Yao et al. BMC Genetics (2018) 19:98
https://doi.org/10.1186/s12863-018-0680-1

BMC Genetics

METHODOLOGY ARTICLE Open Access

An improved statistical model for taxonomic assignment of metagenomics

Yujing Yao[1], Zhezhen Jin[1] and Joseph H Lee[2,3*]

Abstract

Background: With the advances in the next-generation sequencing technologies, researchers can now rapidly examine the composition of samples from humans and their surroundings. To enhance the accuracy of taxonomy assignments in metagenomic samples, we developed a method that allows multiple mismatch probabilities from different genomes.

Results: We extended the algorithm of taxonomic assignment of metagenomic sequence reads (TAMER) by developing an improved method that can set a different mismatch probability for each genome rather than imposing a single parameter for all genomes, thereby obtaining a greater degree of accuracy. This method, which we call TADIP (**T**axonomic **A**ssignment of metagenomics based on **DI**fferent **P**robabilities), was comprehensively tested in simulated and real datasets. The results support that TADIP improved the performance of TAMER especially in large sample size datasets with high complexity.

Conclusions: TADIP was developed as a statistical model to improve the estimate accuracy of taxonomy assignments. Based on its varying mismatch probability setting and correlated variance matrix setting, its performance was enhanced for high complexity samples when compared with TAMER.

Keywords: EM algorithm, Metagenomics, Taxonomic assignment

Background

As the next generation sequencing technologies continue to advance at a rapid pace, it is now possible to identify the presence of microorganisms with greater efficiency and accuracy. Such studies can, in turn, help to explain whether the presence or absence of certain species or specific genus contributes to disease processes of interest. Biological samples taken from different parts of the human body as well as from different environment, such as seawater, soil, etc. can be used to extract DNA, and then those DNA samples can be analyzed as short reads of ~100 base pairs. To analyze these short reads, Basic Local Alignment Search Tool (BLAST) is often used to identify regions of similarity between nucleotide or protein sequences by comparing sequence reads from one sample to sequences in reference databases. It then assesses the significance of matches, and, using a scoring matrix, assigns reads to the taxonomy tree that most likely to represent what had happened in the evolutionary process [1, 2]. Here, a single sequence read can be matched to multiple genomes because of sequence homology across species as well as overlapping of sequences.

BLAST can potentially lead to inaccurate estimates when errors occur in taxonomy assignment in the context of metagenomic analysis [3–5]. To improve the accuracy of taxonomy assignment, several algorithms have been developed to optimize the use of BLAST searches. Metagenome Analyzer called MEGAN is one of the most commonly used analytical tool [4]. MEGAN assigns matched reads to the least common ancestor in the taxonomy tree when there are multiple matches to different genomes [2], and because it assigns short reads to one genome with the best match and ignores relevant biological information with weak statistical significance, MEGAN can lead to false findings. To address this issue, Jiang and colleagues [2] introduced TAMER, which assigns metagenomic sequence reads with a mixture model by estimating the probability for each read generated

* Correspondence: JHL2@cumc.columbia.edu
[2]Sergievsky Center, Taub Institute, and Departments of Epidemiology and Neurology, Columbia University, New York, NY, USA
[3]Sergievsky Center, Columbia University, 630 West 168th Street, P&S Unit 16, New York, NY 10032, USA
Full list of author information is available at the end of the article

from the genomes. Based on analyses of both simulated and real datasets, Jiang and colleagues [2] showed that TAMER had a higher degree of accuracy and efficiency compared to MEGAN. However, because TAMER assigns equal mismatch probability for all genomes, TAMER will experience difficulties when there exists a high degree of complexity among data and a high degree of correlations among microorganisms as in human microbiome samples.

In the present paper, we propose a statistical framework: a Taxonomic Assignment of metagenomics based on Different Probabilities (TADIP) method with the goal of improving the accuracy of the estimates by setting different mismatch probabilities for different candidate genomes. Unlike TAMER which sets the same mismatch probability for different genomes, TADIP extends TAMER to address the biological reality that: (1) different organisms may have different genetic variants at homologous loci; and (2) different organisms coexist within one microbial community. Specifically, TADIP allows the true mismatch probabilities for the genomes to be generated by each genome's own mismatch part plus systemic errors. We also illustrate the use of a burden/variance component test based on a logistic regression model to test the need for a range of setting with varying mismatch probabilities to reflect the complexity of samples. We evaluated TADIP using both simulated and real datasets.

Methods

Data

This study uses the NCBI-NT data from the NCBI website as a reference dataset, and uses BLASTn as the primary analytical tool for data analysis. Following the BLASTn analysis, for each read and its corresponding candidate genome, we mapped and recorded the read serial number, corresponding genome name, taxonomy identification number, matched length, and alignment length. These variables constitute the input file for TADIP, which is consistent with the input file for TAMER [2].

TADIP model

Model parameters

We first summarize known information from the BLAST output, which includes the following: n reads(x_j denotes the j^{th} read), k genomes (genome i denotes the i^{th} genome), L_{ji} denotes alignment length for read j against genome i, M_{ji} denotes matched length. The parameters of interest are: R_i: the true proportion of reads generated from genome i or the probability of a read x_j is generated by genome i, $R_i \geq 0, \sum_{i=1}^{k} R_i = 1$. p_i: the probability of observing a mismatched base pair for genome i. Even if x_j could be generated from genome i, it is unlikely to match 100% because of potential errors involving sequencing, alignment and among others.

