Using bioinformatics for the identification of key peptides to engineer dopamine neurons. Towards a therapy for Parkinson's disease.

William D. Mitchell[1], Rowan P. Orme[2], Sarah R. Hart[2], Rosemary A. Fricker[2]

[1] *Centre for Biological Engineering, Loughborough University, LE11 3TT, United Kingdom*
[2] *Institute for Science & Technology in Medicine, Keele University, ST5 5BG, United Kingdom*

Corresponding Author:

William D. Mitchell[1]

Centre for Biological Engineering, Loughborough University, LE11 3TT, United Kingdom

Email address: *w.mitchell@lboro.ac.uk*

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26443v2 | CC BY 4.0 Open Access | rec: 17 Jan 2018, publ: 17 Jan 2018

Abstract

Parkinson's disease is a widespread condition caused by degeneration of dopamine neurons in the midbrain. A number of proteins are known to be important to signalling mechanisms present in the midbrain during natural dopamine neuron development, and may be utilised to better produce dopamine neurons *in vitro*. Relative expression levels of proteins were obtained from substantia nigra tissue of rats from embryonic days E11 through E14 using isobaric tagging for relative and absolute quantification. This project analysed the dataset obtained, with an emphasis on relative expression levels of proteins across the four-day period. Bioinformatics searching of online databases reduced the dataset from 3325 proteins to a shortlist of five worthy of further investigation. It is hoped that the proteins identified using these techniques will help to refine protocols for the production of dopamine neurons *in vitro*.

Keywords: Proteomics; Bioinformatics; Parkinson's disease; iTRAQ; Dopamine neuron; Neural development; Stem cells

Introduction

Parkinson's disease is a widespread condition caused by degeneration of dopaminergic neurons in the midbrain leading to a lack of motor control (1). Medications and surgical interventions to alleviate symptoms are currently available; however, they grow ineffective and produce involuntary movement as neuron degeneration continues. There is currently no available therapy capable of slowing disease progression or preventing further neuron degeneration. Stem cell based therapies offer a way to replace dead or damaged dopamine neurons and restore motor functionality (2). As the adult neurons involved with motor function do not divide, cells from other sources are required. There are many stem cell based sources currently being explored, with foetal neuronal stem cells, embryonic stem cells, induced pluripotent stem cells, adult neural stem cells, and adult bone marrow stem cells all showing potential as sources for neuron replacement therapy (3). Stem cells must be expanded and differentiated in culture in order to produce adult dopamine neurons and may be manipulated by activating or inhibiting signalling pathways. There are many techniques used to increase the efficiency of producing neurons in culture, one of which is to recreate the signalling mechanisms present in the midbrain during natural dopamine neuron development. A number of peptides have been found to play important roles in these processes, while many are yet to be investigated.

A protein expression data set was generated for developing rat midbrain tissue; the tissue that later develops into the dopamine neurons in the substantia nigra whose degeneration causes Parkinson's disease (4). Previous selection of a candidate from this dataset revealed that vitamin D plays an important role in dopamine neuron development and demonstrated that its controlled delivery improves dopamine neuron yield *in vitro* (5). This project reanalyses the dataset with an emphasis on relative expression levels of proteins across four days of embryonic development in order identify further proteins of interest for the improved production of dopamine neurons *in vitro*.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26443v2 | CC BY 4.0 Open Access | rec: 17 Jan 2018, publ: 17 Jan 2018

Methods

The protein expression dataset previously described in (4) was created using the proteomics technique of isobaric tagging for relative and absolute quantification (iTRAQ). This technique allows the expression levels of proteins from different sources to be determined in a single experiment (6). The samples used to generate the dataset were obtained from the substantia nigra of rats at embryonic days E11, E12, E13 and E14, assigned iTRAQ markers 114, 115, 116 and 117 respectively. Tissues were collected under an establishment licence for Keele University (PEL 40/2407). Protein identification and quantification profiles were originally generated by ProteinPilot and exported as Excel spreadsheets. Protein identification was based on a combination of the number of peptides identified, and the similarity between the observed and expected mass for each peptide. Identified proteins were matched to entries in the NCBI Reference Sequence Database (7), with most proteins consisting of multiple peptide component matches. Details of the dataset generation are provided in (4). Following screening for proteins with total ion score confidence intervals of above 95%, the dataset used for this analysis consisted of expression level data for 3325 NCBI database matched proteins.

