Homeostasis of Metastable Proteins in Alzheimer's Disease

Rishika Kundra

Centre for Misfolding Diseases

Department of Chemistry

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

St. John's College

November 2017

To my loving parents,
And my wonderful husband

“Employ the strategy of the Lotus Sutra before any other”
- Nichiren Daishonin

Declaration

This dissertation is the result of my own work and includes nothing, which is the outcome of work done in collaboration except where specifically indicated in the text. It has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

In accordance with the Department of Chemistry guidelines, this thesis does not exceed 60,000 words (approximately 39,819 words), and it contains less than 150 figures.

Rishika Kundra

November 2017

Acknowledgements

The journey towards this PhD, although requiring a lot of perseverance and rigour, has been one of the most fulfilling experiences of my life. I have been tremendously lucky in being supported by amazing colleagues, friends and family, and I would like to take this opportunity to thank all of them for making this journey truly enjoyable.

I am enormously grateful to my supervisors, Prof. Chistopher M Dobson and Prof. Michele Vendruscolo, for giving me the right direction at the very start of my doctoral studies and instilling in me the confidence needed to pursue them. They have always allowed me the freedom to pursue my scientific curiosities and helped shape my scientific acumen through countless stimulating discussions. I would like to especially thank Michele, for nurturing my scientific intellect through his continuous support. His has always encouraged me to go that extra mile and has been supportive of all my endeavors, all of which have played an enormous role in making this work a truly enjoyable and learning experience, and which I am sure, will always help me progress in my career. In Chris, I have found a true mentor. His scientific brilliance has always inspired me and each of my discussions with him has considerably broadened my scientific horizons. I am deeply indebted to him for being a source of incredible support during the toughest of times. He has played a fundamental role in shaping the researcher that I am today, by always providing ingenious guidance and motivation to pursue questions that fascinate me. I have ever been treated with genuine kindness and I am truly fortunate to have worked with him. I hope I am able to always remember the many things that I have learned from him.

I would like to thank Prof. Richard I Morimoto, for being a wonderful collaborator and guide. This work, and indeed I, have greatly benefitted from his insightful observations and adept advice over the many interactions that we have had.

I would also like to thank Dr. Prajwal Ciryam, whose work on supersaturation helped to lay the basis of the studies presented in this thesis.

Karen and Echo have always been incredibly supportive and I would like to thank them for making my life so much easier with regards to all the administrative work.

I would never have been in Cambridge without the Dr. Manmohan Singh Scholarship. I would like to thank them, and St. John's College for their generous support with the funding, which has allowed me the freedom to pursue my research. My time in Cambridge has been truly enriched by being part of the graduate life at St. John's and I would like to thank Masha and Molly for making my time in Cambridge truly enjoyable. We have progressed on our journeys towards a PhD together, and they have ever been a source of delight and solace.

I am truly fortunate to have been surrounded by amazing friends and family, and I would like to thank Bhawna di, Surbhi, Harshita and Aayushi for their love and support throughout.

My family has always been a source of constant support. Nanaji and Nanima have showered me with incredible love and encouragement to follow my dreams. Mamu, Maami and Massi have always been there for me, and their support has been particularly important in my life. Daddyji had always inspired me to pursue my dreams and his words have guided me throughout my life. I wish Daddyji and Nanaji were here as I pass new milestones.

It has been a pleasure to watch Sarthak, my younger brother, develop into an amazing young man and I wish him the absolute best for pursuing his passions.

Cambridge, and indeed St. John's, has not only provided me with an outstanding academic experience, but also a source of lifelong love, support and fun – my dear husband Kedar. He is my pillar of strength and his endless patience and belief in me have been instrumental in making this work possible.

And to my amazing parents – I would like to offer my heartfelt gratitude for their unconditional love and support. Without them, I wouldn't be here in the first place. They were the ones absolutely convinced that I should be doing a PhD and today, I would like to offer them this document.

Abstract

Alzheimer's disease (AD) is the most common cause of dementia, affecting almost 40 million people worldwide, and it is predicted that this number will rise to nearly 150 million by 2050. It results not only in enormous distress for affected individuals and carers but also a substantial economic burden on society. Although more than 100 years have passed since its discovery, no cure for AD exists, despite enormous efforts in basic and clinical research over the past few decades, due to limited understanding of its underlying mechanisms.

Neurodegenerative disorders, of which AD is an example, are highly complex disorders characterized by extensive neuronal dysfunction associated with the misfolding and aggregation of a specific set of proteins, including amyloid plaques and neurofibrillary tangles in AD. One promising avenue for progress in the field is to improve our understanding of the mechanisms by which cellular dysfunction arises from the initial protein aggregation events.

The studies described in the thesis are based on the recent finding that a large number of proteins are inherently supersaturated, being expressed at concentrations higher than their solubilities, and constituting a metastable subproteome potentially susceptible to aggregation. These studies illustrate the dependence of aggregation prone metastable proteins on the cellular degradation machineries. They also study the role of metastable proteins and their homeostasis complement in the vulnerability of various body and brain tissues to protein aggregation diseases. Using extensive sequencing data and network based systems biology approaches, they elucidate how fundamental physicochemical properties of an individual or group of proteins relate to their biological function or dysfunction.

Contents

List of Figures xvi

1. Metastability of proteins against aggregation in health and disease 1

1.1 Overview 1

1.2 Protein folding, misfolding and protein metastability against aggregation 3

1.3 The regulation of protein folding and misfolding – Protein homeostasis, and its role in neurodegenerative diseases 8

1.4 Protein misfolding disorders 12

1.5 The genetics of Alzheimer's disease 16

1.6 Protein aggregation as a widespread phenomenon 21

1.7 Diagnostic and therapeutic strategies for Alzheimer's disease 23

1.8 Towards an understanding of the molecular origins of neurodegenerative diseases .26

2. A transcriptional signature of Alzheimer's disease is associated with a metastable subproteome at risk of aggregation 29

2.1 Overview 29

2.2 Introduction 30

2.3 Results 32

2.3.1 Analysis of the transcriptional changes in aging and in AD 32

2.3.2 Proteins that aggregate in AD correspond to transcriptionally downregulated gene…… 33

2.3.3 Metastable proteins correspond to transcriptionally downregulated genes in aging and in AD 34

2.3.4 Specific protein homeostasis components correspond to genes downregulated in AD…… 37

2.3.5 Biochemical pathways enriched in metastable proteins are also enriched in proteins corresponding to genes downregulated in AD 39

2.3.6 Widespread downregulation of the metastable subproteome is not a general feature of disease 41

2.4 Discussion 43

2.5 Conclusions 45

2.6 Materials and Methods 46

2.6.1 Array normalization 46

2.6.2 Construction of the linear model 47

2.6.3 Determination of significance and magnitude values 47

2.6.4 Combination of significance and magnitude.....48
2.6.5 Calculation of basal expression levels for supersaturation scores.....49
2.6.6 Multiple hypothesis correction.....50
2.6.7 KEGG analysis.....50
2.6.8 Overlap analysis.....51
2.6.9 Transcription factor analysis.....51
2.6.10 Threshold sensitivity analysis.....52
2.6.11 Sensitivity to Gaussian noise in the supersaturation score.....52

3. Protein homeostasis of a metastable subproteome associated with Alzheimer's disease.....55

3.1 Overview.....55
3.2 Introduction.....56
3.3 Results.....58
3.3.1 Protein homeostasis of a metastable subproteome associated with Alzheimer's disease.....58
3.3.2 Coexpression analysis of the AD metastable subproteome and its associated protein homeostasis components.....59
3.3.3 Identification of an AD metastable network.....61
3.3.4 Identification of the hub genes and of their roles in the AD metastable network. 63
3.3.5 Test of module generality using a consensus network analysis with a visual cortex dataset.....64
3.3.6 Identification of a protein homeostasis complement of the AD metastable subproteome.....66
3.3.7 Relationship with genome wide association studies (GWAS).....70
3.3.8 Consensus network analysis of Alzheimer's, Parkinson's and Huntington's diseases.....72
3.4 Discussion.....73
3.4.1 Specific components of the protein homeostasis system that regulate protein aggregation.....73
3.5 Conclusions.....77
3.6 Materials and Methods.....78
3.6.1 Dataset acquisition.....78
3.6.2 Sample clustering.....78
3.6.3 Generation of a 'Weighted Gene Correlation Network'.....78

4. A Map of Protein Aggregation Homeostasis Identifies the Vulnerability of Cells sand Tissues to Alzheimer's Disease 85
4.1 Overview 85
4.2 Introduction 86
4.3 Results 87
4.3.1 The protein homeostasis response is proportional to the risk of aggregation 87
4.3.2 Brain tissues have a weaker protein aggregation homeostasis than other tissues 88
4.3.3 Vulnerable brain tissues have a weaker protein aggregation homeostasis than non-vulnerable brain tissues 89
4.3.4 Neurons have a weaker protein aggregation homeostasis than non-neuronal brain cell types 90
4.3.5 Proteins in the oxidative phosphorylation pathway have a weaker protein aggregation homeostasis in the brain than in other tissues 91
4.4 Discussion 92
4.5 Conclusions 95
4.6 Materials and Methods 95
5. Perspectives and future directions 99
5.1 Overview 99
5.2 Further enquiries into the regulation of metastable proteins..................................101
5.2.1 Characterizing the mechanism of aggregation of metastable mitochondrial membrane proteins...101
5.2.2 Experimental validation of the protein homeostasis complement of metastable proteins...102
5.2.3 Biomarker development and diagnostics 102
5.2.4 Tissue vulnerability over the course of evolution 102
5.3 Towards a gene list for Alzheimer's Disease...103
A. Additional figures on transcriptional regulation of metastable subproteome… 107
B. Additional figures on protein homeostasis of a metastable subproteome ..117
C. Additional figures on a map of protein aggregation homeostasis 129
References 126

List of Figures

Figure 1.1 Schematic illustration of the multiplicity of conformational states that can be adopted by a polypeptide chain following its biosynthesis and the possible transitions between the different states. 7

Figure 1.2 The proteostasis network (PN). 11

Figure 1.3 Progress in risk gene identification, follow-up, and relevance in Alzheimer's disease research. Identification of novel risk genes by GWAS are shown in green. Follow-up and relevance of the identified risk genes are shown in orange....19

Figure 2.1 Proteins that aggregate in AD correspond to transcriptionally downregulated genes....34

Figure 2.2 Transcriptionally regulated genes in aging and AD correspond to proteins metastable against aggregation. 36

Figure 2.3 The metastability of proteins to aggregation is correlated with the downregulation of the corresponding genes in AD. 38

Figure 2.4 Comparison between downregulated and metastable biochemical pathways and networks. 41

Figure 3.1 Protein homeostasis of a metastable subproteome associated with Alzheimer's disease. 58

Figure 3.2 Identification of the AD metastable network by a coexpression analysis of the AD metastable subproteome and its associated protein homeostasis components. 60

Figure 3.3 Identification of KEGG biochemical pathways enriched in hub genes in the AD metastable network....65

Figure 3.4 Identification of the major components of the protein homeostasis system associated with the AD metastable subproteome....68

Figure 3.5 The majority of GWAS genes are found in the AD metastable network....71

Figure 3.6 Network preservation heatmaps for AD-PD (a), AD-HD (b) and HD-PD (c). 73

Figure 3.7 Schematic representation of the main pathways involved in the protein homeostasis of the proteins metastable to aggregation in Alzheimer's disease. .75

Figure 4.1. Brain tissues have a weaker protein aggregation homeostasis than other tissues. 89
Figure 4.2. Vulnerable brain tissues have a weaker protein aggregation homeostasis than non-vulnerable brain tissues. 90
Figure 4.3. Neurons have a weaker protein aggregation homeostasis than non-neuronal brain cell types. 93
Figure 4.4. The oxidative phosphorylation pathway has a weaker protein aggregation homeostasis than other pathways. 94
Figure A.1 Differences in metastability between transcriptionally regulated proteins in AD are robust against changes in differential expression thresholds. 108
Figure A.2 Differences in metastability between transcriptionally regulated proteins in AD are robust against changes in differential expression thresholds. 109
Figure A.3 Metastability levels are correlated with average expression levels for genes down-regulated in AD. 110
Figure A.4 Metastability of proteins encoded by differentially expressed genes is elevated in AD for a range of expression values. 111
Figure A.5 Differences in metastability between transcriptionally regulated proteins in AD are robust against Gaussian noise in the supersaturation score 112
Figure A.6 Differences in metastability between transcriptionally regulated proteins in aging are robust against Gaussian noise in the supersaturation score. 113
Figure A.7 Elevated metastability of proteins encoded by differentially expressed genes in AD and aging is not dependent on oxidative phosphorylation proteins 114
Figure B.1 Module preservation of MEMPs across different datasets. 118
Figure B.2 GO enrichment analysis for the genes in the three main MEMPs 119
Figure B.3 KEGG pathways enriched in the MEMP-1, MEMP-2 and MEMP3 modules. 120
Figure B.4 Correspondence between the modules of the visual cortex dataset (GSE44771) and the consensus modules. 121
Figure B.5 Consensus eigengene networks for the dorsolateral prefrontal cortex and the visual cortex. 122
Figure B.6 Network representation of the MEMPs showing the hub genes and the main components of the protein homeostasis system linked with the AD metastable subproteome. 123

Figure B.7 Consensus eigengene networks and their differential analysis................124
Figure C.1 The protein homeostasis response is proportional to the risk of aggregation..130
Figure C.2 Histogram showing Δs (difference in slope for body tissues and brain tissues) for 1000 random sets of genes. 131

CHAPTER 1

---

1. Metastability of proteins against aggregation in health and disease

1.1 Overview

Alzheimer's disease (AD), is the most common cause of dementia, affecting almost 40 million people worldwide, a number predicted to rise to nearly 150 million by 2050 (1). This disease causes enormous distress in affected individuals and carers, and represents a substantial economic burden on our society. Although over 100 years have passed since its initial description in 1906, no cure for AD exists despite enormous efforts in basic and clinical research over the past few decades, due to limited understanding of its underlying mechanisms (1-17).

Neurodegenerative diseases, of which AD is an example, are highly complex disorders characterised by extensive neuronal dysfunction associated with the misfolding and aggregation of a specific set of proteins (3-17). A feature common to essentially all these conditions is the presence of protein deposits, including amyloid plaques (formed by the Aβ peptide) and neurofibrillary tangles (formed by the protein

tau) in AD (3-17). AD belongs to a class of protein misfolding disorders, associated with the failure of proteins to correctly maintain their native states, resulting in their transition from the soluble form to highly organised fibrillar aggregates. This event can result in cellular toxicity because of aberrant interactions of the aggregates and of a loss of function of the proteins involved (3-17).

In the majority of the cases, AD manifests itself after 65 years of age (1) (late onset AD), but in relatively few instances (2-10%), the symptoms can occur as early as 20-30 years of age (early onset AD). From a genetics standpoint, early onset AD presents itself as a consequence of rare familial mutations, while late onset AD is believed to have a sporadic nature, whose origins depend on a combination of environmental factors and a complex genetic susceptibility (18). From a molecular point of view, the presence of neurofibrillary tangles and amyloid plaques is a common feature for both these forms of AD. With the advent of next generation sequencing technologies, there have been extensive efforts to identify genetic risk factors for AD. Predominantly, mutations in three genes have been identified to be linked to early onset AD: APP (the amyloid precursor protein, from which the Aβ peptide is produced), PSEN1 and PSEN2 (two proteins making up the proteases that cleave APP) (19-26). These mutations have been shown to affect the metabolism of APP, resulting in the formation of aggregation prone forms of Aβ. However, these mutations only account for early onset AD (18). The sporadic forms of AD are much more complex and studies to establish its genetic risk factors are ongoing (18). The strongest identified risk factor for late onset AD is the ε4 allele of the apolipoprotein E gene (APOE) (27, 28). Genome wide association studies (GWAS) have also identified more than 20 genetic loci to be associated with AD (18). Despite these advances in understanding the genetics of AD, clinical trials for this disease have still failed to achieve the desired effects. Hence, there is an ever-increasing need to understand the molecular origins of AD, to design better therapeutic strategies to overcome this "21st century plague" (13).

One promising avenue for progress in the field is to improve our understanding of the mechanisms by which cellular dysfunction arises from the initial protein aggregation events. Although it was originally believed that protein misfolding and aggregation are typical of few peculiar proteins like Aβ and tau, increasing evidence suggests that

it is instead a widespread phenomenon, with hundreds of different proteins found to aggregate in stress, ageing or disease (11, 29-32).

The studies described in this thesis are based on the recent finding that a large number of proteins are inherently supersaturated (33, 34), being expressed at concentrations exceeding their solubilities in the cellular environment, and constituting a metastable subproteome potentially susceptible to aggregation. These studies employ the analysis of extensive sequencing data and network-based systems biology approaches combined with various bioinformatics techniques to obtain more comprehensive descriptions of the complex molecular origins of the disease. They also elucidate how fundamental physicochemical properties of an individual or group of proteins relate to their biological function or dysfunction.

1.2 Protein folding, misfolding and protein metastability against aggregation

Proteins are essential molecules in all forms of life. The number of proteins in human beings has been estimated to be around 100,000 (14) and most of them assume a native state in order to function. Once in their native states, proteins orchestrate the numerous biochemical functions necessary to sustain a cell. But, to do all this, most proteins need to adopt specific three-dimensional conformations. This process of conversion is called protein folding and leads the extended polypeptide chain synthesized by the ribosome to the functional form, a complex conformation consisting of various secondary structures, including α-helices and β-sheets. Protein folding is a highly complex, multistep process which brings together functional groups, which otherwise could be far away in the primary sequence, into close proximity. The final conformations are fairly flexible to allow them to perform various functions (35). This process is driven by thermodynamics and arrives at a structure that has the lowest free energy, at least at low concentrations. Although most newly synthesized polypeptide chains fold in order to be functional, a range of proteins, including α-synuclein, tau and the islet amyloid peptide, which are significant from the point of view of protein misfolding disorders, are largely

unstructured in solution and often described as intrinsically disordered (12). They can, however, fold into more-defined structures upon interaction with specific binding partners. Proteolysis of larger, natively folded structures can also give rise to intrinsically disordered species, like in the case of the Aβ peptide mentioned above (12).

Protein folding, like any other molecular reaction, is ruled by thermodynamics and kinetics (36). Thermodynamics drives the process towards the global free energy minimum. Kinetics, on the other hand, governs the timescale of the process (similar to biological timescales) (37). Pioneering work from Christian B. Anfinsen's laboratory in the early 1960s showed that denatured ribonuclease A, under native conditions of pH, could reestablish its correct fold *in vitro* and recover the single correct arrangement of disulfide bonds out of all possible combinations (38, 39). These seminal studies not only demonstrated that proteins could fold reversibly, but also showed that the information required for achieving the correctly folded native state of a protein was encoded by its amino acid sequence. These experiments also showed that the native state of the protein is the most thermodynamically stable, representing a global free energy minimum. Furthermore, in the later years of the twentieth century, Cyrus Levinthal made the argument that there were too many possible confirmations accessible to the polypeptide for it to arrive at the native state by random searching. He suggested that proteins fold via specific 'folding pathways' (40). The thermodynamic and the kinetic control of protein folding were initially seen to be mutually exclusive; *i.e.*, thermodynamics would drive the polypeptide chain to arrive at the global free energy minimum in a 'pathway-independent' mechanism, whereas, kinetics would push the folding to happen quickly in a 'pathway-dependent' mechanism (37). This problem came to be known as "Levinthal's Paradox" and led to a series of studies to look for folding pathways. However, it was soon clear that the process of protein folding does not involve a pre-determined sequence of steps but is rather a stochastic process involving the polypeptide chain traversing a 'free energy landscape' to find a global minimum, such that the resulting conformation is stable under physiological conditions. This came to be known as the "new view" (41, 42) of protein folding, as opposed to the earlier classical view (36, 37, 43-45). The new view sees protein folding as a diffusion-like process, involving an ensemble of intermediate structures traversing the free energy landscape in order to find the stable confirmation

(36, 37, 44-46). A polypeptide can traverse the accessible conformations in varying amounts of time depending upon the free energy landscape. If this has the right shape, there are only a small number of conformations that need to be sampled for the polypeptide to arrive at its native conformation (36, 37, 45, 46). The free energy landscape of any given protein is encoded by its amino-acid sequence and, hence, has been carefully selected during the course of evolution to allow for rapid and efficient folding and for the avoidance of misfolding (46). More recent studies have unraveled a more complicated view of this process, suggesting the presence of metastable structures as the final product of protein folding. During the course of folding, the polypeptide can adopt conformations corresponding to local minima in the free energy landscape. The resulting structure is only kinetically stable but not thermodynamically so (47, 48), since the polypeptide chain gets trapped in a kinetically stable local energy minima.

These results concern proteins at low concentrations, when inter-molecular interactions are rare. Dobson and colleagues recently showed that amyloid state might be the most thermodynamically stable state for many different types of proteins under cellular conditions, where their concentrations can be quite high (47, 49). They reported, for a range of different peptides and proteins, at cellular concentrations, that the free energy associated with the amyloid structure is lower than that of the respective native states. Thus, from a biological standpoint, it might be possible that the native states of the proteins have been carefully selected by evolutionary pressure to be kinetically, rather than thermodynamically, stable. Thus, these states are metastable to aggregation. The conversion of these native states to a more thermodynamically stable but less functional amyloid state might be impeded by the presence of a high kinetic energy barrier.

Although the protein molecules and their biological environment have co-evolved to maintain the proteins and peptides in their soluble states, certain circumstances can convert them into nonfunctional and potentially toxic aggregates **(Figure 1.1)**, giving rise to diseases collectively referred to as protein misfolding disorders (7, 12, 50). Both intrinsically disordered systems, such as α-synuclein and Aβ (51, 52) and globular proteins, such as β_2-microglobulin and transthyretin (41, 42) have been linked to protein aggregation diseases.

Although it was well established that the conversion of soluble peptides and proteins to aggregated states involved nucleation and growth steps, the individual steps in the process were not fully understood until recently (53-55). A significant advance in the field was made when Knowles *et al.* reported the development of experimental and analytical techniques to study the kinetics of amyloid fibrils (53). Aggregation initiates with the formation of heterogeneous oligomeric species. However, these relatively disorganized species are capable of undergoing internal reorganization to give rise to rudimentary cross-β structures. Knowles *et al.* described the possible ways in which new aggregates could be formed, *i.e.* from monomer through primary nucleation, via fragmentation of existing fibrils or through a combination of monomers and existing aggregates through secondary nucleation. The combination of theoretical approaches and experimental kinetic measurements utilized by Knowles *et al.* enables the illustration of the importance of different microscopic steps during aggregation (7, 53). This development has transformed the landscape of development of therapeutics for diseases like AD by offering rational strategies based on both qualitative and quantitative understanding of the mechanism underlying fibril formation. Aggregates can also grow without any major structural reorganization to give rise to amorphous aggregates (12, 56).

It has also become increasingly clear, over the past few years that the pathogenicity associated with misfolding disorders of the central nervous system arises mainly from the oligomeric forms generated in the process of aggregation (12, 57-60). Rather than being necessarily pathogenic, can cause harm by serving as a potential reservoir of oligomers that can be released (12, 51, 61-63) or by sequestering several proteins, like those of the protein homeostasis system (8, 12, 64), leading to widespread disruption. There is also evidence of prion like spreading of fibrils within the brain, which can serve to amplify all above scenarios (12, 65-68).

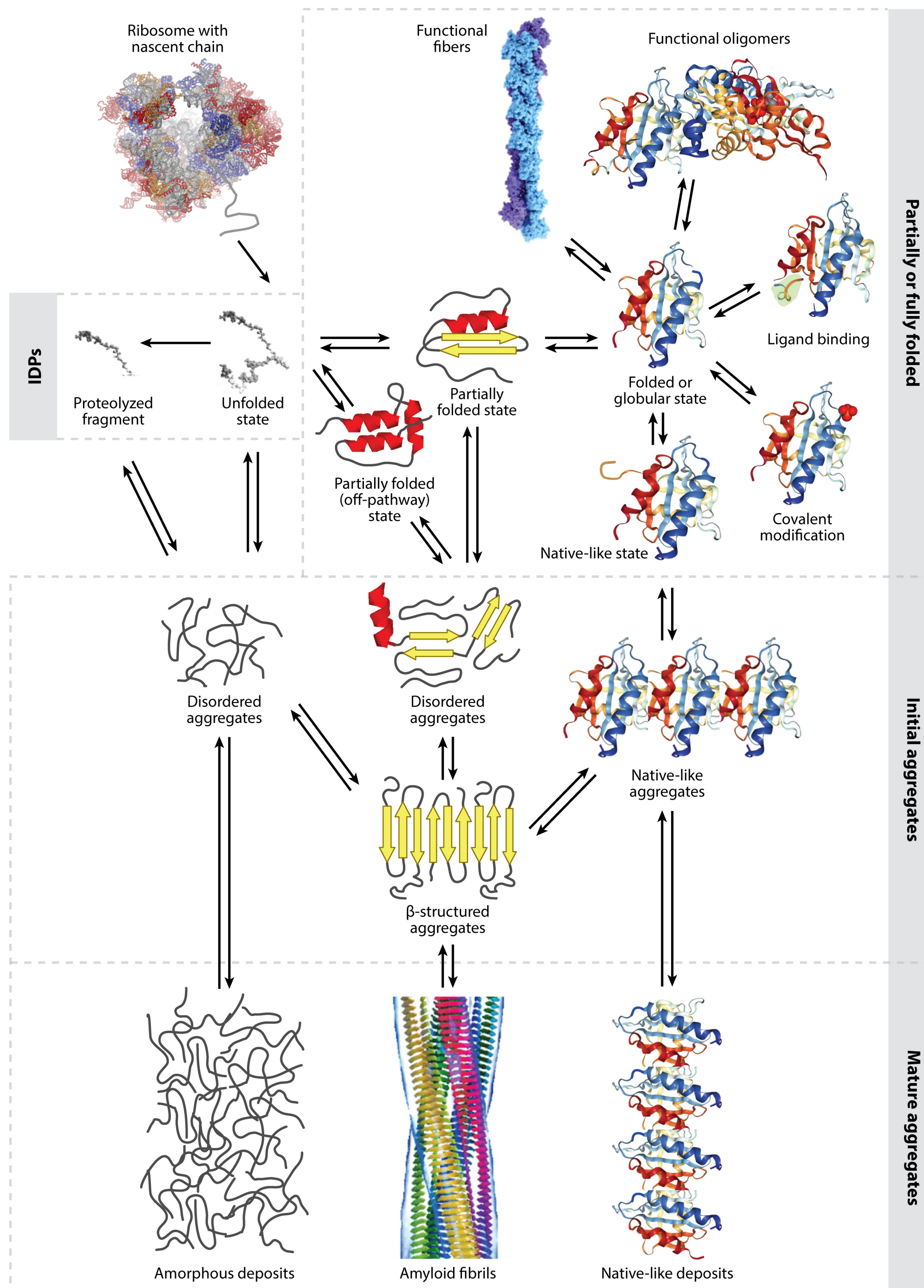


Figure 1.1 Schematic illustration of the multiplicity of conformational states that can be adopted by a polypeptide chain following its biosynthesis and the possible transitions between the different states. Boxes include intrinsically disordered proteins (IDPs), partially or fully folded proteins, initial oligomers, and mature aggregates, respectively. All of these conformational states and their interconversions are carefully regulated in the biological environment by means of the proteostasis network. Protein aggregation can result in the formation of amyloid fibrils (bottom, center), native-like deposits (bottom, right), or amorphous deposits (bottom, left), all of which are associated with pathological states when they form in an uncontrolled manner (12).

It is important to understand, in the wake of tremendously improved knowledge about the fundamentals of protein misfolding and associated pathology in various diseases, that there is unlikely to be a unique toxic agent or a unique cellular mechanism responsible for these diseases (12). It is, hence, extremely important to understand the details of the various interactions and cellular processes involved in the disease to gain an understanding of a global network, disruption in which results in the pathological cascade observed in these diseases.

1.3 The regulation of protein folding and misfolding – Protein homeostasis, and its role in neurodegenerative diseases

As described above, protein folding is a complex process involving an ensemble of intermediate states (36, 37, 44-46). Hence, there are ample opportunities for the formation of misfolded or aggregated structures leading to loss of biological functionality or even toxicity for the cell. In addition, the metastability of a protein may change during its lifecycle, starting with its synthesis at the ribosome, co-translational or post-translational folding to its localization, with post-translational modifications and with interactions with various cellular partners, and ending with its degradation (12). Even proteins that have folded into their native state can unfold through stochastic events and form aggregates (12). There is a high turnover for proteins within the cellular environment, and the native states need to be fully or partially unfolded in order to be degraded, creating another possibility for misfolding or aggregation. Since the proteome is in a constant change of flux, the cells have evolved an equally dynamic, robust protein homeostasis system to control it, comprising of several components like molecular chaperones, cellular trafficking and degradation processes **(Figure 1.2)** (5-8, 35, 69, 70). The ultimate effect of this system is to maintain the solubility of the proteome, either by preventing misfolding of native or unfolded proteins or by degradation of existing misfolded and aggregated species. It maintains a steady state that evolves in accordance with the changing conformations of the proteins. It has been seen that the free energy landscape of the protein itself is dynamic, changing in response to different substrates or molecules or

change in temperature (71-76). Therefore, the protein homeostasis system has the arduous task of maintaining the folding of tens of thousands of protein in the highly crowded cellular environment, where the system itself is in a constant state of flux.

One of the first classic examples of the workings of this system was the discovery of the heat shock response (77). It was seen that cultured cells or whole organisms, on exposure to elevated temperatures, responded by synthesis of a small number of proteins termed as heat shock proteins (Hsps). This response appears to be universal, from simple single celled organisms like bacteria to complex multicellular species like mice. This response allows the cell to deal with the stress of high temperature induced misfolding by increasing the transcription of the Hsp70 molecular chaperone (77-81). Since then, numerous experiments have demonstrated that alterations in the *in vivo* concentrations of various chaperones can modulate the cell's capacity to deal with different stresses (82-86). More than 300 molecular chaperones have been reported for humans (87) and these act at different stages of the lifecycle of a protein – from accompanying the nascent polypeptide chain as it emerges from the ribosome and its proper folding to its subsequent degradation (5, 6, 35, 64, 69, 87). Molecular chaperones, like Hsp70 and Hsp90 systems and the chaperonins (Hsp60), can transiently hide the exposed hydrophobic residues on non-native states that are usually buried in the native conformation, helping in the *de novo* folding of the polypeptide chain (35, 64). Although smaller proteins may fold rapidly *in vitro*, it is often an inefficient process for larger multi domain proteins. Also, larger proteins increase largely in number as we move from prokaryotes to eukaryotes (64). Hence, there is increasing need for a robust protein quality control system as we move up in evolution.

Another major branch of the protein quality control system is the degradation machinery (88). The cellular proteome needs to be constantly remodeled in response to varied demands on the system. Proteins need to be constantly degraded after they have fulfilled their functionality and defective or aggregated species also need to be removed to avoid cellular toxicity. The ubiquitin-proteasome system (UPS) and the autophagic-lysosomal pathway (ALP) are the two major branches of the cellular degradation machinery (88-91). Both these systems intertwine at several junctures to give rise to a highly complex regulatory system for cellular degradation, and utilize

ubiquitin (Ub) as the marker for degradation (88). It was also shown that proteasome inhibition or overexpression induces autophagy, mainly to compensate for the reduced degradative capacity of the proteasome and prevents accumulation of aberrant aggregates (88). Not only is there crosstalk between the degradation machinery, but also between molecular chaperones and the components responsible for cellular degradation. The Hsp70/Hsp90 system can target client proteins to proteasome via the co-chaperone CHIP (92). Molecular chaperones can also recognize specific sequence motifs exposed on peptides and direct them to lysosomes, where they are eventually unfolded and degraded – a process termed as chaperone-mediated autophagy (88-91, 93). Hence, the various components of the protein homeostasis system communicate extensively to maintain an intricate and elegant balance between protein solubility and degradation.

The importance of these systems for different neurodegenerative diseases has been firmly established, with studies showing the decline in protein homeostasis activity to be a central factor in ageing and in the pathology of multiple protein misfolding diseases (5-8, 35, 69, 88-91, 94-96). Also, in the model organism *C. elegans* there is evidence that decline of protein homeostasis is already a feature of early adulthood, with the suppression of the heat shock response at the onset of reproductive age occurring as a result of a genetically programmed event (97). A recent study has also reported a subset of molecular chaperones capable of preventing protein-associated toxicity in a *C. elegans* neurodegenerative disease model, but whose expression is suppressed in ageing and neurodegeneration (87). The widespread disruption of the protein homeostasis system in young healthy worms could represent an early molecular event in the ageing process that disturbs the balance between metastable proteins and the quality control system responsible to maintain their solubility, leading to the accumulation of misfolded species and the onset of neurodegenerative diseases (30, 69, 70). Mutations in two enzymes of the UPS – parkin and ubiquitin C-terminal hydrolase L1, have been shown to be directly associated with Parkinson's disease (PD) (88-91, 98-100). PD is also associated with a loss of 20/26S proteasomal

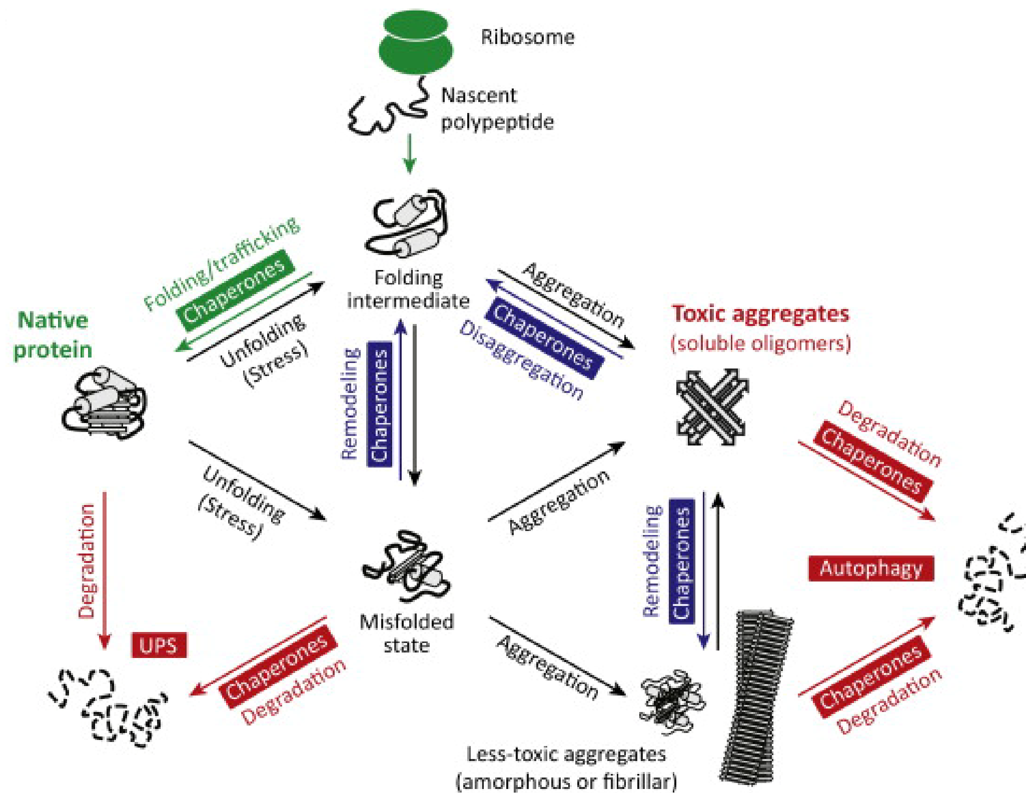


Figure 1.2 The proteostasis network (PN). The PN maintains protein homeostasis by controlling the levels of functional proteins and preventing the formation of toxic aggregates. This is achieved by integrating three branches of the PN: (i) protein synthesis, the chaperone pathways for the folding of newly synthesized proteins and intracellular trafficking (PN branch of biogenesis; green); (ii) the chaperone pathways for the remodelling of misfolded proteins and protein disaggregation (PN branch of conformational maintenance; blue); and (iii) the pathways of protein degradation by the ubiquitin–proteasome system (UPS) and autophagy (PN branch of degradation; red). Toxic aggregates (mainly diffusible, oligomeric states) may be converted to less toxic, insoluble inclusions of amorphous or fibrillar (amyloid-like) structure (6).

subunits and an overall decrease in proteasomal activity in the affected regions of the brain (91, 100, 101). Exogenous expression of aggregating peptides was also shown to severely impair UPS activity in human cell lines (102). The pathology of AD too has been linked to a decrease in abundance or efficiency of components of the cellular degradation machineries. There have also been reports of region specific depletion in proteasomal activity in post-mortem tissues of AD patients, with highly vulnerable brain regions showing a decrease in the activity, but not expression, of the proteasome (103). However, the details of whether this decline in the quality control components precedes, or is a result of, protein misfolding and aggregation events remain elusive.

1.4 Protein misfolding disorders

Given the delicate nature of protein folding and the many sources of conformational instability of the native states of proteins, one would expect a wide array of diseases whose pathology emerges due to protein misfolding. More than 50 human diseases have indeed been associated with protein misfolding and subsequent deposition of aberrant amyloid deposits, spanning across various tissues and organs such as liver, spleen, kidney and muscles among others giving rise to conditions such as type 2 diabetes and atrial amyloidosis (11, 12, 46). Several systemic amyloidosis have also been reported involving the deposition of aggregates across various tissues. Perhaps the most prominent of all these conformational disorders are those in which the aggregates are deposited in the central nervous system, giving rise to neurodegenerative disorders like AD, HD, PD and ALS (11, 12). Their pathology is a result of deposition of aberrant non-functional species or toxic aggregates resulting in neuronal loss.

Furthermore, this rapid increase in the incidence of AD mentioned above can be attributed in part to the increased longevity of people, as age is a major risk factor for conformational disorders of the brain (46). With major advances in medical science and the healthcare system, we are experiencing a demographic shift, which makes it all the more crucial that we make a breakthrough in our understanding of these highly debilitating neurodegenerative disorders (46).

The first reports for these diseases come from the early 19th century, where clinicians have noted various symptomatic manifestations of these diseases, like the deposition of 'lardaceous' material in various different tissues or organs for diseases like systemic amyloidosis (104). Although these earliest studies include some reports about inclusion bodies in the brain, they do not offer any links to neurodegenerative diseases. James Parkinson and George Huntingdon were the first to report the symptoms associated with PD (105) and HD (106) respectively. These pioneer studies report the symptomatic manifestations of the diseases but do not document the presence of aggregates now known to be associated with them. It was Alois Alzheimer's astute observations in 1907 that suggested a link between the presence of

aberrant deposits in the brain and classic symptoms of neurodegeneration, and his work later led to AD being named after him (107-109).

Although we have not yet uncovered a cure for these complex diseases, more than a century (almost 2 centuries for PD and HD) of studies by some of the most ingenious minds across various disciplines has resulted in a huge leap in our understanding of their mechanisms and pathology, although, the ultimate solution still eludes us. Following the initial reports in the early 19th century about the deposition of 'lardaceous' material in different tissues, there was enormous effort to characterize the exact nature of the material. Matthias Schleiden first coined the term 'amyloid' to describe amylaceous plant constituents (104). However, it was Virchow who popularised this term for the deposits being reported in the 19th century. It was based on the observation that these deposits reacted with iodine in a manner similar to starch or *'amylin'* (110). However, there was much confusion about the matter as later studies by Carl Friedreich, August Kekule and Samuel Wilks, among others, showed that the deposits contained chemical substances similar to albuminoids, but not starch or cellulose (104). The importance of the matter is exemplified by the formation of a committee by the Royal Society of London 'on the nature of the so-called Lardaceous Disease and as to the name by which it should be recognized' (111). In the later years of the 19th century, evidence mounted that the deposits were indeed albuminoid - or as we now refer to as proteinaceous. This was followed by numerous attempts to characterise the structure and composition of these deposits (104). The structure of albuminoids, and more generally proteins was not very well understood in the 19th century. Hence, it was almost impossible at the time to understand the specific protein or proteins that make up the amyloid deposits. It was not until Fred Sanger first sequenced the B chain of insulin (112-115), showing for the first time ever that protein molecules have unique amino acid sequences that elucidating the composition of amyloid deposits seemed achievable. However, there was still the difficult issue of isolating and solubilising the proteins from these deposits. As Virchow rightly pointed out in 1863: 'only when we have discovered the means of isolating the amyloid substance, shall we be able to come to any definite conclusion with regard to its nature' (104). Almost a century later, Newcombe and Cohen described a method to isolate, solubilise and re-precipitate amyloid fibrils (116). Thus, huge technical advancements, both in terms of the ability to sequence a protein

and the methodology for its purification, provided a strong impetus to the efforts towards characterizing the amyloid deposits. Two decades after Newcombe and Cohen reported the purification of amyloid fibrils, Glenner and colleagues, in a huge advancement for the field of Alzheimer's research, identified the main constituent of the characteristic amyloid plaques in AD – a peptide now known as amyloid beta or Aβ. They reported that the main constituent of these plaques bears no sequence homology to any protein identified until that time and hence could be used as a biomarker for disease (117). That article probably marked the beginning of a large amount of studies that would be conducted in the later years studying the utility of Aβ as a biomarker and a therapeutic target for AD. Studies in the 1990s also identified α-synuclein and huntingtin as the main constituents of the aggregates in PD (118) and HD (119) respectively.

