Biomarkers for Parkinson's Disease and Alzheimer's Disease

Department of Psychiatry, Medical Sciences Division

Benjamine Liu, Christ Church

Supervisor: Professor Simon Lovestone

Degree: DPhil

Table of Contents **Pages**

List of Tables **Page**

List of Figures

Statement of Potential Conflicts

Somalogic and Proteome Sciences offered an academic discount for their proteomic assays. None of the work in this thesis was funded by Somalogic or Proteome Sciences and these two companies did not influence any of our analysis. However, they did give a discount from their commercial assay per sample costs, which they offer to universities and not-for-profits.

List of Abbreviations

Alzheimer's Disease (AD)
Alzheimer's Disease Neuroimaging Initiative (ADNI)
Area under curve (AUC)
Cerebrospinal fluid (CSF)
Familial Alzheimer's disease (FAD)
Least Absolute Shrinkage and Selection Operator (LASSO)
Levodopa equivalent daily dose (LEDD)
Magnetic resonance imaging (MRI)
Mass spectrometry (MS)
Mild Cognitive Impairment (MCI)
Montreal Cognitive Assessment (MOCA)
National Institute on Aging Alzheimer's Association (NIAAA)
Parkinson's Disease (PD)
Positron emission tomography (PET)
Random Forest (RF)
Relative fluorescence units (RFU)
Slow Off-rate Modified Aptamer (SOMAmers)
Systematic evolution of ligands by exponential enrichment (SELEX)
Unified Parkinson's disease rating scale (UPDRS3)

Abstract

Alzheimer's Disease (AD) and Parkinson's Disease (PD) represent the two most prevalent neurodegenerative diseases in the world, affecting 30 million and 5 million people, respectively. *(1, 2)* As populations age, the global burden of AD and PD will increase, resulting in significant societal and economic implications. By 2050, it is estimated that 1 in 85 people will have AD. *(1)* Biomarkers, surrogate indicators of physiological or pathophysiological states, can be used to guide the diagnosis of diseases, evaluate risk or prognosis, and track therapeutic interventions. In neurodegenerative diseases such as PD and AD, biomarkers can play a crucial role by facilitating earlier diagnosis and the screening of individuals into clinical trials.

Given its function to track various disease states, biomarkers will be fundamentally important as we develop and assess disease-modifying therapies and preventative strategies. *(3)* The advent of high-throughput sequencing technologies and advancements in omics— genomics, proteomics, and transcriptomics—has enabled the exploration of biomarkers for disease with unprecedented scale. Investigators can now move beyond candidate biomarker discovery approaches based upon *a priori* hypotheses. This presents an opportunity to utilize biomarkers to develop more sensitive and specific diagnostics and identify molecular targets that may have been overlooked through candidate approaches. Researchers can also use biomarkers to better define subgroups within PD or AD. These endophenotypes can help characterize the different etiologies that contribute to the development of the diseases, reveal distinct subpopulations within disease categories, and thereby uncover potential therapeutic targets.

The work presented here leverages these new advancements in high variable

capture approaches, specifically proteomics, to identify novel biomarkers and signatures of disease states for PD and AD. My thesis also demonstrates how multivariate analytical tools can be used to interrogate and validate distinguishing signatures found between disease and control states. In the first chapter, I will review the pathophysiology and clinical management of AD and PD. I will also review the AD and PD biomarkers and the various proteomic and computational approaches that enable biomarker discovery at scale. In the second chapter, I introduce a novel biomarker discovery approach that combines the advantages of hypothesis-driven and high variable capture approaches. The third chapter explores biomarkers for PD, using data from plasma and brain. In the fourth chapter, I identify proteins correlated across CSF and plasma and demonstrate how this data may be used in biomarker identification. In summary, I have identified novel biomarker signatures for the diagnosis and prognosis of PD and AD using high variable data capture methods and multivariate computational approaches.

Chapter 1: Review of AD and PD pathophysiology and introduction of the role of biomarkers in neurological disease research

AD is the most common cause of dementia, accounting for 80% of cases in the elderly. *(4)* Dementia is characterized by a cluster of symptoms including memory and cognitive impairment, disturbances in language, psychological and psychiatric changes, and impairments in activities of daily living.

Pathophysiology of AD

While many theories regarding the pathogenesis of AD have been introduced, such as the cholinergic hypothesis, tau hypothesis and inflammation hypothesis, the amyloid cascade hypothesis remains the most explanatory and established model. *(5–7)* The amyloid cascade hypothesis postulates that AD is caused by abnormal deposits of the amyloid plaques in the brain triggering a cascade that results in the formation of neurofibrillary tangles composed of hyperphosphorylated tau, neuritic injury, neuronal dysfunction, and ultimately cell death in AD.

Genetic studies provide evidence for the amyloid cascade hypothesis. An autosomal dominant mutation in the genes that encode amyloid precursor protein (APP*)*, presenilin 1 (*PSEN-1*) and presenilin 2 (*PSEN-2)* lead to familial Alzheimer's disease (FAD) *(8, 9)*. *PSEN-1* and *PSEN-2* encode parts of γ-secretase. Normally, APP is cleaved by alpha-secretase. However, in AD, APP is instead processed by β- and γ-secretase creating abnormal Aβ peptide production. This abnormal production results in an imbalance in the production and clearance of the Aβ peptides causing Aβ to aggregate into soluble oligomers. These oligomers then form fibrils in the insoluble beta-sheet

conformation which are deposited in senile plaques. *(11, 12)* The Aβ42 plaques that appear during later stages attract microglia, which results in the production of proinflammatory cytokines, including TNF-α, IL-1β, and IFN-γ. These cytokines drive nearby astrocytes to make increasing amounts of Aβ42 oligomers, stimulating more Aβ42 production. *(13)*

Aβ aggregates are responsible for the neuronal and vascular degeneration in AD brains. *(14)* Aβ42 monomers activate the neuroprotective signaling of insulin-like growth factor-1 receptor (IGF-1R), while Aβ42 oligomers activate neurodegeneration. These oligomers target many different membrane receptors, such as the scavenger receptor for advanced glycation end products (RAGE), NMDA-glutamate receptor, p75 neurotrophin receptor (p75NTR), α7 nicotinic ACh receptor (α7nAChR), ApoE receptors, formyl peptide receptor-like 1 (FPRL1/2), cellular prion protein (PrPc) acting as an Aβ oligomer receptor, the calcium-sensing receptor (CaSR), the Frizzled receptor, and the insulin receptor. *(15)*

The clearance of Aβ aggregates in the brain occurs by multiple pathways including uptake by astrocytes and microglia, passive removal into the cerebrospinal fluid, sequestration into the vascular compartment by the soluble form of the low-density lipoprotein receptor related protein 1 (LRP1), and proteolytic degradation by proteases such as insulin degrading enzyme (IDE). *(16)*

Management of AD

Currently, available treatment strategies for the management of AD include acetylcholinesterase inhibitors and NMDA receptor antagonists. *(17)* While there is no disease modifying therapy for AD currently available for patients, many groups are

exploring therapies targeting Aβ and tau. By reducing or clearing the levels of Aβ and tau in the brain, these novel therapies aim to limit and halt AD-related neurodegeneration. *(18)*

Pathophysiology of Parkinson's Disease

Parkinson's Disease (PD) affects more than 5 million people worldwide and is characterized by the progressive loss of dopaminergic neurons in the substantia nigra. *(3)*

The clinical presentation of PD is represented by four major components including motor symptoms, cognitive changes, behavioral/neuropsychiatric changes, and symptoms related to autonomic nervous system failure (see **Table 1.1** for non-motor symptoms). *(19)* The main motor features of PD are tremor, bradykinesia, rigidity and postural instability.

Table 1.1: Non-motor symptoms of PD

Category	Symptom
Olfactory	hyposmia / smell loss
Sleep	REM sleep behavior disorder insomnia daytime sleepiness restless leg syndrome
Gastrointestinal	dysphagia hypersalivation constipation swallowing difficulties
Genitourinary	urinary urgency nocturia increased urinary frequency impotence
Neuropsychiatric	anxiety depression visuospatial deficits psychosis apathy aggressiveness disinhibition hallucinations cognitive impairment (dementia) behavioral disorder dysexecutive syndrome
Visual	diplopia blurred vision reading difficulties dry eyes

Non-motor symptoms of PD are broken down by the category. Many of these non-motor symptoms appear before the onset of the classic motor symptoms.

The degeneration of dopamine-producing neurons in PD results in marked impairment of motor control. Mutations in genes that encode proteins in the brain, contribute to dopaminergic neuronal death in PD. Specifically, α-synuclein is altered in PD, resulting in self-aggregation. Aggregated and insoluble α-synuclein make up Lewy Bodies, which are cellular inclusions that are the hallmark of PD. *(20)* Pathways and processes designed to break down abnormal proteins such as the ubiquitin - proteasome system become impaired in PD, further contributing to Lewy Body accumulation. Other processes that may play a role in PD pathogenesis are mitochondrial dysfunction and abnormal oxidative stress through reactive oxygen species, leading to neuronal degeneration. *(21)* In PD, the brain's pontine locus ceruleus and substantia nigra pars compacta are regions affected by depigmentation, neuronal loss, and gliosis.

Management of PD

Currently, available treatment strategies for the management of PD include pharmacotherapy and non-pharmacological alternative approaches such as exercise, education, support groups, speech therapy, and nutritional support. *(22)*

Pharmacological approaches to PD are mainly focused on addressing dopamine deficits or imbalances. Seven categories of drugs are often employed to treat motor symptoms in PD patients including carbidopa/levodopa (Sinemet), dopamine agonists (both ergot and non-ergot types), monoamine oxidase-B (MAO-B) inhibitors, injectable dopamine agonists (apomorphine, or Apokyn), N-methyl-D-Aspartate receptor inhibitors, and anti-cholinergics. *(23)* However, like in AD, no definitive disease-modifying therapy to slow or stop the progression of PD exists.

Diagnosis of PD and AD

Diagnosis of AD

The gold standard for the diagnosis of AD is autopsy-based, requiring a post-mortem pathological evaluation of the brain for the presence and distribution of amyloid plaques and neurofibrillary tangles. In the clinical setting, AD diagnosis is largely based on physical and neurological examinations, medical history, and neuropsychological evaluation. The cornerstone of the clinical diagnosis is a set of consensus criteria established by the National Institute on Aging Alzheimer's Association (NIAAA) workgroup. The NIAAA clinical criteria for the diagnosis of AD dementia are summarized in **Table 1.2.** Laboratory and neuroimaging workup are often employed for investigational purposes.

Table 1.2: NIAAA clinical criteria for the diagnosis of AD dementia

Presence of dementia
Gradual onset of symptoms over months to years
History of progressive cognitive decline
Initial presentation may be amnestic (typical) or non-amnestic (atypical)
No evidence for another cause of cognitive impairment: cerebrovascular disease, other dementia syndromes, or neurological/medical disease

Diagnosis of PD

PD diagnosis relies on clinical criteria: the presence of resting tremor, rigidity, postural instability (gait disturbance) and bradykinesia. Diagnosing PD remains challenging as hallmark PD symptoms (e.g., resting tremor, rigidity etc.) can be observed in other neurodegenerative disorders, leading to a broad differential diagnosis. Despite years of research, the diagnosis and management of PD is handicapped by suboptimal methods for initial detection and prognosis. *(24)* As such, validated biomarkers with high sensitivity and specificity for the disease are critically needed but currently lacking. *(3)* The lack of robust biomarkers is also a major research roadblock, because clinical trial design often demands a target or biomarker to test neuroprotective therapies. *(25)* No single marker is able to predict PD progression with good reliability. Neurologic imaging is not used routinely in PD diagnosis as magnetic resonance imaging (MRI), ultrasonography, and positron emission tomography (PET) based methods lack evidence in diagnosing PD with high reliability.

In PD, there are non-motor, prodromal features (summarized in **Table 1.1**). Olfactory disorders frequently occur in PD, with over 50% of patients experiencing anosmia, 35% of patients suffering from severe hyposmia and 14% of patients experiencing moderate hyposmia. Subtle autonomic disturbances are also an early and frequent sign in PD. For example, almost all PD patients suffer from constipation. Moreover, many PD patients report sleep disruption, which usually begins early in the course of the disease.

The need for biomarkers in AD and PD

In studying neurodegenerative diseases such as PD and AD, biomarkers can play a crucial role by facilitating earlier diagnosis and allowing for appropriate individuals to be screened into clinical trials. The National Institutes of Health's Biomarkers Definitions Working Group defines a biomarker as "**a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.**" *(26)*

Biomarkers are often discussed in the context of two conceptual categories: **biomarkers of trait and state**. Biomarkers of trait are thought of as indicators of risk for disease or some other future clinical outcome while biomarkers of state are thought of as indicators of current disease presence or severity. As such, prognostic or risk-based biomarkers are often included as biomarkers of trait, while diagnostic biomarkers and biomarkers associated with disease severity will often be categorized as biomarkers of state.

Biomarkers allow for the *objective* measurement of a biological process, which allows researchers to better understand disease traits or states beyond more subjective, symptom-based characterizations. Biomarkers can be used to guide the diagnosis of disease, to evaluate risk or prognosis and to characterize disease signatures or pathways modulated in pathogenic processes. As such, biomarkers are fundamental for the development of new drugs and therapies.

For example, in PD, a biomarker that is associated with motor symptom severity would be invaluable with respect to assessing the efficacy of therapeutics meant to address motor symptoms. For preventative therapies, researchers must be able to

objectively characterize the risk of developing a given disease among patient groups with different risk factors, so biomarkers play a role in determining *a priori risk*. In AD, biomarkers can also help researchers better sort patients into clinical trials. For a therapy that targets Aβ, researchers may want to find patients that have a certain level of Aβ pathology in the brain. An easily assayable biomarker that is correlated with Aβ levels in the brain would be useful for helping to select the appropriate patients for such a clinical trial. For drugs that aim to reverse disease processes, a state biomarker that tracks with disease severity can be used as a primary endpoint to understand whether a therapy has any beneficial effect.

In AD, CSF measurements of T-tau, P-tau, and Aβ42 have been found to be differentially expressed across patients with AD, patients with Mild Cognitive Impairment (MCI) due to AD, *(27)* and controls. PET amyloid imaging assays that can successfully detect Aβ in brains are also available. *(28)*

In PD, CSF total α-synuclein levels have been studied by many groups, mostly in the context of a candidate diagnostic biomarker. *(29, 30)* While some groups have reported PD patients exhibiting significantly lower levels compared to normal controls *(31–33)* others have not observed this difference. *(34, 35)* These contradicting findings may be the result of differences in assay sensitivity or cohort characteristics. However, even amongst the studies reporting a significant difference in α-synuclein levels in the CSF of PD patients versus controls, total CSF α-synuclein levels in PD patients and normal controls demonstrate considerable overlap, *(36, 37)* which limits the practical use of α-synuclein as a diagnostic biomarker. Moreover, total CSF α-synuclein does not distinguish PD patients from individuals with other neurodegenerative diseases *(35, 38)*

nor do α-synuclein levels consistently correlate with PD severity. *(32, 35–37, 39)* CSF amyloid-beta, CSF tau, and serum urate have also garnered attention as potential PD biomarkers. However, their utility as biomarkers has also been inconclusive, with many contradicting studies. *(27)*

The utility of a given biomarker is not only influenced by the strength of its association between disease states and traits of interest, but also by how relatively easy it is to measure. While some CSF and imaging-based biomarkers have strong associations with disease states and traits of interest in AD and PD, the relative invasiveness and cost of a lumbar puncture for CSF collection or the injection of radioactive dyes for PET imaging procedures limits their clinical utility. Proteins that can be measured in an easily-accessible biofluid such as urine or blood are attractive as biomarkers because the ease of collection of these bodily fluids allows for widespread clinical adoption. The collection of protein biomarkers in blood or urine is also less expensive and less invasive than imaging biomarkers.

Few easily assayable biomarkers exist for AD and PD. For both diseases, diagnosis still relies on subjective, clinical assessments, with no laboratory-based confirmatory testing available. Because of the subjective nature of PD and AD diagnostic criteria and because clinical symptoms manifest most strongly in later stages of disease, diagnosis in early stages of disease remains a challenge. In PD clinical diagnosis is approximately 80% accurate in patients followed longitudinally with moderate symptoms. However, this accuracy may fall substantially, to approximately 65%, in earlier stages of disease. *(25)* For AD, diagnosis using clinical systems and neuropsychiatric testing often doesn't happen until 10 years after disease onset, when

clinical symptoms are most apparent and when most of the neuronal death and disease progress has already occurred. By identifying cases earlier, a molecular blood test for PD or AD may facilitate the development of preventative therapies and also allow for the testing of interventions in the therapeutic window before dramatic neuronal loss has occurred. In this context, an accessible biomarker for PD or AD could guide clinical care and accelerate therapeutic development.

The advent of high-throughput sequencing technologies, mass spectrometry, and aptamer-based proteomics has enabled the exploration of biomarkers for disease with unprecedented scale. Throughout my thesis, I will use a variety of these large-scale screening techniques and machine learning approaches to identify biomarkers for PD and AD.

I will review these techniques here, along with the computational approaches for biomarker discovery and validation.

Techniques for unbiased screening

Protein biomarkers are attractive from a laboratory testing perspective because they are relatively stable and abundant and can be measured using a variety of approaches.

There are three major types of unbiased screening approaches that allow researchers to measure many proteins at once: antibody-based methods, mass spectrometry-based methods, and aptamer-based methods. Methods that allow for the quantification of many proteins at once are often referred to as proteomic methods.

Antibody-based methods rely on the recognition of a specific protein epitope by an antibody. Antibody-based assays make up the vast majority of protein tests currently used in clinical laboratories. The main advantages of antibody-based methods for biomarker development are that they are relatively easy to measure and translate into clinical settings. However, the main disadvantage of antibody-based methods is the constraint on the number of proteins that one can simultaneously measure from a single sample. These limitations exist because of interference effects from multiplexing a large number of protein assays since these assays fundamentally rely on the ability of a peptide/protein to trigger an immune response. As such, antibody, bead-based multiplex immunoassays currently allow for the quantification of only approximately 100 of more than 200,000 estimated proteins in the proteome.

Mass spectrometry (MS) is an analytical technique that ionizes a given sample so that chemical species within the sample are broken up and separated according to mass and charge. The resulting ionization spectra encodes information that can be used to identify the chemical nature of the sample analyzed, which can then be mapped to the peptide fragments that make up a protein.

The advantage of MS approaches for biomarker discovery is that they are, by definition, unbiased because the technique is applied to and data is collected from the whole sample. However, MS-based techniques have many disadvantages and limitations. First, MS-based approaches require very specific sample preparation protocols that may be difficult to standardize across multiple trial runs. Second, the conversion of raw mass spectral data into interpretable readouts is often very challenging and time consuming. Finally, the results of MS-based approaches are often dominated by the most abundant

protein/peptide in the sample of origin. This limits the ability to identify biomarkers that may be present in lower abundance, which are likely to be overshadowed by high abundance proteins. Despite the fact that MS-based approaches have been available for decades, they have seen very little translation into the clinical realm.

Aptamer-based methods: Aptamers are oligonucleotide or peptide molecules that bind a specific molecular target. Aptamer-based methods often rely on enriching a set of oligonucleotides that bind a certain molecular target and using nucleotide quantification techniques to convert nucleotide signal to protein signal. Recently, a novel platform for proteomics based on an aptamer-based technology has been developed by a company called Somalogic Inc. (Boulder, CO). This platform is based on protein-capture SOMAmers (Slow Off-rate Modified Aptamer), which are chemically modified oligonucleotides with specific affinity to their protein targets, developed by systematic evolution of ligands by exponential enrichment (SELEX). *(40)*

In brief, plasma samples are incubated with the reagent mixes containing SOMAmers to the different proteins in a sample to allow for equilibrium binding of fluorophore-tagged SOMAmers to their protein targets. Next, a series of partitioning and wash steps are used to capture only the SOMAmers that are bound to their cognate proteins. Finally, the protein-bound SOMAmer oligonucleotides are released from the protein complex, captured by complementarity, and quantified using DNA hybridization arrays. After normalization and calibration of signal, readings for each SOMAmer – reported in relative fluorescence units (RFU) – reflect the relative amount of each cognate protein present in the original sample.

The main advantage of this aptamer-based method is that there is no theoretical limit to the number of proteins that can be simultaneously quantified in one sample. Unlike antibody-based methods, the aptamer-based approach is highly scalable and does not suffer from interference effects due to immune responses. In fact, the assay that we have used can simultaneously quantify more than 1000 proteins at once and newer versions of the assay can quantify up to 4000 proteins at once. This assay also allows for a dynamic range of quantification, as it may detect both scarcely abundant and highly abundant proteins. Finally, because aptamers bind to protein epitopes, they are similar to antibody-based methods in that they can be easily translated to clinical settings.

However, aptamers do also have limitations. It is often difficult to ascertain which epitope of the protein the aptamer is actually binding to. This means that aptamers may not be able to truly distinguish different isoforms of proteins, unless SELEX was competitively run to distinguish the two isoforms. So, regardless of high target specificity, aptamers that recognize particular protein targets will often also bind to proteins with a similar structure. In most cases, aptamer generation also requires the availability of purified target proteins, which is not always possible.

In addition to proteomic-based methods, metabolomic and lipidomic-based approaches have been used to identify biomarkers for neurodegenerative disease in the blood and urine.

Metabolomics refers to the quantification of a large number of low-molecular weight species that are involved in metabolic processes in living organisms, often using MS or nuclear magnetic resonance (NMR) techniques. *(41)* The advantage of metabolomics is that it has the potential to detect dynamic changes in metabolites that

result from a physiologic or pathophysiologic process directly. The major limitation of metabolomics is similar to that of the MS-based methods: it is often difficult to convert raw data into interpretable, robust readouts and map them to biological features. Lipidomics represents a subfield of metabolomics, which refers to the detection and quantitation of lipids.

Feature Selection and Computational Approaches for Biomarker Validation

The ability to quantify thousands of proteins at once allows researchers to move beyond univariate approaches for biomarker discovery (i.e., identifying single proteins that are differentially expressed between disease patients and controls). In contrast to univariate approaches, multivariate approaches allow us to identify groups of proteins or protein patterns that distinguish the disease groups of interest.

In order to define protein signatures with clinical and research significance, machine learning is often used to convert protein signals to probability scores that can be used to distinguish and characterize phenotypes of interest using classification algorithms (i.e., cases versus controls or converters versus non-converters). Multivariate and machine learning approaches have the potential to be more diagnostically accurate than classification methods that rely on single protein cutoffs because they can find latent variables in higher dimensional spaces that separate the disease groups of interest with more precision.

The process of identifying a multivariate biomarker signal using machine learning is often broken up into 3 steps. First, a feature selection step is employed to identify and rank proteins that most differentiate the disease groups of interest. Next, a machine-

learning classifier is trained on a dedicated training set. This training step essentially finds the functional parameters that make up the classifier that most distinguishes the categories of interest (i.e., cases versus controls) and minimalizes the classification errors across different crossvalidation partitions. Finally, this classifier is tested on a validation set to assess classification performance.

While there are a variety of different machine-learning algorithms for biomarker discovery, I have chosen to use the method of random decision forests. Random forests (RF) is an ensemble learning method for machine-learning based classification. *(42)* RF operates by constructing and training a number of decision trees to ultimately predict an outcome.

The advantage of a RF is that it is less of a black box approach compared to methods like support vector machines. *(43)* Random decision forests correct for decision trees' habit of overfitting to their training set. As part of random forest model construction, predictors lead to a dissimilarity measure between the observations. A random forest dissimilarity is advantageous because this method handles mixed variable types well (e.g., categorical and continuous data types), is robust to outliers, and is invariant to monotonic transformations of the predictors, compared to methods like logistic regression or support vector machines. There are few drawbacks to the random forest, however some of the downsides include overfitting if training and crossvalidation is not robustly constructed.

However, as we increase the number of proteins we identify for biomarker discovery, we also increase the probability for false discovery. Beyond multiple testing correction approaches, I use a variety of biological validation and novel computational

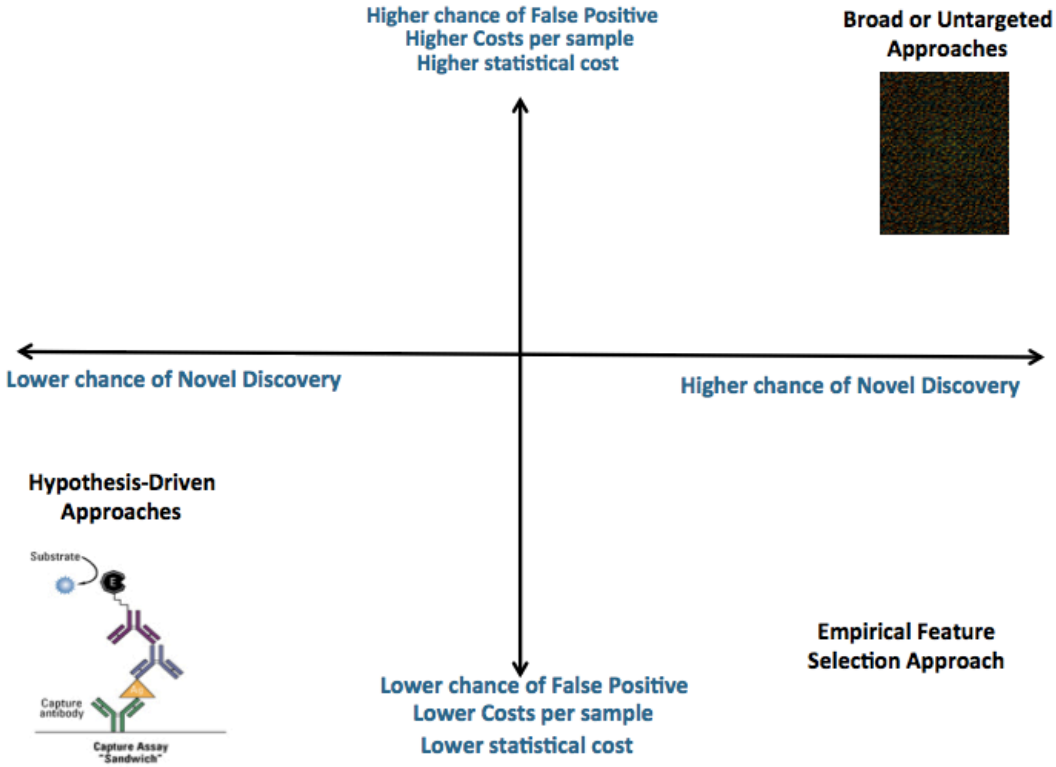
approaches to minimize the chance of false discovery. In the next chapter, I introduce the empirical feature selection approach which combines hypothesis driven approaches with large-scale variable capture approaches to identify biomarkers.

