**ESM Materials**

**Variable selection and validation in multivariate modelling**

Lin Shi1,2,\*, Johan A Westerhuis3,4, Johan Rosén5, Rikard Landberg1,2, Carl Brunius 2

1Department of Molecular Sciences, Swedish University of Agricultural Sciences, Uppsala, Sweden; 5 2Department of Biology and Biological Engineering, Food and Nutrition Science, Chalmers University of Technology, Gothenburg, Sweden; 3Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, Netherlands; 4Metabolomics Center, North-West University, Potchefstroom, South Africa; 5Swedish National Food Agency, Uppsala, Sweden

\*To whom correspondence should be addressed.

Contact: shlin@chalmers.se

**Supplementary Table S1**: Validation characteristics of repeated double cross-validation without (rdCV) or with unbiased variable selection (MUVR) on three data sets.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Method | Original variables | Validation (min) | Minimal-optimal a | All-relevant b | Model performance c | | | | |
| ‘mid’ | nComp d | Prediction fitness | | *p* for permutation e |
| Regression on “**Freelive”** data | | | | | | | | | |
| rdCV-PLS | 1147 | 14 | - | - | - | 4 | R2=0.91 | Q2=0.14 | 0.04 (n=100) |
| MUVR-PLS | 1147 | 38 | 22 | 45 | 31 | 3 | R2=0.81 | Q2=0.61 | 0.001 (n=100) |
| rdCV-RF | 1147 | 110 | - | - | - |  | R2=0.31 | Q2=0.27 | 1.01×10-8 (n=60) |
| MUVR-RF | 1147 | 127 | 8 | 22 | 13 |  | R2=0.58 | Q2=0.43 | 2.7×109 (n=60) |
| Classification on “**Mosquito**” data | | | | | | | | | |
| rdCV-PLS | 738 f | 4 | - | - | - | 2 | 10 misclassification | | 0.04(n=500) |
| MUVR-PLS | 738 f | 6 | 31 | 119 | 61 | 2 | 3 misclassification | | 9.0×10-9(n=500) |
| rdCV-RF | 1678 | 5 | - | - | - |  | 4 misclassification | | 3.5×10-6(n=500) |
| MUVR-RF | 1678 | 8 | 12 | 35 | 20 |  | 4 misclassification | | 5.7×10-6(n=500) |
| Multilevel on “**Crisp**” data | | | | | | | | | |
| rdCV-PLS | 1508 | 4 | - | - | - |  | 8 misclassification | | 0.01(n=100) |
| MUVR-PLS | 1508 | 9 | 30 | 145 | 66 |  | 3 misclassification | | 4.1×10-7(n=100) |
| rdCV-RF | 1508 | 10 |  |  |  |  | 3 misclassification | | 4.1×10-8(n=100) |
| MUVR-RF | 1508 | 11 | 5 | 8 | 6 |  | 4 misclassification | | 1.1×10-8(n=100) |

aNumber of variables selected as minimal-optimal using MUVR ‘min’ model.

bNumber of variables selected as all-relevant using MUVR ‘max’ model;

cModel performance of MUVR ‘mid’ model is presented and was compared with rdCV without variable selection.

dModel optimum of PLS components.

eThe permutation p-value of actual model performance vs random permutation distribution (number of permuted models). The number of permutations per variable selection technique varied due to computation time and to meet the Gaussian distribution.

fVariables with near-zero-variance were excluded when applying PLS core modelling.

**Supplementary Table S2** Features selected from MUVR-PLS that are suggested as urinary biomarkers of whole grain rye intake in free-living subjects

|  |  |
| --- | --- |
| Selected featuresa | Confirmed as putative biomarkersb |
| HNX259.999.1.4530476 | Caffeic acid sulfate |
| HNX368.2198.1.1225227 | Unknown |
| HPX283.9974.1.2967925 | Unknown |
| HPX376.1353.0.70698017 | Unknown |
| RNX109.0527.2.5151808 | HPAA sulfate (fragment) |
| RNX138.0677.2.7376492 | DHPPA derivative (fragment ion) |
| RNX228.063.2.7376049 | DHPPA derivative |
| RNX247.0152.2.2259395 | HHPAA sulfate |
| RNX259.9992.2.4562662 | Caffeic acid sulfate |
| RNX262.0145.2.1115243 | 3,5‐DHPPA sulfate |
| RNX438.0623.6.4936523 | Unknown sulfate |
| RNX438.0625.6.3937316 | Unknown sulfate |
| RNX452.0779.6.5050006 | NA |
| RNX620.1056.2.1157255 | 3,5‐DHPPA sulfate (with adduct) |
| RPX160.0739.1.2547945 | Pimelic acid |
| RPX162.0354.2.1086922 | NA |
| RPX167.0583.2.6159992 | HHPAA |
| RPX207.0532.1.1967309 | Unknown |
| RPX325.1535.2.3103251 | Unknown carnitine |
| RPX340.1647.3.171962 | Unknown carnitine |
| RPX474.2411.5.9182343 | Unknown carnitine |

aFeatures selected by MUVR-PLS, ‘mid’ model.

