OPEN ACCESS Freely available online

PLOS ONE

A Statistical Framework for Accurate Taxonomic Assignment of Metagenomic Sequencing Reads

Hongmei Jiang[1*☯], Lingling An[2,3☯], Simon M. Lin[4,5], Gang Feng[6], Yuqing Qiu[1]

1 Department of Statistics, Northwestern University, Evanston, Illinois, United States of America, **2** Department of Agricultural and Biosystems Engineering, University of Arizona, Tucson, Arizona, United States of America, **3** Interdisciplinary Program in Statistics, University of Arizona, Tucson, Arizona, United States of America, **4** Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, Wisconsin, United States of America, **5** Institute for Clinical and Translational Research, University of Wisconsin at Madison, Madison, Wisconsin, United States of America, **6** Biomedical Informatics Center, Northwestern University, Chicago, Illinois, United States of America

Abstract

The advent of next-generation sequencing technologies has greatly promoted the field of metagenomics which studies genetic material recovered directly from an environment. Characterization of genomic composition of a metagenomic sample is essential for understanding the structure of the microbial community. Multiple genomes contained in a metagenomic sample can be identified and quantitated through homology searches of sequence reads with known sequences catalogued in reference databases. Traditionally, reads with multiple genomic hits are assigned to non-specific or high ranks of the taxonomy tree, thereby impacting on accurate estimates of relative abundance of multiple genomes present in a sample. Instead of assigning reads one by one to the taxonomy tree as many existing methods do, we propose a statistical framework to model the identified candidate genomes to which sequence reads have hits. After obtaining the estimated proportion of reads generated by each genome, sequence reads are assigned to the candidate genomes and the taxonomy tree based on the estimated probability by taking into account both sequence alignment scores and estimated genome abundance. The proposed method is comprehensively tested on both simulated datasets and two real datasets. It assigns reads to the low taxonomic ranks very accurately. Our statistical approach of taxonomic assignment of metagenomic reads, TAMER, is implemented in R and available at http://faculty.wcas.northwestern.edu/~hji403/MetaR.htm.

Citation: Jiang H, An L, Lin SM, Feng G, Qiu Y (2012) A Statistical Framework for Accurate Taxonomic Assignment of Metagenomic Sequencing Reads. PLoS ONE 7(10): e46450. doi:10.1371/journal.pone.0046450

Editor: Haixu Tang, Indiana University, United States of America

Received July 16, 2012; **Accepted** August 30, 2012; **Published** October 1, 2012

Funding: This work was supported by National Science Foundation (DMS 1043080 to HJ, LA, and SL, and DMS 1222592 to LA and HJ) and partially by National Institutes of Health (1UL1RR025011 to SL and UL1RR025741 to GF). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: hongmei@northwestern.edu

☯ These authors contributed equally to this work.

Introduction

Traditional and classical methods of genomics and microbiology allow researchers to study an individual microbial species obtained from the environment by isolating the organism into pure colonies using microbial culture techniques. However, this approach cannot capture the structure of the broader microbial community within the environmental sample, the relative representation of multiple genomes, and their interaction with each other and with the environment. Additionally, a large number of microbial species are very difficult, or impossible, to culture *in vitro* in the laboratory setting. The development of next-generation sequencing has advanced the field of metagenomics by enabling scientists to simultaneously study multiple genomes recovered directly from an environmental sample, thereby bypassing the need for microbial isolation through culturing (see [1] for a review).

In a metagenomic experiment, a sample is usually taken from a natural (e.g., soil and seawater) or a host-associated (e.g., human gut) environment containing micro-organisms organized into communities or microbiomes. DNA is extracted from the environmental sample containing a mixture of multiple genomes and then sequenced without prior separation. The resulting dataset comprises millions of mixed sequence reads from the multiple genomes contained in the sample. Traditionally, DNA has been sequenced using Sanger sequencing technology [2] and the reads generated are routinely 800–1000 base pairs long. However this technology is extremely cumbersome and costly. Recently next-generation sequencers, e.g., Illumina/Solexa, Applied Biosystems' SOLiD, and Roche's 454 Life Sciences sequencing systems, have emerged as the future of genomics with incredible ability to rapidly generate large amounts of sequence data [3,4]. These new technologies greatly facilitate high-throughput while lowering the cost of metagenomic studies. However, the reads generated are of much shorter length making reads assembly and alignment more challenging. For example, Illumina/Solexa and SOLiD generate reads ranging between 35–100 base pairs while Roche 454 reads are approximately 100–400 base pairs in length.

