Capturing the synaptic proteome: approaches for measuring and defining the synapse

Koopmans, Franciscus Theodorus Wilhelmus

2021

VRIJE UNIVERSITEIT

CAPTURING THE SYNAPTIC PROTEOME: APPROACHES FOR MEASURING AND DEFINING THE SYNAPSE

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor of Philosophy aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. C.M. van Praag,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Bètawetenschappen
op maandag 8 november 2021 om 13.45 uur
in een bijeenkomst van de universiteit,
De Boelelaan 1105

door

Franciscus Theodorus Wilhelmus Koopmans

geboren te Beuningen

promotoren:	prof.dr. M. Verhage
	prof.dr. A.B. Smit
copromotoren:	dr. L.N. Cornelisse
	dr. T.M.H. Dijkstra
promotiecommissie:	prof.dr. E.D. Gundelfinger
	prof.dr. P.D. Thomas
	prof.dr. D. Posthuma
	dr. H.D. MacGillavry
	dr. T.V. Pham

Contents

Summary

Synaptic transmission provides the primary mode of communication of neurons in the brain. Synapses are small cell junctions containing a sender and receiver element in which thousands of distinct proteins form macromolecular machines that act together in transducing signals and modulating the strength of neurotransmission. Synapses are fundamental to healthy brain function. The dysregulation of synapses is an important risk factor in many brain disorders and many drugs that treat brain disorders target synaptic proteins. Therefore, synapses are a subject of intense research as improved understanding of their fundamentals is a stepping stone for future research on their function in health and disease.

Studying synapse function is hindered by not knowing all parts of the synaptic machinery, and while the synapse has been studied extensively its exact molecular composition remains elusive. Proteins only expressed in a subset of synapses or with only few copies per synapse may not generate a signal above the limit of detection in typical proteomics approaches which warrants further efforts towards increasing the sensitivity of experimental techniques that can screen synaptic molecules.

This thesis aims to apply a comprehensive approach to characterize synaptic proteins and discover proteins previously not associated with the synapse. Mass spectrometry based high-throughput proteomics and data mining are the experimental approaches used to achieve this goal. The challenge of large-scale characterization of synaptic proteins is approached from multiple angles, advancing both mass spectrometry technology and application thereof.

Chapter 2 first establishes that the probability of detection in proteomics depends on the concentration of the analyte. This causes undersampling of quantitative data on medium to low abundant proteins that results in missing data, which can be substantial among replicates (>30%) and reduces the number of quantifiable proteins. A censoring model coined EBRCT is introduced that takes the pattern of missingness into account in differential expression analysis. Application to a benchmark dataset demonstrates improved performance of the EBRCT model in comparison with alternative models.

Chapter 3 explores an alternative mode of operation, coined WiSIM-DIA, for mass spectrometers and evaluates results together with state-of-the-art SWATH-MS. WiSIM-DIA combines SWATH-MS and wide-SIM (wide selected-ion monitoring) windows to partition the precursor mass-over-charge space to produce high-quality precursor ion chromatograms. This improves MS1 peak area-based quantification in a DIA strategy, in contrast to the SWATH-MS strategy that utilizes MS2 peak areas. Both strategies show strong overlap in the set of quantified peptides, but also exhibit unique advantages.

Chapter 4 is aimed at delineating biochemical impurities from protein constituents of a synaptic subcellular fraction of interest by application of a combination of quantitative proteomics and correlation-based data analysis to a series of related biochemical subfractions. Here, we do not rely on the protein identity list of the respective biochemical subfraction alone, but instead consider the protein abundances relative to related subfractions such as synaptosomes and synaptic membranes. Using canonical PSD proteins as a reference, which are enriched in the PSD biochemical subfraction and relatively low abundant in other subfractions, we searched for proteins that exhibit a strong correlation profile over all samples. The candidate protein list found through this bioinformatics approach indeed contained previously established PSD-enriched proteins among the top scoring proteins, validating the approach. Interestingly, known false-positives from the reference experiment, which was solely based on an affinity purification protocol, yielded low scores in our approach demonstrating the additional power of correlation-based data analysis to our quantitative enrichment approach. We identified multiple candidate proteins that are likely synaptic and validated two using high-resolution microscopy.

Chapter 5 is aimed at using SWATH-MS to quantify levels of hippocampal synaptic proteins of four species, the rodents; mouse and rat, and the primates; marmoset and human. A major challenge for inter-species proteomic comparison stems from a mass spectrometry technicality; peptides with a distinct amino acid sequence exhibit different ionization properties. Consequently, comparing the stoichiometry of some protein between mouse and human is accurate if the exact same peptide sequence is conserved and observable by mass spectrometry. In this study we therefore only used such conserved peptide sequences, a deliberate tradeoff between quantification accuracy and protein coverage, and achieved low technical variability. This enabled reliable detection of many protein abundance differences between species with mostly small fold changes. We used a set of proteins regulated by a strong fear learning paradigm impacting on the hippocampus to represent synaptic plasticity related proteins. Using this set of proteins we asked whether its constituents would belong to the rodent-primate conserved or rather the differentially expressed part of the synaptic proteome. The latter was true. This indicates that within the synaptic proteome those proteins of which expression differences maximally evolved during evolution are overrepresented in the plasticity response.

Chapter 6 describes an evidence-based, expert-curated knowledgebase for the synapse coined SynGO. First, an extensive ontology was developed to define synaptic processes and cellular components. After achieving consensus on this framework within the SynGO consortium, a worldwide collaboration of many expert laboratories in the synapse research community, domain experts systematically described synaptic protein functions and localizations for 1112 synaptic genes. This data was then used to empower bioinformatic analyses of the synapse that were previously not feasible due to a lack of high quality data at such large scale. We found that SynGO genes are exceptionally large, well conserved, and intolerant to mutations (as compared to other genes). Furthermore, a strong enrichment among genes associated with brain disorders was observed. All data was integrated into the Gene Ontology database and an online data analysis platform was developed to facilitate usage of the SynGO knowledgebase.

1

Introduction

General Introduction

Neural networks and synapses

Neurons are the fundamental cellular constituents of the brain, giving rise to its function through an intricate signal processing network. This complex neuronal network is organized at the structural level through connectivity and anatomically distinct areas, and at the smallest scale by composition of molecular machinery. They are part of neuronal circuits in which their axons and dendrites are connected through multitudes of synapses.

Synapses are specialized cell junctions through which (presynaptic) neurons can transduce chemical or electrical signals to a target (postsynaptic) cell. The typical axo-dendritic chemical synapse discussed here contains a presynaptic terminal on the axon of the transmitting neuron and a postsynapse on a dendrite of the receiving neuron, separated by a narrow gap called the synaptic cleft. However, a variety of alternative synaptic arrangements exist, for instance axo-axonic or somato-somatic synapses. An adult human brain contains an estimated 86 billion neurons (Azevedo *et al.,* 2009), supported by billions of glial cells for secondary systems such as supplying energy, and an estimated 100 trillion synapses (Drachman, 2005). These synapses are very small; while their sizes vary, the modal width is 200 nanometer (Mitchell *et al.,* 2012). For reference, the diameter of a human hair is 17~181 micrometer, which is 100~1000 times wider than a synapse, and the synaptic cleft that separates the pre- and post-synapse is only 20 nanometer across.

Each neuronal circuit performs computation by receiving input signals and sending output signals through its network architecture, and integrating and modulating the signals as they travel through all synapses. Signal integration is thus not only controlled by the wiring of the local network, but also at the level of individual synapses where subtle variations in molecular machinery can result in altered neurotransmission properties. Indeed, synapses are considered information processing units (Abbott & Regehr, 2004). The strength of the connections between a pair of neurons can be controlled by the number of synaptic connections, the geospatial location of these synapses (e.g. distance from the neuron's cell body, or soma) or the strength of synapses, which is modulated through previous activity of such a connection (synaptic plasticity) or diffusible chemical signals, which may either increase or decrease synaptic efficacy (Citri & Malenka, 2008; de Jong, Schmitz, *et al.,* 2012; Frischknecht *et al.,* 2009).

Synapses are thought to have evolved over time. Early unicellular organisms (Eukaryotes), which evolved approximately 1.8 billion years ago (and did not form synapses), already possessed basic cellular machinery, such as vesicle trafficking, exocytosis, and signal reception to allow signaling between cells. These basic units were recruited into a synapse that first appeared in multicellular organisms (Metazoa) approximately ~900–1400 million years ago. Over the following hundreds of millions of years, throughout the course of evolution, these basic signaling systems evolved, through genome duplications and subsequent mutation and selection, into refined synaptic and neuronal systems that facilitate complex behavior (Emes & Grant, 2012; Bayes, M. O. Collins, Reig-Viader, *et al.,* 2017). Species that are relatively close to humans in the phylogenetic tree of life, such as rodents, have a strong genetic similarity of synaptic genes (Emes & Grant, 2012) and therefore are often used as model systems for basic scientific research on the molecular mechanisms of the synapse.

Synapses are fundamental to healthy brain function and mutations in synaptic genes are an important risk factor in many brain disorders (Karczewski *et al.,* 2017; Lek *et al.,* 2016). Synapses are subject to intense research and better understanding of their fundamental functions is a stepping stone for future research on their dysfunction in brain disorders. In this thesis I introduce improvements of experimental approaches to identifying protein constituents of synapses, apply these techniques to the mammalian brain and finally put forth an extensive knowledgebase that describes the function and subcellular localization of synaptic proteins. Characterizing the synaptic proteome will enable future studies of synapse function that are currently hindered by not knowing all parts of the synaptic machinery and introducing methodological advancements will pave the way for a more detailed characterization of the synapse in the future.

Information processing in synapses

Synapses transduce and modulate input signals through a number of critical biological processes in the pre- and post-synaptic compartments. The main purpose of the *pre*synapse is to release neurotransmitters as a function of action potential input signals arriving from the neuron through the connected axon. Synaptic vesicles are filled with neurotransmitter and then recruited to the 'active zone', a highly specialized release site at the membrane facing the postsynapse (Schoch & Gundelfinger, 2006; Sudhof, 2012). Upon receiving sufficient input stimulus, these vesicles fuse with the membrane to release their cargo into the synaptic cleft (exocytosis) and as a result, the neurotransmitters diffuse towards the postsynapse. Afterwards, (partially) fused vesicles are recycled (endocytosis) and readied for another cycle of transmitter release (Südhof, 2004; Jahn & Fasshauer, 2012; Saheki & De Camilli, 2012).

The main purpose of the *post*synapse is to sense chemical input signals and convert them into electrical (output) signals of the postsynaptic neuron. Receptor protein complexes embedded in the membranous part of the postsynaptic density, a highly specialized protein scaffold for dynamic signal processing aligned with the presynaptic active zone, are activated by neurotransmitter molecules diffusing across the synaptic cleft. A dynamic system built from various receptors, kinases and substrate proteins is responsible for postsynaptic signal integration (Sheng & Kim, 2011; Sheng & Hoogenraad, 2007; Scannevin & Huganir, 2000). Besides this canonical mode of operation, secondary signaling pathways can be used in parallel to tune synaptic function. For instance, to provide feedback from post- to presynapse (Regehr *et al.,* 2009) or perform signaling from the presynapse to other cell types (Eroglu & Barres, 2010; Neniskyte & Gross, 2017). A visual summary of synaptic processes and cellular components was created as part of the work presented in this thesis and is shown in chapter 6, Figure 1.

Synapses are highly plastic and can adapt their strength in response to neuronal activity. This synaptic plasticity plays an important role in information processing and learning and memory. Both pre- and post-synaptic mechanisms have been identified that modify synaptic properties during neuronal activity (Citri & Malenka, 2008; Roberts & Glanzman, 2003). Numerous types of use-dependent synaptic plasticity have been described, revealing mechanisms that enhance (facilitate) or depress synaptic transmission on time scales in the order of milliseconds to days. Different types of synapses can specialize to interpret (continued) input by implementing various mechanisms of synaptic plasticity.

Short-lasting forms (milliseconds to minutes) of synaptic plasticity are mainly governed by presynaptic mechanisms (de Jong & Verhage, 2009), and are thought to play important roles in adaptions to sensory input and short-lasting forms of memory. Short-term depression can take effect after a single condition stimulus and recover in a matter of seconds, while sustained input (spike trains) may cause stronger effects with increased recovery time. Its counterpart, short-term enhancement (facilitation) increases the probability of presynapses releasing transmitters as a response to incoming action potentials (Zucker & Regehr, 2002; Regehr, 2012).

In long-term plasticity, specific patterns and the timing of pre- and postsynaptic neural activity lead to changes in synaptic efficacy and (neural) excitability that long outlast the events that trigger them. Activity-dependent, long-lasting changes of synaptic strength such as long-term potentiation (LTP) and long-term depression (LTD) are crucial mechanisms for the formation and consolidation of memories (McGaugh, 2000; Martin *et al.,* 2000; Whitlock *et al.,* 2006).

Homeostatic plasticity is not use-dependent on the synaptic level but rather operates on the level of neural circuits and at a much slower timescale, promoting stability of network activity (e.g. preventing runaway effects of increasing synaptic excitability), balancing excitatory versus inhibitory synaptic inputs or normalizing total synaptic strength (Turrigiano & Nelson, 2004).

Uncharted territory: discovering the synaptome

The entire complement of synaptic proteins is referred to as the synaptic proteome, or synaptome. Studying how synapse function emerges from the interactions between the different elements within the synaptome is hindered by not knowing all parts of the synaptic machinery, and while the synapse has been studied extensively its exact molecular composition remains elusive.

Within the synapse, proteins may dynamically assemble into macromolecular machines, also referred to as protein complexes, to perform very specific functions. Each protein may fulfill multiple functions and interact with a multitude of other proteins to do so. Posttranslational modification may be applied to proteins leading to modulation of their functions, enabling or disabling functional domains, thereby dynamically regulating a protein's molecular function. Taken together, at the molecular level synapses are complex dynamic machines with thousands of multifunctional parts that respond and adapt to input.

A complicating factor is that not all synapses are cut from the same cloth. Distinct synapse types that interchange a small set of proteins to attain different signaling properties can be found throughout the brain. Throughout the many structural and functional areas of the mammalian brain, distinct types of neurons have been catalogued and synaptic proteins are differentially expressed across brain areas (Tasic *et al.,* 2018). For example, synapses in the cerebellum express different Tarp family members for modulating postsynaptic AMPA-type glutamate receptors as compared to the hippocampus. Furthermore, proteins may exhibit intrinsic molecular properties that hinder experimental detection such as strong sequence homology with already known other proteins (thus lacking unique parts for identification) or their sequence mostly consisting). An additional challenge is that some proteins are only expressed in a subset of synapses or with only few copies per synapse and therefore may not generate a signal above the limit of detection in typical proteomics approaches. This requires further efforts towards increasing the sensitivity of experimental techniques that can screen synaptic molecules.

Finally, a lot of effort has so far been spent on proteins that are known to be important. As also remarked by Edwards *et al.* (2011); *"75% of protein research focuses on the 10% of proteins that were known before the human genome was mapped"*. I advocate casting a wider net to identify synaptic proteins, or configurations of protein complexes, that have so far escaped our attention but may play an important role in (modulating) the synapse.

Although great progress has been made recently in the identification of neuronal cell types, experimental techniques that can screen all proteins of the synapse from a small area of the brain, or a specific neuronal circuit, are still an active area of research. Pushing the envelope on experimental assays for sensitive large-scale protein identification will enable new insights on what types of synapse configurations exists in the human brain, setting the stage for further functional studies. The next section will discuss such tools and their application to the synaptome.

Proteomics approaches

The goal of proteomics is a comprehensive, quantitative description of protein expression and its changes under the influence of biological perturbations such as disease or drug treatment (N. L. Anderson & N. G. Anderson, 1998). Although there are many paradigms to such research, proteomics in this thesis is focused on the large-scale study of (synaptic) proteins using mass spectrometry.

The bottom-up approach

While the human genome holds circa 22.000 protein coding genes, up to a million unique proteoforms may be created that differ in their primary structure through combinatorial explosion of splice variants (protein isoforms) and post-translational modifications (PTMs) (Smith & Kelleher, 2013; Aebersold, Agar, *et al.,* 2018). This huge collection of unique proteoforms spans a wide range of molecular weights but also gives rise to distinct proteoforms with near indistinguishable molecular weights. Detecting these proteins in their natural state by mass spectrometry is challenging due to a technical limitation; the detectors used in mass spectrometry are more accurate and sensitive for low molecular-weight molecules, and even for small molecules the resolution of current-generation machines may be too low to distinguish all proteoforms (Conrads *et al.,* 2000).

The bottom-up proteomics strategy employs enzymes (e.g. Trypsin) to cleave protein sequences into peptide chunks. The resulting peptide mixture is then analyzed by liquid chromatography coupled to a mass spectrometer to obtain a fingerprint for each peptide that hits the mass detectors. The observed peptide fingerprints are compared to reference databases to computationally infer their amino acid sequence and ultimately infer which proteins were present in the input sample (Altelaar & Heck, 2012; Steen & Mann, 2004). These smaller peptides are easier to identify by mass spectrometers than intact proteins. But a disadvantage is that the digestion of proteins into peptides introduces increased complexity, most proteins yield multiple peptides with potential overlap between proteins, and the inference of the proteins from which the peptides originate is non-trivial (Ma *et al.,* 2012). An overview of the entire workflow is shown in Fig. 1.1 and described in further detail in the next sections.

Data dependent acquisition (DDA)

Mass spectrometers can be configured in different modes to optimize the data acquisition for identification or quantification of peptides (Aebersold & Mann, 2016). Spending most measurement time on fingerprinting peptide identities optimizes the number of proteins that can be identified in an input sample, which is the common purpose of Data Dependent Acquisition (DDA) mode. Alternatively, if the identities of peptides and proteins in the sample are already known, the mass spectrometer can also be optimized for reproducibly quantifying these, as discussed in the next section.

For DDA, the peptide mixture is separated with a gradient of aqueous/organic solvent in high-performance liquid chromatography (HPLC), which effectively sorts the peptides such that those with similar hydrophobic properties co-elute over the column before entering the mass spectrometer (Fig. 1.1B) (Altelaar, Munoz, *et al.,* 2013; Aebersold & Mann, 2016). The analysis time is defined by the length of this gradient and the moment a peptide is detected downstream is referred to as retention time (aka. elution time). The peptides are converted to gas phase ions by electrospray and subsequentially sent to the mass detector. A first mass detector that measures the peptides, here referred to as MS1, scans the entire mass range every few seconds. The top N most abundant peptides (circa 15-25 for modern machines) are each isolated, then fragmented (breaking most bonds between amino acids in the peptide sequence) and sent to another mass detector (here referred to as MS2) for analysis of their mass spectra (Fig. 1.1C). This two-step approach is known as tandem mass spectrometry (MS/MS). To summarize each cycle; 1) peptides eluting at time *t* are scanned in MS1 yielding a mass, charge and signal intensity. 2) top N peptides are isolated, each is sent to MS2 yielding a mass, charge and signal intensity for all fragments of the precursor peptides.

In post-acquisition analysis, data generated for each detected peptide is compared to predicted data profiles for all possible peptides in the reference protein database (e.g. all protein products predicted from the genome of the respective species) resulting in a score that expresses the likelihood of correct match. To account for false-positive matches, a false discovery rate strategy is commonly used to determine at what score threshold it is rare to find false positive results. For this purpose, scores are also generated for spectral matches against a decoy database

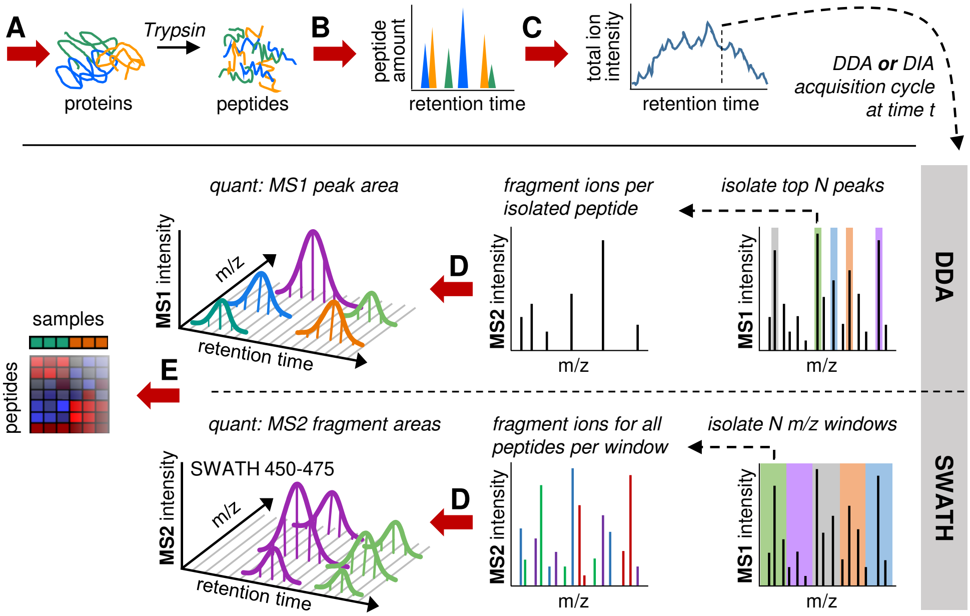

Figure 1.1: Overview of a typical label-free quantitative proteomics workflow. A) In bottom-up proteomics, proteins are first extracted and digested (e.g. by trypsin) into peptides. B) Peptides are separated using chromatography (HPLC). Each peak here represents a peptide. C) As peptides elute from the HPLC over time, they are ionized by electrospray and enter the mass spectrometer. In a continuous cycle (repeated every 1-3 seconds), a full spectrum at the MS1 level is first acquired and then followed by the isolation and fragmentation of selected peptides. Depending on whether DDA or DIA mode is employed; for DDA, the top-most abundant peaks in the MS1 scan are selected and fragmented sequentially while for SWATH/DIA, all peptides within a predefined m/z (mass/charge) window are selected and fragmented (for each window, sequentially). D) After mass spectrometry is done, signal processing software identifies and quantifies peptides in the input sample. For DDA, the MS1 peak areas are used for quantification and the MS2 spectra are used to infer the amino-acid sequence of respective peptides. For SWATH/DIA, elution profiles at MS2-level over time are used to deconvolve the complex spectra generated by co-fragmenting many peptides per m/z window, then used for both identification and quantification. E) The result is an abundance value matrix that shows the relative amount of each peptide in all samples. In this example, differential abundances for some peptides are observed between the samples color-coded as green and orange. From these peptide-level data, protein (relative) abundances can be inferred and used in downstream statistical analyses.

(often generated by simply using reversed sequences from the target database) to obtain a distribution of 'decoy scores' (Shteynberg *et al.,* 2011; Käll *et al.,* 2007). For quantitative analyses, the (relative) abundance value for each peptide is derived from its elution profile at MS1-level (Fig. 1.1D).

Finally, the list of detected peptide sequences is matched to the protein sequences in the reference protein database to infer proteins that were present in the input sample. Peptide sequences sometimes overlap between proteins transcribed from distinct genes and splice variants for the same protein (isoforms) often have strong sequence redundancy. This puzzle, with many pieces that fit in multiple places, is commonly referred to as the protein inference problem (Nesvizhskii & Aebersold, 2005). As a result, there may be some ambiguity if there is peptide evidence for the presence of multiple matching proteins but no unique evidence that separates either. These are then together reported as a 'protein group'. Taken together, discovery proteomics enables high-throughput identification of proteins in a sample by measuring peptides with a mass spectrometer and algorithmic post-processing of the measurement data.

Data independent acquisition (DIA)

DIA (Gillet *et al.,* 2012), also known as SWATH-MS (Liu *et al.,* 2013), is a recent acquisition strategy that enables the quantification of thousands of proteins (B. C. Collins *et al.,* 2017) by fragmenting all precursor ions (peptides) in each acquisition cycle, in contrast to the DDA strategy that selects a few abundant ions for isolation and fragmentation, and a post-acquisition computational approach to deconvolute the overlapping peptide signals. Since the data observed during the experiment have no effect on the peptides selected for fragmentation, this strategy is classified as a data independent acquisition.

During DIA acquisition, a MS1 scan revealing precursor ions eluting over time (same as DDA) is followed by consecutively fragmenting all peptides within stepped m/z windows (e.g. 450~475, then 475~500, etc. where m/z denotes mass-over-charge) until a preconfigured range of precursor masses has been covered. This cycle is continuously repeated throughout peptide elution (Fig. 1.1C) (Gillet *et al.,* 2012; Ludwig *et al.,* 2018). As a result, MS2 data on all peptides in the sample is available. But note that the selection of all peptides in wide m/z windows results in convoluted data signals that heavily lean on downstream software deconvolution and interpretation (Bruderer *et al.,* 2015; Rost *et al.,* 2014), in contrast to DDA mode that has minimal peptide co-elution per MS/MS scan and only MS2 data for a selected subset of peptides (Fig. 1.1D).

While it is difficult to identify peptides in a sample based solely on such DIA data, if the retention time and MS2 fingerprint for peptides in the sample are known *a priori* this data is particularly suited for extracting their quantities. To build such a spectral library for SWATH-MS, a regular DDA is performed on a sample comparable to the sample of interest. This results in an empirical signature of each peptide in the mass-spec (e.g. at that time does it elute, what is the fragmentation pattern in MS2). Alternatively, recent innovations are paving the way for theoretical spectral libraries that are complete and do not require sample nor measurement time, but are not available for all mass-spec platforms yet (Gessulat *et al.,* 2019; Tiwary *et al.,* 2019; Guan *et al.,* 2019).

In post-acquisition data analysis, the multiplexed MS/MS data signals are deconvoluted within each m/z window by finding fragments that co-elute in the retention-time dimension. Fragments of the same peptide should together form a bell-shaped peak over time, coinciding with the moment the peptide eluted over the column and was electro-sprayed into the mass-spec, and match to the peptide fingerprint stored in the respective spectral library. As a result, each peptide in the spectral library is assigned a confidence score and abundance. The former indicates how convincing the empirical evidence for its presence is, the latter is the peak area for each of its fragments (Ludwig *et al.,* 2018).

In conclusion, mass spectrometry can be used for high-throughput protein identification and quantification. DDA is aimed at the former and is (relatively) limited for the latter, whereas DIA is optimized for quantification but requires additional measurement in DDA mode to build a spectral library with peptide fingerprints.

Label-free quantitative proteomics

Mass spectrometry based quantitative proteomics can be used to find differentially expressed proteins between experimental conditions. For instance, given some hippocampal tissues from a set of healthy controls and a group with some disease state, identify those proteins that have significantly altered abundance levels. Several workflows in both sample preparation and acquisition strategies are available to generate quantitative data, here we describe two label-free approaches commonly used in the field and in this thesis. Label-free refers to the lack of chemical (Wiese *et al.,* 2007) and metabolic (Gouw *et al.,* 2010) labeling reagents during sample-prep. In such settings, physiological samples of interest are prepared (wet-lab) and measured (mass-spec) independently.

In label-free DDA, precursor ion intensities are used to quantify peptides (Zhu *et al.,* 2009; Cox *et al.,* 2014). The peak area of each peptide is the area under the curve as a function of its ion intensity (ion count on the detector) over retention time (Fig. 1.1D). In label-free SWATH-MS, the MS2 peak area of all fragments for a peptide are summed together to obtain the peptide abundance (Fig. 1.1D) (Ludwig *et al.,* 2018). These intensities do not directly represent the absolute abundances since the physiological properties of a peptide affects the ionization efficiency in non-linear fashion, obscuring the relationship between peptide copy numbers in the input sample and the observed intensity in mass spectrometry. However, this unknown function is stable thus allowing the 'relative abundances' (peptide intensities) to be compared between measurements but not between peptides. Additional strategies for absolute quantification include the use of synthetic peptides (Kirkpatrick *et al.,* 2005) or in-silico approximation (Schwanhäusser *et al.,* 2011).

In mass spectrometry, peptide abundance and signal quality are strongly correlated. Highly abundant peptides are easier to detect since there are more ions hitting the detector, increasing their signal-to-noise ratio and reducing variation among replicate measurements. This leads to a selection bias towards the peptides of more abundant proteins (Michalski *et al.,* 2011), resulting in more observed values among replicate samples for high abundant peptides and more missing values for low abundant peptides (e.g. the peptide signal quality was too low for detection, or abundance was just below the detection limit) (Zubarev, 2013; Karpievitch *et al.,* 2012). This is commonly referred to as the missing data problem, which will be further discussed in chapter two.

Advantages and limitations of proteomics

Spatial information is mostly lost in the process when using proteomics, in contrast to low-throughput microscopy approaches where a label attached to a protein of interest is visualized at the protein's native position in a cell. As samples are prepared for mass spectrometry, the spatial coordinates of where a protein is in the cell is lost and the output of the experiment simply states the accumulated copy numbers of each protein in the input sample. A common approach to gaining at least a coarse spatial resolution is the application of biochemistry to isolate a subcellular compartment, such as a synaptic vesicle, such that one can infer where proteins identified through proteomics originate. Of course, the resolution of this approach depends on the ability to isolate such subcellular compartments and the presence of impurities (unintentionally co-isolated material). Recent innovations attack this problem using chemical proximity labeling, which adds a tag to a protein located in a subcellular region of interest, and applying a reagent that labels all proteins in close proximity of this tag and finally apply proteomics to identify all labelled proteins (Roux *et al.,* 2013; Firat-Karalar *et al.,* 2014; Myers *et al.,* 2018). This is a useful strategy for high-throughput localization studies at coarse resolution, for high resolution localization studies microscopy is necessary.

Compared to RNA-sequencing (RNA-seq) based technologies, which have been successfully applied recently to identify genes in single brain cells, throughput and sensitivity for proteomics remains modest. Proteomics uses a mass spectrometer to detect protein molecules in a tissue sample while RNA-sequencing is only observing the RNA precursors in the cell body that may be later used to produce proteins. So, with proteomics, we observe the actual molecular machinery of a biological system which is apt for studying molecular neurobiology, whereas the RNA-sequencing approach can only predict protein existence in cells but may be a better fit for large-scale systems-biology.

Key advantages of using proteomics to identify and quantify the actual proteins is that a) the relationship between the amount of observed RNA in a cell and protein copy numbers is not linear, making RNA-seq less suitable for the prediction of the amount and ratio of protein constituents in a cell, b) RNA-seq cannot account for local translation outside of the cell soma, which obscures locally translated synaptic proteins and c) proteins may be expressed as distinct isoforms and each may undergo various Post-Translational-Modifications (PTM), the combination yields a unique proteoform repertoire. Proteomics is able to detect these, although at great difficulty if the isoform or PTM is rare since that amounts to relatively little input material for mass spectrometry.

Capturing the synaptic proteome

Studies of synapse function are hindered by not knowing all parts of their molecular machinery. To work towards a more complete catalog of synaptic protein constituents, mass spectrometry combined with bioinformatic analyses has been successfully applied to characterize sub-synaptic compartments and identify synaptic protein complexes. Previous studies that applied proteomics to biochemical enrichments of synapses (often referred to as synaptosomes) have generated many sets of proteins that are potentially localized in synapses (Filiou *et al.,* 2010; Wilhelm *et al.,* 2014; Tang *et al.,* 2015).

Similarly, biochemical preparations targeting sub-synaptic compartments such as the postsynaptic density (K. W. Li, Hornshaw, der Schors, *et al.,* 2003; K. W. Li, Hornshaw, Van Der Schors, *et al.,* 2004; Jordan *et al.,* 2004; Phillips *et al.,* 2005), active zone (Morciano, Beckhaus, *et al.,* 2009; Abul-Husn *et al.,* 2009; Volknandt & Karas, 2012; Boyken *et al.,* 2013), synaptosomal membranes (K. w. Li *et al.,* 2005; Sialana *et al.,* 2016) or synaptic vesicles (Takamori *et al.,* 2006; Morciano, Burré, *et al.,* 2005; Taoufiq *et al.,* 2020) have generated data on both previously established and potentially novel synaptic proteins. There are many variations to these protocols, some are refinements or adaptions to advancing mass spectrometry capabilities while some are extensions to that aim to increase biochemical purity, for instance by cell-sorting genetically labeled synapses after biochemically enriching synaptosomes (Biesemann *et al.,* 2014).

Additional proteomic approaches that generated data on proteins of the synaptome include the enrichment of protein complexes using antibodies (Schwenk, Harmel, *et al.,* 2009; Klemmer *et al.,* 2009; Dosemeci *et al.,* 2007), genetic tags (Fernández *et al.,* 2009) and proximity labeling assays (Loh *et al.,* 2016; Cijsouw *et al.,* 2018). Quantitative studies that compared protein abundance levels between populations of synapses include an inter-species comparison of postsynaptic proteomes (Bayes, M. O. Collins, Croning, *et al.,* 2012), comparative analysis of protein-levels between regions of the human neocortex (M. Roy *et al.,* 2018), plasticity induced alterations of the synaptome (Rao-Ruiz *et al.,* 2015) and analyses of developmental stages in mammalian brains (Schwenk, Baehrens, *et al.,* 2014; McClatchy *et al.,* 2007).

While many candidate synaptic proteins suggested through proteomics have been confirmed over the past decades, a complete description of the synaptic proteome has not been attained yet. The proteomics studies described here are not without their share of challenges; delineating true- and false-positives in high-throughput proteomics remains challenging and more hypotheses are generated than can be verified through functional studies or low-throughput high-confidence microscopy assays. For example, a source of inaccuracies in proteomic subcellular localization, that ultimately leads to false-positives in synaptic parts lists, are biochemical impurities that arise from co-isolation of non-synaptic compartments which have similar biochemical properties as the target compartment (e.g. a protocol isolating synaptic vesicles may also yield non-synaptic vesicles of a similar size and density). Instead of focusing solely on iterative improvement of biochemical procedures, solutions to such problems may also be found beyond the biochemistry domain. For example by application of alternative experimental designs that include a diversity of control cases in combination with data analysis strategies tailored to this approach.

Interdisciplinary challenges in mass spectrometry and bioinformatics remain that, when addressed, could empower scientific progress in capturing the synaptic proteome. Improving data acquisition and algorithmic processing thereof could shed light on the dark proteome, those proteins present in synapses with low copy numbers or only present in a subpopulation of all synapses that are undetectable in proteomics at current sensitivity capabilities (Zubarev, 2013). Data analysis approaches that accurately separate proteins between experimental conditions, such as enrichment in PSD fraction as compared to control samples, are needed to reduce false positive rates in generating candidate synaptic protein sets.

Compared to the many data resources that have been generated so far, progress on integrating synaptome data has been slow. Independent efforts have been made to collect data as part of the gene ontology (GO) knowledgebase (Ashburner *et al.,* 2000), a database of synapse proteomics data (Pirooznia *et al.,* 2012) and experts in the fields have made use of curated protein lists to empower the analysis of genetic data in synaptic context (Lips *et al.,* 2012). Taking the next step in integrating available synaptic data into a consensus knowledgebase will require a combination of these three approaches; integrating various datasets that describe synaptic proteins (e.g. SynaptomeDB approach by Pirooznia *et al.* (2012)), scrutinizing literature to establish a high quality parts list of synaptic proteins (e.g. as done in Lips et al.) and establishing workflows and protocols for curation of these data to arrive at conclusions on synaptic localization and/or function of a protein (*modus operandi* of the GO consortium).

Outline of this thesis

The general aim of this thesis is the large-scale identification of synaptic proteins through interdisciplinary research that combines synapse research, proteomics and bioinformatics. Chapters two and three focus on method development to advance the sensitivity and specificity of quantitative proteomics by introducing a computational framework for handling missing data and a novel mass spectrometry acquisition strategy. Chapters four and five describe application of proteomics to delineate synaptic subcellular compartments and quantify stoichiometric differences of synaptic proteins between rodent and primate species. Chapter six describes a knowledgebase for the synapse and bioinformatics analyses thereof that demonstrates unique features of synapses and new links between synapses and disease.

Chapter **two** focusses on the missing data problem in label-free proteomics. Here, I first describe how the probability of missingness among replicate measurements in discovery proteomics increases as protein abundance decreases. We then propose a censoring model that takes the pattern of missingness into account in differential expression analysis and compare our model with four alternatives. Compared with four alternative models, the proposed model bests all alternative models when applied to a benchmark data set.

Chapter **three** explores an alternative data independent acquisition strategy for label-free proteomics and compares this to SWATH-MS. WiSIM-DIA combines conventional DIA with wide-SIM (wide selected-ion monitoring) windows to partition the precursor mass-over-charge space to produce high-quality precursor ion chromatograms. This improves MS1 peak area based quantification in a DIA strategy, in contrast to the SWATH-MS strategy that utilizes MS2 peak areas. Both strategies show strong overlap in the set of quantified peptides, but also exhibit unique advantages.

Chapter **four** applies proteomics together with a correlation profiling data analysis strategy to identify proteins enriched in the pre- and post-synapse. We quantified the proteomes of five biochemically isolated mouse brain cellular sub-fractions, with emphasis on synaptic compartments, from three brain regions, hippocampus, cortex and cerebellum. We demonstrated the expected co-fractionation of canonical synaptic proteins belonging to the same functional groups. The enrichment profiles also suggested the presence of many novel pre- and post-synaptic proteins, of which two were experimentally validated.

Chapter **five** applies SWATH-MS in a comparative study of hippocampal synaptic proteomes of rodents and primates. Using a data analysis strategy targeted to peptide sequences conserved among all species in the comparison, relative abundances were accurately compared and mapped to various functional groups of interest. Many differentially expressed proteins were detected between rodent and primate species, mostly with small foldchanges, and these proved highly enriched for plasticity-related proteins.

Chapter **six** describes an evidence-based, expert-curated knowledgebase for the synapse named SynGO. An elaborate ontology was designed for the synapse and used by world-wide domain experts to systematically annotate synaptic protein functions and locations based on published literature. Bioinformatics analyses reveal SynGO genes are exceptionally large, well conserved, and intolerant to mutations. Furthermore, a strong enrichment among genes associated with brain disorders was observed. All data was integrated into the Gene Ontology database and an online data analysis platform was developed to facilitate usage of the SynGO knowledgebase.

References

1. Abbott, L. & Regehr, W. G. Synaptic computation. *Nature* **431,** 796–803 (2004).
2. Abul-Husn, N. S. *et al.* Systems Approach to Explore Components and Interactions in the Presynapse. *Proteomics* **9,** 3303–15. ISSN: 1615-9861 (Electronic) 1615-9853 (Linking) (2009).
3. Aebersold, R., Agar, J. N., *et al.* How many human proteoforms are there? *Nature chemical biology* **14,** 206 (2018).
4. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537,** 347–355 (2016).
5. Altelaar, A. M. & Heck, A. J. Trends in ultrasensitive proteomics. *Current opinion in chemical biology* **16,** 206–213 (2012).

6. Altelaar, A. M., Munoz, J. & Heck, A. J. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature Reviews Genetics* **14,** 35–48 (2013).
7. Anderson, N. L. & Anderson, N. G. Proteome and proteomics: new technologies, new concepts, and new words. eng. *Electrophoresis* **19,** 1853–61. ISSN: 0173-0835. PMID: 9740045 (Aug. 1998).
8. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature genetics* **25,** 25–29 (2000).
9. Azevedo, F. A. C. *et al.* Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. eng. *The Journal of comparative neurology* **513,** 532–41. ISSN: 1096-9861. PMID: 19226510 (Apr. 2009).
10. Bayes, A., Collins, M. O., Croning, M. D., *et al.* Comparative Study of Human and Mouse Postsynaptic Proteomes Finds High Compositional Conservation and Abundance Differences for Key Synaptic Proteins. *PLoS One* **7,** e46683. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking) (2012).
11. Bayes, A., Collins, M. O., Reig-Viader, R., *et al.* Evolution of Complexity in the Zebrafish Synapse Proteome. *Nat Commun* **8,** 14613. ISSN: 2041-1723 (Electronic) 2041-1723 (Linking) (2017).
12. Biesemann, C. *et al.* Proteomic screening of glutamatergic mouse brain synaptosomes isolated by fluorescence activated sorting. *The EMBO journal* **33,** 157–170 (2014).
13. Boyken, J. *et al.* Molecular Profiling of Synaptic Vesicle Docking Sites Reveals Novel Proteins but Few Differences between Glutamatergic and GABAergic Synapses. *Neuron* **78,** 285–97. ISSN: 1097-4199 (Electronic) 0896-6273 (Linking) (2013).
14. Bruderer, R. *et al.* Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues. *Mol Cell Proteomics* **14,** 1400–10. ISSN: 1535-9484 (Electronic) 1535-9476 (Linking) (2015).
15. Cijsouw, T. *et al.* Mapping the proteome of the synaptic cleft through proximity labeling reveals new cleft proteins. *Proteomes* **6,** 48 (2018).
16. Citri, A. & Malenka, R. C. Synaptic plasticity: multiple forms, functions, and mechanisms. *Neuropsychopharmacology* **33,** 18–41 (2008).
17. Collins, B. C. *et al.* Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nature communications* **8,** 1–12 (2017).
18. Conrads, T. P. *et al.* Utility of Accurate Mass Tags for Proteome-Wide Protein Identification. *Anal Chem* **72,** 3349–54. ISSN: 0003-2700 (Print) 0003-2700 (Linking) (2000).

19. Cox, J. *et al.* MaxLFQ Allows Accurate Proteome-Wide Label-Free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction. *Mol. Cell Proteomics* **M113.031591,** 1–37 (July 2014).

20. De Jong, A. P., Schmitz, S. K., *et al.* Dendritic position is a major determinant of presynaptic strength. *Journal of Cell Biology* **197,** 327–337 (2012).

21. De Jong, A. P. & Verhage, M. Presynaptic signal transduction pathways that modulate synaptic transmission. *Current opinion in neurobiology* **19,** 245–253 (2009).

22. Dosemeci, A. *et al.* Composition of the Synaptic PSD-95 Complex. *Mol Cell Proteomics* **6,** 1749–60. ISSN: 1535-9476 (Print) 1535-9476 (Linking) (2007).

23. Drachman, D. A. Do we have brain to spare? eng. *Neurology* **64,** 2004–5. ISSN: 1526-632X. PMID: 15985565 (June 2005).

24. Edwards, A. M. *et al.* Too many roads not taken. *Nature* **470,** 163–165. https://doi.org/10.1038/470163a (Feb. 2011).

25. Emes, R. D. & Grant, S. G. Evolution of Synapse Complexity and Diversity. *Annual Review of Neuroscience* **35,** 111–131. https://doi.org/10.1146/annurev-neuro-062111-150433 (July 2012).

26. Eroglu, C. & Barres, B. A. Regulation of synaptic connectivity by glia. *Nature* **468,** 223–231 (2010).

27. Fernández, E. *et al.* Targeted tandem affinity purification of PSD-95 recovers core postsynaptic complexes and schizophrenia susceptibility proteins. *Molecular systems biology* **5,** 269 (2009).

28. Filiou, M. D. *et al.* Profiling of Mouse Synaptosome Proteome and Phosphoproteome by IEF. *Electrophoresis* **31,** 1294–301. ISSN: 1522-2683 (Electronic) 0173-0835 (Linking) (2010).

29. Firat-Karalar, E. N. *et al.* Proximity interactions among centrosome components identify regulators of centriole duplication. eng. *Current biology : CB* **24,** 664–70. ISSN: 1879-0445. PMID: 24613305 (Mar. 2014).

30. Frischknecht, R. *et al.* Brain Extracellular Matrix Affects AMPA Receptor Lateral Mobility and Short-Term Synaptic Plasticity. *Nat Neurosci* **12,** 897–904. ISSN: 1546-1726 (Electronic) 1097-6256 (Linking) (2009).

31. Gessulat, S. *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. eng. *Nature methods* **16,** 509–518. ISSN: 1548-7105. PMID: 31133760 (June 2019).

32. Gillet, L. C. *et al.* Targeted Data Extraction of the MS/MS Spectra Generated by Data-Independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Molecular & Cellular Proteomics* **11** (2012).

33. Gouw, J. W., Krijgsveld, J. & Heck, A. J. Quantitative proteomics by metabolic labeling of model organisms. *Molecular & cellular proteomics* **9,** 11–24 (2010).

34. Guan, S., Moran, M. F. & Ma, B. Prediction of LC-MS/MS Properties of Peptides from Sequence by Deep Learning. eng. *Molecular & cellular proteomics : MCP* **18,** 2099–2107. ISSN: 1535-9484. PMID: 31249099 (Oct. 2019).

35. Jahn, R. & Fasshauer, D. Molecular machines governing exocytosis of synaptic vesicles. *Nature* **490,** 201–207 (2012).

36. Jordan, B. A. *et al.* Identification and verification of novel rodent postsynaptic density proteins. *Molecular & Cellular Proteomics* **3,** 857–871 (2004).

37. Käll, L. *et al.* Semi-supervised learning for peptide identification from shotgun proteomics datasets. eng. *Nature methods* **4,** 923–5. ISSN: 1548-7091. PMID: 17952086 (Nov. 2007).

38. Karczewski, K. J. *et al.* The ExAC browser: displaying reference data information from over 60 000 exomes. eng. *Nucleic acids research* **45,** D840–D845. ISSN: 1362-4962. PMID: 27899611 (Jan. 2017).

39. Karpievitch, Y., Dabney, A. & Smith, R. Normalization and Missing Value Imputation for Label-Free LC-MS Analysis. *BMC bioinformatics* **13,** S5 (2012).

40. Kirkpatrick, D. S., Gerber, S. A. & Gygi, S. P. The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications. *Methods* **35,** 265–273 (2005).

41. Klemmer, P., Smit, A. & Li, K. Proteomics analysis of immuno-precipitated synaptic protein complexes. *Journal of proteomics* **72,** 82–90 (2009).

42. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. eng. *Nature* **536,** 285–91. ISSN: 1476-4687. PMID: 27535533 (Aug. 2016).

43. Li, K. W., Hornshaw, M. P., Van Der Schors, R. C., *et al.* Proteomics Analysis of Rat Brain Postsynaptic Density. Implications of the Diverse Protein Functional Groups for the Integration of Synaptic Physiology. *J Biol Chem* **279,** 987–1002. ISSN: 0021-9258 (Print) 0021-9258 (Linking) (2004).

44. Li, K. W., Hornshaw, M. P., der Schors, R. C. V., *et al.* Proteomics Analysis of Rat Brain Postsynaptic Density. *Journal of Biological Chemistry* **279,** 987–1002. https://doi.org/10.1074/jbc.m303116200 (Oct. 2003).

45. Li, K. w. *et al.* Organelle proteomics of rat synaptic proteins: correlation-profiling by isotope-coded affinity tagging in conjunction with liquid chromatography-tandem mass spectrometry to reveal post-synaptic density specific proteins. *Journal of proteome research* **4,** 725–733 (2005).

46. Lips, E. S. *et al.* Functional gene group analysis identifies synaptic gene groups as risk factor for schizophrenia. *Molecular psychiatry* **17,** 996–1006 (2012).

47. Liu, Y. *et al.* Quantitative measurements of N-linked glycoproteins in human plasma by SWATH-MS. *Proteomics* **13,** 1247–1256 (2013).

48. Loh, K. H. *et al.* Proteomic analysis of unbounded cellular compartments: synaptic clefts. *Cell* **166,** 1295–1307 (2016).

49. Ludwig, C. *et al.* Data-independent acquisition-based SWATH-MS for quantitative proteomics: a tutorial. eng. *Molecular systems biology* **14,** e8126. ISSN: 1744-4292. PMID: 30104418 (Aug. 2018).

50. Ma, K., Vitek, O. & Nesvizhskii, A. I. A statistical model-building perspective to identification of MS/MS spectra with PeptideProphet. *BMC bioinformatics* **13,** S1 (2012).

51. Martin, S. J., Grimwood, P. D. & Morris, R. G. Synaptic plasticity and memory: an evaluation of the hypothesis. *Annual review of neuroscience* **23,** 649–711 (2000).

52. McClatchy, D. B. *et al.* Quantification of the synaptosomal proteome of the rat cerebellum during post-natal development. *Genome research* **17,** 1378–1388 (2007).

53. McGaugh, J. L. Memory–a century of consolidation. *Science* **287,** 248–251 (2000).

54. Michalski, A., Cox, J. & Mann, M. More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority Is Inaccessible to Data-Dependent LC-MS/MS. *J. Proteome Res.* **10,** 1785–1793 (Apr. 2011).

55. Mitchell, N. *et al.* Sonic hedgehog regulates presynaptic terminal size, ultrastructure and function in hippocampal neurons. *Journal of Cell Science* **125,** 4207–4213. https://doi.org/10.1242/jcs.105080 (May 2012).

56. Morciano, M., Beckhaus, T., *et al.* The Proteome of the Presynaptic Active Zone: From Docked Synaptic Vesicles to Adhesion Molecules and Maxi-Channels. *J Neurochem* **108,** 662–75. ISSN: 1471-4159 (Electronic) 0022-3042 (Linking) (2009).

57. Morciano, M., Burré, J., *et al.* Immunoisolation of two synaptic vesicle pools from synaptosomes: a proteomics analysis. *Journal of neurochemistry* **95,** 1732–1745 (2005).

58. Myers, S. A. *et al.* Discovery of proteins associated with a predefined genomic locus via dCas9-APEX-mediated proximity labeling. eng. *Nature methods* **15,** 437–439. ISSN: 1548-7105. PMID: 29735997 (June 2018).

59. Neniskyte, U. & Gross, C. T. Errant gardeners: glial-cell-dependent synaptic pruning and neurodevelopmental disorders. *Nature Reviews Neuroscience* **18,** 658 (2017).

60. Nesvizhskii, A. I. & Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. eng. *Molecular & cellular proteomics : MCP* **4,** 1419–40. ISSN: 1535-9476. PMID: 16009968 (Oct. 2005).

61. Phillips, G. R. *et al.* Proteomic Comparison of Two Fractions Derived from the Transsynaptic Scaffold. *J Neurosci Res* **81,** 762–75. ISSN: 0360-4012 (Print) 0360-4012 (Linking) (2005).

62. Pirooznia, M. *et al.* SynaptomeDB: an ontology-based knowledgebase for synaptic genes. *Bioinformatics* **28,** 897–899 (2012).

63. Rao-Ruiz, P. *et al.* Time-dependent changes in the mouse hippocampal synaptic membrane proteome after contextual fear conditioning. eng. *Hippocampus* **25,** 1250–61. ISSN: 1098-1063. PMID: 25708624 (Nov. 2015).

64. Regehr, W. G. Short-term presynaptic plasticity. *Cold Spring Harbor perspectives in biology* **4,** a005702 (2012).

65. Regehr, W. G., Carey, M. R. & Best, A. R. Activity-dependent regulation of synapses by retrograde messengers. *Neuron* **63,** 154–170 (2009).

66. Roberts, A. C. & Glanzman, D. L. Learning in Aplysia: looking at synaptic plasticity from both sides. *Trends in neurosciences* **26,** 662–670 (2003).

67. Rost, H. L. *et al.* OpenSWATH Enables Automated, Targeted Analysis of Data-Independent Acquisition MS Data. *Nat Biotechnol* **32,** 219–23. ISSN: 1546-1696 (Electronic) 1087-0156 (Linking) (2014).

68. Roux, K. J., Kim, D. I. & Burke, B. BioID: a screen for protein-protein interactions. eng. *Current protocols in protein science* **74,** 19.23.1–19.23.14. ISSN: 1934-3663. PMID: 24510646 (Nov. 2013).

69. Roy, M. *et al.* Proteomic analysis of postsynaptic proteins in regions of the human neocortex. eng. *Nature neuroscience* **21,** 130–138. ISSN: 1546-1726. PMID: 29203896 (Jan. 2018).

70. Saheki, Y. & De Camilli, P. Synaptic vesicle endocytosis. *Cold Spring Harbor perspectives in biology* **4,** a005645 (2012).

71. Scannevin, R. H. & Huganir, R. L. Postsynaptic organisation and regulation of excitatory synapses. *Nature Reviews Neuroscience* **1,** 133–141 (2000).

72. Schoch, S. & Gundelfinger, E. D. Molecular organization of the presynaptic active zone. *Cell and tissue research* **326,** 379–391 (2006).

73. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473,** 337–342 (2011).

74. Schwenk, J., Baehrens, D., *et al.* Regional Diversity and Developmental Dynamics of the AMPA-Receptor Proteome in the Mammalian Brain. *Neuron* **84,** 41–54. ISSN: 1097-4199 (Electronic) 0896-6273 (Linking) (2014).

