A priori estimation of sequencing effort in complex microbial metatranscriptomes

Toni Monleon-Getino[1*], Jorge Frias-Lopez[2]

[1] Section of Statistics (Department of Genetics, Microbiology, and Statistics). University of Barcelona. Barcelona. Spain. BIOST[3], GRBIO (Research Group in Biostatistics and Bioinformatics).

[2] University of Florida, College of Dentistry, Gainesville, FL 32610, USA.

* Corresponding Author: amonleong@ub.edu

Address: Section of Statistics (Department of Genetics, Microbiology, and Statistics). Faculty of Biology. Avda Digonal 643. 08028 Barcelona (Spain)

Abstract

1. Accurate differential expression of microbial metatranscriptomes based on Next Generation Sequencing depends partly on the depth of the libraries used to perform the analysis. Therefore, estimating the sequencing depth required to sample the metatranscriptome of interest using RNA-seq effectively is an essential first step to both obtain robust results in further analysis and avoiding over-expending once the information contained in the library reaches saturation.

2. Here we present a method to calculate the effort in saturation curves and *a priori* genes prediction using a simulated series of metatranscriptomic/metagenomic matrices. This method is based on the extrapolation rarefaction curve using a Weibull growth model to estimate the maximum number of genes/OTUs as a function of sequencing depth using a machine learning approach. This approach allows us to compute the effort at different confidence intervals and to obtain an approximate *a priori* effort using based on an initial fraction of sequences.

3. The accuracy of the results obtained with simulations and real samples (15 datasets of metatranscriptomes from the oral cavity, RNA sequences consist of vectors of 105-1.5x107 reads depth with a 10000 and 600000 genes size) allows one to use an initial shallowly sequenced sample (in this case 20% of the total amount of reads sampled; accuracy R^2>0.99 simulated samples and 60-93% for real samples) to estimate the expected sequencing effort needed to cover the whole metatranscriptome/ metagenome from the same sample, so can be used to estimate the estimate the sample size. The algorithm containing the proposed method was saved as a function for R.

4. This proposed method of estimation of the maximum number of gene/OTUs, reads to reach 90, 95 and 99% of maximum number of gene/OTUs, using and algorithm based on rarefaction curve + Weibull model + machine learning prediction, is efficient to help researchers to know if the sampling is sufficient or otherwise need to be increased. The analytical pipeline presented here may be successfully used for the in-depth and time-effective characterization of complex microbial communities, representing a useful tool for the microbiome research community.

Keywords: machine learning, metatranscriptomics, metagenomics, NGS, rarefaction curve, sample size, sequencing effort, simulation.

1. Introduction

The study of the human microbiome has dramatically expanded our understanding of the role that microbes play in health and disease. These kinds of studies have been facilitated by the development of technologies for Next Generation Sequencing (NGS), which are capable of generating enough number of sequences as to cover most of the diversity present in the sample. However, capturing the full composition is still a challenge even when estimating the composition by SSU rDNA analysis (Ni, Yan & Yu, 2013; Tamames, de la Peña & de Lorenzo, 2012). Garcia-Ortega & Martinez (2015) using a non-parametric estimator for the number of undetected genes found that on average approximately 10% of the expressed genes per accession remain undetected if individual sequencing libraries are analysed.

The power and accuracy of such experiments depend substantially on the number of reads sequenced, so a crucial step in the experiment design should be to determine the optimal read depth for a particular study or to verify whether one has adequate depth in an actual experiment (Robinson & Storey, 2014).

In the case of RNA-seq studies, most work has been done on assessing sequencing depth on the transcriptome of eukaryotic systems, with a wide range of estimated sequencing depths to cover the full patterns of expression. In the human transcriptome, the estimated numbers of sequencing depth necessary to observe differences in expression profiles vary from 100 to 700 million sequences (Westermann, Gorski & Vogel, 2012; Toung, Morley, Li & Cheung, 2011). In the case of prokaryotic RNA-seq experiments Haas et al. have shown that reads typically produced in a single lane of the Illumina HiSeq sequencer far exceeds the number needed to saturate the annotated transcriptomes of diverse bacteria growing in monoculture (Haas, Chin, Nusbaum, Birren & Livny, 2012).

In NGS technology, saturation would be reached when an increment in the number of reads does not result in additional true expressed transcripts being detected, in the case of metatranscriptome analysis or no additional ORF/OTUs are detected in the case of metagenomic analysis. One way of estimating the point of saturation is by using rarefaction curves, a regular method to assess species richness from the results of sampling. These curves are commonly used in ecology to estimate the species richness as a function of sampling effort. In the case of RNA-seq/DNA-seq, a higher sequencing depth will only make the curve go

on longer but otherwise is comparable to a lower sequencing depth curve for the regions that both cover. Once the curve reaches a plateau, where additional sequencing would only marginally increase the number of transcripts seen, we consider that the curve is saturated. Another useful feature of saturation curves is that we can assess the complexity of the sample. Highly complex communities will have many transcripts being expressed, whereas communities which have low complexity would have a low number of transcripts being expressed.

We have developed a method to calculate the sequencing effort needed to reach the maximum number of existing genes (or operational taxonomic units (OUT) in the case of metagenomics) using rarefaction curves extrapolating from a small initial sequencing depth (10-20%). The method estimates the confidence intervals at 90, 95 or 99% of the maximum sequencing effort.

2. Material and methods

We first simulated more than a thousand of different metatranscriptomic/metagenomic matrices. On those matrices, we computed rarefaction curves using the function iNEXT() from the iNEXT () R library for Interpolation and Extrapolation for Species Diversity (Hsieh, Ma & Chao, 2017). We then used a non-linear growth model to compute the maximum number of genes expected/ OTUs and to estimate sequencing depth (reads) to reach 90, 95 or 99% of the maximum sampling effort.

Finally, using a method based on machine learning we predicted the maximum number of OTUs/genes using only a minimum number of sequencing depth (reads) to reach 90, 95 or 99% of the maximum and the sampling effort needed. All these functionalities were included in some functions of R. The method was tested, as an application thereof, on metatranscriptomic samples of oral microbiome. The results of these are presented in the additional material this article.

2.1 Simulation of metatranscriptomic/ metagenomic matrices

Metatranscriptomic/ metagenomics matrices were simulated as described in Rodriguez-Casado, Monleón-Getino, Cubedo & Ríos-Alcolea (2017) and in Monleón, Rodríguez-Casado & Verde (2019). In the table 1 it is possible to see the general metatranscriptomic/metagenomic matrix (***M***) structure (*n* rows: samples, *p* columns: genes or OTU) obtained after the bioinformatic analysis and that constitutes the starting point of this study. Below is his mathematical formalization and the study of his distribution of probabilities, studied more deeply in previous works (Monleon-Getino, Rodríguez-Casado & Verde, 2019).

Usually, for convenience, we change in ***M*** the notation of *p* by *k*; also during the statistical analysis we use the transpose ***M'***, which shows the samples (e.g. individuals) in the columns and the gene/organism identified in the rows (Table 1).

Table 1: data matrix structure of ***M'*** (metatranscriptomics or metagenomics matrix input).

num	Gene/Taxon	Sample 1	Sample 2	Sample *jth*	Sample *n*	Total
1	*type.1*	m_{11}	m_{12}	…	m_{1n}	$N_{1.}$
2	*type.2*	m_{21}	m_{22}	…	m_{2n}	$N_{2.}$
⋮	⋮	…	…	m_{ij}	…	…
k	*type.k*	m_{k1}	m_{k2}	…	m_{kn}	$N_{k.}$
	Total	$N_{.1}$	$N_{.2}$	...	$N_{.n}$	*N*

As a result of genomic analysis, ***M'*** can be very large and usually has thousands of genes/OTUs, most of them with small frequencies or 0, i.e. ***M'*** is typically a sparse matrix. This matrix is truncated, in the sense that there are characteristics that have not been observed in the sampling.

From the statistical point of view It is very convenient to formalize the probability distribution underlying this matrix structure, so each sample from ***M'*** can be represented by one k-dimensional random vector $X_j; X_j = (m_{1j}, m_{2j}, \dots, m_{kj})$, where m_{kj} represents the number of times that gene/taxa *k* is observed in sample *j*.

