Executable Analysis Document Supporting:

Proteomic Profiling of the Substantia Nigra to Identify Determinants of Lewy Body Pathology and Dopaminergic Neuronal Loss

Part II: LC-MS/MS Data Analysis

Vladislav A. Petyuk[1], Lei Yu[2,3], Heather M. Olson[4], Fengchao Yu[5], Geremy Clair[1], Wei-Jun Qian[1], Joshua M. Shulman[6,7], and David A. Bennett[2,3]

[1]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA
[2]Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago, IL, USA
[3]Department of Neurological Sciences, Rush University Medical Center, Chicago, IL, USA
[4]Enviromental and Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA, USA
[5]Department of Pathology, University of Michigan, Ann Arbor, MI, USA
[6]Departments of Neurology, Molecular & Human Genetics, and Neuroscience, Baylor College of Medicine, Houston, TX, USA
[7]Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, TX, USA

February 25, 2021

Contents

1 Objective 1

2 PCA. Supplementary Figure 3 1

3 Statistical significance analysis 3
- 3.1 Mixed effects model. Figure 3 . . . 3
- 3.2 Top 10 most significant proteins proteins. Figure 4 . . . 4
- 3.3 Top differential proteins for six contrasts. Table 2 . . . 5
- 3.4 Full results of contrasts analysis. Supplementary Table 4 . . . 5

4 GO Term Enrichment 6
- 4.1 Top terms. Table 3 . . . 6
- 4.2 Complete results. Supplementary Table 8 . . . 7
- 4.3 Non-redundant set of GO terms . . . 7
  - 4.3.1 Barplot summary of non-redundant terms. Figure 6 . . . 8
  - 4.3.2 Heatmaps corresponding to individual terms. Figure 7 . . . 9

5 Assessment of the cell type level changes 11
- 5.1 Significance analysis . . . 11
- 5.2 Barplot of cell type contributions. Figure 8 . . . 11
- 5.3 Heatmap of protein markers of endothelial cells. Figure 9 . . . 12

6 Primary factors affecting α-synuclein abundance 13
6.1 Statisical model of α-synuclein abundance 13

7 Causal Network Analysis 13
7.1 Causal Network. Figure 10 14

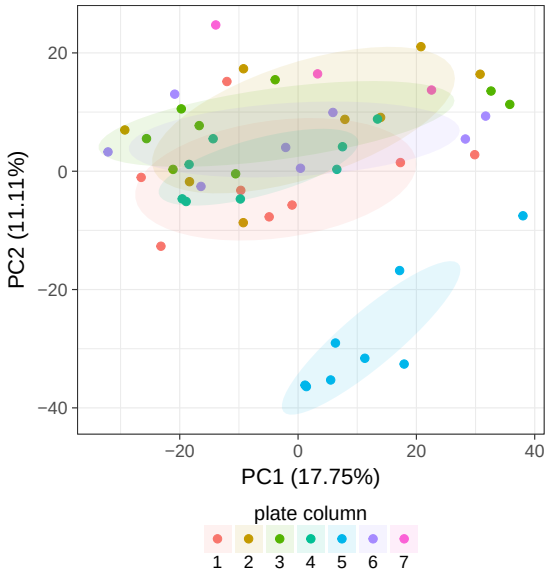
8 Session information 15

1 Objective

The ultimate goal of the project is to explore proteome changes associated with the histopathological features of the Parkinson's disease: Lewy bodies and *Substantia nigra* neuronall loss. This is a new perspective in analysis of Parkinson's disease etiology. As opposed to using clinical diagnosis for study design we rely on molecular and cellular changes presumably preceeding the onset of Parkinson's disease.

2 PCA. Supplementary Figure 3

PCA is a common approach to identify structure within the data. For example, it can help to identify and visualize the extend of the batch effects. Columns of the 96-well plate were identified as a source of technical bias. Thus it should be included into the follow-up statistical models.



Corresponds to **Supplementary Figure 3**. PCA plot of relative protein abundances. Samples colored according to the column of the 96-well plate. The "batch" effect associated with the column position is substantial to warrant it as a covariate in the statistical model.

Percentage of proteins that statistically significantly affected by plate column effect.

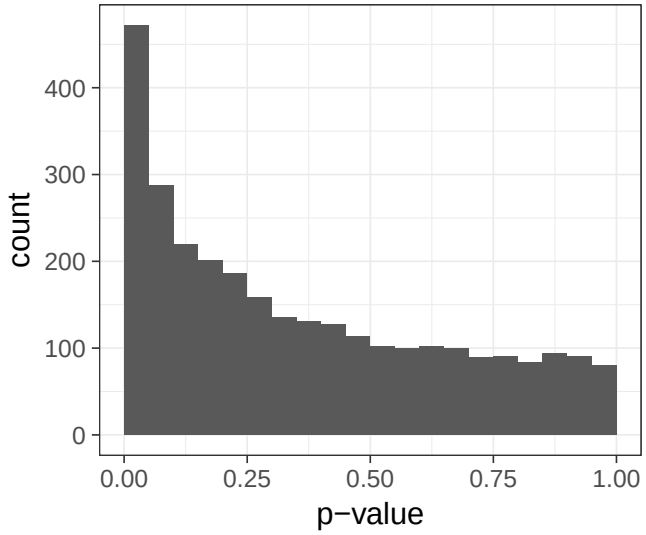
3 Statistical significance analysis

3.1 Mixed effects model. Figure 3

The protein abundances were modeled with linear mixed effects approach. The terms to consider are the subject type as fixed effect and matching group with 96-well plate column as random effects. The full model in Wilkinson-Rogers notation is:

$$protein \sim subject.type + (1|PlateCol) + (1|match.group)$$

The null model excludes the fixed subject type coefficient. The results of ANOVA significance analysis are shown below.