75. Schwenk, J., Harmel, N., *et al.* Functional proteomics identify cornichon proteins as auxiliary subunits of AMPA receptors. *Science* **323,** 1313–1319 (2009).

76. Sheng, M. & Hoogenraad, C. C. The postsynaptic architecture of excitatory synapses: a more quantitative view. *Annu. Rev. Biochem.* **76,** 823–847 (2007).

77. Sheng, M. & Kim, E. The postsynaptic organization of synapses. *Cold Spring Harbor perspectives in biology* **3,** a005678 (2011).

78. Shteynberg, D. *et al.* iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. eng. *Molecular & cellular proteomics : MCP* **10,** M111.007690. ISSN: 1535-9484. PMID: 21876204 (Dec. 2011).

79. Sialana, F. J. *et al.* Mass Spectrometric Analysis of Synaptosomal Membrane Preparations for the Determination of Brain Receptors, Transporters and Channels. *Proteomics* **16,** 2911–2920. ISSN: 1615-9861 (Electronic) 1615-9853 (Linking) (2016).

80. Smith, L. M. & Kelleher, N. L. Proteoform: a single term describing protein complexity. *Nature methods* **10,** 186–187 (2013).

81. Steen, H. & Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nature reviews Molecular cell biology* **5,** 699–711 (2004).

82. Sudhof, T. C. The Presynaptic Active Zone. *Neuron* **75,** 11–25. ISSN: 1097-4199 (Electronic) 0896-6273 (Linking) (2012).

83. Südhof, T. C. The synaptic vesicle cycle. *Annu. Rev. Neurosci.* **27,** 509–547 (2004).

84. Takamori, S. *et al.* Molecular Anatomy of a Trafficking Organelle. *Cell* **127,** 831–846. https://doi.org/10.1016/j.cell.2006.10.030 (Nov. 2006).

85. Tang, B. *et al.* Fmr1 deficiency promotes age-dependent alterations in the cortical synaptic proteome. *Proceedings of the National Academy of Sciences* **112,** E4697–E4706 (2015).

86. Taoufiq, Z. *et al.* Hidden proteome of synaptic vesicles in the mammalian brain. *Proceedings of the National Academy of Sciences* **117,** 33586–33596 (2020).

87. Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563,** 72–78. https://doi.org/10.1038/s41586-018-0654-5 (Oct. 2018).

88. Tiwary, S. *et al.* High-quality MS/MS spectrum prediction for data-dependent and data-independent acquisition data analysis. eng. *Nature methods* **16,** 519–525. ISSN: 1548-7105. PMID: 31133761 (June 2019).

89. Turrigiano, G. G. & Nelson, S. B. Homeostatic plasticity in the developing nervous system. *Nature reviews neuroscience* **5,** 97–107 (2004).

90. Volknandt, W. & Karas, M. Proteomic analysis of the presynaptic active zone. *Experimental Brain Research* **217,** 449–461. https://doi.org/10.1007/s00221-012-3031-x (Feb. 2012).

91. Whitlock, J. R. *et al.* Learning induces long-term potentiation in the hippocampus. *science* **313,** 1093–1097 (2006).

92. Wiese, S. *et al.* Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research. *Proteomics* **7,** 340–350 (2007).

93. Wilhelm, B. G. *et al.* Composition of Isolated Synaptic Boutons Reveals the Amounts of Vesicle Trafficking Proteins. *Science* **344,** 1023–8. ISSN: 1095-9203 (Electronic) 0036-8075 (Linking) (2014).

94. Zhu, W., Smith, J. W. & Huang, C.-M. Mass spectrometry-based label-free quantitative proteomics. *Journal of Biomedicine and Biotechnology* **2010** (2009).

95. Zubarev, R. A. The challenge of the proteome dynamic range and its implications for in-depth proteomics. *Proteomics* **13,** 723–726 (2013).

96. Zucker, R. S. & Regehr, W. G. Short-term synaptic plasticity. *Annual review of physiology* **64,** 355–405 (2002).

2

An Empirical Bayesian Random Censoring Threshold Model Improves Detection of Differentially Abundant Proteins

Frank Koopmans[1][2], L. Niels Cornelisse[1], Tom Heskes[2], Tjeerd M. H. Dijkstra[2]

[1] Department of Functional Genomics, Center for Neurogenomics Cognitive Research,
VU University, Amsterdam, The Netherlands
[2] Machine Learning Group, Institute for Computing and Information Sciences,
Radboud University, Nijmegen, The Netherlands

A challenge in proteomics is that many observations are missing with the probability of missingness increasing as abundance decreases. Adjusting for this informative missingness is required to assess accurately which proteins are differentially abundant. We propose an Empirical Bayesian Random Censoring Threshold (EBRCT) model that takes the pattern of missingness in account in the identification of differential abundance. We compare our model with four alternatives, one that considers the missing values as Missing Completely At Random (MCAR model), one with a fixed censoring threshold for each protein species (fixed censoring model) and two imputation models, k-nearest neighbors (IKNN) and Singular Value Thresholding (SVTI). We demonstrate that the EBRCT model bests all alternative models when applied to the CPTAC study 6 benchmark dataset. The model is applicable to any label-free peptide or protein quantification pipeline and is provided as an R script.

This chapter has been published in *Journal of proteome research 13, no. 9 (2014)*

Introduction

Proteomics has become a popular technique for identifying and quantifying proteins in complex biological samples with applications ranging from fundamental research on molecular pathways to clinical studies on disease biomarkers (Eidhammer *et al.,* 2013). Progress in liquid chromatography, mass spectrometry and software for the identification and quantification of proteins has made it possible to identify thousands of protein species in one discovery proteomics experiment. There are several methods for the quantification of proteins (Neilson *et al.,* 2011; Ong *et al.,* 2002; Ross *et al.,* 2004), in this research we focus on label-free quantification using MS1 peak intensities (Cox & Mann, 2008) which is more accurate than quantification based on spectral counting (Choi *et al.,* 2012; Milac *et al.,* 2012).

Differential abundance analysis is a common application of label-free proteomics with the goal to identify protein species that are differently abundant between case (treatment) and control (reference) conditions. The natural variation between biological samples and from sample preparation (Piehowski *et al.,* 2013) combined with the large number of proteins identified makes this goal difficult. The characteristics of high variability, small number of replicates and large number of variables are shared with gene expression analysis, see Ji & Liu (2010) for a tutorial review. Because of the large fraction of missing data in discovery proteomics data sets, methods developed for gene expression analysis are not directly applicable. Missing data occurs when a protein is detected in one replicate but not the other, a common phenomenon in discovery proteomics where more than 20% of the data can be missing. There are several causes for missingness, ranging from sample preparation to equipment variation (Piehowski *et al.,* 2013). One cause is the stochastic process of peptide selection in Data Dependent Acquisition (DDA), which selects the most abundant peaks in the MS1 spectrum for MS2 fragmentation. Consequently, variations in sample preparation may cause low abundant peptides to be selected for MS2 in one replicate and not in another resulting in missing data. For a review of missing data issues covering also other types of mass spectrometry data see the introduction of the paper by Taylor *et al.* (2013).

In general, low abundant proteins are more frequently missing in experiment replicates. Michalski *et al.* (2011) that more abundant peptides in a mixture are typically targeted for MS2 and eventually identified. As reported in previous studies (Karpievitch, Stanley, *et al.,* 2009), there is a negative correlation between missingness and peptide (or protein) abundance. We show this effect in the table of content graphic for the yeast proteins in the CPTAC study 6 data. For statistical analysis of quantitative proteomics data, one has to decide how to interpret the missing data. Some methods assume that missing data is not informative (Clough *et al.,* 2012) and therefore Missing Completely At Random (MCAR) (Little & Rubin, 2002). This ignores the negative correlation between missingness and abundance. Consequently, the MCAR model tends to overestimate the abundance of proteins with missing values since it is likely that the observed values were higher by chance, see Figure 2.2 below and Figure 1 of Karpievitch, Dabney, *et al.* (2012). Other models assume a fixed detection threshold that determines whether a peptide is observed (Karpievitch, Stanley, *et al.,* 2009; Choi *et al.,* 2012; Paulovich *et al.,* 2010; Taylor *et al.,* 2013). These models capture the negative correlation between missingness and abundance, however as we show below they suffer from poor performance. Taking a page from gene expression analysis (Aittokallio, 2010) we also evaluate two missing value imputation algorithms and show that they perform on average better than the MCAR model. However, their improvement relative to the MCAR model is not consistent (sometimes worse but mostly better) and the estimate of the standard deviation of protein abundance is inaccurate. We introduce a novel model, the Empirical Bayesian Random Censoring Threshold (EBRCT) model that combines censoring with regularization of parameter estimates in a Bayesian fashion. Application of the models to the benchmark CPTAC study 6 data set demonstrates an increase in sensitivity compared to alternative methods.

Materials and Methods

CPTAC study 6 Data Set

To compare the models in a real-world situation we use a data set where we know the set of differentially abundant proteins a priori. The Clinical Proteomic Technology Assessment for Cancer (CPTAC) study 6 data set (Paulovich *et al.,* 2010) consists of five experimental conditions, labeled *A* to *E*, where increasing quantities of 48 Sigma UPS1 proteins were added to a constant yeast background. We do not analyze the quality control conditions that were part of study 6 here. In condition *A* 0.25 fmol/uLof UPS1 protein mixture was spiked in 60 ng/uL of yeast protein mixture and in subsequent conditions the concentration of UPS1 increased with a factor of 3. The experiments were repeated in three laboratories using the same model of mass spectrometer and experimental protocol (Paulovich *et al.,* 2010) with three technical replicates in each laboratory. We focus on two sets of measurements done by two laboratories: *OrbitrapO@65* and *OrbitrapW@56*. We exclude the data from laboratory *OrbitrapE@86* as its results deviate from the other two due to settings of the spectrometer, as noticed in the original study. We only compare four adjacent concentration condition pairs (A-B, B-C, C-D, D-E) to keep the differences in spiked-in protein abundance as small as possible and thereby challenging to distinguish from the yeast background.

We submitted the raw data to MaxQuant (Cox & Mann, 2008) (version 1.3.0.5) for protein identification and label-free quantification. Using default settings, we searched against the UniProt (version 2013-01) yeast proteome fasta database and a fasta database containing the UPS1 sequences. A peptide and protein False Discovery Rate (FDR) of 0.01 was used for high confident identifications. We used MaxQuant's Label-Free Quantification (LFQ) algorithm (Cox, Hein, *et al.,* 2014) to normalize the data using a minimum of 1 for the number of peptides needed for quantification by LFQ. In performing the normalization before the statistical analysis of the missing data we follow the advice of Karpievitch, Dabney, *et al.* (2012) who show that this order performs better than the reverse. In addition to quantification by LFQ we also quantified protein abundance by summing their peptide intensities mainly for the purpose of showing that LFQ improves performance considerably.

The input of our analysis consists of four pairs of data matrices with LFQ values.

We log10 transformed the abundance values as log abundances are often approximately normally distributed (Karpievitch, Dabney, *et al.,* 2012; Taylor *et al.,* 2013). Each data matrix is about 1200 rows (protein species) by 6 columns (2 laboratories and 3 technical replicates). In our analysis we make no distinction between laboratory and replicate and hence consider the data set to have 6 replicates. We can do this as one of the purposes of the CPTAC study is to reduce variability between different laboratories and in case of the two laboratories that we use, the researchers succeeded in their goal. Note that this pooling of laboratories does not favor any of the models as plotted in Fig. 2.4 below. We only analyzed those protein species from each pair of concentration conditions that had at least one observation (and hence maximally five missing values) for each member of the pair. This constraint ensures that the MCAR model can make a prediction for each protein species and makes the statistical comparison of models straightforward. It would be possible to relax this constraint for the EBRCT, IKNN and SVTI models that we introduce later but such an analysis is beyond the scope of this paper.

In Figure 2.1 we report summary statistics of the mean protein abundance split out according to concentration condition (A through E) and UPS1 (foreground) or yeast (background) protein. Means over the six repetitions are calculated for those proteins that have at least one observation in each member of a concentration pair leading to slightly different means for those concentration conditions that occur in more than one pair. This effect is most noticeable for the UPS1 proteins in condition B. Unsurprisingly, one can observe that the abundances of the UPS1 proteins differ between members of a pair whereas the abundances of the yeast proteins stay the same (but see below).

The mean differences over the UPS1 proteins between concentration conditions are reported in the header of each panel. We expect these to equal to $\log_{10}(3) = 0.477$ and indeed the differences scatter around these values. We note that the differences are larger than the expected difference of 0.477 for the A vs B and B vs C pair and smaller for the D vs E pair. Also note that the abundance of the UPS1 proteins is quite different from the background for the C vs D and D vs E pair with already some difference observable for the B vs C comparison. As we believe that in most biological experiments the set of differentially abundant proteins is similar in abundance to the non-differentially abundant (background) proteins, we consider the A vs B comparison the most critical one. The differences in mean over the yeast proteins between concentration conditions are smaller, 0.006, -0.004, -0.034, -0.055 for the A vs B, B vs C, C vs D and D vs E pair resp. Note that the smallest change in UPS1 and the largest (in absolute value) change in yeast mean abundance occurs for the D vs E pair. This is a side-effect from spiking large amounts of protein into the yeast background as noted in the original study and analyzed in detail in Milac *et al.* (2012). Another side-effect is the reduced identification rate of yeast protein species with increasing concentration, see Figure S1 in the Supporting Information. From the same figure one can observe that the number of protein species reported by the LFQ algorithm is about 5% less than reported by summing the intensities, presumably due to the constraint of sharing peptides that forms the basis of the LFQ algorithm.

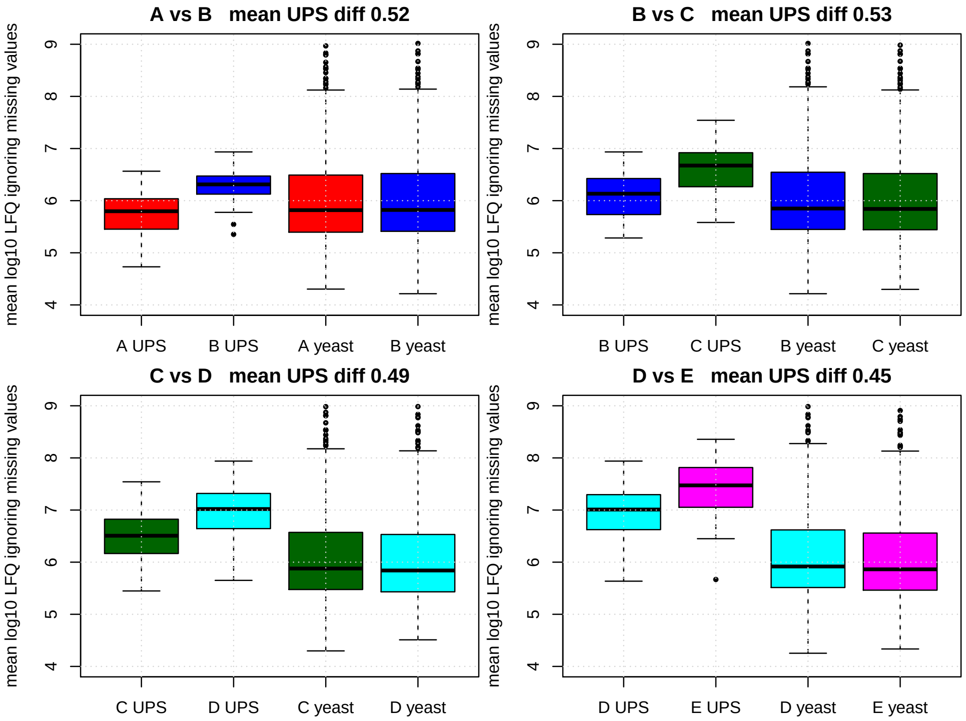

Figure 2.1: Box plots of mean log10 protein abundance values from MaxQuant's LFQ algorithm for each paired comparison split out according to concentration condition (A through E) and UPS1 (foreground) or yeast (background) protein. Concentration conditions are color-coded and per-protein means are calculated ignoring the missing values (MCAR model).

In Table 2.1 we report the fraction missing data. For the UPS1 proteins the fraction missing decreases with increasing concentration, an effect consistent with censoring. The fraction missing is constant for the yeast proteins and hovers around 26%. Further, notice that missingness is higher at the peptide level than at the protein level. Lastly, the fraction missing in the D and E concentration conditions is much lower than yeast background, yet another reason to consider the D vs E concentration pair as less important.

Overview of Random Censoring Threshold Model

In Figure 2.2 we provide a cartoon of a discovery-based proteomics data set from replicate (top panel), through mass-spectrometry (middle panel) to a censoring mechanism causing missingness (bottom panel). One peptide (color-coded green and called "high") is relatively abundant and all its MS1 peaks are observed, one peptide (color-coded orange and called "mid") is medium abundant and some of its MS1 peaks are missing (the MS1 peak is not detected or identified through MS/MS) and one peptide (color-coded magenta and called "low") is low abundant and only one of its MS1 peaks is observed.

	A	B	C	D	E
peptide inten UPS1	0.68	0.59	0.49	0.42	0.35
peptide inten yeast	0.49	0.49	0.48	0.49	0.48
protein inten UPS1	0.61	0.36	0.17	0.04	0.01
protein inten yeast	0.26	0.27	0.26	0.28	0.29
protein LFQ UPS1	0.60	0.35	0.17	0.04	0.02
protein LFQ yeast	0.25	0.27	0.25	0.27	0.27

Table 2.1: Fraction missing data for those protein species that are observed in at least one out of six replicates. Columns are for the indicated concentration condition and rows are for the peptide intensities (top two rows), summed intensities (middle two rows) and LFQ values (bottom two rows). Odd rows contain the fraction missing of UPS1 proteins and even rows of yeast proteins.

The heart of our Empirical Bayesian Random Censoring Threshold (EBRCT) model is detailed in the bottom panel of Figure 2.2 where we indicate the detected protein species with a closed symbol and the missing ones with an open one. Protein abundances can be calculated from peptide intensities with a label-free quantification method that includes the peptide to protein roll-up, for example MaxQuant (Cox & Mann, 2008), LFQuant (Zhang *et al.,* 2012) or OpenMS (Weisser *et al.,* 2013). We posit a random censoring threshold as the cause of missingness. As the censoring thresholds themselves are random, we illustrate a random sample from them with grey horizontal lines. For each protein (both observed and missing) we draw one censoring threshold, this pairing is indicated with thin vertical lines. If the protein abundance falls above its censoring threshold, it is observed and if it falls below its censoring threshold, it is not observed. The solid and dotted lines drawn in the color of each of the proteins denote the mean of the observed (MCAR mean) and the mean of all the data (both observed and missing). As our model takes the censored data into account, the estimates of protein abundance of our model are close to the mean of all the data whereas the mean of only the observed data is biased. The bias depends on the fraction missing: when all data are observed as for the high abundant protein species, the bias is negligible, whereas the bias is large when many replicates are missing as for the low abundant protein species. The same argument is made by Karpievitch, Dabney, *et al.* (2012) in their Figure 1, except that they consider a fixed censoring threshold.

Empirical Bayesian Random Censoring Threshold Model

In this section we summarize the Empirical Bayesian Random Censoring Threshold model (EBRCT), more detail is provided in the Supporting Information. As input we use a pair of log10 transformed protein abundances $y_{jk_1}^{(1)}$ and $y_{jk_2}^{(2)}$ generated by label-free quantification software e.g. MaxQuant (Cox & Mann, 2008), LFQuant (Zhang *et al.,* 2012) or OpenMS (Weisser *et al.,* 2013). The purpose of the statistical analysis is to rank-order the protein species indexed by j according to the estimate of differential abundance between two experimental conditions indexed with superscripts $^{(1)}$ and $^{(2)}$. This purpose is similar to differential gene expression analysis from microarray or RNAseq data see e.g. Soneson and Delorenzi (Soneson

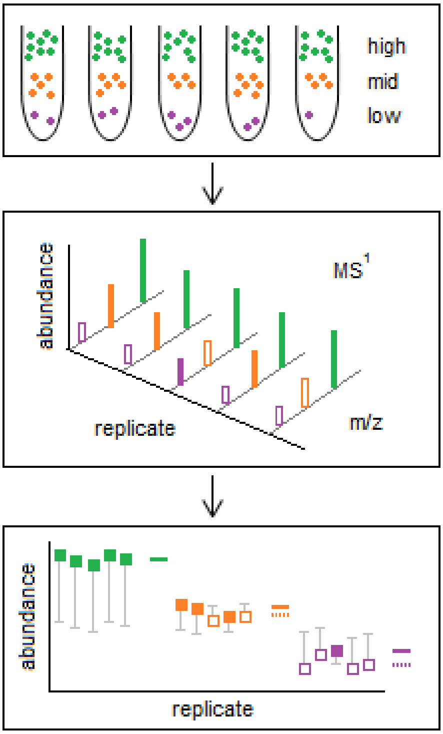

Figure 2.2: A cartoon to illustrate how missing data affects quantitative peptide abundance over replicates. Top panel: five replicate samples with three color-coded peptides each with a different concentration, indicated by the number of dots. green = high, orange = mid and magenta = low abundant. Middle panel: MS1 spectra from the five replicates. Peak height is proportional to concentration. Note the missing data (peak is not detected or identified) denoted by open bars in the mid-abundant peptide (orange) and in the low-abundant peptide (magenta). Bottom panel: estimates of protein abundance from a label-free quantification algorithm that includes the roll-up from peptide- to protein-level. Solid dots denote observed abundances whereas open dots denote missing values. To each protein we link a hypothetical censoring threshold denoted with a short grey line. The observations are above their censoring threshold whereas the missing data fall below. Colored horizontal lines denote the mean abundance as calculated from the observed data only, whereas the dashed lines denote the actual mean (including the missing data)

& Delorenzi, 2013) for review. The data consist of $j \in (1, \dots, J)$ protein species with $k_{1,2} \in (1, \dots, K_{1,2})$ replicates. For simplicity we make no distinction between biological and technical replicates. In a typical discovery-based proteomics experiment J is on the order of 1000 to 5000 and $K_{1,2}$ between 2 and 6. To specify the statistical model we introduce the complete-data abundances $x_{jk}^{(l)}$ with $l \in (1, 2)$ and the censoring thresholds $c_{jk}^{(l)}$. These are related to the observed protein abundances

$y_{jk}^{(l)}$ via the following generative model:

$$y_{jk}^{(l)} = \begin{cases} x_{jk}^{(l)} & \text{if} \quad x_{jk}^{(l)} \geq c_{jk}^{(l)} \quad \text{observed} \\ \text{NA} & \text{if} \quad x_{jk}^{(l)} < c_{jk}^{(l)} \quad \text{censored} \end{cases} \tag{2.1}$$

$$x_{jk}^{(l)} \sim \mathcal{N}(\mu_{x,j}^{(l)}, \sigma_{x,j}^2) \tag{2.2}$$

$$c_{jk}^{(l)} \sim \mathcal{N}(\mu_{c,j}, \sigma_c^2), \tag{2.3}$$

where we coded the missing observations $y_{jk}^{(l)}$ as NA for "Not Available", the convention used by R (R Core Team, 2013) . We provide a graphical summary of the model in plate notation in Figure 2.3. We assume that both the complete-data abundances $x_{jk}^{(l)}$ and the censoring thresholds c_{jk} are normally distributed (denoted by $\mathcal{N}$) and statistically independent. Further, we assume that the abundance of each protein species j has an independent normal distribution with mean $\mu_{x,j}^{(l)}$ and variance $\sigma_{x,j}^2$. Thus, the means can be different between the experimental conditions and the variances are shared. Lastly, we assume that the censoring threshold means $\mu_{c,j}$ are shared between conditions but can differ between protein species and the variance σ_c^2 is shared between all proteins species and conditions. Given a pair of conditions with J protein species each, our model has $4J + 1$ parameters: $2J$ abundance means for conditions 1and 2, J abundance variances for both conditions, J censoring threshold means for both conditions and 1 for the variance of the censoring threshold.

Our choices for the parameters were guided by parsimony and accuracy, i.e. we strive for a model with the fewest number of parameters that detects all differentially abundant proteins. Thus, the abundance means $\mu_{x,j}^{(l)}$ could differ for each protein species and experimental condition, which follows from our goal of detecting the differentially abundant proteins. We opted to pool the abundance variances $\sigma_{x,j}^2$ and censoring threshold means $\mu_{c,j}$ over experimental conditions as this reduced the number of parameters. A version of the model where we did not pool the abundance variances over experimental conditions but did pool the censoring threshold means over both proteins species and experimental conditions lead to inferior performance (see also the section on "Alternative models and Model Comparison Procedure"). Lastly, we pooled the censoring threshold variance σ_c^2 over all data, as this parameter was considered more of a "nuisance".

To estimate these parameters, we formulate a Bayesian variant of the model. We extend the model with conjugate priors on the parameters $\mu_{x,j}^{(l)}$, $1/\sigma_{x,j}^2$, $\mu_{c,j}$ and $1/\sigma_c^2$. Explicitly, we use a conjugate normal-gamma prior (Hoff, 2009) for the

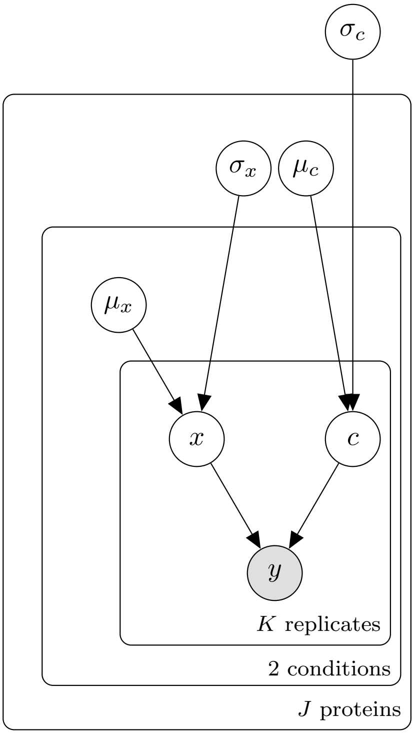

Figure 2.3: A plate version of the EBRCT model of eqs. 2.1, 2.2 and 2.3. Open circles denote latent variables and shaded circles denote observed variables. The boxes denote replicated variables with the replication level indicated in the lower right corner of every box. The arrows denote the dependencies.

complete-data protein abundances x and the censoring threshold c:

$$\mu_{x,j}^{(l)} \sim \mathcal{N}(m_{x0}, \frac{\sigma_{x,j}^2}{k_{x0}}) \tag{2.4}$$

$$\sigma_{x,j}^{-2} \sim \mathcal{G}(a_{x0}, b_{x0}) \tag{2.5}$$

$$\mu_{c,j} \sim \mathcal{N}(m_{c0}, \frac{\sigma_c^2}{k_{c0}}) \tag{2.6}$$

$$\sigma_c^{-2} \sim \mathcal{G}(a_{c0}, b_{c0}), \tag{2.7}$$

with $\mathcal{G}$ denoting the Gamma distribution, parametrized with shape and rate. The prior for the means $\mu_{x,j}$ and μ_c depends on the variances $\sigma_{x,j}^2$ and σ_c^2 which is necessary to obtain conjugacy (Gelman, Carlin, *et al.*, 2004, section 3.3) . We introduced eight hyperparameters for the prior: one mean and one pseudo count for each mean ($\mu_{x,j}$) and inverse variance ($1/\sigma_{x,j}^2$) of the protein abundance and mean (μ_c) and inverse variance ($1/\sigma_c^2$) of the censoring threshold. The four means are denoted by m_{x0}, $1/\sigma_{x0}^2$, m_{c0} and $1/\sigma_{c0}^2$ and the four pseudo counts by k_{x0}, a_{x0}, k_{c0} and a_{c0}. Note that we could have introduced separate priors for the means of each condition but chose not to do so as we are assuming that the number of differentially abundant protein species is small relative to their total number. The same assumption underlies the use of a single abundance variance estimate, pooling over both conditions. The relationship between the second parameter of the Gamma distribution (b_{x0} and b_{c0}) and the means of the inverse variances ($1/\sigma_{x0}^2$ and $1/\sigma_{c0}^2$) is:

$$\begin{aligned} b_{x0} &= \frac{a_{x0}}{2}\frac{1}{\sigma_{x0}^2} \\ b_{c0} &= \frac{a_{c0}}{2}\frac{1}{\sigma_{c0}^2}. \end{aligned}$$

We obtain values for the mean hyper parameters m_{x0} and $1/\sigma_{x0}^2$ from mean and standard deviation of pooled data where we pool all observations together ignoring the missing values. Values for mean hyper parameters m_{c0} and $1/\sigma_{c0}^2$ were obtained similarly but only taking the minimum observed value of each proteins species (pooled over condition) into account. Using the data to specify the prior explains the empirical Bayes moniker of our model.

We specify the pseudo count hyper parameters as fractions of the relevant amount of data, $(K_1+K_2)/2$ for the abundance means, (K_1+K_2) for the abundance variances and the censor threshold means and $J(K_1+K_2)$ for the censor threshold variance. These fractions are denoted by $f_{\mu x}, f_{\sigma x}, f_{\mu c}, f_{\sigma c}$ in the supporting information and can be viewed as a weighting of the prior relative to the data that have a weight of 1 by definition. For the hyper parameter k_{x0} we use a fraction of 0.01 implying a vague prior. To support the goal of accurate differential abundance analysis, we wanted to bias the mean abundance estimates as little as possible hence the small fraction of 0.01. In contrast, for a_{x0} we used a fraction of 2 leading to a strongly informative prior. This was necessary because of an ambiguity in our model in case most data are missing: in that case a range of low values of mean abundance can trade off with a range of big values for the variance of the abundance and keep the likelihood almost the same. The likelihood surface is very flat and the Gibbs sampler converges extremely slowly. In fact this ambiguity exists for all censoring models and can also be viewed as a consequence of an overparametrized model. In case only one or two data points are observed (and the rest missing) it is difficult to reliably fit a model with three (fixed censoring threshold model, see below) or four parameters (random censoring threshold model) without stabilization. Presumably, this ambiguity is also the reason that Taylor *et al.* (2013) only fit their fixed censoring threshold model to those records on the glycomics data with a minimum of three observations. For the hyper parameter k_{c0} we use a fraction of 0.2 implying a weakly informative prior. The results are not very sensitive to this parameter, we found this value to give best performance. Lastly, we use a fraction of 0.01 for a_{c0} leading to a vague prior on the variance of the censoring threshold. As this variance is estimated from all pooled data, its estimate is quite accurate and a vague prior suffices.

We obtain statistical samples from the posterior with a Gibbs sampler as detailed in the supplementary information. Briefly, the sampler obtains initial estimates of each of the 4J + 1 parameters by sampling from eqs. 2.4, 2.5, 2.6 and 2.7. Then it iterates over the following four steps: (1) obtain a sample from the complete data protein abundances $x_{jk}^{(l)}$ and censoring thresholds $c_{jk}^{(l)}$ by sampling from eqs. 2.2 and 2.3. (2) Calculate the sufficient statistics for $x^{(l)}$ and $c^{(l)}$, the means and variances. (3) Update the eight hyper parameters which can be done in closed-form as the priors are conjugate and (4) obtain a sample from the four parameters using eqs. 2.4, 2.5, 2.6 and 2.7. Convergence is generally excellent so we routinely run with 200 burn-in samples and 1000 samples for posterior estimation. We always run three independent Markov chains. The program runs about 80 samples per second for 1,200 protein species and 6 replicates for each condition on a single core of an Intel i7-2635QM CPU running at 2.0 GHz (2011 MacBook Pro 15 inch). As a random censoring threshold is not possible with standard software for posterior sampling like BUGS (Lunn *et al.,* 2012) or JAGS (Plummer *et al.,* 2003) we implemented the sampler in R (R Core Team, 2013) .

Alternative Models and Model Comparison Procedure

We consider four alternatives to our EBRCT model, Missing Completely at Random (MCAR) (Clough *et al.,* 2012), fixed censoring (FCEN) (Karpievitch, Stanley, *et al.,* 2009; Taylor *et al.,* 2013), imputation by k-nearest neighbors (IKNN) (Troyanskaya *et al.,* 2001; Taylor *et al.,* 2013) and Singular Value Thresholding Imputation (SVTI) (Candes & Plan, 2010). There are many imputation methods designed for imputation of microarray data (Aittokallio, 2010) and the purpose of this paper is not a review of imputation models hence we limited our analysis to IKNN and SVTI as they both performed better than MCAR. We also tested SVD imputation (Aittokallio, 2010) but found its performance worse than MCAR and did not consider it further. In an earlier version of this paper we tested a version of the EBRC model where each concentration condition was fitted independently with a single censoring threshold distribution for all protein species. This version performed about intermediate between the two imputation models. As this version of EBRC performed worse than the one we present here, we did not include it in the presentation of the results.

The MCAR model was fitted to each protein species as follows: the mean abundance was calculated for each experimental condition separately ignoring missing values. For the standard deviation calculation both conditions were pooled again ignoring missing values. The fixed censoring model was fitted to each condition and protein species independently by maximizing the logarithm of the likelihood as given by Taylor *et al.* (2013) as the AFT model. The censoring threshold of each protein species was set to the minimum of the data minus 10^{-6} pooled over the pair of concentration conditions. The maximum likelihood fit used the BFGS algorithm encapsulated in R's function "optim". A constrained optimization routine as suggested by Gelman & Hill (2007, section 18.5) was avoided by fitting the natural logarithm of the standard deviation. If one or both of the pair had only one observation the standard deviation was fitted from the data pooled over conditions. Note that this model is prone to overfitting as five parameters are fitted to data containing between 2 and 12 observed values. For the EBRCT no special treatment is necessary when only a single observation is present, as the informative prior on the standard deviation of protein abundance combined with the Gibbs sampler avoids numerical instabilities. The EBRCT model reduces to fixed censoring when (1) the std of the random censoring threshold is zero and (2) all priors are uninformative, except the prior for the std of random censoring threshold. As a control, we ran the EBRCT model with $\sigma_c = 0.01$, $f_{\sigma c} = 100$ and $f_{\mu c} = f_{\mu x} = f_{\sigma x} = 0.01$ and found its performance to be slightly better than fixed censoring, due to the small stabilizing effect from the non-zero prior weights. We could not set the prior weights exactly to zero as that led to numerical problems.

The IKNN model was fitted by (1) imputing the concatenated data frame (12 columns by ~ 1,200 rows) with R package impute (Hastie *et al.,* 2011) version 1.36.0 with parameters $k = 25$ and rowmax $= 0.5$ and (2) calculating means and standard deviations for each condition separately. The SVTI model was fitted similarly, except we used R package imputation (Wong, 2013) version 2.0.1 with parameters lambda $= 1000$, threshold $= 10^{-4}$ and max.iters $= 1000$.

To compare models we calculated the Receiver Operating Characteristic (ROC) using the UPS1 proteins as positives and the yeast proteins as negatives. We quantified performance by the partial Area Under the Curve (pAUC), taking the area under the ROC curve between false positive rates of 0.0 and 0.1. We choose not to use the (full) AUC itself as most models score in the 0.98 to 0.99 range, which is due to the relatively large changes in UPS1 protein abundance between concentration conditions. Using the pAUC focusses attention on the practically relevant regime of small false positive rates (between 0.0 and 0.1) while at the same time not restricting attention to a single value of the false positive rate. Note that a random classifier would give a pAUC of 0.05 and a perfect classifier a score of 0.1. We tested for statistical significance with the bootstrap test for pAUCs as implemented in package pROC (Robin *et al.,* 2011), version 1.5.4. We used a one-sided test and 5000 bootstrap samples.

We used two metrics to order the proteins, the absolute difference of means of protein abundance and the effect size defined as the absolute difference of means divided by the square root of the mean of the variances (Kruschke, 2013). The first measure is akin to the commonly used "fold change" metric as we work with log transformed abundance values. The second measure takes the variability of protein abundance into account and is proportional to the t-statistic for two independent samples. In both differential abundance metrics we used absolute values of the difference, making no use of the information that the abundance of UPS1 proteins is higher in the second member of a paired comparison. We deemed this to be more natural as in typical discovery-based proteomics experiments information on the sign of the abundance difference is not available.

All models are implemented in R (R Core Team, 2013), the code is open-source and available through supporting information. The code was tested with R version 3.0.1 and rstudio 0.98. We provide the summed intensity and LFQ values as an RData file so that readers can recreate our analysis.

Results and Discussion

Comparing Models for Differential Abundance Analysis

To recap, we compare five models for protein abundance on their ability to separate true positives (that are differentially abundant) from false positives (that have the same abundance) between four pairs of concentration conditions. As the models predict both the mean abundance and the standard deviation of abundance of each protein species we compare two metrics for differential abundance analysis: the difference of means and the effect size, defined as the difference of means divided by the root-mean-square of the standard deviations. We use benchmark data from CPTAC study 6 where a set of 48 proteins was spiked in a background proteome from yeast. This data set has relatively large concentration steps of a factor of 3 between consecutive conditions hence we focus on the part of the Receiver Operating Characteristic (ROC) for small false positive rates (FPR < 0.1).

We present the model comparison in three parts. First, an analysis of the performance averaged over all four concentration condition pairs for all five models and both difference metrics. Second, a more detailed comparison of our model with the best three competitors. Third, a discussion of the full ROC curves of all five models with their best difference metrics for both LFQ values and summed MS1 intensities. In Figure 2.4 we compare the mean partial Areas Under the curve (pAUC) averaged over all four concentration pairs for each of five models and two difference metrics. We observe that our EBRCT model with the effect size difference metric performs best followed by the EBRCT, SVTI, IKNN and MCAR models with the mean difference metric. These model-difference metric combinations perform clearly better than the effect size metric of the SVTI, IKNN and MCAR and the FCEN model with both metrics. We discuss the EBRCT effect size model and the SVTI, IKNN and MCAR models with the mean difference metric in light of Figure 2.5 below.

Turning to the worse performing models in Figure 2.4 we note that the effect size metric variants of the SVTI, IKNN and MCAR perform worse than their mean difference metric counterparts. Hence we conclude that the variance estimates of these models degrade their performance. In contrast for the EBRCT and FCEN models, the variance estimates are apparently accurate as in both cases the effect size metric model variants perform slightly better than their mean difference metrics. Lastly, we note that the fixed censoring model performs worst. Defining the mean error as the difference in mean pAUC between perfect performance (0.1) and the observed value, the mean error is 2.52% for the fixed censoring effect size model and 0.51% for the EBRCT effect size model a relative reduction in error of 80%. The stabilizing effect of the prior on the variance estimates in our Bayesian model explains this difference. To explain the performance of our EBRCT model relative to the fixed censoring model in more detail, we ran the analysis with an intermediate model, where we set the std of the censoring threshold to a small value, $\sigma_c = 0.01$, its prior weight to a large value $f_{\sigma c} = 100$ and the prior weight of the mean of the censoring threshold to a small value $f_{\mu c} = 0.01$. This model performed a little worse than the IKNN mean model and a little better than the MCAR mean model. The performance of this intermediate model shows that both the variance stabilisation from the random censoring and from the protein abundance contribute to performance. As the fixed censoring model performs so poorly we did not include

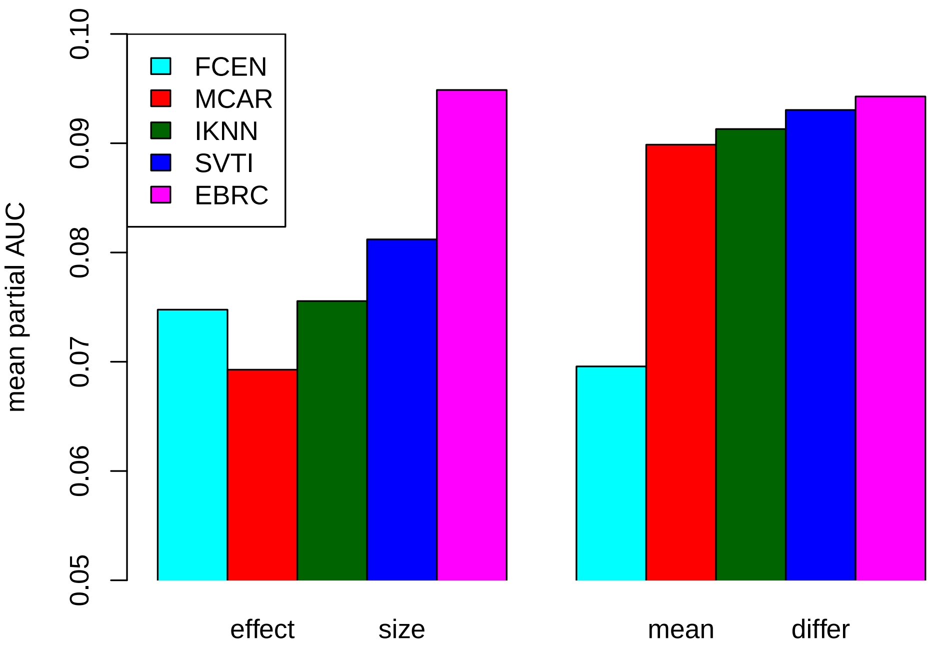

Figure 2.4: Mean partial Areas Under the Curve (pAUC) averaged over all four concentration pairs for each of five models and two difference metrics. The models are color coded and the effect size metric results are plotted left and the mean difference results are plotted on the right.

it in the significance tests that we discuss next. Since these tests are based on bootstrap sampling, we would have to run a large number of bootstrap samples to get a reliable estimate of the p-value.

In Figure 2.5 we show the significance level of a one-sided bootstrap test of the EBRCT model with the effect size metric vs the MCAR, IKNN and SVTI models with the mean difference metric. These are the best performing models in Figure 2.4 and here we analyze performance for each concentration condition pair separately. As can be seen, the EBRCT model significantly bests all competitors in the critical A vs B condition pair. In discussing Figure 2.1 we argued the A vs B condition to be most typical for a proteomics experiment as we expect differentially abundant proteins to have the same mean abundance as background proteins that are not differentially abundant. Further, the EBRCT model is significantly better than the MCAR model for the C vs D and the D vs E pairs and none of the other p values are significant. Thus, while the EBRCT model reduces the mean error relative to the SVTI, IKNN and MCAR models by 25%, 41% and 49% resp, this difference is not always significant.

In Figure 2.6 we show the ROC curves for all four concentration condition pairs and the best performing difference metric for each of the five models. Taking the MCAR model as a reference, as it is computationally the easiest, we observe that the EBRCT model consistently performs better than MCAR over all concentration

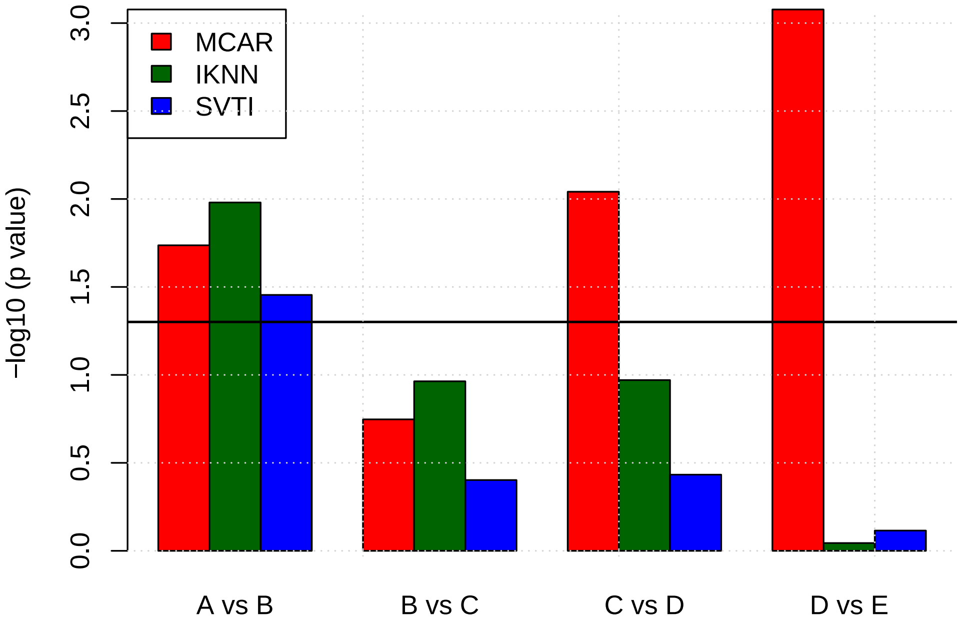

Figure 2.5: $-\log_{10}(pvalue)$ of a one-sided bootstrap test of the EBRCT effect size model vs MCAR, IKNN and SVTI models with the mean difference metric split out according to concentration condition pair. Horizontal line denotes a significance level of $p = 0.05$

condition pairs and for all false positive rates. Imputation with the relatively recent Singular Value Thresholding (SVTI) model results in better mean pAUC performance than MCAR (Figure 2.4) but this advantage comes mainly from the D vs E condition pair, where MCAR performs bad and SVTI good. As we argued in the discussion of Figure 2.1 the D vs E comparison is unusual in that the high concentration of spiked-in UPS1 protein leads to ion suppression of the background protein peak intensity, a situation that is probably not representative of a well-designed discovery proteomics experiment. Moreover, the SVTI model performs worse than the MCAR model in the critical A vs B comparison. What holds true for imputation via SVTI is also true for imputation via k-nearest neighbors, the IKNN model. This model performs a bit worse than SVTI but its performance is more variable over the concentration condition pairs: it performs best in the D vs E comparison and worst in the critical A vs B comparison.

Results in Figure 2.6 are based on abundance quantification via MaxQuant's LFQ algorithm whereas the results in Figure S2 in the supporting information are based on the summed MS1 intensity. Comparing the LFQ-based results in Figure 2.6 with the intensity based results in Figure S2 it is clear that LFQ normalization improves performance: mean pAUC averaged over all models and concentration condition pairs is 0.089 for LFQ and 0.078 for intensity based abundance quantification. However, whereas average performance is worse for intensity-based abundance

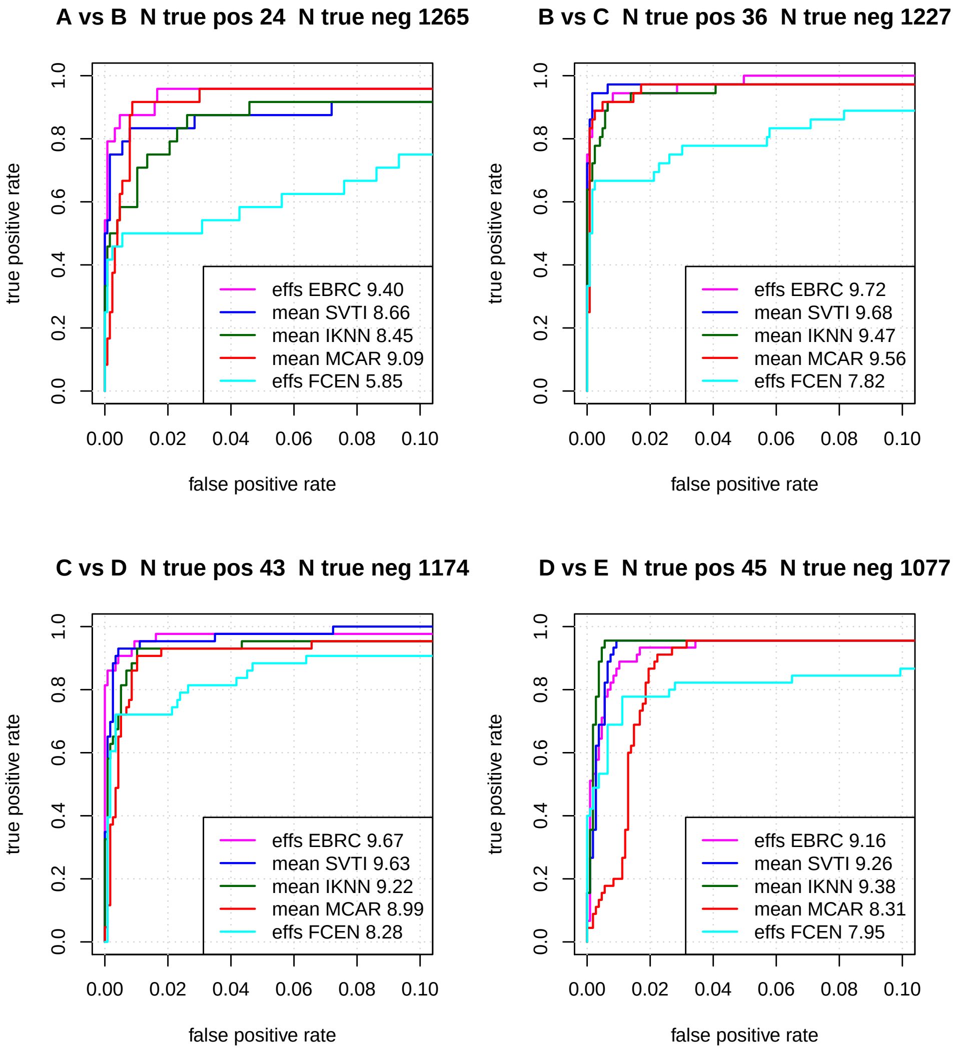

Figure 2.6: Receiver Operating Characteristics (ROCs) of all four paired comparisons based on label-free quantification (LFQ) data. Plot titles capture the paired experiment conditions and the number of true positives (UPS1 proteins) and negatives (yeast proteins) in the data matrices. Partial Areas Under the Curve (pAUCs) of each model are indicated in the legend, pAUC values are multiplied by 100. The best performing difference metric of each of the five models is plotted.

quantification, it performs better in a few cases. First, for the D vs E comparison, performance for intensity-based abundance quantification is better for all models except MCAR. This underscores once more the peculiar condition created by spiking in large concentrations of UPS1 protein. Nevertheless, it is is disappointing that the LFQ algorithm reduces performance relative to simple MS1 intensity summation. By comparison, our EBRC model always improves performance relative to simple MCAR in all concentration condition pairs. Second, performance of the fixed censoring model improves under intensity-based abundance quantification. Apparently, the assumptions of a censoring model are better fulfilled in the intensity data. Some support for the notion that the LFQ algorithm makes estimation of the censoring threshold more difficult comes from the variability of the censoring threshold estimates of the EBRCT model, denoted $\mu_{c,j}$ in eq. 2.6. The standard deviation over protein species and averaged over concentration condition pairs is 0.165 in LFQ-based abundance estimation and 0.125 in intensity-based abundance estimation. Thus, although we follow the advice of Karpievitch, Dabney, *et al.* (2012) in first normalizing (via LFQ in our case) and then dealing with missing data (via EBRCT in our case), our finding suggests that an integrated approach might be better. Lastly, we note that the MCAR model seems to perform particularly bad with intensity-based abundance quantification.

Conclusions

We introduced the Empirical Bayes Random Censoring Threshold (EBRCT) model to detect differentially abundant protein species in discovery-based label-free proteomics experiments by modeling missing data in replicates. We compared it to four alternative models, a model ignoring missing values (MCAR), a model with a fixed censoring threshold for each protein species (FCEN) and two imputation models, k-nearest neighbors (IKNN) and Singular value Thresholding (SVTI). We used the CPTAC study 6 data as benchmark where the differentially abundant proteins are known since the proteins from the UPS1 protein standard were spiked in a background proteome from yeast. We used two metrics to quantify the difference in abundance between a pair of concentration conditions, the mean difference and the effect size. The mean difference metric boils down to a fold change metric since we only consider logarithmically transformed values. The effect size is the mean difference divided by an estimate of the variability, similar to the t-statistic.

