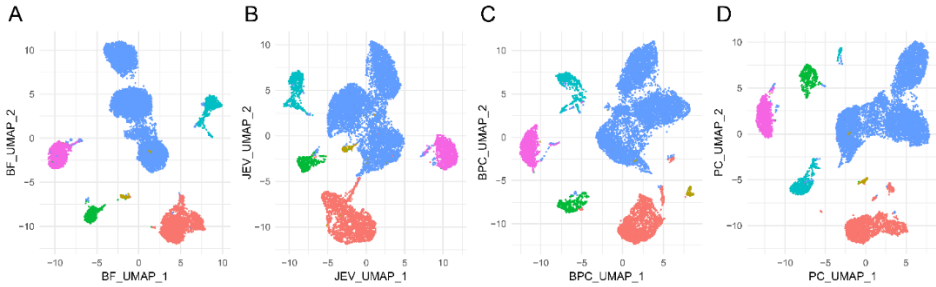

**Supplementary Figure 1:** The distributions of correlation coefficients between the binarized and count-based expressions of every cell (p, x-axis) within each dataset (y-axis). The datasets are grouped by technology.



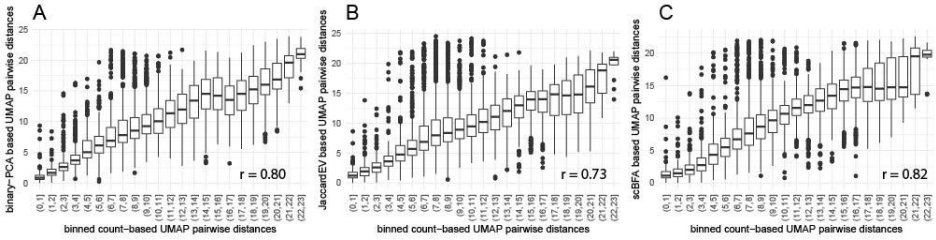

**Supplementary Figure 2: A,B,C)** Every dot is a cell from the PaulHSC dataset. The x-axis represents the correlation coefficient between the binarized and count-based representation. **A)** The y-axis is the detection rate, **B)** the y-axis is the variance of the binarized representation of a gene across all cells, and **C)** the y-axis is the product of detection rate and the variance of non-zero counts.



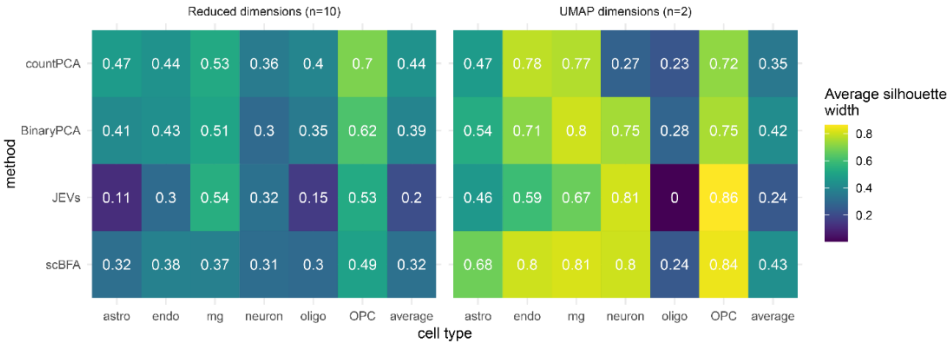

**Supplementary Figure 3:** Comparison of binary-based dimensionality reduction on AD Dataset, all points are colored based on pre-annotated cell types. **A)** First two components from scBFA method. **B)** First two components from the Jaccard similarity eigenvectors. **C)** First two components from binary-based PCA. **D)** First two components from count-based PCA.



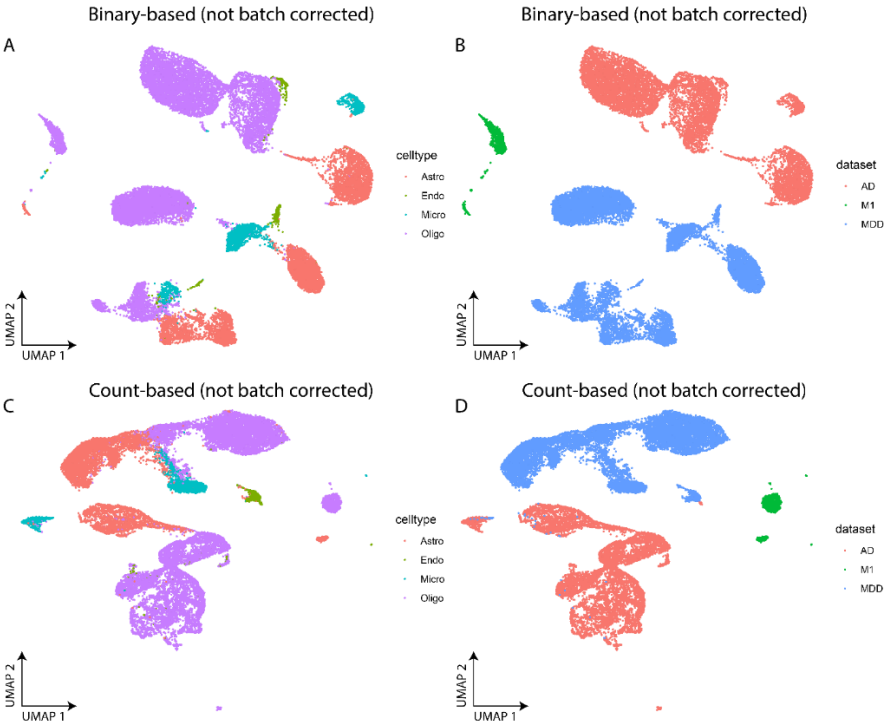

**Supplementary Figure 4:** Comparison of binary-based UMAPs on AD Dataset, all points are colored based on pre-annotated cell types. **A)** UMAP plot based on the ten components from scBFA method. **B)** UMAP plot based on the ten components from the Jaccard similarity eigenvectors. **C)** UMAP plot based on the ten components from binary-based PCA. **D)** UMAP plot based on the ten components from count-based PCA.



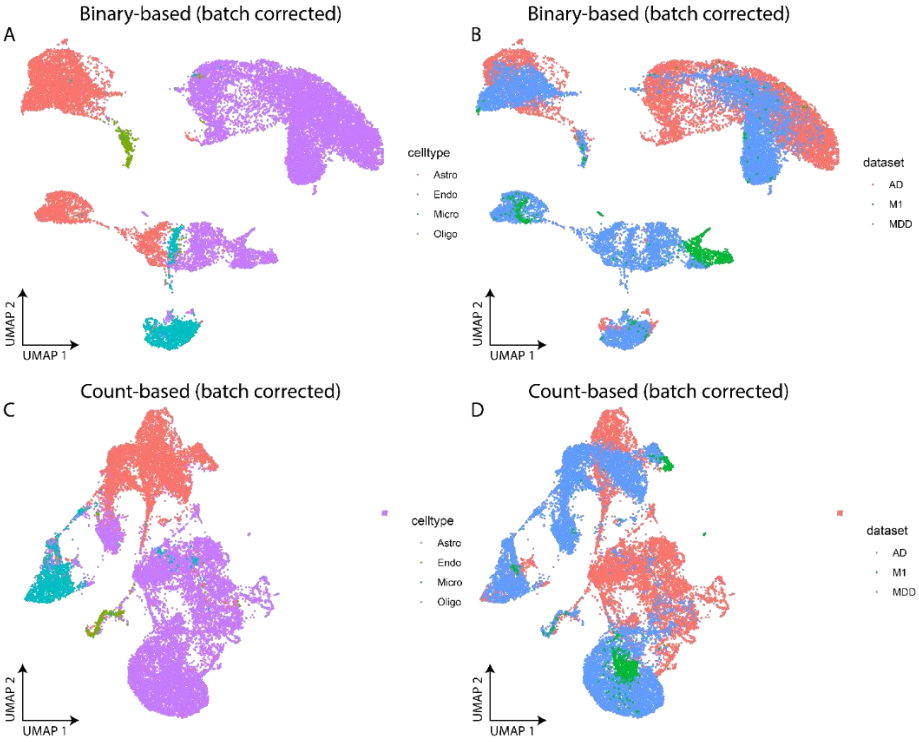

**Supplementary Figure 5:** Association of pairwise eucledian distances between cells from count based UMAP with **A)** binary-PCA based UMAP, **B)** JaccardEV based UMAP and **C)** scBFA based UMAP. First, 5,000 cells were randomly sampled, between which the pairwise eucledian distance was calculated based on the different UMAPs. Based on these pair-wise distances (n = 12,497,500) the pearson correlation was calculated. For plotting 10,000 points were randomly sampled from total number of calculated pair-wise distances.