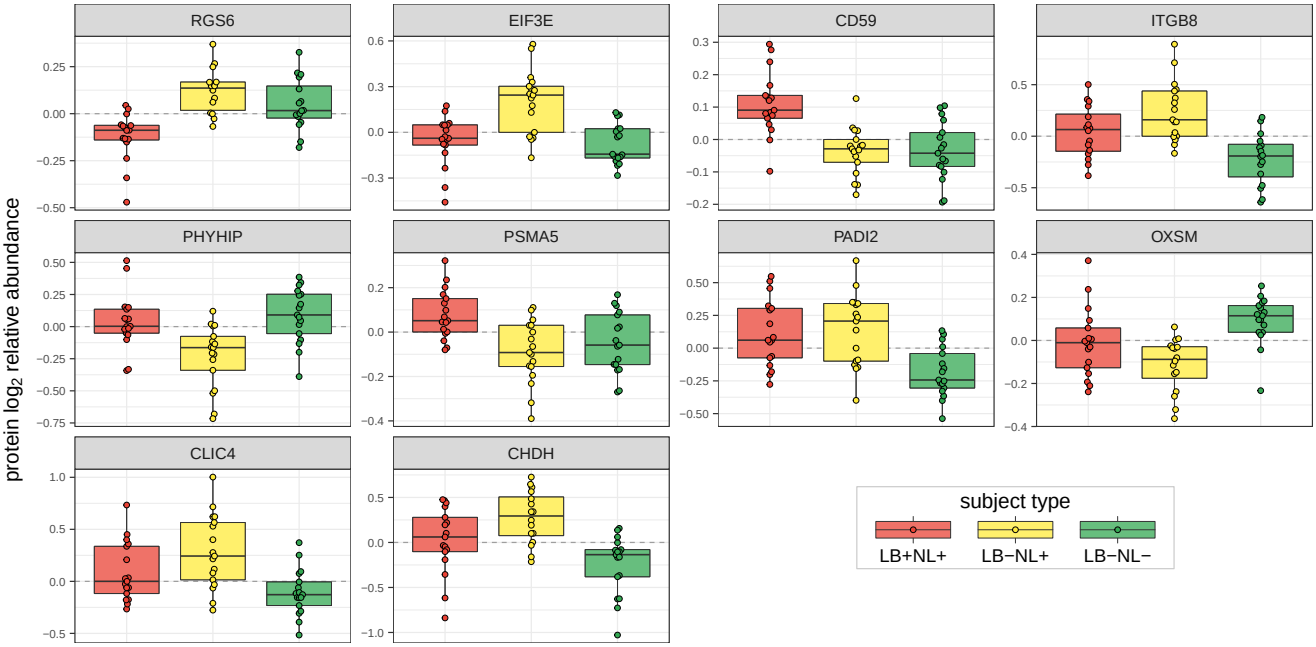
Corresponds to **Figure 3** in the main text. P-value histogram of the mixed effects ANOVA test. Histogram bin width is 0.05. Higher then random number of proteins with low p-values (first bin) indicate on significance of the subject type effect.

Full results of the ANOVA test available in the **Supplementary Table 3**.

protein	p-value	q-value	adj. p-value	LB+NL+	LB-NL+	LB-NL-
RGS6	6.55e-05	0.11	0.19	-0.14	0.11	0.03
EIF3E	1.59e-04	0.13	0.24	-0.08	0.18	-0.10
CD59	2.54e-04	0.13	0.24	0.10	-0.05	-0.05
ITGB8	3.23e-04	0.13	0.24	0.02	0.22	-0.24
PHYHIP	4.64e-04	0.14	0.25	0.07	-0.20	0.12
PSMA5	7.91e-04	0.14	0.25	0.10	-0.07	-0.03
PADI2	7.93e-04	0.14	0.25	0.09	0.12	-0.21
OXSM	8.42e-04	0.14	0.25	-0.00	-0.10	0.10
CLIC4	8.94e-04	0.14	0.25	-0.00	0.20	-0.19
CHDH	9.03e-04	0.14	0.25	0.01	0.26	-0.26

Top 10 statistically significantly varying proteins across the subject types

3.2 Top 10 most significant proteins proteins. Figure 4



Corresponds to **Figure 4** in the main text. Relative abundances of the top 10 most varying proteins across the groups. Proteins ordered according to significance in the ANOVA test. The relative abundances are adjusted for matching group effect and batch effect.

