Supplementary Information

Accurate and robust classification of *Mycobacterium bovis*-infected cattle using peripheral blood RNA-seq data

John F. O'Grady[1], Adriana Ivich[2], Gillian P. McHugo[1], Adnan Khan[1], Thomas J. Hall[1], Sarah L. F. O'Donnell[1,10], Carolina N. Correia[1,11], John A. Browne[1], Valentina Riggio[3,4], James G. D. Prendergast[3,4], Emily L. Clark[3,4], Hubert Pausch[5], Kieran G. Meade[1,6,7], Isobel C. Gormley[8], Eamonn Gormley[7,9], Stephen V. Gordon[6,7,9], Casey S. Greene[2] and David E. MacHugh[1,6,7*]

[1] UCD School of Agriculture and Food Science, University College Dublin, Belfield, Dublin, D04 V1W8, Ireland.

[2] Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

[3] The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Midlothian, EH25 9RG, UK.

[4] Centre for Tropical Livestock Genetics and Health (CTLGH), Roslin Institute, University of Edinburgh, Easter Bush Campus, EH25 9RG, UK.

[5] Animal Genomics, ETH Zurich, Universitaetstrasse 2, 8006, Zurich, Switzerland.

[6] UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield, Dublin, D04 V1W8, Ireland.

[7] UCD One Health Centre, University College Dublin, Belfield, Dublin, D04 V1W8, Ireland.

[8] UCD School of Mathematics and Statistics, University College Dublin, Belfield, Dublin, D04 V1W8, Ireland.

[9] UCD School of Veterinary Medicine, University College Dublin, Belfield, Dublin, D04 V1W8, Ireland.

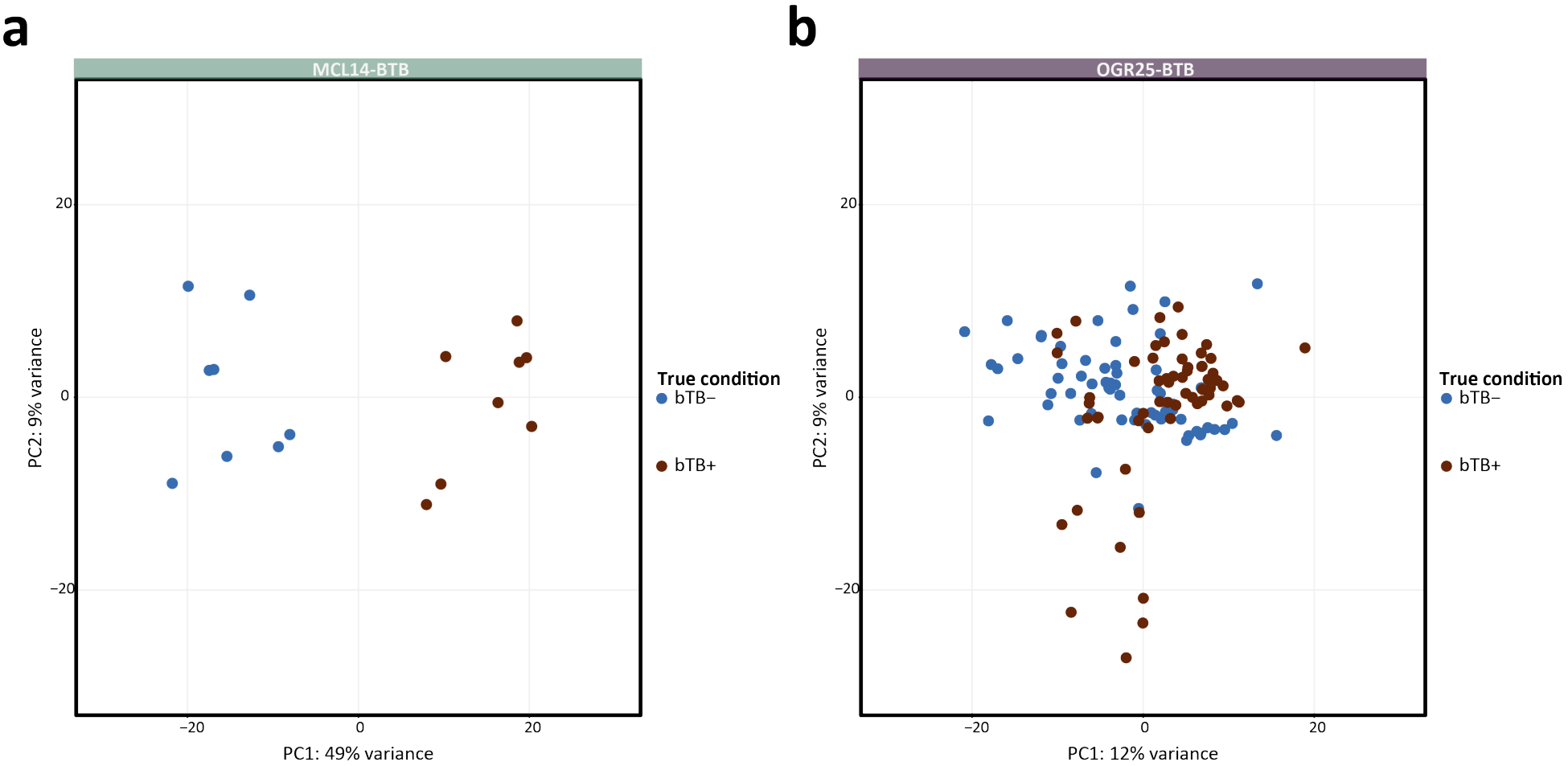
[10] Present address: Irish Blood Transfusion Service, National Blood Centre, James's Street, Dublin, D08 NH5R, Ireland.