Our EBRCT model has a combination of two properties that make it the best performer: (1) censoring as an explanation for missingness and (2) regularization of estimates of single protein species by using of all the data. These properties by themselves are not unique as censoring is also in the fixed censoring (FCEN) model and the imputation models (IKNN and SVTI) also make use of all the data in estimating abundances. We discuss these properties in more detail. First, as just stated, the EBRCT assumes censoring to underly missingness. However, the *random* censoring threshold of the EBRTC model allows to estimate of the censoring threshold with a Bayesian approach. This can be contrasted with alternative Bayesian approaches (Kang & Xu, 2013) where the censoring threshold is assumed known (specified by the manufacturer of the microarray in (Kang & Xu, 2013)). However, the model by Kang & Xu (2013) has two properties that we could incorporate in future versions of our model: (1) it is a hierarchical Bayesian model avoiding the data reuse of our empirical Bayes approach and (2) it has an explicit parametrization of the three groups of proteins species: those that are less abundant than the reference, those that are more abundant and those that are not differentially abundant. Second, the EBRCT model effectively makes use of all the data when estimating the parameters of each protein species as the prior means in eqs. 2.4 to 2.7 are obtained from pooled data. It shares this property with the imputation methods. However, the imputation methods impute missing values with fixed value without an obvious way to inform the downstream statistical analysis that this value was imputed as opposed to observed. In contrast, our EBRCT model does not impute a fixed value for missing values, it estimates a probability density for the missing values thus keeping track of observed and missing data in statistical analysis. Where this pays off is in the effect size metric for quantifying differential abundance: where the EBRCT model performs slightly better with the effect size metric as compared with the mean difference metric, performance of the imputation methods combined with the effect size metric is clearly inferior to the mean difference metric.

Our EBRCT model has a clear performance advantage over a model that ignores missing data (MCAR), significantly besting it in 3 of the 4 comparisons (and non-significantly in the fourth). While the MCAR model is attractive due to its simplicity, its estimates of abundance can be improved both by imputation of missing values and by the EBRCT model. However, we warn against indiscriminate use of the EBRCT model and the imputation methods (IKNN, SVTI) for other purposes than differential abundance analysis, a warning that is generally true for all ways of dealing with missing data (Little & Rubin, 2002). While the MCAR model is reasonably accurate with the mean difference metric, it performs worst among all models considered with the effect size metric. In this light, it is curious to find that Clough *et al.* (2012) find the MCAR model with effect size metric (called "assuming no feature interferences") to perform reasonably well. We can offer a few hypotheses for this difference. First, their data sets do not contain a ground truth hence they only compare models in terms of number of identified protein species. Second, the alternative models they consider, imputation with background intensity and removing protein species from data set are rather coarse models.

While we have shown the EBRCT model to be superior in a statistical sense in separating true from false positives, we refrained from interpreting the randomness of the censoring threshold. We consider the probabilistic censoring threshold as arising from multiple independent stochastic contributions, ranging from variation in sample preparation to mass spectrometry conditions. In our model, this randomness is captured by parameter σ_c the standard deviation of the random censoring threshold which we found to be 0.22 on average (mean over all condition pairs, range from 0.20 to 0.24). The recent book by Eidhammer *et al.* (2013) provides an overview of the many causes of missingness. We are invoking the central limit theorem in support of our assumption of a normal distribution for the censoring threshold. The assumption of normality for the protein abundances is based on the same argument of multiple independent contributions. Other distributional assumptions like a t-distribution can be made and could lead to improved performance at the expense of slower computation.

While the CPTAC study 6 data constitute a good testbed for comparing models of differential protein abundance, a better data set for this purpose would have a larger number of spiked-in proteins (to balance true- and false-positives) with both increasing and decreasing concentrations in a range of concentration ratios on a complex background. Natural extensions to our model for future work include an analysis at the peptide level which combines censoring, normalization and protein roll-up and a comparison of more than two conditions. This latter extension would be similar to an F-test generalisation of the t-test. In its current incarnation our model is complementary to existing protein quantification pipelines (like MaxQuant (Cox & Mann, 2008), LFQuant (Zhang *et al.,* 2012) or OpenMS (Weisser *et al.,* 2013)) and is available as an easy-to-use R script.

Acknowledgements

This research was supported in part by the Netherlands Organization for Scientific Research (NWO) Complexity project 645.000.003 and by an NWO Computational Life Sciences grant (RUMPHI) awarded to TMHD. This paper was significantly improved by suggestions from an anonymous referee.

Additional Information

The authors have declared no conflict of interest.

References

1. Aittokallio, T. Dealing with Missing Values in Large-Scale Studies: Microarray Data Imputation and Beyond. *Briefings in Bioinformatics* **11,** 253–264 (2010).
2. Candes, E. J. & Plan, Y. Matrix Completion with Noise. *Proceedings of the IEEE* **98,** 925–936 (2010).
3. Choi, H. *et al.* SAINT-MS1: Protein-Protein Interaction Scoring Using Label-Free Intensity Data in Affinity Purification-Mass Spectrometry Experiments. *J. Proteome Res.* **11,** 2619–2624 (Apr. 2012).
4. Clough, T. *et al.* Statistical Protein Quantification and Significance Analysis in Label-Free LC-MS Experiments with Complex Designs. *BMC Bioinformatics* **13 Suppl 16,** S6 (2012).
5. Cox, J., Hein, M., *et al.* MaxLFQ Allows Accurate Proteome-Wide Label-Free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction. *Mol. Cell Proteomics* **M113.031591,** 1–37 (July 2014).
6. Cox, J. & Mann, M. MaxQuant Enables High Peptide Identification Rates, Individualized p.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification. *Nature Biotechnology* **26,** 1367–1372 (Dec. 2008).
7. Eidhammer, I. *et al. Computational and Statistical Methods for Protein Quantification by Mass Spectrometry* (Wiley, 2013).
8. Gelman, A., Carlin, J. B., *et al. Bayesian Data Analysis* (Chapman & Hall/ CRC, 2004).

9. Gelman, A. & Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Cambridge University Press, 2007).
10. Hastie, T. *et al. impute: Imputation for microarray data* 2011.
11. Hoff, P. D. *A First Course in Bayesian Statistical Methods* (Springer, 2009).
12. Ji, H. & Liu, X. S. Analyzing Omics Data Using Hierarchical Models. *Nature Biotechnology* **28,** 337–340 (2010).
13. Kang, J. & Xu, E. Y. An Integrated Hierarchical Bayesian Approach to Normalizing Left-Censored microRNA Microarray Data. *BMC Genomics* **14,** 507 (2013).
14. Karpievitch, Y., Stanley, J., *et al.* A Statistical Framework for Protein Quantitation in Bottom-up MS-Based Proteomics. *Bioinformatics* **25,** 2028–2034 (Aug. 2009).
15. Karpievitch, Y., Dabney, A. & Smith, R. Normalization and Missing Value Imputation for Label-Free LC-MS Analysis. *BMC bioinformatics* **13,** S5 (2012).
16. Kruschke, J. K. Bayesian Estimation Supersedes the t Test. *J Exp Psychol Gen* **142,** 573–603 (May 2013).
17. Little, R. J. A. & Rubin, D. B. *Statistical Analysis with Missing Data* (Chichester: Wiley, 2002).
18. Lunn, D. *et al. The BUGS Book: A Practical Introduction to Bayesian Analysis* (CRC Press, 2012).
19. Michalski, A., Cox, J. & Mann, M. More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority Is Inaccessible to Data-Dependent LC-MS/MS. *J. Proteome Res.* **10,** 1785–1793 (Apr. 2011).
20. Milac, T. I., Randolph, T. W. & Wang, P. Analyzing LC-MS/MS Data by Spectral Count and Ion Abundance: Two Case Studies. *Statistics and its Interface* **5,** 75–87 (2012).
21. Neilson, K. A. *et al.* Less Label, More Free: Approaches in Label-Free Quantitative Mass Spectrometry. *Proteomics* **11,** 535–553 (Feb. 2011).
22. Ong, S. E. *et al.* Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Mol. Cell Proteomics* **1,** 376–386 (May 2002).
23. Paulovich, A. G. *et al.* Interlaboratory Study Characterizing a Yeast Performance Standard for Benchmarking LC-MS Platform Performance. *Mol. Cell Proteomics* **9,** 242–254 (Feb. 2010).
24. Piehowski, P. D. *et al.* Sources of Technical Variability in Quantitative LC-MS Proteomics: Human Brain Tissue Sample Analysis. *J. Proteome Res.* **12,** 2128–2137 (May 2013).
25. Plummer, M. *et al. JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling* in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* **124** (2003), 1–10.

26. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2013). `http://www.R-project.org/`.

27. Robin, X. *et al.* pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves. *BMC Bioinformatics* **12,** 77 (2011).

28. Ross, P. L. *et al.* Multiplexed Protein Quantitation in Saccharomyces Cerevisiae Using Amine-Reactive Isobaric Tagging Reagents. *Mol. Cell Proteomics* **3,** 1154–1169 (Dec. 2004).

29. Soneson, C. & Delorenzi, M. A Comparison of Methods for Differential Expression Analysis of RNA-Seq Data. *BMC bioinformatics* **14,** 91 (2013).

30. Taylor, S. L., Leiserowitz, G. S. & Kim, K. Accounting for Undetected Compounds in Statistical Analyses of Mass Spectrometry Omic Studies. *Statistical applications in genetics and molecular biology* **12,** 703–722 (2013).

31. Troyanskaya, O. *et al.* Missing Value Estimation Methods for DNA Microarrays. *Bioinformatics* **17,** 520–525 (2001).

32. Weisser, H. *et al.* An Automated Pipeline for High-Throughput Label-Free Quantitative Proteomics. *J. Proteome Res.* (Feb. 2013).

33. Wong, J. *Imputation: Imputation* 2013.

34. Zhang, W. *et al.* LFQuant: A Label-Free Fast Quantitative Analysis Tool for High-Resolution LC-MS/MS Proteomics Data. *Proteomics* **12,** 3475–3484 (2012).

3

Comparative analyses of data independent acquisition mass spectrometric approaches: DIA, WiSIM-DIA and untargeted DIA Data Independent Analysis, Spectral library, Quantitative proteomics

Frank Koopmans[1], Jenny T. C. Ho[2], August B. Smit[1], Ka Wan Li[1]

[1] Department of Molecular and Cellular Neurobiology, Center for Neurogenomics Cognitive Research, VU University, Amsterdam, The Netherlands
[2] Thermo Fisher Scientific, Hemel Hempstead, UK

Data independent acquisition (DIA) is an emerging technology for quantitative proteomics. Current DIA focusses on the identification and quantitation of fragment ions that are generated from multiple peptides contained in the same selection-window of several to tens of m/z. An alternative approach is WiSIM-DIA, which combines conventional DIA with wide-SIM (wide-selected ion monitoring) windows to partition the precursor m/z space to produce high quality precursor ion chromatograms. However WiSIM-DIA has been under-explored; it remains unclear if it is a viable alternative to DIA. We demonstrate that WiSIM-DIA quantified more than 24000 unique peptides over five orders of magnitude in a single 2 hrs analysis of a neuronal synapse-enriched fraction, compared to 31000 in DIA. There is a strong correlation between abundance values of peptides quantified in both the DIA and WiSIM-DIA datasets. Interestingly, the signal to noise ratio of these peptides was not correlated. We further show that peptide identification directly from DIA spectra identified >2000 proteins, which included unique peptides not found in spectral libraries generated by DDA.

This chapter has been published in *Proteomics 18, no. 1 (2018)*

Main text

LC-MS/MS based quantitative proteomics is the method of choice to measure changes in global protein levels in biological samples. In the past decade, data-dependent acquisition (DDA) has been widely used for this. In DDA, the precursors, usually the top 10-20 peptides per cycle, are sequentially selected from a full mass MS1 scan for fragmentation and acquisition in the MS/MS mode (Hondius *et al.,* 2016; Pandya *et al.,* 2016). Recently, DDA has been optimized to reveal the comprehensive proteome of a single cell type (Bekker-Jensen *et al.,* 2017). However, the stochastic precursor selection of DDA leads to inconsistent detection of peptides. In particular, the under-sampling of medium to low abundant peptides causes high variation across replicates due to the selection of different subsets of peptides. This results in missing peptide identification, which can be substantial among replicates (>30%) and reduces the number of quantifiable proteins (Liu *et al.,* 2004; Michalski *et al.,* 2011).

Data-independent acquisition (DIA, also known as SWATH (Gillet *et al.,* 2012)) is a recent development in quantitative proteomics. It is mainly performed on the high-resolution high mass accuracy mass spectrometers and has been shown to be superior to DDA (Bruderer, Bernhardt, Gandhi, Miladinovic, *et al.,* 2015) by producing a higher number of quantified proteins in shorter analysis time, fewer missing values and lower Coefficients of Variation (CoV) across replicates. In DIA, all peptides within a predefined wide selection-window, which in the original DIA study spanned a 25 m/z range (Gillet *et al.,* 2012), are simultaneously fragmented. The acquisition is repeated sequentially in stepped selection-windows, usually in the 400-1000 m/z range. Generally, the high number of fragments ions generated from multiple peptides contained in the same selection-window complicates the analysis in a classical database search strategy. This problem is circumvented by the use of a reference spectral library, which is generated beforehand by an extensive analysis of the same/similar samples by DDA. The information of the elution time of the peptide and its fragment ions stored in the spectral library defines the identity of the peptide measured in a DIA experiment (Bruderer, Bernhardt, Gandhi & Reiter, 2016; Keller *et al.,* 2016; Tsou *et al.,* 2015; Cox & Mann, 2008). Thus, samples not present in a spectral library in principle cannot be analysed. To circumvent this shortcoming, algorithms have been developed that create a pseudo-DDA dataset from the DIA data (untargeted peptide identification or untargeted DIA (Y. Li *et al.,* 2015; Wang *et al.,* 2015)) for subsequent search in way similar to the classical DDA strategy.

An alternative to DIA is a wide selected-ion monitoring, data-independent acquisition (WiSIM-DIA), which is grossly under-explored. While both DIA and WiSIM-DIA require a spectral library for peptide/protein identification, in contrast to MS2-based DIA method, WiSIM-DIA uses MS1 for quantitation. Previous reports on WiSIM-DIA were performed in an Orbitrap Fusion mass spectrometer (Kiyonami *et al.,* 2014; Bruderer, Bernhardt, Gandhi, Miladinovic, *et al.,* 2015). This method consists of 3-stepped selected-ion monitoring (SIM) scans acquired with 240,000 resolution over a 200 m/z range that covers 400-1000 m/z. In parallel with each SIM scan, peptide fragmentation from selection-windows of 12 m/z were acquired in the ion trap, with acquisition repeated with 17 sequential ion trap MS/MS windows. In comparison, DIA used the Orbitrap for high (60,000) resolution MS1 and 17 sequential MS/MS windows in lower (15,000) resolution. So the quality of MS1 acquisition in WiSIM-DIA was improved compared to DIA by using stepped SIM scans and a higher resolution, while the quality of MS/MS acquisition was favorable for DIA due to the use of the Orbitrap (compared to WiSIM-DIA using Ion Trap for MS/MS). MS/MS data acquired in the low-resolution ion trap were used for identification whereas quantitation was based on the extracted ion chromatogram of the SIM data with a 5 ppm window.

Here, the spectral library could be generated with classical DDA where the MS1 full scan is acquired in the high-resolution Orbitrap, and the fragment ions in the fast but low-resolution ion trap. It is proposed that WiSIM-DIA does not suffer from the drawback of DIA, for example the potential interferences of the large number of fragment ions derived from co-eluting peptides. However, the only application published recently reported the quantitation of about 1100 proteins by WiSIM-DIA(Bruderer, Bernhardt, Gandhi, Miladinovic, *et al.,* 2015), which seems to be on the lower side acquired by a modern MS. Thus, it has remained unclear whether WiSIM-DIA is a viable alternative to DIA.

In this study we used an Orbitrap Fusion Lumos in DDA mode to generate two spectral libraries from the mouse synaptosome, a preparation enriched for proteins of the neuronal synapse (Gonzalez-Lozano *et al.,* 2016), that constitutes the building block of the brain. The tryptic digest of 10 μg synaptosome proteins were fractionated offline using high pH reversed phase cartridges into 8 fractions. Each fraction was subjected to DDA by two separate acquisition strategies: 1) MS1 OT with the fast but low resolution IT for MS/MS (HCD-IT) and , 2) MS1 OT with the high resolution OT MS/MS (HCD-OT). The data was processed using MaxQuant (Cox & Mann, 2008) with 1% FDR at both peptide and protein level.

From the same sample we used 1 μg for DIA with a 2 h LC gradient. Three replicates each for DIA and WiSIM-DIA, were performed. Technically, several parameters can be considered to maximize the DIA output. While a cycle scan time is usually fixed around 3-4 seconds to obtain 6-10 measurement points of a peptide that is needed for quantitation, the width of a selection-window, the accumulation time per selection-window, and the whole m/z range can be varied. The original study opted for a 25 m/z selection-window (Gillet *et al.,* 2012), which may cause peptide fragment ion interferences due to their high complexity. In another extreme, a narrow selection-window of 3 m/z has been proposed as preference for more comprehensive and in depth view of protein profiling in a complex sample (Kang *et al.,* 2017). This is compromised by a shorter acquisition time with potentially reduced sensitivity. Considering the mild protein complexity of the synaptosome fraction of about 5000 proteins contained in the spectral library, we chose the 12 m/z selection-window for both DIA and WiSIM-DIA (see also (S. Li *et al.,* 2017)). The total mass range covered was 400-800 m/z that includes the majority of the peptides (Supplementary Figure S1).

In addition to the classical DDA-based spectral library we generated a spectral library from the DIA data using the recently launched Spectronaut Pulsar software (untargeted DIA at 1% peptide and protein FDR, settings analogous to the MaxQuant DDA analysis), which yielded 17894 unique peptide sequences in 2079 protein groups. This is less than the 27897 and 33673 unique peptide sequences and 4770 and 4989 protein groups represented in the IT and OT spectral libraries, respectively, within the 400-800 m/z range (Figure 3.1A-B, the total number of identified peptides without any m/z filters in each spectral library is shown in Supplementary Figures S1 and S2). Here, we compared the subset of peptides in the 400-800 m/z range to match the DDA spectral library with the acquisition settings for DIA and WiSIM-DIA on the Orbitrap Fusion Lumos.

The samples used for (untargeted) DIA were not fractionated, in contrast to the extensively fractionated samples used exclusively for DDA spectral library construction, which may account for the overall reduced number of identifications by untargeted DIA. Despite the lack of extensive fractionation, untargeted DIA contributed 2901 unique peptides to the spectral library. Interestingly, most of the peptides exclusively identified by untargeted DIA belong to protein groups that were also identified in the IT or OT libraries (or both). This suggests that the untargeted DIA unique peptides may have been lost in the first dimensional high pH reversed phase HPLC separation used for the OT and IT analyses. Alternatively, respective peptide MS/MS spectra quality could be subpar, be part of mixed chimeric spectra in shotgun MS/MS or it may due to the nature of stochastic precursor selection of DDA; while these peptides were present they might not have been selected for MS/MS or the MS/MS could have been triggered far from the peak apex (pseudo MS/MS in untargeted DIA is generated at the apex of the peak). Either way, this argues that DIA data that are generated from routine quantitative analysis might subsequently be added to spectral libraries generated by conventional DDA to increase protein coverage.

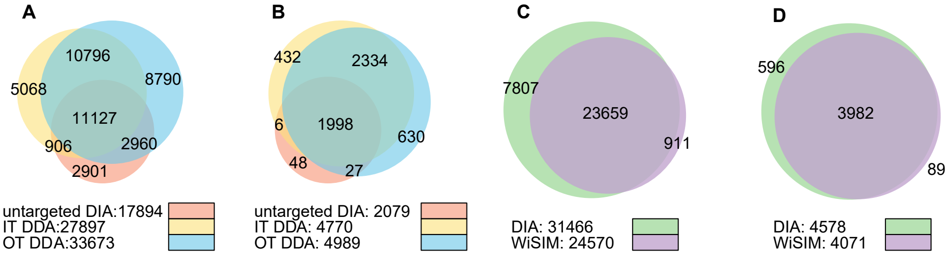

Figure 3.1: For the spectral library Ion Trap (IT) DDA, Orbi Trap (OT) DDA and untargeted DIA acquisition resulted in A) 27897, 33673 and 17894 unique peptide sequences within the 400-800 m/z range and B) 4770, 4989 and 2079 protein groups, respectively. From the merged spectral library, which contains all peptides identified by either IT, OT or untargeted DIA, C) 31466 and 24570 unique peptide sequences and D) 4578 and 4071 protein groups were quantified by DIA and WiSIM-DIA, respectively.

The performance of both DIA and WiSIM-DIA was excellent; 31466 and 24570 unique peptide sequences contained in the merged spectral library were quantified at 1% peptide-level FDR, respectively, with extensive overlap (Figure 3.1 C-D). These peptides map to > 4000 protein groups in the merged spectral library (no protein-level FDR was applied).

From the two spectral libraries generated using MS2 HCD-OT or MS2 CID-IT we observed a slightly higher number of peptides and protein groups identified by OT (Figure 3.1), and a slightly higher coverage of the MS2 HCD-OT library following DIA quantification (Figure 3.2). Thus, despite the slower scan rate of Orbitrap its high resolution and mass accuracy favourably affects the population of identified peptides that can be recovered in the DIA data analysis. It may also underlie the fact that both employed the same MS sector, the Orbitrap, for the measurement. The coverage of protein groups from the untargeted DIA spectral library is 100% (Figure 3.2), which reflects the fact that it is generated from the original DIA data. However, using the WiSIM-DIA data we also quantified 90% and 99% of all peptides and protein groups, respectively, from the untargeted DIA spectral library, which suggests the untargeted DIA approach tends to prioritize peptides that exhibit a clean elution profile with high signal to noise ratio.

The spectral library coverage by WiSIM-DIA is generally lower than that of DIA, which may underlie at least in part that current DIA algorithms are primarily using MS2 fragment intensities to identify spectral library peptides. Algorithm improvements that lead to better utilization of high quality MS1 signals with sub-ppm mass error would improve the recovery rate of spectral library peptides by WiSIM-DIA. This approach would be an extension of the previous described "accurate mass tag" strategy, in which the identities of the peptides based on the LC-FTICR MS1 measurement were validated by LC-MS/MS analysis on a conventional ion trap mass spectrometer(Conrads *et al.,* 2000; Gonzalez-Lozano *et al.,* 2016). A similar approach has been applied to the analysis of phosphorylated human peptides (Mao *et al.,* 2011), and HeLa cell proteome(Ivanov *et al.,* 2017).

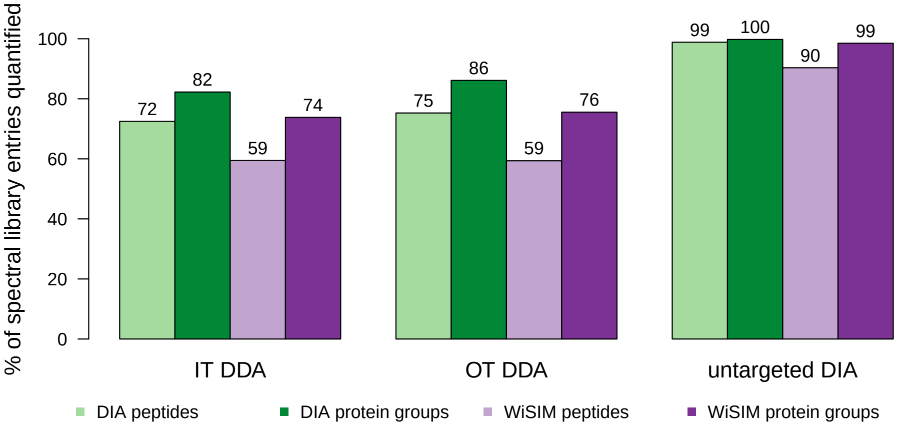

Figure 3.2: Fraction of peptide sequences and protein groups from individual spectral libraries quantified by DIA and WiSIM-DIA. DIA quantifies on average 13% more peptides from each spectral library compared to WiSIM-DIA. Although the number of peptides contributed to the merged spectral library is relatively low for Pulsar (Figure 1A,B), their recovery after quantification is remarkably high (note; most likely the peptides with best MS2 abundance profiles are selected for identification). Analogous figures with count data instead of fractions in Supplementary Figure S3.

The median Coefficient of Variation (CoV) for 25788 peptides quantified in both the DIA and WiSIM-DIA datasets was 9% within three WiSIM-DIA technical replicates and 7% within three DIA technical replicates (Supplementary Figure S4). Evaluating all peptides quantified by DIA (35123) and WiSIM-DIA (26899) resulted in 8% and 9% median CoV, respectively. For both comparisons, student's t-tests reveal the differences between DIA and WiSIM-DIA CoV was statistically significant (p-value < 10-16) albeit with much higher effect size for the former (Cohen's d: 0.29) compared to the latter (Cohen's d: 0.09). Correlation of the abundance values between technical replicates yielded a 0.94 R2 and 0.93 R2 on average for DIA and WiSIM-DIA, respectively. The slightly reduced technical variation of DIA over WiSIM-DIA will likely result in higher sensitivity when performing differential abundance analysis in real-world biological applications.

The signal to noise ratio is a good indicator of the mass spectrometric measurement quality. In DIA mode the signal to noise ratio for fragment ion intensities per precursor is better than those of the corresponding precursor measured in MS1 (Figure 3.3A). On the other hand, WiSIM-DIA yields a better signal to noise ratio for the precursor ion measured in MS1 compared to its MS2 signal to noise ratio. These findings are in accordance to the experimental design that DIA is optimized for MS/MS analysis, and WiSIM-DIA for MS1 measurement. Student's t-tests applied to the log-transformed WiSIM-DIA MS1 and MS2 signal to noise distributions confirmed statistical significance of this comparison (p-value < 10-16) with a medium-large effect size (Cohen's d: 0.67). Analogously, the overall DIA MS2 signal to noise distribution was significantly lower (p-value < 10-16) than its WiSIM-DIA MS1 counterpart but with a relatively small effect size (Cohen's d: 0.34).

In addition to comparing the distributions shown in Figure 3.3A, student's t-tests on log2 DIA MS2 and WiSIM-DIA MS1 signal to noise values for all individual peptides using the triplicate DIA and WiSIM-DIA measurements resulted in 4155 (out of 25780) significantly different peptides at FDR adjusted p-value ≤ 0.01. Of these, 1010 and 3145 peptides showed improved signal to noise in WiSIM-DIA MS1 and DIA MS2, respectively. We found a strong correlation (0.792 R2) between the abundance values of peptides quantified in both the DIA and WiSIM-DIA datasets, as expected (Figure 3.3B). Interestingly, the signal to noise ratio of these peptides was not correlated (0.171 R2) between DIA and WiSIM-DIA (as compared to 0.80 and 0.59 average R2 between DIA and WiSIM-DIA replicates, Supplementary Figure S5). The lack of correlation in signal to noise might be explained by the stepped SIM scans in WiSIM-DIA that clean up the spectra of many peptides, but might reduce the signal for already low abundant peptides (Figure 3.3C). The use of different modes of quantification for WiSIM-DIA (MS1) and DIA (MS2) taken together with their overall similar S/N distributions shown in Figure 3.3A could give rise to subpopulations of peptides that are quantified with higher signal quality in either WiSIM-DIA or DIA, indicating mutually exclusive benefits. Alternatively, observed differences in peptide subsets could arise by chance. Future research could further investigate this hypothesis using extensive datasets that allow for cross-validation of peptides with stark differences in S/N between WiSIM-DIA and DIA.

We conclude that DIA and WiSIM-DIA can quantify more than 31000 and 24000 unique peptides (at 1% peptide-level FDR), respectively, over 5 orders of magnitude in a single 2 hrs analysis with nearly no missing values (0.08% and 0.004% missing peptide values between three technical replicates, respectively). The number of peptides from the spectral libraries recovered by WiSIM-DIA will be improved when its high quality MS1 signal is better taken advantage of by future improvements of analysis software (eg. by relying on accurate retention time and low precursor mass error for matching precursor peaks to the library in absence of high quality fragment spectra(Conrads *et al.,* 2000; Gonzalez-Lozano *et al.,* 2016)). The untargeted DIA spectral library generated from the triplicate 2 h DIA analysis yields nearly 50% of the peptides/proteins contained in the spectral library generated from the 8x2 hrs analysis of the deep MS sequencing of the sample, as well as unique peptides. We anticipate that a narrow selection window of a few m/z (SWATH-ID of 3 m/z(Kang *et al.,* 2017)) analysed in a fast machine such as Q-Exactive HF-X with 40 Hz will generate a untargeted DIA library that might be of competitive quality with the classically generated spectral library, however, with much reduced analysis time, which is also a better match to the subsequent DIA analysis using similar LC-MS/MS parameters.

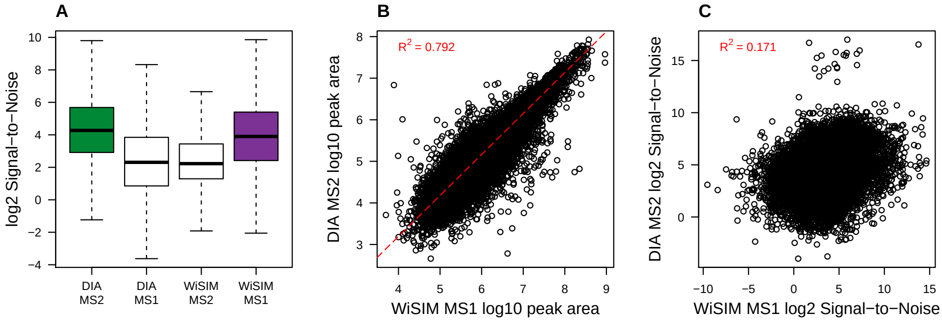

Figure 3.3: Quality of precursor quantification compared between DIA (MS2) and WiSIM-DIA (MS1). A) The signal to noise ratio (S/N) is high for both DIA MS2 and WiSIM-DIA MS1. Analogously, the secondary mode of quantification (MS1 for DIA, MS2 for WiSIM-DIA) is more noisy. B) The abundance values of 25778 precursors quantified by both DIA and WiSIM-DIA are correlated (0.792 R2), as expected. The dashed red line shows linear regression. C) Interestingly, the S/N for these peptides is not correlated (0.171 R2). A subpopulation of precursors is quantified with higher signal quality by DIA MS2 than WiSIM-DIA MS1, and vice versa.

Experimental procedures

Sample preparation for MS acquisition

All animal experiments were performed in accordance with relevant guidelines and regulations of the Vrije Universiteit. The animal ethics committee of the Vrije Universiteit approved the experiments. Hippocampal synaptosomes were prepared from three 3-month-old C57BL6 mice as previously described (Gonzalez-Lozano *et al.,* 2016). The synaptosome was solubilized in 2% SDS, and prepared for MS analysis using the FASP protocol (Pandya *et al.,* 2016) and 30K centicon filters from Millipore. Cysteine residues were derivatized by MMTS. Proteins were digested overnight at

37oC with Sequence Grade Trypsin/Lys-C from Promega. Peptides were speedvac dried, and re-dissolved in 0.1% Formic acid.

Mass Spectrometric acquisition

15ug of mouse synaptosome was fractionated into 8 fractions by high pH reversed phase cartridges following the protocol included in the kit (Pierce™ High pH Reversed-Phase Peptide Fractionation Kit). Peptide fractions were dried and then re-dissolved in 15uL of water containing 0.1% formic acid and spiked with 0.5uL of 10x HRM peptides (Biognosys). 5uL of each fraction was analysed by nanoLC-MS/MS using the Orbitrap Fusion Lumos (Thermo Scientific, San Jose). The spectral libraries were generated by DDA (Data Dependent Acquisition) using MS1 Orbitrap survey scan and either MS2 HCD with detection in the ion trap or MS2 HCD with detection in the Orbitrap. Peptides were separated by nanoLC (Ultimate 3000 RSLCnano, Thermo Scientific). Peptides were loaded on a µ-precolumn (300um ID x 5mm, C18 PepMap100, 5um, Thermo Scientific) at 15uL/min for 3min using 98/2 water/acetonitrile containing 0.05% TFA. After 3 mins the peptides were separated on an EasySpray column (75um ID x 50cm, C18 PepMap, 2um, Thermo Scientific) at 300nL/min using water/acetonitrile/formic acid gradient. The gradient consisted of initial step of 3-8% B over 5min followed by 8-28% over 90min, 28-80%B over 7min, held at 80%B for 4min and then equilibrated for 15min at 3% B, where mobile phase A consisted on water containing 0.1% formic acid and mobile phase B consisted of 80/20 acetonitrile/water containing 0.1% formic acid. Separation was performed at 40°C and the total acquisition time was 150min. The mass spectrometer was fitted with an EasySpray source (Thermo Scientific) and operated in DDA manner. Each DDA cycle consisted of one Orbitrap MS survey scan acquired at 120,000 resolution at m/z 200 and precursors ions meeting user defined criteria such as charge state, monoisotopic precursor selection, intensity and dynamic exclusion were selected for MS2 based on ‘most intense’. Precursor ions were isolated using the quadrupole (1.6Th isolation width) and activated by HCD in the ion routing multipole. In one experiment, fragment ions were detected in the ion trap in rapid scan and in another experiment fragment ions were detected in the Orbitrap at 15,000 resolution (at m/z 200).

1 µg of unfractionated peptides spiked with 1uL of HRM peptides (Biognosys) were analysed by DIA and WiSIM-DIA by nanoLC-MS/MS using the Orbitrap Fusion Lumos. NanoLC conditions and gradients for DIA and WiSIM-DIA were the same as DDA experiments. DIA on the Fusion Lumos consisted of a MS1 scan at 60,000 resolution at m/z 200 followed by sequential quadrupole isolation windows of 12m/z for HCD MS/MS with detection of fragment ions in the Orbitrap at 15,000 resolution at m/z 200. The m/z range covered was 400-800 and the AGC settings for MS/MS was 5e5 target value and 55ms maximum injection time. WiSIM-DIA on the Fusion Lumos consisted of four high resolution selected ion monitoring (SIM) scans (240,000 resolution at m/z 200) with wide isolation windows of 100m/z were used to cover all precursor ions of 400-800 m/z. In parallel each SIM scan, 15 sequential ion trap MS/MS with 7m/z isolation windows were acquired to cover the associated 100m/z SIM mass range. Quantitative information for all precursor ions detected

in four sequential SIM scans is recorded in a single run. All ion trap MS/MS spectra were used to confirm peptide sequences of interest by querying specific fragment ions in the spectral library. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (Keller *et al.,* 2016) partner repository with the dataset identifier PXD006934.

Spectral library generation

MS/MS spectra from each DDA dataset were separately imported into MaxQuant (Cox & Mann, 2008)(version 1.6.0.1) and searched against the Biognosys iRT fasta database and the UniProt mouse proteome (June 2017 release) including both reviewed (Swiss-Prot) and unreviewed (TrEMBL) records of both canonical and isoform sequences. The software does not discriminate between Swiss-Prot and TrEMBL records at any stage; identified protein groups may contain both Swiss-Prot and TrEMBL proteins. Beta-methylthiolation was used as the fixed modification and Methionine oxidation and N-terminal acetylation as variable modifications. The minimum peptide length was set to 6, with at most one miss-cleavage allowed. For both peptide and protein identification a false discovery rate of 1% was set. MaxQuant search results were imported as spectral libraries into Spectronaut with default settings.

The Pulsar search engine integrated in Spectronaut 11 was used to identify peptides and proteins using only the DIA dataset with the exact search engine parameters (fasta database, modifications, peptide length, miss-cleavage, peptide and protein FDR) as listed above for MaxQuant. Finally, a merged spectral library based on search engine results from the IT, OT and DIA datasets was generated using Spectronaut. For each unique precursor, a consensus spectrum of relative fragment ion intensities was composed from all detections of the precursor over all datasets. There was no selection/prioritization by search engine or any other parameters when merging multiple identifications for a precursor. Its consensus iRT (normalized retention time) was computed from the evidence count (number of MS/MS detections) weighted median value. Theoretical m/z values for fragment ions and precursors were used in all spectral libraries. Protein inference on the merged library was performed on the principle of parsimony using the ID picker algorithm (Zhang *et al.,* 2007) as implemented in Spectronaut.

Analysis of DIA and WiSIM-DIA data

Quantitative analysis of DIA and WiSIM-DIA data was performed by using Spectronaut 11 in two separate analyses. Parameter settings of the software were the same for DIA and WiSIM-DIA and Spectronaut does not perform any computational steps particular to WiSIM-DIA. The generated output of each analysis contains qualitative and quantitative peptide-level data for both MS1 and MS2.

Dynamic retention time prediction was selected to enable non-linear alignment of precursor retention times between the (iRT, normalized retention time) spectral library and the DIA / WiSIM-DIA data by segmented regression (Bruderer, Bernhardt, Gandhi, Miladinovic, *et al.,* 2015; Bruderer, Bernhardt, Gandhi & Reiter, 2016). Mass calibration was performed by the software to estimate empirical mass accuracy and tolerances used during peak extraction, with initial tolerances set to +/- 40ppm for Orbitrap and +/- 0.5 Th for Ion Trap. While matching peptide fragment ions to the spectral library by retention time and m/z, Spectronaut can additionally use MS1 peptide elution profiles to disambiguate spectral library matches. The peptide identification score FDR, Q-value in Spectronaut output, was estimated with the mProphet approach (Rost *et al.,* 2014) integrated in Spectronaut using scrambled sequences as decoys (Supplementary Figure S6 shows target/decoy spectral library matching score distributions from Spectronaut). This score indicates the preciseness of the observed peptide match and its respective signature in the spectral library and was used as a qualitative metric in our downstream analysis.

The Spectronaut software computed MS1 peptide abundance as the summed precursor XIC (monoisotopic precursor ion plus isotopic envelope) and the MS2 peptide abundance as the summation of all selected fragment ions. In downstream analysis, peptide quantification for WiSIM-DIA and DIA was based on MS1 and MS2 abundances, respectively. The signal to noise ratio was computed using the same XIC profiles and peak integration boundaries; the value for ‘signal’ was defined as the maximum intensity within the peak integration boundary and the ‘noise’ as the average intensity outside of the peak integration boundary for the full width of the extracted XIC. DIA and WiSIM-DIA analysis results (which contain, among many others; peptide sequence, Q-value, MS1 and MS2 abundance value, MS1 and MS2 S/N) were exported as Spectronaut reports and further processed using the R language for statistical computation (Gillet *et al.,* 2012). In downstream analysis we used Spectronaut’s peptide Q-values to discriminate high confidence peptides within a set of triplicate DIA, or WiSIM-DIA, measurements; only peptides with Q-value ≤ 0.01 in at least 2 out of 3 replicates were used for quantitative analysis.

Acknowledgements

F. Koopmans was funded from the Netherlands Organization for Scientific Research (NWO) Complexity project 645.000.003.

Additional Information

The authors have declared no conflict of interest.

References

1. Bekker-Jensen, D. B. *et al.* An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. *Cell Syst* **4,** 587–599 e4. ISSN: 2405-4712 (Print) 2405-4712 (Linking) (2017).

2. Bruderer, R., Bernhardt, O. M., Gandhi, T., Miladinovic, S. M., *et al.* Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues. *Mol Cell Proteomics* **14,** 1400–10. ISSN: 1535-9484 (Electronic) 1535-9476 (Linking) (2015).

3. Bruderer, R., Bernhardt, O. M., Gandhi, T. & Reiter, L. High-Precision iRT Prediction in the Targeted Analysis of Data-Independent Acquisition and Its Impact on Identification and Quantitation. *Proteomics* **16,** 2246–56. ISSN: 1615-9861 (Electronic) 1615-9853 (Linking) (2016).
4. Conrads, T. P. *et al.* Utility of Accurate Mass Tags for Proteome-Wide Protein Identification. *Anal Chem* **72,** 3349–54. ISSN: 0003-2700 (Print) 0003-2700 (Linking) (2000).
5. Cox, J. & Mann, M. MaxQuant Enables High Peptide Identification Rates, Individualized p.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification. *Nature Biotechnology* **26,** 1367–1372 (Dec. 2008).
6. Gillet, L. C. *et al.* Targeted Data Extraction of the MS/MS Spectra Generated by Data-Independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Molecular & Cellular Proteomics* **11** (2012).
7. Gonzalez-Lozano, M. A. *et al.* Dynamics of the Mouse Brain Cortical Synaptic Proteome during Postnatal Brain Development. *Sci Rep* **6,** 35456. ISSN: 2045-2322 (Electronic) 2045-2322 (Linking) (2016).
8. Hondius, D. C. *et al.* Profiling the Human Hippocampal Proteome at All Pathologic Stages of Alzheimer's Disease. *Alzheimers Dement* **12,** 654–68. ISSN: 1552-5279 (Electronic) 1552-5260 (Linking) (2016).
9. Ivanov, M. V. *et al.* MS/MS-Free Protein Identification in Complex Mixtures Using Multiple Enzymes with Complementary Specificity. *J Proteome Res.* ISSN: 1535-3907 (Electronic) 1535-3893 (Linking) (2017).
10. Kang, Y. *et al.* SWATH-ID: An Instrument Method Which Combines Identification and Quantification in a Single Analysis. *Proteomics.* ISSN: 1615-9861 (Electronic) 1615-9853 (Linking) (2017).
11. Keller, A. *et al.* Opening a SWATH Window on Posttranslational Modifications: Automated Pursuit of Modified Peptides. *Mol Cell Proteomics* **15,** 1151–63. ISSN: 1535-9484 (Electronic) 1535-9476 (Linking) (2016).
12. Kiyonami, R. *et al.* Large-Scale Targeted Protein Quantification Using WiSIM-DIA on an Orbitrap Fusion Tribrid Mass Spectrometer. *Applicatioon note Thermo Scientific,* 1–8 (2014).
13. Li, S. *et al.* Optimization of Acquisition and Data-Processing Parameters for Improved Proteomic Quantification by Sequential Window Acquisition of All Theoretical Fragment Ion Mass Spectrometry. *J Proteome Res* **16,** 738–747. ISSN: 1535-3907 (Electronic) 1535-3893 (Linking) (2017).
14. Li, Y. *et al.* Group-DIA: Analyzing Multiple Data-Independent Acquisition Mass Spectrometry Data Files. *Nat Methods* **12,** 1105–6. ISSN: 1548-7105 (Electronic) 1548-7091 (Linking) (2015).
15. Liu, H., Sadygov, R. G. & Yates J. R., 3. A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics. *Anal Chem* **76,** 4193–201. ISSN: 0003-2700 (Print) 0003-2700 (Linking) (2004).

16. Mao, Y. *et al.* Identification of Phosphorylated Human Peptides by Accurate Mass Measurement Alone. *Int J Mass Spectrom* **308,** 357–361. ISSN: 1387-3806 (Print) 1387-3806 (Linking) (2011).

17. Michalski, A., Cox, J. & Mann, M. More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority Is Inaccessible to Data-Dependent LC-MS/MS. *J. Proteome Res.* **10,** 1785–1793 (Apr. 2011).

18. Pandya, N. J. *et al.* Group 1 Metabotropic Glutamate Receptors 1 and 5 Form a Protein Complex in Mouse Hippocampus and Cortex. *Proteomics* **16,** 2698–2705. ISSN: 1615-9861 (Electronic) 1615-9853 (Linking) (2016).

19. Rost, H. L. *et al.* OpenSWATH Enables Automated, Targeted Analysis of Data-Independent Acquisition MS Data. *Nat Biotechnol* **32,** 219–23. ISSN: 1546-1696 (Electronic) 1087-0156 (Linking) (2014).

20. Tsou, C. C. *et al.* DIA-Umpire: Comprehensive Computational Framework for Data-Independent Acquisition Proteomics. *Nat Methods* **12,** 258–64, 7 p following 264. ISSN: 1548-7105 (Electronic) 1548-7091 (Linking) (2015).

21. Wang, J. *et al.* MSPLIT-DIA: Sensitive Peptide Identification for Data-Independent Acquisition. *Nat Methods* **12,** 1106–8. ISSN: 1548-7105 (Electronic) 1548-7091 (Linking) (2015).

22. Zhang, B., Chambers, M. C. & Tabb, D. L. Proteomic Parsimony through Bipartite Graph Analysis Improves Accuracy and Transparency. *J Proteome Res* **6,** 3549–57. ISSN: 1535-3893 (Print) 1535-3893 (Linking) (2007).

4

Correlation profiling of brain subcellular proteomes reveals co-assembly of synaptic proteins and subcellular distribution

*Frank Koopmans[1][2], *Nikhil J. Pandya[2], Johan A. Slotman[3], Iryna Paliukhovich[2], Adriaan B. Houtsmuller[3], #August B. Smit[2], #Ka Wan Li[2]

*Co-first authors, #co-senior authors

[1]Department of Functional Genomics, [2]Department of Molecular and Cellular Neurobiology, Center for Neurogenomics Cognitive Research, VU University, Amsterdam, The Netherlands
[3]Optical Imaging Center, Department of Pathology, Erasmus Medical Center, Rotterdam, The Netherlands

Protein correlation profiling might assist in defining co-assembled proteins and subcellular distribution. Here, we quantified the proteomes of five biochemically isolated mouse brain cellular sub-fractions, with emphasis on synaptic compartments, from three brain regions, hippocampus, cortex and cerebellum. We demonstrated the expected co-fractionation of canonical synaptic proteins belonging to the same functional groups. The enrichment profiles also suggested neuronal culture we confirmed the postsynaptic localization of PLEKHA5 and ADGRA1. We further detected profound brain region specific differences in the extent of enrichment for some functionally the presence of many novel pre- and post-synaptic proteins. Using super-resolution microscopy on primary associated proteins. This is exemplified by different AMPA receptor subunits and substantial differences in sub-fraction distribution of their potential interactors, which implicated the differences of AMPA receptor complex compositions. This resource aids the identification of proteins partners and subcellular distribution of synaptic proteins.

This chapter has been published in *Scientific reports 7, no. 1 (2017)*

Introduction

Brain function is carried in large by synaptic transmission between neurons. Synaptic transmission relies on the stimulus-dependent release of transmitter from the presynaptic element and receptor-mediated signal perception and integration at the postsynapse. In both elements of the synapse crucial functional assemblies of proteins have been identified underlying synaptic function. These, for instance, include the interaction of t- and v-SNARE proteins STX1/2, SNAP25 and VAMP2 (Ackermann *et al.,* 2015; Bayes, van de Lagemaat, *et al.,* 2011; Milovanovic & Jahn, 2015; Sudhof, 2012), facilitating the Ca2+ influx-induced fusion of docked/primed synaptic vesicles to the membrane of the presynaptic active zone. This event leads to the exocytotic release of glutamate which binds to and activates NMDA- and AMPA-type receptors residing in the opposing post-synaptic membrane (Sheng & Kim, 2011). These receptors in turn are anchored to the postsynaptic density (PSD) via an assembly of scaffolding proteins belonging, among others, to DLG and SHANK families (Zhu *et al.,* 2016).

Using proteomics analysis various molecular machineries governing synaptic sub-function, have been characterized successfully, including the synaptic vesicle, the pre-synaptic active zone and PSD (Bayes, Collins, *et al.,* 2012; Distler *et al.,* 2014; Dosemeci *et al.,* 2007; K. W. Li *et al.,* 2004; Sialana *et al.,* 2016; Takamori *et al.,* 2006; Volknandt & Karas, 2012; Bayes, Collins, *et al.,* 2012). In most cases these topological synaptic subdomains were isolated biochemically, and subsequently subjected to proteomics analysis to define its constituents. A caveat of this is that although the sample is enriched in proteins from the targeted compartment it contains proteins from other compartments, albeit at a lower level. Correlation profiling (K. Li *et al.,* 2005) (also known as fractionation/spatial profiling (Borner *et al.,* 2014; Lund-Johansen *et al.,* 2016)) of the step-wise isolated fractions is a preferred method to deal with protein enrichment in fractions, from which their co-assembly or subcellular localization might be inferred.

Here we applied data-dependent mass spectrometry on five biochemically isolated cellular sub-fractions, with focus on synaptic compartments, from three mouse brain regions, hippocampus, cortex and cerebellum. In each brain region at least 3000 proteins were quantified with a total of 4237 unique proteins in the complete dataset, with a sub-fraction typically yielding ~2000 proteins. Correlation profiling of some well-known synaptic functional groups showed enrichment of pre-synaptic proteins in synaptosome or canonical postsynaptic proteins in the PSD.

The AMPA receptor (AMPAR), a genuine postsynaptic receptor, showed considerable brain region-specific distribution profiles. Also, the recently reported high-confident AMPAR interacting proteins (Chen, Pandya, *et al.,* 2014; Schwenk *et al.,* 2014; K. Li *et al.,* 2005) followed different distribution patterns. These data suggest the presence of spatially segregated, distinct AMPAR sub-complexes. Our analyses further suggested the presence of novel synaptic proteins. We validated PSD localization of PLEKHA5 and ADGRA1 by super-resolution microscopy. Together, the present study provides a rich resource to interrogate the potential co-assembly and subcellular distribution of synaptic proteins in hippocampus, cortex, and cerebellum.

Results and Discussion

Generating a synaptic proteome resource for correlation profiling

We isolated biochemical cellular sub-fractions, with emphasis on synaptic compartments, from three brain regions (Figure 4.1) followed by label-free data dependent MS analysis of each fraction similar to previously described spatial/fractionation correlation proteomics(Borner *et al.,* 2014). Data visualization methods were chosen to emphasize the (1) quantitative fraction distribution of a few selected proteins by their iBAQ values, or the (2) normalized fraction distribution of a large number of proteins to accommodate the profound quantitative differences that span 4 orders of magnitude. First, the experimental reproducibility was addressed. Biological replicates of proteomics analyses were performed on hippocampal microsome (M), P2, synaptosome (SYN), synaptic membrane (SYM) and postsynaptic density (PSD). Three replicates of P2 and synaptic membrane, show each R2 of at least 0.856, which is representative of all synapse sub-fractions (Supplemental Figure S1). Overall, the coefficient of variation of each synaptic fraction ranges from 26-34% (Supplemental Figure S1), as is typical for data-dependent analysis, and has been reported previously for other biological systems (Piehowski *et al.,* 2013).

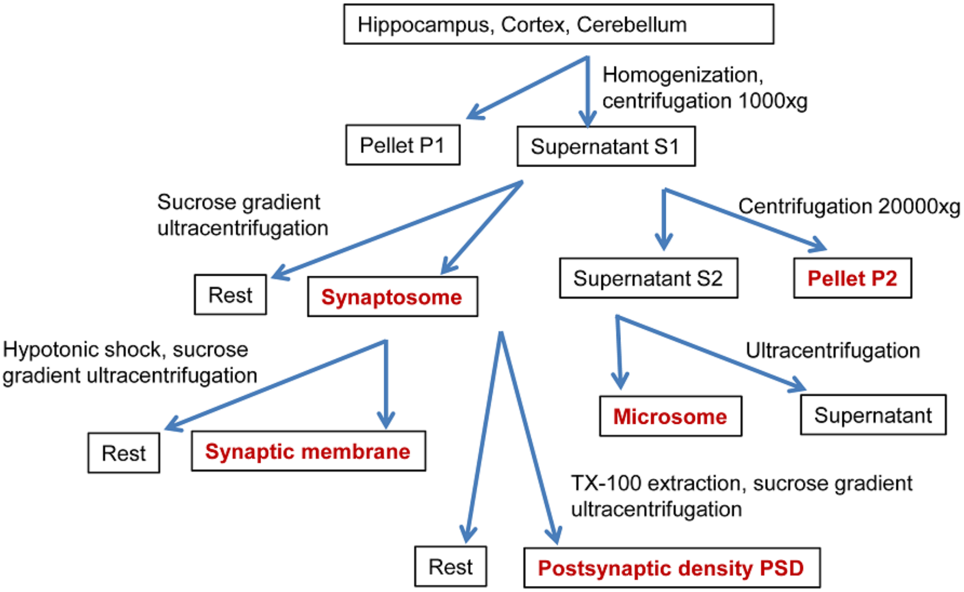
Hippocampus, Cortex, Cerebellum
Homogenization, centrifugation 1000xg
Pellet P1
Supernatant S1
Sucrose gradient ultracentrifugation
Centrifugation 20000xg
Rest
Synaptosome
Supernatant S2
Pellet P2
Hypotonic shock, sucrose gradient ultracentrifugation
Ultracentrifugation
Microsome
Supernatant
Rest
Synaptic membrane
TX-100 extraction, sucrose gradient ultracentrifugation
Rest
Postsynaptic density PSD

Figure 4.1: Biochemical isolation of cellular sub-fractions from three brain regions. Fractions labelled red were collected for proteomics analysis.

