





# **Sequencing**

from Blood to Brain

Jeroen van Rooij

Lay-out & printing: ProefschriftMaken || [www.proefschriftmaken.nl](http://www.proefschriftmaken.nl)

ISBN 978-94-6423-017-8

© copyright Jeroen G. J. van Rooij, 2020

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior permission of the author or the copyright-owning journals for previous published chapters.



# Sequencing

## van bloed tot brein

### Sequencing

#### from Blood to Brain

#### Proefschrift

ter verkrijging van de graad van doctor aan de  
Erasmus Universiteit Rotterdam  
op gezag van de  
rector magnificus

Prof.dr. R.C.M.E. Engels

en volgens besluit van het College voor Promoties.  
De openbare verdediging zal plaatsvinden op  
woensdag 9 december 2020 om 09:30 uur

door

Jeroen Gerardus Johannes van Rooij  
geboren te 's-Hertogenbosch

## **Promotiecommissie:**

Promotor: Prof.dr. A.G. Uitterlinden  
Prof.dr. J.C. van Swieten

Overige leden: Prof.dr. M.A. Ikram  
Prof.dr. S.A. Kushner  
Prof.dr. A.B. Smit

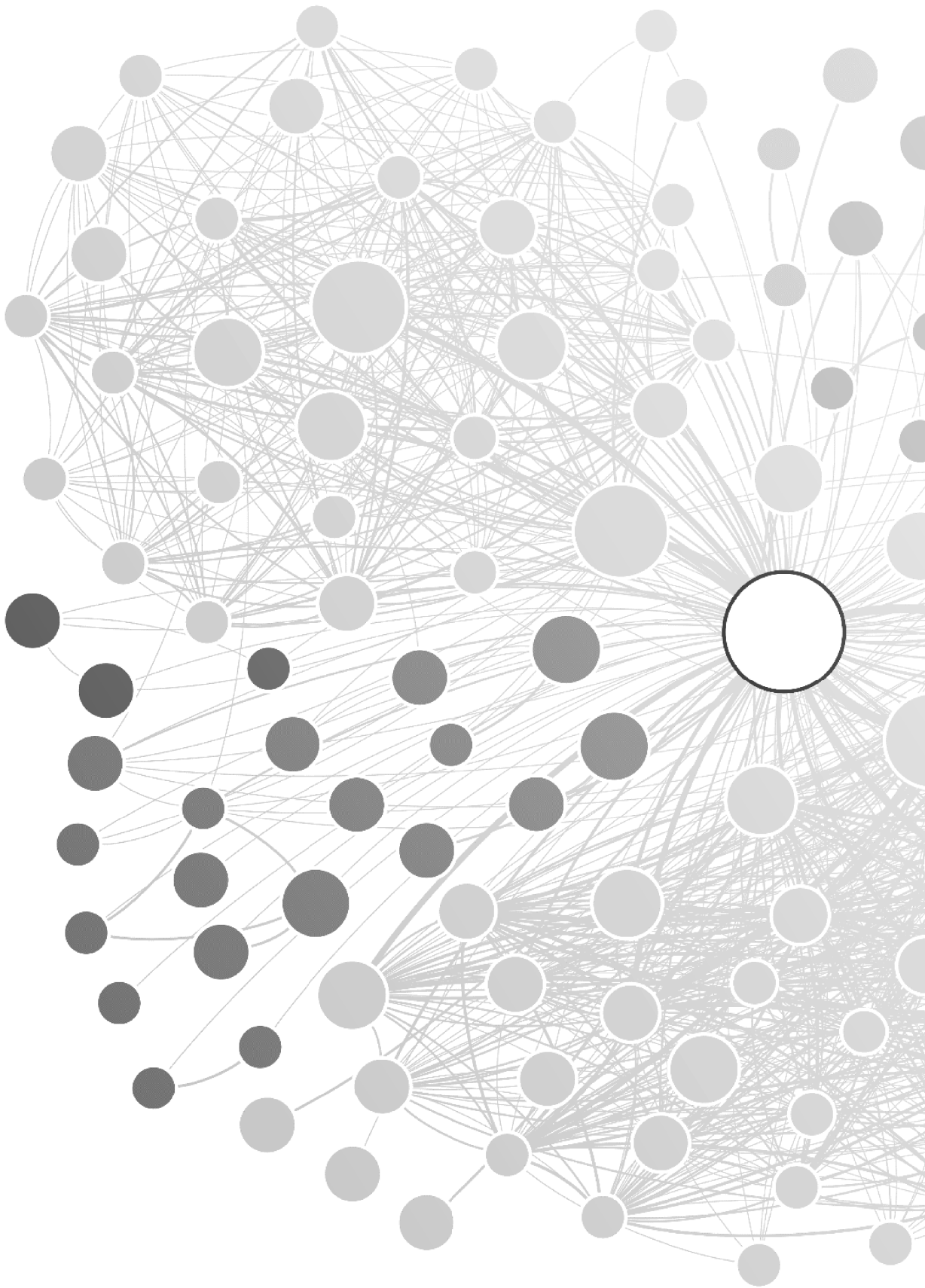
Copromotor: Dr. J.B.J. van Meurs

Paranimfen: Martin Huisman

Dennis Schmitz

## Table of contents

<b>Chapter 1</b>	General introduction	7
<b>Chapter 2</b>	Sequencing blood DNA	29
Chapter 2.1	Population-specific genetic variation in large sequencing datasets; why more data is still better	31
Chapter 2.2	Reduced penetrance of pathogenic ACMG variants in a deeply phenotyped cohort study and evaluation of clinvar classification over time	39
Chapter 2.3	<i>EIF2AK3</i> variants in Dutch patients with Alzheimer's disease	55
<b>Chapter 3</b>	Sequencing RNA blood & brain	73
Chapter 3.1	Evaluation of commonly used analysis strategies for epigenome and transcriptome-wide association studies through replication of large-scale population studies	75
Chapter 3.2	Hippocampal transcriptome profiling combined with protein-protein interaction analysis elucidates Alzheimer's Disease pathways and genes	99
<b>Chapter 4</b>	Sequencing brain DNA	133
Chapter 4.1	Somatic <i>TARDBP</i> variants as cause of Semantic Dementia	135
<b>Chapter 5</b>	General discussion	163
<b>Chapter 6</b>	Appendices	183
6.1.	Summary	185
6.2.	About the author	189
6.3.	Portfolio	191
6.4.	List of Publications	195
6.5.	Dankwoord	205
6.6.	Abbreviations	209





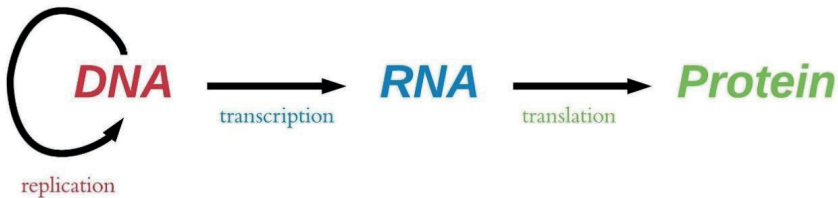
# Chapter 1

## General introduction



## The Central Dogma of Biology

The human genome (DNA) consists of 3 billion building blocks of A, C, G or T (1). Each person inherits two genome copies from their parents and uses these as a blueprint for every gene and protein your cells might require to function (2, 3). Variations in the genome between individuals may affect this blueprint and thus affect how our cells function (4, 5). Some of these variations contribute to the developments of diseases, and are subject of scientific research to help us understand, prevent and treat these diseases. (6, 7). To understand how a genome variant contributes to a disease, we may investigate how our cells use their DNA, something that is described by the central dogma of biology, and illustrated below (8).



**Figure 1.** the central dogma of biology. From left to right; DNA is copied (replication) to provide copies to new cells; genes in the DNA are transcribed (transcription) into RNA when the cell calls upon the function of this gene; RNA is translated (translation) into protein, which is then available to perform whichever task in the cell that was needed.

When creating a new cell, an existing cell produces a copy of its genome by DNA replication (9). The cell then divides in two, and both cells continue with their own genome copies (10). Later, when one of these cells needs to perform a specific function, it can activate the genes needed for that function and construct the proteins required (2, 3). To do this, the cell recognizes and activates the part of the genome containing the required gene (11). It does this by removing chemical methyl molecules around that part of the genome, causing the DNA, which is normally wrapped in itself, to unravel and present the gene for processing (12, 13). This process is called DNA methylation and can be studied by measuring the presence of these methyl-groups across the genome (13). After the gene is accessible to further processing, a molecule called RNA polymerase copies the gene from the DNA to an RNA molecule, a process known as transcription or gene expression (14, 15). This process can be repeated when multiple copies of the gene are required. When the gene is copied, the DNA folds back into itself (13). This process of methylation and transcription can be repeated numerous times, and multiple genes can be transcribed at the same moment (14). Transcription can be studied by extracting and counting the RNA copies of each gene from a cell or group of cells (16). Although the transcription and methylation processes are related, they can be studied separately and each add insights to the molecular workings of diseases (17, 18). Next, the RNA molecule is transported from the nucleus of the cell to the ribosomes where it is translated into a functional protein (19). This process is called translation and is surrounded by a number of chemical changes to the RNA or protein molecule, called post-transcriptional or post-translational modifications (20, 21). These modifications allow for production of multiple forms of the protein from the same blueprint (22). Just like transcription, translation can be studied by extracting all proteins in a cell or group of cells and measuring their abundance (23, 24).

Variants in the DNA are able to influence how a cell functions by interfering with any of the processes described above (25). The most straightforward example is when a DNA variant is located within the gene blueprint. When the gene with this variant is transcribed and translated, the end-product protein is slightly different than without the variant (26). These variants are called coding variations, as they directly impact the code of a protein. Because the code of the protein directly influences its function, these variants sometimes have large influence on the proteins function, and many disease-causing DNA variants were coding variants (27). Other variants on the DNA can interfere with the regulatory processes in the central dogma, for example by changing the binding site of the RNA polymerase in the genome. Such a variant does not change the code of the protein, but can alter the amount or folded of the protein that is produced (5, 28, 29). If the variant influences post-transcriptional or post-translation modifications it can also result in alternate or incorrectly folded protein (30). Finally, this regulatory system itself is regulated by proteins in complex networks of protein-protein interaction, both within a single cell and between cells (31, 32). These networks monitor the cell's state and environment, and will signal to the cell which proteins need to be produced (33). This means that DNA variants in one gene may affect the production and function of other genes. These networks of interaction and activity are considered “dynamic” (as opposed to the more “static” genomic DNA) and are studied in the fields of genomics (methyloomics, transcriptomics, proteomics and others) (3, 5, 29, 34). Finally, these networks are influenced by external factors, for example in age, gender, environment and lifestyle, but also by diseases. Thus, genetic, methylomic, transcriptomic or proteomic changes can contribute to or cause disease, but a disease itself also influences epigenetic, transcriptomic and proteomic changes.

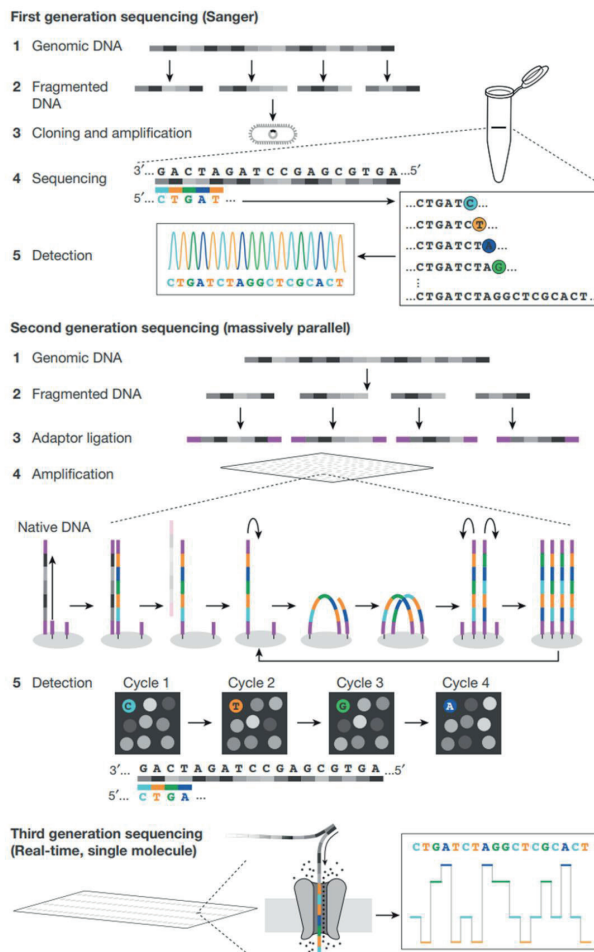
## Genomics and technology

In genomics, research developments are often driven by technological developments (35). Generally, technological improvements allow for more accurate or more simultaneous measurements, which are used to address research questions that couldn't be studied before (36). Examples are equipment, such as the microscope or the computer, but also knowledge-based developments such as biostatistics or bioinformatics, permitting the implementation of new methods (37).

The development that much increased the resolution with which we are able to look at the DNA sequence in the field of genetics was the ability to “sequence” DNA; determining the order or sequence of nucleotides in a DNA fragment. The first sequencing methods were developed around 1976; Sanger sequencing and Maxam-Gilbert sequencing (35, 38). Both methods relied on fragmentation of DNA, either by chemically cleaving fragment at specific bases (Maxam-Gilbert) or by randomly stopping DNA-replication at specific bases (Sanger). In both methods, random-length fragments are produced with a known last nucleotide. By size-separating these fragments using gel electrophoresis, they align according to size, and thus sequence, of the original fragment. This is shown for Sanger sequencing in figure 2 (39). The method progressed and around 1987, several labs were able to produce a ~1000 nucleotide sequence within a day. By sequencing multiple random DNA fragments of the same sample, and overlapping their results, larger DNA sequences could be constructed, this approach was labelled shotgun-sequencing (39). These developments sparked the



Human Genome Project, in 1990, in which large DNA fragments of the human genome were isolated and cloned into bacterial artificial chromosomes, which could be cultured to produce large amounts of purified DNA copies (1). These were then sequenced and ultimately combined to produce the first complete genome sequence in 2004 (1). During this project, nearly every step of the procedure was improved; using different labelled nucleotide terminators to allow single-tube reactions instead of four tubes per fragment; optimizing DNA amplification methods to directly produce sufficient copies of input DNA fragments without need to bacterial cloning and cultures; bead-based purification methods to clean the input DNA; capillary electrophoresis to forego the need for cast gels; as well as other steps in automation, quality control, etc (39). By 2001, several sequencing centers were able to sequence up to 10 million nucleotides per day; four orders of magnitude more than just over a decade ago.



**Figure 2.** figure adopted from publication “DNA sequencing at 40: past, present and future” by Jay Shendure et al (Nature, 19-Oct 2017; PMID 29019985). Schematic representations of first, second and third generation sequencing. One main method is shown for each generation; Sanger Sequencing, Sequencing by Synthesis and NanoPore Sequencing.

In parallel to the developments above, several groups investigated an alternative to the electrophoretic sequencing, which was considered a bottleneck in increasing the throughput of sequencing data further. This alternative was called massively parallel sequencing, which would quickly be known as next generation or second generation sequencing (39). In its most common application, adaptors are ligated to a large amount of random DNA fragments and these fragments are subsequently spread over a 2D surface spotted with fixed primers, to which the adaptors bind. This causes individual DNA fragments to be attached to a surface, allowing for millions of parallel sequencing reactions. Each DNA fragment is amplified through so-called bridge amplification, creating thousands of DNA copies, all constantly bound to surface, resulting in a “cluster” of identical copies of the original DNA fragment (39). Next, through so-called sequencing by synthesis (SBS), a single fluorescently labelled nucleotide is incorporated in each cluster. The fluorescent signal of several thousands of simultaneously incorporated nucleotides can be captured by high-density optical cameras. All reagents are then washed away, and the next nucleotide is incorporated. The camera’s record the sequence of fluorescent signal in each cluster after each cycle of incorporation. Depending on the number of cycles, longer fragments can be sequenced, currently up to ~600nt. These next-generation sequencing devices can sequence millions of DNA fragments in a single experiment, causing the cost of sequencing per nucleotide to drop by another four orders of magnitude between 2007 and 2012 (39). By 2012, most sequencers were from the Illumina company, but other alternatives still exist, usually with slight variations to the described SBS chemistry. These second-generation sequencer can sequence a complete human genome in less than a day for fewer than one millionth the cost of the original human genome sequence (38).

Currently, the third generation of sequencing is inbound. The next step main development is live single molecule sequencing, foregoing the need to stop and detect incorporated bases. This development allows for faster sequencing, and of much larger DNA fragments (39). Two main methods currently exist; PacBio sequencing, which uses individual spotted polymerases that incorporate fluorescent nucleotides, the emitted signal at each incorporation can be detected in real-time; and NanoPore sequencing, which runs a single DNA fragment through an electrified pore, detecting the change in current when each nucleotide passes (39). Through these methods, individual DNA sequence of up to 100,000 nucleotides could be determined (39). However, currently limitations are a lower amount of parallel sequencing reactions compared to second generation sequencers, and a much higher error rate (1-10%, vs < 0.1% in second-generation sequencing) (40). Expected is that when these limitations are relieved, these third-generation machines will become more common. Already now, for specific applications they have become the devices of choice, for example to sequence DNA fragments of high complexity, or for single-RNA molecule sequencing (39).

Since completion of the first whole genome sequence hundreds of thousands of genomes have been sequenced (26). Each genome deviates on approximately 20 million nucleotides from the reference sequence (0.6% of the 3.3 billion nucleotides in the reference) (41). Across all collected genomes a total of 324 million DNA variants have been identified so far (41). About 15 million of these variants are common (present in 1% or more) in the human population (41).

## Genomic studies

Genomic studies are used to research one or more of the genomic layers (genetics, methylomics, transcriptomics, proteomics). These studies can have different designs depending on the research questions that must be answered. Also, a distinction is usually made between genetic studies and studies of dynamic genomics data (methylation, expression and protein abundances). The main difference being that genetic studies can be done on DNA derived from any tissue and at any point in time, as DNA almost doesn't change with age or across tissues (2, 3, 42). In contrast, gene expression or protein abundance changes continuously and must be studied in a relevant tissue at a relevant time point in relation to the disease (17). Below we discuss three commonly used study designs; family-based, case-control and population studies.

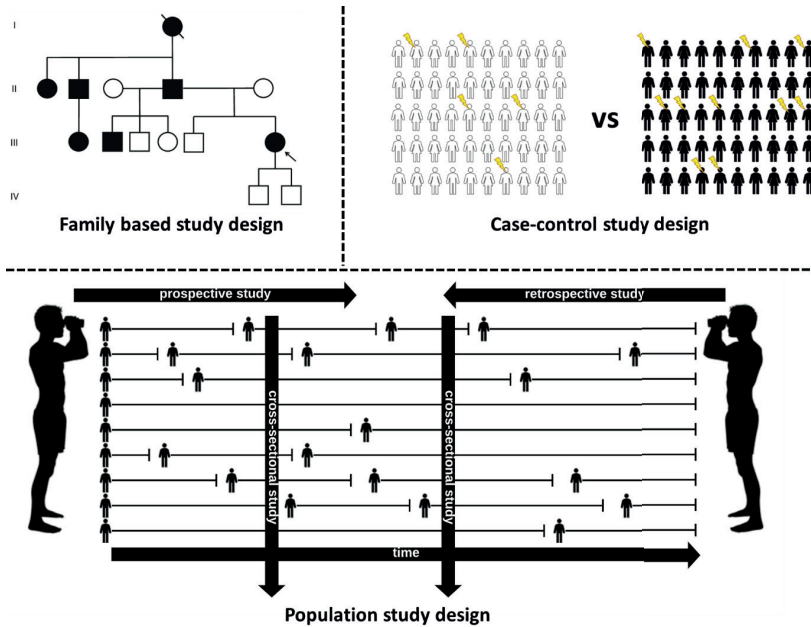
### ***Family-Based Studies***

In such studies, families where multiple relatives suffer from a certain disease are investigated, as shown below in figure 3. This design is specifically used in genetic studies. Usually all the genes (whole exome sequencing; WES) or the complete genome (whole genome sequencing; WGS) is sequenced in multiple family members, and all identified DNA variants per individual are annotated to the reference genome (43). DNA variants are studied whether they are present in all affected relatives and absent in all unaffected relatives of the family (43). In addition, for each variant we annotate their frequency in large datasets of healthy controls and the predicted impact on the function of the protein (27). For example, if a variant is never observed in healthy individuals and it is located in a gene where other genetic variants have been shown to cause a similar disease, it might be more likely that this new variant causes the disease in the studied family (4, 27, 44, 45). Family studies perform well when the disease is clearly inherited across multiple generations and multiple family members have DNA available for sequencing.

### ***Case-Control Studies***

In a case-control design for genetic studies, DNA is extracted and genotyped for a set of unrelated cases and controls. Genotyping is usually done either for a candidate gene or region (by analyzing a single SNP or several SNPs in a gene-wide fashion) or genome-wide by applying SNP arrays or sequencing (WES or WGS). Every DNA variant is identified, annotated to the reference genome and compared between both groups. Variants occurring more frequently in the case group are statistically “associated” to the disease, suggesting that carrying one of these variants increases that person's risk of acquiring the disease (46, 47). The difference between the groups indicating by how much this risk increases. In contrast to family studies, these variants usually also occur in healthy individuals, and only a portion of the cases in the study will carry that specific variant. In general, DNA variants with large deleterious effects are identified in families (as every carrier acquires a disease), whereas variants with smaller effects are identified through case-control studies. When performed in a genome-wide fashion, with either SNP arrays assessing >300k (tagging) SNPs or with WES or WGS, they are also referred to as genome-wide association studies (GWAS) (48). For dynamic genomic studies, the case-control design is the most common study design. In such studies, a tissue relevant to the disease is collected from a set of cases and controls and DNA methylation, RNA expression or protein abundance is measured (34, 49). These

studies must be designed such that the only difference between the cases and controls is the disease of interest, as every other factor might also influence the dynamic genomic data (50). When this is correctly done, every methylated site, expressed gene or protein can be measured and compared between the case and control groups. When a site, gene or protein is significantly different between both groups this indicates an association to the disease process, similar to the DNA GWAS studies (51, 52). However in dynamic genomic data studies this does not necessarily indicate a causal association, as the disease itself may also influence these measurements.



**Figure 3.** commonly used study designs in genomic research. Top-left; a family-based study design, the family tree contains four generations. Affected family members are shown in black, unaffected members by the white shapes. The clear inheritance across multiple members in multiple generation suggest a causal genetic variant. Top-right; a case-control study design. A number of cases (in black) and controls (in white). All participants are tested, for example for a DNA variant, and all tested positive are indicated by the yellow shape. The fraction of positive participants is compared between groups. Bottom; a population study design. A population of individuals is portrayed over time from left to right. People enter and exit the population. At any given time, we can test the population participants and compare cases with controls as a case-control study (often called cross-sectional design). We can also test participants at the start and follow them over time, or test them multiple times over a time period (prospective design). Prospective, repeated testing is the only way to disentangle causal and consequential changes in dynamic genomic studies.

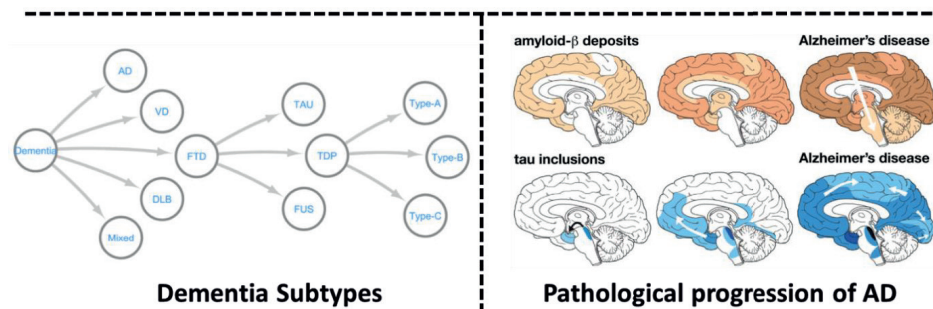
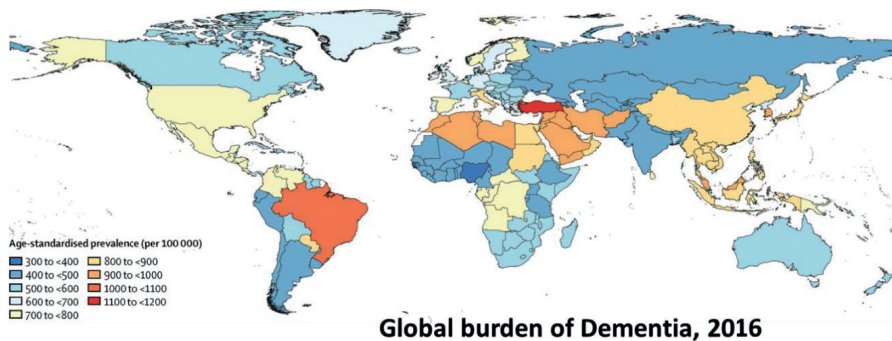
### ***Population-based studies***

Population-based studies are similar to case-control studies, but with a different sampling strategy and an added time-component. In general, a population study follows a large number of randomly selected individuals over time (prospective) as some develop a disease and others don't (53). This allows for repeated measurements before and after the onset of disease, supporting investigation of changes in the time-frame of the disease. The study population can vary, some are random representations of the healthy population, but it can also be a population of patients (54). Adding the time-component is important for the cause-consequence question in dynamic genomic studies, although the required tissue specificity can challenge repeated sampling of healthy study participants.

Genomic studies are used to generate insight into diseases. For example, genetic studies identify genes in which dysfunction causes or contributes to a disease (46, 47). Further investigation of these genes, their biological function and how that dysfunction exactly leads to disease helps understand why certain people get this disease and others do not. In addition, furthering our understanding of the biology behind disease may help in identifying methods to counter this dysfunction and developing treatments (55, 56). Dynamic genomic studies contribute much to this aspect, as they provide insight into the molecular and cellular state of the tissue in which the disease manifests (45, 57). In this thesis, two aspects of genomic studies are investigated; 1) general methodological aspects of such studies, which can be applied to almost any disease and 2) applying these specific study designs and data types to investigate Dementia.

## **Dementia**

Dementia is the collective term for a collection of neurodegenerative diseases (58). Each disease is marked by progressive decline of one or more cognitive domains (e.g., memory, language). Globally 50 million people suffer from dementia, about 70% being Alzheimer's Disease (AD) (58, 59). Other common forms are Frontotemporal Dementia (FTD), Dementia with Lewy Bodies (DLB) and Vascular Dementia (VD) (60, 61). The forms are broadly distinguished by the main affected cognitive domain, for example memory in AD and language or behavior in FTD, often correlated to the region of the brain that is degenerating (62). The causes of dementia are often not known, although genetic factors play a strong causal role in most forms (56). In this thesis, we focus on AD and FTD, specifically the Semantic Dementia form (SD) of FTD.



**Figure 4.** general characteristics of dementia. Top; world overview of the burden of dementia by country. Bottom-left; schematic view of most common dementia subtypes, further detailed for FTD subtypes based on pathology (Tau, TDP or FUS) and TDP-pathology subtypes (A, B or C). Semantic Dementia most commonly manifests as FTD-TDP-Type C. Bottom-right; schematic view of pathological progression in AD, shown for both amyloid pathology and for tau pathology. In short, amyloid pathology start cortical and spreads to the rest of the brain. Tau pathology starts in the entorhinal cortex and spreads to the hippocampus and cortical areas.

### ***Pathological presentation***

Dementia usually starts in a specific region in the brain and spreads to adjacent regions as the disease progresses, as illustrated in figure 4 for AD (63, 64). Affected brain regions typically undergo loss of neurons, resulting in so-called neurodegeneration. Additional features are pathological protein aggregations in specific brain regions, cell types or cellular compartments (65). In addition, these aggregations contain different proteins, and are thus usually characterized by the main component(s) with which the aggregates are stained; amyloid (AD), tau (AD, FTD), TDP (FTD, ALS) or synuclein (PD, DLB) (63-66). Further classification can be done based on cellular subtype or compartment and spatial pattern of pathological protein aggregates (66, 67). However, large pathological variation between affected patients exists, and pathology often becomes of mixed type as the disease progresses (68). Post-mortem pathological classification is the golden standard way of classifying the type and subtype of dementia in a patient. However, based on clinical presentation and evaluation of cerebral spinal fluid (CSF) and imaging (MRI, PET) biomarkers clinical classification can be done during life (54).

One of the earliest and most severely affected brain regions in AD is the hippocampus, involved in memory formation and retrieval (58). Typical AD pathology includes so-called intercellular plaques characterized by Amyloid-beta (AB-plaques) and intracellular neurofibrillary tangles characterized by hyperphosphorylated tau (NFTs) (62). This pathology spreads to the temporal and frontal lobes (language and behavior) and to entire brain in later stages (55, 65). FTD divides into clinical and pathological subtypes (66, 69). Typically, FTD pathology and neurodegeneration starts in the frontal and/or temporal lobes and is characterized by either NFTs or TDP43-positive protein aggregates (TDP43-positive inclusions) (67, 68). Subtypes of TDP43 are based on location and form of TDP43-positive inclusions and dystrophic neurites (DN) (66). Type A has many neuronal cytoplasmic inclusions (NCI) and short DN. Type B has a moderate amount of NCI and few DN. Type C has few NCI but many long DN. Type D has many short DN and shows neuronal intranuclear inclusions (NII) (64, 69).

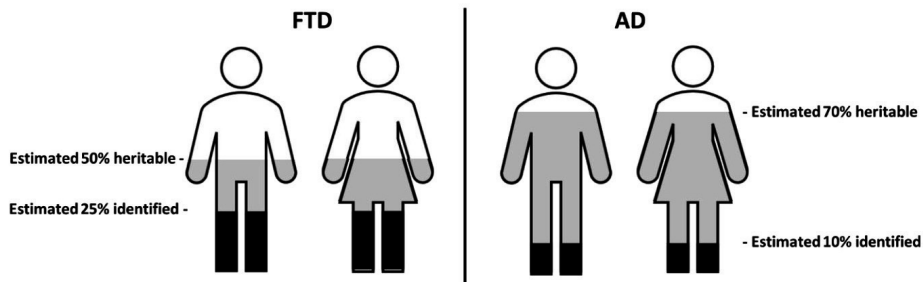
### ***Clinical presentation***

The clinical subtypes of dementia often correlate to the pathological subtypes. AD manifests as neurodegeneration in the hippocampus and patients thus present with progressive memory loss (62, 70). Similarly, in FTD patients the temporal or frontal lobe degenerates and they thus present with symptoms in the language or behavior domains. Several clinical FTD subtypes are defined; behavioral variant FTD (bvFTD), semantic variant FTD (svFTD), non-fluent primary progressive aphasia (nfPPA), motor neuron disease (FTD-MND) and a few other, rarer forms (61).

In this thesis we investigate one of the FTD subtypes; the semantic variant primary progressive aphasia, often referred to as semantic dementia (SD) (71, 72). Clinical presentation of SD starts with impairment of language comprehension and word finding difficulties, disrupting the patient's communication with others and often leading to social isolation (61, 71). In later stages behavioral symptoms usually manifest, for example as compulsive behavior (61, 71). Pathologically, SD manifests as localized unilateral atrophy of the temporal lobe, which in later stages also affects the other temporal lobe (73). Many DN, but few NCI are present in the temporal and frontal cortex and the pathology classifies as TDP Type C (66). Additionally, a large number of NCI are observed in the dentate gyrus region of the hippocampus. The hippocampus (memory) and temporal lobe (language) collaborate to perform speech processing (i.e., retrieving the memory that belongs to an object's name), the main cognitive function disrupted in SD (71, 73). SD is clinically and pathologically relatively homogeneous and the clinicopathological correlation is relatively high (73). Also, SD rarely occurs in familial form and no genetic variants causing SD have been described (61), unlike almost every other form of dementia.

### ***Genetic studies in AD and FTD***

Both AD and FTD are considered complex, multifactorial, diseases with a large heritable component, as shown in figure 5. Both diseases may take familial form, with a highly penetrant variant causing AD or FTD in every carrier (74). In parallel, genetic risk factors in the overall population, where carriers have increased risk of acquiring AD or FTD but can also remain healthy. As shown in figure 5, the total proportion of FTD that is caused by genetic factors is estimated on approximately 50% (46, 54, 61). This is approximately 70% for AD (74). Combining familial variants and population risk factors, we estimate that approximately half of the genetic component of FTD has been identified, against approximately 10% for AD (46, 47, 74).



**Figure 5.** the estimated total and currently identified heritable component of FTD and AD. For FTD, approximately half of all occurrence of disease is estimated to be caused by genetic factors, of which half again is identified. For AD, approximately 70% is estimated to be genetic, of which 10% is currently identified.

Family-based genetic studies in AD and FTD have identified genes with highly penetrant disease-causing variants. For AD, familial variants in *APP*, *PSEN1*, *PSEN2* and *SORL1* make up about 1-2% of the estimated genetic components of the disease (75, 76). For FTD, variants in *C9ORF72*, *GRN*, *MAPT*, *TARDBP*, *CHMP2B* and *VCP* compose ~50% of the genetic contribution to the disease (54, 61, 77, 78). The variants in these genes are often highly penetrant (i.e., almost all carriers of the variant acquire the disease) (44, 54, 61).

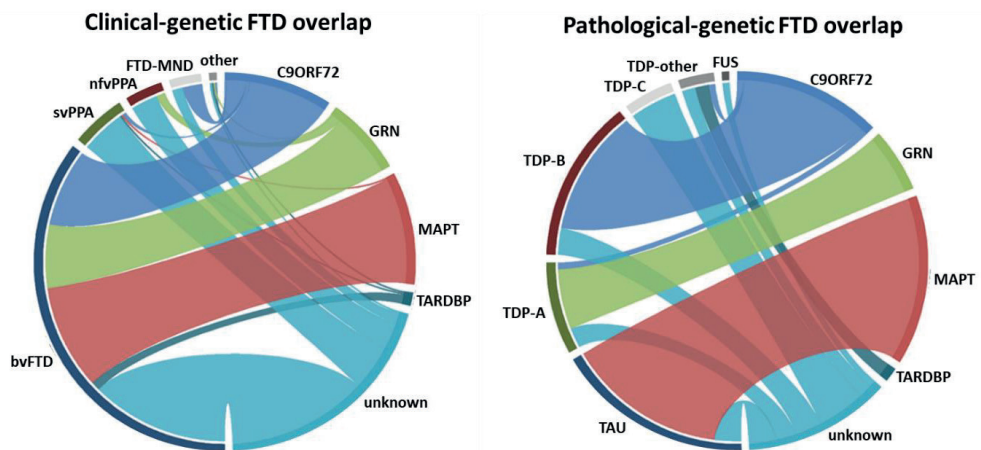
Population-based GWAS on AD and FTD have identified additional genetic factors that increase the risk of acquiring disease. These genetic factors are common in the population and individually have a lower penetrance, meaning that the risk is not increased by a large amount when carrying one such a variant, and individuals may carry the variant without acquiring the disease. However, such GWAS have also highlighted that the so-called “genetic architecture” of complex diseases, such as AD and FTD, consists of many hundreds if not thousands of such common risk-variants. Collectively, such sets of common variants can explain a substantial part of the genetic variance of the disease. The trend in GWAS is therefore to perform reiterative meta-analyses of ever bigger GWAS datasets to identify the growing list of common risk variants which explain increasing amounts of explained genetic variance. One of the most well-known common genetic risk factors is the combination of two genetic variants (rs7412 and rs429358) in the *APOE* gene, which are denoted as e2, e3 or e4, where e3 is most common (79). Heterozygous carriers of the e4 combination (e3/e4) have a 4-fold increased risk of developing AD, and homozygous carriers (e4/e4) have an 11-fold increased risk (79, 80). However, homozygous carriers exist that never acquire the disease. For almost all other known genetic risk factors, the increased risk is usually smaller than 1.5x (56, 74).

The largest population study for AD included 94,437 cases and identified genetic risk factors in 25 genes (47, 74, 81) explaining approximately 31% of the genetic variance for late-onset AD. For FTD the largest population study contained 3,526 FTD patients and 9,402 controls explaining only a modest amount of the genetic variance (46). In this study, patients already carrying a variant in one of the known familial disease genes were excluded. Five additional genes were identified where genetic variants increased the risk of FTD; *RAB38*, *CTSC*, *HLA-DRA*, *HLA-DRB5* and *BTNL2* (46, 78). The change in AD or FTD disease risk of each separate



variant is low with odds ratios ranging from approximately 0.75 to 1.25, and the biological mechanism through which they contribute to the disease is largely unknown.

A large fraction of the identified heritability in FTD stems from a limited number of genes in which many of the familiar cases carry a causal variant. In the Dutch FTD patient population, approximately 37% of patients with positive family history were identified with a genetic variant. Most carried the expanded repeat in *C9ORF72* (21%), 6% carried a pathogenic single nucleotide variant or small insertion or deletion in *MAPT*, 4.5% in *GRN*, 3.5% in *TARDBP* and another 2.5% carried a likely causal variant in *VCP*, *TBK1*, *PSEN1* or *OPTN*. In figure 6, we show the clinical and pathological FTD subtypes of these genetic groups.



**Figure 6.** Clinical or pathological classification of 198 Dutch FTD patients, stratified by the gene in which they carry a genetic defect, when known. Both diagrams indicate on the right-side patients caused by genetic defects in *C9orf72*, *GRN*, *MAPT*, *TARDBP* or patients with unknown genetic or other causes. The left side of the left diagram displays the clinical presentations (left figure); behavioral-variant FTD (bvFTD), semantic-variant primary progressive aphasia (svPPA, also known as SD), non-fluent-variant primary progressive aphasia (nfvPPA), FTD with motor neuron disease (FTD-MND) and other. The left-side of the diagram on the right indicates pathological categories; Tau pathology, TDP pathology type A, B or C, FUS pathology and other. The size of the group is represented by the size of the outer ring fragments. The size of the overlap between genetic and clinical (left) or pathological (right) groups is demonstrated by the size of the connecting bands.

The clinical-genetic diagram shows that the main clinical group; behavioral-variant FTD presents in all main genetic groups. In contrast, semantic-variant and non-fluent-variant primary progressive aphasia present mostly in the group with unknown genetic cause, although nfvPPA can also be caused by *GRN* genetic variants. FTD-motor neuron disease is mostly caused by the *c9orf72* expansion, and the *TARDBP* and unknown genetic groups have the most mixed clinical presentation. The pathological-genetic figure show clear overlap for the main genetic groups; *C9ORF72* presents mostly with TDP-type-B, *GRN* with TDP-type-A and *MAPT* with TAU pathology. Vice-versa, although each main pathological group still contains patients with unknown genetic cause, most patients with a specific pathology are caused by variants in the respective gene. Overall, this overview demonstrates clear

genetic FTD subgroups with distinct pathological, and sometimes clinical, presentation. Nevertheless, in a relatively large groups of patients the suspected genetic defect has not been identified.

### ***Dynamic genomic data studies in AD and FTD***

For both AD and FTD, dynamic genomic studies have been performed comparing the methylation, expression or proteomic patterns in brain tissue of cases with controls. Most genomics studies for either AD or FTD so far have reported decreased activity of neurotransmitter signaling and energy metabolism and increased activity of stress response pathways and epigenetic regulation (51, 52, 57). Due to the dynamic nature of the data, it is difficult to determine which changes represent causal changes and which are consequence of the disease. However, these changes are generally observed in all neurodegenerative tissues and are considered mostly consequential changes, caused by degeneration of neurons and activation of glial cells to cope with the damage to the brain (57, 82). Most dynamic genomic datasets derived for AD or FTD use frozen brain tissues of post-mortem donors, obtained at the end of the disease.

Most dynamic genomic data studies for AD or FTD include a single brain region between cases and controls (12, 52). They statistically compare dynamic genomics data (e.g., each methylated CpG site, gene expression or protein abundances) between both groups, corrected for confounding factors as age and gender. The CpGs, genes or proteins that are significantly different between both groups are further investigated, for example by comparing to other dynamic genomic data studies.

To translate these individual changes to biological and clinical disease insights, the CpGs, genes or proteins are often grouped into biological pathways based on their described functions (83-85). For example, all genes that are involved in response to stress. These changes in biological pathways are easier to interpret than single genes, and can make it easier to compare between different studies or diseases (57, 86). A challenge to this approach is that the gene function is not always known or completely described, and standardized methods to study dynamic genomics data in such a way are still lacking (87, 88).

In addition to studying a single brain region, several studies have collected data in a different design. For example, including cases with different severity of the disease (57, 82). By separately comparing severe and mildly affected cases to control samples it is possible to add some claims on the timeframe of the dynamic genomic data changes throughout the disease process, although not in the same individual. This design is informative, but challenged by the scarcity of early-stage post-mortem brain samples. Other studies collected data from multiple brain regions and are separately comparing these to control brains, followed by investigating the differences in comparison between each region (57). In this way, severely and mildly affected brain regions from the same individual can be compared, which also provides some insight into disease progression (57). A more novel approach is single-cell dynamic genomic data analysis. In these studies, individual cells are derived, measured and compared between brains of cases and of controls (89). These studies show further heterogeneity of cellular activity and function, even within the same brain region of one patient (89).

A special dynamic genomic data study-type is a biomarker study. Here, dynamic genomic data is collected from the blood or CSF to identify biological markers identifying/predicting the disease state. As blood or CSF can be extracted during life and at multiple time points, it permits repeated measurements of patients as the disease progresses (90). The aim of these studies is not necessarily to investigate the underlying biology, but to discover markers that can identify or stratify patients as a tool in the diagnostic procedure (91, 92).

## Study populations in this thesis

Three datasets are studied in this thesis; the Rotterdam Study (RS), the FTD patients enrolled at the department of Neurology (Neurology) and the Dementia patients and controls that donated their brain to the Netherlands Brain Bank (NHB).

The Rotterdam Study cohort is a population-based cohort founded in 1990 to investigate disease and disability in the elderly in the Netherlands (53). The cohort comprises ~15,000 participants that enrolled in 1990, 2000 or 2006. All participants were at least 45 years at enrollment, and undergo extensive research-based measurements every five years, including blood draws (53). Their medical records, measurements and DNA extracted from blood are available for researchers.

The FTD cohort is collected over the last 30 years by the department of Neurology at the Erasmus Medical Center. This cohort includes ~700 FTD patients, and is representative for a clinical FTD population (61, 93). Extensive medical information is collected for these patients, with clinical measurements, MRI imaging, pathology (when available) and often multiple blood and/or CSF draws (54, 61). We selected from this cohort the patients that were diagnosed with Semantic Dementia. Many patients in this cohort have donated their brains to scientific research and are also present in the Dutch brain bank cohort.

The Dutch Brain Bank (NHB) cohort consists of neurological patients and non-demented controls that donated their brain to scientific research (94). For all donors, post-mortem frozen tissue is available for dozens of brain regions, as well as a selected set of clinical and pathological parameters. The biobank can be mined for brain tissues of cases and controls of interest. Over the last 30 years, more than 4,000 brains have been collected by the NHB, including ~900 AD brains and ~200 FTD brains, and is one of the largest such biobanks worldwide (94).

### *Outline of the thesis*

In this thesis, we investigated applications of next-generation sequencing. Either in the form of best practices when using NGS data, or by applying NGS to answer research questions in the AD or FTD field. In **chapter 2.1**, we describe the generation of an exome sequencing population dataset and demonstrate how the majority of genetic variants are population specific. We offer recommendations on the analysis and interpretation of DNA based NGS data from such population-based datasets. This topic is continued in **chapter 2.2** where we investigate the occurrence and interpretation of pathogenic variants in disease-causing genes in DNA NGS data. Then, in **chapter 2.3** we perform a DNA study in several Alzheimer's

Disease families and identify a candidate gene that might cause the disease in two of these families. In **chapter 3.1** we move to dynamic genomic data by investigating the analysis methods used in RNA sequencing and/or DNA methylation studies. We compare commonly used methods and provide recommendations on their use. This topic is continued in **chapter 3.2**, where we studied post-mortem gene expression in hippocampus of AD brains versus control brains. We demonstrate how such an RNA NGS dataset can be used to investigate the biology underlying AD, and how datasets can be compared on biological pathway level. In **chapter 4.1** we combine multiple of these methods and perform a dynamic genomic study on DNA NGS data. We compare the DNA in the brain of semantic dementia patients with DNA from their blood and identify tissue-specific somatic DNA variants. In **chapter 5** we discuss the results obtained by the studies in this thesis, and how these contribute to the field of genomics and dementia. Finally, we outline the most recent and upcoming developments in the genomic field and how these will further research into dementia biology.

## References

1. International Human Genome Sequencing C. Finishing the euchromatic sequence of the human genome. *Nature*. 2004;431(7011):931-45.
2. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518(7539):317-30.
3. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. 2015;348(6235):648-60.
4. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56-65.
5. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, et al. Disease variants alter transcription factor levels and methylation of their binding sites. *Nat Genet*. 2017;49(1):131-8.
6. Ge T, Chen CY, Neale BM, Sabuncu MR, Smoller JW. Phenome-wide heritability analysis of the UK Biobank. *PLoS Genet*. 2017;13(4):e1006711.
7. Brainstorm C, Anttila V, Bulik-Sullivan B, Finucane HK, Walters RK, Bras J, et al. Analysis of shared heritability in common disorders of the brain. *Science*. 2018;360(6395).
8. Felsenfeld G, Groudine M. Controlling the double helix. *Nature*. 2003;421(6921):448-53.
9. Dewar JM, Walter JC. Mechanisms of DNA replication termination. *Nat Rev Mol Cell Biol*. 2017;18(8):507-16.
10. Stiles J, Jernigan TL. The basics of brain development. *Neuropsychol Rev*. 2010;20(4):327-48.
11. Greer EL, Shi Y. Histone methylation: a dynamic mark in health, disease and inheritance. *Nat Rev Genet*. 2012;13(5):343-57.
12. Humphries CE, Kohli MA, Nathanson L, Whitehead P, Beecham G, Martin E, et al. Integrated whole transcriptome and DNA methylation analysis identifies gene networks specific to late-onset Alzheimer's disease. *J Alzheimers Dis*. 2015;44(3):977-87.
13. Zemach A, McDaniel IE, Silva P, Zilberman D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science*. 2010;328(5980):916-9.
14. Amaral PP, Dinger ME, Mercer TR, Mattick JS. The eukaryotic genome as an RNA machine. *Science*. 2008;319(5871):1787-9.
15. Alba M. Replicative DNA polymerases. *Genome Biol*. 2001;2(1):REVIEWS3002.
16. Kavanagh T, Mills JD, Kim WS, Halliday GM, Janitz M. Pathway analysis of the human brain transcriptome in disease. *J Mol Neurosci*. 2013;51(1):28-36.
17. Consortium GT. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45(6):580-5.
18. Hitzemann R, Darakjian P, Walter N, Iancu OD, Searles R, McWeeney S. Introduction to sequencing the brain transcriptome. *Int Rev Neurobiol*. 2014;116:1-19.
19. Wang W, Nag S, Zhang X, Wang MH, Wang H, Zhou J, et al. Ribosomal proteins and human diseases: pathogenesis, molecular mechanisms, and therapeutic implications. *Med Res Rev*. 2015;35(2):225-85.
20. Hebert DN, Molinari M. In and out of the ER: protein folding, quality control, degradation, and related human diseases. *Physiol Rev*. 2007;87(4):1377-408.
21. Khoury GA, Baliban RC, Floudas CA. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci Rep*. 2011;1.
22. Zhang YW, Thompson R, Zhang H, Xu H. APP processing in Alzheimer's disease. *Mol Brain*. 2011;4:3.
23. Marcelli S, Corbo M, Iannuzzi F, Negri L, Blandini F, Nistico R, et al. The Involvement of Post-Translational Modifications in Alzheimer's Disease. *Curr Alzheimer Res*. 2018;15(4):313-35.

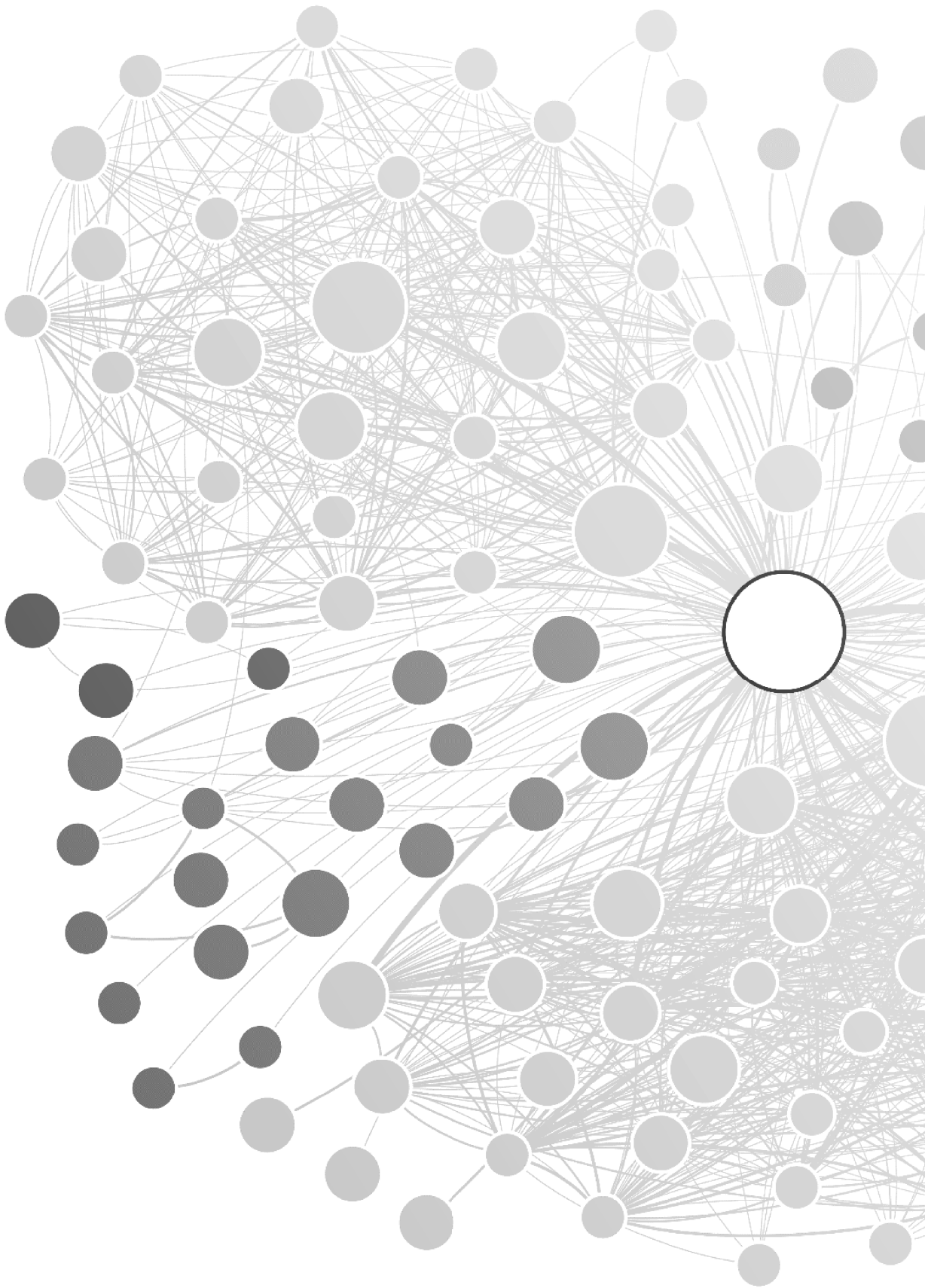
24. Ren RJ, Dammer EB, Wang G, Seyfried NT, Levey AI. Proteomics of protein post-translational modifications implicated in neurodegeneration. *Transl Neurodegener.* 2014;3(1):23.
25. Stranger BE, Dermitzakis ET. From DNA to RNA to disease and back: the 'central dogma' of regulatory disease variation. *Hum Genomics.* 2006;2(6):383-90.
26. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature.* 2016;536(7616):285-91.
27. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet.* 2014;46(3):310-5.
28. Russell R. RNA misfolding and the action of chaperones. *Front Biosci.* 2008;13:1-20.
29. Zhernakova DV, Deelen P, Vermaat M, van Iterson M, van Galen M, Arindrarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat Genet.* 2017;49(1):139-45.
30. Lo R, Weksberg R. Biological and biochemical modulation of DNA methylation. *Epigenomics.* 2014;6(6):593-602.
31. Katakura Y, Okui T, Kishi R, Ikeda T, Miyake H. [Distribution of <sup>14</sup>C-formaldehyde in pregnant mice: a study by liquid scintillation counter and binding to DNA]. *Sangyo Igaku.* 1991;33(4):264-5.
32. Selbach M, Schwanhauser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. Widespread changes in protein synthesis induced by microRNAs. *Nature.* 2008;455(7209):58-63.
33. Roux PP, Topisirovic I. Signaling Pathways Involved in the Regulation of mRNA Translation. *Mol Cell Biol.* 2018;38(12).
34. Peters MJ, Joehanes R, Pilling LC, Schurmann C, Conneely KN, Powell J, et al. The transcriptional landscape of age in human peripheral blood. *Nat Commun.* 2015;6:8570.
35. Pettersson E, Lundeberg J, Ahmadian A. Generations of sequencing technologies. *Genomics.* 2009;93(2):105-11.
36. Gayon J. From Mendel to epigenetics: History of genetics. *C R Biol.* 2016;339(7-8):225-30.
37. Gauthier J, Vincent AT, Charette SJ, Derome N. A brief history of bioinformatics. *Brief Bioinform.* 2018.
38. Garrido-Cardenas JA, Garcia-Maroto F, Alvarez-Bermejo JA, Manzano-Agugliaro F. DNA Sequencing Sensors: An Overview. *Sensors (Basel).* 2017;17(3).
39. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, et al. DNA sequencing at 40: past, present and future. *Nature.* 2017;550(7676):345-53.
40. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet.* 2016;17(6):333-51.
41. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature.* 2015;526(7571):68-74.
42. Bae T, Tomasini L, Mariani J, Zhou B, Roychowdhury T, Franjic D, et al. Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. *Science.* 2018;359(6375):550-5.
43. Wong TH, van der Lee SJ, van Rooij JGJ, Meeter LHH, Frick P, Melhem S, et al. EIF2AK3 variants in Dutch patients with Alzheimer's disease. *Neurobiol Aging.* 2019;73:229 e11- e18.
44. Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.* 2017;45(D1):D840-D5.
45. Wong TH, Chiu WZ, Breedveld GJ, Li KW, Verkerk AJ, Hondius D, et al. PRKAR1B mutation associated with a new neurodegenerative disorder with unique pathology. *Brain.* 2014;137(Pt 5):1361-73.
46. Ferrari R, Hernandez DG, Nalls MA, Rohrer JD, Ramasamy A, Kwok JB, et al. Frontotemporal dementia and its subtypes: a genome-wide association study. *Lancet Neurol.* 2014;13(7):686-99.

47. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet.* 2013;45(12):1452-8.
48. Minikel EV, Vallabh SM, Lek M, Estrada K, Samocha KE, Sathirapongsasuti JF, et al. Quantifying prion disease penetrance using large population control cohorts. *Sci Transl Med.* 2016;8(322):322ra9.
49. van Rooij JGJ, Meeter LHH, Melhem S, Nijholt DAT, Wong TH, Netherlands Brain B, et al. Hippocampal transcriptome profiling combined with protein-protein interaction analysis elucidates Alzheimer's disease pathways and genes. *Neurobiol Aging.* 2019;74:225-33.
50. van Rooij J, Mandaviya PR, Claringbould A, Felix JF, van Dongen J, Jansen R, et al. Evaluation of commonly used analysis strategies for epigenome- and transcriptome-wide association studies through replication of large-scale population studies. *Genome Biol.* 2019;20(1):235.
51. Sekar S, McDonald J, Cuyugan L, Aldrich J, Kurdoglu A, Adkins J, et al. Alzheimer's disease is associated with altered expression of genes involved in immune response and mitochondrial processes in astrocytes. *Neurobiol Aging.* 2015;36(2):583-91.
52. Twine NA, Janitz K, Wilkins MR, Janitz M. Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease. *PLoS One.* 2011;6(1):e16266.
53. Ikram MA, Brusselle GGO, Murad SD, van Duijn CM, Franco OH, Goedegebure A, et al. The Rotterdam Study: 2018 update on objectives, design and main results. *Eur J Epidemiol.* 2017;32(9):807-50.
54. Seelaar H, Kamphorst W, Rosso SM, Azmani A, Masdjedi R, de Koning I, et al. Distinct genetic forms of frontotemporal dementia. *Neurology.* 2008;71(16):1220-6.
55. Selkoe DJ, Hardy J. The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Mol Med.* 2016;8(6):595-608.
56. Van Cauwenbergh C, Van Broeckhoven C, Sleegers K. The genetic landscape of Alzheimer disease: clinical implications and perspectives. *Genet Med.* 2016;18(5):421-30.
57. Wang M, Roussos P, McKenzie A, Zhou X, Kajiwara Y, Brennand KJ, et al. Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer's disease. *Genome Med.* 2016;8(1):104.
58. Scheltens P, Blennow K, Breteler MM, de Strooper B, Frisoni GB, Salloway S, et al. Alzheimer's disease. *Lancet.* 2016;388(10043):505-17.
59. Prince M, Bryce R, Albanese E, Wimo A, Ribeiro W, Ferri CP. The global prevalence of dementia: a systematic review and metaanalysis. *Alzheimers Dement.* 2013;9(1):63-75 e2.
60. Knopman DS, Roberts RO. Estimating the number of persons with frontotemporal lobar degeneration in the US population. *J Mol Neurosci.* 2011;45(3):330-5.
61. Seelaar H, Rohrer JD, Pijnenburg YA, Fox NC, van Swieten JC. Clinical, genetic and pathological heterogeneity of frontotemporal dementia: a review. *J Neurol Neurosurg Psychiatry.* 2011;82(5):476-86.
62. Ballard C, Gauthier S, Corbett A, Brayne C, Aarsland D, Jones E. Alzheimer's disease. *Lancet.* 2011;377(9770):1019-31.
63. Braak H, Braak E. Staging of Alzheimer's disease-related neurofibrillary changes. *Neurobiol Aging.* 1995;16(3):271-8; discussion 8-84.
64. Neumann M, Sampathu DM, Kwong LK, Truax AC, Micsenyi MC, Chou TT, et al. Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science.* 2006;314(5796):130-3.

65. Murray ME, Graff-Radford NR, Ross OA, Petersen RC, Duara R, Dickson DW. Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. *Lancet Neurol.* 2011;10(9):785-96.
66. Mackenzie IR, Neumann M, Baborie A, Sampathu DM, Du Plessis D, Jaros E, et al. A harmonized classification system for FTLN-TDP pathology. *Acta Neuropathol.* 2011;122(1):111-3.
67. Bodea LG, Eckert A, Ittner LM, Piguot O, Gotz J. Tau physiology and pathomechanisms in frontotemporal lobar degeneration. *J Neurochem.* 2016;138 Suppl 1:71-94.
68. Irwin DJ, Cairns NJ, Grossman M, McMillan CT, Lee EB, Van Deerlin VM, et al. Frontotemporal lobar degeneration: defining phenotypic diversity through personalized medicine. *Acta Neuropathol.* 2015;129(4):469-91.
69. Mackenzie IR, Neumann M, Bigio EH, Cairns NJ, Alafuzoff I, Kril J, et al. Nomenclature and nosology for neuropathologic subtypes of frontotemporal lobar degeneration: an update. *Acta Neuropathol.* 2010;119(1):1-4.
70. Jack CR, Jr., Knopman DS, Jagust WJ, Shaw LM, Aisen PS, Weiner MW, et al. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol.* 2010;9(1):119-28.
71. Gorno-Tempini ML, Hillis AE, Weintraub S, Kertesz A, Mendez M, Cappa SF, et al. Classification of primary progressive aphasia and its variants. *Neurology.* 2011;76(11):1006-14.
72. Rascovsky K, Hodges JR, Knopman D, Mendez MF, Kramer JH, Neuhaus J, et al. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain.* 2011;134(Pt 9):2456-77.
73. Hodges JR, Patterson K. Semantic dementia: a unique clinicopathological syndrome. *Lancet Neurol.* 2007;6(11):1004-14.
74. Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat Genet.* 2019;51(3):404-13.
75. Holstege H, van der Lee SJ, Hulsman M, Wong TH, van Rooij JG, Weiss M, et al. Characterization of pathogenic SORL1 genetic variants for association with Alzheimer's disease: a clinical interpretation strategy. *Eur J Hum Genet.* 2017;25(8):973-81.
76. St George-Hyslop PH. Molecular genetics of Alzheimer disease. *Semin Neurol.* 1999;19(4):371-83.
77. Wong TH, Pottier C, Hondius DC, Meeter LHH, van Rooij JGJ, Melhem S, et al. Three VCP Mutations in Patients with Frontotemporal Dementia. *J Alzheimers Dis.* 2018;65(4):1139-46.
78. Olszewska DA, Lonergan R, Fallon EM, Lynch T. Genetics of Frontotemporal Dementia. *Curr Neurol Neurosci Rep.* 2016;16(12):107.
79. Yu JT, Tan L, Hardy J. Apolipoprotein E in Alzheimer's disease: an update. *Annu Rev Neurosci.* 2014;37:79-100.
80. Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R, et al. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *JAMA.* 1997;278(16):1349-56.
81. Bettens K, Sleegers K, Van Broeckhoven C. Genetic insights in Alzheimer's disease. *Lancet Neurol.* 2013;12(1):92-104.
82. Hondius DC, van Nierop P, Li KW, Hoozemans JJ, van der Schors RC, van Haastert ES, et al. Profiling the human hippocampal proteome at all pathologic stages of Alzheimer's disease. *Alzheimers Dement.* 2016;12(6):654-68.
83. Kong W, Mou X, Zhang N, Zeng W, Li S, Yang Y. The construction of common and specific significance subnetworks of Alzheimer's disease from multiple brain regions. *Biomed Res Int.* 2015;2015:394260.



