Executable Analysis Document Supporting:

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

Part I: Study Design

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

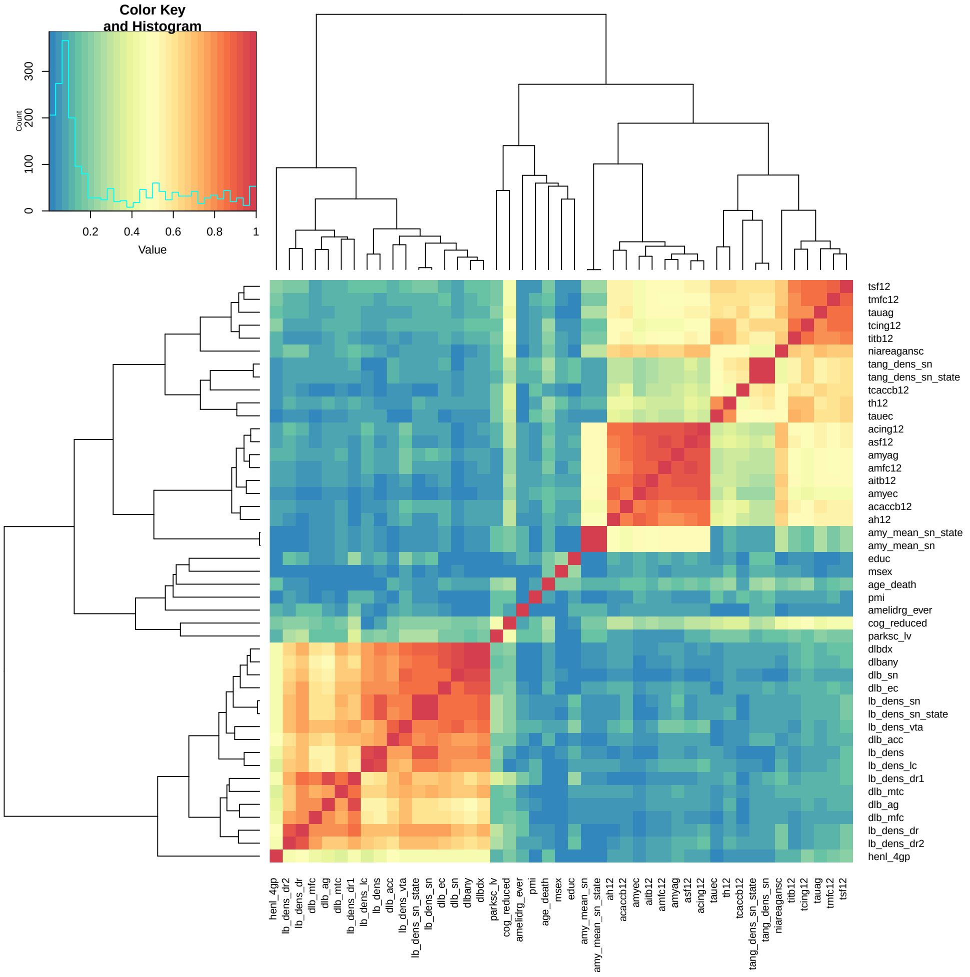
Contents

1 Objective

The purpose of this document is to describe the study design. This includes definition of cases and controls and selection of subjects out of 573 with available frozen substatia nigra.

2 Correlation structure and identificaton of covariates

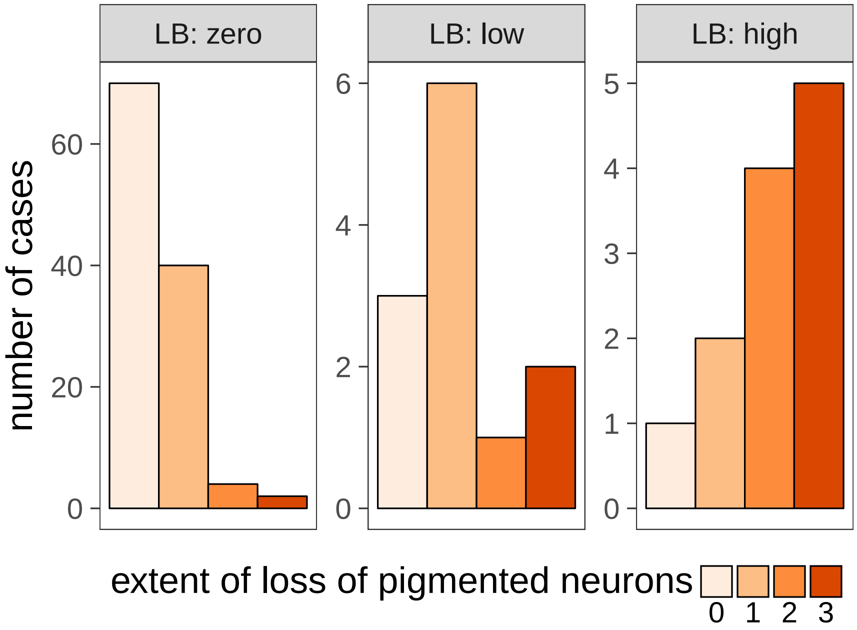
One of the problems with study designs is the presense of confounding factors that correlate with the variable of interest. Their presense usually make the interpretation of the results ambigious as it can be hard to distinguish the real correlate with a protein abundance. The wealth of available clinical information allowed us to discover such (typically hidden) covariates.



