

# **Disease progression and genetic risk factors in the primary tauopathies**

*Dr William Johnson Scotton*

A dissertation submitted in partial fulfilment of the requirements for the degree of

**Doctor of Philosophy**

of

**University College London**

UCL Queen Square Institute of Neurology

26 February 2023



I, William Johnson Scotton, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.



**ਸੰਗੀਤਾ, ਡਾਇਲਨ ਅਤੇ ਈਵਾ ਲਈ। ਤੁਸੀਂ ਹੀ ਮੇਰੀ ਦੁਨੀਆ ਹੋ।**

Sagītā, ḍā'ilana atē īvā la'ī. Tusīm hī mērī dunī'ā hō.



# Abstract

The primary tauopathies are a group of progressive neurodegenerative diseases within the frontotemporal lobar degeneration spectrum (FTLD) characterised by the accumulation of misfolded, hyperphosphorylated microtubule-associated tau protein (MAPT) within neurons and glial cells. They can be classified according to the underlying ratio of three-repeat (3R) to four-repeat (4R) tau and include Pick's disease (PiD), which is the only 3R tauopathy, and the 4R tauopathies the most common of which are progressive supranuclear palsy (PSP) and corticobasal degeneration (CBD). There are no disease modifying therapies currently available, with research complicated by the wide variability in clinical presentations for each underlying pathology, with presentations often overlapping, as well as the frequent occurrence of atypical presentations that may mimic other non-FTLD pathologies. Although progress has been made in understanding the genetic contribution to disease risk in the more common 4R tauopathies (PSP and CBD), very little is known about the genetics of the 3R tauopathy PiD.

There are two broad aims to this thesis; firstly, to use data-driven generative models of disease progression to try and more accurately stage and subtype patients presenting with PSP and corticobasal syndrome (CBS, the most common presentation of CBD), and secondly to identify genetic drivers of disease risk and progression in PiD. Given the rarity of these disorders, as part of this PhD I had to assemble two large cohorts through international collaboration, the 4R tau imaging cohort and the Pick's disease International Consortium (PIC), to build large enough sample sizes to enable the required analyses.

In **Chapter 3** I use a probabilistic event-based modelling (EBM) approach applied to structural MRI data to determine the sequence of brain atrophy changes in clinically diagnosed PSP - Richardson syndrome (PSP-RS). The sequence of atrophy predicted by the model broadly mirrors the sequential spread of tau pathology in PSP post-mortem staging studies, and has potential utility to stratify PSP patients on entry into clinical trials based on disease stage, as well as track disease progression.

To better characterise the spatiotemporal heterogeneity of the 4R tauopathies, I go on to use Subtype and Stage Inference (SuStaIn), an unsupervised machine algorithm, to identify population subgroups with distinct patterns of atrophy in PSP (**Chapter 4**)

and CBS (**Chapter 5**). The SuStaIn model provides data-driven evidence for the existence of two spatiotemporal subtypes of atrophy in clinically diagnosed PSP, giving insights into the relationship between pathology and clinical syndrome. In CBS I identify two distinct imaging subtypes that are differentially associated with underlying pathology, and potentially a third subtype that if confirmed in a larger dataset may allow the differentiation of CBD from both PSP and AD pathology using a baseline MRI scan.

In **Chapter 6** I investigate the association between the *MAPT* H1/H2 haplotype and PiD, showing for the first time that the H2 haplotype, known to be strongly protective against developing PSP or CBD, is associated with an increased risk of PiD. This is an important finding and has implications for the future development of *MAPT* isoform-specific therapeutic strategies for the primary tauopathies.

In **Chapter 7** I perform the first genome wide association study (GWAS) in PiD, identifying five genomic loci that are nominally associated with risk of disease. The top two loci implicate perturbed GABAergic signalling (*KCTD8*) and dysregulation of the ubiquitin proteasome system (*TRIM22*) in the pathogenesis of PiD.

In the final chapter (**Chapter 8**) I investigate the genetic determinants of survival in PiD, by carrying out a Cox proportional hazards genome wide survival study (GWSS). I identify a genome-wide significant association with survival on chromosome 3, within the *NLGN1* gene, which encodes a synaptic scaffolding protein located at the neuronal pre-synaptic membrane. Loss of synaptic integrity with resulting dysregulation of synaptic transmission leading to increased pathological tau accumulation is a plausible mechanism through which NLGN1 dysfunction could impact on survival in PiD.



# Impact Statement

The primary tauopathies (FTLD-tau) are a devastating group of neurodegenerative disorders that take a terrible physical and emotional toll on both patients and their families. Given their midlife onset, these diseases cause a dramatic reduction in life expectancy, resulting in a significant economic impact both for affected families, and the wider economy in general. The loss of productivity for both patients and carers from being pulled out of the workforce, often at the height of their careers, has led to estimates of the economic costs associated with these diseases that are nearly two times higher than for Alzheimer's disease. There are no curative treatments for the primary tauopathies at present, and there is an urgent need to develop new therapies, both from a symptomatic and a disease-modifying perspective.

In the first part of this thesis, I investigate the spatiotemporal heterogeneity of the 4-repeat tauopathies, PSP and CBD, using generative data-driven disease progression models, to try and better stratify these diseases based on both subtype and disease stage at baseline. In the second part of the thesis, I focus on the understudied 3-repeat tauopathy Pick's disease, using both targeted and genome wide approaches to identify genetic determinants of disease risk and progression.

The findings from this thesis are important with direct relevance for the academic community clinicians and for patients and their families. Firstly, I provide evidence that the 4R tauopathies can be accurately staged and subtyped using only baseline structural MRI scans. These findings have direct clinical utility as they are based on MRI data easily available to clinicians, and provide an objective measure of disease status with important implications for stratifying patients into homogenous groups for future clinical trials. Secondly, by building a large international collaboration (the PIC), I am able to show that the *MAPT* H2 haplotype is associated with the risk of Pick's disease, and go on to perform the first genome-wide association studies of both disease risk and survival, identifying novel loci that implicate genes directly involved in disease pathogenesis. These results will be important for the PiD academic community, as they provide directly testable hypotheses for future functional studies in animal and cell models, to probe affected biological pathways. In addition, the PIC, an international collaboration between nearly 40 leading research centres in this disease globally, provides a network to enable coordination of future sample

collection, development of consensus diagnostic criteria, and studies that may enable us to finally find a treatment to prevent this devastating disease.

The results from this thesis have been, or will be, widely disseminated both nationally and internationally through conference presentations, and open-access publications in peer-reviewed journals. The 4R tauopathy disease progression modelling has led to active collaborations between UCL, Manchester, Oxford, Cambridge and UCSF in the United States, with future plans to apply these models to autopsy confirmed Pick's disease imaging in collaboration with UCSF, UPenn and Banner. In addition, I plan to build a staging model for autopsy confirmed CBD using subtype and stage inference modified to use ordinal pathological data, in collaboration with Queen Square Brain Bank. Collaboration through the PIC, with the Mayo Clinic in Florida is also currently ongoing, with future studies planned including the development of tau specific RT-QuIC tracers, and long read sequencing of the *MAPT* locus to better understand the impact of tau isoform ratios in development of the different tauopathies.

# Publications

1. Scotton WJ, Bocchetta M, Todd E, Cash DM, Oxtoby N, VandeVrede L, Heuer H; PROSPECT Consortium, 4RTNI Consortium; Alexander DC, Rowe JB, Morris HR, Boxer A, Rohrer JD, Wijeratne PA. A data-driven model of brain volume changes in progressive supranuclear palsy. *Brain Commun.* 2022 Apr 14;4(3):fcac098.359

UCL Research Paper Declaration Form referencing the doctoral candidate's own published work(s)
---------------------------------------------------------------------------------------------------

1. **For a research manuscript that has already been published** (if not yet published, please skip to section 2)

a) **What is the title of the manuscript?**

A data-driven model of brain volume changes in progressive supranuclear palsy

b) **Please include a link to or doi for the work** 10.1093/braincomms/fcac098

c) **Where was the work published?** Brain Communications

d) **Who published the work?** OUP

e) **When was the work published?** April 14th 2022

f) **List the manuscript's authors in the order they appear on the publication**

W. J. Scotton, M. Bocchetta, E. Todd, D. M. Cash, N. Oxtoby, L. VandeVrede, H. Heuer, PROSPECT Consortium, 4RTNI Consortium, D. C. Alexander, J. B. Rowe, H. R. Morris, A. Boxer, J. D. Rohrer, and P. A. Wijeratne

g) **Was the work peer reviewed?** Yes

h) **Have you retained the copyright?**

Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>)

i) **Was an earlier form of the manuscript uploaded to a preprint server?**

(e.g. medRxiv). If 'Yes', please give a link or doi)

No

If 'No', please seek permission from the relevant publisher and check the box next to the below statement:



*I acknowledge permission of the publisher named under **1d** to include in this thesis portions of the publication named as included in **1c**.*

- 2. For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)

n/a

- 3. For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)

**Conception and design:** WJS, JR, PAW; **Data Collection:** WJS, MB, ET, DC, NO, LV, HH, DCA, JBR, AB, HRM, JR, PAW, PROSPECT Consortium, 4RTNI Consortium; **Execution:** WJS ; **Analysis:** WJS; **Interpretation:** WJS, HM, JR, PAW; **Writing Manuscript:** WJS; **Critical review of manuscript:** WJS, MB, ET, DC, NO, LV, HH, DCA, JBR, AB, HRM, JR, PAW

- 4. In which chapter(s) of your thesis can this material be found?**

Chapter 3

- 5. e-Signatures confirming that the information above is accurate**

*Candidate*

William Scotton

*Date:* 26/02/23

*Supervisor/ Senior Author (where appropriate)*

Peter Wijeratne

*Date* 26/02/23

2. Scotton WJ, Shand C, Todd E, Bocchetta M, Cash DM, VandeVrede L, Heuer H; PROSPECT Consortium, 4RTNI Consortium; Young AL, Oxtoby N, Alexander DC, Rowe JB, Morris HR, Boxer AL, Rohrer JD, Wijeratne PA. Uncovering spatiotemporal patterns of atrophy in progressive supranuclear palsy using unsupervised machine learning. Brain Commun. 2023 Mar 2;5(2):fcad048.

UCL Research Paper Declaration Form  
referencing the doctoral candidate's own published work(s)

1. **For a research manuscript that has already been published** (if not yet published, please skip to section 2)

n/a

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)

- a) **What is the current title of the manuscript?**

Uncovering spatiotemporal patterns of atrophy in progressive supranuclear palsy using unsupervised machine learning A data-driven model of brain volume changes in progressive supranuclear palsy

- b) **Has the manuscript been uploaded to a preprint server?** (e.g. medRxiv; if 'Yes', please give a link or doi)

No

- c) **Where is the work intended to be published?** (e.g. journal names)

Brain Communications

- d) **List the manuscript's authors in the intended authorship order**

W.J. Scotton, C. Shand, E. Todd, M. Bocchetta, D. M. Cash, L. VandeVrede, H. Heuer, PROSPECT Consortium, 4RTNI Consortium, A.L. Young, N. Oxtoby, D.C. Alexander, J.B. Rowe, H.R. Morris, A.L. Boxer, J.D. Rohrer,1, P. A. Wijeratne

- e) **Stage of publication** (e.g. in submission)

Accepted for publication – awaiting DOI and publication date

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)

**Conception and design:** WJS, JR, PAW; **Data Collection:** WJS, MB, ET, DMC, NO, LV, HH, JBR, AB, HRM, JR, PROSPECT Consortium, 4RTNI Consortium; **Execution:** WJS; **Analysis:** WJS; **Plotting figures:** WJS, CS **Interpretation:** WJS, ALY, HM, JR, PAW; **Writing Manuscript:** WJS; **Critical review of manuscript:** WJS, CS, MB, ET, DMC, ALY, NO, LV, HH, DCA, JBR, AB, HRM, JR, PAW

**4. In which chapter(s) of your thesis can this material be found?**

Chapter 4

**5. e-Signatures confirming that the information above is accurate** (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)

*Candidate*

William Scotton

*Date:*

26/02/23

*Supervisor/ Senior Author (where appropriate)*

Peter Wijeratne

*Date*

26/02/23

## Acknowledgements

I would like to start by thanking all of the patients and their families from countries around the world who gave their precious time to the studies that provided data for this PhD. I hope that some of the work included here, even if it is in a small way, contributes to future breakthroughs that are so badly needed to treat these devastating diseases. I also wanted to express my gratitude to my funders the Wellcome Trust, and the UCL Academic Clinical Office, for taking a chance on me and financially supporting this PhD.

The best advice I was given before starting this PhD was from Dr. Mary O'Driscoll in the summer of 2019, who simply said "*It can be very lonely; make sure you build a support network to get you through the difficult times*". The wisdom of this advice only really became apparent in mid-April 2020 when the Covid pandemic was in full flow. As the years have gone by this advice has only become more pertinent, and now I have reached the end and had time to reflect, I realise how lucky I have been to have the support and mentorship of some truly inspirational people.

Firstly, I want to thank my primary supervisor Professor Sir John Hardy. It has been such a privilege to work with you these last few years, and to see your life's work recognised during this period with a knighthood. You have only ever shown me kindness, respect and unwavering support. You have rekindled the passion I had as a younger man for pure scientific enquiry, where someone's worth is not measured by their status or position, but purely by their ideas. As Montaigne once said, "*A virtuous, ordinary life, striving for wisdom but never far from folly, is achievement enough*".

Thank you, Professor Jonathan Rohrer, for welcoming me into your lab and giving me insight into cognitive neurology and clinical research. Your positive approach and sage advice have helped me get through some difficult times, and although things did not necessarily work out as expected, your support for me has never diminished.

Thank you to Professor Huw Morris for teaching me about the genetic and clinical aspects of the atypical parkinsonian syndromes, and supporting me, through your lab, with the genetic analyses in this PhD. And thank you to Dr. Martina Bocchetta and Dr. Andre Altmann for agreeing to be on my thesis committee, guiding and advising me through the process.

To Dr. Peter Wijeratne (“the Mighty PAW”); you are an inspirational person and have become a good friend. For the first six months I literally had no idea what you were talking about, but through the medium of Python I gradually started to glimpse the kernel of meaning. Thank you for the laughter, the words left unsaid (mainly in supervisor meetings) and for the friendship that helped me through the long months of the pandemic. The University of Sussex don’t know how lucky they are. I look forward to our future collaborations and in particular developing the HTTMM.

Thank you to Dr. Maryam Shoai for your friendship and support. You have always been on hand to give advice whenever I have needed, and patiently walk me through all of the different genetic analyses used in this thesis. It has been inspiring to work with someone with your level of intellect and humanity; I understand why John keeps you so close.

Thank you to Dr. Raquel Real for always being there to help, and provide code when I have needed it. You didn’t need to do this, but you did nonetheless despite your own busy schedule which reveals the type of person you truly are. I have been very lucky to work with you.

Thank you to the people at the Institute of Neurology who helped me at various points during the PhD; Dr Kin Mok, Alejandro Martinez-Carrasco and Hannah Macpherson. Thank you to Professor Tammarn Lashley for your supervision both as my research graduate tutor and also in mid-2020 when performing the immunohistochemistry; I enjoyed our varied conversations and also the company during a period when I rarely saw anyone in person apart from my own family.

Thank you to the POND team at CMIC for taking me under your wing; in particular to Professor Danny Alexander, Dr Neil Oxtoby and Dr Cameron Shand for all the help and words of encouragement.

Thank you to Dr Owen Ross, Dr Rebecca Valentino and the rest of the team at the MCJ, and Dr Adam Boxer and the team at UCSF for collaborating on various projects, and putting your trust in me.

To Ann Pennington, thank you for inviting me into your home, and reminding me of the power of good conversation. Our friendship has been a great comfort to me over the last three years and enabled me to maintain perspective on what is truly important.



Thank you to my oldest friend Nick Amery, for taking me climbing on Wednesday evenings when I was in London, and always being the best company.

To my mother, Sue, and father, John - thank you for always being there for me. You have always given me the love and encouragement to follow what I believe is important in life, and instilled in me the importance of kindness, compassion and the hard work. Words cannot convey the debt I owe to you both, but I hope this work goes a small way to repaying your faith in me.

And finally, to my wife, Sangeeta, and my children, Dylan and Eva. You are my world. Thank you for sharing the journey, and putting up with me over the last three years. Sangeeta, your constant encouragement, positivity, and reassurance has been the glue that has held me together, and none of this would have been possible without you. Dylan and Eva, I am so proud of you both and the people you are growing up to be; if there were any positives from the Covid pandemic it was having the chance to spend more time with you. As Ian Carr once so beautifully put it:

*“.. our children, like our possessions and our lives, are lent not given”*



## Contributions statement

I wanted to acknowledge the support and contributions from the many people that were involved in the work laid out in this thesis. I have personally performed the majority of the laboratory work and bioinformatic/data analysis, and written all chapters (and associated manuscripts) with support from my supervisors. However, given the nature of the cohorts I assembled, there are a large number of people that have been involved in data collection across the world who I fully reference in the **Appendix A** and **B**. In the sections below I will detail, by chapter, the individuals who supported me in the data collection and analysis where relevant.

### **Chapter 3-5: Disease Progression Modelling in the 4R tauopathies**

Data analysed in these chapters were collected from eight different cohorts as detailed in **Chapter 2**, and I would like to thank all of those people involved in participant identification, recruitment and testing (**Appendix A**), as well as the patients and their families. I performed all the data analyses in these chapters with support from Dr. Peter Wijeratne. These chapters were written by me, with input on results interpretation and conclusions from Dr. Peter Wijeratne and Professor Jonathan Rohrer. I would also like to thank Cameron Shand who updated and improved the plotting functions in SuStaIn, which greatly helped in making the figures included. Finally, I would like to recognise Dr. Alexandra Young who developed SuStaIn, and Dr. Neil Oxtoby (along with Dr. Nick Firth) who developed the KDE EBM; both provided trouble shooting support and additional code at various stages of the analyses.

### **Chapter 6-8: Genetic analyses in Pick's Disease**

Data analysed in these chapters were collected as part of the Pick's Disease International Consortium (PIC). A large group of researchers and clinicians, as part of the PIC, contributed to data collection and curation and are detailed in the Acknowledgements section (**Appendix B**). All of the 3R/4R tau immunohistochemistry for the UCL samples was performed by myself under the supervision of Professor Tammarn Lashley at Queen Square Brain Bank, while Professor Dennis Dickson and Dr Shanu Roemer performed the equivalent for the North American samples at the Mayo Clinic Jacksonville (MCJ). I am very grateful to Dr. Kin Mok who took the time to teach me the key steps required to call genotypes

from the Neurobooster and Global Screening arrays in GenomeStudio. I plated all UCL samples with support from Dr Hannah Macpherson and Dr Raquel Real, and UCL Genomics carried out the genotyping, providing the raw IDAT data for further analyses. The North American samples were plated and genotyped at the MCJ, and raw IDAT data was shared with me at UCL for further analyses.

Dr. Michael Heckmann performed the original *MAPT* haplotype analysis in **Chapter 6**, and I am very grateful for him taking the time to explain to me the steps required to reproduce the analysis and providing the code to do so. I wrote the chapter in collaboration with Dr. Rebecca Valentino at the MCJ, under the overall supervision of Dr Owen Ross, Professor Sir John Hardy and Professor Jonathan Rohrer.

I performed the data analyses for **Chapter 7** and **Chapter 8** under the supervision of Dr. Maryam Shoi, with input from Professor Sir John Hardy and Professor Jonathan Rohrer on results interpretation and conclusions. I am very grateful to Alejandro Martinez-Carrasco who helped with the fine-mapping and *Coloc* analysis of the lead loci in these chapters, and also to Dr. Raquel Real who shared code that I modified for some of the analyses. I would also like to acknowledge the GP2 teaching materials online (<https://github.com/GP2-TNC-WG/GP2-Beginner-Bioinformatics-for-PD-Genetics>), that provided an invaluable resource for learning the key steps for genome wide association study (GWAS) analysis. In particular thankyou to Dr Sandra Bandres-Ciga at NIH who answered any questions I had (at any time of the day or night) regarding the pipelines included in this amazing resource.

# Table of Contents

<b>ABSTRACT .....</b>	<b>7</b>
<b>IMPACT STATEMENT .....</b>	<b>9</b>
<b>ACKNOWLEDGEMENTS .....</b>	<b>15</b>
<b>CONTRIBUTIONS STATEMENT .....</b>	<b>19</b>
<b>TABLE OF CONTENTS .....</b>	<b>21</b>
<b>LIST OF TABLES.....</b>	<b>25</b>
<b>LIST OF FIGURES.....</b>	<b>23</b>
<b>CHAPTER 1: INTRODUCTION.....</b>	<b>27</b>
1.1 OVERVIEW .....	27
1.2 AN INTRODUCTION TO FTLD AND THE PRIMARY TAUOPATHIES .....	28
1.3 GENETICS .....	49
1.4 MEASURING DISEASE PROGRESSION .....	57
1.5 THESIS AIMS .....	67
<b>CHAPTER 2: COHORTS AND GENERAL METHODOLOGY .....</b>	<b>69</b>
2.1 INTRODUCTION.....	69
2.2 4R TAUOPATHY DISEASE PROGRESSION MODELLING.....	69
2.3 3R TAUOPATHY (PICK’S DISEASE) GENETIC STUDIES .....	91
<b>CHAPTER 3: A DATA-DRIVEN MODEL OF BRAIN VOLUME CHANGES IN PROGRESSIVE SUPRANUCLEAR PALSY RICHARDSON SYNDROME.....</b>	<b>113</b>
3.1 INTRODUCTION.....	113
3.2 METHODS.....	115
3.3 RESULTS.....	119
3.4 DISCUSSION.....	130
3.5 CONCLUSIONS AND FUTURE WORK.....	134
<b>CHAPTER 4: UNCOVERING SPATIOTEMPORAL PATTERNS OF ATROPHY IN PROGRESSIVE SUPRANUCLEAR PALSY USING UNSUPERVISED MACHINE LEARNING.....</b>	<b>135</b>
4.1 INTRODUCTION.....	135
4.2 METHODS.....	137
4.3 RESULTS.....	149
4.4 DISCUSSION.....	166
4.5 CONCLUSION AND FUTURE WORK .....	171
<b>CHAPTER 5: DISTINCT SPATIOTEMPORAL ATROPHY PATTERNS IN CORTICOBASAL SYNDROME ARE ASSOCIATED WITH DIFFERENT UNDERLYING PATHOLOGIES.....</b>	<b>172</b>
5.1 INTRODUCTION.....	172
5.2 METHODS.....	174

5.3 RESULTS.....	183
5.4 DISCUSSION.....	202
5.5 CONCLUSIONS AND FUTURE WORK.....	207
<b>CHAPTER 6: INVESTIGATING THE ASSOCIATION OF <i>MAPT</i> HAPLOTYPES WITH RISK OF PICK’S DISEASE .....</b>	<b>208</b>
6.1 INTRODUCTION.....	208
6.2 METHODS.....	209
6.3 RESULTS.....	215
6.4 DISCUSSION.....	221
6.5 CONCLUSIONS AND FUTURE WORK.....	223
<b>CHAPTER 7: POTENTIAL GENETIC MODIFIERS OF DISEASE RISK IN AUTOPSY-CONFIRMED PICK’S DISEASE: A GENOME-WIDE ASSOCIATION STUDY .....</b>	<b>225</b>
7.1 INTRODUCTION.....	225
7.2 METHODS.....	226
7.3 RESULTS.....	236
7.4 DISCUSSION.....	258
7.5 CONCLUSIONS AND FUTURE WORK.....	262
<b>CHAPTER 8: GENETIC DETERMINANTS OF SURVIVAL IN PICK’S DISEASE: A GENOME-WIDE ASSOCIATION STUDY .....</b>	<b>264</b>
8.1 INTRODUCTION.....	264
8.2 METHODS.....	265
8.3 RESULTS.....	270
8.4 DISCUSSION.....	291
8.5 CONCLUSIONS AND FUTURE WORK.....	294
<b>CHAPTER 9: GENERAL CONCLUSIONS AND FUTURE DIRECTIONS .....</b>	<b>296</b>
9.1 SUMMARY .....	296
9.2 DISEASE PROGRESSION MODELLING IN THE 4R TAUOPATHIES .....	297
9.3 GENETICS ANALYSES IN THE 3R TAUOPATHY PiD .....	302
<b>APPENDIX A - 4R TAU IMAGING COHORT .....</b>	<b>308</b>
<b>APPENDIX B - PICK’S DISEASE INTERNATIONAL CONSORTIUM (PIC) .....</b>	<b>311</b>
<b>BIBLIOGRAPHY.....</b>	<b>319</b>

# List of Figures

Figure 1.1 - The clinico-pathological spectrum of frontotemporal lobar degeneration (FTLD).....	29
Figure 1.2 - The <i>MAPT</i> gene, transcript isoforms and tau protein structure .....	32
Figure 1.3 - Structure-based classification of tauopathies.....	34
Figure 1.4 - Structure of tau filaments in PiD .....	35
Figure 1.5 - Clinical syndromes in progressive supranuclear palsy .....	38
Figure 1.6 - Contribution of clinical features to different PSP phenotypes .....	40
Figure 1.7 - Hypothetical model of disease courses in PSP syndromes.....	41
Figure 1.8 - Structure of tau filaments in CBD and PSP .....	42
Figure 1.9 - Sequential stages of tau pathology in PSP.....	46
Figure 1.10 - Sequential stages of tau pathology in PiD .....	48
Figure 1.11 - Disease causing MAPT mutations.....	49
Figure 1.12 - MAPT mutations and associated pathology .....	50
Figure 1.13 - MAPT mutation location and association with sporadic FTLD-Tau pathology .....	51
Figure 1.14 - Structure of <i>MAPT</i> 17q21.21 locus .....	53
Figure 1.15 - Overview of current data-driven disease progression models.....	64
Figure 1.16 - Conceptual overview of SuStaIn .....	66
Figure 2.1 - Imaging processing pipeline .....	79
Figure 2.2 - Schematic of how EBM works .....	80
Figure 2.3 - Conceptual overview of SuStaIn .....	85
Figure 2.4 - Flowchart of procedures for SuStaIn model fitting .....	89
Figure 2.5 - Global map (A) and table (B) reporting countries and recruitment sites that have contributed Pick's disease tissues to the Pick's disease International Consortium (PIC) to date .....	92
Figure 2.6 - PIC pathological diagnostic criteria for PiD.....	94
Figure 2.7 - Pathological assessments of Pick's disease brains .....	95
Figure 2.8 - Identifying genetic variants by risk allele frequency and genetic effect size. ....	101
Figure 3.1 - Kernel Density Estimation (KDE) mixture models.....	122
Figure 3.2 - Sequence of atrophy progression in PSP Richardson Syndrome .....	123
Figure 3.3 - Sequence of PSP-RS atrophy progression after five-fold cross validation .....	123
Figure 3.4 - Histogram of event-based model staging results for PSP-RS .....	125
Figure 3.5 - Longitudinal consistency of baseline EBM.....	126
Figure 3.6 - Association between predicted EBM stage, PSP Rating Scale score, and disease duration for all scans (baseline and follow-up) .....	127
Figure 3.7 - Association between predicted EBM stage, PSPRS and disease duration at baseline scan only.....	128
Figure 3.8 - Linear models to test association between age at scan and predicted EBM stage.....	129
Figure 4.1 - Association of age at baseline scan with covariate adjusted regional volumes in cases .	151
Figure 4.2 - Association of age at baseline scan with covariate adjusted regional volumes in controls .....	152

Figure 4.3 - Subtype progression patterns of PSP atrophy identified by Subtype and Stage Inference (SuStaIn).....	155
Figure 4.4 - Selecting optimal SuStaIn subtype model given data.....	156
Figure 4.5 - Stage distribution and Subtype probability of two-subtype model .....	157
Figure 4.6 - Average stage of PSP clinical syndromes assigned to the Subcortical SuStaIn subtype	160
Figure 4.7 - Stage progression at follow-up visits by SuStaIn subtype.....	165
Figure 5.1 - Selecting optimal SuStaIn subtype model given data.....	187
Figure 5.2 - Two-subtype model of atrophy progression in CBS identified by Subtype and Stage Inference (SuStaIn).....	188
Figure 5.3 - Three-subtype model of atrophy progression in CBS identified by Subtype and Stage Inference (SuStaIn).....	190
Figure 5.4 - Stage progression at follow-up visits by SuStaIn subtype.....	193
Figure 6.1 - Global map (A) and table (B) of reporting countries and recruitment sites that have contributed Pick's disease tissues to the Pick's disease International Consortium (PIC) to date .....	211
Figure 7.1 - Genetic principal component plots .....	231
Figure 7.2 - Overview of sample quality control for PiD GWAS .....	237
Figure 7.3 - Association plots for PiD .....	244
Figure 7.4 - Regional association plots and recombination rates at suggestive genomic loci .....	246
Figure 7.5 - Conditional analyses adjusted for the lead SNP at each suggestive GWAS loci. ....	247
Figure 7.6 - Fine-mapping of the two lead loci .....	250
Figure 7.7 - Regional plots from PiD GWAS and MetaBrain cis-eQTLs for <i>KCTD8</i> and <i>YIPF7</i> genes .....	252
Figure 7.8 - <i>KCTD8</i> RNA and protein expression levels.....	255
Figure 7.9 - <i>TRIM22</i> RNA and protein expression levels .....	256
Figure 7.10 - <i>YIPF7</i> RNA and protein expression levels.....	257
Figure 8.1 - Genetic principal component plots .....	267
Figure 8.2 - Overview of sample quality control for PiD GWSS .....	271
Figure 8.3 - Distribution of disease duration across PiD cohort .....	272
Figure 8.4 - Association plots for PiD survival GWAS .....	276
Figure 8.5 - Conditional analyses adjusted for the lead SNP at the top two GWSS loci .....	278
Figure 8.6 - Survival curves for genome-wide significant SNPs. Kaplan-Meier curves for PiD survival. ....	280
Figure 8.7 - Fine-mapping of the <i>NLGNI</i> locus.....	282
Figure 8.8 - Transcription factor binding site mapping at the <i>NLGNI</i> locus.....	283
Figure 8.9 - Cell-type specific regulatory element mapping at <i>NLGNI</i> locus .....	284
Figure 8.10 - Enhancer element mapping at <i>NLGNI</i> locus .....	285
Figure 8.11 - <i>NLGNI</i> RNA and protein expression levels .....	288



# List of Tables

Table 2.1 - 4R Tau imaging cohort summary.....	77
Table 2.2 - Overview of PIC contributing sites.....	97
Table 3.1 - Overview of PSP-RS cases by cohort included in study. ....	116
Table 3.2 - PSP-RS EBM baseline demographics.....	120
Table 4.1 - PSP clinical phenotypes and baseline characteristics by contributing cohort. ....	139
Table 4.2 - List of GIF subregions included in each region used as SuStaIn input. ....	142
Table 4.3 - Effect size by region of interest. ....	144
Table 4.4 - SuStaIn algorithm settings for each biomarker.....	146
Table 4.5 - Clinical and baseline characteristics by clinical diagnosis. ....	150
Table 4.6 - Comparison of demographics, clinical diagnosis and test scores between SuStaIn subtypes. ....	158
Table 4.7 - Average stage by clinical syndrome by SuStaIn Subtype.....	160
Table 4.8 - Regional brain volumes by subtype in the 2-subtype model .....	161
Table 4.9 - Comparison of adjusted clinical scores between subtypes. ....	163
Table 4.10 - Longitudinal consistency of subtype assignments .....	164
Table 5.1 - GIF subregions included in each cortical and cerebellar region used as SuStaIn input. ..	177
Table 5.2 - Effect size (Cohen's <i>d</i> ) by region of interest.....	180
Table 5.3 - SuStaIn algorithm settings for each biomarker .....	181
Table 5.4 - Baseline clinical and demographic data (by pathology) .....	185
Table 5.5 - Longitudinal consistency of subtype assignments for two-subtype model.....	192
Table 5.6 - Longitudinal consistency of subtype assignments for three subtype model .....	192
Table 5.7 - Comparison of demographics, pathological diagnosis and clinical test scores between subtypes (two-subtype model).....	195
Table 5.8 - Regional brain volumes by subtype in the two-subtype model .....	196
Table 5.9 - Comparison of demographics, pathological diagnosis and clinical test scores between subtypes (three-subtype model).....	197
Table 5.10 - Regional brain volumes by subtype in the three-subtype model .....	198
Table 5.11 - Comparison of adjusted clinical scores between subtypes in the two-subtype model....	200
Table 5.12 - Comparison of adjusted clinical scores between subtypes in the three-subtype model..	201
Table 6.1 - Genotype counts and frequencies of six common <i>MAPT</i> SNPs in Pick's disease cases and controls. ....	213
Table 6.2 - Clinical characteristics of samples included in <i>MAPT</i> haplotype analysis.....	215
Table 6.3 - Associations of individual <i>MAPT</i> variants with risk of Pick's disease.....	217
Table 6.4 - Associations of individual <i>MAPT</i> variants with age of disease onset and disease duration in Pick's disease subjects .....	217
Table 6.5 - Associations between <i>MAPT</i> haplotypes and risk of Pick's disease. ....	219
Table 6.6 - Associations of <i>MAPT</i> haplotype with age of disease onset and disease duration in Pick's disease cases. ....	220

Table 7.1 - Overview of total PiD samples included, broken down by contributing site.....	228
Table 7.2 - Clinical characteristics of samples included in GWAS. ....	238
Table 7.3 - Candidate variant analysis using GWAS data .....	240
Table 7.4 - <i>MAPT</i> Haplotypes from GWAS .....	241
Table 7.5 - Top independent SNPs at suggestive loci from PiD GWAS .....	245
Table 7.6 - Fine-mapping results of <i>KCTD8</i> and <i>TRIM22</i> loci .....	249
Table 8.1 - Clinical characteristics of samples included in GWSS. ....	272
Table 8.2 - Candidate variant analysis using GWSS data .....	274
Table 8.3 - Top independent genome wide significant SNPs from PiD GWSS .....	275
Table 8.4 - Lead Genomic Loci from PiD GWSS ( $p < 5 \times 10^{-6}$ ) .....	277
Table 8.5 - Colocalisation analysis of MetaBrain / eQTLGen datasets for genomic loci $p < 5 \times 10^{-7}$ ..	286

# Chapter 1: Introduction

## 1.1 Overview

Neurodegenerative diseases are a major and growing cause of mortality worldwide<sup>1</sup>. Specific diseases caused by neurodegeneration include Alzheimer's disease (AD), Parkinson's disease (PD), Dementia with Lewy Bodies (DLB) and Frontotemporal Lobar Degeneration (FTLD). Increasing life expectancy, in part due to more effective treatment of other age-related conditions such as cancer and ischaemic heart disease, combined with population growth, is driving an increase in the global burden of these disorders<sup>2</sup>. This threatens to further increase the pressure on already overstretched healthcare services<sup>3</sup>. Globally the number of people living with dementia, of which neurodegenerative disease is the commonest cause, doubled to 43.8 million between 1990 and 2016<sup>4</sup>. It is projected that by 2050 over 100 million people will be living with dementia<sup>5</sup>. The estimated societal economic cost in 2018 was over one trillion dollars and is forecast to double by 2030<sup>6</sup>. A significant proportion of this financial cost, not to mention the emotional suffering, is borne by the affected individual and their family<sup>7</sup>. Apart from the recent promising data for the anti-amyloid monoclonal antibody Lecanemab in early onset AD<sup>8</sup>, for all other sporadic neurodegenerative diseases there are currently no disease modifying therapies available.

In this thesis I focus on the primary tauopathies, a subgroup of FTLD, that share a common molecular mechanism of pathogenic accumulation of the altered microtubule-associated protein tau (MAPT); in particular Pick's disease (PiD), progressive supranuclear palsy (PSP) and corticobasal degeneration (CBD). These diseases can present anytime within adulthood, with rapidly progressive cognitive, behavioural and/or motor symptoms. They account for approximately 40% of FTLD cases<sup>9</sup>. FTLD is the third most common neurodegenerative cause of dementia after AD and DLB<sup>10–12</sup> and has a similar prevalence to AD in patients younger than 65<sup>9,13,14</sup>.

The overarching aims of my PhD are firstly to use data-driven generative models of disease progression to accurately stage and subtype patients with these diseases using structural MRI, and then to identify genetic risk factors for disease onset and progression.

## 1.2 An introduction to FTLD and the primary tauopathies

### 1.2.1 Terminology and clinical syndromes

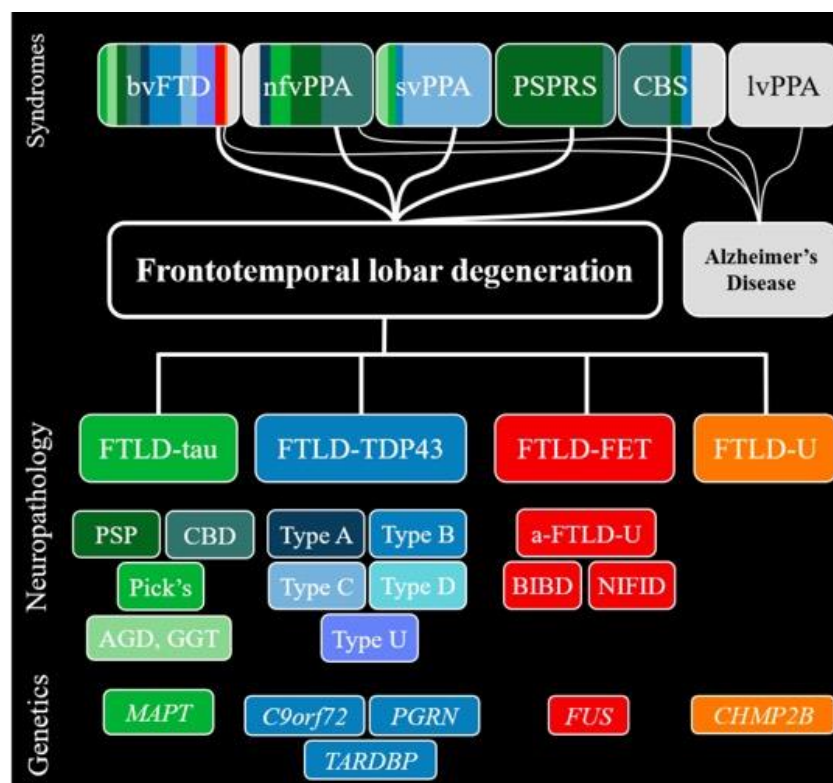
Research into FTLD is plagued by confusing terminology, which is exacerbated by the heterogeneous and often overlapping clinical syndromes that the underlying pathologies can present with. In this thesis I use the current consensus nosology to describe the spectrum of FTLD and associated clinical syndromes. The term FTLD refers to the neuropathological diagnosis characterised by the pathological accumulation of a disease specific protein<sup>15</sup>. I will use the term “*frontotemporal lobar degeneration syndromes*” to refer to the range of clinical syndromes that individuals can present with. These include behavioural variant frontotemporal dementia (bvFTD)<sup>16</sup>, the semantic (svPPA) and non-fluent (nfvPPA) variants of the primary progressive aphasia (PPA)<sup>17</sup>, progressive supranuclear palsy (PSP) syndrome, and corticobasal syndrome (CBS)<sup>18</sup>. The logopenic variant of PPA (lvPPA) is not commonly incorporated into the clinical spectrum of the FTLD syndromes due to its more common association with AD<sup>19</sup>.

### 1.2.2 Frontotemporal lobar degeneration

FTLD encompasses a highly heterogeneous group of disorders both in terms of clinical presentations and neuropathology (**Figure 1.1**). FTLD patients are neuropathologically characterised by relatively localized degeneration of the frontal and temporal lobes<sup>12,20</sup>, though the pathology can also involve subcortical structures such as the amygdala, hippocampus, basal ganglia, thalamus and brainstem<sup>21,22</sup>. This results in a diverse range of clinical syndromes depending on the particular pattern of neurodegeneration. As the pathology progresses through the brain, individual patients can go on to develop overlap syndromes depending on the particular pattern of topographical spread. While some patients can present with progressive changes in behaviour (bvFTD), others can develop language dysfunction presenting as one of the PPAs; most commonly nfvPPA or svPPA. These syndromes often converge as the disease progresses<sup>23</sup>. A proportion of patients with FTLD can present with, or develop,

atypical parkinsonian syndromes such as PSP and CBS, or motor neuron disease / amyotrophic lateral sclerosis (MND/ALS)<sup>19</sup>.

Microscopically, there is micro-vacuolation and neuronal loss with misfolded protein aggregation in neurons and/or glial cells<sup>12</sup>. The current molecular pathological classification splits FTLN into three major subtypes, based on the protein aggregates that are most characteristic: TDP-43 (FTLD-TDP43), tau (FTLD-tau), or FET (FTLD-FET)<sup>15,24,25</sup>. FTLD-FET is characterized by pathologic inclusions of the FET proteins, including FUS (Fused in sarcoma), Ewing sarcoma, and TATA-binding protein-associated factor 2N (TAF15). A fourth subtype, FTLD-UPS, was introduced for very rare forms of FTLD associated with inclusions only labelling for markers of the ubiquitin/proteasome system (UPS). So far these inclusions have only been seen in cases with mutations in *CHMP2B*.<sup>24</sup>.



**Figure 1.1 - The clinico-pathological spectrum of frontotemporal lobar degeneration (FTLD).** The first row shows the clinical syndromes associated with FTLD, with colours showing the proportion of cases associated with each FTLD subtype. The next row shows the neuropathological subtypes of FTLD, and the final row shows the Mendelian mutations so far identified to cause the respective pathologies. Abbreviations – FTD: Frontotemporal dementia, PPA: Primary progressive aphasia, bvFTD: behavioural variant FTD, nfvpPPA: non-fluent variant PPA, svPPA: semantic variant PPA, CBS: corticobasal syndrome, lvPPA: logopenic variant PPA, PSP: Progressive supranuclear palsy, CBD: Corticobasal degeneration, AGD, GGT: Globular glial tauopathy, FTLD-U: FTLD-UPS (ubiquitin proteasome system), a-FTLD-U: atypical FTLD-U, BIBD: Basophilic inclusion body disease, NIFID: Neuronal intermediate filament inclusion disease. *Reproduced with permission from Murley 2020<sup>26</sup>.*

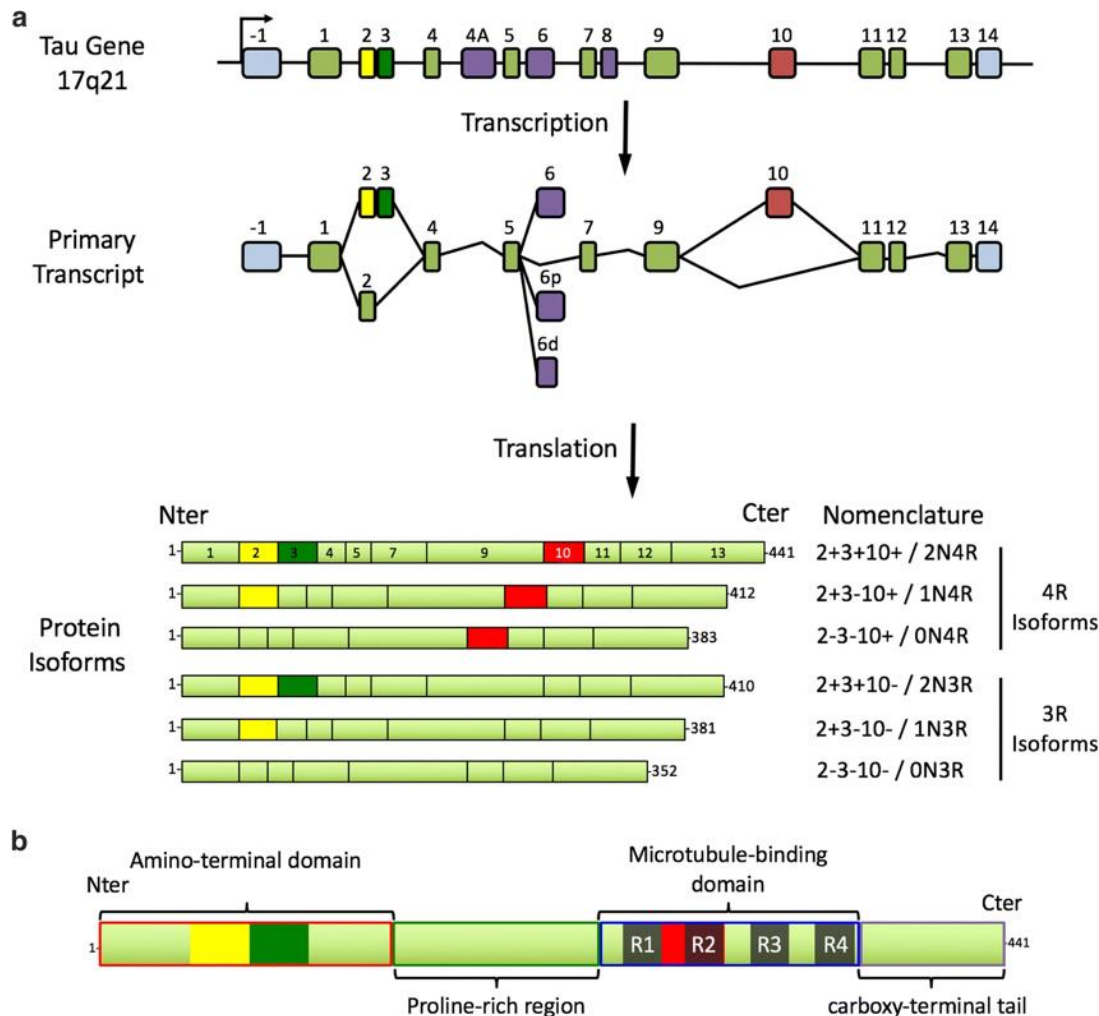
FTLD-tau pathology accounts for approximately 40% of all cases of FTL<sup>23</sup>, and the diseases within this subgroup are referred to as primary tauopathies. AD, in contrast, is classed as a secondary tauopathy as the tau pathology is thought to occur downstream mechanistically of amyloid deposition<sup>27</sup>.

### 1.2.3 Tau biology in health and disease

Tau is a microtubule associated protein that has an important role in microtubule assembly and stability, crucial for regulating axonal transport and maintaining neuronal stability<sup>28,29</sup>. Although predominantly expressed in neurons, it is also found in glial cells (astrocytes and oligodendrocytes), albeit at much lower concentrations<sup>30</sup>. Tau may also interact with the actin cytoskeleton, plasma membrane and cytoplasmic organelles<sup>31,32</sup>. The tau gene (*MAPT*) is located on chromosome 17q21, and six major isoforms, ranging from 352 to 441 amino acids in length, are expressed in the adult human brain (**Figure 1.2**). These isoforms are produced by alternative splicing of messenger RNA (mRNA) transcripts from *MAPT*, differing by the presence or absence of inserts of 29 and 58 amino acids in the N-terminal region (exons 2 and 3, with exon 3 only being transcribed together with exon 2), and in the C-terminal region the inclusion, or exclusion, of the 31 amino acid microtubule binding repeat, encoded by exon 10<sup>33</sup>. Alternative splicing of exons 2 or 3 produces isoforms with either none, one or 2 amino terminal inserts, whereas alternative splicing of exon 10 produces isoforms with either three (3R) or four (4R) microtubule binding regions. 3R tau isoforms predominate in the developing human brain, whereas in healthy adult brain, there is an equal ratio of 3R and 4R tau<sup>34</sup>. Adult 4R tau isoforms bind more strongly to microtubules than 3R, promoting microtubule assembly, which is in keeping with the need for a more dynamic cytoskeleton during nerve cell development<sup>35</sup>.

Phosphorylation has been shown to dynamically regulate the physiological functions of tau<sup>36,37</sup>, and in the primary tauopathies there is excessive and abnormal intracellular accumulation of insoluble fibrils of hyperphosphorylated tau in both neurons and glia<sup>38,39</sup>. Other post-translational modifications (including acetylation, glycosylation and truncation) may also contribute to pathology<sup>40</sup>. The amount and site of phosphorylation, and the ratio of 3R:4R isoforms varies between tauopathy subtype and is likely to explain, at least in part, the variation in inclusion morphology and cellular specificity<sup>23</sup>. This is supported by recent cryo-electron microscopy studies that

confirm disease-specific molecular conformations of tau fibrils across the tauopathy spectrum<sup>41</sup> (which I shall discuss in more detail in the next section). The primary tauopathies can be classified based on the distinct involvement of anatomical areas, the cell types in which abnormal tau accumulates<sup>42</sup>, and the ultrastructural features of the tau filaments<sup>41</sup>. These include PiD characterised by 3R tau immunoreactive neuronal inclusions; PSP, CBD, globular glial tauopathy (GGT), and argyrophilic grain disease (AGD) characterised by 4R tau in morphologically diverse neuronal and glial inclusions, and FTLT-tau due to *MAPT* mutations where the underlying pathology is heterogeneous and dependent on the specific mutation.



**Figure 1.2 - The *MAPT* gene, transcript isoforms and tau protein structure. (A)** The *MAPT* gene consists of 16 exons and six main transcripts resulting in six different protein isoforms in the human brain. Exons 1, 4, 5, 7, 9, 11, 12, and 13 are constitutive, whereas 2, 3, 6, and 10 are subject to alternative splicing. Exons 1 and 14 are in the untranslated regions of the gene. Alternative splicing of exons 2 or 3 produces isoforms with either none, one or two amino terminal inserts, whereas alternative splicing of exon 10 produces isoforms with either three (3R) or four (4R) microtubule binding regions. **(B)** Tau protein structure. Tau has four domains with different biochemical properties: an acidic amino terminal domain (encoded by exons 1-5), a proline rich domain (encoded by exons 7 and 9), the microtubule-binding region (MTBR) with 3-4 repeated sequences (3R tau [R2-R4] and 4R tau [R1-R4]), and a carboxy-terminal domain. Reproduced with permission from Colin et al. 2020<sup>40</sup>.

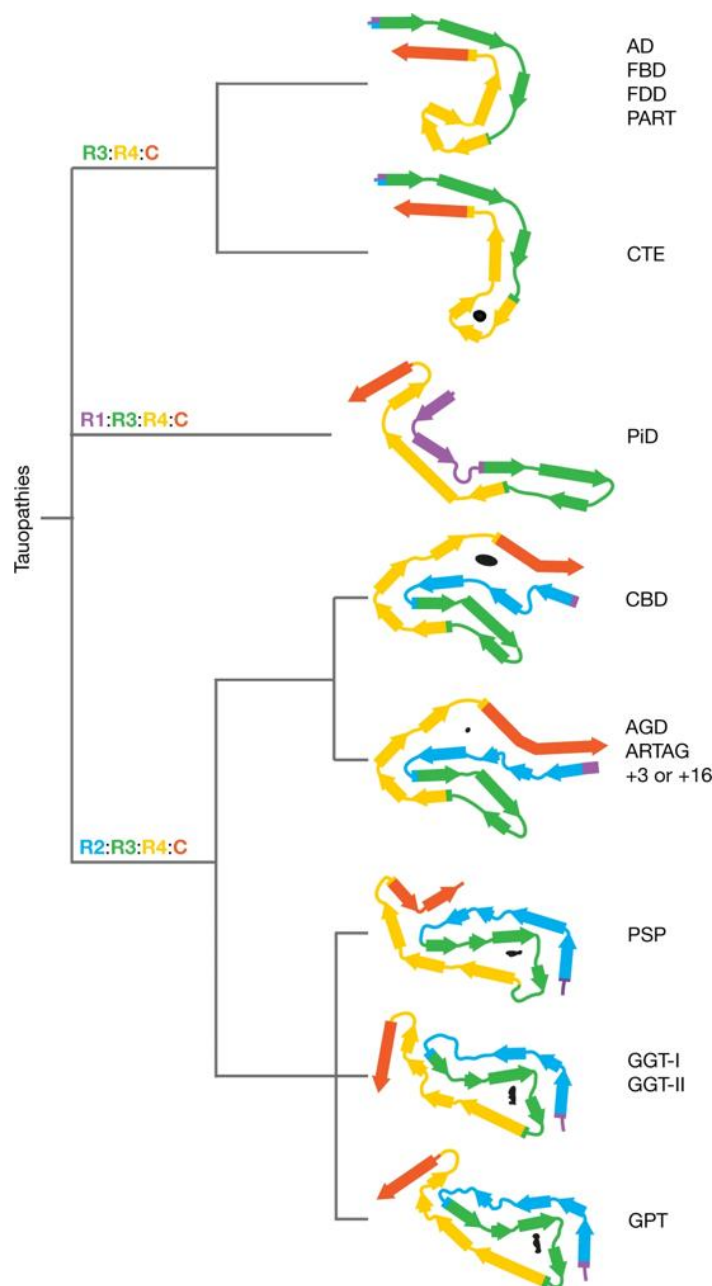
## 1.2.4 A structural classification of the tauopathies

Recently Shi et al. have proposed a structural based classification system for the tauopathies based on the unique filament fold derived from high-resolution cryo-electron microscopy (cryo-EM)<sup>41</sup>. In this study they build on previous work showing unique tau folds in AD (with identical folds in primary age related tauopathy [PART]), chronic traumatic encephalopathy (CTE), PiD and CBD<sup>43-47</sup>, and determine cryo-EM structures for the remaining primary tauopathies (PSP, GGT, AGD, 4R predominant



MAPT mutations, familial Danish dementia [FDD], ageing-related tau astroglipathy [ARTAG] and familial British dementia [FBD]). Overall, there are eight unique tau folds identifiable with cryo-EM across the 14 tauopathies suggesting an overarching hierarchical classification based on tau filament folds that complements the existing classification based on clinical features and post-mortem neuropathological examination (**Figure 1.3**).

In support of the validity of this classification system, there was a direct correlation between structural tau strains and specific tauopathies; individuals with the same disease had the same tau strain, and the tau strains usually differed between disease<sup>48</sup>. This association suggests that therapeutic approaches as well as diagnostics should take into account fibril structure. The question as to whether the fibrils are a consequence of disease, correlated or the actual cause remains to be determined. However, given the misfolding of tau into reproducible fibrils across the range of tauopathies investigated in a total of eight different ways, it seems the most likely explanation is that the variation in fibril type is, at least in part, a component of the main disease causing mechanism. Given the different fibrils have differing exposed surfaces with varying stabilities and flexibilities, it is possible that each fibril type causes cell-specific intracellular damage in a unique way<sup>48</sup>. An alternative explanation for the data is that under physiological conditions normal tau folds into a range of different fibrils, with particular fibrils strains selected for in specific tauopathies in certain cell types. This type of selection has been previously shown in prion diseases, where the infectious prion pool consists of several fibril polymorphisms<sup>49</sup>. Either way this study is a milestone in our understanding of the tauopathies providing a framework for both the development of more specific and sensitive tau biomarkers, and increasing our understanding of the underlying disease pathogenesis. It provides support for the hypothesis of the aggregation of tau (with perhaps different folds determining disease specific cell type vulnerability) in a small number of brain cells, followed by a prion-like spreading of pathological tau through the brain<sup>50</sup> in a specific spatiotemporal pattern<sup>51,52</sup>.

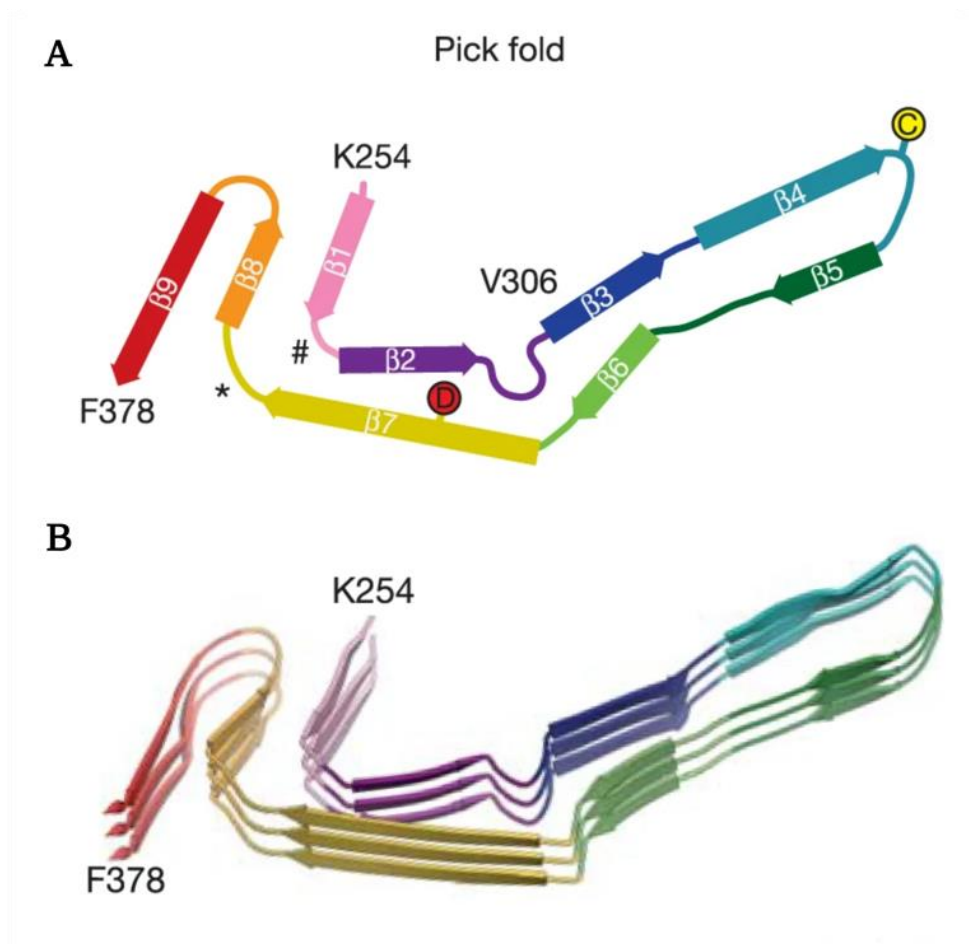


**Figure 1.3 - Structure-based classification of tauopathies.** The dendrogram shows the proposed classification of tauopathies, with the corresponding folds displayed with the first  $\beta$ -strand in R3 oriented horizontally, except for the GGT and GPT folds, which are aligned to the PSP fold. Residues in R1–R4 and in the C-terminal domain are coloured purple, blue, green, gold and orange, respectively. Internal, non-proteinaceous densities are shown in black. AD, Alzheimer’s disease; PiD, Pick’s disease. *Reproduced with permission from Shi et al. 2021<sup>41</sup>.*

Given the focus of this thesis is on PiD, PSP and CBD I will review these specific diseases in more detail below. For more detail on GGT, AGD, PART, ARTAG and CTE see the following excellent reviews of the primary tauopathies<sup>27,30,53</sup>. Ghetti et al.<sup>54</sup> also gives a comprehensive overview of the pathology related to autosomal dominant *MAPT* mutations.

### 1.2.5 3R tauopathies: Pick's disease

PiD is a 3R predominant tauopathy characterised by severe atrophy of the frontal, temporal and to a lesser extent the parietal lobes<sup>55</sup>. It is morphologically and biochemically distinct from all of the other FTL D-tau subtypes, with the presence of neuronal spherical Pick's bodies containing 3R tau, ramified astrocytes and less frequently oligodendroglial globular inclusions<sup>56</sup> throughout the frontotemporal neocortex, white matter and to a lesser extent the basal ganglia and brainstem nuclei<sup>53</sup>. Granular neurons of the dentate fascia and hippocampal neurons are particularly susceptible to Pick's bodies<sup>12</sup>. The tau filaments are characterised by 22-24 nm diameter twisted filaments<sup>57</sup> with 15-19 nm diameter straight tubules, and have a novel fold shown on cryo-electron microscopy unique to PiD<sup>47</sup> (**Figure 1.4**).



**Figure 1.4 - Structure of tau filaments in PiD.** Models based on resolved cryo-EM tau protofilaments structure in PiD. (A) Schematic of the secondary structure elements in the Pick folds depicted as a single rung. The positions of Cys322 (yellow 'C') and Asp348 (red 'D') are highlighted (B) Rendered view of the secondary structure elements in the Pick fold, depicted as three successive rungs. *Figure modified from Falcon et al.<sup>47</sup> with permission.*

Historically PiD referred to a clinical syndrome of progressive behavioural and language decline first described by Arnold Pick in 1892<sup>55</sup>, and became synonymous with the spectrum of FTLN syndromes. It was a further nineteen years before Alois Alzheimer reported the classic pathology of Pick's bodies and ballooned cells<sup>58</sup>. It has only been in more recent years, as the lack of clinico-pathological correlation became clearer, that the label of Pick's Disease has been updated to refer only to those patients at post-mortem that have the classic 3R pathological findings detailed above<sup>25</sup>. Although PiD is rare, more recent studies suggest it accounts for up to 30% of FTLN-tau cases at autopsy<sup>59</sup>, and 10% of all cases of FTLN overall. Given the difficulty of diagnosing PiD in life, the prevalence is unknown, though it has been estimated that it is at least ten times less frequent than AD<sup>60</sup>.

PiD usually develops in those over the age of 55 years, with average survival approximately ten years from symptom onset<sup>61–64</sup>. These studies suggest there is a significant variation in survival with some individuals surviving for more than 20 years<sup>61</sup>. PiD can present across the spectrum of FTLN clinical syndromes, though bvFTD is the most common (up to 84% of cases), followed by svPPA, nvPPA and CBS<sup>65</sup>. In a recent study, 24% of individuals with PiD (5/21) had a clinical diagnosis of CBS, suggesting that CBS may not be as rare presentation of PiD as previously thought<sup>66</sup>. As the disease progresses the clinical syndromes may converge<sup>65,67</sup>. Imaging studies show PiD is an asymmetrical disease that has a predilection for either the right or left hemisphere initially with this asymmetry increasing as the disease progresses<sup>65</sup>. In particular it appears to affect a network of cortical regions including the frontal insula, anterior temporal lobe, anterior cingulate and with lesser involvement of the basal ganglia, posterior temporal and anterior parietal areas<sup>62,67</sup>. The presenting clinical syndrome is likely to relate to the specific region of this network affected first: bvFTD occurs most commonly when there is anterior right hemisphere atrophy, svPPA with left anterior temporal atrophy, nvPPA with left frontal-insula atrophy, and CBS if posterior frontal and parietal regions are affected first<sup>65</sup>. The overlap of clinical syndromes as the disease progresses is likely to due to sequential atrophy of similar hubs in the network.

## **1.2.6 4R tauopathies**

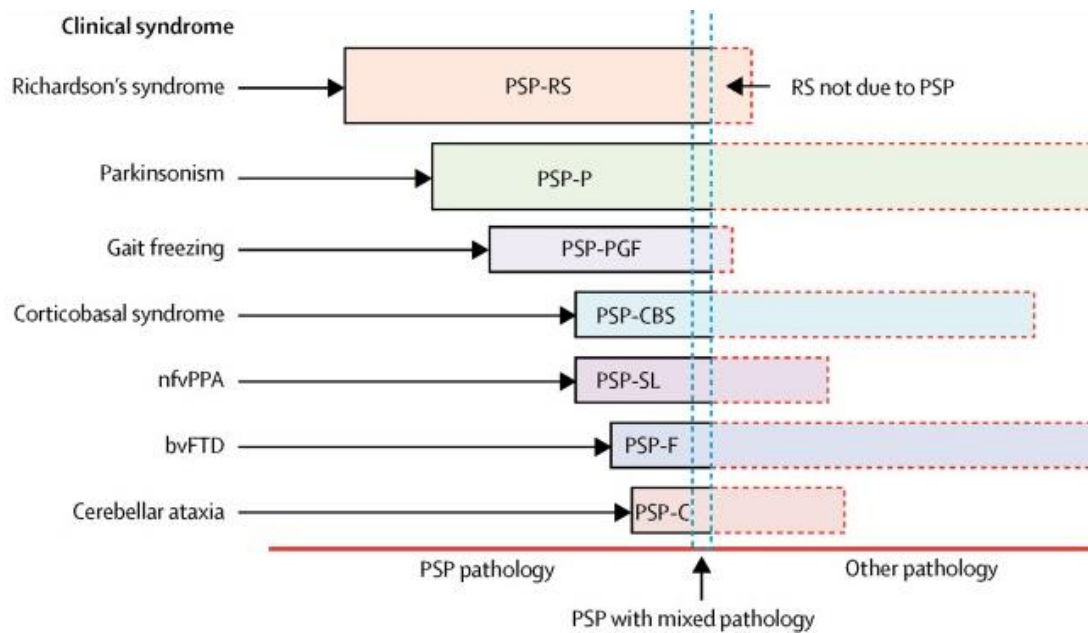
### **1.2.6.1 Progressive supranuclear palsy**

PSP is a 4R tauopathy characterised by neurodegeneration and tau accumulation predominantly in the subthalamic nucleus, globus pallidus, striatum, cerebellum with dentate nucleus, frontal lobes and to a lesser extent in the occipital cortices<sup>51</sup>. The neuropathological diagnosis is based on the presence of a combination of neurofibrillary tangles and tufted astrocytes in subcortical nuclei<sup>68,69</sup>, in addition to oligodendroglial coiled bodies and diffuse cytoplasmic immunoreactivity in neurons<sup>30,70</sup>.

The most common so called “classical” clinical phenotype of PSP, Richardson syndrome (PSP-RS), was first described by Steele, Richardson and Olszewski in 1963<sup>71</sup>. It is characterised by a levodopa unresponsive akinetic-rigid syndrome with a vertical supranuclear gaze palsy, early falls and dementia. Diagnosis of PSP-RS was operationalised in the NINDS–Society for Progressive Supranuclear Palsy criteria of 1996<sup>72</sup>. PSP-RS is associated with PSP pathology in approximately 90% of cases<sup>73</sup>. Epidemiological studies of PSP have been largely based on this classical clinical phenotype, suggesting an annual incidence increasing with age from 1.7 cases per 100,000 at ages 50 to 59 years to 14.7 per 100,000 at 80 to 89 years<sup>74</sup>. Overall prevalence has been estimated between 4-7 per 100,000<sup>14,75,76</sup>. This makes PSP-RS the most common of the FTLT-tau diseases.

In recent years it has become clear from pathological studies that there is significant pathological heterogeneity resulting in a spectrum of clinical presentations; indeed PSP-RS may only account for approximately half of autopsy-confirmed PSP cases<sup>77–79</sup> with the remainder presenting with variant PSP (vPSP) clinical syndromes. These include parkinsonism resembling idiopathic Parkinson’s disease (PSP-P)<sup>78–80</sup>, progressive gait freezing (PSP-PGF)<sup>81–83</sup>, postural instability (PSP-PI)<sup>77,84</sup>, predominant ocular motor dysfunction (PSP-OM)<sup>72,77</sup>, a fronto-behavioural presentation (PSP-F) including bvFTD<sup>85–87</sup>, and a speech / language disorder (PSP-SL) including progressive apraxia of speech (AOS) and nvPPA<sup>88–92</sup>, corticobasal syndrome (PSP-CBS)<sup>93,94</sup> and primary lateral sclerosis (PSP-PLS)<sup>95,96</sup>. It is important to note that although PSP-RS has a high clinico-pathological correlation with underlying PSP pathology<sup>73</sup>, a wider range of pathologies can underlie the vPSP

syndromes (**Figure 1.5**). However, where the PSP syndrome is caused by underlying PSP pathology, the PSP tau filament fold is identical between typical and variant PSP syndrome<sup>41</sup> (**Figure 1.3**), suggesting that the initiating sites of tau pathology are similar in the clinical subtypes with differing subsequent propagation patterns. The PSP tau fold structure is most similar to the GGT fold, with it adopting a markedly different confirmation to the fold seen in CBD.



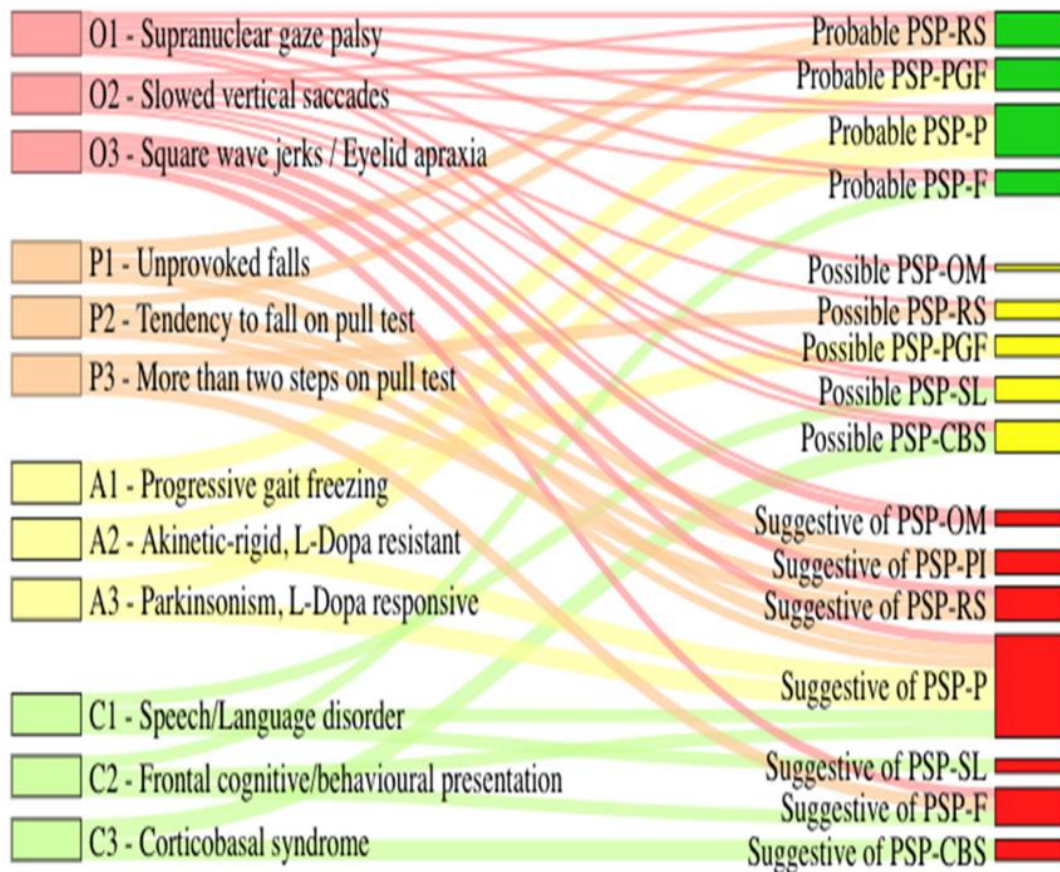
**Figure 1.5 - Clinical syndromes in progressive supranuclear palsy.** Relative proportions of each syndrome (represented by bar length) are speculative. PSP with mixed pathology means meeting the neuropathological criteria for PSP in addition to other neurodegenerative disorders. Other pathology means at post-mortem pathology is demonstrated to not be PSP but rather another pathology such as AD or Parkinson's Disease. *Figure reproduced with permission from Boxer et al. 2017*<sup>97</sup>

This clinical heterogeneity has been operationalised in the 2017 Movement Disorder Society (MDS) PSP clinical diagnostic criteria<sup>98</sup>, to try and improve the clinical detection (sensitivity) of underlying PSP pathology, while achieving high specificity versus alternative diagnoses. These new criteria expand the core functional domains of postural instability and ocular motor dysfunction by adding in two further domains (cognitive dysfunction and akinesia), and also specify in each domain three characteristic core clinical features stratified by presumed levels of certainty (1 [highest], 2 [mid], and 3 [lowest]). Specific combinations of the core clinical features indicate the degree of diagnostic certainty and map to specific PSP syndrome diagnoses (16 in total) (**Figure 1.6**). The “probable” diagnoses have a high specificity for underlying PSP pathology but low sensitivity early in the disease course, while the

“possible” category have a high sensitivity but lower specificity. The introduction of the new category “suggestive of” recognises that early in the disease course key diagnostic features may not be present and so the individual may not meet formal diagnostic criteria. However, this category allows the early identification of individuals with PSP pathology<sup>99</sup> (with the acknowledgement there may be a different underlying pathology driving the presentation in some cases), that may be suitable for biomarker development or inclusion in clinical trials of disease-modifying therapies<sup>100</sup>. It also allows for a change in the diagnostic label between vPSP and typical (PSP-RS) PSP over time. To overcome the issue of individuals being assigned to multiple diagnoses, Grimm et al.<sup>99</sup> proposed the four MAX (multiple allocation extinction)-rules which decrease the average number of diagnoses per individual from 5.4 to 1.1. The first two rules state that diagnostic certainty (probable > possible > suggestive of) and temporal order (1<sup>st</sup> > 2<sup>nd</sup> > 3<sup>rd</sup> diagnosis) should be applied based on the rationale that this reflects increasing specificity of there being underlying PSP pathology. Rule three states that those phenotypes with a higher specificity for PSP pathology should take precedence (PSP-RS > PSP-OM/PSP-PI > others), while rule four states that where more than one of the above rules apply the rule hierarchy is Rule 1 > Rule 2 > Rule 3. Studies support the 2017 criteria having a higher sensitivity for PSP pathology than the previous NINDS criteria (87.9% vs 45.5%)<sup>101,102</sup>, and the “*suggestive of PSP*” clinical category significantly increases the sensitivity for early identification of patients with PSP pathology<sup>99</sup>.

Natural history studies have shown for typical PSP (i.e. PSP-RS) that the mean age of disease onset is 63 to 67 years with a mean disease duration of 6 to 7 years<sup>14,103,104</sup>. The subcortical PSP variants, PSP-P and PSP-PGF, have a similar age of disease onset to PSP-RS, but have a significantly longer disease duration (PSP-RS 6-7 years, PSP-P 9 years, PSP-PGF 13 years), and slower rates of disease progression<sup>81,103–109</sup>. This is supported by a recent pathology study showing that disease duration is significantly longer in PSP subcortical phenotypes compared to PSP-RS and PSP cortical variants<sup>107</sup>. Even when controlling for disease severity at baseline, PSP-RS progresses significantly faster than PSP-P<sup>105</sup>. The molecular pathogenic basis for this clinical variation is still poorly understood, though a recent study by Kovacs et al.<sup>51</sup> shows that significant differences in tau burden and different tau cytopathologies may distinguish clinical subtypes.



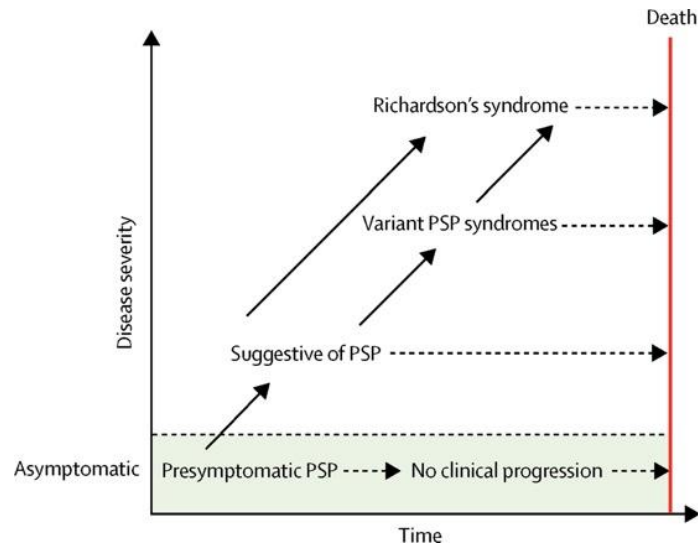


**Figure 1.6 - Contribution of clinical features to different PSP phenotypes.** Sankey plot with core clinic features displayed on the left of the plot and the phenotypic diagnoses according to the MDS 2017 PSP diagnostic criteria (Hoglinger 2017) on the right. The phenotypic diagnoses are coloured according to diagnostic certainty: high ('Probable' - green), medium ('Possible' - yellow) and low ('Suggestive of' - red). O = Ocular motor dysfunction, P = Postural instability, A = Akinesia, C = Cognitive dysfunction. *Reproduced with permission from Street 2022<sup>110</sup>.*

Given the rare sporadic nature of PSP, and the absence of in-vivo biomarkers, little is known about the pre-diagnostic stages of the disease. There is strong evidence in more common sporadic neurodegenerative disorders such as AD and PD that there is a pre-symptomatic phase in which neuropathological changes accumulate before the threshold for manifestation of clinical symptoms is reached<sup>97</sup>. In genetic forms of AD and FTD biomarker changes occur between 10-25 years prior to clinical symptom onset<sup>111–114</sup>. Reports of mild asymptomatic PSP pathology in clinically healthy older people<sup>115–117</sup>, suggest that a similar sequence of events occurs in PSP. A recent study of UK Biobank data suggests that clinical diagnostic features of PSP (motor slowing, cognitive dysfunction and postural instability) are present at least three years before diagnosis suggesting that in addition to a presymptomatic phase there is also a long prediagnostic phase in PSP with subtle changes in motor and cognitive function<sup>118</sup>. A hypothetical model for PSP incorporating the MDS 2017 diagnostic criteria has been



proposed by Boxer et al.<sup>100</sup> (**Figure 1.7**). However, given the lack of symptoms in the pre-symptomatic period and the lack of clear mendelian genetics, combined with the low specificity of early symptoms in the pre-diagnostic phase of PSP has made research in these early stages very difficult.

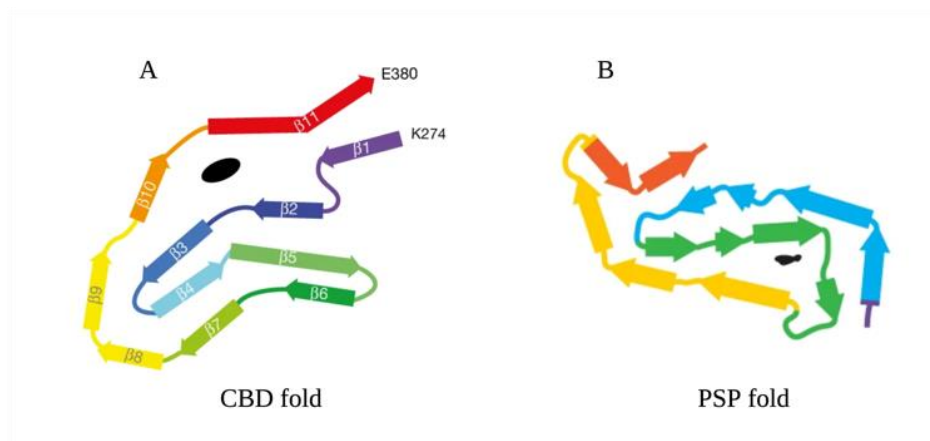


**Figure 1.7 - Hypothetical model of disease courses in PSP syndromes.** Hypothetical model of the clinical trajectories of progressive supranuclear palsy. Progressive supranuclear palsy is considered as a pathological continuum from a presymptomatic phase, through a suggestive phase, to a fully symptomatic phase that, in many cases, would meet research criteria for possible or probable PSP-Richardson's syndrome by the Movement Disorder Society criteria,<sup>10</sup> or a variant PSP syndrome. Not all cases of presymptomatic or suggestive PSP will progress to a PSP phenotype. PSP=progressive supranuclear palsy. *Reproduced with permission from Boxer et al.*<sup>100</sup>

### 1.2.6.2 Corticobasal degeneration

CBD is characterised by the accumulation of hyper-phosphorylated 4R tau in both neurons and glia of the cortical and subcortical regions, and to a lesser extent the brainstem<sup>119,120</sup>. The hallmark glial lesion that distinguishes it from the other primary tauopathies is the large and diffuse astrocytic plaque found in the limbic and subcortical structures<sup>68,121</sup>, though one also finds tau-positive “coiled bodies” and tau inclusions in the white matter. Macroscopically, there is atrophy in the medial surface of the superior frontal gyrus, subcortical white matter atrophy, thinning of the corpus callosum, discoloration of the globus pallidus and invariably depigmentation of the substantia nigra with gliosis and neuronal loss<sup>66</sup>. The subthalamic nucleus may have atrophy, but is usually less affected than in PSP. Although classically CBD is thought to be characterised by asymmetrical atrophy, there is a growing recognition that lack of marked asymmetry does not rule out CBD pathology<sup>66</sup>.

The histopathological features of CBD can overlap with PSP<sup>23,120</sup>, and there has been ongoing debate about the utility of considering these disease as separate clinic-pathological entities<sup>122,123</sup>. However, recent evidence from cryo-EM studies demonstrated a unique tau filament ultrastructure in CBD that is distinct from PSP as well as the other tauopathies<sup>41,46</sup>. Although the ordered core of the CBS tau fold is essentially the same as that in the PSP tau filaments, it adopts a markedly different confirmation. This supports, at least from a pathological perspective, that CBD is a distinct disease.



**Figure 1.8 - Structure of tau filaments in CBD and PSP.** Schematic model based on resolved cryo-EM tau protofilament structure in (A) CBD and (B) PSP. Internal, nonproteinaceous densities are shown in black. Adapted from Zhang *et al.* 2020<sup>46</sup> and Shi *et al.*<sup>41</sup>

CBD was originally described by Rebeiz et al.<sup>124</sup> in the 1960s to describe CBD pathology presenting with a classical CBS; what was thought at the time to be a distinct clinic-pathological entity. Findings from large autopsy studies demonstrated that less than 50% of CBS patients had underlying CBD pathology with alternative underlying pathologies including most commonly AD and PSP<sup>53</sup>. Other pathologies found at post-mortem include FTLN-TDP-43<sup>125</sup>, Lewy body disease (LBD)<sup>126</sup>, primary lateral sclerosis<sup>127</sup>, FTLN-FUS<sup>128</sup>, Creutzfeldt Jacob disease (CJD)<sup>129</sup>, cerebrovascular disease<sup>130</sup> and atypical MSA<sup>131</sup>.

Conversely CBD pathology, in addition to presenting with CBS (37% of cases)<sup>132,133</sup>, can present with PSP-RS (23%)<sup>133,134</sup>, bvFTD (14%)<sup>134–137</sup>, nfvPPA (5%)<sup>138,139</sup>, or an AD amnesic syndrome (8%)<sup>140</sup>. The heterogeneity of clinical syndromes in CBD is related to the diversity of focal cortical atrophy, with the specific clinical syndrome related to the location of the highest burden of tau pathology<sup>141</sup>. Pathologically, CBD can be divided based on the distribution and severity of pathology into three main subtypes which correspond to the main clinical phenotypes: typical CBD, “basal ganglia predominant” CBD and “PSP-like” CBD<sup>66</sup>. Both “basal ganglia predominant” and “PSP-like CBD” are associated with a PSP like syndrome, while typical CBD is associated with a frontal behavioural-spatial (bvFTD like) syndrome. Tau burden in CBD-CBS is greater in the frontal (in particular motor) and parietal cortices, whereas in CBD-PSP it is greater in the cerebellum and medulla. The contribution of co-pathologies is an area of active research, with the suggestion that this may modify clinical phenotypes: TDP-43 pathology, affecting in particular the midbrain tegmentum, is more common in CBD-RS than CBD-CBS<sup>142</sup>.

In the absence of a molecular biomarkers for CBD or co- pathology this has made the definition of sensitive and specific clinical diagnostic criteria for CBD extremely challenging. The most recent diagnostic criteria for CBD are the Armstrong criteria<sup>140</sup>, which define four clinical syndromes of CBD pathology: CBS, PSP syndrome (PSPS), frontal behavioural-spatial syndrome and nfvPPA. CBS is further classified as probable or possible. Based on the combination of clinical phenotypes and associated features individuals can be defined as either probable or possible CBD. The probable CBD label aims to identify individuals with a high probability of CBD pathology, while those with possible CBD are more likely to have mixed pathology. Although the Armstrong criteria are an advance on previous criteria due to including clinical

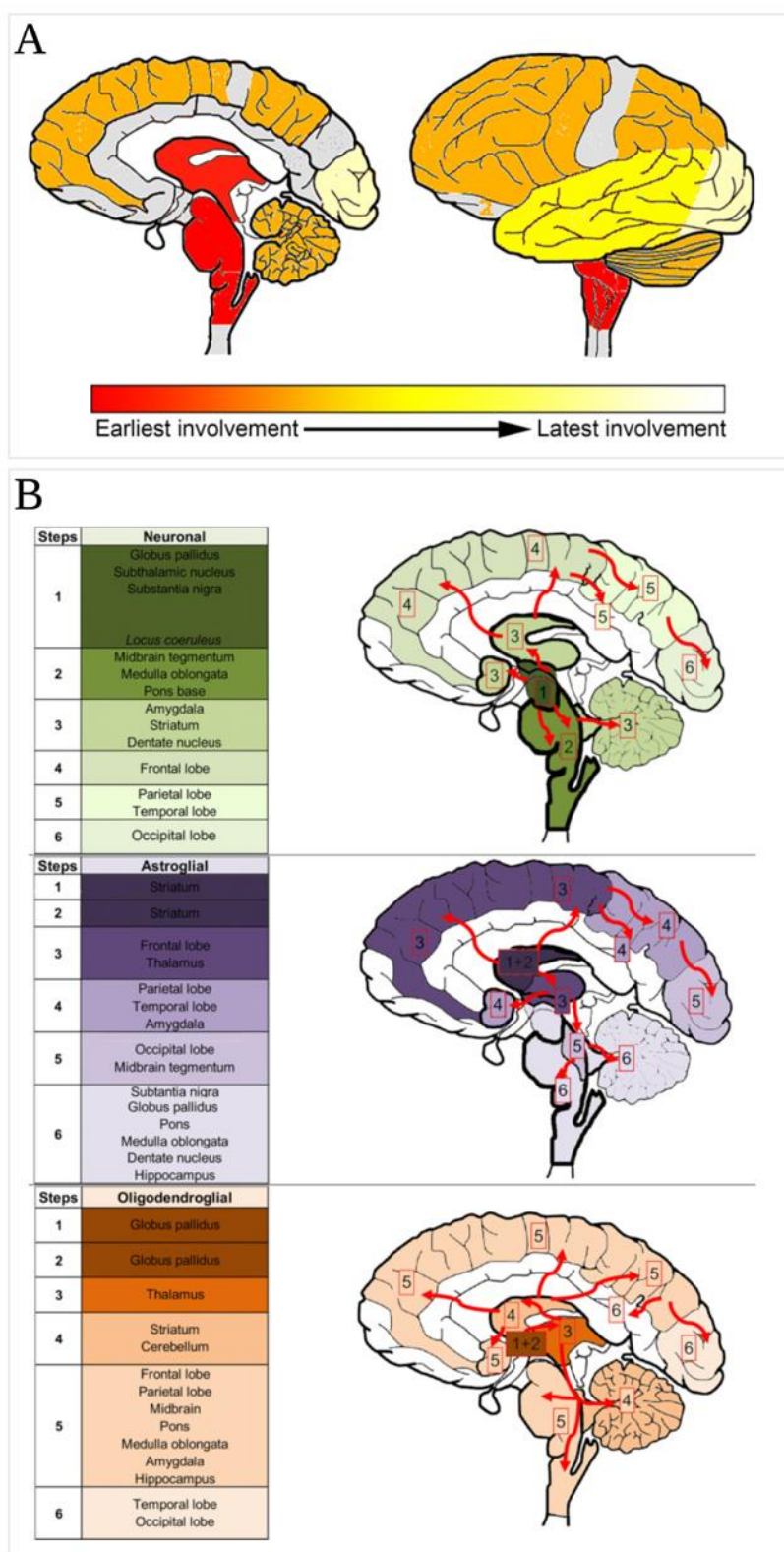
phenotypes other than CBS, they have been shown to have low sensitivity and specificity<sup>143</sup>. The 2017 MDS diagnostic criteria for PSP include CBS as one of the eight possible clinical phenotypes (possible PSP with CBS [PSP-CBS])<sup>98</sup>. This phenotype is included within the probable 4R tauopathy classification, to allow for joint ante-mortem recognition of clinical syndromes that predict with high specificity the presence of underlying PSP or CBD pathology<sup>99</sup>. These criteria include AD biomarkers to exclude AD pathology, the most common non-CBD pathology underlying CBS. With development of tau isoform specific biomarkers alongside AD biomarkers, it will be necessary to incorporate these into updated diagnostic criteria to improve sensitivity and specificity.

### **1.2.7 Pathological staging systems**

There is growing evidence that disease progression in tauopathies is characterised by the accumulation of pathological tau in vulnerable neuronal populations (primary vulnerable cells) and then spread (propagation) to an increasing number of brain regions (secondary vulnerable cells) in a hierarchical and stereotypical pattern<sup>144</sup>. A number of studies suggest that propagation of pathology occurs in a prion-like manner, where conformational ‘seeds’ of the pathological protein are transmitted from cell-to-cell<sup>34,145–147</sup>. Although the so-called “pathogenic spread” model can account for the stereotypical spatial distribution of protein aggregates seen in neuro-pathological staging, equally the older concept of selective neuronal vulnerability could also explain this observation. The proponents of this hypothesis argue that certain neurons are more vulnerable to the accumulation of a pathological protein than others<sup>148</sup>, perhaps due to differences in gene expression profiles, leading to them degenerating earlier than others. Of course, these competing hypotheses are not necessarily mutually exclusive, as it is also possible that “selective vulnerability” to particular tau species/conformers could explain the variation in initiation site of pathology, with subsequent spread then determined by “pathogenic spread” down either anatomical or functional connections, or indeed both. It may also be the case that the relative contribution of these two mechanisms varies between the specific disease.

The recognition of sequential involvement of brain regions in neuropathology studies, has led to the development of staging systems based on the spatial distribution of protein deposits. A recent study in PSP by Kovacs et al.<sup>51</sup> consolidates and builds upon

previous work showing that the dominant tau pathology in incipient cases is astrocytic<sup>115</sup>, and the severity and distribution of 4R tau pathology differs between clinical phenotypes<sup>78,79,94,149</sup>. In PSP-RS they demonstrated that there are six sequential stages of brain region involvement starting with the subthalamic nucleus, with tau pathology spreading out caudally to the cortex and rostrally to the cerebellum. There appeared to be a common early vulnerability pattern across all PSP subtypes centred on the pallido-nigro-luysian axis, that then diverges both spatially and in relation to the predominant cytopathology (neuronal vs glial) between different subtypes (**Figure 1.9**)<sup>51</sup>. Although neuronal tau pathology is important, in the earliest stage they found that both astrocytic and oligodendroglial tau pathology can proceed neuronal pathology in some regions. This finding is difficult to explain directly by anatomical connections within the “pathogenic spread” model, and evidence at the transcriptomic level showing that neurofibrillary tangles are enriched for synaptic and PSP risk genes, whilst tufted astrocytes are enriched for the microglial immune network, suggests that cell-autonomous vulnerability certainly plays a role<sup>150</sup>. The early involvement of astrocytes is also seen in CBD<sup>119</sup>, while there have also been reports of cases of PiD where there is astrocytic tau in the absence of neuronal tau<sup>62,151</sup>.

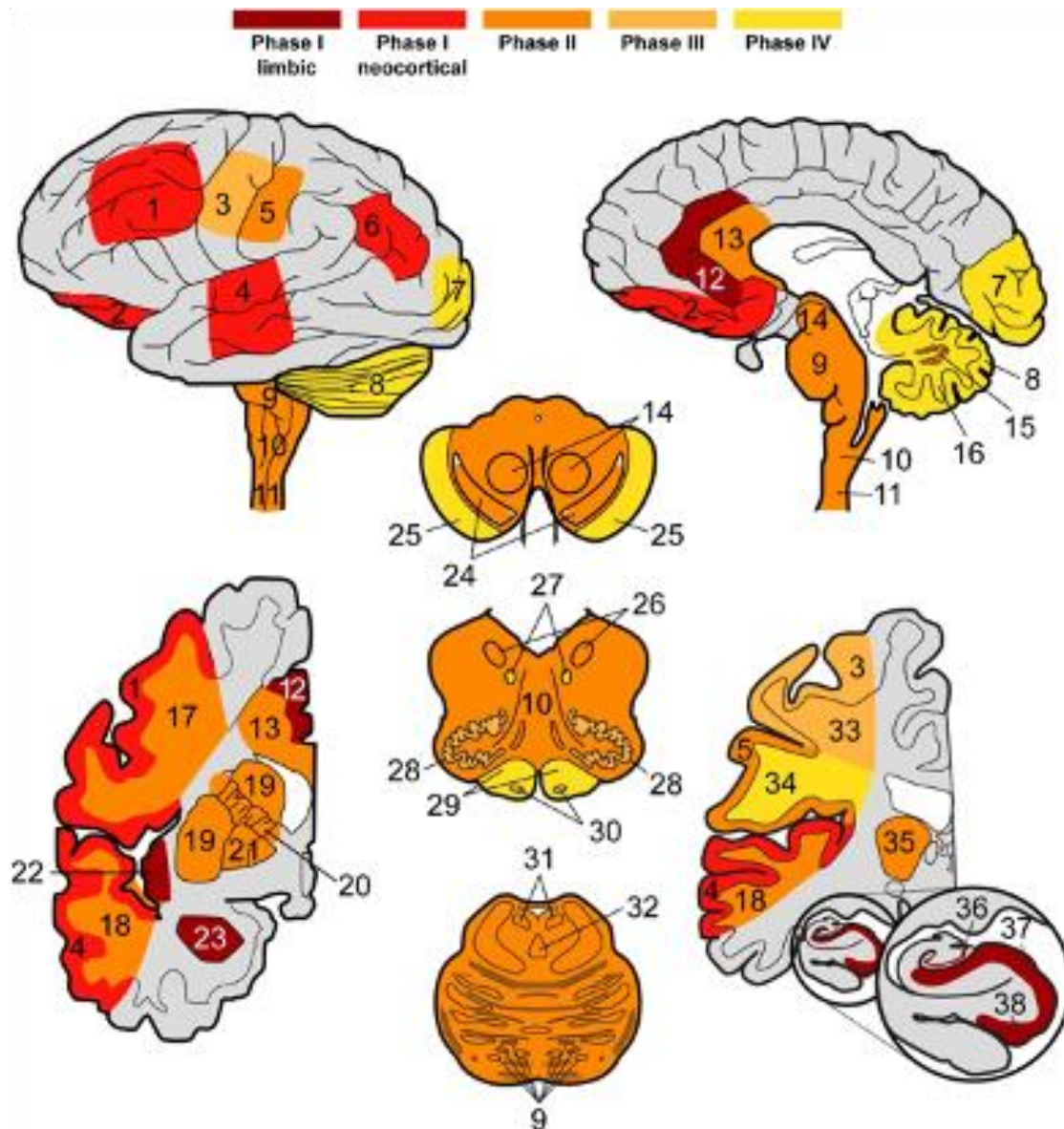


**Figure 1.9 - Sequential stages of tau pathology in PSP.** (A) Heatmap showing development of tau pathology in pooled cases of different PSP clinical subtypes. Based on conditional probability matrix of total tau pathology across pooled cases. Dark red indicates early and yellow-white later involvement. Grey indicates region not included in analysis. (B) Data from conditional probability matrix stratified for accumulation of neuronal, astroglial and oligodendroglial pathologies. Numbers in brackets indicate the order (1 = earliest vs. 6 = latest) specified region becomes abnormal in sequence. *Adapted from Kovacs et al.*<sup>51</sup>.

Similar to PSP, sequential distribution patterns of tau deposition have been recognised in PiD, with four stages of pathology proposed<sup>62</sup> (**Figure 1.10**). Phase 1 shows tau deposition in frontotemporal limbic cortices (limbic: anterior cingulate, insula, amygdala) and neocortical regions (cortical: middle frontal, orbitofrontal cortex, angular gyrus, temporal cortex), Phase 2 spreading into basal ganglia, Phase 3 motor cortex and Phase 4 involves the visual cortex and cerebellum. This pattern suggests the pathology spreads from the frontotemporal lobes to the basal ganglia, then to the motor cortex and the cerebellum potentially via the thalamus, prior to more widespread dispersion via thalamic relays<sup>152</sup>. Of note the majority of the cases in this study had bvFTD, though a more recent imaging study showed that early volume loss in the insula, cingulate, and orbitofrontal cortex was common to both PiD presenting with bvFTD and nvPPA<sup>67</sup>. It will be important to in future work to build on this proposed staging system by including PiD cases presenting with other clinical syndromes, including nvPPA, svPPA and CBS.

In CBD, pathological staging with hierarchical regional involvement has yet to be carried out in a large case series, possibly due to the heterogeneity of clinical phenotypes and the consequent difficulty in identifying a unique sequence of regional involvement<sup>152</sup>. There is evidence from both incidental asymptomatic CBD cases<sup>153,154</sup> and end stage CBS cases, of astrocytic plaque pathology as the most prominent pathology in anterior frontal and parietal regions. Interestingly, there is widespread tau pathology before the onset of clinical symptoms, in a similar pattern to end stage CBD, though less severe. This suggests a threshold of pathological burden in the appropriate anatomical regions needs to be reached before the onset of clinical symptoms<sup>119</sup>. The authors have proposed four stages based on tau lesion load and cell type involved, with the anterior frontal cortex, striatum and subthalamic nucleus affected early.





**Figure 1.10 - Sequential stages of tau pathology in PiD.** Phases are defined by the extent of regional involvement of tau pathology across cases beginning with **Phase I**, where disease is restricted to limbic and neocortical frontotemporal regions and angular gyrus. **Phase II** has additional pathology in associated white matter tracts, subcortical structures, and serotonergic/noradrenergic brainstem nuclei. **Phase III** tau neuropathology is characterized by additional pathology in primary motor cortex and precerebellar nuclei in the medulla. Finally, **phase IV** cases with the most severe tau pathology burden include additional tau deposits in the visual cortex and variably in the cerebellar granular layer and brainstem white matter. Numbering index: 1 = midfrontal cortex; 2 = orbitofrontal cortex; 3 = motor cortex; 4 = superior mid-temporal cortex; 5 = sensory cortex; 6 = angular gyrus; 7 = visual cortex; 8 = cerebellar granule layer; 9 = pons; 10 = medullary reticular formation; 11 = cervical spinal cord; 12 = anterior cingulate gyrus; 13 = corpus callosum; 14 = midbrain red nucleus; 15 = cerebellar dentate gyrus; 16 = cerebellar white matter; 17 = midfrontal cortical white matter; 18 = superior midtemporal cortical white matter; 19 = striatum; 20 = internal capsule; 21 = globus pallidus; 22 = insular cortex; 23 = amygdala; 24 = midbrain substantia nigra; 25 = midbrain crus cerebri; 26 = dorsal motor nucleus vagus; 27 = hypoglossal nucleus; 28 = inferior olive; 29 = medullary pyramids; 30 = arcuate nucleus/pontobulbar body; 31 = locus coeruleus; 32 = raphe nuclei; 33 = motor cortical white matter; 34 = sensory cortical white matter; 35 = thalamus; 36 = hippocampal dentate gyrus; 37 = hippocampal cornu ammonis; 38 = hippocampal entorhinal cortex. *Figure reproduced with permission from Irwin et al<sup>62</sup>.*



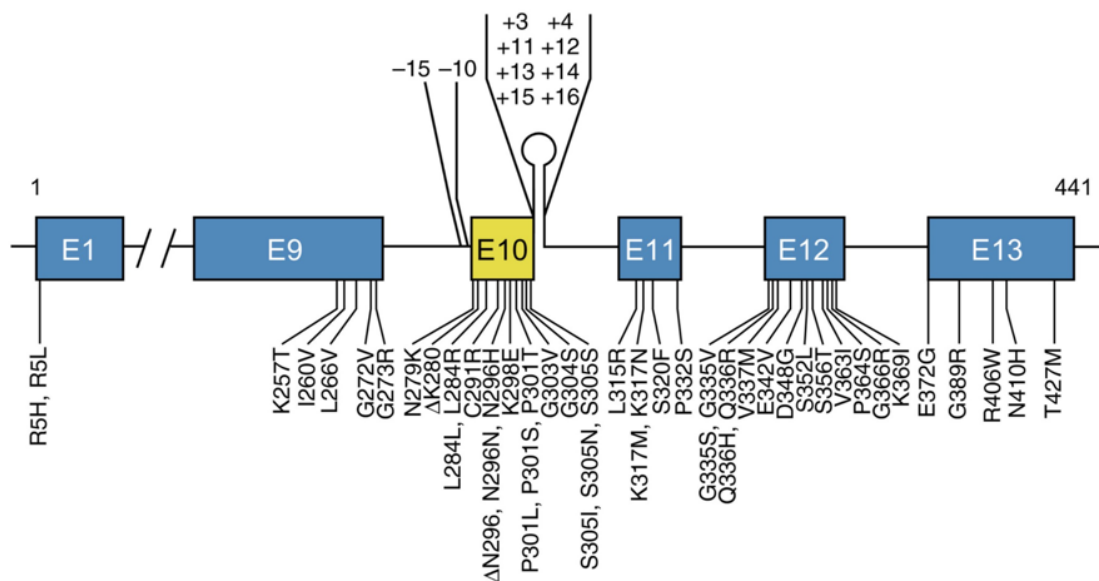
## 1.3 Genetics

### 1.3.1 Familial FTLD-tau due to autosomal dominant

#### *MAPT* mutations

The discovery of autosomal dominant pathogenic mutations in the *MAPT* gene encoding tau in 1998<sup>155–157</sup> proved that tau dysfunction itself was sufficient to cause neuronal toxicity and cell death in the tauopathies.

To date 72 pathogenic autosomal dominant *MAPT* mutations have been reported (<https://www.ftdtalk.org/what-is-ftd/genetics/mapt-mutations/>) (**Figure 1.11**) accounting for approximately 20% of all familial FTLD cases with a strong family history<sup>59,152,158</sup>. N279K, P301L and intron 10<sup>+16</sup> are the most common mutations. In cases of FTLD-tau there is a positive family history of parkinsonism, ALS, dementia or FTD in up to 40%, of which one third having a strong autosomal dominant pattern of inheritance<sup>30,159</sup>.



**Figure 1.11 - Disease causing *MAPT* mutations.** This figure shows forty-seven coding region mutations and ten intronic mutations flanking exon 10 (yellow) in *MAPT*. As of 2023 there are 72 known *MAPT* mutations, 15 of which are not shown here. *Reproduced with permission from Goedert 2018*<sup>33</sup>.

Historically, tauopathies due to *MAPT* mutations were considered a separate pathological entity named frontotemporal dementia with parkinsonism associated with

chromosome 17 (FTDP-17). More recently there has been a move towards classifying them as familial forms of the sporadic FTLT-tau subtypes<sup>160</sup>.

There are two broad categories of mutations depending on whether the mutations influence alternative splicing of pre-messenger RNA or has a primary affect at the protein level<sup>161</sup>. No correlation has been shown between known mutations and post-translational modifications of tau. The specific location and type of the *MAPT* mutation determines the ratio of 3R and 4R tau isoforms deposited (**Figure 1.12**), with the tau pathology is predominantly confined to neurons, or both neurons and glia<sup>54</sup>.

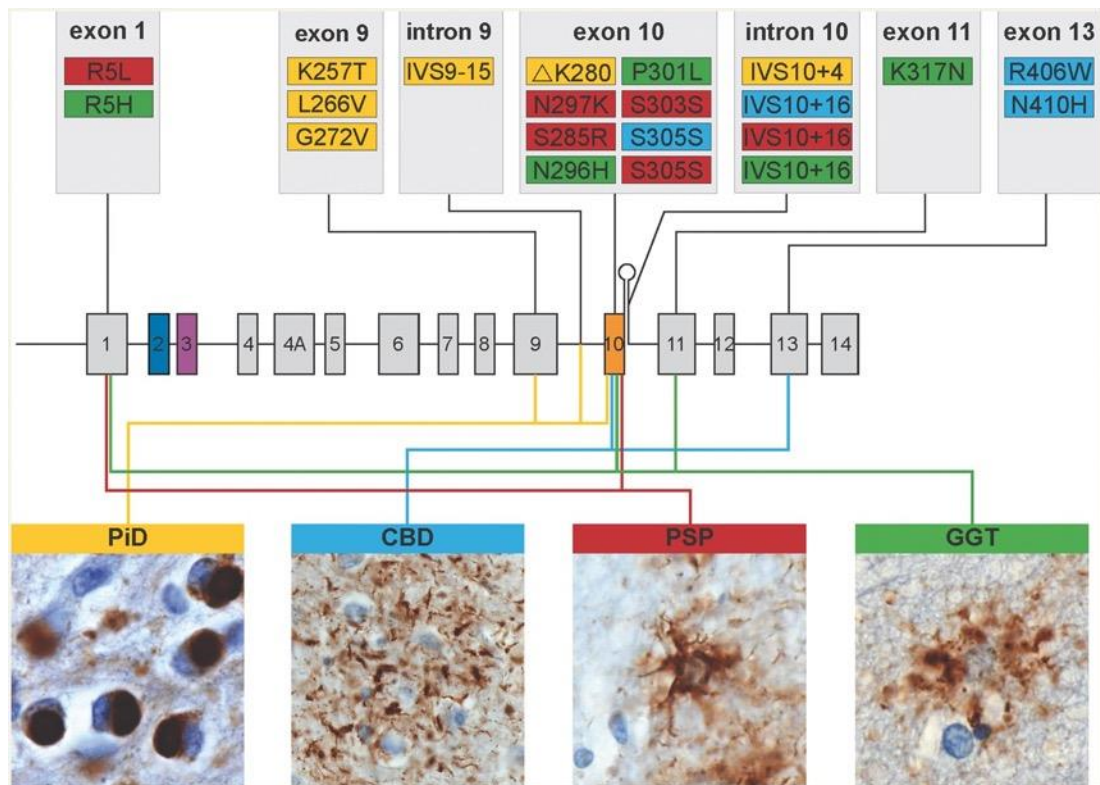
	Exon 1	Exon 9	Intron 9	Exon 10	Intron 10	Exon 11	Exon 12	Exon 13
	R5W, R5L	K257T, L266V, G272V		N279K, L284R, N296N, S305N, S305S	IVS10+3, IVS10+11, IVS10+12, IVS10+13, IVS10+14, IVS10+16	L315R, S320F	Q336R, E342V, K369I	G389R
Western blot bands	Isoform kDa AA 3R 64 381 4R 64 383 4R 68 412	Isoform kDa AA 4R 72 441 4R 68 412 3R 68 410 3R 64 381 3R 60 352		Isoform kDa AA 4R 72 441 4R 68 412 4R 64 383	Isoform kDa AA 4R 72 441 4R 68 412 4R 64 383	Isoform kDa AA 4R 72 441 4R 68 412 3R 68 410 4R 64 383 3R 64 381 3R 60 352	Isoform kDa AA 4R 72 441 4R 68 412 3R 68 410 4R 64 381 3R 60 352	Isoform kDa AA 4R 72 441 4R 68 412 3R 68 410 4R 64 381 3R 60 352
Type of filament		K257T, L266V, G272V: Gives rise to SFs and some twisted filaments with a cross-over spacing of 120 nm and a width of 6-23 nm		N279K: Filaments appear as twisted ribbons with a maximum width of 15 nm and a spacing between crossovers of 130-190 nm. P301L: Twisted ribbons with a maximum width of 15 nm and a cross-over spacing of 130 nm and SF 12 nm wide with stranded appearance	IVS10+3, IVS10+12, IVS10+13, IVS10+16: Twisted ribbons with a maximum width of 15 nm and a cross-over spacing of 130 nm and SF 12 nm wide with stranded appearance.	L315R, S320F: Gives rise to SFs and some twisted filaments with a cross-over spacing of 120 nm and a width of 6-23 nm	Q336R, E342V, K369I: Gives rise to SFs and some twisted filaments with a cross-over spacing of 120 nm and a width of 6-23 nm V337M: Gives rise to PHFs and SFs. PHFs have a diameter of 8-20 nm and a cross-over spacing of approximately 80 nm.	G389R: Gives rise to SFs and some twisted filaments with a cross-over spacing of 120 nm and a width of 6-23 nm R406W: Gives rise to PHFs and SFs. PHFs have a diameter of 8-20 nm and a cross-over spacing of approximately 80 nm.
Cellular inclusions	R5H, R5L: Neuronal and glial	K257T, L266V, G272V: Pick bodies and axonal inclusions. Note: L266V also has glial inclusions.		N279K, L284L, N296N, N296H, P301L, G303V, S305N, S305S: Neuronal and glial	IVS10+3, IVS10+11, IVS10+12, IVS10+13, IVS10+14, IVS10+16: Neuronal and glial	L315R, S320F: Pick bodies and axonal inclusions. Note: L315R also has glial inclusions.	Q336R, E342V, K369I: Pick bodies V337M: Neurofibrillary tangles	G389R: Pick bodies and axonal inclusions R406W: Neurofibrillary tangles

**Figure 1.12 - *MAPT* mutations and associated pathology.** Mutations are grouped by colour according to the predominant pattern of tau isoforms present in inclusions. Grey / Blue = all six isoforms are present, and filaments are paired helical and straight similar to AD. Peach = predominantly 3R tau isoforms. Pick bodies and axonal inclusions predominate, but some glial inclusion present. Green = predominantly 4R tau. Filaments are twisted ribbons and inclusions present in both neurons and glial cells. P301L mutations 3R also present in lower amounts. *Table is reproduced with permission from Ghetti et al.*<sup>54</sup>

PiD 3R tau inclusions are most commonly seen in exon 9 missense mutations (K257T, L226V and G272V) where a reduction in tau binding to microtubules enhances 3R over 4R tau assembly, or in exon 12 (P364S, Q336R, Q336H, K369I, G342V). Mutations in introns 9 (IVS9-15) and 10 (IVS10 + 4), exon 10 ( $\Delta$ K280), 11 (L315R, S320F, P332S), and 13 (G389R, R406W) have also been reported to cause Pick's pathology<sup>54</sup>. These mutations lead to deposits of predominantly 3R tau in neurons.

