# Supplementary Methods, Section 2

# Data analysis in detail

To build the *gene model* and *pathway model* the batch 1 data was bootstrapped 100 times. The bootstrap samples were of the same size as the batch 1 data and sampled with replacement. For each bootstrap sample a RF model was built using 5 fold cross validation to tune the model parameter mtry [1]. mtry is the number of variables randomly selected for consideration at each split in the decision tree. The model parameter that determined the number of trees created within each model ntree was set to 501 throughout; an odd ntree was used to account for any ties. The change in Gini index was used to create variable importance scores. These were ranked across all variables per model and then summed across all bootstrap samples. Variables were ordered by this metric and plotted. The variable importance plot plateaued after approximately 5% of variables and consequently the top 5% of variables were taken forward to the next stage (See Supplementary Figures 1 and 2).

Recursive feature elimination (RFE) was performed on this subset of variables in the original batch 1 data for the *pathway model* and *gene model*. Building of the *demographic model* began at this stage using all variables. The feature elimination was again based on a RF model with ntree = 501 and investigated subsets of variables of all possible sizes. Caret’s ‘pickSizeTolerance’ function was used (tolerance = 5%) to identify a further subset of variables. This function finds a smaller set of variables while maintaining model accuracy [2]. If this subset matched the RFE optimal set, the RFE model was taken forward. Otherwise, the optimum variables were selected using the ‘selectVar’ function in caret and a final RF model was built.

RF models were used throughout for their non-parametric, non-linear properties. Further, the use of bootstrapping in RF modeling and random selection of variables at each decision point decreases the dependence of these models on noise.

# References

[1] Liaw A, Wiener M (2002) Classification and Regression by randomForest. *R News* **2**, 18–22.

[2] Kuhn JM. Contributions from Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z and the R Core Team (2014) *caret: Classification and Regression Training*. R package version 6.0-35.