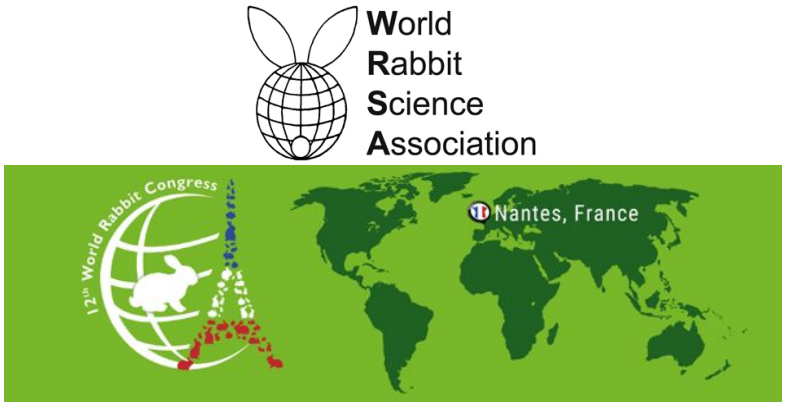


PROCEEDINGS OF THE 12th WORLD RABBIT CONGRESS

Nantes (France) - November 3-5, 2021
ISSN 2308-1910

Session
BREEDING & GENETICS

Piles M., Tusell L., Velasco-Galilea M., Helies V., Drouilhet L., Zemb O., Sánchez J.P., Garreau H.

MACHINE LEARNING ALGORITHMS FOR THE PREDICTION OF FEED EFFICIENCY BASED ON CAECAL MICROBIOTA

Full text of the communication
+
Slides of the presentation

How to cite this paper
Piles M., Tusell L., Velasco-Galilea M., Helies V., Drouilhet L., Zemb O., Sánchez J.P., Garreau H., 2021. Machine learning algorithms for the prediction of feed efficiency based on caecal microbiota. Proceedings 12th World Rabbit Congress - November 3-5 2021 - Nantes, France, Communication BG-21, 4 pp. + presentation

MACHINE LEARNING ALGORITHMS FOR THE PREDICTION OF FEED EFFICIENCY BASED ON CAECAL MICROBIOTA

Piles M.[1]*, Tusell L.[2], Velasco-Galilea M.[1], Helies V.[2], Drouilhet L.[2], Zemb O.,[2] Sánchez J.P.[1], Garreau H.[2]

[1]Animal Breeding and Genetics Program, Institute of Agriculture and Food Research and Technology (IRTA), Torre Marimon, E08140 Caldes de Montbui, Barcelona, Spain
[2]GenPhySE, Université de Toulouse, INRAE, F-31326 Castanet-Tolosan, France
*Corresponding author: miriam.piles@irta.es

ABSTRACT

This study aimed at predicting feed conversion ratio (FCR) of young rabbits from abundances of amplicon sequence variants (ASVs) to improve this trait by selecting animals with the most favorable microbiota and identifying the most relevant microorganisms involved in feed efficiency. Data come from two rabbit populations coming from paternal INRA 1001 line (the G10, selected for 10 generations for decreased residual feed intake and the G0 control produced from frozen embryos of the common ancestor line). There were 296 and 292 FCR data from G10 and G0 individuals, respectively. Phenotypic data were pre-corrected for the systematic effects of group, batch, litter size and sex and the random litter effect. Sequence quality control and chimera removal were performed with the DADA2 pipeline. Samples with less than 5,000 final sequence counts and doubleton ASV were removed. The ASV counts of the final table (including 918 ASVs) were centered log-ratio transformed and corrected for batch effects with a surrogate variable analysis. Nested resampling for hyper-parameter tuning and prediction validation was implemented leading to 25 pairs of training/test sets. Bayesian regression models (Bayesian Lasso, Bayesian Ridge Regression and Reproducing Kernel Hilbert Spaces) and machine learning algorithms (Support vector machine and Elastic net) were fitted to all ASVs leading to an almost null prediction accuracy in all cases. Then, ASVs were ranked for their prediction importance using the permutation accuracy importance score in a Random Forest algorithm based on conditional inference and, different subsets of increasing size (50, 100, 150, 200, 300, 400, 500, All) of the most important ASVs and surrogate variables were used as predictors in the machine learning algorithms. The best performance and the most stable results were obtained with machine learning using the 100 most important ASVs being most of them assigned to order *Clostridiales*. The medians of the Spearman correlation (interquartile range) were 0.33 (0.09) and 0.32 (0.06) for SVM and ENET, respectively.

Key words: feed efficiency, machine learning, caecal microbiota, prediction, selection.

INTRODUCTION

Rabbit gut microbiota plays an important role in production traits (Drouilhet *et al*., 2016) because of its effect on metabolic, nutritional, physiological, and immunological processes. Among production traits, feed efficiency (FE) is one of the most important components of productivity, profitability and sustainability of meat production and, therefore, improving this trait is a priority. One possible strategy could be to change animal's gut microbial composition based on its effect on animal performance. In addition, recent studies indicate that gut microbiota is heritable and could be modified by selection (Velasco-Galilea *et al*., 2018; Crespo-Piazuelo et al., 2019). Therefore, selecting animals with the optimal microbial composition based on its effect on FE could also lead to selection of individuals with genes that promote the presence of those beneficial microorganisms. Selection would be based on the prediction of FE (previously corrected by environmental factors) obtained from high-throughput deep sequencing data of microbial composition. Machine learning (ML) algorithms can be suitable

models because they are efficient for finding generalizable patterns from high-dimensional data in a small number of samples.