The probability distribution of each random vector $X_{i.}$ (vector row) and $X_{.j}$ (vector column) can be associate individually to a multinomial distribution,

$$X_{.j} \sim MN\left(N_{.j}, \theta_{1j}, \ldots, \theta_{kj}\right);\ \forall j = 1, .., n \quad (1)$$

$$X_{i.} \sim MN(i., \theta_{i1}, \ldots, \theta_{in});\ \forall i = 1, .., k \quad (2)$$

The multinomial distribution is a multivariate generalization of the binomial distribution, where the marginal distribution of each X_{ij} is:

$$X_{ij} \sim Bin\left(m_{ij}, \theta_{ij}\right); 1 \leq \theta_{ij} \leq 1;\ \forall j = 1, .., n; \forall i = 1, .., k \quad (3)$$

e.g. if we consider the partition of all sample space Ω^j the j-sample space in *k* parts:

$$A_{1j},\quad A_{2j,}\quad \ldots\quad , A_{kj}$$

One individual selected randomly has the probability θ_{kj} of belongs to the gene/taxon A_{kj} in the partition:

$$\left.\begin{array}{l} P\left(A_{1j}\right) = \theta_{1j} \\ P\left(A_{2j}\right) = \theta_{2j} \\ \vdots \\ P\left(A_{kj}\right) = \theta_{kj} \end{array}\right\} \sum_{i=1}^{k} \theta_{ij} = 1\ ;\ \forall j = 1, .., n \quad (4)$$

If we wish calculate for a sample j the probability of have $N_{.j}$ individuals, m_{1j} belonging to classA_{1j}, m_{2j} belongs to class A_{2j},..., m_{kj} belongs to class A_{kj}, with the restriction

$$\sum_{i\,=\,1}^{k} m_{ij} = N_{.j}\ ;\ \forall j = 1, .., n \quad (5)$$

Furthermore, using the multinomial function of density (mass function) we can calculate this probability, $MN\left(N_{.j}; \theta_j = \left(\theta_{1j}, \theta_{2j}, \ldots, \theta_{kj}\right)\right)$:

$$P\left[\left(A_{1j} = m_{1j}\right) \cap \ldots \cap \left(A_{hj} = n_{kj}\right)\right] = \frac{N_{.j}!}{m_{1j}! m_{2j}! \ldots m_{kj}!} \theta_{1j}^{m_{1j}} \cdot \theta_{2j}^{m_{2j}} \cdot \ldots \cdot \theta_{kj}^{m_{kj}};\ \forall j \quad (6)$$

Where $0 \leq \theta_{ij} \leq 1$ for all i in 1 to *k*, and $\theta_{1j} + \cdots + \theta_{kj} = 1\ (\forall j)$, and if *k= 1* the mass function reduces to the binomial, $\forall j = 1, .., n$.

The conjugate prior of the Multinomial distribution is the Dirichlet distribution, the multivariate generalization of beta distribution. Hence the parameter vector $\theta_k = \left(\theta_{1j}, \theta_{2j}, \ldots, \theta_{kj}\right);\ \forall j$ has a prior distribution given by:

$$\theta_k \sim Dirichlet\left(\alpha_{1j}, \alpha_{2j}, \ldots, \alpha_{kj}\right); \forall j = 1, .., n \quad (7)$$

In (10) the density function is given by:

$$g\left(\theta | \alpha_{1j}, \alpha_{2j}, \ldots, \alpha_{kj}\right) = \frac{\Gamma\left(\sum_i^k \alpha_{ij}\right)}{\prod_i^k (\Gamma\, \alpha_{ij})} \theta_{1j}^{(\alpha_{1j}-1)} \theta_{2j}^{(\alpha_{2j}-1)} \ldots \theta_{kj}^{(\alpha_{kj}-1)};$$

$$\alpha_{ij} > 0; 0 \leq \theta_{ij} \leq 1;\ \sum_i^k \theta_{ij} = 1;\ \forall j = 1,..,n \quad (8)$$

In Bayesian inference, $p(\theta|x)$ is known as posterior distribution and is proportional to likelihood ($p(x|\ \theta)$) *x* prior distribution ($p(x)$), so $p(\theta|x) \propto p(x|\theta) \cdot p(x)$.

The posterior distribution of θ_j given *X* is:

$$\theta_j|x \sim Dirichlet\left(x_{1j} + \alpha_{1j}, x_{2j} + \alpha_{2j}, \ldots, x_{kj} + \alpha_{kj}\right);\ \forall j = 1,..,n \quad (9)$$

Thus, in order to implement a new method that calculates the depth of the sample and conveniently estimates the effort of convenient sampling, as well as whether it is necessary to sequence more samples or not, it can be done by simulating matrices ***M'*** with different values of *k* and *n*, ***M'*** it has to have a multinomial probability distribution. We can simulate directly ***M'*** from the joint posterior Dirichlet distribution, using the function rdirichlet() from the LearnBayes package in R (CRAN, 2018) and the function rmultinom() with probability priori Dirichlet (Monleon-Getino, Rodríguez-Casado & Verde, 2019)

2.2. Calculating rarefaction curves

There are many methods for calculating the rarefaction curve for each ***M'***; here we have chosen to use one of the last ones that are the iNEXT() function of the library of R iNEXT () for Interpolation and Extrapolation for Species Diversity (Hsieh, Ma & Chao, 2016). This library provides simple functions to compute and plot two types (sample-size- and coverage-based) rarefaction and extrapolation of species diversity (based on Hill numbers) for individual-based (abundance) data or sampling-unit based (incidence) data.

Using the function iNEXT(), we calculated the rarefaction curves for each metatranscriptomic/ metagenomic matrix (***M'***) simulated previously.

2.3. Calculating sampling effort

Unfortunately, iNEXT() cannot calculate the maximum number of genes/OTUs and neither estimate the sampling effort, and the reads to reach the 90, 95 and 99% of the maximum number of genes/OTUs in the case of non-saturative rarefaction curves. To address this caveat, we propose a saturative non-linear parametric model.

In this type of studies is common do a previous analysis of the selection of models that fit rarefaction curves, based on previous experience and test a selection of possible non-linear models (Mendez, Monleon-Getino, Jofre & Lucena, 2017) or using Bayesian methods (Monleon-Getino, Rodriguez-Casado, Mendez-Viera, 2017).

Several functions including Weibull, logistic, asymptotic regression through the origin (or 2 parameters Weibull growth model), Gompertz and Michaelis-Menten models were tested here using non-linear regression in order to be used as extrapolations of the rarefaction curves (Mendez, Monleon-Getino, Jofre & Lucena, 2017). The regression analysis was performed using the R-package function nls(), and the model accuracy was tested with the function accuracy() of the R-package rcompanion (R Companion, 2018), that produces a table of fit statistics for multiple models. The model accuracy was tested using Efron's pseudo r-squared, Min.max.accuracy (for minimum, maximum accuracy, more substantial indicates a better fit, and a perfect fit is equal to 1) and root mean square error (RMSE) which has the same units as the predicted values. The Weibull sigmoid model obtained best scores and was selected as a good function that fits and extrapolates rarefaction curve.

The Weibull growth model used in our studies is derived from the one-parameter Weibull function (10), given by:

$$F(x) = 1 - e^{(-x^{\gamma})} \tag{10}$$

Where γ is a shape parameter and $x>0$ and $\gamma>0$. The distribution function has a point of inflection at $(x, F(x)) = \left(\frac{[(\gamma-1)/\gamma]^{1}}{\gamma}, 1 - exp\left(-\left(1 - \gamma^{-1}\right)\right)\right)$. Then the following equation can be used to obtain the sigmoidal curve for empirical use:

$$F(x) = \beta + (\alpha - \beta)F(kx, \theta) \quad (11)$$

Moreover, for the Weibull function of four parameters can be described by function. $F(x) = \alpha - (\alpha - \beta)e^{-(kx)^{\gamma}}$. So, in our case the Weibull growth model of four parameters (Pinheiro, 2018) is described by the function$W(x)$:

$$W(x) = a - be^{-(cx)^{m}} \quad (12)$$

Where $W(x)$ is the potential number of genes/OTUs being expressed at each number of reads (x) and now $a = \alpha$, $b = \alpha - \beta$, $c = \kappa^{\gamma}$ and $m = \gamma$. a, b, c and m are parameters to be estimated and e is the base of the natural logarithms. a is the asymptote of limiting value of the response variable $W(x)$, $\lim_{x\to\infty} W(x) = a$, that represents the maximum number of expressed genes/OTUs. b is the biological constant (lower asymptote), c is the parameter governing the rate at which the response variable approaches its potential maximum a and finally, m is the allometric constant. The four-parameter Weibull growth model is considered a very flexible model in that it can be easily transformed into a 3-, 2- or 1-parameter Weibull growth model to adapt the relation between possible numbers of genes/OTUs being expressed at each sample size (reads). For example, by setting b=a and m=1 from (12), we obtain a 2-parameter Weibull growth model (or Asymptotic regression through the origin model) given by:

$$W(x) = a(1 - e^{(-cx)}) \quad (13)$$

with the same meaning $W(x)$, x, a and c (see 12).

