UNIVERSITY OF TURKU

Examining the feasibility of the negative binomial model in characterizing microbial abundance variation

Wisam Tariq Saleem

Master thesis

December 2020

Department of Mathematics and Statistics

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.

UNIVERSITY OF TURKU

Department of Mathematics and Statistics

Wisam Tariq Saleem: Examining the feasibility of the negative binomial model in characterizing microbial abundance variation

Master thesis, 54 pages

Applied Mathematics

December 2020

The simplicity of a linear model makes it a powerful tool for studying natural phenomena. However, often the assumptions are too limited. In statistical ecology, for instance, we frequently encounter situations where variation in species abundances fluctuate in a non-linear manner and exhibit properties such as heteroscedasticity that are not captured by standard linear models. In this thesis, I investigate the feasibility of the negative binomial model, as a tool in statistical ecology. The negative binomial has many desirable properties in terms of modelling variation in species abundances. I discuss these properties and assess the performance of a specific type of the negative binomial model which is called the traditional negative binomial model in the context of microbial ecology, which is a rapidly emerging application area for statistical models. The analysis is based on openly available data sets from published literature. The thesis concludes by discussing the potentials of the negative binomial models and their challenges.

Keywords: count data, generalized linear models, dispersion, maximum likelihood estimation, goodness of fit, phyloseq object, microbiota, microbiome.

Acknowledgements.

Foremost, I would like to thank my supervisor, Dr Leo Lahti for giving me the opportunity to work in a truly interdisciplinary research field with the freedom and responsibilities of scientific work, and with the necessary amount of guidance. I am thankful for his kind supervision, patience and knowledge.

Besides my supervisor, I would like to express my thanks and deepest appreciation to Professor Kari Auranen for his insightful comments and questions on the thesis.

My sincere appreciation also goes to Dr Katja Pahkala for giving me the chance to do the analysis for the gut microbial data, which is a part from STRIP study (Available at: https://stripstudy.utu.fi/english.html).

Finally, I owe my gratitude to my wonderful wife, Wasna' for her patience and understandable attitude for this journey in our life, and for my kids who are the most beautiful creatures in my world.

TABLE OF CONTENTS

1 INTRODUCTION …….. 1

2 THEORETICAL BACKGROUND..……... 3

2.1 Statistical concepts…..…… ……..…. 3

2.2 Generalized linear regression models…….…….. 5

2.2.1 Exponential family distributions…….. 5

2.2.2 The components of GLM ……..…….. 6

2.3 Statistical inference……...…… 7

2.3.1 Newton-Raphson algorithm……... 7

2.3.2 Iteratively weighted least squares algorithm (IWLS) ….…….. 8

3 COUNT DATA MODELS...…… 9

3.1 Poisson model…….……. 9

3.2 Overdispersion…… 10

3.2.1 Dispersion parameter…….……. 10

3.3 Negative binomial (NB) model…….. 11

3.3.1 The traditional negative binomial (NB2) GLM……..……..... 11

3.3.2 Poisson-gamma mixture ……..... 12

3.4 Goodness of fit (GOF) ……... 14

3.4.1 Likelihood statistics……... 14

3.4.1.1 Akaike information criterion (AIC) …….. 14

3.4.1.2 Bayes information criterion (BIC) …… 15

3.4.1.3 Pearson Chi-square…… 15

3.5 False discovery rate (FDR)……. 16

4 DEMONSTRATION ON GUT MICROBIOM DATA……..…….... 17

4.1 Human adult gut microbiota.…….……. 18

4.2 Visualising and grouping…… 18

4.3 The prevalent taxa and DHARMa R package…….…… 20

4.4 Model implementation.……...……. 22

4.5 The results…….. 22

4.5.1 Log-symmetric taxa…..…… 26

4.5.2 Right-skewed taxa…...…….. 29

4.5.3 Left-skewed taxa…..……. 30

5 DISCUSSION…….……... 31

REFERENCES…….……. 33

APPENDIX 1…….……..……. 38

APPENDIX 2…….……... 39

APPENDIX 3…….……... 42

List of figures

Figure 2.1: Linear model for *Y~ X.* ……... 4

Figure 4.1: Log-symmetric taxa according to CLR plots…… 19

Figure 4.2: Right-skewed taxon according to CLR plot…..……..……..... 19

Figure 4.3: Bimodal taxon according to CLR plot……..…..……..... 20

Figure 4.4: Left-skewed taxon according to CLR plot……...……..……. 20

Figure 4.5: Standardized AIC for the Poisson model for all the prevalent taxa……..……. 24

Figure 4.6: Standardized AIC for the NB2 model for all the prevalent taxa…… 24

Figure 4.7: Fitting the NB2 model to log-symmetric *Bryantella formatexigens et rel*. gene.... 26

Figure 4.8: Fitting the NB2 model to log-symmetric *Subdoligranulum variable at rel*. gene.. 27

Figure 4.9: Fitting the NB2 model to *Lachnospira pectinoschiza et rel.* gene……...…… 28

Figure 4.10: Fitting the NB2 model to right-skewed taxon…… 29

Figure 4.11: Fitting NB2 to left-skewed taxon……. 30

List of tables

Table 4.1: Ten prevalent taxa in with their CLR plots……..……...… 21

Table 4.2: AIC and BIC for model fit for the first 18 prevalent taxa under the Poisson and the NB2 models showing the difference between AIC and BIC…….... 23

Table 4.3: p-values for KS test, larger than 0.01…… 25

1. INTRODUCTION

Wide range of parametric models fall under the umbrella of the generalized linear model (GLM) [1, 2], yet our data of interest are counts and known for their heteroscedasticity, which makes certain parametric models more efficient than other ones. It has been suggested that the negative binomial (NB) model is an efficient parametric candidate to model counts with heteroscedasticity [3 4 5].

Microbial counts are rich with mathematical challenges and thus require certain tools and techniques. Some challenges for modelling species abundance are: 1) counts are rarely equidispersed [6, 7, 5], i.e. the variance is significantly greater than the mean and Poisson regression is not applicable as it would lead to a high chance of false positives; 2) counts often exhibit lots of zeros [5, 8] representing taxa that do not appear in the sample; 3) due to the development in the field, the number of features is almost equal or exceeds the number of samples in the data set [7]; 4) the differences in absolute numbers of counts obtained by the current measurement techniques arise from technical, not biological variation, and hence only information on differences in relative abundances are available [7, 5, 8, 9]. The latter challenge is technical and it has two main remedies, a) normalization, which is a parametric approach[10]; b) rarefication, which is a non-parametric approach [11, 12, 13, 14].

By modelling microbial data, we try to deploy mathematical measures to 1) understand the microbe's behaviour and the relationships between certain features and the abundance of specific species; 2) make reliable predictions about the microbe's behaviour under certain variables or features. To do so we need to look closely at the statistical challenges embedded in microbial data in order to decrease the risk of having overestimated or underestimated associations. It is important to highlight that studying the microbial species abundances has made use of the developments in high-throughput RNA sequencing (RNA-seq) data [15].

From the mathematical point of view, a statistical model can be defined as a set of probability distributions on the sample space S. A parameterised statistical model is a parameter set together with a function, which assigns to each parameter point a probability distribution on S [16]. From the applied view, a statistical model is a description of the probability distribution of random variables, which can be assumed to represent a real-world phenomenon [1, 2, 17, 18].

Furthermore, the parametric model is a model assuming the underlying population to be distributed according to a defined distribution, such as the normal distribution[19].

In the coming chapters, I shall mathematically examine some count models including the Poisson, negative binomial (NB) and briefly about zero-inflated negative binomial (ZINB) models. There are different opinions among researchers about how appropriate the NB models are for modelling the microbial abundances [5, 20, 21, 22]. The topic is extensive and in order to concentrate our efforts, I studied in this thesis the traditional negative binomial model, referred to as the NB2 model. In addition, I briefly review some popular algorithms for estimating the model parameters, including measures for the goodness of fit.

The practical part of the thesis will analyze an openly available microbial dataset to examine the feasibility of the NB2 model for modelling all the taxa with the highest representation across the samples, referred to as the prevalent taxa. The analysis is accompanied with plots and tests that are essential for understanding the behaviour of the NB2 in modelling microbial abundances. Some representative examples are examined more closely.

Finally, I shall try to draw some conclusions about the main advantages and shortcomings of using the NB models for modelling the microbial abundances in the light of recent research. In this work, I use taxa and counts sometimes interchangeably, however, taxa are the specific names of the microorganisms while counts are the observed abundances of those microorganisms.

2. THEORETICAL BACKGROUND

This chapter and the next one provide the theoretical basis for the analysis in order to understand the NB model and its variants. I will first introduce standard terminology of generalized linear models (GLMs) and compare it with the standard linear model. This will cover the methods whose practical performance will then be demonstrated and evaluated in Chapter 4.

2.1 Statistical concepts

Numeric or quantitative variables are measured by numbers and they can be continuous or discrete. Variables that are not numeric indicate groups or labels and are called factors. Sometimes non-numeric variables are called *qualitative* variables. Levels of the factor are determined by the groups within the factor, for example, the variable sex has two levels, male and female. Besides, there are more classifications for variables depending on their semantic meaning. In general, numeric variables can be analyzed by linear models if they satisfied certain requirements.

Linear regression model (LM) is a common technique, with the following simple form:

$$y_i = B_0 + B_1 x_i + \epsilon_i \quad \text{for} \quad i = 1, 2, \dots, n, \tag{2.1}$$

where y is called the *output*, *response* or *dependent* variable. The response here is assumed to be a normally distributed random variable, with n independent components. The expectation of the response is $E(y_i) = \mu_i = B_0 + B_1 x_i,$ and the variance is σ^2. Variable x is called the predictor, B_0 is the y-intercept, B_1 is the slope of the regression line and ϵ is a random variable that represents the error of the model, i.e. it accounts for the random variation in y that is not explained by x. The error term ϵ consists of n independent components which are normally distributed with expectation $E(\epsilon_i) = 0$ and the variance is σ^2.

Equation (2.1) is a linear expression in terms of B_0 and B_1, whereas related to x it can well be for example squared or cubic. For example, the equation $y = B_0 + B_1 x^2$ is still a linear regression model, by contrast, $y = B_0 + e^{B_1 x} + \epsilon$ is not. Furthermore, in multiple linear regression, more than one predictor can be added to the model[1, 2]. Figure 2.1 illustrates an example of linear

regression where the random variable Y systematically depends on the regressor X. All the above assumptions need to be checked from the data to guarantee reliable results.

Equation (2.1) can be written in a matrix form as follows:

$$Y = XB + \epsilon, \tag{2.2}$$

where Y is an n-dimensional vector representing the output, X is called a design matrix or model matrix. It is an $n \times (p + 1)$ matrix in which each row corresponds to one observation while the columns correspond to the predictor variables or features. The first column in the design matrix is a vector of ones corresponding to the intercept B_0. B is a $(p + 1)$- dimensional vector representing the y-intercept B_0 and the coefficients of the model, finally ϵ is an n-dimensional vector of errors. Equation (2.2) defines a multiple linear regression model. In the simple linear regression model, the model matrix is $n \times 2$, where β is a 2-dimensional vector of the intercept and the model coefficient [2, 23, 24, 25].

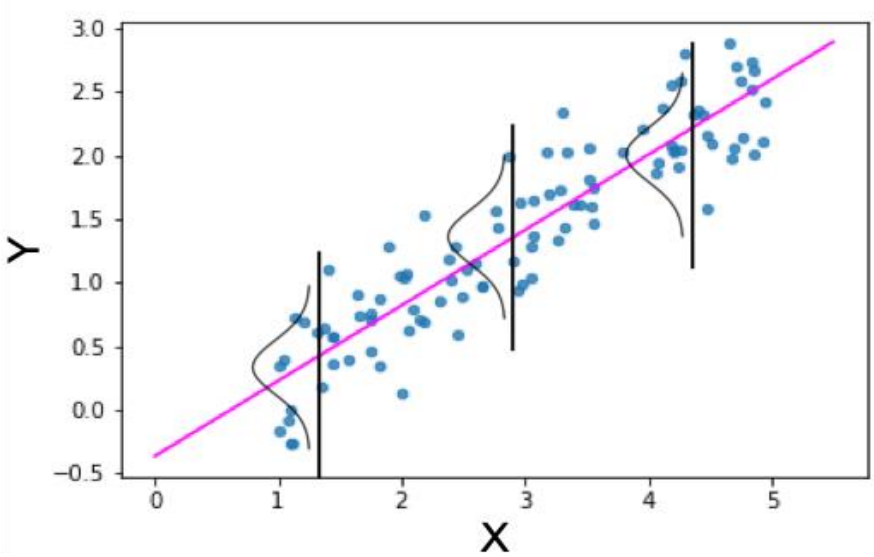


Figure 2.1: Linear model for *Y~X*

Modelling counts means that the response variable is a nonnegative integer. Data presented as counts are widely seen in ecology, directly or indirectly, as the number of birds in a certain area; or the abundance of microbiota in the human gut. The term "microbiota" is referred to the microbial taxa associated with humans to signify the communities of microorganisms within a specific environment[5]. Counts are discrete, which makes the linear model problematic as it is built on the assumptions of normally distributed continuous observations [26].

2.2 Generalized linear regression models

Generalized linear models extend standard linear regression models to include non-normal distribution of the response. Furthermore, the components of the response (y_i) can have unequal variances and need not be continuous, for example, nominal or categorical variables are possible [17, 27]. GLM can be extended to include general additive models (GAM) or more distributions such as the multivariate NB distribution. Here, the models of interest are univariate GLMs.

2.2.1 Exponential family distributions

The exponential family of probability distributions appears in several forms in the GLM literature. The following formulation is used in this thesis:

$$f(y;\theta,\alpha) = exp\{\frac{y\theta - b(\theta)}{a(\alpha)} + c(y,\alpha)\}, \tag{2.3}$$

where θ is the natural or canonical parameter that depends on regressors via a linear predictor, α is the dispersion parameter and functions *a(.)*, *b(.)*, *c(.)* are known and depend on what distribution is being considered. The derivative of $b(.)$ is a monotonic and differentiable function [17, 24, 25].

2.2.2 The components of GLM

The GLM is establishing a relationship between the exponential family of distributions and the model parameters. The main formula, in essence, is Equation (2.1) or (2.2), but the dependent variable is not limited to the normal family of distributions. To understand the GLM, I shall look at the model components as follows: 1) the random component is the response variable $Y = (y_1, \dots, y_n)$ as in Equation (2.1) or (2.2) with the following assumptions: a) $y_1, \dots, y_n$ are mutually independent, b) each y_i belongs to the exponential family distribution[1,2]; 2) the systematic component that represents the systematic part of the process; 3) a smooth and invertible link function $g(.)$ is applied to each component of $E(Y)$, relating it to the linear predictor or the systematic part as follows:

$$g(E(y_i)) = g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}. \quad (2.4)$$

For simplicity, the link function $g(.)$ is called η. In the simple linear model $g(\mu) = \mu$, i.e. $g(.)$ is the identity function. This means that the linear model is a special case of GLMs. However, there are diverse link functions and choosing the appropriate one is depending on the prior knowledge of the problem and the type of the response variable in question [1, 2].

The link function that was used by Nelder and Wedderburn in (1972) was simple [28], later the link function has been generalized to functions that could be numerically estimated and make use of the power of computers[1].

Considering Equations (2.3) and (2.4), η is called the canonical link function which is defined as $\eta = g(\mu) = \theta$. It can be shown that $\mu = b'(\theta)$ and $g = (b')^{-1}$. Use of the canonical link function eases the calculations, but also other functions are possible. Link function should be selected depending on the data and the problem beyond it[1]. Popular canonical link functions include the identity link function $g(\mu) = \mu$, as in the normally distributed response, log link function $g(\mu) = log(\mu)$, as in Poisson distributed response and logit link function $g(\mu) = logit(\mu) = log\left(\frac{\mu}{1-\mu}\right)$ as in the binomially distributed response [1]. For the NB model, the canonical link function is $g(\mu) = -\ log\left(\frac{1}{\alpha\mu} + 1\right)$ that yields canonical negative binomial (NB-C) [3, 4], as I will discuss in the next chapter, subsection (3.3.1).

2.3 Statistical inference

Statistical inference is done by estimating the parameters of the appropriate distribution based on the dataset, in order to make statistical inference about the population. Estimation is often based on the likelihood function which is maximised to find the optimal values of the model parameter. For efficient calculation, a logarithmic likelihood is used. It has been proven that the maximum likelihood estimation (MLE) is unbiased, consistent and asymptotically normal [2, 27, 30].

The log-likelihood function based on the observed counts $y_1, \dots, y_n$ following a distribution that belongs to the exponential family with known dispersion parameter α[1, 2, 27, 31] is:

$$l(\theta_1, \dots, \theta_n ; \alpha, y_1, \dots, y_n) = \sum_{i=1}^{n} \frac{y_i \theta_i - b(\theta_i)}{a(\alpha)} + c(y_i, \alpha)$$

Estimating the parameters via the maximum likelihood function often needs numerical methods. These numerical methods are implemented via algorithms, which are essential in the thesis context; as the R-functions have been used for modelling counts were built upon these methods' mathematical assumptions; the most predominant ones for our scope are explained in what follows.

2.3.1 Newton-Raphson algorithm

Newton-Raphson algorithm is useful for calculating the dispersion parameter of NB and it depends on observed information matrix (OIM). Yet, this context is not GLM as the dispersion parameter is unknown as I will discuss in the next chapter. Newton-Raphson algorithm has drawbacks like not being convergence sometimes, thus IWLS is a modified version for Newton-Raphson method, which has other several modified algorithms like Marquardt [3, 32].

2.3.2 Iteratively weighted least squares (IWLS) algorithm

IWLS is a numerical algorithm to calculate the log-likelihood function; which is essential in every step for estimating the model parameters, coefficients, model fitting, and finally models goodness of fit. IWLS makes use of two-term Taylor expansion for the log-likelihood function and stops when certain accuracy has been achieved and it based on expected information matrix (EIM) [3, 33]. IWLS is implemented as an optimization of Newton-Raphson method with Fisher scoring[1] and called also, Iteratively Re-weighted Least Squares algorithm (IRWLS) for the same algorithm and it works in the framework of the generalized linear models and the exponential family of distributions.

3. COUNT DATA MODELS

Poisson, quasi-Poisson, geometric and NB regression models are the main GLMs for modelling different types of counts such as microbial abundance. Zero-inflated models for counts are also essential to consider.

Ideally, modelling counts starts with Poisson regression, a standard model when the mean and variance of the observations are close to each other or theoretically equal to each other [4]. However, this is not the case in microbial data [7]. Relevant statistical tests should be used to check if we have real overdispersion data or not [3].

3.1 Poisson model

A discrete random variable *Y* is Poisson-distributed with intensity or rate parameter μ, $\mu > 0$, and *t* as the exposure, is defined as follows:

$$f(Y = y;\ \mu) = \frac{e^{-t\mu}(t\mu)^y}{y!}, \quad \mu > 0,\ y = 0, 1, 2, \ldots \tag{3.1}$$

Equation (3.1) is the Poisson probability mass function of Y. Exposure *t* can be defined as the length of time during which the events are recorded. The exposure can be constant or vary between events (reads). Sometimes it can also be an area, distance or population size.

If the length of the exposure period *t* equals to one we get the standard Poisson distribution function as follows:

$$f(Y = y;\ \mu) = \frac{e^{-\mu}(\mu)^y}{y!}, \quad \mu > 0, \qquad y = 0, 1, 2, \ldots$$

The Poisson regression model for counts derives from the Poisson distribution. For observation i, the relationship between the mean of the observations μ_i, coefficient vector β and the covariates or predictors x_i is parameterized as follows:

$$\mu_i = \exp(x_i'\beta) = \exp(x_{1i}\beta_1)\exp(x_{2i}\beta_2) \cdots \exp\left(x_{pi}\beta_p\right)$$

The exponential function ensures that the mean has a positive value.

The standard Poisson distribution has the equidispersion property [3, 4] i.e.

$$E(Y_i = y_i) = var(Y_i = y_i) = \mu_i$$

3.2 Overdispersion

Overdispersion or heterogeneity occurs when the mean is less than the response variance. Therefore the standard Poisson model is not capable of handling overdispersion. Two types of overdispersion in statistics should be considered, 1) apparent overdispersion, for example, the data have outliers and it could be fixed by different techniques in the Poisson framework; 2) real overdispersion when the fixing techniques for apparent overdispersion are not efficient anymore. We should then examine the reason for overdispersion in the data, and consider moving to another model like the negative binomial model [3]. In microbial data, there are many taxonomic counts with very different representations in the samples that make the means of those taxa are quite deviant from their variances [7, 5, 34].

3.2.1 Dispersion parameter

To understand the dispersion parameter, I present it via the quasi-Poisson model. Consider the variance of the quasi-Poisson model as follows:

$$var(Y_i = y_i|x_i) = \alpha\mu_i,$$

where α is called the dispersion parameter, if α = 1, then we have the standard Poisson distribution. The variance-mean relationship in the quasi-Poisson model is simple and not efficient enough to capture heterogeneity in microbial data. The reason might be that the variance is here a linear function of the mean which is a very restricting assumption easily violated in microbial counts. Therefore, we shall examine the NB model.

3.3 Negative binomial (NB) model

The negative binomial model is a unique model with many desirable properties for modelling counts with real overdispersion. The negative binomial model has several parameterizations and many model varieties, which are useful in addressing certain challenges that appear in modelling different types of data [3, 4]. The negative binomial model includes the traditional NB2, NB-C (Canonical), NB1, Geometric, NB-H (Heterogeneous negative binomial), NB-P and several more varieties with sub-varieties [3]. In this thesis, I hold to the traditional NB2 model which derived as either a member of the exponential family of distributions or as a Poisson-gamma mixture model[3].

3.3.1 The traditional negative binomial (NB2) GLM.

In literature, there is a *p-class* of negative binomial (NB-P) GLMs, which have different mean-variance relationships defined as $var(Y = y|\lambda, \alpha) = \lambda + \lambda^p \alpha$.