This research aimed at assessing the suitability of ML algorithms for the prediction of feed efficiency from abundances of amplicon sequence variants (ASVs) and identifying the most relevant microorganisms involved.

MATERIALS AND METHODS

Animal material and experimental design.

The experimental rabbits came from the paternal INRA 1001 line. Two populations were used in this analysis: G10, selected for 10 generations for decreased residual feed intake (RFI) (Drouilhet *et al*., 2016), and G0 control produced from frozen embryos of the ancestor population of the selected line. The 296 G10 and 292 G0 rabbits were produced in 3 batches with a 42 days interval. In each batch, half of the kits were fostered by G0 does and the rest by G10 does. The does adopted alternatively kits from both lines in successive batches. At weaning (32 days), kits were placed in individual cages. More details about the experiment can be found in Garreau *et al*., (2019). Genomic DNA of caecal samples collected from 588 kits was extracted with ZR Soil Microbe DNA MiniPrep^{TM} kit (ZymoResearch, Freiburg, Germany). A fragment containing V4-V5 hypervariable regions of the 16 rRNA gene was amplified with the pair of primers F515Y/R926 (Parada *et al*., 2016) and re-amplified in a limited-cycle PCR to add barcodes of multiplex Nextera® XT kit (Illumina, Inc., San Diego CA, United States) following the manufacturer's instructions. Final libraries were paired-end sequenced in parallel in a MiSeq Illumina 2x250 platform at the Autonomous University of Barcelona.

Bioinformatics

Sequence processing was performed using QIIME2 software (version 2018.6; Bolyen *et al*., 2018). Sequence quality control and chimera removal were performed in a single step with the DADA2 pipeline (Callahan *et al*., 2016), implemented through the q2-dada2 plugin. The output table containing the counts of unique sequences for each sample, i.e., 100% ASVs, was clustered into ASVs with 99% similarity. The ASV table was filtered at: (1) sample level by discarding samples with less than 5,000 final sequence counts and at (2) ASV level by removing the doubleton ones. The ASV counts of the final table (including 918 ASVs) were centered log-ratio transformed using the R package "chemometrics" to account for the compositional nature of microbiota data. Taxonomic assignment of ASVs was conducted by mapping them to the Greengenes reference database.

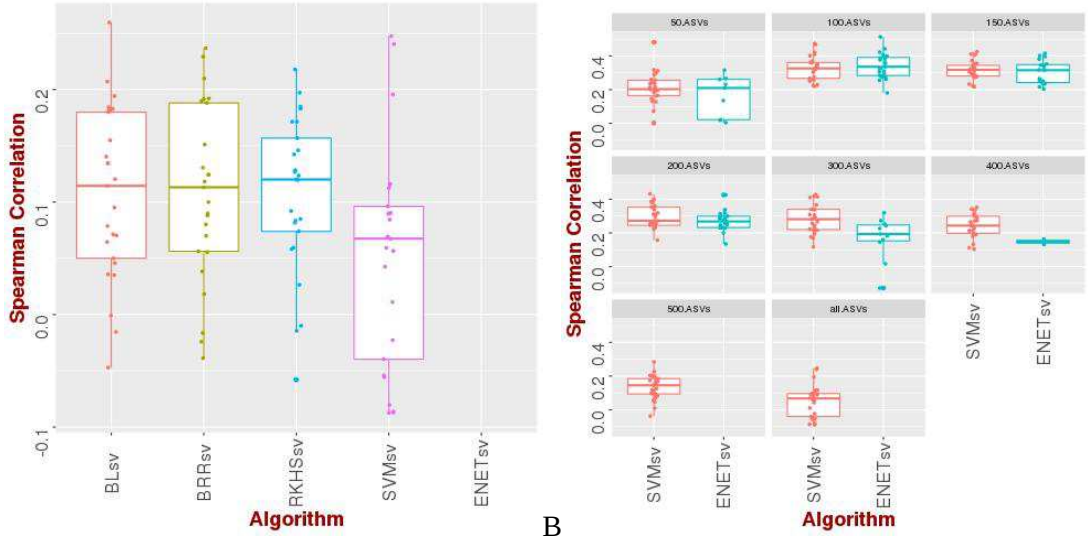
Data and Statistical Analysis

Feed efficiency was measured as feed conversion ratio (FCR), i.e., feed intake divided by body weight gain. The statistical analysis was performed in three steps. In a first step, FCR records were pre-corrected for the systematic effects of group, batch, litter size and sex, and the random effect of litter. Then, a surrogate variable analysis (Leek and Storey, 2007) was performed using the R package "SVA" to include surrogate variables (SV) in the model of prediction which allows accounting for unnoticed factors of variation affecting ASVs abundances. In a second step, the ASVs were ranked for their predictive importance using the permutation accuracy importance score in a Random Forest algorithm based on conditional inference (Strobl *et al*., 2007). In the last step, different subsets of increasing size (i.e., 50, 100, 150, 200, 300, 400, 500 and 918) of the most important ASVs and SV were selected as predictors of FCR using two machine learning algorithms. Support Vector Machine (SVM; Vapnik *et al*., 1999) and Elastic Net (ENET; Zou and Hastie, 2005) algorithms were implemented using the "mlr" R package which allows to compare results from different algorithms under the same conditions and to find the optimal hyper-parameters for each algorithm. Nested resampling for hyper-parameter tuning was implemented. It consisted of 2 nested resampling loops. In the outer resampling loop, a 5-fold cross-validation was repeated 5 times originating 25 pairs of training/testing sets. On each of those outer training sets, hyper-parameter tuning was done in an inner resampling loop of 5-fold cross-validation repeated 2 times using the R-squared performance criterion.