3.3 Top differential proteins for six contrasts. Table 2

To obtain a detailed view on expression profiles acress the three subject types we performed contrast analysis. The differential abundance was tested across six contrasts including three pairwise comparisons and three comparisons of individual subject type vs the other two. Adjustment for multiplicity of comparisons was performed using single-step procedure based on z-statistic implemented in the `multcomp` package.

contrast	protein	estimate	pvalue
LB+NL+ vs LB-NL+	RGS6	-0.25	4.96e-06
	GANAB	-0.15	4.16e-04
	CD59	0.15	4.39e-04
LB+NL+ vs LB-NL-	CD59	0.15	5.81e-04
	SRPK2	-0.27	8.32e-04
	COL4A1	0.56	8.77e-04
LB-NL+ vs LB-NL-	ITGB8	0.47	2.52e-05
	EIF3E	0.28	1.11e-04
	CHDH	0.52	1.56e-04
LB+NL+ vs (LB-NL+ and LB-NL-)	RGS6	-0.21	4.37e-06
	CD59	0.15	2.70e-05
	PSMA5	0.15	6.14e-04
LB-NL+ vs (LB+NL+ and LB-NL-)	EIF3E	0.27	1.29e-05
	PHYHIP	-0.30	9.25e-05
	SNRNP200	0.36	4.37e-04
LB-NL- vs (LB+NL+ and LB-NL+)	PADI2	-0.31	1.81e-04
	ITGB8	-0.37	2.02e-04
	PDCD5	0.19	7.94e-04

Top 3 most significant proteins for each contast. Corresponds to **Table 2** in the main text.

3.4 Full results of contrasts analysis. Supplementary Table 4

The entire test results for all proteins and all contrasts corresponding to **Supplementary Table 4** are saved as a text file with the corresponding name.

4 GO Term Enrichment

4.1 Top terms. Table 3

Top GO terms for each of the contrasts and directions of change.

Contrast	Direction	Ontology	ID	Description	GeneRatio	BgRatio	p.adjust
LB+NL+ vs LB-NL+	up	BP	GO:0034314	Arp2/3 complex-mediated actin nucleation	8/96	19/2798	4.2e-05
		CC	GO:0015629	actin cytoskeleton	16/97	166/2872	8.2e-03
		MF	GO:0008092	cytoskeletal protein binding	26/94	313/2745	4.9e-04
	down	BP	GO:0006397	mRNA processing	22/82	67/2798	2.4e-16
		CC	GO:0071013	catalytic step 2 spliceosome	12/84	25/2872	3.8e-11
		MF	GO:0044822	poly(A) RNA binding	44/80	370/2745	1.7e-17
LB+NL+ vs LB-NL-	up	BP	GO:0030198	extracellular matrix organization	11/51	82/2798	1.4e-05
		CC	GO:0005604	basement membrane	8/51	25/2872	1.3e-07
		MF	GO:0016684	oxidoreductase activity, acting on peroxide as acceptor	5/50	20/2745	4.5e-04
	down	BP	GO:1903829	positive regulation of cellular protein localization	5/35	92/2798	2.8e-01
		CC	GO:0044432	endoplasmic reticulum part	7/37	250/2872	3.3e-01
		MF	GO:0044877	macromolecular complex binding	7/36	293/2745	4.3e-01
LB-NL+ vs LB-NL-	up	BP	GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	13/102	53/2798	4.4e-06
		CC	GO:0071013	catalytic step 2 spliceosome	8/103	25/2872	9.7e-05
		MF	GO:0003723	RNA binding	33/103	443/2745	2.1e-03
	down	BP	GO:0023061	signal release	20/125	133/2798	5.0e-04
		CC	GO:0045202	synapse	32/127	262/2872	4.0e-06
		MF	GO:0030695	GTPase regulator activity	11/119	82/2745	2.6e-02
LB+NL+ vs (LB-NL+ and LB-NL-)	up	BP	GO:0030198	extracellular matrix organization	10/48	82/2798	8.2e-05
		CC	GO:0044420	extracellular matrix component	8/48	34/2872	1.3e-06
		MF	GO:0005198	structural molecule activity	9/48	236/2745	3.6e-01
	down	BP	GO:0016071	mRNA metabolic process	12/52	148/2798	1.4e-03
		CC	GO:0005654	nucleoplasm	17/52	480/2872	5.1e-02
		MF	GO:0044822	poly(A) RNA binding	22/51	370/2745	2.2e-06
LB-NL+ vs (LB+NL+ and LB-NL-)	up	BP	GO:0006397	mRNA processing	22/102	67/2798	1.9e-14
		CC	GO:0005681	spliceosomal complex	14/105	36/2872	4.1e-10
		MF	GO:0044822	poly(A) RNA binding	42/103	370/2745	1.2e-10
	down	BP	GO:0034314	Arp2/3 complex-mediated actin nucleation	11/174	19/2798	1.4e-06
		CC	GO:0045202	synapse	46/176	262/2872	8.2e-10
		MF	GO:0017075	syntaxin-1 binding	6/168	11/2745	1.1e-03
LB-NL- vs (LB+NL+ and LB-NL+)	up	BP	GO:0044255	cellular lipid metabolic process	11/44	241/2798	8.2e-02
		CC	GO:0005783	endoplasmic reticulum	12/44	368/2872	6.1e-02
		MF	GO:0030695	GTPase regulator activity	7/41	82/2745	2.2e-03
	down	BP	GO:0030198	extracellular matrix organization	12/78	82/2798	2.4e-04
		CC	GO:0005604	basement membrane	5/79	25/2872	1.6e-02
		MF	GO:0016491	oxidoreductase activity	14/78	259/2745	1.9e-01

Most significant GO terms for each contrast and direction of change. Corresponds to **Table 3** in the main text.