Approximately 2000 proteins were identified in each sub-fraction, with the exception of PSD in which about 1400 identified proteins were detected, likely reflecting lower sample complexity. The dynamic range of the protein abundances in each fraction spans over four orders of magnitude (Supplemental Figure S1). To reveal the distribution of each protein across the synapse sub-fractions within a brain region, we scaled protein abundances between zero and 100% of their maximum value over all sub-fractions. Hierarchical cluster analysis of 3632 proteins quantified in five hippocampus sub-fractions revealed the presence of sub-fraction-specific groups of proteins (Figure 4.2). Only 166 proteins, ~6% of all proteins quantified in the hippocampus sub-fractions, were found exclusively in P2, whereas the remaining proteins showed a variety of more complex enrichment profiles for the other sub-fractions. The differential distribution of proteins across the fractions was at the basis for subsequent correlation profiling analysis.

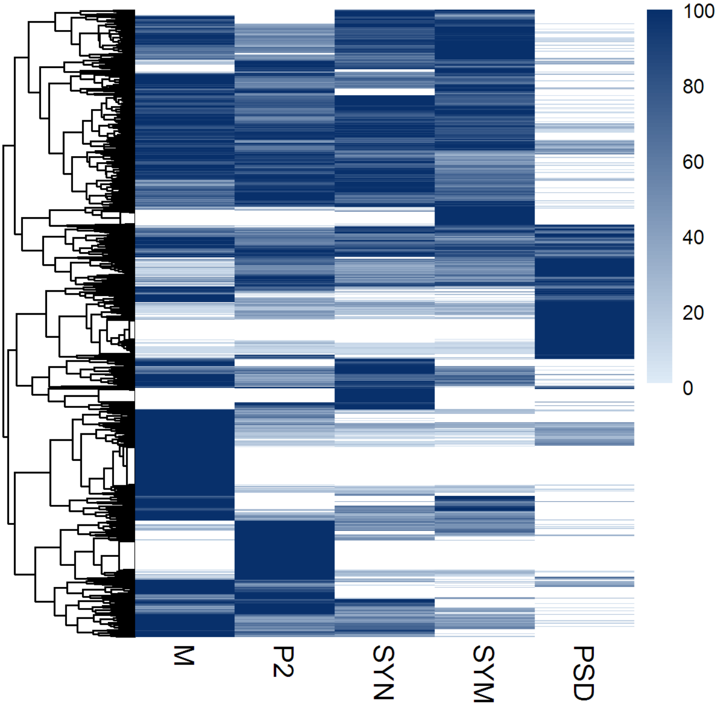

Figure 4.2: Hierarchical clustering of 3632 proteins shows specific enrichment in various subcellular fractions of the hippocampus. Protein abundances were scaled between zero and 100% of their maximum over all sub-fractions.

Next, we carried out the same analysis of cortical and cerebellar synapse sub-fractions and shown in the protein abundance matrices for all the fractions from cortex, cerebellum and hippocampus (Supplemental Table S1). As the analysis of hippocampal samples revealed a low variation of biological replications, the analysis of single samples of cortex and cerebellum should be sufficient. Hierarchical cluster analyses of proteins quantified in the cellular sub-fractions from cortex and cerebellum are shown in Supplementary Figure S2. Immuno-blotting was performed on a selection of canonical pre- and postsynaptic proteins, the presynaptic vesicle protein SYP and the postsynaptic NMDA receptor subunit GRIN2A, and the scaffolding protein DLG4 (PSD95/SAP90) that anchors NMDA receptor and other proteins to the PSD, to validate the data as revealed by mass spectrometry. The similarity of protein distribution profiles between these two different measurement methods is high (Figure 4.3). SYP was found enriched in synaptosome and synaptic membrane and highly depleted or absent in the PSD fraction. GRIN2 and DLG4 were highly enriched in PSD. Thus, immunoblotting confirmed the mass spectrometry based data.

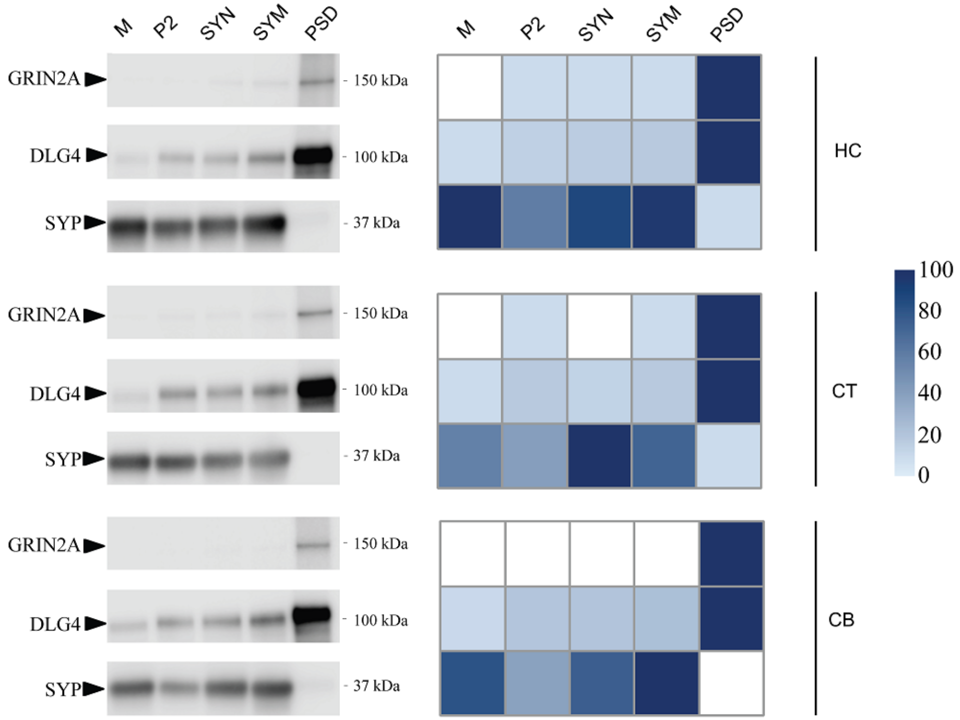

Figure 4.3: Protein abundance distribution over the cellular sub-fractions in hippocampus (HC), cortex (CT) and cerebellum (CB). Left panel is an immunoblot analysis showing the presence of presynaptic protein SYP, and postsynaptic proteins GRIN2A and DLG4 in each sub-fraction. Right panel shows protein abundance in each sub-fraction as observed with proteomics. Mass spectrometry protein abundance values were scaled between zero and 100% of their maximum over all sub-fractions in each panel and accordingly color-coded from light-blue to dark-blue.

Functional grouping by correlation profiling

We next investigated whether proteins might be delineated that assemble into correlated functional groups and are related to known synaptic processes. For this we used the protein abundance profiles over sub-fractions of the hippocampus. Correlation profiling was performed using three 'seeds' that are well-known representatives of SV exocytosis and the PSD, respectively. Protein profiles in the entire catalog, with a Pearson correlation of ≥ 0.9 with two of the three seed proteins, were selected (Figure 4.4, Supplemental Table S2). Proteins strongly correlated with the exocytosis seeds were enriched in synaptosomes, synaptic membranes and the microsomal fraction and were found depleted in P2 and PSD (Figure 4.4A). Conversely, the seed proteins for the PSD group were depleted in synaptosomes, synaptic membranes and microsomes (Figure 4.4B), but with a larger variation in their enrichment for PSD compared to other fractions. Proteins of the PSD are in many cases differentially detected in other fractions, potentially indicative of differences in assembly during routing to the PSD, or alternatively, being part of different functional units within and outside the PSD. Obviously, the inclusion criteria for similar profiles can be adjusted. For example, the correlation coefficient threshold might be adjusted, or one might opt for a more stringent setting with inclusion of more seed proteins. The impact on the number of correlated proteins and trade-off in true/false-positives can be explored with provided data based on user-specified criteria (see example Supplemental Figure S3 and Supplemental Table S2). The correlation profiling of each protein against all other identified proteins in hippocampus, cortex and cerebellum, are shown in Supplementary Table S3-5, respectively.

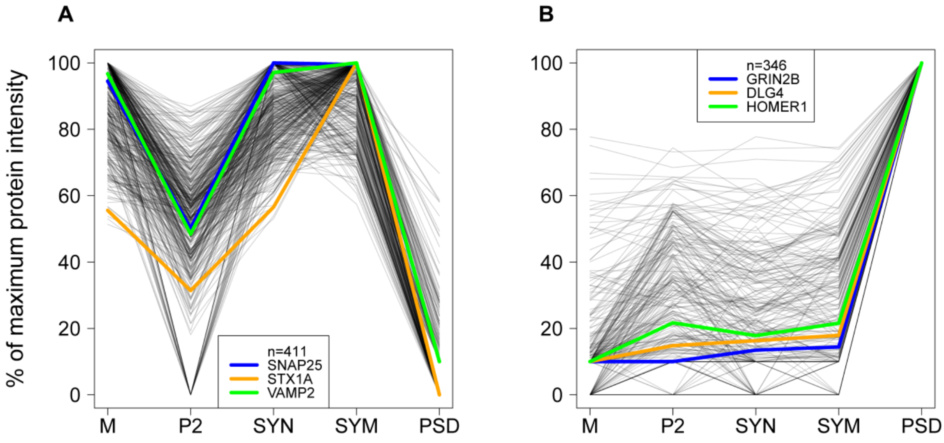

Figure 4.4: Correlation analysis of abundance profiles of proteins over sub-fractions of the hippocampus with selected seed proteins. In each panel, three functionally related seed proteins were chosen typical for each of the specific synaptic processes in A and B. Proteins shown have a Pearson correlation of at least 0.9 with at least two out of these three seed proteins. Protein abundances are scaled between zero and their maximum intensity over all selected sub-fractions. This leads to A) exocytosis, 411 proteins and B) postsynaptic density, 346 proteins.

Subcellular localization by correlation profiling

Correlation profiling may indicate proteins that typically co-localize in a given subcellular localization. An alternative approach is to affinity isolate the organelle of interest for quantitative proteomics analysis. This might yield higher organelle purity than that obtained from biochemical isolation, but the presence of contaminants in an affinity-precipitated sample can still not be excluded (Chen, Koopmans, *et al.,* 2015; Chen, Pandya, *et al.,* 2014). To benchmark our approach, we compared the proteins from GRIN2B-DLG4-HOMER1 set (cf. Figure 4.4B) to the previously reported affinity purified PSD using an anti-DLG4 antibody (Figure 4.5). Out of the top ranking proteins from the affinity isolated PSD (Table 2 from Dosemeci et al (Dosemeci *et al.,* 2007), containing 49 unique proteins), 31 were contained in the GRIN2B-DLG4-HOMER1 set and 8 were enriched in PSD with lower correlation (between 0.5-0.9). Of the remaining 10 proteins, 1 showed a slightly weaker correlation of 0.49, 9 had no to negative correlation. The non-compliance of these proteins may be explained at least in part by their major localization site outside the PSD/spine. For example, tubulins (found in affinity-purified PSD) are abundantly present in the dendritic shaft, i.e. outside the synapse, for long distant transport and mechanical stability, whereas only a small fraction of microtubules may protrude into spine in an activity dependent manner (Kapitein & Hoogenraad, 2015) and was reported to be present within PSD (Yun-Hong *et al.,* 2011). Thus, correlation profiling generates data that also includes information on enrichment and alternative distribution, whereas affinity isolation only addresses co-assembled proteins in the context of the bait.

Of particular interest, different protein family members can differ in their correlation profiles. For example, DLG2, DLG3 and DLG4 displayed the same PSD-enriched profile, different from DLG1. DLG1 was reported involved in multiple functions, i.e. biosynthesis and trafficking of glutamate receptors (Jeyifous *et al.,* 2009; Sans *et al.,* 2001), as well as the recruitment of components of vesicle trafficking machinery either to the plasma membrane or to transport vesicles (Walch, 2013). Also, GRIA3 followed the DLG4 profile, whereas GRIA2 showed a lower correlation. The majority of GRIA2 is present in GluA1/2 receptors that are known to have high membrane mobility. In fact, in hippocampus, there is a sizeable amount of extra-synaptic AMPAR, and AMPAR in synaptic cytosol and dendrites, which may form reserve or recycling pools for activity-dependent plastic changes of PSD-trapped AMPARs (Constals *et al.,* 2015). The latter may explain the observed GRIA1 and GRIA2 in the microsome fraction, resulting in a lower correlation to the GRIN2B-DLG4-HOMER1 profile. Correlation profiling clearly reveals the house keeping proteins, such as ALDOA and GAPDH as not PSD specific (Figure 4.5C).

Glutamate receptors and their interacting proteins

Interestingly, the GRIA subunits forming AMPARs were highly enriched in cortex and cerebellum PSD and to a lesser extend in hippocampus. Compared to the GRIN2B-DLG4-HOMER1 profile, GRIA3 and GRIA4 showed a 0.99 Pearson correlation (median value of these seed proteins) in all three brain regions whereas GRIA1 and GRIA2 showed a lower PSD enrichment in hippocampus reflected by 0.66 and 0.88 correlations, respectively. The percentage of GRIA1 in the PSD fraction of cortex, cerebellum and hippocampus were 74%, 55% and 23%, respectively (Figure 4.6). This indicates a by far higher fraction of AMPARs anchoring to the PSD in cortex and cerebellum. Consequently, a lower percentage of AMPARs in cortex

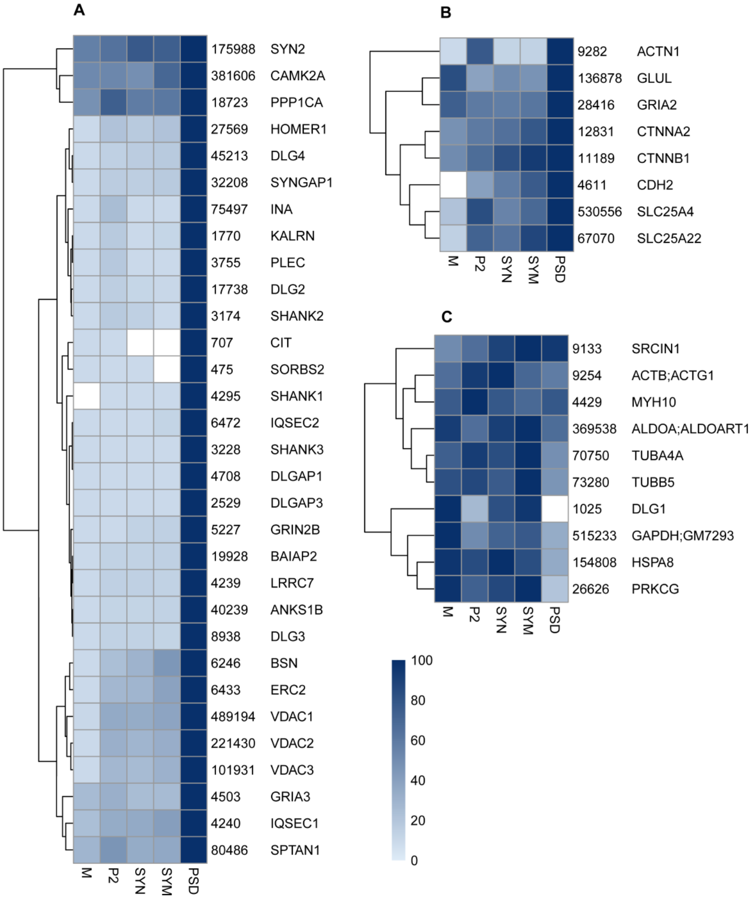

Figure 4.5: Comparison of the 49 PSD-95 (DLG4)-affinity associated proteins as listed in Table 2 from Dosemeci et al. (Dosemeci *et al.*, 2007). A) 31 proteins from this reference set are recovered by correlation analysis with hippocampal sub-fractions (as used for Figure 4b). B) 8 additional reference proteins are recovered using the Pearson correlation threshold from 0.9 to 0.5 (for 2+ reference proteins). C) 10 remaining proteins from the reference set which were not recovered by correlation analysis. Summed iBAQ abundance of each protein is shown alongside each protein name. Protein abundances were scaled between zero and 100% of their maximum over all sub-fractions and color-coded from light-blue to dark-blue according to the heatmap.

and cerebellum might be available for trafficking in and out of the PSD, which may impact on the extent of post-synaptic plasticity.

Equally, AMPAR interactors as previously established (Chen, Pandya, *et al.*, 2014; K. Li *et al.*, 2005) exhibited highly variable distribution patterns across the sub-fractions (Figure 4.6A). For example, CACNG2 was enriched in the PSD, CPT1C and SACM1L were highly enriched in microsome. In hippocampus and cortex, FRRS1L has a more even distribution in microsome and synaptosome but not in PSD. This is in general agreement with the previous studies indicating the co-localization of a population of AMPA receptors with CPT1C intracellularly rather than on the cell membrane (Gratacos-Batlle *et al.,* 2014), whereas CACNG2 traps AMPA receptors at the PSD and modulates channel properties (Bats *et al.,* 2007; Shaikh *et al.,* 2016). Different members of the CACNG family showed pronounced brain-region specific expression differences. CACNG2 is the predominant form of CACNG in the cerebellum residing mainly in the PSD. CACNG8 is more abundant in hippocampus, and is widely distributed across all the examined sub-cellular compartments. This is in agreement with a recent study showing the restricted distribution of CACNG2 mainly at the perforated synapses of pyramidal cells and the synapses of parvalbumin-positive interneurons, whereas CANCG8 was present at various synapse types(Yamasaki *et al.,* 2016). Unlike the differential distribution of AMPAR interactors, the NMDAR subunits and their interactors DLG4, DLGAP4, HOMER1 and SHANK1 showed high enrichment in PSD in all three brain regions (Figure 4.6B).

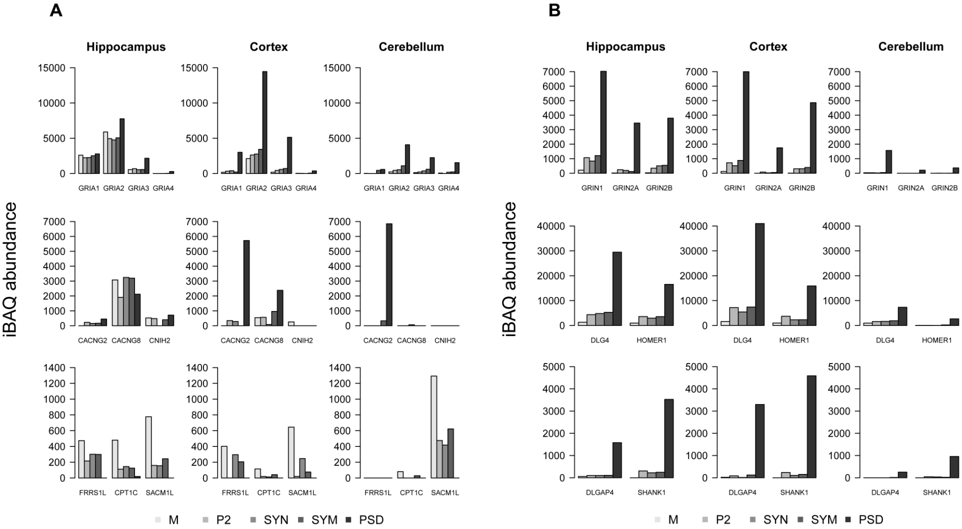

Figure 4.6: The sub-cellular abundance of ionotropic glutamate receptors and their associated proteins in three brain regions. (A) AMPA receptor subunits GRIA1-4, and 6 known auxiliary subunits, compared between sub-fractions and brain regions. Each group in the bar graph reflects the abundance of a protein over sub-fractions. (B) NMDA receptor subunits GRIN1 and GRIN 2A/B, and the interactors DLG4, DLGAP4, HOMER1 and SHANK1. Legend at the bottom reflects the order of sub-fractions and their respective gray scale coding. Protein iBAQ values are computationally approximated absolute abundances.

Postsynaptic density proteins

Proteins that share similar profile have a high chance to be present in the same sub-cellular compartment. We selected two PSD proteins for further study, namely ADGRA1 and PLEKHA5, which showed high correlation to the GRIN2B-DLG4-HOMER1 profile and were found in the PSD fraction of all three brain regions. PLEKHA5

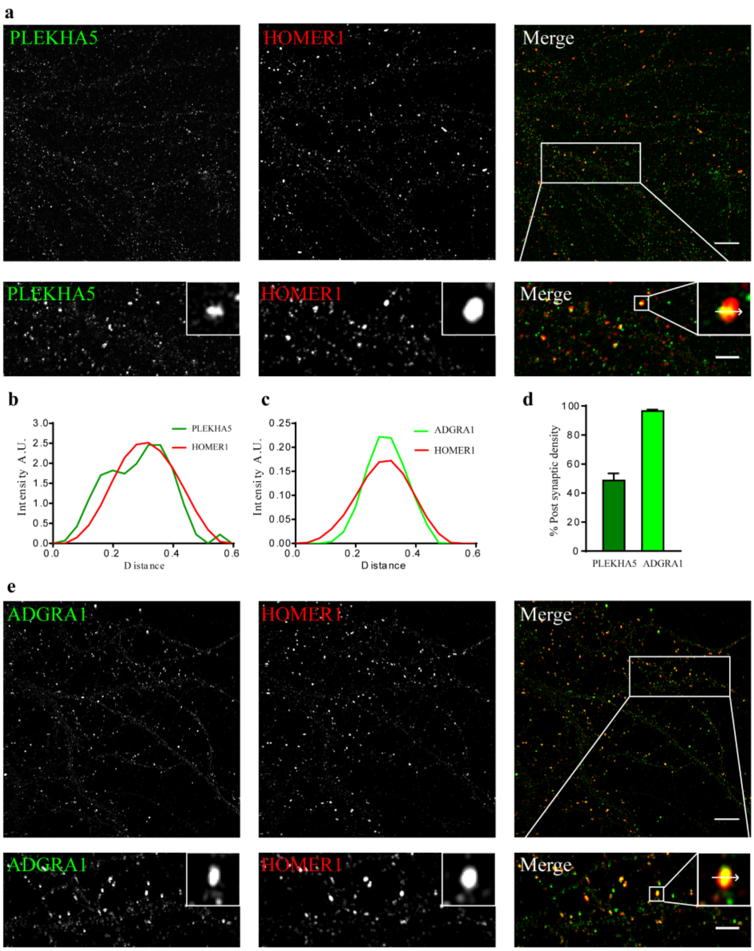

Figure 4.7: Super-resolution imaging microscopy validation of novel PSD-enriched proteins, PLEKHA5 and ADGRA1. a,e) SIM imaging of primary cultured hippocampal neurons at DIV19 for PLEKHA5 and ADGRA1 (green) along with HOMER1 (red), the latter as a marker for the PSD. Lower panel shows zoom in of the marked area (scale bar 2 µm). b,c) Line scan analysis on inset from panel a,e shows co-localization between the three proteins at postsynaptic sites. d) Bar graph showing percentage (mean±sem) of postsynaptic density HOMER1 puncta positive for PLEKHA5 (dark green, 48.61 ± 4.92, n = 10), ADGRA1 (light green 96.54 ± 1.02, n = 7).

was found previously in a PSD preparation (Bayes, van de Lagemaat, *et al.,* 2011), but PSD localization has never been demonstrated. Here, we performed super-resolution structured illumination microscopy and showed that PLEKHA5 co-localizes with the PSD marker Homer (Figure 4.7a-b), thereby confirming it as a PSD protein. PLEKHA5 was present in about 50% of synapses (Figure 4.7d). In contrast, ADGRA1 was exclusively found in the PSD fraction with an iBAQ value 300 fold lower than GRIA1, showing that its copy number in the PSD is very low, and it might have been missed in previous experiments because of this. We observed excellent overlap of ADGRA1 and HOMER1 in primary hippocampal cultures (Figure 4.7e). Line scan analysis of synapses showed nearly perfect co-localization of HOMER1 and ADGRA1 (Figure 4.7c). We further quantified the percentage of ADGRA1 positive synapses and revealed that 96.54% of HOMER1 positive puncta representing synapses were positive for ADGRA1 (Figure 4.7d).

ADGRA1 is present on chromosome 10q26.3 and is a 7-TM domain (Fredriksson *et al.,* 2003) containing protein belonging to the adhesion family of G protein-coupled receptors. It is predicted to have a PDZ binding domain in the C-terminus of the protein (Lagerstrom *et al.,* 2007). It is the only member of the adhesion GPCRs that lacks a cleavable GPCR auto-proteolysis–inducing (GAIN) domain (Hamann *et al.,* 2015). Due to its widespread expression in the brain, it is suggested that ADGRA1 may have an important role in the regulation of neuronal signal transduction. It has been reported in complex with PSD95 (Dlg4) (Fernandez *et al.,* 2009), which is in line with its tight localization within the PSD. Further research on the functions of ADGRA1 might elucidate the processes in which it is involved.

PLEKHA5 is a cytosolic protein belonging to the PLEKHA5 family and is a Pleckstrin Homology (PH) Domain containing protein through which it is involved in binding to Phosphatidylinositol (3, 4, 5)-trisphosphate (PIP3) (Yamada *et al.,* 2012). Previous PSD identification studies from human postsynaptic densities have identified PLEKHA5 in the postsynaptic density preparations, but validation of its PSD localization was not reported (Bayes, van de Lagemaat, *et al.,* 2011). Interestingly, PLEKHA5 belongs to region 12p12 and SNPs associated with this locus are associated with early onset bipolar disorder (Jamain *et al.,* 2014). Like other PH domain containing proteins, PLEKHA5 might get recruited to the plasma membrane upon PIP3 formation in the postsynaptic compartments. The role that it plays in the postsynaptic density remains to be determined.

In conclusion, correlation profiling of synaptic sub-fractions aids to reveal interactome organization and may help to define subcellular structures of interest (Breckels *et al.,* 2016; Lund-Johansen *et al.,* 2016). Typically, immunoblotting of synaptic protein distribution corresponded well with our sub-fraction proteomics data. When comparing our data with a previously reported immuno-purified PSD protein complex, we observed high degree of agreement. Significantly, our large dataset with thousands of proteins covering different synapse sub-fractions from three brain regions allows to interrogate the biochemically-defined spatial distribution of most synaptic proteins in a brain-region specific manner, and may help to generate hypotheses regarding novel protein localizations related to synapse functions.

Materials and Methods

Preparation of cellular sub-fractions

All animal experiments were performed in accordance with relevant guidelines and regulations of the VU University. The animal ethics committee of the VU University approved the experiments. Subcellular fractions were prepared from 3-month-old

C57BL6 mice as described in (von Engelhardt *et al.,* 2010). In brief, mouse hippocampi, cortex and cerebellum were dissected and stored at -80oC until used. The brain regions were pottered separately in homogenization buffer (0.32 M Sucrose, 5 mM HEPES pH 7.4, Protease inhibitor cocktail (Roche)) on a dounce homogenizer (potterS; 12 strokes, 900 rpm) and spun at 1000xg for 10 min at 4oC. Supernatant 1 (S1) was centrifuged at 20,000xg for 20 min to obtain pellet 2 (P2) and supernatant 2 (S2). The S2 fraction was ultracentrifuged at 100,000xg for 2 hrs; the pellet was recovered as microsomal fraction. S1 was subjected to ultracentrifugation in a 0.85/1.2 M sucrose density gradient at 100,000xg for 2 hrs. Synaptosomes were recovered at the interface of 0.85/1.2 M sucrose. The hypotonic shock of synaptosomes in 5 mM HEPES with protease inhibitor for 15 min yielded the synaptic membrane fraction, which was subsequently isolated by sucrose gradient ultracentrifugation as stated above at the interface of 0.85/1.2M fraction. To obtain the PSD, the synaptosome fraction was extracted in 1% Tx-100 for 30 min, layered on top of 1.2/1.5/2 M sucrose, centrifuged at 100,000xg for 2 hrs, and recovered as PSD-I at the interface of 1.5/2 M sucrose. PSD-I was subjected to second extraction in 2% Tx-100 for 30 min, subjected to sucrose gradient ultracentrifugation as stated above, and recovered at the 1.5/2M sucrose interface. The PSD-II fraction was then pelleted in 5 mM HEPES by centrifuging at 100,000xg for 30 min.

Gel separation and in-gel digestion

Gel digestion was performed as described (Chen, Koopmans, *et al.,* 2015; Pandya *et al.,* 2016). Sample was dissolved in Laemmli buffer and boiled at 98 °C for 5 min; 5 μl 30% acrylamide was then added and vortexed for 30 min at room temperature to form a fixed modification of Cys-S-beta-propionamide. Samples were run on a 10% SDS-PAGE gel, which was stopped when the front reached halfway of the gel. The gel was fixed overnight in 40% ethanol/3% phosphoric acid, and stained briefly for about 30 min with colloidal coomassie blue. Each lane was cut into two slices, chopped into 1 mm by 1 mm pieces followed by a sequential incubation in 50% acetonitrile/50 mM NH3HCO3 – 100% acetronitrile – 50 mM NH3 HCO3 - 50% acetonitrile/50 mM NH3HCO3 – 100% acetronitrile. The gel pieces were dried in a speedvac, rehydrated in trypsin solution in 50 mM NH3HCO3 (500 ng per gel slice) at 37 °C overnight, extracted with 200 μL 0.1M acetic acid, and the supernatant was transferred to an Eppendorf tube and dried in a speedvac. The tryptic peptides were dissolved in 17 μL 0.1M acetic acid and analyzed by LC-MS/MS.

MS acquisition and data analysis

Peptides were analyzed by nano-LC MS/MS using an Ultimate 3000 LC system (Dionex, Thermo Scientific) coupled to the TripleTOF 5600 mass spectrometer (Sciex) (Carney *et al.,* 2014). Peptides were trapped on a 5 mm Pepmap 100 C18 column (300 μm i.d., 5μm particlesize, from Dionex) and fractionated on a 200 mm Alltima C18 column (100 μm i.d., 3 μm particle size). The acetonitrile concentration in the mobile phase was increased from 5 to 30% in 90 min, to 40% in 5 min, and to 90% in another 5 min, at a flow rate of 500 nL/min. The eluted peptides were electro-sprayed into the TripleTOF MS. The nano-spray needle voltage was set to

2500V. The mass spectrometer was operated in a data-dependent mode with a single MS full scan (m/z 350−1200, 250 msec) followed by a top 25 MS/MS (85 msec per MS/MS, precursor ion > 90 counts/s, charge state from +2 to +5) with an exclusion time of 16 sec once the peptide was fragmented. Ions were fragmented in the collision cell using rolling collision energy, and a spread energy of 10eV.

The MS raw data were imported into MaxQuant (version 1.5.2.8) (Cox & Mann, 2008), and searched against the UniProt mouse proteome (SwissProt+Trembl February 2016 release) with Cys-S-beta-propionamide as the fixed modification and Methionine oxidation and N-terminal acetylation as variable modifications. For both peptide and protein identification a false discovery rate of 0.01 was set, MaxLFQ normalisation was enabled with a LFQ minimal ratio count of 1. The minimal peptide length was set to 6; further MaxQuant settings were left at default. The MaxQuant search results are provided in Supplementary Table S6. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (Vizcaino *et al.,* 2016) partner repository with the dataset identifier PXD005634.

External contaminants such as immunoglobulins, keratin and trypsin, as well as histones, were excluded from downstream analysis. Next, we collapsed protein groups that shared the same gene name. All (majority) protein accessions in a protein group were matched against the Fasta database, their gene names extracted from the Fasta headers and the set of unique gene names for each protein group was stored. The protein abundance matrix was built by summation of MaxLFQ normalised protein intensities of protein groups that map to the same unique set of genes. These protein intensities were converted to iBAQ pseudo-absolute abundances using the number of digestible peptides provided by MaxQuant and then replicates were merged using their respective mean protein abundances (missing values were disregarded). This data is provided in Supplemental Table S1.

Immunostaining of primary neurons

Primary hippocampal neurons were obtained from E18 rat pups as described previously (Frischknecht *et al.,* 2009). Briefly, 18000 cells were grown in neurobasal medium supplemented with B27 on poly D-Lysine coated coverslips. The cells were used for staining at DIV 14-16. The coverslips were fixed with ice-cold methanol for 10 min, followed by three washes in ddH2O and PBS. The neurons were then blocked and permeablised with blocking buffer (5% FCS, 0.1% Triton X-100, and 0.1% Glycine in phosphate buffered saline, pH 7.4) for 1 hour. Next, the neurons were incubated with anti-ADGRA1 (GPR123) (1 in 250, cat. no. sc-162892, Santa cruz) or anti-PLEKHA5 (1 in 250, cat. no. sc-390311, Santa cruz) and anti-HOMER1 (1 in 1000, cat. no. 160 004, Synaptic systems), diluted in blocking buffer overnight at 4 °C. After three times washing in PBS, the cells were incubated with an alexa conjugated secondary antibodies for 1 hour at room temperature (anti-mouse Alexa 488 (1 in 1000), anti-goat Alexa 488 (1 in 1000), Anti-Guinea pig Alexa 647 (1 in 1000) (Molecular Probes) and subsequently washed and fixed on glass slides (Superfrost Plus, Thermo) using Moviol. Images were taken using a LSM Elyra SIM microscope with 63x Oil immersion lens (N.A. 1.4) and analyzed using ImageJ. Line scan analysis was performed as described in (Frischknecht *et al.,* 2009).

3D-SIM microscopy

Imaging was performed using a Zeiss Elyra PS1 system. 3D-SIM data was acquired using a 63x 1.4NA oil objective. 488 nm, 561 nm, 642 nm, 100mW diode lasers were used to excite the fluorophores together with respectively a BP 495-575 + LP 750, BP 570-650 + LP 750 or LP 655 emission filter. For 3D-SIM imaging a grating was present in the light path. The grating was modulated in 5 phases and 5 rotations, and multiple z-slices were recorded with an interval of 110 nm on an Andor iXon DU 885, 1002x1004 EMCCD camera. Raw images were reconstructed using the Zeiss Zen software.

Reconstructed 3D-SIM images were analyzed with imageJ (Schneider *et al.,* 2012) extended in the FIJI framework (Schindelin *et al.,* 2012). Particles larger than 10 pixels were detected and marked as region of interest (ROI) and mean ADGRA1 or PLEKHA5 signals inside the ROIs were measured. A HOMER1 particle was counted as positive for ADGRA1 or PLEKHA5 if their mean intensity was more than three times above their respective local backgrounds.

Acknowledgements

N.J. Pandya was sponsored by EU International training site ‘Braintrain’ (MCITN-CT-2009-031910); F. Koopmans was funded from the Netherlands Organization for Scientific Research (NWO) Complexity project 645.000.003; K.W. Li and A.B. Smit were in part funded by EU FP7 ‘SynSys’ (grant #242167).

Author Contributions

N.J.P. performed the experiments. N.J.P., F.K., A.B.S. and K.W.L. wrote the manuscript; N.J.P., F.K., and I.P. performed the data analysis of MS data. J.A.S. and N.J.P. performed data analysis of imaging data. N.J.P, F.K., A.B.S. and K.W.L designed the experiments. A.B.H, A.B.S. and K.W.L. supervised the research. All authors reviewed the manuscript.

Additional Information

The authors have declared no conflict of interest.

References

1. Ackermann, F., Waites, C. L. & Garner, C. C. Presynaptic Active Zones in Invertebrates and Vertebrates. *EMBO Rep* **16,** 923–38. ISSN: 1469-3178 (Electronic) 1469-221X (Linking) (2015).
2. Bats, C., Groc, L. & Choquet, D. The Interaction between Stargazin and PSD-95 Regulates AMPA Receptor Surface Trafficking. *Neuron* **53,** 719–34. ISSN: 0896-6273 (Print) 0896-6273 (Linking) (2007).

3. Bayes, A., Collins, M. O., *et al.* Comparative Study of Human and Mouse Postsynaptic Proteomes Finds High Compositional Conservation and Abundance Differences for Key Synaptic Proteins. *PLoS One* **7,** e46683. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking) (2012).

4. Bayes, A., van de Lagemaat, L. N., *et al.* Characterization of the Proteome, Diseases and Evolution of the Human Postsynaptic Density. *Nat Neurosci* **14,** 19–21. ISSN: 1546-1726 (Electronic) 1097-6256 (Linking) (2011).

5. Borner, G. H. *et al.* Fractionation Profiling: A Fast and Versatile Approach for Mapping Vesicle Proteomes and Protein-Protein Interactions. *Mol Biol Cell* **25,** 3178–94. ISSN: 1939-4586 (Electronic) 1059-1524 (Linking) (2014).

6. Breckels, L. M. *et al.* Learning from Heterogeneous Data Sources: An Application in Spatial Proteomics. *PLoS Comput Biol* **12,** e1004920. ISSN: 1553-7358 (Electronic) 1553-734X (Linking) (2016).

7. Carney, K. E. *et al.* Proteomic Analysis of Gliosomes from Mouse Brain: Identification and Investigation of Glial Membrane Proteins. *J Proteome Res* **13,** 5918–27. ISSN: 1535-3907 (Electronic) 1535-3893 (Linking) (2014).

8. Chen, N., Koopmans, F., *et al.* Interaction Proteomics of Canonical Caspr2 (CNTNAP2) Reveals the Presence of Two Caspr2 Isoforms with Overlapping Interactomes. *Biochim Biophys Acta* **1854,** 827–33. ISSN: 0006-3002 (Print) 0006-3002 (Linking) (2015).

9. Chen, N., Pandya, N. J., *et al.* Interaction Proteomics Reveals Brain Region-Specific AMPA Receptor Complexes. *J Proteome Res* **13,** 5695–706. ISSN: 1535-3907 (Electronic) 1535-3893 (Linking) (2014).

10. Constals, A. *et al.* Glutamate-Induced AMPA Receptor Desensitization Increases Their Mobility and Modulates Short-Term Plasticity through Unbinding from Stargazin. *Neuron* **85,** 787–803. ISSN: 1097-4199 (Electronic) 0896-6273 (Linking) (2015).

11. Cox, J. & Mann, M. MaxQuant Enables High Peptide Identification Rates, Individualized p.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification. *Nature Biotechnology* **26,** 1367–1372 (Dec. 2008).

12. Distler, U. *et al.* In-Depth Protein Profiling of the Postsynaptic Density from Mouse Hippocampus Using Data-Independent Acquisition Proteomics. *Proteomics* **14,** 2607–13. ISSN: 1615-9861 (Electronic) 1615-9853 (Linking) (2014).

13. Dosemeci, A. *et al.* Composition of the Synaptic PSD-95 Complex. *Mol Cell Proteomics* **6,** 1749–60. ISSN: 1535-9476 (Print) 1535-9476 (Linking) (2007).

14. Fernandez, E. *et al.* Targeted Tandem Affinity Purification of PSD-95 Recovers Core Postsynaptic Complexes and Schizophrenia Susceptibility Proteins. *Mol Syst Biol* **5,** 269. ISSN: 1744-4292 (Electronic) 1744-4292 (Linking) (2009).

15. Fredriksson, R. *et al.* There Exist at Least 30 Human G-Protein-Coupled Receptors with Long Ser/Thr-Rich N-Termini. *Biochem Biophys Res Commun* **301,** 725–34. ISSN: 0006-291X (Print) 0006-291X (Linking) (2003).

16. Frischknecht, R. *et al.* Brain Extracellular Matrix Affects AMPA Receptor Lateral Mobility and Short-Term Synaptic Plasticity. *Nat Neurosci* **12,** 897–904. ISSN: 1546-1726 (Electronic) 1097-6256 (Linking) (2009).

17. Gratacos-Batlle, E. *et al.* AMPAR Interacting Protein CPT1C Enhances Surface Expression of GluA1-Containing Receptors. *Front Cell Neurosci* **8,** 469. ISSN: 1662-5102 (Linking) (2014).

18. Hamann, J. *et al.* International Union of Basic and Clinical Pharmacology. XCIV. Adhesion g Protein-Coupled Receptors. *Pharmacol Rev* **67,** 338–67. ISSN: 1521-0081 (Electronic) 0031-6997 (Linking) (2015).

19. Jamain, S. *et al.* Common and Rare Variant Analysis in Early-Onset Bipolar Disorder Vulnerability. *PLoS One* **9,** e104326. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking) (2014).

20. Jeyifous, O. *et al.* SAP97 and CASK Mediate Sorting of NMDA Receptors through a Previously Unknown Secretory Pathway. *Nat Neurosci* **12,** 1011–9. ISSN: 1546-1726 (Electronic) 1097-6256 (Linking) (2009).

21. Kapitein, L. C. & Hoogenraad, C. C. Building the Neuronal Microtubule Cytoskeleton. *Neuron* **87,** 492–506. ISSN: 1097-4199 (Electronic) 0896-6273 (Linking) (2015).

22. Lagerstrom, M. C. *et al.* The Evolutionary History and Tissue Mapping of GPR123: Specific CNS Expression Pattern Predominantly in Thalamic Nuclei and Regions Containing Large Pyramidal Cells. *J Neurochem* **100,** 1129–42. ISSN: 0022-3042 (Print) 0022-3042 (Linking) (2007).

23. Li, K. W. *et al.* Proteomics Analysis of Rat Brain Postsynaptic Density. Implications of the Diverse Protein Functional Groups for the Integration of Synaptic Physiology. *J Biol Chem* **279,** 987–1002. ISSN: 0021-9258 (Print) 0021-9258 (Linking) (2004).

24. Li, K. *et al.* Organelle Proteomics of Rat Synaptic Proteins: Correlation-Profiling by Isotope-Coded Affinity Tagging in Conjunction with Liquid Chromatography-Tandem Mass Spectrometry to Reveal Post-Synaptic Density Specific Proteins. *J Proteome Res* **4,** 725–33. ISSN: 1535-3893 (Print) 1535-3893 (Linking) (2005).

25. Lund-Johansen, F. *et al.* MetaMass, a Tool for Meta-Analysis of Subcellular Proteomics Data. *Nat Methods* **13,** 837–40. ISSN: 1548-7105 (Electronic) 1548-7091 (Linking) (2016).

26. Milovanovic, D. & Jahn, R. Organization and Dynamics of SNARE Proteins in the Presynaptic Membrane. *Front Physiol* **6,** 89. ISSN: 1664-042X (Linking) (2015).

27. Pandya, N. J. *et al.* Group 1 Metabotropic Glutamate Receptors 1 and 5 Form a Protein Complex in Mouse Hippocampus and Cortex. *Proteomics* **16,** 2698–2705. ISSN: 1615-9861 (Electronic) 1615-9853 (Linking) (2016).

28. Piehowski, P. D. *et al.* Sources of Technical Variability in Quantitative LC-MS Proteomics: Human Brain Tissue Sample Analysis. *J. Proteome Res.* **12,** 2128–2137 (May 2013).

29. Sans, N. *et al.* Synapse-Associated Protein 97 Selectively Associates with a Subset of AMPA Receptors Early in Their Biosynthetic Pathway. *J Neurosci* **21,** 7506–16. ISSN: 1529-2401 (Electronic) 0270-6474 (Linking) (2001).

30. Schindelin, J. *et al.* Fiji: An Open-Source Platform for Biological-Image Analysis. *Nat Methods* **9,** 676–82. ISSN: 1548-7105 (Electronic) 1548-7091 (Linking) (2012).

31. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 Years of Image Analysis. *Nat Methods* **9,** 671–5. ISSN: 1548-7105 (Electronic) 1548-7091 (Linking) (2012).

32. Schwenk, J. *et al.* Regional Diversity and Developmental Dynamics of the AMPA-Receptor Proteome in the Mammalian Brain. *Neuron* **84,** 41–54. ISSN: 1097-4199 (Electronic) 0896-6273 (Linking) (2014).

33. Shaikh, S. A. *et al.* Stargazin Modulation of AMPA Receptors. *Cell Rep* **17,** 328–335. ISSN: 2211-1247 (Electronic) (2016).

34. Sheng, M. & Kim, E. The Postsynaptic Organization of Synapses. *Cold Spring Harb Perspect Biol* **3.** ISSN: 1943-0264 (Electronic) 1943-0264 (Linking) (2011).

35. Sialana, F. J. *et al.* Mass Spectrometric Analysis of Synaptosomal Membrane Preparations for the Determination of Brain Receptors, Transporters and Channels. *Proteomics* **16,** 2911–2920. ISSN: 1615-9861 (Electronic) 1615-9853 (Linking) (2016).

36. Sudhof, T. C. The Presynaptic Active Zone. *Neuron* **75,** 11–25. ISSN: 1097-4199 (Electronic) 0896-6273 (Linking) (2012).

37. Takamori, S. *et al.* Molecular Anatomy of a Trafficking Organelle. *Cell* **127,** 831–46. ISSN: 0092-8674 (Print) 0092-8674 (Linking) (2006).

38. Vizcaino, J. A. *et al.* 2016 Update of the PRIDE Database and Its Related Tools. *Nucleic Acids Res* **44,** D447–56. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (2016).

39. Volknandt, W. & Karas, M. Proteomic Analysis of the Presynaptic Active Zone. *Exp Brain Res* **217,** 449–61. ISSN: 1432-1106 (Electronic) 0014-4819 (Linking) (2012).

40. von Engelhardt, J. *et al.* CKAMP44: A Brain-Specific Protein Attenuating Short-Term Synaptic Plasticity in the Dentate Gyrus. *Science* **327,** 1518–22. ISSN: 1095-9203 (Electronic) 0036-8075 (Linking) (2010).

41. Walch, L. Emerging Role of the Scaffolding Protein Dlg1 in Vesicle Trafficking. *Traffic* **14,** 964–73. ISSN: 1600-0854 (Electronic) 1398-9219 (Linking) (2013).

42. Yamada, K. *et al.* Identification and Characterization of Splicing Variants of PLEKHA5 (Plekha5) during Brain Development. *Gene* **492,** 270–5. ISSN: 1879-0038 (Electronic) 0378-1119 (Linking) (2012).

43. Yamasaki, M. *et al.* TARP Gamma-2 and Gamma-8 Differentially Control AMPAR Density across Schaffer Collateral/Commissural Synapses in the Hippocampal CA1 Area. *J Neurosci* **36,** 4296–312. ISSN: 1529-2401 (Electronic) 0270-6474 (Linking) (2016).

44. Yun-Hong, Y. *et al.* A Study of the Spatial Protein Organization of the Post-synaptic Density Isolated from Porcine Cerebral Cortex and Cerebellum. *Mol Cell Proteomics* **10,** M110 007138. ISSN: 1535-9484 (Electronic) 1535-9476 (Linking) (2011).

45. Zhu, J., Shang, Y. & Zhang, M. Mechanistic Basis of MAGUK-Organized Complexes in Synaptic Development and Signalling. *Nat Rev Neurosci* **17,** 209–23. ISSN: 1471-0048 (Electronic) 1471-003X (Linking) (2016).

5

Comparative hippocampal synaptic proteomes of rodents and primates: differences in neuroplasticity-related proteins

*Frank Koopmans[1], *Nikhil J. Pandya[1], Sigrid K. Franke[12], Ingrid H.C.M.H. Phillippens[2], Iryna Paliukhovich[1], Ka Wan Li[1], August B. Smit[1]

*Co-first authors

[1]Department of Molecular and Cellular Neurobiology, Center for Neurogenomics Cognitive Research, VU University, Amsterdam, The Netherlands
[2]Biomedical Primate Research Centre, Rijswijk, The Netherlands

Key to the human brain's unique capacities are a myriad of neural cell types, specialized molecular expression signatures and complex patterns of neuronal connectivity. Neurons in the human brain communicate via well over a quadrillion synapses. Their specific contribution might be key to the dynamic activity patterns that underlie primate-specific cognitive function. Recently, functional differences were described in transmission capabilities of human and rat synapses. To test whether unique expression signatures of synaptic proteins are at the basis of this, we performed a quantitative analysis of the hippocampal synaptic proteome of four mammalian species, two primates, human and marmoset, and two rodents, rat and mouse. Abundance differences down to 1.15-fold at an FDR-corrected p-value of 0.005 were reliably detected using SWATH mass spectrometry. The high measurement accuracy of SWATH allowed the detection of a large group of differentially expressed proteins between individual species and rodent versus primate. Differentially expressed proteins between rodent and primate were found highly enriched for plasticity-related proteins.

This chapter has been published in *Frontiers in molecular neuroscience 11 (2018)*

Introduction

The human brain's unique cognitive capacity is probably not only derived from its absolute or relative size, or even its number of neurons and glia, but also involves increased diversity of molecular expression signatures, neural cell types, and typical spatial and temporal development leading to expanded and/or more complex patterns of neuronal connectivity. Humans have specialized neuronal connections and a myriad of neuronal cell types which communicate via an estimated well over quadrillion synapses in the human central nervous system (Silbereis *et al.,* 2016). Specific synaptic connections form the core components of neural circuits and networks, collectively referred to as the connectome (van den Heuvel *et al.,* 2016), and their contribution might be key to the dynamic activity patterns that underlie species-specific cognitive function (Markov *et al.,* 2013; Mesulam, 2000; van den Heuvel *et al.,* 2016).

The mammalian brain is capable of processing information in parallel, at high speed and in a highly adaptive manner. These features are largely governed by fast transmission in highly plastic excitatory glutamatergic synapses. Over the years, many proteins of the synapse, their subcellular enrichment and molecular organization into functional entities has become apparent (Chua, 2014; Pandya *et al.,* 2017). Examples of these functional entities are the resident proteins of the presynaptic vesicle, proteins of the fusion and release machinery (Jahn & Fasshauer, 2012) or smaller functional units, such as those associated to synaptic calcium channels (Muller *et al.,* 2010) or glutamate receptors (Chen *et al.,* 2014; Schwenk *et al.,* 2012). Basic synaptic features, such as vesicle release and receptor-mediated signal transduction, are largely carried by evolutionary strongly sequence-conserved proteins (Bayes *et al.,* 2017). Previous studies have indicated changes in the components of the glutamatergic signaling pathway during primate brain evolution in terms of gene expression, protein expression, and promoter sequence changes (Muntane *et al.,* 2015). An outstanding question however is whether the levels of synaptic proteins that underlie the stoichiometries of protein-protein interactions and govern their function in molecular assemblies, have remained conserved.

Intriguingly, recent findings show that human and mouse synapses do not differ in aspects of basic transmission, but drastically differ functionally in the capability to confer high frequency signals (Testa-Silva, Verhoog, Linaro, *et al.,* 2014). Features of synaptic plasticity may also be differently organized, such as observed in spike time dependent plasticity comparing human and rat synapses (Testa-Silva, Verhoog, Goriounova, *et al.,* 2010). Therefore, in this study we investigated the expression signature of the mammalian hippocampal synaptic proteome of rodents; mouse and rat, and primates; marmoset and human. First, we investigated how the expression of synaptic proteins has evolved between these species. This is relevant as alterations in synaptic function are carried by the synaptic protein interaction network and are likely caused by the underlying changes in expression of proteins. Secondly, we investigated whether evolutionary dictated expression differences might relate to plasticity features of synapses. The hippocampus was selected for its well-known role in learning and memory, and the well-described occurrence of correlated synaptic plasticity features, apparent in long-term potentiation (LTP) or long-term depression (LTD) (Cooper & Lowenstein, 2003). Of technical importance, the hippocampus is a neuro-anatomically distinct structure that can be dissected in a reproducible way from the brain of different mammalian species, giving credence to the comparative analysis of this study (Spijker, 2011).

Proteomics analysis was performed using SWATH mass spectrometry (a type of Data Independent Acquisition, DIA), which has been developed to allow for a complete recording of all fragment ions of all (detectable) peptides in a given sample (Gillet *et al.,* 2012). Key advantages of SWATH are a strong reduction in missing values and lowered Coefficient of Variation between (replicate) measurements compared to traditional shotgun proteomics (Bruderer *et al.,* 2015; Koopmans *et al.,* 2018). SWATH analysis enabled a sensitive comparative analysis of the hippocampal synaptic proteome of four species. We first delineated bona fide pre- and postsynaptic proteins and determined their abundance in known protein complexes. In these, we discriminated synaptic substructures and functionalities, such as elements of the postsynapse, e.g., the postsynaptic density, and the presynaptic release machinery. This analysis revealed that within the inter-species conserved synaptic proteome distinct ratiometric differences are apparent, which are species and/or order specific. Proteins involved in this are enriched for synaptic plasticity function suggesting that selective expression differences subserve plasticity and cognitive function.

Materials and Methods

Animal and human tissue use

The use of rodent brain material was approved by the animal ethics committee of the Vrije Universiteit Amsterdam. The use of marmoset brain material was approved by the Biomedical Primate Research Centre (BPRC) ethics committee before the start of experiments, according to Dutch law. Human hippocampus brain samples with donor consent were obtained from, and used according to the guidelines of, the Dutch Brain Bank.