Heatmap of the correlation structure based on 45 variables.

The clinical variables split into four major clusters. Three of them have high internal correlation correlation and reflect the presense and density of tangles, amyloid plaques and Lewy bodies in the different parts of the brain. Noteworthy, the amyloid and tangles cluster show mutual correlation. The forth cluster contains the rest low correlating variables (e.g. education, sex, post-mortem interval, cognitive skills, etc).

The only strong covariate with the Lewy body presense is the loss of pigmented (dopanimergic) neurons in the substantia nigra (`henl_4gp`). This is not an unexpected finding as both Lewy bodies and loss of dopaminergic neurons in the substantia nigra (by the way, the region was named due to its intense pigmentation) are the hallmarks of Parkinson's disease.



Corresponds to **Figure 1** . Association of neronal loss (henl_4gp) with density of Lewy bodies (lb_dens_sn_state).

The association of neuronal loss and Lewy body presense is highy significant as evident from the the results of Kruskal-Wallis (non-parametric ANOVA) test.

```
kruskal.test(henl_4gp ~ lb_dens_sn_state, data=clinical_metadata_full)


Kruskal-Wallis rank sum test

data:  henl_4gp by lb_dens_sn_state
Kruskal-Wallis chi-squared = 29.334, df = 2, p-value = 4.267e-07
```

```
cor.test(~ henl_4gp + lb_dens_sn,
         data=clinical_metadata_full,
         method='spearman', exact=F, continuity=T)


Spearman's rank correlation rho
```

```
data:  henl_4gp and lb_dens_sn
S = 253800, p-value = 3.611e-08
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.4450163
```

Considerations for study design.

- If we do not match by neuronal loss, then it won't be clear what the protein abundance difference is due to: Lewy bodies or neuronal loss.
- However, neuronal loss is not normally presense in the Lewy body-free population. There might be another reason for neuronal loss (an example is PARK2 mutation). Thus matching cases and controls by neuronal loss may confound Lewy body-associated and some other type of cause of neuronal loss.
- Therefore to deconvolve the proteome changes associated with Lewy bodies from changes associated with neuronal loss or cause of Lewy body-independent neuronal loss we decided to use two types of controls. First type of controls is matched by neuronal loss. Second type, represents normal population and not matched by neuronal loss.

3 Imputing Lewy body, amyloid, tangles density

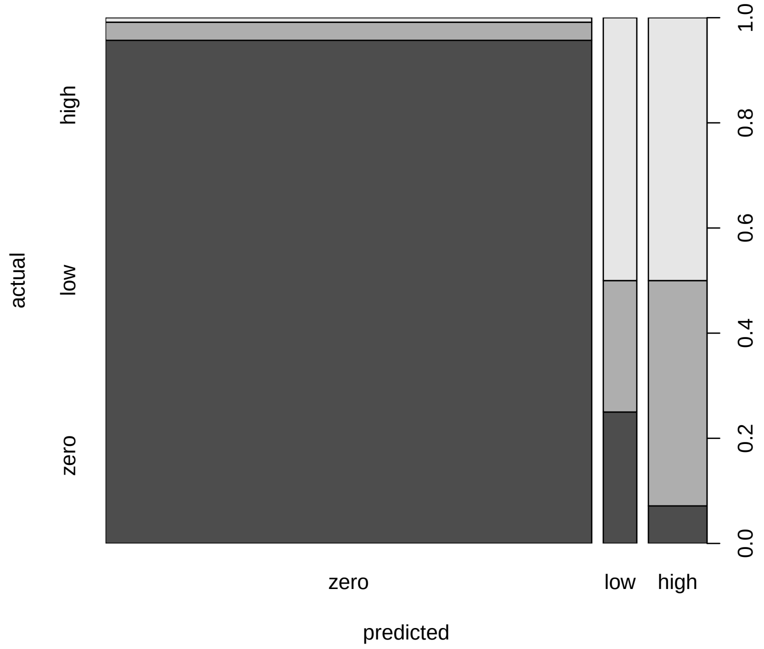
This section describes imputing Lewy bodies, amyloid plaques and neurofibrillary tangle density values in the *substantia nigra* region. The continuous density values were discretized into 3-5 states as high precision values were not necessary and qualitative "low"/"high" descriptors were sufficient for the study design.
Lewy bodies density is a crucial parameter for contrasting cases vs controls. Density of amyloid and tangles is a matching parameters with intent of reducing nuisance variability between paired cases and controls. Therefore accuracy in prediction of amyloid and tangle density is less crucial then of Lewy bodies. In the worst case scenario if the amyloid and tangle density can not be accurately predicted it is equivalent to leaving out those variables of consideration for case/control matching.