4.2 Complete results. Supplementary Table 8

The signicantly enriched GO terms for up- and down-regulated proteins for each six contrasts corresponding to the **Supplementary Table 8** are saved as text files.

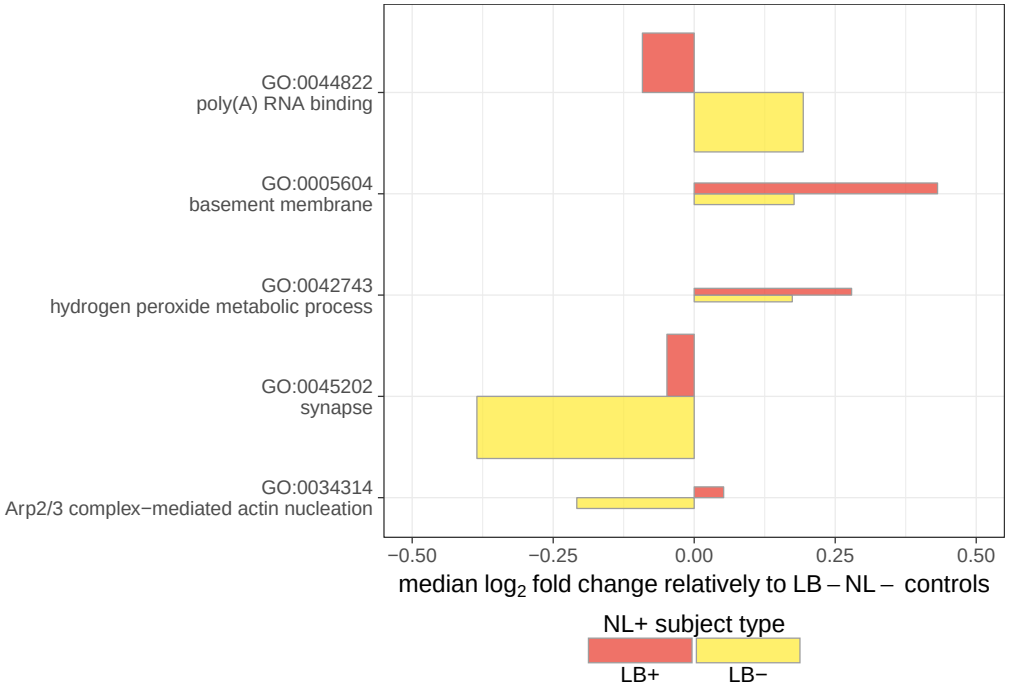
4.3 Non-redundant set of GO terms

To infer a parsimonious set of GO terms we applied an iterative procedure. For each iteration round we selected the most statistically significant over-represented GO term regardless of the contrast and directionality of change. Then proteins corresponding to this top GO term were removed from further consideration. The iterations continued until no GO term is statistically significantly enriched. Using this iterative ontology elimination procedure we selected only five GO terms.

Contrast	Direction	Ontology	ID	Description	GeneRatio	BgRatio	pvalue	p.adjusted
LB+NL+ vs LB-NL+	down	MF	GO:0044822	poly(A) RNA binding	44/80	370/2745	3.8e-19	2.0e-15
LB-NL+ vs (LB+NL+ and LB-NL-)	down	CC	GO:0045202	synapse	46/176	262/2872	5.8e-12	2.6e-08
LB+NL+ vs LB-NL-	up	CC	GO:0005604	basement membrane	8/51	25/2872	4.8e-09	1.8e-05
LB+NL+ vs LB-NL+	up	BP	GO:0034314	Arp2/3 complex-mediated actin nucleation	8/88	19/2798	4.0e-08	1.4e-04
LB+NL+ vs LB-NL-	up	BP	GO:0042743	hydrogen peroxide metabolic process	5/43	18/2798	5.0e-06	1.7e-02

