

## Sequence analysis

# A web application for sample size and power calculation in case-control microbiome studies

Federico Mattiello<sup>1,\*</sup>, Bie Verbist<sup>2</sup>, Karoline Faust<sup>3</sup>, Jeroen Raes<sup>3</sup>,  
William D. Shannon<sup>4</sup>, Luc Bijmans<sup>2</sup> and Olivier Thas<sup>1,5</sup>

<sup>1</sup>Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure Links 653, Gent, 9000, <sup>2</sup>Janssen Pharmaceutica, Turnhoutseweg 30, Beerse, 2340, Belgium, <sup>3</sup>KU Leuven, Laboratory of Molecular Bacteriology and Department of Microbiology and Immunology, Herestraat 49, Leuven, 3000, Belgium, <sup>4</sup>BioRankings, 4041 Forest Park Ave, St.Louis, MO 63108, USA and <sup>5</sup>University of Wollongong, National Institute for Applied Statistics Research Australia (NIASRA), School of Mathematics and Applied Statistics, Australia

\*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on June 10, 2015; revised on January 21, 2016; accepted on February 15, 2016

## Abstract

**Summary:** When designing a case-control study to investigate differences in microbial composition, it is fundamental to assess the sample sizes needed to detect an hypothesized difference with sufficient statistical power. Our application includes power calculation for (i) a recoded version of the two-sample generalized Wald test of the 'HMP' R-package for comparing community composition, and (ii) the Wilcoxon-Mann-Whitney test for comparing operational taxonomic unit-specific abundances between two samples (optional). The simulation-based power calculations make use of the Dirichlet-Multinomial model to describe and generate abundances. The web interface allows for easy specification of sample and effect sizes. As an illustration of our application, we compared the statistical power of the two tests, with and without stratification of samples. We observed that statistical power increases considerably when stratification is employed, meaning that less samples are needed to detect the same effect size with the same power.

**Availability and implementation:** The web interface is written in R code using Shiny (RStudio Inc., 2016) and it is available at <https://fedematt.shinyapps.io/shinyMB>. The R code for the recoded generalized Wald test can be found at <https://github.com/mafed/msWaldHMP>.

**Contact:** Federico.Mattiello@UGent.be

## 1 Introduction

A pivotal aspect of planning a case-control study is the calculation of sample sizes, which is typically based on the costs of data collection and on the need to have sufficient statistical power for detecting a relevant difference between cases and controls. Power can only be computed if the researchers can specify: (i) the smallest relevant deviation from the null hypothesis that is to be detected at some specified significance level, and (ii) a realistic guess of the variability in the sample.

We model microbiome data with the Dirichlet-Multinomial (DM) distribution. It is described by two parameters: the overdispersion parameter ( $\theta$ ) which measures the within-sample excess of

variability *w.r.t.* a multinomial distribution; and the vector of relative abundances ( $\pi$ ). Depending on the research question each element of  $\pi$  refers to a single Operational Taxonomic Unit (OTU) or to a species, genus or any other rank in the microbial taxonomy. In many microbiome studies the null hypothesis can be expressed as  $H_0 = \pi_1 = \pi_2$ , where the indices 1 and 2 refer, respectively, to the controls and cases. The null hypothesis thus expresses equality of community composition. (La Rosa *et al.*, 2012) proposed to test this null hypothesis with the generalized Wald test of Koehler and Wilson (1986). For other applications researchers are interested in differential abundance for  $k \geq 1$  specific OTUs. The null hypotheses of interest are then  $H_0: \pi_{1j} = \pi_{2j}, j = 1, \dots, k$ , where the index  $j$  refers

to OTU  $j$ . Tests developed for RNA-Seq may here be used (McMurdie and Holmes, 2014), as well as the non-parametric Wilcoxon-Mann-Whitney (WMW) test, but correcting for multiple testing is required.

It is usually hard to specify the smallest relevant deviation from  $H_0$ , particularly when the  $\pi$ -vectors are large. In the web application,  $\pi_1$  (controls) is specified by the user in an interactive way with several possible choices, while  $\pi_2$  (cases) is based on  $\pi_1$  with some user-selected OTUs showing a modified relative abundance (referred to as the ‘requested’ OTUs). For the other OTUs (referred to as the ‘unrequested’ OTUs) the relative abundances are altered to make the entries in  $\pi_2$  add up to 1. It is also possible to estimate  $\pi_1$  and  $\theta$  from user-uploaded data. For ease of presentation both vectors  $\pi_1$  and  $\pi_2$  are ordered according to the decreasing order of relative abundances in  $\pi_1$ . Our application allows the simulation-based statistical power to be computed either for a specific pair of sample sizes (balanced or not, named ‘Power Option 1’ in the application), or for a range of sample sizes (only balanced, named ‘Power Option 2’ in the application). Powers for both the generalized Wald test for comparing community composition and the WMW test for comparing relative abundances of specific OTUs are implemented.