One set of selected hyper-parameters was obtained for each outer training set. The learner was fitted on each training set using the selected hyper-parameters and its performance was evaluated on the corresponding testing set. Predictive ability was assessed as the Spearman correlation (SC) between the observed and predicted records in the testing sets. On the other hand, Bayesian regression models (de los Campos *et al*., 2013) such as Bayesian Lasso (BL), Bayesian Ridge Regression (BRR) and Reproducing Kernel Hilbert Spaces (RKHS; Gianola *et al*., 2006) were also implemented in the same 25 pairs of training/test sets using all ASVs and SV as predictors with "BGLR" R package (Pérez & de los Campos, 2014).

RESULTS AND DISCUSSION

Using as predictors all ASVs and SV (Figure 1, panel A), ENET was not able to fit a model because of lack of convergence and SVM had a null prediction ability with a very large variability among sets (the median and interquartile range (IQR) of the SC were -0.07 and 0.14, respectively). Predictive performance was slightly better but still very low for BL, BRR and RKHS algorithms being the median of the SC (IQR) 0.11 (0.13), 0.11 (0.13) and 0.12 (0.08), respectively. When feature selection was performed (Figure 1, panel B), the predictive performance improved significantly. The best performances and the most stable results were obtained with SVM and ENET using the 100 most important ASVs. The medians of the SC (IQR) were in this case 0.33 (0.09) and 0.32 (0.06) for SVM and ENET, respectively.



A B

Figure 1: Boxplots of Spearman correlations between observed and predicted FCR obtained in 25 training/testing datasets with the different algorithms using all ASVs (panel A) or subsets with increasing number of ASVs (panel B).

Taxonomic assignment of representative sequences revealed that most (74) of the ASVs belong to order *Clostridiales*. In animals with low FCR performances, 32 ASVs belonging to order *Clostridiales* (families *Lachnospiraceae* (8), *Ruminococcaceae* (8), *Clostridiaceae* (1) and unknown (15)), 6 ASVs belonging to order *Bacteroidales* (families *Bacteroidaceae* (1), *Rikenellaceae* (2), *S24-7* (2) and unknown (1)) and 2 ASVs belonging to phylum *Tenericutes* (orders *RF39* and *ML615J-28*) were overrepresented. In addition, two completely unknown ASVs were also associated with high efficient animals. For animal with high FCR, 42 ASVs belonging to order *Clostridiales* (families *Lachnospiraceae* (14), *Ruminococcaceae* (15) and unknown (13)), 10 ASVs belonging to order *Bacteroidales* (families *Bacteroidaceae* (3), *Rikenellaceae* (3), *S24-7* (3) and unknown (1)), 2 ASVs belonging to phylum *Tenericutes* (order *RF39*,) 2 ASVs belonging to order *Verrucomicrobiales* (genus *Akkermansia*) and 2 ASVs belonging to phylum *Proteobacteria* (families *Oxalabacteraceae* and *Desulfovibrionales*) were overrepresented.

CONCLUSIONS

Support Vector Machine and Elastic net algorithms enabled the best prediction of FCR when the abundances of the 100 most important ASVs were used as predictive variables. Taxonomic assignment of the representative sequences of these selected ASVs revealed that different species belonging to order *Clostridiales* are involved in feed efficiency.

ACKNOWLEDGEMENTS

This research was supported by the Feed-a-Gene Project funded by the European's Union H2020 Programme under grant agreement EU 633531.

REFERENCES

Bolyen E., Rideout J.R., Dillon M.R. *et al*. 2018. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. *PeerJ Preprints, e27295v1*.

Callahan B.J., McMurdie P.J., Rosen M.J., Han A.W., Johnson A.J.A., Holmes S.P. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. *Nature Meth., 13(7):581.*

de Los Campos G., Hickey J.M., Pong-Wong R., Daetwyler H.D., Calus M.P.L. 2013 Whole-genome regression and prediction methods applied to plant and animal breeding, Genetics, 193, 327-345.

Crespo-Piazuelo D., Migura-Garcia L., Estellé J., Criado-Mesas L., Revilla M., Castelló A., Muñoz M., García-Casco J.M., Fernández A.I., Ballester M., Folch J.M. 2019. Association between the pig genome and its gut microbiota composition, Scientific Reports, 9, 8791.

Drouilhet, L., C. S. Achard, O. Zemb, C. Molette, T. Gidenne, C. Larzul, J. Ruesche, A. Tircazes, M. Segura, T. Bouchez, M. Theau-Clement, T. Joly, E. Balmisse, H. Garreau, and H. Gilbert. 2016. Direct and correlated responses to selection in two lines of rabbits selected for feed efficiency under ad libitum and restricted feeding: I. Production traits and gut microbiota characteristics. *J Anim. Sci*., 94, 38-48.

Gianola D, Fernando R.L., Stella A. 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics, 173, 1761–1776.*

Leek J.T., Scharpf R.B., Bravo H.C., Simcha D., Langmead B., Johnson W.E., Geman D., Baggerly K., Irizarry R.A. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet., 11, 733-739.*

Parada A.E., Needham D.M., Fuhrman J.A. 2016. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Env. Microbiol., 18(5):1403-1414.*

Perez P., de los Campos G. 2014. Genome-wide regression and prediction with the BGLR statistical package. *Genetics, 198, 483-495.*

Strobl C, Boulesteix A.L., Zeileis A., Hothorn T. 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinfo., 8(1), 25.*

Vapnik V.N. 1999. The nature of statistical learning theory. 2nd ed. New York: Springer-Verlag.

