Decoding neuroproteomics: integrating the proteome with the transcriptome and genome

Decoding neuroproteomics: integrating the translatome with the connectome

Rob Kitchen 4/23/14 9:04 PM
Comment [1]: we might want to keep options for titles – could we consider adding something about "cell type specificity" or "connectome" – might be too much in the title

Robert R. Kitchen[1,2], Joel S. Rozowsky[1], Mark B. Gerstein[1,3], Angus C. Nairn[2]

[1] Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA.

[2] Division of Molecular Psychiatry, Abraham Ribicoff Research Facilities, Connecticut Mental Health Center, Yale University School of Medicine, New Haven, CT 06508, USA.

[3] Department of Computer Science, Yale University, New Haven, CT 06520, USA.

Abstract

Whole-proteome analyses have the potential to be a powerful complement to existing or proposed genomic, epi-genomic, and transcriptomic studies. Technological developments in tandem mass-spectrometry now have the potential to allow investigators to profile peptides and proteins at a sufficiently high resolution and coverage to meaningfully complement, even in complex mammalian systems, results obtained from high-throughput transcriptomic studies. In this review we discuss this state of the art in mass-spectrometry proteomics, highlight recent large scale efforts to quantify the proteome of mammalian systems, and cover attempts to integrate proteomic data with functional genomic data for a more holistic approach to measuring gene expression. We discuss ways in which proteomic data and analysis can be made more compatible with the other high-throughput *omics data, both from more meaningful processing of the peptide data themselves to improved downstream integration in gene-expression and variation network analyses. Finally, we address issues regarding profiling the proteome of the central nervous system, paying specific attention to the immense inter- and intra-cellular

heterogeneity in the mammalian brain. This heterogeneity drives the requirement for better integration of protein measurement with functional genomics and imaging and we discuss methodologies being employed to achieve finer resolution in the high-throughput *omics data obtained from neural tissues.

Introduction

Since the completion of the human genome sequencing project there has been a huge amount of community effort devoted to the functional characterisation of the genome, from its structure to its molecular products. Thanks to the astounding pace of technological and methodological innovation there are a wealth of assays available for querying the full gamut of processes accessible via the measurement of nucleic acids including 3-dimensional DNA conformation and interactions [ref, ref] and structural variation [ref], DNA-protein interactions and modifications [ref], RNA transcription [ref], post-transcriptional modifications [ref], and post-transcriptional and -translational regulation [ref, ref, ref]. High-profile, multi-investigator efforts have recently produced much of these genomic data either, in the case of ENCODE, epigenome roadmap, and BrainSpan, to characterise the multi-omic landscape of specific cell-types, tissues, and species or, in the case of 1000 Genomes, gEUVADIS, TCGA, and GTEx to characterise genome and transcriptome across a large set of individuals to better understand their variation in the human population and disease. There have been a number of ambitious initiatives to characterize RNA expression [ref, ref nenad, ref, ref] and localisation [ref allen, gensat] in the central nervous system and across the regions of the brain; each reinforcing observations of significant differences in gene expression between neuronal cell-types, brain regions, developmental stages, and species [ref].

Rob Kitchen 4/30/14 8:12 AM
Comment [2]: BrainSpan, xSpecies

While these research efforts have lent significant insight into the incredible complexity of cellular regulatory processes and their dynamics under perturbation or disease, a notable exception has been tandem mass-spectrometry-based whole-proteome analyses (MS/MS); even very recently awarded projects such as psychENCODE do not include whole-proteome profiling in their core experimental design. There are several reasons why advancements in proteome analyses have lagged compared to the other *omics. Notably it is our inability to in-vitro amplify amino acids leads to more demanding requirements on the technology used for detection; although innovation in amino-acid 'sequencing' appear promising [ref]. Despite this limitation, with recent technological advances it is now possible to reliably obtain quantitative observations of tens of thousands of peptides derived from between 1,000-12,000 proteins [ref].

Given the wealth of additional insight into the biosynthetic state of the cell offered by MS/MS, whole-proteome analysis is an increasingly attractive option for investigators, even those studying complex organisms.