The expression change ratio between embryonic days was calculated for all proteins to allow the comparison of relative protein levels for neighbouring days. Data was then fit to patterns of interest in order to exclude proteins with no significant changes over days E11 through E14. Nine patterns of interest were selected in order to capture peaks or troughs of protein activity over the four-day period (Figure 1). The filter function of Excel was used to specify cut-off values for relative expression values, allowing proteins to be fit to patterns with little manual manipulation. The stringency of pattern fitting is therefore controllable through the selection of cut-off values for each expression change. Figure 1 shows relative protein expression levels where a significant change is considered to be an increase or decrease in expression of at least a factor of two over a single day.

Proteins meeting the expression change conditions were classified by tissue type according to data from the UniProt Knowledgebase (UniProtKB) database. Proteins associated with relevant tissues were carried forward for the next round of analysis while those with no known association were discarded. Following tissue categorisation, proteins were classified according to their molecular function again using data present in the UniProtKB database. Once proteins had been classified by expression pattern, tissue type and molecular function, a shortlist of potentially interesting proteins was produced. Further prioritisation of classification categories was then performed until a manageable shortlist was produced for further investigation.

Results

Classification of proteins according to the expression level patterns reduced the original dataset from 3325 down to 96 proteins of interest. The complete set of expression changes for all proteins is shown in Figure 2. Expression level changes were recorded as the ratio of the expression level on each day relative to the expression level on the following embryonic day. Plotting the $\log_2$ value of this ratio allows the magnitude of expression level changes to be

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26443v2 | CC BY 4.0 Open Access | rec: 17 Jan 2018, publ: 17 Jan 2018

shown symmetrically regardless of the direction of change. An increase or decrease of a factor of two was required in order for a change to be considered significant. Expression level changes above a factor of two lay outside the horizontal red lines, while those under a factor of two are located within the red lines.

The distribution of proteins over the patterns of interest is shown in Figure 3. The majority of proteins featuring a significant change in expression level over the four-day period showed a large decrease in expression from E11 to E12 and were therefore were categorised as pattern A. Proteins in this category are likely involved in neurogenesis which is thought to peak at around E11 (8). The proteins were then categorised according to tissue type and were found to fit into four main groups: neural, blood, other tissues, and ubiquitous (Figure 4). One protein was excluded as it lacked any tissue information on UniProtKB and related databases. The final classification was according to molecular function using the categories: binding, enzyme, enzyme regulator, receptor, structural, transcription factor, translation and transport. The distribution of proteins between these groups is shown in Figure 5.

Following classification by expression pattern, tissue type and molecular function, it became possible to easily further prioritise categories guided by initial literature based research. Proteins matching expressions patterns A, B and E were carried forward as these patterns feature high expression levels on E11 and E12, the time at which peak neurogenesis is thought to occur (8). Carrying forward only these proteins reduced the dataset from 96 to 68 proteins. Proteins were then filtered according to tissue type, further reducing the list from 68 to 24 proteins. Proteins with no known link to neural tissue were removed, as well as ubiquitous proteins, which were considered unlikely to promote dopamine neuron growth specifically. Initial research into biological functions also found that many proteins present in blood as well as neural tissue were primarily associated with biological roles in blood, and so this group was discarded. It was decided not to filter proteins based on their molecular function, as there was much crossover with most proteins involved in multiple functions.

The final stage of the investigation was a literature search of the 24 remaining proteins. This was carried out with the aim of finding known connections to dopamine neurons or general neuron development and was performed manually in order to remove proteins included due to database errors or with unsubstantial evidence. Most proteins discarded during the manual investigation phase were discarded due to a lack of published evidence to support classifications present in online databases. It is likely that false negatives were present within the “blood and neural” and “blood, neural and other” groups and were discarded during filtering according to tissue type. Following this final stage of investigation, the shortlist of five proteins given in Table 1 was produced. The relative expression pattern of each shortlisted protein is provided in Figure 6.

A2M is an inactive form of the large plasma protein A2M, produced by the liver and present in blood. It is capable of binding to brain-derived neurotrophic factor and nerve growth factor (9). CMP-NeuNAc synthase is an enzyme that catalyzes the activation of N-acylneuraminate (NeuNAc) to CMP-N-acylneuraminate (CMP-NeuNAc), a substrate required for the addition of

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26443v2 | CC BY 4.0 Open Access | rec: 17 Jan 2018, publ: 17 Jan 2018

sialic acid (10). Salaic acid is found in high levels in the brain and is essential in synaptogenesis and for enabling neural transmission (11). P2RX4 is found in the central and peripheral nervous systems and has been shown to regulate synaptic strengthening (12). RTN1 is a member of the reticulon family of proteins which aid membrane curvature and have been shown to be involved with neuron differentiation, neuroendocrine secretion (13). GSK-3β is an enzyme capable of negatively regulating the Wnt signalling pathway, a key element of dopamine neuron development (14).