Velasco-Galilea M., Piles M., Viñas M., Rafel O., González-Rodríguez O., Guivernau M., Sánchez J.P. 2018. Determinismo genético de la microbiota intestinal del conejo. In Proc. XIX Reunión de Mejora Genética Animal, 2018 June, León, Spain,

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Series B Stat. Methodol., 67, 301-20.

MACHINE LEARNING ALGORITHMS FOR THE PREDICTION OF FEED EFFICIENCY BASED ON CAECAL MICROBIOTA

Piles M.*[1], Tusell L.[2], Velasco-Galilea M.[1], Helies V.[2], Drouilhet L.[2], Zemb O.,[2] Sánchez J.P. [1], Garreau H.[2]

[1] *Animal Breeding and Genetics, Institute for Food and Agriculture Research and Technology. Barcelona, Spain.*
2GenPhySE, Université de Toulouse, INRAE, F-31326 Castanet-Tolosan, France
[*]Corresponding author: Miriam.piles@irta.es

12th WRC, Nantes 3-5, 2021

IRTA Generalitat de Catalunya

IMPROVING FEED EFFICIENCY (FE)

Selecting animals with the **optimal microbial composition based on its effect on feed efficiency (FE)** is expected to lead to selection of individuals with genes that promote the presence of those beneficial microorganisms.

HOW?

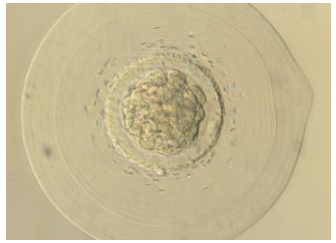
Selection based on prediction of FE from microbial composition using Machine Learning algorithms

OBJECTIVES

- assessing the suitability of ML algorithms for the prediction FE from abundances of amplicon sequence variants (ASVs)

- identifying the most relevant microorganisms involved.

ANIMALS & PHENOTYPIC INFORMATION



G.0

292 G.0 rabbits

10 generations of selection for RFI

G.10

296 G.10 rabbits

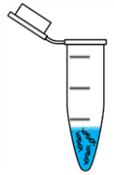
- FE as **FEED CONVERSION RATIO**
- 3 batches
- Crossfostering by G.0 does and G.10 does

BIOINFORMATICS

SAMPLE PROCESSING & SEQUENCING

Collection of cecal samples at day 66

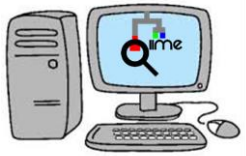
DNA extraction



Library generation

Amplification V4-V5 16S rRNA gene (primers 515Y/926R)

MiSeq sequencing



BIOINFORMATICS

- Filtering sequences & removing quimeric contigs
- Clusterization of final contigs into OTUs
- Filtering & CSS normalization of OUT table

FINAL ASV TABLE: 588 samples & 918 ASVs

Machine learning algorithms

mlr R package

Support Vector Regression (e1071)

- Find a function, f(x), with at most ε-deviation from the target y

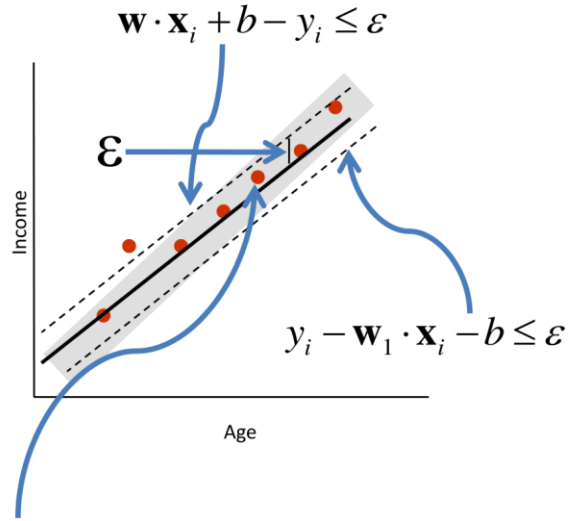
The problem can be written as a convex optimization problem

$$\min \frac{1}{2} \| \mathbf{w} \|^2$$

$$s.t.\ y_i - \mathbf{w}_1 \cdot \mathbf{x}_i - b \le \varepsilon;$$

$$\mathbf{w}_1 \cdot \mathbf{x}_i + b - y_i \le \varepsilon;$$

C: trade off the complexity



What if the problem is not feasible?

We can introduce slack variables (similar to soft margin loss function).

We do not care about errors as long as they are less than ε

Elastic net (cvglmnet)

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1$$

The elastic net penalty

$$J(\beta) = \alpha\|\beta\|^2 + (1-\alpha)\|\beta\|_1$$

$$(\text{with } \alpha = \tfrac{\lambda_2}{\lambda_2+\lambda_1})$$

$$\min_{\beta} \|\mathbf{y}-\mathbf{X}\beta\|^2 \text{ s.t. } J(\beta) \le t.$$

Bayesian Lasso (BL)

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta}_L + \boldsymbol{\varepsilon} \qquad p(\boldsymbol{\beta}_L \mid \tau^2\sigma_\varepsilon^2) = \prod_{j=1}^{p_L} N(\boldsymbol{\beta}_{L,j} \mid 0, \sigma_\varepsilon^2\tau_j^2)$$

$$\boldsymbol{\varepsilon} \sim N(0, Diag\left\{\frac{\sigma_\varepsilon^2}{w_i^2}\right\})$$