bMetabolites that have been suggested as urinary biomarkers of whole grain rye intake in free-living subjects. These results have been reported elsewhere (Hanhineva,K. et al. 2015. Mol. Nutr. Food Res., 59, 2315–2325). Features related with rye consumption were first identified from ‘Freelive’ data using rdcv-PLS and then were subjected to metabolite identification, based on accurate mass and product ion spectrum matching against an in-house library of authentic standards, online databases and literature. NA denotes the discrepancy between features detected by MUVR-PLS and rdcv-PLS. Although two features, i.e. RNX452.0779.6.5050006 and RPX162.0354.2.1086922 were not found by rdcv-PLS, they were significantly correlated with rye intake (r=0.51 and 0.52, respectively), suggesting their potential to be rye biomarkers.

**Supplementary Table S3**: Model performance for MUVR PLS regression on entire ‘Freelive’ data, or on an external testing set (n=18).

|  |  |  |
| --- | --- | --- |
| All dataset (n=58) | Training set (n=40) | External testing set (n=18) |
| R2=0.81 | R2=0.94 |  |
| Q2=0.61 | Q2=0.56 | Q2=0.60 |

**Supplementary Table S4**: Performance of repeated double cross validation (rdCV) models using Borutaa and VSURFb on three authentic data sets.

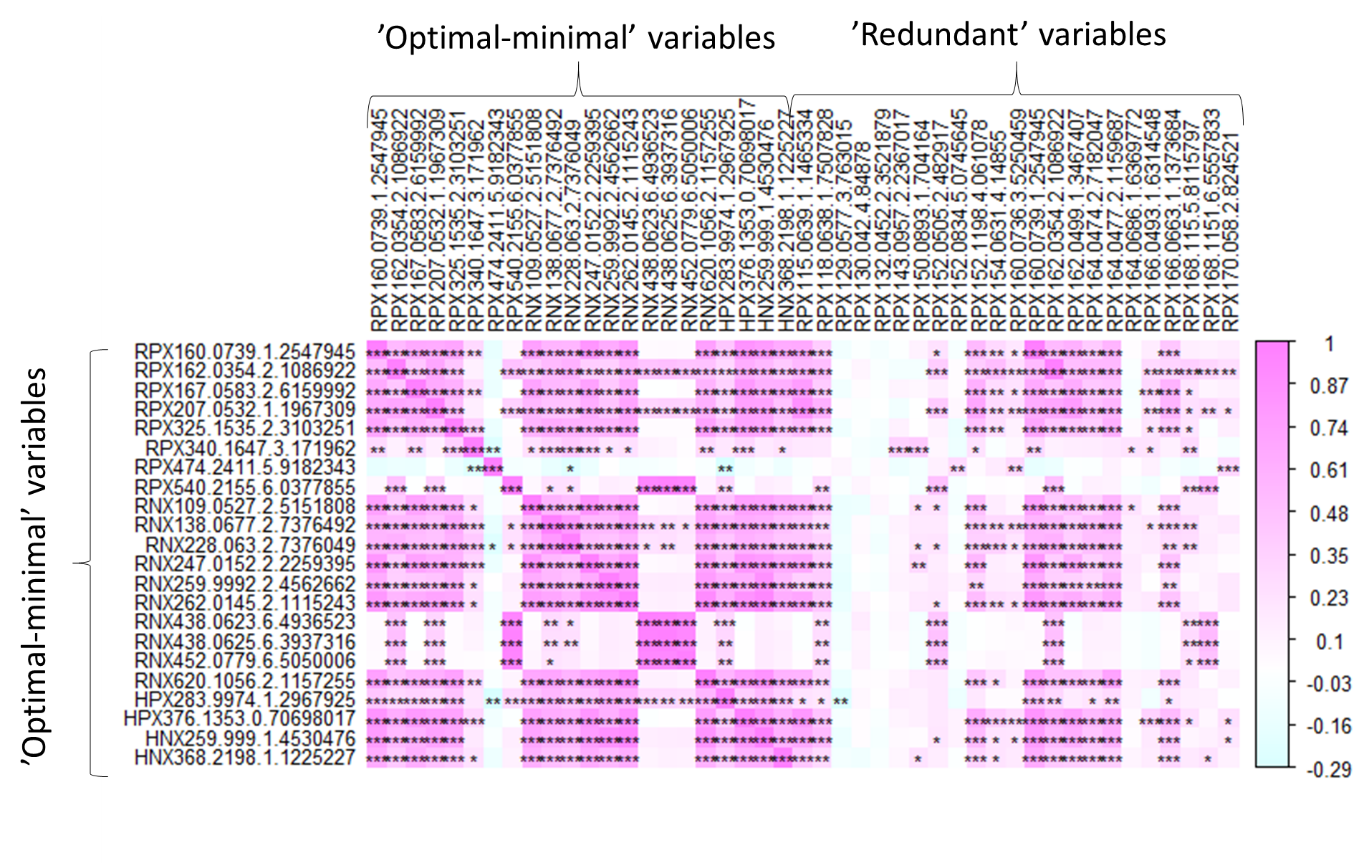
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable selection | Validation | Selection (min) | Selected variables | Model performance | | |
| nCompc | Prediction fitness | |
| Regression using “**Freelive”** data | | | | | | |
| Boruta | rdCV-PLS | 17 | 14 | 2 | R2=0.63 | Q2=0.43 |
| VSURFI | rdCV-PLS | 259 | 32 | 2 | R2=0.67 | Q2=0.36 |
| VSURFP | rdCV-PLS | 259 | 14 | 2 | R2=0.64 | Q2=0.49 |
| Boruta | rdCV-RF | 17 | 14 |  | R2=0.58 | Q2=0.48 |
| VSURFI | rdCV-RF | 259 | 32 |  | R2=0.58 | Q2=0.52 |
| VSURFP | rdCV-RF | 259 | 14 |  | R2=0.62 | Q2=0.59 |
| Classification using “**Mosquito**” data | | | | | | |
| Boruta | rdCV-PLS | 0.2 | 19 | 2 | 3 misclassifications | |
| VSURF-I | rdCV-PLS | 15 | 6 | 2 | 5 misclassifications | |
| VSURF-P | rdCV-PLS | 17 | 5 | 2 | 5 misclassifications | |
| Boruta | rdCV-RF | 0.2 | 19 |  | 2 misclassifications | |
| VSURF-I | rdCV-RF | 15 | 6 |  | 1 misclassifications | |
| VSURF-P | rdCV-RF | 17 | 5 |  | 1 misclassifications | |
| Multilevel using “**Crisp**” data d | | | | | | |
| VSURF-I | rdCV-PLS | <1 | 4 | 1 | 2 misclassifications | |
| VSURF-P | rdCV-PLS | <1 | 4 | 1 | 2 misclassifications | |
| VSURF-I | rdCV-RF | <1 | 4 |  | 4 misclassifications | |
| VSURF-P | rdCV-RF | <1 | 4 |  | 4 misclassifications | |