One goal of metagenomic studies is to identify what genomes are contained in the environmental sample and to estimate their

relative abundance. Identification of genomes is complicated by the mixed nature of multiple genomes in the sample. A widely used approach is assigning the sequence reads to NCBI's taxonomy tree based on sequence read homology alignment with known sequences catalogued in reference databases. The sequence reads are first aligned to the reference sequence databases using a sequence comparison program such as BLAST [5]. Reads which have hits in the database are then assigned to the taxonomy tree based on the best match or multiple high-scoring hits. The challenge of this approach is that hits may be found in multiple genomes for a single read at a given threshold of bit-score or Expect value, due to sequence homology and overlaps associated with similarity among species. Strategy of weighting similarities for multiple BLAST hits has been used to estimate the relative genomic abundance and average size [6]. Another representative and stand-alone analysis tool, MEGAN [7], assigns a read with hits in multiple genomes to their lowest common ancestor (LCA) in the NCBI taxonomy tree. Thus assignments of reads to different ranks of taxonomy tree depend on what threshold for bit-score or Expect value is used. Furthermore, MEGAN assigns reads one at a time. As a consequence, the results have less false positives but lack specificity. Various methods have been proposed to improve the taxonomic assignment of reads by assigning more reads to the lower ranks of taxonomy tree [8–12]. In particular, CARMA3 [10] which is BLAST-based but not LCA-based, uses reciprocal search technique as in SOrt-ITEMS [13] to reduce the number of hits and hence further improves the accuracy of the taxonomic classification.

In this paper, we propose a statistical approach, TAMER, for taxonomic assignment of metagenomic sequence reads. In this approach we first identify a list of candidate genomes using homology searches. A mixture model is then employed to estimate the proportion of reads generated by each candidate genome. Finally, instead of assigning reads one at a time to the taxonomy tree as done by LCA-based methods, reads are assigned to the genomes in a global manner by taking into account both sequence alignment scores and estimated proportion of reads generated by each genome. The proposed method is comprehensively tested on simulated metagenomic data with diverse complexity of microbial community structure and with various read length and also applied to several real world metagenomic datasets. Compared with other available algorithms and tools designated for metagenomic analysis, the proposed approach demonstrates greater accuracy in identification and quantification of multiple genomes in a given sample.

Materials and Methods

Input Data

The proposed homology-based method, TAMER, identifies multiple genomes based on the hits obtained by aligning sequence reads against known reference sequence databases. In this paper we use the NCBI-NT instead of NCBI-NR database as reference sequence source in our data analysis. NT database contains almost all known nucleotide sequences of all known species from NCBI GeneBank, EMBL and DDBJ, while NR database does not have reference sequences for reads generated from intergenic regions. Although BLAST is the traditional sequence comparison and alignment tool for the NT database, computation time is the bottleneck [7,9]. High performance computing infrastructure and fast alignment tools such as BLAT [14] have been recommended when dealing with large megagenomic datasets [15]. The alignment tools developed especially for next generation sequencing technologies retrieve matches with high similarities. In this research we use MegaBLAST (version 2.2.25+) which yields matches with relatively high similarities but is much faster than BLASTn [16]. Notably, our proposed method has great versatility and can also be applied when other alignment tools and other reference databases are used.

When aligning reads to the reference database, there may be multiple hits within one genome for a sequence read. In this situation, the hit with the largest number of identical matches is chosen to represent the corresponding genome. For each read and the corresponding hits in one or multiple genomes, we record the genome name or taxon identification number of the hit, number of matched base pairs, and the alignment length. These consist of the input data for the proposed method which evaluates the likelihood of alignment of a read with a given genome among the list of candidate genomes.