3.1 Imputing Lewy Body density in substantia nigra

The ideal variable to contrast cases vs controls is Lewy body density in *substantia nigra*. However the value for this variable is availabe only for 140 subjects. Given the strong correlation with other variables reflecting Lewy body presense and density in the various brain regions it should be possible to impute the density in the *substantia nigra* region.
We consider predictors that have non-missing values in $>90\%$ of the subjects.

We performed LOOCV with independent feature selection for each round to evaluate the power of the available data to predict Lewy body density state (zero, low, high) in the *substantia nigra* region. There are 138 subjects that have Lewy body density state values and no missing values for the selected predictors.

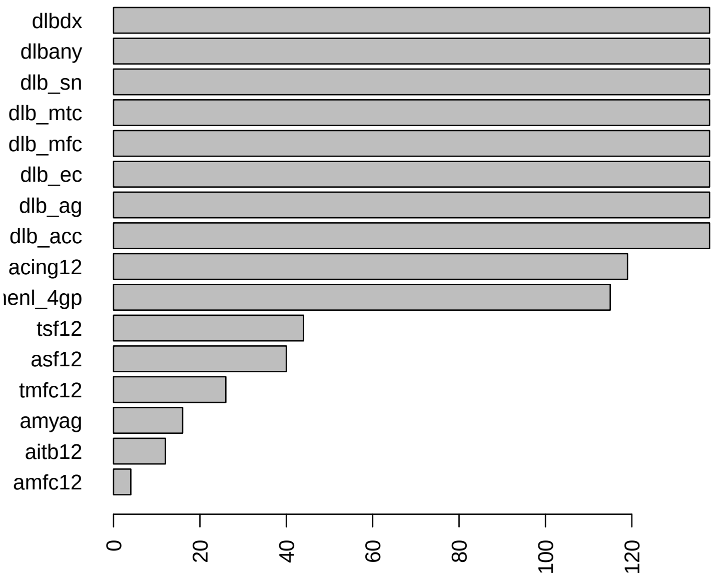
```
Confusion Matrix and Statistics

          Reference
```



Corresponds to **Supplementary Figure 1** . Confusion matrix showing correspondence of known vs predicted states of LB density.

```
Prediction zero low high
      zero  111   4    1
      low     2   2    4
      high    1   6    7

Overall Statistics

               Accuracy : 0.8696
                 95% CI : (0.8017, 0.9208)
    No Information Rate : 0.8261
    P-Value [Acc > NIR] : 0.1055

                  Kappa : 0.5529
 Mcnemar's Test P-Value : 0.7851

Statistics by Class:

                     Class: zero Class: low Class: high
Sensitivity               0.9737    0.16667     0.58333
Specificity               0.7917    0.95238     0.94444
Pos Pred Value            0.9569    0.25000     0.50000
Neg Pred Value            0.8636    0.92308     0.95968
```

```
Prevalence                0.8261    0.08696     0.08696
Detection Rate            0.8043    0.01449     0.05072
Detection Prevalence      0.8406    0.05797     0.10145
Balanced Accuracy         0.8827    0.55952     0.76389
```



The frequency of features selection for the best model summarized over all rounds of LOOCV.

For building the model for the imputation of missing data we selected only those predictors that were consistenly selected across all 138 rounds of LOOCV. Specfically, this includes presense/absense of LB in (alphabetic order) anterior cingulate cortex (dlb_acc), angular gyrus (dlb_ag), entorhinal cortex (dlb_ec), midfrontal cortex (dlb_mfc), middle/superior temporal cortex (dlb_mtc) and substantia nigra (dlb_sn). Also it includes presense/absense of LB anywhere in the brain (dlbany) and type (nigral, limbic or neocortex) of LB disease (dlbdx).

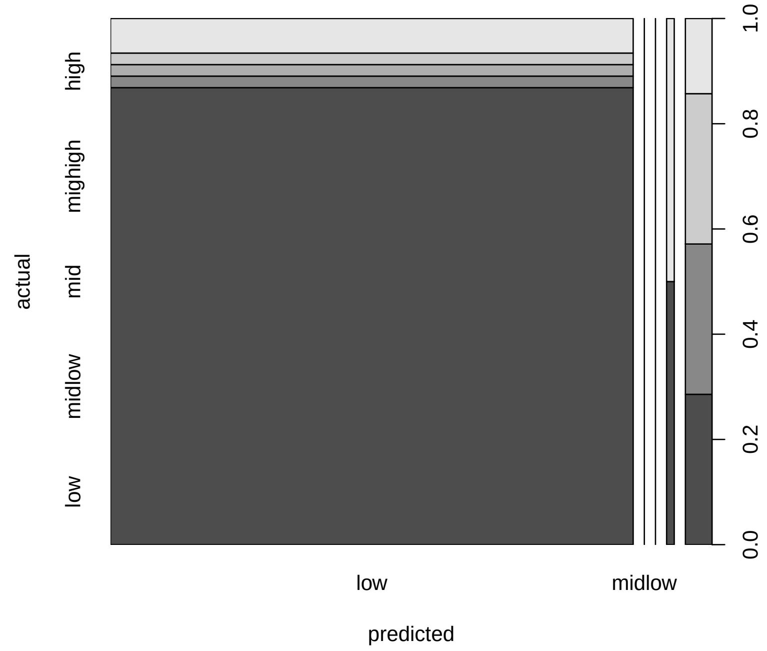
```
[1] "dlb_acc" "dlb_ag"  "dlb_ec"  "dlb_mfc" "dlb_mtc" "dlb_sn"  "dlbany"  "dlbdx"
```

The 140 subjects had actual *substantia nigra* Lewy body density state values. For 571 subjects it was possible to predict making the total number with a value 711 out of 573.

