UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 1385

Proteomics Studies of Subjects with Alzheimer's Disease and Chronic Pain

PAYAM EMAMI KHOONSARI

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2017

ISSN 1651-6206
ISBN 978-91-513-0111-2
urn:nbn:se:uu:diva-331748

Dissertation presented at Uppsala University to be publicly examined in Rosénsalen, Akademiska sjukhuset, Ing 95/96, nbv, Uppsala, Tuesday, 5 December 2017 at 09:00 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English. Faculty examiner: Docent Ann Brinkmalm (Institutionen för neurovetenskap och fysiologi, Sahlgrenska akademin, Sahlgrenska universitets sjukhuset).

Abstract

Emami Khoonsari, P. 2017. Proteomics Studies of Subjects with Alzheimer's Disease and Chronic Pain. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine* 1385. 82 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-0111-2.

Alzheimer's disease (AD) is a neurodegenerative disease and the major cause of dementia, affecting more than 50 million people worldwide. Chronic pain is long-lasting, persistent pain that affects more than 1.5 billion of the world population. Overlapping and heterogenous symptoms of AD and chronic pain conditions complicate their diagnosis, emphasizing the need for more specific biomarkers to improve the diagnosis and understand the disease mechanisms.

To characterize disease pathology of AD, we measured the protein changes in the temporal neocortex region of the brain of AD subjects using mass spectrometry (MS). We found proteins involved in exo-endocytic and extracellular vesicle functions displaying altered levels in the AD brain, potentially resulting in neuronal dysfunction and cell death in AD.

To detect novel biomarkers for AD, we used MS to analyze cerebrospinal fluid (CSF) of AD patients and found decreased levels of eight proteins compared to controls, potentially indicating abnormal activity of complement system in AD.

By integrating new proteomics markers with absolute levels of Aβ42, total tau (t-tau) and p-tau in CSF, we improved the prediction accuracy from 83% to 92% of early diagnosis of AD. We found increased levels of chitinase-3-like protein 1 (CH3L1) and decreased levels of neurosecretory protein VGF (VGF) in AD compared to controls.

By exploring the CSF proteome of neuropathic pain patients before and after successful spinal cord stimulation (SCS) treatment, we found altered levels of twelve proteins, involved in neuroprotection, synaptic plasticity, nociceptive signaling and immune regulation.

To detect biomarkers for diagnosing a chronic pain state known as fibromyalgia (FM), we analyzed the CSF of FM patients using MS. We found altered levels of four proteins, representing novel biomarkers for diagnosing FM. These proteins are involved in inflammatory mechanisms, energy metabolism and neuropeptide signaling.

Finally, to facilitate fast and robust large-scale omics data handling, we developed an e-infrastructure. We demonstrated that the e-infrastructure provides high scalability, flexibility and it can be applied in virtually any fields including proteomics. This thesis demonstrates that proteomics is a promising approach for gaining deeper insight into mechanisms of nervous system disorders and find biomarkers for diagnosis of such diseases.

Keywords: Bioinformatics, microservices, biomarkers, Alzheimer's disease, chronic pain, fibromyalgia, neuropathic pain, spinal cord stimulation, cloud computing, proteomics, metabolomics, software, workflows, data analysis, mass spectrometry

Payam Emami Khoonsari, Department of Medical Sciences, Clinical Chemistry, Akademiska sjukhuset, Uppsala University, SE-75185 Uppsala, Sweden.

ISSN 1651-6206
ISBN 978-91-513-0111-2
urn:nbn:se:uu:diva-331748 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-331748)

To my beloved family

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I Musunuri, S*.; **Emami Khoonsari, P***.; Mikus, M.; Wetterhall, M.; Haggmark-Manberg, A.; Lannfelt, L.; Erlandsson, A.; Bergquist, J.; Ingelsson, M.; Shevchenko, G.; Nilsson, P.; Kultima, K., Increased Levels of Extracellular Microvesicle Markers and Decreased Levels of Endocytic/Exocytic Proteins in the Alzheimer's Disease Brain. *Journal of Alzheimer's disease : JAD* **2016,** *54* (4), 1671-1686.

II **Emami Khoonsari, P**.; Haggmark, A.; Lonnberg, M.; Mikus, M.; Kilander, L.; Lannfelt, L.; Bergquist, J.; Ingelsson, M.; Nilsson, P.; Kultima, K*.; Shevchenko, G*., Analysis of the Cerebrospinal Fluid Proteome in Alzheimer's Disease. *PloS one* **2016,** *11* (3), e0150672.

III **Emami Khoonsari, P**.; Shevchenko, G.; Herman, S.; Musunuri, S.; Remnestål, J.; R., B.; Degerman Gunnarsson, M.; Kilander, L.; Zetterberg, H.; Nilsson, P.; Lannfelt, L.; Ingelsson, M.; Kultima, K., Chitinase-3-like protein 1 (CH3L1) and Neurosecretory protein VGF (VGF) as two novel CSF biomarker candidates for improved diagnostics in Alzheimer's disease. *Manuscript*.

IV Lind, A. L.; **Emami Khoonsari, P**.; Sjodin, M.; Katila, L.; Wetterhall, M.; Gordh, T.; Kultima, K., Spinal Cord Stimulation Alters Protein Levels in the Cerebrospinal Fluid of Neuropathic Pain Patients: A Proteomic Mass Spectrometric Analysis. *Neuromodulation : journal of the International Neuromodulation Society* 2016, 19 (6), 549-62.

V **Emami Khoonsari, P**.; Musunuri, S.; Herman, S.; Svensson, CI.; Lars, T.; Gordh, T.; Kultima, K., Systematic Analysis of the Cerebrospinal Fluid Proteome of Fibromyalgia patients. *Manuscript*.

VI **Emami Khoonsari, P**.; Moreno, P.; Bergmann, S.; Burman, J.; Capuccini, M.; Carone, M.; Cascante, M.; Atauri, P.; Dudova, Z.; Foguet, C.; Gonzalez-Beltran, A.; Hankemeier, T.; Haug, K.; He, S.;

Herman, S.; Johnson, D.; Kale, N.; Larsson, A.; Salek, R.; Neumann, S.; Peters, K.; Pireddu, L.; Rocca-Serra, P.; Roger, P.; Rueedi, R.; Ruttkies, C.; Sadawi, N.; Sansone, S.; Schober, D.; Selivanov, V.; A. Thévenot, E.; van Vliet, M.; Zanetti, G.; Steinbeck, C.; Kultima, K.; Spjuth, O., Interoperable and scalable metabolomics data analysis with microservices. *Manuscript*.

*Author with equal contributions

Reprints were made with permission from the respective publishers.

Related papers not included in this thesis:

I Nikitidou, E.; **Emami Khoonsari, P**.; Shevchenko, G.; Ingelsson, M.; Kultima, K.; Erlandsson, A., Increased Release of Apolipoprotein E in Extracellular Vesicles Following Amyloid-beta Protofibril Exposure of Neuroglial Co-Cultures. Journal of Alzheimer's disease : JAD 2017, 60 (1), 305-321.

II Almandoz-Gil, L.; Welander, H.; Ihse, E.; **Emami Khoonsari, P**.; Musunuri, S.; Lendel, C.; Sigvardson, J.; Karlsson, M.; Ingelsson, M.; Kultima, K.; Bergstrom, J., Low molar excess of 4-oxo-2-nonenal and 4-hydroxy-2-nonenal promote oligomerization of alpha-synuclein through different pathways. Free radical biology & medicine 2017, 110, 421-431.

III Herman, S.; **Emami Khoonsari, P**.; Aftab, O.; Krishnan, S.; Strömbom, E.; Larsson, R.; Hammerling, U.; Spjuth, O.; Kultima, K*.; Gustafsson, M*., Mass spectrometry based metabolomics for in vitro systems pharmacology: pitfalls, challenges, and computational solutions. Metabolomics 2017, 13 (7), 79.

*Author with equal contributions

Contents

s

Abbreviations

2-DE	Two-dimensional gel electrophoresis
2D-PAGE	Two-dimensional polyacrylamide gel electrophoresis
AC	Alternating current
ACN	Acetonitrile
AD	Alzheimer's disease
API	Application programming interface
APP	Amyloid precursor protein
Aβ	Beta-amyloid
CID	Collision-induced dissociation
CNS	Central nervous system
CSF	Cerebrospinal fluid
DC	Direct current
DTT	1,4-dithiothreitol
ELISA	Enzyme linked immunosorbent assay
ESI	Electrospray ionization
FM	Fibromyalgia
FTD	Frontotemporal dementia
FT-ICR	Fourier transform ion cyclotron mass spectrometer
GABA	Gamma-aminobutyric acid
GO	Gene ontology
GUI	graphical user interface
HAc	Acetic acid
HPC	High performance computing
IAA	Iodoacetic acid
IaaS	Infrastructure as a service
IAM	Iodoacetamide
LC	Liquid chromatography
m/z	Mass-to-charge ratio
MALDI	Matrix-assisted laser desorption/ionization
MAP	The microtubule-associated protein
MCI	Mild cognitive impairment
MS	Mass spectrometry
nHPLC	Nano liquid chromatography
OND	Other neurological disorders
PaaS	Platform as a service
PD	Parkinson's disease

PLS-DA	PLS discriminant analysis
PNS	Peripheral nervous system
p-tau	Phosphorylated-tau
QT	Quality threshold
RF	Radio frequency
RT	Retention time
SaaS	Software as a service
SCS	Spinal cord stimulation
SILAC	Stable isotope labeling by amino acids in cell culture
SPE	Solid phase extraction
TEAB	Triethylammonium bicarbonate
TOF	Time-of-flight mass spectrometry
VRE	Virtual research environment
WSE	Weak scaling efficiency

Introduction

Nervous system

The nervous system is arguably the most complex structure in the human body. Composed of billions of neurons, nerves and glial cells, the nervous system is the center of consciousness and responsible for all somatic and autonomic bodily functions. The nervous system receives inputs from the sensory organs which are then processed and interpreted to trigger an appropriate motor output such as muscle movement or regulating body homeostasis. Neurons and nerves are the main building blocks of the nervous system that transmit electrochemical signals, enabling seamless communication throughout the nervous system. These units have a tree-like shape consisting of a round cell body attached to multiple dendrites to receive and one axon to send electrical signals (using multiple terminals) to other cells. These signals are released in the form of transmitters from the axon's terminals and are received using dendrites of the targeted cells and converted back to electrical signals. In humans (as well as in many other higher vertebrates) the nervous system consists of two major parts: the central nervous system (CNS) and the peripheral nervous system (PNS). These two parts together build a highly plastic network, making it possible to receive input from the sensory system, integrate them and finally trigger an appropriate response through motor output (figure 1).

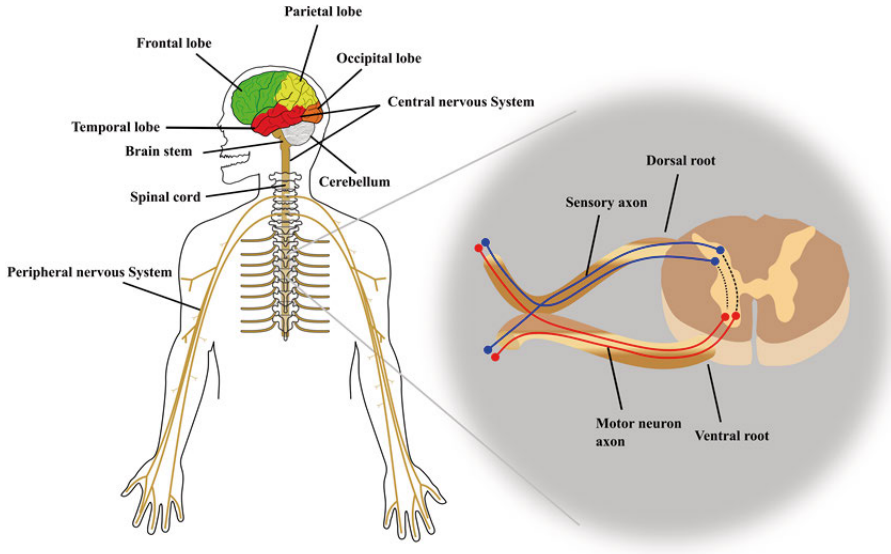

Figure 1. Components of the nervous system. The nervous system consists of central and peripheral nervous system. The central part receives signal and exerts its action through the peripheral part.

Central nervous system

The CNS is the processing part of the nervous system and is responsible for monitoring and coordinating organ function and responding to changes in the external stimuli. The CNS consists of the brain and spinal cord.

The brain

It has been reported that the brain consists of approximately 100 billion neurons and ten times more glial cells [1]. These cells are distributed in major specialized areas of the brain (cortex, cerebellum, basal ganglia and brain stem) that are interconnected and work together to perform tasks such as thinking (cortex), coordination (cerebellum) and breathing. Human brain can also be anatomically divided into multiple lobes that are also related to different brain functions (frontal lobes: judgment and motor function, occipital lobes: visual processing, parietal lobes: sensation and movement, temporal lobes: memory). Although these areas have been associated with distinct functions, these activities can move to different locations as the consequence of brain plasticity. The brain plasticity (called neuroplasticity) is defined as the ability of the brain to reorganize or form synaptic connections. Neuroplasticity allows the brain to develop from the immature brain (infancy) to adulthood as well as to compensate for loss of function in the case of disease or brain injury [2-4]. These sophisticated functions give humans a unique capacity to learn new skills, memorize new experience and adapt to new environments [5, 6].

Spinal cord

The spinal cord is a (approximately) 45 cm long bundle of nerves that begins from the brain stem to the lower part of the spine. The main function of the spinal cord is to conduct sensory and motor information between the brain and rest of the body through peripheral nervous system. Primarily, spinal cord carries messages from motor cortex (using efferent nerves) to the body and transmits the signals from sensory receptors to sensory cortex (using afferent nerves) [7]. In addition, spinal cord contains neural pathways which can perform involuntary reflexes in response to external stimulus. The spinal cord contains white and grey matter. The white matter contains long bundle of axons (coated with myelin) that carry information up and down the spinal cord and between different areas of cerebrum and other parts of the brain. The grey matter, on the other hand, contains masses of cell bodies, dendrites and axon terminals. The grey matter is further divided into ventral roots and dorsal roots. The ventral roots contain axon of motor neurons that receive the information from the brain and send it to the skeletal muscles. The dorsal roots contain sensory axons that send information (through spinal tracts) to the brain. These complex ascending and descending pathways allow the brain the efficiently communicate with rest of the body and exert its function.

Cerebrospinal fluid

The cerebrospinal fluid (CSF) is a biologic fluid and is produced by choroid plexus in the lateral and fourth ventricles with the rate of approximately 500 ml per day (the total volume of CSF in the healthy adult is approximately 150 ml) [8-10]. The primary function of CSF is to protect (mechanically and chemically) the CNS. Specifically, CSF can act as shock absorber (e.g., cushion) which lessens possible impacts to the head. Moreover, CSF can remove waste products from e.g., cerebral metabolism and also maintain homeostasis in the brain by distributing many substances such as hormones to other areas of the brain [11]. As CSF is in direct contact with the brain, it can reflect pathological activity in the CNS [12, 13]. Analysis of CSF can facilitate more accurate diagnosis of several CNS diseases such as brain hemorrhage [14], multiple sclerosis, meningitis [15] and Alzheimer's disease (AD) [16].

Peripheral nervous system

The part of the nervous system that connects the brain and spinal cord to sensory receptors and other organs such as muscles is referred to as PNS. The complex PNS network mainly consists of sensory receptors (chemoreceptors, thermoreceptors, mechanoreceptors and photoreceptors) and motor neurons. More specifically, PNS is often subdivided into two subsystems

called somatic and the autonomic nervous system. The somatic system is mainly associated with voluntary movement of muscles. Whereas the autonomic nervous system regulates certain organs (e.g. cardiac muscle or glands) without voluntary control. To maintain communication between the brain and other body organs, 12 pairs of cranial nerves and 31 pairs of spinal nerves are used in the nervous system. The cranial nerves are mainly associated with sensory and motor function in the head and neck whereas the spinal nerves originate from the spinal cord (dorsal and ventral roots) and carry signal between all the body organs and the spinal cord and the brain. The majority of the mentioned nerves are referred to as mixed since they both carry motor as well as sensory signals thought some cranial nerves such as olfactory nerves are known to be specialized to relay specific sensory data for example related to smell [17, 18].

Neurological disorders

Neurological disorders are diseases that cause abnormality in CNS and PNS. The neurological disorders are very common and pose a large burden on worldwide health. Approximately 1 in 9 people dies due to a nervous system disease [19] and billions of people are affected by these disorders. This number is expected to increase in many of these disorders [20]. Neurological disorders can be either acute or degenerative, causing sudden or gradual loss of a specific or several functions. As essentially all the bodily functions are controlled by the nervous system, these disorders can have a devastating effect on the quality of life. Depending on the area involved, neurological disorders can result in wide range of symptoms such as loss of sensation, pain, altered consciousness as well as abnormalities in memory or cognition. Examples of such diseases are multiple sclerosis, AD, schizophrenia, neuropathic pain, fibromyalgia (FM) and Parkinson's disease (PD) which are often very difficult to diagnose and treat [21, 22].

Biomarkers for neurological disorders

Diagnosis of neurological conditions have traditionally been performed by clinicians through excluding unlikely diseases based on the presence or absence of certain symptoms [23], resulting in a high rate of misdiagnosis of many neurological conditions [24-27]. This clearly indicates a great need for identification of biomarkers for neurological disorders that can provide new insights into the disease pathology as well as offer new possibilities for diagnosing and treatment of affected patients [28, 29]. Finding such markers however, poses great challenges such as limited availability of tissue from the target site as well as complexity of the nervous system [29], all resulting in poor specificity of such markers [30, 31]. The CSF can be regarded as the

main source of biomarkers for nervous system disorders [29]. This, however, requires a lumbar puncture, a relatively invasive procedure and CSF is therefore not as available as other body fluids, such as blood and plasma. In addition, obtaining CSF from healthy controls is difficult since few volunteers want to undergo lumbar puncture. This thesis is an attempt to address some of these problems in biomarker discovery in three neurological disorders: AD (**Paper I, II, and III**), neuropathic pain (**Paper IV**) and FM (**Paper V**).

Alzheimer's disease

Alzheimer's disease is an age-dependent neurodegenerative disorder and the most common form of dementia in the elderly population, accounting for more than 50% of all dementia cases [32]. The first notable symptoms of AD include memory loss, disorientation, and impairment of other cognitive functions. Epidemiological investigations have estimated that the numbers of AD patients will double every 20 years to more than 66 million worldwide in 2030 and 100 million by 2050 [33, 34]. Alzheimer's disease is mainly associated with multiple molecular characteristics, including extracellular amyloid-β (Aβ) plaque deposition and accumulation of intracellular neurofibrillary tangles composed mainly of hyperphosphorylated tau proteins. Whereas 10–15% of all AD cases are caused by dominant mutations in either of three different disease genes (*APP*, *PSEN1*, or *PSEN2*), which all are related to the generation of amyloid-β (Aβ), the vast majority of sporadic cases have a largely unknown etiology. In the normal condition, the amyloid precursor protein (APP) is cleaved by α-secretase resulting in α-APP and C-83. These two products can be further processed to produce APPICD as well P3 which both are believed to be nontoxic. Moreover, APP can also be processed by β- and γ-secretases that instead generates Aβ40 and Aβ42 peptides. The Aβ42 species are more prone to adopt a beta-sheet conformation and can thereby more readily aggregate to oligomers, larger prefibrillar species and insoluble plaques. Especially the prefibrillar species are believed to have neurotoxic properties [35]. Furthermore, the presence of plaques can cause microglial activation which in turn causes production of excessive amount of pro-inflammatory cytokines, stimulating the neurons to produce more Aβ42, resulting in oxidative damage [36, 37].

The microtubule-associated protein (MAP), with six major isoforms [38], is essential for the assembly and stability of the microtubules, an important component of the neuronal cytoskeleton [39]. In the AD brain, abnormally hyperphosphorylated tau accumulate as neurofibrillary tangles. The accumulation of dysfunctional Aβ and tau are believed to mediate the extensive loss of neurons and synapses as well as the inflammatory processes in the AD brain [40].

Despite great progress in defining the pathogenesis of AD, numerous changes in the AD brain still remain to be characterized [41]. Potentially, such knowledge will be important for the development of novel disease bi-

omarkers. Today, measurements of decreased Aβ42 and increased tau and phosphorylated-tau (p-tau) in CSF are used to aid the clinical diagnosis. The combination of these markers have been reported to be indicative of AD with a sensitivity of 71 to 95% and a specificity of 44% to 87% [42, 43]. However, based on recent reports, the sensitivity can in practice be lower at prodromal disease stages, i.e. in patients with mild cognitive impairment (MCI) [44]. Generally, MCI is recognized as the intermediate stage of brain impairment whereby the patients show impaired memory and additional cognitive dysfunctions [45]. Some of these patients have considerable risk of developing dementia and particularly AD, yet they do not meet the clinical criteria for AD [46, 47]. Therefore, research in MCI can likely lead to revised clinical criteria or biomarkers that allow detection and intervention at earlier time.

Finally, in additional to diagnosis, the current biomarkers for AD are poor in prognosticating the disease progression and cannot be used to monitor response to immunotherapy with monoclonal antibodies against Aβ and tau or other treatment strategies that are currently being evaluated. Thus, there is a great need to find new biomarkers that also could be used for these purposes. Our goal in **Paper I, II, and III** was to find proteins altered in AD to understand pathological changes in AD and provide potential biomarkers for early AD diagnosis.

Chronic pain

The International Association for the Study of Pain has defined pain as an "unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage". The pain state is normally divided into two subgroups acute and chronic pain. Acute pain is provoked by specific stimuli and is considered the body's normal reaction to for example physical injuries that is not long lasting. Chronic pain, in contrast, lasts for more than three months and can become progressively more intense and sometime is considered as a disease. The estimated prevalence of chronic pain can be as high as 60% [48-51] and severely affect quality of life (e.g. inability to exercise, sleep and maintain relationships) [52]. Chronic pain can further be categorized to nociceptive (i.e. damage to body tissue) or neuropathic pain (damage to nervous system).