The identification of Aβ as the main component of amyloid plaques and the sequencing of its first 24 residues by Glenner and Wong paved the way for numerous studies into the mechanistic details about the pathogenesis. It was soon shown that Aβ_{1-42} is especially aggregation prone (120) and is associated with disease (121-123). This finding was followed by the identification and sequencing of the APP gene, whose proteolysis gives rise to the Aβ peptide in the cell (124). Although the normal cellular function of Aβ still remains debated to date, mutations in the APP gene (23) and the secretase enzymes involved in cleavage of APP gene were identified in the 1990s to be associated with the familial form of AD (19-22, 24-26).

The early studies of the 19th century, despite the crudeness of the histological methods available at that time, had the insight to group the amyloid deposits together (104), implying that they might bear similarities regardless of their tissue localisation. These were the earliest clues that the amyloid deposits might share structural commonalities. Improvements in the isolation techniques for amyloid aggregates and the development of microscopy and X-ray crystallography were instrumental in transforming the field of protein structures. Although there was direct evidence for similarity between amyloid deposits from different sources, the techniques lacked the resolution needed to dissect their structural details. In 1968, Eanes and Glenner succeeded in using X-ray crystallography to elucidate the structure of amyloid fibrils (125). They reported that fibrils obtained from different sources all shared the antiparallel cross-β structure.

These improvements and advances in our understanding of the structure of amyloid fibrils laid the groundwork for studying the fundamental mechanisms of misfolding and aggregation of proteins.

The observations that proteins can aggregate and that aggregation is a general phenomenon were made much before the advent of the technological and methodological advancements described above. However, no structural information existed at the time to investigate the details of the process. Seminal studies by Bernal, Crowfoot (126), Astbury and Lomax (127) in 1934 provided the first evidence for the existence of both globular and fibrous states of a protein – in their case pepsin. When Eanes and Glenner reported the structure of amyloid fibrils, they suggested that the cross-β structure observed under physiological conditions was not necessarily the same as the *in-vitro* conversion of proteins into fibrils. All these studies offered remarkable scientific advancements, most notably the cross-β structure of amyloids and their structural similarity despite heterogeneity amongst the protein giving rise to them.

There was renewed interest, in the late 20th and the early 21st century, in the general capability of proteins to form amyloids. This was probably inspired by Astbury's initial speculations about the general ability of proteins to fibrilise (128). Work from Guijarro *et al*. showed that fibrils could be formed from the small globular SH3 domain (129). Later, Dobson and colleagues showed that myoglobin, a well known globular protein, could indeed be induced to form aggregates (130). Studying the effects of mutations in known amyloidogenic proteins helped to understand the process of amyloid formations. Studies on lysozyme, the first such example, illustrated that although aggregation promoting mutations rendered the protein more vulnerable, they did not abolish the enzymatic activity of the protein (131). It was observed that these mutations lowered the stability of the β domain, hence promoting aggregation (131). This result also suggested that aggregation could result from a subtle destabilization of the native states and not necessarily involving conformational transformation of the soluble protein, thus elucidating the importance of the competition between alternate conformational states of the protein. Indeed, it is now known that molecular chaperones help tip the balance towards properly folded native states, both *in vitro* and *in vivo*, minimising the possibility of amyloid formation. It is

now evident that a wide range of proteins, unrelated in sequence and structure, can be induced to form aggregates – giving rise to the view that amyloid formation is a generic biophysical phenomenon (12, 46). Indeed, Dobson and colleagues also reported that even non-disease related proteins can be cytotoxic when induced to form amyloids, establishing the common cross-*β* structural similarity of amyloids to be the basis of cellular toxicity (132).

These findings about the generic nature of amyloids and their inherent toxicity, regardless of the specific proteins that form them, provided important insights into the importance of preventing protein aggregation in neurodegenerative diseases. In parallel with the mounting evidence about the failure of the protein homeostasis system during ageing and disease, these studies shed light on the role of the delicate and fragile nature of the balance between protein folding and aggregation that evolution has achieved, but which is highly susceptible to disruption.

1.5 The genetics of Alzheimer's disease

Along with the efforts to study the structure of amyloid fibrils and to identify amyloidogenic proteins, the end of the 20th century also witnessed increasing interest in understanding the genetic basis of these neurodegenerative diseases. It was Gusella's discovery of the Huntingtin gene in 1983 (133) that illustrated first the potential of genetic analysis for neurodegenerative diseases. For AD, one of the first evidence pointing to its genetic dependence was the increased incidence of disease in the families of affected individuals. Evidence also started emerging that AD segregated as autosomal dominant condition (134, 135). The increased incidence of Down's syndrome in families of affected individuals and the increased incidence of dementia of Alzheimer's type in individuals with Down's syndrome over the age of 35 suggested the possibility of a genetic association between the two (136). The heterogeneous nature of AD made the initial attempts at identifying the genetic locus a laborious task, with the earliest reports claiming that the Alzheimer's locus was the same as APP (137, 138) and that there was a duplication of the chromosome 21 in AD (139). Although the duplication was later found to be an error (140) and the subjects used to study linkage of APP and the genetic locus were indeed found to be

chromosome 14 linked (141), it was serendipitous that they still managed to suggest APP as a genetic locus for AD. Subsequent genetic analysis revealed that AD was indeed linked to chromosome 21 (142) but also that the disease was highly heterogeneous (143, 144). The first mutation identified in Aβ was in relation to hereditary cerebral haemorrhage with amyloidosis where Aβ was found to be deposited in the walls of the blood vessels of affected individuals (145). This observation paved the way for the identification of AD specific mutation in the APP genes, and it was soon identified that a point mutation (Val→Ile) at position 717 was highly pathogenic in various AD pedigrees (146). These familial cases have a particular phenotype and the age of onset was approximately 50 years, associating this mutation with the early onset AD. The presence of Lewy bodies and tangles was also reported in subjects with the APP mutation. Although there was sufficient evidence implicating the role of APP mutations in AD, a large number of familial cases were still found to lack any linkage with chromosome 21. Work from St. George-Hyslop and colleagues succeeded in identifying *presenilin1* as another important locus linked with early onset AD (147). Subsequent studies would also identify mutations in *presenilin2* (148, 149).

While the importance of APP in the pathology of early onset AD was well established by the end of 20th century, its physiological role within the cell still remained elusive. It was subsequently discovered that the 695-residue amyloid 14 precursor, one of the three splicing products of the APP gene, was an integral transmembrane protein, with three extracellular domains, one transmembrane domain and a 47 residue cytoplasmic domain (124). APP was also seen to undergo rapid metabolism in further studies (150). Later, Haass *et al.* reported that soluble Aβ is a normal product of cellular metabolism, both *in vitro* and *in vivo* (151). With these studies as the background, there was a whirl of excitement to better understand the production and metabolism of APP and the production of Aβ using the tools of genetic analysis. Studies of APP metabolism in the presence of the pathogenic mutations revealed an increase in the production of Aβ (20, 123, 152, 153). These observations formed the framework for the birth of what is now well known as the 'amyloid hypothesis' of AD. Presenilin

mutations were also shown to be involved in the cleavage of APP in cultured primary mice neurons, with an ultimate increase in Aβ production[i].

Together with the *presenilin* mutations, mutations in the APP gene have been reported for almost all cases of autosomal dominant, early onset AD (18). However, these are responsible for a very small fraction of the reported AD cases (18). Majority of AD cases are sporadic, with a late age of onset and no apparent genetic signature. Linkage studies in the early 1990s from Pericak-Vance *et al.* and Strittmatter *et al.* succeeded in identifying APOEε4 as major locus in the etiology of late onset AD (27, 28). However, the mechanistic details of this association still remain incompletely understood.

Efforts to identify more risk factors for late onset AD (LOAD) through linkage studies failed to fruitful, mainly because these studies relied on knowledge of pre-conceived candidate genes for linkage analysis and thus, the design of the study itself excluded the possibility of identifying novel loci (154). The high cost and laborious nature of the technique also impeded the study of multiple markers simultaneously. With the rising popularity of the microarray technology in the late 1990s and the major implications of the sequencing of the human genome in 2001, it became possible to simultaneously study thousands of genes and associated polymorphisms and their role in various human diseases. These advances allowed for a hypothesis-free approach to analyse the whole genome for allelic polymorphisms associated with a particular condition – which are now popularly known as genome wide association studies or GWASs (18). As seen throughout the historical perspective of AD research, this huge leap in technological advancement provided tremendous momentum to the field, with new loci being reported to date **(Figure 1.3)** (18). However, the problem of massive multiple testing warranted the development of robust statistical models to justify the observed effects.

---

[i] Mutations were also identified in the three Aβ processing enzymes – β-secretase, responsible for the N-terminal cut of Aβ from APP (at APP 671); α-secretase, responsible for the alternative cleavage of the shorter fragment p3 (APP687); and γ-secretase, responsible for releasing both Aβ and p3 from the relevant C-terminal stubs of APP and cleaving intramembranously around APP residues 711–713 (153).

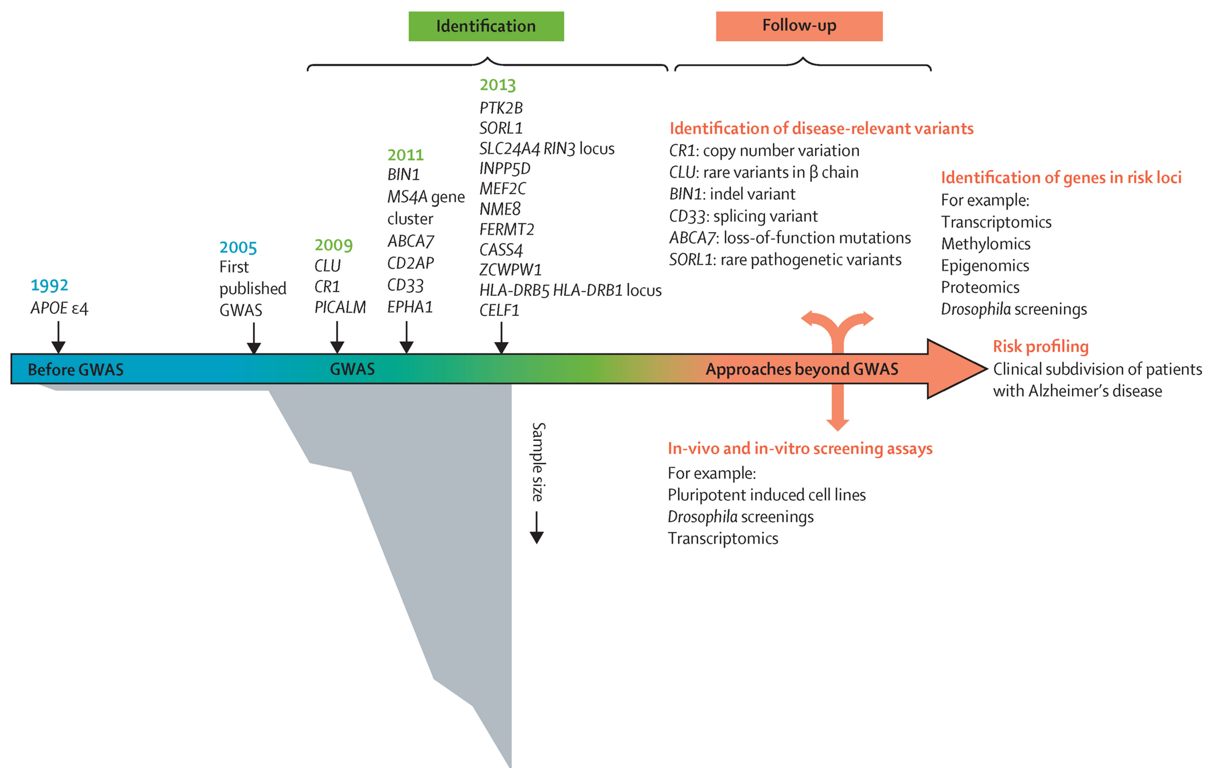


Figure 1.3 Progress in risk gene identification, follow-up, and relevance in Alzheimer's disease research. Identification of novel risk genes by GWAS are shown in green. Follow-up and relevance of the identified risk genes are shown in orange. The grey graph underneath the timeline arrow represents the increase in the sample size of the most recent GWAS in Alzheimer's disease research (18).

Reiman *et al*. reported the first significant AD associated GWAS locus - GRB2-associated binding protein 2 (GAB2) (155). They observed that the effect was more pronounced in conjunction with APOEε4. However, the effect was not consistently replicated in subsequent studies, with it often showing non-significant effect sizes (156-158). The smaller size of the early GWASs was probably a major factor in their failure to identify significant loci. In the following year, Bertram *et al.* reported three novel loci: *ATXN1* (ataxin 1), *CD33* (siglec 3), and an as yet uncharacterized locus on chromosome 14 (GWA_14q31.2) (159). While the latter locus failed to show validation in subsequent case control studies (160), there were studies, albeit only a few, to merit the association of *ATXN1* and *CD33*. *ATXN1* was suggested to modulate the levels of Aβ *in vitro* (161). *CD33* belongs to a family of lectins involved in cell-cell interactions and in regulation of cells of the immune system (162, 163) – both

believed to be involved in the inflammatory reaction observed in AD brains. During this time, several other loci were suggested but failed to reach independent validation. In 2009, two large-scale GWASs from Lambert *et al.* (164) and Harold *et al.* (157) suggested three novel AD associated loci, all of which were replicated in subsequent studies: *CLU* (clusterin; a.k.a. apolipoprotein J), *CR1* (complement component (3b/4b) receptor 1), and *PICALM* (phosphatidylinositol binding clathrin assembly protein). Clusterin is a molecular chaperone, and has been implicated in Aβ fibrillization, clearance and aggregation (165, 166) and also regulation of lipid metabolism in the brain (167). *CR1* is the main receptor of the complement C3b protein, complement activation being suggested as a protective mechanism in AD (168, 169). *PICALM* plays a role in clathrin mediated endocytosis and synaptic transmission (170). Latest additions to the list are *BIN1* (bridging integrator 1[i]; and *EXOC3L2* (exocyst complex component 3-like 2) (171).

Although GWASs have succeeded in breaking the deadlock facing linkage studies and transform the landscape of LOAD genetics over the past decade by suggesting novel genetic associations, the effect sizes of individual genes are extremely small, making it unlikely to explain fully the etiology of LOAD. As for rare variants associated with EOAD, *i.e.* APP, PSEN1 and PSEN2, meticulous functional characterization *in vitro* and *in vivo* of the GWAS genes is required to establish their role in AD. This aspect has proved particularly problematic because unlike the EOAD genes, which assert their effects via mutation in functional genes, GWAS variants are mostly present in genomic regions of no functional consequence. Also, polymorphisms in the non-coding regulatory regions of the genome can have subtle manifestations, such as modification of expression levels of a particular protein, which are particularly difficult to study over a large scale. GWASs have also enumerated genes associated with different physiological processes that were previously never associated with AD, opening up the field for broader interventions. However, the various loci uncovered have, as yet, failed to elucidate a comprehensive mechanism for LOAD. It would probably require a merger of the hypothesis-free GWAS line of enquiry and a rigorous understanding of the fundamental biology of AD to allow us to tackle this 'twenty-first century plague' (13).

---

[i] Originally implicated at subgenome-wide significance by Harold et al. (157)

1.6 Protein aggregation as a widespread phenomenon

Protein misfolding diseases have been historically associated with the aggregation and deposition of specific proteins in the form of amyloid plaques and neurofibrillary tangles, mainly formed by Aβ and tau, respectively, although the deposits may contain other components (172-174). As described above, for almost a century it was believed that the ability to form amyloid fibrils is a property of specific disease associated proteins, conferred by certain aberrant sequence motifs in the polypeptide chain, until the work by Dobson and colleagues, illustrating the generic nature of amyloid state (7, 11, 12, 14, 175). The observation that homopolymers such as polythreonine or polylysine are able to form amyloids *in vitro* proved that amino acid sequences are inherently capable of forming amyloid structures (130). Although specific amino acid sequences were shown to be rather inconsequential in determining the amyloidogenic nature of a polypeptide, they play an important role in determining the propensity to form amyloids, or in other terms the sequence of a polypeptide chain influences its aggregation propensity (175, 176). Indeed, it was shown by Chiti *et al.* that a single amino acid mutation in a 100-residue protein could cause a change of an order of magnitude or more in the aggregation rate of that protein (175). This change of aggregation rate also correlated with the predicted change in other physicochemical properties, such as charge and hydrophobicity (175, 176) – a correlation subsequently found to be validated for a large range of sequences, endorsing the generic nature of aggregation and amyloid formation.

The native and the amyloid states can be considered to be in competition, making the maintenance of the balance between their populations highly significant. Indeed, Vendruscolo and colleagues reported a negative correlation between the expression of human genes and the aggregation propensity of their respective proteins (49). This observation was highly interesting because it suggested that proteins are expressed at the limits of their solubility and that their aggregation propensities have been finely tuned through random mutations and evolutionary selection to allow them to be functional at the required concentrations. However, it also means that proteins have co-evolved with the cellular environment to be sufficiently soluble so as to be

functional, but not more so. Hence, even relatively small increases in the concentration or aggregation propensity, such as those caused by mutations or regulatory process failure, can push the proteins over their solubility edge and cause them to aggregate (49).

The role of protein homeostasis thus becomes even more paramount given an environment in which proteins are expressed at their solubility limits. A possible strategy to maintain the integrity of the proteome would be for each step of protein synthesis and folding to be tightly regulated, such as the formation of misfolded and aggregated species is implausible. Orchestrating the exquisite control of each step in the process, from the synthesis of the polypeptide at the ribosome, to its folding and modification and its subsequent degradation would be exceedingly energy intensive. It would also considerably impede the rate of protein synthesis, as each step would need to be checked meticulously. An alternative situation would be a one in which large amounts of proteins are synthesised and degraded continuously, such that at any given time, enough properly folded structures are present to maintain the functionality of the proteome. However, there would also be a high-energy cost associated with the synthesis of a large number of non-functional proteins and the subsequent identification and degradation of large numbers of misfolded species. Bennink and colleagues provided evidence favouring the latter approach, when they showed that almost 30% of newly synthesised proteins are rapidly degraded in the cell (177). Ubiquitinated and misfolded species were seen to accumulate upon inhibition of the proteasome – shedding light on the role of proteasome in the degradation of newly synthesised proteins.

Given the fact that proteins are expressed at their solubility limits and the existence of an intricate protein homeostasis system, it is probable that any kind of physiological stress can have major effects on proteome integrity. There is increasing evidence that hundreds, even thousands, of proteins unrelated and sequence and structure can aggregate during the course of ageing (29, 178), disease (172-174), or upon heat shock (179, 180) or due to expression of amyloidogenic proteins (31). It is not currently clear whether the widespread aggregation within the cell leads to the formation of amyloids or amorphous species. Independently from these possibilities, the importance of solubility in maintaining a functional proteome is reinforced by the

fact that so many different proteins can aggregate under conditions of stress. This widespread aggregation can also rationalise, in part, the multifaceted pathology of neurodegenerative diseases (33, 181, 182).

As it turns out, proteins associated with neurodegenerative diseases are inherently highly metastable. In their analysis, Vendruscolo and colleagues reported that a large number of proteins, in a healthy state, are expressed at levels higher than their intrinsic solubilities – *i.e.* they are supersaturated (34). Specific proteins that co-aggregate with plaques and tangles were seen to have elevated supersaturation levels and biochemical pathways associated with neurodegeneration were enriched in these supersaturated proteins. The presence of such a metastable subproteome would enhance the risk of widespread aggregation under physiological stress. This analysis highlighted an important link between the two branches of the study of misfolding diseases - the study of biophysical and biochemical phenomena governing protein folding and the widespread cellular disruption characteristic of disease pathology (33, 34).

1.7 Diagnostic and therapeutic strategies for Alzheimer's disease

The heterogeneous and multifactorial nature of AD, along with a lack of clear understanding of its underlying mechanisms presents an immense challenge in the design of diagnostic and therapeutic strategies. Differentiating between AD and other types of dementia has proved to be another roadblock. For the better part of the 20th century, definitive AD diagnosis depended upon post-mortem detection of plaques and tangles along with a history of dementia (183). From a clinical perspective, the definition of AD has evolved drastically over the last decade or so to involve patients with milder symptoms and the acknowledgement of a preclinical phase of the disease – during which, although there might be no phenotypical manifestations, the pathological changes have already initiated in the brain (183-185). Efforts to define this disease have been further complicated by the observation that many older individuals, considered cognitively healthy at the time of death, show AD-related

neuropathological changes in the brain (186-194). There have been increasing efforts to differentiate the clinical term 'Alzheimer's disease', which is based on the presentation of cognitive and behavioural symptoms, from the AD associated neuropathological changes, which refer to histopathological changes seen in the brain at the time of autopsy regardless of the clinical setting (183).

The current diagnostic criteria divide AD into three stages – the preclinical stage, mild cognitive impairment (MCI), and AD dementia (184, 185, 195). The preclinical stage is devoid of any symptoms and can extend over a period of years, even decades, before proceeding to the state of MCI. Although there have been reports that certain CSF biomarkers, such as increased CSF tau and cortical thinning can be detected during the pre-clinical stage (185), in the absence of any symptoms, diagnosing the disease at this initial stage still remains challenging. MCI, as the name suggests, is accompanied by mild reduction in cognitive function, albeit to a level that still allows the subject to be independent and functional (184). Despite all the advances in imaging and biomarkers, diagnosis of MCI still depends on cognitive and functional tests, the interpretation of which is often at the discretion of the physician. MCI cannot yet be diagnosed by a standardised laboratory test (184). Advanced cognitive and behavioural impairment paired with a cognitive or function test and PET scans are the best available means to diagnose AD dementia (195).

Identification of biomarkers for AD has gained much interest due to possibility of early detection and also monitoring disease progression and also act as potential therapeutic targets. Traditional biomarkers like Aβ and tau, although have provided important knowledge about disease pathology, have so far failed to produce tangible results. Reduced Aβ levels in CSF and evidence of Aβ deposits in PET scans have been reported to be a feature of AD (184, 185), although alone, these are insufficient for a definitive AD diagnosis. Increased CSF levels of total and phosphorylated tau have also been seen in disease pathology. Existence of variability in biomarker measurements between different laboratories has also been a limiting factor in effective biomarker development – biomarker levels fluctuate greatly with different disease stages, hence it is imperative to be able to definitively identify various stages of disease progression (196). Rapid advances in imaging techniques, like high-resolution MRI and molecular imaging using PET has greatly augmented the efforts

towards AD diagnosis. PET has completely revolutionised the *in vivo* detection of amyloid deposits, owing to its high sensitivity, which has greatly helped the clinicians in staging AD (196-199). Radioligands capable of binding tau have also helped in studying tau deposition, which has been reported to be a better predictor of cognitive decline than Aβ load (200-203). Overall, however, clinicians and scientists still lack a consensus on diagnostic strategies for AD, mostly due to the fact that the present strategies only work in conjunction with each other and that the interpretation can be highly subjective.

Development of various different biomarker and imaging technologies, accompanied by an exponential increase in literature regarding disease pathology have so far failed to translate into effective therapeutics for AD. Research into the possible mechanisms of AD over the last 3 decades has inspired various therapeutic strategies, but they have so far been highly disappointing (204). The 'amyloid hypothesis' has been the basic premise for most of therapeutic efforts, with strategies involving inhibition of cleavage of APP resulting in lower Aβ production or direct targeting of Aβ and tau using small molecules, resulting in reduction of plaques and tangles (205). High-specificity antibodies (206), especially against Aβ and tau have also been reported to ameliorate certain pathological features in animal models. Despite enormous effort into the development of these strategies and encouraging results in pre-clinical studies, most of these strategies have failed massively in clinical trials, resulting also in enormous financial losses (204, 207). A recent example is solanezumab, a humanised monoclonal antibody shown to preferable bind soluble Aβ and ameliorate its clearance in pre-clinical studies, failed to produce any significant improvement in phase 3 clinical trials in 2014 (207). A review of clinical trials for AD held between 2002-2012 showed that 98% candidate drugs failed in phase 3 (204).

The failure of therapeutic interventions in AD so far can be attributed to two major factors. The first factor is a lack of mechanistic insight into the interaction of therapeutic agent (such as small molecules or antibodies) with the target molecule (such as Aβ or tau), which leads to an almost blind dependence on the final read out, like reduction in amount of deposits. This lack of quantitative understanding might be particularly dangerous in the light of the growing knowledge of the kinetics of protein aggregation, the measurements of which have recently been made possible (53).

Without discerning the exact mechanism of action of a therapeutic candidate, it is impossible to determine the stage of disease at which it would be beneficial, greatly increasing the risk of failure of the subsequent clinical trials.

The second factor is the inability to correctly and definitively diagnose AD, particularly the pre-clinical stage, where the neuropathological changes are just beginning. It is plausible that due to the advanced stage of the disease of most subjects recruited for clinical trials, it is almost impossible for the drug to show any effect as the damage may have reached a point of no return. Thus, rational design of therapeutics, grounded in our increasing knowledge of the physicochemical properties or proteins and the kinetics of the aggregation process, accompanied by development of biomarkers capable of early detection of AD are the need of the hour to allow for the management and/or a cure for this highly debilitating disease. An anticancer drug, bexarotene, was shown to selectively inhibit primary nucleation of Aβ and delay the deposition of toxic species in neuroblastoma cells by Habchi *et al* (208). It was also reported to supress Aβ deposition and associated pathology in a C. *elegans* model. Hence, the rational design of therapeutics will prove instrumental in navigating the current bottleneck of AD therapeutics (209, 210). Another avenue of progress could be the incorporation of the protein homeostasis components as therapeutic targets. Since disease pathology is characterised by widespread aggregation and cellular dysfunction, targeting singular proteins like Aβ at the advanced stages of the disease might be insufficient to reverse the damage caused.

1.8 Towards an understanding of the molecular origins of neurodegenerative diseases

Although the myriad pathological disruptions and varied clinical presentations of neurodegenerative disorders paint a daunting picture in terms of developing effective therapeutic strategies, the study of these disorders in terms of common fundamental physicochemical properties offers a fresh perspective (7, 12, 33, 47). The observation that a large number of proteins can aggregate in these diseases can be rationalised to a large extent by the generic and predictable nature of protein aggregation (7, 12, 33,

34). The inherent metastability against aggregation of a substantial part of the proteome provides novel therapeutic and diagnostic approaches that can look beyond the focus on specific disease-related proteins.

Equally, understanding the mechanisms of regulation of the metastable proteome in healthy cells and organisms is also of great importance. Although different types of stress, environmental disturbances, polymorphisms, altered transcriptional levels can all alter proteome metastability and push proteins towards supersaturation, the protein homeostasis system is still capable of restoring the balance and maintain a functional proteome. Hence knowledge about the regulation of metastable proteins will offer opportunities for novel rational therapeutic approaches aimed at restoring the balance between metastable proteins and their natural quality control systems.

The studies described in this thesis represent an initial attempt to achieve this goal. We have utilised extensive sequencing data, systems biology and bio-informatics approaches to study the behaviour of metastable proteins in AD and to identify a specific protein homeostasis complement (PHC) responsible for their regulation. We have also aimed to study the role of the balance between metastable proteins and their regulation in determining tissue vulnerability to misfolding diseases.

These studies represent an early effort to gain a holistic understanding of multifactorial protein misfolding diseases. Such efforts have become necessary in the wake of failure of various therapeutic strategies. As the 20th century efforts to group these diseases into ‘amyloids’ created a breakthrough in their understanding, the time has now come to look at these diseases in terms if their shared molecular origins, grounded in the knowledge of the generic features of protein aggregation.

CHAPTER 2

---

2. A transcriptional signature of Alzheimer’s disease is associated with a metastable subproteome at risk of aggregation

2.1 Overview

It is well established that widespread transcriptional changes accompany the onset and progression of Alzheimer’s disease. Because of the multifactorial nature of this neurodegenerative disorder and its complex relationship with aging, however, it remains unclear whether such changes are the result of non-specific dysregulation and multi-systems failure, or instead are part of a coordinated response to cellular dysfunction. To address this problem in a systematic manner we performed a meta-analysis of about 1600 microarrays from human central nervous system tissues to identify transcriptional changes upon aging and as a result of Alzheimer’s disease. Our strategy to discover a transcriptional signature of AD revealed a set of

downregulated genes that encode proteins metastable to aggregation in control brains[i]. Using this approach, we identified a small number of biochemical pathways, notably oxidative phosphorylation, enriched in proteins vulnerable to aggregation in control brains and encoded by genes downregulated in Alzheimer's disease. These results suggest that the downregulation of a metastable subproteome may act to mitigate aberrant protein aggregation when protein homeostasis becomes compromised in Alzheimer's disease (181).

2.2 Introduction

Alzheimer's disease (AD) is a neurodegenerative condition responsible for the majority of the reported cases of dementia, (3, 4, 7, 9, 10, 12, 15-17, 101, 211) and its onset and progression have been associated with a multitude of factors, including mitochondrial dysfunction, disruption of the endoplasmic reticulum and membrane trafficking, disturbances in protein folding and clearance, and the activation of the inflammatory response (3, 4, 7, 9, 10, 12, 15-17, 101, 211). As discussed in Chapter 1, it is clear that AD belongs to a class of protein conformational disorders whose characteristic feature is that specific peptides and proteins misfold and aggregate to form amyloid assemblies (7, 16, 17). The presence of such aberrant aggregate species generates a cascade of pathological events, leading to the loss of the ability of protein homeostasis mechanisms to preserve normal biological function and to avoid the formation of toxic species (7, 16, 17).

The appearance of protein aggregates in living systems is increasingly recognized as being common, as described in Section 1.6. Growing evidence indicates that proteins are only marginally stable against aggregation in their native states (7, 49) and that the molecular processes that prevent protein aggregation decline with aging (32, 70, 97, 212). Thus, protein aggregation is emerging as a widespread biological phenomenon, in which hundreds of different proteins can aggregate in aging, stress or disease (29-32, 172-174, 178, 179, 213-215). To understand why some proteins aggregate while

[i] This chapter is based on a published manuscript (181). The findings were obtained in collaboration with Dr. Prajwal Ciryam.

others remain soluble, we recently observed that many proteins in the proteome are insufficiently soluble relative to their expression levels (34). Such proteins are metastable to aggregation as their concentrations exceed their solubilities, i.e. they are supersaturated (33, 34, 216, 217). Upon formation of aggregate seeds by nucleation events, a supersaturated protein will form insoluble deposits until the concentration of its soluble fraction is reduced to match its solubility (33, 34, 216-218). We found that the proteins that co-aggregate with inclusion bodies, those that aggregate in aging, and those in the major biochemical pathways associated with neurodegenerative diseases tend to be supersaturated (34). The observation that these metastable proteins appear to be a common feature in aging, stress and disease prompts the question of whether or not their supersaturation levels are altered in AD. These levels are particularly crucial, as supersaturation represents a major driving force for aggregation (33). The downregulation of supersaturated proteins can thus limit their aggregation in response to compromised protein homeostasis.

In the present study, we examined the experimental information acquired in the last decade about transcriptional changes in AD (219-233). We aimed to determine the relationship between protein supersaturation and the transcriptional changes that occur during normal aging and in AD. We found that distinct but partially overlapping transcriptional changes take place in aging and AD. Moreover, downregulated genes generally correspond to metastable proteins at risk of aggregation, as they are supersaturated and encoded by highly expressed genes. Accordingly, the biochemical pathways downregulated in AD are nearly identical to those previously identified as highly enriched in supersaturated proteins (34, 234). These changes are also accompanied by a transcriptional downregulation of certain components of the protein homeostasis network. The downregulation of genes corresponding to supersaturated proteins may thus represent a specific mechanism to limit widespread aggregation by regulating cellular concentrations in a compromised protein-folding environment.

2.3 Results

2.3.1 Analysis of the transcriptional changes in aging and in AD

A long-standing question is whether AD represents an acceleration of the normal aging process or a qualitatively distinct phenomenon (234). Determining changes in gene expression can offer important insights into this problem. The complications associated with obtaining human tissue samples, however, constrain the extent to which confounding variables such as age, gender and tissue type can be controlled in a transcriptional analysis of AD. In the present work, the control samples (mean 70.8 ± 16.4 years) are younger than the disease samples (mean 81.1 ± 9.5 years), necessitating the use of techniques to account for these disparities (**Table A.1** and **Section 2.6**).

For the human genes examined in our analysis, we constructed a linear model of expression differences across a range of factors (**Section 2.6**). We thus obtained the overall median magnitude and statistical significance of expression changes by combining these individual values across different studies. In this analysis, microarray probes were mapped onto UniProt IDs to determine the corresponding protein (**Section 2.6**). Using this procedure, we determined the transcriptional changes associated with 19,254 genes. An important aspect of this approach is that the effects on gene expression of different factors are considered as additive. Because the occurrence of AD increases with age, Alzheimer's subjects exhibit specific disease-related transcriptional changes in addition to those associated with natural aging. We considered a gene to be differentially expressed if it undergoes a change in expression of at least 10% with a Benjamini-Hochberg-corrected p-value ≤ 0.01. We then tested over 18,000 other combinations of thresholds and found our results to be robust to changes in these thresholds (**Figures A.1** and **A.2**). In the used model here, the aging component is a linear variable, and therefore estimating the magnitude of change requires specifying a range of ages. As the assumption of linearity is expected to hold

best near the average age, we used the change in expression for an age range of approximately two standard deviations, i.e. 25 years.

2.3.2 Proteins that aggregate in AD correspond to transcriptionally downregulated genes

We next asked how the transcriptional changes identified in aging and AD might be associated with protein aggregation. First we considered the set of disease-related amyloid proteins, i.e. those annotated as 'amyloid' in UniProt, which include those associated with neurodegenerative diseases (34). On average, we could not detect an overall connection between amyloid proteins and proteins corresponding to differentially expressed genes (**Figure 2.1 A,B**). We also note, however, that this analysis does not imply that individual genes in the amyloid class may not have important roles in AD. As an example, the downregulation of the APP gene (in our analysis by 9.5%, with p=0.011) has been reported in neurons containing neurofibrillary tangles (235).

We identified, however, a clear signal for another set of proteins associated with AD, namely those that co-aggregate with amyloid plaques (172) and neurofibrillary tangles (173) in human autopsy samples as identified by mass spectrometry. Among the proteins that co-aggregate with plaques (35%, $p = 4.7 \cdot 10^{-3}$) and tangles (41%, $p = 1.7 \cdot 10^{-13}$), a disproportionate number correspond to downregulated genes in AD (**Figure 2.1A**) in addition to those that are downregulated during natural aging (**Figure 2.1C**). Proteins corresponding to genes downregulated in aging are overrepresented among tangle co-aggregators (10%, $p = 2.5 \cdot 10^{-3}$) but not plaque co-aggregators (4%, p = 1.0) (**Figure 2.1C**). By contrast, only an insignificant number of genes encoding proteins aggregating in plaques and tangles was observed to be upregulated in AD (**Figure 2.1B**) or aging (**Figure 2.1D**).

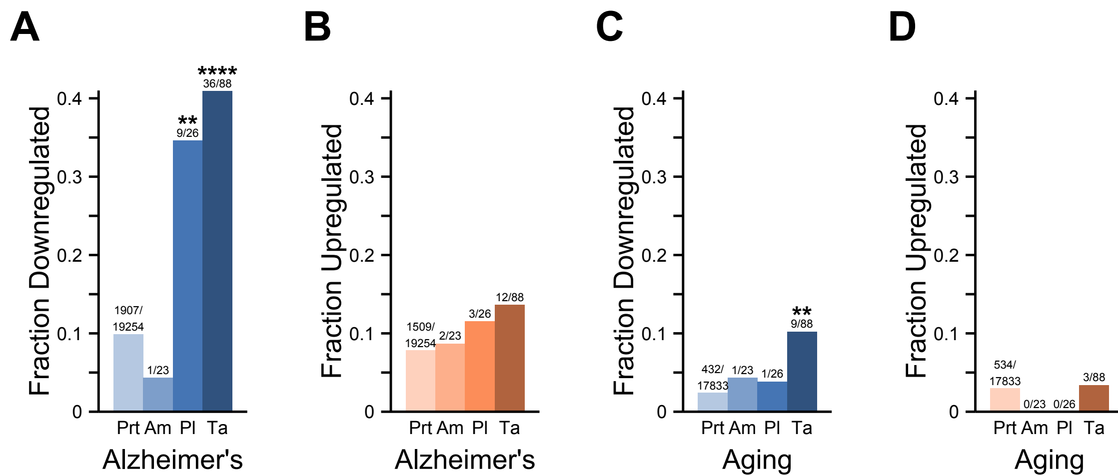


Figure 2.1 Proteins that aggregate in AD correspond to transcriptionally downregulated genes. (A, B) Fraction of proteins corresponding to transcriptionally downregulated (A) or upregulated (B) genes in AD in the whole proteome (Prt, downregulated fraction 1907/19254, upregulated fraction 1509/19254), and for amyloid deposits (A, 1/23, 2/23), plaques (P, 9/26, 3/26) and tangles (T, 36/88, 9/88). (C, D) Fraction of proteins corresponding to transcriptionally downregulated (C) or upregulated (D) genes in aging in the whole proteome (Prt, 432/17833, 534/17833), and for amyloid deposits (A, 1/23, 0/23), plaques (P, 1/26, 0/26) and tangles (T, 9/88, 3/88). The statistical significance of the difference with the proteome (first column) was assessed with a Fisher's exact test with Holm-Bonferroni corrections (***p < 0.001, ****p < 0.0001).

2.3.3 Metastable proteins correspond to transcriptionally downregulated genes in aging and in AD

We next investigated if the fact that so many proteins that co-aggregate with plaques and tangles correspond to genes downregulated in AD could be a consequence of their metastability to aggregation. We previously observed that these metastable proteins tend to be supersaturated, having concentrations exceeding their solubility limits (34). Here, we calculated the metastability of proteins to aggregation in terms of supersaturation scores (σ_u), which represent the risk of proteins aggregating from their unfolded states (34). We assessed proteins corresponding to genes downregulated in AD to be about 8.8-fold (8.8X, $p < 2.2 \cdot 10^{-16}$) more metastable than those for which the expression levels of the corresponding genes do not change significantly in disease (**Figure 2.2A**). Similarly, we found proteins encoded by genes downregulated

in aging to be more metastable (7.4X, $p < 2.2 \cdot 10^{-16}$) than those whose expression does not change (**Figure 2.2B**).

We also found that proteins corresponding to genes upregulated in AD (1.3X, $p = 9.7 \cdot 10^{-13}$) (**Figure 2.2A**) and in aging (1.5X, $p < 8.8 \cdot 10^{-7}$) (**Figure 2.2B**) are modestly, but significantly, more metastable than those with unchanged expression in AD. These upregulated genes are almost exclusively associated with an inflammatory response. For example, of those genes that encode metastable proteins, the most highly upregulated gene (123% increase in expression) in AD is alpha-1 antichymotrypsin, which inhibits serine proteases, particularly those active in inflammation (236).

Despite the fact that only 16% of downregulated genes are common to aging and AD (**Figure 2.2D**), in both cases the transcriptional response appears to be associated with metastability to aggregation (**Figure 2.2A-C**). Indeed, we observed a significant overlap ($p < 2.2 \cdot 10^{-16}$) between the most metastable proteins ($\geq 95^{th}$ percentile), proteins corresponding to genes downregulated in AD, and in aging, as well as between any two of these categories (**Figure 2.2D**). At the intersection of these three groups there is a set of transcriptionally downregulated genes encoding a group of proteins making up a metastable subproteome specific to AD, which is here referred to as the 'metastable subproteome'. By contrast, the most transcriptionally upregulated genes in AD and in aging overlap significantly with each other, but neither group is significantly enriched in genes encoding metastable proteins (**Figure 2.2E**). As a control, we divided the downregulated and upregulated genes into low, medium, and high levels and calculated the supersaturation scores at each of these levels (**Figure 2.3A-D**). Our results indicated a trend towards increasing levels of supersaturation with increasing levels of downregulation in AD (**Figure 2.3A**). This correlation is weaker in aging (**Figure 2.3C**), and weaker still among upregulated genes (**Figure 2.3B,D**). The negative correlation between protein supersaturation and gene downregulation also persists at the individual level for AD, but much less so for aging (**Figure A.3**).