aRandom forest based Boruta algorithm.

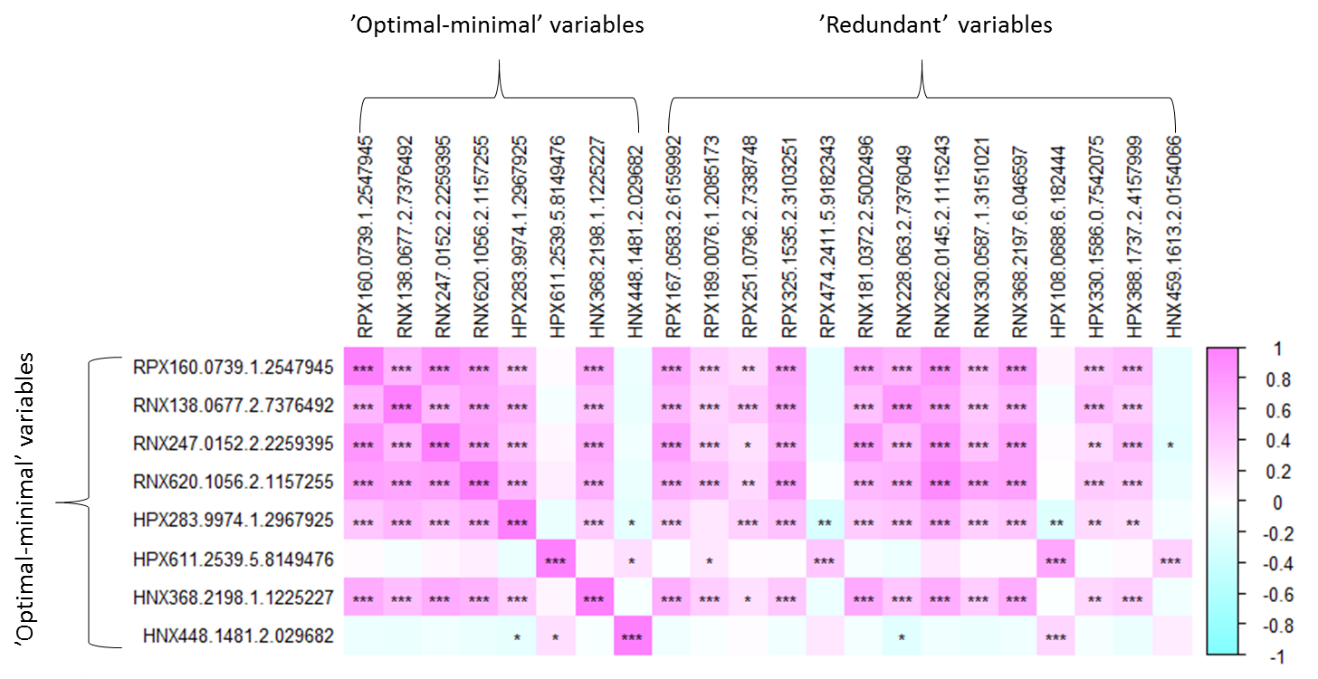
bRandom forest VSURF algorithm. VSURF-I denotes identification set. VSURF-P-denotes prediction set.

cOptimal number of PLS components according to rdCV.