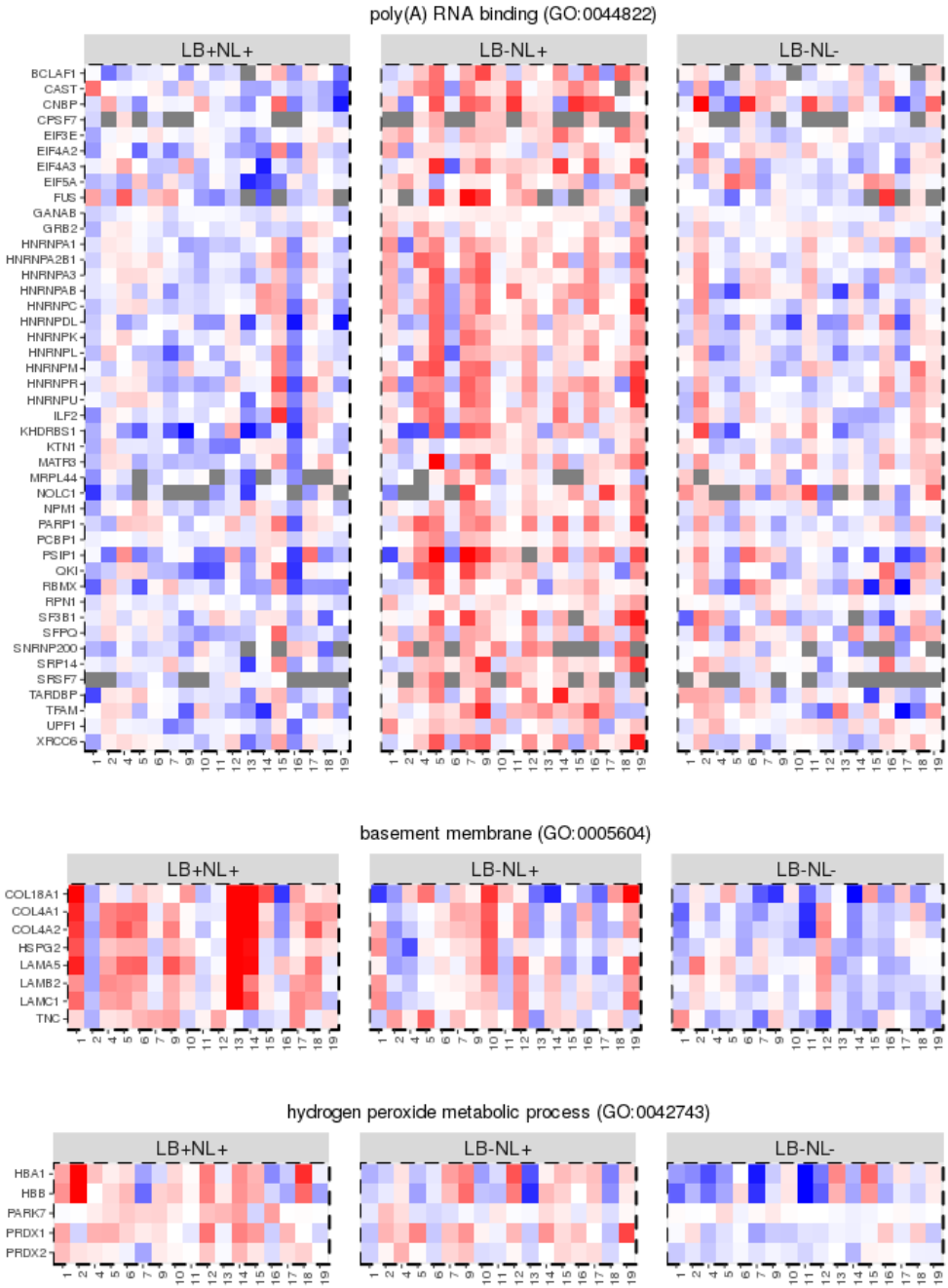
Non-redundant set of GO terms explaning the variance between the subject types.

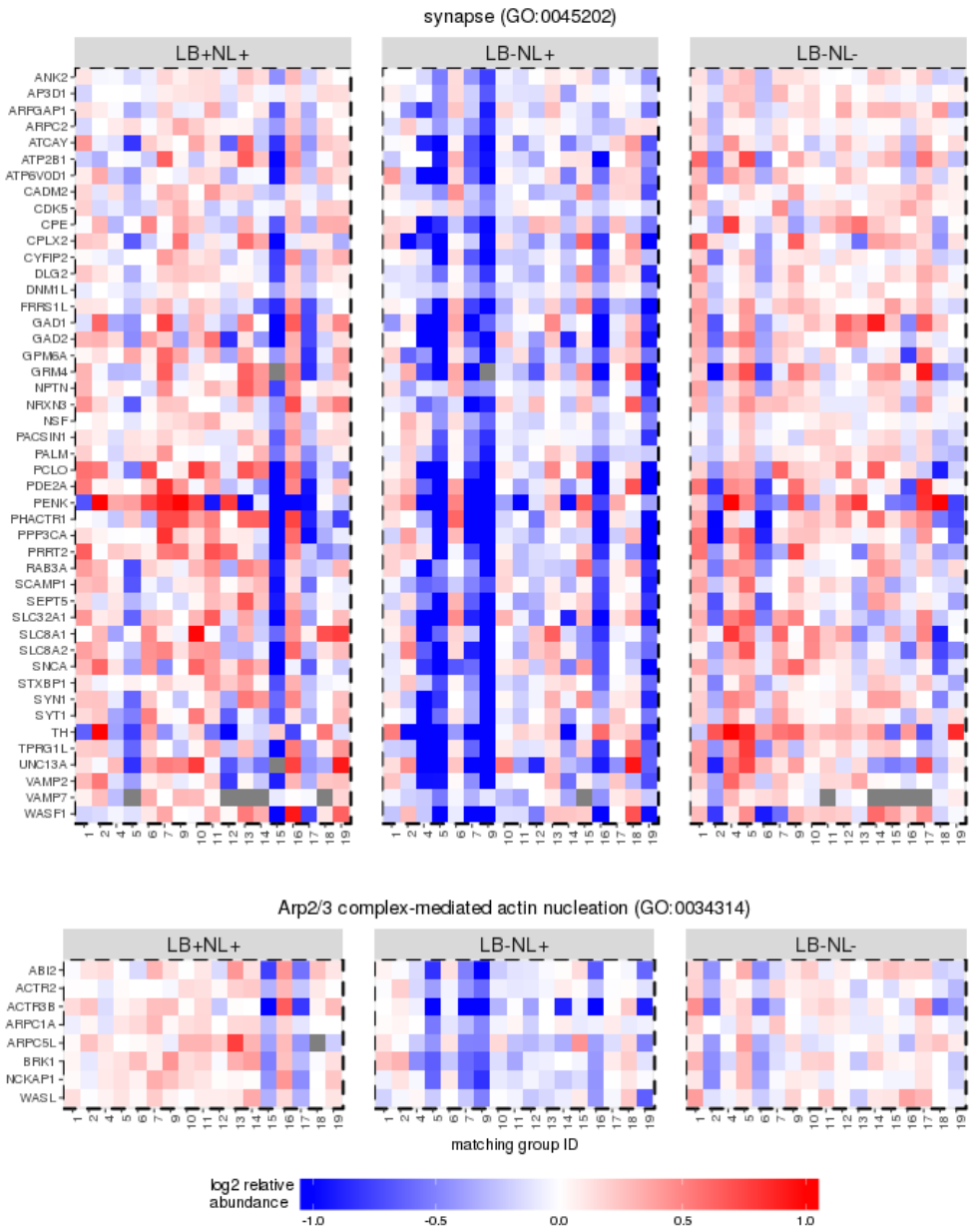
4.3.1 Barplot summary of non-redundant terms. Figure 6