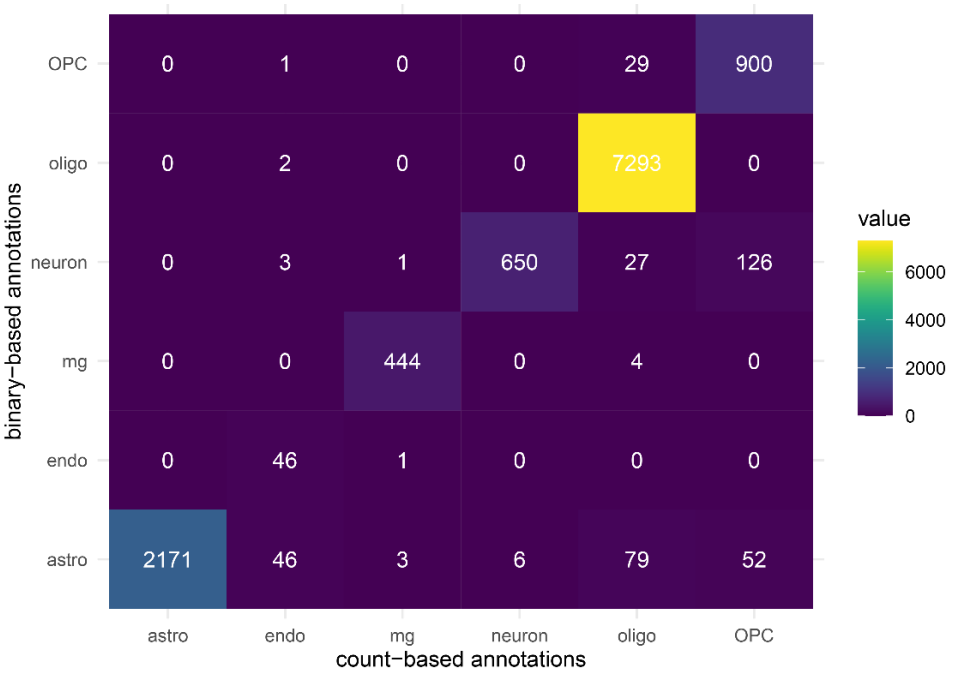

**Supplementary Figure 6:** Silhouette scores of count- and binary-based dimensionality reduction. Silhouette scores were calculated with the reduced dimensions and cell types as clusters. CountPCA and BinaryPCA are PCs obtainend with counts and binarized counts repsecitvely. JEVs are Jaccard eigen values and scBFA were components obtained using binary data and scBFA. The last column represents the average of the whole dataset.



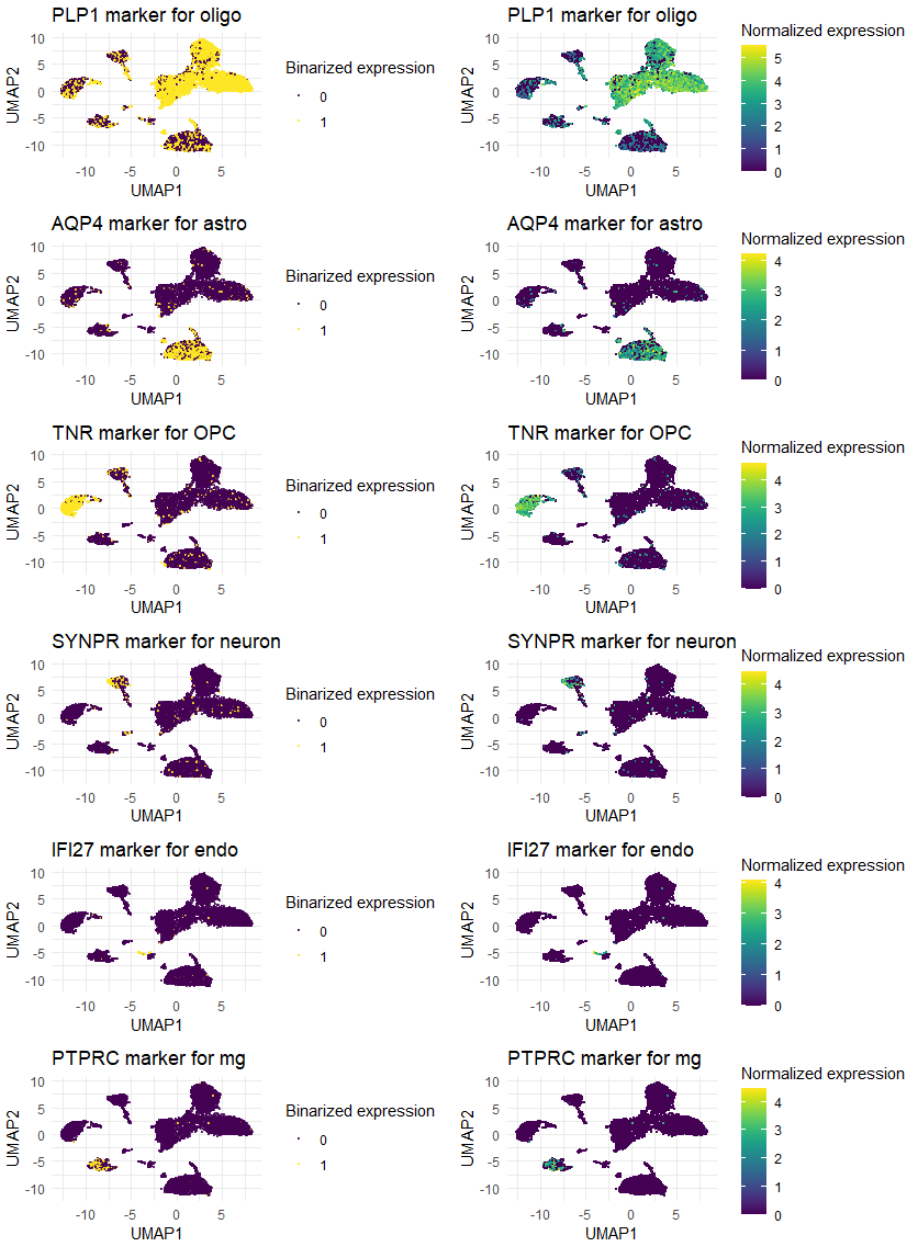

**Supplementary Figure 7:** UMAP plots of three brain datasets. **A)** UMAP plot of three brain datasets where the dataset representation was binary, colors indicate cell type. **B)** UMAP plot of three brain datasets where the dataset representation was binary, colors indicate dataset. **C)** UMAP plot of three brain datasets where the dataset representation was log normalized counts, colors indicate cell type. **D)** UMAP plot of three brain datasets where the dataset representation was log normalized counts, colors indicate dataset.