Mixture Model

Suppose we have a total of n sequence reads $x_1, x_2, \cdots x_n$ which have been aligned to K candidate genomes. For each of the K genomes, there is at least one sequence read having hit. Let L_{ji} denote the alignment length and M_{ji} be the number of matched base pairs when aligning read x_j against genome i. If a read x_j does not have any hit in genome i, then $M_{ji}=0$. Due to differences in genome compositions, a short read is usually aligned to one or a few genomes. Thus, the scoring matrix M below, where rows represent reads and columns represent genomes, is very sparse, i.e., most entries in the matrix are zero.

$$M=\begin{bmatrix} M_{11} & M_{12} & \cdots & M_{1K} \\ M_{21} & M_{22} & \cdots & M_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ M_{n1} & M_{n2} & \cdots & M_{nK} \end{bmatrix} \quad (1)$$

To identify which of the K candidate genomes in the scoring matrix are truly contained in the metagenomic sample, we propose a statistical framework to model the matches between the reads and reference sequences. Let R_i denote the proportion of reads generated from genome $i (i=1,2,\cdots,K)$, where $R_i \geq 0$ and the sum of R_i is 1. If the reads are randomly generated by the K genomes, then the probability that a read x_j is generated by genome i is R_i. Even if a read x_j is generated from genome i, it is possible that the match is not 100% identical due to sequencing errors, alignment errors, and/or single nucleotide polymorphism (SNP). Let p denote the probability of observing a mismatched base pair, then 1- p is the probability of observing a matched base pair. 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 is $R_i p^{L_j - M_{ji}}(1-p)^{M_{ji}}$, where $L_j = \max\{L_{ji}, i=1,\cdots,K\}$ is the maximum alignment length. Then the probability of observing a read x_j in the dataset is

$$\Pr(x_j) = \sum_{i=1}^{K} \left[R_i p^{L_j - M_{ji}} (1-p)^{M_{ji}} \right].$$

Assuming that the reads are independent of each other, the likelihood function of the data is:

$$\ell(p,R_1,\cdots,R_K)=\prod_{j=1}^{n}\Pr(x_j)=\prod_{j=1}^{n}\left\{\sum_{i=1}^{K}\left[R_i p^{L_j-M_{ji}}(1-p)^{M_{ji}}\right]\right\}, \quad (2)$$

where the values of L_j and M_{ji} are observable, and the parameters p and $R_i(i=1,2,...,K)$ are to be estimated.

EM Algorithm

For this mixture model, the expectation maximization (EM) algorithm [17] is used to calculate the maximum likelihood estimation for the parameters p and $R_i(i=1,2,...,K)$. Let $Z=(Z_1,\cdots,Z_n)$ be the latent variables that determine the genome from which each read originate. The aim is to estimate the unknown parameters $\theta=(p,R)$, where $R=(R_1,\cdots,R_K)$. The likelihood function can be written as:

$$\ell(\theta,M,Z)=\prod_{j=1}^{n}\left\{\sum_{i=1}^{K}\left[I(z_j=i)R_i p^{L_j-M_{ji}}(1-p)^{M_{ji}}\right]\right\}$$

where I is an indicator function. As the density function is an exponential family function, the likelihood function can be expressed as:

$$\ell(\theta,M,Z)=\exp\left\{\sum_{j=1}^{n}\sum_{i=1}^{K}\left[I(z_j=i)[\log(R_i)-M_{ji}\log(p/(1-p))+L_j\log p]\right]\right\}$$