Chapter 2: Empirical Feature Selection in Biomarker Discovery

Chapter Summary

Historically, biomarker discovery has been either broad and untargeted—utilising various "omics" technologies to identify markers associated with an outcome of interest, or narrow and targeted—typically analysing a single or small number of analytes based on a hypothesis generated from previous studies. For example, many studies have used targeted approaches to examine blood or CSF levels of Aβ or tau for Alzheimer's since these species are known to be part of the pathophysiology of the disease. *(44)* This targeted approach is contrasted with untargeted approaches where thousands or tens of thousands of chemical species are quantified in an unbiased fashion to determine candidate biomarkers that differentiate the phenotypic groups of interest. *(45)* While broad and untargeted approaches can lead to the identification of new candidate biomarkers, they are also more prone to false discovery than hypothesis driven approaches (**Fig. 2.1**).

Fig. 2.1: Advantages and Disadvantages of Biomarker Discovery Approaches



Hypothesis driven biomarker discovery approaches usually rely on interrogating biochemicals involved in known pathological processes (bottom left). As a result, they are less prone to false discovery compared to untargeted approaches. However, while untargeted approaches (top right) increase the probability of discovering novel markers, they are also more prone to false discovery. The empirical feature selection approach combines the advantages of hypothesis driven research with broad or untargeted approaches.

We wanted to develop an experimental and computational method for biomarker discovery that combines the advantages of hypothesis driven research together with the power of very large variable capture by utilising experimental data from an *in vitro* model for targeted biomarker discovery. In order to do so, we employ an empirically derived feature selection approach, from an in vitro experimental model, to identify blood-based biomarkers for AD.

Introduction

AD is characterized by neuropathological features including amyloid plaques and neurofibrillary tangles composed of hyperphosphorylated tau. The amyloid cascade hypothesis remains the most explanatory model of pathogenesis. *(5, 6)* Studies have suggested that the antagonist of canonical Wnt signalling, Dickkopf-1 (*DKK1*) is involved in the cascade leading to Alzheimer's-related toxicity. *(46–49)* Specifically, Aβ has been shown to induce AD pathogenesis by increasing *DKK1* expression while silencing of *DKK1* and genes induced by both Aβ and DKK1 blocks Aβ neurotoxicity. *(49)*

Many studies have explored the relationship between Wnt signalling and Aβ mediated neurotoxicity in an effort to reveal the underlying mechanisms and pathways that contribute to the development of AD. *(50–52)* A previous study conducted by my supervisor's lab identified a pathway in which Aβ induces a clusterin/p53/DKK1/Wnt-PCP-JNK which drives the upregulation of several genes that mediate the development of AD-like pathological processes. *(49)* Moreover, Aβ and DKK1 induce an overlapping

transcript signature which is present in mouse models of AD, *in vitro* in cell models, and in the brain of AD patients. *(49)*

While studies have explored the effects of *DKK1* overexpression on gene expression, few studies have investigated how overexpression of *DKK1* may impact protein expression at the proteomic scale. This is of critical importance both for the reason that it is proteins that effect cellular change but also because it is proteins that are the principle targets for therapeutic development and biomarker discovery.

This study aims to explore first, whether overexpression of *DKK1* induces a proteomic signature just as it does a transcript signature; second, whether overexpression of *DKK1* modulates biological pathways relevant in AD pathogenesis that have also been previously identified in transcript signatures; and third, whether a protein signature empirically derived from in vitro experimental models might be detectable in people with disease. Specifically, in relation to the latter objective, we wanted to determine whether a protein signature derived from experimental *in vitro* models might function as an AD biomarker set in vivo. While previous studies have used unbiased approaches to nominate protein signatures differentially expressed in AD versus controls, *(53–61)* this empirical, experimentally generated approach might provide further corroborating evidence of DKK1 involvement in AD pathogenesis, as well as help to identify biomarkers that have diagnostic potential and are relevant to known biological mechanisms in AD.

Materials and Methods

Overexpression of DKK1 in HEK293A cells

All wet lab work was conducted by my colleague Elena Ribe and described here. In order to establish a DKK1 induced protein signature a HIS-tagged DKK1 cDNA was synthesized (GENEWIZ, UK) and cloned into the mammalian expression construct pcDNA3.1+ (Invitrogen, UK) and validated by sequencing. HEK293A cells (an adherent strain of HEK293) were cultured in DMEM + 10% FCS in 12 well plates until 80% confluent and transfected with the DKK1 construct or the empty vector control using FuGene 6 according to the manufacturer instructions (Promega, UK). Next day the FCS content of the media was adjusted to 2% and the cells maintained for a further 6 h. Media was then removed, and total cell lysates collected in M-PER (ThermoFisher, UK) for proteomic array studies using the SOMAScan platform (n=5 per condition).

Subjects

Protein measurements were obtained from 320 AD patients and 209 elderly unaffected controls recruited from two studies, including the Maudsley and King's Healthcare Partners Dementia Case Register (DCR), which incorporates the Alzheimer's Research UK (ARUK) cohort and the European Union funded AddNeuroMed (ANM) biomarker study *(62)*. Subjects were required to fast for 2 hours before sample collection. Blood samples were drawn by venepuncture, collected into EDTA glass tubes, and centrifuged within 2 hours of collection. Plasma supernatant was collected, divided into aliquots, and frozen at 80–C until further use. The local ethical committees at each research site approved all protocols and procedures. We obtained informed consent for all

subjects according to the Declaration of Helsinki (1991). Demographical details of the cohort are summarized in **Table 2.1**.

Table 2.1: AD Cohort Characteristics

Condition	Number Female/Male	P-value	Age ± SD	P-value	Presence of APOE ε4 yes/no	P-value
Alzheimer's Disease	320 222/98	3.46e-05[a]	79.01 ± 7.04	2.55e-13[b]	181/139	1.36e-11[a]
Healthy Controls	209 107/102		74.40 ± 6.30		56/153	

The demographic and clinical data describing the AD and healthy control cohorts are reported. Age represents the age (in years) at plasma collection. The proportion of males compared to females, age, and proportion of participants who had at least one APOE ε4 allele differed significantly between the Alzheimer's and healthy control groups. Data was adjusted for these covariates prior to downstream analysis.

[a]Fisher Exact Test, two-sided

[b]Mann-Whitney Test, two sided

Protein Quantification

1128 proteins were quantified using the SOMAScan assay from SomaLogic Inc. (Boulder, CO). The identities of all proteins are listed in **Table 2.2**. The SOMAScan assay is an aptamer-based technology that allows for the quantification of protein levels in biofluid. This technique is made possible by protein-capture SOMAmers (Slow Off-rate Modified Aptamer), which are chemically modified oligonucleotides with specific affinity to their protein targets, developed by SELEX as previously described. *(21, 63)*

The SOMAScan assay is described in detail in technical white papers at www.somalogic.com. In brief, the sample is diluted to three different concentrations, and all three dilutions are assayed, so that the least concentrated dilution can be used to detect the most abundant proteins, and the most concentrated dilutions used to detect the least abundant proteins, within the dynamic range of each protein's individual SOMAmer assay. Diluted samples are then incubated with the SOMAmer mixes (including SOMAmers to 1128 different proteins) to allow for equilibrium binding of fluorophore-tagged SOMAmers to their corresponding proteins. A series of automated partitioning steps are subsequently used to retain only the SOMAmers that are bound to proteins, which are then released from the protein complex prior to quantification. Finally, the previously protein-bound SOMAmer oligonucleotides – now a quantitative proxy for the proteins to which they were bound – are captured by complementarity and quantitated using DNA hybridization arrays. The hybridization arrays are then normalized and calibrated using data from a reference set of pooled plasma samples run on each batch to adjust for batch-to-batch variation. Thus, the normalized, calibrated signal for each SOMAmer reflects the relative amount of each cognate protein present in the original

sample, with findings reported in relative fluorescence units (RFU). All data was first log10 transformed prior to analysis.

Table 2.2: 1128 proteins quantified by SOMAScan

The identities of the 1128 proteins quantified by SOMAScan are shown here along with the corresponding Entrez Gene names.

Protein Name	Entrez Gene name
41	EPB41
14-3-3	YWHAB YWHAE YWHAG YWHAH WHAQ YWHAZ SFN
17-beta-HSD 1	HSD17B1
3HIDH	HIBADH
4-1BB	TNFRSF9
4-1BB ligand	TNFSF9
40S ribosomal protein SA	RPSA
4EBP2	EIF4EBP2
6-Phosphogluconate dehydrogenase	PGD
6Ckine	CCL21
a1-Antichymotrypsin	SERPINA3
a1-Antitrypsin	SERPINA1
a2-Antiplasmin	SERPINF2
a2-HS-Glycoprotein	AHSG
a2-Macroglobulin	A2M
ABL1	ABL1
ABL2	ABL2
ACE2	ACE2
ACTH	POMC
Activated Protein C	PROC
Activin A	INHBA
Activin AB	INHBA INHBB
Activin RIB	ACVR1B
ADAM 9	ADAM9
ADAM12	ADAM12
ADAMTS-4	ADAMTS4
ADAMTS-5	ADAMTS5
Adiponectin	ADIPOQ
Afamin	AFM
Aflatoxin B1 aldehyde reductase	AKR7A2
Aggrecan	ACAN
AGR2	AGR2

AIF1	AIF1
AIP	AIP
AK1A1	AKR1A1
Albumin	ALB
ALCAM	ALCAM
ALK-1	ACVRL1
Alkaline phosphatase bone	ALPL
ALT	GPT
AMHR2	AMHR2
Aminoacylase-1	ACY1
AMNLS	AMN
AMPK a1b1g1	PRKAA1 PRKAB1 PRKAG1
AMPK a2b2g1	PRKAA2 PRKAB2 PRKAG1
AMPM2	METAP2
amyloid precursor protein	APP
AN32B	ANP32B
Angiogenin	ANG
Angiopoietin-1	ANGPT1
Angiopoietin-2	ANGPT2
Angiopoietin-4	ANGPT4
Angiostatin	PLG
Angiotensinogen	AGT
ANGL3	ANGPTL3
ANGL4	ANGPTL4
annexin I	ANXA1
annexin II	ANXA2
annexin VI	ANXA6
Antithrombin III	SERPINC1
Apo A-I	APOA1
Apo B	APOB
Apo D	APOD
Apo E	APOE
Apo E2	APOE
Apo E3	APOE
Apo E4	APOE
AREG	AREG
ARGI1	ARG1
ARI3A	ARID3A
ARMEL	CDNF
ARP19	ARPP19
ARSB	ARSB

ART	AGRP
Artemin	ARTN
ARTS1	ERAP1
Arylsulfatase A	ARSA
ASAH2	ASAH2
ASAHL	NAAA
ASGR1	ASGR1
ASM3A	SMPDL3A
ATP synthase beta chain	ATP5B
ATS1	ADAMTS1
ATS13	ADAMTS13
ATS15	ADAMTS15
AURKB	AURKB
Aurora kinase A	AURKA
Azurocidin	AZU1
b-ECGF	FGF1
b-Endorphin	POMC
b-NGF	NGF
b2-Microglobulin	B2M
B7	CD80
B7-2	CD86
B7-H1	CD274
B7-H2	ICOSLG
BAFF	TNFSF13B
BAFF Receptor	TNFRSF13C
BARK1	ADRBK1
BASI	BSG
BCAM	BCAM
BCAR3	BCAR3
Bcl-2	BCL2
BCL2-like 1 protein	BCL2L1
BCMA	TNFRSF17
BDNF	BDNF
bFGF	FGF2
bFGF-R	FGFR1
BFL1	BCL2A1
BGH3	TGFBI
BGN	BGN
BLC	CXCL13
BMP RII	BMPR2
BMP-1	BMP1

BMP-14	GDF5
BMP-6	BMP6
BMP-7	BMP7
BMP10	BMP10
BMPER	BMPER
BMPR1A	BMPR1A
BMX	BMX
BNP-32	NPPB
BOC	BOC
Bone proteoglycan II	DCN
BPI	BPI
BRF-1	BRF1
BSP	IBSP
BSSP4	PRSS22
BST1	BST1
BTK	BTK
C1-Esterase Inhibitor	SERPING1
C1q	C1QA C1QB C1QC
C1QBP	C1QBP
C1r	C1R
C1s	C1S
C2	C2
C3	C3
C34 gp41 HIV Fragment	Human-virus
C3a	C3
C3adesArg	C3
C3b	C3
C3d	C3
C4	C4A C4B
C4b	C4A C4B
C5	C5
C5a	C5
C5b 6 Complex	C5 C6
C6	C6
C7	C7
C8	C8A C8B C8G
C9	C9
CAD15	CDH15
Cadherin E	CDH1
Cadherin-12	CDH12
Cadherin-2	CDH2

Cadherin-5	CDH5
Cadherin-6	CDH6
Calcineurin	PPP3CA PPP3R1
Calcineurin B a	PPP3R1
calgranulin B	S100A9
Calpain I	CAPN1 CAPNS1
Calpastatin	CAST
calreticulin	CALR
CAMK1	CAMK1
CAMK1D	CAMK1D
CAMK2A	CAMK2A
CAMK2B	CAMK2B
CAMK2D	CAMK2D
CaMKK alpha	CAMKK1
CAPG	CAPG
Carbonic anhydrase 6	CA6
Carbonic anhydrase 9	CA9
Carbonic anhydrase I	CA1
carbonic anhydrase II	CA2
Carbonic anhydrase III	CA3
Carbonic Anhydrase IV	CA4
Carbonic anhydrase VII	CA7
Carbonic Anhydrase X	CA10
Carbonic anhydrase XIII	CA13
Cardiotrophin-1	CTF1
CASA	CSN1S1
Caspase-10	CASP10
Caspase-2	CASP2
Caspase-3	CASP3
Catalase	CAT
CATC	CTSC
CATE	CTSE
Cathepsin A	CTSA
Cathepsin B	CTSB
Cathepsin D	CTSD
Cathepsin G	CTSG
Cathepsin H	CTSH
Cathepsin S	CTSS
Cathepsin V	CTSL2
CATZ	CTSZ
CBG	SERPINA6

CBPE	CPE
CBX5	CBX5
CCL28	CCL28
CD109	CD109
CD22	CD22
CD226	CD226
CD23	FCER2
CD27	CD27
CD30	TNFRSF8
CD30 Ligand	TNFSF8
CD36 ANTIGEN	CD36
CD39	ENTPD1
CD40 ligand soluble	CD40LG
CD48	CD48
CD5L	CD5L
CD70	CD70
CD97	CD97
CDC37	CDC37
CDK1/cyclin B	CDC2 CCNB1
CDK2/cyclin A	CDK2 CCNA2
CDK5/p35	CDK5 CDK5R1
CDK8/cyclin C	CDK8 CCNC
CDON	CDON
CFC1	CFC1
cGMP-stimulated PDE	PDE2A
Chitotriosidase-1	CHIT1
CHK1	CHEK1
Chk2	CHEK2
CHL1	CHL1
CHST2	CHST2
CHST6	CHST6
Chymase	CMA1
cIAP-2	BIRC3
Ck-b-8-1	CCL23
CK-BB	CKB
CK-MB	CKB CKM
CK-MM	CKM
CK2-A1:B	CSNK2A1 CSNK2B
CK2-A2:B	CSNK2A2 CSNK2B
CKAP2	CKAP2
CLC1B	CLEC1B

CLC4K	CD207
CLC7A	CLEC7A
CLF-1/CLC Complex	CRLF1 CLCF1
CLM6	CD300C
Clusterin	CLU
CNDP1	CNDP1
CNTF	CNTF
CNTFR alpha	CNTFR
CNTN2	CNTN2
CO8A1	COL8A1
Coactosin-like protein	COTL1
Coagulation Factor IX	F9
Coagulation Factor IXab	F9
Coagulation Factor V	F5
Coagulation Factor VII	F7
Coagulation Factor X	F10
Coagulation Factor Xa	F10
Coagulation Factor XI	F11
Cofilin-1	CFL1
COLEC12	COLEC12
Collectin Kidney 1	COLEC11
COMMD7	COMMD7
complement factor H-related 5	CFHR5
CONA1	COL23A1
contactin-1	CNTN1
Contactin-4	CNTN4
Contactin-5	CNTN5
COX-2	PTGS2
CPNE1	CPNE1
CRDL1	CHRDL1
Cripto	TDGF1
CRIS3	CRISP3
CRK	CRK
CRP	CRP
CRTAM	CRTAM
CSF-1	CSF1
CSK	CSK
CSK21	CSNK2A1
CTACK	CCL27
CTAP-III	PPBP
CTGF	CTGF

CTLA-4	CTLA4
CXCL16 soluble	CXCL16
Cyclin B1	CCNB1
Cyclophilin A	PPIA
Cyclophilin F	PPIF
Cystatin C	CST3
Cystatin M	CST6
Cystatin-S	CST4
CYTD	CST5
CYTF	CST7
Cytidylate kinase	CMPK1
CYTN	CST1
Cytochrome c	CYCS
Cytochrome P450 3A4	CYP3A4
CYTT	CST2
D-dimer	FGA FGB FGG
DAF	CD55
DAN	NBL1
DAPK2	DAPK2
DBNL	DBNL
DC-SIGN	CD209
DC-SIGNR	CLEC4M
DcR3	TNFRSF6B
DEAD-box protein 19B	DDX19B
DERM	DPT
Desmoglein-1	DSG1
Desmoglein-2	DSG2
DHH	DHH
discoidin domain receptor 1	DDR1
Discoidin domain receptor 2	DDR2
Dkk-4	DKK4
DKK1	DKK1
DKK3	DKK3
DLC8	DYNLL1
DLL1	DLL1
DLL4	DLL4
DLRB1	DYNLRB1
DMP1	DMP1
DnaJ homolog	DNAJC19
dopa decarboxylase	DDC
DPP2	DPP7

DR3	TNFRSF25
DR6	TNFRSF21
DRAK2	STK17B
DRG-1	VTA1
DRR1	FAM107A
DSC3	DSC3
Dtk	TYRO3
DUS3	DUSP3
Dynactin subunit 2	DCTN2
DYRK3	DYRK3
ECM1	ECM1
EDA	EDA
EDAR	EDAR
EF-1-beta	EEF1B2
EG-VEGF	PROK1
eIF-5	EIF5
eIF-5A-1	EIF5A
Elafin	PI3
Elastase	ELANE
EMAP-2	AIMP1
EMR2	EMR2
ENA-78	CXCL5
Endocan	ESM1
Endoglin	ENG
Endostatin	COL18A1
Endothelin-converting enzyme 1	ECE1
ENPP7	ENPP7
Enterokinase	PRSS7
ENTP3	ENTPD3
ENTP5	ENTPD5
Eotaxin	CCL11
Eotaxin-2	CCL24
EP15R	EPS15L1
EphA1	EPHA1
EPHA3	EPHA3
EphA5	EPHA5
EPHAA	EPHA10
EPHB2	EPHB2
EphB4	EPHB4
EphB6	EPHB6
Ephrin-A4	EFNA4

Ephrin-A5	EFNA5
Ephrin-B3	EFNB3
EPI	EREG
Epithelial cell kinase	EPHA2
Epo	EPO
EPO-R	EPOR
ER	ESR1
ERAB	HSD17B10
ERBB1	EGFR
ERBB2	ERBB2
ERBB3	ERBB3
ERBB4	ERBB4
ERK-1	MAPK3
ERP29	ERP29
ESAM	ESAM
Esterase D	ESD
ETHE1	ETHE1
FABP	FABP3
FABPE	FABP5
Factor B	CFB
Factor D	CFD
Factor H	CFH
Factor I	CFI
FAK1	PTK2
FAM107B	FAM107B
Fas ligand soluble	FASLG
FCAR	FCAR
FCG2A/B	FCGR2A FCGR2B
FCG3B	FCGR3B
FCGR1	FCGR1A
FCN1	FCN1
FCN2	FCN2
FCRL3	FCRL3
FER	FER
Ferritin	FTH1 FTL
FETUB	FETUB
FGF-10	FGF10
FGF-12	FGF12
FGF-16	FGF16
FGF-17	FGF17
FGF-18	FGF18

FGF-19	FGF19
FGF-20	FGF20
FGF-4	FGF4
FGF-5	FGF5
FGF-6	FGF6
FGF-8A	FGF8
FGF-8B	FGF8
FGF23	FGF23
FGF7	FGF7
FGF9	FGF9
FGFR-2	FGFR2
FGFR-3	FGFR3
FGFR4	FGFR4
FGR	FGR
Fibrinogen	FGA FGB FGG
Fibrinogen g-chain dimer	FGG
Fibronectin	FN1
Ficolin-3	FCN3
FLRT1	FLRT1
Flt-3	FLT3
Flt3 ligand	FLT3LG
FN1.3	FN1
FN1.4	FN1
Fractalkine/CX3CL-1	CX3CL1
FSH	CGA FSHB
FST	FST
FSTL3	FSTL3
Fucosyltransferase 3	FUT3
FUT5	FUT5
FYN	FYN
G-CSF	CSF3
G-CSF-R	CSF3R
GA733-1 protein	TACSTD2
Galectin-2	LGALS2
Galectin-3	LGALS3
Galectin-4	LGALS4
Galectin-8	LGALS8
GAPDH liver	GAPDH
GAS1	GAS1
GCKR	GCKR
GCP-2	CXCL6

GDF-11	GDF11
GDF-9	GDF9
GDF2	GDF2
Gelsolin	GSN
GFAP	GFAP
GFRa-1	GFRA1
GFRa-2	GFRA2
GFRa-3	GFRA3
GHC2	SLC25A18
GIB	PLA2G1B
GIIE	PLA2G2E
GITR	TNFRSF18
Glucagon	GCG
Glucocorticoid receptor	NR3C1
Glutamate carboxypeptidase	CNDP2
Glutathione S-transferase Pi	GSTP1
Glypican 3	GPC3
GM-CSF	CSF2
GNS	GNS
GOT1	GOT1
GP114	GPR114
gp130 soluble	IL6ST
GP1BA	GP1BA
GPC2	GPC2
GPC5	GPC5
gpIIbIIIa	ITGA2B ITGB3
GPNMB	GPNMB
GPVI	GP6
Granulysin	GNLY
granzyme A	GZMA
Granzyme B	GZMB
Granzyme H	GZMH
GRB2-related adapter protein 2	GRAP2
GREM1	GREM1
GRN	GRN
Gro-a	CXCL1
Gro-b/g	CXCL3 CXCL2
Growth hormone receptor	GHR
GSK-3 alpha/beta	GSK3A GSK3B
GSTA3	GSTA3
GV	PLA2G5

GX	PLA2G10
H6ST1	HS6ST1
HAI-1	SPINT1
Haptoglobin Mixed Type	HP
Hat1	HAT1
HB-EGF	HBEGF
HCC-1	CCL14
HCC-4	CCL16
HCG	CGA CGB
HCK	HCK
HDAC8	HDAC8
HDGR2	HDGFRP2
Hemoglobin	HBA1 HBB
Hemopexin	HPX
Heparin cofactor II	SERPIND1
HGF	HGF
HGFA	HGFAC
HINT1	HINT1
HIPK3	HIPK3
Histone H1.2	HIST1H1C
Histone H2A.z	H2AFZ
HIV-2 Rev	Human-virus
HMG-1	HMGB1
HMGR	HMGCR
hnRNP A/B	HNRNPAB
hnRNP A2/B1	HNRNPA2B1
hnRNP K	HNRNPK
HNRPQ	SYNCRIP
HO-2	HMOX2
HPG-	HPGD
HPLN1	HAPLN1
HPV E7 Type 16	Human-virus
HPV E7 Type18	Human-virus
HRG	HRG
HSP 40	DNAJB1
HSP 60	HSPD1
HSP 70	HSPA1A
HSP 90a/b	HSP90AA1 HSP90AB1
HSP70 protein 8	HSPA8
HTRA2	HTRA2
HVEM	TNFRSF14

I-309	CCL1
I-TAC	CXCL11
iC3b	C3
ICOS	ICOS
IDE	IDE
IDS	IDS
IDUA	IDUA
IF4A3	EIF4A3
IF4G2	EIF4G2
IFN-aA	IFNA2
IFN-g	IFNG
IFN-g R1	IFNGR1
IFN-lambda 1	IL29
IFN-lambda 2	IL28A
IgD	IGHD IGK@ IGL@
IgE	IGHE IGK@ IGL@
IGF-I	IGF1
IGF-I sR	IGF1R
IGF-II receptor	IGF2R
IGFBP-1	IGFBP1
IGFBP-2	IGFBP2
IGFBP-3	IGFBP3
IGFBP-4	IGFBP4
IGFBP-5	IGFBP5
IGFBP-6	IGFBP6
IGFBP-7	IGFBP7
IgG	IGHG1 IGHG2 IGHG3 IGHG4 IGK@ IGL@
IgM	IGHM IGJ IGK@ IGL@
IL-1 R AcP	IL1RAP
IL-1 R4	IL1RL1
IL-1 sR9	IL1RAPL2
IL-1 sRI	IL1R1
IL-10	IL10
IL-10 Rb	IL10RB
IL-11	IL11
IL-11 RA	IL11RA
IL-12	IL12A IL12B
IL-12 Rb1	IL12RB1
IL-12 RB2	IL12RB2
IL-13	IL13

IL-13 Ra1	IL13RA1
IL-15 Ra	IL15RA
IL-16	IL16
IL-17	IL17A
IL-17 RC	IL17RC
IL-17 RD	IL17RD
IL-17 sR	IL17RA
IL-17B	IL17B
IL-17B R	IL17RB
IL-17D	IL17D
IL-17E	IL25
IL-17F	IL17F
IL-18 BPa	IL18BP
IL-18 Ra	IL18R1
IL-18 Rb	IL18RAP
IL-19	IL19
IL-1a	IL1A
IL-1b	IL1B
IL-1F7	IL1F7
IL-1Rrp2	IL1RL2
IL-2	IL2
IL-2 sRa	IL2RA
IL-2 sRg	IL2RG
IL-20	IL20
IL-20 Ra	IL20RA
IL-22	IL22
IL-22BP	IL22RA2
IL-23	IL12B IL23A
IL-23 R	IL23R
IL-27	IL27 EBI3
IL-3	IL3
IL-3 Ra	IL3RA
IL-34	IL34
IL-4	IL4
IL-4 sR	IL4R
IL-5	IL5
IL-5 Ra	IL5RA
IL-6	IL6
IL-6 sRa	IL6R
IL-7	IL7
IL-7 Ra	IL7R

