BITS :: Call for Abstracts 2021 - Oral communication

Type	Oral communication
Session	Algorithms for Bioinformatics
Title	Investigating differential abundance methods in microbiome data: a benchmark study
All Authors	Cappellato M, Baruzzo G, Di Camillo B.
Affiliation	

Department of Information Engineering, University of Padova, Padova.

Motivation

The continuous improvement of high throughput DNA sequencing techniques has enhanced the possibility of studying complex microbial systems. Mining microbiome data, however, requires the development of specific computational methods to extract the information useful for analysing the micro-world of interest. Differential Abundance (DA) analysis in microbiome dataset is now a standard step in downstream analysis [1], focusing on the identification of specific taxonomic features (taxa) that significantly drive differences in microbial composition between experimental groups.
Methods adapted from the RNA-sequencing field were initially used for this type of investigation. Microbiome data show peculiar characteristics, both biological and technical, which motivated the development of new tools based on the compositional approach [2-6].
Despite several studies compare the performance of the DA methods [5, 7-9], there is a lack of investigation on the most recently developed approaches.
Here we exploit a generative model of microbiome synthetic data that takes into account its compositional nature and simulate different scenarios by combining all the possible covariates of interest while maintaining the main characteristics of the datasets, as an ultimate test-bed for the DA methods.

Methods

In this study we focus on established and recent DA methods developed for microbiome analyses (i.e., ALDEx2 [10], eBay [11], ANCOM [12], ANCOM-BC [13], corncob [14], metagenomeSeq [15]) and for differential expression analysis of RNA-seq data (i.e. edgeR [16] and DESeq2[17]).
As a benchmark to assess methods' performance we use simulated data generated by metaSPARSim [18], a recently published simulator able to resemble 16S sequencing data and estimate simulation parameters from real datasets. We simulate microbial count data starting from three real datasets [19-21] characterized by different library size, sequencing technology used, sparsity and amplified hypervariable regions, obtaining a wide range of scenarios. For each simulation, we generated differentially abundant taxa in groups of samples introducing a taxa fold change (FC) varying it in a predefined interval. metaSPARSim simulation procedure preserves the mean–dispersion relationship learned from real scenarios, thus preventing the risk of creating unrealistic abundance distributions.
For consistency with previous comparison [5, 7-9], for each simulated dataset we firstly investigate the effects of three covariates: percentage of DA taxa (5%, 10% and 20%), number of samples in the experimental groups (10, 25, 50 and 100) and library size (half or double of the original). In addition, since at low abundance DA features detection is a difficult task and large biological variability may affect method performance, we also assess methods' results by simulating more/fewer DA features in the low abundance range and changing taxa variability level.
For each dataset and covariate, we first evaluate tools performance in terms of false positive rate (FPR) control under the null hypothesis (i.e., without DA features). Then, recall and FDR are investigated along with their trade-off considered in the PR-curve and the area under PR-curve (AUPR). Information about running times completes the performance overview since computational time can play a key role in method choice (e.g., for large datasets).

Results

The literature reports that methods tend to have high FPR and, therefore, the FDR value higher than the desired threshold (namely 5%). However, in our study ALDEx2, eBay, ANCOM, and ANCOM-BC (followed by corncob and the baseline method Wilcox) do not follow this behaviour. Even at low sample size ALDEx2, eBay corncob and Wilcox, tends to stay below 5%. The proper control of FPR is maintained even when the DA features are simulated at low abundance. In addition, we verify that methods' performance tends to be robust to the library size parameter.
Unsurprisingly, recall is particularly influenced by the sample size. In scenarios with many low abundance features, many samples are needed to overcome 50% of recall. As expected, in the presence of low abundance DA features all methods reveal a noticeable decrease in power. In general, when the number of samples increase edgeR, DEseq, corncob and eBay reach a recall close to 75%.

Again expectedly, the variability simulation parameter significantly impacts the recall, but not the overall observed ranking of methods in methods performance.
Finally, all methods have low computational times, although corncob and ANCOM show a longer execution time as the sample size increases.
Despite there is no method that outperform the other in all the scenarios and covariates, compositional approaches exploited in ANCOM, ANCOM-BC, eBay, ALDEx2 achieve a good trade-off between precision and recall.

