The holobiont 'predictome' of immunocompetence in pigs

J. Calle-García*[1], Y. Ramayo-Caldas[2], L.M. Zingaretti[3], R. Quintanilla[2], M. Ballester[2], M. Pérez-Enciso[1,4]

[1]Centre for Research in Agricultural Genomics (CRAG), 08193, Bellaterra, Spain. [2]Institute of Agrifood Research and Technology (IRTA), 08140, Caldes de Montbui, Barcelona, Spain. [3]Universidad Nacional de Villa María, Córdoba, Argentina. [4]ICREA 08010, Barcelona, Spain
joancallegarcia@gmail.com

Abstract
The objective of this work was to assess the ability of holobiont data (i.e., host's genotype and microbiota) to predict six immunity traits in 400 pigs. We propose the term 'predictome' to mean a systematic study of as many predictive methods as possible. With this spirit, we compare REML, Bayes C and Bayesian reproducing kernel Hilbert space (RKHS) regression with a wide range of priors. We assessed the performance of the models by partitioning the data into three disjoint sets. In total, we run 273 analyses per trait. We find that there does not exist a systematically best prediction method, although our results favor slightly Bayes C. By default, microbiota abundances should not be clustered. An holobiont model performs better than models using only genotype or microbiota data, i.e., use all data at your disposal.

Introduction
The role of gut microbiota composition in complex traits is well documented and it is currently a topic of utmost interest (Camarinha-Silva et al., 2017; Difford et al., 2018; Zhang et al., 2020). However, there exist relatively little evidence on the actual advantage of combining genotype and microbial information for predictive purposes. Moreover, there is a plethora of statistical methods for prediction (Gianola, 2013), but we lack a systematic review on their actual behavior in real holobiont data. In this work, we perform a comprehensive holobiont prediction analysis for important traits related to the immune system in pigs. We call this analysis the 'predictome' to signify a comprehensive set of predictive tools and their performance.

Materials & Methods
Data. We used genotype, phenotype and microbiota data fully described in (Ballester et al., 2020; Ramayo-Caldas et al., 2021), except that here we included additional 16S sequence. Very briefly, data were available from 400 pigs for six immunocompetence traits: immunoglobulins IgM (IGM) and IgG (IGG), two acute-phase proteins, C-reactive protein (CRP) and haptoglobin (HP), gamma-delta T cells (GAMMADELTA) and lymphocyte phagocytic capacity (PHAGO_LYMPH). Data were corrected for fixed effects as described (Ballester et al., 2020) and scaled. After quality control, 41,131 SNPs were retained from the original Porcine 70k GGP Porcine HD Array (Illumina).

Faecal 16S paired-end sequences were processed with QIIME2 (Bolyen et al., 2019). Amplicon Sequence Variants (ASVs) were assigned using *DADA2* R package (Callahan et al., 2016). ASVs present in less than three samples and representing less than 0.001% of the total counts were discarded. Several quality control measures were applied to ensure that 16S reads from both sequencing batches could be merged, e.g., both resulted in same Euclidean distances between samples and number of ASVs were congruent. The average number of reads per sample was over 136,000 and 2,945 ASVs were retained. We studied the effect in the prediction

of clustering by phylogeny at the genus level and by abundance vs. using individual ASV abundances. MEGA package and NJ algorithm was used for genus clustering, and *hclust* R package for abundance clustering. In either approach, 232 clusters of ASVs were obtained.

Statistical analyses. We applied a wide range of models and algorithms to characterize the predictive ability for immunocompetence in pigs. Prediction was measured as the correlation between predicted and observed phenotype averaged over three partitions, each containing ~20% of different records. Three models were compared: a holobiont model that includes genotype (**X**) and abundance data (**B**), a genetic model fitting only genotypes and a microbiota model with abundance data only. For **B**, the ASVs were fitted individually or grouped by phylogeny or by abundance.