IL-8	IL8
IL22RA1	IL22RA1
IL24	IL24
ILT-2	LILRB1
ILT-4	LILRB2
IMB1	KPNB1
IMDH1	IMPDH1
IMDH2	IMPDH2
ING1	ING1
Insulin	INS
Integrin a1b1	ITGA1 ITGB1
Integrin aVb5	ITGAV ITGB5
IP-10	CXCL10
IR	INSR
ITI heavy chain H4	ITIH4
JAG1	JAG1
JAG2	JAG2
JAK2	JAK2
JAM-B	JAM2
JAM-C	JAM3
JAML1	AMICA1
JNK2	MAPK9
K-ras	KRAS
Kallikrein 11	KLK11
Kallikrein 12	KLK12
Kallikrein 13	KLK13
Kallikrein 14	KLK14
Kallikrein 4	KLK4
Kallikrein 5	KLK5
Kallikrein 6	KLK6
Kallikrein 7	KLK7
Kallikrein 8	KLK8
Kallistatin	SERPINA4
Karyopherin-a2	KPNA2
Keratin 18	KRT18
KI2L4	KIR2DL4
KI3L2	KIR3DL2
KI3S1	KIR3DS1
KIF23	KIF23
Kininogen HMW	KNG1
KIRR3	KIRREL3

KLRF1	KLRF1
KPCI	PRKCI
KPCT	PRKCQ
KREM2	KREMEN2
Ku70	XRCC6
KYNU	KYNU
Lactoferrin	LTF
LAG-1	CCL4L1
LAG-3	LAG3
Lamin-B1	LMNB1
Laminin	LAMA1 LAMB1 LAMC1
Layilin	LAYN
LBP	LBP
LCK	LCK
LCMT1	LCMT1
LD78-beta	CCL3L1
LDH-H 1	LDHB
LEAP-1	HAMP
Leptin	LEP
LG3BP	LGALS3BP
LGMN	LGMN
LIF sR	LIFR
LIGHT	TNFSF14
LIMP II	SCARB2
LIN7B	LIN7B
Lipocalin 2	LCN2
Livin B	BIRC7
LKHA4	LTA4H
LRIG3	LRIG3
LRP8	LRP8
LRRT1	LRRTM1
LRRT3	LRRTM3
LSAMP	LSAMP
Luteinizing hormone	CGA LHB
LY86	LY86
LY9	LY9
Lymphotactin	XCL1
Lymphotoxin a1/b2	LTA LTB
Lymphotoxin a2/b1	LTA LTB
Lymphotoxin b R	LTBR
LYN	LYN

LYNB	LYN
Lysozyme	LYZ
LYVE1	LYVE1
M-CSF R	CSF1R
M2-PK	PKM2
Macrophage mannose receptor	MRC1
Macrophage scavenger receptor	MSR1
Mammaglobin 2	SCGB2A1
MAPK14	MAPK14
MAPK2	MAPKAPK2
MAPK5	MAPKAPK5
MAPKAPK3	MAPKAPK3
Marapsin	PRSS27
MASP3	MASP1
MATK	MATK
MATN2	MATN2
MATN3	MATN3
MBD4	MBD4
MBL	MBL2
MCP-1	CCL2
MCP-2	CCL8
MCP-3	CCL7
MCP-4	CCL13
MDC	CCL22
MDHC	MDH1
MDM2	MDM2
MED-1	MED1
MEK1	MAP2K1
MEPE	MEPE
Mesothelin	MSLN
Met	MET
METAP1	METAP1
MFGM	MFGE8
MFRP	MFRP
MIA	MIA
MICA	MICA
MICB	MICB
Midkine	MDK
MIF	MIF
MIP-1a	CCL3
MIP-3a	CCL20

MIP-3b	CCL19
MIP-5	CCL15
MIS	AMH
MK01	MAPK1
MK08	MAPK8
MK11	MAPK11
MK12	MAPK12
MK13	MAPK13
MMEL2	MMEL1
MMP-1	MMP1
MMP-10	MMP10
MMP-12	MMP12
MMP-13	MMP13
MMP-14	MMP14
MMP-16	MMP16
MMP-17	MMP17
MMP-2	MMP2
MMP-3	MMP3
MMP-7	MMP7
MMP-8	MMP8
MMP-9	MMP9
Mn SOD	SOD2
MO2R1	CD200R1
Moesin	MSN
MOZ	KAT6A
MP2K2	MAP2K2
MP2K4	MAP2K4
MPIF-1	CCL23
MRC2	MRC2
MRCKB	CDC42BPB
MSP	MST1
MSP R	MST1R
Myeloperoxidase	MPO
Myoglobin	MB
Myokinase human	AK1
NA	NA
NACA	NACA
NADPH-P450 Oxidoreductase	POR
NAGK	NAGK
NANOG	NANOG
NAP-2	PPBP

NCAM-120	NCAM1
NCAM-L1	L1CAM
NCC27	CLIC1
NCK1	NCK1
NDP kinase B	NME2
Nectin-like protein 1	CADM3
Nectin-like protein 2	CADM1
NET4	NTN4
NEUREGULIN-1	NRG1
Neurotrophin-3	NTF3
Neurotrophin-5	NTF4
NG36	EHMT2
NID2	NID2
Nidogen	NID1
NKG2D	KLRK1
NKp30	NCR3
NKp44	NCR2
NKp46	NCR1
NLGNX	NLGN4X
NMT1	NMT1
Noggin	NOG
Nogo Receptor	RTN4R
NOTC2	NOTCH2
Notch 1	NOTCH1
Notch-3	NOTCH3
NovH	NOV
NPS-PLA2	PLA2G2A
Nr-CAM	NRCAM
NR1D1	NR1D1
NRP1	NRP1
NRX1B	NRXN1
NRX3B	NRXN3
NSF1C	NSFL1C
Nucleoside diphosphate kinase A	NME1
NUDC3	NUDCD3
NXPH1	NXPH1
OBCAM	OPCML
OCAD1	OCIAD1
Olfactomedin-4	OLFM4
OLR1	OLR1

OMD	OMD
ON	SPARC
OPG	TNFRSF11B
OSM	OSM
OX2G	CD200
OX40 Ligand	TNFSF4
P-Cadherin	CDH3
P-Selectin	SELP
p27Kip1	CDKN1B
PA2G4	PA2G4
PACAP-27	ADCYAP1
PACAP-38	ADCYAP1
PAFAH	PLA2G7
PAFAH beta subunit	PAFAH1B2
PAI-1	SERPINE1
PAK3	PAK3
PAK6	PAK6
PAK7	PAK7
PAPP-A	PAPPA
paraoxonase 1	PON1
PARC	CCL18
PARK7	PARK7
PBEF	NAMPT
PCI	SERPINA5
PCNA	PCNA
PCSK7	PCSK7
PD-L2	PDCD1LG2
PDE11	PDE11A
PDE3A	PDE3A
PDE4D	PDE4D
PDE5A	PDE5A
PDE7A	PDE7A
PDE9A	PDE9A
PDGF Rb	PDGFRB
PDGF-AA	PDGFA
PDGF-BB	PDGFB
PDGF-CC	PDGFC
PDK1	PDK1
PDPK1	PDPK1
PECAM-1	PECAM1
Periostin	POSTN

PERL	LPO
Peroxiredoxin-1	PRDX1
Peroxiredoxin-5	PRDX5
Peroxiredoxin-6	PRDX6
Persephin	PSPN
PESC	PES1
PF-4	PF4
PFD5	PFDN5
PGCB	BCAN
PGP9.5	UCHL1
PGRP-S	PGLYRP1
PH	PPY
PHI	GPI
phosphoglycerate kinase 1	PGK1
Phosphoglycerate mutase 1	PGAM1
PIGR	PIGR
PIK3CA/PIK3R1	PIK3CA PIK3R1
PIM1	PIM1
PK3CG	PIK3CG
PKB a/b/g	None
PKC-A	PRKCA
PKC-B-II	PRKCB
PKC-D	PRKCD
PKC-G	PRKCG
PKC-Z	PRKCZ
Plasmin	PLG
Plasminogen	PLG
PLCG1	PLCG1
PlGF	PGF
PLK-1	PLK1
PLPP	PDXP
PLXC1	PLXNC1
PPAC	ACP1
PPase	PPA1
PPIB	PPIB
PPID	PPID
PPIE	PPIE
Prekallikrein	KLKB1
PRKACA	PRKACA
PRL	PRL
Prolactin Receptor	PRLR

Properdin	CFP
prostatic binding protein	PEBP1
Protease nexin I	SERPINE2
Protein C	PROC
Protein disulfide isomerase A3	PDIA3
Protein disulfide-isomerase	P4HB
Protein S	PROS1
Proteinase-3	PRTN3
Prothrombin	F2
PSA	KLK3
PSA-ACT	KLK3 SERPINA3
PSA1	PSMA1
PSA2	PSMA2
PSA6	PSMA6
PSD7	PSMD7
PSMA	FOLH1
PSME1	PSME1
PSME3	PSME3
pTEN	PTEN
PTH	PTH
PTHrP	PTHLH
PTK6	PTK6
PTN	PTN
PTP-1B	PTPN1
PTP-1C	PTPN6
PUR8	ADSL
PYY	PYY
Rab GDP dissociation inhibitor beta	GDI2
RAC1	RAC1
RAD51	RAD51
RAN	RAN
RANK	TNFRSF11A
RANTES	CCL5
RAP	LRPAP1
RASA1	RASA1
Rb	RB1
RBM39	RBM39
RBP	RBP4
RELT	RELT
Renin	REN

resistin	RETN
RET	RET
RGM-C	HFE2
RGMA	RGMA
RGMB	RGMB
ROBO2	ROBO2
ROBO3	ROBO3
ROR1	ROR1
RPS6KA3	RPS6KA3
RS3	RPS3
RS7	RPS7
RSK-like protein kinase	RPS6KA5
RSPO2	RSPO2
RTN4	RTN4
RUXF	SNRPF
SAA	SAA1
SAP	APCS
SARP-2	SFRP1
SBDS	SBDS
sCD163	CD163
sCD4	CD4
SCF sR	KIT
SCGF-alpha	CLEC11A
SCGF-beta	CLEC11A
SDF-1	CXCL12
sE-Selectin	SELE
SE6L2	SEZ6L2
Secretin	SCT
Semaphorin 3A	SEMA3A
Semaphorin 3E	SEMA3E
Semaphorin-6A	SEMA6A
SEPR	FAP
SET	SET
sFRP-3	FRZB
SGTA	SGTA
SH21A	SH2D1A
SHBG	SHBG
SHC1	SHC1
SHP-2	PTPN11
Sialoadhesin	SIGLEC1
sICAM-1	ICAM1

sICAM-2	ICAM2
sICAM-3	ICAM3
sICAM-5	ICAM5
SIG14	SIGLEC14
Siglec-3	CD33
Siglec-6	SIGLEC6
Siglec-7	SIGLEC7
Siglec-9	SIGLEC9
SIRT2	SIRT2
SKP1	SKP1
sL-Selectin	SELL
SLAF5	CD84
SLAF6	SLAMF6
SLAF7	SLAMF7
sLeptin R	LEPR
SLIK5	SLITRK5
SLPI	SLPI
SMAC	DIABLO
SNAA	NAPA
SOD	SOD1
Soggy-1	DKKL1
Somatostatin-28	SST
Sonic Hedgehog	SHH
SORC2	SORCS2
Sorting nexin 4	SNX4
SP-D	SFTPD
SPARCL1	SPARCL1
Sphingosine kinase 1	SPHK1
SPHK2	SPHK2
SPINT2	SPINT2
Spondin-1	SPON1
SPTA2	SPTAN1
sRAGE	AGER
sRANKL	TNFSF11
SRCN1	SRC
SREC-I	SCARF1
SREC-II	SCARF2
SSRP1	SSRP1
ST4S6	CHST15
STAB2	STAB2
Stanniocalcin-1	STC1

sTie-1	TIE1
sTie-2	TEK
STK16	STK16
STRATIFIN	SFN
Stress-induced-phosphoprotein 1	STIP1
STX1a	STX1A
suPAR	PLAUR
Survivin	BIRC5
TACI	TNFRSF13B
TAFI	CPB2
TAJ	TNFRSF19
TAK1-TAB1	MAP3K7 TAB1
TARC	CCL17
tau	MAPT
TBK1	TBK1
TBP	TBP
TCCR	IL27RA
TCPTP	PTPN2
TCTP	TPT1
TEC	TEC
TECK	CCL25
Tenascin	TNC
Testican-1	SPOCK1
Testican-2	SPOCK2
TF	F3
TFF3	TFF3
TFPI	TFPI
TGF-b R II	TGFBR2
TGF-b R III	TGFBR3
TGF-b1	TGFB1
TGF-b2	TGFB2
TGF-b3	TGFB3
TGM3	TGM3
Thrombin	F2
Thrombopoietin Receptor	MPL
Thrombospondin-1	THBS1
Thymidine kinase	TK1
Thyroglobulin	TG
Thyroxine-Binding Globulin	SERPINA7
TIG2	RARRES2

TIMD3	HAVCR2
TIMP-1	TIMP1
TIMP-2	TIMP2
TIMP-3	TIMP3
TLR2	TLR2
TLR4:MD-2 complex	TLR4 LY96
TMA	TPO
TNF sR-I	TNFRSF1A
TNF sR-II	TNFRSF1B
TNF-a	TNF
TNF-b	LTA
TNFSF15	TNFSF15
TNFSF18	TNFSF18
TNR4	TNFRSF4
Topoisomerase I	TOP1
tPA	PLAT
TPSB2	TPSB2
TPSG1	TPSG1
TRAIL R1	TNFRSF10A
TRAIL R4	TNFRSF10D
transcription factor MLR1 isoform CRA_b	LCORL
Transferrin	TF
Transketolase	TKT
TrATPase	ACP5
Triosephosphate isomerase	TPI1
TrkA	NTRK1
TrkB	NTRK2
TrkC	NTRK3
Tropomyosin 1 alpha chain	TPM1
Tropomyosin 2	TPM2
TRY3	PRSS3
Trypsin	PRSS1
Trypsin 2	PRSS2
TS	TYMS
TSG-6	TNFAIP6
TSH	CGA TSHB
TSLP	TSLP
TSLP R	CRLF2
TSP2	THBS2
TSP4	THBS4

TWEAK	TNFSF12
TWEAKR	TNFRSF12A
TXD12	TXNDC12
TYK2	TYK2
UB2L3	UBE2L3
UBC9	UBE2I
UBE2N	UBE2N
Ubiquitin	RPS27A
Ubiquitin+1	RPS27A
UFC1	UFC1
UFM1	UFM1
ULBP-1	ULBP1
ULBP-2	ULBP2
ULBP-3	ULBP3
UNC5H3	UNC5C
UNC5H4	UNC5D
uPA	PLAU
URB	CCDC80
Vasoactive Intestinal Peptide	VIP
VCAM-1	VCAM1
VEGF	VEGFA
VEGF sR2	KDR
VEGF sR3	FLT4
VEGF-C	VEGFC
VEGF121	VEGFA
vWF	VWF
WFKN2	WFIKKN2
WIF-1	WIF1
WISP-1	WISP1
WNT7A	WNT7A
XEDAR	EDA2R
XPNPEP1	XPNPEP1
XTP3A	DCTPP1
YES	YES1
ZAP70	ZAP70

Statistical analysis

Univariate Test of Significance: Student t-test

The student t-test was used to find proteins that were differentially expressed between DKK1 overexpressing cells and control cells. The Bonferroni method was used for multiple testing correction.

Characterizing a DKK1 induced multivariate signal

We used partial least squares regression to characterize a multivariate signature that separated *DKK1* overexpressed cells from control cells. PLS regression was deliberately chosen over related multivariate approaches like principal component analysis (PCA). While PCA finds the components that capture the most variance in your predictors (X), it is agnostic to whether these components are associated with the response variable (Y). This is because PLS performs a simultaneous decomposition of X and Y with the additional constraint of finding latent components that maximize the covariance between the protein predictors and *DKK1* overexpression status. This ensures that PLS finds hidden factors in the data that can be used to predict *DKK1* overexpressing cells from normal cells.

A partial least square regression model was fit to our data using all 1128 proteins as the predictors (X) and the *DKK1* or control status as the response variable (Y). PLS models were built using the R package "PLS" using two components. We then used the calculated coefficients from the resulting PLS regression model predicting for *DKK1* overexpression status to rank proteins. The coefficients corresponding to each protein in the model are a proxy for how much each protein contributes to the signal. The top 100

proteins that contributed to this multivariate signature were reported. The top 100 proteins were chosen for subsequent analysis because they represent approximately the top 10% of all quantified proteins (reducing the size of our feature set). Specifically, we defined a significant coefficient threshold which represented the coefficients that contributed to >95% of the information criterion of the model. The value of the coefficient of the 100th protein exists at the cut-off for the criterion (**Fig. 2.3**).

Pathway Analysis of Top 100 Proteins

We evaluated the top 100 PLS ranked proteins in order to evaluate the biological significance of the *DKK1*-induced signature. We used the DAVID Bioinformatics Resource, version 6.7 Functional Annotation tool and performed enrichment analysis on the KEGG database. *(64, 65)* The 100 proteins were inputted as our 'gene list' while all 1128 proteins quantified in the study were inputted as our 'background gene list.' This analysis assigned probabilities to the distribution of proteins observed in *DKK1* list versus those expected under a random draw of 100 proteins from the 1128 protein set. The top enriched pathways were reported, along with associated p-values (corrected and uncorrected).

Association of DKK1 signature characterized in cell-line experiments in the plasma of AD patients

90 out of the 100 Top ranked proteins in the *DKK1* signature had been previously quantified in the plasma of 320 AD patients and 209 normal controls as part of an earlier study from our group using a previous version of the same assay. *(53)* Empirical Bayes

Method was used to adjust the AD and control patient data for collection site, age, gender, and presence of ApoE4 alleles *(66)* with the R package 'sva.' We used these 90 *DKK1*-induced proteins to fit PLS models using the *DKK1* data and AD data. Specifically, a PLS model was fit using the values of the top 90 proteins quantified in the lysates of *DKK1 overexpressing HEK293 cells* or control data as the predictor variables and the *DKK1* or control label as the response variable. A PLS model was also fit using the values of top 90 proteins quantified in the plasma of AD patients and controls as the predictor variables and the AD versus control label as the response variable. We then calculated the Spearman's correlation between the coefficients of the two models (i.e., the DKK1 versus control model and the AD versus control model). A strong correlation would suggest that the signature induced by *DKK1* overexpression is found in the plasma of AD patients.

Monte Carlo methods to test significance of PLS model coefficient correlations

In order to test the significance of the correlation we found between the two models, we employed a Monte Carlo approach to find the probability of achieving a correlation as high simply by chance when two PLS models are generated from our data sets using the same procedures outlined in the above section. First, we fit partial least square regression models using all 1128 proteins as the predictor variables and using the shuffled DKK1 or control labels as the response variables. Next, we used the PLS regression coefficients to rank the top 90 proteins. PLS models were again fit using the top 90 proteins using the AD data with shuffled AD and control labels. To determine how strongly the shuffled DKK1 signature associates in the blood of AD patients, we

calculated the Spearman's correlation between the PLS model coefficients calculated from the AD data with shuffled AD and Control labels compared to the PLS coefficients calculated from the shuffled DKK1 data using the top 90 ranked proteins from the shuffled DKK1 data. This was done for 100,000 iterations where the DKK1 versus control label and AD versus control label were shuffled. We then compared the correlation of the unshuffled data with the other 100,000 shuffled correlations.

Classification performance of DKK1 induced proteins in predicting AD versus control

RF classifiers were trained using data from the 320 AD samples and 209 control samples. The top 90 proteins nominated from the DKK1 overexpression PLS model were used as the predictors for AD versus control classification. Performance was assessed using crossvalidation, specifically the bootstrap .632+ algorithm (1000 samples). The 'randomForest' package in R was used to train the classifiers. We found that a classifier trained using the 90 proteins achieved an area under the curve of 0.805.

Stability Selection and finding a sparse prediction panel

Stability Selection *(67)* using the Least Absolute Shrinkage and Selection Operator (LASSO) *(68)* method was used to identify consistently important biomarkers and rank them across 10,000 iterations. At each iteration, 30% of the proteins and 10% of the samples were removed and LASSO was used to feature-select for variables on the remaining data. The 90 proteins were ranked by the proportion of iterations in which they were found to have a non-zero coefficient using LASSO.

RF Classifiers were then created using from n=1 to n=90 of the proteins. We found that a 22-protein classifier achieved the highest performance. In order to test the significance of achieving this level of classification performance using only 22 proteins, we generated 1000 random sets of 22 proteins and built classifiers using them as predictors. We then assessed the performance of each of these classifiers using crossvalidation, specifically the bootstrap .632+ algorithm (1000 samples) *(69)*. The significance of the performance of the original 22-protein classifier reflects the proportion of times it had a higher AUC than the other 1000 classifiers built using random sets of 22 proteins. All scripts and analysis can be found at https://drive.google.com/drive/folders/0B2MOTaIrtCx6dFRHN0FZUlVvak.

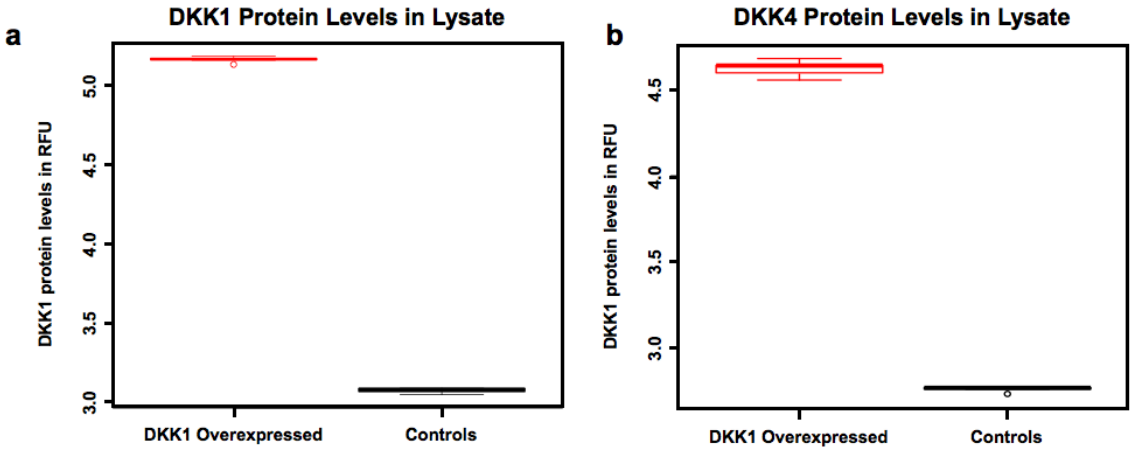
Results

DKK1 and DKK4 were differentially expressed in DKK1 overexpressed cells compared to controls

In order to determine the protein signature induced by DKK1, we first compared lysate from HEK293 cells engineered to over-express human *DKK1* to control cells transfected with empty vector (n=5 in each case) using an aptamer capture proteomics platform. *(21)* In total, 1128 proteins were quantified in these samples using the SOMAScan proteomics platform. We identified proteins with significantly different concentrations in *DKK1* overexpressing and control cells using the Student t-test (**Fig. 2.2**). Two proteins were significantly differentially expressed after multiple testing correction ($P < 4.4 \times 10^{-5}$, the Bonferroni correction level). As expected, DKK1 protein levels were considerably higher in lysate collected from *DKK1* overexpressing cells

compared to controls (123-fold increase in untransformed RFU; **Fig. 2.2a**) but in addition, DKK4 was also found to be significantly higher (**Fig. 2.2b).** Moreover, 61 other proteins were differentially expressed at the uncorrected $p < 0.05$ level (**Table 2.3**).

Fig. 2.2: DKK1 overexpression leads to higher levels of DKK1 in Lysate



The SOMAScan proteomics platform was used to quantify 1128 proteins in lysate from 5 replicates each of DKK1 overexpressed HEK293A cells and HEK293A control cells. **a.** We observed that lysates from the DKK1 overexpressed replicates (red boxplot) had higher protein quantifications of DKK1 (123-fold increase in untransformed RFU) compared to control replicates (black boxplot). **b.** In addition, DKK4 was also found to be overexpressed in DKK1 overexpressing cells compared to controls (75-fold increase in untransformed RFU).