Neuropathic pain

Neuropathic pain is a prevalent complex chronic pain defined as lesion or disease of the PNS and CNS [53]. Neuropathic pain patients can be broadly classified into several categories, such as patients with CNS lesions, multifocal nerve lesions, and peripheral generalized polyneuropathies [54]. The patients often show various symptoms including paraesthesia, thermohypoesthesia and mechanical dynamic allodynia which can occur in a variety of

different combinations in different pain groups. Although pathophysiological mechanisms in neuropathic pain are not fully understood, peripheral sensitization where peripheral neurons become abnormally sensitive and central sensitization, including hyperexcitability in spinal cord neurons (leading to increased activity of neurons in response to stimuli), are thought to be two of the main reasons behind neuropathic pain. Effective treatment of neuropathic pain remains a great unmet medical need. Current pharmacological treatments are often (e.g. amitriptyline, duloxetine [55]) unsatisfactory [54] with patients suffering substantial residual pain and treatment side effects. Electrical neuromodulation by spinal cord stimulation (SCS) is a treatment option for specific neuropathic (and ischemic) pain conditions, leading to pain reduction of e.g. 50–70% of eligible neuropathic pain patients [56]. Although SCS has been used since 1960s and has been shown to change the level of e.g. substance P [57] and gamma-aminobutyric acid (GABA) [58], its mechanism of action, especially on protein level, is not clear to the scientific community. While SCS is a beneficial treatment option, it can lead to certain side effects such as migration, connection failure or breakage [59]. In addition, SCS is not globally available and in some cases (pain in the area not covered by SCS) might not produce desired results for everyone, however, a better understanding of SCS mechanism may trigger further investigations and lead to improved treatment strategies for neuropathic pain. Our goal in **Paper IV** was to find proteins altered by SCS and gain insight into SCS mechanism of action.

Fibromyalgia

The FM syndrome is a chronic pain condition that is recognized by wide spread pain in somatic tissues such as muscles [60]. Fibromyalgia affects approximately 2% of the population aged between 20 and 50 years and is more prevalent in females [61, 62]. The current criteria for FM diagnosis include diffuse soft tissue pain, widespread pain, pain responses in a minimum number of tender points, sleep, fatigue, and morning stiffness [63]. It has been claimed that the sensitivity and specificity of the diagnostic test for FM are approximately 88% and 81%, respectively [62]. Despite recent great efforts by the scientific community, the etiology of FM is not well understood [64]. The recent evidence suggests abnormalities in the neuroendocrine, lowered pain thresholds, and partly environmental and genetic factors [64, 65]. Currently there is no biomarker that can objectively diagnose FM. However, there have been several promising findings including cytokines [66], active peptides [67], blood proteins [68], metabolites [69], etc. [70, 71] that might facilitate diagnosis of FM. In addition, genetic markers, imaging of CNS as well as neurotransmitter and hormone levels have been proposed to be good predictors of FM [72]. However, currently none of the proposed biomarkers have been extensively used in clinic. The new biomarkers can be implemented as part of clinical diagnostic test that can complement the clas-

sical clinical criteria and therefore patients can receive proper treatment. In **Paper V**, we aimed to find potential biomarkers for FM based on CSF samples.

Large scale proteomics for biomarker discovery

Biomarkers play a critical role in improving diagnosis and drug development in health care. However, identification, qualification and validation of diagnostic and prognostic biomarkers require extensive characterization of the targeted samples (e.g. biofluid or tissue) which in turn necessitate applying different methodologies and instruments. Biomarker studies can either be performed in targeted manner with a set of predefined molecules measured by various methods (often called hypothesis driven) or to use an unbiased approach involving large scale detection platforms that are referred to as hypothesis generating methods. In this context, "omics" technologies are used to detect genes (genomics), mRNA (transcriptomics), proteins (proteomics), metabolites (metabolomics) and fluxes (fluxomics), providing an overview of the targeted system as a whole. These strategies have potential applications in many fields including drug development, biomarker discovery [73-76] and personalized medicine [77]. Among these technologies, proteomics has been widely used in clinical biomarker discovery [76, 78, 79]. This includes blood, plasma and CSF biomarkers for AD [80-82] and chronic pain [83-85]. Proteomics was used as the main approach in **Papers I, II, III, IV and V**.

Proteomics

Proteomics, the term introduced by Wilkins *et al* as a complement to genomics [86], represents the identification and quantification of all the proteins present in an organism as well as a description of the molecular basis of pathophysiological processes. Compared to the relatively static genome, the proteome has more variability in the composition and corresponds to a large number of phenotypes. This can be exemplified by the one-to-many relationship between genes and proteins as one human gene on an average can result in more than ten proteins [87]. In addition, slight changes in the level or post-translational modifications, e.g. by environmental factors, can change the expression and function of proteins.

In order to study all proteins expressed in an organism, post-translational modifications, alternative splice products and the broad dynamics range represent some challenges. On the other hand, the field has been facilitated by several advancements during the past decades such as improvements in separation technology, e.g. by the use of two-dimensional polyacrylamide gel electrophoresis (2D-PAGE), nano liquid chromatography (LC) methods [88]

and soft ionization techniques such as electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) to facilitate these challenges. Moreover, development of high-resolution mass spectrometry (MS) instruments, such as Fourier transform ion cyclotron resonance [89] and Orbitrap technology [90], has provided a further improvement in the sensitivity and mass resolving power, enabling identification and quantification of thousands of proteins in a short time. Simultaneously, there have been major developments in the field of immunodiagnostics for detection of proteins, especially related to Enzyme Linked Immunosorbent Assay (ELISA) [91] and multiplex bead-based assays, for detection of hundreds of proteins in a small sample volume. Today, mass spectrometry coupled with separation instruments or immunoassays are indispensable tools to speed up the discovery of new drugs and disease biomarkers for *in vitro* diagnostics [92].

Computational challenges in proteomics

The advent of high throughput proteomics instruments has resulted to generation of massive and complex datasets [93]. The increased size and complexity of datasets led to developing a large number of sophisticated computational tools to analyze this data [93]. This poses two challenges to the users: 1) the use of desktop or workstation computers for analysis is often not sufficient because of the high requirements for memory and processing resources. 2) Installation of different tools and their dependencies as well as chaining them together into a workflow demand substantial knowledge in the relevant areas (such as the operating system or internal functions of the tools).

To speed up the processing of omics data and alleviate the tool installation, high performance computing (HPC) systems are used in for example academic institutes. These systems offer aggregated computing power provided by several computers with high-end hardware (e.g. more processing units as well as memory). Due to availability of large resources, the computationally demanding operations can be performed using these supercomputers in reasonable amount of time which otherwise are infeasible on workstation computers. Due to high demand for using these system (e.g. within an institute) by multiple users, there are normally rigid constraints on the way these resources must be used (e.g. a queue system for using the resources). Furthermore, a multitude of computational tools needs to be installed and updated which normally requires approval and may even be restricted by system administrators. Cloud computing offers a compelling alternative to HPC systems, providing frameworks that can be instantiated on-demand and includes operating systems and software tools. This can facilitate faster, more accurate and reproducible data analysis for proteomics data, leading to more robust and comparable results and ultimately

more reliable biological conclusions. In **Paper VI**, developed a cloud based data analysis framework and showcased its application in “omics” (metabolomics) data handling.

Methods

Shotgun proteomics

Shotgun or bottom-up proteomics refers to a method that can characterize proteins by analysis of enzymatically cleaved peptides, providing an indirect measurement of proteins via their peptides [94]. In contrast, top-down proteomics is used to characterize intact proteins, providing advantage of observing the precise location of post-translation modifications as well as protein isoform determination [95]. However, due to differences in solubility, greater molecular weight (compared to peptides) and largely unknown fragmentation patterns, the bottom-up proteomics is better suited for protein quantification [96]. In a typical shotgun proteomics experiment, the proteins in the mixture first undergo digestion, e.g. by addition of trypsin, resulting in a large number of peptides. After extensive sample preparation, the peptides are then separated by their physiochemical properties by a separation technique, ionized through an electrospray source [97] and entered into the MS instrument where their mass-to-charge ratio (m/z) and their relative abundance will be recorded. Finally, the resulting data will be processed and analyzed to identify and quantify the peptide species that will be used for protein inference.

Sample preparation for shotgun proteomics

The sample mixture should be free of contaminants such as detergents and plastics to reduce unwanted interference and provide more robust identification and quantification measurements. Sample preparation steps should be carefully planned and monitored according to characteristics of the experiment such as sample type and amount in order to minimize contamination and other environmental effects. Typically, this procedure starts at sample collection (assuming that representative biological samples have been selected) where care should be taken to collect the samples under appropriate condition to avoid contaminations as well as protein degradation by immediately freezing the samples. Subsequently, the samples might be subjected to total protein estimation, immunodepletion, protein digestion and sample cleanup. Furthermore, depending on the experiment, the quantification can be performed as labeled or label free quantification. Finally, the LC-MS experi-

ment is performed followed by data pre-processing, peptide identification, and downstream data analysis (figure 2).

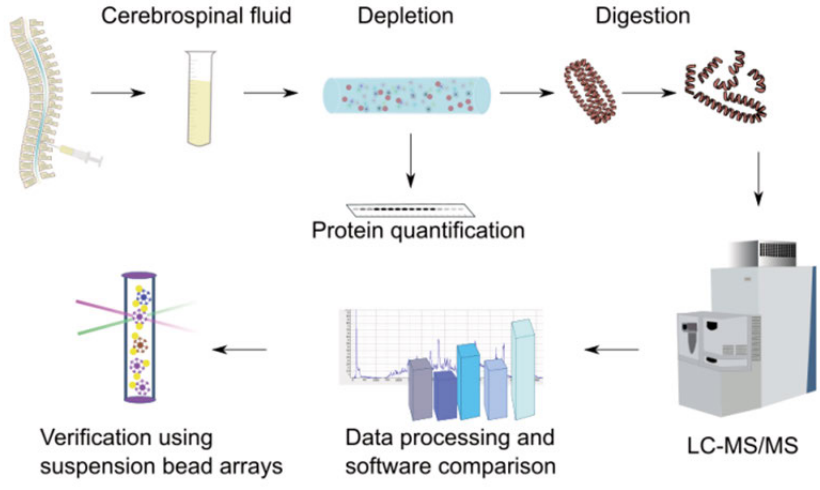

Figure 2. A typical workflow of proteomics experiment for human CSF. The experiment starts from selecting and collecting a sample and is followed by immunodepleting, protein digestion, MS analysis, data processing and here verification of results using an orthogonal technique.

Immunodepletion

The human body fluids are one source for the biomarker discovery. A thorough, systematic examination of these sources can facilitate detection of biomarkers for the early disease diagnosis. However, characterization of human body fluids proteome is a difficult task due to multiple factors including the immense dynamic range of protein concentrations (e.g. 10 to 12 orders of magnitude in CSF [98] and plasma [99]). Depending on the type of sample, usually the top 10-14 most abundant proteins can build up approximately 90% of the total protein content of the sample [100], making it extremely difficult to detect other proteins with the lower concentrations and very wide dynamic range using the current MS technology (which can detect proteins up to four orders of magnitude). This limited of detection is caused by competition between high and low abundant proteins in several stages including digestion, ionization and most importantly at the detector. Therefore, protein enrichment is indispensable to reduce the dynamic range and increase the proteome coverage. Immunodepletion of high abundant proteins using multi-affinity removal system of specific proteins is one method used to achieve lower protein dynamic range. Multi-affinity removal systems commonly use immobilized specific antibodies to remove typically top 7-12 of the highest abundant proteins [101]. This system results in two fractions of proteins (flow through and elute fractions), where the flow through frac-

tion normally is used for the rest of the analyses which can greatly improve characterization of proteome [101-103]. As the flow through fraction does not contain the depleted proteins, the dynamic range of protein concentrations will be lower compared to the crude. Therefore, one can measure relatively low abundant proteins which were not accessible by MS in crude sample. We have used immunodepletion of the 7 and 14 high abundant proteins in the **Papers II and III, IV and V**.

Digestion and sample clean-up

One of the main steps in the sample preparation for shotgun proteomics is the cleavage of proteins to peptides. In most cases, trypsin is used to digest the proteins and convert them to peptides (other alternative enzymes are e.g. Lys-C or Asp-N). Trypsin is a highly specific protease which cleaves at arginine (Arg) or lysine (Lys) residues (with the exception of Lys or Arg bound to carboxyl terminal proline (Pro)), producing peptides in the MS preferred mass range (~600–4,000 Da) [104]. Prior to digestion, the proteins need to be denatured, thus disrupting the protein tertiary structure to make the cleavage sites more accessible to trypsin. This step is normally performed simultaneously along with a reduction step to prevent re-folding. By combining heat and a reducing reagent (commonly 1,4-dithiothreitol (DTT), β-mercaptoethanol, or tris(2-carboxyethyl)phosphine), disulfide bonds will be reduced and renaturation of the proteins will be prevented. Addition of iodoacetamide (IAM) and iodoacetic acid (IAA) will further reduce the potential renaturation (through alkylation of cysteine) [105]. Three types of digestion are common in the proteomics field: in-gel, in-solution, and on-filter digestion. In the gel electrophoresis-based method, proteins are first separated in one or two dimensions followed by in-gel digestion to identify proteins. However, this technique is time-consuming and is normally performed if one additional separation is needed before the MS analysis [106]. In-solution and on-filter digestion (normally used in LC-MS/MS) are simpler and more straightforward, requiring less protein for digestion compared to the gel based method. For the in-solution digestion method, trypsinization is performed after adding ammonium bicarbonate or triethylammonium bicarbonate (TEAB) to the solution, which provides an optimal pH for trypsin. For the on-filter digestion [107] the solution is transferred to a spin filter, followed by exchange of buffer and addition of the enzyme to the filter. After the digestion, the peptides elute from the filter by centrifugation. This allows for the removal of interfering chemicals and small molecules after protein solubilization and before digestion [107]. Finally, since the buffer contains salts, buffers, detergents and contaminants an extensive sample clean-up needs to be performed (after performing labeling in the case of labeled quantification). Several sample cleanup techniques are available, including ultrafiltration [108], precipitation [109], and solid phase extraction (SPE) [110]. The precipitation and solid phase extraction methods were used

in **Papers I, II, III, IV and V**. Precipitation starts by adding a precipitant to the solution, which causes the proteins to precipitate in the suspension and form a pellet through centrifugation. The supernatant is then removed and the pellet is re-dissolved in a buffer, allowing for concentrating and purifying the samples. The SPE columns are another approach to separate interferences from the biological samples using solid particles (sorbent) packed in a SPE column which separates the analyte based on either polarity or ionic interaction. In the first step, the solvent conditions the column and the samples are transferred into the column. In the second step, the unwanted materials are washed off and the column is rinsed to collect the analytes of interest. The choice of solutions as well as the sorbents is dependent on the type of analytes and the matrix. In the case of **Papers II and III, IV and V**, Isolute C18 solid phase extraction column was used along with acidification using acetic acid (HAc) and as eluting solvent Acetonitrile (ACN).

Separation techniques

Proteomics mixtures normally contain a large number of digested proteins, resulting in extremely complex samples. Contemporary MS instruments typically can separate a large number of these peptides based on their masses. However, in the process of ionization, the peptides might compete for getting ionized, lowering the chance of ionization for low abundance or poorly ionizable groups, also known as ion suppression. Furthermore, direct infusion of the sample into the MS can cause saturation effects, preventing accurate measurement of the ions. Therefore, an additional method to separate the biomolecules before ionization can greatly enhance the number and accuracy of the detected biomolecules [111].

There are two categories of techniques available for performing the separation prior to MS analysis, gel-based and gel free. Gel-based methods such as sodium dodecyl sulfate polyacrylamide gels (SDS-PAGE) and 2D-GE are applied on protein level (before digestion) [112] in which the proteins are separated based on their isoelectric points and molecular weights. Briefly, the samples are solubilized and will be loaded on an isoelectric focusing gel in which an electric field will cause the protein to move through the gel until they reach their isoelectric point (pI). The proteins will be re-solubilized in SDS and DTT, causing denaturation of the proteins. SDS also binds to the denatured proteins, imparting negative charge to the proteins which is approximately proportional to their molecular weight. The proteins will then be loaded on a polyacrylamide in an applied electric field, causing the proteins to move towards the positive anode with rates depending on their masses, thereby separating base on mass. Finally, the proteins in the gel will be excised, digested and analyzed by MS. In contrast to gel-based approaches, gel-free or chromatography based approaches represent an attempt to separate the compounds through interactions with e.g. small size particles. In the

subsequent section liquid chromatography (LC) is described. This technique is widely used in proteomics although other options such as ion-exchange, capillary electrophoresis, and size-exclusion chromatography are also common. We used LC to perform peptide separation in **Papers I, II, III, IV and V**.

Liquid Chromatography

High-performance nano liquid chromatography (nHPLC) is a common separation technique in shotgun proteomics. An nHPLC system consists of a column (packed with silica particles attached to alkyl chains C4-C18) connected to one or several mobile phases. The samples are first dissolved in an aqueous solution and transported by a mobile phase onto the column. The mobile phase is moved through the column by pressure created by a pump. As the compounds move through the column, they interact with the stationary phase causing them to elute from the column in different time points depending on their physiochemical properties. Utilizing a gradient of aqueous and organic phase solvents, the ratio of aqueous/organic phase can be adjusted to sequentially (e.g. 100-min gradient from 2% to 50% organic) release the peptides with different affinity to the column. Typically, acidified (formic acid or trifluoroacetic acid) water and methanol/acetonitrile are used for the aqueous and organic phase, respectively. The peptides are then ionized and their mass to charge ratio (m/z) is recorded by the MS.

Compared to traditional high-performance liquid chromatography (HPLC), nHPLC uses particles with a size of about five microns and a typical flow rate of 200-400 nL/min, generating a back pressure ranging between 100-250 bar. As an improved alternative, ultra-performance nano liquid chromatography (UPLC) uses smaller particles (1.7 microns) to achieve higher resolving power and separation speed, thus generating higher back-pressures, typically in ranges of 400-800 bar at flow rates of 200-400 nL/min. Moreover, using a tip size of 1 μm, smaller spray droplets can be generated, resulting in further improvement in ionization, all of which results in increased resolving power compared to nHPLC and nUPLC. In **Papers I, II, III, IV and V** we used nHPLC to perform the separation of peptides prior to MS analysis.

Mass spectrometry

MS-based proteomics is a well-established high-throughput method for identification and quantification of proteins in complex samples. Although MS has a long history of more than a century [113], its application in proteomics was not extensively used earlier than 1980s (before invention of soft ionization techniques [114]). Since then MS has had an extremely rapid progress, leading to the advent of modern instruments with tremendous precision and speed. A typical MS instrument consists of three components, an ion source,

a mass analyzer and a detector. The analytes are first turned into gas phase ions by the ion source and their mass to charge ratios (m/z) are measured by the mass analyzer. Finally, the intensity of each ion with specific m/z is recorded by the detector.

Ionization

Any MS analysis requires the analytes to be ionized and to be in gas phase. Matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization (ESI) are two technologies that are commonly used in MS-based proteomics. These two "soft" ionization techniques enable analysis of large macromolecules such as peptides and proteins. In MALDI, first introduced in 1988 by Hillenkamp [115], the samples are first co-crystallized with a matrix (e.g. dihydrobenzoic acid) on a metal plate. The ultraviolet light from a laser radiation is absorbed by the matrix, resulting in vaporization of the matrix and analyte. The analytes will then receive charge using the matrix as proton donor and receptor. This technique normally results in singly charged ions without fragmentation in gas phase. The MALDI technique is generally more robust for ion suppression [116], producing a high ion yield [117] and a direct correlation between the mass spectra and the levels of the corresponding peptides. However, since LC is one of the main separation techniques being used in shotgun proteomics, coupling LC-MALDI to MS is often desired. However, this setup poses a challenge as the fractions of effluent need to be spotted onto MALDI plates and taken to the MS whereas ESI can be easily coupled to MS and greatly enhances tandem MS as described in the subsequent section.

Electrospray ionization

Electrospray ionization is a soft ionization technique which can transfer ions from solution into the gas phase using high voltage, without resulting in source fragmentation [97]. The process of ionization starts by generating a spray of charged droplets through a needle, which is maintained at a high voltage (3 – 6.0 kV). The charged droplets are reduced in size by evaporation through high temperature and drying gas (e.g. nitrogen) as they move towards the mass analyzer, causing increased charge density on the droplet surface. Finally, the droplets explode as a result of greater Coulombic repulsion [118] than the surface tension, resulting in smaller droplets. This process is repeated, resulting in released ions which pass through a sampling cone or the orifice of a capillary. Electrospray ionization generally results in doubly charged peptides ions (or multiply charged peptides in the case of long peptides), making it possible for subsequent MS detection of large biomolecules such as intact proteins (as m/z of large molecules will be within favorable mass range of the instrument). However, computer algorithms are needed to derive molecular weight of the compounds from multiply charged ions. Electrospray ionization can operate in both positive (protonation) and

negative (deprotonation) models and provide good sensitivity, adaptability to liquid chromatography and tandem MS [119]. In **Papers I, II, III, IV and V** we used ESI in positive mode to ionize the peptides.

Mass analyzers

A mass analyzer is a part of MS that measures m/z ratio of the ionized compounds. Since 1917, when the precursor of the model mass spectrometers was developed [120], many instruments have been invented, proposing new methods for separating ions based on their molecular weights. In 1938, the sector mass analyzer [121] was one of the first MS technologies invented. By this technique, the trajectories of ions into circular paths are bended upon applying a magnetic field perpendicular to the direction of ion motion. Starting from the ion source, ions with potentially different kinetic energy are focused based on their kinetic energy-to-charge ratios. The ions will then pass through a magnetic sector which disperses the ions in space so that the ions with identical m/z can be focused at a slit where they can be measured by a detector. Therefore, holding a constant potential in the electric sector, ions of different m/z can be separated by changing the magnetic field. In 1946, the time-of-flight mass spectrometer (TOF) was invented but was not widely used until the 1980s after MALDI had been invented [122].