We considered three statistical methods: REML, Bayesian RKHS regression (equivalent to GBLUP), and Bayes C. REML was run using ASREML and Bayesian methods, using BGLR (Pérez and de los Campos 2014). In RKHS, a pseudo flat prior and the default mildly informative prior were compared. In Bayes C, the following prior probabilities of a variable entering the model were compared p = 0.01, 0.001 and 0.0001 for **X** and p = 0.1, 0.01 and 0.001 for **B**. Combinations of all models, algorithms, priors and clustering approaches were run for each partition and phenotype. In total, 273 analyses per phenotype were run.

Results

The very big questions: Frequentist or Bayes? Bayes C or GBLUP? Performance in prediction is an objective, pragmatic, and easy way of contrasting statistical approaches. Figure 1 shows that REML is not necessarily worse than Bayesian methods, except in PHAGO_LYMPH**.** The method chosen was not too important except in some phenotypes like PHAGO_LYMPH, where a Bayes C modelling for both **X** and **B** outperformed other options, or haptoglobin where GBLUP methods prevailed.

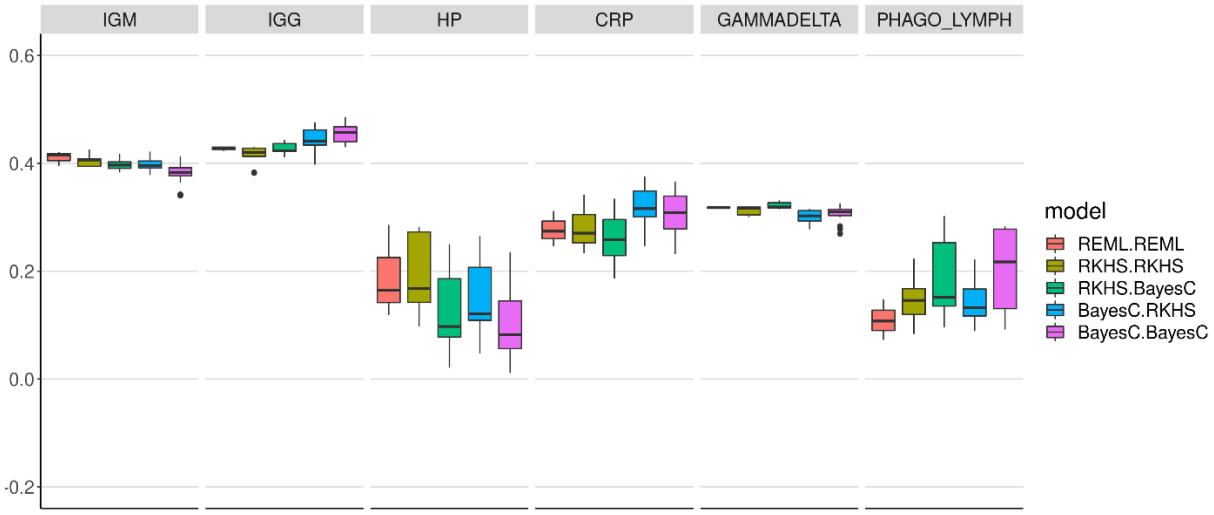

Figure 1. Predictive accuracy of the holobiont model by statistical method and trait across priors and ASV clustering. For each analysis combination, the first term (REML, RKHS, BayesC) refers to the algorithm modelling the genotype (**X**) and the second to the algorithm modeling the microbiota (**B**). For instance, RKHS.BayesC means genotype modeled as RKHS and abundances as Bayes C.

Is it worth using holobiont data? In a recent simulation study (Pérez-Enciso et al. 2021), we argued that using an holobiont model could significantly increase predictive accuracy, provided

both **X** and **B** contribute to phenotypic variance, and that a holobiont model was the best default option. Our results (Figure 2) confirm this conjecture, and we observed that **B** can improve prediction by a sizable margin, e.g., for HP. For CRP and PHAGO_LYMPH, the combined model **X+B** increased correlation by ~ 25% compared to models without **B**.