Dissection of hippocampal tissue

Whole hippocampus was dissected from male mouse brain (C57B6/J; Charles River, France) and from male rats (Wistar; Harlan, The Netherlands) (Table 5.1). Hippocampus of the common marmoset (Callithrix jacchus; Biomedical Primate Research Centre, Rijswijk, The Netherlands) was taken between Bregma -2.00 and +1.50. Human postmortem brain tissue from individuals without neurological disorders was obtained from the Netherlands Brain Bank (NBB), Netherlands Institute for Neuroscience, Amsterdam (Table 1). All brain tissue has been collected from donors with written informed consent for brain autopsy, and approval to use this tissue for research purposes has been obtained by the NBB. Approximately 50mg of tissue was isolated from fresh frozen human hippocampal brain tissue by making slices of 20μm using a cryostat. Brain slices include all hippocampal subregions and a small part of the temporal lobe. The tissue was collected in pre-weighed and cooled Eppendorf tubes. All tissues dissected were stored at -80 °C until later use.

Postmortem delay times of human tissue 4-7h, marmoset < 20 min., rodents <10 min.

Sample	Sex	Age at Dissection	Tissue wet weight (mg)
Mouse-1	M	50 Weeks	47
Mouse-2	M	50 Weeks	43
Mouse-3	M	50 Weeks	35
Mouse-4	M	50 Weeks	46
Mouse-5	M	50 Weeks	45
Rat-1	M	50 Weeks	160
Rat-2	M	50 Weeks	159
Rat-3	M	50 Weeks	148
Rat-4	M	50 Weeks	174
Rat-5	M	50 Weeks	175
Marmoset-1	F	5,6 years	58
Marmoset-2	M	3,2 years	88
Marmoset-3	M	7 years	44
Marmoset-4	M	3,1 years	58
Marmoset-5	F	3,2 years	45
Marmoset-6	F	2,7 years	65
Human-1	F	50-55 years	62
Human-2	M	56-60 years	45
Human-3	M	50-55 years	42
Human-4	M	46-50 years	52
Human-5	M	50-55 years	48

Table 5.1: Overview of all samples used in this study.

Synaptosome isolation

Synaptosomes were isolated as described previously (Pandya *et al.,* 2017). To correct for differences in amount of input material between different species, we used 50 µl of homogenization buffer per mg of tissue to ensure that the homogenization conditions were identical between species. Post-homogenization, the samples were spun at 1000 x g for 10 min. and the supernatant was loaded on a sucrose gradient of 1.2/0.85M followed by ultracentrifugation at 100,000 x g for 2 h. Synaptosomes were collected at the interface of 1.2/0.85M, mixed with 5 ml homogenization buffer and centrifuged at 20,000 x g for 30 min to obtain the synaptosomal pellets, which were stored at -80 °C prior to the FASP procedure.

FASP in-solution digestion of proteins

Samples were digested using the FASP in-solution digestion protocol (Wisniewski *et al.,* 2009) with some modifications according to (Pandya *et al.,* 2017). 10 µg of synaptosomes from each sample was incubated with 75 µl 2% SDS 1 uL 50 mM Tris (2-carboxyethyl)phosphine (TCEP) reducing agent at 55 °C for 1 hour at 900 rpm.

Next, the sample was incubated with 0.5 uL 200 mM methyl methanethiosulfonate (MMTS) for 15 min at RT with shaking, after which samples were transferred to YM-30 filters (Microcon®, Millipore) after addition of 200 µl 8 M Urea in Tris buffer (pH 8.8). Samples were washed with 8M Urea in Tris buffer 5 times by spinning at 14,000 x g for 10 min each followed by 4 washes with 50 mM NH4HCO3. Finally, the samples were incubated with 100 µl of Trypsin overnight in a humidified chamber at 37 °C for 12 h. Digested peptides were eluted from the filter with 0.1% acetic acid. The peptides in solution were dried using a speedvac and stored at -20 °C prior to LC-MS analysis.

SCX fractionation

Peptides obtained from 100 µg of synaptosomes following the FASP procedure were fractionated using strong cation-exchange chromatography as described previously (Gonzalez-Lozano *et al.,* 2016). Peptide samples were loaded onto a 4.6 x 100 mm polysulfoethyl A column (PolyLC) and separated using a non-linear gradient of 60 min at 200 µl/ min solvent A (10mM KH2PO4, 20 % acetonitrile, pH 2.9) and solvent B (solvent A + 500 mM KCl). From 0-10 min, flow of 100% solvent A, from 10- 35 min solvent B was increased to 65%. In the next 5 min, solvent B was increased to 100% followed by 8 min of wash with 100% solvent A. In total 40 fractions of 200 µl each were collected. Fractions were pooled in the following manner: 16-20 min (Fraction 1), 21-22 min (Fraction 2), 23-24 min (Fraction 3), 25-26 min (Fraction 4), 27-28 min (Fraction 5), 28-38 min (Fraction 6). Fraction 6 was desalted using Oasis column prior to LC-MS/MS DDA analysis.

Micro-LC and data-dependent data acquisition mass spectrometry of SCX fractions

Peptides were analyzed by micro LC MS/MS using an Ultimate 3000 LC system (Dionex, Thermo Scientific) coupled to the TripleTOF 5600 mass spectrometer (Sciex). Peptides were trapped on a 5 mm Pepmap 100 C18 column (300 µm i.d., 5µm particle size, Dionex) and fractionated on a 200 mm Alltima C18 column (300 µm i.d., 3 µm particle size). The acetonitrile concentration in the mobile phase was increased from 5 to 18% in 88 min, to 25% at 98 min, 40% at 108 min and to 90% in 2 min, at a flow rate of 5 µL/min. The eluted peptides were electro-sprayed into the TripleTOF MS. The micro-spray needle voltage was set to 5500V. The mass spectrometer was operated in a data-dependent mode with a single MS full scan (m/z 350−1250, 150 msec) followed by a top 25 MS/MS (m/z 200- 1800, 150 msec) at high sensitivity mode in UNIT resolution, precursor ion > 150 counts/s, charge state from +2 to +5) with an exclusion time of 16 sec once the peptide was fragmented. Ions were fragmented in the collision cell using rolling collision energy, and a spread energy of 5eV.

Micro-LC and SWATH mass spectrometry

The conditions used for LC in SWATH MS-based experiments were the same as those of the DDA experiments. SWATH experiments consisted of a parent ion scan of 150 msec followed by SWATH window of 8 Da with scan time of 80 msec, and stepped through the mass range between 450-770 m/z. The total cycle time was about 3.2 sec, which yielded in general 9-10 measurement points across a typical peptide with an elution time of 30 sec. The collision energy for each window was determined based on the appropriate collision energy for a 2+ ion, centered upon the window with a spread of 15 eV. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (Vizcaino *et al.,* 2016) partner repository with the dataset identifier PXD009251.

Analysis of data-dependent acquisition mass spectrometry

LC-MS data measured in DDA mode was analyzed using MaxQuant 1.5.2.8 (Cox & Mann, 2008). An initial search using a 0.07 Da peptide mass tolerance was followed by a correction of systematic mass errors. The calibrated data was then subjected to the main search with a 0.006 Da peptide mass tolerance. The minimum peptide length was set to 6, with at most two miss-cleavages allowed. Methionine oxidation and N-terminal acetylation were set as variable modifications with cysteine beta-methylthiolation set as fixed modification.

MS/MS spectra were searched against the human proteome using the UniProt (release February 2016) FASTA database that includes both reviewed (Swiss-Prot) and unreviewed (TrEMBL) records and both canonical and isoform sequences. The Biognosys iRT FASTA database was also included in order to ensure that iRT peptides were included in the search results, as these were used to normalize retention times in downstream analysis. For both peptide and protein identification a false discovery rate of 0.01 was set. The computation of iBAQ (Schwanhausser *et al.,* 2011) in-silico estimated absolute protein abundances (by dividing the protein intensity by the number of theoretically observable tryptic peptides) was enabled.

SWATH data extraction and analysis

A spectral library was made from the MaxQuant analysis of DDA data using Spectronaut 6.0.6880.14 (Bruderer *et al.,* 2015). The Q-value threshold for peptides imported from the MaxQuant msms.txt output table was set to 0.01, all other settings were left to default.

Next, Spectronaut was used to extract peptide abundances from the raw SWATH data. The retention time prediction type was set to dynamic iRT and profiling peak refinement was enabled. Finally, across-run normalization based on total peak areas was performed by Spectronaut. Peptide abundances were exported as a Spectronaut report and further processed using the R language for statistical computation (Gillet *et al.,* 2012), in which we considered each unique precursor as a peptide (e.g., the same peptide sequence observed with distinct modifications or charge was considered a distinct peptide).

Spectronaut's fragment group Q-values were used to discriminate high confidence peptides. For the validation of SWATH capabilities using three technical replicates, all peptides with a Q-value higher than 10-3 in any sample were removed. For the pairwise comparison between (groups of) species, we used the subset of peptides that are present in all species and quantified with high confidence in either group. The former condition was reached by comparing the in-silico digestion of the respective FASTA database of all species. The latter is formalized as having a Q-value smaller than, or equal to, 10-4 over all samples (while allowing for one outlier within each species) in either group.

Protein abundances were computed by summation of the normalized peak area of their respective peptides. Peptides that map ambiguously to multiple genes in the spectral library were discarded. Finally, the protein abundance matrix was Loess normalized using the normalizeCyclicLoess function from the limma R package (Smyth *et al.,* 2005), which was set to 'fast' and iterations were set to 10.

Synaptic plasticity proteins listed in Supplementary Table S2 of (Rao-Ruiz *et al.,* 2015) were used for the statistical analysis of synaptic plasticity within our SWATH data. Overlap between this set and SWATH quantified proteins from this study can be found in the last column of Supplementary Tables 2, 3 and 4.

Differential Abundance Analysis

Differential abundance analysis between (groups of) species was performed on log transformed protein abundances. Empirical Bayes moderated t-statistics with multiple testing correction by False Discovery Rate (FDR), as implemented by the eBayes and topTable functions from the limma R package, was used. An FDR adjusted p-value threshold of 0.005 was used to discriminate proteins of interest after differential abundance analysis. Gene Ontology enrichment tests were performed using the PANTHER Overrepresentation Test (version 13.1) with the total set of SWATH quantified proteins as the background set (Mi *et al.,* 2017).

Results

Brain tissue was obtained taking age and gender matching into account. Human subjects were on average 53.1 ± 3.7 years of age and animal groups were chosen in line with this, e.g. with mouse and rat 50 weeks of age and marmoset on average 4 years of age (Table 5.1). Whole hippocampus from mouse, rat and marmoset was dissected. To enable direct comparison of the rodent and marmoset hippocampus samples with the much larger human hippocampus, we made 20µm cross section slices of human hippocampus that included all sub-regions. Synaptosome fractions were prepared from independent biological replicates (n=6 for marmoset and n=5 for the other species), as described previously (Pandya *et al.,* 2017).

Spectral library for SWATH analysis

For SWATH analysis, a spectral library that contains a fingerprint (e.g.; m/z and retention time) of each peptide is used for targeted data extraction to obtain peptide abundancy values. A key challenge in building the spectral library is the inclusion of low abundant proteins as the probability of protein detection in discovery proteomics is correlated with their abundance. Therefore, an extensive spectral library was prepared from the synaptosome fraction of each species using Data Dependent Analysis (DDA) proteomics and extensive SCX fractionation to optimize coverage of low abundant proteins.

Since the SWATH approach compares abundance levels of the exact same peptide between samples, we used the DDA data to create a spectral library of detectable peptides in the human synaptic proteome. This was later used to compare abundance values of human synaptic proteins among species. The rationale for using DDA of all species to identify the human synaptic proteome, and not only human samples, lies in the assumption that protein levels may vary among species (for instance, proteins more abundant in mouse have better detection probability in mouse). Using MaxQuant, 681626 MS/MS spectra were searched against the UniProt human proteome, resulting in the identification of 29710 unique peptide sequences and 3937 protein groups. 166 protein groups that contained proteins from different genes were considered ambiguous and removed from the dataset. The dynamic range of synaptic protein levels was investigated using iBAQ abundances (an in-silico estimation of absolute protein abundance using peptide intensity values) obtained from the DDA data. Protein groups that mapped to the same gene were merged by summation of their iBAQ abundances, yielding a final dataset containing absolute abundance levels for in total 3630 proteins (Supplementary Table 1).

From the thousands of identified proteins we confidently assigned 336 proteins to the pre- or postsynapse, or are synaptic without a definite pre- or postsynaptic localization (Figure 5.1). These core synaptic proteins were derived from data from previous analyses (Chua *et al.,* 2010; Weingarten *et al.,* 2014; Wilhelm *et al.,* 2014). These 336 proteins span close to 6 orders of magnitude difference in iBAQ abundance (Supplementary Table 1). In particular, presynaptic proteins belonging to the synaptic vesicle and those serving the actual vesicle to membrane fusion event were highly abundant. Much lower abundant were proteins typical of the postsynapse, such as postsynaptic receptors and their auxiliary subunits.

SWATH workflow and quality control

The SWATH mass spectrometry data were processed in Spectronaut using our synapse spectral library and the resulting qualitative and quantitative peptide data were processed using the R language for statistical computation.

We first assessed the performance of the SWATH proteomics pipeline by determining the technical variation observed in a triplicate measurement of a single mouse synaptosome preparation. Using only peptides identified in the mouse synaptosome samples from the spectral library that were quantified confidently in all three SWATH technical replicates, we calculated protein abundances and applied Loess normalization. The median Coefficient of Variation (CoV) for 5144 peptides was 4% (Supplementary Figure S1A) and the median CoV of 1831 quantified proteins was 3%, while 75% and 99.65% of these were quantified with < 5% and < 11% CoV, respectively (Supplementary Figure S1B). The reproducibility of this workflow was further verified by calculating the correlation of abundance values between sample pairs, yielding an R2 of 0.997, illustrating excellent technical reproducibility (Supplementary Figure S1C).

Pairwise comparison between (sets of) species was performed on the subset of peptides detected with high confidence in either group of samples. For instance, when comparing rodent and primate synaptosome fractions, we used the set of

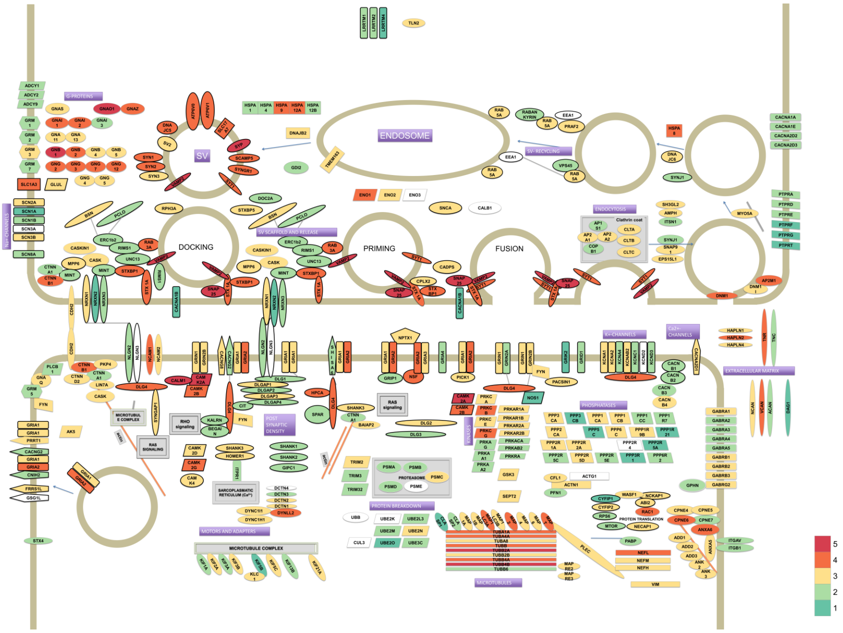

Figure 5.1: Model of the synaptic proteome featuring 336 proteins selected from literature. Protein colors reflect their estimated abundances in the synaptosome fraction of human hippocampus. The iBAQ values (which were approximately log-normal distributed) were log10 transformed to improve visualization of differences. The color legend for depicted log10 iBAQ values is shown in the bottom-right; low abundant proteins are depicted in green, medium in yellow and high abundant proteins are shown in red.

high quality peptides that were found in all rodent and/or all primate samples. Furthermore, we disregarded peptides that mapped ambiguously to multiple genes and peptides that were not conserved in all species of the pairwise comparison.

Protein abundances were computed by the summation of their respective peptide peak areas and Loess normalized. As a result, we confidently quantified 6021 peptides covering 1642 proteins over all samples when comparing rodents with primates, and 8243 peptides covering 2109 proteins when comparing mouse with human.

Finally, the SWATH data for biological replicates in each pairwise comparison of species was subjected to quality control in order to validate reproducibility. The coefficient of determination between pairs of biological replicates was at least 0.97 for all species in all the pairwise comparisons (Supplementary Figure S2). Unsupervised hierarchical clustering of the log-transformed protein abundances showed grouping of samples from the same species (Supplementary Figure S3A). When comparing rodents with primates, the 1642 proteins compared over all species, show a median CoV 8%, 8%, 12% and 9% for mouse, rat, marmoset and human, respectively (Supplementary Figure S3B). The mouse versus human comparison yielded more proteins, i.e., 2109, because within each group the stringent quality criteria for peptides were applied to fewer samples/species. The median protein CoV in this comparison was found to be 10% for both species (Supplementary Figure S3C).

As an internal control for accurate measurements, we first analyzed the ratios of proteins within well-established ribosome protein complexes. Proteins residing in these functional complexes are likely evolutionary well conserved and are predicted to have fixed ratios, which should result in similar stoichiometries when comparing species. Supplementary Figure S5A, B shows the relative protein abundances in each sample, in which we see differences between species for ribosomal proteins. However, the vast majority of values seem to move in a similar pattern among species. And indeed, visualizing the protein-protein correlations among ribosomal proteins (Supplementary Figure S5C, D) indicates that the relative abundance values of proteins in the ribosomal complex are tightly coupled for the vast majority of all 48 quantified proteins from the small and large subunits of the ribosome, internally validating our quantitative approach. While the relative abundance values obtained from label-free proteomics are not accurate estimates of true copy numbers (absolute abundances), we can infer that their stoichiometry (ratios of abundances across species) is fixed if up/down shift between species is tightly correlated. The only exceptions were RPS17 (enriched in rat, unlike other subunits) and RPL14 (anti-correlated with most large subunits of the ribosome).

Comparing synaptic proteomes among species

Differential abundance analysis by empirical Bayes moderated t-statistics was used to compare (sets of) species. Comparing rodents with primates resulted in 381 proteins with higher abundance in primates and 398 proteins with higher abundance in rodents, whereas 862 proteins were not significantly different at an FDR adjusted p-value 0.005 (Figure 5.2A, Supplementary Table 2). Amongst the highest differentially expressed proteins between rodents and primates belong, for instance, the cell adhesion molecules NCAM1 and -2, the sodium channel subunit SCN3B, the Annexins ANXA1 and -2 and the isocitrate dehydrogenases (IDH3A and IHD2), the latter of which are proteins with lower and higher expression in primates respectively (Supplementary Table 2).

When comparing mouse with human, we found 644 proteins with higher abundance in human and 663 proteins with higher abundance in mouse (in total 1307 proteins were changed), whereas 800 proteins were not differentially abundant at FDR adjusted p-value 0.005 (Figure 5.2B, Supplementary Table 3). The fold-changes for proteins that were not differentially abundant were similar across all species comparisons (Supplementary Figure S7, Supplementary Tables 2, 3 and 4). Proteins that were statistically significant had much higher fold-changes and varied between species comparisons (eg; differences in mouse vs human were stronger than in mouse vs rat).

Both statistical tests indicated a large number of proteins that were differentially abundant. In particular, the low variation between biological replicates (see quality control, Supplementary Figure S2-3, Supplementary Table 2-3) allowed us to reliably detect fold changes between species, i.e., the lowest fold change for differentially abundant proteins between rodents vs primates and mouse vs human was 1.148 and 1.184, respectively. Obviously, it remains to be seen whether differences that

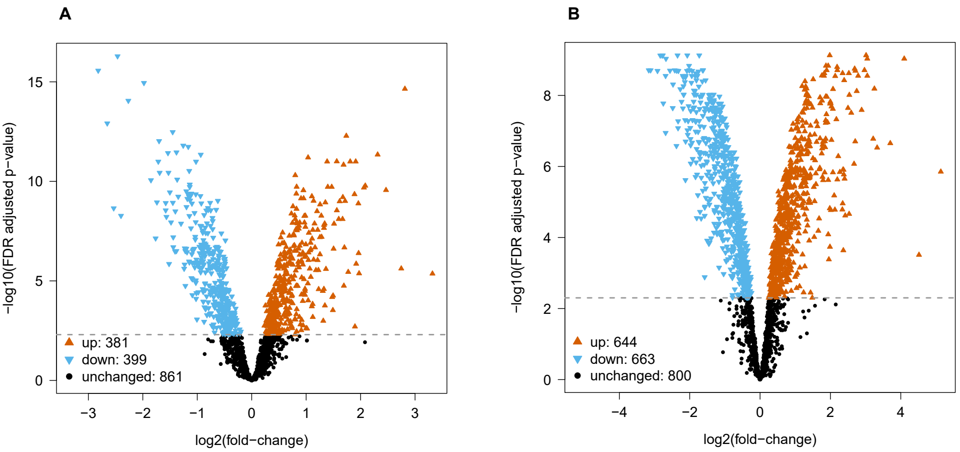

Figure 5.2: Differential abundance analysis for (A) rodent vs primate and (B) mouse vs human. Empirical Bayes moderated t-statistics followed by a FDR adjusted p-value 0.005 cutoff resulted in 399/ 381 proteins with decreased/ increased abundance from rodent to primate, and 663/ 644 proteins with decreased/ increased abundance from mouse to human.

are so small are biologically meaningful. If additional filtering of statistical results by fold-change is desired, we would recommend using the data tables (S. Fig. S7, S. Tables 2-4) for further filtering on fold change.

Overrepresentation analysis of differentially abundant proteins in Figures 5.2A and 5.2B in the Gene Ontology (GO) cellular components, biological processes, molecular functions and PANTHER protein classes yielded no results. Similarly, using subsets of statistically significant hits with large quantitative differences between species (fold-change of at least 2 or 3) yielded no results. Thus, functional annotations of synaptic proteins available in public databases could not explain the many species differences we detected.

However, visualization of proteins differentially expressed in mouse vs human using the synapse model that features 336 proteins (cf. Figure 5.1) suggested interesting expression differences for functionally and structurally related proteins (Supplementary Figure S4). For instance, a downregulation is apparent for synaptic vesicle endocytosis and postsynaptic density proteins while upregulated groups include neurofilaments and extracellular matrix proteins. Given that synaptic proteins of interest in this study are lacking GO annotation coverage at this time we focussed on functionally related groups of proteins that are commonly studied in the synapse field to interpret species differences. We compared protein abundance profiles within functional groups of proteins in the synapse, both in terms of expression levels and correlations thereof.

The presynaptic protein groups

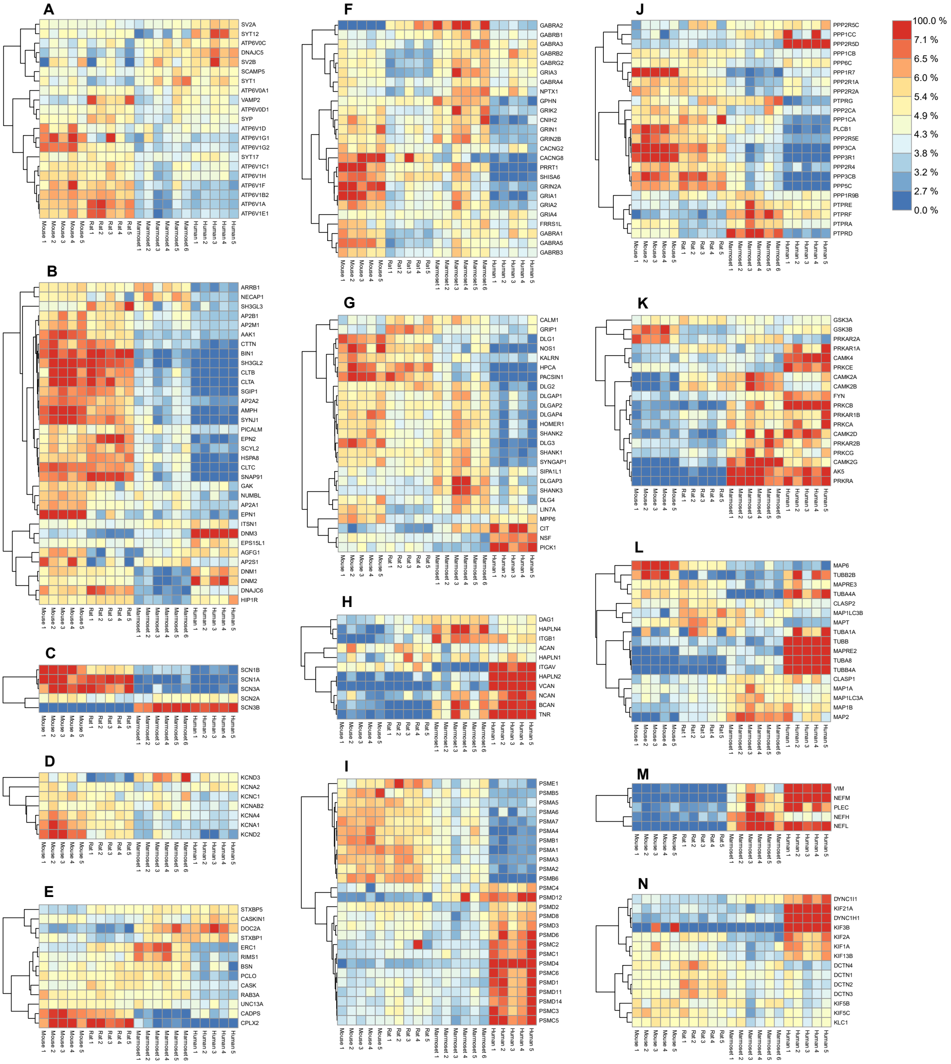

Figure 5.3: Quantified proteins in various functional groups of interest. Abundance values were scaled by their total over all samples to reveal their relative enrichment, if any, between species. The color legend on the top-right shows the protein color gradient from relatively low abundances in blue to relative enrichment in red. (A) Synaptic Vesicle. (B) Endocytosis. (C) Sodium channels. (D) Potassium channels. (E) Presynaptic scaffold. (F) Ligand-gated ion channels and associated proteins. (G) Postsynaptic density. (H) Extracellular matrix. (I) Proteasome. (J) Phosphatases. (K) Kinases. (L) Microtubules. (M) Neurofilaments. (N) Motor proteins. Respective protein-protein Pearson correlation matrices are shown in Supplementary Figure S6.

The synaptic vesicle

Synaptic transmitters are pumped into synaptic vesicles using a proton gradient generated by the vesicular-ATPase. The ATPase is built of the vesicle external V1 domain and the transmembrane V0 domain each consisting of different subunits (ATP6V1- and ATP6V0- subunits, respectively). The v-ATPase shows no differential expression of the proton translocating V0 domain subunits between species. In contrast a higher expression of the ATPase V1 domain subunits is observed in both rodent species (Figure 5.3A). When comparing the expression of ATP6V subunits in mouse and human, the V0 and V1 subunits each show tight co-expression (Supplementary Figure S6A). The stator subunit c, involved in assembly of subunits and regulator of the activity of the v-ATPase, shows a distinct pattern from other subunits.

When inspecting other proteins of the vesicle, three major expression-correlated subgroups can be discerned. Next to the ATP6-ase subgroup, a set of 6 integral synaptic vesicle membrane proteins, and a group containing synaptophysin (SYP) and its interacting protein synaptobrevin (VAMP2) together with two vATP6V0 subunits (Figure 5.3A) is identified. Interestingly, SYP and VAMP2 have been shown previously to interact with the vATP6V0 subunits (Galli *et al.,* 1996). Also, these proteins correlate with the ATP6V1 group. The set of 6 integral vesicle membrane proteins are highly correlated amongst each other and anti-correlated in expression with the ATP6V1 group.

The endocytosis machinery

Core proteins of the synaptic endocytosis machinery (McMahon & Boucrot, 2011) show a differential regulation over species. Notably many endocytosis proteins were abundantly expressed in rodents compared to marmoset and were even lower expressed in human. The dynamins (DNM1-3) are part of a subgroup, of which members are highly expressed in human (Figure effig:ch5fig3B). From this group, the intersectin1 (ITSN1) protein has been investigated regarding its role in membrane fusion and is involved in exocytosis or endocytosis (Gubar *et al.,* 2013). The correlation matrix indicates an anti-correlation with the core endocytosis set (Supplementary Figure S6B). The same holds for the ITSN1-interacting protein EPS15L1. When inspecting the correlation of ITSN1 versus exocytosis proteins, ITSN1 correlates highest with STXBP1 (0.72), STX1A (0.63), SYT1 (0.55), which could suggest a role in exocytosis.

Ion channels

Sodium channels are formed by the major pore forming alpha subunits (SCN1-4A) and the channel modulating beta subunits (SCN1-4B). Four out of five detected subunits in hippocampus have a very strong differential expression profile between rodents and primates (Figure 5.3C, Supplementary Figure S6C). Also, SCN3B and SCN1B, both modulatory subunits of the sodium channels are differentially expressed in rodents and humans. Like sodium channels, K+ channels are also involved in shaping membrane depolarization. Interestingly, like the Na+ channels also the K+ channels show expression differences that relate to neuronal functional differences between rodents and primates (Figure 5.3D, Supplementary Figure S6D).

Presynaptic organization

Organizing proteins of the vesicle release in the presynapse show differential distribution between species. Some of these such as CDPS, CPLX2, DOC2A and STXBP1 show a differential expression between rodents and primates, e.g. with DOC2A high and CPLX2 low in primates (Figure 5.3E). Various presynaptic proteins show the lowest expression in human. A strong positively correlated expression cluster exists for the presynaptic scaffold organizers, Piccolo (PCLO), CASK, Bassoon (BSN), and RIMS1, and the RIMS1 binding protein ERC1. The group of CASKIN1, STXBP1, STXBP5, and DOC2A is anti-correlated with the expression cluster CDPS, UNC13A, complexin2 and Rab3A (Supplementary Figure S6E).

The postsynaptic protein groups

Ligand-gated ion channels

When exploring the ligand gated ion channels, in particular the AMPAR, NMDAR, GABAR and their associated proteins, striking differences in expression patterns between 4 species are observed. In particular, GRIA1, 2, and its auxiliary proteins (Chen *et al.,* 2014; Schwenk *et al.,* 2012) CACNG2, PRRT1, CNIH2 and SHISA6 have low expression in human (Figure 5.3F). This is different for the GABAR subunits for which expression is lowest in rat. A strong expression cluster exists for GABAR α1/β2/γ2, which is considered a typical GABAA receptor subunit composition, and is distinct from GABAR α5/β3 and GABAR α2/β1. Interestingly, GABAR α5/β3 form a genomic subunit cluster (Papadimitriou *et al.,* 2001) and have been described as part of a single functional receptor (Sur *et al.,* 1998) (Supplementary Figure S6F).

Postsynaptic density proteins

A fraction of the postsynaptic density (PSD) proteins show a low abundance in humans compared to rodents and marmoset (Figure 5.3G). There are no PSD proteins that are consistently differentially expressed between rodent and primates. NSF, involved in fusion of AMPARs with the postsynaptic membrane is highest in human. The strongest co-expression cluster is formed by NSF and PICK1, which are well-known interactors and regulators of AMPAR trafficking (Hanley *et al.,* 2002). Also, SIPA1L1, SHANK3, DLGAP3 show co-expression. In agreement with this, SHANK3, DLGAP3 have been shown to interact (Tu *et al.,* 1999), and both proteins bind and recruit SIPA1L1 to synapses with a central coiled-coil region that harbors a leucine zipper motif (Wendholt *et al.,* 2006). The largest correlated cluster contains HOMER1, SHANK1,-2, DLGAPs1,-2, SYNGAP1, DLG2,-3, most of which are known to interact, e.g. Shank1-Homer1 (Tu *et al.,* 1999), SHANK1- DLGAP1 (Im *et al.,* 2003), DLGAP1-DLG4 (Kim *et al.,* 1997). A subgroup is formed by HPCA, NOS1, KALRN, DLG1, PACSIN1 (Figure 5.3J, Supplementary Figure S6G). These proteins are not known to bind each other.

Pan-synaptic protein groups

The extracellular matrix (ECM)

The ECM plays an important role in plasticity processes. Some of the proteins of the ECM are expressed at high levels in human, in particular Versican (VCAN), Hyaluronan link protein HAPLN2 and Tenascin-R, and to some extent Neurocan (NCAN) (Figure 5.3H). These 4 proteins and HAPLN1 form a co-expressed core of the ECM (Supplementary Figure S6H). HAPLN2 and HAPLN4 are known to bind Brevican and Neurocan, respectively (Spicer *et al.,* 2003). Versican was shown to increase expression in humans during progression of Alzheimer's disease (Hondius *et al.,* 2016).

The 26S proteasome

Comparing human and mouse we found that all 10 PSMA-B proteins of the 20S core complex were differentially expressed from the 15 PSMC-D proteins of the 19S regulatory complex (Figure 5.3I, Pearson correlation matrix shown in Supplementary Figure S6I). Although the proteasome is differentially expressed between tissues and during development (Claud *et al.,* 2014), differential expression of 19S and 20S subunits has not been observed previously.

Phosphatases and kinases

Overall, phosphatases are less expressed in primates than in rodents, with the notable exception of PPP2R5d, a phosphatase 2A regulatory subunit B family, which is highly expressed in human (Figure 5.3J). Protein phosphatase 2A is one of the four major Ser/Thr phosphatases, and disruptive mutations in the regulatory subunit PPP2R5d were found causative in prenatal overgrowth and intellectual disability (Loveday *et al.,* 2015). Membrane bound receptor-type tyrosine-protein phosphatases (PTPRs) form a strongly co-expressed group, most abundantly expressed in marmoset, which are anti-correlated with the main group of Ser/Thr phosphatases (Supplementary Figure S6J). The kinases are found higher expressed in primates than in rodents (Figure 5.3K). Interesting individual expression differences exist for some kinases. PRKAR2A, the regulatory subunit of PKA, is low expressed in human, whereas its paralog PRKAR2B is highly expressed. Remarkable expression differences exist between mouse and human for CAMK2A/B, and PRKCB.

The cytoskeleton

In particular, the tubulins TUBA8 and TUBB4A are highly expressed in human (Figure 5.3L) and strongly co-expressed between species (Supplementary Figure S6M). MAPT and MAP6 both bind to stable microtubules and they form a co-expressed pair. Both proteins are well known for affecting cognitive abilities upon changing expression. CLASP1 and 2 are microtubule end-binding proteins (Mimori-Kiyosue *et al.,* 2005), however, they have also been shown to bind to actin and were proposed to link actin to microtubules (Tsvetkov *et al.,* 2007). They show anti-correlated expression suggesting these modulate opposite function with regards to the cytoskeleton species (Supplementary Figure S6M). In contrast to the microtubular network, actin and adhering proteins are not strongly differentially regulated between species. Elements of the filamentous cytoskeleton neurofilaments (NEFL, NEFM) and vimentin

(VIM) are all more abundantly expressed in primates than in rodents. (Figure 5.3M, Supplementary Figure S6M).

Motor proteins

Molecular motors and their adaptors serve transport functions along the cytoskeleton. KIF3b, for membrane organelle transport, and Kif21a, involved in axonal transport, and DYNC1H1, dynein, are all highly expressed in human (Figure 5.3N). Microtubular kinesin motors KIF5b, 5c and their interactor protein KLC1 are strongly co-expressed (Supplementary Figure S6N). The same holds for the entire group of Dynactins (DCTN1-4), viewed as adaptor proteins of Dynein (Urnavicius *et al.,* 2015), a major motor protein of the tubulin-based transport.

Plasticity-related proteins

Cognitive differences among species may be underlain by the differences in the protein composition of the synapses, and that the plastic changes of hippocampal synapse physiology are considered to be at the basis of learning and memory. We thus tested whether variation in synaptic protein abundance between species might specifically involve proteins that are prone to condition-dependent synaptic plasticity. In particular, we tested the hypothesis whether this set of plasticity proteins is differentially expressed between mouse and humans; following the reasoning that synaptic plasticity might relate to a set of proteins that in recent adapted its expression to improve dynamic response to changes in synaptic stimulation. We previously defined a set of 400 synaptic plasticity proteins in mouse hippocampus, which respond to a learning stimulus with changes in expression, using a quantitative iTRAQ proteomics analysis (Rao-Ruiz *et al.,* 2015). Of these synaptic plasticity proteins in the SWATH interspecies comparison data set, 340 were quantified in the rodent with primate comparison of which 188 proteins were differentially expressed at FDR adjusted p-value 0.005, which is a significant difference (Chi-square p-value 1.3×10-03) among all differentially abundant proteins. Analogous pairwise comparison of mouse with human and rat with human also yielded a significant difference for plasticity proteins (Chi-square p-values 6.0×10-04 and 1.2×10-03, respectively). One might reason that plasticity proteins by nature show more variation in expression and therefore might have an increased likelihood of being differentially abundant. Therefore, we analyzed the same synaptic plasticity proteins in the SWATH interspecies comparison of mouse with rat and found a much lower significance level at Chi-square p-value 0.012. Finally, we tested plasticity proteins in the marmoset with human comparison, yielding a significant difference (Chi-square p-value 7.0×10-03). When the group of 340 synaptic plasticity genes is assessed in the group mouse, rat, marmoset versus human (Supplementary Table 5), 175 proteins show differential regulation with a FDR corrected p-value <0.01. In this group of proteins some interesting sets of proteins show lower expression in human, e.g., proteins involved in endocytosis (SNAP91, SYNJ1, CLTA/B/C and Pacsin-1), the ionotropic glutamate receptors (GRIA1 and -2) and auxiliary subunits (PRRT1, Shisa6 and CACNG8). Proteins that show higher expression in human include the extracellular matrix components (TNR, BCAN and NCAN). In conclusion, proteins that show stimulus-dependent changes in expression after fear-learning in the mouse hippocampus are showing a differential abundance between rodents and primates and between marmoset and human, whereas comparison within rodent species does not show this.

Discussion

In this study we used SWATH to quantify levels of hippocampal synaptic proteins of four species, the rodents; mouse and rat, and the primates; marmoset and human. We revealed many protein abundance differences between species with in many instances small fold changes. The quantification accuracy of SWATH and low technical variability of the method, allowed us to very accurately quantify even minor (as low as 1.15 fold) differences across species. Furthermore, our SWATH proteomics required building spectral libraries, which serve as catalogs for synaptic proteins in these four species analyzed. One can use these catalogs to generate protein maps of synapses to gain insight in stoichiometries (cf. Fig 1) or visualize differences in levels between species (cf. Fig S4).

In silico approximations of absolute abundances from label-free data, such as iBAQ (Schwanhausser *et al.,* 2011) or SCAMPI (Gerster *et al.,* 2014) are currently not sufficiently accurate to detect small differences in protein abundance expected between species. As distinct peptide sequences have different mass spectrometric properties, such as ionization efficiency, their ion intensity measured by label-free proteomics can differ up to several orders of magnitude even if the absolute amount of these peptides in the input sample is the same. Including peptide standards in the sample to obtain accurate estimates of absolute abundances (Gerber *et al.,* 2003) is not feasible for thousands of unique peptides present in the synaptic proteome of multiple species. However, physical properties may vary amongst different peptides; identical peptides from different samples behave the same. Therefore, in this study, we considered the set of peptide sequences that are identical in all species (within a given comparison of species) and quantified these confidently in all samples. This allowed comparison of the relative abundance of peptides and proteins with the highest accuracy possible. In the current study we compared synaptic fractions obtained from different species. This may involve different efficiencies of isolating the synaptosome fraction, or proteins therein, which cannot be easily corrected for. We did not observe apparent species differences in the amount of protein isolated in the synaptosome fractions. Another aspect of this study is that it inherently incorporates different postmortem delay times, in particular between human versus marmoset and rodent brains. We cannot rule out that this may have introduced differences in the synaptic proteins isolated.

Key to the human brain's unique capacities is probably is not its absolute or relative size, or even its number of neuronal and glia cells. Instead, this likely involves evolved neural cell types, and expanded and/or more complex patterns of neuronal connectivity. However, these specific evolutionary adaptations likely depend on the increased diversity of molecular expression signatures, both in terms of levels and of cell type specific expression.

Previous transcriptome analyses have reported that there are more genes higher expressed in the adult human brain than in the non-human primate brain, and not in other examined tissues (Caceres *et al.,* 2003; J. Gu & X. Gu, 2003; Khaitovich *et al.,* 2004). Also it was found that there is little evidence for accelerated divergence in gene expression in the human brain (Hsieh *et al.,* 2003). Other studies have reported several interesting findings on human-specific differences in the expression of genes involved in metabolism (Babbitt *et al.,* 2010; Uddin *et al.,* 2008) and on genes that are organized into human-specific co-expressed modules (Konopka *et al.,* 2012; Oldham *et al.,* 2006). Human-specific differences in gene expression have also been reported for groups of genes with developmental time-shifts (Somel *et al.,* 2009), miRNA regulation (Hu *et al.,* 2011; Somel *et al.,* 2009), RNA editing (Li *et al.,* 2013), and transcription factor regulation (Liu *et al.,* 2012). Notwithstanding the usefulness of gene expression analysis, a constraint in unambiguous translation of transcript levels to proteins is the multitude of regulatory steps in protein synthesis and breakdown. Furthermore, differences in gene expression are difficult to translate to synaptic protein expression. Typically, synaptic proteins, as studied here, are well known to be regulated by condition-dependent trafficking, local synthesis and target specific breakdown.

One of the outstanding questions regarding species comparison is whether expression differences of distinct proteins or protein groups might be related to functional differences between rodents and primates. We indeed detected a number of synaptic functional groups that show differential abundance in rodents and primates. An interesting group of proteins to consider are the sodium channels, which are formed by the major pore forming alpha subunits (SCN1-4A) and the channel modulating beta subunits (SCN1-4B). Four out of five subunits detected in hippocampus have a very strong differential expression profile between rodents and primates (cf. Fig 3C). As such these may qualify for distinct physiology of human (mammalian) neurons. For instance, we found SCN3B and SCN1B, both modulatory subunits of the sodium channels highly differentially expressed in rodents and humans (cf. Fig 3C). Given the essential role of these channels in the formation and propagation of action potentials, this might be an important observation. Co-expression studies of SCN1B, -2B, and -3B subunits with the SCN2A subunit in HEK293 cells have shown to shift sodium channel activation and inactivation to more positive membrane potentials (Qu *et al.,* 2001). However, SCN3b is unique in causing increased persistent sodium currents. Because persistent sodium currents are thought to amplify summation of synaptic inputs, expression of this subunit would increase the excitability of the expressing neurons to all of their inputs. This might be specifically the case for human neurons. Interestingly, SCN2A and SCN3B are highly co-expressed over the species, suggestive of a co-expressed pair (cf Fig S6C). If so, the not differentially expressed SCN2A channel might be modulated by SCN3B thereby providing a higher offset for neuronal excitation in humans.

Among proteins with differential level we found neurofilament, which is particularly highly expressed in human brain tissue. Synaptosome fractions are only enriched in synaptic proteins and therefore caution is warranted regarding inferring neurofilament presence at the synapse based on biochemical means only. Neurofilament has been regarded as an impurity of the synaptic fraction (e.g., (Matus *et al.,* 1980)), however, recent immuno-EM data shows its presence at striatal postsynaptic sites (Yuan *et al.,* 2015). We therefore consider it a potential component of the hippocampal synapse from which it can be readily isolated (this study, (Yuan *et al.,* 2015)). Deletion of neurofilaments has been shown to affect synaptic long-term potentiation (Yuan *et al.,* 2015) and neurofilament expressing pyramidal neurons in humans may show higher vulnerability to neurodegenerative disease (Perrot *et al.,* 2008).

Importantly, we demonstrated that, among the many proteins with expression differences, a group of plasticity–related proteins follows a global up-regulation in human hippocampal synapse proteome. One might argue that an important aspect of species evolution is the potential to be adaptive, at the molecular and subcellular level displayed as synaptic plasticity. Plasticity mechanisms enable the brain with the feature to rapidly change the efficacy of synapses, allowing these to be used over a wide dynamic range of signal transmission and act as logical operators. In line with this, synaptic plasticity might relate to a set of proteins that evolved rapidly in recent evolution. Changes in protein abundance might reflect dynamic response to changes in synaptic stimulation, and/ or change the impact a protein has in a larger network. As such this set is expected to be dynamic within a species and consequently may show abundance differences between species. To specifically test this we used a plasticity set of proteins as found regulated by a strong fear learning paradigm impacting on the hippocampus (Rao-Ruiz *et al.,* 2015). Using this set of proteins we asked whether its constituents would belong to the rodent-primate conserved or rather the differentially expressed part of the synaptic proteome. The latter was true. This indicates that within the synaptic proteome those proteins of which expression differences maximally evolved during evolution are overrepresented in the plasticity response.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

FK, NP, KWL, ABS planned and designed the experiments. NP and I.Paliukhovich performed experiments. FK performed data analysis. SK and I.Phillippens provided marmoset samples. FK, KWL, NP and ABS wrote the manuscript. All authors reviewed the manuscript.

Funding

This work was partially sponsored by the EU FP7 Health ‘SynSys’ (grant #242167 to A.B.S., and K.W.L.). F.K. was funded from The Netherlands Organization for Scientific Research (NWO) complexity project 645.000.003. N.P. was funded by EU International training site ‘Braintrain’ (MCITN-CT-2009-031910).

Data Availability Statement

The datasets generated for this study can have been deposited to the ProteomeXchange Consortium via the PRIDE (Vizcaino *et al.,* 2016) partner repository with the dataset identifier PXD009251.

References

1. Babbitt, C. C. *et al.* Both Noncoding and Protein-Coding RNAs Contribute to Gene Expression Evolution in the Primate Brain. *Genome Biol Evol* **2,** 67–79. ISSN: 1759-6653 (Electronic) 1759-6653 (Linking) (2010).
2. Bayes, A. *et al.* Evolution of Complexity in the Zebrafish Synapse Proteome. *Nat Commun* **8,** 14613. ISSN: 2041-1723 (Electronic) 2041-1723 (Linking) (2017).
3. Bruderer, R. *et al.* Extending the Limits of Quantitative Proteome Profiling with Data-Independent Acquisition and Application to Acetaminophen-Treated Three-Dimensional Liver Microtissues. *Mol Cell Proteomics* **14,** 1400–10. ISSN: 1535-9484 (Electronic) 1535-9476 (Linking) (2015).
4. Caceres, M. *et al.* Elevated Gene Expression Levels Distinguish Human from Non-Human Primate Brains. *Proc Natl Acad Sci U S A* **100,** 13030–5. ISSN: 0027-8424 (Print) 0027-8424 (Linking) (2003).
5. Chen, N. *et al.* Interaction Proteomics Reveals Brain Region-Specific AMPA Receptor Complexes. *J Proteome Res* **13,** 5695–706. ISSN: 1535-3907 (Electronic) 1535-3893 (Linking) (2014).
6. Chua, J. J. Macromolecular Complexes at Active Zones: Integrated Nano-Machineries for Neurotransmitter Release. *Cell Mol Life Sci* **71,** 3903–16. ISSN: 1420-9071 (Electronic) 1420-682X (Linking) (2014).
7. Chua, J. J. *et al.* The Architecture of an Excitatory Synapse. *J Cell Sci* **123,** 819–23. ISSN: 1477-9137 (Electronic) 0021-9533 (Linking) (2010).
8. Claud, E. C. *et al.* Differential Expression of 26S Proteasome Subunits and Functional Activity during Neonatal Development. *Biomolecules* **4,** 812–26. ISSN: 2218-273X (Print) 2218-273X (Linking) (2014).
9. Cooper, E. C. & Lowenstein, D. *Hippocampus* (ELS, 2003).
10. Cox, J. & Mann, M. MaxQuant Enables High Peptide Identification Rates, Individualized p.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification. *Nature Biotechnology* **26,** 1367–1372 (Dec. 2008).
11. Galli, T., McPherson, P. S. & De Camilli, P. The V0 Sector of the V-ATPase, Synaptobrevin, and Synaptophysin Are Associated on Synaptic Vesicles in a Triton X-100-Resistant, Freeze-Thawing Sensitive, Complex. *J Biol Chem* **271,** 2193–8. ISSN: 0021-9258 (Print) 0021-9258 (Linking) (1996).
12. Gerber, S. A. *et al.* Absolute Quantification of Proteins and Phosphoproteins from Cell Lysates by Tandem MS. *Proc Natl Acad Sci U S A* **100,** 6940–5. ISSN: 0027-8424 (Print) 0027-8424 (Linking) (2003).

13. Gerster, S. *et al.* Statistical Approach to Protein Quantification. *Mol Cell Proteomics* **13,** 666–77. ISSN: 1535-9484 (Electronic) 1535-9476 (Linking) (2014).

14. Gillet, L. C. *et al.* Targeted Data Extraction of the MS/MS Spectra Generated by Data-Independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. *Molecular & Cellular Proteomics* **11** (2012).

15. Gonzalez-Lozano, M. A. *et al.* Dynamics of the Mouse Brain Cortical Synaptic Proteome during Postnatal Brain Development. *Sci Rep* **6,** 35456. ISSN: 2045-2322 (Electronic) 2045-2322 (Linking) (2016).

16. Gu, J. & Gu, X. Induced Gene Expression in Human Brain after the Split from Chimpanzee. *Trends Genet* **19,** 63–5. ISSN: 0168-9525 (Print) 0168-9525 (Linking) (2003).

17. Gubar, O. *et al.* Intersectin: The Crossroad between Vesicle Exocytosis and Endocytosis. *Front Endocrinol (Lausanne)* **4,** 109. ISSN: 1664-2392 (Print) 1664-2392 (Linking) (2013).

18. Hanley, J. G. *et al.* NSF ATPase and Alpha-/Beta-SNAPs Disassemble the AMPA Receptor-PICK1 Complex. *Neuron* **34,** 53–67. ISSN: 0896-6273 (Print) 0896-6273 (Linking) (2002).

19. Hondius, D. C. *et al.* Profiling the Human Hippocampal Proteome at All Pathologic Stages of Alzheimer's Disease. *Alzheimers Dement* **12,** 654–68. ISSN: 1552-5279 (Electronic) 1552-5260 (Linking) (2016).

20. Hsieh, W. P. *et al.* Mixed-Model Reanalysis of Primate Data Suggests Tissue and Species Biases in Oligonucleotide-Based Gene Expression Profiles. *Genetics* **165,** 747–57. ISSN: 0016-6731 (Print) 0016-6731 (Linking) (2003).

21. Hu, H. Y. *et al.* MicroRNA Expression and Regulation in Human, Chimpanzee, and Macaque Brains. *PLoS Genet* **7,** e1002327. ISSN: 1553-7404 (Electronic) 1553-7390 (Linking) (2011).

22. Im, Y. J. *et al.* Crystal Structure of the Shank PDZ-Ligand Complex Reveals a Class I PDZ Interaction and a Novel PDZ-PDZ Dimerization. *J Biol Chem* **278,** 48099–104. ISSN: 0021-9258 (Print) 0021-9258 (Linking) (2003).

23. Jahn, R. & Fasshauer, D. Molecular Machines Governing Exocytosis of Synaptic Vesicles. *Nature* **490,** 201–7. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (2012).

24. Khaitovich, P. *et al.* Regional Patterns of Gene Expression in Human and Chimpanzee Brains. *Genome Res* **14,** 1462–73. ISSN: 1088-9051 (Print) 1088-9051 (Linking) (2004).

25. Kim, E. *et al.* GKAP, a Novel Synaptic Protein That Interacts with the Guanylate Kinase-like Domain of the PSD-95/SAP90 Family of Channel Clustering Molecules. *J Cell Biol* **136,** 669–78. ISSN: 0021-9525 (Print) 0021-9525 (Linking) (1997).

26. Konopka, G. *et al.* Human-Specific Transcriptional Networks in the Brain. *Neuron* **75,** 601–17. ISSN: 1097-4199 (Electronic) 0896-6273 (Linking) (2012).