Work in several fields has begun to deliver some results in disentangling the immense complexity of the neuronal circuitry of the mammalian brain [ref]. Mapping the complete set of neuronal connections, dubbed the 'connectome', is an area of extremely active research, in which imaging tools such as fMRI, dMRI, and PET are being employed to non-invasively produce huge amounts of data [ref] on the wiring of the brain at different resolutions and under different conditions [ref]. Thanks to advancements in automation, electron microscopy is an increasingly attractive method for tracing neuronal processes through thousands of perfectly stacked images to trace the multitude of connections within the brain [ref, ref]. Further, in-situ hybridisation (ISH) and immunohistochemistry (IHC) has been used to create detailed maps of the spatial expression profiles of individual genes across the human brain [ref allenAtlas, ref]. MALDI technology is capable of allowing peptides and proteins from the same brain tissue slices stained by IHC to MS/MS to more deeply profile the protein contents of precise brain regions and tissues [ref].

However, despite advances in these imaging- and antibody-based methods, MS/MS remains the only method of profiling the protein contents of the tissues of the brain with a throughput and resolving power that comparable to other functional genomic methods. This is important to further unravel the normal and abnormal system-level function of CNS cells, by better understanding the relationship between RNA and protein expression, the roles of post-translational modifications, and the localisation of proteins, especially in the likely 100's of distinct neuronal cell-types each with specific transcriptome/proteome profiles. The purpose of this review is to summarise the potential benefits of wide adoption of MS/MS proteomics, outline several methodological improvements that might facilitate the integration of proteomic data with other functional genomic analyses, and discuss specific considerations with respect to proteome profiling of the CNS.

Biological insights from mass-spectrometry based proteomics

The standard insight offered by MS/MS is one of assaying peptide- or protein-level abundances, an overview of which is available in **Box 1** and further illustrated in **Figure 1**. These methods of spectra acquisition and quantification are complemented by the various options available for the purification of protein from distinct **sub-cellular compartments**, which allows investigators to

specifically and separately assay proteins located in the nucleus, cytosol, cytoskeleton, endoplasmic reticulum, and plasma membrane [ref]. This is similar to the nucleus- and cytoplasm-specific preps available for purifying RNA, but the increased resolution available with MS/MS can improve sensitivity to low abundance or membrane-specific proteins that may otherwise be occluded by highly abundant nuclear and cytosolic proteins. For example, the protein composition of the nucleolus, nucleus, and cytoplasm has been found to be quite exclusive, with only a small number of proteins equally abundant between each compartment [ref].

Extending beyond comparison of steady state protein expression, metabolic labelling of amenable cells (for example cell cultures or yeast) is capable of yielding valuable insight to the **rates of protein turnover** and has revealed that proteins are both more abundant and have a longer half life than do RNAs [ref]. Similar studies of dynamic cellular processes involving protein kinases and phosphatases, regulatory enzymes responsible for signal transduction, and their sites of **phosphorylation** are exclusively accessible by MS/MS proteomics [ref] and provide valuable insight into sometimes subtle regulation of protein function in health and disease [ref].

Besides being the ultimate 'read-out' of the abundance of the molecular products of a genome, the proteome contains a wealth of information about the landscape of well over 90,000 sites [ref] of **post-translational modifications** (PTMs) that are simply inaccessible through the analysis of nucleic acids [ref]. In addition to phosphorylation, PTMs such as acetylation, acylation, deamidation, glycosylation, methylation, and ubiquitination play pivotal roles in regulating almost all cellular processes including energy production and transport, DNA modification, transcription and translation of RNA, and RNA or protein stability [ref]. This information is obtained directly using MS/MS due to the characteristic mass-shift that they cause in the peptide spectra [ref] and, as such, provides not only the exact locations of these modifications but also quantitative measurements of their relative abundance. These peptide sequence and PTM data can be obtained using the sub-cellular fractionation methods mentioned earlier; for example, this approach has been exploited to reveal the fascinating landscape of **histone modifications**, including the simultaneous analysis of the abundance and co-occurrence of various combinations of marks, in the nucleus and their changes during ES cell differentiation [ref]. Indeed, proteomic profiling of the nucleus enables simultaneous quantification of the complete set of expressed **transcription factors**, the relative abundance of which has been shown to provide valuable transcriptional insights [ref] and, despite the obvious lack of information regarding the genomic binding locations of these nuclear proteins, proteomic

profiling has been vital in identifying regulatory proteins [ref].