- 
84. Chi LM, Wang X, Nan GX. In silico analyses for molecular genetic mechanism and candidate genes in patients with Alzheimer's disease. *Acta Neurol Belg.* 2016;116(4):543-7.
  85. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 2002;30(7):1575-84.
  86. Ferrari R, Forabosco P, Vandrovцова J, Botia JA, Guelfi S, Warren JD, et al. Frontotemporal dementia: insights into the biological underpinnings of disease through gene co-expression network analysis. *Mol Neurodegener.* 2016;11:21.
  87. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000;25(1):25-9.
  88. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A.* 2005;102(43):15545-50.
  89. Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, et al. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature.* 2019.
  90. van der Ende EL, Meeter LH, Stingl C, van Rooij JGJ, Stoop MP, Nijholt DAT, et al. Novel CSF biomarkers in genetic frontotemporal dementia identified by proteomics. *Ann Clin Transl Neurol.* 2019;6(4):698-707.
  91. Meeter LH, Kaat LD, Rohrer JD, van Swieten JC. Imaging and fluid biomarkers in frontotemporal dementia. *Nat Rev Neurol.* 2017;13(7):406-19.
  92. Teunissen CE, Elias N, Koel-Simmelink MJ, Durieux-Lu S, Malekzadeh A, Pham TV, et al. Novel diagnostic cerebrospinal fluid biomarkers for pathologic subtypes of frontotemporal dementia identified by proteomics. *Alzheimers Dement (Amst).* 2016;2:86-94.
  93. Rosso SM, Donker Kaat L, Baks T, Joosse M, de Koning I, Pijnenburg Y, et al. Frontotemporal dementia in The Netherlands: patient characteristics and prevalence estimates from a population-based study. *Brain.* 2003;126(Pt 9):2016-22.
  94. Netherlands Brain Bank. <https://www.brainbank.nl> 2019



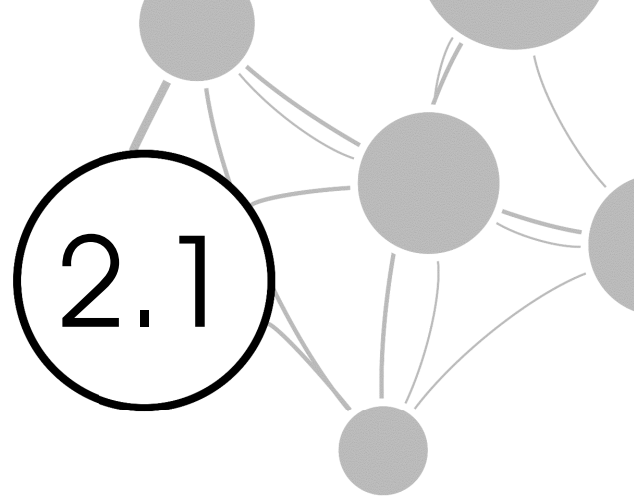


# Chapter 2

## Sequencing blood DNA



# Chapter 2.1



Population-specific genetic variation  
in large sequencing datasets; why  
more data is still better

*Published as a short report in the European Journal of Human Genetics  
(IF=4.3) on October 25<sup>th</sup>, 2017 (PMID:28905877, doi: 10.1038/  
ejhg.2017.110.)*

## **Abstract**

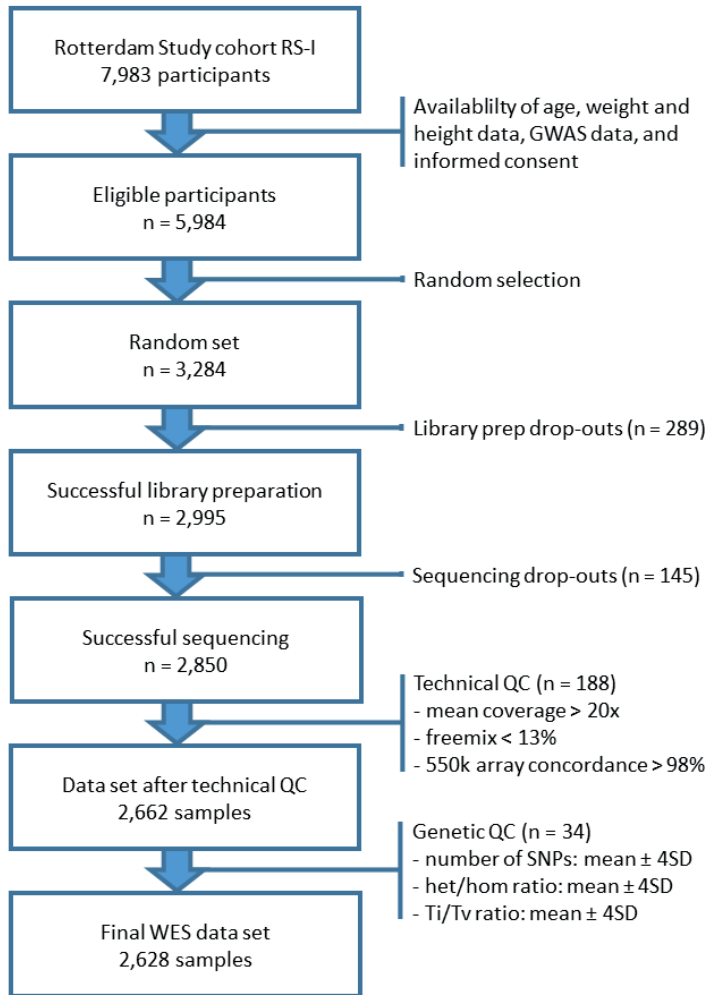
We have generated a next generation whole exome sequencing dataset of 2,628 participants of the population-based Rotterdam Study cohort, comprising 669,737 single nucleotide variants and 24,019 short insertions and deletions. Because of broad and deep longitudinal phenotyping of the Rotterdam Study, this dataset permits extensive interpretation of genetic variants on a range of clinically relevant outcomes, and is accessible as a control dataset. We show that next generation sequencing datasets yield a large degree of population specific variants, which are not captured by other available large sequencing efforts, being ExAC, ESP, 1000G, UK10K, GoNL, and DECODE.

**Keywords;** Rotterdam Study, Next Generation Sequencing, Exome, Population Genetics, Rare Variation

## Introduction

In the era of Next Generation Sequencing (NGS), the use of large population datasets to approximate variant frequencies in control populations has become common practice. The first large population-scale sequencing dataset was generated by the 1000 Genomes Project <sup>(1)</sup>, where an integrated genome-wide map of genetic variation was established for 2,504 individuals of European, American, African and Asian descent. Another approach was made by the NHLBI “Grand Opportunity” Exome Sequencing Project, in which a set of 6,500 European and African Americans samples was exome sequenced <sup>(2)</sup>. The recent Exome Aggregation Consortium (ExAC) is now combining exome sequencing datasets from over 60,000 unrelated individuals from different origins <sup>(3)</sup>. From these large sequencing projects, it became apparent that many variants are population-specific <sup>(3)</sup>. Therefore, several initiatives have generated more local datasets. The UK10K project <sup>(4)</sup> contains 4,000 genomes from the UK, along with 6,000 exomes from individuals with selected extreme phenotypes. A collection of 3,000 Finnish exomes, showed that the Finnish population had more loss-of-function variants and gene knock-outs than non-Finish Europeans <sup>(5)</sup>. GoNL <sup>(6)</sup>, the Dutch reference genome project, provided a local genetic map based on whole genome sequencing of 250 Dutch trios <sup>(7)</sup>. Another local dataset is based on full genomes from 2,636 Icelanders <sup>(8)</sup>. In this isolated population, deleterious variants could reach higher frequencies than in other populations. These initiatives emphasize the importance of local genetic maps to interpret clinical relevance of a potential disease-causing mutation, and indicate the differences in available population datasets that should be considered when these are used in research or clinical practice.

Within the Rotterdam Study cohort, a prospective population-based cohort study on individuals 45 years and older to investigate determinants of disease and disability in the Dutch population <sup>(9)</sup>, we have generated a set of 2,628 exomes for integrative genetic studies of diverse phenotypes and to serve as local reference panel for clinical sequencing efforts.



**Figure 1.** Overview of sample selection and quality control. Out of 5,984 eligible samples, a final random set of 2,628 exomes was generated. QC, quality control; SNP, single nucleotide polymorphism; SD, standard deviation; het/hom ratio, ratio between heterozygous and homozygous positions; Ti/Tv ratio, ratio between transitions and transversions.



## Methods

DNA samples were obtained from the Rotterdam Study, which is a prospective population-based cohort study established in 1990 studying the determinants of disease and disability in Dutch elderly individuals <sup>(9)</sup>. Out of 5,984 eligible participants from the RS-I cohort - based on the availability of height, weight, GWAS data and informed consent - 3,284 subjects were randomly selected, as shown in Fig 1. Baseline characteristics are provided in Supplementary Table 1.

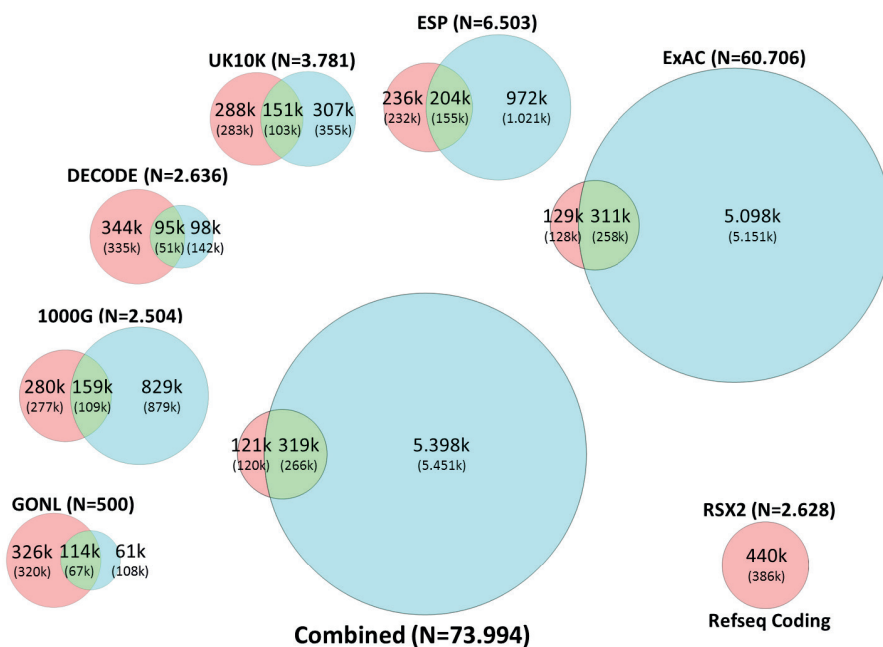
Genomic DNA was prepared from whole blood and processed using the Illumina TruSeq DNA Library preparation (Illumina, Inc., San Diego, CA), followed by exome capture using the Nimblegen SeqCap EZ V2 kit (Roche Nimblegen, Inc., Madison, WI). Paired-end 2 x 100bp sequencing was performed at 6 samples per lane on Illumina HiSeq2000 sequencer using Illumina TruSeq V3 chemistry.

Reads were demultiplexed and aligned to the human reference genome hg19 (UCSC, Genome Reference Consortium GRCh37) using the Burrows-Wheeler alignment tool (BWA version 0.7.3a <sup>(10)</sup>). After indel realignment and base quality score recalibration using the Genome Analysis ToolKit (GATK version 2.7.4 <sup>(11)</sup>) and masking of duplicates (Picard Tools version 1.90 <sup>(12)</sup>), gvcf files were generated using HaplotypeCaller v3.1.1 (GATK) and genotyped using GenotypeGVCFs v3.1.1 (GATK) <sup>(11)</sup>. Raw genotype data was QC-ed and filtered as described in the Supplementary Information.

All detected variants were annotated based on RefSeq annotation (NCBI Reference Sequence Database) using ANNOVAR (version 2014-07-14 <sup>(13)</sup>). The presence and allele frequencies of these variants in various databases: 1000G (v3) <sup>(1)</sup>, ESP (v2) <sup>(2)</sup>, ExAC (v0.3) <sup>(3)</sup>, UK10K (v1407) <sup>(4)</sup>, DECODE (v1501) <sup>(8)</sup> and the Genome of the Netherlands (v4) <sup>(6)</sup> were obtained and compared to our dataset.

## Results

2,628 samples passed technical and genetic quality control and were included in the dataset (Fig. 1), with an average mean depth of coverage of 55x (range 20x to 185x, median coverage of 53x). A total of 669,737 single nucleotide variants (SNVs) and 24,019 short insertions or deletions (indels) were detected, this dataset was denoted Rotterdam Study Exome Sequencing set 2 (RSX2). Of all 669,737 SNVs detected in our RSX2 dataset, 439,633 (66%) were exonic. Of these, 120,677 (27.4%) were not detected in any other public database (ExAC2.0, ESP6500, 1000G, UK10K, DECODE, and GoNL), as shown in Fig. 2. Most of these variants (120,179; 99.6%) were found at a minor allele frequency (MAF) below 1% in our dataset, 65,324 were singletons (54%) and 19,870 were doubletons (17%). The largest overlap with a single dataset was with ExAC2.0 (71% of 439,633 SNVs), followed in descending order by ESP6500 (46%), 1000G (36%), UK10K (34%), GoNL (26%) and DECODE (22%).



**Figure 2.** Overlap of RSX2 with other publically available datasets. Overlap was based on only RefSeq coding SNVs which were detected in at least 1 individual in RSX2 (439,633 SNVs total). The numbers in the Venn diagrams display the number of overlapping SNVs in thousands, the numbers between parenthesis are those SNVs with MAF below 1% (386,341 total). A total of 318,586 SNVs were present in any of the 6 databases (72%). Each individual database yielded a smaller overlap, ranging from 311,017 (Exac, 71%) to 113,627 (GoNL, 26%). Almost all SNVs unique to RSX2 have a MAF < 1% in the RSX2 dataset (120,547; 99.6%).

## Discussion

From 439,633 detected coding variants, 120,179 were absent from all six other population databases. A portion of this absence can be attributed to various biological (ie; ethnical backgrounds, isolated populations or case-series) and technical (whole genome sequencing, exome capturing or filtering strategies and sequencing depth) differences, the remainder is most likely due to population specific variance.

The smallest overlap with DECODE is partly due to the lower sequencing depth and stronger filtering strategy in that dataset, resulting in fewer variants in general. In addition, the genetically isolated status of the Icelandic population warrants fewer genetic variability and smaller overlap with RSX2<sup>(8)</sup>. Despite originating from a similar population, the small overlap with the GoNL database is likely due to its small sample size, reducing power to detect rare variants<sup>(6)</sup>. A larger overlap with UK10K was observed as a result of its large sample size and related population. The differences with the UK10K dataset are largely due to population-specific differences and, the selection of individuals with extreme phenotype in UK10K<sup>(4)</sup>. The 1000G dataset holds many more variants than RSX2, probably caused by whole genome sequencing coverage on coding regions inaccessible by whole exome sequencing, and by the presence of non-Caucasian individuals<sup>(1)</sup>. Similarly, difference in populations and sample size leads to the ESP6500 dataset to be larger than RSX2, although the selection for various case-populations might also be of influence<sup>(2)</sup>. Finally, the greatest dataset of ExAC2.0 contains most variants, as a result of much larger sample size and the inclusion of many different populations<sup>(3)</sup>.

Each dataset present in this comparison contained variants not present in any of the other datasets. These results suggest that, e.g., when filtering or interpreting genetic variants in a WES analysis of a Mendelian disease pedigree, both smaller population-specific datasets (such as RSX2, GoNL, UK10K, and/or deCODE) as well as large aggregation datasets (such as EXAC) contribute information and should be used jointly to filter. Additionally, each database contributes variants not seen elsewhere, suggesting that as many databases as eligible should be considered in these types of analyses. When WES datasets are to be used as controls (e.g., in a case control comparison) note should be taken that some datasets such as UK10K, ESP and EXAC2.0, contain large collections of case-series<sup>(2-4)</sup> and will not provide a good representation of DNA sequence variants of any allele frequency spectrum in the normal population. Given their design and collection strategy, population-based datasets such as RSX2, deCODE and GoNL, might be better suited for this purpose, depending on the disease or trait studied and their estimated prevalence in these databases.

## References

1. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467(7319):1061-73.
2. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*. 2012;337(6090):64-9.
3. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285-91.
4. UK10K WTSI, Hinxton, UK (URL: <http://www.uk10k.org>) [June-2015].
5. Lim ET, Wurtz P, Havulinna AS, Palta P, Tukiainen T, Rehnstrom K, et al. Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet*. 2014;10(7):e1004494.
6. Boomsma DI, Wijmenga C, Slagboom EP, Swertz MA, Karssen LC, Abdellaoui A, et al. The Genome of the Netherlands: design, and project goals. *Eur J Hum Genet*. 2014;22(2):221-7.
7. Genome of the Netherlands C. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet*. 2014;46(8):818-25.
8. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, et al. Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet*. 2015;47(5):435-44.
9. Hofman A, Brusselle GG, Darwish Murad S, van Duijn CM, Franco OH, Goedegebure A, et al. The Rotterdam Study: 2016 objectives and design update. *Eur J Epidemiol*. 2015;30(8):661-708.
10. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2010;26(5):589-95.
11. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-303.
12. <http://broadinstitute.github.io/picard/>) PNPTU.
13. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164.

# Chapter 2.2



## Reduced penetrance of pathogenic ACMG variants in a deeply phenotyped cohort study and evaluation of clinvar classification over time

Jeroen van Rooij<sup>1</sup> (MSc), Pascal Arp<sup>1</sup>(BSc), Linda Broer<sup>1</sup> (PhD), Joost Verlouw<sup>1</sup> (BSc), Frank van Rooij<sup>2</sup> (MSc), Robert Kraaij<sup>1</sup> (PhD), André Uitterlinden<sup>1,2</sup> (PhD), Annemieke J.M.H. Verkerk<sup>1</sup> (PhD)

<sup>1</sup>Department of Internal Medicine, Erasmus Medical Centre, Rotterdam, the Netherlands

<sup>2</sup>Department of Epidemiology, Erasmus Medical Centre, Rotterdam, the Netherlands

*Published on 15<sup>th</sup> of July 2020 in Genetics in Medicine (IF=8.7)  
doi; <https://doi.org/10.1038/s41436-020-0900-8>.*

## Abstract

**Purpose:** We studied the penetrance of pathogenically classified variants in an elderly Dutch population from the Rotterdam study for which deep phenotyping is available. We screened the 59 actionable genes for which reporting of “known” pathogenic variants was recommended by the ACMG, and demonstrate that determining what constitutes as a “known” pathogenic variant can be quite challenging.

**Methods:** We defined known pathogenic as classified pathogenic by both ClinVar and HGMD. In 2,628 individuals, we performed exome sequencing and identified known pathogenic variants. We investigated the clinical records of carriers and evaluated clinical events during 25 years of follow-up for evidence of variant pathogenicity.

**Results:** Out of 3,815 variants detected in the 59 ACMG genes, 17 variants were considered known pathogenic. For 14/17 variants the ClinVar classification had changed over time. Of 24 confirmed carriers of these variants, in only three participants (13%) we observed at least one clinical event possibly caused by the variant.

**Conclusion:** We show that the definition of “known pathogenic” is often unclear and should be approached carefully. Additionally variants marked as known pathogenic do not always have clinical impact on their carriers. Definition and classification of true (individual) expected pathogenic impact should be defined carefully.

**Keywords:** ACMG genes, clinical interpretation, pathogenic variants, exome sequencing, penetrance

## Introduction

Whole Exome Sequencing (WES) is of great value to detect rare, disease-causing genetic variants in affected individuals, and is applied in both diagnostic as well as research settings. However, evaluating whether a variant causes the disease can be challenging, even when this variant is predicted as potentially pathogenic by bioinformatic tools and classified as such in databases as HGMD and/or ClinVar. Increasingly, WES is being applied to large population-based settings with the potential to detect incidental or secondary findings.

Given these developments, the ACMG-AMP (American College of Medical Genetics and Genomics and the Association for Molecular Pathology) has released a set of guidelines on interpretation of genetic variants for clinical interpretation [1].

These guidelines include evidence like variant segregation through the affected individuals' family, previously described presence of other disease-causing variants in the same gene and knowledge of the functional mechanism of this gene in relation to the disease. Variants are classified in five classes based on clinical relevance; 1. Benign, 2. Likely Benign, 3. Uncertain Significance, 4. Likely Pathogenic, 5. Pathogenic [1]. Some databases, like ClinVar, directly follow this classification system [2]. Other databases use their own adaptation of such a classification, such as HGMD [3].

In 2013, Green et al. published a list of 56 genes involving rare monogenetic disorders for which preventive measures and/or treatments were available and recommended reporting to carriers of "incidental or secondary" findings, in clinical exome and genome sequencing data, regardless the diagnostic implication for which the sequencing was ordered [4]. This list was updated by Kalia et al. in 2016, removing one gene and adding four others to a total of 59 genes [5]. However, insufficient knowledge on penetrance of many variants, also in the categories of known pathogenic (KP) or expected pathogenic (EP) variants makes interpretation challenging. Since then various studies have looked into the carrier status of pathogenic gene variants in larger and healthy populations and how pathogenicity scores are defined by different databases [6-10].

Comparing interpretations of 99 variants of different classifications based on the ACMG-AMP guidelines of genetic variants in a Mendelian disease family setting showed a 71% to 92% agreement between 9 clinical laboratories [7]. This indicates that clinical interpretation of genetic variants for the primary outcome (the Mendelian disease segregating in these families) yields similar conclusions for most patients in these diagnostic laboratories. In regard to secondary findings in sequencing datasets from non-family-based sources, investigations of several large population-studies show that between 0.7% and 3.4% of their study population participants carry a KP or EP variant [6, 8-10]. Several of these studies used the list of 56 genes initially reported by Green et al. [9, 10]. Other studies add additional genes considered to have a clear phenotype-genotype relation by clinical genetic specialists, like the 112-114 genes used by Dorschner et al. and Amendola et al. [6, 8]. Most studies reported KP and EP carriers, although Amendola et al. and Jurgens et al. report respectively 0.7% and 0.9% carriers of only KP variants, suggesting almost 1% of the population carries a KP variant in the 56 ACMG genes [6, 9]. Yet, these studies lack an extensive clinical follow-

up with information on health and disease status of the participants. And so, how many of these carriers of KP or EP variants actually have experienced clinically relevant phenotypes due to these variants is not yet clear.

Recent studies have shown that the occurrence of KP variants is higher in the healthy normal population than expected based on the frequency in the Mendelian disease patient-cohorts in which these variants have been originally identified. For example, Minikel *et al.* showed that the prevalence of missense variants in the dominant prion disease gene *PRNP* was 30-fold higher in the general population than expected based on prion's disease prevalence [11]. A similar observation was made for *ASXL1* and other intellectual disability genes by Ropers *et al.* [12]. On a larger scale, Saleheen *et al.* showed that 1,317 genes were predicted to be completely knocked out in at least one of 10,503 adult Pakistani individuals, caused by the large rate of consanguinity in this population, but in many cases without obvious phenotype [13]. Similarly, Lek *et al.* showed that 3,230 genes in their Exome Aggregation Consortium database of 60,706 individuals harbored damaging variants without a currently established disease phenotype [14]. They also showed that each participant carried on average 54 variants that might be considered pathogenic by ClinVar or HGMD, often at higher than expected frequencies, even for homozygous variants in genes for recessive inheritance. Finally, Chen *et al.* identified 13 carriers of severe Mendelian pathogenic variants in a large cohort of nearly 600,000 participants [15], who did not show the expected phenotypes and were considered non-penetrant or resilient to these variants. Results like these show that many potentially pathogenic variants have a lower than expected penetrance in healthy populations and thus should be interpreted with caution.

In our study, we combined WES data with clinical information of 2,628 participants of the longitudinal Rotterdam Study. This is a prospective, population-based cohort study of elderly subjects 45 years and older, living in a suburb of Rotterdam since 1990, and of whom we have almost 30 years of follow-up information from clinical records and detailed physical examination every 4-5 years [16]. In the WES data we evaluated different variant classifications for the 59 ACMG genes, using and comparing ClinVar and HGMD to ascertain known pathogenic variants, and then retrospectively look into the clinical history of carriers to evaluate possible variant pathogenicity and penetrance. Additionally, we analyzed overall changes of variant classification over time in the different database versions of ClinVar, in particular for the identified known pathogenic variants observed in our study population.



## Methods

Details on collection and processing of exome sequencing data from the Rotterdam Study have been described previously [17]. In short, DNA of 2,628 participants was sequenced to an average depth of 56x using NimbleGen SeqCap v2 capture and Illumina's HiSeq2000. Data was processed using BWA, picard, samtools and GATK. Variants were called using GATKs HaplotypeCaller. Variants with a QD < 5 were filtered out. Variants in the 59 ACMG genes were extracted and annotated using Annovar, including Minor Allele Frequency's from the Genome Aggregation Database (GnomAD, Karczewski et al, 2019, unpublished data), CADD (Combined Annotation Dependent Depletion) scores and multiple versions of the ClinVar database, including the most recently available version [2018-03-06] [2, 18]. Variants were annotated to HGMD (v17.3) by batch filtering in the HGMD professional database [3]. No additional filtering was performed based on CADD score or population MAF.

### **Identifying Known Pathogenic variants**

To identify KP variants in our dataset we utilized the largest and most commonly used databases of clinical interpretation of genetic variants; NCBI's Clinical variants database (ClinVar) and the Human Gene Mutation Database (HGMD). We categorized the classifications from both databases for all variants detected in the 59 ACMG genes according to the 5 major classifications outlined in the ACMG-AMP guidelines, to be able to compare classifications in both databases [1]. Specific additional evidence criteria from ClinVar were not assessed at this point.

We added the category for absence from databases with a zero as follows: 0. absent from database, 1. *benign*, 2. *likely/probable benign or likely/probably non-pathogenic*, 3. *unknown, untested or uncertain*, 4. *likely/probable pathogenic* and 5. *pathogenic*. When multiple classifications for the same variant were available in ClinVar, they were averaged (e.g., a 4-4-5 variant is classified as class 4, while a 4- 5- 5 variant is classified as 5). HGMD classifications were coded in a similar manner: 0. absent from database, 3. *NA or Functional Polymorphism (FP)*, 4. *Disease Polymorphism (DP)*, *Disease Functional Polymorphism (DFP)* or *possible Disease Mutations (DM?)*, 5. *Disease Mutations (DM)*. Classes 1 and 2 are not present in HGMD. Variants classified as class 5 in both ClinVar and HGMD were considered KP variants. All KP variants were checked in the latest online ClinVar database (date; April-2020) to confirm the pathogenic classification for the phenotype of which the gene was included in the ACMG recommendations. From this time point, the ClinVar star rating score was extracted for each variant, as well as the number of submissions, as indicated in Table 1.

### **Phenotypic validation of carriers**

Phenotypic events of all study participants are collected weekly by automated linking of the general practitioners' records and diagnoses made by medical specialists, as detailed in the supplemental methods. These events are compared to all medical records, letters from medical specialist and discharge reports. All events were confirmed by trained research assistants. Participants are interviewed about all events at their next study visit [19].

For each KP variant carrier, the events and respective age at event were extracted. For each carrier of a KP variant with an event of interest, four clinicians evaluated the potential causal relationship between the variant and the event, giving consideration to the age at which the event occurred. Ties were broken by the first author. For events marked by a majority all occurrences of this event in the dataset were collected. For each event, the average age at event and the standard deviation were determined. The age at event of the KP carrier was expressed as a z-score, by calculating the number of standard deviations from the average event age across the 2,628 participants with WES data available.

### ***Confirmation by Sanger Sequencing***

All carriers of KP variants classified as class 5 by both ClinVar and HGMD were validated using Sanger sequencing. Primers were designed and produced by Baseclear B.V. (Leiden, The Netherlands). Optimal primer annealing temperature was determined using gradient PCR on control DNA samples. Sanger sequencing of variants in *BRCA1/2* was performed at our department of Clinical Genetics, where these are routinely performed for diagnostic purposes. Sanger sequencing for the other variants was performed by Baseclear B.V. Results were checked manually to verify the variants. Primer sequences and Sanger results are available in supplemental results 1. Variants not confirmed by Sanger sequencing were retained as to not bias further interpretation (two variants in *BRCA2*), as is addressed in the discussion.

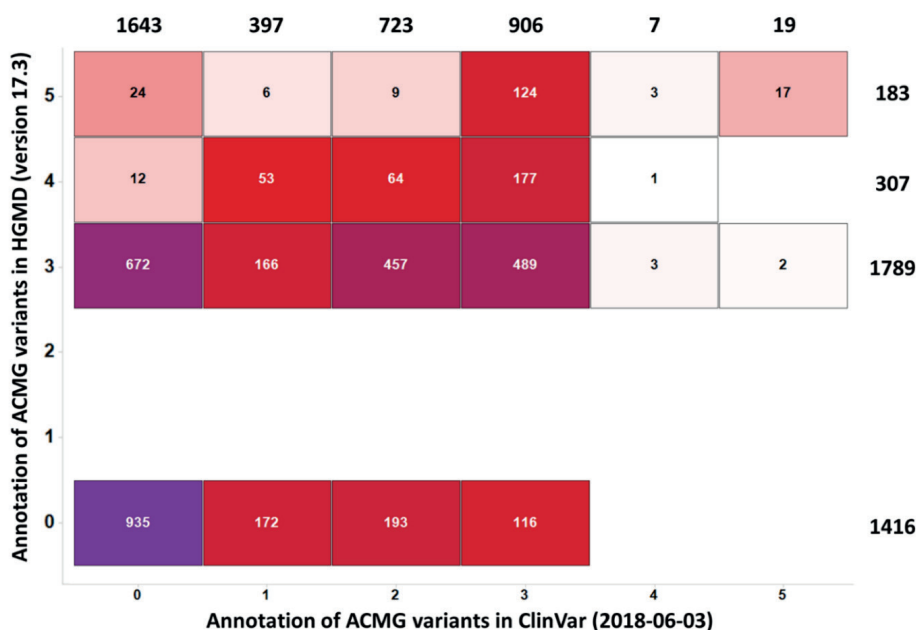
### ***Ethics Statement***

The Rotterdam Study has been approved by the Medical Ethics Committee of the ErasmusMC (registration number MEC 02.1015) and by the Dutch Ministry of Health, Welfare and Sport (Population Screening Act WBO, license number 1071272-159521-PG). This study has been entered into the Netherlands National Trial Register ([www.trialregister.nl](http://www.trialregister.nl)) and into the WHO International Clinical Trials Registry Platform ([www.who.int/ictcp/network/primary/en/](http://www.who.int/ictcp/network/primary/en/)) under shared catalogue number NTR6831. All participants provided written informed consent to participate in the study and to have their information obtained from treating physicians.

## Results

### Identification of known pathogenic variant carriers

Exome sequencing was performed on 2,628 Rotterdam Study (RS) participants and after filtering and QC resulted in a total of 703,990 genomic variants, as was previously described [17]. Of these, 3,815 variants were located in one of the 59 ACMG genes [5]. All these 3,815 variants were classified using both the HGMD and ClinVar databases, resulting in 6 classes: 0 (absent from database), 1 (benign), 2 (likely benign), 3 (uncertain), 4 (likely pathogenic) or 5 (pathogenic) per database.



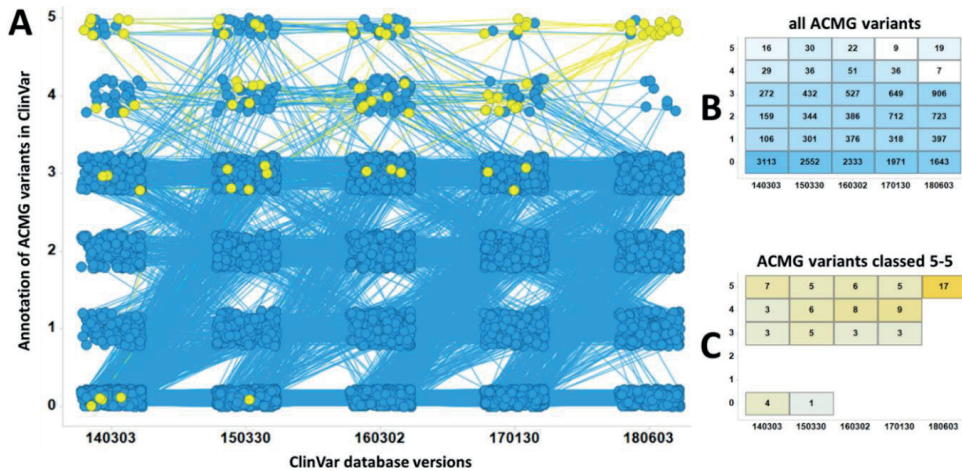
**Figure 1.** classification of clinically relevant variants in 2,628 Rotterdam study participants in the 59 ACMG genes according to ClinVar version 2018 and HGMD. Classes are defined as per the ACMG-AMP guidelines; 1. Benign, 2. Likely benign, 3. Uncertain, 4. Likely pathogenic, 5. Pathogenic. Variants absent from the database are coded as 0. The classifications for HGMD were converted to: NA (class 3), FP (class 3), DP (class 4), DFP (class 4), DM? (class 4) and DM (class 5). For visualization purposes, the variants observed in autosomal recessive genes *ATP7B* and *MUTYH* are not shown. The numbers at the sides are sums for that respective classification.

The 3,815 variants were classified and grouped according to this system as indicated in Figure 1, comparing their classification in both databases. The 119 variants in autosomal recessive genes *MUTYH* or *ATP7B* were excluded from this figure and analyzed separately. Of the resulting 3,696 variants, 935 variants (25%) were absent from both databases. An additional 708 variants (19%) were present in HGMD but not in ClinVar and another 481 variants (13%) were present in ClinVar but not in HGMD. Thus, the remaining 1,691 variants (43%) were classified by both databases. Furthermore, HGMD classifies 183 of these variants (5%) as pathogenic (class 5) versus only 19 by ClinVar (0.5%). In total 17 variants are classified as pathogenic by both of the databases (0.5% of all variants), and are here defined as known pathogenic (KP) variants. In total, 24 participants were confirmed by Sanger validation to carry one of these 17 KP variants (0.9% of all participants). An additional 2 carriers of a single variant in *BRCA2* were identified, but were found to be false positives by Sanger validation. These variants were retained as not to bias further interpretation, but carefully marked in subsequent tables.

Additionally, 8 of the 119 variants in *MUTYH* and *ATP7B* were classified as pathogenic by both HGMD and ClinVar (not shown), but only as autosomal recessive inheritance, thus in homozygous state. In total 50 carriers were observed for any of these 8 variants, all in a heterozygous state. No compound heterozygosity was detected. Heterozygous variants in these genes were not considered as KP and thus they were not followed up further.

### ***Variation in ClinVar clinical classification over time***

We have downloaded ClinVar database versions from the years 2014 until 2018. For HGMD the most recent online version was used (v17.3). Comparing the clinical classification for the 3,815 ACMG variants identified in our study population between ClinVar database versions shows that classification largely changes over time, as shown in figure 2. Firstly, in 2014 only 582 variants were present in ClinVar (16%), versus 2,052 in 2018 (56%), a 3.5-fold increase. This increase was most notable for variants of class 1; benign (3.7-fold increased), class 2; likely benign (4.5-fold increased) and class 3; uncertain significance (3.3-fold increased). Whereas class 5; pathogenic remained almost unchanged (1.2-fold increase) and class 4; likely pathogenic decreased 4,1-fold decrease). The migration of classification for the 17 known pathogenic variants (as classified in version 2018) is marked separately in figure 2. As shown, only between 5-7 of these 17 KP variants were classified as pathogenic at the same time at any given ClinVar version in the previous years. In fact, only 3 of the 17 KP variants remained at class 5 in all tested previous versions of ClinVar. The classification per variant per ClinVar version is indicated in table 1. All variants were confirmed pathogenic at the online version of ClinVar (dated April-2020). Five of the 17 variants received a three star score in ClinVar (reviewed by expert panel), 10 received a two star score (multiple submitters, no conflicting interpretation). A single variant received a one star score (multiple submitters, conflicting interpretation), and one variant received a zero star score (no assertion criteria provided).



**Figure 2.** A) Classification of all variants detected in one of the 59 ACMG genes in 2,628 participants of the Rotterdam Study population according to ClinVar at different time points: March-2014 (date 140303), March-2015 (date 150330), March-2016 (date 160302), January-2017 (date 170130) and June-2018 (date 180603). Each variant is connected by a line between all five versions. Marked in yellow are the 17 known pathogenic variants classified as category 5 by the most recent versions of ClinVar (version 180603) and HGMD (version 17.3). B) the number of variants in each class of each ClinVar database version. C) the class at each database version for the 17 variants that were classified as 5 in ClinVar in 2018 and by HGMD 17.3 (marked yellow in figure A). For visualization purposes, the variants observed in autosomal recessive genes *ATP7B* and *MUTYH* are not shown.

### ***Phenotypic evaluation of known pathogenic carriers***

We extracted 94 ICD10-coded clinical events for the 26 KP carriers, from 9,165 coded clinical events across our 2,628 study participants, in addition to the age at each event, shown in figure 3. In total 18 events (20%) in 10 different individuals were marked by at least one clinical referee as possibly related to the KP variant. Nine events (10%) in 3 carriers (indicated with an asterisk in figure 3) were marked by at least three referees.

### ***Frequency of ICD10 events in entire study population***

Nine ICD10-coded clinical events in three carriers were considered linked to the detected variant. For each we calculated the prevalence and average age in the rest of the Rotterdam study population for which we have WES data available (n=2,628) [17]. The results for these nine events are shown in supplemental table 3. All events occurred commonly in this population; I20:angina pectoris (in 4.9% of the 2,628 participants, average age of the event is 72±8), I21:myocardial infarction (10.5%, average age 79±8), I46:cardiac arrest (4.6%, average age 81±8), I48:atrial fibrillation (19.8%, average age 77±10), I50:heart failure (24.9%, average age 80±8) and R99:death with cause unknown (6.3%, average age 87±7). For all events selected by the referees the age at event was earlier than the average age at event across the 2,628 participants for which WES data was available, although all events fell within 1.5 standard deviation.

## Discussion

From 3,815 variants that we found in 59 reported ACMG genes in WES data of 2,628 participants from the Rotterdam study, we confirmed 24 participants to carry a total of 17 “known” pathogenic (KP) variants, comprising 0.9% of our study population. Two additional carriers of a single variant in *BRCA2* were identified, but this variant proved false positive after Sanger validation, despite passing all exome sequencing QC and filtering criteria. Upon investigation, the variant was supported by a small number of reads and would have been filtered out in single-sample data processing (i.e., the fact of two putative carriers strengthened the variant quality in calling). Thus, this result indicates we should be careful in the way we handle and interpret this kind of data. Validation by Sanger sequencing in our case was required for a reliable result. This is in line with previous findings, where <2% of all variants identified through WES could not be confirmed, and variants of high clinical relevance should be confirmed beyond doubt [20, 21].

The proportion of 0.9% KP carriers is similar to what was found in previous studies [6, 8-10]. Upon investigation by four clinicians, 10 variant carriers (out of 26) were observed with at least one ICD10-coded clinical events deemed possibly related to their KP variant, according to at least one of the referees. Only in three carriers (13%), at least one clinical event was considered to be related to the identified variant by a majority of the referees. In all of these carriers it was difficult to determine if the ICD10-based clinical events were caused by these variants, as these events occur frequently in the population. As a result, no information was reported back to any of the carriers or their relatives.

Gene	Carrier	Sanger	age 60-64	age 65-69	age 70-74	age 75-79	age 80-84	age 85-89	age 90+	Associated Diseases (ACMG)
RET	A1	+				H40	F00 S72	F00		Multiple Endocrine Neoplasia Type 2
PTEN	B1	+	G20	G20						PTEN Hamartoma Tumor Syndrome
KCNQ1	C1	+				S52	S22 S32 <b>I48*</b>			Romano-Ward Long QT Syndromes Types 1, 2, and 3, Brugada Syndrome
KCNQ1	<b>D1*</b>	+	<b>I48*</b>				<b>I46*</b>			
MYBPC3	E1	+		<b>I63*</b>	H25 S22		<b>I50*</b> <b>I50*</b>			
MYL2	F1	+					D47 S22	H25	F00 J44	
	F2	+					S52	C45 C45		
	F3	+				C67	M96			
	<b>F4*</b>	+		C61 C67	<b>I21*</b> <b>I21*</b> <b>I21*</b> <b>I21*</b>			<b>R99*</b>		Hypertrophic cardiomyopathy, Dilated cardiomyopathy
	F5	+	<b>I64*</b>		H25					
	F6	+		S32	S32 S52 S95 M96 S72					
	<b>F7*</b>	+		<b>I50*</b> <b>I20*</b>		C50 I25*				
MYL2	G1	+						<b>I64*</b> F01	F01	
BRC2A	H1	+					J15			
BRC2A	I1	+		S52						
BRC2A	J1	-			I21	I63 I64 S62				Hereditary Breast and Ovarian Cancer
	J2	-			<b>C66*</b> G45		I50 R99			
BRC1A	K1	+						C44 H35	C44 G45 R99	
DSC2	L1	+			S52	H35		S72		Arrhythmic right ventricular cardiomyopathy
DSG2	M1	+					H25 H25	H35	S22 G45 G45 G45 G45 <b>I96*</b>	Familial hypercholesterolemia
LDLR	N1	+		I80						
RYR1	O1	+		H35	<b>I64*</b>	H25	I61			
	O2	+				C18 C19				
	O3	+						C34 C34		Malignant hyperthermia susceptibility
RYR1	P1	+				F00 C18	C18 I26 I80			
RYR1	Q1	+		H40 C34 C34						

**Figure 3.** 26 carriers of 17 KP variants, one shown on each line. The column “Sanger” denotes confirmed (+) (24 samples) or unconfirmed (-) (2 samples) by Sanger sequencing. For each carriers their recorded clinical events are displayed in 5-year intervals. The events are coded using the ICD10 classification system. The last column denotes the primary disease for which the gene was included in the ACMG recommendations. Events marked with a “++” are evaluated by at least 3 of the 5 referees (3 of 4 clinicians or 2 clinicians and the first author) as possibly explained by the variant for which the patient was a carrier. Those carriers are marked by an asterisk and shown in bold. Events marked with a single “+” were marked by only 1 or 2 referees. ICD10 codes in alphabetical order; Neoplasm of C18:colon, C19:rectosigmoid junction, C34:bronchus, C44:skin, C45:mesothelioma, C50:breast, C61:prostate, C66:ureter, C67:bladder. D47:other neoplasm of uncertain behavior. F00:Alzheimer’s disease, F01:Vascular dementia, G20:Parkinson’s disease, G45:transient ischemic attack. H25:cataract, H35:retinopathy, H40:glaucoma. I20:angina pectoris, I21:myocardial infarction, I25:ischemic heart disease, I46:cardiac arrest, I48:atrial fibrillation, I50:heart failure, I61:intercerebral hemorrhage, I63:cerebral infarct, I64:stroke, I80;deep vein thrombosis. J15;pneumonia, J44:chronic obstructive pulmonary disease, J96:respiratory failure. M96:postprocedural skeletal disorder. R99:death of unknown cause. Fractures of; S22:rib, S32:lumbar spine, S52:forearm, S62:wrist, S72:femur, S92:foot.

We consulted two main databases for clinical interpretation; HGMD and ClinVar [2, 3]. Comparing their clinical classification for the ACMG variants identified in our study population we observed disagreement in which variants are classified as pathogenic. In total 17 variants were categorized as class 5 by both databases, 19 in total by ClinVar and 183 in total by HGMD.

Of concern is a large portion of classifications which differ between both databases, such as the 59 variants classified as class 4 or 5 (likely pathogenic or pathogenic) in HGMD and class 1 (benign) in ClinVar. These most likely stem from over-estimation of pathogenicity of HGMD, as has been described before [22, 23]. This disagreement illustrates the challenge of clinically interpreting genetic variants, especially in a research setting and how different individuals, laboratories or databases might reach different conclusions for the same variant. Even when restricting to variants classified as class 5 in both databases, it appears that such variants can be carried without obvious phenotypic consequence.

Additionally, we investigated the clinical classification within ClinVar in different releases over five years (from 2014-2018). We observe that the clinical interpretation of many variants



has changed over time, where many variants moved towards class 1 (benign), 2 (likely benign) or 3 (uncertain significance). Over this period various genomic variant resources have surfaced and impacted variant interpretation, including the gnomAD database which now contains data from 125,748 exomes and 15,708 whole genomes from population studies. Additionally the ACMG/AMP criteria were released during this time frame and has influenced how consistently labs were applying evidence. One example of this is the reclassification for *BRCA1* and *BRCA2* variants over time, most often “downgraded” [24, 25]. Traditionally the classification of (pathogenic) variants was based on the ascertainment from the more severe Mendelian disorders. Now, with more data available from population studies reduced penetrance of variants is becoming clearer as is demonstrated by these kind of variants found in individuals without a Mendelian phenotype [11-14, 26]. By including information about penetrance in healthy populations, the changes in variant classification may stabilize over time.

Although ClinVar contributes greatly to centralizing publicly available clinical genetic information, it does not contain local databases maintained by clinical genetic laboratories. This could result in classification differences of variants between laboratories, and may challenge research efforts to utilize clinical genetic classifications by the more conservative ACMG-AMP criteria. Thus, our definition of a KP variant may be less stringent than used by a clinical genetic laboratory. Furthermore, several of the variants we indicated as KP have limited information available in ClinVar. At the most recently checked online version (April-2020) two variants had a star classification of less than 2. Five additional variants had only one or two submissions in ClinVar at this time. These results demonstrate the need for additional clinical genetic information to completely classify such variants. Nevertheless, we have attempted to retain the most likely true pathogenic variants as possible using publicly available information. We believe that most of these variants would retain their pathogenic classifications under ACMG-AMP evaluation in clinical genetic laboratories. However, it is possible that the percentage of carriers (0.9%) and fraction of expressivity in these carriers (13%) is lower than under complete clinical genetic evaluation.

For the clinical evaluation of our KP carriers we used the ICD10-coded records that report clinical events during standard clinical practice and during Rotterdam Study research participation. We collected 9,165 ICD10-coded events for 2,628 study participants, providing unique insight into the health state of such a typical elderly population. In 0.9% of this population we observed a KP variant, but only 13% of these carriers (0.13% of the whole study population) presented an ICD10-coded event that could be related to the variant. For none of them this effect was obvious. Due to these results, no events were reported back to any of these carriers, and thus we were not able to collect additional, more detailed, phenotypic information.

Our study demonstrated that the definition of a KP variant is ambiguous between databases, but also within different versions of the same database. This might lead to differences in reporting depending on the used evidence for classification. Specifically, information on the occurrence of KP variants in healthy populations is needed to correctly estimate the penetrance of such variants, and this information should be considered in the recommendations. Currently, several studies have demonstrated that approximately 1% of



the population carries a KP defined as such by different databases. Our results based on a thorough clinical follow-up evaluation in subjects 55 years and older linked only 0.13% of events to the presence of a KP variant. This suggests that KP variants are less likely to lead to a phenotype in their carriers, and that such reduced penetrance should be considered when reporting back results to carriers in population-based studies. Overall, our results indicate that reporting back of pathogenic ACMG variants should be approached carefully in these kind of studies.

Several causes for the reduced penetrance could play a role in our population. First, our study population is an elderly population, in which carriers reached late adulthood (55 years or older) despite carrying a potentially pathogenic variant [16]. Therefore, our population contains survival bias and the penetrance of some of these variants might be higher in younger populations. Additionally, these participants were investigated in a research setting, and despite the rigorous phenotype collection in the Rotterdam Study they may have exhibited subtle clues missed during examination, such as subclinical deviations or specific relevant family history, which is often used in ACMG-AMP evaluation but could not be collected in this setting. Conversely, this dataset is representative for many hospital populations in which (secondary) genetic testing is most likely to occur [16]. Secondly, the expected penetrance is not standardly included in the classification of a pathogenic variant. Thus, variants in class 5 can have variable penetrance and those variants we observe in an elderly research population are likely those with lower penetrance. Considering penetrance on top of the five-class system might facilitate more accurate interpretation. Thirdly, such severely reduced penetrance of KP variants in population-based settings could indicate a strong influence of the genomic context of the functional effects of KP variants in such normal healthy population-dwelling subjects. While in Mendelian disease families the penetrance is usually substantially higher, also here penetrance can be variable and also here the genomic context might play a role due to the complex way in which different inherited variants or modifiers can influence the phenotype [27].

### **Conclusion**

We show that the definition of “known pathogenic” is often not clear and should be approached carefully. Variants marked as KP may have (severely) reduced penetrance. Definition and classification of true (individual) expected pathogenic impact should include, for example, the use of multiple data sources, the pathogenicity prediction over time, and an assessment of the penetrance of the variant in healthy control populations.

### **Acknowledgements**

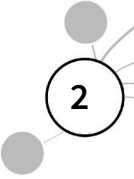
We thank the participants of the ERGO population study for their participation in this research, Emma van de Ende, Merel Mol, Eline van der Valk and Anela Blazevic for interpretation of clinical events in variant carriers and Mila Jhamai, Joost Verlouw and Marijn Verkerk for their help in generating the exome sequencing dataset. We thank Jolande Verkroost-van Heemst for coordinating clinical follow-up data collection and Joyce van Meurs for supporting the project. We thank Sergio Chavez, Wout Deelen and Joan Kromosoeto for supporting and performing the Sanger sequencing experiments.



## References

1. Richards, S., et al., *Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology*. *Genet Med*, 2015. **17**(5): p. 405-24.
2. Landrum, M.J., et al., *ClinVar: improving access to variant interpretations and supporting evidence*. *Nucleic Acids Res*, 2018. **46**(D1): p. D1062-D1067.
3. Stenson, P.D., et al., *The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies*. *Hum Genet*, 2017. **136**(6): p. 665-677.
4. Green, R.C., et al., *ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing*. *Genet Med*, 2013. **15**(7): p. 565-74.
5. Kalia, S.S., et al., *Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics*. *Genet Med*, 2017. **19**(2): p. 249-255.
6. Amendola, L.M., et al., *Actionable exomic incidental findings in 6503 participants: challenges of variant classification*. *Genome Res*, 2015. **25**(3): p. 305-15.
7. Amendola, L.M., et al., *Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium*. *Am J Hum Genet*, 2016. **98**(6): p. 1067-1076.
8. Dorschner, M.O., et al., *Actionable, pathogenic incidental findings in 1,000 participants' exomes*. *Am J Hum Genet*, 2013. **93**(4): p. 631-40.
9. Jurgens, J., et al., *Assessment of incidental findings in 232 whole-exome sequences from the Baylor-Hopkins Center for Mendelian Genomics*. *Genet Med*, 2015. **17**(10): p. 782-8.
10. Olfson, E., et al., *Identification of Medically Actionable Secondary Findings in the 1000 Genomes*. *PLoS One*, 2015. **10**(9): p. e0135193.
11. Minikel, E.V., et al., *Quantifying prion disease penetrance using large population control cohorts*. *Sci Transl Med*, 2016. **8**(322): p. 322ra9.
12. Ropers, H.H. and T. Wienker, *Penetrance of pathogenic mutations in haploinsufficient genes for intellectual disability and related disorders*. *Eur J Med Genet*, 2015. **58**(12): p. 715-8.
13. Saleheen, D., et al., *Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity*. *Nature*, 2017. **544**(7649): p. 235-239.
14. Lek, M., et al., *Analysis of protein-coding genetic variation in 60,706 humans*. *Nature*, 2016. **536**(7616): p. 285-91.
15. Chen, R., et al., *Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases*. *Nat Biotechnol*, 2016. **34**(5): p. 531-8.
16. Ikram, M.A., et al., *The Rotterdam Study: 2018 update on objectives, design and main results*. *Eur J Epidemiol*, 2017. **32**(9): p. 807-850.
17. van Rooij, J.G.J., et al., *Population-specific genetic variation in large sequencing data sets: why more data is still better*. *Eur J Hum Genet*, 2017. **25**(10): p. 1173-1175.
18. Rentzsch, P., et al., *CADD: predicting the deleteriousness of variants throughout the human genome*. *Nucleic Acids Res*, 2019. **47**(D1): p. D886-D894.
19. Leening, M.J., et al., *Methods of data collection and definitions of cardiac outcomes in the Rotterdam Study*. *Eur J Epidemiol*, 2012. **27**(3): p. 173-85.
20. Beck, T.F., et al., *Systematic Evaluation of Sanger Validation of Next-Generation Sequencing Variants*. *Clin Chem*, 2016. **62**(4): p. 647-54.

21. Lincoln, S.E., et al., *A Rigorous Interlaboratory Examination of the Need to Confirm Next-Generation Sequencing-Detected Variants with an Orthogonal Method in Clinical Genetic Testing*. *J Mol Diagn*, 2019. **21**(2): p. 318-329.
22. Cassa, C.A., M.Y. Tong, and D.M. Jordan, *Large numbers of genetic variants considered to be pathogenic are common in asymptomatic individuals*. *Hum Mutat*, 2013. **34**(9): p. 1216-20.
23. Kundu, K., et al., *Determination of disease phenotypes and pathogenic variants from exome sequence data in the CAGI 4 gene panel challenge*. *Hum Mutat*, 2017. **38**(9): p. 1201-1216.
24. Mighton, C., et al., *Correction: Variant classification changes over time in BRCA1 and BRCA2*. *Genet Med*, 2019.
25. Mighton, C., et al., *Variant classification changes over time in BRCA1 and BRCA2*. *Genet Med*, 2019.
26. Narasimhan, V.M., et al., *Health and population effects of rare gene knockouts in adult humans with related parents*. *Science*, 2016. **352**(6284): p. 474-7.
27. Deltas, C., *Digenic inheritance and genetic modifiers*. *Clin Genet*, 2018. **93**(3): p. 429-438.





# Chapter 2.3



## *EIF2AK3* variants in Dutch patients with Alzheimer's disease

Tsz Hang Wong, Sven J<sup>1</sup>. van der Lee<sup>2</sup>, Jeroen G.J van Rooij<sup>3</sup>, Lieke H.H. Meeter<sup>1</sup>, Petra Frick<sup>4</sup>, Shami Melhem<sup>1</sup>, Harro Seelaar<sup>1</sup>, M. Arfan Ikram<sup>2</sup>, Annemieke J. Rozemuller<sup>5</sup>, Henne Holstege<sup>6</sup>, Marc Hulsmans<sup>7</sup>, Andre Uitterlinden<sup>8</sup>, Manuela Neumann<sup>9</sup>, Jeroen J.M. Hoozemans<sup>5</sup>, Cornelia M. van Duijn<sup>2</sup>, Rosa Rademakers<sup>10</sup> and John C. van Swieten<sup>11</sup>

<sup>1</sup> Alzheimer Center and Department of Neurology, Erasmus Medical Center, Rotterdam, The Netherlands.

<sup>2</sup> Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands.

<sup>3</sup> Alzheimer Center and Department of Neurology, Erasmus Medical Center, Rotterdam, The Netherlands; Department of Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands.

<sup>4</sup> DZNE, German Centre for Neurodegenerative Disease, Tübingen, Germany.

<sup>5</sup> Department of Pathology, VU University Medical Center, Amsterdam, The Netherlands.

<sup>6</sup> Alzheimer Center, Department of Neurology, VU University Medical Center, Amsterdam, The Netherlands; Department of Clinical Genetics, VU University Medical Center, Amsterdam, The Netherlands.

<sup>7</sup> Alzheimer Center, Department of Neurology, VU University Medical Center, Amsterdam, The Netherlands; Department of Clinical Genetics, VU University Medical Center, Amsterdam, The Netherlands; Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands.

<sup>8</sup> Department of Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands.

<sup>9</sup> DZNE, German Centre for Neurodegenerative Disease, Tübingen, Germany; Department of Neuropathology, University of Tübingen, Tübingen, Germany.

<sup>10</sup> Department of Neuroscience, Mayo Clinic Florida, Jacksonville, FL, USA.

<sup>11</sup> Alzheimer Center and Department of Neurology, Erasmus Medical Center, Rotterdam, The Netherlands; Alzheimer Center, Department of Neurology, VU University Medical Center, Amsterdam, The Netherlands.

*Published in Neurobiology of Aging (IF=4.4), Jan-2019 doi: 10.1016/j.neurobiolaging.2018.08.016*

## Abstract

Next generation sequencing has contributed to our understanding of the genetics of Alzheimer's disease (AD), and has explained a substantial part of the missing heritability of familial AD. We sequenced 19 exomes from 8 Dutch families with a high AD burden, and identified *EIF2AK3*, encoding for protein kinase RNA-like endoplasmic reticulum kinase (PERK), as a candidate gene. Gene based burden analysis in a Dutch AD exome cohort containing 547 cases and 1070 controls showed a significant association of *EIF2AK3* with AD (OR 1.84 [95% CI 1.07-3.17],  $p$ -value 0.03), mainly driven by the variant p.R240H. Genotyping of this variant in an additional cohort from the Rotterdam study showed a trend towards association with AD ( $p$ -value 0.1). Immunohistochemical staining with pPERK and peIF2 $\alpha$  of three *EIF2AK3* AD carriers showed an increase in hippocampal neuronal cells expressing these proteins compared to non-demented controls, but no difference was observed compared to AD non-carriers. This study suggests that rare variants in *EIF2AK3* may be associated with disease risk in AD.

## Introduction

Alzheimer's disease (AD) is the most common cause of dementia, characterized by progressive decline in memory and other cognitive functions.<sup>1</sup> Genetic factors are strongly linked to AD, and in about 5% of cases an autosomal dominant mode of inheritance has been reported.<sup>2</sup> In autosomal dominant forms of early-onset AD, mutations in  $\beta$ -amyloid precursor protein (*APP*), presenilin 1 (*PSEN1*) and presenilin 2 (*PSEN2*) have been found to be causative genes;<sup>3-7</sup> this accounts for approximately 13% of early-onset AD.<sup>8</sup> In late-onset AD, the  $\epsilon 4$  allele of apolipoprotein E gene have been found to be the most common risk factor.<sup>9</sup>

Neuropathologically, the aggregation of misfolded proteins is a major hallmark of many neurodegenerative disorders.<sup>10</sup> The accumulation of extracellular amyloid plaques and intracellular neurofibrillary tangles are the hallmarks of AD.<sup>11</sup> Previous studies suggest that disrupted protein homeostasis in the endoplasmic reticulum (ER) and activation of unfolded protein response (UPR) may be major drivers in AD pathogenesis.<sup>10, 12</sup> The UPR is induced by three transmembrane proteins in the ER: protein kinase RNA-like endoplasmic reticulum kinase (PERK), Inositol Regulating Enzyme 1 (IRE1) and Activating Transcription Factor 6 (ATF6). Activation of UPR lead to transient suppression of protein synthesis and increased expression of genes aimed to restore the homeostasis of the ER.<sup>10</sup> Pharmacological and genetic manipulation of the UPR pathways in animal studies, in particularly the PERK pathway, has been reported to inhibit neurodegeneration.<sup>13</sup>

Advances in next generation sequencing technology have contributed substantially to our understanding of the genetics of AD. In recent years, studies using whole exome sequencing (WES) and whole genome sequencing reported the association of rare variants in *PLD3*, *ABCA7*, *TREM2* and *SORL1* with an increased risk in AD.<sup>14-18</sup> Furthermore, a large exome micro-array study identified rare coding variants in *PLCG2*, *ABI3* and *TREM2*, explaining a small part of missing heritability in AD.<sup>19</sup> These studies indicate the existence of other rare variants related to the heritability of AD.

In this paper, we performed WES in eight Dutch AD families with probable autosomal dominant inheritance, and identified Eukaryotic Translation Initiation Factor 2 Alpha Kinase 3 (*EIF2AK3*), encoding for PERK, as a candidate AD risk gene in two of these families. Together with previous reports on an increased activation of PERK in AD brain and the involvement of PERK in memory and learning,<sup>20</sup> these findings suggest the possible role of *EIF2AK3* in the pathogenesis of AD.



## Methods

### **Subjects**

Our discovery dataset included 19 AD patients from eight Dutch families with a high AD burden. Each family had at least two patients with AD suggestive of an autosomal dominant inheritance pattern, except one family with an uncertain mode of inheritance due to the early death of both parents. The mean age at disease onset in the families varied from 62.5 to 71.3 years (Table 1). Non-demented first and second-degree family members of each family were also included if available. Using WES, all patients were screened negative for mutations in *PSEN1*, *PSEN2* and *APP*; *APP* copy number mutations were also excluded. For WES, we included DNA samples of at least two patients with AD from each family. Non-demented family members with a minimum age of 65 were used to test for segregation in their respective family.

Patients and family members were recruited after referral to the department of Neurology in the Erasmus Medical Center, or after visiting (nursing) homes. Diagnosis of probable AD was confirmed in all patients according to the National Institute of Neurological and Communicative Disorders and Stroke-Alzheimer's Disease and Related Disorders Association criteria for AD.<sup>21</sup>

To replicate the association of our candidate gene with AD, we used exome data available from 547 AD cases and 1070 controls from three different sites (the Rotterdam Study, Amsterdam Dementia Cohort (ADC-VUmc) and Alzheimer Centrum Erasmus MC (AC-EMC)) included from a Dutch AD exome dataset, previously described by Holstege et al.<sup>18</sup> We then genotyped our candidate variant in 1055 AD cases and 6162 controls from the Rotterdam study;<sup>22</sup> any individuals from the Rotterdam Study included in the exome data were excluded for genotyping.

Our study has been approved by the Medical Ethical Committee of Erasmus Medical Center, and written informed consent was obtained from all participants or their legal representatives.



**Table 1.** Baseline characteristics of the families

Family	Cases	Controls	WES cases	Mean age at onset (range)	Mean age at last visits controls (range)	% female	APOE fraction $\epsilon 2/ \epsilon 3/ \epsilon 4$
NLAD 1	5	8	3	70.4 (60-89)	69.1 (65-77)	46.2	0.2/0.5/0.3
NLAD 2	2	2	2	62.5 (52-73)	69.0 (68-70)	50.0	0/0.25/0.75
NLAD 3	5	2	3	71.3 (68-77)	78.5 (71-86)	71.4	0/0/1
NLAD 4	5	1	2	62.8 (59-65)	66.7 (61-72)	57.1	0/0.12/0.88
NLAD 5	2	0	2	66.0 (NA)	NA	0.0	0/0.75/0.25
NLAD 6	3	1	3	67.7 (64-70)	69.0 (NA)	75.0	0/0.67/0.33
NLAD 7	2	3	2	71.0 (66-76)	73.3 (69-78)	20.0	0/0/1
NLAD 8	2	4	2	64.5 (59-70)	71.4 (70-73)	50.0	0/1/0

Number of patients and controls included from each family. Cases are the total number of included patients with Alzheimer's disease and patients with mild cognitive impairment. Controls contains the total number of included individuals without subjective or objective memory impairment during the last visit. Age at onset is the mean age of first disease onset of all included cases, and the age at last visits is the mean age of all included controls. Age at onset and age at last visits in years. AD, Alzheimer's disease; WES, individuals selected for whole exome sequencing; NA, not available

### **Whole exome sequencing analysis**

Exomes of 19 AD patients from the discovery set, the Rotterdam Study cohort, and the AC-EMC cohort were captured using Nimblegen Seqcap EZ Exome Capture Kit v2. Exomes from the ADC-VUmc cohort were captured using the Nimblegen SeqCap EZ Exome capture kit v3. All data were generated at the Human Genomics Facility (HuGeF; www.glimdna.org) at Erasmus MC Rotterdam, the Netherlands. DNA from each sample was prepared with the Illumina TruSeq Paired-End Library Preparation Kit, and 100 base-pair paired end reads were acquired by sequencing the libraries on a HiSeq 2000. For the Dutch exome dataset, we used the overlapping regions between capture kits during calling of the data. Sequencing reads were aligned to the hg19 human genome assembly using BWA-MEM (version 0.7.3a)<sup>23</sup>, and Picard Tools (version 1.9)<sup>24</sup> were used to mark duplicates and to sort the alignments. Subsequently, Genome Analysis Toolkit (GATK) (version 3.3) was used to perform indel realignment and base quality score recalibration.<sup>25</sup> Haplotype Caller from GATK was used to create gVCF files, and to call variants from these gVCF files. For the exome data from the eight families (discovery set), we used hard filters according to GATK best practices to filter out low quality variants. For the exome data from the three Dutch cohorts, we used Variant Quality Score Recalibration (VQSR) with >99% sensitivity to filter out low-quality variants. Subsequently, Plink was used to calculate Principle component (PC), and outliers on the first two PCs were removed.<sup>26</sup> Related individuals with identity by decent value > 0.1 were also removed from the analysis set. All individuals in the WES data were checked for sex concordance using Plink<sup>26</sup>. Variants from all datasets were annotated using ANNOVAR.<sup>27</sup>

In our discovery set, we used a family based analysis to identify candidate genes from the eight families. Each family was analyzed separately to identify the candidate variants in their respective family. We focused on shared variants among the affected family members which

resulted in an amino acid change. Subsequently, variants with a frequency of 0.5% or lower in 1000 genomes, NHLBI Exome Sequencing Project (ESP), Exome Aggregation Consortium (ExAC), Genome of the Netherlands, and in-house WES data from the Rotterdam Study were selected (Supplementary Table 1).<sup>28-32</sup> If the same variant or different variants in the same gene were identified in at least two families, these variants were selected as candidates for follow up and tested with Sanger sequencing for segregation in their respective families.

