Effect of biomarker identification on power analysis for diagnostics research

Animesh Acharjee (a.acharjee@bham.ac.uk)
University of Birmingham https://orcid.org/0000-0003-2735-7010
Joseph Larkman
University of Birmingham College of Medical and Dental Sciences
Victor Roth Cardoso
University of Birmingham College of Medical and Dental Sciences
Georgios V. Gkoutos
University of Birmingham College of Medical and Dental Sciences

Research

Keywords: Random forest, feature selection, power study, biomarker

Posted Date: April 1st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-19527/v1

Title: Effect of biomarker identification on power analysis for diagnostics research

Animesh Acharjee[1,2,3,*,#], Joseph Larkman[1,2,*], Victor Roth Cardoso[1,2], Georgios V. Gkoutos[1-6]

[1]College of Medical and Dental Sciences, Institute of Cancer and Genomic Sciences, Centre for Computational Biology, University of Birmingham, B15 2TT, UK

[2]Institute of Translational Medicine, University Hospitals Birmingham NHS, Foundation Trust, B15 2TT, UK

[3]NIHR Surgical Reconstruction and Microbiology Research Centre, University Hospital Birmingham, Birmingham B15 2WB, UK.

[4]MRC Health Data Research UK (HDR UK)

[5]NIHR Experimental Cancer Medicine Centre, B15 2TT, Birmingham, UK

[6]NIHR Biomedical Research Centre, University Hospital Birmingham, Birmingham, B15 2TT, UK.

*** Corresponding Author.**

Animesh Acharjee

Tel.: +44 (0)1213718135

E-mail: a.acharjee@bham.ac.uk

ABSTRACT

Background

Biomarker identification is one of the major and important goal of the functional genomics and translational medicine remits. Large scale –omics data are increasing being accumulated and can provide vital means for the identification of biomarkers for the early diagnosis of complex disease and/or patient/diseases stratification for prospective studies. These tasks are clearly interlinked and it is essential that an unbiased and stable methodology is applied in order to address them. Although, recently, many, primarily machine learning based, biomarker identification approaches have been developed, the exploration of potential associations between biomarker identification and the design of future experiments remains a challenge.

Methods

In this study, using both simulated and published experimentally derived (real) datasets. We compared the performance of decision based machine learning approach called Random Forest. Four Random forest based feature selection methods namely, Boruta, Permutation based feature selection without correction, Permutation based feature selection with correction, Backward elimination based feature selection. Moreover, we conducted power analysis to estimate the number of samples required for potential future studies using the derived stable from the previous step.

Results

We presented a number of different RF based stable feature selection methods and compared their performances using simulated as well as published experimentally derived datasets. Across all of the scenarios considered, we found Boruta to be the most stable methodology, whilst Permutation (Raw) offered the largest number of relevant features when allowed to stabilise over a number of iterations. Finally, we developed a web interface (https://joelarkman.shinyapps.io/PowerTools/) to streamline power calculations and aid future study design within a translational medicine context.

Conclusions

We developed a pipeline to discover biomarkers using RF methods. The web interface, “PowerTools” offers the potential for designing appropriate and cost-effective subsequent future omics study designs.

Key words: Random forest, feature selection, power study, biomarker

Introduction

Over the last few years there has been lots of emphasis on the high dimensional omics data generation that includes untargeted -omics datasets like transcriptomics [1,2] metabolomics [3,4], proteomics [5,6], microbiomes [7–9], as well as deep phenotyping [10]. As a consequence, a massive amount of data is routinely accumulated which needs to be integrated and analysed so as to facilitate the identification of the relevant markers. If the markers from -omics datasets are robust, reproducible and indicative then they can be used as a biomarker for patient's stratification [11,12] and can also be useful either as diagnostics or prognostic tools. Selecting relevant markers or features from a high dimensional dataset is defined as feature or variable selection [13] and requires a robust statistical or computational workflow [14].

In the literature there has been lots of interest in the application of machine learning methods on omics datasets for feature selection [14]. Lately, decision tree based statistical machine learning methods, for example, Random Forest (RF) [15], have gained prominence. RF is an ensemble learning method which has been applied successfully on multiple high dimensional omics studies that includes transcriptomics [16], metabolomics [17], methylation [18] and proteomics [19]. The objective of these studies has been either the prediction or the selection of important features that serve as potential biomarkers and that can be potentially employed for patient stratification in future studies. The random forest algorithm is a powerful prediction method that is known to be able to capture complex dependency patterns between the outcome and the covariates. The latter feature makes random forest a promising candidate for developing a prediction method tailored to the challenges of multi-omics data.

In the case of omics analysis, it is important to have feature selection procedures that are more systematic and data driven so to avoid any bias selection. In the literature there is limited evidence that RF can also be employed for this tasks when coupled with other feature selection methods, for example: selecting genes and metabolites [20]; selection of lipids, metabolites [21]. More recently, Degenhardt et al. (2019) performed a simulation study, and also applied on published experimentally derived (real) datasets, different versions of RF methods. However, in that instance, the markers that were identified, were not satisfactorily related to future translational research. That is, the effect/weight of the identified markers was not assessed and validated for their use in similar future studies via 'study design' or 'power analysis'.

There are some earlier studies that have focused on power analysis, for example relating to metabolomics data [22,23] and transcriptomics data [24]. However, those studies are very specific to certain -omics datasets and often failed to properly relate power calculations to putative biomarkers identified via stable feature selection procedures.

In this study, we divided our objectives into two main tasks. For the first module, we performed extensive simulation with RF based feature selection methods such as Boruta [25], permutation based feature selection [26], permutation based feature selection with correction [26], and backward elimination based feature selection [27], both in a regression and classification context, to understand their feature selection capabilities and prediction error. We then conducted similar investigations using experimentally derived data, to examine how the methods perform when dealing with disparate -omics data paradigms. For the second module, we developed a workflow to identify the number of samples required for a future study using the stable biomarkers that were identified in the first task. Finally, we developed a web interface to streamline power calculations and offer a valuable asset for use in translational research, going forward.

Materials and methods

Random Forest (RF)

Random Forest (RF) [15] is an ensemble-based machine learning approach that can handle nonlinear relationships between response and predictor variables and both classification and regression based analysis. Typically, around two-thirds of the samples are used for model training, with the remaining third being held out as the out-of-bag (OOB) samples and used to assess the model's performance. For classification, prediction performance can be quantified in terms of the rate at which OOB samples are misclassified, or the OOB error. For regression, the average distance between OOB predictions and the true continuous response variable can be quantified using the mean squared error (MSE) metric. The contribution from each variable to the final model is quantified as a ranked measure of variable importance (More information in the Additional file 1 methods section).

RF feature selection methods

Four methodologies were utilised for the automatic selection of important features from the aforementioned ranked list RF generates, namely, Boruta [25], permutation based feature selection [26], permutation based feature selection with correction [26], and backward elimination based feature selection [27]. Details of the statistical basis for each of these approaches is outlined in the following section.