Integration of proteomic and transcriptomic data

Integration of expression measurements

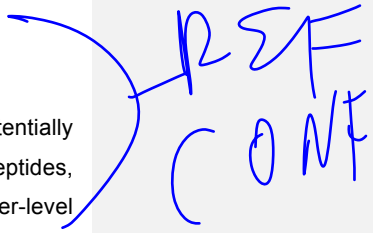
Several high-quality studies have employed MS/MS to profile the protein output of tissues and organisms under a variety of conditions and have proven useful as standalone resources in their own right [ref, ref, ref]. However a large number of MS/MS experiments are performed as validations for findings obtained by chromatin or RNA profiling [ref, ref], similarly RNA-level data has also been used as a 'reverse-validation' of the results obtained by MS/MS [ref]. However such attempts at presenting a so-called 'integrated' analysis often appear haphazard and provide limited utility in terms of the information the validation data add to the original. The situation is further complicated due to the often limited correlation of observed mRNA and protein abundances, even in well matched experiments [refs].

Several studies over recent years have attempted to elucidate this **limited correlation in molecular abundances** of the transcriptome with those of the proteome [ref mark's stuff, ref], but almost all have reported values between 50-70%; certainly not strong enough for measurements of mRNA abundance alone to be considered predictive of protein abundance [ref biton, ref, ref ingolia ref etc]. This is true regardless of the technology or experimental method used to profile the RNA (including microarray [ref], RNA-seq [ref], and ribosome profiling [ref]) or the protein [ref SILAC, ref label-free]. The cause for this poor correlation is more than likely a combination of biological and technical factors. General biological variables including cellular heterogeneity, alternative splicing, differential RNA stability, micro-RNA induced repression, post-translational modifications, protein-turnover, and protein localisation. The ability to resolve such biological variables is confounded by the technical noise introduced due to differences in sample preparation, measurement technology, and data handling used for the transcriptomic and proteomic analyses [ref, ref].

A recent development in transcriptomics has enabled investigators to directly assess translational control, known to be a significant regulatory process that determines the protein output of a transcript [ref], by sequencing the very short fragments of RNA contained within the mono-ribosome complex itself [ref]. This so-called ribosome profiling allows, for the first time, a transcriptome-wide survey of the positions of ribosomes on each transcript and, when compared to the relative abundance of those same transcripts, has introduced the concept of translational

Rob Kitchen 4/23/14 9:23 PM
Comment [3]: could mention Ribo-tag methods, TRAP etc in general here to then go onto profiling

efficiency to mainstream gene expression profiling. This has a large impact for the integration of proteome and transcriptome analyses as the translational efficiency may, once the methodology matures, prove a more reliable indicator of protein abundance than simple RNA expression [ref]. For example, an early study of translational efficiency in yeast revealed that cells can modify their protein output whilst maintaining stable RNA abundances during different stages of meiosis, simply by increasing the density of ribosomes on selected transcripts [ref]. Ribosome profiling has also shown that non-coding RNAs in the cytosol, which are known to be spliced, capped, and polyadenylated in a similar manner to mRNAs [ref], are engaged by the polyribosome [ref], but do not code for protein [ref, ref]. Finally, ribosome profiling has been used to identify sites in the 5' UTR of known transcripts that contain short open reading frames (ORFs). These upstream ORFs (uORFs) can have a variety of regulatory influences on their host transcript and can themselves produce short peptides, either of these scenarios are ripe for further exploration in searching for the relevant peptide sequences by MS/MS.



Improving the compatibility of proteomics with functional genomics

Databases of **peptide identifications** such as PRIDE [ref] and Peptide Atlas are a potentially extremely valuable resource for mapping spectra based on previously identified peptides, however such resources are of limited use to non-experts who may simply desire higher-level information on sites of post-translational modifications and the complement of proteins that are observed in a given disease state, tissue, or cellular compartment [ref]. In the same vein, a very basic, but significant limitation of most studies that attempt to combine RNA and protein level results is that different gene **annotations** are used in the analysis of the RNA compared to the protein, which results in mundane but immediate difficulties in integrating the output of these assays. More significantly, however, is the extent to which the use of different reference annotations limit one's ability to relate observed peptides to potential transcripts, which is necessary for integrative analyses of molecular networks [ref]. Adoption of a common reference would benefit not only improve RNA/protein abundance comparisons, but may also facilitate the integration of peptide/protein abundance information to genome browsers such as those provided by ENSEMBL and UCSC.