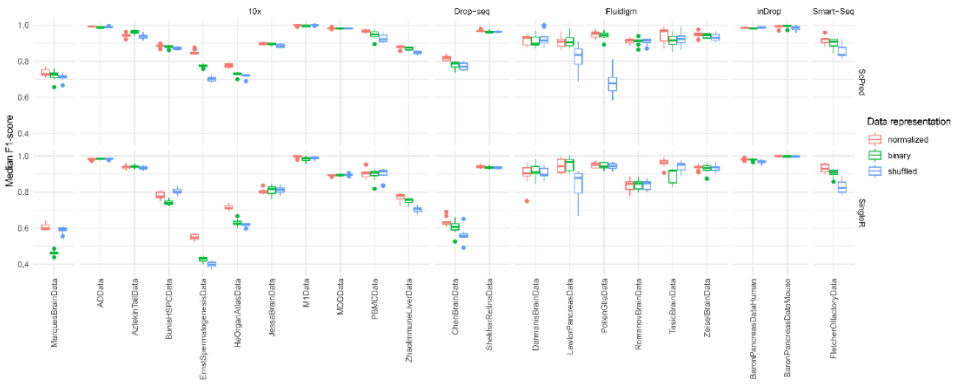

**Supplementary Figure 8 :** UMAP plots of three brain datasets, batch corrected for datasets using Harmony. **A)** UMAP plot of three brain datasets where the dataset representation was binary, colors indicate cell type. **B)** UMAP plot of three brain datasets where the dataset representation was binary, colors indicate dataset. **C)** UMAP plot of three brain datasets where the dataset representation was log normalized counts, colors indicate cell type. **D)** UMAP plot of three brain datasets where the dataset representation was log normalized counts, colors indicate dataset.



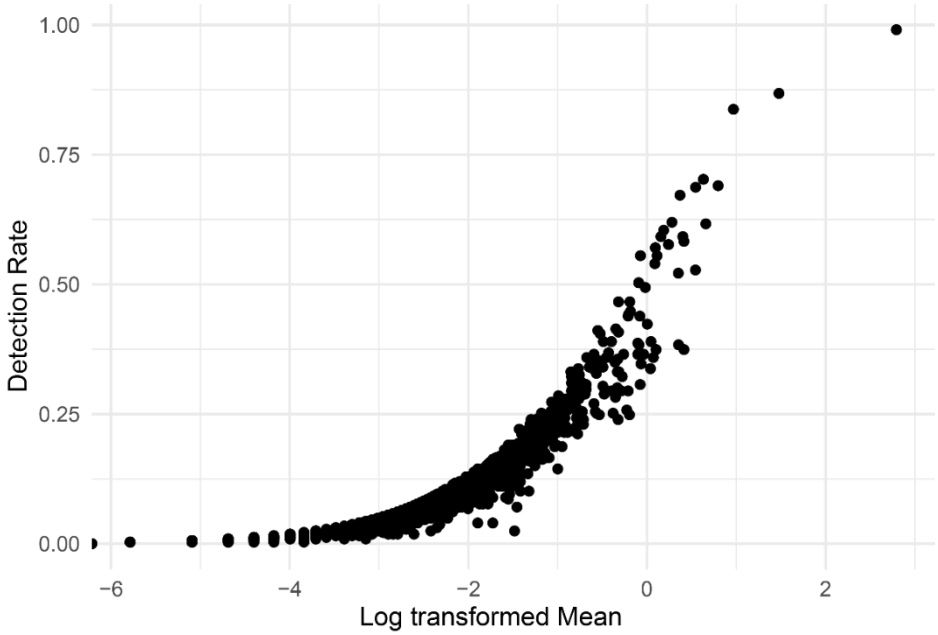

**Supplementary Figure 9:** Heatmap of concordance between binary-based cell type annotations using markers and counts-based cell type annotations using markers.



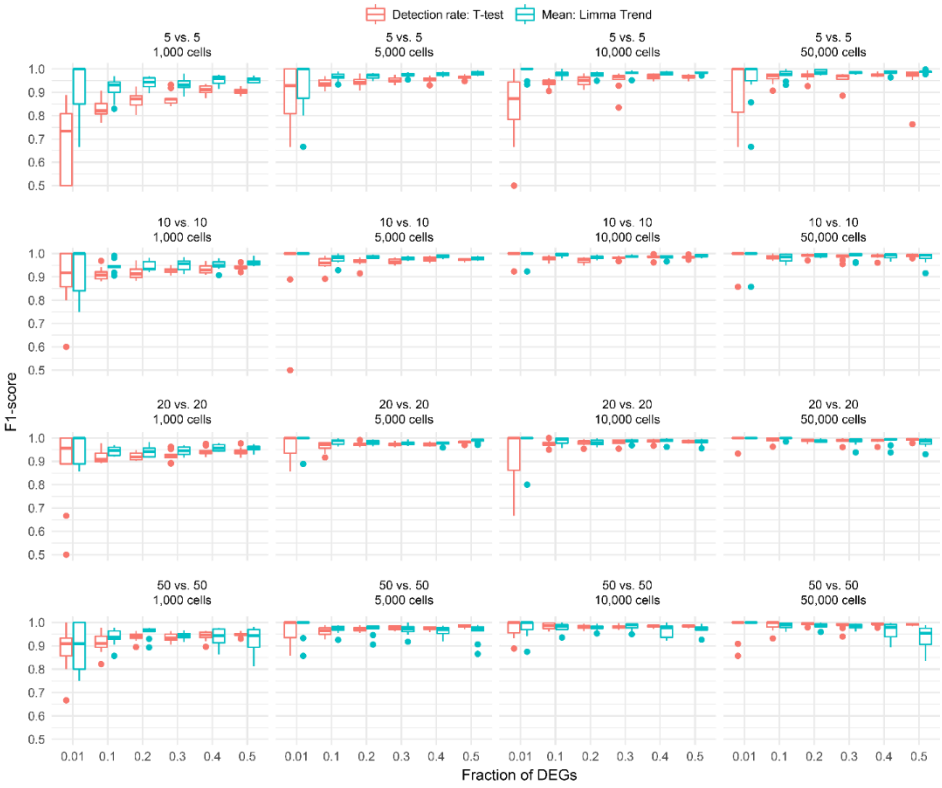

**Supplementary Figure 10:** UMAP plot of the AD Dataset with expressions of marker genes, using binarized and normalized representations.




**Supplementary Figure 11:** Boxplots of the median F1-score of the automatic cell type prediction with different data representations. Both methods (scPred, SingleR) were applied 10 times on each dataset with different reference/target splits. The datasets are represented on the x-axis and the y-axis are median F1-scores.