PSP pathology is more rarely associated with *MAPT* mutations (<10%) compared to greater than 20% in CBD<sup>30</sup>. PSP pathology is mainly associated with *MAPT* mutations in exon 10 (N297K, S285R, S303S, S305S), though also occasionally in intron 10

(IVS10 + 16) and exon 1 (R5L). CBD pathology is associated with mutations in exon 10 (S303S), intron 10 (IVS10 + 16) and exon 13 (N410H) <sup>54,160</sup> (**Figure 1.13**). Mutations in exons 1 and 10, as well as the surrounding introns, are associated with mixed neuronal and glial pathology. The glial pathology manifests as tufted astrocytes, astrocytic plaques and oligodendroglial coiled bodies, similar to that seen in PSP and CBD.



**Figure 1.13 - *MAPT* mutation location and association with sporadic FTLT-Tau pathology.** 3R PiD pathology is associated with point mutations in exon 9, intron 9 and a deletion in exon 10. Point mutations in exon 13 are associated with CBD. Point mutations in exon 10 and intron 10 give rise to increased 4R-tau pathology including CBD, PSP and GGT pathology. *Reproduced with permission from Forrest et al.*<sup>160</sup>

Interestingly both the age of onset<sup>162</sup> and the neuropathology associated with specific *MAPT* mutation can vary<sup>30,54,163</sup> suggesting that there are additional modifying factors that are currently unknown.

Recently a post-mortem study of individuals with microduplications of the 17q21.31 chromosomal region encompassing the *MAPT* gene<sup>164</sup>, has shown that *MAPT* duplication can also cause a primary tauopathy with diverse clinical and neuropathological features. These cases all developed a progressive disorder with severe memory impairment with or without behavioural changes, mimicking AD, and

neuropathological examination demonstrated tau aggregates with a range of 3R and 4R tau isoforms.

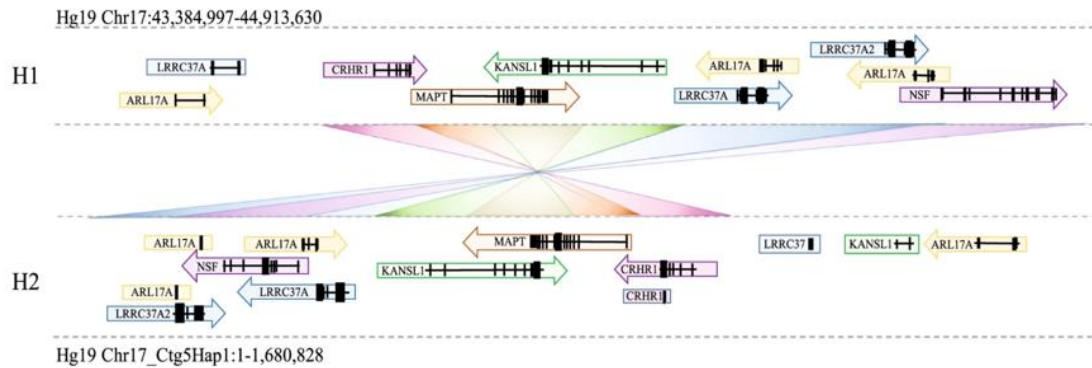
### 1.3.2 Genetic contributions to sporadic FTLD-tau

In genetic studies of familial neurodegenerative disease, the underlying proteinopathy is usually known and related to a specific clinical syndrome (for example familial AD due to *PSEN1*, *PSEN2* or *APP*). As previously discussed, there is significant heterogeneity in the FTLD primary tauopathies both in terms of the underlying pathology and the clinical presentation, complicating genetic studies in these disorders<sup>165</sup>. One approach is to assemble homogenous, pathologically diagnosed cohorts of patients to identify genetic loci that may inform on the underlying molecular pathogenesis<sup>166–169</sup>. This approach greatly increases the power to detect association for a given pathology, though is obviously offset by the rarity of pathologically confirmed cases. Another approach is to identify genetic variation associated with a clinical diagnosis, though this will provide insight on the selective vulnerability of particular brain networks that are affected in a particular syndrome, rather than the underlying pathology<sup>170</sup>. In the following sections I will focus on the genetic studies that have been performed in pathologically confirmed cohorts.

#### 1.3.2.1 *MAPT* Haplotype

The *MAPT* gene at the 17q21.31 locus sits within a 1.5Mb inversion region which is one of the most dynamic and complex regions of the human genome. In addition to *MAPT*, other genes associated with neural function and development are located within the inversion (*LRRC37A*, *LRRC37A2*, *NSF*, *ARL17A*, *ARL17B*, *KANSL1*, *SPPL2C*, and *CRHR1*). Two haplotypes exist at this locus, one in a direct (H1) and the second in an inverted (H2) orientation (**Figure 1.14**), with consistent differences in cortical gene expression demonstrated between the two<sup>171</sup>. These haplotypes were initially defined by Baker et al.<sup>172</sup>, with it later being shown that H1 could be an inversion polymorphism of the ancestral H2 haplotype<sup>173</sup>, with a second more recent inversion leading to the H2 haplotype seen in modern humans<sup>174</sup>. The absence of recombination between the inverted and non-inverted chromosomes has resulted in two haplotypes in complete LD<sup>174</sup>. H1 is the most common haplotype in Europeans with a frequency of 0.8, while H2, which appears to be under positive selection in the Finnish population

(frequency 0.2)<sup>173</sup>, is absent or at a lower frequency in East and South Asian populations (frequency 0-0.009)<sup>175</sup>. Whereas H1 is characterised by a higher frequency of polymorphisms (more than 20 common subhaplotypes)<sup>176</sup>, sequence diversity and variability are limited in the H2 haplotype<sup>177</sup>.



**Figure 1.14 - Structure of *MAPT* 17q21.21 locus.** Two distinct haplotypes, H1 and H2 (inverted with respect to H1), are defined at the locus. The direction of gene orientation in each haplotype are indicated by arrows. Each gene is labelled with a distinct colour and connected with a crossed rectangle between H1 and H2. Reproduced with permission from Bowles *et al.*<sup>175</sup>.

The H1 haplotype is associated with increased risk for a range of neurodegenerative diseases including *APOE*  $\epsilon$ 4-negative AD<sup>178</sup>, CBD<sup>169</sup>, PSP<sup>167,179</sup>, and PD<sup>180-182</sup>. Although it has been assumed that the signal at this locus is driven by the *MAPT* gene, there is growing evidence, at least in PD, that variation in genes within the inversion other than *MAPT* may be important (*KANSL1*<sup>183,184</sup>, *LRR37A/2*<sup>175</sup>). Although intuitively the role of *MAPT* seems more secure in the primary tauopathies, PSP and CBD, the findings in PD suggest that the role of other genes at the 17q21.21 locus should also be investigated. In contrast the evidence that the H2 haplotype is associated with neurodegenerative disease is inconclusive<sup>185,186</sup>. One study reported an association between H2 and familial FTD though this has not been replicated<sup>187</sup> while a second (again not replicated) suggested an association between H2 and clinically diagnosed FTD<sup>188</sup>. Given the high sequence homology in the H2 haplotype, it is predisposed to microdeletions through non-allelic homologous recombination between directly orientated duplications<sup>189</sup>, resulting in developmental delay and learning difficulties. The resulting 17q21.31 microdeletion syndrome (Koolen-de Vries syndrome) is caused by haploinsufficiency of *KANSL1*, a gene that encodes a chromatin modifier that influences gene expression<sup>190,191</sup>.

### 1.3.2.2 PSP and CBD

The first genetic risk factor identified in FTLT-tau disorders was the H1 *MAPT* haplotype in PSP<sup>172</sup>, which was subsequently found to also confer increased risk of CBD<sup>169,192,193</sup>. There have been two large-scale case-control genome-wide association studies (GWAS) to identify common risk variants in PSP. The first identified significant risk variants for PSP at *MAPT* (H1 haplotype and H1c sub-haplotype), *MOBP*, *STX6* and *EIF2AK3*, implicating genes involved in microtubule function, vesicle-membrane fusion, the unfolded protein response and myelin structure<sup>167</sup>. The odds ratio (OR) for PSP in those carrying the H1/H1 haplotype in this study was 5.5, which is higher than the OR for the ApoE  $\epsilon$ 3/4 genotype in AD ( $\sim$ 3.3)<sup>97</sup>. The second GWAS, which added an additional 600 neuropathologically diagnosed cases to the original 2011 GWAS cases, replicated the signals in *MAPT*, *STX6* and *MOBP*, as well as identifying new risk variants at *SLCO1A2* and *RUNX2*<sup>194</sup>, implicating genes involved in osteoblastic differentiation and solute trafficking via transporters at the blood brain barrier. A GWAS comparing PSP-RS with PSP-non-RS groups has shown that common variation at the *TRIM11* locus modifies PSP phenotype<sup>195</sup>. *TRIM 11* has a critical role in the clearance of misfolded proteins via the ubiquitin proteasome system (UPS)<sup>196</sup>, and it has been shown in mouse tauopathy models that tau accumulation is associated with decreased activity of the UPS<sup>197</sup>. More recently the same group have shown that variation close to the *LRRK2* locus is associated with decreased survival in PSP<sup>198</sup>, which is of particular interest given the development of *LRRK2* inhibitors for treatment of PD.

A case-control GWAS in pathologically confirmed cases of CBD showed genome wide associations at *MAPT* (H1 haplotype and H1c sub-haplotypes), as well as support for the *MOBP* locus increasing risk of disease<sup>169</sup>. The overlap in genetic risk between CBD, PSP and other FTLT syndromes has been further supported by a GWAS metanalysis<sup>199</sup>. SNPs tagging the *MAPT* haplotype overlapped between the three disorders, whilst, in addition, SNPs in or near *MOBP*, *CXCR4*, *EGFR*, and *GLDC* showed significant genetic overlap between CBD and PSP. More detailed analysis of *MAPT* sub-haplotypes in both CBD<sup>200</sup> and PSP<sup>201</sup>, suggests that the H1c and H1d subhaplotypes are associated with increased risk of both diseases, while the H2 haplotype is strongly protective. The increased 4R tauopathy disease risk associated with the H1c haplotype may be driven by increased *MAPT* gene expression and 4R tau

isoform ratios<sup>171,181</sup>. The protective association of the H2 haplotype is unknown but may at least in part be due to the increased inclusion of the N-terminal exon 3<sup>202</sup>; it has been shown that inclusion of exon 2 and 10 increase tau aggregation while inclusion of exon 3 decreases it<sup>203</sup>.

### **1.3.2.3 PiD**

Due to the overlap in clinical presentation with other pathologies within the FTLT spectrum, and the lack of in vivo biomarkers to diagnose PiD in life, genetic association studies have been limited. The largest study to date in pathologically confirmed PiD (34 cases, 215 controls), failed to show an association between H1/H2 haplotypes and disease<sup>186</sup>. This finding was in keeping with an earlier study by Russ et. al. which also failed to show an association<sup>185</sup>. There is indirect evidence, however, that there may be as yet unidentified variation at the *MAPT* locus that could contribute to the risk of developing PiD. First is the genetic overlap between PSP, CBD and FTLT syndrome at the *MAPT* locus in the GWAS metanalysis by Yokoyama et al.<sup>199</sup>. Secondly the largest GWAS to date in FTLT syndromes to date showed an association of the *MAPT* SNP rs8070723 (H2 haplotype protective) with bvFTD and progressive non-fluent aphasia subtypes<sup>204</sup>, the two most common clinical presentations of PiD<sup>62,65,67</sup>. Finally (as discussed in Familial FTLT-tau due to autosomal dominant *MAPT* mutations) there are a number of autosomal dominant *MAPT* mutations, most commonly in exons 9, 10, 11, 13, that are associated with Pick bodies at post-mortem.

### **1.3.3 3R Tau predominant pathology in other diseases**

There are two other diseases of interest in relation to 3R tau pathology. The first is myotonic dystrophy, an autosomal dominant disease of which there are two forms, myotonic dystrophy type 1 (DM1) and type 2 (DM2). The DM1 mutation consists of a trinucleotide CTG expansion in the 3' untranslated region of the *DMPK* gene and the DM2 mutation is a tetranucleotide CCTG expansion in the first intron of a gene encoding a nuclear protein (*ZNF9*)<sup>205</sup>. Although tau pathology has been reported in a case of DM2<sup>206</sup>, the link is more established in DM1 where although the brain distribution is heterogenous, the tau pathology has a 3R tau isoform predominance with alternative splicing and reduced inclusion of exon 2, 3 and 10<sup>207-209</sup>. Given the predominance of 0N3R tau in the foetal brain, this suggests that foetal isoforms of tau

are inappropriately expressed in the adult DM1 brain, possibly due to dysregulation of RNA splicing proteins resulting in pathological mis-splicing of tau in the brain<sup>209</sup>. In support of this hypothesis, an in-vitro study using either ectopic expression or siRNA of *MBLNI*, *CUGBP1* and *CELF1* (RNA splicing factors sequestered to pathological RNA foci containing mutant DMPK1 transcripts), implicates toxic gain of function in *CUGBP1* and *CELF1* as the main driver of tau mis-splicing in DM1<sup>208</sup>.

More recently a novel tauopathy has been described in individuals with tuberous sclerosis complex (TSC), an autosomal dominant disorder characterised by developmental delay, epilepsy and neurobehavioural dysfunction<sup>210</sup>. TSC can present with a neurodegenerative syndrome characterised by predominant right anterior temporal lobe atrophy that meets the clinical criteria for bvFTD<sup>211</sup>. “TSC” tauopathy is characterised by neurofibrillary tangles in the cerebral cortex, limbic system, subcortical and brainstem regions, with predominant tau acetylation and 3R tau isoform predominant inclusions. TSC is caused by pathogenic variants in either *TSC1* (coding for hamartin) located on chromosome 9q34 or *TSC2* (coding for tuberin) located on chromosome 16p13<sup>212</sup>. Both of these proteins form a poly-protein complex that inhibits the mammalian target of rapamycin complex 1 (mTORC1), the key regulator in the mTOR pathway that controls protein translation, autophagy and cell growth<sup>213</sup>.



## 1.4 Measuring disease progression

### 1.4.1 Failures of clinical trials

To date the only randomised clinical trials of disease-modifying treatments for the FTLD-tauopathies have been in PSP. Unfortunately, so far, no treatment has been shown to modify the disease course. Early, small ( $n < 100$ ), placebo-controlled trials of coenzyme Q10<sup>214,215</sup> showed no effect on disease progression. The first large international multicentre randomised controlled trial (RCT) testing Riluzole also failed to demonstrate benefit in terms of disease progression or functional outcome<sup>216</sup>. In recent years the focus of disease modifying treatments has shifted towards dysfunction of tau and its pathogenic spread. After it was shown that inhibition of glycogen synthase (GSK)-3 reduced tau phosphorylation in vitro and in vivo<sup>217</sup>, RCTs in lithium (ClinicalTrials.gov NCT00703677), sodium valproate<sup>218</sup> and Tideglusib<sup>219</sup> were carried out, though none showed significant improvement in the primary outcome. Davunetide, a compound that stabilises microtubules, also failed to demonstrate efficacy<sup>220</sup>. Most recently two RCTs of humanised monoclonal antibodies specific for the N-terminus of tau, tilavonemab<sup>221</sup> and gosuranemab<sup>222</sup> were stopped due to meeting futility criteria in phase 2. A trial involving an anti-sense oligonucleotide directly targeting *MAPT* expression (Ionis-MAPTRx, ClinicalTrials.gov: NCT03186989) is currently underway in AD, with a trial potentially occurring after this in *MAPT* mutation carriers. Alternative approaches expected to be trialled in the near future include targeting post translational modifications in tau, as well as tau vaccines<sup>223</sup>.

The failures of all disease modifying therapy trials so far in PSP leads to the question of what the reasons for this might be. For N-terminal anti-tau monoclonal antibodies it may be that either they do not target the specific tau species that drives neurodegeneration and the spread of pathology, or that not enough of the antibody enters the CSF<sup>224</sup>. This possibility is supported by a study in a transgenic mouse tau seeding model<sup>225</sup>; only antibodies targeting the mid-domain epitope of tau suppressed aggregation, whereas N-terminal antibodies did not. A monoclonal antibody UCB0107 (bepranemab), specific to the mid-domain is currently in phase 2 trials (UCB0107, ClinicalTrials.gov Identifier: NCT04658199).

Another issue is the difficulty of early detection of PSP in the absence of reliable biomarkers; early detection of the disease to allow entry into trials is essential to maximise the likelihood that the tau pathology (and consequent neurodegeneration) is mild enough that an intervention would be of adequate benefit<sup>226</sup>. The broadening of the MDS diagnostic criteria for clinical PSP has been a step in the right direction but introduces the complication of differences in rates of progression between different clinical subtypes<sup>105,107,227</sup> which could introduce bias and reduce statistical power.

A related concern is the accuracy of current clinical biomarker endpoints used in trial design. Clinical disease progression is still the gold standard for the primary outcome measure and is usually based on the PSP rating scale in combination with cognitive scales such as the Montreal Cognitive Assessment (MoCA) and functional scales such as the Schwab and England Activities of Daily living (SEADL). Although the PSP rating scale has been shown to be a good independent predictor of survival<sup>228</sup>, clinical biomarkers such as these are affected by intra- and inter-rater variability, as well as fluctuation in patient's clinical state. In addition, the PSP rating scale is designed for the classical presentation of PSP-RS, and the rate of change in score may well be different between different clinical subtypes<sup>229</sup>. Reliable, objective, disease progression markers are therefore required to complement these clinical ratings scales.

## **1.4.2 Methods to measure disease progression in vivo**

Accurate measurement of disease progression in the primary tauopathies (and neurodegenerative diseases in general) is essential to be able to assess the efficacy of therapeutic interventions, and better understand the underlying molecular pathogenesis. As discussed in the previous section, in the absence of reliable objective biomarkers, clinical scales are currently the gold standard to measure disease progression, though come with limitations. There is therefore an urgent need for individualised disease progression models based on clinical and objective biomarkers at baseline to better stratify homogenous populations for future clinical trials<sup>230</sup>. As defined by a recent working group such biomarkers need to “*correlate with or even anticipate clinical progression*” and show “*superiority over clinical measurements in terms of practicability, precision, effect size, or any combination of these*”<sup>229</sup>.

Biomarkers of disease progression can be divided into three main categories: neuroimaging, biological and neurophysiological. Given biological and

neurophysiological biomarkers are outside of the scope of this thesis, see the following reviews for more a more detailed summary of these<sup>97,230,231</sup>.

### **1.4.2.1 Neuroimaging biomarkers**

#### **1.4.2.1.1 Diagnostic**

A growing number of neuroimaging studies have been carried out in autopsy confirmed FTLD-tau subtypes PiD<sup>62,65,67,232</sup>, CBD<sup>137,233–236</sup> and PSP<sup>90,234,236,237</sup>. PiD usually has the most cortical atrophy, and greatest hemispheric asymmetry, with significant “knife-edge” bifrontal and anterior temporal lobe atrophy when presenting as bvFTD. When presenting as nvPPA, the second most common presentation, the atrophy involves the left inferior frontal gyrus, insula and orbitofrontal cortex. Over time the patterns of atrophy tend to merge<sup>65,238</sup>. In CBD there is typically asymmetrical focal atrophy greatest in the medial frontal and parietal lobes, with relative sparing of the anterior temporal lobes, and variable involvement of the subcortical structures<sup>137,233,235</sup>. However, there is a growing recognition that CBD pathology can be quite symmetrical, and the lack of asymmetry on imaging does not necessarily exclude the diagnosis<sup>66</sup>. Notably, the rates of global atrophy in CBD are significantly higher than in other tauopathies<sup>239</sup>. Typical PSP (PSP-RS) is characterised by significant atrophy in the brainstem and subcortical structures with additional involvement of the medial frontal regions<sup>240</sup>.

Given that PSP is the most common of the FTLD-tau pathologies, combined with the fact that PSP-RS has a highest clinico-pathological correlation with underlying PSP pathology<sup>73</sup>, the majority of neuroimaging studies have focused on this disorder. Within these studies the majority have focused on the clinical utility of structural MRI imaging as a diagnostic biomarker. While atrophy of the midbrain and superior cerebellar peduncles is a useful biomarker for differentiating PSP-RS from other parkinsonian syndromes<sup>241–243</sup>, it is not able to differentiate PSP-RS due to underlying PSP pathology from CBD pathology<sup>244,245</sup>, or PSP-RS from variant PSP syndromes<sup>246,247</sup>. The MRI-Parkinsonism Index (MRPI) has been shown to be a reliable biomarker for PSP-RS<sup>157,164</sup>, and shows good diagnostic accuracy for differentiating both PSP and 4R tau pathology (PSP and CBD) from other pathologies<sup>234</sup>. However, it still has the same limitations not allowing discrimination between CBD and PSP pathology, as well as showing poor sensitivity for the vPSP

syndromes. It is worth noting that the majority of the structural MRI imaging biomarker data in PSP has been obtained from patients late in the disease course presenting with PSP-RS<sup>240</sup>. There have only been two imaging studies so far of the vPSP syndromes<sup>249,250</sup> with only the study by Grivalja et al. assessing the accuracy of the MRPI in differentiating between the different syndromes. Further work will be required to assess how useful these markers are earlier on in the disease course in the different clinical subtypes.

Tau-PET imaging is a promising new tool and offers the opportunity for in vivo topographical mapping and quantification of tau pathology, as demonstrated in AD<sup>251</sup>. However, in contrast to in AD, the first generation of tau-radioligands are only thought to bind to 4R tau with low affinity<sup>252</sup>, and exhibit off target binding in the basal ganglia<sup>253</sup>. Newer tau ligands are now in development that aim to reduce off target binding<sup>254</sup>. Recently Tagai et al. have developed a new tau-radioligand, <sup>18</sup>F-PM-PBB3, that appears to bind both 3R and 4R tau, enabling identification of tauopathies in vivo in a range of FTLN syndromes<sup>255</sup>. Although further validation work will be required, this may open up an exciting new chapter in the diagnosis of disease specific tau pathology in vivo.

In the absence of a reliable tau-PET ligand ready for use in clinical practice, the use of structural imaging as an indirect measure of underlying tau pathology is supported by both animal and human studies. A pathology imaging study in PSP and CBD showed that in vivo structural imaging, compared to connectivity measures, better reflected the independent contribution from tau burden and neurodegeneration at autopsy<sup>237</sup>. In an <sup>18</sup>F-AV1451 PET AD study structural atrophy patterns were highly correlated with tau-PET differences at the group level<sup>256</sup>, while it has been shown in PSP that structural imaging outperforms <sup>18</sup>F-AV1451 as a longitudinal biomarker of disease progression<sup>257</sup>. Finally, in a mouse model of AD-like progressive tauopathy it has been demonstrated that propagation of tau pathology from the entorhinal cortex to the hippocampus is associated with tensor-based morphometry related atrophy<sup>258</sup>.

#### **1.4.2.1.2 Longitudinal**

Structural MRI has also shown utility as a biomarker of disease progression in the 4R tauopathies<sup>227,259–264</sup>. There are increased rates of midbrain, frontal and third ventricle atrophy in PSP-RS compared to controls, and given these changes can be picked up in

a one year period they offer a potential objective biomarker readout for clinical trials<sup>265</sup>. More recently a longitudinal study of PSP-RS and vPSP cases showed that there were significant differences both at one-year follow-up and in rates of change of PSPRS and gait-midline score between different clinical syndromes<sup>227</sup>. PSP-subcortical (PSP-parkinsonism, PSP-gait freezing, PSP-postural and PSP-corticospinal) showed the least impairment and slowest progression, which is in keeping with survival studies in neuropathologically confirmed PSP cases which have also demonstrated a longer disease duration in PSP-subcortical cases<sup>107</sup> and slower progression<sup>105</sup>. A recent study in PSP comparing the rate of midbrain atrophy with the change in the tau-PET ligand 18F-AV-1451 SUVR, showed that rate of midbrain atrophy correlated better with change in PSPRS score over a period of 12 months<sup>257</sup>. They concluded that rate of midbrain atrophy would be a more useful longitudinal biomarker for clinical treatment trials in PSP compared to this particular tau-PET tracer.

There have been fewer longitudinal imaging studies in either CBD or in PiD due to a combination of poor clinicopathological correlation, and the absence of in vivo biomarkers to enable accurate diagnosis of the underlying pathology in life. A longitudinal structural imaging study of PSP-RS and CBS, showed that there is greater baseline atrophy with greater longitudinal atrophy rates in the cortical and basal ganglia regions in CBS compared to PSP-RS<sup>240</sup>. In small longitudinal imaging study in pathologically confirmed PiD, rates of frontotemporal atrophy were faster in bvFTD-PiD than nfvPPA-PiD<sup>67</sup>.

#### **1.4.2.2 Data-driven disease progression modelling**

A major challenge in current tauopathy research is to construct models of disease progression using real-world data that are able determine, in vivo, the particular sequence and evolution of biomarker abnormality across the disease course. The ability to use this information to stage and stratify individual patients would help reduce heterogeneity in clinical trials, and more accurately monitor the effect of treatments on disease progression. The FTLT-tauopathies are particularly problematic both due to the heterogeneity of clinical phenotypes (subtype heterogeneity) each with differing rates of progression (temporal heterogeneity), as well the fact that pathology

and neurodegeneration is no-longer in the early stages by the time they are diagnosed clinically, and are enrolled into clinical trials<sup>240</sup>.

In the original hypothetical models of biomarker evolution in AD<sup>266</sup> the order of biomarker transition from normal to abnormal is decided a priori based on aggregated evidence from the literature<sup>267</sup>, rather than being learned directly from the data. Traditionally, longitudinal models attempt to delineate temporal heterogeneity by regressing a biomarker against a prespecified clinical measure; in PSP this can either involve use of the PSPRS or stages based on this score (PSP staging score)<sup>228</sup>. This approach assumes that all individuals have a homogenous phenotype, which as previously discussed is not always the case in the primary tauopathies. Subtype heterogeneity is often modelled using approaches such as clustering<sup>238,268–272</sup> to identify distinct groups, or grouping individuals *a priori* based on post-mortem pathology<sup>65,273–275</sup>. Either the assumption made is that all individuals are at a similar stage of the disease, or this needs to be imposed by selection based on *a priori* biomarker cut-offs.

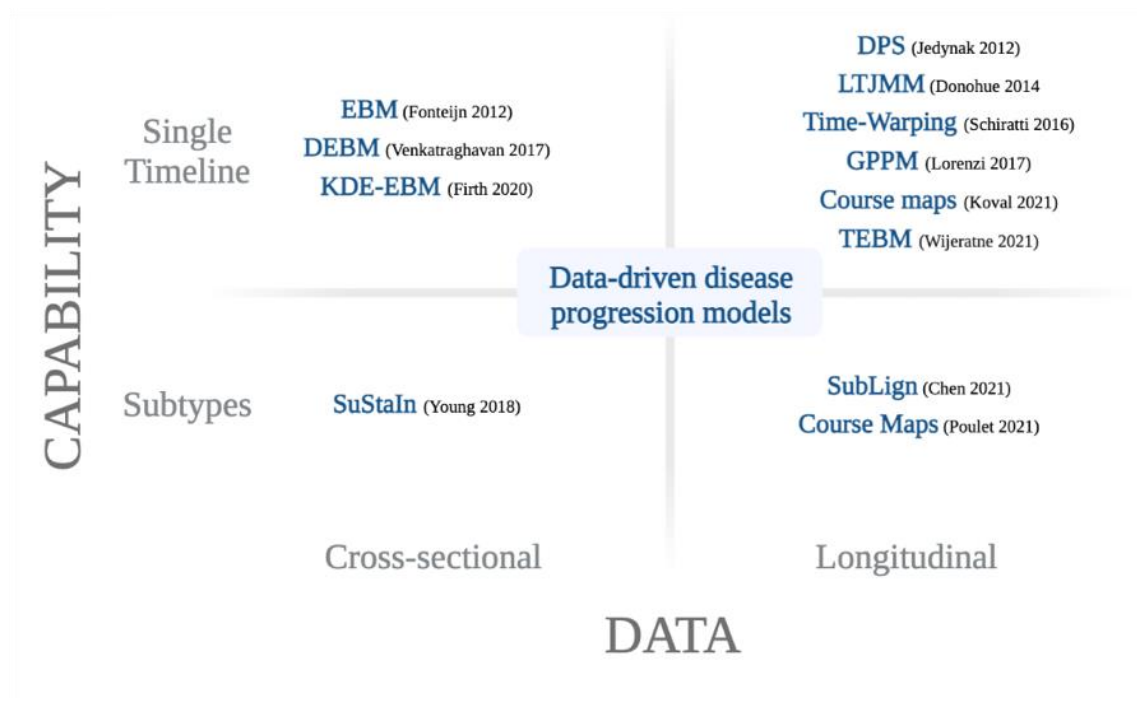
These challenges, combined with increasing availability of large datasets and intense theoretical debate in the neuroscience community regarding hypothetical biomarker models<sup>266,276,277</sup>, has driven the development of data-driven disease progression modelling (DDPM). DDPMs are defined by two key features; firstly, they simultaneously reconstruct the disease timeline and estimate the quantitative disease signature along this timeline and secondly, they are directly informed by the observed biomarker data<sup>278</sup>. For further detail on the history and development of DDPMs please refer to the review by Oxtoby and Alexander<sup>279</sup>. These new models are opening up exciting avenues to investigate neurodegenerative disorders in vivo with non-invasive biomarkers such as MRI, and to infer both temporal progression and disease subtype without a priori explicit biomarker cut points<sup>279,280</sup>.

#### **1.4.2.2.1 The current landscape of DDPMs**

An important step common to all model implementation is data pre-processing which entails firstly controlling for confounding variables (covariates) and secondly handling missing data. Missing data includes missing biomarker (input features) data due to one or more measurements not being performed, or irregularly sampled visits across individuals in the cohort. This step is important as otherwise one risks introducing bias

through the model learning non-disease related patterns that could be due to confounding. Some of the models can include covariates, whereas in others this is not possible. In this situation the covariates of concern can be regressed out, by training regression models on control data, and then adjusting the disease data using the residuals from the control regression. With regard to missing data, two main strategies can be deployed. The easiest approach is to remove those participants that have any missing biomarker or covariate data, though this can be sub-optimal due to it reducing the sample size available for analysis and potentially introducing sampling bias. The other approach can be to impute the missing data, either explicitly (e.g. using group mean values) or implicitly. Bayesian models are able to implicitly impute data by mapping data to probabilities and then dealing with the missing data probabilistically<sup>278</sup>; an example of this is in the event-based model (EBM), where a missing measurement ( $x$ ) is set to  $P(event) | x = 0.5$  which represents maximal uncertainty.

**Figure 1.15** summarises key DDPMs currently in use, categorising them by both model capability and model data requirements. In terms of model capability, they can either estimate a single timeline (top left quadrant) or multiple subtype timelines (bottom left quadrant), whilst for data requirements they can accept either cross-sectional data (pseudo-timeline) or longitudinal data (time-shift). Time-shifting describes deforming the individual data and “*stitching together*” them together into a quantitative template of disease progression.



**Figure 1.15 - Overview of current data-driven disease progression models.** All models estimate a disease timeline with some capable of simultaneously estimating multiple disease subtype timelines (bottom row of quadrant), using either cross-sectional data (creating a pseudo-timeline) or longitudinal data (time-shift). Abbreviations: EBM = event-based model, DEBM – discriminative EBM; KDE-EBM – kernel density estimation EBM, DPS = disease progression score, LTJMM = Latent Time Joint Mixed Model, GPPM = Gaussian Process Progression Model, SuStaIn = Subtype and Stage Inference, SubLign = Subtyping Alignment. TEBM = Temporal Event-Based Model. *Figure adapted from Oxtoby 2022<sup>278</sup>.*

In this PhD I was interested in trying to build a model that would have clinical utility for stratifying individuals at time of clinical trial entry. Given these individuals are likely to have only a baseline biomarker measurement at time of selection, I decided to focus on those DDDPMs that can use cross-sectional data only as input. Therefore, in the following sections I will discuss in more detail these types of DDPMs (the EBM and SuStaIn). For more detail on the other types of DDDPM highlighted in **Figure 1.15** please refer to the comprehensive review by Oxtoby 2022<sup>278</sup>.

#### 1.4.2.2.2 Event Based Modelling

One approach to modelling disease progression is the event-based model (EBM)<sup>281</sup>, a probabilistic data-driven generative model, that can infer the order in which biomarkers become abnormal directly from cross-sectional data by combining information across biomarkers and individuals without reference to a given individual's clinical status<sup>267</sup>. The EBM is able to extract longitudinal information about disease progression, by assuming 1) that in a patient cohort containing a spectrum of disease stages, more individuals will show abnormality in a biomarker

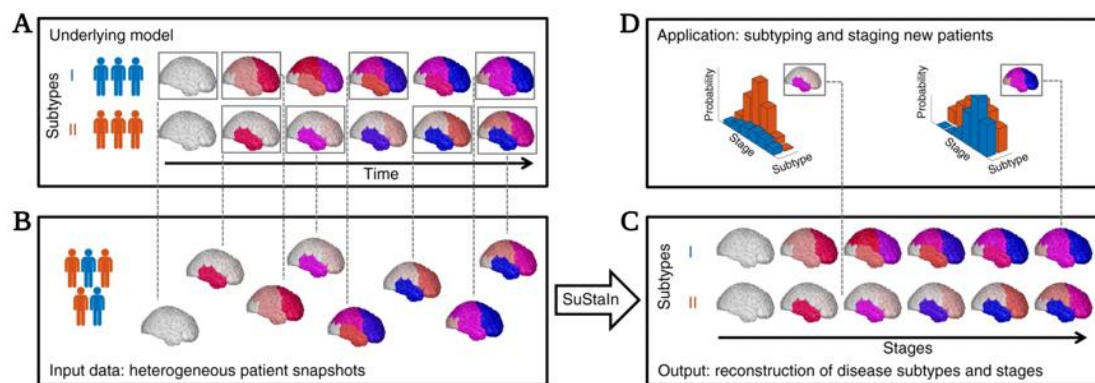


that changes early in the disease course, 2) that there is monotonic progression of an individual biomarker from normal to abnormal and 3) homogenous disease progression with all individuals having a broadly similar disease progression pattern. The EBM is suited to analysis of data from research studies that have strict inclusion and exclusion criteria to enable enrichment for one disease of interest, and provides a powerful, directly interpretable model that can be applied to stage individuals with only a baseline biomarker measure. This approach has been successfully applied to Huntington's disease<sup>267</sup>, sporadic and familial AD<sup>280,282,283</sup>, the posterior cortical atrophy variant of AD<sup>284</sup>, multiple sclerosis (MS)<sup>285</sup>, PD<sup>286</sup> and to ALS<sup>287</sup>, providing a simple and validated method to investigate temporal disease patterns and estimate individuals' disease stage. Recent work has demonstrated the clinical utility of the EBM for screening patients on entry into clinical trials, to improve cohort homogeneity and increase the power to detect a treatment effect<sup>288</sup>.

In Chapter 2 (**The Event Based Model (EBM)**) I will discuss the technical aspects of the EBM in more detail, and summarise the main steps required to fit the model, stage individual patients and for cross validation.

#### **1.4.2.2.3 Subtype and Stage Inference**

A limitation of the EBM is that similar to traditional longitudinal models, the assumption is that all individuals follow a similar disease trajectory i.e. are the same phenotype. To this end the EBM has recently been extended to create new unsupervised machine learning algorithm called Subtype and Stage Inference (SuStain) (Young 2018)<sup>289</sup>, that is able to disentangle temporal and phenotypic heterogeneity to identify population subgroups with distinct patterns of disease progression (**Figure 1.16**). SuStaIn has proven to be a powerful and versatile technique providing insights into subtypes with distinct patterns of biomarker evolution across a range of different diseases including AD<sup>290-292</sup> genetic FTD<sup>290,293</sup>, MS<sup>294</sup>, and chronic lung disease<sup>295</sup>. A recent study applied SuStaIn to tau-PET imaging in AD, identifying four distinct spatiotemporal trajectories of tau progression, with important implications for the notion of "typical AD" and tau pathological staging<sup>296</sup>. The heterogeneity of both the clinical phenotype, as well as underlying tau pathology, in the FTLD-tau disorders, make these disorders ideally suited to investigation with these advanced modelling techniques.



**Figure 1.16 - Conceptual overview of SuStaIn.** (A) shows the underlying model consisting of a patient cohort with two hypothetical disease subtypes. (B) Cross-sectional input data from the cohort contains snapshots of biomarker measurements (in this cases imaging data) from each individual who has an unknown subtype and stage. (C) SuStaIn recovers the underlying set of disease subtypes and their temporal progression patterns via simultaneous clustering and z-score disease progression modelling. (D) The SuStaIn model fitted on the cross-sectional baseline data (C), can then be used to estimate the probability that an unseen biomarker measurement belongs to each stage and subtype. The colour of each region represents the amount of pathology in that region ranging from white (no pathology) to red to magenta to blue (maximum pathology). *Figure reproduced with permission from Young et al<sup>290</sup>.*

In their seminal SuStaIn paper, Young et al.<sup>290</sup> also introduce the linear z-score model, which reformulates the EBM by replacing the instantaneous normal to abnormal events with events that represent the linear transitions between a biomarker's z-scores. This obviates the need for mixture modelling, and allows finer grained inference to be performed across the range of biomarker abnormality.

In Chapter 2 (**Subtype and Stage Inference (SuStaIn)**) I will discuss the technical aspects of SuStaIn in more detail, and summarise the main steps required to fit the model and subtype and stage individual patients.

## 1.5 Thesis aims

As covered in this chapter, the FTLD-tauopathies represent a clinically and pathologically heterogeneous group of disorders. Despite ongoing efforts there are currently no disease modifying therapies for these devastating diseases. With the advent of experimental therapies directly targeting pathological tau, there is an urgent need to develop individualised disease progression models to more accurately stratify patients selected for clinical trials, and monitor disease progression. Probabilistic data-driven disease progression models, validated in more common neurodegenerative diseases such as AD, provide a robust and unbiased method to better understand this heterogeneity. Although in recent years genetic studies in the 4R tauopathies (PSP and CBS) have increased our understanding of the biological determinants of phenotypic variation and disease progression, very little progress has been made in PiD mainly due to its rarity and non-specific clinical presentations. Identification of genetic susceptibility loci for PiD pathology and progression may provide novel mechanistic targets for therapeutic intervention, as well as guide biomarker development to be able to diagnose PiD pathology in life.

Broadly speaking this thesis is split into two parts; the first part is focused on disease progression modelling in the 4R tauopathies (PSP and CBS), and the second part is focused on genetic analysis of the 3R tauopathy, PiD. Within this framework, the main aims of the thesis are:

### **A. Disease progression modelling in the 4R tauopathies:**

1. To develop an image-based model of brain volume biomarker changes in PSP-RS to enable in vivo staging of disease (**Chapter 3**)
2. To investigate the spatiotemporal heterogeneity of atrophy in the 4R tauopathy syndromes and identify population subgroups with distinct patterns of disease progression:
  - a. PSP (**Chapter 4**)
  - b. CBS (**Chapter 5**)

## **B. Genetics analyses in the 3R tauopathy PiD:**

To use genetic studies to understand the biological determinants of disease and progression in PiD; specifically:

1. Investigate the association of *MAPT* haplotypes with risk of PiD (**Chapter 6**).
2. Perform a genome wide association study to identify genetic risk factors for developing PiD (**Chapter 7**)
3. Perform a genome wide survival study to identify genetic risk factors for decreased survival in PiD (**Chapter 8**)

# **Chapter 2: Cohorts and General Methodology**

## **2.1 Introduction**

As discussed in the previous chapter this thesis is broadly split into two main parts: disease progression modelling of the 4R tauopathies (PSP and CBS), and genetic analyses of the 3R tauopathy PiD. As part of the PhD, I had to build two separate cohorts to generate large enough sample sizes to facilitate the analyses required: the 4R tauopathy imaging cohort, and the Pick's disease international consortium (PIC), Chapters 3, 4 and 5 of this thesis use data from the 4R tauopathy imaging cohort, whilst chapters 6, 7 and 8 use data from the PIC. I have therefore split this chapter into two sections, each of which contains the details of the relevant cohort, and general analysis methods used on that dataset.

## **2.2 4R tauopathy disease progression modelling**

### **2.2.1 4R tauopathy imaging cohort**

Individuals with a clinical diagnosis of either “possible” or “probable” PSP (including PSP-RS and vPSP syndromes) or CBS were collected from seven main studies and included in the cohort: the 4R Tauopathy Imaging Initiative Cycle 1 (4RTNI 1; ClinicalTrials.gov: NCT01804452)<sup>259,263</sup>, the 4R Tauopathy Imaging Initiative Cycle 2 (4RTNI 2; ClinicalTrials.gov: NCT02966145), the davunetide randomized control trial (DAV; ClinicalTrials.gov: NCT01056965)<sup>220</sup>, the salsalate clinical trial (SAL; ClinicalTrials.gov: NCT02422485)<sup>297</sup>, the young plasma clinical trial (YP; ClinicalTrials.gov: NCT02460731)<sup>297</sup>, the PROgressive Supranuclear Palsy CorTico-Basal Syndrome Multiple System Atrophy Longitudinal Study (PROSPECT; ClinicalTrials.gov: NCT02778607)<sup>298</sup>, and the University College London Dementia Research Centre (UCL DRC) FTD cohort. Controls were collected from three cohorts with equivalent available data; PROSPECT, the UCL DRC FTD cohort and the Frontotemporal Lobar Degeneration Neuroimaging Initiative dataset (FTLDNI; <http://4rtni-ftldni.ini.usc.edu/>). Controls were defined as no known diagnosis of a neurological or neurodegenerative condition and no known history of memory

complaints. Appropriate ethics was applied for and approved via the relevant trial ethics committees.

To be included in my cohort, all participants needed to have, as a minimum, a clinical diagnosis of either PSP (PSP-RS or vPSP) or CBS, a baseline T1 volumetric MRI on a 1.5 or 3 Tesla scanner, basic demographic data (gender and age at time of scan). Age at symptom onset, age at death, clinical rating scale scores (PSP rating scale, Unified Parkinson Disease Rating Scale [UPDRS], Schwab and England Activities of Daily Living scale [SEADL], and Montreal Cognitive Assessment [MoCA] or Mini-mental State Examination [MMSE] at baseline and follow-up), pathology at autopsy, CSF AD biomarker positivity [ $A\beta$ 1–42, tau, and ptau], amyloid PET positivity (with florbetaben, florbetapir, or Pittsburgh Compound-B), and follow-up scans were also included if available. Amyloid PET scans were collected at participating 4RTNI centres with demonstrated experience in FDA-approved amyloid imaging agents, and positivity was defined by expert visual read by certified staff.

## **2.2.2 Participants, inclusion criteria, clinical and MRI data**

### **2.2.2.1 4RTNI1 / FTLDNI trials**

Participants were recruited as part of two longitudinal observational neuroimaging studies; 4RTNI1 which enrolled PSP and CBS patients and FTLDNI which recruited healthy controls. Both trials were managed by the University of California (UCSF) with patients also recruited at University of California of San Diego (UCSD), University of Toronto (UToronto) and Massachusetts General Hospital (MGH). A common study design and protocol was run at all sites. Patients with PSP met the NINDS-SPSP criteria<sup>72</sup>, while CBS patients met the Armstrong criteria for possible or probable CBS-CBD subtype<sup>140</sup>: all participants had to be aged between 45-90yrs to be considered for inclusion. Participants were evaluated at baseline, 6 and 12 months with an MRI scan, and a clinical assessment that included a PSPRS score, SEADL, UPDRS, MOCA and MMSE.

Three scanner types (all 3T) were used. At UCSF and MGH a Siemens Trio Tim system (Siemens, Iselin, NJ) with a 12-channel receiver head coil was used; whole brain images were acquired with a volumetric magnetization prepared rapid gradient-

echo sequence (MPRAGE; repetition time (TR)/echo time (TE)/inversion time (TI) = 2300/2.98/900 ms,  $\alpha = 9^\circ$ , 1 x 1 mm in-plane resolution, 1 mm slice thickness). Scans at UCSD were acquired on a GE Discovery MR750 system (GE, Milwaukee, WI) equipped with 32-channel head coil, and scans at UToronto were acquired on a GE Signa HDx system equipped with an 8-channel receiver head coil. Whole brain images at both UToronto and UCSD were acquired sagittally with a 3D inversion-recovery prepared spoiled gradient echo imaging pulse sequence (UCSD IR-SPGR; TR/TE/TI = 7.1/3.00/400 ms,  $\alpha = 11^\circ$ , 1 x 1 mm in-plane resolution, 1.2 mm slice thickness; UToronto IR-SPGR; TR/TE/TI = 7.0/2.80/400 ms,  $\alpha = 11^\circ$ , 1 x 1 mm in-plane resolution, 1.2 mm slice thickness)<sup>259</sup>. For each patient, baseline and follow-up MRI were acquired on the same scanner using the same sequence parameters.

A total of 111 patients (62 PSP-RS and 49 CBS) met criteria for inclusion in my cohort. Of these 92 (83%) had a 6-month follow-up scan, and 68 (61%) has a 12-month follow-up scan. 143 age and gender matched controls with a baseline scan were identified from FTL DNI and included in our cohort.

#### **2.2.2.2 4RTNI2 trial**

Participants were recruited from eight sites across North America, as part of a longitudinal observational study of Corticobasal Syndrome (CBS), Progressive Supranuclear Palsy (PSP) or Oligo- or Variant- Progressive Supranuclear Palsy (o/vPSP). The study was managed by the University of California San Francisco (UCSF) with patients also recruited from the University of California of San Diego (UCSD), John Hopkins University, Harvard University Massachusetts General Hospital, the Mayo Clinic Rochester, Columbia University, University of Pennsylvania, and the University of Toronto. A common study design and protocol was run at all sites, and patients were diagnosed with a clinically probable or possible PSP syndrome according to the Movement Disorder Society 2017 PSP diagnostic criteria<sup>98</sup>. All participants had to be aged between 45-80yrs to be considered for inclusion, and were evaluated at baseline, 6, 12 and 24 months with a volumetric MRI brain scan, a clinical assessment that included a PSPRS score, UPDRS, and MOCA, Tau and Amyloid PET scans, eye movement function and retinal imaging.

Four scanner types (all 3T) were used in the 4RTNI2 study. A GE Discovery MR750 system (GE, Milwaukee, WI) equipped with 8-channel head coil; whole brain images

were acquired with a 3D volumetric inversion recovery fast spoiled gradient-echo recalled sequence (IR-FSPGR; repetition time (TR)/echo time (TE)/inversion time (TI) = 2300/2.98/400 ms,  $\alpha = 11^\circ$ , 1.2 mm slice thickness). A GE Signa Premier System (GE, Milwaukee, WI) equipped with 48-channel head coil, whole brain images were acquired with a 3D volumetric inversion recovery fast spoiled gradient-echo recalled sequence (IR-FSPGR; repetition time (TR)/echo time (TE)/inversion time (TI) = 2300/2.98/900 ms,  $\alpha = 8^\circ$ , 1.2 mm slice thickness). A Philips Achieva system (Andover, Massachusetts, United States) equipped with 32-channel head coil, whole brain images were acquired with a volumetric magnetization prepared rapid gradient-echo sequence (MPRAGE; repetition time (TR)/echo time (TE)/inversion time (TI) = 2300/2.98/900 ms,  $\alpha = 9^\circ$ , 1 mm slice thickness). A Siemens Magnetom Prism Fit system with a 32-channel receiver head coil was used; whole brain images were acquired with a volumetric magnetization prepared rapid gradient-echo sequence (MPRAGE; repetition time (TR)/echo time (TE)/inversion time (TI) = 2300/2.91/900 ms,  $\alpha = 9^\circ$ , 1.2 mm slice thickness). For each patient, baseline and follow-up MRI were acquired on the same scanner using the same sequence parameters.

A total of 167 individuals met the criteria for inclusion in the study; 104 PSP cases (60 PSP-RS and 44 PSP-cortical) and 63 CBS cases (19 CBS-AD and 44 CBS-IDT cases). There were a total of 113 follow-up scans at varying timepoints (61 PSP and 52 CBS) between 1-3 years post baseline scan.

### **2.2.2.3 DAV trial**

Patients were recruited from 48 centres in Australia, Canada, France, Germany, the UK, and the USA, between September 2010 and November 2012<sup>220</sup>. For inclusion participants had to be aged between 41-85 at disease onset and meet modified PSP criteria from the national Neuroprotection and Natural History in Parkinson Plus Syndromes (NNIPPS) study for the most common clinical presentation PSP-RS<sup>216</sup>. Specifically, they had to have at least a 12-month history of postural instability or falls during the first three years from disease onset, supranuclear ophthalmoplegia or reduced downward saccade velocity, and prominent axial rigidity. In addition, at time of screening participants were required to have a MMSE score of at least 15, be able to ambulate independently (or walk five steps with minimal assistance), live outside a dementia care facility, have PSP symptoms for either less than five years, or more than



five years with a PSPRS score of no more than 40. More detailed inclusion and exclusion criteria are included in the original study manuscript<sup>220</sup>.

Primary endpoints were the change in PSPRS and SEADL between baseline and twelve-month visit, with secondary outcomes including Clinical Global Impression of Change (CGIC), and MRI measured ventricular volume. MRI data was collected on fourteen different 1.5T or 3T scanners with varying scanner types but consistent sequences based on standards set by the Mayo Clinic's Aging and Dementia Imaging Research laboratory (Rochester, MN, USA). All T1 images in the trial were acquired were one of MPRAGE, Coronal IR-SPGR, or Sagittal IR-SPGR. For each patient, baseline and follow-up MRI were acquired on the same scanner using the same sequence parameters. Details of scanners and acquisition protocols are included in the following references<sup>220,259</sup>.

I had access to data from 226 patients from this trial, of which 187 PSP cases met the criteria for inclusion into the 4R tauopathy imaging cohort with at least a baseline MRI scan and a baseline PSPRS score. Of the 187 included, 177 (98%) had a 12 month follow-up scan.

#### **2.2.2.4 SAL / YP trials**

The SAL trial recruited from the University of California San Francisco (UCSF; San Francisco, CA) Memory and Aging Center and the Oregon Health and Science University (Portland, OR) Parkinson Center & Movement Disorder Program between June 2015 to February 2018. For the YP trial, patients were recruited from UCSF, and the trial ran from June 2015 to August 2017. Individuals included in both studies had PSP-RS as defined by the 2017 International Parkinson and Movement Disorder Society criteria for PSP-RS<sup>98</sup>, were aged 50 to 85 years; had a MMSE score of 14-30, an MRI consistent with PSP, and were on stable medications at least 1 month before screening, except for approved AD and PD medications. For more detail on inclusion and exclusion criteria please refer to original study manuscript<sup>297</sup>.

Given these were phase one open label trials, the primary outcome measure was safety and tolerability. In addition, PSPRS (among other clinical scales) and an MRI were collected at baseline, three and six months follow-up (after drug treatment). Structural MRIs were acquired on a 3T Siemens Trio Trim or a 3T Siemens Prisma-Fit scanner (Siemens Healthineers AG, Erlangen, Germany). On the Trio Tim the following

acquisition protocol was used; T1 MPRAGE sequence with slice thickness 1mm, with TR of 2.3s, TE of 2.98 ms, and T1 of 900ms. The Prisma Fit acquisition protocol was identical to that on the Trio Tim. For each patient, baseline and follow-up MRI were acquired on the same scanner using the same sequence parameters.

In total eight PSP patients from the SAL trial and six from the YP trial had at least a baseline MRI with baseline PSPRS score and were included in our cohort. All participants (100%) had a two follow-up MRIs at three and 6 months with and PSPRS recorded at all visits.

### **2.2.2.5 PROSPECT trial**

The PROSPECT observational study recruits patients from seven main UK study sites; University College London (UCL), Cambridge, Oxford, Newcastle, Manchester, Brighton and Newport. Recruitment started in September 2015 and is ongoing. Inclusion into the study was originally defined for PSP according to the NINDS-SPSP criteria<sup>72</sup>. At the end of baseline recruitment all cases were reclassified according to the 2017 MDS clinical PSP diagnostic criteria<sup>98</sup>. All PSP cases met the criteria for at least “possible” PSP, and were stratified into PSP-RS, PSP cortical (PSP-CBS, PSP-SL, PSP-F) and PSP subcortical (PSP-P, PSP-PGF, PSP-oculomotor); stratifying in this way is an established approach in the PSP research setting<sup>107,195,198,299</sup>. CBS was diagnosed according to the Armstrong criteria<sup>140</sup>. CBS cases with autopsy confirmed with CSF biomarkers consistent with AD were classed as CBS-AD, CBS-CBD or CBS-PSP based on neuropathology at post-mortem, or CBS-indeterminate (IDT) if CSF status / autopsy diagnosis was unknown. Recruited control participants included a spouse or a friend of the case or came through the Join Dementia Research volunteer registry<sup>298</sup>.

Study assessments including a PSPRS score<sup>228</sup>, a modified MDS Unified Parkinson’s Disease Rating Scale (UPDRS)<sup>300</sup>, SEADL<sup>301</sup>, and cognitive tests including the Montreal Cognitive Assessment (MoCA)<sup>302</sup> and Addenbrooke’s Cognitive Examination3 (ACE-III)<sup>303</sup> were performed at baseline and follow-up visits (6, 12, and 24 months). Participants had volumetric weighted MRI on Siemens 3T scanners; either a Magnetom Skyra, Magnetom Prisma, or Tim Trio. Scan protocols were designed at the outset of the study to closely match across centres, based on the international Genetic Frontotemporal Dementia Initiative protocols (MPRAGE, TR 2s, TE 2.93ms,

Flip angle 8 degrees, 1.1mm isotropic)<sup>113</sup>. For each patient, baseline and follow-up MRI were acquired on the same scanner using the same sequence parameters.

82 PSP cases with baseline MRI scans met criteria to be included in the cohort from PROSPECT; 57 PSP cases (35 PSP-RS, 9 PSP-cortical, 14 PSP-subcortical), and 25 CBS cases (1 CBS-CBD, 11 CBS-AD and 13 CBS-4RT). 31 participants had a follow-up MRI at either 1 or 2 years. 36 age and gender matched controls with baseline scans from PROSPECT were included in the control cohort.

#### **2.2.2.6 UCL DRC FTD Cohort**

I reviewed the UCL DRC FTD cohort MRI database to identify patients with a clinical diagnosis of either PSP or CBS, and a good quality T1-weighted MRI scan. Patients were diagnosed as PSP-RS according to the NINDS-SPSP criteria<sup>72</sup> if diagnosis had been before 2017, or as a PSP syndrome (PSP-RS or vPSP) according to the MDS 2017 PSP clinical diagnostic criteria<sup>98</sup> if diagnosed from 2017 onwards. CBS patients were diagnosed according to the Armstrong Criteria<sup>140</sup> as and all met the criteria for “probable or “possible” CBS. All patients included were between the age of 42 and 85 years. All patients had initially undergone a standard clinical assessment at the National Hospital for Neurology and Neurosurgery in a specialist cognitive disorders or movement disorder clinic, depending on their initial clinical presentation. Age and gender matched controls were also identified from this database and included in the control cohort.

T1-weighted MRIs were acquired between 1992 to 2014 on three different scanners: a 1.5T Signa scanner (GE Medical systems, Milwaukee, WI, TR = 12 ms, TI = 650 ms, TE = 5 ms, acquisition matrix =  $256 \times 256$ , spatial resolution = 1.5 mm), a 3T Trio Tim (Siemens, Erlangen, Germany, TR = 2200 ms, TI = 900 ms, TE = 2.9 ms, acquisition matrix =  $256 \times 256$ , spatial resolution = 1.1 mm), and a 3T Prisma scanner (Siemens, Erlangen, Germany, TR = 2,000 ms, TI = 850 ms, TE = 2.93 ms, acquisition matrix =  $256 \times 256$ , spatial resolution = 1.1 mm, acquisition plane=sagittal). A total of 43 patients (25 PSP-RS, 3 PSP-P, 1 PSP-PGF, and 14 CBS), and 141 controls with baseline scans met the criteria for inclusion. 20 of the 43 individuals (47%) had a follow-up scan at varying timepoints between 1-3 years post baseline scan.

### 2.2.2.7 Summary of 4R tauopathy imaging cohort

**Table 2.1** gives a summary of the total number of baseline and follow-up scans included in the 4R tauopathy imaging cohort, broken down by contributing study. I detail in the top row the total number that met the inclusion criteria, and in the second row the number that passed quality control (QC) (I detail the QC process in **Image processing pipeline** below) and were therefore available for analyses. The rest of the rows in this table breakdown the number of scans by clinical diagnosis (PSP vs. CBS). In total there were 561 baseline scans that passed QC; 426 of these had a diagnosis of PSP, and 135 CBS. PSP-RS was the most common PSP clinical syndrome, with 52 cases PSP-Cortical (PSP-C) and 17 cases PSP-Subcortical (PSP-SC); 135 had a diagnosis of CBS with 12 having a pathological diagnosis of CBS-CBD, 6 of CBS-PSP, 34 of CBS-AD and the remaining 83 defined as indeterminate (CBS-IDT). Of those PSP cases that came to post-mortem (31/426) 94% were pathologically confirmed PSP, one had GGT and the other had CBD pathology.

**Table 2.1 - 4R Tau imaging cohort summary**

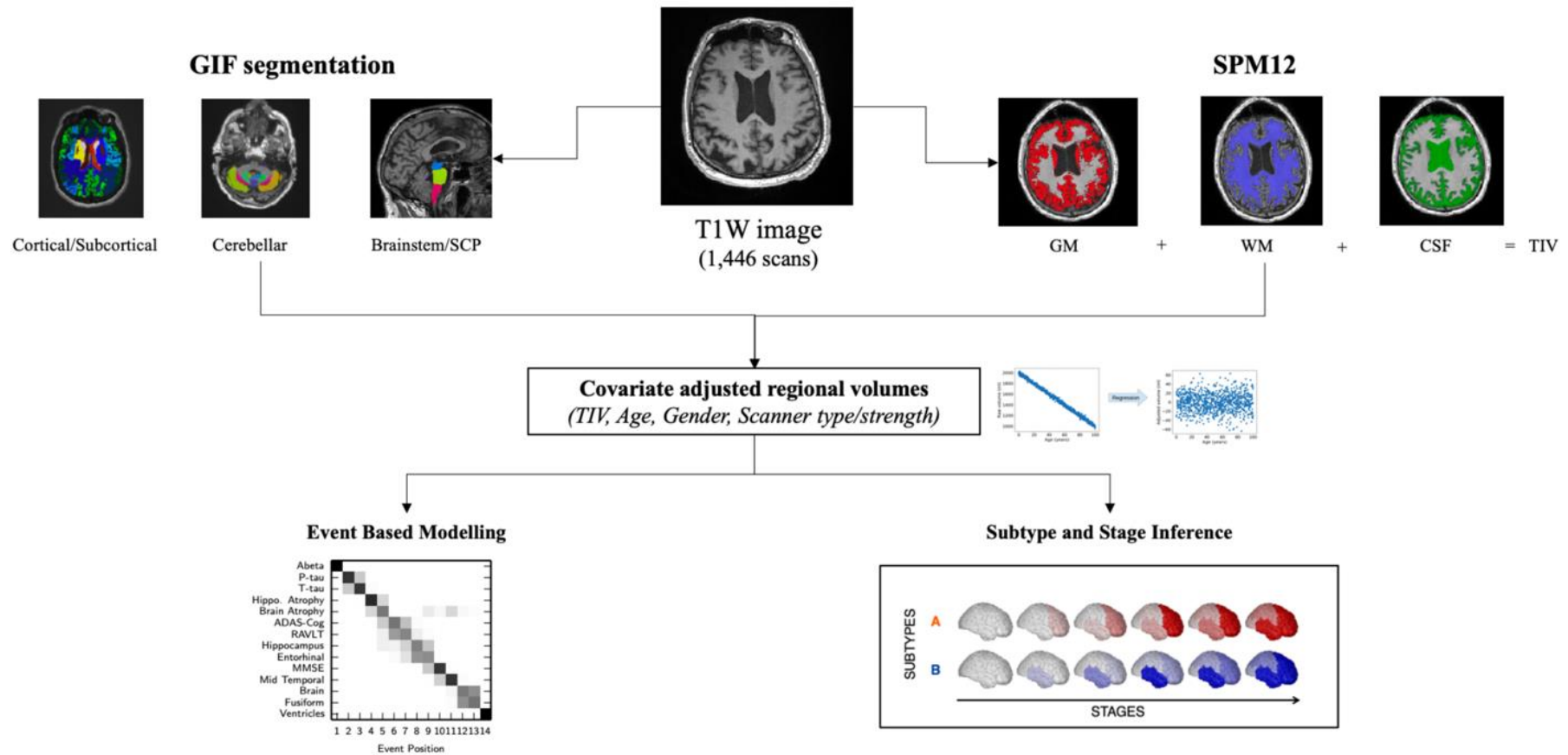
	Controls <sup>a</sup>	All	4RTNI1	4RTNI2	DAV	YP/SAL	PROSPECT	UCL
Inclusion, n (fu visits)	320	604 (522)	111 (142)	167 (113)	187 (177)	14 (28)	82 (34)	43 (28)
Post QC, n (fu visits)	290	561 (480)	103 (137)	160 (109)	173 (157)	13 (24)	74 (31)	38 (22)
PSP, n (fu visits)	-	426 (367)	59 (83)	100 (58)	173 (157)	13 (24)	52 (25)	29 (20)
- PSP-RS	-	357 (329)	59 (83)	59 (34)	173 (157)	13 (24)	30 (17)	23 (14)
- PSP-C	-	52 (25)	-	41 (24)	-	-	9 (0)	2 (1)
- PSP-SC	-	17 (13)	-	-	-	-	13 (8)	4 (5)
- PSP pathology, n (%)	-	31 (94%)	7 (88%) <sup>b</sup>	-	-	-	6 (100%)	18 (95%) <sup>c</sup>
CBS, n (fu visits)	-	135 (113)	44 (54)	60 (51)	-	-	22 (6)	9 (2)
- CBS-CBD <sup>d</sup>	-	12 (13)	8 (12)	-	-	-	1 (1)	3 (0)
- CBS-PSP <sup>d</sup>	-	6 (5)	4 (5)	-	-	-	-	2 (0)
- CBS-AD <sup>e</sup>	-	34 (26)	6 (6)	17 (18)	-	-	10 (2)	1 (0)
- CBS-IDT	-	83 (69)	26 (31)	43 (33)	-	-	11 (3)	3 (2)

Values are Baseline n (n follow-up visits), Pathology n (% PSP).<sup>a</sup> controls included from FTLDNI (143), PROSPECT (36) and UCL DRC FTD (141). <sup>b</sup> one case Globular Glial Tauopathy pathology, <sup>c</sup> one case CBD pathology, <sup>d</sup> autopsy confirmed, <sup>e</sup>AD pathology defined as either autopsy confirmed, positive AD CSF biomarkers (A $\beta$ 1–42, tau, and ptau) or positive Amyloid PET. Abbreviations: 4RTNI1 = 4-repeat tauopathy neuroimaging initiative (Phase 1), 4RTNI2 = 4-repeat tauopathy neuroimaging initiative (Phase 2), DAV = Davunetide trial, , YP = Young plasma trial, SAL = Salsalate trial, PROSPECT = PROgressive Supranuclear Palsy CorTico-Basal Syndrome Multiple System Atrophy Longitudinal Study, UCL = University College London Dementia Research Centre Frontotemporal Dementia cohort.

### 2.2.3 Image processing pipeline

I processed all raw volumetric T1-weighted MR images according to the same pipeline (**Figure 2.1**). In the first quality control (QC) step I visually inspected all raw images to ensure correct acquisition and the absence of artefacts (including head coverage, radiofrequency noise, signal inhomogeneity, susceptibility and motion artifacts). Next, I bias field corrected the raw images that passed QC to correct for magnetic field inhomogeneity, then parcellated the whole brain (cortical and subcortical structures) using the geodesic information flow (GIF) algorithm<sup>304</sup>. This algorithm automatically extracts regions based on the Neuromorphometrics atlas, using an atlas propagation and label fusion strategy. I then extracted subregions of the cerebellum with a version of GIF that uses the Diedrichsen cerebellar atlas<sup>305</sup>: the cerebellar lobules (I-IV, V, VI, VIIa-Crus I, VIIa-Crus II, VIIb, VIIIa, VIIIb, IX and X), the vermis and the deep nuclei (dentate, interposed and fastigial). The volumes of the whole brainstem, medulla, pons, superior cerebellar peduncles (SCP) and midbrain were subsequently segmented using a customised version of the module available in FreeSurfer to accept the GIF parcellation as input for Freesurfer<sup>306</sup>. Total intracranial volume (TIV) was calculated using SPM12 v6225 (Statistical Parametric Mapping, Wellcome Trust Centre for Neuroimaging, London, UK) running under MATLAB R2012b (Math Works, Natick, MA, USA). In the final QC step, I visually inspected all GIF and SPM segmentations to ensure accurate segmentation. QC'd regional volumes selected for inclusion in subsequent disease progression models were controlled for the following covariates: age at scan, sex, scanner type and strength and total intracranial volume (TIV).

I processed a total of 1,446 T1 volumetric MRI scans through this pipeline (1,126 baseline and follow-up scans from cases, and 320 control scans), of which 1,331 passed the QC steps (1,041 case and 290 control scans) and were included in further analyses.

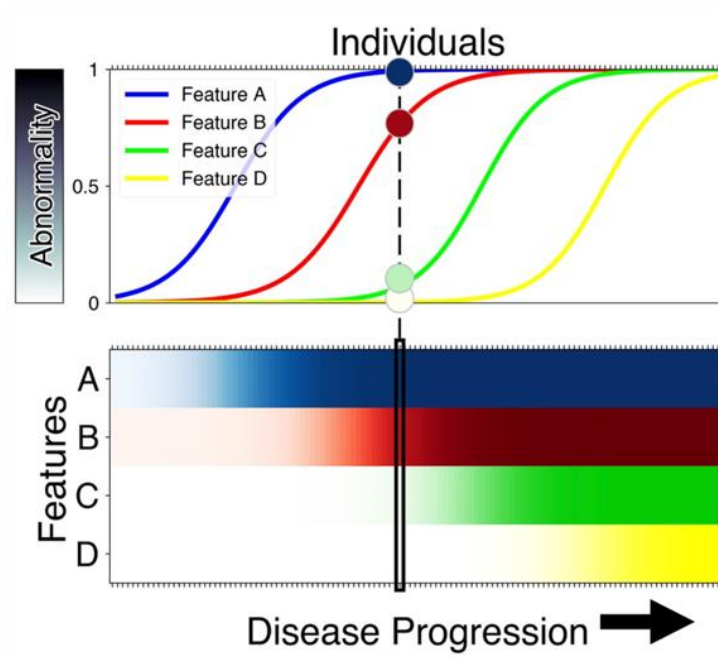


**Figure 2.1 - Imaging processing pipeline.** Cortical/subcortical, cerebellar and brainstem regional volumes were extracted with three different version of GIF from T1 weighted MRI scans. TIV was calculated using SPMv12 software. Covariate adjusted regional volumes were generated by regressing out TIV, age, gender and scanner type/strength and these were used as input data for downstream disease progression modelling using either the EBM or SuStaIn. GIF = geodesic information flow<sup>304</sup>, SPM12 = statistical parametric mapping version 12, GM = grey matter, WM = white matter, CSF – cerebrospinal fluid, TIV = total intracranial volume.

## 2.2.4 Disease progression modelling

### 2.2.4.1 The Event Based Model (EBM)

The EBM (as briefly discussed in **Chapter 1**)<sup>281</sup>, is designed to infer a data-driven, probabilistic sequence in which biomarkers become abnormal from cross-sectional data (**Figure 2.2**). The strengths of the EBM are firstly that it requires no a priori biomarker cut-offs (thresholds) to define abnormality; secondly it requires no a priori staging; and finally it can be completely specified using only moderately sized cross-sectional data. Its ability to generate stages from purely cross-sectional data has shown potential for targeted stratification of patients into clinical trials<sup>278</sup>. Its reliability with moderately sized datasets makes it ideally suited for analysing biomarkers in rare diseases such as the primary tauopathies.



**Figure 2.2 - Schematic of how EBM works.** The EBM is a statistical method for quantifying a sequence of observable abnormalities in a set of disease-relevant biomarkers. It works by assessing at the group level combinations of simultaneously normal and abnormal measurements in multiple biomarkers across individuals at varying stages along the disease course. The top panel shows a hypothetical model of disease progression (left to right on x-axis), with abnormality in biomarkers represented on the y-axis by a probability ranging from 0 (no abnormality) to 1 (maximum abnormality). In this example the biomarkers become abnormal in the sequence A to B to C to D. The bottom panel shows a cross-sectional sample of individuals (lines on the x axis) at different stages in the disease, with corresponding observed combinations of normal (white) and degrees of abnormality (shades of colour) across the four biomarkers. The vertical rectangle highlights an individual near the middle of the disease course with their associated biomarker abnormalities across the four biomarkers shown in the top panel, with higher abnormalities in the early biomarkers (A and B), than in the later biomarkers (C and D). When an individual has a higher abnormality in e.g. biomarker A, than e.g. biomarker B, this provides additional evidence that A changes before B. Reproduced from Oxtoby *et al* 2021<sup>286</sup>.



### 2.2.4.1.1 Model Fitting

The EBM is based on the assumptions of homogenous disease progression and monotonicity: that is all patients have a broadly similar disease progression pattern with a unimodal distribution of orderings, and biomarker change is unidirectional from normal to abnormal i.e. no remission. It also requires that the patient cohort contains a spectrum of disease stages distributed across the whole disease course. A biomarker that changes early in the disease will be abnormal in more individuals, than a biomarker that changes later. An ‘event’ is considered to have occurred when a biomarker (in this study an MRI derived regional volume), has an abnormal value (‘atrophy’) in comparison with the expected values measured in healthy controls. The model then estimates the sequence  $S = S(1), s(2), \dots S(l)$  in which the biomarkers become abnormal where  $S(1)$  is the first biomarker, and  $S(l)$  is the last. Conceptually if biomarker A is usually abnormal when biomarker B is abnormal, but B is often abnormal without A, we infer that B occurs before A in the sequence.

The estimation procedure first fits a two-component univariate mixture model (see Models of event distribution) to control and patient data for each biomarker to generate likelihoods  $P(x_{ij}|E_i)$  (“post-event”) and  $P(x_{ij}|\neg E_i)$  (“pre-event”) of observing the value,  $x_{ij}$ , of biomarker  $i$  for subject  $j$ , given that biomarker  $i$  has or has not become abnormal, respectively. The EBM combines these likelihoods to then calculate the likelihood of the full dataset  $X = x_{ij}: i = 1, \dots, Z; j = 1, \dots, N$  for a given sequence,  $S$ . Assuming independent observations and biomarkers, the likelihood of an ordered sequence  $S$  is:

$$P(\mathbf{X}|\mathbf{S}) = \prod_{j=1}^N \left[ \sum_{k=0}^Z \left( P(k) \prod_{i=1}^k P(x_{ij}|E_i) \prod_{i=k+1}^Z P(x_{ij}|\neg E_i) \right) \right] \quad (1)$$

$j$  iterates over the number of subjects  $N$ , and  $i$  iterates over the number of events  $Z$ .  $P(k)$  refers to the prior likelihood of being at stage  $k$  and in the absence of prior information is treated as uniform to impose as little information as possible on estimated orderings. The estimation procedure then searches for the characteristic ordering,  $\bar{S}$ , which is the sequence that maximises the likelihood of  $P(\mathbf{X}|\mathbf{S})$  in equation

(1)<sup>267</sup>. This is found through a combination of a multiply initialized greedy ascent and Markov Chain Monte Carlo (MCMC) sampling, which samples from the posterior distribution on  $S$ , to find  $\bar{S}$ , which is simply the sequence with the highest (maximum) likelihood. The set of samples from the MCMC sampling also provides information on the uncertainty of the maximum likelihood sequence, which can be visualised on a positional variance diagram<sup>267,281</sup>.

#### 2.2.4.1.2 Models of event distribution

The evaluation of equation (1) requires likelihood models for  $P(x_{ij}|E_i)$  (“post-event”) and  $P(x_{ij}|\neg E_i)$  (“pre-event”); these can be obtained by fitting mixture models to the observed distributions of control and patient regional volumes. A variety of approaches have been used for mixture modelling depending on the specific disease under investigation. In familial AD<sup>281</sup> and Huntington’s disease (HD)<sup>267</sup> with a well-defined control population (no mutation)  $P(x_{ij}|\neg E_i)$  was directly estimated; a Gaussian distribution was first fitted to the control population, followed by fitting a mixture of two Gaussian distributions to the set of volumes from the patients, with one component fixed to the healthy control model’s parameters (statistical mean and standard deviation). The parameters of the second component then provide the model for  $P(x_{ij}|E_i)$ . In sporadic diseases, such as AD, where healthy controls may go on to develop the disease, the control and patient groups are not uniquely defined<sup>280</sup> and so  $P(x_{ij}|\neg E_i)$  cannot be directly estimated. If the biomarkers are normally distributed, one approach is to fit a mixture of two normal distributions using two component Gaussian mixture modelling. The fitting of the model is achieved using a Sequential Least Squares Programming optimizer (SLSQP). To ensure a robust fit, especially if there is significant overlap between the control and case distributions, one can constrain the standard deviations of  $P(x_{ij}|E_i)$  and  $P(x_{ij}|\neg E_i)$  to be less or equal to that of the disease and control group respectively<sup>280</sup>. More recently Firth et al. proposed a non-parametric mixture modelling approach for when biomarker distributions are skewed, that uses kernel density estimate (KDE) distributions in the place of Gaussian distributions<sup>284</sup>. They show that this approach performs at a similar level to the classic EBM (that incorporates Gaussian mixture modelling) with parametric input data, while demonstrating superiority when the data are skewed. It is

particularly useful when biomarker data includes cognitive test scores that are subject to “floor” and “ceiling” effects.

#### 2.2.4.1.3 Patient staging

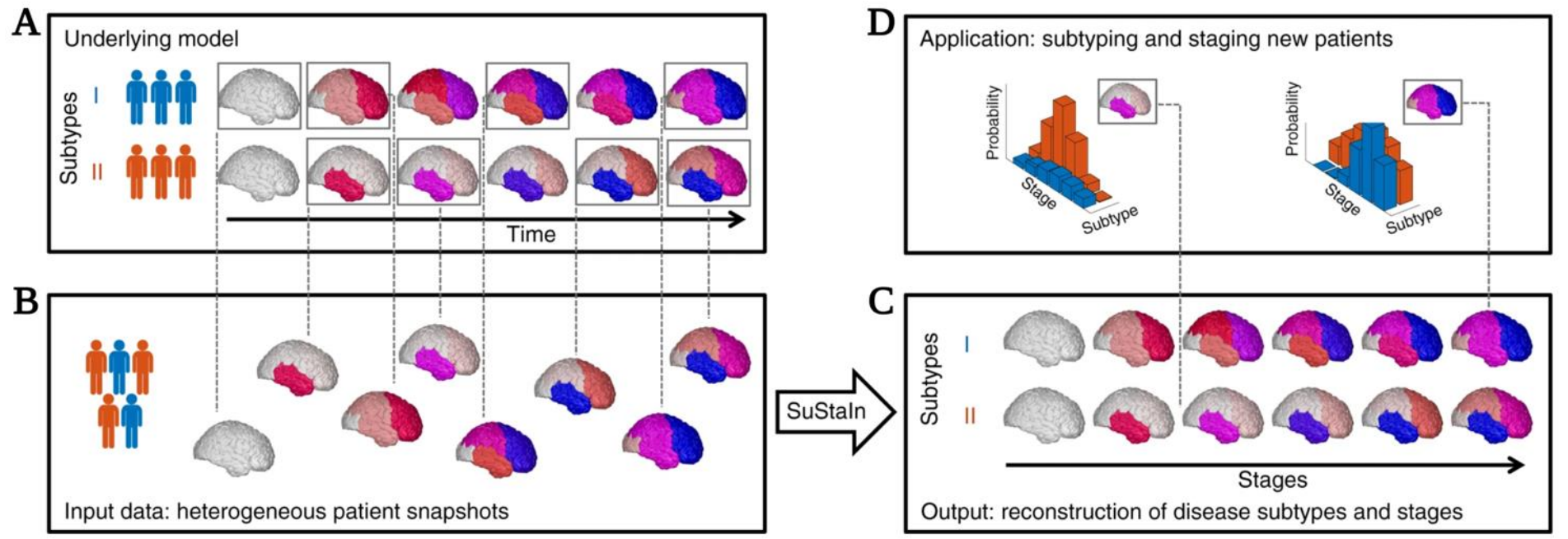
Once the characteristic sequence,  $\bar{S}$ , has been obtained via the EBM, an individual sample  $X_j$  (a vector of all measurements across biomarkers  $i$  for a patient  $j$ ), can be staged by evaluating the stage  $k$  that maximises the likelihood in equation (2) below (Young 2014):

$$\operatorname{argmax}_k P(X_j|\bar{S}, k) = \operatorname{argmax}_k P(k) \prod_{i=1}^k P(x_{ij}|E_i) \prod_{i=k+1}^Z P(x_{ij}|\neg E_i) \quad (2)$$

As before  $P(k)$ , the prior likelihood of being at stage  $k$ , is treated as uniform i.e., no a priori information on a particular stage. The EBM stage, between 1 and the number of biomarkers,  $l$ , of subject  $j$ , is therefore given by the stage  $k$  that maximises equation (2). The assignment of stage  $k$  to a particular patient does not necessarily mean they fit the model exactly, but rather this it is the stage most compatible with their measurements.

#### 2.2.4.2 Subtype and Stage Inference (SuStaIn)

SuStaIn<sup>290</sup> is a generalisation of the EBM concept that augments underlying disease progression modelling with a hierarchical machine learning clustering algorithm. Rather than an EBM, the underlying disease progression model is reformulated as a linear z-score model, with instantaneous normal to abnormal events (as in the EBM) replaced with events that represent the linear transitions between a biomarker's z-scores. **Figure 2.3** demonstrates the concept behind SuStaIn; namely disentangling temporal and phenotypic heterogeneity to identify population subgroups (“subtypes”) with common patterns of disease progression (“stages”). The model fitting iterates through increasing numbers of subtypes  $C$ , estimating the proportion of subjects  $f$  that belong to each subtype, and the order  $S_c$  in which a biomarker reaches each z-score for each subtype  $c = 1 \dots C$ . The optimal number of subtypes  $C$  for the data is determined through ten-fold cross validation. A detailed formulation of the SuStaIn algorithm is provided in Young et al.’s seminal paper published in Nature Comms in 2018<sup>290</sup>. I will summarise the key components of SuStaIn in the following section, based on this paper.



**Figure 2.3 - Conceptual overview of SuStaIn** (A) shows the underlying model consisting of a patient cohort with two hypothetical disease subtypes. (B) Cross-sectional input data from the cohort contains snapshots of biomarker measurements (in this cases imaging data) from each individual who has an unknown subtype and stage. (C) SuStaIn recovers the underlying set of disease subtypes and their temporal progression patterns via simultaneous clustering and z-score disease progression modelling. (D) The SuStaIn model fitted on the cross-sectional baseline data (C), can then be used to estimate the probability that an unseen biomarker measurement belongs to each stage and subtype. The colour of each region represents the amount of pathology in that region ranging from white (no pathology) to red to magenta to blue (maximum pathology). *Figure reproduced with permission from Young et al<sup>290</sup>.*

The following four sections are taken with permission from Young et al. 2018<sup>290</sup>.

#### 2.2.4.2.1 Mathematical model

The linear z-score model underlying SuStaIn is a continuous generalisation of the original EBM<sup>281</sup>, the mathematical formulation of which is given in equation (1) in the previous section. The linear z-score model consists of a set of  $N$  z-score events  $E_{iz}$ , corresponding to a linear increase of biomarker  $i = 1 \dots Z$  to a z-score  $Z_{ir} = Z_{i1} \dots Z_{iR_i}$ ; each biomarker is associated with its own set of z-scores resulting in  $N = \sum R_i$ . In addition, each biomarker has an associated maximum z-score,  $Z_{max}$ , which it reaches by the end of stage  $N$ . Young et al. define a continuous time axis,  $t$ , with arbitrary scaling going from  $t = 0$  to  $t = 1$ . The disease stage,  $k$ , goes from  $t = \frac{k}{N+1}$  to  $t = \frac{k+1}{N+1}$ , with a z-score event  $E_{iz}$  occurring at each stage. As time  $t$  progresses the biomarkers evolve according to a piecewise linear function  $g_{i(t)}$ :

$$g(t) = \begin{cases} \frac{z_1}{t_{E_{z_1}}} t, 0 < t \leq t_{E_{z_1}} \\ z_1 + \frac{z_2 - z_1}{t_{E_{z_2}} - t_{E_{z_1}}} (t - t_{E_{z_1}}), t_{E_{z_1}} < t \leq t_{E_{z_2}} \\ \vdots \\ z_{R-1} + \frac{z_R - z_{R-1}}{t_{E_{z_R}} - t_{E_{z_{R-1}}}} (t - t_{E_{z_{R-1}}}), t_{E_{z_{R-1}}} < t \leq t_{E_{z_R}} \\ z_R + \frac{z_{max} - z_R}{1 - t_{E_{z_R}}} (t - t_{E_{z_R}}), t_{E_{z_R}} < t \leq 1 \end{cases} \quad (3)$$

Therefore, the times  $t_{E_{iz}}$  are determined by the position of the z-score event  $E_{iz}$  in the sequence  $S$ , so if event  $E_{iz}$  occurs in the position  $k$  in the sequence, then  $t_{E_{iz}} = \frac{k+1}{N+1}$ .

The model likelihood for the linear z-score model is formulated by replacing equation (1) in the previous section with:

$$P(\mathbf{X}|\mathbf{S}) = \prod_{j=1}^J \left[ \sum_{k=0}^N \left( \int_{t=\frac{k}{N+1}}^{t=\frac{k+1}{N+1}} \left( P(t) \prod_{i=1}^I P(x_{ij}|t) \right) \partial t \right) \right] \quad (4)$$

where,

$$P(x_{ij}|t) = \text{NormPDF}(x_{ij}, g_i(t), \sigma_i)$$

$\text{NormPDF}(x_{ij}, \mu, \sigma_i)$  is the normal probability distribution function, with mean  $\mu$  and standard deviation  $\sigma$ , evaluated at  $x$ . Similar to the original EBM the prior on disease prior time is treated as uniform. The final SuStaIn model is a mixture of linear z-score models, which leads to the final formulation of the overall model:

$$P(\mathbf{X}|\mathbf{M}) = \sum_{c=1}^C f_c P(\mathbf{X}|\mathbf{S}_c) \quad (5)$$

Where  $C$  is the number of clusters (“subtypes”),  $f$  is the proportion of subjects assigned to the particular cluster (“subtype”),  $\mathbf{S}_c$  is the order in which biomarkers reach each z-score for each cluster, and  $\mathbf{M}$  is the overall SuStaIn model. For the next section I will use the term subtype and cluster interchangeably.

#### 2.2.4.2.2 Model Fitting

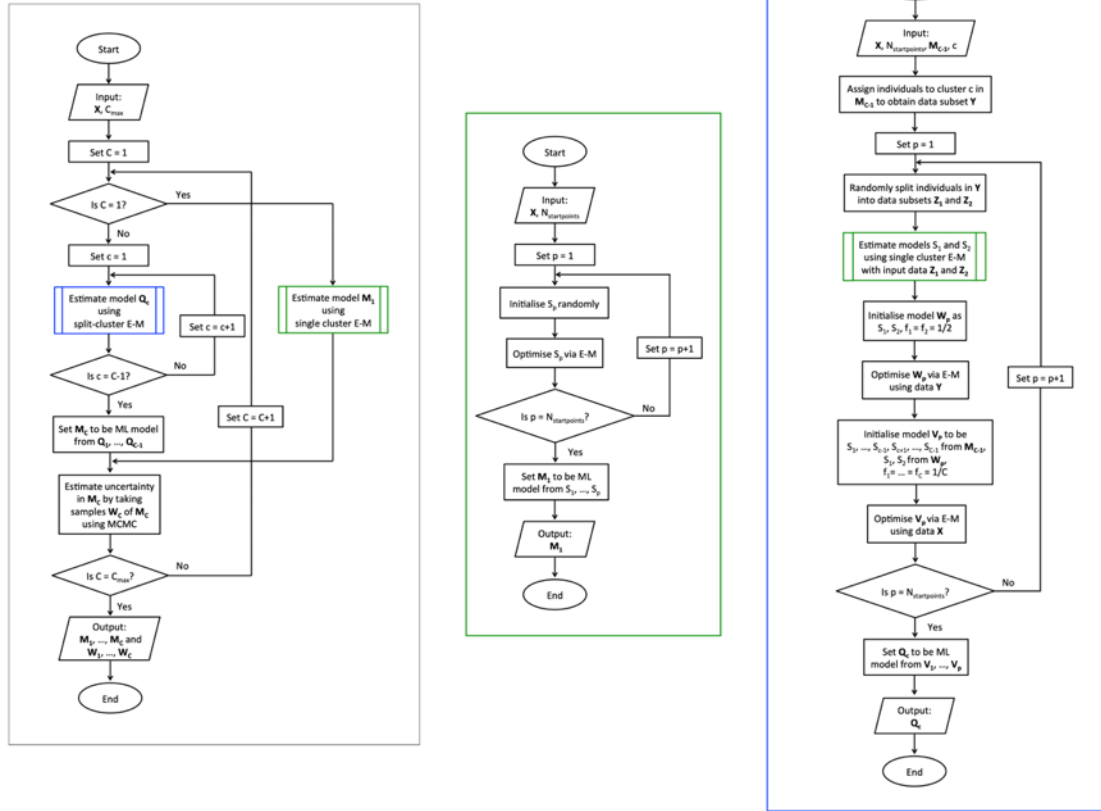
Fitting the SuStaIn model requires simultaneously optimising subtype membership, subtype trajectory and the posterior distributions of both; **Figure 2.4** summarises the main procedures involved in SuStaIn fitting. The model is fitted hierarchically, with the number of clusters estimated by model selection criteria obtained from cross-validation (black box in **Figure 2.4**). The clustering problem is solved sequentially from  $C = 1 \dots C_{max}$ , by initialising the fitting of each  $C$  cluster from the previous  $C-1$  cluster model. For the first cluster ( $C = 1$ ), the single-cluster expectation maximisation (E-M) sub-procedure is used (green box in **Figure 2.4**). Subsequent clusters ( $C > 1$ ) are fit hierarchically by generating  $C-1$  candidate  $C$ -cluster models

using the split-cluster E-M sub-procedure (blue box in **Figure 2.4**); the model with the highest likelihood from these candidate models is then chosen.

The blue box in **Figure 2.4** summarises the split-cluster E-M sub-procedure that is used to generate each of the  $C-I$  candidate  $C$  cluster models. For each of the  $C-I$  clusters the procedure first finds the optimal split of cluster  $c$  into two clusters. To achieve this the data points belonging to cluster  $c$  are randomly split into two separate clusters, and the optimal model parameters for these two clusters are found using the single-cluster E-M procedure (green box in **Figure 2.4**). These parameters are used to initialise the fitting of the two-cluster model to the data subset belonging to cluster  $c$  using E-M. This two-cluster solution is then used together with the other  $C-2$  clusters to initialise the fitting of the  $C$  cluster model. E-M is then used to optimise the  $C$  cluster model, alternating between updating the sequences  $S_c$  and the fractions  $f$  for each cluster. To maximise the chances of finding a global minimum, this procedure is repeated from 25 different start points (i.e. random cluster assignments) to find the maximum likelihood solution.

The single-cluster E-M procedure (green box in **Figure 2.4**) is used find optimal model parameters (that is the sequence  $S$  in which the biomarkers reach each z-score) for a single cluster. In this sub-procedure the sequence  $S$  is initialised randomly and optimised using E-M by going through each z-score event  $E$  in turn to find its optimal position in the sequence relative to the other events. As per the split-cluster E-M sub-procedure, the single-cluster sequence  $S$  is optimised from 25 different random starting sequences to find the maximum likelihood solution.





**Figure 2.4 - Flowchart of procedures for SuStaIn model fitting.** The overall procedure for SuStaIn model fitting is shown in the black box, with the sub-procedures , single cluster E-M and split-cluster E-M, shown in the green and blue boxes respectively.  $M_C$  denotes a SuStaIn model with  $C$  clusters, that contains a set of sequences  $S_1, \dots, S_C$  and subtype fractions  $f_1, \dots, f_C$ . *Figure reproduced from Young et al. 2018*<sup>290</sup>.

#### 2.2.4.2.3 Uncertainty estimation

In addition to estimating the most probable sequence,  $S_C$ , for each subtype, the probability of each possible sequence can be evaluated enabling the relative likelihood of all sequences to be computed. This gives an estimate of the uncertainty in the ordering of  $S_C$ , which can be plotted in a positional variance diagram (similar to the EBM), where different colours indicate the cumulative probability that each region has reached a particular z-score. Given evaluating all the possible sequences involves a search among  $S_n!$  possible sequences which rapidly becomes computationally intractable once  $n$  exceeds  $\sim 10$ , Markov Chain Monte Carlo (MCMC) sampling of the posterior distribution is used to provide an approximation of this uncertainty.

#### 2.2.4.2.4 Cross-validation

$k$ -fold cross-validation<sup>307</sup> is used to both evaluate the optimal number of subtypes, and to evaluate the consistency of the subtype progression patterns. In the original SuStaIn paper<sup>290</sup>, Young et al. evaluate the optimal number of subtypes using the Cross-Validation Information Criterion (CVIC) based on ten-fold cross-validation<sup>308</sup>. The CVIC is defined as the  $CVIC = -2 \times \log(P(\mathbf{X}|\mathbf{M}))$ , where  $P(\mathbf{X}|\mathbf{M})$  is the probability of the data for a particular SuStaIn model,  $\mathbf{M}$ . This evaluates the likelihood of each  $c$ -subtype model from  $c = 1 \dots C$  on the test data for each of the ten folds, subsequently selecting the model with the highest out-of-sample likelihood  $P(\mathbf{X}|\mathbf{M})$ , or equivalently the lowest CVIC across all folds. The CVIC is evaluated for a range of different SuStaIn models (with increasing subtypes), and where the evidence based on the CVIC is not strong, the less complex (less number of subtypes) model is selected. Although the thresholds for this decision on strength of model evidence compared to the N-1 subtype model is subjective, in the original SuStaIn paper they define it as “*a difference of less than 6 between the CVIC and the minimum CVIC across models, or equivalently a difference of less than 3 between the out-of-sample log-likelihood and the minimum out-of-sample log-likelihood across models*”<sup>290</sup>.

To evaluate the consistency of the subtype progression patterns,  $k$ -fold cross validation can also be used. The data is divided into  $k$ -folds and the model re-fitted to each subset of data, with one fold held out for testing each time. The similarity between the progression patterns for the model fitted to each fold and the model fitted to the whole data set can then be quantified using the Bhattacharyya coefficient<sup>309</sup>, which measures the similarity of the distribution of the position of biomarker events in the subtype sequences (0 = maximum dissimilarity, 1 = maximum similarity). The coefficient is evaluated between the position of each biomarker event in the two-subtype progression patterns, and then averaged across the biomarker events and MCMC samples.

## **2.3 3R tauopathy (Pick's disease) genetic studies**

### **2.3.1 Pick's disease International Consortium (PIC)**

Due to the rare and understudied nature of PiD, in collaboration with researchers (Dr Owen Ross, Professor Rosa Rademakers and Professor Dennis Dickson) at Mayo Clinic Brain Bank in Jacksonville, FL, USA (MCJ) I led efforts to establish the world's first international consortium for Pick's disease; the Pick's disease International Consortium (PIC) (**Figure 2.5**). At UCL I was responsible for collecting PiD cases from European and Australasian territories, while the MCJ led the effort to identify and source PiD cases from the North American regions. Inclusion criteria were a neuropathologic diagnosis of PiD with Pick bodies, no evidence of an underlying *MAPT* mutation, and available frozen brain tissue. Exclusion criteria were frontotemporal dementia due to etiology other than a 3R predominant tauopathy or lack of frozen specimens. IRB approval was obtained for the studies at both collection sites (MCJ and UCL) and each individual brain bank had institutional IRB approval for collection and sharing of specimens.

**Figure 2.5** - Global map (A) and table (B) reporting countries and recruitment sites that have contributed Pick's disease tissues to the Pick's disease International Consortium (PIC) to date (01/02/2023). *Samples identified* refers to samples with an archival neuropathological diagnosis of PiD whereas *PiD pathology confirmed* refers to samples confirmed to have a 3R tauopathy consistent with PiD using the PiD neuropathological diagnostic algorithm. Red = countries that have collected and donated Pick's disease tissues.

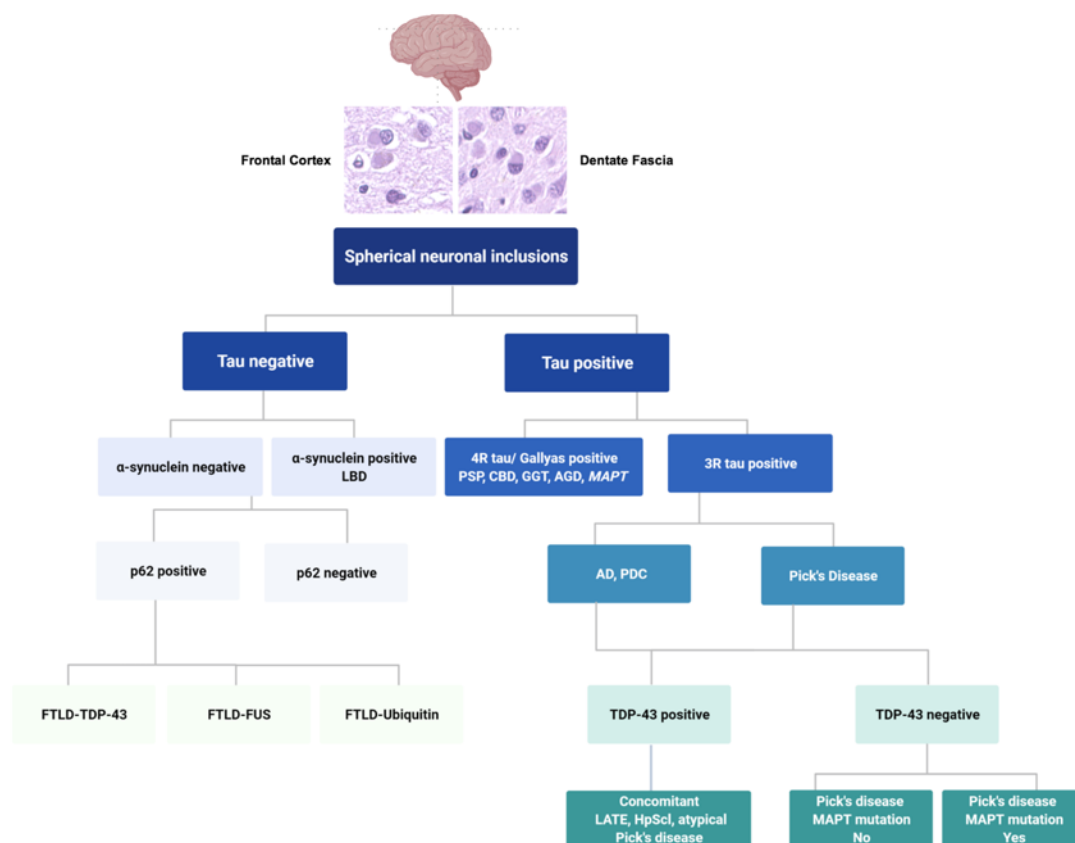
### **2.3.1.1 Neuropathological diagnosis of Pick's disease**

Currently, there are no established diagnostic consensus criteria for the neuropathologic diagnosis of PiD. In many diagnostic centres a neuropathological diagnosis of PiD relies on the presence of argyrophilic, spherical neuronal inclusions using traditional silver staining methods, such as Bielschowsky's or Gallyas-Braak silver staining methods. Both silver staining methods stain Alzheimer's disease (AD) neurofibrillary tangles, yet spherical inclusions in PiD are negative on Gallyas-Braak silver staining<sup>141</sup>. Given that AD and PiD pathology can co-exist in the same patient, the ability to differentiate between the two pathologies using silver staining methods is helpful, especially for centres that do not use immunohistochemistry (IHC) against phosphorylated tau (p-tau) or have isotype specific tau antibodies incorporated into their diagnostic work-up. IHC against epitope-specific tau antibodies further helps to distinguish between AD and PiD features. Despite both spherical inclusions and neurofibrillary tangles staining positive with antibodies against phosphorylated tau (p-tau), antibodies against 3R tau highlight selective 3R tau spherical inclusions in PiD, which is further validated by these inclusions not staining with 4R tau antibodies. This distinction is most pronounced in the granule cell neurons of the hippocampal dentate fascia, which may be used solely to diagnose PiD.

### **2.3.1.2 PIC diagnostic algorithm**

It became clear during the early stages of my PhD that since a harmonized neuropathologic diagnostic scheme did not exist for PiD it was essential to establish a defined set of operational diagnostic criteria that would ensure that submitted PiD cases reflect a 3R-predominant tauopathy. These diagnostic criteria were developed during 2020 through collaboration between myself and Professor Tammarny Lashley at UCL, and Professor Dickson and Dr Shanu Roemer at the MCJ. All potential PiD cases identified by the PIC had as a minimum an archival neuropathologic diagnosis of PiD (i.e. the presence of argyrophilic or p-tau positive spherical inclusions) through neuropathological assessments at their respective brain banks. Each participating centre was requested to submit and report respective 3R and 4R tau staining results for each individual PiD case to the PIC. To fulfil PIC criteria all cases had to have Pick bodies with 3R tau positive and 4R tau negative inclusions on IHC (**Figure 2.6**). The additional presence of ballooned neurons and negative Gallyas staining of inclusions

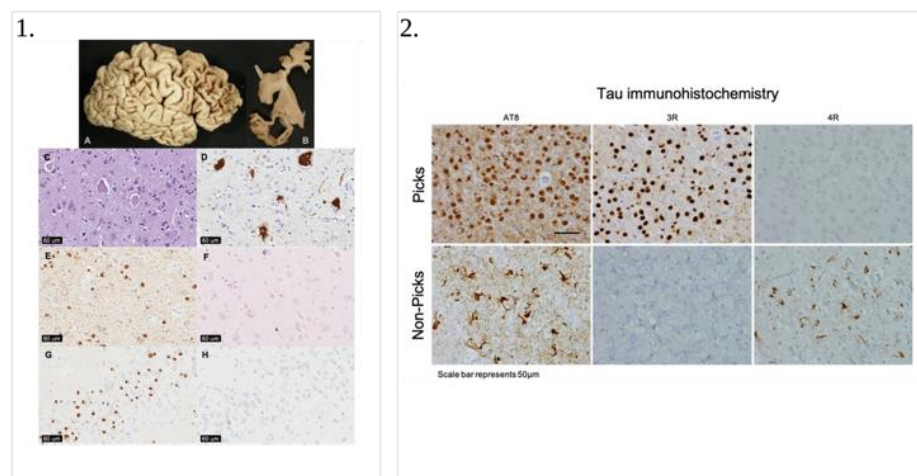
was preferred to confirm diagnosis. If 3R/4R tau IHC had not been performed at their respective brain banks, centres submitted routinely cut sections (up to seven microns)/slides of unstained, formalin fixed paraffin embedded (FFPE) tissue from hippocampal, frontal or temporal lobe regions for 3R and 4R tau IHC assessments. At MCJ all cases were stained using hematoxylin and eosin (H&E), thioflavin-S, a modified Bielschowsky or Gallyas silver staining method. IHC was performed using antibodies against phosphorylated tau (CP13 at MCJ and AT8 at UCL), 3R and 4R tau (see next section) and phosphorylated TDP-43 (Cosmo) (**Figure 2.7**). Select cases were stained for amyloid- $\beta$  ( $\text{a}\beta$ ),  $\alpha$ -synuclein (NACP), sequestome (p62), fused in sarcoma (FUS), and ubiquitin. Cases submitted to UCL were examined by myself and Professor Tammarny Lashley at Queen Square Brain Bank, and those submitted to the Mayo Clinic Brain Bank by Professor Dennis Dickson and Dr. Shanu Roemer, all using the PIC diagnostic algorithm.



**Figure 2.6 - PIC pathological diagnostic criteria for PiD.** These criteria were developed in collaboration with Professor Dennis Dickson and Dr Shanu Roemer at the Mayo Clinic Jacksonville. Abbreviations: LBD = Lewy body dementia, PSP = progressive supranuclear palsy, CBD = corticobasal degeneration, GGT = globular glial tauopathy, AGD = argyrophilic grain disease, AD = Alzheimer's disease, PDC = Parkinsonism Dementia Complex, LATE = limbic predominant age-related TDP-43 pathology, HpScl = hippocampal sclerosis.

### 2.3.1.3 Neuropathologic methods for 3R/4R tau IHC

Sections were deparaffinized in xylene and rehydrated using graded alcohols. IHC for all antibodies required pressure cooker pre-treatment for 10 minutes in citrate buffer pH 6.0. Endogenous peroxidase activity was blocked in 0.3% H<sub>2</sub>O<sub>2</sub> in methanol for 10 minutes and non-specific binding blocked with 10% dried milk solution. Tissue sections were first incubated with primary antibodies; either AT8 (p-tau, 1:600; Thermo); RD3 (3R p-tau, 1:800; Millipore) or RD4 (4R p-tau, 1:4000, Millipore). AT8 and RD4 were added to sections and incubated at room temperature for 1 hour, while RD3 was added to sections and incubated for 48 hours at 4°C. Of note, before applying the RD3 antibody, the tissue was treated with potassium permanganate 0.25% for 15 minutes, followed by oxalic acid for 3 minutes. Following this the tissue sections were treated with the secondary antibody; biotinylated anti-goat IgG Antibody (1:200; VectorLabs) for 30 minutes followed by Avidin-Biotin complex (30 minutes; Dako)<sup>310</sup>. Colour was developed with di-aminobenzidine/H<sub>2</sub>O<sub>2</sub>.



**Figure 2.7 - Pathological assessments of Pick's disease brains (1)** At the MCJ (A) The superior and dorsolateral surfaces of the frontal cortex and temporal lobe often show severe circumscribed 'knife-edge' edge atrophy. (B) Coronal sections of the brain show markedly dilated ventricles, cortical atrophy, and hippocampal affection. (C) Enlarged, amorphous ballooned neurons. (D) In regions with severe astrogliosis and neuronal loss, staining against αβ-crystallin may highlight ballooned neurons. (E) Phosphorylated tau antibodies highlight spherical cytoplasmic neuronal inclusions and may also show marked neuropil staining, especially in cases with concomitant Alzheimer's type pathology. (F) Gallyas silver stains may stain isolated glial lesions or neurofibrillary tangles; however, Pick bodies do not show any significant degree of silver staining. (G) 3R tau staining of the dentate fascia of the hippocampus show strong immunoreactivity of spherical inclusions. (H) 4R tau staining of the dentate fascia show negative spherical inclusion, although isolated neurofibrillary tangles may stain positive. Images are from Pick cases submitted to MCJ. **(2)** Tau IHC performed at UCL (same procedure followed at the MCJ). The top row shows a Pick's disease case that was positive for AT8 and 3R-tau immunoreactive Pick bodies. The bottom row shows a non-Pick's disease case (that was originally pathologically diagnosed with Pick's disease) that was positive for AT8 and 4R-tau but negative for 3R tau immunoreactive Pick bodies. Images are from Pick cases submitted to UCL.

#### 2.3.1.4 Total cases collected by the PIC

Frozen brain tissue from cerebellum or prefrontal cortex were obtained from each case. Baseline demographic information was collected for all subjects (age at symptom onset (AAO) and age at death (AAD) for PiD patients, age at blood collection for controls, and sex). In addition to basic demographic information, the PIC also collected information related to family histories, clinical outcomes (e.g. behavioural and language impairments, presence/absence of parkinsonism, upper and lower motor neurone deficits, Mini-Mental State Examination and Clinical Dementia Rating), and pathological information (e.g. Thal phase, Braak stage, and brain weight,) for each individual case, as well as noting whether other tissues and brain imaging data were available.

As of the 1<sup>st</sup> February 2023 the total number of cases identified with an archival neuropathologic diagnosis of PiD (i.e. the presence of argyrophilic or p-tau positive spherical inclusions) was 449 (178 in the UCL cohort and 271 in the MCJ cohort). Of these 365 (156 at UCL and 209 at the MCJ) cases met the PIC neuropathological diagnostic criteria for PiD (**Figure 2.6**). The breakdown of samples by recruitment site is detailed in **Table 2.2**. Not all of these samples were available at the time of analysis for the *MAPT* haplotype analysis (**Chapter 6**) or the case-control (**Chapter 7**) and survival (Chapter 8) GWASs. I will detail the samples, broken down by contributing site, used in each of these studies in the respective chapters.

#### 2.3.1.5 Controls

For the *MAPT* haplotype analysis in **Chapter 6** 1,312 controls were collected from the Mayo clinic, FL (N=881) or Rochester, MN (N=431). Control subjects were deemed neurologically healthy by neurologists at Mayo Clinic.

For GWAS analysis controls were collected from two sources: 989 from the Global Parkinson's Genetics Program (GP2)<sup>311</sup>, and 457 from the Invasive Fungal Infection and GENetics (IFIGEN) cohort<sup>312</sup>.