27. Koopmans, F. *et al.* Comparative Analyses of Data Independent Acquisition Mass Spectrometric Approaches: DIA, WiSIM-DIA, and Untargeted DIA. *Proteomics* **18.** ISSN: 1615-9861 (Electronic) 1615-9853 (Linking) (2018).

28. Li, Z. *et al.* Evolutionary and Ontogenetic Changes in RNA Editing in Human, Chimpanzee, and Macaque Brains. *RNA* **19,** 1693–702. ISSN: 1469-9001 (Electronic) 1355-8382 (Linking) (2013).

29. Liu, X. *et al.* Extension of Cortical Synaptic Development Distinguishes Humans from Chimpanzees and Macaques. *Genome Res* **22,** 611–22. ISSN: 1549-5469 (Electronic) 1088-9051 (Linking) (2012).

30. Loveday, C. *et al.* Mutations in the PP2A Regulatory Subunit B Family Genes PPP2R5B, PPP2R5C and PPP2R5D Cause Human Overgrowth. *Hum Mol Genet* **24,** 4775–9. ISSN: 1460-2083 (Electronic) 0964-6906 (Linking) (2015).

31. Markov, N. T. *et al.* Cortical High-Density Counterstream Architectures. *Science* **342,** 1238406. ISSN: 1095-9203 (Electronic) 0036-8075 (Linking) (2013).

32. Matus, A. *et al.* Brain Postsynaptic Densities: The Relationship to Glial and Neuronal Filaments. *J Cell Biol* **87,** 346–59. ISSN: 0021-9525 (Print) 0021-9525 (Linking) (1980).

33. McMahon, H. T. & Boucrot, E. Molecular Mechanism and Physiological Functions of Clathrin-Mediated Endocytosis. *Nat Rev Mol Cell Biol* **12,** 517–33. ISSN: 1471-0080 (Electronic) 1471-0072 (Linking) (2011).

34. Mesulam, M. Brain, Mind, and the Evolution of Connectivity. *Brain Cogn* **42,** 4–6. ISSN: 0278-2626 (Print) 0278-2626 (Linking) (2000).

35. Mi, H. *et al.* PANTHER Version 11: Expanded Annotation Data from Gene Ontology and Reactome Pathways, and Data Analysis Tool Enhancements. *Nucleic Acids Res* **45,** D183–D189. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (2017).

36. Mimori-Kiyosue, Y. *et al.* CLASP1 and CLASP2 Bind to EB1 and Regulate Microtubule Plus-End Dynamics at the Cell Cortex. *J Cell Biol* **168,** 141–53. ISSN: 0021-9525 (Print) 0021-9525 (Linking) (2005).

37. Muller, C. S. *et al.* Quantitative Proteomics of the Cav2 Channel Nano-Environments in the Mammalian Brain. *Proc Natl Acad Sci U S A* **107,** 14950–7. ISSN: 1091-6490 (Electronic) 0027-8424 (Linking) (2010).

38. Muntane, G. *et al.* Analysis of Synaptic Gene Expression in the Neocortex of Primates Reveals Evolutionary Changes in Glutamatergic Neurotransmission. *Cereb Cortex* **25,** 1596–607. ISSN: 1460-2199 (Electronic) 1047-3211 (Linking) (2015).

39. Oldham, M. C., Horvath, S. & Geschwind, D. H. Conservation and Evolution of Gene Coexpression Networks in Human and Chimpanzee Brains. *Proc Natl Acad Sci U S A* **103,** 17973–8. ISSN: 0027-8424 (Print) 0027-8424 (Linking) (2006).

40. Pandya, N. J. *et al.* Correlation Profiling of Brain Sub-Cellular Proteomes Reveals Co-Assembly of Synaptic Proteins and Subcellular Distribution. *Sci Rep* **7,** 12107. ISSN: 2045-2322 (Electronic) 2045-2322 (Linking) (2017).

41. Papadimitriou, G. N. *et al.* GABA-A Receptor Beta3 and Alpha5 Subunit Gene Cluster on Chromosome 15q11-Q13 and Bipolar Disorder: A Genetic Association Study. *Am J Med Genet* **105,** 317–20. ISSN: 0148-7299 (Print) 0148-7299 (Linking) (2001).

42. Perrot, R. *et al.* Review of the Multiple Aspects of Neurofilament Functions, and Their Possible Contribution to Neurodegeneration. *Mol Neurobiol* **38,** 27–65. ISSN: 0893-7648 (Print) 0893-7648 (Linking) (2008).

43. Qu, Y. *et al.* Differential Modulation of Sodium Channel Gating and Persistent Sodium Currents by the Beta1, Beta2, and Beta3 Subunits. *Mol Cell Neurosci* **18,** 570–80. ISSN: 1044-7431 (Print) 1044-7431 (Linking) (2001).

44. Rao-Ruiz, P. *et al.* Time-Dependent Changes in the Mouse Hippocampal Synaptic Membrane Proteome after Contextual Fear Conditioning. *Hippocampus* **25,** 1250–61. ISSN: 1098-1063 (Electronic) 1050-9631 (Linking) (2015).

45. Schwanhausser, B. *et al.* Global Quantification of Mammalian Gene Expression Control. *Nature* **473,** 337–42. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (2011).

46. Schwenk, J. *et al.* High-Resolution Proteomics Unravel Architecture and Molecular Diversity of Native AMPA Receptor Complexes. *Neuron* **74,** 621–33. ISSN: 1097-4199 (Electronic) 0896-6273 (Linking) (2012).

47. Silbereis, J. C. *et al.* The Cellular and Molecular Landscapes of the Developing Human Central Nervous System. *Neuron* **89,** 248–68. ISSN: 1097-4199 (Electronic) 0896-6273 (Linking) (2016).

48. Smyth, G. K., Michaud, J. & Scott, H. S. Use of Within-Array Replicate Spots for Assessing Differential Expression in Microarray Experiments. *Bioinformatics* **21,** 2067–75. ISSN: 1367-4803 (Print) 1367-4803 (Linking) (2005).

49. Somel, M. *et al.* Transcriptional Neoteny in the Human Brain. *Proc Natl Acad Sci U S A* **106,** 5743–8. ISSN: 1091-6490 (Electronic) 0027-8424 (Linking) (2009).

50. Spicer, A. P., Joo, A. & Bowling R. A., J. A Hyaluronan Binding Link Protein Gene Family Whose Members Are Physically Linked Adjacent to Chondroitin Sulfate Proteoglycan Core Protein Genes: The Missing Links. *J Biol Chem* **278,** 21083–91. ISSN: 0021-9258 (Print) 0021-9258 (Linking) (2003).

51. Spijker, S. in *Neuroproteomics* 13–26 (Springer, 2011).

52. Sur, C. *et al.* Rat and Human Hippocampal Alpha5 Subunit-Containing Gamma-Aminobutyric AcidA Receptors Have Alpha5 Beta3 Gamma2 Pharmacological Characteristics. *Mol Pharmacol* **54,** 928–33. ISSN: 0026-895X (Print) 0026-895X (Linking) (1998).

53. Testa-Silva, G., Verhoog, M. B., Goriounova, N. A., *et al.* Human Synapses Show a Wide Temporal Window for Spike-Timing-Dependent Plasticity. *Front Synaptic Neurosci* **2,** 12. ISSN: 1663-3563 (Electronic) 1663-3563 (Linking) (2010).

54. Testa-Silva, G., Verhoog, M. B., Linaro, D., *et al.* High Bandwidth Synaptic Communication and Frequency Tracking in Human Neocortex. *PLoS Biol* **12,** e1002007. ISSN: 1545-7885 (Electronic) 1544-9173 (Linking) (2014).

55. Tsvetkov, A. S. *et al.* Microtubule-Binding Proteins CLASP1 and CLASP2 Interact with Actin Filaments. *Cell Motil Cytoskeleton* **64,** 519–30. ISSN: 0886-1544 (Print) 0886-1544 (Linking) (2007).

56. Tu, J. C. *et al.* Coupling of mGluR/Homer and PSD-95 Complexes by the Shank Family of Postsynaptic Density Proteins. *Neuron* **23,** 583–92. ISSN: 0896-6273 (Print) 0896-6273 (Linking) (1999).

57. Uddin, M. *et al.* Distinct Genomic Signatures of Adaptation in Pre- and Postnatal Environments during Human Evolution. *Proc Natl Acad Sci U S A* **105,** 3215–20. ISSN: 1091-6490 (Electronic) 0027-8424 (Linking) (2008).

58. Urnavicius, L. *et al.* The Structure of the Dynactin Complex and Its Interaction with Dynein. *Science* **347,** 1441–1446. ISSN: 1095-9203 (Electronic) 0036-8075 (Linking) (2015).

59. van den Heuvel, M. P., Bullmore, E. T. & Sporns, O. Comparative Connectomics. *Trends Cogn Sci* **20,** 345–61. ISSN: 1879-307X (Electronic) 1364-6613 (Linking) (2016).

60. Vizcaino, J. A. *et al.* 2016 Update of the PRIDE Database and Its Related Tools. *Nucleic Acids Res* **44,** D447–56. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (2016).

61. Weingarten, J. *et al.* The Proteome of the Presynaptic Active Zone from Mouse Brain. *Mol Cell Neurosci* **59,** 106–18. ISSN: 1095-9327 (Electronic) 1044-7431 (Linking) (2014).

62. Wendholt, D. *et al.* ProSAP-Interacting Protein 1 (ProSAPiP1), a Novel Protein of the Postsynaptic Density That Links the Spine-Associated Rap-Gap (SPAR) to the Scaffolding Protein ProSAP2/Shank3. *J Biol Chem* **281,** 13805–16. ISSN: 0021-9258 (Print) 0021-9258 (Linking) (2006).

63. Wilhelm, B. G. *et al.* Composition of Isolated Synaptic Boutons Reveals the Amounts of Vesicle Trafficking Proteins. *Science* **344,** 1023–8. ISSN: 1095-9203 (Electronic) 0036-8075 (Linking) (2014).

64. Wisniewski, J. R. *et al.* Universal Sample Preparation Method for Proteome Analysis. *Nat Methods* **6,** 359–62. ISSN: 1548-7105 (Electronic) 1548-7091 (Linking) (2009).

65. Yuan, A. *et al.* Neurofilament Subunits Are Integral Components of Synapses and Modulate Neurotransmission and Behavior in Vivo. *Mol Psychiatry* **20,** 986–94. ISSN: 1476-5578 (Electronic) 1359-4184 (Linking) (2015).

6

SynGO: an evidence-based, expert-curated knowledgebase for the synapse

Frank Koopmans[12], SynGO consortium, L. Niels Cornelisse[1], #August B. Smit[2], #Matthijs Verhage[1]

#co-senior and co-corresponding authors

[1]Department of Functional Genomics, [2]Department of Molecular and Cellular Neurobiology, Center for Neurogenomics Cognitive Research, VU University, Amsterdam, The Netherlands

Synapses are fundamental information processing units of the brain and synaptic dysregulation is central to many brain disorders ('synaptopathies'). However, systematic annotation of synaptic genes and ontology of synaptic processes are currently lacking. We established SynGO, an interactive knowledgebase that accumulates available research about synapse biology using Gene Ontology (GO) annotations to novel ontology terms: 87 synaptic locations and 179 synaptic processes. SynGO annotations are exclusively based on published, expert-curated evidence. Using 2922 annotations for 1112 genes, we show that synaptic genes are exceptionally well conserved and less tolerant to mutations than other genes. Many SynGO terms are significantly overrepresented among gene variation associated with intelligence, educational attainment, ADHD, autism and bipolar disorder and among de novo variants associated with neurodevelopmental disorders including schizophrenia. SynGO is a public, universal reference for synapse research and an online analysis-platform for interpretation of large scale -omics data (`https://syngoportal.org` and `http://geneontology.org`).

This chapter has been published in *Neuron 103, no. 2 (2019)*

Introduction

Synapses are information processing units of the brain that provide the foundation for higher level information integration in dendrites, neurons and networks. Use-dependent changes in synaptic strength (synaptic plasticity) are firmly established as main underlying principles of cognitive processes, such as memory formation and retrieval, perception, sensory processing, attention, associative learning, and decision making (Abdou *et al.,* 2018; Groschner *et al.,* 2018; Kandel, 2001; Petersen & Crochet, 2013; Ripolles *et al.,* 2018). Based on both genetic and neurobiological evidence, synaptic dysregulation is widely recognized as an important component of risk in many brain disorders (termed 'synaptopathies' (Boda *et al.,* 2010; Bourgeron, 2015; Grant, 2012; Monday & Castillo, 2017)), such as autism spectrum disorder (ASD), attention deficit hyperactivity disorder (ADHD), schizophrenia, Alzheimer's disease and Parkinson's disease (Arnsten *et al.,* 2012; Bourgeron, 2015; De Rubeis *et al.,* 2014; Fromer *et al.,* 2014; Heutink & Verhage, 2012; Hong *et al.,* 2016; Selkoe, 2002; Soukup *et al.,* 2018; Spires-Jones & Hyman, 2014; Sudhof, 2008). Despite these intense investigations and a large variety of research efforts focused synaptic proteins and on their subcellular organization and specific functions, only sparse efforts have been made to establish systematic resources for synapse biology in health and disease. In particular, the ontology of synaptic processes has been poorly defined, which has precluded the systematic annotation of synaptic proteins/genes.

Gene Ontology (GO) is the most widely used resource for gene function annotations. The resource has two components: (i) the ontology, a framework of definitions called 'terms' to describe gene functions and locations and their relationships, and (ii) GO annotations, statements linking genes to specific terms (Ashburner *et al.,* 2000; The Gene Ontology, 2018). The ontology is divided into three aspects: (i) molecular function (MF), defining the molecular activities of gene products (e.g., protein kinase activity); (ii) Cellular Component, defining where they are active (e.g., on synaptic vesicle); and (iii) Biological Process, defining the processes that they carry out (e.g., synaptic vesicle exocytosis). Relationships between CC terms generally specify how smaller structures are parts of larger ones. Relationships between BP terms specify how sub-processes contribute to larger ones. The accuracy of GO annotations depends on (i) how well the ontology represents Molecular Function, Cellular Component (CC) and Biological Process (BP) terms for given systems, e.g., synapses; and (ii) how well experimental evidence supports the annotations.

Using existing annotations to synaptic GO terms and synaptic gene sets, several studies have shown that synaptic genes, i.e., genes encoding synaptic proteins, are significantly enriched in genetic variation associated with several brain traits (Savage *et al.,* 2018; Zwir *et al.,* 2018) and have produced valuable leads to understand the role of synapse function and dysfunction in these traits (De Rubeis *et al.,* 2014; Fromer *et al.,* 2014; Mattheisen *et al.,* 2015; Pedroso *et al.,* 2012; Thapar *et al.,* 2016). However, it is evident that the lack of systematic annotation of synaptic genes also limits progress. Available resources, including GO, have only limited representations of synapse biology, and lacked a comprehensive ontology of synaptic processes and subcellular locations in the synapse. Rather than capturing current understanding of the synapse, existing resources are biased by uneven and patchy coverage of different aspects of synapse biology. Moreover, existing resources include data that have not been curated by synapse experts and a large fraction of the data has been aggregated in an unsupervised manner, for example by automated text mining, or by large-scale experiments that result in high rates of false-positives, such as bulk proteomics analyses and yeast two-hybrid studies. Thresholds for inclusion are not systematically defined and are typically set quite low. Together these shortcomings limit the impact of such resources and may engender incorrect conclusions, for instance in studies reporting associations between genetic findings and synapses and between synapses and brain related traits..

To overcome these limitations, we established SynGO, a partnership between the GO Consortium and 15 synapse expert laboratories in Europe, North America and Asia, for the systematic annotation of synaptic proteins. SynGO experts have developed an extensive ontology to represent synaptic locations (87 terms) and synaptic processes (179 terms) and generated almost 3000 annotations of synaptic genes/proteins to these terms, based on a novel comprehensive evidence tracking system that classifies evidence according to experiment types, model systems and target engagement types (gene modifications, antibody binding etc.), using only published data sets. Using SynGO, we observed that synaptic genes are exceptionally well conserved, relatively much more intolerant to mutations than non-synaptic genes and are associated with many brain traits, such as IQ and educational attainment, and brain disorders such as ASD, ADHD and bipolar disorder. SynGO provides a unique, publicly accessible knowledgebase (https://syngoportal.org) as a universal reference for synapse research and education, and for enrichment studies on genomic associations, mRNA profiling and proteomic data.

Results

SynGO ontologies provide comprehensive frameworks for synaptic gene annotation

To systematically annotate synaptic genes, we designed a generic synapse model as a conceptual starting point, defining locations at the synapse and processes related to the synapse, and refined this model iteratively until consensus was reached among expert laboratories worldwide (Fig. 6.1). Subsequently, we created GO terms for Cellular Components (CC) and Biological Processes (BP) for synapses and defined their relationships. At the top level of the CC hierarchy (Fig. 6.2A), synaptic proteins can be described as localized to the presynapse, the postsynapse, the synaptic cleft, the extra-synaptic space and synaptic membranes (the latter term is used when no distinction is possible between pre- and postsynaptic membranes). From these high-level terms, up to 4 additional hierarchical levels were defined for pre- or postsynaptic cytosol or membrane, or organelles within these compartments. The SynGO CC ontology adds substantial precision to the preexisting GO ontology that contained 13 terms directly connected to the central ‘synapse’ term (and 19 additional terms). SynGO maintained only two of these 13 terms (Fig. 6.2A, green symbols) and excluded 11 (Fig. 6.2A, purple symbols). Some of the GO

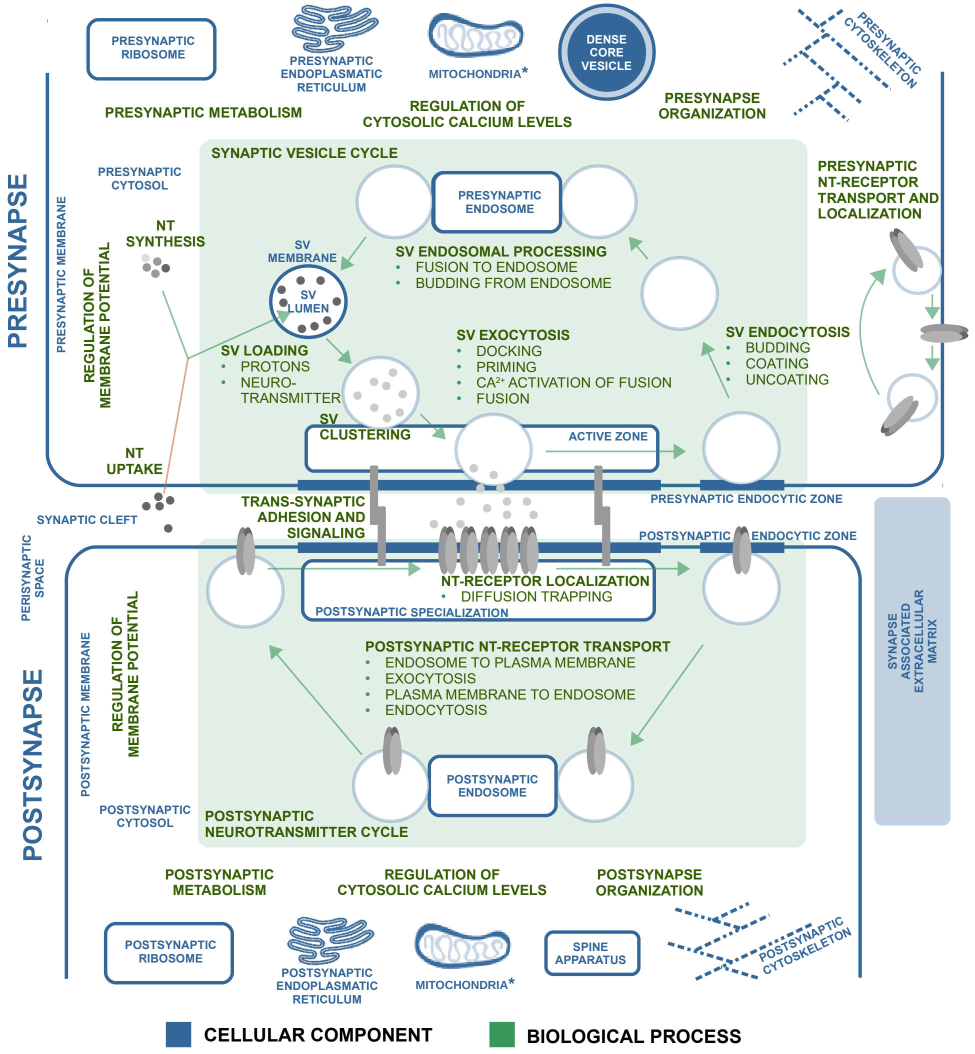

Figure 6.1: Conceptual framework of synapse ontology in SynGO. The top-level Cellular Component (location, shown in green) and Biological Process (function, shown in blue) terms are depicted in a schematic representation of a synapse. For the full set of ontology terms, which also include all sub-classifiers that further specialize terms shown here, see Figure 2 and Supplementary Table 2. *The mitochondrion is depicted for completeness, but is not part of SynGO ontology (see text).

terms were replaced by similar but more precise terms, e.g. ”presynaptic active zone dense projection” (GO) by ”presynaptic active zone” (SynGO), others were replaced with more specific terms further down in the hierarchical SynGO ontology, e.g. instead of ”symmetric synapse” and ”excitatory synapse”, we created a general term ”postsynaptic specialization” with first level subclassifiers ”postsynaptic specialization of symmetric synapse” and ”postsynaptic density”. All together, 142 SynGO CC ontology terms were designed for accurate annotation of synaptic localizations (Table S2). To visualize this elaborate ontology hierarchy and provide a standardized visualization of SynGO annotations, all CC terms populated with gene annotations in SynGO 1.0 (92/142 terms) were plotted in a circular fashion with the highest hierarchical term (synapse) in the center and each layer of subclasses in outward concentric rings (Fig. 6.2C, see Table S2 for all term names). SynGO did not define mitochondria as part of a specific synaptic CC, as mitochondrial proteins are already well annotated (Calvo *et al.,* 2016; Smith & Robinson, 2018).

BP terms for synaptic processes and their relationships were also defined consistently with existing GO-terms, with pre- and postsynaptic processes, synaptic organization, synaptic signaling, axonal/dendritic transport, and metabolism as main terms, with up to 5 levels of subclasses (Fig. 6.2B). In total, the BP ontology features 256 terms of which 212 are new. 192 of these BP ontology terms were populated with gene annotations in SynGO 1.0 and visualized in a sunburst plot (Fig. 6.2D, analogous to Fig. 6.2B, see Table S2 for all term names). Hence, these novel CC- and BP-ontologies provide a substantial innovation and also a substantially increased precision for the ontology of the synapse. Together, these ontologies provide a comprehensive structure for the systematic annotation of synaptic genes and for future computational models of synapse biology and pathophysiology.

SynGO is based on expert annotation and systematic evidence tracking

Currently available synaptic protein lists contain many unsupervised inclusions, in particular from large-scale, automated experiments expected to have substantial false positive rates. SynGO established a systematic evidence tracking protocol and annotation by synapse experts only, based exclusively on published experimental data (PubMed). The SynGO workflow (Fig. S1) was implemented in a web-interface and used by synapse experts to annotate synaptic genes. To systematically track evidence, classifications were designed for the model systems used (Fig. S2). For synaptic localization (CC), microscopy and biochemical studies were defined as the main experimental classes, each with several sub-classes. For functional studies, experimental classes were defined based on perturbation type and the methodology (assay) used to detect the consequences, again with several sub-classes (Fig. S2). These classifications were made coherent with the Evidence and Conclusions Ontology (ECO) (Giglio *et al.,* 2018), and new ECO terms were defined. Together, these three dimensions of evidence, (i) model system/preparation, (ii) experimental perturbation and (iii) assay, provide a systematic, coherent and detailed definition of the evidence to annotate synaptic genes.

Detailed reference to these three dimensions of evidence was stored as part of each annotation (PubMed ID, figure numbers, panels, see Table S3), providing a detailed rationale for each annotation, which can be reviewed by SynGO users. For any given study, annotations were made for the species used and these were subsequently mapped to the consensus human ortholog using HUGO Gene Nomenclature Committee (HGNC) data resource (Yates *et al.,* 2017). Annotations for orthologous genes in different species were possible and encouraged, yielding multiple annotations to the same consensus human ortholog originating from different species.

In addition, we applied SynGO annotations in GO Phylogenetic Annotation (Gaudet *et al.,* 2011) to infer annotations to evolutionarily-related genes, using the experimentally-supported SynGO annotations as evidence. In this process, an expert biocurator reviewed all experimentally-supported GO annotations for all members of a gene family in >100 species in the context of a phylogenetic tree and inferred functions of experimentally uncharacterized genes in tens of other organisms. In the current SynGO 1.0 we did not systematically annotate different splice forms of single genes, because systematic evidence for splice site-specific subcellular localizations or functions is currently sparse. In cases where studies used different approaches to reach the same conclusion, multiple annotations for the same gene to the same CC or BP terms were made frequently and were encouraged. Similarly, when evidence existed for annotating a single gene to multiple CC or BP terms (multiple locations or functions), multiple annotations were made and encouraged. Following standard GO annotation practice, the same gene/protein may be annotated at different levels along the SynGO hierarchical ontology tree. For instance, initial evidence may indicate that a protein is involved in synaptic transmission (SynGO term chemical synaptic transmission; GO:0007268), a subsequent study may reveal the protein regulates presynaptic secretion (SynGO term synaptic vesicle exocytosis; GO:0016079) and the most recent study may show that the protein regulates vesicle priming (SynGO term synaptic vesicle priming; GO:0016082).

Annotations completed by expert laboratories first passed through a quality control pipeline by the SynGO support team (Fig. S1) and were then added either directly to the SynGO database (`https://syngoportal.org`) or returned to the expert laboratories if further editing was required. These annotations were also deposited in the Gene Ontology annotation repository (`http://geneontology.org`) as GO-CAM models (The Gene Ontology, 2018). GO-CAM is an extension of the standard GO annotation format that allows more expressive annotations, e.g. specifying the cell type using Cell Ontology terms (The Gene Ontology, 2018), and multiple pieces of evidence for a single annotation. Together, this evidence tracking system, including detailed reference to the evidence (PMID, figure, panel), provides an excellent framework for comprehensive, transparent annotation of synaptic genes.

SynGO 1.0 provides 2922 expert-curated annotations on 1112 synaptic genes

Using the three dimensions of evidence tracking (model system/preparation, experimental perturbation and assay), 2922 expert-curated annotations were generated using a cumulative candidate synaptic gene lists from published (Lips *et al.,* 2012;

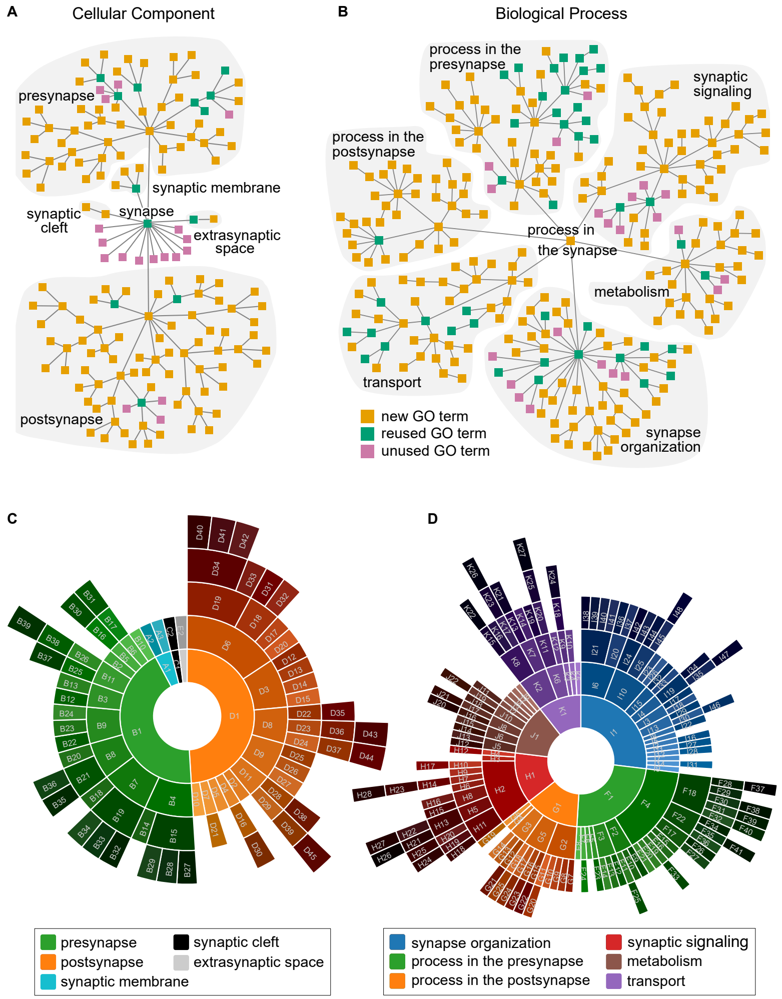

Figure 6.2: Increased resolution in synaptic ontology terms. Comparison between new terms in SynGO (orange) and pre-existing synapse ontology terms in GO (green and purple) for A) Cellular Components (CC, locations) and B) Biological Processes (BP, functions). SynGO adds resolution by creating increasingly detailed terms in a consistent systematic for Cellular Component (129 new terms) and Biological Process (212 new terms). Some existing GO terms identical to SynGO ontologies were re-used (green nodes, 13 for CC and 44 for BP) and some existing GO synapse-related terms that did not overlap with the SynGO ontologies were discarded or replaced (purple nodes, 18 for CC and 22 for BP). Supplementary Table 1 contains a complete list of pre-existing GO terms indicated in green and purple. SynGO ontology terms shown in panels A and B (in orange or green) that were populated with at least one gene annotation in SynGO v1.0 were visualized as 'sunburst plots', an alternative representation of tree structures, for C) Cellular Components and D) Biological Processes. The top-level terms in these CC and BP ontology trees, 'synapse' and 'process in the synapse' respectively, are represented by a white circle in the center of the sunburst. Terms on the second level of the ontology term tree, previously highlighted in A and B, are color coded as indicated in the legend. Subclassifiers in outer circles are shown in progressive darker colors. Supplementary Table 2 contains the complete list of SynGO ontology terms matching the sunburst plots.

Ruano *et al.,* 2010) and unpublished data resources (EU-funded projects EUROSPIN and SYNSYS, see acknowledgements), proteomic data and specific input from expert laboratories. The annotations were subjected to quality control and, typically after iterative optimization, deposited in the SynGO database and the central Gene Ontology knowledgebase (The Gene Ontology, 2018), see Fig S1. In total, we found compelling evidence for 1112 unique synaptic genes. These were admitted to the SynGO 1.0 knowledgebase. The full list of 1112 genes/proteins can be downloaded from `https://syngoportal.org`. For most genes, both subcellular localization (CC) and Biological Process (BP) evidence was found (60%, Fig. S3A), for the remaining 40%, evidence was lacking for either CC or BP and only one term was included. A core set of synaptic proteins was annotated to ≥3 CC or BP terms (Fig. S3B). Most evidence was obtained from studies of rodent species (Fig. S3C) of either intact tissue or cultured neurons (Fig. S3D). Microscopy and biochemical fractionation were the two main assay types used to make CC annotations, whereas BP annotations were based on a larger array of assay types assessing synaptic function (Fig. S3E). Together, these 2922 expert-curated annotations on 1112 synaptic genes, with a core set annotated to ≥3 CC or BP terms, provide an excellent annotation collection for descriptive studies, functional analyses of synaptic genes and gene enrichment studies.

The structure of synaptic genes is very different from other genes

As a first descriptive analysis, we compared basic structural features of SynGO-annotated synaptic genes with other genes. Human gene features were extracted from BioMart (GRCh38.p12) and Ensembl web services. Interestingly, synaptic genes were found to be different from other (non-SynGO) genes in many respects. Synaptic genes were on average more than twice as long as other genes (2.6 fold of non-SynGO genes, Fig. 6.3A), with 1.6-fold longer cDNA (Fig. 6.3B). The number of known protein coding transcripts was 1.7-fold higher (Fig. 6.3C) and the sequence of introns + exons (immature transcript length) for protein coding transcripts was more than 2 fold longer (Fig. 6.3D). Protein coding transcripts for synaptic genes also contained 1.4 fold more introns (Fig. 6.3E) and these were 1.7 fold longer (Fig. 6.3F).

To compare SynGO genes to other brain-expressed genes, we defined two control gene sets: (A) brain-enriched genes: 6600 genes with the most brain-enriched expression patterns, i.e., maximal expression difference between brain and other tissues (Ganna *et al.,* 2016); and (B) ‘top N’ genes most highly expressed in brain, with N equal to the number of unique genes in the SynGO set (1112). Differences between SynGO genes and control sets A and B were generally smaller in comparisons of gene size, introns and cDNA length, but still highly significant (Fig. S4A-L). Finally, we tested the possibility that SynGO annotated genes have a higher structural/topological complexity than other genes, especially more transmembrane regions (TMR), and that this may explain the observed differences between SynGO genes and others. A TMR prediction algorithm (A. Krogh *et al.,* 2001) indicated that SynGO annotated genes indeed encode significantly more proteins with at least one

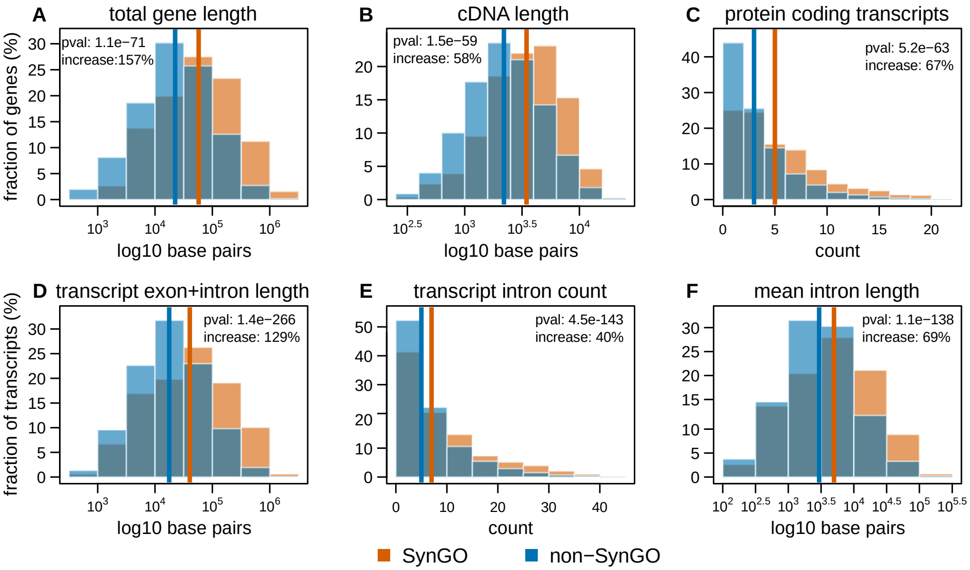

Figure 6.3: Gene features compared between synaptic genes and the rest of the genome. A) Total gene length, B) cDNA length, C) number of known protein coding splice variants, D) total length of protein coding transcripts, E) number of introns in protein coding transcripts and F) mean length of introns in protein coding transcripts. Vertical lines indicate median values for respective data distributions, which were also used to compute the percentage increase for synaptic genes. Two-sample student's t-test were applied to log transformed data to confirm overall distributions are significantly distinct, a Wilcoxon rank-sum test was used for the count data in panels C and E, "pval" in each panel denotes the resulting p-values. Analogous comparison between SynGO and brain-enriched or brain most-expressed genes is shown in Supplementary Figure S4.

TMR (35.2% versus 29.7% for the whole genome; p-value = 6.1e-5, using a two-sided Fisher exact test). However, when comparing SynGO annotated proteins to all membrane proteins, SynGO proteins are still significantly different to a similar extent and in all aspects indicated in Fig 3 and Fig S4A-L, see Fig S4M-R.

We also investigated the complexity of isoform expression of synaptic genes in cerebellar neurons using recently published full-length RNA sequencing data (Gupta *et al.*, 2018). Synaptic genes expressed a higher number of distinct isoforms, as compared to non-SynGO genes, per equal read counts, than non-synaptic genes (Fig. S5). We also analysed the number of posttranslational modifications, as important determinants of cell signalling, by testing the number of experimentally verified modifications obtained from dbPTM (Huang *et al.*, 2016) and UniProt (U. Consortium, 2018) per protein and per amino acid (to correct for difference in average protein length; Fig. S6). The incidence of all major modifications, phosphorylation, ubiquitination, acetylation and S-nitrosylation appear to be all significantly higher in synaptic proteins as compared to other proteins. However, these observations might emerge, at least in part, from the fact that synaptic proteins are more extensively studied experimentally.

Synaptic genes emerged earlier in evolution than other genes, primarily in three major waves

We tested when SynGO genes emerged in evolution relative to other genes. We found that their evolution follows a pattern that differs substantially from the overall pattern for all human genes (Fig. 6.4A). Specifically, SynGO genes evolved primarily in three "waves" of innovation, during which modern-day synaptic genes were gained at a faster rate than other human genes.

The first wave of emergence of SynGO genes, was prior to the last eukaryotic common ancestor (LECA), approximately 1800 million years (Mya) (Kumar *et al.,* 2017). While LECA was unicellular and obviously did not form synapses, it did possess cellular machinery that would later be co-opted for the synapse, such as vesicle trafficking, exocytosis and signal reception.

The second wave was prior to the last common ancestor of the eumetazoa (multicellular animals) and corresponds with the first appearance of the synapse. Among SynGO genes gained during this wave, we found strong enrichments for pre- and postsynaptic membranes and the postsynaptic density (Fig. S7B) and weak enrichments for a few synaptic processes (Fig. S7C).

The third wave was prior to the last common ancestor of vertebrates, suggesting significant synaptic evolution in this period. SynGO genes gained during this last wave are enriched again for the postsynaptic density and now also the active zone; and for more specific, largely regulatory processes: regulation of synaptic organization, synapse adhesion, modulation of synaptic signaling, and regulation of postsynaptic neurotransmitter receptors (Fig. S7E). By this time, approximately 450 Mya, about 95% of all SynGO genes were already in place, with very few additional synaptic genes appearing after that point.

A similar trend, albeit with smaller differences, was observed when gene duplication events were not weighted (Fig. S7). Figure 4B shows one of the few exceptions to this rule: the carnitine palmitoyltransferase gene family expanded via a gene duplication prior the last common ancestor of placental mammals, resulting in an additional, neuron-specific paralog found only in placental mammals (CPT1C), whereas other amniotes have only two paralogs (CPT1A, CPT1B) expressed primarily in other tissues. CPT1C is localized to the endoplasmic reticulum in neurons and has been shown to directly regulate the levels of AMPA receptors in the postsynapse (Fado *et al.,* 2015).

Overall, however, our analysis indicates that the synapse is highly conserved among modern vertebrates, as suggested before (Emes *et al.,* 2008), and that 95% of the human synaptic genes in SynGO 1.0 are shared among vertebrates. As the invertebrates C. elegans and D. melanogaster have been important model organisms in synapse biology, we also explored how many paralogs emerged in these invertebrates and how many in the vertebrate lineage (until humans) for any shared gene. For both invertebrates, we found that almost 30% of all genes have a 1:1 relationship with human genes (one paralog identified in each species, Fig. 6.4C). For most genes, more than a single paralog is identified ('many') with one a single paralog in C. elegans and D. melanogaster (many;1) or more than one in all species (many:many, Fig. 6.4C). Interestingly for synaptic genes, we found fewer 1:1 relationships and more many:1 and many:many (Fig. 6.4C). This indicates that synaptic genes underwent gene duplication at a higher rate than other genes after the vertebrate/invertebrate bifurcation.

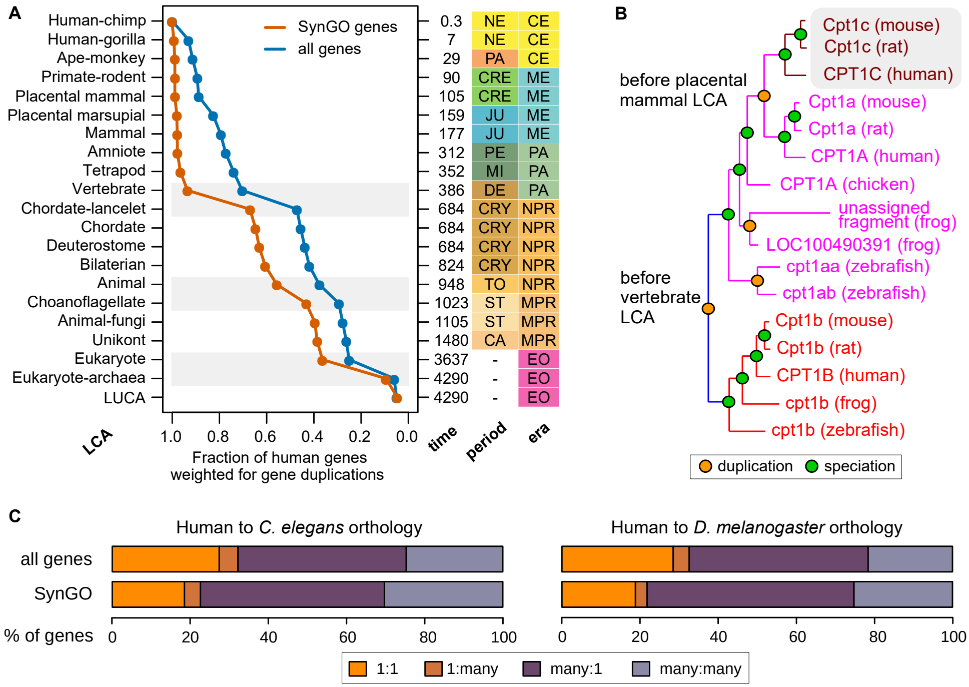

Figure 6.4: Synaptic genes are exceptionally well conserved. (A) Cumulative distribution of synaptic genes (orange) and all human genes (blue), by gene age. Highlighted areas (grey) show periods of rapid gain of synaptic genes. Ages (time in Million Years Ago) are obtained from dating of gene duplication events (relative to speciation events) in PANTHER gene trees (Mi *et al.,* 2019). Clades are shown on the y-axis, their names on the left and estimated speciation times on the right. LCA: Last Common Ancestor. LUCA: Last Universal Common Ancestor. Eras; CE: Cenozoic, ME: Mesozoic, PA: Paleozoic, NPR: Neo-Proterozoic, MPR: Meso-Proterozoic, EO: Eoarchean. Periods; NE: Neogene, PA: Paleogene, CRE: Cretaceous, JU: Jurassic, PE: Pennsylvanian, MI: Mississipian, DE: Devonian, CRY: Cryogenian, TO: Tonian, ST: Stenian, CA: Calymmian. Note that unlike the phylostratigraphic approach (Domazet-Lošo *et al.,* 2007), ages reflect not simply the oldest traceable gene age, but explicitly consider gene duplication, by adding a fractional count for each duplication event along the evolutionary path to a modern gene (see Methods for details). This is critical due to the prevalence of gene duplication in the evolution of eukaryotic genomes. (B) Evolution of the family of genes containing CPT1C (highlighted in grey), a synaptic gene annotated in SynGO. There are three tissue-specific isoforms in this family; CPT1A (liver), CPT1B (muscle) and CPT1C (brain). The latter is only found in placental mammals. C) Orthology relations between human genes and their counterparts in Caenorhabditis elegans and Drosophila melanogaster were classified by the number of paralogs matching respective organisms. For example, the many-to-one group contains all human genes that have undergone gene duplication from their ancestral gene while the given model organism has not.

Synaptic gene expression is enriched in the brain

We predicted that expression levels of SynGO genes is higher in the brain than in other tissues. To test this, we compared tissue specific expression using different gene-sets in GTEx v7 (G. T. Consortium *et al.,* 2017). Brain enrichment was computed by dividing the number of transcripts detected in brain over those in other tissues, expressed as log2 fold change (see Methods) and plotted against the expression level of this transcript in brain. As shown in Fig S8A, expression of SynGO genes is generally higher in brain than in other tissues, although some SynGO genes are in fact de-enriched in brain (below horizontal line at zero). SynGO genes with high expression levels in the brain are, on average, enriched to a similar extent as those with lower expression levels in the brain (Fig. S8A-B). We compared brain expression enrichment for different SynGO CC and BP terms. Several terms within these ontologies, especially in BP, are predicted to be highly brain specific, e.g., trans-synaptic signaling, active zone assembly or postsynaptic density organization, whereas others are expected to be similar to terms outside the synapse and outside the brain, e.g., phosphatase and kinase pathways. Indeed, specific analyses of individual SynGO terms in CC and BP ontologies revealed a large degree of heterogeneity among proteins annotated for different terms (Fig. S8C-D). The pre- and postsynaptic plasma membranes and especially the postsynaptic density contain proteins that are highly significantly enriched in brain (Fig. S8C). Active zones and synaptic vesicles, but not dense core vesicles, also contain significantly enriched proteins (Fig. S8C). For BP, a strong enrichment was observed for most major synaptic processes except metabolism and transport (Fig. S8D). Taken together, these data indicate that expression of SynGO genes is higher in brain than in other tissues, especially for ‘synapse-specific’ locations/functions.

Synaptic proteins are exceptionally intolerant to mutations

The frequency of coding variants in the general population is an indication of the functional constraints. To test whether SynGO genes have the same loss-of-function mutation incidence as other genes, we used the probability of being loss-of-function intolerant (pLI) obtained from the Exome Aggregation Consortium (ExAC, (Karczewski *et al.,* 2017). The pLI was compared between all SynGO genes and other genes. A major difference in loss-of-function intolerance was observed; SynGO genes are exceptionally intolerant to loss-of-function mutations relative to non-SynGO, brain-enriched and ‘top N’ most highly brain expressed control genes (Fig. 6.5A-C). The distribution of high pLI values was similar among different CC and BP terms (Fig. 6.5D-E). In the CC ontology, pLI scores were particularly high (mean value ≥ 0.7) for PSD and active zone genes (which also contribute to parent terms). Interestingly, the synaptic vesicle and dense core vesicle annotated genes showed much lower pLI scores (mean value ≤ 0.5). Taken together, these data indicate that synaptic genes are exceptionally intolerant to loss-of-function mutations, suggesting that functional constraints and evolutionary selection pressure on synaptic genes are much stronger than for other genes.

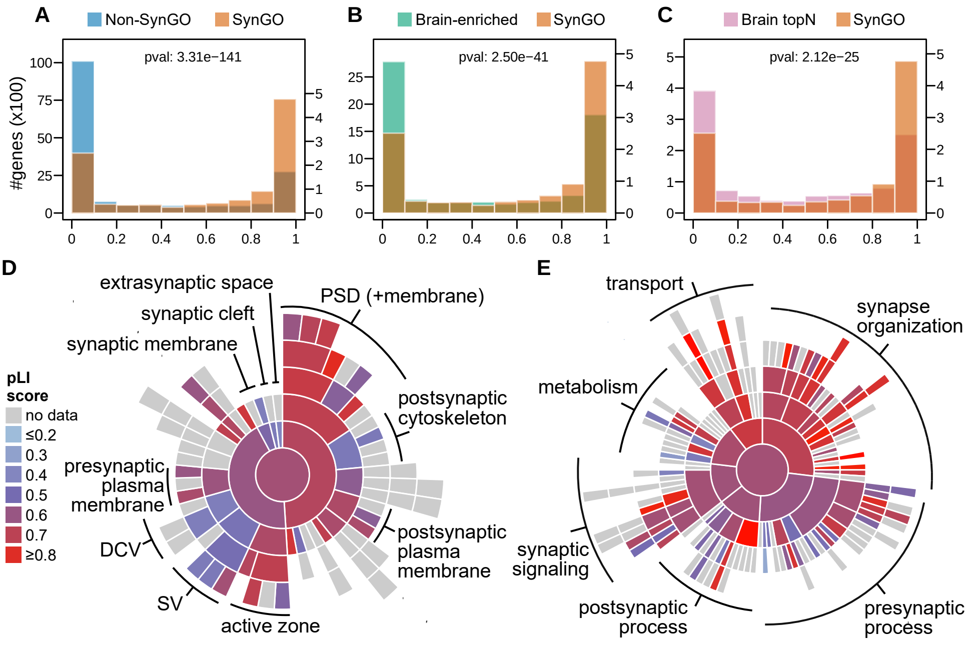

Figure 6.5: Gene pLI scores, indicating probability of intolerance to Loss of Function (LoF) mutation. pLI scores compared between synaptic genes and A) rest of the genome, B) brain enriched genes and C) 1112 genes most highly expressed in brain. Two-sample Wilcoxon signed-rank test p-values indicate that overall distributions are significantly different (denoted as "pval" in panels A-C). Mean pLI scores for respective synaptic genes annotated against D) SynGO Cellular Component terms and E) Biological Process terms are visualized in a sunburst plot, for terms with at least 5 unique annotated genes with a pLI score. Terms where annotated genes are typically LoF tolerant are shown in blue, while terms with mostly LoF intolerant genes are shown in red. Note that the CC and BP sunburst plots are aligned with Figures 2C and 2D, respectively.

Synaptic proteins annotated to closely related SynGO terms are more likely to interact

SynGO proteins annotated to the same ontology term or to closely related terms are predicted to often be in the same protein complexes or be involved in the same process and are thus more likely to interact. This prediction was tested using protein-protein interaction data available through StringDB v10.5 (Jeanquartier *et al.*, 2015), using the 'high confidence' interaction filter. Proteins reported to be in the same protein complexes were significantly overrepresented in synaptic genes annotated against the same CC term in SynGO (Fig. S9A) and also for the same BP term (Fig. S9B). Hence, synaptic proteins annotated for the same CC or BP term are much more likely to interact and, vice versa, interacting synaptic proteins are much more likely to have the same localization or be part of a similar process.

Different synaptic preparations contain largely overlapping synaptic protein collections

SynGO enables the analysis of existing, large-scale proteomics data from biochemical preparations enriched for synaptic components. We extracted data from 19 well-described and quantitative proteomic studies on 3 biochemical preparations enriched for synaptic components: (A) synaptosome fractions (7 studies, (Arnsten *et al.,* 2012; Biesemann *et al.,* 2014; Chang *et al.,* 2015; Filiou *et al.,* 2010; Moczulska *et al.,* 2014; Monday & Castillo, 2017; Pardinas *et al.,* 2018)); (B) postsynaptic density fractions (PSD, 6 studies, (Bayes, Collins, *et al.,* 2012; Arnsten *et al.,* 2012; Bayes, van de Lagemaat, *et al.,* 2011; Collins *et al.,* 2006; Monday & Castillo, 2017; Roy *et al.,* 2018)) and (C) active zone or docked vesicle fractions (5 studies, (Abul-Husn *et al.,* 2009; Boyken *et al.,* 2013; Hong *et al.,* 2016; Morciano *et al.,* 2005; Phillips *et al.,* 2005)). Synaptosome studies have identified between 894 and 3331 proteins (Fig. 6.6A). These protein collections contained between 17 and 39% of the SynGO CC annotated proteins. Together, 80% of proteins with a SynGO CC annotation were detected in at least one of the synaptosome preparations. PSD analyses typically identified smaller numbers of components, up to 1207 (Roy *et al.,* 2018).

A consensus set of proteins identified in at least three proteomic datasets per compartment contains 2621 unique proteins for synaptosome, 791 for PSD and 88 for active zone. The PSD components showed a large degree of overlap (90%) with the synaptosome consensus set, with only 76 proteins exclusively identified in the PSD consensus set (Fig. 6.6B). 73% (1906 proteins) of the synaptosome consensus set is not found in the PSD consensus set, 78% (2033 proteins) is not found in SynGO 1.0 and in total 61% (1596 proteins) of the synaptosome consensus set was not found in either PSD, active zone or the SynGO database.

Active zone preparations yielded smaller numbers of proteins, maximally 249 (Fig. 6.6A). These protein collections contained between 35 and 62% of SynGO annotated proteins, slightly more than synaptosome and postsynaptic density percentages. A total of 2084 proteins currently lacking SynGO 1.0 Cellular Component annotation were identified in at least three proteomics datasets of synaptosome, active zone or PSD subcellular fractions (Fig. 6.6B).