### **Sanger sequencing**

We used Primer 3<sup>33</sup> to design primers for candidate variants. PCR amplification was performed using Qiagen Taq DNA polymerase (Qiagen, CA, USA). Direct sequencing of PCR products was performed using Big Dye Terminator chemistry ver. 3.1 (Applied Biosystems), and run on an ABI3130 genetic analyzer and an ABI3730xl genetic analyzer (Applied Biosystems, CA, USA). The sequences were analyzed with Sequencher software, version 4.5 (Genecodes, VA, USA) and Seqscape version 2.6 (Applied Biosystems, CA, USA).

### **Genotyping of rs147458427 variant in EIF2AK3**

The variant rs147458427 (p.R240H) was genotyped using TaqMan SNP Genotyping Assays and genotypes of rs147458427 were determined using TaqMan Allelic discrimination. Signals were read with the Taqman 7900HT (Applied Biosystems Inc.) and analyzed using the Sequence Detection System 2.4 software (Applied Biosystems Inc.). To evaluate genotyping accuracy, all heterozygous calls were typed twice to confirm genotypes. Single variant association effects for AD association were calculated using R (version 3.2.3) “seqMeta” tool v.1.6.0 adjusting for gender. *APOE* status was added as covariate in the secondary analysis.

### **Statistical analysis of the candidate genes in the Dutch exome dataset**

Single variant association effect for AD association was calculated using R (version 3.2.3) “seqMeta” tool v.1.6.0 adjusting for gender. Burden test was calculated for our top candidate gene in the family-based analysis using burdenMeta function in “seqMeta” tool v.1.6.0. Only variants with minor allele frequency (MAF)  $\leq 1\%$  in ExAC was included in the burden test, adjusting for gender. In the secondary analysis, we performed these analyses on our top candidate gene, adjusting for gender and *APOE* status.

### **Histology and immunohistochemistry**

The Netherlands Brain Bank performed brain autopsy according to their Legal and Ethical Code of Conduct. Tissue blocks of three *EIF2AK3* carriers (two from family NLAD 1 and one from family NLAD 4) were taken from all cortical areas, hippocampus, amygdala, basal ganglia, substantia nigra, pons, medulla oblongata, cerebellum, and cervical spinal cord. They were embedded in paraffin blocks and subjected to routine staining with haematoxylin and eosin, periodic acid-Schiff reaction and silver staining. Immunohistochemistry was performed with antibodies directed against phosphorylated pancreatic endoplasmic reticulum kinase (pPERK) (sc-32577, Santa Cruz biotechnology, CA, 1:12800) and phosphorylated eukaryotic initiation factor-2 $\alpha$  (peIF2 $\alpha$ ) (SAB4504388, Sigma-Aldrich, St. Louis, MO, 1:100). We performed staining of pPERK and peIF2 $\alpha$  on the frontal, temporal and hippocampal regions of our three pathological-confirmed AD *EIF2AK3* carriers, three AD non-carriers, and three non-demented controls. Immunohistochemical staining of the neurons with pPERK and peIF2 $\alpha$  were scored with a semi-quantitative method using a modified version of the scale

developed by Stutzbach et al and Hoozemans et al: Negative (-): no cells stained, rare (+): 1–3 cells stained, ++: 4–20 cells stained or up to 10 percent of cells stained, +++: 20+ cells stained or 11 to 30 percent of cells stained, ++++: high density of stained cells (> 30 percent) in almost every field of the section.<sup>34,35</sup> In the frontal and temporal regions, the average number of positive stained cells per field were counted in nine different fields of the cortical layer at 20x magnification. In the hippocampus, we used a different scoring method as this region is often severely affected in AD with extensive neuronal loss. We counted the total number of neurons with a nucleus, as well as the number of these neurons containing pPERK or pEIF2 $\alpha$  staining to calculate the percentage of stained neurons. We focused on Cornu Ammonis 1 (CA1) and subiculum, as these contain the largest number of positive stained cells, and calculated the average percentage of stained cells per field in three different fields of CA1 and subiculum, each at 40x magnification.

We used Mann-Whitney U test to examine the difference between AD *EIF2AK3* carriers and non-carriers. All tests are two-sided significant, and a *p*-value below 0.05 was assumed as being statistically significant.

### **Immunoblot analysis**

Post-mortem fresh-frozen brain tissue of frontal cortex from three carriers of *EIF2AK3* mutations (III:15 and III:18 from family NLAD 1 and III:7 from family NLAD 4, Supplementary Figure 1) and three AD cases were extracted from the frontal cortex with buffers of increasing strength.<sup>36</sup> Briefly, grey matter was extracted at 5 ml/g (volume/weight) with low salt buffer (10mM Tris, pH 7.5, 5mM EDTA, 1mM DTT, 10% sucrose, and a cocktail of protease inhibitors), high salt-Triton buffer (low salt + 1% Triton<sup>TM</sup> X-100 + 0.5M NaCl), myelin floatation buffer (30% sucrose in low salt + 0.5M NaCl), and sarkosyl (SARK) buffer (1% N-lauroylsarcosine in low salt + 0.5M NaCl). The SARK insoluble material was extracted in 0.25 ml/g urea buffer (7M urea, 2M thiourea, 4% 3-[(3- cholamidopropyl) dimethylammonio]-1-propanesulphonate (CHAPS), 30mM Tris, pH 8.5). Proteins were resolved by 7.5% SDS-PAGE and transferred to PVDF membranes (Millipore). Following transfer, membranes were blocked with Tris buffered saline containing 3% powdered milk and probed with the antibody p-PERK (sc-32577, Santa Cruz). Primary antibodies were detected with horseradish peroxidase-conjugated anti-mouse or anti-rabbit IgG (Jackson ImmunoResearch), and signals were visualized by a chemiluminescent reaction (Millipore) and the Chemiluminescence Imager Stella 3200 (Raytest).

## Results

### **Family based exome analysis of the discovery set**

In our discovery analysis of 19 AD patients from eight families, we found an average of 91 (range 26-136) candidate variants per family after filtering (Supplementary table 1). Combining the candidate variants of the eight families, we found 101 variants in 36 candidate genes, with some genes showing many variants shared among families (Supplementary Table 2). We excluded the *MUC* genes as potential candidate as these are reported as frequent hitters in many WES datasets.<sup>37</sup> We selected the gene *EIF2AK3*, encoding for pancreatic endoplasmic reticulum kinase (PERK) as top candidate gene,<sup>7</sup> based on its involvement in memory and learning, and on its neurodegenerative role in AD and other neurodegenerative diseases,<sup>12,</sup>

<sup>38</sup>

The first *EIF2AK3* variant, p.R240H (rs147458427), was heterozygous in four affected individuals (including one with mild cognitive impairment) of family NLAD 1, and in one non-demented, 72-year old cousin of the proband (Supplementary Figure 1). This variant had a CADD score of 31 and a frequency of  $8.00 \times 10^{-04}$  in ExAC. The second *EIF2AK3* variant, p.N286S (rs150474217), had a low CADD score of 0.002 and a frequency of  $3.00 \times 10^{-05}$  in ExAC, and was confirmed in four patients with AD from family NLAD 4 and in one non-demented, 72-year old individual at last visit. One sibling with memory complaints and a normal Mini mental state examination score, did not carry the variant. Two of three patients with AD in family NLAD 4 carried homozygous *APOE*  $\epsilon 4$ ; the third patient was heterozygous for *APOE*  $\epsilon 4$ . All patients were diagnosed with early onset AD.

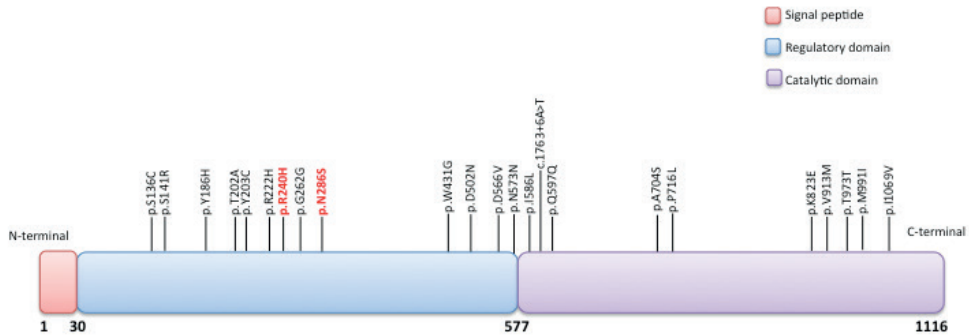
Sanger sequencing on the remaining variants in the 32 candidate genes shared among the eight families (*MUC* genes excluded) confirmed variants in 15 genes (Supplementary table 2). Segregation analysis of the variants in these 15 genes in their respective family did not show perfect segregation for most variants; the segregation in some variants could not be tested due to limited samples from related individuals.

### **Evaluation of *EIF2AK3* variants in Dutch cohorts**

To determine the genetic association of *EIF2AK3* in AD, we performed gene-based burden analysis of *EIF2AK3* variants on the Dutch AD WES dataset. We detected 23 *EIF2AK3* variants in this dataset (Figure 1 and Supplementary Table 3), of which 19 had an allele frequency <1% in ExAC; 17 of these rare variants were missense mutations. Burden test of all variants in *EIF2AK3* with MAF <1% in ExAC showed an increased risk for AD (OR=1.84; 95%CI 1.07-3.17,  $p=0.03$ ). Single variant analysis showed more carriers of variant p.R240H in cases (OR = 4.22; 95%CI 1.06 - 16.80,  $p=0.04$ ), but the nominal significant did not sustain the Bonferroni correction (Supplementary Table 3). We then performed a second analysis with *APOE* as additional covariate showing the frequency of *EIF2AK3*-carriers with at least one copy *APOE*  $\epsilon 4$  is 62% (16/26). The single variant analysis of p.R240H (OR=4.47,  $p=0.04$ ) and the burden analysis (OR=1.9,  $p=0.025$ ) were similar to the analysis without *APOE* as covariate.

As the variant p.R240H showed a suggestive signal with a high CADD score, we genotyped this variant in an independent cohort from the Rotterdam study containing 1055 cases and 6162 controls. We found an increased frequency in AD cases compared to controls

(OR=3.03; 95%CI 0.78-11.48,  $p=0.10$ ), and an association with AD after adjusting for *APOE* as additional covariate (OR=2.57; 95%CI 0.69-9.51,  $p=0.16$ ), however, in both cases the results were not statistically significant.



**Figure 1.** Schematic representation of *EIF2AK3* gene and relative position of the *EIF2AK3* variants found in the present study. The gene *EIF2AK3* contains 1116 amino acids and is composed of a signal peptide, a regulatory domain and a catalytic domain. Variants highlighted in red are found in the family based analysis.

### **Immunohistochemistry and immunoblot analysis**

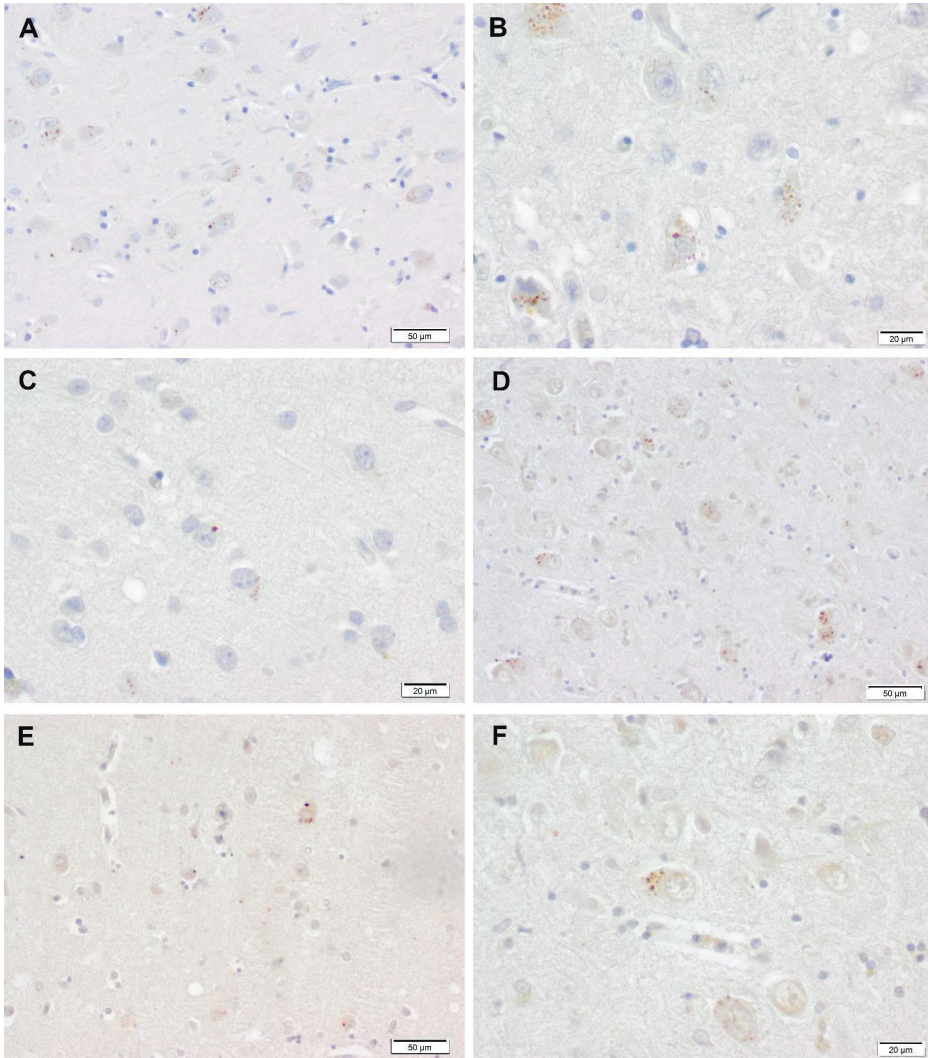
In our *EIF2AK3* carriers, many neurons with positive staining for pPERK and pEIF2 $\alpha$  were seen in the hippocampus, as well as a low to moderate number of positively stained neurons in the frontal and temporal cortex (Table 2). The activated pPERK and pEIF2 $\alpha$  staining in neurons were punctate shaped and were located in the cytoplasm, as reported in previous studies (Figure 2A-F).<sup>34, 35</sup> One carrier (III:18) from family 1 had severe neuronal loss in the CA regions and subiculum. Overall, the staining of pEIF2 $\alpha$  was more prominent than pPERK (Figure 2A,2D). All elderly non-demented controls showed a low to moderate degree of pPERK staining in the hippocampus. *EIF2AK3* carriers had significantly more positive staining than non-demented controls in the hippocampus ( $p=0.04$ ) and temporal region ( $p=0.03$ ). For pEIF2 $\alpha$ , a trend for more positive staining was only observed in the hippocampus of *EIF2AK3* carriers compared to non-demented controls ( $p=0.07$ ). We found no difference in all examined regions when comparing *EIF2AK3* carriers with AD non-*EIF2AK3* carriers; all *EIF2AK3* carriers had Braak stages 6 with extensive tau pathology in the hippocampus, frontal, temporal and parietal cortices.

We used western blot analysis with a series of buffers with increasing strength to solubilize proteins to investigate biochemical alteration of pPERK. One band of approximately 140 kDa in low salt, representing pPERK, was found in both *EIF2AK3* mutation carriers and AD cases. We found no differences in banding and solubility of pPERK between carriers of *EIF2AK3* and AD non-*EIF2AK3* carriers (Supplementary Figure 2).

**Table 2.** Scoring of inclusions for pEIF2 $\alpha$  and pPERK antibodies.

ID	Braak stage	Age at death	PMD	pEIF2 $\alpha$			pPERK		
				Frontal	Temporal	Hippocampus	Frontal	Temporal	Hippocampus
Carrier III:15 (R240H)	6	83	5:30	-	++	++++	-	+	+++
Carrier III:18 (R240H)	6	91	4:20	+	++	+++	+	+	++++
Carrier III:7 (N286S)	6	70	6:20	+	+++	++++	+	++	++++
AD Non-EIF2AK3 carrier 1	5	95	7:00	+	++	++++	-	+	+++
AD Non-EIF2AK3 carrier 2	5	62	4:40	+	+++	++++	-	+	+++
AD Non-EIF2AK3 carrier 3	5	71	5:50	+	+	++++	-	-	+++
ND control 1	4	96	4:10	-	++	+++	-	-	++
ND control 2	2	80	4:25	-	++	++	-	-	++
ND control 3	2	90	5:45	-	+	++	-	-	+

Semiquantitative scoring of inclusions for pEIF2 $\alpha$  and pPERK for carriers with *EIF2AK3* variants, Alzheimer's disease non-*EIF2AK3* controls and non-demented controls. -, negative; +, rare; ++, low density (up to 10%); +++, moderate density (11-30%); +++++, high density, >30%). AD, Alzheimer's disease; ND, Non-demented; PMD, Post mortem delay



**Figure 2.** Immunohistochemical staining of pPERK and pelf2 $\alpha$  in the AD cases with *EIF2AK3* mutations. Activated pPERK and pelf2 $\alpha$  was found in the hippocampus and temporal regions (A-F). High numbers of pPERK stained cells were observed in the cornu ammonis (A) and subiculum (B) of the hippocampus, and lesser numbers were found in the temporal cortex (C). Abundant neurons with pelf2 $\alpha$  staining were also found in the hippocampus (D), and a moderate number of stained cells were found in the frontal cortex (E). Cytoplasmic pelf2 $\alpha$  staining is punctate shaped (F), and is similar to the pPERK staining (B). Scale bars have been added to the figures.

## Discussion

This is the first study to investigate the role of rare variants in *EIF2AK3* in patients with AD. We performed whole exome sequencing in eight Dutch families with a high burden of AD, and identified *EIF2AK3* as a candidate gene in two families. Subsequently, gene based analysis in an independent Dutch WES cohort showed suggestive association of *EIF2AK3* with AD. These effects seemed to be mainly driven by variant p.R240H. Although pPERK and pEIF2 $\alpha$  staining was more prominent in *EIF2AK3* carriers than in controls, it was similar to AD non-*EIF2AK3* carriers.

We identified two distinct variants in *EIF2AK3* segregating with AD in two different families, although unaffected carriers found in each family suggested incomplete penetrance; however, they may still develop AD at an older age. The association of an *EIF2AK3* variant with AD has been reported previously, wherein one SNP (rs7571971) in *EIF2AK3* was associated with AD in *APOE*  $\epsilon$ 4 carriers, but not independent of *APOE*,<sup>39</sup> however, to date, no studies have examined the association of rare variants in *EIF2AK3* with the risk of AD. The gene burden test of *EIF2AK3* in our Dutch AD exome dataset supported this association of rare variants with AD ( $p=0.03$ ), in which it was mainly driven by the variant p.R240H with a CADD score of 31, but we were unable to confirm the association between p.R240H and AD in an additional cohort from the Rotterdam study, although there was a trend towards association with AD. A possible explanation for the lack of significance is the relatively small sample size for this rare variant. Notably, the high frequency of *APOE*  $\epsilon$ 4 carriers among the *EIF2AK3* carriers in the two families and in the Dutch AD exome dataset further support an association of *EIF2AK3* variant with AD in *APOE*  $\epsilon$ 4 carriers as indicated by Liu et al,<sup>39</sup> although similar results were found for the association tests with and without *APOE* as covariate. Studies with larger sample sizes are needed to examine the effects of rare variants in *EIF2AK3* on the risk of developing AD.

The potential significance of *EIF2AK3* variants in our families also lies in the fact that PERK is a transmembrane protein involved in learning, memory and unfolded protein response (UPR).<sup>20, 40</sup> Our hypothesis was that variants in *EIF2AK3* may enhance PERK signaling, resulting in increased phosphorylation of tau by glycogen synthase kinase 3 $\beta$  (GSK3 $\beta$ ) and amyloidogenesis (by BACE1). Previous studies have indicated that PERK-eIF2 $\alpha$  signaling is involved in the modulating of tau phosphorylation and APP processing in AD,<sup>35, 40, 41</sup> but that it is also correlated with the level of tau pathology in Progressive Supranuclear Palsy and AD.<sup>34, 35</sup> pPERK immunoreactivity also colocalized with GSK3 $\beta$  in neuronal cells, which is involved in tau phosphorylation.<sup>35, 41</sup> Treatment with a PERK-inhibitor (GSK2606414) in transgenic mice with frontotemporal lobar degeneration and overexpression of p.P301L mutation resulted in reduced GSK3 $\beta$ -levels and tau phosphorylation compared to transgenic mice without PERK inhibitor treatment.<sup>42</sup> Moreover, *PSEN1* (5XFAD) mutated mice with PERK haploinsufficiency had lower levels of Beta-secretase 1 (BACE1) than those with normal PERK levels, resulting in lower amyloid-beta peptides levels and plaque burden, as well as fewer memory deficits and cholinergic neurodegeneration.<sup>40</sup> Reduced synaptic plasticity and spatial memory deficits were found in APP/PS1 AD model mice with PERK haploinsufficiency.<sup>43</sup> Although these studies supported a role of PERK signaling in the pathogenesis of AD, functional experiments are needed to confirm the effect of *EIF2AK3* variants.



The increase of PERK-eIF2 $\alpha$  signaling in the *EIF2AK3* carriers is supported by the more positive staining of pPERK and pEIF2 $\alpha$  compared to non-demented controls, indicating an increased activation of UPR. This increased UPR has also been observed in AD and PSP patients in previous studies.<sup>34, 44</sup> However, we did not find any differences in pPERK and pEIF2 $\alpha$  staining between *EIF2AK3* carriers and AD non-*EIF2AK3* carriers, suggesting *EIF2AK3* mutation carriers might not induce more UPR activation than other AD patients. A possible explanation is that *EIF2AK3* mutation carriers may trigger UPR activation early in the disease process, without the ability to observe this at the end stage AD.

The main limitation of our study is the family-based analysis used to identify the candidate genes; we only selected genes containing rare variants in at least two families for follow-up. We cannot rule out the possibility that other possible candidates in the families were missed. However, this method has previously been successfully used by Cruchaga et al, resulting in the identification of the genetic association of *PLD3* with AD.<sup>14</sup> Furthermore, *EIF2AK3* was the only gene in our candidate list involved in the pathogenesis of AD. Another limitation is the limited available samples of related cases and (old) non-demented controls in some families to analyze segregation; some non-demented controls may still develop dementia at older age. Finally, the frequency of *APOE*  $\epsilon$ 4 is high in some families, and *APOE*  $\epsilon$ 4 segregates with the disease in some of them. This is also true for family 4, in which variant p.N285S was found; four patients and one individual with memory complaints carried at least one copy of *APOE*  $\epsilon$ 4. However, all four patients carrying p.N285S and *APOE*  $\epsilon$ 4 had early onset AD, indicating a possible additional effect of genetic variation in *EIF2AK3* on the risk of AD among *APOE*  $\epsilon$ 4 carriers, as indicated in a previous study.<sup>39</sup> Future analyses in larger case-control studies are necessary to confirm this association.

In conclusion, our study showed that rare variants in *EIF2AK3* may be associated with an increased risk of AD based on segregation among the patients with AD in two families and a gene-based analysis in the Dutch WES cohort. Immunohistochemistry confirmed more activation of UPR, characterized by increased pPERK and pEIF2 $\alpha$  in AD patients compared to non-demented controls, but not between *EIF2AK3* carriers and AD non-carriers. Further studies are needed to investigate the full contribution of rare variants in *EIF2AK3* in the development of AD.

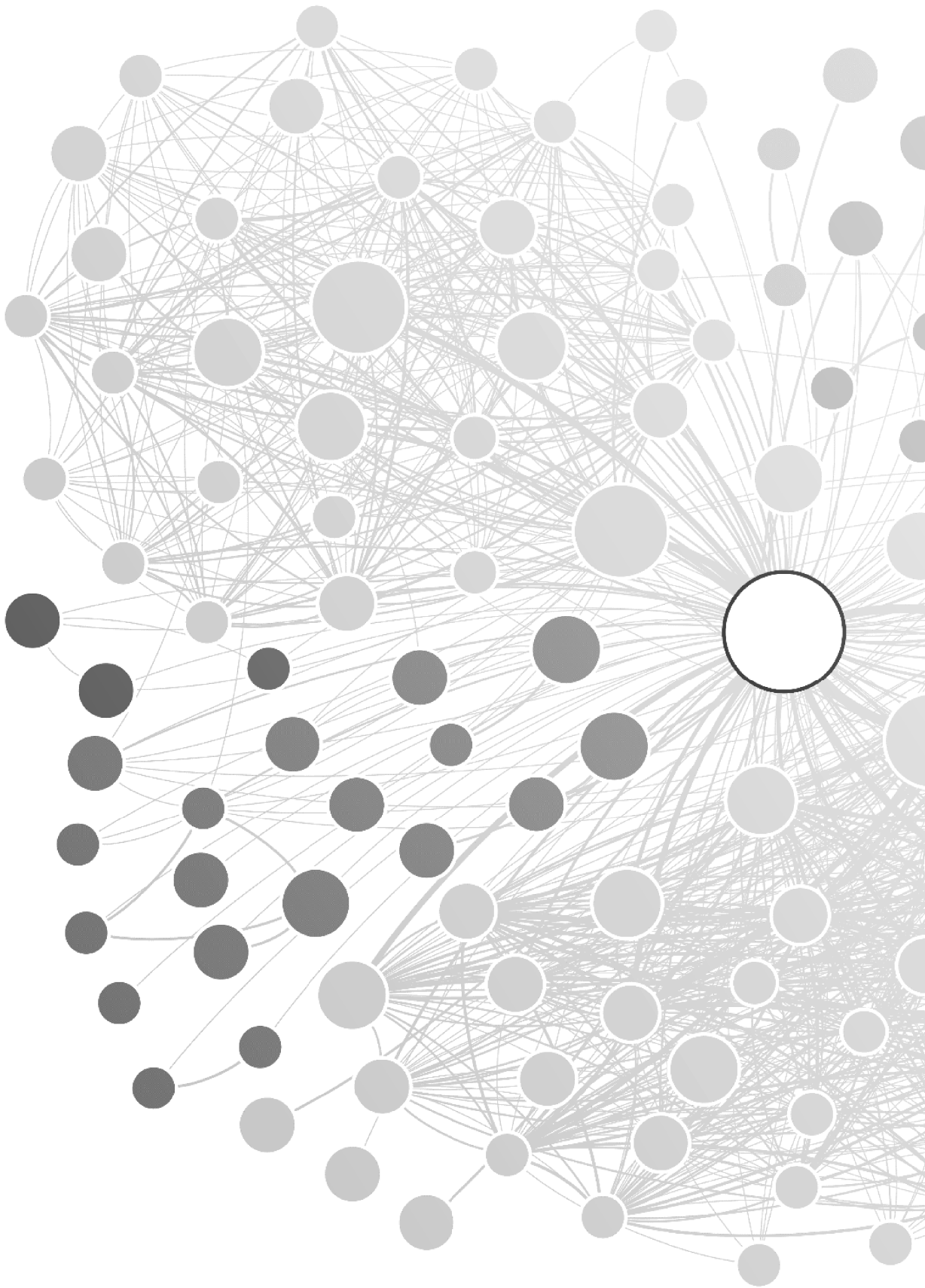
## References

1. Ballard C, Gauthier S, Corbett A, Brayne C, Aarsland D, Jones E. Alzheimer's disease. *Lancet* 2011;377:1019-1031.
2. St George-Hyslop PH. Molecular genetics of Alzheimer disease. *Semin Neurol* 1999;19:371-383.
3. Goate A, Chartier-Harlin MC, Mullan M, et al. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature* 1991;349:704-706.
4. Levy E, Carman MD, Fernandez-Madrid IJ, et al. Mutation of the Alzheimer's disease amyloid gene in hereditary cerebral hemorrhage, Dutch type. *Science* 1990;248:1124-1126.
5. Levy-Lahad E, Wasco W, Poorkaj P, et al. Candidate gene for the chromosome 1 familial Alzheimer's disease locus. *Science* 1995;269:973-977.
6. Rogaeve EI, Sherrington R, Rogaeve EA, et al. Familial Alzheimer's disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer's disease type 3 gene. *Nature* 1995;376:775-778.
7. Sherrington R, Rogaeve EI, Liang Y, et al. Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature* 1995;375:754-760.
8. Campion D, Dumanchin C, Hannequin D, et al. Early-onset autosomal dominant Alzheimer disease: prevalence, genetic heterogeneity, and mutation spectrum. *Am J Hum Genet* 1999;65:664-670.
9. Farrer LA, Cupples LA, Haines JL, et al. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *JAMA* 1997;278:1349-1356.
10. Hetz C, Mollereau B. Disturbance of endoplasmic reticulum proteostasis in neurodegenerative diseases. *Nat Rev Neurosci* 2014;15:233-249.
11. Braak H, Braak E. Neuropathological staging of Alzheimer-related changes. *Acta Neuropathol* 1991;82:239-259.
12. Scheper W, Hoozemans JJ. The unfolded protein response in neurodegenerative diseases: a neuropathological perspective. *Acta Neuropathol* 2015;130:315-331.
13. Smith HL, Mallucci GR. The unfolded protein response: mechanisms and therapy of neurodegeneration. *Brain* 2016;139:2113-2121.
14. Cruchaga C, Karch CM, Jin SC, et al. Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. *Nature* 2014;505:550-554.
15. Cuyvers E, De Roeck A, Van den Bossche T, et al. Mutations in ABCA7 in a Belgian cohort of Alzheimer's disease patients: a targeted resequencing study. *Lancet Neurol* 2015;14:814-822.
16. Guerreiro R, Wojtas A, Bras J, et al. TREM2 variants in Alzheimer's disease. *N Engl J Med* 2013;368:117-127.
17. Pottier C, Hannequin D, Coutant S, et al. High frequency of potentially pathogenic SORL1 mutations in autosomal dominant early-onset Alzheimer disease. *Mol Psychiatry* 2012;17:875-879.
18. Holstege H, van der Lee SJ, Hulsman M, et al. Characterization of pathogenic SORL1 genetic variants for association with Alzheimer's disease: a clinical interpretation strategy. *Eur J Hum Genet* 2017;25:973-981.
19. Sims R, van der Lee SJ, Naj AC, et al. Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. *Nat Genet* 2017;49:1373-1384.
20. Rozpedek W, Markiewicz L, Diehl JA, Pytel D, Majsterek I. Unfolded Protein Response and PERK Kinase as a New Therapeutic Target in the Pathogenesis of Alzheimer's Disease. *Curr Med Chem* 2015;22:3169-3184.

21. McKhann GM, Knopman DS, Chertkow H, et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* 2011;7:263-269.
22. Ikram MA, Brusselle GGO, Murad SD, et al. The Rotterdam Study: 2018 update on objectives, design and main results. *Eur J Epidemiol* 2017;32:807-850.
23. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754-1760.
24. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078-2079.
25. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297-1303.
26. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;81:559-575.
27. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;38:e164.
28. Genome of the Netherlands C. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* 2014;46:818-825.
29. Genomes Project C, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature* 2015;526:68-74.
30. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;536:285-291.
31. Tennessen JA, Bigham AW, O'Connor TD, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 2012;337:64-69.
32. van Rooij JGJ, Jhamai M, Arp PP, et al. Population-specific genetic variation in large sequencing data sets: why more data is still better. *Eur J Hum Genet* 2017;25:1173-1175.
33. Untergasser A, Cutcutache I, Koressaar T, et al. Primer3--new capabilities and interfaces. *Nucleic Acids Res* 2012;40:e115.
34. Stutzbach LD, Xie SX, Naj AC, et al. The unfolded protein response is activated in disease-affected brain regions in progressive supranuclear palsy and Alzheimer's disease. *Acta Neuropathol Commun* 2013;1:31.
35. Hoozemans JJ, van Haastert ES, Nijholt DA, Rozemuller AJ, Eikelenboom P, Scheper W. The unfolded protein response is activated in pretangle neurons in Alzheimer's disease hippocampus. *Am J Pathol* 2009;174:1241-1251.
36. Neumann M, Sampathu DM, Kwong LK, et al. Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science* 2006;314:130-133.
37. Fuentes Fajardo KV, Adams D, Program NCS, et al. Detecting false-positive signals in exome sequencing. *Hum Mutat* 2012;33:609-613.
38. Ohno M. PERK as a hub of multiple pathogenic pathways leading to memory deficits and neurodegeneration in Alzheimer's disease. *Brain Res Bull* 2017.
39. Liu QY, Yu JT, Miao D, et al. An exploratory study on STX6, MOBP, MAPT, and EIF2AK3 and late-onset Alzheimer's disease. *Neurobiol Aging* 2013;34:1519 e1513-1517.
40. Devi L, Ohno M. PERK mediates eIF2alpha phosphorylation responsible for BACE1 elevation, CREB dysfunction and neurodegeneration in a mouse model of Alzheimer's disease. *Neurobiol Aging* 2014;35:2272-2281.

41. Nijholt DA, Nolle A, van Haastert ES, et al. Unfolded protein response activates glycogen synthase kinase-3 via selective lysosomal degradation. *Neurobiol Aging* 2013;34:1759-1771.
42. Radford H, Moreno JA, Verity N, Halliday M, Mallucci GR. PERK inhibition prevents tau-mediated neurodegeneration in a mouse model of frontotemporal dementia. *Acta Neuropathol* 2015;130:633-642.
43. Ma T, Trinh MA, Wexler AJ, et al. Suppression of eIF2alpha kinases alleviates Alzheimer's disease-related plasticity and memory deficits. *Nat Neurosci* 2013;16:1299-1305.
44. Hoozemans JJ, Veerhuis R, Van Haastert ES, et al. The unfolded protein response is activated in Alzheimer's disease. *Acta Neuropathol* 2005;110:165-172.







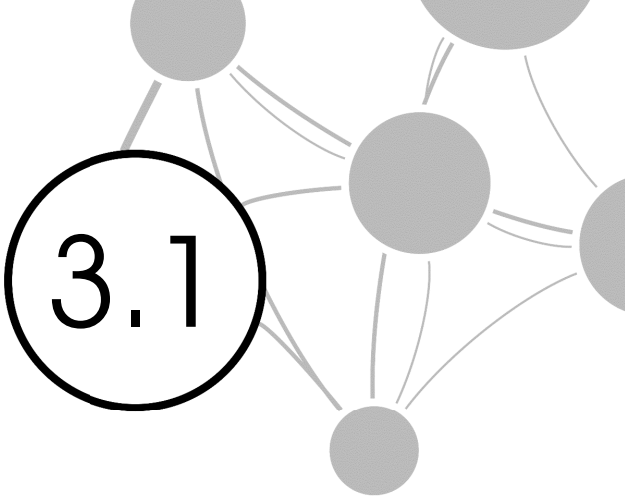
# Chapter 3

## Sequencing RNA blood & brain





# Chapter 3.1



## Evaluation of commonly used analysis strategies for epigenome and transcriptome-wide association studies through replication of large-scale population studies

Jeroen van Rooij<sup>\*1</sup>, Pooja R. Mandaviya<sup>\*1,2</sup>, Annique Claringbould<sup>3</sup>, Janine F. Felix<sup>4</sup>, Jenny van Dongen<sup>5</sup>, Rick Jansen<sup>6</sup>, Lude Franke<sup>7</sup>, BIOS consortium, Peter A.C. 't Hoen<sup>#8,9</sup>, Bas Heijmans<sup>#10</sup>, Joyce B.J. van Meurs<sup>#1</sup>

<sup>1</sup> Department of Internal Medicine, Erasmus MC, University Medical Center Rotterdam, the Netherlands

<sup>2</sup> Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, Maastricht, the Netherlands

<sup>3</sup> Faculty of Medical Sciences, University of Groningen, Groningen, the Netherlands

<sup>4</sup> The Generation R Study Group, Department of Epidemiology, Department of Pediatrics, Erasmus MC, University Medical Center Rotterdam, the Netherlands

<sup>5</sup> Department of Biological Psychology, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

<sup>6</sup> Department of Psychiatry, VU University Medical Center, Amsterdam, the Netherlands

<sup>7</sup> Department of Genetics, University of Groningen, Groningen, the Netherlands

<sup>8</sup> Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands

<sup>9</sup> Centre for Molecular and Biomolecular Informatics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center Nijmegen, Nijmegen, the Netherlands

<sup>10</sup> Molecular Epidemiology, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, the Netherlands

\*These authors contributed equally as first authors

#These authors contributed equally as last authors

Published on 14<sup>th</sup> of November 2019 in *Genome Biology* (IF = 14). PMID: 31727104, doi: 10.1186/s13059-019-1878-x

## Abstract

**Background:** A large number of analysis strategies are available for DNA methylation (DNAm) array and RNA-seq datasets, but it is unclear which strategies are best to use. We compare commonly used strategies and report how they influence results in large cohort studies.

**Results:** We tested the associations of DNAm and RNA expression with age, BMI and smoking in four different cohorts ( $n \sim 2,900$ ). By comparing strategies against the base model on the number and percentage of replicated CpGs for DNAm analyses or genes for RNA-seq analyses in a leave-one cohort out replication approach, we find the choice of normalization method and statistical test does not strongly influence the results for DNAm array data. However, adjusting for cell counts or hidden confounders substantially decreases the number of replicated CpGs for age and increases the number of replicated CpGs for BMI and smoking. For RNA-seq data, the choice of normalization method, gene expression inclusion threshold and statistical test does not strongly influence the results. Including five principal components or excluding correction of technical covariates or cell counts decreases the number of replicated genes.

**Conclusions:** Results were not influenced by normalization method or statistical test. However, the correction method for cell counts, technical covariates, principal components and/or hidden confounders does influence the results.

## Background

Epigenomics and transcriptomics are important tools to investigate molecular mechanisms of disease etiology. Unlike the genome, the epigenome and transcriptome are dynamic and differ across tissues and over time [1-4].

Consequently, an epigenome-wide or transcriptome-wide association study (EWAS or TWAS, respectively) is influenced by more biological and technical factors than a genome-wide association study (GWAS). As a result, EWAS and TWAS methods are less standardized and do not always present the same results. For example, EWASs comparing current smokers with never smokers resulted in different significant CpGs and different numbers of significant CpGs per study, independent of sample size [5-15]. Similarly, TWASs comparing current smokers with never smokers found different numbers of associated genes [16-19]. Although these studies took place in different populations, they also used different analytical strategies, which could explain part of the variation in results.

For DNA methylation (DNAm) array data, previous studies compared different normalization methods [20-24]. Wu *et al.* concluded that most normalization methods performed similarly in association analyses when there was a strong association between CpGs and the exposure of interest [20]. To investigate the performance of DNAm values, Du *et al.* compared the use of beta-values with M-values in two samples and concluded that M-values had better statistical properties, whereas beta-values were more biologically interpretable [25]. Furthermore, white blood cell (WBC) counts are often used as important confounder adjustments for EWASs in whole blood. Cell counts estimated using the Houseman method [26] are commonly used when measured cell counts are not available. However, since the Houseman method is based on only six reference individuals [27], thorough investigation of this method based on large-scale DNAm data is needed. Lastly, principal components (PCs), surrogate variables (SVs) or unobserved covariates (also known as hidden confounders (HCs)) are commonly used methods to adjust for unmeasured hidden (technical or biological) confounders. Estimation of HCs using CATE has been suggested to outperform covariate adjustment using PCs or SVs [28, 29].

For RNA sequencing (RNA-seq) data, Li *et al.* compared a range of normalization methods, and concluded that the commonly used options (e.g. DESeq/edgeR) provided the highest accuracy at the cost of decreased sensitivity compared to options with more specific applications [30]. When sufficient replicates ( $n > 4$ ) per group were used, all methods performed similarly. Li *et al.* also compared normalization methods and concluded that commonly used options performed similarly, although some specific methods performed better for short (35bp) read lengths and/or when alignment quality was low [30]. Several studies focused on other aspects of the analysis procedure such as the gene database used for quantifications (i.e. RefSeq, UCSC and Ensembl) or sequencing platform and flowcell effect on results [31-33]. However, a comprehensive examination of multiple steps and combinations of analysis options is still lacking.

Most of these previous studies focused on a specific aspect of the procedure using simulated data or small datasets. To provide a complete evaluation of analysis strategies, we analyzed,

replicated and compared analysis strategies composed of commonly used normalization, correction and association options in four large population-based datasets of the BIOS project, which have both DNAm array and RNA-seq data available [34, 35]. Because of this design, we can replicate results across cohorts and evaluate analysis strategies based on their replication performance. Our evaluation will help researchers select the optimal strategy and reduce unnecessary variation across studies. In addition, information about strategy differences will be helpful when comparing studies where different analysis strategies are used.

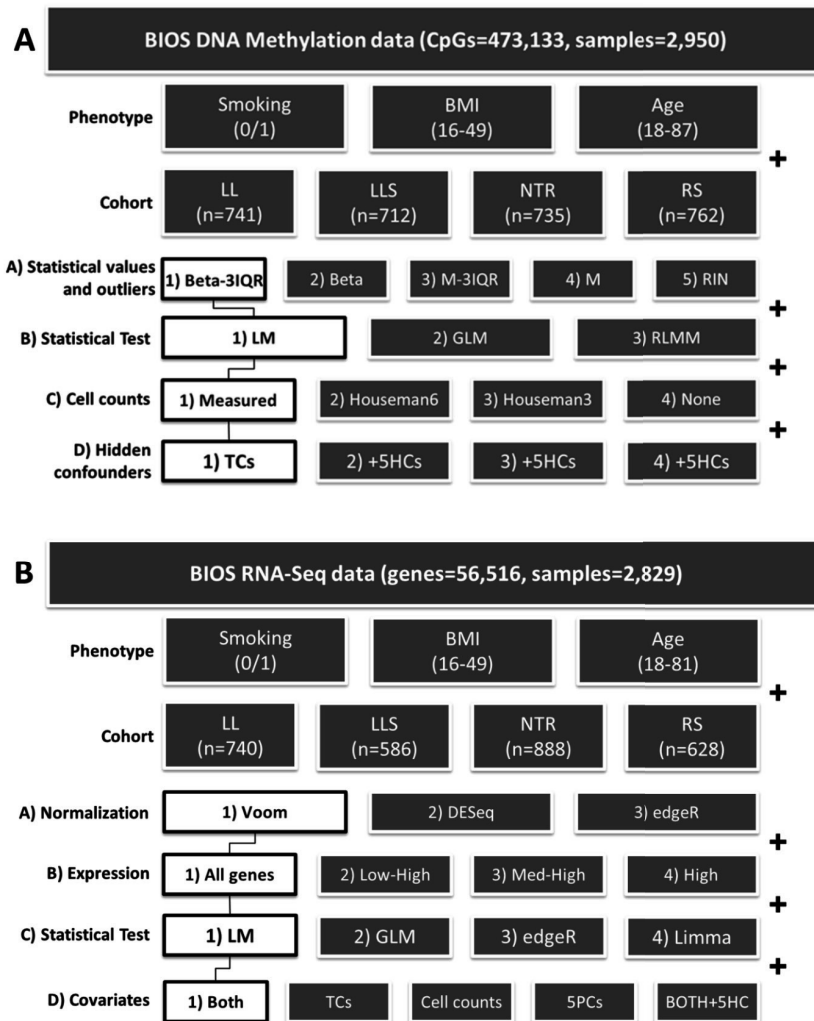
## Results

Table 1 shows phenotypic characteristics for the four cohorts analyzed. To accommodate the differences in characteristics of the cohorts, cohorts were meta-analyzed. Figure 1 shows the various analysis strategies under evaluation. We selected a base model for DNAm and RNA-seq analysis comprised of one option in each category. Then, per category we swapped the option in the base model with the alternatives and evaluate the replication performance against the base model. The categories for DNAm were; A) DNAm value preprocessing, B) Statistical test, C) Cell counts, D) Hidden confounders. The categories for RNA-seq were; A) Normalization method, B) Expression inclusion threshold, C) Statistical test, D) Technical Covariates.

**Table 1.** characteristics of the four main cohorts at time of blood draw. All entries represent averages with standard deviations unless otherwise indicated.

	Phenotypes				DNA methylation				RNA-seq			
	LL	LLS	NTR	RS	LL	LLS	NTR	RS	LL	LLS	NTR	RS
	n=761	n=790	n=1,866	n=768	n=741	n=712	n=735	n=762	n=740	n=579	n=882	n=628
Age	45 ± 13	58 ± 8	37 ± 14	68 ± 6	46 ± 13	59 ± 7	40 ± 15	68 ± 7	45 ± 13	59 ± 7	38 ± 15	69 ± 6
Sex (%male)	0.42	0.48	0.33	0.43	0.42	0.48	0.35	0.43	0.42	0.47	0.35	0.43
Smoking (%current)	0.15	0.11	0.19	0.10	0.15	0.12	0.19	0.10	0.15	0.13	0.18	0.09
BMI	25 ± 4	25 ± 4	24 ± 4	28 ± 4	25 ± 4	25 ± 3	25 ± 4	28 ± 4	25 ± 4	25 ± 3	25 ± 4	28 ± 4
Lymp (%of cells)	34 ± 8	29 ± 7	35 ± 9	36 ± 8	35 ± 7	29 ± 7	35 ± 9	36 ± 8	34 ± 8	29 ± 7	35 ± 9	36 ± 8
Mono (%of cells)	9 ± 2	5 ± 2	8 ± 3	7 ± 2	9 ± 2	6 ± 2	8 ± 3	7 ± 2	9 ± 2	6 ± 2	9 ± 3	7 ± 2
Gran (%of cells)	57 ± 8	63 ± 7	56 ± 9	57 ± 8	57 ± 8	63 ± 7	57 ± 9	57 ± 9	57 ± 8	63 ± 7	56 ± 9	57 ± 8

Each analysis strategy was meta-analyzed across three cohorts and replicated in the fourth, in all four combinations (the so-called “leave-one-out method”). Both meta-analysis and replication were defined by Bonferroni correction ( $p < 0.05$ ) for the number of CpGs/genes tested. Below, we first describe the performance of the base model for methylation and expression data. Then we describe, per category, how the various options affected the number of replicated signals (as a measure of sensitivity) and percentage of replicated signals (as a measure of true-positive rate in the discovery) and the overlap of significant CpGs/genes between analysis strategies. All results are Bonferroni corrected.



**Figure 1a.** Overview of DNA methylation analysis steps and commonly used options. We identified four steps in the procedure which often vary in literature; A) DNAm value preprocessing, B) Statistical test, C) Cell count correction, D) Hidden confounder correction. We selected one combination of options and then varied these a single step at the time. These models were applied to age, BMI and smoking. Each model was meta-analyzed in each combination of three discovery and one replication cohorts. Average replication rate and number of replicated genes of these four analyses were used to evaluate strategies. The base model is connected by the black line and includes Beta-3IQR dataset, an LM model, measured cell counts correction, known technical confounder correction (TCs) (plate and row) and applying Bonferroni correction. HCs: hidden confounders, calculated after regressing out technical covariates (2), cell-counts (3) or both (4).

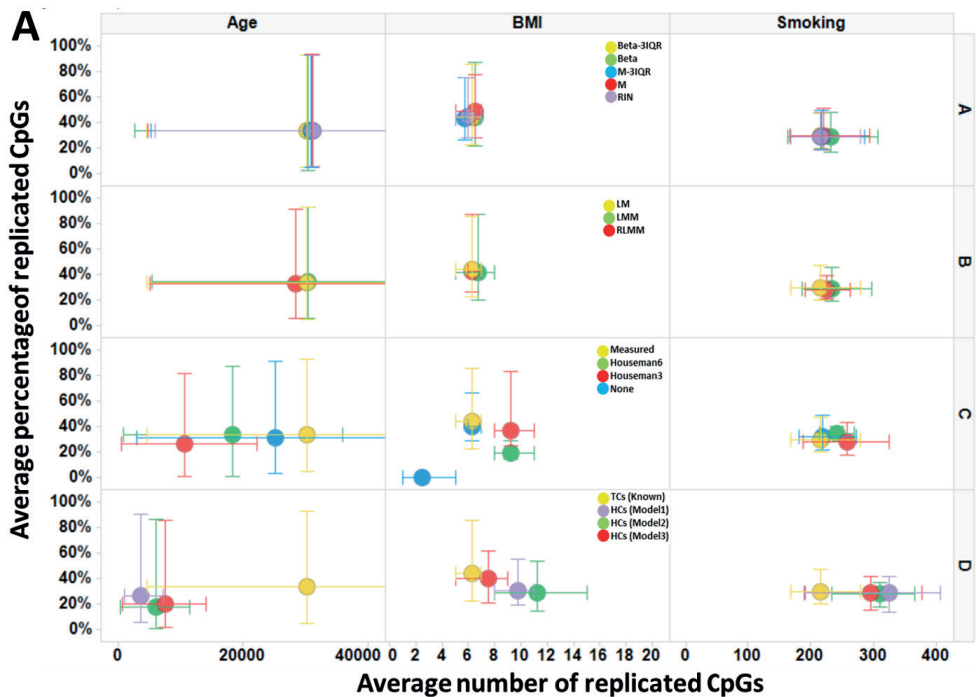
**Figure 1b.** Overview of gene expression analysis steps and commonly used options. We identified four steps in the procedure which often vary in literature; A) Normalization, B) Expression, C) Tests and D) technical covariates. We selected one combination of options and then varied these a single step at the time. These models were applied to age, BMI and smoking. Each model was meta-analyzed in each combination of three discovery and one replication cohorts. Average replication rate and number of replicated genes of these four analyses were used to evaluate strategies. The base model is connected by the black line; Voom normalization, including all genes, a LM for statistical analysis, including technical covariates and cell-counts and applying Bonferroni correction.

### **DNA methylation strategy performance**

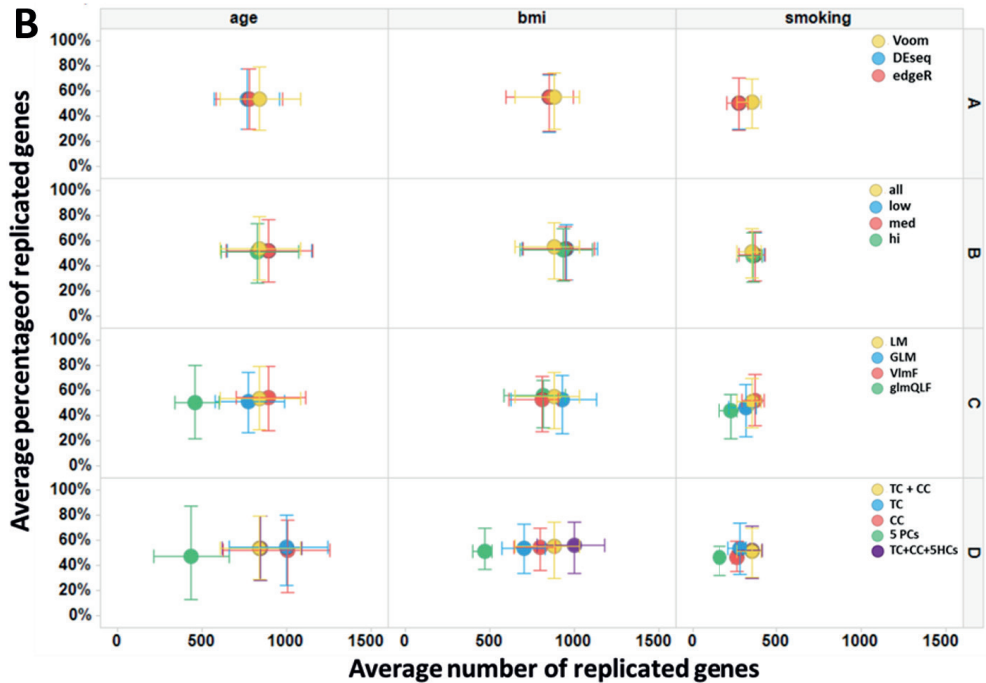
The base model included using normalized beta values and removing outliers based on the three interquartile range strategy (beta-3IQR), a linear model (LM), measured cell counts and technical covariates, as described in more detail in the methods. This resulted in an average of 30,275 significantly replicated CpGs for age (range 4,621 - 59,087), 6 replicated CpGs for BMI (range 5 - 7) and 217 replicated CpGs for smoking (range 168 - 279). The corresponding replication rates were on average 40% for age (range 5% - 93%), 52% for BMI (range 23% - 86%) and 31% for smoking (range 20% - 47%). All summary results are shown in figure 2a, figure 3a and Additional file 3. Below we describe per category how different options influenced these results.

- A. *DNAm value preprocessing*: For age, all normalization methods showed similar replication rates and slightly higher replication number compared to the base model. The same was observed for smoking, except that the RIN method performed more similar to the base model than the beta, M or M-3IQR methods. The replicated number and rate of CpGs were largely the same across methods. For BMI, given the small numbers of CpGs (e.g. 6 for the base model), it was difficult to robustly compare results.
- B. *Statistical tests*: Compared to the base model, a linear mixed model (LMM) reported a slightly higher number of replicated hits for age and smoking. The robust linear mixed model (RLMM) reported lower numbers of replicated CpGs for age and similar number of replicated CpGs for smoking. Replication rates were nearly identical to the LM base model for all exposures. The replicated CpGs were shared across methods.
- C. *Cell count adjustment*: Without correction for cell counts, fewer replicated CpGs were found for age (83% compared to the number of replicated CpGs in the base model), but no differences were seen for BMI and smoking (Figure 2a). For age, adjusting for Houseman imputed cell counts substantially decreased the number of significantly replicated CpGs; Houseman6 resulted in 18,368 CpGs for age (61% of the base model) and Houseman3 resulted in 10,678 CpGs for age (35% of the CPGs compared to the base model). The replication rate with Houseman6 was similar as compared to the base model, but Houseman3 resulted in a slightly lower replication rate as compared to the base model. For smoking, using Houseman imputed cell counts resulted in a slightly higher number of replicated CpGs; Houseman6 resulted in 243 CpGs (112% compared to the base model), while Houseman3 resulted in 259 CpGs (119% compared to the base model). When examining the overlap between the CpGs in the different cell count adjustment strategies across all 4 cohorts (figure 3a) for smoking, we observed that a total of 652 CpGs were common for all cell count adjustment methods. In addition, a relatively large number of CpGs were only observed by Houseman6 and 3, respectively (312 and 220 CpGs).

*Correction for Hidden Confounders (HCs):* HCs were calculated in three additional models (model 1 being the base model); model 2) HCs independent of the described covariates, but not measured differential cell counts; model 3) HCs independent of the described covariates, but not known technical covariates; model 4) using HCs independent of the exposure of interest, age, sex, known technical covariates and measured differential cell counts. For age, adjusting for 5 HCs resulted in a decreased number of significantly replicated CpGs: 7,509 in model 4 (25% compared to the base model), 6,054 in model 3 (20% compared to the base model) and 3,621 in model 2 (12% compared to the base model). In contrast, for BMI and smoking, these three HCs models showed an increase in number of significantly replicated CpGs: 8, 9 and 10 for BMI and 297 (137% of the base model), 311 (143% of the base model) and 325 (150% of the base model) for smoking in models 4, 3 and 2, respectively. Thus, for age, a large number of CpGs were not detected when correcting for HCs, while for smoking and BMI, a number of CpGs were found only when using HCs correction. The replication rates were very similar across all models.







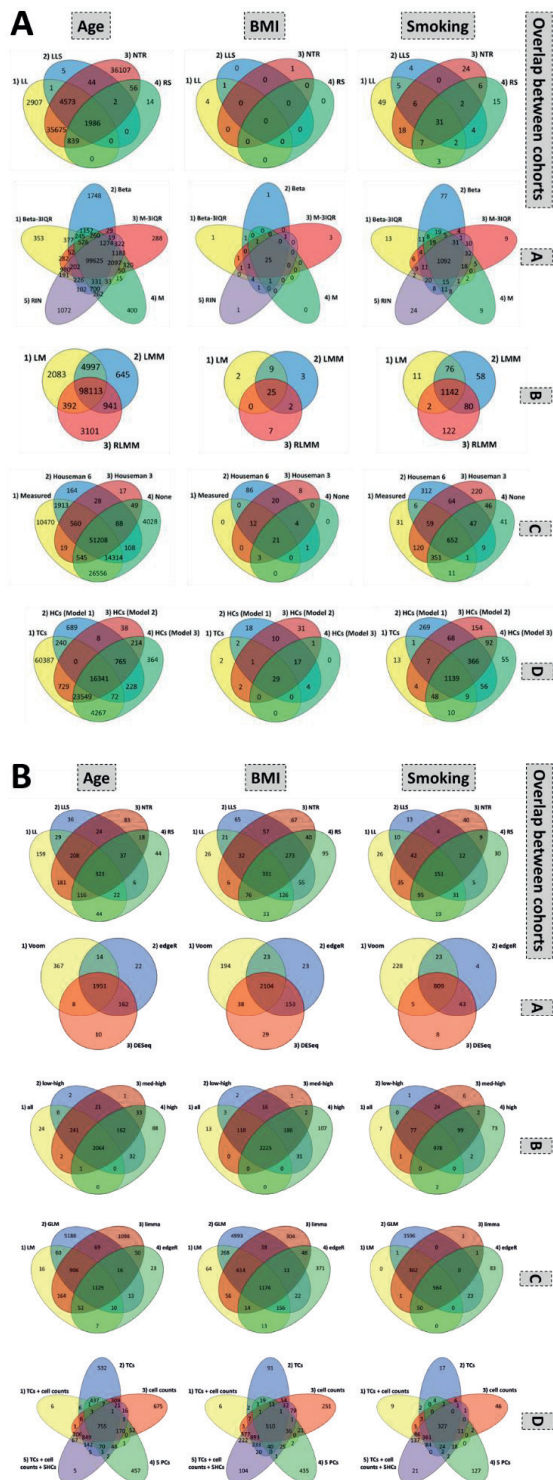
**Figure 2a.** The number (x-axis) and percentage (y-axis) of replicated CpGs for age, BMI and smoking (shown in columns). Per row, each step of the analysis strategy is displayed. The yellow model is the reference model and remains the same in each column and row: Beta-3IQR dataset, standard linear model (LM), measured cell counts correction and known technical confounders (bisulfite conversion plate and array row) correction (TCs). The circles are average Bonferroni-corrected replication results. The bars indicate the range of the four leave-one-out analyses. In each row, the other (non-yellow) colors represent alternative options: A) Datatypes: Beta without exclusion of outliers in green, M-values in red, M-values with outlier exclusion using the 3IQR method in blue and RIN in purple. B) Statistical models: linear mixed models (LMM) in green and robust linear mixed models (RLMM) in red. C) Cell count adjustment: Houseman6 in green, Houseman3 in red and none in blue (see methods for details) D) Hidden confounders (HCs) correction; Model 1 in purple, Model 2 in green and Model 3 in red (see methods for details).

**Figure 2b.** The number (x-axis) and percentage (y-axis) of replicated genes for age, BMI and smoking (shown in columns). Per row, each step of the analysis strategy is displayed. The yellow model is the reference model and remains the same in each column and row: Voom normalization, including all genes, standard linear model (LM), correcting for technical covariates (TC) and cell counts (CC). The circles are average Bonferroni-corrected replication results. The bars indicate the range of the four leave-one-out analyses. In each row, the other (non-yellow) colors represent alternative options: A) Normalization methods: DESeq normalization in blue and edgeR in red. B) gene inclusion: removing very low-expressed genes (blue), low-expressed genes (red) or medium-expressed genes (green). C) Statistical models: A limma linear model Fit in red (limma), a standard GLM in blue and the edgeR GLM adaptation in green. D) Covariates: correcting solely for technical covariates (TC; blue) or cell-counts (CC; red) or replacing both for the first 5 principal components (5PCs; green), the last option is by adding 5 hidden confounders (HCs) to the technical covariates and cell counts (5HCs; purple).

### **RNA sequencing strategy performance**

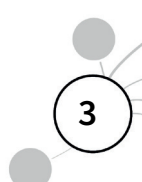
The base model (Voom normalization, no expression inclusion threshold, LM, technical covariates and measured cell counts) resulted on average in 842 significantly replicated genes for age (range 610 - 1082), 881 replicated genes for BMI (range 651 - 1029) and 354 replicated genes for smoking (range 268 - 409). The corresponding mean replication rates were 54% for age (range 28% - 80%), 55% for BMI (range 30% - 74%) and 51% for smoking (range 30% - 69%). Below we describe per category how different options influenced these results, as available in Additional file 3 and shown in figures 2b and 3b.

- A. *Normalization method*: The DESeq and edgeR normalization methods reported slightly lower number of replicated genes with the same replication rate compared to the base model (93% and 91% of the base model, respectively). The normalization method did not influence which genes were replicated. This pattern was observed for all three exposures.
- B. *Gene expression inclusion criteria*: Including low (average CPM > 1 in 20% of samples) and higher expressed genes (1.low) or medium (average CPM > 1) and higher expressed genes (2.med) provided slightly more replicated genes for age (both 107% compared to the base model) at a similar replication rate. The most stringent threshold (3. hi) also resulted in a similar replication number (98% compared to the base model) and percentage (98% compared to the base model). Mostly the same genes were replicated regardless of the inclusion threshold.
- C. *Statistical tests*: limma's linear model fit (limma) test resulted in slightly more replicated genes, at the cost of a lower replication rate (lower specificity). The glmQLF test from edgeR showed a lower number of replicated genes. GLM showed nearly the same results as the base model. These findings were consistent across the exposures, with smaller differences for BMI.
- D. *Covariates*: For age, correcting solely for technical covariates or cell-counts resulted in a large increase (119% compared to the base model) in replicated genes. For BMI and smoking, the number of replicated genes, as well as the replication rate decreased when removing these covariates. Correcting for five principal components instead of technical covariates or cell-counts decreased the number of replicated signals to 51%, 53% and 46% of the base model for age, BMI and smoking, respectively. Similarly the replication rate decreased to 87%, 96% and 96% for age, BMI and smoking compared to the base model, respectively. Conversely, five hidden confounders added to the technical covariates and cell-counts in the base model increased the replication number to 100.4%, 114% and 101.4% compared to the base model for age, BMI and smoking, and increased the replication rate to 107%, 103% and 103% of the base model for age, BMI and smoking, respectively. In addition to finding fewer replicated genes after PC correction, the identified genes were not the same as the base model, and other methods did not observe these genes. Similarly when adding five HCs, many genes identified in the model with HCs were not observed in the other models, but the difference was smaller than for the model including PCs.