Boruta

Boruta [25] compares the feature importance values estimated for the real predictor variables, with variables generated by permutation of the real variables across observations. Variables generated by permutation are termed "shadow" variables. For each run, a RF is trained using a double length set of predictor variables comprised of an equal number of true and shadow variables. For each of the real predictor variables, a statistical test is conducted comparing its importance with that of the maximum importance value achieved by a shadow variable. Variables with significantly larger or smaller importance values are declared by the algorithm as important or unimportant, respectively. In subsequent runs, all unimportant and shadow variables are removed, and the process is repeated until all variables have been classified or a specified maximum number of runs have taken place.

Permutation based feature selection

Standard permutation testing is an established approach for approximating a significance level or threshold for the selection of a subset of associated features from a RF model [21,26]. The iterative nature of this approach ultimately establishes a null distribution of importance values expected to be observed by chance alone. Using this null distribution, features are determined to be important only if the probability of achieving their observed importance value or one more extreme (p-value), is below a specified threshold. Permutation testing was implemented using both raw (uncorrected) p-values as well corrected ones for multiple hypothesis testing via the Benjamini-Hochberg procedure (corrected) [28]. For both implementations, a threshold p-value of 0.05 was used to determine statistical significance.

Backward elimination based feature selection

Recursive feature elimination (RFE) [27] is an approach which endeavours to determine the smallest subset of variables that produce an effective model with good prediction accuracy. The methodology involves the iterative fitting of RF, where upon each iteration, a specified

proportion of the variables with smallest variable importance are discarded. This process is applied recursively until only a single variable remains available for input. At each iteration, model performance is assessed in terms of out-of-bag error, when RF is used in a classification capacity, or mean squared error (MSE) for regression forests. The set of variables leading to the production of the RF with the smallest error, or within a certain range of the minimum, are ultimately selected. This procedure was facilitated for classification forests using the base implementation of the method in the R package varSelRF [27]. Additionally, the constituent functions of varSelRF were modified by the present study to accept a continuous y variable input and to use MSE for model assessment, to facilitate feature selection when RF is used in a regression capacity

Statistical analysis strategy

Module 1 - Stable feature selection

We divided our analysis methodology into two modules. In module one, we incorporated a double cross validation (CV) procedure [29–32], which is summarised in Figure 1 (Module 1)**.** First, a training:testing ratio of 75:25 is used to generate outer train and test data subsets, respectively. The outer train subset is then subject to a further tenfold CV, where one-tenth of the data (inner train) is used for hyper parameter optimisation (additional information is provided in the Additional file 1 methods). The remaining nine-tenths are considered the inner test subset which are then used to complete 100 iterations of feature selection, where upon each iteration, a RF is trained and each of the aforementioned methods identify a subset of important features. Stable features are then defined as those identified by a particular method in a number of iterations greater than a specified stringency value. In the present study, features selected in >5/100 iterations are determined low stringency (LS) stable features and those selected in at least 90/100 iterations, high stringency (HS) stable features. Stable features are subset from the outer test data and used to quantify properly the predictive power (R^2) and prediction error (OOB/MSE) of the model using independent data. The entire procedure is repeated four times, each with a modified outer loop split, such that each sample appears once within the outer test data subset. The values specifying predictive power, prediction error, and the frequency with which each feature is selected by each method, are then averaged across the four outer loop repeats. The full double CV procedure was applied to simulated data and three published

experimentally derived (real) data cohorts. Due to sample size limitations, CV was not conducted for published experimentally derived classification data 2.

Module 2 – Power Analysis

We implemented a flexible approach to facilitate power analysis and sample size determination, based on functions designed and described by Blaise et al., (2016) [23]. We included both datasets with a continuous outcome variable (regression) and those with a two-group (binary) classification outcome. Furthermore, the correlation structure of input data was explicitly modelled, in order to capture any multicollinearity between variables. An overview of the approach is represented schematically in Figure 1 (Module 2) and described briefly in the following section. (More information in the Additional file 1 methods section).

In summary, data is first simulated using a multivariate log-normal distribution then iteratively subset to produce data with a specified series of sample sizes. In the case of regression, for each of the variables assessed, a continuous outcome is generated that relates to the assessed variable with a Pearson correlation value equal to that of its relation to the real outcome (effect size). The simulated outcome variable is then regressed against each variable to produce a set of p-values describing each variable's association with the outcome.

For the two-group (binary) classification case, two datasets are produced for each specified sample size (one for each group) and a specified Cohens d effect size [33] is introduced to one of them for the currently iterated variable and its highly correlated partners. A one-way ANOVA is then conducted for each variable, comparing the intra and inter group variances, and producing a set of p-values describing the variables with statistically significant variances.

In either case, true/false positive metrics are then determined by comparing the set of statistically significant variables to a set containing the variable chosen for analysis and its highly correlated partners.

Datasets

We used simulated data and published experimentally derived datasets to understand methods and their performance in both a classification and regression context. Please refer to the Table 1 for detail data information.

Simulated data

Simulation data featuring correlated predictor variables and a quantitative outcome variable were generated using a nonlinear regression model, as previously reported by the reference literature [26,34]. Specifically, the simulation strategy and equations outlined by Degenhardt et al., (2019) [26] were adapted and implemented in the present study, with minor parameter modifications. A single simulation scheme was incorporated in which sixty (six groups of ten) correlated variables were generated, alongside additional uncorrelated variables, to produce a dataset of 5000 predictor variables.

First, six uniformly distributed variables, x_1, x_2, x_3, x_4, x_5 and x_6, were sampled individually from $U(0,1)$. The correlated predictor variables were then generated according to the equation:

$$V_i^{(j)} = x_i + (0.01 + \frac{0.5(j-1)}{n-1}) \cdot N(0,0.3)$$

for $j = 1,\ldots,10$ and $i = 1,\ldots,6$, where $V_i^{(j)}$ denotes the jth variable in group i and the correlation between $V_i^{(j)}$ and x decreases as j increases. Conversely, values for the uncorrelated predictor variables were simply sampled from the uniform distribution $U(0,1)$.

The variables $x_1, x_2$and x_3 were also used to generate the quantitative outcome variable y via the equation:

$$y = 0.25exp\ (4x_1) + \frac{4}{1 + exp(-20(x_2 - 0.5))} + 3x_3 + N(0,0.2)$$

where y correlates decreasingly with variables $x_1, x_2$and x_3.

The simulation scheme outlined above was repeated iteratively to produce a final simulation dataset of 200 observations, 5000 predictor variables and a quantitative outcome variable correlated with only the first three groups of correlated predictor variables. Additionally, the quantitative outcome variable was adapted for a binary classification context, whereby observations with a y value below the mean of the outcome variable set were assigned to group

one, and those with a value above the mean assigned to group two. Consequently, the same simulation protocol was used to facilitate the assessment of feature selection and power analysis in both a classification and regression context.

Published experimentally derived (real) datasets

A summary of the published experimentally derived datasets is presented in Table 1. For more detailed methods please refer to method section of the respective articles.

Table 1: List of the published datasets used in this study. In each of the RF modes two datasets was considered and compared with published results.