[11] Present address: Children's Health Ireland, 32 James's Walk, Rialto, Dublin, D08 HP97, Ireland.

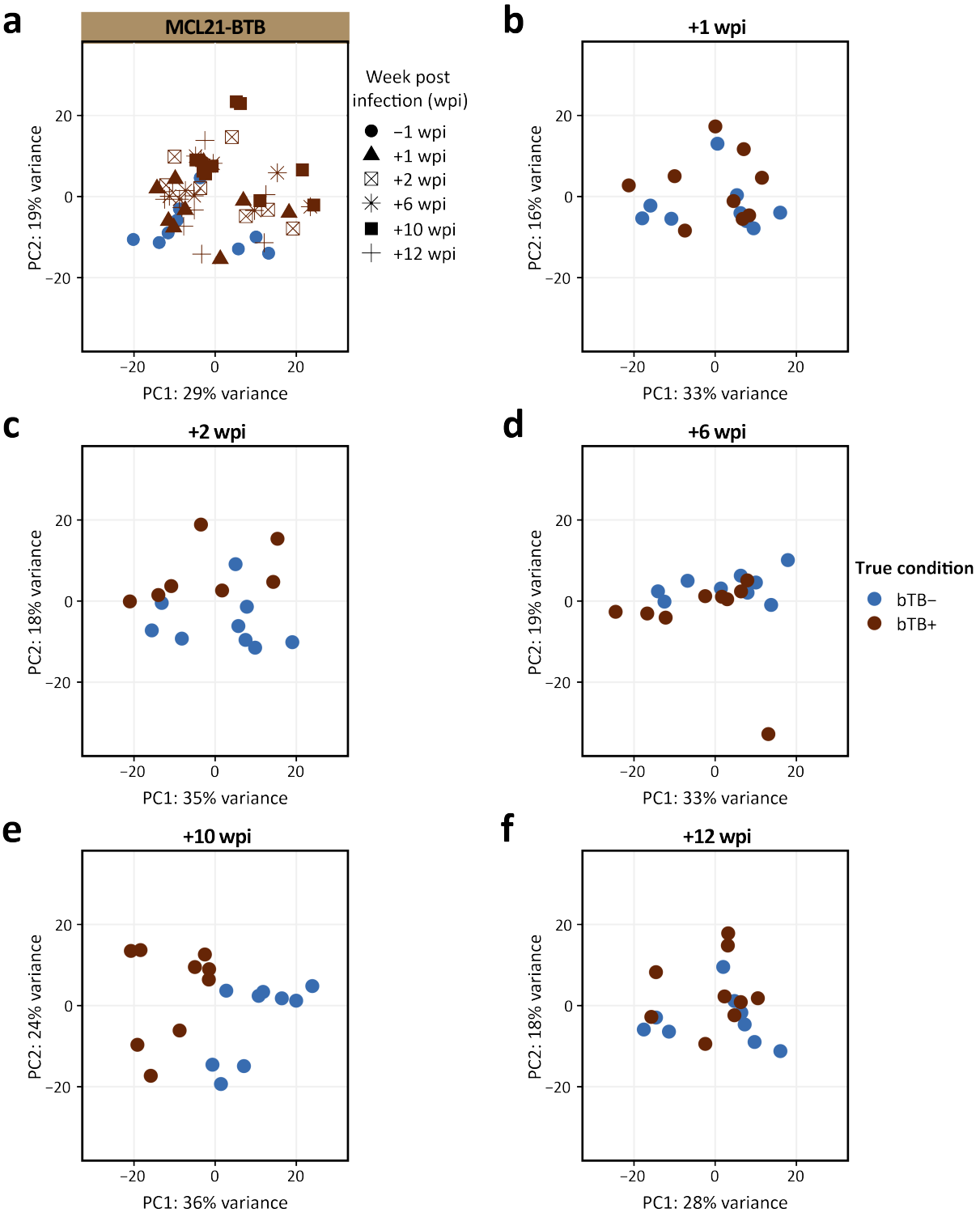
[*] Corresponding author

E-mail address: david.machugh@ucd.ie

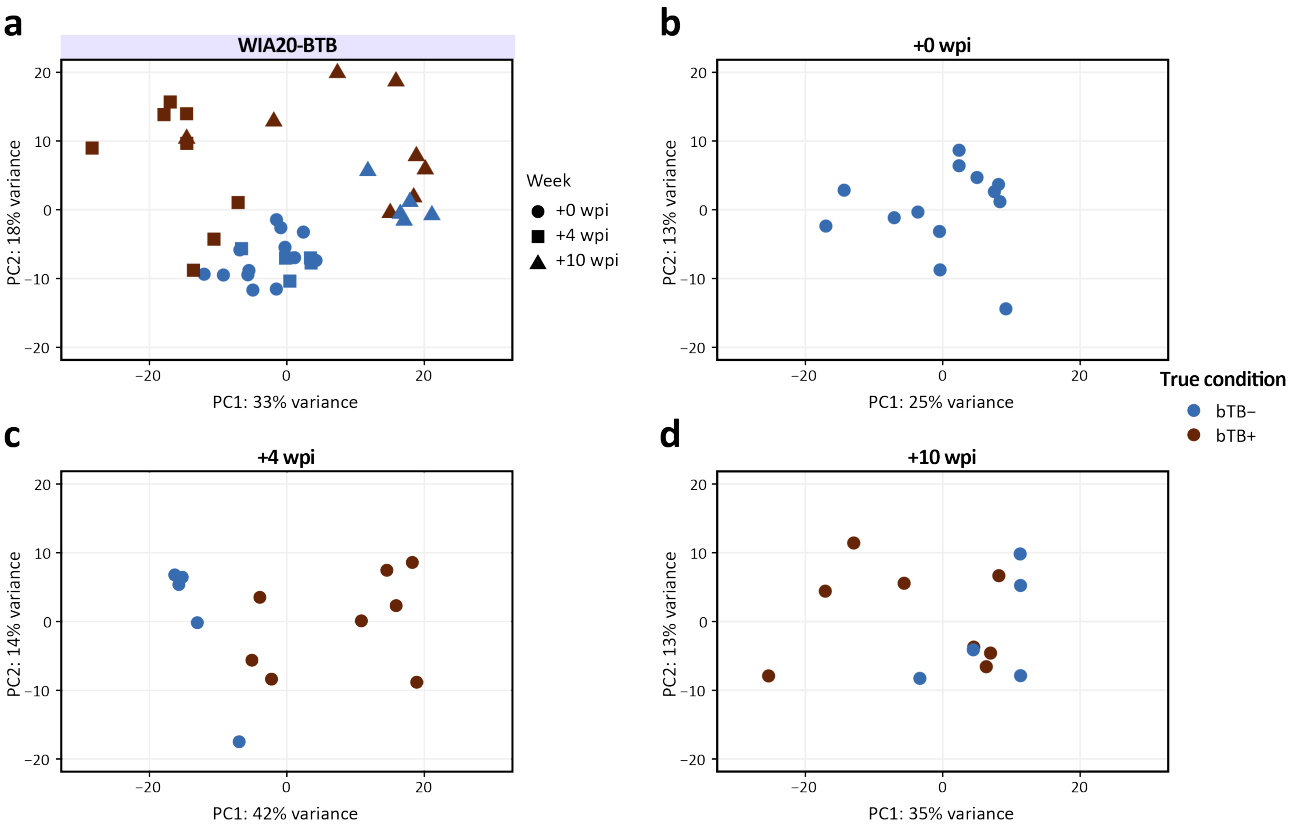