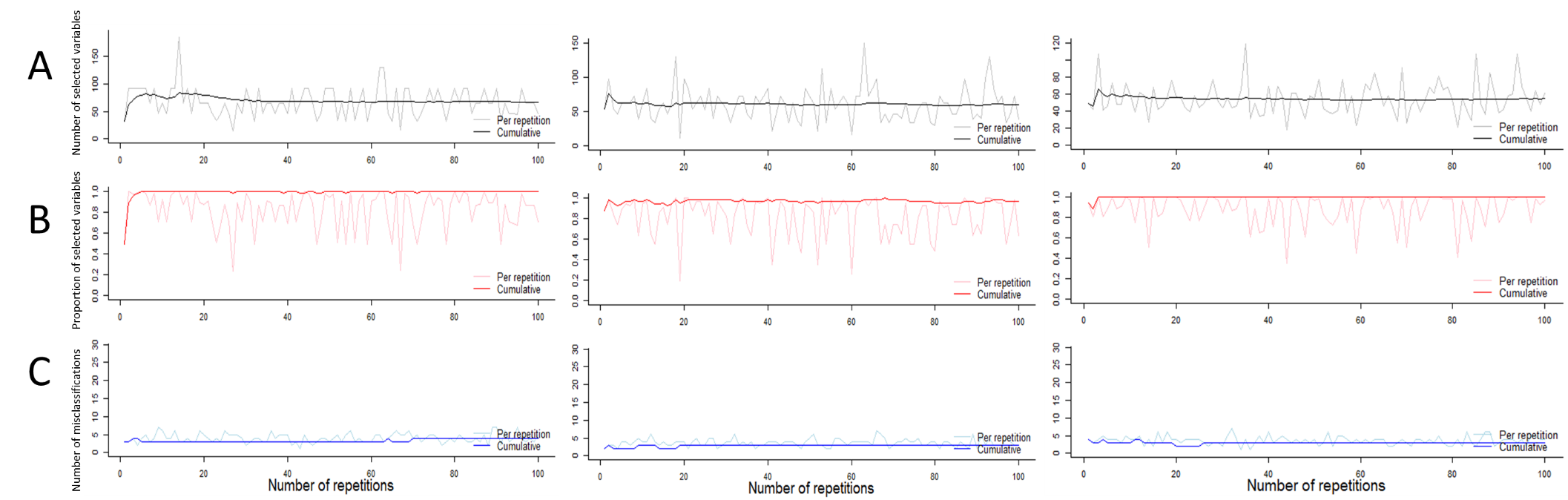
dBoruta was unable to select any variables in multilevel analysis.



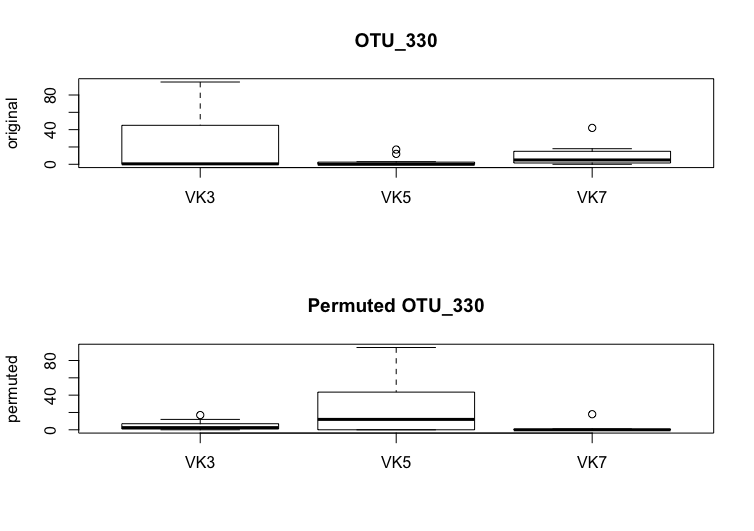
**Supplementary Fig. S1.** Correlations between optimal-minimal variable set selected from ‘min’ model of MUVR-PLS regression on ‘**Freelive’** data and redundant variables that belong to the all-relevant variable set selected in the MUVR “max” model, but not belonging to the minimal-optimal variable set. \*\*\* denotes Bonferroni adjusted p<0.001. \*\* denotes Bonferroni adjusted p<0.01. \* denotes Bonferroni adjusted p<0.05. There are 3 variables in MUVR PLS ‘redundant set’ that are not highly correlated with optimal ones. This observation in fact indicates the redundant variables might be important from a biological point of view, but do not provide additional information in model prediction.



**Supplementary Fig. S2.** Correlations between optimal-minimal variable set selected from ‘min’ model of MUVR-RF regression on ‘Freelive’ data and redundant variables that belong to the all-relevant variable set selected in the MUVR “max” model, but not belonging to the minimal-optimal variable set. \*\*\* denotes Bonferroni adjusted p<0.001. \*\* denotes Bonferroni adjusted p<0.01. \* denotes Bonferroni adjusted p<0.05.