Most of the modern TOF instruments work through acceleration of ions via a fixed potential into a drift region (fixed length). Assuming the same kinetic energy for all the ions, the time they take to reach the detector can be used to calculate the velocities and subsequently the mass through a kinetic energy formula ($0.5\ mv^2$). In addition, most of the TOF instruments use an electrostatic mirror by which the ions are reflected to the ion detector, compensating for small differences in their kinetic energy as well as increase in mass resolving power and accuracy (as the same ions with small kinetic energy difference will have the almost identical energy before reaching the detector) [123].

The 1970s witnessed the invention of the quadrupole mass analyzer, one of the most widely used instruments in the MS history. Quadrupole mass analyzers use a combination of radio frequency (RF) and direct current (DC) voltages on four rods to change the trajectories of ion with a given m/z. By changing the magnitude of the RF and DC voltages, one can pass ions with a specific m/z through the detector while neutralizing the other ions [124]. Closely related to quadrupole, ion traps follow the same principle, with the difference that the electric field is applied in three dimensions, with the help of two end-cap and one ring shape electrodes (3D ion traps [125]). As the voltage increases, the ions of higher m/z are ejected through the end-cap opening before they are detected by the detector. Following the sample principles, linear ion traps consist of two trapping elements and a central section [125]. The ions are trapped in the central section radially and axially via RF and DC. The unstable ions will then be ejected and detected via an alternat-

ing current (AC) voltage. In the same decade, the Fourier transform ion cyclotron MS (FT-ICR) was developed and became one of the most powerful instruments in terms of mass resolving power and mass accuracy [126]. In a FT instrument, the ions are trapped in an ICR cell within a magnetic field. The ions orbit with cyclotron frequencies which are inversely proportional to their m/z ratios. An RF voltage will be applied perpendicular to the magnetic field on excitation plates causing the ions to excite to higher radii orbits. The detectors on the detection plates will then record image currents induced by oscillating field of ions which will be transformed to a mass spectrum by Fourier transformation (figure 3). The orbitrap [127], introduced in 2000, is one of the newest instruments and has a similar trapping function as FT-ICR. However, orbitraps use an electrostatic field generated by an outer electrode and an inner axial spindle shaped to make ions orbiting around the spindle and performing harmonic oscillations at the frequency proportional to $(m/z)^{0.5}$. The image current will then be detected and transformed to a mass spectrum via Fourier transformation. In **Papers I, II, III, IV and V** we used FT-ICR to measure m/z and intensity of peptides.

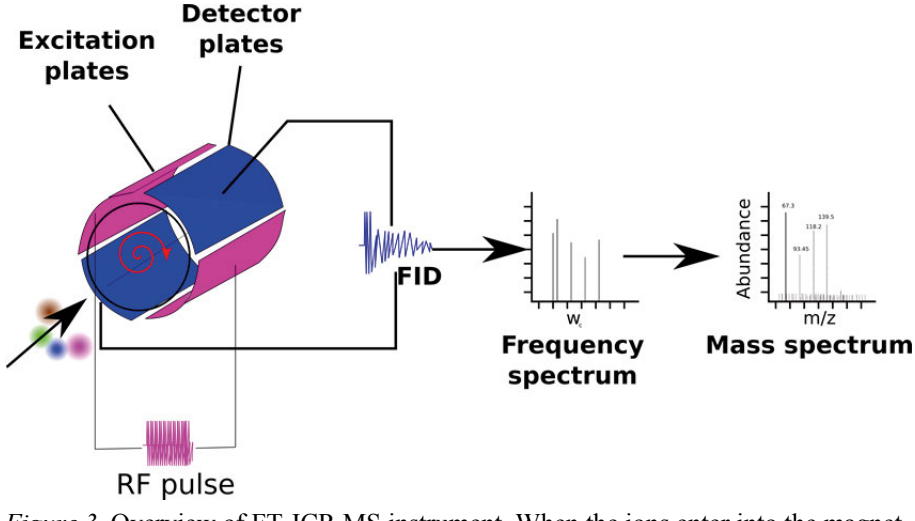

Figure 3. Overview of FT-ICR MS instrument. When the ions enter into the magnetic field and their path will be bend into a circular motion (with ion cyclotron frequency). This frequency of rotation of the ions will be recorded by the detection plates and the transform to a mass spectrum.

Tandem mass spectrometry (MS/MS)

Tandem mass spectrometry or MS/MS is a technique that can provide fragment ions or structural information about the compound of interest via two stages of MS analysis (MS^2). In the first stage the ions are separated based on their m/z ratio. A precursor ion will then be selected for fragmentation

usually by letting it collide with a neutral gas. The resulting product ions will be separated based on their m/z in the second stage of the MS analysis.

Typically, two types of instrumentation designs can be used to perform MS/MS, tandem in space and tandem in time. Tandem in space refers to the design where two instruments are coupled with a connector maintained in high vacuum. The precursor ions are selected in the first instrument, fragmented, and resultant product ions are measured in the second mass analyzer. Various instruments can be coupled to perform tandem in space e.g. quadrupole, time-of-flight as well as linear ion trap and FT-ICR. In contrast, tandem in time is performed using an ion trap where precursor selection and measuring product ions occur over time in the same trap. Furthermore, the fragmentation can be achieved by multiple techniques, such as collision-induced dissociation (CID) and photodissociation. The collision-induced dissociation is the most common technique used in proteomics. Generally, the precursor ions are accelerated to a high kinetic energy and collide with static target neutral gas molecules such as helium, resulting in a conversion of the translational energy into internal energy and decomposition [128]. The cleavage of the amide bonds will result in different fragments of molecules which can be further characterized by e.g. database searching. A common CID is called higher energy collisional dissociation (HCD) in which the fragmentation occurs in HCD cells and ions are transferred to the c-trap (in orbitrap instruments) to measure their masses. In **Papers I, II, III, IV and V** we used MS^2 to obtain fragmentation information about the peptides and utilized this to identify peptide sequences.

Quantification of proteins and peptides

The main goal of biomarker discovery and clinical proteomics is to provide accurate quantification of proteins/peptides that can be used for diagnostic and/or prognostic assessment of patient condition. Quantitative proteomics can be performed using several methods, including two-dimensional gel electrophoresis (2-DE) and MS [129] or a combination of those. Although not inherently quantitative due to issues such as chromatography reproducibility, ionization efficiency and missing peptide abundance, MS-based proteomics protocols have been designed to provide quantitative information [130]. Relative quantification is the most common method of MS-based quantification and is achieved by comparing the levels of peptides/proteins across different sample types (e.g. disease and healthy control). Relative quantification can be performed label free or by using stable isotope labeling by amino acids in cell culture (SILAC) or other peptide labeling techniques, such as dimethyl labeling. In absolute quantification a known quantity of a stable isotope-labeled standard peptide is added to the samples and the MS signal will be compared to the synthetic peptide [131]. We have used label free quantification of peptides in **Papers I, II, III, IV and V** and dimethyl

labeling of peptides in **Paper IV**. In the subsequent section, label free approach and dimethyl labeling are described.

Label free quantification

Label free quantification is more straightforward and less costly compared to labeling techniques. Currently the quantification in label free is performed using either the precursor ion intensity (MS^1) or the product ions (fragment spectra). Quantification using MS^1 typically involves integrating over extracted ion chromatograms (XICs) for each peptide and compare it across the samples. Whereas MS^2-based quantification involves either counting of the product ions for each peptide (called spectral count) or quantifying using product ions intensities [132], with the assumption that more MS^2 spectra might correspond to higher protein abundance.

Stable-isotope dimethyl labeling

Stable isotope dimethyl labeling is a fast and relatively inexpensive chemical labeling method in which the primary amines of the peptide are converted to dimethylamines via reductive amination using formaldehyde and cyanoborohydride [133], a process which is fast and generates no significant side products [134]. This method provides a single label for all the peptides upon arginine cleavage and two labels for peptides upon lysine cleavage, making it compatible with proteolytic peptides. Using dimethyl labeling, a mass shift of at least 4 Da can be induced between different sample via light (CH_2O), intermediate (CD_2O), and heavy ($^{13}CD_2O$) isotopomeric tags (peptide triplets). The differently labeled peptides will then be detected, quantified and compared within each sample. In this way, the run to run variation (that often arises in label free studies) can be reduced which potentially increases the quantification accuracy.

Data analysis

The output of an MS analysis contains MS^1 scans which provide information about the precursor ions and MS^n scans which can be used to characterize the structure of the analyte of interest. However, this data requires extensive processing to turn the raw data into biological information. To do so, specialized software programs are commonly required. Commercial programs are attractive solutions, as they provide user-friendly environments as well as a robust behavior (e.g. by preventing wrong parameter selection). Whereas open source programs offer more flexibility in terms of possibilities to modify existing algorithms. A careful selection of proper programs for data processing is crucial, since different programs have been shown to produce different and, in some cases, contradictory results [135, 136]. However, irre-

spective of the selected solution, there is a number of processing MS analysis steps that needs to be performed. These steps are illustrated in figure 4 and will be explained more thoroughly in following sections below.

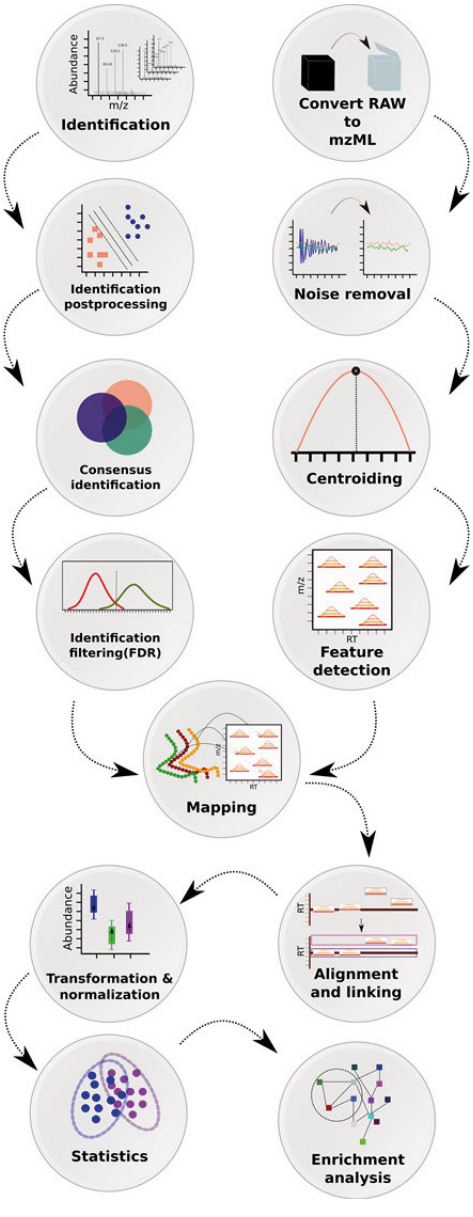

Figure 4. A typical workflow of downstream analysis of MS data using open source tools. For quantification, the MS files (vendor specific format) are first converted to an Open source format (e.g. mzML). The data is then reduced through several steps (noise and background reduction, centroiding, feature detection). Simultaneously, the identification is performed on raw data using several search engines and the result is aggregated (merged). The identification result is then mapped to the feature detection result. The retention time shift is corrected and the corresponding features across the samples are matched. The abundances are log transformed and normalized. Finally, statistical and enrichment analyses are performed.

Data conversion

The output of MS instruments is normally encoded in a vendor specific format. However, the majority of the available open source tools require the data to be in open source format. The most commonly used software to convert the data is called "msconvert" from ProteoWizard [137]. ProteoWizard has collected libraries from several vendors and use them to convert the data into the mzML format which is readable by various open source tools.

Pre-processing

Mass spectrum are normally pre-processed in order to filter out non-relevant signals from the true signals. The output data from MS is often affected by the presence of a high level of noise which can originate from e.g. the electrical system or from other unwanted chemical sources. This can cause arising of the baseline and extraneous peaks [138], potentially producing bias in peak detection.

Noise filtering and background reduction

Typically, a filtering method such as Gaussian filter or Savitzky Golay filter is applied, in which a local polynomial is fitted through the subset of the data and the central point of the fitted polynomial will be determined as the smooth value [139]. In addition, there are several methods for performing baseline reduction [140, 141], with top-hat filter being one of most well-known methods in which the signal is subtracted from its morphological opening.

Centroiding

Mass spectrometry data can be collected from the instrument in either profile (a collection of signals) or centroid mode (as discrete signals) which is performed by the vendor software. Although the profile mode data contains more information about the actual peak shape, the majority of algorithms have been built to work on the centroid mode as it has significantly smaller size and potentially less noise compared to the profile mode [142]. Beside the vendor specific software, several other methods have been developed to perform data centroiding, such as peak picking using wavelet technique [143] or cubic spline interpolation [144]. The choice of algorithm is dependent on the instrument in use and experiment specific properties [145]. These algorithms attempt to find cluster of related signals which are then aggregated to one peak, therefore significantly reducing the size of the MS spectrum as well as increasing the signal to noise ratio. In **Papers I, II, III and V**, we used a method based on cubic spline interpolation [144] to perform centroiding of the MS data.

Feature detection

Another data reduction step in MS pre-processing is called feature detection. Feature detection refers to the process of quantifying the signals generated by an ion in a region of the LC MS map. As peptides exhibit specific patterns in the m/z and retention time dimensions (isotopic pattern and elusion profile), characterizing these patterns will give an estimate of peptide amount in the samples. Although there are several methods to perform feature detection, most of the algorithms create mass traces (consecutive scans potentially related to one peptide) and combine co-eluting mass traces which show a plausible isotope pattern [146] to create a feature for a peptide (some algorithms might perform these steps simultaneously [144, 147]). Several curve fitting methods can be applied on both elution and isotopic pattern to find the features as well as to resolve conflicting features (e.g. overlapping mass trace). In **Papers I, II, III and V** we used FeatureFinderCentroided [144] tool from OpenMS [148] to detect and quantify the features in the MS runs. In **Paper IV**, the feature detection was performed by a commercial software, DecyderMS (GE healthcare), which to the best of our knowledge treats the MS spectrum as an image and performs image recognition to find the features.

Retention time correction

Retention time (RT) drift is another source of variation which occurs specially when using LC to separate the peptides prior to detection. This phenomenon causes a shift (linear or nonlinear) in the retention time and the elution profile of a peptide across different MS runs. This shift can result in missing values across the runs as well as deviation in feature intensities. The most common causes of RT variability are changes in mobile phase composition, temperature, stationary phase surface of chromatography column as well as irreversible binding of analyte components to the surface of stationary phase. Several alignment methods have been developed, such as algorithms based on total ion chromatograms [149], parametric time warping [150], mass bind based alignment [151] and alignment based on landmark selection [152]. These alignment algorithms are either based on the profile/centroided raw data or the processed feature data. Based on retention time information of a set of data points (using algorithm specific parameter), the algorithms fit a model [144], by use of pairwise distance [153] or clustering methods [154] to compute a transformation which maps all the data to a common RT scale, thus correcting the shift and distortion between MS runs. In **Papers I, II, III and V** we used the MapAlignerIdentification [144] tool from OpenMS [148] to perform the alignment. In **Paper IV**, the alignment was performed by commercial software, DecyderMS (GE healthcare).

Feature grouping (peak matching)

To be able to compare the abundances across the samples, the quantified features must be matched in the way that a group of features represents the abundance of e.g. a peptide across different samples. There are multiple clustering methods to perform the matching, such as MultiAlign [155], quality threshold (QT) [144] and XCMS specific matching algorithms [156]. However, in principle, these methods attempt to group the corresponding peaks across the sample given a user or automatically defined boundary (m/z and RT) and derive a score which can be used to reduce the data into a set of reliable matched features.

Each of the mentioned steps can be performed using a different set of tools e.g. DeconTools (https://omics.pnl.gov/software/decontools-decon2ls) for peak picking, and MassSpecWavelet [157] for performing peak detection. However, it is often preferred to perform the analysis using software frameworks or packages such as OpenMS [148], MaxQuant [158] to avoid format conversion for each step of the MS analysis. In **Papers I, II, III and V** we used OpenMS [148] to perform the preprocessing and in **Paper IV** we used DecyderMS to perform label free and MSQuant [159] to perform labelled preprocessing.

Identification

The identification of the peptides is performed using the combination of the m/z of the precursor ion and the fragmentation data of the precursor. Fragmentation with low energy dissociation normally results in breakage of CO – NH bond between amino acids (although with higher energy, breakage at CH – CO and NH – CH is also possible), resulting mainly in b- and y- type ions [104]. The b ions occur when the charge is retained on the N terminal part of the peptide after the fragmentation (therefore the N terminal part of the peptide is preserved). Whereas the y ions extend from the C terminal where the charge is located after fragmentation (the N terminal is lost). Since the mass difference between the successive b and y ions corresponds to mass of a single amino acid residual, peptide sequence can be inferred and assigned to the peaks by calculating the mass difference. This approach is referred to as *de novo* sequencing and can greatly enhance the reliability of the resulting identifications especially when a sequence database is not available. However, this approach requires presence of dominant peaks in the spectrum (with high signal to noise ratio) and can also be computationally slow when applied on thousands of spectra. However, with the development of new algorithms [160, 161], this approach can be regarded as a viable choice, assisting the typical database searching algorithms.

Database searching is perhaps the most common way of performing identification. Typically, a protein database is provided to the algorithm which performs *in-silico* digestion and fragmentation on the sequence database.

The resulting predicted spectra will be compared against the experimental spectra to find and score the matches. The search engines normally use the precursor mass with a defined mass window to limit the potential matches. The predicted MS^2 spectrum from the limited list will then be compared against the experimental spectrum, taking into account several criteria for scoring the matches such as precursor mass and product ion tolerance, enzyme specificity, missed cleavages, number of charges, and post-translational modifications. There are multiple tools to perform the identification of peptides such as MASCOT [162], SEQUEST [163], PEAKS DB [161], OMSSA [164], X! Tandem [165] and Andromeda [166]. Although the general principle of these algorithms is similar, the methods for handling the peaks (peak selection), searching and scoring the peptide can cause differences in the resulting peptide identifications [167]. In order to provide a measure of false positives, false discovery rate (FDR) is often calculated for the matches in which the search is performed on a decoy database (e.g. reversed or shuffled sequences of proteins) and the target database at the same time. The ratio between the false (decoy) matches and the true (target) matches is regarded as FDR (higher than a specific score). Various cutoffs can be used (e.g. 0.05 or 0.01 translating to 95% or 99% correct matches, respectively) to accept a certain FDR for the hits [168, 169]. However, as multiple scores can give rise to the same FDR as well as the issue of not being monotonically increasing with the matches scores, often a q-value is preferred and assigned to each spectrum with certain score, meaning that for a specific database match there is a certain number of highly ranked matches that are likely to be wrong (e.g. 1 in 20 in the case of 0.05). The q-value can be interpreted similar to p-values therefore providing statistically sound measure to apply different cutoffs on it [168]. In the **Paper I and II** we used OMSSA and X! Tandem identification results. In **Paper III and V**, we used MS-GF+ [170] and in **Paper IV** we used Mascot search engine.

Another aspect of identification is to reconstruct the proteins from the peptides providing an identification confidence value for proteins. As mentioned before, for peptides matches, FDR values based on the target/decoy approach are normally reported. However, since no protein level information is present, the protein FDR must be estimated from the list of peptides [171]. The most commonly used protein inference tools are ProteinProphet [172], Fido [173] and MSBayesPro [174]. The in-depth discussion of these algorithms is however well beyond the scope of this thesis. Briefly, the identification results are first post-processed (e.g. with Percolator) to assign probability to the PSMs. The PSMs are then assembled to proteins, considering the proteins that are matched by one peptide as well as the peptides that matches to multiple proteins. Finally, the probability that a protein is present in the sample is based upon the evidence of corresponding peptides. We used Fido to perform protein FDR estimation in **Papers III and V**.

Combined identification

Since different search engines can produce different (sometimes conflicting) outputs, the result of multiple search engines can be combined using various algorithms [175]. Doing so, the identification result can be potentially enhanced by filtering the conflicting results as well as using hits that are found only by a number of engines [176] (but not all) and therefore increasing the number of matches [177] (especially if some search engines report more than one peptide matches for each spectrum). At the first sight, using the best score of any search engine seems like an obvious choice. However, aggregating the identification result requires a similar score type generated by all the search engines. Algorithms, such as PeptideProphet [178] have been developed to transform the original search engine scores to a probability score which is comparable across the search engines and is less subjective and prone to error compared to introducing cutoffs on the original scores. Since different search engines produce different score types (and perhaps multiple scores per search engines e.g. Xcorr and ΔC_n in SEQUEST [163]), the algorithm first combine multiple score type to a single hyperscore or a discriminant score. Two distributions will then be fitted through the data (e.g. a joint distribution of Gaussian distribution for the correct scores and gamma distribution for the incorrect scores) and the probability is calculated using Bayes' law (e.g. the ratio of probability of having a specific score and being a correct hit and the probability of having a specific score). This probability value can then be regarded as comparable across the search engines and to combine the result of identification using many algorithms such as SeqSim (using similarity of peptide sequences via identity matrix or PAM30MS substitution matrix), shared peak count (via calculating the ion ladder similarity) and average scoring of several search engines [177]. The resulting similarity score will then be aggregated to a consensus score, using the probability score of each search engine and by taking the similarity score from the other engines into account. The resulting identification should then be assigned to the quantified features (in case of MS^1 quantification), facilitating the downstream analysis on the peptides. We used the similarity score approach in the **Paper I and II** to combine OMSSA and X! Tandem identification results. In **Paper III and V**, we used MS-GF+ and in **Paper IV** we used Mascot search engine without combining the identification from other engines.