Mode RF method	*Article reference*	*Pubmed ID*	*Number of samples (N)*	*Number of features (p)*	*Outcome variable*
Regression	[21]	*28185575*	*73*	*196*	*Relative liver weight*
	[35]	*28190990*	*40*	*219*	*3 months infant milk amount*
Classification	[21]	*28185575*	*73*	*196*	*Relative liver weight class (below or above the mean value)*
	[36]	*27176004*	*68*	*414*	*Colorectal cancer (CRC) stages*

Software and Code Availability

We used R (https://www.r-project.org) v3.5.0 software for statistical computing. Different packages were used for RF methods and are listed in Table 2. The web interphase was made using R shiny app (http://shiny.rstudio.com/) and is available at: https://joelarkman.shinyapps.io/PowerTools/. All other scripts are provided in the Github: https://github.com/joelarkman/RF-FeatureSelection-PowerAnalysis.

Table 2: List of the methods and R packages used for this study.

Method	**R package used**	**References**
Random Forest	randomForest	[15]

Random Forest (Optimised for memory)	ranger	[37]
Boruta	Boruta	[25]
Hyperparameter selection	Caret	[38]
Permutation based feature selection	pomona	[26]
Recursive feature elimination (RFE)	vaSelRF	[27]

Results

Regression Mode

Simulation Data

We performed regression considering the complex relations created within the 5000 predictor variables and continuous response (y). Figure 2(A) illustrates the relationships for the first 120 features as a correlation plot, showing the intra-group correlation for each of the predictor variables as well as the correlation between each feature and outcome variable (y). As designed, the first sixty variables formed six highly correlated clusters, whilst the remaining features showed no pattern of correlation between variables. The first three correlated groups exhibited a clear correlation with the outcome variable that decreased in intensity both within and between correlated groups, from V1-V30. All other predictor variables exhibited a negligible correlation with the continuous outcome variable.

Results from RF in regression mode on simulated data are summarised in Figure 2(B) where feature selection is considered in terms of both HS and LS for stability. In the HS environment, Boruta and RFE each identified V1-V20; Permutation (Corrected) identified V1-V21; whilst Permutation (Raw) identified V1-V29. In the LS environment, Boruta, RFE and Permutation (Corrected) all increased from their HS tally to select 29 true positive features (V1-V29).

Following identification of the stable features selected by each method, the performance of validation models trained using only these features and held out data, were quantified in terms of predictive accuracy (R-squared). The results of this analysis are shown in Figure 2(C), where generally, the predictive performance varied negligibly between the models.

We conducted power analysis using a subset of the simulation data containing only the six groups of correlated features (V1-V60) and a single non-correlated variable (V4500). The results from this analysis are presented graphically in SF2(A), where the groups containing the 20 HS stable features selected by Boruta are indicated with an asterisk. The six correlated feature groups were successfully detected by the function and grouped together. The member from each group with the largest effect size were then assessed for power at a broad range of sample sizes using the continuous outcome variable scheme outlined in Study Design (Module two). We compared the features representing the three groups used to generate y (V1-30) in terms of their effect sizes and sample size necessary to achieve maximal power. V1, chosen to represent correlated group one (V1-V10), had an effect size of 0.82, and achieved power=1 at a sample size of ~60; the second y related group (V11-V20) was represented by V19, had an effect size of 0.49 and achieved power=1 with ~140 samples; whilst group three (V21-30) was represented by V21, had an effect size of 0.38 and achieved power=1 at a sample size of ~225.

Examination of the power calculations for the non-y-related features revealed similarly negligible effect sizes for V36 (representing correlated group 3) and V4500. Consequently, the power values calculated for these variables remained close to zero across the full range of sample sizes considered. Contrastingly, V50 and V52 (representing correlated groups 5 and 6, respectively) each produced an effect size 0.14, by chance. Consequently, significant power values could be obtained for these variables at sufficiently large sample sizes. We determined a sample size of ~1990 necessary to observe power=1.

Published experimentally derived metabolomics data for regression

We applied similar strategies on a public -omics dataset by Acharjee et al., (2016) [21], featuring lipid metabolites (Positive DI-MS Lipids) and using relative liver weight as a continuous outcome variable (y). Six metabolites of interest were identified by the original work and were thus considered known by the present study.

Boruta selected the largest number of HS features, including three previously known features of interest (Table 3, SF3A). In the LS environment, Boruta selected an additional forty features. In addition, the HS Boruta model achieved the highest R-squared value, even slightly exceeding the value it achieved under LS (Fig. 3A). Permutation (Raw) also achieved strong stability across iterations (Table 3), exhibited consistent predictive performance (R-squared) across validation models (Fig. 3A), and identified two known metabolites under HS. RFE and Permutation (Corrected) both exhibited poor stability across iterations, retaining only six and three HS features, respectively, despite the methods identifying 94 and 51 features each, under LS (Table 3).

The results for power analysis of -omics data 1 are presented in Figure 3(B). The function reduced the 46 HS Boruta features to seven highly correlated groups, and power calculations were performed for the group member with the largest observed effect size. Five potential biomarkers with an effect size in excess of 0.6 emerged from this analysis, for which maximal power values were observed at a sample size between 35 and 45. The two remaining features obtained effect sizes of 0.4 and 0.27 and achieved max power at a sample size of 75 and 620, respectively. The relationship between each of the assessed variables and the continuous outcome variable are displayed in Figure 4(A). All seven variables were deemed statistically significant (P-value < 0.05), whilst adjusted R-squared values ranged from 0.77 for 'SM(39:7)' to just 0.06 for 'PC(33:2) or PE(36:2). The three previously known metabolites selected by Boruta: 'PE(42:4)', 'PC(40:5)' and 'PC(42:9)', all correlated highly with the putative biomarker with the largest effect size 'SM(39:7)'.

Published experimentally derived lipidomics data for regression

The results from a second application of the regression approach using lipidomic data from 3 months old infants are summarised in Table 3 and SF3(B). Previous work identified three lipids: PC(35:2), SM(36:2) and SM(39:1)[35] and were thus considered known. In our analysis, the three known metabolites were observed by all four feature selection methods in the LS environment, whilst 'SM(36:2)' and 'SM(39:1) were retained by Boruta and Permutation (Raw) in the HS environment. RFE and Permutation (Corrected) struggled for stability for this data, only retaining a single feature: 'SM(39:1)', under HS. In total, eight features were selected by Boruta under HS and used to conduct power analysis.

The power function produced three correlated groups represented by each of the observed known metabolites: 'SM(36:2)' & 'SM(39:1)', and the novel potential biomarker: 'PC(34:2)'. The three features achieved effect sizes 0.74, 0.66 and 0.68, respectively. Power calculations performed similarly for all three groups, achieving maximal power between 35 and 45 samples (SF4B). The relationship between the three subset features and the outcome variable are displayed in Figure 4(B). Once again, all exhibited a statistically significant relationship (P-value <0.05) but this time produced the less emphatic adjusted R-squared values of 0.53, 0.43, and 0.44, respectively.