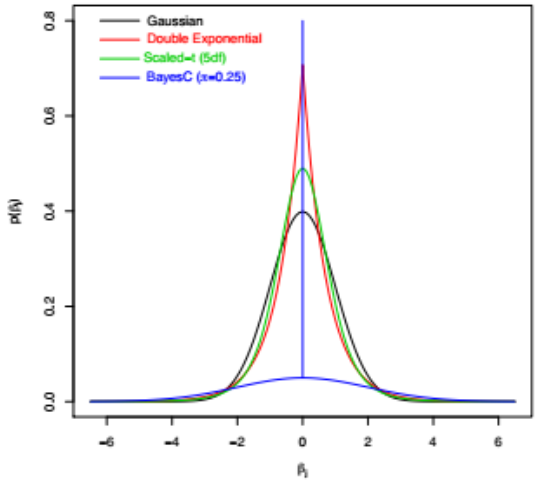


Figure 1: Prior Densities of Regression Coefficients Implemented in **BGLR**. All the densities displayed correspond to random variables with null mean and unit variance.

Bayesian Ridge Regression (BRR)

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta}_R + \boldsymbol{\varepsilon} \qquad p(\boldsymbol{\beta}_R \mid \sigma_{\boldsymbol{\beta}_R}^2) = \prod_{j=1}^{p_R} N(\boldsymbol{\beta}_{R,j} \mid 0, \mathbf{I}\sigma_{\boldsymbol{\beta}_R}^2)$$

$$\boldsymbol{\varepsilon} \sim N(0, Diag\left\{\frac{\sigma_\varepsilon^2}{w_i^2}\right\})$$

Reproducing Kernel Hilbert Space with kernel averaging, 3 Gaussian kernels used (RKHS)

$$\begin{cases} \mathbf{y} = \mathbf{1}\mu + \mathbf{g} + \boldsymbol{\varepsilon} \\ \begin{pmatrix} \boldsymbol{\varepsilon} \\ \mathbf{g} \end{pmatrix} \middle| \sigma_\varepsilon^2, \sigma_g^2 \sim N\left[\mathbf{0}, \begin{pmatrix} \mathbf{I}\sigma_\varepsilon^2 & \mathbf{0} \\ \mathbf{0} & inv(\mathbf{K})*\sigma_g^2 \end{pmatrix}\right] \end{cases}$$

$$\mathbf{y} = \mathbf{g} + \mathbf{e} = \mathbf{K}_1 f_1 + \mathbf{K}_2 f_2 + \mathbf{K}_3 f_3 + \mathbf{e}$$

$$Var(\mathbf{g}) = \sigma_{f_1}^2 inv(K_1) + \sigma_{f2}^2 inv(K_2) + \sigma_{f3}^2 inv(K_3)$$

$$\mathbf{K}_j = K_j(\mathbf{x}_i, \mathbf{x}_{i'}) = \exp(-h_j \|\mathbf{x}_i - \mathbf{x}_{i'}\|^2)$$

Statistical Analyses

1st step: Data Pre-correction and Surrogate Variable Analysis

FCR = group + batch + litter size + sex + litter (random) + e

2nd step: Gene ranking based on permutation accuracy importance score

Random Forest algorithm based on conditional inference

Data (**X**, **y**)

X is the **519 x 918 predictor matrix of ASVs**

y is the vector of adjusted **FCR data**

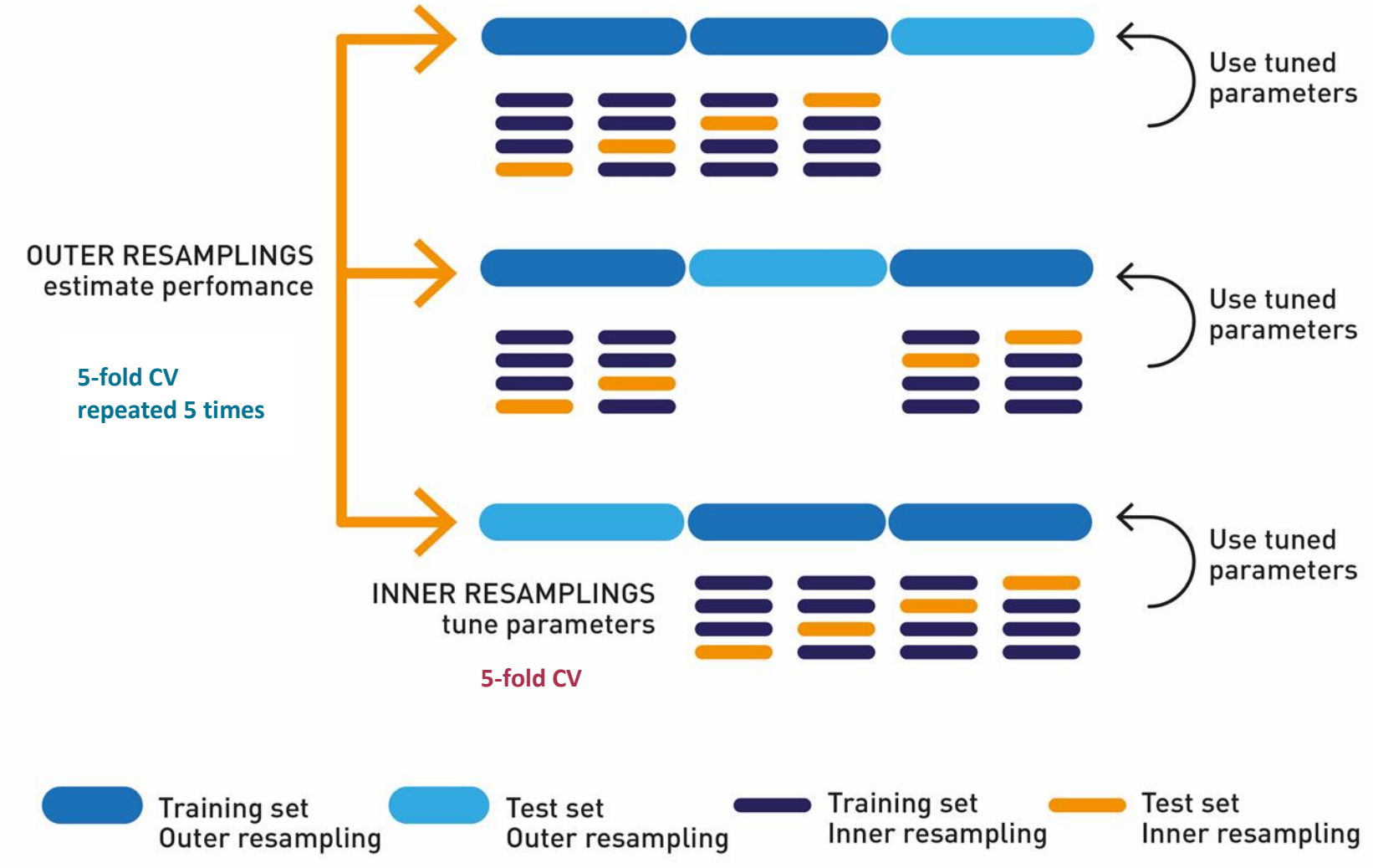
3rd step: Prediction using machine learning algorithms