Taken together, these data indicate that SynGO aids in dissecting overlap and differences in large synaptic protein sets that were purified in different synaptic preparations. Many proteins identified in such fractions await experimental validation before they can be annotated to SynGO CC and BP terms.

Synaptic genes are enriched among genes associated with various brain traits

Results from large scale genetic studies are often used to test for association of a trait of interest with a set of functionally related genes. Such tests gain power with a higher confidence definition of the gene sets used. We predicted that expert-curated, evidence-based SynGO genes show robust associations with experimental data on brain traits and that SynGO gene sets are more strongly associated than existing synapse gene sets. We tested this prediction on genome-wide associa-

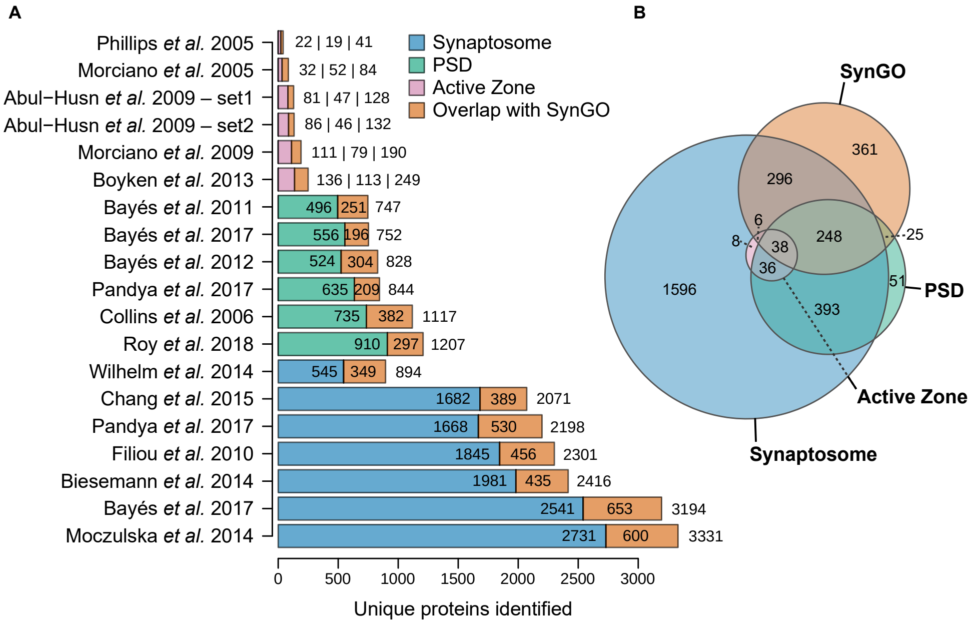

Figure 6.6: Representation of SynGO proteins in large scale proteomic analyses of synaptic (sub-) fractions. Proteins identified in a selection of published proteomic analyses of biochemically purified synaptic fractions (synaptosomes, postsynaptic densities (PSD) and active zone) were analyzed for SynGO annotated proteins. A) The number of unique proteins detected in the selected studies, blue: synaptosomes; green: PSD; pink: active zone, orange: subset of proteins that are CC annotated in SynGO. B) overlap among SynGO CC annotated proteins (orange) and 'consensus sets' for synaptosome (blue), PSD (green) or active zone (pink), defined as proteins identified in at least three datasets described in panel A (matching respective compartments). Supplementary Table 4 details the selected proteomics studies and their identified proteins.

tion study (GWAS) data for three continuous traits, educational attainment (EA) (J. J. Lee *et al.*, 2018), Intelligence Quotient (IQ) (Savage *et al.*, 2018) and human height (Wood *et al.*, 2014), and for five brain disorders, ADHD (Demontis *et al.*, 2016), autism spectrum disorder (ASD) (Grove *et al.*, 2019) schizophrenia (Pardinas *et al.*, 2018), bipolar disorder (Psychiatric, 2011) and major depression (Wray *et al.*, 2018). The association with gene-sets based on SynGO genes and previously annotated synaptic genes in GO were compared for these traits to three control gene sets: all other genes, other genes with similar brain-enriched expression and genes with similar (high) conservation. Two analysis methods were used, MAGMA (de Leeuw *et al.*, 2015) and linkage disequilibrium score (LDSC) regression analysis (B. K. Bulik-Sullivan *et al.*, 2015). These two methods have similar goals, yet rely on different assumptions and statistical algorithms. LDSC tests for enrichment of SNP-based heritability for various traits in gene-sets, while MAGMA tests whether gene-level genetic association with the various traits is stronger in specific gene-sets. Both methods account for confounders like gene size and linkage disequilibrium in different ways.

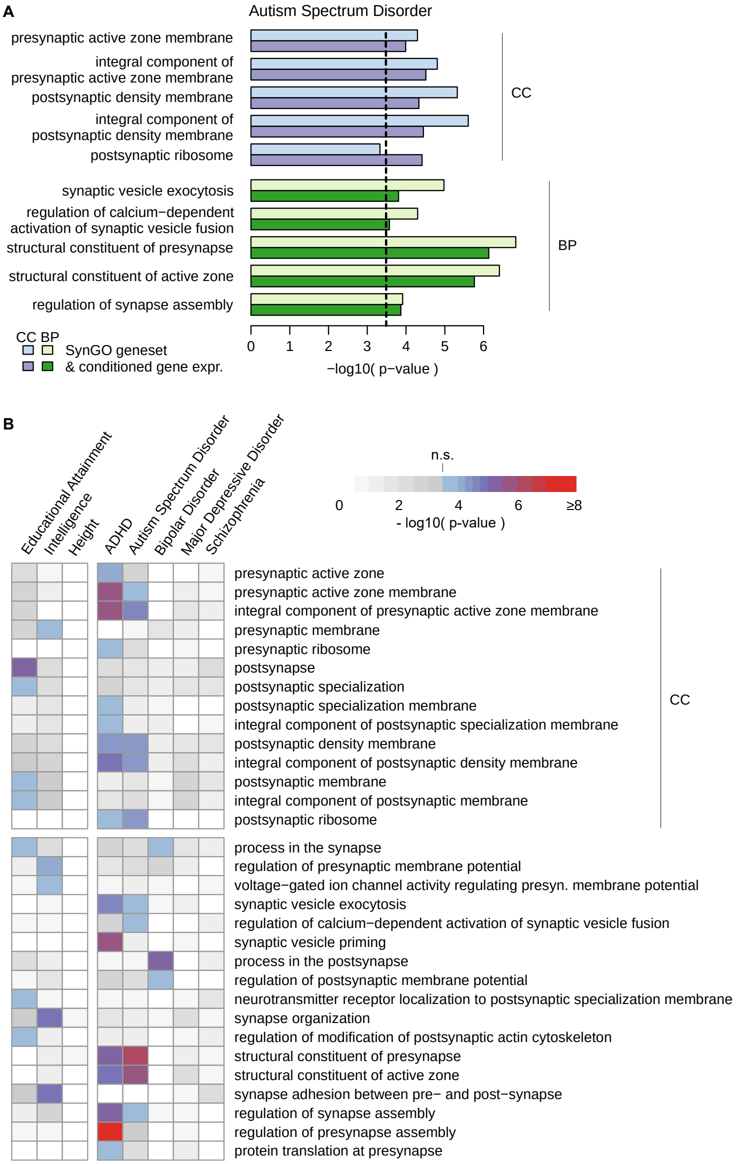

Figure 6.7: Enrichment study of SynGO genesets in GWAS. A) Magma analysis of Autism Spectrum Disorder revealed enrichment of SynGO Cellular Components (light blue) and Biological Processes (light green). Conditioning by gene expression values (GTEx) typically reduced the signal, except for postsynaptic ribosome, as visualized in dark blue and dark green. Only SynGO ontology terms significant after Bonferroni correction at α 0.05 (Pbon=0.05/154, vertical dashed line) in the latter analysis are shown. B) Overview of significantly enriched SynGO ontology terms in various GWAS. P-values from Magma analysis, with conditioning by gene expression values, were color-coded from blue to red for all ontology terms significant after Bonferroni correction at α 0.05. Additional studies are available in Supplementary Figure S10 and Supplementary Table 6.

Fig 7A shows gene-set analyses using MAGMA for ASD. We observed a highly significant association of the sets involving presynaptic active zone and the postsynaptic density (CC-terms), for presynaptic functions and synapse assembly (BP-terms; Fig. 6.7A). These associations remained significant, albeit typically less strongly, when conditioned on brain gene expression values (Fig. 6.7A, dark colors), or conditioned on homology conservation scores (Fig. S10A-B). Interestingly, one set of SynGO genes, the postsynaptic ribosome genes, was not significant when compared to all other genes, but became significant when conditioned on brain-expressed genes. Hence, gene-set analysis for SynGO genes in ASD GWAS data reveals new and highly significant associations with pre- and postsynaptic compartments and presynaptic processes.

Similar analyses were performed for all other traits listed above (Fig. 6.7B). SynGO genes were significantly associated with educational attainment, especially genes annotated with postsynaptic localizations and processes. Five SynGO ontology terms were associated with intelligence, but none were associated with human height. Furthermore, many ontology terms were associated with ADHD, especially ontologies involving locations and functions related to the presynaptic active zone and presynaptic assembly (Fig. 6.7B). Finally, strong associations of both pre- and postsynaptic terms were observed for ASD, and for postsynaptic processes with bipolar disorder (Fig. 6.7B). Very similar conclusions were reached when additionally conditioning on homology conservation scores (Fig. S10A-B) and when LDSC regression analysis was used instead of MAGMA (Fig. S10C-D).

Taken together, SynGO genes are strongly enriched in GWAS results for brain-related traits, with new links becoming manifest between ASD and the synapse; ADHD and presynaptic genes; educational attainment and postsynaptic processes and several other links between synaptic genes and bipolar disorder or intelligence.

Synaptic genes are enriched among de novo protein-coding variants for four brain disorders

In addition to GWAS studies, exome sequence studies of de novo coding variation have recently become available, allowing us to perform enrichment studies in SynGO genes among all de novo coding variation detected from several brain disorder patient populations. We tested for enrichment in SynGO genes of protein truncating (PTV) and missense mutations that were previously reported to be associated with 4 brain diseases: Developmental Delay (DD, 4293 trios), Intellectual Disability (ID, 971 trios), ASD (3982 trios) and Schizophrenia (SCZ, 1024 trios), with non-syndromic Congenital Heart Defect (CHD, 1487 trios) and unaffected siblings (UNAFF SIB, 2216 trios) as non-affected classes (see Table S7 for all references). PTV and missense mutations were filtered if they were present in the ExAC reference database (Lek *et al.,* 2016), and de novo enrichment in each group was compared against a mutation model that estimates the expected mutation rate among each gene set. SynGO gene enrichment was compared to previously annotated synaptic genes in GO and to matched brain-enriched genes: control gene sets with similar brain enrichment/specificity and gene size exactly matching SynGO genes. SynGO genes were robustly enriched for all 4 disease classes (Fig. 6.8A-B), most strongly for ID (>2 fold enriched), but also for DD (1.6 fold enriched), ASD (1.4 fold enriched) and SCZ (1.3 fold enriched). All these enrichments for SynGO genes were substantially stronger than for synaptic genes previously annotated in GO, especially for DD and ID (Fig. 6.8A). PTVs and missense mutations in SynGO genes were not enriched for CHD-NS and in unaffected siblings (Fig. 6.8A).

To test the distribution of these enrichments within SynGO ontology terms, we plotted the enrichment p-values for each term as false colour values in SynGO CC and BP ontologies (Fig. 6.8C-D, Table S7). Highly enriched gene sets were unevenly distributed among locations and processes. For subcellular locations (CC) the strongest associations were observed in postsynaptic density and active zone, together with pre- and post-synaptic plasma membrane terms (Fig. 6.8C). For Biological Processes (BP), the strongest associations accumulated in presynaptic processes, especially synaptic vesicle exocytosis and generation of the presynaptic membrane potential, with further association in postsynaptic processes and synapse organization (Fig. 6.8D). Together these data show that SynGO genes were strongly enriched for de novo PTV and missense variation in all four brain disorders. Importantly, SynGO genes are more robustly enriched than GO-genes previously annotated to the synapse.

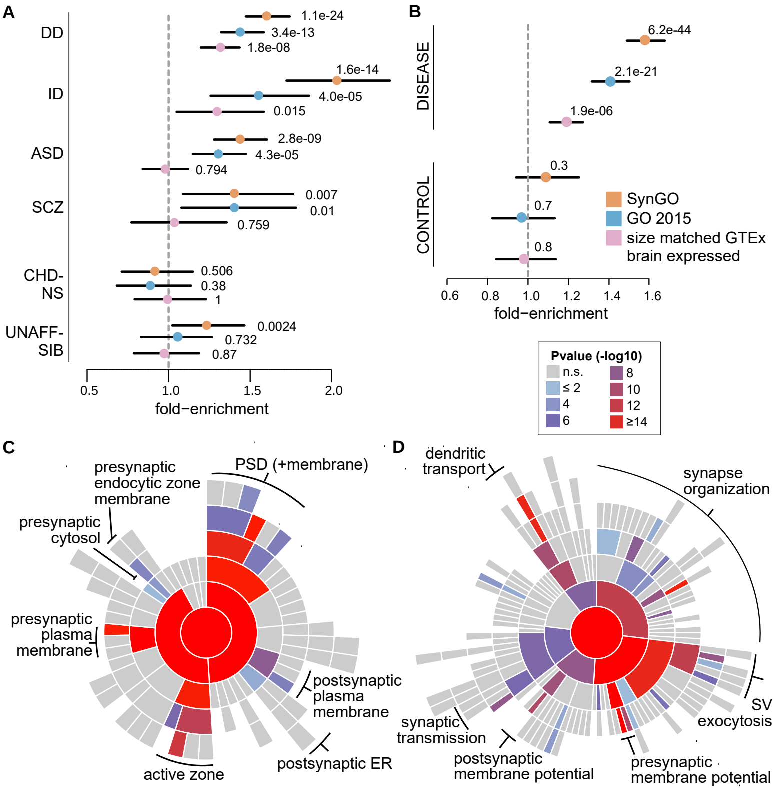

Figure 6.8: Enrichment for protein truncating (PTV) and missense mutations in SynGO genes. A) synaptic genes are more enriched for PTV and missense mutations among patients with brain disorders compared to the control set of GTEx brain expressed genes of equal size and compared to pre-existing synaptic annotations in GO. For each comparison the p-values from a binomial test against mutation model expectation are shown as text, their median fold-enrichment as a circle (color coded by gene set) and the 10~90% quantile of fold-enrichment as a horizontal line. Patient populations with brain disorders: Developmental Delay (DD), Intellectual Disability (ID), Autism (ASD) and Schizophrenia (SCZ). As a control group we included patient populations with non-syndromic Coronary Heart Disease (CHD-NS) or unaffected siblings (UNAFF-SIB). B) Group-level effects were tested for the patient populations described in panel A. The median disease p-value per ontology term (with at least 5 unique annotated genes) was visualized for C) Cellular Components and D) Biological Processes. Note that the CC and BP sunburst plots are aligned with Figures 2C and 2D, respectively.

Discussion

This study describes SynGO, the first comprehensive knowledgebase that provides an expert community consensus ontology of the synapse. The ontology and annotations accumulated in SynGO provide a comprehensive definition of synapses, new unique features of synapses, new links between synapses and brain disorders and excellent future perspectives as an up-to-date interactive community resource. We deliver proof of principle application of SynGO 1.0 for the analysis of gene/protein properties, evolutionary conservation, mRNA expression, loss of function tolerance, protein-protein interaction, enrichment in GWAS data for brain-related traits and brain disorders, and in rare de novo coding variation for neurodevelopmental disorders including schizophrenia.

SynGO provides a major step forward in defining synapses

Adequately defining a biological system like the synapse requires a coherent and logical definition of its components, their relationships and how biological functions emerge from these. The SynGO ontology is the first ontology to provide such definitions coherently for the synapse. The SynGO 1.0 ontology has defined 87 Cellular Component (CC) and 179 Biological Process (BP) terms, designed in consensus by expert laboratories worldwide. Previous models suffered from the lack of a coherent, top-down design of synapse-related ontology terms and relations. Consequently, many heterogeneous terms, both specific and general, were positioned directly under the master term 'synapse' (see Fig. 6.2A-B).

Defining synapses adequately also requires the underlying annotations to be accurate and reliable. SynGO is exclusively based on published, expert-curated evidence and detailed classification of this evidence. This is a substantial innovation that provides accountability for decisions made by experts and allows for structured discussions and resolving annotation disputes, in particular in the web-based SynGO resource (https://syngoportal.org). Moreover, different types of evidence can now be integrated in statistical models in a differential manner. For instance, evidence that is considered very strong can be given a higher weight than evidence less so. Finally, providing evidence-tracking tools to (future) expert contributors engages the synapse research community, ensuring that SynGO annotations are based on solid evidence. Hence, the new SynGO evidence tracking system provides a fundamental step forward for annotation accuracy, transparency and expert-engagement, and a solid basis for future refinements in a biology-driven overall synaptic ontology framework.

Using SynGO 1.0 annotations, we show that the SynGO ontology indeed defines the synapse adequately. We show that (i) SynGO genes are indeed more evolutionary conserved than other genes (Fig. 6.4), as previously shown (Emes *et al.,* 2008), and (ii) that synaptic genes are indeed brain enriched, with brain-specific aspects of synapses particularly enriched, as opposed to generic aspects, like transport and metabolism (Fig. S8). Furthermore, (iii) SynGO proteins documented to interact in published protein-protein interaction data are much more likely to be annotated to the same ontology terms (Fig. S9). Finally, (iv) enrichment of synaptic genes among genes associated with all tested traits in GWAS data (Fig. 6.7) and among rare variants causing neurodevelopmental disorders (Fig. 6.8), is without exception stronger for SynGO genes than for gene-sets previously annotated to the synapse. Together these four groups of observations confirm that SynGO defines synapses adequately, consistent with previous findings, and consistently outperforms previous gene set resources used in gene-set analyses.

While the definition of a synapse is now becoming accurate and reliable, the definition of synaptic genes remains precarious. No cellular compartment operates in isolation. Components move in and out and no gene product, also not of SynGO genes, is expressed exclusively in the synapse. Since GO annotations for location (CC) and process (BP) are independent, genes that regulate synaptic function do not necessarily have to be located in the synapse. In principle, this opens the possibility of annotating for instance transcription factors that regulate expression of synaptic genes. SynGO 1.0 currently only lists few of these examples, but it will eventually be useful to include such genes in SynGO annotation. Such genes can be easily excluded from an analysis by filtering for CC terms, i.e., only genes that have a confirmed synaptic location will be retained. Other regulatory aspects of synapse function may include proteins derived from the extracellular matrix, axon, dendrite or glia, which are not yet accommodated in SynGO 1.0.

Taken together, SynGO provides a comprehensive definition of the synapse with new, elaborate and consensus ontologies, accurate and transparent evidence tracking and close to 3000 validated annotations. SynGO is ready to serve as a universal reference in synapse biology and for enrichment studies using –omics data, but also to form a fundamental component of future computational models to help understand synaptic computation principles in the brain and their dysregulation in disease.

SynGO discovers unique features of synaptic genes and new disease links

In addition to adequately defining synapses, SynGO also allowed us to identify several novel features of synapses and synaptic genes/proteins. First, we show that synaptic genes are structurally very different from other genes (Fig. 6.3). Second, nearly all synaptic genes have evolved prior to the last common ancestor of all vertebrates, >450M years ago, much earlier than the average for other human genes (Fig. 6.4). Third, synaptic genes are exceptionally intolerant to mutations (Fig. 6.5). We find that synaptic genes have accumulated more coding and non-coding sequence, which may have served to expand their transcriptional regulatory repertoire and diversification of functions of the encoded proteins. Moreover, larger genes with more intron-exon boundaries may have given rise to more alternatively spliced variants; a prediction that may soon become validated with the introduction of new long-read RNA sequencing. Also, mechanisms of gene duplication and splicing have generated expansion of synaptic gene diversity. Interestingly, as synaptic genes are found highly intolerant to mutation this diversification must have come with incorporating new essential synaptic functions, such as in features of plasticity, contributing to accelerating computational capabilities of the brain during evolution.

Synaptic dysregulation is central to many brain disorders (‘synaptopathies’).

SynGO analyses described here strengthen the links between synapses and many brain traits (Fig. 6.7-8). Many SynGO CC and/or BP terms are enriched among genes associated with educational attainment, intelligence, ADHD, ASD and bipolar disorder. In particular, analysis of SynGO suggests a link between educational attainment and postsynaptic processes. Furthermore, these analyses provide better insights in links between ADHD and both pre- and postsynaptic genes, between ASD and presynaptic genes (in addition to the well-known links to the PSD, see (Bourgeron, 2015)) and between bipolar disorder and postsynaptic genes. One informative achievement of SynGO analyses is that, due to detailed structure of the SynGO ontology, genetic risk for each disease was mapped to specific synaptic locations and processes. The mapping resolution to specific terms is currently limited by the small number of genes/proteins annotated in some sub-classes in levels 3 and down. More synapse research is necessary to drive this refinement to saturation and allow more specific and definitive associations between genetic risk for brain disorders and distinct synaptic locations and processes.

SynGO is expected to grow as an expert community effort

Although SynGO 1.0 contains 2922 annotations, this is still only a fraction of all relevant information available in scientific literature. Only for a core set of proteins, SynGO 1.0 contains three or more annotations per protein. A concerted effort by all experts involved in synapse research will help to uncover a larger fraction of available information on synapses and further improve the impact of SynGO. The publicly accessible SynGO portal has been optimized to make such efforts with a user-friendly interface and stored credits for each annotator.

SynGO 1.0 contains 2922 annotations against 1112 genes, but proteomics studies of synaptic preparations implicate a few thousand proteins in synapses (Fig. 6.6). An unknown fraction of these synaptic candidate proteins will prove to be bona fide synaptic, for which the experimental evidence is currently lacking. It is important to note that biochemical purifications cannot purify synapses or synaptic compartments to completeness and some candidate proteins will remain false positives. SynGO 1.0 does not include these candidates by default to avoid low confidence analyses with SynGO data. However, they can be downloaded from the SynGO database for validation studies. SynGO is also working together with UniProt (UniProt, 2018) to accumulate information on available antibodies to facilitate this validation.

Using the public SynGO interface (`https://syngoportal.org`), SynGO ontologies and gene annotations can be used for enrichment analyses of any new data set (genomic, mRNA or protein) and differences between experimental and control groups can be computed and visualized using SynGO visualization tools (Fig. 6.1, Fig. 6.2C-D). The SynGO ontologies and annotations are also fully integrated into the central GO resource (`http://geneontology.org`), and are made available as part of standard GO releases, so that this information is automatically included in all of the myriad analysis environments and tools that use the GO. SynGO annotations are available as both standard GO annotations (`http://geneontology.org/docs/go-annotations/`) and as GO-CAM models (`https://geneontology.`

cloud/browse/g:SynGO).

Proteins that function in different types of synapses are systematically annotated in SynGO. However, SynGO 1.0 and currently published data do not yet provide sufficient resolution to define individual synaptic proteomes (synaptomes) down to specific synapse populations, which will be important to predict function, e.g. being facilitating or depressing, or being inhibitory or excitatory, and to identify changes in disease. Biochemical purifications or other systematic studies of specific synapse populations will be required to establish such specific synaptomes. Until such data become available, the currently available single cell mRNA resources can be a proxy to define which synaptic genes are expressed in specific neuronal populations. Hence, continued research in the synapse field provides excellent opportunities to further improve and expand SynGO, while, conversely, SynGO can provide the conceptual framework and be a key hypothesis generator for such future studies.

The approach described here, including the novel evidence tracking and multimodal analyses, may also provide a foundation for higher fidelity annotation of other systems, other parts of neurons, other brain cells or non-neuronal cells and systems. Eventually, such efforts will provide a more complete picture of biological processes and common themes, e.g. in secretion principles or signal detection/integration, between synapses and other systems.

Conclusion

Taken together, SynGO provides the scientific community with a public data resource for universal reference in synapse research, which is fully integrated in the Gene Ontology resource (http://geneontology.org), and ready for online gene enrichment analyses. By the engagement of the synapse research community, SynGO aims at reaching saturation to establish a truly comprehensive definition of the synapse. SynGO already brings together many expert laboratories, but actively seeks participation of additional experts to annotate new synaptic genes and/or refine existing annotations. A user-friendly interface (https://syngoportal.org) supports submission of such contributions, which will be reviewed by domain experts before being admitted to SynGO.

Acknowledgements

SynGO was funded by The Stanley Center for Psychiatric Research at The Broad Institute of MIT and Harvard. SynGO was built on previous efforts (synaptic parts lists) funded by EU (EUROSPIN: HEALTH-F2-2009-241498 and SYNSYS: HEALTH-F4-2010-242167). A.B.P was supported by Spanish grants ref. BFU2012-34398, BFU2015-69717-P (partially funded by FEDER funds of the EU), Ramón y Cajal Fellowship, ref. RYC-2011-08391, the European FP People: Marie Curie Action: Career Integration Grant, ref. 304111 and the CERCA Program/Generalitat de Catalunya. M.R.K. was supported by DFG CRC779 project B08, EU-JPND STAD and Leibniz Foundation SAW. P.De C. was supported by NIH NS36251; E.D.G and D.C.D were supported by DFG CRC779 project B09. N.B. was supported by the German Federal Ministry of Education and Research (ERA-NET Neuron Synpathy) and an ERC Advanced grant from the EU (SynPrime). M.V. was supported by an ERC Advanced grant from the EU (ERC-ADG-322966-DCVfusion).

Author contributions

- Designed the study: G.F., S.E.H., F.K., P.v.N., P.D.T, A.B.S. and M.V.
- Designed ontologies and reached consensus: all authors
- Implemented ontologies and evidence in GO, GO-training and quality control: R.F. B.K., R.L., H.M., P.G. and D.O-S.
- Annotated synaptic genes (>50): M.A-A, J.J.E.C., T.C., L.N.C., R.J.F., H.L.G., P.S.McPh, C.I., A.P.H. de J., H.J., M.K., N.L., H.MacG., P.v.N., M.N., V.O'C , R.P., K-H.S., R.F.G.T., C.V., R.R.V. and J.v.W.
- Supervised annotations: C.B., À.B., T.B., N.B., J.J.E.C., D.C.D., E.D.G., C.H., R.L.H., R.J., P.S.K., E.K., M.R.K., P.S.McPh., V.O'C, T.A.R. and C.S.
- Performed annotation QC: F.K. and P.v.N.
- Performed phylogenetic annotation: M.F., H.M., P.G.
- Performed bioinformatics analyses: A.B., D.P.H., F.K., H.T., K.T. and K.W.
- Supervised bioinformatics analyses: B.M.N., D.P., P.D.T., A.B.S, and M.V.
- Designed and built SynGO portal: F.K., with input of P.v.N., A.B.S and M.V.
- Generated figures: F.K. with input of A.B., D.P.H, P.v.N., K.T. and K.W.
- Wrote the paper: M.V., with input of T.C.B., F.K., A.B.S., P.D.T. and all expert laboratories

Conflict of Interest Statement

The authors declare no competing interests. M.S. and C.H. were employees of Genentech, a member of the Roche Group. S.E.H. serves on the Boards of Voyagers Therapeutics and Q-State Biosciences and on the scientific advisory boards of Janssen and BlackThorn.

Materials and Methods

Synaptic gene ontologies and integration into GO

Ontology terms in SynGO v1.0 were compared to pre-existing synaptic ontologies in the GO database prior to the starting date of SynGO (2015-01-01). A snapshot of the GO database representing the state at 2015-01-01 was obtained from `http://purl.obolibrary.org/obo/go/releases/2014-12-22/go.obo` (the last release in 2014) and converted into a directed graph using the iGraph R package (http://igraph.org). To construct the CC and BP graphs in Fig 2 we first created a tree from the SynGO v1.0 ontologies and classified terms that were present in the GO snapshot as 'reused'. Next, pre-existing synapse related terms that were not used by SynGO, indicated as purple nodes in Fig 2, were defined as subclassifiers of these 'reused' terms within the GO snapshot. Finally, we restricted resulting terms to match the scope of SynGO v1.0 (typical glutamatergic and GABA-ergic synapses). Terms that further specialize parent terms into serotonergic-, dopaminergic-, cholinergic-synapses, neuromuscular junctions, or 'regulation of' terms, were not taken into account in this evaluation of candidate terms for re-use by SynGO. Graphs in Fig 2 were visualized using a force-directed layout algorithm in Cytoscape (Shannon *et al.,* 2003).

SynGO ontologies and annotations were integrated into the existing ontologies within the GO database and will continuously be updated as the SynGO project expands synaptic ontologies and adds annotations in the future. These GO ontologies are available in the 'goslim_synapse' subset, its most recent version is always available at http://purl.obolibrary.org/obo/go/subsets/goslim_synapse.obo. Respective SynGO annotations are translated when exported to GO, e.g., annotations against 'process in the presynapse' are stored in GO as 'biological_process(GO:0008150) occurs_in presynapse(GO:0098793)'. The identifier of such terms that only exist in SynGO starts with "SYNGO:", whereas terms also available in GO have identifiers that start with "GO:" (as seen in the SynGO terms list in Table S2). SynGO annotations as integrated into GO are available through existing GO tools and websites, analysis on the SynGO subset is possible by filtering for annotations with the 'contributor=SynGO' property. All data from the SynGO consortium together with purpose-built analysis tools and community engagement are available through the SynGO website at https://syngoportal.org.

Gene expression data

The "brain-expressed" control set consists of genes that were expressed in significantly higher levels in brain compared to other tissues in Genotype Tissue Expression Consortia (GTEx) data (Ganna *et al.,* 2016). The control set with "brain topN" was defined as the N highest expressed genes in brain, where N was set to the number of unique genes annotated in SynGO v1.0. The highest expressed genes were computed by ranking the average gene-expression levels (in RPKM) from all brain samples in GTEx (G. T. Consortium *et al.,* 2017) version 6 (GTEx_Analysis_v6_RNA-seq_RNA-SeQCv1.1.8_gene_rpkm.gct.gz). For the brain enrichment analysis of synaptic genes in Fig S8 we computed the mean fold change comparing brain to all other tissues for each gene in the GTEx (version 7) data set. To examine enrichment, we applied a Wilcoxon Rank-Sum test for each SynGO ontology containing at least 5 genes. We used a one-sided hypothesis test in order to test whether the genes in the annotation are more brain expressed than expected under the null.

Gene features

Gene features described in Fig 3 and S4 were extracted from the BioMart (Smedley *et al.,* 2015) Ensembl Human genes GRCh38.p12 dataset and the Ensembl REST API Endpoints (release 95). Total gene length was computed using the `start_position` and `end_position` BioMart attributes (gene start and end, in base pairs). All known splice variants per gene were obtained through BioMart, from which the number of protein coding splice variants were counted using the `transcript_biotype` attribute. cDNA length was extracted from gene sequences provided through the Ensembl REST API with `'mask_feature=1'` parameter, and analogously all transcript exonic and intronic regions were obtained.

Isoform counts from full-length RNA sequencing

From our recent publication (Gupta *et al.,* 2018) we isolated full-length long reads that were expressed in neuronal subtypes, namely external granular layer neurons, internal granular layer neurons and Purkinje cells and had been attributed to a spliced protein coding gene. Subsequently, we considered only genes that had 20 or more such reads and split this gene list into two subsets: those annotated in SynGO and its complement. These groups differed substantially in the number of reads per gene. In order to normalize this, we randomly selected 10 full-length reads for each gene, resulting in two gene lists (SynGO and non-SynGO) with exactly 10 reads each. We then counted the number of distinct isoforms that these 10 reads described for each gene and repeated this subsampling process 1000 times.

Conservation of synaptic genes

Cumulative distribution of genes by gene age: Gene trees, covering ~95% of human genes, were obtained from the PANTHER resource (Mi *et al.,* 2018). Gene duplication events were dated relative to the earliest speciation node descending from the duplication. Trees were then pruned to contain only human paralogs, and the root of the tree (this ensures that fractional gene counts will add up to the total number of human genes). Each human gene was then traced back through the pruned tree to the root of the tree, and the number of branches was counted; this gives the total number of duplications (plus one, for the root) along the path to the root. Then, for each human gene, for each duplication (and root node) along the path from the gene to the root, a fractional count of 1/total was added to the count of genes that evolved at the date of that node. This process yields a count of human genes gained over each period of evolution, including gene duplication events. Estimated speciation times were taken from the TimeTree resource(Kumar *et al.,* 2017). The tree of CPT1C-related genes was obtained from the PANTHER website and can be accessed, together with additional information about the sequences and a multiple sequence alignment, at `http://pantherdb.org/treeViewer/treeViewer.jsp?book=PTHR22589&species=agr`. For enrichment analysis of synaptic genes at different periods of evolution, we extracted reconstructed ancestral genomes from the Ancestral Genomes resource [PMID: 30371900], and used the set of human ”proxy genes” for each ancestral gene. The specific ancestral genomes were obtained from the following URLs:

- `http://ancestralgenomes.org/species/genes/` `(list:genes/Metazoa-Choanoflagellida/Homo%20sapiens)`
- `http://ancestralgenomes.org/species/genes/` `(list:genes/Bilateria/Homo%20sapiens)`
- `http://ancestralgenomes.org/species/genes/` `(list:genes/Craniata-Cephalochordata/Homo%20sapiens)`
- `http://ancestralgenomes.org/species/genes/` `(list:genes/Euteleostomi/Homo%20sapiens)`

For each ontology term we applied a 1-sided Fisher exact test with 'greater than' hypothesis to compare genes only found in the 'after' set with all genes in the 'before' set. To find enriched terms within the entire SynGO ontology, we first selected the most specific term where each 'gene cluster' (unique set of genes) is found and then applied multiple testing correction using False Discovery Rate (FDR) on the subset of terms that contain these 'gene clusters'. For human-C. elegans and human-D. melanogaster orthologs, we used the "ancestral genome comparison" functions available in the Ancestral Genomes resource, to obtain the genes in each genome (e.g. human) that descend from each gene in the bilaterian common ancestor ("inparalogs"). We used this information to match up inparalog groups in the two genomes being compared, to obtain sets of orthologs between those genomes; e.g. the inparalog group of human gene(s) that descend from a given bilaterian ancestral gene are all orthologs of the inparalog group of C. elegans gene(s) that descend from that same ancestral gene. We classified each ortholog set as either 1:1, 1:many, many:1 or many:many depending on the number of inparalogs in each organism (i.e. whether there were gene duplications after speciation). We then calculated the proportion of genes (either all genes, or only SynGO genes, with at least one ortholog between human and a given model organism) that are in each type of ortholog set.

Large scale protein-protein interaction data

StringDB (Szklarczyk *et al.,* 2015) 10.5 human interactions were filtered by combined score (700, high confidence) and experimental evidence (400, medium confidence). StringDB PPIs then were matched to SynGO HGNC annotated genes by gene symbol, or alternative names for cases without a match. The distance between a pair of SynGO genes was defined as their path distance. For the CC model, the path distance between a membrane term and it's integral, anchored or extrinsic sub-classes (e.g., from SV membrane to anchored component of SV membrane) was set to zero. For the null distribution we computed all path distances within the CC or BP graph between any pair of all SynGO genes.

Proteomics of synaptic fractions

Proteins identified in selected proteomics studies shown in Fig 6 were mapped to human gene identifiers (HGNC) using the `https://www.uniprot.org` ID mapping service and mapping tables provided through `https://www.genenames.org`

(Table S4). Keratins were considered an external contaminant and therefore excluded from downstream analysis. The Venn diagram was generated using the ‘eulerr’ R package.

GWAS datasets

GWAS summary statistics for 8 traits were collected from the following resources; ADHD (Martin *et al.,* 2018), Autism Spectrum Disorder, Bipolar Disorder (Bipolar *et al.,* 2018) and Major Depressive Disorder(Wray *et al.,* 2018) from `https://www.med.unc.edu/pgc/results-and-downloads`, Educational Attainment(J. J. Lee *et al.,* 2018) from `https://www.thessgac.org/data`, Height (Wood *et al.,* 2014) from `https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files`, Intelligence (Savage *et al.,* 2018) from `https://ctg.cncr.nl/software/summary_statistics`, Schizophrenia (Pardinas *et al.,* 2018) from `http://walters.psycm.cf.ac.uk/`.

Magma gene-set analysis

First MAGMA gene analysis (de Leeuw *et al.,* 2015) was performed using the 1000 Genome Phase3 reference panel for European population by assigning SNPs to genes within a 2kb upstream and 1kb downstream window for 20,319 genes. The default model (SNP-wide mean) was used. Then MAGMA gene-set analyses were then performed for SynGO and original synaptic GO terms. For SynGO, one additional set with all SynGO genes was added, and in total 154 terms with at least 5 annotated (unique) genes were tested. For original GO, 5 additional sets; all synaptic genes, all BP genes, all CC genes, presynapse and postsynapse were added, and in total 96 terms with at least 5 annotated (unique) genes were tested. The gene set analyses were performed with the following three conditions for each trait: 1) no additional covariate, 2) conditioning on brain and average expression across all tissue types based on GTEx v7 RNA-seq dataset (G. T. Consortium *et al.,* 2017), 3) conditioning on brain and average expression, and the level of conservation of the genes. GTEx v7 RNA-seq data was obtained from `https://gtexportal.org`. The homology conservation scores in Fig S10 represent the level of conservation of genes, measured by the number of species with homolog genes using 65 species available through BioMart. Bonferroni correction was performed for each analysis separately (Pbon=0.05/154 for SynGO and 0.05/96 for GO). Statistical results are available in Table S6.

LDSC geneset analysis

To assess the contribution of each SynGO term to disease/phenotype heritability, we applied Stratified LD-Score Regression (S-LDSC) (Finucane *et al.,* 2015; Gazal *et al.,* 2017) to binary gene set annotations constructed with a ±100KB window around each gene as done in previous work (Finucane *et al.,* 2015; Zhu & Stephens, 2018). In our analyses, we conditioned on the 75 functional annotations in the baseline-LD model (Gazal *et al.,* 2017), an annotation containing all 23,987 protein-coding genes with a ±100KB window, as well as brain-enriched genes (see above), and a continuous annotation representing the conservation score of each gene. For each gene set from SynGO or pre-existing synaptic GO annotations, we assessed the statistical significance of the gene set annotations standardized effect size $\tau*$, (defined as the proportionate change in per-SNP heritability associated to a one standard deviation increase in the value of the annotation, conditioned on other annotations included in the model (Gazal *et al.,* 2017)) based on Bonferroni correction. Statistical results are available in Table S6.

Data and software availability

All data from the SynGO consortium together with purpose-built analysis tools and community engagement are available through the SynGO website at `https://syngoportal.org`.

References

1. Abdou, K. *et al.* Synapse-Specific Representation of the Identity of Overlapping Memory Engrams. *Science* **360,** 1227–1231. ISSN: 1095-9203 (Electronic) 0036-8075 (Linking) (2018).
2. Abul-Husn, N. S. *et al.* Systems Approach to Explore Components and Interactions in the Presynapse. *Proteomics* **9,** 3303–15. ISSN: 1615-9861 (Electronic) 1615-9853 (Linking) (2009).
3. Arnsten, A. F., Wang, M. J. & Paspalas, C. D. Neuromodulation of Thought: Flexibilities and Vulnerabilities in Prefrontal Cortical Network Synapses. *Neuron* **76,** 223–39. ISSN: 1097-4199 (Electronic) 0896-6273 (Linking) (2012).
4. Ashburner, M. *et al.* Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium. *Nat Genet* **25,** 25–9. ISSN: 1061-4036 (Print) 1061-4036 (Linking) (2000).
5. Bayes, A., Collins, M. O., *et al.* Comparative Study of Human and Mouse Postsynaptic Proteomes Finds High Compositional Conservation and Abundance Differences for Key Synaptic Proteins. *PLoS One* **7,** e46683. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking) (2012).
6. Bayes, A., van de Lagemaat, L. N., *et al.* Characterization of the Proteome, Diseases and Evolution of the Human Postsynaptic Density. *Nat Neurosci* **14,** 19–21. ISSN: 1546-1726 (Electronic) 1097-6256 (Linking) (2011).
7. Biesemann, C. *et al.* Proteomic Screening of Glutamatergic Mouse Brain Synaptosomes Isolated by Fluorescence Activated Sorting. *EMBO J* **33,** 157–70. ISSN: 1460-2075 (Electronic) 0261-4189 (Linking) (2014).
8. Bipolar, D. *et al.* Genomic Dissection of Bipolar Disorder and Schizophrenia, Including 28 Subphenotypes. *Cell* **173,** 1705–1715 e16. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking) (2018).
9. Boda, B., Dubos, A. & Muller, D. Signaling Mechanisms Regulating Synapse Formation and Function in Mental Retardation. *Curr Opin Neurobiol* **20,** 519–27. ISSN: 1873-6882 (Electronic) 0959-4388 (Linking) (2010).

10. Bourgeron, T. From the Genetic Architecture to Synaptic Plasticity in Autism Spectrum Disorder. *Nat Rev Neurosci* **16,** 551–63. ISSN: 1471-0048 (Electronic) 1471-003X (Linking) (2015).

11. Boyken, J. *et al.* Molecular Profiling of Synaptic Vesicle Docking Sites Reveals Novel Proteins but Few Differences between Glutamatergic and GABAergic Synapses. *Neuron* **78,** 285–97. ISSN: 1097-4199 (Electronic) 0896-6273 (Linking) (2013).

12. Bulik-Sullivan, B. K. *et al.* LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies. *Nat Genet* **47,** 291–5. ISSN: 1546-1718 (Electronic) 1061-4036 (Linking) (2015).

13. Calvo, S. E., Clauser, K. R. & Mootha, V. K. MitoCarta2.0: An Updated Inventory of Mammalian Mitochondrial Proteins. *Nucleic Acids Res* **44,** D1251–7. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (2016).

14. Chang, R. Y. *et al.* SWATH Analysis of the Synaptic Proteome in Alzheimer's Disease. *Neurochem Int* **87,** 1–12. ISSN: 1872-9754 (Electronic) 0197-0186 (Linking) (2015).

15. Collins, M. O. *et al.* Molecular Characterization and Comparison of the Components and Multiprotein Complexes in the Postsynaptic Proteome. *J Neurochem* **97 Suppl 1,** 16–23. ISSN: 0022-3042 (Print) 0022-3042 (Linking) (2006).

16. Consortium, G. T. *et al.* Genetic Effects on Gene Expression across Human Tissues. *Nature* **550,** 204–213. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (2017).

17. Consortium, U. UniProt: The Universal Protein Knowledgebase. *Nucleic acids research* **46,** 2699 (2018).

18. de Leeuw, C. A. *et al.* MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLoS Comput Biol* **11,** e1004219. ISSN: 1553-7358 (Electronic) 1553-734X (Linking) (2015).

19. De Rubeis, S. *et al.* Synaptic, Transcriptional and Chromatin Genes Disrupted in Autism. *Nature* **515,** 209–15. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (2014).

20. Demontis, D. *et al.* Whole-Exome Sequencing Reveals Increased Burden of Rare Functional and Disruptive Variants in Candidate Risk Genes in Individuals with Persistent Attention-Deficit/Hyperactivity Disorder. *J Am Acad Child Adolesc Psychiatry* **55,** 521–3. ISSN: 1527-5418 (Electronic) 0890-8567 (Linking) (2016).

21. Domazet-Lošo, T., Brajković, J. & Tautz, D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in Genetics* **23,** 533–539 (2007).

22. Emes, R. D. *et al.* Evolutionary Expansion and Anatomical Specialization of Synapse Proteome Complexity. *Nat Neurosci* **11,** 799–806. ISSN: 1097-6256 (Print) 1097-6256 (Linking) (2008).

23. Fado, R. *et al.* Novel Regulation of the Synthesis of Alpha-Amino-3-Hydroxy-5-Methyl-4-Isoxazolepropionic Acid (AMPA) Receptor Subunit GluA1 by Carnitine Palmitoyltransferase 1C (CPT1C) in the Hippocampus. *J Biol Chem* **290,** 25548–60. ISSN: 1083-351X (Electronic) 0021-9258 (Linking) (2015).

24. Filiou, M. D. *et al.* Profiling of Mouse Synaptosome Proteome and Phosphoproteome by IEF. *Electrophoresis* **31,** 1294–301. ISSN: 1522-2683 (Electronic) 0173-0835 (Linking) (2010).

25. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics* **47,** 1228 (2015).

26. Fromer, M. *et al.* De Novo Mutations in Schizophrenia Implicate Synaptic Networks. *Nature* **506,** 179–84. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (2014).

27. Ganna, A. *et al.* Ultra-Rare Disruptive and Damaging Mutations Influence Educational Attainment in the General Population. *Nat Neurosci* **19,** 1563–1565. ISSN: 1546-1726 (Electronic) 1097-6256 (Linking) (2016).

28. Gaudet, P. *et al.* Phylogenetic-Based Propagation of Functional Annotations within the Gene Ontology Consortium. *Brief Bioinform* **12,** 449–62. ISSN: 1477-4054 (Electronic) 1467-5463 (Linking) (2011).

29. Gazal, S. *et al.* Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature genetics* **49,** 1421 (2017).

30. Giglio, M. *et al.* ECO, the Evidence & Conclusion Ontology: Community Standard for Evidence Information. *Nucleic Acids Res.* ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (2018).

31. Grant, S. G. Synaptopathies: Diseases of the Synaptome. *Curr Opin Neurobiol* **22,** 522–9. ISSN: 1873-6882 (Electronic) 0959-4388 (Linking) (2012).

32. Groschner, L. N. *et al.* Dendritic Integration of Sensory Evidence in Perceptual Decision-Making. *Cell* **173,** 894–905 e13. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking) (2018).

33. Grove, J. *et al.* Identification of Common Genetic Risk Variants for Autism Spectrum Disorder. *Nature genetics* **51,** 431–444 (2019).

34. Gupta, I. *et al.* Single-Cell Isoform RNA Sequencing Characterizes Isoforms in Thousands of Cerebellar Cells. *Nat Biotechnol.* ISSN: 1546-1696 (Electronic) 1087-0156 (Linking) (2018).

35. Heutink, P. & Verhage, M. Neurodegeneration: New Road Leads Back to the Synapse. *Neuron* **75,** 935–8. ISSN: 1097-4199 (Electronic) 0896-6273 (Linking) (2012).

36. Hong, S. *et al.* Complement and Microglia Mediate Early Synapse Loss in Alzheimer Mouse Models. *Science* **352,** 712–716. ISSN: 1095-9203 (Electronic) 0036-8075 (Linking) (2016).

37. Huang, K. Y. *et al.* dbPTM 2016: 10-Year Anniversary of a Resource for Post-Translational Modification of Proteins. *Nucleic Acids Res* **44,** D435–46. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (2016).

38. Jeanquartier, F., Jean-Quartier, C. & Holzinger, A. Integrated Web Visualizations for Protein-Protein Interaction Databases. *BMC Bioinformatics* **16,** 195. ISSN: 1471-2105 (Electronic) 1471-2105 (Linking) (2015).

39. Kandel, E. R. The Molecular Biology of Memory Storage: A Dialogue between Genes and Synapses. *Science* **294,** 1030–8. ISSN: 0036-8075 (Print) 0036-8075 (Linking) (2001).

40. Karczewski, K. J. *et al.* The ExAC Browser: Displaying Reference Data Information from over 60 000 Exomes. *Nucleic Acids Res* **45,** D840–D845. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (2017).

41. Krogh, A. *et al.* Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes. *J Mol Biol* **305,** 567–80. ISSN: 0022-2836 (Print) 0022-2836 (Linking) (2001).

42. Kumar, S. *et al.* TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol* **34,** 1812–1819. ISSN: 1537-1719 (Electronic) 0737-4038 (Linking) (2017).

43. Lee, J. J. *et al.* Gene Discovery and Polygenic Prediction from a Genome-Wide Association Study of Educational Attainment in 1.1 Million Individuals. *Nat Genet* **50,** 1112–1121. ISSN: 1546-1718 (Electronic) 1061-4036 (Linking) (2018).

44. Lek, M. *et al.* Analysis of Protein-Coding Genetic Variation in 60,706 Humans. *Nature* **536,** 285–91. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (2016).

45. Lips, E. S. *et al.* Functional Gene Group Analysis Identifies Synaptic Gene Groups as Risk Factor for Schizophrenia. *Mol Psychiatry* **17,** 996–1006. ISSN: 1476-5578 (Electronic) 1359-4184 (Linking) (2012).

46. Martin, J. *et al.* A Genetic Investigation of Sex Bias in the Prevalence of Attention-Deficit/Hyperactivity Disorder. *Biol Psychiatry* **83,** 1044–1053. ISSN: 1873-2402 (Electronic) 0006-3223 (Linking) (2018).

47. Mattheisen, M. *et al.* Genome-Wide Association Study in Obsessive-Compulsive Disorder: Results from the OCGAS. *Mol Psychiatry* **20,** 337–44. ISSN: 1476-5578 (Electronic) 1359-4184 (Linking) (2015).

48. Mi, H. *et al.* PANTHER Version 14: More Genomes, a New PANTHER GO-Slim and Improvements in Enrichment Analysis Tools. *Nucleic Acids Res.* ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (2018).

49. Mi, H. *et al.* PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic acids research* **47,** D419–D426 (2019).

50. Moczulska, K. E. *et al.* Deep and Precise Quantification of the Mouse Synaptosomal Proteome Reveals Substantial Remodeling during Postnatal Maturation. *J Proteome Res* **13,** 4310–24. ISSN: 1535-3907 (Electronic) 1535-3893 (Linking) (2014).

51. Monday, H. R. & Castillo, P. E. Closing the Gap: Long-Term Presynaptic Plasticity in Brain Function and Disease. *Curr Opin Neurobiol* **45,** 106–112. ISSN: 1873-6882 (Electronic) 0959-4388 (Linking) (2017).

52. Morciano, M. *et al.* Immunoisolation of Two Synaptic Vesicle Pools from Synaptosomes: A Proteomics Analysis. *J Neurochem* **95,** 1732–45. ISSN: 0022-3042 (Print) 0022-3042 (Linking) (2005).

53. Pardinas, A. F. *et al.* Common Schizophrenia Alleles Are Enriched in Mutation-Intolerant Genes and in Regions under Strong Background Selection. *Nat Genet* **50,** 381–389. ISSN: 1546-1718 (Electronic) 1061-4036 (Linking) (2018).

54. Pedroso, I. *et al.* Common Genetic Variants and Gene-Expression Changes Associated with Bipolar Disorder Are over-Represented in Brain Signaling Pathway Genes. *Biol Psychiatry* **72,** 311–7. ISSN: 1873-2402 (Electronic) 0006-3223 (Linking) (2012).

55. Petersen, C. C. & Crochet, S. Synaptic Computation and Sensory Processing in Neocortical Layer 2/3. *Neuron* **78,** 28–48. ISSN: 1097-4199 (Electronic) 0896-6273 (Linking) (2013).

56. Phillips, G. R. *et al.* Proteomic Comparison of Two Fractions Derived from the Transsynaptic Scaffold. *J Neurosci Res* **81,** 762–75. ISSN: 0360-4012 (Print) 0360-4012 (Linking) (2005).

57. Psychiatric, G. C. B. D. W. G. Large-Scale Genome-Wide Association Analysis of Bipolar Disorder Identifies a New Susceptibility Locus near ODZ4. *Nat Genet* **43,** 977–83. ISSN: 1546-1718 (Electronic) 1061-4036 (Linking) (2011).

58. Ripolles, P. *et al.* Intrinsically Regulated Learning Is Modulated by Synaptic Dopamine Signaling. *Elife* **7.** ISSN: 2050-084X (Electronic) 2050-084X (Linking) (2018).

59. Roy, M. *et al.* Proteomic Analysis of Postsynaptic Proteins in Regions of the Human Neocortex. *Nat Neurosci* **21,** 130–138. ISSN: 1546-1726 (Electronic) 1097-6256 (Linking) (2018).

60. Ruano, D. *et al.* Functional Gene Group Analysis Reveals a Role of Synaptic Heterotrimeric G Proteins in Cognitive Ability. *Am J Hum Genet* **86,** 113–25. ISSN: 1537-6605 (Electronic) 0002-9297 (Linking) (2010).

