Statistical comparison of relative proportions of bacterial genetic sequences

Jose F. Garcia-Mazcorro[1], Jan S. Suchodolski[1], Jörg M. Steiner[1] and Bradley J. Barney[2]

[1]College of Veterinary Medicine, Department of Small Animal Clinical Sciences, Texas A&M University, College Station, TX, USA

[2]College of Science and Mathematics, Department of Mathematics & Statistics, Kennesaw State University, Kennesaw, GA, USA

Correspondence: Bradley J. Barney (bjbstat@gmail.com), and Jose F. Garcia-Mazcorro (jgarcia@cvm.tamu.edu).

Abstract

The intestinal tract is inhabited by hundreds of different types of bacteria, which have the potential of enhancing health or disease in the host. Several current technologies are capable of identifying these bacteria by determining the order of nucleotides (sequencing) in their DNA sequence with an unprecedented coverage. These technologies can provide two types of data sets: 1) the raw genetic sequences (not discussed here), and 2) the relative proportions of sequences, which are calculated by dividing the number of sequences obtained from a given bacterial group by the total number of sequences obtained. This dependent variable (relative proportions of sequences) is continuous but constrained between 0 and 100%, and has a nested architecture (bacterial species within a genus within a Family within an Order within a Class within a Phylum). We discuss different alternatives (both parametric and non-parametric) to analyze this data set, with emphasis on the use of SAS 9.2. PROC MIXED can be used but skewed residuals are commonly encountered (data is usually not normally distributed). PROC GLIMMIX with a beta distribution can also be used; however, the beta distribution assumes that the total proportion of 100 is divided between two groups. The Dirichlet distribution is a generalization of the beta distribution that allows a proportion to be divided between two or more groups, but SAS does not currently provide this option. Future analyzes are needed and ongoing to empirically determine the most appropriate statistical method to compare relative proportions of bacterial genetic sequences.

Introduction

The intestinal tract in mammals and other animals is a long muscular canal extending from the mouth to the anus where assimilation of nutrients occurs. In part due to the constant availability of nutrients, the intestinal tract contains more than 1000 different species of commensal bacteria belonging to over 10 distinct bacterial phyla (Rajilic-Stojanovic *et al.*, 2007). Because of the wide variety and high number of microorganisms present in the intestinal tract (up to 10^{11} per gram of intestinal contents), as well as their close interlink with the host, the intestinal microbial ecosystem is considered one of the most complex microbial ecosystems on Earth.

The intestinal microbiota (*micro*: little, *biota*: life) can be defined as all the microorganisms living in the intestinal tract. The intestinal microbiota can be identified using culture techniques, which rely on the growth of the microorganisms in selective culture media. Culture techniques are relatively inexpensive and allow for characterization of phenotypes (i.e. the result of genes). However, culture techniques are time consuming and often difficult to perform, mainly because most intestinal microbes are very sensitive to oxygen. The characterization of the intestinal microbiota is important because intestinal microorganisms are often involved in multiple physiological processes in health and disease (Stecher & Hardt, 2008).

Another way of identifying intestinal microorganisms is by 'reading' the base pair composition (sequencing) in their DNA. Sequencing bacterial genes from a complex microbial ecosystem can yield two types of data sets: the actual bacterial sequences (not discussed here), and the relative proportions of sequences, which are calculated by dividing the number of sequences obtained from a given bacterial group by the total number of sequences obtained. This dependent variable

(relative proportions of genetic sequences) is continuous but constrained between 0 and 100%, and has a nested architecture (bacterial species within a genus within a Family within an Order within a Class within a Phylum). This short communication deals with two main questions: what statistical procedures can we use to analyze relative proportions of bacterial sequences? Also, is there a best way to do so?

Inferences about a population proportion π

There are two main assumptions about the relative proportions of bacterial sequences to keep in mind: 1) there is a real unknown abundance of each group of microorganisms in nature, and 2) the abundance of bacterial genetic sequences represents the abundance of the microorganisms themselves. There are some issues with these assumptions but for our purposes they will be considered to be valid. Now, in a binomial experiment each trial results in one of two outcomes (usually labeled arbitrarily success or failure), with π being the probability of success and (1- π) being the probability of failure. Letting *y* denote the number of successes in *n* sample trials, the sample proportion is: $p = y/n$. For our purposes, this can be thought as *y* being the number of sequences of **Y**our **F**avorite **B**acteria ('success' group) and *n* being the total number of sequences obtained. Also, this can be thought as *y* being the number of animals harboring YFB and *n* being the total number of sampled animals. However, the latter approach has the disadvantage that it is possible and in fact very common to find YFB in all different populations of animals, and yet find differences in their relative abundance across these populations. Please note that in this communication we only address the statistical comparison of relative proportions of bacterial sequences and not the proportions of animals harboring or not harboring a specific bacterial

group. The following example shows how we can calculate a 95% confidence interval of the proportion of YFB in healthy animals.

