PowerTools: A web based user-friendly tool for future translational study design

Animesh Acharjee (✉ a.acharjee@bham.ac.uk)

University of Birmingham <https://orcid.org/0000-0003-2735-7010>

Joseph Larkman

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

Victor Roth Cardoso

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

Georgios V. Gkoutos

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

Methodology

Keywords: power study, sample size, translational research, omics

Posted Date: February 18th, 2020

DOI: <https://doi.org/10.21203/rs.2.23833/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background

Biomarker identification is one of the major goals of functional genomics and translational medicine research. The advent of NGS lead to a constant and exponential increase of large datasets that have the potential of providing the means for novel biomarker identification for the early diagnosis of complex diseases and/or for patient/disease stratification. Once a biomarker has been identified, a validation study is necessary to assess its value. A study design that considers its appropriateness and cost-effectiveness is paramount. The calculation of a sample size is a challenge that needs to be addressed.

Methods

The workflow of our tool, termed PowerTools, is based on based on the method described by Blaise et al., (2016) [1]. For a given number of input data sets, a simulation step with the random multivariate normal distribution including correlation is considered. As a next step, datasets of variable sizes are generated by random selection of samples. Based on the outcome variable, either classification or regression modes are available. For binary classification ANOVA and linear regression test can be performed and then performance matrices can be evaluated.

Results

We developed an online framework to streamline power calculations to aid future omics study designs within a translational medicine research context. We make our code freely available on GitHub [2] and we have provided a web interface that can be accessed at online [3].

Conclusions

PowerTools offers the potential for designing appropriate and cost-effective subsequent omics study designs.

Background

Over the last few years there has been a lot of emphasis on the high dimensional omics data generation that includes untargeted omics datasets such as transcriptomics [4], [5] metabolomics [6], [7], proteomics [8], [9], microbiomes [10] [11] [12] and deep phenotyping [13]. Vast amount of data is routinely accumulated which needs to be integrated and analysed to facilitate the identification of the relevant markers. If the identified markers from the various omics datasets are robust, reproducible and indicative then they can be used as a biomarker for patient's stratification [14], [15] and can also be useful either as diagnostics or prognostic tools.

To validate those biomarkers experimentally, a study needs to be carefully designed and more often than not encompass features that are sometimes arbitrary. Earlier studies have focused on generating power

analysis outputs for such scenarios using different omics datasets, for example metabolomics data [16], [1] and transcriptomics data [17]. However, those studies are very specific to these omics datasets and often fail to relate power calculations to the relevant biomarkers that were identified. In this study, we developed a web based interface tool, termed PowerTools, to streamline power calculations and offer a valuable asset for use in translational research.

PowerTools forms a flexible webtool to facilitate power analysis and sample size determination, based on a method described by Blaise et al., (2016) [1]. It can take as input two types of response (outcome) variables; regression (continuous variables) and binary classification (class variables) outcomes. Furthermore, the correlation structure of predictor variables was explicitly modelled, in order to capture any multi-collinearity between variables. To increase the potential adaption of our tool, we used the R statistical software environment (<https://www.r-project.org/>) to implement functionality. Additionally, the redesigned functions incorporated comprehensive progress messages and error notation, to improve their usability. Furthermore, the R implementation, presented here, improves the functionality of the original functions in two key respects. Firstly, each variable is automatically assessed using its true effect size (i.e. in the case of regression, the true effect size of a variable is estimated as its correlation with the true outcome variable, whilst in a binary classification, the observed Cohen's d effect size [8] is computed). Secondly, our approach caters for highly correlated variables to be optionally grouped together and only the member of each group with the largest effect size to be used for assessment thereby facilitating the identification of a smaller subset of potential biomarkers.

Materials And Methods

PowerTools Workflow

PowerTools accepts as input a set of –omics biomarkers associated with an outcome variable. Based on the outcome variable, either a binary or continuous class, it performs a simulation with a random multivariate normal distribution. The design of the workflow also considers potential correlations between the biomarkers. Figure 1 depicts graphically the PowerTools's workflow. ANOVA is performed for the case of binary classification whilst linear regression tests are performed for the case of continuous outcomes. As a last step, performance matrices are generated and evaluated. For detail simulation methods, please refer to the Blaise et al., 2016 [1].

Datasets

Our case studies were based on previously published experimental datasets, presented in Table 1.

Table 1: List of the published datasets used in this study.

<i>PowerTools Mode</i>	<i>Type of omics data</i>	<i>Number of Biomarkers</i>	<i>Number of samples</i>	<i>Outcome variable</i>	<i>Article reference (pubmed ID)</i>
Regression	Lipidomics	3	40	3 months infant milk amount	[18] (28190990)
Classification	Cytokine	3	51	Multi organ dysfunction (Yes/No)	[19] (31857590)

Software and Code Availability

We used the R v3.5.0 software for statistical computing [20]. The web interface was constructed using R shiny app [21] and is available online [3]. All our input and supplementary files can be found on our GitHub repository [2].