When $p = 1$, we have, $var(Y = y|\lambda, \alpha) = \lambda + \lambda\alpha$ and the negative binomial is named NB1,

whereas with $p = 2$, we have,

$$var(Y = y|\lambda, \alpha) = \lambda + \lambda^2 \alpha, \tag{3.2}$$

and the negative binomial is called NB2[3, 35, 36].

The traditional negative binomial model NB2 GLM can be derived from the canonical NB-C model, whereas the latter can be derived directly from the negative binomial probability mass function. However, deriving the NB2 GLM is not straightforward and it has been explained in literature such as the *Negative binomial regression* by Joseph M. Hilbe[3].

For simplicity, I shall explain here the variance derivation of the NB2 GLM, i.e. Equation (3.2), but not the derivation of the link function of the NB2 model and I shall keep things brief since I might be slipped away from the scope of my thesis.

The negative binomial probability mass function of the random variable *y* is defined as follows:

$$f(y; r, p) = \binom{y+r-1}{r-1} p^r (1-p)^y, \tag{3.3}$$

where the random variable *y* denotes the number of failures before achieving the *r*th success and the probability of success in every single trial is p [3, 4].

Rewriting the probability function (3.3) in the form of the exponential family of distributions as in Equation (2.3), the following parameters will be identified:

$\theta = ln(1-p)$, $b(\theta) = -r\,(\ln(p))$ and $a(\alpha) = 1$.

According to the theory of GLMs, taking the first and the second partial derivative of $b(\theta)$ with respect to θ yield the mean (λ) and the variance (var) respectively for the NB-C, as follows:

$$\lambda = \frac{r(1-p)}{p}, \quad var = \frac{r(1-p)}{p^2}. \tag{3.4}$$

Denoting $\alpha = \frac{1}{r}$ and parametrizing the mean and variance in Expression (3.4) using the expressions of λ and α will yield our variance of interest in Expression (3.2). The probability function (3.3) can be parametrized by using the expressions of λ and α to produce the canonical negative binomial NB-C probability mass function as follows:

$$f(y;\lambda,\alpha) = \binom{y + \frac{1}{\alpha} - 1}{\frac{1}{\alpha} - 1} \left(\frac{1}{1+\alpha\lambda}\right)^{1/\alpha} \left(\frac{\alpha\lambda}{1+\alpha\lambda}\right)^{y}$$

According to Joseph M. Hilbe the NB-C has not been used for any research project [3].

The link function of the NB-C; $\eta = -ln(\frac{1}{\alpha\lambda} + 1)$ depends on α and λ. However, for the NB2 model, there is an efficient way to modify the canonical link function to so-called log-link function that depends on λ only, as discussed in literature [3]. Anyway, even with the canonical link function, the dispersion parameter α is known[3] because the NB2 model is a GLM in this derivation.

3.3.2 Poisson- gamma mixture.

The Poisson- gamma mixture model can be derived from the Poisson model in several ways [37]. If one defines $Y \sim Poisson\,(Y|\mu)$, and $\mu \sim gamma\ (\lambda, \alpha)$, the marginal distribution of Y is found to be:

$$f(Y = y|\lambda,\alpha) = \frac{\Gamma(y+\alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1}+\lambda}\right)^{\alpha^{-1}} \left(\frac{\lambda}{\alpha^{-1}+\lambda}\right)^{y}, \tag{3.5}$$

where $\lambda, \alpha > 0$ and $y = 0,1,2,\ldots$

The marginal distribution from Equation (3.12) is NB2 [3, 38] with:

$$E(Y = y|\lambda, \alpha) = \lambda,$$

$$var(Y = y|\lambda, \alpha) = \lambda + \lambda^2\alpha$$

The NB2 derived from Poisson-gamma mixture has two parameters to be estimated; the mean λ and the dispersion α [3, 39 40]. The latter derivation of the NB2 model has been used widely in modelling ecological data[41].

Equation (3.5) with unknown α is not a true GLM anymore according to Julian J. Faraway [27], since it has two unknown parameters to be estimated, the mean and the dispersion. The NB model with a constant dispersion belongs to the exponential families of distributions with a single parameter [3, 32, 39].

To sum up, the traditional NB2 model can be derived, 1) as a Poisson-gamma mixture, which is essential to estimate the dispersion parameter; 2) from the NB-C (canonical)[3, 39].

These two derivations of the traditional NB2 model give a giant leap for the NB2 model efficiency since we can estimate the dispersion parameter from the data by the Poisson-gamma mixture form of the NB2 model and integrate it in the NB2 GLM form, the latter form is important to make use of the GLMs rules for parameters estimation, standard errors, model fitting and goodness of fit [3, 39].

Typically, the estimation of the dispersion parameter from the data is done by the MLE approach [32], however, due to the importance of estimating the dispersion parameter I refer to another method suggested by Cameron and Trivedi and they call it, auxiliary ordinary least square (OLS) regression without an intercept [4].

Another challenge for modelling microbial data is that the data often include lots of zeros. Therefore, a more realistic model should deal with that. It is crucial to use appropriate statistical tests for zero-inflation since the problem happens when the number of zeros in the observed data is higher than the number of zeros in the predicted model, i.e. the model is underfitting zeros then we need a zero-inflated count model. A zero-inflated count model is a two-part discrete model, containing binary and counts parts [5, 8, 3, 4].

The zero-inflated count model parameters are estimated by the MLE approach and the model which is related to my work in modelling microbial abundances is the zero-inflated negative binomial (ZINB) model with the algorithm called expected-maximization (EM) [8].

3.5 Goodness of fit (GOF)

Residuals describe the deviance of the predicted data from the observed data. Residual analysis means performing the appropriate tests and plots to examine whether there is a tolerable error in the model fit [32]. The basic method would evaluate the Euclidean distance between each observed and predicted datum in a linear model. More advanced methods exist in GLM for counts that consist of certain tests like Pearson Chi-square test, and different formulae which work individually for Poisson GLM, NB2 GLM and Poisson-gamma model [3, 39, 32].

Comparing different models is necessary; otherwise, the goodness of fit tests are not informative about the feasibility of the chosen model. These tests are making use of the theory of GLM. The topic can be approached by the following groups of tests.

3.5.1 Likelihood statistics

Likelihood statistics are testing how different models could maximize the likelihood of their parameters. Some of the tests are as follows:

3.5.1.1 Akaike information criterion (AIC).

The Akaike information criterion (AIC) is defined as follows:

$$AIC = -2ln\,L + 2\,P,$$

where $ln\,L$ is the log-likelihood and P is the number of the parameters in the model, such as the mean and dispersion parameters in addition to the regression coefficients [5, 39]. AIC could be accessed from most R functions for fitting GLM like *MASS::glm.nb()* and *stats::glm()*, and by *stats::AIC()* [32]. Also, AICcmodavg R package [42] is dedicated for AIC calculation. A smaller AIC indicates a better predictive ability of the model, given the data.

3.5.1.2 Bayes information criterion (BIC).

The Bayes information criterion (BIC) is defined as follows:

$$BIC = -2 \ln L + P \ln n,$$

where *n* is the sample size [5, 39]. BIC could be accessed by R function *stats::BIC()*. A smaller BIC indicates a better predictive ability of the model, given the data. The Bayes information criterion is similar to AIC, plus making use of the sample size in the penalty term as in Equation.

3.5.1.3 Pearson Chi-square.

The Pearson Chi-square is a useful measure when the mean and variance are specified correctly, then $E\left[\sum_{i=1}^{n}\left(\frac{(y_i-\mu_i)^2}{VAR\ (y_i)}\right)\right] = n$, where *n* is the sample size. The general formula for the Pearson Chi-square test statistics follows as:

$$P = \sum_{i=1}^{n}\left(\frac{(y_i-\hat{\mu}_i)^2}{VAR\ (y_i)}\right), \tag{3.6}$$

where the asymptotic distribution of P is Chi-squared with $n-p$ degrees of freedom. Assuming the NB2 model, $VAR\ (y_i) = \hat{\mu}_i + \hat{\mu}_i{}^2\alpha$, where $\hat{\mu}$ is the estimated mean from the data. As P in Equation (3.6) closer to the sample size *n*, indicates larger evidence for the model, given the data.

The Pearson Chi-square test statistics indeed reflects the underlying variability in the data, the simplest situation is the Poisson model; presuming an accurate specification of μ_i. Then, if $\sum_{i=1}^{n}\left(\frac{(y_i-\hat{\mu}_i)^2}{\hat{\mu}_i}\right) > n-p$, it implies an overdispersed data and when $\sum_{i=1}^{n}\left(\frac{(y_i-\hat{\mu}_i)^2}{\hat{\mu}_i}\right) < n-p$ is an underdispersed data[3, 39].

Many more goodness of fit (GOF) methods are available for instance likelihood ratio test, deviance residual and Deviance information criterion DIC with Bayesian estimation.

Finally, I use Kolmogorov-Smirnov (KS) test for goodness of fit. It is a non-parametric statistical test used to decide if a sample derives from a population with a defined distribution or not. The

null hypothesis of KS when the distribution of the data is defined, while the alternative hypothesis when the data do not follow a specified distribution [43, 44]. In the next chapter, I use KS test to check if the NB2 model for modelling species abundances is well supported by the data or not.

3.6 False discovery rate (FDR)

After checking the model fit, p-values are used to examine the hypotheses about which taxon is affected by certain conditions. However, by modelling microbial counts, we work with the whole community of microorganisms. Therefore, p-values are produced simultaneously at different levels of analysis wherever there are hypotheses to be tested. The process of hypothesis testing depending on p-values will generate a cumulative error which is known in statistics as a false discovery rate (FDR) [41]. Since FDR is a marginal topic in the thesis that I am not going to explain it in details, nevertheless it is necessary to understand criticism to the NB model as I will discuss in Chapter 5 [22].

There are several methods to control the FDR [45], the main one for microbial data is the method by Benjamini & Hochberg [5, 46, 47], however, choosing the best model that describes the data is crucial to reduce the FDR.

4. DEMONSTRATION ON GUT MICROBIOM DATA

The species abundances are statistical counts which are produced by amplifying and sequencing certain highly variable areas in the genes 16srRNA [8, 48]. Depending on their similarity, the reads are clustered to form the Operational Taxonomic Unit (OTU) or lately amplicon sequence variant (ASV) [49]. These OTUs are further interpreted by continuously updated databases [50] like Genome Taxonomy Database (GTDB) to extract species names from them, finally.

All the previous steps pose different statistical challenges. Methods to overcome those challenges are out of the scope of this thesis. I shall start where the species abundances, species names and the features to be examined, are available [5, 41]. By modelling microbial data, we examine which microbes are differentiated among, or between different features according to the species abundance. The term "microbiome" is defined as the collection of the microbial taxa or microbes and their genes, the entire microbial communities. The term "microbiome" is to signify the organisms and all of their related genomes[5].

The demonstration was performed with 4.0.2 (2020-06-22) (R: A language and Environment for Statistical Computing) Running under: Windows 10 x64 (build 18363). Several CRAN packages were used like *MASS* [32] *ggplot2* [51] and *Bioconductor packages like phyloseq* [12], *microbiome* [52] and *DESeq2* [53].

The phyloseq object is one of the most beneficial dataset forms for studying and visualizing the microbial data and has been supported by many different R packages. It consists of several tables in different forms of data structure that should be connected properly. The main data structures in phyloseq are as follows: 1) the operational taxonomic unit (OTU Table) as a matrix of taxonomic counts or abundances of taxa; 2) taxonomic table (Taxonomy Table) as a matrix of taxonomic names as characters and; 3) metadata table (Sample Data) as a data frame that contains the variables of interest whose effect on the taxonomic counts one wants to make statistical inference about.[12, 54].

4.1 Human adult gut microbiota

For the case study, I have used a gut microbiome dataset available as an R data file. The file, referred to as atlas1006, is in the form of phyloseq setup that summarizes the intestinal microbiota in 1151 samples of 1006 western adults with no reported health complications[55]. Atlas1006 consists of, 1)130 operational taxonomic units represent the abundances of 130 genes; 2) 130 taxonomic names represent the names of the previous 130 taxa at three taxonomic ranks; Phylum, Family and Genus; 3) 1151 sample data consists of 10 variables, i.e. it is a data frame of size 1151×10. The OTU table is a technical name in human gut microbiota atlas rather than a real biological name since this dataset was made with a different technique than what has been used recently. However, microbiome research is an active field that is developing quickly and there are always new methods for extracting counts from (RNA-seq). Sample metadata includes information on age, sex, nationality, DNA extraction method, project, diversity, body-mass index (bmi) group, subject, time and sample. It is important to highlight that the lowest level in the taxonomic rank of human gut microbiota atlas is Genus, then I use gene abundances in this demonstration and all the names in this chapter are genes not species.

4.2 Visualization and grouping

A histogram plot for every taxon from the human gut microbiota atlas would help to examine the underlying data distribution. However, microbial counts are compositional [5, 9, 56] that we transformed to $\log_{10}$ [55] or centred log-ratio (CLR) [57, 58]; which compared to $\log_{10}$ has been shown to remove the compositionality bias and reduce skewness. Statistical models for dealing with the compositional property of microbial data is under active research, yet there is no universal approach. Taxonomic units can be classified into five different compositional patterns: 1) log-normal or log-symmetric distribution pattern like *Akkermansia* and *Anaerostipes caccae et rel*.; 2) right-skewed distribution pattern like *Streptococcus bovis et rel*.; 3) left-skewed distribution pattern like *Faecalibacterium prausnitzii et rel*; 4) bimodal distribution pattern with two distinct peaks like *Prevotella oralis et rel*.; 5) rare (very low abundance) like *Serratia* [55].

Consider the CLR-abundance plots in Figure 4.1, Figure 4.2, Figure 4.3 and Figure 4.4. The plots are showing the four main groups of microbial abundance in human gut microbiota atlas excluding the rare taxa.

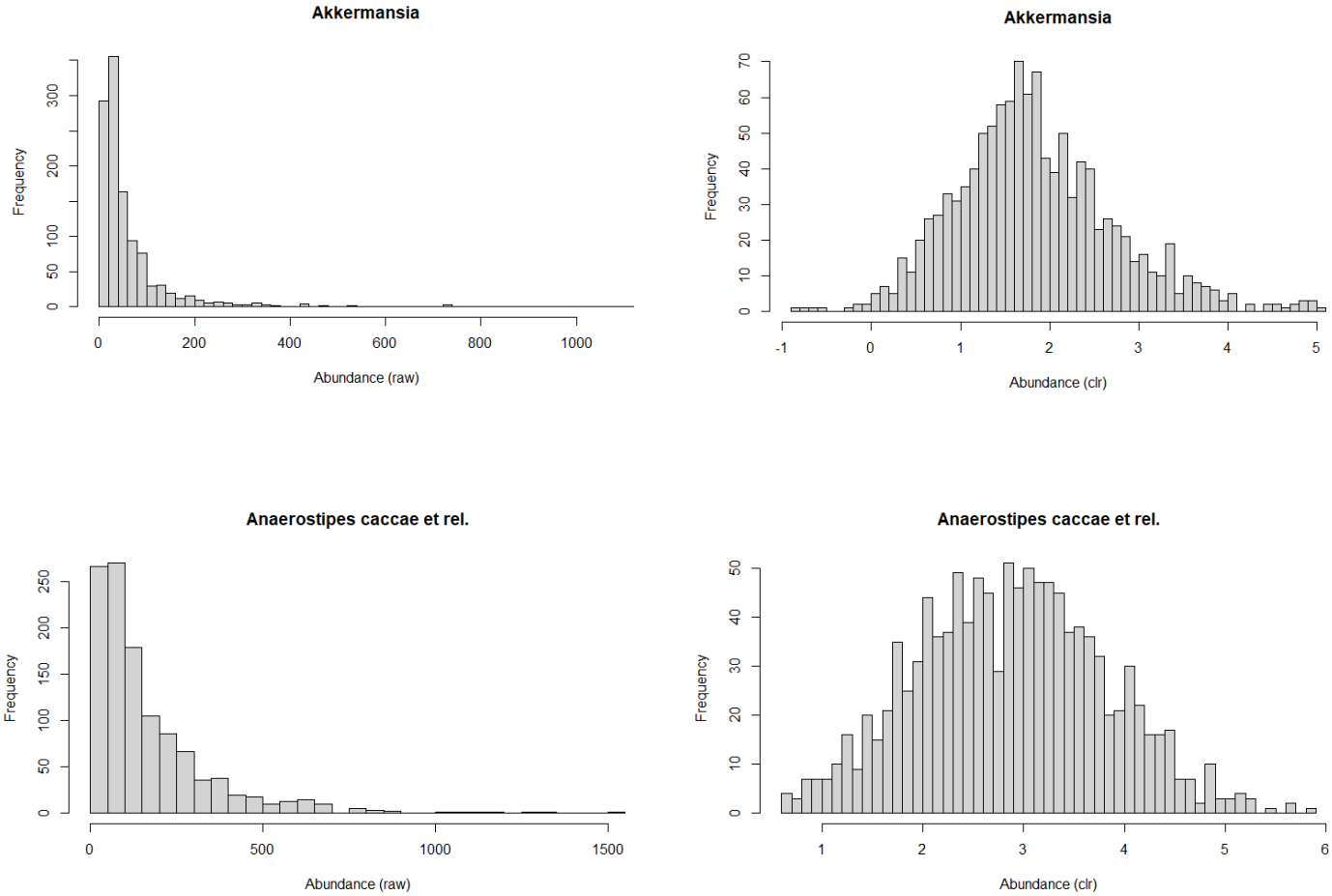


Figure 4.1: Log-symmetric taxa according to CLR plots

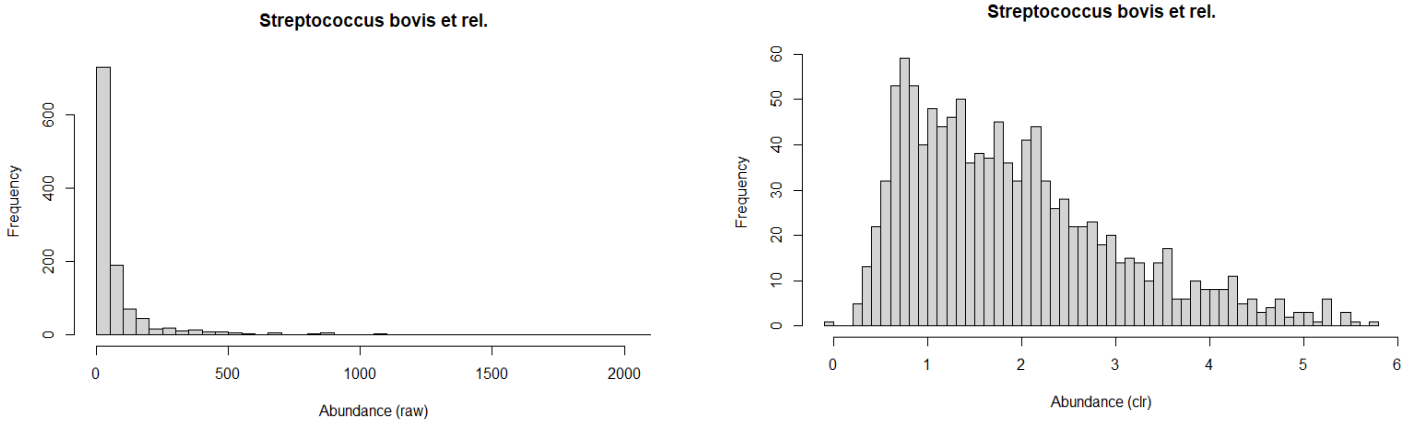


Figure 4.2: Right-skewed taxon according to CLR plot

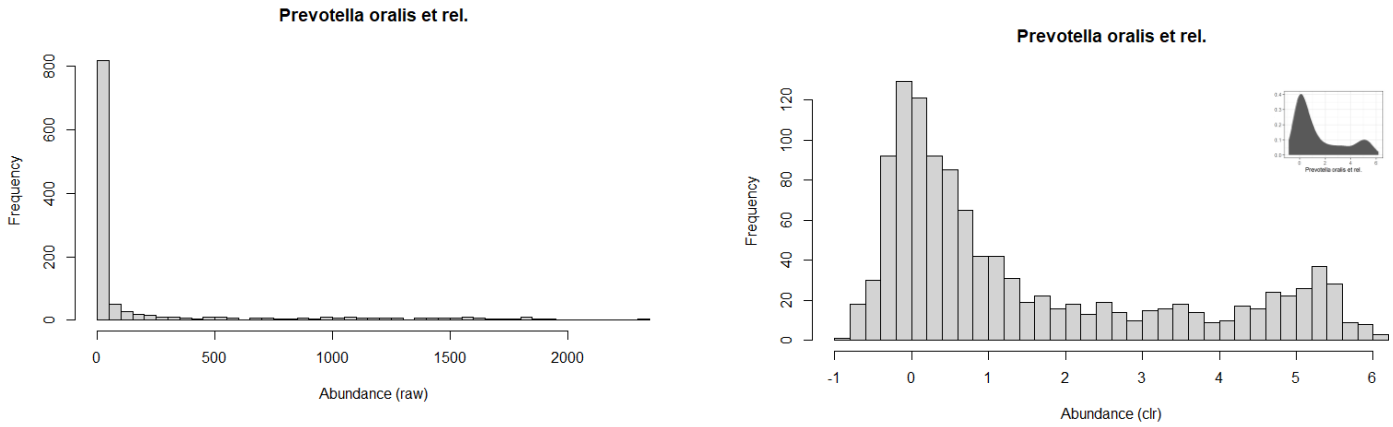


Figure 4.3: Bimodal taxon according to CLR plot

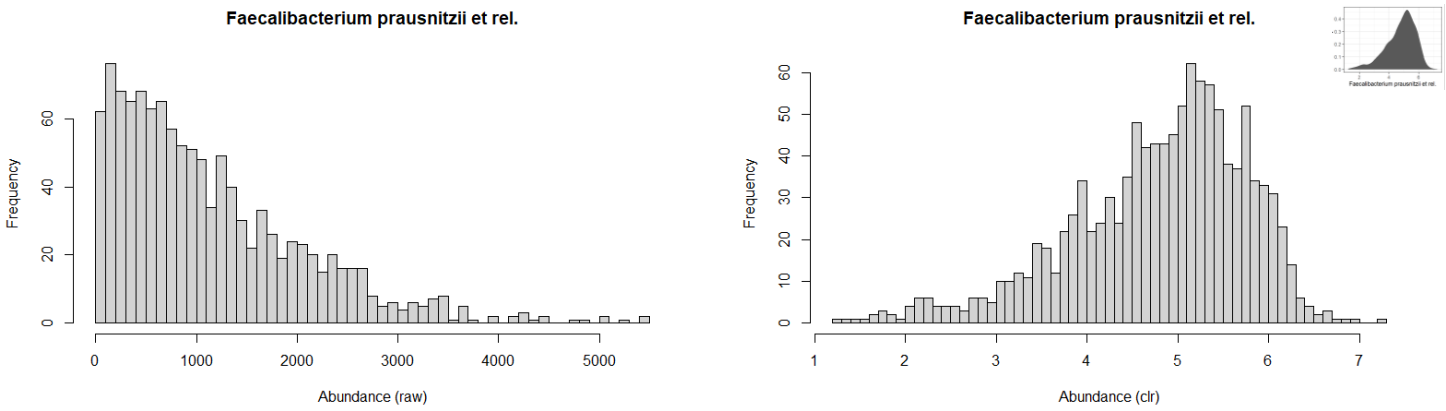


Figure 4.4: Left-skewed taxon according to CLR plot

4.3 The prevalent taxa and DHARMa R package

I examine the prevalent taxa in the human gut microbiota atlas. These prevalent taxa are in Table 4.1 and Appendix 1. The prevalent taxa have been chosen by their relative abundance values across the samples because of the compositional property of microbial data.