Likelihood

Because alignment lengths for sequence reads are nearly the same, let $L_j = \max\{L_{ji}, i = 1, 2, \ldots, k\}$, then the probability that a read x_j is generated by genome i with M_{ji} matched base pairs, and $L_j - M_{ji}$ mismatched base pairs would be $R_i p_i^{L_j - M_{ji}} (1-p_i)^{M_{ji}}$. Therefore, the probability of observing a read x_j from the sample is:

$$Pr(x_j) = \sum_{i=1}^{k} R_i p_i^{L_j - M_{ji}} (1-p_i)^{M_{ji}} \quad (1)$$

Let $\boldsymbol{\theta} = (\mathbf{p}, \mathbf{R})$, $\mathbf{p} = (p_1, \ldots, p_k)^T$, $\mathbf{R} = (R_1, R_2, \ldots, R_k)^T$, and let $\mathbf{D} = (\mathbf{x}, \mathbf{L}, \mathbf{M})$ where $\mathbf{x} = (x_1, \ldots, x_n)^T$, $\mathbf{L} = (L_1, \ldots, L_n)^T$, $\mathbf{M}$ denotes matched lengths for n reads against k genomes. Assuming that the sequence reads are independent and identically distributed random variables, the likelihood and log likelihood of $\boldsymbol{\theta}$ are:

$$L(\boldsymbol{\theta}|\mathbf{D}) = \prod_{j=1}^{n} \left[\sum_{i=1}^{k} R_i p_i^{L_j - M_{ji}} (1-p_i)^{M_{ji}} \right]$$

$$l(\boldsymbol{\theta}|\mathbf{D}) = \sum_{j=1}^{n} log \left[\sum_{i=1}^{k} R_i p_i^{L_j - M_{ji}} (1-p_i)^{M_{ji}} \right].$$

EM algorithm

The maximization of the log-likelihood is not straightforward. As in [2], an Expectation-Maximization (EM) algorithm is used to obtain the maximum likelihood estimators of $\boldsymbol{\theta}$ by introducing latent variables $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)^T$. The latent variable determines the genome from which a sequence read originates, for example, for each read j and each genome i:

$$f(z_j) = \begin{cases} 1 & \text{if } z_j = i \\ 0 & \text{if } z_j \neq i \end{cases}$$

Then the probability of observing a read x_j from the sample becomes:

$$Pr(x_j, z_j) = \sum_{i=1}^{k} R_i p_i^{L_j - M_{ji}} (1-p_i)^{M_{ji}} I_{z_j = i} \quad (2)$$

The log likelihood when the latent variables are observed is:

$$\begin{aligned} l(\boldsymbol{\theta}|\mathbf{D}, \mathbf{Z}) &= \sum_{j=1}^{n} log \left[\sum_{i=1}^{k} R_i p_i^{L_j - M_{ji}} (1-p_i)^{M_{ji}} I_{z_j = i} \right] \\ &= \sum_{j=1}^{n} \sum_{i=1}^{k} log \left[R_i p_i^{L_j - M_{ji}} (1-p_i)^{M_{ji}} \right] I_{z_j = i} \end{aligned} \quad (3)$$

Base on (3), the following EM algorithm can be used. In the E-step of algorithm, let T_{ji} denote the conditional

distribution of the latent variable Z_j given the current estimates of parameters $\boldsymbol{\theta}^{(t)} = (\mathbf{p}^{(t)}, \mathbf{R}^{(t)})$, then

$$T_{ji}^{(t)} = Pr\left(Z_j = i|\boldsymbol{\theta}^{(t)}, \mathbf{D}\right) = \frac{R_i^{(t)} p_i^{(t)L_j - M_{ji}} \left(1-p_i^{(t)}\right)^{M_{ji}}}{\sum_{l=1}^{k} R_l^{(t)} p_l^{(t)L_j - M_{jl}} \left(1-p_l^{(t)}\right)^{M_{jl}}} \quad (4)$$

The expectation of the log likelihood (3) with respect to $Pr(Z_j = i|\ \boldsymbol{\theta}^{(t)}, \mathbf{D})$ is: $Q = E_{\mathbf{Z}|\boldsymbol{\theta}^{(t)},\mathbf{D}}[l(\boldsymbol{\theta}|\mathbf{D},\mathbf{Z})] = E_{\mathbf{Z}|\boldsymbol{\theta}^{(t)},\mathbf{D}}[\sum_{j=1}^{n}\sum_{i=1}^{k} log(R_i p_i^{L_j - M_{ji}}(1-p_i)^{M_{ji}}) I_{z_j=i}]$

$$= \sum_{j=1}^{n}\sum_{i=1}^{k} T_{ji}^{(t)} \left[log\left(R_i p_i^{L_j - M_{ji}}(1-p_i)^{M_{ji}}\right)\right] \quad (5)$$

In the M-step, the expected log likelihood (5) can be maximized, which yields

$$R_i^{(t+1)} = \frac{1}{n}\sum_{j=1}^{n} T_{ji}^{(t)} \quad (6)$$

$$p_i^{(t+1)} = 1 - \frac{\sum_{j=1}^{n} T_{ji}^{(t)} M_{ji}}{\sum_{j=1}^{n} T_{ji}^{(t)} L_j} \quad (7)$$

Repeat the E-step and M-step until the convergence of the estimated values of parameters to obtain the estimate of $\boldsymbol{\theta}$.

Taxonomic assignment of reads

Given the above information, we can then estimate the probability that a read x_j is generated from genome i by:

$$P_{ji} = \frac{R_i p_i^{L_j - M_{ji}}(1-p_i)^{M_{ji}}}{\sum_{l=1}^{k} R_l p_l^{L_j - M_{jl}}(1-p_l)^{M_{jl}}} \quad (8)$$

for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, n$. Then the read x_j is assigned to the genome for which the maximum estimated value of probability is reached.

Hypotheses testing

The null hypothesis $H_0 : p_1 = p_2 = p_3 = \ldots = p_k$ for all p_i (or $\mathbf{p} = \mathbf{p_0}$ where $\mathbf{p} = (p_1, \ldots, p_k)^T$) can then be tested against the alternative hypothesis H_1: at least one pair of p_i is not equal (or $\mathbf{p} \neq \mathbf{p_0}$). If H_0 is true, then the TAMER method is valid to use. On the other hand, if H_0 is rejected, then the TAMER method is not valid, and the TADIP method is appropriate.

Wald test

Using the estimate $\hat{\mathbf{p}}$ from the TADIP method, the Wald test statistics $W = (\mathbf{p} - \mathbf{p_0})^T \times \Sigma^{-1} \times (\mathbf{p} - \mathbf{p_0})$ can be used, when the variance-covariance matrix Σ of $\mathbf{p}$ is known. W follows a χ^2 distribution with $k - 1$ degree of freedom under H_0.

In practice, however, the variance-covariance matrix (Σ) is often unknown. The dimension of the variance-covariance matrix (Σ) is equal to the number of genomes, and the number of genomes can be large for most metagenomic samples. As a result, it is computationally difficult to obtain the inverse of the estimate even when it is estimated using moment estimation methods for example, and its inverse is very much likely to be singular because of sparsity. Moreover, the Wald test often has poor power under sparse alternatives [6]. To circumvent this, we present two alternative tests that can be implemented in practice.

