



Multi-part strategy for testing differential taxa abundance in sequencing data: A simulation study with an application to a microbiome study

Daniela Cianci^{a,*}, Sebastian Tims^b, Guus Roeselers^b, Rachid El Galta^c, Sophie Swinkels^b

^a Julius Center for Health Sciences and Primary Care, Department of Data Science & Biostatistics, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands

^b Danone Nutricia Research, Utrecht, The Netherlands

^c Former employee of Danone Nutricia Research, Utrecht, The Netherlands

ARTICLE INFO

Keywords:

Microbiome
Multi-part strategy
Simulations
Taxon abundance testing

ABSTRACT

Comparing the microbiome across study arms is a recurrent goal in many studies. Standard statistical methods are often used for this purpose, however, they do not always represent the best choice in this context given the characteristics of microbiota sequencing data, e.g., non-negative, highly skewed counts with a large number of zeros.

A multi-part strategy, that combines a two-part test (as described by Wagner et al., 2011), a Wilcoxon sum-rank test, a Chi-square and a Barnard's test was explored to compare the taxa abundance between study arms. The choice of the test is based on the data structure. The type I error of the multi-part strategy was evaluated by using a simulation study and the method was applied to real data. The script to perform the analysis with the multi-part approach is provided in the statistical software SAS.

Several scenarios were simulated and in all of them the type I error was not inflated. Based on the statistical differences resulting from the two-part test (as described by Wagner et al., 2011) and the multi-part strategy (as proposed in this article), different biological implications can be extracted from the same comparison in the same data set.

In the comparison of taxa abundance between study arms, we showed that careful attention needs to be paid on the data structure, in order to be able to choose an appropriate analysis method. Our approach selects the most suitable test according to the type of data observed, maintains a good type I error and is easily applicable by using the SAS macro provided.

1. Introduction

Due to the increased application of high-throughput (amplicon) and shotgun sequencing in microbial ecology, the analysis of large sequence data sets has become commonplace, particularly in human microbiome research. However, the development of statistical tools necessary to accurately analyse the complex data generated and to test biological hypotheses is an area that requires more attention. High-throughput sequencing data typically consist of zeros and a few cases with values above zero. The total number of reads obtained for a sample does not reflect the absolute number of microbes present, since the sample is just a fraction of the original environment. Since the relative abundances sum to 1 and are non-negative, the relative abundances represent compositional data. The 'zeros' in the data do not necessarily represent

absence of certain microbial taxa; due to differential efficiency of the sequencing process, some rare microbial taxa might not be captured and thereby result in zero read counts. Sequencing data sets are inherently multi-dimensional with potentially thousands of microorganisms present in a single sample. Multiple testing is common, as researchers aim to compare the abundance of taxa detected between sample groups or study arms. Relatively small sample sizes are also a frequently observed limitation of microbiota studies given the costs or difficulty to collect samples.

Many studies aim at comparing the microbiome composition across study arms. Standard statistical methods, such as *t*-test or Wilcoxon test, are usually used to investigate differences in taxa or gene function abundances between two study arms (e.g., groups receiving different interventions) (Li, 2015; Xia and Sun, 2017). However, given the

* Corresponding author.

E-mail address: d.cianci@umcutrecht.nl (D. Cianci).

<https://doi.org/10.1016/j.mimeth.2023.106810>

Received 27 March 2023; Received in revised form 17 August 2023; Accepted 18 August 2023

Available online 20 August 2023

0167-7012/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

complex structure of the data, these methods are often not appropriate as the required assumptions are not always met or due to the sparsity level (large number of zeros). Two-part approaches constitute an improvement to the analysis of these data (Wagner et al., 2011; Gleiss et al., 2015). Specifically, the two-part method suggested by Wagner et al., 2011 combines two test statistics: the first test statistic compares the proportion of zeros between the two study arms (binomial), while the second statistic is used for comparing the mean or the median of the continuous part. Both the square of the test statistic for the binomial part and the square of the test statistic for the continuous part, asymptotically follow a Chi-square distribution with 1 degree of freedom (d.f.). If both components of the two-part test statistic are defined (i.e., the proportion is not 0 and not 1 in both samples), then an asymptotic *p*-value can be derived from a Chi-square distribution with 2d.f., including the continuity correction for small sample size. Otherwise, the undefined component is set to zero, and the resulting two-part statistic is assumed to follow a Chi-square distribution with 1d.f.. This two-part approach seems to work when the non-zero counts are >10 in both study arms, where non-zero count refers to the number of times a specific species is observed. For example, if we compare the microbiota in fecal samples obtained from 100 subjects within two study arms of equal size and if sequences affiliated to the bacterial genus *Bacteroides* are observed in samples from 20 subjects within group A and 40 subjects within group B, we will say that for the genus *Bacteroides* the zero counts are 30 and 10 in group A and B respectively, and the non-zero counts are 20 and 40. If the sample size is too small, the test statistics on the continuous part does not follow the normal approximation and the sum of the two test statistics will not produce a Chi-square (Wagner et al., 2011). However, in microbiota sequencing data, this situation is quite common and a large proportion of zero counts is usually observed. The gut microbiota is usually comprised of two types of bacterial taxa: a relatively small amount of highly abundant and prevalent taxa and a relatively large amount of taxa with low abundance and low prevalence. The distribution of both types of taxa contributes to unique microbiota composition (Ruan et al., 2020). This phenomenon is even more pronounced when the gut microbiome of infants is investigated, as infants experience a highly dynamic colonization succession in the first months of life (Wopereis et al., 2014). For this reason, we felt the need to expand the two-part method and make it more generalizable and applicable to high-throughput sequence data.

In this paper, we propose a multi-part strategy, that combines a two-part test (as described by Wagner et al., 2011), a Wilcoxon sum-rank test, a Chi-square and a Barnard's test, according to the distribution of each taxa. We investigated the hypothesis that the type I error of the multi-part method is not worse than the type I error of the two-part statistics. Additionally we explored the appropriateness of the continuity adjustment and how the two methods work on a real data example. We recommend that, in the comparison of taxa abundances between study arms, it is essential to first decide which test should be used based on the amount of zero and non-zero counts. In addition, with the use of simulations we investigated the type I error (i.e., the possibility to draw false positive conclusions and thereby falsely assuming bacterial taxa to be different between study groups) of the proposed approach and assessed the impact of the continuity adjustment for small sample sizes per study arm. An example in which this approach is applied on a real data set is also shown and the SAS program used for the analysis is provided.