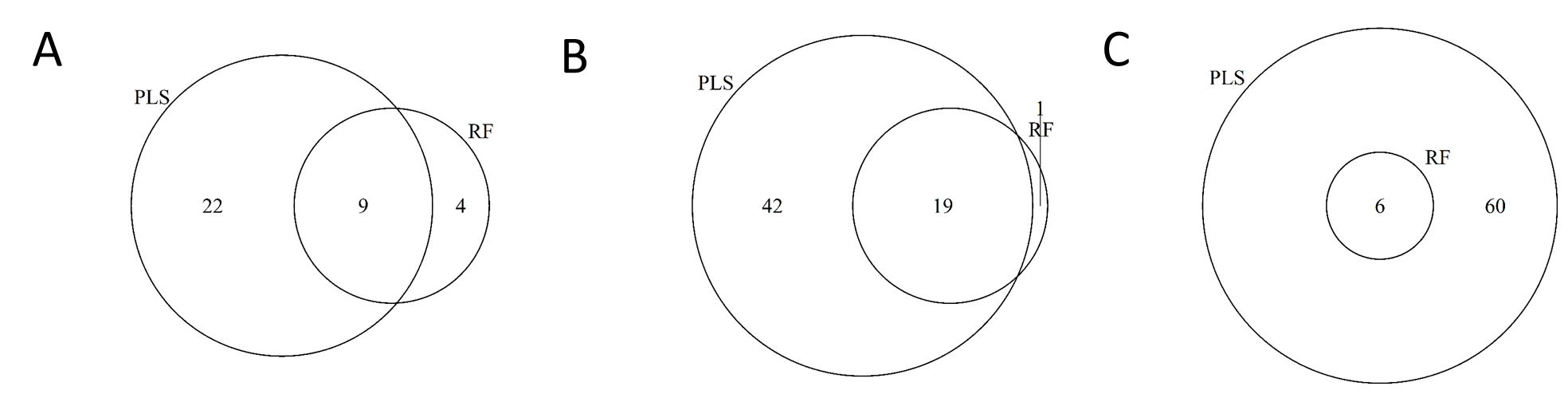
**Supplementary Fig. S3.**  Example of stability of modelling outcomes with the number of repetitions (MUVR-PLS performed on ‘**Mosquito’** data). Per repetition (light color) and cumulative (dark color) values are shown for the number of selected variables (A), the proportion of variables from the final cumulative consensus model selected (B) and number of misclassifications (C). Results are shown for *varRatio*=0.5 (left), 0.75 (middle) and 0.9 (right) and 0 ≤ repetitions ≤ 100. Visual interpretation for stability of modelling performance over the number of repetitions can be generated using the MUVR “*plotStability*” function.



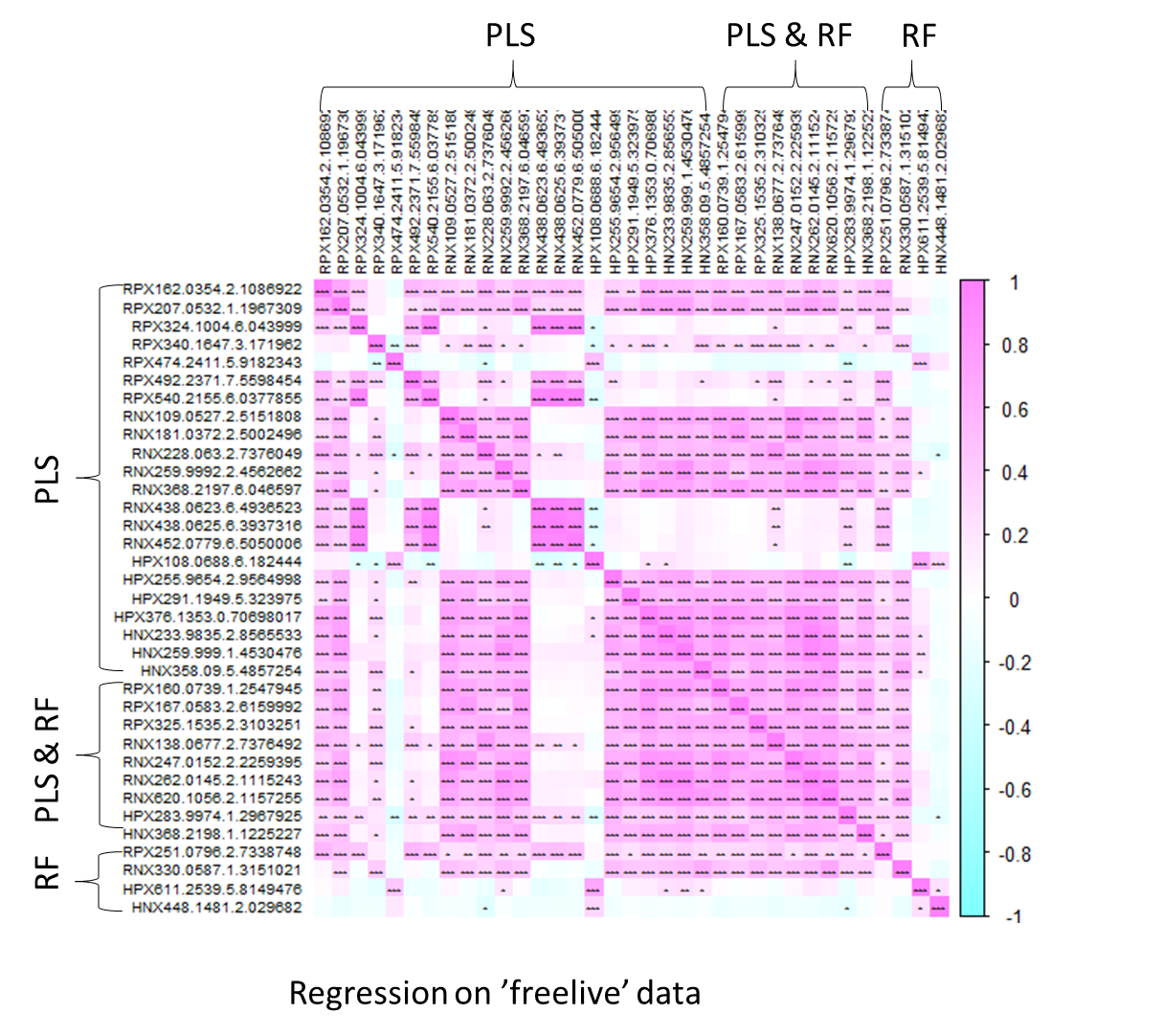
**Supplementary Fig. S4.** Example of systematic patterns caused by random permutations (based on “**Mosquito**” data, classification). The original variable OUT-330 (upper) was permuted (lower) and subsequently identified in the MUVR-PLS ‘min’ model as belonging to the optimal class.