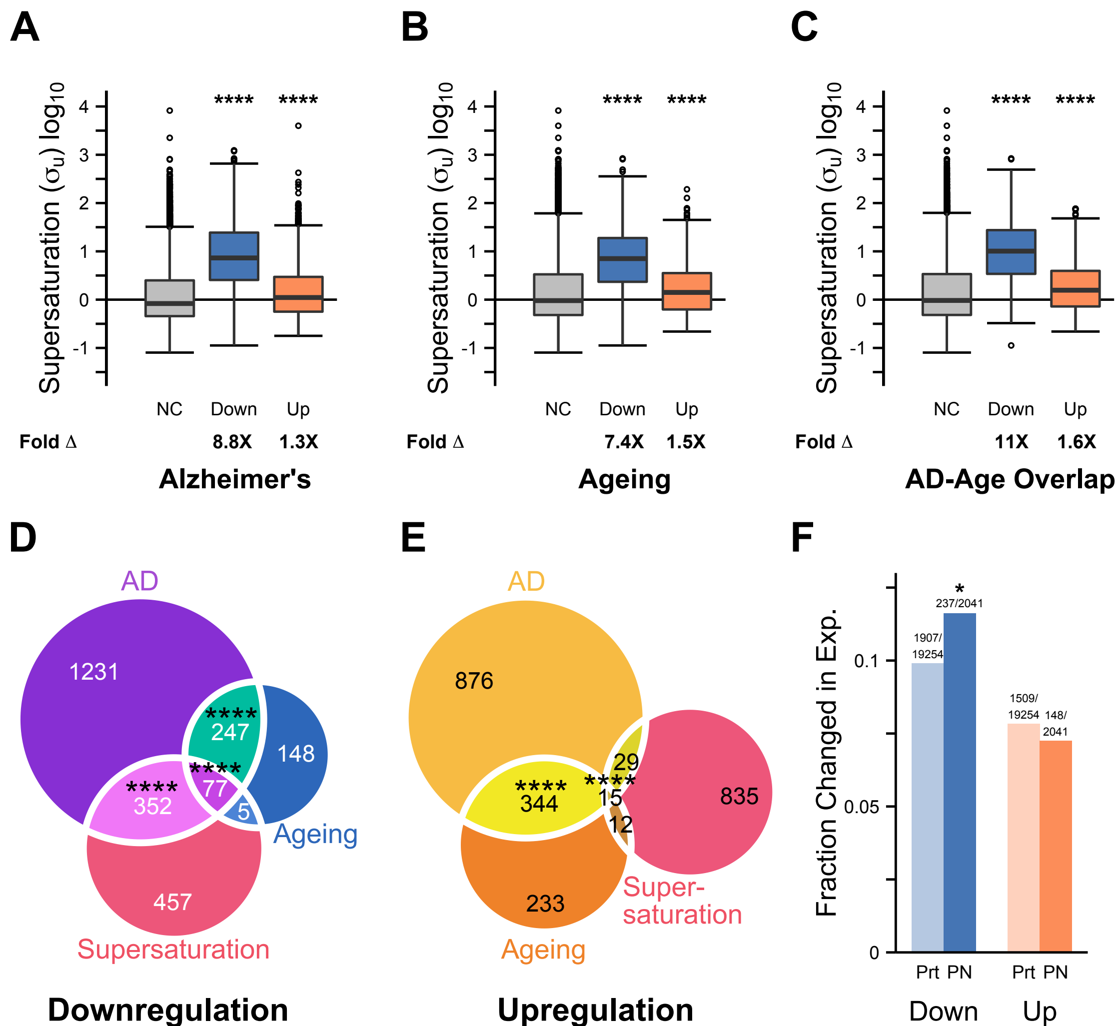


Figure 2.2 Transcriptionally regulated genes in aging and AD correspond to proteins metastable against aggregation. (A-C) Assessment of the metastability to aggregation of the proteins associated with differentially expressed genes in (A) AD, (B) aging and (C) the overlap between the two groups. The median fold difference in supersaturation (or measure of metastability to aggregation) is indicated by 'Fold Δ'. 'NC', 'Down' and 'Up' indicate, respectively, no change in expression, downregulation and upregulation. (D-E) Overlap between the 5% most supersaturated proteins and the corresponding genes either (D) downregulated or (E) upregulated in aging and AD. The number of proteins in each subset is indicated. (F) Fraction of genes downregulated (blue) and upregulated (orange) in the whole proteome (Prt, downregulated fraction 1907/19254, upregulated fraction 1509/19254) and the protein homeostasis network (PN, 1,509/19254, 148/2041). For A-C, ****$p \leq 0.0001$, one-sided Wilcoxon/Mann-Whitney test with Holm-Bonferroni correction. For D-E, *$p \leq 0.05$, **** $p \leq 0.0001$, one-sided Fisher's exact test with Holm-Bonferroni correction).

Elevated supersaturation scores of differentially expressed genes may result from an easier detection of the differences in highly expressed genes than in genes of low expression. To control for this possibility, however, we excluded high expression genes from our analysis, finding the median supersaturation of proteins corresponding to differentially expressed genes to be elevated even after this procedure (**Figure A.4**). We also tested the robustness of our results against changes in the details of our analysis. We found that our results on the metastability of the proteins corresponding to differentially expressed genes are stable across a wide range of thresholds for the defining the groups of upregulated and downregulated genes (**Figures A.1 and A.2**), and also against the introduction of Gaussian noise into the supersaturation score (**Figures A.5 and A.6**).

2.3.4 Specific protein homeostasis components correspond to genes downregulated in AD

As we have discussed above, widespread downregulation of genes corresponding to metastable proteins may represent a general mechanism to maintain protein homeostasis upon aging and AD. An additional transcriptional response, however, may also involve specific components of the protein homeostasis network (212). Following a recent study that showed an enrichment in genes downregulated in aging in this network (212), we examined whether or not particular subnetworks in the overall protein homeostasis network correspond to genes particularly downregulated in aging and AD (**Figure 2.2F**). We found a significant number of protein homeostasis network genes in the 'trafficking' subnetwork to be downregulated in AD (14%, $p=1.1\cdot10^{-2}$).

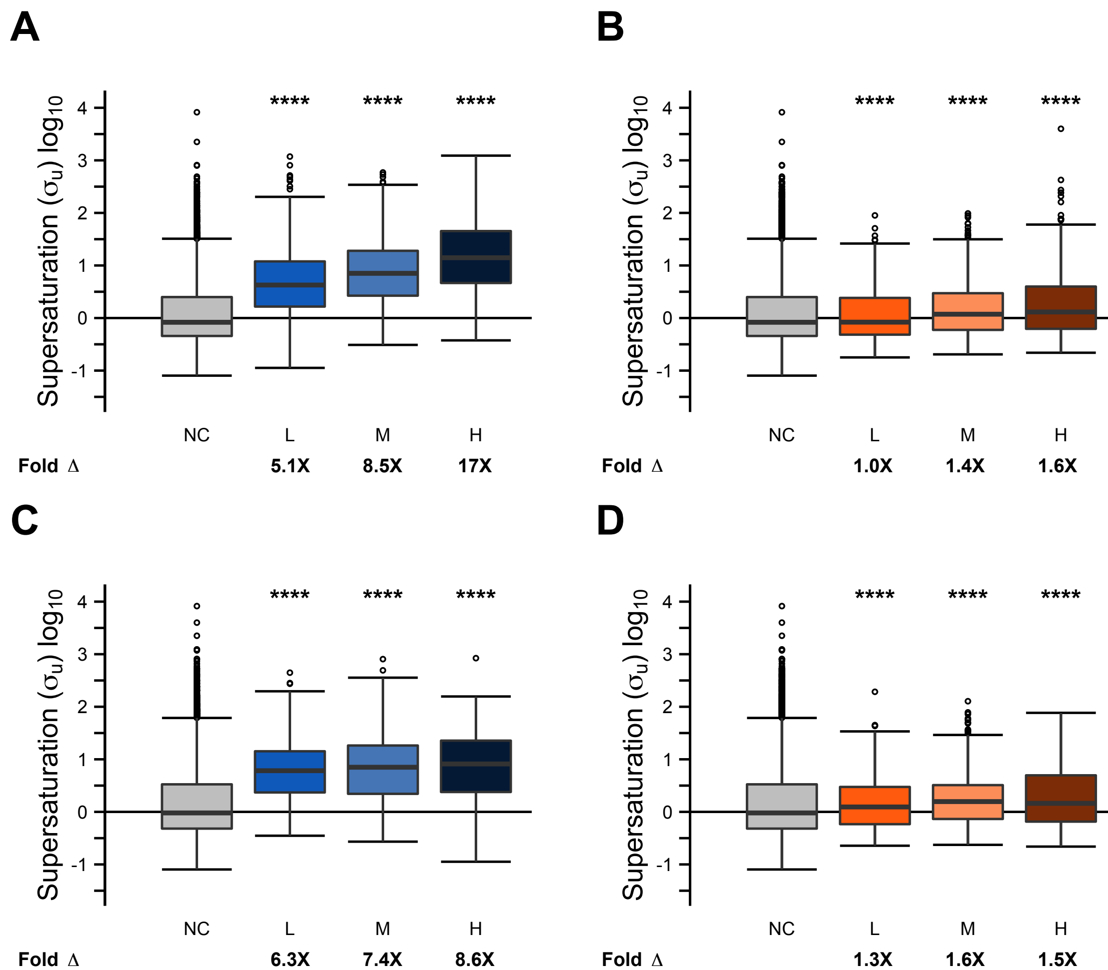


Figure 2.3 The metastability of proteins to aggregation is correlated with the downregulation of the corresponding genes in AD. Metastability levels, assessed by supersaturation scores, for proteins associated with differentially expressed genes: (A) downregulated in AD, (B) upregulated in AD, (C) downregulated in aging, and (D) upregulated in aging. Differentially expressed genes are divided into thirds (Low 'L,' Medium 'M,' High 'H') based on the fold change of expression. The median fold difference in supersaturation is indicated by 'Fold Δ'. 'NC' indicates no change in expression. ****$p \leq 0.0001$, one-sided Wilcoxon/Mann-Whitney test with Holm-Bonferroni correction.

We then investigated whether or not the cell is endowed with transcriptional mechanisms to regulate the solubility burden in register with the protein homeostasis capacity. If so, there may be transcriptional regulators that coordinate such a response by modulating protein homeostasis. To determine in particular if specific transcription factors and histone modifiers are upregulated or downregulated in AD and aging, we generated a map of transcriptional regulators and their targets using ENCODE's regulator binding site data (237). Here, we considered a gene to be regulated by a

particular transcription factor or histone modifier if the regulator has a binding site at least half of which is within 1000 base pairs of the start codon of the gene itself. We identified 23 transcription factors and histone modifiers associated with a significant number of genes downregulated in AD, including EGR1 (238), NRF1 (239), and REST (240). By contrast, we found only one regulator associated with a significant number of genes downregulated in aging, the histone modifier EZH2. In addition, four regulators were found to be associated with a significant number of genes upregulated in AD and none was found to be associated with a significant number of genes upregulated in aging.

2.3.5 Biochemical pathways enriched in metastable proteins are also enriched in proteins corresponding to genes downregulated in AD

To determine the functional implications of the transcriptional regulation of metastable proteins in AD, we conducted an unbiased search of the entire set of 284 pathways in the KEGG database (241), a repository of biochemical pathways and protein networks. We found a close correspondence between the pathways downregulated in AD and pathways that we previously found to be supersaturated based on independent data (33, 34) (**Figure 2.4**). Remarkably, most of these KEGG pathways fall along a band in which increasing metastability levels correspond to increasing downregulation (**Figure 2.4**, purple circles). The overlap between metastable and downregulated pathways is highly significant ($p=8.7\cdot10^{-11}$). Among the simultaneously metastable and downregulated KEGG pathways, we found oxidative phosphorylation (OP), Parkinson's disease (PD), Huntington's disease (HD), Alzheimer's disease (AD), non-alcoholic fatty liver disease (NAFLD), cardiac muscle contraction (CMC), nicotine addiction (NA), GABAergic synapse (GABA) and pathogenic *E. coli* infection (PEcI). These results identify pathological (AD, PD, HD and NAFLD) and functional (OP, CMC, PEcI) networks and pathways enriched in physiological complexes, as well as pathways involved in neuronal signalling (NA, GABA). In particular, our analysis identified certain proteins in the oxidative phosphorylation pathway as being particularly metastable, including all the components of the mitochondrial ATP synthase complex for which we have data,

consistent with the reported involvement of this complex in AD (242). In addition, 43% of the genes in our analysis that encode for mitochondrial ATP synthase complex are transcriptionally repressed. The most repressed is the alpha subunit of the F_1 catalytic core (whose expression is reduced by 26% in AD), which has been observed to accumulate in degenerating neurons in AD and to be associated with neurofibrillary tangles (243). We also verified that, although oxidative phosphorylation is central to the pathways downregulated in AD, the signal for metastability in AD and aging is robust against the exclusion of proteins in this pathway from our analysis (**Figure A.7**).

In this comprehensive analysis of KEGG pathways, we also found other pathways that are significantly enriched in either metastable proteins (**Figure 2.4**, red circles) or in proteins corresponding to downregulated genes (**Figure 2.4**, blue circles), but not both. However, the large majority of these pathways have significance values that are lower than the average jointly metastable and downregulated pathway (**Figure 2.4**, purple circles), the exceptions being the 'ribosome,' which is highly metastable but not downregulated, and the 'synaptic vesicle cycle', 'proteasome' and 'retrograde endocannabinoid signalling', which are downregulated but not metastable. A similar analysis for upregulated pathways in AD did not provide particularly significant results, although one may expect genes associated with the immune response to be upregulated, as for example complement C1q subcomponent subunit C and plasma protease C1 inhibitor in the 'complement and coagulation cascade' pathway.

Thus, the observation that in AD there is a highly specific downregulation of metastable biochemical pathways and networks suggests the presence of a robust transcriptional response to protein aggregation in AD.

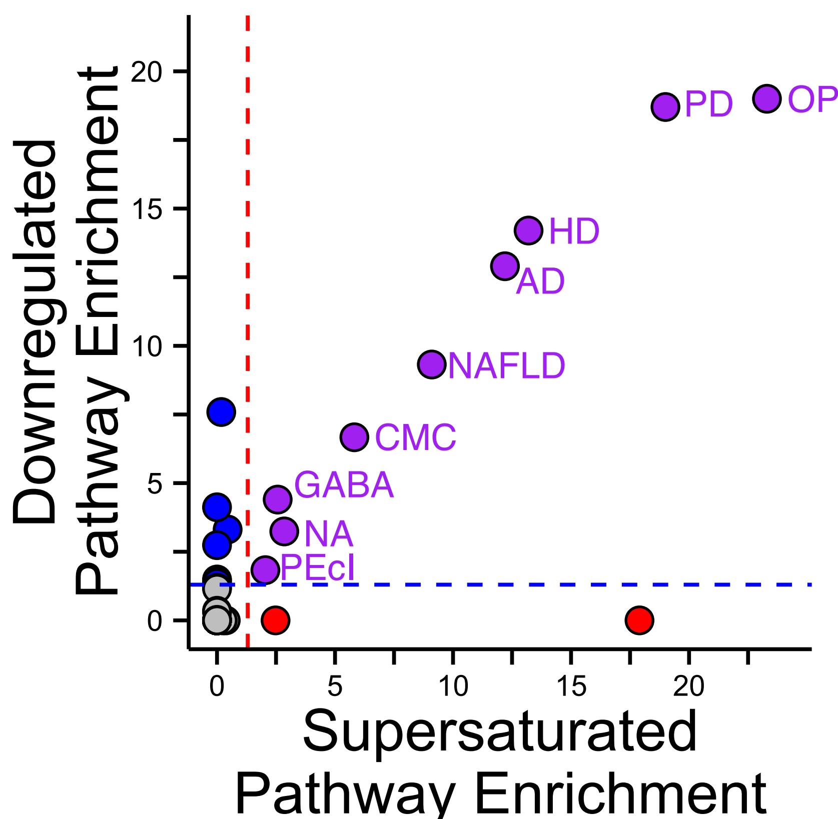


Figure 2.4 Comparison between downregulated and metastable biochemical pathways and networks. We found that the biochemical pathways and networks downregulated in AD correspond closely to those enriched in supersaturated proteins (purple circles). Using the KEGG classification, these biochemical pathways and networks are: 'oxidative phosphorylation' (OP), 'Parkinson's disease' (PD), 'Huntington's disease' (HD), 'Alzheimer's disease' (AD), 'non-alcoholic fatty liver disease' (NAFLD), 'cardiac muscle contraction' (CMC), 'nicotine addiction' (NA), 'GABAergic synapse' (GABA) and 'pathogenic E. coli infection' (PEcI).

2.3.6 Widespread downregulation of the metastable subproteome is not a general feature of disease

Because the genes corresponding to the metastable subproteome are, on average, highly expressed, we considered the possibility that their widespread downregulation could be a general feature of cellular dysfunction in disease. It this were the case, any process that disrupts normal cellular function could impair transcription, preferentially affecting those genes that are highly expressed. To investigate this possibility, we

performed a meta-analysis of expression changes in another cognitive disorder, clinical depression. We considered 470 microarrays, including 239 from control patients and 231 from those with clinical depression (**Table A.1**). As with our analysis of AD, we restricted our analysis to brain samples from cases in which the gender and age (for which we controlled) were known, and GEO database series that included at least 10 total cases. Among the 19,190 genes for which we evaluated changes in expression, we found 7 genes downregulated and 11 genes upregulated in clinical depression at the thresholds of 10% change in expression and $p \leq 0.01$. Overall, we did not observe the same widespread transcriptional repression of the metastable subproteome found in AD, and we found no KEGG pathways significantly enriched in proteins corresponding to those genes differentially expressed in AD.

We then considered the possibility that we had only identified a small number of genes as being differentially expressed in clinical depression because of low statistical power. Although our meta-analysis of clinical depression included only 22% as many arrays as that of AD, this is unlikely to explain the fact that only 0.6% as many genes are differentially regulated in clinical depression. In addition, our separate analysis for aging provided a control to assess the statistical power of the clinical depression dataset relative to that for AD. At the thresholds of 10% change in expression and $p \leq 0.01$, we found 196 genes downregulated and 122 genes upregulated in aging in the clinical depression dataset. This is 23% as many genes as we found differentially regulated in aging based on the AD dataset, consistent with the smaller number of microarrays in the clinical depression analysis. As further control, we reanalysed these data after relaxing the significance threshold for differential expression to $p \leq 0.05$. At this threshold, we found 24 genes downregulated and 17 genes upregulated in clinical depression, and 569 genes downregulated and 291 genes upregulated in aging. At the relaxed threshold, the KEGG pathway for 'olfactory transduction' was enriched in proteins corresponding both to genes downregulated ($p=4.5 \bullet 10^{-3}$) and genes upregulated ($p=4.9 \bullet 10^{-2}$) in clinical depression. Only 'mineral absorption' was enriched in proteins corresponding to genes upregulated in aging in the clinical depression dataset. We also assessed the overall relationship between metastability and transcriptional regulation, and found little correlation between the two.

2.4 Discussion

A major area of investigation into the molecular origins of AD concerns the chemical and physical instability of the proteins associated with pathology, and the mechanisms by which the cell responds to such a situation. A number of studies have reported biophysical features, environmental conditions, and molecular partners that promote or repress the initial aggregation of specific proteins (7, 16, 49, 172-174). More recently, it has been recognised that the regulation of many other proteins is disrupted as a consequence of these initial aggregation events (29-31, 33, 34, 178, 179, 212-215). In a complementary approach, the origins of AD have been studied by analysing the transcriptional response associated with its onset and progression (219-233). These studies have revealed that this transcriptional response involves genes corresponding to proteins that cause the disease and those associated with the cellular processes engaged in combating it (219-233).

In the present study we have brought together these two approaches, finding that the transcriptional changes that occur in AD can be rationalized, at least in part, on the basis of the presence of an AD-specific 'metastable subproteome' at risk of aggregation (**Figure 2.2**). This metastable subproteome is defined as the overlap between the proteins that are most supersaturated and that correspond to highly expressed genes, and those encoded by genes most transcriptionally downregulated in aging and in AD (**Figure 2.2D**). These proteins are intrinsically at risk of aggregation and, as we found here, tend to be the target of the transcriptional response in aging and AD. These results are consistent with previous observations that the expression of oxidative phosphorylation genes is suppressed in AD (244, 245), but suggest in addition that that such suppression may be part of a broader response to the disease.

Having previously shown that the proteins associated with AD tend to be metastable to aggregation because they are supersaturated (33, 34), we have now reported a response to this intrinsic metastability of the proteome in the face of disruptions to protein homeostasis through the transcriptional downregulation of their respective genes. The close correspondence of the biochemical pathways associated with metastability and those downregulated in AD (**Figure 2.4**) supports this conclusion,

as do the tendency for plaque and tangle co-aggregators to correspond to downregulated genes (**Figure 2.1**) and the high overall metastability level of proteins encoded by downregulated genes (**Figure 2.2**). We found these results to be stable against a range of potentially confounding factors, including the choice of thresholds for differential expression (**Figures A.1 and A.2**), noise in the supersaturation score (**Figures A.5 and A.6**), and the large contribution of oxidative phosphorylation (**Figure A.7**).

Analysis of the transcriptional response to the collapse of protein homeostasis in terms of a metastable subproteome at risk of aggregation has also enabled us to address another central question about the progression of AD, namely the way in which changes occurring in this disease are related to the natural process of aging. These results indicate that aging and AD are very different at the transcriptional level, as over three quarters of the transcriptional changes that occur in AD do not occur in aging (**Figure 2.2D,E**). In addition, many cellular processes downregulated in AD are not significantly downregulated in aging (**Figure A.2**). While the differences between regulation in aging and AD are profound, there are important commonalities, as shown by the significant overlap in the specific transcriptional changes that occur in AD and in aging (**Figure 2.3**). AD therefore appear to involve an acceleration in the decline of protein homeostasis associated with aging, and also an extension of its scope and significance. Overall, such acceleration makes the metastable subproteome that we have identified in this work more susceptible to aggregation. This conclusion offers an explanation of why a transcriptional downregulation of genes corresponding to metastable proteins is observed in both aging and AD.

We also observe that these phenomena are unlikely to be a general feature of cellular dysfunction. Our results indicate that a different transcriptional response is present in the case of clinical depression, consistent also with results for epilepsy derived considering the differentially expressed genes in hippocampal samples from five patients with mesial temporal lobe epilepsy (MTLE) with hippocampus sclerosis (HS) (246). In that study, 518 genes were found to be differentially expressed between the subjects. Functional enrichment using DAVID showed enrichment for KEGG pathways associated with neuroactive ligand receptor interaction, drug metabolism and cytokine interaction, among others. The KEGG pathways of oxidative

phosphorylation, and of Alzheimer's, Parkinson's and Huntington's diseases were not, however, seen in epilepsy.

The findings that we have reported here, therefore, suggest that the widespread downregulation of genes corresponding to metastable proteins at risk of aggregation may represent a step in the strategy for cellular regulation in the face of disruptions in protein homeostasis. More generally, understanding the physicochemical implications of transcriptional regulation in aging, AD and other protein misfolding disorders has important implications both for a fundamental biological understanding of the origins of the disease and for clinical practice. Since the maintenance of protein homeostasis is an essential function in the cell, determining how the overall proteome composition is managed and modulated is a central question in biology. At the same time, understanding endogenous strategies for handling supersaturated, metastable and potentially misfolding proteins may provide an avenue for improved therapies. If widespread aggregation is associated with AD, then determining how to regulate this phenomenon is of great value and practical importance.

2.5 Conclusions

We have shown that AD is associated with the transcriptional regulation of a metastable subproteome at risk of aggregation. The presence of these poorly soluble proteins in the cellular environment is inherently dangerous, in particular because these proteins tend to cluster into specific biochemical pathways, and that only limited molecular chaperones and other protective resources are available at any given time to prevent their misfolding and aggregation. In conjunction with new insights into the molecular chaperone functions and the regulation of protein translation and degradation, our results indicate that the study of protein metastability may clarify how failures in maintaining proteins in their normal functional states could result in protein aggregation and in multifactorial disorders such as AD.

Despite the great complexity of aging processes and neurodegenerative disorders, protein solubility may underlie many aspects of their resultant cellular dysfunction. In

this work we have adopted this idea to investigate how the levels of poorly soluble proteins are regulated, finding that the overall transcriptional response to AD is associated with a global downregulation of the expression of the genes encoding proteins that are metastable to aggregation. We anticipate that interventions that target the metastable subproteome at risk of aggregation that we have identified in this work may provide novel opportunities for the early diagnosis and treatment of AD.

2.6 Materials and Methods

2.6.1 Array normalization

We performed array normalization using the limma, affy, gcrma, and text itxps packages for the statistical programming language R. Affymetrix gene arrays read using the ReadAffy function in affy were normalized using the GC Robust Multi-array Average (GCRMA) using the gcrma function, which uses estimates of cross-hybridization based on the GC content of mismatch (MM) probes. It is a modification of the robust multi-array average (RMA) algorithm using the rma function in xps, which was used here to normalize Affymetrix exon arrays. GCRMA cannot be used on these arrays because they do not have the MM probes needed to estimate cross-hybridization. We read Illumina arrays using the read.ilmn function in LIMMA and background corrected with the neqc function in LIMMA. Pre-processed arrays, if they have accompanying significance values, were filtered to remove those with a significance score <0.95 (where significance scores were available), and expression values scaled to $\log_{10}$. We treated each two-colour array that we encountered in the clinical depression dataset as two separate arrays, analysing them using the backgroundCorrect, normalizeWithinArrays, and normalizeBetweenArrays (by the Aquantile method) functions. One other significant difference between our analysis of two-colour arrays and one-color arrays is that when performing backgroundCorrect, we used the 'minimum' method for two-colour arrays while we used the 'normexp' method for one-color arrays. The reason for the use of the 'minimum' method here is to eliminate negative values, which are not compatible with the normalization we perform to correct for the fact that each channel in a given two-colour array is on a

single chip exposed to the same sample. For the clinical depression meta-analysis, we grouped certain series together.

2.6.2 Construction of the linear model

We fitted a linear model to the expression of each gene that included the co-factors of tissue type, gender, age, and disease status in all cases, and subject ID and technical replication when these were relevant. If there was technical replication, we used the function duplicateCorrelation in limma to account for this replication. From LIMMA, we then used the function lmFit to generate the fit, and the function eBayes to generate statistical significance values.

2.6.3 Determination of significance and magnitude values

We obtained adjusted p-values (i.e., q-values) using the Benjamini-Hochberg method (247). In addition, we obtained the coefficients for the disease status and age co-factors to estimate the magnitude of the contribution of these parameters to gene expression. We converted Probe IDs to human reviewed UniProt IDs, based on a mapping of those probes that unambiguously mapped to a single UniProt ID. If multiple probes mapped to a single UniProt ID, we used the median parameter p-value and coefficient. Because aging is a continuous variable, the magnitude of the expression change attributable to aging is the product of the aging cofactor coefficient and some age range. In this study, we used an age range of 25 years for two reasons. First, this equals approximately two standard deviations of the age distribution, which is a reasonable range over which to assume linearity. Second, this value reflects an age range from about 63 years to about 88 years, a period over which the prevalence of AD increases dramatically.

For the clinical depression meta-analysis, the ages of the control samples were mean 50.3 ± 12.8 years and those of the disease samples were mean 49.4 ± 15.6 years. We used an age range of 25 years for these samples, as well, for consistency and because this was also approximately two standard deviations of the age distribution in the

clinical depression meta-analysis. As described below, varying the magnitude threshold significance of aging has the same effect as changing the age range used for our analysis, and the results are robust against such changes.

2.6.4 Combination of significance and magnitude

There are several methods to combine significances across a series of studies. Here, the question was whether or not the change in expression of a gene corresponding to a given protein is statistically significant. In general, we had 10 p-values because we analysed 10 microarray studies separately, although in some cases there were fewer than 10 p-values because some genes are represented in some arrays but not others. The goal was to estimate the probability of obtaining a set of co-factor coefficients in each of the studies that are at least as extreme as those observed, assuming the null hypothesis that there is no change in gene expression is valid. One way to address this issue is to combine p-values from various studies. Perhaps the most popular method to combine p-values is Fisher's method, which in essence yields a significant result if at least one of the studies can reject the null hypothesis (248). By contrast, Pearson's method can be interpreted as assessing a result as insignificant if at least one of the studies fails to reject the null hypothesis (249). Stouffer's method is attractive because it is somewhat less sensitive to extreme values (250). In this method, p-values are first converted to Z-scores, which are standard normal variables. These Z-scores can be combined to give a composite Z-score, based on the property that the sum of k standard normal variables has mean 0 and variance $p(k)$. This sum can then be converted unambiguously back into a p-value. A common variant of this method was proposed by Liptak and involves weighting each individual Z-score by the sample size of the study, an approach that has been shown to be superior by simulation. In the current analysis, we used Liptak's method (241). This method requires that the p-values be one-tailed, and although the p-values obtained from the LIMMA functions lmFit and eBayes are two-tailed, they can be converted into one-tailed p-values. To obtain a combined magnitude, we used the median of magnitudes per co-factor per gene.

2.6.5 Calculation of basal expression levels for supersaturation scores

We estimate metastability using our previously defined supersaturation scores, and estimated basal mRNA expression in control subjects. Because of normalization differences, it is challenging to obtain values for basal expressions by combining data from different studies, and so we selected a single study to obtain these levels. For AD, the study GSE44772 (**Table A.1**) included the most control samples (299), but employed the Rosetta/Merck Human 44k 1.1 microarray (232), in which expression values reported for each array are relative to the expression of a pooled background array, thus making between-gene comparisons impossible. The study GSE1297 (**Table A.1**) used the Affymetrix Human Genome U133 array (220), which reports raw array expression values that we are then able to renormalize. This Affymetrix array is also the most commonly used array in the GEO database among those arrays represented in this analysis. The study GSE1297 also has a relatively large number of control samples (74), although this number is smaller than that available in the study GSE44772. However, given that the Affymetrix platform is more common, better characterized, and amenable to re-normalization within this analysis, we estimated basal expression levels from the control expression values in GSE1297. These values are the log2-average of all the samples, as obtained from the LIMMA function lmFit. For clinical depression, the GSE54562/GSE54563/GSE54564 set of series included the most control samples (56,63), but GSE53987 had the advantage of deriving all 55 of its samples from the same series. It also used a similar platform to that for basal expression AD, Affymetrix Human Genome U133 Plus 2.0 Array (251).

We also note that the use of proteome-level mass spectrometry (32), when applied to brain tissues, could provide a quantitative way to measure supersaturation levels directly as the ratio between the actual soluble and total amounts of individual proteins observed *in vivo*

2.6.6 Multiple hypothesis correction

In addition to the multiple hypothesis correction described above, the following families of tests were corrected using the Holm-Bonferroni method (252): 1) Overlap of aging and disease downregulation with most supersaturated proteins, 2) Overlap of aging and disease upregulation with most supersaturated proteins, 3) Enrichment of disease-related upregulated and downregulated proteins in disease-related amyloid proteins, plaque co-aggregators, tangle co-aggregators, the most supersaturated proteins (included in this family and family 1 above) and the proteostasis network, 4) Enrichment of age-related upregulated and downregulated proteins in disease-related amyloid proteins, plaque co-aggregators, tangle co-aggregators, the most supersaturated proteins and the proteostasis network, 5) Supersaturation scores of proteins upregulated and downregulated in disease, 6) Supersaturation scores of proteins upregulated and downregulated in aging, 7) Supersaturation scores of proteins upregulated and downregulated in disease divided into low, medium and high categories, 8) Supersaturation scores of proteins upregulate and downregulated in aging divided into low, medium and high categories, 9) Disease-related downregulation of subcategories of the protein homeostasis network, and 10) Overlap of KEGG pathways for upregulation and downregulation in aging and disease. KEGG pathway enrichment was corrected using the Holm-Bonferroni method (252), considering aging upregulation, aging downregulation, disease upregulation and disease downregulation each as a separate family. Transcription factor target enrichment was corrected using the Benjamini-Hochberg method, considering aging upregulation, aging downregulation, disease upregulation and disease downregulation each as a separate family. Analyses that excluded oxidative phosphorylation genes were considered as separate families. Analyses of clinical depression and AD were included in separate families.

2.6.7 KEGG analysis

KEGG analysis was performed by first assembling a database of the components of each KEGG pathway (241) from publicly available data. The KEGG gene identifiers were then converted to UniProt ID to make it possible to compare them to the rest of

our data. Enrichment was calculated using a one-sided Fisher's exact test (248) and corrected using the Holm-Bonferroni method (252). Results for the most metastable proteins made use of previously published supersaturation scores, but two aspects of the current analysis of that data differed. First, the new method of deriving KEGG pathways resulted in the analysis of 85 KEGG pathways not analysed in the previous study. Second, the new method used a one-sided Fisher's exact test instead of the modified EASE score to calculate significance. This resulted in some differences in the pathways identified as being enriched in metastable proteins.

2.6.8 Overlap analysis

The significance of the overlaps between aging, AD, and metastability were calculated using a one-sided Fisher's Exact Test, corrected using the Holm-Bonferroni method. For the significance of the triple intersection, the p-value was estimated as being less than or equal to the minimum p-value of any double overlap.

2.6.9 Transcription factor analysis

We used transcription factor binding site data from the ENCODE database (237) to identify transcription factors whose targets are enriched in the genes that we identified as being differentially expressed in aging and AD. ENCODE provides the genome address for binding sites for each transcription factor (237). We defined a gene as being regulated by a transcription factor if its binding site was less than 1000 nucleotides upstream of its start codon. Using this method, we generated a map of transcription factors and their targets. We converted the identifiers for the target genes to human-reviewed UniProt ACs and did the same for the UniProt IDs in our expression analysis. We then used a one-sided Fisher's exact test to determine the significance of enrichment, correcting this p-value using the Benjamini-Hochberg method (247).

2.6.10 Threshold sensitivity analysis

To test the sensitivity of our results for aging and AD to variations in the threshold for differential expression, we varied the expression change threshold between 0.5% and 50% and the significance threshold between $p=10^{-20}$ and $p=1$, for a total of 18,100 combinations of thresholds. At these thresholds, we determined which genes were upregulated, downregulated, or unchanged in expression. For those threshold levels at which there were at least 5 genes in each category, we then recalculated the median fold difference in supersaturation between the proteins encoded by upregulated/downregulated genes and those encoded by genes unchanged in expression, as well as the corresponding statistical significance. Because the aging results scale linearly with the age range selected, changing the magnitude threshold for aging has the same effect as varying the width of the age range used to calculate expression changes in aging.

2.6.11 Sensitivity to Gaussian noise in the supersaturation score

In a method similar to that we previously described (34), we introduced random error into the supersaturation scores we calculated, drawn from 34 increasingly wide Gaussian distributions with standard deviation ranging from 1.1X error to 100X error. At each level, we performed 100 independent trials. At each level, we calculated the median fold difference in supersaturation and the corresponding significance, and then performed a one-sided Wilcoxon/Mann-Whitney test on these sets of median fold differences and p-values to assess whether they were significantly greater than 1 or less than 0.05, respectively.

CHAPTER 3

---

3. Protein homeostasis of a metastable subproteome associated with Alzheimer's disease

3.1 Overview

As described in Chapters 1 and 2, Alzheimer's disease is a neurodegenerative disorder associated with protein aggregation, whose origins have been linked to the dysregulation of a set of highly expressed and aggregation prone proteins that make up a metastable subproteome. Under normal conditions, the protein homeostasis system prevents effectively protein aggregation by controlling these metastable proteins (Section 1.3). Although it is well established that such regulatory mechanisms become progressively impaired with ageing, resulting in an accumulation of protein deposits, the specific nature of such impairment remains to be fully defined. Through a gene coexpression analysis, here we identify the endosomal-lysosomal and ubiquitin-proteasome systems, and more generally the protein trafficking and

clearance mechanisms, as key components of the protein homeostasis system that maintains the metastable proteins in their functional states (182)[i].

3.2 Introduction

Neurodegenerative diseases are highly complex disorders characterised by extensive neuronal dysfunction associated with protein misfolding and aggregation (3-12, 15-17). A feature common to essentially all of these conditions is the presence of abnormal protein deposits, including amyloid plaques and neurofibrillary tangles in Alzheimer's disease, and Lewy bodies in Parkinson's disease (3-12, 15-17). It is increasingly recognized that the formation of such deposits, rather than being an unusual process involving only a small number of proteins, may represent a widespread phenomenon (11, 12), with hundreds of different proteins found to aggregate under stress conditions, in ageing or in disease (7, 29-32, 179, 213) (Section 1.6).

To rationalize these observation, it has been recently shown that a large number of proteins are inherently supersaturated in the cellular environment (33, 34), as they are expressed at concentrations higher than their solubilities (47, 49), and therefore constitute a metastable subproteome potentially susceptible to aggregation (31-34). It has also been observed that proteins that have been reported to co-aggregate with plaques, tangles and Lewy bodies, tend to be supersaturated (33, 34). Therefore, despite their heterogeneous and multifactorial nature, neurodegenerative conditions, including Alzheimer's, Parkinson's and Huntington's diseases, and amyotrophic lateral sclerosis (ALS), share the important common attribute of protein supersaturation (33, 34, 253).

Given the intrinsic propensity of proteins to aggregate, it is not surprising that we are endowed with a powerful array of defense mechanisms whose role is to preserve protein homeostasis by helping to maintain proteins in their soluble states and to promote the degradation of those that misfold and aggregate (5-8, 35, 69, 70, 212,

---

[i] The findings reported in this chapter are based on a published manuscript (182).

254). The progressive decline of the efficacy of these regulatory processes upon ageing is likely to contribute to the increased susceptibility of the elderly population to age-associated neurodegenerative disorders (5-8, 35, 69, 70, 95, 96, 212, 254, 255) (Section 1.3).

As discussed in Chapter 2, the proteins within the metastable subproteome that are also transcriptionally downregulated in Alzheimer's disease may be particularly significant for the pathology of this disorder (181), it is important to determine the detailed mechanisms of their regulation by the protein homeostasis system. Our goal here, therefore, is to identify the specific components of the protein homeostasis system that regulate a recently identified metastable subproteome associated with Alzheimer's disease (181). To achieve this goal we adopted the strategy of determining the association between groups of genes by probing their genetic interactions, an approach that is based on the observation that many functionally related genes are coexpressed (256, 257). For example, genes encoding for the various different components of a protein complex tend to have similar expression patterns 30, (256, 258), and, if groups of genes are regulated by common mechanisms, then they may be expected to be coexpressed (258).

We have therefore constructed a weighted gene correlation network (259, 260) of this metastable subproteome and of the overall protein homeostasis system (212) to gain a systems-level understanding of the transcriptional relationship between these two sets of proteins. By following this approach, we have identified the protein homeostasis components corresponding to the metastable subproteome specifically associated with Alzheimer's disease (181). Our results show that the genes corresponding to this metastable subproteome are tightly coexpressed with specific components of the ubiquitin-proteasome and the endosomal-lysosomal pathways, thereby suggesting that metastable proteins with a high risk of aggregation tend to be closely regulated by the trafficking and degradation machineries.

3.3 Results

3.3.1 Protein homeostasis of a metastable subproteome associated with Alzheimer's disease.

Alzheimer's disease is associated with widespread transcriptional changes (95, 96, 220, 231, 261, 262), which can be rationalized in part by the presence of a set of aggregation-prone proteins in the proteome (181). Many of the proteins involved in this metastable subproteome are components of the mitochondrial respiratory chain, an observation consistent with the well-characterized mitochondrial disruption associated with neurodegenerative disorders and specifically with Alzheimer's disease (263). We refer to the proteins expressed by this subset of genes as the 'AD metastable subproteome'. The primary aim of the present study is to understand the different ways in which these metastable proteins are controlled, as illustrated schematically in **Figure 3.1**. Our goal is thus to identify the specific components of the protein homeostasis system that are most closely involved in the regulation of the AD metastable subproteome.