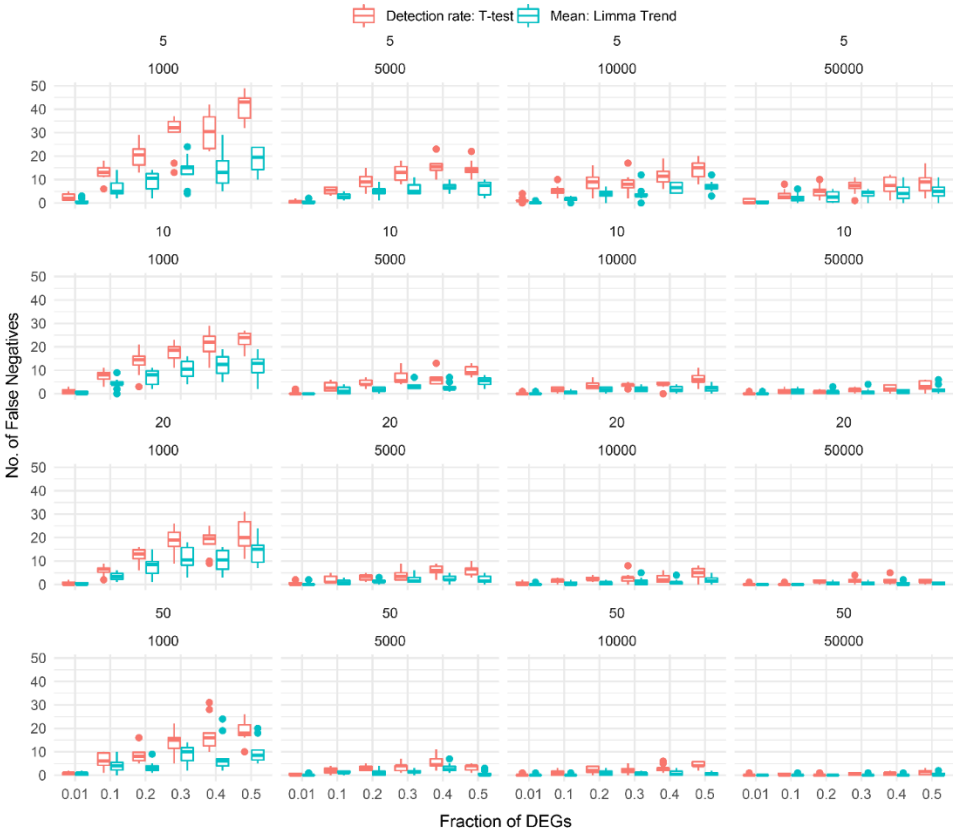



**Supplementary Figure 12:** Scatterplot of detection rate (y-axis) vs mean expression(x-axis) for all genes (n = 30.062) of one individual. The Spearman's rank correlation (across all genes) was ≥ 0.99 for all individuals. Note, spearman's rank correlation was used as this association between detection rate and log transformed mean is known to be non-linear, but their ranks are linearly correlated.

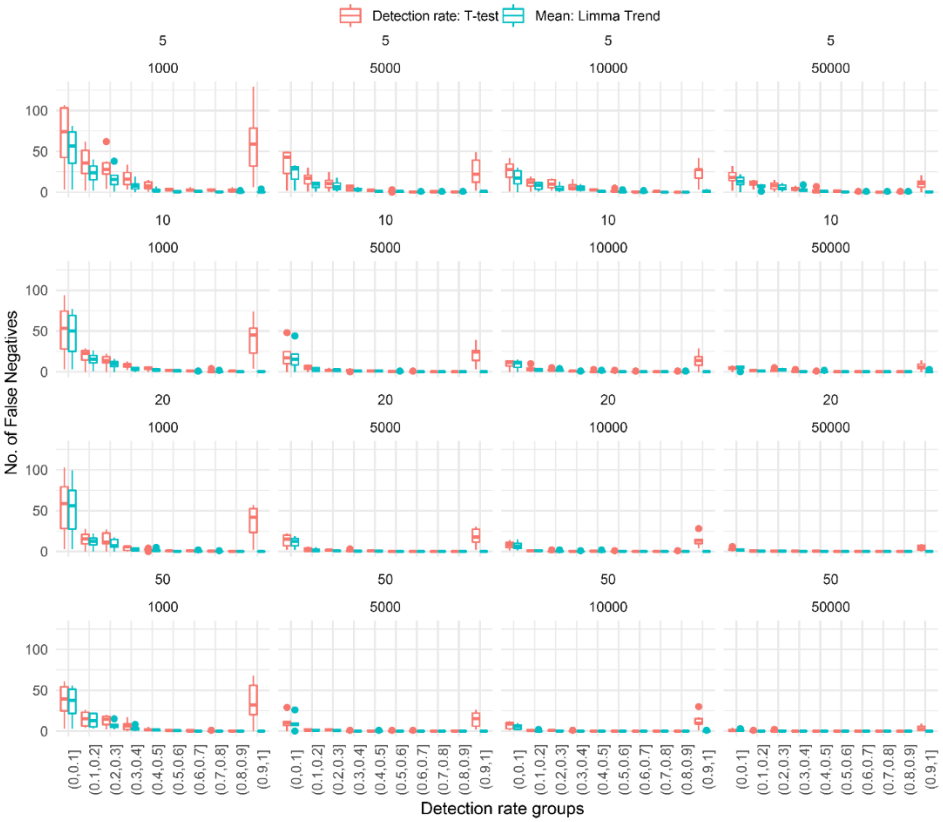



**Supplementary Figure 13:** F1-score of 960 simulated datasets indicating the accuracy of detecting differentially expressed genes in simulated pseudobulk data when either count or binarized data are used. The x-axis represents the fraction of simulated differentially expressed genes. The y-axis represents the F1-score. The top-left panel represents a comparison of 5 vs. 5 samples in a simulated dataset of 1.000 cells, meaning that each sample was comprised of 100 cells. E.g in the bottom left panel each sample was comprised of 10 cells.

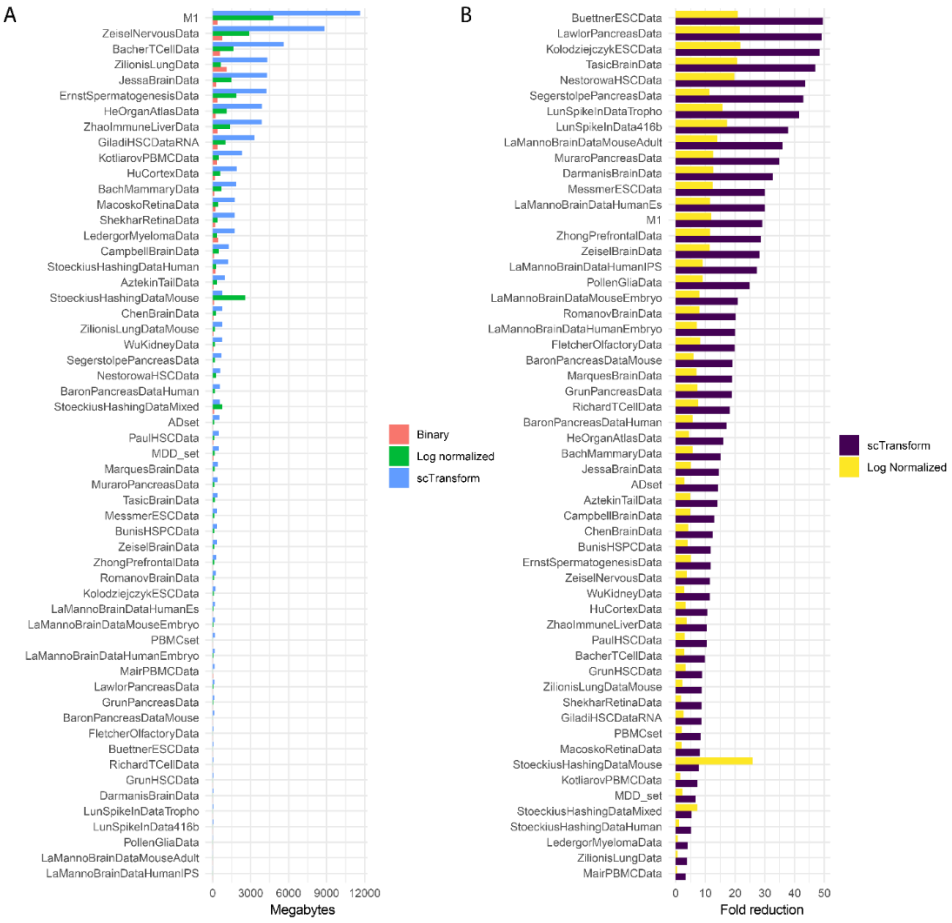

**Supplementary Figure 14:** FPR of 960 simulated datasets indicating the accuracy of detecting differentially expressed genes in simulated pseudobulk data when either count or binarized data are used. The x-axis represents the fraction of simulated differentially expressed genes. The y-axis represents the FPR. The top-left panel represents a comparison of 5 vs. 5 samples in a simulated dataset of 1.000 cells, meaning that each sample was comprised of 100 cells. E.g in the bottom left panel each sample was comprised of 10 cells.