Logistic regression model

One simple test is to use logistic regression models for the mismatch probabilities. Recall that p_i represents the true mismatch probability of genome i which consists of system errors that include sequencing errors, alignment errors and SNPs. The system errors ought to be the same for all genomes in the datasets processing steps, but SNPs are not [7]. Assuming (1) true mismatch probability for each genome to be the sum of a fixed part which comprise average system errors and (2) the genome dependent part denotes different adjustment (SNPs) for each genome i, the following logistic regression model can be used: $\text{logit}(\mathbf{p}) = \boldsymbol{\alpha} + \mathbf{V} \times \boldsymbol{\beta}$ where the matrix $\mathbf{V}$ takes value 1 in the diagonal and $1/(k-1)$ in the non-diagonal. The model yields an estimate of fixed part $\boldsymbol{\alpha}$ which is exactly the logit ($\mathbf{p_0}$), and an estimate of $\boldsymbol{\beta}$ which denotes the coefficient of random part of mismatch probabilities for different genomes. Then test hypotheses are equivalent to the null hypothesis where H_0: $\boldsymbol{\beta} = 0$ against the alternative hypothesis where H_1:$\boldsymbol{\beta} \neq 0$. We can then apply two collapsing tests that are easier to implement than the Wald test to metagenomic samples.

Burden test

The burden test collapses information for multiple random variants into a single score [8] with the assumption that a large proportion causal variants effects are in the same direction and magnitude (See the variance component test below for violation of the assumptions). The magnitude of effects is adjusted by weight where$\mathbf{w} = (w_1, \cdots, w_n)$ [9], i.e., $\beta_j = \beta_0 \times w_j$, $j = 1, \ldots, n$. Let v_{ij} denote the (i, j)th element of the matrix $\mathbf{V}$, then the logistic regression model can be expressed as $\text{logit}(\mathbf{p}) = \boldsymbol{\alpha} + \mathbf{C} \times \beta_0$, where

$\mathbf{C} = (C_1, \ldots, C_k)$, with $C_j = \sum_{i=1}^{n} w_j \times v_{ij}$. The burden test is to test where H_0: $\beta_0 = 0$ with following test statistic

$$Q_{burden} = \left(\mathbf{C}^T \times (\mathbf{p}-\mathbf{p_0})\right)^2 = \left[\sum_{j=1}^{k}\left(\sum_{i=1}^{n} w_j \times v_{ij}\right) \times \left(\hat{p}_j - p_0\right)\right]^2 \quad (9)$$

which follows a χ^2 distribution with one degree of freedom under H_0. The weight w_j can be generated from the Beta function where, $w_j = Beta(\hat{p}_j, a_1, a_2)$ [10]. Throughout this paper, the empirical value for burden test parameters of the Beta function is set to a_1=96, a_2=100.

Variance component test

When the assumptions of the burden tests are violated, which are not infrequent, we can use the variance component test method to deal with differences in directions and magnitudes to take into account both positive or negative effects [9]. Here, we present a variance component test based on the kernel method.

In this variance component test, it is assumed that β_j follows a distribution with mean 0 and variance $w_j^2\tau$. Therefore, we test $\tau = 0$ for the equality of all β_j, $j = 1, \ldots, n$. The test statistic is:

$$Q_{VCT} = (\mathbf{p}-\mathbf{p_0})^T \times \mathbf{K} \times (\mathbf{p}-\mathbf{p_0}) = \sum_{j=1}^{k}\left(\sum_{i=1}^{n} w_j \times v_{ij}\right)^2 \times \left(\hat{p}_j - p_0\right)^2 \quad (10)$$

where $\mathbf{K} = \mathbf{V^T W V}$. Under the $H_0 : \tau = 0$, the test statistic Q_{VCT} follows a mixture degree of freedom and it can be asymptotically calculated by $\sum_{i=1}^{n} \lambda_i \chi_{1,i}^2$. $\chi_{1,i}^2$ means independent χ_1 variables, and λ_i are eigenvalues of $\mathbf{P_0^{1/2} K P_0^{1/2}}$, where $\mathbf{P_0}$ is the inverse of variance under the null and p-values can be calculated using the Davies Method [10, 11]. Again, the weight w_j is often generated from the Beta function, where $w_j = Beta(\hat{p}_j, a_1, a_2)$. In this paper, the empirical values for the variance component test parameters were set to a_1 at 96 and a_2 at 100.

Simulation study

We performed simulation studies using MetaSim, the first sequencing simulator for metagenomics. MetaSim can generate a collection of reads with genome profile settings that mimic existing sequencing platforms such as Illumina, Roche 454, and Sanger [12]. To generate realistic simulated data, MetaSim introduces sequencing errors, SNPs, indels, inversions, translocations, copy number variants (CNVs), short tandem repeats (STRs); consequently, these features generate different mismatch probabilities [13].

To test our models, we generated datasets with the same as well as different mismatch probabilities. Most existing genomic next-generation sequencing simulation tools can be set to generate data with the same mismatch probability at the nucleotide base level. For example, MetaSim uses a fixed value of error rate for the same nucleotide base in one platform for a single run using a dataset. Empirically, the error rates for Roche 454 ranged from 1.07 to 1.7%, while those for Illumina ranged from 0.0034 to 1% [14]. Since we need to set the probabilities at the genome level to be the same, we assumed that the dataset in one sample generated from one approach ("fixed error model", henceforth) yielded the same mismatch probability at the genome level with a small range of error rates. When the datasets were generated from multiple approaches with different error rates ("varying-error model", henceforth), however, different mismatch probabilities at the genome level were expected.

Evaluation of burden test and variance component test

We compare the performance of the burden test and variance component test by estimating empirical type I error and power using simulated datasets. To evaluate type I errors, we generated 100 datasets using an one fixed error model, where all genomes in the datasets were assumed to have the same mismatch probability. The burden test and variance component tests were used with weights generated from the Beta function where $Beta(\hat{p}_j, 96, 100)$. The empirical type I error rate was estimated using the proportion of p values less than $\alpha = 0.05$. To evaluate power, we generated 100 datasets with 1000 reads, where half of the dataset of one genome being generated from the one error-rate approach and the other half generated from the other error-rate approach (specifically "varying-two-error model", henceforth). Therefore, two genomes in the datasets were assumed to have two different mismatch probabilities. Again, the burden test and variance component tests were used with weights generated from the Beta function where $Beta(\hat{p}_j, 96, 100)$. The empirical power was estimated using the proportion of p values less than $\alpha < 0.05$.