2. Materials and methods

2.1. Multi-part method

The method we propose allows testing pairwise differential taxa abundance in sequencing data. It combines a two-part test, a Wilcoxon sum-rank test, a Chi-square and a Barnard's test, according to the abundance distribution of each taxa. The decision process is based on

the amount of *observed* zero and non-zero counts in each of the two study arms and the number of *expected* zero and non-zero counts in a 2×2 tables. The expected values are calculated under the null hypothesis of no association between study arms and the presence/absence of a taxon. For example, in a sample of 100 subjects with two study arms of equal size, if *Bacteroides* are present in 60 out of the 100 samples then expected non-zero counts and the expected zero counts are 30 and 20, respectively, in both study groups (see Table S1). In some situations, depending on the test resulting from the decision process, the relative abundance data of the taxa are used while in other cases the data are dichotomized in presence and absence data.

The microbiota data can be classified in five possible categories: i) Highly abundant, when the expected zero counts for at least one study arm is <5 expected whereas both study arms have ≥ 5 expected non-zero counts; ii) Semi-abundant, when all expected counts ≥ 5 and both study arms have ≥ 10 observed non-zero counts; iii) Low abundant, when all expected counts ≥ 5 and at least one study arm has <10 observed non-zero counts; iv) Infrequent, when at least one study arm has <5 expected non-zero counts and v) Absent, when the taxon was not observed in any subjects (Table S2). The multi-part test consists of five possible options, each option corresponding to one of the possible data categories described above. The multi-part strategy is explained below and summarized in Fig. 1.

Wilcoxon rank-sum test. The Wilcoxon rank sum test is used to determine whether the two samples are likely to derive from the same populations (i.e., the two populations have the same shape; Wilcoxon, 1945). It is performed when the count of *expected* zero is, in at least one arm, <5 and, at the same time, both *expected* non-zero counts are >5 (highly abundant data).

Two-part statistic. In the two-part test, first the proportions of zeros in the two study arms are compared (binomial part), and then the medians of the non-zero data in the two study arms are compared (continuous part). A global *p*-value is produced as output. This approach is used when the *observed* non-zero counts are greater or equal than 10 in both study arms (Wagner et al., 2011) (semi-abundant data). In the context of the multi-part strategy, we will refer to the two-part statistics not as the complete strategy described by Wagner et al., 2011, but only to the situation where both parts of the two-part statistics are defined.

Chi-square test. The Chi-square test is used to compare the proportion of zeros in the two study arms (Agresti, 2007). Data are dichotomized and the Chi-square test is applied if at least one of the *observed* non-zero counts is <10 and if all *expected* counts are >5 (low abundant data).

Barnard's test. Barnard's test is performed to compare the proportion of zeros in the two study arms when the number of counts are small. An unconditional tests such as the Barnard's test was chosen when the number of counts were small because it is considered generally more powerful than conditional tests and Barnard's test has been shown to have the highest power among the unconditional tests (Lydersen et al., 2009). If at least one *expected* non-zero count is <5, data are dichotomized and the Barnard's test is performed (infrequent data).

No test. No test is performed when only zero counts are *observed*, which can occur for rare microbial taxa (taxon absent).

The SAS program used for the analysis, including the correction for multiple testing, is presented in the Supplemental material (Supplemental material 1: SAS code).

Because of the large number of taxa and therefore of comparisons, correction for multiple testing is applied. Multiple testing error is handled by controlling the positive false discovery rate (pFDR) (Benjamini and Hochberg, 1995, J.D. Storey, 2003). The *q* value provides a measure of each feature's significance, taking into account the fact that hundreds of microbial taxa are simultaneously being tested. *q* values are calculated by estimating the proportion of true null hypothesis tests (π_0) among all tests. The bootstrap method described by Storey et al. (2004) is used to estimate π_0 . The bootstrap procedure is chosen because it appears to perform reasonably well, even when the alternative hypotheses are correlated (Tsai et al., 2003), as it is the case in our context.

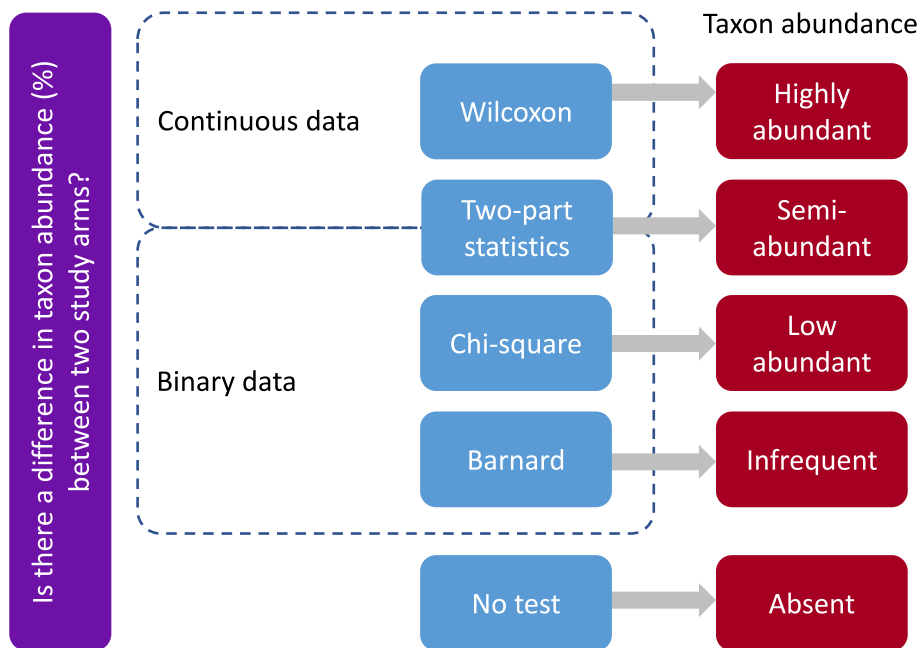


Fig. 1. Overview of the multi-part strategy choices. Taxon abundance is classified as: i) Highly abundant, when at least one study arm has <5 expected zero counts whereas both study arms have ≥ 5 expected non-zero counts of; ii) Semi-abundant, when all expected counts ≥ 5 and both study arms have ≥ 10 observed non-zero counts; iii) Low abundant, when all expected counts ≥ 5 and at least one study arm has <10 observed non-zero counts; iv) Infrequent, when at least one study arm has <5 expected non-zero counts and v) Absent, when the taxon was not observed in any subjects. For each category a specific test is proposed.

2.2. Simulation study

Simulations were used to compare the type I error (i.e., the possibility to draw false positive conclusions) of the two-part test (suggested by Wagner et al., 2011) and the multi-part method (proposed in this article). Through simulations it was checked whether the type I error was not inflated, i.e., the probability of rejecting the null hypothesis when it is actually true does not increase (thereby falsely assuming differences in bacterial taxa between study groups). Continuity adjustment is usually applied when a discrete distribution is approximated by a continuous distribution; an example is when the sample size is small. Currently, the continuity adjustment is always applied in the two-part method (Wagner et al., 2011) with small sample sizes and in this article it was evaluated whether this is always needed by comparing the type I error of the two-part test statistic with continuity adjustment versus without continuity adjustment.