The community would also benefit from a resource providing **quantitative data** on peptide or protein abundance, such as the Plant Proteome Database [ref] or the Encyclopedia of Proteome Dynamics [ref]. Human gene expression atlases are well populated with in-situ and whole-transcriptome RNA abundance data [ref, ref], and although there are efforts to map

protein expression in human tissues by immunohistochemistry [ref] the real power of such resources lies in the combination of this expression localisation data and concurrent relative abundance measurements of thousands of genes. Unfortunately public quantitative proteomic datasets are currently lacking both in volume and in standardisation, particularly in terms of the processing of the raw spectra and the methods of obtaining and normalising the peptide- or protein-level abundances. A resource that combined peptide identification/quantification data in terms of a genome annotation that is consistent with genomics and transcriptomics would be of high value to the community, not least in providing a framework with which to tackle open questions in proteomics such as the non-uniform coverage of peptides belonging to the same protein.

Although it may not have a profound effect on the results of a single, internally consistent, analysis, the lack of data/analysis standardisation further increases the difficulty of comparing proteomic quantifications across studies, which is a key determinant of the utility of the kinds of resources described above. Recommendations from the Human Proteome Organisation (HUPO) Brain Proteome Project (BPP) for dissemination of MS/MS data include storing the list of identified spectral peaks along with the corresponding peptide sequence and modifications as the most 'sensible' unit of measurement [ref]. However this recommendation is particularly vulnerable to issues regarding the **identification and selection of the peaks** from the raw spectra. The choice of software for peak-picking, peptide identification, and quantification as well as the selection of related parameters, such as the potential PTMs to be used in the peptide identification, remains a significant source of data-loss and -variability in MS/MS [ref]. Until the situation improves regarding the processing of spectral peaks, it is necessary to retain the raw MS/MS output in public resources to facilitate development of open-source analysis software and re-analysis of collections of published datasets.

We have just recently started to see the application of **experimental and computational analysis methods** that aim to more tightly integrate analyses of the proteome with the transcriptome. Experimental methodologies such as ribosome profiling, single molecule RNA-sequencing, and top-down MS/MS profiling of intact proteins enable greater selectivity and sensitivity to molecules that are actively involved in the process of protein production and greater specificity to the exact structure of these transcripts and isoforms. Similarly, a small number of recent efforts to leverage the transcript sequence information obtained from RNA-seq to improve peptide identification MS/MS analyses [ref, ref] have resulted in the production of somewhat basic software tools for the direct integration of such datasets [ref]. Utilising ribosome footprinting to identify coding sequences and translation

initiation sites has produced moderate increases in the yield of peptide spectra in the same samples [ref, ref].

Benefits of integrating functional genomic and proteomic profiling of the CNS

Systematic integration of information obtained from various functional genomics assays has been crucial for deeper characterisation and understanding of the complex cellular machinery. Integration, for example of DNA variants and chromatin signals with transcriptomics has been extremely valuable for gaining a deeper intuition for the dynamics cellular processes through genetic, epigenetic, and post-transcriptional regulation. Integration of transcriptomic and proteomic data has the potential to be just as powerful for monitoring the sometimes subtle effects or dysfunction of protein production and localisation underlying neurodevelopment and disorder. In addition to the use of MS/MS proteomic data in validating or comparing to RNA abundance or translational efficiency, a major attraction of proteomics lies in the data obtained from the peptide sequences, which enable proteome-wide validation of genomic and transcriptomic variants, allelic imbalance, and isoform identification (**Figure 2**).