Classification Mode

Simulation Data

The simulated dataset was modified to produce a binary classification outcome variable (y) and utilised for feature selection (Table 3, SF5A). Under HS, whilst all methods selected all of the variables from the first group used to generate y (V1-V10), only a single variable from the second group (V11-V20) was identified by Boruta and Permutation (Corrected) and none were identified by RFE. Permutation (Raw) was the most successful method under HS, identifying V11-V19 from group 2. No method selected any variables from the third group (V21-V30). In the LS environment, Boruta exhibited the greatest stability, increasing its tally of true positive features to 29 whilst selecting only 9 false positive variables. RFE, Permutation (Raw) and Permutation (Corrected) also increased their true positive tally to 29, but selected 201, 465 and 110 false positive variables, respectively (Table 3).

We conducted power analysis on the eleven true positive features stably selected by Boruta under HS, the results from this part of the analysis are displayed in SF5(B). The power function correctly identified two groups of features to assess from those selected (V1-V10 and V11) and calculated Cohen's d effect sizes for each variable. V2 was chosen at random to represent the first group as several features observed equally large effect sizes. Both of the assessed features exhibited effect sizes greater than the 0.8 threshold for a large effect defined by Cohen (1998)[33], with the value of 1.82 observed for V2 and 0.88 for V13. Consequently, maximum power values emerged for both features at a sample size <20.

Published experimentally derived metabolomcis for classification

To further assess the performance the classification approach, we once again utilised data from Acharjee et al. (2016) [21], with a modified binary classification outcome variable: below or above the mean value of the relative liver weight (Table 3 and SF6B). Boruta and Permutation (Raw) selected similar numbers of features in both HS and LS environments. Under HS, both methods selected two lipids known a priori and ~20 novel features of interest. Under LS, the methods each identified an additional known lipid alongside ~70 potential novel features likely featuring a number of concealed false positives. RFE observed comparatively few features in the LS context and observed none at all under HS. Conversely, Permutation (Corrected) selected eight HS features, more than double the equivalent number identified during regression analysis. The 26 HS stable features selected by Boruta were used to conduct power analysis. Two groups of correlated features were produced and represented by 'PC(32:0) or PE(35:0)' with a Cohen's d effect size of 1.92 and 'PC-O(35:5) or PC-P(35:4)' with an effect size of 2.24. These values exceed Sawilowski's descriptors of 'very large' and 'huge', respectively [39]. Maximum power values were achieved for each group at a sample size of ~10-20 (SF6B).

Published experimentally derived transcriptomics data 2 for classification

Finally, we applied the binary classification procedure to a genomics cohort featuring four healthy control samples and individuals from stages 1-4 of colorectal cancer (CRC). The analysis was conducted as a series of pairwise classifications seeking to identify the biomarkers that distinguished most effectively the healthy controls from samples from each CRC stage (Control vs stage 1 in table 3 and other stages in SF7). Across all four pairwise comparisons, Boruta selected a similar number of features in both the HS and LS environments. Permutation (Raw) selected far fewer features under HS compared to LS but remained reasonably sensitive, successfully selecting a number of HS stable features in all but the control vs stage 1 analysis. RFE exhibited low sensitivity across all comparisons, failing to identify any HS features and a maximum of 14 LS features. Similarly, Permutation (Raw) was limited in its selection of HS features despite selecting the second largest number of features under LS, thus indicating poor stability of the selections between iterations.

The HS stable features selected by Boruta in each of the four analyses were provided to the binary classification power functions described in study design (Module 2). In each case the function simplified the provided features into correlated groups and identified the feature from each group determined to possess the largest Cohen's d effect size. In addition, t-tests were conducted for each feature, and statistically significant differences were observed between binary classes, in all cases (SF8).

Power calculations were conducted and this process determined that at their observed effect size, the three features identified in the control vs stage one comparison could achieve maximum power using <10 case samples; whilst the sample sizes necessary for the features selected in the three other comparisons ranged between <5 for the largest effect features and >1280 for the smallest (SF9). In each comparison scenario, the effect size of all features were greater than the 'very large' threshold of 1.0 [39].

Table 3: List of the methods, stringency and criteria used to evaluate methods is listed for both regression and classification.

RF Methods	**Stringency**	**Criteria**	**Regression**			**Classification**		
			Simulation	**Metabolomics**	**Lipidomics**	**Simulation**	**Lipidomics**	**Transcriptomics (stage1)**
RFE	High	TP/Known	20	1	1	10	0	-
		FP/Novel	0	5	0	0	0	0

	Low	TP/Known	29	3	3	29	2	-
		FP/Novel	19	91	8	201	18	14
Boruta	High	TP/Known	20	3	2	11	2	-
		FP/Novel	0	43	6	0	24	19
	Low	TP/Known	29	3	3	29	3	-
		FP/Novel	1	83	34	9	10	39
Permutation (Raw)	High	TP/Known	29	2	2	19	2	-
		FP/Novel	0	24	7	0	10	0
	Low	TP/Known	29	3	3	29	3	-
		FP/Novel	98	68	47	465	35	132
Permutation (Corrected)	High	TP/Known	21	2	1	11	2	-
		FP/Novel	0	1	0	0	6	0
	Low	TP/Known	29	3	3	29	3	-
		FP/Novel	8	48	26	110	46	66

Web tool

To streamline power calculations and provide an accessible package fit for translational medicine, we produced ‘PowerTools’, an interactive open-source web application written in R code using the Shiny framework (http://shiny.rstudio.com/) (Figure 5). The tool is capable of performing efficient simulation-based power calculations for regression and two-group classification datasets from various -omics disciplines. The web interface allows easy specification of sample sizes, quick access to function parameters and is equipped with help

information and example datasets to maximise the tool's accessibility. Confusion matrix result values are presented as both a customisable plot and as raw data tables, which can be easily downloaded using the intuitive user interface. A user guide is added in the Additional file 2.

Discussion

In the present study, we first examined the performance of four RF feature selection methods in the context of regression and classification, using both simulated and published datasets. The known underlying correlation structure and relationship with outcome variable (y) of the simulated data, facilitated assessment of the methods in terms of the number of true and false positive features they identified.

In the regression context, all methods were stable in their selection of only true positive features under HS, whereas Boruta alone maintained near perfect specificity under LS. In contrast, Permutation (Raw) selected an additional 98 false positive features under LS vs its HS tally. These findings suggest that multiple iterations of feature selection, combined with a high stringency threshold for stability, effectively eliminate any false positive features selected by the methods by chance. The high stability of the features selected by Boruta produced the smallest difference in performance between stringency contexts, suggesting its use in scenarios where the large computational runtime of multiple iterations is infeasible. The gulf in LS performance supplemenet between Boruta and permutation testing corroborates previous findings using similar simulation procedures [26], and likely relates to the robust identification of features achieved by the former approaches.