Post-processing of identification

Most of the mentioned identification engines use multiple orthogonal scores to discriminate between correct and incorrect peptide-spectrum matches (PSM) identification [179], thus independent scores can reduce overlap between correct and incorrect PSM assignments. The idea behind the post-processing of identification results is to consider the combination effect of

these to re-rank the list of PSM and assigned probability based on the newly formed list. Tools such as Percolator [180], PeptideProphet [178] and Nokoi [181] use machine learning algorithms to improve the separation between correct and incorrect PSM based on several features normally reported by identification engines (e.g. mass error, variable modifications etc.). These methods often result in improved numbers of confident peptide identifications compared with using original independent scores (at identical q-value). We used Percolator in **Papers III and V**, to post process identification results.

Quantification of proteins

Shotgun proteomics experiments is based on identification and relative quantification of peptides. However, scientists typically have interest in the absolute concentration of proteins. Since MS does not inherently measure absolute concentration of peptides, statistical methods have been developed to estimate abundance of proteins based on peptide information. Some of these methods include exponentially modified protein abundance index (emPAI) [182], the average or summation MS intensity for the three (or more) most intense peptides [183] and normalized spectral index (SI_N) [132]. Although the mathematical definition of these approaches is different, they share the same basic idea, that is: as the amount of protein increases, the number of identified peptides or quantified peptides also increases. Briefly, emPAI calculates the number of observed peptides divided by the number of observable peptides per protein which is then multiplied by ten and subtracted by one. The second approach, aggregates the abundance of the most intense identified peptide features for each protein either by summation or taking the average. Finally, SI_N calculates a protein abundance by considering the number of peptides as well as number of tandem spectra and their intensities. All the mentioned methods have been reported to result in acceptable correlation to the actual protein concentration. However, the selection of proper method depends on various parameters such as protein amount in the samples or MS instrument settings. We chose to quantify the proteins using the summation/average of MS intensity in **Papers III and V**. Whereas in **Paper I, II and IV** we performed the statistical analysis on peptides without aggregating them into proteins.

Normalization

Although the aforementioned steps greatly reduce the noise, the MS experiments are still susceptible to systematic bias. While, the relative abundance of peptides/proteins are considered to be result of biological changes, systematic bias originates from nonbiological sources including sample preparation and instrumentation [184]. The systematic bias results in incomparability of the samples and must be taken into account to follow a reliable statisti-

cal analysis. The process by which this unknown systematic variation is removed is referred to as normalization and can be performed by various methods including global normalizations: median normalization, cyclic loess normalization [185] and local normalization methods using internal standards [186]. The choice of most of the global normalization methods are based certain assumptions [187] such as 1) The different samples are comparable especially with regard to their protein amounts. 2) Only a few proteins have different abundances between the sample groups. 3) The distribution of abundance ratios between different sample groups is symmetric. However, the local normalization using internal standard is typically based on the assumption that the variation in internal standards is due to systematic bias. In any case, the choice of normalization method is based on experience and validity of the assumptions. In **Paper I, III, and IV** we used median normalization whereas in **Paper V** we used, cyclic loess normalization. In **paper II**, we evaluated three normalization methods, including median, reference and a method based on internal standard.

Univariate statistical testing

Considering a typical biomarker discovery experiment, at this stage of the analysis the peptide intensities are ready for performing statistical testing. The testing can be performed on aggregated abundances or directly on the peptide level intensities using any statistical testing methods (e.g. t-test or Mann–Whitney U test) to derive a p-value for the respective peptides or proteins. Although regular statistical testing can suit straightforward experimental designs, more complicated experimental designs (e.g. time series and experiments with many experimental conditions) cannot be handled by these methods. Regression models can help handling increasingly complicated experimental design. However, these methods are also limited as they have fewer number of observations needed to converge towards the true underlying variance. As a solution, approaches similar to Linear Models for Microarray Data (limma) [188] are often used to fit multiple linear models to the peptides and calculate statistic of interest. Since these approaches use regression analysis, they allow for complex experimental design but also provide the possibility that the sample variance can be adjusted towards the expected variance (using information from other proteins) to increase the precision of the estimate thus partly compensating for small sample size. In **Papers I, II, III, IV and V** we used limma for performing statistical testing.

Multivariate statistical analysis

Multivariate methods provide a framework to make use of all the protein abundance simultaneously to extract the effect which is the result of the relationship among proteins (using for example covariance or correlation). Prin-

cipal component analysis (PCA), partial least squares discriminant analysis (PLS-DA) and a closely related method (sparse PLS-DA) [189] as well as learning methods for classification such as random forest [190] are among commonly used methods to extract the variation in the data. Briefly, PCA and PLS-DA seek the projection of the original data on a lower dimension while preserving most of the information. However, PCA achieves this in an un-biased manner (without class information of samples) whereas PLS-DA incorporates information about class of samples in the analysis. Random forest on the other hand, attempts construct many decision trees from different subset of the data and make the prediction by averaging the predictions from individual trees. We used PCA in **Paper I, II, V**, (sparse) PLS-DA in **Paper III and VI** and feature selection based on random forest in the **Paper V**.

Pathway and enrichment analysis

Although finding the proteins with statistically significantly altered levels and validating them can be regarded as the main goal of biomarker discovery, one can perform pathway enrichment or gene ontology (GO) analysis to obtain a global overview of the pathways associated with the altered proteins. These methods typically work by deriving a p-value by e.g. Fisher's exact or Chi-square test, showing whether the overlap of functions represented in a target protein set with the functions in a background set is statistically significant. In practice, there are many databases providing tools and certain level of abstraction in GO for performing such an over-representation analysis (e.g. KEGG [191], STRING [192], and Gene Ontology (GO) Consortium [193]). In **Papers I and IV** we used STRING [192] to perform the enrichment analysis.

Cloud computing for omics data analysis

Cloud computing is generally defined as on-demand network access (e.g. over the internet) to a shared pool of computer resources that can be easily available and released upon request. Cloud computing offers several advantages over traditional HPC systems, including: 1) users can configure their own computers resources (e.g. operating systems and required tools). 2) The resources can be accessed via the internet from virtually any places with multiple devices. 3) Cloud computing offers dynamic resources, meaning that the users can move the tools to stronger computers (scale up) or simply add more computers to their working environment (scale out).

Cloud Computing architecture is broadly classified into front-end and back-end. The front-end and back-end are connected to each other through a network (the internet or virtual network). The front-end comprises the user's devices and the applications (such as interface) needed to access the

cloud system. The back-end consists of the resources required to provide the cloud computing services that include computers, data storage, protocols, management software (to configure the infrastructure), hypervisor (to create virtual computers), network (to communicate with and within the cloud system) and more.

Cloud computing can be divided into four models: private, public, hybrid and community. Private cloud is normally delivered for example within an institute to internal users, offering additional security and privacy as well as more control over the cloud system while at the same time preserving other benefits that cloud computing provides. In the public cloud, the services are delivered by a third-party cloud service provider. This is a business model which the resources can be purchased on demand. However, this is normally offered as pay-as-you-go computing where users only pay when the resources are used. The hybrid model combines public and private cloud such that the critical analysis such as the ones concerning personal data can be performed on the private and the rest on the public cloud, offering benefits of both models. The community model is closely related to both private and public cloud. In this model several for example institutes (or organizations) offer the cloud system to be shared within and between these institutions.

Regardless of model, the cloud services are normally provided in three different forms: infrastructure as a service (IaaS), platform as a service (PaaS) and software as a service (SaaS). In IaaS, infrastructure components such as computers, storage and networking are provided and managed over the internet. In PaaS on the other hand, the operating systems and development tools are also provided and managed for the users. Finally, in SaaS, the software (such as a web application) is provided and can be accessed over internet. Each of these forms offer certain benefits such as fully controlling the computing resources (IaaS), agility and efficiency in developing applications rather than focusing on managing the infrastructure (PaaS) and finally, reduced time and cost for managing an application (SaaS).

Various models of cloud computing are extensively used in omics. An example can be MCW proteomics tools [194], ProteoCloud [195], Trans-Proteomic Pipeline [196], XCMS ONLINE [197], Chorus (chorusproject.org), The Metabolomics Workbench (www.metabolomicsworkbench.org) and many other applications [198-201]. All of these tools provide virtual environments that scale with computational demands and ease of access to certain applications. As pointed out, the large number of required applications in omics analysis has necessitated integrative workflow engines (such as KNIME [202] and Galaxy [203]) that enable the assembly of analysis modules to form computational workflows [204]. However, many of the mentioned cloud based applications provide limited flexibility in terms of building and using standardized and customizable workflows. Furthermore, customization and installation of these applications require high level of

expertise in cloud computing discipline. In this context, microservices provides an easy and more portable way to deliver these applications to the users.

Microservices

Microservice architecture is an approach in developing software systems in which complex tools are divided into smaller services, each of them is isolated and loosely-coupled units that can be executed independently in its own process. These services are portable, started on-demand and can communicate with other microservices using agnostic communication protocol. The microservice architecture offers several benefits including, highly focused (increased productivity especially for large development teams) and modular design, independent initiation and execution (autonomous behavior) and technology agnostic (can use multiple programming languages, development fragment, etc.). Microservice architecture is not a new concept but with advent of containerization it has gained popularity within the software engineering community. Containerization allows an application to be wrapped in a dedicated environment that includes all the dependencies such as files, environment variables and libraries. Containerization works by virtualization of the host operating system such that the system resources can be shared between different independent applications. More specifically, the traditional virtual machines (a software that emulates one or more computers) in which each virtual computer runs its own kernel and requires static allocation of hardware resources. The virtual machines tie up the resources regardless of whether they use it not. Containers, on the other hand, use resources in a shared manner, eliminated the need of fixed allocation of resources.

Creating and running of containers rely on container manager software which access different interfaces on the host kernel to access virtualization features (e.g. Docker [205], LXD and CoreOS rkt). Using the microservice architecture and containers, different independent tools can be chained together to make a full computation workflow that can efficiently scale in the elastic cloud environment. Although running a few containers on local computers using container manager is possible, as the application becomes more complex and the size of datasets grow, automated arrangement and management of software containers as well as interfaces to run them will be needed. In **Paper VI**, we use Kubernetes container orchestrator to administrate the containers (e.g. initiation and scheduling). The Helm package manager (https://github.com/kubernetes/helm) was used to deploy the jobs and services. Since Kubernetes does not cover provision of for example virtual machines, KubeNow (https://github.com/kubenow/KubeNow) was used to create the backbone of the e-infrastructure (e.g. file system storage, networks, configuration of DNS, etc.). The Galaxy workflow engine [203], Luigi [https://github.com/spotify/luigi] and Jupyter Notebooks [206] were

adapted to run on top of the e-infrastructure as the main interfaces. In addition, various metabolomics software packages were included, as for example: for MS analysis: OpenMS [207], XCMS [156], CAMERA [208], MetFrag [209] and ropls [210]; For nuclear magnetic resonance spectroscopy (NMR): "rnmr1d" (https://github.com/phnmnl/container-rnmr1d); For fluxomics: RaMID (https://github.com/seliv55/RaMIDcor), MIDcor (https://github.com/seliv55/midcor), Iso2Flux (https://github.com/cfoguet/iso2flux) and Escher Fluxomic [211]. All these tools were containerized using Docker. The continuous integration of all tools were ensured using Jenkins [212]. For code deposition, GitHub (https://github.com) was used. Finally, the tested containers were deposited into Biocontainers [213] and a private repository (figure 5). This resulted in (**Paper VI**) providing a novel e-infrastructure to set up a complete software suite that can be used to create and execute scientific workflows using containers, offering substantial flexibility and scalability to perform omics analysis on the cloud. The e-infrastructure is available through Phenome and Metabolome aNalysis (PhenoMeNal). Furthermore, this e-infrastructure was adapted and used to perform the pre-processing steps in **Paper V**.

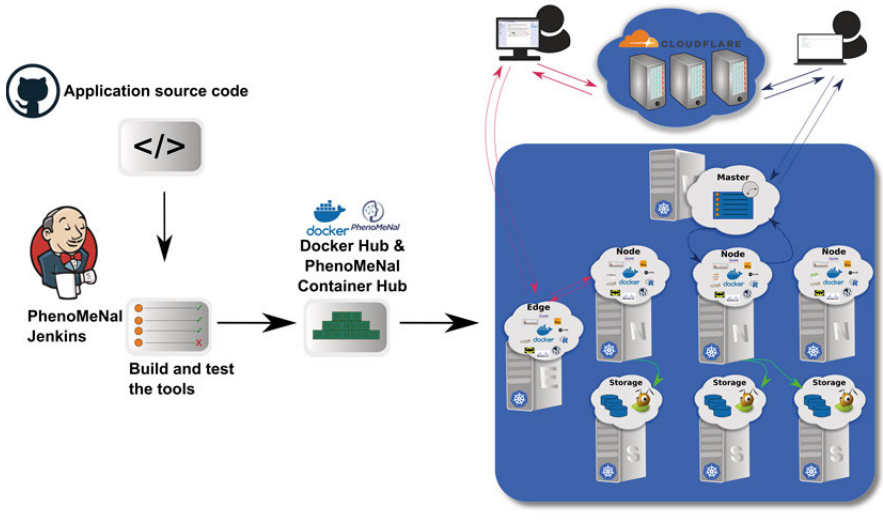

Figure 5. The components of the microservice structure. The applications source codes are deposited in GitHub and then automatically built and tested by Jenkins and will be available in multiple repositories. The Users can interact with the system through API or graphical interface to create computational workflows containing several jobs. The jobs in the workflow are then submitted to Kubernetes which performs the container orchestration.

Validation of mass spectrometry findings

Although MS provides accurate information about the analytes in the samples with a wide dynamic range, it might encounter a sensitivity problem when e.g. a protein is present at an extremely low concentration. In this case, quantification using MS might not provide as precise information as it does for proteins with higher abundance. Antibody-based validation is often used as an orthogonal technique to methodologically validate the MS findings. However, antibody-based experiments are not inherently discovery based. Therefore, using MS to perform the discovery phase of the study and antibody to perform the verification of the findings can provide deeper and more reliable inside into the biological system. In this work, we used membrane-based western blot (**Paper I**) and bead-based assays (**Paper II and IV**) to validate MS results.

Western blot

Western blot is a semi-quantitative technique used to detect specific target proteins in a given sample. Western blot uses SDS-PAGE to separate the proteins based on their molecule size and structure. The proteins will then be blotted onto a membrane and stained by specific antibodies. The unreacted sites on the membrane will be blocked to reduce unspecific binding of proteins. Afterwards, the primary antibody will be incubated with the membrane to bind to the proteins of interest. A secondary antibody (enzyme conjugated) is then incubated with the membrane and finally, a reaction with the enzyme bound to the secondary antibody will generate protein bands. The intensity of the protein band can be quantified by various image processing algorithms, facilitating the comparison of the protein of interest across several groups. Thanks to the high resolution of gel electrophoresis and sensitivity of immunoassay, protein amounts as low as 1 ng can be detected and quantified using western blots [214]. However, the western blot experiment should be carefully controlled by e.g. adding positive, negative, blank, and loading controls as well as validating antibodies [215] (figure 6).

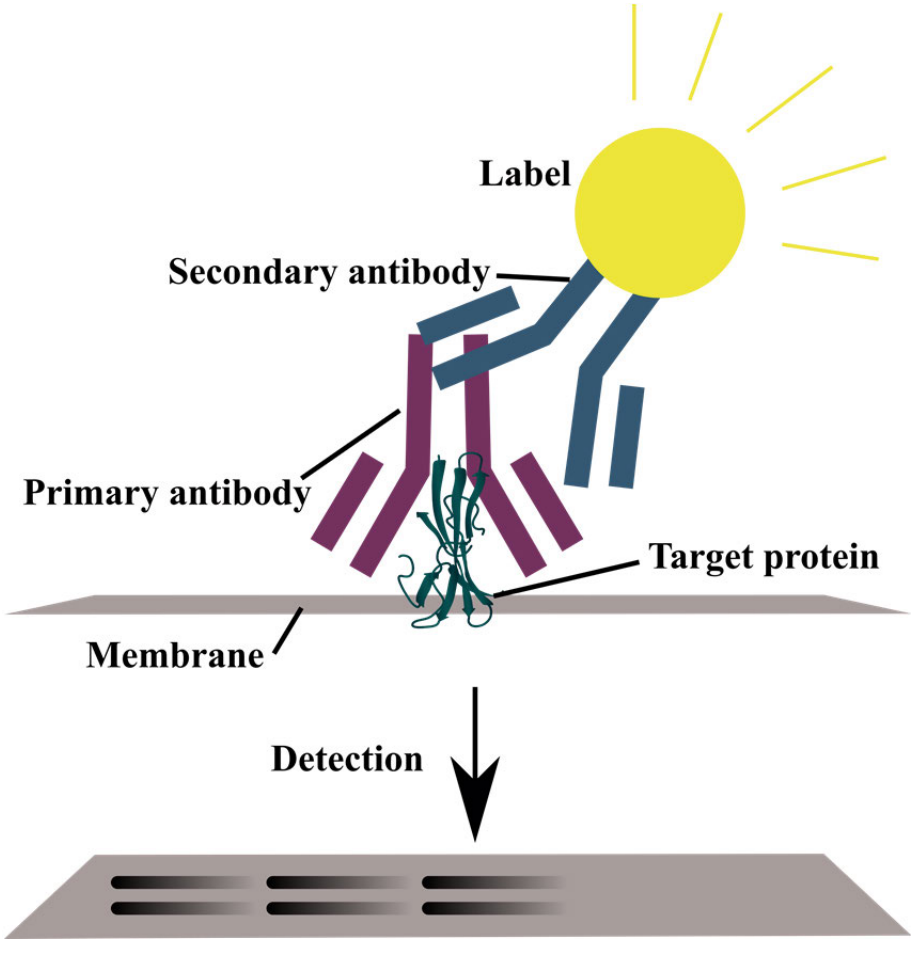

Figure 6. Overview of western blot. The proteins are blotted onto the membrane and are capture by primary antibodies. The labeled secondary antibody will be added to capture the primary antibody which is used for detection.

Bead-based multiplex assay

Multiplexing refers to the process of detecting several analytes in a single cycle of an assay. Bead-based assays are among the most recent methods of performing antibody-based multiplexing [216] which can provide simultaneous detection of several proteins in a small volume of material as well as providing statistically more accurate quantitative information about the proteins through the use of data from a great number of beads. The basic principle of this method is to color-code (ratio of red and infrared fluorescent dyes) numerous beads for each analyte of interest in which a resulting color-coded bead will be specific for an analyte. The beads coated with analyte specific capture antibodies are mixed and added to a well of say a 96-well plate in which they bind to the protein of interest. Fluorescence labelled reporter/detecting antibodies, which bind the capture antibody, will then be

added. The beads will be scanned by two lasers where the first laser reads the beads color-code and the second laser detects the amount of the fluorescence labelled reporter antibody, providing fast and accurate quantitative information [217] (figure 7).

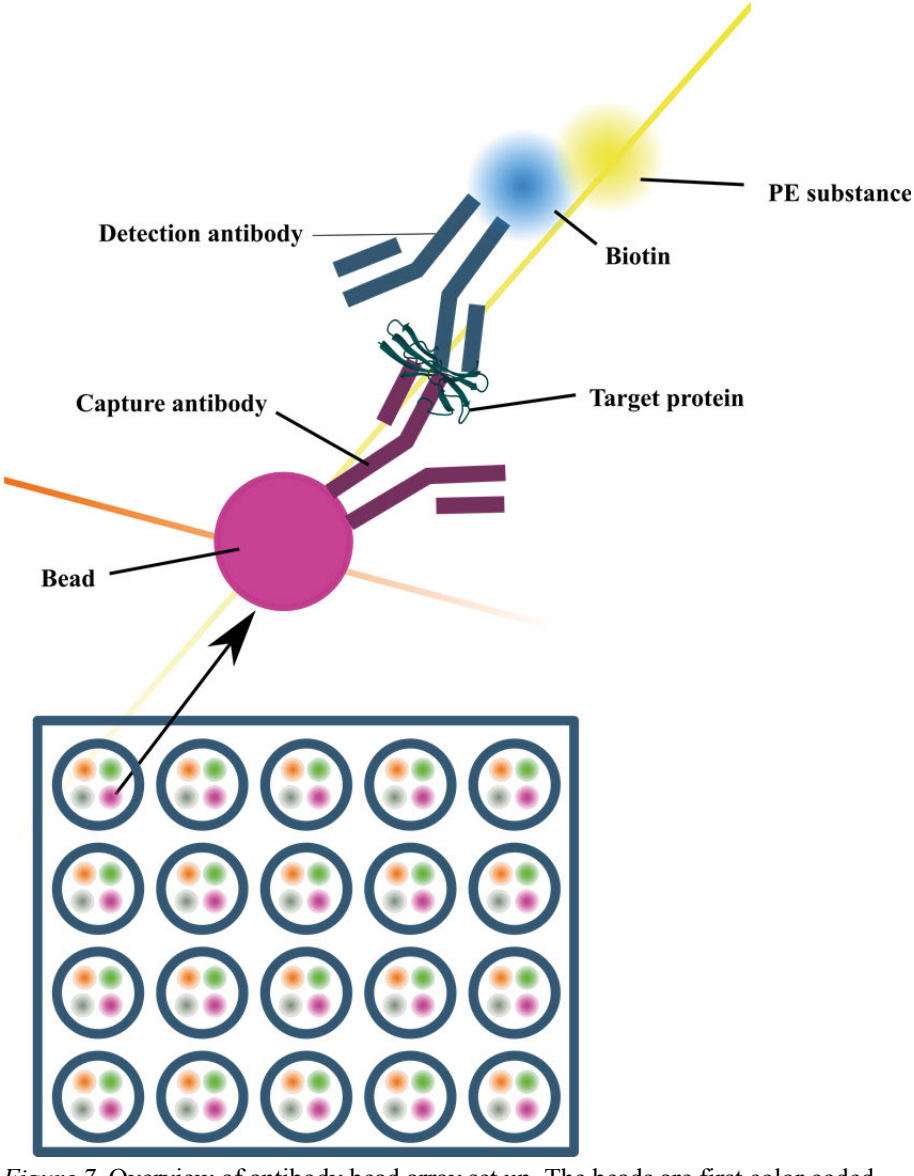

Figure 7. Overview of antibody bead array set up. The beads are first color coded and then coated with capture antibody. The detection antibody attached to PE (Phycoerythrin) substance will then be added. Two lasers will read the code of the bead as well as the magnitude of the PE-derived signal.