The expression levels of A2M, CMP-NeuNAc synthase, P2RX4 and RTN1 all matched pattern A, showing a peak of expression levels on E11 followed by a sharp decrease for E12. GSK-3β presented a more complicated expression pattern and was matched to patterns B, D, E and I during automated pattern fitting. Manual inspection of the data showed that this protein exhibited two peaks in expression levels, one at E12 and another at E14.

Discussion

Categorisation of proteins based on their expression patterns over E11 to 14 allowed an initially large dataset of 3325 proteins to be quickly reduced to 96. Further categorisation of proteins according to tissue type and molecular function using data from online databases allowed a shortlist of five proteins to be generated with minimal manual literature research.

Patterns featuring periods with no expression change (e.g. E12 to E13 and E13 to E14 in pattern A) did not have the lack of an expression change enforced during automated pattern fitting in order to include proteins with multiple significant changes in the same direction. These proteins would otherwise not be captured using patterns with only two expression level states (high and low). Although this solution successfully included proteins with expression patterns featuring multiple significant changes, some proteins were also matched to patterns that did not accurately describe their true expression levels, as occurred with GSK-3β. A more complete solution to pattern matching would be to use patterns featuring three or four states as shown in Figure 7; however, the number of possible patterns in these cases increases the complexity of the method (16 possible patterns with two states, 81 possible patterns with three states, 256 possible patterns with four states).

A limitation of pattern fitting as performed in this study is that expression levels on E10 and E15 must be extrapolated from data present for E11 through E14. Future studies may benefit from utilising the full eight samples possible in iTRAQ to gain information over a longer period (15). This approach has added flexibility as the additional samples may also be utilised to increase the resolution on days of interest (two samples per day giving 12-hour expression windows for example). As a result of E10 and E15 being unknown, there exists a positive bias within the pattern matching method towards types A, D, H and I as these patterns feature a doubling or halving condition followed by or preceding a day for which there is no data. There is also a negative bias away from types B, C and F, as data must pass two expression change conditions in order to positively match these expression patterns.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26443v2 | CC BY 4.0 Open Access | rec: 17 Jan 2018, publ: 17 Jan 2018

The final manual stage of short listing is essential as the automated classification of proteins, as well as their initial identification from sequence data, relies entirely on online protein databases. Correctly matching mass spectrometry data to proteins in online databases is known to be a primary limitation of mass spectrometry based proteomics due to the large number of names used simultaneously for many proteins, as well as the wide range of available resources (16). This issue has been somewhat addressed through the use of the UniProtKB database, a collaborative effort between the European Bioinformatics Institute (EBI), the Protein Information Resource (PIR) and the Swiss Institute of Bioinformatics (SIB) (17).

Conclusion

This analysis of relative protein expression levels across four key days of embryonic development coupled with data from online proteomics databases demonstrates a technique to obtain a shortlist of proteins with a minimal requirement for manual literature research. It is hoped that the proteins and peptides identified using these methods will help to refine protocols for the production of dopamine neurons *in vitro*.

Acknowledgements

This research was funded by an EPSRC Centre for Doctoral Training studentship awarded to WM. The dataset used in the analysis was generated from research funded by Parkinson's UK.