Table 2.3: The 63 differentially expressed proteins between DKK1 overexpressing and control cells

Protein	Gene Name	Coefficient	P-Value
DKK1	DKK1	178.220	1.45E-15
DKK4	DKK4	82.674	6.94E-09
D-dimer	FDP	-6.717	1.57E-04
SAP	SAP	4.511	2.05E-03
LAG-1	LAG1	4.862	2.08E-03
RGM-C	HFE2	4.311	2.71E-03
CHST2	CHST2	4.146	3.24E-03
MAPKAPK3	MAPKAPK3	-3.989	4.08E-03
ULBP-2	ULBP2	-3.673	6.28E-03
Ferritin	FT	-3.975	6.58E-03
M-CSF R	CSF1R	3.601	7.29E-03
CLF-1/CLC Complex	CLF-1/CLC Complex	-3.557	7.44E-03
ROBO2	ROBO2	3.640	1.01E-02
Glucocorticoid receptor	NR3C1	3.257	1.23E-02
Endostatin	COL15A1	3.214	1.24E-02
TARC	TARC	3.250	1.34E-02
Angiotensinogen	AGT	-3.154	1.35E-02
PTP-1B	PTPN1	-3.141	1.48E-02
CONA1	CONA1	3.086	1.53E-02
BTK	BTK	3.361	1.55E-02
Endothelin-converting enzyme 1	ECE1	3.055	1.57E-02
TPSB2	TPSB2	-3.102	1.73E-02
IL-12	IL12	2.974	1.83E-02
IL-12 RB2	IL12RB2	-3.044	1.91E-02
CD109	CD109	3.039	1.93E-02
HAI-1	HAI1	2.938	1.99E-02
TEC	TEC	2.956	2.06E-02
HGFA	HGFA	2.910	2.35E-02

Coactosin-like protein	COTL1	-2.856	2.38E-02
Siglec-9	SIGLEC9	-2.962	2.74E-02
IL-18 Rb	IL18RB	2.918	2.78E-02
MIS	MIS	-2.742	2.83E-02
BFL1	BFL1	-3.090	2.86E-02
GV	GV	2.640	2.98E-02
PDK1	PDK1	-2.718	3.01E-02
Glypican 3	GPC3	2.621	3.07E-02
GPVI	GPVI	2.590	3.22E-02
GFRa-1	GFRA1	-2.663	3.42E-02
WFKN2	WFKN2	2.577	3.42E-02
PKC-A	PKCA	2.570	3.47E-02
AGR2	AGR2	-2.610	3.49E-02
CD226	CD226	2.554	3.49E-02
P-Selectin	SELP	2.551	3.51E-02
CHL1	CHL1	2.721	3.76E-02
Coagulation Factor X	F10	2.698	3.76E-02
Glutamate carboxypeptidase	FOLH1	2.803	3.80E-02
Caspase-2	CASP2	2.458	3.95E-02
SCGF-beta	CLEC11A	-2.448	4.17E-02
VEGF sR2	VEGFSR2	2.544	4.20E-02
IL-20	IL20	2.447	4.23E-02
sTie-1	sTie-1	2.411	4.27E-02
TrkA	TRKA	2.418	4.29E-02
FGF-4	FGF4	2.446	4.32E-02
CNDP1	CNDP1	2.422	4.33E-02
BNP-32	BNP32	2.428	4.34E-02
40S ribosomal protein SA	RPSAb	-2.655	4.37E-02
TRAIL R1	TRAILR1	2.755	4.56E-02
CLC1B	CLC1B	-2.689	4.62E-02
C2	C2	2.367	4.65E-02
HPLN1	HPLN1	2.357	4.71E-02
MK01	MK01	-2.360	4.72E-02

complement factor H-related 5	CFHR5	2.598	4.73E-02
Cadherin-2	CDH2	-2.328	4.96E-02

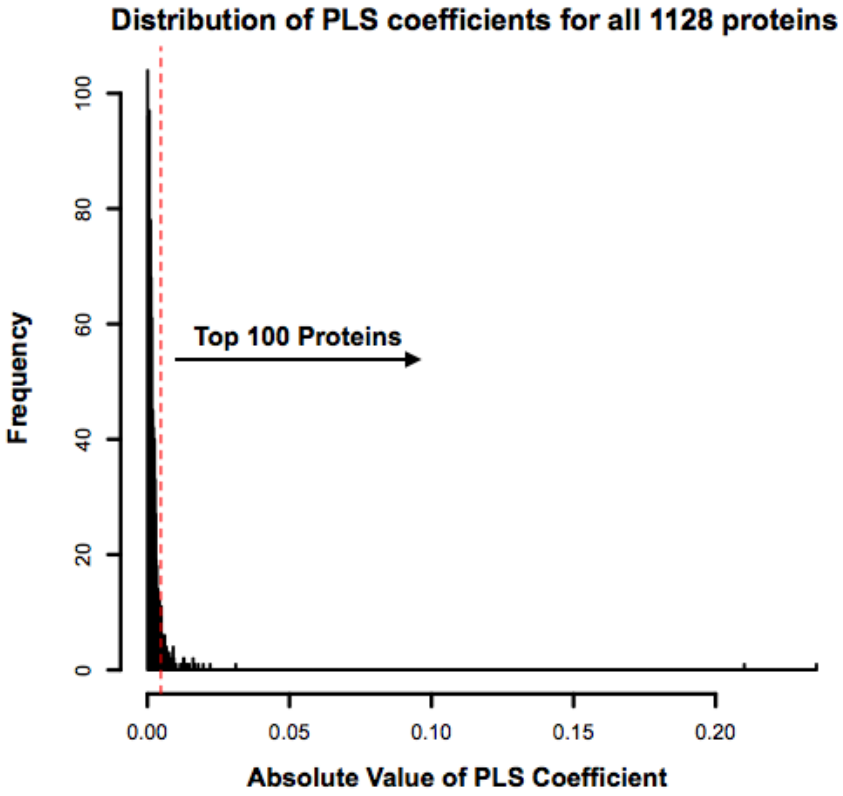
63 proteins were differentially expressed between DKK1 overexpressing cells and controls cells ($p < 0.05$). Protein identities, coefficients, and p-values are displayed.

Overexpression of DKK1 induces a multivariate proteomic signature that is enriched in AD pathways

Previous studies from my supervisor's lab have shown that overexpression of *DKK1* induces a clusterin/p53/DKK1/Wnt-PCP-JNK pathway, which subsequently drives the upregulation of several genes which mediate the development of AD-like pathological processes. *(49)* We wanted to characterize how *DKK1* overexpression may induce multivariate changes in proteomic expression, and whether some of the same biological pathways may be affected by these changes.

To identify the proteins that most strongly mapped to a multivariate signature that distinguished DKK1 overexpressed replicates from control replicates, we used partial least squares regression (PLS). This technique finds components that maximize the covariance between the protein predictors and *DKK1* overexpression status (**Fig. 2.4a**). In brief, PLS finds hidden factors in the data that can be used to predict DKK1 overexpressed cells from normal cells. A PLS regression model was fitted to our data using all 1128 proteins as the predictors and the DKK1 or control status as the response variable. The calculated coefficients for the resulting PLS regression model predictive for DKK1 overexpression status were used to rank proteins by how much that protein contributed to the multivariate signature. We ranked the top 100 proteins that contributed to this multivariate signature (**Table 2.4**), which puts these proteins at the greater than top 10% of all quantified proteins. A histogram plotting the distribution of PLS coefficients also shows that the value of the PLS coefficient of the 100th protein exists at natural cutoff before the density of proteins with lower coefficient values drastically increases (dotted red line, **Fig. 2.3**).

Fig. 2.3: Distribution of PLS coefficients of all 1128 proteins



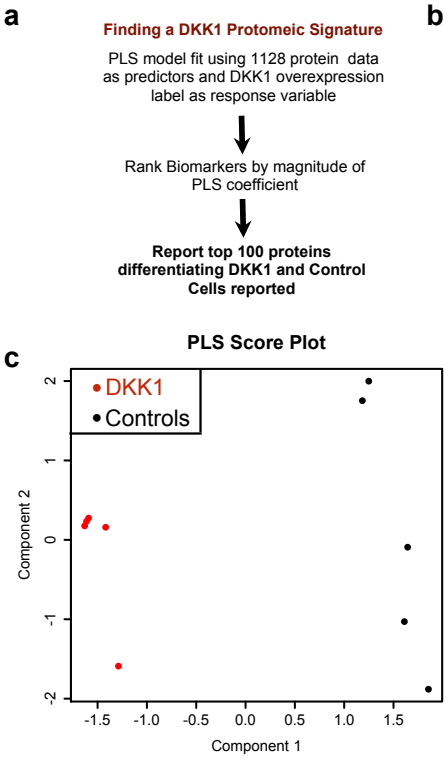
The distribution the coefficients (absolute value) of each protein in PLS regression model fitted to the DKK1 overexpression data is shown in a histogram. The overexpression status, DKK1 or Control, was used as the response variable and the protein data levels were used as the predictor variables. Proteins with larger coefficients in the PLS model contributed most to the signature separating DKK1 overexpressing cells and controls. Values to the right of the dotted red represent the top 100 proteins ranked by the absolute value of the PLS coefficient. The top 100 proteins were chosen for subsequent analysis because they represent approximately the top 10% of all quantified proteins and the value of the PLS coefficient of the 100th protein exists at natural cutoff before the density of proteins with lower coefficient values drastically increases.

The PLS scores plot strongly supports the notion of a signature that differentiates DKK1 samples from controls (**Fig. 2.4c**). DKK1 and DKK4 were also the two proteins with the highest PLS regression coefficient, suggesting that this multivariate signature is weighted highly on DKK1.

In order to assess the biological significance of this signature, we evaluated the 100 top ranked proteins for biological pathway enrichment using the DAVID Bioinformatics Resource, version 6.7 Functional Annotation tool, performed on the KEGG database *(64, 65)* (**Fig. 2.4b**). Notably, Alzheimer's Disease was our top ranked pathway ($p = 7.35 \times 10^{-4}$), while other pathways previously implicated in AD pathogenesis, including the p53 pathway ($p = 0.0248$), the Wnt signalling pathway ($p = 0.018$, all branches), and the long term potentiation pathway ($p = 0.0318$) were also found to be statistically enriched. *(49, 70)* We note that the significance of our enriched pathways does not pass more conservative multiple testing corrections (**Fig. 2.4b**). This is unsurprising because the SOMAscan assay only quantifies ~1000 proteins (which make up our background gene list in DAVID), limiting our power to see significant pathway enrichments after multiple testing correction. However, we had previously identified many of the exact same pathways in gene expression studies of DKK1 overexpressing rat primary neurons, adding to our confidence of the results. *(49)*

Fig. 2.4: A DKK1 induced multivariate signature is enriched in pathways found to influence AD pathogenesis

a



b

KEGG Pathway	Proteins	P-value (uncorrected)	Benjamini corrected p-value
hsa05010:Alzheimer's disease	P99999, P04406, P49841, P28482, Q99714, Q08209, P27361, P63098	7.35E-04	3.03E-01
hsa04062:Chemokine signaling pathway	P49841, Q05397, P02775, P28482, P41240, P19875, P27361, P19876, P49840, P25098, P29353, Q05513	1.09E-02	3.44E-01
hsa04310:Wnt signaling pathway	Q8N474, P49841, O94907, Q9UBT3, Q08209, Q13557, P63098	1.18E-02	2.74E-01
hsa05222:Small cell lung cancer	P99999, P02751, Q05397, P35354, P24941, P46527	2.16E-02	3.61E-01
hsa04115:p53 signaling pathway	P99999, P05121, O96017, P24941, P07996	2.48E-02	3.42E-01
hsa04720:Long-term potentiation	P28482, Q08209, P27361, Q13557, P63098	3.18E-02	3.61E-01
hsa04012:ErbB signaling pathway	P49841, Q05397, P28482, P27361, P16333, P46527, Q13557, P29353	3.48E-02	3.36E-01
hsa04650:Natural killer cell mediated cytotoxicity	P04141, Q9UBN6, Q06124, P28482, Q08209, P27361, O00220, P63098, P29353	4.22E-02	3.51E-01
hsa04360:Axon guidance	Q96MS0, P49841, Q05397, P28482, Q08209, P27361, P16333, P63098	5.05E-02	3.73E-01
hsa04660:T cell receptor signaling pathway	P04141, P49841, P28482, Q08209, P27361, P16333, O15264, P63098	5.05E-02	3.73E-01
hsa04370:VEGF signaling pathway	Q05397, P35354, P28482, Q08209, P27361, O15264, P63098	5.29E-02	3.59E-01
hsa04662:B cell receptor signaling pathway	P49841, P28482, Q08209, P27361, P63098	5.87E-02	3.68E-01
hsa04114:Oocyte meiosis	P28482, Q08209, P27361, P24941, Q13557, P63098	6.15E-02	3.54E-01
hsa04210:Apoptosis	P99999, Q9UBN6, Q08209, Q92851, O00220, P63098	7.07E-02	3.73E-01
hsa04722:Neurotrophin signaling pathway	P49841, Q06124, P28482, P41240, P27361, Q13557, O15264, P29353	7.76E-02	3.75E-01
hsa05220:Chronic myeloid leukemia	Q06124, P28482, P27361, P46527, P29353	9.48E-02	4.26E-01

a. In order to characterize a proteomic signature that was induced by DKK1 overexpression, a PLS regression model was fit to data using the 1128 proteins as the predictor variables and the overexpression status (DKK1 or control) as the categorical response variable. Proteins were ranked by the magnitude of the coefficient of the fitted PLSR model. The magnitude of the coefficient serves as a proxy for how much the corresponding protein contributes to the signature that distinguishes DKK1 overexpressed versus control replicates. The Top 100 PLS-ranked proteins are reported in **Table 2.4** and used in our downstream analysis. **b.** We evaluated the 100 top ranked proteins for biological pathway enrichment using the DAVID Bioinformatics Resource, version 6.7 Functional Annotation tool performed on the KEGG database. The most enriched pathways from the proteins induced by the DKK1 signature are shown. The

proteins that make up each pathway along with p-values are also shown (uncorrected and corrected for multiple testing). **c.** This PLS proteomic signature differentiated DKK1 replicates from control replicates. A PLS scores plot using 1st and 2nd components show that the 5 DKK1 samples (in red) cluster away from the 5 control samples.

Table 2.4: The Top 100 proteins ranked by PLS

PLS Rank	PLS regression Coefficient	Protein Name	Gene Name
1	-0.23554	Dkk1	DKK1
2	-0.21009	Dkk4	DKK4
3	0.03117	C3adesArg	C3
4	0.02213	Kallikrein 7	KLK7
5	0.01976	C3	C3
6	0.01781	Fibrinogen	FGA FGB FGG
7	0.01663	MK13	MAPK13
8	0.01643	TGM3	TGM3
9	0.01625	HIPK3	HIPK3
10	0.01608	CTAP-III	PPBP
11	0.01604	HMG-1	HMGB1
12	0.01466	Lactoferrin	LTF
13	0.01421	CSK	CSK
14	0.01344	BARK1	ADRBK1
15	-0.01331	Hat1	HAT1
16	0.01315	GAPDH liver	GAPDH
17	0.01300	Alkaline phosphatase bone	ALPL
18	0.01283	GSK-3 alpha/beta	GSK3A GSK3B
19	0.01277	eIF-5A-1	EIF5A
20	0.01260	PKC-Z	PRKCZ
21	0.01248	IMDH2	PPIF
22	0.01226	PSD7	PSMD7
23	0.01218	Chk2	CHEK2
24	0.01196	Lysozyme	LYZ
25	0.01176	Cytochrome c	CYCS
26	0.01007	BRF-1	SET
27	0.00978	COX-2	PTGS2
28	0.00942	Histone H1.2	HIST1H1C
29	0.00933	Coagulation Factor V	F5
30	0.00921	Stanniocalcin-1	STC1

31	0.00918	SHP-2	PTPN11
32	0.00916	FER	FER
33	0.00916	ARP19	ARPP19
34	0.00909	FAK1	PTK2
35	-0.00858	Topoisomerase I	TOP1
36	0.00840	HSP 40	DNAJB1
37	-0.00835	TBP	TBP
38	0.00817	AMPM2	METAP2
39	0.00793	RAP	LRPAP1
40	0.00788	Ku70	XRCC6
41	0.00774	Fibrinogen g-chain dimer	FGG
42	0.00770	GM-CSF	CSF2
43	-0.00763	M2-PK	PKM2
44	0.00758	discoidin domain receptor 1	DDR1
45	0.00749	PIGR	PIGR
46	0.00722	PA2G4	PA2G4
47	0.00703	Albumin	ALB
48	0.00688	CAPG	CAPG
49	0.00686	MK01	MAPK1
50	0.00681	kallikrein 5	KLK5
51	0.00669	DLL1	CPNE1
52	-0.00669	CHST2	CHST2
53	0.00668	annexin I	ANXA1
54	0.00662	RAN	RAN
55	0.00642	Coagulation Factor VII	F7
56	0.00635	Thrombospondin-1	THBS1
57	0.00633	UBC9	UBE2I
58	0.00623	Eotaxin	SHC1
59	0.00618	annexin II	ANXA2
60	0.00610	CAMK2D	CAMK2D
61	0.00608	UFC1	UFC1
62	0.00604	ERK-1	MAPK3
63	-0.00603	PAI-1	SERPINE1

64	-0.00601	Histone H2A.z	H2AFZ
65	0.00592	CYTT	CST2
66	-0.00592	CONA1	COL23A1
67	0.00572	C1QBP	C1QBP
68	0.00571	transcription factor MLR1 isoform CRA_b	LCORL
69	0.00569	a1-Antichymotrypsin	SERPINA3
70	0.00569	Desmoglein-2	DSG2
71	-0.00561	Gro-b/g	CXCL3 CXCL2
72	0.00558	Sorting nexin 4	SNX4
73	-0.00551	FN1.4	FN1
74	0.00546	NCC27	CLIC1
75	0.00544	TIMP-1	TIMP1
76	0.00544	PSA-ACT	KLK3 SERPINA3
77	0.00540	SARP-2	SFRP1
78	-0.00537	Calcineurin	PPP3CA PPP3R1
79	0.00537	TRAIL R4	TNFRSF10D
80	0.00537	CDK2/cyclin A	CDK2 CCNA2
81	0.00532	CKAP2	CASP10
82	-0.00519	DLRB1	DYNLRB1
83	-0.00508	PD-L2	PDCD1LG2
84	0.00504	NCK1	NCK1
85	0.00504	Clusterin	CLU
86	0.00500	MIS	AMH
87	0.00500	p27Kip1	CDKN1B
88	-0.00497	ROBO3	ROBO3
89	-0.00496	AURKB	AURKB
90	0.00494	NACA	NACA
91	0.00493	Coactosin-like protein	COTL1
92	0.00492	BFL1	BCL2A1
93	0.00492	phosphoglycerate kinase 1	PGK1
94	-0.00487	TRAIL R1	TNFRSF10A
95	0.00487	IL-18 BPa	IL18BP
96	-0.00486	M-CSF R	CSF1R

97	0.00481	TCPTP	PTPN2
98	-0.00479	IL-6	IL6
99	0.00478	NMT1	NMT1
100	0.00475	ERAB	HSD17B10

The top 100 proteins contributing the PLS signature that distinguished *DKK1* overexpressing and control cells are reported. Proteins were ranked by their corresponding coefficient for the fitted PLS model.

The DKK1 induced signature is found in the plasma of AD patients

Having characterized a DKK1 induced proteomic signature, we wanted to explore whether the same signature could be found in the plasma of AD patients. Finding this signal in the plasma of AD patients would not only provide evidence for its relevance to AD pathogenesis, but may also suggest its potential role as a diagnostic and prognostic peripheral biomarker signature for AD.

1016 proteins, including both DKK1 and DKK4, had been previously quantified in 320 plasma AD samples and 209 control samples using an earlier version of the SOMAScan assay. Protein data was first adjusted age, gender, and presence of APOE ε4 alleles using an empirical Bayes approach. *(66)*

Both DKK1 and DKK4 were not found to be differentially expressed in the plasma of AD patients and Controls. Because it is likely that simple univariate protein differences fail to capture to complexity of neurological diseases such as AD and may be especially difficult to detect in peripheral blood plasma, we sought to determine whether the DKK1-induced multivariate signature was found in the plasma of AD patients.

Out of the 100 proteins that made up our DKK1 induced proteomic signature, 90 proteins had been previously quantified in the plasma of these AD patients. PLS models were built using the levels of the 90 proteins in the plasma of AD and control patients as predictor variables and disease status (AD or controls) as the response variable.

In order to determine whether the multivariate DKK1 signature was found in the blood of AD patients, we calculated the Spearman's correlation between the PLS model coefficient for the 90 proteins fitted to the DKK1 data compared to the AD data. The correlation measures how strongly the DKK1 signature is found in the AD data. The

calculated Spearman's correlation of rho=0.217 had a p-value=.04, suggesting that our *DKK1* overexpression signature was indeed found in the blood of AD patients.

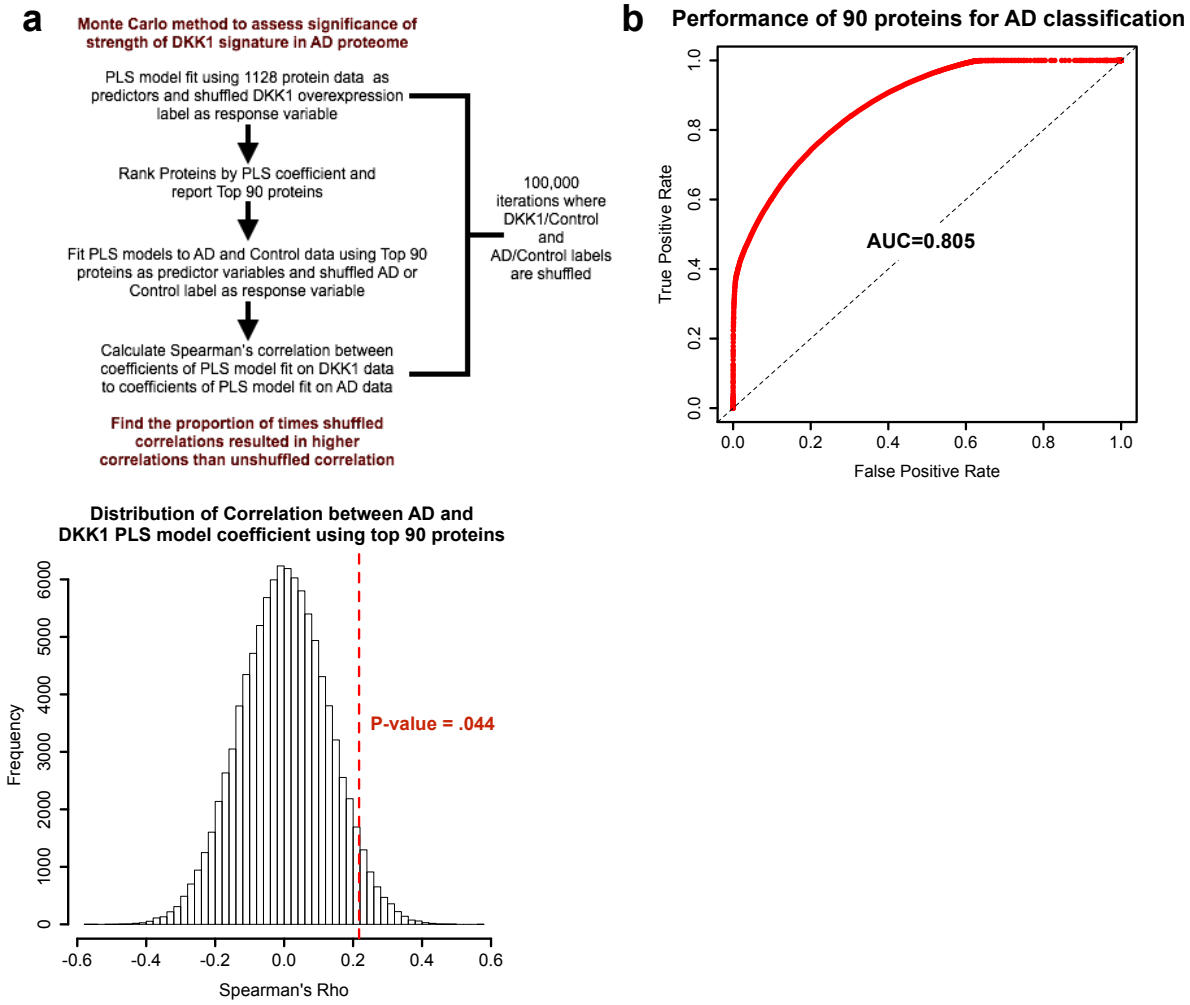
However, it is possible that a correlation as strong or stronger than rho=0.217 between two random PLS models could emerge by chance alone. In order to determine the probability to obtaining a Spearman's correlation as strong as the one we observed using the procedures we employed (i.e. using PLS coefficients from the PLS model fitted to DKK1 overexpressing cells to nominate top proteins, then building the PLS model on the AD data using the top ranked proteins from the first model, then calculating correlation between the coefficients from the two PLS models), we used a Monte Carlo permutation based approach (**Fig. 2.5a**).

First, we fitted PLS regression models using all 1128 proteins as the predictor variables and shuffled DKK1 or control labels as the response variables. Next, we used the PLS regression coefficients to rank the top 90 proteins. PLS models were again fitted using the top 90 proteins using the AD data with shuffled AD or control labels. To determine how strongly the randomly generated "DKK1" signature correlates in the blood of AD patients, we calculated the Spearman's correlation between the PLS model coefficients calculated from the AD data with shuffled AD and control labels compared to the PLS coefficients calculated from the shuffled DKK1 data using the top 90 ranked proteins from the shuffled DKK1 data. This was done for 100,000 iterations where we shuffled the DKK1 versus control and AD versus control labels. Only 4.4% of the iterations resulted in a stronger correlation than was found for the un-permuted data. Therefore, the correlation between the DKK1 signature in the cell line experiment and plasma of AD patients had a significance of .044.

The DKK1 induced signature has diagnostic power in predicting AD patients from normal controls

We wanted to determine whether the 90 proteins had predictive power in distinguishing AD samples from controls. RF classifiers were built using the 90 proteins as predictors and performance was assessed using 1000 iterations of the bootstrap .632+ crossvalidation algorithm. *(69)* The classifiers performed with an AUC of .805, sensitivity of 80%, and specificity of 74% (**Fig. 2.5b**).

Fig. 2.5: The DKK1 induced signature is found in the plasma of AD patients



a. Monte Carlo methods were used to assess the significance of the correlation between the DKK1 induced signature found in the overexpressed cell lines and blood of AD patients (rho=0.217). The histogram plots the distribution of the correlations calculated for the 100,000 iterations where DKK1 versus control labels and AD versus control labels were shuffled. Only 4.4 percent of the iterations achieved a correlation larger than rho=0.217 (dotted red line). Therefore, the p-value associated with the strength of the correlation between the DKK1 overexpression signal found in cells and the signal found in AD patients is p=.044. **b.** We wanted to determine whether the 90 proteins had

predictive power in distinguishing AD samples from controls. RF classifiers were built using the 90 proteins as predictors and performance was assessed using 1000 subsamples of bootstrap .632+ algorithm. The classifiers performed with an AUC of .805 with sensitivity of 80% and specificity of 74%.