**Figure 3a.** CpG Overlaps. The three 4-way Venn diagrams on top indicate the overlap in CpGs for each of the individual cohorts. These are based on the base model, using Bonferroni correction. The four diagrams below indicate the overlap between the strategies for each step, shown here for age, BMI and smoking. These are the same strategies as shown in figure 2a. Yellow always represents the base model, the green, red, blue and purple colors belong to alternative strategies. A) Beta values dataset in green, M-3IQR in blue, M in red and RIN in purple. B) Houseman6 imputed cell counts in green, Houseman3 imputed cell counts in red and no cell count correction in blue. D) Hidden confounders (HCs) correction: Model 1 (HCs independent of the exposure of interest, age, sex, known technical covariates, but not measured differential cell counts) in purple, Model 2 (HCs independent of the exposure of interest, age, sex, measured differential cell counts, but not known technical covariates) in green and Model 3 (independent of the exposure of interest, age, sex, known technical covariates and measured differential cell counts) in red.

**Figure 3b.** Gene Overlaps; The three 4-way Venn diagrams on top indicate the overlap in genes for each of the individual cohorts. These are based on the base model, using Bonferroni correction. The four diagrams below indicate the overlap between the strategies for each step, shown here for age, BMI and smoking. These are the same strategies as shown in figure 2b. Yellow always represents the base model, the blue, green and red colors belong to alternative strategies. A) DESeq normalization in blue, edgeR in red. B) Removing very low-expressed genes (blue), low-expressed genes (red) or medium-expressed genes (green). C) A limma linear model Fit in red, a standard GLM in blue and the edgeR GLM adaptation in green. D) Correcting for only technical covariates (blue), only cell-counts (red), only cell-counts and measured differential cell counts) in green and principal components (green).



### ***FDR instead of Bonferroni correction***

In addition to the comparisons described above, all analyses were also repeated using FDR correction in the discovery analysis instead of Bonferroni correction. All analyses using FDR showed a higher number of replicated CpGs and genes, at the cost of a much smaller replication rate. For example, for the base model for age, 30,275 CpGs and 842 genes were replicated at replication rates of 40% and 47%, respectively, when using Bonferroni correction. When using FDR correction, the number of CpGs increased by 18% and the replication rate decreased by 18%. Similarly, the number of genes increased by 98% and the replication rate decreased by 20%.

### ***METAL or GWAMA for meta-analysis***

As the GWAMA tool requires input that is not provided by some RNA expression statistical methods, we opted to use only METAL for the RNAseq analysis. For those RNAseq models where both could be run; the results were identical.

### ***Evaluation using different p-value cut-offs***

The results for additional p-value cutoffs (FDR, uncorrected  $<1 \times 10^{-8}$  and uncorrected  $<0.05$ ) are available in Additional files 2 and 3. Less stringent cutoffs led to an increase in absolute numbers of replicated signals but at a decreased relative replication rate for both DNAm and RNAseq. Most models responded similarly to this change and the respective performance between methods did not change.

For BMI and smoking in the DNAm analyses, the lowest threshold  $P < 0.05$  showed fewer replicated CpGs as compared to the other three thresholds. This was caused by a 333-fold increase of significant CpGs in discovery meta-analysis for BMI and an 8.6-fold increase for smoking when we used lowest threshold in comparison to FDR threshold. In contrast, the discovery meta-analysis showed only a 1.12-fold increase of significant CpGs for age. As a result, the Bonferroni threshold for replication was strongly increased, and most of the previously replicated CpGs did not survive this threshold.

For the normalization options (A) and covariate correction options (D) in RNAseq analyses, the respective differences between the options were unchanged depending on p-value cutoff. For the gene inclusions thresholds (B), it showed that including only the most highly expressed genes yields a slightly higher replication rate using the uncorrected p-value threshold. For the statistical test comparison (C), using lower p-value thresholds (FDR and uncorrected) provided a more pronounced difference between the models.

### ***Categorical analyses for age and BMI***

For DNAm and RNAseq, when we used age/BMI as categorical instead of continuous exposures, the differences between methods remained largely the same. However, the categorical models consistently resulted in a lower number and percentage of significantly replicated CpGs/genes as compared to the continuous models. The only exception was in the hidden confounders (HCs) correction model for age, where the categorical models resulted in larger number of significantly replicated CpGs/genes as compared to continuous models.

## Discussion

We evaluated commonly used analysis strategies for population-based datasets for DNA methylation and RNA sequencing in almost 3,000 participants from four Dutch cohorts. For each step in the analysis procedure, we compared commonly used options and reported their influence on the exposure of interest. These results will aid in comparing studies with different analysis strategies and can help in the choice between alternative analysis strategies.

The four included cohorts differed on some important parameters (e.g.; age). As a combined dataset would not have easily been able to distinguish true age-effects from batch effects between age-differing cohorts, we decided to run cohort-level analyses first and then meta-analyze the datasets, as is commonly done in meta-analyses of ‘omics’ data [36]. As these exposure-differences will also result in different power between cohorts for each exposure, we meta-analyzed each combination of three cohorts and replicated in the fourth [37]. Therefore, when a cohort of low power for a exposure performs poorly as replication cohort, while a powerful cohort for that exposure replicated many signals, these effects were averaged out and provided a reasonable aggregated performance of each strategy [38].

For DNA methylation data, our evaluation lead to the following considerations/recommendations;

*DNAm value preprocessing:* There were no large differences between the different methylation values. We suggest to use beta-3IQR in order to avoid spurious findings based on DNA methylation outliers, but we do not expect another option to have a large influence on the results.

*Statistical tests:* The theoretical advantage of using an RLMM over LM or LMM is considered to be that it is less sensitive to exposure and methylation outliers and heteroscedasticity. However, LM, LMM and RLMM provided nearly identical results, and the analysis running time for RLMM is considerably longer. Therefore, LM or LMM approaches might be preferred as they are simple and widely used base-R functions.

*Cell count adjustment:* Beforehand we expected that differential cell counts are a major influence on DNA methylation data measured from whole blood [27]. Indeed, we observed large influence of cell counts on age, but not on BMI or smoking. These results were in line with previous work which also found that adjusting or not adjusting for blood cell counts had no substantial impact on EWASs of BMI and smoking [39]. For all exposures, we observed influence of Houseman6/3 cell counts on the analysis, with a larger deviation from the measured cell counts (base model) for Houseman3 than Houseman6. Therefore, we recommend the adjustment for measured cell counts if available. If not, the Houseman6 estimated six cell counts could be used for exposures other than age.

*Correction for HCs:* Adjusting for 5 HCs substantially influenced the results. For age, adjusting for 5 HCs substantially decreased the number of replicated CpGs. For BMI and smoking,

adjusting for 5 HCs seemed to improve the results by improving the number of replicated CpGs. Therefore, for exposures other than age, adjusting for HCs is highly recommended in order to remove unknown variation from the data.

For RNA expression data, our evaluation lead to the following considerations/recommendations;

*Normalization method:* There was no large influence of normalization methods. The Voom method resulted in slightly more replicated genes and is recommended.

*Gene expression inclusion threshold:* The gene inclusion threshold displayed minimal influence on the results. To be complete, it is suggested to include and report all genes in the dataset.

*Statistical method:* In our datasets, the standard LM/GLM models performed similarly to the custom limma/edgeR methods. However, it is possible that datasets of smaller sample sizes (e.g. fewer than 20 samples) benefit more from the custom methods. For larger datasets, the standard, widely-used LM and GLM are easier to use and could provide easier compatibility with other applications (e.g. meta-analysis).

*Covariates:* In our results, correcting for PCs did not improve performance, and is not recommended when technical covariates and/or cell counts are available. In our datasets, the PCs correlated to the technical covariates, cell counts and in some occasions to the exposures (mostly age), this likely led to overcorrection when PCs were added on top of these covariates. Correcting for 5 hidden confounders on top of the base model improved the results for all exposures, and is recommended to use. When doing so, care should be taken that the hidden confounders are not correlated to the exposure of interest (or a confounder which is correlated to the exposure) which could remove true results. At current, adjusting for confounders using HCs is not standard practice in RNA-seq analysis, but should be implemented more widely based on these findings. Additionally, we did not use the Bacon package to correct for inflation of test statistics, as this is not yet widely used for RNAseq data. However, applying bacon correction on RNAseq data is becoming more common and should be considered in future RNAseq studies [29].

### ***Evaluation using different p-value cut-offs***

For all models, we observed a balance with more stringent p-value cutoffs resulting in fewer replicated signals, but a larger replication rate. In general, we recommend using Bonferroni corrected p-values with a cutoff of  $p < 0.05$ . The FDR corrected p-values can provide an alternative. Decreasing the p-value threshold stringency always leads to increased false positives and thus a lower replication rate. Using uncorrected p-value cutoffs (whether nominal 0.05 or a too conservative  $1E-8$ ) is not recommended.

For DNAm, the differences between methods were similar for all thresholds, and the main conclusions did not change. For RNAseq, these results further show that the GLM and edgeR's glmQLF models are more conservative (lower number but higher percentage of replicated

signals) while limma's linear model fit is more liberal (higher number but lower percentage of replicated signals) compared to the base model. The LM model is still recommended.

### ***Categorical analyses for age and BMI***

To assess whether strategies are influenced by the continuous or categorical definition of the exposure, we analyzed age and BMI both as continuous and categorical (i.e. highest versus lowest tertiles) exposures of interest. All models responded similarly to the categorical exposure in comparison to the continuous exposure, showing lower number and percentage of replicated signals, indicating lower power for categorical exposures. For both DNAm and RNAseq analyses, we observed differences in performance between models only with HCs correction. The models with 5 HCs for age performed worse when we used age as a categorical variable with highest vs lowest tertiles and excluded the middle tertile. Likely, these results indicate that HCs are insufficiently adjusted for age when it is included as a categorical variable (compared to continuous). Overall, these results seem robust for categorical/continuous exposure definitions, but do emphasize that HCs correction may be challenging when working with categorical exposures.. For continuous variables and most categorical variables (e.g. BMI tertiles and smoking), using HCs performed best and is still recommended.

Although most of the differences we observed between strategies were consistent across exposures and cohorts, these results might not be applicable to all other DNAm array or RNA-seq studies. For example, we have studied three exposures for which we could observe relatively large differences in blood methylation or expression, with the exception of BMI in methylation. We observed differences in performance between exposures, for example when correcting for different cell counts, HCs or PCs in age, or the low number of replicated CpGs for BMI. As such, a universally optimal model could not be defined and performance of these different strategies needs to be confirmed for other exposures. However, performance differences between many strategies were consistent across exposures (specifically BMI and smoking), individual cohorts and DNAm/RNA-seq datasets, and will likely hold even in other exposures or datasets.

In this study, we have compared multiple analysis strategies on four cohorts and suggested a base model to reduce heterogeneity between studies. The most ideal validation would be to re-analyze a number of published studies using this optimal model and demonstrate a decrease in heterogeneity between results of previous analyses and those with the new model. However, to our knowledge, for none of the studies we investigated this was possible, due to lack of publically available phenotypic information or lack of publically available individual level DNAm/RNAseq data. As it may not always be possible to share such data publicly, this further shows the need for more standardized DNAm/RNAseq methods, so results between studies can be compared more easily.

Similarly, we studied four relatively large population-based studies. Results obtained from smaller studies, or other types of populations, for example patients or samples of extreme exposures, might yield different results and require alternative strategies. These comparisons were beyond the scope of our study, which focused on commonly used strategies. Our results might be most generalizable to population-based DNAm and RNA-Seq studies. Finally, our

study lacked a gold standard, which will have limited our ability to distinguish strategies with many false positives from strategies with a high sensitivity. Despite these factors, we evaluated the consistent influences of analysis strategies and options, and reported analysis suggestions for both datatypes. We hope that these results will aid other researchers in selecting an appropriate analysis strategy and/or in evaluating the impact, a certain strategy might have had on the observed results.

## Conclusions

Based on our findings, for DNA methylation studies we recommend to correct for measured cell counts when available and include additional hidden confounders (independent of cell-counts and technical covariates) in the statistical model. We suggest using Beta-3IQR values and the LM statistical test for DNAm studies, although alternatives will yield similar results and can also be used. For RNA sequencing studies, we recommend using hidden confounders in addition to technical covariates and measured cell counts. The use of principal components is not recommended. We recommend using the Voom normalization method, and suggest to include all genes in the analysis (independent of expression level). Finally we suggest using a LM or GLM statistical model for large studies and a custom method like limma/edgeR for smaller studies. Our results show a large difference in replication results between cohorts, and therefore using replication in DNAm or RNA-seq analysis is also recommend.



## Methods

### **Data generation**

Generation of the BIOS gene expression dataset was described previously [34, 35]. In short, DNA and RNA were collected from 3,296 unrelated participants of six Dutch populations as described below. Analyses were restricted to four large cohorts; LifeLines (LL), Leiden Longevity Study (LLS), Netherlands Twin Register (NTR) and Rotterdam Study (RS). We included 2,950 participants with DNAm array data and 2,829 participants with RNA-seq data. Characteristics for these cohorts are described in table 1.

### **DNA methylation data**

Whole blood was used to isolate genomic DNA. 500 ng of genomic DNA was bisulfite converted using the EZ DNA Methylation kit (Zymo Research, Irvine, CA, USA). Methylation profiling was then performed using Infinium Illumina HumanMethylation 450k arrays according to the manufacturer's protocol. Quality control of the samples was performed using MethylAid [40]. Probes with either a high detection P value ( $> 0.01$ ), low bead count ( $< 3$  beads), or low success rate (missing in  $> 5\%$  of the samples) were set to missing. Samples were excluded from the analysis if they contained an excess of missing probes ( $> 5\%$ ). Imputation was performed per cohort, subsequently, to impute the missing values [41]. The raw beta-values were normalized using functional normalization [22] as implemented in the minfi package [42]. The normalized beta-values were  $\log_2$  transformed to produce M-values [42].

### **RNA-seq data**

Total RNA was derived from whole blood, depleted of globin transcripts using Ambion GLOBINclear and subsequently processed using the Illumina TruSeq v2 library preparation kit. On average 40 million paired-end reads of 50bp were generated per participant using Illumina's HiSeq 2000. Samples were demultiplexed using CASAVA and aligned to the hg19 reference genome using STAR [43]. Alignments were sorted, read groups were added using picard [44] and gene expression was quantified using featureCounts [45]. We selected participants for which all covariates were available (sex, age, BMI, smoking status and measured cell counts). Raw count matrices per cohort were used for analysis.

### **Base model and analysis**

The main steps in epigenomic and transcriptomic analyses often vary between studies, as shown in figures 1a and figure 1b, respectively. First, we compiled a base model with a single option from each step in figure 1a and 1b. These options were then replaced, one at a time, in the various analysis strategies. These strategies were applied to three exposures of interest (age, BMI and smoking status) in each cohort (LL, LLS, NTR and RS). Every combination of three discovery cohorts was meta-analyzed and replicated in the remaining cohort (leave-one-out method). The average number and percentage of replicated CpGs/genes were calculated from these four results and were used to evaluate the performance of each strategy. Age, sex, measured percentages of WBC counts (granulocytes, lymphocytes and monocytes) and technical covariates specified below, were included as covariates unless specified otherwise. Replication analyses were always Bonferroni corrected. Meta-analyses was performed using GWAMA (DNAm array data) [46] or METAL (RNA-seq data) [47].

### **DNA methylation array specific analysis strategies**

The technical covariates used for each DNAm array analysis were bisulfite conversion plate and array row. All analyses were corrected for inflation and bias using the Bacon package [29], which estimates empirical null distribution using the Bayesian method. The following steps were investigated in detail (see figure 1a).

- A. *Methylation values:* We investigated five types of DNAm values, namely (1) beta values, representing the percentage of methylation between 0 (unmethylated) and 1 (methylated) [25]; (2) beta-3IQR values, where beta values of outlier samples per methylation CpG were removed (replaced with NAs) using the 3 interquartile range (IQR) strategy, i.e. any beta value below quartile  $(Q_1 - 3 \times \text{IQR})$  or above  $Q_3 + 3 \times \text{IQR}$  was removed [48]; (3) M-values, calculated as the  $\log_2$  ratio of the methylated probe intensity and unmethylated probe intensity [49]; (4) M-3IQR values, where M-values of outlier samples per methylation CpG were removed using the 3xIQR strategy as described above [48]; (5) RIN (rank-based inverse normal transformation) values, wherein beta-values for each sample were ranked and replaced with the corresponding standard normal quantiles in order to create a normal distribution [50]. We selected beta-3IQR values for the base model.
- B. *Statistical tests:* We investigated three types of linear models: (1) Linear regression model (LM), (2) Linear regression mixed model (LMM) and (3) Robust linear regression mixed model (RLMM). We selected LM for the base model.
- C. *Cell count correction:* (1) For the base model, we used the percentages of differential measured cell counts of granulocytes, lymphocytes and monocytes. This base model was compared with 3 other models: (2) a model without cell count correction, (3) a model adjusted for the cell subtypes imputed with the reference-based Houseman method [26], using the default percentage counts of all six imputed cell types; granulocytes, monocytes, NK cells, B cells, CD4+ and CD8+ T-lymphocytes. We refer to this as “Houseman6”, (4) a model adjusted for the same imputed cell counts, but using three instead of six cell types; granulocytes, monocytes and lymphocytes (sum of NK cells, B cells, CD4+ and CD8+ T-lymphocytes) in order to match with measured cell counts of the base model. We refer to this as “Houseman3”.
- D. *Hidden confounder (HCs) correction;* (1) For the base model, we used known technical confounder correction (bisulfite conversion plate and array row). This base model was compared with three more models that were corrected for HCs calculated from the CATE package [28, 29]. These were calculated per cohort per exposure. (2) We calculated 5 HCs independent of the exposure of interest (BMI or smoking), age, sex and known technical covariates. However, we did not regress out measured differential cell counts, and therefore we assume that the HCs reflect cell counts. This model contained age, sex, technical confounders and 5 HCs as covariates. (3) HCs were calculated by regressing out the exposure of interest, age, sex and also measured differential cell counts. In this case, we did not regress out known technical confounders and therefore these HCs are thought to reflect technical confounders. This model contained age, sex, measured differential cell counts and 5 HCs as covariates. (4) HCs were calculated by regressing out

not only the exposure of interest, age and sex, but also measured differential cell counts and known technical covariates. In this case, HCs can be regarded as any more potential hidden biological or technical confounders that might influence the data in addition to the differential cell counts and technical confounders' correction. This model contained age, sex, measured differential cell counts, known technical confounders and 5 HCs as covariates.

### ***RNA sequencing specific analysis strategies***

All RNA-seq strategies were corrected for technical covariates; sequencing batch (flow cell) and average GC percentage in the reads, in addition to the biological covariates mentioned before. We compared the following steps in detail (see also figure 1b).

- A. Normalization method; Three commonly used RNA-seq normalization methods; (1) Voom, (2) edgeR and (3) DESeq, were investigated. The edgeR and DESeq methods adopted a Trimmed mean of M-values normalization (TMM) [51, 52]. Voom adopted edgeR's normalization but first raised zeros to a minimum value of 1 and performed a log transformation [53]. We selected Voom for the base model.
- B. Expression inclusion criteria; We varied the genes allotted to normalization using four common inclusion CPM (counts per million) thresholds of gene expression. (1) All genes expressed at any level in at least one sample were included. (2) All genes with a CPM  $\geq 1$  in  $\geq 20\%$  of the samples were included. (3) Genes with an average CPM  $\geq 1$  across all samples were included. (4) All genes with an average CPM  $\geq 10$  across all samples were included. In the base model, all genes were included (option 1).
- C. Statistical tests; We used four commonly used statistical tests. (1) A default linear model (LM) [54]. (2) A default generalized linear model (GLM) with negative binomial distribution. (3) The linear model fitfunction of the limma package, which was a weighted linear model where genes with a large variance (e.g. genes with very low expression) had lower weights. (4) The edgeR's generalized linear model fit (glmQLF), which used a negative binomial distribution followed by a log ratio likelihood (LR) test. Options 3 and 4 were RNA-seq specific hierarchical models that take into account differences in variance estimates across genes. [51, 53]. Option 1 was included in the base model. Option 4 was also run on the Voom normalized dataset. Option 2 and 3 were run on the edgeR normalized dataset as the negative binomial distribution did not apply after Voom's log transformation.
- D. Technical correction; We used five commonly used approaches to correct for technical factors. (1) We included technical covariates (GC percentage and flow cell) and measured cell-counts. (2) Corrected only for technical covariates. (3) Corrected only for cell-counts. (4) Replaced technical covariates and cell-counts by the first five principal components PCs, calculated per cohort using the prcomp function in R. (5) Added five hidden confounders to the technical covariates and cell counts. Hidden confounders were calculated per cohort per exposure, and were adjusted for the respective exposure, age, sex, technical covariates and cell-counts.

### ***Evaluating strategy performance***

In each analysis, three of the four cohorts were meta-analyzed in the discovery and the fourth cohort was used for replication. We repeated for each combination of three discovery and one replication cohort. The number of significantly replicated CpGs/genes were obtained for each repetition, as well as the percentage of CpGs/genes from discovery that reached replication (replication rate). For both the number and percentage of replicated signals, the average of the four combinations was calculated and used to evaluate performance of each strategy. We compared each strategy to the base model and looked for consistent differences in replication number or percentage across exposures.

### ***Categorical analyses for age and BMI***

In order to investigate whether an optimal analysis strategy is dependent on whether the independent variable is continuous or categorical, we expanded our association analyses on age and BMI by converting them into tertiles. We used the highest and lowest tertiles to define the categories. The results of these categorical analyses were compared with the results of the continuous analyses where age and BMI were used as continuous measures. For DNAm, we did not analyze BMI into categorical exposure because the numbers of significantly replicated CpGs were already small for the continuous models (average of <12 CpGs) when a Bonferroni threshold was used for multiple testing. This made it difficult to draw conclusions when comparing different methods within continuous models, and therefore would have made it even more difficult to compare results between categorical models.

### ***Evaluation using different p-value cut-offs***

For all the comparisons mentioned, both discovery and replication results were Bonferroni corrected. In addition to using the Bonferroni threshold for the discovery results, we applied three other thresholds to evaluate the robustness of the approaches: (1) Benjamini-Hochberg FDR threshold (FDR p-value <0.05) (2) highest threshold (uncorrected p-value threshold <1x10<sup>-8</sup>) and (3) lowest threshold (uncorrected p-value threshold <0.05). Differences between models were compared between p-value thresholds to establish that the models show similar (respective) results independent of p-value thresholds.

In addition, for each strategy, we performed a meta-analysis of all four cohorts for DNA methylation and RNA expression. Overlaps in CpGs/genes between all strategies per step were determined using Venn diagrams to ascertain if the same CpGs/genes were identified between strategies [55].

## Declarations

### ***Author's Contributions***

Authors JvR and PM performed all analyses and wrote the manuscript. The BIOS datasets were generated and QC'ed by the BIOS Consortium, as is described in detail here; <http://www.bbmri.nl/acquisition-use-analyze/bios/>, including details on contributions of all consortium members. JvR, PM, AC, JF, JvD, RJ, LF, PH, BH and JvM contributed to analysis and interpretation through regular calls and revisions of the manuscript.

### ***Competing Interests***

The authors have no competing interests to report.

### ***Funding***

This work was partially funded by BBMRI-NL, a research infrastructure financed by the Netherlands Organization for Scientific Research (NWO project 184.021.007).

### ***Ethics Approval***

Each biobank received ethical approval for their population study. No additional ethical approval was required.

### ***Data Availability***

The datasets from BIOS are available from the European Genome-Phenome Archive by accession number EGAS00001001077 (<https://www.ebi.ac.uk/ega/studies/EGAS00001001077>). Alternative option to access the data are available through the BIOS website; <https://www.bbmri.nl/acquisition-use-analyze/bios/> [35].

## References

1. Heyn, H., et al., *Distinct DNA methylomes of newborns and centenarians*. Proc Natl Acad Sci U S A, 2012. **109**(26): p. 10522-7.
2. Lork, K., et al., *DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns*. Genome Biol, 2014. **15**(4): p. r54.
3. Consortium, G.T., et al., *Genetic effects on gene expression across human tissues*. Nature, 2017. **550**(7675): p. 204-213.
4. Peters, M.J., et al., *The transcriptional landscape of age in human peripheral blood*. Nat Commun, 2015. **6**: p. 8570.
5. Joehanes, R., et al., *Epigenetic Signatures of Cigarette Smoking*. Circ Cardiovasc Genet, 2016. **9**(5): p. 436-447.
6. Breitling, L.P., et al., *Tobacco-smoking-related differential DNA methylation: 27K discovery and replication*. Am J Hum Genet, 2011. **88**(4): p. 450-7.
7. Breitling, L.P., et al., *Smoking, F2RL3 methylation, and prognosis in stable coronary heart disease*. Eur Heart J, 2012. **33**(22): p. 2841-8.
8. Wan, E.S., et al., *Smoking-Associated Site-Specific Differential Methylation in Buccal Mucosa in the COPD Gene Study*. Am J Respir Cell Mol Biol, 2015. **53**(2): p. 246-54.
9. Zeilinger, S., et al., *Tobacco smoking leads to extensive genome-wide changes in DNA methylation*. PLoS One, 2013. **8**(5): p. e63812.
10. Shenker, N.S., et al., *DNA methylation as a long-term biomarker of exposure to tobacco smoke*. Epidemiology, 2013. **24**(5): p. 712-6.
11. Shenker, N.S., et al., *Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking*. Hum Mol Genet, 2013. **22**(5): p. 843-51.
12. Guida, F., et al., *Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation*. Hum Mol Genet, 2015. **24**(8): p. 2349-59.
13. Qiu, W., et al., *The impact of genetic variation and cigarette smoke on DNA methylation in current and former smokers from the COPD Gene study*. Epigenetics, 2015. **10**(11): p. 1064-73.
14. Gao, X., et al., *DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies*. Clin Epigenetics, 2015. **7**: p. 113.
15. Wan, E.S., et al., *Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome*. Hum Mol Genet, 2012. **21**(13): p. 3073-82.
16. Huan, T., et al., *A whole-blood transcriptome meta-analysis identifies gene expression signatures of cigarette smoking*. Hum Mol Genet, 2016. **25**(21): p. 4611-4623.
17. Vink, J.M., et al., *Differential gene expression patterns between smokers and non-smokers: cause or consequence?* Addict Biol, 2017. **22**(2): p. 550-560.
18. Beineke, P., et al., *A whole blood gene expression-based signature for smoking status*. BMC Med Genomics, 2012. **5**: p. 58.
19. Verdugo, R.A., et al., *Graphical modeling of gene expression in monocytes suggests molecular mechanisms explaining increased atherosclerosis in smokers*. PLoS One, 2013. **8**(1): p. e50888.
20. Wu, M.C., et al., *A systematic assessment of normalization approaches for the Infinium 450K methylation platform*. Epigenetics, 2014. **9**(2): p. 318-29.

21. Wang, T., et al., *A systematic study of normalization methods for Infinium 450K methylation data using whole-genome bisulfite sequencing data*. *Epigenetics*, 2015. **10**(7): p. 662-9.
22. Fortin, J.P., et al., *Functional normalization of 450k methylation array data improves replication in large cancer studies*. *Genome Biol*, 2014. **15**(12): p. 503.
23. Pidsley, R., et al., *A data-driven approach to preprocessing Illumina 450K methylation array data*. *BMC Genomics*, 2013. **14**: p. 293.
24. Marabita, F., et al., *An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform*. *Epigenetics*, 2013. **8**(3): p. 333-46.
25. Du, P., et al., *Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis*. *BMC Bioinformatics*, 2010. **11**: p. 587.
26. Houseman, E.A., et al., *DNA methylation arrays as surrogate measures of cell mixture distribution*. *BMC Bioinformatics*, 2012. **13**: p. 86.
27. Reinius, L.E., et al., *Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility*. *PLoS One*, 2012. **7**(7): p. e41361.
28. Wang, J., et al., *Confounder adjustment in multiple hypothesis testing*. . arXiv:1508.04178, 2015.
29. van Iterson, M., et al., *Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution*. *Genome Biol*, 2017. **18**(1): p. 19.
30. Li, P., et al., *Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data*. *BMC Bioinformatics*, 2015. **16**: p. 347.
31. Zhao, S. and B. Zhang, *A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification*. *BMC Genomics*, 2015. **16**: p. 97.
32. Bullard, J.H., et al., *Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments*. *BMC Bioinformatics*, 2010. **11**: p. 94.
33. Robles, J.A., et al., *Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing*. *BMC Genomics*, 2012. **13**: p. 484.
34. Zhernakova, D.V., et al., *Identification of context-dependent expression quantitative trait loci in whole blood*. *Nat Genet*, 2017. **49**(1): p. 139-145.
35. Bonder, M.J., et al., *Disease variants alter transcription factor levels and methylation of their binding sites*. *Nat Genet*, 2017. **49**(1): p. 131-138.
36. Copetti, M., et al., *Advances in meta-analysis: examples from internal medicine to neurology*. *Neuroepidemiology*, 2014. **42**(1): p. 59-67.
37. George, N.I., et al., *An Iterative Leave-One-Out Approach to Outlier Detection in RNA-Seq Data*. *PLoS One*, 2015. **10**(6): p. e0125224.
38. Evangelou, E. and J.P. Ioannidis, *Meta-analysis methods for genome-wide association studies and beyond*. *Nat Rev Genet*, 2013. **14**(6): p. 379-89.
39. Heiss, J.A. and H. Brenner, *Impact of confounding by leukocyte composition on associations of leukocyte DNA methylation with common risk factors*. *Epigenomics*, 2017. **9**(5): p. 659-668.
40. van Iterson, M., et al., *MethylAid: visual and interactive quality control of large Illumina 450k datasets*. *Bioinformatics*, 2014. **30**(23): p. 3435-7.
41. Hastie, T.T., R.; Narasimhan, B.; Chu, G., *impute: impute: Imputation for microarray data. R package version 1.56.0*. . 2018.
42. Aryee, M.J., et al., *Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays*. *Bioinformatics*, 2014. **30**(10): p. 1363-9.
43. Dobin, A. and T.R. Gingeras, *Optimizing RNA-Seq Mapping with STAR*. *Methods Mol Biol*, 2016. **1415**: p. 245-62.
44. Picard, *Picard Toolkit*. 2018.

45. Liao, Y., G.K. Smyth, and W. Shi, *featureCounts: an efficient general purpose program for assigning sequence reads to genomic features*. *Bioinformatics*, 2014. **30**(7): p. 923-30.
46. Magi, R. and A.P. Morris, *GWAMA: software for genome-wide association meta-analysis*. *BMC Bioinformatics*, 2010. **11**: p. 288.
47. Willer, C.J., Y. Li, and G.R. Abecasis, *METAL: fast and efficient meta-analysis of genomewide association scans*. *Bioinformatics*, 2010. **26**(17): p. 2190-1.
48. Upton, G. and I. Cook, *Understanding Statistics*. 1997.
49. Wood, A.R., et al., *Defining the role of common variation in the genomic and biological architecture of adult human height*. *Nat Genet*, 2014. **46**(11): p. 1173-86.
50. Beasley, T.M., S. Erickson, and D.B. Allison, *Rank-based inverse normal transformations are increasingly used, but are they merited?* *Behav Genet*, 2009. **39**(5): p. 580-95.
51. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data*. *Bioinformatics*, 2010. **26**(1): p. 139-40.
52. Anders, S. and W. Huber, *Differential expression analysis for sequence count data*. *Genome Biol*, 2010. **11**(10): p. R106.
53. Law, C.W., et al., *voom: Precision weights unlock linear model analysis tools for RNA-seq read counts*. *Genome Biol*, 2014. **15**(2): p. R29.
54. Core Team, R., *R: A Language and Environment for Statistical Computing*. R Core Team; Vienna 2015., 2015.
55. Heberle, H., et al., *InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams*. *BMC Bioinformatics*, 2015. **16**: p. 169.



# Chapter 3.2



## Hippocampal transcriptome profiling combined with protein-protein interaction analysis elucidates Alzheimer's Disease pathways and genes

Jeroen G.J. van Rooij<sup>1,2</sup>, Lieke H.H. Meeter<sup>1</sup>, Shami Melhem<sup>1</sup>, Diana A.T. Nijholt<sup>1</sup>, Tsz Hang Wong<sup>1</sup>, Netherlands Brain Bank<sup>3</sup>, Annemieke Rozemuller<sup>4</sup>, Andre G. Uitterlinden<sup>2</sup>, Joyce C. van Meurs<sup>2,†</sup>, John C. van Swieten<sup>1,†</sup>

<sup>1</sup> Department of Neurology, Erasmus Medical Center, Rotterdam, the Netherlands

<sup>2</sup> Department of Internal Medicine, Erasmus Medical Center, Rotterdam, the Netherlands

<sup>3</sup> Netherlands Institute for Neuroscience, Amsterdam, the Netherlands

<sup>4</sup> Department of Pathology, VU University Medical Center, Amsterdam, the Netherlands

<sup>†</sup>These authors contributed equally

*Published on October 29<sup>th</sup> 2018 in Neurobiology of Aging (IF=4.4).  
PMID: 30497016, doi: 10.1016/j.neurobiolaging.2018.10.023*

## Abstract

Transcriptomics in Alzheimer's disease (AD) brains compared to healthy controls provides crucial information about disease pathophysiology. We performed whole transcriptome sequencing on hippocampus of 20 AD cases and 10 age- and sex-matched cognitively healthy controls. We grouped 735 of 2,716 differentially expressed genes in 33 modules based on protein-protein interaction (PPI) data. Enrichment analysis of these modules showed involvement in signal transduction, transport, response to stimulus and several metabolic pathways. 16 modules interacted with previously described AD disease genes. 48% of the differentially expressed genes replicated in another dataset, as well as 64% of enriched biological processes. Clustering genes by PPI data before gene set enrichment identifies specific biological processes involved in AD, providing additional details to traditional gene set enrichment analysis. Additional data, like large gene co-expression networks and unbiased annotation databases are needed to gather complete and robust networks and provide more insight in AD biology.

## Introduction

Alzheimer's Disease (AD) is a neurodegenerative disorder with progressive loss of memory and other cognitive domains, currently affecting over 40 million individuals worldwide (Prince, et al., 2013, Scheltens, et al., 2016). Previous studies have shown neurodegenerative changes in the hippocampus an estimated 15-20 years before symptom onset (Boyle, et al., 2013, Karran, et al., 2011, Murray, et al., 2011). The main pathological features are amyloid plaques and tau tangles in variable severity and localization throughout the brain (Braak and Braak, 1995, Holtzman, et al., 2016, Jellinger, 2008, Selkoe and Hardy, 2016, Thal, et al., 2014, Tomiyama, 2010). Several genetic loci with an effect on incidence and age at onset have been identified, although their exact pathophysiological mechanisms remain largely unknown (Bekris, et al., 2010, Lambert, et al., 2013).

In the last decade, transcriptomic studies on post-mortem AD brain tissue have been performed to further our understanding of AD biology (Kavanagh, et al., 2013, Sutherland, et al., 2011). However, large heterogeneity in study design (i.e.; choice of brain region, case definition, laboratory protocol, quantification method, analysis procedure and mode of reporting) limit our ability to compare results clearly across studies. Most transcriptomic studies determine differentially expressed genes between cases and controls in an affected brain region and report top genes and enriched pathways (Ashburner, et al., 2000, Gene Ontology, 2015, Ogata, et al., 1999). Although genes and pathways vary between studies, usually a decrease in synaptic transmission, mitochondrial function and cytoskeleton biology are reported, while pathways involved in immune response, inflammation and apoptosis are upregulated in AD (Liang, et al., 2008, Ray and Zhang, 2010, Sekar, et al., 2015, Twine, et al., 2011). Recently, protein-protein interaction networks, multiple-omics data analysis or gene co-expression network analysis are utilized to provide more extensive and robust insights in these results (Chi, et al., 2016, C. Humphries, et al., 2015, C.E. Humphries, et al., 2015, Kong, et al., 2015, Kong, et al., 2014). Most notably, a large study by Wang et al investigated gene expression levels through microarrays in more than a thousand brain samples, spread across 19 regions in 125 individuals, including 17 hippocampus controls and 38 cases of varying (braak and CERAD) stages (Wang, et al., 2016). Comparing co-expression modules between individuals of varying AD stages they identified dysregulated networks and functions, and suggested that some of those originated from early disease stages and might reflect causal mechanisms. Analytical tools like PPI and co-expression networking permit us to mine deeper into complex transcriptomic data, and to study genes as part of a larger network, rather than as individual entities. This allows interpretation of the AD transcriptome in its entire complexity, and will ultimately aid in understanding how individual genes, risk factors and pathways interplay to a complex disease phenotype.

We compared whole transcriptome sequencing of 20 AD cases with 10 age- and sex-matched cognitively healthy controls. We report all differentially expressed genes and replicate these in a second independent dataset (van der Brug H, 2017). We show that the identified pathways are usually represented in AD expression literature (Chi, et al., 2016, C.E. Humphries, et al., 2015, Liang, et al., 2008, Ray and Zhang, 2010, Sekar, et al., 2015, Twine, et al., 2011, Wang, et al., 2016).

Then, we implement protein-protein interaction data and gene network clustering, we identify subsets of functionally annotated gene modules, representing components of those commonly detected pathways (van Dongen and Abreu-Goodger, 2012, von Mering, et al., 2003). We chose this

approach as the AD network is large and traditional enrichment analysis only uncovers the main involved biological processes. By extracting specific subsets of genes this network can be studied in more detail. We replicate these modules in a second study and use the modules to reconstruct the AD transcriptome. Finally, we investigate which modules interact with known AD risk genes and which genes are the strongest interactors in the network <sup>(Van Cauwenberghe, et al., 2016)</sup>.

Our results indicate that using PPI and networking analysis provides a detailed overview of the biological processes underlying AD, and provide insight into gene and pathway contributions to AD biology.

## Methods

### Subject selection

20 AD brains were selected from the Netherlands Brain Bank (Braak and Braak, 1995, Mirra, et al., 1991), equally divided into homozygous carriers with and without ApoE4 (10 e44, 9 e33 and 1 e32 carrier). These AD cases were matched for age and gender with brains from 10 non-demented cognitively healthy controls (6 e33 and 4 e32 carriers, Table 1).

**Table 1.** Study sample characteristics. An asterisk denotes statistically significant difference compared to controls. All values represent means with standard deviations unless otherwise indicated. "Cases\_QC" indicates metrics after removing two outlier cases.

	Controls	Cases	Cases_QC
Number	10	20	18
Gender (%Male)	50%	30%	44%
Age ( $\pm$ SD)	76 $\pm$ 12	75 $\pm$ 7	75 $\pm$ 7
Braak	1.5 $\pm$ 1.3	5.5 $\pm$ 0.5*	5.6 $\pm$ 0.5*
amyloid	0.9 $\pm$ 1.1	2.9 $\pm$ 0.3*	2.9 $\pm$ 0.3*
pmd	551 $\pm$ 297	348 $\pm$ 108*	329 $\pm$ 98*
pH	6.6 $\pm$ 0.3	6.3 $\pm$ 0.3*	6.3 $\pm$ 0.3*
brain weight	1319 $\pm$ 240	1045 $\pm$ 119*	1035 $\pm$ 113*
apoe (32/33/44)	4/6/0	1/9/10*	1/8/9*

### RNA collection

The hippocampus of all brains were sectioned in 8 - 10 30 um sections using a cryostat (Thermo Fischer HM560 at -20C). The dentate gyrus and cornu amonis were macro-dissected from the hippocampus tissue and dissolved in lysis buffer of the Qiagen AllPrep DNA/RNA/miRNA Universal kit (Qiagen; Cat No. 80224). Samples were never thawed before reaching the lysis buffer. Total RNA was isolated using the manufacturers protocol.

### Sequencing

Library prep was performed using Illumina's TruSeq RNAseq library prep, using the manufacturers protocol, including polyA-tail selection and acoustic shearing. Sequencing was performed at the Human Genomics facility (HUGE-F, www.glimdna.org) on a HiSeq2000 at 2x50bp, with all samples randomly assigned to a maximum of 4 per lane.

### Data processing

Demultiplexing was performed using CASAVA (v1.8.2), followed by adaptor pruning using trim-o-matic (v0.33) and genome alignment to hg19 using STAR (v2.3.0) (Bolger, et al., 2014, Dobin, et al., 2013). Then, read sorting, pairing and reordering was performed using picard (v1.90), and read and alignment quality control (QC) was performed using fastQC (v0.11.3). Transcript quantification (counts) was performed using featurecounts (v1.4.3) against all 57,820 gene features in GENCODE (version date; 2013-12-05) (Harrow, et al., 2012, Liao, et al., 2014).

### **Data analysis**

Counts were normalized using the edgeR (v3.8.6) trimmed mean of M-values (TMM) method to counts per million (CPM) values, and all low-abundant features were omitted (<1 CPM in 75% of samples). Principal components (PCs) were calculated using “prcomp” in R, and then plotted to visually identify sample outliers. Statistical analysis was performed per gene using the exactTest function in edgeR, reporting FDR corrected p-values and 2log fold changes. The statistical analysis was adjusted for age, gender and the first 2 principal components to correct for cell composition differences (McCarthy, et al., 2012, Robinson, et al., 2010). We combined FDR-corrected p-values and log fold changes to a separate differential expression score (DE score);  $(-\log_{10}(pFDR))/10 * ((\text{SQRT}(\log FC * \log FC))/3) = \text{DE score}$ , with a maximum score of 1 per category, and retained genes with a DE score  $\geq 0.10$  for further analysis (scores shown in Supplemental Table 1).

### **Protein-protein interaction (PPI) modules**

We downloaded the STRING human protein interaction database (v10) and extracted all experimental, co-expression or database interactions scored  $\geq 500$  (von Mering, et al., 2003). This interaction network was imported to Cytoscape (v3.4.0) and subjected to the Markov Clustering Algorithm (MCL, inflation factor 2.0) to identify gene modules (Morris, et al., 2011, Smoot, et al., 2011). In short; MCL clusters graphical data to retain sets of genes (modules) with more interaction within the module than to the rest of the network, while retaining as much of the original network as possible (Enright, et al., 2002, van Dongen and Abreu-Goodger, 2012). MCL revolves around one main parameter; the inflation factor, which determines the module sizes. We ran MCL with various inflation factors, aiming for modules no larger than 100 genes (to separate biological processes by module) but not smaller than 10 (to allow module enrichment analysis) (Subramanian, et al., 2005). The final module enrichment analysis was done with inflation factor 2.0. Each gene can only be assigned to a single module. Per gene, we calculated an overall gene contribution score by multiplying the DE score and absolute number of interactions.

### **Replication of DE genes and modules**

We identified 3 possible replication datasets on GEO, with an RNA-Seq expression matrix of AD brain tissue and non-demented controls (GSE53697, GSE67333, GSE95587) (Magistri, et al., 2015, Plouhinec, et al., 2014, van der Brug H, 2017). Since GSE53697 focusses on different tissue (Dorsolateral Prefrontal Cortex) and GSE67333 was small in sample size (n=8), we used dataset GE95587 for further replication of our discovery findings. Each dataset was analyzed as described for the discovery set; normalizing in edgeR, correcting for age, gender and 2 PCs, resulting in FDR corrected p-values per gene per dataset. For each gene DE scores were calculated, and a network was based on PPI data and subsequent modules where clustered and annotated using MCL and webgestalt.

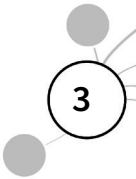
### **Gene set enrichment**

Gene set enrichments were performed using Webgestalt (v27-1-17) against KEGG pathways and GO-terms (Gene Ontology, 2001, Ogata, et al., 1999). For GO enrichment the “noRedundant” terms were used. All enrichments were FDR (Benjamini-Hochberg) corrected, using a threshold of  $p < 0.05$  for statistical significance. Different background gene lists were used depending on the specific analysis, as shown in figure 1. For the modules identified by MCL clustering, the first three enriched Gene Ontology Biological Process (GOBP) terms were extracted and

mapped back on the Gene Ontology family tree (shown in supplementary figure 3) (Ashburner, et al., 2000, Carbon, et al., 2009) Based on the tree the modules are grouped to GO Biological Processes (denoted as branches), and annotated by their specific (sub-)terms.

### ***Mapping known AD genes***

We selected a list of 27 known genetic risk factors known to be involved in AD, compiled from all known AD GWAS loci (i.e. BIN1, SORL1) and AD Mendelian causal genes (i.e. APP, PSEN1) as reviewed recently (Lambert, et al., 2013, Van Cauwenberghe, et al., 2016). All interactions between these genes and the modules were extracted from STRING (independent of DE score of the AD gene), distinguishing between experimental and database interactions, using a cutoff of  $\geq 500$ . All AD gene - module connection of at least two interactions were reported.



## RESULTS

### ***Study sample characteristics***

The demographic data of the AD group did not differ from the control group, with a mean age of death of 75, and 30% male (table 1). Mean brain weight was lower (1045 grams) in the AD than in the control group (1319 grams). Braak and CERAD stage was higher in the AD group than the control group (5.5 and 2.9 versus 1.5 and 0.9, respectively). Post-mortem delay was significantly shorter for cases, 348 versus 551 minutes and brain pH was lower (6.3 versus 6.6) in AD cases. All sequencing quality and alignment QC metrics were similar between groups, with an average of 48,772,000 reads per sample. QC by PCA showed two outliers, which could not be resolved using PCs and other correction factors in the statistical model. These two cases showed significant upregulation in approximately 150 genes compared to both the 10 controls and the other 18 cases. The highest upregulated gene was *TTR*, which is highly expressed in the choroid plexus. Presence of the choroid plexus was confirmed using routine staining and both cases were excluded from further analysis.

### ***Generating protein-protein interaction (PPI) modules of differentially expressed genes***

Using a differential expression score (DE score) of  $\geq 0.1$ , 2,716 genes (19% of 14,564 detected genes) were selected for analysis, as shown in figure 2. 1,671 were downregulated (62%), 1,045 upregulated. In total 8,676 interactions occurred in this dataset, representing 1,610 genes (59%). STRING interactions were either based on “database” (5,467, 63%), “experimental” (2,915, 34%) or “co-expression” (294, 3%). The network was clustered into 33 gene modules, ranging from 10 - 90 genes per module as shown in table 2. 735 genes (46%) were assigned to one or more module(s), with 87% (7,514) of all interactions contained within these modules. Figure 3 shows the various modules and how the remaining 23% of interactions is distributed between modules. The expression table, gene-module assignments and interaction lists can be found in the supplements.



**Table 2.** A) Overview of 33 gene modules, number of genes and interactions, the dominant interaction type (db=STRING-database, exp=STRING-experimental, coe=STRING-coexpression), strongly centralized genes (Hub genes, defined as being involved in at least 30% of all interactions within the module) and the amount of interactions to other modules. The nine columns “organic substance metabolic process” to “other biological processes” represent the main biological processes for each identified gene ontology branch. Per module, the branch(es) in which the three most enriched terms are located are shown, each plus sign representing an enriched biological process in that branch. The last columns indicate how many replication modules overlap with each discovery modules, and what the highest percentage of overlap with a single module is. B) Overview of AD expression studies reporting enriched pathways or GO biological processes in their main tables. Crosses indicate enrichment of terms in this particular branch.

Module	Number of Genes	Interactions within Module	main interaction type	Hubs ≥ 30% interactions in module)	Interactions to other Modules	Organic substance metabolic process	Signal Transduction	Transport	Regulation of biological process	cellular metabolic process	cellular component organization	other metabolic processes	response to stimulus	other biological processes	Replication Modules	Highest Overlap
M1	90	1238	98% db		110		+++								1	56%
M2	52	181	85% db		136			+	+					+	8	12%
M3	39	107	81% db	CREBBP	157	+							++		6	18%
M4	35	62	69% exp	HDAC1	62	++			+						2	34%
M5	32	334	90% db		15			+	+		+				1	34%
M6	31	223	96% db		43			+	++						1	48%
M7	30	177	70% db		71			+	+					+	2	23%
M8	29	56	79% exp	CALM1	34	+				++					2	31%
M9	27	90	59% db		47			+	+					+	0	-
M10	26	113	85% db		116			+			+			+	1	35%
M11	21	55	78% db		102	+	+				+				2	29%
M12	22	83	51% db	RPA2	71	+++									2	23%
M13	20	41	61% exp	YWHAZ, YWHAB	77										3	30%
M14	19	41	78% db	ACTN2	58			+			+			+	1	5%
M15	19	50	58% db	PPP2CA, PPP2R1A	69					++				+	1	10%
M16	18	52	73% db	VAMP2	83				+++						0	-
M17	17	29	52% db	MAPK1	104				++				+		1	41%
M18	18	42	55% exp	NFKB1, RELA	103			+					+	+	2	50%
M19	17	64	77% coe	MRTO4	24	++					+				1	24%
M20	17	117	98% db		15			+	++						1	88%
M21	15	109	61% db		47				++			+			1	60%
M22	15	103	55% db		31						++			+	1	60%
M23	14	69	97% db		24						+	+			1	50%
M24	13	35	97% db		17					+		++			0	-
M25	13	27	56% exp	PRKCA, PRKCB	92	+									1	8%

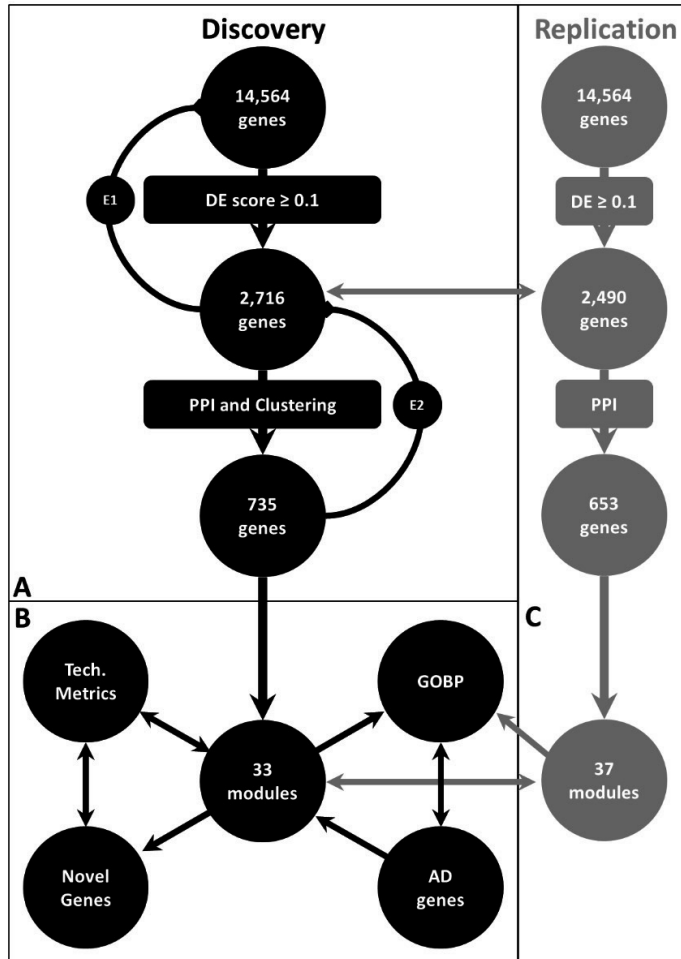
Module	Number of Genes	Interactions within Module	main interaction type	Hubs ≥ 30% interactions in module)	Interactions to other Modules	Organic substance metabolic process	Signal Transduction	Transport	Regulation of biological process	cellular metabolic process	cellular component organization	other metabolic processes	response to stimulus	other biological processes	Replication Modules	Highest Overlap
M26	12	57	51% exp		51	++						+			0	-
M27	12	21	67% db	CDK5	35									+	2	42%
M28	11	31	100% db	ENTPD3	9	+				++					1	45%
M29	11	37	78% db	TUBA4A	25										1	45%
M30	10	17	53% coe	GAPDH	36					+		++			2	40%
M31	10	12	58% db	FOS	19										1	10%
M32	10	62	45% db	ATP5A1, ATP5B, ATP5C1	12			++		+					0	-
M33	10	22	50% db	SMARCA4, SMARCC1	22	+					+				1	10%

**Table 2b.** Replication of Gene Ontology Biological Process Branches in AD expression literature

	Organic substance metabolic process	Signal Transduction	Transport	Regulation of biological process	cellular metabolic process	cellular component organization	other metabolic processes	response to stimulus	other biological processes
Ray et al, 2010		X	X	X	X	X		X	
Liang et al, 2008	X	X	X	X	X	X		X	X
Sekar et al, 2015	X	X		X			X	X	X
Twine et al, 2011	X		X	X	X	X	X	X	X
Chi et al, 2016		X						X	X
(IAD) Humphries et al, 2015	X	X					X	X	
Wang et al, 2016	X	X	X	X		X	X		X

### Functional enrichment of GO-Terms and KEGG pathways

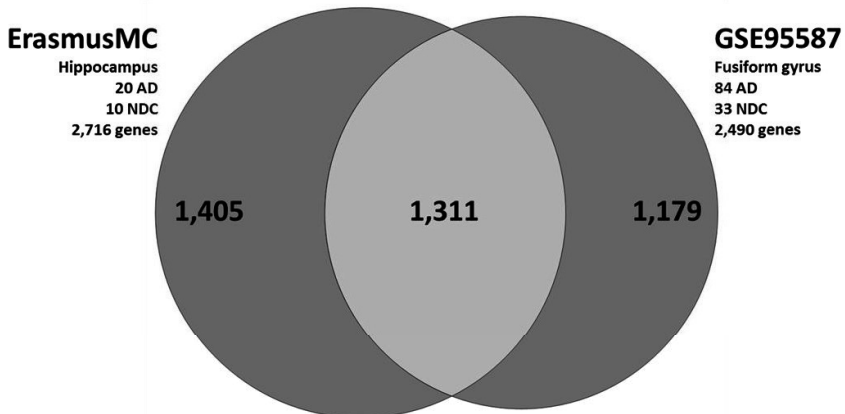
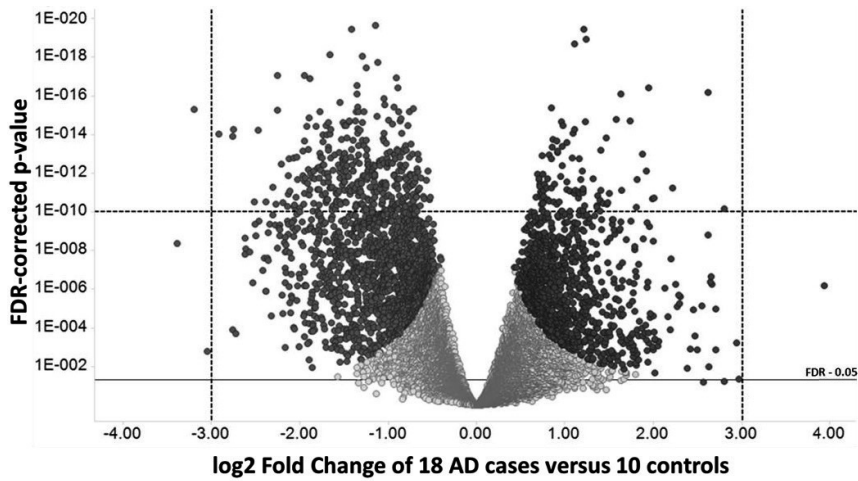
Functional enrichment analysis was split in three steps, as shown in figure 1. i) The 2,716 DE genes were compared to all 14,564 detected genes, to assess AD-specific enriched terms and pathways. The first of 143 significant GO Biological Process terms were: “neurotransmitter transport”, “modulation of synaptic transmission” and “regulation of transmembrane transport”. ii) The 735 modulated genes were compared to 2,716 DE genes, to assess further AD-specific enrichment or gene bias by the PPI analysis. The first of 66 significant biological process terms were “G-protein coupled receptor signaling pathways”, “circulatory system process” and “regulation of transmembrane transport”. iii) Enrichment of each individual module to all 735 modulated genes, to identify module-specific biological processes within the case-control landscape. This resulted in a range of 0 - 76 significant GO-terms per module. An overview of all enrichment results are displayed in supplementary table 2.



**Figure 1.** Flowchart of data analysis. A) extracting differentially expressed genes from all genes, followed by PPI analysis and module clustering. E1; gene set enrichment analysis of 2,716 DE genes versus all 14,564 detected genes (background). E2; gene set enrichment analysis of 735 modulated genes compared to all 2,716 DE genes. B) Gene modules are evaluated based on technical properties, interaction with known genes and functional annotation. Novel high-contributing genes are determined from these evaluations. Known AD genes are described in relation to these modules. E3; gene set enrichment analysis of each module compared to 735 modulated genes. C) Similar values are obtained for the replication cohort, resulting in 2,490 DE genes, of which 653 are clustered into 37 modules. Replication is based on overlapping DE genes and overlapping modules.

### Replication

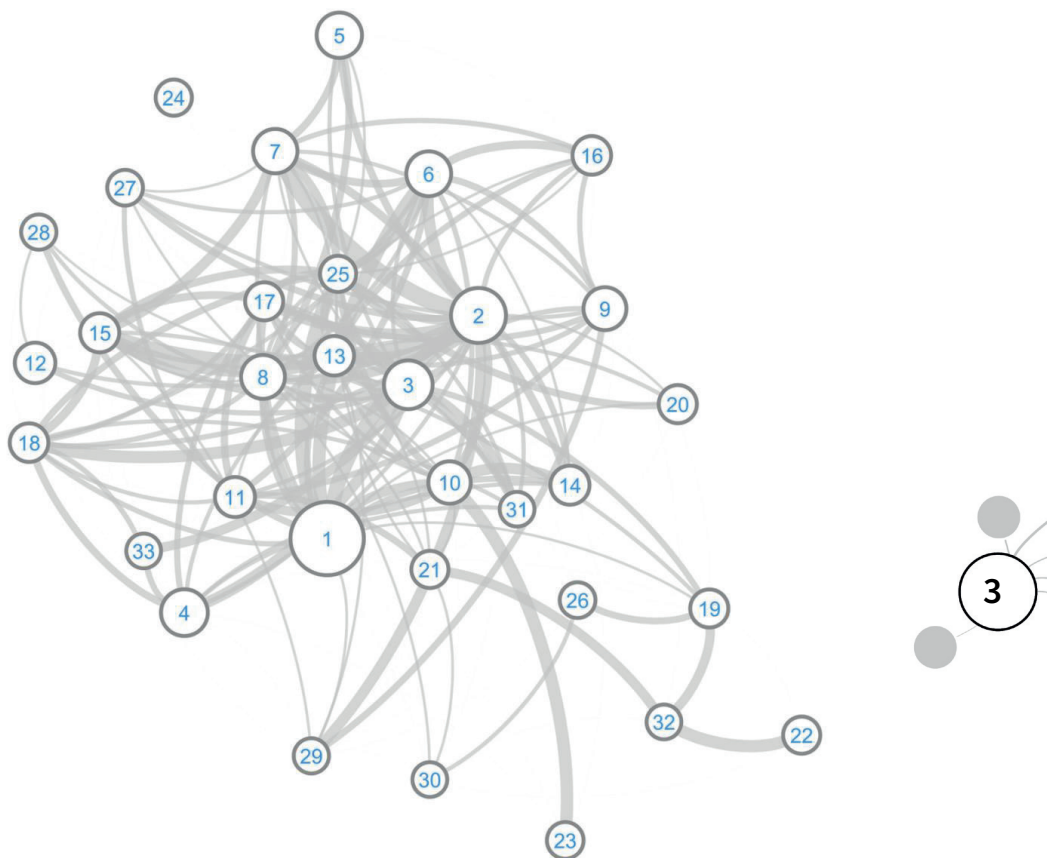
The replication dataset GSE53697 consisted of fusiform gyrus tissue from 84 AD cases and 33 controls. Of 2,716 DE genes in discovery, 2,098 were replicated (FDR < 0.05, 77%), containing 1,311 of the discovery DE genes (DE > 0.1, 48% replication rate), as shown in figure 2. GSE95587 all DE genes (2,490) were clustered into 37 modules, representing 653 genes. Of 97 significantly enriched GOBP terms in 2,716 discovery DE genes, 62 terms are replicated (FDR < 0.05, 64%) in the 2,490 replication DE genes.



**Figure 2.** Volcano plot of 14,564 analyzed protein-coding genes. Each dot is a gene, those dark-grey pass the 0.1 DE score threshold. Upper score limits (set to maximum of 1) are displayed by dotted lines. The solid line displays the default FDR corrected  $\geq 0.05$  threshold. The Venn diagram displays the number of overlapping DE genes between the discovery and replication cohorts.

### ***Annotation of gene modules***

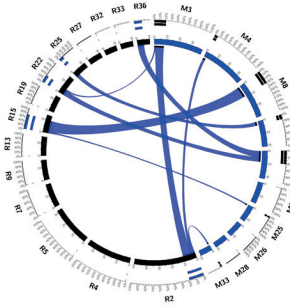
A total of 23 of the 33 discovery modules (70%) are predominantly (>50%) made up of database-based interactions, though 32 of 33 modules (97%) combine experimental with database-driven evidence (table 2). 10 modules (30%) consist mostly (>50% of interactions) of up-regulated genes, though 30 modules (91%) contain both up- and down-regulated genes. All separately clustered modules are displayed similarly in supplementary figure 2. The first three GOBP terms of all discovery and replication modules are mapped to the GO tree, as shown in supplementary figure 3.



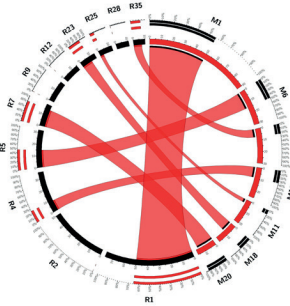
**Figure 3.** Protein-protein interaction network of 1,610 differentially expressed genes. Circles represent gene modules, connections represent interactions between modules. The width of the connections determine the amount of interactions.

The tree was divided in eight main branches; “Organic substance metabolic process”, “Signal Transduction”, “Transport”, “Regulation of biological process”, “Cellular metabolic process”, “Cellular component organization”, “Other metabolic processes” and “Response to stimulus”. The remaining terms are grouped under 9<sup>th</sup> branch; “Other biological processes”. Table 2 shows per module in which branch its three GOBP terms are located, the branches are detailed in supplementary figure 4. In total 84 GOBP terms of discovery are mapped to the GOBP tree, 38 of these replicated as a top three term for a replication module (45%). The replication modules are represented by 90 terms on the GOBP tree. Of the 735 genes in 33 discovery modules, 263 genes replicated into a replication module (36%), as shown in figure 4.