When applied to published experimental data, it was impossible to distinguish variables truly associated with the outcome from false positives. Consequently, the subset of features selected by each method were assessed only in terms of their stability, numerosity and the predictive performance of their resultant validation models. Across the two published experimental datasets considered for regression, Boruta once again exhibited the best stability, producing the smallest difference in feature subsets selected between stringency states. In addition, Boruta selected the largest number of stable HS features from the first dataset, consistent with the

notion that the method uses an all relevant approach [25] and has selected the largest number of biomarker candidates in previous comparison studies [40].

In each of the published experimental data applications, RFE identified the smallest subset of HS stable features. However, this was expected as by design the method seeks the smallest subset of predictive features [27]. Consequently, RFE exhibited the worst stability, producing a 15-fold difference in the number HS/LS features found for the first dataset and an 11-fold difference for the second.

In terms of classification , both simulated and published experimental datasets observed fewer stable feature selections by each method, suggesting more stringent selection criteria for classification. However, the relative performance of each method was similar, with Boruta again exhibiting the greatest stability, across all data considered. For the simulated data, all methods once again selected only true positive variables under HS, with Permutation (Raw) selecting the largest number.

For the first published experimental dataset, Boruta and Permutation (Raw) exhibited similar stability, whilst Boruta selected the largest number of marker candidates. The second published experimental dataset performed similarly between each pairwise comparison of controls and CRC stage; Boruta was the method that selected HS features in all four comparisons and remained consistent in its selections under LS. Throughout, RFE exhibited the poorest stability, in fact failing to select a single HS feature across all three data contexts, again due to its low stability between repeated iterations.

Throughout module one, CV was conducted in order to optimise hyperparameters (SF1) and produce validation models to assess the subset of features selected by each of the assessed methods. Within each dataset, we observed limited variation in validation model performance metrics, supporting similar findings from the literature where feature selection has been evaluated in terms of model performance and stability [40,41]. The RF algorithm is capable of compensating for noisy features without a large decrease in model performance, thus validation model performance alone is an ineffective measure of the relative performance of RF feature selection approaches [41]. Furthermore, the variance in predictive performance values between CV repeats were greater when using variables selected under HS (e.g. Fig.2C). This effect can

be understood with respect to their being greater differences, between outer loop repeats, in the variables meeting the HS selection criteria, compared to the equivalent differences in those which surpass the LS selection criteria.

For our second module, we sought to combine our stable feature selection protocol with a novel simulation-based approach to facilitate power calculations for future study design. Further to previous efforts by Guo et al., 2010 [24], who compared the power achieved by a multitude of classifiers when presented with diverse sets of data, we focused singularly on RF but expanded into both classification and regression domains. Furthermore, we ensured the true correlation structure of input data was captured, applying a methodology originally used in the context of metabolic phenotyping [23]. Once more, we incorporated automated effect size calculations, and grouped similar variables before filtering by effect size, to identify a small subgroup of high effect putative biomarkers for which to quantify power.

We validated the performance of our power calculations with respect to the values achieved using the simulated regression data. The power values predicted for each feature, at the number of samples actually used in module one for selection, matched with the observed empirical power achieved by the most stable feature selection methods (SF2B). Furthermore, as expected, we consistently observed smaller necessary sample sizes for the features with largest effect values. This observation was well illustrated for CRC stage 3 from the second classification dataset [36] (SF9C), where effect sizes ranged between 4.78 and 1.65, and the sample size necessary to obtain maximal power ranged from <5 to >1280.

We explored general trends in the sample sizes necessary to achieve good power, concluding less than 40 samples necessary to observe max power at an effect size of 0.8 in regression mode and no more than 10 samples necessary for a Cohen's d effect of 3.0, for classification. The values at these effect sizes are reported as they appeared consistently amongst the stable features selected by Boruta in module one of the present study. In almost all cases, fewer samples were necessary to observe max power for the features selected during classification than the equivalent chosen in regression mode. This observation corroborates the greater stringency observed for classification mode feature selection, in module one.

Lastly, we developed an interactive open-source web application: 'PowerTools', to facilitate not only the estimation of the number of samples required for future study, but to group similar features and determine the effect size associated with potential biomarkers. Whilst other computational tools for power analysis have been produced previously, many have been limited in their accessibility: providing only raw functions [23,42]; relating only to specific study designs, such as case-control microbiome studies [43]; or are now defunct [24]. We believe that our workflow and approach is generalised across multiple different –omics datasets and will help in the translational community to interpret the stability and future design aspects of the biomarkers.

Conclusion

In this paper, we presented a number of different RF based stable feature selection methods and compared their performances using simulated as well as published experimentally derived datasets. Across all of the scenarios considered we found Boruta to be the most stable methodology, whilst Permutation (Raw) offered the largest number of relevant features when allowed to stabilise over a number of iterations. We determined that the decision about which approach to pursue should be weighed against the computational requirements and runtime necessary. Finally, we extended and linked the effect size of the stable markers with future study, using a web-based interface.

Declarations

Data availability

All the data is available in the respective published papers and also in the github repository.

Author's contributions

JL carried out data analysis, developed web interface and was involved in the drafting of the manuscript. AA conceived and designed the data analytics strategy. VRC contributed to the Additional file and revised the manuscript. AA and GVG supervised the project and the analysis. All authors contributed to the writing of the manuscript.

Funding

This study was supported by the National Institute for Health Research (NIHR) Surgical Reconstruction and Microbiology Research Centre (SRMRC), Birmingham. GVG also acknowledges support from H2020-EINFRA (731075) and the National Science Foundation (IOS:1340112) as well as support from the NIHR Birmingham ECMC, the NIHR Birmingham Biomedical Research Centre and the MRC HDR UK. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Medical Research Council or the Department of Health.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All the data used in this study is available from the respective published papers as well as from our GitHub repository [2].

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This study was supported by the National Institute for Health Research (NIHR) Surgical Reconstruction and Microbiology Research Centre (SRMRC), Birmingham. GVG also acknowledges support from H2020-EINFRA (731075) and the National Science Foundation (IOS:1340112) as well as support from the NIHR Birmingham ECMC, the NIHR Birmingham Biomedical Research Centre and the MRC HDR UK. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Medical Research Council or the Department of Health.

Abbreviations

ANOVA: Analysis of variance

CBGS: Cambridge Baby Growth Study

RF : Random forest

SM: Sphingomyelin

PC: phosphatidylcholine

REFERENCES

1. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10: 57–63.
2. Clark TA. Genomewide Analysis of mRNA Processing in Yeast Using Splicing-Specific Microarrays. Science. 2002; 296: 907–10.
3. McGrath CM, Young SP. Can metabolomic profiling predict response to therapy? Nat Rev Rheumatol. 2019; 15: 129–30.
4. Patti GJ, Yanes O, Siuzdak G. Metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Biol. 2012; 13: 263–9.
5. Domon B. Mass Spectrometry and Protein Analysis. Science. 2006; 312: 212–7.
6. Martens L. Proteomics Databases and Repositories. In: Wu CH, Chen C, Eds. Bioinformatics for Comparative Proteomics [Internet]. Totowa, NJ: Humana Press; 2011 [cited 10 July 2019]: 213–27. Available at: http://link.springer.com/10.1007/978-1-60761-977-2_14

7. Cani PD. Human gut microbiome: hopes, threats and promises. Gut. 2018; 67: 1716–25.

8. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nat Rev Genet. 2012; 13: 260–70.

9. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The Human Microbiome Project. Nature. 2007; 449: 804–10.

10. Robinson PN. Deep phenotyping for precision medicine. Hum Mutat. 2012; 33: 777–80.

11. Azuaje F. Artificial intelligence for precision oncology: beyond patient stratification. npj Precision Onc. 2019; 3: 6.

12. Mischak H, Allmaier G, Apweiler R, et al. Recommendations for Biomarker Identification and Qualification in Clinical Proteomics. Science Translational Medicine. 2010; 2: 46ps42-46ps42.

13. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007; 23: 2507–17.

14. Bravo-Merodio L, Williams JA, Gkoutos GV, Acharjee A. -Omics biomarker identification pipeline for translational medicine. J Transl Med. 2019; 17: 155.

15. Breiman L. Random Forests. Machine Learning. 2001; 45: 5–32.

16. Alexe G, Monaco J, Doyle S, et al. Towards Improved Cancer Diagnosis and Prognosis Using Analysis of Gene Expression Data and Computer Aided Imaging. Exp Biol Med (Maywood). 2009; 234: 860–79.

17. Smolinska A, Hauschild A-C, Fijten RRR, Dallinga JW, Baumbach J, van Schooten FJ. Current breathomics—a review on data pre-processing techniques and machine learning in metabolomics breath analysis. J Breath Res. 2014; 8: 027105.

18. Wilhelm T. Phenotype prediction based on genome-wide DNA methylation data. BMC Bioinformatics. 2014; 15: 193.

19. Swan AL, Mobasheri A, Allaway D, Liddell S, Bacardit J. Application of Machine Learning to Proteomics Data: Classification and Biomarker Identification in Postgenomics Biology. OMICS: A Journal of Integrative Biology. 2013; 17: 595–610.

20. Acharjee A, Kloosterman B, de Vos RCH, et al. Data integration and network reconstruction with ~omics data using Random Forest regression in potato. Analytica Chimica Acta. 2011; 705: 56–63.

21. Acharjee A, Ament Z, West JA, Stanley E, Griffin JL. Integration of metabolomics, lipidomics and clinical data using a machine learning method. BMC Bioinformatics. 2016; 17: 440.

22. Billoir E, Navratil V, Blaise BJ. Sample size calculation in metabolic phenotyping studies. Brief Bioinform. 2015; 16: 813–9.

23. Blaise BJ, Correia G, Tin A, et al. Power Analysis and Sample Size Determination in Metabolic Phenotyping. Anal Chem. 2016; 88: 5179–88.

24. Guo Y, Graber A, McBurney RN, Balasubramanian R. Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms. BMC Bioinformatics. 2010; 11: 447.

25. Kursa MB, Rudnicki WR. Feature Selection with the **Boruta** Package. J Stat Soft [Internet]. 2010 [cited 10 July 2019]; 36. Available at: http://www.jstatsoft.org/v36/i11/

26. Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. Briefings in Bioinformatics. 2019; 20: 492–503.

27. Diaz-Uriarte R. GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. BMC Bioinformatics. 2007; 8: 328.

28. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995; 57: 289–300.

29. Hendriks MMWB, Smit S, Akkermans WLMW, et al. How to distinguish healthy from diseased? Classification strategy for mass spectrometry-based clinical proteomics. Proteomics. 2007; 7: 3672–80.

30. Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society: Series B (Methodological). 1974; 36: 111–33.

31. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006; 7: 91.

32. Acharjee A. Comparison of Regularized Regression Methods for ~Omics Data. Metabolomics [Internet]. 2012 [cited 10 July 2019]; 03. Available at: https://www.omicsonline.org/comparison-of-regularized-regression-methods-for-omics-data-2153-0769.1000126.php?aid=32360

33. Cohen J. Statistical Power Analysis for the Behavioral Sciences. [Internet]. Hoboken: Taylor and Francis; 1988 [cited 10 July 2019]. Available at: http://www.123library.org/book_details/?id=107447

34. Chen Z, Zhang W. Integrative Analysis Using Module-Guided Random Forests Reveals Correlated Genetic Factors Related to Mouse Weight. Roth FP, Ed. PLoS Comput Biol. 2013; 9: e1002956.

35. Acharjee A, Prentice P, Acerini C, et al. The translation of lipid profiles to nutritional biomarkers in the study of infant metabolism. Metabolomics. 2017; 13: 25.

36. Chen X, Deane NG, Lewis KB, et al. Comparison of Nanostring nCounter® Data on FFPE Colon Cancer Samples and Affymetrix Microarray Data on Matched Frozen Tissues. Wang X, Ed. PLoS ONE. 2016; 11: e0153784.

37. Wright MN, Ziegler A. ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Soft [Internet]. 2017 [cited 10 July 2019]; 77. Available at: http://www.jstatsoft.org/v77/i01/

38. Kuhn M. Building Predictive Models in R Using the caret Package. J Stat Soft [Internet]. 2008 [cited 10 July 2019]; 28. Available at: http://www.jstatsoft.org/v28/i05/

39. Sawilowsky SS. New Effect Size Rules of Thumb. J Mod App Stat Meth. 2009; 8: 597–9.

40. Kursa MB. Robustness of Random Forest-based gene selection methods. BMC Bioinformatics. 2014; 15: 8.

41. Fortino V, Kinaret P, Fyhrquist N, Alenius H, Greco D. A Robust and Accurate Method for Feature Selection and Prioritization from Multi-Class OMICs Data. Arthur J, Ed. PLoS ONE. 2014; 9: e107801.

42. Vieth B, Ziegenhain C, Parekh S, Enard W, Hellmann I. powsimR: power analysis for bulk and single cell RNA-seq experiments. Hofacker I, Ed. Bioinformatics. 2017; 33: 3486–8.

43. Mattiello F, Verbist B, Faust K, et al. A web application for sample size and power calculation in case-control microbiome studies. Bioinformatics. 2016; 32: 2038–40.