Corresponds to **Figure 6** in the main text. Non-redundant GO terms. X axis correspond to a median log2 fold change of the corresponding proteins. Control group representing general population (LB-NL-) serves as a reference. The rectangle height (Y axis direction) is proportional to the number of proteins mapped to the term.

4.3.2 Heatmaps corresponding to individual terms. Figure 7





Corresponds to **Figure 7** in the main text. Relative abundances of the individual proteins mapping to the non-redundant significant GO terms.

5 Assessment of the cell type level changes

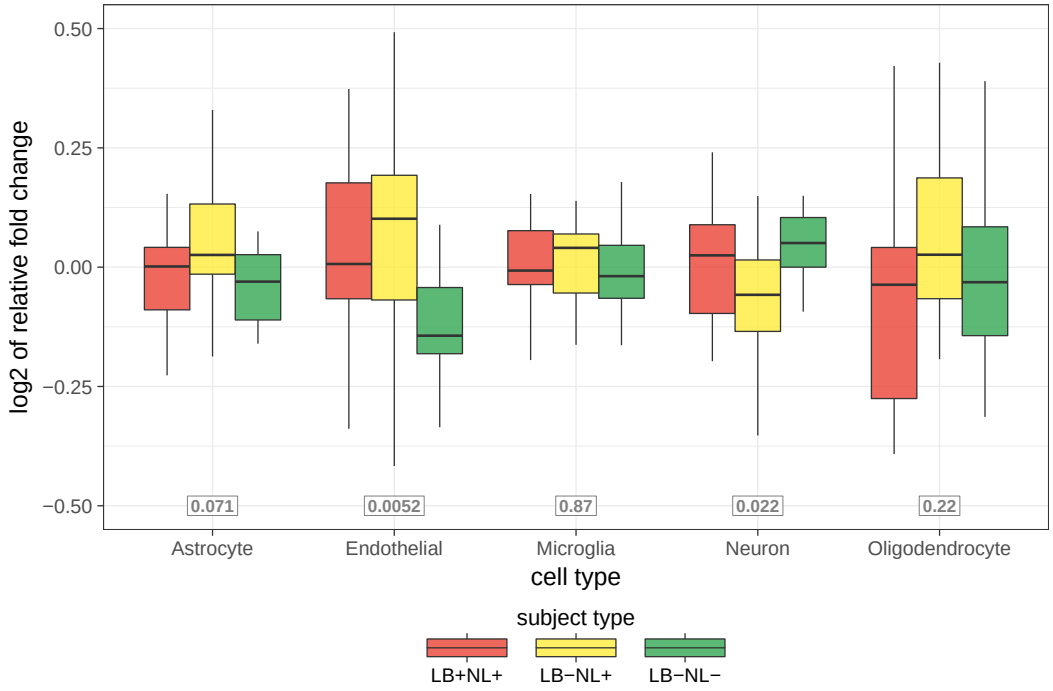
The data on cell-specific gene expression was downloaded from Barres Lab. Specifically, the excel file with processed data is available here. The corresponding text file is available in the package's "extdata" folder. The mouse genes were mapped to protein orthologues using ENSEMBL database.

5.1 Significance analysis

Statistical test (Kruskal-Wallis) to check if any cell type changes in amount across the subject types. The p-values are in the legend to the **Figure 8** in the main text. Given the data, the endothelial cells are the most strongly associated with the neuronal loss regardless of the Lewy bodies.