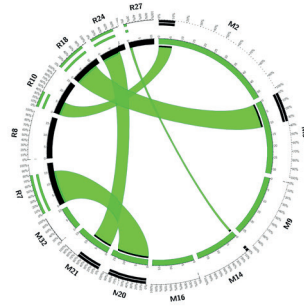
**Branch 1; Organic substance metabolic process**



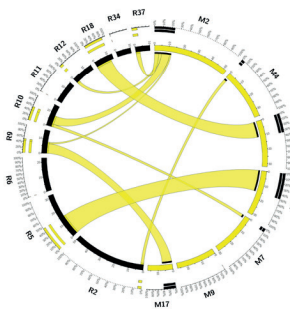
**Branch 2; Signal Transduction**



**Branch 3; Transport**



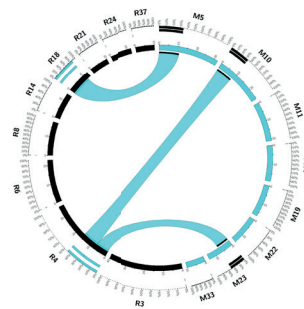
**Branch 4; Regulation of biological process**



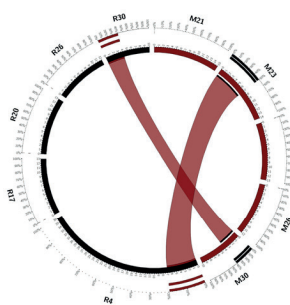
**Branch 5; Cellular metabolic process**



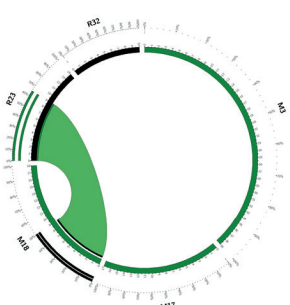
**Branch 6; Cellular component organization**



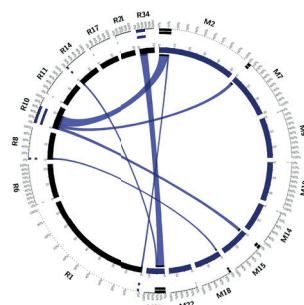
**Branch 7; Other metabolic processes**



**Branch 8; Response to stimulus**



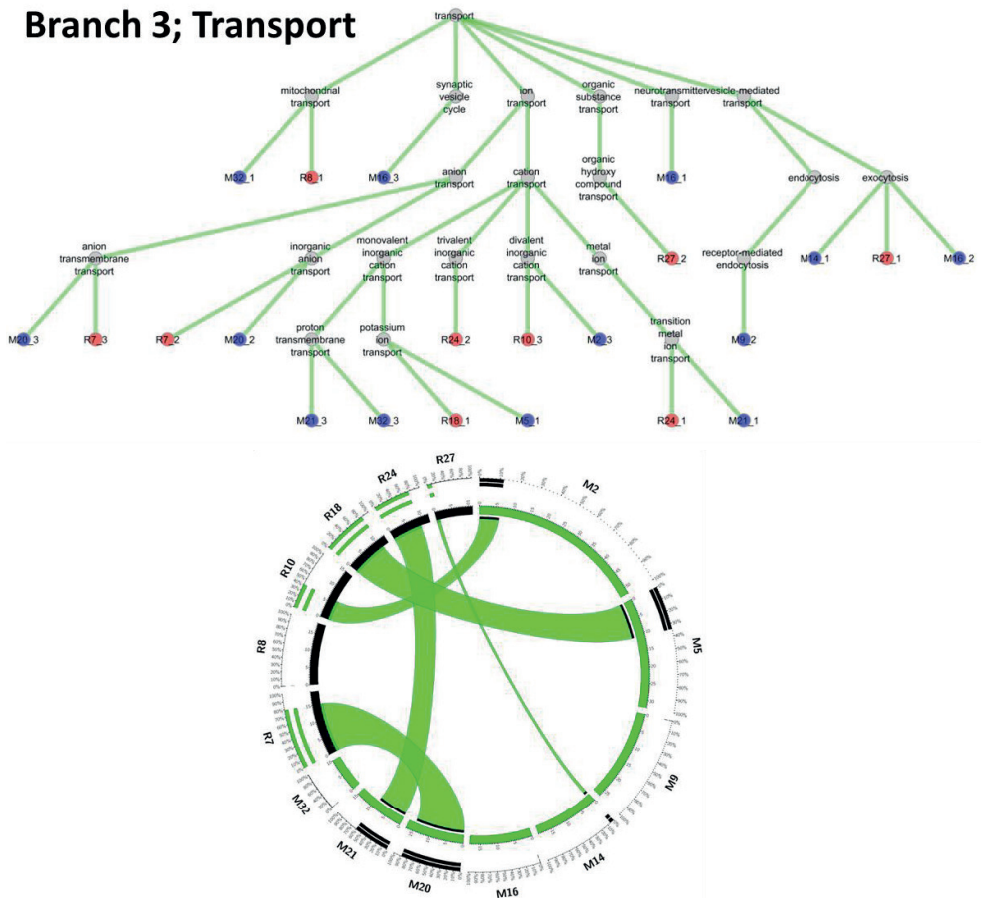
**Branch 9; Other biological processes**



**Figure 4.** Per branch, the overlap of genes between discovery and replication modules. Connections between modules represent overlapping genes, the width indicates the amount of overlap.

Figure 5 shows Branch 3; “Transport” in more detail. The tree shows the annotated discovery and replication modules. For example, the first term of module 5 (M5\_1) and replication module 18 (R18\_1) are both “potassium ion transport”. The plot shows that these modules have 11 genes in common, driving this enrichment in both. R18 has another two genes involved in potassium ion transport, which are not in any discovery module. Similarly, M5 holds another 19 genes involved in this term, not represented by any of the replication modules. This shows that although the term is represented by a module in both datasets, the underlying genes overlap only partially.

### Branch 3; Transport



**Figure 5.** Overview of Branch 4; Transport, and all discovery and replication modules annotated to this branch. Gray circles represent GOBP terms, blue circles are discovery modules, red circles are discovery modules. Connections are parent-children GOBP relationships or significant enrichments of modules to GOBP terms. The circular plot shows overlapping genes between gene modules in this branch, the width indicating the number of overlapping genes. The other branches are displayed in supplementary figure 4.

A second example, M14, M16 and R27 map to “exocytosis”. Only a single gene overlaps M14 and R27, while respectively 8, 14 and 4 genes in these modules are involved with “exocytosis”. Analyzing figure 5 further shows that some modules are nearly identical. I.e.; M20 and R7 or M21 and R24, while other modules are unique to either dataset, like M9, M16 and R8. Figures for the other 8 branches are displayed in supplementary figure 4. The module overlap plots are displayed per branch in figure 4.

### **Interaction with AD genes**

Of the 27 AD genes, 25 were expressed and analyzed, as shown in table 3. Three genes (11%) showed a DE-score of  $\geq 0.1$  and are included in the complete interaction network; *CD2AP* (score 0.18), *MEF2C* (-0.29) and *PTK2B* (-0.50), none of these were assigned to a module. In replication dataset GSE95587 *MEF2C* and *PTK2B* are replicated (DE score of -0.39 and -0.13, respectively). In total 233 interactions between DE-genes and known AD genes were present, 166 of these are database-driven (71%), 55 are experimental (24%), 12 were both database and experimental (5%) and none were based on co-expression. 16 of 33 modules held at least 2 database- or 2 experimental-driven interactions with an AD gene, in five of these both interaction types were present. These include *PTK2B* with module 7 (16 database and 2 experimental interactions) and module 10 (2 database and 3 experimental interactions), *MAPT* with module 2 (3 database and 4 experimental interactions), *BIN1* with module 9 (2 database and 5 experimental interactions) and *PICALM* with module 9 (2 database and 2 experimental interactions). The largest number of interactions were present between module 1 and APP (61 database interactions). Module 9 and module 15 were connected to most AD genes; *BIN1*, *HLA-DRB1*, *HLA-DRB5* and *PICALM* to module 9 and *BIN1*, *MEF2C*, *PICALM* and *MAPT* to module 13. Inversely, *APP* was connected to the most modules among the AD genes; 1, 11, 14, 17, 18 and 23. Nine genes did not interact with any module; *CASS4*, *CD33*, *CR1*, *FERMT2*, *MS4A6A*, *RIN3*, *SLC24A4*, *SORL1* and *ZCWPW1*.

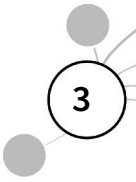
### **Identifying high-contributing genes**

To identify the most relevant genes in each module, we calculated an “contribution” score multiplying the number of interactions and DE score for each gene. This contribution ranged from 0 to 60, with an average of 2.9. Within the modules, the average score is 4.5, compared to 1.5 for all 875 unmodulated genes. 110 genes scored  $\geq 10.0$ . In table 4, the ten most contributing modulated and unmodulated gene are shown. Of note is that *PTK2B* is one of the known AD genes, ranking 48th according to the contribution score but not assigned to a single module in discovery. The ten highest-contributing unmodulated genes were all replicated and reach similar high rank in the replication network (range 13-181). Four of these ten genes were assigned to a module in the replication, as shown in table 4.



**Table3.** Overview of AD genes and their interactions with gene modules. For each detected gene the Fold Change and FDR corrected p-value is shown, with the corresponding Differential Expression (DE) score (fold change and p-value each scaled from 0 to 1 and then multiplied). All metrics are shown for the discovery and replication set (GSE95587). For each AD gene the main interacting discovery module is shown, with the number of interactions between that module and the AD gene. The last column indicates the branches in which that module is active.

Gene	Discovery				Replication				Main Module	Number of Interactions	Module Terms
	Fold Change	FDR p-value	DE score	Fold Change	FDR p-value	DE score					
PTK2B	-1.50	1.6E-13	-0.50*	-0.51	1.3E-04	-0.13*	M7	18	Signal Transduction, Response to stimulus, Other biological processes.		
MEF2C	-1.08	1.2E-08	-0.29*	-0.65	1.1E-09	-0.39*	-	-	-		
CD2AP	0.80	1.3E-07	0.18*	0.20	3.4E-02	0.02	-	-	-		
PSEN2	-0.58	1.3E-04	-0.08	-0.30	3.5E-05	-0.09	-	-	-		
MS4A6A	0.64	5.0E-04	0.07	0.45	8.0E-02	0.03	-	-	-		
INPP5D	0.55	2.5E-04	0.07	0.56	3.4E-05	0.17*	-	-	-		
ZCWPW1	0.41	3.3E-03	0.03	0.12	1.9E-01	0.01	-	-	-		
RIN3	0.48	8.9E-03	0.03	0.57	1.7E-05	0.18*	-	-	-		
CASS4	0.51	6.0E-02	0.02	0.33	1.2E-01	0.02	-	-	-		
MAPT	-0.33	2.2E-02	-0.02	-0.10	1.3E-01	-0.01	M2	7	Transport, Regulation of biological process, Other biological processes.		
APP	-0.26	9.2E-03	-0.02	-0.14	2.8E-02	-0.02	M1	61	Signal Transduction (3x).		
APOE	-0.38	5.8E-02	-0.02	-0.07	6.6E-01	0.00	-	-	-		
CR1	0.53	1.5E-01	0.01	0.93	8.1E-05	0.25*	-	-	-		
ABCA7	0.36	6.7E-02	0.01	0.18	1.8E-01	0.01	M2	3	Transport, Regulation of biological process, Other biological processes.		
PICALM	0.18	1.3E-01	0.01	0.21	2.3E-03	0.04	M9	4	Transport, Regulation of biological process, Other biological processes.		
SLC24A4	0.23	3.6E-01	0.00	-0.15	1.9E-01	-0.01	-	-	-		
CD33	0.23	3.8E-01	0.00	0.38	3.8E-02	0.04	-	-	-		

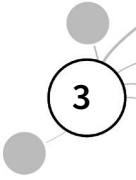


Gene	Discovery				Replication				Main Module	Number of Interactions	Module Terms
	Fold Change	FDR p-value	DE score	Fold Change	FDR p-value	DE score					
HLA-DRB1	0.15	6.1E-01	0.00	0.34	5.4E-02	0.03	M9	7	Transport, Regulation of biological process, Other biological processes.		
PSEN1	0.10	5.3E-01	0.00	0.24	2.6E-03	0.04	-	-	-		
HLA-DRB5	-0.16	7.2E-01	0.00	0.08	8.4E-01	0.00	M9	7	Transport, Regulation of biological process, Other biological processes.		
FERMT2	0.07	7.3E-01	0.00	0.06	5.8E-01	0.00	-	-	-		
CELF1	-0.05	6.6E-01	0.00	-0.02	8.5E-01	0.00	M8	3	Organic substance metabolic process, cellular metabolic process (2x).		
BIN1	0.04	8.1E-01	0.00	0.23	1.7E-02	0.03	M9	7	Transport, Regulation of biological process, Other biological processes.		
SORL1	0.03	8.9E-01	0.00	-0.01	8.8E-01	0.00	-	-	-		
CLU	0.01	9.8E-01	0.00	0.00	1.0E+00	0.00	M14	7	Transport, Cellular component organization, Other biological processes.		
EPHA1	ND	-	-	-	-	-	-	-	-		
NME8	ND	-	-	-	-	-	-	-	-		

**Table4.** Overview of 10 modulated and 10 unmodulated genes with the highest predicted contribution on the AD network, calculated by multiplying DE score and absolute number of interactions in the network. The number of interactions, DE score and subsequent contribution score are shown for discovery and replication analysis. Contribution rank indicates the ranks of contribution scores (separately for discovery and replication). Finally the assigned module, if any, is reported, and for discovery modules the branches in which it is active.

set	Gene	Discovery					Replication					
		interactions	DE score	Contribution Score	Contribution Rank	Module	interactions	DE score	Contribution Score	Contribution Rank	Module	Discovery Module Terms
	GNG2	108	0.56	60.2	1	M1	91	0.22	20.4	15	R1	Signal Transduction (3x).
	SST	41	0.97	39.9	2	M1	32	0.32	10.1	49	R1	Signal Transduction (3x).
	PRKACA	142	0.22	31.6	3	M2	-	0.09	-	-	-	Signal Transduction, Regulation of biological process, Other biological processes.
	MAPK1	120	0.25	29.7	4	M17	133	0.18	23.8	6	R9	Regulation of biological process (2x), Response to stimulus.
	OPRK1	40	0.73	29.2	5	M1	-	0.02	-	-	-	Signal Transduction (3x).
	CNR1	43	0.65	28.2	6	M1	35	0.19	6.6	115	R1	Signal Transduction (3x).
	CXCR4	57	0.49	27.7	7	M1	43	0.91	39.2	1	R1	Signal Transduction (3x).
	PNOC	38	0.68	25.9	9	M1	31	0.38	11.8	36	R1	Signal Transduction (3x).
	FOS	41	0.61	24.9	10	M31	50	0.22	10.9	45	-	-
	RGS4	37	0.67	24.7	11	M1	31	0.73	22.7	8	R1	Signal Transduction (3x).

Genes in a Module



set	Gene	Discovery					Replication					
		interactions	DE score	Contribution Score	Contribution Rank	Module	interactions	DE score	Contribution Score	Contribution Rank	Module	Discovery Module Terms
	PLCB1	60	0.44	26.4	8	-	52	0.32	16.4	23	R1	-
	PTK2B	48	0.50	24.0	13	-	41	0.13	5.5	151	R35	-
	YWHAH	46	0.45	20.9	19	-	46	0.30	13.9	27	-	-
	PAK1	36	0.56	20.1	21	-	40	0.53	21.1	13	-	-
	DNM1	28	0.64	17.8	34	-	33	0.15	4.9	181	-	-
	PRKE	30	0.51	15.2	49	-	32	0.29	9.2	58	R29	-
	NSF	26	0.58	15.1	50	-	14	0.38	5.3	155	-	-
	SYT1	24	0.63	15.0	51	-	16	0.31	5.0	174	-	-
	MAP2K1	41	0.36	14.8	52	-	39	0.17	6.7	109	R9	-
	SNCA	39	0.34	13.4	69	-	41	0.28	11.7	37	-	-

Genes Not in a Module

## Discussion

Our study identified 33 DE-modules representing 735 differentially expressed genes in AD post-mortem hippocampus. We here report a first list of relevant pathways and high-contributing genes, as well as details on how such a network can be constructed and interpreted. These modules represent specific biologically relevant components of 9 main GOBP branches involved in AD pathophysiology. These modules and branches are investigated to provide insight and structure to AD biology. Our results also suggest that individual genes or modules should be considered in the larger network, as many interact with one another. Finally, we show that 48% of DE genes replicate, compared to 64% of enriched biological processes, suggesting that replication based on biological processes is more robust than individual genes.

### **AD transcriptional changes and replication**

We observed 2,716 DE genes in hippocampus AD cases versus controls. Gene set enrichment analyses revealed 97 significantly enriched biological processes, the main one being “neurotransmitter transport”. Our replication dataset GSE95587 held 84 cases and 33 controls of the fusiform gyrus, 2,490 genes showed differential expression, including 1,311 (48%) of the DE genes from discovery <sup>(Brettschneider, et al., 2015, Chang, et al., 2016)</sup>. Of 97 biological processes, 62 replicated, showing high similarity between both studies, despite being separate tissues. The higher replication rate of enriched biological processes suggests that different genes can reach DE in different studies, but represent the same biological process. 36% of genes assigned to a module in discovery also gain a module in replication, compared to 45% of gene module enriched biological processes. This further suggests more robust replication in biological processes compared to individual genes. However, the replication rate for both genes and biological processes has decreased, likely as not all genes are included in the PPI database (41% of 2,716 DE genes did not have any interaction) or lost in the clustering method.

### **Gene modules and Gene Ontology Branches**

Enrichment of the discovery and replication gene modules provided a GOBP tree with differentially expressed terms. These could be assigned to 8 main branches and one “other” branch, as shown in supplementary figure 3. These branches are often observed in literature of AD expression studies, as shown in table 2 <sup>(Chi, et al., 2016, C.E. Humphries, et al., 2015, Liang, et al., 2008, Ray and Zhang, 2010, Sekar, et al., 2015, Twine, et al., 2011, Wang, et al., 2016)</sup>. In the same table we can see that most modules partake in multiple branches, suggesting that these branches are not independent. Similarly, some modules are retained from discovery to replication as nearly the same set of genes (ie; M1, M20 or M21). Most modules however, share only a part of their genes with one or two replication modules, as described for potassium ion transport modules M5 and R18, or exocytosis modules M14, M16 and R27 <sup>(Musunuri, et al., 2016, Vitvitsky, et al., 2012)</sup>. This suggests that for most biological processes, a small set of core genes will replicate while others vary between studies. However, enrichment of the main biological process is maintained, and a module can be assigned to represent this process in each study. This shows that overlap on enriched biological processes or terms, or enrichment on firstly obtained gene modules, is more robust than directly comparing genes.

Finally, a small number of modules is not represented in the replication cohort. These are M9 (first term; “receptor-mediated endocytosis”), M16 (“exocytosis”), M24 (“glycerolipid metabolic process”) and M32 (“mitochondrial transport”). The terms for M16 and M32 are detected in replication modules (but by other genes). It might possibly be that the brain region (hippocampus vs fusiform gyrus) of these specific patients differs on endo/exocytosis and lipid metabolism, or else these are random variations between studies (Di Paolo and Kim, 2011, Kelly and Ferreira, 2007, Musunuri, et al., 2016).

As shown in table 2, enrichment in most of these branches is reported in other AD transcriptomic studies. Not always the exact same terms are observed in each study, possibly as a result of some technical variation per dataset, methods used (ie; micro-arrays or RNA-Seq) or biological differences between cases or brain region studied. As these branches and terms are detected in most studies, it suggests these belong to a general regulatory network of neurodegeneration, relatively robust across samples and brain regions.

### **Summary of AD-involved GOBP Branches**

Branch 1; Organic substance metabolic process. This branch represents 10 modules and 13 replication modules. Most terms revolve around “DNA metabolic process”, including DNA repair and metabolism. RNA transcription, translation and post-translational modification processes are also included in this branch. Most modules overlap only partially, suggesting a large variation or turnover of genes in this branch (Bucholtz and Demuth, 2013, Lillenes, et al., 2016).

Branch 2; Signal Transduction. This branch contains 7 modules and 11 replication modules. Some of the largest modules are in this branch, including modules 1 of both datasets (about 50% overlap), both representing g-protein coupled receptors. Most terms in this branch represent transmembrane neurotransmitter genes. Overlap between discovery and replication modules is high, suggesting these genes work in larger complexes and therefore often co-express together. The clear distinction in various subsets of protein-coupled receptors also indicates clear and complete annotations for these genes, with well-defined functional classes (Kandimalla and Reddy, 2017, Rajmohan and Reddy, 2017).

Branch 3; Transport. As described, the transport branch splits into ion transport (further to cation and anion transport) and a subset of vesicle mediated transport (endo and exocytosis) which is more dominant in the discovery dataset. Transport contains 8 discovery and replication modules (Kelly and Ferreira, 2007, Musunuri, et al., 2016).

Branch 4; Regulation of biological processes. Many of the modules in this branch come from second or third enriched terms. There is large overlap between of this branch to other branches. For example, modules M5 and R18 were involved in potassium ion transport in branch 3, and are here both in “regulation of transmembrane transport”, both enrichments are driven by largely the same genes. This sporadic overlap is also represented in any, small overlaps between the 7 discovery and 10 replication modules. Finally, all modules in this branch also have annotations in other branches, suggesting this branch is largely complimentary to the others.

Branch 5; Cellular metabolic process. This branch contains 6 discovery and 7 replication modules. The only overlap is between M28 and R26 and M30 and R30, although these modules are not enriched for the same terms. Many of the terms involve nucleoside phosphate metabolic processes or RNA processing and suggest a role similar to branch1; organic substance metabolic process (Ansoleaga, et al., 2015).

Branch 6; Cellular component organization. This branch contains several enrichment on cellular structure and function. For example mitochondrion function or axon development, but also microtubule and actin organization (Cabral Fontela, et al., 2017; Yan, et al., 2013). We again observe M5 and R18 sharing a small overlap on term “protein complex oligomerization”. Furthermore M10 and M23 seem to combine into R4, involved in extracellular structure organization. This branch holds 8 discovery and 9 replication modules with few overlaps.

Branch 7; Other metabolic processes. A small branch with only 5 discovery and 5 replication modules. It contains terms like lipid metabolism, methylation and nitrogen compound metabolism. Only two overlaps between M23 and R4 and between M30 and R30, both of which have been seen on the other branches (Di Paolo and Kim, 2011; Liu and Zhang, 2014).

Branch 8; Response to stimulus. Although often observed in AD literature, this branch only contains M3 and M17 with no overlap, M18 with large overlap to R23, and R32 (again no overlap). These modules represent various subsets of response to stress or other stimuli. M3 represents part of the inflammation/immune response often reported in AD (Rozpedek, et al., 2015).

Branch 9; other biological processes. This last branch contains 9 discovery and 9 replication modules and holds all remaining GOBP enrichments. The terms range across various processes and provide small overlaps between modules. Some terms overlap with other branches, like microtubule-based movement or cytokine production.

### **Interactions with AD genes**

Of 27 AD genetic risk factor genes, only three were differentially expressed in our dataset, of which two replicated (*MEF2C* and *PTK2B*). Lack of association for the other genes, suggests their roles might be earlier in the disease process, where the genetic mutations might influence onset of the disease. Nevertheless, several modules hold interactions with these AD genes, suggesting overlap in biological function. 10 AD genes interacted at least twice with a gene module, as shown in table 3. *HLA-DRB1*, *HLA-DRB5*, *BIN1* and *PICALM* interact with M9 and are involved in endocytosis (*HLA-DRB1*, *HLA-DRB5*, *BIN1*, *PICALM*) and microtubule-based movement (*BIN1*) (Baig, et al., 2010; Zhou, et al., 2014). *ABCA7* and *MAPT* interact with M2, involved in ion transport and signaling. *APP* interacts with M1, both are involved in signal transduction (Cheng, et al., 2014; Cirrito, et al., 2008). *PTK2B* is differentially expressed in both discovery and replication (DE -0.50 and -0.13, respectively) and interacts with M7, M10 and M25 (Beecham, et al., 2014; Han, et al., 2017). M7 and M10 are involved in cell surface receptor signaling and M25 in protein modification, *PTK2B* is also associated to those biological processes. *CELF1* interacts with M8 into RNA processing and protein modification and *CLU* interacts with M14 into exocytosis and actin-based filament organization. These interactions suggest a role for

some of these genes in the later stages of AD, and do not represent the typical associations of these genes in a causal inference (Lambert, et al., 2013, Van Cauwenbergh, et al., 2016).

### **Predicting high-contributing genes**

An overall contribution-score was calculated per gene to identify those with large contribution to the AD network; genes with many interactions and large differential expression in AD, see table 4. These genes have potential to mediate large portions of AD biology, and disruptions in their functions might lead to more widespread consequences than in other genes. They are relevant candidates for Mendelian or GWAS genetic studies, either by directly playing on the biological mechanisms, or by mediating defects in other genes, for example AD risk loci. Several high-contributing genes are not assigned to a module, suggesting that genes with ties to multiple gene groups (i.e.; involved in various different processes with distinct gene groups) cannot clearly be assigned to a single module by the clustering algorithm, and end up unassigned as a result. An example of this is the known AD risk gene *PKT2B*, which is differentially expressed and interacts with 48 other DE genes, but cannot be assigned to a single module. Therefore, examining contribution and connectivity separately from the module assignment is relevant and adds importance to the overall interpretation of the expression data.

### **Limitations on PPI + MCL method**

Although extracting gene modules of specific GO Biological Processes provides extra information to traditional enrichment analysis, a number of factors can be improved to make this method more efficient. Firstly, the PPI networks are comprised of existing databases (von Mering, et al., 2003). The use of external data in a Bayesian-kind of approach is useful, but generates bias to well-known genes and biological processes (Gillis, et al., 2014, Schaefer, et al., 2015). At the same time, genes without a currently known function tend to be downplayed. For example, genes unknown in STRING or GOBP will be largely ignored in this analysis. Completing these databases, and adding additional unbiased datasets to use as reference will allow for more complete networks and modules, as well as more robust ones. Furthermore, there are no clear guidelines on protein interaction cutoffs parameters, MCL clustering thresholds or consensus on functionally annotating a gene module. As network properties will change between studies (i.e.; network density and size), it will be challenging to determine a golden standard. Nevertheless some consensus is emerging; 1. prioritizing or limiting to experimental interactions types, or not using text-mining based types (Szkarczyk, et al., 2017, von Mering, et al., 2003); 2. altering the MCL inflation factor to generated modules of 10-100 genes and not much smaller or larger (Subramanian, et al., 2005, van Dongen and Abreu-Goodger, 2012). 3. replicating in additional studies, preferably on a functional annotation level as Gene Ontology (Ashburner, et al., 2000, Gene Ontology, 2015). With improving interaction databases, the quality and type of an interaction can be used as weights in the clustering analysis. In future studies also the DE score and direction of effect could be taken into account during clustering, which wasn't done here as effect estimates can vary widely between individual studies.



### ***Limitations and strengths of this study***

This study was designed as a cross-sectional case-control analysis, making it hard to distinguish between AD-specific causal expression differences versus changes caused by neurodegeneration. Due to the late stage of disease in these samples, many of the observed differences are likely caused by neurodegeneration. Despite correction by PCs, some differences between cases and controls are likely caused by different cell-type compositions; fewer neurons in cases compared to controls as an effect of AD. Our sample size of 20 cases and 10 controls is not optimal to robustly detect all deviations in AD, having more cases than controls also leads to more down- than up-regulated genes, biasing the results towards downregulated pathways.

Our method provides an overview of dysregulated pathways in AD, while maintaining resolution to investigate individual gene contributions to specific pathways and the whole network in general. We show that the PPI and MCL clustering approach identifies clearly defined functional gene modules, which can be combined into a whole AD transcriptomic overview. With more and better input datasets, like large-scale gene co-expression data and unbiased gene annotation databases, as well as other brain regions and earlier disease stages in AD, this method can aid in constructing a complete transcriptomic AD network and in interpreting the impact of a single gene's loss of function by mutation or other relevant risk factors.

## References

- Ansoleaga, B., Jove, M., Schluter, A., Garcia-Esparcia, P., Moreno, J., Pujol, A., Pamplona, R., Portero-Otin, M., Ferrer, I. 2015. Deregulation of purine metabolism in Alzheimer's disease. *Neurobiol Aging* 36(1), 68-80. doi:S0197-4580(14)00517-X [pii]  
10.1016/j.neurobiolaging.2014.08.004.
- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25(1), 25-9. doi:10.1038/75556.
- Baig, S., Joseph, S.A., Tayler, H., Abraham, R., Owen, M.J., Williams, J., Kehoe, P.G., Love, S. 2010. Distribution and expression of picalm in Alzheimer disease. *J Neuropathol Exp Neurol* 69(10), 1071-7. doi:10.1097/NEN.0b013e3181f52e01.
- Beecham, G.W., Hamilton, K., Naj, A.C., Martin, E.R., Huentelman, M., Myers, A.J., Corneveaux, J.J., Hardy, J., Vonsattel, J.P., Younkin, S.G., Bennett, D.A., De Jager, P.L., Larson, E.B., Crane, P.K., Kamboh, M.I., Kofler, J.K., Mash, D.C., Duque, L., Gilbert, J.R., Gwirtsman, H., Buxbaum, J.D., Kramer, P., Dickson, D.W., Farrer, L.A., Frosch, M.P., Ghetti, B., Haines, J.L., Hyman, B.T., Kukull, W.A., Mayeux, R.P., Pericak-Vance, M.A., Schneider, J.A., Trojanowski, J.Q., Reiman, E.M., Alzheimer's Disease Genetics, C., Schellenberg, G.D., Montine, T.J. 2014. Genome-wide association meta-analysis of neuropathologic features of Alzheimer's disease and related dementias. *PLoS Genet* 10(9), e1004606. doi:10.1371/journal.pgen.1004606  
PGENETICS-D-14-00298 [pii].
- Bekris, L.M., Yu, C.E., Bird, T.D., Tsuang, D.W. 2010. Genetics of Alzheimer disease. *J Geriatr Psychiatry Neurol* 23(4), 213-27. doi:23/4/213 [pii]  
10.1177/0891988710383571.
- Bolger, A.M., Lohse, M., Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15), 2114-20. doi:btu170 [pii]  
10.1093/bioinformatics/btu170.
- Boyle, P.A., Wilson, R.S., Yu, L., Barr, A.M., Honer, W.G., Schneider, J.A., Bennett, D.A. 2013. Much of late life cognitive decline is not due to common neurodegenerative pathologies. *Ann Neurol* 74(3), 478-89. doi:10.1002/ana.23964.
- Braak, H., Braak, E. 1995. Staging of Alzheimer's disease-related neurofibrillary changes. *Neurobiol Aging* 16(3), 271-8; discussion 8-84. doi:0197458095000216 [pii].
- Brettschneider, J., Del Tredici, K., Lee, V.M., Trojanowski, J.Q. 2015. Spreading of pathology in neurodegenerative diseases: a focus on human studies. *Nat Rev Neurosci* 16(2), 109-20. doi:nrn3887 [pii]  
10.1038/nrn3887.
- Bucholtz, N., Demuth, I. 2013. DNA-repair in mild cognitive impairment and Alzheimer's disease. *DNA Repair (Amst)* 12(10), 811-6. doi:S1568-7864(13)00167-5 [pii]  
10.1016/j.dnarep.2013.07.005.
- Cabrales Fontela, Y., Kadavath, H., Biernat, J., Riedel, D., Mandelkow, E., Zweckstetter, M. 2017. Multivalent cross-linking of actin filaments and microtubules through the microtubule-associated protein Tau. *Nat Commun* 8(1), 1981. doi:10.1038/s41467-017-02230-8  
10.1038/s41467-017-02230-8 [pii].

- Carbon, S., Ireland, A., Mungall, C.J., Shu, S., Marshall, B., Lewis, S., Ami, G.O.H., Web Presence Working, G. 2009. AmiGO: online access to ontology and annotation data. *Bioinformatics* 25(2), 288-9. doi:btn615 [pii]  
10.1093/bioinformatics/btn615.
- Chang, Y.T., Huang, C.W., Chen, N.C., Lin, K.J., Huang, S.H., Chang, W.N., Hsu, S.W., Hsu, C.W., Chen, H.H., Chang, C.C. 2016. Hippocampal Amyloid Burden with Downstream Fusiform Gyrus Atrophy Correlate with Face Matching Task Scores in Early Stage Alzheimer's Disease. *Front Aging Neurosci* 8, 145. doi:10.3389/fnagi.2016.00145.
- Chen, C.H., Zhou, W., Liu, S., Deng, Y., Cai, F., Tone, M., Tone, Y., Tong, Y., Song, W. 2012. Increased NF-kappaB signalling up-regulates BACE1 expression and its therapeutic potential in Alzheimer's disease. *Int J Neuropsychopharmacol* 15(1), 77-90. doi:S1461145711000149 [pii]  
10.1017/S1461145711000149.
- Cheng, X., Wu, J., Geng, M., Xiong, J. 2014. Role of synaptic activity in the regulation of amyloid beta levels in Alzheimer's disease. *Neurobiol Aging* 35(6), 1217-32. doi:S0197-4580(13)00605-2 [pii]  
10.1016/j.neurobiolaging.2013.11.021.
- Chi, L.M., Wang, X., Nan, G.X. 2016. In silico analyses for molecular genetic mechanism and candidate genes in patients with Alzheimer's disease. *Acta Neurol Belg* 116(4), 543-7. doi:10.1007/s13760-016-0613-6  
10.1007/s13760-016-0613-6 [pii].
- Cirrito, J.R., Kang, J.E., Lee, J., Stewart, F.R., Verges, D.K., Silverio, L.M., Bu, G., Mennerick, S., Holtzman, D.M. 2008. Endocytosis is required for synaptic activity-dependent release of amyloid-beta in vivo. *Neuron* 58(1), 42-51. doi:S0896-6273(08)00124-4 [pii]  
10.1016/j.neuron.2008.02.003.
- Di Paolo, G., Kim, T.W. 2011. Linking lipids to Alzheimer's disease: cholesterol and beyond. *Nat Rev Neurosci* 12(5), 284-96. doi:nrn3012 [pii]  
10.1038/nrn3012.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1), 15-21. doi:bts635 [pii]  
10.1093/bioinformatics/bts635.
- Enright, A.J., Van Dongen, S., Ouzounis, C.A. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30(7), 1575-84.
- Foote, M., Zhou, Y. 2012. 14-3-3 proteins in neurological disorders. *Int J Biochem Mol Biol* 3(2), 152-64.
- Gene Ontology, C. 2001. Creating the gene ontology resource: design and implementation. *Genome Res* 11(8), 1425-33. doi:10.1101/gr.180801.
- Gene Ontology, C. 2015. Gene Ontology Consortium: going forward. *Nucleic Acids Res* 43(Database issue), D1049-56. doi:gku1179 [pii]  
10.1093/nar/gku1179.
- Gillis, J., Ballouz, S., Pavlidis, P. 2014. Bias tradeoffs in the creation and analysis of protein-protein interaction networks. *J Proteomics* 100, 44-54. doi:S1874-3919(14)00038-4 [pii]  
10.1016/j.jpro.2014.01.020.
- Han, Z., Huang, H., Gao, Y., Huang, Q. 2017. Functional annotation of Alzheimer's disease associated loci revealed by GWASs. *PLoS One* 12(6), e0179677. doi:10.1371/journal.pone.0179677  
PONE-D-16-48239 [pii].

- Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J.M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigo, R., Hubbard, T.J. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* 22(9), 1760-74. doi:22/9/1760 [pii] 10.1101/gr.135350.111.
- Ho Kim, J., Franck, J., Kang, T., Heinsen, H., Ravid, R., Ferrer, I., Hee Cheon, M., Lee, J.Y., Shin Yoo, J., Steinbusch, H.W., Salzet, M., Fournier, I., Mok Park, Y. 2015. Proteome-wide characterization of signalling interactions in the hippocampal CA4/DG subfield of patients with Alzheimer's disease. *Sci Rep* 5, 11138. doi:srep11138 [pii] 10.1038/srep11138.
- Holtzman, D.M., Carrillo, M.C., Hendrix, J.A., Bain, L.J., Catafau, A.M., Gault, L.M., Goedert, M., Mandelkow, E., Mandelkow, E.M., Miller, D.S., Ostrowitzki, S., Polydoro, M., Smith, S., Wittmann, M., Hutton, M. 2016. Tau: From research to clinical development. *Alzheimers Dement* 12(10), 1033-9. doi:S1552-5260(16)30019-X [pii] 10.1016/j.jalz.2016.03.018.
- Humphries, C., Kohli, M.A., Whitehead, P., Mash, D.C., Pericak-Vance, M.A., Gilbert, J. 2015. Alzheimer disease (AD) specific transcription, DNA methylation and splicing in twenty AD associated loci. *Mol Cell Neurosci* 67, 37-45. doi:S1044-7431(15)00084-6 [pii] 10.1016/j.mcn.2015.05.003.
- Humphries, C.E., Kohli, M.A., Nathanson, L., Whitehead, P., Beecham, G., Martin, E., Mash, D.C., Pericak-Vance, M.A., Gilbert, J. 2015. Integrated whole transcriptome and DNA methylation analysis identifies gene networks specific to late-onset Alzheimer's disease. *J Alzheimers Dis* 44(3), 977-87. doi:P1144083608527M4 [pii] 10.3233/JAD-141989.
- Jellinger, K.A. 2008. Neuropathological aspects of Alzheimer disease, Parkinson disease and frontotemporal dementia. *Neurodegener Dis* 5(3-4), 118-21. doi:000113679 [pii] 10.1159/000113679.
- Kandimalla, R., Reddy, P.H. 2017. Therapeutics of Neurotransmitters in Alzheimer's Disease. *J Alzheimers Dis* 57(4), 1049-69. doi:JAD161118 [pii] 10.3233/JAD-161118.
- Karran, E., Mercken, M., De Strooper, B. 2011. The amyloid cascade hypothesis for Alzheimer's disease: an appraisal for the development of therapeutics. *Nat Rev Drug Discov* 10(9), 698-712. doi:nrd3505 [pii] 10.1038/nrd3505.
- Kavanagh, T., Mills, J.D., Kim, W.S., Halliday, G.M., Janitz, M. 2013. Pathway analysis of the human brain transcriptome in disease. *J Mol Neurosci* 51(1), 28-36. doi:10.1007/s12031-012-9940-0.
- Kelly, B.L., Ferreira, A. 2007. Beta-amyloid disrupted synaptic vesicle endocytosis in cultured hippocampal neurons. *Neuroscience* 147(1), 60-70. doi:S0306-4522(07)00310-7 [pii] 10.1016/j.neuroscience.2007.03.047.
- Kong, W., Mou, X., Zhang, N., Zeng, W., Li, S., Yang, Y. 2015. The construction of common and specific significance subnetworks of Alzheimer's disease from multiple brain regions. *Biomed Res Int* 2015, 394260. doi:10.1155/2015/394260.

- Kong, W., Zhang, J., Mou, X., Yang, Y. 2014. Integrating gene expression and protein interaction data for signaling pathway prediction of Alzheimer's disease. *Comput Math Methods Med* 2014, 340758. doi:10.1155/2014/340758.
- Kravitz, E., Gaisler-Salomon, I., Biegan, A. 2013. Hippocampal glutamate NMDA receptor loss tracks progression in Alzheimer's disease: quantitative autoradiography in postmortem human brain. *PLoS One* 8(11), e81244. doi:10.1371/journal.pone.0081244  
PONE-D-13-34784 [pii].
- Krzyzanowska, A., Garcia-Consuegra, I., Pascual, C., Antequera, D., Ferrer, I., Carro, E. 2015. Expression of regulatory proteins in choroid plexus changes in early stages of Alzheimer disease. *J Neuropathol Exp Neurol* 74(4), 359-69. doi:10.1097/NEN.0000000000000181.
- Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., DeStafano, A.L., Bis, J.C., Beecham, G.W., Grenier-Boley, B., Russo, G., Thorton-Wells, T.A., Jones, N., Smith, A.V., Chouraki, V., Thomas, C., Ikram, M.A., Zelenika, D., Vardarajan, B.N., Kamatani, Y., Lin, C.F., Gerrish, A., Schmidt, H., Kunkle, B., Dunstan, M.L., Ruiz, A., Bihoreau, M.T., Choi, S.H., Reitz, C., Pasquier, F., Cruchaga, C., Craig, D., Amin, N., Berr, C., Lopez, O.L., De Jager, P.L., Deramecourt, V., Johnston, J.A., Evans, D., Lovestone, S., Letenneur, L., Moron, F.J., Rubinsztein, D.C., Eiriksdottir, G., Sleegers, K., Goate, A.M., Fievet, N., Huentelman, M.W., Gill, M., Brown, K., Kamboh, M.I., Keller, L., Barberger-Gateau, P., McGuinness, B., Larson, E.B., Green, R., Myers, A.J., Dufouil, C., Todd, S., Wallon, D., Love, S., Rogaeva, E., Gallacher, J., St George-Hyslop, P., Clarimon, J., Lleo, A., Bayer, A., Tsuang, D.W., Yu, L., Tsolaki, M., Bossu, P., Spalletta, G., Proitsi, P., Collinge, J., Sorbi, S., Sanchez-Garcia, F., Fox, N.C., Hardy, J., Deniz Naranjo, M.C., Bosco, P., Clarke, R., Brayne, C., Galimberti, D., Mancuso, M., Matthews, F., European Alzheimer's Disease, I., Genetic, Environmental Risk in Alzheimer's, D., Alzheimer's Disease Genetic, C., Cohorts for, H., Aging Research in Genomic, E., Moebus, S., Mecocci, P., Del Zompo, M., Maier, W., Hampel, H., Pilotto, A., Bullido, M., Panza, F., Caffarra, P., Nacmias, B., Gilbert, J.R., Mayhaus, M., Lannefelt, L., Hakonarson, H., Pichler, S., Carrasquillo, M.M., Ingelsson, M., Beekly, D., Alvarez, V., Zou, F., Valladares, O., Younkin, S.G., Coto, E., Hamilton-Nelson, K.L., Gu, W., Razquin, C., Pastor, P., Mateo, I., Owen, M.J., Faber, K.M., Jonsson, P.V., Combarros, O., O'Donovan, M.C., Cantwell, L.B., Soininen, H., Blacker, D., Mead, S., Mosley, T.H., Jr., Bennett, D.A., Harris, T.B., Fratiglioni, L., Holmes, C., de Bruijn, R.F., Passmore, P., Montine, T.J., Bettens, K., Rotter, J.I., Brice, A., Morgan, K., Foroud, T.M., Kukull, W.A., Hannequin, D., Powell, J.F., Nalls, M.A., Ritchie, K., Lunetta, K.L., Kauwe, J.S., Boerwinkle, E., Riemenschneider, M., Boada, M., Hiltunen, M., Martin, E.R., Schmidt, R., Rujescu, D., Wang, L.S., Dartigues, J.F., Mayeux, R., Tzourio, C., Hofman, A., Nothen, M.M., Graff, C., Psaty, B.M., Jones, L., Haines, J.L., Holmans, P.A., Lathrop, M., Pericak-Vance, M.A., Launer, L.J., Farrer, L.A., van Duijn, C.M., Van Broeckhoven, C., Moskvina, V., Seshadri, S., Williams, J., Schellenberg, G.D., Amouyel, P. 2013. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* 45(12), 1452-8. doi:ng.2802 [pii]  
10.1038/ng.2802.
- Liang, W.S., Dunckley, T., Beach, T.G., Grover, A., Mastroeni, D., Ramsey, K., Caselli, R.J., Kukull, W.A., McKeel, D., Morris, J.C., Hulette, C.M., Schmechel, D., Reiman, E.M., Rogers, J., Stephan, D.A. 2008. Altered neuronal gene expression in brain regions differentially affected by Alzheimer's disease: a reference data set. *Physiol Genomics* 33(2), 240-56. doi:00242.2007 [pii]  
10.1152/physiolgenomics.00242.2007.
- Liang, Z., Liu, F., Grundke-Iqbal, I., Iqbal, K., Gong, C.X. 2007. Down-regulation of cAMP-dependent protein kinase by over-activated calpain in Alzheimer disease brain. *J Neurochem* 103(6), 2462-70. doi:JNC4942 [pii]

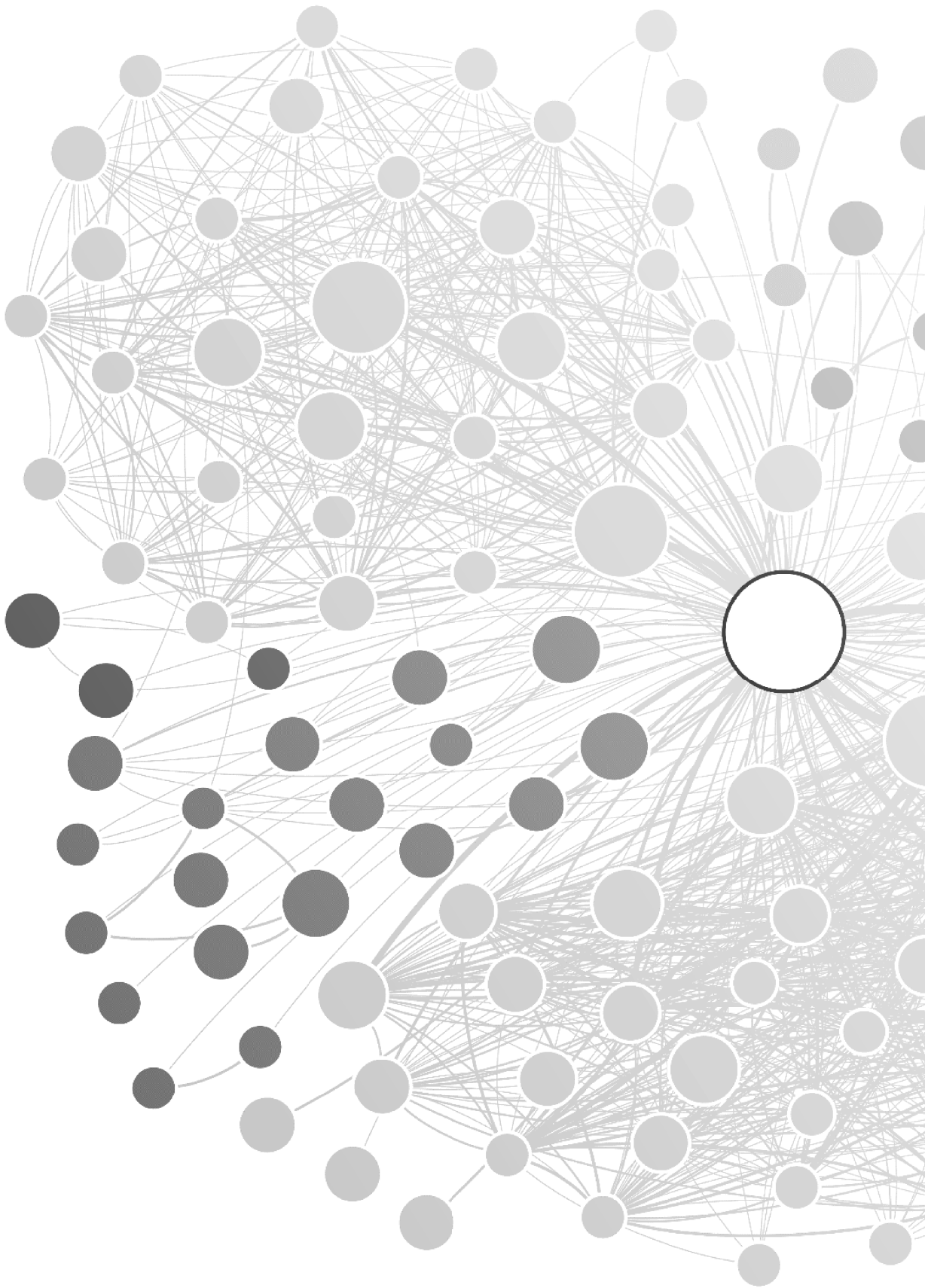
- 10.1111/j.1471-4159.2007.04942.x.
- Liao, Y., Smyth, G.K., Shi, W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30(7), 923-30. doi:btt656 [pii]  
10.1093/bioinformatics/btt656.
- Lillenes, M.S., Rabano, A., Stoen, M., Riaz, T., Misaghian, D., Mollersen, L., Esbensen, Y., Gunther, C.C., Selnes, P., Stenset, V.T., Fladby, T., Tonjum, T. 2016. Altered DNA base excision repair profile in brain tissue and blood in Alzheimer's disease. *Mol Brain* 9(1), 61. doi:10.1186/s13041-016-0237-z  
10.1186/s13041-016-0237-z [pii].
- Liu, Q., Zhang, J. 2014. Lipid metabolism in Alzheimer's disease. *Neurosci Bull* 30(2), 331-45. doi:10.1007/s12264-013-1410-3.
- Magistri, M., Velmeshev, D., Makhmutova, M., Faghihi, M.A. 2015. Transcriptomics Profiling of Alzheimer's Disease Reveal Neurovascular Defects, Altered Amyloid-beta Homeostasis, and Deregulated Expression of Long Noncoding RNAs. *J Alzheimers Dis* 48(3), 647-65. doi:JAD150398 [pii]  
10.3233/JAD-150398.
- McCarthy, D.J., Chen, Y., Smyth, G.K. 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* 40(10), 4288-97. doi:gks042 [pii]  
10.1093/nar/gks042.
- McFerrin, M.B., Chi, X., Cutter, G., Yacoubian, T.A. 2017. Dysregulation of 14-3-3 proteins in neurodegenerative diseases with Lewy body or Alzheimer pathology. *Ann Clin Transl Neurol* 4(7), 466-77. doi:10.1002/acn3.421  
ACN3421 [pii].
- Mirra, S.S., Heyman, A., McKeel, D., Sumi, S.M., Crain, B.J., Brownlee, L.M., Vogel, F.S., Hughes, J.P., van Belle, G., Berg, L. 1991. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part II. Standardization of the neuropathologic assessment of Alzheimer's disease. *Neurology* 41(4), 479-86.
- Morris, J.H., Apeltsin, L., Newman, A.M., Baumbach, J., Wittkop, T., Su, G., Bader, G.D., Ferrin, T.E. 2011. clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics* 12, 436. doi:1471-2105-12-436 [pii]  
10.1186/1471-2105-12-436.
- Murray, M.E., Graff-Radford, N.R., Ross, O.A., Petersen, R.C., Duara, R., Dickson, D.W. 2011. Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. *Lancet Neurol* 10(9), 785-96. doi:S1474-4422(11)70156-9 [pii]  
10.1016/S1474-4422(11)70156-9.
- Musunuri, S., Khoonsari, P.E., Mikus, M., Wetterhall, M., Haggmark-Manberg, A., Lannfelt, L., Erlandsson, A., Bergquist, J., Ingelsson, M., Shevchenko, G., Nilsson, P., Kultima, K. 2016. Increased Levels of Extracellular Microvesicle Markers and Decreased Levels of Endocytic/Exocytic Proteins in the Alzheimer's Disease Brain. *J Alzheimers Dis* 54(4), 1671-86. doi:JAD160271 [pii]  
10.3233/JAD-160271.
- Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 27(1), 29-34. doi:gkc043 [pii].

- Perez-Palma, E., Bustos, B.I., Villaman, C.F., Alarcon, M.A., Avila, M.E., Ugarte, G.D., Reyes, A.E., Opazo, C., De Ferrari, G.V., Alzheimer's Disease Neuroimaging, I., Group, N.-L.N.F.S. 2014. Overrepresentation of glutamate signaling in Alzheimer's disease: network-based pathway enrichment using meta-analysis of genome-wide association studies. *PLoS One* 9(4), e95413. doi:10.1371/journal.pone.0095413  
PONE-D-13-13386 [pii].
- Plouhinec, J.L., Roche, D.D., Pegoraro, C., Figueiredo, A.L., Maczkowiak, F., Brunet, L.J., Milet, C., Vert, J.P., Pollet, N., Harland, R.M., Monsoro-Burq, A.H. 2014. Pax3 and Zic1 trigger the early neural crest gene regulatory network by the direct activation of multiple key neural crest specifiers. *Dev Biol* 386(2), 461-72. doi:S0012-1606(13)00655-6 [pii]  
10.1016/j.ydbio.2013.12.010.
- Prince, M., Bryce, R., Albanese, E., Wimo, A., Ribeiro, W., Ferri, C.P. 2013. The global prevalence of dementia: a systematic review and metaanalysis. *Alzheimers Dement* 9(1), 63-75 e2. doi:S1552-5260(12)02531-9 [pii]  
10.1016/j.jalz.2012.11.007.
- Qiao, H., Foote, M., Graham, K., Wu, Y., Zhou, Y. 2014. 14-3-3 proteins are required for hippocampal long-term potentiation and associative learning and memory. *J Neurosci* 34(14), 4801-8. doi:34/14/4801 [pii]  
10.1523/JNEUROSCI.4393-13.2014.
- Qureshi, H.Y., Han, D., MacDonald, R., Paudel, H.K. 2013a. Overexpression of 14-3-3zeta promotes tau phosphorylation at Ser262 and accelerates proteosomal degradation of synaptophysin in rat primary hippocampal neurons. *PLoS One* 8(12), e84615. doi:10.1371/journal.pone.0084615  
PONE-D-13-25836 [pii].
- Qureshi, H.Y., Li, T., MacDonald, R., Cho, C.M., Leclerc, N., Paudel, H.K. 2013b. Interaction of 14-3-3zeta with microtubule-associated protein tau within Alzheimer's disease neurofibrillary tangles. *Biochemistry* 52(37), 6445-55. doi:10.1021/bi400442d.
- Rajmohan, R., Reddy, P.H. 2017. Amyloid-Beta and Phosphorylated Tau Accumulations Cause Abnormalities at Synapses of Alzheimer's disease Neurons. *J Alzheimers Dis* 57(4), 975-99. doi:JAD160612 [pii]  
10.3233/JAD-160612.
- Ray, M., Zhang, W. 2010. Analysis of Alzheimer's disease severity across brain regions by topological analysis of gene co-expression networks. *BMC Syst Biol* 4, 136. doi:1752-0509-4-136 [pii]  
10.1186/1752-0509-4-136.
- Robinson, M.D., McCarthy, D.J., Smyth, G.K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1), 139-40. doi:btp616 [pii]  
10.1093/bioinformatics/btp616.
- Rozpedek, W., Markiewicz, L., Diehl, J.A., Pytel, D., Majsterek, I. 2015. Unfolded Protein Response and PERK Kinase as a New Therapeutic Target in the Pathogenesis of Alzheimer's Disease. *Curr Med Chem* 22(27), 3169-84. doi:CMC-EPUB-69609 [pii].
- Sancesario, G.M., Bernardini, S. 2015. How many biomarkers to discriminate neurodegenerative dementia? *Crit Rev Clin Lab Sci* 52(6), 314-26. doi:10.3109/10408363.2015.1051658.
- Schaefer, M.H., Serrano, L., Andrade-Navarro, M.A. 2015. Correcting for the study bias associated with protein-protein interaction measurements reveals differences between protein degree distributions from different cancer types. *Front Genet* 6, 260. doi:10.3389/fgene.2015.00260.
- Scheltens, P., Blennow, K., Breteler, M.M., de Strooper, B., Frisoni, G.B., Salloway, S., Van der Flier, W.M. 2016. Alzheimer's disease. *Lancet* 388(10043), 505-17. doi:S0140-6736(15)01124-1 [pii]

- 10.1016/S0140-6736(15)01124-1.
- Sekar, S., McDonald, J., Cuyugan, L., Aldrich, J., Kurdoglu, A., Adkins, J., Serrano, G., Beach, T.G., Craig, D.W., Valla, J., Reiman, E.M., Liang, W.S. 2015. Alzheimer's disease is associated with altered expression of genes involved in immune response and mitochondrial processes in astrocytes. *Neurobiol Aging* 36(2), 583-91. doi:S0197-4580(14)00633-2 [pii]  
10.1016/j.neurobiolaging.2014.09.027.
- Selkoe, D.J., Hardy, J. 2016. The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Mol Med* 8(6), 595-608. doi:emmm.201606210 [pii]  
10.15252/emmm.201606210.
- Shi, J., Qian, W., Yin, X., Iqbal, K., Grundke-Iqbal, I., Gu, X., Ding, F., Gong, C.X., Liu, F. 2011. Cyclic AMP-dependent protein kinase regulates the alternative splicing of tau exon 10: a mechanism involved in tau pathology of Alzheimer disease. *J Biol Chem* 286(16), 14639-48. doi:M110.204453 [pii]  
10.1074/jbc.M110.204453.
- Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., Ideker, T. 2011. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 27(3), 431-2. doi:btq675 [pii]  
10.1093/bioinformatics/btq675.
- Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102(43), 15545-50. doi:0506580102 [pii]  
10.1073/pnas.0506580102.
- Sutherland, G.T., Janitz, M., Kril, J.J. 2011. Understanding the pathogenesis of Alzheimer's disease: will RNA-Seq realize the promise of transcriptomics? *J Neurochem* 116(6), 937-46. doi:10.1111/j.1471-4159.2010.07157.x.
- Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., Jensen, L.J., von Mering, C. 2017. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res* 45(D1), D362-D8. doi:gkw937 [pii]  
10.1093/nar/gkw937.
- Thal, D.R., Attems, J., Ewers, M. 2014. Spreading of amyloid, tau, and microvascular pathology in Alzheimer's disease: findings from neuropathological and neuroimaging studies. *J Alzheimers Dis* 42 Suppl 4, S421-9. doi:P0835582M025W416 [pii]  
10.3233/JAD-141461.
- Tomiyama, T. 2010. [Involvement of beta-amyloid in the etiology of Alzheimer's disease]. *Brain Nerve* 62(7), 691-9. doi:1416100713 [pii].
- Twine, N.A., Janitz, K., Wilkins, M.R., Janitz, M. 2011. Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease. *PLoS One* 6(1), e16266. doi:10.1371/journal.pone.0016266.
- Van Cauwenberghe, C., Van Broeckhoven, C., Sleegers, K. 2016. The genetic landscape of Alzheimer disease: clinical implications and perspectives. *Genet Med* 18(5), 421-30. doi:gim2015117 [pii]  
10.1038/gim.2015.117.
- van der Brug H, H.M., Cao Y. 2017. Heterogeneity in neurodegenerative disease (GSE95587). NCBI GEO.
- van Dongen, S., Abreu-Goodger, C. 2012. Using MCL to extract clusters from networks. *Methods Mol Biol* 804, 281-95. doi:10.1007/978-1-61779-361-5\_15.



- Vitvitsky, V.M., Garg, S.K., Keep, R.F., Albin, R.L., Banerjee, R. 2012. Na<sup>+</sup> and K<sup>+</sup> ion imbalances in Alzheimer's disease. *Biochim Biophys Acta* 1822(11), 1671-81. doi:S0925-4439(12)00165-2 [pii] 10.1016/j.bbadis.2012.07.004.
- von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., Snel, B. 2003. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* 31(1), 258-61.
- Wang, M., Roussos, P., McKenzie, A., Zhou, X., Kajiwara, Y., Brennand, K.J., De Luca, G.C., Crary, J.F., Casaccia, P., Buxbaum, J.D., Ehrlich, M., Gandy, S., Goate, A., Katsel, P., Schadt, E., Haroutunian, V., Zhang, B. 2016. Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer's disease. *Genome Med* 8(1), 104. doi:10.1186/s13073-016-0355-3 10.1186/s13073-016-0355-3 [pii].
- Wang, X., Bao, X., Pal, R., Agbas, A., Michaelis, E.K. 2010. Transcriptomic responses in mouse brain exposed to chronic excess of the neurotransmitter glutamate. *BMC Genomics* 11, 360. doi:1471-2164-11-360 [pii] 10.1186/1471-2164-11-360.
- Whitfield, D.R., Vallortigara, J., Alghamdi, A., Howlett, D., Hortobagyi, T., Johnson, M., Attems, J., Newhouse, S., Ballard, C., Thomas, A.J., O'Brien, J.T., Aarsland, D., Francis, P.T. 2014. Assessment of ZnT3 and PSD95 protein levels in Lewy body dementias and Alzheimer's disease: association with cognitive impairment. *Neurobiol Aging* 35(12), 2836-44. doi:S0197-4580(14)00434-5 [pii] 10.1016/j.neurobiolaging.2014.06.015.
- Yan, M.H., Wang, X., Zhu, X. 2013. Mitochondrial defects and oxidative stress in Alzheimer disease and Parkinson disease. *Free Radic Biol Med* 62, 90-101. doi:S0891-5849(12)01823-0 [pii] 10.1016/j.freeradbiomed.2012.11.014.
- Yuki, D., Sugiura, Y., Zaima, N., Akatsu, H., Takei, S., Yao, I., Maesako, M., Kinoshita, A., Yamamoto, T., Kon, R., Sugiyama, K., Setou, M. 2014. DHA-PC and PSD-95 decrease after loss of synaptophysin and before neuronal loss in patients with Alzheimer's disease. *Sci Rep* 4, 7130. doi:srep07130 [pii] 10.1038/srep07130.
- Zhou, Y., Hayashi, I., Wong, J., Tugusheva, K., Renger, J.J., Zerbinatti, C. 2014. Intracellular clusterin interacts with brain isoforms of the bridging integrator 1 and with the microtubule-associated protein Tau in Alzheimer's disease. *PLoS One* 9(7), e103187. doi:10.1371/journal.pone.0103187 PONE-D-14-15245 [pii].





# Chapter 4

## Sequencing brain DNA



# Chapter 4.1



## Somatic *TARDBP* variants as cause of Semantic Dementia

Jeroen van Rooij<sup>1,2</sup>, Merel Mol<sup>1</sup>, Shamiram Melhem<sup>1</sup>, Pelle van der Wal<sup>2</sup>, Pascal Arp<sup>2</sup>, Francesca Paron<sup>4</sup>, Laura Donker Kaat<sup>1,3</sup>, Harro Seelaar<sup>1</sup>, Netherlands Brain Bank, Suzanne SM Miedema<sup>5</sup>, Takuya Oshima<sup>6</sup>, Bart JL Eggen<sup>6</sup>, André Uitterlinden<sup>2</sup>, Joyce van Meurs<sup>2</sup>, Ronald E van Kesteren<sup>5</sup>, August B Smit<sup>5</sup>, Emanuele Buratti<sup>4</sup>, John C van Swieten<sup>1</sup>

<sup>1</sup> Department of Neurology, Erasmus Medical Center, Rotterdam, The Netherlands

<sup>2</sup> Department of Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands

<sup>3</sup> Department of Clinical Genetics, Erasmus Medical Center, Rotterdam, The Netherlands

<sup>4</sup> International Centre for Genetic Engineering and Biotechnology (ICGEB), Trieste, Italy

<sup>5</sup> Center for Neurogenomics and Cognitive Research, VU University, Amsterdam, The Netherlands

<sup>6</sup> Department of Biomedical Sciences of Cells & Systems, section Molecular Neurobiology, University of Groningen, University Medical Center Groningen (UMCG), Groningen, The Netherlands.

*Manuscript in Press in BRAIN (IF= 11.3), December 2020*

## Abstract

The etiology of late-onset neurodegenerative diseases is largely unknown. Here we investigated whether de novo somatic variants for semantic dementia can be detected, thereby arguing for a more general role of somatic variants in neurodegenerative disease. Semantic dementia is characterized by a non-familial occurrence, early onset (< 65 years), focal temporal atrophy and TDP-43 pathology. To test whether somatic variants in neural progenitor cells during brain development might lead to semantic dementia, we compared deep exome sequencing data of DNA derived from brain and blood of 16 semantic dementia cases. Somatic variants observed in brain tissue and absent in blood were validated using amplicon sequencing and digital PCR. We identified two variants in exon one of the TARDBP gene (L41F and R42H) at low level (1-3%) in cortical regions and in dentate gyrus in two semantic dementia brains, respectively. The pathogenicity of both variants is supported by demonstrating impaired splicing regulation of TDP-43 and by altered subcellular localization of the mutant TDP-43 protein. These findings indicate that somatic variants may cause semantic dementia as a non-hereditary neurodegenerative disease, which might be exemplary for other late-onset neurodegenerative disorders.