3.2 Imputing amyloid density in substantia nigra

We considered density of amyoid plaques in *substantia nigra* is one of the criteria for case/control pairing. The actual value is available for 150 subjects. To impute the value for the rest of the samples, we appied the same strategy as with Lewy bodies mentioned in the section above.
Considered predictors for imputing `amy_mean_sn_state` that have values available for > 90% of subjects.

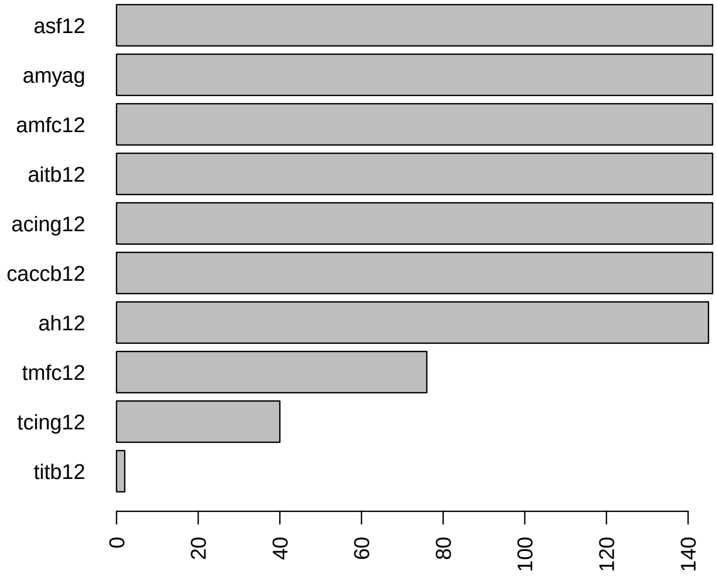
The results of cross-validaton, confusion matrix and performance metrix summarized below.



Confusion matrix showing correspondence of known vs predicted states of amyloid density.

```
Confusion Matrix and Statistics

          Reference
Prediction low midlow mid mighigh high
   low     119      3   3       3    9
   midlow    0      0   0       0    0
   mid       0      0   0       0    0
   mighigh   1      0   0       0    1
   high      2      2   0       2    1

Overall Statistics

               Accuracy : 0.8219
                 95% CI : (0.7501, 0.8802)
    No Information Rate : 0.8356
    P-Value [Acc > NIR] : 0.7174

                  Kappa : 0.1592
 Mcnemar's Test P-Value : NA

Statistics by Class:
```

```
                     Class: low Class: midlow Class: mid Class: mighigh Class: high
Sensitivity              0.9754       0.00000    0.00000        0.00000    0.090909
Specificity              0.2500       1.00000    1.00000        0.98582    0.955556
Pos Pred Value           0.8686           NaN        NaN        0.00000    0.142857
Neg Pred Value           0.6667       0.96575    0.97945        0.96528    0.928058
Prevalence               0.8356       0.03425    0.02055        0.03425    0.075342
Detection Rate           0.8151       0.00000    0.00000        0.00000    0.006849
Detection Prevalence     0.9384       0.00000    0.00000        0.01370    0.047945
Balanced Accuracy        0.6127       0.50000    0.50000        0.49291    0.523232
```



The frequency of features selection for the best model summarized over all rounds of LOOCV.

For building the model for the imputation of missing data we selected only those predictors that were consistenly selected across all 146 rounds of LOOCV. Specifically, this includes average amyloid load in calcarine cortex region (acaccb12), cingulated region (acing12), inferior temporal gyrus region (aitb12), midfrontal gyrus (amfc12), angular gyrus (amyag) and superior frontal gyrus (asf12).