Supplementary Figures



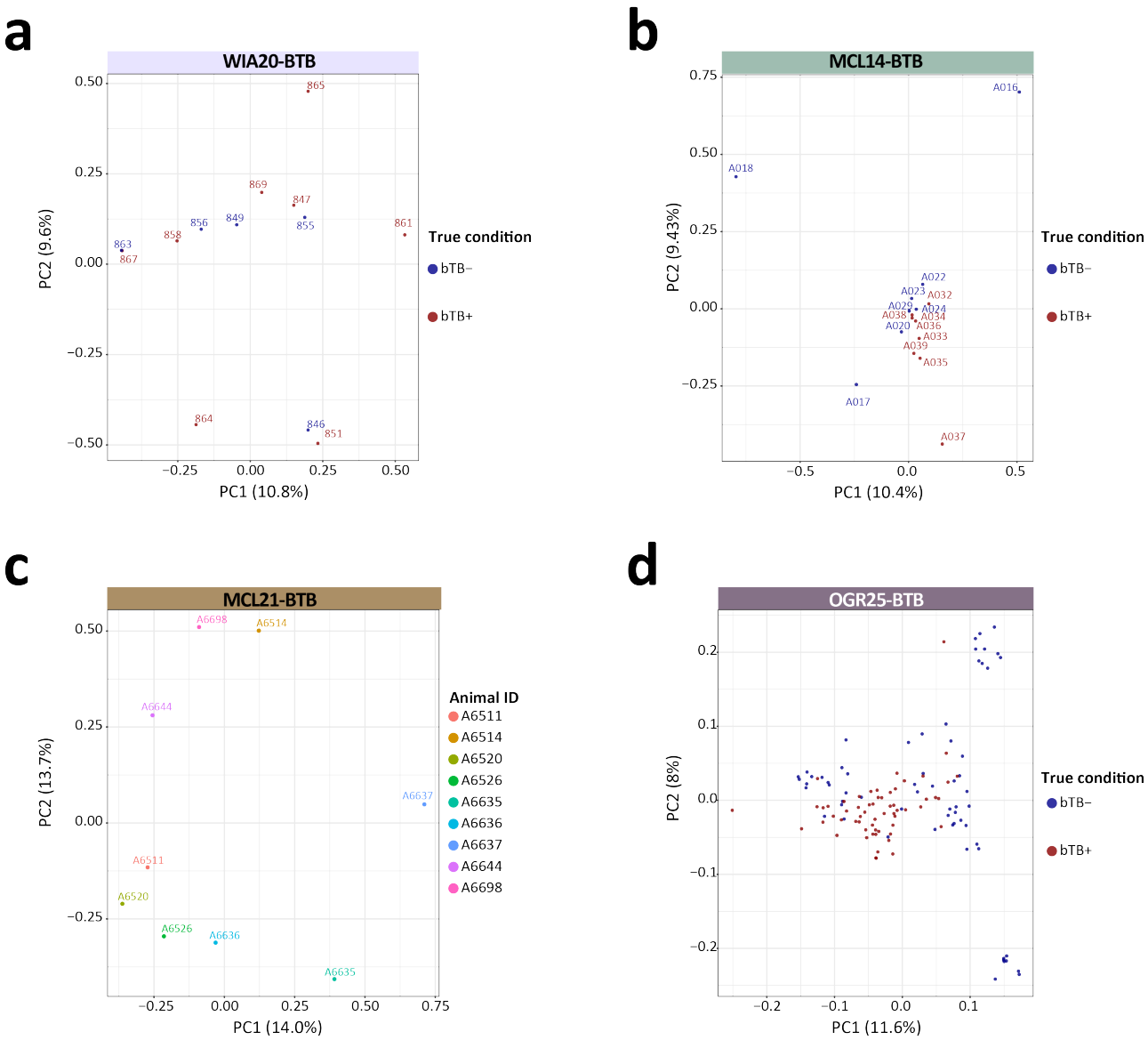
Supplementary Figure S4.1: Principal component analysis (PCA) of the top 750 most variable genes identified in (**a**) all *n* = 16 RNA-seq samples from the MCL14-BTB dataset and (**b**) all *n* = 123 RNA-seq samples from the OGR25-BTB dataset after variance stabilising transformation (VST) using DESeq2 (Love *et al.* 2014). Principal components 1 and 2 (PC1 and PC2) are plotted. Animal data points are shaped based on their experimental condition.