**Keywords;** semantic dementia; somatic variants; TARDBP; TDP-43

## Introduction

Multifactorial etiology, including genetic and environmental factors, has been invoked to explain most late-onset neurodegenerative diseases. Only a small percentage of cases with autosomal dominant inheritance is caused by germline variants in specific genes, for example *PSEN1* and *APP* variants in Alzheimer's disease, *MAPT* and *GRN* in frontotemporal dementia and *C9orf72* and *TARDBP* in both amyotrophic lateral sclerosis and frontotemporal dementia (Ferrari *et al.*, 2019; Greaves and Rohrer, 2019; Clarimon *et al.*, 2020). There is an increasing interest in the potential pathogenic role of *de novo* variants in patients with neurodegenerative diseases with a negative family history (Leija-Salazar *et al.*, 2018; Lodato and Walsh, 2019). A few cases with *de novo* germline variants have been identified in early-onset Alzheimer's disease (Nicolas *et al.*, 2018). For neurodevelopmental diseases, low-level ( $\leq 20\%$  of cells) somatic variants in *mTOR*, *AKT3* and *CCND* arising from the ventricular or subventricular zone have been identified by deep sequencing of candidate genes in affected brain tissue (Lee *et al.*, 2012; Lin *et al.*, 2012; Veltman and Brunner, 2012; Miller *et al.*, 2013; Poduri *et al.*, 2013; Hu *et al.*, 2014; Jamuar *et al.*, 2014; Kovacs *et al.*, 2014; Mirzaa *et al.*, 2014; Rogalski *et al.*, 2014; Bushman *et al.*, 2015; Lim *et al.*, 2015; Lodato *et al.*, 2015; Sala Frigerio *et al.*, 2015; Wiseman *et al.*, 2015; Hoekstra *et al.*, 2016; Kim *et al.*, 2016; Takata *et al.*, 2016). The hypothesis is that post-zygotic variants (after fertilization) or late-somatic variants during brain development might explain the sporadic presentation of neurodegenerative diseases with a negative family history.

The most ideal approach to determine the role of late-somatic variants in neurodegenerative diseases would be the comparison between blood- and brain-derived DNA within the same patients. However, brain tissue for DNA isolation was often not available during life, and DNA derived from blood was often not collected during life in deceased patients. Recent brain-derived DNA studies without matched DNA samples from blood have tried to detect somatic variants in Alzheimer's and Parkinson's disease (Beck *et al.*, 2004; Lin *et al.*, 2012; Proukakis *et al.*, 2014; Bushman *et al.*, 2015; Lodato *et al.*, 2015; Sala Frigerio *et al.*, 2015; Wiseman *et al.*, 2015; Coxhead *et al.*, 2016; Hoekstra *et al.*, 2016; Lee *et al.*, 2018; Lodato *et al.*, 2018; Mokretar *et al.*, 2018; Nicolas *et al.*, 2018; Park *et al.*, 2019; Wei *et al.*, 2019). A higher number of low-level mosaic variants in causative genes (*APP*, *SNCA*) in DNA of Alzheimer's disease or Parkinson's disease brains compared to controls (Lee *et al.*, 2018; Mokretar *et al.*, 2018). Only the study by Park *et al.* performing deep sequencing of hippocampal formation *and* matched blood tissues found an enrichment of somatic DNA variation in the Tau signaling pathway in Alzheimer's disease patients compared to controls (Park *et al.*, 2019). Specifically, a single carrier of a somatic variant in *PIN1* was suggested as potential causal factor in the respective Alzheimer's disease patient (Park *et al.*, 2019).

In the present study, we uniquely investigated the presence of low-level somatic variants in the temporal cortex and dentate gyrus of brains of semantic dementia patients, which were absent in their blood-derived DNA. Semantic dementia is a well-defined clinical and pathological subtype of frontotemporal dementia, mostly occurring before the age of 65 (Hodges *et al.*, 1992; Irish *et al.*, 2012; Mesulam *et al.*, 2014). The disease is characterized by a very circumscribed asymmetric atrophy of the anterior temporal cortex, suggesting a very local disease process (Mummery *et al.*, 2000; Kumfor *et al.*, 2016). Severe neuronal loss with

pathological TDP-43 protein accumulation in neurites and neurons in the temporal cortex and dentate gyrus of the hippocampus are the defining salient and consistent neuropathological features of semantic dementia, most commonly classified as FTD-TDP type C (Davies *et al.*, 2005; Mackenzie *et al.*, 2011; Leyton *et al.*, 2016; Neumann and Mackenzie, 2019). Semantic dementia has a sporadic, non-familial occurrence, and a current lack of mechanistic insight in the disease process precludes a therapeutic strategy. We performed deep exome sequencing (310x-658x) of middle temporal gyrus and dentate gyrus tissue of semantic dementia patients with pathologically confirmed FTD-TDP type C, and compared data with blood DNA samples of the same patients. We identified somatic *TARDBP* variants in the brains of two semantic dementia patients that were absent in blood. These variants were validated using custom amplicon panel sequencing and digital droplet PCR. In addition, we confirmed the disruptive effects of these *TARDBP* variants by demonstrating altered cellular distribution of the mutant TDP-43 proteins. Our results indicate that somatic variants in *TARDBP* contribute to semantic dementia pathogenesis.



## Methods

### ***Patient tissue DNA collection***

For the present study, we used fresh-frozen brain samples from 16 semantic dementia patients with confirmed FTD-TDP type C pathology, obtained from the Netherlands Brain Bank (table 1) (Mackenzie and Neumann, 2017). Informed consent was obtained from all patients for brain autopsy and the use of tissue and clinical information for research purposes. DNA was extracted from fresh frozen brain samples of middle temporal gyrus (n=14) and from the dentate gyrus (n=13). From all cases, DNA from blood was available, obtained during life in 12 patients from the Dutch frontotemporal dementia study and extracted from blood obtained at the time of autopsy in the remaining four cases (Seelaar et al., 2008; Seelaar et al., 2011). The average age at death was 69 (range 62-74), 50% of patients were female. Medical records and neuroimaging (either CT or MRI) were collected and reviewed, if available. For 14 patients the left hemisphere of the brain was fresh-frozen for research, versus the right hemisphere for two patients.

**Table1.** Patient characteristics. Contains clinical and pathological information on the patients examined in this study. Pathological diagnosis, as extracted from the reports from the Netherlands Brain Bank. The most affected side of the brain is reported according following post-mortem pathological examination. Brain tissue side: the side of the brain fresh frozen and used in this study. Dominant side yes/no; whether the side studied was the one most affected according to neuroimaging (NA = not applicable, as both sides were equally affected). MTG and DG indicate whether the middle temporal gyrus and dentate gyrus were available and included in the study.

Patient	Sex	Age at onset	Disease duration	Dominant side pathology	Brain tissue side	Dominant side	MTG	DG
SD01	F	60	10	both	left	no	yes	no
SD02	M	48	14	left	left	yes	no	yes
SD03	F	60	8	both	left	NA	yes	no
SD04	F	45	20	both	left	NA	yes	yes
SD05	M	56	10	both	left	no	yes	yes
SD06	M	51	12	both	left	no	yes	yes
SD07	F	53	11	both	left	no	no	yes
SD08	M	57	12	left	left	yes	yes	yes
SD09	F	63	11	both	left	NA	yes	yes
SD10	M	55	13	left	right	no	yes	yes
SD11	M	51	15	both	left	no	yes	yes
SD12	F	60	12	both	left	no	yes	yes
SD13	F	63	9	left	right	no	yes	yes
SD14	M	57	15	both	right	no	yes	yes
SD15	F	66	8	left	left	yes	yes	no
SD16	M	61	13	both	left	yes	yes	yes

### **Whole exome sequencing**

Blood-derived DNA of 16 patients and brain-derived DNA from middle temporal gyrus (n=14) and/or dentate gyrus (n=13) of semantic dementia brains (n=16) was captured using Nimblegen's SeqCap or MedExome library prep kits and sequenced to an average depth of 139x, 496x and 395x respectively. Reads were mapped to the hg19 reference genome using BWA and processed using picard and GATK, following best practices. Candidate variants were called using thresholds to detect variants present in the brain (>5 reads), but absent in blood ( $\leq 1$  read). The next three filtering steps for candidate variants were; 1) a custom signal to noise filter (S2N  $\geq 5$ ) as described in the supplementary methods, 2) a minor allele frequency less than 0.01% in the ExAC database and 3) a CADD score above 10.

### **Validation amplicon panel sequencing**

We validated a selection of candidate variants (present in brain, absent in blood) to confirm true-positive variants, and two candidate genes (*GRN* and *TARDBP*) to exclude false negatives, by amplicon panel sequencing of the same DNA samples used in the discovery whole exome sequencing. All candidate variants in these targets were included in a custom amplicon panel (SWIFT, product code SW CP-ER6161) and sequenced to an average depth of 1,601x on a MiSeq v3 with 600 cycles. A second round of amplicon panel sequencing was carried out for further classification of somatic variants of interest in DNA from additional cortex regions (middle frontal gyrus, superior parietal lobe) and cerebellum of two semantic dementia brains, and in DNA from middle temporal gyrus of 66 non-demented control brains from the Netherlands Brain Bank. Data analysis of the panel was done similarly to the discovery. Candidate somatic variants were validated when; 1) read depth in the validation was at least 100, 2) the variant allele count was at least 20 in DNA of the brain, 3) the variant allele frequency was at least 1% in DNA of the brain, 4) variant allele frequency was absent in blood of the same patient.

### **Validation of *TARDBP* variants**

We performed additional validation using digital droplet PCR of two *TARDBP* somatic variant carriers. In short, custom LNA FAM+HEX probes for each variant were designed and optimized by TATAA Biocenter (Göteborg, Sweden). Synthetic DNA fragments (gBlock™) with these variants were generated to serve as positive controls and as a dilution ladder for technical evaluation of the assay. Negative controls were water and DNA of middle temporal gyrus from two unrelated non-demented controls. Each assay was tested on five brain regions of the carrier (medial temporal gyrus, medial frontal gyrus, superior parietal gyrus, dentate gyrus and cerebellum), blood and the two negative controls. Droplets were generated using Bio-Rad's Droplet Generation Oil for Probes (cat#1863005) in combination with the qPCR Droplet PCR supermix (no dUTP, Bio-Rad cat#1863024) on a Bio-Rad QX200 Droplet Generator. The PCR plate was measured using the QX200 Droplet Reader (Bio-Rad) and analyzed with the Quantasoft Analysis Pro software (Bio-Rad). Reactions with less than 10,000 accepted droplets were not utilized in the analysis. Sensitivity rates of the assays were established using 0.1%, 1.0% and 2.5% spiked positive control gBlock™ mutation fragments and subsequently used to estimate variant allele frequencies by the ratio of FAM-positive droplets over HEX-positive droplets.

### **Germline variants in blood and brain**

To exclude (*de novo*) germline variants in twelve FTD (*CHMP2B*, *DPP6*, *FUS*, *GRN*, *MAPT*, *OPTN*, *SQSTM1*, *TARDBP*, *TBK1*, *TREM2*, *UNC13A* and *VCP*) candidate genes we performed regular germline variant calling using GATK's Haplotypecaller using best practices (van Rooij *et al.*, 2017; Ferrari *et al.*, 2019; Greaves and Rohrer, 2019; Clarimon *et al.*, 2020). Variants were annotated using Annovar and were manually evaluated based on exonic function, CADD score, frequency in GnomAD, variant allele frequency and presence in the other tissues of the same patient.

### **Functional analysis of somatic TARDBP variants**

The functional impact of both somatic *TARDBP* variants on the TDP-43 protein was assessed by a previously published add-back splicing assay and by immuno-fluorescent microscopy of TDP-43 in HeLa cells (D'Ambrogio *et al.*, 2009). In short; the splicing assay contains a minigene construct containing *CFTR* exon 9 carrying a mutation (C155T) in an exonic splicing enhancer sequence in order to have an approximately 50% of in- or out-splicing of exon 9. Using wild type TDP-43 as positive control, and complete loss-of-function F4L mutated TDP-43 as negative control, the relative impact of L41F and R42H on TDP-43 function could be ascertained. To obtain p-values, an unpaired t-test was carried out using GraphPad software (GraphPad Software, La Jolla California, USA). For the immunofluorescence assays, HeLa cells were transfected with wild type TDP-43 or with TDP-43 carrying variants L41F or R42H. Nuclei were located by chromatic staining of DAPI, and co-localization of TDP-43 is identified by FLAG-TDP-43 protein, as published previously (Mompean *et al.*, 2017). FLAG TDP-43 staining was quantified using regions of interest for nuclear and cytoplasmic signal using Fiji\_ImageJ software. The percentage of nuclear and cytoplasmic fluorescent signal was measured for nine cells each for the wild type, L41F and R42H TDP-43 expressing cells. Statistical tests were performed using 2 way-ANOVA in GraphPad for nuclear-cytoplasmic TDP-43 localization within each cell-line, as well as between the wild type and the L41F or R42H TDP-43 transfected cells.

### **Cell-type specificity of somatic R42H TARDBP variant**

For the R42H *TARDBP* variant carrier, we performed Fluorescence-Activated Nuclear Sorting (FANS) on the frontal lobe and parietal lobe, then isolated DNA from the nuclei with QIAamp DNA Micro Kit (QIAGEN, Germany). Using NeuN and Olig2 as cell surface markers, we separated neurons (NeuN-positive) and oligodendrocytes (Olig2-positive) from microglia, astrocytes and any other nuclei (Double negative). Parietal cortex tissue from a dementia patient unrelated to this study was similarly sorted and used as negative control. Each resulting DNA sample was amplicon sequenced and analyzed using the described procedures.

## Results

### ***Deep whole exome sequencing***

All DNA samples from middle temporal gyrus (n=14), dentate gyrus (n=13) and blood (n=16) were sequenced to an average depth of 496 (range 429-658), 395 (range 310-520) and 139 (range 72-229), respectively.

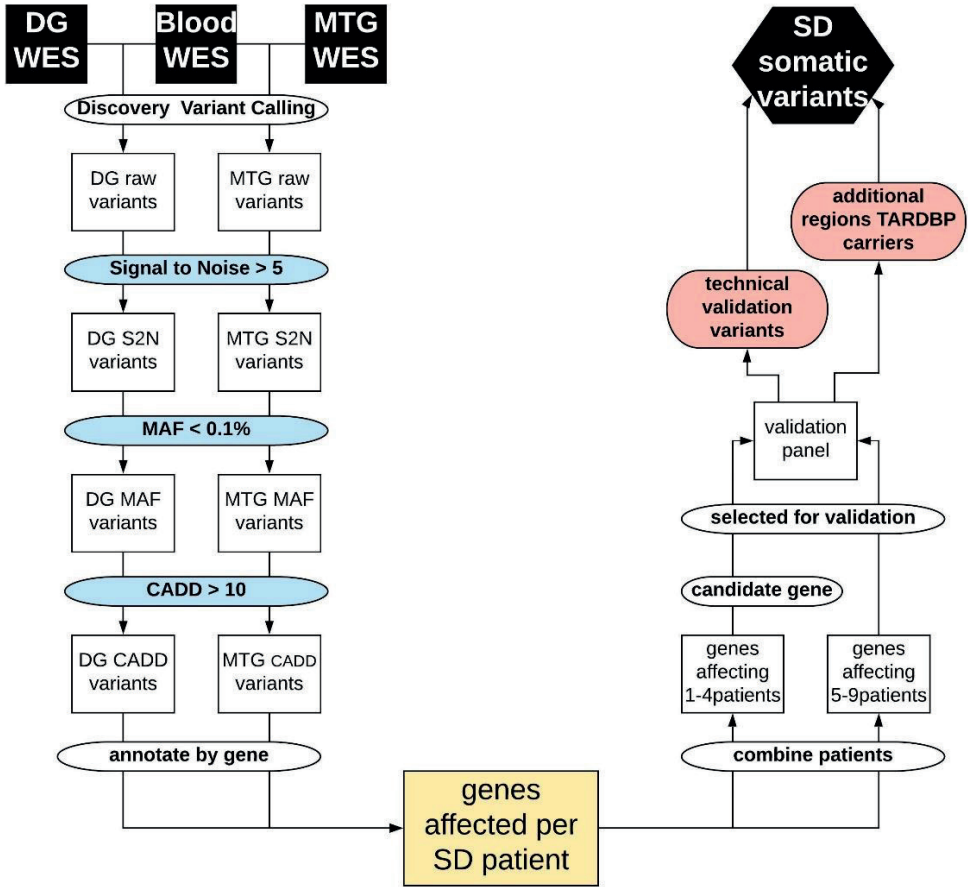
### ***Exclusion of causal germline variants***

Germline variant analysis in the whole exome sequencing data of all semantic dementia patients did not result in known pathogenic variants in any of the 12 known frontotemporal dementia genes. One patient was identified as germline carrier of the V90A variant in *TARDBP*, which was also reported in controls and thus considered of uncertain significance (supplemental table 1) (Borroni *et al.*, 2010; Lattante *et al.*, 2013; Caroppo *et al.*, 2016).

### ***Discovery and validation of somatic variants in semantic dementia brains***

After signal to noise, minor allele frequency and CADD score filtering we retained on average 172 variants for dentate gyrus and 57 for middle temporal gyrus per patient (figure 1 and supplemental figure 1). We detected variants in 1,450 genes from the dentate gyrus and/or middle temporal gyrus of at least one semantic dementia patient and absent in blood. To confirm true-positive variants, we selected a set of 305 variants for validation in a panel of amplicon sequencing based on one of the two following criteria: 1) somatic variants present in at least five brains (resulting in 252 variants in a total of 128 genes), or 2) variants in candidate genes involved in neurodevelopmental or neurodegenerative diseases (resulting in 53 variants in 51 genes present in 1-4 brains). Amongst the 51 candidate genes fulfilling the second criterion were single variant carriers in *TARDBP* (R42H) and in *GRN*.

We identified a total of eight true-positive variants in the panel of amplicon sequencing ( $\geq 100\times$  depth in both brain and blood, variant observed  $\geq 20$  times in the brain, variant allele frequency of  $\geq 1\%$  in brain and  $\leq 1\%$  in blood). Seven of those were previously detected with exome sequencing, whereas the eighth variant was not detected in exome sequencing but identified through rescreening of the *TARDBP* gene in the amplicon sequencing data (table 2). The nonsynonymous variant (R42H) in *TARDBP*; chr1:11073909-G/A with a CADD score of 20 was the most significantly replicated variant (271 out of 18,990 sequenced fragments in middle temporal gyrus, and none out of 5,126 fragments in blood) and completely absent from gnomad (variant allele frequency of 1.4% in the middle temporal gyrus of a single semantic dementia brain).

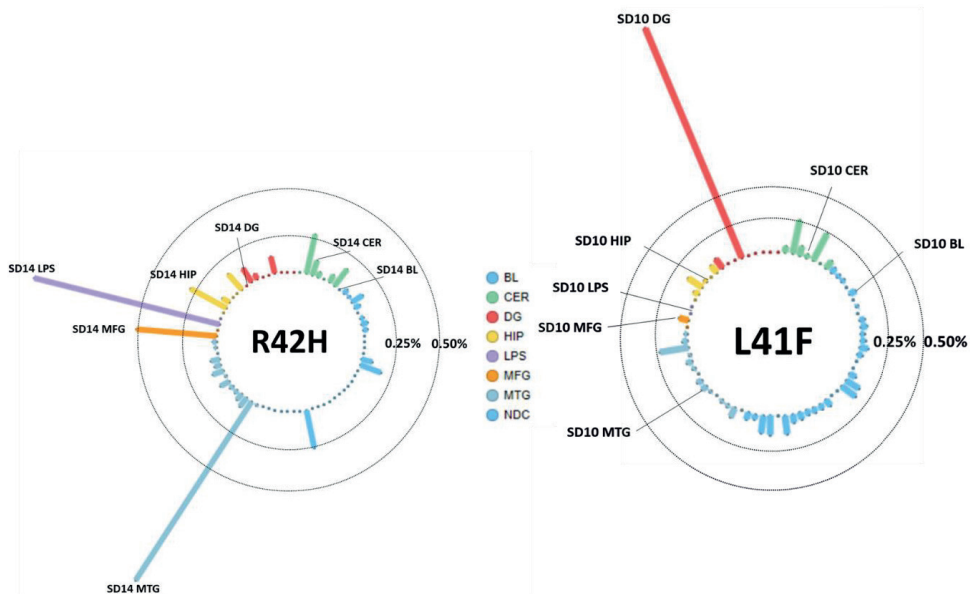


**Figure 1.** flowchart of data filtering and analysis. Starting from top-left; raw somatic variant calling using blood and dentate gyrus (DG) or medial temporal gyrus (MTG) deep exome sequencing data (WES), signal to noise filter (S2N), minor allele frequency filter (MAF), CADD score filter, annotating and grouping per gene, resulting in the genes affected in each SD patient. On the right side; grouping genes affecting multiple patients (>5) or affecting candidate genes in fewer patients (1-4) to be included in the validation amplicon panel. To excluded false negative findings in the WES data in FTD-TDP known germline causal genes *GRN* and *TARDBP*, all exons in these genes were included in the validation panel. The first validation round was performed on the same tissues as the discovery WES to confirm true positive variants from the WES, or identify false negative findings in *GRN* or *TARDBP*. The second round of validation further classified true positive variants in additional brain tissues and non-demented controls.

**Table 2.** Summary of results of 8 validated somatic variants. All eight variants passed the validation criteria. The middle panel shows the read counts for variant and wildtype from exome sequencing. The right panel shows the counts for the amplicon panel. S2N = signal to noise, MAF = minor allele frequency, VAF = variant allele frequency. P-values are obtained by fisher exact tests of the counts between blood and brain.

variant	gene	function	Sample	Tissue	S2N	MAF	CADD	WES-Blood				WES-Brain				Panel-Blood				Panel-Brain			
								REF	ALT	VAF	p-value	REF	ALT	VAF	p-value	REF	ALT	VAF	p-value	REF	ALT	VAF	p-value
chr1:11073909:G/A	TARDBP	nonsyn	16-079	MTG	9.8	0	20	87	0	0.000	765	10	0.013	6E-01	5126	0	0.000	18719	271	0.014	9E-29		
chr1:11073905:C/T	TARDBP	nonsyn	12-124	DG	NA	0	28	138	0	0.000	1037	0	0.000	1E+00	9349	4	0.000	7533	152	0.020	3E-47		
chr9:130928555:T/G	CIZ1	nonsyn	12-128	DG	20.8	0	16	39	0	0.000	323	54	0.143	5E-03	134	0	0.000	346	35	0.092	3E-05		
chr17:4883215:A/G	CAMTA2	nonsyn	12-072	DG	11.1	0	18	58	1	0.017	420	42	0.091	7E-02	146	1	0.007	1058	24	0.022	4E-01		
chr17:34192311:T/G	HEATR9	nonsyn	12-128	DG	7.6	0	13	72	0	0.000	278	21	0.070	2E-02	3232	24	0.007	18031	329	0.018	3E-06		
chr17:701119726:C/A	SOX9	nonsyn	16-079	DG	9.0	0	20	56	1	0.018	405	15	0.036	7E-01	1676	5	0.003	1133	76	0.063	9E-24		
chr19:51133405:A/G	SYT3	nonsyn	12-128	DG	8.5	0	14	14	0	0.000	227	25	0.099	4E-01	103	0	0.000	1593	29	0.018	4E-01		
chr19:51133405:A/G	SYT3	nonsyn	12-072	DG	12.4	0	14	21	0	0.000	236	40	0.145	9E-02	289	1	0.003	1829	29	0.016	2E-01		

A second non-synonymous variant in the same exon; chr1:11073905-C/T (L41F) in the *TARDBP* gene was detected in the dentate gyrus of another patient with variant allele frequency of 2.0% in the amplicon panel sequencing data (152 out of 7,533 fragments,  $p=2.8E-47$ ,  $OR=47$ ,  $95\% CI=[18-175]$  compared to blood). This variant with a CADD score of 28 was also absent from the population databases, and was not observed in blood-derived DNA or any of other brain regions of the same patient (figure 2). Both variants observed in a single patient each were taken forward for further validation by digital PCR and functional testing, as germline variants in *TARDBP* are known to cause frontotemporal dementia and/or amyotrophic lateral sclerosis with TDP-43 pathology (Borrioni *et al.*, 2010; Lattante *et al.*, 2013; Caroppo *et al.*, 2016).



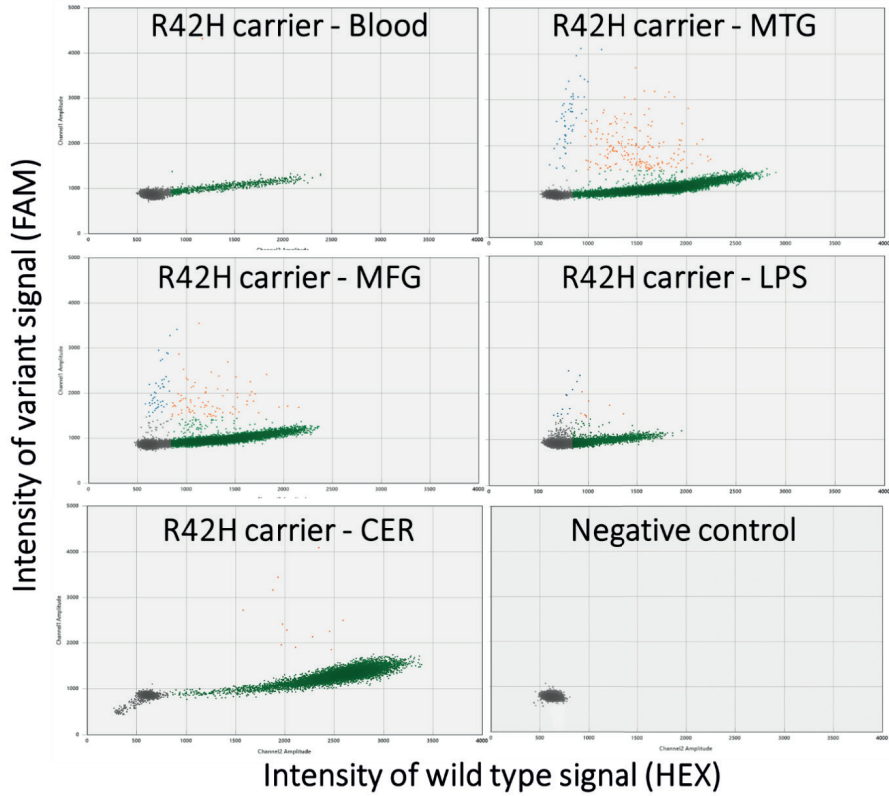
**Figure 2.** allele frequencies for L41F and R42H in all tested amplicon panel samples. Each column is a sample, the tissues represented by color; blood (BL, blue), cerebellum (CER, green), dentate gyrus (DG, red), hippocampus (HIP, orange), middle temporal gyrus (MTG, purple), middle frontal gyrus (MFG, salmon) and superior parietal lobe (LPS, pink). The vertical axis shows the variant allele frequency in that respective tissue, with lines representing the 0.25% and 0.50% thresholds. The tissues with highest VAF are labelled with the patient identifier and respective tissue.

### **Validation of TARDBP variant R42H by amplicon sequencing, digital droplet PCR**

After confirming presence of this variant in middle temporal gyrus (271 fragments out of 18,990,  $p=8.9E-29$ ) and absence in 5,123 sequenced fragments from blood, we validated this variant in other cortical regions of the same brain. We observed this variant with similar frequency in the parietal lobe (1.2%, 12 out of 973 fragments,  $p=2.3E-10$ ), in the frontal lobe (0.5%, 11 out of 2,122 fragments,  $p=1.3E-6$ ), and at lower frequency in the hippocampus (0.3%, 8 out of 3021 fragments,  $p=3.6E-4$ ) and cerebellum (0.1%, 3 out of 2881 fragments,  $p=4.7E-2$ ), although the variant allele frequency observed in hippocampus and cerebellum were within the range observed in the other samples, as shown in figure 2a. The variant was not observed among temporal cortex samples of 66 non-demented controls (0.03%, total 28 fragments out of 106,635, likely representing random sequencing errors). The R42H variant was then sequenced in only neuronal nuclei (NeuN-positive), oligodendrocyte nuclei (Olig2-positive) or the nuclear fraction containing, amongst others, astrocytes and microglia (and other NeuN/Olig2 double negative CNS cell nuclei) in both frontal and parietal lobe of the R42H carrier to an average depth of 3579x. The R42H variant was detected in 2.4% of the neurons in the parietal lobe and 1.1% in the frontal lobe (74 and 42 fragments out of 3093 and 3806 in total, respectively. These frequencies were doubled compared the bulk parietal and frontal tissue (1.2% and 0.5%, respectively). The variant was not observed in the control sample ( $<0.1\%$ ) and at 3-4 times lower frequencies in the oligodendrocytes or double negative nuclear fraction ( $<0.5\%$  in the parietal lobe and  $<0.4\%$  in the frontal lobe, respectively).

Validation using digital droplet PCR confirmed the amplicon sequencing results, as shown by the allelic discrimination plots (figure 3). The variant was observed in 242 droplets out of 13,048 non-empty droplets (variant allele frequency=1.9%) in the temporal lobe which was significantly higher than the negative controls; blood of the same patient (variant allele frequency=0.1%, 1 out of 809 droplets,  $p=1.1E-5$ ) and temporal lobe of 2 non-demented control (variant allele frequency=0.04%, 3 out of 7,698 droplets,  $p=1.5E-44$ ). Similarly, the variant was observed at significantly higher levels compared to the controls in the frontal lobe (variant allele frequency=1.3%, 126 droplets out of 9,549,  $p=6.7E-4$  and  $p=1.1E-28$ ) and parietal lobe (variant allele frequency=0.6%, 21 out of 3,697 droplets,  $p=0.16$  and  $p=3.5E-8$ ) and cerebellum (variant allele frequency=0.1%, 13 out of 9,231 droplets  $p=1.0$  and  $p=0.04$ ).





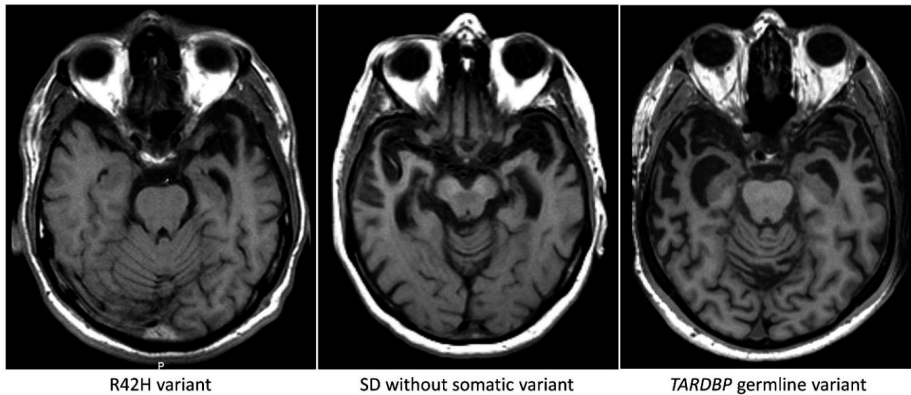
**Figure 3.** allelic discrimination plots of the digital droplet PCR for the R42H *TARDBP* somatic variant. Each marker represents a single droplet and its respective wild type (horizontal axis) and variant (vertical axis) signal intensity. Five different tissues of the carrier are tested; blood, middle temporal gyrus (MTG), middle frontal gyrus (MFG), lateral parietal lobe (LPS), cerebellum (CER) and a negative control of water is shown. The gray droplets are considered empty, green droplets are wild type only, orange is both wild type and variant alleles, and in blue are droplets harboring only the variant allele.

### ***Validation of TARDBP variant L41F by amplicon sequencing, digital droplet PCR***

The second *TARDBP* somatic variant in the same exon was detected in the dentate gyrus with a variant allele frequency of 2.0% in the amplicon panel sequencing data (152 out of 7,533 fragments,  $p=2.8E-47$ ) compared to blood (figure 2b). The variant was not observed among temporal cortex samples of 66 non-demented controls (0.04%, total 39 fragments out of 106,632, likely representing random sequencing errors). Validation with digital droplet PCR confirmed absence of the variant in blood, cerebellum, frontal lobe, temporal lobe and parietal lobe. Due to the low quantity of DNA from laser-capture microdissection-derived dentate gyrus, this tissue could not be tested using dPCR. This may have also influenced the WES result, in which many PCR duplicates were observed for the dentate gyrus data. We did not find any other somatic variants in the *TARDBP* gene in any of the other semantic dementia brains (average coverage across the gene of 1,116) and also not in middle temporal gyrus of non-demented control samples (average coverage of 103x across the gene).

### ***Clinicopathological description of the two cases with somatic TARDBP variants***

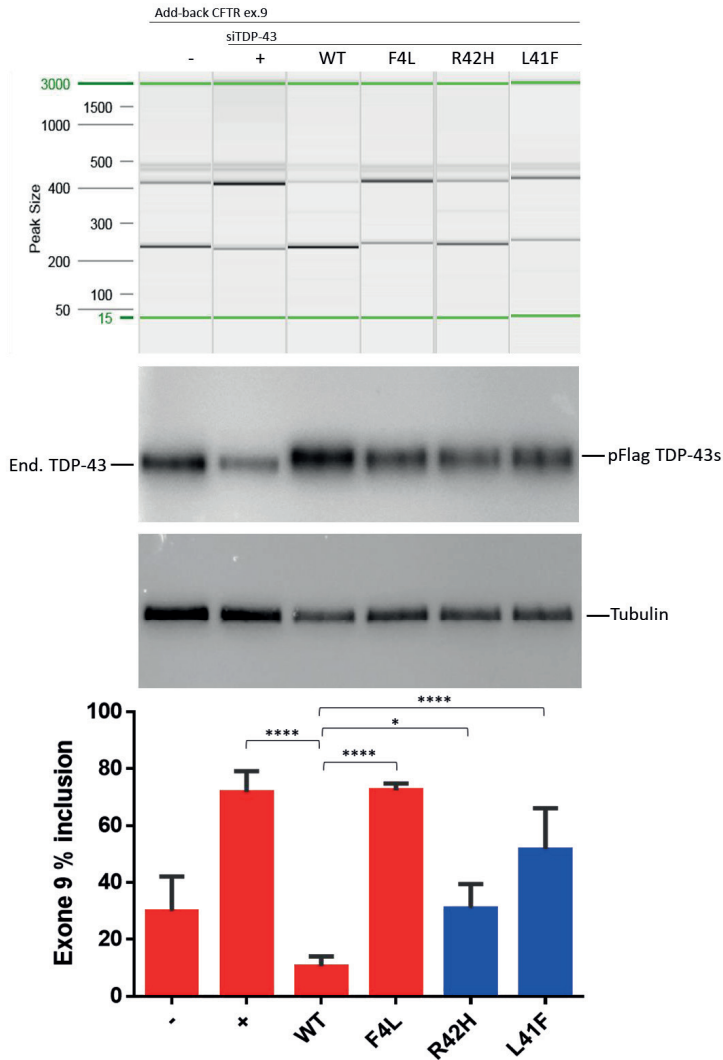
Both patients carrying the *TARDBP* L41F or R42H somatic variant developed progressive problems with word finding and language comprehension, and visual agnosia at the age of 55 and 57, respectively. Compulsive-obsessive behavior, loss of initiative and emotional lability were salient features in both patients, similar to the other 14 patients. Profound left-sided temporal atrophy was observed by neuroimaging (CT, MRI) two and three years after onset in both *TARDBP* carriers, in contrast to asymmetric but bilateral atrophy in the other semantic dementia patients (figure 4). Neuropathological examination after death (68 and 72 years respectively) showed severe anterior temporal atrophy, left more pronounced than right in the L41F carrier and more symmetrical in R42H. Microscopically, neuropathological changes were consistent with TDP-pathology type C, with severe neuron loss, gliosis in the temporal cortex with long thick threads and round cytoplasmic inclusions in granular cells of the hippocampus. For the L41F carrier, DNA of the middle temporal gyrus from the right hemisphere was available in the Netherlands Brain Bank and used for all DNA analyses, for the carrier of the R42H variant this was the middle temporal gyrus of the left hemisphere.



**Figure 4.** Axial T1-weighted MRI of the SD patient carrying somatic variant R42H, showing profound left sided temporal atrophy three years after disease onset. Pathological examination 15 years after disease onset showed atrophy of both temporal poles. The middle picture is from a patient without a somatic variant (four years after onset) showing atrophy of both temporal lobes. The right picture is a patient with the germline (p.I383V) *TARDBP* variant, showing a similar atrophy pattern bilaterally (four years after onset).

### **Functional analysis of *TARDBP* variants**

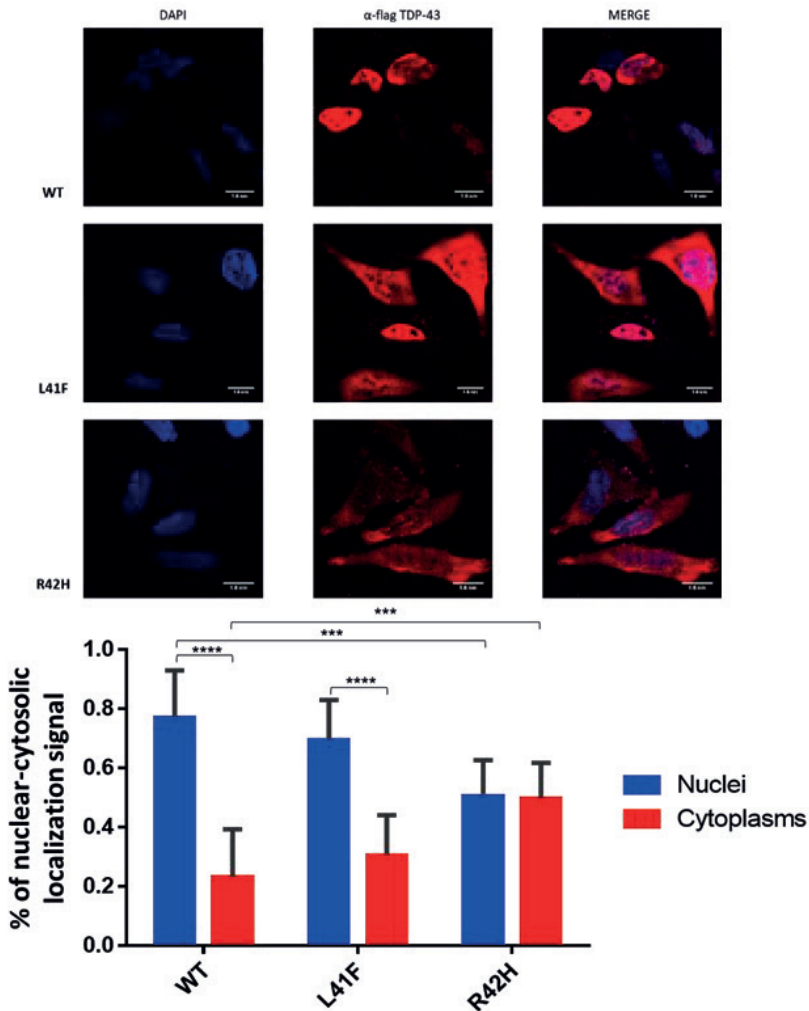
*TARDBP* is a protein involved in RNA splicing (Buratti and Baralle, 2001; D'Ambrogio *et al.*, 2009). Therefore, the impact of both *TARDBP* variants on the activity and localization TDP-43 was established in two assays; splicing regulation and cellular localization. The splicing assay contains a minigene construct containing *CFTR* exon 9 carrying a mutation (C155T) in an exonic splicing enhancer sequence in order to have an approximately 50% of in- or out-splicing of exon 9. The splicing is mediated by TDP-43 binding to the UG-repeat sequences near the 3' start site. Thus, when the function of TDP-43 is lost upon targeted siRNA treatment, a decrease to approximately 20% of exon 9 skipping is observed. Exon 9 skipping is then rescued by adding back wild type TDP-43 (WT) whose transcript has been made resistant to siRNA treatment. As negative control, we used a construct containing a TDP-43 that carries variant F4L, which is also resistant to the siRNA treatment but cannot bind RNA (Buratti and Baralle, 2001). In the presence of these positive and negative controls, the impact of uncharacterized TDP-43 variants can then be evaluated by comparing the amount of exon 9 skipping of each expressed variant. Both variants significantly decreased exon 9 skipping compared to wild type TDP-43, as shown in figure 5.



**Figure 5.** From left to right, the first two lanes show the baseline measurement with both splicing in and out of exon 9 in the absence (-) or presence of TDP-43 siRNA (+). Lane 3 shows that addition of si-resistant wild type TDP-43 can rescue the splicing functionality (WT) but this cannot be achieved by a TDP-43 carrying the F4L mutation that does not allow the protein to bind RNA (lane 4). Lanes 5 and 6 show the results obtained after the addition of mutated TDP-43 carrying the predicted damaging variants (R42H and L41F). In the Western blots below, we show equal expression of the flagged-TDP-43 WT and mutants (pFlag-TDP-43s) following knock down of the endogenous protein (end. TDP-43). Tubulin was used as an internal control. The upper figure shows the gel, the lower figure quantifies the ratio of CFTR exon 9 inclusion. The SD and p-values are reported for three independent experiments. Unpaired t-test was performed for statistical analysis (\*,  $P < 0.05$ ).

Splicing impairment was stronger for the L41F variant than for the R42H variant, in accordance with the predicted impact with CADD scores of 28 and 20, respectively. The impact on TDP-43 function was smaller for both variants compared to the siRNA-resistant TDP-43 variant F4L, which blocks RNA binding completely. Immunofluorescent staining demonstrated significantly altered localization of the R42H mutant TDP-43 protein compared to wild type TDP-43 (figure 6). In the wild type cells, 78% of the fluorescent signal was nuclear

(n=9), versus 71% for the L41F cells ( $p = 0.54$ ) and 52% for the R42H cells ( $p = 0.0004$ ). Only in the R42H TDP-43 expressing cells was TDP-43 no longer significantly localized in nuclei compared to cytoplasm. Region of interest measurements and statistical results are supplied in supplemental table 2.



**Figure 6.** Impact of *TARDBP* variants on the localization of flagged-TDP-43 wild type and mutant proteins overexpressed in Hela cells. The overexpressed proteins were visualized using anti-flag polyclonal antibody in a 100nm/pixel field. Scale bar = 10 nm. The first row shows wild type flag TDP-43, followed by flagged TDP-43s carrying both variants; L41F and R42H. The first column shows DAPI staining to indicate the chromatin in the nucleus in blue. The second column shows TDP-43 stained in red with a flag-specific antibody. The last column overlaps both figures, demonstrating TDP-43 localization in the nucleus for WT TDP-43, whilst localizing also in the cytoplasm for both TDP43 with variant R42H and L41F. In the bar plots below, fluorescent TDP-43 signal is quantified in the nucleus and cytoplasm for nine cells of each line. The average ratio of nuclear and cytosolic signal is plotted and compared between groups. P-values cutoffs are  $<0.0001$  (\*\*\*) or  $0.0001 < 0.001$  (\*\*\*) as calculated by 2-way ANOVAs between the groups illustrated.



## Discussion

The present study identified the occurrence of two low-level pathogenic somatic variants in the *TARDBP* gene in brains of patients with semantic dementia. These two variants in the first exon of the gene are absent from public databases and significantly affect TDP-43 function and localization. Moreover, the temporal lobe atrophy observed by MRI neuroimaging three years after onset in one of the two somatic *TARDBP* variant carriers resembled classical frontotemporal dementia due to germline *TARDBP* variants.

The observed low level (1-3%) of *TARDBP* somatic variants in brain-derived DNA was in accordance with the hypothesis that somatic variants occurred in one or more clones of neurons acquired in a single neural progenitor cells during brain development. Subsequently, the pathophysiological process arising from neurons carrying the somatic variants would then result in focal neurodegeneration later in life. The low percentage may further be attributed to by selective loss of neurons that carried the somatic variants in the affected brain region. The presence of somatic variants shared by (a) clone(s) of neurons in the temporal cortex or dentate gyrus was in contrast to recent studies, which investigated post-mitotic somatic mosaicism (pathogenic single-nucleotide variants and somatic copy-number variations) of known germline disease genes in individual cells (Lee *et al.*, 2018; Lodato *et al.*, 2018; Mokretar *et al.*, 2018). Such post-mitotic somatic variants increased with age in the latter studies and were found in significantly higher number in Alzheimer's disease or Parkinson's disease brains compared to controls (Lee *et al.*, 2018; Lodato *et al.*, 2018; Mokretar *et al.*, 2018). Although these somatic DNA variations for age-associated brain diseases were potential interesting, their causal role could not be determined for sure (Lodato *et al.*, 2018).

Post-mitotic variants are a less likely cause for semantic dementia patients as the disease occurs at a relatively young onset age (< 65 years) and its prevalence does not increase with age (Hodges *et al.*, 2010; Landin-Romero *et al.*, 2016). Therefore, our sequencing of DNA from bulk tissue, aiming to identify variants shared by neurons, and estimating their variant-allele-frequencies resembled the study of Park *et al.* in which somatic variants were found per brain region (hippocampal formation) in both Alzheimer's disease patients and controls (Park *et al.*, 2019).

The presence of single somatic variants (R42H and L41F) in the *TARDBP* gene in several neocortical regions (temporal, frontal and parietal) or dentate gyrus strongly points to the initial occurrence of somatic mosaicism in a single neural progenitor cell (Zilles *et al.*, 2013; Palomero-Gallagher and Zilles, 2019). Somatic variants in neurons arising from the ventricular or subventricular zone have also been shown in childhood or adult neurological diseases (Lee *et al.*, 2012; Lim *et al.*, 2015). By using blood-derived DNA from the same patients as control tissue, we could exclude somatic variants occurring from non-ectodermal lineages (Leija-Salazar *et al.*, 2018). Somatic *TARDBP* variants could be excluded from 66 non-demented controls by using temporal cortex-derived DNA. As the specific somatic variant (R42H) was absent in both hippocampus and cerebellum of the same patient, the variant must have occurred in neural progenitor cells of the lateral segment of the pallium, which develops into the neocortex (Zilles *et al.*, 2013; Palomero-Gallagher and Zilles, 2019). The variant was

enriched (twice as frequent compared to bulk cells) in the neuronal subpopulation of the parietal and frontal lobes, further suggesting the neural origin. A low signal (less than 20% of signal in the neuronal fraction) of the variant in the other nuclear fractions is a likely due to some residual neuronal nuclei present in the NeuN-negative fraction. Based on these results, we estimate that the R42H variant is present in 5.6%, 4.8% and 2.2% of the neurons in the temporal, parietal and frontal lobes, respectively. The second variant (L41F) was only detected in the hippocampus, suggesting that it occurred in neural progenitor cells of the medial segment of the pallium. The asymmetric onset of the disease pathology in these cases did not necessarily require the occurrence of the somatic variants after developmental separation of both hemispheres, as germline variants have also been associated with other asymmetric neurodegenerative disease processes (Stiles and Jernigan, 2010; Caroppo *et al.*, 2016; Gonzalez-Sanchez *et al.*, 2018). Although of interest, due to the collection procedure in the Netherlands Brain Bank, freezing only one hemisphere, the occurrence of absence of the variants in the other hemisphere could not be tested. The similarity in clinical and pathological phenotype (i.e., severe temporal atrophy, TDP-43 positive inclusions) between the somatic *TARDBP* variant carriers and germline *TARDBP* variant carriers supports the potential pathogenicity of these variants (Caroppo *et al.*, 2016; Gonzalez-Sanchez *et al.*, 2018).

Both *TARDBP* variants identified (L41F and R42H) are located in the first exon of *TARDBP* and are non-synonymous changes predicted to impact the N-terminal domain of the protein with CADD scores of 28 and 20 respectively (Chang *et al.*, 2012; Zhang *et al.*, 2013; Sasaguri *et al.*, 2016). Both variants are absent in human germline population databases ExAC and gnomAD; in fact, only eight germline variants in the first exon of *TARDBP* (amino acid 1-79) are described in the gnomAD database (120,000 participants), all extremely rare (<0.003%, 20 carriers across all eight variants combined). Our findings, identifying somatic variants in the N-terminal domain (amino acid 41 and 42) of *TARDBP*, are in contrast with all germline *TARDBP* gene variants for familial amyotrophic lateral sclerosis, and occasionally for familial frontotemporal dementia, reported in the glycine-rich region (GRR domain) between amino acids 262 and 414 of the TDP-43 protein (Barmada and Finkbeiner, 2010; Borroni *et al.*, 2010; Lattante *et al.*, 2013; Caroppo *et al.*, 2016; Wang *et al.*, 2016).

Our functional assays convincingly demonstrate a disruptive effect of both variants on normal *TARDBP* protein function. The impact on TDP-43 activity via *CFTR* minigene splicing was stronger for L41F than for R42H, with approximately 75% and 40% decrease of TDP-43 activity compared to wild type (D'Ambrogio *et al.*, 2009; Mompean *et al.*, 2017). Also, the redistribution of mutant TDP-43 in HeLa cells, from mostly nuclear in unaffected control to both cytoplasmic and nuclear for the R42H variants, supports the cellular pathogenicity. Together, both assays suggest that a correctly folded N-terminal domain of TDP-43 is required for nuclear localization and function, and that neurons carrying these somatic variants have dysfunctional TDP-43 and redistribution of TDP-43 protein to the cytoplasm as observed in frontotemporal dementia and amyotrophic lateral sclerosis brains (Chang *et al.*, 2012; Ihara *et al.*, 2013; Zhang *et al.*, 2013; Qin *et al.*, 2014; Romano *et al.*, 2015; Sasaguri *et al.*, 2016; Mompean *et al.*, 2017; Weskamp and Barmada, 2018). The resulting impact on TDP-43 function in shuttling RNA from the nucleus to the cytoplasm might lead to the protein aggregates observed in semantic dementia brains and subsequent pathogenicity for

the cells and tissue in which the variants are present (Barmada and Finkbeiner, 2010; Igaz *et al.*, 2011).

It is unclear how dysfunction of a small percentage of affected neurons (2-6%, double the variant allele frequency) would lead or contribute to extensive degeneration of the temporal lobe and widespread pathology (10-15% of neurons) in the dentate gyrus. Potentially, neuronal dysfunction within one brain region can accumulate until neuron-neuron signaling is sufficiently impaired to functionally disrupt the entire region. Another consideration is that the current study considers mosaicism in bulk DNA of all neurons in the temporal lobe and/or dentate gyrus, whereas many subtypes of neurons exist in these regions, leaving the possibility that the small number of affected neurons in these patients are enriched for a specific neuronal subtype. In Alzheimer's disease, for instance, a selective loss of parvalbumin-positive GABAergic interneurons (~3% of the total neuronal population) has been observed (Brady and Mufson, 1997), and the selective dysfunction of these neurons has been causally linked to global brain network changes and progressive amyloid pathology (Verret *et al.*, 2012; Iaccarino *et al.*, 2016; Hijazi *et al.*, 2019), indicating that small populations of affected neurons can indeed contribute to more widespread neurodegenerative processes. The challenges in interpreting selective neuronal dysfunction in the context of widespread neurodegeneration are exemplary of the overall discussion on how neurodegeneration starts and progresses (often differently between patients) throughout the brain, regardless of initial cause of the disease. Further work is needed to fully understand these processes and place the contribution of developmental and post-mitotic somatic DNA variation in the context of disrupted brain function. Cell-specific studies of semantic dementia brains carrying these somatic *TARDBP* variants may determine in which neuronal subtypes the somatic variants were present.

An important question that remains is why somatic variants were not found in all 14 brains with semantic dementia. There are several potential explanations, some of which include limitations of this study. 1) The bioinformatic filtering steps (absent in blood, CADD score > 10) may have been too stringent and removed potentially causing somatic variants, 2) pathogenic non-coding variants may have been not detected by the present exome sequencing, and low-level copy number variants missed by the present approach, 3) causal somatic variants may have become undetectable (disappeared) due to neuron loss in medial temporal gyrus during the neurodegenerative process, 4) the disease may have originated from causal somatic variants that were only present in the temporal cortex or dentate gyrus opposite to the side of the examined fresh-frozen brain samples, even though we expected that somatic variants occurred prior to the hemisphere separation in brain development, 5) multifactorial genetic or non-genetic factors may be responsible for most of the semantic dementia cases. Finally, we may have overlooked relevant variant in the WES data by first focusing on shared variants or damaging variants in candidate genes, which may be less likely true variants. Additionally, the low-level (<0.5%) error rate of the sequencing requires stringent filtering which may exclude variants that could be detected through panel sequencing, and further investigation of the data may uncover additional relevant variants, as was observed for the L41F variant. Pathogenicity of the remaining six variants confirmed by panel sequencing validation must be validated by future studies.



An interesting issue is whether somatic variants present in TDP-43 related genes may trigger dysregulation in the TDP-43 pathway. In analogy to this, Park et al reported a significant enrichment of somatic variants in the PI3K-AKT, MAPK and AMPK pathways in Alzheimer's disease brains versus control brains (Park *et al.*, 2019). Using a KEGG pathway overrepresentation analysis, they hypothesized that multiple disease-causing somatic variants converge onto pathways that potentially affect tau phosphorylation. In our view, the next step would be to perform amplicon panel sequencing of a set of FTD-TDP related genes on both semantic dementia brains and controls in order to detect potential additional causal somatic variants in the TDP-43 pathway. Moreover, investigating other series of SD brains may support our findings, and may give a better estimation of their frequency in semantic dementia. Finally, the present findings raise the question whether somatic variants may be causative in other types of frontotemporal dementia, for example somatic variants in *MAPT* causing sporadic Pick's disease. Overall, it seems warranted to carry out such targeted deep sequencing in all well-defined dementia subtypes.

Finally, although our unbiased deep sequencing approach yielded a substantial number of false-positive variants, despite extensive efforts to identify the most likely true variants, it also resulted in the detection of true-positive variants in a well-known candidate gene causative for frontotemporal dementia with TDP-43 pathology. In our view, future studies may choose between two alternative approaches: 1) targeted deep sequencing of bulk tissue of a large number of candidate genes in one way or another related to the pathophysiology, or 2) single-cell whole genome sequencing generating more reliable data on true-positive variants.

In conclusion, low-level somatic pathogenic variants in the *TARDBP* gene are an underlying genetic cause of non-familial semantic dementia. This phenomenon needs investigation in other cases of semantic dementia, as well as in other early-onset neurodegenerative diseases, for example non-familial frontotemporal dementia with tau pathology. Moreover, in other neurodegenerative diseases, such as Alzheimer's disease or Parkinson's disease, somatic variants may also play a causal or contributing role, and deserve further investigation. Further investigation of somatic variants in known disease genes is warranted, specifically in patients without positive family history and with clearly defined focal neurodegeneration. Our findings have implications for understanding of neurodegenerative disease and the specific role of germline versus somatic variants therein. Also, negative germline variant testing might be insufficient for some diseases, and may require DNA from the appropriate tissue instead to detect somatic variants in order to determine disease causes. Finally, studying the properties of somatic disease-causing genetic variants may reveal novel underlying disease processes and point towards new therapeutic strategies.

***Acknowledgements***

We would like to thank the Netherlands Brain Bank and all donors that have provided the material to perform this research. Several authors of this publication are members of the European Reference Network for Rare Neurological Diseases – Project ID No 739510.

***Data Availability***

All main results are available through the (supplemental) tables in the manuscript. Additional data from the raw results are available on request from the authors.

***Funding information***

We would like to thank the funding agencies for this project; Netherlands Organization for Scientific Research (NWO) through the ZonMw Memorabel grants (project #733050811, #733050816), the Alzheimer Nederland organization, the Gieskes-Strijbis Foundation, AriSLA (project PathensTDP) and the Beneficientia Stiftung from Lichtenstein.

## References

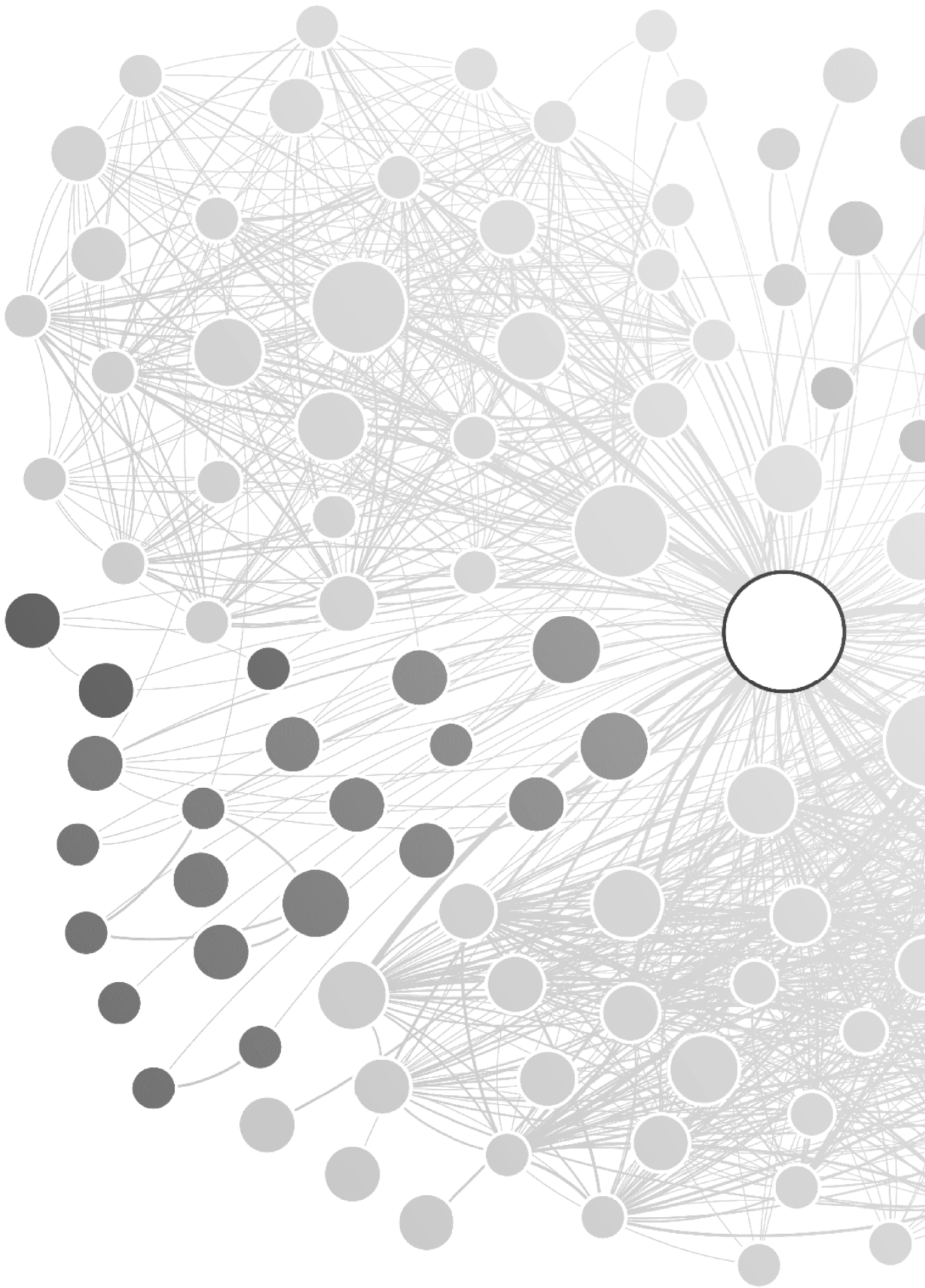
- Barmada SJ, Finkbeiner S. Pathogenic TARDBP mutations in amyotrophic lateral sclerosis and frontotemporal dementia: disease-associated pathways. *Rev Neurosci* 2010; 21(4): 251-72.
- Beck JA, Poulter M, Campbell TA, Uphill JB, Adamson G, Geddes JF, *et al.* Somatic and germline mosaicism in sporadic early-onset Alzheimer's disease. *Hum Mol Genet* 2004; 13(12): 1219-24.
- Borroni B, Archetti S, Del Bo R, Papetti A, Buratti E, Bonvicini C, *et al.* TARDBP mutations in frontotemporal lobar degeneration: frequency, clinical features, and disease course. *Rejuvenation Res* 2010; 13(5): 509-17.
- Brady DR, Mufson EJ. Parvalbumin-immunoreactive neurons in the hippocampal formation of Alzheimer's diseased brain. *Neuroscience* 1997; 80(4): 1113-25.
- Buratti E, Baralle FE. Characterization and functional implications of the RNA binding properties of nuclear factor TDP-43, a novel splicing regulator of CFTR exon 9. *J Biol Chem* 2001; 276(39): 36337-43.
- Bushman DM, Kaeser GE, Siddoway B, Westra JW, Rivera RR, Rehen SK, *et al.* Genomic mosaicism with increased amyloid precursor protein (APP) gene copy number in single neurons from sporadic Alzheimer's disease brains. *Elife* 2015; 4.
- Caroppo P, Camuzat A, Guillot-Noel L, Thomas-Anterion C, Couratier P, Wong TH, *et al.* Defining the spectrum of frontotemporal dementias associated with TARDBP mutations. *Neurol Genet* 2016; 2(3): e80.
- Chang CK, Wu TH, Wu CY, Chiang MH, Toh EK, Hsu YC, *et al.* The N-terminus of TDP-43 promotes its oligomerization and enhances DNA binding affinity. *Biochem Biophys Res Commun* 2012; 425(2): 219-24.
- Clarimon J, Moreno-Grau S, Cervera-Carles L, Dols-Icardo O, Sanchez-Juan P, Ruiz A. Genetic architecture of neurodegenerative dementias. *Neuropharmacology* 2020; 168: 108014.
- Coxhead J, Kurzawa-Akanbi M, Hussain R, Pyle A, Chinnery P, Hudson G. Somatic mtDNA variation is an important component of Parkinson's disease. *Neurobiol Aging* 2016; 38: 217 e1- e6.
- D'Ambrogio A, Buratti E, Stuani C, Guarnaccia C, Romano M, Ayala YM, *et al.* Functional mapping of the interaction between TDP-43 and hnRNP A2 in vivo. *Nucleic Acids Res* 2009; 37(12): 4116-26.
- Davies RR, Hodges JR, Kril JJ, Patterson K, Halliday GM, Xuereb JH. The pathological basis of semantic dementia. *Brain* 2005; 128(Pt 9): 1984-95.
- Ferrari R, Manzoni C, Hardy J. Genetics and molecular mechanisms of frontotemporal lobar degeneration: an update and future avenues. *Neurobiol Aging* 2019; 78: 98-110.
- Gonzalez-Sanchez M, Puertas-Martin V, Esteban-Perez J, Garcia-Redondo A, Borrego-Hernandez D, Mendez-Guerrero A, *et al.* TARDBP mutation associated with semantic variant primary progressive aphasia, case report and review of the literature. *Neurocase* 2018; 24(5-6): 301-5.
- Greaves CV, Rohrer JD. An update on genetic frontotemporal dementia. *J Neurol* 2019; 266(8): 2075-86.
- Hijazi S, Heistek TS, Scheltens P, Neumann U, Shimshek DR, Mansvelder HD, *et al.* Early restoration of parvalbumin interneuron activity prevents memory loss and network hyperexcitability in a mouse model of Alzheimer's disease. *Mol Psychiatry* 2019.
- Hodges JR, Mitchell J, Dawson K, Spillantini MG, Xuereb JH, McMonagle P, *et al.* Semantic dementia: demography, familial factors and survival in a consecutive series of 100 cases. *Brain* 2010; 133(Pt 1): 300-6.
- Hodges JR, Patterson K, Oxbury S, Funnell E. Semantic dementia. Progressive fluent aphasia with temporal lobe atrophy. *Brain* 1992; 115 ( Pt 6): 1783-806.