Papers I-VI

Aims

Paper I: To perform a proteomic profiling of *post mortem* AD brains and compare it with non-AD brains as well as with brains from other neurological diseases to find proteins involved in AD pathology.

Paper II: To detect novel protein CSF markers that can distinguish between AD and healthy elderly controls. To evaluate the impact of software selection and normalization methods upon the results.

Paper III: To detect novel protein CSF markers that can be used for early detection of AD and compare the result to that of traditional AD biomarkers.

Paper IV: To analyze the biological changes produced by SCS in CSF of neuropathic pain patients and find proteins potentially involved in pain signaling.

Paper V: To find new potential biomarkers in chronic pain state, fibromyalgia, with the aim of facilitating more accurate diagnosis of this disease.

Paper VI: To provide a flexible and scalable framework that can be used to perform large scale metabolomics data processing using microservice architecture.

Tissues and biofluids

In **Paper I**, temporal neocortex brain samples from ten ADs, five controls and six cases with other neurological diseases (OND) were included. The AD brains were selected based on Braak staging criteria [218], reflecting the progress of the disease process when it comes to Aβ and tau distribution in different brain regions.

In **Paper II**, CSF samples were collected by aspiration lumbar puncture from ten AD patients and ten healthy control samples. The patients were selected based on several criteria including cognitive examinations, brain

imaging and routine laboratory testing. The clinical characteristics of AD patients were: average Aβ42 420±117, tau 652±376, and p-tau 132±112 (mean±SD ng/l).

In **Paper III**, CSF samples from 76 AD patients, 71 MCI patients and 45 controls were included. The available clinical follow up patients were used to stratify the MCI patients into three groups: the ones that were followed for approximately 2-6 years and were finally diagnosed with AD (n=10), the patients that remained MCI based on the latest follow up information (n=14) and patients that were not followed (n=47). The patient selection criteria were similar to that of **Paper II**. The clinical characteristics of patients are summarized in table 1.

Table 1. *Clinical* characteristics *of patients and controls*

Group	Aβ42 ng/l (Median[range])	tau ng/l (Median[range])	p-tau ng/l (Median[range])
AD	405 (160 - 1160)	617 (160 - 1720)	82 (28 - 220)
MCI/no follow-up	702 (240 - 1500)	285 (112 - 1370)	48 (22 - 184)
CONTROL	676 (337 - 1343)	414 (202 - 1121)	63 (29 - 122)
MCI/AD converter	404 (350 - 639)	604 (358 - 998)	73 (35 - 114)
MCI/non-AD converter	918 (234 - 1360)	321 (82 - 650)	51 (24 - 86)

In **Paper IV**, CSF samples from 14 neuropathic pain patients before and after SCS were collected. The patient selection was based on grading system by Treede et al. [219] with indication of neuropathic pain and lesion confirmed by MRI and neurophysiological investigation. Further inclusion criteria were 18–70, ≥3 months since SCS implantation, BMI of 19–28 kg/m^2 and having no major other health problems.

In **Paper V**, CSF samples from 39 FM patients and 38 controls were included. The FM samples were females of 47.2±9.2 age and the controls were five females and 33 males of 47.9±14.4 age. The FM patients were selected based on 1990 criteria of the American College of Rheumatology (ACR). The patients were excluded if verified with organic cause of the pain or having a record of anxiety disorders, psychotic disorders, dementia, epilepsy, seizure disorders, alcohol or drug.

In **Paper VI**, we did not aim to investigate any specific conditions or biological question. However, the following samples were used in the demonstrators in the paper: 44 samples from human renal proximal tubule cells in the demonstrator 1, 37 CSF samples from multiple sclerosis patients in the demonstrator 2, 132 urine samples of type 2 diabetes mellitus in the demonstrator 3, and samples from HUVEC cells under hypoxia in the demonstrator 4.

Results and discussions

Paper I

Reduced activity of exocytosis and endocytosis in AD brain

In order for a cell to communicate with other cells as well as to import required substances and remove wastes from inside, specialized transport processes must be used. Two of these processes are referred to as exocytosis and endocytosis. In this study, using brain tissue, we found that out of 163 significantly different proteins between AD and control, 121 proteins were associated with regulation of vesicle mediated transport, synaptic transmission, exocytosis and endocytosis processes and SNARE binding. Out of the proteins involved in endo-exocytosis processes, altered levels of VAMP2, syntaxin-1A (STX1A), synaptosomal-associated protein 25 (SNAP-25), clathrin heavy chain 1 (CLH1) and adaptor protein complex AP-2 subunit alpha-1 and alpha-2 (AP2A1 and AP2A2) were reproduced in the replicated experiment or verified by antibody bead arrays. These proteins are the major component of the endo-exocytosis pathways and altered levels of them will most likely result in disturbed activity of the vesicle mediated transport pathway. In the brain, astrocytes use endo-exocytosis to secrete for example gliotransmitter (to facilitate neuron-glial communication), such as glutamate and d-serine, which have been reported to be both increased and decreased in AD brain [220-222]. Microglia, on the other hand, internalize soluble Aβ peptides via endocytosis and facilitate transmission of tau between neurons via exocytosis in AD brain [223-226]. Endocytosis and exocytosis are also vital for the neurons to maintain neurotransmitters release and vesicle recycling which have been shown to be altered in AD brain [227-229]. The cell origin of this observation (altered endo-exocytosis) is not clear to us. However, the diminished activity of endo-exocytosis pathway can ultimately alter communication between neurons as well as other glial cells in the brain through reducing secretion of for example neurotransmitters in presynaptic side and uptake of these molecules in the postsynaptic side of the synapse.

Elevated levels of proteins of extracellular vesicles in AD brain

Emerging evidence indicates that extracellular vesicles (EVs) are important to maintain efficient intracellular communication and homeostasis in the brain [230]. The EVs provides invaluable information for biomarker discovery as their content depends upon the parent cell [231]. In this study, we found that a majority of the proteins (n=41) displaying elevated levels in AD brain were associated with EVs. More specifically, we detected and verified an increased level of CD9 (an extracellular vesicle marker) in AD compared to controls. In addition, we observed and confirmed increased levels of HSP72 and TALDO, which both have been associated to extracellular vesi-

cles. The proposed roles for EVs in AD has been very different depending on the context. For example, they have been shown to carry enzymatically active neprilysin which can degrade Aβ in the brain [232]. They have also been shown to induce synaptotoxicity and Aβ assembly [233]. On the other hand, EVs from CSF have been reported to contain proteins derived from the brain as well as high amount of tau and p-tau [234], which is why they may have the potential to serve as biomarkers. Altogether, these observations might signify that the EV content might be affected in AD brain, making them very interesting targets both for drug and biomarker discovery.

Paper II

Quantification and identification comparisons

As described, MS data processing and analysis involve using multiple quantification and identification components. These components are offered by many different software packages, making the tool selection a challenging task. Further problems arise when the result of applying these tools are different when it comes to quantification and identification [235, 236]. However, when tested for reproducibility of the quantification in our study the five programs (PEAKS, MaxQuant, OpenMS, Sieve and DecyderMS) were found to be very similar and to have high correlation with each other. The main difference between the programs was found to be identification performance. We found that the PEAKS program identified and mapped considerably higher number of peptides as compared to the other programs, mostly due to combination of *de novo* sequencing and regular database searching. Despite identifying substantially higher number of peptides using PEAKS, the number of identified proteins was similar to the other programs. We found that PEAKS search engine can characterize each unique protein by multiple peptides. After PEAKS, MaxQuant, OpenMS, Sieve and DecyderMS, respectively identified more peptides. Aside from the peptide identification, most of the proteins were identified by one program, implying that to gain appropriate peptide/protein coverage, multiple programs can be employed for the identification. It is worth noting that although the quantification performances were similar, the difference in identification can at the end lead to different protein quantification as it affects the selection of peptides to be aggregated to protein abundance. Therefore, a plausible approach would be to have a method that enable combination of quantification and identification from different tools. Finally, we noted that many of the peptides were in fact identified by most of tools but were assigned with low scores. Thus post-processing methods such as Percolator [180] can potentially further improve the overlap of the identification between different engines.

Low-abundance proteins in CSF of Alzheimer's disease patients

We observed that the total protein amount in the depleted CSF fraction (the flow-through fraction) showed a statistically significant decrease (p-value = 0.046) in AD patients (n = 10) as compared to controls (n = 9). Moreover, the fraction between unbound and bound proteins was 24.8±5.5% and 17.5±2.4% in the depleted CSF fraction for the control and AD patients, respectively. The difference between the groups was statistically significant (p-value = 0.003). No significant difference was found for the proteins captured and released by the column, indicating that approximately the same amount of proteins was depleted from both AD and controls. This difference can either be explained by the loss of several proteins in the depletion procedure, the age difference between the study groups (the elderly healthy controls were on an average nine years older compared to AD) or a general difference in protein amount between AD and non-AD. In any case, this led to violation of the assumption that the same total amount of proteins is present in the samples which further affected the downstream analysis.

Effect of normalization on biological conclusions

We started this experiment with the assumption that the total amount of proteins between different sample type (i.e. AD and controls) is similar. With this in mind, we used global normalization (e.g. median normalization) to reduce systematic variation often observed in MS data [237]. We then selected the proteins with altered abundance between AD and controls to validate using antibodies. However, we observed inconsistencies between MS and antibody results. Specifically, the measured relative differences between AD and controls for a number of proteins were found to be contradictory between MS and antibody-based technique. In addition, the correlation between two methods was not satisfactory. We found the reason to be improper normalization, due to a violation of the assumption about the symmetrical distribution of proteins with increased and decreased abundance between the groups. It is worth noting that the violation of this assumption in the investigated CSF samples was not clear before removal of the seven most abundant proteins prior to MS analysis, which were contributing to more than 75% of the total protein mass. We then performed a local normalization based on a spiked-in protein, which resulted to substantial improvement in correlation between the two methods as well as considerable reduction of fold changes comparing the AD to the C group and a more consistent result to that of antibody-based technique (compared to global normalization). This improvement was seen irrespective of the program used for the MS quantification. This observation does not suggest that the global normalization methods should be completely abandoned but rather indicates that the choice of normalization method is important and should be carefully evaluated for example using internal standards or orthogonal methods.

Alzheimer's disease associated proteins

Finally, we found eight proteins that could be verified using two technologies (MS and antibody bead arrays). These proteins included leucine-rich α2 glycoprotein (LRG), apolipoprotein M (ApoM), complement C1q subcomponent (subunit B, C) (C1QB and C1QC), complement C1S (C1S), EGF-containing fibulin-like extracellular matrix protein 1 (fibulin-3, FIBL3), receptor-type tyrosine-protein phosphatase zeta (PTPRZ) and seizure protein 6 homolog (SEZ6), all showing lower relative abundance in AD compared to controls. Additionally, using PCA on antibody bead array data, we observed that the disease signature of these eight proteins greatly matched with that of represented by MS differentially abundant peptides data, indicating reproducibly of the presented proteins using an independent method. These proteins represent a variety of different functions such as cell adhesion, migration, morphology and immune system. Furthermore, three of these proteins are clearly associated to complement activation (C1QB, C1QC and C1S). It is well known that inflammatory and immune system play important roles in AD brain [238]. The immune system in the brain is highly complex and involves expression of proinflammatory genes, leading to activation of immune cells. The complement system is part of the immune system and is involved in inflammatory response [239, 240]. In AD brain, the complement components are believed to have a rather protective role, mainly in clearance of amyloid, and their activation occurs at early stage of AD. Current evidence based on our study as well as recent research on CSF of AD patients clearly suggest the complement components to have sufficient diagnostic value to be used as novel markers in the clinic [241-246].

Paper III

Prediction of conversion of MCI patients to AD

As mentioned before, a diagnosis of AD can, at specialized centers, be made with a relatively high sensitivity and specificity. However, further improved methods for early diagnostics are needed to enable earlier therapeutic intervention in the future [247]. In particular, it is important to detect AD-specific changes already at earlier disease stages, before dementia has developed. Among such patients with SCI or MCI, the currently used CSF biomarkers have a fairly good predictive value [43]. In this study, we could show that, by combining the proteome level information and traditional markers (Aβ42, tau and p-tau), an improved diagnostic accuracy (92%) could be reached as compared to the use of the classical markers alone (proteomics: 75% and traditional markers: 83%) for detecting conversion of MCI to AD. Our approach resulted in only two out of 14 MCI non-converters being misclassified as AD but no misclassification of MCI-AD converters as non-converting MCI. Our present accuracy was in the range which was presented

by multiple studies [248-254]. However, our approach had certain benefits such as performing model training only on the data from patients without follow-up information. Furthermore, many of the mentioned reports used multiple criteria such as traditional markers as well as neuroimaging data which can be costly to perform in small clinics. Proteomics markers, on the other hand, can be set up in an assay, providing similar accuracy but quicker and cheaper diagnostic test.

Two potential biomarkers for AD

A challenge in biomarker discovery is to find molecules that are reproducible in multiple independent cohorts. According to recent reports [255], a large portion of presented biomarkers are cohort specific and thus not replicated in other studies largely due to quality of the data as well as lack stringency in statistics. In our work, we present two of our top protein candidates as potential biomarkers for AD: namely chitinase-3-like protein 1 (CH3L1), with higher level in AD compared to controls, and Neurosecretory protein VGF (VGF), with lower level in AD and more importantly MCI compared to controls. The CH3L1 protein is involved in IL-13-induced inflammation [256, 257] whereas VGF is involved in neuroplasticity, learning and memory [258-260]. Both of these proteins showed to greatly contribute to the prediction accuracy and both have been replicated in a multitude of different studies in CSF, making them less likely to be cohort specific [261-267]. In addition, one of the reported proteins in **Paper II** (C1S) was also replicated in this study, pointing again to alteration in complement system in AD. It is worth noting that, the rest of the reported proteins (n=55) were also valid findings, however, to maintain high stringency and robustness they were not presented as potential biomarkers for AD.

Paper IV

Evidence of changes in proteome of neuropathic pain patients after using spinal cord stimulator

The usage of SCS in pain modulation was motivated by gate control theory, hypothesizing that, sustained stimulation of A-β fibers through dorsal column can eventually reduce the nociceptive pain signal from the periphery [268]. Since then it has been several reports showing altered level of neurotransmitters, amino acids and peptides [269-271] as the result of using SCS. In this study, we showed altered level of twelve proteins in the CSF of neuropathic pain patients after SCS. These proteins include clusterin (CLUS), gelsolin (GELS), mimecan (MIME), angiotensinogen (ANGT), secretogranin-1 (SCG1), amyloid beta A4 protein (A4), apolipoprotein C1 (APOC1), apolipoprotein E (APOE), contactin-1 (CNTN1), neural cell adhesion molecule L1-like protein (NCHL1), dickkopf-related protein 3 (DKK3)

and neurosecretory protein VGF (VGF). Some of these proteins represent interesting findings in the context of SCS mechanism of action while others are novel findings. For example, overexpression of A4, a well-known protein in the AD context, has been shown to prevent neuropathic pain [272] while we also observed that A4 was approximately 52% increased after using SCS. Along the same line APOC1 was recently found to have higher abundance in CSF of neuropathic pain patients compared to controls [273]. Whereas, we found decreased abundance of this protein in CSF of patients after stimulation, indicating a possible beneficial effect on apolipoprotein system [274, 275]. Based on our results, the SCS can systematically affect the nervous system in the proteome level which provides further explanations about the protective effect of this method.

Paper V

Protein alternations in fibromyalgia patients

There is a growing evidence of changes in CSF milieu of FM patients that can be related to neuroinflammation [273, 276, 277]. In this study, we found changes in four proteins in CSF of FM patients compared to controls which partly verified previous findings about type of changes in CSF of FM patients: Apolipoprotein C-III (APOC3), Galectin-3-binding protein (LG3BP), Malate dehydrogenase, cytoplasmic (MDHC) and ProSAAS (PCSK1).

Changes in apolipoproteins in fibromyalgia

We already discussed the changes in the apolipoproteins in **Paper IV**. In this study, we also found higher level of one of the apolipoproteins called APOC3. While, this represents a novel finding in the context of fibromyalgia, APOC3 has been associated with arterial stiffness and cardiovascular events [278-280]. Although a direct connection between APOC3 and FM cannot be established, the observation of cardiovascular events in FM patients [281-284] can be the result of changes in apolipoproteins which is reflected in CSF of FM patients [285, 286].

Evidence of altered cytokine milieu and neuroendocrine

The LG3BP protein is known to regulate pro-inflammatory signaling [287, 288] which has also been proposed to decrease the expression of pro-inflammatory cytokines [289, 290]. We found elevated level of LG3BP in CSF of FM patients which possibly indicates that there is a lower level of certain cytokines as reported by others [291, 292]. Furthermore, we found higher level of PCSK1 which is known to inhibit proprotein convertase 1 [293], which in turn cleaves prohormones into their active hormonal form [294, 295], leading to neuroendocrine abnormalities [296]. Together, this

provides an indication of disturbance in the immune system regulation in patients with FM.

Paper VI

Virtual research environment

We, within PhenoMeNal consortium, have built a virtual research environment (VRE) that facilitates scalable and flexible data analysis for large scale metabolomics experiments using microservice architecture. The intention of having a VRE is to provide a convenient way of setting up the necessary components for initiation of the e-infrastructure. The VRE can be set up using a web interface on major cloud providers such as Amazon Web Services and Google Cloud Platform and can be accessed via application programming interface (API) or via a web-based graphical user interface (GUI). Currently the VRE includes the Galaxy workflow engine [203] for building computation pipelines using a user interface, Luigi [https://github.com/spotify/luigi] that provides higher flexibility in terms of pipeline creation and a coding environment based on Jupyter Notebooks [206]. In the current version, there are more than 160 different modules that can be used to create a pipeline, all of which are open source and available in a public repository such as GitHub. We demonstrate the capability of this VRE using four demonstrators. The PhenoMeNal web portal is publicly available through https://portal.phenomenal-h2020.eu, allowing the users to initiate the VRE on any public cloud provider.

Demonstrators

The purpose of demonstrators was to showcase the capability of microservice architecture to handle large and real scenario datasets. In the first demonstrator, we showed the capability of the microservices to handle increasing large amount of data (528 MS samples). We achieved 88% weak scaling efficiency (WSE), compared with the ideal case of 100% in which the running time of the analysis remains constant while the workload increased. This indicates remarkable computational performance achieved on a large dataset requiring more than 1000 jobs to be executed. In the second demonstrator, we showed capability of microservices to provide a framework to design interoperable computational workflows. We showcased the interoperability using a complete MS data pre-processing and downstream analysis workflow which consisted of eight different tools working with each other that were used to produce biologically valid results. The workflow was used on multiple sclerosis samples and resulted in discovery of regulated metabolites that were part of the tryptophan metabolism (alanyl-

tryptophan and indoleacetic acid) and endocannabinoids (linoleoyl ethanolamide) that are known to be important in multiple sclerosis [297-299]. Finally, in the third and fourth demonstrators we aimed to show that the application domain of microservices is not limited to MS. We created two more fully automated workflows for NMR and ^{13}C metabolic flux analysis (fluxomics). The result of the NMR analysis showed that pattern of the metabolite expression was different between type 2 diabetic and controls. Whereas the fluxomics resulted in description of the fluxes through the reactions accounting for glycolysis and pentose phosphate pathway. This clearly indicates that microservice architecture is domain-agnostic and is not limited to any certain domain. Altogether, we demonstrated that microservices enable efficient processing of large datasets and at the same time allows substantial level of interoperability (through applying data and metadata standards). Finally, as we demonstrated, microservices are not restricted to certain domains (e.g. metabolomics) and are generally applicable to virtually any field (such as proteomics or physics), making it possible for all disciplines to benefit from cloud computing platforms.

Methodological aspects

Proteomics

We used shotgun proteomics approach to discover biomarkers for AD (**Paper I, II, III**) and chronic pain (**Paper IV and V**). While proteomics is a powerful approach, capable of characterizing protein expression, modification, function as well as protein degradation, it does not give direct information from other levels of molecular changes in body (for example metabolites). Approaches such as genomics and metabolomics also hold much promises in biomarker discovery in both AD [300, 301] and chronic pain conditions [302, 303]. Combination of these approaches allows investigation of disease mechanisms from a wider view, including gene expression patterns, proteins and their interactions as well as the body's entire metabolism. Ultimately, accurate diagnosis of complex diseases can be facilitated using multifaceted targets discovered by complementary omics approaches [304].

Proteomics methods

Three major methods were used to extract protein level information in this thesis: MS (**Papers I, II, III, V and IV**), bead-based multiplex assay (**Paper I and II**), and western blot (**Paper I**). The major difference between MS and the two other technologies is that MS-based shotgun proteomics involves identifying proteins via their (or their fragments) masses while western blot and bead-based assays benefit from using antibodies.

In a biomarker discovery context, MS is ideal for rapid, large scale proteomic analysis of samples under investigation. Mass spectrometry is also not constrained by hypothesis concerning which proteins are present or altered in the samples. The MS-based experiments normally result in identification of thousands of proteins with relatively high specificity. On the contrary, immunoassays are constrained by choice of antibodies. In addition, these assays suffer from unspecific binding of antibodies, severely affecting specificity of the resulting protein abundance [305].

The MS analysis poses some challenges and has a number of limitations. The MS-based proteomics experiments require extensive sample preparation whereas the immunoassays require much less samples preparation, offering higher throughput than MS. The analytical sensitivity of immunoassays are typically higher than MS, enabling them to extract information from low abundant proteins which cannot be done with a regular MS-based experiment. Finally, the MS data processing including quantification and identification demands substantial computational resources, tools and expertise, making it prone to mistakes. On the other hand, analyzing immunoassays data is normally straightforward and faster than MS.

For these reasons, proteomics based biomarker discovery experiments can benefit from combining these two methods (similar to **Paper I and II**), by performing the discovery phase using MS and confirming the findings using immunoassays, providing the advantages high specificity and sensitivity without requiring rigid hypothesis about protein changes.