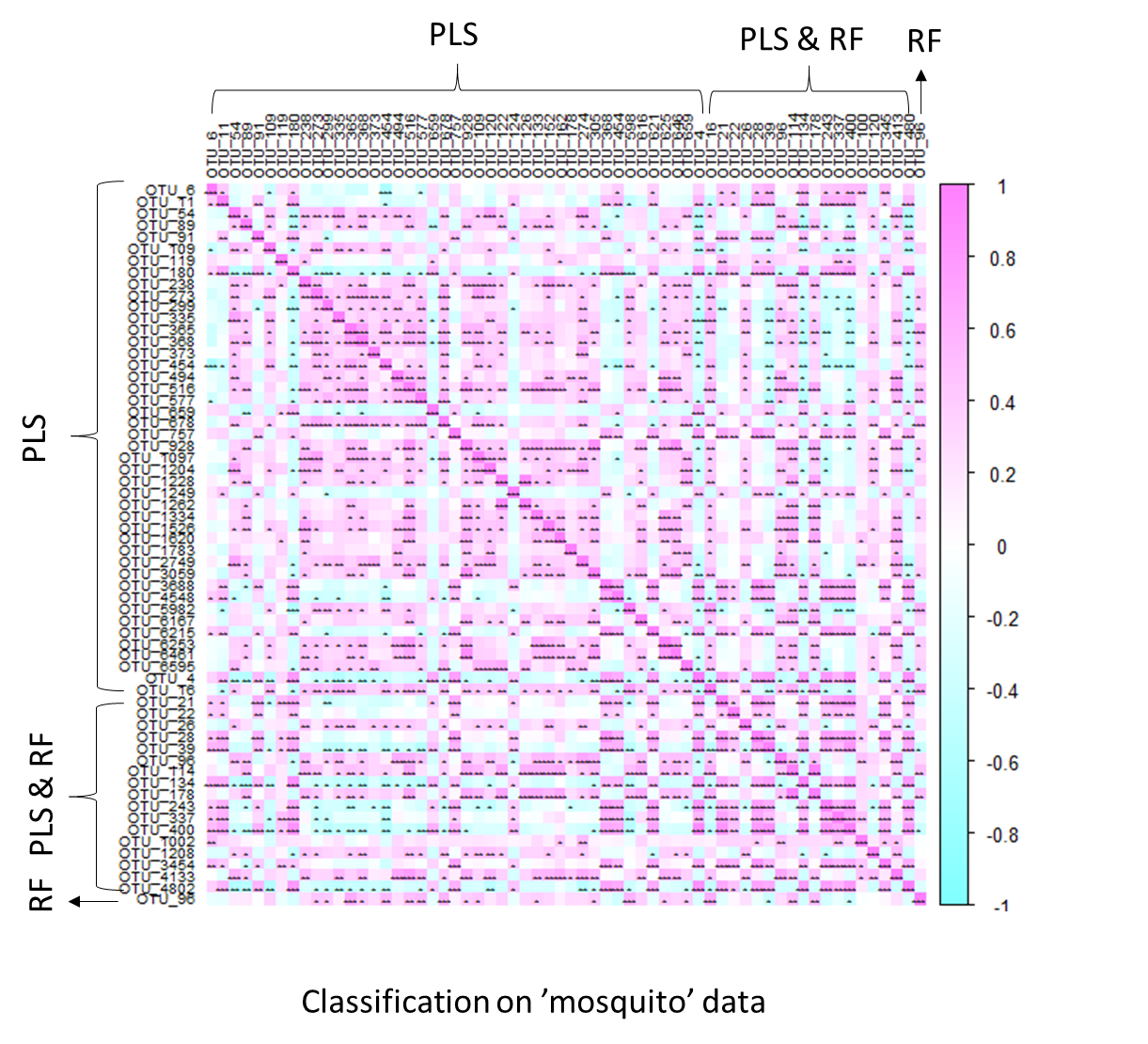
**Supplementary Fig. S5**. Predictive results of minimal-optimal models (MUVR “min” models) conducted for three datasets using PLS (left) and Random Forest (right): A) Regression using “**Freelive**” data. Prediction estimates are shown in grey for each repetition and in black for the prediction estimates averaged over all repetitions. B) Classification using “**Mosquito**” data. Each swimlane represents one observation. Class probabilities are colour-coded by class and presented per repetition (smaller dots) and averaged over all repetitions (larger dots) and misclassified samples are circled. C) Multilevel analysis using “**Crisp**” data. Samples are matched row-wise between upper and lower half for the positive and negative effect matrix, respectively. All upper half predictions should be negative and all lower half predictions positive for ideal classification. Prediction estimates are shown in grey for each repetition, and in black for the prediction estimates averaged over all repetitions and misclassified samples are circled. Visual interpretation for prediction results of optimal models can be generated using the MUVR “*plotMV*” function. PLS: partial least squares regression; RF: random forest.