- Hoekstra JG, Hipp MJ, Montine TJ, Kennedy SR. Mitochondrial DNA mutations increase in early stage Alzheimer disease and are inconsistent with oxidative damage. *Ann Neurol* 2016; 80(2): 301-6.
- Hu WF, Chahrour MH, Walsh CA. The diverse genetic landscape of neurodevelopmental disorders. *Annu Rev Genomics Hum Genet* 2014; 15: 195-213.
- Iaccarino HF, Singer AC, Martorell AJ, Rudenko A, Gao F, Gillingham TZ, *et al.* Gamma frequency entrainment attenuates amyloid load and modifies microglia. *Nature* 2016; 540(7632): 230-5.
- Igaz LM, Kwong LK, Lee EB, Chen-Plotkin A, Swanson E, Unger T, *et al.* Dysregulation of the ALS-associated gene TDP-43 leads to neuronal death and degeneration in mice. *J Clin Invest* 2011; 121(2): 726-38.
- Ihara R, Matsukawa K, Nagata Y, Kunugi H, Tsuji S, Chihara T, *et al.* RNA binding mediates neurotoxicity in the transgenic *Drosophila* model of TDP-43 proteinopathy. *Hum Mol Genet* 2013; 22(22): 4474-84.
- Irish M, Addis DR, Hodges JR, Piguat O. Considering the role of semantic memory in episodic future thinking: evidence from semantic dementia. *Brain* 2012; 135(Pt 7): 2178-91.
- Jamuar SS, Lam AT, Kircher M, D’Gama AM, Wang J, Barry BJ, *et al.* Somatic mutations in cerebral cortical malformations. *N Engl J Med* 2014; 371(8): 733-43.
- Kim J, Kim KM, Noh JH, Yoon JH, Abdelmohsen K, Gorospe M. Long noncoding RNAs in diseases of aging. *Biochim Biophys Acta* 2016; 1859(1): 209-21.
- Kovacs GG, Adle-Biassette H, Milenkovic I, Cipriani S, van Scheppingen J, Aronica E. Linking pathways in the developing and aging brain with neurodegeneration. *Neuroscience* 2014; 269: 152-72.
- Kumfor F, Landin-Romero R, Devenney E, Hutchings R, Grasso R, Hodges JR, *et al.* On the right side? A longitudinal study of left- versus right-lateralized semantic dementia. *Brain* 2016; 139(Pt 3): 986-98.
- Landin-Romero R, Tan R, Hodges JR, Kumfor F. An update on semantic dementia: genetics, imaging, and pathology. *Alzheimers Res Ther* 2016; 8(1): 52.
- Lattante S, Rouleau GA, Kabashi E. TARDBP and FUS mutations associated with amyotrophic lateral sclerosis: summary and update. *Hum Mutat* 2013; 34(6): 812-26.
- Lee JH, Huynh M, Silhavy JL, Kim S, Dixon-Salazar T, Heiberg A, *et al.* De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. *Nat Genet* 2012; 44(8): 941-5.
- Lee MH, Siddoway B, Kaeser GE, Segota I, Rivera R, Romanow WJ, *et al.* Somatic APP gene recombination in Alzheimer’s disease and normal neurons. *Nature* 2018; 563(7733): 639-45.
- Leija-Salazar M, Piette C, Proukakis C. Review: Somatic mutations in neurodegeneration. *Neuropathol Appl Neurobiol* 2018; 44(3): 267-85.
- Leyton CE, Britton AK, Hodges JR, Halliday GM, Kril JJ. Distinctive pathological mechanisms involved in primary progressive aphasia. *Neurobiol Aging* 2016; 38: 82-92.
- Lim JS, Kim WI, Kang HC, Kim SH, Park AH, Park EK, *et al.* Brain somatic mutations in MTOR cause focal cortical dysplasia type II leading to intractable epilepsy. *Nat Med* 2015; 21(4): 395-400.
- Lin MT, Cantuti-Castelvetri I, Zheng K, Jackson KE, Tan YB, Arzberger T, *et al.* Somatic mitochondrial DNA mutations in early Parkinson and incidental Lewy body disease. *Ann Neurol* 2012; 71(6): 850-4.
- Lodato MA, Rodin RE, Bohrsen CL, Coulter ME, Barton AR, Kwon M, *et al.* Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* 2018; 359(6375): 555-9.
- Lodato MA, Walsh CA. Genome aging: somatic mutation in the brain links age-related decline with disease and nominates pathogenic mechanisms. *Hum Mol Genet* 2019; 28(R2): R197-R206.
- Lodato MA, Woodworth MB, Lee S, Evrony GD, Mehta BK, Karger A, *et al.* Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* 2015; 350(6256): 94-8.

- Mackenzie IR, Neumann M. Reappraisal of TDP-43 pathology in FTLD-U subtypes. *Acta Neuropathol* 2017; 134(1): 79-96.
- Mackenzie IR, Neumann M, Baborie A, Sampathu DM, Du Plessis D, Jaros E, *et al.* A harmonized classification system for FTLD-TDP pathology. *Acta Neuropathol* 2011; 122(1): 111-3.
- Mesulam MM, Rogalski EJ, Wieneke C, Hurley RS, Geula C, Bigio EH, *et al.* Primary progressive aphasia and the evolving neurology of the language network. *Nat Rev Neurol* 2014; 10(10): 554-69.
- Miller ZA, Mandelli ML, Rankin KP, Henry ML, Babiak MC, Frazier DT, *et al.* Handedness and language learning disability differentially distribute in progressive aphasia variants. *Brain* 2013; 136(Pt 11): 3461-73.
- Mirzaa GM, Enyedi L, Parsons G, Collins S, Medne L, Adams C, *et al.* Congenital microcephaly and chorioretinopathy due to de novo heterozygous KIF11 mutations: five novel mutations and review of the literature. *Am J Med Genet A* 2014; 164A(11): 2879-86.
- Mokretar K, Pease D, Taanman JW, Soenmez A, Ejaz A, Lashley T, *et al.* Somatic copy number gains of alpha-synuclein (SNCA) in Parkinson's disease and multiple system atrophy brains. *Brain* 2018; 141(8): 2419-31.
- Mompean M, Romano V, Pantoja-Uceda D, Stuani C, Baralle FE, Buratti E, *et al.* Point mutations in the N-terminal domain of transactive response DNA-binding protein 43 kDa (TDP-43) compromise its stability, dimerization, and functions. *J Biol Chem* 2017; 292(28): 11992-2006.
- Mummery CJ, Patterson K, Price CJ, Ashburner J, Frackowiak RS, Hodges JR. A voxel-based morphometry study of semantic dementia: relationship between temporal lobe atrophy and semantic memory. *Ann Neurol* 2000; 47(1): 36-45.
- Neumann M, Mackenzie IRA. Review: Neuropathology of non-tau frontotemporal lobar degeneration. *Neuropathol Appl Neurobiol* 2019; 45(1): 19-40.
- Nicolas G, Acuna-Hidalgo R, Keogh MJ, Quenez O, Steehouwer M, Lelieveld S, *et al.* Somatic variants in autosomal dominant genes are a rare cause of sporadic Alzheimer's disease. *Alzheimers Dement* 2018; 14(12): 1632-9.
- Palomero-Gallagher N, Zilles K. Cortical layers: Cyto-, myelo-, receptor- and synaptic architecture in human cortical areas. *Neuroimage* 2019; 197: 716-41.
- Park JS, Lee J, Jung ES, Kim MH, Kim IB, Son H, *et al.* Brain somatic mutations observed in Alzheimer's disease associated with aging and dysregulation of tau phosphorylation. *Nat Commun* 2019; 10(1): 3090.
- Poduri A, Evrony GD, Cai X, Walsh CA. Somatic mutation, genomic variation, and neurological disease. *Science* 2013; 341(6141): 1237758.
- Proukakis C, Shoaee M, Morris J, Brier T, Kara E, Sheerin UM, *et al.* Analysis of Parkinson's disease brain-derived DNA for alpha-synuclein coding somatic mutations. *Mov Disord* 2014; 29(8): 1060-4.
- Qin H, Lim LZ, Wei Y, Song J. TDP-43 N terminus encodes a novel ubiquitin-like fold and its unfolded form in equilibrium that can be shifted by binding to ssDNA. *Proc Natl Acad Sci U S A* 2014; 111(52): 18619-24.
- Rogalski EJ, Rademaker A, Wieneke C, Bigio EH, Weintraub S, Mesulam MM. Association between the prevalence of learning disabilities and primary progressive aphasia. *JAMA Neurol* 2014; 71(12): 1576-7.
- Romano V, Quadri Z, Baralle FE, Buratti E. The structural integrity of TDP-43 N-terminus is required for efficient aggregate entrapment and consequent loss of protein function. *Prion* 2015; 9(1): 1-9.
- Sala Frigerio C, Lau P, Troakes C, Deramecourt V, Gele P, Van Loo P, *et al.* On the identification of low allele frequency mosaic mutations in the brains of Alzheimer's disease patients. *Alzheimers Dement* 2015; 11(11): 1265-76.

- Sasaguri H, Chew J, Xu YF, Gendron TF, Garrett A, Lee CW, *et al.* The extreme N-terminus of TDP-43 mediates the cytoplasmic aggregation of TDP-43 and associated toxicity in vivo. *Brain Res* 2016; 1647: 57-64.
- Seelaar H, Kamphorst W, Rosso SM, Azmani A, Masdjedi R, de Koning I, *et al.* Distinct genetic forms of frontotemporal dementia. *Neurology* 2008; 71(16): 1220-6.
- Seelaar H, Rohrer JD, Pijnenburg YA, Fox NC, van Swieten JC. Clinical, genetic and pathological heterogeneity of frontotemporal dementia: a review. *J Neurol Neurosurg Psychiatry* 2011; 82(5): 476-86.
- Stiles J, Jernigan TL. The basics of brain development. *Neuropsychol Rev* 2010; 20(4): 327-48.
- Takata A, Ionita-Laza I, Gogos JA, Xu B, Karayiorgou M. De Novo Synonymous Mutations in Regulatory Elements Contribute to the Genetic Etiology of Autism and Schizophrenia. *Neuron* 2016; 89(5): 940-7.
- van Rooij JGJ, Jhamai M, Arp PP, Nouwens SCA, Verkerk M, Hofman A, *et al.* Population-specific genetic variation in large sequencing data sets: why more data is still better. *Eur J Hum Genet* 2017; 25(10): 1173-5.
- Veltman JA, Brunner HG. De novo mutations in human genetic disease. *Nat Rev Genet* 2012; 13(8): 565-75.
- Verret L, Mann EO, Hang GB, Barth AM, Cobos I, Ho K, *et al.* Inhibitory interneuron deficit links altered network activity and cognitive dysfunction in Alzheimer model. *Cell* 2012; 149(3): 708-21.
- Wang W, Wang L, Lu J, Siedlak SL, Fujioka H, Liang J, *et al.* The inhibition of TDP-43 mitochondrial localization blocks its neuronal toxicity. *Nat Med* 2016; 22(8): 869-78.
- Wei W, Keogh MJ, Aryaman J, Golder Z, Kullar PJ, Wilson I, *et al.* Frequency and signature of somatic variants in 1461 human brain exomes. *Genet Med* 2019; 21(4): 904-12.
- Weskamp K, Barmada SJ. TDP43 and RNA instability in amyotrophic lateral sclerosis. *Brain Res* 2018; 1693(Pt A): 67-74.
- Wiseman FK, Al-Janabi T, Hardy J, Karmiloff-Smith A, Nizetic D, Tybulewicz VL, *et al.* A genetic cause of Alzheimer disease: mechanistic insights from Down syndrome. *Nat Rev Neurosci* 2015; 16(9): 564-74.
- Zhang YJ, Caulfield T, Xu YF, Gendron TF, Hubbard J, Stetler C, *et al.* The dual functions of the extreme N-terminus of TDP-43 in regulating its biological activity and inclusion formation. *Hum Mol Genet* 2013; 22(15): 3112-22.
- Zilles K, Palomero-Gallagher N, Amunts K. Development of cortical folding during evolution and ontogeny. *Trends Neurosci* 2013; 36(5): 275-84.









# Chapter 5

## General discussion



## Introduction

The introduction of next-generation sequencing around the year 2005 led to an explosion of genomic data (1-3). With further decreasing costs and increasing familiarity with sequencing, the list of applications grows (2-5). Next-generation sequencing (mostly whole exome sequencing; WES) is now commonplace in clinical practice, and genetic testing (mostly with array technology) made its way into society through direct-to-consumer companies (6-9). The number of DNA sequenced samples worldwide runs into the millions, permitting insight into the frequencies of common and rare variants in all genes throughout the human population at a resolution which was previously unavailable (1, 10-13). This shift to “bigger data” increases the strain on data handling and analysis (14). This strain will further increase with the application of dynamic data (i.e.; data changing over tissues or time) measures in cells or tissues, such as transcriptome or methylome sequencing, potentially multiplying the number of meaningful datasets generated per individual (15, 16). In this chapter, developments in NGS methodology and the application on neurodegenerative disorders over the last 5 years are discussed and how the work in this thesis has contributed. Finally, expected developments over the next 5-10 years will be discussed, and how these may further change the direction of the field of genomics in medicine.

### ***Novel genetic findings in dementia***

One of the first applications of next-generation sequencing was found in family studies, driven by the many years of successful linkage analysis of Mendelian disorders (based on co-segregation of the disease with DNA markers). Rather than having to select and test the correct candidate genes beforehand, requiring prior knowledge, family-based sequencing have yielded dozens of potentially clinically relevant variants per family based on co-segregation (17-20). This has requested a shift in analysis method; instead of looking for a specific genetic variant (hypothesis driven testing), the analysis starts unbiased and evaluates all identified genetic variants or all genetic variants in linkage regions and their potential to cause the disease (17, 18, 20). Overall, this approach aids identification of novel disease-causing genes, which in turn provides insight into the biology behind the disease (17, 18, 20).

Over the last 5-10 years, work by the author and co-authors has utilized NGS in families with dementia to identify novel, causal genetic variants (17-20). Previous work by Wong *et al.* identified a variant in the *PRKAR1B* gene causing a unique neuropathological phenotype in one large family with dementia/parkinsonism (17). The variant was absent from population control databases and from disease control groups, and the neuropathology was hallmarked by intermediate neurofilament neuronal inclusions stained positive for the *PRKAR1B* protein (17). Similar work combining WES of eight families burdened with Alzheimer's Disease (AD) (represented in this thesis in chapter 2.3) identified rare variants in the *EIF2AK3* gene as the cause in two families (18). Further analysis in a large group of unrelated AD patients demonstrated increased burden of rare *EIF2AK3* variants, suggesting a genetic contribution of *EIF2AK3* to AD also outside these families (18). Using WES data of a large group of unrelated patients, rare variants in the *SORL1* gene were also found to be more abundant in cases than in controls (19). Extensive analysis of these variants demonstrated an increased risk for AD, low frequency in the control population and higher predicted damaging scores

in certain domains of the gene (19). Additionally, Mol *et al.* identified a causal variant in the *STUB1* gene in a single large family with ataxia, parkinsonism and cognitive decline. Although variants in this gene were previously recessively associated with spinocerebellar ataxia, the *STUB1* variant in our family supported a dominant mode of inheritance as genetic cause of the disease (21).

In total, in 25% of patients with frontotemporal dementia (FTD) and 10% of patients with AD the causal genetic variants are known. This leaves an estimated 25% of genetically unresolved FTD and 60% of unresolved AD. This difference seems reflected also by the presentation of each disease; FTD seems more often familial and nearly all identified genetic causes are mendelian pathogenic variants. In contrast, only 1-2% of AD is caused by known familial factors in *APP*, *PSEN1*, *PSEN2* and *SORL1*, whereas genetic risk factors seem to play a larger role, such as the *APOE* e4 locus. By studying families or series of cases with dementia but without known gene defect, additional genetic defects that might cause FTD or AD can be identified. Knowing in which genes these variants contribute to neurodegeneration yields biological insights (22, 23). Specifically, these results point towards pathways in which dysfunction causes neurodegenerative disease susceptibility, hinting towards underlying causal biology (22, 23). These small insights contribute to the understanding of the disease's biology and ultimately aid in preventing, halting or treating the disease.

Sharing data and collaborating with other research groups is the essential next step to solve genetic causes in many families with fewer cases or even only a single proband to study. Of specific note are large-scale studies combining multiple existing sequencing datasets. One large collaboration by Janssen *et al.* combined PD datasets and demonstrated that rare variants in lysosomal storage genes increase PD risk (24, 25). In addition to this finding, this was also one of the first large studies to group genes by function in a gene burden analysis (24, 25). A similar large effort by Pottier *et al.* used whole genome sequencing (WGS) to identify novel rare variants in *DPP6*, *UNC13A* and *HLA-DQA2* that increase risk for FTD (26, 27). A large collaboration driven by Holstege *et al.* aims to combine all European AD sequencing datasets (19). These collaborations are well-suited to identify genetic risk factors of incomplete penetrance, which are often missed by family-based studies. They also request the use of standardized workflows and facilitate replication between research groups. For example, the burden testing approach, coupled by pathway- or otherwise biological- gene grouping, will allow further ascertainment of the rare variant contribution for these diseases. This in turn sheds light on the underlying biology, as we learn which dysfunctions appears to contribute to the disease. This understanding is needed for the next step; interfering with the disease process and slow or prevent the disease course.

### ***Population DNA sequencing demonstrates incomplete penetrance***

One application of NGS is the generation of large-scale sequencing databases of many unaffected and unrelated individuals (10, 28, 29). Of particular interest in this context is applying NGS in longitudinal population-based cohort studies with very rich phenotype data on health and disease. One such dataset was generated in the Rotterdam Study, as is detailed in chapter 2.1. A main addition of these datasets is that it allows to investigate how many healthy individuals are also carriers of variants in disease-causing genes. This helps to determine the likelihood that when a person is identified as carrier of a specific gene

defect, it actually results in the disease associated with that gene (30). This likelihood can be denoted as the prevalence of disease among all variant carriers in (part of) that gene, a property dubbed the penetrance of a genetic variant. In general, variants with higher penetrance are more likely to be pathogenic or disease-causing. So far, pathogenicity of a genetic variant was mostly determined on the frequency of the variant in cases, combined with interpretation of the expected molecular impact of the variant on the gene's function (31). Collections of "proven" pathogenic variants are recorded in clinical genetic databases by each hospital, and efforts are ongoing to centralize this information, for example in the ClinVar database. Predictions of the expected impact of a variant on the gene are now automated through tools like SIFT, PolyPhen or CADD. Yet, identifying a variant as being truly pathogenic is non-trivial and can be done based on several criteria (2, 31).

Chapter 2.1 of this thesis describes the generation of one of these population-based sequencing databases (32). Furthermore, in chapter 2.3 this database of WES in 2,628 subjects of the Rotterdam Study is analyzed by screening 59 specific genes for pathogenic variants. These 59 genes are recommended by the American College of Medical Genetics (ACMG) for screening as a secondary result of clinical WES, as variants therein are proven to cause specific diseases and those diseases can be treated or prevented. This study demonstrates that 1% of the Rotterdam Study population (n=26) carries potentially disease-causing variants in one of these 59 genes, according to the publicly available clinical genetic database ClinVar and HGMD. Under the recommendations of the ACMG, carriers of these "actionable" variants are eligible to receive genetic counseling with information on this genetic finding and may be subjected to screening or additional clinical follow-up. However, we demonstrated that only 13% of these carriers have experienced symptoms that might be related to their variant. Thus, 87% of carriers appeared unaffected, suggesting that the penetrance of these variants is relatively low. This could mean that the clinical reporting of such variants to their carriers might place burden on those patients which is not always warranted. This observation is supported by similar data coming out of several collaborations investigating the penetrance of variants in specific disease-causing genes. One of the first of these studies was performed by Minikel *et al.* and regarded variants in the *PRNP* gene and their penetrance in causing prion's disease (33). The authors collected over 60,000 population controls to demonstrate that the prevalence of "pathogenic" *PRNP* variants was 30x higher than the prevalence of prion's disease (33), again indicating a much lower penetrance as expected for many of such variants. They investigated each *PRNP* variant and provided penetrance estimates ranging from 0.1% to 100% per variant on the lifetime risk of prion's disease. The population database generated by this effort (the Exome Aggregation Consortium; ExAC) is open access for other researchers to inspect variant prevalence of their variant/gene of interest (29). A similar observation was made for *ASXL1* and other intellectual disability genes by Ropers *et al.* (34). Over the last years, the database generated in chapter 2.1 of WES data in the Rotterdam Study was used for various efforts to assess association of variants or genes with disease, for example to validate the relation of variants in the *SOLR1* or *EIF2AK3* genes with AD (32).

Similarly, the penetrance of pathogenic variants for neurodegenerative diseases has been re-evaluated recently. For example, van der Lee *et al.* used six large datasets to investigate the association of previously reported pathogenic variants in *PLD3* with AD (35). However, this

association between *PLD3* and AD could not be replicated in these cohorts. The same was true for the role of variants in *TMEM230* in causing PD, which could not be replicated by Giri *et al.* (36). These efforts help distinguish true findings from false positives, and availability of large-scale sequencing datasets is necessary to perform these validation studies.

### ***Transcriptome sequencing growing in utility***

One approach following the sequencing of DNA is the sequencing of RNA, so-called transcriptomics. Transcriptomic studies often complement genetic studies, by investigating how the transcriptome changes in patients and how disease-causing genes play a role (37-39). As the transcriptome is “dynamic”, i.e.; the measured gene expression values change over time, tissues or disease status, this type of data is called dynamic data. The requirement to extract RNA from disease-affected tissue creates challenges for large-scale studies, for example because the tissue cannot be safely sampled, the sampling is invasive or labor-intensive or because when the patient is identified, the affected tissue is already too far damaged. Concurrently, the approach of RNA collection and analysis varies between studies (40), partly due to different research questions requiring alternate approaches, and the lack of standardized methods (40). These factors challenge reproducible analysis of dynamic sequencing data, which is true also for dynamic genomics data generated with other methods such as DNA methylation data by arrays. Therefore, efforts aiming to provide standardized analysis will facilitate larger-scale studies in producing more robust results (41).

In chapter 3.1, commonly used analysis methods for RNA-sequencing and DNA methylation arrays are compared. The results showed that some analysis options yield highly similar results across different datasets, while other options strongly influence the results (40). These methodology-induced changes may confound any actual biological changes that are of interest in any given study. These benchmarking efforts help in standardizing our analysis of dynamic data, allowing for more robust analysis pipelines, as exist for genetic data (42, 43). Perhaps the most relevant observation in this effort were the large differences in results between the four cohorts studied, even when methods were as harmonized as possible. This observation illustrates the requirement for replication and validation when analyzing dynamic genomic data, perhaps even more so than for genetic studies. The study in chapter 3.2 applies transcriptome sequencing on hippocampus tissue from AD patients and compares this with hippocampus of age- and sex-matched non-demented controls (44). The results show, in accordance with previous scientific literature, an enormous shift in the transcriptome in the brains of AD patients compared to controls (45-47). Most of these changes are probably the consequence of neurodegenerative processes, in line with other neurodegenerative studies examining dynamic data (45-47). Unfortunately, our data did not show altered expression of most of the genes in which genetic variants cause AD (48). As the genetic risk variants for AD must preclude the disease, this result could suggest that whichever biological dysfunction(s) these variants caused took place earlier in the disease process and were no longer detectable at the end-stage of disease. Another option is that these genes play their role in other tissues than we studied, or that the effect cannot be observed on the RNA expression level. Because transcriptome data is dynamic (i.e.; changes over time, tissues and disease status), the fact that the disease itself affects the dynamic genomic data is a common challenge faced by dynamic data studies; the difficulty to separate cause from consequence (44-47). This challenge is worse when tissues are derived in later stages of the

disease or when the impact of the disease is larger on the tissue in which it occurs. Therefore, this issue is specifically complicated in neurodegenerative diseases, as nearly all accessible brain tissues are derived post-mortem, after many years of neurodegeneration (49). This issue complicates the interpretation of the study results, and therefore our understanding of neurodegenerative disease biology. This in turn challenges the identification of treatment options capable of acting on causal pathways of the disease. Part of these challenges can be addressed by studying cell and animal models of earlier phases of AD, which permits to test interactions or specific hypotheses in live cells, and correlate these findings to post-mortem human dynamic genomic studies, hopefully illuminating enough pieces to uncover the whole puzzle. A few specific applications of dynamic genomic data might further our understanding of neurodegenerative disease. Firstly, we currently do not understand the causal mode of action of genetic variants in many of the identified genes. Combining dynamic genomic data of cellular models with and without these variants with protein-protein interaction modeling derived from post-mortem human dynamic genomic studies might create biological insight into the causal changes underlying the disease. The next step could be to carefully target these processes in cellular or animal models and see if this alleviates the disease onset or progression, ultimately leading to some form of intervention. Secondly, dynamic genomic data from either post-mortem brain or fluid (blood, CSF) studies might be used to understand heterogeneity between neurodegenerative patients, providing insight into the extensive clinical and pathological range with which patients may present. When an intervention or treatment strategy becomes available, these molecular subtypes would be useful to evaluate which patients might benefit from such a treatment. Already such dynamic genomic studies from post-mortem human brain tissue are being performed to map and understand the molecular differences between patients (with similar clinical or pathological presentation) with different genetic causes, and these findings are relayed to understand the impact and function of the causal gene, as well as to point towards other likely candidate genes for further genetic screening.

### ***Somatic DNA sequencing reveals further complexity***

Tissue-specific DNA sequencing is a novel NGS development, aimed at identifying somatic DNA variants (50-52). The clinical and biological study and relevance of somatic DNA variants has long been limited to the field of cancer, where they lead to clonal outgrowth of tumor cells (53). However, recent evidence showed that germline de-novo (present in germline of offspring, but not parents) variants causing specific diseases, for example intellectual disability (54). De-novo variants are acquired as somatic variants in one of the reproductive cell lineages of the parents (oocytes and sperm) and transmitted as a de-novo germline variant in the offspring (54). These de-novo variants are then present in all tissues of the offspring, as these cells are further derived from these germline cells with this new variant. From then onward, this variant can no longer be distinguished from other inherited germline variants, and can in turn be transmitted to next generations. If a somatic variant arises later in the development, for example when a cell divides during development of the brain, this variant will be present only in the subsequent cells, and thus only in that tissue, for example a specific lobe of the brain (54). Theoretically, one could carry genetic variants known to cause neurodegenerative disease, but only in subsets of cells in our brain (55, 56). Such a “somatic” variant could nevertheless cause a neurodegenerative disease, whilst escaping detection in blood-based germline DNA sequencing (55, 56). This variant

would also not be transmitted to offspring, thus prohibiting familial aggregation and escape detection as a typical Mendelian disease. In addition, the moment at which the somatic variant was acquired during brain development determined the amount of cells and tissues affected and may subsequently influence its clinical and pathological presentation (50, 56).

In chapter 4 of this thesis, the contribution of somatic variants in the brain of Semantic Dementia (SD) patients was investigated. SD is hallmarked by local atrophy of the temporal lobe and pathology stained positive for the *TDP-43* protein, which is also seen in other forms of FTD (55-63). However, the atrophy is uniquely localized and the specific type of *TDP-43* pathology is unique to SD brains (58, 59). Most importantly, no genetic variants are known to cause SD, and familial aggregation is rare (57, 63). Thus, these characteristics of SD (non-heritable, focal onset, relatively homogeneous pathological presentation) meet the expected features for a somatic variant as cause of the disease. Although somatic genetic variants have been shown to cause neurodevelopmental disorders, our study was one of the first to suggest and study such a cause for neurodegenerative disorders. The novelty in this approach is that genetic variants can contribute to disease in a different way than usually studied, and that tissue-specific DNA variations must also be considered when investigating disease genetics. Several lines of evidence indicate that we indeed identified such somatic variants in one gene underlying SD. First, sequencing DNA of the medial temporal lobe and/or dentate gyrus brain regions of 16 patients and comparing it to sequenced DNA of blood of the same patients, revealed somatic brain variants in the *TARDBP* gene in two SD patients. Secondly, germline variants in this gene, albeit in a different domain, are described to cause ALS or behavioral-variant FTD (64). Thirdly, functional assays demonstrated that both variants disrupt the function and localization of *TARDBP*'s protein and unique *TDP-43* pathology has been identified for SD (65-67). That we did not identify similar somatic variants in the other patients might mean that additional variants are to be found, or that somatic variants are the cause only in a subset of patients. This study is one of the first to investigate somatic variants in a neurodegenerative disorder, and the first to demonstrate variants in a known neurodegenerative disease-causing gene somatically in the brains of patients. In addition to identifying the possible cause for these two SD patients, upcoming research like this illustrates a paradigm shift, outside the field of cancer, where the genomic DNA sequence itself can no longer be seen as a stable information entity, Genomic DNA thus falls in the category of dynamic data next to, for example, epigenetic modifications such as DNA methylation.



## Future directions

The discussion in these paragraphs looks ahead to future directions in sequencing-based methods and applications and how these may lead to applications in a clinical setting. NGS has found many diverse applications in biology and medicine, ranging from cancer genome sequencing to family-based analysis in Mendelian disease, RNA and single cell sequencing, down to microbiome sequencing and archeological genome sequencing. Much effort so far has been spent on generating NGS data on patients with Mendelian disorders, in an attempt to understand the cause of their disease. NGS-based applications will likely also develop into tools that might be useful in predicting and subsequently preventing diseases, in the form of personalized medicine. For example, by early genetic screening for breast cancer mutations, breast cancer polygenic risk and genetic variants causing adverse drug responses or food intolerance. Regarding the topics discussed in this thesis, three specific expected clinically-oriented developments will be discussed; individualized preventive genome sequencing, dynamic genomic data sequencing of patients, and somatic variant sequencing.

### ***Individualized preventive genome sequencing***

In parallel to the continuous growth of genetic knowledge from scientific research, a growth in the expectation from society in the application of genetic information can also be observed. Individuals appreciate the benefit of DNA testing in disease prevention, evaluating lifestyle factors and other medically useable information (pharmacogenomics, blood typing, etc.) (6, 8, 68-70). This originates from an individual's desire to be autonomous and take initiative regarding their own health (care) (71-73). This development reflects that genetic research yields directly relevant clinical results, and that this output is recognized outside of a research setting (72-74). As the costs of sequencing continues to decrease and our ability to handle large volumes of data improves, a point will be reached where pre-emptively sequencing and storing the genome sequence of every (consenting) adult becomes commonplace (73, 75, 76). An individual's genome can be utilized for personalized clinical trajectories. For example, one can stratify individuals (early on) by increased or decreased genetic risk in population screening programs, and/or one can test for susceptibility of treatable/preventable diseases and schedule regular clinical check-ins, accompanied by advise on diet or other modifiable lifestyle factors based on genetic susceptibility for traits as obesity or addiction, one can optimize therapeutic treatment based on genetic variants influencing drug metabolism, or one can better select subjects most suitable for organ transplant and better monitor rejection (6, 8, 9, 69, 71, 73, 74). These developments have many aspects to be discussed between the related groups; patients, doctors, counsellors, policy makers etc., but are likely to shift our approach from indication-driven health care (when disease has occurred already) to data-driven health care (to prevent or delay diseases from occurring). While many recognize socio-economic and health-economic advantages of this shift, the interest and preference of the individual patient and citizen should weigh heavily (71, 72, 77). Currently, pre-emptive genome sequencing moves towards the clinic, in the form of clinical trials. Two well-known trials are MedSeq and BabySeq, in which participants are randomized to receive standard care with or without extensive genetic screening and return of possibly relevant results (73, 78). The first results show that sequencing uncovers previously unknown disease risk and can permit early disease detection (79, 80). However, clinically relevant variants are also observed in participants without apparent phenotype,

and screening may be counter-productive in those participants (79, 80). Additional (laboratory) tests are often needed to validate pathogenicity in variant carriers, which may be invasive and/or expensive. Additionally, the patient might experience concern due to their genetic results, possibly unwarranted if not all reported carriers experience a pathogenic consequence (79-81). These first studies show promise, but also that more investigation is required to optimize the workflow of pre-emptive clinical genetic testing.

### ***Dynamic genomic data of patients***

Dynamic genomic data can involve assessing one's DNA methylation profile, one's RNA expression profile, proteomics profile, mitochondrial profile, metabolomics profile etc. Such an individualized genomics analysis has not yet been implemented in the clinic, partly caused by the requirement of collecting specific tissue, which is invasive, and in part because the causal relation between these profiles and the disease is often not clearly established (82-84). Therefore, for tissues that are easy to collect, such as blood or cerebral spinal fluid, dynamic genomic data profiles (so far mostly proteomic profiles) have been utilized as disease biomarkers, to identify patients in which disease has already manifested (85-87). Improvements to this development should allow for earlier disease detection, perhaps permitting to halt disease progression in early stages, as well as stratification of subtypes of patients, for example patients that might be responsive to certain treatment from those in which the treatment is unlikely to be beneficial. In general, such biomarkers are developed by measuring and comparing all proteins, genes or other dynamic genomic data units in cases versus controls. From these results the most predictive genes are determined, which are then translated into a targeted assay for clinical use (85-87). Although this approach has been successful, the most efficacious prediction models are those that retain all measurable information, using advanced clustering algorithms or classification through machine learning (88-91). Specifically, machine-learning based methods are highly suitable for classification of patients based on large datasets. Such methods would require an initial large dataset ( $n$  depending on the disease at study, and the required precision of prediction, but containing at least 100 samples per prediction group). Further developments in the field of dynamic genomics data collection and analyses might permit the collection of complete dynamic genomics profiles more robustly and quickly, which could then replace or complement the targeted assays currently used (82, 84). The added benefit of using an untargeted method is that it can re-evaluate each gene and its weights in disease prediction over time, and adjust the biomarker without the need to redesign a targeted assay. Thus, the performance of such an untargeted biomarker assay could improve over time, as data on more patients becomes available for evaluation of the assay. Improvements include increased accuracy in predicting cases from controls, but may also be extended to further classifications, such as prediction of response to treatment or prediction of subclasses of patients. The first applications of these assays are probably in blood-derived dynamic population datasets, of which data already exist. In such datasets the methodological challenges can be addressed and validated by prediction of several diseases measurable and common in these datasets, for example diabetes or coronary heart disease. Main challenges to be addressed include the replicability of such dynamic genomic biomarkers in multiple patient populations through replication and clinical validation of predictive parameters such as sensitivity and specificity. Further technical developments in the processing of blood samples would be permit creation of more specific biomarkers. For example, by isolating only brain-derived cells from whole

blood and inferring changes in those individual cells back to brain-phenotypes. Similar to personalized genome or exome sequencing, pilots to use (recurrent) dynamic genomic data as biomarker are ongoing. In one main example, the integrative personal omics profile (iPOP) project, a single individual received recurrent multi-omics profiling throughout a period of 400 days (92). This included transcriptome sequencing, but also proteomics, metabolomics and auto-antibody screening. The multiple omics datasets provided insights in the onset and progression of two separate infections and a diagnosis of type-2-diabetes. Matching of genome sequencing of this person with transcriptome sequencing data uncovered extensive allele-specific expression, which varied throughout healthy and disease states. This proof-of-principle project demonstrates recurrent dynamic genomic measurements of a single individual to provide insight into health status, and prompted clinical and lifestyle trajectory changes in the subject (92).

### ***Somatic DNA variant sequencing in patients***

Most of our current DNA screening is done on blood, to test for the presence of germline variants. This means that for most diseases, somatic variants are generally not tested. For some patients, their disease could be caused by somatic variants that have arisen during development of the diseased tissue. Evidence for the clinical relevance of these somatic variants outside the field of cancer is growing, mostly in developmental disorders, but also in age-related disorders (93-95). Although it is largely unclear for how many patients, and for which diseases, somatic DNA variants play a causal role, research projects sequencing tissues of patients will provide insight into the scope of these somatic variants in disease. (50-52, 96-99). For at least some diseases, it can be expected that somatic variant testing will be added to the clinical genetic testing, by immediately sequencing DNA of the relevant tissue. Ease of access to the diseased tissue will determine the first diseases in which this type of testing might occur, for example in blood hematopoietic disorders or disorders in which treatment includes removal of affected scarred, fibrotic or degenerative tissue, for example in treatment of certain cardiac disorders. For diseases in which the affected tissue is difficult to ascertain, such as neurodegenerative disorders, targeted tests for somatic variants could be developed. For example, by extracting cell-free DNA or brain-derived vesicles from blood or CSF and testing for specific somatic variants, such as the *TARDBP* variants identified in this thesis. Significant strides would need to be made in technological developments before such tests could be realized. However, for some diseases it might be highly relevant to identify the underlying genetic cause of the disease, for example when treatment options are specific to a genetic defect. If, for example, a treatment was developed to remedy haploinsufficiency of *GRN* or the expanded *C9ORF72* repeat, it would be relevant to know if a patient carried these defects somatically, and could still benefit from such treatment even in the case of a negative germline genetic test based on blood-derived DNA.

A further extension in this direction of testing is the development of single-cell based genetic and genomic methods. These might provide us with insights into the contribution of a specific form of somatic variants; post-mitotic somatic DNA variants. Especially in post-mitotic cells, such as neurons, this class of somatic variants continues to accumulate over time. Recently, the term “genosenium” was introduced, referring to this accumulation of somatic variants in our cells as a source of ageing, which can be studied by single cell sequencing (50). With the number of somatic variants per cell going in to the thousands over an individual’s lifespan,

it seems almost impossible that these do not influence the function and health of cells and tissues (50, 96). In diseases where somatic variants contribute significantly, we may need to adjust our DNA screening to the relevant tissues and age, when possible, effectively turning genome sequence DNA variation into another dynamic datatype. Over the next years more insight into the portion of disease caused by somatic variants can be expected, and for diseases where the relevant tissue is relatively easy to collect, tissue-specific DNA screening may arise.

## **Conclusions**

In summary, the ability to generate large quantity of genetic and dynamic genomic data is growing. With this abundance of data, novel applications and research questions become available and uncover insight into genetics and diseases. However, more data and higher resolution also increases the risk of false findings, and it can be challenging to understand the limits of a dataset in providing answers. Therefore, it becomes even more important to use robust data collection, cleaning and analysis strategies, as well as proper validation and replication. Clinically, the application of genetic and dynamic genomics data holds great potential, but must also be utilized with appropriate reservation to prevent overly data-driven conclusions. In a time where major developments occur constantly, and these developments tend to cross disciplinary boundaries, close communication between clinical and research efforts is required to guide their implementation as efficiently as possible.

## References

1. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74.
2. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285-91.
3. Pettersson E, Lundeberg J, Ahmadian A. Generations of sequencing technologies. *Genomics*. 2009;93(2):105-11.
4. Gauthier J, Vincent AT, Charette SJ, Derome N. A brief history of bioinformatics. *Brief Bioinform*. 2018.
5. Gayon J. From Mendel to epigenetics: History of genetics. *C R Biol*. 2016;339(7-8):225-30.
6. Lu M, Lewis CM, Traylor M. Pharmacogenetic testing through the direct-to-consumer genetic testing company 23andMe. *BMC Med Genomics*. 2017;10(1):47.
7. Schwartz LM, Woloshin S. Medical Marketing in the United States, 1997-2016. *JAMA*. 2019;321(1):80-96.
8. Schaper M, Schicktanz S. Medicine, market and communication: ethical considerations in regard to persuasive communication in direct-to-consumer genetic testing services. *BMC Med Ethics*. 2018;19(1):56.
9. Kayser M. Forensic DNA Phenotyping: Predicting human appearance from crime scene material for investigative purposes. *Forensic Sci Int Genet*. 2015;18:33-48.
10. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-9.
11. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, et al. NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res*. 2014;42(Database issue):D975-9.
12. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol*. 2016;70:214-23.
13. Tsui B, Dow M, Skola D, Carter H. Extracting allelic read counts from 250,000 human sequencing runs in Sequence Read Archive. *Pac Symp Biocomput*. 2019;24:196-207.
14. Cook CE, Bergman MT, Finn RD, Cochrane G, Birney E, Apweiler R. The European Bioinformatics Institute in 2016: Data growth and integration. *Nucleic Acids Res*. 2016;44(D1):D20-6.
15. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun*. 2017;8:14049.
16. Consortium GT. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45(6):580-5.
17. Wong TH, Chiu WZ, Breedveld GJ, Li KW, Verkerk AJ, Hondius D, et al. PRKAR1B mutation associated with a new neurodegenerative disorder with unique pathology. *Brain*. 2014;137(Pt 5):1361-73.
18. Wong TH, van der Lee SJ, van Rooij JGJ, Meeter LHH, Frick P, Melhem S, et al. EIF2AK3 variants in Dutch patients with Alzheimer's disease. *Neurobiol Aging*. 2019;73:229 e11- e18.
19. Holstege H, van der Lee SJ, Hulsman M, Wong TH, van Rooij JG, Weiss M, et al. Characterization of pathogenic SORL1 genetic variants for association with Alzheimer's disease: a clinical interpretation strategy. *Eur J Hum Genet*. 2017;25(8):973-81.
20. Wong TH, Pottier C, Hondius DC, Meeter LHH, van Rooij JGJ, Melhem S, et al. Three VCP Mutations in Patients with Frontotemporal Dementia. *J Alzheimers Dis*. 2018;65(4):1139-46.
21. Synofzik M, Schule R, Schulze M, Gburek-Augustat J, Schweizer R, Schirmacher A, et al. Phenotype and frequency of STUB1 mutations: next-generation screenings in Caucasian ataxia and spastic paraplegia cohorts. *Orphanet J Rare Dis*. 2014;9:57.

22. Ferrari R, Manzoni C, Hardy J. Genetics and molecular mechanisms of frontotemporal lobar degeneration: an update and future avenues. *Neurobiol Aging*. 2019;78:98-110.
23. Ferrari R, Forabosco P, Vandrovцова J, Botia JA, Guelfi S, Warren JD, et al. Frontotemporal dementia: insights into the biological underpinnings of disease through gene co-expression network analysis. *Mol Neurodegener*. 2016;11:21.
24. Robak LA, Jansen IE, van Rooij J, Uitterlinden AG, Kraaij R, Jankovic J, et al. Excessive burden of lysosomal storage disorder gene variants in Parkinson's disease. *Brain*. 2017;140(12):3191-203.
25. Jansen IE, Gibbs JR, Nalls MA, Price TR, Lubbe S, van Rooij J, et al. Establishing the role of rare coding variants in known Parkinson's disease risk loci. *Neurobiol Aging*. 2017;59:220 e11- e18.
26. Pottier C, Ren Y, Perkerson RB, 3rd, Baker M, Jenkins GD, van Blitterswijk M, et al. Genome-wide analyses as part of the international FTLD-TDP whole-genome sequencing consortium reveals novel disease risk factors and increases support for immune dysfunction in FTLD. *Acta Neuropathol*. 2019;137(6):879-99.
27. Blauwendraat C, Wilke C, Simon-Sanchez J, Jansen IE, Reifschneider A, Capell A, et al. The wide genetic landscape of clinical frontotemporal dementia: systematic combined sequencing of 121 consecutive subjects. *Genet Med*. 2018;20(2):240-9.
28. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56-65.
29. Karczewski KJ, Weisburd B, Thomas B, Solomonson M, Ruderfer DM, Kavanagh D, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res*. 2017;45(D1):D840-D5.
30. Chen R, Shi L, Hakenberg J, Naughton B, Sklar P, Zhang J, et al. Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nat Biotechnol*. 2016;34(5):531-8.
31. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46(3):310-5.
32. van Rooij JGJ, Jhamai M, Arp PP, Nouwens SCA, Verkerk M, Hofman A, et al. Population-specific genetic variation in large sequencing data sets: why more data is still better. *Eur J Hum Genet*. 2017;25(10):1173-5.
33. Minikel EV, Vallabh SM, Lek M, Estrada K, Samocha KE, Sathirapongsasuti JF, et al. Quantifying prion disease penetrance using large population control cohorts. *Sci Transl Med*. 2016;8(322):322ra9.
34. Ropers HH, Wienker T. Penetrance of pathogenic mutations in haploinsufficient genes for intellectual disability and related disorders. *Eur J Med Genet*. 2015;58(12):715-8.
35. van der Lee SJ, Holstege H, Wong TH, Jakobsdottir J, Bis JC, Chouraki V, et al. PLD3 variants in population studies. *Nature*. 2015;520(7545):E2-3.
36. Giri A, Mok KY, Jansen I, Sharma M, Tesson C, Mangone G, et al. Lack of evidence for a role of genetic variation in TMEM230 in the risk for Parkinson's disease in the Caucasian population. *Neurobiol Aging*. 2017;50:167 e11- e13.
37. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, et al. Disease variants alter transcription factor levels and methylation of their binding sites. *Nat Genet*. 2017;49(1):131-8.
38. Peters MJ, Joehanes R, Pilling LC, Schurmann C, Conneely KN, Powell J, et al. The transcriptional landscape of age in human peripheral blood. *Nat Commun*. 2015;6:8570.
39. Zhernakova DV, Deelen P, Vermaat M, van Iterson M, van Galen M, Arindrarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat Genet*. 2017;49(1):139-45.

40. van Rooij J, Mandaviya PR, Claringbould A, Felix JF, van Dongen J, Jansen R, et al. Evaluation of commonly used analysis strategies for epigenome- and transcriptome-wide association studies through replication of large-scale population studies. *Genome Biol.* 2019;20(1):235.
41. Robinson MD, Vitek O. Benchmarking comes of age. *Genome Biol.* 2019;20(1):205.
42. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297-303.
43. Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat Protoc.* 2015;10(10):1556-66.
44. van Rooij JGJ, Meeter LHH, Melhem S, Nijholt DAT, Wong TH, Netherlands Brain B, et al. Hippocampal transcriptome profiling combined with protein-protein interaction analysis elucidates Alzheimer's disease pathways and genes. *Neurobiol Aging.* 2019;74:225-33.
45. Humphries CE, Kohli MA, Nathanson L, Whitehead P, Beecham G, Martin E, et al. Integrated whole transcriptome and DNA methylation analysis identifies gene networks specific to late-onset Alzheimer's disease. *J Alzheimers Dis.* 2015;44(3):977-87.
46. Sekar S, McDonald J, Cuyugan L, Aldrich J, Kurdoglu A, Adkins J, et al. Alzheimer's disease is associated with altered expression of genes involved in immune response and mitochondrial processes in astrocytes. *Neurobiol Aging.* 2015;36(2):583-91.
47. Twine NA, Janitz K, Wilkins MR, Janitz M. Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease. *PLoS One.* 2011;6(1):e16266.
48. Van Cauwenbergh C, Van Broeckhoven C, Sleegers K. The genetic landscape of Alzheimer disease: clinical implications and perspectives. *Genet Med.* 2016;18(5):421-30.
49. Netherlands Brain B. <https://www.brainbank.nl> 2019 [
50. Lodato MA, Rodin RE, Bohrsen CL, Coulter ME, Barton AR, Kwon M, et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science.* 2018;359(6375):555-9.
51. Park JS, Lee J, Jung ES, Kim MH, Kim IB, Son H, et al. Brain somatic mutations observed in Alzheimer's disease associated with aging and dysregulation of tau phosphorylation. *Nat Commun.* 2019;10(1):3090.
52. van den Akker EB, Pitts SJ, Deelen J, Moed MH, Potluri S, van Rooij J, et al. Uncompromised 10-year survival of oldest old carrying somatic mutations in DNMT3A and TET2. *Blood.* 2016;127(11):1512-5.
53. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* 2017;45(D1):D777-D83.
54. Lelieveld SH, Reijnders MR, Pfundt R, Yntema HG, Kamsteeg EJ, de Vries P, et al. Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat Neurosci.* 2016;19(9):1194-6.
55. Verheijen BM, Vermulst M, van Leeuwen FW. Somatic mutations in neurons during aging and neurodegeneration. *Acta Neuropathol.* 2018;135(6):811-26.
56. Bae T, Tomasini L, Mariani J, Zhou B, Roychowdhury T, Franjic D, et al. Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. *Science.* 2018;359(6375):550-5.
57. Ferrari R, Hernandez DG, Nalls MA, Rohrer JD, Ramasamy A, Kwok JB, et al. Frontotemporal dementia and its subtypes: a genome-wide association study. *Lancet Neurol.* 2014;13(7):686-99.
58. Neumann M, Sampathu DM, Kwong LK, Truax AC, Micsenyi MC, Chou TT, et al. Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science.* 2006;314(5796):130-3.

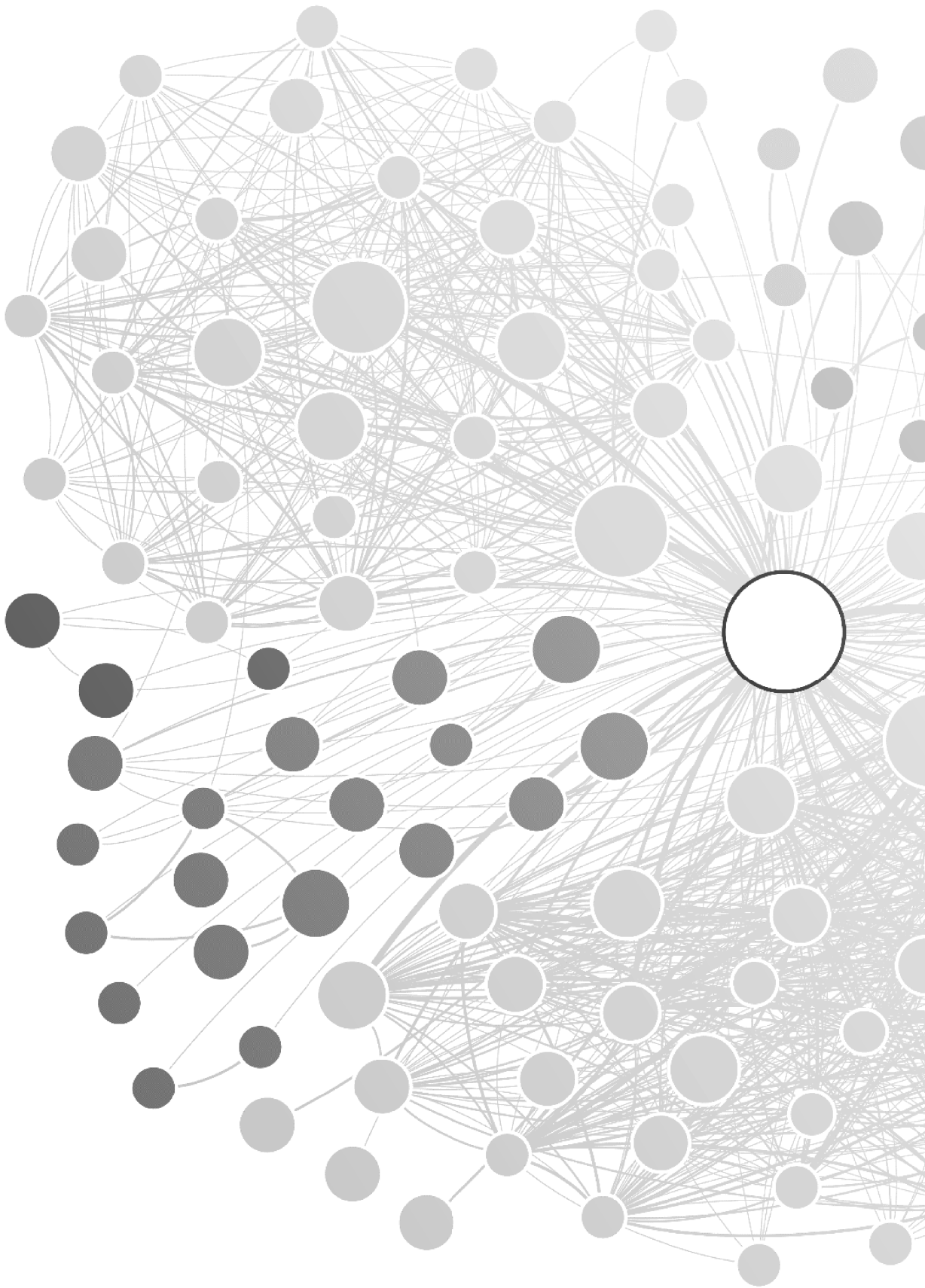
59. Mackenzie IR, Neumann M, Baborie A, Sampathu DM, Du Plessis D, Jaros E, et al. A harmonized classification system for FTLT-DTP pathology. *Acta Neuropathol.* 2011;122(1):111-3.
60. Bodea LG, Eckert A, Ittner LM, Piguot O, Gotz J. Tau physiology and pathomechanisms in frontotemporal lobar degeneration. *J Neurochem.* 2016;138 Suppl 1:71-94.
61. Irwin DJ, Cairns NJ, Grossman M, McMillan CT, Lee EB, Van Deerlin VM, et al. Frontotemporal lobar degeneration: defining phenotypic diversity through personalized medicine. *Acta Neuropathol.* 2015;129(4):469-91.
62. Mackenzie IR, Neumann M, Bigio EH, Cairns NJ, Alafuzoff I, Kril J, et al. Nomenclature and nosology for neuropathologic subtypes of frontotemporal lobar degeneration: an update. *Acta Neuropathol.* 2010;119(1):1-4.
63. Hodges JR, Patterson K. Semantic dementia: a unique clinicopathological syndrome. *Lancet Neurol.* 2007;6(11):1004-14.
64. Lattante S, Rouleau GA, Kabashi E. *TARDBP* and *FUS* mutations associated with amyotrophic lateral sclerosis: summary and update. *Hum Mutat.* 2013;34(6):812-26.
65. D'Ambrogio A, Buratti E, Stuani C, Guarnaccia C, Romano M, Ayala YM, et al. Functional mapping of the interaction between TDP-43 and hnRNP A2 in vivo. *Nucleic Acids Res.* 2009;37(12):4116-26.
66. Buratti E, Baralle FE. Characterization and functional implications of the RNA binding properties of nuclear factor TDP-43, a novel splicing regulator of CFTR exon 9. *J Biol Chem.* 2001;276(39):36337-43.
67. Mompean M, Romano V, Pantoja-Uceda D, Stuani C, Baralle FE, Buratti E, et al. Point mutations in the N-terminal domain of transactive response DNA-binding protein 43 kDa (TDP-43) compromise its stability, dimerization, and functions. *J Biol Chem.* 2017;292(28):11992-2006.
68. Aly SM, Sabri DM. Next generation sequencing (NGS): a golden tool in forensic toolkit. *Arch Med Sadovej Kryminol.* 2015;65(4):260-71.
69. Boyd SD. Diagnostic applications of high-throughput DNA sequencing. *Annu Rev Pathol.* 2013;8:381-410.
70. Berg JS, Agrawal PB, Bailey DB, Jr., Beggs AH, Brenner SE, Brower AM, et al. Newborn Sequencing in Genomic Medicine and Public Health. *Pediatrics.* 2017;139(2).
71. Roberts JS, Robinson JO, Diamond PM, Bharadwaj A, Christensen KD, Lee KB, et al. Patient understanding of, satisfaction with, and perceived utility of whole-genome sequencing: findings from the MedSeq Project. *Genet Med.* 2018;20(9):1069-76.
72. Lupo PJ, Robinson JO, Diamond PM, Jamal L, Danysh HE, Blumenthal-Barby J, et al. Patients' perceived utility of whole-genome sequencing for their healthcare: findings from the MedSeq project. *Per Med.* 2016;13(1):13-20.
73. Vassy JL, Lautenbach DM, McLaughlin HM, Kong SW, Christensen KD, Krier J, et al. The MedSeq Project: a randomized trial of integrating whole genome sequencing into clinical medicine. *Trials.* 2014;15:85.
74. Vassy JL, Christensen KD, Schonman EF, Blout CL, Robinson JO, Krier JB, et al. The Impact of Whole-Genome Sequencing on the Primary Care and Outcomes of Healthy Adult Patients: A Pilot Randomized Trial. *Ann Intern Med.* 2017;167(3):159-69.
75. Christensen KD, Vassy JL, Phillips KA, Blout CL, Azzariti DR, Lu CY, et al. Short-term costs of integrating whole-genome sequencing into primary care and cardiology settings: a pilot randomized trial. *Genet Med.* 2018;20(12):1544-53.
76. Schwarze K, Buchanan J, Taylor JC, Wordsworth S. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genet Med.* 2018;20(10):1122-30.



77. Arora NS, Davis JK, Kirby C, McGuire AL, Green RC, Blumenthal-Barby JS, et al. Communication challenges for nongeneticist physicians relaying clinical genomic results. *Per Med*. 2016;14(5):423-31.
78. Holm IA, Agrawal PB, Ceyhan-Birsoy O, Christensen KD, Fayer S, Frankel LA, et al. The BabySeq project: implementing genomic sequencing in newborns. *BMC Pediatr*. 2018;18(1):225.
79. Holm IA, McGuire A, Pereira S, Rehm H, Green RC, Beggs AH, et al. Returning a Genomic Result for an Adult-Onset Condition to the Parents of a Newborn: Insights From the BabySeq Project. *Pediatrics*. 2019;143(Suppl 1):S37-S43.
80. Machini K, Ceyhan-Birsoy O, Azzariti DR, Sharma H, Rossetti P, Mahanta L, et al. Analyzing and Reanalyzing the Genome: Findings from the MedSeq Project. *Am J Hum Genet*. 2019;105(1):177-88.
81. Christensen KD, Phillips KA, Green RC, Dukhovny D. Cost Analyses of Genomic Sequencing: Lessons Learned from the MedSeq Project. *Value Health*. 2018;21(9):1054-61.
82. Gonorazky HD, Naumenko S, Ramani AK, Nelakuditi V, Mashouri P, Wang P, et al. Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. *Am J Hum Genet*. 2019;104(5):1007.
83. Fresard L, Smail C, Ferraro NM, Teran NA, Li X, Smith KS, et al. Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. *Nat Med*. 2019;25(6):911-9.
84. Baslan T, Hicks J. Unravelling biology and shifting paradigms in cancer with single-cell sequencing. *Nat Rev Cancer*. 2017;17(9):557-69.
85. van der Ende EL, Meeter LH, Stingl C, van Rooij JGJ, Stoop MP, Nijholt DAT, et al. Novel CSF biomarkers in genetic frontotemporal dementia identified by proteomics. *Ann Clin Transl Neurol*. 2019;6(4):698-707.
86. Voellenkle C, van Rooij J, Cappuzzello C, Greco S, Arcelli D, Di Vito L, et al. MicroRNA signatures in peripheral blood mononuclear cells of chronic heart failure patients. *Physiol Genomics*. 2010;42(3):420-6.
87. Teunissen CE, Elias N, Koel-Simmelink MJ, Durieux-Lu S, Malekzadeh A, Pham TV, et al. Novel diagnostic cerebrospinal fluid biomarkers for pathologic subtypes of frontotemporal dementia identified by proteomics. *Alzheimers Dement (Amst)*. 2016;2:86-94.
88. Taskesen E, Reinders MJ. 2D Representation of Transcriptomes by t-SNE Exposes Relatedness between Human Tissues. *PLoS One*. 2016;11(2):e0149853.
89. Bron EE, Smits M, Papma JM, Steketee RME, Meijboom R, de Groot M, et al. Multiparametric computer-aided differential diagnosis of Alzheimer's disease and frontotemporal dementia using structural and advanced MRI. *Eur Radiol*. 2017;27(8):3372-82.
90. Vieira S, Pinaya WH, Mechelli A. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neurosci Biobehav Rev*. 2017;74(Pt A):58-75.
91. Mamoshina P, Vieira A, Putin E, Zhavoronkov A. Applications of Deep Learning in Biomedicine. *Mol Pharm*. 2016;13(5):1445-54.
92. Chen R, Mias GI, Li-Pook-Tham J, Jiang L, Lam HY, Chen R, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*. 2012;148(6):1293-307.
93. Garcia-Nieto PE, Morrison AJ, Fraser HB. The somatic mutation landscape of the human body. *Genome Biol*. 2019;20(1):298.
94. Risques RA, Kennedy SR. Aging and the rise of somatic cancer-associated mutations in normal tissues. *PLoS Genet*. 2018;14(1):e1007108.
95. Juan-Mateu J, Paradas C, Olive M, Verdura E, Rivas E, Gonzalez-Quereda L, et al. Isolated cardiomyopathy caused by a DMD nonsense mutation in somatic mosaicism: genetic normalization in skeletal muscle. *Clin Genet*. 2012;82(6):574-8.

96. Lodato MA, Walsh CA. Genome aging: somatic mutation in the brain links age-related decline with disease and nominates pathogenic mechanisms. *Hum Mol Genet.* 2019;28(R2):R197-R206.
97. D’Gama AM, Walsh CA. Somatic mosaicism and neurodevelopmental disease. *Nat Neurosci.* 2018;21(11):1504-14.
98. Rodin RE, Walsh CA. Somatic Mutation in Pediatric Neurological Diseases. *Pediatr Neurol.* 2018;87:20-2.
99. Evans MA, Sano S, Walsh K. Cardiovascular Disease, Aging, and Clonal Hematopoiesis. *Annu Rev Pathol.* 2019.







# Chapter 6

## Appendices



## 6.1. Summary

Sequencing is used to determine the order (sequence) of nucleotides in genomic fragments (DNA or RNA). Initially used to study specific DNA fragments, for example a gene associated with a specific disease, advancements in technology permit whole genome or transcriptome sequencing in a single experiment. This improvement permits novel applications of genetic information. For example, genome sequencing can be used in a clinical setting, to test for risk of certain diseases, and then screen for or prevent these diseases. Simultaneously, such an untargeted approach increases unexpected findings, which should be approached carefully. In general, collecting and analyzing more data, requires more careful and critical handling of that data. In this thesis, we discuss several applications and considerations of next-generation sequencing; the method of collecting large amounts of sequencing data in a single experiment.

First, in **chapter 2** we speak about the use of sequencing to detect germline (inherited from the parents) genetic variants. Generally, we isolate DNA from the blood of a person, and extract from it all fragments that code for genes, which make up approximately 1.5% of the whole human genome. Using a next-generation sequencing device, we determine the sequence of millions of these fragments, map where they belong on the whole genome sequence and look for deviations compared to the so-called “reference genome”, a representation of the average human genome sequence. This results in a list of ~25,000 coding genetic variants per person, which can be shared between many persons (common variants) or present in only a single or a few carriers (rare variants). **Chapter 2.1** described this process for a population of 2,628 samples from the Rotterdam Study cohort, a local population-study which investigates disease and disability among the elderly in the Netherlands. This chapter describes the steps taken to generate this information per-person and how to combine this data for many persons. It also reports on how to evaluate if the data generated for a person is “good”, i.e., all variants present in that person’s genes are detected without observing many false findings. This report provides practical guidelines to help the generation of similar datasets by other researchers. Additionally, in chapter 2.1 this dataset of coding variants in Dutch individuals was compared to other similar population-level datasets. The results showed that each dataset harbored many genetic variants that were absent in all other populations. These are almost exclusively rare variants only observed in one or a few persons with uncertain biological and clinical relevance. In **chapter 2.2** this topic continues by attempting to identify the most clinically relevant genetic variants in the dataset described in chapter 2.1. This manuscript applies recent recommendations from the genetics field to investigate disease causing or so-called “pathogenic” variants in a set of 59 predetermined genes, for which we know that pathogenic variants cause preventable diseases. Within the 2,628 participants of the Rotterdam Study, 24 carried a variant that fulfilled the recommended criteria to be reported to their carriers for further testing. Due to the design of this study, we had access to life-long clinical information of these carriers, and observed that at most three carriers experienced a disease that could be caused by their pathogenic variant. The study concludes that when testing in such an unbiased manner, we will identify many seemingly pathogenic variants that will not result in disease, and thus this practice should be considered carefully. In **chapter 2.3** the same method is used to identify novel pathogenic variants in Dutch families suffering from Alzheimer’s Disease (AD).