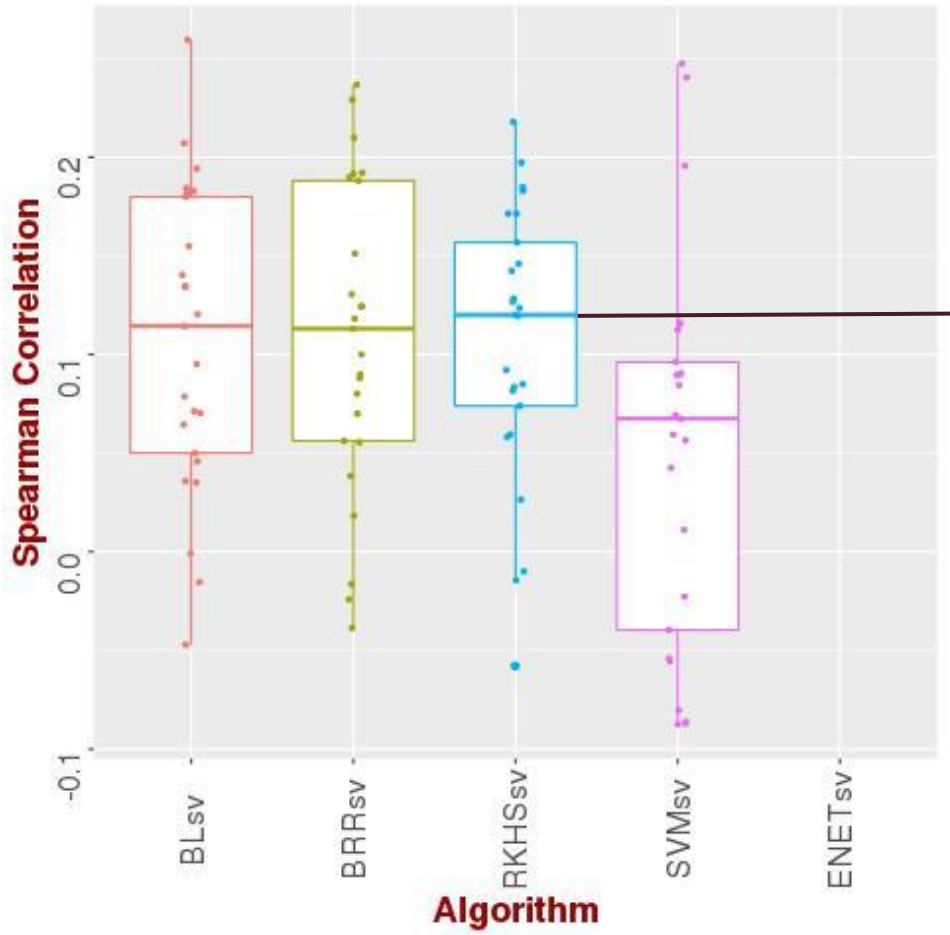
Data (**X**, **y**)