Four types of taxa parameters were considered: i) a taxon was observed in $>95\%$ of subjects; ii) a taxon was observed in 50% of the subjects (and it was not detected in the remaining subjects); iii) a taxon was observed in 5–10% of the subjects; and iv) a taxon was observed in $<5\%$ of the subjects. It was assumed that each of the four taxa types (t_1 , t_2 , t_3 and t_4) occurred in the population with a frequency of w_i , under the constraint

$$\sum_i w_i = 1 \quad (1)$$

with $i = 1, 2, 3, 4$.

The prevalence of each taxon in the population was indicated with p_i and the expected relative abundance with r_i . With L_{\min} and L_{\max} the expected minimum and maximum total counts in a stool sample were specified, that is the library sizes. The sample size of each study arm is indicated by N_k , with $k = 1, 2$. The parameters used in the simulations are summarized in Table S3.

Six different scenarios were simulated. For each scenario 10,000 independent datasets were simulated. Each dataset consists of abundances of a random taxon drawn from a mixture of the four taxa types (t_1 , t_2 , t_3 and t_4) according to pre-specified weights w_1 , w_2 , w_3 and w_4 .

For each dataset with taxa type t_i the abundances were randomly generated as follows. For each subject 1) the library size L was randomly drawn from a uniform distribution between L_{\min} and L_{\max} 2) a binary

outcome X from a Bernoulli distribution was generated with probability p_i , where p_i is the prevalence of non-zero counts for the sampled taxa type t_i in the population. If $X = 0$ then the abundance count was set to zero, $y = 0$, otherwise the abundance count y was randomly sampled from a binomial distribution with size L and probability $= r_i$, where r_i is the expected relative abundance of the sampled taxa type in the population. The relative abundance is calculated by dividing the abundance count (y) by the library size (L).

Throughout the whole simulation the following parameters were fixed: $L_{\min} = 17,542$, $L_{\max} = 273,346$, $p_1 = 0.04$, $p_2 = 0.08$, $p_3 = 0.50$, $p_4 = 1.00$, $r_1 = 0.01$, $r_2 = 0.02$, $r_3 = 0.15$ and $r_4 = 0.10$. N was set to 50 when the type I error of multi-part testing strategy for the six different taxa scenarios was evaluated. Additional simulations were run: i) to compare the type I errors of the multi-part statistic with and without the continuity adjustment using a sample size of 10 subjects per study arm; and ii) to compare the type I error of the two-part test statistic versus multi-part testing strategy, without continuity adjustment. For the latter comparison, taxa data were simulated using $N = 50$ per study arm in two scenarios: one with mainly infrequent data and specifically with the following frequencies $w_1 = 0.07$, $w_2 = 0.13$, $w_3 = 0.1$, $w_4 = 0.7$; one with a less skewed distribution, i.e., $w_1 = 0.1$, $w_2 = 0.3$, $w_3 = 0.1$, $w_4 = 0.5$.

After the simulated dataset has been created, the differential abundance between study arms was tested based on the multi-part strategy. p -values were therefore obtained and, for the nominal significance levels $\alpha = 0.01$, 0.05 and 0.1 , the type I error was calculated as the proportion of p -values less or equal than α . For some scenarios the two-part statistics (Wagner et al., 2011) was performed and the type I error was compared to the one of the multi-part strategy.

2.3. Real data set

The two-part test and the multi-part method were also empirically compared using real data from the Mercurius clinical study (L. M. Breij et al., 2019) registered in the Dutch Trial Register as NTR3683 (www.trialregister.nl). In this study, healthy term born fully formula fed infants were randomised to one of two arms a Test arm in which the formula was comprised of large, milk phospholipid-coated lipid droplets (Nuturis®) containing a dairy-vegetable lipid mixture, and a Control arm which received the formula a standard vegetable oil-based formula.

Both intervention formulas were supplemented with a specific prebiotic mixture consisting of short-chain galacto-oligosaccharides (scGOS) and long-chain fructo-oligosaccharides (lcFOS) mixture in a 9:1 ratio at a 0.8 g/100 mL level (Fig. S1). A group of exclusively breastfed infants served as a reference. Stool samples were collected at baseline, 3 months and 12 months of age and the microbiota composition was determined by 16S rRNA gene amplicon sequencing. DNA extraction from the fecal samples, followed by the sequencing of the V3-V5 region of the 16S rRNA gene and the bioinformatic analysis of the sequence reads was performed as described previously (van den Elsen et al., 2019). For each stool sample the amount of each taxon was measured by sequence counts. Relative abundance was derived from the absolute abundance because it is more reasonable to draw inference regarding the abundance of a taxon in the ecosystem using its relative abundance in the specimen (Mandal et al., 2015). In the Mercurius study, the differences in the microbiota at baseline (0–35 days of age) are expected to be less pronounced between study arms as the microbiota is in a very early stage (L. M. Breij et al., 2019). For the 12-month visit it can be expected that all infants are weaned and therefore are receiving a high variety of solid foods, not related to the study arm they are in. At visit 4, at which the infants are three months of age and are either exclusively formula fed (Test or Control formula) or exclusively breastfed, we expect to find the most study product driven differences. Therefore, in this article we focus on describing the results of visit 4.

The taxa parameters were classified into five classes according to their presence/absence in a 2×2 contingency table for two-samples comparison and at the same time accordingly to the five possible choices described in *multi-part method* (see Fig. 1): i) Highly abundant; ii) Semi-abundant; iii) Low abundant; iv) Infrequent and v) Absent. For a detailed description of this classification please refer to *Materials and methods Real data*. Both two-part and multi-part methods were applied to Mercurius study data.

3. Results

3.1. Simulation study

Table 1 summarizes estimates of Type I error for six different scenarios using a sample size of 50 subjects per study arm. Scenarios I–IV represent situations where abundance counts are 100% according to taxa type 1, 2, 3, and 4, respectively. Simulated taxa data in scenarios V consist of 7% type 1, 13% type 2, 10% type 3 and 70% type 4, resembling the taxa data structure observed in the Venus study (L. Breij, 2016; Shek, 2017). In scenarios I, II, III, IV and V (1) the multi-part strategy for testing differential abundance between study arms was applied without continuity adjustment of the two-part statistic, while the continuity adjustment was computed in scenario V (2). V (1) and V (2) show therefore a comparison of the multi-part statistic with and without

continuity adjustment in the two-part statistics.