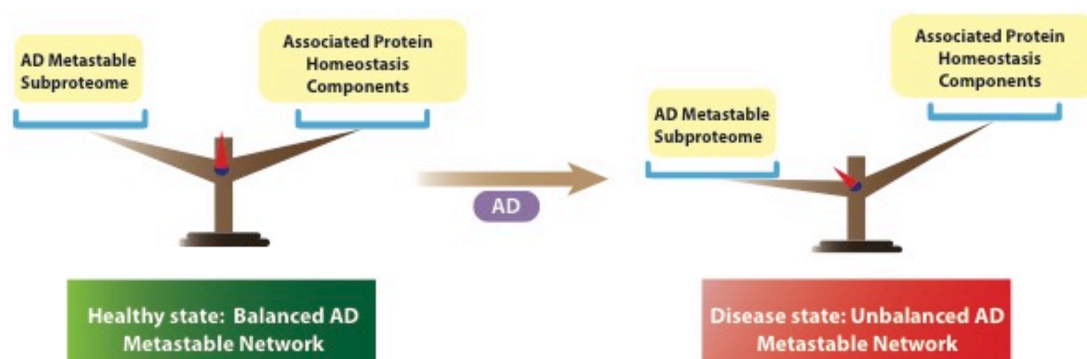


Figure 3.1 Protein homeostasis of a metastable subproteome associated with Alzheimer's disease. In the healthy state the metastable subproteome associated with Alzheimer's disease (denoted 'AD metastable subproteome') is effectively regulated by a series of protein homeostasis mechanisms (denoted 'Associated protein homeostasis components'). In a disease state, this balance is compromised and protein misfolding and aggregation results in the widespread formation of aberrant deposits.

3.3.2 Coexpression analysis of the AD metastable subproteome and its associated protein homeostasis components.

Since the proteins in the AD metastable subproteome are intrinsically aggregation-prone, we searched for the specific protein homeostasis components that maintain the solubility and folding of these proteins. We therefore set out to identify an 'AD metastable network' as a network of genes that encode for: (1) the AD metastable subproteome, and (2) its associated protein homeostasis components (**Figure 3.1**).

In order to identify the AD metastable network, we carried out a weighted gene correlation network analysis (WGCNA) (259, 260) of the set of metastable proteins that we previously identified (181) and of the known components of the overall protein homeostasis system (212) (**Section 3.6**). WGCNA is a robust method of performing gene coexpression analysis that has been shown to be particularly effective when large microarray datasets are available (259). As our aim was to study how metastable proteins are regulated across health and disease, we pooled together extensive microarray data obtained from post mortem brain tissues of patients diagnosed with late-onset Alzheimer's disease and of matched controls (232) (**Table B.1** and **Section 3.6**). WGCNA uses the Pearson's coefficient of correlation between each pair of genes and their 'topological overlap', which is a measure of their connectivity based on their shared neighbors, to identify biologically meaningful groups of coexpressed genes; these groups are called 'modules' and labeled by different colors (260) (**Table B.2**, and **Section 3.6**).

As WGCNA captures the underlying network structure in large-scale gene expression studies, it has been used to study the global changes associated with a range of disease states, with the preservation of groups of coexpressed genes across species,

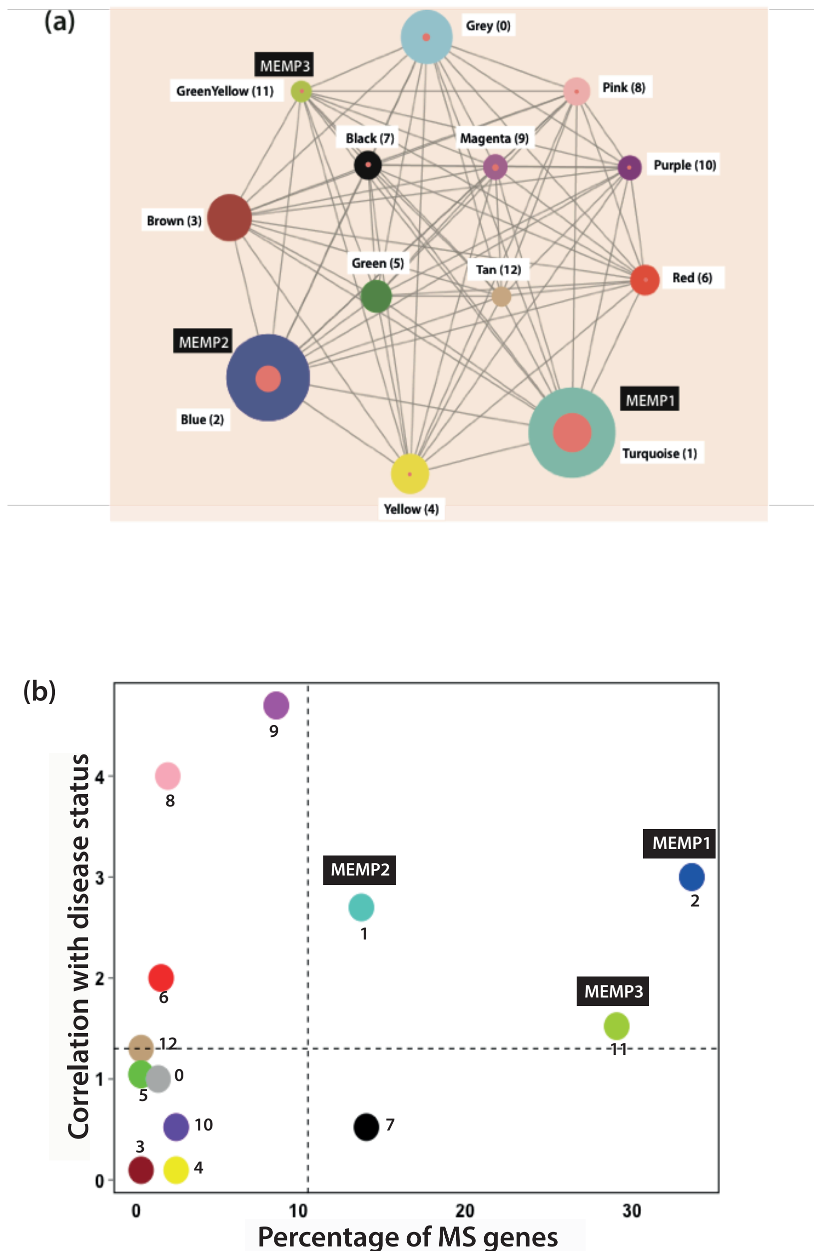


Figure 3.2 Identification of the AD metastable network by a coexpression analysis of the AD metastable subproteome and its associated protein homeostasis components. (a) By using a weighted gene correlation network analysis (WGCNA) we carried out a hierarchical clustering of genes on the basis of their topological overlap. Modules of coexpressed genes are labeled by numbers and shown in different colors, with the size of the circles corresponding to the number of genes in the module. The number of genes encoding for the metastable subproteome is highlighted in pink within each module (Table B.2). The vast majority of metastable proteins are found in three specific 'modules enriched in metastable proteins' (MEMPs), which are referred to as MEMP-1 (Blue, 220 metastable proteins), MEMP-2 (Turquoise, 91 metastable proteins), MEMP-3 (GreenYellow, 10 metastable proteins) and MEMP-4 (Black, 10 metastable proteins); each of the other modules had between 0 and 4 metastable proteins. (b) Analysis of module eigengenes.

Each module is represented by a circle, which is labelled by number and shown in colour, as in panel (a). The x-axis is the percentage of metastable genes in each module and the y-axis the negative log10 of the p-values for the correlation of each module with the disease status (Section 3.6). The horizontal dashed line marks a p-value of 0.05 and the vertical dashed line marks the 10% value. Hence we identify the three modules shown also in panel (a), MEMP1, MEMP2 and MEMP3, as the only modules that have a high percentage of metastable proteins and significant correlation to disease status.

and with the identification of hub genes associated with particular traits (95, 262, 264-267). We observed that the genes encoding for metastable proteins and protein homeostasis components are organized into well-defined modules (**Figure 3.2a**, with the genes encoding for proteins in the metastable subproteome shown in pink in each module), where each module consists of tightly coexpressed genes. We found that the majority of metastable proteins belong to four specific modules, and we refer to them as 'modules enriched in metastable proteins' (MEMPs) - MEMP-1 (Blue), MEMP-2 (Turquoise), MEMP-3 (GreenYellow) and MEMP-4 (Black) **(Figure 3.2a)**, which consisted of 659, 688, 35 and 74 genes, with 220, 91, 10 and 10 genes corresponding to metastable proteins, respectively **(Table B.2)**. We also found several clusters containing no or very few metastable proteins (4 at most); such clusters consisted of tightly coexpressed groups of protein homeostasis genes whose expression levels do not correlate well with the metastable proteins.

3.3.3 Identification of an AD metastable network.

As the coexpression analysis described above revealed the existence of several distinct modules, we asked how the modules are related to each other, as it is likely that closely related modules are functionally related. To this end, we performed an analysis to identify a module eigengene (ME) for each module, which is the first principal component (PC) of the expression values across genes in each module (**Section 3.6**). The ME therefore provides a representative value for the expression of a group of genes in a particular module (260). This approach offers a significant advantage in correlating two modules, as it eliminates the problem of multiple testing and noise by reducing the number of comparisons to just one instead of several

hundreds. The higher the value of the Pearson's coefficient of correlation between two MEs, the more closely the two modules are related (**Section 3.6**).

We therefore identified which of these modules showed the most significant relationship to the disease status by looking at the correlation between the MEs and disease status (**Section 3.6**). We thus found six modules (**Figure 3.2b**) to be significantly correlated with disease status, of which three modules (Magenta, Pink and Red) had very few metastable genes (4, 1 and 1, respectively). We excluded these modules from further analysis since they mostly contained components of the protein homeostasis machinery whose expression levels do not correlate well with the metastable genes. The other three modules (MEMP-1, MEMP-2 and MEMP-3), perhaps not surprisingly, were those most enriched for metastable genes that we described in the previous (**Figure3.2b**).

As these three MEMPs were the only modules both significantly correlated with disease status and significantly enriched for genes encoding for metastable proteins (**Figure 3.2b**), we chose them for further analysis. They are also, in fact, closely related to each other based on the correlation of their MEs with a Pearson's correlation coefficient of 0.78 between MEMP-1 and MEMP-2 and 0.68 between MEMP-2 and MEMP-3 **(Table B.3)**. In order to control for possible biases of the modules because of the use of a particular dataset, we cross-validated the results of module detection with a hippocampal gene expression dataset as the hippocampus is among the regions typically affected in Alzheimer's disease (220). We observed that the MEMPs, along with most of the other modules, were well preserved between the two datasets (**Figure B.1**).

We then performed a gene ontology enrichment analysis to characterize these modules, finding that protein ubiquitination was the most enriched GO term for MEMP1, MEMP2 and MEMP3 (**Figure B.2**). We also asked if the genes contained in these modules are overrepresented in any biochemical pathway. To this end, by analyzing the KEGG biochemical pathways (268), we found that they are strongly overrepresented in the pathways associated with the AD metastable subproteome and the ubiquitin-proteasome and endosomal-lysosomal systems (**Figure B.3**).

Based on these results, we identified the 'AD metastable network' as the set of genes in MEMP1, MEMP2 and MEMP3.

3.3.4 Identification of the hub genes and of their roles in the AD metastable network.

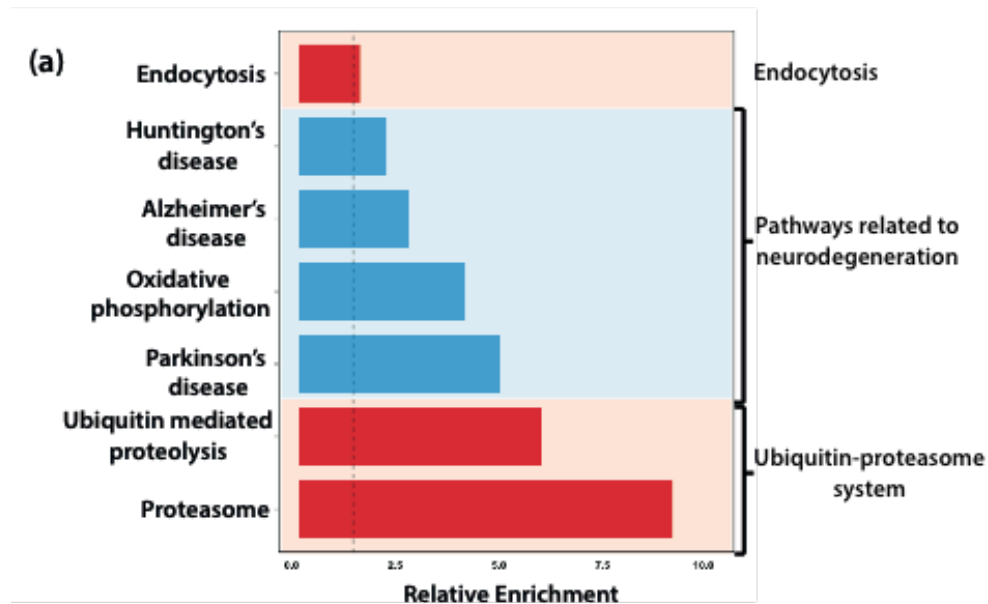
Since any given module is comprised of a large number of genes, it is important to identify the most highly connected genes within a particular module, as these central or 'hub' genes are more likely to be functionally related compared to genes that are less connected. To achieve this goal, we defined the 'module membership' score (MM), by using the intramodular connectivity (kME, **Section 3.6**), which is a measure of how strongly connected*, i.e.* coexpressed, a given gene is to all the other genes in a module (260). Hub genes were defined as those genes having an absolute kME value greater than 0.8.

The hub genes in the AD metastable network were found to be highly enriched in the KEGG biochemical pathways of cellular degradation (proteasome and ubiquitin-mediated proteolysis) and trafficking in addition to those previously associated with metastable proteins such as oxidative phosphorylation, Alzheimer's disease, Parkinson's disease, and Huntington's disease **(Figure 3.3a)**. These results are fully consistent with those reported above for the full list of genes in the AD metastable network (**Figure B.3**). In any given module, a high mean MM value indicates how tightly coexpressed the genes are within that module. We observed that the genes encoding for the AD metastable subproteome in the AD metastable network have a significantly high mean MM value that is significantly greater than that of other genes in that module **(Figure 3.3b)**. In addition, more than two-thirds of the genes encoding for metastable proteins in these modules were hub genes, indicating their central importance in their respective modules.

3.3.5 Test of module generality using a consensus network analysis with a visual cortex dataset.

We next sought to determine whether the modules that we identified are general or instead specific to the dataset or brain region that we analyzed. To check the robustness of the modules identified in this study, we constructed a consensus network (**Section 3.6**) using WGCNA on another dataset from the visual cortex of Alzheimer's disease patients and healthy controls (232), along with the dataset for the dorsolateral prefrontal cortex used previously (**Section 3.6**), to examine whether or not our network is preserved. To assess the level of preservation, we used the 'consensus network' construction, which identifies groups of genes that are tightly coexpressed across multiple studies (269). The consensus module eigengenes (consMEs) represent modules in each of the two sets (269) (**Section 3.6**). Each gene is assigned to a single consensus module but there are two sets of consMEs for each module as a given module can have a different expression profile in the two datasets. We found that all of the modules identified in our study have a consensus counterpart in the visual cortex dataset, indicating that the module structure in the two datasets is similar (**Figure B.4**).

We then constructed the two sets of eigengene dendrograms and eigengene heatmaps based on the consMEs (one for each study) and the results indicate that the overall modular structure in the two sets is quite similar. The preservation heatmap shows the preservation network, defined as one minus the absolute difference of the eigengene networks in the two data sets (**Figure B.5a**). The overall degree of preservation between the two networks is 0.87 and the mean preservation of relationships for each eigengene is consistently high for all the modules except the 'Red' one, as shown by the preservation heatmap and bar plot (**Figure B.5b**), thus indicating that the modules identified in the analysis detailed in the study are highly robust.



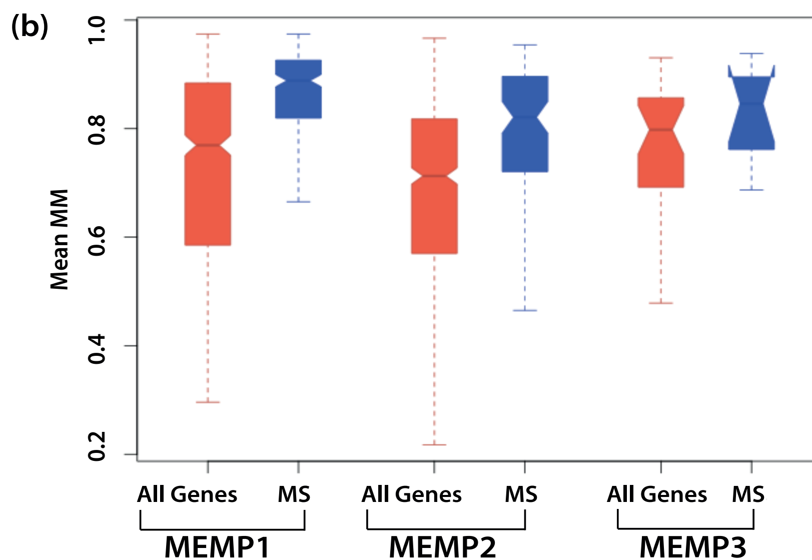


Figure 3.3 Identification of KEGG biochemical pathways enriched in hub genes in the AD metastable network. (a) KEGG biochemical pathways (268) enriched in hub genes (|kME| >0.8) in the AD metastable network; the dotted line indicates p=0.05 (Pr: proteasome, UMP: ubiquitin mediated proteolysis, OP: oxidative phosphorylation, PD: Parkinson's disease; AD: Alzheimer's disease, HD: Huntington's disease; Endo: endocytosis). (b) A comparison of the mean module membership (MM) values of the metastable proteome (MS, blue) and of all genes (orange) shows that genes corresponding to metastable proteins are highly coexpressed.

These results suggest that the difference between a healthy state and a disease state does not involve a reorganisation of the modules, but rather a variation in the expression levels of specific genes within the modules. In the following we therefore carried out further investigations to identify such genes.

3.3.6 Identification of a protein homeostasis complement of the AD metastable subproteome.

We next asked the central question of this work—How is the AD metastable subproteome regulated? To answer this question, we analysed which components of the protein homeostasis system are coexpressed with the AD metastable subproteome in the AD metastable network, as we expect that the knowledge of such components could offer insight into the regulation of these metastable proteins **(Figures 3.4** and **B.6)**. To this end, we identified the most important hub genes by visualizing the modules; we used the Cytoscape software for this purpose (270). The top 10% of all hub gene interactions based on their topological overlap were visualized, with those involved in at least 50 of these interactions shown in the centre **(Figure 3.4a)**. We observed ten genes related to trafficking and five genes related to the ubiquitin-proteasome pathway as the most connected hub genes **(Figure 3.4a)**. These results are consistent with the observation that protein trafficking and degradation are essential in the regulation of protein homeostasis of Alzheimer's disease (89-91, 271-274). Our analysis also identifies other components (autophagy, metabolism, signaling, and protein synthesis; **Figure. 3.4b**), although more extensive data will be needed to clarify their association with the metastable subproteome in greater detail.

3.3.6.1 Endosomal-lysosomal system

Although the present analysis of the hub genes is aimed primarily at identifying the main processes within the AD metastable network, it may also be informative to consider their possible specific roles in AD. Among the hub genes associated with trafficking, we found RAB6A, a small GTPase that helps mediate retrograde transport from the Golgi apparatus to the endoplasmic reticulum (ER), which has an increased expression level in AD brains (275). To explain this finding, it has been

suggested that this protein is involved in a regulatory mechanism that responds to increased protein accumulation (275). In addition, overexpression of RAB1, another small GTPase closely related to RAB6A, was shown to alleviate ER stress in yeast models of PD (276). Hence, RAB6A, which is a central gene in the AD metastable network described in this work, could play an important role in the regulation of the metastable proteins by directing them toward the endosomal–lysosomal degradation machinery, thereby preventing their accumulation in the cytoplasm. Another two genes in the group that we found are ATP6V1H, which encodes a protein subunit of a vacuolar ATPase involved in clathrin-mediated endocytosis (277, 278) and whose role in regulating lysosomal pH has been recently been linked to neurodegeneration (279), and ATL1, which is involved in ER trafficking (280, 281). In fact, all 10 genes that we found to be related to trafficking are part of the endosomal–lysosomal system. Specifically, SH3GL2, SLC9A6, and CLTA are localized in the endocytic vesicle membrane (282, 283), and NSF is involved in vesicle-mediated transport and acts as a fusion protein through the SNARE proteins (284). Our results, therefore, indicate the importance of the endosomal–lysosomal system in controlling the metastable subproteome. These findings extend the well-known role of this system in the processing of Aβ (285) to the regulation of a broader range of aggregation-prone proteins.

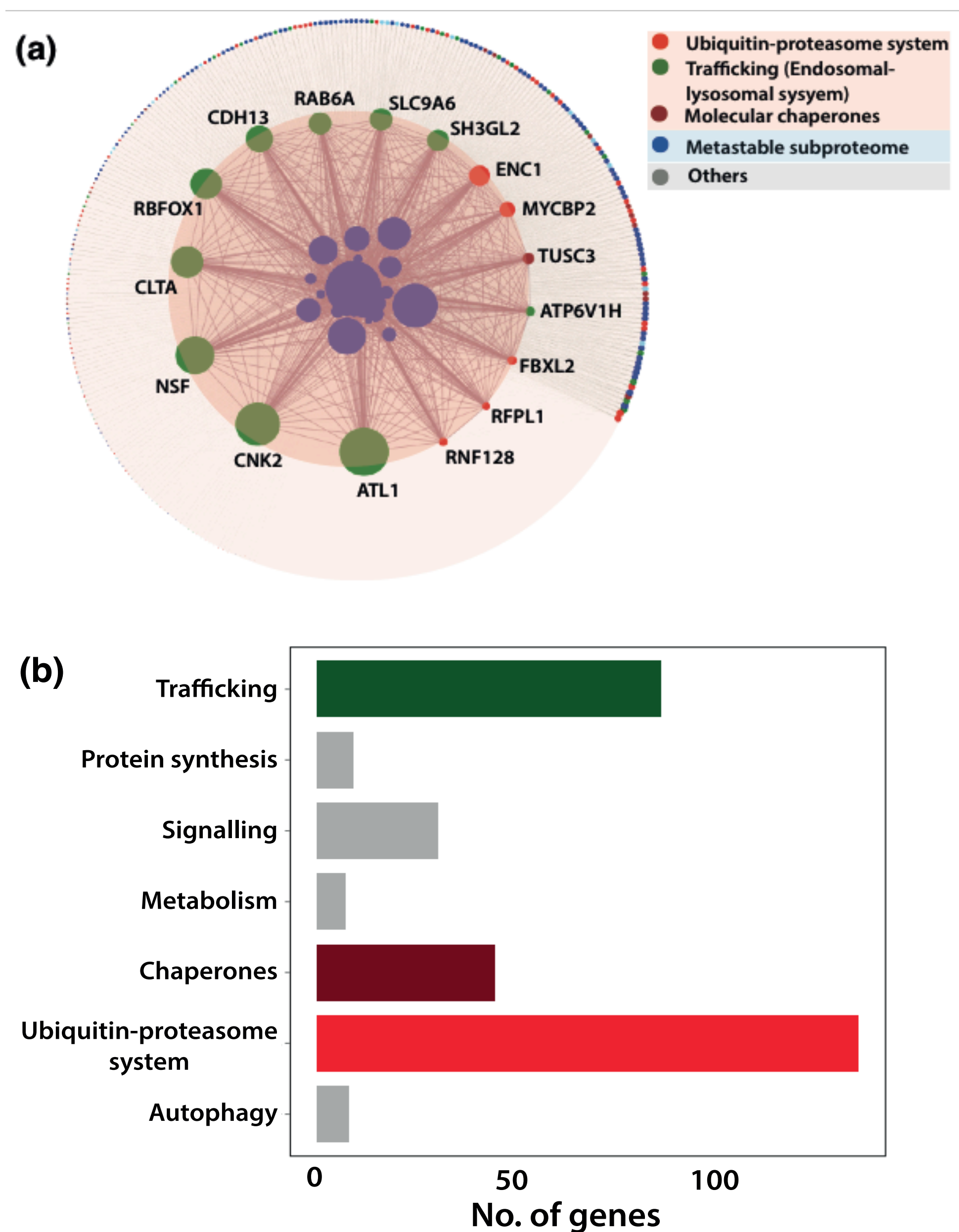


Figure 3.4 Identification of the major components of the protein homeostasis system associated with the AD metastable subproteome. (a) Network representation of the AD metastable network showing the hub genes of the protein homeostasis system (green, red, and dark red circles, middle ring; Table 3.1) and the hub genes of the AD metastable subproteome (blue circles, inner ring). This analysis reveals in particular the importance of the ubiquitin–proteasome 577 (red) and trafficking (green) systems in the regulation of aggregation-prone proteins in AD. We visualized the top 10% of the hub gene interactions, with 578 those genes involved in at least 50 interactions shown in the inner and middle rings (see also Fig. S6). The sizes of the nodes correspond to their degrees of connectivity. (b) Protein homeostasis components within the hub genes of the AD metastable network. The major components in A are shown in the same color code; additional components (autophagy, metabolism, signaling, and protein synthesis) are also shown.

Hub Genes	Protein	Known Function
ATP6V1H	V-type proton ATPase subunit H	Clathrin coated endocytosis, formation of endosomes
SH3GL2	Endophilin-A1	Synaptic vesicle endocytosis
SLC9A6	Sodium/hydrogen exchanger 6	Exchange of protons across early and recycling endosome membrane
RAB6A	Ras-related protein Rab-6A	Retrograde transport from golgi to ER, transport from endosome to plasma membrane
CDH13	Cadherin-13	Regulation of endosytosis
RBFOX1	RNA binding protein fox-1 homolog 1	RNA binding protein, regulated alternative splicing events
CLTA	Clathrin light chain A	Major protein of the polyhedral coat of coated pits and vesicles
NSF	Vesicle-fusing ATPase	SNARE binding, regulation of exocytisis
CNK2	Connector enhancer of kinase suppressor of ras 2	Adaptor protein, regulation of signal transduction
ATL1	Atlastin-1	ER to golgi vesicle transfer
ENC1	Ectoderm-neural cortex protein 1	Proteasomal ubiquitin-independent protein catabolic process
MYCBP2	E3 ubiquitin-protein ligase MYCBP2	Ubiquitin ligase, Protein ubiquitination
FBXL2	F-box/LRR-repeat protein 2	Ubiquitin ligase, Protein ubiquitination
RFPL1	Ret finger protein-like 1	Zinc ion binding
RNF128	E3 ubiquitin-protein ligase RNF128	Ubiquitin ligase, ubiquitin-dependent protein catabolic process
TUSC3	Tumor suppressor candidate 3	Magnesium transporter

Table 3.1 List of hub genes used to identify the components of the protein homeostasis system associated with the AD metastable subproteome. These hub genes are shown in Figure 3.4 and are reported here together with their corresponding proteins and their known functions. The list of hub genes corresponding to the AD metastable subproteome is reported in Table B.5.

3.3.6.2 Ubiquitin-proteasome system

Among the genes associated with the ubiquitin–proteasome system, we found ENC1, which is an actin binding protein that has been reported to modulate the aggregation of mutant huntingtin under ER stress (286). MYCBP2, FBXL2, and RNF128 are E3 ubiquitin ligases and are essential components of the ubiquitin-dependent degradation

of proteins (287-289). These results indicate that the metastable proteins are likely to be regulated upstream of the proteasomes at the ubiquitin ligase stage.

Molecular chaperones. We also found a number of components of molecular chaperone networks coexpressed with the AD metastable subproteome (Table B.4). Such components include co-Hsp70/Hsp90 species, which are known to assist the Hsp70/Hsp90 system to degrade protein aggregates (230, 290). Among such molecular chaperones, we found DNAJC6, a J-domain cochaperone with a role in HSC70-mediated uncoating of the clathrin-coated vesicles in neurons by recruiting HSC70. Also seen as hub genes were TOR1A, with chaperone activity and a member of the AAA family of ATPases, and ERLEC1, which has a role in ER quality control (291, 292).

Taken together, these results indicate that the components of the AD metastable subproteome, which consists of proteins inherently at risk for aggregation, tend to be highly coexpressed with multiple components of the protein homeostasis system. These results illustrate how during the course of AD, when a dysregulation and collapse of these systems is increasingly likely to occur, these metastable proteins are likely to represent an enhanced risk due to the dysfunction of the regulatory mechanisms associated with their folding, transport, and degradation.

3.3.7 Relationship with genome wide association studies (GWAS)

To further assess the significance of our analysis, we compared our results with genetic loci identified by GWAS. These studies have reported that several loci associated with the trafficking and degradation systems are closely associated with AD (18, 293). In particular, seven GWAS genes (PICALM, SORL1, CD33, BIN1, CD2AP, ABCA7, and RIN3) are associated with the endosomal– lysosomal system, and two GWAS genes (CLU and PTK2B) are associated with the ubiquitin–proteasome pathway (18, 293).

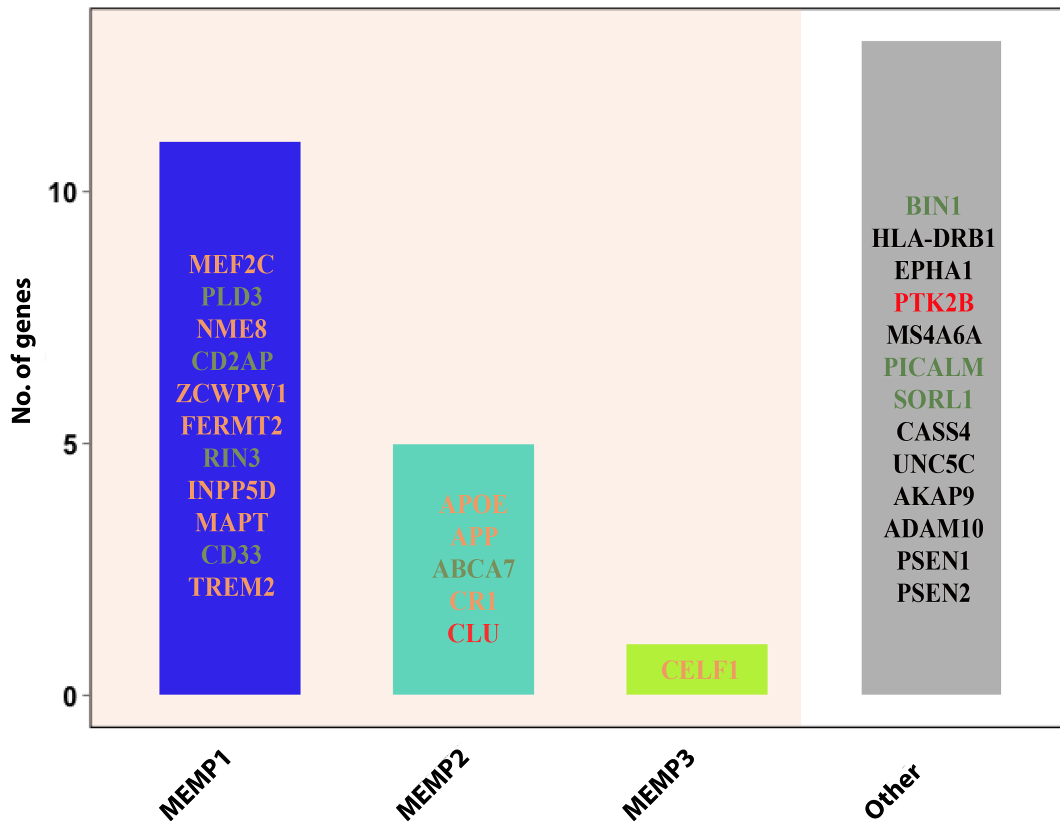


Figure 3.5 The majority of GWAS genes are found in the AD metastable network. Number of genes identified by GWAS (18, 293) that are present in the AD metastable network or in the other modules described in this work. 17 out of 28 genes identified by GWAS are present in the AD metastable network. The names of genes are shown in their respective modules. The genes shown in red belong to the ubiquitin-proteasome system and those in green belong to the trafficking system.

These results are highly consistent with the conclusions of the present study, as 17 GWAS genes (among the 28 that we considered) are present in the AD metastable network identified in this work (**Figure. 3.5**). This consistency is remarkable, as the GWAS strategy, where genes are typically associated with disease on the basis of single nucleotide polymorphism (SNP) statistics, is independent from the one that we have used here to associate genes with disease through the combination of their coexpression and the metastability for aggregation of their products. These two approaches are therefore complementary, as a coexpression analysis can identify a large number of genes and therefore reveal the biochemical pathways involved in the disease and help rationalize the specific roles of the GWAS genes but may not capture

important relationships, such as in the present case the role of ADAM10, PSEN1, and PSEN2 in the processing and regulation of APP **(Figure 3.5)**.

3.3.8 Consensus network analysis of Alzheimer's, Parkinson's and Huntington's diseases

As noted above, the phenomenon of protein misfolding and aggregation is a common feature of many neurodegenerative disorders, including Alzheimer's, Parkinson's and Huntington's diseases, and ALS. Although these diseases are characterised by a variety of different clinical manifestations and features, there is increasing interest in understanding the extent to which they share common molecular origins (3-12, 15-17). To address this question in the present context, we investigated whether or not the regulation of the metastable proteins, in terms of their coexpression with specific protein homeostasis components, is similar across Alzheimer's, Parkinson's and Huntington's diseases.

Since oxidative phosphorylation is the most significantly enriched pathway among the metastable genes, we analysed the coexpression of genes involved in this specific pathway and in the protein homeostasis components. We built a consensus network for gene expression data from hippocampal tissue (220), *substantia niagra* (294) and prefrontal cortex (292) (**Table B.1**), obtained post mortem from patients diagnosed with Alzheimer's, Parkinson's and Huntington's diseases, respectively, and age matched controls. The network heatmaps indicate the correlation of various eigengenes within the Alzheimer's, Parkinson's and Huntington's networks (**Figure B.7**), and the preservation heatmaps (**Figure 3.6**) reveal that the overall preservation of the three networks is highly significant (shown in red). The mean preservation of the three networks exceeds 0.7 in all three cases (**Figure B.7**), indicating that the global structures of the coexpression networks are similar for the three diseases. These results thus suggest that the differences between these diseases may be found in the dysregulation of specific genes within the consensus network (**Figures 3.6** and **B.7**).

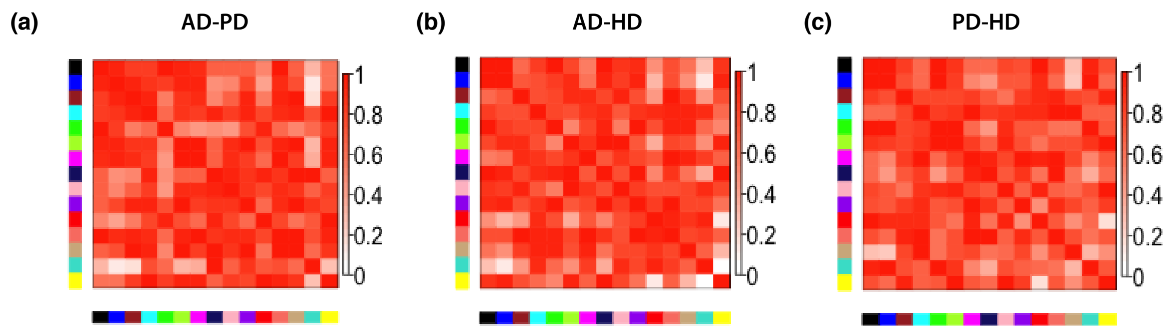


Figure 3.6 Network preservation heatmaps for AD-PD (a), AD-HD (b) and HD-PD (c). The rows and columns represent the module eigengene for different modules in each network. Preservation is defined as one minus the absolute difference of the eigengene networks in the two data sets. Red denotes high preservation and white denotes low preservation.

3.4 Discussion

3.4.1 Specific components of the protein homeostasis system that regulate protein aggregation

In this work, we have taken the view that a major hallmark of ageing and neurodegenerative diseases is the progressive impairment of the balance between protein aggregation and its control by the protein homeostasis system, which leads to the characteristic accumulation of aberrant protein aggregates (3-12, 15-17, 29-35, 69, 70, 95, 96, 179, 181, 212, 213, 254, 255) (**Figure 3.1**). In this context, we have previously reported that large numbers of proteins are inherently metastable to aggregation because of their elevated expression levels relative to their solubilities (33, 34). We have also observed a specific transcriptional downregulation of genes encoding these proteins in AD (181), as well as a tissue-specific vulnerability to Alzheimer's disease caused by an imbalance between aggregation-prone proteins and their protein homeostasis regulators (254).

To identify the components of the protein homeostasis mechanism that controls a set of metastable proteins associated with Alzheimer's disease, in this study we have

analysed together a set of proteins that are inherently metastable to aggregation (181) and a set of proteins that make up the overall protein homeostasis system (212). Our analysis started from a metastable subproteome corresponding to the overlap between genes encoding for proteins that are supersaturated and transcriptionally downregulated in Alzheimer's disease but not in ageing (181). We then constructed an 'AD metastable network' composed of genes encoding this set of metastable proteins together with the corresponding components of the protein homeostasis system. We have found that this specific AD metastable network consists of well-defined modules of coexpressed genes **(Figure 3.2)**, enabling us to identify key players of the ubiquitin-proteasome and endosomal-lysosomal systems, along with some specific molecular chaperones **(Figure 3.4)**.

The systems level approach that we have adopted in this work provides an understanding of the regulation of the AD metastable subproteome as a whole, as opposed to the regulation of individual proteins by specific components of the protein homeostasis system. Our results show that, from the list of about 2000 components of the protein homeostasis system (212), just a relatively small number of specific proteins in the degradation and trafficking machinery along with specific molecular chaperones are primarily responsible for handling the metastable proteins with a high propensity to misfold (**Figure 3.7**).

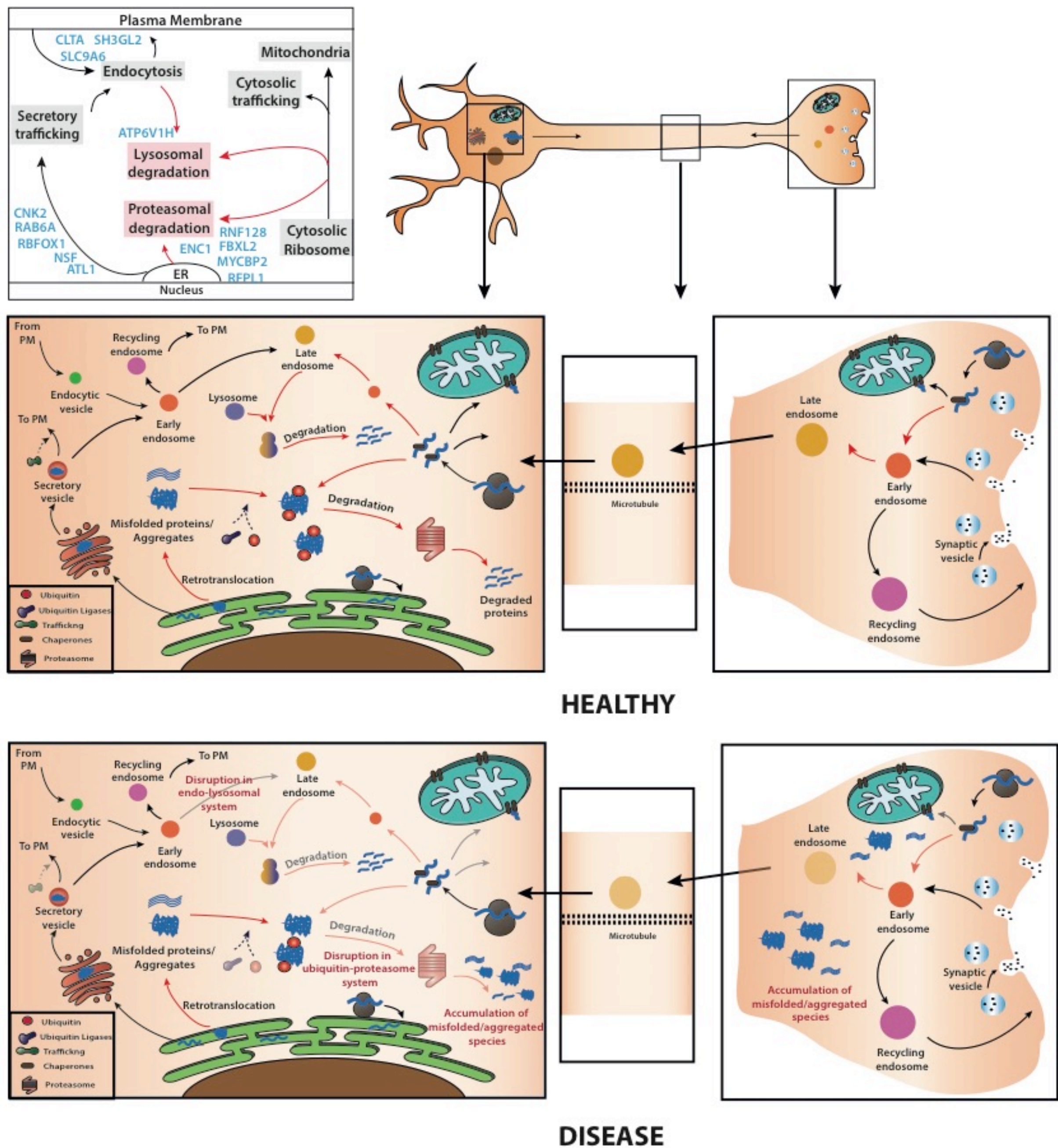


Figure 3.7 Schematic representation of the main pathways involved in the protein homeostasis of the proteins metastable to aggregation in Alzheimer's disease. (a) Healthy state. (b) Disease state. Our results identify the protein degradation network, in particular the ubiquitin-proteasome system and trafficking, as key control mechanisms in the homeostasis of proteins metastable to aggregation in Alzheimer's disease.