Figures

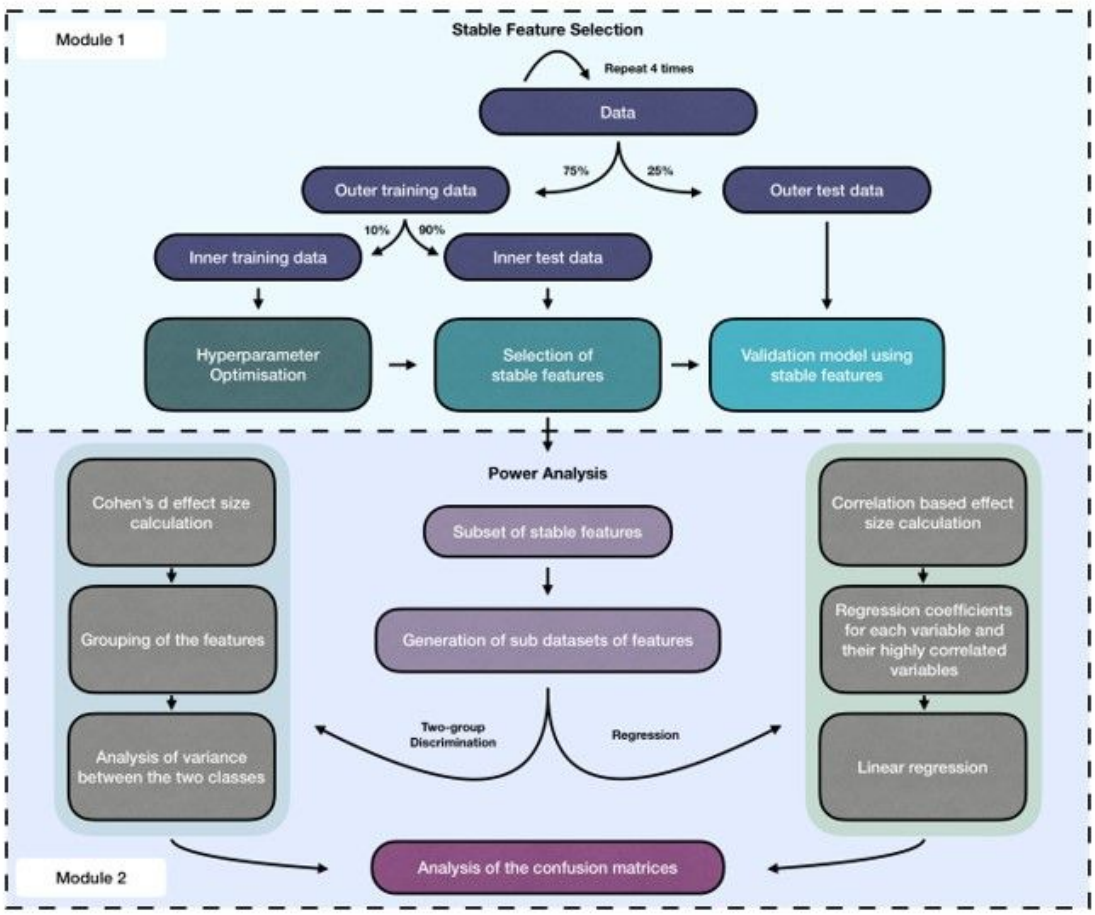

Figure 1

Schematic diagram of the simulation set up and the published experimentally derived (real) data analysis.

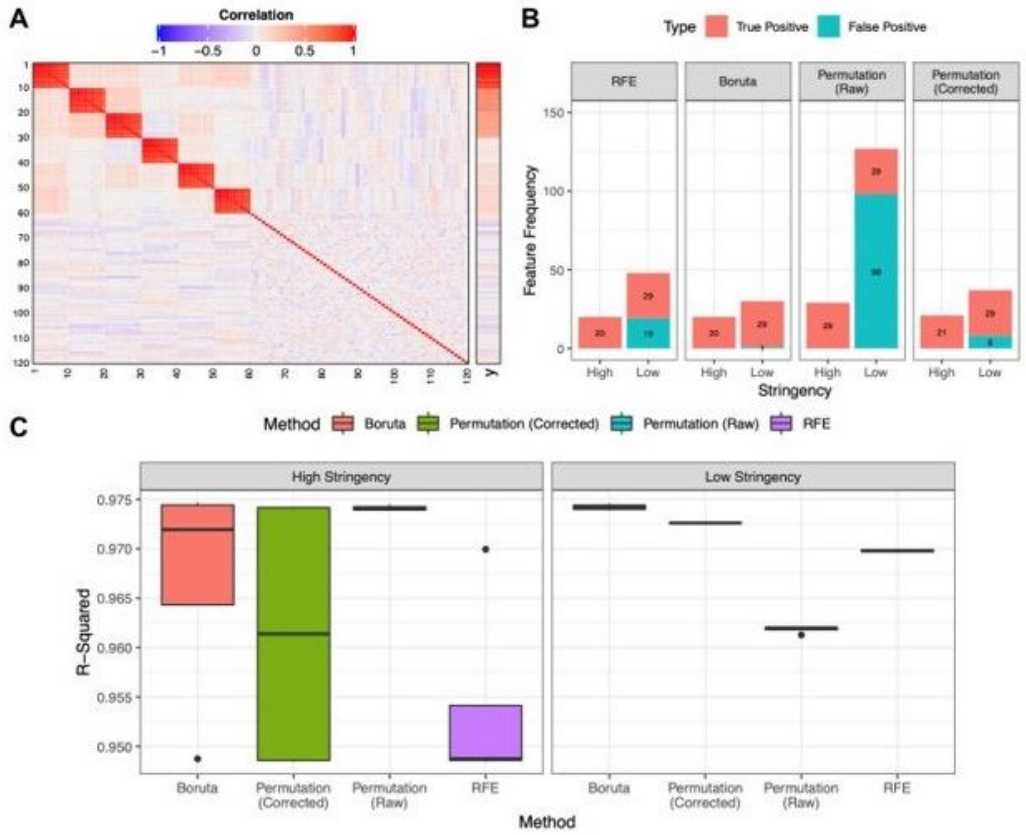

Figure 2

Results from the simulation study in RF regression mode. (A) The structure of the simulated predictor data and the association with outcome variable (y) is described. Only V1-V120 are shown of full dataset featuring 5000 variables. (B) The number of features stably selected by each approach in at least 5/100 iterations (Low Stringency) or a minimum of 90/100 iterations (High Stringency) are shown. True positive: V1-V30, False positive: V3-V5000. Values describing the number of times each feature is chosen by a particular approach are averaged across those achieved after 100 iterations for each of the four inner loop test datasets. (C) The variance in predictive accuracy (R-Squared), across all four outer loop cross-validation repeats, is shown for RFs trained using only the high or LS stable features selected by each feature selection approach using the relevant inner loop test dataset.