It was checked whether the test chosen by the strategy to analyse the data was matching with the simulation characteristics. The *p*-values for testing differential abundance between study arms in scenarios I, II, III and IV following the multi-part strategy corresponded predominately to Wilcoxon statistics, two-part, Chi-square and Barnard's respectively (data not shown) meaning that scenario I corresponds mainly to highly abundant taxa, scenario II to semi-abundant taxa, scenario III to low abundant taxa and scenario IV to infrequent taxa. In the Type I Error columns the significance level (or nominal type I error), *alpha*, was set to 1%, 5% and 10%, indicating that the risk of concluding that a difference exists while there is no actual difference is set at no >1%, 5% and 10% respectively. For each simulated scenario it was calculated how often a statistically significant difference was found, at the different *alpha* levels. Overall, the observed type I error did not exceed the *alpha* level indicating that the multi-part strategy provided a reasonable type I error, regardless to the data characteristics and consequently the test chosen. However, for sparse data (see scenario IV) the approach seemed to be rather conservative as the Barnard's test was predominantly used. The simulation results show that the two-part test statistic without the continuity adjustment had a good type I error even for small sample sizes as 10 subjects per study arm. In Fig. 2, the reference line indicates a perfect match between the nominal type I error and the actual (observed) type I error. The actual type I error without continuity correction (yellow line) is very close to the reference line, while the nominal type I error of the test with continuity adjustment (blue line) is always greater than the actual type I error. This indicates that the two-part test statistic with a continuity adjustment had a conservative type I error, meaning that *p*-values tend to be too high.

The type I error of the multi-part testing strategy and of the two-part test statistic were compared in the situation of many zeros observed (0.7 and 0.5 Barnard's data type, i.e., 70% and 50% of the taxa is present in <5% of the subjects). The statistic of the suggested multi-part strategy had a good type I error and outperformed the two-part test statistic (Wagner et al., 2011), as it is shown in Fig. 3. In Fig. 3 the type I error of the two-part test statistic versus the multi-part testing strategy (without continuity adjustment) is compared for two simulated scenarios. The improvement of the type I error of the multi-part compared to the two-part approach, is particularly clear in the left part of each plot, where the nominal type I error is between 0 and 0.1, since values in the range 0.05 to 0.1 are usually used as significance levels in hypothesis testing.

3.2. Real data

In the Mercurius study, for most taxa, several zeros are observed (absent, infrequent and low abundant; Table 2) due to the high inter-individual variation of gut bacteria between subjects and the early age of the subjects. The other taxa are more common and very few zero are

Table 1
Type I error of multi-part testing strategy for different taxa scenarios.

Scenario	Proportion of simulated taxa type				Type I Error at <i>alpha</i>			adjustment
	W ₁	W ₂	W ₃	W ₄	1%	5%	10%	
I	1	0	0	0	0.009	0.045	0.094	no
II	0	1	0	0	0.010	0.051	0.106	no
III	0	0	1	0	0.008	0.048	0.088	no
IV	0	0	0	1	0.003	0.034	0.084	no
V (1)	0.07	0.13	0.1	0.7	0.008	0.047	0.098	no
V (2)	0.07	0.13	0.1	0.7	0.004	0.035	0.079	yes

W_i indicates the frequency of the occurrence of each taxon parameter in the population ($\sum w = 1$). Four types of taxa parameters were considered: 1) a taxon was observed in >95% of subjects; 2) a taxon was observed in 50% of the subjects (and it was not detected in the remaining subjects); 3) a taxon was observed in 5–10% of the subjects; and 4) a taxon was observed in <5% of the subjects. The type I error is calculated for each scenario at three *alpha* levels: 1%, 5% and 10%. Adjustment indicates whether the continuity adjustment was computed or not. *N* = 50. Details on the simulation parameters are provided in section *Materials and methods – Simulation study*.

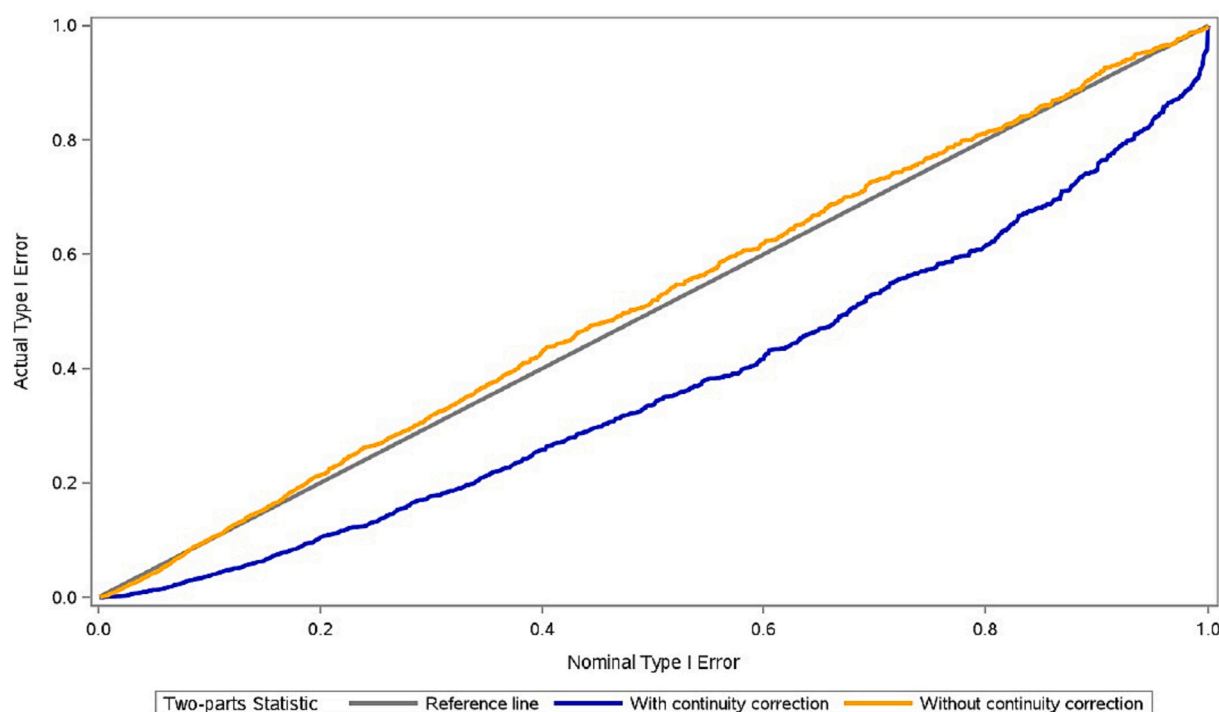


Fig. 2. Type I error of two-part test statistic with and without continuity adjustment for small sample sizes ($N = 10$ per study arm).

observed (semi and highly abundant; Table 2). As described above, the focus will be on visit 4. At this visit, a big fraction of the taxa is infrequent. Example distributions for taxa, one for each abundance category, from the real data set is shown in Fig. 4.