2.4 Estimation of the amount of sequencing (reads) needed to cover the total expected microbial metatranscriptome/metagenome (confidence band)

The maximum potential number of genes/OTUs being expressed and its 95% confidence band were used as an estimation of the asymptote of limiting value in a Weibull growth model of four (12) or two parameters (13). Using this Weibull parametric model we estimated the amount of sequencing needed to cover 90, 95 and 99% of the total expected metagenome/metatranscriptome in the samples and its 95%

confidence interval, based only on the first 1 million sequences for each sample. We used R (v. 3.6) to perform all the calculations described below.

Parameters in the Weibull growth model were estimated using the nls (Non-linear regression), nls2 (Non-linear regression with brute force (CRAN, 2018b) and minpack.lm (R Interface to the Levenberg-Marquardt non-linear Least-Squares) packages. The option $\sim Ssweibull(x; a, b, c, m)$ was used for the four-parameter Weibull growth model, and ~ *SsasympOrig(x; a, b)* was used for the two-parameter Weibull model. In order to initialize the parameters a "brute-force" algorithm has been used, and then the parameters have been optimized until those that maximize the adjustment value have been optimized; the "brute-force" algorithm returns the nls object corresponding to the starting values (CRAN, 2018b).

2.5. A priori genes/OTUs prediction using a few amounts of initial total reads

We used different algorithms to fit a regression model to predict the potential number of genes/OTUs, effort/reads to reach 90, 95, 99% of the maximum number of genes/OTUs by using only the first 10-20% of sequences (reads). A first strategy was used a classical linear regression of the function lm() and optimized using function step() to perform the stepwise model selection, and model validation was performed by using the function cv.lm(data, model, m) from the DAAG library (Maindonald & Braun, 2010; Maindonald & Braun, 2019) This function gives internal and cross-validation measures of predictive accuracy for multiple linear regression.

Two other strategies have been using so-called machine learning algorithms such as support vector machines (SVM) and Extreme Gradient Boosting (XGBoost), where we use the training data (with multiple features) x_i (here the genes/OTUs in each deep sequencing) to predict a target variable y_i (maximum number of genes/OTUs).

Support Vector Machines (SVM) is a data classification method that separates data using hyperplanes, which is useful in the case of regression (Cortes & Vapnik, 1995). The concept of SVM is very intuitive and easily understandable. If we have labelled data, SVM can be used to generate multiple separating

hyperplanes such that the data space is divided into segments and each segment contains only one kind of data. SVM technique is generally useful for data which has non-regularity which means, data whose distribution is unknown. We used the function SVM() in R to do the calculation (Chang & Lin, 2017).

Extreme Gradient Boosting, which is an efficient implementation of the gradient boosting framework from Chen and Guestrin (2016). Gradient boosting is a state-of-the-art prediction technique that sequentially produces a model in the form of linear combinations of simple predictors—typically decision trees—by solving an infinite-dimensional convex optimization problem. XBoost() from library Xboost() in R (Chen & Guestrin, 2016), permits the calculation of this predicted method.

In order to check the accuracy of the different models is common use the coefficient of determination (R^2 or R-squared), the mean absolute error (MAE) and the root-mean-square error (RMSE) (Hyndman & Koehler, 2006).

R^2 it is the percentage of the response variable variation that is explained by the model:

$$R^2 = \text{Explained variation / Total variation} \quad (14)$$

R^2 is always between 0 and 1, 0 indicates that the model explains none of the variability of the response data around its mean. 1 indicates that the model explains all the variability of the response data around its mean.

RMSE a frequently used measure of the differences between values (sample and population values) predicted by a model or an estimator and the values observed. The RMSE represents the sample standard deviation of the differences between predicted and observed values.

$$RMSE = \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n} \quad (15)$$

Where n is the number of pairs of observations, $\hat{y}_i$ the value predicted and y_i the observed value.

Mean Absolute Error (*MAE*) is the average vertical distance between each point and the $Y=X$ line:

$$MAE = \frac{\sum_{i=1}^{n} |(\hat{y}_i - y_i)|}{n} \quad (16)$$

Where n is the number of pairs of observations, $\hat{y}_i$ the value predicted and y_i the observed value.

2.6. Metatranscriptome databases used in actual application of the method

We have used metatranscriptome datasets from three different sources as actual application of the proposed method. The first set was generated in our lab as described in Yost, Duran-Pinedo, Teles, Krishnan & Frias-Lopez. (2015) and is available at the Human Oral Microbiome Database (HOMD) server under the submission number 20141024 (ftp://ftp.homd.org/publication_data/20141024/RNA/). The second dataset was generated by Benítez-Páez, Belda-Ferre , Simón-Soro & Mira. (2014) and is available at the MG-RAST server by accessing to the "Oral Metatranscriptome" project, id 935 (http://metagenomics.anl.gov/linkin.cgi?project=935). The third dataset was generated by Jorth, Turner, Gumus, Nizam, Buduneli & Whiteley. (2014) and is available at DNAnexus study number SRP033605 (http://sra.dnanexus.com/studies/SRP033605). All databases were cleaned up of rRNA sequences bioinformatically and in the case of SRP033605 we also removed low-quality sequences from the query files. Fast clipper and fastq quality filter from the Fastx-toolkit (http://hannonlab.cshl.edu/fastx toolkit/) were used to remove sequences shorter than 50bp with quality score >20 in >80% of the sequence.

3. Results and Discussion

3.1. Metatranscriptomic/ Metagenomic matrix simulation, rarefaction computation and estimation of parameters

Our focus is to study the transcriptome of whole complex microbial communities rather than individual transcriptomes, using the oral community as a model. The oral microbiome is one of the best characterized human body sites (Paster, Boches, Galvin, Ericson, Lau, Levanos et al, 2001; Marsh, 2006; Socransky, Haffajee, Cugini, Smith & Kent, 1998; Haffajee, Socransky, Patel & Song, 2008; Peterson, Snesrud, Liu, Ong, Kilian, Schork et al, 2013; Belda-Ferre, Alcaraz, Cabrera-Rubio, Romero, Simón-Soro, Pignatelli et al., 2012), comprising an extremely complex and highly organized biofilm community (Kolenbrander,

2000; Kolenbrander, Andersen, Blehert, Egland, Foster and Palmer, 2002). More than 600 bacterial species have been identified in the oral cavity [Paster, Boches, Galvin, Ericson, Lau, Levanos et al, 2001; Dewhirst, Chen, Izard, Paster, Tanner, Yu W-H et al., 2010]. Many oral bacterial species have not yet been cultivated, and the only information we possess about them derives from their 16S rRNA phylogenetic affiliation.

In the current study, we investigated the mathematical Weibull model proposed, using a nonlinear regression modeling. This model is a generalization of the asymptotic growth model in that it reduces when the parameter m is unity (see Methods section).

Using an R script (see Supplementary Material) we simulated 1587 metatranscriptomic/metagenomic matrices containing more than 9^9 reads, with random sizes of the number of genes/OTUs (min = 267, max=339319) and reads (min = 550, max=6823774), and always 3 samples (replicas). The simulations had a high computational cost of more than 2 weeks and were carried out on a Linux Xeon SP 4114 2.2 GHz computer server with 40 cores. This information has been collected in a data frame for further analysis.

Table 2: Estimations of parameters of interest using a set of 1587 simulations by means of a multinomial model of a metatranscriptomic/ metagenomic matrices

Estimations of parameters of interest	Mean	Minimum	Maximum
Maximum number of gene/otu observed	183312	6	926470
Effort computed using iNEXT()+Weibull model	72.399%	1.208%	100%
Reads to reach 99% maximum number gene/OTUs	5788494	80	31779350

A rarefaction curve using the 1587 cases simulated was computed using the function iNEXT() and the vector obtained (n=100 points, x=reads, y=genes/OTUs) was saved and used later to compute a) maximum of the number of genes/OTUs b) sampling effort to reach maximum number of gene/OTUs (minimum=1%, maximum =100%; see Table 2) and furthermore the c) reads to reach 90, 95 and 99% maximum number of genes/OTUs. This last part (points a), b), c) was done using an estimation based on a Weibull model commented in section 2.3 using non-linear regression.