**Table 2.2 - Overview of PIC contributing sites.**

		Number of samples*	
Contributing Site	Country	Samples identified	PiD Pathology confirmed
UCL Institute of Neurology cohort (Europe / Australia)		178	156
Netherlands Brainbank, Amsterdam	Netherlands	36	33
UCL Queen Square Brain Bank, London	UK	18	18
Neurological Tissue Bank, Biobanc-Hospital Clínic-IDIBAPS, Barcelona	Spain	14	14
Manchester Brain Bank, Manchester	UK	11	11
Cambridge Brain Bank, Cambridge	UK	25	11
London Neurodegenerative Diseases Brain Bank, London	UK	8	8
South West Dementia Brain Bank, University of Bristol	UK	12	8
Sydney Brain Bank, Sydney	Australia	8	8
Neurobiobank München, Munich	Germany	7	7
Oxford Brain Bank, Oxford	UK	7	7
Newcastle Brain Tissue Resource, Newcastle-upon-Tyne	UK	7	7
IBB NeuroBioBank, University of Antwerp, Antwerp	Belgium	5	5
Neuro-CEB Biobank France, Paris	France	5	5
Victorian Brain Bank, Florey Institute, Parkville, VIC	Australia	5	5
Douglas-Bell Canada Brain Bank, Montreal, QC	Canada	4	3
The Brain Bank at Karolinska Institutet, Stockholm	Sweden	3	3
Fondazione IRCCS Istituto Neurologico Carlo, Milan	Italy	2	2
DZNE e.V. Standort Tübingen	Germany	1	1
Mayo Clinic, Jacksonville cohort (North America)		271	209
Mayo Clinic, Jacksonville, FL/Rochester, MN	USA	63	59
University of California, San Francisco, CA	USA	32	21
University of Pennsylvania, Philadelphia, PA	USA	18	18
Massachusetts General Hospital, Boston, MA	USA	22	15
Northwestern University, Chicago, IL	USA	15	14
John Hopkins University, Baltimore, MD	USA	13	13
Indiana University, Bloomington, IN	USA	12	12
Banner Sun Health Research, Sun City, AZ	USA	10	9
Sunnybrook Health Research, Toronto, ON	Canada	11	9
Columbia University, New York City, NY	USA	9	9
Emory University, Atlanta, GA	USA	8	7
UT Southwestern Medical Center, Dallas, TX	USA	10	7
University of British Columbia Hospital, Vancouver, BC	Canada	5	5
Houston Methodist Hospital, Houston, TX	USA	4	4
Bryan Brain Bank and Biorepository, Duke University, Durham, NC	USA	9	3
Krembil Research Institute, University of Toronto, Toronto, ON	Canada	2	2
UCLA - Sepulveda, Los Angeles, CA	USA	7	1
Parkwood Institute, University of Toronto, Toronto, ON	Canada	1	1
Oregon Health & Science University, Portland, OR	USE	20	Not stained
Total PIC samples		449	365

\* Total samples identified and pathologically confirmed as of 1<sup>st</sup> February 2023

## 2.3.2 Genetic Analysis

### 2.3.2.1 DNA preparation

For the majority of cases, DNA was extracted at their respective collection site (MCJ or UCL). At UCL genomic DNA was extracted from either frozen brain tissue or whole blood lymphocytes using the Kleargene XL Nucleic Acid Purification kit (LGC, Germany). At MCJ genomic DNA was extracted from frozen brain tissue from PiD cases and from peripheral blood lymphocytes from control subjects using an automated or manual method. Automated DNA extractions were carried out using Autogen Tissue Kit reagents according to manufacturer protocols and were processed on the Autogen FlexSTAR+ (both Autogen, Holliston, MA, USA). Manual extractions were completed using QIAamp DNA Mini Kits (Qiagen, MD, USA). DNA quality for all samples was assessed with a NanoDrop 8000 spectrophotometer (ThermoFisher Scientific, USA) and absorbance ratios for 260/280 and 260/230 were between 1.7-2.2 and 2.0-2.2, respectively. Sydney Brain Bank extracted DNA from peripheral blood using a Qiagen DNA extraction kit, UPenn and UCSF extracted DNA from their cases using QIAamp DNA mini kits.

### 2.3.2.2 Genotyping

For the *MAPT* haplotype analysis (**Chapter 6**) we performed targeted genotyping of the six *MAPT* haplotype tagging variants in 2020 / early 2021 using PCR. For the GWAS analysis (**Chapter 7** and **Chapter 8**) we performed genome-wide genotyping using microarrays from mid 2021 to mid 2022. I was responsible for the genotyping of all UCL samples, and Dr Rebecca Valentino in Owen Ross's laboratory at the MCJ was responsible for genotyping of all the MCJ samples. I will therefore discuss each of the genotyping methods in detail below.

#### 2.3.2.2.1 Targeted genotyping

The *MAPT* H2 haplotype-tagging variant rs8070723, in addition to the five common *MAPT* variants (rs1467967, rs242557 [the H1C haplotype-tagging variant], rs3785883, rs2471738, and rs7521) were genotyped. These six variants together define the H1-subhaplotypes allowing analysis of *MAPT* subhaplotype structure<sup>176,202</sup>. European and Australasian cases in the UCL cohort were genotyped using KASPTM

SNP genotyping assays on the Hydrocycler2 system (LGC Genomics, Hoddesdon, Herts, UK) according to manufacturer instructions, and were read on a PHERAStar FSX plate reader (BMG Labtech, Cary, NC, USA). Genotypes were called using Kraken KlusterKaller<sup>TM</sup> software (LGC Genomics, Hoddesdon, Herts, UK). North American cases and all controls were genotyped using TaqMan SNP genotyping assays on an ABI 7900HT Fast Real-Time PCR system (Applied Bio-systems, Foster City, CA, USA). *MAPT* variants were genotyped according to manufacturer instructions (primer sequences available upon request). Genotypes were called using TaqMan Genotyper Software v1.3 (Applied Bio-systems, Foster City, CA, USA).

#### **2.3.2.2.2 Whole-genome microarray genotyping**

All samples from MCJ (North American samples), Sydney (Australasia samples) and IFIGEN controls were genotyped by the local teams on the Illumina (*Illumina, San Diego, CA, USA*) Global Screening Array version 3 (GSA) (<https://www.illumina.com/products/by-type/microarray-kits/infinium-global-screening.html>), which covers c.654,000 variants. All UCL samples (European samples) were genotyped at UCL Genomics on the Illumina NeuroBooster Array (NBA); I carried out the DNA sample quality control and preparation for genotyping including the NanoDrop and Qubit fluorometry and plated all samples which were then delivered to UCL Genomics for genotyping. The NBA consists of an Illumina Global Diversity Array v1 (<https://www.illumina.com/products/by-type/microarray-kits/infinium-global-diversity.html>) backbone (c.1.8 million variants), with bespoke SNP coverage of known variants associated with neurodegenerative disease (95,000 variants) ([https://github.com/GP2code/Neuro\\_Booster\\_Array](https://github.com/GP2code/Neuro_Booster_Array)). This array has been designed with three aims in mind: 1) to improve imputation quality across populations of diverse continental ancestries to facilitate risk loci discovery; 2) to improve imputation of known GWAS loci for neurodegenerative disease; 3) to identify coding and low frequency variants. GP2 controls were also genotyped on the Illumina NeuroBooster Array (NBA) by the local teams within the GP2 network.

#### **2.3.2.3 Genome-wide genotype calling**

Raw IDAT files were shared with me at UCL, and for each of the genotyping platforms (NBA and GSA) I performed genotype calling separately using GenomeStudio version 2.0 (Illumina), based on the protocol published by Guo et al.<sup>313</sup> after training by Dr

Kin Mok. I manually re-clustered common variants ( $MAF > 1\%$ ) on the autosomal chromosomes with a GenTrain score between 0.4 – 0.7 (inclusive). I did not manually re-cluster variants on the sex chromosomes or rare variants ( $MAF < 1\%$ ), as I later excluded these in PLINK. All variants tagging *MAPT* mutations on the NBA array were checked individually and re-clustered if necessary. For the NBA I used the in-house NIH cluster file (based on 2000 GP2 samples), and for the GSA I used the Illumina cluster file provided online (<https://support.illumina.com/downloads/infinium-global-screening-array-v3-0-product-files.html>). Raw data from different genotyping batches were combined to improve accuracy of clustering of the intensity data. All UCL cases and GP2 controls were screened for known *MAPT* mutations covered by the NBA and were excluded from downstream analysis if a *MAPT* mutation was identified. All MCJ cases were screened for known *MAPT* mutations before genotyping and only included if negative.

#### **2.3.2.4 Genome-wide association studies**

In this section I will first give an overview of genome wide association studies (GWAS) with the key steps required for a successful study, and then provide more detail on some of the key statistical methods used in this thesis. For details on the application of these techniques in the PiD case-control and survival GWASs, and the methods specific to each study, please refer to **Chapter 7** and **Chapter 8** respectively.

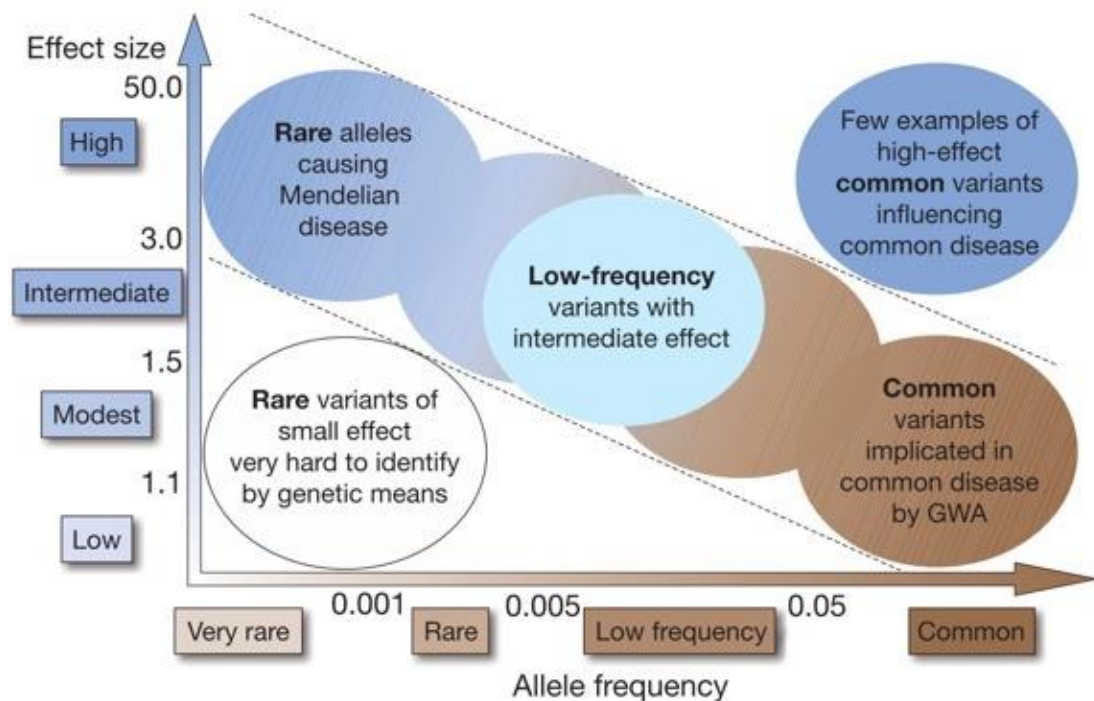
##### **2.3.2.4.1 GWAS - an overview**

GWASs are now a well-established technique to identify genetic variability underlying the risk for disease and have been facilitated by advanced in single nucleotide polymorphism (SNP) arrays that capture a significant proportion of common variation in the genome<sup>314</sup>. The concept was first proposed by Neil Risch and Kathleen Merikangas as technique to identify common disease-associated variants in complex disorders<sup>315</sup>. In recent years several GWASs have been successfully carried out in neurodegenerative disease, leading to the identification of many susceptibility loci associated with disease (see **Chapter 1**).

The International HapMap consortium (HapMap) founded in 2002, was crucial to the development of GWA studies. By determining common patterns of DNA sequence variation in the human genome across populations of differing ancestry and developing

of technology to make SNP genotyping more reliable, faster and cheaper, the possibility of GWA SNP testing became a reality<sup>316</sup>. The HapMap project produced a minimal set of informative SNPs to tag genetic variation across the human genome<sup>317</sup>.

The underlying rationale for GWA studies is the ‘*common disease, common variant*’ (CDCV) hypothesis that hypothesises that allelic variants present in 1-5% of the population account for, at least in part, the risk associated with common disorders<sup>318,319</sup> (**Figure 2.8**). A number of consequences follow necessarily from this hypothesis. Firstly, if common genetic variants influence probability of disease, then the penetrance (or effect size) of an individual variant must be small compared to rare variants (mutations) causing rare disorders. This means that population prevalence of a disease and allele frequency are correlated, and common variants by definition will not have a high penetrance<sup>320</sup>. Following from this, if common alleles have low penetrance, but common disorders show heritability, then many common variants are likely to influence disease susceptibility. The frequency of an allele in the population combined with the individual effect size for that allele are important factors to consider when planning a GWAS.



**Figure 2.8 - Identifying genetic variants by risk allele frequency and genetic effect size.** Variants identified by GWAS tend to lie in the middle to bottom right of the graph. *Image from Manolio et al. 2009*<sup>314</sup>.

In contrast to the CDCV hypothesis, the multiple rare variant (MRV) hypothesis posits that the genetic risk component of common complex diseases is due to MRVs)<sup>321</sup>. In neurodegenerative disease it is likely that both the CDCV and the MRV hypothesis can be true at the same locus, with both common and rare risk alleles driving genetic risk at the same loci<sup>322</sup>.

Unlike traditional linkage studies which analyse family members, a GWAS is conducted in unrelated individuals. GWASs usually focus on common variants, also known as single nucleotide polymorphisms (SNPs), with a minor allele frequency of greater than 1%. The main steps in performing a GWAS are 1) Study design 2) Data pre-processing and QC 3) Association analysis and 4) Post-analytic visualisation. For a more detailed review, I refer the reader Bush et al.<sup>320</sup> for an overview of GWASs and to Reed et al.<sup>323</sup> for a practical guide to performing the analysis.

### **1) Study Design**

The phenotype of interest in a GWA study can be either categorical (usually case vs control), or quantitative. Quantitative traits, from a statistical perspective, are often preferred as they increase the power to detect a genetic effect. In the tauopathies, early studies have focused on a case/control categorical outcome variable<sup>167,169,194</sup> as it is rare to have endophenotypes that have well established quantitative measures. This has been changing in recent years with studies analysing genetic determinants of clinical progression and survival<sup>198,324,325</sup>. For case/control studies a major component of success is the definition of rigorous phenotype criteria: where cases are defined according to pathology (as in the case for this study), standardised rule-based pathology criteria are essential to ensure homogeneity of those included in the cohort.

### **2) Data pre-processing and QC**

Once genotypes from the array have been called the next step is both sample level and variant level quality control. Sample level QC involves identifying call rate and heterozygosity outliers (high or low heterozygosity may indicate poor sample quality), and biologic sex mismatches. Optional steps at this stage can then include filtering based on relatedness and removing ancestry outliers. Variant level QC steps include assessing variant missingness (high missingness suggests failure of that particular genotype probe), missingness by case vs. control as well as by haplotype to ensure that these are not unbalanced between groups (can lead to false positive associations).

Additional steps can include filtering based on minor allele frequency (MAF) prior to imputation (as rare variants can be difficult to call reliably with imputation), and filtering of controls based on violation of Hardy Weinberg Equilibrium (HWE) assumptions (can be an indication of inbreeding, population substructure or genotyping error). Filtering is usually limited to controls in the case of HWE, because in cases its violation may be due to true disease association<sup>326</sup>.

Population stratification can occur when there is an admixture of two or more different ancestral populations in a study, where each subpopulation differs both in terms of phenotype prevalence and allele frequencies. This phenomenon is well documented to cause spurious associations between the phenotype of interest and non-causal SNPs<sup>326</sup> and needs to be controlled for in any GWA analysis. A common, fast and effective way to adjust for population substructure is principal component analysis (PCA) (see Principal component analysis)<sup>327</sup> which generates principle component values that can be thought of as an “ethnicity” score. These can then be used as covariates in the subsequent association analyses to adjust for ancestry effects (population stratification) in the data.

#### **4) Imputation**

Modern chip arrays capture upwards of one million SNPs which vary in at least 1% of the general population, though this is only a fraction of the total number of SNPs that are present in the average human genome (more than 84 million)<sup>328</sup>. The association signal may therefore be missed if the causal variant has not been genotyped. Genotype imputation is a technique that allows missing SNPs to be inferred based on the observed genotypes at neighbouring SNPs, greatly increasing the number that can be tested for association, so increasing the power of a given study. Taking advantage of the linkage disequilibrium (LD) patterns and haplotype frequencies from whole genome sequenced (WGS) reference panels (such as from HapMap, 1000 genomes or the Haplotype Reference Consortium [HRC]), genotypes not directly genotyped on an array can be probabilistically estimated<sup>329</sup>.

The first step in imputation is to pre-phase the genotypes (haplotype estimation) using available genotypes. The second step then involves the actual imputation of the missing genotypes. One needs to check study data with the imputation reference SNP list to ensure that there is overlap before imputing. Popular algorithms for imputation

include IMPUTE2, MaCH/minimac, BEAGLE and fastPHASE<sup>330</sup>. Web-based imputation servers such as the Michigan Imputation Server<sup>331</sup> or the Sanger imputation server<sup>332</sup> can be used if one wants to use the HRC reference panel.

Once imputation has been performed, it is important to assess imputation quality, filtering out imputed genotypes with high degrees of uncertainty. The most common measure used for this filtering is  $R^2$ , which is the value associated with the linear model regressing each imputed SNP on regional typed SNPs. Generally, these genotype calls are made by specifying a genotype probability threshold of  $> 0.8$  (a so called “hard call”). Although these estimated genotypes can be used as true genotypes in the subsequent analyses this can lead to false association results, particularly for rare variants that are often purely imputed and can be called common allele homozygotes. Best practice now is to use the genotype probabilities to account for the uncertainty in genotype state in the association analysis.

## **5) Association analysis**

Quantitative phenotypes (traits) are usually analysed using generalised linear models (GLM): either Analysis of Variance (ANOVA) or multiple linear regression. ANOVA tests the null hypothesis that there is no difference between the trait means of any of the genotype groups, while linear regression generates regresses each SNP separately on a given phenotype, with covariate adjustment for patient-level clinical and demographic factors, as well as the principal components to account for population stratification. Both methods assume that the trait is normally distributed, with common variance (groups of homoscedastic), and that the groups are independent. If the assumption of normality is not met, then a transformation (such as a log transformation) of the original trait values may be required<sup>326</sup>.

For case / control studies where the outcome is a binary categorical variable, logistic regression can be used. In logistic regression the outcome is transformed using a logistic function (logit) that predicts the probability of being in the case group for a given genotype<sup>320</sup>. Both linear and logistic regression allow adjustment for covariates that are known to influence the trait, and so reduce false positive associations. Perhaps the most important covariate to adjust is population stratification (as discussed previously), which can be done at this stage by inserting the principal components generated from a PCA.



For both quantitative and categorical trait analysis there are a number of ways that the genotype data can be encoded for the association tests. Allelic association tests assess the association between one allele of the SNP and the trait, whilst genotypic tests assess the association between genotypes and the trait. For genotypic tests the genotypes for an individual SNP can be grouped into classes (models) such as multiplicative, additive, dominant or recessive<sup>333</sup> each of which make different assumptions about the genetic effect. The underlying genetic model of association for each SNP used in the analysis will impact of the results. For GWAS, where there are a large number of SNPs with generally uncharacterised relationships to the outcome, it is common practice to assume that the disease risk from individual SNPs is roughly additive i.e. that heterozygote risk is intermediate between the two homozygote risks (the minor allele dosage is coded as 0, 1, or 2)<sup>326</sup>. Although counting alleles rather than genotypes in the additive model will increase power, it is not usually recommended as it requires the assumption of HWE in cases and controls<sup>334</sup>.

Given in a GWAS millions of statistical tests are performed (one for each SNP), and each test has its own false positive probability, the cumulative probability of finding a false positive with modern genotyping arrays is therefore much higher. The most common approach to correct for this multiple testing is the use of the Bonferroni correction, which adjusts the significance level ( $\alpha$ ) from 0.05 to  $0.05/k$ , where  $k$  is the total number of statistical tests conducted. For a GWAS using 1 million SNPs this would set the significance level at  $5 \times 10^{-8}$  (the so called genome wide significant threshold), a conservative estimate as the assumption is that each association test is independent. Statistical power in a GWAS is affected by the allele frequency, sample size, loci effect size, and the standard deviation (SD) of the trait.

## **6) Post-analytic visualisation**

After SNPs with significant association are identified, it is common practice to report the chromosome and base pair location based on the relevant genome build that has been used. Specifically, the following are reported; SNP name, chromosome number, base pair location (with details of specific genome build), the  $\beta$  coefficient estimate (or odds ratio) from the regression model, the corresponding standard error and the associated  $p$  value<sup>323</sup>.

There are several plots commonly used to both visualise GWAS results and perform quality control. A Q-Q plot is a probability plot that is used to visualise the extent to which the observed distribution of the SNP-level test statistics follows the expected null distribution. The expected  $-\log(p)$  (a normal distribution) is drawn as a line, and the lambda ( $\lambda$ ) statistic (otherwise known as the genomic inflation factor), gives a formal measure of deviation from the null line (indication of bias within the data).  $\lambda$  scales with sample size, so it is important to report the  $\lambda_{1000}$  (the inflation factor of a study of 1000 cases and 1000 controls). A value of  $<1$  suggests the study is underpowered, a value of close to 1 suggests there is little bias in the data observed, while a value of  $>1.5$  is usually the threshold used to indicate there is systemic bias, such as population stratification, that has not been controlled for. A Manhattan plot is a common way to visualise the  $p$  values of all SNPs in the GWAS by chromosome location. Each point corresponds to an individual SNP where the x-axis value represents the gene coordinate, with the y-axis giving the  $-\log(p)$  value (the larger the value the smaller the  $p$  value). A solid line is then overlain on the plot to represent the Bonferroni corrected significance threshold. Visual inspection of the plot allows for identification of SNPs with small  $p$  values, where the surrounding SNPs have low values, which can indicate a false positive result. If PCA has been performed to control for population stratification, a Scree plot can be a useful method to visualise how much of the variance in the data is accounted for by each principal component. Usually, the majority of the variance should be captured by the first few principal components; if you do not see this it indicates an issue with either the data itself, or up-stream processing of the data.

## 2.3.2.4.2 Statistical Methods

### 2.3.2.4.2.1 Principal component analysis

PCA is a method of data reduction that reduces the complexity in high-dimensional data while retaining trends and patterns. Similar to clustering<sup>335</sup>, it is an unsupervised learning technique, that is able to find patterns without prior knowledge about the groups that the samples come from. PCA reduces data by geometrically projecting them onto lower dimensions called principal components (PCs), with the aim of trying to encode the maximum variance of the underlying data with the minimal number of PCs<sup>336</sup>. A principal component consists of an eigenvector representing directionality, with a corresponding eigenvalue representing the variance in the data in that direction. The first PC contains the most variance, with each subsequent PC geometrically orthogonal (linearly independent and uncorrelated) and with less variance. The PC selection process involved maximising the correlation ( $r^2$ ) between the data and their lower dimensional projection, and is mathematically equivalent to performing linear regression<sup>337,338</sup>. The total number of eigenvectors and eigenvalues (PCs) represents the number of dimensions (input variables) in the underlying data.

### 2.3.2.4.2.2 Logistic (and linear) regression

Logistic regression is a mathematical modelling approach that can be used to describe the relationship between several predictors and a binary or categorical dependent variable. In addition to predicting the value of a categorical variable (for example patient survival), this method can also predict the associated probability.

To understand how logistic regression works it is first necessary to understand linear regression, where the dependent variable is continuous. For linear regression a line of best fit for the data is derived using least squares, a method that minimises the sum of squared distances from each data point to the line (also called residuals). The method is formalised in the following equation:

$$Y = B_0 + B_1 \cdot X$$

Where the beta coefficient of the slope ( $B_1$ ) indicates the increase in the dependent variable ( $Y$ ) associated with each unit increase in the independent variable ( $X$ ). The intercept ( $B_0$ ) indicates the mean of the dependent variable when the value of the independent variable is zero.

Linear regression does not however work well for when the independent variable is a binary/categorical. One approach to predict a categorical variable is logistic regression, which takes its name from the type of curve it uses to fit the data; rather than fitting a straight line (as in linear regression), a smooth sigmoid curve is fit to the data. Logistic regression models the log odds ratio  $\ln(p/(1 - p))$ , where  $p$  is the probability of event occurring, as a linear combination of the independent variable(s). The equation representing this modifies the above linear regression equation by replacing  $Y$  on the right side of the equation with the odds ratio:

$$\ln(p/(1 - p)) = B_0 + B_1.X$$

Modelling the log odds ratio allows estimation of the probability of class membership (the values of the categorical variable e.g. if survival, alive or dead) using a linear relationship, similar to linear regression<sup>339</sup>. We can transform the log odds back to a probability as  $p(t) = 1/(1 + \exp(-t))$ , where  $t = B_0 + B_1X$ . This function produces a S-shaped sigmoid curve, with a steepness controlled by  $B_1$  that maps the linear function back to probabilities between zero to one.

In linear regression models, the fit can be assessed by using  $R^2$ , which estimates the proportion of variability explained by the regression model. It is calculated as one minus the sum of squared errors in the proposed model divided by the sum of squared errors in null model. When the model has multiple independent variables the adjusted  $R^2$  is more useful; it is adjusted for the number of covariates in the model, and will only increase if each additional variable reduces the overall error of the predictions. For logistic regression models, pseudo  $R^2$  methods that approximate the  $R^2$  from linear regression models are used to estimate goodness of model fit. A popular method is Nagelkerke  $R^2$  which estimates the improvement of the fitted model over the null model<sup>340</sup>.

#### **2.3.2.4.2.3 Survival (time to event) analysis**

With time to event data (TTE) the outcome of interest is not only whether the event has occurred, but also when that event occurred. Logistic and linear regression are not able to include both the event and time components in the outcome of the model, and are also not able to handle censoring; a unique type of missing data where subjects do not experience the event of interest during the follow-up time of the data collection period<sup>341</sup>.

The survival function, which computes the probability of meeting the outcome of interest beyond the timepoint  $t$ , can be estimated by a Kaplan Meier curve. Each step on the KM curve represents the proportion of individuals that have met the outcome at that time point, while the bars represent that individuals who have been censored (observation period has ended before the individual met the outcome). The KM curve estimates the probability of surviving to the end of the observed time period, conditioned on surviving up to the beginning of the time period<sup>341</sup>. The most commonly used method for comparing the survival curves between the survival curves of two groups is the log-rank test. Although the KM estimator is the most common approach for describing overall survival distributions and comparing them across groups, it is not able to account for covariates, and so if one's goal is to assess the effect of different risk factors on survival, multivariate statistical modelling is required.

These models can be broadly divided into two categories: proportional hazard (PH) approaches and accelerated failure time (AFT) models. In this thesis I use the Cox PH (CPH) model, a semi-parametric proportional hazards approach (Chapter 8). For a more detailed review of the other PH hazard approaches, such as the Exponential, Weibull or Gompertz models, and AFT models I refer the reader to an excellent summary by Bradburn et al<sup>342</sup>. Given the CPH model is a form of multivariate regression it provides a way to estimate survival, while at the same time allowing analysis of the effect of covariates on outcome. The fact that is semi-parametric, means that it has less strict assumptions about the underlying TTE data distribution than the parametric PH models. In essence the CPH model describes the relationship between the event incidence, as expressed by the hazard function, and the covariates<sup>342</sup>. The hazard is defined as the instantaneous event probability at a given time i.e. the probability that an individual under observation experiences the event at a timepoint  $t$ . This is in contrast to the odds ratio and relative risk which only refer to cumulative probabilities and the total number of events for a single defined timepoint over an entire.

The CPH model is expressed mathematically as:

$$h(t) = h_o(t) \times \exp \{b_1x_1 + b_2x_2 + \dots + b_px_p\}$$

Where the hazard function  $h(t)$  is dependent on a set of  $p$  covariates ( $x_1, x_1 \dots x_p$ ), whose relative impact is related to the size of the respective beta coefficients ( $b_1, b_1 \dots b_p$ ).  $h_o$  refers to the baseline hazard, and is value of the hazard when the dependent variable(s)  $x$  is equal to zero. The beta coefficients from the CPH model or on a log scale, and the exponents of the coefficients give the hazard ratio (HR). The HR refers to the ratio of the two hazard functions (the hazard functions for the two possible outcomes encoded in the dependent variable of interested). An important assumption in the CPH model is that of proportional hazards; that is that the hazard ratios for the independent variable are constant over time. It is important to verify that this assumption for all independent variables included in a CPH model.

As with logistic regression, pseudo- $R^2$  methods<sup>340</sup> have to be used to calculate goodness of fit in CPH models. As previously discussed, these calculate the  $R^2$  based on the improvement in likelihood between the fitted model and the model without dependent (predictor) variables (the null model).

### 2.3.2.5 Post-GWAS analysis

#### 2.3.2.5.1.1 Functional Mapping and Annotation of GWAS (FUMA)

To annotate the GWAS summary statistics in **Chapter 7** and **Chapter 8** I used the online platform Functional Mapping and Annotation (FUMA) of Genome- Wide Association Studies (<https://fuma.ctglab.nl/>) (version 1.3.6)<sup>343</sup>. FUMA utilises information from multiple biological resources to enable functional annotation of GWAS results, gene prioritisation and interactive visualisation. The SNP2GENE function in FUMA identifies independent SNPs and risk loci based on user defined thresholds, annotates these SNPs with available functional data and then maps them to genes, facilitating insight into the directional biological implications of significant genetic associations.

Using the provided GWAS summary statistics, FUMA identifies independent SNPs and their surrounding genomic loci based on the linkage disequilibrium (LD) structure. Independent significant SNPs (and SNPs in LD) are then annotated for functional consequences on gene functions (based on Ensembl genes (build 85) using

ANNOVAR<sup>344</sup>), deleteriousness score (CADD score<sup>345</sup>), potential regulatory functions (RegulomeDB score<sup>346</sup>) and 15-core chromatin state predicted by ChromHMM<sup>347</sup> for 127 tissue/cell types<sup>348,349</sup>, effects on gene expression using eQTLs of various tissue types and 3D structure of chromatin interactions with Hi-C data<sup>343</sup>. These SNPs are also cross-checked with the GWAS Catalog<sup>350</sup>, to identify whether they have previously been reported in association with other phenotypes.

Annotated SNPs are then mapped to genes based on positional, expression quantitative trait loci (eQTLs), and chromatin interaction information. For positional mapping, SNPs are mapped to genes based on physical proximity (up to 10 kb away). eQTL mapping is based on *cis*-eQTLs for genes within 1Mb of the locus, utilising information from four main data repositories: Genotype-Tissue Expression v8 database (GTExv8)<sup>351</sup>, Blood eQTL browser<sup>352</sup>, BIOS QTL browser<sup>353</sup> and BRAINEAC<sup>354</sup>. Chromatin interaction mapping maps SNPs to genes when there is a significant chromatin interaction between genes and GWAS significant loci, using Hi-C data from 14 tissue types<sup>355</sup>.

#### 2.3.2.5.1.2 Colocalisation analysis

In the GWAS/GWSS analyses (**Chapter 7** and **Chapter 8**) I carried out colocalisation analysis using the *coloc* R package for all genes within  $\pm 1$ Mb of the lead genomic loci SNP (version 5.1.0; <https://cran.r-project.org/web/packages/colocr/index.html>)<sup>356</sup>. This method assesses whether two association signals are consistent with a shared causal variant. By integrating GWAS and expression quantitative trait loci (eQTL) summary statistics, one can use this method to assess which gene at a given GWAS locus is potentially mediating the disease signal.

*Coloc* uses a Bayesian inference approach to estimate the posterior probability corresponding to one of five hypotheses (H0, H1, H2, H3, H4), related, in this thesis, to the expression of the tested gene (*cis*-eQTL, trait 1) and tested PiD outcome (disease risk or survival event, trait 2):

- H0: No association with either trait.
- H1: Association with expression of the gene, but not the PiD trait.
- H2: Association with the PiD trait, but not expression of the gene.
- H3: Association with the PiD trait and expression of the gene, with distinct causal variants

- H4: Association with the PiD disease trait and expression of the gene, with a shared causal variant.

A high H4 probability ( $> 0.85$ ) is considered as significant support for colocalization between the GWAS and the cis-eQTL traits i.e. signal colocalisation. *Coloc* calculates Bayes factors under the assumption that there is a single casual variant at a locus; it is therefore important to perform a conditional analysis at the lead genomic loci to ensure that there are no additional independent signals present. In this thesis I use GCTA-COJO software (version 1.93.0 beta for Linux; <https://yanglab.westlake.edu.cn/software/gcta/#Overview>) to perform association analyses conditioned on SNPs of interest. Further details on this method can be found in the original paper<sup>357</sup>.



# Chapter 3: A data-driven model of brain volume changes in Progressive Supranuclear Palsy Richardson Syndrome

The work in this chapter was published in *Brain Communications* in April 2022:

Scotton WJ, Bocchetta M, Todd E, Cash DM, Oxtoby N, VandeVrede L, Heuer H; PROSPECT Consortium, 4RTNI Consortium; Alexander DC, Rowe JB, Morris HR, Boxer A, Rohrer JD, Wijeratne PA. A data-driven model of brain volume changes in progressive supranuclear palsy. *Brain Commun.* 2022 Apr 14;4(3):fcac098.<sup>358</sup>

Scientific commentary is provided in the same journal edition by Franzmeier et al:

Franzmeier N, Hoglinger GU. Inferring the sequence of brain volume changes in progressive supranuclear palsy using MRI. *Brain Commun.* 2022 May 12;4(3):fcac113.<sup>359</sup>

## 3.1 Introduction

No effective disease modifying treatment has yet been proven for PSP, despite recent successful clinical trials<sup>221,224</sup>. Clinical trials in PSP can be complicated by variable disease stage at trial entry, highlighting the importance of stratifying patients into homogenous cohorts based on disease stage with similar rates of disease progression. Although the PSP Rating Scale has been shown to be a good independent predictor of survival<sup>228</sup>, and is used as the primary endpoint in clinical trials, such clinical biomarkers are only indirect measures of the biological stage of disease, and are affected by intra- and inter-rater variability, as well as fluctuation in patients' clinical state. Reliable and individualised disease progression markers are therefore required to complement clinical ratings scales<sup>229</sup>.

Structural MRI reveals significant atrophy in the brainstem and subcortical structures in PSP-RS, with additional involvement of the cortical structures<sup>240</sup>. Increased rates of atrophy in these regions can be detected over a 12-month period<sup>259,265</sup>, offering a potential biomarker readout for clinical trials. While there are new tau PET tracers emerging that show potential in the 4R tauopathies, these are not yet validated for use in the clinic setting<sup>255,360</sup> and in the absence of a validated tau PET tracer for PSP, structural MRI offers an indirect measure of underlying tau pathology in vivo. Indeed, a previous study in PSP showed that in vivo structural imaging reflected the independent contributions from tau burden and neurodegeneration at autopsy<sup>237</sup>, while

the link in Alzheimer's Disease is well established<sup>256,361</sup>. The recent pathology staging system for PSP defines six sequential stages of tau pathology progression, starting with the subthalamic nucleus, spreading out caudally to the cortex and rostrally to the cerebellum<sup>51</sup>. This has been validated in an independent cohort with increasing pathological stage correlating with clinical severity<sup>362</sup>. However, the order in which brain regions show evidence of increased atrophy in vivo is currently unknown.

One approach to estimating the sequence of atrophy progression is event-based modelling (EBM)<sup>281</sup>, using a probabilistic data-driven generative model to infer the order in which biomarkers become abnormal. The EBM can be built from cross-sectional data by combining severity information across biomarkers and individuals without reference to a given individual's clinical status. The EBM allows inference of longitudinal information about disease progression by assuming there is a monotonic progression of an individual biomarker from normal to abnormal (even if this progression is non-linear), so that in a patient cohort containing a spectrum of disease stages, more individuals will necessarily show abnormality in a biomarker that changes early in the disease course. This approach has been successfully applied to Huntington's disease<sup>267</sup>, sporadic and familial Alzheimer's disease<sup>280,282,283</sup>, Parkinson's disease<sup>286</sup>, multiple sclerosis<sup>285</sup>, the posterior cortical atrophy variant of Alzheimer's disease<sup>284</sup>, and to amyotrophic lateral sclerosis<sup>287</sup>, providing a simple and validated method to investigate temporal disease patterns and estimate individuals' disease stage. Recent work has demonstrated the clinical utility of the EBM for screening patients on entry into clinical trials, to improve cohort homogeneity and increase the power to detect a treatment effect<sup>288</sup>.

The aim of this chapter was to define the progression of brain atrophy in clinically diagnosed PSP-RS by developing an EBM that takes cross-sectional structural MR imaging as input. I hypothesised that there is a consistent sequence in which brain regions become atrophic in PSP-RS in vivo, in keeping with the sequential stages of PSP pathology proposed by Kovacs et al.<sup>51</sup> in their post-mortem staging study. I predicted that the image-based EBM stage would be correlated with clinical disease severity as measured by the PSP Rating Scale.

## 3.2 Methods

### 3.2.1 Participants and clinical data collected

Data from individuals with a clinical diagnosis of possible or probable PSP-Richardson Syndrome were collected from the 4R tau imaging cohort. The details of this cohort that I built as part of this PhD are covered in more detail in **Chapter 2**. There were seven main sources for PSP cases included in this cohort: the 4R Tauopathy Imaging Initiative Cycle 1 (4RTNI 1; ClinicalTrials.gov: NCT01804452)<sup>259,263</sup>, the 4R Tauopathy Imaging Initiative Cycle 2 (4RTNI 2; ClinicalTrials.gov: NCT02966145), the davunetide randomized control trial (DAV; ClinicalTrials.gov: NCT01056965)<sup>220</sup>, the salsalate clinical trial (SAL; ClinicalTrials.gov: NCT02422485)<sup>297</sup>, the young plasma clinical trial (YP; ClinicalTrials.gov: NCT02460731)<sup>297</sup>, the PROgressive Supranuclear Palsy CorTico-Basal Syndrome Multiple System Atrophy Longitudinal Study (PROSPECT; ClinicalTrials.gov: NCT02778607), and the University College London Dementia Research Centre (UCL DRC) FTD cohort. Control data were collected from three sources: the Frontotemporal Lobar Degeneration Neuroimaging Initiative dataset (FTLDNI; <http://4rtni-ftldni.ini.usc.edu/>), PROSPECT, and the UCL DRC FTD Cohort. Controls were defined as no known diagnosis of a neurological or neurodegenerative condition, and no known history of memory complaints.

Further details on these individual cohorts are included in **Chapter 2**, and a summary of the demographics of the PSP-RS cases within each cohort is included in **Table 3.1**. Appropriate ethics was applied for and approved via the relevant trial and research ethics committees. For inclusion in this study all patients had to have, as a minimum, a baseline T1-weighted volumetric MRI on a 1.5T or 3T scanner, with basic demographic data (age at time of scan, gender), and disease duration at time of the scan (time from symptom onset to MRI scan) if available. 12-month follow-up scans, if available, were also included in the study, as were PSP Rating scale scores. Given original trial analyses failed to show any treatment effect (including no change in volumetric MRI measurements) in the davunetide<sup>220</sup>, salsalate and young plasma trials<sup>297</sup>, I combined data from each study's treatment and placebo groups. Longitudinal data (both 12-month follow-up MRI and PSP Rating Scale) were used for validation of the staging system produced by the baseline EBM.

**Table 3.1 - Overview of PSP-RS cases by cohort included in study.**

<b>Baseline Demographics</b>	<b>Total</b>	<b>4RTNI1</b>	<b>4RTNI2</b>	<b>DAV</b>	<b>SAL/YP</b>	<b>PROSPECT</b>	<b>UCL</b>	<b>Controls</b>
Inclusion, n (12 mths)	365 (275)	62 (50)	43 (33) <sup>a</sup>	187 (177)	14 (0)	36 (12)	23 (3)	289 <sup>b</sup>
Post-QC, n (12 mths)	341 (255)	59 (50)	43 (33)	173 (157)	13 (0)	30 (12)	23 (3)	260
Gender, % female	48%	54%	47%	51%	48%	43%	30%	57%
Age at first scan, yrs	67.9 (6.8)	70.5 (7.4)	69.2 (7.2)	67.4 (6.6)	69.4 (4.0)	67.2 (8.5)	66.1 (4.9)	62.6 (9.8)
Disease duration <sup>c</sup> , yrs	4.1 (3.1)	5.4 (3.9)	4.6 (3.6)	15% > 5 years <sup>d</sup>	-	2.9 (1.9)	3.5 (1.9)	-

Values are Baseline n (n 12 month follow-up visits) or mean (SD) unless otherwise stated. <sup>a</sup> at time of analysis only 43 of the total 59 PSP-RS cases from 4RTNI2 had been received and processed

<sup>b</sup> control cohort consists of healthy controls from FTLDNI, PROSPECT and UCL with no evidence of neurological disease and otherwise fit and healthy. At time of analysis only 289 of 320 controls finally included in the 4R tau imaging cohort had been received and processed. <sup>c</sup> time from first symptom to first scan <sup>d</sup> for cases included in Davunetide trial disease, duration was only recorded as greater or less than 5 years since disease onset.

### 3.2.2 MRI acquisition and image processing

Detail of the MRI image processing pipeline is discussed in more detail in **Image processing pipeline**. In brief, raw volumetric T1 MRI images were all processed by the same pipeline. Scans first underwent visual quality control (QC) to ensure correct acquisition and the absence of major artefacts. Next, raw images that passed QC were bias field corrected for magnetic field inhomogeneity, and the whole brain (cortical and subcortical structures) parcellated using the geodesic information flow (GIF) algorithm<sup>304</sup>. This automatically extracts regions based on the Neuromorphometrics atlas (Neuromorphometrics, Inc.), using an atlas propagation and label fusion strategy<sup>363,364</sup>. Subregions of the cerebellum were then automatically extracted with GIF based on the Diedrichsen cerebellar atlas: the cerebellar lobules (I-IV, V, VI, VIIa-Crus I, VIIa-Crus II, VIIb, VIIa, VIIb, IX and X), the vermis and the deep nuclei (dentate, interposed and fastigial)<sup>304,305</sup>. The whole brainstem, medulla, pons, superior cerebellar peduncles (SCP) and midbrain were subsequently segmented using a customised version of the module available in FreeSurfer to accept the GIF parcellation as input for FreeSurfer<sup>306</sup>. Total intracranial volume (TIV)<sup>365</sup> was calculated using SPM12 v6225 (Statistical Parametric Mapping, Wellcome Trust Centre for Neuroimaging, London, UK) running under MATLAB R2012b (Math Works, Natick, MA, USA). All segmentations were visually inspected to ensure accurate segmentation.

### 3.2.3 Biomarker selection

In this study I use the term biomarker to refer to image-based regional brain volumes that show a significant difference between cases and healthy controls (two-tailed t-test of mean difference in covariate adjusted volumes). Given the focus of this study was to test the hypothesis that the sequence of atrophy in PSP-RS is in keeping with the sequence of tau pathology at post-mortem as shown by Kovacs et al.<sup>51</sup>, I chose nineteen regions of interest (ROI) for inclusion that most closely matched those used in their study; four brainstem (medulla, pons, superior cerebellar peduncle [SCP], and midbrain), three cerebellar (cerebellar cortex, deep nuclei and vermis), seven subcortical (thalamus, globus pallidus [GP], striatum [caudate and putamen], ventral diencephalon [DC], thalamus, hippocampus and amygdala) and five cortical (frontal, insula, temporal, parietal and occipital). Regions that had a right and left label were

combined. All ROIs were controlled for the following covariates using linear regression on the control cohort: age at scan, sex, scanner type and TIV. Linear regressions of age against predicted EBM stage were also performed (after EBM model fitting) for cases and controls separately to confirm that there was no residual correlation after adjustment. All regions selected for inclusion showed a significant difference in covariate adjusted volumes between cases and controls (Bonferroni corrected threshold of  $p < 3 \times 10^{-3}$ ) under a two-tailed t-test.

### 3.2.4 The Event Based Model

Given the utility of the EBM method to model a disease process as a sequence of events at which individual biomarkers become abnormal, I fitted an EBM to the selected ROIs from all baseline scans to calculate the maximum likelihood sequence of event orderings. Details of the EBM, including the mathematical formulation of estimating the event distributions and model fitting are provided in **Chapter 2**. To obtain the likelihood models for the event distributions ( $P(x_{ij}|E_i)$  “post-event” and  $P(x_{ij}|\neg E_i)$  “pre-event”), I used the KDE method described by Firth et al.<sup>284</sup>. Due to the absence of prior information, a uniform prior likelihood ( $P(k)$ ) was used to impose as little information as possible on estimated orderings. I then staged all individuals based on their baseline scan, using the method laid out in Patient staging.

### 3.2.5 Cross validation of event sequence

Although the MCMC sampling (as detailed in **Model Fitting** section below) of the posterior distribution gives some information on the uncertainty of the event ordering predicted by the EBM, previous work shows it tends to underestimate this uncertainty<sup>280</sup>. Bootstrapping is an additional method that tends to give a more liberal estimate of the uncertainty in the ordering<sup>366</sup>. I first performed cross-validation of the maximum likelihood sequence generated by the EBM, by re-estimating the model on 100 bootstrap samples of the original data (sampling with replacement). I then performed repeated stratified five-fold cross-validation<sup>307</sup> as an additional check on the robustness of the model. This involved refitting the model on 80% of the cohort data and testing accuracy on the held out 20% for each of ten five-fold random partitions, giving a total of 50 cross-validation folds/models, which are averaged to find the final model sequence.

### 3.2.6 Longitudinal validation

I investigated the longitudinal consistency of the staging produced by the EBM, based on the predictions that, firstly, given PSP is a progressive disease, the EBM stage should increase over time, and secondly that increasing EBM stage should be associated with both increasing PSP Rating Scale score (the main clinical measure of disease severity) and also disease duration, especially during later model stages where there is more widespread atrophy. I first staged patients using the baseline EBM based on their 12-month follow-up scan (255 cases) and compared this with predicted stage based on their baseline scan. The follow-up data was processed using the same pipeline as the baseline scans to produce the same ROI biomarkers at 12-months. To test for the relationship of PSP Rating Scale score with baseline EBM stage, a linear mixed effects model was fit to the data using the *lme4* package<sup>367</sup> in R Studio (version 1.4.1106), with EBM defined stage as the independent variable and PSP rating scale score as the dependent variable. 241 baseline and 232 12-month follow-up scans (473 total) had a corresponding PSP rating scale score. Subject Id was modelled as a random effect (random intercept) due to some subjects having two MRI scans at different time points. Significance was calculated using the *lmerTest* package<sup>368</sup> which applies Satterthwaite's method to estimate degrees of freedom and generate p-values for mixed models. In addition, I analysed disease duration (time from first symptom to MRI scan) as a function of predicted EBM stage (87 baseline and 43 twelve-month follow-up scans had disease duration recorded) using the same method. To confirm that baseline EBM stage was also correlated with both PSPRS score and disease duration I fitted a linear model for each as a function of EBM stage.

## 3.3 Results

### 3.3.1 Subject characteristics

**Table 3.2** summarises the key demographic data for the cohort included in the study. 929 MRI images were processed from a total of 654 subjects: 365 with a clinical diagnosis of PSP-RS (of which 275 had 12-month follow-up scans) and 289 controls. Of the PSP-RS cases 26 (8%) had a pathological diagnosis after coming to post-mortem: 24 (92%) showed tau pathology consistent with PSP, while 2 cases had non-PSP tau pathology (one CBD and one GGT) and were therefore excluded from the

analysis. After stringent quality control with visual inspection of all images for the remaining 363 cases (pre- and post- processing), 341 PSP-RS cases (of which 255 had 12-month follow-up scans) and 260 control scans were included for the analysis. Reasons for scans failing quality control included poor quality of the raw T1 image (usually due to movement artefacts) or inaccurate segmentations with the GIF or / and SPM algorithms. 70% (241/341) of the cases included had a PSP rating scale score at baseline and follow-up, as well as recorded age, gender, scanner type and TIV. At baseline the PSP-RS cohort had an older average age (67.9 years, standard deviation [SD]  $\pm$  6.8) compared to healthy controls (62.8 years,  $SD \pm 9.4$ ,  $t = -7.4$ ,  $p < 0.01$ ). Disease duration data (time from diagnosis to baseline visit [average years,  $\pm$  SD]) was available for 87/341 cases and showed an average length of 4.1 years ( $SD \pm 3.1$ ). There was a higher proportion of females in the control group compared to the PSP-RS group (male / female, 112/148 vs 176/165 respectively,  $\chi^2 = 4.3$ ,  $p = 0.04$ ).

**Table 3.2 - PSP-RS EBM baseline demographics.**

Baseline Demographics	PSP-RS	Controls	<i>p</i> value
Baseline, n (12 mths)	341 (255)	260	-
Gender, M/F (% female)	176/165 (48%)	112/148 (57%)	0.04 <sup>a</sup>
Age at first scan, yrs	67.9 (6.8)	62.8 (9.4)	<0.001 <sup>b</sup>
Disease duration, years <sup>c</sup>	4.1 (3.1)	-	-
Pathology (% PSP)	24 (92%) <sup>d</sup>	-	-
PSP Rating Scale	38.9 (12.9) <sup>e</sup>	-	-
UPDRS	30.6 (15.1)	-	-
MOCA	20.7 (5.1)	-	-

Values are Baseline n (n 12 month follow-up visits) or mean (SD) unless otherwise stated. <sup>a</sup>  $\chi$  square. <sup>b</sup> unpaired two-tailed *t*-test. <sup>c</sup> time from first symptom to first scan. <sup>d</sup> % of all cases pre-QC. Two cases (GGT and CBD removed from subsequent analysis). <sup>e</sup> 70% (241/341) of baseline cases included had a PSP rating scale score. Abbreviations: EBM = Event based model, PSP-RS = Progressive Supranuclear Palsy Richardson Syndrome



### 3.3.2 Sequence of atrophy progression

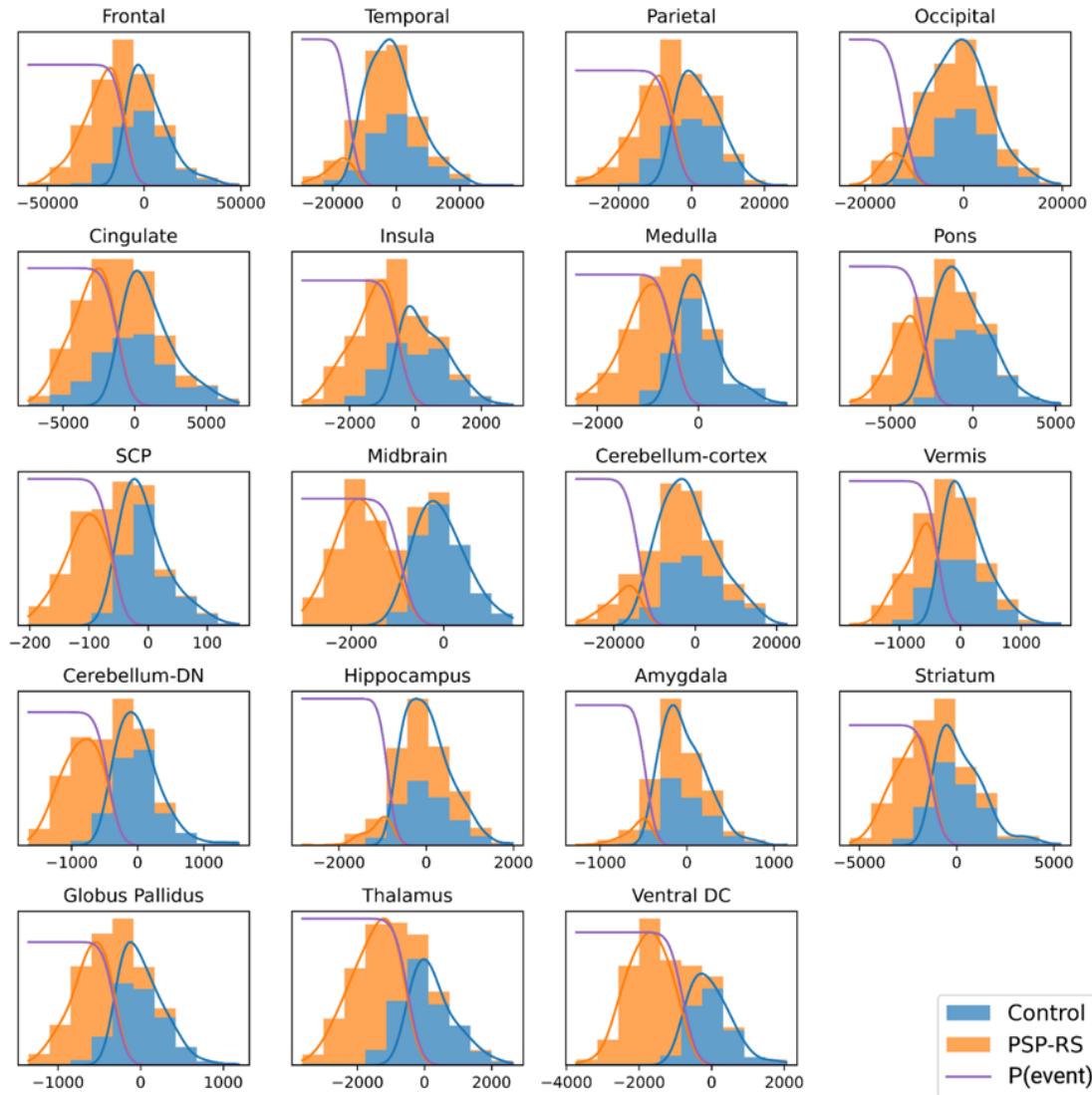
**Figure 3.1** shows histograms of the healthy control (HC) and covariate adjusted PSP-RS ROI biomarker distributions, with KDE mixture model fits and line showing probability of an event. These fits provide the parameters for the normal and abnormal likelihoods,  $P(x_{ij}|E_i)$  and  $P(x_{ij}|\neg E_i)$ , respectively, that are then used to calculate the maximum likelihood sequence of the full dataset. At baseline all nineteen ROI selected for inclusion in the model showed a significantly smaller covariate adjusted volume in PSP-RS compared to controls.

The positional variance diagram in **Figure 3.2A** shows the most likely sequence in which these regions become atrophic, as estimated by the EBM, as well as the uncertainty in this sequence (based on MCMC sampling of the posterior distributions). The maximum likelihood sequence was estimated using PSP-RS cases only, based on the rationale that PSP is a rare disease, and it is very unlikely for my cohort of controls to have asymptomatic PSP. Indeed, it is more likely the controls would have a common disorder such as AD rather than PSP, and I did not want this to confound the sequence estimation hence the exclusion. The EBM estimated that the earliest atrophy occurs in the brainstem and subcortical regions followed by progression caudally into the superior cerebellar peduncle and deep cerebellar nuclei, and rostrally to the cortex. The sequence of cortical atrophy progresses in an anterior to posterior direction, beginning in the insula and then frontal lobe before spreading to the temporal, parietal and finally the occipital lobe (**Figure 3.2C**) The high colour intensity of each square and their presence predominantly on the diagonal of the positional variance diagram indicates that the model has a high degree of certainty regarding their positions in the overall sequence.

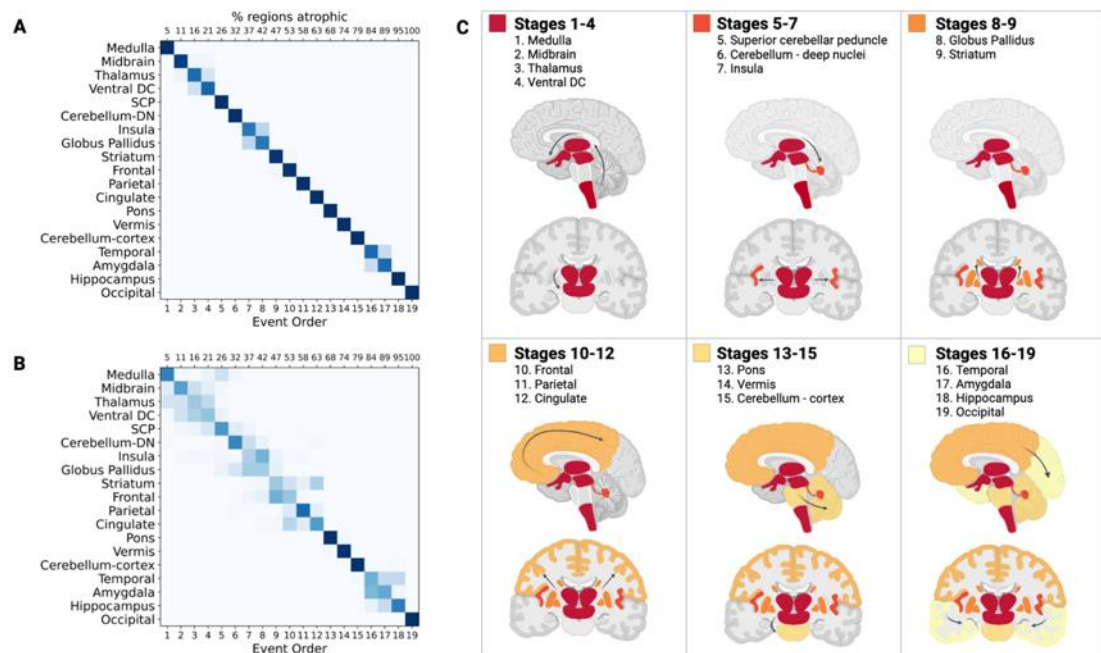
### 3.3.3 Cross validation of event sequence

**Figure 3.2B** shows positional variance of the maximum likelihood sequence re-estimated by bootstrapping of the data (random resampling with replacement 100 times) and refitting the model. The positional variance diagram for the bootstrapped results represents the proportion of bootstrap samples in which the event  $i$  (y axis) appears at position  $k$  (x axis) of the maximum likelihood sequence. The sequence ordering is generally preserved, though as one would expect with this more

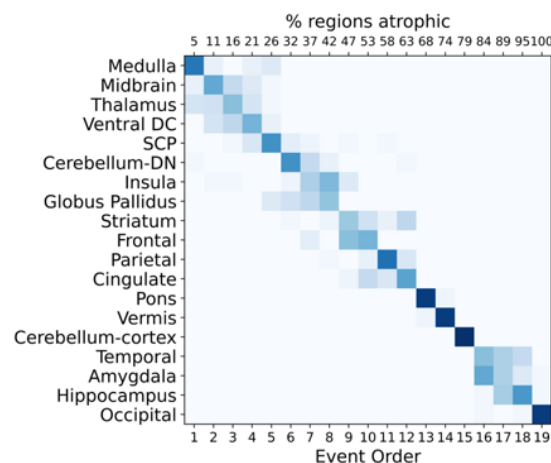
conservative estimate of uncertainty, there is increased uncertainty in the relative positions early in the sequence from stage two (midbrain) to stage 4 (ventral diencephalon), and in the middle from stage nine (striatum) to stage thirteen (pons). Using repeated stratified 5-fold cross-validation (**Figure 3.3**) as an alternative method to assess model robustness (both in terms of the sequence and uncertainty in the sequence), the maximum likelihood sequence is preserved with similar uncertainty in relative positions when visually compared to the bootstrapping method (**Figure 3.2B**)



**Figure 3.1 - Kernel Density Estimation (KDE) mixture models.** Healthy controls (blue) and PSP-RS (orange) volume biomarker distributions, and corresponding KDE mixture model fits. The purple line represents the probability that an event has occurred  $P(x_{ij} | E_i)$ . Note that the volumes are covariate corrected.



**Figure 3.2 - Sequence of atrophy progression in PSP Richardson Syndrome.** (A) Regional volume biomarker positional variance diagram showing the sequence of atrophy progression in PSP-RS. (B) Re-estimation of positional variance after cross-validation of the maximum likelihood event sequence by bootstrap resampling (100 bootstraps). For figures (A) and (B) the vertical ordering on the y-axis (from top to bottom) shows the maximum likelihood sequence estimated by the EBM (earliest to latest event). The bottom x-axis shows EBM stage while the top x-axis represents the percentage of regions atrophic (abnormal) at each stage. Colour intensity of the squares represents the posterior confidence in each biomarker's position in the sequence, from either (A) MCMC samples of the posterior or (B) bootstrapping. SCP = superior cerebellar peduncle, Ventral DC = ventral diencephalon. Note that because these volumes are covariate adjusted the control distribution will be centred at zero. (C) Graphic representation of the event sequence with relevant region transitioning from healthy (grey) to unhealthy (coloured). Dark red = first regions to atrophy, Light yellow = last regions to atrophy. Created with *BioRender.com*.



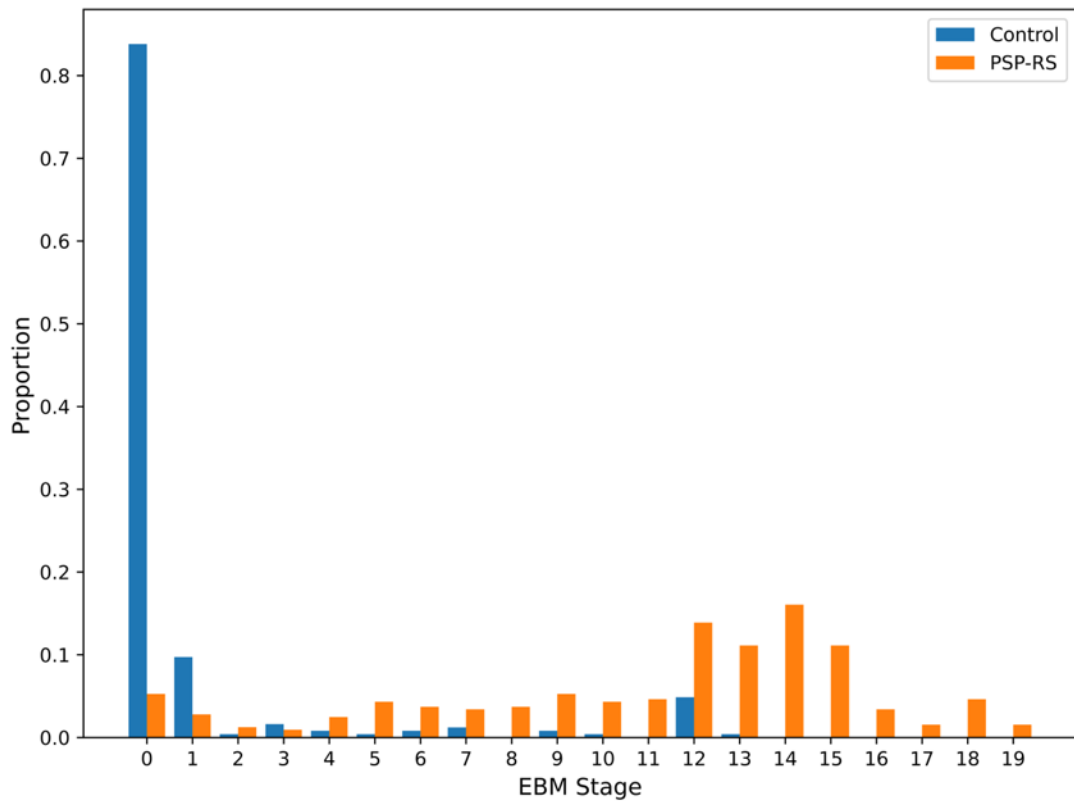
**Figure 3.3 - Sequence of PSP-RS atrophy progression after five-fold cross validation.** Re-estimation of positional variance after cross-validation of the maximum likelihood event sequence across 50-folds (10 repeats and 5-folds). The vertical ordering on the y-axis (from top to bottom) shows the maximum likelihood sequence estimated by the EBM (earliest to latest event). The bottom x-axis shows EBM stage while the top x-axis represents the percentage of regions atrophic (abnormal) at each stage. Colour intensity of the squares represents the posterior confidence in each biomarker's position in the sequence after cross-validation. SCP = superior cerebellar peduncle, Ventral DC = ventral diencephalon. Note that because these volumes are covariate adjusted the control distribution will be centred at zero.

### 3.3.4 Patient staging

**Figure 3.4** shows the proportion of subjects at each EBM defined stage (PSP-RS and HC). Patient staging results were evaluated using the maximum likelihood sequence (**Figure 3.2A**) of regional atrophy for PSP-RS subjects as described in the **Methods** section. As one would expect the HC cohort is clustered at the early stages with greater than 80% at Stage 0 (i.e., no event occurred), while the PSP-RS cases are distributed more evenly across stages with the highest proportion in the middle to late stages. This suggests that the cohort of PSP cases gathered from multiple different studies were temporally heterogenous which supports the importance of accurately staging patients using objective biomarkers.

Using a threshold of Stage 2 (medulla and midbrain atrophic) the model was able to correctly classify subjects as PSP-RS versus healthy control with an overall accuracy of 90% (with a sensitivity and specificity of 91% and 90% respectively). Although not the focus of this model the high classification accuracy provided by the EBM further demonstrates its clinical validity.

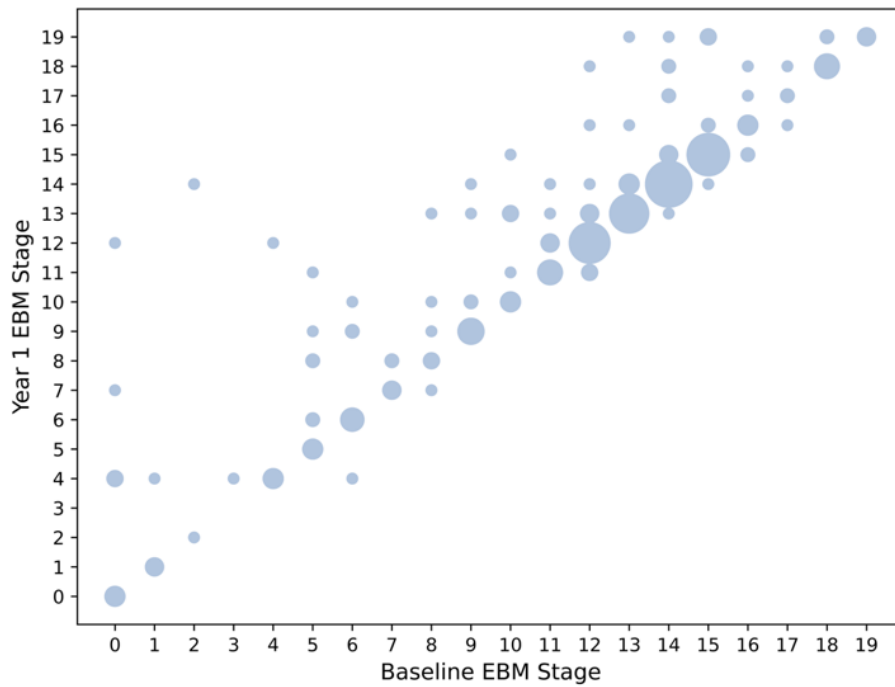
Outliers were present in both the HC and PSP-RS groups: specifically, 10 (4%) of PSP-RS cases were at Stage 0, while 14 controls were at Stage 10 or greater (5%). Visual inspection of these HCs suggested that the segmentations were accurate, but that there were non-specific covariate adjusted decreased volumes in regions including the hippocampus with relative sparing of the brainstem and subcortical structures, suggesting that these could potentially represent people with preclinical Alzheimer's disease.



**Figure 3.4 - Histogram of event-based model staging results for PSP-RS.** Healthy controls in blue and PSP-RS cases in orange. Each bar represents the proportion of patients in each category at each EBM stage. Each EBM stage on x-axis represents the occurrence of a new biomarker transition event. Stage 0 corresponds to no events having occurred and Stage 19 corresponds to all events having occurred. Events are ordered by the maximum likelihood sequence for the whole PSP-RS population as shown in **Figure 3.2A**.

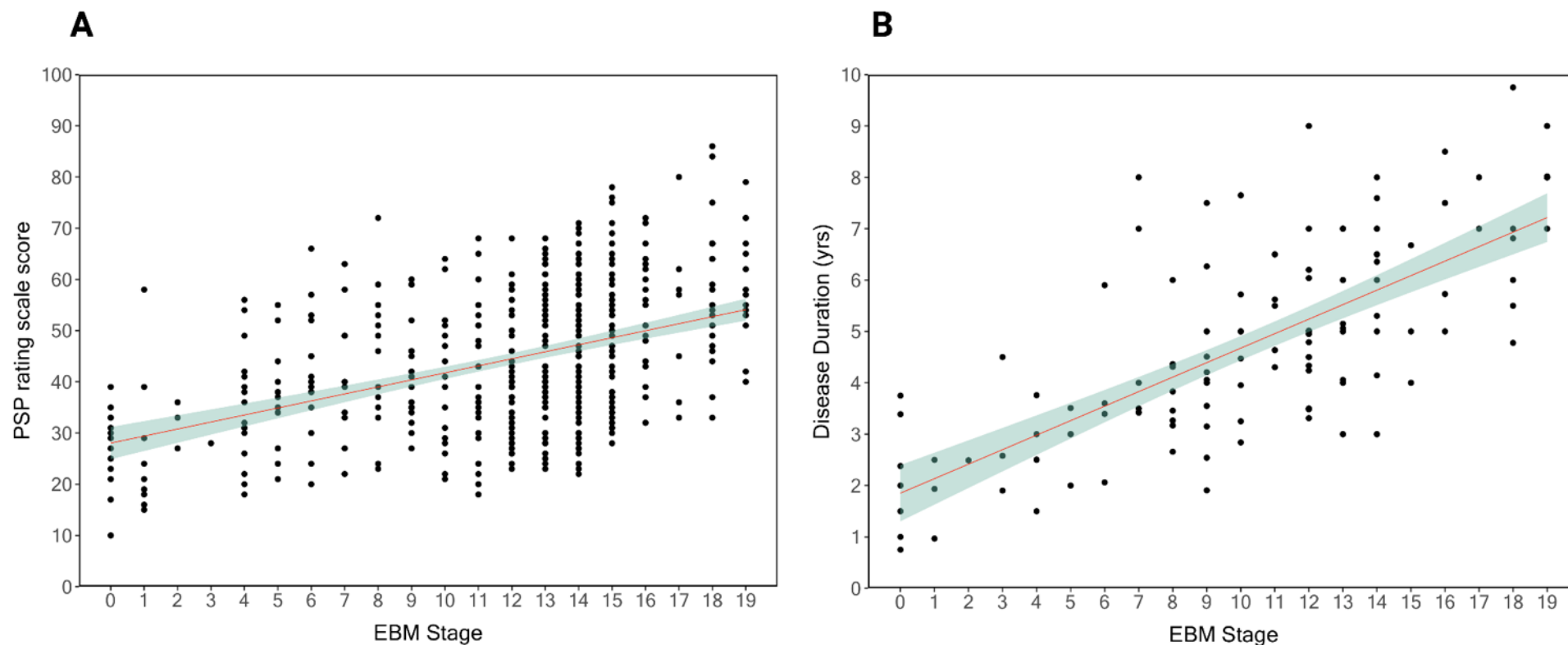
### 3.3.5 Longitudinal consistency

To test the validity of the EBM I first tested the hypothesis that a valid model will produce non-decreasing disease stages for individuals from baseline to follow-up, within the bounds of model uncertainty. **Figure 3.5** compares each PSP-RS subject's EBM stage at baseline with their stage at 12-month follow-up (255 cases had both a baseline and 12-month follow-up scan). Overall, on this metric the EBM shows good longitudinal consistency with each subjects EBM stage generally increasing or remaining stable at 12-months follow-up: 245/255 cases (96%) either stayed at the same stage or progressed. For these cases the average stage progression over 12 months was 1 stage. Of the ten PSP cases that reverted in stage, nine only dropped one stage while one dropped two stages.

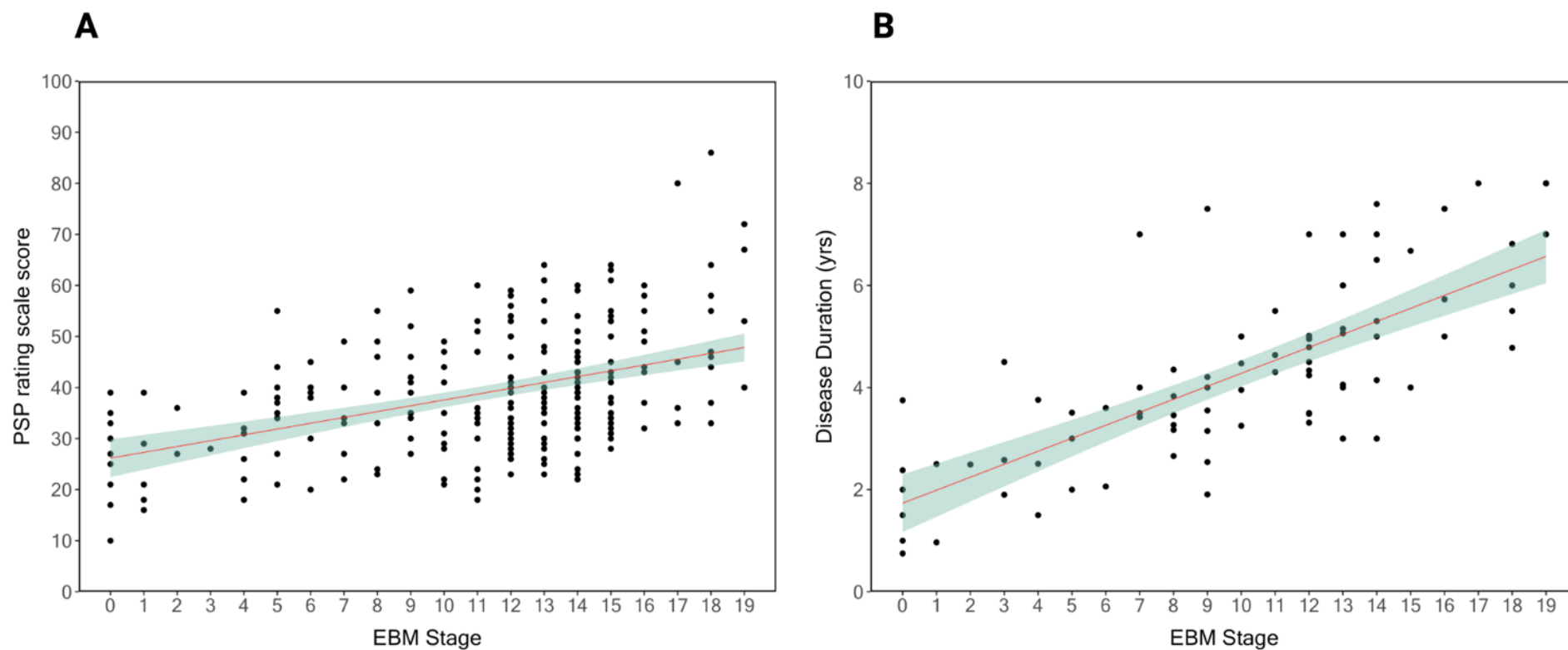


**Figure 3.5 - Longitudinal consistency of baseline EBM.** Scatter plot showing predicted stage at baseline (x-axis) versus predicted stage at 12 months (y-axis) for those PSP-RS subjects with a follow-up scan ( $n = 255$ ). The area of a circle is weighted by the number of subjects at each point.

To further validate the EBM, I first modelled PSP rating scale as a function of predicted EBM stage using a linear mixed model (**Figure 3.6A**). EBM stage was modelled as a fixed effect while Subject Id was modelled as random effect due to some subjects having two MRI scans at different time points. I found a significant fixed effect of EBM stage on predicted PSP rating scale ( $\beta=1.46$ , 95% CI 1.2-1.8,  $p<0.001$ ) and a conditional  $R^2$  of 0.56. I then modelled disease duration (years) as a function of predicted EBM stage, which showed a significant fixed effect ( $\beta=0.29$ , 95% CI 0.24-0.34,  $p<0.001$ ) and a conditional  $R^2$  of 0.68 (**Figure 3.6B**). When fitting linear models for PSPRS score and disease duration versus predicted EBM stage on baseline scans only (**Figure 3.7A** and **Figure 3.7B** respectively), there was also a significant association albeit with a lower adjusted  $R^2$  (PSPRS vs EBM stage at baseline:  $\beta=1.14$ , 95% CI 0.84-1.44,  $p<0.001$ ), adjusted  $R^2$  0.18, disease duration vs EBM stage at baseline: ( $\beta=0.25$ , 95% CI 0.20-0.30,  $p<0.001$ , adjusted  $R^2$  0.39). To check that I had adequately adjusted for age I also ran linear models of age as a function of predicted EBM stage for cases (**Figure 3.8A**) and controls separately (**Figure 3.8B**). There was no association between EBM stage and age in either the case ( $\beta=0.19$ , 95% CI=0.13-0.25,  $p=0.12$ , adjusted  $R^2=0.017$ ) or control group ( $\beta=-0.27$ , 95% CI=-0.66-0.12,  $p=0.18$ , adjusted  $R^2=0.003$ ).

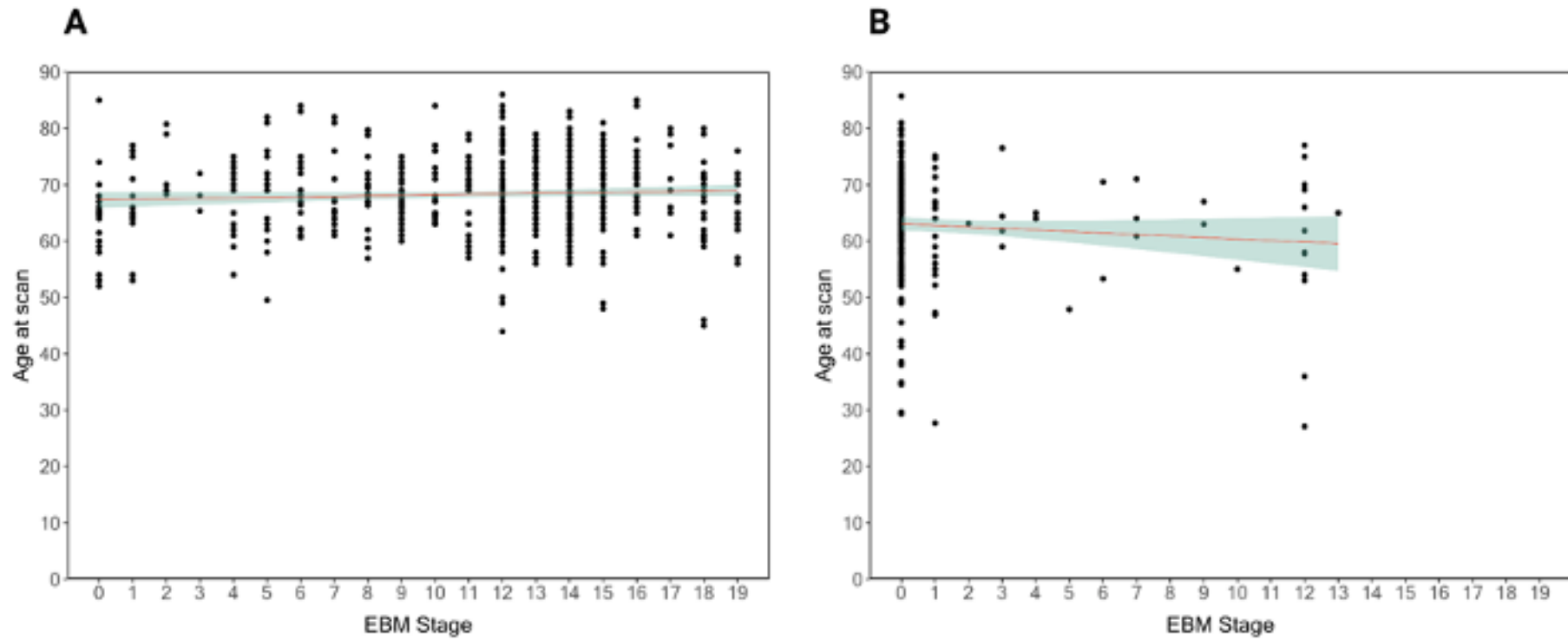


**Figure 3.6 - Association between predicted EBM stage, PSP Rating Scale score, and disease duration for all scans (baseline and follow-up).** (A) PSP Rating Scale score versus EBM stage\* ( $\beta=1.46$ , 95% CI 1.2-1.8,  $p<0.001$ , conditional  $R^2$  0.56 (marginal 0.22)) (B) Disease duration (years) versus EBM stage\*\* ( $\beta=0.29$ , 95% CI 0.24-0.34,  $p<0.001$  and a conditional  $R^2$  of 0.68 (marginal 0.41)). For both (A) and (B) the line represents the linear fixed effect model fit to all subjects, and 95% confidence intervals. Subject Id was modelled as a random effect (random intercept) due to some subjects having two MRI scans at different time points. Significance was calculated using Satterthwaite's method to estimate degrees of freedom and generate p-values. \*473 scans (241 baseline and 232 twelve-month follow-up) with PSPRS score \*\* 130 scans (87 baseline and 43 twelve-month follow-up) with disease duration.



**Figure 3.7** - Association between predicted EBM stage, PSPRS and disease duration at baseline scan only. **(A)** PSP Rating Scale score versus EBM stage\* ( $\beta=1.14$ , 95% CI 0.84-1.44,  $p<0.001$ , adjusted R<sup>2</sup> 0.18). **(B)** Disease duration (years) vs EBM stage\*\* ( $\beta=0.25$ , 95% CI 0.20-0.30,  $p<0.001$ , adjusted R<sup>2</sup> 0.39). For both **(A)** and **(B)** the line represents the linear model fit with 95% confidence intervals.\* 241 baseline scans with PSPRS score \*\* 87 baseline scans with disease duration.





**Figure 3.8 - Linear models to test association between age at scan and predicted EBM stage.** (A) for cases ( $\beta=0.19$ , 95% CI=0.13-0.25,  $p=0.12$ , adjusted  $R^2=0.017$ ) (B) for controls ( $\beta=-0.27$ , 95% CI=-0.66-0.12,  $p=0.18$ , adjusted  $R^2=0.003$ ). For (A) and (B) the line represents the linear model fit with 95% confidence intervals.

## 3.4 Discussion

The principal result of this study is that a probabilistic data-driven method reveals, *in vivo*, the sequence in which brain regions become atrophic in PSP-RS. I established this sequence from cross-sectional data and went on to demonstrate the validity of this model longitudinally. Ninety-six percent remained in the same stage or progressed to a later stage over twelve-months. The model derived staging correlated with both clinical severity and disease duration.

### 3.4.1 Ordering of biomarkers

The order of regional atrophy revealed by the EBM (**Figure 3.2**) broadly mirrors the sequential spread of tau pathology in PSP proposed by Kovacs et al.<sup>51</sup>. The earliest atrophy in my model occurs in the brainstem and subcortical regions followed by progression caudally into the superior cerebellar peduncle and deep cerebellar nuclei, and rostrally to the cortex. The sequence of cortical atrophy progresses in an anterior to posterior direction, beginning in the frontal lobe before then spreading to the temporal, parietal and finally the occipital lobe. In the absence of external data on which to validate the model, I explored the generalisability and robustness of the model using two different validation methods: bootstrap cross validation and repeated stratified five-fold cross-validation. These demonstrate that even with a more conservative estimate of uncertainty, the sequence of atrophy is largely conserved (**Figure 3.2B** and **Figure 3.3**). There remains uncertainty early on between the relative positions of the midbrain, thalamus, ventral DC and SCP, in the middle between the striatum, frontal, parietal, and cingulate lobes, and the pons, and at the end of the sequence between the temporal lobe, amygdala, and hippocampus. This heterogeneity is of interest, and a motivation for future work.

It is difficult, however, to make a direct comparison between my *in vivo* findings and post-mortem tau histopathology staging for two reasons: firstly, in this study I am measuring atrophy rather than tau pathology directly, and although there is evidence that atrophy on structural imaging is associated with tau pathology<sup>237,369</sup> it is unlikely to directly correlate with histopathological scores of tau accumulation across neuronal and glial cell populations. Secondly, two of the regions identified to have the earliest tau pathology in Kovacs' study are the subthalamic nucleus (STN) and the substantia

nigra (SN), regions that are not individually segmented by the GIF algorithm used in this study. These are subsumed within the ventral diencephalon (ventral DC) segmentation in the Neuromorphometrics atlas, along with the hypothalamus. Although not specific for the STN and SN, reassuringly this region does occur early in the sequence (**Figure 3.2A**), and after cross validation one can see (**Figure 3.2B** and **Figure 3.3**) that after the medulla there is uncertainty as to the exact ordering of the midbrain, thalamus, and ventral DC.

The majority of cross-sectional imaging studies in PSP-RS, have focused on the clinical utility of structural MR imaging as a diagnostic biomarker to differentiate PSP from both PD and other atypical parkinsonian disorders<sup>229</sup>. These studies usually only give a group level overview of regional atrophy at baseline, as opposed to the sequence of atrophy changes that I have demonstrated in this study. Even so midbrain atrophy is commonly seen in PSP-RS at baseline, with relative sparing of the pons<sup>370–372</sup>, and the pons to midbrain ratio has high specificity and sensitivity for the diagnosis of pathologically confirmed PSP<sup>373</sup>. SCP atrophy is also evident early in the disease course<sup>374</sup> and has led to the development of the MR Parkinsonism Index (MRPI) for differentiation PSP-RS from other causes of parkinsonism<sup>375</sup>. Atrophy of subcortical structures including the striatum, globus pallidus and thalamus has also been observed in group-level studies<sup>90,244,262,376–378</sup>, as well as involvement of frontal lobe<sup>379–381</sup>. Together these findings are consistent with the sequence of atrophy that the EBM produces, but this study is the first in PSP-RS, to the best of our knowledge, that orders these regions relative to each other.

The placement of the medulla first in the sequence is interesting as the medulla is not widely mentioned in the PSP imaging literature. It is however clear that tau pathology is consistently seen in the medulla at post-mortem<sup>69,149</sup>, with Kovacs<sup>51</sup> placing it at Stage 2 in their pathological staging system. More recently, perhaps due to the advent of automated segmentation techniques for the brainstem, its involvement has been shown in PSP-RS using MRI<sup>298,371,372,382</sup>. The early involvement of the thalamus in our EBM sequence is also supported both by pathological studies<sup>51</sup> where tau pathology been shown to occur in all cases, and structural MRI studies that demonstrate atrophy: in particular the pulvinar, dorsomedial, and anterior nuclei<sup>371,383</sup>. In future work it will be interesting to investigate differential involvement of the thalamic nuclei in the

different PSP subtypes, and their position in the event ordering relative to downstream atrophy events.

### 3.4.2 Patient staging

This EBM demonstrates that there is significant variability in terms of the stage of PSP-RS patients at baseline (**Figure 3.3**) and provides an intrinsic staging mechanism by which to stratify patients more accurately in terms of their temporal position in the disease course. This is supported by the association between EBM stage and disease duration (both at all timepoints and only at baseline) in those subjects for which disease duration was recorded (**Figure 3.6B** and **Figure 3.7B**).

Uncertainty in the model assigned stage is dependent on the degree of overlap between the HC and PSP-RS biomarker distributions, as well as the accuracy of a given person's biomarker measurement<sup>267</sup>. Imaging biomarkers are known to be associated with a high degree of variance, some of which can be explained by different scanners used, the age and gender, and variation in individual TIV. I tried to control for this by regressing these out as covariates. Linear modelling of age against predicted EBM stage for cases and controls (**Figure 3.8A** and **Figure 3.8B**) showed no association, supporting the validity of this approach.

Although the purpose of this study was to identify the sequence of regional atrophy in PSP-RS from cross-sectional data, rather than classify subjects as cases versus controls, using a threshold of stage 2 (medulla and midbrain atrophic) the model was able to correctly classify subjects as PSP-RS versus healthy control with an overall categorisation accuracy of 90%. This accuracy is similar to that seen in other MRI studies using simple group wise comparisons of midbrain volume between cases and controls<sup>382</sup> and gives confidence that the EBM sequence is a valid representation of disease progression. This is further supported by the fact that ninety-six percent of cases either stayed at the same stage or progressed to a higher stage over a twelve-month period. In addition, predicted subject EBM stage is significantly correlated ( $p<0.01$ ) with a validated measure of clinical disease severity (PSP Rating Scale), as well as disease duration ( $p<0.01$ ), demonstrating the clinical relevance of our MRI-based fine-grained staging system. However, unlike a clinical rating score, the EBM also provides insights into the underlying progression of brain volume changes, and given it is probabilistic, a natural way to incorporate uncertainty into the staging.

### 3.4.3 Limitations

There are several assumptions made when building an EBM, which must be considered when interpreting these results. The EBM assumes that all patients have a broadly similar disease progression pattern with a unimodal distribution of orderings. I restricted analysis to those patients with a diagnosis of PSP-RS, to try and exclude some of the heterogeneity in clinical phenotype associated with PSP pathology<sup>98</sup>. Those cases included from the 4RTNI1, Davunetide and SAL / YP cohorts were diagnosed with probable PSP-RS according to the NINDS criteria, though it is possible that at least some of these cases may meet the 2017 diagnostic criteria for non-RS clinical phenotypes. In the PROSPECT study 10% of PSP cases diagnosed under the NINDS criteria were relabelled as a non-RS phenotype when the 2017 MDS criteria were applied<sup>298</sup>. Given the sensitivity of the EBM to sample heterogeneity, and the variation in pathology staging by phenotype<sup>51,362</sup>, investigation of PSP phenotype heterogeneity using subtype and stage inference (SuStaIn)<sup>290</sup> may provide finer grained patient stratification and is worth pursuing.

The EBM staging has no explicit timescale<sup>267</sup>, although it can predict what stage the patient is within the sequence of biomarker abnormalities, it is unable in itself to extract information on the time taken to transition between states. When given longitudinal data the model currently treats repeated measures as if they are independent i.e. from separate individuals, thus losing information on temporal covariance that could further inform on the ordering of events. Recently, a new generative model called the Temporal Event-Based Model (TEBM) has been developed<sup>384</sup> to accommodate longitudinal data, which is able to learn both individual-level trajectories within the sequence of biomarker abnormalities as well as the time to transition between each event. Applied to my dataset the TEBM may provide insights into the transition times between each stage defined by this study.

Although PSP-RS has been shown to be highly correlated with underlying PSP pathology<sup>73</sup>, in rare cases other pathologies such as CBD can present with PSP-RS and imaging is unable to differentiate the underlying pathology<sup>385</sup>. Of the 365 PSP-RS cases selected for image processing, 24/26 (92%) of cases that came to post-mortem had PSP pathology, while one had GGT and the other CBD pathology (these were excluded from the analysis). Although a small sample size this correlation between

PSP-RS and underlying PSP pathology is in keeping with previous studies<sup>73</sup>. In the absence of a sensitive and specific tau-PET ligand, or indeed any other biomarker, for PSP pathology, there is not an easy way around this clinic-pathological disconnect, and until such time the inclusion of patients in clinical trials based on a clinical diagnosis of PSP-RS is likely to continue.

Another limitation, though not unique to this study, is that the MRIs of different patients were acquired across a range of centres and on different scanners. It is well known that scanners can differ from each other in relation to imaging quality, signal homogeneity and image contrast which can lead to bias<sup>265</sup>. Stringent visual quality controls were applied to both the raw images and post segmentation scans, the GIF algorithm bias corrects for field inhomogeneity, and I also controlled for scanner type by introducing it as a covariate in the linear regression. In addition, previous analyses on the davunetide dataset (which had the highest number of different scanners) scanner type showed no significant effect on atrophy rates<sup>386</sup>. Furthermore, the use of different scanners at multiple sites is a realistic scenario for clinical trials in rare diseases such as PSP, and so scanner heterogeneity combined with the large sample size in this study supports stronger generalisability of the findings.

### **3.5 Conclusions and future work**

In this chapter I have uncovered the *in vivo* sequence of brain atrophy in a large series of individuals with PSP-RS using a probabilistic data-driven model of brain volume changes, that mirrors the recent post-mortem brain histopathology staging proposed by Kovacs et al<sup>51</sup>. It provides an objective, *in vivo* staging system that is longitudinally consistent and correlates with clinical measures of disease severity and disease duration. This approach has potential utility to stratify PSP patients on entry into clinical trials based on disease stage, and complement existing clinical outcome measures to track disease progression. Future work should focus firstly on validating this model in an independent cohort of PSP-RS patients, and secondly on applying more advanced models such as Subtype and Stage Inference (SuStaIn) to better understand the clinical heterogeneity of the variant PSP subtypes and whether jointly modelling temporal and phenotypic heterogeneity will provide finer grained patient stratification.

# Chapter 4: Uncovering spatiotemporal patterns of atrophy in progressive supranuclear palsy using unsupervised machine learning

The work in this chapter has been published in *Brain Communications* in March 2023:

Scotton WJ, Shand C, Todd E, Bocchetta M, Cash DM, VandeVrede L, Heuer H, PROSPECT Consortium, 4RTNI Consortium, Young AL, Oxtoby N, Alexander DC, Rowe JB, Morris HR, Boxer A, Rohrer JD, Wijeratne PA. Uncovering spatiotemporal patterns of atrophy in progressive supranuclear palsy using unsupervised machine learning *Brain Commun.* 2023 March 02: fcad048.

## 4.1 Introduction

Many neurodegenerative diseases are complicated by poor clinico-pathological correlation, with the underlying pathology often manifesting as a range of different, and often overlapping, clinical syndromes. Defining disease phenotypes based on common underlying biological mechanisms, as opposed to clinical phenotype, is an important step towards enriching clinical trials with patients that are most likely to benefit from the medicine being investigated, especially as therapeutics increasingly target these biological mechanisms. Recent advances in machine learning have enabled analysis of multidimensional data to classify and stage groups with similar data-driven features (such as spatiotemporal atrophy patterns on MRI)<sup>290</sup> rather than just on common clinical features, providing new tools to tackle the problem of clinical heterogeneity.

Progressive supranuclear palsy (PSP), a neurodegenerative disease defined pathologically by the aggregation and spread of 4-repeat tau protein in neurons, astrocytes, and oligodendrocytes<sup>387</sup>, shows significant differences in severity and neuroanatomical distribution of pathology<sup>51</sup>, resulting in a range of clinical phenotypes involving language, behaviour and movement abnormalities<sup>77</sup>. No effective disease-modifying treatment has yet been proven for PSP, despite increasingly available clinical trials<sup>221,224</sup>. Clinical progression appears to be dependent on progressive spreading of the 4-repeat (4R) tau pathology within the brain, with a recent pathology staging system<sup>51</sup> defining six sequential stages of progression for the most common clinical phenotype PSP-Richardson syndrome (PSP-RS), starting in the pallido-nigro-luysian system and spreading rostrally via striatum and amygdala to the cerebral cortex

(frontal > temporo-parietal > occipital), and caudally to the medulla oblongata, pons and cerebellum<sup>359</sup>. Although the molecular pathogenic basis for clinical variation is still poorly understood, this study suggests that differences in tau burden and different tau cytopathologies may distinguish clinical subtypes.

The Movement Disorder Society (MDS) 2017 PSP diagnostic criteria<sup>98</sup> were introduced to try and account for variant PSP clinical phenotypes (vPSP) and increase the sensitivity and specificity for diagnosis of early PSP pathology. The criteria categorise symptomatology into four clinical domains (ocular motor dysfunction, postural instability, akinesia and cognitive dysfunction), with differing combinations of these symptoms defining a range of clinical PSP syndromes. Although the most common clinical presentation of PSP is Richardson syndrome (PSP-RS), variant clinical syndromes (vPSP) may account for up to 50% of individuals with PSP pathology<sup>77–79</sup>. The vPSP syndromes include subcortical variants (including PSP-parkinsonism [PSP-P] and PSP-primary gait freezing [PSP-PGF]) and cortical variants (PSP-frontal [PSP-F], PSP-corticobasal syndrome [PSP-CBS], and PSP-speech / language syndrome [PSP-SL]). These new criteria have higher sensitivity than the previous NINDS criteria (87.9% vs 45.5%)<sup>102,388</sup>, and the “*suggestive of PSP*” clinical category significantly increases the sensitivity for early identification of patients with PSP pathology<sup>99</sup>. One potential issue with the MDS diagnostic criteria as first defined was that patients can be assigned multiple phenotypes according to clinical symptomatology<sup>389</sup>. The introduction of the MAX (multiple allocation extinction)-rules<sup>99</sup> helps to allocate patients to one phenotype, though its application to clinically diagnosed PSP patients appears to lead to an over representation of PSP-RS versus vPSP syndromes such as PSP-P and PSP-PGF<sup>105,390</sup>. This has important implications for clinical trials given that the subcortical variants (PSP-P and PSP-PGF) have better survival with longer disease durations (PSP-P 9 years, PSP-PGF 13 years vs PSP-RS 6-7 years) and slower rates of disease progression<sup>78,103–109</sup>. Improved quantification of the progression of pathological brain changes across the PSP phenotypic spectrum in living patients will be essential to the success of future therapeutic trials<sup>230,359</sup>.

In **Chapter 3** I use a probabilistic data-driven modelling approach (event-based modelling [EBM]) to characterise the in vivo sequence of brain atrophy in PSP-RS<sup>358</sup>, and show that the order of regional atrophy broadly mirrors the sequential spread of tau pathology proposed by Kovacs et al.<sup>51</sup>. However, the EBM assumes sample



homogeneity, making it unsuitable to investigate the full spectrum of PSP phenotype heterogeneity. Subtype and Stage Inference (SuStaIn), an unsupervised machine learning algorithm<sup>290</sup>, has been developed to identify data-driven disease subtypes with distinct temporal progression patterns, and can do so using only cross-sectional data. The trained model can then be used to subtype and stage new individuals. This ability to disentangle both phenotypic and temporal heterogeneity from cross-sectional biomarkers, distinguishes SuStaIn from traditional approaches that focus on either one or the other. The SuStaIn subtypes account for temporal heterogeneity, enabling more accurate subtype assignment than traditional clustering algorithms. This algorithm has been successfully applied to Alzheimer's disease (AD)<sup>290,291</sup>, Multiple Sclerosis (MS)<sup>294</sup>, as well as in genetic frontotemporal dementia<sup>290,293</sup>, providing important insights into distinct data-driven subtypes of disease progression. This type of disease progression modelling approach is ideally suited to disentangling the clinical and pathological heterogeneity of PSP.

In this chapter I apply the SuStaIn algorithm to cross-sectional MRI data from a large international cohort of clinically diagnosed PSP patients (including PSP-RS and vPSP syndromes), to identify imaging subtypes across the PSP clinical spectrum with distinct sequences of atrophy progression. I validate the observed subtypes and stages using a subset of longitudinal imaging data and then characterise the clinical features of each subtype to gain insight into the relationship between PSP pathology, atrophy patterns and clinical presentation.

Of note, in this chapter I use the term *syndrome* when referring to PSP clinical syndrome/phenotype (as defined in the MDS 2017 diagnostic criteria)<sup>98</sup>, and *subtype* when referring to MRI-based subtypes identified by SuStaIn.

## 4.2 Methods

### 4.2.1 Participants and clinical data collected

Data from individuals with a clinical diagnosis of “possible” or “probable PSP” (including PSP-RS and vPSP syndromes), as per the MDS 2017 PSP diagnostic criteria<sup>98</sup>, were identified from the 4R Tau imaging cohort. The details of this cohort, that I built as part of this PhD, are covered in more detail in Chapter 2 (**4R tauopathy imaging cohort**). In brief, this cohort consisted of PSP cases collected from seven

main sources; the 4R Tauopathy Imaging Initiative Cycle 1 (4RTNI 1; ClinicalTrials.gov: NCT01804452)<sup>259,263</sup>, the 4R Tauopathy Imaging Initiative Cycle 2 (4RTNI 2; ClinicalTrials.gov: NCT02966145), the davunetide randomized control trial (DAV; ClinicalTrials.gov: NCT01056965)<sup>220</sup>, the salsalate clinical trial (SAL; ClinicalTrials.gov: NCT02422485)<sup>297</sup>, the young plasma clinical trial (YP; ClinicalTrials.gov: NCT02460731)<sup>297</sup>, the PROgressive Supranuclear Palsy CorTico-Basal Syndrome Multiple System Atrophy Longitudinal Study (PROSPECT; ClinicalTrials.gov: NCT02778607)<sup>298</sup>, and the University College London Dementia Research Centre (UCL DRC) FTD cohort. Control data were collected from three sources: the Frontotemporal Lobar Degeneration Neuroimaging Initiative dataset (FTLDNI; <http://4rtni-ftldni.ini.usc.edu/>), PROSPECT, and the UCL DRC FTD Cohort. Controls were defined as no known diagnosis of a neurological or neurodegenerative condition and no known history of memory complaints. Further details on the individual cohorts, including recruitment and diagnostic criteria and MRI acquisition protocols are included in **Chapter 2. Table 4.1** summarises the demographics and clinical data for all PSP cases included in this study, broken-down by individual contributing cohort.

**Table 4.1 - PSP clinical phenotypes and baseline characteristics by contributing cohort.**

	Controls	All	4RTNI1	4RTNI2	DAV	YP/SAL	PROSPECT	UCL
Baseline, n (fu visits)	290	426 (367)	59 (83)	100 (58)	173 (157)	13 (24)	52 (25)	29 (20)
- PSP-RS	-	357 (329)	59 (83)	59 (34)	173 (157)	13 (24)	30 (17)	23 (14)
- PSP-C	-	52 (25)	-	41 (24)	-	-	9 (0)	2 (1)
- PSP-SC	-	17 (13)	-	-	-	-	13 (8)	4 (5)
Gender, % female	56%	48%	54%	48%	51%	48%	42%	31%
Age at first scan, y	62.5 (9.4)	68.4 (6.8)	70.2 (7.2)	68.3 (7.0)	67.6 (6.4)	70.0 (3.8)	70.0 (8.1)	67.3 (5.2)
Age at first symptom, y <sup>a</sup>	-	64.1 (7.5)	64.7 (7.6)	63.2 (7.5)	-	-	66.3 (8.0)	62.6 (5.1)
Disease duration, y <sup>a, b</sup>	-	4.5 (3.1)	5.4 (3.9)	4.8 (3.3)	-	-	3.8 (2.4)	4.3 (2.6)
Pathology, n (% PSP)	-	31 (94%)	7 (88%) <sup>c</sup>	-	-	-	6 (100%)	18 (95%) <sup>d</sup>

This table summarises all cases post-QC. Values are mean (SD) apart from Gender % female, Baseline n (n follow-up visits), Pathology n (% PSP). <sup>a</sup> note incomplete data for disease duration / age at first symptom. <sup>b</sup> time from first symptom to first scan. <sup>c</sup> one case Globular Glial Tauopathy pathology. <sup>d</sup> one case CBD pathology. Abbreviations: Abbreviations: 4RTNI1 = 4-repeat tauopathy neuroimaging initiative (Phase 1), 4RTNI2 = 4-repeat tauopathy neuroimaging initiative (Phase 2), DAV = Davunetide trial, , YP = Young plasma trial, SAL = Salsalate trial, PROSPECT = PROgressive Supranuclear Palsy CorTico-Basal Syndrome Multiple System Atrophy Longitudinal Study, UCL = University College London Dementia Research Centre frontotemporal dementia cohort

Appropriate ethical committee approval was applied for and approved via each of the individual trial and research ethics committees. To be included all participants needed to have, as a minimum, a clinical diagnosis of PSP (PSP-RS or vPSP), a baseline T1 volumetric MRI on a 1.5 or 3 tesla scanner, basic demographic data (gender and age at time of scan). Clinical rating scale scores (PSP rating scale, Unified Parkinson Disease Rating Scale [UPDRS], Schwab and England Activities of Daily Living scale [SEADL], and Montreal Cognitive Assessment [MoCA] or Mini-mental State Examination [MMSE] at baseline and follow-up), pathology at autopsy, and follow-up scans were also included if available. As detailed in previous work<sup>358</sup>, original trial analyses failed to show any treatment effect (including no change in volumetric MRI measurements) in the SAL, YP and DAV trials, so data were combined from each study's treatment and placebo arms. Longitudinal data were used to validate the consistency of SuStaIn's subtype and stage assignments at follow-up.

Given the PROSPECT and 4RTNI2 trials only assessed cognitive function using the MOCA (as opposed to the MMSE for the other trials), raw MOCA scores were converted to MMSE scores using the validated method used by Lawton et al<sup>391</sup>. For missing data in clinical scales, I used an adjusted mean score if at least 80% of the assessment was complete, as per Jabbari et al<sup>298</sup>.

## 4.2.2 MRI acquisition and image processing

The MRI acquisition and image processing procedures are described in detail in Chapter 2 (**Image processing pipeline**). Briefly, cortical and subcortical structures were parcellated using the geodesic information flow algorithm (GIF)<sup>304</sup>, which automatically extracts regions based on the Neuromorphometrics atlas (Neuromorphometrics, Inc.), using an atlas propagation and label fusion strategy<sup>363,364</sup>. Subregions of the cerebellum were extracted using GIF based on the Diedrichsen atlas<sup>305</sup>. I subsequently segmented the medulla, pons, superior cerebellar peduncles (SCPs) and midbrain using a customised version of a module available in FreeSurfer to accept the GIF parcellation of the whole brainstem as input<sup>306</sup>. Volumes for 24 grey-matter regions were calculated; four brainstem (medulla, pons, superior cerebellar peduncle [SCP] and midbrain), three cerebellar (cerebellar cortex, dentate nucleus and vermis), nine subcortical (thalamus, globus pallidus (GP), caudate, putamen, ventral diencephalon (DC), thalamus, hippocampus, amygdala and nucleus accumbens [NA])

and eight cortical (basal forebrain, cingulate, frontal anterior, frontal posterior, insula, temporal, parietal and occipital) regions. Regions that had a right and left label were combined. A list of the GIF subregions included in each cortical region are included in **Table 4.2**. I calculated total intracranial volume (TIV) using SPM12 v6225 (Statistical Parametric Mapping, Wellcome Trust Centre for Neuroimaging, London, UK) running under MATLAB R2012b (Math Works, Natick, MA, USA)<sup>365</sup>. I visually inspected all images to ensure accurate segmentation.

**Table 4.2 - List of GIF subregions included in each region used as SuStaIn input.**

<b>Regions included in SuStaIn</b>	<b>GIF Subregions</b>
Frontal Anterior	Frontal operculum, central operculum, frontal pole, gyrus rectus, middle frontal cortex, subcallosal area, superior frontal gyrus medial segment, superior frontal gyrus, middle frontal gyrus, opercular part of the inferior frontal gyrus, orbital part of the inferior frontal gyrus, triangular part of the inferior frontal gyrus, anterior orbital gyrus, medial orbital gyrus, lateral orbital gyrus, posterior orbital gyrus
Frontal Posterior	Precentral gyrus, precentral gyrus medial segment, supplementary motor cortex
Temporal	Entorhinal area, fusiform gyrus, parahippocampal gyrus, inferior temporal gyrus, middle temporal gyrus, superior temporal gyrus, temporal pole, planum polare, planum temporale, transverse temporal gyrus
Parietal	Precuneus, parietal operculum, supramarginal gyrus, superior parietal lobule, angular gyrus, postcentral gyrus, postcentral gyrus medial segment
Occipital	Cuneus, calcarine cortex, lingual gyrus, occipital fusiform gyrus, superior occipital gyrus, inferior occipital gyrus, middle occipital gyrus, occipital pole
Insula	Anterior insula, posterior insula
Amygdala	Amygdala
Cingulate	Anterior cingulate gyrus, middle cingulate gyrus, posterior cingulate gyrus
Medulla	Medulla
Pons	Pons
Superior Cerebellar Peduncles	Superior cerebellar peduncles
Midbrain	Midbrain
Ventral Diencephalon	Ventral Diencephalon (GIF segmentation includes subthalamic nucleus, substantia nigra and hypothalamus)
Thalamus	Thalamus
Nucleus Accumbens	Nucleus accumbens
Globus Pallidus	Globus Pallidus
Putamen	Putamen
Caudate	Caudate
Cerebellar Cortex	Lobules I/IV, V, VI, VIIA-Crus I, VIIA-Crus II, VIIB, VIIIA, VIIB, IX, X
Dentate	Dentate nucleus

All volumes were corrected for TIV, scanner field strength (1.5T or 3T), scanner manufacturer, age at baseline scan and sex, using linear regression on the control population and then propagating this model to the PSP population, to generate covariate-adjusted regional volumes. To confirm that age effects on regional brain volumes had been successfully regressed out, linear models were fit to assess for any residual association between individual covariate adjusted regional volumes and age at scan.