- **Initialization step.** Initialize the values of p and $R_i(i=1,2,\cdots,K)$, call them $p^{(0)}$ and $R_i^{(0)}$. For instance, let the reads be equally distributed among the K genomes, i.e., $R_i^{(0)}=1/K$, and let $p^{(0)}=0.05$.
- **E-step.** Assuming the current estimate of the parameter is $\theta^{(t)}$, then the conditional distribution of Z_j is:

$$T_{ji}^{(t)}:=\Pr(z_j=i|M;\theta^{(t)})=\frac{R_i^{(t)}(p^{(t)})^{L_j-M_{ji}}(1-p^{(t)})^{M_{ji}}}{\sum_{v=1}^{K}R_v^{(t)}(p^{(t)})^{L_j-M_{jv}}(1-p^{(t)})^{M_{jv}}}. \quad (3)$$

Then the E-step result is:

$$Q(\theta|\theta^{(t)})=E[\log(\ell(\theta,M,Z))]=\sum_{j=1}^{n}\sum_{i=1}^{K}T_{ji}^{(t)}\left[\log(R_i)-M_{ji}\log(p/(1-p))+L_j\log p\right]$$

- **M-step.** As the parameters can be maximized independently, we get:

$$R^{(t+1)}=\arg\max_R Q(\theta|\theta^{(t)})=\arg\max_R\left\{\sum_{i=1}^{K}\left[(\log R_i)\sum_{j=1}^{n}T_{ji}^{(t)}\right]\right\}.$$

This gives $R_i^{(t+1)}=\frac{1}{n}\sum_{j=1}^{n}T_{ji}^{(t)}\quad(i=1,2,\cdots,K)$.

The probability of observing a mismatched base pair is estimated as:

$$p^{(t+1)}=1-\frac{\sum_{j=1}^{n}\sum_{i=1}^{K}M_{ji}T_{ji}^{(t)}}{\sum_{j=1}^{n}\sum_{i=1}^{K}L_j T_{ji}^{(t)}}$$

- **Iteration step.** Repeat the E-step and the M-step until all the parameters converge, i.e., $|p^{(t+1)}-p^{(t)}|<\varepsilon$ and $|R_i^{(t+1)}-R_i^{(t)}|<\varepsilon$ for $i=1,2,\cdots,K$ and for some pre-specified small number of ε.

The estimates of $R_i(i=1,2,\cdots,K)$ reflect the proportion of reads generated from each of the K candidate genomes. If $R_i=0$, then the corresponding genome i is not contained in the sample. If we observe an inequality $R_i>R_{i'}$ for two genomes i and i', then we conclude that the sample contains more reads generated from genome$i$than genome i'. However the values of R_i do not give information on which reads are generated by which genomes. Next we show how to assign reads to the K candidate genomes and the taxonomy tree.

Taxonomic Assignment of Reads

To assign each read to the taxonomic tree, we first estimate how likely it is generated by a specific genome. The probability that read x_j is generated by genome i is estimated by.

$$P_{ji}:=\frac{R_i p^{L_j-M_{ji}}(1-p)^{M_{ji}}}{\sum_{v=1}^{K}R_v p^{L_j-M_{jv}}(1-p)^{M_{jv}}}$$

for $i=1,2,\cdots,K$ and $j=1,2,\cdots,n$. Then read x_j is assigned to the genome for which the maximum probability is reached, i.e., read x_j is assigned to genome $i_{\max}$ where $i_{\max}=\arg\max\{P_{ji},i=1,\cdots,K\}$. An assignment matrix $A=[a_{ji}]_{n\times K}$ can be constructed based on the read assignment, where $a_{ji}=1$ if read x_j is assigned to genome i, and $a_{ji}=0$ otherwise. Then the total number of reads assigned to genome i is $\sum_{j=1}^{n}a_{ji}$.

The proposed method, TAMER, applies to the candidate genomes to which the sequence reads have hits. Note the majority of the candidate genomes identified after performing BLAST are at the low ranks of the taxonomy tree, i.e., most of the genomes are species or substrings of species. Once a read is assigned to a specific genome, we also consider that it is assigned to taxa with higher taxonomic ranks. For example, suppose a read is assigned to *Escherichia coli str. K-12 substr. MG1655*. When we summarize reads assigned at different taxonomic ranks, this read is treated as that it is assigned to *Escherichia. coli* at rank Species, to *Escherichia* at rank Genus, to *Enterobacteriaceae* at rank of Family, and so on.