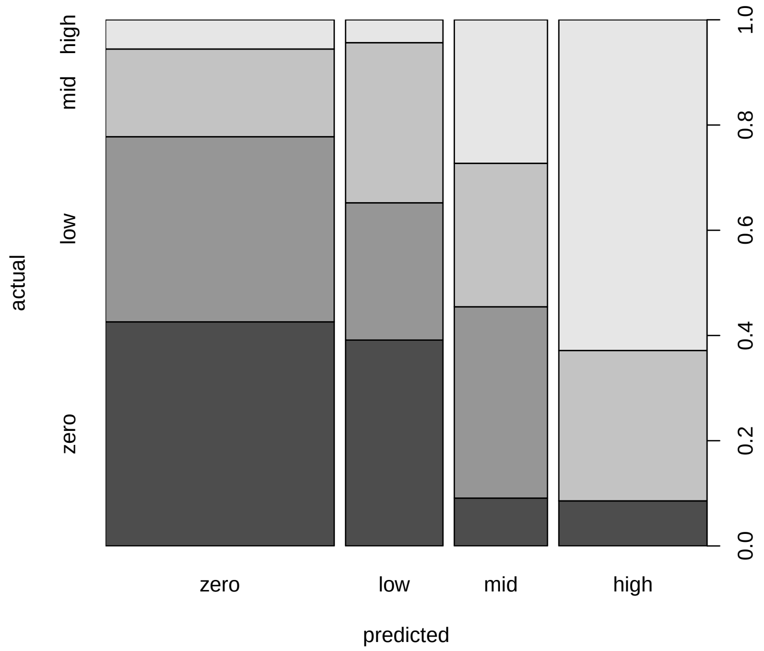
```
[1] "acaccb12" "acing12"  "aitb12"   "amfc12"   "amyag"    "asf12"
```

The 150 subjects had actual *substantia nigra* amyloid mean density state values. For 491 subjects it was possible to predict making the total number with a value 641 out of 573. Note that the estimate of prediction accuracy appears quite high because most of the subjects fall into the "zero" density group.

3.3 Imputing tangle density in substantia nigra

We considered density of tangles in *substantia nigra* is one of the criteria for case/control pairing. The actual value is available for 136 subjects. To impute the value for the rest of the samples, we appied the same strategy as with Lewy bodies mentioned in the section above.
Considered predictors for imputing `tang_dens_sn_state` that have values available for > 90% of subjects.

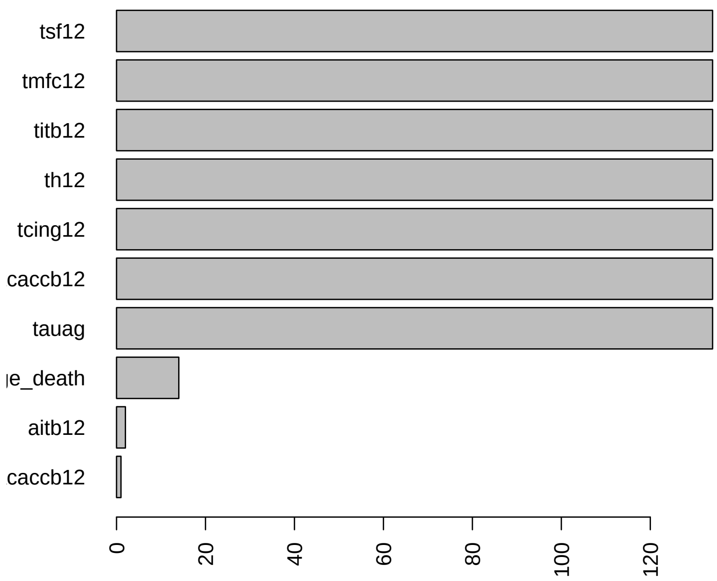
The results of cross-validaton, confusion matrix and performance metrix summarized below.



Confusion matrix showing correspondence of known vs predicted states of tangle density.

```
Confusion Matrix and Statistics

          Reference
Prediction zero low mid high
      zero   23  19   9    3
      low     9   6   7    1
      mid     2   8   6    6
      high    3   0  10   22

Overall Statistics

               Accuracy : 0.4254
                 95% CI : (0.3404, 0.5137)
    No Information Rate : 0.2761
    P-Value [Acc > NIR] : 0.0001451
```

```
                 Kappa : 0.2286
 Mcnemar's Test P-Value : 0.1208045

Statistics by Class:

                     Class: zero Class: low Class: mid Class: high
Sensitivity               0.6216    0.18182    0.18750      0.6875
Specificity               0.6804    0.83168    0.84314      0.8725
Pos Pred Value            0.4259    0.26087    0.27273      0.6286
Neg Pred Value            0.8250    0.75676    0.76786      0.8990
Prevalence                0.2761    0.24627    0.23881      0.2388
Detection Rate            0.1716    0.04478    0.04478      0.1642
Detection Prevalence      0.4030    0.17164    0.16418      0.2612
Balanced Accuracy         0.6510    0.50675    0.51532      0.7800
```



The frequency of features selection for the best model summarized over all rounds of LOOCV.

For building the model for the imputation of missing data we selected only those predictors that were consistenly selected across all 134 rounds of LOOCV. Specifically, this includes tangle density in inferior temporal gyrus region (titb12), superior frontal gyrus (tsf12) cingulated region (tcing12), hippocampus (th12) and calcarine cortex region (tcaccb12).

```
[1] "titb12"   "tsf12"    "tcing12"  "th12"     "tcaccb12"
```

The 136 subjects had actual *substantia nigra* amyloid mean density state values. For 467 subjects it was possible to predict making the total number with a value 603 out of 573.