I defined a biomarker in this study as an image-based regional volume extracted using GIF. To reduce dimensionality, I carried out pairwise comparisons between healthy volunteers and patients at baseline visit, and selected MRI regions whose differences between groups were associated with a moderate to large effect size (Cohen's  $d$  effect size of  $\geq 0.6$  for standardized mean differences between the cases and controls). Applying this threshold to the regional volumes segmented by GIF (see Chapter 2 **Image processing pipeline**) resulted in selection of 20 regions of interest (ROI) that were then included for downstream analysis (**Table 4.3**). Adjusted regional volumes for the ROI were converted into z scores relative to the control group by subtracting the mean of the control group from each patient's ROI volume and dividing by the standard deviation of the control group. Given regional brain volumes decrease with disease progression, the z scores become negative as the disease progresses; z scores were therefore multiplied by -1, to give positive z scores that increase with disease progression. I then used the z scored data as input to SuStaIn.

**Table 4.3 - Effect size by region of interest.**

<b>Region of Interest</b>	<b>Cohen's <i>d</i></b>
Midbrain	2.56
Ventral diencephalon	1.97
SCP	1.73
Pons	1.72
Thalamus	1.67
GP	1.58
Dentate	1.53
Putamen	1.51
Medulla	1.50
Insula	1.27
Frontal Posterior	1.21
Frontal Anterior	1.10
NA	0.87
Parietal	0.85
Cerebellar cortex	0.79
Caudate	0.78
Amygdala	0.74
Cingulate	0.73
Occipital	0.62
Temporal	0.60
Corpus callosum	0.57 <sup>a</sup>
Vermis	0.53 <sup>a</sup>
Basal forebrain	0.41 <sup>a</sup>
Hippocampus	0.36 <sup>a</sup>

Cohen's *d* calculated as the standardised mean difference between adjusted the regional volume of that that region of interest (ROI) in cases vs controls. <sup>a</sup> a Cohen's *d* threshold of greater than or equal to 0.6 was used to select ROI as input for SuStaIn algorithm resulting in 20 biomarkers being included in model.



### 4.2.3 Subtype and Stage Inference

SuStaIn is a probabilistic machine learning algorithm that simultaneously clusters individuals into groups (subtypes) and infers a trajectory of change associated with each group; that trajectory defines the disease stage (degree of disease progression within a subtype) of each individual within the corresponding group. SuStaIn requires only cross-sectional data as input, although can exploit longitudinal data for training if available. I have summarised the main details of the algorithm in Chapter 2 (**Subtype and Stage Inference (SuStaIn)**). In summary, each subtypes' progression pattern is described using a piecewise linear z score model, expressing a trajectory with a series of stages, that each correspond to a single biomarker (regional brain volume in this case) reaching a new z score. Importantly, the number of SuStaIn stages is determined by the number of biomarkers (the product of the number of ROIs and number of z score thresholds per ROI) provided as input. SuStaIn optimises both the subtype membership and the ordering in which different biomarkers reach different z-scores in each subtype (for example one, two or three standard deviations away from the control mean for that ROI) using a data likelihood function. SuStaIn has been applied to a range of different neurodegenerative diseases<sup>290–294</sup>, providing fine-grained patient stratification along both temporal and phenotypic axes, as well as informing on the underlying heterogeneity of disease progression in these different disorders.

Only PSP cases (clinically diagnosed) were used to fit SuStaIn. **Table 4.4** provides a summary of the Z-score settings, MCMC iterations and number of random starting sequences used for the SuStaIn algorithm. I estimated model uncertainty using 100,000 Markov Chain Monte Carlo (MCMC) iterations, and in the single-cluster expectation maximisation procedure I optimised the single-cluster sequence from 24 different random starting sequences to find the maximum likelihood solution. I determined the optimal number of subtypes using information criteria calculated through ten-fold cross-validation (cross-validation information criteria; CVIC), to balance internal model accuracy with model complexity<sup>290</sup>. Where there was no strong evidence for an additional subtype in the model using the CVIC, I assessed the average log-likelihood across folds for the additional subtype and if there was no improvement selected the most parsimonious model i.e. the model with fewer subtypes.

Finally, I used the fitted SuStaIn model to calculate the probability that each individual falls at each stage of each subtype, and assigned individuals to their maximum likelihood stage and subtype based on their baseline scan (as described in Young et al.<sup>290</sup>). I visualised the subtype progression patterns identified by SuStaIn using BrainPainter software<sup>392</sup>, which I modified to include the brainstem segmentations.

**Table 4.4 - SuStaIn algorithm settings for each biomarker.**

<b>Biomarker</b>	<b>R</b>	<b><math>Z_{max}</math></b>
Frontal Anterior	1	4
Frontal Posterior	2	4
Temporal	1	4
Parietal	2	5
Occipital	1	4
Insula	2	4
Amygdala	1	4
Cingulate	1	3
Medulla	3	5
Pons	3	5
Scp	3	5
Midbrain	3	6
Ventral DC	3	6
Thalamus	3	6
NA	2	4
GP	3	5
Putamen	3	4
Caudate	1	4
Cerebellar cortex	1	4
Dentate	3	5

**R** is the number of z-scores included for biomarker  $i$ , and  **$Z_{max}$**  is the max. z-score modelled for biomarker  $i$ . Total number of biomarkers ( $i$ ) for model = 42. Cmax (the maximum number of subtypes fitted) = 4. Model uncertainty was estimated using 100,000 Markov Chain Monte Carlo (MCMC) iterations. In the single-cluster expectation maximisation procedure the single-cluster sequence was optimised from 24 different random starting sequences to find the maximum likelihood solution.

#### 4.2.4 Assigning individuals to subtypes and stages

I computed individuals' stage based on their average stage, weighted by the probability of belonging to each stage of each subtype. Individuals that were assigned to either SuStaIn stage 0 (i.e. no atrophy on imaging compared to controls) or stage 41 (end stage i.e. all ROI maximum atrophy) were labelled “*no subtype*”. All other individuals were labelled as “*subtypable*” and were then assigned to their most probable subtype.

## 4.2.5 Statistical Analyses

For all analyses I stratified PSP cases into PSP-RS, PSP-Subcortical, and PSP-Cortical groups based on their baseline clinical diagnosis. The PSP-Subcortical group includes individuals with PSP-P, and PSP-PGF; the PSP-Cortical group includes cases with PSP-SL, PSP-F and PSP-CBS.

### 4.2.5.1 Clinical phenotype and baseline characteristics

Pairwise comparisons of baseline characteristics were performed between all PSP cases and controls, PSP syndrome (PSP-RS, PSP-Cortical and PSP-Subcortical) vs all PSP cases, and each PSP syndrome against each other, using  $t$ -tests for continuous variables and  $\chi^2$  tests for categorical variables. Statistical significance was reported at a level of  $p < 0.05$ , and at the Bonferroni corrected level of  $p < 0.001$  to correct for multiple comparisons (44 items).

### 4.2.5.2 Association between Subtype assignment and covariates

To assess for any residual association between covariates (TIV, scanner field strength, scanner manufacturer, age at baseline scan, sex, cohort and SuStaIn stage) and SuStaIn subtype, I fitted a logistic regression model to the data using the *lm()* function from the *stats* package (version 3.6.2).

### 4.2.5.3 Subtype characterisation

Overall differences between subtypes were first assessed independently of stage, with individuals classified as “no subtype” (i.e stage 0 or 41) excluded from analysis. To compare whether there were any differences between subtypes, I performed  $t$ -tests for continuous variables and  $\chi^2$  tests for categorical variables (and post-hoc pairwise comparisons for clinical syndrome vs SuStaIn subtype using the *chisq.multicomp()* function from the *RVAideMemoire* R package version 0.9-81-2).

To test for associations between clinical scores (PSP rating scale, UPDRS, SEADL and MMSE) and subtype, I accounted for SuStaIn stage, age, and sex by fitting a linear model (clinical test score ~ subtype + stage + age + sex) for each clinical test score. Statistical significance was reported uncorrected at a level of  $p < 0.05$ , and at the Bonferroni corrected level of  $p < 0.005$  for demographic variables (11 items) and for clinical scores (10 items), to account for multiple comparisons.

To assess average stage by clinical syndrome by SuStaIn subtype, I performed a one-way ANOVA (Mean stage ~ PSP syndrome + Sustain baseline subtype) with the *aov()* function the *stats* package (version 3.6.2). Tukey post-hoc significant differences were then calculated to identify the level of significance.

Finally, I tested for differences in all baseline regional volumes of interest between the different SuStaIn subtypes using two-tailed unpaired *t*-tests, with statistical significance reported at a level of  $p < 0.05$ , both uncorrected for and corrected for multiple comparisons (Bonferroni correction).

All statistical analyses were performed either in R (version 4.0.5) or Python (version 3.7.6).

## **4.2.6 Longitudinal validation**

I used the SuStaIn model fitted on the cross-sectional baseline data to assign maximum likelihood stage and subtype to all follow-up scans (at all time points). These scans were used to validate the stability of subtypes, and to assess stage progression, based on the hypothesis that individuals should remain in the same subtype but should advance to higher stages over time (or at least remain at the same stage). I defined Subtype stability as the proportion of individuals assigned to the same subtype at follow-up(s) or progressed from Stage 0 into a subtype, compared to the total number of individuals. I assessed stage progression by comparing SuStaIn stage at baseline and follow-up(s) for all individuals. Specifically, I calculated the proportion of individuals that advanced or stayed at the same stage at follow-up scan.

## 4.3 Results

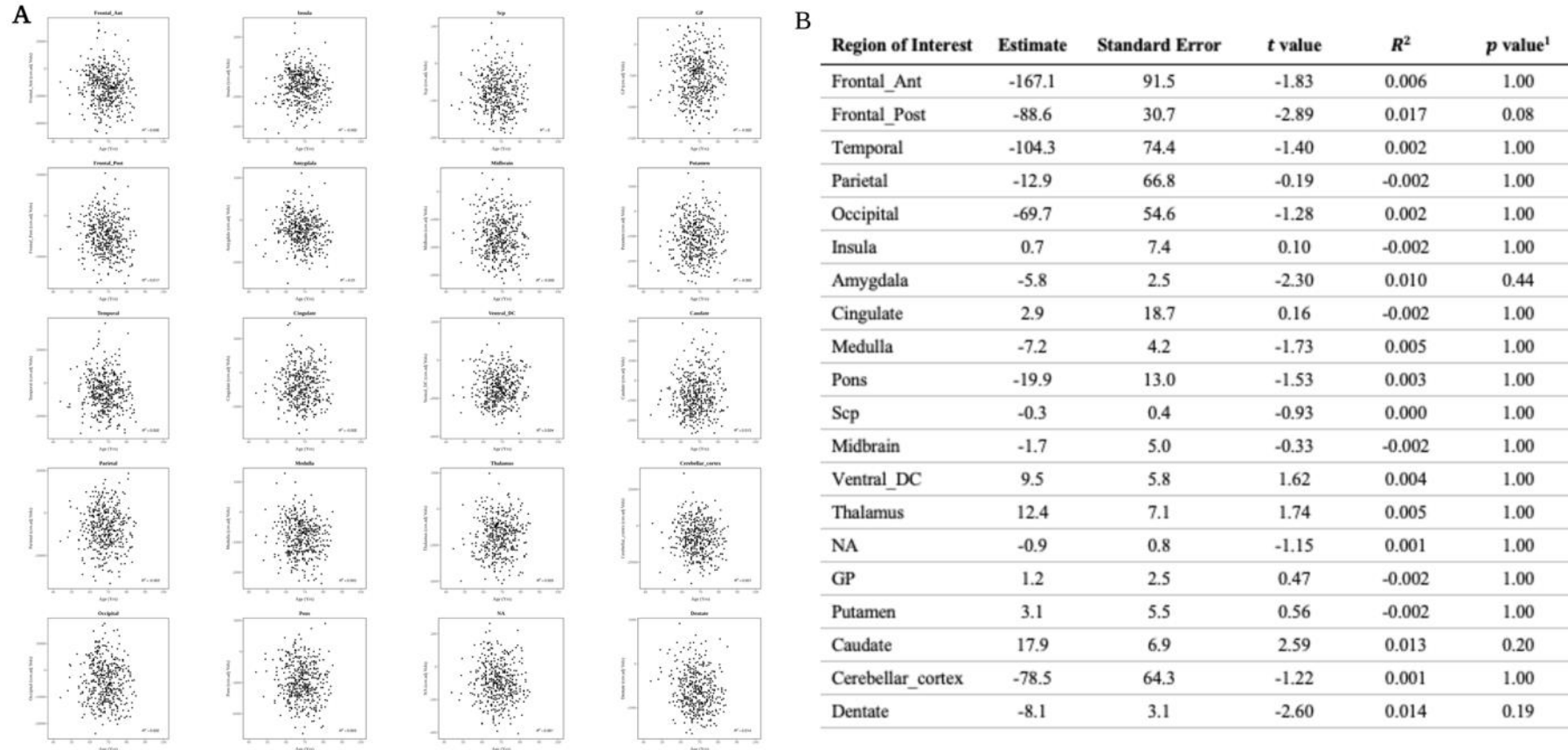
### 4.3.1 Participants

**Table 4.5** summarises the key baseline clinical features for individuals included in the study. For a breakdown of this data by each contributing cohort please refer to **Table 4.1**. I collected a large imaging cohort of PSP cases; 1083 MRI images were included from a total of 716 individuals: 426 with a clinical diagnosis of PSP (with 367 follow-up scans) and 290 controls. Of the PSP cases, 357 (84%) were diagnosed with PSP-RS, 52 (12%) with a PSP-Cortical syndrome (PSP-SL, PSP-F or PSP-CBS), and 17 (4%) with a PSP-Subcortical syndrome (PSP-P or PSP-PGF). 31 (7%) of the PSP cases had a pathological diagnosis after coming to post-mortem, of which 29 (94%) showed 4R tau pathology consistent with PSP, whereas two cases that presented with PSP-RS had non-PSP tau pathology (one corticobasal degeneration [CBD] and one globular glial tauopathy [GGT]). Given the focus of this study was understanding the clinical heterogeneity of clinically diagnosed PSP, both cases were included in the analysis. Overall, the PSP cases at baseline had an older average age compared to controls (68.5 years [ $SD \pm 6.8$ ] vs 62.5 years [ $SD \pm 62.5$ ],  $p < 0.001$ ), though were matched for gender. I confirmed that despite the difference in age between cases and controls, age effects had been effectively regressed out of the regional covariate adjusted volumes for both cases (**Figure 4.1**) and controls (**Figure 4.2**). There were significant differences in baseline clinical scores between the different clinical PSP phenotypes. The highest PSP rating scale score (measure of motor predominant disease burden) was seen in PSP-RS ( $37.2 \pm 13.2$ ), followed by PSP subcortical syndromes ( $26.3 \pm 10.7$ ), with PSP cortical syndromes the least impaired ( $20.6 \pm 20$ ). There was, however, a large variation in this score for the cortical syndromes (PSP-SL  $11.9 \pm 10.8$ , PSP-F  $19.5 \pm 13.2$ , and PSP-CBS  $56.4 \pm 11.9$ ). In keeping with the increased motor predominant disease burden (higher PSP rating scale score) in the PSP-RS and PSP-Subcortical cases, the UPDRS was significantly higher in these cases vs PSP-Cortical cases (PSP-Subcortical  $36.2 \pm 12.9$ , PSP-RS  $29.3 \pm 14$ , PSP-Cortical  $18.8 \pm 17.6$ ,  $p < 0.05$  for each comparison). The PSP-Subcortical cases had a better MMSE score on average compared to the PSP-RS and PSP-Cortical syndromes ( $22.4 \pm 1.2$  vs  $25.7 \pm 3.7$  and  $24.7 \pm 4.6$ ,  $p < 0.001$  for each comparison). There was no significant difference in MMSE between the latter two syndromes.

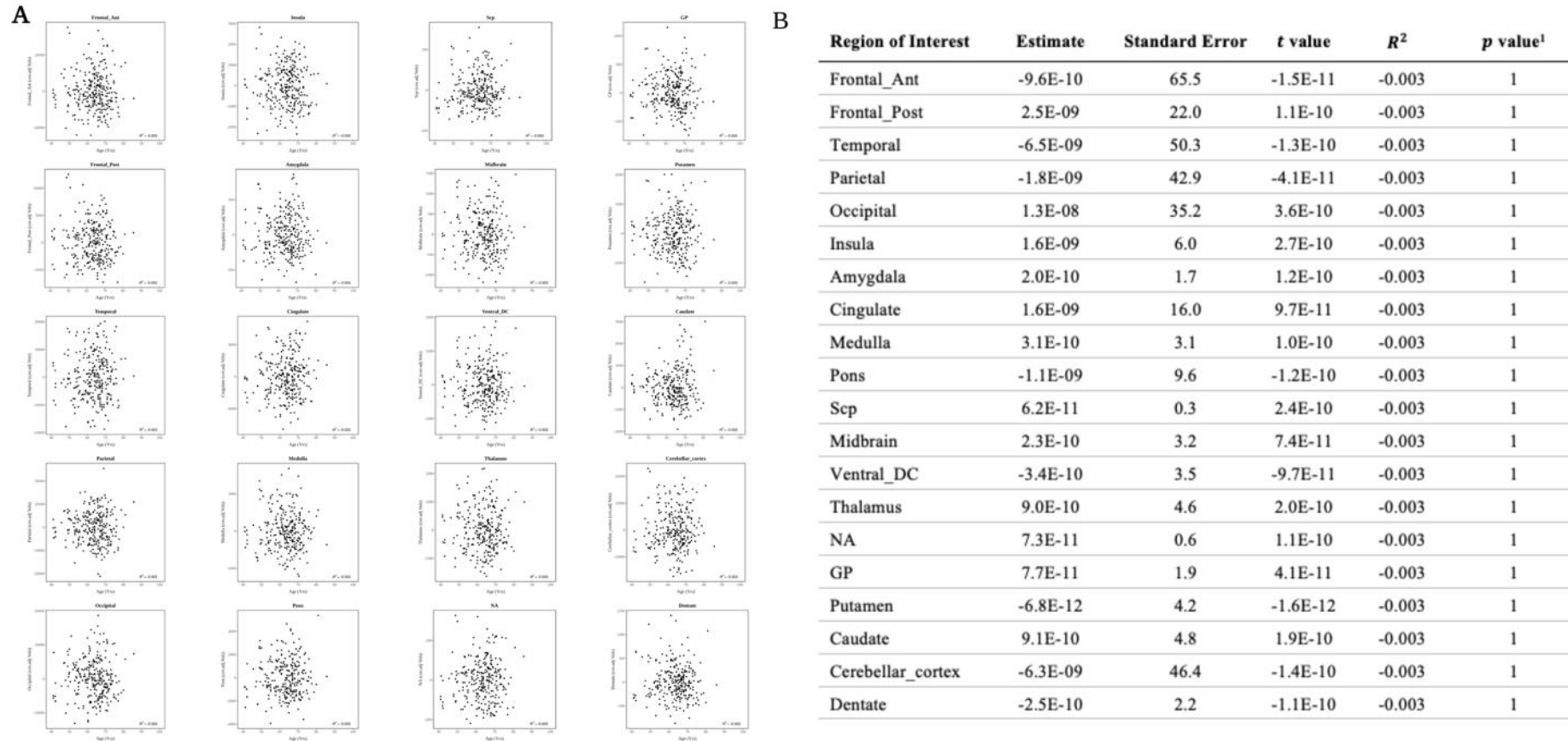
**Table 4.5 - Clinical and baseline characteristics by clinical diagnosis**

	Study Group									
	PSP by Subgroup			PSP-Cortical				PSP-Subcortical		
	Controls	All	PSP-RS	PSP-SL	PSP-F	PSP-CBS	All Cortical	PSP-P	PSP-PGF	All Subcortical
Baseline, n (fu visits)	290	426 (367)	357 (329)	35 (23)	10 (2)	7 (0)	52 (25)	7 (8)	10 (5)	17 (13)
Sex, % female	57%	49%	49%	46%	40%	71%	48%	14%	60%	41%
Age first scan, y	62.5 (9.4)	68.5 (6.8) <sup>d</sup>	68.1 (6.6)	68.6 (7.7)	68.8 (8.6)	77.6 (2.7)	69.9 (7.9)	71.5 (7.7)	71.6 (7.1)	71.5 (7.1)
Age at first symptom, y <sup>a</sup>	-	64.1 (7.5)	63.5 (7.2)	64.1 (8.1)	64.6 (7.7)	72.7 (3.8)	65.4 (8.0)	64.5 (10.3)	66.6 (5.4)	65.7 (7.6)
Disease duration, y <sup>a, b</sup>	-	4.7 (3.2)	4.6 (3.3)	4.5 (2.8)	4.19 (3.0)	4.9 (2.0)	4.5 (2.7)	6.9 (3.5)	5.1 (2.9)	5.8 (3.2)
Pathology, % PSP	-	31 (94%) <sup>c</sup>	26 (92%)	-	1 (100%)	1 (100%)	2 (100%)	4 (100%)	1 (100%)	5 (100%)
PSP rating scale score	-	34.9 (15.0)	37.2 (13.2) <sup>e,h,i</sup>	11.9 (10.8)	19.5 (13.2)	56.4 (11.9)	20.6 (20.0) <sup>f, h</sup>	33.8 (7.89)	23 (10.4)	26.3 (10.7) <sup>e,i</sup>
- History	-	7.8 (3.9)	8.3 (3.6) <sup>h,i</sup>	3.4 (3.3)	5 (2.8)	10.7 (3.4)	4.8 (4.2) <sup>f, h</sup>	8.5 (3.9)	4.4 (3.1)	5.7 (3.7) <sup>i</sup>
- Mentation	-	3.4 (2.6)	3.5 (2.3) <sup>i</sup>	1.8 (1.7)	5.2 (4.2)	4.8 (2.6)	2.8 (2.2)	4 (2.6)	1.0 (0.9)	1.9 (1.3) <sup>e,i</sup>
- Bulbar	-	2.5 (1.8)	2.7 (1.8) <sup>h</sup>	1.8 (1.7)	1.2 (1.4)	2.3 (1.9)	1.8 (1.7) <sup>e, h</sup>	2.7 (2.5)	1.6 (1.1)	1.9 (1.7)
- Ocular motor	-	7.6 (4.0)	8.4 (3.5) <sup>e,h,j,l</sup>	1.3 (0.6)	2.4 (1.9)	10.4 (2.4)	2.9 (2.2) <sup>f,h,j,l</sup>	5.8 (3.3)	4.6 (3.5)	4.9 (3.4) <sup>e,h,j,l</sup>
- Limb motor	-	4.3 (2.9)	4.5 (2.7) <sup>i</sup>	2.4 (2.2)	1.2 (1.0)	10.9 (2.9)	3.5 (3.2)	3.5 (1.3)	2.9 (1.4)	3.1 (1.3) <sup>e,i</sup>
- Gait and midline	-	9.2 (5.2)	9.9 (4.7) <sup>h</sup>	2.2 (2.1)	3.6 (3.2)	17.3 (3.0)	4.8 (3.5) <sup>f,h,k</sup>	9.5 (2.1)	8.6 (3.5)	8.9 (3.1) <sup>k</sup>
SEADL	-	54.8 (24.8)	53.8 (24.0) <sup>i</sup>	66.2 (33.1)	51.4 (13.5)	26.6 (13.5)	58.5 (30.2)	57.5 (20.6)	68.9 (14.5)	65.4 (16.6) <sup>e,i</sup>
UPDRS	-	27.3 (18.0)	29.3 (14.7) <sup>g</sup>	9.3 (8.9)	11.5 (10.9)	65.6 (12.2)	18.8 (17.6) <sup>g,k</sup>	35.5 (9.3)	36.4 (14.7)	36.2 (12.9) <sup>e,k</sup>
MMSE	-	25.5 (3.8)	25.7 (3.7) <sup>j</sup>	27 (2.6)	21.9 (5.3)	19.4 (4.0)	24.7 (4.6) <sup>l</sup>	22 (0.8)	22.6 (1.3)	22.4 (1.2) <sup>f,j,l</sup>

Values are mean (SD), apart from Gender % female, Baseline n (n follow-up visits), Pathology n (% PSP). Pairwise comparisons between groups were performed using *t* tests for continuous variables and  $\chi^2$  tests for categorical variables. a note incomplete data for disease duration / age at first symptom. b time from first symptom to first scan. c 2 cases not PSP pathology (1 CBD, 1 GGT). d PSP all vs Controls. Statistically significant at  $p < 0.05$ , corrected for multiple comparisons (44 comparison,  $p < 0.00114$ ). e PSP [subgroup] vs PSP All. Statistically significant at  $p < 0.05$ , uncorrected for multiple comparisons. f PSP [subgroup] vs PSP All. Statistically significant at  $p < 0.05$ , corrected for multiple comparisons (44 comparison,  $p < 0.00114$ ). g PSP-RS vs PSP-C. Statistically significant at  $p < 0.05$ , uncorrected for multiple comparisons. h PSP-RS vs PSP-C. Statistically significant at  $p < 0.05$ , corrected for multiple comparisons (44 comparison,  $p < 0.00114$ ). i PSP-RS vs PSP-SC. Statistically significant at  $p < 0.05$ , uncorrected for multiple comparisons. j PSP-RS vs PSP-SC. Statistically significant at  $p < 0.05$ , corrected for multiple comparisons (44 comparison,  $p < 0.00114$ ). k PSP-C vs PSP-SC. Statistically significant at  $p < 0.05$ , uncorrected for multiple comparisons. l PSP-C vs PSP-SC. Statistically significant at  $p < 0.05$ , corrected for multiple comparisons (44 comparison,  $p < 0.00114$ ). Abbreviations: CBD = cortico-basal degeneration, GGT = Globular Glial Tauopathy, MMSE = Mini-Mental State Examination, PSP-C = PSP-cortical (includes PSP-frontal, PSP-speech/language disorder and PSP-corticobasal syndrome), PSP-RS = PSP Richardson syndrome, PSP-SC = PSP subcortical (includes PSP parkinsonism and PSP-progressive gait freezing), SEADL = Schwab and England Activities of Daily Living, UPDRS = Unified Parkinson's Disease Rating Scale



**Figure 4.1 - Association of age at baseline scan with covariate adjusted regional volumes in cases.** (A) Each scatter plot (20 regions in total) shows every PSP case (n=426) plotted as a function of age at baseline scan (x axis) and covariate adjusted regional volume (y axis). (B) Table summarising results of fitting linear models for each region of interest covariate adjusted volume by age at scan (linear model = cov. adj. regional vol. ~ age at baseline scan). The  $R^2$  represents the proportion of the variation in the covariate adjusted regional volume that is explained by age at baseline scan.<sup>1</sup> p values were Bonferroni corrected for multiple comparisons.



**Figure 4.2 - Association of age at baseline scan with covariate adjusted regional volumes in controls.** Each scatter plot (20 regions in total) shows every PSP case (n=426) plotted as a function of age at baseline scan (x axis) and covariate adjusted regional volume (y axis). **(B)** Table summarising results of fitting linear models for each region of interest covariate adjusted volume by age at scan (linear model = cov. adj. regional vol. ~ age at baseline scan). The  $R^2$  represents the proportion of the variation in the covariate adjusted regional volume that is explained by age at baseline scan. <sup>1</sup>p values were Bonferroni corrected for multiple comparisons



### 4.3.2 Spatiotemporal subtypes of PSP

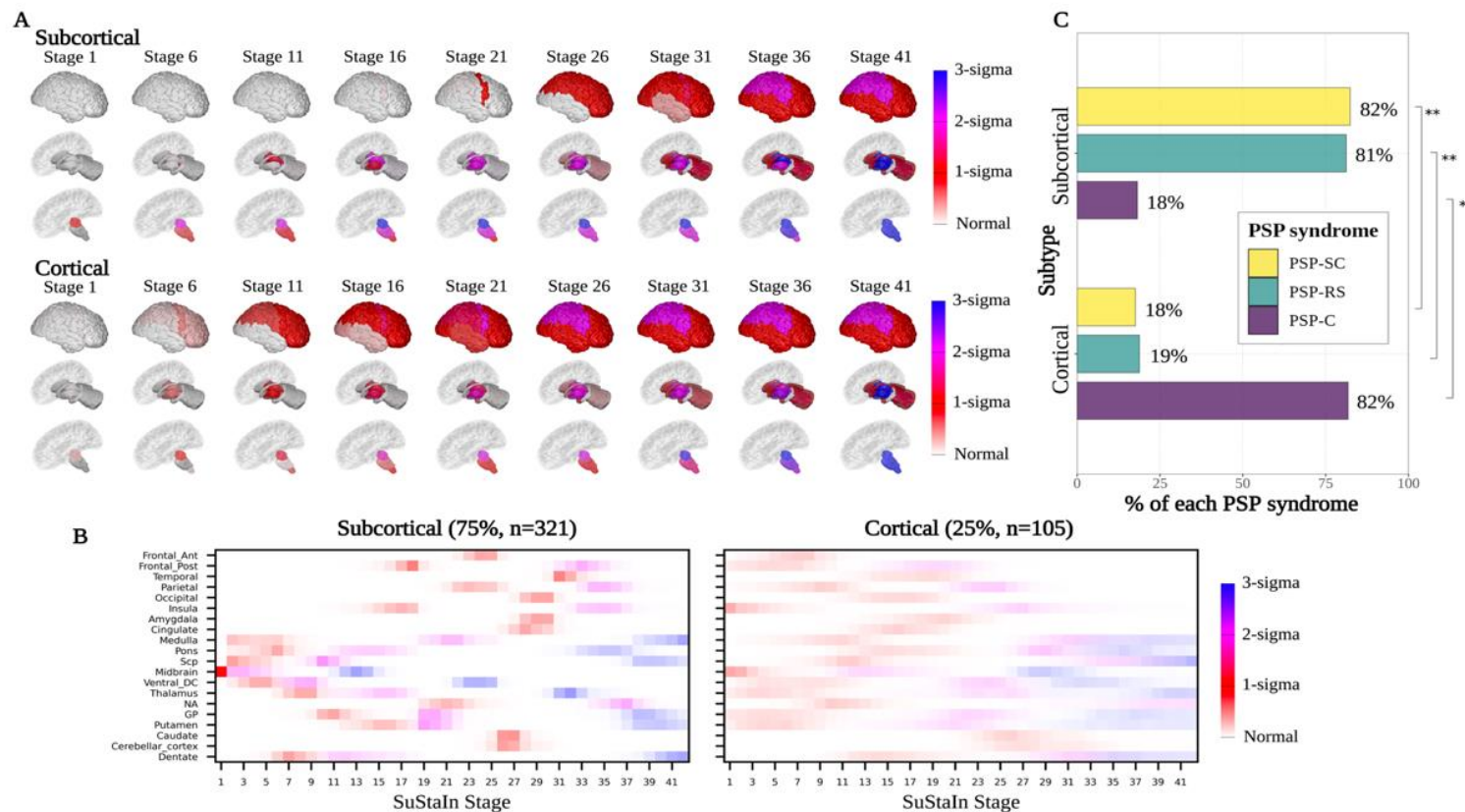
I fitted SuStaIn using PSP cases only, based on the rationale that PSP is a rare disease, and it is very unlikely for our cohort of controls to have asymptomatic PSP. Indeed, it is more likely the controls would have a more common neurodegenerative disorder such as Alzheimer’s disease rather than PSP, and I did not want this to confound the subtype and stage inference estimation hence the exclusion.

SuStaIn identified two imaging subtypes with distinct patterns of regional atrophy evolution (**Figure 4.3A** and **Figure 4.3B** for positional variance diagrams [PVDs]). **Figure 4.4** shows the log likelihoods after 10-fold cross validation with the associated CVIC demonstrating that the two-subtype model was the most parsimonious. Based on the earliest MRI abnormalities seen in the SuStaIn defined trajectories, I labelled the first the “*Subcortical*” subtype and the second the “*Cortical*” subtype. The *Subcortical* subtype (75% of the cases) has atrophy in the midbrain followed by the other brainstem structures (medulla, pons, and superior cerebellar peduncles [SCP]), and the ventral diencephalon (ventral DC) at early SuStaIn stages. The atrophy then progresses caudally to the dentate nucleus of the cerebellum, and rostrally to the thalamus and lentiform nucleus (globus pallidus [GP] and putamen) before spreading to the cortex (after Stage 13). Cortical atrophy progresses in an anterior to posterior direction, beginning in the insula and posterior frontal lobe, before spreading to the temporal, parietal and finally the occipital lobe. The *Cortical* subtype (25% of cases) has more generalised atrophy in the early SuStaIn stages, with the midbrain and insula affected first, then the frontal lobes (posterior > anterior), thalamus, ventral DC and the basal ganglia all affected at a similar time (before stage 13). Interestingly the end stage atrophy pattern is similar for both subtypes.

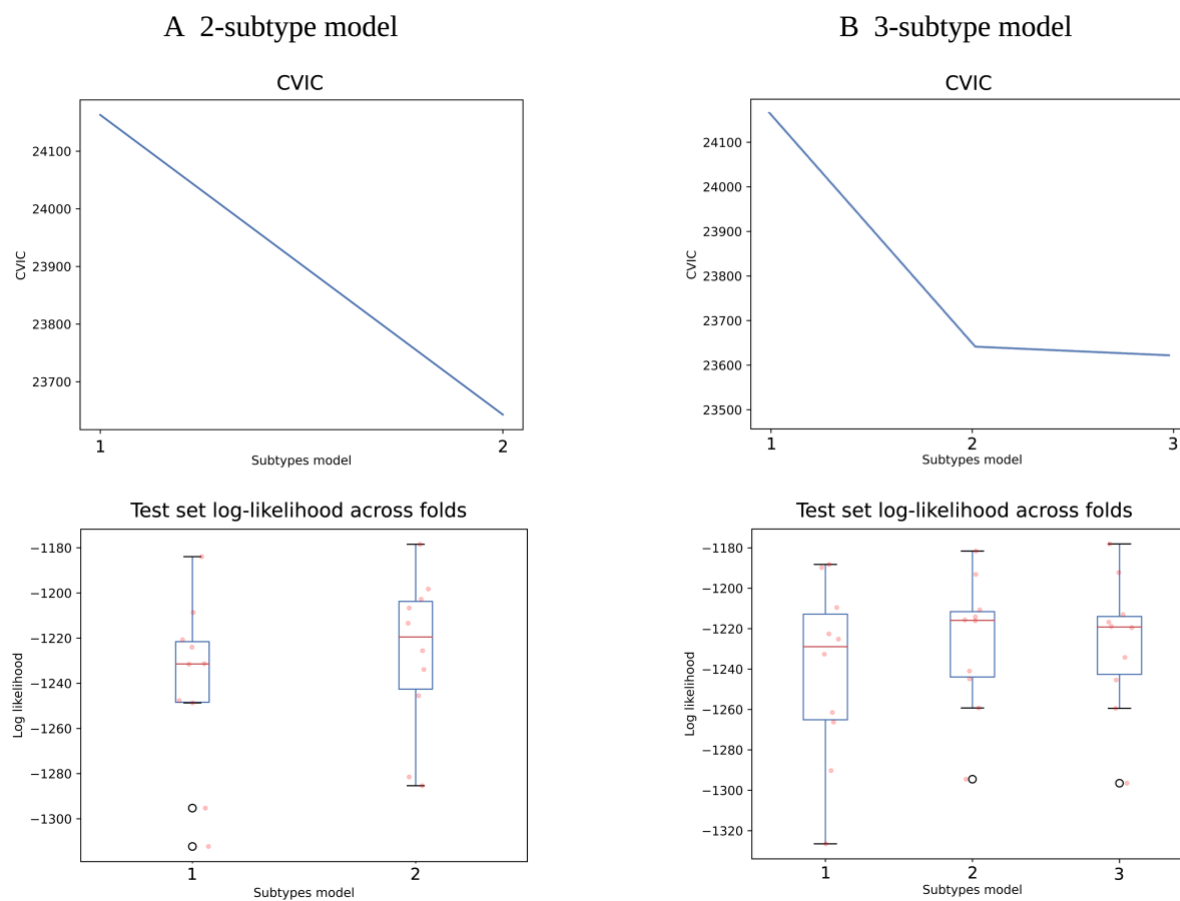
Overall, 20 of the 426 scans (5%) were not subtypable at baseline, and so were excluded from subtype post-hoc analyses. Three of these individuals had a clinical diagnosis of PSP-RS and were at Stage 41, and 17 were at Stage 0 (9 PSP-RS and 8 PSP-Cortical [all PSP-SL]).

I fitted a logistic regression model to assess for any residual association between SuStaIn subtype, regressed covariates, and SuStaIn stage (SuStaIn subtype ~ SuStaIn stage + TIV + age at first scan + sex + scanner + scanner field strength + scanner manufacturer). This demonstrated a remaining association between SuStaIn subtype

and age at first scan ( $z = 2.8$ ,  $p = 6 \times 10^{-3}$ ), General Electric 3T scanner ( $z = -3.0$ ,  $p = 3 \times 10^{-3}$ ), Phillips 3T scanner ( $z = -2.5$ ,  $p = 0.01$ ), and the 4RTNI2 cohort ( $z = 3.6$ ,  $p = 3 \times 10^{-4}$ ). There was no dependency of subtype on stage ( $z = -0.1$ ,  $p = 0.91$ ) with a similar distribution of stages across each subtype (**Figure 4.5**).

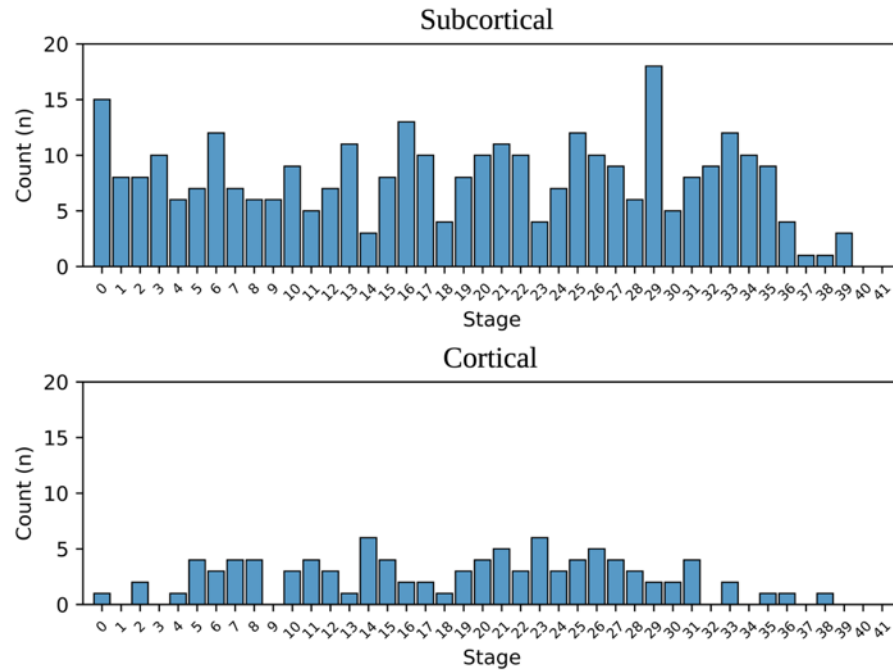


**Figure 4.3 - Subtype progression patterns of PSP atrophy identified by Subtype and Stage Inference (SuStaIn).** (A) Spatial distribution and severity of atrophy at each SuStaIn stage by Subtype. Each row (Subcortical top, Cortical bottom) represents a subtype progression pattern identified by SuStaIn consisting of a set of stages at which brain volumes in PSP cases reach different z-scores relative to controls. (B) Assignment of PSP clinical syndromes to each SuStaIn subtype. Size of bar (x-axis) represents percentage of cases labelled with that PSP syndrome assigned to that SuStaIn subtype (y-axis). A Pearson Chi Square test was performed with post-hoc pairwise comparisons for clinical syndrome vs SuStaIn subtype using the *chisq.multcomp()* function from the *RVaideMemoire* R package version 0.9-81-2 ( $\chi^2$  (2, N = 406) = 81.8,  $p = 2.2 \times 10^{-16}$ ). (C) Positional variance diagrams for SuStaIn subtypes. These represent the uncertainty in the subtype progression patterns for each region. Each region (y-axis) is shaded based on the probability a particular z score is reached at a particular SuStaIn stage (x-axis). Z scores range from zero (white), one (red), two (pink) to three (blue). Abbreviations: Frontal\_Ant = anterior frontal lobe, Frontal\_Post = posterior frontal lobe (supplementary motor cortex and pre-central gyrus), SCP = superior cerebellar peduncle, NA = nucleus accumbens, GP = globus pallidus. \*Statistically significant at  $p < 0.05$ , uncorrected for multiple comparisons. \*\*Statistically significant at  $p < 0.05$ , corrected for multiple comparisons. Visualisations in (A) were generated using the BrainPainter software<sup>392</sup>, modified to include the brainstem segmentations.

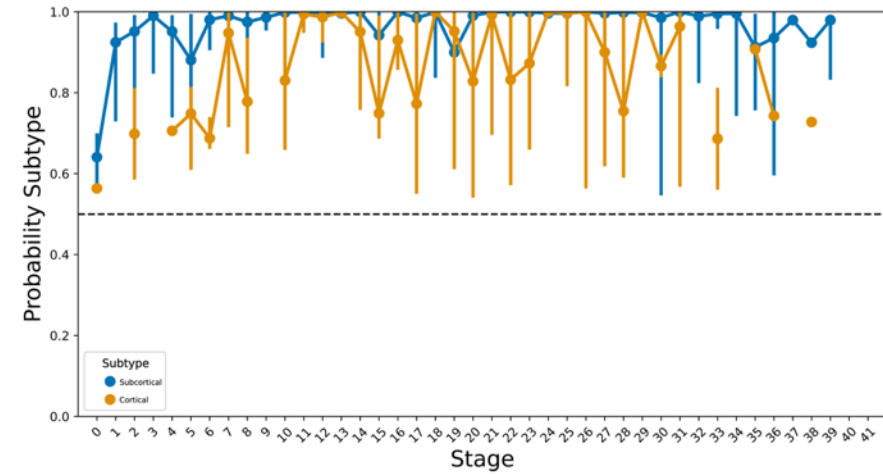


**Figure 4.4 - Selecting optimal SuStaIn subtype model given data.** Top row shows the cross-validation information criterion (CVIC) plots for (A) the 2-subtype model and (B) the 3-subtype model. The bottom row shows the log-likelihood across 10 CV folds for each model. The CVIC for the 3-subtype model for the 3-subtype model is very similar to the 2-subtype model and so for parsimony I selected the 2-subtype model as the best description of the data.

### A. Stage distribution by subtype



### B. Subtype probability by stage



**Figure 4.5 - Stage distribution and Subtype probability of two-subtype model** (A) Stage distribution by Subtype for subtypable baseline scans (B) Subtype assignment probability by stage for subtypable baseline scans. Dots = median probability, vertical-coloured lines = 95% confidence intervals. The horizontal dotted line represents 50% assignment probability. If any stage had less than 50% probability of assignment to subtype would raise concern that a different underlying disease sequence had been appended to another to make one sequence by SuStaIn.

### 4.3.3 Association between PSP clinical syndromes and subtype

I compared the subtype assignments (Subcortical vs Cortical) for clinical PSP syndromes (PSP-RS vs PSP-Cortical and PSP-Subcortical). **Figure 4.3** (and **Table 4.6**) show the percentage of each of these clinical syndromes assigned to each subtype. The *Subcortical* SuStaIn subtype was significantly enriched for PSP-RS and PSP-Subcortical syndromes; 81% of PSP-RS cases ( $p = 2 \times 10^{-6}$ ) and 82% of PSP-Subcortical cases ( $p = 0.007$ ) were assigned to the SuStaIn *Subcortical* subtype respectively. The *Cortical* SuStaIn subtype was enriched for the PSP-Cortical syndromes; 81% of PSP Cortical syndromes ( $p = 2 \times 10^{-5}$ ) were assigned to the SuStaIn *Cortical* subtype.

**Table 4.6 - Comparison of demographics, clinical diagnosis and test scores between SuStaIn subtypes.**

	Subcortical subtype	Cortical subtype	<i>p</i> value
All scans, n	321 (75.4)	105 (24.6)	-
Subtypable scans, n	302 (74.4)	104 (25.6)	0.07 <sup>a</sup>
Average subtype probability <sup>b</sup>	0.94 (0.1)	0.85 (0.2)	<0.005 <sup>c</sup>
Sex, % female	49%	48%	0.82
Age first scan, y	68.1 (6.3)	70.0 (8.1)	0.02
Age at first symptom, y <sup>d</sup>	63.8 (6.9)	65.5 (8.3)	0.14
Disease duration, y <sup>d,e</sup>	4.4 (3.1)	5.0 (3.2)	0.26
PSP syndrome, n			<0.005 <sup>c</sup>
- PSP-RS	280 (81%)	65 (19%)	-
- PSP-C	8 (18%)	36 (82%)	-
- PSP-SC	14 (82%)	3 (18%)	<0.005 <sup>c</sup>
PSP rating scale	37.0 (13.6)	30.3 (16.9)	<0.005 <sup>c</sup>
SEADL	53.7 (23.8)	56.3 (26.6)	0.39
UPDRS	30.0 (16.2)	22.8 (21.0)	<0.005 <sup>c</sup>
MMSE	25.5 (3.6)	24.9 (4.2)	0.20

Values are mean (SD) or n (%), apart from Sex = % female. Pairwise comparisons between groups were performed using *t*-tests for continuous variables and  $\chi^2$  tests for categorical variables. <sup>a</sup> All scans vs. subtypable scans. <sup>b</sup> Subtype probability = the probability of assignment for an individual case to given subtype. <sup>c</sup> Statistically significant at  $p < 0.05$ , corrected for multiple comparisons (11 comparisons,  $p$  value < 0.005). <sup>d</sup> note incomplete data for disease duration / age at first symptom. <sup>e</sup> time from first symptom to first scan. Abbreviations: PSP-RS = PSP Richardson syndrome, PSP-SC = PSP-subcortical (includes PSP-parkinsonism and PSP-progressive gait freezing), PSP-C = PSP-cortical (includes PSP-frontal, PSP-predominant speech/language disorder and PSP-predominant corticobasal syndrome), SEADL = Schwab and England Activities of Daily Living, UPDRS = Unified Parkinson's Disease Rating Scale, MMSE = Mini-Mental State Examination

#### 4.3.4 Subtype demographics and clinical characteristics

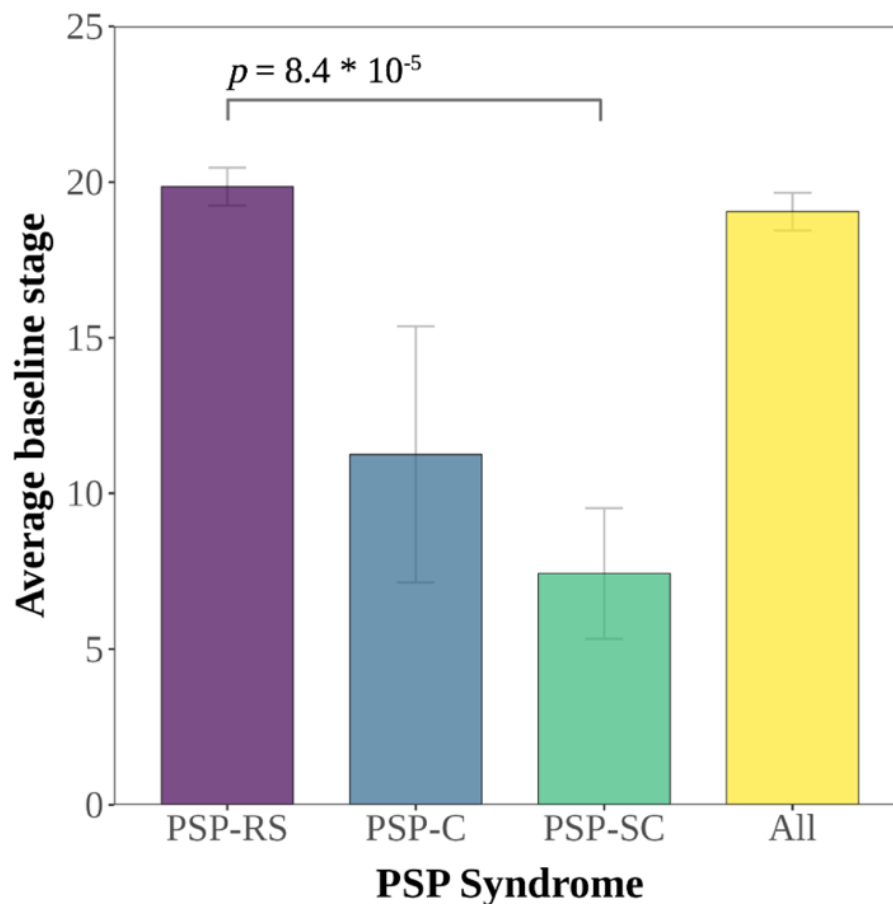
**Table 4.6** gives an overview of demographics, clinical diagnosis, and test scores by subtype. 74% of the subtypable scans were assigned to the *Subcortical* subtype, with an average subtype probability assignment of 0.94 compared to 0.85 for the *Cortical* subtype ( $t = -6.5$ ,  $p < 0.005$ ). Those in the *Subcortical* subtype were both slightly younger at symptom onset (63.8 [SD  $\pm$  6.9] years vs 65.5 [SD  $\pm$  8.3]), and at time of baseline scan (68.1 [SD  $\pm$  6.3] years vs 70.0 [SD  $\pm$  8.1]), though this did not reach statistical significance in either case. Pairwise comparisons of clinical scores, demonstrated that PSP rating scale score (37.0 [SD  $\pm$  13.6] vs 30.3 [SD  $\pm$  16.9],  $t = -3.7$ ,  $p < 0.005$ ) and UPDRS (30.0 [SD  $\pm$  16.2] vs 22.8 [SD  $\pm$  21.0]) were higher (i.e. more severe motor predominant disease burden) in the *Subcortical* subtype. Average MMSE was similar between subtypes (*Subcortical* subtype 25.5 [SD  $\pm$  3.6] vs *Cortical* subtype 24.9 [SD  $\pm$  4.2],  $t = -1.3$ ,  $p < 0.20$ ).

The average stage for subtypable individuals within each subtype was similar (19.0 [SD  $\pm$  10.5] for *Subcortical* vs 18.3 [SD  $\pm$  9.1] for *Cortical*,  $\beta = 8 \times 10^{-6}$ ,  $p = 0.85$ ) (**Table 4.7**). However, PSP-Subcortical cases (82%) assigned to the *Subcortical* SuStaIn subtype were on average at a lower stage (7.4 [SD  $\pm$  5.8]) compared to PSP-RS cases assigned to either the *Subcortical* (19.9 [SD  $\pm$  10.2],  $p < 0.003$ ) (**Figure 4.6**) or the *Cortical* subtype (18.9 [SD  $\pm$  8.7],  $p < 0.003$ ). I then tested whether the PSP-Subcortical and PSP-RS cases assigned to the *Subcortical* subtype showed differences in rate of progression (defined as change in subtype per year). PSP-Subcortical cases in the *Subcortical* subtype progressed on average 0.66 stages per year, compared to 1.86 stages per year for the PSP-RS cases ( $t = 2.49$ , 95% CI 0.1 – 2.4,  $p = 0.046$ ). One PSP-Subcortical case progressed from stage 12 (no cortical involvement) to stage 14 (insula and posterior frontal lobe abnormal), while two cases had more extensive cortical involvement at baseline (stage 16 and stage 26 respectively at baseline, and stage 16 and stage 27 at follow-up).

**Table 4.7- Average stage by clinical syndrome by SuStaIn Subtype**

	Subcortical subtype	Cortical subtype	P value <sup>a</sup>
All	19.0 (10.5)	18.3 (9.1)	0.85
PSP-RS	19.9 (10.2)	18.9 (8.7) <sup>b</sup>	0.98
PSP-Subcortical	7.4 (5.8) <sup>a, b, c</sup>	12.0 (11.3)	0.98
PSP-Cortical	11.3 (8.9)	17.9 (9.6)	0.54

Mean stage (SD) at baseline. One way ANOVA (Mean stage ~ PSP syndrome + Sustain baseline subtype). PSP syndrome  $p = 8.1 \times 10^{-6}$ , Sustain baseline subtype  $p = 0.98$ . <sup>a</sup> Tukey HSD was used for post-hoc multiple comparisons. <sup>b</sup> statistically significant at  $p < 0.05$ , corrected for multiple comparisons. <sup>c</sup> one case PSP-Subcortical assigned to the Subcortical subtype case progressed from stage 12 (confined to subcortical regions to 14 at follow-up (insula and posterior frontal lobe abnormal). Two PSP-Subcortical cases assigned to the Subcortical subtype had more extensive cortical involvement at baseline (stage 16 and stage 26 respectively at baseline, and stage 16 and stage 27 at follow-up). Abbreviations: PSP-RS = PSP Richardson syndrome, PSP subcortical includes PSP-parkinsonism and PSP-progressive gait freezing, PSP-cortical includes PSP-frontal, PSP-speech/language disorder and PSP-corticobasal syndrome.



**Figure 4.6 - Average stage of PSP clinical syndromes assigned to the Subcortical SuStaIn subtype.** Mean SuStaIn stage for PSP syndrome, with associated standard error (SE) bars. ANOVA: PSP syndrome  $F = 12.1$ ,  $p = 8 \times 10^{-6}$ ; SuStaIn subtype;  $F = 1 \times 10^{-4}$ ,  $p = 0.97$ . Tukey post-hoc honest significance test: in Subcortical SuStaIn subtype PSP-RS vs PSP-SC; estimate = -12.4,  $p = 8.4 \times 10^{-5}$ . No other post-hoc comparisons significant. Abbreviations: PSP-RS = PSP Richardson syndrome, PSP-SC = PSP Subcortical (includes PSP-parkinsonism and PSP-progressive gait freezing), PSP-C = PSP-cortical (includes PSP-frontal, PSP-speech/language disorder and PSP-corticobasal syndrome).



I went on to compare the regional unadjusted baseline volumes in the two-subtype model to check whether there were appreciable differences present at baseline scan (**Table 4.8**). This demonstrated that those assigned to the *Subcortical* subtype had significantly lower mean volumes at baseline in the medulla, pons, midbrain, ventral diencephalon and the dentate nucleus. In contrast the *Cortical* subtype had lower mean volumes in the frontal, temporal, parietal, occipital, insula, amygdala, NA, and cingulate at baseline scan. Interestingly, there was no difference in volumes between the two subtypes in the cerebellar vermis and dentate, caudate, hippocampus, amygdala and nucleus accumbens, suggesting that these regions form an early atrophy signature common to both subtypes.

**Table 4.8 - Regional brain volumes by subtype in the 2-subtype model**

Region	Subcortical	Cortical	<i>p</i> value <sup>a</sup>
Frontal Anterior	128935 (13578)	126452 (16259)	2.50 x 10 <sup>-8</sup>
Frontal Posterior	34150 (4371)	32601 (4930)	4.21 x 10 <sup>-9</sup>
Temporal	117607 (10436)	109989 (13026)	4.18 x 10 <sup>-8</sup>
Parietal	85183 (9049)	76286 (10901)	4.56 x 10 <sup>-11</sup>
Occipital	68530 (7469)	63927 (8536)	5.02 x 10 <sup>-7</sup>
Cingulate	26734 (2915)	26268 (3409)	1.09 x 10 <sup>-4</sup>
Insula	9525 (1255)	9193 (1316)	2.06 x 10 <sup>-3</sup>
Corpus callosum	18084 (2796)	17992 (2673)	1.49 x 10 <sup>-5</sup>
Medulla	4597 (646)	4809 (590)	1.93 x 10 <sup>-2</sup>
Pons	12703 (1634)	13984 (1600)	4.72 x 10 <sup>-8</sup>
SCP	196 (45)	230 (47)	5.67 x 10 <sup>-12</sup>
Midbrain	5172 (759)	5778 (703)	4.35 x 10 <sup>-13</sup>
Cerebellar cortex	86575 (8384)	88222 (8993)	1.37x 10 <sup>-7</sup>
Vermis	4359 (532)	4418 (491)	0.11
Dentate	2699 (376)	3023 (399)	0.01
Hippocampus	7361 (768)	7187 (662)	1.00
Amygdala	3301 (339)	3114 (372)	0.20
NA	1051 (111)	1042 (126)	1.00
Caudate	6248 (821)	6177 (901)	1.00
GP	3367 (349)	3383 (373)	1.73 x 10 <sup>-6</sup>
Putamen	7759 (726)	7557 (811)	2.50 x 10 <sup>-8</sup>
Thalamus	9727 (1002)	10110 (926)	4.21 x 10 <sup>-9</sup>
Basal forebrain	982 (98)	975 (106)	4.18 x 10 <sup>-8</sup>
Ventral DC	7837 (857)	8438 (973)	4.56 x 10 <sup>-11</sup>

Values are mean volume (mm<sup>3</sup>) for that brain region (SD). Two-tailed, unpaired *t*-tests performed. <sup>a</sup> statistically significant at *p* < 0.05, corrected for multiple comparisons (Bonferroni). Abbreviations: SCP = superior cerebellar peduncles, DC = diencephalon, GP = globus pallidus, NA = nucleus accumbens

### 4.3.5 Association between stage, subtype, and disease severity

**Table 4.9** shows the relationship between clinical test scores and SuStaIn subtype and stage across all subtypable cases, accounting for age at first scan and sex. Performance on the total PSP rating scale score (and History, Bulbar, Ocular motor, and Gait midline subscores) was worse in the *Subcortical* subtype and related to stage, suggesting that these scores decline with disease progression in both subtypes, but the overall scores are worse in the *Subcortical* subtype. UPDRS score was worse in the *Subcortical* subtype but unrelated to SuStaIn stage, suggesting that test performance shows a stronger decline with disease progression in the *Subcortical* subtype. Worsening SEADL, PSP rating scale Limb motor and Mentation subscores were associated with increasing SuStaIn stage, suggesting these scores decline with disease progression in both subtypes. All these associations survived Bonferroni correction for multiple comparisons. MMSE showed no difference between subtypes and was not associated with SuStaIn stage.

**Table 4.9 - Comparison of adjusted clinical scores between subtypes.**

	SuStaIn subtype		SuStaIn stage		Subtype with worse score	Change with Sustain Stage
	<i>t</i> value	<i>p</i> value	<i>t</i> value	<i>p</i> value		
PSP rating scale score						
- <i>Total</i>	-4.12	5 x 10 <sup>-5b</sup>	5.21	3 x 10 <sup>-7b</sup>	Subcortical subtype	Worsens
- <i>History</i>	-3.98	8 x 10 <sup>-5b</sup>	4.21	3 x 10 <sup>-5b</sup>	Subcortical subtype	Worsens
- <i>Mentation</i>	-0.86	0.39	3.74	2 x 10 <sup>-4b</sup>		Worsens
- <i>Bulbar</i>	-2.17	0.03 <sup>a</sup>	3.42	7 x 10 <sup>-4b</sup>	Subcortical subtype	Worsens
- <i>Ocular motor</i>	-4.56	7 x 10 <sup>-6b</sup>	3.90	1 x 10 <sup>-4b</sup>	Subcortical subtype	Worsens
- <i>Limb motor</i>	-0.40	0.69	3.25	8 x 10 <sup>-6b</sup>		Worsens
- <i>Gait and midline</i>	-4.02	7 x 10 <sup>-5b</sup>	3.12	0.002 <sup>b</sup>	Subcortical subtype	Worsens
SEADL	1.03	0.30	-5.57	5 x 10 <sup>-8b</sup>		Worsens
UPDRS	-2.67	0.009 <sup>b</sup>	1.70	0.08	Subcortical subtype	
MMSE	-1.03	0.31	-1.42	0.16		

Linear model of clinical score ~ subtype + stage + age + sex. <sup>a</sup>. Statistically significant at  $p < 0.05$ , uncorrected for multiple comparisons. <sup>a</sup> statistically significant at  $p < 0.05$ , uncorrected for multiple comparisons. <sup>b</sup>statistically significant at  $p < 0.05$ , corrected for multiple comparisons (10 items,  $p < 0.0125$ ). Abbreviations: SEADL = Schwab and England Activities of Daily Living, UPDRS = Unified Parkinson's Disease Rating Scale, MMSE = Mini-Mental State Examination.

### 4.3.6 Longitudinal consistency of subtypes

Given I used cross-sectional MRI data to infer distinct longitudinal trajectories, evaluating how well longitudinal data fits the SuStaIn model is a key aspect of validation. I tested this in two ways; firstly, by assessing whether subtype assignments were longitudinally stable, and secondly by testing whether individuals progressed to later stages at follow-up. A total of 355 follow-up scans (355/367) were subtypable at follow-up from 289 PSP cases (224 had one follow-up scan, 64 had two and one individual had three). Mean follow-up time was 0.91 years with a SD of 0.38 years.

Overall, the SuStaIn subtype assignments showed good stability at follow-up (**Table 4.10**), with 95% (347 out of 367 visits) either remaining in the same subtype or progressing from the normal appearing (not subtypable) group to the *Subcortical* or *Cortical* subtypes. 97% (265/273) of PSP scans assigned to the *Subcortical* subtype at previous scan remained in that subtype at follow-up; 2% (five scans) switched to the *Cortical* subtype at follow-up scan, while 1% (three scans) reverted to *Normal* (not subtypable). Of those scans assigned to the *Cortical* subtype, 96% (78/81) showed stable subtype assignment, while 4% (three scans) switched to the *Subcortical* subtype.

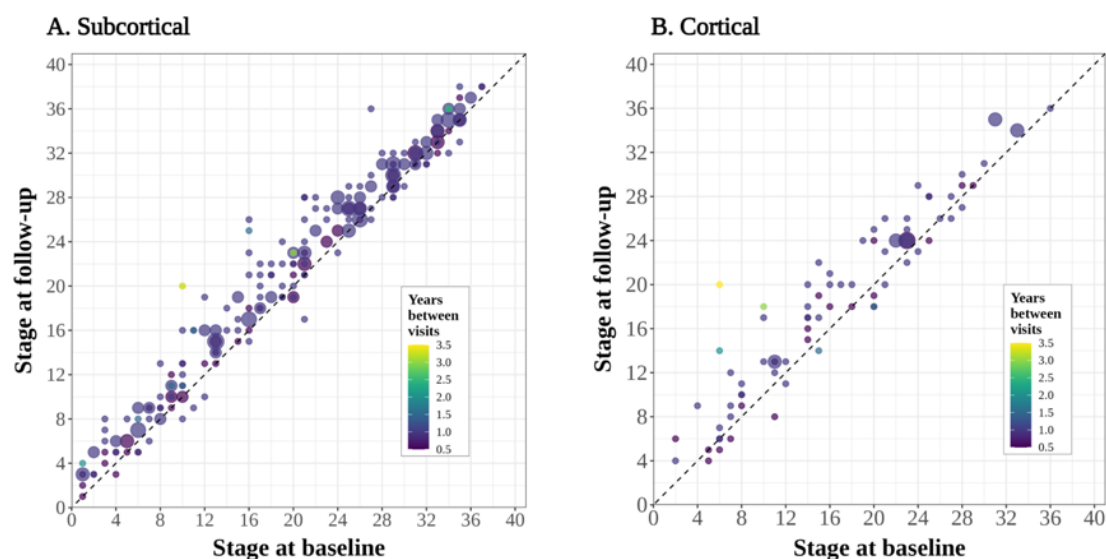
**Table 4.10 - Longitudinal consistency of subtype assignments**

Classification at previous visit	Classification at follow-up visit		
	Normal appearing <sup>a</sup>	Subcortical subtype	Cortical subtype
Normal appearing <sup>a</sup>	9 (69%) <sup>a</sup>	<b>3 (23%)<sup>b</sup></b>	<b>1 (8%)<sup>b</sup></b>
Subcortical subtype	3 (1%)	<b>265 (97%)<sup>b</sup></b>	5 (2%)
Cortical subtype	0 (0%)	3 (4%)	<b>78 (96%)<sup>b</sup></b>

<sup>a</sup>Normal appearing = not subtypable. Note that this only includes 13 individuals that were not subtypable at baseline and had a follow-up scan. An observation is longitudinally consistent (<sup>b</sup>) if individuals remain in the same group or progress from the normal-appearing group to Subcortical or Cortical subtype at follow-up visit. Entries indicate the number of visits n, with the % of the total individuals in classification at previous visit in classification at follow-up in brackets. Longitudinally consistent observations highlighted in bold.

I next tested how SuStaIn stage progressed over time (**Figure 4.7**) by comparing assigned stage at follow-up to baseline stage. As expected, the majority of individuals (90%) either progressed in stage (75%, 318/355), or stayed at the same stage (15%, 53/355) i.e. are on or above the line  $y = x$ . For those individuals assigned to the *Subcortical* subtype, 92% stayed at same stage or progressed (17% and 75% respectively); for those diagnosed with PSP-RS 92% stayed at same stage or progressed, for PSP-Cortical 100%, and for PSP-Subcortical 100%. In the *Cortical*

subtype 83% stayed at the same stage or progressed (9% and 74% respectively) in the *Cortical* subtype; the breakdown for clinical phenotypes assigned to this Subtype, was 83.9% for PSP-RS, PSP-Subcortical 100% and PSP-Cortical 77%.



**Figure 4.7 - Stage progression at follow-up visits by SuStaIn subtype.** Scatter plots of (A) Subcortical subtype (B) Cortical subtype showing predicted stage at baseline (x-axis) versus predicted stage at follow-up scan (y-axis) for those PSP cases with a follow-up scan ( $n = 355$ ). The area of the circle is weighted by the number of scans at each point, and the colour of the circle represents the time (years) between visits.

## 4.4 Discussion

The clinical heterogeneity of PSP is increasingly recognized<sup>98,359</sup> and although post-mortem studies<sup>51,78,79</sup> suggest that this heterogeneity is related to differences in the severity and neuroanatomical distribution of pathology, there is an urgent need to better delineate this variability in vivo. To this end I applied SuStaIn to a large PSP MRI data set, encompassing the spectrum of PSP clinical syndromes, and empirically identified two subtypes characterised by distinct temporal patterns of atrophy. I referred to these two subtypes as *Subcortical* and *Cortical* based on the earliest regions to show abnormality. Clinical scores of disease severity worsened with increasing stage and the *Subcortical* subtype was associated with more severe disease compared to the *Cortical* subtype, as measured by PSP Rating Scale. It is worth noting, however, that the PSP rating scale was originally designed to assess disease severity in PSP-RS, and as such is heavily weighted towards measuring motor predominant disease burden. It is therefore unlikely to be a good measure of overall disease burden in the cortical PSP syndromes<sup>393</sup>, and the scores in these cases relative to the PSP-RS / PSP-Subcortical variants need to be interpreted with caution. As expected, the *Cortical* subtype was enriched for patients clinically diagnosed with PSP-Cortical variants (PSP-CBS, PSP-F, PSP-SL). The *Subcortical* subtype was enriched for patients clinically diagnosed with both PSP-RS and PSP-Subcortical variants (PSP-P, PSP-PGF), though on average the PSP-Subcortical variants were at a lower subtype stage with a slower rate of progression compared to the PSP-RS cases. The distinct patterns of MRI atrophy in these subtypes provides unique insights into disease mechanisms across the disease course and supports the stratification of patients by subtype in clinical trials. The differences in atrophy patterns between the two subtypes are evident in the raw baseline regional volume scores (Table 4.7), further supporting that these differences can be picked up at the time of baseline scan.

The *Subcortical* and *Cortical* SuStaIn subtypes share some common early features (**Figure 4.3**), though are distinguished by much earlier cortical involvement in the latter. Both subtypes demonstrate early involvement of the midbrain and ventral diencephalon (which includes the subthalamic nucleus and substantia nigra), in keeping with previous post-mortem studies<sup>51,78</sup>, and several neuropathological studies of incidental or early-stage PSP cases<sup>64,116,117,394,395</sup>). In contrast to the *Subcortical* subtype, the *Cortical* subtype has concomitant atrophy in the insula, frontal lobes,

thalamus and basal ganglia in these early stages. The subsequent cortical atrophy in the *Cortical* subtype then progresses to the parietal, occipital and cingulate, and finally the temporal lobe. At equivalent stages the subcortical atrophy progresses to the medulla, SCP, dentate and GP, followed by the amygdala, and finally the caudate and cerebellar cortex. In contrast in the *Subcortical* subtype atrophy has already progressed through the whole brainstem, SCP and dentate nucleus, before it reaches the insula and posterior frontal lobe (supplementary motor cortex). Once it reaches the cortex the sequence of cortical atrophy is broadly similar to the *Cortical* subtype.

The sequence of atrophy in the *Subcortical* subtype broadly mirrors the sequential spread of tau pathology proposed by Kovacs et al.<sup>51</sup> in their post-mortem PSP-RS staging system and is consistent with in vivo disease progression models of PSP-RS<sup>358,396</sup>. In keeping with this, in this study 81% of PSP-RS cases were assigned to this SuStaIn subtype. In **Chapter 3** I used an event-based modelling approach with PSP-RS cases to identify this sequence, while Saito et al.<sup>396</sup> applied SuStaIn to a small cohort of PSP-RS and CBS cases. They found that the optimal model consisted of two subtypes, with one subtype associated with CBS and the other with PSP-RS. Given the small sample size of the Saito study, and the absence of any vPSP cases in the sample, they were unable to extract finer grained information on PSP heterogeneity.

An important difference in our study was that in addition to the majority of PSP-RS cases (81%) being assigned to the *Subcortical* subtype, the same was also true for the PSP-Subcortical (PSP-P and PSP-PGF) cases (82%). This implication of this is that atrophy in PSP-Subcortical variants progress along the same trajectory as PSP-RS cases. Previous work shows that PSP-P and PSP-PGF cases develop similar clinical phenomenology to PSP-RS cases in the later stages of the disease course<sup>397–399</sup>, albeit at a slower rate<sup>105,227,399</sup> resulting in longer survival times<sup>107</sup>. Post-hoc analysis of the *Subcortical* SuStaIn subtype supports this, with the PSP-SC cases at an earlier stage in the sequence than PSP-RS cases (**Figure 4.6**) (7.4 vs 19.9,  $t = 12.4$ ,  $p = 8.4 \times 10^{-5}$ ), with a slower stage progression per year (0.66 vs 1.86,  $t = 2.49$ , 95% CI 0.1 – 2.4,  $p = 0.046$ ).

Traditional cross-sectional imaging studies show that the PSP-Subcortical variants usually have less severe atrophy in the midbrain, medulla, and SCP relative to PSP-RS cases<sup>298,387,400–402</sup>. Although this may seem at odds with SuStaIn assigning PSP-RS and PSP-Subcortical to the same trajectory, it can be reconciled by the finding that at

the average stage for PSP-Subcortical variants (7.4) the z-score has only reached 2 sigma (2 standard deviations from controls) for the midbrain, and 1 sigma for the medulla and SCP, versus 3, 2 and 2 sigma respectively by stage 20 (the average stage for the PSP-RS cases). This suggests that on average PSP-SC cases have less atrophy in those regions at the time of baseline scan compared to PSP-RS cases, which would be consistent with the cross-sectional imaging findings that do not account for disease stage heterogeneity. It is currently unclear as to why the PSP-Subcortical variants show a less aggressive disease, though it may be at least partly due to protective genetic variants<sup>195</sup>.

There was also a strong assignment of PSP-Cortical cases (PSP-F, PSP-SL and PSP-CBS) to the *Cortical* subtype, which drove a lower average PSP rating scale and UPDRS score compared to the *Subcortical* subtype. As previously discussed, the *Cortical* subtype had early cortical atrophy predominantly affecting the insula, and frontal lobes alongside subcortical involvement, which is in keeping with cross-sectional MRI studies of PSP-Cortical variants<sup>298,400,403,404</sup>. Although there are few longitudinal imaging studies of these variants due to their rarity, a recent retrospective cohort of PSP-SL demonstrated that the majority of these cases developed symptoms typical of PSP-RS as their disease progressed<sup>404</sup>. My data supports this finding given the end stage atrophy pattern of the *Cortical* SuStaIn subtype is very similar to the *Subcortical* atrophy pattern. Further work is required to better understand what is driving the difference between the *Subcortical* and *Cortical* subtype atrophy patterns, especially early in the disease trajectory. The early cortical involvement in the *Cortical* subtype may, at least in part, be due to higher astroglial or oligodendroglial pathology relative to neuronal tau pathology early in the disease course<sup>51</sup>, though this is an area of ongoing research.

#### **4.4.1 Limitations**

This study has a number of limitations which highlight opportunities for future work. Although I built, to the best of my knowledge, the largest international cohort to date of PSP cases with baseline and follow-up imaging, the sample size for vPSP cases was still small (52 PSP-Cortical and 17 PSP-Subcortical variants). Whilst the SuStaIn algorithm uses a cross-validation framework, ideally one should have a separate training and test set to validate results. Given the low numbers for the vPSP syndromes



specifically, I opted not to do this with this dataset so as not to further underpower the model to find patterns associated with these syndromes. Future work will be required to confirm the validity of the two-subtype model on external PSP datasets. We may find that there are further distinct subtypes as vPSP numbers increase, though for this dataset the two-subtype model was optimal. Related to this, I decided to group the PSP-Cortical (PSP-F, PSP-S/L and PSP-CBS) and PSP-Subcortical (PSP-P and PSP-PGF) syndromes together in the post hoc analyses. This is an established approach in the PSP research setting<sup>107,195,198,298</sup>, though it will be interesting to look at finer grained SuStaIn subtype:clinical syndrome associations when I have larger sample sizes for the vPSP syndromes.

I collected MRI scans from a range of international centres across a number of different scanners. It is well known that using data from different scanners can introduce bias into downstream analyses, through variations in imaging quality, signal homogeneity and image contrast<sup>98</sup>. In addition to stringent visual QC of raw images and post-segmentation scans, I also regressed out both scanner manufacturer and field strength when generating adjusted regional volumes to try and account for this. In multi-centre clinical trials of a rare disease such as PSP, the use of multiple scanner types is the rule rather than the exception. I believe that this study's inclusion of cases across multiple scanner types (albeit adjusted for in the analysis) supports stronger generalisability of the findings to the wider clinical setting.

Another limitation was the use of the Lawton et al<sup>391</sup> method to convert MoCA scores to MMSE. This method has only been validated in Parkinson's disease, though has been used previously in a PSP study<sup>360</sup>. Given that the MMSE is likely to be less sensitive than the MoCA in PSP<sup>405</sup>, the cognitive impairment estimated in this study is likely to be on the conservative side.

SuStaIn fits data based on the assumption that there are a distinct set of trajectories, though it is possible that there is a spectrum of disease progression patterns within the data<sup>290</sup>. In this situation the identified SuStaIn subtype trajectory could have been created by appending unrelated disease trajectories into one subtype<sup>291</sup>. When assessing the two-subtype model I checked the average subtype probability by stage by subtype (**Figure 4.5B**), based on the hypothesis that if a separate sequence is appended to another, I would expect the average probability assignment to drop below 50% at some point in the sequence. This was not the case in our model supporting that

these are indeed two distinct trajectories. In addition, I found that the majority of individuals remained in the same subtype and progressed to later stages in that subtype which supports the model validity. The high association between PSP-Subcortical syndromes and the *Subcortical* subtype at baseline could also be due to this caveat i.e. they do not necessarily progress through the stages of the subtype to cortical involvement. However, when assessing stage at follow-up, one of the PSP-SC cases progressed from a baseline stage where there is only subcortical involvement (< stage 13) to a follow-up stage where the cortex started to be involved (> stage 13). Two other cases already had cortical involvement at baseline, progressing to more extensive cortical involvement at follow-up (**Table 4.7**). Taken together these findings give confidence that in my model appending of different disease trajectories into one is unlikely to be the case.

Another consideration is that SuStaIn has no explicit timescale and is only able to extract information on the relative position the individual is in the sequence within a given subtype. A recently developed generative model called the temporal event-based model (TEBM), addresses this issue, by using longitudinal information to extract transition times between events<sup>384</sup>. Work to integrate this framework into the SuStaIn algorithm is ongoing. It will be interesting in future work to see whether a temporal SuStaIn model identifies a third subtype, given the finding in this study that the majority of PSP-SC cases follow the same trajectory as PSP-RS cases though with a slower stage progression rate.

## 4.5 Conclusion and future work

The SuStaIn model provides data-driven evidence for the existence of two spatiotemporal subtypes of atrophy in clinically diagnosed PSP, giving insights into the relationship between PSP pathology and clinical syndrome. These image-based subtypes are differentially enriched for PSP clinical syndromes and show different clinical characteristics. The results suggest that the PSP-RS and PSP-Subcortical syndromes share a similar trajectory of atrophy, though the latter tend to be at an early stage at diagnosis and progress at a slower rate. Being able to accurately subtype and stage PSP patients at baseline has important implications for screening patients on entry into clinical trials, as well as for tracking disease progression. Future work should focus on validating these results in larger datasets with a higher number of vPSP syndromes that ideally have autopsy-confirmed PSP pathology, extracting information on time to transition between subtype stages and assessing the clinical relevance of these imaging subtypes in real-world setting

# **Chapter 5: Distinct spatiotemporal atrophy patterns in corticobasal syndrome are associated with different underlying pathologies**

The contents of this chapter form part of a manuscript which is currently in preparation for submission.

## **5.1 Introduction**

Corticobasal syndrome (CBS) is characterised by a progressive levodopa resistant asymmetric akinetic-rigid syndrome, and cortical features including apraxia, cortical sensory loss, cognitive dysfunction, and alien-limb phenomenon<sup>406</sup>. Oculomotor dysfunction and postural instability often co-exist in individuals with CBS<sup>140</sup>. Dysphagia, dysarthria and aphasia are also common<sup>407</sup>. Mean survival is 5-8 years<sup>387</sup> and there are currently no disease modifying treatments.

Although CBS was first described in individuals with corticobasal degeneration (CBD) pathology at post-mortem<sup>124</sup>, the Armstrong diagnostic criteria<sup>140</sup> have limited predictive value for identifying CBD pathology<sup>143</sup> in life. Autopsy studies demonstrate considerable underlying pathological heterogeneity in those who present clinically with CBS<sup>408</sup>. CBD pathology only accounts for 50% of all clinically diagnosed CBS patients<sup>409</sup>, with the others having Alzheimer's disease (AD), progressive supranuclear palsy (PSP), Pick's disease (PiD), globular glial tauopathy (GGT), transactive response DNA binding protein 43 (TDP-43) proteinopathy, dementia with Lewy bodies (DLB) and Creutzfeldt-Jakob disease at post-mortem<sup>123,137,138,408,410,411</sup>.

The emergence of amyloid and tau PET tracers, alongside CSF and now plasma biomarkers for AD<sup>412,413</sup> enables identification of CBS secondary to underlying AD pathology. Biomarkers for 4R tau (CBD, PSP, GGT), 3R tau (PiD),  $\alpha$ -synuclein (DLB), and TDP-43 are less well developed in comparison. Structural MRI studies of CBS cases with post-mortem pathology show that at the group level there are differences in the cross-sectional pattern of atrophy between some pathologies (CBD, AD, PSP and TDP-43). It is unclear, however, to what extent these differences are driven by differences in disease stage at time of MRI versus pathology specific

differences, given that the studies either do not correct for underlying disease stage<sup>137,233</sup>, or use MMSE as a proxy for stage<sup>235</sup>. Grouping individuals based on a similar atrophy pattern from cross-sectional MRI without fully accounting for disease stage is sub-optimal, as those belonging to the same subgroup are likely to have different spatial patterns of atrophy as the disease progresses and so would appear different (and vice-versa)<sup>294</sup>. Predicting the pathology underlying CBS is therefore difficult due to the lack of both clinico-pathological correlation, and specific biomarkers. Developing individualised disease progression models of pathological brain changes in CBS that predict this underlying heterogeneity will be critical to the success of clinical trials for emerging disease modifying therapies<sup>100,387,414,415</sup>.

In recent years, advances in machine-learning have provided tools to disentangle this *phenotypic* (clinical subtype) and *temporal* (pathological stage) heterogeneity. One such algorithm, Subtype and Stage Inference (SuStaIn)<sup>290</sup>, combines disease progression modelling with hierarchical clustering to identify probabilistic data-driven disease subtypes with distinct temporal progression patterns, using only cross-sectional data. The fitted disease model can be applied to unseen patient data, to infer what subtype and stage that individual belongs to. SuStaIn was originally applied to structural MRIs in Alzheimer's disease (AD) uncovering unique patterns of atrophy that showed an improved prediction of dementia conversion compared to stage or subtype only models. More recent work includes identifying four distinct patterns of tau accumulation in AD<sup>291</sup>, three subtypes of amyloid accumulation<sup>292</sup>, and redefining multiple sclerosis based on MRI-subtypes that predict disability progression and response to treatment<sup>294</sup>. The clinical and pathological heterogeneity of CBS makes it ideally suited to modelling using SuStaIn.

The aim of this study was to uncover imaging subtypes of CBS based solely on a data-driven assessment of atrophy patterns, to test the hypothesis that modelling disease subtype and stage jointly would provide information on the underlying pathology. To this end I used the SuStaIn algorithm with cross-sectional structural MRI data from a large international cohort of clinically diagnosed CBS patients. I further compared the clinical phenotypes and associated pathology in each SuStaIn subtype to gain insight into the relationship between atrophy, underlying pathology and clinical features.

## 5.2 Methods

### 5.2.1 Study cohorts and clinical data

MRI and clinical data from individuals with a clinical diagnosis of “possible” or “probable” CBS as per the Armstrong 2013 criteria<sup>140</sup> were collected from seven main cohorts; the 4R Tauopathy Imaging Initiative Cycle 1 (4RTNI 1; ClinicalTrials.gov: NCT01804452),<sup>259,263</sup> the 4R Tauopathy Imaging Initiative Cycle 2 (4RTNI 2; ClinicalTrials.gov: NCT02966145), the davunetide randomized control trial (DAV; ClinicalTrials.gov: NCT01056965),<sup>220</sup> the salsalate clinical trial (SAL; ClinicalTrials.gov: NCT02422485),<sup>297</sup> the young plasma clinical trial (YP; ClinicalTrials.gov: NCT02460731),<sup>297</sup> the PROgressive Supranuclear Palsy CorTico-Basal Syndrome Multiple System Atrophy Longitudinal Study (PROSPECT; ClinicalTrials.gov: NCT02778607),<sup>298</sup> and the University College London Dementia Research Centre (UCL DRC) FTD cohort. Controls were collected from three cohorts with equivalent available data; PROSPECT, the UCL DRC FTD cohort and the Frontotemporal Lobar Degeneration Neuroimaging Initiative dataset (FTLDNI; <http://4rtni-ftldni.ini.usc.edu/>). For further information regarding to the recruitment, diagnostic criteria and MRI scanner acquisition protocols please see Chapter 2 (**Participants, inclusion criteria, clinical and MRI data**). Appropriate ethical approval was acquired through application to each of the individual trial and research ethics committees.

For study inclusion all participants needed to have, as a minimum, a clinical diagnosis of “possible” or “probable” CBS according to the 2013 Armstrong criteria, a baseline T1 volumetric MRI on a 1.5 or 3 Tesla scanner, and basic demographic data (gender and age at time of scan). Clinical rating scale scores (PSP rating scale, Unified Parkinson Disease Rating Scale [UPDRS], Schwab and England Activities of Daily Living scale [SEADL], and Montreal Cognitive Assessment [MoCA] or Mini-mental State Examination [MMSE] at baseline and follow-up), pathology at autopsy, CSF AD biomarker positivity [ $A\beta$ 1–42, tau, and ptau], amyloid PET positivity (with florbetaben, florbetapir, or Pittsburgh Compound-B), and follow-up scans were also included if available. Amyloid PET scans were collected at participating 4RTNI centers with demonstrated experience in FDA-approved amyloid imaging agents, and positivity was defined by expert visual read by certified staff.

As detailed in previous work<sup>358</sup>, original trial analyses failed to show any treatment effect (including no change in volumetric MRI measurements) in the SAL, YP and DAV trials, so data were combined from each study's treatment and placebo arms. Longitudinal data were used to validate the consistency of SuStaIn's subtype and stage assignments at follow-up.

For missing clinical scale data, an adjusted mean score was used if at least 80% of the assessment was complete<sup>298</sup>. Given the PROSPECT and 4RTNI2 trials only assessed cognitive function using the MOCA (as opposed to the MMSE for the other trials), raw MOCA scores were converted to MMSE scores using the validated method used by Lawton et al<sup>391</sup>.

## 5.2.2 MRI acquisition and image processing

The MRI acquisition protocols, and image processing pipeline have been described in detail previously in Chapter 2 (**Image processing pipeline**). To summarise, cortical and subcortical structures were automatically parcellated using geodesic information flows algorithm (GIF)<sup>304</sup>, a multi-atlas segmentation propagation approach based on the on the Neuromorphometrics atlas (Neuromorphometrics, Inc.) Subregions of the cerebellum were parcellated using GIF based on the Diedrichsen atlas<sup>305</sup>, and the brainstem structures were subsequently segmented using a version of the brainstem module available in FreeSurfer, customised to accept the GIF parcellation of the whole brainstem as input<sup>306</sup>. Volumes for 24 grey-matter regions were calculated; four brainstem (medulla, pons, superior cerebellar peduncle [SCP] and midbrain), three cerebellar (cerebellar cortex, dentate nucleus and vermis), nine subcortical (thalamus, globus pallidus (GP), caudate, putamen, ventral diencephalon (DC), thalamus, hippocampus, amygdala and nucleus accumbens [NA]) and eight cortical (basal forebrain, cingulate, frontal anterior, frontal posterior, insula, temporal, parietal and occipital) regions. A list of GIF subregions included in each cortical region is detailed in **Table 5.1**. I calculated total intracranial volume (TIV) using SPM12 v6225 (Statistical Parametric Mapping, Wellcome Trust Centre for Neuroimaging, London, UK) running in MATLAB R2012b (Math Works, Natick, MA, USA)<sup>365</sup>. I then visually inspected all segmentations were to ensure accurate segmentation. Regional volumes were corrected for scanner field strength (1.5T or 3T), scanner manufacturer, sex, age at baseline scan and TIV, by performing a linear regression on the control

population and then propagating this model to the CBS population, to generate covariate-adjusted regional volumes.



**Table 5.1 - GIF subregions included in each cortical and cerebellar region used as SuStaIn input.**

<b>Regions included in SuStaIn</b>	<b>GIF Subregions</b>
Frontal Anterior	Frontal operculum, central operculum, frontal pole, gyrus rectus, middle frontal cortex, subcallosal area, superior frontal gyrus medial segment, superior frontal gyrus, middle frontal gyrus, opercular part of the inferior frontal gyrus, orbital part of the inferior frontal gyrus, triangular part of the inferior frontal gyrus, anterior orbital gyrus, medial orbital gyrus, lateral orbital gyrus, posterior orbital gyrus
Frontal Posterior	Precentral gyrus, precentral gyrus medial segment, supplementary motor cortex
Temporal	Entorhinal area, fusiform gyrus, parahippocampal gyrus, inferior temporal gyrus, middle temporal gyrus, superior temporal gyrus, temporal pole, planum polare, planum temporale, transverse temporal gyrus
Parietal	Precuneus, parietal operculum, supramarginal gyrus, superior parietal lobule, angular gyrus, postcentral gyrus, postcentral gyrus medial segment
Occipital	Cuneus, calcarine cortex, lingual gyrus, occipital fusiform gyrus, superior occipital gyrus, inferior occipital gyrus, middle occipital gyrus, occipital pole
Insula	Anterior insula, posterior insula
Amygdala	Amygdala
Corpus Callosum	Corpus Callosum
Medulla	Medulla
Pons	Pons
Superior Cerebellar Peduncles	Superior cerebellar peduncles
Midbrain	Midbrain
Ventral Diencephalon	Ventral Diencephalon (GIF segmentation includes subthalamic nucleus, substantia nigra and hypothalamus)
Thalamus	Thalamus
Caudate	Caudate
Globus Pallidus	Globus Pallidus
Putamen	Putamen
Cerebellar Cortex	Lobules I/IV, V, VI, VIIA-Crus I, VIIA-Crus II, VIIB, VIIIA, VIIB, IX, X
Dentate	Dentate nucleus

### 5.2.3 Biomarker selection and Z-scoring of data

I defined a biomarker in this study as an image-based regional volume (24 regions as described in previous section). I carried out pairwise comparisons between healthy controls and cases at baseline visit, and selected covariate adjusted regional volumes where the difference between the two groups was associated with a moderate to large effect size (Cohen's  $d$  effect size of 0.6 for standardized mean differences between the cases and controls). This resulted in the selection of 19 regions of interest (ROI) that I then included in downstream analysis (**Table 5.2**); four brainstem (medulla, pons, SCP and midbrain), two cerebellar (cerebellar cortex and dentate nucleus), seven subcortical (thalamus, globus pallidus (GP), caudate, putamen, ventral diencephalon (DC), thalamus, and amygdala) and six cortical (frontal anterior, frontal posterior, insula, temporal, parietal and occipital) regions. Regions that had a right and left label were combined. I converted covariate adjusted regional volumes for these 19 ROIs into z scores relative to the control group by subtracting the mean of the control group from each patient's ROI volume and dividing by the standard deviation of the control group. Given regional brain volumes decrease with disease progression, the z scores become negative as the disease progresses; I therefore multiplied the z scores by -1, to give positive z scores that increase with disease progression. I used this z scored data as input to SuStaIn.

### 5.2.4 Subtype and Stage Inference

SuStaIn is a probabilistic machine learning algorithm that simultaneously clusters individuals into groups (subtypes) and infers a trajectory of change associated with each group; that trajectory defines the disease stage (degree of disease progression within a subtype) of each individual within the corresponding group<sup>290</sup>. It only requires cross-sectional data as input, though is able to exploit longitudinal data for training if available. More details on this algorithm are provided in Chapter 2 (**Subtype and Stage Inference (SuStaIn)**). In summary, each subtypes' progression pattern is described using a piecewise linear z score model, expressing a trajectory with a series of stages, that each correspond to a single biomarker (regional brain volume in this case) reaching a new z score. The number of SuStaIn stages is determined by the number of biomarkers (the product of the number of ROIs and number of z score thresholds per ROI) provided as input. SuStaIn optimises both the subtype

membership and the ordering in which different biomarkers reach different z-scores in each subtype (for example one, two or three standard deviations away from the control mean for that ROI) using a data likelihood function.

I fitted the SuStaIn model on the baseline imaging data for CBS cases; model uncertainty was estimated using 100,000 Markov Chain Monte Carlo (MCMC) iterations and in the single-cluster expectation maximisation procedure (**Figure 2.4**) the single-cluster sequence was optimised from 24 different random starting sequences to find the maximum likelihood solution. **Table 5.3** provides a summary of the Z-score settings, MCMC iterations and number of random starting sequences used for the SuStaIn algorithm. The optimal number of subtypes was determined using information criteria calculated through ten-fold cross-validation (cross-validation information criteria; CVIC), to balance internal model accuracy with model complexity. In cases where the evidence for a more complex model (more subtypes) was not strong (defined as per Young et al.<sup>290</sup> as a difference of less than 6 between CVIC and the minimum CVIC across models, or equivalently a difference of less than 3 between the out-of-sample log-likelihood and the minimum out-of-sample log-likelihood across models), I selected the less complex model (fewer subtypes) to avoid overfitting<sup>416</sup>.

I then used the fitted model to calculate the probability that each individual falls at each stage of each subtype, and individuals were assigned to their maximum likelihood subtype and stage (as per Young et al.<sup>290</sup>). Subtype progression patterns identified by SuStaIn were visualized using BrainPainter<sup>392</sup>, that I modified to include brainstem segmentations.

**Table 5.2 - Effect size (Cohen's *d*) by region of interest.**

Region of Interest	Cohen's <i>d</i>
Putamen	1.62
Frontal Posterior	1.49
Midbrain	1.44
Thalamus	1.43
Parietal	1.37
Insula	1.36
Globus pallidus	1.22
Pons	1.13
Amygdala	1.07
Ventral DC	1.06
SCP	1.05
Temporal	0.90
Occipital	0.89
Frontal Anterior	0.88
Caudate	0.85
Dentate	0.78
Corpus callosum	0.73
Medulla	0.72
Cerebellar cortex	0.66
NA	0.59 <sup>a</sup>
Basal forebrain	0.59 <sup>a</sup>
Cingulate	0.50 <sup>a</sup>
Vermis	0.49 <sup>a</sup>
Hippocampus	0.45 <sup>a</sup>

Cohen's *d* calculated as the standardised mean difference between adjusted the regional volume of that that region of interest (ROI) in cases vs controls. <sup>a</sup> a threshold of greater than or equal to 0.6 was used to select ROI as input for SuStaIn algorithm resulting in 19 biomarkers being included in model. Abbreviations: DC = diencephalon, SCP = superior cerebellar peduncles, NA = nucleus accumbens

## 5.2.5 Statistical analysis

Individuals assigned to either SuStaIn stage 0 (i.e. no atrophy on imaging compared to controls) or Stage 40 (end stage i.e. all ROI maximum atrophy) were labelled “*no subtype*”. All other individuals were labelled as “*subtypable*” and I assigned these to their most probable subtype and stage. In addition, I stratified all CBS cases by underlying pathology (or absence of); CBS-PSP, CBS-CBD, CBS-AD or CBS-Indeterminate (CBS-IDT). While CBS-PSP and CBS-PSP could only be diagnosed by post-mortem pathology, cases were assigned to CBS-AD category either by post-mortem pathology, or if they had a positive AD biomarker in life (CSF Tau/A-Beta 1-

42 or Amyloid PET). All other cases without a post-mortem diagnosis or a positive AD biomarker were assigned CBS-IDT.

**Table 5.3 - SuStaIn algorithm settings for each biomarker**

Biomarker	R	$Z_{max}$
Frontal Anterior	2	5
Frontal Posterior	3	5
Temporal	2	5
Parietal	3	6
Occipital	2	5
Insula	2	5
Amygdala	2	4
Corpus Callosum	1	3
Medulla	2	4
Pons	2	4
SCP	2	4
Midbrain	3	6
Ventral DC	3	5
Thalamus	2	5
Caudate	1	4
GP	2	4
Putamen	3	5
Cerebellar cortex	1	3
Dentate	2	4

R is the number of z-scores included for biomarker  $i$ , and  $Z_{max}$  is maximum z-score modelled for biomarker  $i$ . Total number of biomarkers ( $i$ ) for model = 40.  $C_{max}$  (the maximum number of subtypes fitted) = 3. Model uncertainty was estimated using 100,000 Markov Chain Monte Carlo (MCMC) iterations. In the single-cluster expectation maximisation procedure the single-cluster sequence was optimised from 24 different random starting sequences to find the maximum likelihood solution.

### 5.2.5.1 Baseline characteristics

I performed pairwise comparisons of baseline characteristics between all CBS cases and controls, CBS pathological diagnosis (CBS-CBD, CBS-PSP, CBS-AD and CBS-IDT) vs all CBS cases, and each CBS pathology grouping against each other, using two-tailed unpaired  $t$ -tests for continuous variables and  $\chi^2$  tests for categorical variables. Statistical significance was reported at a level of  $p < 0.05$ , both uncorrected for and corrected for multiple comparisons (Bonferroni correction).

### 5.2.5.2 Association between Subtype assignment and covariates

I tested for any residual association between covariates (scanner field strength, scanner manufacturer, sex, age at baseline scan and TIV) and SuStaIn subtype, by fitting a logistic regression model to the data using the *lm()* function from the R stats package (version 3.6.2).

### 5.2.5.3 Subtype characterisation

First, I assessed the overall differences between subtypes independently of stage, excluding individuals classified as “no subtype” (Stage 0 or Stage 40). Two-tailed unpaired  $t$ -tests were performed for continuous variables and  $\chi^2$  tests for categorical variables followed by post-hoc pairwise comparisons for CBS pathology vs SuStaIn subtype using the *chisq.multicomp()* function from the *RVAideMemoire* R package (version 0.9-81-2).

To test for associations between clinical scores (PSP rating scale, UPDRS, SEADL and MMSE) and SuStaIn predicted subtype, a linear mixed effects model was fit to the data using the *lme4* package<sup>367</sup> in R Studio (version 1.4.1106). Subject Id was modelled as a random effect (random intercept) due to some subjects having two MRI scans at different time points. I accounted for SuStaIn subtype and stage, age, and sex by fitting a linear mixed effects model (Clinical score ~ subtype + stage + (1 | ID) + AAS + gender) for each clinical test score. Significance was calculated using the *lmerTest* package<sup>368</sup> which applies Satterthwaite’s method to estimate degrees of freedom and generate p-values for mixed models. Statistical significance was reported at a level of  $p < 0.05$ , and at the Bonferroni corrected level of  $p < 0.005$  for demographic variables (11 items), and  $p < 0.0125$  for clinical scores (10 variables), to account for multiple comparisons.

To assess average stage by clinical syndrome by SuStaIn subtype, I performed a one-way ANOVA (Mean stage ~ CBS pathology + Sustain baseline subtype) with the *aov()* function the *stats* package (version 3.6.2). Tukey post-hoc significant differences were then calculated to identify the level of significance.

Finally, I tested for differences in all baseline regional volumes of interest between the different SuStaIn subtypes using two-tailed unpaired *t* tests, with statistical significance reported at a level of  $p < 0.05$ , both uncorrected for and corrected for multiple comparisons (Bonferroni correction).

All statistical analyses were performed either in R (version 4.0.5) or Python (version 3.7.6).

## 5.2.6 Longitudinal validation

I used the longitudinal imaging data to validate the stability of subtypes, and to assess stage progression, based on the hypothesis that individuals should remain assigned to the same subtype but advance to higher stages over time (or at least remain at the same stage). Subtype stability was defined as the proportion of individuals that were assigned to the same subtype at follow-up(s) or progressed from Stage 0 (not subtypable) to a higher stage and subtype (i.e. became subtypable). To assess stage progression, I compared SuStaIn stage at baseline and follow-up(s) for all individuals and calculated the proportion of individuals that either advanced to a higher stage or stayed at the same stage at follow-up.

## 5.3 Results

### 5.3.1 Minimal difference in demographic and clinical characteristics between CBS pathologies at baseline

**Table 5.4** summarises the key baseline demographic and clinical features for CBS cases and controls included in this study. In total this study included 504 MRI images from a total of 387 individuals; 135 had a clinical diagnosis of CBS (with 113 follow-up scans) and 252 controls. Of those diagnosed with CBS, 52 (39%) received a pathological diagnosis; 12 were CBS-CBD, 6 were CBS-PSP, 34 were CBS-AD and

83 were CBS-IDT. There were no data available on co-pathologies in those that received a pathological diagnosis.

Overall, the CBS cases had an older average age at time of first scan compared to controls (66.4 years,  $SD \pm 7.7$  vs 62.3 years,  $SD \pm 9.2$ ,  $p < 0.05$ , corrected for multiple comparisons), though were matched for gender. Disease duration at time of first scan was lower in the CBS-CBD group compared to CBS-AD and CBS-IDT (3.4 years,  $SD \pm 1.6$  vs 4.9 years,  $SD \pm 3.2$  vs 5.2 years,  $SD \pm 2.9$ ,  $p < 0.05$  for all uncorrected for multiple comparisons) though this did not survive Bonferroni correction.

Regarding clinical scores, the only statistically significant difference between pathology groups was in the Bulbar sub-score of the PSP rating scale which was lower in the CBS-CBD group compared to CBS-IDT (3.4 years,  $SD \pm 1.6$  vs 5.2 years,  $SD \pm 2.9$ ,  $p < 0.05$  uncorrected for multiple comparisons). There was a trend towards increased motor predominant disease burden (PSP rating scale score and UPDRS) in the CBS-PSP group compared to the other groups but this did not reach statistical significance. There was also no difference between the SEADL and MMSE scores between pathological groups.



**Table 5.4 – Baseline clinical and demographic data (by pathology)**

	Controls	All CBS	CBS-CBD	CBS-PSP	CBS-AD	CBS-IDT
Baseline, n (fu visits)	252	135 (113)	12 (13)	6 (5)	34 (26)	83 (69)
Sex, % female	57%	51%	50%	83%	38%	54%
Age first scan, y	62.3 (9.2) <sup>c</sup>	66.4 (7.7) <sup>c</sup>	64.8 (6.2)	70.4 (5.7)	66.5 (7.7)	66.3 (8.0)
Age at first symptom, y <sup>a</sup>	-	61.5 (8.7)	65.2 (7.0)	60 (2.83)	61.5 (8.0)	61.2 (9.4)
Disease duration, y <sup>a, b</sup>	-	4.9 (2.9)	3.4 (1.6) <sup>d, e, f</sup>	4.3 (1.6)	4.9 (3.2) <sup>e</sup>	5.2 (2.9) <sup>f</sup>
PSP rating scale score	-	26.3 (13.9)	26.7 (15.0)	33.8 (8.7)	24.9 (12.8)	26.5 (14.6)
History	-	5.6 (3.2)	6.7 (3.9)	6.8 (4.5)	5.0 (2.6)	5.7 (3.3)
Mentation	-	3.1 (2.8)	2 (1.5)	3.8 (1.3)	3.7 (3.4)	3.0 (2.4)
Bulbar	-	1.7 (2.1)	0.9 (1.2) <sup>f</sup>	1.0 (0.8)	1.4 (2.3)	2.0 (2.1) <sup>f</sup>
Ocular motor	-	2.3 (3.5)	2.9 (3.8)	3.8 (2.2)	1.7 (2.2)	2.4 (4.0)
Limb motor	-	7.7 (3.7)	7.7 (3.7)	9.2 (2.1)	7.3 (3.7)	7.7 (3.9)
Gait and midline	-	5.9 (5.0)	6.6 (5.1)	9.3 (7.2)	5.7 (5.2)	5.7 (4.9)
SEADL	-	57.8 (25.5)	55.7 (20.7)	42.5 (54.4)	53.2 (27.6)	61.0 (24.7)
UPDRS	-	32.0 (17.2)	34.3 (13.0)	47.2 (26.4)	31.2 (20.2)	31.0 (15.7)
MMSE	-	23.8 (5.9)	23.3 (7.5)	19.2 (8.6)	22.0 (7.4)	25.0 (4.3)

Values are mean (SD), apart from Gender % female, Baseline n (n follow-up visits), Pathology n (% PSP). Pairwise comparisons between groups were performed using t tests for continuous variables and  $\chi^2$  tests for categorical variables. <sup>a</sup> note incomplete data for disease duration / age at first symptom. <sup>b</sup> time from first symptom to first scan. <sup>c</sup> CBS all vs Controls. Statistically significant at  $p < 0.05$ , corrected for multiple comparisons. <sup>d</sup> CBS [pathology group] vs All CBS. Statistically significant at  $p < 0.05$ , corrected for multiple comparisons. <sup>e</sup> CBS-CBD vs CBS-AD. Statistically significant at  $p < 0.05$ , uncorrected for multiple comparisons. <sup>f</sup> CBS-CBD vs CBS-IDT. Statistically significant at  $p < 0.05$ , uncorrected for multiple comparisons. Abbreviations: PSP = progressive supranuclear palsy, CBD = corticobasal degeneration, AD = Alzheimer's disease, IDT – indeterminate pathology, SEADL = Schwab and England Activities of Daily Living, UPDRS = Unified Parkinson's Disease Rating Scale, MMSE = Mini-Mental State Examination

### 5.3.2 Spatiotemporal subtypes of CBS

Given CBS is such a rare disease (3 / 100,000 estimated prevalence<sup>417</sup>) I fitted SuStaIn using CBS cases only, based on the rationale that it is very unlikely any of our controls had asymptomatic CBS. Indeed, it is more likely that the controls would have a more common neurodegenerative disorder such as AD, which may confound subtype and stage inference, further supporting the exclusion.

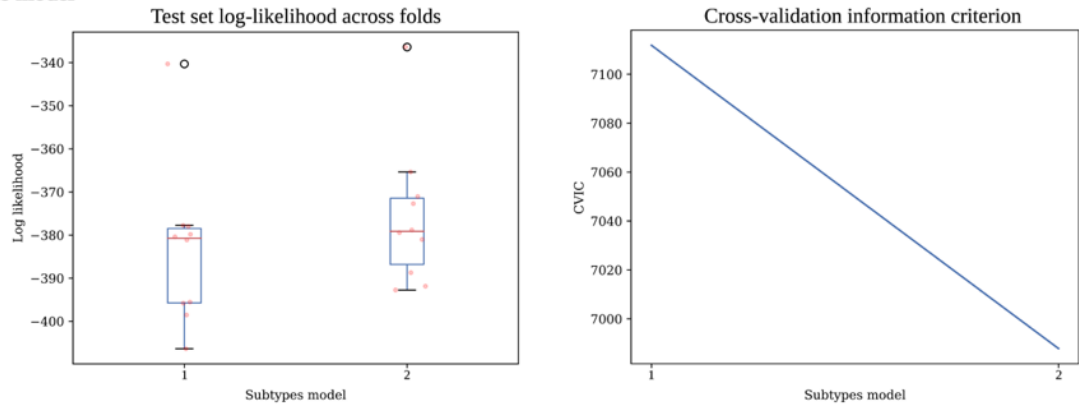
#### 5.3.2.1 Selection of optimum number of subtypes

I started with the hypothesis that there would be three distinct subtypes of atrophy in the CBS cohort. Comparing the out-of-sample log likelihoods and CVIC for the three-subtype model and the two-subtype model demonstrated that the two-subtype model (**Figure 5.1A**) best fitted the data, as the CVIC difference for the three-subtype versus the two-subtype model (**Figure 5.2B**) was greater than six (and the difference in log-likelihoods was greater than three). Given that the study was likely to be underpowered with only 135 cases, I decided to investigate both models to compare the disease progression patterns and clinical phenotypes.

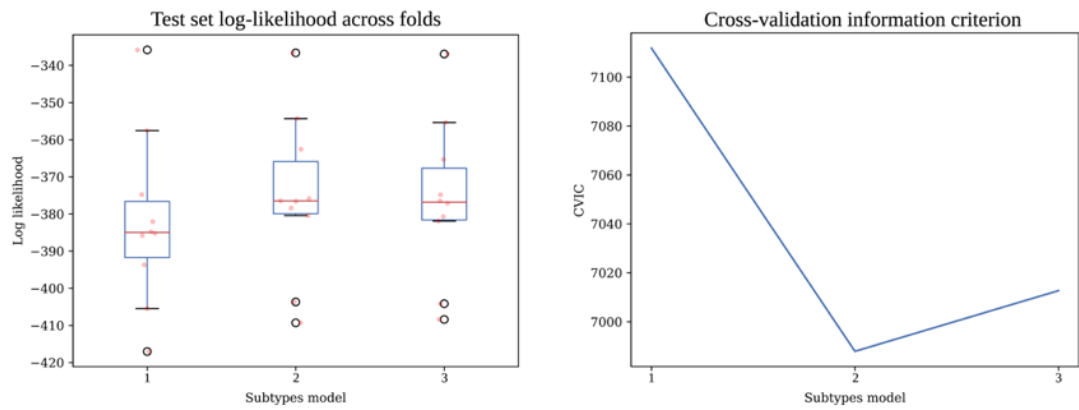
#### 5.3.2.2 Two-subtype model

The two-subtype model was the most parsimonious, with the lowest CVIC after 10-fold cross validation (**Figure 5.1**). Based on the earliest MRI abnormalities seen in the SuStaIn defined trajectories, I named the first the *Subcortical* subtype and the second the *Fronto-parieto-occipital* subtype (**Figure 5.2A.** and **Figure 5.2B** for positional variance diagrams [PVD]). The *Subcortical* subtype (46% of cases) starts with atrophy in the SCP of the cerebellum and the midbrain, followed by the pons, medulla, ventral diencephalon, dentate nucleus and thalamus. The atrophy then progresses to the posterior frontal lobe and the insula, posteriorly to the parietal and occipital lobes and anteriorly to the anterior frontal lobes, before finally affecting the temporal lobes. In contrast, in the *Fronto-parieto-occipital* subtype (54% of cases) the earliest atrophy starts in the parietal lobe and posterior frontal lobe, followed by the insula, occipital and then temporal lobe. Atrophy in the basal ganglia (putamen and GP) also occurs earlier on in this subtype than the *Subcortical* subtype, while the brainstem, thalamus and ventral diencephalon become atrophic later in sequence.