**Supplementary Fig. S6**. Venn diagrams representing the optimal variables selected by MUVR “Min” models using PLS and RF modelling, and their overlap. Optimal variables were selected from three authentic omics data sets: A. Regression using “**Freelive**” data; B: Classification using “**Mosquito**” data and; C: Multilevel analysis using “**Crisp**” data. PLS: partial least squares regression; RF: random forest.



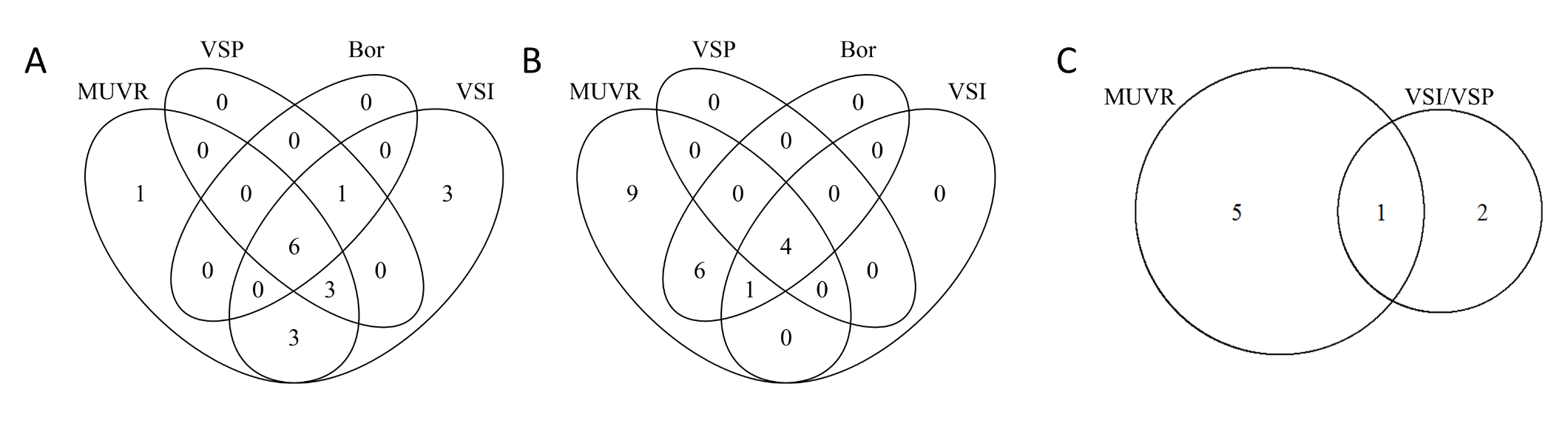
**Supplementary Fig. S7**. Correlations between variables selected by using MUVR-PLS modelling and MUVR-random forest modeling on ‘**Freelive’** data. PLS & RF represents the variables selected by both PLS and RF. PLS: partial least squares regression; RF: random forest. \*\*\* denotes Bonferroni adjusted p<0.001. \*\* denotes Bonferroni adjusted p<0.01. \* denotes Bonferroni adjusted p<0.05.