All p -values and q -values of both methods are shown in Table S4 for the Mercurius study data set. In this data set enrichment of commensals such as *Bifidobacterium* and *Bacteroides* was associated with the presence of human milk oligosaccharides (HMOs) and/or with the specific prebiotic scGOS/lcFOS mixture. P -values, between the study arms, for taxa that were detected in the majority of the samples were not excessively impacted by the method used, since the two methods basically overlap for more common taxa. Several taxa that are infrequent or low abundant, however, did show large differences in the p -values produced by the two methods (Table S4; Table 3). For instance, *Fusobacterium* seems to be statistically significantly different and more prevalent at three months of age in the infants in the Breastfed arm compared to the Control arm when using the two-part statistics ($p = 0.020$), while the *Proteus* genus seems the least prevalent in the infants in the Test arm compared to the Control arm ($p = 0.018$). However, the p -values resulting from the multi-part strategy for the *Fusobacterium* and *Proteus* genera, in the same comparisons do not implicate any statistically significant difference for these taxa ($p = 0.339$ and $p = 0.433$, respectively). Although in this particular study the comparison from which the difference in *Proteus* presence would not have been regarded an actual finding due controlling for the pFDR ($q = 0.149$), this could have been different in other comparisons or studies if the p -value distribution over all taxa would have been different.

The multi-part strategy results indicate that the *Anaerococcus* genus and even the highly abundant *Veillonella* genus were statistically significantly different, taking pFDR into account, between Test and the Breastfed subjects. Moreover, the multi-part strategy uncovers several infrequent or low abundant taxa to be different in both Test versus breastfed and Control versus Breastfed comparisons, such as *Scardovia*, *Shewanella*, *Peptostreptococcaceae* Other, and *Halomonas*. Furthermore, several infrequent taxa, such as *Lactococcus*, *Dermabacter*, *Peptoniphilus*, and *Campylobacter*, as well as the low abundant taxon *Enterobacter* were found to be different in Control versus Breastfed comparison only

(Fig. S2).

4. Discussion

The type I error of the multi-part strategy has been evaluated by using a simulation study and acceptable results have been observed. Investigating the type I error is relevant because the strategy we propose envisage a choice among several tests to compare taxa abundance between study arms. The simulations included scenarios where the complete multi-part strategy was needed and scenarios where the multi-part approach corresponded predominately to one component of the strategy (Barnard's, Chi-square, two-part or Wilcoxon statistics). For all possible choices of the multi-part strategy, it was shown that the nominal type I error was still maintained. The results of our method were quite conservative when the Barnard's test, used when the data are extremely sparse, was predominantly applied. When differences in the all microbiome are investigated – without taxa specific a priori hypothesis – the context is usually explorative and therefore it may be preferred to choose for a higher level of α , especially in this situation where the test is pretty conservative. The condition in which many zeros are observed was further investigated and the type I error of the multi-part testing strategy and of the two-part test statistic were compared. The multi-part strategy outperformed the two-part test statistic in terms of type I error.

The impact of the continuity adjustment for a small sample size per study arm has been assessed as well in the simulation study. The type I error for the two-part test statistic without the continuity adjustment was good even for small sample sizes, whereas it was too conservative for the two-part test statistic with a continuity adjustment. Therefore, we conclude that it is preferable to not apply the continuity adjustment in the two-part statistics of the proposed multi-part method.

In the SAS code provided to perform the complete analysis, correction for multiple testing is applied because of the large number of comparisons. Multiple testing error is handled by controlling for the positive false discovery rate (pFDR) (Benjamini and Hochberg, 1995; J. D. Storey, 2003). pFDR relies on the assumption that p -values are uniformly distributed under the null hypothesis. In this paper we showed