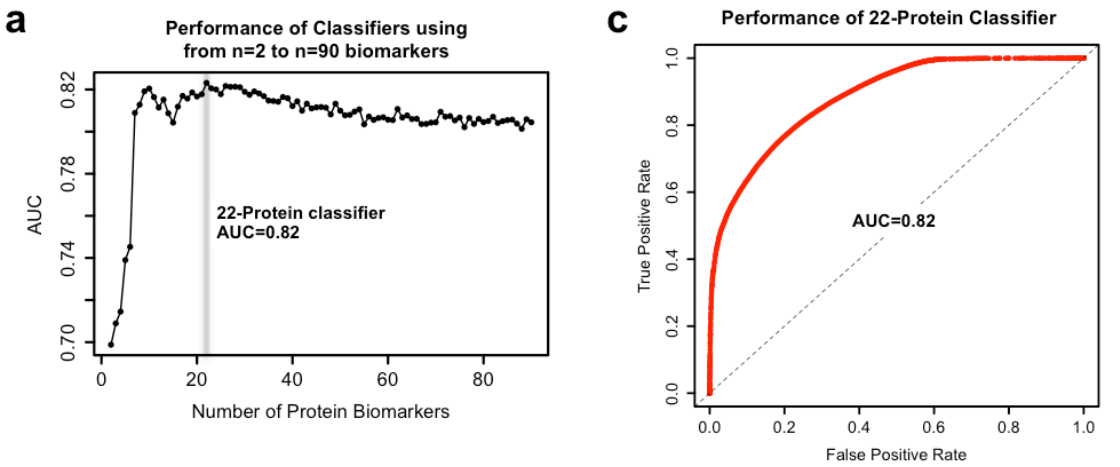
Next, we wanted to see whether we could further reduce down these 90 proteins to a smaller panel that could achieve high diagnostic accuracy (**Fig. 2.6a**). From a practical standpoint, a smaller protein panel is less expensive and easier to implement in clinical settings. From a statistical perspective, a robust and "sparse" protein panel removes redundant or noisy signals and reduces the probability of over-fitting, thus improving the performance of a diagnostic classifier. To find this "sparse" panel, we used Stability Selection, *(67)* a meta-statistical tool that identifies consistently important features by repeated sub-sampling of the data with LASSO (Least Absolute Shrinkage and Selection Operator) *(68)*. LASSO is a method that eliminates redundant signals to create a sparse model of the data. Within Stability Selection, the 90 candidate biomarkers were ranked by the number of times LASSO included the protein in the model across the 10,000 iterations. At each iteration, 10% of the samples and 30% of the proteins were randomly removed, and LASSO was used to identify a sparse set of proteins using the remaining jack-knifed data.

RF classifiers were then built using from n=1 to n=90 biomarkers (Fig 4A) and classification performance was assessed using 1000 iterations of the .632+ bootstrap method. We found that a panel of 22 biomarkers (**Fig. 2.6b, 2.6c**) resulted in the highest classification performance ranked by AUC (AUC=.82, Sensitivity=80%, Specificity=77%). Interestingly, DKK4 was identified among the top ranked biomarkers while Clusterin, the well-replicated AD genetic risk variant and blood biomarker, was also among the 22 proteins making up the classifier.*(71–74)* The presence of DKK4 in the 22-protein classifier corroborates our finding that a DKK1-related signature is found in the plasma of AD patients. The fact that DKK4 is among the 22 proteins and *not*

DKK1 is a consequence of our goal to find a sparse diagnostic signal. Because DKK1 and DKK4 are highly correlated (**Fig. 2.6d**), they encode redundant information from a statistical perspective. Since the LASSO penalizes redundant signals in the data, DKK1 was ranked lowly after already including DKK4 in the model.

It is possible that random sets of 22 proteins could potentially predict AD patients with high probability. To test whether we could achieve this level of performance by chance alone, we created 1000 random sets of 22 proteins. We then built classifiers using each of these random sets of proteins. We found that no randomly generated set of 22 proteins performed as well as our 22-protein classifier nominated from the DKK1 overexpression analysis. This suggests that the proteins nominated from the DKK1 overexpression analysis have high diagnostic significance.

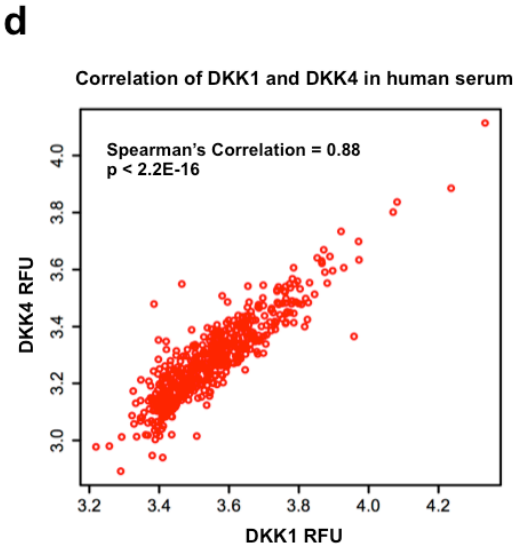
Fig. 2.6: The 90 proteins can be reduced to a smaller 22-protein classifier with high diagnostic performance



b

The 22 proteins in the classifier

Protein Name
DKK4
GADPH liver
Stanniocalcin-1
discoidin domain receptor 1
Albumin
FN1-4
PSA-ACT
BFL1
CAMK2D
Clusterin
Ku70
Kallikrein-7
M-CSF-R
annexin-2
Cytochrome c
IL-6
Fibrinogen
UBC9
Lysozyme
MK13
AURKB
Alkaline phosphatase bone



a. We sought to determine whether we could reduce down the 90-proteins to a smaller panel to improve diagnostic performance. The 90 biomarkers were ranked using Stability Selection with the LASSO method. RF classifiers were then build using n=2 to n=90 of the proteins. We found that a panel of 22 proteins achieved the highest diagnostic performance. **b.** The identities of the 22 proteins that make up the highest performing classifier are shown. Interestingly, DKK4 is among the 22-proteins in the classifier, suggesting that the DKK1 signature is contributing to the diagnostic performance**. c.** The

ROC curve for 22-protein classifier is shown. Performance was assessed using 1000 subsamples of bootstrap .632+ crossvalidation algorithm. The 22-protein classifier achieved an AUC of .82 (Sensitivity=80%, Specificity=77%). **d.** DKK1 and DKK4 are highly correlated in human plasma (rho=0.88, $p < 2.2E-16$), meaning they encode overlapping or redundant diagnostic information. The reason why DKK4 is among the 22 proteins in the classifier and not DKK1 is due to the fact that LASSO penalizes redundant signals in the data. Since DKK4 is already included in the LASSO model, DKK1 would be not included as it does not provide much additional information beyond what DKK4 provides.

Discussion

In this study, we employed an empirically derived feature selection approach, driven by experimental data from an *in vitro* model, for targeted biomarker discovery using the *a priori* information that DKK1 overexpression is a hallmark of Aβ induced neuropathology. We hypothesized that DKK1 overexpression would modulate protein expression in pathways implicated in AD and induce a plasma-based signature that had diagnostic power in distinguishing AD patients from normal controls.

We quantified 1128 proteins in 5 lysate replicates of DKK1 overexpressing cells and control cells transfected with empty vector and found that *DKK1* overexpression resulted in a unique multivariate signature which induced biological pathways previously implicated in AD. The two most changing proteins included not only DKK1 but also its family member, DKK4. That it was this isoform and not the other isoforms in the family is in line with the observations that DKK1 and DKK4 are the most similar in terms of biological actions. *(75)* This is consistent with studies conducted by my supervisor's lab finding that DKK1 and DKK4 both upregulated *EGR1* and *c-FOS* in primary neurons, while the other isoforms in the family did not. Interestingly, the top enriched pathways induced by *DKK1* overexpression included the Alzheimer's Disease pathway, the Wnt signalling pathway, the p53 pathway, and the long-term potentiation pathway. Previous studies, including those from Simon Lovestone's lab, have found that p53 and Wnt signalling pathways are part of the downstream cascades involved in Aβ-mediated toxicity. *(49–52)* Studies have also shown that Aβ peptides impair long-term potentiation and affect axon guidance in many disease models of AD. *(76, 77)* The protein signatures found in DKK1 overexpressed cells were associated with signatures in the plasma of

Alzheimer's disease patients. Finally, the top proteins that made up this signature had significant diagnostic power in distinguishing AD patients from normal controls.

The proteins we identify as a signature consequent to DKK1 expression participate in pathways known to be involved in AD pathogenesis and hence might be considered as prime targets for therapeutic intervention. Moreover, the signature we observe in vitro is reflected in vivo in blood from people with AD compared to aged controls without known disease. These data therefore add validity to the previously identified Aβ-p53-DKK1-Wnt pathway that also includes the well-replicated genetic risk variant and putative biomarker, Clusterin, which was also among the proteins making up the 22-protein classifier. Moreover, as this protein signature is detectable and significantly altered in disease then these proteins might also function as diagnostic, theragnostic, or other companion biomarkers for AD.

Sattlacker *et al.* 2014 used a purely computational pipeline in the same human plasma data set to develop a diagnostic classifier. In that study, the team identified 13 proteins that predicted AD vs Control with a AUC of 0.70. Only 1 protein, FN1-4, in the 13-protein classifier in the Sattlacker paper overlapped with 22-protein classifier we nominated using the in vitro feature selection method we have described here. Our 22-protein classifier also achieved higher diagnostic performance (AUC = .82), demonstrating some of the potential benefits of how an in vitro selection approach can help converge towards meaningful diagnostic signatures more efficiently than purely computational approaches.

Beyond neurodegeneration this study has wider implications for experimental design in biomarker discovery. In general terms, biomarker discovery tends to be either

broad and untargeted or narrow and targeted. The former, hypothesis independent approach, utilises one of the increasing number of very large variable biological analytes platforms (the various 'omics technologies) and the latter typically analyses a single or small number of analytes based on a hypothesis generated from previous studies. In the study we report here we have in effect combined both approaches combining the power of hypothesis driven research together with the power of very large variable capture. This alternative approach exploits the advantages of hypothesis driven analysis based on experimentally derived data together with the unarguable power of a platform technology that is rapidly scaling to increase its proteomic coverage.

While the methods included in this study are useful for biomarker discovery, there are major limitations to this study, specifically as it relates to the cells we ran the experiment with. The study was conducted with HEK293 cells, which are derived from kidney. We had chosen this cell type because of the high efficiency of transient transfection and the fact that this cell line has an intact canonical WNT pathway, thus being amenable to DKK1 modulation. However, ideally, it would have been desirable to have used human neuronal data. Replicating this experiment using human neuronal data, for example, using iPSC-derived neural stem cells from a patient with AD, would be a better model. Nevertheless, the methods presented here can be applied to any disease model system.

In summary we have generated using experimental models *in vitro* a protein signature of DKK1 expression using the SOMAmer proteomic platform technology. Pathway analysis of this protein signature reflects processes known to be involved in AD pathogenesis and pathways previously identified as being generated by both DKK1 and

Aβ in cell models and present in animal models of amyloid pathology and in brain from people with AD. Moreover, the multivariate protein signature established *in vitro* is detectable *in vivo* in blood in people with AD. Together, this combination of experimentally derived, empirically driven data together with large scale proteomics further validates the Aβ-p53-DKK1-Wnt pathway in AD and adds weight to this pathway not only for understanding of disease pathogenesis and therapeutic target identification but now also for biomarker discovery.

Chapter 3: Protein Biomarkers for Parkinson's Disease

Chapter Summary

Parkinson's disease (PD) is a progressive neurodegenerative disorder affecting five million people worldwide. Current diagnosis of PD relies almost entirely on clinical examination, with no laboratory-based confirmatory testing available. Thus, a molecular test for PD could guide clinical care and accelerate therapeutic development. In this study, we used an aptamer-based screening platform and a mass spectrometry assay to measure levels of 1129 serum proteins and of 206 proteins, respectively, in a cohort of 161 PD and 66 normal control individuals. Using univariate methods, we identified 1) 44 diagnostic markers that differentiated PD from controls and 2) 63 endophenotypic markers that differentiated PD patients with increasingly severe PD symptoms ('Poor' PD and 'Good' PD groups), assessed by clinically validated motor and cognitive scores. The biomarkers nominated by univariate methods however are not, in aggregate, statistically significant, as we would expect 56 significant proteins just by chance. A feature selection approach using RF identified a 22-protein machine learning classifier that predicted PD from controls with an AUC =0.91 and a 24-protein classifier that predicted Good versus Poor PD patients with AUC=.89. We wanted to see whether this same diagnostic signature is found in the brain. We were able to quantify 21 out of the 22 proteins nominated from the serum analysis in the brain. The 21-protein classifier trained using the proteins nominated from the serum data performed with over AUC >.83, suggesting that the serum PD signature is also validated in the brain.

Introduction

Affecting more than five million people globally, PD is characterized by the progressive loss of dopaminergic neurons in the substantia nigra, resulting in a clinical syndrome defined by bradykinesia, rigidity, tremor, and postural instability. *(3)* Current medical practice for the diagnosis of PD relies almost entirely on clinical examination, with no laboratory-based confirmatory testing available. Clinical diagnosis is approximately 80% accurate in patients followed longitudinally with moderate symptoms. However, this accuracy may fall substantially, to approximately 65%, in earlier stages of disease. *(25)* In this context, a molecular test for PD could guide clinical care and accelerate therapeutic development. Indeed, in AD, this laboratory-test-corroborated approach has recently been adopted in several clinical trials, using the cerebrospinal fluid (CSF) biomarkers tau and amyloid-beta.

To date, however, no reliable laboratory-based test for the confirmation of PD exists. While reduction of CSF α-synuclein protein levels in PD subjects has been demonstrated by multiple groups, diagnostic accuracy for models using CSF α-synuclein as a biomarker is still limited. *(31, 34, 38)* Moreover, although an FDA-approved, radioligand-based imaging test (DaTSCAN) exists for confirmation of PD diagnosis, time and expense have prevented its widespread adoption in clinical settings. *(24)* Thus, the development of a blood-based marker for PD diagnosis could transform screening and diagnosis for PD.

We sought to identify a protein signature in the blood for early PD diagnosis. Within PD patients, we also wanted to identify signatures that differentiated patients with more severe PD motor and cognitive symptoms compared to patients with less severe

symptoms. To identify these protein signatures, we used an aptamer screen of more than 1129 proteins and a MS based screen of 206 proteins. We then validated these proteins signatures using postmortem brain tissue to determine if serum proteins discovered in early-stage PD could also be found in the brain tissue of late-stage PD in the anterior cingulate cortex. Biomarkers that are common across disease stages and tissue-types, are more likely to be biologically or clinically significant.

Materials and Methods

Sample Selection for Serum Data

Samples were collected from two PD study cohorts in the UK, including the Oxford Discovery Cohort and the Proband Cohort. *(78)* We selected a total of 227 PD and control subjects from Oxford Discovery and Proband cohorts in the UK (**Table 3.1**).

Definition of PD endophenotypes and severity

We characterized PD endophenotypes according to a combined, standardized sum of cognitive (Montreal Cognitive Assessment: MOCA) and motor (Unified Parkinson's disease rating scale: UPDRS3). *(79, 80)* This measure captures two dimensions of Parkinson's Disease, cognition and motor disease severity. Scores were calculated using the following procedure. First, for the entire PD population that was available from each cohort, we standardized their MOCA and UPDRS3 scores and summed them together. We then selected subjects based on the z-score distribution of scores: (i) **Good PD**: subjects with relatively normal cognitive/motor scores (top 15%, n = 66); (ii) **Poor PD:** subjects

with poor cognitive/motor scores (bottom 10%, n = 66); (iii) **Middle PD**: subjects with intermediate cognitive/motor scores (15-90%, n = 29; note that these subjects were selected from Oxford Discovery only). One of the Good PD subjects was excluded from analysis as he was found to be diagnosed with dystonic tremor instead of PD. We also selected a control group from the Oxford Discovery cohort (n = 66). Proband did not have a control arm. All groups were gender matched and age-matched within 2.5 years.

There are limitations and downsides to how we defined the good, poor and middle PD groups. In particular, the middle PD group's UPDRS3 scores have very high variance and actually the mean UPDRS3 score of the middle group is worse than the Poor PD. The study team defined the subgroups as the sum of standardized MOCA and UPDRS3 scores. Perhaps a better way to define PD subgroups is to examine motor dimensions and cognitive dimensions separately, forming a good, middle and poor group based on UPDRS3 and MOCA scores individually.

Table 3.1: Sample characteristics for the 227 subjects selected across Oxford Discovery and Proband cohorts

		PD			
	Controls	Good PD	Middle PD	Poor PD	Significance d
N	66	66	29	66	
Age	70.6 (± 6.7)	70.4 (± 6.4)	68.8 (± 6.2)	71.0 (± 6.4)	0.53
Sex, Count(% Male)	46 (70%)	47 (71%)	21 (75%)	47 (71%)	0.97
MOCA[a]	25.7 (± 3.1)	28.1 (± 1.2)	24.7 (± 2.9)	19.0 (± 4.0)	<0.001
UPDRS 3[b]	--	12.2 (± 4.8)	23.4 (± 7.4)	19.0 (± 3.8)	<0.001
LED[C]	0.0	290 (± 210)	279 (± 205)	320.5 (± 182.2)	0.46
Disease Dur (yrs)	--	1.3 (± 0.9)	1.2 (± 0.9)	1.5 (± 0.9)	0.21

a MOCA: Montnial Cognitive Assessment, Normal: 30 - 26, MCI: 25 - 21, Dementia:< 20.
b: UPDRS3: Unified Parkinson', Disease Rating Scale, part 3, ,cores 0(best] - 56 (worst) c: Levodopa equivalent dose: LED
d: Group differences determined using an ANOVA. Significance calculated for Good vs Poor PD endophenotypes

Sample Selection for Brain data

We selected the post-mortem brain tissue of 24 PD subjects and 13 controls obtained from Parkinson's UK Brain Bank. Prior to extraction, the brains were homogenized in liquid nitrogen and the brain powder was stored at −80 °C. Brain powder was homogenized with 1 mL of lysis buffer. In order to prevent protein degradation, protease inhibitor cocktail (10 μL) was added during the sample preparation.

We analyzed the anterior cingulate cortex of the brain, as this area of the brain would be affected by alpha synucleinopathy but not as diseased as the substantia nigra in the late stage of PD. These individuals were matched for age, gender, and disease duration as well as other indicators of tissue quality including post-mortem interval, and RNA-integrity (see **Table 3.2**). This sample selection and sample preparation was conducted by Jess Ash, a postdoc in my supervisor's lab. I was only responsible for the analysis and did not have any role in the sample preparation or selection.

Table 3.2: Sample characteristics for the 37 subjects selected from Parkinson's UK Brain Bank

	PD (N = 24)	Controls (N =13)	Sig.
Sample Size (by Gender)	18 M, 6 F	11 M, 2 F	
Mean Age in yrs (sd)	78.2 (11.4)	79.2 (6.0)	0.75
Disease Dur in yrs (sd)	8 (2.3)	0	0.18
Postmortem Interval in hrs (sd) (< 24 hrs ideal)	17.5 (5.2)	17.2 (8.2)	0.13
RNA integrity number (sd) (>7 better)	7.0 (0.7)	7.4 (1.1)	0.27

Note: T-tests were used to determine significance levels.

Description of Proteomics Platform

We used both the Somalogic aptamer assay and Proteome Sciences mass spectrometry assay *(81)* to quantify proteins in the serum of both PD and control subjects. We quantified all serum samples using both platforms. We quantified brain samples using the Somalogic, aptamer-based platform. *(21)*

Aptamer-based proteomics platform

All samples were assayed together, with operators blinded to disease status. Proteins were quantified using SOMAScan, an aptamer-based technology from SomaLogic Inc. (Boulder, CO). This proteomics platform is made possible by protein-capture SOMAmers (Slow Off-rate Modified Aptamer), chemically modified oligonucleotides with specific affinity to their protein targets, developed by *in vitro* selection (SELEX).

The specific steps of the SOMAScan assay have been previously outlined and described in detail in technical white papers at www.somalogic.com. In brief, plasma samples are incubated with the reagent mixes containing SOMAmers to 1129 different proteins to allow for equilibrium binding of fluorophore-tagged SOMAmers to their protein targets. Next, a series of partitioning and wash steps are used to capture only the SOMAmers that are bound to their cognate proteins. Finally, the protein-bound SOMAmer oligonucleotides are released from the protein complex, captured by complementarity, and quantified using DNA hybridization arrays.

To adjust for batch-to-batch variation, the hybridization arrays are normalized and calibrated using data from a reference set of pooled plasma samples run on each batch.

Thus, the normalized and calibrated signal for each SOMAmer—reported in relative fluorescence units (RFU)—reflects the relative amount of each cognate protein present in the original sample.

Pre-processing and normalization of sample data

All plasma samples in this study were assayed in two runs, along with 13 plasma calibrator samples, hybridization controls and 2 buffer (no protein) control samples. Run quality control standards were derived from metrics obtained during assay development, and pre-processing and normalization methods are described in detail in a technical white paper. In brief, sample data was first normalized to eliminate hybridization artifacts, using "spiked in" hybridization controls. Median normalization was subsequently applied for each sample to remove other potential assay biases within the run. For the SOMAScan assay, the hybridization control and median scale factors are expected to be in the range of 0.4-2.5 (±1.32 on $\log_2$ scale). 2 PD samples had median normalization scale factors outside the acceptable range. These sample was eliminated from our analysis.

MS-Based Proteomics Platform

Proteome Sciences, PLC (Surrey, UK), uses a gel-free, triple-stage mass-spectrometry workflow coupled with Tandem Mass Tags (TMTs) for quantitation of the relative abundance of peptide fragments that when compared with database information can be used to identify proteins. This technique produces results for highly-abundant proteins and identifies protein isoforms. For Phase 1, 6,887 unique peptide sequences

translated to 2,038 protein groups, with a total of 206 remaining for analysis after quality control.

Statistical Methods

Univariate Protein Analysis for PD versus Controls

Regression models were run associating the concentrations of each of the 1129 aptamer-assayed proteins and 206 MS assayed proteins with group status (PD versus controls) adjusting for the covariates, age at the serum visit, gender, medication (Levodopa equivalent daily dose: LEDD), and cohort (Discovery versus Proband), Proteins were reported if they had an associated p-value at $p<.05$

Univariate Protein Analysis for PD endophenotype analysis

Regression models were run associating the concentrations of each of the 1129 aptamer-assayed proteins and 206 MS assayed proteins with PD endophenotype group status (Good versus Poor PD) adjusting for the covariates, age at the serum visit, gender, and medication (Levodopa equivalent daily dose: LEDD). Proteins were reported if they had an associated p-value at $p<.05$.

Construction and Evaluation of Diagnostic Classifiers

To obtain a multivariate signature, we first used an Empirical Bayesian approach to adjust the data for covariates including Age, Gender, LEDD, and Study group. A feature selection approach based upon minimizing out of bag error during RF classification training was used to identify the most important protein markers that were significantly associated with group status (PD versus Controls). Subsequently, we trained

and tested RF classifiers using these proteins and empirically determined the features that resulted in the highest cross-validated performance in differentiating PD from Controls using from n=1 to n=87 of the proteins that passed our multivariate threshold. Final models were analyzed for Oxford Discovery data only given it was the only cohort with a control arm. Classification results were reported from 1000 repeats of 5-fold cross validation.

Construction and Evaluation of Endophenotypic Classifiers

The same methods were employed for the training of an endophenotypic classifier. The only difference is that we used protein predictors to predict for PD endophenotype status (e.g,. Good versus Poor) rather than PD versus Control.

Validation of Serum Classifier in Brain

We wanted to determine whether the 22 proteins that made up the serum classifier also had diagnostic power in the brain. 21 out of the 22 proteins were found in the brain-based Somalogic assay.

After adjusting the brain expression data by controlling for Age, Gender, and LEDD values, a machine learning classifier based off of self-organizing maps using the 21 proteins nominated from the serum data was trained and crossvalidated on the brain expression data. Classification performance was assessed using 1000 repeats of 5-fold crossvalidation.

Monte Carlo Methods to determine significance of our classification performance

We wanted to determine the probability of achieving the classification performance that the 21-protein classifier achieved in brain by chance. As such, we employed a Monte Carlo, permutation-based approach. We generated 1000 random sets of 21 proteins. We then assessed the performance of each of these classifiers on the brain expression data. Only 6 out of the 1000 randomly generated sets of 21 proteins performed with an AUC > 0.83, suggesting that the significance of the performance of our serum nominated proteins set is p~.006. This relatively high performance suggests that the brain expression data validates our serum biomarkers.

Biological Pathway Analysis

We evaluated significant proteins for biological pathway enrichment using the DAVID Bioinformatics Resource, version 6.7, Functional Annotation tool, performed on the PANTHER database. The significant proteins were inputted as our 'gene list' while all 1129 proteins quantified in the study were inputted as our 'background gene list', for purposes of assigning probabilities to the distribution of proteins observed in our candidate PD biomarker list versus those expected under a random draw of the number of significant proteins from the full 1129 protein set.

Results

44 candidate serum proteins differentiate PD from normal controls using the aptamer-based platform and 26 proteins differentiate PD from normal controls using the MS assay.

To identify proteins that had significantly different concentrations in PD versus normal controls samples, models were run associating the concentrations of each of the 1129 aptamer-assayed proteins and 206 MS-assayed proteins with group status (PD versus controls) adjusting for the covariates, age at the serum visit, gender, medication (Levodopa equivalent daily dose: LEDD), and cohort (Discovery versus Proband),

A total of 44 proteins that were significantly differentiated between PD cases and controls in the aptamer platform (Controls: n = 65; PD: n = 159; $p \leq 0.05$; **Table 3.3**). Prostate-specific antigen complexed to a1-antichymotrypsin showed the strongest association with PD. These p-values are uncorrected so out of 1129 proteins we ran tests on, we expect ~56 proteins to pass the significant threshold just by chance. This suggests these results are not significant statistically.