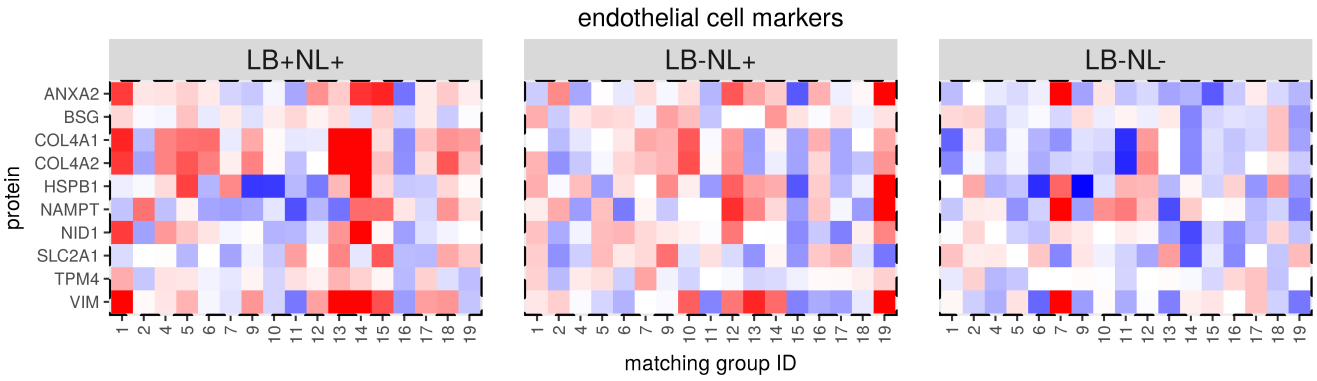
	Astrocyte	Endothelial	Microglia	Neuron	Oligodendrocyte
p-value	0.071	0.0052	0.87	0.022	0.22

5.2 Barplot of cell type contributions. Figure 8



Corresponds to **Figure 8** in the main text. Estimates of cell-type level changes based on the sets of marker proteins. Gross cell-level changes were estimated based on the median relative abundance value of the marker protein sets.

5.3 Heatmap of protein markers of endothelial cells. Figure 9



Corresponds to **Figure 9** in the main text. Relative abundances of the endothelial marker proteins. The colorkey of relative abundances is the same as on Figure 7.

6 Primary factors affecting α-synuclein abundance

6.1 Statisical model of α-synuclein abundance

Modeling α-synuclein abundance. The terms to consider are the presense of Lewy bodies (term LB), dopaminergic neuronal loss (term pigmented_NL) assessed by melanin deposits and overal neronal cell relative amount estimated using neuronal protein markers (term neuron). The random effect terms included plate column and matching group. The full model in Wilkinson-Rogers notation is:

$$SNCA \sim LB + pigmented\_NL + neuron + (1|PlateCol) + (1|match.group)$$

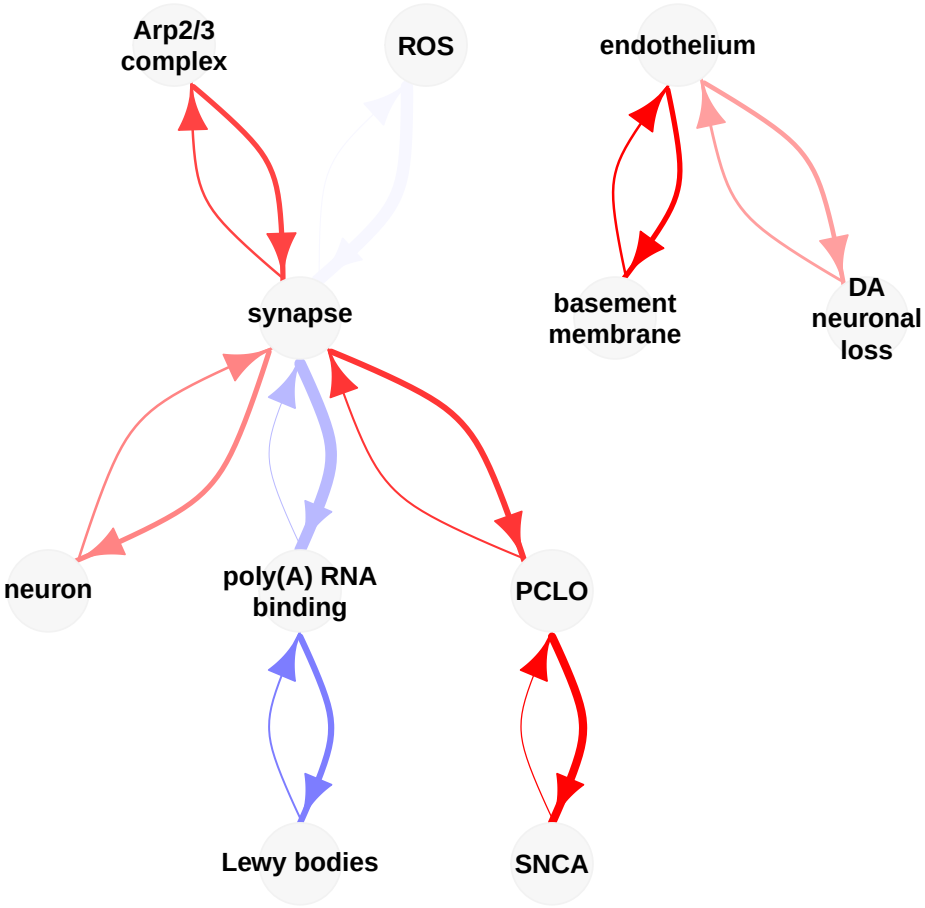
	term description	p-value	effect size (log2)
LB	Lewy body presence	0.017	0.28
pigmented_NL	DA neuronal loss	0.027	-0.18
neuron	relative amount of neuronal cells	1.3e-05	1.77

Relation of α-synuclein abundance and neuropathological parameters.