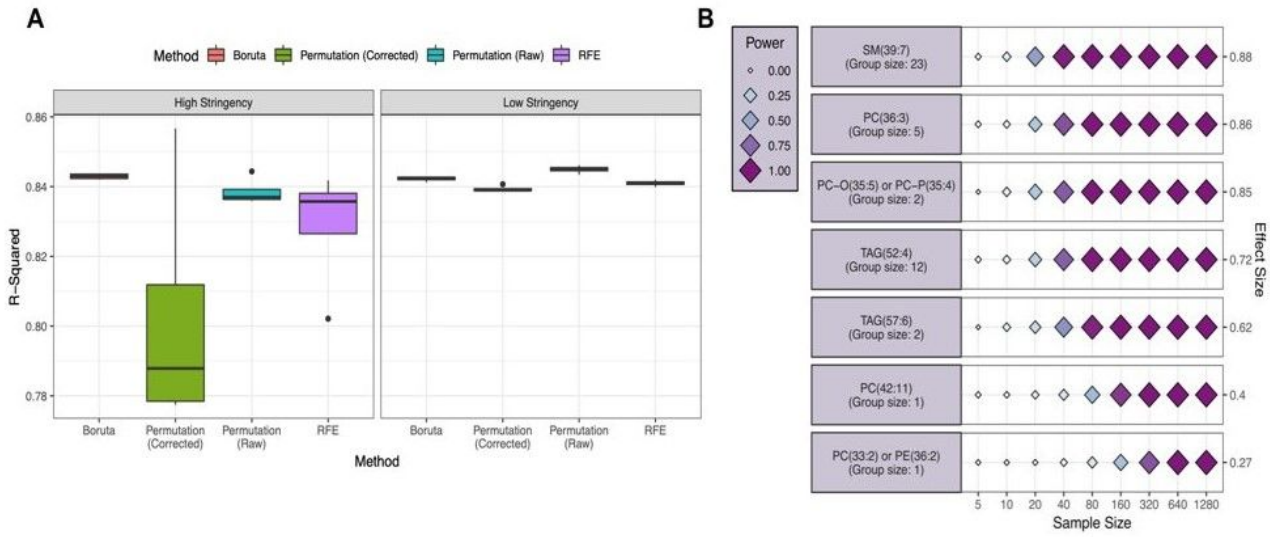

Figure 3

Validation model performance and power analysis of published experimentally derived data 1, regression mode . (A) Boxplots displaying the variance in the observed R-squared value of validation models trained using the stable features selected by each feature selection approach, across four outer-loop CV repeats. Values are shown for models trained using either the features selected by each approach in at least 5/100 iterations (Low Stringency) or a minimum of 90/100 iterations (High Stringency). (B) The three groups of correlated features identified by the power function are represented by the group member with the largest observed effect size. The effect size of each assessed variable is shown along the y axis and a series of sample sizes along the x axis. Power values determined for each effect/sample size combination using a simulated dataset with the same correlation structure as input data and displayed using variably sized/coloured rhombi.

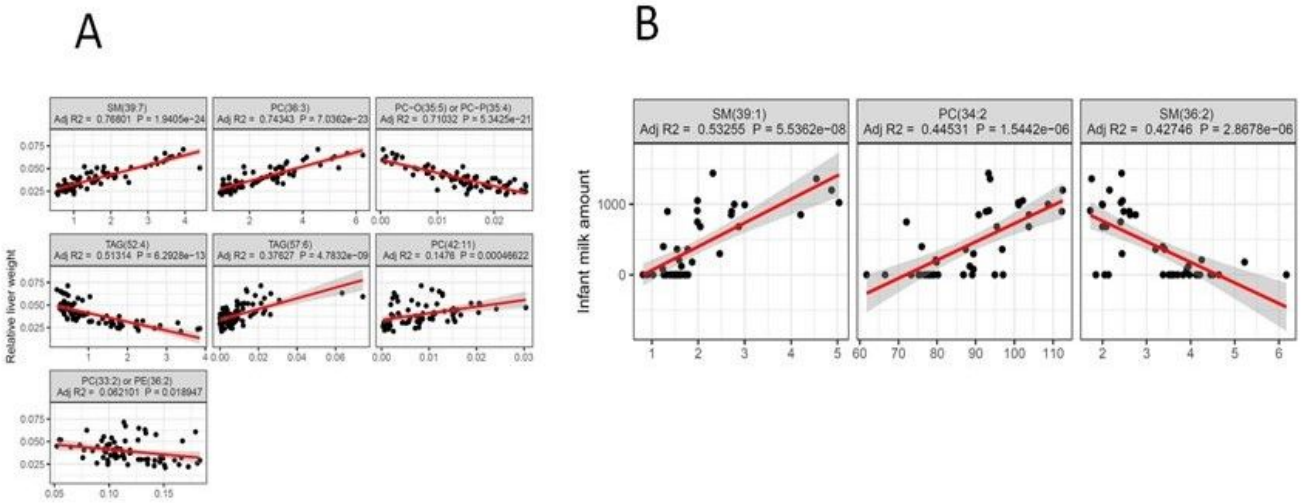

Figure 4

Results from public dataset identified by the module 1 of the workflow is listed above with probability values <0.05. (A) Stable metabolic markers and their variance explained with relative liver weight is shown. (B) Lipids associated with amount of milk in the 3m old infants are listed.

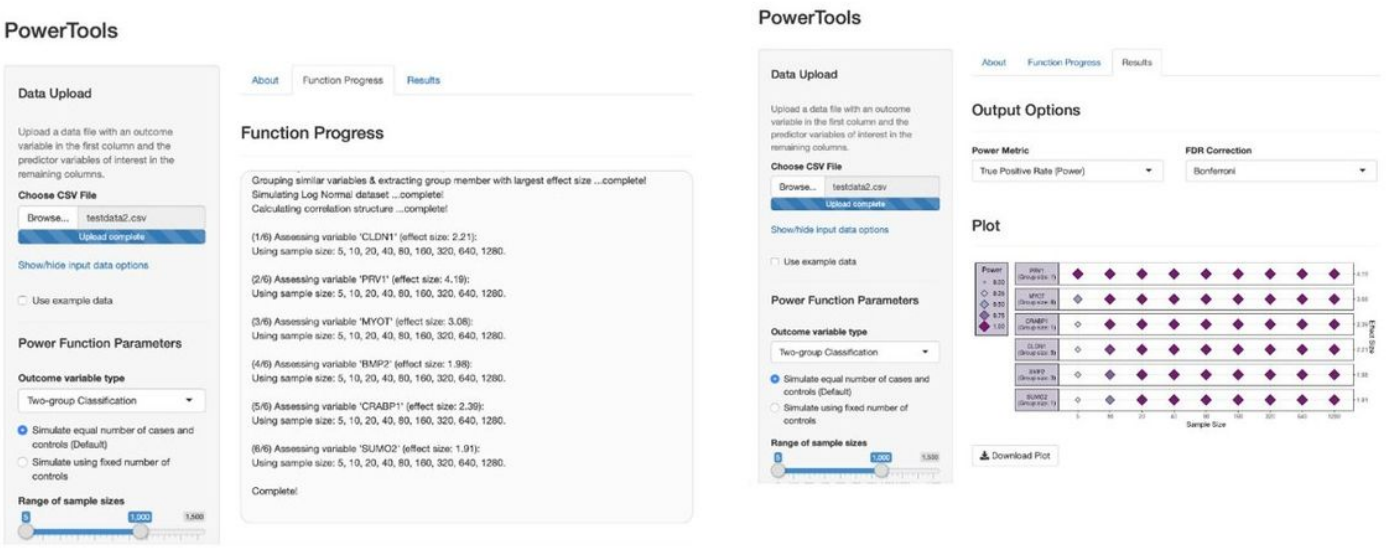

Figure 5

Screenshots of the open-source web application ‘PowerTools’, for efficient and accessible simulation based power calculations.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterial1.docx
- SupplementaryMaterial2.docx