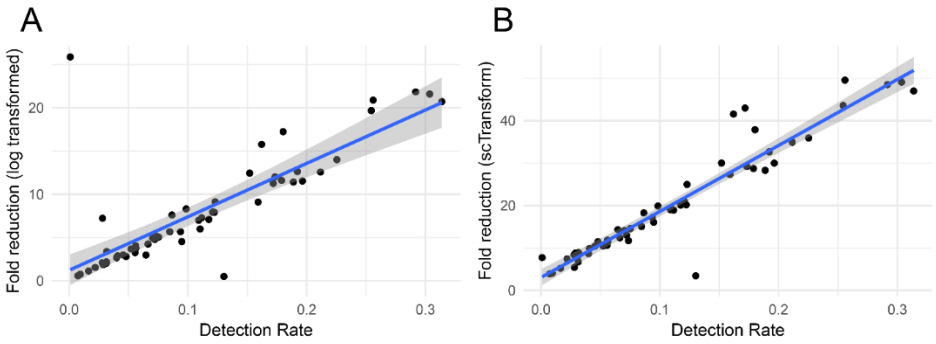

**Supplementary Figure 15:** FNR of 960 simulated datasets indicating the accuracy of detecting differentially expressed genes in simulated pseudobulk data when either count or binarized data are used. The x-axis represents the fraction of simulated differentially expressed genes. The y-axis represents the FNR. The top-left panel represents a comparison of 5 vs. 5 samples in a simulated dataset of 1.000 cells, meaning that each sample was comprised of 100 cells. E.g in the bottom left panel each sample was comprised of 10 cells.



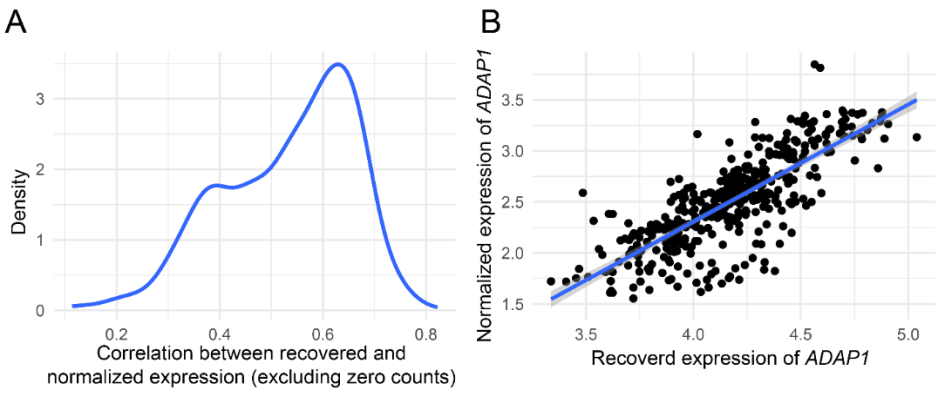

**Supplementary Figure 16:** Number of false negatives of 960 simulated datasets indicating how well simulated differentially expressed genes in pseudobulk data can be found back when either count data is used or binarized data. The x-axis represent the detection rate groups. E.g. simulated genes with a detection rate between 0 and 0.1 belong the first group (0,0.1]. The y-axis represent the number of false negatives.




**Supplementary Figure 17:** Storage requirements for the different data representations. **A)** For each dataset (y-axis) the required storage required in megabytes (x-axis). **B)** Fold reduction(x-axis) for all datasets(y-axis). Fold reduction of bit-stored relative to scTransform is purple. Fold reduction of bit-stored relative to log normalized is yellow.



**Supplementary Figure 18:** Association of detection rate with fold reduction. Scatter plost where each dot is a dataset, the x-axis represents the detection rate and the y-axis is the fold reduction of bit-stored relative to **A)** log normalized and **B)** scTransform.




**Supplementary Figure 19: A)** Density plot of the correlation coeffiecient between recovered expression values and normalized expression values of the non-zero counts. **B)** Scatter plot showing the recovered expression of *ADAP1* (x-axis), and the normalized expession values of *ADAP1* (y-axis) from the AD dataset. All zero counts are excluded, as these artifically inflate the correlation coefficient.

# Discussion

# 9.1. General discussion

While we have become adept at identifying genetic risk variants associated with various diseases, understanding these variants remains challenging. As GWASs (genome-wide association studies) continue to grow in sample size, they continue to identify more risk variants, including those with smaller effect sizes. However, for most genetic risk variants, we still do not know how they modulate disease risk[1,2]. In this thesis, we present two computational approaches (*geneset-QTLs, gsQTL,* and *differential-correlation-QTLs, dcQTL*) aimed at providing additional context to the consequences of genetic risk variants.

Furthermore, technological advancements have made it feasible to generate population-scale multi-condition single cell transcriptome (scRNAseq) datasets. However, effective methods for analysing these large scale datasets are still limited. In this work, we present cell-projected phenotypes that exploit the between-individual phenotypic variation to characterize within-individual cellular variation. Additionally, with the introduction of such datasets, the number of cells is often prioritized over sequencing depth[3,4], resulting in increased sparsity. To address this issue, we advocate the use of a binarized representation of gene expression in this thesis to handle the sparser nature of these datasets.

# 9.2. Binarized single-cell RNAseq data

## 9.2.1 Loss of information when binarizing single-cell RNAseq data

As reported in **chapter 8**, zero counts are the most abundant observation in most scRNAseq datasets, followed by single counts (i.e. count equal to 1). In many cases, the percentage of observations exceeding 1 is below 5% (**Fig. 1**). Consequently, binarization preserves much of the overall signal, because only a small fraction of data points (those counts >1) is mapped to one value. In other words, the main “loss-of-information” from binarization arising from collapsing counts that are greater than zero into a single category (i.e., 1) is restricted to about 5% of the data. We have seen in **chapter 8** that this has nearly no consequences on any of the downstream tasks. This might be understood by that a single cell is never measured in isolation, i.e. there are many cells that share similar profiles that in the downstream analysis serve as pseudo-replicates for estimating a gene’s expression level. As we have shown that the detection rate of a gene (the fraction of cells in which it is expressed) provides a robust proxy for its average expression across these pseudo-replicates, focusing on the detection rate effectively captures the signal across all expression levels in most scRNAseq datasets, thus even when counts above 1 are collapsed into a single category.