61. Savage, J. E. *et al.* Genome-Wide Association Meta-Analysis in 269,867 Individuals Identifies New Genetic and Functional Links to Intelligence. *Nat Genet* **50,** 912–919. ISSN: 1546-1718 (Electronic) 1061-4036 (Linking) (2018).

62. Selkoe, D. J. Alzheimer’s Disease Is a Synaptic Failure. *Science* **298,** 789–91. ISSN: 1095-9203 (Electronic) 0036-8075 (Linking) (2002).

63. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res* **13,** 2498–504. ISSN: 1088-9051 (Print) 1088-9051 (Linking) (2003).

64. Smedley, D. *et al.* The BioMart Community Portal: An Innovative Alternative to Large, Centralized Data Repositories. *Nucleic Acids Res* **43,** W589–98. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (2015).

65. Smith, A. C. & Robinson, A. J. MitoMiner v4.0: An Updated Database of Mitochondrial Localization Evidence, Phenotypes and Diseases. *Nucleic Acids Res.* ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (2018).

66. Soukup, S. F., Vanhauwaert, R. & Verstreken, P. Parkinson's Disease: Convergence on Synaptic Homeostasis. *EMBO J* **37.** ISSN: 1460-2075 (Electronic) 0261-4189 (Linking) (2018).

67. Spires-Jones, T. L. & Hyman, B. T. The Intersection of Amyloid Beta and Tau at Synapses in Alzheimer's Disease. *Neuron* **82,** 756–71. ISSN: 1097-4199 (Electronic) 0896-6273 (Linking) (2014).

68. Sudhof, T. C. Neuroligins and Neurexins Link Synaptic Function to Cognitive Disease. *Nature* **455,** 903–11. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (2008).

69. Szklarczyk, D. *et al.* STRING V10: Protein-Protein Interaction Networks, Integrated over the Tree of Life. *Nucleic Acids Res* **43,** D447–52. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (2015).

70. Thapar, A. *et al.* Psychiatric Gene Discoveries Shape Evidence on ADHD's Biology. *Mol Psychiatry* **21,** 1202–7. ISSN: 1476-5578 (Electronic) 1359-4184 (Linking) (2016).

71. The Gene Ontology, C. The Gene Ontology Resource: 20 Years and Still GOing Strong. *Nucleic Acids Res.* ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (2018).

72. UniProt, C. UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (2018).

73. Wood, A. R. *et al.* Defining the Role of Common Variation in the Genomic and Biological Architecture of Adult Human Height. *Nat Genet* **46,** 1173–86. ISSN: 1546-1718 (Electronic) 1061-4036 (Linking) (2014).

74. Wray, N. R. *et al.* Genome-Wide Association Analyses Identify 44 Risk Variants and Refine the Genetic Architecture of Major Depression. *Nat Genet* **50,** 668–681. ISSN: 1546-1718 (Electronic) 1061-4036 (Linking) (2018).

75. Yates, B. *et al.* Genenames.Org: The HGNC and VGNC Resources in 2017. *Nucleic Acids Res* **45,** D619–D625. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (2017).

76. Zhu, X. & Stephens, M. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nature communications* **9,** 1–14 (2018).

77. Zwir, I. *et al.* Uncovering the Complex Genetics of Human Temperament. *Mol Psychiatry.* ISSN: 1476-5578 (Electronic) 1359-4184 (Linking) (2018).

7

Discussion

Discussion

The general aim of this thesis is the large-scale identification of synaptic proteins through interdisciplinary research that combines synapse research, proteomics and bioinformatics.

The first two chapters describe innovations in mass spectrometry data acquisition and analysis. A censoring model coined EBRCT is introduced in chapter 2 that improves differential expression analysis in the presence of many missing values. Chapter 3 explores an alternative mode of operation for mass spectrometers to perform data independent acquisition, coined WiSIM-DIA. Mutually exclusive advantages are discovered in a comparison with the commonly used SWATH-MS approach. Chapter 4 applies proteomics to a series of biochemical subfractions of the synapse to discover novel protein constituents. Synapses from rodent and primate species are compared using quantitative proteomics in chapter 5, revealing proteins that exhibit small changes in protein expression. Finally, chapter 6 presents the results of a community effort to describe the synapse (SynGO), by annotating the location and function of its constituents based on published literature and using this information for bioinformatics analyses of gene metadata and links between synapses and brain disorders.

Novel proteomics approaches to uncover the synaptome

Advances in label-free proteomics

Mass spectrometry is the key technology behind modern proteomics analysis. Sensitivity and dynamic range are limiting factors in mass spectrometry driven proteomics that reduce the probability of detection and signal quality for low abundant protein species (Karpievitch *et al.,* 2009; Michalski *et al.,* 2011). Research in this thesis focuses on the synapse. To enable studies on an increasingly larger proportion of the synaptic proteome and work with increasingly specific subpopulation of synapses I aimed at advancing proteomics data acquisition and analysis. When biochemically isolating synapses, it is estimated that each microgram of a synaptosomal preparation contains 9̃.95 million synapses (Wilhelm *et al.,* 2014). Because routine experiments on state-of-the-art mass spectrometers use input material in the order of 500ng (Bekker-Jensen *et al.,* 2020). Consequently, the output signal in mass spectrometric analyses of synaptosomes averages over millions of synapses. One can get by with less input material at the cost of only seeing the top-most abundant proteins. This can be a problem if the phenotype in a study only affects a specific subpopulation of synapses because signal from the phenotype would be 'diluted'.

Missingness in label-free proteomics One approach to improving proteomics near the detection limit, that I studied in this thesis, considers models of missingness that deal with 'missing data' mostly observed for peptides with low signal intensity. Chapter 2 describes the Empirical Bayesian Random Censoring Threshold (EBRCT) model which improves the detection of differentially abundant proteins as compared to alternative approaches. At the time of publication this model showed modest gains on the benchmarked dataset as compared to alternative methods.

However, the mass spectrometry field has progressed steadily and therefore this dataset does not represent the current state of affairs. In particular, improvements in mass-spec hardware and the introduction of both DIA and improved match-between-runs algorithms, the number of missing values in datasets generated in recent times has become much lower.

Quantitative datasets with lower signal-to-noise ratio in combination with algorithms that infer missing values from the actual mass-spec data, instead of only using patterns in abundances (censoring model), have proven to be an effective strategy. For instance, the data independent acquisition (DIA) dataset described in chapter 3 has only <0.1% missing values and state-of-the-art match-between-runs approaches for DDA datasets show <0.2% missing values (Shen *et al.,* 2018).

With few missing values, the choice of algorithm one uses to fill these gaps in the dataset has hardly any effect on the overall outcome. Thus, simply treating unobserved data as Missing Completely At Random (MCAR), which is free of any assumption on missing values and computationally inexpensive, is now commonplace for state-of-the-art datasets with few missing values. Future research that marry both approaches would be interesting; for instance, by using informed censoring models to estimate whether the match-between-runs 'imputations from raw data' are unlikely values (potential technical errors).

Alternative acquisition strategy for DIA Another approach to expand the set of quantifiable peptides in proteomics is described in chapter 3. A novel precursor ion-centric approach, WiSIM-DIA, is introduced which combines the traditional DDA strategy (that uses MS1 for label-free quantification) with SWATH-MS. High resolution MS1 data is acquired in a few sequentially isolated m/z windows, spending a relatively large amount of resources (measurement time) to optimize precursor peak area quality for label-free quantification. Peptides are fragmented analogous to SWATH-MS, but as a tradeoff consequential to the MS1-focussed strategy at slightly lower quality.

Interestingly, the dataset generated revealed that, whereas regular SWATH-MS outperformed WiSIM-DIA, there is no correlation between the signal-to-noise ratio (S/N) of the exact same peptides measures by both approaches while the overall S/N distributions of each approach are similar (chapter 3, Figure 3A). This suggests there is a subpopulation of peptides that are quantified with higher signal quality in either WiSIM-DIA or DIA, indicating mutually exclusive benefits to each approach. Huang *et al.* (2020) recently confirmed that it is indeed beneficial to consider both MS1 and MS2 quantitative data in DIA approaches. Future research could extend this to a computational model that integrates not only quantitative signals from MS1 and MS2, but also takes respective confidence scores into account in spirit of Käll *et al.* (2019). A low prior confidence should be placed in low abundant peptide features and peptides with high variation among replicate measurements. Finally, an informed censoring model could be integrated following our observations in chapter 2 and discussion in the previous section.

Current state of the art for MS1-focussed quantification strategies is a refined execution of the approach introduced in chapter 3; the BoxCar approach uses many MS1 m/z windows in combination with a spectral library (Meier, Geyer, *et al.,* 2018).

Resulting quantification accuracy and coverage of the proteome is unprecedented at 10.000 proteins quantified in mouse cerebellum tissue within 100 minutes, suggesting that MS1 centric approaches have a lot of potential as the mass accuracy and quality of HPLC separation in novel mass spectrometers advances.

Future steps in MS data acquisition methods With the advent of mass spectrometers that can be controlled during operation by custom software, new avenues to data acquisition have been opened. For example, recently introduced real-time software can analyze mass spectrometry spectra during acquisition, running a search engine live to obtain peptide quality metrics, and then directly control whether the respective peptide fragments are selected for further analysis in MS3 (Schweppe *et al.,* 2020). While this approach was developed for the improvement of TMT labeling based quantification, it will be interesting to further explore adoptions to the label-free proteomics. For instance, in DDA mode triggering another MS/MS event for an eluting precursor depending on live search engine results (e.g. regarding ambiguous spectra and repetition with alternative collision energy).

Refining DIA acquisition methods using a more software involved approach in configuring the mass spectrometer should be possible. Earlier publications have demonstrated that multiplexing SWATH-MS windows improves the deconvolution of peptide signals (Amodei *et al.,* 2019). The variable-window approach that adapts the SWATH-MS acquisition to the sample complexity (in m/z space) also results in improved signal-to-noise ratios (Schilling *et al.,* 2017). Taken together, it will be interesting to configure more complex DIA acquisition patterns, in both mass-over-charge and retention time dimensions, in a data-driven approach in order to push the boundary of the level of detection and quantification.

Finally, recent innovations in the addition of ion mobility spectrometry to DIA have shown strong improvements in both sensitivity and quantification accuracy (Meier, Brunner, *et al.,* 2020). I expect this technology will find its way into future synaptome studies.

Proteomics to explore synaptic subcellular protein localization

Chapter 4 aims to discover novel synaptic proteins through bioinformatics applied to mass spectrometry analyses of synaptic subcellular compartments. Biochemical purification protocols yield crude enrichments of the targeted subcellular compartment of interest together with any biochemical impurities that exhibit similar biochemical properties (e.g., when using density gradients, impurities may simply have the same density as the target compartment). Past efforts have applied this approach to synaptic proteins; isolate biochemical subfractions enriched for synaptic proteins and then apply proteomics to find protein identities (K. W. Li *et al.,* 2003; Takamori *et al.,* 2006; Dosemeci *et al.,* 2007; Volknandt & Karas, 2012; Bayés *et al.,* 2012; Sialana *et al.,* 2016).

The approach put forth in chapter 4 does not rely on the protein identity list of the respective biochemical subfraction alone, but instead considers protein abundances relative to other subfractions. Whereas previous studies used relative enrichments to identify proteins associated with some subfraction compared to another (Boyken *et al.,* 2013), the approach in chapter 4 integrates data from many subfractions.

Using canonical PSD proteins as a reference, which should be relatively enriched in the PSD biochemical subfraction, candidate PSD proteins were ranked by their correlation profile over many synaptic subfractions. A major advantage over the traditional approach is that one can prioritize candidate proteins by some enrichment score, instead of a qualitative metric for identification (contaminants can be highly abundant and confidently identified as well).

The candidate protein list found through this bioinformatics approach indeed contains previously established PSD-enriched proteins among the top scoring proteins, validating the approach (chapter 4, Figure 5A). Interestingly, known false-positives established in the reference experiment, which was solely based on an affinity purification protocol, yielded low scores in our approach demonstrating the additional power of the experimental design combined with correlation-based data analysis (chapter 4, Figure 5C). Both candidate proteins from this proteomics study that were validated by high-resolution microscopy were later annotated in SynGO (chapter 6), contributing experimental evidence for underrepresented proteins to increase our understanding of the synaptic proteome.

Bioinformatics analyses of proteomics applied to synapses was also used in chapter 6, but in contrast to chapter 4 it is based on previously published datasets. There, I summarized results from 19 previous studies (chapter 6, Figure 6) that performed biochemical purification of synaptic subfractions, and contrasted the set of synaptic proteins put forth in each study against the set of confirmed synaptic proteins (SynGO, which was not available at the time of publishing chapter 4). From the relatively small proportion of SynGO confirmed synaptic proteins in each of these datasets (2̃5%) it is apparent that it is indeed challenging to delineate true- and false-positive synaptic protein constituents in these approaches that entirely rely on biochemical purity, which speaks to the approach taken in chapter 4. However, the proportion of synaptic proteins in these preparations may be underrepresented here because the SynGO knowledgebase is not (yet) complete and this analysis only considers protein identities (and not copy numbers). To put this into context, Wilhelm *et al.* (2014) estimated 5̃8% of particles observed in synaptosomal preparations (P2 fraction followed by a Ficoll gradient) by fluorescence microscopy are synaptosomes. Taken together, more elaborate experimental designs and/or further post-hoc analysis is needed to generate high confidence candidates from such biochemical preparations.

No experiment is perfect. However, while there will be false positives and false negatives in the approach to subcellular compartment interrogation discussed in chapter 4, an excellent signal to noise ratio was found in this application. It would be interesting to explore how this experimental design scales when combined with current state-of-the-art mass spectrometry and expansion of the biochemical subfractions used (e.g. generate PSD fractions in a series of 1̃0 increasing degrees of stringency for purification).

Much increased sensitivity in current technology will improve coverage for low abundant proteins and require less input material. The latter is particularly welcome for subcellular fractions that require strong biochemical enrichment such as the PSD. Further, working with less input material enables proteomic characterization of subfields of a brain region of interest (e.g. focus on CA1 instead of hippocampus as a whole). In contrast, the technology used in chapter 4 required pooling of hippocampi from multiple mice to obtain sufficient input material for mass spectrometric measurement. Repeating the experiments from chapter 4 at large scale with great technical ability is a painstaking effort that could yield a valuable resource for the scientific community. Analogous to the widely used large-scale in situ hybridization imaging data available in the Allen Brain Atlas, this could prove a useful reference for proteins of interest in everyday research and as input for bioinformatics intersection with other -omics datasets.

Proteomics applied to inter-species synapse comparison

In chapter 5 we used SWATH-MS to quantify levels of hippocampal synaptic proteins of four species, the rodents; mouse and rat, and the primates; marmoset and human. A major challenge for inter-species proteomic comparison stems from a mass spectrometry technicality; peptides with a distinct amino acid sequence exhibit different ionization properties. For example, 2 distinct peptides with the exact same copy numbers in a sample may yield different ion intensity counts. Consequentially, comparing the stoichiometry of a given protein between mouse and human is accurate only if the exact same peptide sequence is conserved and observable by mass spectrometry. In this study we therefore only used such conserved peptide sequences, a deliberate tradeoff between quantification accuracy and protein coverage, and achieved low technical variability.

This enabled reliable detection of many differentially abundant proteins between species, but mostly with small fold changes. These were overrepresented for proteins previously linked to synaptic plasticity (Rao-Ruiz *et al.,* 2015). Interestingly, we later learned through phylogenetic analyses of synaptic genes described in chapter 6 (Figure 4) that synaptic genes have been strongly conserved, for more than 100 million years prior to the emergence of rodents. So, it is not surprising that there are no huge changes across species and only minor fine-tuning of the synaptic machinery at this relatively late stage of human evolution.

Quantified proteins were mapped to various functional groups of interest to aid interpretation of observed species differences (chapter 5, Figure 3). Functionally related proteins generally exhibit correlated abundance patterns over species, confirming technical validity of the quantitative proteomics methodology. Technical merit aside, a major caveat in the interpretation of data from such experimental designs is that we cannot account for systematic differences between species induced during biochemistry. The mass-spec readout describes differences in the input biochemical preparations, which themselves are surrogates for native protein populations. The human brain generally features larger synapses and has different fatty acid composition as compared to rodents, consequently one cannot completely rule out that such physiological differences may have its effect on sample preparation.

Electron Microscopy (EM) introspection of synaptosomes prepared with similar protocols have in the past shown that for both mouse and human brain, intact synapses are purified, but a selection bias towards synaptic subpopulations or against biochemical impurities is very difficult to assess when only a limited set of synapses in the sample is visualized by EM in such studies (Schrimpf *et al.,* 2005; Biesemann *et al.,* 2014; Benito *et al.,* 2018; Tang *et al.,* 2015). Taken together, we have made every effort to account for technical issues in this species comparison, from wet-lab to quantitative proteomics methodology, in order to accurately reflect evolutionary differences in synaptic molecular machinery.

When performing overrepresentation analysis of the differentially abundant proteins using the Gene Ontology (GO) database, the SynGO knowledgebase (chapter 6) had not yet been created, and no enriched biological functions or subcellular compartments were found. Given that synaptic proteins of interest in this study were lacking GO annotation coverage at the time we instead created groups of functionally related proteins that are commonly studied in the synapse field to interpret species differences. We compared protein abundance profiles within these groups and successfully identified subsets of proteins with diverging abundance patterns.

Reflecting on this study it becomes apparent that the established SynGO project clearly addresses the need for functional interpretation of –omics datasets, saving individual research groups the trouble to collect the adequate literature data for interpretation. Interestingly, the functional protein group definitions that powered the data interpretation in this study do have some unique merits that stem from the fact that we mixed biological function and molecular function. For the latter, simply assuming that protein presence in synaptosome preparations indicates synaptic localization and combining this observation with known molecular function (e.g. from GO database) to define groups such as ‘motor proteins’. Such functional groups would not translate 1:1 to SynGO 1.0 due to stringent experimental evidence requirements for protein annotation and lack of SynGO ontology terms that directly translate to the groups discussed in chapter 5. Future enhancements of the SynGO annotation system, such as more extensive annotation models, might address the latter issue and are discussed below.

The synapse knowledgebase: SynGO

A solid foundation for future research of the synapse has been laid with the SynGO paper (chapter 6), which brings together piecewise literature-based knowledge of the synapse community in a comprehensive annotation framework to further our understanding of synapses. This interactive knowledgebase, available at `https://syngoportal.org`, accumulates available research about synapse biology using Gene Ontology (GO) annotations.

The SynGO knowledgebase was designed as a high quality resource at the scale of thousands of genes. By requiring hard evidence, and expert curation thereof, to establish synaptic localization for each annotated gene a quality standard was set. Annotations that solely rely on data from large-scale hypothesis generating experiments or non-neuronal model systems were not accepted. Consequently, these criteria demanded literature review and discussion of thousands of papers. The SynGO consortium, an assembly of a 15 expert synapse groups collaborating with the GO consortium, was uniquely positioned to rise to the challenge. We first established systematics to maximize the robustness and efficiency of the annotation and quality control workflows and then applied a divide-and-conquer approach in which we assigned sets of established and previously hypothesized candidate genes to their respective synapse domain experts for literature review, assessment of experimental evidence and eventual annotation. As a result, a comprehensive ontology of the synapse was designed and used for 2922 annotations against 1112 unique genes. The SynGO knowledgebase can serve as a universal reference in synapse biology but also enable bioinformatic interpretation of -omics data in synaptic context. For instance, over-representation analysis of proteomics/transcriptomics/genomics results to find synaptic locations/functions where statistical signal from the –omics study accumulates.

It has been surprisingly difficult to find high resolution localization evidence in literature for some proteins that have long been established as synaptic. For example, synaptic vesicle v-ATPase protein complexes have been identified decades ago (Jahn & Sudhof, 1994; Galli *et al.,* 1996) but high-resolution experimental evidence for localization of some of its subunits in synapses was lacking until recently (Wilhelm *et al.,* 2014; Abbas *et al.,* 2020). A drawback of the stringent criteria for inclusion in SynGO is that we were unable to annotate some proteins because evidence in synaptic context is lacking. For example, adaptor protein complex AP-2 is known to play an important role in clathrin mediated endocytosis (Haucke *et al.,* 2011; Kirchhausen *et al.,* 2014) but experimental evidence for specific subunit proteins in synaptic context is sparse, limiting SynGO annotation to only a fraction of all subunits. Experimental assays that describe the function of AP-2 subunits may use non-neuronal model systems (Kovtun *et al.,* 2020) or describe one subunit (Morgan *et al.,* 2000) while for other subunits no literature can be found. Some use model systems such as Caenorhabditis elegans (worm) were used to study AP-2 subunits (Gu *et al.,* 2013) and as a result, translation of these experimental results to mammalian synapses may not be immediately clear in case high-confidence localization evidence of said subunits in mammalian synapses has not been shown. Taken together, strict annotation guidelines, tracking experimental evidence and reluctance to "infer" synaptic annotations are key features that establish SynGO as a quality data resource, but also pause the inclusion of some well-known synaptic genes into the knowledgebase until experimental evidence in mammalian synaptic context becomes available.

Nearly all synaptic genes have evolved prior to the last common ancestor (LCA) of all vertebrates (circa 386 million years ago), which is much earlier than the set of all human genes (chapter 6, Figure 4). SynGO annotated genes were found to be structurally very different from the rest of the human genome and, to a lesser extent, the set of all brain expressed genes; synaptic genes are larger and more complex (chapter 6, Figure 3). So evolution of synaptic genes has not halted since the LCA of all vertebrates simply because we only find sparse accumulation of new synaptic genes since. Instead, the accumulation of more coding and non-coding sequences into synaptic genes may have served to expand their transcriptional regulatory repertoire and diversification of functions of the encoded proteins. A strong enrichment among genes associated with brain disorders was found for SynGO annotated genes (chapter 6, Figures 7 and 8). Further, synaptic genes proved exceptionally intolerant to mutations (chapter 6, Figure 5).

Towards completeness

As a first step within SynGO, we should work towards completeness. Not all synaptic genes are in the SynGO database yet, some may have been missed in the initial annotation efforts and at the same time there may be newly published experimental evidence for synaptic genes that was previously lacking. From a knowledge discovery perspective, this can be seen as a breadth-first approach where priority was placed on including as many known synaptic genes as possible. To achieve this, extended efforts are now required to annotate more of the pre-existing literature and keep this process going as more experimental evidence is published in the future, demanding additions and corrections to the knowledgebase as the synapse field advances.

For those proteins where only low confidence evidence is available (e.g. Western blot in a PSD biochemical fraction, or computationally inferred protein-protein-interactions that link a protein to the synapse), additional experimental data should be generated to clarify their location and/or function in the synapse. Resources permitting, an example of strengthening the knowledgebase is the addition of high-resolution microscopy for accurate subcellular localization of all genes currently lacking cellular component annotation with strong experimental evidence, or annotated against a low-resolution ontology term such as 'synapse'.

Another aspect is the inclusion of all synaptic locations and functions for each gene. Some are currently only annotated in either ontology domain and many synaptic proteins are known to play a role in multiple functions and/or be present in multiple subcellular compartments, demanding multiple biological process or cellular component annotations. In contrast to expanding gene coverage, this can be seen as a depth-first expansion of the knowledgebase.

Multifunctional proteins versus canonical function

An interesting dilemma that arose while annotating synaptic proteins that may be located in multiple subcellular compartments relates to what is the 'canonical' localization, or function, of a synaptic protein in case of multiple annotations? For example, the Gria1 protein (aka. GluA1) is currently annotated in SynGO as both on the pre- and post-synaptic membrane whereas it is commonly considered to be primarily a postsynaptic protein. One could speculate that the evidence provided for presynaptic localization is accurate but only reflects a small minority population of synapses, making the annotation no less factual but complicating the interpretation of 'what does this protein do?' nonetheless. For instance, running an over-representation analysis on a dataset with proteins that are predominantly expressed in the post-synapse, while also having presynaptic annotations, will question whether the experimental phenotype of said dataset is a pre- or post-synaptic phenomenon simply because ratio-metric distribution for annotations is lacking. This limitation holds for both SynGO and the GO database since both have the same ontology systematics.

Solving this dilemma does not only pertain to the annotation systematics but is mostly limited by a lack of data. If we would include an expression ratio or score to each annotation, this would still need to be based on experimental evidence, which is still lacking at this point. As an intermediate solution, the domain expert that creates or updates an annotation could classify one location and one function per synaptic protein as 'canonical'. This approach relies on the annotator's ability to make a compelling argument for some annotation to be considered canonical based on published literature. This approach is limited by ambiguity (e.g. What are the exact rules and can these apply to all annotations?) and human bias (e.g. selected literature and interpretation thereof by annotator), but until we find a large-scale approach to generating such data this is an appealing interim solution.

Sub-synaptic organelles

Mitochondria are localized and active inside the synapse, but as of now not yet part of the SynGO ontology because mitochondrial proteins are already well annotated (Calvo *et al.,* 2016; Smith & Robinson, 2019). However, as interaction between synaptic function and mitochondria is further unraveled, we may find synapse specific interacting proteins that warrant the introduction of specific ontology terms for mitochondria, while also having a mechanism that allows the integration of a generic 'mitochondrial protein resource' into SynGO. A similar systematic should be applied to ribosomes and proteasomes. Crucially, this requires a strict definition of 'synapse-specific <function>'.

Synapses are not isolated compartments

So far, all proteins annotated in SynGO are strictly required to have confirmed localization within synapses. However, synapses do not operate in isolation and some proteins that operate in pathways directly tied to synaptic processes cannot be annotated under these strict criteria. Although action potentials are the main input that presynaptic terminals (of chemical synapses) operate on to trigger neurotransmitter release, the synaptic system is dynamically modulated through post- to pre-synapse feedback loops (Regehr *et al.,* 2009) and interaction with adjacent glial cells (Eroglu & Barres, 2010; Neniskyte & Gross, 2017). Another topic of interest to SynGO that reaches beyond the synaptic compartment is the communication between synapse and soma, both the signaling pathways and the mechanisms of transport (Deisseroth *et al.,* 2003; Panayotis *et al.,* 2015; Yagensky *et al.,* 2016).

So future extensions to SynGO may consider relaxing the requirements for inclusion in SynGO to proteins with either confirmed synaptic localization or annotated to a synaptic biological process. In downstream application of SynGO data, users may always choose to filter by CC annotation if they are not interested in such proteins. In order to keep the scope of the SynGO project firmly focused on the synapse, and prevent 'scope creep', a key step will be the formalization of what encompasses and differentiates 'synaptic' and 'synapse related' processes.

SynGO version 2: Annotation of functional relationships between synaptic proteins

We ultimately aim for a quantitative description of synaptic molecular mechanisms with increasingly high resolution, which will be instrumental in bottom-up investigation of disorders that lead to disrupted brain function. To take this next step, we have to move from the current SynGO ontology systematics, where each ontology term is a 'geneset' (set of annotated genes), towards causal models that capture causal relationships between proteins and the molecular functions that these act upon. In spirit of this idea, examples of such models can often be found in informal notation in literature, illustrated as some cartoon figure that depicts the molecular mechanism of the respective subject of study or review. This knowledge of molecular mechanisms should be formalized in SynGO2 in defining a standardized approach, following the scientific rigor and engagement of domain experts as in the original SynGO community effort here described in chapter 6. Thomas *et al.* (2019) recently introduced GO-CAM, a structured framework ideally suited to implement the proposed SynGO2 in.

This will open up new avenues for the interpretation of genetics (e.g. GWAS) and 'omics (e.g. quantitative proteomics) data in the context of synapses. Where current analyses are limited to over-representation analyses of genesets, the proposed causal models enable network analyses that reason over the causal graph's structure. This enables, for example, the discovery of genes that are both enriched in some risk score, e.g., according to a GWAS, and causally related in the synapse by checking the local neighborhood of each synaptic protein in the causal models for neighbors that also have a high risk score (Reyna *et al.,* 2020). In contrast, we may not find any enrichment in this hypothetical example using the traditional geneset approach if these genes only share common annotations against an ontology term with many annotated genes, e.g., presynapse, a term that contains hundreds of genes. If we achieve a further increase in the level of detail of the SynGO knowledgebase by integrating functional relationships between individual proteins into maps of molecular mechanisms, we can make the next leap in resolving power.

The most prominent challenges to achieve this are the human resources required to create and curate the annotation models, and acquiring a comparable level of detailed data for all synaptic proteins. Regarding the former, domain expertise is a necessity for making detailed models that reflect the current scientific state of affairs throughout a vast catalogue of synaptic machineries and processes. For SynGO version 1, it took an international consortium of synapse experts over a year to create and curate all annotations currently in the knowledgebase. The more detailed we strive to make the annotation models, the more reliance on the respective domain expert to infer from a body of literature. After all, not every synaptic protein is equally well studied. Which gives rise to the second challenge, i.e., identifying currently available data and exploring what additional experimental data needs to be generated.

Using cell types for synapse types

Ultimately, SynGO should not only describe 'the canonical chemical synapse' but also appreciate the heterogeneity of synapses in the brain and describe how synapse types relate to signal integration properties. In this section I discuss a few challenges moving forward.

Many distinct neuronal cell types which encode different repertoires of genes have so far been identified by single cell transcriptomics (Tasic *et al.,* 2016; Poulin *et al.,* 2016). The repertoire of available proteins available for the construction of pre- and post-synaptic compartments connected to a neuron is determined by its cell type. But using single cell RNA sequencing (scRNA-seq) data on neurons alone may not be sufficient for the accurate identification of synapse types. For example, local translation near synapses may affect synaptic protein levels (Nakahata & Yasuda, 2018; Holt *et al.,* 2019) and such local translation may not be captured by RNA sequencing of the neuron.

Furthermore, not all protein identities available to a neuron (according to identified RNA sequences) have to be expressed in each of its synapses. Synapses of the same neuron are known to form a heterogeneous population of synapses with various efficacies (Grant & Fransén, 2020), so future ultra-sensitive proteomics assays might be used to investigate the diversity in molecular composition of their synaptomes. Perhaps some subpopulations of synapses from the same neuron encode for protein X and others for protein Y. Or, conversely, perhaps synaptic proteomes of the same neuron prove to be quite homogenous in the future (and heterogeneous phenotypes are just a function of activity/interaction applied to the same protein partslist), in which case scRNA-seq would be a strong predictor of synaptic proteomes.

To address this, additional large-scale experimental data is required to unravel protein composition of within populations of synapses. Whether one uses biochemical purification combined with proteomics or imaging to observe protein identities in some synapse population, preserving association with the respective neuron's cell type is pivotal if the data is to be used for investigating synapse types and ultimately modeling cell circuits. One idea is to use genetic approaches to express a unique synaptic molecule (e.g. minor modification of a synaptic protein) in a neuronal cell type of interest, or add a unique tag to each (known) neuronal cell type, then isolate synapses and apply mass-spec. Because of the genetic tag included inside each synapse, its associated neuron (type) could be inferred in the proposed experiment. This allows for comparison between neuronal cell types, but not among individual synapses of the same neuron, for which imaging would be required.

A first step that can be taken today is the interpretation of single-cell RNA sequencing data together with SynGO to identify combinations of pre- or post-synaptic proteins uniquely associated with some neuronal (sub)class. This would predict which synaptic genes are available for building synapses at each neuron type/subclass and thereby predict 'synapse types'. Potential challenges are local translation and synapse heterogeneity within the same cell type as discussed above. Once SynGO2 is under construction, the same principle could be applied. If we intersect the detailed causal models with scRNA-seq datasets we may be able to identify variations of the same pathway as expressed in the brain, presumably with optimized properties for the respective neuronal circuit, thereby moving beyond the current understanding in our field.

Outlook

Mass spectrometry and bioinformatics have proven to be a powerful combination in synaptic proteome studies. I have shown application to inter-species comparisons and discovery of novel synaptic proteins in this thesis, in concert with incremental improvements to methodological approaches, but we must continue to push the envelope and characterize the synaptome in greater detail to empower further studies on synapse function.

Future ambitions for synapse proteomics include the routine characterization of (known synaptic) proteins that are now rarely or never observed in mass spectrometry studies ('dark proteome'). This might be achieved by advancements in sensitivity, a more general objective of mass spectrometry development, but alternative protocols or acquisition strategies like middle-down or top-down should not be overlooked. Additional dimensions of proteome data that are not included in this thesis but are exciting avenues for future synapse research include; mapping post-translational modifications (PTM), protein-protein interactions (PPI) and increasing sensitivity such that we can study smaller populations of cells and synapses.

The SynGO manuscript characterized unique properties of synaptic genes, such as their increased complexity compared to other (brain expressed) genes, strong evolutionary conservation, intolerance to mutation and association to brain disorders. The systematic analysis of synaptic genes at this scale was for the first time made possible by an expert-curated knowledgebase of the synapse. As a next step beyond expansion of the current knowledgebase, we should aim for a systematic description of molecular interactions in increasingly high resolution. Adhering to the evidence driven and expert-curated approach, this will be a monumental task that is only possible through community-wide efforts.

Moving towards causal annotation models and a complete coverage of the synapse that also includes cross-compartment pathways will demand the generation of new experimental data on fundamental molecular mechanisms of the synapse. Taken together, this roadmap will lead to a resource instrumental in bottom-up investigation of disorders that lead to disrupted brain function and efforts to unravel the architecture of brain circuitry.

References

1. Abbas, Y. M. *et al.* Structure of V-ATPase from the mammalian brain. *Science* **367,** 1240–1246 (2020).
2. Amodei, D. *et al.* Improving Precursor Selectivity in Data-Independent Acquisition Using Overlapping Windows. eng. *Journal of the American Society for Mass Spectrometry* **30,** 669–684. ISSN: 1879-1123. PMID: 30671891 (Apr. 2019).
3. Bayés, *et al.* Comparative Study of Human and Mouse Postsynaptic Proteomes Finds High Compositional Conservation and Abundance Differences for Key

Synaptic Proteins. *PLoS ONE* **7** (ed Dunaevsky, A.) e46683. https://doi.org/10.1371/journal.pone.0046683 (Oct. 2012).

4. Bekker-Jensen, D. B. *et al.* A compact quadrupole-orbitrap mass spectrometer with FAIMS interface improves proteome coverage in short LC gradients. *Molecular & Cellular Proteomics* **19,** 716–729 (2020).
5. Benito, I., Casañas, J. J. & Montesinos, M. L. Proteomic Analysis of Synaptoneurosomes Highlights the Relevant Role of Local Translation in the Hippocampus. *Proteomics* **18,** 1800005 (2018).
6. Biesemann, C. *et al.* Proteomic screening of glutamatergic mouse brain synaptosomes isolated by fluorescence activated sorting. *The EMBO journal* **33,** 157–170 (2014).
7. Boyken, J. *et al.* Molecular Profiling of Synaptic Vesicle Docking Sites Reveals Novel Proteins but Few Differences between Glutamatergic and GABAergic Synapses. *Neuron* **78,** 285–97. ISSN: 1097-4199 (Electronic) 0896-6273 (Linking) (2013).
8. Calvo, S. E., Clauser, K. R. & Mootha, V. K. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. eng. *Nucleic acids research* **44,** D1251–7. ISSN: 1362-4962. PMID: 26450961 (Jan. 2016).
9. Deisseroth, K. *et al.* Signaling from synapse to nucleus: the logic behind the mechanisms. *Current opinion in neurobiology* **13,** 354–365 (2003).
10. Dosemeci, A. *et al.* Composition of the Synaptic PSD-95 Complex. *Molecular & Cellular Proteomics* **6,** 1749–1760. https://doi.org/10.1074/mcp.m700040-mcp200 (July 2007).
11. Eroglu, C. & Barres, B. A. Regulation of synaptic connectivity by glia. *Nature* **468,** 223–231 (2010).
12. Galli, T., McPherson, P. S. & De Camilli, P. The Vo sector of the V-ATPase, synaptobrevin, and synaptophysin are associated on synaptic vesicles in a Triton X-100-resistant, freeze-thawing sensitive, complex. *Journal of Biological Chemistry* **271,** 2193–2198 (1996).
13. Grant, S. G. & Fransén, E. The Synapse Diversity Dilemma: Molecular Heterogeneity Confounds Studies of Synapse Function. *Frontiers in Synaptic Neuroscience* **12,** 45 (2020).
14. Gu, M. *et al.* AP2 hemicomplexes contribute independently to synaptic vesicle endocytosis. *Elife* **2,** e00190 (2013).
15. Haucke, V., Neher, E. & Sigrist, S. J. Protein scaffolds in the coupling of synaptic exocytosis and endocytosis. *Nature Reviews Neuroscience* **12,** 127–138 (2011).
16. Holt, C. E., Martin, K. C. & Schuman, E. M. Local translation in neurons: visualization and function. *Nature structural & molecular biology* **26,** 557–566 (2019).

17. Huang, T. *et al.* Combining Precursor and Fragment Information for Improved Detection of Differential Abundance in Data Independent Acquisition. *Molecular & Cellular Proteomics* **19,** 421–430 (2020).

18. Jahn, R. & Sudhof, T. C. Synaptic vesicles and exocytosis. *Annual review of neuroscience* **17,** 219–246 (1994).

19. Käll, L. *et al.* Integrated identification and quantification error probabilities for shotgun proteomics. *Molecular & Cellular Proteomics* **18,** 561–570 (2019).

20. Karpievitch, Y. *et al.* A Statistical Framework for Protein Quantitation in Bottom-up MS-Based Proteomics. *Bioinformatics* **25,** 2028–2034 (Aug. 2009).

21. Kirchhausen, T., Owen, D. & Harrison, S. C. Molecular structure, function, and dynamics of clathrin-mediated membrane traffic. *Cold Spring Harbor perspectives in biology* **6,** a016725 (2014).

22. Kovtun, O. *et al.* Architecture of the AP2/clathrin coat on the membranes of clathrin-coated vesicles. *Science advances* **6,** eaba8381 (2020).

23. Li, K. W. *et al.* Proteomics Analysis of Rat Brain Postsynaptic Density. *Journal of Biological Chemistry* **279,** 987–1002. https://doi.org/10.1074/jbc.m303116200 (Oct. 2003).

24. Meier, F., Brunner, A.-D., *et al.* diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nature Methods* **17,** 1229–1236 (2020).

25. Meier, F., Geyer, P. E., *et al.* BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. eng. *Nature methods* **15,** 440–448. ISSN: 1548-7105. PMID: 29735998 (June 2018).

26. Michalski, A., Cox, J. & Mann, M. More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority Is Inaccessible to Data-Dependent LC-MS/MS. *J. Proteome Res.* **10,** 1785–1793 (Apr. 2011).

27. Morgan, J. R. *et al.* A conserved clathrin assembly motif essential for synaptic vesicle endocytosis. *Journal of Neuroscience* **20,** 8667–8676 (2000).

28. Nakahata, Y. & Yasuda, R. Plasticity of spine structure: local signaling, translation and cytoskeletal reorganization. *Frontiers in synaptic neuroscience* **10,** 29 (2018).

29. Neniskyte, U. & Gross, C. T. Errant gardeners: glial-cell-dependent synaptic pruning and neurodevelopmental disorders. *Nature Reviews Neuroscience* **18,** 658 (2017).

30. Panayotis, N. *et al.* Macromolecular transport in synapse to nucleus communication. *Trends in neurosciences* **38,** 108–116 (2015).

31. Poulin, J.-F. *et al.* Disentangling neural cell diversity using single-cell transcriptomics. *Nature neuroscience* **19,** 1131–1141 (2016).

32. Rao-Ruiz, P. *et al.* Time-dependent changes in the mouse hippocampal synaptic membrane proteome after contextual fear conditioning. eng. *Hippocampus* **25,** 1250–61. ISSN: 1098-1063. PMID: 25708624 (Nov. 2015).

33. Regehr, W. G., Carey, M. R. & Best, A. R. Activity-dependent regulation of synapses by retrograde messengers. *Neuron* **63,** 154–170 (2009).

34. Reyna, M. A. *et al.* Pathway and network analysis of more than 2500 whole cancer genomes. *Nature communications* **11,** 1–17 (2020).

35. Schilling, B., Gibson, B. W. & Hunter, C. L. Generation of High-Quality SWATH® Acquisition Data for Label-free Quantitative Proteomics Studies Using TripleTOF® Mass Spectrometers. eng. *Methods in molecular biology (Clifton, N.J.)* **1550,** 223–233. ISSN: 1940-6029. PMID: 28188533 (Feb. 2017).

36. Schrimpf, S. P. *et al.* Proteomic analysis of synaptosomes using isotope-coded affinity tags and mass spectrometry. *Proteomics* **5,** 2531–2541 (2005).

37. Schweppe, D. K. *et al.* Full-featured, real-time database searching platform enables fast and accurate multiplexed quantitative proteomics. eng. *Journal of proteome research.* ISSN: 1535-3907. PMID: 32126768 (Mar. 2020).

38. Shen, X. *et al.* IonStar enables high-precision, low-missing-data proteomics quantification in large biological cohorts. eng. *Proceedings of the National Academy of Sciences of the United States of America* **115,** E4767–E4776. ISSN: 1091-6490. PMID: 29743190 (May 2018).

39. Sialana, F. J. *et al.* Mass spectrometric analysis of synaptosomal membrane preparations for the determination of brain receptors, transporters and channels. *PROTEOMICS* **16,** 2911–2920. https://doi.org/10.1002/pmic.201600234 (Nov. 2016).

40. Smith, A. C. & Robinson, A. J. MitoMiner v4.0: an updated database of mitochondrial localization evidence, phenotypes and diseases. eng. *Nucleic acids research* **47,** D1225–D1228. ISSN: 1362-4962. PMID: 30398659 (Jan. 2019).

41. Takamori, S. *et al.* Molecular Anatomy of a Trafficking Organelle. *Cell* **127,** 831–846. https://doi.org/10.1016/j.cell.2006.10.030 (Nov. 2006).

42. Tang, B. *et al.* Fmr1 deficiency promotes age-dependent alterations in the cortical synaptic proteome. *Proceedings of the National Academy of Sciences* **112,** E4697–E4706 (2015).

43. Tasic, B. *et al.* Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nature neuroscience* **19,** 335–346 (2016).

44. Thomas, P. D. *et al.* Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. *Nature genetics* **51,** 1429–1433 (2019).

45. Volknandt, W. & Karas, M. Proteomic analysis of the presynaptic active zone. *Experimental Brain Research* **217,** 449–461. https://doi.org/10.1007/s00221-012-3031-x (Feb. 2012).

46. Wilhelm, B. G. *et al.* Composition of Isolated Synaptic Boutons Reveals the Amounts of Vesicle Trafficking Proteins. *Science* **344,** 1023–8. ISSN: 1095-9203 (Electronic) 0036-8075 (Linking) (2014).

47. Yagensky, O., Kalantary Dehaghi, T. & Chua, J. J. E. The roles of microtubule-based transport at presynaptic nerve terminals. *Frontiers in synaptic neuroscience* **8,** 3 (2016).

Acknowledgements

Jumping into neuroscience has been an amazing journey into the unknown. The opportunity to work, learn and grow as both a scientist and human being, while hopefully contributing to the greater good in the process, has been a gift.

This long overdue thesis was made possible by the support of many people. First of all, many thanks to Niels for the leap of faith in hiring me and setting off this journey. In the first two years I enjoyed a part-time stay at Tom's department in Nijmegen, which laid the foundation to the data science presented in this thesis. Tom, many thanks for your guidance and wisdom, I wish there was time to learn more. Tjeerd, it's been a blast learning all about bioinformatics, you have been a great mentor.

The CNCR in Amsterdam has been home away from home ever since. Matthijs, thanks for fostering an environment of scientific excellence in which we can all learn and thrive at FGA. The lab meetings and sharp discussions helped open the door to synapse biology for me (from the most important side of the synapse, of course). I'm happy that, after some years, we found a fitting project to collaborate on that ultimately led to a successful paper. Guus, your incredible dedication and scientific ability continue to inspire. I've enjoyed our discussions in years past about virtually everything related to our research and beyond. Somehow we always find solutions to tough problems and have a great time doing it. To Ka Wan, proteomics star and fearless innovator who is always exploring new things (easy!) and looking past minor hurdles on our path (doesn't matter!). Thanks for your teachings and open-mindedness, we make a great team and together we keep pushing the envelope.

To the many collaborators and colleagues, over the many years; it's been an absolute pleasure and I am certain your future is bright. Roel and Iryna, thanks for enabling all those proteomics projects. Remco, thanks for sharing your mastery in proteomics, friendship and many Balistos. David, thanks for the joyful collaboration and neuropathology adventures. Ning, the footsteps of your tremendous work have been followed by many. Joerg, a kindred spirit I found along the way. Pim, thanks for sharing beers and climbing mount SynGO with me. Nikhil, the optimism and energy as we spent crazy hours and weekends to push each other, and science, forward has made a lasting impression. Sophie and Miguel, thanks for being great colleagues and friends; partners in crime in everything from scientific endeavors to sky-diving and traveling to the US and China.

I am not one for the celebration of 'milestones', or personal goals for that matter, but rather cherish everyday life, the road travelled upon and the company I keep. I am grateful for the patience, support and many moments of joy shared with those around me. A shout-out to my homies; Sharon, Fleur and Judith, you made phs4 feel like home and I miss having y'all around. Cheers to living the good life in Nimma and beyond with my mates Willem, Roy, Roberto, Sebas and the entire Barry-bar clan. And finally a heartfelt thank you to my entire family, the foundation of my life.

Curriculum Vitæ

Frank Koopmans

13-06-1982	Born in Beuningen, The Netherlands

Education

2000–2004 Bachelor’s degree in Information & Communication Technology
Avans Hogeschool, Den Bosch, The Netherlands

2004–2009 Master’s degree in Computer Science
Radboud University, Nijmegen, The Netherlands

Thesis Title: ‘Tracking Local Community Evolution’
Analyze cluster evolution in complex graphs, such as online social networks, to model external influence on local communities.
Thesis supervisor: prof. Theo van der Weide

2009–2010 Master’s degree in Information Sciences
Radboud University, Nijmegen, The Netherlands

Thesis Title: ‘Broader Perception For Local Community Identification’
Introduce and validate improvements to local network community identification algorithms, presented at the KDIR2010 conference.
Thesis supervisor: prof. Theo van der Weide

2011–2021 PhD research presented in this thesis
Thesis Title: ‘Capturing the synaptic proteome’

Interdisciplinary research that combines synapse research, proteomics and bioinformatics.

Departments of Functional Genomics & Molecular and Cellular Neurobiology, VU University, Amsterdam, The Netherlands
Supervisors: Dr. Niels Cornelisse, Dr. Ka Wan Li, prof. Matthijs Verhage, prof. Guus Smit

Machine Learning Group, Radboud University, Nijmegen, The Netherlands
Supervisors: Dr. Tjeerd Dijkstra, prof. Tom Heskes

List of Publications

First author works

Koopmans, F.; Cornelisse, L. N.; Heskes, T.; Dijkstra, T. M., Empirical bayesian random censoring threshold model improves detection of differentially abundant proteins. *Journal of proteome research* 2014, *13* (9), 3871-3880.

Pandya, N. J.*; **Koopmans, F***.; Slotman, J. A.; Paliukhovich, I.; Houtsmuller, A. B.; Smit, A. B.; Li, K. W., Correlation profiling of brain sub-cellular proteomes reveals co-assembly of synaptic proteins and subcellular distribution. *Scientific reports* 2017, *7* (1), 1-11.

Koopmans, F.; Ho, J. T.; Smit, A. B.; Li, K. W., Comparative analyses of data independent acquisition mass spectrometric approaches: DIA, WiSIM-DIA, and untargeted DIA. *Proteomics* 2018, *18* (1), 1700304.

Koopmans, F*.; Pandya, N. J.*; Franke, S. K.; Phillippens, I. H.; Paliukhovich, I.; Li, K. W.; Smit, A. B., Comparative hippocampal synaptic proteomes of rodents and primates: differences in neuroplasticity-related proteins. *Frontiers in molecular neuroscience* 2018, *11*, 364.

Koopmans, F.; van Nierop, P.; SynGO consortium; Cornelisse, L. N.; Smit, A. B.; Verhage, M., SynGO: an evidence-based, expert-curated knowledge base for the synapse. *Neuron* 2019, *103* (2), 217-234. e4.

Hondius, D. C.*; **Koopmans, F***.; Leistner, C.; Pita-Illobre, D.; Peferoen-Baert, R. M.; Marbus, F.; Paliukhovich, I.; Li, K.K.; Rozemuller, A. J. M.; Hoozemans, J. J. M.; Smit, A. B., The proteome of granulovacuolar degeneration and neurofibrillary tangles in Alzheimer's disease. *Acta Neuropathologica* 2021, *141* (3).

Co-authored works

Li, K. W.; Chen, N.; Klemmer, P.; **Koopmans, F**.; Karupothula, R.; Smit, A. B., Identifying true protein complex constituents in interaction proteomics: the example of the DMXL2 protein complex. *Proteomics* 2012, *12* (15-16), 2428-2432.

Chen, N.; Pandya, N. J.; **Koopmans, F**.; Castelo-Székelv, V.; van der Schors, R. C.; Smit, A. B.; Li, K. W., Interaction proteomics reveals brain region-specific AMPA receptor complexes. *Journal of proteome research* 2014, *13* (12), 5695-5706.

Chen, N.; **Koopmans, F**.; Gordon, A.; Paliukhovich, I.; Klaassen, R. V.; van der Schors, R. C.; Peles, E.; Verhage, M.; Smit, A. B.; Li, K. W., Interaction proteomics of canonical Caspr2 (CNTNAP2) reveals the presence of two Caspr2 isoforms with overlapping interactomes. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 2015, *1854* (7), 827-833.

Geerts, C. J.; Mancini, R.; Chen, N.; **Koopmans, F**.; Li, K. W.; Smit, A. B.; Van Weering, J. R.; Verhage, M.; Groffen, A. J., Tomosyn associates with secretory vesicles in neurons through its N-and C-terminal domains. *PloS one* 2017, *12* (7).

He, E.; Lozano, M. A. G.; Stringer, S.; Watanabe, K.; Sakamoto, K.; den Oudsten, F.; **Koopmans, F**.; Giamberardino, S. N.; Hammerschlag, A.; Cornelisse, L. N., MIR137 schizophrenia-associated locus controls synaptic function by regulating synaptogenesis, synapse maturation and synaptic transmission. *Human molecular genetics* 2018, *27* (11), 1879-1891.

Gupta, I.; Collier, P. G.; Haase, B.; Mahfouz, A.; Joglekar, A.; Floyd, T.; **Koopmans, F**.; Barres, B.; Smit, A. B.; Sloan, S. A., Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nature biotechnology* 2018, *36* (12), 1197-1202.

Brouwer, M.; Farzana, F.; **Koopmans, F**.; Chen, N.; Brunner, J. W.; Oldani, S.; Li, K. W.; van Weering, J. R.; Smit, A. B.; Toonen, R. F., SALM1 controls synapse development by promoting F-actin/PIP2-dependent Neurexin clustering. *The EMBO journal* 2019, *38* (17).

Gonzalez-Lozano, M. A.; **Koopmans, F**., Data-Independent acquisition (SWATH) mass spectrometry analysis of protein content in primary neuronal cultures. In *Springer: Neuroproteomics* 2019, pp 119-127.

Gonzalez-Lozano, M. A.; **Koopmans, F**.; Paliukhovich, I.; Smit, A. B.; Li, K. W., A fast and economical sample preparation protocol for interaction proteomics analysis. *Proteomics* 2019, *19* (9), 1900027.

van der Spek, S. J.; **Koopmans, F**.; Paliukhovich, I.; Ramsden, S. L.; Harvey, K.; Harvey, R. J.; Smit, A. B.; Li, K. W., Glycine Receptor Complex Analysis Using Immunoprecipitation-Blue Native Gel Electrophoresis-Mass Spectrometry. *Proteomics* 2020, p.1900403.

Gonzalez-Lozano, M. A.; **Koopmans, F**.; Sullivan, P. F.; Protze, J.; Krause, G.; Verhage, M.; Li, K. W.; Liu, F.; Smit, A. B., Stitching the synapse: Cross-linking mass spectrometry into resolving synaptic protein interactions. *Science Advances* 2020, *6* (8).

Li, K. W.; Gonzalez-Lozano, M. A.; **Koopmans, F**.; Smit, A. B., Recent Developments in Data Independent Acquisition (DIA) Mass Spectrometry: Application of Quantitative Analysis of the Brain Proteome. *Front Mol Neurosci* 2020, *13* (564446).