Comparison of TAMER and TADIP methods

To determine whether or not TADIP improves the accuracy of taxonomy assignment, we simulated three benchmark datasets with low (2 genomes, simLC), medium (9 genomes, simMC), high (15 genomes, simHC) complexity. We generated these three benchmark datasets under the

Table 1 Simulation results for evaluation of the tests

Tests	Burden Test	Variance Component Test
α	0.05	0.02
$1-\beta$	0.99	1.00

Type I error and power of 100 simulation datasets with 1000 reads per set

Table 2 Results for simulation study: Long reads. The proportions of reads correctly (TP) and incorrectly (FP) assigned to taxonomy tree at different ranks of two methods with average length of 500 bp

	simLC				simMC				simHC			
	TAMER		TADIP		TAMER		TADIP		TAMER		TADIP	
Rank	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
Species	1.0000	0.0000	1.0000	0.0000	0.9229	0.1731	0.9229	0.1004	0.9565	0.0273	0.9565	0.0165
Genus	1.0000	0.0000	1.0000	0.0000	0.9233	0.1731	0.9233	0.0999	0.9644	0.0194	0.9644	0.0087
Family	1.0000	0.0000	1.0000	0.0000	0.9995	0.0965	0.9995	0.0767	0.9838	0.0000	0.9838	0.0000
Order	1.0000	0.0000	1.0000	0.0000	0.9995	0.0965	0.9995	0.0766	0.9838	0.0000	0.9838	0.0000
Class	1.0000	0.0000	1.0000	0.0000	1.0000	0.0960	1.0000	0.0747	0.9838	0.0000	0.9838	0.0000
Phylum	1.0000	0.0000	1.0000	0.0000	1.0000	0.0960	1.0000	0.0747	0.9838	0.0000	0.9838	0.0000
Kingdom	1.0000	0.0000	1.0000	0.0000	1.0000	0.0960	1.0000	0.0747	0.9838	0.0000	0.9838	0.0000

varying-error model with two different read lengths: 500 bp and 150 bp. First, we generated each dataset from a number of genomes that included 10,000 reads with the average length of 500 bp ('long reads,' henceforth). Second, we generated the same set of simulated data with an average read length of 150 bp ('short reads,' henceforth). With these simulated datasets, we compared the performance of TAMER and TADIP by estimating the proportion of correct assignment (true positive (TP)) and the proportion of incorrect assignment (false positive (FP) at different taxonomy ranks. Incorrect assignment includes reads that were aligned incorrectly or overmatched. Here, TP represents the number of correctly assigned reads / the total number of reads (10,000). FP represents the number of incorrectly assigned reads / the total number of reads (10,000).

Real data study

Using TADIP, we examined eight oral datasets from human oral cavity (http://www.mg-rast.org) [15], and 11 gut datasets (https://www.ncbi.nlm.nih.gov) [16] generated from two real metagenomic studies. The oral cavity study compared the metagenomics in four groups: healthy controls who never had caries against patients who had been treated caries; those who had active caries; and those who had cavities. Each group contributed two samples. The datasets included ~ 2 million reads in total, and the average read length was 425 ± 117 bp, representing the long reads. The smallest samples had ~ 70,000 reads, while the largest sample had ~ 465,000 reads [15]. In addition, we examined 11 human gut data, obtained from a study of Crohn's disease, an inflammatory bowel disease (https://www.ncbi.nlm.nih.gov) [16]. Crohn's disease results in changes of microbial community in the human gut [17]. This dataset comprised seven healthy donors and four donors with Crohn's disease. The whole genome reads were generated using the Illumina platform, and the average length of the whole genome is 119 bp, representing the short reads [16].

Results

Type I error and power

Table 1 shows that the burden test was less conservative than the variance component test where the empirical type I error for the burden test was 0.05 and that for variance component test was 0.02. Table 1 further shows that two tests were valid and powerful, despite the relatively small sample size. Specifically, the power estimate

Table 3 Results for simulation study: Short reads. The proportions of reads correctly (TP) and incorrectly (FP) assigned to taxonomy tree at different ranks of two methods with average length of 150 bp

	simLC				simMC				simHC			
	TAMER		TADIP		TAMER		TADIP		TAMER		TADIP	
Rank	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
Species	0.7102	0.0000	0.7102	0.0000	0.6704	0.0398	0.6704	0.0065	0.7113	0.0242	0.7289	0.0055
Genus	0.7102	0.0000	0.7102	0.0000	0.6704	0.0398	0.6704	0.0051	0.7113	0.0242	0.7289	0.0055
Family	0.7102	0.0000	0.7102	0.0000	0.7102	0.0000	0.7102	0.0000	0.7002	0.0353	0.7002	0.0342
Order	0.7102	0.0000	0.7102	0.0000	0.7102	0.0000	0.7102	0.0000	0.7002	0.0353	0.7002	0.0342
Class	0.7102	0.0000	0.7102	0.0000	0.7102	0.0000	0.7102	0.0000	0.7337	0.0018	0.7337	0.0018
Phylum	0.7102	0.0000	0.7102	0.0000	0.7102	0.0000	0.7102	0.0000	0.7337	0.0018	0.7337	0.0018
Kingdom	0.7102	0.0000	0.7102	0.0000	0.7102	0.0000	0.7102	0.0000	0.7337	0.0018	0.7337	0.0018