Example 1

Your Favorite Bacteria (YFB) in the intestinal tract of dogs is known to have a positive effect on intestinal health. Therefore, with the purpose of using the proportion of YFB as an indicator of intestinal health and to compare it with a population of diseased dogs, researchers at the Gastrointestinal Laboratory want to calculate a confidence interval of the proportion of YFB in healthy dogs. For this purpose, they sampled intestinal contents from ten healthy dogs and obtained a total of 1,290 bacterial sequences. From these sequences, a total of 340 were YFB. From this data,

$$\mathrm{p} = \frac{340}{1290} = 0.26, \text{ therefore, } (1 - \mathrm{p}) = 0.74$$

To calculate the standard error (s.e.) of p,

$$\mathrm{s.e.} = \sqrt{\frac{(0.26)(0.74)}{1290}} = 0.012$$

Note that this standard error is not particularly informative because the size of *n* (number of bacterial sequences) is inversely proportional to the standard error. In other words, the higher the number of sequences the smaller the standard error, independently from the actual proportions. This is something important to keep in mind because it is not uncommon to analyze many

(>10,000) bacterial sequences. For example, if we change the number of sequences above from 1,290 to 12,900, the standard error gets three times lower (0.004). Therefore, using these many sequences (12,900), yet the same proportion (26%) of YFB, the 95% confidence interval of the proportion of YFB in healthy dogs is only:

$$0.26 \pm 1.96(0.004)\ or\ 0.26 \pm 0.008\ or\ 26\% \pm 0.8\%$$

On the contrary, when using the original 1,290 sequences, the 95% confidence interval of the proportion of YFB in healthy dogs looks more reasonable:

$$0.26 \pm 1.96(0.012)\ or\ 0.26 \pm 0.024\ or\ 26\% \pm 2.4\%$$

From this data, one could conclude with high confidence that the proportion of YFB in healthy dogs will be between 23.6% and 28.4%. As one can appreciate, this is also not very informative because the proportion of YFB was obtained from intestinal samples from different healthy dogs (without regards to breed, age, diet, or environment, all factors that could potentially affect the proportion of YFB). For this proportion to be useful, one would need to sample a very specific population of dogs (e.g., young Chihuahua dogs under the same diet). Nonetheless, even though one would sample a specific population of dogs, unless the calculated proportion is compared with another population of animals (see below), a confidence interval of a proportion of any bacterial group is useless by itself because it cannot be compared with the proportions in other studies. This is due to numerous factors associated with both sequencing and pre-sequencing procedures.

Inferences about the difference between two population proportions, $\pi_1 - \pi_2$

As shown above, we can calculate a confidence interval for the proportion of YFB but this is not particularly informative, unless you have sampled a very specific population of individuals. Another interesting question is whether we can compare two binomial parameters belonging to two different populations. The most common scenario where one might want to compare two proportions of bacterial sequences is between a healthy and a diseased population of animals. For this type of comparison, we assume that independent random samples are drawn from two binomial populations with unknown parameters, π_1 and π_2. If y_1 of YFB are observed for the random sample of size n_1 from population 1 (healthy animals) and y_2 of YFB are observed for the random sample of size n_2 from population 2 (diseased animals), then the sample proportions p1 and p2 are:

$$\text{p1} = \frac{y1}{n1} \qquad and \qquad \text{p2} = \frac{y2}{n2}$$

Example 2

A devastating disease that causes bloody diarrhea is known to be associated with changes in the proportions of YFB. In order to investigate this potential association, researchers at the Gastrointestinal Laboratory have obtained intestinal samples from ten healthy dogs and ten dogs suffering from the disease that causes bloody diarrhea. These are the numbers of bacterial sequences obtained from these two populations of dogs:

	Bacterial sequences in dogs with bloody diarrhea	Bacterial sequences in healthy dogs
All other bacteria	216	114
YFB	392	413

In this example, the proportion of bacterial sequences of YFB is higher in healthy dogs, so let's make 'Healthy dogs' population number 1.