Results

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. The tool is capable of performing efficient simulation-based power calculations for regression and binary classification datasets from various omics disciplines. The web interface caters the estimation of sample sizes, quick access to function parameters and is complemented with help information and example datasets. Performance matrix or confusion matrix result values are presented as both a customisable plot as well as raw data tables, which can be downloaded using the user interface. A screenshot of PowerTools is presented in the Figure 2.

Case studies

PowerTools was applied to perform power analysis using previously published freely available omics datasets. To assess the two different modes, regression and classification, we have employed the data published by Acharjee et al., 2017 [18] and Bravo-Merodio et al., 2019 [19].

Regression mode case

In this category, the outcome variable considered was the amount of the milk given to the infants in the Cambridge Baby Growth Study (CBGS). In this case, we used PowerTools in the regression mode. A previous study [18] identified three lipids: PC(35:2), SM(36:2) and SM(39:1) and were thus considered for a potential future design study individually.

To achieve maximum power ($P=1$) for all the biomarkers, 40 samples are required. Usage of 20 samples will allow for these markers to achieve 0.50 to 0.75 power. The three biomarker features SM(39:1),

SM(36:2) and PC(35:2) achieved effect sizes of 0.74, 0.66 and 0.63, respectively. The relationship between each of the assessed variables and the continuous outcome variable are depicted in Figure 3.

Classification Mode case

We used three physiological features (decrease neutrophil CD62L and CD63 expression as well as monocyte CD63 expression and frequency) [19] as potential biomarkers for multi organ dysfunction (MODS) development. These features were identified by Bravo-Merodio et al., 2019 [19] as biomarker of immune response to trauma in 51 patients at three different post-injury time points (ultra-early (≤ 1 h), 4–12 h, 48–72 h). Following a power analysis, we found CD62L requires 40 samples in each category to achieve a power of 0.86. The lowest effect size estimated was for Monocyte count (0.7).

Discussion

We developed PowerTools, an interactive open-source web application to facilitate the estimation of the number of samples required for future studies and to group similar features determining the effect size associated with potential biomarkers. In addition to that, PowerTools offers the potential of not only calculating the future sample size but also have a positive impact on optimising time and cost for future experiments. A limitation of our tool is that it caters only for binary classifications. In a future version, we plan to add multi class classification options.

Whilst there are a few examples of computational tools for power analysis, by and large they have been limited in their applications and data types. For example, some have been developed to provide only raw functions [1] [22], others are directly related to specific study designs, such as case-control microbiome studies [23] and some are not currently being maintained [17].

Conclusion

PowerTools forms an interactive open-source web application that utilises an intuitive visual representation to cater for the estimation of the number of samples required for potential future studies. We believe that our workflow and approach is generalised across multiple different –omics datasets and will help in the translational and precision medicine community to interpret the stability and future design aspects of potential biomarkers.

Declarations

Data Availability

All the data used in this study is available from the respective published papers as well as from our GitHub repository [2]. The step by step procedure can be found in the supplementary material.

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 analysis strategy. VRC contributed to the supplementary material 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

MODS: Multi Organ Dysfunction

SM: Sphingomyelin

PC: phosphatidylcholine

References

1. Blaise BJ, Correia G, Tin A, Young JH, Vergnaud A-C, Lewis M, et al. Power Analysis and Sample Size Determination in Metabolic Phenotyping. *Analytical Chemistry*. 2016;88(10):5179-88. doi:10.1021/acs.analchem.6b00188.
2. Larkman Jea. PowerTools 2020 [Available from: <https://github.com/gkoutos-group/PowerTools>. Accessed on 13 February 2020.
3. Larkman J. PowerTools 2019 [Available from: <https://joelarkman.shinyapps.io/PowerTools/>. Accessed on 30 January 2020.
4. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*. 2009;10(1):57-63. doi:10.1038/nrg2484.
5. Clark TA, Sugnet CW, Ares M. Genomewide Analysis of mRNA Processing in Yeast Using Splicing-Specific Microarrays. *Science*. 2002;296(5569):907-10. doi:10.1126/science.1069415.
6. McGrath CM, Young SP. Can metabolomic profiling predict response to therapy? *Nature Reviews Rheumatology*. 2019;15(3):129-30. doi:10.1038/s41584-018-0136-z.
7. Patti GJ, Yanes O, Siuzdak G. Metabolomics: the apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology*. 2012;13(4):263-9. doi:10.1038/nrm3314.
8. Domon B, Aebersold R. Mass Spectrometry and Protein Analysis. *Science*. 2006;312(5771):212-7. doi:10.1126/science.1124619.
9. Martens L. Proteomics Databases and Repositories. In: Wu CH, Chen C, editors. *Bioinformatics for Comparative Proteomics*. Totowa, NJ: Humana Press; 2011. p. 213-27.
10. Cani PD. Human gut microbiome: hopes, threats and promises. *Gut*. 2018;67(9):1716-25. doi:10.1136/gutjnl-2018-316723.
11. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics*. 2012;13(4):260-70. doi:10.1038/nrg3182.
12. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The Human Microbiome Project. *Nature*. 2007;449(7164):804-10. doi:10.1038/nature06244.
13. Robinson PN. Deep phenotyping for precision medicine. *Human Mutation*. 2012;33(5):777-80. doi:10.1002/humu.22080.
14. Azuaje F. Artificial intelligence for precision oncology: beyond patient stratification. *npj Precision Oncology*. 2019;3(1):6. doi:10.1038/s41698-019-0078-1.