Table 3.3: Univariate proteins that differentiated PD cases versus controls using aptamer-based assay (No. of proteins = 44, $P \leq 0.05$, uncorrected p-values)

Protein	Group Coefficient	Group Significance
PSA-ACT	0.079	0.00007
Proteinase-3	0.108	0.004
M-CSF R	-0.066	0.008
ALCAM	-0.036	0.011
P-Selectin	-0.058	0.011
IL-4	-0.054	0.012
kallikrein 8	-0.052	0.015
CONAl	0.046	0.016
Alkaline phosphatase, bone	0.096	0.016
PH	-0.123	0.016
C036 ANTIGEN	-0.066	0.018
Aminoacylase-1	-0.093	0.019
al-Antitrypsin	0.059	0.019
Haptoglobin, Mixed Type	0.174	0.019
IL-7 Ra	0.052	0.02
Growth hormone receptor	-0.055	0.023
5CF sR	-0.053	0.023
NCAM-Ll	-0.047	0.024
IMOH2	0.026	0.025
PCNA	0.085	0.03
Persephin	-0.041	0.032
BPI	0.125	0.032
NCC27	0.05	0.033
IL-6 sRa	-0.046	0.034
C0109	-0.047	0.034
Elafin	0.069	0.034
IL-1 R4	-0.069	0.035
SPTA2	0.033	0.035
SGTA	0.07	0.036
IL-27	0.053	0.036
6-Phosphogluconate dehydrogenase	0.122	0.037
Lactoferrin	0.097	0.038
Flt3 ligand	0.04	0.041
EphAS	0.041	0.042

BSTl	-0.09	0.043
Cyclin Bl	0.051	0.044
sE-Selectin	-0.057	0.045
C9	0.032	0.045
TPSB2	-0.09	0.046
al-Antichymotrypsin	0.029	0.046
ALT	-0.032	0.047
Karyopherin-a2	0.031	0.049
Transferrin	0.031	0.05
CYTD	-0.095	0.05

In the MS analysis of the 206 proteins, a total of 26 proteins that were significantly different between PD cases and controls (Controls: n = 65; PD: n = 126; $p \leq 0.05$; **Table 3.4**). Similar to the SomaLogic findings, Alpha-1 antitrypsin and Alpha-1 antichromotrypsin were significantly unregulated in PD relative to controls. These p-values are uncorrected so out of 206 proteins we ran tests on, we expect ~10 proteins to pass the significant threshold just by chance. This suggests these results are moderately statistically significant, as we found 26 significant proteins.

Table 3.4: Univariate proteins that differentiated PD cases and controls using mass spec assay (No. of proteins = 26, $P \leq 0.05$, uncorrected p-values)

Gene	Protein Name	Group coefficient	Group Significance
CRP	C-reactive protein	1.009	0.001
C4BPA	C4b-bindingprotein alpha chain	0.139	0.001
FGA	Fibrinogenalpha chain	0.356	0.002
LTF	Lactotransferrin	0.408	0.002
W.JF	von Willebrand factor	0.305	0.005
C4BPB	lsoform 2 of C4b-binding protein beta chain	0.146	0.007
ACTB	Act in, cytoplasmic 1	0.244	0.008
AlBG	Alpha-1B-glycoprotein	0.100	0.009
LRGl	Leucine-rich alpha-2-glycoprotein	0.184	0.016
HP	Haptoglobin	0.304	0.018
AHSG	Alpha-2-HS-glycoprotein	-0.117	0.023
SERPINA3	Alpha-1-a ntichymotrypsi n	0.111	0.027
SERPINAl	Alpha-1-antitrypsin	0.107	0.028
SEPP1	Selenoprotein P	-0.105	0.029
APOC2	Apolipoprotein C-2	-0.202	0.031
PROS1	Vitamin K-dependent protein S	0.094	0.033
SAAl	Serum amyloid A-1 protein	0.666	0.033
APOA2	Apolipoprotein A-2	-0.110	0.034
FS	Coagulation factorV	-0.098	0.034
FNl	lsoform 14 of Fibronectin	0.245	0.034
CA2	Carbonic anhydrase2	0.308	0.038
PLTP	Phospholipid transfer protein	-0.118	0.039
COMP	Cartilage oligomericmatrixprotein	-0.165	0.041
PONl	Serum paraoxonase/arylesterase1	-0.152	0.042
F12	CoagulationfactorXII	0.1S6	0.044
FETUB	Fetuin-B	0.131	0.046

Training a multivariate classifier to predict PD samples from Controls

Next, we wanted to determine whether a combination of multiple markers could predict PD samples from controls with high accuracy. To obtain a multivariate signature, we first used an Empirical Bayesian method to adjust the data for covariates including Age, Gender, LEDD, and Study. A feature selection approach based upon minimizing out of bag error during RF classification training was used to identify the most important protein markers that were significantly associated with group status (PD versus Controls). Subsequently, we trained and tested random forest classifiers using these proteins and empirically determined the features that resulted in the highest crossvalidated performance in differentiating PD from Controls. Final models were analyzed for Oxford Discovery data only given it was the only cohort with a control arm.

For the aptamer-based assay, 87 proteins initially passed our thresholding step for multivariate significance. A final 22-protein classifier differentiated PD from Controls with an area under the curve (AUC) of 0.91 (for protein list, see **Table 3.5**). For the MS-based assay, there were a total number of 34 proteins that initially passed our thresholding step for multivariate significance with a final 24-protein classifier obtaining a performance of AUC = 0.82 (for protein list, see **Table 3.6**). The ROC comparing both SomaLogic and Proteome Sciences is shown in **Fig. 3.1** below.

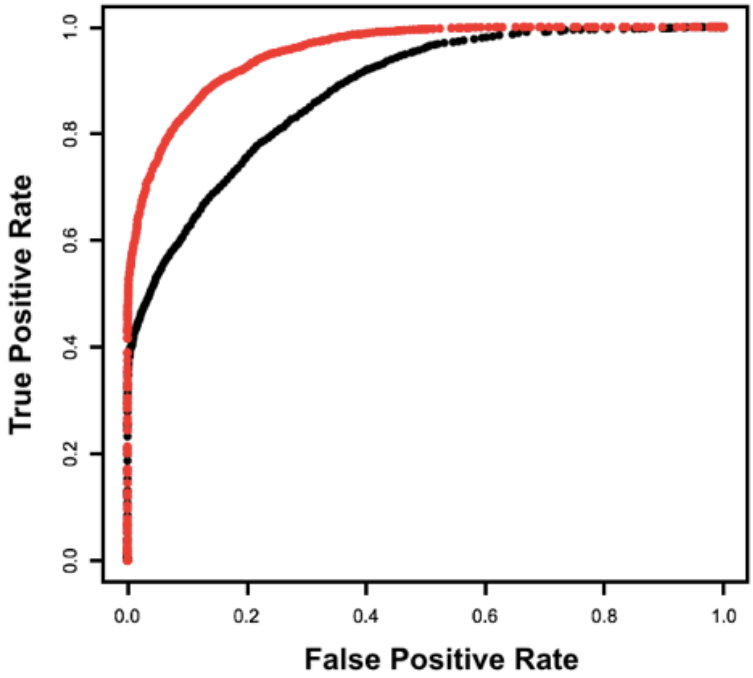
Table 3.5: Proteins that make up multivariate classifier for predicting PD cases from controls for the aptamer assay.

Gene Name	Protein Name
PSA-ACT	PSA:alpha-1-antichymotrypsin complex
CYTD	Cystatin-D
PI3	Elafin
C3A	C3a anaphylatoxin
CCL28	CC motif chemokine 28
DDX19B	ATP-dependent RNA helicase DDX19B
PROC	Activated Protein C
PRDX1	Peroxiredoxin-1
SPTAN1	Spectrin alpha chain, non-erythrocytic 1
TNFRSF13C	Tumor necrosis factor receptor 13C
TYMS	Thymidylate synthase
KLK5	Kallikrein-5
DDC	Aromatic-l-amino-acid decarboxylase
CST2	Cystatin-SA
CD36	Platelet glycoprotein 4
AT3	Antithrombin-111
ACP5	Tartrate-resistant acid phosphatase type 5
DMP1	Dentin matrix acidic phosphoprotein1
AIF1	Allograft inflammatory factor 1
PH	Pancreatic hormone
HDGFL2	Pancreatic Hepatoma-derived growth factor-related protein 2
HSPA8	Heat shock cognate 71 kDa protein

Table 3.6: Proteins that make up the multivariate classifier predicting PD cases for the MS assay

Gene	Protein Name
FGA	Fibrinogen alpha chain
FBLN1	Fibulin-1
CA2	Carbonic anhydrase 2
APOC2	Apolipoprotein C-2
LRG1	Leucine-rich alpha-2-glycoprotein
APOB	Apolipoprotein B
CHl1	Neural cell adhesion molecule L1-like protein
ACTB	Actin cytoplasmic 1
TGFBI	Transforming Growth factor-beta-induced protein ig-h3
C8B	Complement component C8 beta chain
BTD	Biotinidase
LCP1	Complement Clr subcomponent-like protein
LCP1	Plastin-2
CFHR2	Complement factor H-related protein 2
SRGN	Serglycin
A2M	Alpha-2-macroglobulin
KLKB1	Plasma kallikrein
AFM	Afamin
SELENOP	Selenoprotein P
VTN	Vitronectin
LTF	isoform Delta F of Lactotransferrin
THBS1	thrombospondin-1
APMAP	Adipocyte plasma membrane-associated protein
IGLC7	lambda-7 chain C region

Fig. 3.1: ROC Curve comparing Somalogic (red, AUC = .91) and Proteome Sciences (black, AUC = .82) for PD versus controls



The ROC curve from the Somalogic derived classifier (red trace) is compared with the ROC curve from the Proteome Sciences derived classifier (black trace). We found that the Somalogic derived classifier performed with higher classification performance in predicting PD from control samples.

63 candidate serum proteins differentiate Good PD and Poor PD samples using the aptamer platform and 20 proteins differentiate Good PD and Poor PD controls using the MS assay.

To identify proteins that had significantly different concentrations in the good PD endophenotype group versus the poor PD endophenotype group, regression models were run associating the concentrations of each of the 1129 aptamer-assayed proteins and 206 MS-assayed proteins with group status (Good PD versus Poor PD) adjusting for the covariates, age at the serum visit, gender, medication (Levodopa equivalent daily dose: LEDD), and cohort.

Results from the aptamer assay revealed a total of 63 significant proteins across studies (Good PD: n = 65; Poor PD: n = 65; $p \leq 0.05$; **Table 3.7**). Five of these proteins also distinguished PD from controls including Mast/stem cell growth factor receptor Kit, Prostate-specific antigen complexed to a1-antichymotrypsin, Growth Hormone Receptor, P-selectin, Interleukin-6 receptor subunit alpha.

Table 3.7: Univariate proteins that differentiated Good PD versus Poor PD across studies using aptamer-based assay

Gene	Protein Name	Group Coefficient	Group Significance
CNDPl	Beta-Ala-His dipeptidase	-0.075	0.0007
ERBBl	Epidermal growth factor receptor	-0.046	0.001
CD86	T-lymphocyte activation antigen CD86	-0.047	0.002
GZMH	Granzyme H	0.142	0.002
ER	Estrogenreceptor	0.031	0.003
SOD2	Superoxide dismutase [Mn], mitochondrial	-0.052	0.004
TG	Thyroglobulin	0.079	0.005
GPC3	Glypican-3	-0.064	0.005
AHSG	Alpha-2-HS-glycoprotein	-0.034	0.006
FGF20	Fibroblast growth factor 20	-0.05	0.006
SCFR	Mast/stem cell growth factor receptor	-0.062	0.006
TRAF	Tumor necrosis factor receptor superfamllymember6B	-0.062	0.008
BGN	Biglycan	-0.054	0.009
MIS	Muelerian-inhibiting factor	0.053	0.012
PSA-ACT	PSA:alpha-1-antichymotrypsin complex	0.051	0.012
CD200R1	Cellsurface glycoprotein CD200 receptor 1	-0.035	0.012
CAMK2B	Calcium/calmodulin-dependent protein kinase type II subunit beta	0.094	0.013
RETN	Resistin	0.046	0.014
KLKB1	Plasma kallikrein	-0.029	0.015
IL12RB2	interleukin-12 receptor subunit beta-2	-0.054	0.015
ENG	Endoglin	-0.048	0.017
SELP	P-Selectin	-0.051	0.018
CDKN2C	Cyclin-dependent kinase inhibitor 18	-0.134	0.018
GHR	Growth honmone receptor	-0.051	0.018
VEGFR2	Vascular Endothelial Growth factor receptor 2	-0.024	0.019
ANXA1	AnnexinAl	0.079	0.02
FGF1	Fibroblast growth factor 1	0.041	0.021

IL17B	lnterleukin-17B	0.024	0.022
IL6R	interleukin-6 receptor subunit alpha Fibroblast growth factor 1	-0.045	0.023
NCAM2	Neural cell adhesion molecule ll	-0.048	0.024
UFM1	Ubiquitin-fold modifiers	0.07	0.025
GSTP1	Glutathione S-transferase P	0.022	0.026
LYZ	Lysozyme C	0.057	0.027
DDX19B	ATP-dependent RNA helicase DDX19B	-0.08	0.027
KIF23	Kinesin-like protein KIF23	-0.047	0.028
NANOG	Homeobox protein NANOG	0.065	0.029
GREM1	Gremlin-1	0.072	0.029
TYRO3	Tyrosine-protein kinase receptor TYR03	-0.023	0.029
AIF1	Allograft inflammatory factor 1	0.032	0.03
EPO	Erythropoietin	0.071	0.03
MEK1	Dual specificity mitogen-activated protein kinase	0.061	0.03
MAP2K1	Inosine-5'-monophosphate dehydrogenase 1	0.109	0.033
CYST3	Cystatin-C	0.033	0.034
TFF3	Trefoil factor 3	0.05	0.034
SET	Protein SET	-0.029	0.035
TNFSF14	Tumor Necrosis factor ligand superfamily member 14	0.069	0.036
EPS15L1	Epidermal growth factor receptor substrate 15-like 1	0.046	0.037
HNRNPA2B1	Heterogeneous Nuclear Ribonucleoproteins A2/81	0.089	0.037
IL2RA	interleukin-2 receptor subunit alpha	-0.037	0.038
IL18RAP	lnterleukin-18receptor accessory protein	-0.026	0.038
NOTCH1	Neurogenic locus notch homolog protein 1	-0.025	0.039
RPSA	40S ribosomal protein SA	0.025	0.04
HMGR	3-hydroxy-3-methylglutaryl-coenzyme reductase	-0.052	0.043
XRCC6	X-ray repair cross-complementing protein 6	0.097	0.044
HG	Hemoglobin	-0.154	0.044
IL1RAP	Interleukin-1 Receptor accessory protein	-0.058	0.044
Efnb3	Ephrin-B3	0.045	0.045
DLRB1	Dylein Light Chain roadblock-type 1	0.047	0.048
CLF1/CLC Complex	Cytokine receptor-like factor 1:Cardiotrophin-like cytokine factor 1 Complex	0.062	0.048

CDH5	Cadherin-5	-0.033	0.049
ERBB3	Receptor tyrosine-protein kinase erbB-3	-0.053	0.049
LRRTM1	Leucine-rich repeat transmembrane neuronal protein 1	0.054	0.049
RBM39	RNA-binding protein 39	0.04	0.049

Using the MS-assay, we identified a total of 20 significant proteins (Good PD: n = 62; Poor PD: n = 64; $p \leq 0.05$; **Table 3.8**) that were differentially expressed across good PD and poor PD groups. Eight of these proteins also distinguished PD from controls including Apolipoprotein C-II, Haptoglobin, Alpha-1-antichymotrypsin, Apolipoprotein A-II, Alpha-1B- glycoprotein, Serum paraoxonase/arylesterase 1, Selenoprotein P, Actin, cytoplasmic 1.

Table 3.8: 20 proteins were differentially expressed between Good PD versus Poor PD groups using the Mass Spec assay

Gene	Protein	Group coefficient	Group Significance
APOC2	Apolipoprotein C-2	-0.255	0.002
CNDP1	Beta-Ala-His dipeptidase	-0.242	0.002
ECMl	Extracellular matrix protein 1	0.225	0.002
HP	Haptoglobin	340	0.004
ALB	Serum albumin	-0.343	0.004
ORM1	Alpha-18-glycoprotein	0.089	0.005
SELENOP	Selenoprotein P	-0.118	0.007
ITIH3	lsoform 2ofInter-alpha-trypsin inhibitor heavy chain	0.17	0.008
PONl	Serum paraoxonase/arylesterase 1	-0.168	0.009
LCAT	Phosphatidylcholine-sterol acyltransferase	-0.082	0.018
AGT	Angiotensinogen	-0.096	0.022
SERPINA3	Alpha -1-antichymotrypsin	0.103	0.029
ig lambda chain V-I 1 I region LOI	ig lambda chain V-I 1 I region LOI	0.28	0.031
SERPINA4	Kallistatin	-0.083	0.033
APOA2	Apolipoprotein A-II	-0.092	0.036
EFEMPl	lsoform 2of EGF-containing fibulin-like extracellular	0.102	0.039
SERPINAS	Plasma serine protease inhibitor	-0.088	0.041
C9	Complement component C9	0.112	0.042
CAMP	Cathelicidin antimicrobial peptide	-0.168	0.042
ACTB	Actin, cytoplasmic 1	-0.159	0.043

Training a multivariate classifier to predict Good PD from Poor PD

We used the same multivariate workflow to train a classifier that predicted Good and Poor PD samples. For the aptamer-based assay, a total number of 78 proteins initially passed our thresholding step for multivariate significance and a final 24-protein classifier achieved an AUC of 0.89 (for protein list, see **Table 3.9**). For the MS assay, 31 proteins passed our thresholding step for multivariate significance and a final 14-protein classifier was trained that achieved an AUC of 0.83 (for protein list, see **Table 3.10**). The ROC is shown in **Fig. 3.2** below comparing classifiers for the aptamer-based SomaLogic and MS-based Proteome Sciences assay.

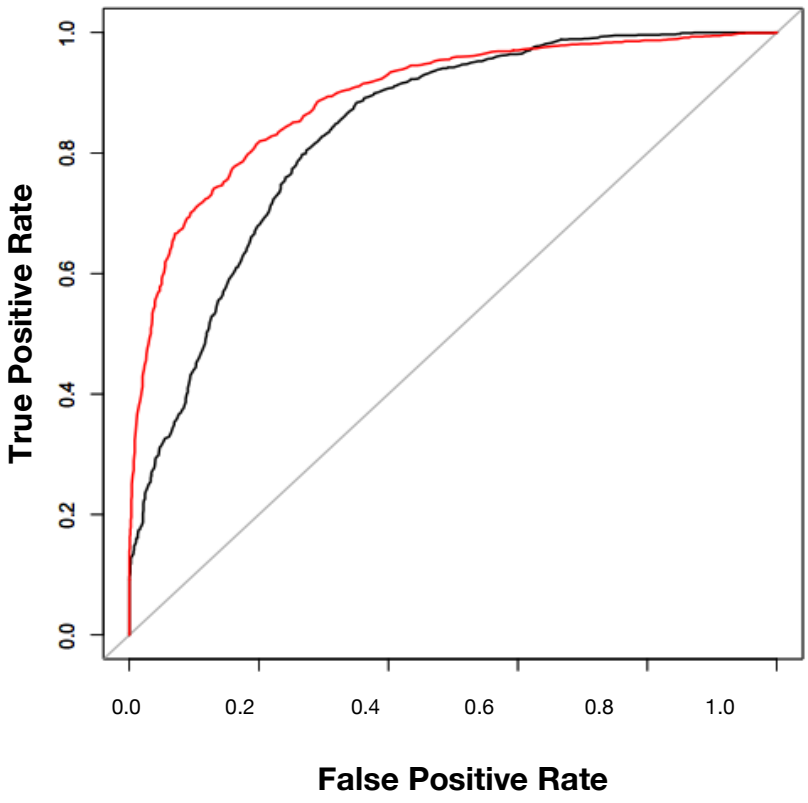
Table 3.9: Proteins that make up multivariate classifier that predicts Good versus Poor PD in the aptamer assay

Gene	Protein Name
GZMH	Granzyme H
ERBB3	Receptor tyrosine-protein kinase erbB-3
CDKN1B	Cyclin-dependent kinase inhibitor 1B
TG	Thyroglobulin
LRRTM1	Leucine-rich repeat transmembrane neuronal protein 1
LTA4H	Leukotriene A-4 hydrolase
BGN	Biglycan
SOD2	Superoxide dismutase [Mn], mitochondrial
ACVR1B	Activin receptor type-1B
CD86	T-lymphocyte activation antigen CD86
FGF20	Fibroblast growth factor 20
TIE1	Tyrosine-protein kinase receptor Tie-1
TNFRSF6B	Tumor necrosis factor receptor superfamily member 6B
SELL	L-Selectin
PA2G4	Proliferation-associated protein 2G4
CDKl/cyclin B	Cyclin-dependent kinase 1:G2/mitotic-specific cyclin-Bl complex
CLF-1/CLCComplex	Cytokine receptor-like factor 1:Cardiotrophin-like cytokine factor 1Complex
IL3	interleukin-3
C7	Complement component C7
FER	Tyrosine-protein kinase Fer
CCL15	C-C motif chemokine 15
ULBP2	NKG20 ligand 2
FGF1	Fibroblast growth factor l
OCIAD1	OCIA domain-containing protein1

Table 3.10: Proteins making up multivariate classifier predicting for Good versus Poor PD in MS assay

Gene	Protein Name
IG	lg lambda chain-111
ECMl	Extracellular matrix protein 1
APOC2	Apolipoprotein C-2
SERPINDl	Heparin cofactor 2
EFEMP1	lsoform 2 of EGF-containing fibulin-like extracellular matrix protein 1
APOA2	Apolipoprotein A-2
HP	Haptoglobin
CP	Ceruloplasmin
PON1	Serum paraoxonase/arylesterase 1
SELENOP	Selenoprotein P
APOL1	lsoform 3 of Apolipoprotein L-1
LYVE1	Lymphatic vessel endothelial hyaluronic acid receptor 1
C9	Complement component C9
IGHM	lg mu heavy chain disease protein

Fig. 3.2: Discovery Phase 1 ROC Curve comparing SomaLogic (red, AUC = .89) and Proteome Sciences (black, AUC = .83) for Good PD versus Poor PD



The ROC curve from the Somalogic derived classifier (red trace) is compared with ROC curve from the Proteome Sciences derived classifier (black trace). We found that the Somalogic derived classifier performed with higher classification performance in predicting Good PD vs Poor PD samples.

141 proteins were differentially expressed in the brain of PD patients versus controls

Somalogic increased the number of proteins quantified in its assay from 1129 (which was used in our serum quantifications) to 1,310, which was used to quantify brain samples. Linear regression models associating each protein with PD versus control status were run, adjusting for age at death, gender, processing date, and medication. 141 proteins were found to be significantly different in the brain tissue of late-stage PD versus controls (PD: n = 24, Control: n = 13; $p \leq 0.05$).

Pathway analysis of differentially expressed proteins include

We wanted to see whether the differentially expressed proteins were enriched in any particular pathways. KEGG pathway analysis of these proteins showed an enrichment in the Toll-like receptor signaling pathway, which has been previously linked to Parkinson's Disease

Table 3.11: KEGG pathway analysis of 141 differentially expressed proteins

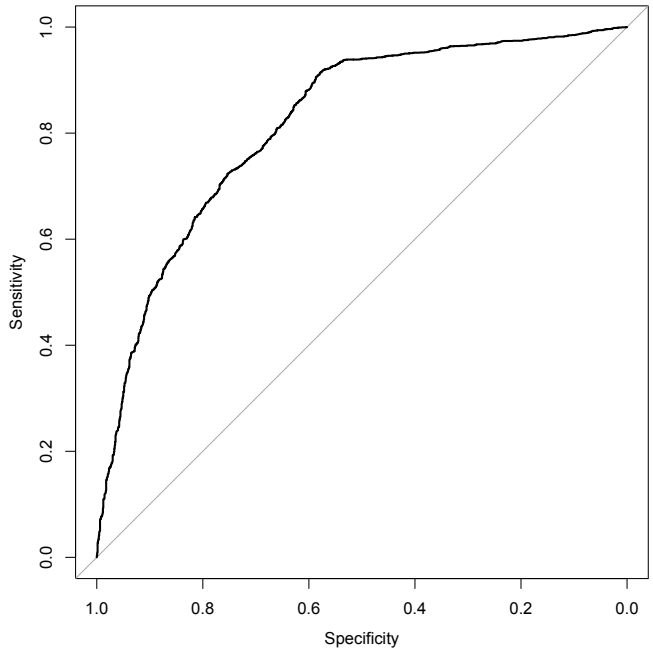
Pathway	PValue	Protein (Uniprot IDs)
Toll-like receptor signaling pathway	0.031525	Q9Y243, P42224, P31751, Q9Y6Y9, P45985, O00206, Q16539, O60603, P36507, O15264, P31749
Tight junction	0.073351	P19784, Q9Y243, P31751, P67870, P11171, P31749
Fc epsilon RI signaling pathway	0.083243	Q9Y243, P14555, P04141, P31751, P45985, Q16539, P36507, O15264, P31749
Progesterone-mediated oocyte maturation	0.092142	Q9Y243, P07900, P08238, P31751, Q16539, O15264, P31749

Validating our Serum Markers in Brain:

We wanted to use the brain expression data in order to help validate whether the proteins that make up the 22-protein classifier in serum has predictive power in brain. 21 out of the 22 proteins that made up the serum classifier are also found in the newer 1310-plex version of the assay.

After adjusting the brain expression data by controlling for Age, Gender, and LEDD values, a machine learning classifier based on self-organizing maps using the 21 proteins nominated from the serum data was trained and crossvalidated on the brain expression data. The classifier performed with 89% Sensitivity, 61% Specificity, AUC 0.83 (**Fig. 3.3**).

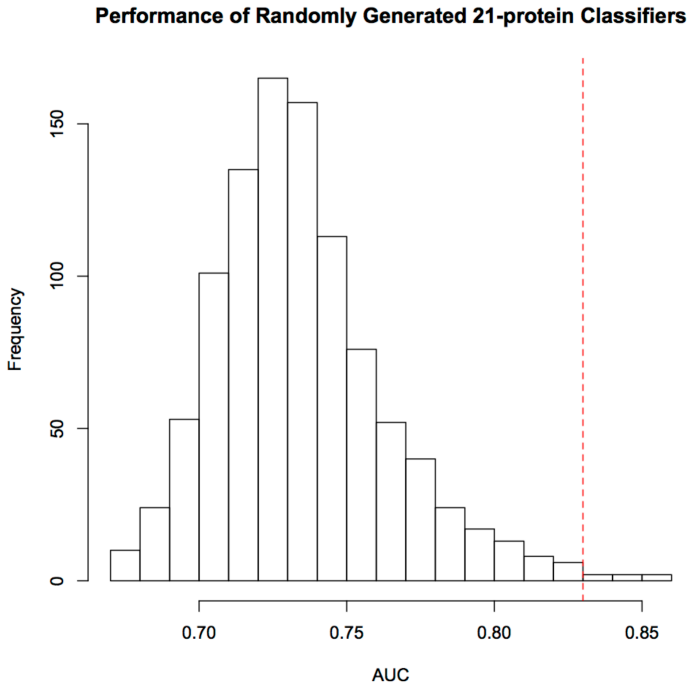
Fig. 3.3: ROC Curve of 21-protein classifier on brain expression data



A 21-protein classifier was tested on the brain expression data. The classifier performed with 89% Sensitivity, 61% Specificity, AUC = 0.83.