References

1. Orme R, Fricker-Gates RA, Gates MA (2009) Ontogeny of Substantia Nigra Dopamine Neurons: Birth, life and death of dopaminergic neurons in the substantia nigra. Springer.
2. Lindvall O, Björklund A (2004) Cell therapy in Parkinson's disease. NeuroRx 1(4): 382–93.
3. Fricker-Gates RA, Gates MA (2010) Stem cell-derived dopamine neurons for brain repair in Parkinson's disease. Regen Med. 5(2): 267–78.
4. Orme RP, Gates MA, Fricker-Gates RA (2010) A multiplexed quantitative proteomics approach for investigating protein expression in the developing central nervous system. J Neurosci Methods 191(1): 75–82.
5. Orme RP, Bhangal MS, Fricker RA (2013) Calcitriol imparts neuroprotection in vitro to midbrain dopaminergic neurons by upregulating GDNF expression. PLoS One. 8(4):e62040.
6. Wiese S, Reidegeld KA, Meyer HE, Warscheid B (2007) Protein labelling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research. Proteomics. 7(3): 340–50.
7. Acland A, Agarwala R, Barrett T, Beck J, Benson DA, et al. (2014) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 42(D1):D7–17.
8. Gates MA, Torres EM, White A, Fricker-Gates RA, Dunnett SB (2006) Re-examining the ontogeny of substantia nigra dopamine neurons. Eur J Neurosci. 23(5): 1384–90.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26443v2 | CC BY 4.0 Open Access | rec: 17 Jan 2018, publ: 17 Jan 2018

9. Skornicka EL, Shi X, Koo PH (2002) Comparative binding of biotinylated neurotrophins to α2-macroglobulin family of proteins: Relationship between cytokine-binding and neuro-modulatory activities of the macroglobulins. J Neurosci Res. 67(3): 346–53.
10. Lawrence SM, Huddleston KA, Tomiya N, Nguyen N, Lee YC, et al. (2001) Cloning and expression of human sialic acid pathway genes to generate CMP-sialic acids in insect cells. Glycoconj J. 18(3): 205–13.
11. Wang B (2009) Sialic acid is an essential nutrient for brain development and cognition. Annu Rev Nutr. 29(3) :177–222.
12. Baxter AW, Choi SJ, Sim JA, North RA (2011) Role of P2X4 receptors in synaptic strengthening in mouse CA1 hippocampal neurons. Eur J Neurosci. 34(2): 213–20.
13. Hens J, Nuydens R, Geerts H, Senden NHM, Van De Ven WJM, et al. (1998) Neuronal differentiation is accompanied by NSP-C expression. Cell Tissue Res. 292(2): 229–37.
14. Castelo-Branco G, Arenas E (2006) Function of Wnts in dopaminergic neuron development. Neurodegener Dis. 3(1–2):5–11.
15. Fuller HR, Morris GE (2012) Quantitative Proteomics Using iTRAQ Labeling and Mass Spectrometry. Integr Proteomics. 347-362.
16. Bell AW, Deutsch EW, Au CE, Kearney RE, Beavis R, et al (2009) A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. Nat Methods. 6(6):423–30.
17. Consortium TU (2008) The Universal Protein Resource. Nucleid Acid Res. 36: D190–5.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26443v2 | CC BY 4.0 Open Access | rec: 17 Jan 2018, publ: 17 Jan 2018

Accession Number	Protein Name	Expression Pattern	Tissue Type	Molecular Function
gi|158138551	Alpha-2-macroglobulin precursor (A2M)	A	neural, other	binding, enzyme regulator
gi|68059163	N-acylneuraminate cytidylyltransferase (CMP-NeuNAc synthase)	A	neural	enzyme
gi|149063348	Purinergic receptor P2X, ligand-gated ion channel 4, isoform CRA_d (P2RX4)	A	neural	receptor
gi|16758732	Reticulon-1 (RTN1)	A	neural	transport
gi|125374	Glycogen synthase kinase 3 beta (GSK-3β)	B, D, E, I	neural	binding, enzyme, enzyme regulator, receptor

239
240 Table 1 - Proteins selected for the final shortlist based on expression pattern, tissue type and molecular
241 function.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26443v2 | CC BY 4.0 Open Access | rec: 17 Jan 2018, publ: 17 Jan 2018