15. Mischak H, Allmaier G, Apweiler R, Attwood T, Baumann M, Benigni A, et al. Recommendations for Biomarker Identification and Qualification in Clinical Proteomics. *Science Translational Medicine*. 2010;2(46):46ps2-ps2. doi:10.1126/scitranslmed.3001249.
16. Billoir E, Navratil V, Blaise BJ. Sample size calculation in metabolic phenotyping studies. *Brief Bioinform*. 2015;16(5):813-9. doi:10.1093/bib/bbu052.
17. 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(1):447. doi:10.1186/1471-2105-11-447.
18. Acharjee A, Prentice P, Acerini C, Smith J, Hughes IA, Ong K, et al. The translation of lipid profiles to nutritional biomarkers in the study of infant metabolism. *Metabolomics*. 2017;13(3):25. doi:10.1007/s11306-017-1166-2.
19. Bravo-Merodio L, Acharjee A, Hazeldine J, Bentley C, Foster M, Gkoutos GV, et al. Machine learning for the detection of early immunological markers as predictors of multi-organ dysfunction. *Scientific Data*. 2019;6(1):328. doi:10.1038/s41597-019-0337-6.
20. Team RC. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2018.
21. McPherson WCaJCaJAaYXaJ. shiny: Web Application Framework for R. 2019.
22. Vieth B, Ziegenhain C, Parekh S, Enard W, Hellmann I. powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics*. 2017;33(21):3486-8. doi:10.1093/bioinformatics/btx435.
23. Mattiello F, Verbist B, Faust K, Raes J, Shannon WD, Bijmens L, et al. A web application for sample size and power calculation in case-control microbiome studies. *Bioinformatics*. 2016;32(13):2038-40. doi:10.1093/bioinformatics/btw099.

Figures

Figure 1

A schematic diagram of the workflow. The evaluation starts with a dataset with biomarkers of interest. Different steps are executed depending on the type of data. For both cases the effect size is calculated, for binary classification with Cohen's d and for regression with correlation. The following steps for binary classification are: grouping of feature and analysis of variance; for regression a regression coefficient is calculated and a linear regression is performed. In both cases there is a final step for analysis of decision matrices.

PowerTools

By Joseph Larkman

Data Upload

Upload a data file with an outcome variable in the first column and the predictor variables of interest in the remaining columns.

Choose CSV File

Browse...

testdata2.csv

Upload complete

[Show/hide input data options](#)

☐ Use example data

Power Function Parameters

Outcome variable type

Two-group Classification

☒ Simulate equal number of cases and controls (Default)

☐ Simulate using fixed number of controls

Range of sample sizes

5

1,000

1,500

AboutFunction ProgressResults

Function Progress

Grouping similar variables & extracting group member with largest effect size ...complete!

Simulating Log Normal dataset ...complete!

Calculating correlation structure ...complete!

(1/6) Assessing variable 'CLDN1' (effect size: 2.21):
Using sample size: 5, 10, 20, 40, 80, 160, 320, 640, 1280.

(2/6) Assessing variable 'PRV1' (effect size: 4.19):
Using sample size: 5, 10, 20, 40, 80, 160, 320, 640, 1280.

(3/6) Assessing variable 'MYOT' (effect size: 3.08):
Using sample size: 5, 10, 20, 40, 80, 160, 320, 640, 1280.

(4/6) Assessing variable 'BMP2' (effect size: 1.98):
Using sample size: 5, 10, 20, 40, 80, 160, 320, 640, 1280.

(5/6) Assessing variable 'CRABP1' (effect size: 2.39):
Using sample size: 5, 10, 20, 40, 80, 160, 320, 640, 1280.

(6/6) Assessing variable 'SUMO2' (effect size: 1.91):
Using sample size: 5, 10, 20, 40, 80, 160, 320, 640, 1280.

Complete!

Figure 2

Screenshots of the open-source web application PowerTools for simulation-based power calculations.

Figure 3

Biomarkers PC(35:2), SM(36:2) and SM(391:1) are used as in for a potential study design. The effect size of each assessed variable is shown along the y axis and a series of sample sizes along the x axis. The various size/colour rhombi indicate the power values for each effect/sample size combination using a simulated dataset automatically derived from the real dataset. A) Estimation of the sample size with individual features (lipids) To achieve maximum power ($P=1$) for all the biomarkers 40 samples are required. Based on this, 20 will achieve 0.50 to 0.75 power. B) Estimation of the number of the samples required for a future study considering the group of the three lipids. Following clustering of the lipids identified, two clusters were determined necessitating 80 samples to achieve maximum power ($P=1$).

Supplementary Files

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

- [20200131SupplementaryMaterial1.docx](#)