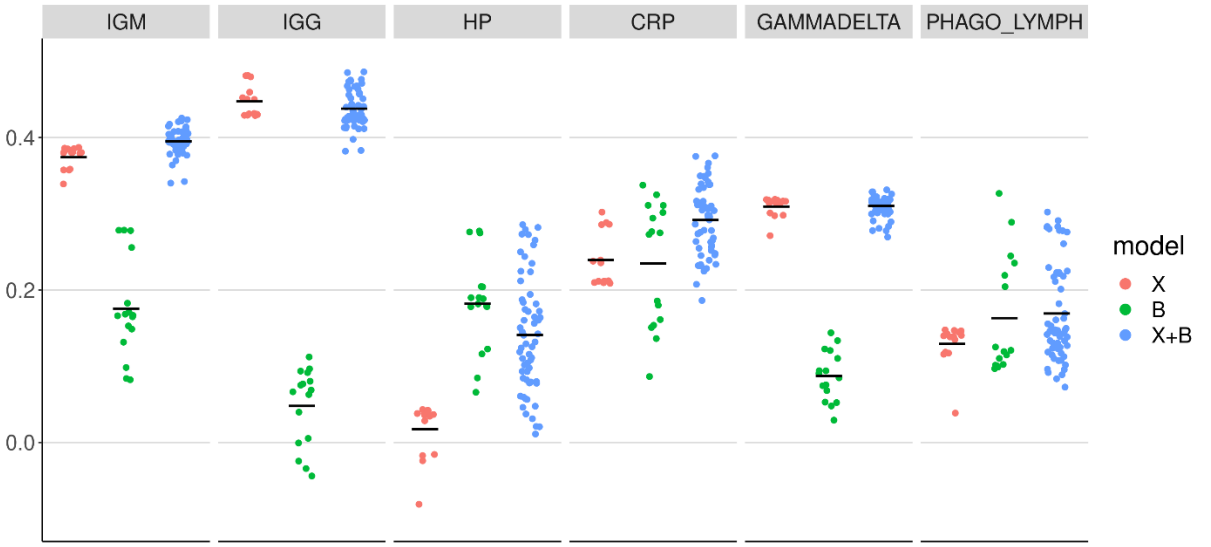

Figure 2. Correlation between predicted and observed phenotypes with genotype (X), microbiota (B) and holobiont (X+B) models. Each dot represents one analysis with different algorithm, prior and microbiota clustering combinations. The horizontal black line is the mean.