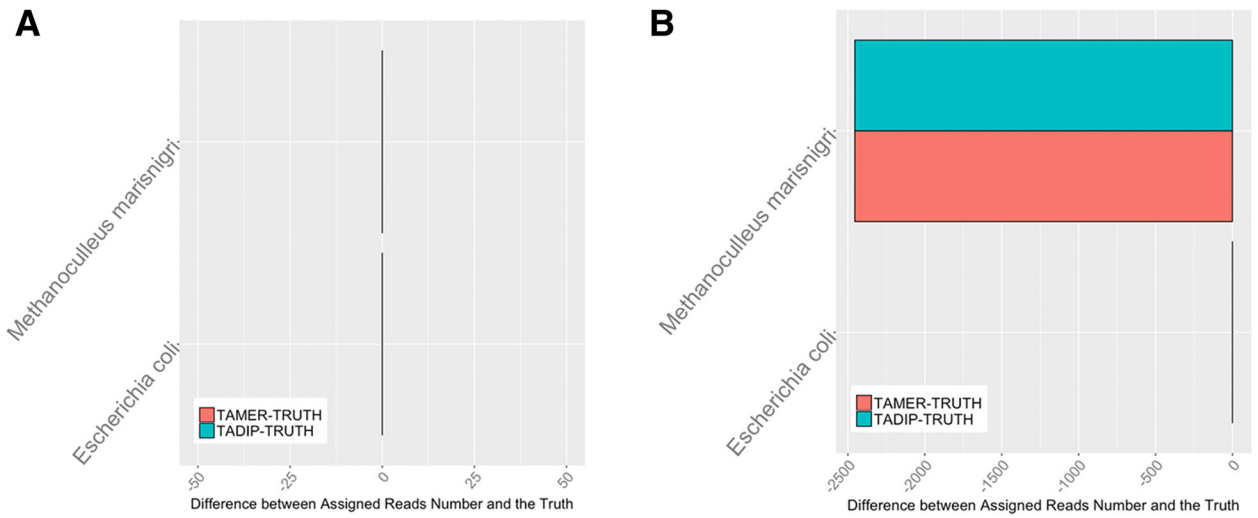

Fig. 1 Results for low level complexity simulation study. Numbers of reads assigned using TAMER and TADIP were compared with the true values at the level of Species for the simLC with long read length (**a**) and short read length (**b**)

for the burden test was 0.99, while that for the variance component test was 1.00.

Comparison of TAMER and TADIP methods

We compared the performance of TAMER and TADIP using three datasets with low, medium and high levels of complexity as defined by the number of genomes (i.e., simLC, simMC, and simHC). Datasets with different mismatch probabilities were tested for both long and short reads. Table 2 shows the results for the long reads, and Table 3 represents the results for short reads.

Low level complexity with two genomes

For datasets with long reads, the simulation study revealed that, at the level of Species, both TAMER and TADIP assigned 100.0% of the reads correctly, and had 0.00% false assignments. It is evident that the numbers of reads that TAMER and TADIP assigned were close to the true values at the level of Species (Fig. 1). However, the *p*-values of the burden test and variance component test were less than 0.05. This is because of different taxonomic composition, different genomes were assigned to the sample based on TAMER and TADIP, though these genomes belong to the same species. When the datasets with short reads were examined in Table 3, the rates of TP were lower than those in long reads, and this is most likely due to low matching rate from BLAST. Both TAMER and TADIP assigned 71.0% of the reads correctly at the at all rank levels and had 0.00% false assignments. The *p*-values of the burden test and variance component test were 0.131 and 0.009, respectively (Table 4). The test results indicate that TAMER and TADIP are both appropriate to use in this simplest setting.

Table 4 Simulation results of hypothesis testing for three benchmark data sets. Test results of burden test and variance component test indicating the need of different mismatch probabilities setting in the models of these simulation samples

Group	Tests	Burden Test	Variance Component Test
simLC	Long reads	0.0004	0.02
	Short reads	0.131	0.009
simMC	Long reads	< 0.0001	< 0.0001
	Short reads	0.002	< 0.0001
simHC	Long reads	< 0.0001	< 0.0001
	Short reads	< 0.0001	< 0.0001

Medium level complexity with nine genomes

For the datasets with long reads, TAMER assigned 92.3% of the reads correctly at the level of Species and had 17.3% false assignments. On the other hand, TADIP assigned 92.3% of the reads correctly and has 10.0% false assignments (Table 2). Note that the sum of TP and FP is greater than 1 under this setting due to the occurrence of multiple assignments. We then counted the numbers of reads that TAMER and TADIP assigned were close to the true value at the level of Species (Fig. 2). Particularly, the assignment number of *Escherichia coli*, Francisella tularensis and Shigella dysenteriae using TADIP was notably close to the truth. We then tested the datasets with short reads. TAMER assigned 67.0% of the reads correctly at the level of Species, and had 3.98% false assignments, while TADIP assigned 67.0% of the reads correctly and has 0.65% false assignments (Table 3). The assignment number of *Escherichia coli*, Shigella dysenteriae using TADIP in this simulation was notably close to the truth, however, the detected number of Francisella tularensis, *Pasteurella multocida*, Pseudomonas

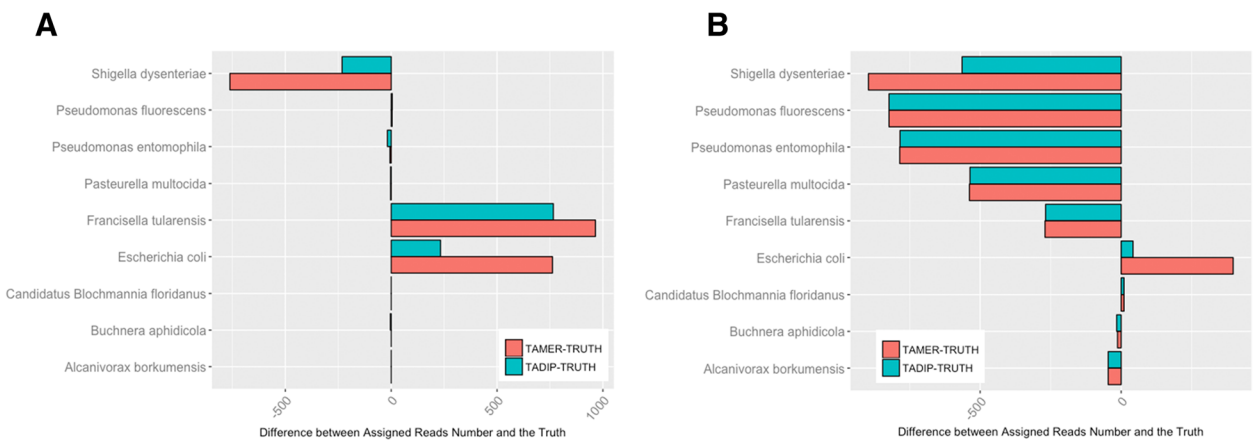

Fig. 2 Results for medium level complexity simulation study. Numbers of reads assigned using TAMER and TADIP were compared with the true values at the level of Species for the simMC with long read length (**a**) and short read length (**b**)

entomophila, *Pseudomonas fluorescens* are less than the truth both for TAMER and TADIP as shown in Fig. 2, which contribute to the decrease of TP. The *p*-values for the burden test and variance component test of two simulation studies were less than 0.05, suggesting that it was appropriate to use TADIP in this setting.