We wanted to determine the significance of achieving a performance at this level. As such, we employed a permutation-based approach where we generated 1000 random sets of 21 proteins. We then assessed the performance of each of these classifiers on the brain expression data. Only 6 out of the 1000 randomly generated sets of 21 proteins performed with an AUC > 0.83 (**Fig 3.4**), suggesting that the significance of the performance of our serum nominated proteins set is p~.006. This relatively high performance suggests that the brain expression data validates our serum biomarkers.

Fig. 3.4: Distribution of AUC of the 1000 randomly generated 21-protein classifiers. Red line at AUC=0.83.

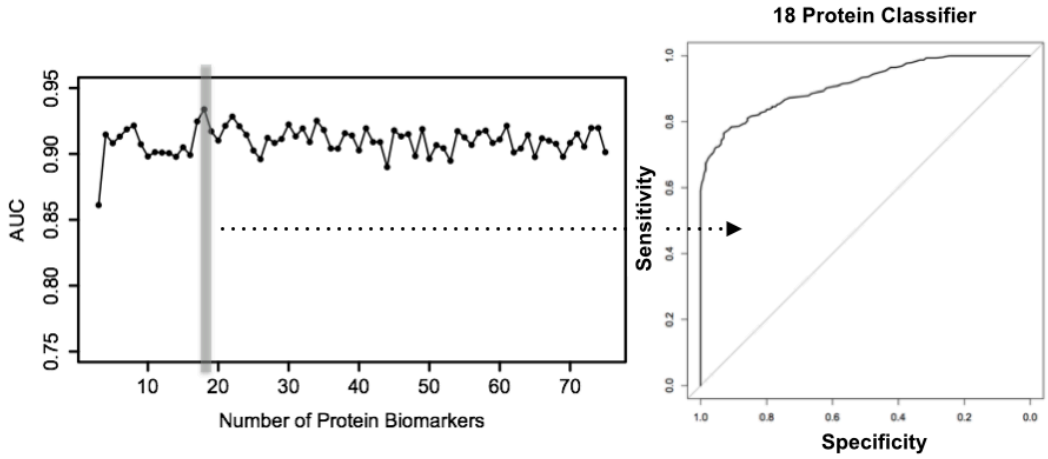


In order to test the significance of the 21-protein classifier, we used a permutation-based approach. We generated 1000 random sets of 21 proteins and trained classifiers using these 1000 set of proteins. Only 6 of the 1000 randomly generated set of 21 protein classifiers achieved a performance greater than 0.83. This means the significance of our performance is p ~ .006.

Using our Brain Expression Data to nominate biomarkers in Serum

We wanted to see whether the reverse was true, that is, if the PD signature found in brain was found in plasma. To do so, we first identified multivariate protein signatures that differentiated PD brains from Normal brains. As done previously, we employed a feature selection approach based upon minimizing prediction error to rank the most important protein markers that were significantly associated with group status (PD versus Controls). RF classifiers were then built using the top ranked protein markers and we then found that an 18-protein biomarker set resulted in the highest performance. An 18-protein random forest classifier was then trained and crossvalidated on the brain expression data. This classifier performed with AUC > .92.

Fig. 3.5: ROC Curve of 18-protein classifier



We found that an 18-protein classifier achieved the highest performance in predicting PD vs control in brain (AUC > 0.92).

We then wanted to test the Brain expression classifier in serum. We then tested this classifier on the serum expression data, which performed no better than chance AUC~.55.

Ten proteins were found to be differentially expressed between PD and controls in both brain and serum

Finally, we wanted to access the multivariate markers that overlapped between in brain and serum. We observed that 10 proteins were differentially expressed between PD and control samples in both the brain and serum in the same relative direction (**Table 3.12)**. A broad summary of this study is summarized in **Table 3.13**.

Table 3.12: Intersection Between Serum and Brain Biomarkers nominated by multivariate methods

SomaId	TargetFullName	Target	UniProt
SL004458	Elafin	Elafin	P19957
SL008190	Spectrin alpha chain, non-erythrocytic 1	SPTA2	Q13813
SL000272	Antithrombin-III	Antithrombin III	P01008
SL000055	Cadherin-1	Cadherin E	P12830
SL004782	Tumor necrosis factor-inducible gene 6 protein	TSG-6	P98066
SL010456	Cystatin-SN	CYTN	P01037
SL014009	Hepatoma-derived growth factor-related protein 2	HDGR2	Q7Z4V5
SL005228	NKG2D ligand 2	ULBP-2	Q9BZM5
SL000541	Plasminogen	Plasminogen	P00747
SL008099	Macrophage-capping protein	CAPG	P40121

Table 3.13: Summary of Results

Outcome	SomaLogic (Aptamer-based)	Proteome Sciences (Mass spectrometry)
Serum: PD versus Controls		
No. of sig. proteins (Univariate - Regression)	44	26
No. of sig. proteins (Multivariate-Threshold)	87	34
Multivariate Classifier (Random Forest)	22 protein classifier AUC= 0.91, SD = 0.04	24 protein classifier AUC= 0.82, SD = 0.05
Serum: Good PD versus Poor PD		
No. of sig. proteins (Univariate - Regression)	63	20
No. of sig. proteins (Multivariate-Threshold)	78	31
Multivariate Classifier (Random Forest)	24 protein classifier AUC= 0.89, SD = 0.06	14 protein classifier AUC= 0.83, SD = 0.08
Post-mortem Brain Tissue: PD versus Controls		
No. of sig. proteins (Univariate - Regression)	141	NA
No. of sig. proteins (Multivariate-Threshold)	126	NA
Multivariate Classifier (Random Forest)	18	NA

Discussion

Currently, no biochemical tests for confirmation of PD exist. Instead, diagnosis relies almost entirely on the physician's history and clinical exam, with an estimated accuracy of <80%. Thus, the development of a reliable, blood-based assay for confirmation of PD diagnosis would be instrumental in both the clinical care of PD patients and in subject selection for clinical trials of potential PD-modifying drugs.

Here we employ proteomics profiling of PD and control serum samples, specifically an aptamer-based screen 1129 proteins and a MS screen of 206 serum proteins, to identify diagnostic and endophenotypic biomarkers of Parkinson's disease. In our univariate analysis, we found 44 proteins that were differentially expressed between PD and Control samples and 63 proteins that were differentially expressed between good and poor PD endophenotypes using the aptamer-based assay at the uncorrected $p < .05$ significance level. We would expect ~56 proteins to be significant just by chance, when conducting this univariate test with 1129 proteins, suggesting the univariate results are not significant. Using the MS assay, we identified 26 proteins that were differentially expressed between PD and Controls and 20 proteins that were differentially expressed between poor and good PD groups at the uncorrected $p < .05$ significance level. We would expect ~10 proteins to be significant just by chance if we tested 206 proteins, suggesting that there was moderate statistical enrichment. However, when taken in aggregate, the univariate analysis *does not* nominate statistically significant biomarkers when accounting for multiple testing.

As such we also used multivariate methods, employing machine-learning to identify classifiers and validating the significance of these classifiers using robust crossvalidation. We found that a biomarker panel of 22-proteins could predict PD versus

controls with an AUC of .91. A separate biomarker panel of 24 proteins can predict PD patients with high disease burden compared to patients with low disease burden with an AUC =.89. We were able to validate the serum diagnostic protein signature using data from 18 PD and 11 Normal Control post-mortem brain samples. A classifier trained using 21 out of the 22 proteins from the serum classifier performed with an AUC =.83.

However, we found that the most dominant protein signature in the brain that distinguished PD from controls was not found in the serum. We imagine that this might be the case because the brains came from late PD patients that had an average disease duration of more than 8 years, while serum samples came from PD patients with disease onset less than 1.5 years. Therefore, this suggests that our serum nominated diagnostic signatures more accurately reflect early stage PD while brain nominated markers represent more late-stage.

In this study, we used two different proteomics platforms, an aptamer-based method and a MS-based method. The aptamer-based method identified multivariate protein biomarker signatures that were able to classifier PD vs controls and Poor vs Good PD patients with higher diagnostic accuracy. While the aptamer-based method quantified a greater number of proteins and identified a greater number of significant proteins, the results from a univariate perspective were not significant.

Further replication is needed to ensure that our results are generalizable outside of our sites and the UK. However, these results suggest that "near-proteomic" profiling of blood from PD patients may be a powerful approach both for the development of clinical tools and for gaining insight into pathophysiological mechanisms in this currently incurable disease.

Chapter 4: Exploring correlations between CSF and Blood protein concentrations

Chapter Summary

Little is known about the relationship between the concentrations of proteins in CSF and blood. We applied a highly-multiplexed, aptamer-based proteomic assay to quantify the correlations of 1124 proteins in matched CSF and serum samples of 40 subjects. 55 proteins had a significant Spearman's correlation (p-value <0.00005, Bonferonni level) across CSF and serum. In order to confirm these results, we calculated Spearman's correlations across matched CSF and plasma from an additional 18 patients from an independent cohort. 49 out of the 55 proteins replicated with significant correlations across CSF and serum ($p<.05$). In order to demonstrate the utility of using this information in biomarker discovery, we quantified to levels of these 49 proteins in the serum of 80 PD patients and 65 Normal Controls (NC). We identified Growth Hormone Receptor and PARC as novel blood biomarkers for PD as they were both significantly lower in the blood PD patients compared to controls. PD patients with lower Growth Hormone Receptor in their blood also had worse Motor symptoms as measured by UPDRS3 scores adjusting for disease duration, gender, and LEDD medication. Both these results replicated in an independent cohort of 96 PD and 45 NC subjects. We present here the first large scale characterization of protein associations across CSF and blood and demonstrate how it may guide biomarker discovery in brain-related disorders.

Introduction

Biomarkers, surrogate indicators of physiological or pathophysiological states, can be used to guide the diagnosis of diseases, evaluate risk or prognosis, and track therapeutic interventions. Because CSF is directly in contact with the extracellular space of the brain, proteins in CSF may reflect biochemical changes in the brain that are involved in disease processes. For this reason, researchers studying brain-based diseases such as AD have traditionally searched for biomarkers in CSF. *(27)* Indeed, studies have found CSF levels of tau and β-amyloid to be promising diagnostic biomarkers for AD. *(36, 59, 82)*

However, collecting CSF, which requires a lumbar puncture, is significantly more invasive than collecting whole blood, plasma, or serum, creating a substantial barrier to widespread clinical translation. As a consequence, many groups have investigated the potential of plasma-based protein biomarkers in brain diseases such as AD and PD. *(24, 81)*

Understanding the relationship between a given protein's expression in plasma or serum and CSF has many implications for biomarker discovery and, beyond that, for biomedical research and clinical medicine more generally. Previous studies have used correlation analysis to investigate the relationship between specific proteins in CSF and plasma. For example, studies have explored the correlation between amyloid beta protein 1-40 ($A\beta_{40}$), amyloid beta protein 1-42 ($A\beta_{1-42}$), apolipoprotein E (ApoE), and alpha1-antichymotrypsin in matched CSF and plasma samples from AD patients using sandwich enzyme-linked immunosorbent assays. *(83)* However, as in the examples cited, existing studies are often conducted on the scale of single proteins, limiting their usefulness to a

researcher seeking to understand the biofluid expression patterns of other *de novo* candidate markers.

Many studies have examined the relationship between gene expression across the transcriptome in the blood, brain, and CSF. These eQT/pQTL type of studies have helped to inform the relationships between different upstream and downstream processes. *(84–86)*

From the standpoint of understanding which CSF markers may be followed in the blood, and which plasma markers may reflect central nervous system (CNS) processes, a reference set of matched CSF and plasma samples characterized at the level of the proteome would be a valuable resource. In the present study, we measured the levels of 1124 proteins in matched plasma and CSF samples of 40 subjects and present one of the first attempts to examine CSF and plasma correlations at the proteomic level. We then used this information to demonstrate how it may guide biomarker discovery for brain disorder, using PD as a proof of concept.

Materials and Methods

Subjects

Matched CSF and serum samples were collected from 40 research subjects enrolled in the Oxford Discovery PD Biomarkers project and collected at the same time point. An independent validation set matched CSF and plasma samples were collected from 18 subjects as part of a Bristol Myers Squibb Study.

PD subjects

80 PD and 65 NC serum samples were collected from subjects in the Oxford Parkinson's Disease Biomarker Study by the study researchers. An independent validation cohort of 96 PD and 45 NC plasma samples were collected from subjects enrolled in Parkinson's Disease Biomarker Research Projects at the University of Pennsylvania.

CSF Collection

In brief, samples were collected through a routine lumbar puncture and placed into clear polypropylene tubes. Specific protocols for CSF collection followed those of the international Alzheimer's Disease Neuroimaging Initiative (ADNI) biorepository program (http://www.adni-info.org/Scientists/ADNIStudyProcedures.aspx). CSF was then aliquoted into 0.5mL samples in 1.5mL cryogenic tubes for storage at -80°C until use.

Protein Quantification

Proteins were quantified using the SOMAScan assay from SomaLogic Inc. (Boulder, CO). The SOMAScan assay is an aptamer-based technology that allows for the quantification of protein levels in biofluid. This technique is made possible by protein-capture SOMAmers (Slow Off-rate Modified Aptamer), which are chemically modified oligonucleotides with specific affinity to their protein targets, initially developed by *in vitro* selection (SELEX) as previously described. *(40)*

The SOMAScan assay is described in detail in technical white papers at www.somalogic.com. In brief, the biological fluid of interest (i.e. plasma or CSF) is diluted to three different concentrations, and all three dilutions are assayed, so that the least concentrated dilution can be used to detect the most abundant proteins, and the most concentrated dilutions used to detect the least abundant proteins, within the dynamic range of each protein's individual SOMAmer assay. Diluted samples are then incubated with the SOMAmer mixes (including SOMAmers to 1125 different proteins) to allow for equilibrium binding of fluorophore-tagged SOMAmers to their corresponding proteins. A series of automated partitioning steps are subsequently used to retain only the SOMAmers that are bound to proteins, which are then released from the protein complex prior to quantitation. Finally, the previously-protein-bound SOMAmer oligonucleotides – now a quantitative proxy for the proteins to which they were bound -- are captured by complementarity and quantitated using DNA hybridization arrays. The hybridization arrays are then normalized and calibrated using data from a reference set of pooled plasma samples run on each batch to adjust for batch-to-batch variation. Thus, the normalized, calibrated signal for each SOMAmer reflects the relative amount of each cognate protein present in the original sample, with findings reported in relative fluorescence units (RFU).

Statistical Methods

All of the data analysis and plots were analyzed and generated using R (R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org Version 3.02). Spearman's correlation coefficient was calculated for CSF and plasma

values of each protein. Linear Regression models were used to determine the significance of the associations between proteins markers and PD-related outcome variables.

Results

Analysis of 1124 proteins in CSF and plasma identifies 55 proteins with highly correlated levels in both biofluids

We calculated Spearman correlation coefficients for SOMAScan measurements from matched serum and CSF samples from 40 subjects. In order to reduce technical variability, subjects had CSF and serum drawn on the same day. We found that 55 proteins had significant correlations ($p<0.00005$, Bonferonni correction level) across plasma and CSF levels (**Table 4.1**), all with Spearman's correlation coefficient greater than 0.6.

49 of the 55 proteins replicate with significant correlations in an independent cohort of 18 subjects.

Matched CSF and plasma were collected from an additional 18 subjects in order to further validate the significance of the correlations of the previously identified proteins. 9 of these samples came from individuals with AD and 9 samples came from normal controls. 49 of the 55 proteins replicated by achieving a significant correlation ($p<.05$) (**Table 4.1**).

Demonstration of Utility: Blood-Based Biomarker Discovery for Parkinson's Disease

We wanted to demonstrate how this data could guide researchers interested in conducting biomarker discovery. An *a priori* hypothesis may suggest proteins that are correlated with CSF are more likely to be biological relevant biomarkers to brain disorders. As such we investigated whether these 49 proteins might be differentially expressed in PD.

6 out of the 49 proteins are differentially expressed in the serum of 80 PD and 65 NC subjects (at p<.05).

First, we examined whether any of the 49 proteins were significantly differentially expressed in the serum of 80 PD and 65 NC subjects. Regression models associating each of the 49 proteins with PD or Control status were run, controlling for Age, Gender, and LEDD. We found that 6 out of 49 proteins were differentially expressed at p<.05, including Growth hormone receptor, Luteinizing hormone, FSH, PARC, CYTD, ENPP7.

2 of these 6 proteins, Growth Hormone Receptor and PARC replicated in in an independent cohort of 96 PD and 45 NC plasma samples

We wanted to see whether the results from these 6 proteins replicated in an independent cohort of 96 PD and 45 NC subjects. Regression models associating each of the 6 proteins with PD or Control status were run, controlling for Age, Gender, and LEDD. 2 of the 6 proteins replicated, including Growth Hormone Receptor (p<.002) and PARC (p<.007).

PD patients with lower levels of Growth Hormone Receptor in plasma have worse motor symptoms measured by UPDRS3 ($p<.018$)

Finally, we wanted to determine whether Growth Hormone Receptor was also associated with other PD endpoints. Specifically, we wanted to determine whether PD subjects with lower Growth Hormone Receptor concentrations had worse motor symptoms (measured by UPDRS3 scores).

Regression models were run associating UPDRS scores with Growth Hormone Receptor concentrations controlling for Disease Duration, Gender, and LEDD. We found that PD patients with lower levels of Growth Hormone Receptor had significantly motor symptoms ($p<.018$).

In order to see whether this result replicated, we ran the same analysis in an independent cohort of 96 PD and 45 NC samples. We found that the results replicated in this independent cohort ($p<.007$).

Discussion

In this study, we report the levels of 1124 proteins in matched plasma and CSF samples of 40 subjects. We found 55 proteins that were significantly correlated across serum and CSF samples at the Bonferonni, multiple correction level ($p<.00005$), and 49 of these correlations replicated in an independent set of 18 matched CSF and plasma samples.

As a proof of concept to demonstrate the utility of these correlations in biomarker discovery, we performed a targeted proteomic analysis of PD biomarkers using the 49 proteins we've identified to be significantly correlated across CSF and blood. This

targeted analysis helps to identify a novel PD biomarker, Growth Hormone Receptor, which is found in lower concentrations in PD patients compared to Normal Controls. Moreover, among people with PD, those with lower Growth Hormone Receptor levels have worse motor symptoms. This suggests that Growth Hormone Receptor is not only a potential diagnostic marker of PD, but also an endophenotypic marker of motor disease severity.

There are many limitations to this study. First, serum blood was collected for the Oxford samples, while all the other blood samples were from plasma. Ideally, we would have used all serum or plasma to reduce extraneous sources of variability. However, the fact that we still see replication of our correlation results despite the differences in blood source shows the robustness of these blood-CSF associations. Second, relatively small sample sizes were available for patients who had matched CSF and plasma. Future studies should explore this type of analysis with larger samples sizes.

In summary, we find that a statically significant number of proteins show correlated CSF and plasma expression in a proteome-level analysis. All proteins shown passed the most conservative, Bonferroni correction for significance, though the absolute number of proteins that passed our significance threshold was modest. Large proteins are not meant to pass the blood/brain barrier so there is a question of whether these proteins are expected to be correlated across CSF and blood. Nevertheless, the data and results summarized here may serve as a useful reference for current and future studies to develop biomarkers in brain-based diseases.

Table 4.1: Correlations across CSF and Blood Samples

Protein Name	SomaID	Correlation (40 subject cohort)	Significance (40-subject cohort)	Correlation (18 subject cohort)	Significance (18 subject cohort)
CRP	SL000051	0.909193246	0	0.541795666	0.021974787
Elafin	SL004458	0.845028143	0	0.729618163	0.000861403
FCG2AB	SL017613	0.891557223	0	0.463364293	0.0545448
HCG	SL001766	0.856472795	0	0.742002064	0.000631404
IgD	SL000460	0.919136961	0	0.86996904	0
IL1RAcP	SL004588	0.824765478	0	0.570691434	0.014971567
Leptin	SL000498	0.841275797	0	0.537667699	0.023158691
PDGFRb	SL004155	0.859796427	1.20E-12	0.453044376	0.060707966
Trypsin 2	SL010388	0.826305175	5.16E-11	0.625709944	0.005478492
IgE	SL000461	0.82371705	6.68E-11	0.91744066	0
Siglec 9	SL005219	0.810488792	2.33E-10	0.708978328	0.00138548
Lysozyme	SL000510	0.783151181	2.34E-09	0.347781218	0.157527546
BCMA	SL004672	0.818761726	4.88E-09	0.795665635	0.000111862
Hemopexin	SL000440	0.815572233	9.25E-09	0.624355005	0.00672993
Trypsin	SL000603	0.808255159	2.07E-08	0.618163055	0.007428688
SAA	SL000572	0.807879925	2.13E-08	0.667698658	0.003181836
sRAGE	SL003680	0.807692308	2.16E-08	0.541795666	0.021974787
ARTS1	SL007729	0.804878049	2.66E-08	0.626418989	0.006509214
MICA	SL005199	0.80206379	3.20E-08	0.688338493	0.002135796
SLAF7	SL016928	0.789868668	5.93E-08	0.744066047	0.000598234
FSH	SL000428	0.789493433	6.03E-08	0.667698658	0.003181836
PCSK7	SL014069	0.735716299	6.36E-08	0.481940144	0.044675703
ENPP7	SL009045	0.727989118	1.02E-07	0.684210526	0.002318886
PAPPA	SL002755	0.769418386	1.25E-07	0.591331269	0.011168007
BST1	SL008644	0.768292683	1.30E-07	0.814241486	4.39E-05
MICB	SL005200	0.722735589	1.39E-07	0.793601651	0.000121898
Ckb81	SL003301	0.759099437	1.73E-07	0.793601651	0.000121898
sICAM1	SL002922	0.745778612	2.60E-07	0.706914345	0.001449288
IL17sR	SL004850	0.738273921	3.30E-07	0.832817337	6.55E-06
HCC.4	SL003300	0.733771107	3.82E-07	0.535603715	0.023769364
Growth hormone receptor	SL005168	0.700877153	4.75E-07	0.626418989	0.006509214
RTN4	SL008309	0.712382739	8.06E-07	0.655314757	0.003985461
Luteinizing hormone	SL000506	0.710131332	8.77E-07	0.585139319	0.01221611
MSP	SL005202	0.709756098	8.89E-07	0.90505676	0

PARC	SL003323	0.704502814	1.09E-06	0.517027864	0.029860689
Haptoglobin. Mixed Type	SL000437	0.68408462	1.14E-06	0.742002064	0.000631404
MPIF1	SL003302	0.694371482	1.61E-06	0.735810114	0.000739486
SIG14	SL014292	0.693245779	1.69E-06	0.758513932	0.000401645
IGFBP 1	SL000462	0.692870544	1.71E-06	0.710376966	0.000954305
sLeptin.R	SL003184	0.688180113	2.07E-06	0.673890609	0.002832295
Renin	SL000565	0.684803002	2.37E-06	0.600929352	0.008351853
Siglec 3	SL005215	0.66641651	4.98E-06	0.82868937	1.28E-05
HCC1	SL003329	0.650124303	5.62E-06	0.492260062	0.039827442
ILT4	SL005191	0.663227017	5.65E-06	0.50877193	0.03293506
Adiponectin	SL004258	0.656285178	7.46E-06	0.779153767	0.000209577
Desmoglein 2	SL004857	0.641583798	8.15E-06	0.725490196	0.00095102
IgM	SL000468	0.650093809	9.52E-06	0.696594427	0.001804372
PPAC	SL008063	0.647091932	1.07E-05	0.680082559	0.002514381
TPSB2	SL010617	0.646716698	1.09E-05	0.799793602	9.34E-05
CYTD	SL008382	0.637148218	1.57E-05	0.583075335	0.012582565
Chitotriosidase 1	SL006029	0.630206379	2.04E-05	0.578947368	0.013342029
SPINT2	SL001897	0.629268293	2.11E-05	-0.120743034	0.632523225
IL 1 R4	SL004146	0.622138837	2.75E-05	0.374613003	0.126305597
SHBG	SL005102	0.62195122	2.77E-05	0.502579979	0.035400021
FCN2	SL006542	0.600928837	4.13E-05	0.128998968	0.609108914

Conclusion: The role of biomarkers in discovering treatments for AD and PD

My DPhil aimed to explore how high variable capture approaches, specifically proteomics, and multivariate computational methods could be applied to identify biomarkers and signatures of disease for PD and AD.

In the first chapter, I gave an overview of the proteomic and computational approaches that enable biomarker discovery at scale. In the second chapter, I introduced a novel biomarker discovery approach that combines the advantages of hypothesis-driven and high variable capture approaches. Specifically, we employed an empirically derived feature selection approach, from an *in vitro* experimental model of AD (DKK1 overexpression), to identify blood-based biomarkers for AD. This empirical feature selection approach can be applied to not only identify biomarkers that are, by definition, related to a pathological process of interest, but also can be used to validate the relevance of a disease model in humans. The third chapter explored biomarkers for PD, using data from plasma and brain. I present machine-learning and Monte Carlo approaches to define endophenotypes within a disease and to interrogate whether the PD-related brain signature is also found in plasma and vice versa. In the fourth chapter, I identified proteins correlated across CSF and plasma and demonstrate how this data can be used in biomarker identification.

Throughout the DPhil, I have introduced robust computational pipelines that can be used to regularize the highly dimensional data generated from proteomic and metabolomics assays. As we move beyond candidate approaches for biomarker discovery, these computational pipelines will become increasingly important to identify and validate disease signatures, distinguishing signal amongst noise. However, more

importantly, these computational approaches can be used to define subgroups within a disease. For example, these disease signatures can identify subtle differences between patients with the same diseases, that will help stratify patients for clinical trials or therapeutic treatments.