3.4.1.1 Endosomal-lysosomal and ubiquitin-proteasome regulation of the proteins involved in oxidative phosphorylation

The expression levels of most of the components of the protein homeostasis complement identified in this study have been previously seen to be decreased with ageing (181). Hence, during ageing and disease, with the suppression of the protein homeostasis system, these proteins could become particularly vulnerable to aggregation because of their inherent metastability. Since these proteins perform fundamental functions, including in particular energy metabolism through oxidative phosphorylation (181), their aggregation could result in triggering a cascade of events contributing to disease pathology and ultimately to neuronal death. In addition, such dysregulation poses a pronounced threat to the neurons due to their post-mitotic state and increased dependence on mitochondria for energy production. Indeed, there is a substantial overlap in the genes involved in the pathways associated with oxidative phosphorylation and Alzheimer's, Huntington's and Parkinson's diseases, indicating again that the proteins encoded by these genes are highly metastable and hence are significant in the context of disease pathology. Mitochondria play a central role in ageing and in regulating cell death (263, 295), as well as in the overall maintenance of cellular health. Whether mitochondrial dysfunction is the cause or effect of the disease pathology is still, however, unclear. Mitochondria have been shown to interact with aggregation-prone proteins, including α-synuclein and Aβ. More specifically, Aβ was shown to be localized on the mitochondrial membrane in a transmembrane arrested form, possibly disrupting protein import into the mitochondria (296, 297).

Our results also point to a possible dependence of the proteins in the respiratory chain complex on the endosomal-lysosomal system, in particular RAB6A, ATP6V1H, ATL1, SH3GL2, SLC9A6 and CLTA, and on the ubiquitin-proteasome system, especially the E3 ubiquitin ligases MYCBP2, FBXL2 and RNF128 (**Figure 3.4**). These indications are consistent with the observation in yeast that accumulation of mitochondrial proteins in the cytoplasm leads to activation of an unfolded protein response (298). Furthermore, recent studies have reported the presence of polyubiquitinated mitochondrial proteins suggesting that they are substrates of the ubiquitin proteasome system (299, 300), and that in yeast the expression of the proteasome is upregulated upon cytoplasmic accumulation of mitochondrial proteins

(298, 301). If mitochondrial import is disrupted and these metastable proteins therefore accumulate in the cytoplasm, the cell responds by clearing them through degradation. If, however, this happens in an environment where protein homeostasis is compromised, these proteins would be particularly at danger of aggregation. We observed, in addition, a similar pattern of co-expression of genes encoding for mitochondrial membrane proteins across Alzheimer's, Huntington's and Parkinson's diseases, indicating that even though the initial cause of dysregulation might be different, these diseases are likely to share common molecular mechanisms at a later stage of progression, with regulation of mitochondrial membrane proteins playing an important role.

The finding that a metastable subproteome that is specifically associated with Alzheimer's disease is primarily regulated by the protein trafficking and degradation systems provides important insights into the control of protein misfolding in this disease. These results suggest that in a setting of compromised protein folding the maintenance of proteins in their soluble states may move away from regulating conformations and towards regulating concentrations.

3.5 Conclusions

We have described specific components of the protein homeostasis system that regulates a metastable subproteome associated with Alzheimer's disease. This analysis has revealed the central roles of the ubiquitin-proteasome and endosomal-lysosomal degradation pathways, whose relevance to AD is well known (89-91, 271-274), in the maintenance of a pool of proteins prone to aggregation. By identifying a series of regulatory pathways associated with Alzheimer's disease, these findings also help rationalize the roles in the disease of the individual genes resulting from genome-wide association studies (GWAS). We anticipate that an increasingly detailed understanding of the mechanisms of regulation of metastable proteins will contribute significantly to the development of therapeutic strategies aimed at promoting the maintenance of aggregation-prone proteins in their soluble states.

3.6 Materials and Methods

3.6.1 Dataset acquisition

Microarray data for brain tissues from post mortem Alzheimer's disease patients and healthy controls were downloaded from the Gene Expression Omnibus (GEO) database (302). The following datasets were used for analysis (**Table B.1**): GSE44770, containing tissues derived through autopsy from the dorsolateral prefrontal cortex (PFC) region obtained late-onset Alzheimer's disease (LOAD) patients and from healthy controls; GSE44771, containing tissues derived through autopsy from the visual cortex (VC) region obtained from LOAD patients and from healthy controls; GSE1297, containing hippocampal gene expression data from LOAD patients and from healthy controls; GSE33000, containing dorsolateral pre-frontal cortex tissue from Huntington's disease patients and from healthy controls obtained from the Harvard Brain Tissue Resource Center (HBTRC); GSE20292, containing post-mortem brain tissue from the *substantia niagra* of Parkinson's disease patients and from healthy controls. Using the GEOquery package, data were downloaded into R, and checked for missing values (259).

3.6.2 Sample clustering

Samples in each dataset were hierarchically clustered within GEOquery to detect outliers. One sample from GSE44771 (GSM1090949) and one sample from GSE20292 (GSM508732) were found to be outliers and hence removed from further analysis.

3.6.3 Generation of a 'Weighted Gene Correlation Network'

A distance measure commonly used for coexpression analysis is based on the Pearson's coefficient of correlation; in this approach, gene pairs with a coefficient of correlation below a given cut-off value (e.g. 0.8) are considered as not correlated. However, this kind of 'hard thresholding' may be insensitive to subtle and yet

important expression patterns (303). We therefore employed the 'Weighted Gene Correlation Network Analysis' (WGCNA) method (259, 260), which uses a 'soft thresholding' and the concept of topological overlap or shared neighbours to identify clusters of coexpressed genes. The soft thresholding method assigns a weight to each pair of interacting genes and uses such weights, along with the topological overlap to identify modules of coexpressed genes in the expression data (259, 260).

The construction of a 'Weighted Gene Correlation Network' was performed using the R package for WGCNA (259). Absolute values of Pearson's coefficient of correlation were calculated for the expression values of each gene pair across all microarray samples. WGCNA uses a power function to transform the coexpression similarities (given by a similarity matrix $S = [s_{ij}]$) into connection strengths (given by an adjacency matrix $A = [a_{ij}]$)

$$a_{ij} = |s_{ij}|^{\beta} \quad (1)$$

where β is the soft thresholding power. In unweighted networks, the entries a_{ij} of the adjacency matrix are either 1 or 0, indicating whether or not a pair of nodes is connected. In weighted networks, the values are real numbers ranging from 0 to 1. Due to the noise in microarray data and the limited number of samples, we weighted the Pearson's coefficients of correlation by taking their absolute values and raising them to the power β. To chose the value of β we observed that many biological networks, especially gene expression networks have been found to exhibit approximate scale free topology (304) *i.e.* the connectivity distribution $p(k)$ for each node k follows a power law, $p(k) = k^{-\gamma}$, with exponent γ. This 'scale-free' relationship indicates that there are a few nodes that are highly connected while others have much fewer connections. Through these considerations we chose β=6 (260). This procedure results in a weighted network in which the continuous nature of the gene expression values is preserved (as opposed to unweighted networks); the results are robust with respect to the choice of β, as opposed to the high sensitivity to the cutoff value of unweighted networks.

3.6.3.1 Identification of modules in the Weighted Gene Correlation Network

Modules were defined as groups of genes having high correlation and high topological overlap (260). The topological overlap of two nodes refers to their interconnectedness, which is measured as the number of shared neighbors between two nodes. It provides a similarity measure that has been shown to be very useful in biological networks (305), and was used here as the basis for average linking hierarchical clustering to identify modules of coexpressed genes.

3.6.3.2 Module eigengenes

The module eigengene (ME), which is defined as the first principal component of a given module, can be considered as a representative of the gene expression profiles in a module (259). The connectivity of a gene i with a module k ($MM_k(i)$) is defined as the Pearson's coefficient of correlation of the expression value of that gene with the ME of the module. It is a measure of module membership (MM) for a particular gene. Specifically,

$$MM_k(i) = cor(e(i), E_k) \quad (2)$$

where $MM_k(i)$ is a measure of MM for gene i with respect to module k, e(i) is the expression profile of gene i and E_k is the eigengene of module k. The intramodular connectivity (kME) is defined as the connectivity of a gene within its own module. The ME is also used to calculate the Pearson's coefficient of correlation and the associated student p-value of each module with disease status; the disease status is encoded as binary information for disease or healthy.

3.6.3.3 Module preservation and consensus analysis

WGCNA provides various measures of module preservation statistics, which assess whether or not the interconnections among the genes within a module, and connectivity patterns of individual modules (for example, intramodular hub gene

status) are preserved between two datasets. To assess the preservation of our disease associated modules found in the PFC dataset (the network that we analyzed) and in a hippocampal gene expression dataset (test network), we used the modulePreservation function in the WGCNA R package (269). In brief, this function provides an average measure of several preservation statistics generated through many permutations of the data, the $Z_{summary}$ value. In general, modules with $Z_{summary}$ scores >10 are interpreted as strongly preservation (that is, densely connected, distinct, and reproducible modules), $Z_{summary}$ scores between 2 and 10 are weak to moderately preserved, and $Z_{summary}$ scores <2 are not preserved (269). Another way to look at module preservation is to rank the modules by their overall preservation in the test set which gives a relative measure of module preservation. Median rank is a measure which relies on observed preservation statistics rather than the permutation Z-statistics (269). It is calculated as described previously (269).

Consensus analysis is a way to identify modules present in several independent datasets. Consensus modules group together genes densely connected in all conditions and are defined from the clustering of consensus similarity

$$\text{Modules}_{\text{consensus}} = \min(\text{Network 1, Network 2}) \quad (3)$$

Consensus modules are by construction present (i.e. preserved) in all input datasets. If a module identified in a reference dataset is strongly preserved in test datasets, it would also be a consensus module among the reference and test datasets. Each consensus module has one eigengene per dataset. Eigengene correlation helps to visualize the overall network structure and also to compare a given network between different datasets. An eigengene network (A_{ij}) is defined as a signed network with a soft thresholding power of 1. A preservation network ($Pres_{ij}$) measures the correlation of eigengene correlation amongst different networks (306):

$$Pres_{ij}^{(1,2\ldots)} = 1 - [\max\left(A_{ij}^{(1)}, A_{ij}^{(2)}, \ldots\right) - \min\left(A_{ij}^{(1)}, A_{ij}^{(2)}, \ldots\right)] \quad (4)$$

where $Pres_{ij}^{(1,2\ldots)}$ is the preservation network for any networks 1 and 2. The overall

mean preservation of eigengene networks is given by (306)

$$D^{(1,2\dots)} = mean_{i<j} P_{i,j}^{(1,2\dots)} \tag{5}$$

CHAPTER 4

---

4. A Map of Protein Aggregation Homeostasis Identifies the Vulnerability of Cells sand Tissues to Alzheimer's Disease

4.1 Overview

As described earlier in this thesis, Alzheimer's disease is associated with a set of metastable aggregation prone proteins (Chapter 2) (33, 34, 181), which are regulated by components of the cellular trafficking and degradation systems (Chapter 3) (182). Here, to understand why disease-associated protein deposits form in certain tissues but not in others, we analyzed the ability of different cell and tissue types to respond to the presence of these aggregation-prone proteins by studying the balance between the expression of genes encoding metastable proteins and their protein homeostasis complement. We thus found that protein aggregation homeostasis is weaker in

neurons than in other cell types, and also weaker in brain tissues than in other body tissues. These results provide quantitative evidence that the defective regulation of protein aggregation is part of the molecular origins of Alzheimer's disease.

4.2 Introduction

As we have discussed, a large fraction of the proteome is inherently metastable within the cellular environment (33, 34, 181, 253), as the concentrations of these proteins exceed their intrinsic solubilities (49). Proteins that co-aggregate with plaques and tangles, and Lewy bodies were also found to be metastable. Biochemical pathways associated with neurodegenerative diseases were found to be enriched in these metastable proteins, indicating an important common feature of widespread protein metastability for these multifactorial neurodegenerative diseases. A fraction of these metastable proteins were reported to be specifically transcriptionally downregulated in AD, constituting an AD specific metastable subproteome (181), as discussed in Chapter 2. These proteins have been reported to be associated with specific components of the protein trafficking and clearance mechanisms, specifically the endosomal-lysosomal and the ubiquitin-proteasome systems, detailed in Chapter 3 (307). Thus, these components are crucial for the maintenance of the cellular homeostasis of a set of metastable proteins prone to aggregation in AD.

The protein quality control mechanisms have been shown to be progressively impaired upon ageing and neurodegenerative diseases, with the simultaneous increase in accumulation of aggregated species. Hence, it is crucial to maintain the balance between aggregation-prone proteins and their associated protein homeostasis components in order to maintain overall cellular health. Although ageing affects neurons as much as cells in other tissues, most of the diseases associated with protein aggregation concern the central nervous system. While there have been several studies about the selective regional vulnerability within brain tissues, it is not clear whether the brain itself is more vulnerable to aggregation diseases compared to other body tissues.

The aim of the study is to investigate whether the brain tissues, in a healthy state, are more susceptible to protein aggregation diseases compared to other body tissues, based on the balance between the expression of aggregation prone proteins and their regulatory systems. To achieve this goal, we determined the expression levels of metastable proteins associated with AD and the associated protein homeostasis components across 77 different healthy human tissues, including several brain tissues. We show that the brain tissues have a higher expression of the metastable proteins but lower expression of the associated homeostasis components compared to other tissues. Thus, due to the inherent lower expression of the homeostasis components associated with metastable proteins, the brain, as an organ, is ill-equipped to deal with these highly aggregation-prone metastable proteins, providing an explanation for the vulnerability of neurons to protein aggregation diseases.

4.3 Results

4.3.1 The protein homeostasis response is proportional to the risk of aggregation

Since aggregation-prone proteins are intrinsically metastable even in the absence of disease, we studied healthy brain tissues (265) to identify their associated protein homeostasis components using an approach based on coexpression network analysis (259). We used the subset of metastable proteins that are specifically associated with AD (MS), as described in Chapter 2 (181, 307) as our aim was to study the vulnerability of different tissues to AD. Using this weighted coexpression approach, we identified the protein homeostasis components (PHC) associated with these metastable proteins (**Section 3.6 and 4.6**). We used extensive microarray data across 77 different healthy human tissues (308) to study the vulnerability of various tissues based on the relative expression of these genes. We measured the relative expression of the MS and associated PHC in each tissue **(Figure 4.1 and C.1)** and found a correlation between the average expression of the MS and the average expression of its PHC. We also defined a “protection factor”, *s*, as the slope of the best-fit line for

the relative expression of the MS and its PHC (**Section 4.6**). The *s* score for MS and PHC across different tissues is 1.02, indicating a close overall proportionality between the protein aggregation risk and its mechanism of control **(Figure C.1)**. Thus, these results reveal the presence of a robust response to the presence of aggregation-prone proteins.

4.3.2 Brain tissues have a weaker protein aggregation homeostasis than other tissues

To understand the origin of the scatter around the overall correlation between PHC and MS levels shown in Figure C.1, we looked in more detail to the tissue-specific response. We thus found that the brain tissues and the body tissues separate into two distinct clusters **(Figure 4.1)**. Brain tissues have an elevated expression of metastable genes, but not of their protein homeostasis counterpart. In body tissues, the PHC expression grows more rapidly than the MS expression (s = 1.33). In brain tissues, the opposite is true (s = 0.88).

These results indicate that brain tissues, overall, have a weaker response to the presence of aggregation-prone proteins. There have been several studies to analyse the differential vulnerability of specific brain regions to different types of stress (309-313), however, there are no reports about the vulnerability of the brain as an organ. Our results show that the brain tissues are less capable of regulating the aggregation-prone proteins due to a weaker protein aggregation homeostasis compared to rest of the body tissues.

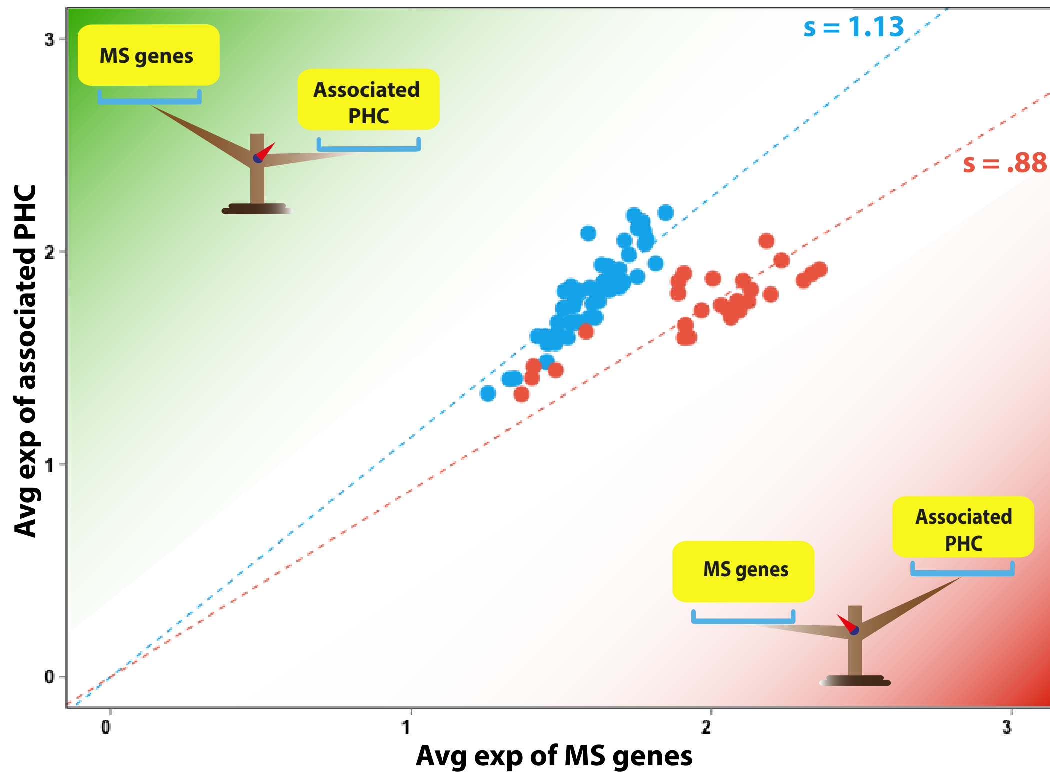


Figure 4.1. Brain tissues have a weaker protein aggregation homeostasis than other tissues. Scatter plot showing the average expression of metastable genes (MS) and the average expression of the associated protein homeostasis components (PHC) in different tissues. The brain tissues (orange) have a smaller slope than the body tissues (blue), indicating that the brain tissues have a lower protection against protein aggregation relative to other tissues.

4.3.3 Vulnerable brain tissues have a weaker protein aggregation homeostasis than non-vulnerable brain tissues

The greater vulnerability of brain tissues to protein aggregation suggests the presence of a link between protein aggregation homeostasis and neurodegenerative diseases, including in particular Alzheimer's disease. In order to investigate this relationship in more detail, we differentiated between tissues that are vulnerable and those that are resistant to Alzheimer's disease, as assessed by the Braak staging (314). We mapped various brain tissues to the respective Braak stages as reported previously (254). Amongst these, there were 7 brain tissues corresponding to different Braak stages

present in the microarray dataset used in the current study. We analysed the expression of the MS and the associated PHC in these tissues and compared them to the non-Braak tissues. We found that brain tissues corresponding to Braak regions have a lower slope than non-Braak regions **(Figure 4.2)**.

These results link closely protein aggregation homeostasis with tissue vulnerability to Alzheimer's disease.

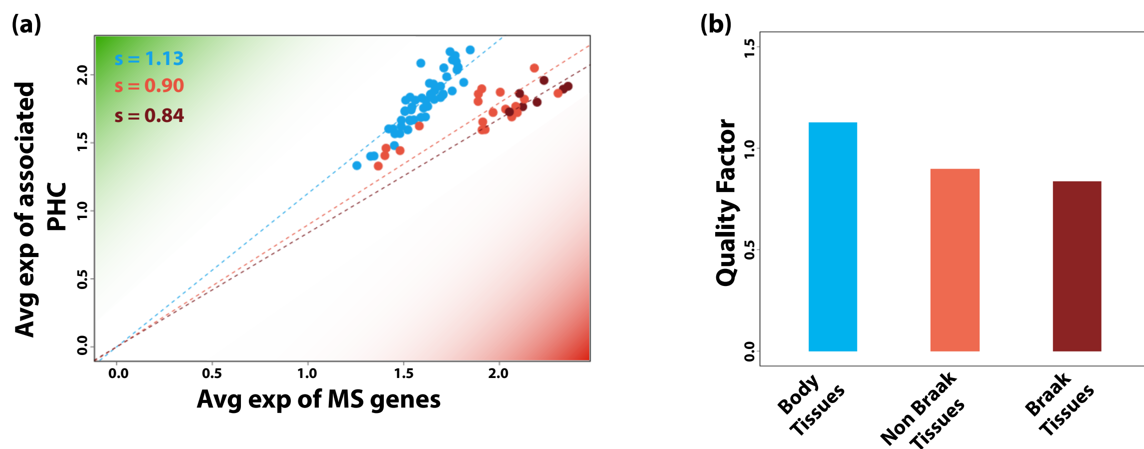


Figure 4.2. Vulnerable brain tissues have a weaker protein aggregation homeostasis than non-vulnerable brain tissues. (a) Brain tissues corresponding to different Braak regions have a lower slope than non-Braak regions. (b) Bar plot depicting the slopes in different tissues.

4.3.4 Neurons have a weaker protein aggregation homeostasis than non-neuronal brain cell types

The results presented above about the role of protein aggregation homeostasis in tissue vulnerability to Alzheimer's disease suggest a possible origin for the increased vulnerability of neurons to this disease with respect to other non-neuronal brain cell types. In order to verify whether or not neurons are less able than other brain cell types to regulate protein aggregation, we compared the expression of metastable genes (MS) and their associated protein homeostasis components (PHC) in neurons,

astrocytes, microglia and oligodendrocytes, using extensive single cell RNA sequencing data(315). We found that neurons have the lowest slope among these cell types (**Figure 4.3**).

These results indicate that neurons are the cells most vulnerable to protein aggregation in the brain.

4.3.5 Proteins in the oxidative phosphorylation pathway have a weaker protein aggregation homeostasis in the brain than in other tissues

In order to better understand the molecular origins of the increased vulnerability of certain brain tissues to Alzheimer's disease we considered the homeostasis of oxidative phosphorylation, since this is a specific process associated with metastable proteins (34, 181, 307). Oxidative phosphorylation has also been extensively linked with neurodegenerative diseases, although the exact nature of disruption is still unknown. We divided the genes associated with the Oxidative Phosphorylation pathway in the KEGG database (268) into two parts: those that are metastable and those that are not. We then calculated the protection factor (s) for the oxidative phosphorylation genes and their associated PHC for both groups across all tissues. The difference in s values between the body and brain tissues (Δs) is much higher in the group of oxidative phosphorylation genes that are also metastable, whereas there is almost no difference in the slopes for non-metastable oxidative phosphorylation genes. These results suggest that, even from within the oxidative phosphorylation pathway, those genes that are also metastable pose a higher risk for the brain tissues as these are less well prepared to face the challenge posed by the presence of these aggregation-prone proteins (**Figure 4.4**).

These results confirm previous observations about the close association between oxidative phosphorylation, protein aggregation, and neurodegenerative processes.

4.4 Discussion

In this chapter, we have shown that the balance between protein aggregation and its regulation through the protein homeostasis system is crucial in determining the vulnerability of cells or tissues to protein aggregation diseases. We have previously reported the presence of inherently aggregation prone proteins even in a healthy state (33, 34). We also observed that a subset of these aggregation prone proteins that is associated with AD is regulated closely by components of the endo-lysosomal system and the ubiquitin-proteasome pathway, identifying a protein homeostasis complement for these aggregation prone metastable proteins (181, 307). There is evidence that the protein homeostasis system gets progressively impaired with ageing and neurodegenerative diseases and this disrupts the balance between protein folding and aggregation within the cellular environment, leading to accumulation of aberrant aggregates.

Our results show that, already in a healthy state, the protein aggregation response is proportional to the risk of aggregation **(Figure 4.1 and C.1)**. We have taken a view that the strength of this correlation is an important factor in predicting the capability of different tissues or cells to handle the threat posed by the presence of aggregation prone proteins. We used data from healthy brain tissues to identify the protein homeostasis components of the MS based on coexpression network construction. We then defined the protection factor (s) as the slope between the MS and the associated PHC for different tissues and cells. We observed that body tissues have an s value higher than 1, suggesting that these tissues are well equipped to regulate these aggregation prone proteins. The brain tissues, however, have an s value of less than 1. Also, from among the brain tissues, those that correspond to Braak regions have an even lower s value. These results show that brain, as an organ, is less capable compared to other body tissues in terms of dealing with threat of protein aggregation. The protection factor s serves as a good measure of predicting the vulnerability to protein aggregation as it decreases when we move from body tissues to resistant brain tissues to highly vulnerable brain tissues. It also provides probable insights into the molecular origins of Alzheimer's disease as neurons have the least s score compared to other types of cells found in the CNS.

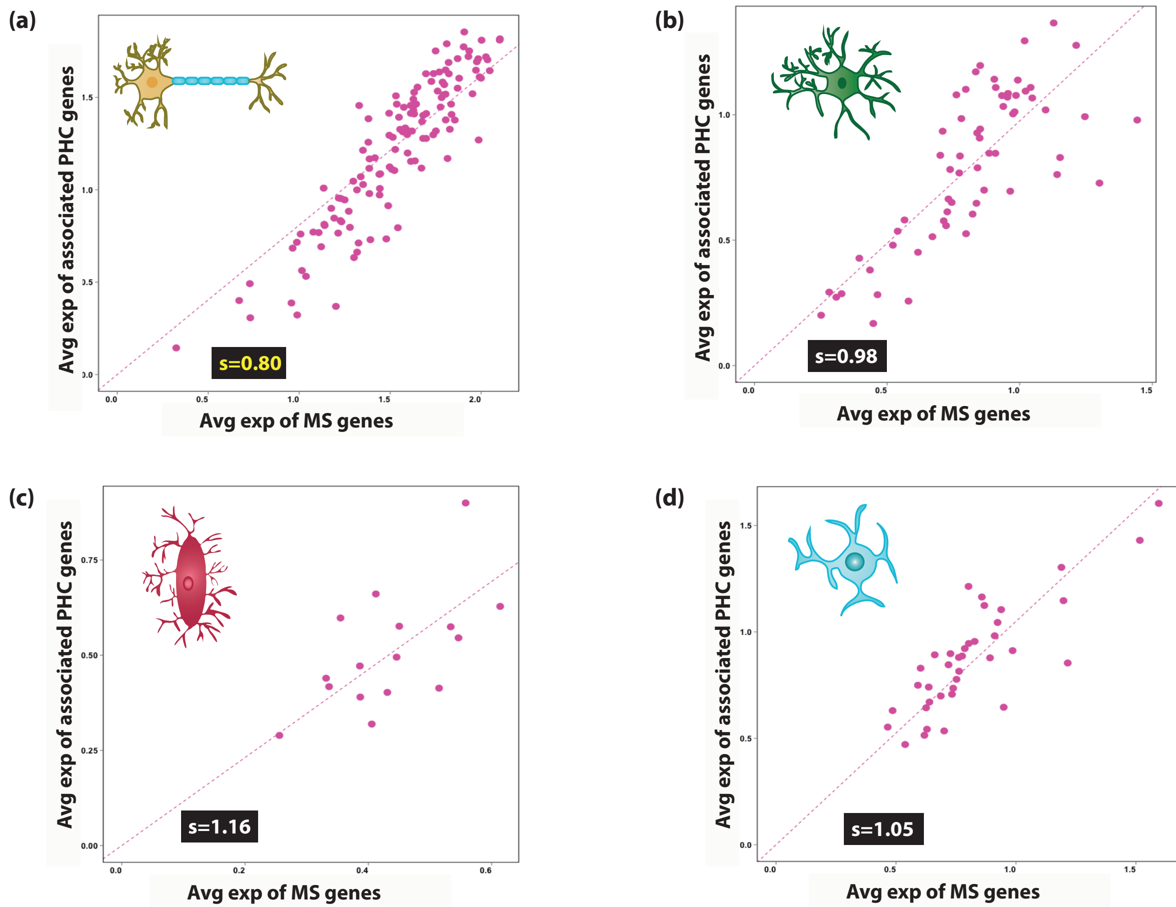


Figure 4.3. Neurons have a weaker protein aggregation homeostasis than non-neuronal brain cell types. Scatter plots depicting the expression of metastable genes (MS) and their associated protein homeostasis components (PHC) in (a) neurons, (b) astrocytes, (c) microglia and (d) oligodendrocytes. Each dot represents a single cell.

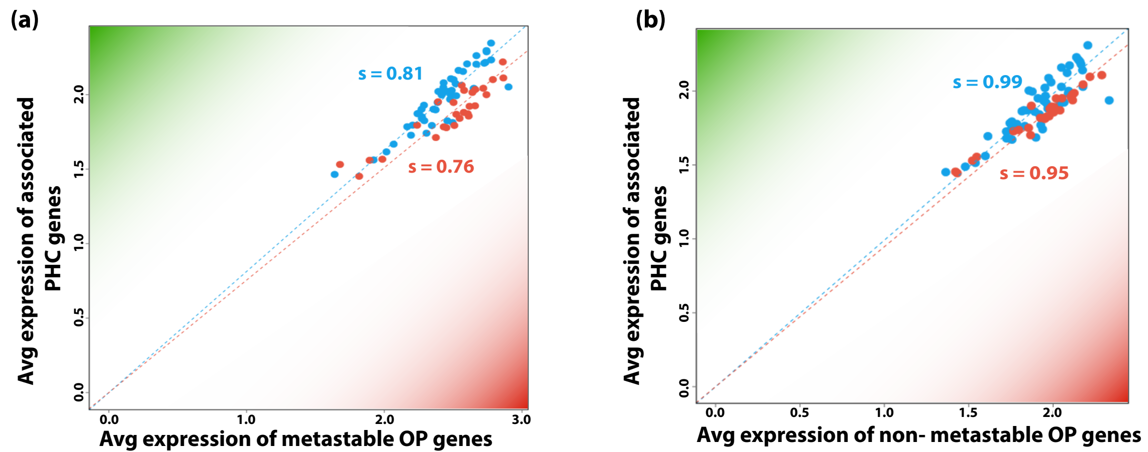


Figure 4.4. The oxidative phosphorylation pathway has a weaker protein aggregation homeostasis than other pathways. (a) Average expression of metastable genes within the KEGG Oxidative Phosphorylation pathway and their associated PHC. (b) Average expression of all non-metastable genes within the Oxidative Phosphorylation pathway and their associated PHC. The slope for these critical genes is lower (<1) throughout the body, not just in the brain and the difference in slopes between the body brain tissues is much more pronounced in the metastable Oxidative Phosphorylation genes.

Our results show that the brain, as an organ is vulnerable to protein aggregation even in the absence of disease. It has been reported that the human brain went through an accelerated cortical reorganization compared to other primates during the course of evolution (316). There is a possibility that this accelerated evolution of the brain placed increased pressure on critical genes in the brain to evolve their expression accordingly, not providing enough time for the simultaneous evolution of the regulatory components of the protein homeostasis system. Thus, when faced with the challenges of stress or aggregation prone proteins, brain tissues are more likely to suffer from protein aggregation diseases. Our results also show that neurons are the most vulnerable cells within the brain. Taken together, our results provide important insights into the molecular origins of AD and the vulnerability of brain, more specifically different regions and cells of the brain, to the threat posed by protein aggregation.

4.5 Conclusions

We have shown that brain tissues are particularly vulnerable to Alzheimer's disease because of their relatively poor ability to regulate protein aggregation. Furthermore, we have also shown that this ability is particularly lacking in neurons with respect to other brain cells, thus providing insight into the molecular origins of this neurodegenerative disease.

4.6 Materials and Methods

4.6.1 Dataset Acquisition

4.6.1.1 Healthy Brain Tissues

Microarray data for healthy brain tissues was acquired from the Allen Brain Atlas (265). Gene expression data for 6 healthy human brains was available, across 900 different tissues. Data was scaled and normalised using the 'scale' function in R. As Allan Brain Atlas uses multiple probes for each genes, the 'collapseRows' (317) function of the WGCNA package in R was used to get a single expression value for each gene across all samples. The expression values for each gene was then averaged across all six brains to arrive at the final expression value associated with each gene.

4.6.1.2 Tissue Specific Data

We obtained tissue specific data across various human tissues from a previously published dataset (308). Certain cancerous tissues and cell lines were removed from the analysis as our aim was to study the expression levels in a healthy state, leaving 77 tissues for the study. From these, 27 were characterised as neural tissues and 50 body tissues **(Table C.1)**.

For the Braak and Non Braak analysis, brain regions (from the 27 brain tissues) were assigned to either Braak or Non-Braak as described previously (254). Briefly, brain regions in the Allen Brain Atlas were matched with the closest regions mentioned in the original paper (314). 7 tissues were found to correspond to different Braak stages **(Table C.1)**

4.6.1.3 Cell type specific data

Single cell RNAseq data for 8 different cell types found in the brain was obtained from a published dataset (315). Data was scaled and normalised in R.

4.6.2 Coexpression network construction

We used the data obtained from the Allen Brain Atlas to construct a coexpression network for the metastable genes associated with Alzheimer's disease (MS) (182)and the genes associated with the protein homeostasis system to get the protein homeostasis complement (PHC) of the MS. WGCNA (259, 260) was used to construct the coexpression network as described previously (307). Briefly, WGCNA is a clustering algorithm based on hierarchical clustering but which uses 'soft thresholding' and the concept of topological overlap or shared neighbours to identify clusters of coexpressed genes. The soft thresholding method assigns a weight to each pair of interacting genes and uses such weight along with the topological overlap to identify modules of coexpressed genes in the expression data. The construction of a 'Weighted Gene Correlation Network' was performed using the R package for WGCNA. The names of MS and associated PHC are given in **Table S2**.

4.6.3 Calculation of 'Protection Factor (s)'

Our aim was to study the balance between the expression level of aggregation prone proteins and components associated with their regulation. Thus, we defined a "protection factor, s" as the slope of the best-fit line for the relative expression of the MS and its PHC across different tissue or cell types. A line was fitted based on linear

regression between the expression of MS and the PHC. The protection factor is a measure of the strength of the balance between the MS and its PHC.

4.6.4 Statistical testing

To evaluate the significance of our results, we used random sets of genes corresponding in number to the MS and PHC and calculated the s for them. We then calculated the difference in s between the body and brain tissues (Δs). We repeated these 1000 times to have a frequency distribution of Δs values for random sets of genes. The Δs for our genes of interest is 0.25, which is more than two SD away from the random sets of genes **(Figure C.2)**. Thus, our s values are highly robust and significant. The statistical testing was performed using the Scipy module in Python.

CHAPTER 5

---

5. Perspectives and future directions

5.1 Overview

The phenomenon of protein folding is fundamental to living systems and its failure is implicated in many human diseases (4, 6, 7, 9-12, 14, 15). Misfolding and aggregation can cause not just a loss of protein functionality, but also toxic, gain of function effects. The generic nature of amyloid formation suggests a general vulnerability of proteins to these conformational disorders (7, 11, 12). While the characteristic proteins in misfolding diseases, like Aβ, tau or α-synuclein might play a key role in initiating disease pathology, recent studies have started to identify the role of widespread protein aggregation in the pathology and progression of disease (31, 32, 34, 47). However, previously, these studies had not been able to elucidate the reasons behind he widespread aggregation observed in misfolding diseases like AD, PD, HD and ALS amongst others. Recently, it has been reported that a large set of proteins is inherently metastable to aggregation, even in a healthy state (34). These proteins, termed supersaturated because they are expressed a levels exceeding their intrinsic solubilities, are particularly aggregation prone due to their fundamental physico-

chemical properties. Intriguingly, these metastable proteins were also found to be enriched in biochemical pathways associated with neurodegeneration (33, 34).

The studies presented in this thesis build upon these observations by further analysing the role of these metastable proteins in AD and their regulation by the natural defences of the cell. These studies show that there is a specific overall transcriptional downregulation of metastable genes in AD, suggesting a potential cellular response to attenuate widespread aggregation in the face of compromised protein folding environment (181). The protein homeostasis system acts as a natural buffer to maintain the solubility of the proteome and also to clear any aggregated or misfolded species. It is a dynamic system capable of responding to cellular stresses, as exemplified by the heat shock response, and orchestrates a highly co-ordinated response between different components to maintain a healthy proteome within the cell (6, 8, 35, 88). However, these protective responses might be subject to disruption, as observed in ageing or misfolding diseases. The deficiency of protein homeostasis response could be causal to the disease or a result of overwhelming protein misfolding and aggregation. Either way, it is absolutely essential to understand the relationship between aggregation-prone proteins and their regulation in a healthy environment in order to restore this delicate balance in the face of misfolding diseases like AD. The results presented in this thesis identify the importance of cellular degradation and trafficking mechanisms in the regulation of AD associated metastable proteins (182), most of which are mitochondrial membrane proteins. Although mitochondrial damage and disruption has long since been associated with neurodegenerative diseases, the details of the process remain elusive. Studying the metastability of mitochondrial proteins is extremely important, as it can have major implications in understanding the details of mitochondrial disruption in neurodegenerative diseases like AD.

Thus the study of protein metastability could help to rationalise the widespread cellular disruption associated with disease pathology and also offer novel avenues for therapeutic interventions. The studies presented in this thesis also suggest that it is not just individual aggregating proteins, but the balance between protein aggregation and its regulation that is critical to maintain a healthy functional proteome. If the human brain, as an organ, is particularly vulnerable to protein misfolding diseases by virtue of possessing a weak inherent response to metastable protein aggregation, it is all the

more crucial that we understand the mechanisms of regulation of these metastable proteins to be able to intervene and rescue the neurons and potentially reverse or stall disease progression.

5.2 Further enquiries into the regulation of metastable proteins

5.2.1 Characterizing the mechanism of aggregation of metastable mitochondrial membrane proteins

As compared to cytosolic proteins, the misfolding and aggregation of membrane proteins has been poorly understood. This is particularly striking, as it is known that the hydrophobic nature of membrane proteins makes them particularly vulnerable to aggregation (318). It has been observed that hundreds of membrane proteins aggregate upon ageing in *C. elegans* (32). Indeed, a recent study reported that an α-helical protein, *Escherichia coli* lactose permease, is capable of forming amyloid like fibrils under destabilizing conditions (319). However, difficulty in purification and recombinant production are probably the major reasons behind the lack of detailed understanding of aggregation of these proteins (319). Mitochondrial membrane proteins are especially interesting due to the well acknowledged role of mitochondrial dysfunction in neurodegenerative diseases (263) and also because most of these are metastable to aggregation. One starting point to study these proteins would be to see if they aggregate upon overexpression in model organisms. This observation would provide a rationale to further isolate and characterize the aggregates and also to study the mechanistic details of their aggregation. Another interesting avenue would be to look for metastable mitochondrial membrane protein aggregation in disease models, as this could help explain the nature of mitochondrial disruption in neurodegenerative diseases.