4 Inclusion Criteria

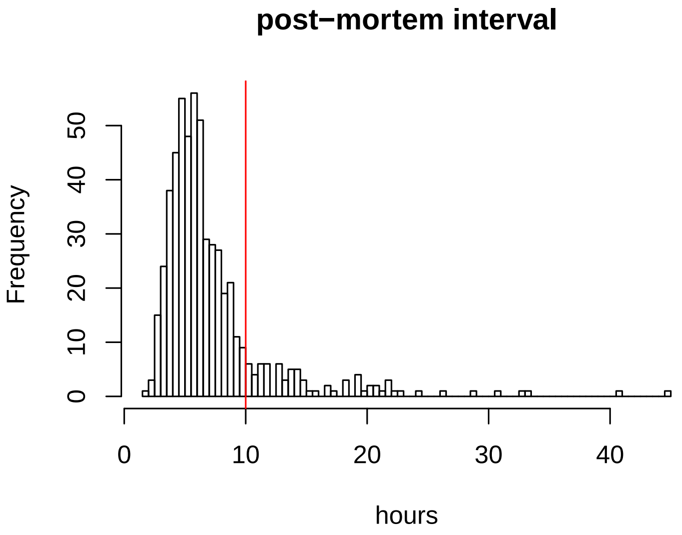
The rational for excluding some objects is reduction of cohort variance. This allows reduction matching and treating the cohort as relatively homogeneous. The ROS and MAP cohort are overwhelmingly represented by white/caucasians. Thus the other races were excluded.

```
table(clinical_metadata_full$race)


                  White                   Black Native American, Indian
                    564                       8                       1
               Asian/PI
                      0
```

The white/caucasians are overwhelmingly non-hispanic. Therefore we exclude subjects identified themselves as of hispanic origin.

```
table(clinical_metadata_full$spanish)


    Hispanic Not Hispanic
           9          555
```

The distribution of post-mortem interval (PMI) follows log-normal distribution. The median value is 6 hours.

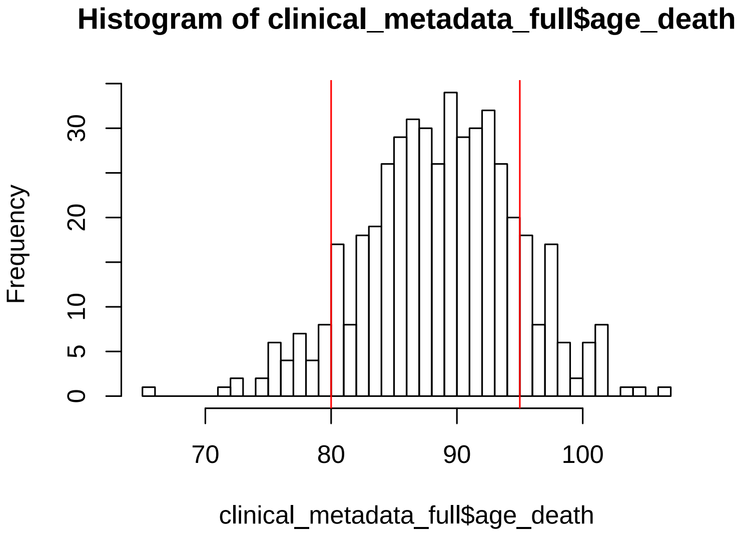


The distribution of post-mortem interval. Subjects with PMI >= 10 hours (red line) were excluded.

We excluded subjects with more or equal to 10 hours of PMI.

```
table(clinical_metadata_full$pmi < 10)


FALSE  TRUE
   77   478
```

Finally, we exluded subjects with extreme values of age at death. The age at death ranges from 65.9931554 to 106.4996578 with median value of 88.991102 years.



The distribution of age at death. Only subjects with > 80 and < 95 age at death were retained.

We retained subjects from 80 to 95 years at death, constituting 78% of the cohort.

Since most subjects have "low" density of amyloid this was considered as inclusion criterion.

A few subjects were excluded due to unavailable substantia nigra.

Overall applying four inclusion criteria we retained 46.9% or 269 out of 573.

5 Definitions for cases and controls

Cases were defined as subjects with Lewy bodies spread out through the entire brain (`dbldx == 3`) and having "high" state of density of Lewy bodies in *substantia nigra* and. Controls are the subjects completely lacking Lewy bodies anywhere in the brain (`dlbany == 0`) and having "zero" density state of Lewy body density in the *substantia nigra*.

Number of subjects satisfying case and control criteria are 29 and 211, respectively.

6 Paired Matching

As we mentioned before, the presense of Lewy bodies correlates with neuronal loss. Therefore to avoid confounding those two tissue-level differences with proteomics data we devised two types of controls. First type of control include matching by neuronal loss. Second type of controls disregard neuronal loss and reflect more general population.