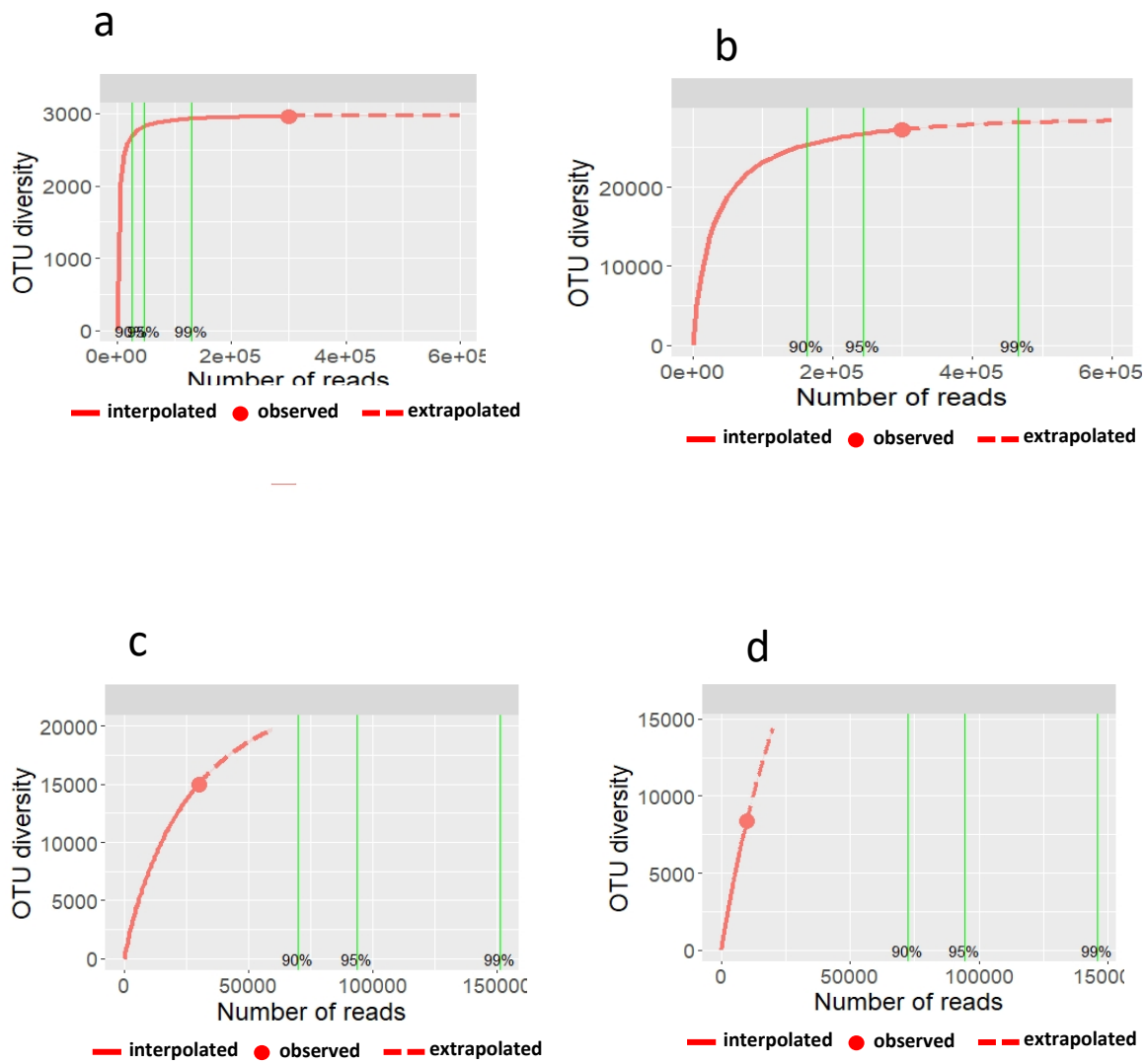


Figure 1: Calculation of the number of gene/OTUs versus number of read using function PILI3() of the library library(BDSbiost3).

Four examples of the results obtained are shown in Figure 1. The results obtained can distinguished in four different types of rarefaction curves a) over-sampling curves: minimum sampling effort to obtain the maximum amount of genes/OTUs in a quick rarefaction curve; b) correct sampling curves: medium sampling effort to obtain the maximum amount of genes/OTUs in a saturative rarefaction curve; c) under-sampling curves: maximum sampling effort to obtain the maximum amount of genes/OTUs in a non-observed saturative rarefaction curve; d) very under-sampling: very maximum sampling effort to obtain the maximum amount of genes/OTUs in a non-observed saturative rarefaction curve. Moreover, in the curves of Figure 1 we can distinguish the vertical lines of the reads to reach the 90, 95 and 99% of the maximum number of gene/OTUs

3.2. A priori genes/OTUs prediction using a few amounts of total reads

Using the data simulated and the parameters estimated previously we fit a regression to predict the potential number of genes/OTUs and the reads to reach 90, 95, 99% of the maximum number of gene/OTUs using only the first 20% of sequences (reads). To carry out this method we have been used three algorithms (linear model (lm), Extreme Gradient Boosting (XB) and support vector machine (SVM)) to predict values commented in the section. Several predictors were tested to predict the maximum number of genes/OTUs as a function of the first 20% of sequences (reads), using the simulated data; for which several good predictors were detected, such as the asymptote using a 4-parameter Weibull model or other similar and well-known models such as the logistics curve model (Mendez, Monleon-Getino, Jofre & Lucena , 2017). Other predictors used were the minimum-maximum number of genes/OTUs observed and finally the minimum-maximum number of reads observed (see Table 3, central column; model 1 and model 2 and supplementary material).

After testing the prediction of the models proposed using the three prediction algorithms indicated above (lm, Xboost and SVM) it was found that the results of the prediction of interest (maximum number of genes/ OTUS and reads to reach 90, 95, 99% of the maximum number of gene/OTUs) for the total curve

with the 1587 simulated samples was very similar, with a $R^2 > 0.99$, which indicates a possible over-fitting (see table 3, right column).

To validate the method and the models, first we used only the first 20 points of the rarefaction curve (reads of the 20% of the total amount of the curve obtained) and secondly, we divided the total number of rarefaction curves simulated (n=1587) and the estimated parameters (maximum number of genes/OTUs, sampling effort, etc.) into two parts using cross-validation. 1) training set: the 70% was used to train and estimate the prediction models (lm, XB and SVM) and 2) test set: the 30% was used to check the fitting of the model and its capacity to predict the maximum number of genes/OTUs, reads to reach 90, 95, 99% of the maximum number of gene/OTUs using only the first 20% of sequences (reads).

Table 3: Model accuracy for maximum number of genes/ OTUS prediction using only a 20% of total reads in a simulation of 1587 metatranscriptomic/metagenomic genomic sequences.

Model name	**Predictors used in the model (independent variables, X_i)**	**Results (R^2) with different algorithms of prediction**
Model 1	• Asymptote estimated using a logistic function • Asymptote estimated using a Weibull 4 parameters function • "Observed" minimum number of reads of the 20% vector • "Observed" maximum number of reads of the 20% vector	SVM = 0.9964754 LM = 0.9990069 Xboost = 0.999999
Model 2	• Asymptote estimated using a Weibull 4 parameters function • "Observed" minimum number of reads of the 20% vector • "Observed" maximum number of reads of the 20% vector	SVM = 0.9964423 LM = 0.9981882 Xboost 0.9999981

We used 300 random re-samplings, and a significant computational effort was made to obtain the predictions using models 1 and 2. We determined that the XB and lm are useful methods to predict the maximum number of genes/OTUs using only the 20% of depth sequencing. To prove the accuracy of the

method we used the mean absolute error (MAE), root square mean error (RSME) and the coefficient of determination (R^2) between estimation using Weibull model with 100% of the rarefaction curve and the 20% of it.