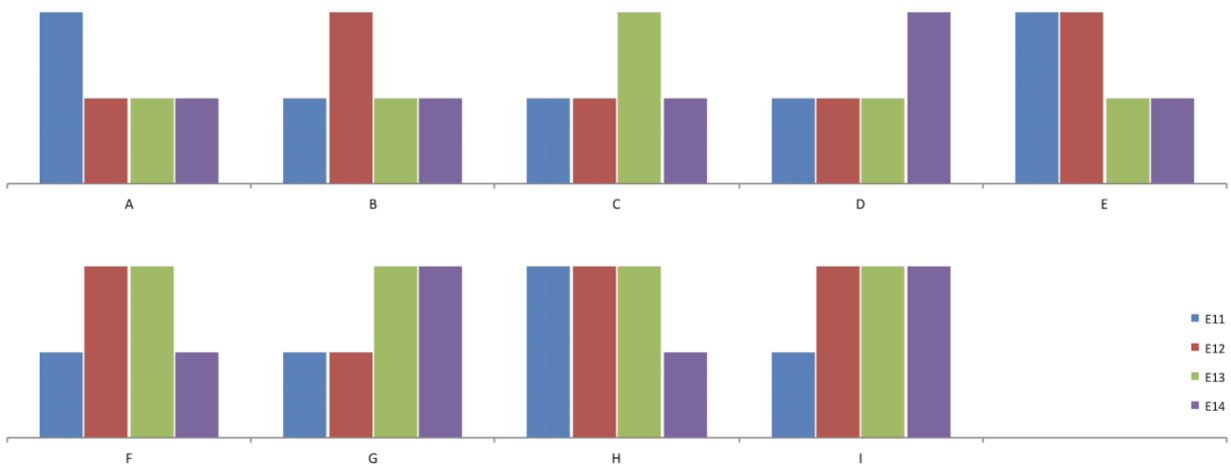


Figure 1 - Expression patterns of interest based upon expression levels relative to neighbouring days.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26443v2 | CC BY 4.0 Open Access | rec: 17 Jan 2018, publ: 17 Jan 2018

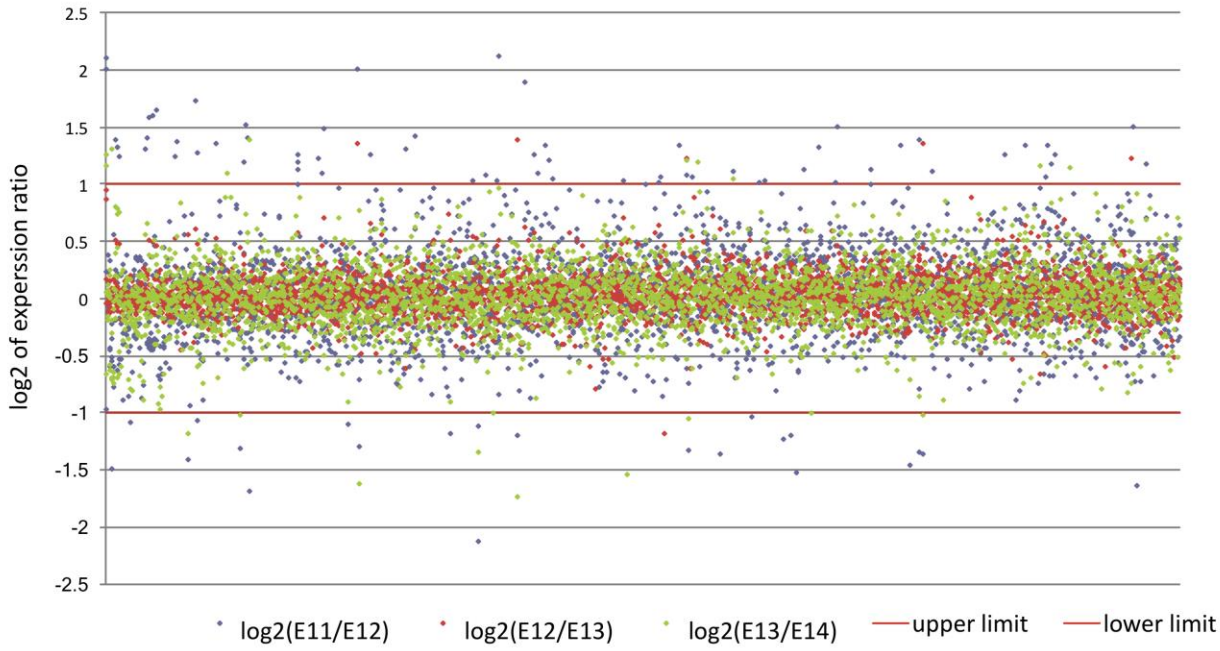


Figure 2 - Expression change ratios from E11 to E12, E12 to 13 and E13 to E14. The majority of expression changes were small and lie inside the cut-off points, plotted as red horizontal lines.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26443v2 | CC BY 4.0 Open Access | rec: 17 Jan 2018, publ: 17 Jan 2018