By sequencing the genes of 19 AD patients from 8 families and comparing the identified genetic variants, first between family members, then between families, variants in the *EIF2AK3* gene were identified as the most likely cause of disease in two AD families. Follow-up analyses of this specific gene, including in the Rotterdam Study dataset described in the previous chapters, demonstrated that rare variants in this gene were more often observed in AD cases versus non-demented controls. Thus, this study identified a novel gene in which genetic variants cause AD.

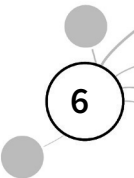
Next, in **chapter 3** the sequencing method is used to study the transcriptome. A transcript is a gene that is copied of the genome in order to produce its corresponding protein. The transcript is in so-called RNA form, which can be reverted to its DNA sequence and sequenced as a DNA fragment. RNA sequencing has an additional complexity, as multiple transcripts of the same gene can be present in the same cell, if that cell has need of multiple copies of the respective protein. Thus, the amount of RNA fragments for each gene indicate the gene's activity in that cell or tissue (when many cells are taken). This also means that RNA-sequencing of blood or brain of the same person will yield different results. First, in **chapter 3.1** RNA-sequencing is used on RNA fragments extracted from blood of participants from the Rotterdam Study and three other similar population studies, for a total of 2,800 participants. Within this dataset, a large number of different methods to analyze RNA-Sequencing, and another similar datatype; DNA-methylation array, data were applied to determine their relative influences on the results. Each analysis aimed to determine which genes were associated with increased age, smoking status and/or increased BMI, as these are phenotypes with large suspected impact on the transcriptome (or methylome) which might be detected in blood. These results show that a large number of analysis options do not have a large influence on the interpretation of results, but some have, such as correction for so-called principal components, and this should be considered when interpreting the results of such a study. Next, in **chapter 3.2** the transcriptome was sequenced in hippocampus brain tissues from patients with Alzheimer's Disease and non-demented age- and sex-matched controls. For this study, the hippocampus was selected as it is responsible for memory formation and retrieval, the main cognitive domain affected in AD. These results showed enormous differences in gene activity between the cases and the controls, with more than 40% of all detected genes affected, spanning more than a hundred different biological pathways.

Finally, in **chapter 4.1** both approaches from chapters 2 and 3 are combined, by sequencing the DNA from the disease-relevant tissue and investigating DNA variants which are present in the brain tissue of dementia patients, but not in their blood. These so-called somatic variants occur during cell division by incomplete DNA replication or later in life due to DNA damage. As they are not present in DNA from the blood, regular testing for pathogenic variants, such as described in chapter 2, will not identify such variants. In this chapter, a series of sixteen Semantic Dementia patients, a neurodegenerative disease with homogeneous clinical and pathological presentation but without any known familial occurrence, was studied for somatic brain variants. In two of the sixteen patients, DNA from the brain revealed possibly pathogenic somatic variants in the *TARDBP* gene, in which germline variants detected in blood are already known to cause neurodegenerative diseases ALS and FTD, with similar but



distinct pathological presentation. This study demonstrated for the first time that somatic variants in a neurodegenerative disease gene can also cause dementia.

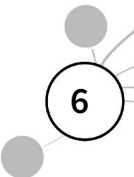
Finally, **chapter 5** discussed in more detail the above findings and their context in other similar research conducted in the last years, including those in which the author of this thesis participated. In this chapter an outline for the expected future directions and considerations of sequencing in research and clinical practice is discussed.





## 6.2. About the author

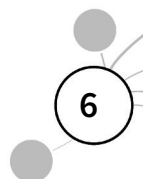
Jeroen Gerardus Johannes was born in s’Hertogenbosch on the 6<sup>th</sup> of March in 1988. He graduated from the Avans Academy for Technology, Health and Environment in 2009 with a major in biomedical research and minor in bioinformatics. His graduation thesis was on miRNA involvement in cardiac tissue of ischemic or idiopathic heart disease performed under Prof. Fabio Martelli in the Policlinico San Donato (Milan). He remained there for another year working on miRNA detection from sequencing data, before accepting a position as bio-informatician in the group of Prof. Uitterlinden at the Erasmus Medical Center (EMC) in February of 2010. In 2015, this position moved towards a PhD project in collaboration with Prof. John van Swieten at the department of Neurology. During this period, the author completed a Master’s program in Genetic Epidemiology at the Erasmus University Rotterdam. In 2020, the author received the CHARGE Early Career Achievement award for his efforts in the CHARGE (Cohorts for Heart and Aging Research in Genomic Epidemiology) consortium. After his defense, the author remains at EMC in a Postdoc position at the departments of Internal Medicine, working on personalized disease prevention through genetic evaluation (the GOALL project) and Neurology, continuing the semantic dementia research shown in chapter 4.1.





### 6.3. Portfolio

Courses and Training	Year	ECTS
SNPs and Human Diseases	2011	1.4
Study Design	2012	4.3
Analysis of microarray and RNA SEQ expression data	2013	2.0
Biostatistical Methods II: Classical Regression Models	2013	4.3
Family Based Genetic Analysis	2013	1.4
Bayesian Statistics	2015	1.4
Genetic-epidemiology Research Methods	2015	5.1
Psychiatric Epidemiology	2015	1.1
Repeated Measurements in Clinical Studies	2015	1.4
Advances in Genome-Wide Association Studies	2016	1.4
Course on R	2016	1.4
Epidemiology of Infectious Diseases	2016	1.4
Microscopic Image Analysis: From Theory to Practice	2016	0.8
Photoshop and Illustrator for PhD-students and other researchers	2016	0.3
Programming with Python	2016	1.0
Psychopharmacology	2016	1.4
Quality of Life Measurements	2016	0.9
Advanced Topics in Clinical Trials	2017	1.9
Advanced Topics in Decision-making in Medicine	2017	2.4
Diagnostic Research	2017	1.4
Medical Demography	2017	1.1
Research Integrity	2017	0.3
<i>remaining courses to complete master in genetic epidemiology</i>	2013-2018	17.2
<i>research training to complete master in genetic epidemiology</i>	2017-2018	73.6
<b>Total</b>		<b>128</b>



Conferences/Symposia - Speaker	Year	ECTS
1000 Genomes User Meeting - Ann Arbor (abstract)	2012	1.0
ADES User meetings - Cardiff (invited)	2015	1.0
ADES User meetings - Lille (invited)	2016	1.0
Alzheimercafé wetenschap en dementie - Rotterdam (invited)	2019	1.0
CHARGE International Meetings - Houston (abstract)	2020	1.0
CHARGE International Meetings - Los Angeles (abstract)	2014	1.0
CHARGE International Meetings - Rotterdam (abstract)	2013	1.0
Clinical Oncogenetics Refereeravond - Rotterdam (invited)	2019	1.0
Dutch Society of Human Genetics - Veldhoven (abstract)	2019	1.0
Erasmus Mini Symposium GOALL - Rotterdam (invited)	2019	1.0
Erasmus Mini Symposium GOALL - Rotterdam (invited)	2020	1.0
ERGO Exome Dataset Release Mini Symposium - Rotterdam (invited)	2013	1.0
ERGO RNA Dataset Release Mini Symposium - Rotterdam (invited)	2017	1.0
ESHG Meetings - Berlin (digital, abstract)	2020	1.0
Internal Medicine Science Days - Antwerp (abstract)	2015	1.0
Transcriptome Sequencing Mini Symposium - Utrecht (invited)	2014	1.0
RIVM Sequencing Symposium - Bilthoven (invited)	2012	1.0

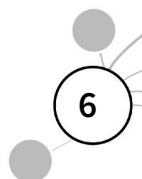
Conferences/Symposia - Poster	Year	ECTS
Alzheimer Association International Conference - Toronto	2016	1.0
American Society of Human Genetics - Houston	2019	1.0
CHARGE International Meetings - Boston	2012	1.0
CHARGE International Meetings - Charlottesville	2016	1.0
CHARGE International Meetings - Rotterdam	2018	1.0
CHARGE International Meetings - St. Louis	2019	1.0
ESHG Meetings - Glasgow	2015	1.0
ESHG Meetings - Göttenburg	2019	1.0
ESHG Meetings - Kopenhagen	2017	1.0
ESHG Meetings - Milaan	2018	1.0
Internal Medicine Science Days - Antwerp 2012-2016	2012	4.0
Internal Medicine Science Days - Sint Michielsgestel 2018-2020	2018	3.0
International Consortium for FTD - Munchen	2016	1.0

<b>Total</b>		<b>35</b>
--------------	--	-----------

Teaching - Courses	Year	ECTS
NIHES - Erasmus Summer Program 2013-2019	2013	1.0
NIHES - Next Generation Sequencing Course 2012-2020	2012	4.5
Avans University Breda - Supervisor Minor Bioinformatics 2012-2016	2012	2.0
MOLMED - SNP Course 2012 - 2019	2012	1.5

Teaching - Students	Year	ECTS
Avans University Breda - Dennis Schmitz	2013	2.0
Avans University Breda - Mariëlla Klein	2013	2.0
Avans University Breda - Robert Nooijens	2013	1.0
University Leiden - Annelies Smouter	2013	1.0
Avans University Breda - Joost Verlouw	2015	2.0
Avans University Breda - Tom de Laat	2015	2.0
Avans University Breda - Theo de Vet	2015	2.0
University Utrecht - Coco Versluijs	2017	2.0
University of Geneva - Merel van der Thiel	2017	1.0
Avans University Breda - Michiel van Berkel	2019	1.0
Technical University Delft - Simone Smits	2019	2.0
University Amsterdam - Robin Groenenboom	2020	1.0
Erasmus University Rotterdam - Merel Mol (PhD-student)	2018	8.0
Erasmus University Rotterdam - Bahar Sedaghatik-hayat (PhD-student)	2019	5.0
Erasmus University Rotterdam - Vivi Zhou (PhD-student)	2019	3.0

<b>Total</b>		<b>44</b>
--------------	--	-----------







## 6.4. List of Publications

1. **MicroRNA signatures in peripheral blood mononuclear cells of chronic heart failure patients;** Voellenkle C, *van Rooij J*, Cappuzzello C, Greco S, Arcelli D, Di Vito L, Melillo G, Rigolini R, Costa E, Crea F, Capogrossi MC, Napolitano M, Martelli F.; *Physiol Genomics*. 2010 Aug;42(3):420-6. doi: 10.1152/physiolgenomics.00211.2009. Epub 2010 May 18.
2. **The dystrophin gene and cognitive function in the general population;** Vojinovic D, Adams HH, van der Lee SJ, Ibrahim-Verbaas CA, Brouwer R, van den Hout MC, Oole E, *van Rooij J*, Uitterlinden A, Hofman A, van IJcken WF, Aartsma-Rus A, van Ommen GB, Ikram MA, van Duijn CM, Amin N.; *Eur J Hum Genet*. 2015 Jun;23(6):837-43. doi: 10.1038/ejhg.2014.183. Epub 2014 Sep 17.
3. **The transcriptional landscape of age in human peripheral blood;** Peters MJ, Joehanes R, Pilling LC, Schurmann C, Conneely KN, Powell J, Reinmaa E, Sutphin GL, Zhernakova A, Schramm K, Wilson YA, Kobes S, Tukiainen T; NABEC/UKBEC Consortium, Ramos YF, Garing HH, Fornage M, Liu Y, Gharib SA, Stranger BE, De Jager PL, Aviv A, Levy D, Murabito JM, Munson PJ, Huan T, Hofman A, Uitterlinden AG, Rivadeneira F, *van Rooij J*, Stolk L, Broer L, Verbiest MM, Jhamai M, Arp P, Metspalu A, Tserel L, Milani L, Samani NJ, Peterson P, Kasela S, Codd V, Peters A, Ward-Caviness CK, Herder C, Waldenberger M, Roden M, Singmann P, Zeilinger S, Illig T, Homuth G, Grabe HJ, Valzke H, Steil L, Kocher T, Murray A, Melzer D, Yaghoobkar H, Bandinelli S, Moses EK, Kent JW, Curran JE, Johnson MP, Williams-Blangero S, Westra HJ, McRae AF, Smith JA, Kardia SL, Hovatta I, Perola M, Ripatti S, Salomaa V, Henders AK, Martin NG, Smith AK, Mehta D, Binder EB, Nylocks KM, Kennedy EM, Klengel T, Ding J, Suchy-Dicey AM, Enquobahrie DA, Brody J, Rotter JI, Chen YD, Houwing-Duistermaat J, Kloppenburg M, Slagboom PE, Helmer Q, den Hollander W, Bean S, Raj T, Bakhshi N, Wang QP, Oyston LJ, Psaty BM, Tracy RP, Montgomery GW, Turner ST, et al.; *Nat Commun*. 2015 Oct 22;6:8570. doi: 10.1038/ncomms9570.
4. **PLD3 variants in population studies;** van der Lee SJ, Holstege H, Wong TH, Jakobsdottir J, Bis JC, Chouraki V, *van Rooij JG*, Grove ML, Smith AV, Amin N, Choi SH, Beiser AS, Garcia ME, van IJcken WF, Pijnenburg YA, Louwersheimer E, Brouwer RW, van den Hout MC, Oole E, Eiriksdottir G, Levy D, Rotter JI, Emilsson V, O'Donnell CJ, Aspelund T, Uitterlinden AG, Launer LJ, Hofman A, Boerwinkle E, Psaty BM, DeStefano AL, Scheltens P, Seshadri S, van Swieten JC, Gudnason V, van der Flier WM, Ikram MA, van Duijn CM.; *Nature*. 2015 Apr 2;520(7545):E2-3. doi: 10.1038/nature14038.
5. **Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture;** Zheng HF, Forgetta V, Hsu YH, Estrada K, Rosello-Diez A, Leo PJ, Dahia CL, Park-Min KH, Tobias JH, Kooperberg C, Kleinman A, Styrkarsdottir U, Liu CT, Uggla C, Evans DS, Nielson CM, Walter K, Pettersson-Kymmer U, McCarthy S, Eriksson J, Kwan T, Jhamai M, Trajanoska K, Memari Y, Min J, Huang J, Danecek P, Wilmot B, Li R, Chou WC, Mokry LE, Moayyeri A, Claussnitzer M, Cheng CH, Cheung W, Medina-Gómez C, Ge B, Chen SH, Choi K, Oei L, Fraser J, Kraaij R, Hibbs MA, Gregson CL, Paquette D, Hofman A, Wibom C, Tranah GJ, Marshall M, Gardiner BB, Cremin K, Auer P, Hsu L, Ring

- S, Tung JY, Thorleifsson G, Enneman AW, van Schoor NM, de Groot LC, van der Velde N, Melin B, Kemp JP, Christiansen C, Sayers A, Zhou Y, Calderari S, *van Rooij J*, Carlson C, Peters U, Berlivet S, Dostie J, Uitterlinden AG, Williams SR, Farber C, Grinberg D, LaCroix AZ, Haessler J, Chasman DI, Giulianini F, Rose LM, Ridker PM, Eisman JA, Nguyen TV, Center JR, Nogues X, Garcia-Giralt N, Launer LL, Gudnason V, Mellström D, Vandenput L, Amin N, van Duijn CM, Karlsson MK, Ljunggren Ö, Svensson O, Hallmans G, Rousseau F, Giroux S, Bussière J, Arp PP, et al.; Nature. 2015 Oct 1;526(7571):112-7. doi: 10.1038/nature14878. Epub 2015 Sep 14.
6. **Uncompromised 10-year survival of oldest old carrying somatic mutations in DNMT3A and TET2;** van den Akker EB, Pitts SJ, Deelen J, Moed MH, Potluri S, *van Rooij J*, Suchiman HE, Lakenberg N, de Dijcker WJ, Uitterlinden AG, Kraaij R, Hofman A, de Craen AJ, Houwing-Duistermaat JJ, van Ommen GJ; Genome of The Netherlands Consortium, Cox DR, van Meurs JB, Beekman M, Reinders MJ, Slagboom PE.; Blood. 2016 Mar 17;127(11):1512-5. doi: 10.1182/blood-2015-12-685925. Epub 2016 Jan 29.
  7. **Novel Genetic Variants for Cartilage Thickness and Hip Osteoarthritis;** Castaño-Betancourt MC, Evans DS, Ramos YF, Boer CG, Metrustry S, Liu Y, den Hollander W, *van Rooij J*, Kraus VB, Yau MS, Mitchell BD, Muir K, Hofman A, Doherty M, Doherty S, Zhang W, Kraaij R, Rivadeneira F, Barrett-Connor E, Maciewicz RA, Arden N, Nelissen RG, Kloppenburg M, Jordan JM, Nevitt MC, Slagboom EP, Hart DJ, Lafeber F, Styrkarsdottir U, Zeggini E, Evangelou E, Spector TD, Uitterlinden AG, Lane NE, Meulenbelt I, Valdes AM, van Meurs JB.; PLoS Genet. 2016 Oct 4;12(10):e1006260. doi: 10.1371/journal.pgen.1006260. eCollection 2016 Oct.
  8. **Quantifying prion disease penetrance using large population control cohorts;** Minikel EV, Vallabh SM, Lek M, Estrada K, Samocha KE, Sathirapongsasuti JF, McLean CY, Tung JY, Yu LP, Gambetti P, Blevins J, Zhang S, Cohen Y, Chen W, Yamada M, Hamaguchi T, Sanjo N, Mizusawa H, Nakamura Y, Kitamoto T, Collins SJ, Boyd A, Will RG, Knight R, Ponto C, Zerr I, Kraus TF, Eigenbrod S, Giese A, Calero M, de Pedro-Cuesta J, Haik S, Laplanche JL, Bouaziz-Amar E, Brandel JP, Capellari S, Parchi P, Pileggi A, Ladogana A, O'Donnell-Luria AH, Karczewski KJ, Marshall JL, Boehnke M, Laakso M, Mohlke KL, Kähler A, Chambert K, McCarroll S, Sullivan PF, Hultman CM, Purcell SM, Sklar P, van der Lee SJ, Rozemuller A, Jansen C, Hofman A, Kraaij R, *van Rooij JG*, Ikram MA, Uitterlinden AG, van Duijn CM; Exome Aggregation Consortium (ExAC), Daly MJ, MacArthur DG.; Sci Transl Med. 2016 Jan 20;8(322):322ra9. doi: 10.1126/scitranslmed.aad5169.
  9. **Nonsynonymous Variation in NKPD1 Increases Depressive Symptoms in European Populations;** Amin N, Belonogova NM, Jovanova O, Brouwer RW, *van Rooij JG*, van den Hout MC, Svishcheva GR, Kraaij R, Zorkoltseva IV, Kirichenko AV, Hofman A, Uitterlinden AG, van IJcken WF, Tiemeier H, Axenovich TI, van Duijn CM.; Biol Psychiatry. 2017 Apr 15;81(8):702-707. doi: 10.1016/j.biopsych.2016.08.008. Epub 2016 Aug 11.
  10. **Excessive burden of lysosomal storage disorder gene variants in Parkinson's disease;** Robak LA, Jansen IE, *van Rooij J*, Uitterlinden AG, Kraaij R, Jankovic J; International

Parkinson's Disease Genomics Consortium (IPDGC), Heutink P, Shulman JM.; *Brain*. 2017 Dec 1;140(12):3191-3203. doi: 10.1093/brain/awx285.

11. **Population-specific genetic variation in large sequencing data sets: why more data is still better**; *van Rooij JGJ*, Jhamai M, Arp PP, Nouwens SCA, Verkerk M, Hofman A, Ikram MA, Verkerk AJ, van Meurs JBJ, Rivadeneira F, Uitterlinden AG, Kraaij R.; *Eur J Hum Genet*. 2017 Oct;25(10):1173-1175. doi: 10.1038/ejhg.2017.110. Epub 2017 Jul 19.
12. **Characterization of pathogenic SORL1 genetic variants for association with Alzheimer's disease: a clinical interpretation strategy**; Holstege H, van der Lee SJ, Hulsman M, Wong TH, *van Rooij JG*, Weiss M, Louwersheimer E, Wolters FJ, Amin N, Uitterlinden AG, Hofman A, Ikram MA, van Swieten JC, Meijers-Heijboer H, van der Flier WM, Reinders MJ, van Duijn CM, Scheltens P.; *Eur J Hum Genet*. 2017 Aug;25(8):973-981. doi: 10.1038/ejhg.2017.87. Epub 2017 May 24.
13. **Exome-Wide Meta-Analysis Identifies Rare 3'-UTR Variant in ERCC1/CD3EAP Associated with Symptoms of Sleep Apnea**; van der Spek A, Luik AI, Kocovska D, Liu C, Brouwer RWW, *van Rooij JGJ*, van den Hout MCGN, Kraaij R, Hofman A, Uitterlinden AG, van IJcken WFJ, Gottlieb DJ, Tiemeier H, van Duijn CM, Amin N.; *Front Genet*. 2017 Oct 18;8:151. doi: 10.3389/fgene.2017.00151. eCollection 2017.
14. **Exome-sequencing in a large population-based study reveals a rare Asn396Ser variant in the LIPG gene associated with depressive symptoms**; Amin N, Jovanova O, Adams HH, Dehghan A, Kavousi M, Vernooij MW, Peeters RP, de Vrij FM, van der Lee SJ, *van Rooij JG*, van Leeuwen EM, Chaker L, Demirkan A, Hofman A, Brouwer RW, Kraaij R, Willems van Dijk K, Hankemeier T, van IJcken WF, Uitterlinden AG, Niessen WJ, Franco OH, Kushner SA, Ikram MA, Tiemeier H, van Duijn CM.; *Mol Psychiatry*. 2017 Apr;22(4):634. doi: 10.1038/mp.2016.141. Epub 2016 Aug 9.
15. **Identification of context-dependent expression quantitative trait loci in whole blood**; Zhernakova DV, Deelen P, Vermaat M, van Iterson M, van Galen M, Arindrarto W, van 't Hof P, Mei H, van Dijk F, Westra HJ, Bonder MJ, *van Rooij J*, Verkerk M, Jhamai PM, Moed M, Kielbasa SM, Bot J, Nooren I, Pool R, van Dongen J, Hottenga JJ, Stehouwer CD, van der Kallen CJ, Schalkwijk CG, Zhernakova A, Li Y, Tigchelaar EF, de Klein N, Beekman M, Deelen J, van Heemst D, van den Berg LH, Hofman A, Uitterlinden AG, van Greevenbroek MM, Veldink JH, Boomsma DI, van Duijn CM, Wijmenga C, Slagboom PE, Swertz MA, Isaacs A, van Meurs JB, Jansen R, Heijmans BT, 't Hoen PA, Franke L.; *Nat Genet*. 2017 Jan;49(1):139-145. doi: 10.1038/ng.3737. Epub 2016 Dec 5.
16. **Disease variants alter transcription factor levels and methylation of their binding sites**; Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, van Iterson M, van Dijk F, van Galen M, Bot J, Sliker RC, Jhamai PM, Verbiest M, Suchiman HE, Verkerk M, van der Breggen R, *van Rooij J*, Lakenberg N, Arindrarto W, Kielbasa SM, Jonkers I, van 't Hof P, Nooren I, Beekman M, Deelen J, van Heemst D, Zhernakova A, Tigchelaar EF, Swertz MA, Hofman A, Uitterlinden AG, Pool R, van Dongen J, Hottenga JJ, Stehouwer CD, van der Kallen CJ, Schalkwijk CG, van den Berg LH, van Zwet EW, Mei

- H, Li Y, Lemire M, Hudson TJ; BIOS Consortium, Slagboom PE, Wijmenga C, Veldink JH, van Greevenbroek MM, van Duijn CM, Boomsma DI, Isaacs A, Jansen R, van Meurs JB, 't Hoen PA, Franke L, Heijmans BT.; *Nat Genet.* 2017 Jan;49(1):131-138. doi: 10.1038/ng.3721. Epub 2016 Dec 5.
17. **Establishing the role of rare coding variants in known Parkinson's disease risk loci;** Jansen IE, Gibbs JR, Nalls MA, Price TR, Lubbe S, *van Rooij J*, Uitterlinden AG, Kraaij R, Williams NM, Brice A, Hardy J, Wood NW, Morris HR, Gasser T, Singleton AB, Heutink P, Sharma M; International Parkinson's Disease Genomics Consortium.; *Neurobiol Aging.* 2017 Nov;59:220.e11-220.e18. doi: 10.1016/j.neurobiolaging.2017.07.009. Epub 2017 Aug 2.
  18. **Lack of evidence for a role of genetic variation in TMEM230 in the risk for Parkinson's disease in the Caucasian population;** Giri A, Mok KY, Jansen I, Sharma M, Tesson C, Mangone G, Lesage S, Bras JM, Shulman JM, Sheerin UM; International Parkinson's Disease Consortium (IPDGC), Dáez-Fairen M, Pastor P, Martá MJ, Ezquerro M, Tolosa E, Correia-Guedes L, Ferreira J, Amin N, van Duijn CM, *van Rooij J*, Uitterlinden AG, Kraaij R, Nalls M, Simón-Sánchez J.; *Neurobiol Aging.* 2017 Feb;50:167.e11-167.e13. doi: 10.1016/j.neurobiolaging.2016.10.004. Epub 2016 Oct 11.
  19. **A C6orf10/LOC101929163 locus is associated with age of onset in C9orf72 carriers;** Zhang M, Ferrari R, Tartaglia MC, Keith J, Surace EI, Wolf U, Sato C, Grinberg M, Liang Y, Xi Z, Dupont K, McGoldrick P, Weichert A, McKeever PM, Schneider R, McCorkindale MD, Manzoni C, Rademakers R, Graff-Radford NR, Dickson DW, Parisi JE, Boeve BF, Petersen RC, Miller BL, Seeley WW, van Swieten JC, *van Rooij J*, Pijnenburg Y, van der Zee J, Van Broeckhoven C, Le Ber I, Van Deerlin V, Suh E, Rohrer JD, Mead S, Graff C, Oijerstedt L, Pickering-Brown S, Rollinson S, Rossi G, Tagliavini F, Brooks WS, Dobson-Stone C, Halliday GM, Hodges JR, Piguat O, Binetti G, Benussi L, Ghidoni R, Nacmias B, Sorbi S, Bruni AC, Galimberti D, Scarpini E, Rainero I, Rubino E, Clarimon J, Lleó A, Ruiz A, Hernández I, Pastor P, Díez-Fairen M, Borroni B, Pasquier F, Deramecourt V, Lebouvier T, Pernecky R, Diehl-Schmid J, Grafman J, Huey ED, Mayeux R, Nalls MA, Hernandez D, Singleton A, Momeni P, Zeng Z, Hardy J, Robertson J, Zinman L, Rogaeva E; International FTD-Genomics Consortium (IFGC).; *Brain.* 2018 Oct 1;141(10):2895-2907. doi: 10.1093/brain/awy238.
  20. **Whole-Genome Linkage Scan Combined With Exome Sequencing Identifies Novel Candidate Genes for Carotid Intima-Media Thickness;** Vojinovic D, Kavousi M, Ghanbari M, Brouwer RWW, *van Rooij JGJ*, van den Hout MCGN, Kraaij R, van Ijcken WFJ, Uitterlinden AG, van Duijn CM, Amin N.; *Front Genet.* 2018 Oct 9;9:420. doi: 10.3389/fgene.2018.00420. eCollection 2018.
  21. **Three VCP Mutations in Patients with Frontotemporal Dementia;** Wong TH, Pottier C, Hondius DC, Meeter LHH, *van Rooij JGJ*, Melhem S; Netherlands Brain bank, van Minkelen R, van Duijn CM, Rozemuller AJM, Seelaar H, Rademakers R, van Swieten JC.; *J Alzheimers Dis.* 2018;65(4):1139-1146. doi: 10.3233/JAD-180301.

22. **Neuropsychiatric Symptoms Complicating the Diagnosis of Alzheimer's Disease: A Case Report**; Eikelboom WS, *van Rooij JGJ*, van den Berg E, Coesmans M, Jiskoot LC, Singleton E, Ossenkuppele R, van Swieten JC, Seelaar H, Papma JM.; J Alzheimers Dis. 2018;66(4):1363-1369. doi: 10.3233/JAD-180700.
23. **A systematic analysis highlights multiple long non-coding RNAs associated with cardiometabolic disorders**; Ghanbari M, Peters MJ, de Vries PS, Boer CG, *van Rooij JGJ*, Lee YC, Kumar V, Uitterlinden AG, Ikram MA, Wijmenga C, Ordovas JM, Smith CE, van Meurs JBJ, Erkeland SJ, Franco OH, Dehghan A.; J Hum Genet. 2018 Apr;63(4):431-446. doi: 10.1038/s10038-017-0403-x. Epub 2018 Jan 30.
24. **Potential genetic modifiers of disease risk and age at onset in patients with frontotemporal lobar degeneration and GRN mutations: a genome-wide association study**; Pottier C, Zhou X, Perkinson RB 3rd, Baker M, Jenkins GD, Serie DJ, Ghidoni R, Benussi L, Binetti G, López de Munain A, Zulaica M, Moreno F, Le Ber I, Pasquier F, Hannequin D, Sánchez-Valle R, Antonell A, Lladó A, Parsons TM, Finch NA, Finger EC, Lipka CF, Huey ED, Neumann M, Heutink P, Synofzik M, Wilke C, Rissman RA, Slawek J, Sitek E, Johannsen P, Nielsen JE, Ren Y, van Blitterswijk M, DeJesus-Hernandez M, Christopher E, Murray ME, Bieniek KF, Evers BM, Ferrari C, Rollinson S, Richardson A, Scarpini E, Fumagalli GG, Padovani A, Hardy J, Momeni P, Ferrari R, Frangipane F, Maletta R, Anfossi M, Gallo M, Petrucelli L, Suh E, Lopez OL, Wong TH, *van Rooij JGJ*, Seelaar H, Mead S, Caselli RJ, Reiman EM, Noel Sabbagh M, Kjolby M, Nykjaer A, Karydas AM, Boxer AL, Grinberg LT, Grafman J, Spina S, Oblak A, Mesulam MM, Weintraub S, Geula C, Hodges JR, Piguet O, Brooks WS, Irwin DJ, Trojanowski JQ, Lee EB, Josephs KA, Parisi JE, Ertekin-Taner N, Knopman DS, Nacmias B, Piaceri I, Bagnoli S, Sorbi S, Gearing M, Glass J, Beach TG, Black SE, Masellis M, Rogaeva E, Vonsattel JP, Honig LS, Kofler J, Bruni AC, Snowden J, Mann D, Pickering-Brown S, et al.; Lancet Neurol. 2018 Jun;17(6):548-558. doi: 10.1016/S1474-4422(18)30126-1. Epub 2018 Apr 30.
25. **Rare coding variants in genes encoding GABA(A) receptors in genetic generalised epilepsies: an exome-based case-control study**; May P, Girard S, Harrer M, Bobbili DR, Schubert J, Wolking S, Becker F, Lachance-Touchette P, Meloche C, Gravel M, Niturad CE, Knaus J, De Kovel C, Toliat M, Polvi A, Iacomino M, Guerrero-López R, Baulac S, Marini C, Thiele H, Altmaller J, Jabbari K, Ruppert AK, Jurkowski W, Lal D, Rusconi R, Cestèle S, Terragni B, Coombs ID, Reid CA, Striano P, Caglayan H, Siren A, Everett K, Møller RS, Hjalgrim H, Muhle H, Helbig I, Kunz WS, Weber YG, Weckhuysen S, Jonghe P, Sisodiya SM, Nabbout R, Franceschetti S, Coppola A, Vari MS, Kasteleijn-Nolst Trenité D, Baykan B, Ozbek U, Bebek N, Klein KM, Rosenow F, Nguyen DK, Dubeau F, Carmant L, Lortie A, Desbiens R, Clément JF, Cieuta-Walti C, Sills GJ, Auce P, Francis B, Johnson MR, Marson AG, Berghuis B, Sander JW, Avbersek A, McCormack M, Cavalleri GL, Delanty N, Depondt C, Krenn M, Zimprich F, Peter S, Nikanorova M, Kraaij R, *van Rooij J*, Balling R, Ikram MA, Uitterlinden AG, Avanzini G, Schorge S, Petrou S, Mantegazza M, Sander T, LeGuern E, Serratosa JM, Koeleman BPC, Palotie A, Lehesjoki AE, Nothnagel M, Nürnberg P, Maljevic S, Zara F, Cossette P, Krause R, Lerche H; Epicure Consortium; EuroEPINOMICS CoGIE Consortium, et al.; Lancet Neurol. 2018 Aug;17(8):699-708. doi: 10.1016/S1474-4422(18)30215-1. Epub 2018 Jul 17.

26. **A rare missense variant in RCL1 segregates with depression in extended families**; Amin N, de Vrij FMS, Baghdadi M, Brouwer RWW, *van Rooij JGJ*, Jovanova O, Uitterlinden AG, Hofman A, Janssen HLA, Darwish Murad S, Kraaij R, Stedehouder J, van den Hout MCGN, Kros JM, van IJcken WFJ, Tiemeier H, Kushner SA, van Duijn CM.; *Mol Psychiatry*. 2018 May;23(5):1120-1126. doi: 10.1038/mp.2017.49. Epub 2017 Mar 21.
27. **Large-scale whole-exome sequencing association studies identify rare functional variants influencing serum urate levels**; Tin A, Li Y, Brody JA, Nutile T, Chu AY, Huffman JE, Yang Q, Chen MH, Robinson-Cohen C, Macé A, Liu J, Demirkan A, Sorice R, Sedaghat S, Swen M, Yu B, Ghasemi S, Teumer A, Vollenweider P, Ciullo M, Li M, Uitterlinden AG, Kraaij R, Amin N, *van Rooij J*, Kutalik Z, Dehghan A, McKnight B, van Duijn CM, Morrison A, Psaty BM, Boerwinkle E, Fox CS, Woodward OM, Köttgen A.; *Nat Commun*. 2018 Oct 12;9(1):4228. doi: 10.1038/s41467-018-06620-4.
28. **Whole-Exome Sequencing in Age-Related Macular Degeneration Identifies Rare Variants in COL8A1, a Component of Bruch's Membrane**; Corominas J, Colijn JM, Geerlings MJ, Pauper M, Bakker B, Amin N, Lores Motta L, Kersten E, Garanto A, Verlouw JAM, *van Rooij JGJ*, Kraaij R, de Jong PTVM, Hofman A, Vingerling JR, Schick T, Fauser S, de Jong EK, van Duijn CM, Hoyng CB, Klaver CCW, den Hollander AI.; *Ophthalmology*. 2018 Sep;125(9):1433-1443. doi: 10.1016/j.ophtha.2018.03.040. Epub 2018 Apr 26.
29. **Rare gene deletions in genetic generalized and Rolandic epilepsies**; Jabbari K, Bobbili DR, Lal D, Reinthaler EM, Schubert J, Wolking S, Sinha V, Motameny S, Thiele H, Kawalia A, Altmaller J, Toliat MR, Kraaij R, *van Rooij J*, Uitterlinden AG, Ikram MA; EuroEPINOMICS CoGIE Consortium, Zara F, Lehesjoki AE, Krause R, Zimprich F, Sander T, Neubauer BA, May P, Lerche H, Nürnberg P.; *PLoS One*. 2018 Aug 27;13(8):e0202022. doi: 10.1371/journal.pone.0202022. eCollection 2018.
30. **Genome-wide analyses as part of the international FTLD-TDP whole-genome sequencing consortium reveals novel disease risk factors and increases support for immune dysfunction in FTLD**; Pottier C, Ren Y, Perkerson RB 3rd, Baker M, Jenkins GD, van Blitterswijk M, DeJesus-Hernandez M, *van Rooij JGJ*, Murray ME, Christopher E, McDonnell SK, Fogarty Z, Batzler A, Tian S, Vicente CT, Matchett B, Karydas AM, Hsiung GR, Seelaar H, Mol MO, Finger EC, Graff C, Oijerstedt L, Neumann M, Heutink P, Synofzik M, Wilke C, Prudlo J, Rizzu P, Simon-Sanchez J, Edbauer D, Roeber S, Diehl-Schmid J, Evers BM, King A, Mesulam MM, Weintraub S, Geula C, Bieniek KF, Petrucelli L, Ahern GL, Reiman EM, Woodruff BK, Caselli RJ, Huey ED, Farlow MR, Grafman J, Mead S, Grinberg LT, Spina S, Grossman M, Irwin DJ, Lee EB, Suh E, Snowden J, Mann D, Ertekin-Taner N, Uitti RJ, Wszolek ZK, Josephs KA, Parisi JE, Knopman DS, Petersen RC, Hodges JR, Piguet O, Geier EG, Yokoyama JS, Rissman RA, Rogava E, Keith J, Zinman L, Tartaglia MC, Cairns NJ, Cruchaga C, Ghetti B, Kofler J, Lopez OL, Beach TG, Arzberger T, Herms J, Honig LS, Vonsattel JP, Halliday GM, Kwok JB, White CL 3rd, Gearing M, Glass J, Rollinson S, Pickering-Brown S, Rohrer JD, Trojanowski JQ, Van Deerlin V, Bigio EH, Troakes C, Al-Sarraj S, Asmann Y, Miller BL, Graff-Radford NR, Boeve BF, Seeley WW, et al.; *Acta Neuropathol*. 2019 Jun;137(6):879-899. doi: 10.1007/s00401-019-01962-9. Epub 2019 Feb 9.

31. **Novel CSF biomarkers in genetic frontotemporal dementia identified by proteomics;** van der Ende EL, Meeter LH, Stingl C, *van Rooij JGJ*, Stoop MP, Nijholt DAT, Sanchez-Valle R, Graff C, Öjjerstedt L, Grossman M, McMillan C, Pijnenburg YAL, Laforce R Jr, Binetti G, Benussi L, Ghidoni R, Luijder TM, Seelaar H, van Swieten JC.; *Ann Clin Transl Neurol.* 2019 Mar 7;6(4):698-707. doi: 10.1002/acn3.745. eCollection 2019 Apr.
32. **Evaluation of commonly used analysis strategies for epigenome- and transcriptome-wide association studies through replication of large-scale population studies;** *van Rooij J*, Mandaviya PR, Claringbould A, Felix JF, van Dongen J, Jansen R, Franke L; BIOS consortium, 't Hoen PAC, Heijmans B, van Meurs JBJ.; *Genome Biol.* 2019 Nov 14;20(1):235. doi: 10.1186/s13059-019-1878-x.
33. **Hippocampal transcriptome profiling combined with protein-protein interaction analysis elucidates Alzheimer's disease pathways and genes;** *van Rooij JGJ*, Meeter LHH, Melhem S, Nijholt DAT, Wong TH; Netherlands Brain Bank, Rozemuller A, Uitterlinden AG, van Meurs JG, van Swieten JC.; *Neurobiol Aging.* 2019 Feb;74:225-233. doi: 10.1016/j.neurobiolaging.2018.10.023. Epub 2018 Oct 29.
34. **EIF2AK3 variants in Dutch patients with Alzheimer's disease;** Wong TH, van der Lee SJ, *van Rooij JGJ*, Meeter LHH, Frick P, Melhem S, Seelaar H, Ikram MA, Rozemuller AJ, Holstege H, Hulsman M, Uitterlinden A, Neumann M, Hoozemans JJM, van Duijn CM, Rademakers R, van Swieten JC.; *Neurobiol Aging.* 2019 Jan;73:229.e11-229.e18. doi: 10.1016/j.neurobiolaging.2018.08.016. Epub 2018 Aug 24.
35. **Exome Sequencing Analysis Identifies Rare Variants in ATM and RPL8 That Are Associated With Shorter Telomere Length;** van der Spek A, Warner SC, Broer L, Nelson CP, Vojinovic D, Ahmad S, Arp PP, Brouwer RWW, Denniff M, van den Hout MCGN, *van Rooij JGJ*, Kraaij R, van IJcken WFJ, Samani NJ, Ikram MA, Uitterlinden AG, Codd V, Amin N, van Duijn CM.; *Front Genet.* 2020 Apr 30;11:337. doi: 10.3389/fgene.2020.00337. eCollection 2020.
36. **Clinical and pathologic phenotype of a large family with heterozygous STUB1 mutation;** Mol MO, *van Rooij JGJ*, Brusse E, Verkerk AJMH, Melhem S, den Dunnen WFA, Rizzu P, Cupidi C, van Swieten JC, Donker Kaat L.; *Neurol Genet.* 2020 Mar 23;6(3):e417. doi: 10.1212/NXG.0000000000000417. eCollection 2020 Jun.
37. **Validation of the BOADICEA model and a 313-variant polygenic risk score for breast cancer risk prediction in a Dutch prospective cohort;** Lakeman IMM, Rodríguez-Girondo M, Lee A, Ruitter R, Stricker BH, Wijnant SRA, Kavousi M, Antoniou AC, Schmidt MK, Uitterlinden AG, *van Rooij J*, Devilee P. *Genet Med.* 2020 Jul 6. doi: 10.1038/s41436-020-0884-4. Online ahead of print.
38. **Reduced penetrance of pathogenic ACMG variants in a deeply phenotyped cohort study and evaluation of ClinVar classification over time;** *van Rooij J*, Arp P, Broer L, Verlouw J, van Rooij F, Kraaij R, Uitterlinden A, Verkerk AJMH. *Genet Med.* 2020 Jul 15. doi: 10.1038/s41436-020-0900-8. Online ahead of print.

39. **Underlying genetic variation in familial frontotemporal dementia: sequencing of 198 patients.** Mol MO, *van Rooij JGJ*, Wong TH, Melhem S, Verkerk AJMH, Kievit AJA, van Minkelen R, Rademakers R, Pottier C, Kaat LD, Seelaar H, van Swieten JC, Dopfer EGP. *Neurobiol Aging*. 2020 Jul. doi: 10.1016/j.neurobiolaging.2020.07.014. Online ahead of print.
40. **C9orf72, AAO and ancestry help discriminating behavioural from language variants in FTLD cohorts.** Costa B, Manzoni C, Bernal-Quiros M, Kia DA, Aguilar M, Alvarez I, Alvarez V, Andreassen O, Anfossi M, Bagnoli S, Benussi L, Bernardi L, Binetti G, Blackburn D, Boada M, Borroni B, Bowns L, Bråthen G, Bruni AC, Chiang HH, Clarimon J, Colville S, Conidi ME, Cope TE, Cruchaga C, Cupidi C, Di Battista ME, Diehl-Schmid J, Diez-Fairen M, Dols-Icardo O, Durante E, Flisar D, Frangipane F, Galimberti D, Gallo M, Gallucci M, Ghidoni R, Graff C, Grafman JH, Grossman M, Hardy J, Hernández I, Holloway GJ, Huey ED, Illán-Gala I, Karydas A, Khoshnood B, Kramberger MG, Kristiansen M, Lewis PA, Lleó A, Madhan GK, Maletta R, Maver A, Menendez-Gonzalez M, Milan G, Miller B, Mol MO, Momeni P, Moreno-Grau S, Morris CM, Nacmias B, Nilsson C, Novelli V, Öijerstedt L, Padovani A, Pal S, Panchbhaya Y, Pastor P, Peterlin B, Piaceri I, Pickering-Brown S, Pijnenburg YA, Puca AA, Rainero I, Rendina A, Richardson AM, Rogaeva E, Rogelj B, Rollinson S, Rossi G, Rossmeyer C, Rowe JB, Rubino E, Ruiz A, Sanchez-Valle R, Sando SB, Santillo AF, Saxon J, Scarpini E, Serpente M, Smirne N, Sorbi S, Suh E, Tagliavini F, Thompson JC, Trojanowski JQ, Van Deerlin VM, Van der Zee J, Van Broeckhoven C, *van Rooij J*, Van Swieten JC, Veronesi A, Vitale E, Waldö ML, Woodward C, Yokoyama J, Escott-Price V, Polke JM, Ferrari R; International FTD-Genetics Consortium (IFGC). *Neurology*. 2020 Sep 17. doi: 10.1212/WNL.0000000000010914. Online ahead of print.
41. **Novel TUBA4A variant associated with familial frontotemporal dementia.** M.O. Mol, T.H. Wong, S. Melhem, A.J.M. Rozemuller, Netherlands Brain Bank, C. Fallini, J.E. Landers, L. Donker Kaat, H. Seelaar, *J.G.J. van Rooij*, J.C. van Swieten. Under review at *Neurology; Genetics*, Oct-2020.
42. **Distinctive pattern of temporal atrophy in patients with frontotemporal dementia and the I383V variant in TARDBP.** M.O. Mol, S.W.R Nijmeijer, *J.G.J. van Rooij*, R.M.L. van Spaendonk, Y.A.L. Pijnenburg, S.J. van der Lee, R. van Minkelen, L. Donker Kaat, A.J.M. Rozemuller, M.R. Janse van Mantgem, W. van Rheenen, M.A. van Es, J.H. Veldink, F.A.M. Hennekam, M.W. Vernooij J.C. van Swieten, P.E. Cohn-Hokke, H. Seelaar, E.G.P. Dopfer. Under review at *JNNP*, Oct-2020.
43. **Genotyping On ALL patients (GOALL); clinical implementation of high-throughput genotyping arrays.** B. Vijlbrief, D. Houtman, J.E.Klapwijk. L. Broer, J.M.H. Verkerk, H. Adams, M.F. van Dooren, R. Hofstra, A.G. Uitterlinden, S.R. Riedijk, *J.G.J. van Rooij*. Manuscript in preparation, Oct-2020.
44. **The genetics of atypical femur fractures – a systematic review.** Wei Zhou, *Jeroen G.J. van Rooij*, Peter R. Eberling, Annemieke J.M.H. Verkerk, M. Carola Zillikens. Manuscript in preparation, Oct-2020.



45. **Polygenic Risk Score and its potential to improve diagnostic ability in knee and hip osteoarthritis.** B. Sedaghati-khayat, C. Boer, L. Broer, J. Verkerk, J. Runhaar, S. Bierma-Zeinstra, *J. van Rooij*, J. van Meurs. Manuscript in preparation, Oct-2020.
46. **Detecting mutations in breast cancer genes by GSA and PMDA array platforms/analysis.** J. de Vries, L. Broer, A. van de Ouweland, C. Vallerga, M. Collee, B. Sedaghati-khayat, M. Honing, J. van Meurs, A. Uitterlinden, *J. van Rooij*, J. Verkerk. Manuscript in preparation, Oct-2020.
47. **Modeling of the cartilage pathology in MPS VI using human induced pluripotent stem cells.** M. Broeders, *J.G.J. van Rooij*, T.J.M. van Gestel, CA Smith, SJ Kimber, Esmee Oussoren, JMP van den Hout, AT van der Ploeg, Roberto Narcisi, WWM Pijnappel. Manuscript in preparation, Oct-2020.
48. **Genetics subtype- and cell type- specific protein signatures in FrontoTemporal Dementia.** Suzanne SM Miedema, Frank TW Koopmans, Pim van Nierop, Kevin Menden, Christina F de Veij Mestdagh, *Jeroen van Rooij*, Andrea B Ganz, Merel Mol, Iryna Paliukhovich, Shamiram Melhem, RiMOD-FTD (Risk and Modifying factors in Fronto-Temporal Dementia) consortium, Ka Wan Li, Henne Holstege, Patrizia Rizzu, Ronald E van Kesteren, John C van Swieten, Peter Heutink, August B Smit. Manuscript in preparation, Oct-2020.
49. **Somatic TARDBP variants as cause of Semantic Dementia.** *Jeroen van Rooij*, Merel Mol1 Shamiram Melhem, Pelle van der Wal, Pascal Arp, Francesca Paron, Laura Donker Kaat, Harro Seelaar, Netherlands Brain Bank, Suzanne SM Miedema, Takuya Oshima, Bart JL Eggen, André Uitterlinden, Joyce van Meurs, Ronald E van Kesteren, August B Smit, Emanuele Buratti, John C van Swieten. Paper in press, BRAIN, Dec-2020.



## 6.5. Dankwoord

It should be obvious that this thesis was only possible due to the contribution of many, many people. Each person having added something to one of the projects, an experiment, analysis, insight or just by being part of some of the many discussions surrounding this work. But also people whom provided insights or suggestions surrounding personal development; how to deal with certain people or issues, how to organize you time and work, which things to pursue and when to know to let something go. Although those aspects are not directly represented in this thesis, I feel that this has been the most valuable result of these last years. For this, I would like to thank everyone involved, and apologize if any name was forgotten, they are so many. *Thank you all!*

Geachte Prof. Uitterlinden, beste **André**, vanzelfsprekend mag ik jou als eerste bedanken, voor het vertrouwen dat je me gaf toen ik tien jaar geleden bij jouw groep begon. Onder jouw begeleiding mocht ik doorstromen van bio-informaticus tot PhD-student, en nu ook als Postdoc. Jij stond altijd klaar met opbouwende suggesties en aanvullende ideeën; alles is leuk en interessant, alles is mogelijk en vergewist van jouw support had ik alle vrijheid om mijn eigen werk te ontplooiën. Het plezier dat je toont in je werk is aanstekelijk, en toont dat je welzeker van je hobby je werk kunt maken. Dank voor al je ondersteuning.

Geachte Prof. van Swieten, beste **John**, onze samenwerking was als de eerste race in een nieuwe auto. Eerst wat onwennig, maar na wat aanpassingen aan de motor komt daar toch een mooi resultaat. Jij bent altijd bezig met hoe het onderzoek praktisch iets kan bijdragen voor de patiënt, hoe ingewikkeld het ook wordt, er moet altijd een praktisch nut aan zitten. Deze focus is een van de belangrijkste lessen die ik heb geleerd van onze samenwerking, en zal mij altijd bijblijven. Dank voor al het vertrouwen dat je me gaf.

Geachte Dr. van Meurs, beste **Joyce**, aan jou de ondankbare taak mijn wilde ideeën, ingewikkelde plaatjes en andere afleidingen weer terug de goede richting op te sturen. Hoewel voor jou soms vermoeiend, was het voor mij ontzettend waardevol mijn ideeën op iemand uit te kunnen proberen; jij was altijd de persoon naar wie ik als eerste ging met problemen of vragen, en gelukkig altijd bereid te helpen. Dank voor al je geduld.

Dear Prof. Ikram, Prof Smit. and Prof. Kushner, dear Arfan, Guus and Steven. Thank you for participating in my reading committee and taking the time to read and comment on this thesis, this is really appreciated. Beste **Arfan**, jouw kennis van alles dat met epidemiologie te maken heeft is indrukwekkend, en ik heb altijd genoten van onze samenwerking en jou altijd nuttige en vriendelijke suggesties. Beste **Guus**, onze discussies over dynamische data analyses en interpretatie waren zodanig stimulerend, dat ik het niet erg vond ervoor naar Amsterdam te komen. Dear **Steven**, I have always been impressed by the work you do in your lab and your beautiful and well-delivered presentations.

Dear Prof. Buratti, Prof. Hardy and Dr. Seelaar, thank you all for participating in my thesis committee. I'm glad you agreed to be part of my defense, and your input on our collaboration over the years. Dear **Emanuele**, your kind and immediate responses to anything we send your way are great, we could not have completed our SD manuscript without your expertise

and inputs. Dear **John**, we only met a few times, but I really enjoyed our interactions during the (PER)ADES meetings, you have some of the best anecdotes. Beste **Harro**, het is me soms een raadsel hoe je zo warrig kunt zijn en toch zoveel gedaan krijgt, het is altijd fijn om je erbij te hebben.

Beste **Martin**, ten eerste zeer veel dank dat je mijn paranimf wilde zijn, er was geen twijfel wie ik hier als eerste voor zou vragen. Het is een geruststellende gedachte dat je er altijd bent, of ik nu stoom af moet blazen of even niet weet wat ik ergens mee moet doen, je hoort het altijd rustig aan en probeert daarna mee te denken aan een oplossing. Je vriendschap is zeer belangrijk voor mij. Beste **Dennis**, wat ben ik blij dat je al die jaren geleden onze groep koos voor je stage. Jij was mijn eerste student, en ik denk dat ik net zo veel van jou heb geleerd als andersom. Hoewel we elkaar niet iedere week zien, beschouw ik je al een goede vriend. Jij hebt dezelfde domme en simpele humor als ik. Ik ben erg dankbaar dat je dit samen met me wilde doen, en kijk uit naar je eigen promotie. Beste **Puck**, jij bent een van de vriendelijkste mensen die ik ken, ook al zul je dat zelf niet zo zien. Het klimmen en onze trips daarin samen zijn een groot deel van mijn leven geworden, en ik ben blij dat ik dit samen met jou kan doen. Hopelijk duurt het nog even voordat alles hebben uitgeklimmen. Lieve **Iris**, ondanks de tegenslagen laatste jaren zet jij altijd door, en kom je er uiteindelijk sterker uit. Ik heb altijd erg genoten van onze klimsessies, en hoop dat we deze snel weer mogen hervatten. Lieve **Anja**, het leven gaat soms snel, alles veranderd, maar na 20 minuten koffie met Anja is alles weer zoals vroeger. Dank je voor je altijd frisse blik tijdens onze koffiemomenten of drankjes bij suf. Liefs ook voor Duncan en Noah. Beste **Djawad**, de barbecues bij je zelfgebouwde huis zijn altijd heerlijk. Hopelijk mogen er nog vele volgen. Veel liefs ook voor Eliza en Susan.

Lieve **Shami**, het is bijzonder hoeveel jij doet voor ons onderzoek. In je eentje voorzie je verschillende onderzoeklijnen van experimenten, en altijd met een lach. Je vrolijkheid is altijd erg aanstekelijk, en ik denk met plezier terug aan onze tijd in het lab. Zonder jou had dit proefschrift er niet zo geweest. Weet ook dat het idee van mijn kaft ontstond omdat ik de mensen achter de schermen, waarbij ik met name aan jou dacht, een prominente plek wilde geven. Blijf altijd awesome! Beste **Annemieke**, jij bezit het vermogen om mensen op hun gemak te stellen en hun zorgen met je te delen. Iets waar ik, en ik denk vele anderen, met enige regelmaat gebruik van heb gemaakt. Je bent een fijn mens, en, ondanks dat onze stijlen soms wat clashen, denk ik dat er erg mooie dingen uit kunnen komen. Ons artikel is een van de projecten waar ik met het meeste plezier aan heb gewerkt. Beste **Joost**, na Stephan heb jij het sequencing werk overgenomen. Je verzet erg veel en goed werk achter de schermen, wat vaker genoemd zou mogen worden. Ik vind dat je je werkt goed doet, en het is prettig te weten dat je op zo iemand kunt terugvallen. Beste **Marijn**, ook voor jou geldt dat je ontzettend veel werk verzet, wat niet altijd voldoende gewaardeerd wordt. Zolang alles werkt hoor je niets, en zodra het misloopt wordt er snel veel geklaagd, iets waar ik mijzelf, helaas, ook schuldig aan maak. Vrijwel geen enkele analyse uit dit boek had gekund zonder jou inspanningen, waarvoor enorm veel dank. Beste **Ramazan**, altijd als ik even pauze nodig heb kijk ik of ik jou ergens van je werk kan houden. Onze gesprekken zijn zelden nuttig, maar altijd gemakkelijk, en misschien is dat ook weer nuttig.

Bedankt ook aan mijn kamergenootjes van 2238. Ondanks mijn gemopper soms vind ik het altijd erg gezellig. Beste **Tsz**, ik ben enorm onder de indruk van hoe je jezelf altijd nieuwe skills aanleerde, waaronder alles in de genetica. Dank dat je aan mij dacht bij je eigen promotie, en hopelijk kom je snel terug in Rotterdam voor een spelletje, als Emily het goed vindt. Lieve **Leonie**, het is gek je niet meer tegenover me te hebben zitten, of om samen op vrijdag (iets te laat) te gaan klimmen. Hopelijk blijf je af en toe aanhaken, en anders kunnen we Mike vast strikken voor een barbecue. Veel liefs ook voor Mike en Liva. Lieve **Merel**, ik vind het ontzettend fijn met je samen te werken, je doet het fantastisch, ondanks je rare begeleider. Lieve **Emma, Jacky, Jessica**, dank voor de gezelligheid op onze kamer, en het aanhoren van mijn geklaag tussen meetings. Hetzelfde voor mijn voorgaande of nieuwe kamergenootjes; **Dina, Elise, Janneke, Judy, Lauren, Lieke, Lize, Lucia, Lynn, Marjolein** en **Sophie**, het was er nooit saai (of stil). Uiteraard ook dank voor de bijdrage van alle andere mensen bij of rondom het Alzheimercentrum; **Amel, Dorothee, Eric, Fenne, Frank-Jan, Hannah, Janne, Laura, Rebecca, Sanne, Willem**.

Then, all of my colleagues from the genetic lab, there are so many, it is hard to decide where to start. Thankfully, when I'm not sure what to do I always find Eline, who has all the answers. Dear **Eline**, thank you for all your help surrounding my PhD, it is very comforting to have you around. Dear **Robert**, you are who brought me in to internal medicine and guided me the first years into the field of sequencing. I may not be the easiest person to supervise, but I think you did very well. Dear **Bahar**, sometimes I wonder how far you get from my office before you happily get distracted by a new idea or result. Your enthusiasm is admirable, catching and never boring, please don't let my grumpiness diminish that. Dear **Vivi**, I have a lot of respect for your willingness to learn new things and accept criticism. There are few people with whom I can be this direct, and I look forward to seeing you develop your skills in the future. Next, I would like to thank our technicians; **Michael, Mila, Pascal, Pawan, Pelle, Ramazan, Sarah, Sergio, Thomas** and **Zara**. You are the backbone of this whole group. How you all keep the facility running, develop novel methods in the lab and participate in the research meetings and project is exemplary to working as a team. I had the pleasure to work in the lab during my PhD, and would be completely unable to do what you do in such an organized and fast manner. Followed by the (bio-)informatics people; **Costanza, Djawad, Linda, Jard, Joost, Joost, Marijn** and **Stephan**. All scientists agree that you need good data to get good results, but few realize the subtlety and skill that goes into creating good data. Your tireless efforts ensure that our researchers and our clients are fed with the best data available. Then, all the (former) PhD-students and PostDocs that analyze and interpret all this data, resulting in so many interesting discussions and publications; **Annelies, Ariadne, Artemis, Carolina, Cindy, Denise, Elisabeth, Enisa, Ester, Fatimeh, Fjorda, Gaby, Ingrid, JinLuan, Josje, Katerina, Komal, Lieke, Ling, Lisette, Marjolein, Marjolein, Masa, Melissa, Natalia, Olja, Pooja, Ruolin, Samuel, Vid** and **Zografia**. Finally, all these efforts must be coordinated, in addition to the people already mentions; Dear **Fernando**, your drive and science is excellent, as well as your taste in beers, I look forward to another basketball game, maybe next time a little more coordinated. Dear **Carola**, you bring a much-appreciated clinical view to our research, and I like how you don't shy away from complicated topics. Thank you for your trust by involving you in part of your work.

Dan, nog een aantal mensen van verschillende afdelingen en centra, die over de jaren hebben bijgedragen aan mijn promotie-werk. Van de afdeling interne geneeskunde; **Anela, Annelies, Bram, Edward, Eline, Frank, Franka, Gretchen, Jenny, Jeroen, Leo, Marcel, Marjolein, Patrick, Rens, Robin, Samantha, Sander, Theo, Marcel** en alle andere (ex-)collega's, dank voor alle borrels, labdagen en andere gezelligheid. Van de epi, **Frank**; vanwege al het fantastische werk je doet met onze dataverzamelingen. **Abbas, Albert, Annemarie, Ashley, Bruno, Cornelia, Janine, Maryam, Mohsen, Najaf, Shazhad** en **Trudy** en alle andere onderzoekers en deelnemers van de ERGO studie. In het specifiek, **Hieab, Paul, Sven** en **Symen**, voor de boeiende discussies en ideeën tijdens verschillende binnen- en buitenlandse meetings. Onze trip in LA was memorabel. Verder mijn collega's vanuit het BIOS consortium; **Annie, Bas, Dasha, Freerk, Jenny, Leon, Lude, Maarten, Peter-Bram, Rick**, en **Maarten**, dank voor al het mooie en belangrijke werk dat jullie daar verzetten. Beste **Bas**, jouw enthousiasme met betrekking tot complexe genomische analyses is aanstekelijk, ik heb erg genoten van onze samenwerkingen rondom de BIOS data. To my colleagues from the CHARGE and Sanford collaborations; **Bruce, Chris, Dylan, Henry, James, Jerry, Jim** and **Cassie** and all those other people doing fantastic work, the CHARGE meetings are a highlight each year. Thanks to the people from the AD genetics consortia; **Alfredo, Henne, Iris** and **Marc**. As well as from the FTD genomic collaborations; **Raff, Cyril, Francesca, Javier, Kawan Peter, Pim, Raff, Ronald, Rosa** and **Suzanne**.

Een laatste groep, voor mij van persoonlijk belang; alle studenten die het hebben aangedurfd onder mijn begeleiding te werken; **Annelies, Bahar, Coco, Dennis, Emilia, Joost, Mariella, Madina, Merel, Merel, Michiel, Robert, Robin, Tom, Theo, Simone** en **Vivi**. Jullie hebben mij geleerd dat iedere persoon anders benaderd moet worden, en dat iedereen kennis bevat die je zelf niet hebt, wat gewaardeerd moet worden.

Als laatste, mijn familie; Pa, Ma, Erik, Saskia, weet dat jullie nooit ver zijn van mijn gedachten, de wens om jullie trots te maken is mijn grootste motivatie. Lieve Opa, Oma, Annemieke, jullie waren vast ook trots geweest.

## 6.6. Abbreviations

AB	amyloid beta
AD	Alzheimer's Disease
bvFTD	behavioral variant FTD
CBS	cortical basal syndrome
CpG	adjacent CG on the genome, can be methylated
CSF	cerebral spinal fluid
DLB	Dementia with Lewy-Bodies
DN	dystrophic neurites, form of TDP pathology
DNA	deoxyribonucleic acid
FTD	FrontoTemporal Dementia
FUS	fused-in sarcoma protein, product of FUS
GWAS	genome-wide association study
MND	motor neuron disease
MRI	magnetic resonance imaging
NCI	neuronal cytoplasmic inclusions, form of TDP pathology
nfPPA	non-fluent primary progressive aphasia
NFT	neurofibrillary tangle, form of tau pathology
NGS	next generation sequencing
NII	neuronal intranuclear inclusions, form of TDP pathology
NHB	Netherlands brain bank
PCR	polymerase chain reaction
PD	Parkinson's Disease
PET	positron emission tomography
PSP	progressive supranuclear palsy
RNA	ribonucleic acid
RS	Rotterdam study
SD	semantic dementia
svFTD	semantic variant FTD
TAU	tau protein, product of MAPT
TDP43	tar-dna binding protein, product of TARDBP
VD	vascular dementia
WES	whole exome sequencing
WGS	whole genome sequencing