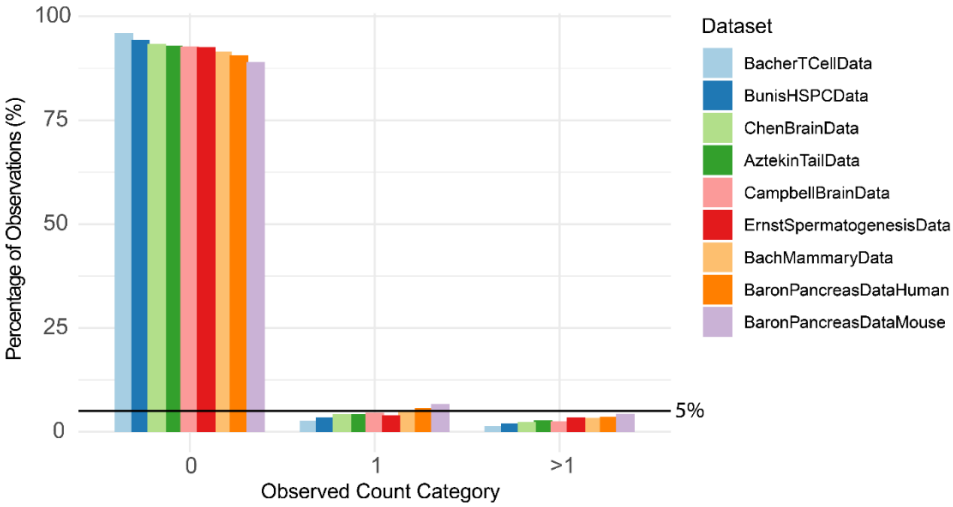



**Figure 1:** Distribution of scRNAseq read counts across multiple datasets. The x-axis shows the observed count category: 0, 1, or >1. The y-axis indicates the percentage of total observations within each dataset. Bars are grouped by dataset.

## 9.2.2 Future of binarized single-cell RNAseq data

In **chapters 7** and **8**, we provide an overview of the various analytical tasks that can be performed using binarized scRNAseq data, hoping to inform tool developers about the potential of binarized scRNAseq. Since then, additional methods utilizing binarized scRNAseq data have emerged, including: **1)** Multi-omics integration[5], **2)** quantum gene regulatory networks[6], **3)** clustering of spatially resolved transcriptomics data[7], and **4)** cell state identification[8]. The feasibility of using binarized scRNAseq data primarily depends on the capture rates of sequencing protocols. Currently, the efficiency of capturing, converting, and amplifying poly-adenylated mRNA is estimated to be between 10 and 40%[9]. Significant technological advancements are needed to improve the efficiency. At which point exactly using binarized scRNAseq data may no longer be feasible remains to be investigated.

In **chapter 8**, we emphasize the importance of developing specialized bit-aware implementations of scRNAseq methods. These implementations have the potential to greatly reduce the computational resources needed for methods and analyses, including deep learning methods[10]. This reduction is crucial for handling increasingly large datasets. For example, when analysing datasets with over 100 million cells, the computational demands of current methods will likely exceed the capabilities of many existing systems. Bit-aware implementations offer a viable solution to this problem by optimizing resource usage and enabling the analysis of these massive datasets.

In summary, the sparsity of scRNAseq data is likely a persistent challenge. As the number of cells being analysed continues to grow, the computational demands increase accordingly. Binarization of scRNAseq data, particularly through specialized bit-aware implementations, may offer a viable solution to meet these increasing demands.

## 9.3. Unveiling disease heterogeneity within and between individuals

In **Chapter 6** we introduce *cell projected phenotypes*, where we show that each cell within an individual can be transcriptionally affected by conditions like Alzheimer's to varying degrees. We extended this to entire cell types, demonstrating that people with similar phenotypes might exhibit a diverse range of affected cell types. For instance, some individuals might experience slight effects across all brain cell types due to AD, while others may show impacts only in specific cell types, like microglia, yet display similar overall phenotypes.

Differential correlation analyses, as discussed in **Chapters 2** and **4**, are particularly well-suited for a similar personalized approach. In these chapters, we conduct differential correlation analyses to compare correlations across different contexts—for example, healthy versus diseased states (**Chapter 2**) and individuals with protective versus risk genetic variants (**Chapter 4**). Notably, the discovery that genetic variants are linked to differential correlations suggests that each person possesses a unique network of functionally related biomolecules, a nuance often overlooked by current methods. scRNA-seq data would be highly valuable in this context. Creating such personalized, cell type-specific correlation networks requires a substantial number of deeply sequenced cells to confidently calculate correlations for a specific cell type within an individual, underscoring the importance of technological advancements.

In summary, these individualized, cell type-specific approaches have already enhanced, and will continue to enhance our understanding of disease heterogeneity, emphasizing the necessity for strategies that consider each person's unique genetic, molecular, and cellular landscapes.

## 9.4. Somatic variant profiles of single-cells

In **chapter 5**, we identified somatic mutations specific to excitatory neurons by calling variants from reads obtained from a scRNAseq experiment and contrasting them with individual genomes constructed using whole-genome sequencing data. The ability to confidently call variants from scRNAseq depends on the expression level of the respective transcripts and the number of cells measured for a cell type per individual, as this is the determining factor of the coverage. Consequently, we were limited to calling variants from the most abundant cell type: excitatory neurons. Ideally, this analysis would also be performed for other cell types to investigate differences in somatic mutation "hotspots" between cell types.

Additionally, in this work, the reads were prepared using Chromium Single Cell 3′ Reagent Kits, which allow for a very limited range of the genome to call variants for. This limitation could be addressed by using full read-length methods[11] or newly developed techniques such as vasa-seq[12]. However, a limitation persists: since the biomolecules we measure are mRNAs, any nucleotide differences from the germline sequence could result from either transcription errors or somatic mutations. Therefore, a more ideal approach would be the use of single-cell DNA measurements[13].

In **chapter 5** we find associations between age and the accumulation of somatic mutations, which is thought to be caused by oxidative stress and DNA repair inefficiencies[14]. Especially in the context of neurodegeneration, it would be interesting to first investigate how these somatic mutations arise and how they contribute to the disease. For instance, does everyone with Alzheimer disease have somatic mutations that contribute to the disease, or is this only true for a subset of individuals? Could this indicate a specific subtype of Alzheimer?

## 9.5. Interpretation and prioritization of genetic risk variants

In **chapter 2**, **3** and **4** we provide new approaches to identify and investigate putative downstream effects of genetic risk variants.

In **chapter 2**, we prioritize genes by the number of changes in their context-specific correlations, positing that significant shifts in correlations from a neutral to a diseased context indicate the importance of a gene in the disease. Specifically, we focus on genes within 1Mb of genetic risk variants that exhibit the most changes in correlations. This method allows for the investigation of gene associations in the diseased context using gene set enrichment analysis to identify potentially disrupted processes. In **chapter 5**, our focus shifts to proteins. We begin with a univariate protein-QTL analysis to identify associations between individual variants and proteins, followed by a differential correlation analysis to examine context-specific correlations between protein pairs in individuals with protective or risk variants. This approach reveals that a single variant can influence the associations between multiple proteins, often highlighting a central protein significant to the risk variant, thus prioritizing genes or proteins beyond those within 1Mb of the variant. In **chapter 3**, we introduce gene set-QTLs (gsQTLs) to directly link individual risk variants to gene sets. By reducing the genes in a gene set to a single vector representing their shared variance, we associate these vectors with genetic risk variants, thereby establishing a direct connection between the variant and the gene set. This approach offers a more targeted analysis compared to traditional methods requiring multiple risk variants linked to multiple genes.

These three approaches share the common goal of providing additional context to the putative downstream consequences of genetic risk variants. In a biological system, genes and proteins never act alone; there are complex interactions among various biomolecules. Knowing that only a single gene is behaving differently reveals very little about the broader changes occurring. While the approaches presented here may be more challenging to interpret (such as understanding the mechanistic pathways that result in an entire gene set having "reduced activity" associated with a genetic risk variant), we believe they complement standard QTL analyses rather than replace them. These methods should be used in conjunction to provide a more comprehensive understanding of the biological implications of genetic risk variants.