High level complexity with 15 genomes

For the datasets with long reads, TAMER assigned 96.4% of the reads correctly at the level of Genus, and had 1.94% false assignments. On the other hand, TADIP assigned 96.4% of the reads correctly, and had 0.87% false assignments (Table 2). It is evident that the numbers of reads that TAMER and TADIP assigns were close to the true values at the level of Species, even though there were differences in subspecies or strain. The assignment number of Escherichia and Shigella in this simulation using TADIP is closer to the truth (Fig. 3). For the dataset with short reads, TAMER assigns 71.1% of the reads correctly at the level of Species, and has 2.42% false assignments. TADIP assigns 72.9% of the reads correctly and has 0.55% false assignments (Table 3). The assignment number of Escherichia and Shigella using TADIP is also closer to the truth (Fig. 3). In the meanwhile, the assigned number is less than the truth over half of the species both for TAMER and TADIP. The *p*-values of burden test and variance component test were close to 0 and less than 0.05, respectively.

As the complexity increases, TADIP performs better than TAMER, especially for generating lower levels of FPs. The three levels of complexity for simulation showed that TADIP performed better than TAMER, when mismatch probabilities were different across datasets. However, the enhancement is weakened when the length of the reads

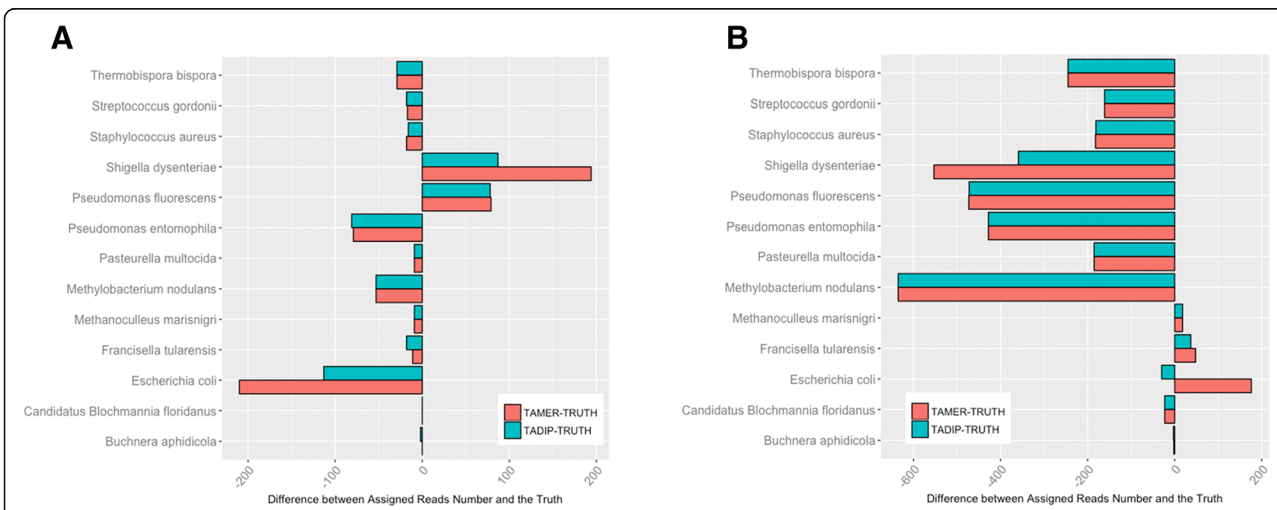

Fig. 3 Results for high level complexity simulation study. Numbers of reads assigned using TAMER and TADIP were compared with the true values at the level of Species for the simHC with long read length(**a**) and short read length (**b**)

Fig. 4 Results for the study of oral metagenomics data. Numbers of reads assigned using TAMER and TADIP for the representative Classes of the eight oral samples

was shorter such that the likelihood of low matching rate of each assignment increased.

Oral metagenomics

We identified approximately 2500 species in these eight oral samples, comprising two controls and six patients (grouped by treated caries, active caries, cavities), and the number of identified species varied by sample ranging from 700 to 1400. We estimated the proportions of reads assigned to the dominant Classes based on TADIP as shown in Additional file 1: Figure S1 and Table S1. Under the TADIP analysis, we can observe that in general diseased samples had more Bacteroidia, Fusobacteriia at the rank level of Classes than the healthy samples. On the other hand, controls had more Gammaproteobacteria that were seem to be absent in five diseased samples. In addition, the proportions of Betaproteobacteria and Actinobacteria behaved oddly in the first and last patient samples, and some other Class such as Epsilonproteobacteria, Coriobacteriia appeared in samples with caries or cavities. As the eight samples were chosen with a range of clinical features, there was a large variation among the samples, indicating the individuation in oral samples. Under the TAMER analysis, most of estimated proportions of dominant Classes were similar with the result of TADIP, except for evident differences in the bar plot of first patient with treated caries (Fig. 4). TAMER identified less Bacilli and Fusobacteria and Gammaproteobacteria than TADIP. It is of interest to note that the results from TADIP were more consistent to the results in the original published [15], i.e., the amount of Gammaproteobacteria of that patient ought to be the largest of all eight samples. The result of burden test also showed that there were significance differences between the mismatch probabilities among different genomes.

Gut metagenomics

Around 300 dominant species were assigned in 11 gut samples consisting seven controls and four patients with Crohn's disease. Fig. 5 shows that the estimated proportions of reads assigned to the dominant Phylums based on TADIP. Under the TADIP analysis, the four major Phylum in the human gut were Firmicutes, Bacteroidetes, Actinobacteia and Proteobacteria, replicating the earlier published results [16]. In our analysis, Verrucomicrobia was in high abundance compared with Proteobacteria, and Actinobacteia in three healthy samples and one patient sample (Fig. 5). This agrees with some findings in human gut that Verrucomicrobia can be occasionally observed [3, 18]. In general, the proportions of the phylum Proteobacteria and Actinobacteia were higher in Crohn's disease patients than in healthy controls. In addition, the proportions of Firmicutes and Bacteroidetes were unusual in the samples from the first two Crohn's disease patients, indicating that they might have been over-represented or depleted. These observations on samples from Crohn's disease patients agree with previous findings [19–24], supporting the notion that the causes of Crohn's disease among patients vary widely. Under the TAMER analysis, most of estimated proportions of dominant Phylums were similar to the results from the TADIP analysis where fewer Phylums were detected in the second diseased sample and the sixth control sample (Additional file 1: Figures S2-S3 and Tables S2-S3).