Useful plots and tests for examining the models fit are coming from the R package called: residual diagnostics for hierarchical (multi-level/mixed) regression models (DHARMa) [59]. DHARMa consists of plots and tests to compare the expected to the observed data with appropriate tests such as the dispersion test, outlier test and Kolmogorov-Smirnov (KS) test for goodness of fit.

DHARMa's plots and tests for all the reported prevalent taxa in gut microbiota atlas are in Appendix 3. To some extent, the DHARMa package is also useful to tackle the challenge of having the number of features exceeding the sample size by simulation-based approach [59].

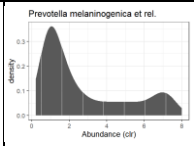
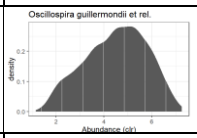
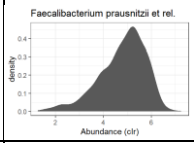
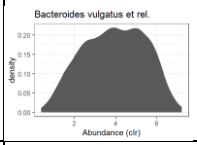
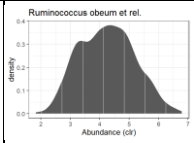
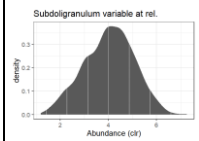
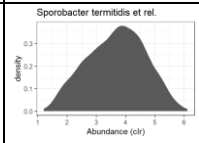
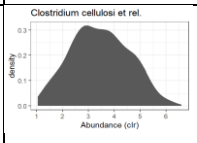
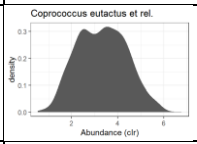
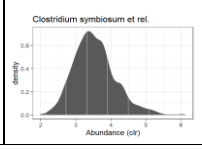
SL	Taxa	Abundance (CLR)
1	*Prevotella melaninogenica et rel.*	
2	*Oscillospira guillermondii et rel.*	
3	*Faecalibacterium prausnitzii et rel.*	
4	*Bacteroides vulgatus et rel.*	
5	*Ruminococcus obeum et rel.*	
6	*Subdoligranulum variable at rel.*	
7	*Sporobacter termitidis et rel.*	
8	*Clostridium cellulosi et rel.*	
9	*Coprococcus eutactus et rel.*	
10	*Clostridium symbiosum et rel.*	

Table 4.1: Ten prevalent taxa with their CLR plots

4.4 Model implementation

I built two models from the human gut microbiota atlas, the NB2 and the standard Poisson models. The dependent variable is the abundance of every prevalent taxon. I use the raw counts as the dependent variable in the model implementation, not the compositional values. Using compositional values would be problematic itself and needs more justifications which are out of the scope of this work. The explanatory or independent variables are (time) and (nationality). The variable (time) is a continuous variable and (nationality) is a nominal variable with the following six factors as follows: "CentralEurope", "EasternEurope", "Scandinavia", "SouthEurope", "UKIE" and "US".

The R packages I deployed to build the models are 1) *stats* [60], its main function is *stats::glm()* for the Poisson and the NB2 GLM model, the latter model is the NB2 model with one parameter where the dispersion parameter is known; 2) *MASS* [32] and its main function is *MASS::glm.nb()* which I used it to estimate the dispersion parameter from the data under the NB2 model.

4.5 The results

After fitting the Poisson and the NB2 models, I calculated their AIC and BIC in Table 4.2 and Appendix 2. Besides, I calculated the difference between each AIC value for the Poisson and the NB2 model, i.e.

$$AIC\ for\ the\ Poisson\ model - AIC\ for\ the\ NB2\ model,$$

the same for BIC values in the sixth and seventh columns respectively, see Table 4.2 and Appendix 2.

Taxa	AIC for Poisson	BIC for Poisson	AIC for the NB2	BIC for the NB2	Difference between AICs for Poisson and NB2	Difference between BICs for Poisson and NB2
Prevotella melaninogenica et rel.	4,594,856	4,594,891	15,743	15,778	4,579,113	4,579,113
Oscillospira guillermondii et rel.	1,351,615	1,351,651	17,854	17,889	1,333,761	1,333,762
Faecalibacterium prausnitzii et rel.	790,414	790,450	17,933	17,968	772,481	772,482
Bacteroides vulgatus et rel.	1,214,112	1,214,147	17,047	17,082	1,197,065	1,197,065
Ruminococcus obeum et rel.	489,489	489,525	16,593	16,628	472,896	472,897
Subdoligranulum variable at rel.	526,996	527,031	16,400	16,435	510,596	510,596
Sporobacter termitidis et rel.	397,416	397,452	15,633	15,669	381,783	381,783
Clostridium cellulosi et rel.	499,238	499,273	15,486	15,521	483,752	483,752
Coprococcus eutactus et rel.	336,460	336,495	14,996	15,031	321,464	321,464
Clostridium symbiosum et rel.	165,440	165,475	14,435	14,471	151,005	151,004
Clostridium orbiscindens et rel.	126,795	126,830	14,151	14,186	112,644	112,644
Prevotella oralis et rel.	685,010	685,046	12,633	12,668	672,377	672,378
Butyrivibrio crossotus et rel.	124,109	124,144	13,904	13,939	110,205	110,205
Dorea formicigenerans et rel.	105,804	105,839	13,745	13,780	92,059	92,059
Allistipes et rel.	239,730	239,765	14,037	14,072	225,693	225,693
Bifidobacterium	325,636	325,671	13,833	13,868	311,803	311,803
Uncultured Clostridiales I	367,886	367,921	13,889	13,924	353,997	353,997
Anaerostipes caccae et rel.	145,204	145,239	13,595	13,631	131,609	131,608

Table 4.2: AIC and BIC for model fit for 18 prevalent taxa under the Poisson and the NB2 models showing the difference between the AIC and BIC.

Table 4.2 and Appendix 2 give an initial insight into the general performance of the NB2 model. It is clear that the NB2 model is more consistent with the data than Poisson according to both AIC and BIC, since both AIC and BIC for the NB2 model are less than their peers for the Poisson. This is noticed even more clearly in Figures 4.5 and 4.6 which show the histograms of the standardized AIC values for the Poisson and the NB2 models.

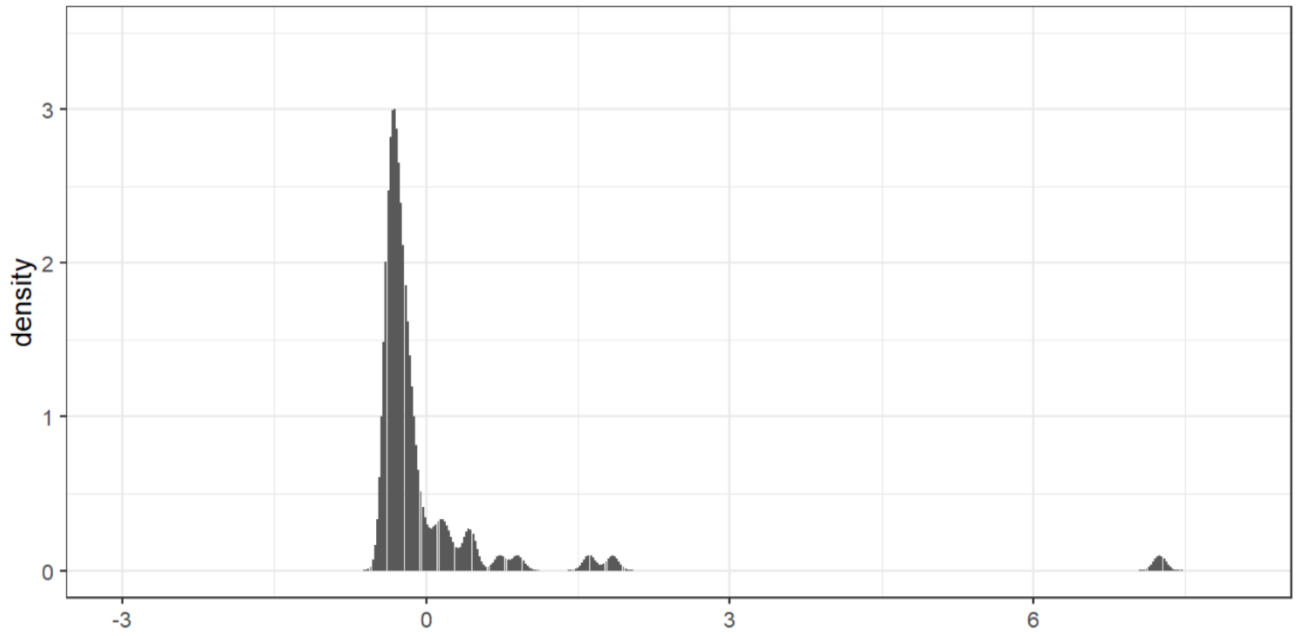


Figure 4.5: Standardized AIC for the Poisson model for all the prevalent taxa

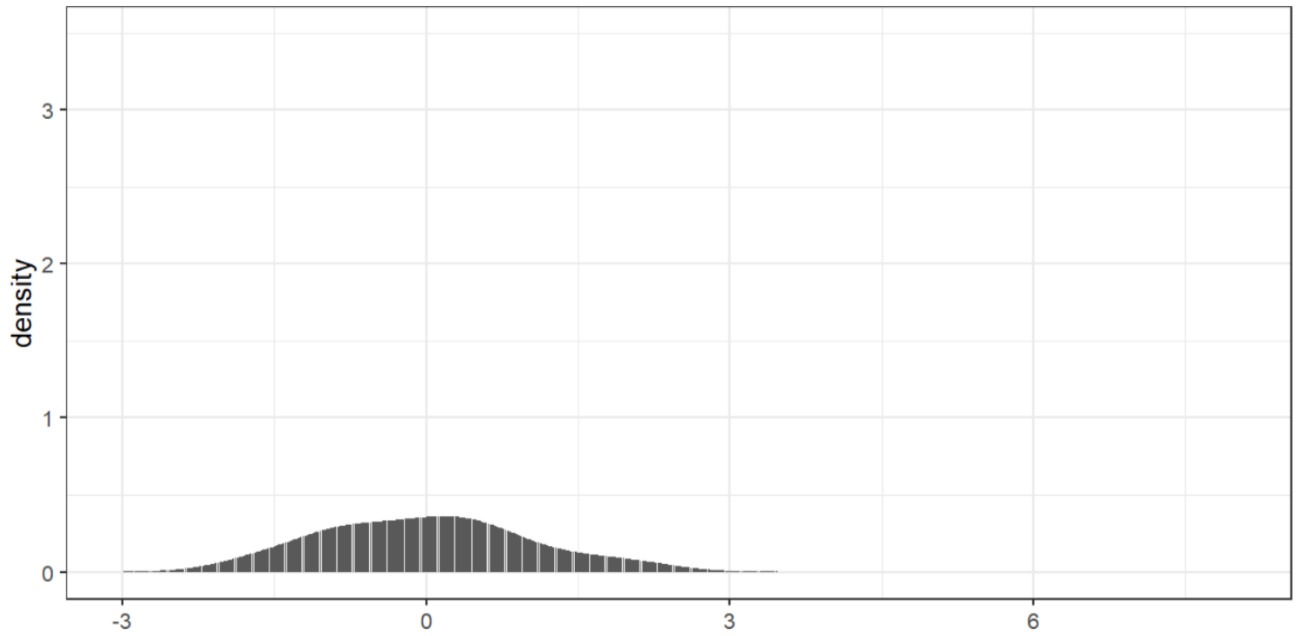


Figure 4.6: Standardized AIC for the NB2 model for all the prevalent taxa

Further inspection can be done by plotting the residual and uniform quantile-quantile plots obtained from DHARMa package, together with the Kolmogorov-Smirnov (KS) test for goodness of fit, dispersion test and outlier test [59]. Residual and uniform quantile-quantile plots with the tests for each prevalent taxon can be checked from Appendix 3, where there are more information about the goodness of fit of the NB2 model to microbial abundance despite when the KS GOF test null hypothesis is rejected.

Table 4.3 hereby shows the p-values of the KS GOF test for the NB2 models that I have implemented in Section 4.4. The null hypothesis of the KS test states that the abundances of the genes follow the negative binomial distribution and the alternative hypothesis means the distribution of the abundances are not specified as a negative binomial distribution. I report in Table 4.3 the p-values which are larger than 0.01, i.e. the null hypothesis only reported. It assessed that the abundance of those taxa in Table 4.3 is indeed following the NB2 distribution i.e. the NB2 model is feasible for modelling the abundance of those taxa.

SL	Taxa	p-values for KS test
1	Faecalibacterium.prausnitzii.et.rel.	0.502
2	Bryantella.formatexigens.et.rel.	0.431
3	Lachnospira.pectinoschiza.et.rel.	0.261
4	Ruminococcus.bromii.et.rel.	0.14
5	Subdoligranulum.variable.at.rel.	0.075
6	Dorea.formicigenerans.et.rel.	0.03
7	Clostridium.orbiscindens.et.rel.	0.029
8	Oscillospira.guillermondii.et.rel.	0.023
9	Eubacterium.rectale.et.rel.	0.019
10	Butyrivibrio.crossotus.et.rel.	0.015
11	Sporobacter.termitidis.et.rel.	0.012

Table 4.3: p-values for KS test, larger than 0.01.

The goodness of fit tests that are suitable for our purposes are diverse and included in several R packages and publications [21]. To narrow down our approach, representatives from each group of

counts are chosen according to Section 4.2, yet the models are the same as in Section 4.4. Bimodal counts like *Prevotella* group are avoided in this section as they violate the assumptions of the NB2 model, but they are in Appendix 3.

4.5.1 Log-symmetric taxa.

The log-symmetric distribution[61] pattern underlying the gene abundance can be seen in many taxa in human gut microbiota atlas so that I considered three examples in this section.

The compositional CLR plot of *Bryantella formatexigens et rel.* gene in Figure 4.7 (the upper line) supports the log-symmetric pattern. However, modelling the raw counts of *Bryantella formatexigens et rel.* gene by the NB2 model shows a better predictive ability for AIC and BIC of the NB2 model than the Poisson model, see Appendix 2 the gene: *Bryantella formatexigens et rel.*

The uniform quantile-quantile plot with KS test and the residual plot in Figure 4.7 (the lower line) give more positive shreds of evidence for the NB2 model fitting.

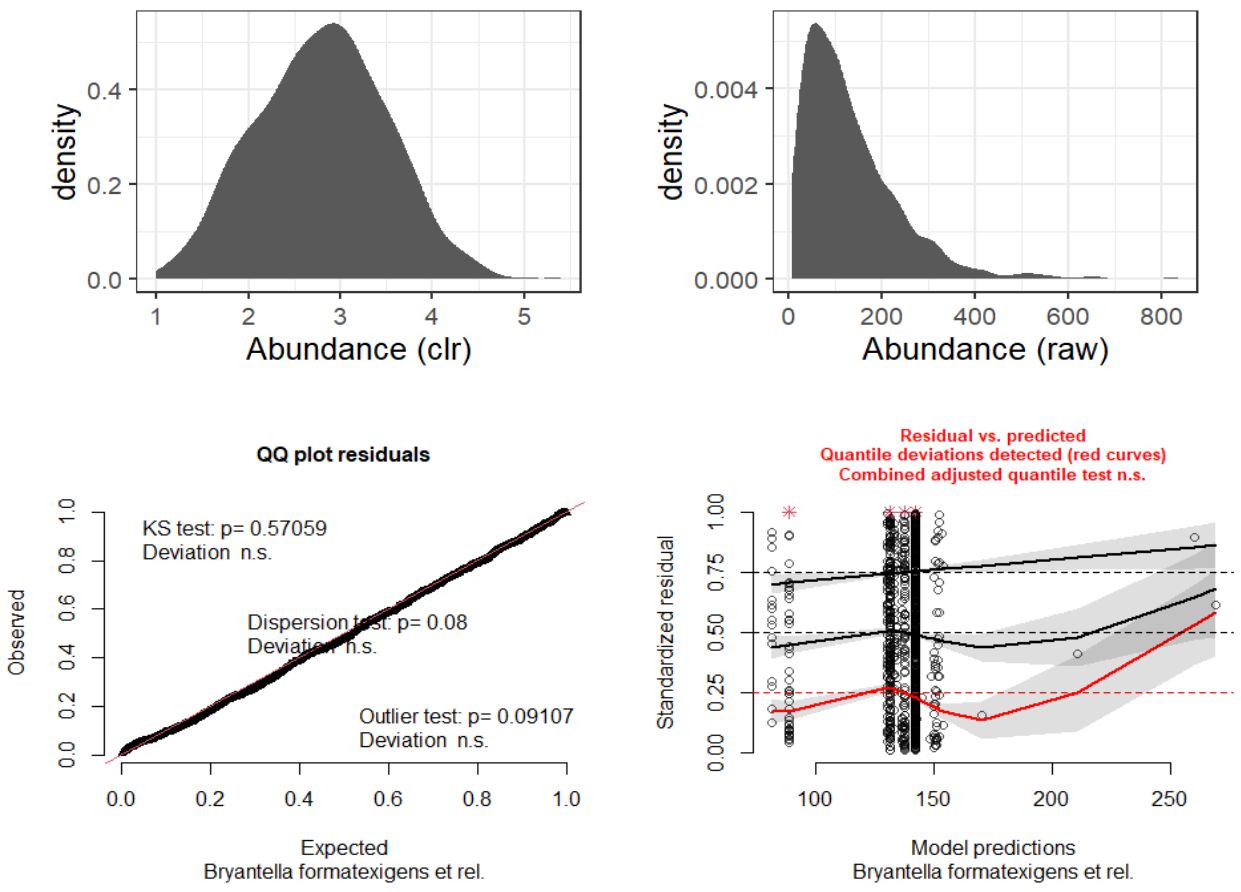


Figure 4.7: Fitting the NB2 model to log-symmetric *Bryantella formatexigens et rel*. gene.

The other example of the log-symmetric taxa is the *Subdoligranulum variable at rel*. gene. The NB2 model shows better predictive performance according to AIC and BIC than the Poisson model, see Table 4.2 the gene: *Subdoligranulum variable at rel*. Furthermore, KS test and the two plots in Figure 4.8 (the lower line) support the NB2 model approach.