X is the **65 x p predictor matrix of ASVs**

Subsets of *p* predictors ***p* = 50, 100, 150, 200, 918**

y is the vector of adjusted **FCR data**

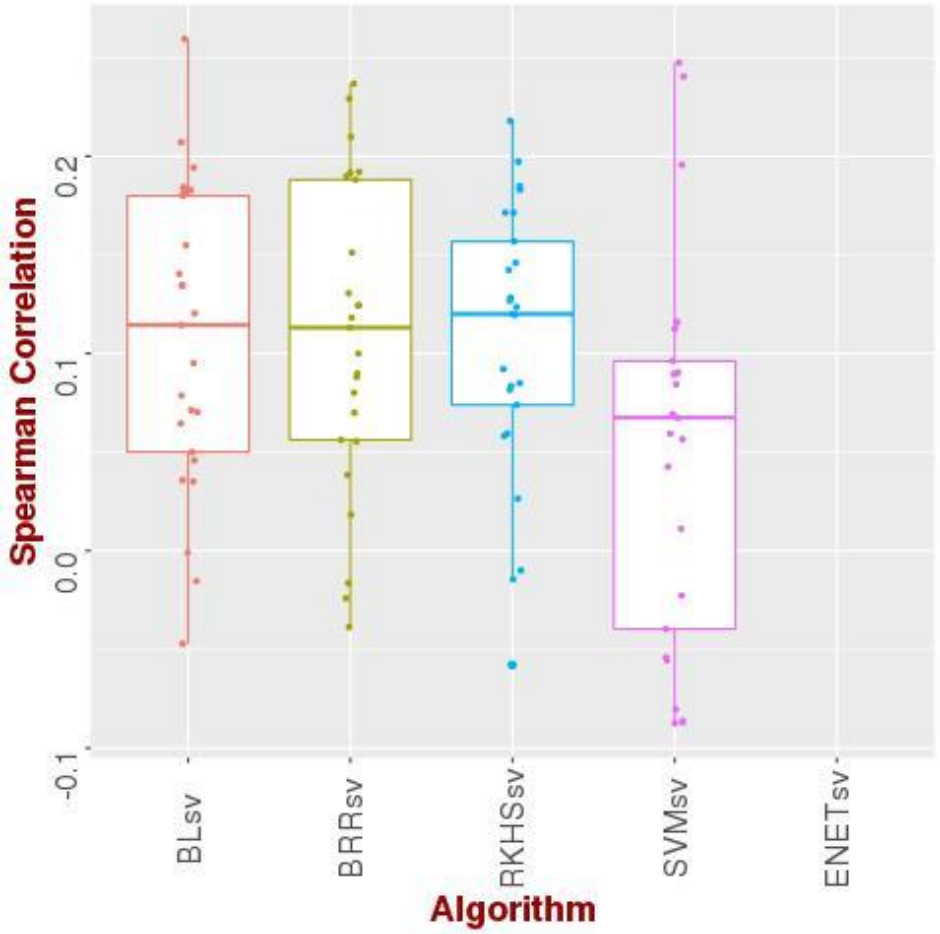

OUTER RESAMPLINGS
estimate perfomance
5-fold CV
repeated 5 times
Use tuned
parameters
Use tuned
parameters
Use tuned
parameters
INNER RESAMPLINGS
tune parameters
5-fold CV
Training set
Outer resampling
Test set
Outer resampling
Training set
Inner resampling
Test set
Inner resampling

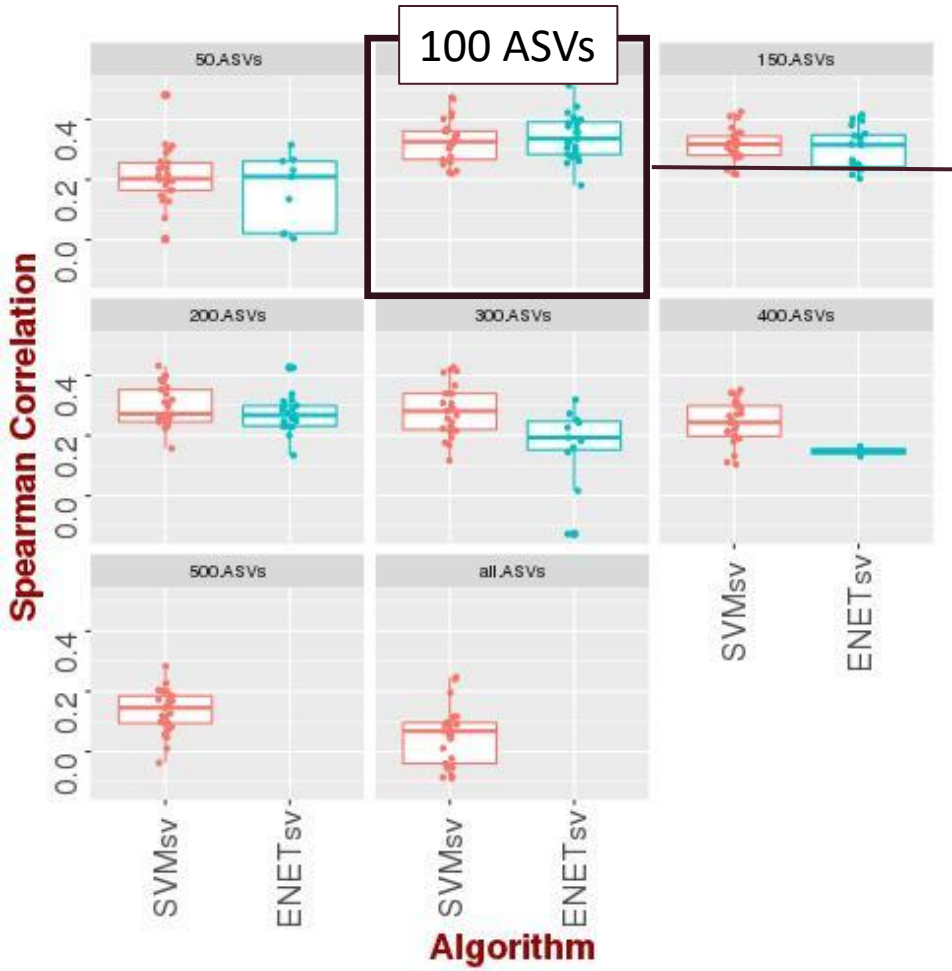


0.12 (0.08)