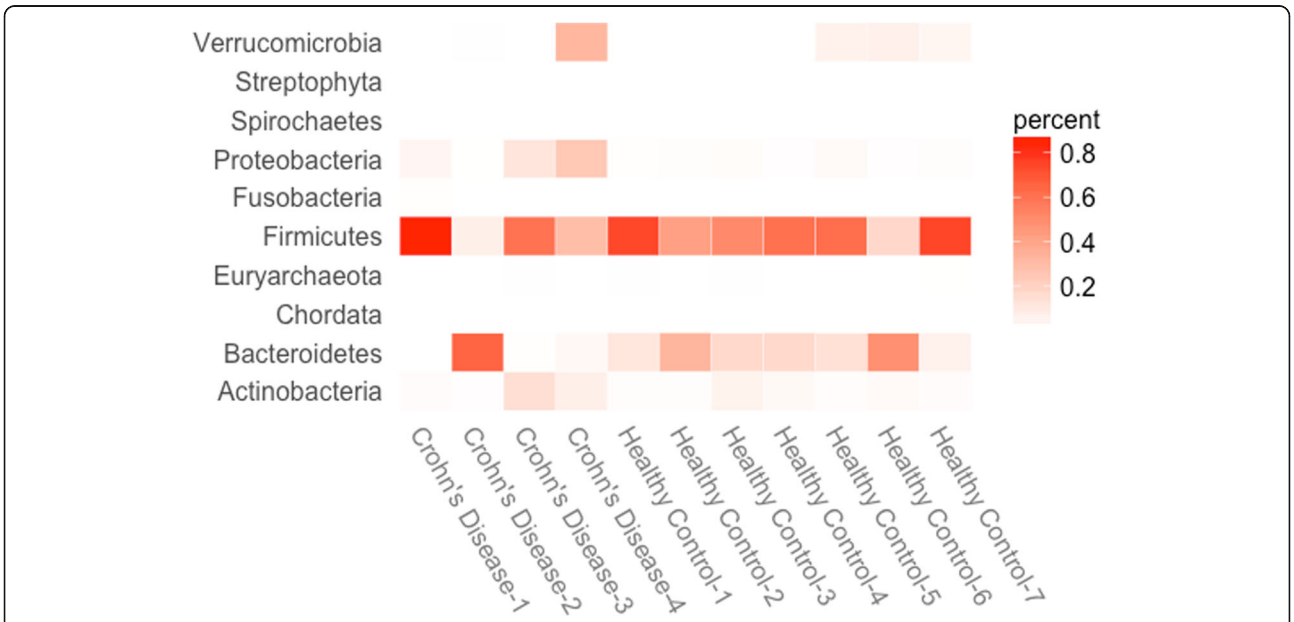

Fig. 5 Results for the study of gut metagenomics data. Heat maps for representative Phylum show the estimated proportion of reads assigned to each of the 11 samples based on the TADIP model

The burden test showed that there were significance differences at the significance level of 5%, indicating that the mismatch probabilities among genomes were significant.

Discussion

We have proposed a statistical framework called TADIP to improve the estimate accuracy of taxonomy assignments in metagenomic data. This approach extends the TAMER method and allows efficient analysis of metagenomics data by allowing different mismatch probabilities to different candidate genomes, rather than using one common mismatch probability for an array of different genomes.

It has been shown that TAMER performs better than MEGAN at high taxonomic ranks and estimates with a greater degree of accuracy at the genus level or even at the species level [2]. However, TAMER does not perform well in samples with a high degree of complexity. Our study has demonstrated that TADIP performed better for high complexity samples, because TADIP takes into account the correlations among different genomes with different mismatch probabilities. Using simulated and real datasets, we showed that TADIP can overcome some of the problems faced by complex samples. When we tested both simulation and real data incorporating significantly different mismatch probabilities among samples, for highly complicated samples, we showed that the burden test and variance component tests may yield different results because of the opposite assumptions, especially when sample size is large and these is a high degree of complexity. A future work will apply a bootstrap method to further explore this problem by resampling the original sequence reads with replacement for the statistical inference.

BLAST is considered as a general-purpose tool of the preprocessing step for metagenomics study, in which the set of DNA sequences is compared against publicly available databases [4]. However, we can also use some other efficient alignment tools such as BOWTIE [25] for short reads. Output from these alignment tools can serve as input for TADIP. The results from BLAST, Bowtie, and TADIP are presented in Additional file 1.

We note that selection of the right match algorithm is important. TADIP, TAMER and MEGAN rely on homologous searches of the sequence reads in the reference databases, and these algorithms do not perform well when the reads are generated from new genomes and the reference databases contain limited genome data. When a limited set of genomic data is available on novel genus, order or higher level in the evolutionary tree, algorithms that employ the sequence composition approach that characterizes sequence reads phylogenetically, such as PhyloPhythiaS and Phymm, performed substantially better than other algorithms [26, 27].

Conclusions

TADIP is developed as a statistical model to improve the estimate accuracy of taxonomy assignments. TADIP allows a varying mismatch probability setting and a correlated variance matrix setting to mimic the biological reality (i.e., truth). It performs better than TAMER for high complexity samples, especially in samples that contain different species, where different mismatch probabilities are likely to be abundant.