Allelic diversity across the human population is well known to influence brain development [ref], for example humans suffering microcephaly frequently carry a premature stop mutation in the gene ASPM, which is localised in the mitotic spindle, leading to a truncation of the protein and restricted growth of the cerebral cortex [ref]. Thus, even small modifications to the structure, abundance, or localisation of RNAs and proteins in the brain can have profound consequences. Integrated analyses of DNA- and RNA-sequence data obtained from the same individuals have resulted in a large number of discoveries relating to ADAR-mediated adenosine to inosine (A-to-I) editing of the transcriptome [ref]. **RNA-editing** appears to have played a significant role in human brain evolution [ref, ref, ref] and of the trio of ADAR proteins responsible for the post-transcriptional A-to-I modification, the third (ADAR3) is exclusively expressed in the brain [ref]. Several mis-sense RNA-edit sites in the AMPA receptor have been shown to alter the downstream behaviour of this protein, are edited at specific stages of human brain development [ref], and are required for normal brain function and phenotype in mice [ref]. Another example is the serotonin receptor 5-HT_{2C}R (HTR2C), which contains numerous RNA-editing sites that alter both the expressed protein sequence [ref] and cause an order of magnitude reduction in efficacy in the interaction of the receptor with its G proteins [ref]. Although these RNA-DNA differences (RDDs) tend to occur in intergenic and intronic regions,

several important instances RDDs have been found to alter the protein product of the transcript [ref brainspan]. Unfortunately, reliable identification of RDDs is technically very challenging and has led to erroneous false-positive identifications [ref], however MS/MS an extremely attractive tool for unbiased validation of mis-sense RDDs, which is capable of detecting not only the presence or absence of an RDD but provides quantitative data on the abundance of the edited and unedited copies of the proteins. The use of a proteome-wide, independent verification of mis-sense RDDs could equally be of great benefit to biochemical analyses of RNA-editing, such as ICE-seq [ref], in which inosine nucleotides are directly identified through chemical modification.

Within a given individual the **allele-specific expression** (ASE) of RNA is the result of epigenetic regulatory processes that are common across species and tissues [ref]. Detection of ASE relies on the discovery of heterozygous genomic variants that lead to an imbalance in the abundance of RNA produced from each parental allele. A recent survey of the mouse CNS revealed 1,300 genes exhibit an allelic imbalance in expression and, interestingly, during brain development this bias appears to favour the maternal allele, while in the adult a bias toward the paternal allele was observed between brain regions [ref]. Detection of allelic expression at the protein level relies on mis-sense mutations of one allele relative to the other, however such events occur with sufficient frequency to make MS/MS validations worthwhile. For example, an analysis of allele-specific protein expression in yeast reported that around 10% of heterozygous coding loci exhibit an allelic bias [ref], however the correlation of this bias to that observed at the mRNA level in the same genes was fairly poor (<0.35). In addition to verification of ASE events, an interesting application of MS/MS has been to assess allele-specific transcription factor binding (ASB), in which regions of the genome that are known to be heterozygous are purified and the collection of proteins bound to these regions are subsequently profiled by MS/MS in order to identify differential binding [ref].

Quantitative trait loci (QTL) have been extensively profiled in a variety of tissues using transcriptomics in terms of the relation of genomic variants to RNA expression changes (eQTL) and more recently, variants have also been related to protein abundance changes (pQTL) [ref]. In the prefrontal cortex, for example, such variants have been found to affect the expression of more than 100 genes [ref]. An analysis of individuals genotyped in the HapMap project reported that almost two-thirds of 185 detected cis-acting pQTLs were not found in a complementary analysis of the RNA [ref].

Alternative RNA-splicing (AS) is well known to be a highly tissue-specific process [ref] that greatly increases the complexity of the potential set of RNA molecules produced from multi-

exon genes. Splicing in the brain has been extensively profiled using transcriptomics and specific instances of alternative transcript useage have been implicated in neuropsychiatric disorders [ref]. During CNS cell development, for example, alternative splicing (AS) in neuronal progenitor stem cells has revealed different isoform usage at different stages of maturation to the final neuronal state [ref]. The differential expression of DISC1 isoforms that are associated with schizophrenia [ref], not to mention the translocation itself [ref]. Validation of splicing events by MS/MS is quite common, either in terms of determining the contribution of known isoforms in a given experiment [ref], or verifying the existence of novel exon junctions identified by RNA-sequencing [ref].