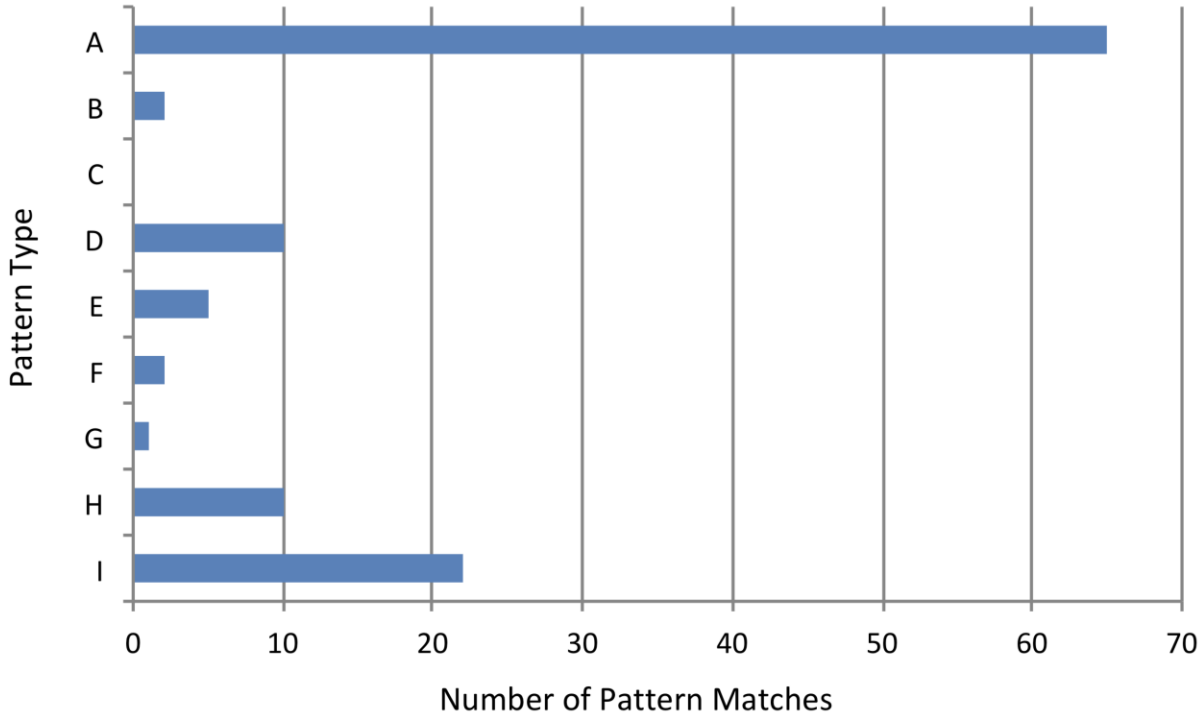


Figure 3 - The majority of proteins were found to fit pattern A, showing peak expression at E11 followed by a decrease at E12.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26443v2 | CC BY 4.0 Open Access | rec: 17 Jan 2018, publ: 17 Jan 2018