$$\text{p1} = \frac{413}{(114+413)} = 0.784 \qquad \text{p2} = \frac{392}{(216+392)} = 0.645$$

The estimated standard error is

$$\sqrt{\frac{(.784)(1-.784)}{527} + \frac{(.645)(1-.645)}{608}} = .0264$$

Therefore, the 95% confidence interval is

$$(.784 - .645) - 1.96(.0264) \leq \pi 1 - \pi 2 \leq (.784 - .645) + 1.96(.0264)$$

or

$$.087 \leq \pi 1 - \pi 2 \leq .191$$

Therefore the 95% confidence interval for $\pi 1 - \pi 2$ is (0.784-0.645) +/- 1.96(0.0264), or (0.087,0.191), which indicates that we are 95% confident that the percentage of all sequences that are of YFB in healthy dogs is between 8.7% and 19.1% greater than the percentage of all sequences that are of YFB in dogs with bloody diarrhea.

Example 3

The Gastrointestinal Laboratory designs a study to compare the effectiveness of a new drug ('Happy Intestines') that supposedly enhances intestinal health by raising the proportion of YFB. To investigate this, the researchers enrolled a total of 10 identical dogs (i.e., same breed, age, diet, and type of environment), divided them into two groups (placebo and the new drug), and administered the treatment for 30 days. Intestinal samples were collected after either treatment was discontinued. After all samples were collected, sequencing was performed, yielding the following number of sequences:

	Bacterial sequences in dogs consuming new drug	Bacterial sequences in dogs consuming placebo
YFB	146	123
Other bacteria	49	56

In this example we want to answer the following question: does the new drug 'Happy Intestines' leads to a higher proportion of YFB when compared to placebo? For this, we denote the proportion of YFB in the new drug group as π_1 and the proportion of YFB in the placebo group as π_2. Our hypotheses are:

$$Ho: \pi_1 - \pi_2 \leq 0$$

$$Ha: \pi_1 - \pi_2 > 0$$

We will reject *Ho* if the test statistic *z* is greater than $z_{0.05} = 1.645$. From the data provided above we can compute the estimates

$$\mathrm{p1} = \frac{146}{195} = .749 \quad and \quad \mathrm{p2} = \frac{123}{179} = .687$$

Using these estimates we can compute the test statistics to be

$$z = \frac{\mathrm{p1} - \mathrm{p2}}{\sqrt{\frac{\mathrm{p1}(1-\mathrm{p1})}{\mathrm{n1}} + \frac{\mathrm{p2}(1-\mathrm{p2})}{\mathrm{n2}}}} = \frac{.749 - .687}{\sqrt{\frac{.749(.251)}{195} + \frac{.687(.313)}{179}}} = 1.33$$

Since z=1.33 is smaller than 1.645, we do not have evidence to support the hypothesis that the new drug 'Happy Intestines' leads to a higher proportion of YFB when compared to placebo.

Inferences about several proportions: Chi-square goodness-of-fit test

The comparison of two binomial parameters is helpful when one needs to compare the relative proportions of bacterial sequences from only two distinct populations (e.g. healthy and diseased). However, it is also common to analyze proportions from more than two populations (for example, the proportions of YFB in young healthy, young diseased, adult healthy, adult diseased). Following previous examples, we define y_1 and n_1 to be, respectively, the number of YFB sequences and the total number of sequences from a random sample of population 1. We analogously define y_2 and n_2, y_3 and n_3, up to y_k and n_k. In order to use such sample information to make inferences regarding $\pi_1, \pi_2, \ldots, \pi_k$, we can use the chi-square goodness-of-fit test to test the null hypothesis that the proportion is the same in each of the *k* populations. The chi-square test focuses on *y* and n-*y* for each population. The conditions for this test are that:

1) the expected number of bacterial sequences of each bacterial group (e.g., YFB and not YFB) are at least 10 for each population (e.g., $n_1\pi_1$, $n_1(1-\pi_1) \geq 10$);

2) each subject from a population has the same true underlying proportion of each bacterial group as the other subjects in the population (which is unlikely, see final note below); and

3) each bacterial sequence is independent from the others, both within and among different subjects.