5.2.2 Experimental validation of the protein homeostasis complement of metastable proteins

Our cells are equipped with a robust and highly dynamic protein homeostasis system to deal with protein misfolding and aggregation and the system is highly responsive to changes in cellular environment due to a variety of stresses. Disruption of the protein homeostasis system is, however, a hallmark of neurodegenerative diseases (6, 8). It is, hence, imperative to understand the regulation of metastable proteins by components of the protein homeostasis system in order to gain insight into their natural regulation within the cellular environment and use the knowledge to develop new therapeutic strategies aimed at enhancing the natural defence systems of the body. The work presented in this thesis is an initial attempt to identify these components. Experimental validation of these regulators in models of neurodegenerative diseases would pave the way for better understanding of the process.

5.2.3 Biomarker development and diagnostics

As discussed in Section 1.7, there is an immediate need to develop robust biomarkers capable of detecting not only disease progression in AD, but also the early pre-symptomatic phase of the disease. Currently available biomarkers do not necessarily give a definitive diagnosis and mostly work in conjunction with imaging techniques. The measurement of CSF Aβ and tau is also a fairly invasive technique, which makes it unsuitable as a routine screening measure. Since metastable proteins were seen to be transcriptionally downregulated in AD, they could offer another way of monitoring diseases progression. It would also be interesting to see the exact ages where this transcriptional downregulation is observed in animal models of disease. If it is seen in young animal models of disease, it could also offer an opportunity to detect the pre-symptomatic disease stages, providing a larger window to test drug efficacy.

5.2.4 Tissue vulnerability over the course of evolution

The presence of a large number of highly aggregation prone metastable proteins in our bodies prompts an important question – why have these proteins been allowed to

exist, given that they clearly pose a threat to the health of the proteome? It could be that it is the result of random mutations over the course of evolution or it could be a result of a necessary feature required by the protein sequences to be optimally functional. Indeed, as discussed in section 1.6, Vendruscolo and colleagues reported that the expression levels of human genes and the aggregation propensity of the respective polypeptide chains encoded by them are inversely correlated (49). This result suggests that protein sequences have carefully evolved over the years through random mutations and natural selection to be just soluble enough to be functional, but not more so. Hence, any disturbance in the cellular environment can cause be dangerous. The protein homeostasis system plays an important role in maintaining the solubility of proteins. The results discussed in Chapter 4 highlights the importance of the balance between the expression of these metastable proteins and the associated protein homeostasis components in determining the vulnerability of different cells and tissues to protein misfolding disorders, suggesting that brain, as on organ is particularly vulnerable. Building upon these observations it would be highly interesting to see how this balance has been maintained or evolved over the course of evolution and it might help to rationalise why humans specifically suffer from these highly debilitating neurodegenerative diseases.

5.3 Towards a gene list for Alzheimer's Disease

Our results so far have provided important insights into the mechanisms of regulation of protein misfolding in Alzheimer's disease, suggesting that protein aggregation can be reduced either by altering the conformations or the concentrations of metastable proteins. These findings suggest that in a setting of compromised protein folding the balance of these processes may move away from regulating conformations and towards regulating concentrations, by reducing expression and enhancing degradation. These concepts are very exciting, as they provide a clear strategy for identifying the key components of comprehensive list of genes associated with AD - the "AD gene list". This list will have a huge potential to revolutionize the field of AD, as it will help integrate different aspects of the disease, specifically:

a. It will facilitate the rational design of therapeutic strategies aimed at restoring the balance between aggregation-prone proteins and their corresponding natural quality control mechanisms.
b. It will help develop much-needed innovative diagnostic tests for AD. Preliminary findings suggest that we can distinguish disease mice from healthy controls prior to any detectable symptoms, potentially opening the way for an early diagnostic test for AD[i].
c. It will help create better disease models of AD. The AD models currently available do not allow us to gain a comprehensive understanding of complex multifactorial diseases like AD as almost all of them are based on genetic mutations of a very small number of genes, and hence fail to capture the cascade of molecular events associated with the disease. We expect that our gene list will enable better design of model systems to study the various pathological events associated with diseases like AD.

This approach, based on the biophysical properties of the proteins they encode and the protein homeostasis mechanisms responsible for their regulation, is an exciting endeavor, as it will allow a non-reductionist understanding of this multifactorial disease.

---

[i] This refers to preliminary unpublished studies done in collaboration with Dr. Darren Logan and Dr. Gabriela Sanchez-Andrade from The Wellcome Trust Sanger Institute, Hinxton, UK.

Appendix A

A. Additional figures on transcriptional regulation of metastable subproteome

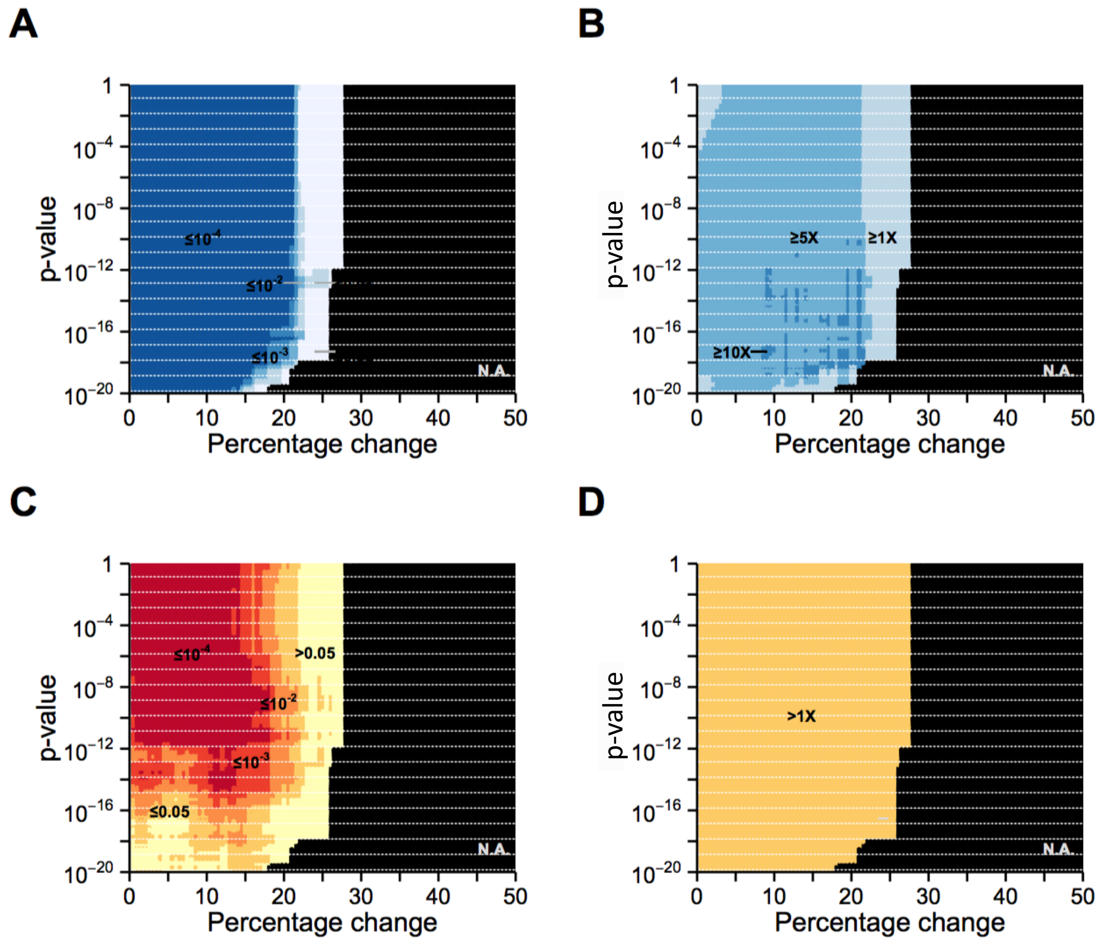


Figure A.1 Differences in metastability between transcriptionally regulated proteins in aging are robust against changes in differential expression thresholds. A range of values for thresholds of minimum percentage change (0.5–50%) and P value (10−20 to 1) was used to determine which genes are increased, decreased, or unchanged in expression upon aging. A total of 18,100 combinations were considered. Supersaturation scores were then calculated for the proteins corresponding to differentially expressed genes. The corresponding protein supersaturation was assessed in terms of (A and C) P value and (Band D) median fold difference. This analysis was performed for down-regulated (A and B) and up-regulated (C and D) genes.

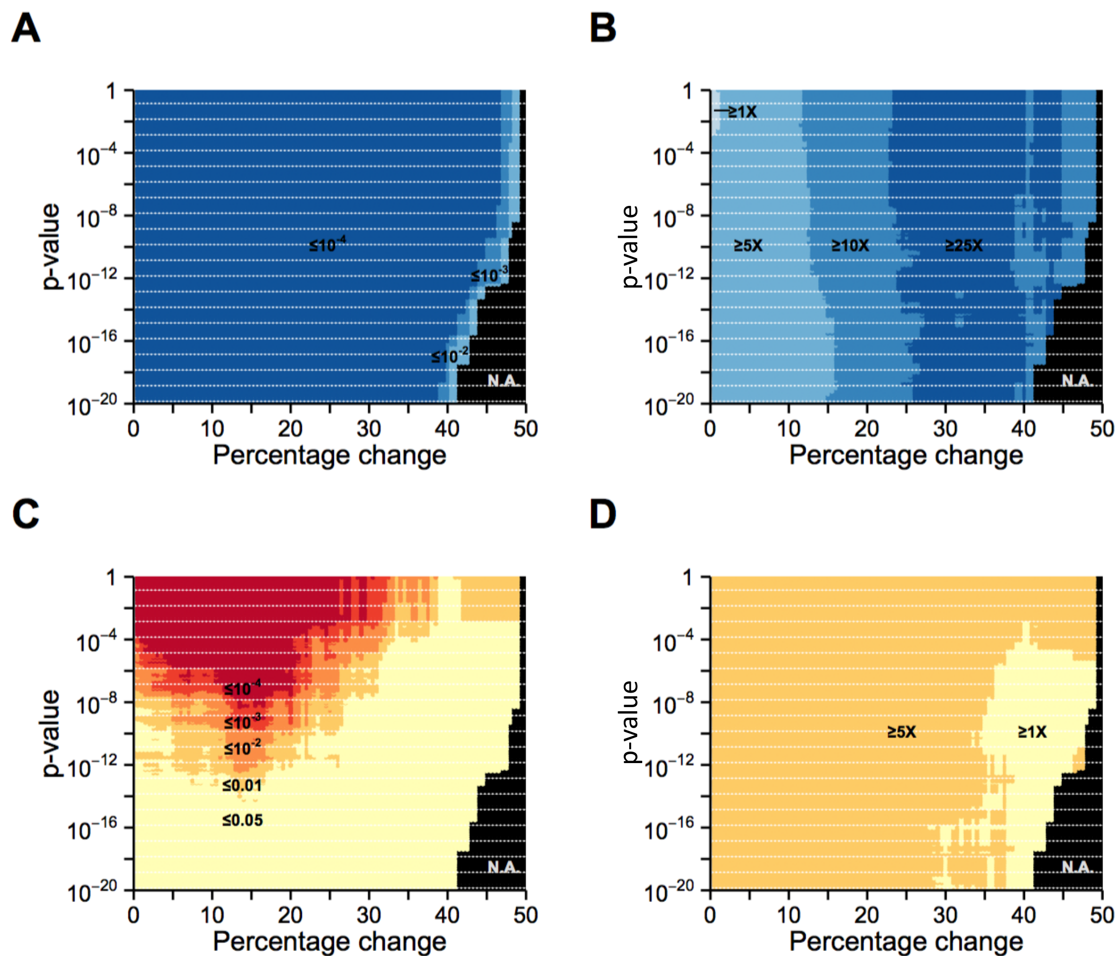


Figure A.2 Differences in metastability between transcriptionally regulated proteins in AD are robust against changes in differential expression thresholds. A range of values for thresholds of minimum percentage change (0.5–50%) and P value (10−20 to 1) was used to determine which genes are increased, decreased, or unchanged in expression in AD. A total of 18,100 combinations were considered. Supersaturation scores were then calculated for the proteins corresponding to differentially expressed genes. The corresponding protein supersaturation was assessed in terms of (A and C) P value and (Band D) median fold difference. This analysis was performed for down-regulated (A and B) and up-regulated (C and D) genes.

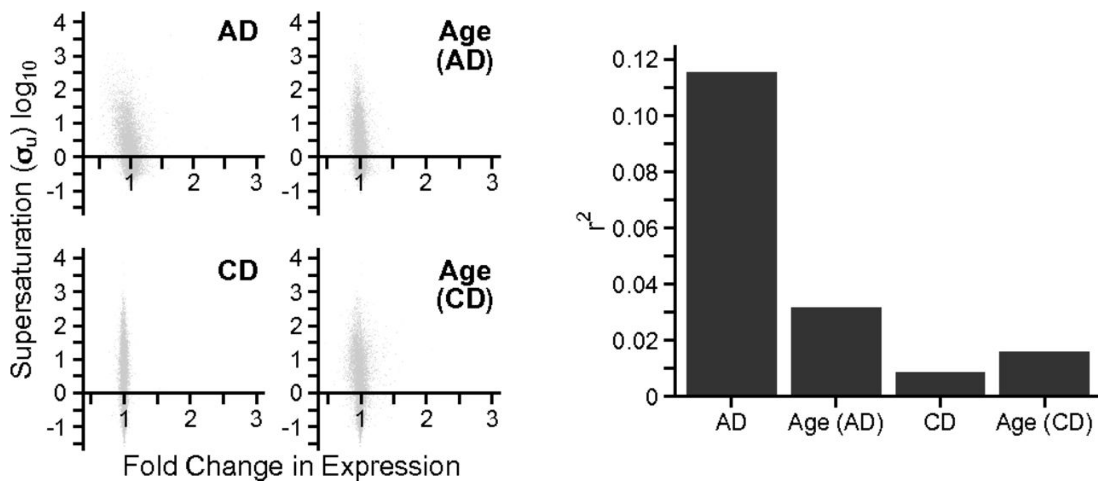


Figure A.3 Metastability levels are correlated with average expression levels for genes down-regulated in AD. (Left) Plot of protein supersaturation scores against the fold change in expression for the corresponding genes in AD (AD, Upper Left), aging based on the AD studies [Age (AD), Upper Right], clinical depression (CD, Lower Left), and aging based on the clinical depression studies [Age (CD), Lower Right]. (Right) Pearson's correlation coefficient (r2) for the categories plotted (Left).

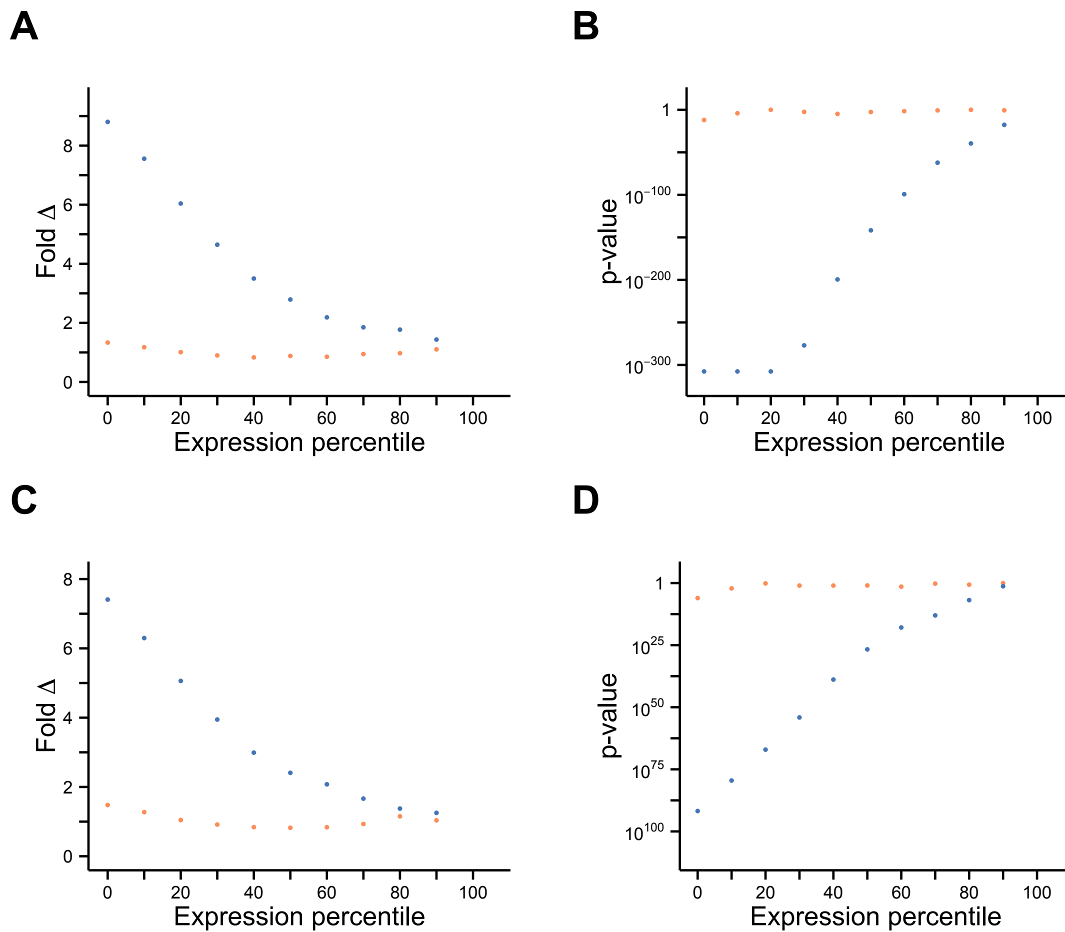


Figure A.4 Metastability of proteins encoded by differentially expressed genes is elevated in AD for a range of expression values. Supersaturation of proteins associated with AD (A and B) and ageing (C and D) was determined after restricting the genes of interest to those above a range of expression levels plotted by expression percentile rank. (A and C) Fold Δ and (B and D) P value are plotted. Blue points represent values for down-regulated genes; orange points represent values for up-regulated genes. The median fold difference in supersaturation is indicated by Fold Δ. P values are calculated using the one-sided Wilcoxon/Mann–Whitney test with Holm–Bonferroni correction.

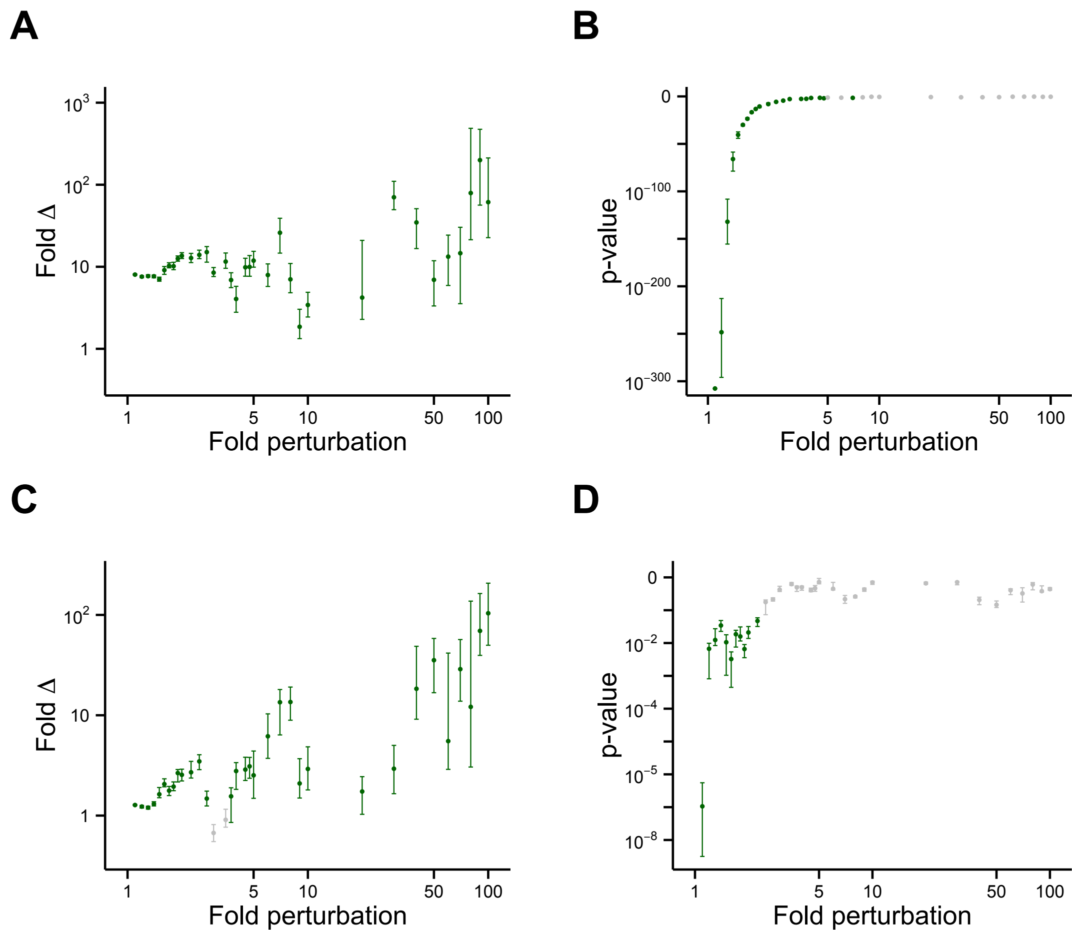


Figure A.5 Differences in metastability between transcriptionally regulated proteins in AD are robust against Gaussian noise in the supersaturation score. Test of the robustness of the significance of the (A and C) median fold difference and (B and D) P value of supersaturation for proteins transcriptionally (A and B) down-regulated or (C and D) up-regulated in AD. Gaussian noise was introduced 100 independent times into the proteome scores at levels ranging from 1.1× to 100× (where 1× signifies no noise). Tests were performed at each noise level to determine whether the 100 median fold differences obtained were significantly greater than 1 and the 100 P values obtained were significantly below 0.05. For down-regulated genes, supersaturation (A) median fold difference is robust up to 100× and (B) P value is robust up to 7×. For up-regulated genes, supersaturation (C) median fold difference is robust up to 100× and (D) P value is robust up to 2.25×. Error bars indicate interquartile ranges; green points indicate $P \leq 0.05$ by the one-sided Wilcoxon/Mann–Whitney test.

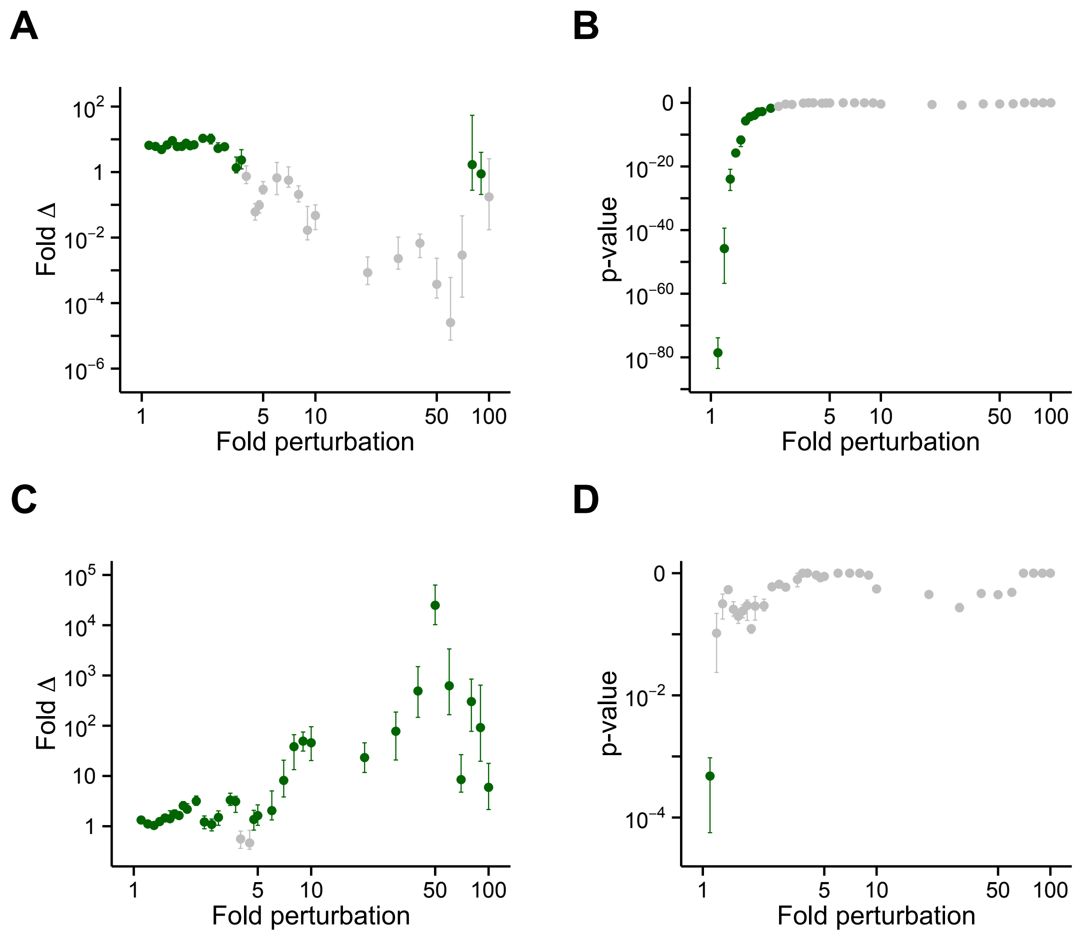


Figure A.6 Differences in metastability between transcriptionally regulated proteins in aging are robust against Gaussian noise in the supersaturation score. Test of the robustness of the significance of the (A and C) median fold difference and (B and D) P value of supersaturation for proteins transcriptionally (A and B) down-regulated or (C and D) up-regulated in aging (AD dataset). Gaussian noise was introduced 100 independent times into the proteome scores at levels ranging from 1.1× to 100× (where 1× signifies no noise). Tests were performed at each noise level to determine whether the 100 median fold differences obtained were significantly greater than 1 and the 100 P values obtained were significantly below 0.05. For down-regulated genes, supersaturation (A) median fold difference is robust up to 3.75× and (B) Pvalue is robust up to 2.25×. For up-regulated genes, supersaturation (C) median fold difference is robust up to 100× and (D) P value is robust up to 1.1×. Error bars indicate interquartile ranges; green points indicate $P \leq 0.05$ by the one-sided Wilcoxon/Mann–Whitney test.

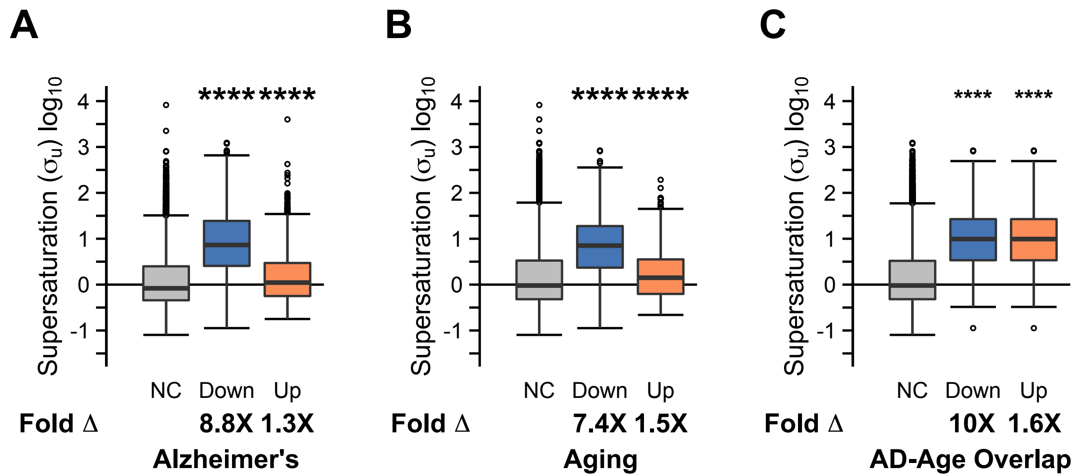


Figure A.7 Elevated metastability of proteins encoded by differentially expressed genes in AD and aging is not dependent on oxidative phosphorylation proteins. Supersaturation of proteins associated with differentially expressed genes in (A) AD, (B) aging, and (C) the overlap between the two, but with those proteins found in the KEGG pathway for oxidative phosphorylation excluded. The median fold difference in supersaturation is indicated by Fold Δ. NC indicates genes that do not change significantly in expression. ****$P \leq 0.0001$, one-sided Wilcoxon/Mann–Whitney test with Holm–Bonferroni correction. Whiskers range from the lowest to highest value data points within 150% of the interquartile ranges.

Series	Platform	Refs.	Control Samples	Disease Samples	Disease Set
GSE1297	GPL96	(220)	74	87	AD
GSE5281	GPL96	(222, 223)	9	22	AD
GSE15222	GPL2700	(224)	187	174	AD
GSE26927	GPL6255	(228)	7	11	AD
GSE29378	GPL6947	(231)	32	31	AD
GSE29652	GPL570	(226)	6	12	AD
GSE36980	GPL6244	(230)	47	32	AD
GSE37263	GPL5175	(225)	8	8	AD
GSE44772	GPL4272	(320)	299	388	AD
GSE12654	GPL8300	(Iwamoto, Mol Psychiatry 2004)	15	11	CD
GSE53987	GPL570	(Lanz, PLoS One, 2015)	55	50	CD
GSE54562, GSE54563, GSE54564	GPL6947	(Chang et al, PLoS One, 2014)	56	56	CD
GSE54565, GSE54566	GPL570	(Chang et al, PLoS One, 2014)	29	30	CD
GSE54567, GSE54568, GSE54571, GSE54572	GPL570	(Chang et al, PLoS One, 2014)	54	54	CD
GSE24095	GPL10907	(Duric et al, Nat Med, 2010)	30	30	CD

Table A.1 List of the studies used for the microarray meta-analyses carried out in this work for Alzheimer's disease and clinical depression

Appendix B

B. Additional figures on protein homeostasis of a metastable subproteome

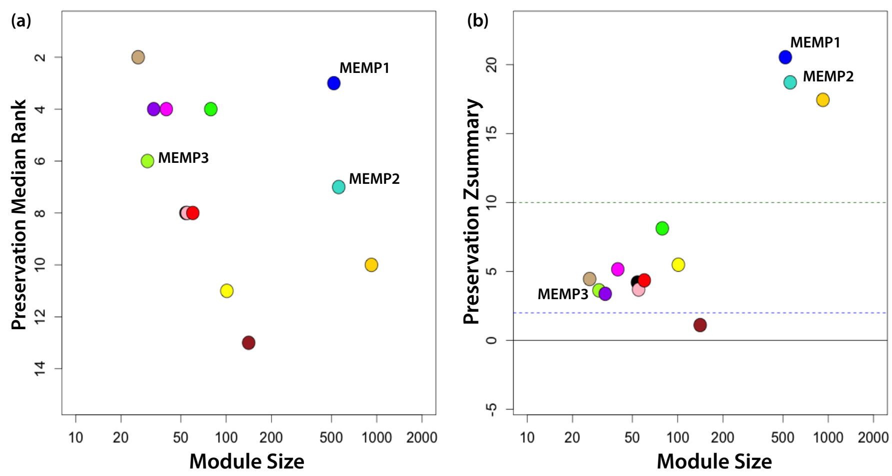
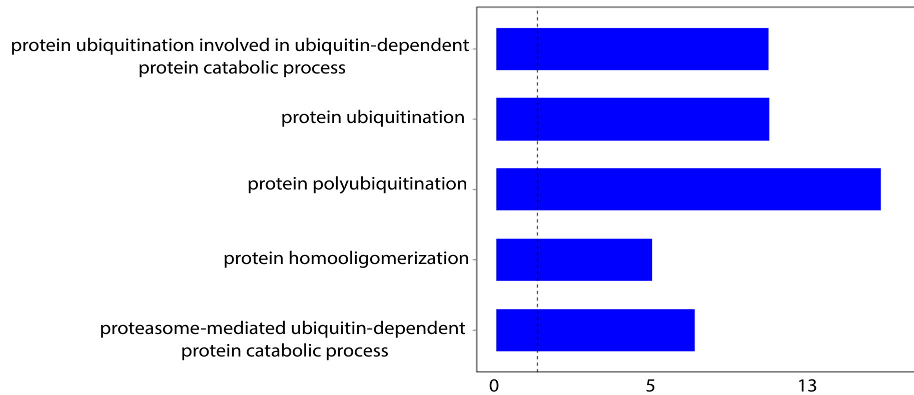
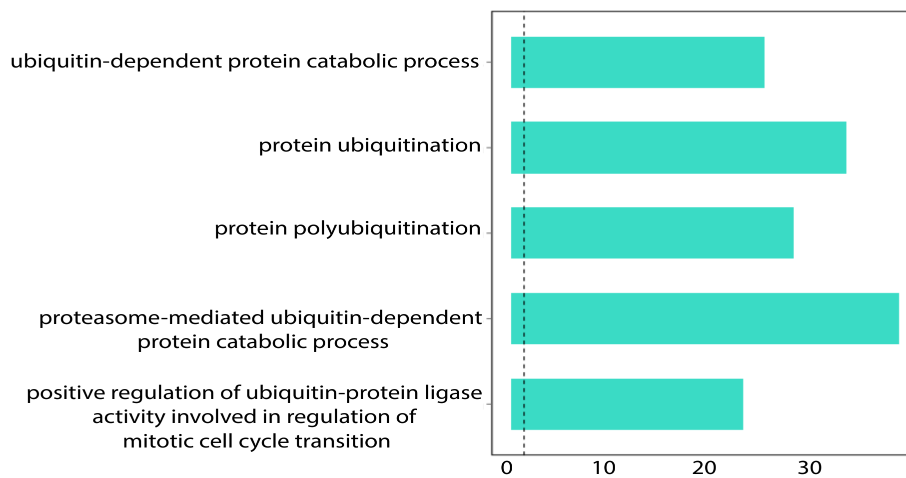


Figure B.1 Module preservation of MEMPs across different datasets. (a) Preservation median rank, and (b) preservation Zsummary scores of various modules.





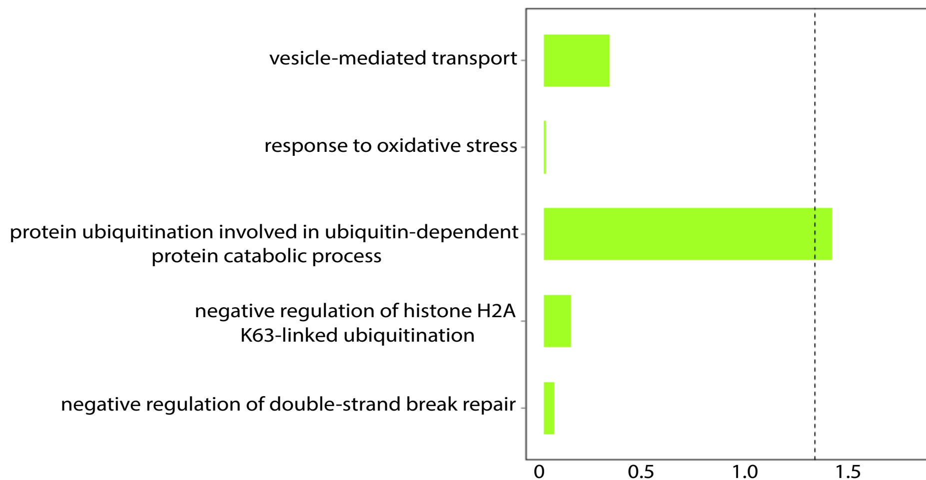


Figure B.2 GO enrichment analysis for the genes in the three main MEMPs (MEMP1, MEMP2 and MEMP3), which identifies specific components of the ubiquitin-proteasome and endosomal-lysosomal systems in the regulation of the metastable subproteome.

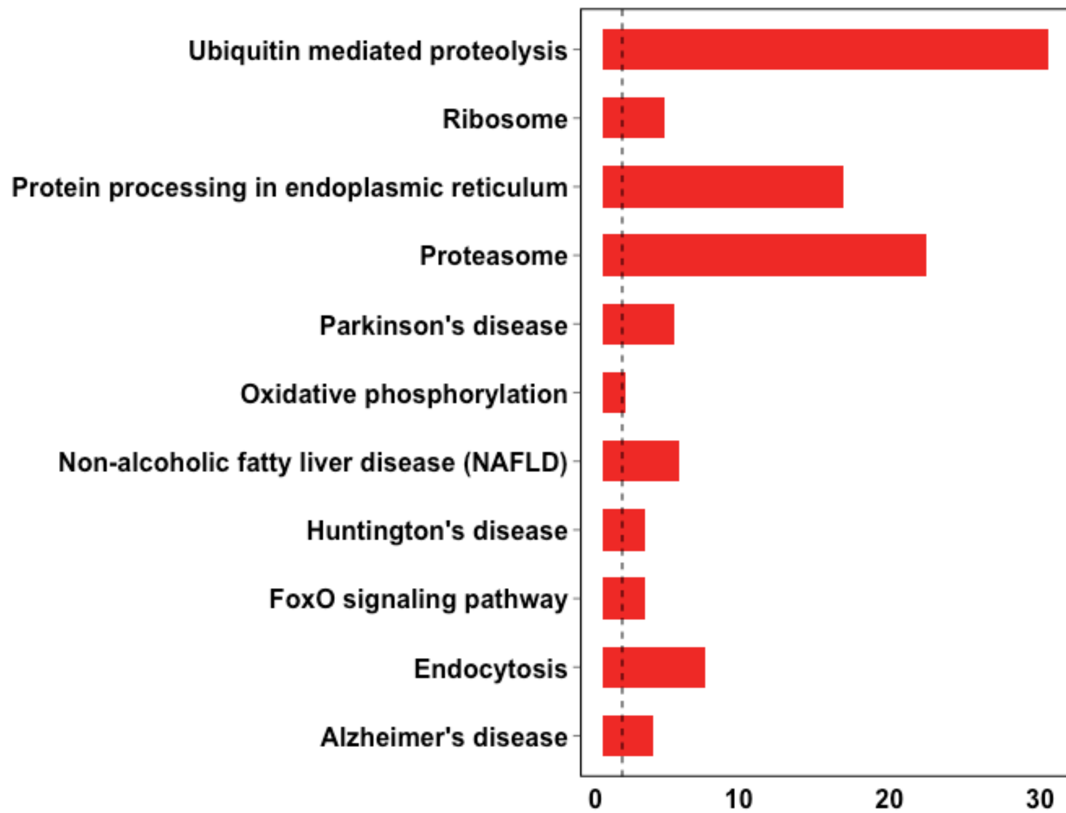


Figure B.3 KEGG pathways enriched in the MEMP-1, MEMP-2 and MEMP3 modules. UMP: ubiquitin mediated proteolysis, Pr: proteasome, Rb: ribosome, PD: Parkinson's disease, AD: Alzheimer's disease, HD: Huntington's disease, Endo: endocytosis, Adipo Sg: adipocytokine signaling pathway.

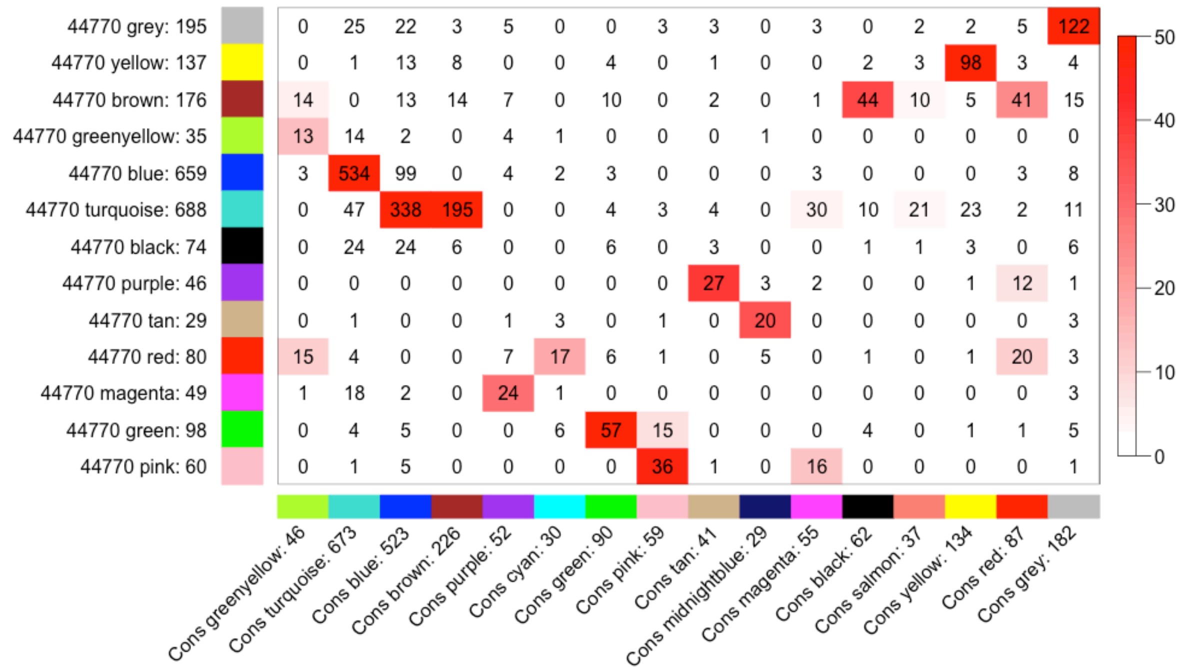


Figure B.4 Correspondence between the modules of the visual cortex dataset (GSE44771) and the consensus modules. Each row of the table corresponds to one of the modules of the visual cortex dataset (labelled by colour as well as text), and each column corresponds to one consensus module. Numbers in the table indicate the gene counts in the intersections of the corresponding modules. Colouring of the table encodes −log(p), with p being the Fisher's exact test p-value for the overlap of the two modules. The table indicates that most of the modules of the visual cortex dataset have a consensus counterpart.