Mass spectrometry quantification methods

In **Paper IV**, we used two MS quantification methods, stable isotope dimethyl labeling and label free. Both methods provide certain advantages for quantification and identification. The quantification by stable isotope dimethyl labeling provides the advantage of having lower run to run variation as well as faster processing time due to having multiple samples in one MS run (run refers to a complete analysis of a sample by MS). This is particularly beneficial when the cohorts consist of paired samples (such as **Paper IV**) in which for example before and after treatment samples are from the same individual. In label free quantification, these samples might be affected by for example ionization efficiency or changed conditions in LC column. However, label free approach normally results in higher number of identified proteins due to having to deal with a less complex mixture than labeled experiments. Although the total MS running time is faster for labelled experiments than label free, an additional labeling step as well as extra experimental planning are needed in sample preparation. In addition, extra data analysis step needs to be perform for labelled data to correct for isotope distribution as well as retention time shift within labels. In summary, when more accurate relative quantification is required or fractionation has been performed, labelling is desired otherwise, label free experiments can provide

similarly accurate quantification but with more straightforward sample preparation.

Data analysis

Data analysis is indeed a key step in the biomarker discovery experiments. Selection of pre-processing and statistical methods can affect the end results and thus the biological conclusions.

Open source vs. commercial tools

In **Paper II and IV**, we used a mix of open source and commercial tools. While, the open source tools offer the great advantage of flexibility in selecting the parameters as well as the possibility of editing the tools, they might pose challenges to non-expert users. First, typically the users need to create computational workflows which demand substantial understanding of all the data processing steps (e.g. noise reduction, peak picking and feature detection to name a few). Commercial tools on the other hand generally provide user interfaces which are convenient to use for most of the users, along with predefined workflows. Second, open source tools expose a large selection of parameters to users (e.g. more 40 different parameters in the case of **Paper III**) to optimize, making it extremely prone to mistakes. Commercial tools often limit this selection to a small set of important parameters while the other settings (if exist) will be set automatically. Lastly, the version control in some open source tools is poorly maintained, making the data analysis irreproducible for example after changes in the tools are made. This is often not the problem as most tools from major vendors are subjected to scheduled release. In summary, open source tools provide flexible algorithm that can generally be adapted to work with many different experimental settings and with data from various instruments. Whereas the commercial tools are more robust and easier to use for non-expert users.

Data processing steps

As mentioned above the MS data processing consists of multiple steps that demand employing various algorithms. Output of each step can change the data and thus the behavior of the next steps:

1. The noise reduction and centroiding step should be wisely performed in order to not remove the real signals. In **Paper I, II, III and V**, we only performed centroiding as the result of noise removal was found to severely affect the result of feature detection (as the real signals were removed). Accordingly, in many of the high-resolution instruments, we found centroiding to sufficiently reduce the noise, thus the noise removal step can be safely skipped. However, for low-resolution instrument, the level of noise should be determined (e.g. by visually examining the intensities) and appropriate noise reduction steps should be performed.

2. In **Paper I, II, III, IV and V**, we performed retention time correction in order to reduce chromatography time shift. Based on experience from **Paper II**, overestimation of retention time shift can have drastic effects on the feature grouping step (more than underestimation time shift), allowing unrelated peptides from different proteins to be grouped. This ultimately results in incorrect quantification and therefore inaccurate biological conclusions.
3. Although it is desirable to combine the identification results from several search engines, in **Paper III**, we found that it might not consistently improve the identification results. This was found to be due to different scoring algorithms and weighting procedures implemented in different search engines (i.e. while several search engines assign high scores to correct PSMs, one inferior engine might decide on low scores for most of these matches). Therefore, to avoid losing identification strength, multiple search engines should be evaluated and post-processing of the identification should be performed.
4. In **Paper I, III, IV and V**, we used global normalization methods whereas in **Paper II**, we present local normalization as being more appropriate than global ones. First, in **Paper II**, the assumption of global normalization was violated whereas we did not observe this in other papers. Secondly, not all normalization methods are applicable to all data. In our analysis, the choice of normalization methods was based on the amount of variation in the levels of internal standard as well as our quality control samples.
5. In **Paper III and V**, we performed both univariate and multivariate data analysis. The choice of statistical method is critical to find relevant biological information about the data. While, univariate analysis looks at one variable at a time, the multivariate approaches look at and account for relationships between the multiple variables. Based on **Paper III and V**, the results of univariate and multivariate methods may not always completely overlap. In addition, univariate approaches are easier to interpret and are not affected by uninformative variables while the multivariate methods use complementary information from multiple variables but can be negatively influenced by the presence high number of uninformative variables. Therefore, performing both can offer benefit of easy interpretation as well as characterizing complex relationships between the variables.
6. In **Paper I, II, and IV**, we performed the statistical analysis on peptide level abundances whereas in **Paper III and V**, protein level abundances were used. Statistical analysis on the peptides is considered to be a liberal approach as several tests are performed on several peptides for each protein, making it more possible to find a statistically significant change. Statistical analysis on the proteins on the other hand, can provide more accurate measurement of concentration of the protein in the samples but

is more conservative. We clearly observed that in **Paper IV**, where for one protein, some of the peptides showed inconsistency in their relative levels (some had higher and some lower average abundance in SCS ON compared to OFF state) but were found to be statistically significant. However, when the peptide abundances were aggregated into protein abundance, it was not statistically significant. In general, selection of appropriate quantification method should be based on for example internal standards as well as having other level of information in mind (e.g. inconsistencies in peptide level when using protein quantification).

7. In **Paper VI**, we present the benefits of cloud computing over for example HPC in omics data analysis. However, it is important to note that while cloud computing targets the problems that can be divided into multiple tasks with minimum efforts, HPC intends to speed up analysis on extremely large sets in which one instance of a tool requires large amount of memory and processing power. Therefore, selection of a proper platform to perform that analysis depends on the nature of tool and the amount of workload. The steps such as feature grouping better fit HPC systems (as it demands large amount of memory and cannot be trivially parallelized) while the steps such as feature detection are perfectly suited for parallelized computation and cloud computing.

In summary, the selection of pre-processing and downstream analysis methods should always be performed based on the study data to avoid introducing data analysis artifacts and over/underestimating level of changes between the conditions under investigation.

Conclusions and future perspectives

The neurological diseases are one of the most prevalent type of disorders affecting human health. The nervous system is responsible for coordinating all the functions in the human body, the disorders in this system can severely decrease quality of life. Unfortunately, many of these diseases cannot be diagnosed in early stages, preventing the possibility for early intervention. This will often result in irreversible changes in the nervous system which could otherwise be prevented or at least controlled. Accurate and early diagnosis of these diseases require plenty of markers to reflect the degree of impairment in different cells. However, due to complexity and heterogeneity of cells in the nervous system, finding clinically relevant markers becomes difficult, necessitating a significant effort in the biomarker discovery field to overcome all the challenges posed by complex neurological disorders. The ultimate platform for performing biomarker discovery as well as to characterize different conditions in the body should include genomics, transcriptomics, proteomics and metabolomics. Among these, proteomics can provide a great level of information about the changes in biological systems when exposed to environmental and genetic factors. However, like other "omics", several challenges must be tackled to perform a successful and informative proteomics experiment, including sample selection, optimizing and standardizing sample preparation, selecting an appropriate platform to carry out the experiment and handling the resulting data.

In this thesis, the application of proteomics in finding biomarkers and eliciting pathological information for AD and two chronic pain conditions (neuropathic pain and FM) were discussed. In **Paper I**, evidence of hampered activity in vesicle transport mechanism as well as importance of micro vesicles for biomarker discovery in AD was presented and discussed. In **Paper II and III**, the consequence selecting different data analysis methods were present and discussed. In addition, ten potential biomarkers were present, some of which were novel in the AD context. Furthermore, we showed that early diagnosis of AD can be greatly improved using novel proteins. In **Paper IV**, we found proteins that significantly contribute to the beneficial effects of SCS in relieving neuropathic pain signs which was novel to the field. In **Paper V**, four novel proteins were presented as biomarkers for FM, some of which also indicated altered cytokine milieu and neuroendocrine in FM. Finally, In **Paper VI**, an e-infrastructure for performing scalable and flexible large scale data analysis was demonstrated that can be used to facili-

tate, amount others, biomarker discovery. In conclusion, the thesis provides a multidisciplinary overview of proteomics including experimental design, instrumentation, data analysis and computational solutions and demonstrated its significant role of biomarkers discovery.

Acknowledgements

I owe many thanks to a lot of people who have helped during my PhD studies and supported me to bring this thesis to its completion.

I would like to express my sincere gratitude and appreciation to my main supervisor Kim Kultima. Kim, words cannot express all I wish to say after the continuous support, inspiration and encouragement I received from you during my studies and in my personal life. As I have always said, I should write another thesis only about the things you taught me. If at any time during my life I call myself a researcher that is because of having you as a wonderful advisor and a great friend. Thank you for everything.

Beside my main advisor, I am very grateful to my first co-supervisor, Martin Ingelsson for tremendous guidance and support. Martin, it was my absolute honor to be guided by you. Your guidance was of enormous help for me in all the time of research and writing of this thesis. Your dedication and vast knowledge will be always inspiring me to be a better researcher.

I would like to thank and appreciate my second co-supervisor, Magnus Wetterhall for all the help specially in the beginning of my studies. Magnus, your help has been always substantial and key to our research. Although short, but the time I was guided by you was so valuable and helped me to pave my way into the field of mass spectrometry. Thank you for all the help.

I would like to give my special thanks to Ola Spjuth for all the great help and encouragement he gave to me. Ola, I must say I have been always motivated by your dedication and hard work. I'm so proud to be collaborating with you and hope that it will continue in the future.

I would like to express my thanks to Stephanie Herman for revising my thesis and giving me valuable inputs during my studies. I want to thank my good friend Shibu Krishnan for the immense help and favors he did to me.

I would like to thank my colleagues, Marco Capuccini, Anders Larsson and Jon Ander Novella, at the department of Pharmaceutical Biosciences in Uppsala for helping me to learn cloud computing. I also would like to give many thanks to Steffen Neumann, Christoph Ruttkies and Kristian Peters from

Leibniz Institute of Plant Biochemistry for being very friendly and helpful during my research retreat in Germany.

My special thanks to all of my colleagues, Ganna Shevchenko, Sravani Musunuri, Efthymia Chantzi, Elena Ossipova, Anne-Li Lind, Elisabeth Nikitidou and Sandy Abujrais for their enormous contributions and supports in the projects. It has been my pleasure working with you. You are all great researchers and with having you onboard, I'm quite sure, we can make the world a better place.

Special thanks go to Uppsala Berzelii Center and PhenoMeNal for giving me the opportunity to collaborate with a lot of experts in different fields, helping me to broaden my knowledge in various areas of science.

I would like to Mats Gustafsson and Eva Freyhult for all the interesting discussions and helps about statistics. Your inputs have been so valuable and key to our work. I appreciate all the help from you.

I thank our fantastic collaborators, Anna Erlandsson, Anna Häggmark, Jonas Bergquist, Lars Lannfelt, Lena Kilander, Lenka Katila, Marcus Sjödin, Maria Lönnberg, Maria Mikus, Peter Nilsson, Torsten Gordh, Camilla Svensson and Fredrik Nikolajeff for all the help and amazing and valuable discussions.

I thank all of my friends and colleagues at medical sciences department, Uppsala Berzelii Center and PhenoMeNal, for all the help and support I received from you during my studies.

I am proud to have a great family who provided me with a lot of motivation and support in my life. I would to thank to my father, Mohammad Ali, my mother, Soheila, my brother, Iman, and my sister, Shabnam for all the goodness they have done for me.

At last but not the least, I would like to give my warmest thank to my beloved wife, Elham and my son, Taha, for all the love, encouragement and support they gave to me. Without you, my success wouldn't have been possible. I'm so lucky and thankful to have you in my life. I love you!

References

1. Herculano-Houzel, S., *The Human Brain in Numbers: A Linearly Scaled-up Primate Brain.* Front Hum Neurosci, 2009. **3**.
2. Liu, K.K.L., et al., *Plasticity of brain wave network interactions and evolution across physiologic states.* Front Neural Circuits, 2015. **9**.
3. Kolb, B. and R. Gibb, *Brain Plasticity and Behaviour in the Developing Brain*, in *J Can Acad Child Adolesc Psychiatry*. 2011. p. 265-76.
4. Cai, L., et al., *Brain plasticity and motor practice in cognitive aging.* Front Aging Neurosci, 2014. **6**.
5. Green, C.S. and D. Bavelier, *Exercising Your Brain: A Review of Human Brain Plasticity and Training-Induced Learning.* Psychol Aging, 2008. **23**(4): p. 692-701.
6. Colom, R., et al., *Human intelligence and brain networks.* Dialogues Clin Neurosci, 2010. **12**(4): p. 489-501.
7. Nógrádi, A. and G. Vrbová, *Anatomy and Physiology of the Spinal Cord.* 2013.
8. Matsumae, M., et al., *Research into the Physiology of Cerebrospinal Fluid Reaches a New Horizon: Intimate Exchange between Cerebrospinal Fluid and Interstitial Fluid May Contribute to Maintenance of Homeostasis in the Central Nervous System*, in *Neurol Med Chir (Tokyo)*. 2016. p. 416-41.
9. Jurado, R. and H.K. Walker, *Cerebrospinal Fluid.* 1990.
10. Edsbagge, M., et al., *Spinal cerebrospinal fluid volume in healthy elderly individuals.* Clin Anat, 2011. **24**(6): p. 733-40.
11. Pollay, M., *The function and structure of the cerebrospinal fluid outflow system*, in *Cerebrospinal Fluid Res*. 2010. p. 9.
12. Stover, J.F., et al., *Neurotransmitters in cerebrospinal fluid reflect pathological activity.* Eur J Clin Invest, 1997. **27**(12): p. 1038-43.
13. Cepok, S., et al., *Patterns of cerebrospinal fluid pathology correlate with disease progression in multiple sclerosis.* Brain, 2001. **124**(Pt 11): p. 2169-76.
14. Nagy, K., et al., *Cerebrospinal fluid analyses for the diagnosis of subarachnoid haemorrhage and experience from a Swedish study. What method is preferable when diagnosing a subarachnoid haemorrhage?* Clin Chem Lab Med, 2013. **51**(11): p. 2073-86.
15. Rodewald, L.E., et al., *Relevance of common tests of cerebrospinal fluid in screening for bacterial meningitis.* J Pediatr, 1991. **119**(3): p. 363-9.
16. Olsson, B., et al., *CSF and blood biomarkers for the diagnosis of Alzheimer's disease: a systematic review and meta-analysis.* Lancet Neurol, 2016. **15**(7): p. 673-684.
17. Catala, M. and N. Kubis, *Gross anatomy and development of the peripheral nervous system.* Handb Clin Neurol, 2013. **115**: p. 29-41.

18. Fairless, R. and S.C. Barnett, *Olfactory ensheathing cells: their role in central nervous system repair.* Int J Biochem Cell Biol, 2005. **37**(4): p. 693-9.
19. Bergen, D.C. and D. Silberberg, *Nervous system disorders: a global epidemic.* Arch Neurol, 2002. **59**(7): p. 1194-6.
20. Thakur, K.T., et al., *Neurological Disorders*, in *Mental, Neurological, and Substance Use Disorders: Disease Control Priorities, Third Edition (Volume 4)*. 2016, The International Bank for Reconstruction and Development / The World Bank(c) 2016 International Bank for Reconstruction and Development / The World Bank.: Washington (DC).
21. DiNunzio, J.C. and R.O. Williams, 3rd, *CNS disorders--current treatment options and the prospects for advanced therapies.* Drug Dev Ind Pharm, 2008. **34**(11): p. 1141-67.
22. Schneider, L.S., et al., *Clinical trials and late-stage drug development for Alzheimer's disease: an appraisal from 1984 to 2014.* J Intern Med, 2014. **275**(3): p. 251-83.
23. Stone, J., *Functional neurological disorders: the neurological assessment as treatment.* Neurophysiol Clin, 2014. **44**(4): p. 363-73.
24. Stores, G., *Clinical diagnosis and misdiagnosis of sleep disorders*, in *J Neurol Neurosurg Psychiatry*. 2007. p. 1293-7.
25. Pope, J.V. and J.A. Edlow, *Avoiding Misdiagnosis in Patients with Neurological Emergencies.* Emerg Med Int, 2012. **2012**.
26. Kohshi, K., et al., *Cerebrospinal vascular diseases misdiagnosed as decompression illness: the importance of considering other neurological diagnoses.* Undersea Hyperb Med, 2017. **44**(4): p. 309-313.
27. Butler, C. and A.Z.J. Zeman, *Neurological syndromes which can be mistaken for psychiatric conditions.* 2005.
28. Polivka, J., et al., *Current status of biomarker research in neurology*, in *EPMA J*. 2016.
29. Dunckley, T., K.D. Coon, and D.A. Stephan, *Discovery and development of biomarkers of neurological disease.* Drug Discov Today, 2005. **10**(5): p. 326-34.
30. McGhee, D. and C.E. Counsell, *115 Systematic review of biomarkers for disease progression in Parkinson's disease.* 2012.
31. Uddin, L.Q., et al., *Progress and roadblocks in the search for brain-based biomarkers of autism and attention-deficit/hyperactivity disorder*, in *Transl Psychiatry*. 2017. p. e1218-.
32. Terry, R.D., *Alzheimer's disease and the aging brain.* J Geriatr Psychiatry Neurol, 2006. **19**(3): p. 125-8.
33. Banerjee, S., *The macroeconomics of dementia--will the world economy get Alzheimer's disease?* Arch Med Res, 2012. **43**(8): p. 705-9.
34. *2016 Alzheimer's disease facts and figures.* Alzheimers Dement, 2016. **12**(4): p. 459-509.
35. Heo, C., et al., *Effects of the monomeric, oligomeric, and fibrillar Abeta42 peptides on the proliferation and differentiation of adult neural stem cells from subventricular zone.* J Neurochem, 2007. **102**(2): p. 493-500.
36. Kumar, A., A. Singh, and Ekavali, *A review on Alzheimer's disease pathophysiology and its management: an update.* Pharmacol Rep, 2015. **67**(2): p. 195-203.

37. Rosenmann, H., *Immunotherapy for targeting tau pathology in Alzheimer's disease and tauopathies.* Curr Alzheimer Res, 2013. **10**(3): p. 217-28.
38. Iqbal, K., et al., *Tau pathology in Alzheimer disease and other tauopathies.* Biochim Biophys Acta, 2005. **1739**(2-3): p. 198-210.
39. Craddock, T.J.A. and J.A. Tuszynski, *On the Role of the Microtubules in Cognitive Brain Functions.* 5, 2010.
40. Swerdlow, R.H., *Pathogenesis of Alzheimer's disease*, in *Clin Interv Aging*. 2007. p. 347-59.
41. Selkoe, D.J., *Alzheimer's disease: genes, proteins, and therapy.* Physiol Rev, 2001. **81**(2): p. 741-66.
42. Beach, T.G., et al., *Accuracy of the Clinical Diagnosis of Alzheimer Disease at National Institute on Aging Alzheimer's Disease Centers, 2005–2010.* J Neuropathol Exp Neurol, 2012. **71**(4): p. 266-73.
43. Hansson, O., et al., *Association between CSF biomarkers and incipient Alzheimer's disease in patients with mild cognitive impairment: a follow-up study.* Lancet Neurol, 2006. **5**(3): p. 228-34.
44. Blennow, K., et al., *Clinical utility of cerebrospinal fluid biomarkers in the diagnosis of early Alzheimer's disease.* Alzheimers Dement, 2015. **11**(1): p. 58-69.
45. Petersen, R.C., *Early Diagnosis of Alzheimer's Disease: Is MCI Too Late?* Curr Alzheimer Res, 2009. **6**(4): p. 324-30.
46. Kelley, B.J. and R.C. Petersen, *Alzheimer's disease and mild cognitive impairment.* Neurol Clin, 2007. **25**(3): p. 577-609, v.
47. Gauthier, S., et al., *Mild cognitive impairment.* Lancet, 2006. **367**(9518): p. 1262-70.
48. Schopflocher, D., P. Taenzer, and R. Jovey, *The prevalence of chronic pain in Canada*, in *Pain Res Manag*. 2011. p. 445-50.
49. Phillips, C.J., *The Cost and Burden of Chronic Pain*, in *Rev Pain*. 2009. p. 2-5.
50. Johannes, C.B., et al., *The prevalence of chronic pain in United States adults: results of an Internet-based survey.* J Pain, 2010. **11**(11): p. 1230-9.
51. Bouhassira, D., E. Luporsi, and I. Krakowski, *Prevalence and incidence of chronic pain with or without neuropathic characteristics in patients with cancer.* Pain, 2017. **158**(6): p. 1118-1125.
52. Dureja, G.P., et al., *Prevalence of chronic pain, impact on daily life, and treatment practices in India.* Pain Pract, 2014. **14**(2): p. E51-62.
53. Bouhassira, D., et al., *Prevalence of chronic pain with neuropathic characteristics in the general population.* Pain, 2008. **136**(3): p. 380-7.
54. Baron, R., *Neuropathic pain: a clinical perspective.* Handb Exp Pharmacol, 2009(194): p. 3-30.
55. Moulin, D., et al., *Pharmacological management of chronic neuropathic pain: Revised consensus statement from the Canadian Pain Society*, in *Pain Res Manag*. 2014. p. 328-35.
56. Pluijms, W.A., et al., *Pain relief and quality-of-life improvement after spinal cord stimulation in painful diabetic polyneuropathy: a pilot study.* Br J Anaesth, 2012. **109**(4): p. 623-9.
57. Linderoth, B., et al., *Dorsal column stimulation induces release of serotonin and substance P in the cat dorsal horn.* Neurosurgery, 1992. **31**(2): p. 289-96; discussion 296-7.