A recent result that potentially has a significant impact on both the future analyses of the CNS and interpretation of proteomics data is the finding that most genes, in a given cell-type or tissue, tend to express only a single dominant transcript/isoform [ref]. The myelin basic protein (MBP), for example, expresses a completely different isoform in the brain compared to all other healthy human tissues [ref]. This has a direct impact on the so-called interactome, in which differential isoform usage between CNS cell types and during the progression of neurological disorders affects protein-protein interactions, as exemplified in the Autism Spliceform Interaction Network [ref]. Currently full-length transcript [ref, ref] and isoform [ref] profiling technologies are immature [ref], however an early example of full-length transcript profiling in the brain directly observed the various isoforms of neurexin, showing their production mediates distinct protein interactions across the synapse [ref]. Such investigations will only become more common, and have great potential to significantly simplify the process of integrating and interpreting genome-wide measurements of RNA and protein in the near future.

Finally, there has been a lot of recent activity in identifying and ascribing potential functional roles to **fusion transcripts** and **non-protein-coding** regions of the genome. Efforts including ENCODE [ref] and others [ref], have reported that the union of all RNA molecules detected across a variety of tissues, cell-lines, and conditions infer that more than 75% of genomic DNA is at some point transcribed to RNA. This 'pervasive transcription' has caused some controversy and confusion, not least when considering whether the presence of these molecules may imply they have a functional role in cellular processes [ref]. In terms of understanding the functional output of genomes of previously un-annotated organisms, transcriptomics alone is insufficient to accurately define the cohort of protein coding sequences, even when combined with ribosome profiling; high-throughput proteome profiling by MS/MS again provides the most useful avenue to this.

There have been many proteomic studies of the CNS by MS/MS, and these have been

Rob Kitchen 5/1/14 3:48 PM
Comment [4]: not brain

enumerated in previous reviews [ref, ref, ref]. More recent applications of MS/MS to the CNS include the profiling of neural tissue in model organisms such as the fruit fly [ref] and neurodegeneration in zebrafish [ref], as well as profiling cultures of specific CNS cell-types in primary culture such as neurons [ref] and oligodendroglial cells [ref]. Fluorescence activated cell sorting (FACS) of microglia has provided more than 100 genes that are enriched compared to neurons and other oligodendrocytes [ref] and enabled MS/MS assessment of synaptic proteins [ref, ref]. A novel method, fluorescence activated nuclei sorting (FANS), in combination with an antibody against the neuronal-specific splicing protein, NeuN, has been successfully applied to purify neuronal nuclei from primary tissue in order to provide a quantitative comparison of the abundance of nuclear proteins with astrocytes and oligodendrocytes [ref].

Challenges in assessing the proteome of the CNS and implications for future studies in quantitative neuroproteomics

There exist two particularly challenging issues regarding to proteomic profiling of the CNS. Both cellular heterogeneity and the dynamics of protein turnover are critical components of the long term adaptation of specific classes of neurons, specifically at the synapse, to external stimuli such as stress, drugs of abuse, and neurodegenerative illness. There are methods to monitor protein dynamics, but these are not particularly applicable to the mammalian CNS.

Monitoring rates of protein production and degradation can be achieved fairly straightforwardly in cultured cells by pulse-labelling followed by SILAC proteomics [ref]. Similarly, nascent peptide chains can be captured as they are synthesised by the ribosome using a biotin-puromycin labelling approach [ref]. However these approaches do not lend themselves to assessment of rates of protein turnover in mammalian and human tissues. Integration with transcriptome profiling, specifically ribosome profiling, is an increasingly attractive proxy to direct measurement of protein synthesis. Similarly, non-invasive imaging methods are potentially valuable methods by which differential rates of protein synthesis can be repeatedly observed in the mammalian CNS [ref], both over long and relative short periods of time.

The major obstacle to any genome/transcriptome/proteome wide study of neuronal cells derives from their immense heterogeneity. The complexity of the human brain is reflected in ~86 billion neurons, and at least an equal number of glial cells [ref], which can be further subdivided into hundreds of different types based on their morphology, connectivity, and molecular and electrophysiological properties. All of the different CNS cell types develop and