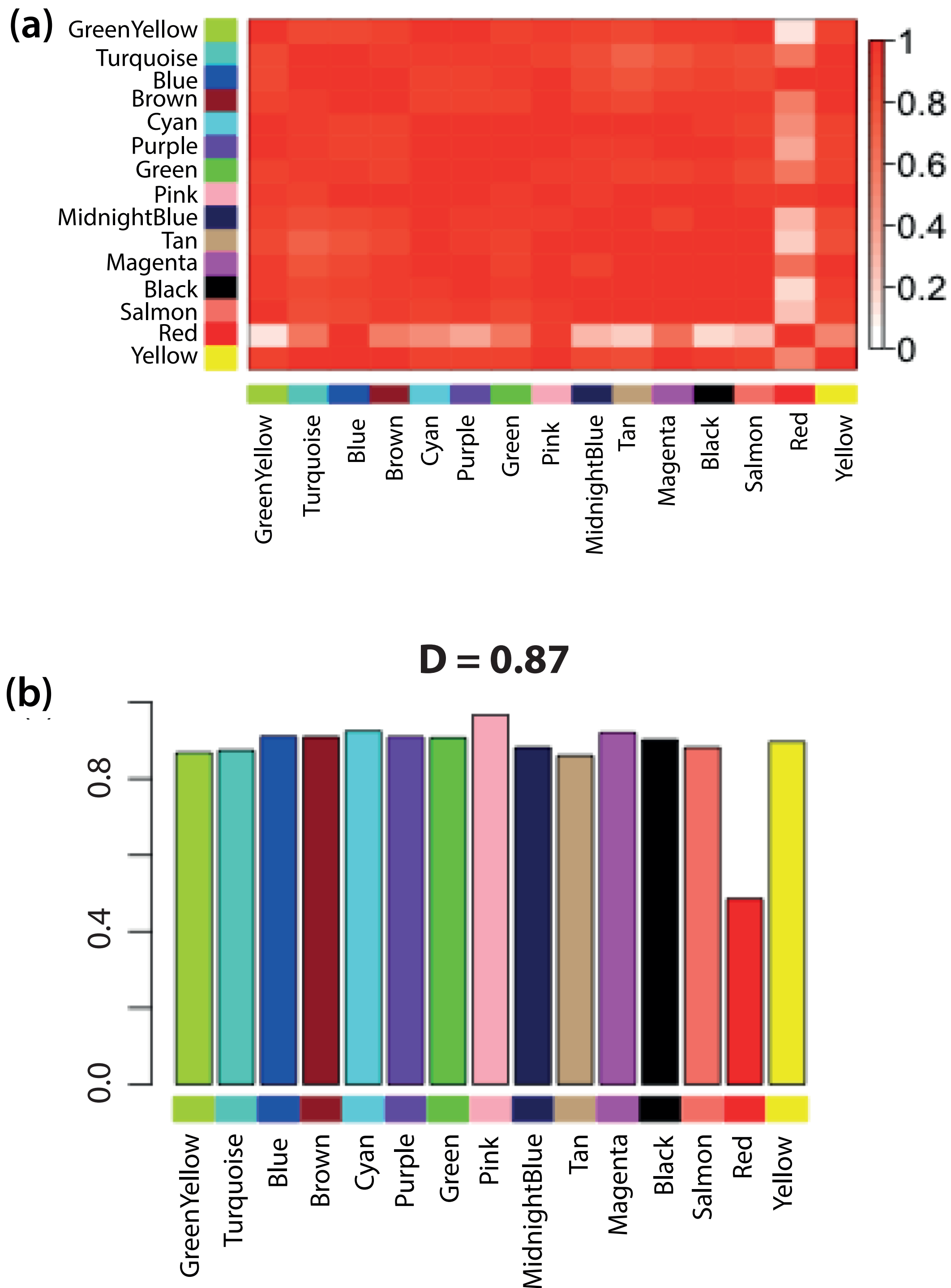


Figure B.5 Consensus eigengene networks for the dorsolateral prefrontal cortex and the visual cortex. (a) Heatmap of the preservation network, defined as one minus the absolute difference of the eigengene networks in the two data sets. (b) Mean preservation of adjacency for each of the eigengenes to all other eigengenes. D denotes the mean preservation of eigengene networks among the datasets. $\boldsymbol{D}^{(1,2\ldots)} = \boldsymbol{mean}_{i<j}\boldsymbol{P}_{i,j}^{(1,2\ldots)}$.

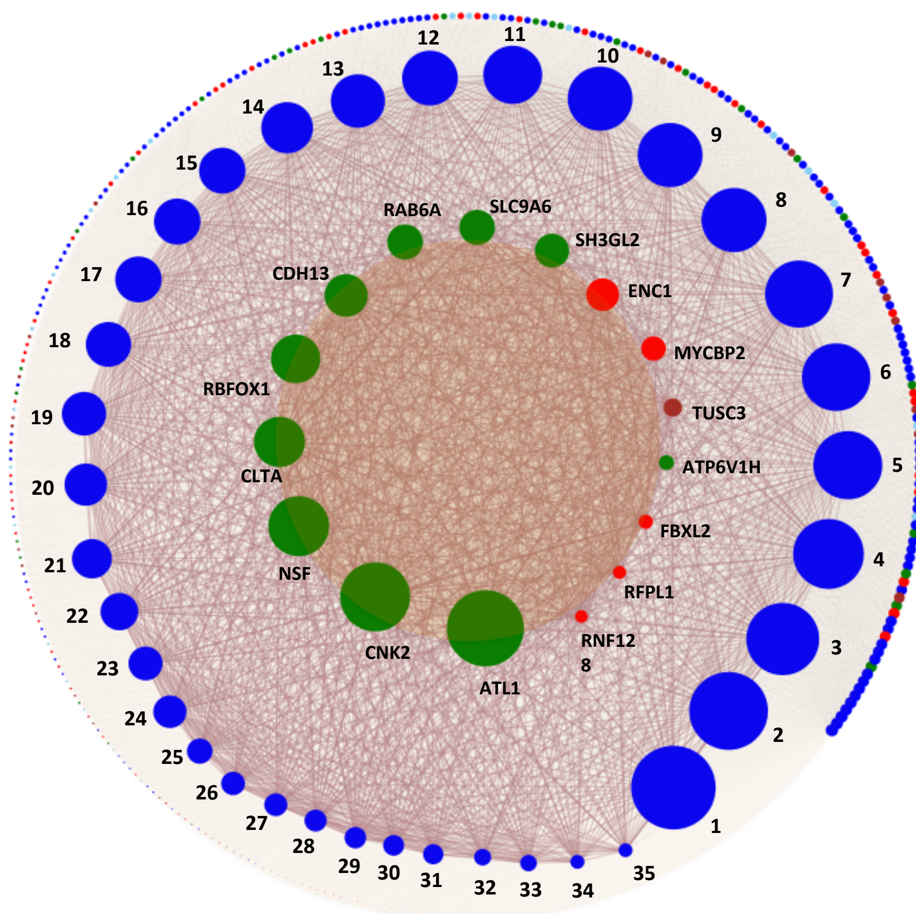


Figure B.6 Network representation of the MEMPs showing the hub genes and the main components of the protein homeostasis system linked with the AD metastable subproteome. This analysis reveals in particular the importance of the ubiquitin-proteasome (red) and trafficking (green) systems in the regulation of aggregation-prone proteins in Alzheimer's disease. The top 10% of the hub gene interactions were visualized, with those genes involved in at least 50 interactions shown in the centre. The sizes of the nodes correspond to their degrees of connectivity. The metastable genes are shown in blue. Table B.5 reports the names of the metastable genes according to the numerical labels shown here.

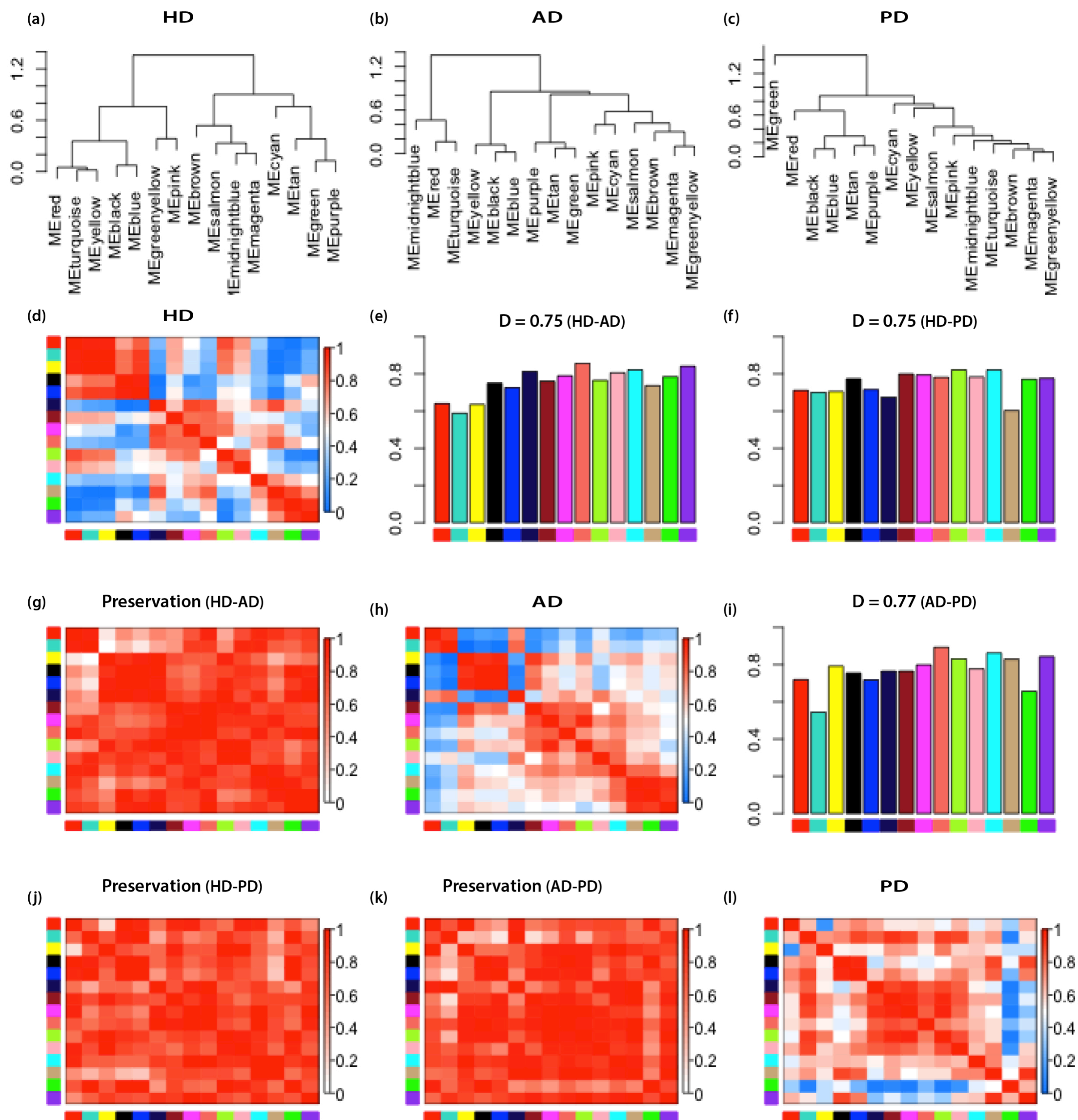


Figure B.7 Consensus eigengene networks and their differential analysis. (a-c) Dendrograms (clustering trees) of the consensus module eigengenes in the three datasets. (d, h, l) Eigengene network heatmaps - red denotes high adjacency (positive correlation) and blue denotes low adjacency. (g, j, k) Heatmaps of the preservation network, defined as one minus the absolute difference of the eigengene networks in the two data sets. (e, f, i) Mean preservation of adjacency for each of the eigengenes to all other eigengenes.

Disease	Dataset	Tissue	Number of Disease Samples	Number of Controls
Alzheimer's Disease	GSE44770	Pre-Frontal Cortex	118	73
Alzheimer's Disease	GSE44771	Visual Cortex	124	69
Alzheimer's Disease	GSE1297	Hippocamplus	22	9
Huntington's Disease	GSE33000	Dorsolateral Pre-Frontal Cortex	157	157
Parkinson's Disease	GSE20292	Substantia Niagra	11	18

Table B-1 List of datasets used for the analysis described in Chapter 3

Module	Genes encoding for metastable proteins	Total genes
black	10	74
blue	220	659
brown	0	176
green	0	98
greenyellow	10	35
grey	2	195
magenta	4	49
pink	1	60
purple	1	46
red	1	80
tan	0	29
turquoise	91	688
yellow	3	137

Table B-2 Total number of genes and number of genes encoding for metastable proteins in different modules

	ME black	**ME blue**	ME brown	ME green	**ME greenyellow**	ME grey	ME magenta	ME pink	ME purple	ME red	ME tan	**ME turquoise**	ME yellow
ME black	1	0.72	-0.56	0.51	0.11	0.053	-0.14	-0.11	0.36	0.3	-0.03	0.87	-0.85
ME blue	0.72	1	0.0078	0.015	**0.68**	0.12	-0.76	-0.45	-0.0076	-0.41	-0.45	**0.78**	-0.42
ME brown	-0.56	0.0078	1	-0.5	0.66	0.29	-0.42	0.031	-0.59	-0.69	-0.28	-0.56	0.64
ME green	0.51	0.015	-0.5	1	-0.27	0.0047	0.47	0.64	0.16	0.65	0.22	0.28	-0.49
ME greenyellow	0.11	0.68	0.66	-0.27	1	0.055	-0.79	-0.36	-0.32	-0.7	-0.48	**0.23**	0.092
ME grey	0.053	0.12	0.29	0.0047	0.055	1	-0.17	0.43	-0.25	-0.16	0.064	-0.26	0.061
ME magenta	-0.14	-0.76	-0.42	0.47	-0.79	-0.17	1	0.59	0.22	0.83	0.54	-0.36	-0.16
ME pink	-0.11	-0.45	0.031	0.64	-0.36	0.43	0.59	1	-0.25	0.44	0.3	-0.48	0.092
ME purple	0.36	- 0.0076	-0.59	0.16	-0.32	-0.25	0.22	-0.25	1	0.61	0.75	0.42	-0.64
ME red	0.3	-0.41	-0.69	0.65	-0.7	-0.16	0.83	0.44	0.61	1	0.71	0.086	-0.59
ME tan	-0.03	-0.45	-0.28	0.22	-0.48	0.064	0.54	0.3	0.75	0.71	1	-0.17	-0.33
ME turquoise	0.87	0.78	-0.56	0.28	0.23	-0.26	-0.36	-0.48	0.42	0.086	-0.17	1	-0.72
ME yellow	-0.85	-0.42	0.64	-0.49	0.092	0.061	-0.16	0.092	-0.64	-0.59	-0.33	-0.72	1

Table B-3 Pearson's correlation coefficients between various module eigengenes

Molecular chaperone	Module
FKB1B_HUMAN	blue
TUSC3_HUMAN	blue
FAXC_HUMAN	blue
AUXI_HUMAN	blue
TTC9A_HUMAN	blue
ERLEC_HUMAN	blue
DNJC5_HUMAN	blue
AFG32_HUMAN	blue
DJC18_HUMAN	blue
MKKS_HUMAN	blue
HSPB3_HUMAN	blue
OPA1_HUMAN	blue
TOR1A_HUMAN	blue
SRP68_HUMAN	blue
DJC12_HUMAN	blue
TOM34_HUMAN	blue
DJC27_HUMAN	blue
PFD4_HUMAN	blue
PSMG1_HUMAN	blue
TTC36_HUMAN	blue
TOM70_HUMAN	turquoise
FKBP3_HUMAN	turquoise
TMX3_HUMAN	turquoise
TIM14_HUMAN	turquoise
TTC33_HUMAN	turquoise
PFD3_HUMAN	turquoise
TTC3_HUMAN	turquoise
CDC27_HUMAN	turquoise
CD37L_HUMAN	turquoise
AN13C_HUMAN	turquoise
SACS_HUMAN	turquoise
THIO_HUMAN	turquoise
TCPZ_HUMAN	turquoise
ERO1B_HUMAN	turquoise
KLC1_HUMAN	turquoise
CLGN_HUMAN	turquoise
RPAP3_HUMAN	turquoise
DNJA3_HUMAN	turquoise
PFD1_HUMAN	turquoise
GRP75_HUMAN	turquoise
TTC13_HUMAN	turquoise
FKB10_HUMAN	turquoise
CABIN_HUMAN	turquoise
TBCE_HUMAN	turquoise

Table B-4 List of molecular chaperones in the MEMPs

Number	Uniprot ID
1	SV2B_HUMAN
2	NEUS_HUMAN
3	MOAP1_HUMAN
4	NFL_HUMAN
5	GLRB_HUMAN
6	NPTN_HUMAN
7	AP180_HUMAN
8	GP158_HUMAN
9	SCG1_HUMAN
10	NP1L5_HUMAN
11	CRYM_HUMAN
12	GASP1_HUMAN
13	7B2_HUMAN
14	RTN1_HUMAN
15	MAGE1_HUMAN
16	GASP2_HUMAN
17	AF1Q_HUMAN
18	DCLK1_HUMAN
19	NBEA_HUMAN
20	KCRU_HUMAN
21	BEX5_HUMAN
22	ENOG_HUMAN
23	AT2B1_HUMAN
24	AMPH_HUMAN
25	SYT13_HUMAN
26	TSN13_HUMAN
27	SNP25_HUMAN
28	NP1L2_HUMAN
29	XK_HUMAN
30	OPCM_HUMAN
31	ELOV4_HUMAN
32	LDB2_HUMAN
33	SCN2A_HUMAN
34	RIFK_HUMAN
35	PLK2_HUMAN

Table B-5 Names of the metastable genes according to the numerical labels.

Appendix C

C. Additional figures on a map of protein aggregation homeostasis

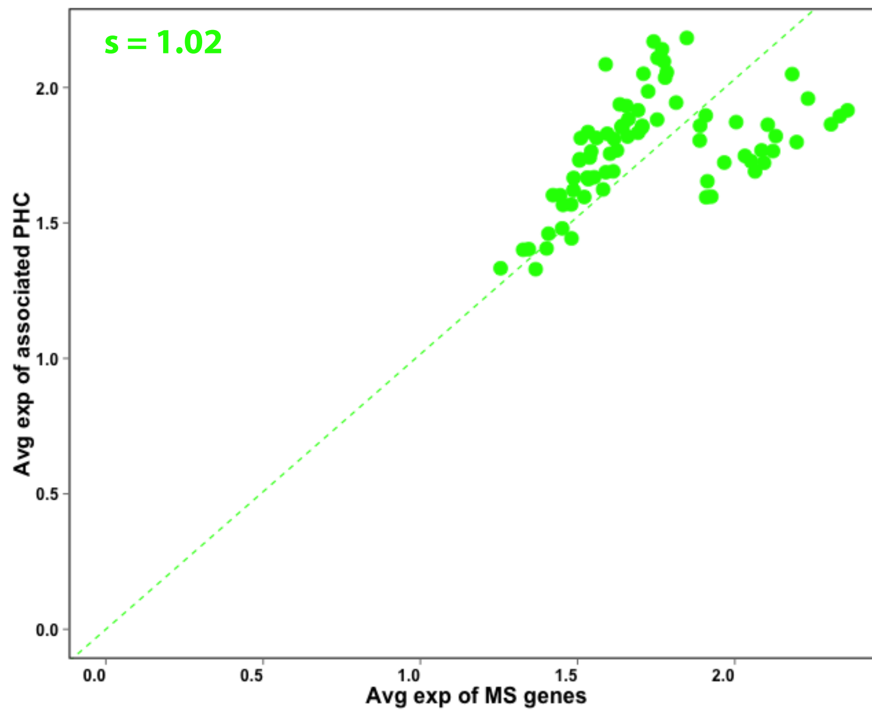


Figure C.1 The protein homeostasis response is proportional to the risk of aggregation. Scatter plot showing the average expression of metastable genes (MS) and the average expression of the associated protein homeostasis components (PHC) in different tissues. The correlation between the average expression of the MS and the average expression of the its PHC reveals the presence of a robust response to the presence of aggregation-prone proteins.

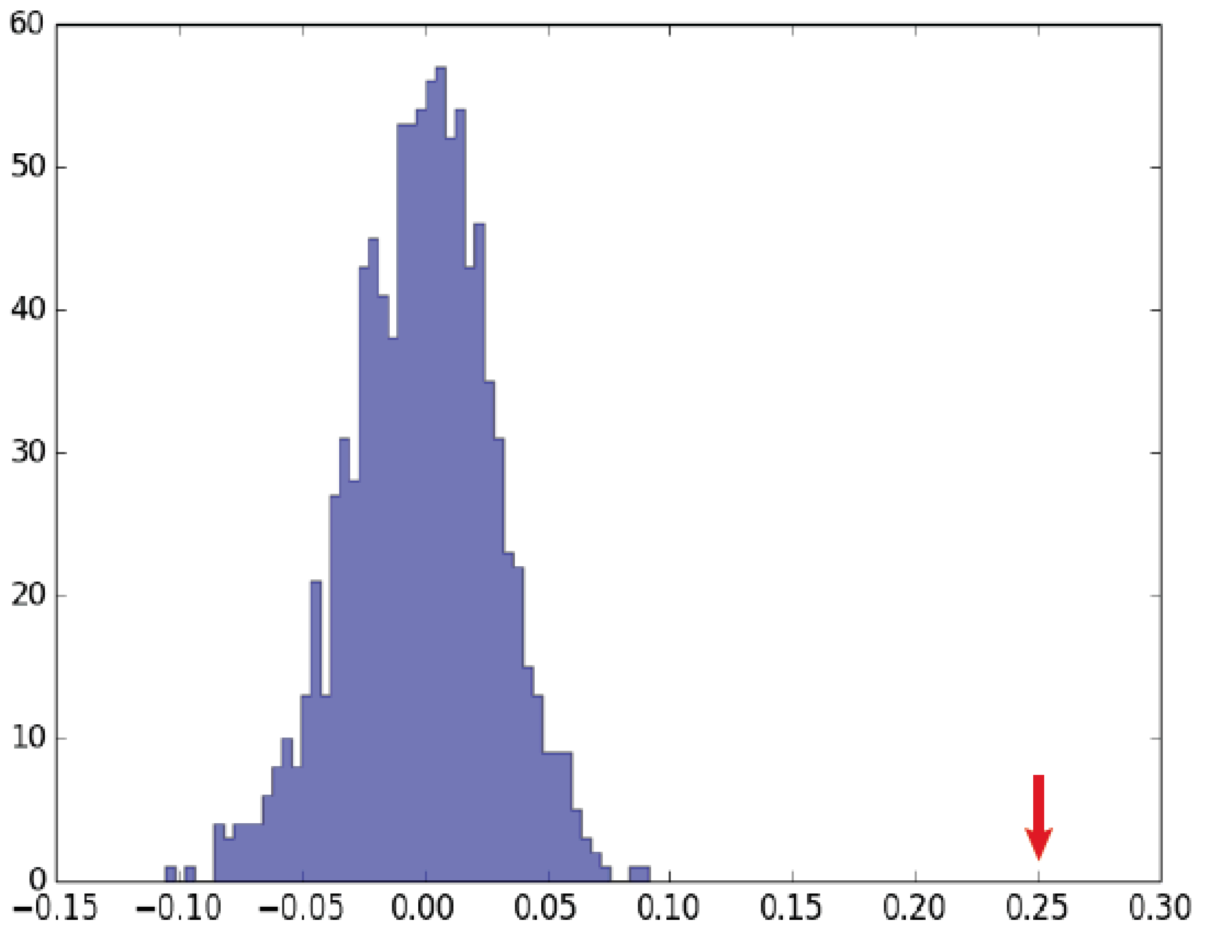


Figure C.2 Histogram showing Δs (difference in slope for body tissues and brain tissues) for 1000 random sets of genes. The delta slope for our genes of interest (genes that are supersaturated and downregulated only in AD, and the associated PHC) is 0.25 (shown by red arrow).

Tissue	**Classification (1)**	**Classification (2)**
Adipocyte	Body	-
AdrenalCortex	Body	-
Adrenalgland	Body	-
Appendix	Body	-
AtrioventricularNode	Body	-
BDCA4+_DentriticCells	Body	-
Bonemarrow	Body	-
BronchialEpithelialCells	Body	-
CD105+_Endothelial	Body	-
CD14+_Monocytes	Body	-
CD19+_BCells(neg._sel.)	Body	-
CD33+_Myeloid	Body	-
CD34+	Body	-
CD4+_Tcells	Body	-
CD56+_NKCells	Body	-
CD71+_EarlyErythroid	Body	-
CD8+_Tcells	Body	-
CardiacMyocytes	Body	-
FetalThyroid	Body	-
Fetalliver	Body	-
Fetallung	Body	-
Heart	Body	-
Kidney	Body	-
Liver	Body	-
Lung	Body	-
Lymphnode	Body	-
Ovary	Body	-
Pancreas	Body	-
PancreaticIslet	Body	-
Placenta	Body	-
Prostate	Body	-
Salivarygland	Body	-
SkeletalMuscle	Body	-
Skin	Body	-
SmoothMuscle	Body	-
Testis	Body	-
TestisGermCell	Body	-
TestisIntersitial	Body	-
TestisLeydigCell	Body	-
TestisSeminiferousTubule	Body	-
Thymus	Body	-
Thyroid	Body	-
Tongue	Body	-

Tonsil	Body	-
Trachea	Body	-
Uterus	Body	-
UterusCorpus	Body	-
WholeBlood	Body	-
colon	Body	-
small_intestine	Body	-
Amygdala	Brain	Braak
Caudatenucleus	Brain	Non Braak
Cerebellum	Brain	Non Braak
CerebellumPeduncles	Brain	Non Braak
CiliaryGanglion	Brain	Non Braak
CingulateCortex	Brain	Braak
DorsalRootGanglion	Brain	Non Braak
Fetalbrain	Brain	Non Braak
GlobusPallidus	Brain	Non Braak
Hypothalamus	Brain	Braak
MedullaOblongata	Brain	Non Braak
OccipitalLobe	Brain	Braak
OlfactoryBulb	Brain	Non Braak
ParietalLobe	Brain	Non Braak
Pons	Brain	Non Braak
PrefrontalCortex	Brain	Braak
Spinalcord	Brain	Non Braak
SubthalamicNucleus	Brain	Non Braak
SuperiorCervicalGanglion	Brain	Non Braak
TemporalLobe	Brain	Braak
Thalamus	Brain	Braak
TrigeminalGanglion	Brain	Non Braak
Wholebrain	Brain	Non Braak
pineal_day	Brain	Non Braak
pineal_night	Brain	Non Braak
Pituitary	Brain	Non Braak
retina	Brain	Non Braak

Table C.1 List of different tissues used in the analysis alongwith their classification

References

1. 2015 Alzheimer's disease facts and figures. *Alzheimer's and Dementia : the Journal of the Alzheimer's Association* 11:332-384.
2. Krstic D & Knuesel I (2013) Deciphering the mechanism underlying late-onset Alzheimer disease. *Nature Reviews Neurology* 9:25-34.
3. Selkoe DJ & Hardy J (2016) The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Molecular Medicine* 8(6):595-608.
4. Holtzman DM, Morris JC, & Goate AM (2011) Alzheimer’s disease: the challenge of the second century. *Science Translational Medicine* 3(77):77sr71.
5. Balch WE, Morimoto RI, Dillin A, & Kelly JW (2008) Adapting proteostasis for disease intervention. *Science* 319:916-919.
6. Hipp MS, Park S-H, & Hartl FU (2014) Proteostasis impairment in protein-misfolding and-aggregation diseases. *Trends in Cell Biology* 24(9):506-514.
7. Knowles TP, Vendruscolo M, & Dobson CM (2014) The amyloid state and its association with protein misfolding diseases. *Nature Reviews Molecular Cell biology* 15(6):384-396.
8. Labbadia J & Morimoto RI (2015) The biology of proteostasis in aging and disease. *Annual Review of Biochemistry* 84:1-30.
9. Eisenberg D & Jucker M (2012) The amyloid state of proteins in human diseases. *Cell* 148(6):1188-1203.
10. De Strooper B & Karran E (2016) The cellular phase of Alzheimer's disease. *Cell* 164:603-615.
11. Chiti F & Dobson CM (2006) Protein misfolding, functional amyloid, and human disease. *Annual Review of Biochemistry* 75:333-366.
12. Dobson C (2017) Amyloid formation, protein homeostasis, and human disease: A summary of progress over the last decade. *Annual Review of Biochemistry* 86:1-42.
13. Dobson CM (2015) Alzheimer’s disease: addressing a twenty-first century plague. *Rendiconti Lincei* 26:251-262.
14. Dobson CM (1999) Protein misfolding, evolution and disease. *Trends Biochem. Sci.* 24:329-332.
15. Querfurth HW & LaFerla FM (2010) Mechanisms of disease: Alzheimer’s disease. *New England Journal of Medicine* 362(4):329-344.
16. Selkoe D, Mandelkow E, & Holtzman D (2012) Deciphering Alzheimer disease. *Cold Spring Harbor Perspectives in Medicine* 2(1):a011460.
17. Hardy J & Selkoe DJ (2002) The amyloid hypothesis of Alzheimer's disease: Progress and problems on the road to therapeutics. *Science* 297(5580):353-356.
18. Cuyvers E & Sleegers K (2016) Genetic variations underlying Alzheimer's disease: evidence from genome-wide association studies and beyond. *The Lancet Neurology* 15:857-868.
19. Borchelt DR*, et al.* (1996) Familial Alzheimer's Disease–linked Presenilin 1 variants elevate Aβ1–42/1–40 ratio in vitro and in vivo. *Neuron* 17:1005-1013.

20. Citron M, *et al.* (1997) Mutant presenilins of Alzheimer's disease increase production of 42-residue amyloid beta-protein in both transfected cells and transgenic mice. *Nature Medicine* 3:67-72.
21. Cruts M & Van Broeckhoven C (1998) Presenilin mutations in Alzheimer's disease. *Human Mutation* 11:183-190.
22. Duff K, *et al.* (1996) Increased Amyloid-β42(43) in Brains of Mice Expressing Mutant Presenilin 1. in *Nature*, pp 710-713.
23. Goate A, *et al.* (1991) Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature* 349:704-706.
24. Jankowsky JL, *et al.* (2004) Mutant presenilins specifically elevate the levels of the 42 residue beta-Amyloid peptide in vivo: evidence for augmentation of a 42-specific gamma-secretase. *Human Molecular Genetics* 13:159-170.
25. Lemere CA, *et al.* (1996) The E280A presenilin 1 Alzheimer mutation produces increased A beta 42 deposition and severe cerebellar pathology. *Nature Medicine* 2:1146-1150.
26. Scheuner D, *et al.* (1996) Secreted amyloid beta-protein similar to that in the senile plaques of Alzheimer's disease is increased in vivo by the presenilin 1 and 2 and APP mutations linked to familial Alzheimer's disease. *Nature Medicine* 2:864-870.
27. Pericak-Vance MA, *et al.* (1991) Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. *American Journal of Human Genetics* 48:1034-1050.
28. Strittmatter WJ, *et al.* (1993) Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proceedings of the National Academy of Sciences* 90:1977-1981.
29. David DC, *et al.* (2010) Widespread protein aggregation as an inherent part of aging in C. elegans. *PLoS biology* 8:e1000450.
30. Gidalevitz T, Ben-Zvi A, Ho KH, Brignull HR, & Morimoto RI (2006) Progressive disruption of cellular protein folding in models of polyglutamine diseases. *Science* 311(5766):1471-1474.
31. Olzscha H, *et al.* (2011) Amyloid-like aggregates sequester numerous metastable proteins with essential cellular functions. *Cell* 144(1):67-78.
32. Walther DM, *et al.* (2015) Widespread proteome remodeling and aggregation in aging C. elegans. *Cell* 161(4):919-932.
33. Ciryam P, Kundra R, Morimoto RI, Dobson CM, & Vendruscolo M (2015) Supersaturation is a major driving force for protein aggregation in neurodegenerative diseases. *Trends in Pharmacological Sciences* 36:72-77.
34. Ciryam P, Tartaglia GG, Morimoto RI, Dobson CM, & Vendruscolo M (2013) Widespread aggregation and neurodegenerative diseases are associated with supersaturated proteins. *Cell Reports* 5:781-790.
35. Hartl FU, Bracher A, & Hayer-Hartl M (2011) Molecular chaperones in protein folding and proteostasis. *Nature* 475(7356):324-332.
36. Dobson CM, Sali A, & Karplus M (1998) Protein folding: A perspective from theory and experiment. *Angewandte Chemie - International Edition* 37:868-893.
37. Dill K & Chan HS (1997) From Levinthal to pathways to funnels. *Nature structural biology* 4:10-19.
38. Anfinsen (1973) Principles that govern the folding of protein chains. *Science* 181:223-230.

39. Anfinsen CB, Haber E, Sela M, & White FH (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America* 47:1309-1314.
40. Levinthal C (1968) Are there pathways for protein folding? *J. Chim. Phys.* 65:44-45.
41. Eakin CM, Berman AJ, & Miranker AD (2006) A native to amyloidogenic transition regulated by a backbone trigger. *Nature Structural and Molecular Biology* 13:202-208.
42. Sekijima Y*, et al.* (2005) The biological and chemical basis for tissue-selective amyloid disease. *Cell* 121:73-85.
43. Baldwin RL (1995) The nature of protein folding pathways: The classical versus the new view. *Journal of Molecular Biology* 5:103-109.
44. Bryngelson JD, Onuchic JN, Socci ND, & Wolynes PG (1995) Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins: Structure, Function, and Bioinformatics* 21:167-195.
45. Wolynes PG, Onuchic JN, & Thirumalai D (1995) Navigating the folding routes. *Science* 267:1619-1620.
46. Dobson CM (2003) Protein folding and misfolding. *Nature* 426:884-890.
47. Baldwin AJ*, et al.* (2011) Metastability of native proteins and the phenomenon of amyloid formation. *JACS* 133:14160-14163.
48. Gazit E (2002) The "correctly folded" state of proteins: Is it a metastable state? *Angewandte Chemie - International Edition* 41:257-259.
49. Tartaglia GG, Pechmann S, Dobson CM, & Vendruscolo M (2007) Life on the edge: a link between gene expression levels and aggregation rates of human proteins. *Trends in Biochemical Sciences* 32(5):204-205.
50. Dobson CM (2017) The amyloid phenomenon and its links with human disease. *Cold Spring Harbor Perspectives in Biology* 9.
51. Cremades N*, et al.* (2012) Direct observation of the interconversion of normal and toxic forms of α-synuclein. *Cell* 149:1048-1059.
52. Lee J, Culyba EK, Powers ET, & Kelly JW (2011) Amyloid-β Forms fibrils by nucleated conformational conversion of oligomers. *Nature Chemical Biology* 7:602-609.
53. Knowles TPJ*, et al.* (2009) An analytical solution to the kinetics of breakable filament assembly. *Science* 326:1533-1537.
54. Arosio P, Knowles TPJ, & Linse S (2015) On the lag phase in amyloid fibril formation. *Physical Chemistry Chemical Physics* 17:7606-7618.
55. Jarrett JT, Berger EP, & Lansbury PT (1993) The carboxy terminus of the ß Amyloid protein Is critical for the seeding of amyloid formation: implications for the pathogenesis of Alzheimer’s disease. *Biochemistry* 32:4693-4697.
56. Plakoutsi G*, et al.* (2005) Evidence for a mechanism of amyloid formation involving molecular reorganisation within native-like precursor aggregates. *Journal of Molecular Biology* 351:910-922.
57. Collinge J & Clarke AR (2007) A general model of prion strains and their pathogenicity. *Science* 318:930-936.
58. Roberts HL & Brown DR (2015) Seeking a mechanism for the toxicity of oligomeric alpha-synuclein. *Biomolecules* 5:282-305.
59. Benilova I, Karran E, & De Strooper B (2012) The toxic Aβ oligomer and Alzheimer's disease: an emperor in need of clothes. *Nature Neuroscience* 15:349-357.

60. Guerrero-Muñoz MJ, Gerson J, & Castillo-Carranza DL (2015) Tau oligomers: The toxic player at synapses in Alzheimer's disease. *Frontiers in Cellular Neuroscience* 9:464.
61. Lesné S*, et al.* (2006) A specific amyloid-beta protein assembly in the brain impairs memory. *Nature* 440:352-357.
62. Martins IC*, et al.* (2008) Lipids revert inert Abeta amyloid fibrils to neurotoxic protofibrils that affect learning in mice. *The EMBO Journal* 27:224-233.
63. Koffie RM*, et al.* (2009) Oligomeric amyloid beta associates with postsynaptic densities and correlates with excitatory synapse loss near senile plaques. *Proceedings of the National Academy of Sciences* 106:4012-4017.
64. Balchin D, Hayer-Hartl M, & Hartl FU (2016) In vivo aspects of protein folding and quality control. *Science* 353:aac4354.
65. Prusiner SB (2013) Biology and genetics of prions causing neurodegeneration. *Annual Review of Genetics* 47:601-623.
66. Brettschneider J, Del Tredici K, Lee VM-Y, & Trojanowski JQ (2015) Spreading of pathology in neurodegenerative diseases: a focus on human studies. *Nature Reviews Neuroscience* 16:109-120.
67. Walker LC & Jucker M (2015) Neurodegenerative Diseases: Expanding the Prion Concept. *Annual Review of Neuroscience* 38:87-103.
68. Aguzzi A & Lakkaraju AKK (2016) Cell biology of prions and prionoids: A status report. *Trends in Cell Biology* 26:40-51.
69. Douglas PM & Dillin A (2010) Protein homeostasis and aging in neurodegeneration. *The Journal of Cell Biology* 190(5):719-729.
70. Ben-Zvi A, Miller Ea, & Morimoto RI (2009) Collapse of proteostasis represents an early molecular event in Caenorhabditis elegans aging. *Proceedings of the National Academy of Sciences of the United States of America* 106:14914-14919.
71. Boehr DD, Mcelheny D, Dyson HJ, & Wright PE (2006) The Dynamic energy landscape of dihydrofolate reductase catalysis. *Science* 313:1638-1642.
72. Freire E (1999) The propagation of binding interactions to remote sites in proteins: Analysis of the binding of the monoclonal antibody. *Proceedings of the National Academy of Sciences* 96:10118-10122.
73. Guo W, Lampoudi S, & Shea J-E (2004) Temperature dependence of the free energy landscape of the src-SH3 protein domain. *Proteins* 55:395-406.
74. Hammarstrom P (2003) Prevention of transthyretin amyloid disease by changing protein misfolding energetics. *Science* 299:713-716.
75. Swain JF & Gierasch LM (2006) The changing landscape of protein allostery. *Current Opinion in Structural Biology* 16:102-108.
76. Tsai CC-JJ, Ma B, & Nussinov R (1999) Folding and binding cascades: Shifts in energy landscapes. *Proceedings of the National Academy of Sciences* 96:9970-9972.
77. Lindquist S (1986) The heat-shock response. *Annual Review of Biochemistry* 55:1151-1191.
78. Cummings CJ*, et al.* (2001) Over-expression of inducible HSP70 chaperone suppresses neuropathology and improves motor function in SCA1 mice. *Human Molecular Genetics* 10:1511-1518.
79. Klucken J, Shin Y, Masliah E, Hyman BT, & McLean PJ (2004) Hsp70 reduces α-synuclein aggregation and toxicity. *Journal of Biological Chemistry* 279:25497-25502.