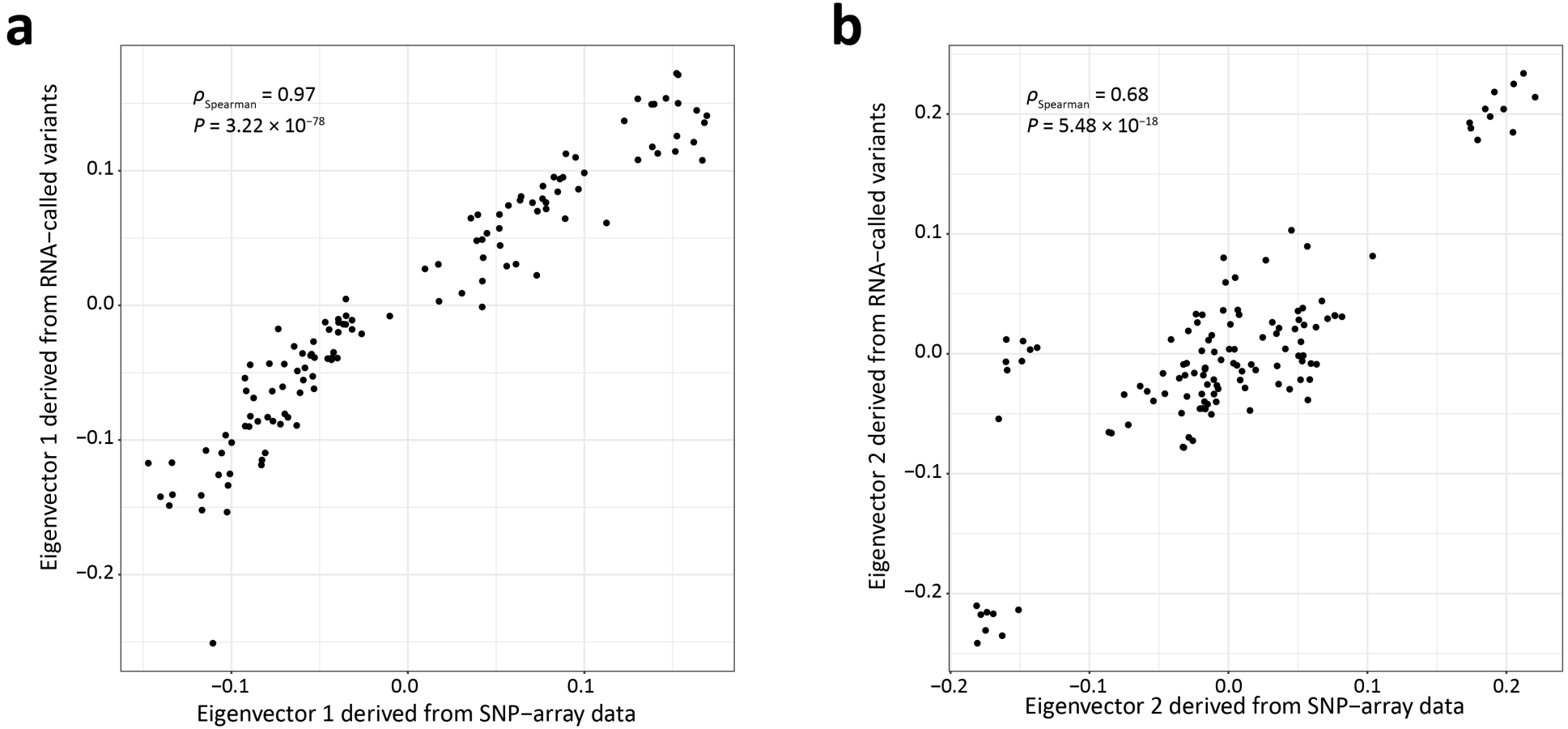
Supplementary Figure S4.2: Principal component analysis (PCA) of the top 750 most variable genes identified for (**a**) all *n* = 52 RNA-seq samples from the MCL21-BTB dataset after variance stabilising transformation (VST) using DESeq2. Principal components 1 and 2 (PC1 and PC2) are plotted. Animal data points are coloured based on their experimental condition and shaped based on their sampling time point, measured in terms of weeks post-infection (wpi). (**b**) PCA of animals sampled at +1 wpi compared to animals sampled at −1 wpi. (**c**) PCA of animals sampled at +2 wpi compared to animals sampled at −1 wpi. (**d**) PCA of animals sampled at +6 wpi compared to animals sampled at −1 wpi. (**e**) PCA of animals sampled at +10 wpi compared to animals sampled at −1 wpi. (**f**) PCA of animals sampled at +12 wpi compared to animals sampled at −1 wpi.