are integrated into functional networks within very precise constraints; deviations from this normal course of development can lead to a variety of disorders. Furthermore, the sub-cellular localisation of RNAs and proteins as well as the rapid and cell-type specific production of specific genes are fundamental to neuronal development, function, and disease [ref]. Proteins localised, for example, at the **post-synaptic density** (PSD) are of interest in the fields of addiction and substance abuse [ref]. The abundance of particular PSD proteins is dependent on cell-type and brain region [ref], however transcriptomic analysis are sub-optimal for assaying such differences due to the confound introduced by the RNA trafficking and/or local translation at synapses or potentially in axons [ref]. Proteomic profiling can be used to directly access the protein complement of the PSD, however such purifications can introduce significant variability in the measured abundances of some of these protein [ref]. Similarly, **density/gradient centrifugation** for nuclear/organelle purification offers sub-cellular resolution [ref], but this will always be confounded by inter-cellular variability unless also applied to homogeneous collections of cells.

The molecular diversity both between and within the traditional neurotransmitter classes in the mammalian CNS has been observed not only based on the absolute presence or absence of proteins, but also in their relative abundance levels, further supporting the need for unbiased, comprehensive, and quantitative measurement of isoform expression at a higher resolution than is available at the whole-tissue level [ref]. Existing whole-tissue analyses [ref brainspan, nenad] are not sensitive to this small scale inter- or intra-cellular variability and suffer from confounded and diluted signals from not only different classes of neurons but also from the high proportion of glia. There are currently a variety of approaches being used to overcome this issue of heterogeneity in the mammalian brain, including **ISH** atlases of spatial and temporal RNA expression [ref allen] and the somewhat limited **immunohistochemistry** atlases of spatial protein abundance [ref]. This approach is severely limited however as it is very low-throughput, is subject to variable antibody specificity and is, at best, semi-quantitative.

Analysis of single or small numbers of neural cells, followed by transcriptional profiling is an attractive alternative, providing individual quantitative measurement of all RNAs in a very small physical volume of tissue. However LCM of neural tissue is not guaranteed to result in single-cell specificity due to the close proximity and overlapping processes of neuronal cells, especially between the layers of the cortex [ref]. Additionally, the throughput of such an approach is still too low to meaningfully assess, for example, the response of collections of neurons in a given brain region to experimental variables such as the treatment effect a drug, especially when there is a requirement for multiple biologically independent subjects. Current

proteomic technologies would require larger numbers of cells to be collected by LCM in order to obtain enough material, further exacerbating the issues of contamination due to non-target cell types.

The ultimate application of LCM for proteomics is to enable **single-cell** protein analyses. For several years, researchers have been using qPCR to assess transcript abundances obtained from individual cells and very recently RNA-sequencing of single cells has expanded these analyses to the whole transcriptome. What has been found mirrors observations at a whole-tissue level [ref], in that a given gene in a given cell typically transcribes a single isoform [ref]. Moreover, very recent observations have even suggested that mammalian cells express these isoforms randomly from a single allele [ref], revealing perhaps novel regulatory mechanisms within the cell to produce this behaviour. There is no doubt that with the continued development of MS-based methods for low sample input, new and exciting biology will emerge that advances our understanding of transcriptional and translational programmes both within the cell and in cell-to-cell signalling [ref]. However, as is the case for LCM, the utility of single-cell analyses for studying the effects of brain development, malfunction, and effect of chemical treatment, is very limited due to the issues of sample-throughput and resolution to extremely low abundance molecules. Fortunately, there are other approaches to studying not single neurons, but single populations of neurons that are more compatible with existing technologies and do not suffer the significant issues regarding throughput and contamination as do LCM / single cell techniques.

An extremely elegant means of obtaining quantitative spatial expression measurements for a specific gene is through the creation of libraries of bacterial artificial chromosomes (BAC) containing fluorescent markers downstream of target regulatory elements (enhancer/promoters) that themselves lie upstream of the desired gene [ref]. Such an approach, when collected for multiple regulatory elements, or when used in combination with ISH/IHC spatial expression profiles can be used to reveal cell-type specific promoter activity and RNA expression [ref]. A complementary method for obtaining all cytosolic RNA expressed in a given cell-type leads directly from this identification of cell-type specific promoters, in which overexpression of a **GFP-labelled ribosomal protein** under the control of one such cell-type-specific promoter allows purification by IP of the transcripts bound by the polyribosome in the desired cells [ref]. The principal advantage of this approach is that cellular material is obtained in the same way as if profiling a standard tissue extract, except the introduction of the eGFP-IP removes RNAs from non-target cell-types. Moreover, due to the labelling of the ribosomal protein, strong GFP signal is also observed in the nucleolus of target cell types enabling FACS purification of target nuclei