Sometimes biological samples can be classified into strata. A multicenter trial is a typical example, but stratification may also arise when age, gender or any other baseline information is used in the design stage to account for population heterogeneity. Power calculation for stratified designs is also implemented in the web application, and we illustrate its use with enterotypes as strata. Enterotypes can be described as microbiome profiles to which a sample can be said to belong; more details on the topic can be found in Arumugam *et al.* (2011). Although enterotypes have to be used with care as a stratification factor (they can only be identified after sequencing), power calculations are still correct as long as differences between cases and controls are not confounded by the enterotype. Moreover, over-sampling may be needed to ensure that the required number of subjects is acquired in the least frequent stratum, or when accounting for drop-out during the study.

## 2 Methods and implementation

Powers are calculated by means of a Monte Carlo approach in which, for given sample sizes, data of  $k$  OTUs are randomly generated from a DM distribution with the parameters specified in the interface. The total numbers of reads (library size) are randomly drawn with replacement from the library sizes observed in the HMP data (‘stool’ is default, but user can change it).

For testing equality of community composition the generalized Wald test gives a single  $P$ -value. Hence, its power is computed as the average number of rejections among the Monte Carlo-generated datasets. The WMW test, instead, is applied to each OTU individually and the resulting  $P$ -values are multiplicity adjusted with the Benjamini and Hochberg (1995) method, so as to control the False Discovery Rate. Here two quantities are reported: (i) the average number of times there was *at least one rejection* among the requested OTUs (requested by the user to be differentially abundant between the two groups), and (ii) the average power among the requested OTUs. The latter is interpreted as the expected power for an arbitrary requested OTU.

In case of stratification, since observations of different strata are mutually independent, the generalised Wald test can be performed for each stratum individually, and then the single-stratum statistics may simply be summed to form the overall test statistic. As each

single-stratum test statistic asymptotically has a  $\chi^2$  null distribution, the overall one has also a  $\chi^2$  null distribution, with degrees of freedom being the sum of the individual degrees of freedom. The same construction is used for combining the WMW test statistics (sum of squared standardized WMW statistics).

*Shiny* (RStudio Inc., 2016) was introduced by RStudio for developing interactive web applications from within R (R Core Team, 2015). It is mainly composed of two files: ui.R, in which the user interface is defined, and server.R that contains scripts producing the outputs. The generalized Wald test of the HMP package (La Rosa *et al.*, 2016) has been recoded to speed up computations, whereas for the WMW test we employed the ‘wilcox.test’ function of the standard R installation. The web page is divided in two sections: a sidebar on the left, where all parameters can be set, and the main page, which is further divided into tabs. The user can specify: stratification (‘yes’ or ‘no’); the ‘Shape’ of  $\pi_1$  with several choices, including all body sites from the HMP (‘Stool’ is the default), and possibility to use user-uploaded data; the number of OTUs to be considered ( $k$ ); the  $\theta$  parameter (only if ‘Shape’ is one of the theoretical ones; for HMP body sites and for user-uploaded data  $\theta$  is estimated); two sets of requested OTUs for which relative abundances in  $\pi_2$  are modified relative to  $\pi_1$  with a user-specified percentage for each of the two sets of OTUs; the number of Monte Carlo runs used for power calculations; the significance level ( $\alpha$ ); sample sizes of the two groups (for ‘Power Option 1’); range of sample sizes for the power *vs.* sample size graph (‘Power Option 2’); *include or not* the WMW test. A PDF report with all inputs and outputs of the application can be downloaded, including a list of the requested OTUs

## 3 Results

Using HMP stool samples data, we compared the powers of the generalized Wald test ( $\alpha = 0.10$ ) with and without stratification by enterotype. The sample size was set to 40 in both groups. The four most abundant OTUs of  $\pi_2$  were set 20, 20, 40 and 40% higher than those in  $\pi_1$ . The resulting power was much larger for the stratified Wald test, going from 0.008 *without* to 0.393 *with* stratification. The power of the WMW test went up from 0.145 to 0.94. These results were based on 1000 Monte Carlo simulations.

## 4 Conclusion

We developed a web interface that allows exploring the power/sample sizes behaviour in microbiome case-control studies with or without stratification. We demonstrated that our application can estimate power when stratified tests are applied.

**Funding:** We acknowledge support of IAP research network grant (P07/06) from the Belgian government (Belgian Science Policy) and of Ghent University (Multidisciplinary Research Partnership Bioinformatics: from nucleotides to networks). We also want to thank Doris Vandeputte who ran the enterotype detection pipeline on the HMP Stool data.

**Conflict of Interest:** none declared.

## References

- Arumugam, M. *et al.* (2011) Enterotypes of the human gut microbiome. *Nature*, 473, 174–180.
- Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc., Ser. B*, 57, 289–300.

- Koehler, K.J. and Wilson, J.R. (1986) Chi-square tests for comparing vectors of proportions for several cluster samples. *Commun. Stat. Theory Methods*, 15, 2977–2990.
- La Rosa, P.S. et al. (2012) Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLoS One*, 7, e52078.
- La Rosa, P.S. et al. (2016). HMP: hypothesis testing and power calculations for comparing metagenomic samples from HMP. R package version 1.4.2.
- McMurdie, P.J. and Holmes, S. (2014) Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.*, 10, e1003531
- R Core Team. (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- RStudio Inc. (2016). *shiny: Web Application Framework for R*. R package version 0.13.0.