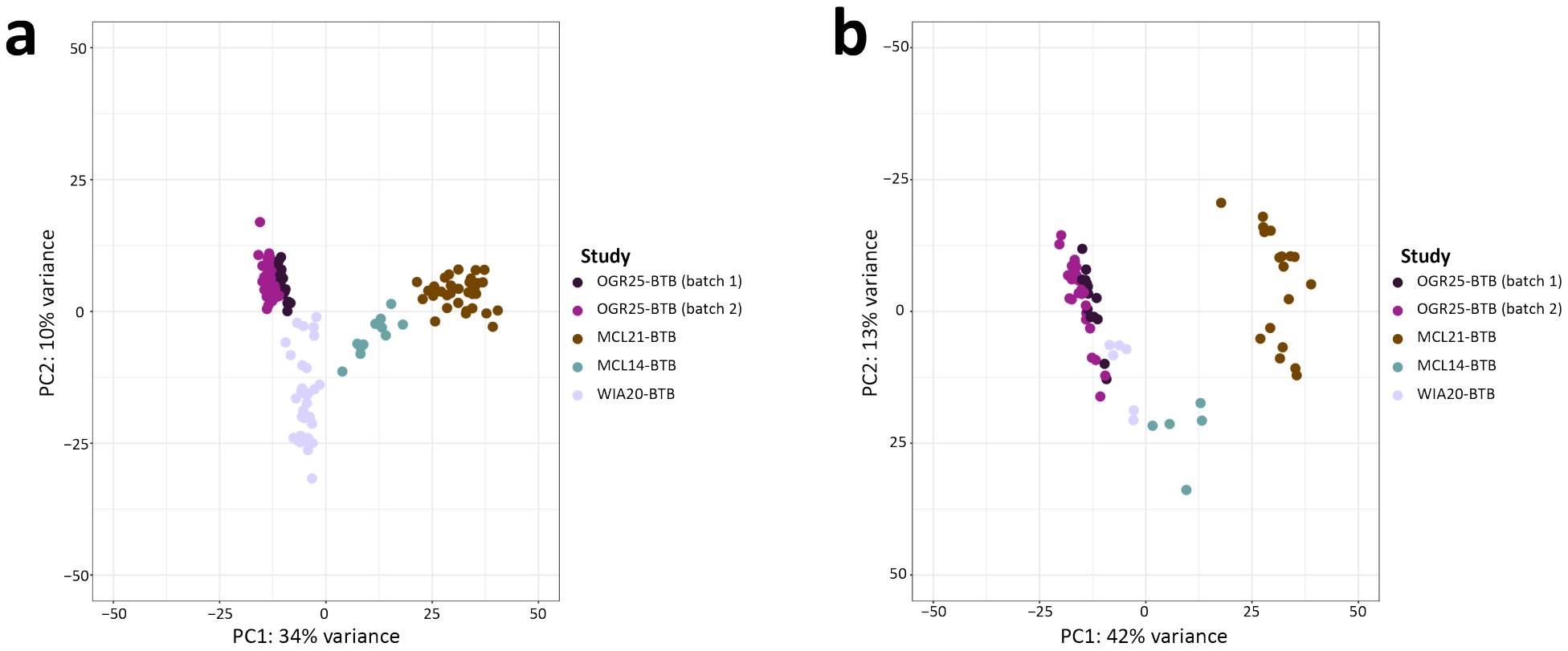
Supplementary Figure S4.3: Principal component analysis (PCA) of the top 750 most variable genes identified for (**a**) all *n* = 39 RNA-seq samples from the WIA20-BTB dataset after variance stabilising transformation (VST) using DESeq2. Principal components 1 and 2 (PC1 and PC2) are plotted. Animal data points are coloured based on their experimental condition and shaped based on their sampling time point, measured in terms of weeks post-infection (wpi). (**b**) PCA of animals sampled at +0 wpi. Note: all *n* = 8 animals experimentally infected with *M. bovis* were sampled prior to inoculation and are therefore considered control (bTB−) animals at this time point. (**c**) PCA of animals sampled at +4 wpi compared to animals sampled at 0 wpi. (**d**) PCA of animals sampled at +10 wpi compared to animals sampled at 0 wpi.