In neurodegenerative diseases such as PD and AD, biomarkers will play a crucial role by facilitating earlier diagnosis and the screening of individuals into clinical trials. While AD and PD represent the two most prevalent neurodegenerative diseases in the world, affecting 30 and 5 million people, respectively, there are still no disease modifying therapies available. While billions of dollars have been invested in clinical trials for AD and PD, all of the studies in the past 10 years have failed to show the desired improvements in cognitive or motor endpoints to receive approval. *(87)* However, in many studies, certain subpopulations do seem to respond to treatment compared to others. As such, researchers are increasingly interested in incorporating biomarkers into the clinical trials in an attempt to characterize the subgroups that are more likely to respond to a given medication, given the target pathways that the drug modulates. Many of the computational approaches introduced in this work can be used to define these subgroups in a similar manner employed in Chapter 3 where we identified biomarkers across good and poor PD endophenotypes.

While we identify both biomarkers of state and trait, there is still work that needs to be done in respect to developing biomarkers that can help to improve drug development success rates. A disease modifying therapy must, by definition, alter the expected progression of disease for a patient. As such, it is important to have biomarkers that characterize not only the disease state where a patient begins, but where they are

likely to end up. In particular, a surrogate biomarker that both predicts disease trajectory but also tracks with disease severity would be invaluable, especially if it is easily assayable. The identification of this sort of marker would allow a researcher to test a drug's efficacy by observing whether it modulates the trajectory of the surrogate marker along with the endpoint of interest.

However, few studies, including the ones conducted in this thesis, are designed to allow for the discovery of these types of surrogate markers of disease state. Most studies measure proteins at one time point and then explore which proteins are differentially expressed across fast or slow decliners. Ideally, a study would need to measure candidate biomarkers across different time points and find the biomarkers that change across time that have a significant association with disease measures of interest (e.g., proteins that decrease as cognitive scores decrease). While some studies have done this for single proteins, few have done this at the proteome level and at enough time points to generate appropriate power to discover the relevant signatures, due to the high costs and potential high patient burden for participating in such a study. The Deep and Frequent Phenotyping Study, led by Professor Simon Lovestone and colleagues will be one of the first to measure biomarkers across all many modalities (CSF, blood, imaging with proteomic, metabolomics and transcriptomic phenotyping) at frequent time intervals to map the expected trajectory of disease. This will allow researchers not only to find biomarkers that track with disease severity across different biomarker modalities, but also to better characterize the different etiologies that contribute to the development of the diseases and thereby reveal potential therapeutic targets.

Moreover, there are many opportunities to combine proteomics-level data with

transcriptomics, genomics datasets, leveraging eQTL/pQTL types of analysis. By combining genetic, transcript, and protein level data, we may be able to develop not only more robust biomarkers and diagnostics, but also better reveal underlying etiologies of these diseases.

A number of recent scientific advances can enhance and supplement biomarker discovery. As discussed in Chapter 2, iPSCs derived from patients with AD can be used as a disease model in biomarker discovery. CRISPR can be used to edit the genome to analyze that markers that are modulated in a disease model. These methods expand the toolkit of researchers to create more accurate and relevant disease models and modulate known disease pathways to help elucidate novel markers of pathological processes.

Emergence of Digital Biomarkers and Continuous Behavior Measurement

The advent of mobile phone applications and wearable devices presents an exciting opportunity to enhance the real-time monitoring of patient behaviors and symptoms. These devices may empower researchers to collect objective measures of motor symptoms or cognitive states at greater frequencies. For example, in the context of Parkinson's Disease, objective measures of motor symptom severity using wearables and a phone's own sensors/gyroscopes can help researchers better characterize a patient's tremors. *(88)* These objective measurements of activity can be collected at more regular intervals and coupled with clinically validated surveys to empower researchers with a more complete representation of symptoms. One can imagine these measurements as 'digital biomarkers' that can track different disease symptoms.*(89, 90)*

In both PD and AD, cognitive and motor rating scales are often used as endpoints

for clinical trials. While these rating scales have been used across many decades of studies, they have many shortcomings including rater bias, subjectivity, and high variance. Moreover, because these scales are often administered in person, they are not particularly scalable. As such, cognitive measurements such as the MMSE or MOCA and motor scales such as the UPDRS3 are often only administered on a monthly or bi-yearly time scale. These measurements are used to represent a patient's disease severity for a time period proportional to the frequency of administration. As a result, changes in cognitive/motor state and variability of symptoms across time are not as robustly collected as researchers would like. Collecting measurements at these infrequent time intervals make the data vulnerable to misrepresentative variance. For example, rating scales may misrepresent a PD patient who happens to have particularly bad motor symptoms during the rating scale administration day.

In contrast, a sensor that collects real-time motor data passively from PD patients can empower researchers with a level of sensitivity previously unprecedented, allowing researchers to develop a more objective and stable sense of a patient's motor or cognitive symptoms. Because no human intervention is potentially needed during the administration of the test, this type of characterization is highly scalable and can be administered with greater frequency. For cognitive scales, a survey or mobile app equivalent of an MMSE that may be administered through a mobile device without human intervention and taps into various components could similarly allow for a better understanding of patients' cognitive states. The availability of these tools can empower researchers to develop more sensitive and specific biomarkers. While much work is still needed to develop and characterize these digital markers, we can imagine that these

digital tools will not only enhance biochemical biomarker discovery, but also help researchers better characterize the environmental factors that may influence disease progression and development.

As our populations age, the global burden of AD and PD will only increase, having huge societal and economic implications. By 2050, it is estimated that 1 in 85 people will have AD. Biomarkers will play a crucial role in the early identification and subsequent development of new therapies. In this thesis, we explore how high variable capture assays and computational approaches can be used to develop biomarker signatures at scale that can help diagnose and subcategorize patients with AD and PD. We look forward to translating these signatures to develop better diagnostics and treatments for patients with AD and PD.

References

1. R. Brookmeyer, E. Johnson, K. Ziegler-Graham, H. M. Arrighi, Forecasting the global burden of Alzheimer's disease, *Alzheimer's Dement.* **3**, 186–191 (2007).

2. B. Zheng, Z. Liao, J. J. Locascio, K. A. Lesniak, S. S. Roderick, M. L. Watt, *et al.* A Potential Therapeutic Target for Early Intervention in Parkinson's Disease, *Sci. Transl. Med.* **2**, 52–73 (2010).

3. T. B. Sherer, Biomarkers for Parkinson's Disease, *Sci. Transl. Med.* **3**, 79–93 (2011).

4. R. Anand, K. D. Gill, A. A. Mahdi, Therapeutics of Alzheimer's disease: Past, present and future, *Neuropharmacology* **76**, 27–50 (2014).

5. G. G. Glenner, C. W. Wong, Alzheimer's disease: Initial report of the purification and characterization of a novel cerebrovascular amyloid protein, *Biochem. Biophys. Res. Commun.* **120**, 885–890 (1984).

6. J.-P. Brion, J. Flament-Durand, P. Dustin, Alzheimer's disease and tau proteins, *Lancet* **328**, 1098 (1986).

7. J. A. Hardy, G. A. Higgins, Alzheimer's disease: the amyloid cascade hypothesis, *Science.* **256**, 184–185 (1992).

8. D. Campion, J.-M. Flaman, A. Brice, D. Hannequin, B. Dubois, C. Martin, *et al.,* Mutations of the presenilin I gene in families with early-onset Alzheimer's disease, *Human Molecular Genetics.* ***1995,*** *2373-2377* (1995).

9. A. Goate, M.-C. Chartier-Harlin, M. Mullan, J. Brown, F. Crawford, L. Fidani, *et al.,* Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease, *Nature* **349**, 704 (1991).

10. A. Kumar, A. Singh, Ekavali, A review on Alzheimer's disease pathophysiology and

its management: an update, *Pharmacol. Reports* **67**, 195–203 (2015).

11. S. Salomone, F. Caraci, G. M. Leggio, J. Fedotova, F. Drago, New pharmacological strategies for treatment of Alzheimer's disease: focus on disease modifying drugs, *Br. J. Clin. Pharmacol.* **73**, 504–517 (2012).

12. J. Hardy, The amyloid hypothesis for Alzheimer's disease: a critical reappraisal, *J. Neurochem.* **110**, 1129–1134 (2009).

13. W.-Y. Wang, M.-S. Tan, J.-T. Yu, L. Tan, Role of pro-inflammatory cytokines released from microglia in Alzheimer's disease, *Ann. Transl. Med.* **3**, 136 (2015).

14. A. D. Roth, G. Ramírez, R. Alarcón, R. Von Bernhardi, Oligodendrocytes damage in Alzheimer's disease: Beta amyloid toxicity and inflammation, *Biol. Res.* **38**, 381–387 (2005).

15. I. Dal Prà, A. Chiarini, L. Gui, B. Chakravarthy, R. Pacchiana, E. Gardenal, *et al.,* Do Astrocytes Collaborate with Neurons in Spreading the "Infectious" Aβ and Tau Drivers of Alzheimer's Disease?, *Neurosci.* **21**, 9–29 (2014).

16. W. Q. Qiu, M. F. Folstein, Insulin-degrading enzyme and amyloid-beta peptide in Alzheimer's disease: review and hypothesis, *Neurobiol. Aging* **27**, 190–198 (2017).

17. G. Silvestrelli, A. Lanari, L. Parnetti, D. Tomassoni, F. Amenta, Treatment of Alzheimer's disease: From pharmacology to a better understanding of disease pathophysiology, *Mech. Ageing Dev.* **127**, 148–157 (2006).

18. W. F. Goure, G. A. Krafft, J. Jerecic, F. Hefti, Targeting the proper amyloid-beta neuronal toxins: a path forward for Alzheimer's disease immunotherapeutics, *Alzheimers. Res. Ther.* **6**, 42 (2014).

19. M. Löhle, A. Storch, H. Reichmann, Beyond tremor and rigidity: non-motor features

of Parkinson's disease, *J. Neural Transm.* **116**, 1483 (2009).

20. J. P. L. Daher, Interaction of LRRK2 and α-Synuclein in Parkinson's Disease BT - Leucine-Rich Repeat Kinase 2 (Springer International Publishing, Cham, 2017), pp. 209–226.

21. L. Gold, D. Ayers, J. Bertino, C. Bock, A. Bock, E. N. Brody, *et al.,* Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery, *PLoS One* **5**, 15004 (2010).

22. P. Hickey, M. Stacy, Available and emerging treatments for Parkinson's disease: a review, *Drug Des. Devel. Ther.* **5**, 241–254 (2011).

23. W. H. Oertel, Recent advances in treating Parkinson's disease, *F1000Research* **6**, 260 (2017).

24. J. K. Qiang, Y. C. Wong, A. Siderowf, H. I. Hurtig, S. X. Xie, V. M.-Y. Lee, *et al.,* Plasma Apolipoprotein A1 as a Biomarker for Parkinson's Disease, *Ann. Neurol.* **74**, 119–127 (2013).

25. A. Shtilbans, C. Henchcliffe, Biomarkers in Parkinson's disease: an update, *Curr. Opin. Neurol.* **25** (2012).

26. B. D. W. Group, Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework, *Clin. Pharmacol. Ther.* **69**, 89–95 (2001).

27. B. Olsson, R. Lautner, U. Andreasson, A. Öhrfelt, E. Portelius, M. Bjerke, *et al,,* CSF and blood biomarkers for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis, *Lancet Neurol.* **15**, 673–684 (2017).

28. C. Marcus, E. Mena, R. M. Subramaniam, Brain PET in the Diagnosis of Alzheimer's Disease, *Clin. Nucl. Med.* **39**, 413–426 (2014).

29. P. Eusebi, D. Giannandrea, L. Biscetti, I. Abraha, D. Chiasserini, M. Orso, *et al,,* Diagnostic utility of CSF α-synuclein species in Parkinson's disease: protocol for a systematic review and meta-analysis, *BMJ Open* **6** (2016).

30. L. Gao, H. Tang, K. Nie, L. Wang, J. Zhao, R. Gan, et al., Cerebrospinal fluid alpha-synuclein as a biomarker for Parkinson's disease diagnosis: a systematic review and meta-analysis, *Int. J. Neurosci.* **125**, 645–654 (2015).

31. B. Mollenhauer, V. Cullen, I. Kahn, B. Krastins, T. F. Outeiro, I. Pepivani, *et al.,* Direct quantification of CSF α-synuclein by ELISA and first cross-sectional study in patients with neurodegeneration, *Exp. Neurol.* **213**, 315–325 (2008).

32. T. Tokuda, S. A. Salem, D. Allsop, T. Mizuno, M. Nakagawa, M. M. Qureshi, *et al.,* Decreased α-synuclein in cerebrospinal fluid of aged individuals and subjects with Parkinson's disease, *Biochem. Biophys. Res. Commun.* **349**, 162–166 (2006).

33. Z. Hong, M. Shi, K. A. Chung, J. F. Quinn, E. R. Peskind, D. Galasko, *et al.,* DJ-1 and α-synuclein in human cerebrospinal fluid as biomarkers of Parkinson's disease, *Brain* **133**, 713–726 (2010).

34. M. J. Park, S.-M. Cheon, H.-R. Bae, S.-H. Kim, J. W. Kim, Elevated Levels of α-Synuclein Oligomer in the Cerebrospinal Fluid of Drug-Naïve Patients with Parkinson's Disease, *J Clin Neurol* **7**, 215–222 (2011).

35. P. G. Foulds, O. Yokota, A. Thurston, Y. Davidson, Z. Ahmed, J. Holton, *et al.,* Post mortem cerebrospinal fluid α-synuclein levels are raised in multiple system atrophy and distinguish this from the other α-synucleinopathies, Parkinson's disease and Dementia with Lewy bodies, *Neurobiol. Dis.* **45**, 188–195 (2012).

36. B. Mollenhauer, J. J. Locascio, W. Schulz-Schaeffer, F. Sixel-Döring, C.

Trenkwalder, M. G. Schlossmacher, α-Synuclein and tau concentrations in cerebrospinal fluid of patients presenting with parkinsonism: A cohort study, *Lancet Neurol.* **10**, 230–240 (2011).

37. M. Shi, J. Bradner, A. M. Hancock, K. A. Chung, J. F. Quinn, E. R. Peskind, *et al.,* Cerebrospinal fluid biomarkers for Parkinson disease diagnosis and progression, *Ann. Neurol.* **69**, 570–580 (2011).

38. F. Tateno, R. Sakakibara, T. Kawai, M. Kishi, T. Murano, Alpha-synuclein in the Cerebrospinal Fluid Differentiates Synucleinopathies (Parkinson Disease, Dementia With Lewy Bodies, Multiple System Atrophy) From Alzheimer Disease, *Alzheimer Dis. Assoc. Disord.* **26** (2012).

39. T. Tokuda, M. M. Qureshi, M. T. Ardah, S. Varghese, S. A. S. Shehab, T. Kasai, *et al.,* Detection of elevated levels of α-synuclein oligomers in CSF from patients with Parkinson disease, *Neurol.* **75**, 1766–1770 (2010).

40. C. Tuerk, L. Gold, Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase, *Science.* **249**, 505-510 (1990).

41. G. A. Nagana Gowda, D. Raftery, Biomarker Discovery and Translation in Metabolomics, *Curr. Metabolomics* **1**, 227–240 (2013).

42. T. K. Ho, in *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, ICDAR '95. (IEEE Computer Society, Washington, DC, USA, 1995), p. 278.

43. C. Cortes, V. Vapnik, Support-vector networks, *Mach. Learn.* **20**, 273–297 (1995).

44. B. Olsson, R. Lautner, U. Andreasson, A. Öhrfelt, E. Portelius, M. Bjerke, *et al.,* CSF and blood biomarkers for the diagnosis of Alzheimer's disease: a systematic review and

meta-analysis, *Lancet Neurol.* **15**, 673–684 (2016).

45. M. Sattlecker, S. J. Kiddle, S. Newhouse, P. Proitsi, S. Nelson, S. Williams, *et al.,* Alzheimer's disease biomarker discovery using SOMAscan multiplexed protein technology, *Alzheimer's Dement. J. Alzheimer's Assoc.* **10**, 724–734 (2015).

46. A. Caricasole, A. Copani, F. Caraci, E. Aronica, A. J. Rozemuller, A. Caruso, *et al.,* Induction of Dickkopf-1, a Negative Modulator of the Wnt Pathway, Is Associated with Neuronal Degeneration in Alzheimer's Brain, *J. Neurosci.* **24**, 6021–6027 (2004).

47. S. A. Purro, E. M. Dickins, P. C. Salinas, The Secreted Wnt Antagonist Dickkopf-1 is Required for Amyloid β-Mediated Synaptic Loss, *J. Neurosci.* **32**, 3492–3498 (2012).

48. M. C. Rosi, I. Luccarini, C. Grossi, A. Fiorentini, M. G. Spillantini, A. Prisco, *et al.,* Increased Dickkopf-1 expression in transgenic mouse models of neurodegenerative disease, *J. Neurochem.* **112**, 1539–1551 (2010).

49. R. Killick, E. M. Ribe, R. Al-Shawi, B. Malik, C. Hooper, C. Fernandes, *et al.,* Clusterin regulates beta-amyloid toxicity via Dickkopf-1-driven induction of the wnt-PCP-JNK pathway, *Mol Psychiatry* **19**, 88–98 (2014).

50. G. V. De Ferrari, N. C. Inestrosa, Wnt signaling function in Alzheimer's disease, *Brain Res. Rev.* **33**, 1–12 (2000).

51. N. C. Inestrosa, E. Arenas, Emerging roles of Wnts in the adult nervous system, *Nat Rev Neurosci* **11**, 77–86 (2010).

52. E. M. Dickins, P. C. Salinas, Wnts in action: from synapse formation to synaptic maintenance, *Front. Cell. Neurosci.* **7**, 162 (2013).

53. M. Sattlecker, S. J. Kiddle, S. Newhouse, P. Proitsi, S. Nelson, S. Williams, *et al.,* Alzheimer's disease biomarker discovery using SOMAscan multiplexed protein

technology, *Alzheimer's Dement.* **10**, 724–734 (2014).

54. S. Kiddle, M. Sattlecker, P. Proitsi, A. Simmons, E. Westman, C. Bazenet, *et al.,* Candidate Blood Proteome Markers of Alzheimer's Disease Onset and Progression: A Systematic Review and Replication Study, *J. Alzheimer's Dis.* **5**, 102-105 (2013).

55. W. T. Hu, D. M. Holtzman, A. M. Fagan, L. M. Shaw, R. Perrin, S. E. Arnold, *et al.,* Plasma multianalyte profiling in mild cognitive impairment and Alzheimer disease, *Neurology* **79**, 897–905 (2012).

56. O. SE, G. Xiao, R. Barber, A serum protein–based algorithm for the detection of alzheimer disease, *Arch. Neurol.* **67**, 1077–1081 (2010).

57. S. E. O'Bryant, G. Xiao, R. Barber, R. Huebinger, K. Wilhelmsen, M. Edwards, *et al.,* A Blood-Based Screening Tool for Alzheimer's Disease That Spans Serum and Plasma: Findings from TARC and ADNI, *PLoS One* **6**, 28092 (2011).

58. S. E. O'Bryant, G. Xiao, R. Barber, C. M. Cullum, M. Weiner, J. Hall, *et al.,* Molecular Neuropsychology: Creation of Test-Specific Blood Biomarker Algorithms, *Dement. Geriatr. Cogn. Disord.* **37**, 45–57 (2014).

59. S. Lista, F. Faltraco, D. Prvulovic, H. Hampel, Blood and plasma-based proteomic biomarker research in Alzheimer's disease, *Prog. Neurobiol.* **101–102**, 1–17 (2013).

60. C. Noelker, H. Hampel, R. Dodel, Blood-Based Protein Biomarkers for Diagnosis and Classification of Neurodegenerative Diseases, *Mol. Diagn. Ther.* **15**, 83–102 (2011).

61. M. Thambisetty, S. Lovestone, Blood-based biomarkers of Alzheimer's disease: challenging but feasible, *Biomark. Med.* **4**, 65–79 (2010).

62. S. Lovestone, P. Francis, I. Kloszewska, P. Mecocci, A. Simmons, H. Soininen, et al., on behalf of the A. Consortium, AddNeuroMed—The European Collaboration for the

Discovery of Novel Biomarkers for Alzheimer's Disease, *Ann. N. Y. Acad. Sci.* **1180**, 36–46 (2009).

63. T. M. Dewey, A. Mundt, G. J. Crouch, M. C. Zyzniewski, B. E. Eaton, New Uridine Derivatives for Systematic Evolution of RNA Ligands by Exponential Enrichment, *J. Am. Chem. Soc.* **117**, 8474–8475 (1995).

64. D. W. Huang, B. T. Sherman, R. A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat. Protoc.* **4**, 44–57 (2008).

65. D. W. Huang, B. T. Sherman, R. A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, *Nucleic Acids Res.* **37**, 1–13 (2009).

66. W. E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods, *Biostat.* **8**, 118–127 (2007).

67. N. Meinshausen, P. Bühlmann, Stability selection, *J. R. Stat. Soc. Ser. B Statistical Methodol.* **72**, 417–473 (2010).

68. R. Tibshirani, Regression shrinkage and selection via the lasso: a retrospective, *J. R. Stat. Soc. Ser. B Statistical Methodol.* **73**, 273–282 (2011).

69. B. Efron, R. Tibshirani, Improvements on Cross-Validation: The .632+ Bootstrap Method, *J. Am. Stat. Assoc.* **92**, 548–560 (1997).

70. Y. Huang, X. Sun, G. Hu, An integrated genetics approach for identifying protein signal pathways of Alzheimer's disease, *Comput. Methods Biomech. Biomed. Engin.* **14**, 371–378 (2011).

71. J.-C. Lambert, S. Heath, G. Even, D. Campion, K. Sleegers, M. Hiltunen, *et al.,* Genome-wide association study identifies variants at CLU and CR1 associated with

Alzheimer's disease, *Nat Genet* **41**, 1094–1099 (2009).

72. M. Thambisetty, A. Simmons, L. Velayudhan, Association of plasma clusterin concentration with severity, pathology, and progression in alzheimer disease, *Arch. Gen. Psychiatry* **67**, 739–748 (2010).

73. M. Thambisetty, S. M. Resnick, L. Ferrucci, D. Wong, Y. Zhou, Clusterin is a Plasma Biomarker of Severity, Progression and Pathology in Alzheimer's Disease, *Alzheimer's Dement. J. Alzheimer's Assoc.* **5**, 1 (2015).

74. K. van Dijk, W. Jongbloed, W. van de Bergh, P. Scheltens, S. Mulder, P. Eikelenboom, *et al.,* Clusterin: An early biomarker for Alzheimer's disease?, *Alzheimer's Dement. J. Alzheimer's Assoc.* **9**, 198–199 (2015).

75. V. E. Krupnik, J. D. Sharp, C. Jiang, K. Robison, T. W. Chickering, L. Amaravadi, *et al.,* Functional and structural diversity of the human Dickkopf gene family, *Gene* **238**, 301–313 (1999).

76. Q.-S. Chen, B. L. Kagan, Y. Hirakura, C.-W. Xie, Impairment of hippocampal long-term potentiation by Alzheimer amyloid β-peptides, *J. Neurosci. Res.* **60**, 65–72 (2000).

77. A. Auffret, V. Gautheron, M. P. Mattson, J. Mariani, C. Rovira, Progressive Age-Related Impairment of the Late Long-Term Potentiation in Alzheimer's Disease Presenilin-1 Mutant Knock-in Mice, *J. Alzheimers. Dis.* **19**, 1021–1033 (2010).

78. M. Lawton, F. Baig, M. Rolinski, C. Ruffman, K. Nithi, M. T. May, *et al.,* Parkinson's Disease Subtypes in the Oxford Parkinson Disease Centre (OPDC) Discovery Cohort, *J. Parkinsons. Dis.* **5**, 269–279 (2015).

79. Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, *et al.,* The Montreal Cognitive Assessment, MOCA: A Brief Screening Tool For

Mild Cognitive Impairment, *J. Am. Geriatr. Soc.* **53**, 695–699 (2005).

80. S. Greffard, M. Verny, A. Bonnet, Motor score of the unified parkinson disease rating scale as a good predictor of lewy body–associated neuronal loss in the substantia nigra, *Arch. Neurol.* **63**, 584–588 (2006).

81. A. Hye, J. Riddoch-Contreras, A. L. Baird, N. J. Ashton, C. Bazenet, R. Leung, *et al.,* Plasma proteins predict conversion to dementia from prodromal disease., *Alzheimers. Dement.* **10**, 799–807 (2014).

82. P. D. Mehta, T. Pirttilä, S. P. Mehta, E. A. Sersen, P. S. Aisen, H. M. Wisniewski, Plasma and cerebrospinal fluid levels of amyloid β proteins 1-40 and 1-42 in Alzheimer disease, *Arch Neurol* **57** (2000).

83. J. B. Toledo, L. M. Shaw, J. Q. Trojanowski, Plasma amyloid beta measurements - a desired but elusive Alzheimer's disease biomarker, *Alzheimers. Res. Ther.* **5**, 8 (2013).

84. D. Melzer, J. R. B. Perry, D. Hernandez, A.-M. Corsi, K. Stevens, I. Rafferty, *et al.,* A Genome-Wide Association Study Identifies Protein Quantitative Trait Loci (pQTLs), *PLOS Genet.* **4**, 1000072 (2008).

85. L. Wu, S. I. Candille, Y. Choi, D. Xie, L. Jiang, J. Li-Pook-Than, *et al.,* Variation and genetic control of protein abundance in humans, *Nature* **499**, 79 (2013).

86. J. Jakubowska, E. Hunt, M. Chalmers, D. Leader, M. McBride, A. F. Dominiczak, System level visualization of eQTLs and pQTLs, *BMC Bioinformatics* **6**, 15 (2005).

87. L. S. Schneider, F. Mangialasche, N. Andreasen, H. Feldman, E. Giacobini, R. Jones, *et al.,* Clinical trials and late-stage drug development for Alzheimer's disease: an appraisal from 1984 to 2014, *J. Intern. Med.* **275**, 251–283 (2014).

88. A. J. Espay, P. Bonato, F. Nahab, W. Maetzler, J. M. Dean, J. Klucken, *et al.,*

Technology in Parkinson disease: Challenges and Opportunities, *Mov. Disord.* **31**, 1272–1282 (2016).

89. J. Torous, J. Rodriguez, A. Powell, The New Digital Divide For Digital Biomarkers, *Digit. Biomarkers* **1**, 87–91 (2017).

90. B. M. Bot, C. Suver, E. C. Neto, M. Kellen, A. Klein, C. Bare, *et al.,* The mPower study, Parkinson disease mobile data collected using ResearchKit, **3**, 160011 (2016).