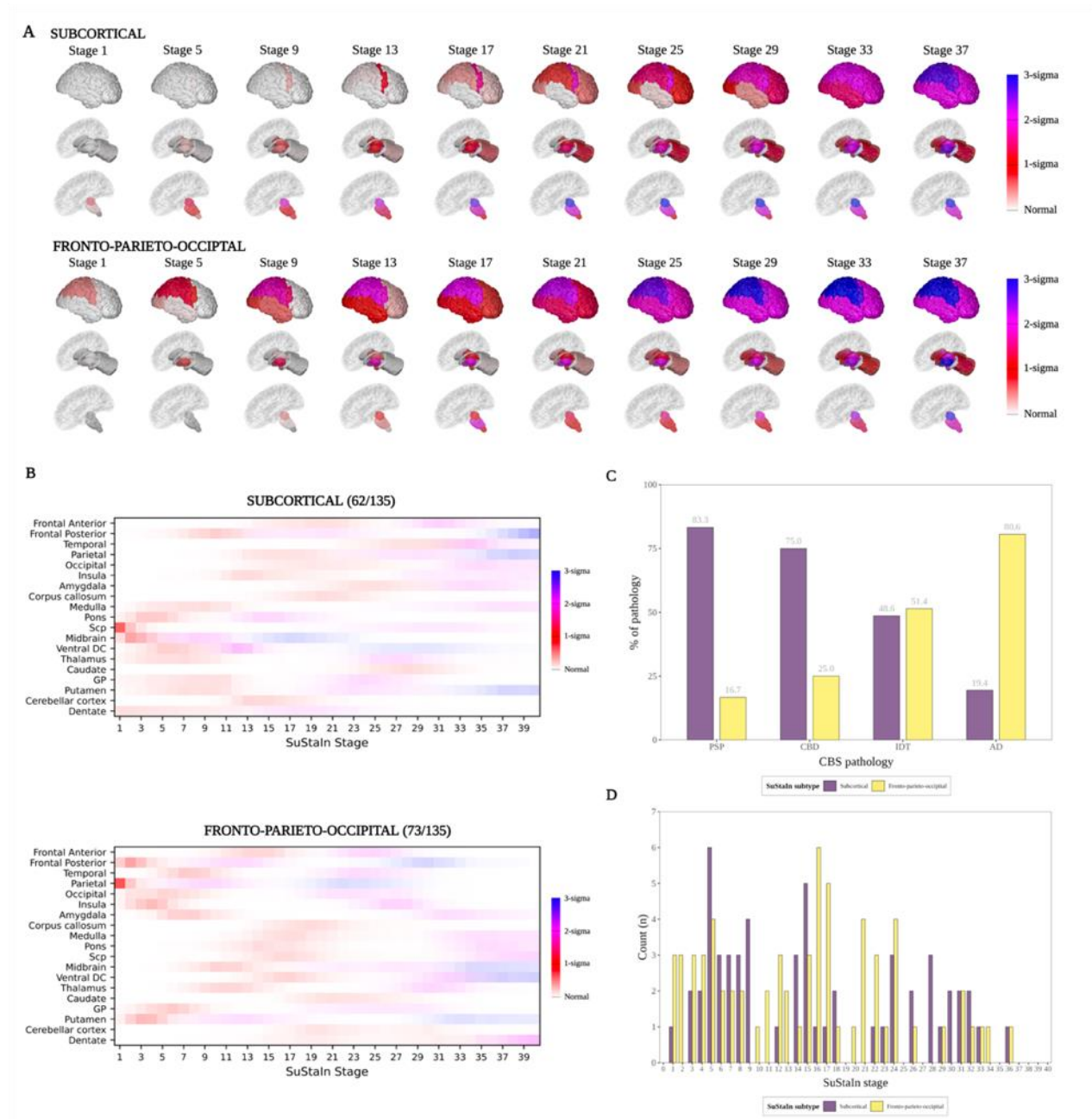
### A. 2-subtypes model



### B. 3-subtypes model



**Figure 5.1 - Selecting optimal SuStaIn subtype model given data.** The plots on the left of the figure show the test set log-likelihood across ten cross validation folds for (A) the two-subtype model and (B) the three-subtype model. The plots on the right show the cross-validation information criterion (CVIC) for each of the models as detailed above. The fact that the test set log-likelihoods drop and the CVIC increased with the addition of a third subtype (B) suggests that the two-subtype model is the most parsimonious and best for the data.



**Figure 5.2 - Two-subtype model of atrophy progression in CBS identified by Subtype and Stage Inference (SuStaIn).** (A) Spatial distribution and severity of atrophy at each SuStaIn stage by Subtype. Each row (Subcortical top, Fronto-parieto-occipital bottom) represents a subtype progression pattern identified by SuStaIn consisting of a set of stages at which brain volumes in CBS cases reach different z-scores relative to controls. (B) Positional variance diagrams for each SuStaIn subtype. These represent the uncertainty in the subtype progression patterns for each region. Each region (y-axis) is shaded based on the probability a particular z score is reached at a particular SuStaIn stage (x-axis). Z scores range from zero (white), one (red), two (pink) to three (blue). (C) Assignment of CBS pathology to each SuStaIn subtype. Size of bar (x-axis) represents percentage of cases labelled with that PSP syndrome assigned to that SuStaIn subtype (y-axis). PSP = PSP pathology at post-mortem, CBD = at post-mortem, AD = AD pathology at post-mortem or a positive AD biomarker (CSF or Amyloid PET) during life. (D) SuStaIn stage distribution by Subtype at baseline scan.

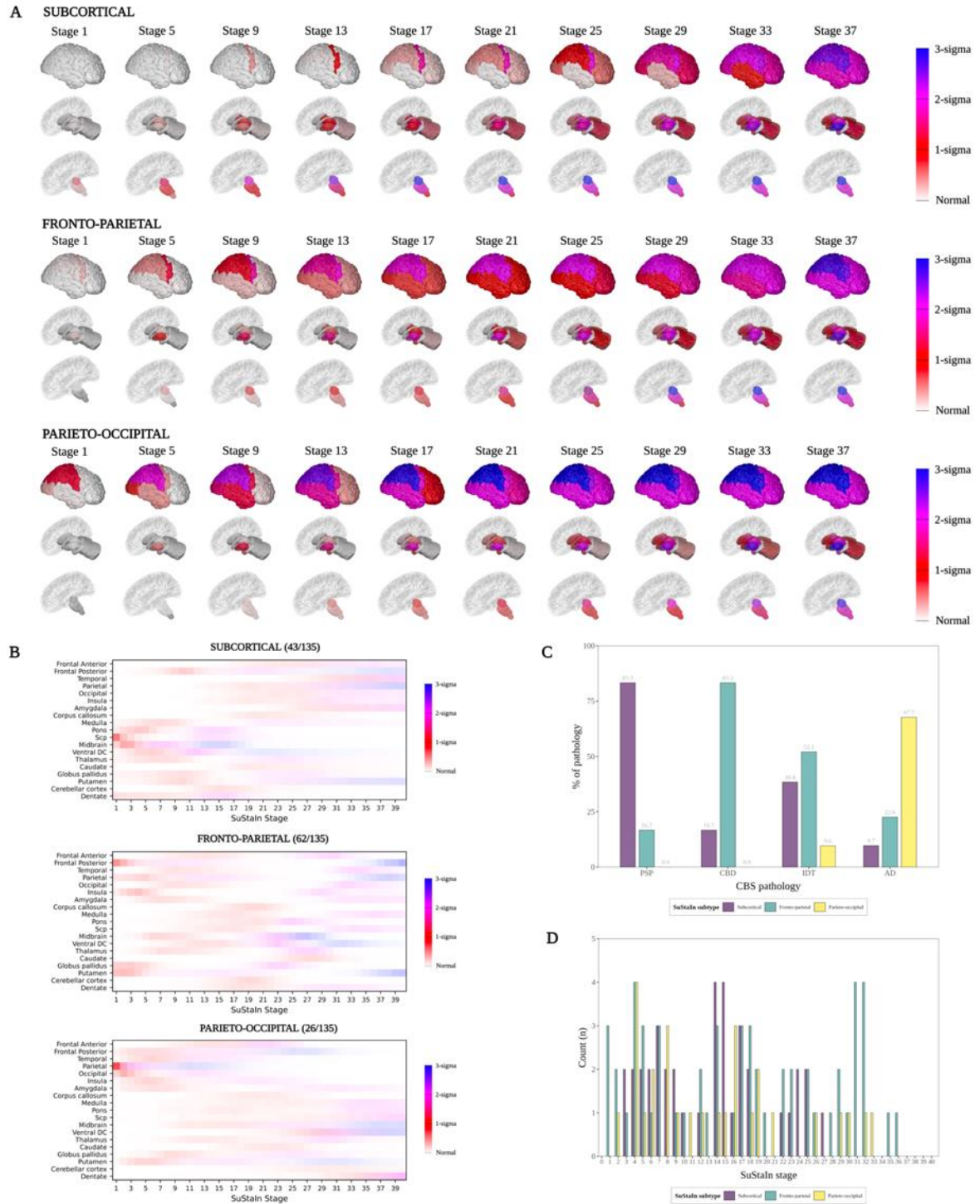
Overall, 12 of the 135 individuals (9%) were not subtypable by the two-subtype model, and so were excluded from subtype post-hoc analysis. Three of these individuals had a pathological diagnosis of AD (CBS-AD) and nine were CBS-IDT. Interestingly of the nine CBS-IDT, six had negative AD biomarkers and were therefore a pathology other than AD.

I fitted a logistic regression model to assess for any residual association between SuStaIn subtype, SuStaIn stage, regressed covariates and cohort (SuStaIn subtype ~ SuStaIn stage + TIV + age at first scan + sex + scanner field strength + scanner manufacturer + cohort). Apart from age at first scan (younger in *Fronto-parietal-occipital* subtype [ $z = 2.2$ ,  $p = 0.03$ ]) there was no dependency of subtype on any of the other covariates including SuStaIn stage which showed a similar distribution of stages across each subtype (**Figure 5.2C**).

### 5.3.2.3 Three-subtype model

In the three-subtype model (**Figure 5.3A** and **Figure 5.3B** for the PVDs) the *Subcortical* subtype (32% of cases) was also present with a very similar trajectory of atrophy to the *Subcortical* subtype in the two-subtype model. The second subtype I named the *Fronto-parietal* subtype (46% of cases) which had earliest atrophy in the posterior frontal and basal ganglia regions, followed closely by the insula and parietal regions. The midbrain and thalamus were affected next followed by the temporal and occipital lobes. The third, *Parieto-occipital* (19%) subtype, showed the most posterior atrophy with the parietal and occipital lobes affected first followed by the posterior frontal lobe and putamen, then the insula amygdala and temporal lobe.

13 of the cases (9.6% of all cases) were not subtypable at baseline; three of these had a pathological diagnosis of CBS-AD, and the other ten were CBS-IDT. Six of the ten CBS-IDT cases were negative for AD biomarkers. There was similar distribution of stages across each subtype (**Figure 5.3B**).



**Figure 5.3 - Three-subtype model of atrophy progression in CBS identified by Subtype and Stage Inference (SuStaIn). (A)** Spatial distribution and severity of atrophy at each SuStaIn stage by Subtype. Each row (Subcortical top, Fronto-parietal middle and Parieto-occipital bottom) represents a subtype progression pattern identified by SuStaIn consisting of a set of stages at which brain volumes in CBS cases reach different z-scores relative to controls. **(B)** Positional variance diagrams for each SuStaIn subtype. These represent the uncertainty in the subtype progression patterns for each region. Each region (y-axis) is shaded based on the probability a particular z score is reached at a particular SuStaIn stage (x-axis). Z scores range from zero (white), one (red), two (pink) to three (blue). **(C)** Assignment of CBS pathology to each SuStaIn subtype. Size of bar (x-axis) represents percentage of cases labelled with that PSP syndrome assigned to that SuStaIn subtype (y-axis). PSP = PSP pathology at post-mortem, CBD = at post-mortem, AD = AD pathology at post-mortem or a positive AD biomarker (CSF or Amyloid PET) during life. **(D)** SuStaIn stage distribution by Subtype at baseline scan.

### 5.3.3 Longitudinal consistency of models

To validate the models' inference of subtype longitudinal trajectories from the baseline MRI data, I used the fitted SuStaIn models to subtype and stage the follow-up MRI data. A total of 103 follow-up (103/113) scans were subtypable for both the two- and three-subtype models from a total of 63 CBS cases (47% of all CBS cases in cohort; 23 cases had one follow-up scan, 37 had two follow-up scans and two had three follow-up scans). The ten scans not subtypable at follow-up were also not subtypable at baseline scan. Mean follow-up time was 0.87 years with a SD of 0.45 years. I first assessed whether individual subtype assignments were stable at follow-up scan, and then tested whether individuals within their subtype progressed to later stages at follow-up.

#### 5.3.3.1 SuStaIn subtype assignments were stable at follow-up

Overall, the two-subtype model showed the highest subtype assignment stability with 98% of those with subtypable follow-up scans (101/103) remaining in the same subtype at follow-up or progressing to a subtype from being non-subtypable at baseline (one case) (**Table 5.5**). Two cases assigned to the *Fronto-parieto-occipital* subtype switched to the *Subcortical* subtype at follow-up (both CBS-AD). The average probability with which SuStaIn assigned individuals to the subtypes at baseline was high; 0.92 (SD  $\pm$  0.1) for the *Subcortical* subtype and 0.94 (SD  $\pm$  0.1) for the *Fronto-parieto-occipital* subtype.

For the three-subtype model, 93% (96/103) of cases showed subtype assignment stability (**Table 5.6**); five cases switched from the *Subcortical* subtype to the *Fronto-parietal* subtype (all CBS-IDT and negative for AD biomarkers) at follow-up, and two switched from the *Fronto-parietal* to the *Parieto-occipital* subtype (one was CBS-AD, and the other CBS-IDT). The average probability of subtype assignment at baseline was slightly lower than the two-subtype model; 0.87 (SD  $\pm$  0.2) for the *Subcortical* subtype, 0.81 (SD  $\pm$  0.1) for the *Fronto-parietal* subtype and 0.79 (SD  $\pm$  0.2) for the *Parieto-occipital* subtype.

**Table 5.5 - Longitudinal consistency of subtype assignments for two-subtype model.**

Classification previous visit	Classification follow-up visit		
	Normal appearing <sup>a</sup>	Subcortical subtype	Fronto-parieto-occipital subtype
Normal appearing <sup>a</sup>	10 (91%)	0 (0%)	<b>1 (9%)<sup>b</sup></b>
Subcortical subtype	0 (0%)	<b>36 (100%)<sup>b</sup></b>	0 (0%)
Fronto-parieto-occipital subtype	0 (0%)	2 (3%)	<b>64 (97%)<sup>b</sup></b>

<sup>a</sup> Normal appearing = not subtypable. Note that this only includes 11 individuals that were not subtypable at baseline and had a follow-up scan. An observation is longitudinally consistent (<sup>b</sup>) if individuals remain in the same group or progress from the normal-appearing group to a SuStaln subtype at follow-up visit. Entries indicate the number of visits n, with the % of the total individuals in classification at previous visit in classification at follow-up in brackets. Longitudinally consistent observations highlighted in bold.

**Table 5.6 - Longitudinal consistency of subtype assignments for three subtype model**

Classification previous visit	Classification follow-up visit			
	Normal appearing <sup>a</sup>	Subcortical subtype	Fronto-parietal subtype	Parieto-occipital subtype
Normal appearing <sup>a</sup>	10 (91%)	0 (0%)	0 (0%)	<b>1 (9%)<sup>b</sup></b>
Subcortical subtype	0 (0%)	<b>21 (80%)<sup>b</sup></b>	5 (20%)	0 (0%)
Fronto-parietal subtype	0 (0%)	0 (0%)	<b>53 (96%)<sup>b</sup></b>	2 (4%)
Parieto-occipital subtype	0 (0%)	0 (0%)	0 (0%)	<b>21 (100%)<sup>b</sup></b>

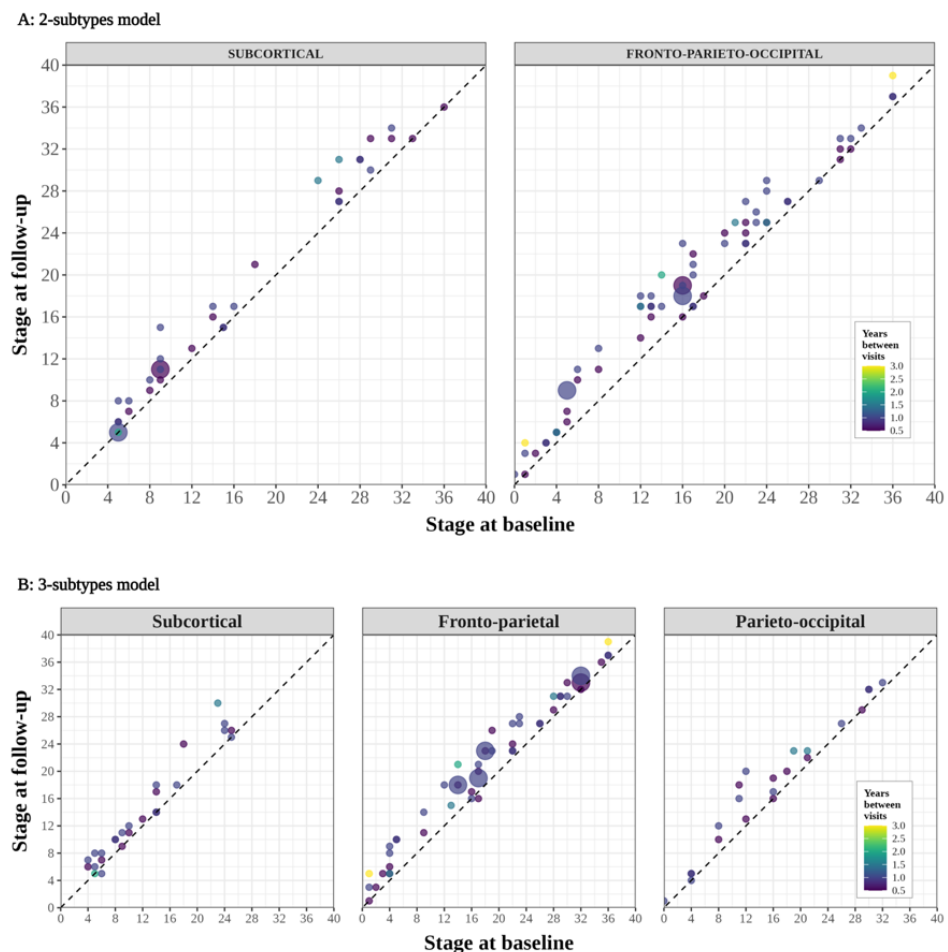
<sup>a</sup> Normal appearing = not subtypable. Note that this only includes 11 individuals that were not subtypable at baseline and had a follow-up scan. An observation is longitudinally consistent (<sup>b</sup>) if individuals remain in the same group or progress from the normal-appearing group to a SuStaln subtype at follow-up. Entries indicate the number of visits n, with the % of the total individuals in classification at previous visit in classification at follow-up in brackets. Longitudinally consistent observations highlighted in bold.



### 5.3.3.2 Individuals consistently progressed to higher stages at follow-up

In the two-subtype model a 100% of subtypable individuals either stayed at the same stage (15%, 15/103) or progressed to a higher stage (85%, 88/103) (**Figure 5.4A**). The *Fronto-parieto-occipital* subtype had a slightly higher percentage progressing to a higher stage at follow-up (88%, 59/67) compared to the *Subcortical* subtype (81%, 29/36).

In the three-subtype model 98% stayed at the same stage or progressed (11%, 11/103 and 87%, 90/103 respectively) (**Figure 5.4B**). 2 individuals (2%) (both CBS-CBD, one assigned to the *Fronto-parietal* and one assigned to the *Subcortical* subtype) dropped one stage at follow-up.



**Figure 5.4 - Stage progression at follow-up visits by SuStaIn subtype.** Scatter plots of each subtype for (A) the two-subtype model (B) the three-subtype model showing predicted stage at baseline (x-axis) versus predicted stage at follow-up scan (y-axis) for those subtypable CBS cases with a follow-up scan ( $n = 103$ ). The area of the circle is weighted by the number of scans at each point, and the colour of the circle represents the time (years) between visits.

### 5.3.4 Subtypes were differentially enriched for underlying CBS pathologies though there was minimal difference in clinical characteristics between subtypes

I then compared the subtype assignments for the different pathological groupings to assess whether SuStaIn imaging subtypes provided information on underlying pathology. The subtypes in both the two- and three- subtype models were differentially enriched for CBS pathology groupings.

In the two-subtype model the *Subcortical* subtype appears to be associated with four-repeat Tau (4RT) pathology and the *Fronto-parieto-occipital* subtype with AD pathology (**Figure 5.2C**). 83.4% of CBS-PSP cases (5/6) and 75% of the CBS-CBD cases (9/12) were assigned to the *Subcortical* subtype, whereas 80.6% of CBS-AD cases (25/35) were assigned to the *Fronto-parieto-occipital* subtype (**Table 5.7**). CBS cases with no information on underlying pathology (CBS-IDT) were relatively evenly split between the two subtypes (49% assigned to *Subcortical* vs. 52% to *Fronto-parieto-occipital* subtype). There was little difference in baseline demographic and clinical scores between the two subtypes. When looking at regional unadjusted baseline volumes in the two-subtype model (**Table 5.8**) the *Fronto-parieto-occipital* subtype had significantly lower mean volumes in the temporal, parietal, occipital lobes compared to the *Subcortical* subtype. In contrast the *Subcortical* subtype had significantly lower volumes in the midbrain, pons, SCP, dentate and the ventral diencephalon.

In the three-subtype model, the addition of a third subtype appears to start to separate CBS-CBD from CBS-PSP pathology with CBS-AD pathology predominantly assigned to the third subtype (**Figure 5.3C**). In those with CBS-CBD, 82% (10/12) were assigned to the *Fronto-parietal* subtype with the remainder assigned to the *Subcortical* subtype, whilst in CBS-PSP 83% are assigned to the *Subcortical* subtype and 17% to the *Fronto-parietal* subtype. Neither of the CBS-4RT pathologies (PSP and CBD) were assigned to the *Parieto-occipital* subtype. In contrast the majority of the CBS-AD cases were assigned to the *Parieto-occipital* subtype (68%, 21/31) with 22% (7/31) assigned to the *Fronto-parietal* subtype and 10% (3/31) assigned to the *Subcortical* subtype (**Table 5.9**). Comparing regional unadjusted baseline volumes in the three-subtype model (**Table 5.10**) the *Subcortical* subtype has the lowest volumes

of the three subtypes in the midbrain, SCP, pons, dentate and the ventral diencephalon, whilst the *Parieto-occipital* subtype had the lowest volumes in the temporal, parietal, occipital lobes and the hippocampus. The *Fronto-parietal* subtype showed a similar pattern to the Subcortical subtype relative to the *Parieto-occipital* subtype (i.e. lower volumes in the midbrain, SCP, pons, dentate and ventral diencephalon / higher volumes in the temporal, parietal, occipital and hippocampal regions) though less severe. The *Fronto-parietal* subtype had the lowest volumes in the caudate, basal ganglia and putamen of the 3 subtypes though this did not reach statistical significance.

**Table 5.7 - Comparison of demographics, pathological diagnosis and clinical test scores between subtypes (two-subtype model).**

	Subcortical	Fronto-parieto-occipital	P value
All scans, n	62 (45.9)	73 (54.8)	-
Subtypable scans, n	56 (45.5)	67 (54.5)	0.77 <sup>a</sup>
Average subtype probability <sup>b</sup>	0.92 (0.1)	0.94 (0.1)	0.33
Sex, % female	50%	55%	0.56
Age first scan, y	68.3 (7.9)	65.4 (7.2)	0.03 <sup>c</sup>
Age at first symptom, y <sup>d</sup>	64.0 (9.3)	60.3 (7.7)	0.06
Disease duration, y <sup>d, e</sup>	4.4 (2.7)	5.1 (2.8)	0.18
CBS pathology, n			-
- CBS-CBD	9 (75%)	3 (25%)	-
- CBS-PSP	5 (83%)	1 (17%)	-
- CBS-AD	6 (19%)	25 (81%)	-
- CBS-IDT	36 (49%)	38 (51%)	<0.001 <sup>f</sup>
PSP rating scale	27.8 (13.6)	24.8 (14.6)	0.31
SEADL	58.5 (22.7)	55.9 (28.5)	0.62
UPDRS	33.2 (17.7)	31.1 (17.5)	0.55
MMSE	23.8 (4.9)	23.5 (7.0)	0.82

Values are mean (SD) or n (%), apart from Sex = % female. Pairwise comparisons between groups were performed using t tests for continuous variables and  $\chi^2$  tests for categorical variables. <sup>a</sup>all scans vs. subtypable scans. <sup>b</sup>subtype probability = the probability of assignment for an individual case to given subtype. <sup>c</sup>statistically significant at  $p < 0.05$ , uncorrected for multiple comparisons. <sup>d</sup>note incomplete data for disease duration / age at first symptom. <sup>e</sup>time from first symptom to first scan. <sup>f</sup>statistically significant at  $p < 0.05$ , corrected for multiple comparisons. Abbreviations: CBS = corticobasal syndrome, CBD = corticobasal degeneration, PSP = progressive supranuclear palsy, AD = Alzheimer's disease, IDT = pathology indeterminate, SEADL = Schwab and England Activities of Daily Living, UPDRS = Unified Parkinson's Disease Rating Scale, MMSE = Mini-Mental State Examination

**Table 5.8 - Regional brain volumes by subtype in the two-subtype model**

Region	Subcortical	Fronto-parieto-occipital	<i>p</i> value
Frontal Anterior	128935 (13578)	126452 (16259)	0.36
Frontal Posterior	34150 (4371)	32601 (4930)	0.07
Temporal	117607 (10436)	109989 (13026)	4.6 x 10 <sup>-4a</sup>
Parietal	85183 (9049)	76286 (10901)	2.5 x 10 <sup>-6a</sup>
Occipital	68530 (7469)	63927 (8536)	0.02
Cingulate	26734 (2915)	26268 (3409)	0.41
Insula	9525 (1255)	9193 (1316)	0.16
Amygdala	3301 (339)	3114 (372)	4.3 x 10 <sup>-3a</sup>
Corpus callosum	18084 (2796)	17992 (2673)	0.85
Medulla	4597 (646)	4809 (590)	0.06
Pons	12703 (1634)	13984 (1600)	2.7 x 10 <sup>-5a</sup>
SCP	196 (45)	230 (47)	5.4 x 10 <sup>-5a</sup>
Midbrain	5172 (759)	5778 (703)	1.3 x 10 <sup>-5a</sup>
Ventral DC	7837 (857)	8438 (973)	3.9 x 10 <sup>-4a</sup>
Thalamus	9727 (1002)	10110 (926)	0.03
Caudate	6248 (821)	6177 (901)	0.65
GP	3367 (349)	3383 (373)	0.81
Putamen	7759 (726)	7557 (811)	0.15
Cerebellar_cortex	86575 (8384)	88222 (8993)	0.30
Dentate	2699 (376)	3023 (399)	9.4 x 10 <sup>-6a</sup>
Hippocampus	7361 (768)	7187 (662)	0.19
Vermis	4359 (532)	4418 (491)	0.52
NA	1051 (111)	1042 (126)	0.68
Basal forebrain	982 (98)	975 (106)	0.73

Values are mean volume (mm<sup>3</sup>) for that brain region (SD). Two-tailed, unpaired *t* tests performed. <sup>a</sup> statistically significant at *p* < 0.05, corrected for multiple comparisons (Bonferroni). Abbreviations: SCP = superior cerebellar peduncles, DC = diencephalon, GP = globus pallidus, NA = nucleus accumbens

**Table 5.9 - Comparison of demographics, pathological diagnosis and clinical test scores between subtypes (three-subtype model)**

	Subcortical	Fronto-parietal	Parieto-occipital	<i>p</i> value
All scans, n	43 (32%)	62 (46%)	30 (22%)	-
Subtypable scans, n	38 (31%)	56 (46%)	28 (23%)	0.77 <sup>a</sup>
Average subtype probability <sup>b</sup>	0.87 (0.2)	0.81 (0.1)	0.79 (0.2)	0.07
Sex, % female	58%	48%	47%	0.36
Age first scan, y	68.5 (6.6)	66.3 (8.4)	64.9 (7.2)	0.15
Age at first symptom, y <sup>d</sup>	63.3 (7.2)	62.8 (10.1)	58.6 (7.2)	0.15
Disease duration, y <sup>d, e</sup>	5.0 (3.1)	4.5 (2.7)	4.9 (2.5)	0.71
CBS pathology, n				-
- CBS-CBD	2 (17%)	10 (83%)	0 (0%)	-
- CBS-PSP	5 (83%)	1 (17%)	0 (0%)	-
- CBS-AD	3 (10%)	7 (22%)	21 (68%)	-
- CBS-IDT	28 (38%)	38 (52%)	7 (10%)	<0.05 <sup>f</sup>
PSP rating scale	28.5 (13.7)	26.0 (15.3)	24.3 (12.6)	0.55
SEADL	57.8 (22.2)	59.8 (25.7)	51.2 (31.1)	0.43
UPDRS	35.0 (18.5)	30.6 (15.8)	31.5 (20.1)	0.58
MMSE	23.9 (4.6)	25.3 (4.5)	20.1 (8.8)	<0.05 <sup>f</sup>

Values are mean (SD) or n (%), apart from Sex = % female. Group comparisons were performed using a linear model for continuous variables (continuous variable ~ SuStaIn subtype) and  $\chi^2$  tests for categorical variables. <sup>a</sup> all scans vs. subtypable scans. <sup>b</sup> subtype probability = the probability of assignment for an individual case to given subtype. <sup>c</sup> statistically significant at  $p < 0.05$ , uncorrected for multiple comparisons. <sup>d</sup> note incomplete data for disease duration / age at first symptom. <sup>e</sup> time from first symptom to first scan. <sup>f</sup> statistically significant at  $p < 0.05$ , corrected for multiple comparisons. Abbreviations: CBS = corticobasal syndrome, CBD = corticobasal degeneration, PSP = progressive supranuclear palsy, AD = Alzheimer's disease, IDT = pathology indeterminate, SEADL = Schwab and England Activities of Daily Living, UPDRS = Unified Parkinson's Disease Rating Scale, MMSE = Mini-Mental State Examination

**Table 5.10 - Regional brain volumes by subtype in the three-subtype model**

Region	Subcortical	Fronto-parietal	Parieto-occipital	<i>p</i> value
Frontal Anterior	131521 (11495)	124941 (14110)	127361 (20123)	0.12
Frontal Posterior	34650 (4185)	32087 (3831)	33841 (6407)	0.03
Temporal	119008 (9222)	112067 (10712)	108583 (16717)	1.6 x 10 <sup>-3a</sup>
Parietal	87136 (7631)	79395 (8404)	72733 (13935)	1.4 x 10 <sup>-7a</sup>
Occipital	69087 (7654)	66254 (7431)	61235 (9222)	5.8 x 10 <sup>-4a</sup>
Cingulate	27080 (2484)	26101 (3154)	26437 (4053)	0.35
Insula	9807 (1109)	9129 (1192)	9136 (1594)	0.03
Amygdala	3358 (272)	3112 (337)	3156 (474)	4.4 x 10 <sup>-3a</sup>
Corpus callosum	18127 (2414)	17642 (2575)	18774 (3287)	0.20
Medulla	4507 (644)	4794 (616)	4820 (570)	0.051
Pons	12513 (1471)	13700 (1834)	13985 (1435)	4.3 x 10 <sup>-4a</sup>
SCP	182 (40)	225 (47)	237 (43)	1.2 x 10 <sup>-6a</sup>
Midbrain	5074 (756)	5596 (720)	5894 (722)	1.4 x 10 <sup>-5a</sup>
Ventral DC	7738 (845)	8116 (852)	8843 (1010)	1.1 x 10 <sup>-5a</sup>
Thalamus	9686 (932)	9900 (979)	10342 (950)	0.02
Caudate	6244 (740)	6123 (862)	6337 (1031)	0.56
GP	3414 (323)	3292 (352)	3491 (403)	0.043
Putamen	7893 (681)	7449 (722)	7723 (926)	0.020
Cerebellar cortex	86043 (7531)	86829 (9419)	90988 (8156)	0.05
Dentate	2652 (416)	2893 (355)	3142 (397)	6.8 x 10 <sup>-6a</sup>
Vermis	4378 (500)	4330 (543)	4540 (439)	0.20
Hippocampus	7579 (686)	7186 (728)	7003 (597)	2.3 x 10 <sup>-3a</sup>
NA	1068 (99)	1016 (116)	1076 (141)	0.04
Basal forebrain	981 (93)	978 (93)	979 (132)	0.99

Values are mean volume (mm<sup>3</sup>) for that brain region (SD). Group comparisons were performed using a linear model for continuous variables (continuous variable ~ SuStaIn subtype) and  $\chi^2$  tests for categorical variables. <sup>a</sup> statistically significant at  $p < 0.05$ , corrected for multiple comparisons (Bonferroni). Abbreviations: SCP = superior cerebellar peduncles, DC = diencephalon, GP = globus pallidus, NA = nucleus accumbens

### 5.3.5 Association between stage, subtype, and clinical disease severity

I went on to assess the association between stage, subtype, and clinical disease severity in the both the two- and three- subtype model, controlling for age and gender.

In the two-subtype model (**Table 5.11**) only the Gait and Midline PSP rating scale subscore was different between the *Subcortical* and *Fronto-parieto-occipital* subtype (worse in the *Subcortical* subtype:  $t = -2.04$ ,  $p = 0.04$ ). Worsening Total PSP rating scale score (and History, Bulbar and Oculomotor subscores), and MMSE score were associated with increasing SuStai stage, suggesting these scores decline with disease progression in both subtypes.

In the three-subtype model (**Table 5.12**) the main difference to the two-subtype model was that there was no-longer a significant difference in Limb motor subscores between the subtypes, while significant differences between performance on the MMSE became apparent in the *Parieto-occipital* subtype ( $t = -3.11$ ,  $p = 2.4 \times 10^{-3}$ ).

**Table 5.11 – Comparison of adjusted clinical scores between subtypes in the two-subtype model**

	SuStaIn subtype		SuStaIn stage		Subtype with worse score	Change with Sustain Stage
	t value	p value	t value	p value		
PSP rating scale score						
- <i>Total</i>	-0.63	0.27	2.32	0.02 <sup>a</sup>		<i>Worsens</i>
- <i>History</i>	-1.11	0.78	1.99	0.04 <sup>a</sup>		<i>Worsens</i>
- <i>Mentation</i>	0.61	0.55	1.38	0.17		
- <i>Bulbar</i>	-0.35	0.72	4.00	1 x 10 <sup>-4b</sup>		<i>Worsens</i>
- <i>Ocular motor</i>	-0.62	0.54	2.46	0.02 <sup>a</sup>		<i>Worsens</i>
- <i>Limb motor</i>	0.13	0.89	-0.25	0.80		
- <i>Gait and midline</i>	-2.04	0.04 <sup>a</sup>	0.34	0.73	<i>Subcortical subtype</i>	
SEADL	-0.31	0.75	-0.94	0.34		
UPDRS	-0.01	0.99	0.88	0.38		
MMSE	-0.19	0.85	-4.20	5 x 10 <sup>-5b</sup>		<i>Worsens</i>

Linear mixed model of Clinical score ~ subtype + stage + (1 | ID) + AAS + gender. Significance was calculated using Satterthwaite's method to estimate degrees of freedom and generate p-values. Includes 226 scans (123 baseline and 103 follow-up scans and varying timepoints). <sup>a</sup> statistically significant at  $p < 0.05$ , uncorrected for multiple comparisons (10 items,  $p < 0.0125$ ). <sup>b</sup> statistically significant at  $p < 0.05$ , corrected for multiple comparisons (10 items,  $p < 0.0125$ ). Abbreviations: FPO = fronto-parieto-occipital, SEADL = Schwab and England Activities of Daily Living, UPDRS = Unified Parkinson's Disease Rating Scale, MMSE = Mini-Mental State Examination.



**Table 5.12 – Comparison of adjusted clinical scores between subtypes in the three-subtype model**

	SuStaIn subtype (Fronto-Parietal*)		SuStaIn subtype (Parieto-Occipital*)		SuStaIn stage		Subtype with worse score	Change with Sustain Stage
	<i>t</i> value	<i>p</i> value	<i>t</i> value	<i>p</i> value	<i>t</i> value	<i>p</i> value		
PSP rating scale score								
- <i>Total</i>	-1.24	0.22	-0.83	0.41	2.46	0.01 <sup>a</sup>		<i>Worsens</i>
- <i>History</i>	-0.97	0.34	-0.91	0.37	2.02	0.05 <sup>b</sup>		<i>Worsens</i>
- <i>Mentation</i>	-0.57	0.57	1.42	0.16	1.44	0.15		
- <i>Bulbar</i>	-0.05	0.96	-1.14	0.25	3.68	3.7 x 10 <sup>-4a</sup>		<i>Worsens</i>
- <i>Ocular motor</i>	-1.39	0.17	-1.58	0.12	2.72	7.6 x 10 <sup>-3a</sup>		<i>Worsens</i>
- <i>Limb motor</i>	-0.03	0.98	-0.51	0.61	0.22	0.83		
- <i>Gait and midline</i>	-1.58	0.11	-1.07	0.29	1.72	0.09		
SEADL	0.44	0.66	-1.16	0.25	-1.13	0.26		
UPDRS	-0.75	0.46	-0.24	0.81	1.05	0.30		
MMSE	1.35	0.18	-3.11	2.4 x 10 <sup>-3a</sup>	-3.79	2.4 x 10 <sup>-4a</sup>	<i>Parieto-occipital</i>	<i>Worsens</i>

Linear mixed model of Clinical score ~ subtype + stage + (1 | ID) + AAS + gender. Significance was calculated using Satterthwaite's method to estimate degrees of freedom and generate p-values. Includes 226 scans (123 baseline and 103 follow-up scans at varying timepoints). \*Named SuStaIn subtype compared to Subcortical subtype. <sup>a</sup>. Statistically significant at  $p < 0.05$ , corrected for multiple comparisons (10 items,  $p < 0.0125$ ). <sup>b</sup>. Statistically significant at  $p < 0.05$ , uncorrected for multiple comparisons (10 items,  $p < 0.0125$ ). Abbreviations: SEADL = Schwab and England Activities of Daily Living, UPDRS = Unified Parkinson's Disease Rating Scale, MMSE = Mini-Mental State Examination

## 5.4 Discussion

Clinical-pathological studies have emphasised the pathological heterogeneity of CBS with diverse diseases such as CBD, AD, PSP and TDP-43 proteinopathy presenting clinically as CBS<sup>418</sup>. Accurately identifying the underlying pathology in CBS in life, as well as accurately measuring individuals' disease progression to assess the effectiveness of therapeutic interventions, is increasingly important as we enter the era of protein-specific therapies for neurodegenerative diseases. Although a range of validated CSF, PET and blood-based biomarkers are now available for identifying AD pathology<sup>419</sup>, there are still no such biomarkers available for 4RT or TDP-43 pathology. To be able to both subtype and stage patients with CBS using only a structural MRI scan would have significant implications for clinical trial stratification and disease progression monitoring.

To address this, I applied an unsupervised machine learning algorithm (SuStaIn) to a relatively large cohort of clinically diagnosed CBS cases, uncovering imaging subtypes based solely on a data-driven assessment of cross-sectional atrophy patterns. I then used the fitted model to subtype and stage individuals both at baseline and follow-up, demonstrating an association between clinical measures of disease progression and increasing data-driven model stage. I also showed that subtype assignment was stable at follow-up scan, and individuals consistently progressed to higher stages, supporting the model's longitudinal validity.

Prior studies have retrospectively assessed both structural<sup>137,233,235</sup> and FDG-PET imaging<sup>420</sup> at a group level, as correlates of CBS pathology. Three of these studies<sup>137,233,420</sup> take no account of disease stage in their analysis and so are limited by the inherent assumption that all subjects are at a common disease stage (no temporal heterogeneity). The study by Whitwell et al.<sup>235</sup> uses the MMSE score as a proxy for disease stage, which although an improvement, assumes that not only is the MMSE score similarly affected across the different pathologies for a given stage of disease, but that it is sensitive and specific for marker of disease progression. In addition, none of these clinico-pathological studies include longitudinal imaging follow-up and provide little information on the earliest regions in the brain affected by disease within the different pathological subtypes. By using SuStaIn to jointly model both disease stage and subtype simultaneously, I am able to better account for this temporal

heterogeneity, highlighting the regions that are affected earliest in the disease course for each imaging subtype, whilst also providing a fine-grained staging model within each subtype that allowed staging of individual patients.

It is important to note that the model was agnostic to underlying pathology, and I only used the pathology information post-hoc, to test the hypothesis that these imaging subtypes would provide information on the underlying pathology. In support of this hypothesis, the subtypes were differentially associated with underlying pathology; the data best supported a two-subtype model, with 4RT (PSP or CBD) confirmed cases being most commonly assigned to the *Subcortical* subtype (83% of PSP and 75% of CBD respectively), and AD cases being most commonly assigned to the *Fronto-parieto-occipital* subtype (81% of cases). The *Subcortical* subtype (46% of cases) was characterised by early atrophy of the SCP, midbrain and dentate nucleus, followed by the basal ganglia, remaining brainstem structures and the thalamus, with the posterior frontal lobe the first cortical structure to become abnormal. This early involvement of the brainstem and subcortical structures in CBS-4RT is in keeping with previous work that shows that more severe atrophy is found in these regions in CBS-PSP and CBS-CBD compared to controls and CBS-AD<sup>137</sup>. In contrast, the *Fronto-parieto-occipital* subtype demonstrates earliest atrophy in the parietal region closely followed by the posterior frontal, insular and occipital lobes. The basal ganglia, similar to the *Subcortical* subtype, are involved early in the sequence, as one might expect given these individuals have presented with a cortico-basal syndrome. The fact that AD pathology is strongly assigned to this subtype is also in keeping with published clinico-pathological imaging studies, where CBS-AD demonstrates the most severe atrophy in the parietal and posterior frontal regions<sup>137,233,235</sup>.

The two-subtype model best explained the data in this cohort, as evidenced by the cross-validation log likelihoods and CVIC in **Figure 5.1**. The three-subtype model was underpowered with several of the different subtype stages only having a single individual assigned. Despite these caveats, further analysis of the three-subtype model was interesting. There was a suggestion that adding a third subtype allowed differentiation of PSP from CBD pathology, albeit at a loss of specificity for AD pathology. Given the availability of sensitive and specific AD biomarkers, this may allow for identification of these cases that do not map to the most “AD-like” subtype, thus enriching the other subtypes for 4RT pathology. PSP pathology was still strongly

assigned to the *Subcortical* subtype (83.3% of cases), though 75% of CBD cases were now assigned to the new *Fronto-parietal* subtype. Neither 4RT pathologies were assigned to the *Parieto-occipital* subtype, which had a very similar sequence of atrophy to the *Fronto-parieto-occipital* subtype from the two-subtype model. 68% of AD pathology was assigned to this *Parieto-occipital* subtype, with 23% assigned to the new *Fronto-parietal* subtype. The sequence of atrophy on the *Fronto-parietal* subtype demonstrates earliest involvement of the posterior frontal lobe and the basal ganglia with early involvement of the parietal and insula, which is consistent with imaging in autopsy confirmed CBD cases<sup>137,233,235</sup>. Interestingly this subtype also showed later involvement of the temporal lobe compared to the *Parieto-occipital* subtype, another feature that has been shown to differentiate CBS-CBD from CBS-AD<sup>235</sup>. In keeping with the *Parieto-occipital* subtype being more strongly associated with AD pathology, analysis of regional volumes at baseline demonstrated that the hippocampal and temporal (as well as parietal and occipital) regions were more atrophic compared to the *Fronto-parietal* subtype at presentation (**Table 5.10**). Further support for this is that the MMSE was significantly lower in the *Parieto-occipital* subtype (20.1, SD  $\pm$  8.8,  $t = -2.3$ , Bonferroni corrected  $p = 0.02$ ) compared to the other subtypes (23.9 SD  $\pm$  4.6, 25.3 SD  $\pm$  4.9 for the *Subcortical* and *Fronto-parietal* subtypes respectively).

When comparing clinical scores between subtypes there was minimal difference; in the two-subtype model only the Limb-motor PSP rating scale sub-score was different (lower in the *Subcortical* subtype), whilst as mentioned above only the MMSE showed a difference between subtypes in the three-subtype model (lowest in the *Parieto-occipital* subtype). This is perhaps unsurprising given the lack of clinical difference between the different pathology groups at baseline (**Table 5.4**), and the association of these pathologies with different subtypes identified by the model. The lack of clinical features that accurately identify the different underlying pathologies in CBS is well known<sup>409</sup> and is likely a reflection of the similar spatial patterns of underlying pathology (whatever that pathology may be) leading to a similar constellation of clinical symptoms and signs. It is suggested that more fine-grained cognitive tests, such as the Benson figure copy<sup>137</sup>, may be more sensitive to underlying AD pathology, but I unfortunately did not have access to these more detailed cognitive assessments

for this cohort. Instead, for cognitive scores I had to rely on the MMSE (or converted MOCA to MMSE) with the associated loss in diagnostic sensitivity.

Overall, the fitted SuStaIn model showed strong subtyping and staging capabilities. In the two-subtype model, assignments were longitudinally consistent at 101 out of 103 (98%) of follow-up visits with 101 individuals staying within the same subtype and one individual progressing from not-subtypable to subtypable. The two individuals who changed from the *Fronto-parietal-occipital* subtype to *Subcortical* at follow-up were only weakly assigned at baseline (0.43 and 0.58). From a staging perspective, individuals consistently moved to higher stages over time in both subtypes, with no cases dropping to a lower stage at follow-up scan. As expected, the three-subtype model the subtypes were slightly less stable, which likely reflects the increased uncertainty in assignment due to lower sample sizes in each cohort. 96 of the 103 (93%) follow-up visits were assigned to the same subtype; five individuals in the *Subcortical* subtype switched to the *Fronto-parietal* subtype (only one of these cases had pathology and was CBD), while two individuals in the *Fronto-parietal* subtype switched to the *Parieto-occipital* subtype (both had AD pathology). An interesting observation was the variability in model predicted stage of CBS cases at baseline scan seen in both the two- and three-subtype model (**Figure 5.2D** and **Figure 5.3D**). This baseline temporal variability has important implications in terms of accurately stratifying patients at time of recruitment to clinical trials. The SuStaIn algorithm provides an intrinsic subtyping and staging mechanism for individual patients which may help with this process.

### 5.4.1 Limitations

An important limitation of this study is that, although I built a large imaging cohort from the perspective of CBS (135 cases with 113 follow-up visits), this is still small for a suitably powered SuStaIn analysis. In this study I decided to combine regions with a right and left label, and also to then select features based on Cohen's  $d$  greater than 0.6, to try and reduce the number of features included in the model, and so maximise power to detect subtypes with the available sample size. It is known that CBS-CBD, in particular, is characterised by asymmetric atrophy, at least later in the disease course<sup>235,418</sup>. However, there are also volumetric imaging studies demonstrating that many patients with CBD fall within the range of asymmetry seen

in healthy controls<sup>137</sup>. There is also a growing recognition from neuropathological studies that a lack of asymmetry does not exclude an diagnosis of underlying CBD pathology<sup>66</sup>. By combining right and left labels for cortical regions I am likely to have reduced the sensitivity for detecting a “CBD” like subtype in particular, as the effect size for a given region affected by CBD pathology would be diluted by the less severe atrophy in the contralateral hemisphere. It is interesting that the three-subtype model hinted at a third subtype in the data, associated with CBD pathology, even with right and left cortical labels combined. In future work I would like to expand the CBS imaging cohort through collaboration with additional sites, to investigate whether adding right and left labels separately increases the sensitivity to identify finer grained pathological specific subtypes. As it stands the two-subtype model is likely a conservative estimate of the true number of unique imaging subtypes in CBS.

A related, but separate issue is the lack of pathology or amyloid biomarker data for 74 of the cases (categorised as CBS-IDT). Although the focus of this study was to identify CBS imaging subtypes and stages a priori, I wanted to test post-hoc the assignment of the different pathologies to the these identified subtypes to test the hypothesis that joint modelling of disease stage and subtype would provide additional information on underlying pathology. The difficulty of interpreting these results is compounded by the fact that I had no data on TDP-43 pathology, which is known to account for up to 14% of cases of CBS<sup>66</sup>. It is an interesting observation that of the 12 cases that were not subtypable at baseline, three cases were CBS-AD, nine were CBS-IDT, and of these six had negative biomarkers for AD and so were pathology other than AD. One might speculate that given all of the cases with 4RT pathology were subtypable that these un-subtypable cases could have a different underlying pathology such as TDP-43. A good test of the pathology association with subtype will be testing whether those that come to post-mortem in the future match the expected pathology based on their subtype assignment. Another potential for follow-up work would be to subtype and stage new autopsy confirmed cases with our fitted model and assess whether they are assigned to the expected phenotypes.

It is increasingly recognized that co-pathologies are common in neurodegenerative disease and may modify clinical presentations<sup>66</sup>. For example, concurrent AD pathology in PSP may modify the clinical presentation from PSP-Richardson syndrome to that of CBS<sup>421</sup>. Given I did not have data on the presence or absence of

co-pathologies in this study, one has to be careful when drawing conclusions based on the primary pathological diagnosis. Conversely, it could perhaps be argued that one of the explanations for the assignment of a minority of cases with a given pathology to subtype more strongly associated with a different pathology (e.g. in the two-subtype model CBS-4RT cases being assigned to *Fronto-parieto-occipital* subtype, or in the three-subtype model CBS-AD cases being assigned to the *Subcortical* or *Fronto-parietal* subtype) could be due to the presence of co-pathologies that are “unseen” in our data. It would be interesting to stage and subtype individuals in a validation cohort with well-defined pathology (including co-pathologies) to test this hypothesis.

## 5.5 Conclusions and future work

In this chapter I provide data-driven evidence for the existence of two distinct and stable spatiotemporal subtypes of atrophy in clinically diagnosed CBS, by jointly modelling disease stage and subtype using cross-sectional structural MRI. Underlying CBS pathology is differentially associated with these subtypes giving insights into the relationship between CBS pathology and the topographical distribution of atrophy. In addition, our model provides an intrinsic staging and subtyping mechanism by which individual patients can be more accurately stratified according to disease stage within each subtype. In the absence of sensitive and specific biomarkers for the range of different pathologies in CBS, being able to accurately subtype and stage CBS patients at baseline has important implications for screening patients on entry into clinical trials, as well as for tracking disease progression.

Future work should focus on validating these results in larger datasets, ideally with detailed pathological data, to investigate whether (as suggested in this work) there is a third imaging subtype that is able to differentiate between the 4R tau pathologies. In addition, it will also be important to test newer iterations of the SuStaIn algorithm under development with the aim of extracting information on time to transition between subtype stages, which will have direct utility in disease progression monitoring for clinical trials.

# Chapter 6: Investigating the association of *MAPT* haplotypes with risk of Pick's Disease

The contents of this chapter form part of a manuscript which is currently in preparation for submission.

## 6.1 Introduction

Pick's disease (PiD) is a rare and predominantly sporadic subtype of frontotemporal lobar degeneration (FTLD) and represents approximately 3-12% of all dementias worldwide<sup>422–424</sup>. Although there are no clinical diagnostic criteria for PiD, it typically develops in individuals approximately 55 years of age and presents with behavioural change, impaired cognition and occasionally motor difficulties<sup>53,61,63–65,67,425</sup>. PiD is a relatively rapidly progressive disease and patients die approximately 10 years after disease onset<sup>61–65,238</sup>. Current therapies are available to treat clinical symptoms, but no treatments are available to prevent disease onset or progression.

Neuropathologically, PiD is classified by severe frontotemporal, knife-blade like cortical atrophy and microscopically the presence of ballooned neurons and argyrophilic inclusion “Pick bodies” in frontal and temporal regions<sup>61</sup>. Characteristic Pick bodies consist of hyperphosphorylated 3-repeat (3R) tau aggregate proteins which are encoded by the *MAPT* gene on chromosome 17<sup>425,426</sup>, and for this reason PiD is classified as a 3R tauopathy. *MAPT* codes for six major tau isoforms in the adult human brain, and this is determined by alternative splicing of exons 2, 3, and 10 influencing the number of repeat domains in the N-terminus and C-terminus<sup>427</sup>. More specifically, alternative splicing leading to exon 10 exclusion results in 3-repeat units in the microtubule binding C-terminal domain, generating 3R tau proteins<sup>428</sup>.

Rare mutations in *MAPT* have been reported in four individuals with PiD-like pathology<sup>429–432</sup>; however, this data is inconsistent as larger, independent cohorts of PiD cases do not report *MAPT* mutations<sup>433</sup>. The *MAPT* gene also has two well characterized common haplotypes, H1 and H2, which developed from a 900kb ancestral genetic inversion event<sup>172</sup>. Not only has *MAPT* H1 consistently been associated with an increased risk of 4-repeat (4R) primary tauopathies such as progressive supranuclear palsy (PSP) and corticobasal degeneration (CBD), but the



H1 haplotype is also the strongest genetic risk factor for both diseases<sup>169,193</sup>. To date, no such haplotype association has been shown in 3R tauopathy of PiD and this may be due to the limited available sample size<sup>185,186</sup>.

Due to its rare prevalence and the inability to diagnose it clinically in life, PiD is an understudied neurodegenerative disease, and its genetic etiology is unknown. As previously mentioned, the few studies of *MAPT* haplotype in PiD that have been conducted have been very small and underpowered, failing to find an association of either the *MAPT* H1 or H2 haplotype with disease risk. Moreover, limited access to 3R tauopathy samples has stalled research advancement in understanding how *MAPT* haplotypes and isoforms influence disease risk/pathology and has prevented progression of developing isoform-specific therapies.

To address this, in collaboration with Mayo Clinic Jacksonville (MCJ) I established the Pick's disease International Consortium (PIC), to collect data from pathologically confirmed PiD cases from sites worldwide. My primary aim in this study was to use samples collected via the PIC to evaluate the association of the *MAPT* H1/H2 haplotype with PiD risk, age of onset (AAO), and disease duration (DD). As a secondary aim, I assessed the associations of the *MAPT* H1 subhaplotypes with disease outcomes.

## 6.2 Methods

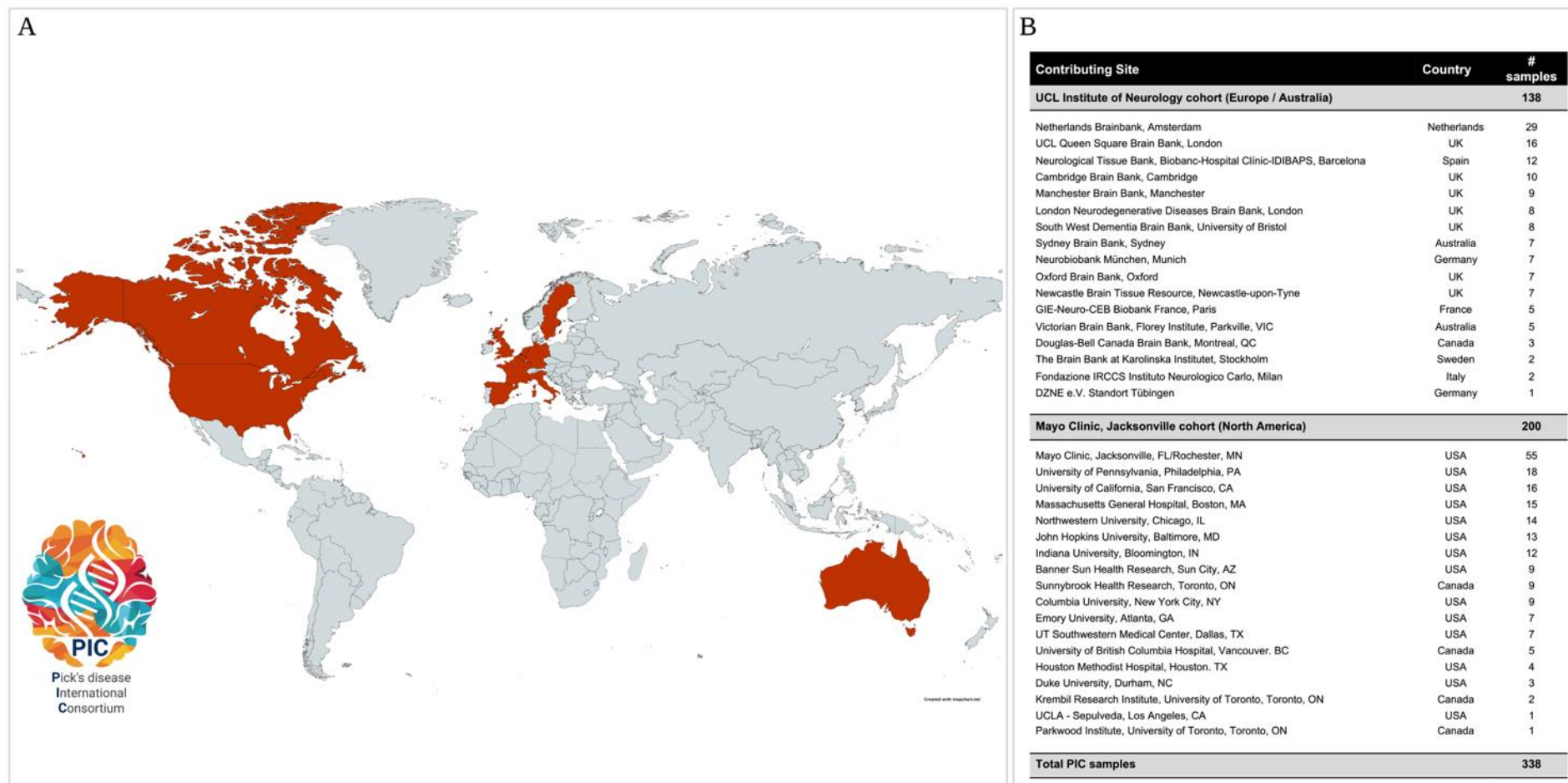
### 6.2.1 Study Subjects

At the time of this study, 338 neuropathologically confirmed PiD cases were available from the PIC, recruited from 36 sites (**Figure 6.1**). Please refer to Chapter 2 (**Pick's disease International Consortium (PIC)**) for details on the neuropathological diagnostic criteria of cases and recruitment. Of the 338 samples, 200 were from the North America cohort (MCJ) and 138 from European/Australasian cohort (UCL). Frozen brain tissue from cerebellum or prefrontal cortex were obtained from each case. All subjects were unrelated and self-reported Caucasian and non-Hispanic. Baseline demographic information was collected for all subjects (age at symptom onset (AAO) and age at death (AAD) for PiD patients, age at blood collection for controls, and sex). Disease duration (DD) was calculated from the difference between AAD and AAO for a subset of 309 PiD cases. In addition to basic demographic information, the PIC also

collected information related to family histories, clinical outcomes (e.g. behavioural and language impairments, presence/absence of parkinsonism, upper and lower motor neurone deficits, Mini-Mental State Examination and Clinical Dementia Rating), and pathological information (e.g. Thal phase, Braak stage, and brain weight,) for each individual case, as well as noting whether other tissues and brain imaging data were available. Cases were removed if a rare *MAPT* mutation was identified. Peripheral blood was provided from 1,312 controls from the Mayo Clinic in Jacksonville, FL (MCJ) (N=881) or Rochester, MN (N=431). Control subjects were deemed neurologically healthy by neurologists at Mayo Clinic.

### **6.2.2 Neuropathological diagnosis of Pick's disease**

As detailed in Chapter 2 (**Neuropathological diagnosis of Pick's disease**) there are no established diagnostic criteria for the neuropathological diagnosis of PiD, and so I developed a diagnostic algorithm as part of the PhD, in collaboration with Professor Dickson and Dr. Shanu Roemer at the MCJ, to ensure that I had a homogenous cohort of cases with known 3R tau PiD pathology (see **PIC diagnostic algorithm**) 365 cases from the PIC were initially identified, and underwent neuropathological assessment as detailed in Chapter 2 (**Neuropathologic methods for 3R/4R tau IHC**). Of these 338 cases were confirmed as PiD neuropathologically according to the algorithm and included in the study.



**Figure 6.1** - Global map (A) and table (B) of reporting countries and recruitment sites that have contributed Pick's disease tissues to the Pick's disease International Consortium (PIC) to date. Note not all pathology confirmed samples included in MAPT analysis due to either ancestry or samples not collected at time of data freeze for analysis Red = countries that have collected and donated Pick's disease tissues.

### 6.2.3 DNA Preparation

For the majority of samples, DNA was extracted from each subject at their respective collection site (UCL or MCJ) as detailed in Chapter 2 (**DNA preparation**). I was responsible for the DNA extraction from all UCL samples, and Dr. Rebecca Valentino for all MCJ samples.

### 6.2.4 *MAPT* SNP Genotyping

The *MAPT* H2 haplotype-tagging variant rs8070723 was genotyped in all cases and controls to associate with disease. In addition, the five common *MAPT* variants (rs1467967, rs242557 [the H1C haplotype-tagging variant], rs3785883, rs2471738, and rs7521) which along with rs8070723 define H1-subhaplotypes were genotyped to assess *MAPT* subhaplotype structure<sup>176,202</sup>. Further details of the genotyping methods used at UCL and MCJ are provided in Chapter 2 (**Targeted genotyping**).

All cases were assessed for population stratification using available whole genome SNP genotyping data. After standard genotyping data quality control steps, I performed a principal components analysis (PCA), merged all cases with the HapMap reference dataset, and identified any cases of non-white European ancestry which were excluded from further analysis. The details of PCA analysis to determine ancestry are covered in Chapter 7 (**Quality control and imputation**). Allele and genotype frequencies for each variant are detailed in **Table 6.1**.

Table 6.1 - Genotype counts and frequencies of six common *MAPT* SNPs in Pick's disease cases and controls.

Variant	Minor allele count and frequency	Major allele count and frequency	Genotype 1 count and frequency	Genotype 2 count and frequency	Genotype 3 count and frequency
<b>rs1467967</b>					
Pick's disease	G: 190 (28.1%)	A: 486 (71.9%)	AA: 176 (52.1%)	AG: 134 (39.6%)	GG: 28 (8.3%)
Controls	G: 849 (32.4%)	A: 1775 (67.6%)	AA: 603 (46.0%)	AG: 569 (43.4%)	GG: 140 (10.7%)
<b>rs242557</b>					
Pick's disease	A: 236 (34.9%)	G: 440 (65.1%)	GG: 146 (43.2%)	GA: 148 (43.8%)	AA: 44 (13.0%)
Controls	A: 965 (36.8%)	G: 1659 (63.2%)	GG: 548 (41.8%)	GA: 563 (42.9%)	AA: 201 (15.3%)
<b>rs3785883</b>					
Pick's disease	A: 114 (16.9%)	G: 562 (83.1%)	GG: 231 (68.3%)	GA: 100 (29.6%)	AA: 7 (2.1%)
- Controls	A: 472 (18.0%)	G: 2152 (82.0%)	GG: 879 (67.0%)	GA: 394 (30.0%)	AA: 39 (3.0%)
<b>rs2471738</b>					
Pick's disease	T: 136 (20.1%)	C: 540 (79.9%)	CC: 215 (63.6%)	CT: 110 (32.5%)	TT: 13 (3.8%)
- Controls	T: 542 (20.7%)	C: 2082 (79.3%)	CC: 842 (64.2%)	CT: 398 (30.3%)	TT: 72 (5.5%)
<b>rs8070723</b>					
Pick's disease	G: 196 (29.0%)	A: 480 (71.0%)	AA: 167 (49.4%)	AG: 146 (43.2%)	GG: 25 (7.4%)
Controls	G: 603 (23.0%)	A: 2021 (77.0%)	AA: 784 (59.8%)	AG: 453 (34.5%)	GG: 75 (5.7%)
<b>rs7521</b>					
Pick's disease	A: 278 (41.1%)	G: 398 (58.9%)	GG: 117 (34.6%)	GA: 164 (34.6%)	AA: 57 (16.9%)
Controls	A: 1223 (46.6%)	G: 1401 (53.4%)	GG: 385 (29.3%)	GA: 631 (48.1%)	AA: 296 (22.6%)

## 6.2.5 Statistical Analysis

Single-variant associations with risk of PiD were evaluated using logistic regression models that were adjusted for age and sex. Odds ratios (ORs) and 95% confidence intervals (CIs) were estimated and correspond to each additional minor allele. Single-variant associations with AAO and DD in PiD patients were examined using linear regression models that were adjusted for sex and series (UCL or MCJ) (AAO analysis) or sex, AAO, and series (DD analysis). Regression coefficients (referred to as  $\beta$ ) and 95% CIs were estimated and are interpreted as the increase in the mean AAO or DD corresponding to each additional copy of the minor allele. For all single-variant associations, analysis involving rs8070723 (the H2-tagging variant) was considered as the primary analysis (to test the hypothesis that it is associated with risk of PiD), with results for the five remaining variants considered as secondary and I present these for completeness.

Associations between six-variant *MAPT* haplotypes and risk of PiD were assessed using score tests for association under a logistic regression framework<sup>434</sup>, where tests were adjusted for age and sex. ORs and 95% CIs were estimated and correspond to each additional copy of the given haplotype. In analysis of PiD patients, associations of six-variant *MAPT* haplotypes with AAO and DD were assessed using score tests for association under a linear regression framework<sup>434</sup>, where tests were adjusted for sex and series (AAO analysis) or sex, AAO, and series (DD analysis).  $\beta$ -coefficients and 95% CIs were estimated and are interpreted as the increase in the mean AAO or DD corresponding to each additional copy of the given haplotype. Haplotypes occurring in <1% of subjects in a specific analysis were excluded from that analysis.

$p$ -values <0.008 (6 tests) were considered as statistically significant after Bonferroni correction in the primary analysis involving the *MAPT* rs8070723 variant. In secondary analysis assessing associations between *MAPT* haplotypes and outcomes,  $p$ -values  $\leq 0.0028$  (18 tests) were considered as statistically significant after Bonferroni correction in the disease-association analysis, and  $p$ -values  $\leq 0.0031$  (16 tests due to two haplotypes below the 1% frequency threshold) were considered as statistically significant in the AAO and DD analyses.  $p$ -values  $\leq 0.05$  were considered as significant in all remaining analysis. All statistical tests were two-sided. Statistical

analyses were performed using R Statistical Software (version 4.1.2; R Foundation for Statistical Computing, Vienna, Austria).

## 6.3 Results

I identified a total of 338 pathologically confirmed PiD cases from the PIC for inclusion in the study. A summary of clinical characteristics is provided in **Table 6.2**. In the combined series the median age at death was 69 years (minimum-maximum; 40-100 years) which was similar in both the North American and European/Australasian series (70 years (40-100 years) and 69 years (41-88 years) respectively). The age at disease onset for the combined series was 58 years (33-80 years) which was consistent across series, with a median disease duration of 10 years (2-25 year). Interestingly there was a higher proportion of males in the combined series (60.7% compared to 39.3% females; this difference was more pronounced in the European series than the North American series (66.7% versus 56.5% males respectively). The controls were well matched from an age perspective, though had a lower proportion of males (46.6%). Both age and gender were controlled for in subsequent regression analysis to minimise the effects of these covariates on results.

**Table 6.2 - Clinical characteristics of samples included in *MAPT* haplotype analysis.**

Variable	MCJ PiD series <sup>1</sup>	UCL PiD series <sup>2</sup>	Combined PiD series	Controls
Sample (n)	200	138	338	1,312
Age (years)	70 (40, 100)	69 (41, 88)	69 (40, 100)	69 (45, 92)
Age of disease onset (years)	58 (36, 80)	58 (33, 80)	58 (33, 80)	n/a
Disease duration (years)	10 (3, 25)	10 (2, 20)	10 (2, 25)	n/a
Sex				
- Male	113 (56.5%)	92 (66.7%)	205 (60.7%)	611 (46.6%)
- Female	87 (43.5%)	46 (33.3%)	133 (39.3%)	701 (53.4%)

<sup>1</sup> North American samples, <sup>2</sup> Europe and Australasia samples. The sample median (minimum, maximum) is given for age. Age represents age at death in Pick's disease cases and age at blood draw in controls. Age at disease onset and disease duration information was not available for 29 PiD cases. Abbreviations: MCJ = Mayo Clinic Jacksonville, UCL = University College London

### 6.3.1 Association results for the *MAPT* rs8070723 H2 allele

#### rs8070723 and PiD risk

There was a significant association between the *MAPT* rs8070723 H2 allele and an increased risk of PiD in the overall series (OR: 1.35, 95% CI: 1.12-1.64,  $p=0.0021$ ) (Table 6.3), with minor allele frequencies (MAFs) of 29.0% in the 338 PiD patients and 23.0% in the 1,312 controls. For comparison the MAF in the non-Finnish European population on GnomAD (RRID:SCR\_014964; <http://gnomad.broadinstitute.org/>) was 19.7%, 19.8% in Southern Europeans and 22.8% in North-western Europeans.

#### rs8070723 and AAO / DD

*MAPT* rs8070723 was not associated with AAO ( $\beta$ : -0.54, 95% CI: -1.94 to 0.87,  $p=0.45$ ) or DD ( $\beta$ : 0.25, 95% CI: -0.46 to 0.96,  $p=0.50$ ) (Table 6.3). Single-variant associations with PiD are also shown for the other five *MAPT* variants used to define *MAPT* haplotypes in Table 6.3 and Table 6.4 respectively. Of note, there was no association between rs242557 (the H1c haplotype defining SNP) and risk of PiD (OR: 0.94, 95% CI: 0.79-1.12,  $p=0.51$ ). None of the other *MAPT* variants showed an association with disease risk, AAO or DD in PiD.



**Table 6.3 - Associations of individual *MAPT* variants with risk of Pick's disease**

Variant	Minor Allele Frequency			Association with PiD	
	PiD (N=338)	Controls (N=1,312)	NFE	OR (95% CI)	p value
rs1467967	28.1%	32.4%	33.5%	0.83 (0.68, 1.00)	0.046
rs242557	34.9%	36.8%	37.5%	0.94 (0.79, 1.12)	0.51
rs3785883	16.9%	18.0%	17.9%	0.91 (0.72, 1.15)	0.42
rs2471738	20.1%	20.7%	22.0%	0.96 (0.78, 1.18)	0.70
rs8070723	29.0%	23.0%	19.7%	1.35 (1.12, 1.64)	<b>0.0021</b>
rs7521	41.1%	46.6%	53.1%	0.81 (0.69, 0.96)	0.018

ORs, 95% CIs, and p-values result from logistic regression models that were adjusted for age and sex. ORs correspond to each additional minor allele of the given variant.  $p < 0.008$  are considered as statistically significant after applying a Bonferroni correction for multiple testing OR=odds ratio; CI=confidence interval, NFE = Non-Finnish Europeans, PiD = Pick's disease

**Table 6.4 - Associations of individual *MAPT* variants with age of disease onset and disease duration in Pick's disease subjects**

Variant	Association with age of disease onset			Association with disease duration	
	MAF (N=309)	$\beta$ (95% CI)	p	$\beta$ (95% CI)	p value
rs1467967	28.8%	32.4%	33.5%	-0.11 (-0.80, 0.59)	0.76
rs242557	34.6%	36.8%	37.5%	-0.42 (-1.07, 0.24)	0.22
rs3785883	16.8%	18.0%	17.9%	0.08 (-0.79, 0.94)	0.86
rs2471738	19.9%	20.7%	22.0%	0.01 (-0.77, 0.80)	0.98
rs8070723	29.6%	23.0%	19.7%	0.25 (-0.46, 0.96)	0.50
rs7521	40.8%	46.6%	53.1%	-0.40 (-1.05, 0.26)	0.23

$\beta$  values, 95% CIs, and p-values result from linear regression models that were adjusted for sex and series (age of disease onset analysis) or sex, age of disease onset, and series (disease duration analysis).  $\beta$  values are interpreted as the change in the mean value of the given outcome (age of disease onset or disease duration) corresponding to each additional copy of the minor allele of the given variant.  $p < 0.008$  are considered as statistically significant after applying a Bonferroni correction for multiple testing.  $\beta$ =regression coefficient; CI=confidence interval.

### 6.3.2 Association results for the *MAPT* H1 subhaplotypes

#### H1 subhaplotypes and PiD risk

In secondary analysis, an evaluation of associations between six-variant *MAPT* haplotypes and risk of PiD is displayed in **Table 6.5**. As with the single-variant analysis, the H2 haplotype was associated with an increased risk of PiD (OR: 1.34, 95% CI: 1.11-1.63,  $p = 0.0028$ ); the slight difference between the two numerical estimates is due to the two different analysis approaches. Additionally, a nominally significant ( $p \leq 0.05$ ) protective association was noted for the rare H1f haplotype (0.0% in PiD, 1.2% in controls,  $p = 0.049$ ), with a slightly weaker finding noted for H1b (OR: 0.76, 95% CI: 0.58-1.00,  $p = 0.051$ ). There were no other notable associations between *MAPT* haplotypes and risk of PiD (all  $p \geq 0.15$ ).

#### H1 subhaplotypes and AAO / DD

Associations of *MAPT* haplotypes with AAO and DD in PiD patients are shown in **Table 6.6**. None of the six-variant *MAPT* haplotypes were significantly associated with AAO or DD after correcting for multiple testing ( $p < 0.0031$  Bonferroni corrected). However, nominally significant associations were observed with AAO for H1b ( $\beta$ : 2.66, 95% CI: 0.63 to 4.70,  $p = 0.011$ ), H1i ( $\beta$ : -3.66, 95% CI: -6.83 to -0.48,  $p = 0.025$ ) and H1u ( $\beta$ : -5.25, 95% CI: -10.42 to -0.07,  $p = 0.048$ ), and with a shorter DD for H1x ( $\beta$ : -3.73, 95% CI: -6.98 to -0.48,  $p = 0.025$ ).

Table 6.5 - Associations between *MAPT* haplotypes and risk of Pick's disease.

Haplotype	<i>MAPT</i> variant						Haplotype frequency		Association with PiD	
	rs1467967	rs242557	rs3785883	rs2471738	rs8070723	rs7521	PiD (N=338)	Controls (N=1312)	OR (95% CI)	p value
H1b	G	G	G	C	A	A	13.1%	16.0%	0.76 (0.58, 1.00)	0.051
H1c	A	A	G	T	A	G	10.2%	11.3%	0.93 (0.70, 1.25)	0.65
H1d	A	A	G	C	A	A	7.4%	7.1%	0.99 (0.68, 1.42)	0.94
H1e	A	G	G	C	A	A	9.8%	9.0%	1.03 (0.74, 1.42)	0.87
H1f	G	G	A	C	A	A	0.0%	1.2%	n/a <sup>1</sup>	0.049
H1g	G	A	A	C	A	A	0.7%	1.1%	0.43 (0.11, 1.65)	0.22
H1h	A	G	A	C	A	A	4.0%	4.1%	0.95 (0.57, 1.57)	0.85
H1i	G	A	G	C	A	A	3.9%	4.4%	0.98 (0.60, 1.61)	0.95
H1l	A	G	A	C	A	G	3.6%	3.0%	1.11 (0.67, 1.84)	0.69
H1m	G	A	G	C	A	G	2.9%	2.9%	1.00 (0.56, 1.78)	0.99
H1o	A	A	A	C	A	A	1.1%	2.3%	0.53 (0.23, 1.26)	0.15
H1p	G	G	G	T	A	G	1.1%	1.5%	0.82 (0.33, 2.04)	0.66
H1r	A	G	G	T	A	G	0.7%	1.1%	0.63 (0.20, 2.01)	0.44
H1u	A	A	G	C	A	G	2.4%	2.4%	1.11 (0.58, 2.11)	0.75
H1v	G	G	A	T	A	G	2.2%	1.2%	1.50 (0.70, 3.21)	0.30
H1x	G	A	A	T	A	G	1.3%	1.3%	1.06 (0.44, 2.56)	0.91
H1y	A	A	A	T	A	G	1.4%	1.6%	0.85 (0.34, 2.07)	0.71
H2	A	G	G	C	G	G	28.5%	22.7%	1.34 (1.11, 1.63)	<b>0.0028</b>

ORs, 95% CIs, and p-values result from score tests of association that were adjusted for age and sex. <sup>1</sup>Indicates a haplotype that was not observed in PiD patients, making estimation of an OR impossible.  $p < 0.0028$  are considered as statistically significant after applying a Bonferroni correction for multiple testing. OR=odds ratio; CI=confidence interval, PiD = Pick's disease

**Table 6.6 - Associations of *MAPT* haplotype with age of disease onset and disease duration in Pick's disease cases.**

Haplotype	Haplotype frequency (%), N=309	Association with age of disease onset		Association with disease duration	
		$\beta$ (95% CI)	<i>p</i> value	$\beta$ (95% CI)	<i>p</i> value
H1b	13.3%	2.66 (0.63, 4.70)	0.011	-0.03 (-1.07, 1.02)	0.96
H1c	10.0%	1.63 (-0.61, 3.86)	0.15	0.08 (-1.05, 1.22)	0.88
H1d	7.2%	0.79 (-1.79, 3.38)	0.55	-0.91 (-2.21, 0.39)	0.17
H1e	9.3%	0.52 (-1.94, 2.98)	0.68	0.52 (-0.72, 1.76)	0.41
H1h	4.0%	2.03 (-1.57, 5.64)	0.27	-0.45 (-2.27, 1.37)	0.63
H1i	4.1%	-3.66 (-6.83, -0.48)	0.025	-0.90 (-2.53, 0.72)	0.28
H1l	3.5%	-1.75 (-5.42, 1.92)	0.35	0.43 (-1.42, 2.28)	0.65
H1m	3.1%	-1.25 (-5.33, 2.84)	0.55	0.94 (-1.11, 3.00)	0.37
H1o	1.2%	0.05 (-6.91, 7.00)	0.99	0.03 (-3.47, 3.52)	0.99
H1p	1.0%	-5.65 (-12.60, 1.30)	0.11	0.17 (-3.36, 3.69)	0.93
H1u	2.2%	-5.25 (-10.42, -0.07)	0.048	-2.40 (-5.03, 0.22)	0.074
H1v	2.1%	-1.74 (-6.61, 3.13)	0.48	1.91 (-0.54, 4.35)	0.13
H1x	1.4%	-5.39 (-11.84, 1.07)	0.10	-3.73 (-6.98, -0.48)	0.025
H1y	1.5%	-0.70 (-6.93, 5.54)	0.83	1.82 (-1.31, 4.95)	0.26
H1z	1.6%	-1.81 (-8.02, 4.40)	0.57	-0.08 (-3.20, 3.05)	0.96
H2	29.4%	-0.62 (-2.03, 0.79)	0.39	0.22 (-0.49, 0.93)	0.54

$\beta$  values, 95% CIs, and *p*-values result from score tests of association that were adjusted for sex and series (age of disease onset analysis) or sex, age of disease onset, and series (disease duration analysis).  $\beta$  values are interpreted as the change in the mean value of the given outcome (age of disease onset or disease duration) corresponding to each additional copy of the given haplotype. *p* < 0.0031 are considered as statistically significant after applying a Bonferroni correction for multiple testing.  $\beta$ =regression coefficient; CI=confidence interval.

## 6.4 Discussion

PiD is a rare sporadic 3R tauopathy that presents primarily as a behavioural or language variant of frontotemporal dementia<sup>61–64,67</sup>. Little is known regarding the etiology or underlying pathobiology of the disease. To date, no genetic variation has been shown to associate with disease risk, although a handful of cases with PiD-like 3R predominant tau pathology have been suggested to be caused by rare *MAPT* mutations; G389R<sup>429</sup>, Q336R<sup>430</sup>, G272V<sup>431</sup> and Q336H<sup>432</sup>. In contrast to the majority of *MAPT* mutations, at least two of these mutations (Q336R and Q336H) lead to increased binding of tau to microtubules (MT) and their subsequent hyperstabilisation<sup>435</sup>. This MT hyperstabilisation is also seen in hereditary spastic paraplegia due to spastin mutations resulting in increased MT longevity due to decreased MT severing<sup>436</sup>. In the present study I have shown that the common *MAPT* H2 haplotype, strongly protective in 4R-tauopathy, is associated with an increased risk of PiD (3R tauopathy). This was only possible by establishing and creating a global consortium (PIC) to increase the number of available pathologically defined PiD cases. Previous genetic studies in PiD were underpowered with only 34 cases and 33 cases respectively<sup>185,186</sup> and failed to show an association between the *MAPT* H2 haplotype and disease risk. A ten-fold increase in sample size was needed to establish *MAPT* H2 as a risk factor for in PiD.

Previous research in frontotemporal dementia linked to chromosome 17 with tau pathology (FTDP17t) has clearly demonstrated that mutations in the 5' splice site of *MAPT* exon 10 can increase the incorporation of the exon into mRNA (10+ transcripts) thus increasing 4R isoform production. This emphasises how important exon 10 alternative splicing regulation is as its dysregulation is sufficient to drive tangle formation and consequent neurodegeneration<sup>156,172</sup>. Given the association of *MAPT* H2 with a 3R-tauopathy, and its protection in 4R-tauopathy, it is possible that the *MAPT* H1 and H2 haplotypes increase the expression of 4R and 3R tau respectively. Previous studies have investigated the haplotype influence on *MAPT*/tau expression but given the presence of six different isoforms in human brain it has remains to be conclusively proven<sup>437–439</sup>. There is a suggestion, however, using allele specific expression assays in cell models, that H1 chromosomes express significantly more exon 10+ transcripts (i.e 4R tau) than H2 chromosomes<sup>171,440</sup>. The genetic

predisposition herein described would appear to support the hypothesis that the pathologic effects of the H1-H2 haplotypes is via isoform specific expression differences. This may have implications in the determination of therapeutic strategies that have focused on either overall lowering of tau expression or specifically targeting the lowering of 4R-tau or increasing 3R-tau isoforms. The overall balance of the 3R and 4R forms of tau would appear to be important for the primary tauopathies but does not in itself explain the mixed pathology observed in AD, although it is tempting to suggest an overall expression of total tau may be underlying the mixed pathology.

In addition to *MAPT*, other genes are located within the inversion on Chr17q21.31 (*LRRC37A*, *LRRC37A2*, *NSF*, *ARL17A*, *ARL17B*, *KANSL1*, *SPPL2C*, and *CRHR1*). An additional explanation for the differing associations of the *MAPT* H1 and H2 haplotypes in the 4R and 3R tauopathies could be that the signals are being driven by alternative genes to *MAPT* in a haplotype specific manner. In support of this hypothesis, multiple genes other than *MAPT* are differentially regulated by the inversion haplotypes in a tissue specific manner<sup>437</sup>. There is also growing evidence that the H1 association signal at Chr17q21.31 in Parkinson's disease (PD), previously thought to be due to *MAPT*, may actually be driven by *KANSL1*<sup>183,184</sup> and/or *LRRC37A2*<sup>175</sup>. Interestingly, in association studies in Asian populations, in which the H2 haplotype is absent, there is no evidence of a signal at Chr17q21<sup>441</sup>. It is therefore possible that the H2 association signal in PiD may be also driven by H2 specific regulation of genes within the haplotype inversion other than *MAPT*. To test this hypothesis, future studies to investigate the effect of the H2 haplotype on expression of these genes may be of interest. Alongside *KANSL1* and *LRRC37A2* which have been implicated in PD pathogenesis, *N-ethylmaleimide sensitive fusion protein (NSF)* is also an interesting gene at this locus and may warrant further investigation. A recent study using proximity-labelling proteomics to chart tau interactomes in primary neurons found that tau interacts with NSF, modulates its localisation, and reduces its activity in a dose dependent manner<sup>442</sup>. They go on to show that pathogenic tau from brain lysates of human AD can suppress NSF activation. Given NSF's role in synaptic function and memory maintenance via AMPA receptor stabilisation at the post-synaptic membrane, they propose a mechanism by which pathogenic tau, via NSF inhibition, impairs synaptic function under disease associated conditions.

The results of this study provide evidence that the *MAPT* H2 haplotype is associated with an increased risk of PiD. In addition, I observed nominally significant associations that were observed with risk of PiD AAO, and DD, though these will require validation in larger studies. Although the strengths of this study (pathological diagnosis of PiD based on new consensus pathological derived criteria, and a large sample size for this disease) are important to highlight, there are several limitations which are important to consider. Due to the scarcity of PiD samples there was no possibility of defining an external validation dataset. I decided to maximize power in the primary study by including all the samples, which made it unfeasible to build a replication series. As a result, validation of these findings in an independent dataset will be important. Despite including all the PiD patients identified through the PIC (n=338) the sample remains relatively small (though a factor of 10 larger than any study so far carried out), which coupled with the relatively rare nature of many of the *MAPT* haplotypes, results in a low power to detect disease associations. Therefore, the possibility of a type II error (i.e. false-negative finding) is important to consider, and we cannot conclude that there is no true association between a given haplotype and risk of PiD simply due to a non-significant p-value in this study. Another limitation was that the controls were defined clinically rather than neuropathologically which means that some cases may have had underlying neuropathology that was not yet manifesting in a clinical syndrome. Additionally, without available genome-wide SNP data for controls, genetic principal components could not be regressed out, and so it is possible that population stratification could have had an influence on our results. However, we used the case genetic principal components to exclude any cases of non-European ancestry and our control *MAPT* H1-H2 frequencies were in keeping with published data<sup>443</sup>. The MAF for rs8070723 (H2 tagging variant) in the Mayo controls was 0.23, compared to 0.19 in the NFE population on GnomAD. This suggests, that if anything, that the H2 association with risk of PiD is more likely to be under- rather than over-estimated in this study. The fact that the association was still identified gives a degree of confidence that this is indeed a true effect.

## 6.5 Conclusions and future work

PiD is a rare and understudied disease with a devastating impact on both patients and their families. Through collaboration and building of the PIC, we have for the first

time a rare opportunity to engage in studies that may tease out the underlying pathobiology in PiD. In this chapter I have detailed the development of consensus neuropathological criteria for diagnosis of PiD pathology, there application to brain tissue samples collected as part of the PIC and used these samples to show for the first time that the *MAPT* H2 haplotype is associated with increased risk of developing PiD. As a primary tauopathy, there is the possibility that the identification of genetic variables, such as *MAPT* H2, involved in PiD pathology will inform into the other more common tau-related disorders, PSP, CBD, and potentially AD. Larger scale unbiased studies to explore genome-wide or structural genetic variation in PiD are now warranted. Furthermore, resolving the genetic determinants of PiD may help both in establishing diagnostic criteria and elucidating the dysfunctional pathways, which I hope will help to direct future therapeutic intervention strategies for this devastating disease.



# Chapter 7: Potential genetic modifiers of disease risk in autopsy-confirmed Pick's disease: a genome-wide association study

## 7.1 Introduction

PiD is characterised by severe “knife-blade” frontotemporal lobar atrophy and classified neuropathologically by the presence of ballooned neurons and argyrophilic inclusion Pick bodies. These eponymous Pick bodies contain hyperphosphorylated 3-repeat tau aggregates, leading to its designation as a 3-repeat tauopathy, in contrast to the 4-repeat tauopathies such as progressive supranuclear palsy (PSP), corticobasal degeneration (CBD), argyrophilic grain disease (AGD) and globular glial tauopathy (GGT). A recently proposed structure-based classification of the tauopathies, derived using cryo-electron microscopy, demonstrates that these 3R tau aggregates consist of a distinct disease-specific molecular conformation of tau fibrils in PiD<sup>41</sup>.

As discussed in more detail in Chapter 1 (**Tau biology in health and disease**) Tau is encoded by the MAPT gene on chromosome 17, with six major isoforms in the adult human brain<sup>33</sup> generated through alternative splicing of exons 2, 3, and 10. Rare mutations in the MAPT gene can cause Pick's-like 3R pathology<sup>54</sup>, though there has been no systematic study of large cohorts to ascertain their true prevalence in PiD. The genomic architecture of the MAPT locus is characterised by two haplotypes resulting from a 900kb inversion (H1) or non-inversion (H2) polymorphism<sup>173</sup>. Inheritance of the H1 haplotype has long been known to be a risk factor for both PSP (OR = 5.46)<sup>167,172</sup> and CBD<sup>169,193</sup> (OR = 3.45). In the previous chapter (Chapter 6), I have shown through direct genotyping of the MAPT H2 tagging SNP (rs8070723), that conversely the H2 haplotype is associated with an increased risk of Pick's disease (OR 1.35).

The rarity of PiD, combined with the difficulty of diagnosing the underlying pathology in life (due to lack of in vivo biomarkers and low clinico-pathological correlations), has impeded large scale genetic studies in this disease<sup>185,186</sup>. To overcome these issues, as part of the PhD I established the Pick's International Consortium (PIC) (see **Pick's disease International Consortium (PIC)**) and collected the largest number of PiD

cases to date to enable better powered genetic studies. There have been numerous GWASs in PSP; two case-control studies<sup>167,194</sup>, one investigating the genetic determinants of PSP phenotype<sup>195</sup>, and the most recent focused on association between genetic variation and survival<sup>198</sup>. There have been no equivalent studies yet performed in PiD.

The aim of this study was to perform a GWAS using autopsy-confirmed cases from the PIC, to identify genes that modify the risk for developing PiD, and so better understand the underlying pathophysiology with the hope that this will lead to new therapeutic approaches.

## 7.2 Methods

### 7.2.1 Study design and participants

Due to the rarity of PiD, and the consequent sample size, I was only able to perform a discovery case-control GWAS. Future work will need to validate the findings from this work in an independent PiD cohort.

At the time of this study, 321 neuropathologically confirmed PiD cases were available from the PIC (see **Pick's disease International Consortium (PIC)**), recruited from 31 international clinical or pathological research centres in the UK, France, Italy, Netherlands, Germany, Italy, Spain, Sweden, Australia, United States and Canada (Table 7.1). Of the 321 cases, 151 were from the UCL cohort and 170 from MCJ. For inclusion, all cases had to meet the strict PIC diagnostic criteria for PiD (as defined in Neuropathological diagnosis of Pick's disease); as a minimum there needed to be the presence of Pick bodies with 3R tau positive and 4R tau negative inclusions. The additional presence of ballooned neurons and positive Gallyas staining was preferred to confirm diagnosis. All samples were screened for the known *MAPT* mutations; the MCJ samples by direct sequencing, and the UCL samples by using the NeuroBooster Array that covers all known *MAPT* variants. Clinical and demographic data was collected for all cases, and included age at symptom onset, age at death and gender. This information was used to calculate the total disease duration, defined as age at death – age at symptom onset. Age at symptom onset was defined as the age at which first symptoms appeared, including initial cognitive dysfunction in judgment, language, or memory, or changes in behaviour or personality. Healthy controls with

no clinical signs of neurological disease were collected with the aim of having a ~1:3 ratio (cases:controls), a similar age (defined as age at blood draw), and similar gender balance to cases. Controls were obtained from two sources; the Global Parkinson's Genetics Program (GP2)<sup>311</sup>, and the Invasive Fungal Infection and GENetics (IFIGEN) cohort<sup>312</sup>.

The appropriate institutional review boards for each site approved the study, and written informed consent was obtained for each participant.

**Table 7.1 - Overview of total PiD samples included, broken down by contributing site.**

Contributing Site	Country	# samples
<b>UCL Institute of Neurology cohort (Europe / Australia)</b>		<b>151</b>
Netherlands Brainbank, Amsterdam	Netherlands	33
UCL Queen Square Brain Bank, London	UK	18
Neurological Tissue Bank, Biobanc-Hospital Clínic-IDIBAPS, Barcelona	Spain	14
Manchester Brain Bank, Manchester	UK	11
Cambridge Brain Bank, Cambridge	UK	11
London Neurodegenerative Diseases Brain Bank, London	UK	8
South West Dementia Brain Bank, University of Bristol	UK	8
Sydney Brain Bank, Sydney	Australia	8
Neurobiobank München, Munich	Germany	7
Oxford Brain Bank, Oxford	UK	7
Newcastle Brain Tissue Resource, Newcastle-upon-Tyne	UK	7
Neuro-CEB Biobank France, Paris	France	5
Victorian Brain Bank, Florey Institute, Parkville, VIC	Australia	5
Douglas-Bell Canada Brain Bank, Montreal, QC	Canada	3
The Brain Bank at Karolinska Institutet, Stockholm	Sweden	3
Fondazione IRCCS Istituto Neurologico Carlo, Milan	Italy	2
DZNE e.V. Standort Tübingen	Germany	1
<b>Mayo Clinic, Jacksonville cohort (North America)</b>		<b>170</b>
Mayo Clinic, Jacksonville, FL/Rochester, MN	USA	55
University of California, San Francisco, CA	USA	21
University of Pennsylvania, Philadelphia, PA	USA	18
Massachusetts General Hospital, Boston, MA	USA	15
Northwestern University, Chicago, IL	USA	14
Banner Sun Health Research, Sun City, AZ	USA	9
Sunnybrook Health Research, Toronto, ON	Canada	9
Columbia University, New York City, NY	USA	9
Emory University, Atlanta, GA	USA	7
University of British Columbia Hospital, Vancouver, BC	Canada	5
Houston Methodist Hospital, Houston, TX	USA	4
Krembil Research Institute, University of Toronto, Toronto, ON	Canada	2
UCLA - Sepulveda, Los Angeles, CA	USA	1
Parkwood Institute, University of Toronto, Toronto, ON	Canada	1
<b>Total PIC samples</b>		<b>321</b>

### 7.2.2 DNA preparation

For the majority of samples, DNA was extracted from each subject at their respective collection site (UCL or MCJ) as detailed in Chapter 2 (**DNA preparation**). I was responsible for the DNA extraction from all UCL samples, and Dr. Rebecca Valentino for all MCJ samples.

### 7.2.3 Genotyping

As detailed in Chapter 2 (**Whole-genome microarray genotyping**), all samples from MCJ (North American samples), Sydney (Australasia samples) and IFIGEN controls were genotyped by the local teams on the Illumina Global Screening Array version 3 (GSA). GP2 controls were genotyped on the Illumina NeuroBooster Array (NBA) by the local teams. All UCL samples (European samples) were genotyped at UCL Genomics on the Illumina NeuroBooster Array; I carried out the DNA sample quality control and preparation for genotyping including the NanoDrop and Qubit fluorometry and plated all samples which were then delivered to UCL Genomics for genotyping.

Raw IDAT files were shared with me at UCL, and I called genotypes using GenomeStudio version 2.0 (Illumina), as detailed in Chapter 2 (**Genome-wide genotype calling**). All UCL cases and GP2 controls were screened for known *MAPT* mutations covered by the NBA and were excluded from downstream analysis if a *MAPT* mutation was identified. MCJ samples had already been screened for *MAPT* mutations before being included in the study.

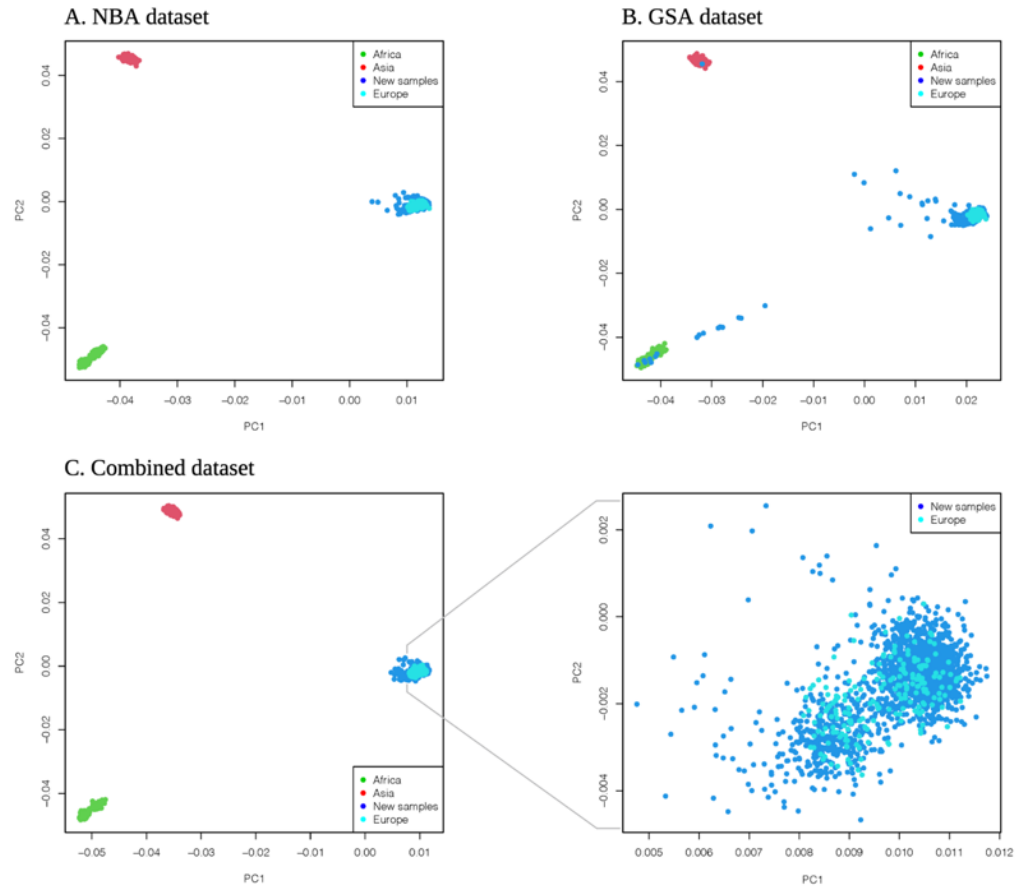
### 7.2.4 Quality control and imputation

I performed standard quality control procedures in PLINK (v1.9)<sup>444</sup> and R (4.0.5, 2021-03-31) at the individual sample level and then the SNP level. I performed all the quality control (QC) steps detailed below separately for the NBA and GSA samples. Post-QC each dataset was imputed separately, then merged post-imputation based on overlapping SNPs for downstream analysis.

Samples with a low overall genotyping rate (<98%), related individuals (Identity-By-Descent PIHAT>0.1), and heterozygosity outliers (>2SDs away from the mean) were removed, as were individuals where clinically reported biological sex did not match genetically determined sex. I also excluded variants with a missingness rate > 5%,

minor allele frequency (MAF)  $< 0.01$  and Hardy-Weinberg equilibrium (HWE)  $p < 1 \times 10^{-6}$ . I performed principal component analysis (PCA) (see Principal component analysis) on a linkage disequilibrium (LD) pruned set of variants (removing SNPs with an  $r^2 > 0.5$  in a 50kb sliding window shifting five SNPs at a time), after merging with the HapMap3 reference panel (**Figure 7.1A** and **Figure 7.1B**) for NBA and GSA datasets respectively). Individuals who deviated more than six standard deviations (6SD) from the mean of the first 10 principal components of the HapMap3 CEU population were excluded from the analysis.

I imputed the two genotyping array datasets (NBA and GSA) separately against the Haplotype Reference Consortium (HRC) reference panel (version r1.1 2016; <http://www.haplotype-reference-consortium.org/>) in the Michigan Imputation Server (RRID:SCR\_017579; <https://imputationserver.sph.umich.edu>) using Minimac4 (version 1.0.0; <https://genome.sph.umich.edu/wiki/Minimac4>). Imputed variants were excluded if the imputation information  $R^2$  was  $\leq 0.7$  and the genotype posterior probability was  $\leq 0.9$ , to ensure that only high-quality genotype calls were retained for further analysis. I then merged the two datasets based on shared variants, and variants with missingness  $> 5\%$  and minor allele frequencies  $< 1\%$  were also excluded. Ancestry was then rechecked on the merged dataset using the same procedure as detailed above, excluding any further samples that were greater than 6SD from the mean of the first 10 principal components of the CEU population (**Figure 7.1C**). After this final extraction of European-ancestry samples I re-calculated the first ten principal components and used as these as covariates in the association analysis



**Figure 7.1 - Genetic principal component plots.** First two principal components of the NBA (A) and GSA (B) and combined dataset (C) plotted against the HapMap3 Genome Reference Panel. These data include cases and controls. The NBA and GSA datasets are plotted before samples more than 6SD from the mean of the first 10 principal components were removed. For the combined dataset (C) all samples are within 6SD of the CEU reference population (outliers removed), which is expanded in the figure to the right.

## 7.2.5 Association analysis

I used logistic regression (see **Logistic (and linear) regression**) to test the additive genetic model of each variant for association with disease status using PLINK v1.9. I applied stepwise logistic regression (*step* function in R *stats* package [version 3.6.2]) to identify the optimal number of covariates to include in the GWAS (from gender, genotype array, age and PC1-10) that minimised the AIC and maximised the  $R^2$ . Using this method, the final covariates used in the logistic GWAS were gender, genotype array, age (at death for cases and age at blood draw for controls), and three principal components (PC1, 8 and 10), which gave an AIC of 1027 and an  $R^2$  of 0.40. I defined a genome-wide significant threshold at  $p < 5 \times 10^{-8}$ , with a threshold of  $p < 5 \times 10^{-6}$  for a suggestive (nominal) association. Variant positions are reported on human genome version 37 (GRCh37/hg19).

## 7.2.6 Genomic risk loci definition and gene mapping

I used FUMA (described in more detail in **Functional Mapping and Annotation of GWAS (FUMA)**) to annotate and functionally map the variants included in the GWAS<sup>343</sup>. I defined genomic risk loci around variants with  $p < 5 \times 10^{-6}$ , and included all variants correlated ( $R^2 > 0.6$ ) with the most significant variant. Genomic risk loci within 250 kilobases (kb) of each other were incorporated into the same locus. The individual genomic risk loci were mapped genes using positional and eQTL mapping; for positional mapping all variants within 10kb of a gene in the genomic risk locus were assigned to that gene. For eQTL mapping, FUMA maps variants on the basis of significant eQTL interactions in the PsychEncode, BloodeQTL, CMC, and GTEXv8 (Brain and whole blood) data repositories. To identify whether any of the lead SNPs were eQTLs for gene expression I used GTEXv8. Regional association plots were generated in LocusZoom (RRID:SCR\_021374; <https://my.locuszoom.org/>), and LDProxy was used to identify any deleterious variants in high linkage disequilibrium (LD) with variants of interest (<https://ldlink.nci.nih.gov/?tab=ldproxy>)<sup>445</sup> in European populations (excluding the Finnish population).



### 7.2.7 Conditional analysis

To understand whether there were one or more variants at the same locus contributing to the signal at each genomic risk locus, I performed conditional analysis on single SNPs using a conditional and joint association approach<sup>446</sup>. I used the GWAS summary statistics and the AMP-PD cohort ( $n = 10,418$ ) as the reference sample for linkage disequilibrium estimation. The reference sample went through the same QC steps as described above for the PiD cohort. I used CGTA-COJO software (v.1.93.0 beta for Linux; <https://yanglab.westlake.edu.cn/software/gcta/#Overview>)<sup>357</sup> to perform association analyses conditioned on SNPs of interest.

### 7.2.8 Fine-mapping and functional annotation

To nominate causal variants, fine-mapping was applied using SuSiE (v. 0.12.27; <https://github.com/stephenslab/susieR>)<sup>447</sup> and FINEMAP (v.1.3.1; <http://www.christianbenner.com/>)<sup>448</sup> on all variants within 1Mb of the lead variant of each genomic risk loci. A credible set, in the context of a multiple regression model, is defined as a subset of variants that has a 95% probability of  $p$ , or greater, of containing at least one effect variant (i.e. a variant with a non-zero regression coefficient). SuSiE calculates the credible set using a simple model-fitting algorithm, Iterative Bayesian Stepwise Selection (IBSS), that fits a “single effects” regression model at each step<sup>449</sup>. FINEMAP on the other hand uses a Shotgun Stochastic Search (SSS) algorithm that explores the causal configuration space by concentrating effort on those configurations with non-negligible probability<sup>448</sup>. To compute the SNPs correlation matrix necessary to infer the SNPs credible set, I used the 503 European ancestry individuals from the Phase 3 of the 1000 Genomes Project.

The *echolocator* R package (V. 1.4; <https://github.com/RajLabMSSM/echolocator>) was used to report the Union Credible Set SNPs (UCS), which is the union of all tool-specific CS<sub>95%</sub>, as well as the Consensus SNPs, which are those nominated from the two fine mapping tools.

To further investigate cis and trans regulatory mechanisms in these nominated genomic regions, each locus was mapped to brain cell type specific enhancer-promoter interactome data, to regulatory elements data from the FANTOM5 (RRID:SCR\_002678) project<sup>450,451</sup>, and to functional DNA elements from the

ENCODE dataset (RRID:SCR\_006793, <https://www.encodeproject.org/>)<sup>348</sup> using the *echolocatoR* R package as detailed above.

## 7.2.9 Colocalisation analysis

To investigate whether there was an overlap between the GWAS loci that reached nominal significance ( $p < 5 \times 10^{-6}$ ) and expression quantitative trait loci (eQTLs), a colocalization analysis was performed using the *coloc* R package for all genes within  $\pm 1$  Mb of the lead genomic loci SNP (version 5.1.0; <https://cran.r-project.org/web/packages/colocr/index.html>)<sup>356</sup>. More detail on this method is given in Chapter 2 (**Colocalisation analysis**). Given *coloc* calculates Bayes factors under the assumption that there is a single casual variant at a locus, I first performed conditional analysis (see **Conditional analysis**), to confirm that there were no additional independent signals and so ensure that this assumption of a single casual variant was met.

I used cortex-specific cis-eQTLs from MetaBrain (<https://www.metabrain.nl/>)<sup>452</sup>; this is the largest meta-analysis of brain eQTLs studies available to date, providing more confident effect sizes and significance estimates, hence increasing the power and the certainty when doing colocalisation analyses. *Coloc* was run using default priors; these are the prior probabilities that any random SNP in the region is associated with trait 1 or trait 2,  $p_1=10^{-4}$  and  $p_2=10^{-4}$ . A threshold of  $p_{12}=5 \times 10^{-6}$  was used for the  $p_{12}$  prior, which is the probability that a SNP in the region is associated with both traits. Loci with a posterior probability of hypothesis 4 (PP.H4)  $\geq 0.85$  were considered as significant evidence of colocalization between the GWAS and the eQTL traits (one shared causal variant). I used a more conservative threshold for PP.H4 than used in the original work by Giambartolomei et al<sup>356</sup> (0.75), in line with more recent work where a threshold of 0.85<sup>453</sup> or even 0.90<sup>446</sup> has been used on GWAS data.

In addition, to investigate whether the nominated loci regulate alternative splicing, a similar approach was followed using cortex splicing QTLs (sQTLs) from the GTExv8<sup>351</sup> containing all variant-gene association from 255 individuals, based on LeafCutter (version 0.2.9 RRID:SCR\_017639; <https://davidaknowles.github.io/leafcutter/>)<sup>454</sup>.