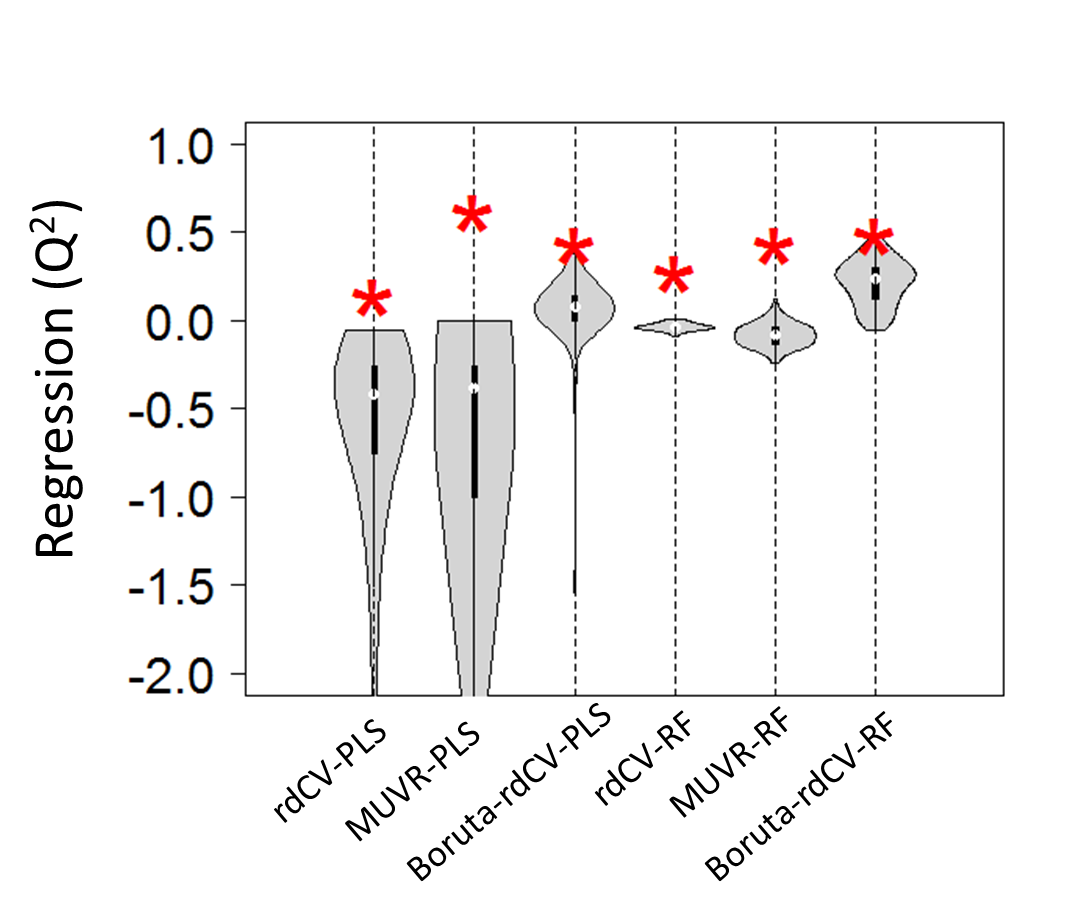
**Supplementary Fig. S8**. Correlations between variables selected by using MUVR-PLS modelling and MUVR-random forest modeling on ‘**Mosquito’** data. PLS & RF represents the variables selected by both PLS and RF. PLS: partial least squares regression; RF: random forest. \*\*\* denotes Bonferroni adjusted p<0.001. \*\* denotes Bonferroni adjusted p<0.01. \* denotes Bonferroni adjusted p<0.05.



**Supplementary Fig. S9**. Correlations between variables selected by using MUVR-PLS modelling and MUVR-random forest modeling on ‘**Crisp’** data. PLS & RF represents the variables selected by both PLS and RF. PLS: partial least squares regression; RF: random forest. \*\*\* denotes Bonferroni adjusted p<0.001. \*\* denotes Bonferroni adjusted p<0.01. \* denotes Bonferroni adjusted p<0.05.



**Supplementary Fig. S10**. Venn diagrams representing the variables selected by MUVR-RF (optimal variables from the “min” model), Boruta, VSURF, and their overlap. Optimal variables were selected from three authentic omics data sets: A. Regression using “**Freelive**” data; B: Classification using “**Mosquito**” data and; C: Multilevel analysis using “**Crisp**” data. VSURF-I denotes identification set. VSURF-P-denotes prediction set. Variables selected by VSURF-I and VSURF-P were identical. No variable was selected by Boruta algorithm with default parameter settings. RF: random forest.



**Supplementary Fig. S11**. Performance of repeated double cross-validation models (rdCV) without variable selection or with variable selection using either MUVR or Boruta. Models were built from actual data and random permutations for the “**Freelive**” regression task. The performance distributions of random permutations are represented as violin plots with the red asterisks representing actual model performance. In total, 100 permutations were performed for MUVR-PLS and rdCV-PLS and 60 for random forest (p<0.0001). For Boruta, 100 permutations were performed (p>0.05). PLS: partial least squares regression; RF: random forest.