The results of the statistical modeling indicate that the LB presense explain the SNCA abundance. However, given the data, the most strongest term correlating with SNCA is the amount of neuronal cells.

7 Causal Network Analysis

7.1 Causal Network. Figure 10



Corresponds to **Figure 10** in the main text. The representation of the causal model and associated uncertainties inferred by bootstrap analysis. Certainty in edge presence is denoted by opaqueness (opaque - more likely). Edge color represents directionality of the effect (red - positive, blue - negative). Confidence in directionality is depicted by the edge width. Given the data, the structural modeling does not suggest α-synuclein as the major driver of LB formation.

8 Session information

All software and respective versions used in this document, as returned by sessionInfo() are detailed below.

- R version 3.3.2 (2016-10-31), x86_64-pc-linux-gnu
- Locale: LC_CTYPE=en_US.UTF-8, LC_NUMERIC=C, LC_TIME=en_US.UTF-8, LC_COLLATE=en_US.UTF-8, LC_MONETARY=en_US.UTF-8, LC_MESSAGES=C, LC_PAPER=en_US.UTF-8, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=en_US.UTF-8, LC_IDENTIFICATION=C
- Base packages: base, datasets, graphics, grDevices, grid, methods, parallel, stats, stats4, utils
- Other packages: AnnotationDbi 1.36.2, Biobase 2.34.0, BiocGenerics 0.20.0, BiocParallel 1.8.2, BiocStyle 2.2.1, bnlearn 4.0, Category 2.40.0, clusterProfiler 3.2.14, doParallel 1.0.10, DOSE 3.0.10, dplyr 0.5.0, dynamicTreeCut 1.63-1, fastcluster 1.1.20, foreach 1.4.3, gelnet 1.2.1, ggbeeswarm 0.5.0, ggplot2 2.2.1, GOstats 2.40.0, graph 1.52.0, igraph 1.0.1, IRanges 2.8.2, iterators 1.0.8, knitr 1.14, lme4 1.1-12, lubridate 1.6.0, MASS 7.3-45, Matrix 1.2-7.1, MSnbase 2.0.2, multcomp 1.4-5, mvtnorm 1.0-6, mzR 2.8.1, org.Hs.eg.db 3.4.0, pcaMethods 1.66.0, ProtGenerics 1.6.0, qvalue 2.6.0, Rcpp 0.12.9, ReactomePA 1.18.1, reshape2 1.4.1, S4Vectors 0.12.2, survival 2.40-1, TH.data 1.0-8, tidyr 0.6.1, vp.misc 0.1, WGCNA 1.51, xtable 1.8-2
- Loaded via a namespace (and not attached): acepack 1.4.1, ade4 1.7-4, affy 1.52.0, affyio 1.44.0, annotate 1.52.1, AnnotationForge 1.16.1, assertthat 0.1, backports 1.0.5, base64enc 0.1-3, beeswarm 0.2.3, BiocInstaller 1.24.0, bitops 1.0-6, Boruta 5.0.0, caTools 1.17.1, checkmate 1.8.2, cluster 2.0.5, codetools 0.2-15, colorspace 1.3-2, data.table 1.10.4, DBI 0.5-1, digest 0.6.9, DO.db 2.9, evaluate 0.10, fastmatch 1.0-4, FField 0.1.0, fgsea 1.0.2, foreign 0.8-67, formatR 1.4, Formula 1.2-1, gdata 2.17.0, genefilter 1.56.0, glmnet 2.0-5, GO.db 3.4.0, GOSemSim 2.0.4, gplots 3.0.1, graphite 1.20.1, gridExtra 2.2.1, GSEABase 1.36.0, gtable 0.2.0, gtools 3.5.0, Hmisc 4.0-2, htmlTable 1.9, htmltools 0.3.5, htmlwidgets 0.8, impute 1.48.0, KernSmooth 2.23-15, labeling 0.3, lattice 0.20-34, latticeExtra 0.6-28, lazyeval 0.2.0, limma 3.30.13, magrittr 1.5, MALDIquant 1.14, matrixStats 0.50.2, memoise 1.0.0, minqa 1.2.4, munsell 0.4.3, mzID 1.12.0, nlme 3.1-128, nloptr 1.0.4, nnet 7.3-12, outliers 0.14, plyr 1.8.4, preprocessCore 1.36.0, R6 2.2.0, randomForest 4.6-12, ranger 0.4.0, rappdirs 0.3.1, RBGL 1.50.0, RColorBrewer 1.1-2, RCurl 1.95-4.8, reactome.db 1.58.0, Rgraphviz 2.18.0, ROCR 1.0-7, rpart 4.1-10, RSQLite 1.1-2, sandwich 2.3-4, scales 0.4.1, splines 3.3.2, stringi 1.1.2, stringr 1.2.0, tibble 1.2, tools 3.3.2, varSelRF 0.7-5, vipor 0.3.2, vsn 3.42.3, XML 3.98-1.5, zlibbioc 1.20.0, zoo 1.7-14