58. Duggan, A.W. and F.W. Foong, *Bicuculline and spinal inhibition produced by dorsal column stimulation in the cat.* Pain, 1985. **22**(3): p. 249-59.
59. Shim, J.H., *Limitations of spinal cord stimulation for pain management.* Korean J Anesthesiol, 2015. **68**(4): p. 321-2.
60. Wolfe, F., et al., *The American College of Rheumatology 1990 Criteria for the Classification of Fibromyalgia. Report of the Multicenter Criteria Committee.* Arthritis Rheum, 1990. **33**(2): p. 160-72.
61. Walitt, B., et al., *The Prevalence and Characteristics of Fibromyalgia in the 2012 National Health Interview Survey*, in *PLoS One*. 2015.
62. Nampiaparampil, D.E. and R.H. Shmerling, *A review of fibromyalgia.* Am J Manag Care, 2004. **10**(11 Pt 1): p. 794-800.
63. Borchers, A.T. and M.E. Gershwin, *Fibromyalgia: A Critical and Comprehensive Review.* Clin Rev Allergy Immunol, 2015. **49**(2): p. 100-51.
64. Okifuji, A. and B.D. Hare, *Management of Fibromyalgia Syndrome: Review of Evidence*, in *Pain Ther*. 2013. p. 87-104.
65. Bradley, L.A., *Pathophysiology of Fibromyalgia.* Am J Med, 2009. **122**(12 Suppl): p. S22.
66. Bazzichi, L., et al., *Cytokine patterns in fibromyalgia and their correlation with clinical manifestations.* Clin Exp Rheumatol, 2007. **25**(2): p. 225-30.
67. Russell, I.J., et al., *Elevated cerebrospinal fluid levels of substance P in patients with the fibromyalgia syndrome.* Arthritis Rheum, 1994. **37**(11): p. 1593-601.
68. Sugimoto, C., et al., *Mucosal-associated invariant T cell is a potential marker to distinguish fibromyalgia syndrome from arthritis.* PLoS One, 2015. **10**(4): p. e0121124.
69. Malatji, B.G., et al., *A diagnostic biomarker profile for fibromyalgia syndrome based on an NMR metabolomics study of selected patients and controls*, in *BMC Neurol*. 2017.
70. Wahlen, K., et al., *Systemic alterations in plasma proteins from women with chronic widespread pain compared to healthy controls: a proteomic study.* J Pain Res, 2017. **10**: p. 797-809.
71. Stensson, N., et al., *Alterations of anti-inflammatory lipids in plasma from women with chronic widespread pain - a case control study.* Lipids Health Dis, 2017. **16**(1): p. 112.
72. Ablin, J.N., D. Buskila, and D.J. Clauw, *Biomarkers in fibromyalgia.* Curr Pain Headache Rep, 2009. **13**(5): p. 343-9.
73. Kell, D.B., *Systems biology, metabolic modelling and metabolomics in drug discovery and development.* Drug Discov Today, 2006. **11**(23-24): p. 1085-92.
74. Gerhold, D.L., R.V. Jensen, and S.R. Gullans, *Better therapeutics through microarrays.* Nat Genet, 2002. **32 Suppl**: p. 547-51.
75. Swami, M., *Proteomics: A discovery strategy for novel cancer biomarkers.* Nature Reviews Cancer, 2010. **10**(9): p. 597-597.
76. He, Q.Y. and J.F. Chiu, *Proteomics in biomarker discovery and drug development.* J Cell Biochem, 2003. **89**(5): p. 868-86.
77. Chen, R. and M. Snyder, *Promise of Personalized Omics to Precision Medicine.* Wiley Interdiscip Rev Syst Biol Med, 2013. **5**(1): p. 73-82.

78. Crutchfield, C.A., et al., *Advances in mass spectrometry-based clinical biomarker discovery*, in *Clin Proteomics*. 2016.
79. Cominetti, O., et al., *Proteomic Biomarker Discovery in 1000 Human Plasma Samples with Mass Spectrometry.* 2015.
80. Perneczky, R. and L.H. Guo, *Plasma Proteomics Biomarkers in Alzheimer's Disease: Latest Advances and Challenges.* Methods Mol Biol, 2016. **1303**: p. 521-9.
81. Finehout, E.J., et al., *Cerebrospinal fluid proteomic biomarkers for Alzheimer's disease.* Ann Neurol, 2007. **61**(2): p. 120-9.
82. Baird, A.L., S. Westwood, and S. Lovestone, *Blood-Based Proteomic Biomarkers of Alzheimer's Disease Pathology.* Front Neurol, 2015. **6**.
83. Olausson, P., et al., *Protein alterations in women with chronic widespread pain – An explorative proteomic study of the trapezius muscle.* Scientific Reports, 2015. **5**.
84. Nwagwu, C.D., et al., *Biomarkers for Chronic Neuropathic Pain and their Potential Application in Spinal Cord Stimulation: A Review.* Transl Perioper Pain Med, 2016. **1**(3): p. 33-8.
85. Backryd, E., et al., *Multivariate proteomic analysis of the cerebrospinal fluid of patients with peripheral neuropathic pain and healthy controls - a hypothesis-generating pilot study.* J Pain Res, 2015. **8**: p. 321-33.
86. Wilkins, M.R., et al., *From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis.* Biotechnology (N Y), 1996. **14**(1): p. 61-5.
87. Kellner, R., *Proteomics. Concepts and perspectives.* Fresenius J Anal Chem, 2000. **366**(6-7): p. 517-24.
88. Wang, K., C. Huang, and E. Nice, *Recent advances in proteomics: towards the human proteome.* Biomed Chromatogr, 2014. **28**(6): p. 848-57.
89. Marshall, A.G., C.L. Hendrickson, and G.S. Jackson, *Fourier transform ion cyclotron resonance mass spectrometry: a primer.* Mass Spectrom Rev, 1998. **17**(1): p. 1-35.
90. Zubarev, R.A. and A. Makarov, *Orbitrap mass spectrometry.* Anal Chem, 2013. **85**(11): p. 5288-96.
91. Vashist, S.K., *Immunodiagnostics: Major Advances and Future Insights.* Journal of Biochips & Tissue Chips, 2013. **3**(2).
92. Vitzthum, F., et al., *Proteomics: from basic research to diagnostic application. A review of requirements & needs.* J Proteome Res, 2005. **4**(4): p. 1086-97.
93. Berger, B., J. Peng, and M. Singh, *Computational solutions for omics data.* Nat Rev Genet, 2013. **14**(5): p. 333-46.
94. Wu, C.C. and M.J. MacCoss, *Shotgun proteomics: tools for the analysis of complex biological systems.* Curr Opin Mol Ther, 2002. **4**(3): p. 242-50.
95. Catherman, A.D., O.S. Skinner, and N.L. Kelleher, *Top Down Proteomics: Facts and Perspectives.* Biochem Biophys Res Commun, 2014. **445**(4): p. 683-93.
96. Gregorich, Z.R. and Y. Ge, *Top-down Proteomics in Health and Disease: Challenges and Opportunities.* Proteomics, 2014. **14**(10): p. 1195-210.
97. Ho, C., et al., *Electrospray Ionisation Mass Spectrometry: Principles and Clinical Applications*, in *Clin Biochem Rev*. 2003. p. 3-12.

98. Cunningham, R., et al., *Protein changes in immunodepleted cerebrospinal fluid from transgenic mouse models of Alexander disease detected using mass spectrometry.* J Proteome Res, 2013. **12**(2): p. 719-28.
99. Anderson, N.L. and N.G. Anderson, *The human plasma proteome: history, character, and diagnostic prospects.* Mol Cell Proteomics, 2002. **1**(11): p. 845-67.
100. Pernemalm, M., R. Lewensohn, and J. Lehtio, *Affinity prefractionation for MS-based plasma proteomics.* Proteomics, 2009. **9**(6): p. 1420-7.
101. Tu, C., et al., *Depletion of Abundant Plasma Proteins and Limitations of Plasma Proteomics.* J Proteome Res, 2010. **9**(10): p. 4982-91.
102. Qian, W.J., et al., *Enhanced detection of low abundance human plasma proteins using a tandem IgY12-SuperMix immunoaffinity separation strategy.* Mol Cell Proteomics, 2008. **7**(10): p. 1963-73.
103. Huang, L., et al., *Immunoaffinity separation of plasma proteins by IgY microbeads: meeting the needs of proteomic sample preparation and analysis.* Proteomics, 2005. **5**(13): p. 3314-28.
104. Steen, H. and M. Mann, *The ABC's (and XYZ's) of peptide sequencing.* Nat Rev Mol Cell Biol, 2004. **5**(9): p. 699-711.
105. Hustoft, H.K., et al., *Critical assessment of accelerating trypsination methods.* J Pharm Biomed Anal, 2011. **56**(5): p. 1069-78.
106. Lopez-Ferrer, D., et al., *Sample treatment for protein identification by mass spectrometry-based techniques.* TrAC Trends in Analytical Chemistry, 2006. **25**(10): p. 996-1005.
107. Wisniewski, J.R., et al., *Universal sample preparation method for proteome analysis.* Nat Methods, 2009. **6**(5): p. 359-62.
108. Lai, C.C. and G.R. Her, *Analysis of phospholipase A2 glycosylation patterns from venom of individual bees by capillary electrophoresis/electrospray ionization mass spectrometry using an ion trap mass spectrometer.* Rapid Commun Mass Spectrom, 2000. **14**(21): p. 2012-8.
109. Jiang, L., L. He, and M. Fountoulakis, *Comparison of protein precipitation methods for sample preparation prior to proteomic analysis.* J Chromatogr A, 2004. **1023**(2): p. 317-20.
110. Rappsilber, J., Y. Ishihama, and M. Mann, *Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics.* Anal Chem, 2003. **75**(3): p. 663-70.
111. Pitt, J.J., *Principles and Applications of Liquid Chromatography-Mass Spectrometry in Clinical Biochemistry*, in *Clin Biochem Rev*. 2009. p. 19-34.
112. Schuhmacher, M., et al., *Direct isolation of proteins from sodium dodecyl sulfate-polyacrylamide gel electrophoresis and analysis by electrospray-ionization mass spectrometry.* Electrophoresis, 1996. **17**(5): p. 848-54.
113. III, J.R.Y., *A century of mass spectrometry: from atoms to proteomes.* Nature Methods, 2011. **8**: p. 633-637.
114. Griffiths, J., *A Brief History of Mass Spectrometry.* 2008.
115. Karas, M. and F. Hillenkamp, *Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons.* Anal Chem, 1988. **60**(20): p. 2299-301.

116. Yang, Y., et al., *A Comparison of nLC-ESI-MS/MS and nLC-MALDI-MS/MS for GeLC-Based Protein Identification and iTRAQ-Based Shotgun Quantitative Proteomics*, in *J Biomol Tech*. 2007. p. 226-37.
117. Zaluzec, E.J., D.A. Gage, and J.T. Watson, *Matrix-assisted laser desorption ionization mass spectrometry: applications in peptide and protein characterization.* Protein Expr Purif, 1995. **6**(2): p. 109-23.
118. Banerjee, S. and S. Mazumdar, *Electrospray Ionization Mass Spectrometry: A Technique to Access the Information beyond the Molecular Weight of the Analyte.* Int J Anal Chem, 2012. **2012**.
119. Eberlin, M.N., *Electrospray ionization mass spectrometry: a major tool to investigate reaction mechanisms in both solution and the gas phase.* Eur J Mass Spectrom (Chichester), 2007. **13**(1): p. 19-28.
120. McLafferty, F.W., *A century of progress in molecular mass spectrometry.* Annu Rev Anal Chem (Palo Alto Calif), 2011. **4**: p. 1-22.
121. Rubakhin, S.S. and J.V. Sweedler, *A mass spectrometry primer for mass spectrometry imaging.* Methods Mol Biol, 2010. **656**: p. 21-49.
122. Vestal, M.L., *Modern MALDI time-of-flight mass spectrometry.* J Mass Spectrom, 2009. **44**(3): p. 303-17.
123. Mamyrin, B.A., et al., *The mass-reflectron, a new nonmagnetic time-of-flight mass spectrometer with high resolution.* Soviet Journal of Experimental and Theoretical Physics, 1973. **37**: p. 45.
124. Syed, S.U., et al., *Quadrupole mass filter: design and performance for operation in stability zone 3.* J Am Soc Mass Spectrom, 2013. **24**(10): p. 1493-500.
125. Douglas, D.J., A.J. Frank, and D. Mao, *Linear ion traps in mass spectrometry.* Mass Spectrom Rev, 2005. **24**(1): p. 1-29.
126. Bogdanov, B. and R.D. Smith, *Proteomics by FTICR mass spectrometry: top down and bottom up.* Mass Spectrom Rev, 2005. **24**(2): p. 168-200.
127. Hu, Q., et al., *The Orbitrap: a new mass spectrometer.* J Mass Spectrom, 2005. **40**(4): p. 430-43.
128. McLuckey, S.A., *Principles of collisional activation in analytical mass spectrometry.* J Am Soc Mass Spectrom, 1992. **3**(6): p. 599-614.
129. Elliott, M.H., et al., *Current trends in quantitative proteomics.* J Mass Spectrom, 2009. **44**(12): p. 1637-60.
130. Ong, S.E. and M. Mann, *Mass spectrometry-based proteomics turns quantitative.* Nat Chem Biol, 2005. **1**(5): p. 252-62.
131. Bantscheff, M., et al., *Quantitative mass spectrometry in proteomics: a critical review.* Anal Bioanal Chem, 2007. **389**(4): p. 1017-31.
132. Griffin, N.M., et al., *Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis.* Nat Biotechnol, 2010. **28**(1): p. 83-9.
133. Hsu, J.L., et al., *Stable-isotope dimethyl labeling for quantitative proteomics.* Anal Chem, 2003. **75**(24): p. 6843-52.
134. Jentoft, N. and D.G. Dearborn, *Labeling of proteins by reductive methylation using sodium cyanoborohydride.* J Biol Chem, 1979. **254**(11): p. 4359-65.
135. Perez-Riverol, Y., et al., *Open source libraries and frameworks for mass spectrometry based proteomics: a developer's perspective.* Biochim Biophys Acta, 2014. **1844**(1 Pt A): p. 63-76.

136. McHugh, L. and J.W. Arthur, *Computational methods for protein identification from mass spectrometry data.* PLoS Comput Biol, 2008. **4**(2): p. e12.
137. Kessner, D., et al., *ProteoWizard: open source software for rapid proteomics tools development*, in *Bioinformatics*. 2008. p. 2534-6.
138. Yang, C., Z. He, and W. Yu, *Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis*, in *BMC Bioinformatics*. 2009. p. 4.
139. Savitzky, A. and M.J. Golay, *Smoothing and differentiation of data by simplified least squares procedures.* Analytical chemistry, 1964. **36**(8): p. 1627-1639.
140. Barkauskas, D.A. and D.M. Rocke, *A general-purpose baseline estimation algorithm for spectroscopic data.* Anal Chim Acta, 2010. **657**(2): p. 191-7.
141. Williams, B., et al. *An algorithm for baseline correction of MALDI mass spectra*. in *Proceedings of the 43rd annual Southeast regional conference - Volume 1*. 2005. ACM.
142. Zhang, J., et al., *Review of Peak Detection Algorithms in Liquid-Chromatography-Mass Spectrometry*, in *Curr Genomics*. 2009. p. 388-401.
143. Lange, E., et al., *High-accuracy peak picking of proteomics data using wavelet techniques.* Pac Symp Biocomput, 2006: p. 243-54.
144. Weisser, H., et al., *An automated pipeline for high-throughput label-free quantitative proteomics.* J Proteome Res, 2013. **12**(4): p. 1628-44.
145. Rafiei, A. and L. Sleno, *Comparison of peak-picking workflows for untargeted liquid chromatography/high-resolution mass spectrometry metabolomics data analysis.* Rapid Commun Mass Spectrom, 2015. **29**(1): p. 119-27.
146. Senko, M.W., S.C. Beu, and F.W. McLaffertycor, *Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions.* J Am Soc Mass Spectrom, 1995. **6**(4): p. 229-33.
147. Tautenhahn, R., C. Bottcher, and S. Neumann, *Highly sensitive feature detection for high resolution LC/MS.* BMC Bioinformatics, 2008. **9**: p. 504.
148. Kohlbacher, O., et al., *TOPP--the OpenMS proteomics pipeline.* Bioinformatics, 2007. **23**(2): p. e191-7.
149. Tomasi, G., F. van den Berg, and C. Andersson, *Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data.* Journal of Chemometrics, 2004. **18**(5): p. 231-241.
150. Eilers, P.H., *Parametric time warping.* Anal Chem, 2004. **76**(2): p. 404-11.
151. Bylund, D., et al., *Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography-mass spectrometry data.* J Chromatogr A, 2002. **961**(2): p. 237-44.
152. Krebs, M.D., et al., *Alignment of gas chromatography–mass spectrometry data by landmark selection from complex chemical mixtures.* Chemometrics and Intelligent Laboratory Systems, 2006. **81**(1): p. 74-81.
153. Prakash, A., et al., *Signal maps for mass spectrometry-based comparative proteomics.* Mol Cell Proteomics, 2006. **5**(3): p. 423-32.

154. Lange, E., et al., *A geometric approach for the alignment of liquid chromatography—mass spectrometry data.* Bioinformatics, 2007. **23**(13): p. i273-i281.
155. LaMarche, B.L., et al., *MultiAlign: a multiple LC-MS analysis tool for targeted omics analysis.* BMC Bioinformatics, 2013. **14**: p. 49.
156. Smith, C.A., et al., *XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification.* Anal Chem, 2006. **78**(3): p. 779-87.
157. Du, P., W.A. Kibbe, and S.M. Lin, *Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching.* Bioinformatics, 2006. **22**(17): p. 2059-65.
158. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.* Nat Biotechnol, 2008. **26**(12): p. 1367-72.
159. Mortensen, P., et al., *MSQuant, an open source platform for mass spectrometry-based quantitative proteomics.* J Proteome Res, 2010. **9**(1): p. 393-403.
160. Ma, B., *Novor: Real-Time Peptide de Novo Sequencing Software*, in *J Am Soc Mass Spectrom*. 2015. p. 1885-94.
161. Zhang, J., et al., *PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification**, in *Mol Cell Proteomics*. 2012.
162. Perkins, D.N., et al., *Probability-based protein identification by searching sequence databases using mass spectrometry data.* Electrophoresis, 1999. **20**(18): p. 3551-67.
163. Ducret, A., et al., *High throughput protein characterization by automated reverse-phase chromatography/electrospray tandem mass spectrometry.* Protein Sci, 1998. **7**(3): p. 706-19.
164. Geer, L.Y., et al., *Open mass spectrometry search algorithm.* J Proteome Res, 2004. **3**(5): p. 958-64.
165. Craig, R. and R.C. Beavis, *TANDEM: matching proteins with tandem mass spectra.* Bioinformatics, 2004. **20**(9): p. 1466-7.
166. Cox, J., et al., *Andromeda: a peptide search engine integrated into the MaxQuant environment.* J Proteome Res, 2011. **10**(4): p. 1794-805.
167. Balgley, B.M., et al., *Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy.* Mol Cell Proteomics, 2007. **6**(9): p. 1599-608.
168. Kall, L., et al., *Posterior error probabilities and false discovery rates: two sides of the same coin.* J Proteome Res, 2008. **7**(1): p. 40-4.
169. Reiter, L., et al., *Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry.* Mol Cell Proteomics, 2009. **8**(11): p. 2405-17.
170. Kim, S. and P.A. Pevzner, *MS-GF+ makes progress towards a universal database search tool for proteomics.* Nat Commun, 2014. **5**: p. 5277.
171. The, M., A. Tasnim, and L. Käll, *How to talk about protein□level false discovery rates in shotgun proteomics*, in *Proteomics*. 2016. p. 2461-9.
172. Nesvizhskii, A.I., et al., *A statistical model for identifying proteins by tandem mass spectrometry.* Anal Chem, 2003. **75**(17): p. 4646-58.

173. Serang, O., M.J. MacCoss, and W.S. Noble, *Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data.* J Proteome Res, 2010. **9**(10): p. 5346-57.
174. Li, Y.F., et al., *A bayesian approach to protein inference problem in shotgun proteomics.* J Comput Biol, 2009. **16**(8): p. 1183-93.
175. Shteynberg, D., et al., *Combining Results of Multiple Search Engines in Proteomics**, in *Mol Cell Proteomics*. 2013. p. 2383-93.
176. Kapp, E.A., et al., *An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis.* Proteomics, 2005. **5**(13): p. 3475-90.
177. Nahnsen, S., et al., *Probabilistic consensus scoring improves tandem mass spectrometry peptide identification.* J Proteome Res, 2011. **10**(8): p. 3332-43.
178. Keller, A., et al., *Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search.* Anal Chem, 2002. **74**(20): p. 5383-92.
179. MacCoss, M.J., C.C. Wu, and J.R. Yates, 3rd, *Probability-based validation of protein identifications using a modified SEQUEST algorithm.* Anal Chem, 2002. **74**(21): p. 5593-9.
180. Kall, L., et al., *Semi-supervised learning for peptide identification from shotgun proteomics datasets.* Nat Methods, 2007. **4**(11): p. 923-5.
181. Gonnelli, G., et al., *A decoy-free approach to the identification of peptides.* J Proteome Res, 2015. **14**(4): p. 1792-8.
182. Ishihama, Y., et al., *Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein.* Mol Cell Proteomics, 2005. **4**(9): p. 1265-72.
183. Silva, J.C., et al., *Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition.* Mol Cell Proteomics, 2006. **5**(1): p. 144-56.
184. Callister, S.J., et al., *Normalization Approaches for Removing Systematic Biases Associated with Mass Spectrometry and Label-Free Proteomics.* J Proteome Res, 2006. **5**(2): p. 277-86.
185. Ballman, K.V., et al., *Faster cyclic loess: normalizing RNA arrays via linear models.* Bioinformatics, 2004. **20**(16): p. 2778-86.
186. Sysi-Aho, M., et al., *Normalization method for metabolomics data using optimal selection of multiple internal standards*, in *BMC Bioinformatics*. 2007. p. 93.
187. Wang, D., et al., *Comparison of different normalization assumptions for analyses of DNA methylation data from the cancer genome.* Gene, 2012. **506**(1): p. 36-42.
188. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.
189. Le Cao, K.A., S. Boitard, and P. Besse, *Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems.* BMC Bioinformatics, 2011. **12**: p. 253.
190. Svetnik, V., et al., *Random forest: a classification and regression tool for compound classification and QSAR modeling.* J Chem Inf Comput Sci, 2003. **43**(6): p. 1947-58.

191. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.
192. von Mering, C., et al., *STRING: a database of predicted functional associations between proteins.* Nucleic Acids Res, 2003. **31**(1): p. 258-61.
193. *The Gene Ontology (GO) database and informatics resource*, in *Nucleic Acids Res*. 2004. p. D258-61.
194. Halligan, B.D., et al., *Low cost, scalable proteomics data analysis using Amazon's cloud computing services and open source search algorithms.* J Proteome Res, 2009. **8**(6): p. 3148-53.
195. Muth, T., et al., *ProteoCloud: a full-featured open source proteomics cloud computing pipeline.* J Proteomics, 2013. **88**: p. 104-8.
196. Slagel, J., et al., *Processing Shotgun Proteomics Data on the Amazon Cloud with the Trans-Proteomic Pipeline**, in *Mol Cell Proteomics*. 2015. p. 399-404.
197. Tautenhahn, R., et al., *XCMS Online: a web-based platform to process untargeted metabolomic data.* Anal Chem, 2012. **84**(11): p. 5035-9.
198. Judson, B., et al. *Cloud IaaS for Mass Spectrometry and Proteomics: On-Demand Coupling of Cloud Computing to Experimental Facilities*. in *Proceedings of the 8th Workshop on Scientific Cloud Computing*. 2017. ACM.
199. Kaján, L., et al., *Cloud Prediction of Protein Structure and Function with PredictProtein for Debian.* Biomed Res Int, 2013. **2013**.
200. Fusaro, V.A., et al., *Biomedical Cloud Computing With Amazon Web Services*, in *PLoS Comput Biol*. 2011.
201. de Bruin, J.S., A.M. Deelder, and M. Palmblad, *Scientific Workflow Management in Proteomics*, in *Mol Cell Proteomics*. 2012.
202. Fillbrunn, A., et al., *KNIME for reproducible cross-domain analysis of life science data.* J Biotechnol, 2017.
203. Goecks, J., A. Nekrutenko, and J. Taylor, *Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.* Genome Biol, 2010. **11**(8): p. R86.
204. Tommaso, P.D., et al., *Nextflow enables reproducible computational workflows.* Nature Biotechnology, 2017. **35**: p. 316-319.
205. Merkel, D., *Docker: lightweight Linux containers for consistent development and deployment.* Linux J., 2014. **2014**(239): p. 2.
206. Kluyver, T., et al., *Jupyter Notebooks – a publishing format for reproducible computational workflows.* 2016.
207. Sturm, M., et al., *OpenMS - an open-source software framework for mass spectrometry.* BMC Bioinformatics, 2008. **9**: p. 163.
208. Kuhl, C., et al., *CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets.* Anal Chem, 2012. **84**(1): p. 283-9.
209. Wolf, S., et al., *In silico fragmentation for computer assisted identification of metabolite mass spectra.* BMC Bioinformatics, 2010. **11**: p. 148.
210. Thevenot, E.A., et al., *Analysis of the Human Adult Urinary Metabolome Variations with Age, Body Mass Index, and Gender by Implementing a Comprehensive Workflow for Univariate and OPLS Statistical Analyses.* J Proteome Res, 2015. **14**(8): p. 3322-35.

211. King, Z.A., et al., *Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways.* PLoS Comput Biol, 2015. **11**(8): p. e1004321.
212. Smart, J.F., *Jenkins: The Definitive Guide.* 2011.
213. Leprevost, F.D., et al., *BioContainers: An open-source and community-driven framework for software standardization.* Bioinformatics, 2017.
214. Mahmood, T. and P.C. Yang, *Western Blot: Technique, Theory, and Trouble Shooting*, in *N Am J Med Sci*. 2012. p. 429-34.
215. Egelhofer, T.A., et al., *An assessment of histone-modification antibody quality.* Nat Struct Mol Biol, 2011. **18**(1): p. 91-3.
216. Ellington, A.A., et al., *Antibody-Based Protein Multiplex Platforms: Technical and Operational Challenges.* Clin Chem, 2010. **56**(2): p. 186-93.
217. Elshal, M.F. and J.P. McCoy, *Multiplex Bead Array Assays: Performance Evaluation and Comparison of Sensitivity to ELISA.* Methods, 2006. **38**(4): p. 317-23.
218. Mirra, S.S., et al., *The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part II. Standardization of the neuropathologic assessment of Alzheimer's disease.* Neurology, 1991. **41**(4): p. 479-86.
219. Treede, R.D., et al., *Neuropathic pain: redefinition and a grading system for clinical and research purposes.* Neurology, 2008. **70**(18): p. 1630-5.
220. Madeira, C., et al., *d-serine levels in Alzheimer's disease: implications for novel biomarker development.* Transl Psychiatry, 2015. **5**: p. e561.
221. Hashimoto, K., et al., *Possible role of D-serine in the pathophysiology of Alzheimer's disease.* Prog Neuropsychopharmacol Biol Psychiatry, 2004. **28**(2): p. 385-8.
222. Fayed, N., et al., *Brain glutamate levels are decreased in Alzheimer's disease: a magnetic resonance spectroscopy study.* Am J Alzheimers Dis Other Demen, 2011. **26**(6): p. 450-6.
223. Ramirez, A.I., et al., *The Role of Microglia in Retinal Neurodegeneration: Alzheimer's Disease, Parkinson, and Glaucoma.* Front Aging Neurosci, 2017. **9**.
224. Mandrekar, S., et al., *Microglia Mediate the Clearance of Soluble Aβ through Fluid Phase Macropinocytosis.* J Neurosci, 2009. **29**(13): p. 4252-62.
225. Dickson, D.W., *Microglia in Alzheimer's Disease and Transgenic Models : How Close the Fit?*, in *Am J Pathol*. 1999. p. 1627-31.
226. Asai, H., et al., *Depletion of microglia and inhibition of exosome synthesis halt tau propagation.* Nat Neurosci, 2015. **18**(11): p. 1584-93.
227. Sze, C.I., et al., *Loss of the presynaptic vesicle protein synaptophysin in hippocampus correlates with cognitive decline in Alzheimer disease.* J Neuropathol Exp Neurol, 1997. **56**(8): p. 933-44.
228. Sze, C.I., et al., *Selective regional loss of exocytotic presynaptic vesicle proteins in Alzheimer's disease brains.* J Neurol Sci, 2000. **175**(2): p. 81-90.
229. Masliah, E., et al., *Topographical distribution of synaptic-associated proteins in the neuritic plaques of Alzheimer's disease hippocampus.* Acta Neuropathol, 1994. **87**(2): p. 135-42.

230. Krämer-Albers, E.-M. and W.P. Kuo-Elsner, *Extracellular Vesicles: Goodies for the Brain|[quest]|*. Neuropsychopharmacology, 2016. **41**(1): p. 371-372.
231. Ban, L.A., N.A. Shackel, and S.V. McLennan, *Extracellular Vesicles: A New Frontier in Biomarker Discovery for Non-Alcoholic Fatty Liver Disease.* Int J Mol Sci, 2016. **17**(3): p. 376.
232. Katsuda, T., K. Oki, and T. Ochiya, *Potential application of extracellular vesicles of human adipose tissue-derived mesenchymal stem cells in Alzheimer's disease therapeutics.* Methods Mol Biol, 2015. **1212**: p. 171-81.
233. Joshi, P., et al., *Extracellular vesicles in Alzheimer's disease: friends or foes? Focus on abeta-vesicle interaction.* Int J Mol Sci, 2015. **16**(3): p. 4800-13.
234. Vella, L.J., A.F. Hill, and L. Cheng, *Focus on Extracellular Vesicles: Exosomes and Their Role in Protein Trafficking and Biomarker Potential in Alzheimer's and Parkinson's Disease*, in *Int J Mol Sci*. 2016.
235. Cappadona, S., et al., *Current challenges in software solutions for mass spectrometry-based quantitative proteomics.* Amino Acids, 2012. **43**(3): p. 1087-108.
236. Valikangas, T., T. Suomi, and L.L. Elo, *A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation.* Brief Bioinform, 2017.
237. Kultima, K., et al., *Development and evaluation of normalization methods for label-free relative quantification of endogenous peptides.* Mol Cell Proteomics, 2009. **8**: p. 2285-95.
238. Shen, Y., L. Yang, and R. Li, *What does complement do in Alzheimer's disease? Old molecules with new insights*, in *Transl Neurodegener*. 2013. p. 21.
239. Charles A Janeway, J., et al., *The complement system and innate immunity.* 2001.
240. Veerhuis, R., H.M. Nielsen, and A.J. Tenner, *Complement in the brain.* Mol Immunol, 2011. **48**(14): p. 1592-603.
241. Toledo, J.B., et al., *Low levels of cerebrospinal fluid complement 3 and factor H predict faster cognitive decline in mild cognitive impairment.* Alzheimer's Research & Therapy, 2014. **6**(3): p. 36.
242. Mulder, C., et al., *CSF markers related to pathogenetic mechanisms in Alzheimer's disease.* J Neural Transm (Vienna), 2002. **109**(12): p. 1491-8.
243. Hu, W.T., et al., *CSF complement 3 and factor H are staging biomarkers in Alzheimer's disease*, in *Acta Neuropathol Commun*. 2016.
244. Daborg, J., et al., *Cerebrospinal fluid levels of complement proteins C3, C4 and CR1 in Alzheimer's disease.* J Neural Transm (Vienna), 2012. **119**(7): p. 789-97.
245. Bradt, B.M., W.P. Kolb, and N.R. Cooper, *Complement-dependent proinflammatory properties of the Alzheimer's disease beta-peptide.* J Exp Med, 1998. **188**(3): p. 431-8.
246. Aiyaz, M., et al., *Complement activation as a biomarker for Alzheimer's disease.* Immunobiology, 2012. **217**(2): p. 204-15.
247. Dubois, B., et al., *Timely Diagnosis for Alzheimer's Disease: A Literature Review on Benefits and Challenges*, in *J Alzheimers Dis*. p. 617-31.

248. Anoop, A., et al., *CSF Biomarkers for Alzheimer's Disease Diagnosis.* Int J Alzheimers Dis, 2010. **2010**.
249. Rosen, C., et al., *Fluid biomarkers in Alzheimer's disease - current concepts.* Mol Neurodegener, 2013. **8**: p. 20.
250. Suk, H.I. and D. Shen, *Deep learning-based feature representation for AD/MCI classification.* Med Image Comput Comput Assist Interv, 2013. **16**(Pt 2): p. 583-90.
251. Callahan, B.L., et al., *Predicting Alzheimer's disease development: a comparison of cognitive criteria and associated neuroimaging biomarkers.* Alzheimer's Research & Therapy, 2015. **7**(1): p. 68.
252. Spies, P.E., et al., *A prediction model to calculate probability of Alzheimer's disease using cerebrospinal fluid biomarkers.* Alzheimers Dement, 2013. **9**(3): p. 262-8.
253. Johnson, P., et al., *Genetic algorithm with logistic regression for prediction of progression to Alzheimer's disease.* BMC Bioinformatics, 2014. **15**(16).
254. Challis, E., et al., *Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI.* Neuroimage, 2015. **112**: p. 232-43.
255. Scherer, A., *Reproducibility in biomarker research and clinical development: a global challenge.* Biomark Med, 2017. **11**(4): p. 309-312.
256. Libreros, S., R. Garcia-Areas, and V. Iragavarapu-Charyulu, *CHI3L1 plays a role in cancer through enhanced production of pro-inflammatory/pro-tumorigenic and angiogenic factors.* Immunol Res, 2013. **57**(0): p. 99-105.
257. He, C.H., et al., *Chitinase 3-like 1 Regulates Cellular and Tissue Responses via IL-13 Receptor α2.* Cell Rep, 2013. **4**(4): p. 830-41.
258. Riedl, M.S., et al., *Proteomic Analysis Uncovers Novel Actions of the Neurosecretory Protein VGF in Nociceptive Processing.* J Neurosci, 2009. **29**(42): p. 13377-88.
259. Hunsberger, J.G., et al., *Antidepressant actions of the exercise-regulated gene VGF.* Nat Med, 2007. **13**(12): p. 1476-82.
260. Alder, J., et al., *Brain-Derived Neurotrophic Factor-Induced Gene Expression Reveals Novel Actions of VGF in Hippocampal Synaptic Plasticity.* J Neurosci, 2003. **23**(34): p. 10800-8.
261. Kester, M.I., et al., *Cerebrospinal fluid VILIP-1 and YKL-40, candidate biomarkers to diagnose, predict and monitor Alzheimer's disease in a memory clinic cohort.* Alzheimer's Research & Therapy, 2015. **7**(1): p. 59.
262. Wennstrom, M., et al., *The Inflammatory Marker YKL-40 Is Elevated in Cerebrospinal Fluid from Patients with Alzheimer's but Not Parkinson's Disease or Dementia with Lewy Bodies.* PLoS One, 2015. **10**(8): p. e0135458.
263. Choi, J., H.W. Lee, and K. Suk, *Plasma level of chitinase 3-like 1 protein increases in patients with early Alzheimer's disease.* J Neurol, 2011. **258**(12): p. 2181-5.
264. Rosen, C., et al., *Increased Levels of Chitotriosidase and YKL-40 in Cerebrospinal Fluid from Patients with Alzheimer's Disease.* Dement Geriatr Cogn Dis Extra, 2014. **4**(2): p. 297-304.
265. Kang, K., H.-W. Lee, and U. Yoon, *Plasma levels of lipocalin 2 and chitinase 3-like 1 protein in patients with amnestic mild cognitive*

impairment and Alzheimer's disease. Alzheimer's & Dementia: The Journal of the Alzheimer's Association. **9**(4): p. P860.
266. Gispert, J.D., et al., *The APOE ε4 genotype modulates CSF YKL-40 levels and their structural brain correlates in the continuum of Alzheimer's disease but not those of sTREM2*, in *Alzheimers Dement (Amst)*. 2017. p. 50-9.
267. Brinkmalm, G., et al., *A Parallel Reaction Monitoring Mass Spectrometric Method for Analysis of Potential CSF Biomarkers for Alzheimer's Disease.* PROTEOMICS – Clinical Applications: p. 1700131-n/a.
268. Jeon, Y.H., *Spinal Cord Stimulation in Pain Management: A Review.* Korean J Pain, 2012. **25**(3): p. 143-50.
269. Fontana, F., et al., *Opioid peptide response to spinal cord stimulation in chronic critical limb ischemia.* Peptides, 2004. **25**(4): p. 571-575.
270. Linderoth, B., et al., *Release of neurotransmitters in the CNS by spinal cord stimulation: survey of present state of knowledge and recent experimental studies.* Stereotact Funct Neurosurg, 1993. **61**(4): p. 157-70.
271. Cui, J.G., et al., *Spinal cord stimulation attenuates augmented dorsal horn release of excitatory amino acids in mononeuropathy via a GABAergic mechanism.* Pain, 1997. **73**(1): p. 87-95.
272. Kotulska, K., et al., *APP overexpression prevents neuropathic pain and motoneuron death after peripheral nerve injury in mice.* Brain Res Bull, 2010. **81**(4-5): p. 378-84.
273. Lind, A.-L., et al., *Affinity Proteomics Applied to Patient CSF Identifies Protein Profiles Associated with Neuropathic Pain and Fibromyalgia.* 2017.
274. Niederberger, E., et al., *Proteomics in Neuropathic Pain Research.* Anesthesiology: The Journal of the American Society of Anesthesiologists, 2017. **108**(2): p. 314-323.
275. Melemedjian, O.K., et al., *Proteomic and functional annotation analysis of injured peripheral nerves reveals ApoE as a protein upregulated by injury that is modulated by metformin treatment*, in *Mol Pain*. 2013. p. 14.
276. Vaeroy, H., et al., *Elevated CSF levels of substance P and high incidence of Raynaud phenomenon in patients with fibromyalgia: new features for diagnosis.* Pain, 1988. **32**(1): p. 21-6.
277. Bäckryd, E., et al., *Evidence of both systemic inflammation and neuroinflammation in fibromyalgia patients, as assessed by a multiplex protein panel applied to the cerebrospinal fluid and to plasma*, in *J Pain Res*. 2017. p. 515-25.
278. Wyler von Ballmoos, M.C., B. Haring, and F.M. Sacks, *The risk of cardiovascular events with increased apolipoprotein CIII: A systematic review and meta-analysis.* J Clin Lipidol, 2015. **9**(4): p. 498-510.
279. van Capelleveen, J.C., et al., *Apolipoprotein C-III Levels and Incident Coronary Artery Disease Risk: The EPIC-Norfolk Prospective Population Study.* Arterioscler Thromb Vasc Biol, 2017. **37**(6): p. 1206-1212.
280. Gaudet, D., et al., *Antisense Inhibition of Apolipoprotein C-III in Patients with Hypertriglyceridemia.* http://dx.doi.org/10.1056/NEJMoa1400283, 2015.

281. Tsai, P.S., Y.C. Fan, and C.J. Huang, *Fibromyalgia is associated with coronary heart disease: a population-based cohort study.* Reg Anesth Pain Med, 2015. **40**(1): p. 37-42.
282. Ablin, J.N., et al., *Association between fibromyalgia and coronary heart disease and coronary catheterization.* Clin Cardiol, 2009. **32**(6): p. E7-11.
283. Su, C.H., et al., *Increased Risk of Coronary Heart Disease in Patients with Primary Fibromyalgia and Those with Concomitant Comorbidity-A Taiwanese Population-Based Cohort Study.* PLoS One, 2015. **10**(9): p. e0137137.
284. Lee, J.H., et al., *Arterial Stiffness in Female Patients With Fibromyalgia and Its Relationship to Chronic Emotional and Physical Stress.* Korean Circ J, 2011. **41**(10): p. 596-602.
285. Walldius, G. and I. Jungner, *Apolipoprotein B and apolipoprotein A-I: risk indicators of coronary heart disease and targets for lipid-modifying therapy.* J Intern Med, 2004. **255**(2): p. 188-205.
286. Khuseyinova, N. and W. Koenig, *Apolipoprotein A-I and risk for cardiovascular diseases.* Curr Atheroscler Rep, 2006. **8**(5): p. 365-73.
287. DeRoo, E.P., et al., *The role of galectin-3 and galectin-3–binding protein in venous thrombosis*, in *Blood*. 2015. p. 1813-21.
288. Jeon, S.-B., et al., *Galectin-3 Exerts Cytokine-Like Regulatory Actions through the JAK–STAT Pathway.* 2010.
289. Ohshima, S., et al., *Galectin 3 and its binding protein in rheumatoid arthritis.* Arthritis Rheum, 2003. **48**(10): p. 2788-95.
290. McInnes, I.B. and G. Schett, *Cytokines in the pathogenesis of rheumatoid arthritis.* Nature Reviews Immunology, 2007. **7**(6): p. 429-442.
291. Sturgill, J., E. McGee, and V. Menzies, *Unique Cytokine Signature in the Plasma of Patients with Fibromyalgia.* Journal of Immunology Research, 2014. **2014**: p. 5.
292. Kadetoff, D., et al., *Evidence of central inflammation in fibromyalgia-increased cerebrospinal fluid interleukin-8 levels.* J Neuroimmunol, 2012. **242**(1-2): p. 33-8.
293. Basak, A., et al., *Inhibitory specificity and potency of proSAAS-derived peptides toward proprotein convertase 1.* J Biol Chem, 2001. **276**(35): p. 32720-8.
294. Bandsma, R., et al., *From diarrhea to obesity in prohormone convertase 1/3 deficiency – Age dependent clinical, pathological and enteroendocrine characteristics.* J Clin Gastroenterol, 2013. **47**(10).
295. Ugleholdt, R., et al., *Prohormone Convertase 1/3 Is Essential for Processing of the Glucose-dependent Insulinotropic Polypeptide Precursor.* 2006.
296. Crofford, L.J., *Neuroendocrine abnormalities in fibromyalgia and related disorders.* Am J Med Sci, 1998. **315**(6): p. 359-66.
297. Lovelace, M.D., et al., *Current Evidence for a Role of the Kynurenine Pathway of Tryptophan Metabolism in Multiple Sclerosis.* Front Immunol, 2016. **7**.
298. Lim, C.K., et al., *Kynurenine pathway metabolomics predicts and provides mechanistic insight into multiple sclerosis progression.* Scientific Reports, 2017. **7**.

299. Amirkhani, A., et al., *Interferon-beta affects the tryptophan metabolism in multiple sclerosis patients.* Eur J Neurol, 2005. **12**(8): p. 625-31.
300. Sato, Y., et al., *Identification of a new plasma biomarker of Alzheimer's disease using metabolomics technology[S]*, in *J Lipid Res*. 2012. p. 567-76.
301. Kang, J., J. Lu, and X. Zhang, *Metabolomics-based promising candidate biomarkers and pathways in Alzheimer's disease.* Pharmazie, 2015. **70**(5): p. 277-82.
302. Slade, G.D., et al., *Cytokine Biomarkers and Chronic Pain: Association of Genes, Transcription, and Circulating Proteins with Temporomandibular Disorders and Widespread Palpation Tenderness.* Pain, 2011. **152**(12): p. 2802-12.
303. Finco, G., et al., *Can Urine Metabolomics Be Helpful in Differentiating Neuropathic and Nociceptive Pain? A Proof-of-Concept Study*, in *PLoS One*. 2016.
304. Neagu, M., C. Longo, and S. Ribero, *Omics Landscape in Disease Biomarkers Discovery.* Dis Markers, 2016. **2016**.
305. Kingsmore, S.F., *Multiplexed protein measurement: technologies and applications of protein and antibody arrays.* Nat Rev Drug Discov, 2006. **5**(4): p. 310-20.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 1385

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-331748

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2017