### 7.2.10 Assessment of gene transcript and protein expression of lead genes

Brain expression profiles of gene transcripts and encoded proteins highlighted by the GWAS were assessed, using a range of different publicly available online data sources. Bulk RNA and protein expression was assessed using The Human Protein Atlas (RRID:SCR\_006710; <https://www.proteinatlas.org/>)<sup>455</sup>. This resource provides consensus RNA expression data derived from the GTExv8 RNA-seq, Human Protein Atlas (HPA) RNA-seq and the FANTOM CAGE datasets. Protein expression by tissue is based on tissue profiles generated from 6120 antibodies with more than five million immunohistochemistry-based images covering 5067 human genes, corresponding to approximately 25% of the human genome<sup>456</sup>.

Cell specific RNA expression was investigated using the Brain RNA-Seq database (RRID:SCR\_017483; <https://www.brainrnaseq.org/>)<sup>457</sup>, and single cell RNA-seq data provided by DropViz (<http://dropviz.org/>)<sup>458</sup>. Data in the Brain RNA-seq dataset was generated from healthy temporal lobe samples resected from 14 patients intra-operatively. The cell types sequenced from these samples were mature astrocytes (n=12), microglia (n=3), oligodendrocytes (n=3) and neuron (n=1). DropViz provides gene expression on 690,000 individual cells derived from nine different regions of the adult mouse brain. Predicted protein interaction networks were investigated using the STRING database (RRID:SCR\_005223; <https://string-db.org/>).

### 7.2.11 Candidate variant analysis

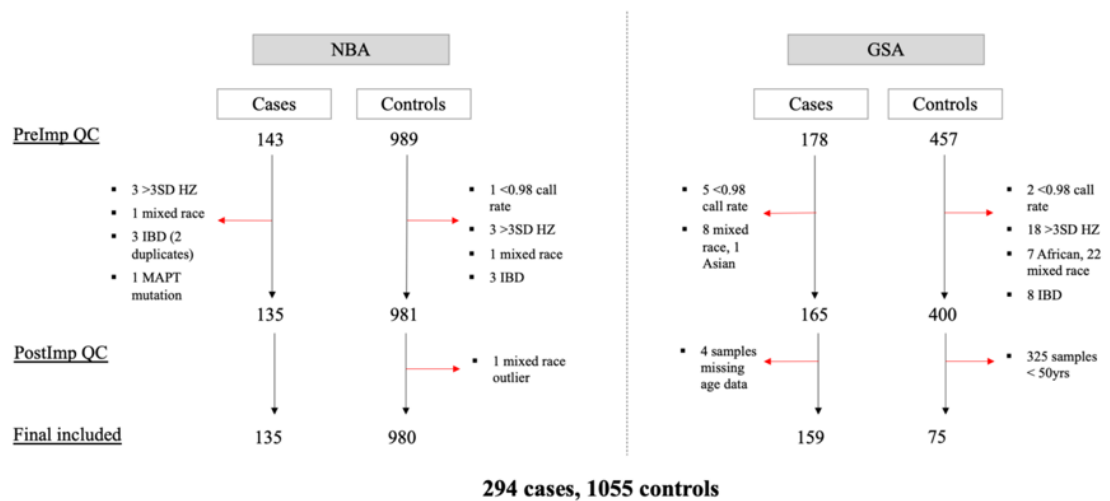
Specific variants that have previously been identified in related diseases were extracted to check whether they showed any association with risk of PiD. This included variants identified in the primary tauopathies (PSP<sup>167,195,298</sup>, CBD<sup>169</sup> and primary aged-related tauopathy (PART)<sup>459</sup> and clinically diagnosed FTD<sup>204</sup>. I also checked for an association between *MAPT* haplotypes and risk of PiD, by extracting the six *MAPT* variants that define the H1-subhaplotype structure (see ***MAPT* SNP Genotyping**). In particular, I wanted to confirm the association of the H2 haplotype with risk of PiD that I demonstrated in Chapter 6 when directly genotyping the six *MAPT* sub-haplotype defining SNPs. Finally, I checked whether ApoE genotype<sup>460</sup> and TMEM106B<sup>166</sup> were associated with risk of PiD disease. APOE genotypes were

inferred from the imputed genotypes of rs7412 and rs429358 variants. I also tested for an association between either ApoE genotype or *MAPT* H2 haplotype, age at onset and survival (age at death - age at onset).

## 7.3 Results

### 7.3.1 Cohort characterisation

A total of 321 autopsy confirmed PiD cases were considered for inclusion from 31 different brainbanks in the PIC (**Table 7.1**). Of these, 143 cases were genotyped on the NBA (all from the UCL series), and 178 cases were genotyped on the GSA (171 from the Mayo series and 7 from Sydney collected as part of the UCL series). 1446 controls were considered for inclusion (989 from NIH that were genotyped on the NBA, and 457 from IFIGEN genotyped on the GSA). Samples excluded at each stage of the QC process are summarised in **Figure 7.2**. After thorough quality control and filtering, 294 cases (135 NBA and 159 GSA), and 1055 controls (980 NBA and 75 GSA), covering 6,316,457 variants were available for association analysis. Due to the young average age of the IFIGEN controls compared to the GSA genotyped cases (Mean, SD; 37.7 years  $\pm$  12.1 vs 70.1 years  $\pm$  7.6;  $t = -31$ ,  $p < 0.05$ ), I only selected GSA controls older than 50yrs for selected for inclusion. This resulted in 75 GSA controls being selected (55.7 years  $\pm$  5.6) that were more closely matched in age, while still leaving enough GSA genotyped controls to allow inclusion of array type as a covariate to regress out array batch effects in the association analysis. Demographics and basic clinical characteristics of the samples included for analysis are summarised in (**Table 7.2**). The median age of onset for cases was 58 years (min-max; 33-80 years), median age at death 70 years (40-90 years) with a median survival of 10 years (3-25 years). Interestingly there were a higher percentage of men (186/294, 63.3%) compared to woman (108/294, 36.7%) in the overall cohort, with this difference more pronounced in the UCL series (67.4% males vs. 59.7% in the MCJ series).



**Figure 7.2 - Overview of sample quality control for PiD GWAS.** Summary of samples excluded at each stage of quality control (QC) pre-imputation (PreImp) and post-imputation (PostImp). Final number of samples included in GWAS detailed at bottom of figure.

Table 7.2 - Clinical characteristics of samples included in GWAS.

Variable	MCJ PiD series <sup>1</sup> (GSA)	UCL PiD series (NBA)	Combined PiD series	Controls <sup>2</sup>
Sample (n)	159	135	294	1055
Age (years)	70 (40, 90)	69 (41, 88)	70 (40, 90)	65 (50, 98)
Age of disease onset (years)	59 (36, 79)	58 (33, 80)	58 (33, 80)	N/A
Disease duration (years)	10 (3, 25)	10 (3, 20)	10 (3, 25)	N/A
Sex				
- Male	95 (59.7%)	91 (67.4%)	186 (63.3%)	558 (52.9%)
- Female	64 (40.3%)	44 (32.6%)	108 (36.7%)	497 (47.1%)

The sample median (minimum, maximum) is given for age. Age represents age at death for Pick’s disease cases and age at blood draw in controls. <sup>1</sup> Includes 7 Australasian samples from the UCL series which were also genotyped on the GSA-v3. <sup>2</sup> NBA ctls: n = 980, median age (min, max) = 66 (55, 98), GSA ctls: n = 75, median age (min, max) = 54 (50, 72). Abbreviations: GSA = Illumina Global Screening Array, PiD = Pick’s disease, MCJ = Mayo Clinic Jacksonville, NBA = Illumina NeuroBooster Array, UCL = University College London

### 7.3.2 *MAPT* mutations

All samples included from MCJ and Sydney had negative *MAPT* mutation screening. All UCL samples were screened for known *MAPT* mutations covered by the NBA at the GenomeStudio genotype calling stage. This identified four samples with a total of five rare *MAPT* variants; K280del (GRCh37/hg19 Chr17:44087694 AAG>---), A239T (Chr17:44073923 G>A; p.Ala239Thr), S318L (Chr17:44061123 C>T, p.Ser319Leu) and V363I (Chr17:44096073 G>A; p.Val363Ile) in the same case, and Q230R (Chr17:44060859 A>G; p.Gln305Arg). Of these four samples, two of them failed QC (A239T IBD, S318L and V363I > 3SD HZ) so were excluded from the association analysis anyway. The K280del was excluded from downstream analysis, while the Q230R was included given this is likely a benign polymorphism in *MAPT* with a MAF of 0.05 in the European (non-Finish) population (GnomAD; <https://gnomad.broadinstitute.org/>). The MAF for this variant tagging Q230R in this dataset was 0.06 for both cases and controls, supporting the fact that this is unlikely to be a deleterious variant in PiD.

### 7.3.3 Targeted assessment of candidate variants

Variants were extracted that have either previously been associated with tauopathies or related diseases, to test for an association with PiD (**Table 7.3**) This confirmed the association for the *MAPT* H2 haplotype (OR: 1.52, 95% CI: 1.18 - 1.97,  $p = 0.001$ ) that I previously demonstrated in Chapter 6 through direct genotyping rs8070723 (see **Association results for the *MAPT* rs8070723 H2 allele**). Six tagging SNPs were used to infer H2 and H1 subhaplotypes in *MAPT*; rs8070723 (H2 tagging variant), rs1467967, rs242557 (H1C haplotype-tagging variant), rs3785883, rs2471738, and rs7521 There was 100% concordance between the direct genotyping and chip-based imputation of these six SNPs. Looking at the frequency of *MAPT* H1 and H2 haplotypes one can see that there is a small, but significant increase (Chi square:  $\chi = 6.04$ ,  $df = 2$ ,  $p = 0.003$ ) in both H1/H2 heterozygotes (45.6% PiD cases vs. 36.1% controls) and H2/H2 homozygotes (6.8% PiD cases vs. 5% controls) (**Table 7.4**). Of the other 17 variants tested, none passed the analysis-wide significance threshold ( $p = 0.05/18$ ; 0.003), though *MOBP* showed a nominal association (OR: 0.76, 95% CI: 0.59 – 0.98,  $p = 0.03$ ) which did not survive Bonferroni correction. Neither the ApoE4 genotype or the *MAPT* H2 haplotype were associated with age of onset or survival.

Table 7.3 - Candidate variant analysis using GWAS data

Disease	Chr	SNP	Nearest gene	OR (95% CI)	MAF (cases)	MAF (controls)	MAF (total cohort)	p value
<b>MAPT (H2)</b>	<b>17</b>	<b>rs8070723</b>	<b>MAPT</b>	<b>1.52 (1.18 - 1.97)</b>	<b>0.29</b>	<b>0.23</b>	<b>0.24</b>	<b>0.001<sup>a</sup></b>
MAPT (H1c)	17	rs242557	MAPT	1.02 (0.80 - 1.20)	0.35	0.34	0.34	0.86
AD	19	rs429358	ApoE	1.40 (1.02 - 1.93)	0.15	0.13	0.13	0.04
AD	19	rs7412	ApoE	0.97 (0.66 - 1.43)	0.09	0.09	0.09	0.89
PSP	1	rs564309	TRIM11	1.31 (0.88 - 1.96)	0.09	0.09	0.09	0.19
PSP	12	rs2242367	SLC2A13	1.11 (0.86 - 1.42)	0.29	0.26	0.27	0.43
PSP	1	rs1411478	STX6	1.08 (0.86 - 1.35)	0.42	0.41	0.41	0.51
<b>PSP</b>	<b>3</b>	<b>rs1768208</b>	<b>MOBP</b>	<b>0.76 (0.59 - 0.98)</b>	<b>0.25</b>	<b>0.28</b>	<b>0.28</b>	<b>0.03<sup>b</sup></b>
PSP	2	rs7571971	EIF2AK3	0.96 (0.74 - 1.25)	0.28	0.26	0.26	0.77
FTD-TDP	7	rs1990622	TMEM106B	1.24 (0.98 - 1.56)	0.42	0.39	0.40	0.07
FTD	11	rs302668	RAB38	1.15 (0.90 - 1.45)	0.37	0.35	0.35	0.26
FTD	11	rs16913634	RAB38/CTSC	0.95 (0.63 - 1.45)	0.09	0.09	0.09	0.81
FTD	6	rs9268877	HLA-DRA/DRB5	1.04 (0.83 - 1.31)	0.42	0.44	0.44	0.72
FTD	6	rs9268856	HLA-DRA/DRB5	0.91 (0.70 - 1.19)	0.25	0.24	0.24	0.50
FTD	6	rs1980493	BTNL2	1.04 (0.74 - 1.46)	0.15	0.14	0.14	0.82
PART	4	rs56405341	JADE1	1.01 (0.79 - 1.31)	0.27	0.29	0.28	0.91
CBD	2	rs963731	SOS1	0.63 (0.37 - 1.08)	0.05	0.05	0.05	0.09
CBD	8	rs643472	lnc-KIF13B-1	1.13 (0.74 - 1.25)	0.23	0.22	0.27	0.35

Logistic regression additive model adjusted for gender, age, 3 PCs (PC1, 8 and 10) were used to study the association of candidate loci with risk of PiD in the total cohort (294 cases, 1055 controls). <sup>a</sup> significant after correction for multiple comparisons (Bonferroni:  $p=0.05/18=0.003$ ). <sup>b</sup> significant without correction for multiple comparisons. Abbreviations: CI = confidence interval, MAF = minor allele frequency, OR = odds ratio. Bold text highlights statistically significant variants.



**Table 7.4 - MAPT Haplotypes from GWAS**

	H1/H1	H1/H2	H2/H2
Cases	47.6% (140)	45.6% (134)	6.8% (20)
Controls	58.9% (621)	36.1% (381)	5.0% (53)
Total	56.4% (761)	38.2% (515)	5.4% (73)

Samples size (n) = 294 controls, 1055 cases within 6SD of CEU. MAPT haplotypes derived from rs8070723 H2 tagging SNP (minor allele G = H2). Chi square:  $\chi = 6.04$ , df = 2,  $p = 0.003$

### 7.3.4 Association Results

Using a case-control logistic regression model, adjusting for gender, genotype array, age and three genetic principal components (PCs 1, 8 and 10) to account for population substructure, I assessed the role of 6,316,457 variants in the risk of developing PiD. The genomic inflation factor ( $\lambda$ ) was 0.99 ( $\lambda_{1000} = 0.97$ ) demonstrating there was no confounding by population stratification. No disease-associated variants reached genome-wide significance ( $p < 5 \times 10^{-8}$ ), but there were suggestive associations ( $p < 5 \times 10^{-6}$ ) at five genomic loci (**Figure 7.3**), with the leading SNP at each these loci shown in **Table 7.5**.

**Figure 7.4** shows more detailed regional association plots for each of the five genomic loci with suggestive associations. Conditional analyses were performed on the lead variant at each of these five loci to confirm that there were no additional independent signals.

#### rs11216197 (*KCTD8*)

The most significant SNP, rs11216197 on Chromosome 4, is an intronic variant located in *KCTD8* (Odds Ratio (OR) = 7.53, 95% Confidence Interval (CI) = 3.62-15.65,  $p = 6.37 \times 10^{-8}$ ) (**Figure 7.4B**). The *KCTD8* gene encodes a potassium tetramerisation domain that facilitates GABA<sub>B</sub> receptor expression in axonal terminals and contributes to presynaptic excitation by GABA<sub>B</sub> receptors<sup>461,462</sup>. c.500kb downstream of the lead variant there are three genes: *YIPF7*, *GUF1* and *GNPDA2*. rs112161979 is an eQTL for *GNPDA2* in blood (GTExv8;  $p = 6.4 \times 10^{-5}$ ) and nerve-tibial tissue (GTExv8;  $p = 6.3 \times 10^{-5}$ ), and for *GUF1* in blood (GTExv8;  $p = 1.9 \times 10^{-4}$ ), though not for either gene in the brain. *GNPDA2* (Glucosamine-6-Phosphate Deaminase 2) encodes an enzyme that controls flux through the hexosamine biosynthetic pathway; increased flux through this pathway modulates the unfolded protein response<sup>463</sup>, and also mediates endoplasmic reticulum (ER) stress resistance and longevity<sup>464,465</sup>. The *GUF1* (GTP Binding Elongation Factor) gene encodes a

GTPase that triggers back-translocation of the elongating ribosome during mitochondrial protein synthesis. Homozygous variants in *GUF1* have been associated with West syndrome, an infantile disorder characterised by infantile spasms, hypsarrhythmia and developmental regression<sup>466</sup>. *YIPF7* encodes a protein that forms the alpha subunit of the YIPF complex 1 that localises to the early compartment of the Golgi apparatus (ERGIC) and is recycled between the endoplasmic reticulum and the cis-Golgi<sup>467</sup>.

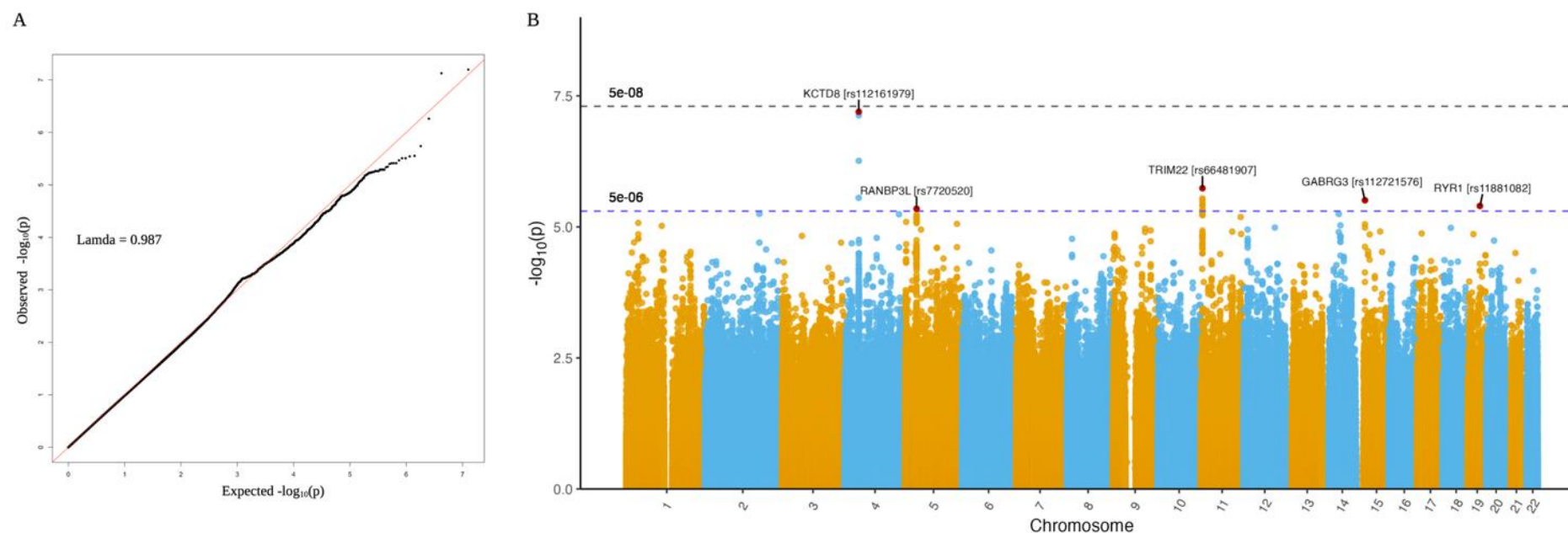
#### **rs66481907 (*TRIM22*)**

The next suggestive association was on Chromosome 11; the lead variant in this region was rs66481907 (OR = 2.10, 95% CI = 1.54-2.84,  $p = 1.83 \times 10^{-6}$ ), an intronic SNP located in the *TRIM22* gene (**Figure 7.4B**). *TRIM22* is a member of the tripartite motif-containing (TRIM) superfamily, all of which have an E3 ubiquitin ligase function, and are involved in a wide range of cellular processes including degradation of misfolded proteins<sup>468</sup>, and regulation of the NLRP3 inflammasome signalling pathway<sup>469</sup>. *TRIM22* and *TRIM5* also have been shown to have antiretroviral activity against a wide range of viruses including HIV, Influenza A, Hepatitis B and C<sup>470</sup>, and play an important role in the innate immune response to infection<sup>471</sup>. Interestingly the rs66481907 SNP is a sQTL for *TRIM22* in nerve-tibial tissue (GTExv8; intron id 5708603:5709053:clu\_7256,  $p = 1.2 \times 10^{-8}$ , is associated with both non-coding transcripts with a retained intron, as well transcripts targeted for nonsense mediated decay in ENSEMBL (RRID:SCR\_002344; ENSEMBL) and has a Combined Annotation Dependent Depletion (CADD) score of 10.02 placing it in the top 10% most deleterious variants in the genome.

#### **rs112721576 (*GABRG3*), rs11881082 (*RYR1*) and rs7720520 (*RANBP3L*)**

The final three suggestive genomic loci with variants showing nominal significance were Chromosome 15 (lead SNP rs112721576, an intronic variant in *GABRG3* (OR = 3.73, 95% CI = 2.14-6.48,  $3.10 \times 10^{-6}$ )), Chromosome 19 (lead SNP rs11881082, a splice site variant in *RYR1* (OR = 2.96, 95% CI = 1.87-4.69,  $4.00 \times 10^{-6}$ ), and Chromosome 5 (lead SNP rs7720520, an intergenic variant close to *RANBP3L* (OR = 1.76, 95% CI = 1.38-2.23,  $4.50 \times 10^{-6}$ )) (**Figure 7.4C-E**). The rs11881082 SNP on Chromosome 19 was an eQTL for *RYR1* in the amygdala (GTExv8;  $p = 3.36 \times 10^{-5}$ ), anterior cingulate cortex (GTExv8;  $p = 4.22 \times 10^{-11}$ ), nucleus accumbens (GTExv8;

$p = 1.05 \times 10^{-8}$ ), putamen (GTExv8;  $p = 3.00 \times 10^{-5}$ ) and the caudate (GTExv8;  $2.48 \times 10^{-11}$ ). No eQTLs were identified for the other two variants. *GABRG3* encodes the GABA Type A receptor subunit gamma, mutations in which have been shown to be associated with autism<sup>472,473</sup>. *RYR1* encodes the Ryanodine Receptor Type 1, an intracellular calcium channel that mediates the release of calcium from the intracellular stores, and although primarily expressed in the sarcoplasmic reticulum of muscle<sup>474</sup>, it is also expressed at lower levels in the brain; in particular the hippocampus and cerebellum<sup>475</sup>.

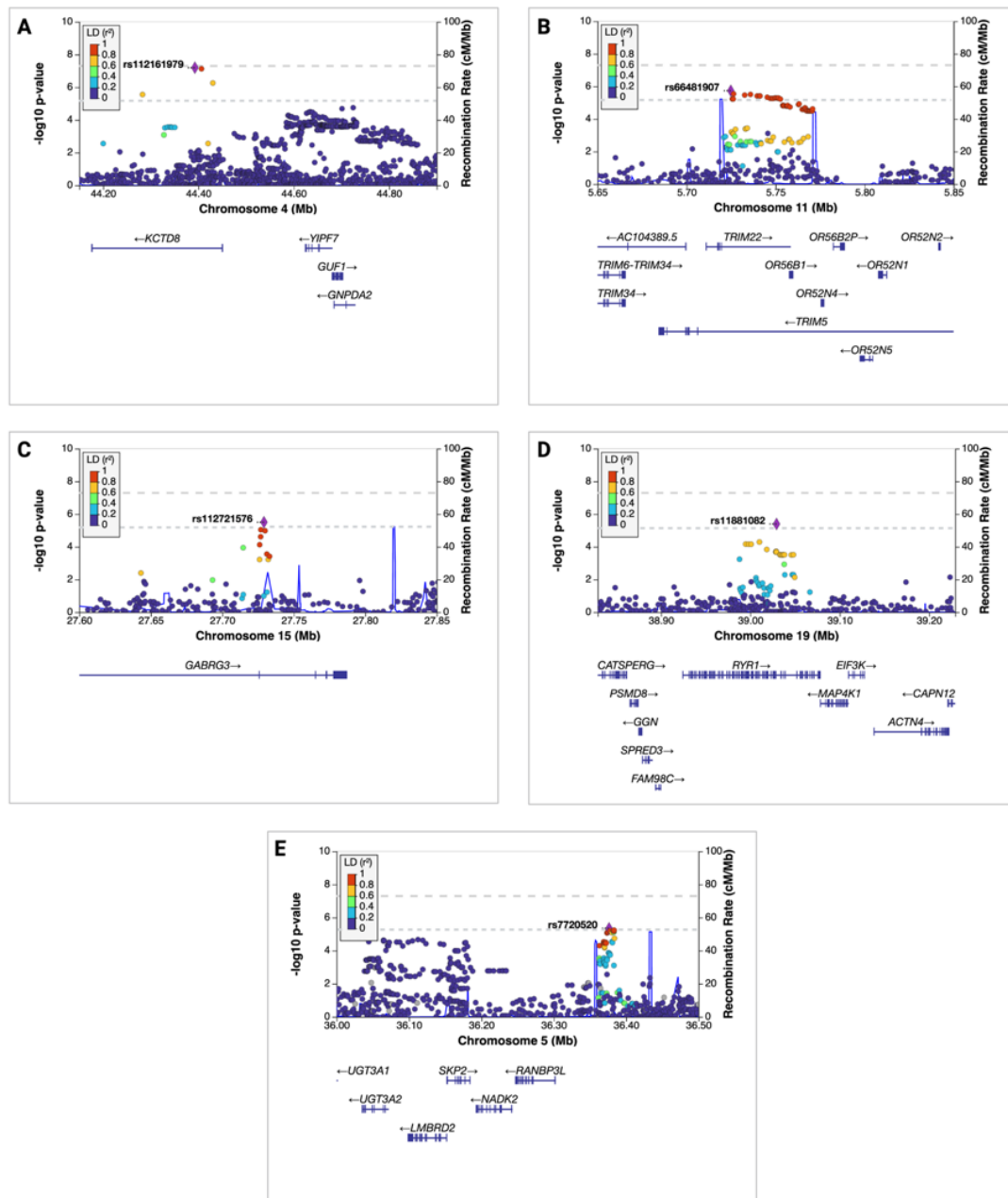


**Figure 7.3 - Association plots for PiD.** (A) Quantile-quantile (QQ) plot based on 6,316,457 variants after imputation. The genomic inflation factor (lambda  $\lambda_{gc}$ ) was 0.987. (B) Manhattan plot showing  $-\log_{10}(P)$  values from logistic regression of imputed variants corrected for age, gender, array and three principal components (PC1, 8 and 10). Red dots indicate the variant (rsID and nearest gene labelled) with the lowest  $p$  value at each genomic locus that reached nominal significance ( $p < 5 \times 10^{-6}$  indicated by the blue dashed line). Genome-wide significance was set at  $p < 5 \times 10^{-8}$  and indicated by the grey dashed line.

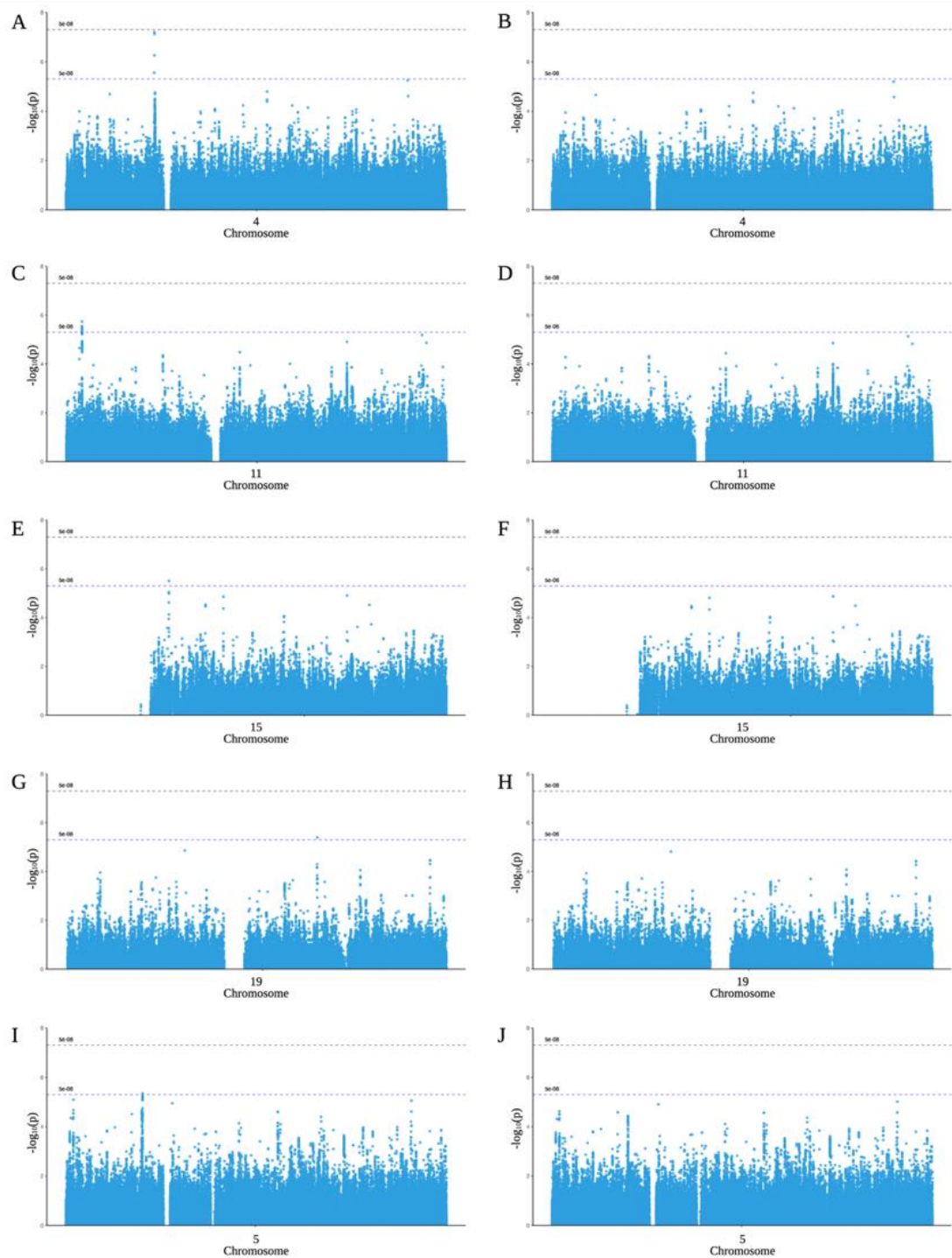
Table 7.5 - Top independent SNPs at suggestive loci from PiD GWAS

					MAF				
Chr	BP	SNP	Nearest gene	Minor allele	Cases	Controls	Population*	OR (95% CI)	p value
4	44,392,571	rs112161979	KCTD8	A	0.040	0.013	0.016	7.53 (3.62-15.65)	6.37 x 10 <sup>-8</sup>
11	5,724,803	rs66481907	TRIM22	A	0.206	0.126	0.120	2.10 (1.54-2.84)	1.83 x 10 <sup>-6</sup>
15	27,729,149	rs112721576	GABRG3	G	0.045	0.029	0.038	3.73 (2.14-6.48)	3.10 x 10 <sup>-6</sup>
19	39,029,201	rs11881082	RYR1	A	0.070	0.050	0.053	2.96 (1.87-4.69)	4.00 x 10 <sup>-6</sup>
5	36,376,351	rs7720520	RANBP3L	G	0.421	0.328	0.337	1.76 (1.38-2.23)	4.50 x 10 <sup>-6</sup>

Logistic regression additive model adjusted for gender, age, array and 3 PCs (PC1, 8 and 10) was used to study the association between 6,316,457 variants and risk of PiD (294 cases, 1055 controls). \*European Non-Finnish population (GnomAD v2.1.1). Abbreviations: CI = confidence interval, MAF = minor allele frequency, OR = odds ratio



**Figure 7.4 - Regional association plots and recombination rates at suggestive genomic loci. (C-E)** Regional association plots at 4: 44392571 (rs112161979) (**A**), 11: 5724803 (rs66481907) (**B**), 15: 27729149 (rs112721576) (**C**), 19:39029201 (rs11881082) (**D**), and 5:36376351 (rs7720520) (**E**). The index variants are indicated by a purple diamond and corresponding rsID. Linkage disequilibrium between the index variant and nearby variants, as measured by  $r^2$ , is colour-coded (dark blue:  $0 \leq r^2 < 0.2$ ; light blue:  $0.2 \leq r^2 < 0.4$ ; green:  $0.4 \leq r^2 < 0.6$ ; orange:  $0.6 \leq r^2 < 0.8$ ; red:  $0.8 \leq r^2 \leq 1$ ; grey: no  $r^2$  available). All plots were generated in <http://locuszoom.org/>.



**Figure 7.5 - Conditional analyses adjusted for the lead SNP at each suggestive GWAS loci.** For chromosome 4 (A) Unconditioned Manhattan plot (B) Manhattan plot conditioned on the *KCTD8* variant rs112161979. For chromosome 11 (C) Unconditioned Manhattan plot and (D) Manhattan plot conditioned on the *TRIM22* variant rs66481907. For chromosome 15 (E) Unconditioned Manhattan plot and (F) Manhattan plot conditioned on the *GABRG3* variant rs112721576. For chromosome 19 (G) Unconditioned Manhattan plot and (H) Manhattan plot conditioned on the *RYR1* variant rs11881082. For chromosome 5 (I) Unconditioned Manhattan plot and (J) Manhattan plot conditioned on the intergenic variant rs7720520 near *RANBP3L*.

### 7.3.5 Fine-mapping and functional annotation

I decided to focus on the top two lead genomic loci (Chromosome 4 and 11) for this analysis. Based on the single causal variant assumption (supported by the conditional analysis detailed in **Figure 7.5**) statistical fine-mapping was performed at these two loci with FINEMAP<sup>448</sup> and SuSiE<sup>447</sup>.

No consensus causal SNPs (supported by both fine-mapping techniques) were identified at the lead locus (Chromosome 4). However, SuSiE nominated rs990356 as the likely causal SNP (posterior probability 1), a 3' UTR variant in *YIPF7* (Yip1 Domain Family Member 7) (**Table 7.6** and **Figure 7.6A**). Mapping the Chromosome 4 locus against genomic regulatory elements did not show any significant signals.

Fine-mapping of the Chromosome 11 locus also failed to demonstrate a consensus causal SNP across the two fine-mapping algorithms. SuSiE nominated three SNPs as potentially causal, while FINEMAP nominated one (**Table 7.6** and **Figure 7.6B**). Of particular interest was the rs7397032 SNP identified by SuSiE; this is in high LD ( $D'$  0.95,  $R^2$  0.86) with the lead SNP from the GWAS (rs66481907), is a 3'UTR variant in *TRIM22*. Mapping this region against available genomic regulatory elements demonstrated that the lead SNP, and surrounding SNPs in high LD, sit within a conserved transcription factor binding site, supporting that this locus is involved in transcriptional regulation of surrounding genes (Bottom panel in **Figure 7.6B** ).



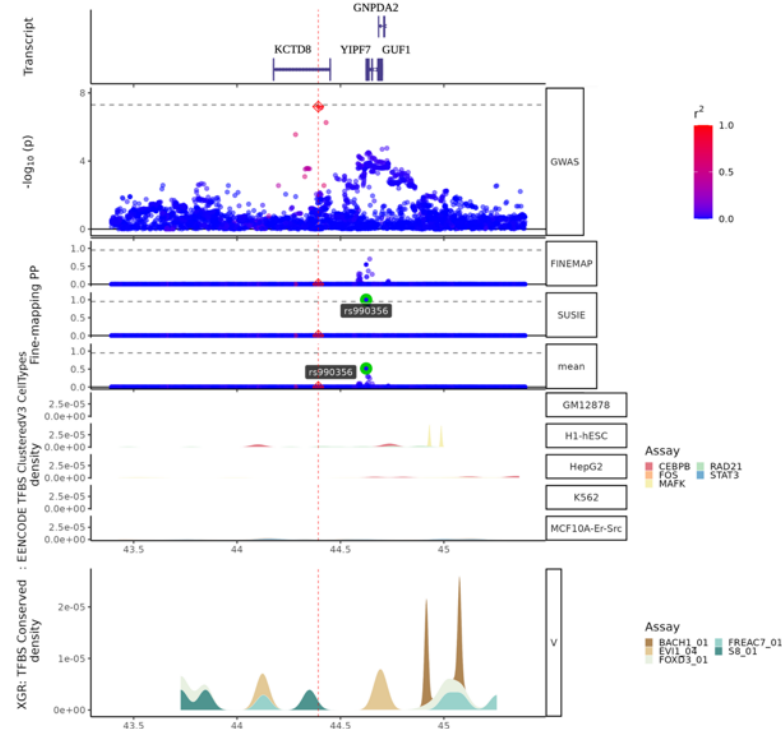
**Table 7.6 - Fine-mapping results of *KCTD8* and *TRIM22* loci**

SNP	CHR	N	T_stat	FINEMAP.CS	FINEMAP.PP	SUSIE.CS	SUSIE.PP	Support	Consensus	Mean.PP
rs990356	4	2668	3.80	1	0	2	1	1	FALSE	0.33
rs11038885	11	2692	4.32	1	0	3	1	1	FALSE	0.25
rs7397032	11	2688	4.29	1	0	2	1	1	FALSE	0.25
rs76253329	11	2678	2.26	1	0.96697	0	0	1	FALSE	0.15

Abbreviation = N, Sample size to do fine-mapping; t\_stat = test statistic; CS = Credible Set; PP = Posterior Probability. Mean.PP = the mean posterior probability from the two fine-mapping posterior probabilities

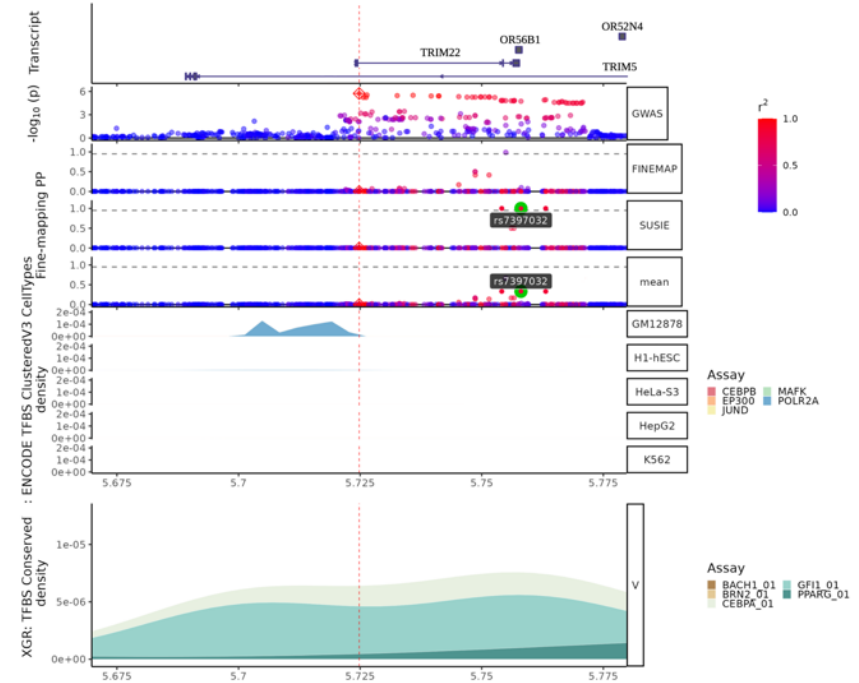
A

Locus: KCTD8 (SNPs=5,711; zoom=3x)



B

Locus: TRIM22 (SNPs=6,997; zoom=20x)

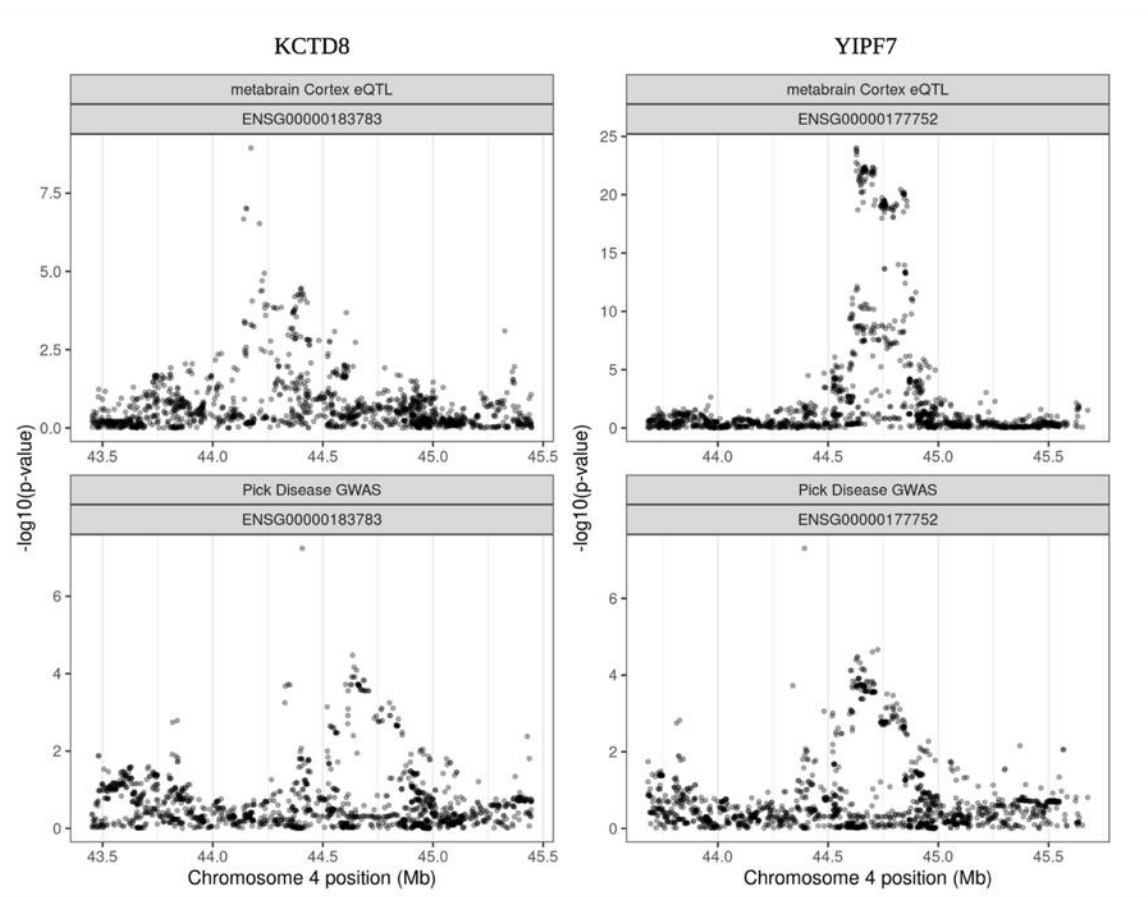


**Figure 7.6 - Fine-mapping of the two lead loci.** (A) shows the chromosome 4 locus (*KCTD8*) and (B) shows the chromosome 11 (*TRIM22*) locus. Transcript plots are shown in the top row. The second row shows the GWAS results with the  $\log_{10} p$  value for each SNP (x-axis). The next three rows show the cross tools fine-mapping output (FINEMAP and SuSiE) and the consensus results (mean) respectively. For the fine-mapping results the x-axis represents the per SNP posterior probability (PP). The final two rows show transcription factor binding site (TFBS) data for the region. Row six shows TFBS clustered by cell type from ENCODE and the final row the TFBS density conserved across species from XGR (human/mouse/rat).

### 7.3.6 Colocalisation analysis

To try to further delineate the effects of the chromosome 4 and 11 loci on regulation of gene expression, colocalisation analysis was performed using cortical cis-eQTLs from the MetaBrain dataset<sup>452</sup>. There was no evidence of colocalisation (defined as a PP.H4 of  $> 0.85$ ) between the two lead PiD GWAS loci and eQTLs for genes within  $\pm 1$ Mb of the locus. At the chromosome 4 locus *KCTD8* had a PP.H4 of colocalisation of 0.02, while the three genes downstream *YIPF7*, *GUF1* and *GNPDA2* had PP.H4s of 0.74, 0.81 and 0.72 respectively. At the chromosome 11 locus, *TRIM22* and *TRIM5* both had PP.H4s of 0.04. Given the suggestion (though not consensus) that the causal SNP for the chromosome 4 signal may be mediated by rs990356 in the 3' region of *YIPF7*, regional association plots for the eQTL signal from *KCTD8* and *YIPF7* and the PiD GWAS signal were plotted (**Figure 7.7**). Visual inspection of these plots demonstrates that the GWAS signal is more closely aligned with the *YIPF7* than the *KCTD8* eQTL signal and suggests that the GWAS signal at the chromosome 4 locus could be mediated by dysregulation of *YIPF7* gene expression. However, this association did not meet the predefined threshold of certainty for colocalisation and so with the current sample size this cannot be confirmed. Given the chromosome 11 lead SNP (rs66481907) is a sQTL for *TRIM22* in nerve-tibial tissue, I wanted to explore whether this region could have a role in alternative splicing of the gene in cortical tissue, so performed colocalisation analysis using cortex sQTLs from the GTExv8 dataset<sup>351</sup>. Again, there was no evidence for colocalisation of the GWAS signal in this region and sQTLs for *TRIM22* in cortical tissue (PP.H4: min 0.036, max 0.055).

Given that there was not a genome wide significant signal at either of the loci, it is possible that this study was underpowered to find colocalisation support for dysregulated gene expression in these regions. Equally, it may be possible that dysregulation of gene expression by these variants is not their mechanism of action. Although there was a suggestion of an association for *YIPF7* (PP.H4 0.74), *GUF1* (PP.H4 0.81) and *GNPDA2* (PP.H4 0.72) at the chromosome 4 locus, the current data does not provide strong support for colocalisation in cortical tissue. Likewise, there was no evidence of colocalisation of the chromosome 11 signal with cortex sQTLs for *TRIM22*.



**Figure 7.7 - Regional plots from PiD GWAS and MetaBrain cis-eQTLs for *KCTD8* and *YIPF7* genes.** The PP.H4 of there being a shared causal variant associated with both PiD (bottom panels) and regulation of gene expression (top panels) was 0.02 for *KCTD8* (left side) and 0.74 for *YIPF7* (right side).

### 7.3.7 Transcript and protein expression of suggestive genes

Given the suggestion that the causal SNP at the chromosome 4 locus was located in *YIPF7*, I investigated the gene and protein expression profile of this gene in addition to *KCTD8*, alongside *TRIM22* located at the chromosome 11 locus.

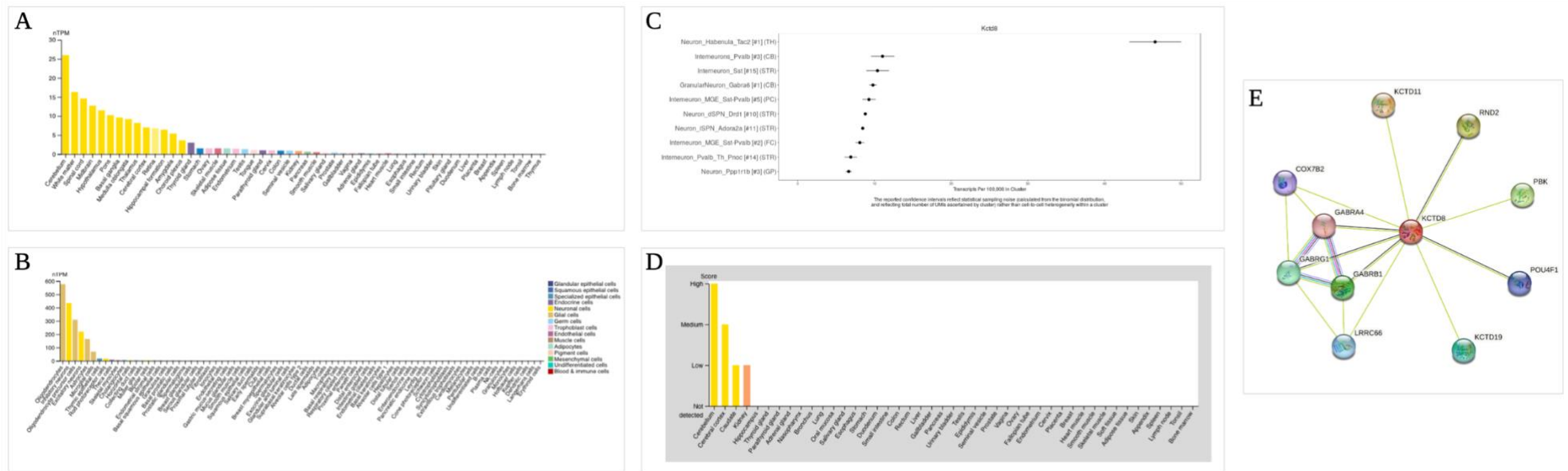
*KCTD8* transcript expression is enriched in the central nervous system (**Figure 7.8A**); the cerebellum has the highest expression levels (normalised transcripts per million (nTPM) 16), with expression in the cerebral cortex, amygdala and hippocampus at 7.0, 6.4 and 5.4 nTPM respectively. Human single cell RNA-seq data shows that oligodendroglia are the most enriched cell type (576.7 nTPM) followed by inhibitory neurons (435.3 nTPM), then excitatory neurons (219.9 nTPM). Astrocytes and microglia are the least enriched of the brain cell types (164.2 and 69.3 nTPM respectively) (**Figure 7.8B**). Mouse scRNA data suggests that it is specifically

habenula neurons that most highly express *KCTD8* (**Figure 7.8C**). At the protein level, in keeping with RNA expression levels, immunohistochemistry demonstrates that *KCTD8* is most highly expressed in the cerebellum and the cerebral cortex (**Figure 7.8D**). **Figure 7.8E** shows the predicted protein interaction network for *KCTD8* protein, highlighting its interactions with GABRG1, GABRA4 and GABRB1 all of which are subunits of the GABA<sub>a</sub> receptor in the human brain.

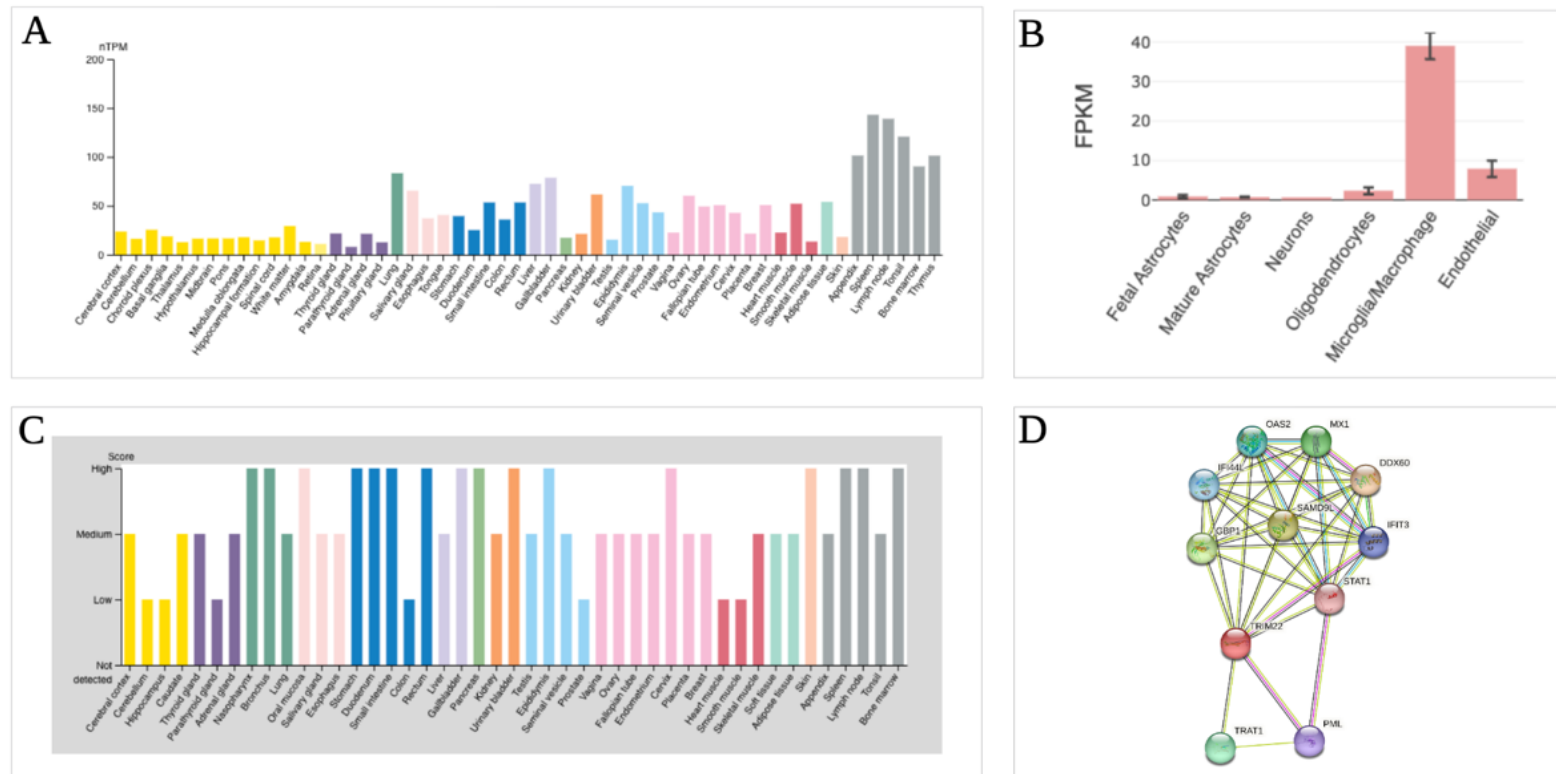
The *TRIM22* gene on the other hand has low tissue specificity and is ubiquitously expressed throughout the body, with highest expression in lymphoid tissues (spleen nTPM 143.1) (**Figure 7.9A**). It is, however, still expressed in the brain with transcript expression in the brain ranging from 12.8 nTPM in the thalamus to 29.3 nTPM in the white matter. This low specificity is also reflected at the level of protein expression, with levels in the brain ranging from low in the cerebellum and hippocampus to medium in the cerebral cortex and caudate (**Figure 7.9B**). The Brain RNA-seq data shows that *TRIM22* is predominantly expressed microglia within the brain (FPKM;  $39.07 \pm 3.36$ ), with little expression in neurons (FPKM; 0.632) (**Figure 7.9C**). Unfortunately, *TRIM22* is not included in the mouse scRNA dataset so I was unable to interrogate single cell RNA specificity. **Figure 7.9D** shows the predicted protein interaction network for *TRIM22*. Of interest is the interaction with promyelocytic leukaemia protein (*PML* also known as *TRIM19*) a protein that promotes clearance of misfolded proteins (including mutant ataxin-7 in Spinal Cerebellar Ataxia Type 7)<sup>476</sup> via the ubiquitin proteasome system (UPS)<sup>477</sup>. *PML* has been shown to colocalise with *TRIM22* in nuclear complexes upon IFN- $\gamma$  induced *TRIM22* expression<sup>478</sup>.

Although *YIPF7* is predominantly expressed in skeletal muscle (nTPM 84.9), with low transcript levels in the bulk brain tissue (cerebellum 5.4 nTPM, cerebral cortex 1.3 nTPM) (**Figure 7.10A**), interestingly antibody staining in the Human Protein Atlas suggests that the protein is present in the brain with high levels detected in glial cells in the hippocampus and medium levels in neurons of the cerebral cortex, caudate and cerebellum (**Figure 7.10B**). However, this discrepancy between RNA and protein expression does raise concerns about the specificity of the antibody used in IHC and will need to be confirmed in follow-up studies. In addition, although overall RNA expression is low in the brain, its expression is enriched specifically within excitatory and inhibitory neurons (**Figure 7.10C**). Further investigation of *YIPF7*'s protein interaction network yields some interesting findings (**Figure 7.10D**). Firstly, it is

predicted to interact with both *YIF1A* and *YIF1B* to form the YIPF complex 1<sup>467</sup>, that is localised to the early compartment between the endoplasmic reticulum (ER) and the Golgi, and participates in the anterograde recycling of proteins from the ER to the cell membrane. *YIF1A* has been shown to bind *VAPB*; the *VAPB* P56S mutation is known to cause amyotrophic lateral sclerosis<sup>479</sup>. Homozygous mutations in *YIF1B* have recently been shown to cause a progressive encephalopathy with global developmental delay and cognitive impairment<sup>480</sup>. Secondly, *YIPF7* is also predicted to interact with *dystrophica myotonica-protein-kinase* (*DMPK*), mutations in which cause myotonia dystrophy type 1 (DM1). This is of particular interest given the known association between DM1 and predominantly 3R-tau pathology in the brain at post-mortem<sup>207,481</sup>.



**Figure 7.8 - *KCTD8* RNA and protein expression levels.** Comparison of (A) Bulk RNA expression across different tissue types (normalised transcripts per million (nTPM)) (B) single cell RNA (scRNA) expression (nTPM) across different cell types (C) Mouse scRNA expression (nTPM) across cell subtypes (D) Protein expression (score: high, medium, low or not detected on immunohistochemistry) by tissue type (E) Protein-protein interaction network of KCTD8 (predicted interactions: green lines = text mining, black = co-expression; known interactions: pink = experimentally determined, light blue = curated databases; purple = protein homology ). (A-B) Consensus data from Human Protein Atlas, (C) data from DropViz (D) IHC data from Human Protein Atlas (E) Protein interaction data from string-db.org.



**Figure 7.9 - *TRIM22* RNA and protein expression levels.** Comparison of (A) Bulk RNA expression across different tissue types (normalised transcripts per million (nTPM)) (B) Brain RNA expression (nTPM) across different cell types (C) Protein expression (score: high, medium, low or not detected on immunohistochemistry) by tissue type (D) Protein-protein interaction network of *TRIM22* (predicted interactions: green lines = text mining, black = co-expression; known interactions: pink = experimentally determined, light blue = curated databases; purple = protein homology). (A) Consensus data from Human Protein Atlas, (B) data from BrainRNA-seq.org (C) IHC data from Human Protein Atlas (D) Protein interaction data from string-db.org.



**Figure 7.10 - YIPF7 RNA and protein expression levels.** Comparison of (A) Bulk RNA expression across different tissue types (normalised transcripts per million (nTPM)) (B) Brain RNA expression (nTPM) across different cell types (B) Protein expression (score: high, medium, low or not detected on immunohistochemistry) by tissue type (C) Single cell RNA (scRNA) expression clusters for brain cell types (D) Protein-protein interaction network of YIPF7 (predicted interactions: green lines = text mining, black = co-expression; known interactions: pink = experimentally determined, light blue = curated databases; purple = protein homology ). (A) Consensus data from Human Protein Atlas, (B) IHC data from Human Protein Atlas (C) Consensus scRNA data from Human Protein Atlas (D) Protein interaction data from string-db.org.

## 7.4 Discussion

In this chapter using samples collected through the PIC I have conducted the first GWAS in 294 autopsy confirmed PiD cases to identify risk factors for developing the disease. Unfortunately, no disease-associated common variants reached genome-wide significance. Given a previous GWAS of 219 autopsy-confirmed corticobasal degeneration (a 4R tauopathy) did identify significant disease-associated common variants<sup>169</sup>, this may indicate that the contribution to disease risk in PiD is governed by genetic modifiers with smaller effect sizes that this study was underpowered to detect. Although no SNPs reached genome wide significance, there were five suggestive genomic loci with nominal association with disease; the top two genomic loci were located on chromosome 4 in the *KCTD8* gene (rs112161979,  $p = 6.37 \times 10^{-8}$ ), and chromosome 11 in the *TRIM22* gene (rs66481907,  $p = 1.83 \times 10^{-6}$ ) respectively.

The association of the *MAPT* H2 haplotype with risk of PiD was confirmed through the candidate variant analysis. A limitation of the direct genotyping method used in **Chapter 6**, was that without genome-wide SNP data for the controls it was not possible to regress out genetic principal components, and so there was a possibility that population stratification could have influenced the results. This study shows that the association remains after controlling for population stratification, and actually becomes slightly stronger after accounting for this (OR: 1.52, 95% CI: 1.18 - 1.97,  $p = 0.001$  vs. OR: 1.35, 95% CI: 1.12-1.64,  $p = 0.0021$  respectively). There was no evidence that the H2 haplotype was associated with age at onset, or survival. The *ApoE4* genotype was also not associated with age of onset or survival in PiD, which is of interest given that previous work has suggested that there is a lower age of onset in FTLT-tau cases<sup>482</sup> with the ApoE4 allele. However, in their study the FTLT-tau cohort contained both individuals with *MAPT* mutations (53/104) and sporadic FTLT-tau (51/104) with no break-down of 3R versus 4RT pathology. It is therefore possible that, especially given the findings in this study, that this lower age of onset may be driven by non-3R tau pathology.

Variants known to be associated with the 4R tauopathies, PSP and CBD, are not associated with PiD, with the exception of *MOBP* (OR 0.76, 95% CI 0.59 - 0.98,  $p = 0.03$ ), though this effect did not survive correction for multiple testing. Larger sample sizes will be needed to prove or refute the association of *MOBP* with PiD. At present

this data suggests that different variants drive disease risk in the 3R and 4R tauopathies, implying that there are different underlying pathophysiological mechanisms in PiD compared to the 4R tauopathies.

The identification of a deltaK280 mutation in one of the cases was confirmed by direct sequencing, and re-checking the post-mortem pathology confirmed that this case met the PIC criteria for PiD pathology, with minimal AD co-pathology. The pathogenicity of the deltaK280 mutation, first described in 1999<sup>483</sup>, is a topic of long-standing and ongoing debate within the field. Although cell-based work shows it is likely to be fibrillogenic (at least for the 4R transcripts) and decreases the 4R/3R tau isoform ratio<sup>484–486</sup>, segregation with underlying 3R tau pathology has yet to be shown and in one deltaK280 case the underlying pathology was shown to be AD rather than PiD<sup>487</sup>. My finding of a deltaK280 mutation in a case with 3R tau PiD pathology, in the absence of 4R tau on IHC, and minimal AD pathology is in keeping with the hypothesis that this mutation abolishes a splice enhancer element in exon 10 leading to increased 3R tau transcripts. However, in the absence of data showing segregation within family members the debate on deltaK280 pathogenicity in PiD pathology remains open.

The lead signal at the chromosome 4 locus comes from an intronic SNP located within the *KCTD8* gene. There are no deleterious coding variants in LD with the lead SNP, no colocalisation of the GWAS signal in this region with brain eQTLs of *KCTD8* or the three genes c.500kb downstream (*YIPF7*, *GNPDA2* or *GUF1*), and fine mapping was inconclusive with regards to a consensus causal SNP. However, given the colocalisation analysis was likely to be underpowered due to the absence of genome-wide significance in the GWAS data, we cannot exclude that this variant does not actually affect expression of these downstream genes. *YIPF7* is an interesting potential candidate gene at this locus given its predicted interactions with proteins that when mutated are known to cause neurodevelopmental disorders as well as ALS. The predicted interaction with the DM1 gene *DMPK* is also intriguing given the known presence of 3R tau at post-mortem in these patients. However, with the data in this study it is not possible to confirm the relevance of this gene, and will require better powered genetic studies alongside functional studies to confirm or refute its relevance.

The *KCTD8* gene, a member of the potassium channel tetramerisation domain family, that acts as an auxiliary subunit facilitating GABA<sub>b</sub> receptor expression<sup>462</sup>, is highly

expressed within the brain, predominantly in neurons and oligodendrocytes, and the protein it encodes has been shown using IHC to be present in the cortex. scRNA analysis suggests its expression is specifically enriched in habenula neurons, which is supported by in situ hybridisation analysis of *KCTD* receptor transcripts in the mouse brain showing its particular abundance in the medial habenula followed by the subiculum of the hippocampus<sup>488</sup>. The habenula and the subiculum are of interest with regards to PiD. The habenula is affected by neurodegeneration in behavioural variant FTD (bvFTD) (the most common presentation of PiD) showing a 29% lower volume compared to controls<sup>489</sup>, and its degeneration can lead to perseveration or disinhibition and impulsivity<sup>22</sup>, symptoms commonly seen in bvFTD. The subiculum is commonly affected by Pick's pathology with high densities of Pick bodies found in this part of the hippocampus at post-mortem<sup>490</sup>. The *KCTD* family of proteins are currently poorly characterised, though are increasingly recognised to be involved in a range of neurocognitive, neurodevelopmental and neuropsychiatric disorders. These include mutations in *KCTD3* (global developmental delay, seizures and cerebellar hypoplasia)<sup>491,492</sup>, *KCTD13* (autism and schizophrenia)<sup>493–495</sup>, *KCTD17* (myoclonus-dystonia)<sup>496–498</sup>, *KCTD12* (bipolar 1 disorder)<sup>499</sup>, and *KCTD7* which can cause either a severe progressive myoclonic epilepsy syndrome (EPM3)<sup>500</sup> or neuronal ceroid lipofuscinosis<sup>501</sup> depending on the specific mutation. *KCTD8*, specifically, acts as an auxiliary subunit of *GABA<sub>b</sub>* receptors and has been shown to facilitate their axonal expression in habenula cholinergic neurons<sup>462</sup>. GABA deficits in FTD have long been recognised, with evidence that GABAergic neurons are markedly reduced in FTD at post-mortem<sup>502</sup>, with PET<sup>503</sup> and MRS imaging studies demonstrating GABAergic deficits in vivo<sup>504</sup>. Overall, the localisation of the lead SNP, and the gene expression profiles described above make *KCTD8* the most plausible candidate gene at the chromosome 4 locus, though further work will be needed both to validate this signal, and also delineate the mechanisms by which it contributes to PiD pathology.

The GWAS signal at chromosome 11 implicates *TRIM22*. The lead SNP rs66481907 is a sQTL in nerve tibial tissue for *TRIM22*, is located in alternatively spliced transcripts with retained introns (non-functional) and others targeted for nonsense mediated decay, and has a CADD score of 10.02 placing it in the top 10% most deleterious variants in the genome. *TRIM22* is expressed both at the transcript and protein level in the brain and is enriched within microglia in contrast to *KCTD8* which

is predominantly expressed in neurons. The TRIM family of proteins, the majority of which have E3 ubiquitin ligase activity, have a wide range of functions within cellular processes including eliminating misfolded proteins (via autophagy<sup>505,506</sup>, the ubiquitin proteasome system (UPS)<sup>196,468</sup>, and endoplasmic-reticulum associated degradation (ERAD)<sup>507</sup>), antiviral activity<sup>508</sup>, and regulation of the NF- $\kappa$ B/NLRP3 inflammasome pathway<sup>509</sup>. Mutations in TRIMs are increasingly recognised as a cause of a wide range of disease pathologies including a more aggressive phenotype in PSP (*TRIM11/17*)<sup>195</sup>, cerebral small vessel disease (*TRIM47*)<sup>510</sup>, and limb girdle muscular dystrophies (*TRIM32*)<sup>511</sup>. Although *TRIM22* was first identified through its anti-viral properties in HIV infection<sup>512</sup>, more recent work has demonstrated its role autophagy through interaction with autophagy regulators ULK1 and Beclin1<sup>505,506</sup>, as well as effectively promoting elimination of misfolded proteins via the UPS during cell transformation<sup>513</sup>. Consistent with this role is the finding that *TRIM19/PML*, which promotes clearance of misfolded proteins (including ataxin-7 in SCA7) via the proteasome<sup>476</sup>, colocalises with *TRIM22* in nuclear bodies under IFN- $\gamma$  stimulation<sup>478</sup>. Overall, *TRIM22* is a biologically plausible candidate gene for risk of PiD, based on the hypothesis that variation at the chromosome 11 locus modifies the function of the protein (potentially through nonsense mediate decay or alternative splicing of gene transcripts), leading to decreased degradation of toxic 3R tau protein via the UPS and / or the autophagy pathway(s).

There are a number of limitations to this study. Ideally a GWAS should have a two-stage design with a discovery phase that identifies a “genome-wide significant” signal, followed by a replication phase where this signal remains significant after correction for multiple testing. Despite collecting all of the known autopsy-confirmed cases of PiD worldwide, the sample size was still small by GWAS standards. The study is likely underpowered even for this discovery GWAS, and there was certainly no way to conduct a two-stage study to validate the findings. The lack of clinico-pathological correlation meant it was also not possible to meaningfully enrich for PiD pathology from the main FTD clinical presentations. This would require in vivo biomarkers to distinguish PiD pathology from 4R tauopathies (including PSP and CBD) and also from TDP-43 pathology at a minimum. Given the estimated number of autopsy-confirmed cases per year that will be collected from the PIC brainbank network is only 10-15 per year, without these biomarkers it will not realistically be feasible to replicate

the findings from this study within the next 5-10 years. I hope that not only will this work stimulate hypotheses for further functional and biomarker studies, but also that the PIC will provide a network for neuroscience centres to coordinate collection of future cases with a standardised set of pathological diagnostic criteria.

Another limitation of this study was that apart from direct genotyping of the six *MAPT* variants, there was no direct genotyping of lead SNPs at the suggestive loci and thus genotypes were inferred from imputed data. Although stringent QC thresholds were used, including an  $R^2$  of  $\geq 0.7$  and the genotype posterior probability was  $\geq 0.9$  for imputed genotypes, small errors in the frequency of minor alleles in cases versus controls could have a significant impact on variant effect size and therefore the  $p$  value. This is especially true for the lead variant on chromosome 4 which was a relatively rare variant with minor allele frequency of 0.011 (1.1%) in controls (0.016 [1.6%] in European Non-Finish population in GnomAD), and 0.040 (4%) in cases. In future work it will be necessary to directly sequence the suggestive regions highlighted in this study both to confirm the genotypes, and also check (especially for the chromosome 4 locus) whether the GWAS signal is actually tagging a deleterious mutation.

## 7.5 Conclusions and future work

In this chapter I have performed the first GWAS with the aim of identifying the genetic drivers of disease risk in PiD. The data confirms that the *MAPT* H2 haplotype is associated with PiD, as opposed to the more common H1 haplotype in PSP and CBD. Known risk variants for the 4R tauopathies are not associated with disease, which suggests that the underlying genetic architecture of disease risk for PiD is distinct. This has important implications for the future development of therapeutics to treat PiD, and emphasises the need for PiD specific biomarkers to identify these individuals in life. *KCTD8*, the most plausible gene at the lead locus, modulates GABA<sub>b</sub> receptor expression within anatomically relevant regions of the brain and implicates dysregulation of the GABAergic neurons as an important driver of disease pathology. This is supported by the other suggestive association within the *GABRG3* gene (a GABA<sub>a</sub> receptor subunit) on chromosome 15. In addition, common variation in *TRIM22* may also play a role in disease pathogenesis, potentially through perturbation

of the UPS and its ability to eliminate toxic tau species, representing a potential target for disease modifying therapies.

Future work should focus on further GWASs with larger sample sizes to confirm or refute the findings from this study, whole genome sequencing of the lead loci to identify if rare deleterious variants are driving the signals here, and functional studies to unravel how these genes may be contributing to PiD risk.

# Chapter 8: Genetic determinants of survival in Pick's disease: a genome-wide association study

The contents of this chapter form part of a manuscript which is currently in preparation for submission.

## 8.1 Introduction

Pick's disease (PiD) is a rare and sporadic neurodegenerative tauopathy characterised by the aggregation of three-repeat (3R) tau in neuronal argyrophilic Pick bodies, predominantly in the frontal and temporal regions<sup>61</sup>. PiD usually develops in those over the age of 55 years, with average survival approximately ten years from symptom onset<sup>61–65,67</sup>. There is, however, significant variation in survival with some individuals surviving for more than 20 years<sup>61</sup>.

In progressive supranuclear palsy (PSP), a related tauopathy characterised by deposition of four repeat (4R) tau, tau seeds from PSP cases injected into primates induce a PSP syndrome with prion-like seeding and trans-synaptic spread of pathology<sup>514</sup>. Although less well studied than PSP, and characterised by deposition of a different isoform of tau (3R), “prion-like” seeding of tau followed by its spread through the brain could also underlie disease pathogenesis in PiD. Indeed, neuropathological studies are consistent with the sequential spread of tau pathology in PiD, with Irwin et al.<sup>62</sup> proposing that 3R tau deposition progresses in a sequential pattern starting in the limbic/paralimbic and neocortical regions progressing to the subcortical regions, followed by the primary motor cortex and finally the visual cortex. Clinical disease progression mirrors the sequential spread of pathology through the brain, with differing regional neuronal susceptibility to pathology proposed to account for the specific temporal sequence. Variation in survival between individuals may be explained, at least in part, by differing genetic susceptibility or resistance to the spread of the underlying pathology through different brain regions.

The majority of genome-wide studies in neurodegenerative disease have focused on identifying genetic susceptibility loci that increase the risk of disease (case-control studies) and have provided powerful insights into the underlying pathways that are associated with disease. In **Chapter 7** apply this methodology to identify genomic loci



that are implicated in PiD pathogenesis. However, there is increasing recognition that the genetic risk factors for developing disease may be different to the risk factors for clinical progression and survival. In PSP genetic variation at the *LRRK2* locus is associated with survival though not disease risk<sup>198</sup>, while in Parkinson's disease (PD) there is minimal overlap between variants associated with PD risk and PD motor or cognitive progression<sup>515</sup>. Given that future clinical trials in PiD will likely focus on disease modifying therapies that slow or halt disease progression in individuals that have already developed the disease (i.e. diagnosed with PiD), it is important to understand whether there are unique genetic risk factors for disease progression and survival independent of those for disease risk. This understanding may provide important biological insights that spur new hypotheses for therapeutic intervention.

The aim of this study was to identify genetic determinants of survival (symptom onset to death) in PiD cases of European ancestry, by performing a Cox-proportional hazards genome wide survival study (GWSS).

## 8.2 Methods

### 8.2.1 Study design and participants

For the initial selection, pathology confirmed PiD cases identified through the PIC (see **Total cases collected by the PIC**) were recruited from 31 international clinical or pathological research centres in the UK, France, Italy, Netherlands, Germany, Italy, Spain, Sweden, Australia, United States and Canada. For inclusion, all cases had to meet the strict PIC diagnostic criteria for PiD (**Figure 2.6**); as a minimum there needed to be the presence of Pick bodies with 3R tau positive and 4R tau negative inclusions. The additional presence of ballooned neurons and positive Gallyas staining was preferred to confirm diagnosis. All samples were screened for the known *MAPT* mutations; the MCJ samples by direct sequencing, and the UCL samples by using the NBA that covers all known *MAPT* variants. Clinical and demographic data was collected for all cases, and included age at symptom onset, age at death and gender. This information was used to calculate the total disease duration, defined as age at death – age at symptom onset. Age at symptom onset was defined as the age at which first symptoms appeared, including initial cognitive dysfunction in judgment, language, or memory, or changes in behaviour or personality. Only samples that had

clinical information to allow calculation of disease duration (age at symptom onset and age at death) were included in the study.

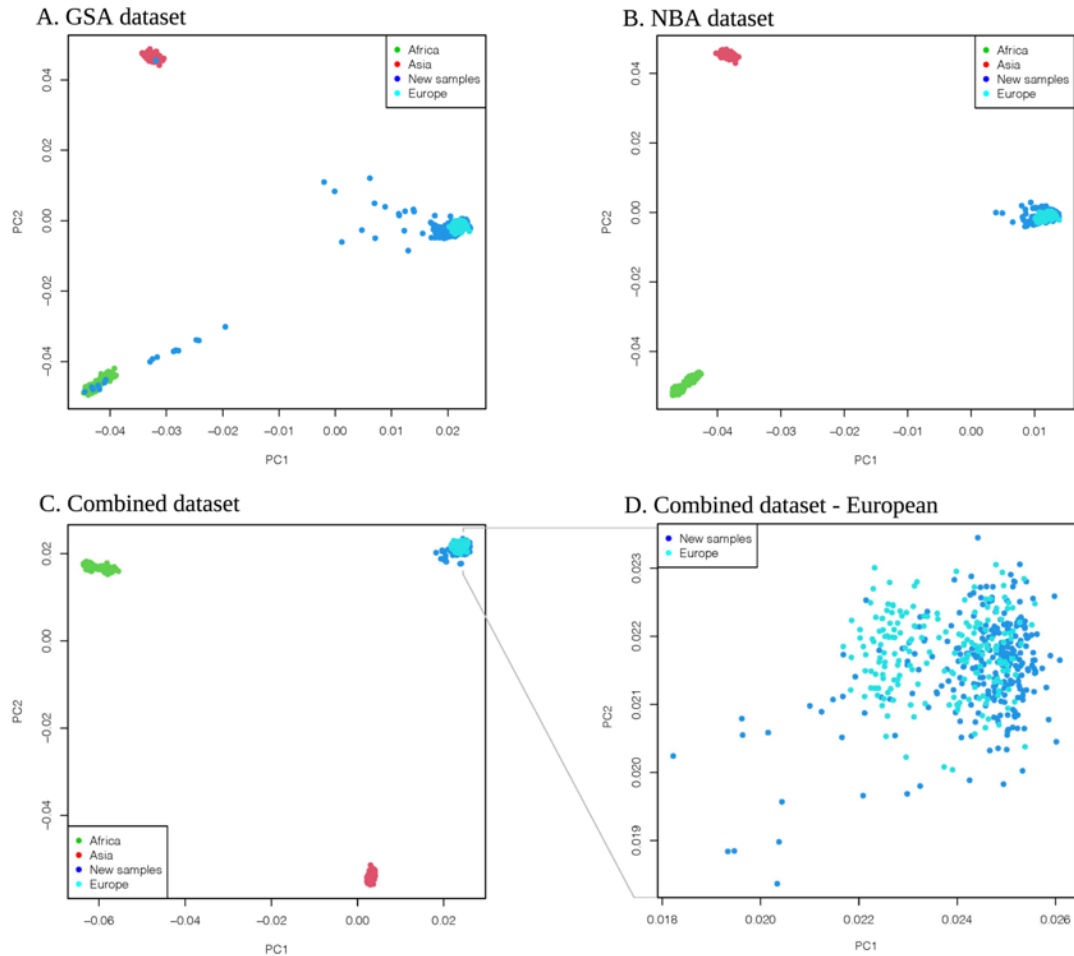
The appropriate institutional review boards for each site approved the study, and written informed consent was obtained for each participant.

### **8.2.2 DNA extraction and Genotyping**

DNA was extracted and samples were genotyped as detailed in Chapter 2 (**DNA preparation** and **Whole-genome microarray genotyping**).

### **8.2.3 Quality control and imputation**

I performed standard pre-imputation quality control, imputation and post-imputation quality control as detailed in Chapter 7 (**Quality control and imputation**) The only QC step that differed in this study was the application of the variant missingness rate threshold. Due to the fact that the sample size was smaller in this study than the case-control GWAS, I applied a more stringent variant missing rate (<1%) threshold to try and prevent false positive signals. As before for each array I generated genetic principal components to identify sample ancestry, and excluded any sample greater than 6SD from the CEU population in HAPMAPv3 (Figure **8.1**).



**Figure 8.1 - Genetic principal component plots.** First two principal components of the GSA (A) and NBA (B) and combined dataset (C) plotted against the HapMap3 Genome Reference Panel. The NBA and GSA datasets are plotted before samples more than 6SD from the mean of the first 10 principal components were removed. For the combined dataset (C) all samples are within 6SD of the CEU reference population (outliers removed), which is expanded in the figure to the right.

## 8.2.4 Survival analysis

I carried out a time-to-event genome-wide survival study (GWSS) in R (version 4.05), using the Cox proportional hazards (CPH) function in the *survival* package (version 3.2.11; RRID:SCR\_021137; <https://CRAN.R-project.org/package=survival>), with an event defined as death which was regressed against each SNP along with covariates. This results in each SNP having a hazard ratio (HR), 95% confidence interval (CI) and a  $p$  value. I tested all SNPs for adherence to the proportional hazards assumption. The Schoenfeld residuals for all reported SNPs were not time dependent (Cox.zph  $p > 0.05$ ), demonstrating that the proportional hazards assumption was met. To identify the optimal number of covariates to include in the final model I performed a stepwise CPH regression analysis using the *stepwiseCox* function in the *StepReg* package, with

default settings (version 1.4.3; <https://cran.r-project.org/web/packages/StepReg/index.html>). The best model was selected such that the Akaike Information Criterion (AIC) was minimised, and the amount of variance ( $R^2$ ) explained by the model was maximised. Using this method, the final covariates selected were array, age at onset, and three principal components (PC2, 6 and 8) which gave an AIC of 2483 and an  $R^2$  of 0.13. I defined a genome-wide significant threshold at  $p < 5 \times 10^{-8}$ , with a threshold of  $p < 5 \times 10^{-6}$  for a suggestive (nominal) association. Variant positions are reported on human genome version 37 (GRCh37/hg19). I generated Manhattan plots of the PiD survival GWAS, and Kaplan-Meier (KM) survival curves of lead SNPs using R (version 4.0.5) and the `survfit` function (for KM survival curves) from the *survival* package (version 3.4.0, <https://cran.r-project.org/web/packages/survival/index.html>).

## 8.2.5 Genomic risk loci definition and gene mapping

As detailed in Chapter 2 (**Functional Mapping and Annotation of GWAS (FUMA)**), I used FUMA to annotate and functionally map the variants included in the GWAS. I used the same settings as previously detailed in Chapter 7 (**Genomic risk loci definition and gene mapping**).

## 8.2.6 Conditional analysis

To understand whether there were one or more variants at the same locus contributing to the signal at each genomic risk locus, I performed conditional analysis on genome wide loci single SNPs with a  $p < 5 \times 10^{-7}$  using a conditional and joint association approach<sup>446</sup>, as detailed in Chapter 7 (**Conditional analysis**).

## 8.2.7 Fine-mapping and functional annotation

I performed fine-mapping and functional annotation of the lead loci using the same methods as detailed in Chapter 7 (**Fine-mapping and functional annotation**). For the fine-mapping in addition to using SuSiE and FINEMAP, I also used coloc (version 5.1.0; <https://cloud.r-project.org/web/packages/coloc/index.html>)<sup>356</sup> and PolyFun-SuSiE (v. 1.0; <https://github.com/omerwe/polyfun>)<sup>516</sup>. Coloc, like SuSiE and FINEMAP, is a purely statistical fine-mapping method that does not incorporate functional data into the model. Coloc calculates an approximate Bayes factor (ABF)

for all the variants in a defined region around the index variant ( $\pm 500\text{kb}$ ); variants are ordered by their posterior probabilities and sequentially added to the credible set until the cumulative sum is  $>0.95$  (95% credible set). In contrast to the other three methods, PolyFun prioritises variants in enriched functional annotations by specifying prior causal probabilities in proportion to the predicted per-SNP heritabilities<sup>516</sup> (these priors are then used as input to statistical fine-mapping algorithms (in this case to SuSiE, hence the name PolyFun-SuSiE).

## 8.2.8 Colocalisation analysis

To investigate whether there was an overlap between the lead GWSS loci (which I defined in this study as any loci with lead SNP  $p < 5 \times 10^{-7}$ ) and expression quantitative trait loci (eQTLs), I performed a colocalisation analysis using the *coloc* R package for all genes within  $\pm 1\text{Mb}$  of the lead genomic loci SNP (version 5.1.0; <https://cran.r-project.org/web/packages/colocr/index.html>)<sup>356</sup>. More detail on this method is given in Chapter 2 (**Colocalisation analysis**). Given *coloc* calculates Bayes factors under the assumption that there is a single casual variant at a locus, I first performed conditional analysis, as detailed in Chapter 7 (**Conditional analysis**), to confirm that there were no additional independent signals and so ensure that this assumption of a single casual variant was met.

I used both cortex-specific cis-eQTLs from MetaBrain (<https://www.metabrain.nl/>)<sup>452</sup> and blood-specific cis-eQTL from the eQTLGen dataset (<https://www.eqtlgen.org/>)<sup>517</sup>. *Coloc* was run using default priors; these are the prior probabilities that any random SNP in the region is associated with trait 1 or trait 2,  $p_1=10^{-4}$  and  $p_2=10^{-4}$ . A threshold of  $p_{12}=5 \times 10^{-6}$  was used for the  $p_{12}$  prior, which is the probability that a SNP in the region is associated with both traits. Loci with a posterior probability of hypothesis 4 ( $\text{PP.H4} \geq 0.85$ ) were considered as significant evidence of colocalization between the GWSS and the eQTL traits (one shared causal variant). A more conservative threshold for PP.H4 than used in the original work by Giambartolomei et al<sup>356</sup> (0.75), in line with more recent work where a threshold of 0.85<sup>453</sup> or even 0.90<sup>446</sup> has been used for GWAS data.

## 8.2.9 Assessment of gene transcript and protein expression of lead genes

I assessed brain expression profiles of gene transcripts and encoded proteins highlighted by the GWSS, using the same methods described in Chapter 7 (Assessment of gene transcript and protein expression of lead genes).

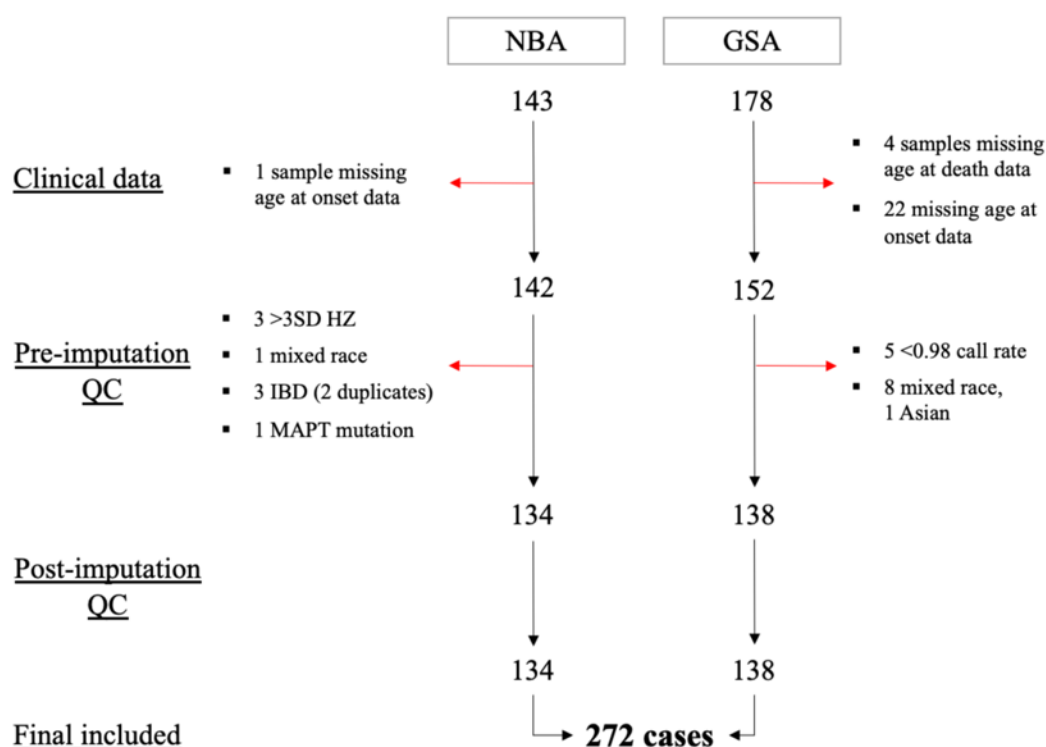
## 8.2.10 Candidate variant analysis

I extracted specific variants that have previously been identified in related diseases to check whether they showed any association with risk of PiD. This included variants identified in the primary tauopathies (PSP<sup>167,195,198</sup>, CBD<sup>169</sup> and primary aged-related tauopathy (PART)<sup>459</sup>) and clinically diagnosed FTD<sup>204</sup>. I checked for an association between *MAPT* haplotypes and PiD survival, by extracting the six *MAPT* variants that define the H1-subhaplotype structure (see **MAPT SNP Genotyping**). I also checked whether *ApoE* genotype<sup>460</sup> and *TMEM106B*<sup>518</sup> were associated with PiD survival. *APOE* genotypes were inferred from the imputed genotypes of rs7412 and rs429358 variants. Finally, I checked whether the lead SNPs from the five suggestive genomic loci from the PiD case-control GWAS (**Table 7.5**) were associated with survival.

# 8.3 Results

## 8.3.1 Cohort characterisation

At the time of this study 321 autopsy confirmed PiD cases from 31 recruitment sites, diagnosed pathologically according to the PIC neuropathological criteria (**Figure 2.6**), were available for inclusion in the study (**Table 7.1**). Of these, 143 cases were genotyped on the NBA (all from the UCL series), and 178 cases were genotyped on the GSA (171 from the Mayo series and 7 from Sydney collected as part of the UCL series). After quality control (pre and post imputation) and imputation, 272 cases (134 NBA and 138 GSA), covering 3,227,157 variants were available for survival analysis (**Figure 8.2**).



**Figure 8.2 - Overview of sample quality control for PiD GWSS.** Summary of samples excluded at each stage of quality control (QC) pre-imputation (PreImp) and post-imputation (PostImp). Final number of samples included in GWSS detailed at bottom of figure.

Demographics and basic clinical characteristics of the samples included in the GWSS are summarised in **Table 8.1**. The median age of disease onset was 58 years (min-max; 33-80 years), the median age at death of 69 years (min-max; 40-90 years) resulting in a median disease duration of 10 years (min-max; 3-25 years). There was very little variation between these figures across the two arrays, with both having a similar variation in disease duration. **Figure 8.3** shows the distribution of disease duration for the whole cohort, confirming a wide variation in survival for PiD. There was a higher percentage of men (172/272, 63.2%) vs women (100/272, 36.8%), with the difference more pronounced in the UCL series than the MCJ series (67.2% vs 59.4% respectively).

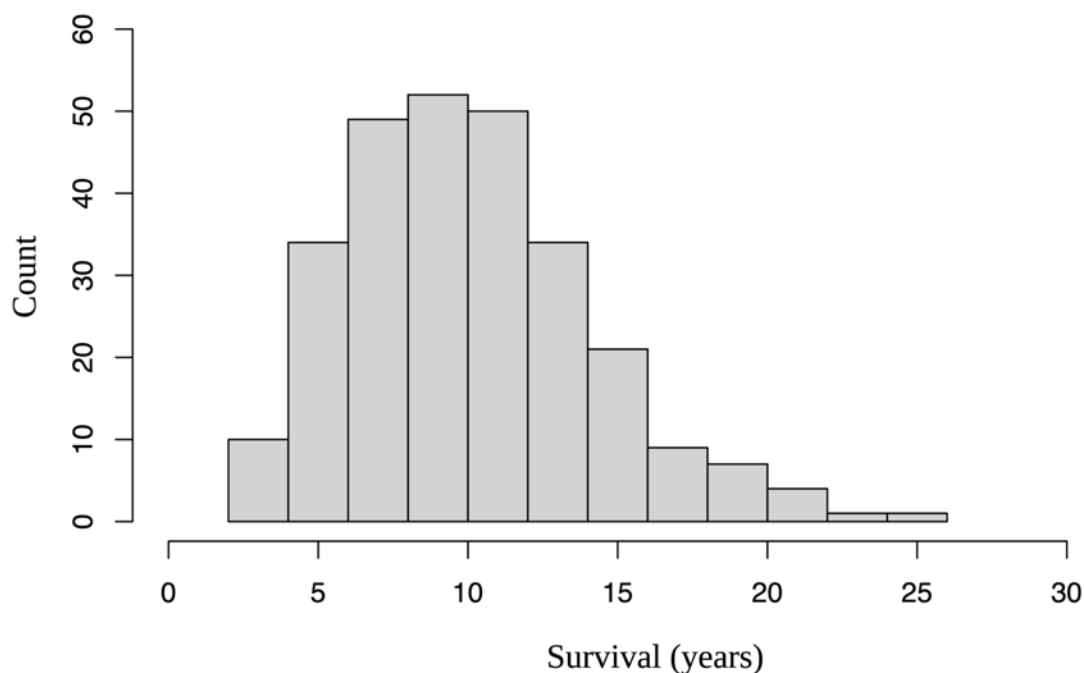
All samples included from MCJ and Sydney had negative *MAPT* mutation screening. All UCL samples were screened for known *MAPT* mutations covered by the Neurobooster array at the GenomeStudio genotype calling stage. The *MAPT* mutations identified and those excluded from further analysis have already been detailed in Chapter 7 (***MAPT* mutations**).

**Table 8.1 - Clinical characteristics of samples included in GWSS.**

	MCJ PiD series <sup>1</sup> (GSA)	UCL PiD series (NBA)	Combined PiD series
Sample (n)	138	134	272
Age (years)	70 (40, 90)	69 (41, 88)	69 (40, 90)
Age of disease onset (years)	59 (36, 79)	58 (33, 80)	58 (33, 80)
Disease duration (years)	10 (3, 25)	10 (3, 20)	10 (3, 25)
Sex			
- Male	82 (59.4%)	90 (67.2%)	172 (63.2%)
- Female	56 (40.6%)	44 (32.8%)	100 (36.8%)

The sample median (minimum, maximum) is given for age. Age represents age at death for Pick's disease cases.

<sup>1</sup>Includes 8 Australasian samples from the UCL series which were also genotyped on the GSA-v3.. Abbreviations: GSA = Illumina Global Screening Array v3, PiD = Pick's disease, MCJ = Mayo Clinic Jacksonville, NBA = Illumina NeuroBooster Array, UCL = University College London



**Figure 8.3 - Distribution of disease duration across PiD cohort.** Histogram of survival times (disease duration defined as time from first symptom onset to death in years) with survival plotted on the  $x$ -axis and count on the  $y$ -axis.



### 8.3.2 Targeted assessment of candidate variants

Variants were extracted that have either previously been associated with tauopathies and related diseases or identified in the case-control GWAS (**Table 7.5**) to test for an association with PiD. The results are summarised in **Table 8.2**. Of the 23 variants tested only rs7720520 (chromosome 5, nearest gene *RANDBP3L*), that was the lead SNP at one of the nominally associated genomic loci in the PiD case-control (HR 1.76, 95% CI 1.38-2.23,  $p = 4.5 \times 10^{-6}$ ) showed an association with survival (HR 1.24, 95% CI 1.04-1.47,  $p = 0.02$ ) though this did not survive correction for multiple comparisons (Bonferroni threshold:  $p < 0.002$ )

Table 8.2 - Candidate variant analysis using GWSS data.

Disease	Chr	BP	SNP	Effect allele	Nearest gene	HR	95% CI	Effect Allele Frequency		
								PiD survival	NFE	p value
Tauopathy	17	44081064	rs8070723	G (H2)	<i>MAPT</i>	0.99	0.80-1.21	0.3015	0.1972	0.91
Tauopathy	17	44019712	rs242557	A (H1c)	<i>MAPT</i>	1.02	0.84-1.22	0.3314	0.3725	0.87
AD	19	45411941	rs429358	C	<i>ApoE</i>	1.01	0.79-1.29	0.1562	0.1486	0.95
AD	19	45412079	rs7412	T	<i>ApoE</i>	0.88	0.67-1.14	0.0978	0.0767	0.33
PSP	1	228585562	rs564309	A	<i>TRIM11</i>	0.95	0.71-1.26	0.0882	0.0989	0.70
PSP	12	40413698	rs2242367	A	<i>SLC2A13</i>	0.97	0.80-1.18	0.2907	0.2820	0.78
PSP	1	180962282	rs1411478	A	<i>STX6</i>	1.04	0.88-1.24	0.4320	0.5881	0.64
PSP	3	39523003	rs1768208	T	<i>MOBP</i>	1.22	0.99-1.49	0.2371	0.2714	0.06
PSP	2	88895351	rs7571971	T	<i>EIF2AK3</i>	0.97	0.79-1.19	0.2708	0.2726	0.76
FTD-TDP	7	12283787	rs1990622	G	<i>TMEM106B</i>	1.11	0.94-1.31	0.4185	0.4055	0.21
FTD	11	87876911	rs302668	C	<i>RAB38</i>	1.04	0.88-1.24	0.3759	0.3569	0.63
FTD	11	87934068	rs16913634	A	<i>RAB38/CTSC</i>	0.94	0.67-1.32	0.0787	0.0965	0.74
FTD	6	32431147	rs9268877	A	<i>HLA-DRA/DRB5</i>	0.90	0.76-1.07	0.4265	0.4436	0.23
FTD	6	32429719	rs9268856	A	<i>HLA-DRA/DRB5</i>	1.14	0.89-1.45	0.2335	0.2759	0.29
FTD	6	32363215	rs1980493	C	<i>BTNL2</i>	0.81	0.56-1.17	0.1434	0.1530	0.26
PART	4	130085480	rs56405341	A	<i>JADE1</i>	0.86	0.71-1.05	0.2731	0.2896	0.14
CBD	2	39216873	rs963731	T	<i>SOS1</i>	1.11	0.75-1.66	0.0515	0.0483	0.59
CBD	8	29153777	rs643472	C	<i>lnc-KIF13B-1</i>	1.05	0.86-1.28	0.2331	0.2296	0.63
PiD	4	44392571	rs112161979	A	<i>KCDT8</i>	1.08	0.69-1.70	0.0392	0.016	0.74
PiD	11	5724803:	rs66481907	A	<i>TRIM22</i>	1.07	0.86-1.32	0.2118	0.120	0.56
PiD	15	27729149	rs112721576	G	<i>GABRG3</i>	0.83	0.55-1.25	0.0466	0.038	0.37
PiD	19	39029201	rs11881082	A	<i>RYR1</i>	1.14	0.83-1.56	0.0757	0.053	0.43
PiD	5	36376351	rs7720520	G	<i>RANBP3L</i>	1.24	1.04-1.47	0.4240	0.337	0.02 <sup>a</sup>

Cox proportional hazards model adjusted for age at onset, array and 3 PCs (PC2, 6 and 8) was used to study the association of candidate loci survival (time from symptom onset to death) in PiD (272 cases).<sup>a</sup> significant uncorrected for multiple comparisons (Bonferroni:  $p=0.05/23=0.002$ ). Abbreviations: CI = confidence interval, NFE = Non-Finnish Europeans (GnomAD), HR = hazard ratio

### 8.3.3 Identification of genetic determinants of survival in Pick's disease

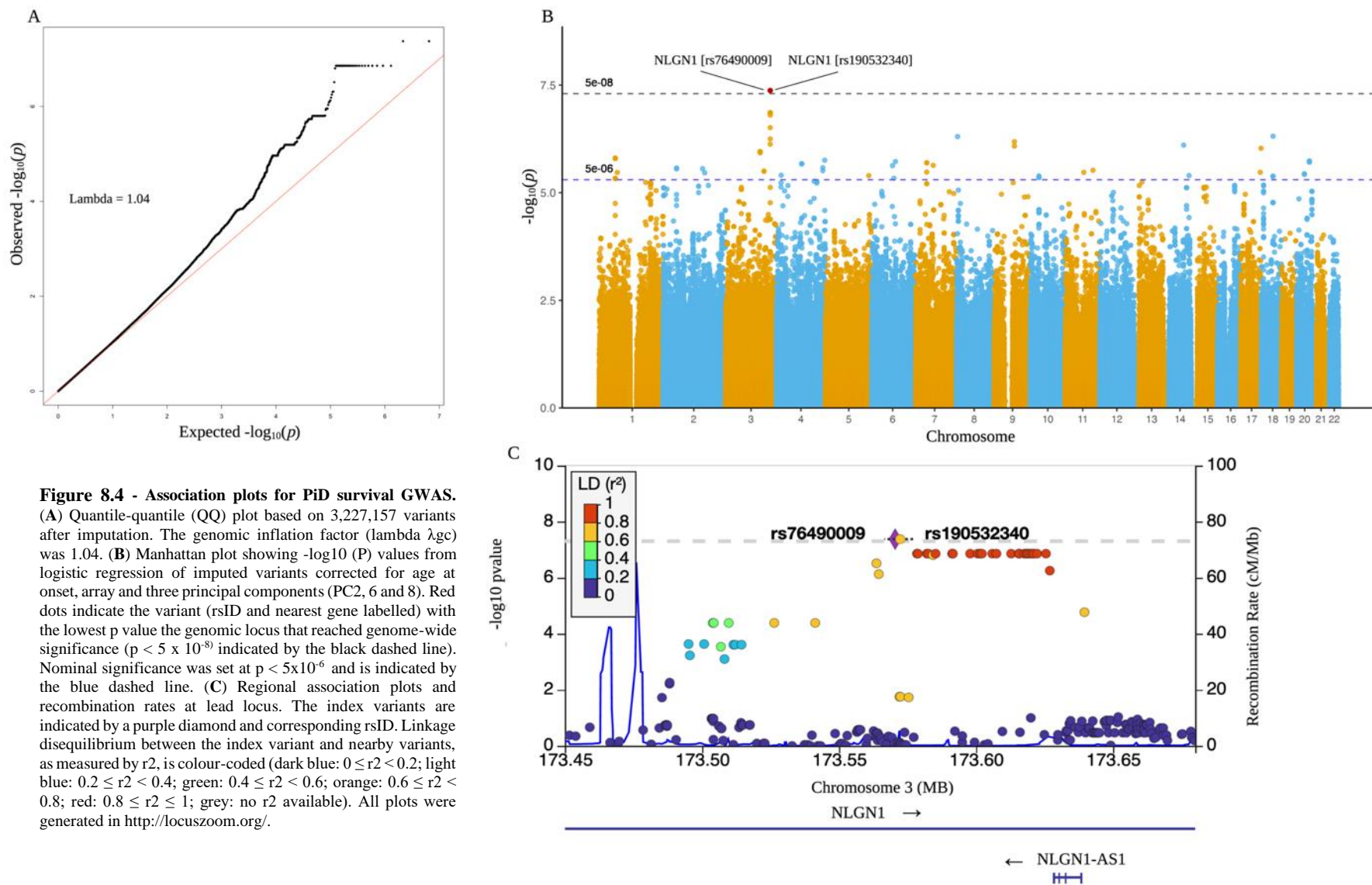
Using a CPH survival model, adjusting for genotyping array, age at onset and three genetic principal components (PCs 2, 6 and 8) to account for population substructure, I assessed the role of 3,227,157 variants in survival for those with underlying PiD pathology. The genomic inflation factor ( $\lambda$ ) was 1.04 suggesting minimal confounding from population stratification (**Figure 8.4A**). There was a genome-wide significant association signal at chromosome 3 (**Figure 8.4B**), with two SNPs in the *NLGN1* gene reaching the genome-wide significance threshold (**Table 8.3**). All genomic loci with a nominally significant lead SNP ( $p < 5 \times 10^{-6}$ ) identified using FUMA are summarised in **Table 8.4**.

**Table 8.3 - Top independent genome wide significant SNPs from PiD GWSS**

Chr	BP	SNP	Effect Allele	Nearest gene	HR	95% CI	Effect allele frequency		<i>p</i> value
							PiD Cohort	NFE	
3	173570502	rs76490009	C	NLGN1	6.06	3.11-11.82	0.0184	0.0169	4.23E-08
3	173572356	rs190532340	A	NLGN1	6.06	3.11-11.82	0.0184	0.0201	4.23E-08

Cox proportional hazards model adjusted for gender, age at onset, 3 PCs (PC2, 6 and 8) were used to study the association of candidate loci with survival (time from symptom onset to death) in PiD (272 cases). Abbreviations: BP = base pair, Chr = Chromosome, CI = confidence interval, NFE = Non-Finnish Europeans (GnomAD v2.1.1), OR = odds ratio

Both of the lead SNPs, rs76490009 (GRCh37/hg19 genome assembly, chr3:173570502) and rs190532340 (GRCh37/hg19 genome assembly, chr3:173572356) (**Figure 8.4C**) had the same HR and *p* value (HR 6.06, 95% CI 3.11-11.82,  $p = 4.23 \times 10^{-8}$ ) (**Table 8.3**). The SNPs were in high linkage disequilibrium as measured by  $D'$  (1) though had a lower  $R^2$  (0.6) due to differences in the underlying population minor allele frequencies (MAF) (rs76490009 = 0.0169 vs rs190532340 = 0.0201). In the PiD cohort the MAFs for each was identical (0.0185).