Current efforts on the interpretation and prioritization of genetic risk variants are increasingly utilizing deep learning. According to the central dogma of molecular

biology, all information is stored in DNA. As such, deep learning models are developed that use parts of the DNA sequence, typically a window around the TSS of a gene, to predict gene expression[15]. By altering the input sequence and observing the effect on gene expression (in silico mutagenesis), variants with the largest predicted impact can be prioritized. A challenge in this approach is determining the appropriate size of the receptive field or context to examine. Current state-of-the-art models include a context of up to 100kb. However, biological examples show that variants on one chromosome can affect genes on another, indicating the need for even broader contexts. This might extend beyond DNA to include interactions with other genes and biomolecules, as well as environmental and disease factors, all of which significantly influence gene expression.

## 9.6. Cell projected phenotypes for interpretation and prioritization of genetic risk variants

With *cell projected phenotypes*, we calculate a disease manifestation score for various cell types across individuals. An interesting approach would be to identify associations between these manifestation scores and SNPs. For example, a specific SNP might be linked to individuals with higher Inhibitory-neuron AD scores, suggesting a greater vulnerability to AD in their inhibitory neurons. Which would provide additional context to the downstream consequences of genetic risk variants.

Another application of cell projected phenotypes for interpretation and prioritization of genetic risk variants is to adapt the method for calculating cell projected genotypes, similar as previously done by Rumker et al[16]. Instead of associating cell neighbourhoods with phenotypes like AD diagnosis or amyloid-beta load, this method would test for the enrichment of certain genotypes. This approach enables the identification of genotype-associated cell states, providing additional and valuable context to the consequences of genetic risk variants.

## 9.7. Future of cell projected phenotypes

In **chapter 6**, we introduce cell-projected phenotypes, an approach that considers within-individual cellular heterogeneity and quantifies the degree of transcriptional association of a single cell with the phenotypic characteristics of the individual. This approach determines whether a cell is affected by the individual's phenotypic characteristic or remains unaffected. We also extended this to entire cell types, demonstrating that different individuals are affected by AD in different cell types to varying degrees. For instance, we found that cognitive impairment is particularly correlated with astrocytes and oligodendrocytes being affected, while affected microglia are associated with increased amyloid-beta load. This suggests cell type-specific components of AD and potentially indicates the existence of subtypes.

Since **Chapter 6** presents a computational approach, we currently lack empirical confirmation that neurons assigned high tau tangle scores are indeed more affected by tau pathology. To validate our computational predictions, one strategy would be to acquire spatial data from the same donors that have high cell-specific

tau tangle scores. By projecting these scores onto spatially resolved cells (similar to the methodology employed in SPAGE[17]), we can assess the maturation stages[18] of neurofibrillary tangle development within these cells, ranging from pre-tangle neurons to just before dying and leaving behind ghost tangles. If we observe that our computed score increases in tandem with the maturation of tangle development, this would provide biological validation for our approach.

An interesting use case for cell-projected phenotypes is the identification of disease-specific archetypical cell states. While significant effort is currently directed towards creating reference atlases of cell types[19], a reference atlas could also be developed to include disease-specific archetypical cell states. Such an atlas could be invaluable for disease prediction or diagnostics, particularly if disease-specific archetypical cell states are present in peripheral blood mononuclear cells (PBMCs), as this would offer a non-invasive method for diagnosing many diseases.

## 11.8. Concluding remarks

In this thesis, we have addressed one of the main challenges in scRNAseq analyses: its sparsity. Through our work, we aimed to demonstrate the utility of binarizing scRNAseq data and inspire researchers to incorporate this approach into their methods. We have explored novel ways to investigate genetic risk variants. The key takeaway is that in a biological system, genes and proteins never act alone; there are complex interactions between them. To truly understand genetic consequences, we should therefore incorporate many different contexts. Additionally, we have developed a novel framework for analysing scRNAseq data that leverages between-individual phenotypic heterogeneity to gain a better understanding of within-individual cellular dynamics, showing that individuals with similar (known) phenotypic characteristics can have very distinct cell type manifestation profiles, underscoring the importance of the contexts considered.

## References

1. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From Association to Function. *Am. J. Hum. Genet.* **102**, 717–730 (2018).

2. Pierce, S. E. *et al.* Post-GWAS knowledge gap: the how, where, and when. *npj Park. Dis. 2020 61* **6**, 1–5 (2020).

3. Zhang, M. J., Ntranos, V. & Tse, D. Determining sequencing depth in a single-cell RNA-seq experiment. *Nat. Commun. 2020 111* **11**, 1–11 (2020).

4. Schmid, K. T. *et al.* scPower accelerates and optimizes the design of multi-sample single cell transcriptomic studies. *Nat. Commun. 2021 121* **12**, 1–18 (2021).

5. Misra, R., Ferrena, A. & Zheng, D. Facilitate integrated analysis of single cell multiomic data by binarizing gene expression values. *bioRxiv* 2024.02.22.581665 (2024) doi:10.1101/2024.02.22.581665.

6. Roman-Vicharra, C. & Cai, J. J. Quantum gene regulatory networks. *npj Quantum Inf. 2023 91* **9**, 1–8 (2023).

7. Lin, S. *et al.* Complete spatially resolved gene expression is not necessary for identifying spatial domains. *Cell Genomics* **4**, (2024).

8. Jansma, A. *et al.* High order expression dependencies finely resolve cryptic states and subtypes in single cell data. *bioRxiv* 2023.12.18.572232 (2023) doi:10.1101/2023.12.18.572232.

9. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* **9**, 1–12 (2017).

10. Dürichen, R., Rocznik, T., Renz, O. & Peters, C. Binary Input Layer: Training of CNN models with binary input data.

11. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc. 2013 91* **9**, 171–181 (2014).

12. Salmen, F. *et al.* High-throughput total RNA sequencing in single cells using VASA-seq. *Nat. Biotechnol. 2022 4012* **40**, 1780–1793 (2022).

13. Luquette, L. J. *et al.* Single-cell genome sequencing of human neurons identifies somatic point mutation and indel enrichment in regulatory elements. *Nat. Genet. 2022 5410* **54**, 1564–1571 (2022).

14. Lodato, M. A. *et al.* Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science (80-. ).* **359**, 555–559 (2018).

15. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods 2021 1810* **18**, 1196–1203 (2021).

16. Rumker, L. *et al.* Identifying genetic variants that influence the abundance of cell states in single-cell data. *bioRxiv* 2023.11.13.566919 (2023) doi:10.1101/2023.11.13.566919.

17. Abdelaal, T., Mourragui, S., Mahfouz, A. & Reinders, M. J. T. SpaGE: Spatial Gene Enhancement using scRNA-seq. *Nucleic Acids Res.* **48**, e107–e107 (2020).

18. Moloney, C. M., Lowe, V. J. & Murray, M. E. Visualization of neurofibrillary tangle maturity in Alzheimer's disease: A clinicopathologic perspective for biomarker research. *Alzheimer's Dement.* **17**, 1554–1574 (2021).

19. Jones, R. C. *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, (2022).

# Acknowledgements

I can wholeheartedly say that I enjoyed these past four years!

Marcel and Ahmed, Thank you for an incredible four years! With you as my supervisors, I always felt that my personal scientific development was the priority above anything else, and I truly appreciate that! Leen and Roderick, thank you for inspiring me to pursue a career in science and for your guidance and inspiration! Even now, when I encounter challenges in my projects, I often think, 'What would Roderick say? Manolis, thank you for your enthusiasm and for reigniting my passion for science! Marunka, thank you for your invaluable support surrounding administrative matters! Ruud thank you for your technical assistance!