Example 4

Researchers at the Gastrointestinal Laboratory speculate that the proportion of YFB is different between healthy and diseased animals. In addition, they think that this difference may be related to the age of the animals. In order to investigate this hypothesis, researchers obtained intestinal samples from 20 dogs (5 from each population) and have obtained the following number of sequences:

	Bacterial sequences in young healthy	Bacterial sequences in young diseased	Bacterial sequences in adult healthy	Bacterial sequences in adult diseased
YFB	326	125	222	486
Other bacteria	125	451	315	223

The most commonly used statistical technique used to analyze several proportions is the Chi-square goodness-of-fit test. Using the data set provided in Table, we obtained a significant p-value ($p<0.0001$), suggesting that at least one of the cell's proportions differs from the expected value. However, even though we have reached statistical significance, it is neither

straightforward nor accurate to determine exactly the population of dogs in which the proportion of YFB was significantly higher or lower. For example, using the data provided above we can calculate the proportions of YFB and other bacteria in all sampled populations:

	Proportions in young healthy	Proportions in young diseased	Proportions in adult healthy	Proportions in adult diseased
YFB	0.72	0.22	0.41	0.69
Other	0.28	0.78	0.59	0.31

Based on these proportions, one could conclude that young healthy animals have a high proportion of YFB, which tend to be lower during disease. On the contrary, adult healthy animals have a low proportion of YFB, which tend to be elevated during disease. One disadvantage of this analysis is that even though the possibility of an interaction between age and health status is clear, this hypothesis is not directly tested.

The Friedman's test

The Friedman's test (method of ranks) is a non-parametric statistical test developed by the American economist Milton Friedman in the 1930's. Similarly to the parametric ANOVA, it is commonly used to detect differences among different treatments or populations. One important disadvantage of the method of ranks is that it does not allow for testing interactions, because *without quantitative measurements, 'interaction', in the sense used in the ordinary analysis of variance, is meaningless* (Friedman, 1936). Another disadvantage of the Friedman's test is that it is not possible to analyze more than one independent variable (e.g., treatment and age). Also, the Friedman's test is incapable of modeling potential correlations among repeated measures.

Despite these disadvantages, the Friedman's test is a useful non-parametric alternative of an ANOVA and has been used recently to compare relative proportions of bacterial sequences before, during, and after administration of probiotics (Garcia-Mazcorro *et al.*, 2011).

PROC MIXED in SAS 9.2

Another option to analyze relative proportions of bacterial sequences is to consider this dependent variable as continuous and employ an ANOVA. To investigate the use of an ANOVA to analyze this data set, we analyzed the relative proportions of bacterial families found in a recent study (Garcia-Mazcorro *et al.*, 2011) using the MIXED procedure in SAS 9.2 (Table 1). In this study, 14 bacterial families were found in at least two time points (before, during, or after consumption of probiotics) in at least half the subjects (healthy dogs). The following code was used in SAS 9.2:

```
ods graphics on;
libname aglm 'c:\Your_Folder';
proc mixed data= aglm.Your_Data method=reml;
class dog time;
model Bacterial_Family= time / e3 ddfm=kr solution residual;
random dog;
run;
```

Briefly with regards to this code, many studies show that the default estimation method of the covariance parameters, the Restricted/Residual Maximum Likelihood (REML), has many advantages over other methods. The fit statistics based on REML can be used to compare different covariance models based on the same mean model; the fit statistics based on ML can be used to compare different mean models on the same covariance model (Mixed Model Analyses using SAS® course notes, 2008). The default covariance structure used by the procedure is Variance Components or simple structure (not shown in the code). This is an independent and

equal variance structure, where the within-subject error correlation is zero. This is usually not a reasonable structure for repeated measures data because the repeated measures within a subject are often correlated.

Table 1 Summary of statistical analysis of relative proportions of bacterial sequences from a recent study by Garcia-Mazcorro *et al*. (2011). The data set used for these analyzes included 36 observations for each bacterial family (each observation is a proportion of the bacterial family) from 12 dogs at three time points each (before, during and after administration of probiotics).

Bacterial family	Percentage of zeros in data set	p-value ANOVA†	Studentized Residuals	Comments	P-value (Friedman's test)
Clostridiaceae	None	0.7624	Normal		0.7624*
Ruminococceae	None	0.3892	Normal		0.3892*
Lachnospiraceae	None	0.6976	Little skewed		0.6976*
Erysipelotrichaceae	None	0.4657	Not normal		0.4657*
Eubacteriaceae	8% (3/36)	0.9186	Not normal		0.0388
Coriobacteriaceae	11% (4/36)	0.0626	Little skewed	One clear outlier was removed	0.1482
Fusobacteriaceae	22% (8/36)	0.0767	Not normal	Two outliers were not removed	0.0128
Streptococcaceae	33% (12/36)	0.5544	Not normal		0.9294
Veillonaceae	36% (13/36)	0.4452	Not normal		0.0626
Bacteroidaceae	39% (14/36)	0.9420	Not normal		0.6065
Prevotellaceae	39% (14/36)	0.9758	Not normal		0.1522
Enterococcaeae	47% (17/36)	0.1316	Not normal	One clear outlier was removed	0.2319
Lactobacillaceae	58% (21/36)	0.3330	Not normal	One clear outlier was removed	0.5308
Enterobacteriaceae	61% (22/36)	0.1630	Not normal		0.5404