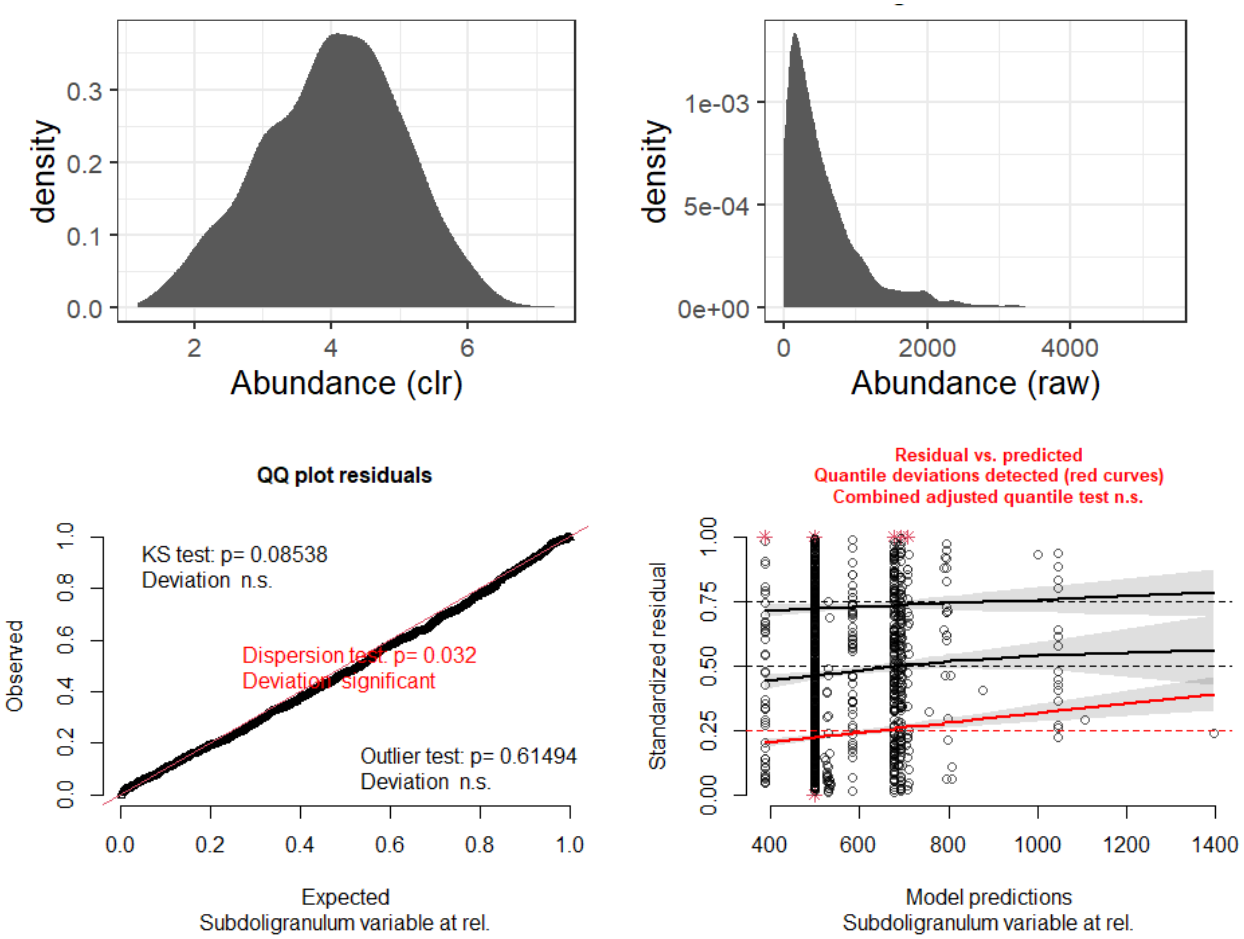


Figure 4.8: Fitting the NB2 model to log-symmetric *Subdoligranulum variable at rel*. gene.

The gene *Lachnospira pectinoschiza et rel*. shows a good model fit of the NB2 model despite that the symmetry is not exact of the gene compositional values in CLR plot, see Figure 4.9 (the upper line) which is the case in many other alike taxa. However, the AIC and BIC give better predictive for the NB2 model than the Poisson, see Appendix 2 the gene: *Lachnospira pectinoschiza et rel*. Further support in the uniform quantile-quantile plot with KS GOF test and residual plot in Figure 4.9 (the lower line).

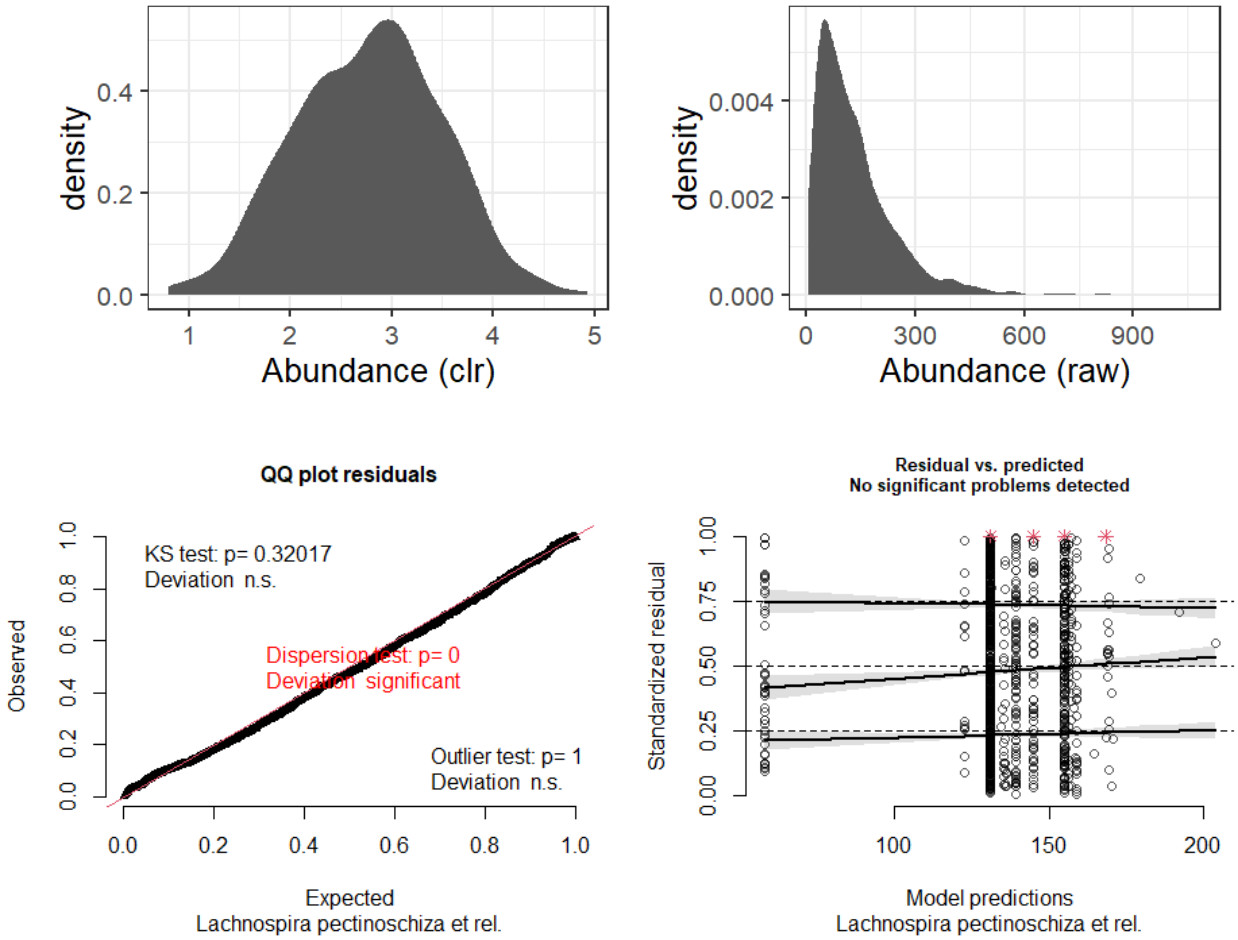


Figure 4.9: Fitting the NB2 model to *Lachnospira pectinoschiza et rel.* gene.

4.5.2 Right-skewed taxa

The *Streptococcus bovis et rel.* gene is an example of the right-skewed taxon. The *Streptococcus bovis et rel.* gene compositional values according to CLR plot can show its pattern, see Figure 4.10 (the upper line). Modelling the raw abundance of *Streptococcus bovis et rel*. gene by the NB2 model indicates better predictivity for AIC and BIC than the Poisson model, see Appendix 2 the gene: *Streptococcus bovis et rel.* But, the uniform quantile-quantile plot with KS GOF test and the residual plot altogether in Figure 4.10 (the lower line) suggest that the NB2 model is not good for modelling the *Streptococcus bovis et rel.* gene abundance. Indeed, the NB2 model is not good enough for modelling the right-skewed taxon due to the high heteroscedasticity that the NB2 model is unable to capture.

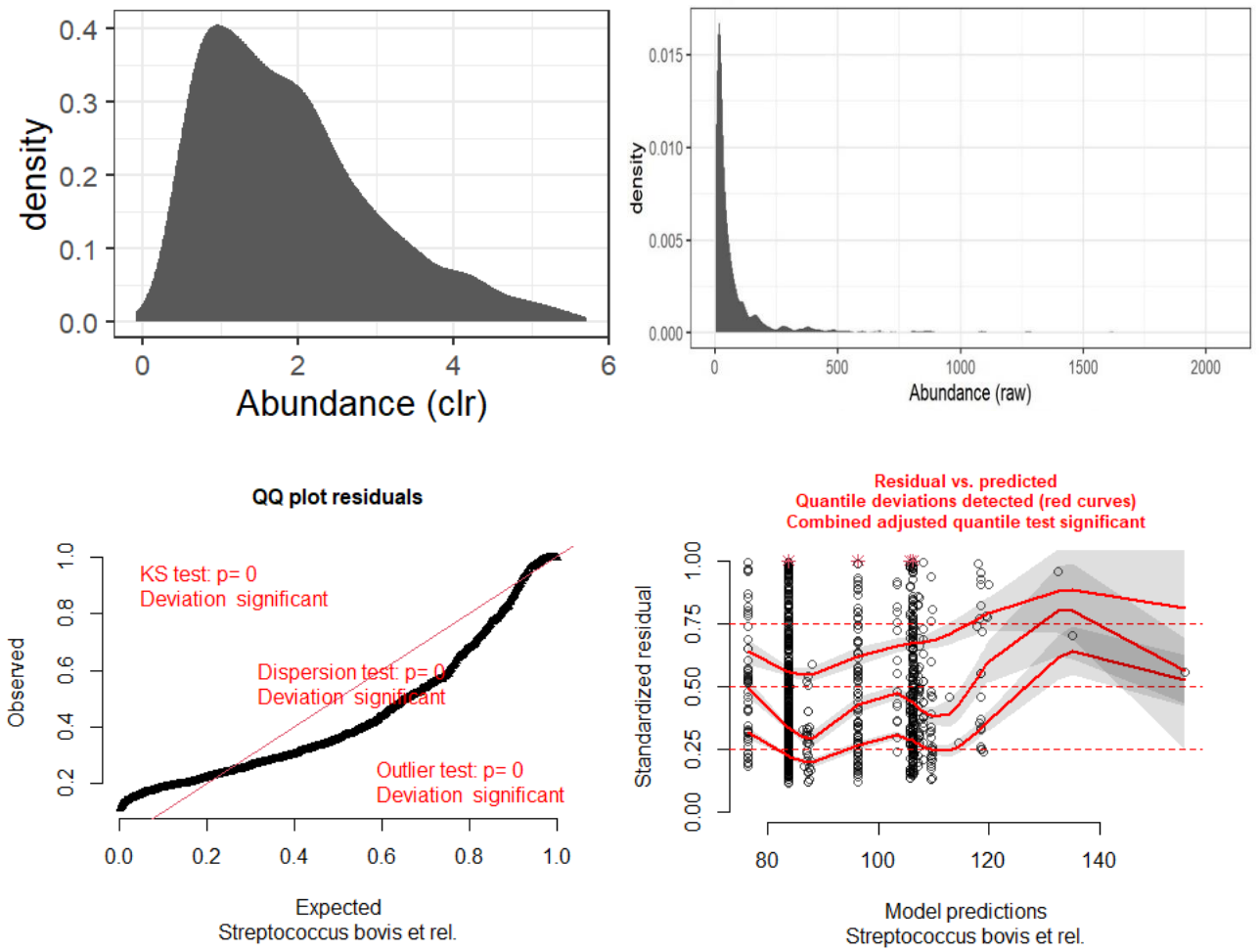


Figure 4.10: Fitting the NB2 model to right-skewed taxon

4.5.3 Left-skewed taxa.

The *Faecalibacterium prausnitzii et rel*. gene compositional pattern is an example of left-skewed taxa as it is proposed via the *Faecalibacterium prausnitzii et rel*. gene's CLR plot in Figure 4.11 (the upper line). Modelling the *Faecalibacterium prausnitzii et rel.* gene raw abundance by the NB2 and Poisson model suggests better predictive for the NB2 model according to AIC and BIC, see Table 4.2 the gene: *Faecalibacterium prausnitzii et rel*. Meanwhile, Figure 4.11 (the lower line) shows good results from the uniform quantile-quantile plot with KS GOF test and the residual plot.

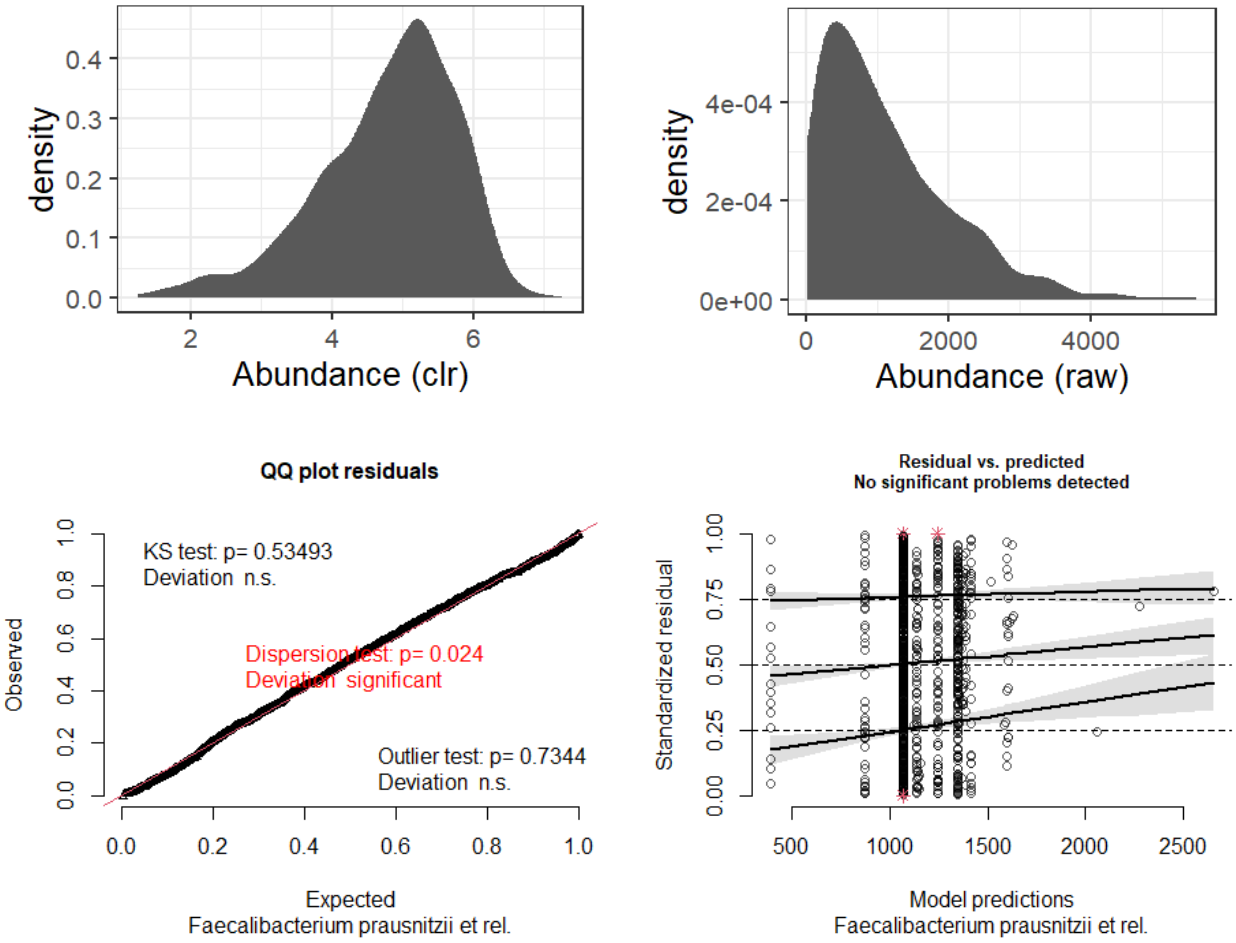


Figure 4.11: Fitting NB2 to left-skewed taxon.

Finally, I think I shall stop here as the theme is interdisciplinary and it is easy to slip out from the main topic as my main concern has been to examine the feasibility of the NB2 model for modelling microbial counts. Checking the results of AIC and BIC for all the prevalent genes in Figure 4.5 & 4.6 and the numbers in Appendix 2 and finally plots in Appendix 3, can help us to observe the behaviour of NB2 in modelling microbial counts.

5. DISCUSSION

Ecological data constitute counts with heteroscedasticity, and fitting the best model to such data is not straightforward, which I found as a motivational challenge in the thesis. However, the NB model has the required adaptation to make statistical inference for ecological data with heteroscedasticity, due to the NB model several varieties and the way these varieties could integrate themselves to different challenges [3] [39]. Our traditional NB2 model is more feasible than the standard Poisson and quasi-Poisson, that is mainly due to NB2's variance-mean relationship which is a quadratic, the property that gives NB2 more flexibility to capture heteroscedasticity in the data, this potential is even better; to some extent; for NB-P [21, 62].

Microbial ecology is an active field for bioscientists, statisticians, mathematicians and computer scientists, and the standard modelling frameworks often need to be further adjusted to fit the specifics of each application domain. However, the aim of this thesis must be specific; which is very challenging itself because in literature there are different requirements related to each aspect of microbial data such as compositionality and multiple testing. There are diverse approaches from the field of bioinformatics to overcome compositionality and the multi-testing challenges that could be further incorporated into the model [5, 41]. Meanwhile, in practice, we must work with these challenges simultaneously. Therefore I dealt with these challenges briefly, sometimes with plots like CLR, and mostly with citations. Microbiome study is an interdisciplinary and dynamic field of research [63].

Here is an example of one challenge at the beginning of the analysis which is estimating the dispersion parameter simultaneously for thousands of taxa in the dataset, keep in mind our counts are compositional. Estimating the dispersion parameter has five common scenarios, 1) the dispersion is constant for all taxa; 2) it differs between taxa but it is fixed for every taxon under all conditions; 3) the dispersion could be different for all taxa and conditions; 4) it could be a function to the mean; 5) the dispersion is a function to the mean with additional variability between-counts [21]. All the previous scenarios need to be dealt with, and they affect the results and implementation speed. However, in my demonstration, I used the second approach, because it seems to provide a feasible trade-off between simplicity and flexibility. Further research could be done to perform model comparison between the alternative models but this is out of the scope for this thesis work.

Another challenge for modelling microbial counts is so-called zero-inflation that described briefly at the end of Chapter 3, where the Negative binomial model still has some handlings to offer like the (ZINB) models [5,8]. Zero-inflation has not been shown in human gut microbiota atlas, but it has a clear role in other microbial datasets, like the vaginal taxon *Lactobacillus vaginalis*[5]. Modelling zero-inflated data could be done by *pscl::zeroinfl()* [5, 64].

An interesting paper published in April 2020 claimed that the negative binomial and zero-inflated negative binomial models are poorly controlling FDR at its nominal level [22]. This paper suggests moving to nonparametric methods that do not depend on the distribution assumptions, like Wilcoxon rank-sum test and R package ANOVA-Like Differential Expression tool for high throughput sequencing data (ALDEx2) [65, 66]. However, a closer look into that paper and its references suggest that the source of the problem is the compositional property of microbial data from one side, and the thousands of hypothesis which are being tested simultaneously from the other side.

To sum up my experience in this thesis I can refer to these challenges, which are problematic to assess them in details in the thesis, the challenges as follows: 1) integrate the relative abundances of the species in the model implementation; 2) estimate a more feasible dispersion parameter and; 3) assess reliable GOF tests which can be more sensitive to FDR.

I believe some of the good examples of R-packages that work with all the previous statistical challenges can be seen in some of the popular ones such as *DESeq2* [67], *edgeR* [68]. These packages incorporate the NB models, keep developing their tools to reduce errors and enhance model results. A newly formulated R-package named gamma-Poisson generalized linear model (glmGamPoi) [69] is an inspiring example of applying the negative binomial model to reduce time and memory usage in comparison to other commonly used packages in the field.

REFERENCES

1. McCullagh, P. & Nelder, J. A. *Generalized Linear Models, Second Edition (Monographs on Statistics and Applied Probability 37)*. *Cambridge University Press* (1989).

2. Agresti, A. Foundations of Linear and Generalized Linear Models. *John Wiley & Sons, INC* 473 (2015).

3. Hilbe, J. M. *Negative binomial regression*. (Cambridge University Press, 2007).

4. Cameron, A. Colin, and P. K. T. *Regression Analysis of Count Data*. (Cambridge University Press, 1998).

5. Xia, Yinglin, Jun Sun, and D.-G. C. *Statistical Analysis of Microbiome Data with R*. (Springer, 2018).

6. White, J. R., Nagarajan, N. & Pop, M. Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. *PLoS Comput. Biol.* **5**, (2009).

7. Pendegraft, A. H., Guo, B. & Yi, N. Bayesian hierarchical negative binomial models for multivariable analyses with applications to human microbiome count data. *PLoS One* **14**, 1–23 (2019).

8. Xinyan Zhang, Himel Mallick, N. Y. Zero-Inflated Negative Binomial Regression for Differential Abundance Testing in Microbiome Studies. *J. Bioinforma. Genomics* 1 (2016).

9. Gloor, G. B., Macklaim, J. M., Pawlowsky-glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional : And This Is Not Optional. **8**, 1–6 (2017).

10. Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. 1–18 (2017).

11. Mcmurdie, P. J. & Holmes, S. Waste Not , Want Not : Why Rarefying Microbiome Data Is Inadmissible. **10**, (2014).

12. McMurdie, P. J. & Holmes, S. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **8**, e61217 (2013).

13. Kim, B. *et al.* Deciphering Diversity Indices for a Better Understanding of Microbial Communities. **27**, 2089–2093 (2017).

14. Shade, A. *et al.* Conditionally rare taxa disproportionately contribute to temporal changes in

microbial diversity. *mBio* **5**, (2014).

15. McClure, R. *et al.* Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Res.* **41**, e140–e140 (2013).

16. McCullagh, P. What is a statistical model? *Annals of Statistics* **30**, (2002).

17. Dobson, A. *An Introduction to Generalized Linear Models, Second Edition*. (2001).

18. Chen, H. & Lindsey, J. K. Applying Generalized Linear Models. *Technometrics* **40**, (1998).

19. Efron, B. & Hastie, T. *Computer age statistical inference: Algorithms, evidence, and data science*. (2016).

20. Love, M. I., Anders, S. & Hu-, W. *Differential analysis of count data – the DESeq2 package*. (2016).

21. Mi, G., Di, Y. & Schafer, D. W. Goodness-of-fit tests and model diagnostics for negative binomial regression of RNA sequencing data. *PLoS One* **10**, 1–16 (2015).