The results of the validations of the three prediction methods (XB, lm, and SVM) and model 1 are presented in the Figures 2, 3 (prediction of maximum number of genes/OTUs) and 4, 5 (prediction of reads to reach 95 maximum number of OTU) where is shown absolute error (MAE), RMSE bands (mean and 95% and 99% confidence) and R^2 for the 300 random resampling test sets. Is possible appreciate that lm and XB in all situations (estimation of the maximum number of genes/OTUs; number of reads to reach 95% of the maximum number of gene/OTUs) are the best methods.

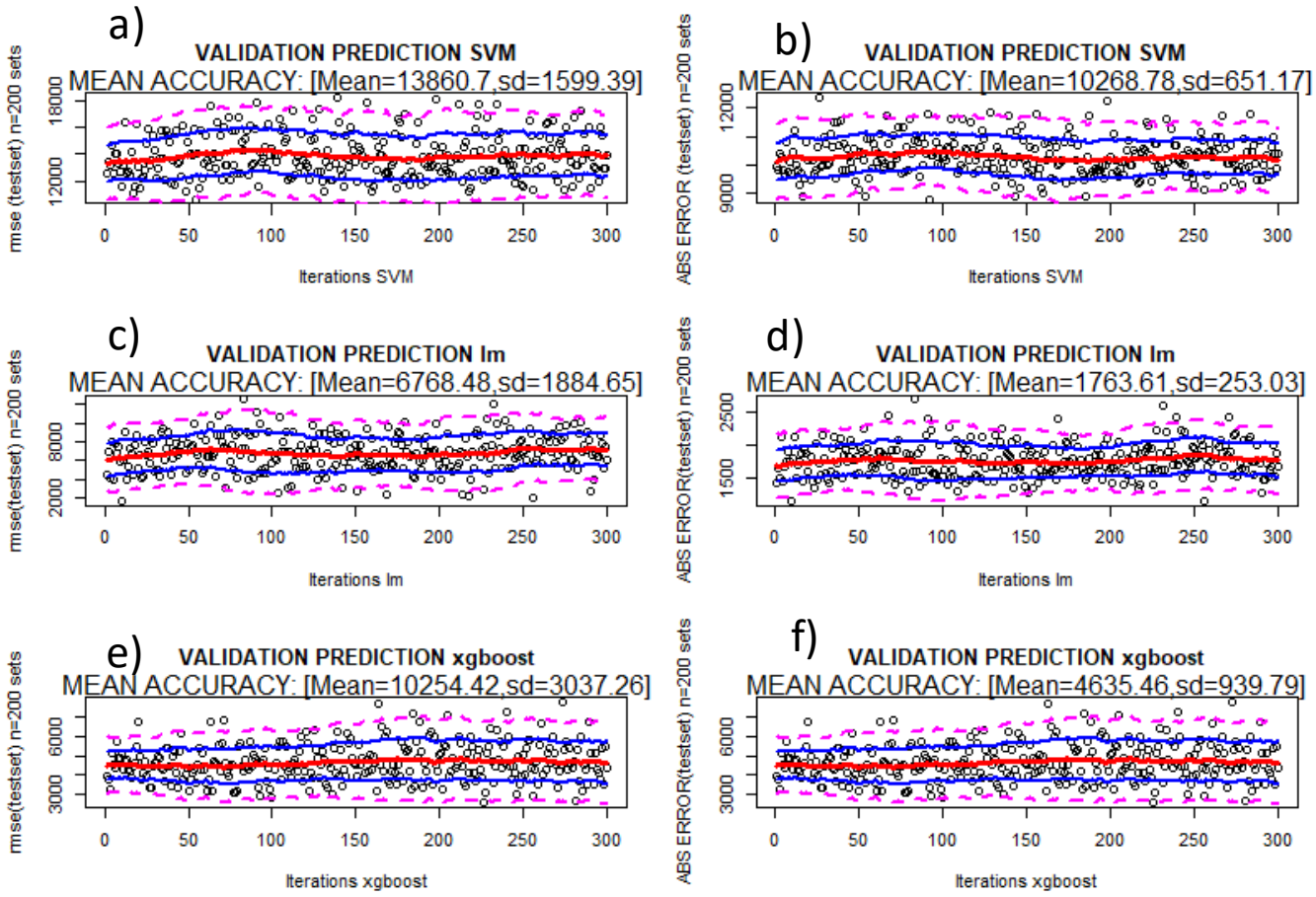


Figure 2: RMSE and absolute error bands (mean (red), 95% (blue) and 99% confidence (magenta)) of different methods [a) Support vector Machine, (b) linear regression model, and c) XBoost] using 20% of

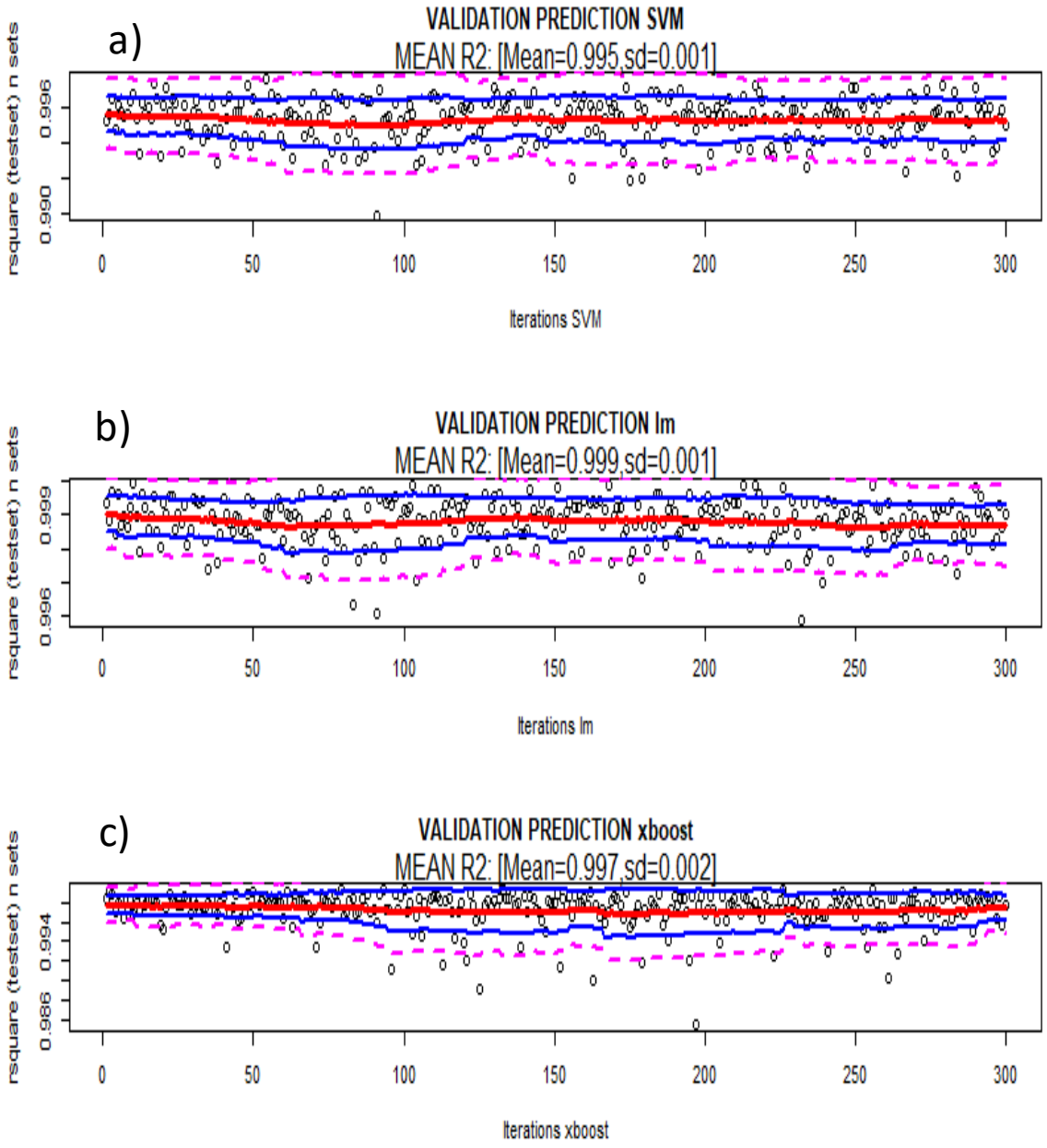
371 depth sequencing (reads) to predict maximum number of otus/genes. 300 random resamples were
372 performed.