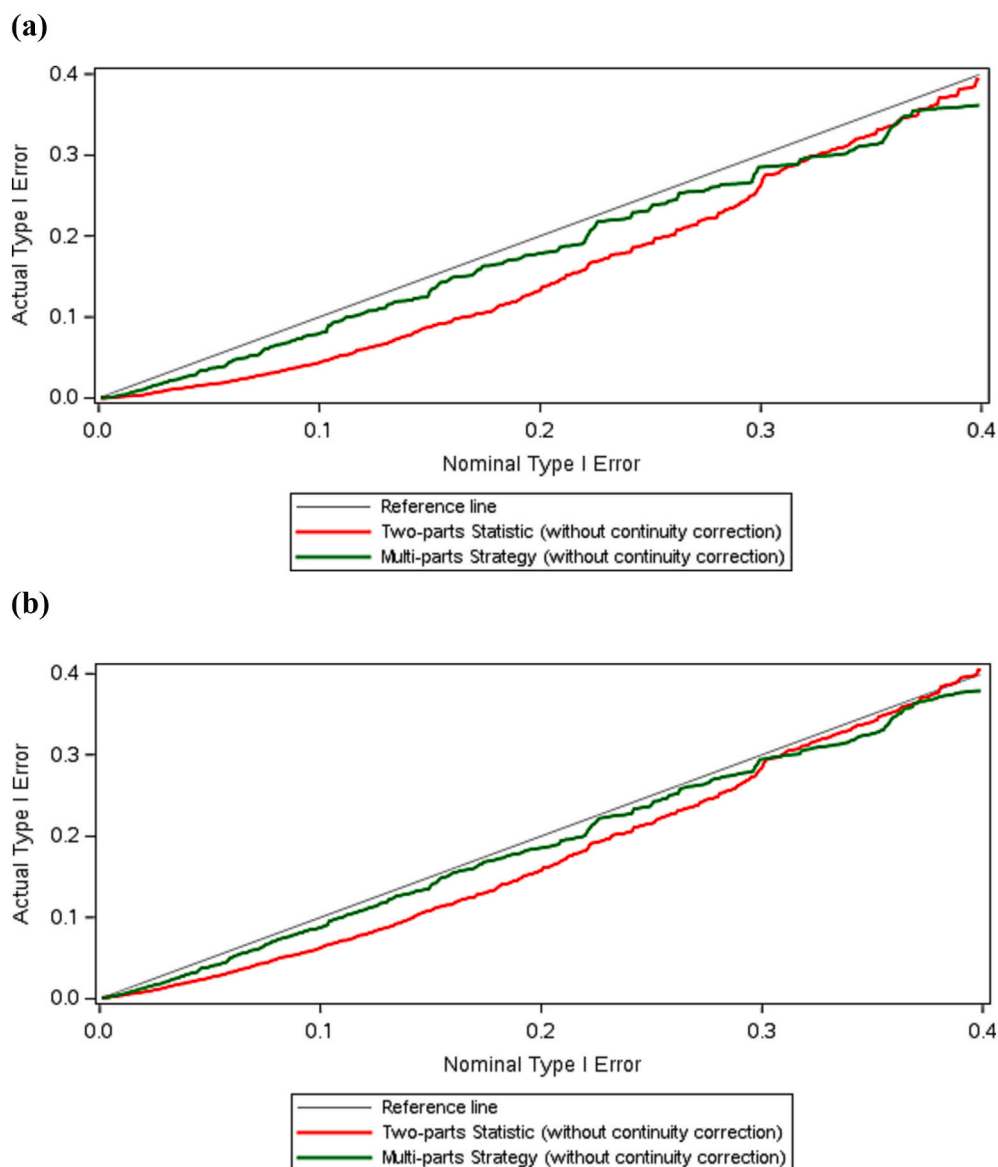


Fig. 3. Type I error of two-part test statistic versus multi-part testing strategy, without continuity adjustment. Taxa data were simulated using $N = 50$ per study arm and a) $w_1 = 0.07$, $w_2 = 0.13$, $w_3 = 0.1$, $w_4 = 0.7$; b) $w_1 = 0.1$, $w_2 = 0.3$, $w_3 = 0.1$, $w_4 = 0.5$.

Table 2

Distribution of bacterial genus level taxa by comparison and visit.

		Frequency of non-zeros abundance/OTUs										All
		Highly Abundant ¹		Semi abundant ²		Low abundant ³		Infrequent ⁴		Absent ⁵		
		N	%	N	%	N	%	N	%	N	%	N
Study visit 1	Comparison C versus B	5	3.1	19	11.8	15	9.3	81	50.3	41	25.5	161
	T versus B	5	3.1	20	12.4	15	9.3	79	49.1	42	26.1	161
	T versus C	4	2.5	26	16.1	20	12.4	73	45.3	38	23.6	161
4	C versus B	8	5.0	30	18.6	26	16.1	70	43.5	27	16.8	161
	T versus B	8	5.0	28	17.4	32	19.9	64	39.8	29	18.0	161
	T versus C	7	4.3	44	27.3	26	16.1	56	34.8	28	17.4	161
6	C versus B	20	12.4	32	19.9	20	12.4	68	42.2	21	13.0	161
	T versus B	18	11.2	33	20.5	23	14.3	63	39.1	24	14.9	161
	T versus C	19	11.8	35	21.7	23	14.3	59	36.6	25	15.5	161

1) Highly abundant, when at least one study arm has <5 expected zero counts whereas both study arms have ≥ 5 expected non-zero counts; 2) Semi-abundant, when all expected counts ≥ 5 and both study arms have ≥ 10 observed non-zero counts 3) Low abundant, when all expected counts ≥ 5 and at least one study arm has <10 observed non-zero counts 4) Infrequent, when at least one study arm has <5 expected non-zero counts; and 5) Absent, when the taxon was not observed in any subjects. T: Test arm; C: Control arm and B: Breastfed arm.

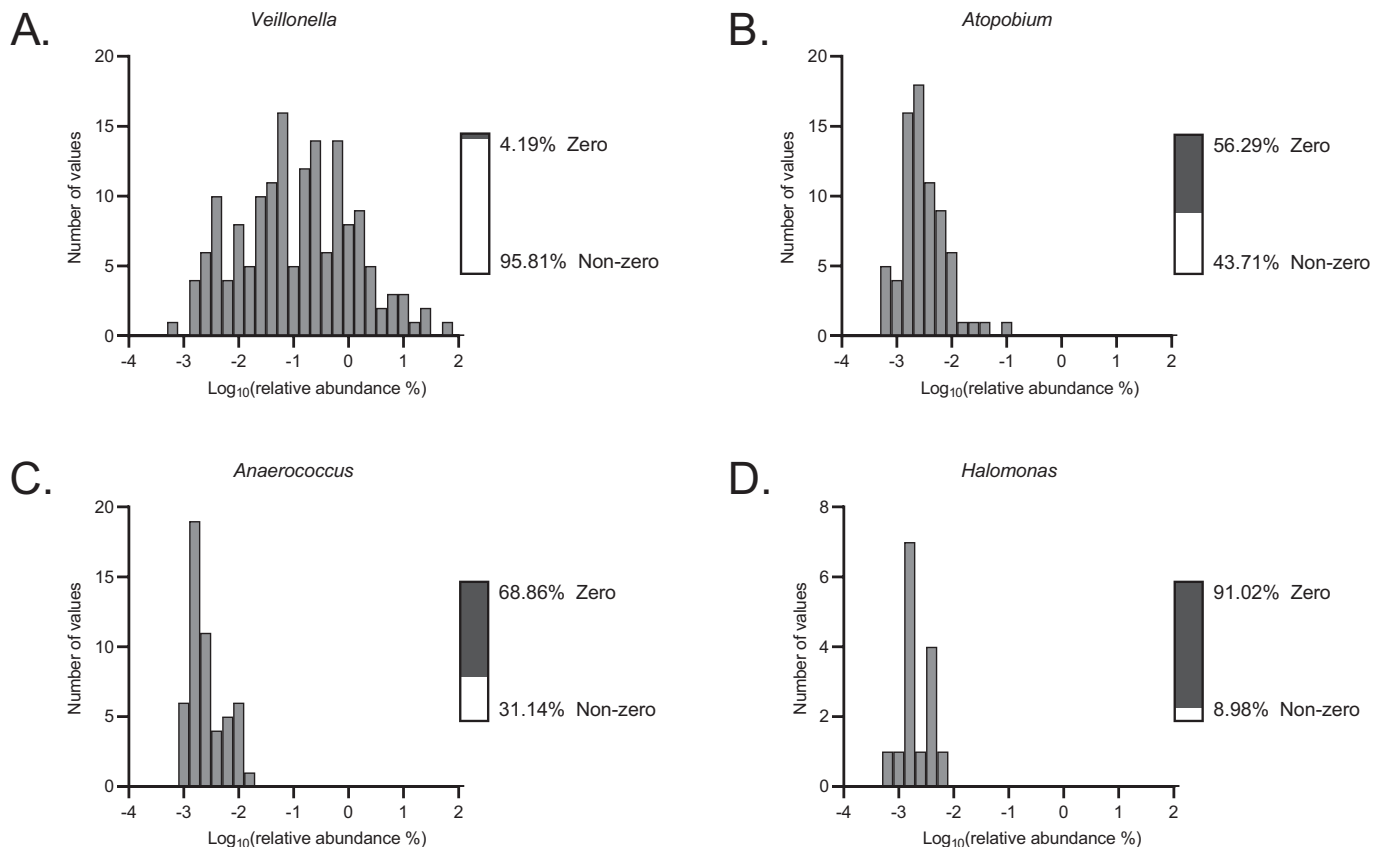


Fig. 4. Histograms of exemplary taxa from the Mercurius study, one for each abundance category: A) Highly abundant; B) Semi abundant; C) Low abundant; D) Infrequent.

that the proposed strategy preserves a reasonable type error in general and hence the uniformity assumption may be reasonable, even in presence of sparse data taxa (with many zeros), in contrast to other existing methods.