Info

References:
[1] Calle ML. Statistical Analysis of Metagenomics Data. Genomics Inform. 2019 Mar;17(1):e6. doi: 10.5808/GI.2019.17.1.e6.
[2] Gloor GB, Reid G. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. Can J Microbiol. 2016 Aug;62(8):692-703. doi: 10.1139/cjm-2015-0821.
[3] Gloor GB, Wu JR, Pawlowsky-Glahn V, Egozcue JJ. It's all relative: analyzing microbiome data as compositions. Ann Epidemiol. 2016 May;26(5):322-9. doi: 10.1016/j.annepidem.2016.03.003.
[4] Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional. Front Microbiol. 2017 Nov 15;8:2224. doi: 10.3389/fmicb.2017.02224.
[5] Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vázquez-Baeza Y, Birmingham A, Hyde ER, Knight R. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 2017 Mar 3;5(1):27. doi: 10.1186/s40168-017-0237-y
[6] Quinn TP, Erb I, Richardson MF, Crowley TM. Understanding sequencing data as compositions: an outlook and review. Bioinformatics. 2018 Aug 15;34(16):2870-2878. doi: 10.1093/bioinformatics/bty175.
[7] Hawinkel S, Mattiello F, Bijnens L, Thas O. A broken promise: microbiome differential abundance methods do not control the false discovery rate. Brief Bioinform. 2019 Jan 18;20(1):210-221. doi: 10.1093/bib/bbx104.
[8] Calgaro M, Romualdi C, Waldron L, Risso D, Vitulo N. Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. Genome Biol. 2020 Aug 3;21(1):191. doi: 10.1186/s13059-020-02104-1.
[9] Lin H, Peddada SD. Analysis of microbial compositions: a review of normalization and differential abundance analysis. NPJ Biofilms Microbiomes. 2020 Dec 2;6(1):60. doi: 10.1038/s41522-020-00160-w.
[10] Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome. 2014 May 5;2:15. doi: 10.1186/2049-2618-2-15.
[11] Liu T, Zhao H, Wang T. An empirical Bayes approach to normalization and differential abundance testing for microbiome data. BMC Bioinformatics. 2020 Jun 3;21(1):225. doi: 10.1186/s12859-020-03552-z.
[12] Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb Ecol Health Dis. 2015 May 29;26:27663. doi: 10.3402/mehd.v26.27663.
[13] Lin H, Peddada SD. Analysis of compositions of microbiomes with bias correction. Nat Commun. 2020 Jul 14;11(1):3514. doi: 10.1038/s41467-020-17041-7.
[14] Martin BD, Witten D, Willis AD. MODELING MICROBIAL ABUNDANCES AND DYSBIOSIS WITH BETA-BINOMIAL REGRESSION. Ann Appl Stat. 2020 Mar;14(1):94-115. doi: 10.1214/19-aoas1283.
[15] Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nat Methods. 2013 Dec;10(12):1200-2. doi: 10.1038/nmeth.2658.
[16] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139-40. doi: 10.1093/bioinformatics/btp616.
[17] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. doi: 10.1186/s13059-014-0550-8.
[18] Patuzzi I, Baruzzo G, Losasso C, Ricci A, Di Camillo B. metaSPARSim: a 16S rRNA gene sequencing count data simulator. BMC Bioinformatics. 2019 Nov 22;20(Suppl 9):416. doi: 10.1186/s12859-019-2882-6.
[19] Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012 Jun 13;486(7402):207-14. doi: 10.1038/nature11234.
[20] He X, Parenti M, Grip T, Lönnerdal B, Timby N, Domellöf M, Hernell O, Slupsky CM. Fecal microbiome and metabolome of infants fed bovine MFGM supplemented formula or standard formula with breast-fed infants as reference: a randomized controlled trial. Sci Rep. 2019 Aug 12;9(1):11589. doi: 10.1038/s41598-019-47953-4.
[21] Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Andrews E, Ajami NJ, Bonham KS, Brislawn CJ, Casero D, Courtney H, Gonzalez A, Graeber TG, Hall AB, Lake K, Landers CJ, Mallick H, Plichta DR, Prasad M, Rahnavard G, Sauk J, Shungin D, Vázquez-Baeza Y, White RA 3rd; IBDMDB Investigators, Braun J, Denson LA, Jansson JK, Knight R, Kugathasan S, McGovern DPB, Petrosino JF, Stappenbeck TS, Winter HS, Clish CB, Franzosa EA, Vlamakis H, Xavier RJ, Huttenhower C.

Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019 May;569(7758):655-662. doi: 10.1038/s41586-019-1237-9.	
Figure	
-	
Availability	-
Corresponding Author	
Name, Surname	Barbara, Di Camillo
Email	barbara.dicamillo@unipd.it
Submitted on	29.04.2021