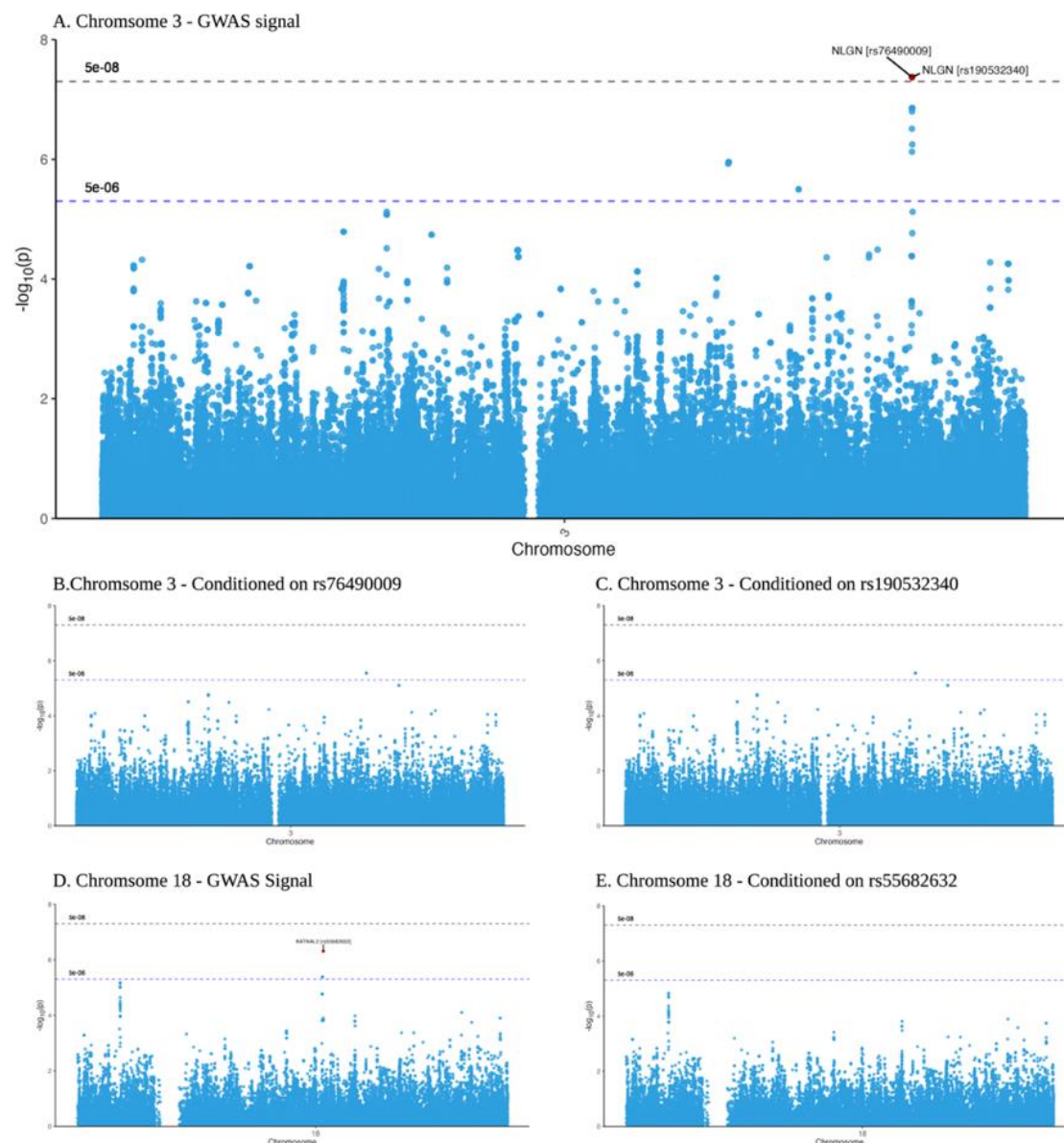


**Table 8.4 - Lead Genomic Loci from PiD GWSS ( $p < 5 \times 10^{-6}$ )**

Chr	BP	SNP	Effect Allele	Nearest gene	HR	95% CI	Effect allele frequency		
							PiD Survival	NFE	pvalue
3	173570502	rs76490009	C	NLGN1	6.06	3.11-11.82	0.0185	0.0169	4.23E-08
18	44593746	rs55682632	G	KATNAL2	4.89	2.63-9.07	0.0222	0.0246	4.86E-07
8	3552391	rs77405284	C	CSMD1	7.98	3.55-17.93	0.0129	0.0154	4.97E-07
9	78557508	rs5022354	A	PCSK5	2.38	1.69-3.36	0.0762	0.0956	6.53E-07
14	77460200	rs12435060	G	IRF2BPL:RP11-7F17.1	2.78	1.85-4.18	0.0500	0.0506	7.86E-07
17	78864497	rs2672883	T	RPTOR	7.29	3.3-16.11	0.0130	0.0201	9.30E-07
3	134279279	rs62271502	T	CEP63	3.45	2.10-5.69	0.0277	0.0200	1.11E-06
1	62587111	rs140486433	T	INADL	6.88	3.13-15.11	0.0130	0.0112	1.56E-06
4	187824750	rs423093	C	MRPS36P2	1.99	1.50-2.64	0.1259	0.1267	1.76E-06
20	49520302	rs6096175	A	ADNP	7.72	3.33-17.89	0.0111	0.0187	1.85E-06
6	91596780	rs56302597	A	MAP3K7	3.32	2.03-5.45	0.0332	0.0587	1.89E-06
7	42503417	rs61638218	A	AC027269.2	4.61	2.46-8.67	0.0203	0.0130	2.02E-06
4	98276739	rs996455	T	STPG2:RP11-681L8.1	4.12	2.30-7.41	0.0241	0.0387	2.13E-06
7	67661041	rs35946692	A	RP5-945F2.3	2.77	1.82-4.23	0.0463	0.0480	2.32E-06
6	80307871	rs113616685	C	SH3BGRL2	5.21	2.63-10.35	0.0148	0.0200	2.35E-06
4	180441783	rs76896758	T	RP11-404J23.1	2.32	1.63-3.29	0.0648	0.0745	2.63E-06
2	51957838	rs11563091	G	AC007682.1	5.31	2.64-10.65	0.0167	0.0390	2.69E-06
2	153497962	rs186630669	T	FMNL2	7.69	3.28-18.04	0.0111	0.0099	2.77E-06
11	108445649	rs11212714	A	EXPH5	4.89	2.51-9.51	0.0185	0.0215	3.02E-06
3	149283200	rs149355495	G	WWTR1	5.20	2.6-10.41	0.0167	0.0242	3.16E-06
17	75838777	rs62077154	A	FLJ45079	5.58	2.7-11.52	0.0148	0.0166	3.35E-06
11	72433925	rs61894163	A	ARAP1	6.08	2.84-13.01	0.0130	0.0196	3.36E-06
1	70496017	rs111722349	C	LRRC7:RP11-181B18.1	3.87	2.19-6.85	0.0185	0.0237	3.38E-06
2	161292848	rs116362068	T	RBMS1	5.71	2.74-11.91	0.0148	0.0097	3.45E-06
20	29989958	rs62204599	T	DEFB121	3.43	2.04-5.78	0.0297	0.0392	3.63E-06
4	18216699	rs76182690	T	LCORL	4.10	2.25-7.47	0.0223	0.0167	3.92E-06
14	98419836	rs140025898	G	C14orf64	3.35	2.00-5.59	0.0316	0.0239	3.99E-06
5	168519124	rs144408570	G	SLIT3	7.22	3.12-16.73	0.0112	0.0148	4.00E-06
10	33452029	rs111383576	C	NRP1	3.87	2.17-6.87	0.0241	0.0296	4.07E-06
6	85531998	rs192911302	C	TBX18	6.08	2.81-13.17	0.0130	0.0121	4.62E-06

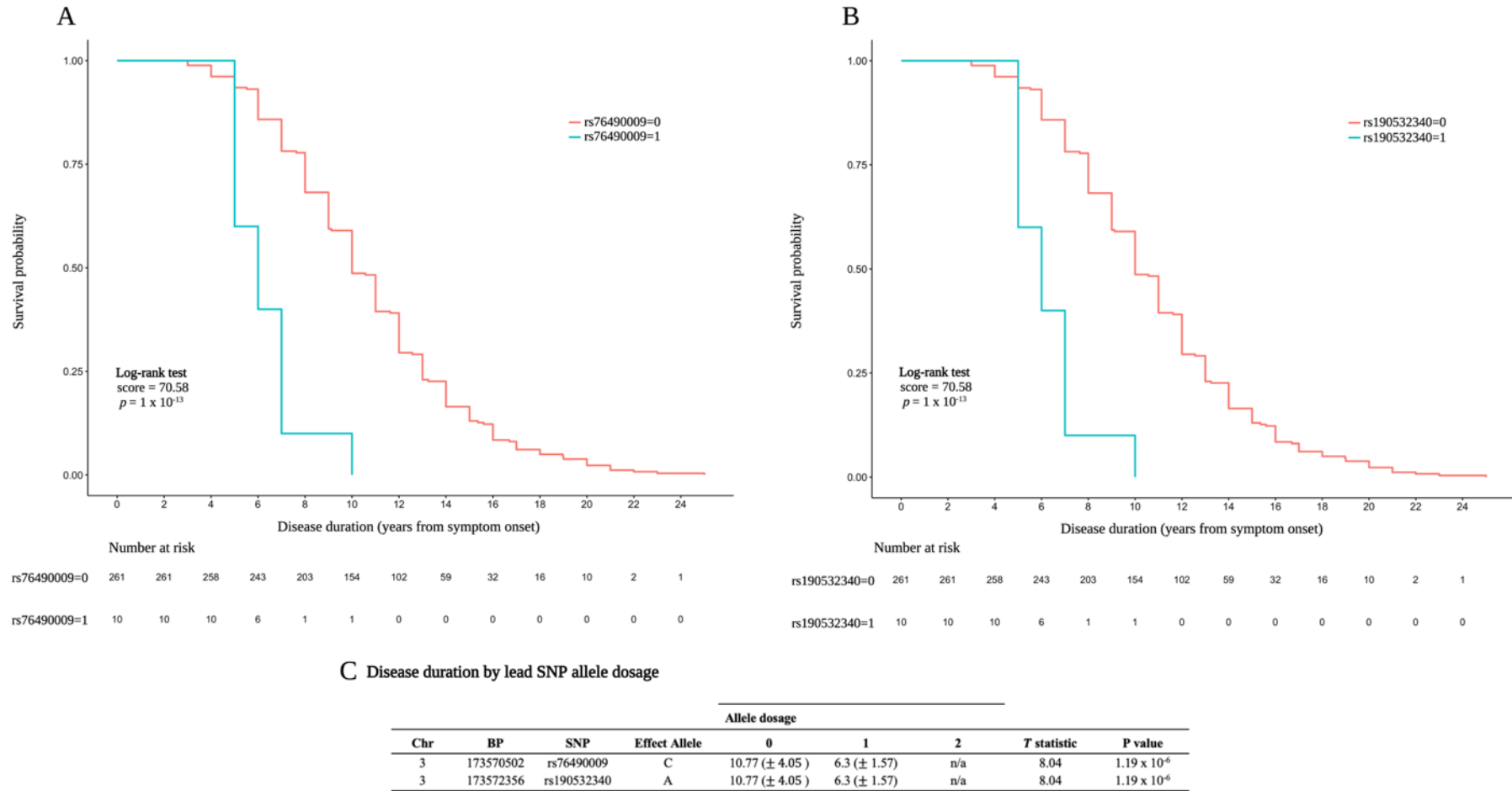
Lead SNP for each genomic locus defined using FUMA standard setting. Cox proportional hazards model adjusted for gender, age at onset and three PCs (PC2, 6 and 8) were used to study the association of candidate loci with survival (time from symptom onset to death) in PiD in the total cohort (272 cases). Abbreviations: BP = base pair, Chr = Chromosome, CI = confidence interval, NFE = Non-Finnish Europeans (GnomAD v2.1.1), OR = odds ratio.

There were no proxy coding variants in high linkage disequilibrium with either of the two lead variants on chromosome 3 (using the LD Proxy tool). Conditional analyses at the chromosome 3 locus adjusting for both rs76490009 and rs190532340 in turn, did not reveal any other independent SNPs contributing to the signal at the chromosome 3 locus (**Figure 8.5A-C**). I also performed conditional analysis for the other locus on chromosome 18 (rs55682632 an intronic variant in *KATNAL2*; HR 4.89, 95% CI 2.63-9.07,  $p = 4.86 \times 10^{-7}$ ), which again did not reveal any other independent SNPs contributing to the chromosome 18 signal (**Figure 8.5D-E**).



**Figure 8.5 - Conditional analyses adjusted for the lead SNP at the top two GWSS loci.** For chromosome 3 (A) Unconditioned Manhattan plot (B) Manhattan plot conditioned on the *NLGN1* variant rs76490009 and (C) conditioned on the *NLGN1* variant rs190532340. For chromosome 18 (D) Unconditioned Manhattan plot and (E) Manhattan plot conditioned on the *KATNAL2* variant rs55682632.

The minor allele (defined as the effect allele) for both of the genome wide significant SNPs on chromosome 3 were associated with worsening survival (**Figure 8.6**); 6.30 yrs SD  $\pm$  1.57 yrs for those carrying one effect allele versus 10.77 yrs SD  $\pm$  4.05 for those not carrying an effect allele (unpaired two-tailed t test, t statistic = 8.04, p =  $1.19 \times 10^{-6}$ ). No individuals survived longer than 10 years in the heterozygous state (CPH log rank test = 70.58 on 6 df, p =  $3 \times 10^{-13}$ ). Using the summary statistics from the PiD case-control GWAS, there was no association between either SNP and risk of PiD (rs76490009: OR 1.08, 95% CI 0.46-2.53, p = 0.86 and rs190532340: OR 1.03, 95% CI 0.44-2.28, p = 0.99).



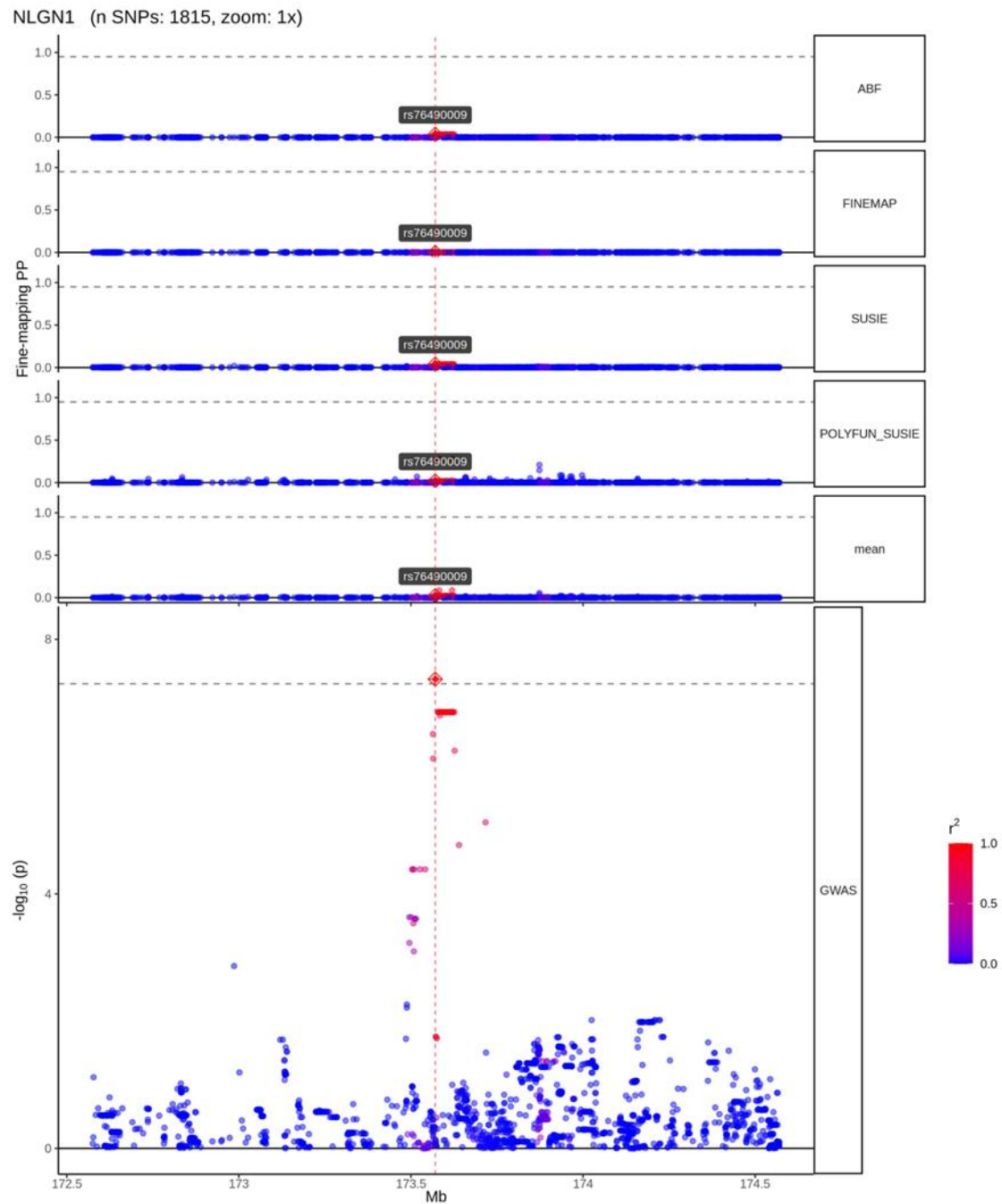
**Figure 8.6 - Survival curves for genome-wide significant SNPs.** Kaplan-Meier curves for PiD survival (disease duration, yrs from symptom onset to death) based on (A) rs76490009 and (B) rs190532340 allele number. Statistical analysis was conducted using Cox proportional hazards models in 272 PiD samples. Both SNPs met the proportional hazards assumption. (C) Table showing mean survival by allele dosage for each SNP. Unpaired two-tailed *t*-test performed to compare means between two allele dosage groups.



### 8.3.4 Fine-mapping and functional annotation

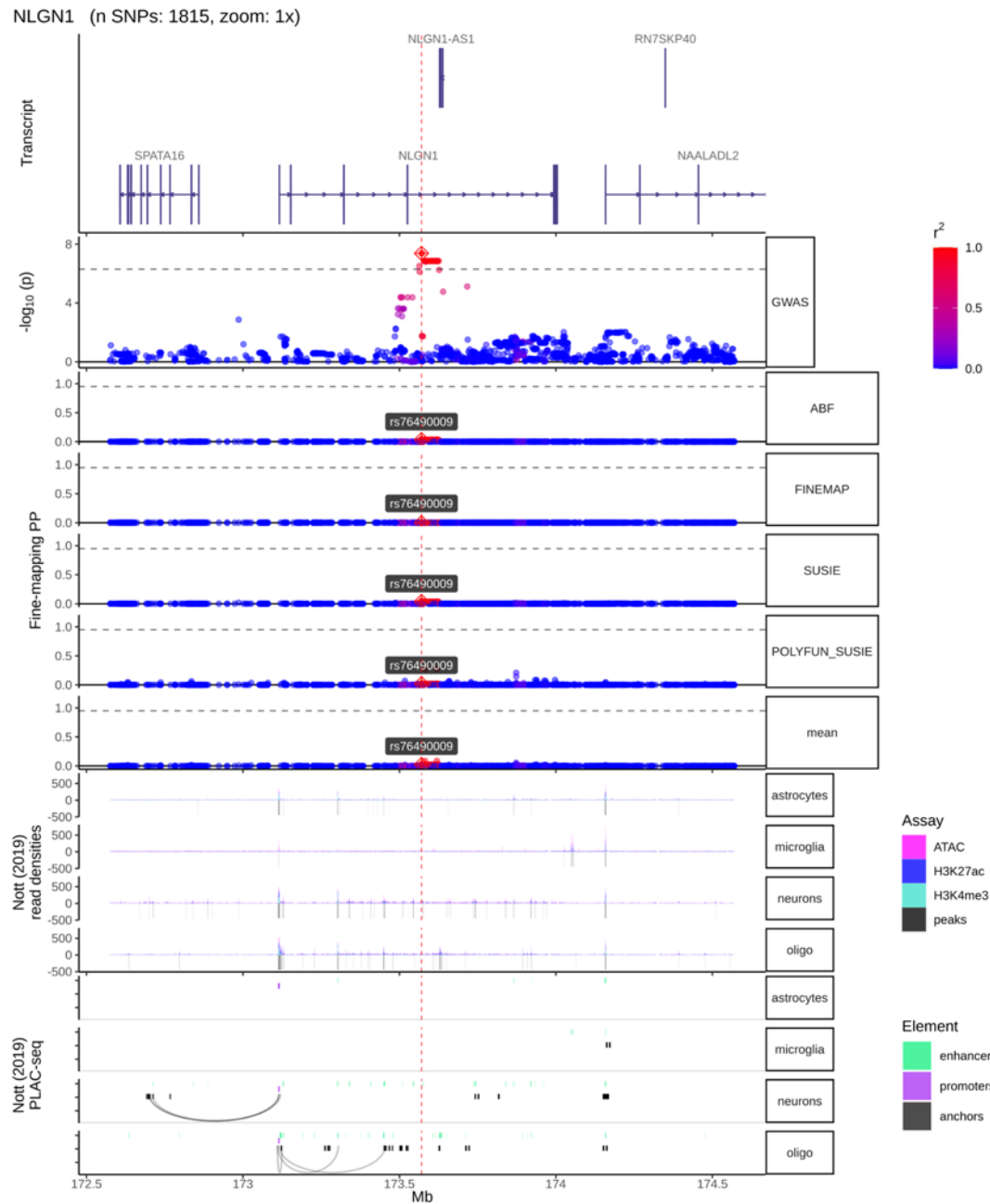
Given the conditional analysis results, statistical and functional fine-mapping was performed under the single variant assumption on the genome-wide significant locus on chromosome 3. No consensus SNPs were identified (**Figure 8.7**), though given rs76490009 had a higher CADD score (8.76) versus rs190532340 (0.39) I nominated this putative causal SNP out of the two.

The potential effects of the lead GWSS loci in controlling gene expression were then investigated using available genomic regulatory elements datasets. The chromosome 3 locus was mapped against 1) functional DNA elements from the ENCODE dataset (**Figure 8.8**) 2) cell-type specific enhancer marks (**Figure 8.9**) and 3) regulatory elements data from the FANTOM5 dataset (**Figure 8.10**) respectively). The only finding of note was that the genome wide signal is located at a transcription factor binding site within the NLGN1 gene (**Figure 8.8**), suggesting that variation at this site may affect transcription factor binding and therefore dysregulate gene expression. In support of this there is a proxy SNP in high linkage disequilibrium (rs73035447,  $R^2=0.94$ ,  $D'=1$ ) with the lead SNP (rs76490009), that in the Regulome database (RegulomeDB) has high chromatin accessibility in the brain based on DNase-seq and is predicted to be part of a transcription factor binding motif.

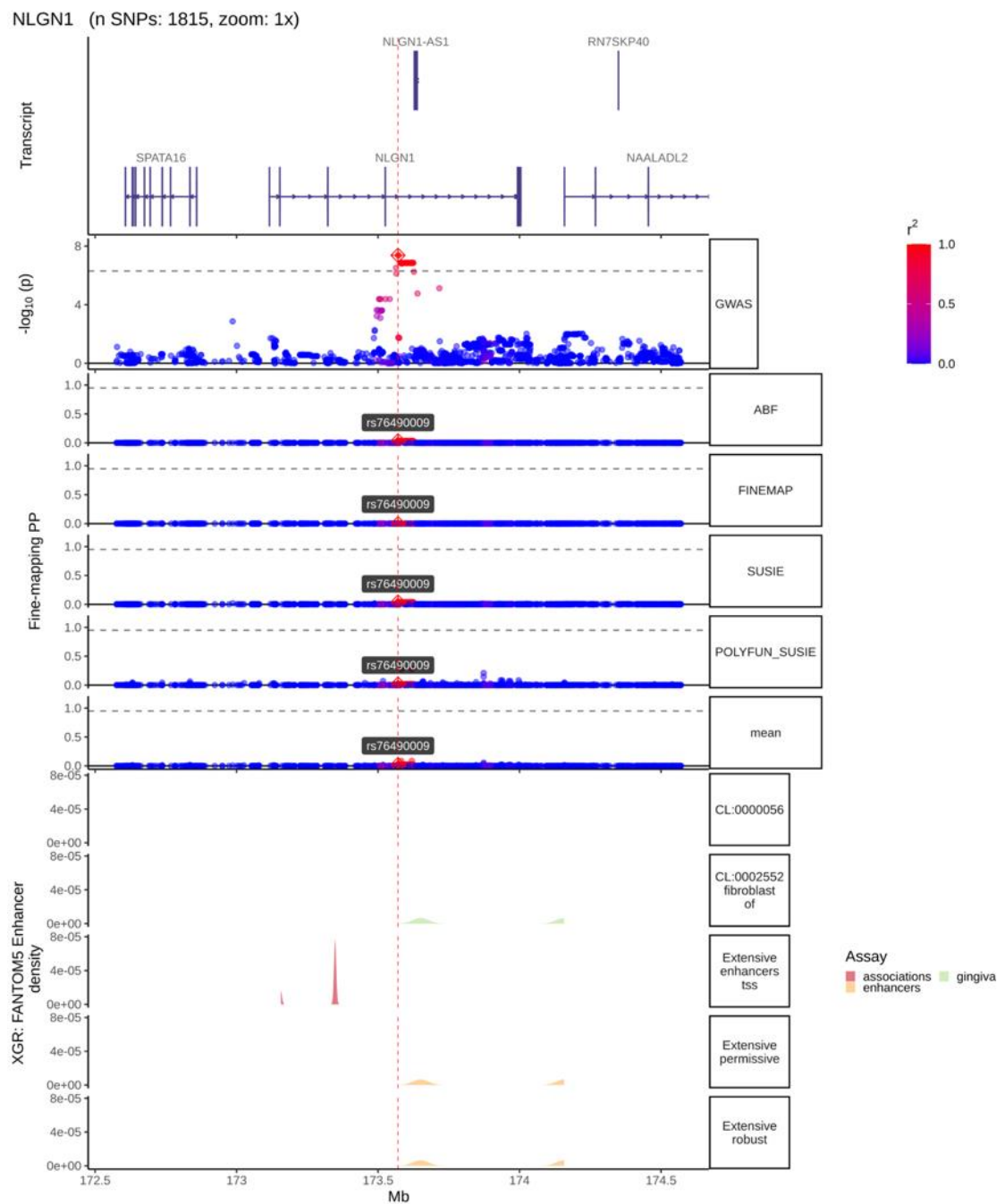


**Figure 8.7 - Fine-mapping of the *NLGN1* locus.** The first five rows show the cross tools fine-mapping output (ABF, FINEMAP, SUSIE and POLYFUN-SUSIE) and the mean values across all tools with the fine-mapping posterior probabilities (PP) plotted on the x-axes.





**Figure 8.9 - Cell-type specific regulatory element mapping at *NLGN1* locus.** Cell type specific regulatory element marks from Nott dataset (bottom two rows) with top seven rows showing GWAS loci, transcript and fine-mapping plots respectively.



**Figure 8.10 - Enhancer element mapping at NLGN1 locus.** The bottom row shows the FANTOM5 enhancer marks from the FANTOM project, with GWAS loci, transcript and fine-mapping plots respectively on top seven rows.

### 8.3.5 Colocalisation analysis

Using FUMA I mined available eQTL datasets, but this failed to show that either of the two lead SNPs were associated with altered gene expression of *NLGN1* or any nearby genes. I therefore performed co-localisation analysis to further evaluate whether there was an overlap between SNPs modulating survival and gene expression. Given the absence of additional signals at the top two genomic loci, confirmed through conditional analysis, the single variant assumption require for colocalisation analysis was met. eQTLs were obtained from the eQTLGen and MetaBrain Cortex datasets, and the *coloc* package was used to test for colocalisation with genes within  $\pm 1\text{mb}$  of the GWSS loci with a  $p$  value  $< 5 \times 10^{-7}$ . Of note, *NLGN1* was present in the MetaBrain cortex data but not in the eQTLGen data. There were no colocalisation signals for any gene at the chromosome 3 locus (GWSS signal,  $p = 4.23 \times 10^{-8}$ ) or at the chromosome 18 locus (GWSS signal,  $p = 4.86 \times 10^{-7}$ ) (Table 8.5).

**Table 8.5 - Colocalisation analysis of MetaBrain and eQTLGen datasets for genomic loci  $p < 5 \times 10^{-7}$**

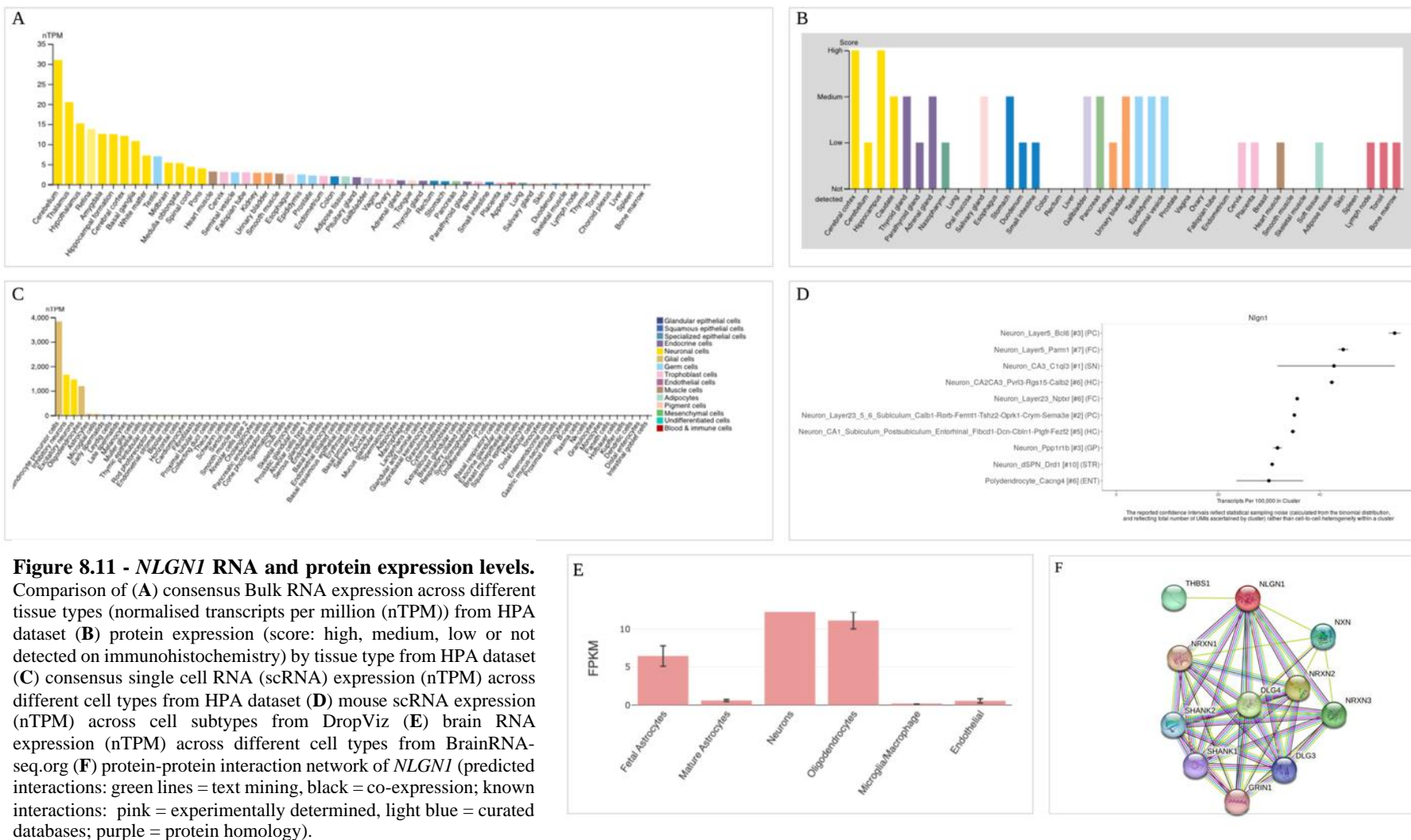
Dataset	Gene	PP.H0	PP.H1	PP.H2	PP.H3	PP.H4
eQTL_CORTEX	<i>SPATA16</i>	0.759	0.138	0.08	0.014	0.010
eQTL_CORTEX	<i>RNF165</i>	0.679	0.184	0.099	0.027	0.012
eQTL_CORTEX	<i>NAALADL2</i>	0.676	0.165	0.118	0.029	0.012
eQTL_CORTEX	<i>IER3IP1</i>	0.669	0.150	0.136	0.030	0.015
eQTL_CORTEX	<i>SKOR2</i>	0.614	0.142	0.185	0.043	0.016
eQTL_CORTEX	<i>ST8SIA5</i>	0.540	0.131	0.247	0.06	0.022
eQTL_CORTEX	<i>PSTPIP2</i>	0.512	0.141	0.252	0.069	0.026
eQTL_CORTEX	<i>ATP5F1A</i>	0.488	0.134	0.276	0.076	0.026
eQTL_CORTEX	<i>CSMD1</i>	0.448	0.237	0.194	0.103	0.019
eQTL_CORTEX	<i>NLGN1</i>	0.337	0.070	0.463	0.096	0.033
eQTL_CORTEX	<i>ELOA2</i>	0.398	0.088	0.369	0.082	0.063
eQTL_CORTEX	<i>ZBTB7C</i>	0.009	0.002	0.781	0.153	0.055
eQTL_CORTEX	<i>PIAS2</i>	0.000	0.000	0.777	0.182	0.041
eQTL_CORTEX	<i>LOXHD1</i>	0.029	0.007	0.739	0.183	0.041
eQTL_CORTEX	<i>C18orf25</i>	0.048	0.013	0.70	0.193	0.046
eQTL_CORTEX	<i>SMAD2</i>	0.013	0.003	0.727	0.180	0.076
eQTL_CORTEX	<i>HAUS1</i>	0.000	0.000	0.736	0.202	0.062
eQTL_CORTEX	<i>HDHD2</i>	0.000	0.000	0.726	0.161	0.113
eQTL_CORTEX	<i>KATNAL2</i>	0.000	0.000	0.682	0.160	0.158
ciseQTL_blood	<i>PIAS2</i>	0.000	0.000	0.679	0.162	0.159
ciseQTL_blood	<i>IER3IP1</i>	0.000	0.000	0.769	0.173	0.057
ciseQTL_blood	<i>RNF165</i>	0.000	0.000	0.738	0.212	0.050
ciseQTL_blood	<i>PSTPIP2</i>	0.000	0.000	0.710	0.196	0.094
ciseQTL_blood	<i>ATP5F1A</i>	0.000	0.000	0.746	0.212	0.041
ciseQTL_blood	<i>HAUS1</i>	0.000	0.000	0.731	0.208	0.061
ciseQTL_blood	<i>C18orf25</i>	0.000	0.000	0.747	0.207	0.046
ciseQTL_blood	<i>LOXHD1</i>	0.000	0.000	0.757	0.198	0.044
ciseQTL_blood	<i>KATNAL2</i>	0.000	0.000	0.764	0.178	0.058
ciseQTL_blood	<i>HDHD2</i>	0.000	0.000	0.764	0.174	0.062
ciseQTL_blood	<i>SMAD2</i>	0.000	0.000	0.753	0.195	0.053
ciseQTL_blood	<i>CSMD1</i>	0.000	0.000	0.639	0.328	0.033
ciseQTL_blood	<i>ZBTB7C</i>	0.000	0.000	0.722	0.161	0.117
ciseQTL_blood	<i>UTP18</i>	0.050	0.014	0.662	0.183	0.092
ciseQTL_blood	<i>HDHD2-AS</i>	0.000	0.000	0.766	0.174	0.061
ciseQTL_blood	<i>Lnc-CTIF-8</i>	0.000	0.000	0.753	0.195	0.053

eQTL\_CORTEX = MetaBrain cortex eQTLs, ciseQTL\_blood = eQTLGen blood eQTLs. PP.H = posterior probability of hypothesis 0,1,2,3, or 4.

### 8.3.6 Gene and protein expression

Given the fact that the lead SNP was located within *NLGNI*, combined with its location in a transcription factor binding site, and the lack of evidence for it modulating expression of any genes within 1Mb, made it the most likely candidate gene at the chromosome 3 locus. I therefore went on to investigate *NLGNI*'s gene and protein (Neuroigin1) expression profiles in the brain.

*NLGNI* RNA transcript expression is enriched within the central nervous system (**Figure 8.11A**); the cerebellum has the highest expression levels (normalised transcripts per million (nTPM) 31.0), with the amygdala, hippocampus and cerebral cortex all showing similar levels (12.6 nTPM, 12.5 nTPM and 12.1 nTPM respectively). Protein expression mirrors gene expression with high Neuroigin1 protein expression, as measured through antibody staining, in the cerebral cortex and the hippocampus (**Figure 8.11B**). Consensus human scRNA data from the HPA shows that gene expression is highly enriched in oligodendrocyte precursor cells (3834.34 nTPM), inhibitory and excitatory neurons (1658.7 nTPM and 1462.4 nTPM respectively) and oligodendrocytes (1191.2 nTPM) (**Figure 8.11C**). This enrichment within neurons and oligodendrocytes is supported by cell type data from the Brain RNA-seq dataset (**Figure 8.11E**). Mouse scRNA data suggests that the gene is specifically enriched within Layer 5 neurons (Neuron\_Layer5\_bcl6 and Neuron\_Layer5\_Parm1) in the posterior and frontal cortex respectively, as well as CA3 neurons in the hippocampus (Neuron\_CA2CA3\_Pvrl3-Rgs15-Calb2) and the substantia nigra (Neuron\_CA3\_C1ql3) (**Figure 8.11E**). The protein interaction network for Neuroigin1 shows it interacts with Neurexin 1, a pre-synaptic cell surface protein, as well as other post-synaptic proteins including DLG4 (PSD-95), SHANK1, SHANK2 and GRIN1 (**Figure 8.11F**).





### 8.3.7 Neuroligin 1

On balance, the evidence supports *NLGN1* as the most likely candidate gene at the chromosome 3 locus, with the possibility that the SNP at this locus modulates transcription factor binding, thus altering expression of the gene.

Neuroligin 1, the protein coded for by the *NLGN1* gene, belongs to the neuroligin superfamily; a group of trans-synaptic cell adhesion molecules anchored at the post-synaptic membrane, that act as synaptic organisers involved in the functional maturation and identity specification of the synapse<sup>519</sup>. Their main binding partners are neurexins on the pre-synaptic membrane<sup>520</sup>. The neuroligin-neurexin bridge stabilises the synapse as well as functionally coupling the post-synaptic density with the pre-synaptic complex that is responsible for neurotransmitter release<sup>521–523</sup> and endocytosis<sup>524</sup>. Mutations in the genes of these proteins have been demonstrated in a range of neurodevelopmental disorders including schizophrenia and autism<sup>519,525,526</sup>.

Neuroligin1, specifically, is usually expressed in neurons at post-synaptic excitatory synapse densities<sup>523</sup> and has been implicated in a range of neurological and neuropsychiatric disorders including schizophrenia, autism and stroke<sup>527–529</sup>. The relevance of Neuroligin1 (and its binding partner Neurexin1) in AD and other tauopathies has been emphasised in a number of studies over the last decade; a truncating mutation in the gene has been shown to cause AD pathology<sup>530</sup>, amyloid- $\beta$  protein oligomers were shown to bind Neuroligin1 in vitro and modulate synaptic integrity<sup>531,532</sup>, and there is evidence that Neuroligin1 levels can be decreased in different brain regions in Alzheimer's disease as well PiD, CBD and PSP<sup>533,534</sup>. Work in rat models suggests that neuroinflammation triggered by amyloid deposition results in the epigenetic suppression of Neuroligin1, and subsequent impairment of synaptic function and memory<sup>535</sup>.

Overall *NLGN1* constitutes a promising candidate gene for modifying survival in PiD, though from this study it is unclear whether the lead SNP's effect is to increase or decrease *NLGN1* expression. Further work will be required both to validate the finding in an independent PiD cohort, as well as functional studies in model systems (such as induced pluripotent stem cells) to better understand the effect of the lead SNP on *NLGN1* expression and how this perturbs the underlying biological pathway to impact on PiD survival.

### 8.3.8 Potential candidate genes

Finally, I searched for potential candidate genes with a  $p$  value near the genome-wide significant threshold that could be relevant to PiD survival. 30 independent genomic loci were identified where the lead SNP had a suggestive  $p$  value  $< 5 \times 10^{-6}$  (**Table 8.3**). The nearest genes to some of these variants are involved in pathways known to contribute to neurodegeneration, including complement (*CSMD1*)<sup>536</sup> and autophagy (*RPTOR*)<sup>537,538</sup>, as well as pathways involved in neuronal microtubule stabilisation (*KATNAL2*, *ADNP*)<sup>539</sup>, and dendrite development (*NRPI*)<sup>540</sup>. *CSMD1* is a complement inhibitor, that in its native state opposes the complement cascade in neural tissue, preventing synaptic pruning by microglia<sup>541</sup>. Common variation at *CSMD1* is a risk factor for both schizophrenia<sup>542</sup> and is also associated with immediate episodic memory in healthy adults<sup>543</sup>. Recently heterozygous mutations in this gene have also been shown to cause familial Parkinson's disease<sup>544</sup>.

Of particular interest with regards to PiD and tauopathy are *KATNAL2* on chromosome 18 (lead SNP rs55682632, HR 4.89, 95% CI 2.63-9.07,  $p = 4.86 \times 10^{-7}$ ) and *ADNP* on chromosome 20 (lead SNP rs6096175, HR 7.72, 95% CI 3.33-17.89,  $p = 1.85 \times 10^{-6}$ ). The protein encoded by *KATNAL2*, though poorly characterised, is part of the Katanin family which function as microtubule-severing enzymes regulating the development of neuronal protrusions through cytoskeletal rearrangements<sup>545</sup>. In normal physiological conditions microtubules are protected from the basal microtubule severing capacity of Katanin by microtubule-bound tau<sup>546,547</sup>. The gene has been consistently associated with autistic spectrum disorders (ASD), with 49 mutations so far identified in patients with ASD and other neurodevelopmental disorders<sup>548</sup>. *ADNP* (activity dependent neuroprotective protein) is essential for brain formation and function<sup>549</sup>, and its expression has been found to be dysregulated in a range of neurological and neurodegenerative disorders including schizophrenia<sup>550,551</sup>, Parkinson's disease<sup>441</sup> and Alzheimer's disease<sup>552</sup>. *ADNP* deficiency in mice causes pathological tau hyperphosphorylation concomitant with cognitive dysfunction<sup>553</sup>, and heterozygous mutations in the gene cause a neurodevelopmental disorder with a tauopathy evident at post-mortem<sup>554</sup>. Interestingly, *ADNP* is correlated with higher 3R Tau expression<sup>555</sup>, and the *ADNP*-derived peptide NAP (Davunetide) which has been shown to prevent microtubule degradation by recruiting tau to microtubules, preferentially binds 3R tau<sup>556</sup>. The relevance of these genes to PiD survival is still to

be determined, and future studies with larger sample sizes are needed to confirm or refute these associations.

## 8.4 Discussion

In this chapter I have conducted genome-wide survival study using 272 samples from the PIC and shown that variation at the *NLGN1* locus on chromosome 3 is a genetic determinant of disease duration in PiD. For those individuals carrying the minor allele at the lead SNP (rs76490009), none survived more than 10 years from symptom onset, with an average survival of 6.3 years. Mapping the loci to regulatory elements data from the FANTOM5 demonstrates that this region is located at a transcription factor binding site within *NLGN1*, suggesting that the lead SNP may modify transcription of the gene.

Candidate variants that have been previously reported in association with primary or secondary tauopathy risk did not show an association with PiD survival. In addition, there was no association for rs2242367, the variant that has been previously shown to modify PSP survival<sup>198</sup>. The *MAPT* H2 haplotype, which I have previously shown to increase the risk of PiD (**Chapter 6**), as well as the lead genomic loci in the PiD case-control GWAS, were also not associated with PiD survival suggesting that the risk conferred by these loci affects pathways upstream of those implicated in decreased survival. These findings have important implications for the development of disease modifying therapeutics for slowing PiD progression. Firstly, they suggest that therapeutics targeting aberrant pathways in the 4R tauopathies (such as PSP) may not be efficacious in individuals with PiD. Secondly, in the absence of biomarkers to identify individuals at increased risk of developing PiD, any future clinical trial in PiD will involve recruitment of those that already have a clinical diagnosis with the aim of slowing disease progression/prolonging survival. Given the genetic variants associated with survival appear to be different to disease risk, these findings suggest that multiple different pathways may need to be targeted to treat the disease.

*NLGN1* is a plausible candidate gene for modulating survival in PiD. It is highly expressed in the brain in regions relevant to PiD pathology (amygdala, hippocampus and cerebral cortex)<sup>62</sup>, with highest expression within neurons, the cell types with the highest burden of 3R tau pathology at post-mortem<sup>62</sup>. Interestingly mouse scRNA data shows that Layer V neurons are the most enriched neuronal type for *NLGN1*

expression. This finding is in keeping with the neurons in this layer showing a high density of Pick bodies in human post-mortem samples<sup>62,557,558</sup>, as well as the observation of early loss of Von Economo neurons (Layer 5 neurons) in behavioural variant FTD (the most common presentation of PiD)<sup>559</sup>. At the protein level, Neuroligin1 is decreased in the hippocampi of AD patients and importantly this decrease is greater than in those with amnesic cognitive impairment (aMCI)<sup>534</sup> and correlates with degree of cognitive impairment. The decrease in Neuroligin1 is greater in magnitude than PSD-95 (another post-synaptic scaffolding protein) suggesting that this finding is not simply a result of greater neuronal loss in the AD and aMCI cases. Decreased Neuroligin1 levels compared to controls has also been shown in the primary tauopathies, with the greatest decrease occurring in PiD specifically<sup>533</sup>. The discovery of a truncating heterozygous *NLGN1* mutation in a familial case of AD shows that *NLGN1* is sufficient to cause neurodegeneration, giving credence to the assertion that it may play a role in progression of PiD<sup>530</sup>.

Neuroligin1, through binding to pre-synaptic Neurexin1, and its interaction with PSD-95 in the post-synaptic density, stabilises the synapse and is essential for normal synaptic transmission and plasticity<sup>519,560</sup>. The Neuroligin1:PSD-95 post-synaptic complex modulates pre-synaptic neurotransmitter (NT) release via Neurexin1<sup>561</sup>, as well as regulating pre-synaptic NT endocytosis<sup>524</sup>. In mice, synaptic activity has been shown to reduce accumulation of pathological tau through stimulation of the autophagic-lysosomal degradation pathway<sup>562</sup>. Conversely, in the same study when they inhibited synaptic activity there was a worsening of tau pathology, with accumulation of Tau oligomers in swollen lysosomes and induction of further deterioration of synapses. Given that knock out of *NLGN1* in mice leads to significant impairment in synaptic transmission<sup>563,564</sup>, a possible mechanism explaining the association with PiD survival is that dysregulation of *NLGN1* will impair normal synaptic activity and so lead to worsening tau pathology in individuals with PiD. PSD-95, which dynamically regulates Neuroligin1 function and trafficking to the cell membrane<sup>565</sup>, can itself form a complex with tau<sup>566</sup>, and in a P301L mouse model pTau has been shown to disrupt the binding of PSD-95 to nitrous oxide synthase (NOS) in the post-synaptic density leading to impairment of NOS function<sup>567</sup>. It is possible that a similar mechanism with regards to Neuroligin1 may occur in the presence of hyperphosphorylated tau which is known to mis-localise to the presynaptic dendritic

spines<sup>568</sup>. This would lead to a vicious cycle where hyperphosphorylated tau impairs Neuroligin1 function, leading to decreased synaptic activity which in turn would lead to impaired tau degradation, accumulation and then further dysregulation of NLGN1 via PSD-95.

One other interesting observation with regards to Neuroligin1 is that it contains an extracellular acetylcholinesterase-like domain that forms the interaction site with Neurexin1 at the pre-synaptic membrane<sup>569</sup>. Excess expression of AChE *in-vitro* decreases the association between Neurexin1 and Neuroligin1<sup>570,571</sup> and inhibits Neuroligin1 mediated synaptic integrity. Given the long-standing clinical observation that AChE inhibitors are ineffective in the FTD syndromes (in particular bvFTD the main presentation of PiD) and can lead to worsening of disinhibition and compulsive behaviour, opens up the intriguing possibility that this effect may be mediated by AChE inhibitors competitive inhibition of (already dysfunctional) Neuroligin binding to Neurexin1 via its AchE-like domain.

A strength of this study is the fact that disease duration, which captures the entire clinical disease course from symptom onset to death, was used as the outcome measure for analysis. This has advantages over the use of longitudinal rate of change in clinical scores<sup>325,572</sup>, which only capture specific timepoints in the disease course, and can be subject to inter- and intra-rater variability<sup>573,574</sup>. In addition, given all individuals included had a pathological diagnosis of PiD using rigorous and standardised diagnostic criteria, diagnostic accuracy was not an issue in this study.

On the other hand, I acknowledge that there are potential limitations with this study. Firstly, as with the case-control GWAS (**Chapter 7**), there was no replication cohort available with which to validate the findings from this study. Through the PIC I collected all available pathology confirmed PiD cases worldwide (321 at the time of performing study), but due to some cases having missing clinical data only 294 were available for inclusion, and due to exclusion of 22 cases during the QC process only 272 were available for analysis. This sample size is small by GWAS standards and makes it likely the study was underpowered. It also made it impossible to perform a two-stage study to replicate the findings. Despite this I hope that the discovery of the genome-wide significant *NLGN1* locus, with a plausible testable hypothesis for the mechanism by which Neuroligin1 modulates survival, will stimulate future work on biomarker and functional studies to confirm or refute the involvement of synaptic

pathology in PiD. In addition, the suggestive association of the *KATNAL2*, *ADNP* and *CSMD1* genes at nominally significant loci, all of which are involved in biologically plausible pathways, provides additional hypotheses that should be tested in future studies. Since performing this study, a further 45 PiD cases have been collected by the PIC network, though given that the estimated number of PiD cases that will be collected a year is c.10-15, it will be difficult to perform a replication study within the next few years. A second limitation was that apart from direct genotyping of the six *MAPT* variants, there was no direct genotyping of lead SNPs at the lead loci. I applied stringent QC thresholds, including an  $R^2$  of  $\geq 0.7$  and a genotype posterior probability of  $\geq 0.9$  for imputed genotypes, but acknowledge that small errors in the frequency of imputed minor alleles have a significant impact on variant effect size and therefore the  $p$  value. In future work it will be necessary to directly sequence the *NLGN1* locus to confirm the genotypes found by micro array genotyping. Perhaps the most effective way to do this will through large-scale analysis of whole genome sequencing of the PIC samples, as this will also allow us to identify any deleterious rare coding variants that may associated with PiD survival. Finally, due to the MCJ and UCL cohorts being genotyped on two different chips (the GSA and NBA respectively) I had to impute the two datasets separately and merge based on common SNPs. A disadvantage of this approach is that there was a loss of total number of SNPs in the final analysis, due to the GSA having less dense genome coverage than the NBA, which may have led to missing potential signals in less well covered parts of the genome. Again, by carrying out future whole genome sequencing of the samples this limitation will be mitigated.

## 8.5 Conclusions and future work

This is the first study to use disease duration to look for genetic risk factors that modify survival in PiD. Using a cox proportional hazards model, I identified a genome-wide significant variant located in the *NLGN1* gene on chromosome 3, which when present in the heterozygous state reduced survival by nearly 4 years. In addition, I identified several other possible loci that may be involved in disease progression; in particular *CSMD1*, mutations in which can cause Parkinson's disease; *KATNAL2* which is involved in microtubule cleavage, and *ADNP*, mutations in which cause a neurodevelopmental disorder characterised by the accumulation of 3R tau at post-mortem. Loss of synaptic integrity with resulting dysregulation of synaptic

transmission leading to increased pathological tau accumulation is a plausible mechanism through which *NLGN1* dysfunction could impact on survival in PiD. It will be necessary to replicate the findings in an independent cohort of PiD cases, as well as sequence the cases from this study to confirm the accuracy of imputed genotypes as well as identify if any rare deleterious mutations or indels could be driving the signal at this locus. Although future studies with larger sample sizes would be preferable to allow identification of associations with suggestive variants of smaller effect sizes and allele frequencies, in the absence of in vivo biomarkers to diagnose Pick's pathology in life this will prove difficult due to the rarity of samples. Future studies focused on identification of biomarkers for PiD will be essential to accelerate research progress in this disease, and I hope this study will help stimulate hypotheses to this end. Functional studies in *in-vitro* model systems, as well as animal models, to investigate the interplay of *NLGN1*, synaptic dysfunction and neuronal tau accumulation in PiD will be necessary to further our knowledge of disease pathogenesis. And finally, investigation of the *NLGN1* loci as potential genetic determinant of disease progression and survival in related tauopathies (AD, PSP and CBD) will be important, as this may identify a common target across these disorders that could be amenable to therapeutic intervention.

# Chapter 9: General conclusions and future directions

## 9.1 Summary

The primary tauopathies (FTLD-tau) are a heterogeneous group of disorders both pathologically and in terms of clinical presentation. The most common of them, PSP, has a prevalence of 4-7 per 100,000 of the population<sup>14,75,76,417</sup> with healthcare costs previously estimated at up to £43,000 per patient per year<sup>575</sup>. CBS has a prevalence of c.2 per 100,000<sup>417</sup>, while the prevalence of PiD is currently unknown. Aside from the terrible toll these diseases take both physically and emotionally on both patients and their families, there is also a significant economic impact worsened by the fact that they have a younger age of onset than other neurodegenerative diseases, such as AD. One key study run by the Association for Frontotemporal Degeneration revealed that the economic costs associated with FTD were nearly two times higher than for AD, most likely due to loss of productivity for both patient and carers from being pulled out of the workforce at the height of their careers<sup>576</sup>. There are no curative treatments for the primary tauopathies at present, and there is an urgent need to develop new therapies, both from a symptomatic and a disease-modifying perspective. The development of therapies is hampered by the wide variability in clinical presentations for each underlying pathology, with presentations often overlapping, as well as the frequent occurrence of atypical presentations that may mimic other non-FLTD pathologies (most commonly AD). Although progress has been made in understanding the genetic contribution to disease risk in the more common 4R tauopathies (PSP and CBS), very little is known about the genetics of the 3R tauopathy PiD.

In this thesis I have focused firstly on trying to better delineate disease progression in vivo in the 4R tauopathies (PSP and CBS), and secondly to understand the genetic drivers of disease risk and progression in the PiD. Given the rarity of these diseases, I had to build two separate cohorts (the 4R tau imaging cohort and the PIC), through collaboration with a large number of international groups, to facilitate the analyses laid out in the aims of this thesis. I hope that these cohorts will provide a framework for ongoing international collaboration, and allow researchers to build on the findings



from this thesis, to develop new biomarkers and identify therapeutic targets to treat these devastating diseases.

## **9.2 Disease progression modelling in the 4R tauopathies**

### **9.2.1 PSP-RS event based modelling**

The first aim of this thesis was to develop an image-based model of brain volume biomarker changes in PSP-RS to enable to in vivo staging of disease. In **Chapter 3** I used a probabilistic EBM applied to cross-sectional structural MRI scans to determine the sequence of brain atrophy in clinically diagnosed PSP-RS. I used a combination of 12-month follow-up scans, and a validated clinical rating scale (the PSP rating scale) to demonstrate the longitudinal consistency and the utility of the EBM's staging system. The order of regional atrophy broadly mirrors the sequential spread of tau pathology proposed by Kovacs et al.<sup>51</sup> in their post-mortem PSP staging study, and confirms that the brainstem and cortical structures are affected early in the disease, while demonstrating how atrophy (and presumably pathology) spreads via the thalamus to the frontal cortex before progressing in a caudal to rostral direction with the occipital lobe affected last. My work presents an advance for clinical PSP research, as it opens up the possibility to stage disease and track disease progression using easily available, objective structural MRI-based measures in living patients<sup>359</sup>.

This work does however pose further questions for future research. Although the EBM provides information on the sequence of atrophy in PSP-RS and allows probabilistic staging of individuals at baseline, it remains to be shown whether its power to detect change in brain volumes over time is superior to a classical volumetric change analysis. A structural MRI study by Hoglinger et al.<sup>265</sup> employing atlas-based volumetry in PSP patients demonstrated that combined analysis of annualised percentage volume change in three regions with the highest effect sizes (third ventricle, frontal lobe and midbrain), allowed the reduction in sample size required to detect a 50% treatment effect over one year by 65% compared to the PSP rating scale score. It will be interesting to compare these findings with an equivalent analysis using the EBM. It may be, however, that the real utility of the EBM in the clinical trial setting comes from using the predicted stage as an objective biomarker to stratify individuals at entry

to minimise disease stage heterogeneity. In support of this application a recent study by Oxtoby et al.<sup>288</sup> used an ADNI-trained EBM to stage baseline data from the ADCS-MCI clinical trial, and in a post-hoc analysis identified a data driven subgroup with more severe cognitive impairment who showed a clearer treatment response than observed in the full cohort. Following a similar approach using the trained EBM from this thesis to stage data from recent PSP clinical trials could prove informative.

An important difference between my work and the PSP pathological staging system proposed by Kovacs et al.<sup>51</sup>, is that theirs is based on post-mortem 4R tau pathology as opposed to in vivo atrophy. This may, at least in part, explain the differences in the order of events in some regions between the two models. For example, the GP is the eighth of 19 stages in the EBM, but the first of six stages in the tau pathology staging, while the medulla which is first in the EBM and second in the pathology staging. Differential aggregation of tau in neurons, astrocytes and oligodendroglia may contribute to this discrepancy, as might atrophy in brain areas distal to, but not colocalised with, tau pathology. Although not yet validated for clinical use, there are promising second generation tau PET tracers that show potential in the 4R tauopathies<sup>255</sup>. Once a sensitive and specific tau PET tracer has been identified, it will be important to perform a similar EBM analysis to ascertain whether tau pathology demonstrates an identical or differential sequential pattern of brain region involvement in PSP-RS. Similarities or differences between these two imaging modalities will be highly informative as to how volume loss, as assessed on structural MRI, and progressive spreading of tau pathology are related to each other.

Another consideration is that the EBM only provides an ordering of biomarker events, and contains no information on the time between events. Understanding the time to transition between stages predicted by the EBM will also be important, and progress in this area has been made with the development of the temporal EBM (TEBM)<sup>384</sup>. The TEBM is able to learn both an individual's stage, as well as progression risk, with the potential to stratify patients at clinical trial entry based on disease stage and predicted rate of progression. Applying this model to PSP is an important area for future studies.

## 9.2.2 Subtype and stage inference in the 4R tauopathies

The second aim of this thesis was to investigate the spatiotemporal heterogeneity of atrophy in the 4R tauopathy clinical syndromes, and identify population subgroups with distinct patterns of disease progression. To do this I applied a novel unsupervised machine learning algorithm, SuStaIn<sup>290</sup>, to cross-sectional structural MRI data firstly from clinically diagnosed PSP patients (including PSP-RS and vPSP syndromes) (**Chapter 4**), and secondly to clinically diagnosed CBS patients (**Chapter 5**).

### 9.2.2.1 PSP

In 426 clinically diagnosed PSP cases (of which 69 had vPSP syndromes), when accounting for both disease stage and subtype heterogeneity with SuStaIn, the data supported two imaging subtypes: a *Subcortical* subtype and a *Cortical* subtype. There was a strong association between clinical diagnosis and SuStaIn subtype with 82% of PSP-Subcortical cases and 81% of PSP-Richardson cases assigned to the *Subcortical* subtype and 82% of PSP-Cortical cases assigned to the *Cortical* subtype. Increasing stage was associated with worsening clinical scores, while the *Subcortical* subtype was associated with worse clinical severity scores compared to the *Cortical* subtype (PSP rating scale and Unified Parkinson's Disease Rating Scale). Validation experiments showed that subtype assignment was longitudinally stable (95% of scans were assigned to the same subtype at follow-up) and individual staging was longitudinally consistent with 90% remaining at the same stage or progressing to a later stage at follow-up. The finding that PSP-SC (PSP-P and PSP-PGF) cases are assigned to the same subtype as PSP-RS, but on average at an earlier stage, and progressing more slowly through stages, is in keeping with previous studies of these variant syndromes<sup>77,78,104,105,107,399</sup>. The PSP-Cortical variants (PSP-F, PSP-SL and PSP-CBS) appeared to be most commonly assigned to the *Cortical* subtype with more generalised atrophy in the early stages, with early involvement of the insula and the frontal lobes.

An interesting question arises as to why c.10% of PSP-RS and PSP-SC cases were assigned to the *Cortical* subtype and a similar percentage of PSP-C to the *Subcortical* subtype. For the PSP cases from 4RTNI1, DAV and pre-2017 UCL PSP cases I labelled them PSP-RS, given they had all been diagnosed as probable PSP (vertical supranuclear gaze palsy plus postural instability and falls within the first year of symptom onset) according to the NINDS-SPSP criteria. These criteria have been

shown to have a low sensitivity for variant PSP syndromes<sup>98</sup> and so it seemed reasonable to do this. However, it is possible that some of these cases actually had a variant syndrome rather than PSP-RS; in particular a PSP-C variant given that the NINDS-SPSP criteria do not include a cortical functional domain component. Another possibility is that the classification of these minority cases to the alternative subtype, could represent the uncertainty in clinical diagnosis with cases “forced” into a procrustean diagnostic category. If this is the case the SuStaIn imaging subtype may represent the “true” underlying disease subtype. Alternatively, if one assumes that the clinical diagnostic category is correct, the assignment of these cases may represent the underlying atrophy heterogeneity for those with the same clinical label.

There are a number of areas that may be worth pursuing in future studies. The first is to increase the number of variant PSP syndromes included in the model (ideally pathology confirmed cases), to see whether there are more imaging subtypes within the PSP clinical spectrum. In particular, it would be interesting to see whether there are additional subtypes within the *Cortical* subtype, that this study did not have the power to resolve. Work by Whitwell et al.<sup>249</sup> shows that the PSP-SL variant in particular has more severe atrophy in the supplementary motor and precentral regions, with less atrophy in brainstem structures than the PSP-F and PSP-CBS variant. Given there was no correction for disease stage (only age at scan) in the analyses, these differential atrophy patterns may simply be due to disease stage heterogeneity between groups. Running SuStaIn, better powered for these variants, may allow us to unravel these discrepancies. A second area worth further investigation is the finding from this study of the apparent slower progression through SuStaIn stages of PSP-SC variants assigned to the *Subcortical* subtype. Similar to the EBM, SuStaIn only learns the relative ordering of events/stages, and does not provide an explicit timescale of transition times between events. I inferred this slower progression through staging follow-up scans with the fitted model, and calculating the rate of progression (change in stage per year) for these cases in a post-hoc analysis. The TEBM, discussed in the previous section, is currently being integrated into the SuStaIn algorithm to allow identification of both subtype and progression rate, and would provide the ideal tool to try and show this slower rate of progression directly<sup>384</sup>. It may also help to identify whether a proportion of PSP cases fit into the recently described long duration/protracted course PSP category<sup>109,577</sup>, where up to 24% autopsy confirmed

PSP cases have a disease duration of greater than 10 years. Interestingly although PSP-P was the most common clinical diagnosis in these cases<sup>577</sup>, it only accounted for one third of cases, suggesting that the clinical diagnostic criteria do not adequately predict disease duration in these cases.

#### 9.2.2.2 CBS

The aim of this study (**Chapter 5**) was to uncover imaging subtypes of CBS based solely on a data-driven assessment of atrophy patterns, to test the hypothesis that modelling disease subtype and stage jointly would provide information on the underlying pathology.

Applying SuStaIn to 135 CBS cases from the 4R tau imaging cohort, the data best supported a two subtype model; a *Subcortical* subtype with an atrophy progression pattern characterised by early brainstem and subcortical involvement, and a *Fronto-parieto-occipital* subtype with early involvement of the parietal, posterior frontal and occipital regions. 81% of CBS-AD cases were assigned to the *Fronto-parieto-occipital* subtype, in keeping with previous autopsy work showing that CBS-AD is characterised by atrophy in these areas. Conversely 83% of CBS-PSP and 75% of CBS-CBD cases were assigned to the *Subcortical* subtype, again in keeping with the previous studies<sup>137,235</sup>. There was a suggestion that there could be a third subtype that is able to differentiate CBS-CBD from CBS-PSP; splitting the *Fronto-parieto-occipital* subtype into a *Fronto-parietal* (“CBD like”) and a *Parieto-occipital* (“AD-like”). However, using the CVIC criteria for model selection, I was unable to justify selecting this three-subtype model, over the more parsimonious two-subtype model. It will be important to build a larger pathology confirmed CBS cohort to confirm the existence of this third subtype with SuStaIn, and if proven to test the fitted model on unseen data from an independent pathology confirmed CBS cohort. A model that is able to accurately predict underlying pathology in CBS would have direct clinical utility for disease management and prognosis, as well as for selecting patients for entry into clinical trials for pathology specific disease modifying therapeutics.

Although Ling et al.<sup>119</sup> have proposed characteristic features of pathological progression in CBD, there is currently no consensus pathological staging system. It would be worth exploring whether we can use SuStaIn to better understand the progression and spatial heterogeneity of CBD pathology directly. SuStaIn has recently

been adapted to accommodate ordinal data<sup>578</sup>, and has been used to assess the progression and heterogeneity of TDP-43 accumulation in a range of TDP43 proteinopathies<sup>579</sup>. I would like to perform a similar analysis using CBD pathology data, with the aim of building a data-driven pathological staging system. Should a single sequence of pathology progression exist in CBD this could be compared to the third imaging subtype identified in **Chapter 5**.

As with PSP, the development of 4R specific tau PET tracers will be important to enable direct visualisation of pathology in life. The recent identification of disease specific tau filament folds<sup>41</sup> will hopefully facilitate this development. If these tracers can be developed, it will be important to compare the subtypes and sequences of 4R tau pathology identified by SuStaIn with the atrophy patterns identified in this thesis. Another approach, should these tau PET tracers become available, would be to train a machine learning classifier (such as decision tree classifiers or support vector machines) on 4R tau PET tracer positive cases to try and distinguish between PSP and CBS pathology in vivo.

## **9.3 Genetics analyses in the 3R tauopathy PiD**

### **9.3.1 The association of MAPT haplotypes with risk of PiD**

The first aim in this part of the PhD was to test the hypothesis that the *MAPT* haplotype structure was associated with the risk of developing PiD (**Chapter 6**). It has long been known that the *MAPT* *H1* is associated with increased risk of the 4R tauopathies PSP (H1 OR 5.5, H1c OR 2.1) and CBD (H1 OR 3.4, H1c 1.5)<sup>169,193</sup>, but equivalent studies looking to replicate this finding in PiD have been underpowered due to rarity of autopsy confirmed cases<sup>185,186</sup>. Through establishing the PIC, I was able to increase the available sample size by a factor of ten compared to these previous studies (338 cases versus 34 cases). The finding that the H2 haplotype is associated with increased risk of PiD is important as it contrasts with the strong protective effect that this haplotype has in PSP and CBD, and may inform future *MAPT* isoform-specific therapeutic strategies for the primary tauopathies.

It is currently unclear how the *MAPT* H1/H2 haplotypes differentially drive disease risk in the 4R and 3R tauopathies, and future work should focus on trying to understand how each haplotype is involved in disease specific pathogenesis. It will be important to compare tau isoform expression between the different tauopathies using long read sequencing, to test the hypotheses that haplotype status modulates the 3R-4R tau isoform ratio (H1 decreased and H2 increases), as well as overall levels of *MAPT* gene expression (H1c increases). There is conflicting evidence regarding the effect of the H1c subhaplotype on overall expression levels; while some studies suggest its pathogenic role in PSP may be related to an increase in overall *MAPT* expression<sup>580,581</sup>, other work suggests that this association is likely the result of a technical artefact<sup>582</sup>. The increased inclusion of exon 3 is consistently identified on the H2 background<sup>582,583</sup>, highlighting the need to also understand the role of N-terminal splicing better in the tauopathies, and how this relates to background haplotype status and the resulting disease. Another area that merits further investigation is the effect that the H2 haplotype has on the expression of other genes located within the haplotype inversion. The recent evidence that the H1 association signal in PD may actually be driven by *KANSL1*<sup>183,184</sup> and/or *LRRC37A2*<sup>175</sup>, opens up the possibility that haplotype specific effects at this locus in the 3R and 4R tauopathies may at least partially driven by genes other than *MAPT*.

### 9.3.2 A genome-wide association study in PiD

In **Chapter 7** I addressed the second aim of this part of the PhD; namely to identify the genetic determinants of disease risk in PiD using a genome wide approach. The association of the *MAPT* H2 haplotype with risk of disease was confirmed using this approach, and importantly after accounting for population stratification, the OR was slightly higher than using the targeted genotyping approach in **Chapter 6** (OR 1.52  $p = 0.001$  vs 1.35  $p = 0.003$  respectively). Although no genomic loci reached genome wide significance, likely due to the study being underpowered (sample size of 294 PiD cases), I identified five suggestive genomic loci with a with nominal association with risk of disease ( $p < 5 \times 10^{-6}$ ). The lead variant was located on chromosome 4 in the *KCTD8* gene (rs112161979,  $p = 6.37 \times 10^{-8}$ ) that encodes an auxiliary subunit of GABA<sub>b</sub> receptors, implicating dysregulated GABAergic signalling in the pathogenesis of PiD. The second genomic locus was located on chromosome 11 within *TRIM22* (rs66481907,  $p = 1.83 \times 10^{-6}$ ), which encodes an E3 ubiquitin ligase involved in a wide range of cellular process including misfolded protein clearance via the UPS system, antiviral activity and regulation of the NF- $\kappa$ B/NLRP3 inflammasome pathway. The TRIM family of proteins are of particular interest given the increasing recognition of their involvement in a wide range of neurodegenerative disease; *TRIM11* is associated with the more aggressive Richardson syndrome phenotype in PSP<sup>195</sup>, while TRIM19/PML has been shown to promote clearance of misfolded ataxin-7 in SCA7<sup>476</sup>. Although the lead variant (rs66481907) is a sQTL for *TRIM22* in nerve tibial tissue, I was unable to find an equivalent effect of the variant on splicing in brain tissue. It is possible that due to the smaller sample size for cortical sQTLs in GTEx, combined with the sub-genome wide signal in the GWAS, that I was simply underpowered to detect this effect.

It is important to note that the lead SNPs at these two genomic loci are intronic and are not associated with coding changes in expressed proteins, making it difficult to know which gene at these loci is driving the association signals. For the lead variant on Chromosome 4 (rs112161979) there was no colocalisation with brain eQTLs for *KCTD8* or *YIPF7*, *GNPDA2* or *GUF1* 500kb downstream. At the *TRIM22* locus the lead variant has a CADD score of 10.02 and is located in alternatively spliced *TRIM22* transcripts with retained introns (non-functional) and others targeted for nonsense



mediated decay suggesting that it is likely to mediate its effects through dysregulation of *TRIM22* expression.

### 9.3.3 A genome-wide survival study in PiD

In **Chapter 8** I addressed the final aim of the thesis, by performing a genome wide survival study using a CPH model to identify genetic variants associated with survival (defined as symptom onset to death). I discovered a genome wide association signal at the *NLGN1* locus on chromosome 3, which encodes a synaptic scaffolding protein located at the pre-synaptic membrane in neurons. Heterozygous carriers of the minor allele had a more aggressive disease course, with none surviving longer than 10 years (the median survival time in the cohort). As with the variants identified in the case-control GWAS, the lead variant here (rs76490009) is an intronic SNP. I was unable to identify any brain eQTL colocalisation signals, though it is possible given its location in a TF binding site that variation at this locus alters gene expression through perturbing TF binding. Further functional studies in cell and animal models will be needed to investigate this hypothesis, as well as potential mechanisms for *NLGN1* mediated pathogenesis in neuronal dendrites. The importance of PSD-95 in Neuroligin1 function and trafficking to the cell membrane<sup>565</sup> is well established, and given it has been shown in mouse models that PSD-95 function can be perturbed by hyperphosphorylated tau<sup>566,567</sup>, this provides a potentially interesting pathway to investigate in future studies.

There were several other candidate loci with nominal associations with survival that may also warrant further investigation. In particular *KATNA2L* which encodes a protein involved in microtubule cleavage<sup>545</sup>, and *ADNP* a neuroprotective protein, mutations in which cause a neurodevelopmental disorder with predominantly 3R tauopathy at post-mortem<sup>554</sup>.

Importantly there was no overlap between variants associated with risk of PiD and survival, suggesting that different biological pathways are involved in each of these processes. In addition, the *LRRK2* locus which has been shown to be a genetic determinant of PSP survival<sup>198</sup> is not associated with survival in PiD. This, in combination with the finding that none of the variants known to be associated with other tauopathies are associated with PiD, suggests that the genetic risk for PiD (both disease and survival risk) is unique among the tauopathies. This has important

implications for the development of disease modifying therapies and emphasises the need to develop sensitive and specific biomarkers that are able to identify PiD in life to facilitate pathology specific clinical trials.

### **9.3.4 General limitations and considerations**

The difficulty in identifying the causal variant and so the causal gene is a challenge common to all GWAS studies<sup>584</sup>. It will be vital to perform functional studies in cell and animal models to understand the functional consequences of genetic variation at the loci identified in this thesis and how these influence PiD risk and progression. An additional general limitation is that the genetic analyses in this thesis were restricted to white Caucasians. This means these results cannot be extrapolated to other populations, given there are likely to be different LD structures, allele frequencies, causal variants and effect sizes in other populations<sup>441</sup>. For genetic studies to have a positive impact on clinical practice and to reduce healthcare inequalities, participation needs to be widened to include a more diverse set of populations. In other neurological disorders this process is already well underway and include the Global Parkinson's Genetics Program (GP2)<sup>311</sup>, and the ADAMS study in MS (<https://app.mantal.co.uk/adams>). Genetic studies of disease risk and progression in a more diverse group of populations are clearly also needed in PiD, though again in the absence of disease specific biomarkers, the rarity of autopsy samples will make this challenging in the foreseeable future.

Ideally replication studies will be required to validate the findings from both GWASs. However due to the rarity of autopsy confirmed PiD samples, and the fact that only 10-15 new cases per year are estimated to be collected by the sites included within the PIC, this will prove difficult to achieve in a reasonable time frame. Only through identification of a PiD specific biomarker, such as a 3R tau PET tracer or RT-QuIC for 3R-tau, will we be able to significantly increase the recruitment of new PiD cases to enable a well powered replication study. Another consideration is that GWASs can only detect common variation, and it is possible that there are also rare variants with larger effect sizes, as well as structural variants, that could contribute to disease risk and progression. With that in mind, in my view the next steps should include targeted sequencing of the lead genomic loci for all samples to confirm the accuracy of the

genotyping, and also check whether there are any deleterious rare coding variants that are being tagged by the GWAS signals.

A final consideration, given the difficulties in finding additional PiD samples, is to look at other neurodegenerative disorders in which 3R tau predominates to see whether common pathogenic mechanisms can be identified that may inform on PiD pathogenesis. It is known that there are autosomal dominant *MAPT* mutations that can cause a 3R predominant pathology<sup>54</sup>; studies on the effect of *MAPT* haplotype background on the 3R/4R tau ratio in these mutations will be important. The presence of 3R predominant tau pathology in myotonic dystrophy type 1 and 2<sup>207–209</sup>, and tuberous sclerosis<sup>210</sup> is also of interest, especially because these diseases are caused by autosomal dominant mutations in non-*MAPT* genes. Given their rarity an international collaboration would need to be established, with a focus initially on better characterising the post-mortem brain pathology. An advantage of investigating these diseases is that they are caused by autosomal dominant mutations with a high penetrance, which would enable identification of cases with a high level of certainty using genetic testing, and allow the study of the presymptomatic period providing a window in the earliest stages of the disease process in the brain. Assuming that the predominant brain pathology is indeed 3R tau in these diseases, this approach could facilitate 3R tau biomarker development, and allow increased recruitment to clinical trials to test disease modifying therapies specifically targeting 3R tau pathology.

# Appendix A - 4R tau imaging cohort

## 1. 4RTNI Consortium

**Adam Boxer** - Department of Neurology, Memory and Aging Center,  
University of California, San Francisco, CA, USA.

**Lawren VandeVrede** - Department of Neurology, Memory and Aging Center,  
University of California, San Francisco, CA, USA.

**Hilary Heuer** - Department of Neurology, Memory and Aging Center,  
University of California, San Francisco, CA, USA.

**Bradley F. Boeve** - Department of Neurology, Mayo Clinic, Rochester, MN  
55905, USA

**Brad C. Dickerson** - Departments of Neurology and Psychiatry,  
Frontotemporal Disorders Unit and Alzheimer's Disease Research Center,  
Boston Massachusetts USA.

**Carmela M. Tartaglia** - Tanz Centre for Research in Neurodegenerative  
Diseases, University of Toronto, Toronto, Canada.

**Irene Litvan** - Department of Neurosciences, University of California San  
Diego, La Jolla, California, USA.

**Murray Grossman** - Department of Neurology, University of Pennsylvania,  
Philadelphia, USA.

**Alex Pantelyat** – Department of Neurology. School of Medicine, Johns  
Hopkins University, Baltimore, MD, USA.

**Edward D. Huey** - Department of Psychiatry and Neurology, Columbia  
University, New York, New York, USA.

**David J. Irwin** - Penn Center for Neurodegenerative Disease Research,  
University of Pennsylvania School of Medicine, Philadelphia, PA, USA.

**Anne Fagan** - Department of Neurology, Washington University School of  
Medicine, St Louis, MO, USA.

**Suzanne L. Baker** - Molecular Biophysics and Integrated Bioimaging,  
Lawrence Berkeley National Laboratory, Berkeley, CA, USA.

**Arthur W. Toga** - Laboratory of Neuro Imaging, Stevens Neuroimaging and  
Informatics Institute, Keck School of Medicine of USC, University of Southern  
California, Los Angeles, CA, United States.

## 2. PROSPECT Consortium

**Huw R. Morris**, FRCP, PhD - Department of Clinical and Movement Neurosciences, UCL (University College London) Queen Square Institute of Neurology, London, United Kingdom.

**James B. Rowe**, FRCP, PhD - Cambridge University Department of Clinical Neurosciences and Cambridge University Hospitals NHS Trust; Medical Research Council Cognition and Brain Sciences Unit, Cambridge UK.

**Alyssa A. Costantini**, MSc - Department of Clinical and Movement Neurosciences, UCL (University College London) Queen Square Institute of Neurology, London, United Kingdom.

**Henry Houlden**, FRCP, PhD - Department of Clinical and Movement Neurosciences, UCL (University College London) Queen Square Institute of Neurology, London, United Kingdom; Movement Disorders Centre, UCL Queen Square Institute of Neurology, London, United Kingdom; Department of Neuromuscular Diseases, UCL Queen Square Institute of Neurology, London, United Kingdom.

**Christopher Kobylecki**, FRCP, PhD - Department of Neurology, Manchester Academic Health Science Centre, Salford Royal NHS (National Health Service) Foundation Trust, University of Manchester, Manchester, United Kingdom.

**Michele T. M. Hu**, FRCP, PhD - Division of Neurology, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, United Kingdom

**Nigel Leigh**, FRCP, PhD - Department of Neuroscience, Brighton and Sussex Medical School, Brighton, United Kingdom.

### **3. UCL FTD cohort**

**Jonathan D. Rohrer**, FRCP, PhD - Dementia Research Centre,  
Department of Neurodegenerative Disease, UCL Queen Square Institute of  
Neurology, University College London, London, UK.

**Jason Warren**, FRCP, PhD - Dementia Research Centre, Department of  
Neurodegenerative Disease, UCL Queen Square Institute of Neurology,  
University College London, London, UK.

**Martina Bocchetta**, PhD - Dementia Research Centre, Department of  
Neurodegenerative Disease, UCL Queen Square Institute of Neurology,  
University College London, London, UK.

**David Cash**, PhD - Dementia Research Centre, Department of  
Neurodegenerative Disease, UCL Queen Square Institute of Neurology,  
University College London, London, UK.

# Appendix B - Pick's disease International Consortium (PIC)

## Consortia Members

Name	Degree	Department
William J. Scotton	Mbioch., MRCP.	Dementia Research Centre, Department of Neurodegenerative Disease, University College London, Queen Square Institute of Neurology, London, UK
Rebecca R. Valentino	Ph.D.	Department of Neuroscience, Mayo Clinic, Jacksonville, FL 32224, USA
William J. Scotton	Mbioch., MRCP.	Dementia Research Centre, Department of Neurodegenerative Disease, University College London, Queen Square Institute of Neurology, London, UK
Shanu F. Roemer	M.D.	Department of Neuroscience, Mayo Clinic, Jacksonville, FL 32224, USA
Tammaryn Lashley	Ph.D.	Queen Square Brain Bank for Neurological Disorders, University College London, Queen Square Institute of Neurology London, UK
		Department of Neurodegenerative Disease, University College London, Queen Square Institute of Neurology, London, UK
Michael G. Heckman	M.Sc.	Division of Clinical Trials and Biostatistics, Mayo Clinic, Jacksonville, FL 32224, USA
Maryam Shoi	Ph.D.	Department of Neurodegenerative Disease, University College London, Queen Square Institute of Neurology, London, UK
Alejandro Martinez-Carrasco	M.Sc.	Department of Clinical and Movement Neurosciences, University College London, Queen Square Institute of Neurology, London, UK
Nicole Tamavaka	B.Sc.	Department of Neuroscience, Mayo Clinic, Jacksonville, FL 32224, USA
Ronald L. Walton	B.Sc.	Department of Neuroscience, Mayo Clinic, Jacksonville, FL 32224, USA
Matthew C. Baker	B.Sc.	Department of Neuroscience, Mayo Clinic, Jacksonville, FL 32224, USA
Hannah L. Macpherson	M.Sc.	Department of Neurodegenerative Disease, University College London, Queen Square Institute of Neurology, London, UK
Raquel Real	Ph.D.	Department of Clinical and Movement Neurosciences, University College London, Queen Square Institute of Neurology, London, UK
Alexandra I. Soto-Beasley	M.Sc.	Department of Neuroscience, Mayo Clinic, Jacksonville, FL 32224, USA
Kin Mok	Ph.D., FRCP	Department of Neurodegenerative Disease, University College London, Queen Square Institute of Neurology, London, UK
		UK Dementia Research Institute at UCL, London, UK
		Division of Life Science, State Key Laboratory of Molecular Neuroscience, Molecular Neuroscience Center, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China
		Hong Kong Center for Neurodegenerative Diseases, Hong Kong Science Park, Hong Kong, China.
Tamas Revesz	Ph.D., FRCPath	Queen Square Brain Bank for Neurological Disorders, University College London, Queen Square Institute of Neurology London, UK
		Department of Neurodegenerative Disease, University College London, Queen Square Institute of Neurology, London, UK
Thomas T. Warner		Queen Square Brain Bank for Neurological Disorders, University College London, Queen Square Institute of Neurology London, UK
		Department of Clinical and Movement Neurosciences, University College London, Queen Square Institute of Neurology, London, UK

Zane Jaunmuktane		Queen Square Brain Bank for Neurological Disorders, University College London, Queen Square Institute of Neurology London, UK
		Department of Clinical and Movement Neurosciences, University College London, Queen Square Institute of Neurology, London, UK
Bradley F. Boeve	M.D.	Department of Neurology, Mayo Clinic, Rochester, MN 55905, USA
Elizabeth A. Christopher	M.B.A.	Department of Neuroscience, Mayo Clinic, Jacksonville, FL 32224, USA
Michael DeTure	Ph.D.	Department of Neuroscience, Mayo Clinic, Jacksonville, FL 32224, USA
Ranjan Duara	M.D.	Wien Center for Alzheimer's Disease and Memory Disorders, Mount Sinai Medical Center Miami Beach, FL
Neill R. Graff-Radford	M.D.	Department of Neurology, Mayo Clinic, Jacksonville, FL 32224, USA
Keith A. Josephs	M.D.	Department of Neurology, Mayo Clinic, Rochester, MN 55905, USA
David S. Knopman	M.D.	Department of Neurology, Mayo Clinic, Rochester, MN 55905, USA
Shunsuke Koga	M.D., Ph.D.	Department of Neuroscience, Mayo Clinic, Jacksonville, FL 32224, USA
Melissa E. Murray	Ph.D.	Department of Neuroscience, Mayo Clinic, Jacksonville, FL 32224, USA
Kelly E. Lyons	Ph.D.	University of Kansas Medical Center, Parkinson's Disease & Movement Disorder Division, Kansas City, KS. 66160
Rajesh Pahwa	M.D.	University of Kansas Medical Center, Parkinson's Disease & Movement Disorder Division, Kansas City, KS. 66160
Joseph E. Parisi	M.D.	Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55905, USA
Ronald C. Petersen	M.D., Ph.D.	Department of Neurology, Mayo Clinic, Rochester, MN 55905, USA
Jennifer Whitwell	Ph.D.	Department of Radiology, Mayo Clinic, Rochester, MN 55905, USA
Lea T. Grinberg	M.D., Ph.D.	Department of Neurology, Memory and Aging Center, University of California San Francisco, San Francisco, CA, USA
Bruce Miller	M.D.	Department of Neurology, Memory and Aging Center, University of California San Francisco, San Francisco, CA, USA
Athena Schlereth		Department of Neurology, Memory and Aging Center, University of California San Francisco, San Francisco, CA, USA
William W. Seeley	M.D.	Department of Neurology, Memory and Aging Center, University of California San Francisco, San Francisco, CA, USA
Salvatore Spina	M.D., Ph.D.	Department of Neurology, Memory and Aging Center, University of California San Francisco, San Francisco, CA, USA
Murray Grossman	M.D., Ph.D.	Department of Neurology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA
David J. Irwin	M.D.	Department of Neurology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA
Edward B. Lee	M.D., Ph.D.	Department of Pathology and Laboratory Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA
EunRan Suh	Ph.D.	Department of Pathology and Laboratory Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA
John Q. Trojanowski	M.D., Ph.D.	Department of Pathology and Laboratory Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA
Vivianna M. Van Deerlin	M.D., Ph.D.	Department of Pathology and Laboratory Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA
David A. Wolk	M.D.	Department of Neurology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA 19104, USA
Theresa R. Connors	B.S.	Neuropathology Service, C.S. Kubik Laboratory for Neuropathology, Massachusetts General Hospital/Harvard Medical School, Boston, MA, USA



Patrick M. Dooley	B.A.	Neuropathology Service, C.S. Kubik Laboratory for Neuropathology, Massachusetts General Hospital/Harvard Medical School, Boston, MA, USA
Matthew P. Frosch	M.D., Ph.D.	Neuropathology Service, C.S. Kubik Laboratory for Neuropathology, Massachusetts General Hospital/Harvard Medical School, Boston, MA, USA
Derek H. Oakley	M.D.	Neuropathology Service, C.S. Kubik Laboratory for Neuropathology, Massachusetts General Hospital/Harvard Medical School, Boston, MA, USA
Iban Aldecoa	M.D., Ph.D.	Pathology, BDC, Hospital Clinic de Barcelona, Barcelona, Spain
		University of Barcelona, Barcelona, Spain
		Neurological Tissue Bank, Biobanc-Hospital Clínic-FRCB-IDIBAPS, Barcelona, Spain
Mircea Balasa	M.D., Ph.D.	Alzheimer's Disease and other Cognitive Disorders Unit, Neurology Department, Hospital Clinic, Barcelona, Spain
		Barcelona Clinical Research Foundation-August Pi i Sunyer Biomedical Research Institute (FRCB-IDIBAPS), Barcelona, Spain
Ellen Gelpi	M.D., Ph.D.	Division of Neuropathology and Neurochemistry, Department of Neurology, Medical University of Vienna, Vienna, Austria.
Sergi Borrego-Écija	M.D.	Alzheimer's Disease and other Cognitive Disorders Unit, Neurology Department, Hospital Clinic, Barcelona, Spain
		Barcelona Clinical Research Foundation-August Pi i Sunyer Biomedical Research Institute (FRCB-IDIBAPS), Barcelona, Spain
		University of Barcelona, Barcelona, Spain
Rosa Maria de Eugenio Huéllamo	M.D.	Hospital de Palamós, Carrer Hospital, 36, 17230 Palamós, Girona, Spain
Jordi Gascon-Bayarri	M.D.	Servei de Neurologia, Hospital Universitari de Bellvitge. Institut d'Investigació Biomèdica de Bellvitge (Idibell). L'Hospitalet de Llobregat, Spain.
Raquel Sánchez-Valle	M.D., Ph.D.	Alzheimer's Disease and other Cognitive Disorders Unit, Neurology Department, Hospital Clinic, Barcelona, Spain
		Barcelona Clinical Research Foundation-August Pi i Sunyer Biomedical Research Institute (FRCB-IDIBAPS), Barcelona, Spain
		University of Barcelona, Barcelona, Spain
Pilar Sanz-Cartagena	M.D., Ph.D.	Hospital de Mataró, Carr. de Cirera, 230, 08304 Mataró, Barcelona, Spain.
Gerard Piñol-Ripoll	M.D., Ph.D.	Unitat Trastorns Cognitius (Cognitive Disorders Unit), Clinical Neuroscience Research, IRBLleida, Santa Maria University Hospital, Lleida, Spain
Laura Molina-Porcel	M.D., Ph.D.	Neurological Tissue Bank, Biobanc-Hospital Clínic-FRCB-IDIBAPS, Barcelona, Spain
		Alzheimer's Disease and other Cognitive Disorders Unit, Neurology Department, Hospital Clinic, Barcelona, Spain
		Barcelona Clinical Research Foundation-August Pi i Sunyer Biomedical Research Institute (FRCB-IDIBAPS), Barcelona, Spain
Eileen H. Bigio	M.D.	Mesulam Center for Cognitive Neurology & Alzheimer's Disease, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
		Department of Pathology, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
Margaret E. Flanagan	M.D.	Mesulam Center for Cognitive Neurology & Alzheimer's Disease, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
		Department of Pathology, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
Tamar Gefen	Ph.D.	Mesulam Center for Cognitive Neurology & Alzheimer's Disease, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
		Department of Psychiatry and Behavioral Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

Emily J . Rogalski	Ph.D.	Mesulam Center for Cognitive Neurology & Alzheimer's Disease, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
		Department of Psychiatry and Behavioral Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
Sandra Weintraub	Ph.D.	Mesulam Center for Cognitive Neurology & Alzheimer's Disease, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
		Department of Psychiatry and Behavioral Sciences, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
Javier Redding-Ochoa	M.D.	Johns Hopkins School of Medicine, Baltimore, MD, USA
Koping Chang	M.D.	Johns Hopkins School of Medicine, Baltimore, MD, USA
Juan C. Troncoso	M.D.	Johns Hopkins School of Medicine, Baltimore, MD, USA
Stefan Prokop	M.D.	Fixel Institute for Neurological Diseases, University of Florida, Gainesville, FL, USA.
Kathy L. Newell	M.D.	Department of Pathology and Laboratory Medicine, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA.
Bernardino Ghetti	M.D.	Department of Pathology and Laboratory Medicine, Indiana University School of Medicine, Indianapolis, Indiana 46202, USA.
Matthew Jones	M.D., FRCP.	Cerebral Function Unit, Manchester Centre for Clinical Neurosciences, Salford Royal NHS Foundation Trust, UK.
		Division of Neuroscience, School of Biological Sciences, University of Manchester, UK.
Anna Richardson	B.Sc., FRCP	Cerebral Function Unit, Manchester Centre for Clinical Neurosciences, Salford Royal NHS Foundation Trust, UK.
		Division of Neuroscience, School of Biological Sciences, University of Manchester, UK.
Andrew C. Robinson	Ph.D.	Division of Neuroscience, Faculty of Biology, Medicine and Health, School of Biological Sciences, The University of Manchester, Salford Royal Hospital, Salford, M6 8HD, UK.
		Geoffrey Jefferson Brain Research Centre, Manchester Academic Health Science Centre (MAHSC), Manchester, UK.
Federico Roncaroli	M.D.	Division of Neuroscience, Faculty of Biology, Medicine and Health, School of Biological Sciences, The University of Manchester, Salford Royal Hospital, Salford, M6 8HD, UK.
		Geoffrey Jefferson Brain Research Centre, Manchester Academic Health Science Centre (MAHSC), Manchester, UK.
Julie Snowden	Ph.D.	Cerebral Function Unit, Manchester Centre for Clinical Neurosciences, Salford Royal NHS Foundation Trust, UK.
		Division of Neuroscience, School of Biological Sciences, University of Manchester, UK.
Kieren Allinson	FRCPath.	Histopathology Box 235 Cambridge University Hospital NHS Foundation Trust, Cambridge Biomedical Campus, Hills Road, Cambridge, CB2 0QQ
Oliver Green	B.Sc.	Histopathology Box 235 Cambridge University Hospital NHS Foundation Trust, Cambridge Biomedical Campus, Hills Road, Cambridge, CB2 0QQ
James B. Rowe	Ph.D.	Cambridge University Department of Clinical Neurosciences and Cambridge University Hospitals NHS Trust, Cambridge, UK
		Medical Research Council Cognition and Brain Sciences Unit, Cambridge, UK
Poonam Singh	M.Phil.	Histopathology Box 235 Cambridge University Hospital NHS Foundation Trust, Cambridge Biomedical Campus, Hills Road, Cambridge, CB2 0QQ
Thomas G. Beach	M.D., Ph.D.	Civin Laboratory of Neuropathology, Banner Sun Health Research Institute, Sun City, AZ 85351, USA
Geidy E. Serrano	Ph.D.	Civin Laboratory of Neuropathology, Banner Sun Health Research Institute, Sun City, AZ 85351, USA

Xena E. Flowers	B.S.	Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University, New York, NY, USA
James E. Goldman	M.D., Ph.D.	Department of Pathology and Cell Biology, Columbia University Irving Medical Center, Vagelos College of Physicians and Surgeons, Columbia University, New York, NY 10032, USA.
Allison C. Heaps	M.Sc.	Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University, New York, NY, USA
Sandra P. Leskinen	M.A.	Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University, New York, NY, USA
Andrew F. Teich	M.D., Ph.D.	Department of Pathology and Cell Biology, Columbia University, New York, NY, USA
		Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University, New York, NY, USA
Sandra E. Black	M.D., FRCPC.	Department of Medicine, Division of Neurology, Sunnybrook Health Sciences Centre and University of Toronto, Hurvitz Brain Sciences Research Program, Sunnybrook Research Institute
Julia L. Keith	M.D., FRCPC.	Laboratory Medicine and Molecular Diagnostics, Sunnybrook Health Sciences Centre, and Laboratory Medicine and Pathobiology, University of Toronto
Mario Masellis	M.D., Ph.D., FRCPC.	Department of Medicine, Division of Neurology, Sunnybrook Health Sciences Centre and University of Toronto, Hurvitz Brain Sciences Research Program, Sunnybrook Research Institute
Istvan Bodi	FRCPath.	Clinical Neuropathology Department, King's College Hospital NHS Foundation Trust, London, UK
		London Neurodegenerative Diseases Brain Bank, Department of Basic and Clinical Neuroscience, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
Andrew King	FRCPath.	Clinical Neuropathology Department, King's College Hospital NHS Foundation Trust, London, UK
		London Neurodegenerative Diseases Brain Bank, Department of Basic and Clinical Neuroscience, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
Safa-Al Sarraj	FRCPath.	Clinical Neuropathology Department, King's College Hospital NHS Foundation Trust, London, UK
		London Neurodegenerative Diseases Brain Bank, Department of Basic and Clinical Neuroscience, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
Claire Troakes	Ph.D.	London Neurodegenerative Diseases Brain Bank, Department of Basic and Clinical Neuroscience, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
Glenda M. Halliday	Ph.D.	University of Sydney Brain and Mind Centre and Faculty of Medicine and Health School of Medical Sciences
John R. Hodges	M.D.	University of Sydney Brain and Mind Centre and Faculty of Medicine and Health School of Medical Sciences
Jillian J. Kril	Ph.D.	University of Sydney Faculty of Medicine and Health School of Medical Sciences
John B. Kwok	Ph.D.	University of Sydney Brain and Mind Centre and Faculty of Medicine and Health School of Medical Sciences
Olivier Piguet	Ph.D.	University of Sydney Brain and Mind Centre and Faculty of Science School of Psychology
Marla Gearing	Ph.D.	Dept. of Pathology and Laboratory Medicine, Dept. of Neurology, and Goizueta Alzheimer's Disease Center Brain Bank; Emory University School of Medicine, Atlanta, GA USA
Thomas Arzberger	M.D.	Department of Psychiatry and Psychotherapy, University Hospital, Ludwig-Maximilians-University Munich, Germany
Sigrun Roeber	M.D.	Center for Neuropathology and Prion Research, Ludwig-Maximilians-University Munich, Germany

Johannes Attems	M.D.	Newcastle Brain Tissue Resource, Translational and Clinical Research Institute, Newcastle University, Newcastle upon Tyne, NE4 5PL, UK
Christopher M. Morris	Ph.D.	Newcastle Brain Tissue Resource, Translational and Clinical Research Institute, Newcastle University, Newcastle upon Tyne, NE4 5PL, UK
Alan J. Thomas	Ph.D.	Newcastle Brain Tissue Resource, Translational and Clinical Research Institute, Newcastle University, Newcastle upon Tyne, NE4 5PL, UK
Bret M. Evers	M.D., Ph.D.	University of Texas Southwestern Medical Center, Dallas, TX 75390
Charles L. White, 3rd	M.D.	University of Texas Southwestern Medical Center, Dallas, TX 75390
Naguib Mechawar	Ph.D.	Douglas Hospital Research Centre
Anne A. Sieben	MD, Ph.D.	Laboratory of Neurology, Translational Neurosciences, Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium.
		IBB-NeuroBiobank BB190113, Born Bunge Institute, Antwerp, Belgium
		Department of Pathology, Antwerp University Hospital, Antwerp, Belgium
		Department of Neurology, Ghent University Hospital, Ghent University, Belgium
Patrick P. Cras	MD., Ph.D.	Laboratory of Neurology, Translational Neurosciences, Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium.
		IBB-NeuroBiobank BB190113, Born Bunge Institute, Antwerp, Belgium
		Department of Neurology, Antwerp University Hospital - UZA, Antwerp, Belgium
Bart B. De Vil	M.Sc., Eng.	Laboratory of Neurology, Translational Neurosciences, Faculty of Medicine and Health Sciences, University of Antwerp, Antwerp, Belgium.
		IBB-NeuroBiobank BB190113, Born Bunge Institute, Antwerp, Belgium
		Department of Neurology, Antwerp University Hospital - UZA, Antwerp, Belgium
Peter Paul P.P. . De Deyn	MD, Ph.D.	Laboratory of Neurochemistry and Behavior, Experimental Neurobiology Unit, University of Antwerp, Universiteitsplein 1, 2610 Antwerpen, Belgium
Charles Duyckaerts	M.D., Ph.D.	Laboratoire de Neuropathologie Escourolle, Hôpital de la Salpêtrière, AP-HP, & Alzheimer Prion Team, ICM, 47 Bd de l'Hôpital, 75651 CEDEX 13 Paris, France.
Isabelle Le Ber	M.D., Ph.D.	Inserm U1127, CNRS UMR 7225, Sorbonne Université, Paris Brain Institute (ICM), Hôpital Pitié-Salpêtrière, Paris, France
		Centre de référence des démences rares ou précoces, Hôpital Pitié-Salpêtrière, Paris, France
Danielle Seihean	M.D., Ph.D.	Laboratoire de Neuropathologie Escourolle, Hôpital de la Salpêtrière, AP-HP, & ICM, 47 Bd de l'Hôpital, 75651 CEDEX 13 Paris, France.
Sabrina Turbant-Leclerc	Ph.D.	Inserm U1127, CNRS UMR 7225, Sorbonne Université, Paris Brain Institute (ICM) Hôpital Pitié-Salpêtrière, Paris, France.
Ian R. MacKenzie	M.D.	Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, BC Canada V6T 2B5
Catriona McLean	M.D., Ph.D.	Department of Anatomical Pathology Alfred Heath, Melbourne, Victoria, 3004, Australia
		Victorian Brain Bank, The Florey Institute of Neuroscience of Mental Health, Parkville, Victoria, 3052, Australia
Matthew D. Cykowski	M.D.	Department of Pathology and Genomic Medicine, Houston Methodist Research Institute and Weill Cornell Medicine, Houston, TX.
John F. Ervin		Department of Neurology, Duke University Medical Center, Durham, USA
Shih-Hsiu J. Wang	M.D., Ph.D.	Department of Neurology, Duke University Medical Center, Durham, USA
Caroline Graff	Ph.D.	Division for Neurogeriatrics, Centre for Alzheimer Research, Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden.

		Unit for Hereditary Dementias, Karolinska University Hospital Solna, Stockholm, Sweden.
Inger Nennesmo	M.D.	Dept of laboratory Medicine Huddinge Karolinska Institutet, Stockholm Sweden.
		Dept of Pathology, Karolinska University Hospital Solna, Stockholm, Sweden.
Rashed M. Nagra	Ph.D.	Human Brain and Spinal Fluid Resource Center, Brentwood Biomedical Research Institute, Los Angeles, CA, United States
James Riehl	B.S.	UCLA - Sepulveda, Los Angeles, CA
Gabor G. Kovacs	M.D., Ph.D.	Tanz Centre for Research in Neurodegenerative Disease (CRND) and Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada.
		Laboratory Medicine Program and Krembil Brain Institute, University Health Network, Toronto, ON, Canada.
Giorgio Giaccone	M.D.	Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy
Benedetta Nacmias	Ph.D.	Department of Neuroscience, Psychology, Drug Research and Child Health University of Florence, Florence, Italy
		IRCCS Fondazione Don Carlo Gnocchi, Florence, Italy
Manuela Neumann	M.D.	Molecular Neuropathology of Neurodegenerative Diseases, German Center for Neurodegenerative Diseases (DZNE), Tübingen, Germany
		Department of Neuropathology, University Hospital of Tübingen, Tübingen, Germany
Lee-Cyn Ang	M.D.	Department of Pathology and Laboratory Medicine, London Health Sciences Centre, London, ON, Canada
		Schulich School of Medicine and Dentistry, Western University, London, ON, Canada
Elizabeth C. Finger	M.D.	Department of Clinical Neurological Sciences, Western University, London, ON, Canada.
		Schulich School of Medicine and Dentistry, Western University, London, ON, Canada.
Cornelis Blauwendraat	Ph.D.	Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA
Mike A. Nalls	Ph.D.	Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA
		Center for Alzheimer's and Related Dementias, National Institutes of Health, Bethesda, MD, USA
		Data Tecnica International LLC, Washington, DC, USA
Andrew B. Singleton	Ph.D.	Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA
Dan Vitale	M.Sc.	Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, USA
		Center for Alzheimer's and Related Dementias, National Institutes of Health, Bethesda, MD, USA
		Data Tecnica International LLC, Washington, DC, USA
Cristina Cunha	Ph.D.	Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal
Agostinho Carvalho	Ph.D.	Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal
		ICVS/3B's - PT Government Associate Laboratory, Braga/Guimarães, Portugal
Zbigniew K. Wszolek	M.D.	Department of Neurology, Mayo Clinic, Jacksonville, FL 32224, USA
Huw R. Morris	Ph.D., FRCP	Department of Clinical and Movement Neurosciences, University College London, Queen Square Institute of Neurology, London, UK
Rosa Rademakers	Ph.D.	VIB-UAntwerp Center for Molecular Neurology, University of Antwerp, Antwerp 2610, Belgium

		Department of Neuroscience, Mayo Clinic, Jacksonville, FL 32224, USA
John A. Hardy	Ph.D.	UK Dementia Research Institute at UCL, London, UK
		Reta Lila Weston Institute, University College London, Queen Square Institute of Neurology, London, UK
		Department of Neurodegenerative Disease, University College London, Queen Square Institute of Neurology, London, UK
		Institute for Advanced Study, The Hong Kong University of Science and Technology, Hong Kong, China
Dennis W. Dickson	M.D.	Department of Neuroscience, Mayo Clinic, Jacksonville, FL 32224, USA
Jonathan D. Rohrer	Ph.D., FRCP	Dementia Research Centre, Department of Neurodegenerative Disease, University College London, Queen Square Institute of Neurology, London, UK
Owen A. Ross	Ph.D.	Department of Neuroscience, Mayo Clinic, Jacksonville, FL 32224, USA
		Department of Clinical Genomics, Mayo Clinic, Jacksonville, FL 32224, USA

# Bibliography

1. Erkinen MG, Kim MO, Geschwind MD. Clinical neurology and epidemiology of the major neurodegenerative diseases. *Cold Spring Harb Perspect Biol.* 2018;10(4):1-44.
2. Feigin VL, Nichols E, Alam T, et al. Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* 2019;18(5):459-480.
3. Nyaaba GN, Stronks K, De-Graft Aikins A, Kengne AP, Agyemang C. Tracing Africa's progress towards implementing the Non-Communicable Diseases Global action plan 2013-2020: a synthesis of WHO country profile reports. *BMC Public Health.* 2017;17(1):1-13.
4. Nichols E, Szeke CEI, Vollset SE, et al. Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* 2019;18(1):88-106.
5. Brookmeyer R, Johnson E, Ziegler-graham K, Arrighi HM. Forecasting the global burden of Alzheimer's disease. 2007;3:186-191.
6. Wimo A, Guerchet M, Ali GC, et al. The worldwide costs of dementia 2015 and comparisons with 2010. *Alzheimer's Dement.* 2017;13(1):1-7.
7. Winblad B, Amouyel P, Andrieu S, et al. Defeating Alzheimer's disease and other dementias: A priority for European science and society. *Lancet Neurol.* 2016;15(5):455-532.
8. van Dyck CH, Swanson CJ, Aisen P, et al. Lecanemab in Early Alzheimer's Disease. *N Engl J Med.* 2022;9-21.
9. Ratnavalli E, Brayne C, Dawson K, Hodges JR. The prevalence of frontotemporal dementia. *Neurology.* 2002;58(11):1615-1621.
10. Harvey RJ, Skelton-Robinson M, Rossor MN. The prevalence and causes of dementia in people under the age of 65 years. *J Neurol Neurosurg Psychiatry.* 2003;74(9):1206-1209.
11. Van Der Flier WM, Scheltens P. Epidemiology and risk factors of dementia. *Neurol Pract.* 2005;76(SUPPL. 5):2-7.
12. Cairns NJ, Bigio EH, Mackenzie IRA, et al. Neuropathologic diagnostic and nosologic criteria for frontotemporal lobar degeneration: Consensus of the Consortium for Frontotemporal Lobar Degeneration. *Acta Neuropathol.* 2007;114(1):5-22.
13. Mercy L, Hodges JR, Dawson K, Barker RA, Brayne C. Incidence of early-onset dementias in Cambridgeshire, United Kingdom. *Neurology.* 2008;71(19):1496-1499.
14. Coyle-Gilchrist ITS, Dick KM, Patterson K, et al. Prevalence, characteristics, and survival of frontotemporal lobar degeneration syndromes. *Neurology.* 2016;86(18):1736-1743.
15. Mackenzie IRA, Neumann M, Bigio EH, et al. Nomenclature and nosology for

- neuropathologic subtypes of frontotemporal lobar degeneration: an update. *Acta Neuropathol.* 2010;119:1-4.
16. Rascovsky K, Hodges JR, Knopman D, et al. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. *Brain.* 2011;134(9):2456-2477.
  17. Gorno-Tempini ML, Hillis AE, Weintraub S, et al. Classification of primary progressive aphasia and its variants. *Neurology.* 2011;76(11):1006-1014.
  18. Rohrer et al. Launchpad Report - Frontotemporal Dementia, Neurological and Neuropsychiatric Diseases Working Group - 2019.
  19. Convery R, Mead S, Rohrer JD. Review: Clinical, genetic and neuroimaging features of frontotemporal dementia. *Neuropathol Appl Neurobiol.* 2019;45(1):6-18.
  20. Marshall CR, Hardy CJD, Volkmer A, et al. Primary progressive aphasia: a clinical approach. *J Neurol.* 2018;265(6):1474-1490.
  21. Garibotto V, Borroni B, Agosti C, et al. Subcortical and deep cortical atrophy in Frontotemporal Lobar Degeneration. *Neurobiol Aging.* 2011;32(5):875-884.
  22. Bocchetta M, Malpetti M, Todd EG, Rowe JB, Rohrer JD. Looking beneath the surface: the importance of subcortical structures in frontotemporal dementia. *Brain Commun.* 2021;3(3).
  23. Mackenzie IRA, Neumann M. Molecular neuropathology of frontotemporal dementia: insights into disease mechanisms from postmortem studies. *J Neurochem.* 2016;138:54-70.
  24. Neumann M, Mackenzie IRA. Review: Neuropathology of non-tau frontotemporal lobar degeneration. *Neuropathol Appl Neurobiol.* 2019;45(1):19-40.
  25. Mackenzie IRA, Neumann M, Bigio EH, et al. Nomenclature for neuropathologic subtypes of frontotemporal lobar degeneration: Consensus recommendations. *Acta Neuropathol.* 2009;117(1):15-18.
  26. Murley AG. Behavioural disinhibition in the syndromes associated with frontotemporal lobar degeneration. Cambridge, University of Cambridge 2020.
  27. Götz J, Halliday G, Nisbet RM. Molecular Pathogenesis of the Tauopathies. *Annu Rev Pathol Mech Dis.* 2019;14(1):239-261.
  28. Lee G, Leugers CJ. Tau and tauopathies. In: *Progress in Molecular Biology and Translational Science.* Vol 107. Elsevier B.V.; 2012:263-293.
  29. Lee VMY, Goedert M, Trojanowski JQ. Neurodegenerative tauopathies. *Annu Rev Neurosci.* 2001;24(1):1121-1159.
  30. Forrest SL, Halliday GM, McCann H, et al. Heritability in frontotemporal tauopathies. *Alzheimer's Dement Diagnosis, Assess Dis Monit.* 2019;11:115-124.
  31. Buée L, Bussi re T, Bu e-Scherrer V, Delacourte A, Hof PR. Tau protein isoforms, phosphorylation and role in neurodegenerative disorders. *Brain Res Rev.* 2000;33(1):95-130.