† P-value for treatment effect (before, during, or after administration of probiotics) in PROC MIXED using the code described in main text.

* These p-values are the only ones in this column that are not from a Friedman's test. These p-values are from an ANOVA for repeated measures in Prism5 (GraphPad Software, Inc.). Interestingly, in Prism5 an ANOVA for independent measures (no repeated) is the same as an ANOVA for repeated measures (GraphPad Prism5.0 User's Guide).

PROC GLIMMIX in SAS 9.2

As noticed above, the use of a linear mixed model using PROC MIXED is not very useful to analyze relative proportions of bacterial sequences because this data set is usually not normally distributed (a high percentage of zeros is common in most bacterial groups). Another option to

analyze relative proportions of sequences is to use a generalized linear model using PROC GLIMMIX in SAS 9.2. There are many different distributions (for both discrete and continuous response variables) that can be analyzed using the GLIMMIX procedure. For our purposes, the most adequate distribution to use for comparison of relative proportions of bacterial sequences is the Beta distribution. The Beta distribution is akin to the binomial distribution in that it assumes the existence of only two proportions (e.g. YFB and all other bacteria). Note that for using the Beta distribution in PROC GLIMMIX, the original data set should contain proportions between 0 and 1 (instead of between 0 and 100%, which can be analyzed using PROC MIXED). Theoretically, it is possible to use a beta distribution in PROC GLIMMIX using, for example, the following code:

```
ods graphics on;
libname aglm 'c:\ Your_Folder';
proc glimmix data= aglm.Your_Data plots=studentpanel;
class dog time;
model Bacterial_Family = time/dist=beta e3 ddfm=kr solution ;
random dog;
run;
```

which is very similar in syntax to the code for PROC MIXED:

```
ods graphics on;
libname aglm 'c:\Your_Folder';
proc glimmix data= aglm.Your_Data plots=;
class dog time;
model Bacterial_Family= time / e3 ddfm=kr solution residual;
random dog;
run;
```

However, the use of PROC GLIMMIX often requires advanced statistical expertise. For example, a RANDOM statement in PROC MIXED defines the G-side random effects and a

REPEATED statement specifies the covariance structure among the residuals (i.e., the R matrix in mixed model theory notation). In PROC MIXED, you can use both statements to model correlated errors given random effects. However, one need to be careful in using both because often one statement (for example, the RANDOM statement) captures most of the variability in your data, and the other statement might not be needed. On the contrary, a RANDOM statement in PROC GLIMMIX defines the Z matrix of the mixed model, the random effects in the vector, the structure of G, and the structure of R. Also, the RANDOM_residual_ statement (not available in PROC MIXED) in PROC GLIMMIX indicates a residual-type (R-side) random component that defines the R matrix. In other words, an R-side effect in the GLIMMIX procedure is equivalent to a REPEATED effect in the MIXED procedure (for more details, the reader is referred to Statistical Analysis with the GLIMMIX procedure, course notes, 2007). Collaborations with professional statisticians are crucial for sound statistical analysis using this procedure in SAS.

The Dirichlet distribution

Thus far we have assumed that all the intestinal microbiota is composed of two bacterial groups, Your Favorite Bacteria and all other bacteria. As mentioned in the introduction, the intestinal tract contains up to 10^{11} microorganisms belonging to more than 1000 bacterial species, and therefore it is also of interest to have a statistical procedure that can allow the proportions of bacterial groups to be divided among more than two groups. The Dirichlet distribution is a useful alternative for this (Melo *et al.*, 2009). In contrast to general or generalized linear models, where we have one dependent variable (e.g., proportion of bacteria in YFB), the Dirichlet distribution can model multiple dependent variables. Let y_{ig} be the proportion of all the bacteria for subject *i*