22. Hawinkel, S., Rayner, J. C. W., Bijnens, L. & Thas, O. Sequence count data are poorly fit by the negative binomial distribution. *PLoS One* **15**, 1–16 (2020).

23. Rencher, A. C. & Schaalje, G. B. *Linear Models in Statistics*. *Linear Models in Statistics* (2007).

24. Krijnen, W. P. Applied Statistics for Bioinformatics using R. *GNU Free Document License* (2009).

25. Casella, G., Fienberg, S. & Olkin, I. *An Introduction to Statistical Learning*. *Springer Texts in Statistics* (2013).

26. Warton, D. I., Lyons, M., Stoklosa, J. & Ives, A. R. Three points to consider when choosing a LM or GLM test for count data. *Methods Ecol. Evol.* **7**, 882–890 (2016).

27. Faraway, J. J. *Extending the Linear Model with R. Second Edition (2nd ed.).* (Chapman and Hall/CRC, 2016).

28. Nelder, J. A. & Wedderburn, R. W. M. Composite Link Functions in Generalized Linear Models Author (s): R . Thompson and R . J . Baker Published by : Wiley for the Royal Statistical Society Stable. *J. R. Stat. Soc.* **135**, 370–384 (1972).

29. Cheek, P. J., McCullagh, P. & Nelder, J. A. *Generalized Linear Models, 2nd Edn. Applied*

Statistics **39**, (1990).

30. Cox, D. R. & Hinkley, D. V. *Theoretical Statistics*. *Cambridge University Press* (1974).

31. Agresti, A. Introduction to Generalized Linear Models. *Categ. Data Anal.* 115–164 (2003).

32. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S Fourth edition by*. *World* **53**, (2002).

33. Fox, J. & Weisberg, S. *CAR - An R Companion to Applied Regression*. *Thousand Oaks CA: Sage.* (2019).

34. Xia, Y. & Sun, J. Hypothesis testing and statistical analysis of microbiome. *Genes Dis.* **4**, 138–148 (2017).

35. Cameron, A. C. & Trivedi, P. K. Econometric models based on count data. Comparisons and applications of some estimators and tests. *J. Appl. Econom.* **1**, (1986).

36. Greene, W. Functional Form and Heterogeneity in Models for Count Data. *Found. Trends® Econom.* **1**, 113–218 (2007).

37. Greenwood, M. & Yule, G. U. An Inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents. *J. R. Stat. Soc.* **83**, 255–279 (1920).

38. Greene, W. H. *Econometric Analysis 4th Ed.* (2000).

39. Strawderman, R. L., Cameron, A. C. & Trivedi, P. K. Regression Analysis of Count Data. **94**, 984 (1999).

40. Raschke, C. & Greene, W. H. Erratum to Functional forms for the negative binomial model for count data [Economics Letters 99, (2008), 585-590]. *Economics Letters* **107**, (2010).

41. Holmes, S. & Huber, W. *Modern Statistics for Modern Biology*. (Cambridge University Press, 2019).

42. Mazerolle, M. J. AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c). (2020).

43. Capon, J. On the Asymptotic Efficiency of the Kolmogorov-Smirnov Test. *J. Am. Stat. Assoc.* **60**, 843–853 (1965).

44. Lilliefors, H. W. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance

Unknown. *J. Am. Stat. Assoc.* **62**, 399–402 (1967).

45. Dudoit, Sandrine., and M. J. van der. L. *Multiple Testing Procedures with Applications to Genomics*. (Springer, 2008).

46. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).

47. Jiang, L. *et al.* Discrete False-Discovery Rate Improves Identification of Differentially Abundant Microbes. *mSystems* **2**, e00092-17 (2017).

48. Callahan, B. J. *et al.* Bioconductor Workflow for Microbiome Data Analysis : from raw reads to community analyses [version 2 ; referees : 3 approved] Referee Status : 1–48 (2019).

49. Barnes, C. J. *et al.* Comparing DADA2 and OTU clustering approaches in studying the bacterial communities of atopic dermatitis. 1–10 (2020).

50. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).

51. Wilkinson, L. ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H. *Biometrics* **67**, (2011).

52. Leo Lahti and Sudarshan Shetty. microbiome R package (2012-2019).

53. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).

54. Callahan, B. J., Sankaran, K., Julia, A., Mcmurdie, P. J. & Susan, P. Workflow for Microbiome Data Analysis : from raw reads to community analyses . (2019).

55. Lahti, L., Salojärvi, J., Salonen, A., Scheffer, M. & De Vos, W. M. Tipping elements in the human intestinal ecosystem. *Nat. Commun.* **5**, 4344 (2014).

56. Luz Calle, M. Statistical analysis of metagenomics data. *Genomics and Informatics* **17**, (2019).

57. Jones, M. C. & Aitchison, J. The Statistical Analysis of Compositional Data. *J. R. Stat. Soc. Ser. A* **150**, (1987).

58. Shanmugam, R. *Applied compositional data analysis: with worked examples in R*. *Journal of Statistical Computation and Simulation* **89**, (2019).

59. Hartig, F. DHARMa: Residual Diagnostics for Hierarchical Regression Models. *The Comprehensive R Archive Network* (2020).

60. R Core Team. R Core Team (2014). R: A language and environment for statistical computing. *R Found. Stat. Comput. Vienna, Austria.* (2014).

61. Vanegas, L. H. & Paula, G. A. Log-symmetric distributions: Statistical properties and parameter estimation. *Brazilian J. Probab. Stat.* **30**, (2016).

62. Di, Y., Schafer, D. W., Cumbie, J. S. & Chang, J. H. The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Stat. Appl. Genet. Mol. Biol.* **10**, (2011).

63. Shetty, S. A. & Lahti, L. Microbiome data science. *J. Biosci.* **44**, 1–6 (2019).

64. Jackman, S. pscl: classes and methods for R developed in the political science computational laboratory. United States Studies Centre, University of Sydney. Sydney, New South Wales, Australia. *Front. Ecol. Evol.* **5**, (2017).

65. Fernandes, A. D. *et al.* Unifying the analysis of high-throughput sequencing datasets: Characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2**, (2014).

66. Gloor, G. B., Macklaim, J. M. & Fernandes, A. D. Displaying Variation in Large Datasets: Plotting a Visual Summary of Effect Sizes. *J. Comput. Graph. Stat.* **25**, 971–979 (2016).

67. Love, M., Anders, S. & Huber, W. Analyzing RNA-seq data with DESeq2. *Bioconductor* **2**, 1–63 (2017).

68. Soneson, C. & Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* **14**, 91 (2013).

69. Ahlmann-Eltze, C. & Huber, W. glmGamPoi: Fitting Gamma-Poisson Generalized Linear Models on Single Cell Count Data. *bioRxiv* (2020).

APPENDICES

Appendix 1

The rest of the most prevalent taxa, that part of them in Table (1).

SL	Taxa	SL	Taxa
11	*Clostridium orbiscindens et rel.*	38	*Papillibacter cinnamivorans et rel.*
12	*Prevotella oralis et rel.*	39	*Eubacterium rectale et rel.*
13	*Butyrivibrio crossotus et rel.*	40	*Ruminococcus gnavus et rel.*
14	*Dorea formicigenerans et rel.*	41	*Bacteroides ovatus et rel.*
15	*Allistipes et rel.*	42	*Ruminococcus lactaris et rel.*
16	*Bifidobacterium*	43	*Bacteroides plebeius et rel.*
17	*Uncultured Clostridiales I*	44	*Anaerotruncus colihominis et rel.*
18	*Anaerostipes caccae et rel.*	45	*Escherichia coli et rel.*
19	*Clostridium leptum et rel.*	46	*Bacteroides splachnicus et rel.*
20	*Uncultured Clostridiales II*	47	*Roseburia intestinalis et rel.*
21	*Bryantella formatexigens et rel.*	48	*Tannerella et rel.*
22	*Lachnospira pectinoschiza et rel.*	49	*Streptococcus mitis et rel.*
23	*Clostridium sphenoides et rel.*	50	*Sutterella wadsworthia et rel.*
24	*Ruminococcus callidus et rel.*	51	*Clostridium (sensu stricto)*
25	*Ruminococcus bromii et rel.*	52	*Uncultured Mollicutes*
26	*Outgrouping clostridium cluster XIVa*	53	*Prevotella tannerae et rel.*
27	*Bacteroides fragilis et rel.*	54	*Clostridium colinum et rel.*
28	*Streptococcus bovis et rel.*	55	*Collinsella*
29	*Eubacterium ventriosum et rel.*	56	*Anaerovorax odorimutans et rel.*
30	*Clostridium nexile et rel.*	57	*Eubacterium biforme et rel.*
31	*Lachnobacillus bovis et rel.*	58	*Clostridium stercorarium et rel.*
32	*Eubacterium hallii et rel.*	59	*Bacteroides stercoris et rel.*
33	*Dialister*	60	*Phascolarctobacterium faecium et rel.*
34	*Parabacteroides distasonis et rel.*	61	*Oxalobacter formigenes et rel.*
35	*Bacteroides uniformis et rel.*	62	*Lactobacillus plantarum et rel.*
36	*Clostridium difficile et rel.*	63	*Mitsuokella multiacida et rel.*
37	*Akkermansia*	64	*Lactobacillus gasseri et rel.*

Appendix 2

AIC and BIC for model fit for the rest of the prevalent taxa under the Poisson and the NB2 models.

Taxa	AIC for Poisson	BIC for Poisson	AIC for NB2	BIC for NB2	Difference between AICs for Poisson and NB2	Difference between BICs for Poisson and NB2
Clostridium leptum et rel.	153,498	153,533	13,485	13,520	140,013	140,013
Uncultured Clostridiales II	169,622	169,658	13,280	13,315	156,342	156,343
Bryantella formatexigens et rel.	84,314	84,350	13,011	13,046	71,303	71,304
Lachnospira pectinoschiza et rel.	90,451	90,486	13,012	13,047	77,439	77,439
Clostridium sphenoides et rel.	78,638	78,673	12,866	12,901	65,772	65,772
Ruminococcus callidus et rel.	113,176	113,211	12,835	12,871	100,341	100,340
Ruminococcus bromii et rel.	158,074	158,109	12,834	12,869	145,240	145,240
Outgrouping clostridium cluster XIVa	104,393	104,428	12,698	12,733	91,695	91,695
Bacteroides fragilis et rel.	160,551	160,586	12,372	12,407	148,179	148,179
Streptococcus bovis et rel.	189,523	189,558	12,294	12,329	177,229	177,229
Eubacterium ventriosum et rel.	108,452	108,487	12,274	12,309	96,178	96,178
Clostridium nexile et rel.	77,721	77,756	12,159	12,194	65,562	65,562
Lachnobacillus bovis et rel.	88,482	88,517	12,059	12,095	76,423	76,422
Eubacterium hallii et rel.	79,573	79,608	11,898	11,933	67,675	67,675
Dialister	289,146	289,181	10,888	10,923	278,258	278,258
Parabacteroides distasonis et rel.	93,003	93,038	11,756	11,791	81,247	81,247
Bacteroides uniformis et rel.	123,340	123,375	11,566	11,601	111,774	111,774
Clostridium difficile et rel.	223,043	223,078	11,534	11,570	211,509	211,508

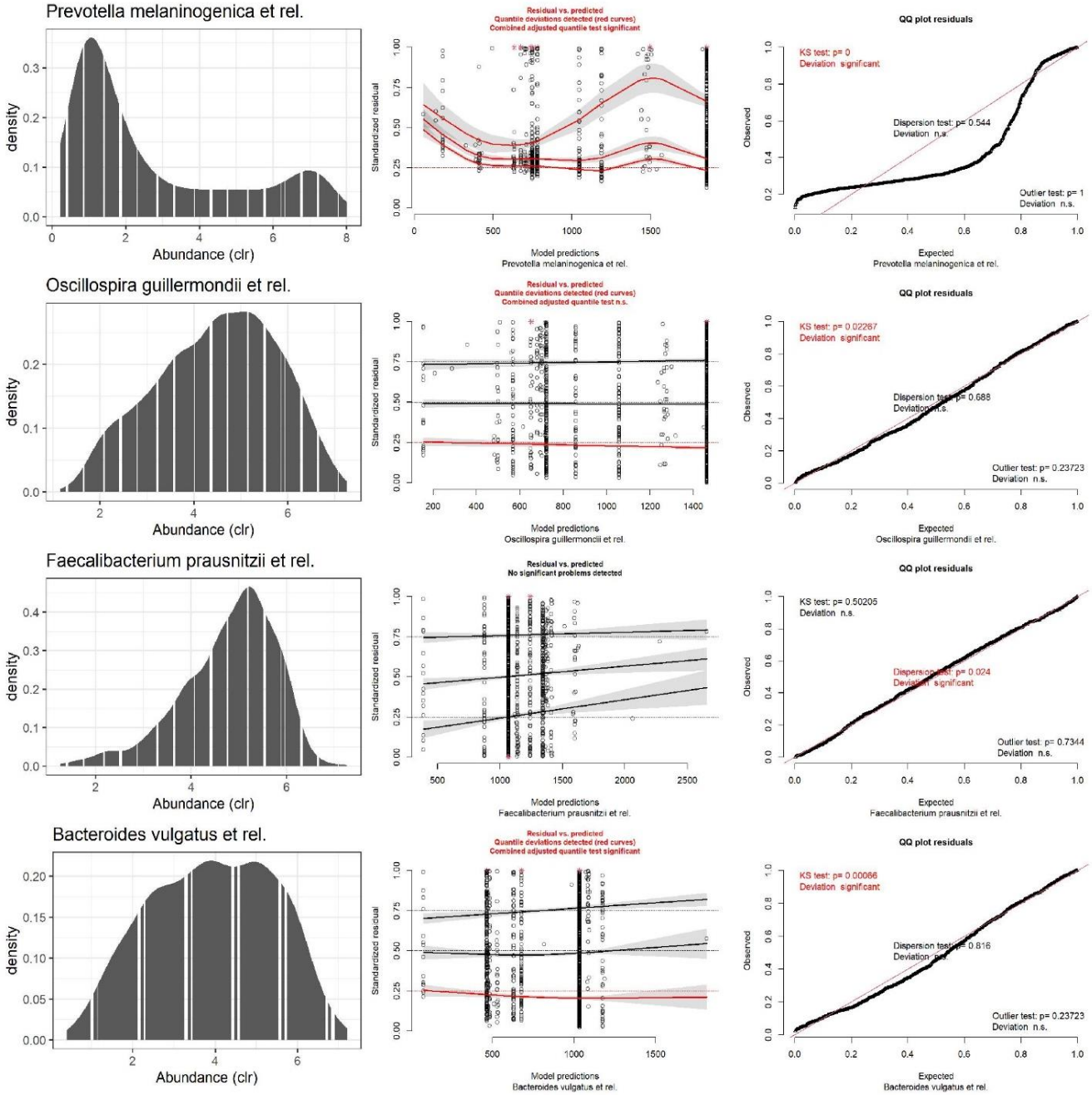
Akkermansia	87,122	87,157	11,537	11,572	75,585	75,585
Papillibacter cinnamivorans et rel.	49,434	49,469	11,300	11,335	38,134	38,134
Eubacterium rectale et rel.	46,181	46,216	11,270	11,305	34,911	34,911
Ruminococcus gnavus et rel.	53,829	53,865	11,332	11,367	42,497	42,498
Bacteroides ovatus et rel.	65,230	65,265	11,211	11,246	54,019	54,019
Ruminococcus lactaris et rel.	93,706	93,741	10,810	10,845	82,896	82,896
Bacteroides plebeius et rel.	49,718	49,754	10,498	10,533	39,220	39,221
Anaerotruncus colihominis et rel.	43,417	43,452	10,174	10,209	33,243	33,243
Escherichia coli et rel.	166,998	167,033	9,405	9,440	157,593	157,593
Bacteroides splachnicus et rel.	28,102	28,137	9,878	9,913	18,224	18,224
Roseburia intestinalis et rel.	34,364	34,399	10,058	10,094	24,306	24,305
Tannerella et rel.	28,823	28,858	9,917	9,952	18,906	18,906
Streptococcus mitis et rel.	51,555	51,591	10,037	10,072	41,518	41,519
Sutterella wadsworthia et rel.	54,931	54,967	9,831	9,866	45,100	45,101
Clostridium (sensu stricto)	31,255	31,290	9,486	9,521	21,769	21,769
Uncultured Mollicutes	75,466	75,501	9,749	9,785	65,717	65,716
Prevotella tannerae et rel.	38,685	38,720	9,641	9,676	29,044	29,044
Clostridium colinum et rel.	34,295	34,330	9,339	9,375	24,956	24,955
Collinsella	49,948	49,983	9,236	9,271	40,712	40,712
Anaerovorax odorimutans et rel.	15,732	15,767	8,532	8,567	7,200	7,200
Eubacterium biforme et rel.	38,059	38,094	8,834	8,869	29,225	29,225
Clostridium stercorarium et rel.	20,556	20,591	8,660	8,695	11,896	11,896
Bacteroides stercoris et rel.	20,086	20,121	8,531	8,566	11,555	11,555

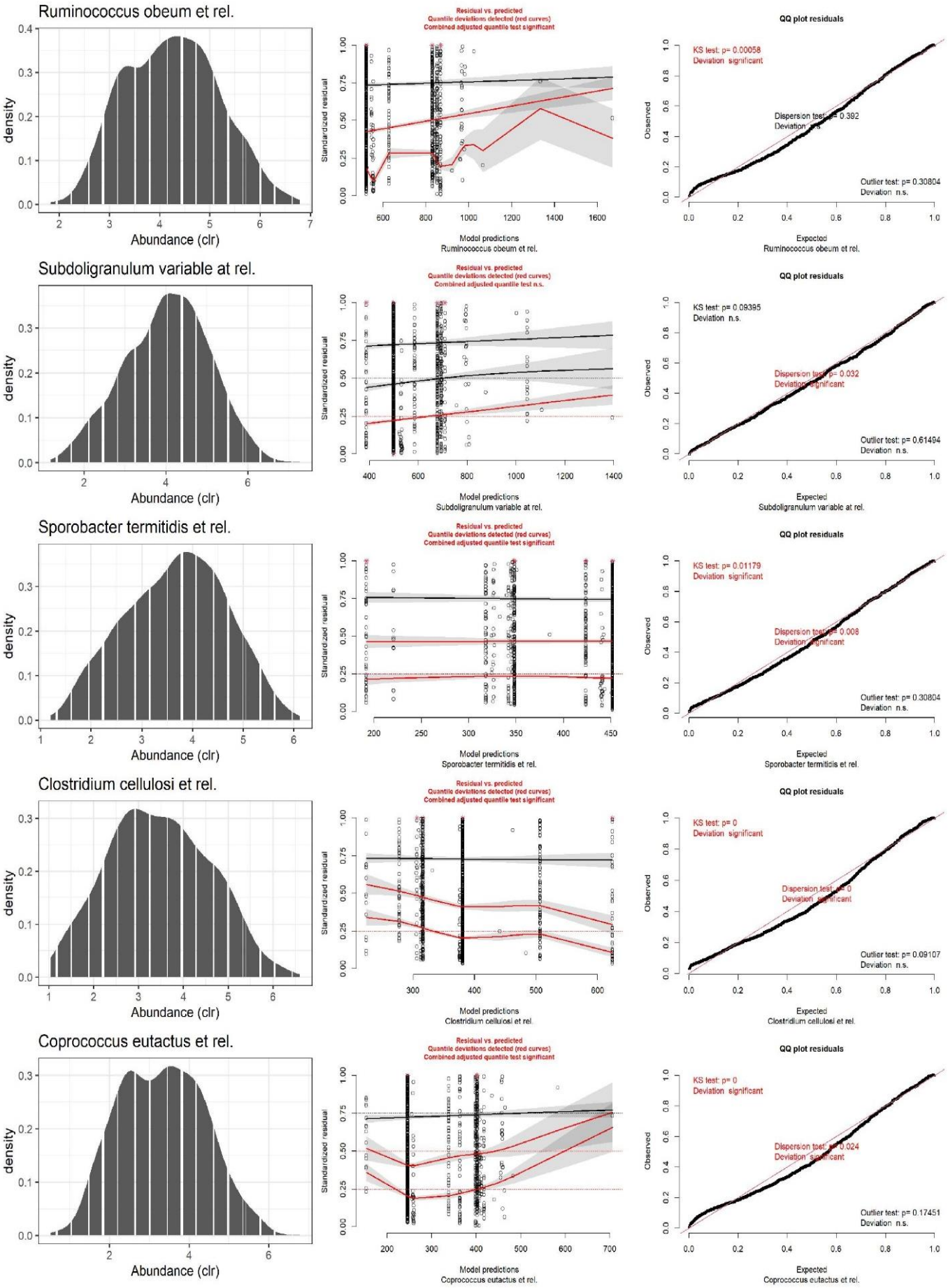
Phascolarctobacterium faecium et rel.	30,270	30,306	8,451	8,486	21,819	21,820
Oxalobacter formigenes et rel.	20,538	20,573	8,322	8,357	12,216	12,216
Lactobacillus plantarum et rel.	11,515	11,550	7,358	7,394	4,157	4,156
Mitsuokella multiacida et rel.	62,765	62,800	6,414	6,449	56,351	56,351
Lactobacillus gasseri et rel.	20,714	20,749	7,451	7,486	13,263	13,263