373

374 This final lm model (model 1) for predict max number of genes/OTUs has a RMSE = 6769, MAE= 1763
375 and R^2 = 0.999 between observed and predicted values (Figure 2(b,c) and Figure 3(b). This final lm model
376 (model 1) for predict reads to reach 95% of the maximum number of genes/OTUs has a RMSE = 41283,
377 MAE= 16124 and R^2 = 1 between observed and predicted values (Figure 4(b,c) and Figure 5(b).
378 The final XB model estimated for predict max number of genes/OTUs has a RMSE = 10254, MAE= 4635
379 and R^2 = 0.997 between observed and predicted values (Figure 2(e,f) and Figure 3(c). This final XB model
380 estimated for predict reads to reach 95% of the maximum number of genes/OTUs has a RMSE = 102112,
381 MAE=49946 and R^2 = 0.999 between observed and predicted values (Figure 4(e,f) and Figure 5(c).

382

383



384

385 **Figure 3: Coefficient of determination** (R^2) bands (mean (red), 95% (blue) and 99% confidence
386 (magenta)) of the different methods used [a) Support vector Machine, (b) linear regression model, and c)
387 XBoost] using 20% of depth sequencing (reads) to predict maximum number of genes/OTUs. 300 random
388 resamples were performed.

389

390

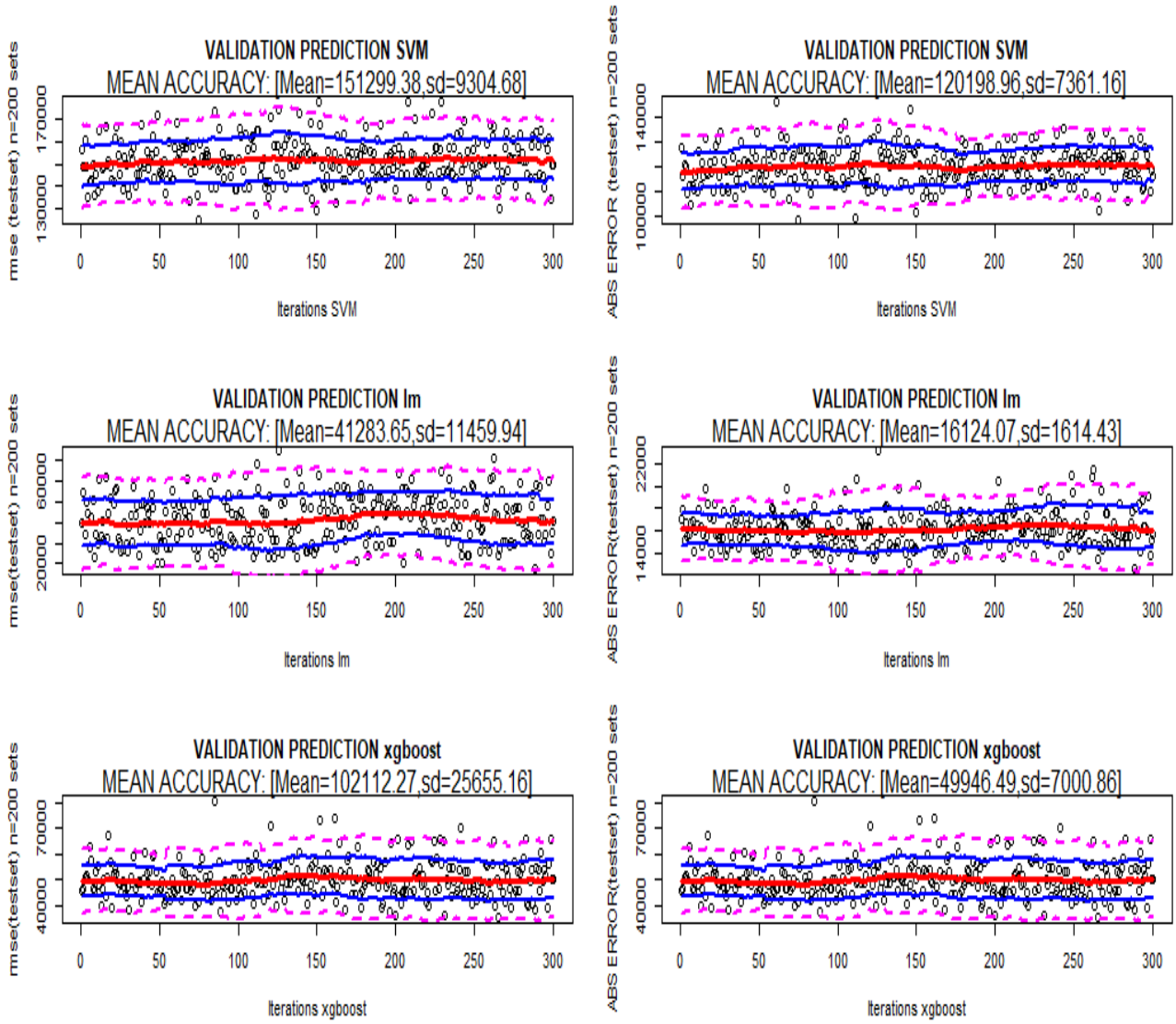


Figure 4: RMSE and absolute error bands (mean (red), 95% (blue) and 99% confidence (magenta)) of different methods [a) Support vector Machine, (b) linear regression model, and c) XBoost] using 20% of depth sequencing (reads) to predict the reads to reach 95% of maximum number of genes/OTUs. 300 random resamples were performed.

Finally, an XB model 1 including the total amount of simulated data (n= 1587) was estimated and saved. The R^2 of all data and prediction models (lm, XB and SVM) are presented in Figure 6 . This model will be used to predict the described parameters of interest (max num of genes/OTUs: Figure 6 a,b,c, reads to reach 95% of the maximum number of genes/OTUs: Figure 6 d,e,f , effort, etc.). Also, the confidence interval (95%) can be computed. To obtain the 95% confidence of the prediction we have used a "bagging" method

thanks to the use of XB model, which basically means creating the same model many times (that has randomness in it.