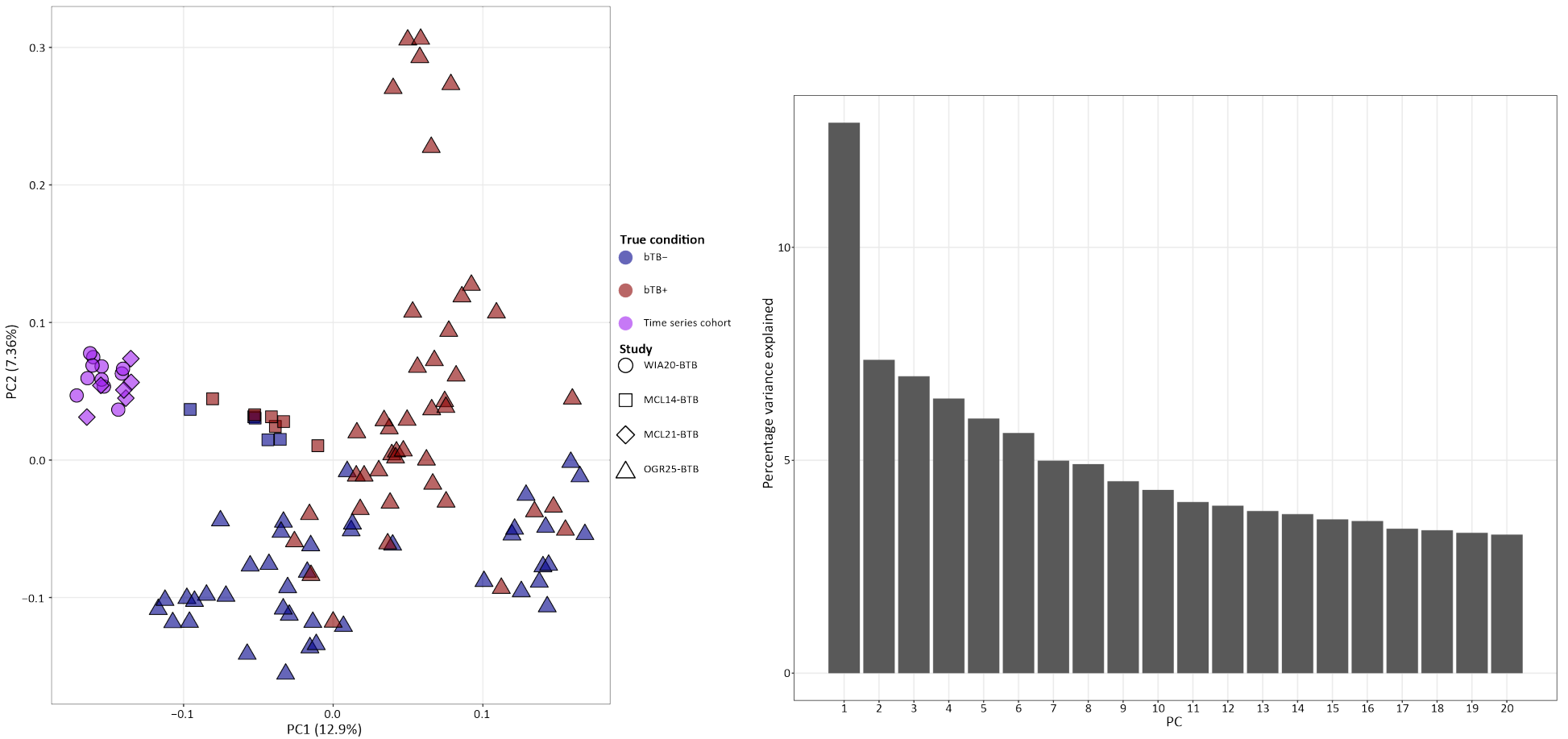
Supplementary Figure S4.4: (**a**) Genotype principal component analysis (PCA) of 29,740 genome-wide SNPs called from the RNA-seq data in the WIA20-BTB dataset. Animals are coloured based on whether they were experimentally infected with *M. bovis* or not. (**b**) Genotype PCA of 29,740 genome-wide SNPs called from the RNA-seq data in the MCL14-BTB dataset. Animals are coloured based on their experimental designation. (**c**) Genotype PCA of 29,740 genome-wide SNPs called from the RNA-seq data in the MCL21-BTB dataset. Animals are coloured based on their designated animal ID. (**d**) Genotype PCA of 29,740 genome-wide SNPs called from the OGR25-BTB dataset. Animals are coloured based on their experimental designation. For all panels, principal components 1 and 2 (PC1 and PC2) are plotted.