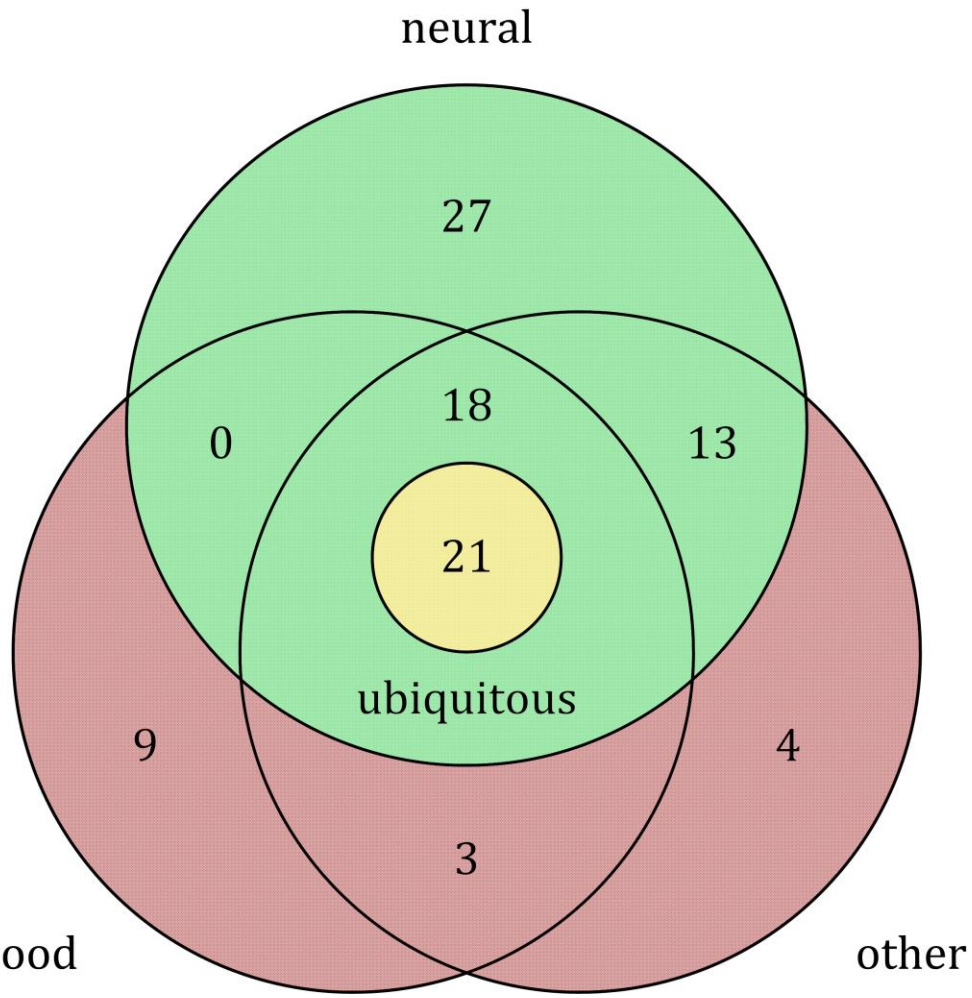


Figure 4 – Protein distribution across tissue types. Most proteins were linked to neural tissue, while many were also associated with blood and other tissues.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26443v2 | CC BY 4.0 Open Access | rec: 17 Jan 2018, publ: 17 Jan 2018

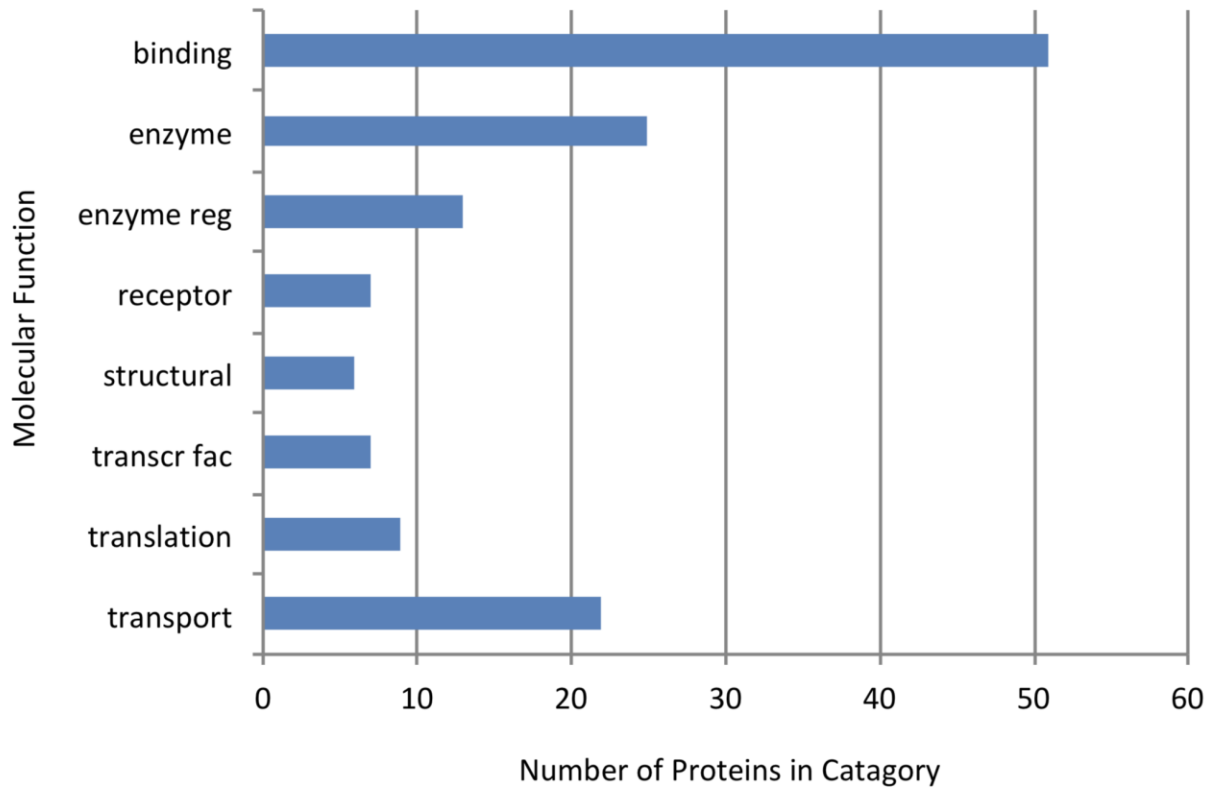
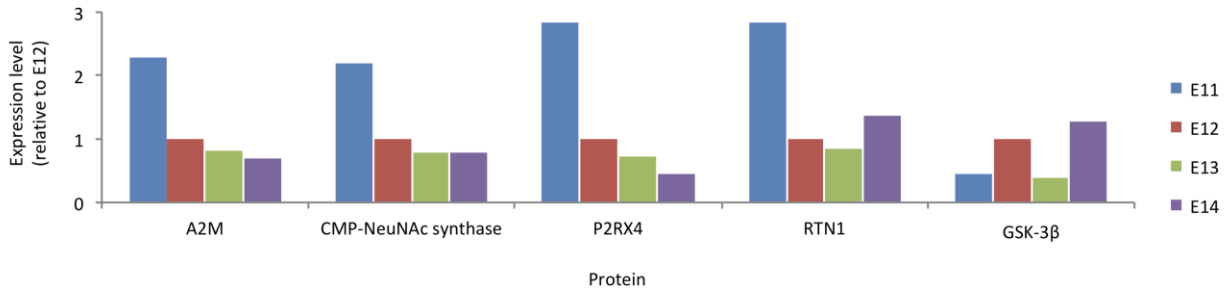