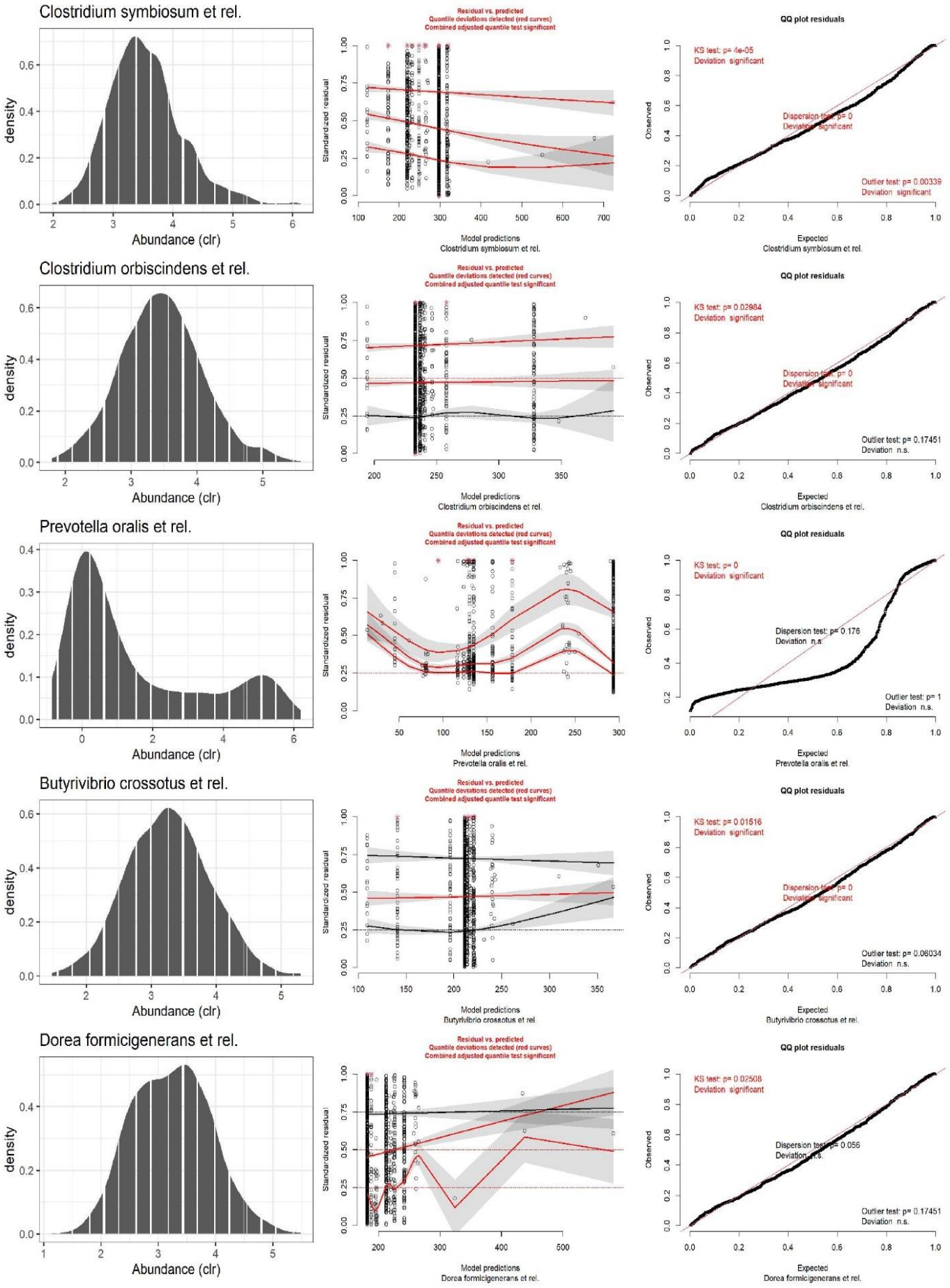
Appendix 3

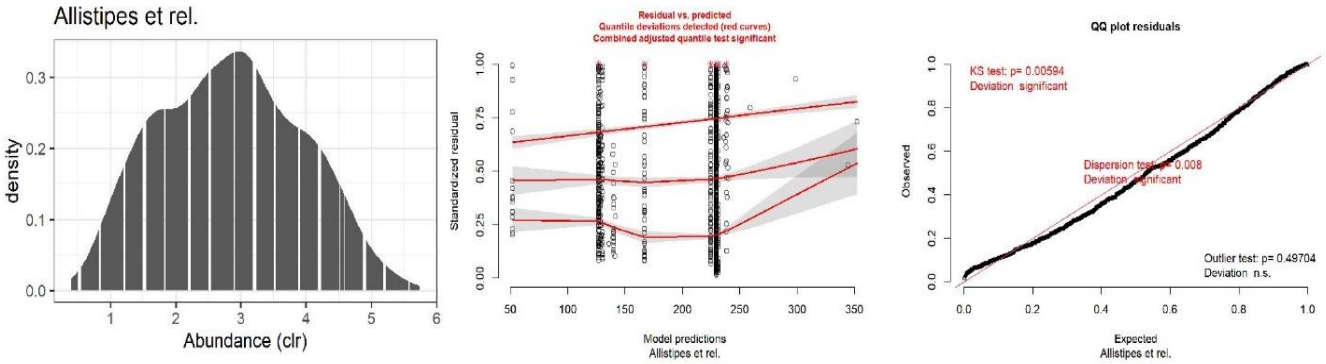
CLR, Q-Q and residual plots for the prevalent counts.

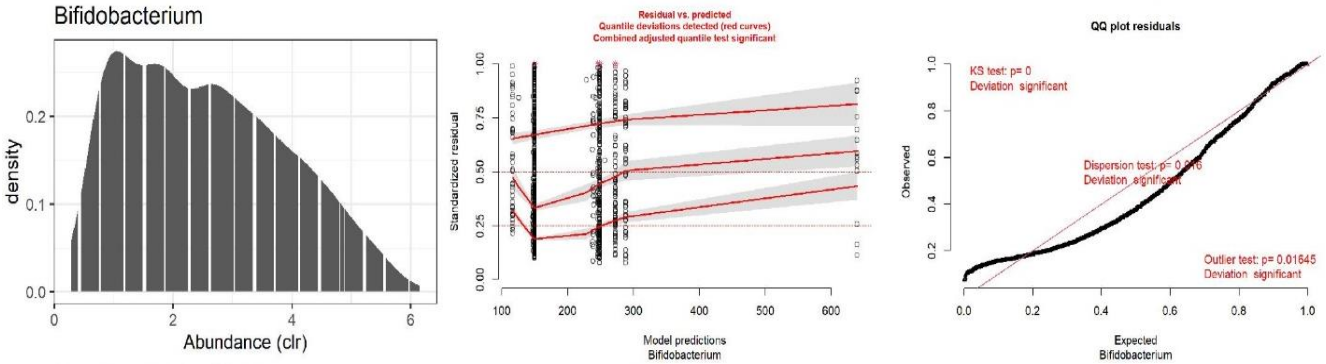
Q-Q plot includes Kolmogorov-Smirnov (KS) test for goodness of fit, dispersion test and outlier test.

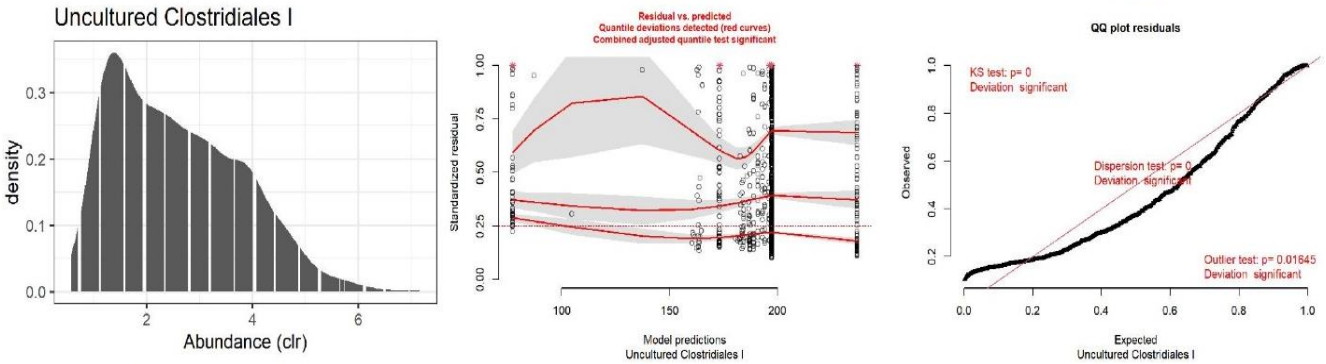


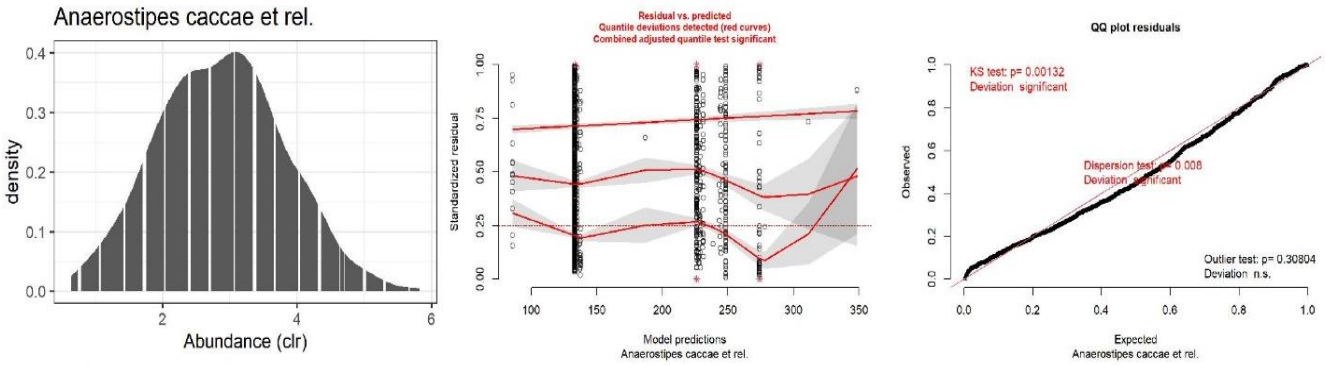

Ruminococcus obeum et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Ruminococcus obeum et rel.
QQ plot residuals
KS test: p= 0.00058
Deviation significant
Dispersion test: p= 0.392
Deviation n.s.
Outlier test: p= 0.30804
Deviation n.s.
Observed
Expected
Ruminococcus obeum et rel.
Subdoligranulum variable at rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test n.s.
Standardized residual
Model predictions
Subdoligranulum variable at rel.
QQ plot residuals
KS test: p= 0.09385
Deviation n.s.
Dispersion test: p= 0.032
Deviation significant
Outlier test: p= 0.61494
Deviation n.s.
Observed
Expected
Subdoligranulum variable at rel.
Sporobacter termitidis et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Sporobacter termitidis et rel.
QQ plot residuals
KS test: p= 0.01179
Deviation significant
Dispersion test: p= 0.008
Deviation significant
Outlier test: p= 0.30804
Deviation n.s.
Observed
Expected
Sporobacter termitidis et rel.
Clostridium cellulosi et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Clostridium cellulosi et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.09107
Deviation n.s.
Observed
Expected
Clostridium cellulosi et rel.
Coprococcus eutactus et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Coprococcus eutactus et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0.024
Deviation significant
Outlier test: p= 0.17451
Deviation n.s.
Observed
Expected
Coprococcus eutactus et rel.

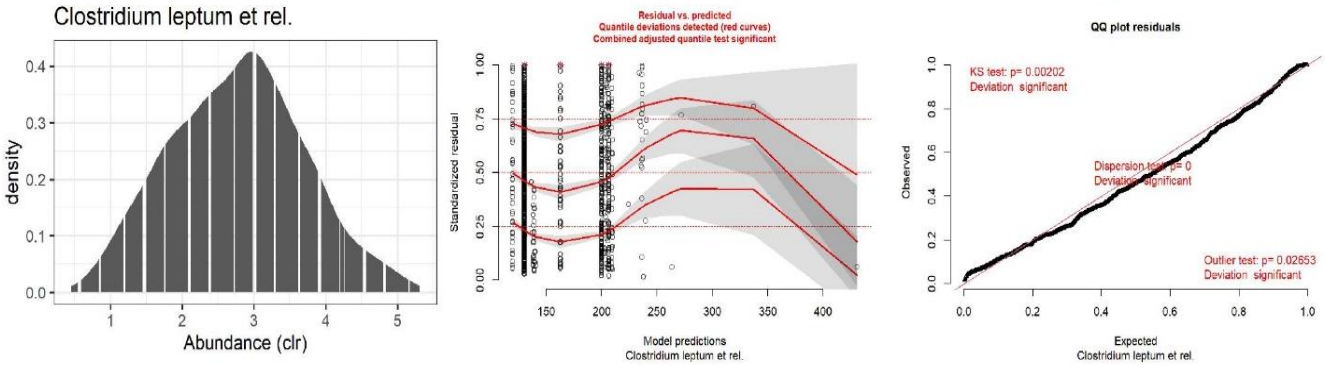

Clostridium symbiosum et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Clostridium symbiosum et rel.
QQ plot residuals
KS test: p= 4e-05
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.00339
Deviation significant
Observed
Expected
Clostridium symbiosum et rel.
Clostridium orbiscindens et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Clostridium orbiscindens et rel.
QQ plot residuals
KS test: p= 0.02984
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.17451
Deviation n.s.
Observed
Expected
Clostridium orbiscindens et rel.
Prevotella oralis et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Prevotella oralis et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0.176
Deviation n.s.
Outlier test: p= 1
Deviation n.s.
Observed
Expected
Prevotella oralis et rel.
Butyrivibrio crossotus et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Butyrivibrio crossotus et rel.
QQ plot residuals
KS test: p= 0.01516
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.06034
Deviation n.s.
Observed
Expected
Butyrivibrio crossotus et rel.
Dorea formicigenerans et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Dorea formicigenerans et rel.
QQ plot residuals
KS test: p= 0.02508
Deviation significant
Dispersion test: p= 0.056
Deviation n.s.
Outlier test: p= 0.17451
Deviation n.s.
Observed
Expected
Dorea formicigenerans et rel.

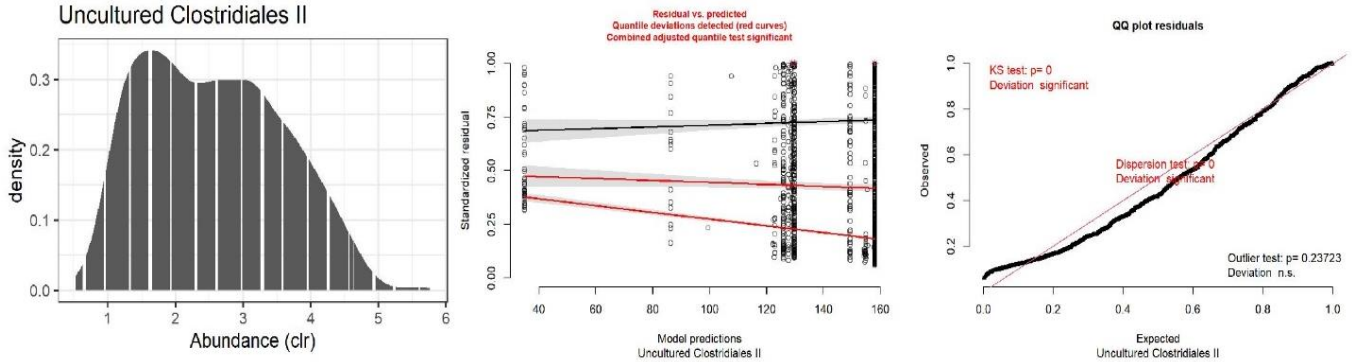

Allistipes et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Allistipes et rel.
QQ plot residuals
KS test: p= 0.00594
Deviation significant
Dispersion test: p= 0.008
Deviation significant
Outlier test: p= 0.49704
Deviation n.s.
Observed
Expected
Allistipes et rel.

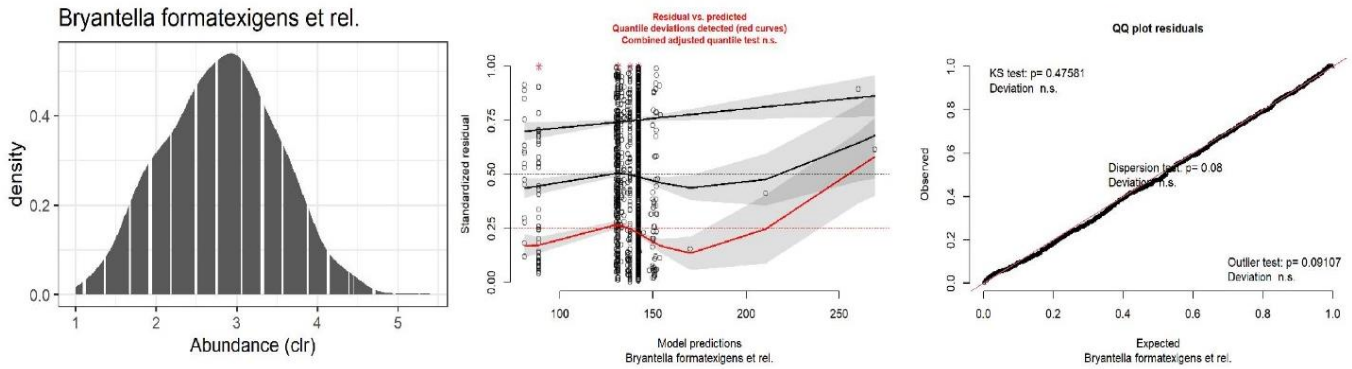

Bifidobacterium
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Bifidobacterium
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.01645
Deviation significant
Observed
Expected
Bifidobacterium

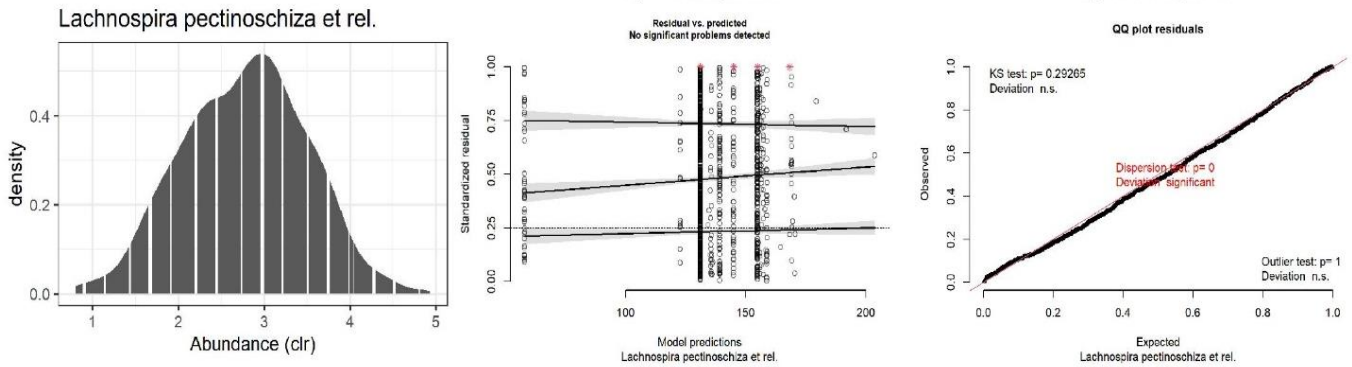

Uncultured Clostridiales I
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Uncultured Clostridiales I
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.01645
Deviation significant
Observed
Expected
Uncultured Clostridiales I

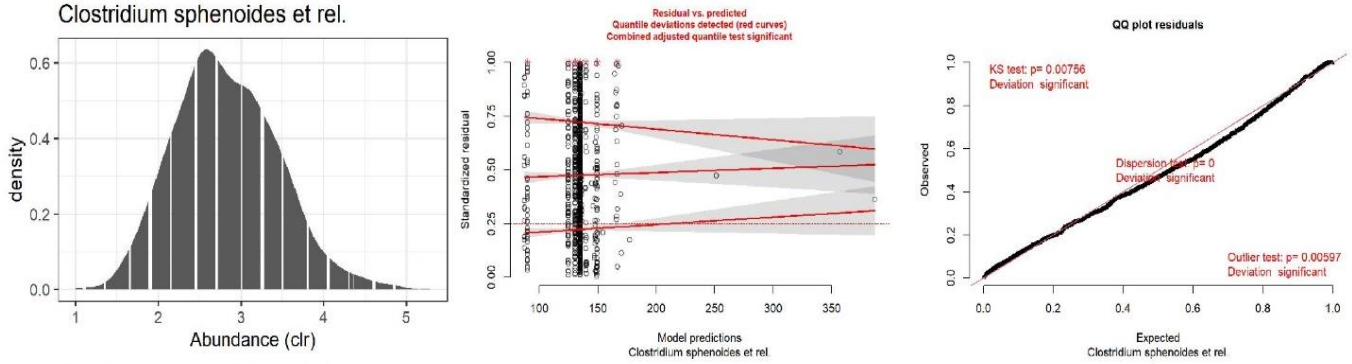

Anaerostipes caccae et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Anaerostipes caccae et rel.
QQ plot residuals
KS test: p= 0.00132
Deviation significant
Dispersion test: p= 0.008
Deviation significant
Outlier test: p= 0.30804
Deviation n.s.
Observed
Expected
Anaerostipes caccae et rel.