Nicco, thank you for your guidance during my Master's thesis project and for your valuable input in the projects that followed! And now, we're direct colleagues again in Amsterdam, with many more projects to come! Jasper, I often pride myself on being early in the office, but you're usually already there with the lights off, so I only realize you're around once I reach the door, thereby making it the first surprise of my day whether you're there or not. I appreciate your down-to-earth perspective on most things, while also admiring the random topics you wholeheartedly like to discuss. Even though we tackle challenges in completely different ways, we still manage to understand one another, and I truly value that connection. Bram, Late into my own PhD journey, you appeared, but soon proved yourself as a friend I deeply value. Bound by our shared passion for research, we quickly became friends. Open and honest in every conversation, your transparency keeps inspiring me to embrace authenticity. Laughter is something we never lack, especially when even "that" word makes its inevitable appearance in your banter. Friendship with you feels easy and natural. Sincerely, I am grateful for your friendship, which has added a lightness and depth to my final steps as a PhD candidate. Timo, you were the first Master's student I supervised, and it was great to witness your academic growth from that stage all the way into your PhD. I enjoyed guiding your research, and later on, sharing an office with you. I'll never forget that one morning you overslept when I was supposed to pick you up for a conference, thanks for keeping things entertaining! I wish you all the best, and I look forward to seeing all the great things you accomplish. Swier, I often find myself drifting into your office whenever I'm a bit bored, ready to chat about anything and everything and occasionally something work-related. As for life after your PhD, I'm genuinely curious to see where you go, maybe hand-puppet making with Mr. Lucky? Rickard, I look back on my time in Boston with pleasure, in part due to the fun activities we did together, like bouldering and visiting a Boston Red Sox game!

To everyone in the bioinformatics group, thank you for creating such an awesome work environment. I am truly grateful to have been part of such a fantastic group!

To everyone in the office in Leiden, thank you not only for creating such an amazing work environment but also for challenging me with difficult biological questions!

Papa, Mama, dank jullie wel voor jullie onvoorwaardelijke steun. En dank jullie wel dat jullie, zelfs nadat ik met twee opleidingen vroegtijdig was gestopt, toch in mij bleven geloven en mij ook nog steunden toen ik bio-informatica wilde studeren. Jaimy, Bep, Claudia, dank jullie wel voor jullie onvoorwaardelijke steun en voor het feit dat jullie altijd voor me klaarstaan. Carmen, de laatste twee jaar van mijn PhD leken ineens een stuk makkelijker te gaan. Hoe zou dat toch komen?

# Curriculum Vitae

## Education

2020-2025 **Ph.D. Bioinformatics**

Delft University of Technology, Delft, The Netherlands

*Promotors:* Prof. dr. ir. M.J.T. Reinders

Dr. A. Mahfouz

2019-2020 **Master of Science in Computer Science (track: Bioinformatics)**

Leiden University, Leiden, The Netherlands

2015-2018 **Bachelor of Science in Bioinformatics**

Leiden University of Applied Sciences, Leiden, The Netherlands

## Experience

2024-present **Amsterdam UMC, Amsterdam, The Netherlands**

Postdoctoral Researcher, Department of Human Genetics

2023-present **Massachusetts Institute of Technology, Cambridge, USA**

Visiting Researcher, Computational Biology Group, Computer Science and Artificial Intelligence Lab.

2020-2024 **Delft University of Technology, Delft, The Netherlands**

Ph.D. Candidate, The Delft Bioinformatics Lab

2020-2024 **Leiden University Medical Center, Leiden, The Netherlands**

Visiting Ph.D. Candidate, Department of Human Genetics

| | |
|---|---|
| 2018-2020 | **Leiden University Medical Center, Leiden, The Netherlands**<br>Senior Analyst, Department of Cell and Chemical Biology |
| 2016-2018 | **PricewaterhouseCoopers, Amsterdam, The Netherlands**<br>Working Student Tax Technology and Data Analytics |

# List of Publications

## In This Thesis

1. **Bouland,G.A**., Mahfouz,A. and Reinders,M.J.T. (2021) Differential analysis of binarized single-cell RNA sequencing data captures biological variation. NAR Genomics Bioinforma., 3.

2. **Bouland,G.A**., Mahfouz,A. and Reinders,M.J.T. (2023) Consequences and opportunities arising due to sparser single-cell RNA-seq datasets. Genome Biol. 2023 241, 24, 1–10.

3. Zhang,M., **Bouland,G.A**., Holstege,H. and Reinders,M.J.T. (2022) Identifying aging and Alzheimer's disease associated somatic mutations in excitatory neurons from the human frontal cortex using whole genome sequencing and single cell RNA sequencing data. medRxiv, 10.1101/2022.05.25.22275538.

4. **Bouland,G.A**., Tesi,N., Mahfouz,A. and Reinders,M.J.T. (2024) gsQTL: Associating genetic risk variants with gene sets by exploiting their shared variability. bioRxiv, 10.1101/2024.09.13.612853.

5. **Bouland,G.A**., Marinus,K.I., Kesteren,R.E. van, Smit,A.B., Mahfouz,A. and Reinders,M.J.T. (2023) Single-cell RNA sequencing data reveals rewiring of transcriptional relationships in Alzheimer's Disease associated with risk variants. medRxiv, 10.1101/2023.05.15.23289992.

## Other Publications

6. Canouil,M., **Bouland,G.A**., Bonnefond,A., Froguel,P., 't Hart,L.M. and Slieker,R.C. (2019) NACHO: an R package for quality control of NanoString nCounter data. Bioinformatics, 36, 970–971.(*shared first authors*)

7. **Bouland,G.A**., Beulens,J.W.J., Nap,J., van der Slik,A.R., Zaldumbide,A., 't Hart,L.M. and Slieker,R.C. (2020) CONQUER: an interactive toolbox to understand functional consequences of GWAS hits. NAR Genomics Bioinforma., 2.

8. **Bouland,G.A**., Beulens,J.W.J., Nap,J., van der Slik,A.R., Zaldumbide,A., 't Hart,L.M. and Slieker,R.C. (2022) Diabetes risk loci-associated pathways are shared across metabolic tissues. BMC Genomics 2022 231, 23, 1–9.

9. Slieker,R.C., Donnelly,L.A., Fitipaldi,H., **Bouland,G.A**., Giordano,G.N., Åkerlund,M., Gerl,M.J., Ahlqvist,E., Ali,A., Dragan,I., et al. (2021) Replication and cross-validation of type 2 diabetes subtypes based on clinical variables: an IMI-RHAPSODY study. Diabetologia, 64, 1982–1989.

10. Slieker,R.C., Donnelly,L.A., Fitipaldi,H., **Bouland,G.A**., Giordano,G.N., Åkerlund,M., Gerl,M.J., Ahlqvist,E., Ali,A., Dragan,I., et al. (2021) Distinct Molecular Signatures of Clinical Clusters in People With Type 2 Diabetes: An IMI-RHAPSODY Study. Diabetes, 70, 2683–2693.

11. Slieker,R.C., Donnelly,L.A., Akalestou,E., Lopez-Noriega,L., Melhem,R., Güneş,A., Abou Azar,F., Efanov,A., Georgiadou,E., Muniangi-Muhitu,H., et al. (2023) Identification of biomarkers for glycaemic deterioration in type 2 diabetes. Nat. Commun. 2023 141, 14, 1–18.

12. ’t Hart,L.M., de Klerk,J.A., **Bouland,G.A**., Peerlings,J.H.D., Blom,M.T., Cramer,S.J., Bijkerk,R., Beulens,J.W.J. and Slieker,R.C. (2024) Small RNA sequencing reveals snoRNAs and piRNA-019825 as novel players in diabetic kidney disease. Endocrine, 86, 194–203.

13. Slieker,R.C., Münch,M., Donnelly,L.A., **Bouland,G.A**., Dragan,I., Kuznetsov,D., Elders,P.J.M., Rutter,G.A., Ibberson,M., Pearson,E.R., et al. (2024) An omics-based machine learning approach to predict diabetes progression: a RHAPSODY study. Diabetologia, 67, 885–894.