From the real data example, the similarities and differences seem to relate to the abundances of the taxa with differential outcomes between the two-part statistics and the proposed multi-part method. As infants in both infant formula arms were consuming the specific scGOS/lcFOS mixture the largest effects were expected to be found between either formula arm versus the Breastfed arm. Enrichment of the more abundant and/or prevalent commensals such as *Bifidobacterium* and *Bacteroides* was associated with the presence of human milk oligosaccharides (HMOs) and/or with the specific prebiotic scGOS/lcFOS mixture, which is not surprising and in accordance with several previous studies. Indeed, these taxa were detected in the majority of the samples and the chosen statistical test hardly lead to differences in interpretation. Several rare or low abundant taxa, however, did show differences in their significance according to the chosen statistical test. Although such taxa might not be present in all subjects in each study arm, their results can nevertheless invoke new hypotheses or research directions when these taxa are found to be rarer or even completely absent in certain subject populations. Especially when previously reported studies highlight an important biological role for a certain taxon. For instance, *Fusobacterium* seems to be more prevalent at three months of age in the infants in the Breastfed arm compared to the Control arm when using the two-part statistics. This bacterial genus harbours the species *Fusobacterium nucleatum*, which has been shown to human inhibitory receptor TIGIT (Gur et al., 2015), which could steer the biological interpretation of these result towards immunological implications. Moreover, the *Proteus* genus, which could harbour the species *Proteus mirabilis* that is known to interact with monocytes upon intestinal injury (Seo et al.,

2015), seems the least prevalent in the infants in the Test arm compared to the Control arm and therefore hints at another immunological aspect. However, when using the multi-part strategy, the *Fusobacterium* and *Proteus* genera would not be considered to be different between any of the study arms, hence not suggesting a difference in immunological response upon consuming these infant formulas.

The multi-part strategy results would put more emphasis on the metabolic activity of the microbes in relation to the anaerobic environment of the gut. Both the *Anaerococcus* genus and the *Veillonella* genus were found to be statistically significantly different between Test and the Breastfed subjects when using the multi-part strategy. Species from these genera have previously been linked to decreasing oxygen levels and suggested to reflect a switch towards more butyrate production and less lactate production (Bäckhed et al., 2015). Moreover, the multi-part strategy uncovers more rare low abundant taxa to be different in both Test versus breastfed and Control versus Breastfed comparisons. Which would affect the focus of best-after-breast type of research, when a microbiota modulating effect is considered to be an important factor. For abundant taxa the test used in both the two-part statistics and the multi-part strategy, will basically coincide, while the choice of the test to be used will be different for low abundant taxa. Results will therefore differ between the tests and could lead to a different aim or hypothesis for new research, i.e., focus of future work could be put on different microbial species depending on the initial analysis method chosen. The biological implications that could be derived from the two methods would lead to different follow-up of such a microbiota investigation. From a biological point of view, it is hard to decide which method produces more reliable results as differential outcomes are mainly with regards to infrequent or low abundant taxa and more independent studies with same dietary interventions would be needed to confirm the findings. Based on the results of the simulation study, we recommend

Table 3

P-values, for comparisons between the study arms, for taxa that showed differences in the *p*-values produced by the methods applied: two-part test and the multi-part strategy. A difference was defined as followed: the *p*-value in one of the methods is below the α of 0.05, while the *p*-value in the other method is above the α of 0.05. Taxa names in red indicates that the difference remains after controlling for pFDR (i.e., the lower *p*-value has a corresponding *q*-value of <0.05). Blue to red indicates low to high *p*-values.

Comparison	Genus	Two-part <i>p</i> -value	Multi-part <i>p</i> -value
Control vs Breastfed	<i>Campylobacter</i>	0.056	0.031
	<i>Dermabacter</i>	0.092	0.031
	<i>Enterobacter</i>	0.055	0.024
	<i>Fusobacterium</i>	0.02	0.339
	<i>Halomonas</i>	NA	0
	<i>Lactococcus</i>	NA	0.001
	Other	NA	0.001
	<i>Peptoniphilus</i>	0.059	0.018
	<i>Scardovia</i>	NA	0
	<i>Shewanella</i>	NA	0
Test vs Breastfed	<i>Acinetobacter</i>	0.027	0.169
	<i>Faecalibacterium</i>	0.084	0.028
	<i>Granulicatella</i>	0.101	0.041
	<i>Halomonas</i>	NA	0
	<i>Incertae Sedis</i>	0.099	0.04
	<i>Lactococcus</i>	NA	0.021
	Other	NA	0.002
	<i>Peptoniphilus</i>	0.081	0.029
	<i>Proteus</i>	0.019	0.054
	<i>Ralstonia</i>	NA	0.034
	<i>Scardovia</i>	NA	0
	<i>Shewanella</i>	NA	0
	<i>Varibaculum</i>	0.025	0.084
Test vs Control	<i>Aquabacterium</i>	0.074	0.025
	<i>Enterobacter</i>	0.085	0.027
	<i>Flavonifractor</i>	0.076	0.031
	<i>Fusobacterium</i>	0.063	0.028
	<i>Gordonibacter</i>	NA	0.008
	<i>Granulicatella</i>	0.035	0.051
	<i>Proteus</i>	0.018	0.433
	<i>Roseburia</i>	0.126	0.042
	<i>Stenotrophomonas</i>	0.034	0.109
	<i>Subdoligranulum</i>	0.008	0.066

using the multi-part strategy, since it outperformed the two-part test statistic in terms of type I error.

In conclusion, the multi-part strategy without continuity adjustment was demonstrated to be an appropriate and pretty straightforward method for testing differential taxa abundance in sequencing data, especially for low abundant taxa. However, this strategy allows only to test for pairwise comparison at one time point. An expansion of this method that would allow to compare multiple groups and several time points would be very useful.