Figure 5 - Proteins were found to be mostly associated with binding activity, with many proteins being involved in multiple molecular functions. Enzyme regulator has been abbreviated to “enzyme reg” and transcription factor has been abbreviated to “transcr fac”.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26443v2 | CC BY 4.0 Open Access | rec: 17 Jan 2018, publ: 17 Jan 2018



261
262
263 Figure 6 - Relative expression levels for proteins on the final shortlist, normalised to their expression
264 level on embryonic day 12.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26443v2 | CC BY 4.0 Open Access | rec: 17 Jan 2018, publ: 17 Jan 2018

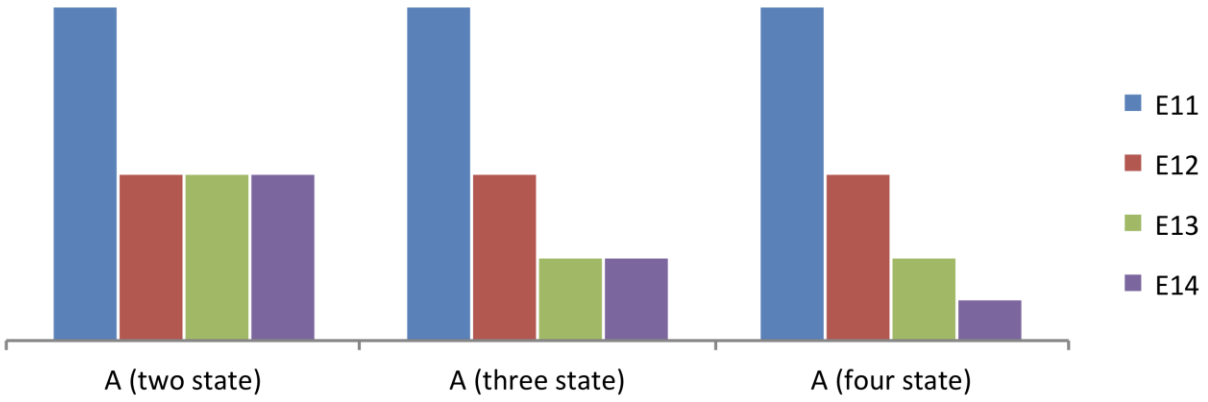


Figure 7 - Expression patterns A (three state) and A (four state) are included when fitting data to pattern A (two state) provided conditions for unchanging days are unenforced.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26443v2 | CC BY 4.0 Open Access | rec: 17 Jan 2018, publ: 17 Jan 2018