6.1 Selecting type 1 controls

The variables that had to exactly match between the cases and type 1 controls include sex, neuronal loss, density of amyloid and tangles in *substantia nigra*.

```
[1] "msex"                       "henl_4gp"                   "amy_mean_sn_state_imputed"
[4] "tang_dens_sn_state_imputed"
```

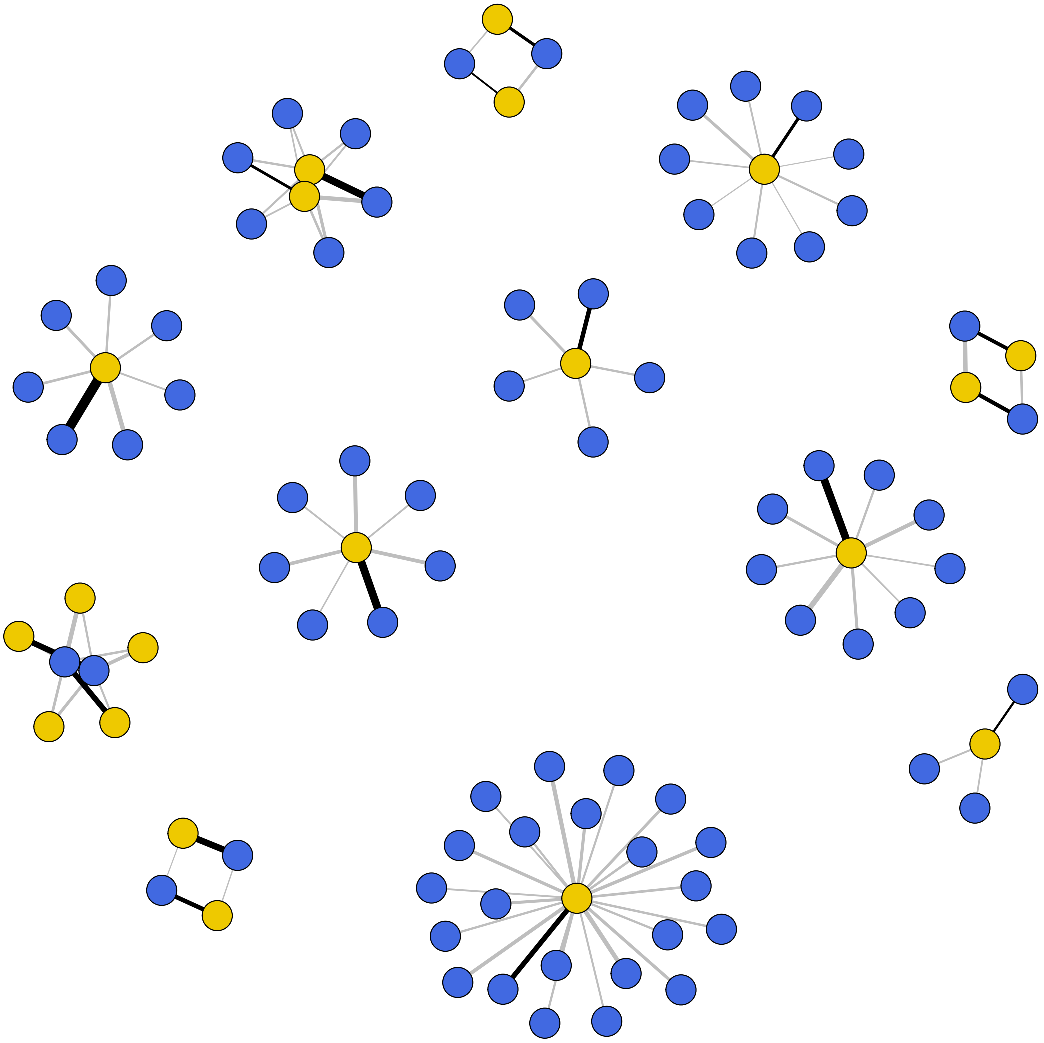
Using this approach, 17 cases out of 29 were matched to type 1 controls.