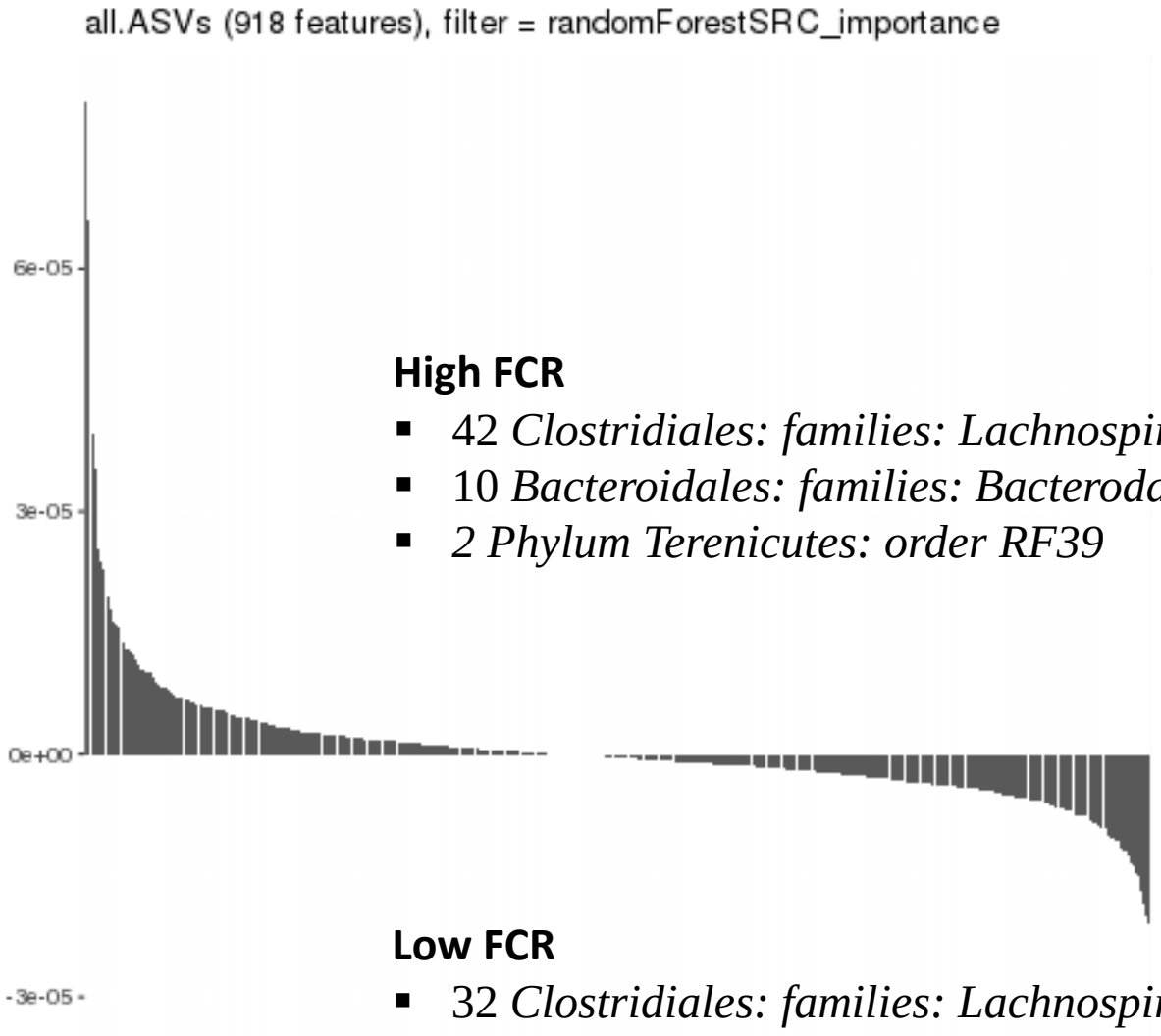
IRTA

Generalitat de Catalunya


Spearman Correlation
0.2
0.1
0.0
-0.1
BLsv
BRRsv
RKHSsv
SVMsv
ENETsv
Algorithm


100 ASVs
50.ASVs
150.ASVs
200.ASVs
300.ASVs
400.ASVs
500.ASVs
all.ASVs
Spearman Correlation
0.0
0.2
0.4
SVMsv
ENETsv
Algorithm

0.32 (0.06)



Greengenes reference database

High FCR

- 42 *Clostridiales: families: Lachnospiraceae, Ruminococcaceae, Clostridiaceae*
- 10 *Bacteroidales: families: Bacterodaceae, Rikenellaceae*
- *2 Phylum Terenicutes: order RF39*

Low FCR

- 32 *Clostridiales: families: Lachnospiraceae, Ruminococcaceae, Clostridiaceae*
- 6 *Bacteroidales: families: Bacterodaceae, Rikenellaceae*
- *2 Phylum Terenicutes: orders RF39 and ML615J-28*

TO TAKE HOME MESSAGES

- Support Vector Machine and Elastic net algorithms using 100 ASVs enabled the best prediction of FCR.

- Different species belonging to order *Clostridiales* are involved in feed efficiency.

WE SHARE OUR SCIENCE TO FEED THE FUTURE

IRTA

Generalitat de Catalunya