Additional file

Additional file 1: Figure S1. | Heat maps for the representative Classes of oral metagenomics data based on the TADIP (A) model and the TAMER (B) model. **Figure S2.** | Heat maps for the representative Species of gut metagenomics data based on the TADIP (A) model and the TAMER (B) model. **Figure S3.** | Numbers of reads assigned using TAMER and TADIP for the representative Phylums of gut metagenomics data. **Figure S4.** | Comparison of Blast and Bowtie for the study of oral metagenomics data. **Table S1.** | Tables for the estimated proportion of reads assigned to representative Classes of oral metagenomics data based on the TADIP model and the TAMER model. **Table S2.** | Tables for the estimated proportion of reads assigned to representative Phylum of gut metagenomics data(disease)based on the TADIP model and the TAMER model. **Table S3.** | Tables for the estimated proportion of reads assigned to representative Phylum of gut metagenomics data(control)based on the TADIP model and the TAMER model. Supplementary Code | BLAST. Supplementary Code | Bowtie. Supplementary Code | R package for TADIP and Hypothesis Testing. (DOCX 4876 kb)

Abbreviations

BLAST: Basic local alignment search tool; CNVs: Copy number variations; FP: False positives; MEGAN: Metagenome analyzer; simHC: Simulated data with high level complexity; simLC: Simulated data with low level complexity; simMC: Simulated data with medium level complexity; SNPs: Single nucleotide polymorphisms; STR: Short tandem repeats; TADIP: Taxonomic assignment of metagenomics base on different probabilities; TAEC: Taxonomic analysis by elimination and correction; TAMER: Taxonomic assignment of metagenomic sequence reads; TP: True positives

Funding

This study was supported in part by the funding from the NIH/NIA (AG054186 for ZJ, JL; AG051876 for JL). The funding agreement ensured the authors' independence in designing the study, collecting, analyzing and interpreting the data, and writing the manuscript.

Availability of data and materials

Simulated data were generated as specified in the Methods section. The de-identified publicly available data can be obtained from the MG-RAST Project ID mgp128 and NCBI BioProject ID numbers 82,111 and 175,224.

Authors' contributions

ZJ, YY and JL conceived and designed experiments; YY and ZJ performed the experiments; YY, JL and ZJ analyzed and interpreted the data; YY, JL and ZJ drafted the manuscript; All authors read and approved the final version of the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Department of Biostatistics, Columbia University, New York, NY, USA. [2]Sergievsky Center, Taub Institute, and Departments of Epidemiology and Neurology, Columbia University, New York, NY, USA. [3]Sergievsky Center, Columbia University, 630 West 168th Street, P&S Unit 16, New York, NY 10032, USA.

Received: 18 February 2018 Accepted: 2 October 2018
Published online: 29 October 2018

References

1. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.
2. Jiang H, An L, Lin SM, Feng G, Qiu Y. A statistical framework for accurate taxonomic assignment of metagenomic sequencing reads. PLoS One. 2012; 7(10):46450.
3. Sohn MB, An L, Pookhao N, Li Q. Accurate genome relative abundance estimation for closely related species in a metagenomic sample. BMC Bioinf. 2014;15(1):242.
4. Huson DH, Auch AF, Qi J, Schuster SC. Megan analysis of metagenomic data. Genome Res. 2007;17(3):377–86.
5. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
6. Fan J, Liao Y, Yao J. Power enhancement in high-dimensional cross-sectional tests. Econometrica. 2015;83(4):1497–541.
7. Brookes AJ. The essence of snps. Gene. 1999;234(2):177–86.
8. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89(1):82–93.
9. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. Am J Hum Genet. 2014;95(1):5–23.
10. Zhao N, Chen J, Carroll IM, Ringel-Kulka T, Epstein MP, Zhou H, Zhou JJ, Ringel Y, Li H, Wu MC. Testing inmicrobiome-profiling studies with mirkat, the microbiome regression-based kernel association test. Am J Hum Genet. 2015;96(5):797–807.
11. Davies RB. The distribution of a linear combination of χ2 random variables. Appl Stat. 1980;29(3):323–33.
12. Richter DC, Ott F, Auch AF, Schmid R, Huson DH. Metasim—a sequencing simulator for genomics and metagenomics. PLoS One. 2008;3(10):3373.
13. Pattnaik S, Gupta S, Rao AA, Panda B. Sinc: an accurate and fast error-model based simulator for snps, indels and cnvs coupled with a read generator for short-read sequence data. BMC Bioinf. 2014;15(1):40.
14. Jia B, Xuan L, Cai K, Hu Z, Ma L, Wei C. Nessm: a next-generation sequencing simulator for metagenomics. PLoS One. 2013;8(10):75448.
15. Belda-Ferre P, Alcaraz LD, Cabrera-Rubio R, Romero H, Simon-Soro A, Pignatelli M, Mira A. The oral metagenome in health and disease. ISME J. 2012;6(1):46.
16. Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, Reyes JA, Shah SA, LeLeiko N, Snapper SB, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol. 2012;13(9):79.
17. Khor B, Gardet A, Xavier RJ. Genetics and pathogenesis of inflammatory bowel disease. Nature. 2011;474(7351):307.
18. Dubourg G, Lagier J-C, Armougom F, Robert C, Audoly G, Papazian L, Raoult D. High-level colonisation of the human gut by verrucomicrobia following broad-spectrum antibiotic treatment. Int J Antimicrob Agents. 2013;41(2):149–55.
19. Thota VR, Dacha S, Natarajan A, Nerad J. Eggerthella lentabacteremia in a crohn's disease patient after ileocecal resection. Future Microbiol. 2011;6(5):595–7.
20. P´erez-Brocal V, Gar´cıa-Lopez R, V'azquez-Castellanos JF, Nos P, Beltr´an B, Latorre A, Moya A. Study of the viral and microbial communities associated with crohn's disease: a metagenomics approach. Clin Transl Gastroenterol. 2013;4(6):36.
21. Sartor RB. Microbial influences in inflammatory bowel diseases. Gastroenterology. 2008;134(2):577–94.
22. Liu Y, Van Kruiningen HJ, West AB, Cartun RW, Cortot A, Colombel J-F. Immunocytochemical evidence of listeria, escherichia coli, and streptococcus antigens in crohn's disease. Gastroenterology. 1995;108(5):1396–404.
23. Man SM, Kaakoush NO, Mitchell HM. The role of bacteria andpattern-recognition receptors in crohn's disease. Nat Rev Gastroenterol Hepatol. 2011;8(3):152.
24. Frank DN, Amand ALS, Feldman RA, Boedeker EC, Harpaz N, Pace NR. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. Proc Natl Acad Sci. 2007;104(34):13780–5.
25. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie2. Nat Methods. 2012;9(4):357.
26. Patil KR, Haider P, Pope PB, Turnbaugh PJ, Morrison M, Scheffer T, McHardy AC. Taxonomic metagenome sequence assignment with structured output models. Nat Methods. 2011;8(3):191–2.
27. McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I. Accurate phylogenetic classification of variable-length dna fragments. Nat Methods. 2007;4(1):63–72.