32. Ittner A, Ittner LM. Dendritic Tau in Alzheimer's Disease. *Neuron*. 2018;99(1):13-27.
33. Goedert M. Tau filaments in neurodegenerative diseases. *FEBS Lett*. 2018;592(14):2383-2391.
34. Goedert M, Eisenberg DS, Crowther RA. Propagation of Tau Aggregates and Neurodegeneration. *Annu Rev Neurosci*. 2017;40:189-210.
35. Goedert M, Jakes R. Expression of separate isoforms of human tau protein: Correlation with the tau pattern in brain and effects on tubulin polymerization. *EMBO J*. 1990;9(13):4225-4230.
36. Noble W, Hanger DP, Miller CCJ, Lovestone S. The importance of tau phosphorylation for neurodegenerative diseases. *Front Neurol*. 2013;4 JUL(July):1-11.
37. Sergeant N, Bretteville A, Hamdane M, et al. Biochemistry of Tau in Alzheimer's disease and related neurological disorders. *Expert Rev Proteomics*. 2008;5(2):207-224.
38. Ballatore C, Lee VMY, Trojanowski JQ. Tau-mediated neurodegeneration in Alzheimer's disease and related disorders. *Nat Rev Neurosci*. 2007;8(9):663-672.
39. Lee VMY, Brunden KR, Hutton M, Trojanowski JQ. Developing therapeutic approaches to tau, selected kinases, and related neuronal protein targets. *Cold Spring Harb Perspect Med*. 2011;1(1).
40. Colin M, Dujardin S, Schraen-Maschke S, et al. From the prion-like propagation hypothesis to therapeutic strategies of anti-tau immunotherapy. *Acta Neuropathol*. 2020;139(1):3-25.
41. Shi Y, Zhang W, Yang Y, et al. Structure-based classification of tauopathies. *Nature*. 2021;598(7880):359-363
42. Kovacs GG. Tauopathies. *Handb Clin Neurol*. 2018;145:355-368.
43. Fitzpatrick AWP, Falcon B, He S, et al. Cryo-EM structures of tau filaments from Alzheimer's disease. *Nature*. 2017;547(7662):185-190.
44. Falcon B, Zhang W, Schweighauser M, et al. Tau filaments from multiple cases of sporadic and inherited Alzheimer's disease adopt a common fold. *Acta Neuropathol*. 2018;136(5):699-708.
45. Falcon B, Zivanov J, Zhang W, et al. Novel tau filament fold in chronic traumatic encephalopathy encloses hydrophobic molecules. *Nature*. 2019;568(7752):420-423.
46. Zhang W, Tarutani A, Newell KL, et al. Novel tau filament fold in corticobasal degeneration. *Nature*. 2020;580(7802):283-287.
47. Falcon B, Zhang W, Murzin AG, et al. Structures of filaments from Pick's disease reveal a novel tau protein fold. *Nature*. 2018;561(7721):137-140.
48. Stahlberg H, Riek R. Structural strains of misfolded tau protein define different diseases. *Nature*. 2021;598(7880):264-265.
49. Li J, Browning S, Mahal SP, Oelschlegel AM, Weissmann C. Darwinian evolution of prions in cell culture. *Science (80- )*. 2010;327(5967):869-872.

50. Clavaguera F, Akatsu H, Fraser G, et al. Brain homogenates from human tauopathies induce tau inclusions in mouse brain. *Proc Natl Acad Sci U S A*. 2013;110(23):9535-9540.
51. Kovacs GG, Lukic MJ, Irwin DJ, et al. Distribution patterns of tau pathology in progressive supranuclear palsy. *Acta Neuropathol*. 2020;140(2):99-119.
52. Saito Y, Ruberu NN, Sawabe M, et al. Staging of argyrophilic grains: An age-associated tauopathy. *J Neuropathol Exp Neurol*. 2004;63(9):911-918.
53. Irwin DJ. Tauopathies as clinicopathological entities. *Park Relat Disord*. 2016;22:S29-S33.
54. Ghetti B, Oblak AL, Boeve BF, Johnson KA, Dickerson BC, Goedert M. Invited review: Frontotemporal dementia caused by microtubule-associated protein tau gene (MAPT) mutations: a chameleon for neuropathology and neuroimaging Frontotemporal dementia caused by microtubule-associated protein tau gene (MAPT) mutations: a chame. *Neuropathol Appl Neurobiol*. 2015;41:24-46.
55. Pick A. Über die Beziehungen der senilen Hirnatrophie zur Aphasie. *Prag. Med. Wochenschr*. 1892. 17:165–67.
56. Kovacs GG. Molecular pathology of neurodegenerative diseases: principles and practice. *J Clin Pathol*. 2019;72:725-735.
57. Arima K. Ultrastructural characteristics of tau filaments in tauopathies: Immuno-electron microscopic demonstration of tau filaments in tauopathies. In: *Neuropathology*. Vol 26. Neuropathology; 2006:475-483.
58. Alzheimer A. Über eigenartige Krankheitsfalle des späteren Alters. *Z. Gesamte Neurol. Psychiatr*. 1911. 4:356–85.
59. Josephs KA, Hodges JR, Snowden JS, et al. Neuropathological background of phenotypical variability in frontotemporal dementia. *Acta Neuropathol*. 2011;122(2):137-153.
60. Constantinidis J. Pick Dementia: Anatomoclinical Correlations and Pathophysiological Considerations. In: Vol 19. Karger Publishers; 2015:72-97.
61. Choudhury P, Scharf EL, Paolini MA, et al. Pick's disease: clinicopathologic characterization of 21 cases. *J Neurol*. 2020;267(9):2697-2704.
62. Irwin DJ, Brettschneider J, McMillan CT, et al. Deep clinical and neuropathological phenotyping of Pick disease. *Ann Neurol*. 2016;79(2):272-287.
63. Piguet O, Halliday GM, Reid WGJ, et al. Clinical phenotypes in autopsy-confirmed pick disease. *Neurology*. 2011;76(3):253-259.
64. Yokota O, Tsuchiya K, Arai T, et al. Clinicopathological characterization of Pick's disease versus frontotemporal lobar degeneration with ubiquitin/TDP-43-positive inclusions. *Acta Neuropathol*. 2009;117(4):429-444.
65. Rohrer JD, Lashley T, Schott JM, et al. Clinical and neuroanatomical signatures of tissue pathology in frontotemporal lobar degeneration. *Brain*. 2011;134(9):2565-2581.

66. Koga S, Josephs KA, Aiba I, Yoshida M, Dickson DW. Neuropathology and emerging biomarkers in corticobasal syndrome. *J Neurol Neurosurg Psychiatry*. 2022;93(9):919-929.
67. Whitwell JL, Tosakulwong N, Schwarz CC, et al. Longitudinal anatomic, functional, and molecular characterization of Pick disease phenotypes. *Neurology*. 2020;95(24):e3190-e3202.
68. Dickson DW. Neuropathologic differentiation of progressive supranuclear palsy and corticobasal degeneration. *J Neurol Suppl*. 1999;246(2).
69. Hauw JJ, Daniel SE, Dickson D, et al. Preliminary NINDS neuropathologic criteria for steele-richardson-olszewski syndrome(progressive supranuclear palsy). *Neurology*. 1994;44(11):2015-2019.
70. Kovacs GG. Invited review: Neuropathology of tauopathies: Principles and practice. *Neuropathol Appl Neurobiol*. 2015;41(1):3-23.
71. Steele JC, Richardson JC, Olszewski J. Progressive Supranuclear Palsy: A Heterogeneous Degeneration Involving the Brain Stem, Basal Ganglia and Cerebellum With Vertical Gaze and Pseudobulbar Palsy, Nuchal Dystonia and Dementia. *Arch Neurol*. 1964;10(4):333-359.
72. Litvan I, Agid Y, Calne D, et al. Clinical research criteria for the diagnosis of progressive supranuclear palsy (Steele-Richardson-Olszewski syndrome): Report of the NINDS-SPSP International Workshop. *Neurology*. 1996;47(1):1-9.
73. Osaki Y, Ben-Shlomo Y, Lees AJ, et al. Accuracy of clinical diagnosis of progressive supranuclear palsy. *Mov Disord*. 2004;19(2):181-189.
74. Wenning GK, Litvan I, Tolosa E. Milestones in atypical and secondary Parkinsonisms. *Mov Disord*. 2011;26(6):1083-1095.
75. Schrag A, Ben-Shlomo Y, Quinn NP. Prevalence of progressive supranuclear palsy and multiple system atrophy: A cross-sectional study. *Lancet*. 1999;354(9192):1771-1775.
76. Swallow DMA, Zheng CS, Counsell CE. Systematic Review of Prevalence Studies of Progressive Supranuclear Palsy and Corticobasal Syndrome. *Mov Disord Clin Pract*. 2022;9(5):604-613.
77. Respondek G, Stamelou M, Kurz C, et al. The phenotypic spectrum of progressive supranuclear palsy: A retrospective multicenter study of 100 definite cases. *Mov Disord*. 2014;29(14):1758-1766.
78. Williams DR, Holton JL, Strand C, et al. Pathological tau burden and distribution distinguishes progressive supranuclear palsy-parkinsonism from Richardson's syndrome. *Brain*. 2007;130:1566-1576.
79. Williams DR, De Silva R, Paviour DC, et al. Characteristics of two distinct clinical phenotypes in pathologically proven progressive supranuclear palsy: Richardson's syndrome and PSP-parkinsonism. *Brain*. 2005;128(6):1247-1258.
80. Birdi S, Rajput AH, Fenton M, et al. Progressive supranuclear palsy diagnosis and confounding features: Report on 16 autopsied cases. *Mov Disord*. 2002;17(6):1255-1264.

81. Williams DR, Holton JL, Strand K, Revesz T, Lees AJ. Pure akinesia with gait freezing: A third clinical phenotype of progressive supranuclear palsy. *Mov Disord.* 2007;22(15):2235-2241.
82. Facheris MF, Maniak S, Scaravilli F, Schüle B, Klein C, Pramstaller PP. Pure akinesia as initial presentation of PSP: A clinicopathological study. *Park Relat Disord.* 2008;14(6):517-519.
83. Compta Y, Valldeoriola F, Tolosa E, Rey MJ, Martí MJ, Valls-Solé J. Long lasting pure freezing of gait preceding progressive supranuclear palsy: A clinicopathological study. *Mov Disord.* 2007;22(13):1954-1958.
84. Kurz C, Ebersbach G, Respondek G, Giese A, Arzberger T, Höglinger GU. An autopsy-confirmed case of progressive supranuclear palsy with predominant postural instability. *Acta Neuropathol Commun.* 2016;4(1):120.
85. Kaat LD, Boon AJW, Kamphorst W, Ravid R, Duivenvoorden HJ, Van Swieten JC. Frontal presentation in progressive supranuclear palsy. *Neurology.* 2007;69(8):723-729.
86. Hassan A, Parisi JE, Josephs KA. Autopsy-proven progressive supranuclear palsy presenting as behavioral variant frontotemporal dementia. *Neurocase.* 2012;18(6):478-488.
87. Han HJ, Kim H, Park JH, et al. Behavioral changes as the earliest clinical manifestation of progressive supranuclear palsy. *J Clin Neurol.* 2010;6(3):148-151.
88. Boeve B, Dickson D, Duffy J, Bartleson J, Trenerry M, Petersen R. Progressive nonfluent aphasia and subsequent aphasic dementia associated with atypical progressive supranuclear palsy pathology. *Eur Neurol.* 2003;49(2):72-78.
89. Mochizuki A, Ueda Y, Komatsuzaki Y, Tsuchiya K, Arai T, Shoji S. Progressive supranuclear palsy presenting with primary progressive aphasia - Clinicopathological report of an autopsy case. *Acta Neuropathol.* 2003;105(6):610-614.
90. Josephs KA, Whitwell JL, Dickson DW, et al. Voxel-based morphometry in autopsy proven PSP and CBD. *Neurobiol Aging.* 2008;29(2):280-289.
91. Cilia R, Rossi C, Frosini D, et al. Dopamine transporter spect imaging in corticobasal syndrome. *PLoS One.* 2011;6(5).
92. Josephs KA, Duffy JR, Strand EA, et al. Clinicopathological and imaging correlates of progressive aphasia and apraxia of speech. *Brain.* 2006;129:1385-1398.
93. Tsuboi Y, Josephs KA, Boeve BF, et al. Increased tau burden in the cortices of progressive supranuclear palsy presenting with corticobasal syndrome. *Mov Disord.* 2005;20(8):982-988.
94. Ling H, de Silva R, Massey LA, et al. Characteristics of progressive supranuclear palsy presenting with corticobasal syndrome: A cortical variant. *Neuropathol Appl Neurobiol.* 2014;40(2):149-163.
95. Nagao S, Yokota O, Nanba R, et al. Progressive supranuclear palsy presenting as primary lateral sclerosis but lacking parkinsonism, gaze palsy, aphasia, or dementia. *J Neurol Sci.* 2012;323(1-2):147-153.

96. Josephs KA, Katsuse O, Beccano-Kelly DA, et al. Atypical progressive supranuclear palsy with corticospinal tract degeneration. *J Neuropathol Exp Neurol*. 2006;65(4):396-405.
97. Boxer AL, Yu JT, Golbe LI, Litvan I, Lang AE, Höglinger GU. Advances in progressive supranuclear palsy: new diagnostic criteria, biomarkers, and therapeutic approaches. *Lancet Neurol*. 2017;16(7):552-563.
98. Höglinger GU, Respondek G, Stamelou M, et al. Clinical diagnosis of progressive supranuclear palsy: The movement disorder society criteria. *Mov Disord*. 2017;32(6):853-864.
99. Grimm MJ, Respondek G, Stamelou M, et al. How to apply the movement disorder society criteria for diagnosis of progressive supranuclear palsy. *Mov Disord*. 2019;34(8):1228-1232.
100. Boxer AL, Yu J-T, Golbe LI, Litvan I, Lang AE, Höglinger GU. New diagnostics and therapeutics for progressive supranuclear palsy HHS Public Access. *Lancet Neurol*. 2017;16(7):552-563.
101. Shoeibi A, Litvan I, Juncos JL, et al. Are the International Parkinson disease and Movement Disorder Society progressive supranuclear palsy (IPMDS-PSP) diagnostic criteria accurate enough to differentiate common PSP phenotypes? *Park Relat Disord*. 2019;69(April):34-39.
102. Respondek G, Grimm M, Piot I, et al. Validation of the Movement Disorder Society Criteria for the Diagnosis of 4-Repeat Tauopathies. *Mov Disord*. 2020;35(1):171-176.
103. Nath U, Ben-Shlomo Y, Thomson RG, Lees AJ, Burn DJ. Clinical features and natural history of progressive supranuclear palsy: A clinical cohort study. *Neurology*. 2003;60(6):910-916.
104. Dell'Aquila C, Zoccolella S, Cardinali V, et al. Predictors of survival in a series of clinically diagnosed progressive supranuclear palsy patients. *Park Relat Disord*. 2013;19(11):980-985.
105. Shoeibi A, Litvan I, Tolosa E, Ser T del, Lee E. Progression of two Progressive Supranuclear Palsy phenotypes with comparable initial disability. *Park Relat Disord*. 2019;66(November 2018):87-93.
106. Jellinger KA. Different Tau Pathology Pattern in Two Clinical Phenotypes of Progressive Supranuclear Palsy. *Neurodegener Dis*. 2008;5(6):339-346.
107. Guasp M, Molina-Porcel L, Painous C, et al. Association of PSP phenotypes with survival: A brain-bank study. *Parkinsonism Relat Disord*. 2021;84:77-81.
108. O'Sullivan SS, Massey LA, Williams DR, et al. Clinical outcomes of progressive supranuclear palsy and multiple system atrophy. *Brain*. 2008;131(Pt 5):1362-1372.
109. Couto B, Martinez-Valbuena I, Lee S, et al. Protracted Course-Progressive Supranuclear Palsy (PC-PSP). *Eur J Neurol*. 2022;(April):1-12.
110. Street DJ. Progression and variation of Progressive Supranuclear Palsy.. Cambridge, University of Cambridge 2022
111. Chu SA, Flagan TM, Staffaroni AM, et al. Brain volumetric deficits in MAPT mutation carriers: a multisite study. *Ann Clin Transl Neurol*. 2021;8(1):95-110.

112. Luo J, Agboola F, Grant E, et al. Sequence of Alzheimer disease biomarker changes in cognitively normal adults: A cross-sectional study. *Neurology*. 2020;95(23):e3104-e3116.
113. Rohrer JD, Nicholas JM, Cash DM, et al. Presymptomatic cognitive and neuroanatomical changes in genetic frontotemporal dementia in the Genetic Frontotemporal dementia Initiative (GENFI) study: a cross-sectional analysis. *Lancet Neurol*. 2015;14(3):253-262.
114. Venneri A, De Marco M. Reduced monoaminergic nuclei MRI signal detectable in pre-symptomatic older adults with future memory decline. *Sci Rep*. 2020;10(1):1-11.
115. Yoshida K, Hata Y, Kinoshita K, Takashima S, Tanaka K, Nishida N. Incipient progressive supranuclear palsy is more common than expected and may comprise clinicopathological subtypes: a forensic autopsy series. *Acta Neuropathol*. 2017;133(5):809-823.
116. Nogami A, Yamazaki M, Saito Y, et al. Early stage of progressive supranuclear palsy: A neuropathological study of 324 consecutive autopsy cases. *J Nippon Med Sch*. 2016;82(6):266-273.
117. Dugger BN, Hentz JG, Adler CH, et al. Clinicopathological outcomes of prospectively followed normal elderly brain bank volunteers. *J Neuropathol Exp Neurol*. 2014;73(3):244-252.
118. Street D, Whiteside D, Rittman T, Rowe JB. Prediagnostic Progressive Supranuclear Palsy – Insights from the UK Biobank. 2022.
119. Ling H, Kovacs GG, Vonsattel JPG, et al. Astroglial pathology predominates the earliest stage of corticobasal degeneration pathology. *Brain*. 2016;139(12):3237-3252.
120. Dickson DW, Bergeron C, Chin SS, et al. Office of rare diseases neuropathologic criteria for corticobasal degeneration. *J Neuropathol Exp Neurol*. 2002;61(11):935-946.
121. Litvan I, Grimes DA, Lang AE, et al. Clinical features differentiating patients with postmortem confirmed progressive supranuclear palsy and corticobasal degeneration. *J Neurol*. 1999 Sep;246 Suppl:II1-5.
122. Höglinger GU. Is it Useful to Classify Progressive Supranuclear Palsy and Corticobasal Degeneration as Different Disorders? No. *Mov Disord Clin Pract*. 2018;5(2):141-144.
123. Ling H, Macerollo A. Is it Useful to Classify PSP and CBD as Different Disorders? Yes. *Mov Disord Clin Pract*. 2018;5(2):145-148.
124. Rebeiz JJ, Kolodny EH, Richardson EP. Corticodentatonigral Degeneration With Neuronal Achromasia. *Arch Neurol*. 1968;18(1):20-33. doi:10.1001/archneur.1968.00470310034003
125. Tartaglia MC, Sidhu M, Laluz V, et al. Sporadic corticobasal syndrome due to FTLTDP. *Acta Neuropathol*. 2010;119(3):365-374.
126. Kasanuki K, Josephs KA, Ferman TJ, et al. Diffuse Lewy body disease manifesting as corticobasal syndrome A rare form of Lewy body disease. *Neurology*. 2018;91(3):E268-E279.

127. Murakami A, Koga S, Dickson DW. Asymmetrical Primary Lateral Sclerosis Presenting as Corticobasal Syndrome. *J Neuropathol Exp Neurol*. 2022;81(2):154-156.
128. Bieniek KF, Josephs KA, Lin WL, Dickson DW. Neuronal intermediate filament inclusion disease may be incorrectly classified as a subtype of FTLD-FUS. *Free Neuropathol*. 2020;1(507):1-15.
129. Lee W, Simpson M, Ling H, McLean C, Collins S, Williams DR. Characterising the uncommon corticobasal syndrome presentation of sporadic Creutzfeldt-Jakob disease. *Park Relat Disord*. 2013;19(1):81-85.
130. Koga S, Roemer SF, Kasanuki K, Dickson DW. Cerebrovascular pathology presenting as corticobasal syndrome: An autopsy case series of “vascular CBS.” *Cogn Emot J*. 2016;30(7):1289-1303.
131. Aoki N, Boyer PJ, Lund C, et al. Atypical multiple system atrophy is a new subtype of frontotemporal lobar degeneration: frontotemporal lobar degeneration associated with  $\alpha$ -synuclein. *Acta Neuropathol*. 2015;130(1):93-105. doi:10.1007/s00401-015-1442-z
132. Gibb WR, Luthert PJ, Marsden CD. Clinical and pathological features of corticobasal degeneration. *Adv Neurol*. 1990;53:51-4.
133. Litvan I, Grimes DA, Lang AE. Phenotypes and prognosis: clinicopathologic studies of corticobasal degeneration. *Adv Neurol*. 2000;82:183-96.
134. Murray R, Neumann M, Forman MS, et al. Cognitive and motor assessment in autopsy-proven corticobasal degeneration. *Neurology*. 2007;68(16):1274-1283.
135. Grimes DA, Lang AE, Bergeron CB. Dementia as the most common presentation of cortical-basal ganglionic degeneration. *Neurology*. 1999;53(9):1969-1974.
136. Bergeron C, Davis A, Lang AE. Corticobasal ganglionic degeneration and progressive supranuclear palsy presenting with cognitive decline. *Brain Pathol*. 1998;8(2):355-365.
137. Coughlin DG, Litvan I. Progressive supranuclear palsy: Advances in diagnosis and management. *Park Relat Disord*. 2020;73(2):105-116.
138. Josephs KA, Petersen RC, Knopman DS, et al. Clinicopathologic analysis of frontotemporal and corticobasal degenerations and PSP. *Neurology*. 2006;66(1):41-48.
139. Kertesz A, Martinez-Lage P, Davidson W, Munoz DG. The corticobasal degeneration syndrome overlaps progressive aphasia and frontotemporal dementia. *Neurology*. 2000;55(9):1368-1375.
140. Armstrong MJ, Litvan I, Lang AE, et al. Criteria for the diagnosis of corticobasal degeneration. *Neurology*. 2013;80(5):496-503.
141. Murray ME, Kouri N, Lin WL, Jack CR, Dickson DW, Vemuri P. Clinicopathologic assessment and imaging of tauopathies in neurodegenerative dementias. *Alzheimer's Res Ther*. 2014;6(1).
142. Koga S, Kouri N, Walton RL, et al. Corticobasal degeneration with TDP-43 pathology presenting with progressive supranuclear palsy syndrome: a distinct clinicopathologic subtype. *Acta Neuropathol*. 2018;136(3):389-404.

143. Alexander SK, Rittman T, Xuereb JH, Bak TH, Hodges JR, Rowe JB. Validation of the new consensus criteria for the diagnosis of corticobasal degeneration. *J Neurol Neurosurg Psychiatry*. 2014;85(8):925-929.
144. Fu H, Hardy J, Duff KE. Selective vulnerability in neurodegenerative diseases. *Nat Neurosci*. 2018;21(10):1350-1358.
145. Walker LC, Jucker M. Neurodegenerative Diseases: Expanding the Prion Concept. *Annu Rev Neurosci*. 2015;38:87-103.
146. Brettschneider J, Tredici K Del, Lee VM-Y, Trojanowski JQ. Spreading of pathology in neurodegenerative diseases: a focus on human studies. *Nat Rev Neurosci*. 2015;16(2):109-120.
147. Mudher A, Colin M, Dujardin S, et al. What is the evidence that tau pathology spreads through prion-like propagation? *Acta Neuropathol Commun*. 2017;5.
148. Walsh DM, Selkoe DJ. A critical appraisal of the pathogenic protein spread hypothesis of neurodegeneration. *Nat Rev Neurosci*. 2016;17(4):251-260.
149. Dickson DW, Ahmed Z, Algom AA, Tsuboi Y, Josephs KA. Neuropathology of variants of progressive supranuclear palsy. *Curr Opin Neurol*. 2010;23(4):394-400.
150. Allen M, Wang X, Serie DJ, et al. Divergent brain gene expression patterns associate with distinct cell-specific tau neuropathology traits in progressive supranuclear palsy. *Acta Neuropathol*. 2018;136(5):709-727.
151. Kovacs GG, Xie SX, Robinson JL, et al. Sequential stages and distribution patterns of aging-related tau astrogliopathy (ARTAG) in the human brain. *Acta Neuropathol Commun*. 2018;6(1):50.
152. Forrest SL, Kril JJ, Halliday GM. Cellular and regional vulnerability in frontotemporal tauopathies. *Acta Neuropathol*. 2019;138:705-727.
153. Martínez-Maldonado A, Luna-Muñoz J, Ferrer I. Incidental corticobasal degeneration. *Neuropathol Appl Neurobiol*. 2016;42(7):659-663.
154. Milenkovic I, Kovacs GG. Incidental corticobasal degeneration in a 76-yearold woman. *Clin Neuropathol*. 2013;32(1):69-72.
155. Poorkaj P, Bird TD, Wijsman E, et al. Tau is a candidate gene for chromosome 17 frontotemporal dementia. *Ann Neurol*. 1998;43(6):815-825.
156. Hutton M, Lendon CL, Rizzu P, et al. Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature*. 1998;393(6686):702-704.
157. Spillantini MG, Murrell JR, Goedert M, Farlow MR, Klug A, Ghetti B. Mutation in the tau gene in familial multiple system tauopathy with presenile dementia. *Proc Natl Acad Sci U S A*. 1998;95(13):7737-7741.
158. Raffaele F, Claudia M, John H. Genetics and molecular mechanisms of frontotemporal lobar degeneration: an update and future avenues. *Neurobiol Aging*. 2019;78:98-110.
159. Wood EM, Falcone D, Suh ER, et al. Development and validation of pedigree classification criteria for frontotemporal lobar degeneration. *JAMA Neurol*. 2013;70(11):1411-1417.



160. Forrest SL, Kril JJ, Stevens CH, et al. Retiring the term FTDP-17 as MAPT mutations are genetic forms of sporadic frontotemporal tauopathies. *Brain*. 2018;141(2):521-534.
161. Spillantini MG, Goedert M. Tau pathology and neurodegeneration. *Lancet Neurol*. 2013;12(6):609-622.
162. Moore KM, Nicholas J, Grossman M, et al. Age at symptom onset and death and disease duration in genetic frontotemporal dementia: an international retrospective cohort study. *Lancet Neurol*. 2020;19(2):145-156.
163. Giannini LAA, Ohm DT, Rozemuller AJM, et al. Isoform-specific patterns of tau burden and neuronal degeneration in MAPT-associated frontotemporal lobar degeneration. *Acta Neuropathol*. 2022;144(6):1065-1084.
164. Wallon D, Boluda S, Rovelet-Lecrux A, et al. Clinical and neuropathological diversity of tauopathy in MAPT duplication carriers. *Acta Neuropathol*. 2021;142(2):259-278.
165. Carr JS, Sirkis DW, Yokoyama JS. Genetic contributions to sporadic frontotemporal dementia. In: *Genetics, Neurology, Behavior, and Diet in Dementia*. Elsevier; 2020:71-88.
166. Van Deerlin VM, Sleiman PMA, Martinez-Lage M, et al. Common variants at 7p21 are associated with frontotemporal lobar degeneration with TDP-43 inclusions. *Nat Genet*. 2010;42(3):234-239.
167. Höglinger GU, Melhem NM, Dickson DW, et al. Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy. In: *Nature Genetics*. Vol 43. ; 2011:699-705.
168. Pottier C, Zhou X, Perkerson RB, et al. Potential genetic modifiers of disease risk and age at onset in patients with frontotemporal lobar degeneration and GRN mutations: a genome-wide association study. *Lancet Neurol*. 2018;17(6):548-558.
169. Kouri N, Ross OA, Dombroski B, et al. Genome-wide association study of corticobasal degeneration identifies risk variants shared with progressive supranuclear palsy. *Nat Commun*. 2015;6.
170. Ferrari R, Wang Y, Vandrovicova J, et al. Genetic architecture of sporadic frontotemporal dementia and overlap with Alzheimer's and Parkinson's diseases. *J Neurol Neurosurg Psychiatry*. 2017;88(2):152-164.
171. Myers AJ, Pittman AM, Zhao AS, et al. The MAPT H1c risk haplotype is associated with increased expression of tau and especially of 4 repeat containing transcripts. *Neurobiol Dis*. 2007;25(3):561-570.
172. Baker M, Litvan I, Houlden H, et al. Association of an extended haplotype in the tau gene with progressive supranuclear palsy. *Hum Mol Genet*. 1999;8(4):711-715.
173. Stefansson H, Helgason A, Thorleifsson G, et al. A common inversion under selection in Europeans. *Nat Genet*. 2005;37(2):129-137.
174. Donnelly MP, Paschou P, Grigorenko E, et al. The Distribution and Most Recent Common Ancestor of the 17q21 Inversion in Humans. *Am J Hum Genet*. 2010;86(2):161-171.

175. Bowles KR, Pugh DA, Liu Y, et al. 17q21.31 sub-haplotypes underlying H1-associated risk for Parkinson's disease are associated with LRRC37A/2 expression in astrocytes. *Mol Neurodegener.* 2022;17(1):1-21.
176. Pittman AM, Myers AJ, Abou-Sleiman P, et al. Linkage disequilibrium fine mapping and haplotype association analysis of the tau gene in progressive supranuclear palsy and corticobasal degeneration. *J Med Genet.* 2005;42(11):837-846.
177. Zody MC, Jiang Z, Fung HC, et al. Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat Genet.* 2008;40(9):1076-1083.
178. Jun G, Ibrahim-Verbaas CA, Vronskaya M, et al. A novel Alzheimer disease locus located near the gene encoding tau protein. *Mol Psychiatry.* 2016;21(1):108-117.
179. Chen JA, Chen Z, Won H, et al. Joint genome-wide association study of progressive supranuclear palsy identifies novel susceptibility loci and genetic correlation to neurodegenerative diseases. *Mol Neurodegener.* 2018;13(1):1-11.
180. Bandrés-Ciga S, Price TR, Barrero FJ, et al. Genome-wide assessment of Parkinson's disease in a Southern Spanish population. *Neurobiol Aging.* 2016;45:213.e3.
181. Desikan RS, Schork AJ, Wang Y, et al. Genetic overlap between Alzheimer's disease and Parkinson's disease at the MAPT locus. *Mol Psychiatry.* 2015;20(12):1588-1595.
182. Nalls MA, Pankratz N, Lill CM, et al. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat Genet.* 2014;46(9):989-993.
183. Soutar MPM, Melandri D, Callaghan BO, et al. Regulation of mitophagy by the NSL complex underlies genetic risk for Parkinson's disease at 16q11.2 and MAPT H1 loci. *Brain.* 2022 Dec 19;145(12):4349-4367.
184. Soto-Beasley AI, Walton RL, Valentino RR, et al. Screening non-MAPT genes of the Chr17q21 H1 haplotype in Parkinson's disease. *Park Relat Disord.* 2020;78:138-144.
185. Russ C, Lovestone S, Baker M, et al. The extended haplotype of the microtubule associated protein tau gene is not associated with Pick's disease. *Neurosci Lett.* 2001;299(1-2):156-158.
186. Morris HR, Baker M, Yasojima K, et al. Analysis of tau haplotypes in Pick's disease. *Neurology.* 2002;59(3):443-445.
187. Ghidoni R, Signorini S, Barbiero L, et al. The H2 MAPT haplotype is associated with familial frontotemporal dementia. *Neurobiol Dis.* 2006;22(2):357-362.
188. Kaivorinne AL, Krüger J, Kuivaniemi K, et al. Role of MAPT mutations and haplotype in frontotemporal lobar degeneration in Northern Finland. *BMC Neurol.*
189. Steinberg KM, Antonacci F, Sudmant PH, et al. Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nat Genet.* 2012;44(8):872-880.

190. Zollino M, Orteschi D, Murdolo M, et al. Mutations in KANSL1 cause the 17q21.31 microdeletion syndrome phenotype. *Nat Genet.* 2012;44(6):636-638.
191. Koolen DA, Kramer JM, Neveling K, et al. Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome. *Nat Genet.* 2012;44(6):639-641.
192. Di Maria E, Tabaton M, Vigo T, et al. Corticobasal degeneration shares a common genetic background with progressive supranuclear palsy. *Ann Neurol.* 2000;47(3):374-377.
193. Houlden H, Baker M, Morris HR, et al. Corticobasal degeneration and progressive supranuclear palsy share a common tau haplotype. *Neurology.* 2001;56(12):1702-1706.
194. Sanchez-Contreras MY, Kouri N, Cook CN, et al. Replication of progressive supranuclear palsy genome-wide association study identifies SLCO1A2 and DUSP10 as new susceptibility loci. *Mol Neurodegener.* 2018;13(1):37.
195. Jabbari E, Woodside J, Tan MMX, et al. Variation at the TRIM11 locus modifies progressive supranuclear palsy phenotype. *Ann Neurol.* 2018;84(4):485-496.
196. Chen L, Zhu G, Johns EM, Yang X. TRIM11 activates the proteasome and promotes overall protein degradation by regulating USP14. *Nat Commun.* 2018;9(1).
197. Myeku N, Clelland CL, Emrani S, et al. Tau-driven 26S proteasome impairment and cognitive dysfunction can be prevented early in disease by activating cAMP-PKA signaling HHS Public Access. *Nat Med.* 2016;22(1):46-53.
198. Jabbari E, Koga S, Valentino RR, et al. Genetic determinants of survival in progressive supranuclear palsy: a genome-wide association study. *Lancet Neurol.* 2021;20(2):107-116.
199. Yokoyama JS, Karch CM, Fan CC, et al. Shared genetic risk between corticobasal degeneration, progressive supranuclear palsy, and frontotemporal dementia. *Acta Neuropathol.* 2017;133(5):825-837.
200. Valentino RR, Koga S, Walton RL, et al. MAPT subhaplotypes in corticobasal degeneration: assessing associations with disease risk, severity of tau pathology, and clinical features. *Acta Neuropathol Commun.* 2020;8(1):1-9.
201. Heckman MG, Brennan RR, Labbé C, et al. Association of mapt subhaplotypes with risk of progressive supranuclear palsy and severity of tau pathology. *JAMA Neurol.* 2019;76(6):710-717.
202. Allen M, Kachadoorian M, Quicksall Z, et al. Association of MAPT haplotypes with Alzheimer's disease risk and MAPT brain gene expression levels. *Alzheimer's Res Ther.* 2014;6(4):39.
203. Zhong Q, Congdon EE, Nagaraja HN, Kuret J. Tau isoform composition influences rate and extent of filament formation. *J Biol Chem.* 2012;287(24):20711-20719.
204. Ferrari R, Hernandez DG, Nalls MA, et al. Frontotemporal dementia and its subtypes: A genome-wide association study. *Lancet Neurol.* 2014;13(7):686-699.

205. Day JW, Ranum LPW. RNA pathogenesis of the myotonic dystrophies. *Neuromuscul Disord*. 2005;15(1):5-16.
206. Maura CA. Similar brain tau pathology in DM2 / PROMM and DM1 / Steinert disease. *Neurology*. 2005;2:1-3.
207. Sergeant N, Sablonnière B, Schraen-Maschke S, et al. Dysregulation of human brain microtubule-associated tau mRNA maturation in myotonic dystrophy type 1. *Hum Mol Genet*. 2001;10(19):2143-2155.
208. Dhaenens CM, Tran H, Frandemich ML, et al. Mis-splicing of Tau exon 10 in myotonic dystrophy type 1 is reproduced by overexpression of CELF2 but not by MBNL1 silencing. *Biochim Biophys Acta - Mol Basis Dis*. 2011;1812(7):732-742.
209. Jiang H, Mankodi A, Swanson MS, Moxley RT, Thornton CA. Myotonic dystrophy type 1 is associated with nuclear foci of mutant RNA, sequestration of muscleblind proteins and deregulated alternative splicing in neurons. *Hum Mol Genet*. 2004;13(24):3079-3088.
210. Hwang JHL, Perloff OS, Gaus SE, et al. Tuberous sclerosis complex is associated with a novel human tauopathy. *Acta Neuropathol*. 2022;145(1):1-12.
211. Olney NT, Alquezar C, Ramos EM, et al. Linking tuberous sclerosis complex, excessive mTOR signaling, and age-related neurodegeneration: a new association between TSC1 mutation and frontotemporal dementia. *Acta Neuropathol*. 2017;134(5):813-816.
212. De Vries PJ, Whittemore VH, Leclezio L, et al. Tuberous Sclerosis Associated Neuropsychiatric Disorders (TAND) and the TAND checklist. *Pediatr Neurol*. 2015;52(1):25-35.
213. Feliciano DM. The Neurodevelopmental Pathogenesis of Tuberous Sclerosis Complex (TSC). *Front Neuroanat*. 2020;14(July):1-18.
214. Stamelou M, Reuss A, Pilatus U, et al. Short-term effects of coenzyme Q10 in progressive supranuclear palsy: A randomized, placebo-controlled trial. *Mov Disord*. 2008;23(7):942-949.
215. Apetauerova D, Scala SA, Hamill RW, et al. CoQ10 in progressive supranuclear palsy: A randomized, placebo-controlled, double-blind trial. *Neurol Neuroimmunol NeuroInflammation*. 2016;3(5).
216. Bensimon G, Ludolph A, Agid Y, et al. Riluzole treatment, survival and diagnostic criteria in Parkinson plus disorders: The NNIPPS Study. *Brain*. 2009;132(1):156-171.
217. Ferrer I, Barrachina M, Puig B. Glycogen synthase kinase-3 is associated with neuronal and glial hyperphosphorylated tau deposits in Alzheimer's disease, Pick's disease, progressive supranuclear palsy and corticobasal degeneration. *Acta Neuropathol*. 2002;104(6):583-591.
218. Leclair-Visonneau L, Rouaud T, Debilly B, et al. Randomized placebo-controlled trial of sodium valproate in progressive supranuclear palsy. *Clin Neurol Neurosurg*. 2016;146:35-39.
219. Tolosa E, Litvan I, Höglinger GU, et al. A phase 2 trial of the GSK-3 inhibitor tideglusib in progressive supranuclear palsy. *Mov Disord*. 2014;29(4):470-478.

220. Boxer AL, Lang AE, Grossman M, et al. Davunetide in patients with progressive supranuclear palsy: A randomised, double-blind, placebo-controlled phase 2/3 trial. *Lancet Neurol.* 2014;13(7):676-685.
221. Dam T, Boxer AL, Golbe LI, et al. Safety and efficacy of anti-tau monoclonal antibody gosuranemab in progressive supranuclear palsy: a phase 2, randomized, placebo-controlled trial. *Nat Med.* 2021;27(8):1451-1457.
222. Caouette D. Biogen Reports Top-Line Results from Phase 2 Study in Progressive Supranuclear Palsy. <https://investors.biogen.com/news-releases/news-release-details/biogen-reports-top-line-results-phase-2-study-progressive>. 2019
223. Jadhav S, Avila J, Schöll M, et al. A walk through tau therapeutic strategies. *Acta Neuropathol Commun.* 2019;7(1):22.
224. Höglinger GU, Litvan I, Mendonca N, et al. Safety and efficacy of tilavonemab in progressive supranuclear palsy: a phase 2, randomised, placebo-controlled trial. *Lancet Neurol.* 2021;20(3):182-192.
225. Courade JP, Angers R, Mairet-Coello G, et al. Epitope determines efficacy of therapeutic anti-Tau antibodies in a functional assay with human Alzheimer Tau. *Acta Neuropathol.* 2018;136(5):729-745.
226. Grossman M. Comment Lessons learned from a progressive supranuclear palsy trial. *Lancet Neurol.* 2021;20(3):162-163.
227. Pavone C, Weigand SW, Ali F, et al. Longitudinal clinical decline and baseline predictors in progressive supranuclear palsy. *Park Relat Disord.* 2023;107(January):105290.
228. Golbe LI, Ohman-Strickland PA. A clinical rating scale for progressive supranuclear palsy. *Brain.* 2007;130(6):1552-1565.
229. van Eimeren T, Antonini A, Berg D, et al. Neuroimaging biomarkers for clinical trials in atypical parkinsonian disorders: Proposal for a Neuroimaging Biomarker Utility System. *Alzheimer's Dement Diagnosis, Assess Dis Monit.* 2019;11:301-309.
230. Stamelou M, Giagkou N, Höglinger GU. One decade ago, one decade ahead in progressive supranuclear palsy. *Mov Disord.* 2019;34(9):1284-1293.
231. Meeter LH, Kaat LD, Rohrer JD, Van Swieten JC. Imaging and fluid biomarkers in frontotemporal dementia. *Nat Rev Neurol.* 2017;13(7):406-419.
232. Perry DC, Brown JA, Possin KL, et al. Clinicopathological correlations in behavioural variant frontotemporal dementia. *Brain.* 2017;140(12):3329-3345.
233. Josephs KA, Whitwell JL, Boeve BF, et al. Anatomical differences between CBS-corticobasal degeneration and CBS-Alzheimer's disease. *Mov Disord.* 2010;25(9):1246-1252.
234. Illán-Gala I, Nigro S, VandeVrede L, et al. Diagnostic Accuracy of Magnetic Resonance Imaging Measures of Brain Atrophy Across the Spectrum of Progressive Supranuclear Palsy and Corticobasal Degeneration. *JAMA Netw open.* 2022;5(4):e229588.
235. Whitwell JL, Jack CR, Boeve BF, et al. Imaging correlates of pathology in corticobasal syndrome. *Neurology.* 2010;75(21):1879-1887.

236. Carlos AF, Tosakulwong N, Weigand SD, et al. Histologic lesion type correlates of magnetic resonance imaging biomarkers in four-repeat tauopathies. *Brain Commun.* 2022;4(3).
237. Spina S, Brown JA, Deng J, et al. Neuropathological correlates of structural and functional imaging biomarkers in 4-repeat tauopathies. *Brain.* 2019;142(7):2068-2081.
238. Whitwell JL, Przybelski SA, Weigand SD, et al. Distinct anatomical subtypes of the behavioural variant of frontotemporal dementia: A cluster analysis study. *Brain.* 2009;132(11):2932-2946.
239. Whitwell JL, Jack CR, Parisi JE, et al. Rates of cerebral atrophy differ in different degenerative pathologies. *Brain.* 2007;130(4):1148-1158.
240. Whitwell JL, Höglinger GU, Antonini A, et al. Radiological biomarkers for diagnosis in PSP: Where are we and where do we need to be? *Mov Disord.* 2017;32(7):955-971.
241. Quattrone A, Morelli M, Williams DR, et al. MR parkinsonism index predicts vertical supranuclear gaze palsy in patients with PSP-parkinsonism. *Neurology.* 2016;87(12):1266-1273.
242. Massey LA, Jäger HR, Paviour DC, et al. The midbrain to pons ratio. *Neurology.* 2013;80:1856-1861.
243. Morelli M, Arabia G, Novellino F, et al. MRI measurements predict PSP in unclassifiable parkinsonisms: A cohort study. *Neurology.* 2011;77(11):1042-1047.
244. Massey LA, Micallef C, Paviour DC, et al. Conventional magnetic resonance imaging in confirmed progressive supranuclear palsy and multiple system atrophy. *Mov Disord.* 2012;27(14):1754-1762.
245. Whitwell JL, Jack CR, Parisi JE, et al. Midbrain atrophy is not a biomarker of progressive supranuclear palsy pathology. *Eur J Neurol.* 2013;20(10):1417-1422.
246. Huppertz HJ, Möller L, Südmeyer M, et al. Differentiation of neurodegenerative parkinsonian syndromes by volumetric magnetic resonance imaging analysis and support vector machine classification. *Mov Disord.* 2016;31(10):1506-1517.
247. Möller L, Kassubek J, Südmeyer M, et al. Manual MRI morphometry in Parkinsonian syndromes. *Mov Disord.* 2017;32(5):778-782.
248. Oba H, Yagishita A, Terada H, et al. New and reliable MRI diagnosis for progressive supranuclear palsy. *Neurology.* 2005;64(12):2050-2055.
249. Whitwell JL, Tosakulwong N, Botha H, et al. Brain volume and flortaucipir analysis of progressive supranuclear palsy clinical variants. *NeuroImage Clin.* 2020;25.
250. Grijalva RM, Pham NTT, Huang Q, et al. Brainstem Biomarkers of Clinical Variant and Pathology in Progressive Supranuclear Palsy. *Mov Disord.* 2021:1-12.
251. Wang YT, Edison P. Tau Imaging in Neurodegenerative Diseases Using Positron Emission Tomography. *Curr Neurol Neurosci Rep.* 2019;19(7).

252. Whitwell JL. Tau Imaging in Parkinsonism: What Have We Learned So Far? *Mov Disord Clin Pract*. 2018;5(2):118-130.
253. Lemoine L, Leuzy A, Chiotis K, Rodriguez-Vieitez E, Nordberg A. Tau positron emission tomography imaging in tauopathies: The added hurdle of off-target binding. *Alzheimer's Dement Diagnosis, Assess Dis Monit*. 2018;10:232-236.
254. Leuzy A, Chiotis K, Lemoine L, et al. Tau PET imaging in neurodegenerative tauopathies—still a challenge. *Mol Psychiatry*. 2019;24(8):1112-1134.
255. Tagai K, Ono M, Kubota M, et al. High-Contrast In Vivo Imaging of Tau Pathologies in Alzheimer's and Non-Alzheimer's Disease Tauopathies. *Neuron*. 2021;109(1):42-58.e8.
256. Joie R La, Visani A V., Baker SL, et al. Prospective longitudinal atrophy in Alzheimer's disease correlates with the intensity and topography of baseline tau-PET. *Sci Transl Med*. 2020;12(524):5732.
257. Whitwell JL, Tosakulwong N, Schwarz CG, et al. MRI Outperforms [18F]AV-1451 PET as a Longitudinal Biomarker in Progressive Supranuclear Palsy. *Mov Disord*. 2019;34(1):105-113.
258. Fung CW, Guo J, Fu H, Figueroa HY, Konofagou EE, Duff KE. Atrophy associated with tau pathology precedes overt cell death in a mouse model of progressive tauopathy. *Sci Adv*. 2020;6(42):8098-8114.
259. Dutt S, Binney RJ, Heuer HW, et al. Progression of brain atrophy in PSP and CBS over 6 months and 1 year. *Neurology*. 2016;87(19):2016-2025.
260. Josephs KA, Xia R, Mandrekar J, et al. Modeling trajectories of regional volume loss in progressive supranuclear palsy. *Mov Disord*. 2013;28(8):1117-1124.
261. Josephs KA, Whitwell JL, Boeve BF, et al. Rates of cerebral atrophy in autopsy-confirmed progressive supranuclear palsy. *Ann Neurol*. 2006;59(1):200-203.
262. Whitwell JL, Xu J, Mandrekar JN, Gunter JL, Jack CR, Josephs KA. Rates of brain atrophy and clinical decline over 6 and 12-month intervals in PSP: Determining sample size for treatment trials. *Park Relat Disord*. 2012;18(3):252-256.
263. Zhang Y, Walter R, Ng P, et al. Progression of microstructural degeneration in progressive supranuclear palsy and corticobasal syndrome: A longitudinal diffusion tensor imaging study. *PLoS One*. 2016;11(6):1-13.
264. Paviour DC, Price SL, Lees AJ, Fox NC. MRI derived brain atrophy in PSP and MSA-P: Determining sample size to detect treatment effects. *J Neurol*. 2007;254(4):478-481.
265. Höglinger GU, Schöpe J, Stamelou M, et al. Longitudinal magnetic resonance imaging in progressive supranuclear palsy: A new combined score for clinical trials. *Mov Disord*. 2017;32(6):842-852.
266. Jack CR, Knopman DS, Jagust WJ, et al. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol*. 2010;9(1):119-128.
267. Wijeratne PA, Young AL, Oxtoby NP, et al. An image-based model of brain volume biomarker changes in Huntington's disease. *Ann Clin Transl Neurol*.

2018;5(5):570-582.

268. Nettiksimmons J, Beckett L, Schwarz C, Carmichael O, Fletcher E, DeCarli C. Subgroup of ADNI normal controls characterized by atrophy and cognitive decline associated with vascular damage. *Psychol Aging*. 2013;28(1):191-201.
269. Nettiksimmons J, Harvey D, Brewer J, et al. Subtypes based on cerebrospinal fluid and magnetic resonance imaging markers in normal elderly predict cognitive decline. *Neurobiol Aging*. 2010;31(8):1419-1428.
270. Noh Y, Jeon S, Lee JM, et al. Anatomical heterogeneity of Alzheimer disease Based on cortical thickness on MRIs. *Neurology*. 2014;83(21):1936-1944.
271. Racine AM, Kosciak RL, Berman SE, et al. Biomarker clusters are differentially associated with longitudinal cognitive decline in late midlife. *Brain*. 2016;139(8):2261-2274.
272. Zhang X, Mormino EC, Sun N, Sperling RA, Sabuncu MR, Yeo BTT. Bayesian model reveals latent atrophy factors with dissociable cognitive trajectories in Alzheimer's disease. *Proc Natl Acad Sci U S A*. 2016;113(42):E6535-E6544.
273. Rohrer JD, Geser F, Zhou J, et al. TDP-43 subtypes are associated with distinct atrophy patterns in frontotemporal dementia. *Neurology*. 2010;75(24):2204-2211.
274. Whitwell JL, Jack CR, Parisi JE, et al. Does TDP-43 type confer a distinct pattern of atrophy in frontotemporal lobar degeneration? *Neurology*. 2010;75(24):2212-2220.
275. Whitwell JL, Dickson DW, Murray ME, et al. Neuroimaging correlates of pathologically defined subtypes of Alzheimer's disease: A case-control study. *Lancet Neurol*. 2012;11(10):868-877.
276. Aisen PS, Petersen RC, Donohue MC, et al. Clinical core of the Alzheimer's disease neuroimaging initiative: Progress and plans. *Alzheimer's Dement*. 2010;6(3):239-246.
277. Jack CR, Knopman DS, Jagust WJ, et al. Tracking pathophysiological processes in Alzheimer's disease: An updated hypothetical model of dynamic biomarkers. *Lancet Neurol*. 2013;12(2):207-216.
278. Oxtoby NP. Data-Driven Disease Progression Modelling. 2022:1-24. <http://arxiv.org/abs/2211.05786>.
279. Oxtoby NP, Alexander DC. Imaging plus X: Multimodal models of neurodegenerative disease. *Curr Opin Neurol*. 2017;30(4):371-379.
280. Young AL, Oxtoby NP, Daga P, et al. A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain*. 2014;137(9):2564-2577.
281. Fonteijn HM, Modat M, Clarkson MJ, et al. An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. *Neuroimage*. 2012;60(3):1880-1889.
282. O'Connor A, Weston PSJ, Pavisic IM, et al. Quantitative detection and staging of presymptomatic cognitive decline in familial Alzheimer's disease: A retrospective cohort analysis. *Alzheimer's Res Ther*. 2020;12(1):1-9.



283. Oxtoby NP, Young AL, Cash DM, et al. Data-driven models of dominantly-inherited Alzheimer's disease progression. Supplementary material. *Brain*. 2018;141(5):1529-1544.
284. Firth NC, Primativo S, Brotherhood E, et al. Sequences of cognitive decline in typical Alzheimer's disease and posterior cortical atrophy estimated using a novel event-based model of disease progression. *Alzheimer's Dement*. 2020;16(7):965-973.
285. Eshaghi A, Marinescu R V., Young AL, et al. Progression of regional grey matter atrophy in multiple sclerosis. *Brain*. 2018;141(6):1665-1677.
286. Oxtoby NP, Leyland L-A, Aksman LM, et al. Sequence of clinical and neurodegeneration events in Parkinson's disease progression. *Brain*. February 2021.
287. Gabel MC, Broad RJ, Young AL, et al. Evolution of white matter damage in amyotrophic lateral sclerosis. *Ann Clin Transl Neurol*. 2020;7(5):722-732.
288. Oxtoby NP, Shand C, Cash DM, Alexander DC, Barkhof F. Targeted Screening for Alzheimer's Disease Clinical Trials Using Data-Driven Disease Progression Models. *Front Artif Intell*. 2022;5(May):1-9.
289. Young AL, Bocchetta M, Cash DM, et al. Subtype and stage inference identifies distinct atrophy patterns in genetic frontotemporal dementia that MAP onto specific MAPT mutations . *Alzheimer's Dement*. 2020;16(S1):e042996.
290. Young AL, Marinescu R V., Oxtoby NP, et al. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nat Commun*. 2018;9(1).
291. Vogel JW, Young AL, Oxtoby NP, et al. Four distinct trajectories of tau deposition identified in Alzheimer's disease. *Nat Med*. April 2021:1-11.
292. Collij LE, Salvadó G, Wottschel V, et al. Spatial-Temporal Patterns of Amyloid- $\beta$  Accumulation: A Subtype and Stage Inference Model Analysis. 2022.
293. Young AL, Bocchetta M, Russell LL, et al. Characterizing the Clinical Features and Atrophy Patterns of MAPT-Related Frontotemporal Dementia With Disease Progression Modeling. 2021.
294. Eshaghi A, Young AL, Wijeratne PA, et al. Identifying multiple sclerosis subtypes using unsupervised machine learning and MRI data. *Nat Commun*. 2021;12(1):1-12.
295. Lynch DA, Hawkes DJ, Alexander DC, Hurst JR, Investigators C. Disease Progression Modeling in Chronic Obstructive Pulmonary Disease. 2020;201(3):294-302.
296. Vogel JW, Young AL, Oxtoby NP et al. Four distinct trajectories of tau deposition identified in Alzheimer's disease. *Nat Med*. 2021 May;27(5):871-881.
297. VandeVrede L, Dale ML, Fields S, et al. Open-Label Phase 1 Futility Studies of Salsalate and Young Plasma in Progressive Supranuclear Palsy. *Mov Disord Clin Pract*. 2020;7(4):440-447.

298. Jabbari E, Holland N, Chelban V, et al. Diagnosis Across the Spectrum of Progressive Supranuclear Palsy and Corticobasal Syndrome. *JAMA Neurol.* 2020;77(3):377-387.
299. Jabbari E, Woodside J, Tan MMX, et al. The genetic and clinico-pathological profile of early-onset progressive supranuclear palsy. *Mov Disord.* July 2019;mds.27786.
300. Goetz CG, Tilley BC, Shaftman SR, et al. Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Mov Disord.* 2008;23(15):2129-2170.
301. Schwab, R.S. and England, A.C. (1969) Projection Technique for Evaluating Surgery in Parkinson's Disease. In: Billingham, F.H. and Donaldson, M.C., Eds., Third Symposium on Parkinson's Disease, Churchill Livingstone, Edinburgh, 152-157.
302. Nasreddine ZS, Phillips NA, BÃ©dirian V, et al. The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. *J Am Geriatr Soc.* 2005;53(4):695-699.
303. Mioshi E, Dawson K, Mitchell J, Arnold R, Hodges JR. The Addenbrooke's Cognitive Examination revised (ACE-R): A brief cognitive test battery for dementia screening. *Int J Geriatr Psychiatry.* 2006;21(11):1078-1085.
304. Cardoso MJ, Modat M, Wolz R, et al. Geodesic Information Flows: Spatially-Variant Graphs and Their Application to Segmentation and Fusion. *IEEE Trans Med Imaging.* 2015;34(9):1976-1988.
305. Diedrichsen J, Balsters JH, Flavell J, Cussans E, Ramnani N. A probabilistic MR atlas of the human cerebellum. *Neuroimage.* 2009;46(1):39-46.
306. Iglesias JE, Van Leemput K, Bhatt P, et al. Bayesian segmentation of brainstem structures in MRI. *Neuroimage.* 2015;113:184-195.
307. Efron B, Gong G. A Leisurely Look at the Bootstrap , the Jackknife , and Cross-Validation. *The American Statistician.* 1983; 37 (1):36-48
308. Gelman A, Hwang J, Vehtari A. Understanding predictive information criteria for Bayesian models. *Stat Comput.* 2014;24(6):997-1016.
309. Bhattacharyya, A. On a Measure of Divergence between Two Statistical Populations Defined by Their Probability Distributions. *Bulletin of the Calcutta Mathematical Society.* 1943; 35: 99-109.
310. Toomey CE, Heywood W, Benson BC, Packham G, Mills K, Lashley T. Investigation of pathology, expression and proteomic profiles in human TREM2 variant postmortem brains with and without Alzheimer's disease. *Brain Pathol.* 2020;30(4):794-810.
311. Parkinson TG, Program G. GP2: The Global Parkinson's Genetics Program. *Mov Disord.* 2021;36(4):842-851.
312. Doni A, Parente R, Laface I, et al. Serum amyloid P component is an essential element of resistance against *Aspergillus fumigatus*. *Nat Commun.* 2021;12(1):1-15.

313. Guo Y, He J, Zhao S, et al. Illumina human exome genotyping array clustering and quality control. *Nat Protoc.* 2014;9(11):2643-2662.
314. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature.* 2009;461(7265):747-753.
315. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996 Sep 13;273(5281):1516-7.
316. Belmont JW, Boudreau A, Leal SM, et al. A haplotype map of the human genome. *Nature.* 2005;437(7063):1299-1320.
317. McVean G, Spencer CCA, Chaix R. Perspectives on human genetic variation from the HapMap project. *PLoS Genet.* 2005;1(4):0413-0418.
318. Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends Genet.* 2001;17(9):502-510.
319. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet.* 2001;69(1):124-137.
320. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. *PLoS Comput Biol.* 2012;8(12):e1002822.
321. Hardy J, Singleton A. Genomewide Association Studies and Human Disease. *N Engl J Med.* 2009;17360(23):1759-1768.
322. Singleton A, Hardy J. A generalizable hypothesis for the genetic architecture of disease: Pleomorphic risk loci. *Hum Mol Genet.* 2011;20(R2):158-162.
323. Reed E, Nunez S, Kulp D, Qian J, Reilly MP, Foulkes AS. A guide to genome-wide association analysis and post-analytic interrogation. *Stat Med.* 2015;34(28):3769-3792.
324. Tan MMX, Lawton MA, Jabbari E, et al. Genome-wide association studies of cognitive and motor progression in Parkinson's disease. *Mov Disord.* 2021 Feb;36(2):424-433.
325. Moss DJH, Tabrizi SJ, Mead S, et al. Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study. *Lancet Neurol.* 2017;16(9):701-711.
326. Balding DJ. A tutorial on statistical methods for population association studies. *Nat Rev Genet.* 2006;7(10):781-791.
327. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006;38(8):904-909.
328. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature.* 2015;526(7571):68-74.
329. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. *Annu Rev Genomics Hum Genet.* 2009;10:387-406. d
330. Naj AC. Genotype Imputation in Genome-Wide Association Studies. *Curr Protoc Hum Genet.* 2019;102(1).
331. Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet.* 2016;48(10):1284-1287.

332. Durbin R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics*. 2014;30(9):1266-1272.
333. Lewis CM. Genetic association studies: design, analysis and interpretation. *Brief Bioinform*. 2002;3(2):146-153.
334. Sasieni PD. From Genotypes to Genes: Doubling the Sample Size. *Biometrics*. 1997;53(4):1253.
335. Altman N, Krzywinski M. Points of Significance: Clustering. *Nat Methods*. 2017;14(6):545-546.
336. Lever J, Krzywinski M, Altman N. *Nat Methods*. 2017;14(6):641–642.
337. Altman N, Krzywinski M. Points of Significance: Simple linear regression. *Nat Methods*. 2015;12(11):999-1000.
338. Krzywinski M, Altman N. Points of Significance: Multiple linear regression. *Nat Methods*. 2015;12(12):1103-1104.
339. Lever J, Krzywinski M, Altman N. Points of Significance: Logistic regression. *Nat Methods*. 2016;13(7):541-542.
340. Walker DA, Smith TJ. Nine pseudo R2 indices for binary logistic regression models. *J Mod Appl Stat Methods*. 2016;15(1):848-854.
341. Dey T. Survival analysis—time-to-event data and censoring. *Nat Methods*. 2022;19(8):903.
342. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival Analysis Part II: Multivariate data analysis- An introduction to concepts and methods. *Br J Cancer*. 2003;89(3):431-436.
343. Watanabe K, Taskesen E, Van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun*. 2017;8(1):1-10.
344. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):1-7.
345. 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):310-315.
346. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res*. 2012;22(9):1790-1797.
347. Ernst J, Kellis M. ChromHMM: Automating chromatin-state discovery and characterization. *Nat Methods*. 2012;9(3):215-216.
348. Dunham I, Kundaje A, Aldred SF, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74.
349. Luo Y, Hitz BC, Gabdank I, et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res*. 2020;48(D1):D882-D889.
350. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*. 2014;42(D1):1001-1006.

351. Ardlie KG, DeLuca DS, Segrè A V., et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* (80- ). 2015;348(6235):648-660.
352. Westra HJ, Peters MJ, Esko T, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet.* 2013;45(10):1238-1243.
353. Zhernakova D V., Deelen P, Vermaat M, et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat Genet.* 2017;49(1):139-145.
354. Ramasamy A, Trabzuni D, Guelfi S, et al. Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nat Neurosci.* 2014;17(10):1418-1428.
355. Schmitt AD, Hu M, Jung I, et al. A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Rep.* 2016;17(8):2042-2059.
356. Giambartolomei C, Vukcevic D, Schadt EE, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* 2014;10(5).
357. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A tool for genome-wide complex trait analysis. *Am J Hum Genet.* 2011;88(1):76-82.
358. Scotton WJ, Bocchetta M, Todd E, et al. A data-driven model of brain volume changes in progressive supranuclear palsy. *Brain Commun.* 2022;4(3).
359. Franzmeier N, Hoglinger GU. Inferring the sequence of brain volume changes in progressive supranuclear palsy using MRI. *Brain Commun.* 2022:1-3.
360. Brendel M, Barthel H, Van Eimeren T, et al. Assessment of 18F-PI-2620 as a Biomarker in Progressive Supranuclear Palsy. *JAMA Neurol.* 2020;77(11):1408-1419.
361. Ossenkoppele R, Lyoo CH, Sudre CH, et al. Distinct tau PET patterns in atrophy-defined subtypes of Alzheimer's disease. *Alzheimer's Dement.* 2020;16(2):335-344.
362. Briggs M, Allinson KSJ, Malpetti M, Spillantini MG, Rowe JB, Kaalund SS. Validation of the new pathology staging system for progressive supranuclear palsy. *Acta Neuropathol.* 2021;141(5):787-789.
363. Perlaki G, Horvath R, Nagy SA, et al. Comparison of accuracy between FSL's FIRST and Freesurfer for caudate nucleus and putamen segmentation. *Sci Rep.* 2017;7(1):1-9.
364. Johnson EB, Gregory S, Johnson HJ, et al. Recommendations for the use of automated gray matter segmentation tools: Evidence from Huntington's disease. *Front Neurol.* 2017;8(OCT):519.
365. Malone IB, Leung KK, Clegg S, et al. Accurate automatic estimation of total intracranial volume: A nuisance variable with less nuisance. *Neuroimage.* 2015;104:366-372.
366. Efron, B. Bootstrap Methods: Another Look at the Jackknife. In: Kotz, S., Johnson, N.L. (eds) Breakthroughs in Statistics. Springer Series in Statistics.

- Springer, New York, NY.1992; [https://doi.org/10.1007/978-1-4612-4380-9\\_41](https://doi.org/10.1007/978-1-4612-4380-9_41)
367. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw.* 2015;67(1):1-48.
  368. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest Package: Tests in Linear Mixed Effects Models. *J Stat Softw.* 2017;82(1):1-26.
  369. La Joie R, Visani A V., Lesman-Segev OH, et al. Association of APOE4 and clinical variability in Alzheimer disease with the pattern of tau- and amyloid-PET. *Neurology.* December 2020;10.1212/WNL.0000000000011270.
  370. Cosottini M, Ceravolo R, Faggioni L, et al. Assessment of midbrain atrophy in patients with progressive supranuclear palsy with routine magnetic resonance imaging. *Acta Neurol Scand.* 2007;116(1):37-42.
  371. Bocchetta M, Iglesias JE, Chelban V, et al. Automated brainstem segmentation detects differential involvement in atypical parkinsonian syndromes. *J Mov Disord.* 2020;13(1):39-46.
  372. Sjöström H, Granberg T, Hashim F, Westman E, Svenningsson P. Automated brainstem volumetry can aid in the diagnostics of parkinsonian disorders. *Park Relat Disord.* 2020;79:18-25.
  373. Massey LA, Jager HR, Paviour DC, et al. The midbrain to pons ratio: A simple and specific MRI sign of progressive supranuclear palsy. *Neurology.* 2013;80(20):1856-1861.
  374. Paviour DC, Price SL, Stevens JM, Lees AJ, Fox NC. Quantitative MRI measurement of superior cerebellar peduncle in progressive supranuclear palsy. *Neurology.* 2005;64(4):675-679.
  375. Quattrone A, Nicoletti G, Aguglia U. MR imaging index for differentiation of progressive supranuclear palsy from Parkinson disease and the Parkinson variant of multiple system atrophy. *Radiology.* 2008;246(1):214-221.
  376. Messina D, Cerasa A, Condino F, et al. Patterns of brain atrophy in Parkinson's disease, progressive supranuclear palsy and multiple system atrophy. *Park Relat Disord.* 2011;17(3):172-176.
  377. Saini J, Bagepally BS, Sandhya M, et al. Subcortical structures in progressive supranuclear palsy: Vertex-based analysis. *Eur J Neurol.* 2013;20(3):493-501.
  378. Looi JCL, Macfarlane MD, Walterfang M, et al. Morphometric analysis of subcortical structures in progressive supranuclear palsy: In vivo evidence of neostriatal and mesencephalic atrophy. *Psychiatry Res - Neuroimaging.* 2011;194(2):163-175.
  379. Brenneis C, Seppi K, Schocke M, Benke T, Wenning GK, Poewe W. Voxel based morphometry reveals a distinct pattern of frontal atrophy in progressive supranuclear palsy. *J Neurol Neurosurg Psychiatry.* 2004;75(2):246-249.
  380. Cordato NJ, Pantelis C, Halliday GM, et al. Frontal atrophy correlates with behavioural changes in progressive supranuclear palsy. *Brain.* 2002;125(4):789-800.
  381. Josephs KA, Whitwell JL, Eggers SD, Senjem ML, Jack CR. Gray matter correlates of behavioral severity in progressive supranuclear palsy. *Mov Disord.* 2011;26(3):493-498.

382. Pyatigorskaya N, Yahia-Cherif L, Gaurav R, et al. Multimodal Magnetic Resonance Imaging Quantification of Brain Changes in Progressive Supranuclear Palsy. *Mov Disord.* 2020;35(1):161-170.
383. Padovani A, Borroni B, Brambati SM, et al. Diffusion tensor imaging and voxel based morphometry study in early progressive supranuclear palsy. *J Neurol Neurosurg Psychiatry.* 2006;77(4):457-463.
384. Wijeratne PA, Alexander DC. Learning transition times in event sequences: The temporal event-based model of disease progression. Springer; 2021:583-595.
385. Whitwell JL, Jack CR, Parisi JE, et al. Midbrain atrophy is not a biomarker of progressive supranuclear palsy pathology. *Eur J Neurol.* 2013;20(10):1417-1422.
386. Tsai RM, Lobach I, Bang J, et al. Clinical correlates of longitudinal brain atrophy in progressive supranuclear palsy. *Park Relat Disord.* 2016;28:29-35.
387. Stamelou M, Respondek G, Giagkou N, Whitwell JL, Kovacs GG, Höglinger GU. Evolving concepts in progressive supranuclear palsy and other 4-repeat tauopathies. *Nat Rev Neurol.* 2021;0123456789.
388. Ali F, Martin PR, Botha H, et al. Sensitivity and Specificity of Diagnostic Criteria for Progressive Supranuclear Palsy. *Mov Disord.* 2019;34(8):1144-1153. doi:10.1002/MDS.27619
389. Höllerhage M, Klietz M, Höglinger GU. Disease modification in Parkinsonism: obstacles and ways forward. *J Neural Transm.* 2022;129(9):1133-1153.
390. Frank A, Peikert K, Linn J, Brandt MD, Hermann A. MDS criteria for the diagnosis of progressive supranuclear palsy overemphasize Richardson syndrome. *Ann Clin Transl Neurol.* 2020;7(9):1702-1707.
391. Lawton M, Kasten M, May MT, et al. Validation of conversion between minimal state examination and montreal cognitive assessment. *Mov Disord.* 2016;31(4):593.
392. Marinescu R V, Eshaghi A, Alexander DC, Golland P. BrainPainter: A software for the visualisation of brain structures, biomarkers and associated pathological processes.
393. Piot I, Schweyer K, Respondek G, et al. The Progressive Supranuclear Palsy Clinical Deficits Scale. *Mov Disord.* 2020;35(4):650-661. doi:10.1002/mds.27964
394. Evidente VGH, Adler CH, Sabbagh MN, et al. Neuropathological findings of PSP in the elderly without clinical PSP: Possible incidental PSP? *Parkinsonism Relat Disord.* 2011;17(5):365-371.
395. Sakai K, Yamada M. Early-stage progressive supranuclear palsy with degenerative lesions confined to the subthalamic nucleus and substantia nigra. *Neuropathology.* 2011;31(1):77-81.
396. Saito Y, Kamagata K, Wijeratne PA, et al. Temporal Progression Patterns of Brain Atrophy in Corticobasal Syndrome and Progressive Supranuclear Palsy Revealed by Subtype and Stage Inference (SuStaIn). *Front Neurol.* 2022;13(February).

397. Necpál J, Borsek M, Jeleňová B. “Parkinson’s disease” on the way to progressive supranuclear palsy: a review on PSP-parkinsonism. *Neurol Sci.* 2021;42(12):4927-4936.
398. Srulijes K, Mallien G, Bauer S, et al. In vivo comparison of Richardson’s syndrome and progressive supranuclear palsy-parkinsonism. *J Neural Transm.* 2011;118(8):1191-1197.
399. Jecmenica-Lukic M, Petrovic IN, Pekmezovic T, Kostic VS. Clinical outcomes of two main variants of progressive supranuclear palsy and multiple system atrophy: A prospective natural history study. *J Neurol.* 2014;261(8):1575-1583.
400. Whitwell JL, Tosakulwong N, Botha H, et al. Brain volume and flortaucipir analysis of progressive supranuclear palsy clinical variants. *Neuroimage Clin.* 2020;25:102152.
401. Longoni G, Agosta F, Kostić VS, et al. MRI measurements of brainstem structures in patients with Richardson’s syndrome, progressive supranuclear palsy-parkinsonism, and Parkinson’s disease. *Mov Disord.* 2011;26(2):247-255.
402. Agosta F, Kostić VS, Galantucci S, et al. The in vivo distribution of brain tissue loss in Richardson’s syndrome and PSP-parkinsonism: a VBM-DARTEL study. *Eur J Neurosci.* 2010;32(4):640-647.
403. Santos-Santos MA, Mandelli ML, Binney RJ, et al. Features of patients with Nonfluent/agrammatic primary progressive aphasia with underlying corticobasal degeneration or progressive supranuclear palsy pathology. *JAMA Neurol.* 2016;73(6):733.
404. Whitwell JL, Stevens CA, Duffy JR, et al. An Evaluation of the Progressive Supranuclear Palsy Speech/Language Variant. *Mov Disord Clin Pract.* 2019;6(6):452-461.
405. Fiorenzato E, Weis L, Falup-Pecurariu C, et al. Montreal Cognitive Assessment (MoCA) and Mini-Mental State Examination (MMSE) performance in progressive supranuclear palsy and multiple system atrophy. *J Neural Transm.* 2016;123(12):1435-1442.
406. Kouri N, Murray ME, Hassan A, et al. Neuropathological features of corticobasal degeneration presenting as corticobasal syndrome or Richardson syndrome. *Brain.* 2011;134(11):3264-3275.
407. Kouri N, Whitwell JL, Josephs KA, Rademakers R, Dickson DW. Corticobasal degeneration: A pathologically distinct 4R tauopathy. *Nat Rev Neurol.* 2011;7(5):263-272.
408. Boeve BF, Maraganore DM, Parisi JE, et al. Pathologic heterogeneity in clinically diagnosed corticobasal degeneration. *Neurology.* 1999;53(4):795-795.
409. Parmera JB, Rodriguez RD, Studart Neto A, Nitrini R, Brucki SMD. Corticobasal syndrome: A diagnostic conundrum. *Dement Neuropsychol.* 2016;10(4):267-275.
410. Wadia PM, Lang AE. The many faces of corticobasal degeneration. *Park Relat Disord.* 2007;13(SUPPL. 3):336-340.



411. Marsili L, Dickson DW, Espay AJ. Globular Glial Tauopathy May be Mistaken for Corticobasal Syndrome—Pointers for the Clinician. *Mov Disord Clin Pract*. 2018;5(4):439-441.
412. Zetterberg H, Schott JM. Biomarkers for Alzheimer's disease beyond amyloid and tau. *Nat Med* 2019 252. 2019;25(2):201-203.
413. Thijssen EH, La Joie R, Strom A, et al. Plasma phosphorylated tau 217 and phosphorylated tau 181 as biomarkers in Alzheimer's disease and frontotemporal lobar degeneration: a retrospective diagnostic performance study. *Lancet Neurol*. 2021;20(9):739-752.
414. VandeVrede L, Ljubenkova PA, Rojas JC, Welch AE, Boxer AL. Four-Repeat Tauopathies: Current Management and Future Treatments. *Neurotherapeutics*. 2020;17(4):1563-1581.
415. Boeve BF, Boxer AL, Kumfor F, Pijnenburg Y, Rohrer JD. Advances and controversies in frontotemporal dementia: diagnosis, biomarkers, and therapeutic considerations. *Lancet Neurol*. 2022;21(3):258-272.
416. Kass R, Raftery A. Bayes Factors. *J Am Stat Assoc*. 1995;90(430):773-795.
417. Swallow DMA, Counsell CE. Prevalence of Progressive Supranuclear Palsy and Corticobasal Syndrome in Scotland. *Neuroepidemiology*. 2022;56(4):291-297.
418. Constantinides VC, Paraskevas GP, Paraskevas PG, Stefanis L, Kapaki E. Corticobasal degeneration and corticobasal syndrome: A review. *Clin Park Relat Disord*. 2019;1:66-71.
419. Zetterberg H, Bendlin BB. Biomarkers for Alzheimer's disease—preparing for a new era of disease-modifying therapies. *Mol Psychiatry*. 2021;26(1):296-308.
420. Pardini M, Huey ED, Spina S, et al. FDG-PET patterns associated with underlying pathology in corticobasal syndrome. *Neurology*. 2019;92(10):E1121-E1135.
421. Koga S, Ikeda A, Dickson DW. Deep learning-based model for diagnosing Alzheimer's disease and tauopathies. *Neuropathol Appl Neurobiol*. 2021;(May):1-12.
422. Knopman DS, Mastri AR, Frey WH, Sung JH, Rustan T. Dementia lacking distinctive histologic features: A common non-Alzheimer degenerative dementia. *Neurology*. 1990;40(2):251-256.
423. Brun A. Frontal lobe degeneration of non-Alzheimer type revisited. *Dementia*. 1993;4(3-4):126-131.
424. Barker WW, Luis CA, Kashuba A, et al. <00002093-200210000-00001.Pdf>. *Alzheimer Dis Assoc Disord*. 2002;16(4):203-212.
425. McKhann GM. Clinical and Pathological Diagnosis of Frontotemporal Dementia. *Arch Neurol*. 2001;58(11):1803.
426. Dickson DW, Kouri N, Murray ME, Josephs KA. Neuropathology of frontotemporal lobar degeneration-Tau (FTLD-Tau). *J Mol Neurosci*. 2011;45(3):384-389.
427. Goedert M, Spillantini MG, Jakes R, Rutherford D, Crowther RA. Multiple

- isoforms of human microtubule-associated protein tau: sequences and localization in neurofibrillary tangles of Alzheimer's disease. *Neuron*. 1989;3(4):519-526.
428. Goedert M, Wischik CM, Crowther RA, Walker JE, Klug A. Cloning and sequencing of the cDNA encoding a core protein of the paired helical filament of Alzheimer disease: Identification as the microtubule-associated protein tau. *Proc Natl Acad Sci U S A*. 1988;85(11):4051-4055.
  429. Murrell JR, Spillantini MG. Tau Gene Mutation G389R Causes a Tauopathy with Abundant Pick Body-like Inclusions and Axonal Deposits. *J Neuropathol Exp Neurol*. 1999 Dec;58(12):1207-26.
  430. Pickering-Brown SM, Baker M, Nonaka T, et al. Frontotemporal dementia with Pick-type histology associated with Q336R mutation in the tau gene. *Brain*. 2004;127(6):1415-1426.
  431. Bronner IF, Ter Meulen BC, Azmani A, et al. Hereditary Pick's disease with the G272V tau mutation shows predominant three-repeat tau pathology. *Brain*. 2005;128(11):2645-2653.
  432. Tacik P, DeTure M, Hinkle KM, et al. A novel Tau Mutation in exon 12, P. Q336H, causes hereditary pick disease. *J Neuropathol Exp Neurol*. 2015;74(11):1042-1052.
  433. Valentino RR, Heckman MG, Johnson PW, et al. Association of Mitochondrial DNA Genomic Variation With Risk of Pick Disease. *Neurology*. 2021;96(13):e1755-e1760.
  434. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet*. 2002;70(2):425-434.
  435. Xia Y, Nasif L, Giasson BI. Pathogenic MAPT mutations Q336H and Q336R have isoform-dependent differences in aggregation propensity and microtubule dysfunction. *J Neurochem*. 2021;158(2):455-466.
  436. Zempel H, Mandelkow EM. Tau missorting and spastin-induced microtubule disruption in neurodegeneration: Alzheimer Disease and Hereditary Spastic Paraplegia. *Mol Neurodegener*. 2015;10(1):1-12.
  437. de Jong S, Chepelev I, Janson E, et al. Common inversion polymorphism at 17q21.31 affects expression of multiple genes in tissue-specific manner. *BMC Genomics*. 2012;13(1):458.
  438. Myers AJ, Gibbs JR, Webster JA, et al. A survey of genetic human cortical gene expression. *Nat Genet*. 2007;39(12):1494-1499.
  439. Valenca GT, Srivastava GP, Oliveira-Filho J, et al. The role of MAPT haplotype H2 and isoform 1N/4R in parkinsonism of older adults. *PLoS One*. 2016;11(7):1-12.
  440. Caffrey TM, Joachim C, Paracchini S, Esiri MM, Wade-Martins R. Haplotype-specific expression of exon 10 at the human MAPT locus. *Hum Mol Genet*. 2006;15(24):3529-3537.
  441. Foo JN, Chew EGY, Chung SJ, et al. Identification of Risk Loci for Parkinson Disease in Asians and Comparison of Risk between Asians and Europeans: A

- Genome-Wide Association Study. *JAMA Neurol.* 2020;77(6):746-754.
442. Prikas E, Paric E, Asih PR, et al. Tau target identification reveals NSF - dependent effects on AMPA receptor trafficking and memory formation . *EMBO J.* 2022;41(18):1-24.
  443. Nalls MA, Blauwendraat C, Vallerga CL, et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson’s disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* 2019;18(12):1091-1102.
  444. 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(3):559-575. d
  445. Boughton AP, Welch RP, Flickinger M, et al. LocusZoom.js: interactive and embeddable visualization of genetic association study results. *Bioinformatics.* 2021;37(18):3017-3018.
  446. Real R, Martinez-Carrasco A, Reynolds RH, et al. Association between the LRP1B and APOE loci in the development of Parkinson’s disease dementia. *Brain.* 2023 May 2;146(5):1873-1887.
  447. Wang G, Sarkar A, Carbonetto P, Stephens M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J R Stat Soc Ser B Stat Methodol.* 2020;82(5):1273-1300.
  448. Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics.* 2016;32(10):1493-1501.
  449. Zou Y, Carbonetto P, Wang G, Stephens M. Fine-mapping from summary data with the “Sum of Single Effects” model. *PLoS Genet.* 2022;18(7):1-24. doi:10.1371/journal.pgen.1010299
  450. Andersson R, Gebhard C, Miguel-Escalada I, et al. An atlas of active enhancers across human cell types and tissues. *Nature.* 2014;507(7493):455-461.
  451. Nott A, Holtman IR, Coufal NG, et al. Brain cell type – specific enhancer – promoter interactome maps and disease-risk association. 2019 Nov 29;366(6469):1134-1139.
  452. Klein N de, Tsai EA, Vochteloo M, et al. Brain expression quantitative trait locus and network analysis reveals downstream effects and putative drivers for brain-related diseases. *Nat Genet.* 2023 Mar;55(3):377-388.
  453. Krohn L, Heilbron K, Blauwendraat C, et al. Genome-wide association study of REM sleep behavior disorder identifies polygenic risk and brain expression effects. *Nat Commun.* 2022 Dec 5;13(1):7496.
  454. Li YI, Knowles DA, Humphrey J, et al. Annotation-free quantification of RNA splicing using LeafCutter. *Nat Genet.* 2018;50(1):151-158.
  455. Uhlén M, Fagerberg L, Hallström BM, et al. Tissue-based map of the human proteome. *Science* 2015 Jan 23;347(6220):1260419
  456. Berglund L, Björling E, Oksvold P, et al. A gene-centric human protein atlas for expression profiles based on antibodies. *Mol Cell Proteomics.* 2008;7(10):2019-2027.

457. Zhang Y, Sloan SA, Clarke LE, et al. Purification and Characterization of Progenitor and Mature Human Astrocytes Reveals Transcriptional and Functional Differences with Mouse. *Neuron*. 2016;89(1):37-53.
458. Saunders A, Macosko EZ, Wysoker A, et al. Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell*. 2018;174(4):1015-1030.e16.
459. Farrell K, Kim SH, Han N, et al. Genome-wide association study and functional validation implicates JADE1 in tauopathy. *Acta Neuropathol* 2021 1431. 2021;143(1):33-53.
460. Kunkle BW, Grenier-Boley B, Sims R, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates A $\beta$ , tau, immunity and lipid processing. *Nat Genet*. 2019;51(3):414-430.
461. Zheng S, Abreu N, Levitz J, Kruse AC. Structural basis for KCTD-mediated rapid desensitization of GABAB signalling. *Nature*. 2019;567(7746):127-131.
462. Ren Y, Liu Y, Zheng S, Luo M. KCTD8 and KCTD12 Facilitate Axonal Expression of GABA B Receptors in Habenula Cholinergic Neurons. *J Neurosci*. 2022;42(9):1648-1665.
463. Wang Z V., Deng Y, Gao N, et al. Spliced X-box binding protein 1 couples the unfolded protein response to hexosamine biosynthetic pathway. *Cell*. 2014;156(6):1179-1192.
464. Ruegenberg S, Horn M, Pichlo C, Allmeroth K, Baumann U, Denzel MS. Loss of GFAT-1 feedback regulation activates the hexosamine pathway that modulates protein homeostasis. *Nat Commun*. 2020;11(1):1-16.
465. Horn M, Denzel SI, Srinivasan B, et al. Hexosamine Pathway Activation Improves Protein Homeostasis through the Integrated Stress Response. *iScience*. 2020;23(3):100887.
466. Alfaiz AA, Müller V, Boutry-Kryza N, et al. West syndrome caused by homozygous variant in the evolutionary conserved gene encoding the mitochondrial elongation factor GUF1. *Eur J Hum Genet*. 2016;24(7):1001-1008.
467. Shaik S, Pandey H, Thirumalasetti SK, Nakamura N. Characteristics and Functions of the Yip1 Domain Family (YIPF), Multi-Span Transmembrane Proteins Mainly Localized to the Golgi Apparatus. *Front Cell Dev Biol*. 2019;7(July).
468. Zhu Y, Afolabi LO, Wan X, Shim JS, Chen L. TRIM family proteins: roles in proteostasis and neurodegenerative diseases. *Open Biol*. 2022;12(8).
469. Deng YJ, Huang ZX, Zhou CJ, et al. Gene profiling involved in immature CD4+ T lymphocyte responsible for systemic lupus erythematosus. *Mol Immunol*. 2006;43(9):1497-1507. doi:10.1016/J.MOLIMM.2005.07.039
470. Forlani G, Accolla RS. Tripartite motif 22 and class II transactivator restriction factors: Unveiling their concerted action against retroviruses. *Front Immunol*. 2017;8(OCT):1-7.
471. Ozato K, Shin DM, Chang TH, Morse HC. TRIM family proteins and their emerging roles in innate immunity. *Nat Rev Immunol*. 2008;8(11):849-860.

472. Yang S, Guo X, Dong X, et al. GABAA receptor subunit gene polymorphisms predict symptom-based and developmental deficits in Chinese Han children and adolescents with autistic spectrum disorders. *Sci Rep*. 2017;7(1):1-9.
473. Wang L, Li J, Shuang M, et al. Association study and mutation sequencing of genes on chromosome 15q11-q13 identified GABRG3 as a susceptibility gene for autism in Chinese Han population. *Transl Psychiatry*. 2018;8(1).
474. Tilgen N, Zorzato F, Halliger-Keller B, et al. Identification of four novel mutations in the C-terminal membrane spanning domain of the ryanodine receptor 1: Association with central core disease and alteration of calcium homeostasis. *Hum Mol Genet*. 2001;10(25):2879-2887.
475. Liu X, Betzenhauser MJ, Reiken S, et al. Role of leaky neuronal ryanodine receptors in stress- induced cognitive dysfunction. *Cell*. 2012;150(5):1055-1067.
476. Janer A, Martin E, Muriel MP, et al. PML clastosomes prevent nuclear accumulation of mutant ataxin-7 and other polyglutamine proteins. *J Cell Biol*. 2006;174(1):65-76.
477. Guo L. A cellular system that degrades misfolded proteins and protects against neurodegeneration. *Mol Cell*. 2014.
478. Forlani G, Tosi G, Turrini F, Poli G, Vicenzi E, Accolla RS. Tripartite motif-containing protein 22 interacts with class II transactivator and orchestrates its recruitment in nuclear bodies containing TRIM19/PML and Cyclin T1. *Front Immunol*. 2017;8(MAY).
479. Kuijpers M, Yu K Lou, Teuling E, Akhmanova A, Jaarsma D, Hoogenraad CC. The ALS8 protein VAPB interacts with the ER-Golgi recycling protein YIF1A and regulates membrane delivery into dendrites. *EMBO J*. 2013;32(14):2056-2072.
480. AlMuhaizea M, AlMass R, AlHargan A, et al. Truncating mutations in YIF1B cause a progressive encephalopathy with various degrees of mixed movement disorder, microcephaly, and epilepsy. *Acta Neuropathol*. 2020;139(4):791-794.
481. Caillet-Boudin ML, Fernandez-Gomez FJ, Tran H, Dhaenens CM, Buee L, Sergeant N. Brain pathology in myotonic dystrophy: When tauopathy meets spliceopathy and RNAopathy. *Front Mol Neurosci*. 2014;6(JAN).
482. Koriath C, Lashley T, Taylor W, et al. ApoE4 lowers age at onset in patients with frontotemporal dementia and tauopathy independent of amyloid- $\beta$  copathology. *Alzheimer's Dement Diagnosis, Assess Dis Monit*. 2019;11:277-280.
483. Rizzu P, Van Swieten JC, Joosse M, et al. High prevalence of mutations in the microtubule-associated protein tau in a population study of frontotemporal dementia in the Netherlands. *Am J Hum Genet*. 1999;64(2):414-421.
484. D'Souza I, Poorkaj P, Hong M, et al. Missense and silent tau gene mutations cause frontotemporal dementia with parkinsonism-chromosome 17 type, by affecting multiple alternative RNA splicing regulatory elements. *Proc Natl Acad Sci U S A*. 1999;96(10):5598-5603.
485. D'Souza I, Schellenberg GD. Regulation of tau isoform expression and dementia. *Biochim Biophys Acta - Mol Basis Dis*. 2005;1739(2):104-115.

486. Van Swieten JC, Bronner IF, Azmani A, et al. The  $\Delta$ K280 mutation in MAP tau favors exon 10 skipping in vivo. *J Neuropathol Exp Neurol*. 2007;66(1):17-25.
487. Momeni P, Pittman A, Lashley T, et al. Clinical and pathological features of an Alzheimer's disease patient with the MAPT  $\Delta$ K280 mutation. *Neurobiol Aging*. 2009;30(3):388-393.
488. Metz M, Gassmann M, Fakler B, Schaeren-Wiemers N, Bettler B. Distribution of the auxiliary GABAB receptor subunits KCTD8, 12, 12b, and 16 in the mouse brain. *J Comp Neurol*. 2011;519(8):1435-1454. doi:10.1002/CNE.22610
489. Bocchetta M, Gordon E, Marshall CR, et al. The habenula: An under-recognised area of importance in frontotemporal dementia? *J Neurol Neurosurg Psychiatry*. 2016;87(8):910-912.
490. Hof PR, Bouras C, Perl DP, Morrison JH. Quantitative neuropathologic analysis of Pick's disease cases: cortical distribution of Pick bodies and coexistence with Alzheimer's disease. *Acta Neuropathol*. 1994;87(2):115-24.
491. Alazami AM, Patel N, Shamseldin HE, et al. Accelerating novel candidate gene discovery in neurogenetic disorders via whole-exome sequencing of prescreened multiplex consanguineous families. *Cell Rep*. 2015;10(2):148-161.
492. Faqeih EA, Almannai M, Saleh MM, AlWadei AH, Samman MM, Alkuraya FS. Phenotypic characterization of KCTD3-related developmental epileptic encephalopathy. *Clin Genet*. 2018;93(5):1081-1086.
493. Michaelson JJ, Shi Y, Gujral M, et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell*. 2012;151(7):1431-1442.
494. Weiss lauren. Association between Microdeletion and Microduplication at 16p11.2 and Autism. *N Engl J Med*. 2008 Feb 14;358(7):667-75.
495. McCarthy SE, Makarov V, Kirov G, et al. Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet*. 2009;41(11):1223-1227.
496. Mencacci NE, Rubio-Agusti I, Zdebik A, et al. A Missense Mutation in KCTD17 Causes Autosomal Dominant Myoclonus-Dystonia. *Am J Hum Genet*. 2015;96(6):938-947.
497. Graziola F, Stregapede F, Travaglini L, et al. A novel KCTD17 mutation is associated with childhood early-onset hyperkinetic movement disorder. *Park Relat Disord*. 2019;61(December 2018):4-6.
498. Marcé-Grau A, Correa M, Vanegas MI, et al. Childhood onset progressive myoclonic dystonia due to a de novo KCTD17 splicing mutation. *Park Relat Disord*. 2019;61(October 2018):7-9. doi:10.1016/j.parkreldis.2019.01.004
499. Lee MTM, Chen CH, Lee CS, et al. Genome-wide association study of bipolar i disorder in the Han Chinese population. *Mol Psychiatry*. 2011;16(5):548-556.
500. Metz KA, Teng X, Coppens I, et al. KCTD7 deficiency defines a distinct neurodegenerative disorder with a conserved autophagy-lysosome defect. *Ann Neurol*. 2018;84(5):766-780.
501. Wang Y, Cao X, Liu P, et al. KCTD7 mutations impair the trafficking of lysosomal enzymes through CLN5 accumulation to cause neuronal ceroid lipofuscinoses. *Sci Adv*. 2022;8(31):1-16.

502. Ferrer I. Neurons and their dendrites in frontotemporal dementia. *Dement Geriatr Cogn Disord*. 1999;10 Suppl 1:55-60.
503. Leuzy A, Zimmer ER, Dubois J, et al. In vivo characterization of metabotropic glutamate receptor type 5 abnormalities in behavioral variant FTD. *Brain Struct Funct*. 2016;221(3):1387-1402.
504. Murley AG, Rouse MA, Simon Jones P, et al. GABA and glutamate deficits from frontotemporal lobar degeneration are associated with disinhibition. *Brain*. 2021;143(11):3449-3462.
505. Mandell MA, Jain A, Arko-Mensah J, et al. TRIM Proteins Regulate Autophagy and Can Target Autophagic Substrates by Direct Recognition. *Dev Cell*. 2014;30(4):394-409.
506. Kimura T, Mandell M, Deretic V. Precision autophagy directed by receptor regulators - emerging examples within the TRIM family. *J Cell Sci*. 2016 Mar 1;129(5):881-91.
507. Zhang L, Afolabi LO, Wan X, Li Y, Chen L. Emerging Roles of Tripartite Motif-Containing Family Proteins (TRIMs) in Eliminating Misfolded Proteins. *Front Cell Dev Biol*. 2020;8(August):1-9.
508. Giraldo MI, Hage A, van Tol S, Rajsbaum R. TRIM Proteins in Host Defense and Viral Pathogenesis. *Curr Clin Microbiol Reports*. 2020;7(4):101-114.
509. Deng N-H, Zhou Z-X, Liu H-T, et al. TRIMs: Generalists Regulating the NLRP3 Inflammasome Signaling Pathway. *DNA Cell Biol*. 2022;41(3):262-275.
510. Mishra A, Duplaà C, Vojinovic D, et al. Gene-mapping study of extremes of cerebral small vessel disease reveals TRIM47 as a strong candidate. *Brain*. 2022;145(6):1992-2007.
511. Johnson K, De Ridder W, Töpf A, et al. Extending the clinical and mutational spectrum of TRIM32 -related myopathies in a non-Hutterite population. *J Neurol Neurosurg Psychiatry*. 2019;90(4):490-493.
512. Tissot C, Mechti N. Molecular cloning of a new interferon-induced factor that represses human immunodeficiency virus type I long terminal repeat expression. *J Biol Chem*. 1995;270(25):14891-14898.
513. Chen L, Brewer MD, Guo L, Wang R, Jiang P, Yang X. Enhanced degradation of misfolded proteins promotes tumorigenesis. *Cell Rep*. 2017 Mar 28;18(13):3143-3154.
514. Darricau M, Katsinelos T, Raschella F, et al. Tau seeds from patients induce progressive supranuclear palsy pathology and symptoms in primates. *Brain*. 2022 Nov 16;awac428.
515. Tan MMX, Lawton MA, Jabbari E, et al. Genome-Wide Association Studies of Cognitive and Motor Progression in Parkinson's Disease. *Mov Disord*. 2021;36(2):424-433.
516. Weissbrod O, Hormozdiari F, Benner C, et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat Genet*. 2020;52(12):1355-1363.
517. Vösa U, Claringbould A, Westra HJ, et al. Large-scale cis- and trans-eQTL

- analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet.* 2021;53(9):1300-1310.
518. Van Deerlin VM, Sleiman PMA, Martinez-Lage M, et al. Common variants at 7p21 are associated with fronto? temporal lobar degeneration with TDP-43 inclusions. *Nat Publ Gr.* 2010;42:42-44.
  519. Südhof TC. Neuroligins and neurexins link synaptic function to cognitive disease. *Nature.* 2008;455(7215):903-911.
  520. Craig AM, Kang Y. Neurexin-neuroligin signaling in synapse development. *Curr Opin Neurobiol.* 2007;17(1):43-52.
  521. Cao X, Tabuchi K. Functions of synapse adhesion molecules neurexin/neuroligins and neurodevelopmental disorders. *Neurosci Res.* 2017;116:3-9.
  522. Ichtchenko K, Hata Y, Nguyen T, et al. Neuroligin 1: A splice site-specific ligand for  $\beta$ -neurexins. *Cell.* 1995;81(3):435-443.
  523. Song JY, Ichtchenko K, Südhof TC, Brose N. Neuroligin 1 is a postsynaptic cell-adhesion molecule of excitatory synapses. *Proc Natl Acad Sci U S A.* 1999;96(3):1100-1105.
  524. Luo JK, Melland H, Nithianantharajah J, Gordon SL. Postsynaptic Neuroligin-1 Mediates Presynaptic Endocytosis During Neuronal Activity. *Front Mol Neurosci.* 2021;14(October):1-11.
  525. Cast TP, Boesch DJ, Smyth K, Shaw AE, Ghebria M, Chanda S. An autism-associated mutation impairs neuroligin-4 glycosylation and enhances excitatory synaptic transmission in human neurons. *J Neurosci.* 2021;41(3):392-407.
  526. Kasem E, Kurihara T, Tabuchi K. Neurexins and neuropsychiatric disorders. *Neurosci Res.* 2018;127:53-60. d
  527. Owczarek S, Bang ML, Berezin V. Neurexin-neuroligin synaptic complex regulates schizophrenia-related DISC1/Kal-7/Rac1 “signalosome.” *Neural Plast.* 2015;2015.
  528. Nakanishi M, Nomura J, Ji X, et al. Functional significance of rare neuroligin 1 variants found in autism. *PLoS Genet.* 2017;13(8):1-28.
  529. Shen H, Chen Z, Wang Y, et al. Role of Neurexin-1 $\beta$  and Neuroligin-1 in Cognitive Dysfunction after Subarachnoid Hemorrhage in Rats. *Stroke.* 2015;46(9):2607-2615.
  530. Tristán-Clavijo E, Camacho-Garcia RJ, Robles-Lanuza E, et al. A truncating mutation in Alzheimer’s disease inactivates neuroligin-1 synaptic function. *Neurobiol Aging.* 2015;36(12):3171-3175.
  531. Brito-moreira J, Lourenco M V, Oliveira MM, et al. Interaction of amyloid- (A) oligomers with neurexin 2 and neuroligin 1 mediates synapse damage and memory loss in mice. *Jounal Biol Chem.* 2017;292:7327-7337.
  532. Dinamarca MC, Di Luca M, Godoy JA, Inestrosa NC. The soluble extracellular fragment of neuroligin-1 targets A $\beta$  oligomers to the postsynaptic region of excitatory synapses. *Biochem Biophys Res Commun.* 2015;466(1):66-71.



533. Camporesi E, Lashley T, Gobom J, et al. Neuroligin-1 in brain and CSF of neurodegenerative disorders: investigation for synaptic biomarkers. *Acta Neuropathol Commun.* 2021;9(1):1-15.
534. Dufort-Gervais J, Provost C, Charbonneau L, et al. Neuroligin-1 is altered in the hippocampus of Alzheimer's disease patients and mouse models, and modulates the toxicity of amyloid-beta oligomers. *Sci Rep.* 2020;10(1):1-16.
535. Bie B, Wu J, Yang H, Xu JJ, Brown DL, Naguib M. Epigenetic suppression of neuroligin 1 underlies amyloid-induced memory deficiency. *Nat Neurosci.* 2014;17(2):223-231.
536. Kraus DM, Elliott GS, Chute H, et al. CSMD1 Is a Novel Multiple Domain Complement-Regulatory Protein Highly Expressed in the Central Nervous System and Epithelial Tissues. *J Immunol.* 2006;176(7):4419-4430.
537. Sancak Y, Peterson TR, Shaul YD, et al. The rag GTPases bind raptor and mediate amino acid signaling to mTORC1. *Science.* 2008 Jun 13;320(5882):1496-501
538. Son SM, Park SJ, Stamatakou E, Vicinanza M, Menzies FM, Rubinsztein DC. Leucine regulates autophagy via acetylation of the mTORC1 component raptor. *Nat Commun.* 2020;11(1):1-13.
539. Ververis A, Christodoulou A, Christoforou M, Kamilari C, Lederer CW, Santama N. A novel family of katanin-like 2 protein isoforms (KATNAL2), interacting with nucleotide-binding proteins Nubp1 and Nubp2, are key regulators of different MT-based processes in mammalian cells. *Cell Mol Life Sci.* 2016;73(1):163-184.
540. Kohno T, Ishii K, Hirota Y, et al. Reelin-Nrp1 interaction regulates neocortical dendrite development in a context-specific manner. *J Neurosci.* 2020;40(43):8248-8261.
541. Baum ML, Wilton DK, Muthukumar A, et al. CUB and Sushi Multiple Domains 1 (CSMD1) opposes the complement cascade in neural tissues. *bioRxiv.* 2020;1:2020.09.11.291427.
542. Ripke S, Sanders AR, Kendler KS, et al. Genome-wide association study identifies five new schizophrenia loci. *Nat Genet.* 2011;43(10):969-978.
543. Athanasiu L, Giddaluru S, Fernandes C, et al. A genetic association study of CSMD1 and CSMD2 with cognitive function. *Brain Behav Immun.* 2017;61:209-216.
544. Ruiz-Martínez J, Azcona LJ, Bergareche A, Martí-Massó JF, Paisán-Ruiz C. Whole-exome sequencing associates novel CSMD1 gene mutations with familial Parkinson disease. *Neurol Genet.* 2017;3(5):1-6.
545. Lynn NA, Martinez E, Nguyen H, Torres JZ. The Mammalian Family of Katanin Microtubule-Severing Enzymes. *Front Cell Dev Biol.* 2021;9(August):1-18.
546. Qiang L, Yu W, Andreadis A, Luo M, Baas PW. Tau protects microtubules in the axon from severing by katanin. *J Neurosci.* 2006;26(12):3120-3129.

547. Siahaan V, Krattenmacher J, Hyman AA, et al. Kinetically distinct phases of tau on microtubules regulate kinesin motors and severing enzymes. *Nat Cell Biol.* 2019;21(9):1086-1092.
548. Zheng J, Long F, Cao X, Xiong B, Li Y. Knockout of *Katnal2* Leads to Autism-like Behaviors and Developmental Delay in Zebrafish. *Int J Mol Sci.* 2022;23(15).
549. Pinhasov A, Mandel S, Torchinsky A, et al. Activity-dependent neuroprotective protein: A novel gene essential for brain formation. *Dev Brain Res.* 2003;144(1):83-90.
550. Merenlender-Wagner A, Malishkevich A, Shemer Z, et al. Autophagy has a key role in the pathophysiology of schizophrenia. *Mol Psychiatry.* 2015;20(1):126-132.
551. Dresner E, Agam G, Gozes I. Activity-dependent neuroprotective protein (ADNP) expression level is correlated with the expression of the sister protein ADNP2: Deregulation in schizophrenia. *Eur Neuropsychopharmacol.* 2011;21(5):355-361.
552. Yang MH, Yang YH, Lu CY, et al. Activity-dependent neuroprotector homeobox protein: A candidate protein identified in serum as diagnostic biomarker for Alzheimer's disease. *J Proteomics.* 2012;75(12):3617-3629.
553. Vulih-Shultzman I, Pinhasov A, Mandel S. Activity-Dependent Neuroprotective Protein Snippet NAP Reduces Tau Hyperphosphorylation and Enhances Learning in a Novel Transgenic Mouse Model. *Journal of Pharmacology and Experimental Therapeutics.* 2007, 323 (2) 438-449
554. Grigg I, Ivashko-Pachima Y, Hait TA, et al. Tauopathy in the young autistic brain: novel biomarker and therapeutic target. *Transl Psychiatry.* 2020;10(1).
555. Schirer Y, Malishkevich A, Ophir Y, Lewis J, Giladi E, Gozes I. Novel marker for the onset of frontotemporal dementia: Early increase in Activity-Dependent Neuroprotective Protein (ADNP) in the face of tau mutation. *PLoS One.* 2014;9(1).
556. Ivashko-Pachima Y, Maor-Nof M, Gozes I. NAP (davunetide) preferential interaction with dynamic 3-repeat Tau explains differential protection in selected tauopathies. *PLoS One.* 2019;14(3):1-20.
557. A SR, M G G, M C. Comparison of the neurofibrillary pathology in Alzheimer's disease and familial presenile dementia with tangles. *Acta Neuropathol.* 1996;92:42-48.
558. Seeley WW. Selective functional, regional, and neuronal vulnerability in frontotemporal dementia. *Curr Opin Neurol.* 2006;23(1):1-7.
559. Seeley WW, Carlin DA, Allman JM, et al. Early frontotemporal dementia targets neurons unique to apes and humans. *Ann Neurol.* 2006;60(6):660-667.
560. Li H, Guo R, Guan Y, Li J, Wang Y. Modulation of Trans-Synaptic Neurexin–Neurologin Interaction in Pathological Pain. *Cells.* 2022;11(12):1-15.
561. Futai K, Kim MJ, Hashikawa T, Scheiffele P, Sheng M, Hayashi Y. Retrograde modulation of presynaptic release probability through signaling mediated by PSD-95-neurologin. *Nat Neurosci.* 2007;10(2):186-195.

562. Akwa Y, Gondard E, Mann A, et al. Synaptic activity protects against AD and FTD-like pathology via autophagic-lysosomal degradation. *Mol Psychiatry*. 2018;23(6):1530-1540.
563. Varoqueaux F, Aramuni G, Rawson RL, et al. Neuroligins Determine Synapse Maturation and Function. *Neuron*. 2006;51(6):741-754.
564. Jiang M, Polepalli J, Chen LY, Zhang B, Südhof TC, Malenka RC. Conditional ablation of neuroligin-1 in CA1 pyramidal neurons blocks LTP by a cell-autonomous NMDA receptor-independent mechanism. *Mol Psychiatry*. 2017;22(3):375-383.
565. Jeong J, Pandey S, Li Y, Badger JD, Lu W, Roche KW. PSD-95 binding dynamically regulates NLGN1 trafficking and function. *Proc Natl Acad Sci U S A*. 2019;116(24):12035-12044.
566. Prikas E, Ahel H., Stefanoska K et al. Interaction between the guanylate kinase domain of PSD-95 and the proline-rich region and microtubule binding repeats 2 and 3 of tau. *Biochem Cell Biol*. 2021 Oct;99(5):606-616.
567. Park L, Hochrainer K, Hattori Y, et al. Tau induces PSD95–neuronal NOS uncoupling and neurovascular dysfunction independent of neurodegeneration. *Nat Neurosci*. 2020;23(9):1079-1089.
568. Hoover BR, Reed MN, Su J, et al. Tau Mislocalization to Dendritic Spines Mediates Synaptic Dysfunction Independently of Neurodegeneration. *Neuron*. 2010;68(6):1067-1081.
569. Hunter JW, Mullen GP, McManus JR, Heatherly JM, Duke A, Rand JB. Neuroligin-deficient mutants of *C. elegans* have sensory processing deficits and are hypersensitive to oxidative stress and mercury toxicity. *DMM Dis Model Mech*. 2010;3(5-6):366-376.
570. Andres C, Beerli R, Friedman A, et al. Acetylcholinesterase-transgenic mice display embryonic modulations in spinal cord choline acetyltransferase and neurexin I $\beta$  gene expression followed by late-onset neuromotor deterioration. *Proc Natl Acad Sci U S A*. 1997;94(15):8173-8178.
571. Xiang YY, Dong H, Yang BB, Macdonald JF, Lu WY. Interaction of Acetylcholinesterase with Neurexin-1 $\beta$  Regulates Glutamatergic Synaptic stability in Hippocampal neurons. *Mol Brain*. 2014;7(1):1-17.
572. Iwaki H, Blauwendraat C, Leonard HL, et al. Genetic risk of Parkinson disease and progression: An analysis of 13 longitudinal cohorts. *Neurol Genet*. 2019;5(4).
573. Downing SM. Reliability: On the reproducibility of assessment data. *Med Educ*. 2004;38(9):1006-1012.
574. Crossley J, Davies H, Humphris G, Jolly B. Generalisability: A key to unlock professional assessment. *Med Educ*. 2002;36(10):972-978.
575. McCrone P, Payan CAM, Knapp M, et al. The economic costs of progressive supranuclear palsy and multiple system atrophy in France, Germany and the United Kingdom. *PLoS One*. 2011;6(9):1-9.

576. Galvin JE, Howard DH, Denny SS, Dickinson S, Tatton N. The social and economic burden of frontotemporal degeneration. *Neurology*. 2017;89(20):2049-2056.
577. Lukic MJ, Respondek G, Kurz C, et al. Long-Duration Progressive Supranuclear Palsy: Clinical Course and Pathological Underpinnings. *Ann Neurol*. 2022 Oct;92(4):637-649.
578. Young AL, Vogel JW, Aksman LM, et al. Ordinal SuStaIn: Subtype and Stage Inference for Clinical Scores, Visual Ratings, and Other Ordinal Data. *Front Artif Intell*. 2021;4(August):1-13.
579. Young AL. Data-driven neuropathological staging and subtyping of TDP-43 proteinopathies. *Brain*. 2023 May 8;awad145.
580. Nalls MA, Plagnol V, Hernandez DG, et al. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: A meta-analysis of genome-wide association studies. *Lancet*. 2011;377(9766):641-649.
581. Gibbs JR, van der Brug MP, Hernandez DG, et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in Human Brain. *PLoS Genet*. 2010;6(5):29.
582. Trabzuni D, Wray S, Vandrovicova J, et al. MAPT expression and splicing is differentially regulated by brain region: Relation to genotype and implication for tauopathies. *Hum Mol Genet*. 2012;21(18):4094-4103.
583. Lai MC, Bechy AL, Denk F, et al. Haplotype-specific MAPT exon 3 expression regulated by common intronic polymorphisms associated with Parkinsonian disorders. *Mol Neurodegener*. 2017;12(1):1-16.
584. Coetzee SG, Pierce S, Brundin P, Brundin L, Hazelett DJ, Coetzee GA. Enrichment of risk SNPs in regulatory regions implicate diverse tissues in Parkinson's disease etiology. *Sci Rep*. 2016;6(April):1-11.