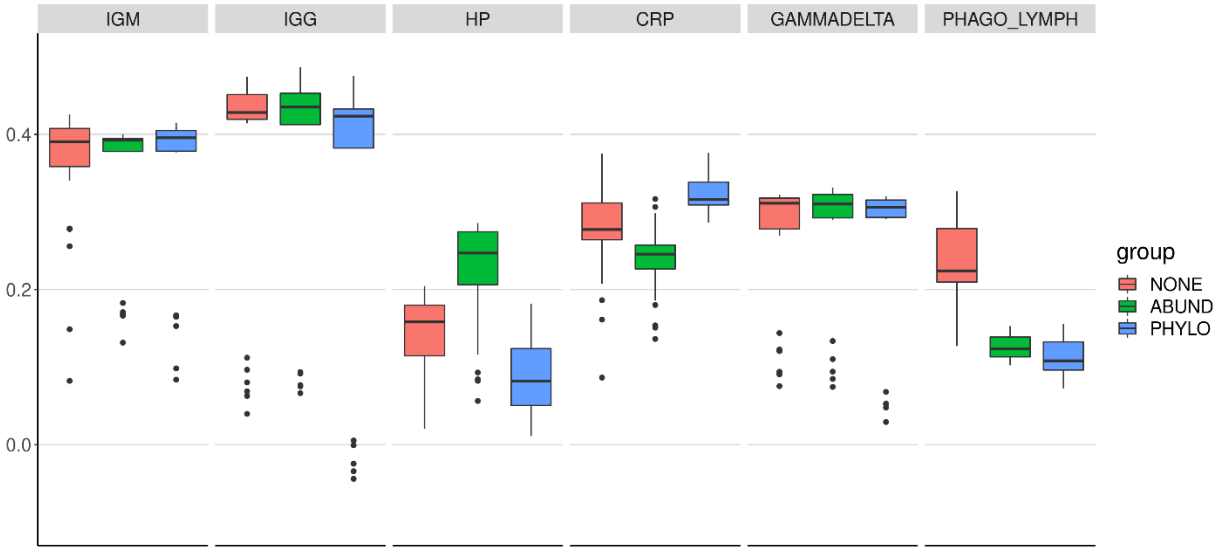

Figure 3. Microbiota clustering effect on prediction. "NONE": no clustering; "PHYLO": clustering at the genus level; "ABUND": clustering by abundance.

Should microbial abundances be grouped? Microbial abundances display highly leptokurtic distributions that require careful quality control, filtering and transformation steps. An option to minimize risks is to group them. Is that useful for predictive purposes? Again, there is not a single best option, although clustering worsens performance for most traits (Figure 3). Clustering by abundance yielded a 50% increase in HP prediction correlation, and clustering by phylogeny increased CRP results by 15%.

Discussion

The term "hologenome" (Zilber-Rosenberg and Rosenberg 2008) was coined to describe the joint action of the genome and of the microbiome on a phenotype. We confirm that microbiota data can improve prediction, sometimes by a large margin, in important phenotypes related to immunocompetence. Here we systematically explored the optimal statistical approaches for prediction, what we have called 'predictome'. In all, we can conclude the following:

1. There does not exist a systematically best prediction method, although our results favor Bayes C. Don't feel culprit though if you choose REML or Bayesian methods.
2. A holobiont model tends to perform better than either genotype or microbiota only models, i.e., use all data at your disposal.
3. By default, we recommend not clustering microbial abundances, despite their highly extreme distributions.

The explosion of statistical methods for holobiont based prediction has neither a clear 'winner' nor a clear 'loser'.

References

Ballester, M., Ramayo-Caldas, Y., González-Rodríguez, O., Pascual, M., , *et al*. (2020). *Scientific Reports*, *10*(1), 1–15. https://doi.org/10.1038/s41598-020-75417-7

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, *et al*. (2019). *Nature Biotechnology*, *37*(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, *et al*. (2016). *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Camarinha-Silva, A., Maushammer, M., Wellmann, R., Vital, M., Preuss, S., *et al*. (2017). *Genetics*, *206*(3), 1637–1644. https://doi.org/10.1534/genetics.117.200782

Difford, G. F., Plichta, D. R., Løvendahl, P., Lassen, J., Noel, S. J., , *et al*. (2018). *PLoS Genetics*, *14*(10), e1007580. https://doi.org/10.1371/journal.pgen.1007580

Gianola, D. (2013). *Genetics*, *194*(3), 573–596. https://doi.org/10.1534/genetics.113.151753

Pérez-Enciso, M., Zingaretti, L. M., Ramayo-Caldas, Y., & de los Campos, G. (2021). *Genetics, Selection, Evolution: GSE*, *53*(1), 1–20. https://doi.org/10.1186/s12711-021-00658-7

Pérez, P., & de los Campos, G. (2014). *Genetics*, *198*(2), 483–495. https://doi.org/10.1534/genetics.114.164442

Ramayo-Caldas, Y., Zingaretti, L. M., Pérez-Pascual, D., Alexandre, P. A., Reverter, A., *et al*. (2021). *Animal Microbiome*, *3*(1), 1–11. https://doi.org/10.1186/s42523-021-00138-9

Zhang, Q., Difford, G., Sahana, G., Løvendahl, P., Lassen, J., *et al*. (2020). *The ISME Journal*, *14*(8), 2019–2033. https://doi.org/10.1038/s41396-020-0663-x

Zilber-Rosenberg, I., & Rosenberg, E. (2008). *FEMS Microbiology Reviews*, *32*(5), 723–735. https://doi.org/10.1111/j.1574-6976.2008.00123.x