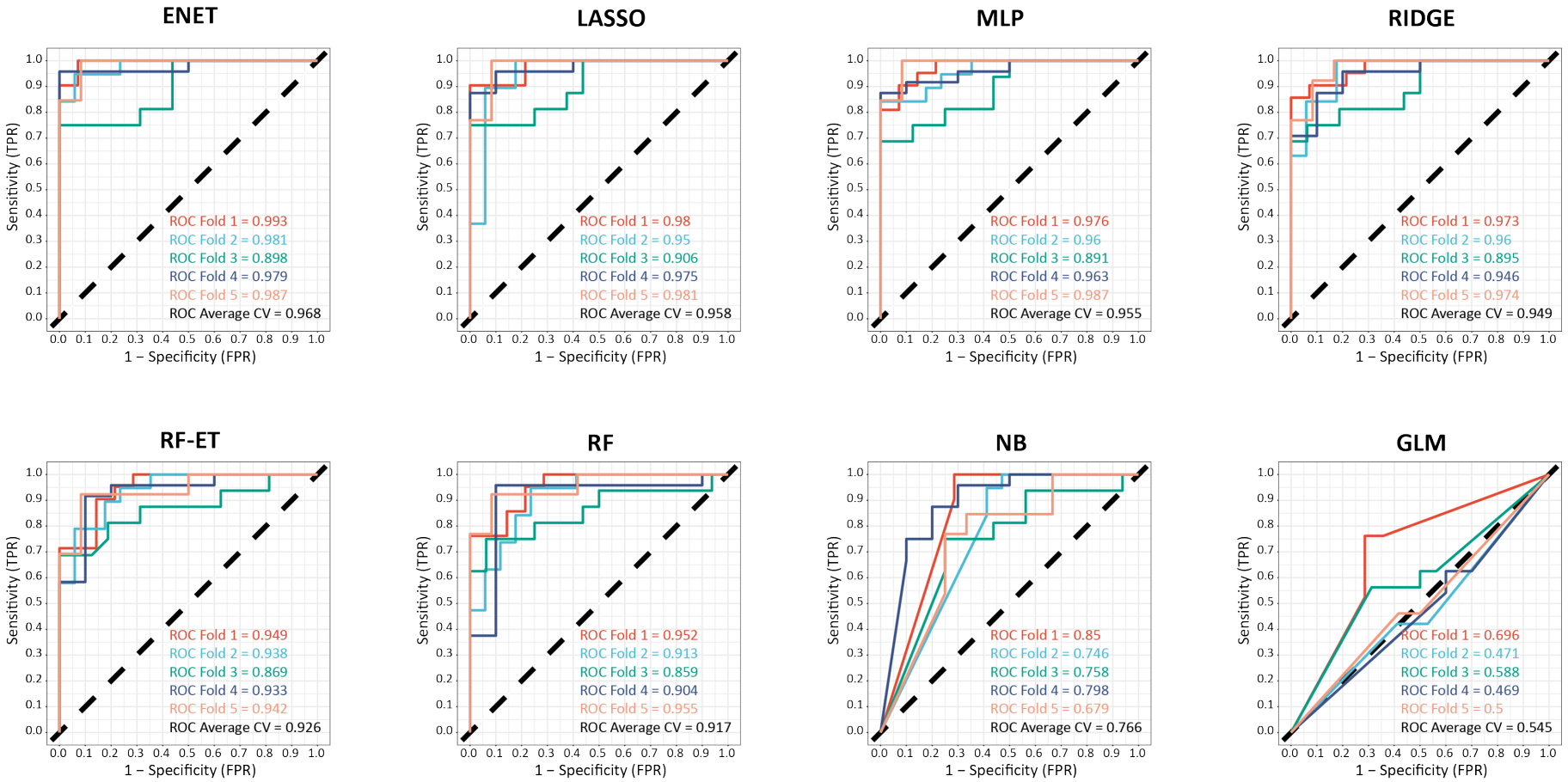
Supplementary Figure S4.5: (**a**) Spearman correlation (ρ) between Eigenvector 1 coordinates for all n = 123 animals from the OGR25-BTB dataset derived from a principal component analysis (PCA) of 29,740 genome-wide SNPs and Eigenvector 1 derived from a PCA of 34,272 pruned genome-wide SNP array data reported for the same set of animals in O'Grady *et al.* (2025) . (**b**) Same as **a** but for Eigenvector 2 coordinates.



Supplementary Figure S4.6: (**a**) Principal component analysis (PCA) of the top 750 most variable genes identified in the training set after variance stabilising transformation (VST) using DESeq2. Principal components 1 and 2 (PC1 and PC2) are plotted. Animals are coloured based on their designated study. (**b**) A PCA of the top 750 most variable genes identified in the testing set after VST using DESeq2. PC1 and PC2 are plotted, and animals are coloured based on their designated study.



Supplementary Figure S4.7: Genotype principal component analysis (PCA) of 29,740 genome-wide SNPs identified in the training set. Principal components 1 and 2 (PC1 and PC2) are plotted. Animals are coloured depending on their experimental designation or if they are derived from a time series experiment. Animals are shaped based on their designated study. The barplot depicts the proportion of variance explained by each PC out of the top 20 PCs.



Supplementary Figure S4.8: Area under the receiver operating characteristic curve (AUROC) plots for eight machine learning (ML) models in each cross-validation fold in the training set. ENET, elastic-net; MLP, multi-layered perceptron; LASSO, least absolute shrinkage and selection operator; RIDGE, ridge-penalised regression; RF-ET, random forest (extra trees); RF, random forest; NB, naïve Bayes; and GLM, generalised unpenalized logistic regression model.

References

Love M.I., Huber W. & Anders S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 15, 550.

O'Grady J.F., McHugo G.P., Ward J.A., Hall T.J., Faherty O'Donnell S.L., Correia C.N., Browne J.A., McDonald M., Gormley E., Riggio V., Prendergast J.G.D., Clark E.L., Pausch H., Meade K.G., Gormley I.C., Gordon S.V. & MacHugh D.E. (2025) Integrative genomics sheds light on the immunogenetics of tuberculosis in cattle. *Communications Biology* 8, 479.