6.2 Selecting type 2 controls

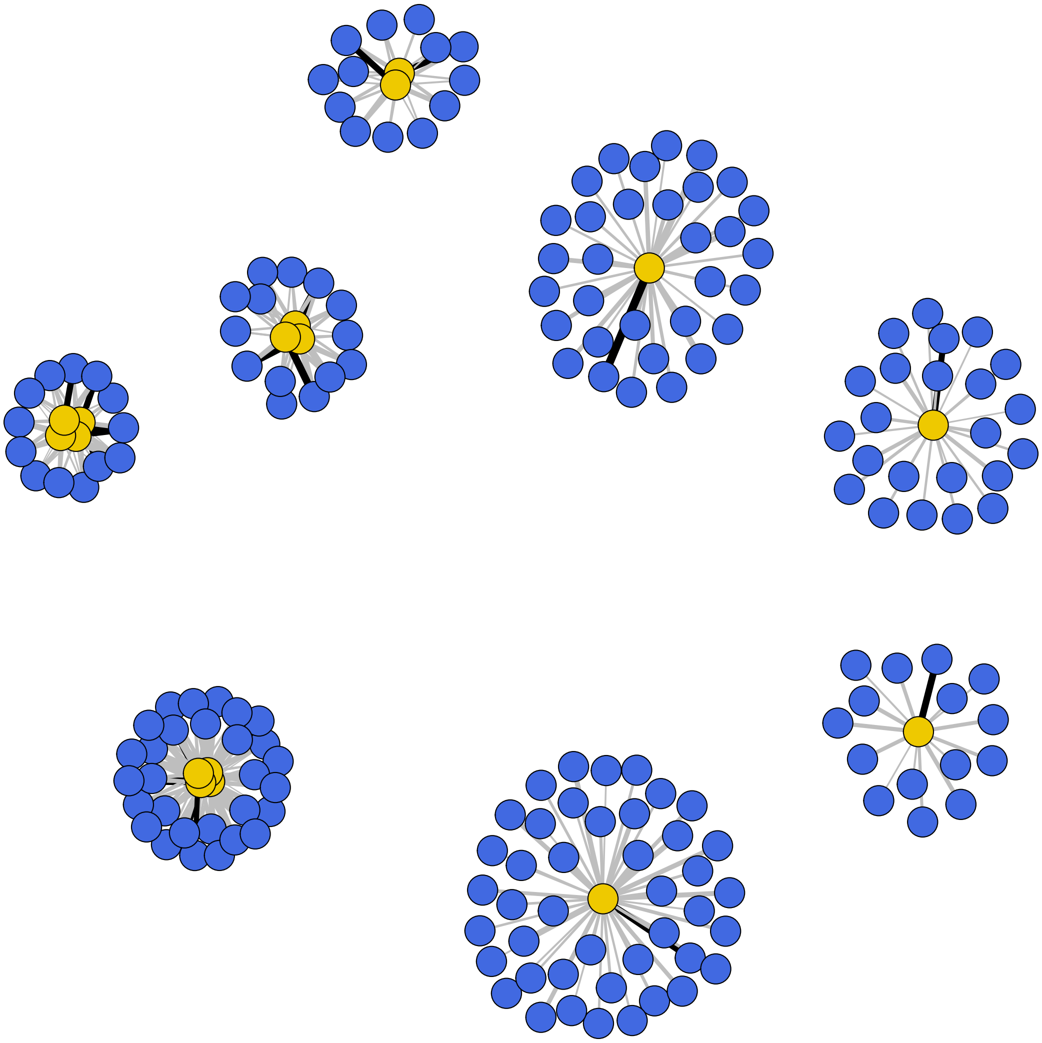
Variables that had to exactly match between the cases and type 2 controls are the same as above, except the neuronal loss. Note, that for the second of matches we considered only the cases that found a match in the first round. Conversly, the controls that were denoted as type 1 were excluded from the consideration in the second round.

```
[1] "msex"                       "amy_mean_sn_state_imputed"  "tang_dens_sn_state_imputed"
```

Using this approach, 17 cases out of 29 were matched to type 2 controls.



Corresponds to **Supplementary Figure 2** . Matching round 1. Cases and controls represented as a bipartite graph with yellow and blue nodes, respectively. Thicker edges correspond to better case/control matching. Edges corresponding to optimal case/control matching highlighted with black.



Matching round 2. Cases and controls represented as a bipartite graph with yellow and blue nodes, respectively. Thicker edges correspond to better case/control matching. Edges corresponding to optimal case/control matching highlighted with black.

6.3 Finalizing Matches

The selection table of 17 groups of 3 (case matched by two types of controls). The final 51 subjects (17 cases (LB+NL+), 17 controls type 1 (LB-NL+), 17 controls type 2 (LB-NL-)) along with nigral LB and neuronal loss neurpathologies reported in **Supplementary Table 1** .

7 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, methods, parallel, stats, utils
- Other packages: Boruta 5.0.0, caret 6.0-68, digest 0.6.9, doParallel 1.0.10, doRNG 1.6, foreach 1.4.3, ggbeeswarm 0.5.0, ggplot2 2.2.1, gplots 3.0.1, igraph 1.0.1, iterators 1.0.8, knitr 1.14, lattice 0.20-34, LewyBodies.SN.Proteomics.StudyDesign4Pub 1.0, pkgmaker 0.22, randomForest 4.6-12, ranger 0.4.0, RColorBrewer 1.1-2, registry 0.3, reshape2 1.4.1, rngtools 1.2.4
- Loaded via a namespace (and not attached): assertthat 0.1, beeswarm 0.2.3, BiocStyle 2.2.1, bitops 1.0-6, car 2.1-2, caTools 1.17.1, class 7.3-14, codetools 0.2-15, colorspace 1.3-2, e1071 1.6-7, evaluate 0.10, formatR 1.4, gdata 2.17.0, grid 3.3.2, gtable 0.2.0, gtools 3.5.0, highr 0.6, KernSmooth 2.23-15, labeling 0.3, lazyeval 0.2.0, lme4 1.1-12, magrittr 1.5, MASS 7.3-45, Matrix 1.2-7.1, MatrixModels 0.4-1, mgcv 1.8-15, minqa 1.2.4, munsell 0.4.3, nlme 3.1-128, nloptr 1.0.4, nnet 7.3-12, pbkrtest 0.4-6, plyr 1.8.4, quantreg 5.29, Rcpp 0.12.9, scales 0.4.1, SparseM 1.74, splines 3.3.2, stats4 3.3.2, stringi 1.1.2, stringr 1.2.0, tibble 1.2, tools 3.3.2, vipor 0.3.2, xtable 1.8-2