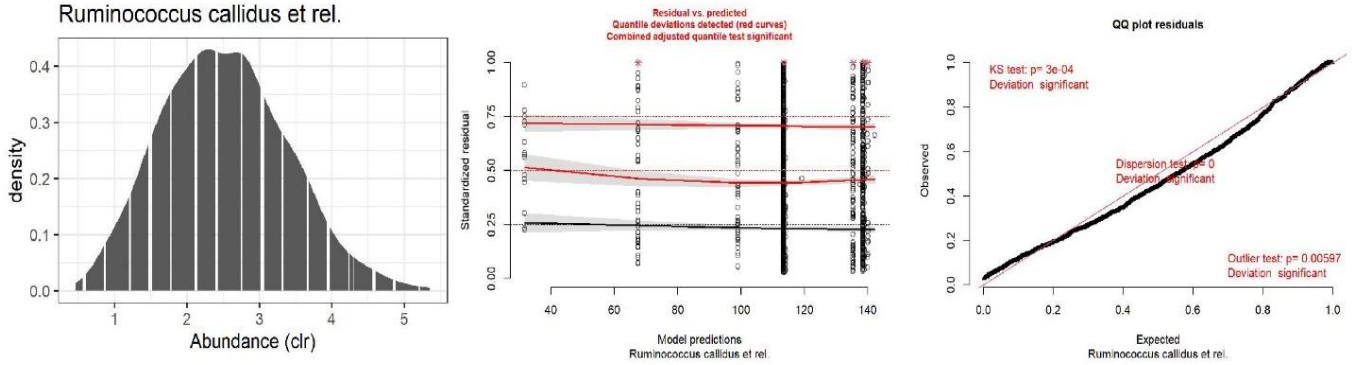

Clostridium leptum et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Clostridium leptum et rel.
QQ plot residuals
KS test: p= 0.00202
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.02653
Deviation significant
Observed
Expected
Clostridium leptum et rel.

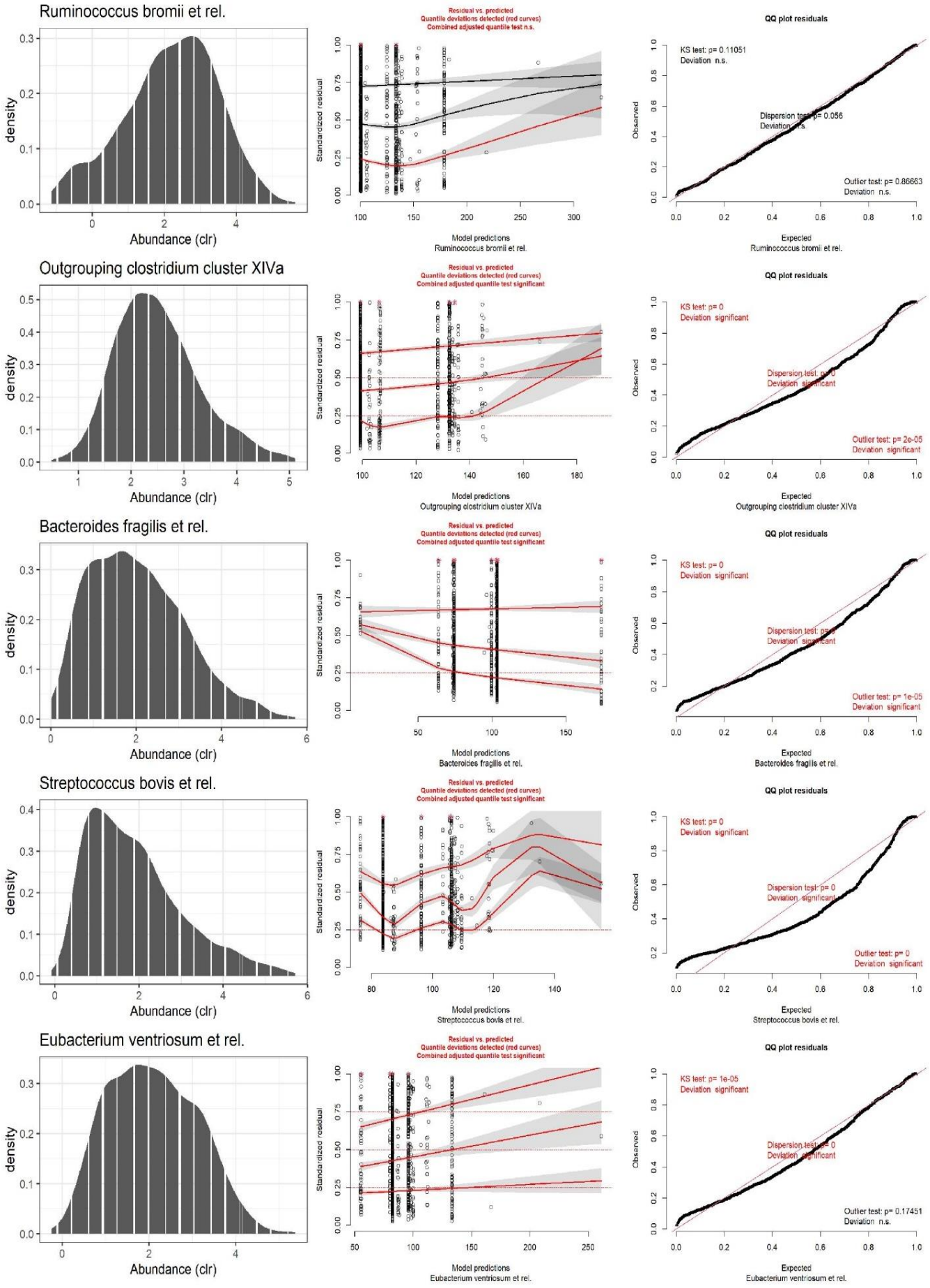

Uncultured Clostridiales II
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Uncultured Clostridiales II
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.23723
Deviation n.s.
Observed
Expected
Uncultured Clostridiales II

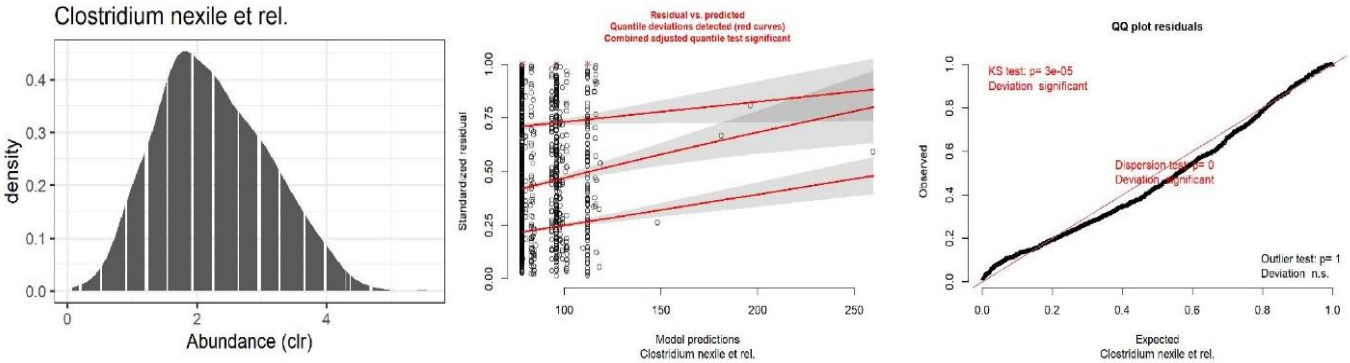

Bryantella formatexigens et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test n.s.
Standardized residual
Model predictions
Bryantella formatexigens et rel.
QQ plot residuals
KS test: p= 0.47581
Deviation n.s.
Dispersion test: p= 0.08
Deviation n.s.
Outlier test: p= 0.09107
Deviation n.s.
Observed
Expected
Bryantella formatexigens et rel.

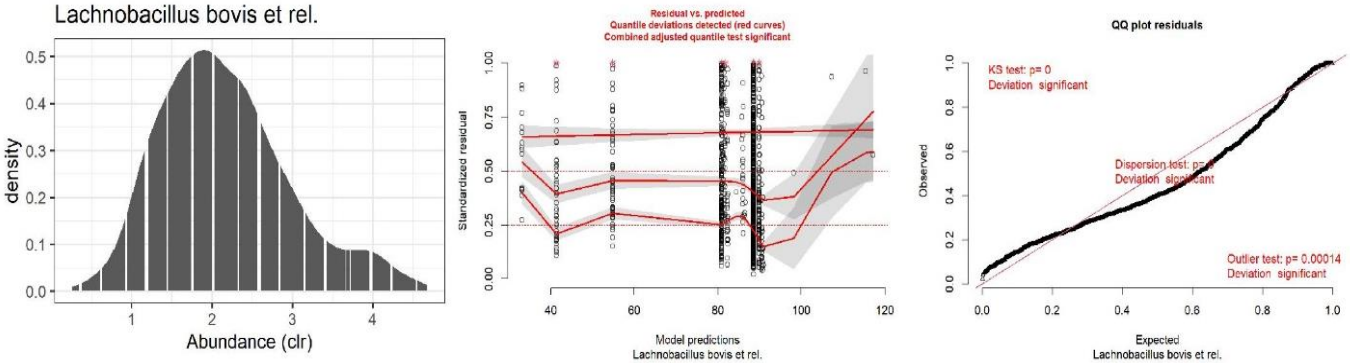

Lachnospira pectinoschiza et rel.
density
Abundance (clr)
Residual vs. predicted
No significant problems detected
Standardized residual
Model predictions
Lachnospira pectinoschiza et rel.
QQ plot residuals
KS test: p= 0.29265
Deviation n.s.
Dispersion test: p= 0
Deviation significant
Outlier test: p= 1
Deviation n.s.
Observed
Expected
Lachnospira pectinoschiza et rel.

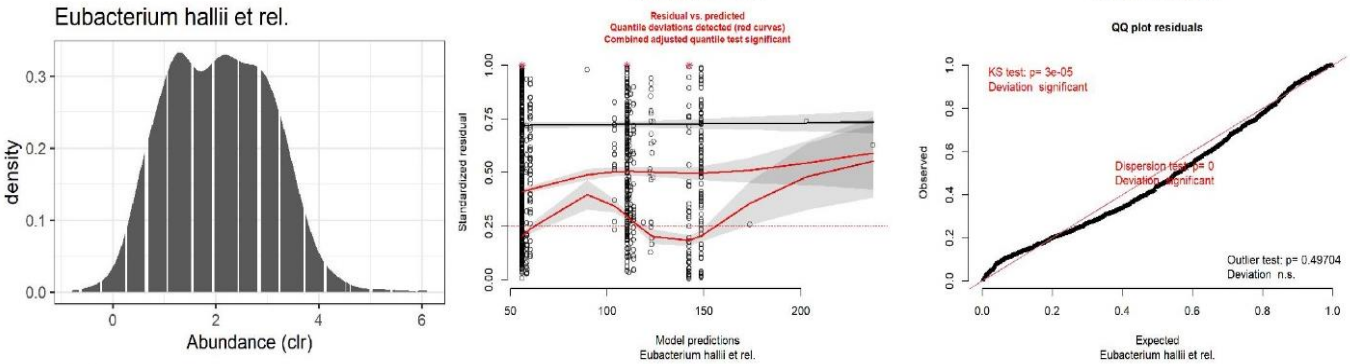

Clostridium sphenoides et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Clostridium sphenoides et rel.
QQ plot residuals
KS test: p= 0.00756
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.00597
Deviation significant
Observed
Expected
Clostridium sphenoides et rel.

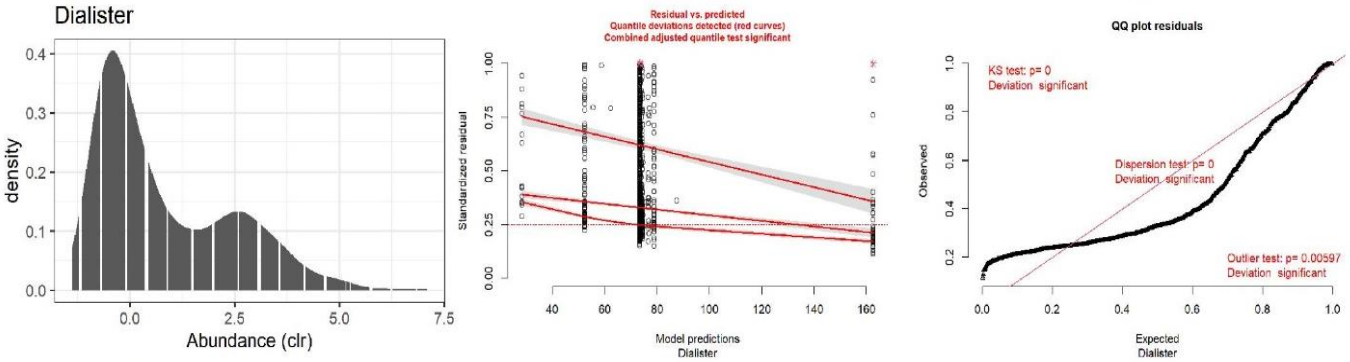

Ruminococcus callidus et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Ruminococcus callidus et rel.
QQ plot residuals
KS test: p= 3e-04
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.00597
Deviation significant
Observed
Expected
Ruminococcus callidus et rel.

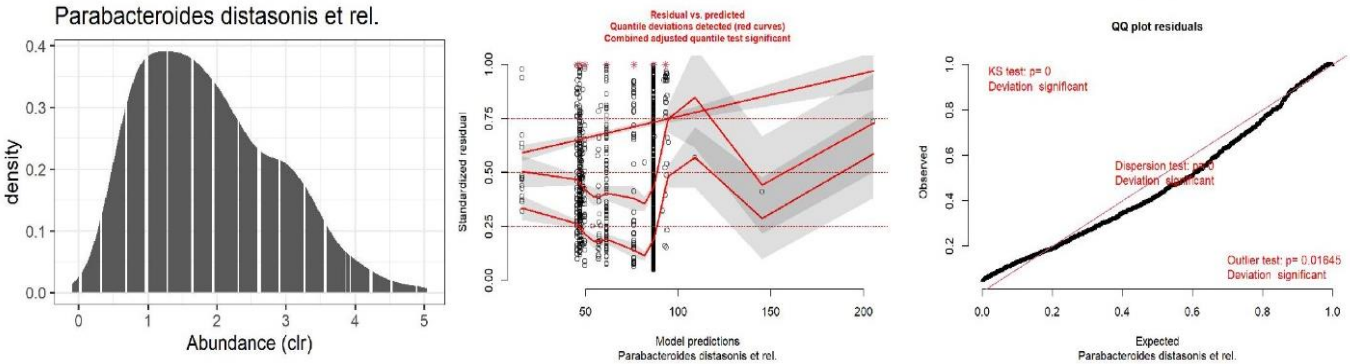

Ruminococcus bromii et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test n.s.
Standardized residual
Model predictions
Ruminococcus bromii et rel.
QQ plot residuals
KS test: p= 0.11051
Deviation n.s.
Dispersion test: p= 0.056
Deviation n.s.
Outlier test: p= 0.86663
Deviation n.s.
Observed
Expected
Ruminococcus bromii et rel.
Outgrouping clostridium cluster XIVa
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Outgrouping clostridium cluster XIVa
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 2e-05
Deviation significant
Observed
Expected
Outgrouping clostridium cluster XIVa
Bacteroides fragilis et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Bacteroides fragilis et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 1e-05
Deviation significant
Observed
Expected
Bacteroides fragilis et rel.
Streptococcus bovis et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Streptococcus bovis et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0
Deviation significant
Observed
Expected
Streptococcus bovis et rel.
Eubacterium ventriosum et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Eubacterium ventriosum et rel.
QQ plot residuals
KS test: p= 1e-05
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.17451
Deviation n.s.
Observed
Expected
Eubacterium ventriosum et rel.

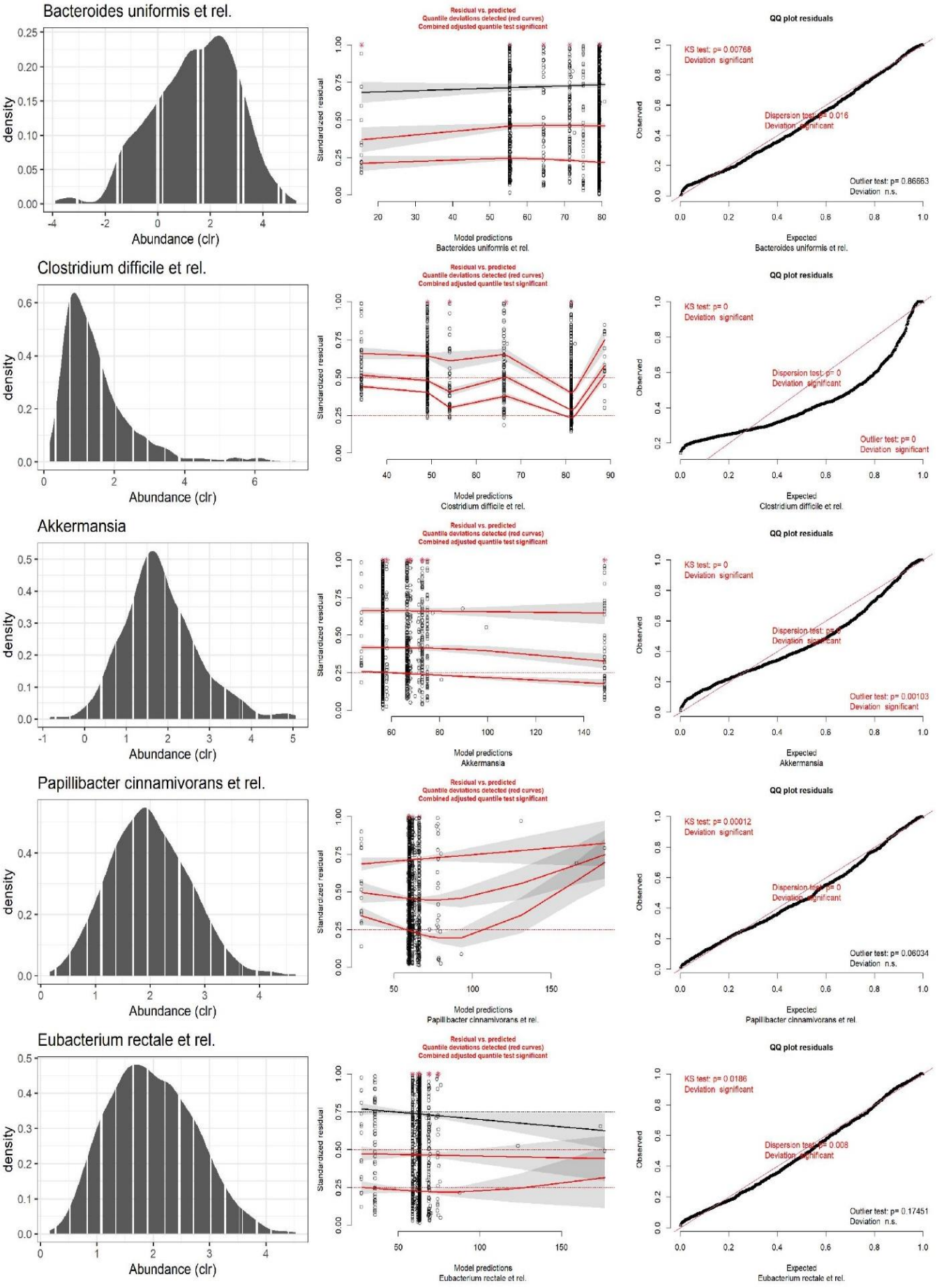

Clostridium nexile et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Clostridium nexile et rel.
QQ plot residuals
KS test: p= 3e-05
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 1
Deviation n.s.
Observed
Expected
Clostridium nexile et rel.

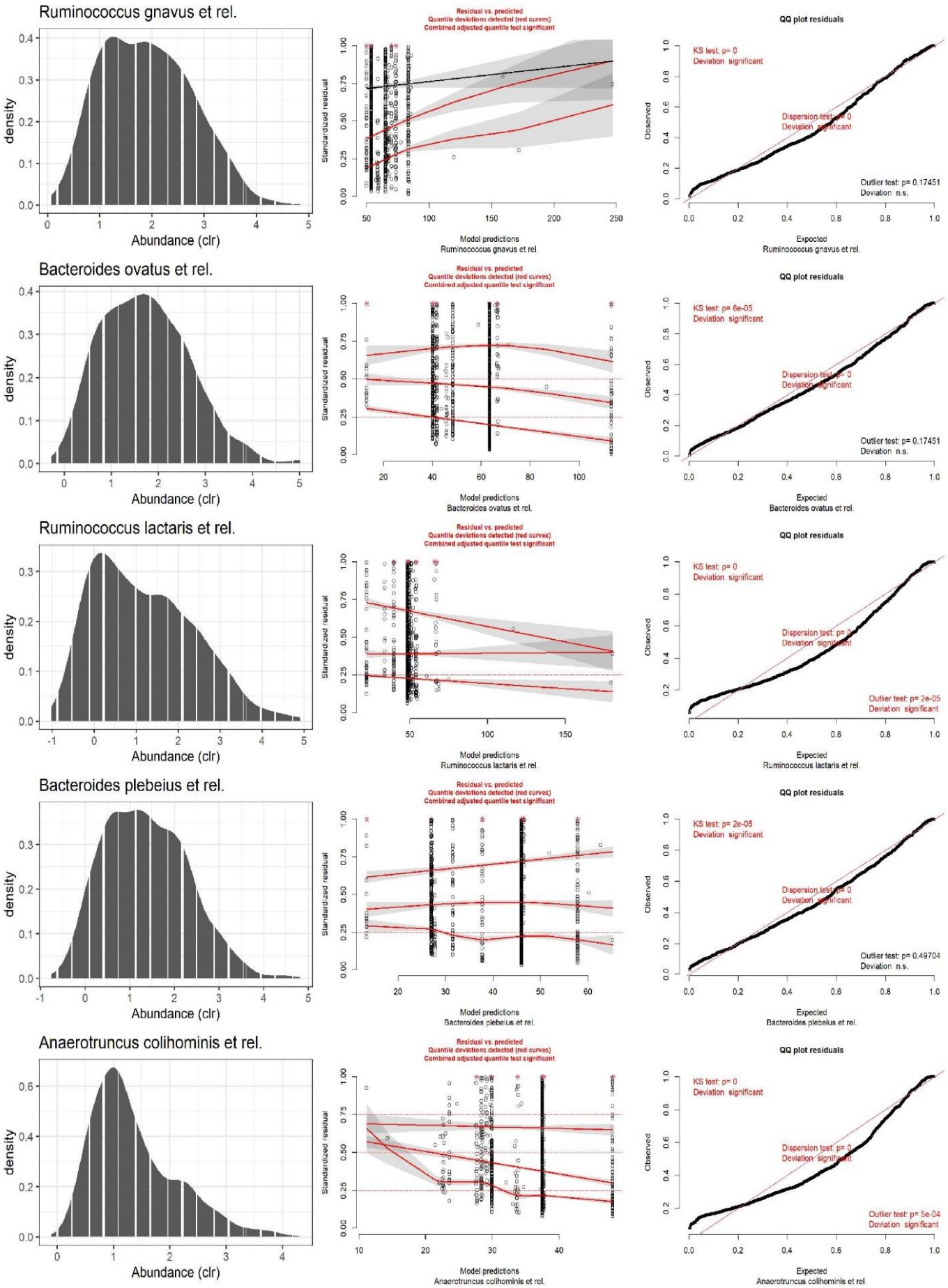

Lachnobacillus bovis et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Lachnobacillus bovis et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.00014
Deviation significant
Observed
Expected
Lachnobacillus bovis et rel.