80. Krobitsch S & Lindquist S (2000) Aggregation of huntingtin in yeast varies with the length of the polyglutamine expansion and the expression of chaperone proteins. *Proceedings of the National Academy of Sciences* 97:1589-1594.
81. Warrick JM*, et al.* (1999) Suppression of polyglutamine-mediated neurodegeneration in Drosophila by the molecular chaperone HSP70. *Nature Genetics* 23:425-428.
82. Glover JR & Lindquist S (1998) Hsp104, Hsp70, and Hsp40: A novel chaperone system that rescues previously aggregated proteins. *Cell* 94:73-82.
83. Lee DH, Sherman MY, & Goldberg aL (1996) Involvement of the molecular chaperone Ydj1 in the ubiquitin-dependent degradation of short-lived and abnormal proteins in Saccharomyces cerevisiae. *Molecular and Cellular Biology* 16:4773-4781.
84. Nijtmans LG*, et al.* (2000) Prohibitins act as a membrane-bound chaperone for the stabilization of mitochondrial proteins. *The EMBO Journal* 19:2444-2451.
85. Nishihara K, Kanemori M, & Yanagi H (2000) Overexpression of trigger factor prevents aggregation of recombinant proteins in escherichia coli. *Applied and Environmental Microbiology* 66:884-889.
86. Tam Stephen GRG, Spiess Christoph Spiess, Frydman Judith (2006) The chaperonin TRiC controls polyglutamine aggregation and toxicity through subunit-specific interactions. *Nature Cell Biology* 8:1155-1162.
87. Brehme M*, et al.* (2014) A chaperome subnetwork safeguards proteostasis in aging and neurodegenerative disease. *Cell Reports* 9:1135-1150.
88. Dikic I (2017) Proteasomal and autophagic degradation systems. *Annual Review of Biochemistry* 86:1-32.
89. Ihara Y, Morishima-Kawashima M, & Nixon R (2012) The ubiquitin–proteasome system and the autophagic–lysosomal system in Alzheimer disease. *Cold Spring Harbor perspectives in medicine* 2(8):a006361.
90. Nixon RA (2013) The role of autophagy in neurodegenerative disease. *Nature Medicine* 19(8):983-997.
91. Vilchez D, Saez I, & Dillin A (2014) The role of protein clearance mechanisms in organismal ageing and age-related diseases. *Nature Communications* 5:5659.
92. Lackie RE*, et al.* (2017) The Hsp70/Hsp90 chaperone machinery in neurodegenerative diseases. *Frontiers in Neuroscience* 11:1-23.
93. Kaushik S & Cuervo AM (2012) Chaperone-mediated autophagy: A unique way to enter the lysosome world. *Trends in Cell Biology* 22:407-417.
94. Dahlmann B (2007) Role of proteasomes in disease. *BMC Biochemistry* 8:S3.
95. Miller JA, Oldham MC, & Geschwind DH (2008) A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging. *Journal of Neuroscience* 28(6):1410-1420.
96. Ricciarelli R*, et al.* (2004) Microarray analysis in Alzheimer's disease and normal aging. *IUBMB life* 56(6):349-354.
97. Labbadia J & Morimoto RI (2015) Repression of the heat shock response is a programmed event at the onset of reproduction. *Molecular Cell* 59(4):639-650.
98. Leroy E*, et al.* (1998) The ubiquitin pathway in Parkinson's disease. *Nature* 395:451-452.

99. Shimura H, *et al.* (2001) Ubiquitination of a new form of a-synuclein by parkin from human brain: Implications for Parkinson's disease. *Science* 293:263-269.
100. McNaught KSP, Belizaire R, Isacson O, Jenner P, & Olanow CW (2003) Altered proteasomal function in sporadic Parkinson's disease. *Experimental Neurology* 179:38-46.
101. Rubinsztein DC (2006) The roles of intracellular protein-degradation pathways in neurodegeneration. *Nature* 443:780-786.
102. Bence NF, Sampat RM, & Kopito RR (2001) Impairment of the ubiquitin-proteasome system by protein aggregation. *Science* 292:1552-1555.
103. Keller JN, Hanni KB, & Markesbery WR (2000) Impaired proteasome function in Alzheimer's disease. *Journal of Neurochemistry* 75:436-439.
104. Kyle RA (2001) Amyloidosis: A convoluted story. *British Journal of Haematology* 114:529-538.
105. Parkinson J (1817) An essay on the shaking palsy. *Whittingham and Rowland for Sherwood, Neely, and Jones*.
106. Huntington G (1872) On Chorea. *Med Surg Reports* 26:109-112.
107. Alzheimer A (1907) Über eine eigenartige Erkrankung der Hirnrinde. *Allgemeine Zeitschrift für Psychiatrie und phychish-Gerichtliche Medizin, (Berlin)* 64:146-148.
108. Alzheimer A (1911) Uber eigenartige Krankheitsfalle des spateren Alters. *Zeitschrift fur die gesamte Neurologie und Psychiatrie* 4:356-385.
109. Stelzmann RA, Norman Schnitzlein H, & Reed Murtagh F (1995) An english translation of alzheimer's 1907 paper, "ber eine eigenartige erkankung der hirnrinde". *Clinical Anatomy* 8:429-431.
110. Virchow R (1971) Lecture XVII. Amyloid degeneration. Inflammation. in *Cellular pathology as based upon physiological and pathological histology*, pp 409-437.
111. Marcet W, Wilks S, Bristowe JS, Andrew J, & Dickinson WH (1871) Report of the Committee on Lardaceous Disease. *Transactions of the Pathological Society of London*, pp 1-12.
112. Sanger F & Thompson EOP (1953) The Amino-acid Sequence in the Glycyl Chain of Insulin: 1. The identification of lower peptides from partial hydrolysates. *Biochemical Journal* 53:353-366.
113. Sanger F & Thompson EOP (1953) The Amino-acid Sequence in the Glycyl Chain of Insulin 2. The investigation of peptides from enzymatic hydrolysates. *Biochemical Journal* 53:366-374.
114. Sanger F & Tuppy H (1951) The Amino-acid Sequence in the Phenylalanine Chain of Insulin 1. The identification of lower peptides from partial hydrolysates. *Biochemical Journal* 49:463-481.
115. Sanger F & Tuppy H (1951) The Amino-acid Sequence in the Phenylalanyl Chain of Insulin 2. The investigation of peptides from enzymatic hydrolysates. *Biochemical Journal* 49:481-490.
116. Newcombe DS & Cohen AS (1965) Solubility characteristics of isolated amyloid fibrils. *Biochimica et Biophysica Acta (BBA) - General Subjects* 104:480-486.
117. Glenner GG & Wong CW (1984) Alzheimer's disease: Initial report of the purification and characterization of a novel cerebrovascular amyloid protein. *Biochemical and Biophysical Research Communications* 120:885-890.

118. Polymeropoulos MH, *et al.* (1997) Mutation in the alpha-Synuclein Gene Identified in Families with Parkinson's Disease. *Science* 276:2045-2047.
119. MacDonald ME, *et al.* (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* 72:971-983.
120. Snyder SW, *et al.* (1994) Amyloid-beta aggregation: selective inhibition of aggregation in mixtures of amyloid with different chain lengths. *Biophysical Journal* 67:1216-1228.
121. Näslund J, *et al.* (1994) Relative abundance of Alzheimer A beta amyloid peptide variants in Alzheimer disease and normal aging. *Proceedings of the National Academy of Sciences* 91:8378-8382.
122. Roher aE, *et al.* (1993) β-Amyloid-(1-42) is a major component of cerebrovascular amyloid deposits: Implications for the pathology of Alzheimer disease. *Proceedings of the National Academy of Sciences* 90:10836-10840.
123. Suzuki N, *et al.* (1994) An increased percentage of long amyloid beta protein secreted by familial amyloid beta protein precursor (beta APP717) mutants. *Science* 264:1336-1340.
124. Kang J, *et al.* (1987) The precursor of Alzheimer's disease amyloid A4 protein resembles a cell-surface receptor. *Nature* 325:733-736.
125. Eanes ED & Glenner CG (1968) X-ray diffraction studies on amyloid filaments. *J Histochem Cytochem* 16:673-677.
126. Bernal JD & Crowfoot D (1934) X-ray photographs of crystalline pepsin. *Nature* 133:794-795.
127. Astbury WT & Lomax R (1934) X-ray photographs of crystalline pepsin. *Nature* 133:795.
128. Astbury WT, Dickinson S, & Bailey K (1935) The X-ray interpretation of denaturation and the structure of the seed globulins. *Biochemical Journal* 29:2351-2360.2351.
129. Iaki Guijarro J, Sunde M, Jones JA, Campbell ID, & Dobson CM (1998) Amyloid fibril formation by an SH3 domain. *Biochemistry* 95:4224-4228.
130. Fandrich M, Fletcher MA, & Dobson CM (2001) Amyloid fibrils from muscle myoglobin. *Nature* 410:165-166.
131. Booth DR, *et al.* (1997) Instability, unfolding and aggregation of human lysozyme variants underlying amyloid fibrillogenesis. *Nature* 385:787-793.
132. Bucciantini M, *et al.* (2002) Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* 416:507-511.
133. Gusella JF, *et al.* (1983) A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* 306:234-238.
134. Cook RH, Ward BE, & Austin JH (1979) Studies in aging of the brain: IV. Familial Alzheimer disease: Relation to transmissible dementia, aneuploidy, and microtubular defects. *Neurology* 29:1402-1412.
135. Nee LE, *et al.* (1983) A family with histologically confirmed Alzheimer's disease. *Archives of Neurology* 40:203-208.
136. Oliver C & Holland AJ (1986) Down's Syndrome and Alzheimer's disease: a review. in *Psychological Medicine*, pp 307-322.
137. St.George-Hyslop PH, *et al.* (1987) The genetic defect causing familial Alzheimer's disease maps on chromosome 21. *Science* 235:885-890.

138. Tanzi RE, *et al.* (1987) Amyloid beta protein gene, cDNA, mRNA distribution, and genetic linkage near the Alzheimer locus. *Science (New York, N.Y.)* 235:880-884.
139. Delabar J-m, *et al.* (1987) Beta amyloid gene duplication in Alzheimer's disease and karyotypically normal Down syndrome. *Science* 235:1390-1392.
140. St George-Hyslop PH, *et al.* (1987) Absence of duplication of chromosome 21 genes in familial and sporadic Alzheimer's disease. *Science* 238:664-666.
141. St George-Hyslop P, *et al.* (1992) Genetic evidence for a novel familial Alzheimer's disease locus on chromosome 14. *Nature Genetics* 2:330-334.
142. Goate AM, *et al.* (1989) Predisposing Locus for Alzheimer'S Disease on Chromosome 21. *The Lancet* 333:352-355.
143. Schellenberg GD, *et al.* (1988) Absence of linkage of chromosome 21q21 markers to familial Alzheimer's disease. *Science* 241:1507-1510.
144. St George-Hyslop PH, *et al.* (1990) Genetic linkage studies suggest that Alzheimer's disease is not a single homogeneous disorder. *Nature* 347:194-197.
145. van Duinen SG, *et al.* (1987) Hereditary cerebral hemorrhage with amyloidosis in patients of Dutch origin is related to Alzheimer disease. *Proceedings of the National Academy of Sciences* 84:5991-5994.
146. Mullan M, *et al.* (1993) Clinical comparison of Alzheimer's disease in pedigrees with the codon 717 Val ---> Ile mutation in the amyloid precursor protein gene. *Neurobiol Aging* 14:407-419.
147. Sherrington R, *et al.* (1995) Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature* 375:754-760.
148. Levy-Lahad E, *et al.* (1995) A familial Alzheimer's disease locus on chromosome I. *Science* 269:970-973.
149. Rogaev EI, *et al.* (1995) Familial Alzheimer's disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer's disease type 3 gene. *Nature* 376:775-778.
150. Weidemann A, *et al.* (1989) Identification, biogenesis, and localization of precursors of Alzheimer's disease A4 amyloid protein. *Cell* 57:115-126.
151. Haass C, *et al.* (1992) Amyloid beta-peptide is produced by cultured cells during normal metabolism. *Nature* 359:322-325.
152. Cai X-D, Golde TE, & Younkin SG (1993) Release of excess amyloid b protein from a mutant amyloid b protein precursor. *Science* 259:514-516.
153. Haass C & Selkoe D (1993) Cellular processing of beta-Amyloid precursor protein and
the genesis of amyloid p-peptide. *Cell* 75:1039-1042.
154. Bertram L, Lill CM, & Tanzi RE (2010) The genetics of Alzheimer disease: back to the future. *Neuron* 68(2):270-281.
155. Reiman EM, *et al.* (2007) GAB2 alleles modify Alzheimer's risk in APOE 4 carriers. *Neuron* 54:713-720.
156. Chapuis J, *et al.* (2008) Association study of the GAB2 gene with the risk of developing Alzheimer's disease. *Neurobiology of Disease* 30:103-106.
157. Harold D, *et al.* (2009) Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nature Genetics* 41:1088-1093.
158. Sleegers K, *et al.* (2009) Common variation in GRB-associated Binding Protein 2 (GAB2) and increased risk for Alzheimer dementia. *Human Mutation* 30:338-344.

159. Bertram L, *et al.* (2008) Genome-wide Association Analysis Reveals Putative Alzheimer's Disease Susceptibility Loci in Addition to APOE. *American Journal of Human Genetics* 83:623-632.
160. Bettens K, *et al.* (2010) Follow-up study of susceptibility loci for Alzheimer's disease and onset age identified by genome-wide association. *Journal of Alzheimer's Disease* 19:1169-1175.
161. Zhang C, *et al.* (2010) Loss of function of ATXN1 increases amyloid β-protein levels by potentiating β-secretase processing of β-amyloid precursor protein. *Journal of Biological Chemistry* 285:8515-8526.
162. Crocker PR, Paulson JC, & Varki A (2007) Siglecs and their roles in the immune system. *Nature Reviews Immunology* 7:255-266.
163. von Gunten S & Simon H-U (2006) Sialic acid binding immunoglobulin-like lectins may regulate innate immune responses by modulating the life span of granulocytes. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 20:601-605.
164. Lambert J-C, *et al.* (2009) Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nature Genetics* 41:1094-1099.
165. DeMattos RB, *et al.* (2004) ApoE and clusterin sooperatively Suppress Abeta Levels and Deposition: Evidence that ApoE Regulates Extracellular Abeta Metabolism In Vivo. *Neuron* 41:193-202.
166. DeMattos RB, *et al.* (2002) Clusterin promotes amyloid plaque formation and is critical for neuritic toxicity in a mouse model of Alzheimer's disease. *Proceedings of the National Academy of Sciences* 99:10843-10848.
167. Nuutinen T, Suuronen T, Kauppinen A, & Salminen A (2009) Clusterin: A forgotten player in Alzheimer's disease. *Brain Research Reviews* 61:89-104.
168. Khera R & Das N (2009) Complement receptor 1: Disease associations and therapeutic implications. *Molecular Immunology* 46:761-772.
169. Wyss-Coray T, *et al.* (2002) Prominent neurodegeneration and increased plaque formation in complement-inhibited Alzheimer's mice. in *Proceedings of the National Academy of Sciences*, pp 10837-10842.
170. Tebar F, Bohlander SK, & Sorkin A (1999) Clathrin assembly lymphoid myeloid leukemia (CALM) protein: localization in endocytic-coated pits, interactions with clathrin, and the impact of overexpression on clathrin-mediated traffic. *Molecular Biology of the Cell* 10:2687-2702.
171. Seshadri S, *et al.* (2010) Genome-wide analysis of genetic loci associated with Alzheimer disease. *JAMA* 303:1832-1840.
172. Liao L, *et al.* (2004) Proteomic characterization of postmortem amyloid plaques isolated by laser capture microdissection. *Journal of Biological Chemistry* 279:37061-37068.
173. Wang Q, *et al.* (2005) Proteomic analysis of neurofibrillary tangles in Alzheimer disease identifies GAPDH as a detergent-insoluble paired helical filament tau binding protein. *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 19:869-871.
174. Xia Q, *et al.* (2008) Proteomic identification of novel proteins associated with Lewy bodies. *Frontiers in bioscience* 13:3850-3856.
175. Chiti F, Stefani M, Taddei N, Ramponi G, & Dobson CM (2003) Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* 424:805-808.

176. DuBay KF, *et al.* (2004) Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. *Journal of Molecular Biology* 341:1317-1326.
177. Schubert U, *et al.* (2000) Rapid degradation of a large fraction of newly synthesized proteins by proteasomes. *Nature* 404:770-774.
178. Reis-Rodrigues P, *et al.* (2012) Proteomic analysis of age-dependent changes in protein solubility identifies genes that modulate lifespan. *Aging Cell* 11:120-127.
179. Chapman E, *et al.* (2006) Global aggregation of newly translated proteins in an Escherichia coli strain deficient of the chaperonin GroEL. *Proceedings of the National Academy of Sciences of the United States of America* 103:15800-15805.
180. Mogk A, *et al.* (1999) Identification of thermolabile Escherichia coli proteins: prevention and reversion of aggregation by DnaK and ClpB. *The EMBO Journal* 18:6934-6949.
181. Ciryam P, *et al.* (2016) A transcriptional signature of Alzheimer's disease is associated with a metastable subproteome at risk for aggregation. *Proceedings of the National Academy of Sciences of the United States of America* 113(17):4753-4758.
182. Kundra R, Ciryam P, Morimoto RI, Dobson CM, & Vendruscolo M (2017) Protein homeostasis of a metastable subproteome associated with Alzheimer ’ s disease. *PNAS* 114(28):E5703–E5711.
183. Hyman BT, *et al.* (2012) National institute on aging-Alzheimer's association guidelines for the neuropathologic assessment of Alzheimer's disease: A practical approach. *Alzheimer's and Dementia* 123:1-13.
184. Albert MS, *et al.* (2011) The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's and Dementia* 7:270-279.
185. Sperling RA, *et al.* (2011) Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's and Dementia* 7:280-292.
186. Bennett DA, Schneider JA, Wilson RS, Bienias JL, & Arnold SE (2004) Neurofibrillary tangles mediate the association of amyloid load with clinical Alzheimer disease and level of cognitive function. *Archives of Neurology* 61:378-384.
187. Bouros C, Hof PR, Giannakopoulos P, Michel JP, & Morrison JH (1994) Regional distribution of neurofibrillary tangles and senile plaques in the cerebral-cortex of elderly patients: A quantitative-evaluation of a one-year autopsy population from a geriatric hospital. *Cerebral Cortex* 4:138-150.
188. Braak H & Braak E (1997) Frequency of stages of Alzheimer-related lesions in different age categories. *Neurobiology of Aging* 18:351-357.
189. Caselli RJ, Walker D, Sue L, Sabbagh M, & Beach T (2010) Amyloid load in nondemented brains correlates with APOE e4. *Neuroscience Letters* 473:168-171.
190. Crystal H, *et al.* (1988) Clinico-pathologic studies in dementia: Nondemented subjects with pathologically confirmed alzheimer's disease. *Neurology* 38.

191. Dickson DW, *et al.* (1992) Identification of normal and pathological aging in prospectively studied nondemented elderly humans. *Neurobiology of Aging* 13:179-189.
192. Price JL, *et al.* (2009) Neuropathology of nondemented aging: Presumptive evidence for preclinical Alzheimer disease. *Neurobiology of Aging* 30:1026-1036.
193. Price JL & Morris JC (1999) Tangles and plaques in nondemented aging and "preclinical" Alzheimer's disease. *Annals of Neurology* 45:358-368.
194. Schmitt FA, *et al.* (2000) "Preclinical" AD revisited: neuropathology of cognitively normal older adults. *Neurology* 55:370-376.
195. McKhann GM, *et al.* (2011) The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's and Dementia* 7:263-269.
196. Hane FT, Lee BY, & Leonenko Z (2017) Recent progress in Alzheimer's disease research, Part 1: Pathology. *Journal of Alzheimer's Disease* 57:1-28.
197. Klunk WE, *et al.* (2004) Imaging brain amyloid in Alzheimer's disease with Pittsburgh compound-B. *Annals of Neurology* 55:306-319.
198. Maruyama M, *et al.* (2013) Imaging of tau pathology in a tauopathy mouse model and in alzheimer patients compared to normal controls. *Neuron* 79:1094-1108.
199. Villemagne VL, *et al.* (2013) Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: A prospective cohort study. *The Lancet Neurology* 12:357-367.
200. Fodero-Tavoletti MT, *et al.* (2011) 18F-THK523: A novel in vivo tau imaging ligand for Alzheimer's disease. *Brain* 134:1089-1100.
201. Gordon BA, *et al.* (2016) The relationship between cerebrospinal fluid markers of Alzheimer pathology and positron emission tomography tau imaging. *Brain* 139:2249-2260.
202. Harada R, *et al.* (2016) 18F-THK5351: A novel PET radiotracer for imaging neurofibrillary pathology in Alzheimer disease. *Journal of Nuclear Medicine : Official Publication, Society of Nuclear Medicine* 57:208-214.
203. Villemagne VL, Fodero-Tavoletti MT, Masters CL, & Rowe CC (2015) Tau imaging: early progress and future directions. *The Lancet Neurology* 14:114-124.
204. Cummings JL, Morstorf T, & Zhong K (2014) Alzheimer's disease drug-development pipeline: few candidates, frequent failures. *Alzheimer's research & therapy* 6:37.
205. Alavez S, Vantipalli MC, Zucker DJS, Klang IM, & Lithgow GJ (2011) Amyloid-binding compounds maintain protein homeostasis during ageing and extend lifespan. *Nature* 472:226-229.
206. Citron M (2010) Alzheimer's disease: strategies for disease modification. *Nature Reviews Drug discovery* 9:387-398.
207. Doody RS, *et al.* (2014) Phase 3 trials of solanezumab for mild-to-moderate Alzheimer's disease. *The New England Journal of Medicine* 370:311-321.
208. Habchi J, *et al.* (2016) An anti-cancer drug suppresses the primary nucleation reaction that initiates the formation of toxic Aβ aggregates associated with Alzheimer's disease. *Science Advances* e1501244:1-13.

209. Aprile FA, *et al.* (2017) Selective targeting of primary and secondary nucleation pathways in Aβ42 aggregation using a rational antibody scanning method. *Science Advances* 3.
210. Habchi J, *et al.* (2017) Systematic development of small molecules to inhibit specific microscopic steps of Aβ42 aggregation in Alzheimer's disease. *Proceedings of the National Academy of Sciences* 114:E200-E208.
211. Glass CK, Saijo K, Winner B, Marchetto MC, & Gage FH (2010) Mechanisms underlying inflammation in neurodegeneration. *Cell* 140(6):918-934.
212. Brehme M, *et al.* (2014) A chaperome subnetwork safeguards proteostasis in aging and neurodegenerative disease. *Cell Reports* 9(3):1135-1150.
213. Koga H, Kaushik S, & Cuervo AM (2011) Protein homeostasis and aging: The importance of exquisite quality control. *Ageing research reviews* 10(2):205-215.
214. Koplin A, *et al.* (2010) A dual function for chaperones SSB-RAC and the NAC nascent polypeptide-associated complex on ribosomes. *Journal of Cell Biology* 189(1):57-68.
215. Narayanaswamy R, *et al.* (2009) Widespread reorganization of metabolic enzymes into reversible assemblies upon nutrient starvation. *Proceedings of the National Academy of Sciences* 106(25):10147-10152.
216. Hofrichter J, Ross PD, & Eaton WA (1976) Supersaturation in sickle cell hemoglobin solutions. *Proceedings of the National Academy of Sciences* 73(9):3035-3039.
217. Ikenoue T, *et al.* (2014) Heat of supersaturation-limited amyloid burst directly monitored by isothermal titration calorimetry. *Proceedings of the National Academy of Sciences* 111(18):6654-6659.
218. Muta H, *et al.* (2014) Supersaturation-limited amyloid fibrillation of insulin revealed by ultrasonication. *Journal of Biological Chemistry* 289(26):18228-18238.
219. Ho L, *et al.* (2001) Altered expression of a-type but not b-type synapsin isoform in the brain of patients at high risk for Alzheimer's disease assessed by DNA microarray technique. *Neuroscience Letters* 298(3):191-194.
220. Blalock EM, *et al.* (2004) Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proceedings of the National Academy of Sciences* 101(7):2173-2178.
221. Umemura K, *et al.* (2006) Autotaxin expression is enhanced in frontal cortex of Alzheimer-type dementia patients. *Neuroscience Letters* 400(1):97-100.
222. Liang WS, *et al.* (2007) Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. *Physiological Genomics* 28(3):311-322.
223. Liang WS, *et al.* (2008) Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons. *Proceedings of the National Academy of Sciences* 105(11):4441-4446.
224. Webster JA, *et al.* (2009) Genetic control of human brain transcript expression in Alzheimer disease. *The American Journal of Human Genetics* 84(4):445-458.
225. Tan MG, *et al.* (2010) Genome wide profiling of altered gene expression in the neocortex of Alzheimer's disease. *Journal of Neuroscience Research* 88(6):1157-1169.

226. Simpson JE*, et al.* (2011) Microarray analysis of the astrocyte transcriptome in the aging brain: relationship to Alzheimer's pathology and APOE genotype. *Neurobiology of Aging* 32(10):1795-1807.
227. Cooper-Knock J*, et al.* (2012) Gene expression profiling in human neurodegenerative disease. *Nature Reviews Neurology* 8:518-530.
228. Durrenberger PF*, et al.* (2012) Selection of novel reference genes for use in the human central nervous system: a BrainNet Europe Study. *Acta Neuropathologica* 124(6):893-903.
229. Antonell A*, et al.* (2013) A preliminary study of the whole-genome expression profile of sporadic and monogenic early-onset Alzheimer's disease. *Neurobiology of Aging* 34(7):1772-1778.
230. Hokama M*, et al.* (2014) Altered expression of diabetes-related genes in Alzheimer's disease brains: the Hisayama study. *Cerebral Cortex* 24:2476-2488.
231. Miller JA, Woltjer RL, Goodenbour JM, Horvath S, & Geschwind DH (2013) Genes and pathways underlying regional and cell type changes in Alzheimer's disease. *Genome Medicine* 5(5):48.
232. Zhang B*, et al.* (2013) Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer’s disease. *Cell* 153(3):707-720.
233. Ding B*, et al.* (2014) Gene expression profiles of entorhinal cortex in Alzheimer’s disease. *American Journal of Alzheimer's disease and other Dementias*:1533317514523487.
234. Yankner BA, Lu T, & Loerch P (2008) The aging brain. *Annual Reviews Pathology* 3:41-66.
235. Ginsberg SD, Hemby SE, Lee VM, Eberwine JH, & Trojanowski JQ (2000) Expression profile of transcripts in Alzheimer's disease tangle-bearing CA1 neurons. *Annals of Neurology* 48(1):77-87.
236. Baker C, Belbin O, Kalsheker N, & Morgan K (2007) SERPINA3 (aka alpha-1-antichymotrypsin). *Frontiers in Bioscience* 12:2821-2835.
237. Consortium EP (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57-74.
238. Koldamova R*, et al.* (2014) Genome-wide approaches reveal EGR1-controlled regulatory networks associated with neurodegeneration. *Neurobiology of disease* 63:107-114.
239. Sheng B*, et al.* (2012) Impaired mitochondrial biogenesis contributes to mitochondrial dysfunction in Alzheimer’s disease. *Journal of neurochemistry* 120(3):419-429.
240. Lu T*, et al.* (2014) REST and stress resistance in ageing and Alzheimer's disease. *Nature* 507(7493):448-454.
241. Whitlock M (2005) Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology* 18(5):1368-1373.
242. Terni B, Boada J, Portero-Otin M, Pamplona R, & Ferrer I (2010) Mitochondrial ATP-Synthase in the entorhinal cortex Is a target of oxidative Stress at stages I/II of Alzheimer's disease pathology. *Brain Pathology* 20(1):222-233.
243. Sergeant N*, et al.* (2003) Association of ATP synthase α-chain with neurofibrillary degeneration in Alzheimer’s disease. *Neuroscience* 117(2):293-303.

244. Chandrasekaran K, Hatanpää K, Rapoport SI, & Brady DR (1997) Decreased expression of nuclear and mitochondrial DNA-encoded genes of oxidative phosphorylation in association neocortex in Alzheimer disease. *Molecular Brain Research* 44(1):99-104.
245. Manczak M, Park BS, Jung Y, & Reddy PH (2004) Differential expression of oxidative phosphorylation genes in patients with Alzheimer's disease. *Neuromolecular Medicine* 5(2):147-162.
246. Johnson MR*, et al.* (2015) Systems genetics identifies Sestrin 3 as a regulator of a proconvulsant gene network in human epileptic hippocampus. *Nature Communications* 6.
247. Benjamini Y & Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*:289-300.
248. Fisher RA (1925) *Statistical methods for research workers* (Genesis Publishing Pvt Ltd).
249. Pearson ES (1938) The probability integral transformation for testing goodness of fit and combining independent tests of significance. *Biometrika* 30(1-2):134-148.
250. Stouffer SA, Suchman EA, DeVinney LC, Star SA, & Williams Jr RM (1949) *The American soldier: adjustment during army life.* (Princeton University Press, Princeton, NJ.).
251. Lanz TA*, et al.* (2015) STEP levels zre unchanged in pre-frontal cortex and associative striatum in post-mortem human brain samples from subjects with schizophrenia, bipolar disorder and major depressive disorder. *PloS One* 10(3): e0121744.
252. Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*:65-70.
253. Ciryam P*, et al.* (2017) Spinal motor neuron protein supersaturation patterns are associated with inclusion body formation in ALS. *Proceedings of the National Academy of Sciences*:201613854.
254. Freer R*, et al.* (2016) A protein homeostasis signature in healthy brains recapitulates tissue vulnerability to Alzheimer's disease. *Science Advances* 2(8):e1600947.
255. Khurana V*, et al.* (2017) Genome-scale networks link neurodegenerative disease genes to α-synuclein through specific molecular pathways. *Cell Systems* 4(2):157-170. e114.
256. Eisen MB, Spellman PT, Brown PO, & Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* 95(25):14863-14868.
257. Stuart JM, Segal E, Koller D, & Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302(5643):249-255.
258. Heyer LJ, Kruglyak S, & Yooseph S (1999) Exploring expression data: identification and analysis of coexpressed genes. *Genome Research* 9(11):1106-1115.
259. Langfelder P & Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9(1):559.
260. Zhang B & Horvath S (2005) A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology* 4(1):1128.

261. Wu G, *et al.* (2011) Altered expression of autophagic genes in the peripheral leukocytes of patients with sporadic Parkinson's disease. *Brain Research* 1394:105-111.
262. Miller JA, Horvath S, & Geschwind DH (2010) Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. *Proceedings of the National Academy of Sciences* 107(28):12698-12703.
263. Lin MT & Beal MF (2006) Mitochondrial dysfunction and oxidative stress in neurodegenerative diseases. *Nature* 443(7113):787-795.
264. Voineagu I, *et al.* (2011) Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 474(7351):380-384.
265. Hawrylycz MJ, *et al.* (2012) An anatomically comprehensive atlas of the adult human brain transcriptome. *Nature* 489(7416):391-399.
266. Oldham MC, *et al.* (2008) Functional organization of the transcriptome in human brain. *Nature Neuroscience* 11(11):1271-1282.
267. Oldham MC, Horvath S, & Geschwind DH (2006) Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proceedings of the National Academy of Sciences* 103(47):17973-17978.
268. Kanehisa M, Goto S, Furumichi M, Tanabe M, & Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research* 38(S1):D355-D360.
269. Langfelder P, Luo R, Oldham MC, & Horvath S (2011) Is my network module preserved and reproducible? *PLoS Computational Biology* 7(1):e1001057.
270. Shannon P, *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* 13(11):2498-2504.
271. Treusch S, *et al.* (2011) Functional links between Aβ toxicity, endocytic trafficking, and Alzheimer's disease risk factors in yeast. *Science* 334(6060):1241-1245.
272. Cirrito JR, *et al.* (2008) Endocytosis is required for synaptic activity-dependent release of amyloid-β in vivo. *Neuron* 58(1):42-51.
273. Kaushik S & Cuervo AM (2015) Proteostasis and aging. *Nature Medicine* 21(12):1406-1415.
274. Matus S, Glimcher LH, & Hetz C (2011) Protein folding stress in neurodegenerative diseases: a glimpse into the ER. *Current Opinion in Cell Biology* 23(2):239-252.
275. Scheper W, *et al.* (2007) Rab6 is increased in Alzheimer's disease brain and correlates with endoplasmic reticulum stress. *Neuropathology and Applied Neurobiology* 33(5):523-532.
276. Cooper AA, *et al.* (2006) α-Synuclein blocks ER-Golgi traffic and Rab1 rescues neuron loss in Parkinson's models. *Science* 313(5785):324-328.
277. Molina MF, *et al.* (2011) Decreased expression of ATP6V1H in type 2 diabetes: a pilot report on the diabetes risk study in Mexican Americans. *Biochemical and Biophysical Research Communications* 412(4):728-731.
278. Geyer M, Fackler OT, & Peterlin BM (2002) Subunit H of the V-ATPase involved in endocytosis shows homology to β-adaptins. *Molecular Biology of the Cell* 13(6):2045-2056.
279. Colacurcio DJ & Nixon RA (2016) Disorders of lysosomal acidification—The emerging role of v-ATPase in aging and neurodegenerative disease. *Ageing Research Reviews* 32:75-88.

280. Namekawa M, *et al.* (2007) Mutations in the SPG3A gene encoding the GTPase atlastin interfere with vesicle trafficking in the ER/Golgi interface and Golgi morphogenesis. *Molecular and Cellular Neuroscience* 35(1):1-13.
281. Zhu P-P, *et al.* (2003) Cellular localization, oligomerization, and membrane association of the hereditary spastic paraplegia 3A (SPG3A) protein atlastin. *Journal of Biological Chemistry* 278(49):49063-49071.
282. Brett CL, Wei Y, Donowitz M, & Rao R (2002) Human Na(+)/H(+) exchanger isoform 6 is found in recycling endosomes of cells, not in mitochondria. *American Journal of Physiology, Cell Physiology* 282:C1031-C1041.
283. Brodsky FM, Chen C-y, Towler MC, & Wakeham DE (2001) Biological basket weaving: formation and function of clathrin-coated vesicles. *Annual Review of Cell and Developmental Biology* 17:517-568.
284. Rothman JE (1994) Mechanisms of intracellular protein transport. *Nature* 372(6501):55-63.
285. Sannerud R, *et al.* (2016) Restricted location of PSEN2/γ-secretase determines substrate specificity and generates an intracellular Aβ pool. *Cell* 166(1):193-208.
286. Lee H, *et al.* (2016) ENC1 modulates the aggregation and neurotoxicity of mutant huntingtin through p62 under ER stress. *Molecular Neurobiology* 53(10):6620-6634.
287. Han S, *et al.* (2008) Pam (Protein associated with Myc) functions as an E3 ubiquitin ligase and regulates TSC/mTOR signaling. *Cellular Signalling* 20(6):1084-1091.
288. Chen BB, Glasser JR, Coon TA, & Mallampalli RK (2012) F-box protein FBXL2 exerts human lung tumor suppressor-like activity by ubiquitin-mediated degradation of cyclin D3 resulting in cell cycle arrest. *Oncogene* 31:2566-2579.
289. Kuchay S, *et al.* (2013) FBXL2- and PTPL1-mediated degradation of p110-free p85β regulatory subunit controls the PI(3)K signalling cascade. *Nature Cell Biology* 15:472-480.
290. Arndt V, *et al.* (2010) Chaperone-assisted selective autophagy is essential for muscle maintenance. *Current Biology* 20(2):143-148.
291. Torres GE, Sweeney AL, Beaulieu J-M, Shashidharan P, & Caron MG (2004) Effect of torsinA on membrane proteins reveals a loss of function and a dominant-negative phenotype of the dystonia-associated ΔE-torsinA mutant. *Proceedings of the National Academy of Sciences* 101(44):15650-15655.
292. Hosokawa N, *et al.* (2008) Human XTP3-B forms an endoplasmic reticulum quality control scaffold with the HRD1-SEL1L ubiquitin ligase complex and BiP. *Journal of Biological Chemistry* 283(30):20914-20924.
293. Hinz FI & Geschwind DH (2016) Molecular genetics of neurodegenerative dementias. *Cold Spring Harbor Perspectives in Biology* doi: 10.1101/cshperspect.a023705.
294. Zhang Y, James M, Middleton FA, & Davis RL (2005) Transcriptional analysis of multiple brain regions in Parkinson's disease supports the involvement of specific protein processing, energy metabolism, and signaling pathways, and suggests novel disease mechanisms. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 137(1):5-16.
295. Danial NN & Korsmeyer SJ (2004) Cell death: critical control points. *Cell* 116:205-219.

296. Anandatheerthavarada HK, Biswas G, Robin M-A, & Avadhani NG (2003) Mitochondrial targeting and a novel transmembrane arrest of Alzheimer's amyloid precursor protein impairs mitochondrial function in neuronal cells. *The Journal of Cell Biology* 161(1):41-54.
297. Devi L, Prabhu BM, Galati DF, Avadhani NG, & Anandatheerthavarada HK (2006) Accumulation of amyloid precursor protein in the mitochondrial import channels of human Alzheimer's disease brain is sssociated with mitochondrial dysfunction. *The Journal of Neuroscience* 26:9057-9068.
298. Wrobel L*, et al.* (2015) Mistargeted mitochondrial proteins activate a proteostatic response in the cytosol. *Nature* 524(7566):485-488.
299. Heo J-M & Rutter J (2011) Ubiquitin-dependent mitochondrial protein degradation. *The International Journal of Biochemistry and Cell Biology* 43(10):1422-1426.
300. Itakura E*, et al.* (2016) Ubiquilins chaperone and triage mitochondrial membrane proteins for degradation. *Molecular Cell* 63(1):21-33.
301. Wang X & Chen XJ (2015) A cytosolic network suppressing mitochondria-mediated proteostatic stress and cell death. *Nature* 524(7566):481-484.
302. Edgar R, Domrachev M, & Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* 30(1):207-210.
303. Carter SL, Brechbühler CM, Griffin M, & Bond AT (2004) Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics* 20:2242-2250.
304. Barabási A-L (2009) Scale-free networks: a decade and beyond. *Science* 325(5939):412-413.
305. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, & Barabási A-L (2002) Hierarchical organization of modularity in metabolic networks. *Science* 297(5586):1551-1555.
306. Langfelder P & Horvath S (2007) Eigengene networks for studying the relationships between co-expression modules. *BMC Systems Biology* 1(1):54.
307. Kundra R, Ciryam P, Morimoto RI, Dobson CM, & Vendruscolo M (2016) Protein homeostasis of a metastable subproteome associated with Alzheimer's disease. *Proceedings of the National Academy of Sciences* 114.
308. Su AI*, et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences* 101(16):6062-6067.
309. Mattsson N, Schott JM, Hardy J, Turner MR, & Zetterberg H (2016) Selective vulnerability in neurodegeneration: insights from clinical variants of Alzheimer's disease. *J Neurol Neurosurg Psychiatry* 87(9):1000-1004.
310. Wang X & Michaelis EK (2010) Selective neuronal vulnerability to oxidative stress in the brain. *Front Aging Neurosci* 2:12.
311. Surmeier DJ, Obeso JA, & Halliday GM (2017) Selective neuronal vulnerability in Parkinson disease. *Nature Reviews Neuroscience* 18(2):101-113.
312. Saxena S & Caroni P (2011) Selective neuronal vulnerability in neurodegenerative diseases: from stressor thresholds to degeneration. *Neuron* 71(1):35-48.
313. Brockington A*, et al.* (2013) Unravelling the enigma of selective vulnerability in neurodegeneration: motor neurons resistant to degeneration in ALS show

distinct gene expression characteristics and decreased susceptibility to excitotoxicity. *Acta Neuropathol* 125(1):95-109.

314. H. Braak EB (1991) Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathologica* 82:239-259.
315. Darmanis S*, et al.* (2015) A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences* 112(23):7285-7290.
316. He Z*, et al.* (2017) Comprehensive transcriptome analysis of neocortical layers in humans, chimpanzees and macaques. *Nature Neuroscience* 20:6-9.
317. Jeremy A Miller CC, Peter Langfelder, Daniel H Geschwind, Sunil M Kurian, Daniel R Salomon, Steve Horvath (2011) Strategies for aggregating gene expression data: The collapseRows R function. *BMC Bioinformatics* 12(322):1471.
318. Rath A & Deber CM (2007) Membrane protein assembly patterns reflect selection for non-proliferative structures. *FEBS letters* 581(7):1335-1341.
319. Stroobants K*, et al.* (2017) Amyloid-like fibrils from an alpha-Helical transmembrane protein. *Biochemistry* 56(25):3225-3233.
320. Zhang B*, et al.* (2013) Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* 153:707-720.