that are in the *g*th bacterial group, for each subject *i* and for each bacterial group *g*=1,2,…,*k*. With only two possible groups, y_{i2} must equal 1-y_{i1}, meaning that one of the measurements is redundant and thus we can focus on the (univariate) y_{i1}'s while ignoring the y_{i2}'s. In general, y_{ik} is always equal to 1-y_{i1}-y_{i2}-…-$y_{\_i(k-1)}$, so the last component is redundant but we need at least *k*-1 components. When we have more than two groups, it is possible to analyze the data using a series of two-group comparisons, each time using the beta distribution. However, it is statistically preferable to analyze them simultaneously using the Dirichlet distribution, which extends the beta distribution to allow the total number of elements to belong to more than two groups of interest. In intestinal microbial ecology, this is exactly what is needed to analyze the proportions of bacterial sequences, given the expected close interactions among different bacterial groups in nature. Suppose we are considering three bacterial groups: YFB, Additional Favorite Bacteria (AFB), and everything else. A possible multivariate response would be y_{i1}=0.40, y_{i2}=0.25, y_{i3}=0.35 (meaning for subject *i* that 40% is YFB, 25% is group AFB, and the remaining 35% is everything else). Unfortunately, the Dirichlet distribution is not available in SAS 9.2 or 9.3.

Summary and conclusion

The intestinal microbiota can be defined as all the microorganisms living in the intestinal tract. The intestinal microbiota can be identified using culture or culture-independent techniques. A commonly used culture-independent technique is sequencing (reading the base pair composition in the DNA). Sequencing of microbial genes yields two data sets: the actual sequences, and the relative proportions of sequences, which are obtained by dividing the number of sequences obtained from a given bacterial group by the total number of sequences obtained.

Based on our experience dealing with relative proportions of bacterial sequences, and the information contained in this communication, we conclude that the best approach to analyze relative proportions of bacterial sequences should be one that: 1) allows the inclusion of more than one independent variable, 2) allows for interactions between independent variables, 3) can model correlations among repeated measurements, and 4) allows the proportions of bacterial groups to be divided among more than two groups. In addition, it would be useful if there would be a way to model a hierarchical structure among the bacterial groups (bacterial species within a genus within a Family within an Order within a Class within a Phylum). This will be especially interesting in microbial ecology for finding functional clusters of microorganisms, as 16S rRNA gene-based approaches only provide information about phylogenetic clustering. A generalized linear mixed model with a Dirichlet distribution seems to be a promising alternative for doing so.

Final note about the binomial distribution

Because we assume that the abundance of bacterial sequences represent the abundance of the microorganisms themselves, it is unlikely that each bacterial sequence is independent from every other bacterial sequence. Also, in order to combine numbers of sequences from different subjects (see examples above), one needs to assume that each subject has the same underlying true values for the overall proportion, which in our experience is also very unlikely. New sequencing technologies can provide us with thousands of sequences from a single or small number of individuals, which helps to accurately measure the proportions for each dog. However, we must remember that there might be wide dog-to-dog variability in the relative abundance of different

intestinal microorganisms., For example, even if we knew the sequence of every bacterium in a dog's intestinal tract, so that we knew exactly the proportion of sequences from each bacterial group, there is still uncertainty about what the average would be across all dogs because not every dog need have the same overall proportions of sequences. If we combine all of the sequences from different dogs to estimate the proportion for all dogs, we must remember this: many (or even unlimited) sequences from each of only a few dogs is no substitute for having possibly fewer sequences per dog but from many more dogs.

List of cited references

Friedman M (1936) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Annals of Mathematical Statistics* 29-46.

Garcia-Mazcorro JF, Lanerie DJ, Dowd SE, Paddock CG, Grutzner N, Steiner JM, Ivanek R & Suchodolski JS (2011) Effect of a multi-species synbiotic formulation on fecal bacterial microbiota of healthy cats and dogs as evaluated by pyrosequencing. *FEMS Microbiology Ecol*ogy doi: 10.1111/j.1574-6941.2011.01185.x

Melo TFN, K.L.P. V & A.J. L (2009) Some restriction tests in a new class of regression models for proportions. *Computational Statistics and Data Analysis* **53**: 3972-3979.

Rajilic-Stojanovic M, Smidt H & de Vos WM (2007) Diversity of the human gastrointestinal tract microbiota revisited. *Environmental Microbiology* **9**: 2125-2136.

Stecher B & Hardt WD (2008) The role of microbiota in infectious disease. *Trends in Microbiology* **16**: 107-114.