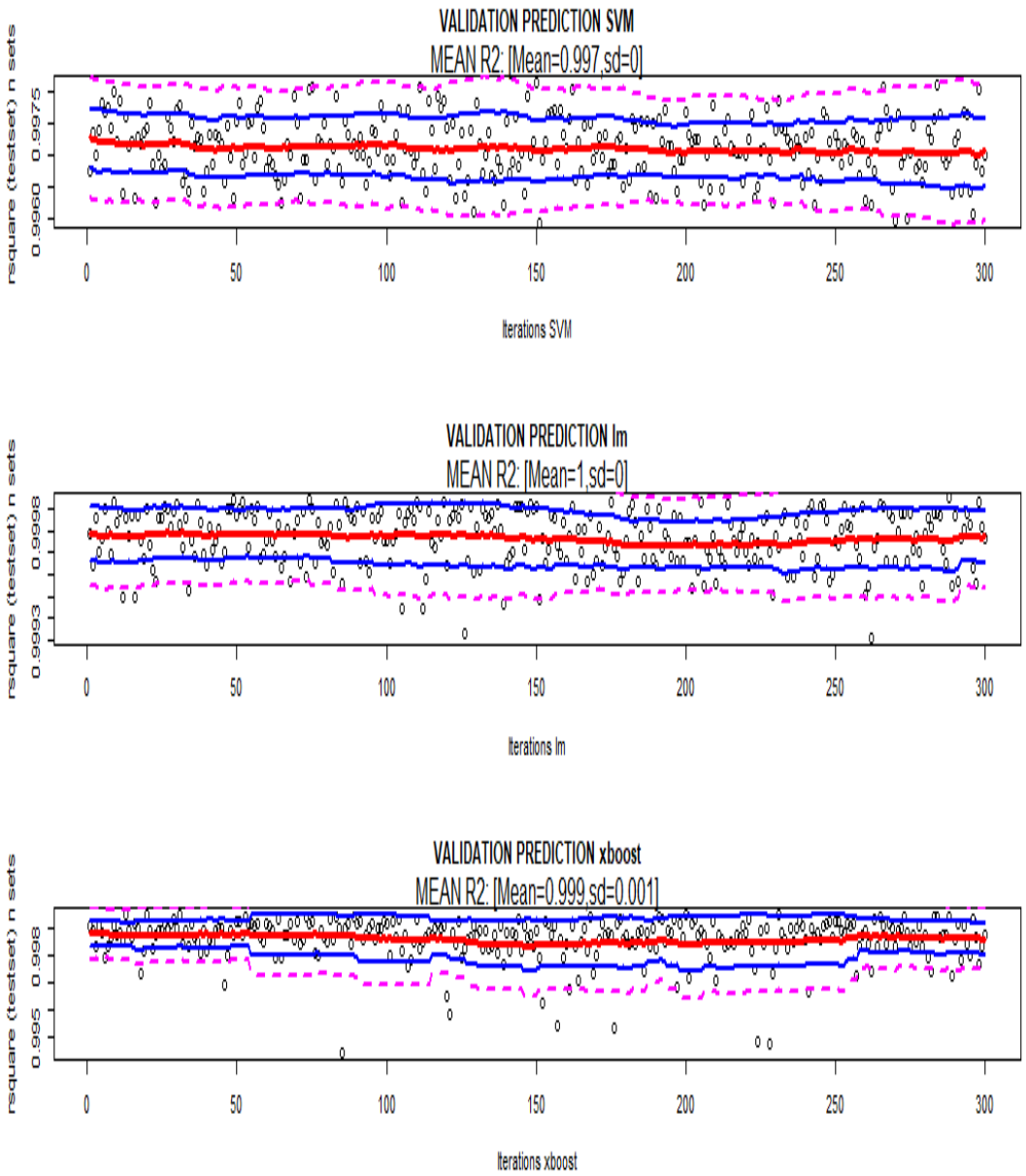


Figure 5: Coefficient of determination (R^2) bands (mean (red), 95% (blue) and 99% confidence (magenta)) of the different methods used [a) Support vector Machine, (b) linear regression model and c) XBoost] using 20% of depth sequencing (reads) to predict the reads to reach 95% of maximum number of genes/OTUs. 300 random resamples were performed

Finally using 100 subsamples we can obtain the prediction mean and the 95% by means of the function ci.mean() of the library(Publish) for R. This final XB model estimated for predict max number of genes/OTUs has a MAE= 107 and R^2 = 0.9964754 between observed and predicted values (Figure 6c). This final XB model estimated for predict reads to reach 95% of the maximum number of genes/OTUs has a MAE= 1380 and R^2 = 0.9999996 between observed and predicted values (Figure 6g).

3.3. Application of the method proposed to real data

An example of external validation (real data not used before) was used to check the algorithms developed previously. To this end, we used a set of 15 datasets of metatranscriptomes from the oral cavity. These RNA sequences consist of vectors of 105-1.5x107 reads depth with a 10000 and 600000 genes size, most of them with saturation but in some cases with a definite no saturation examples. We used these sequences to validate the method and predict the maximum number of genes and the number of reads to reach 95% of the maximum number of genes using all number of reads or only a percentage of it (3%, 20%, and 60% of reads depth). The function monle.predict.max() was developed in order to compute this type of incomplete transcriptomic vectors (X=sequensing depth, Y=genes).

The results of this validation are shown in the supplementary material and reflect that the model used, based on a Weibull model of four parameters, fits perfectly and is possible to estimate correctly the parameters of interest (maximum number of genes, reads depth to reach 95% of maximum genes).

When only a percentage (3%, 20% and 60% of reads depth) of the transcriptomic vector was used the results are equally quite acceptable to predict the maximum amount of genes/OTU and medium acceptable to predict reads to reach 95% of the maximum number of gene/OTU. The prediction for the maximum number of genes was considered acceptable when the maximum number of genes is inside the XB bagging 95% prediction interval. In the same way, the prediction to reads depth to reach 95% of the maximum number of genes prediction was considered acceptable when it is inside the XB bagging 95%prediction interval or between 90-99% interval calculate using the 100% reads depth of the transcriptomic.

When a 3% (10^5-$5x10^5$ reads) was used to predict the parameters of interest, 12/15 (80%) curves to predict the number of genes and 6/15 (33%) curves to predict reads to reach 95% maximum genes were acceptable. When a 20% (105-3x106 reads) was used to predict the parameters of interest, 14/15 (93%) curves to predict the number of genes and 9/15 (60%) curves to predict reads to reach 95% maximum genes were acceptable. When a 60% (105-1x107 reads) were used to predict the parameters of interest, 14/15 (90%) curves to predict the number of genes and 9/15 (60%) curves to predict reads to reach 95% maximum genes were acceptable.

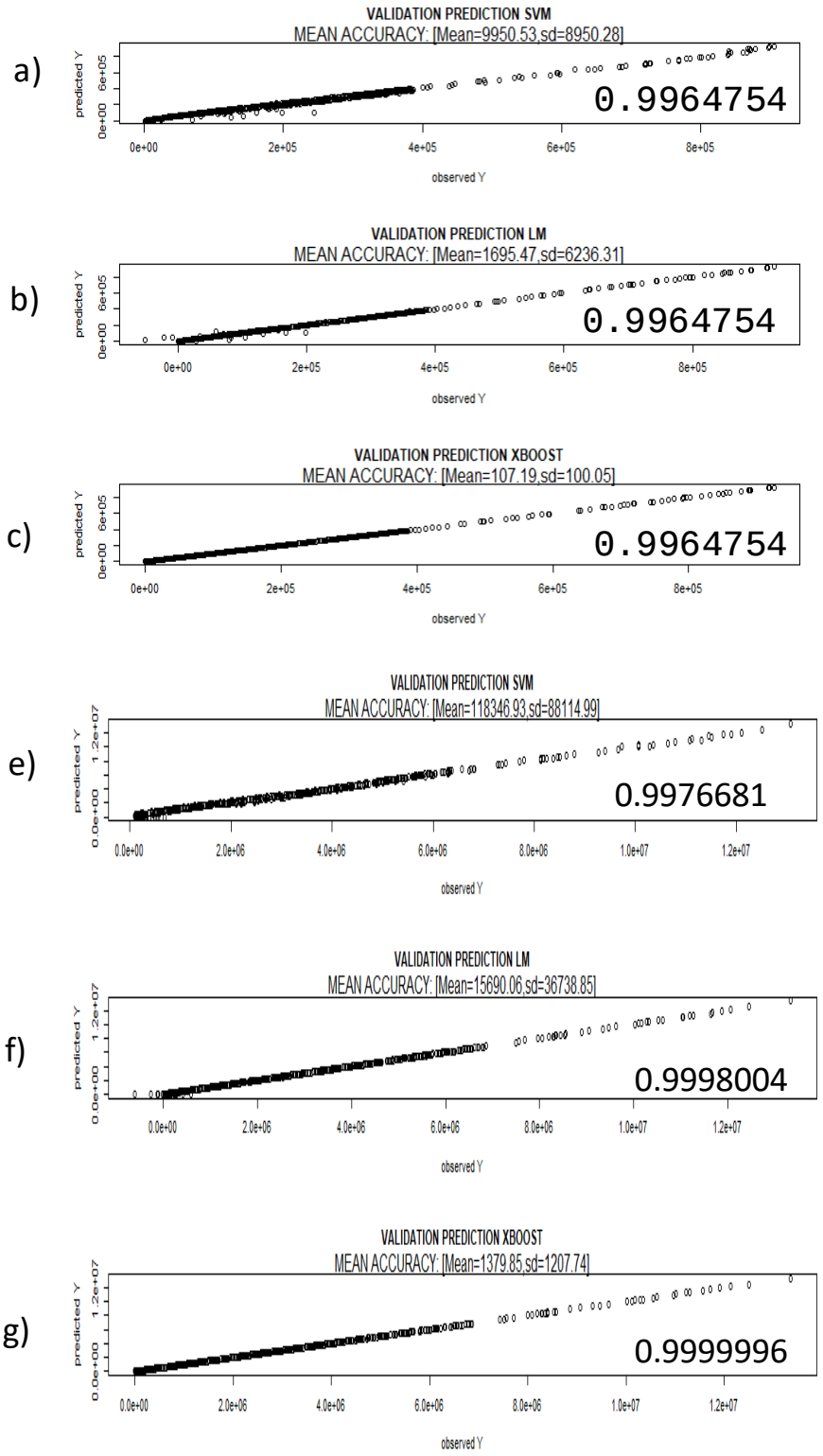


Figure 6: Prediction of maximum number of gene/OTU (a, b, c) and reads to reach 95% maximum gene/OTUs (d, e, f) using a SVM, lm and XGBoost model and only the 20% of the reads versus observed value. All the samples were used (n=1556).