In this paper, we show that, in the comparison of taxa abundance between study arms, careful attention needs to be paid on the data structure, in order to be able to choose an appropriate analysis method. Applying the same test to compare all taxa may result too simplistic. A two-part approach is an improvement compared to *t*-test or Wilcoxon test but may produce too many false positive results. The multi-part method we propose overcomes these limitations. Our approach selects

the most suitable test according to the type of data observed, maintains a good type I error and is easily applicable by using the SAS macro provided in the Supplemental material.

Authors' contributions

DC and SS conceived the ideas and designed methodology; REG improved the methodology, developed and performed the simulation plan; ST supervised the data collection; REG and ST analysed the data; DC and ST led the writing of the manuscript; GR and SS revised the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Daniela Cianci reports article publishing charges was provided by Danone Nutricia Research. Sebastian Tims, Guus Roeselers and Sophie Swinkels reports financial support was provided by Danone Nutricia Research. Sebastian Tims, Guus Roeselers and Sophie Swinkels reports a relationship with Danone Nutricia Research that includes: employment.

Data availability

Data will be made available on request.

Acknowledgements

The authors would like to thank the Danone Nutricia Research statistical programmers involved in this projects for their valuable support, Marieke Berkeveld-Abrahamse for the helpful comments on the manuscript and the MERCURIUS study team.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.mimet.2023.106810>.

References

- Agresti, A., 2007. *An Introduction to Categorical Data Analysis*. John Wiley & Sons Hoboken.
- Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., Jun, W., 2015. Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe* 17 (5), 690–703. <https://doi.org/10.1016/j.chom.2015.04.004>.
- Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Ser. B (Methodol.)* 57, 289–300. <https://doi.org/10.2307/2346101>.
- Breijl, L., 2016. An innovative infant formula with large, phospholipid-coated lipid droplets supports an adequate growth in healthy, term infants. In: *Nutrition and Growth Congress*.
- Breijl, L.M., Abrahamse-Berkeveld, M., Vandenplas, Y., Jespers, S.N.J., De Mol, A.C., Khoo, P.C., Hokken-Koelega, A.C.S., 2019. An infant formula with large, milk phospholipid-coated lipid droplets containing a mixture of dairy and vegetable lipids supports adequate growth and is well tolerated in healthy, term infants. *Am. J. Clin. Nutr.* <https://doi.org/10.1093/ajcn/nqy322>.
- Gleiss, A., Dakna, M., Mischak, H., Heinze, G., 2015. Two-group comparisons of zero-inflated intensity values: the choice of test statistic matters. *Bioinformatics* 31 (14), 2310–2317. <https://doi.org/10.1093/bioinformatics/btv154>.
- Gur, C., Ibrahim, Y., Isaacson, B., Yamin, R., Abed, J., Gamliel, M., Mandelboim, O., 2015. Binding of the Fap2 protein of fusobacterium nucleatum to human inhibitory receptor TIGIT protects tumors from immune cell attack. *Immunity* 42 (2), 344–355. <https://doi.org/10.1016/j.immuni.2015.01.010>.
- Li, H., 2015. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annu. Rev. Stat. Appl.* 2, 73–94. <https://doi.org/10.1146/annurev-statistics-010814-020351>.
- Lydersen, S., Fagerland, M.W., Laake, P., 2009. Recommended tests for association in 2×2 tables. *Stat. Med.* <https://doi.org/10.1002/sim.3531>.
- Mandal, S., Van Treuren, W., White, R.A., Eggesbø, M., Knight, R., Peddada, S.D., 2015. Analysis of composition of microbiomes: a novel method for studying microbial

- composition. *Microb. Ecol. Health Dis.* 26, 27663. <https://doi.org/10.3402/mehd.v26.27663>.
- Ruan, W., Engevik, M.A., Spinler, J.K., Versalovic, J., 2020. Healthy human gastrointestinal microbiome: composition and function after a decade of exploration. In: *Digestive Diseases and Sciences*. Springer. <https://doi.org/10.1007/s10620-020-06118-4>, March 1.
- Seo, S.U., Kamada, N., Muñoz-Planillo, R., Kim, Y.G., Kim, D., Koizumi, Y., Núñez, G., 2015. Distinct commensals induce interleukin-1 β via NLRP3 Inflammasome in inflammatory monocytes to promote intestinal inflammation in response to injury. *Immunity* 42 (4), 744–755. <https://doi.org/10.1016/j.immuni.2015.03.004>.
- Shek, L., 2017. An innovative infant milk formula with large, phospholipid-coated lipid droplets supports an adequate growth and is well-tolerated in healthy, term Asian infants. In: *Nutrition and Growth Congress*.
- Storey, John D., 2003. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Stat.* 31 (6), 2013–2035. <https://doi.org/10.1214/aos/1074290335>.
- Storey, J.D., Taylor, J.E., Siegmund, D., 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* 66 (1), 187–205. <https://doi.org/10.1111/j.1467-9868.2004.00439.x>.
- Tsai, C.-A., Hsueh, H., Chen, J.J., 2003. Estimation of false discovery rates in multiple testing: application to gene microarray data. *Biometrics* 59 (4), 1071–1081. Retrieved from. <http://www.ncbi.nlm.nih.gov/pubmed/14969487>.
- van den Elsen, L.W.J., Tims, S., Jones, A.M., Stewart, A., Stahl, B., Garssen, J., Van't Land, B., 2019. Prebiotic oligosaccharides in early life alter gut microbiome development in male mice while supporting influenza vaccination responses. *Benefic. Microbes* 10 (3), 279–291. <https://doi.org/10.3920/BM2018.0098>.
- Wagner, B.D., Robertson, C.E., Harris, J.K., Price, C., Janoff, E., 2011. Application of two-part statistics for comparison of sequence variant counts. *PLoS One* 6 (5), e20296. <https://doi.org/10.1371/journal.pone.0020296>.
- Wilcoxon, F., 1945. Individual comparisons by ranking methods. *Biom. Bull.* 1 (6), 80–83. Retrieved from. <http://www.jstor.org/stable/3001968>.
- Wopereis, H., Oozeer, R., Knipping, K., Belzer, C., Knol, J., 2014. The first thousand days - intestinal microbiology of early life: establishing a symbiosis. In: *Pediatric Allergy and Immunology*. Blackwell Publishing Ltd. <https://doi.org/10.1111/pai.12232>.
- Xia, Y., Sun, J., 2017. Hypothesis testing and statistical analysis of microbiome. *Genes Dis.* <https://doi.org/10.1016/j.gendis.2017.06.001>. Chongqing yi ke da xue, di 2 lin chuang xue yuan Bing du xing gan yan yan jiu suo. (2017, September 1).