for assessment of DNA, histone modifications, transcription factor binding, or nascent RNA transcription in the specific cell-type of interest [ref, ref]. The obvious disadvantage to these approaches is that they are only compatible with systems amenable to transfection, such as cultures or rodent models, and so of limited utility for directly studying human neurobiology. Furthermore, although the FACS approach can be used to profile proteins in the nuclei of the selected cell-type, cytosolic proteins are much more difficult to obtain by such a method thanks to the extensive processes of neuronal cells. For application to human neurobiology, and as a refinement of the FACS ideology, it may become possible to exploit proteins at the plasma membrane of specific neuronal subtypes for purification by antibody pulldown. Use of such cell-surface markers may be able to enable cell-type specific analyses of RNA and protein.

<<BOX 1>>The various methodologies for contemporary MS/MS have been comprehensively covered elsewhere including excellent reviews on the mass-spectrometer technologies [ref], computational analysis of spectra [ref], and specific examples of proteomics applied to the CNS [ref]. For a given sample, the tens to hundreds of thousands of peptides, typically the products obtained using trypsin digestion, are individually quantified, but the abundances from all peptides derived from a given protein can be aggregated to facilitate protein-level expression analyses in addition to the simpler peptide-level comparisons. Briefly, MS/MS experiments are distinguished by two choices, illustrated in **Figure 1**; the first being the method of quantification of the observed protein products and the second whether the experiment should be hypothesis-driven or hypothesis-free. Labelled MS/MS analysis methods such as Stable Isotope Labeling by Amino acids in cell Culture (SILAC) [ref], Stable Isotope Labeling of Mammals (SILAM) [ref], and Isobaric Tag for Relative and Absolute Quantitation (iTRAQ) [ref] are commonly employed for assessing differential abundance of proteins across phenotypes, conditions, or treatments; while label-free abundance estimates [ref] also allow relative quantitation of peptides to each other. Absolute quantitation can be achieved, in an increasingly high-throughput manner, using stable isotope dilution (SID) of a number of target proteins by spiking in synthetic labelled sequences that are exact analogues of the target sequences [ref]. Improvements to both labelled and label-free quantitation can be achieved by restricting the MS/MS scans to pre-defined ranges; this approach, termed Selected Reaction Monitoring (SRM) or Multiple Reaction Monitoring (MRM) [ref, ref], essentially allows the instrument to more accurately measure the

abundance of selected peptides by spending a greater fraction of total instrument time monitoring a smaller list of pre-defined peptides. Recently, however a renewed interest in data-independent spectra acquisition [ref], driven by faster and more accurate mass-spectrometers, has led to the development of software such as OpenSWATH which claims to be capable of detecting 30% more proteins than conventional label-free acquisition [ref].**<<END BOX 1>>**

<<START ABBREVIATIONS>>

MS/MS	Tandem mass-spectrometry
MALDI	Matrix Assisted Laser Desorption/Ionisation
iTRAQ	Isobaric Tag for Relative and Absolute Quantitation
SILAC/M	Stable Isotope Labeling of Culture/Mammals
SID	Stable Isotope Dilution
S/MRM	Selected/Multiple Reaction Monitoring
RNA-seq	Second-generation, massively parallel RNA-sequencing
fMRI	Functional Magnetic Resonance Imaging
dMRI	Diffusion Magnetic Resonance Imaging
PET	Positron Emission Tomography
ISH	In-Situ Hybridisation
IHC	Immunohistochemistry
ES cell	Embryonic Stem cell
PTM	Post-translational modification
[u]ORF	[upstream] Open reading frame
RDD	RNA-DNA differences, arising due to RNA-editing by ADAR
ASE/B	allele specific expression/binding
e/pQTL	expression/protein quantitative trait loci
ENCODE	Encyclopedia of DNA Elements project
gEUVADIS	Genetic European Variation in Health and Disease project
TCGA	The Cancer Genome Atlas project

GTEx	Genotype-Tissue Expression project

<<END ABBREVIATIONS>>