3.3. Final conclusions

This proposed method of estimation of the maximum number of gene/OTUs, reads to reach 90, 95 and 99% of maximum number of gene/OTUs, using and algorithm based on rarefaction curve + Weibull model + machine learning prediction, is efficient to help researchers to know if the sampling is sufficient or otherwise need to be increased. It needs to be used with precaution to predict the sequencing depth, especially with the non-saturation observed samples; sometimes the proposed model can cause predictive problems, but in most cases, it works. More efforts can be used with real sequences and typologies to validate completely this model and methodology based on simulation.

Estimating the sequencing depth required to adequately sample the metatranscriptome/ metagenome of interest using RNA-seq and Shotgun is an essential first step to both obtain robust results in further analysis and avoiding over-expending once the information contained in the library reaches saturation. Our method allows one to use an initial shallowly sequenced sample (in this case 20% of the total amount of reads sampled) to estimate the expected sequencing effort needed to cover the whole metatranscriptome/ metagenome from the same sample, so can be used to estimate the estimate the sample size. This limited initial number of sequences is low enough that with the current NGS methods allows for the estimate of considerable number of samples at a low cost.

AUTHOR S' CONTRIBUTIONS

T.M.G. developed the procedure. R. wrote the code.

T.M.G. and J.F.L. conducted the analyses and wrote the manuscript.

DATA ACCESSIBILITY

All code is open source and available in Github. All of these functionalities showed in the Figure 1 (rarefaction curve, Weibull non-linear model, effort estimation, extrapolation of the maximum number of

gene/OTUs, reads to reach 90, 95, 99% for the maximum number of gene/OTUs) used has been compiled in a new two functions in R: PILI3() and monle.predict.max() and added to the library BDSbiost3 and can be found at the repository https://github.com/amonleong/BDSbiost3

ACKNOWLEDGMENT

Our thanks to Andreu Paytuvy and Walter Sanseverino from Sequencia-Biotech for his advice with metagenomics. Also our thanks to Javier Mendez from Departament of Microbiology of the University of Barcelona for his advice with the methods.

REFERENCES

Belda-Ferre P, Alcaraz LD, Cabrera-Rubio R, Romero H, Simón-Soro A, Pignatelli M, et al. The oral metagenome in health and disease. ISME J. 2012;6:46–56.

Benítez-Páez A, Belda-Ferre P, Simón-Soro A, Mira A. Microbiota diversity and gene expression dynamics in human oral biofilms. BMC Genomics. 2014;15:311.

Chang C-C, Lin C-J. LIBSVM: A Library for Support Vector Machines. ACM Trans Intell Syst Technol. 2011;2:27:1–27:27.

Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. ArXiv160302754 Cs. 2016;785–94.

Cortes C, Vapnik V. Support-Vector Networks. Mach Learn. 1995;20:273–97.

CRAN. 2018b. nls2: Non-linear regression with brute force version 0.2 from CRAN [Internet]. [cited 2018 Aug 14]. Available from: https://rdrr.io/cran/nls2/

CRAN. LearnBayes: Functions for Learning Bayesian Inference version 2.15.1 from [Internet]. [cited 2018 Aug 14]. Available from: https://rdrr.io/cran/LearnBayes/

497 Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner ACR, Yu W-H, et al. The human oral microbiome. J
498 Bacteriol. 2010;192:5002–17.

499 García-Ortega LF, Martínez O. How Many Genes Are Expressed in a Transcriptome? Estimation and
500 Results for RNA-Seq. PLOS ONE. 2015;10:e0130262.

501 Haas BJ, Chin M, Nusbaum C, Birren BW, Livny J. How deep is deep enough for RNA-Seq profiling of
502 bacterial transcriptomes? BMC Genomics. 2012;13:734.

503 Haffajee AD, Socransky SS, Patel MR, Song X. Microbial complexes in supragingival plaque. Oral
504 Microbiol Immunol. 2008;23:196–205.

505 Hsieh TC, Ma KH, Chao A. iNEXT: an R package for rarefaction and extrapolation of species diversity
506 (Hill numbers). Methods Ecol Evol. 2016;7:1451–6.

507 Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. Int J Forecast. 2006;22:679–
508 88.

509 Jorth P, Turner KH, Gumus P, Nizam N, Buduneli N, Whiteley M. Metatranscriptomics of the Human Oral
510 Microbiome during Health and Disease. mBio. 2014;5:e01012-14.

511 José Pinheiro. 2018. Mixed-Effects Models in S and S-PLUS. Springer [Internet]. [cited 2018 Aug 14].
512 Available from: https://www.springer.com/us/book/9780387989570

513 Kolenbrander PE, Andersen RN, Blehert DS, Egland PG, Foster JS, Palmer, R J J. Communication among
514 oral bacteria. Microbiol Mol Biol Rev MMBR. 2002;66:486–505.

515 Kolenbrander PE. Oral microbial communities: biofilms, interactions, and genetic systems. Annu Rev
516 Microbiol. 2000;54:413–37.

517 Maindonald, J.H. and Braun, W.J. 2019. Data Analysis and Graphics Data and Functions (DAAG).
518 Avalaible at: https://cran.r-project.org/web/packages/DAAG/DAAG.pdf

519 Maindonald, J.H. and Braun, W.J.(3rd edn 2010) "Data Analysis and Graphics Using R"

520 Marsh PD. Dental plaque as a biofilm and a microbial community - implications for health and disease.
521 BMC Oral Health. 2006;6 Suppl 1:S14–.

522 Mendez J, Monleon-Getino A, Jofre J, Lucena F. Use of non-linear mixed-effects modelling and regression
523 analysis to predict the number of somatic coliphages by plaque enumeration after 3 hours of incubation. J
524 Water Health. 2017;15:706–17.

525 Monleon-Getino A, Rodriguez-Casado CI, Mendez-Viera J. Sample size in metagenomics, a bayesian
526 approach using BDSbiost3 for R. Proc CEB 2017. Sevilla, Spain; 2017.

527 Monleon-Getino T., Rodríguez-Casado C.I., Verde P.E. 2019. The Shannon entropy ratio: a Bayesian
528 biodiversity index applied to the measure of uncertainty in metagenomic communities (putative
529 enterotypes). Jornal of Advanced statistics ("in press")

530 Ni J, Yan Q, Yu Y. How much metagenomic sequencing is enough to achieve a given goal? Sci Rep
531 [Internet]. 2013 [cited 2014 Jul 16];3. Available from:
532 http://www.nature.com/srep/2013/130611/srep01968/full/srep01968.html

533 Paster BJ, Boches SK, Galvin JL, Ericson RE, Lau CN, Levanos VA, et al. Bacterial diversity in human
534 subgingival plaque. J Bacteriol. 2001;183:3770–83.

535 Peterson SN, Snesrud E, Liu J, Ong AC, Kilian M, Schork NJ, et al. The dental plaque microbiome in
536 health and disease. PloS One. 2013;8:e58487.

537 R Companion: The Handbook for Biological Statistics [Internet]. [cited 2018 Aug 14]. Available from:
538 https://rcompanion.org/rcompanion/a_02.html

539 Robinson DG, Storey JD. subSeq: Determining Appropriate Sequencing Depth Through Efficient Read
540 Subsampling. Bioinformatics. 2014;30:3424–6.

541 Rodríguez-Casado, Monleón-Getino T, Cubedo M, Ríos-Alcolea M. A priori groups based on
542 Bhattacharyya distance and partitioning around medoids algorithm (PAM) with applications to
543 metagenomics. IOSR Journal of Mathematics. 2017; 13(3): 24-32.

544 Socransky SS, Haffajee AD, Cugini MA, Smith C, Kent, R L J. Microbial complexes in subgingival plaque.
545 J Clin Periodontol. 1998;25:134–44.

546 Tamames J, de la Peña S, de Lorenzo V. COVER: a priori estimation of coverage for metagenomic
547 sequencing. Environ Microbiol Rep. 2012;4:335–41.

548 Toung JM, Morley M, Li M, Cheung VG. RNA-sequence analysis of human B-cells. Genome Res.
549 2011;21:991–8.

550 Westermann AJ, Gorski SA, Vogel J. Dual RNA-seq of pathogen and host. Nat Rev Microbiol.
551 2012;10:618–30.

Yost S, Duran-Pinedo AE, Teles R, Krishnan K, Frias-Lopez J. Functional signatures of oral dysbiosis during periodontitis progression revealed by microbial metatranscriptome analysis. Genome Med. 2015;7:27.

552
553