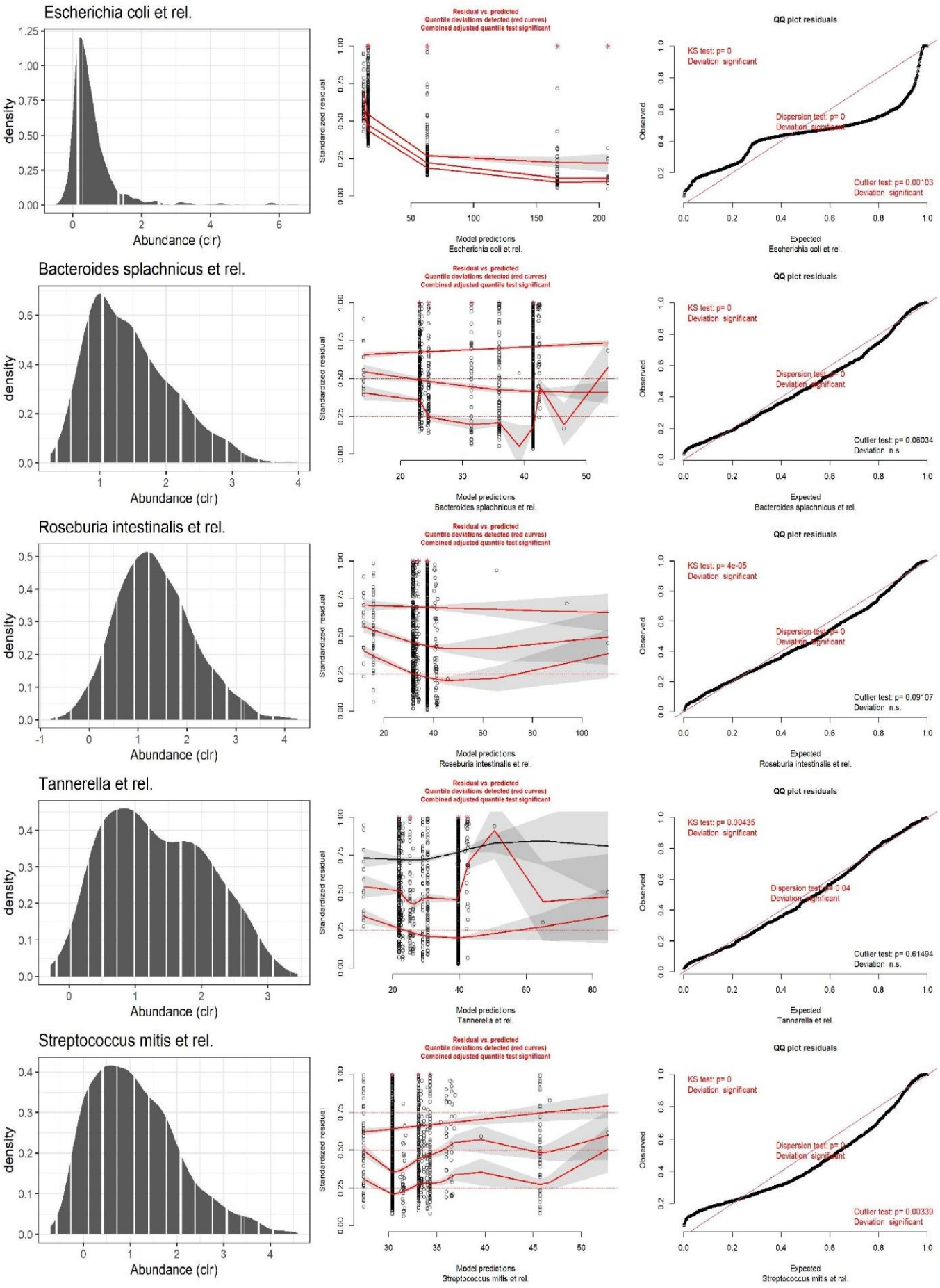

Eubacterium hallii et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Eubacterium hallii et rel.
QQ plot residuals
KS test: p= 3e-05
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.49704
Deviation n.s.
Observed
Expected
Eubacterium hallii et rel.

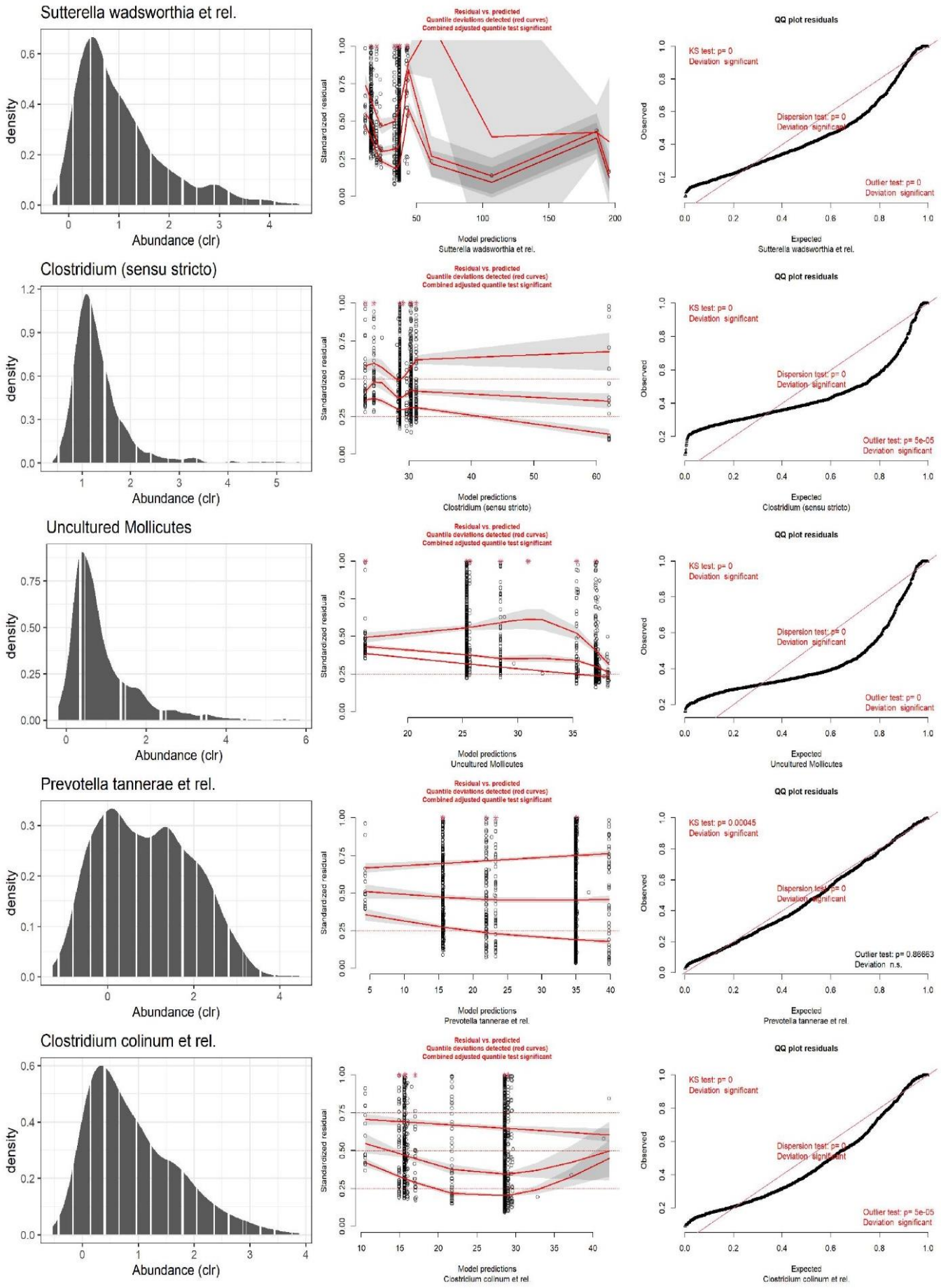

Dialister
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Dialister
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.00597
Deviation significant
Observed
Expected
Dialister

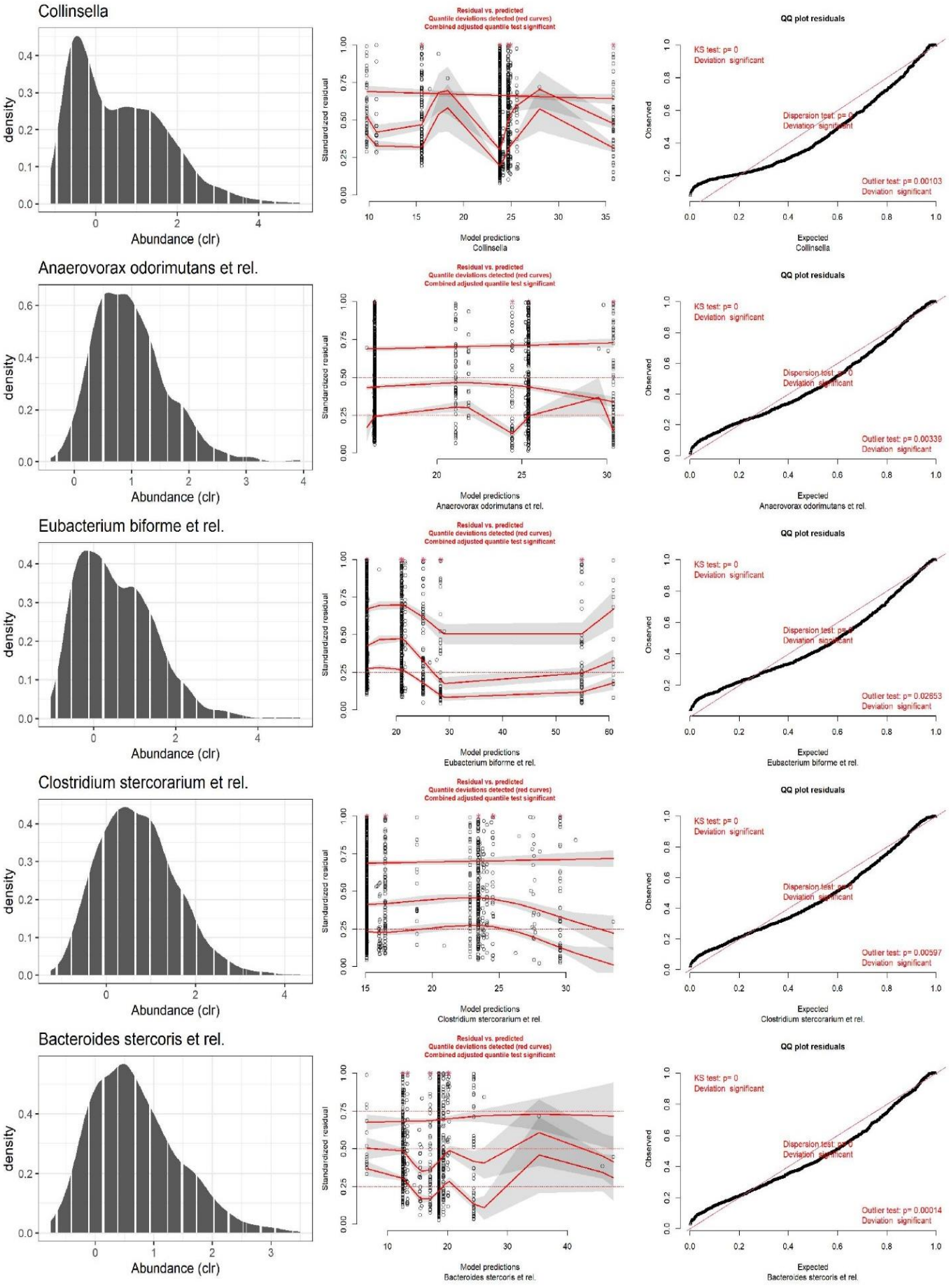

Parabacteroides distasonis et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Parabacteroides distasonis et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.01645
Deviation significant
Observed
Expected
Parabacteroides distasonis et rel.

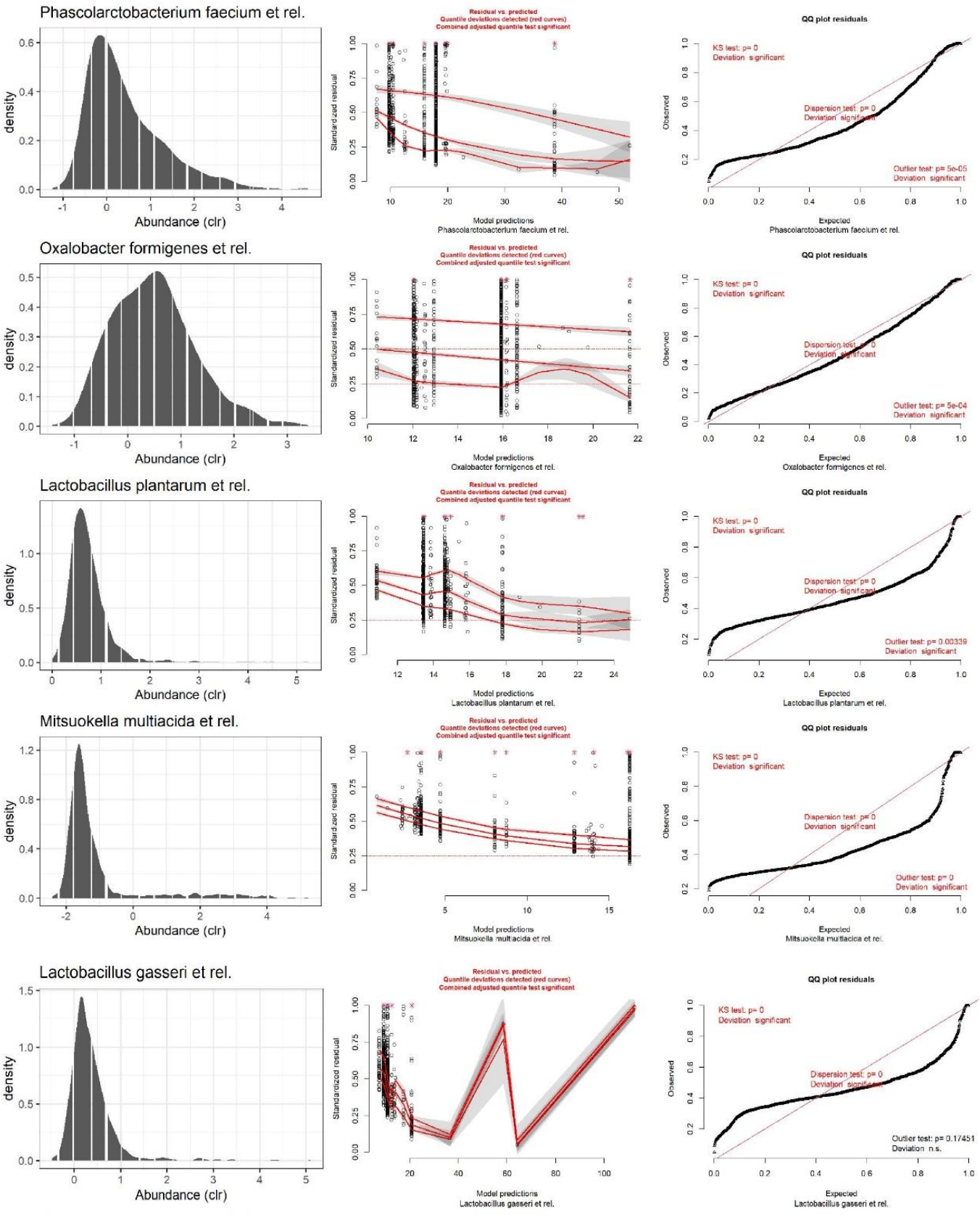

Bacteroides uniformis et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Bacteroides uniformis et rel.
QQ plot residuals
KS test: p= 0.00768
Deviation significant
Dispersion test: p= 0.016
Deviation significant
Outlier test: p= 0.86663
Deviation n.s.
Observed
Expected
Bacteroides uniformis et rel.
Clostridium difficile et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Clostridium difficile et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0
Deviation significant
Observed
Expected
Clostridium difficile et rel.
Akkermansia
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Akkermansia
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.00103
Deviation significant
Observed
Expected
Akkermansia
Papillibacter cinnamivorans et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Papillibacter cinnamivorans et rel.
QQ plot residuals
KS test: p= 0.00012
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.06034
Deviation n.s.
Observed
Expected
Papillibacter cinnamivorans et rel.
Eubacterium rectale et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Eubacterium rectale et rel.
QQ plot residuals
KS test: p= 0.0186
Deviation significant
Dispersion test: p= 0.008
Deviation significant
Outlier test: p= 0.17451
Deviation n.s.
Observed
Expected
Eubacterium rectale et rel.


Ruminococcus gnavus et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Ruminococcus gnavus et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.17451
Deviation n.s.
Observed
Expected
Ruminococcus gnavus et rel.
Bacteroides ovatus et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Bacteroides ovatus et rel.
QQ plot residuals
KS test: p= 6e-05
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.17451
Deviation n.s.
Observed
Expected
Bacteroides ovatus et rel.
Ruminococcus lactaris et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Ruminococcus lactaris et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 2e-05
Deviation significant
Observed
Expected
Ruminococcus lactaris et rel.
Bacteroides plebeius et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Bacteroides plebeius et rel.
QQ plot residuals
KS test: p= 2e-05
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.49704
Deviation n.s.
Observed
Expected
Bacteroides plebeius et rel.
Anaerotruncus colihominis et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Anaerotruncus colihominis et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 5e-04
Deviation significant
Observed
Expected
Anaerotruncus colihominis et rel.


Escherichia coli et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Escherichia coli et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.00103
Deviation significant
Observed
Expected
Escherichia coli et rel.
Bacteroides splachnicus et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Bacteroides splachnicus et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.06034
Deviation n.s.
Observed
Expected
Bacteroides splachnicus et rel.
Roseburia intestinalis et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Roseburia intestinalis et rel.
QQ plot residuals
KS test: p= 4e-05
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.09107
Deviation n.s.
Observed
Expected
Roseburia intestinalis et rel.
Tannerella et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Tannerella et rel.
QQ plot residuals
KS test: p= 0.00435
Deviation significant
Dispersion test: p= 0.04
Deviation significant
Outlier test: p= 0.61494
Deviation n.s.
Observed
Expected
Tannerella et rel.
Streptococcus mitis et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Streptococcus mitis et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.00339
Deviation significant
Observed
Expected
Streptococcus mitis et rel.


Sutterella wadsworthia et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Sutterella wadsworthia et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0
Deviation significant
Observed
Expected
Sutterella wadsworthia et rel.
Clostridium (sensu stricto)
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Clostridium (sensu stricto)
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 5e-05
Deviation significant
Observed
Expected
Clostridium (sensu stricto)
Uncultured Mollicutes
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Uncultured Mollicutes
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0
Deviation significant
Observed
Expected
Uncultured Mollicutes
Prevotella tannerae et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Prevotella tannerae et rel.
QQ plot residuals
KS test: p= 0.00045
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.88663
Deviation n.s.
Observed
Expected
Prevotella tannerae et rel.
Clostridium colinum et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Clostridium colinum et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 5e-05
Deviation significant
Observed
Expected
Clostridium colinum et rel.


Collinsella
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Collinsella
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.00103
Deviation significant
Observed
Expected
Collinsella
Anaerovorax odorimutans et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Anaerovorax odorimutans et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.00339
Deviation significant
Observed
Expected
Anaerovorax odorimutans et rel.
Eubacterium biforme et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Eubacterium biforme et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.02653
Deviation significant
Observed
Expected
Eubacterium biforme et rel.
Clostridium stercorarium et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Clostridium stercorarium et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.00597
Deviation significant
Observed
Expected
Clostridium stercorarium et rel.
Bacteroides stercoris et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Bacteroides stercoris et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.00014
Deviation significant
Observed
Expected
Bacteroides stercoris et rel.


Phascolarctobacterium faecium et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Phascolarctobacterium faecium et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 5e-05
Deviation significant
Observed
Expected
Phascolarctobacterium faecium et rel.
Oxalobacter formigenes et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Oxalobacter formigenes et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 5e-04
Deviation significant
Observed
Expected
Oxalobacter formigenes et rel.
Lactobacillus plantarum et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Lactobacillus plantarum et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.00339
Deviation significant
Observed
Expected
Lactobacillus plantarum et rel.
Mitsuokella multiacida et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Mitsuokella multiacida et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0
Deviation significant
Observed
Expected
Mitsuokella multiacida et rel.
Lactobacillus gasseri et rel.
density
Abundance (clr)
Residual vs. predicted
Quantile deviations detected (red curves)
Combined adjusted quantile test significant
Standardized residual
Model predictions
Lactobacillus gasseri et rel.
QQ plot residuals
KS test: p= 0
Deviation significant
Dispersion test: p= 0
Deviation significant
Outlier test: p= 0.17451
Deviation n.s.
Observed
Expected
Lactobacillus gasseri et rel.