Estimates of Relative Genome Abundance

The number of sequence reads generated by a genome is proportional not only to the number of copies of that genome in the metagenomics sample but also to the length of the genome [6]. Similar to [18], the relative genome abundance can be computed for known genomes which are present in the sample. Let $G_i$denote the actual length of the genome$i$in base pairs. Suppose there are $C_i$copies of genome$i$in the sample. Assuming uniform distribution of reads across the multiple genomes, we have.

$$R_i = \frac{C_i G_i}{\sum_{h=1}^{K} (C_h G_h)}.$$

Then the relative abundance of genome i (i.e., relative copy number) in the sample can be calculated by.

$$\frac{C_i}{\sum_{h=1}^{K} C_h} = \frac{R_i/G_i}{\sum_{h=1}^{K} (R_h/G_h)}.$$

Algorithm Implementation

All algorithms developed in this research are implemented in R, a free software environment for statistical computing and graphics [19]. The R source codes are available at http://faculty.wcas.northwestern.edu/~hji403/MetaR.htm. For practical implementation, the scoring matrix M in equation (1) could require a huge storage space when the total number of reads is large. Recognizing that M is a sparse matrix, substantial memory requirement reductions can be achieved by storing only the non-zero matching scores. For the zero entries of M_{ji},their influence on estimating the parameters is nominal because we have $p^{L_j - M_{ji}}(1-p)^{M_{ji}} = p^{L_j} \approx 0$when $M_{ji} = 0$, for a small value of p(e.g., 0.02^35 = 3.4e-60). With the use of sparse matrix technique, detecting multiple genomes via the mixture model becomes very efficient. For example, the computational time for a dataset of 150,000 reads with average read length of 100 bp is about 2 ~ 3 minutes on a laptop with 8 GB RAM and 2 core 3.06 GHz CPU.

Simulation Studies

Due to the complexity of metagenomic data, simulation studies with verifiable results are crucial to benchmark TAMER and conduct comparisons with other existing methods. For the analysis by MEGAN the default parameters are used.

Simulation study 1. MetaSim [20], a sequencing simulator for genomics and metagenomics, is used to generate sequence reads for simulation studies. Four benchmark simulation datasets with low (2 genomes, simLC), medium (9 genomes, simMC), high (11 genomes, simHC), and super high (100 genomes, simSC) complexity are used. The first three setups were designed by [20] in conjunction with [21]. We use the simulation study with 100 genomes to reflect the high complex structure of some microbial communities which may contain hundreds even thousands of species. Here, we present the simulation studies using reads with an average read length of 100 bp for all four simulation studies, thereby mimicking next-generation sequencing short reads. For the medium and high complexity, we also perform simulation study using average read length of 400 bp. In this simulation study, we compare the performance of TAMER with MEGAN.

Simulation study 2. To compare TAMER with CARMA3 [10], we use the same evaluation dataset as in [10]. This CARMA3 evaluation dataset consists of 25,000 metagenomic reads which are randomly simulated from 25 bacterial genomes with an average read length of 265 bp. The online version of CARMA3, WebCARMA (http://webcarma.cebitec.uni-bielefeld.de/), with default parameters is used for taxonomic classification. We also perform the taxonomic analysis using TAMER and MEGAN, and compare their performance with CARMA3. When BLASTx and NR database are used, CARMA3 gives better taxonomic assignment than MEGAN [10]. Therefore we only present the results by MEGAN using MegaBLAST and NT database in this study.

Real Datasets

TAMER is also applied to two sets of actual metagenomic data. Archived metagenomic datasets are accessible from several sources including the NCBI short read archive [22], CAMERA [23], and the MG-RAST server [24]. In this paper we analyze data from eight oral samples and two seawater samples.

The eight oral samples downloaded from the MG-RAST server were examined in a human metagenome oral cavity study [25]. They represent different degrees of oral health with two samples for each of the four status, healthy controls (never with caries), treated for past caries, active caries, and cavities. There are totally about 2 million reads. The smallest sample has about 70,000 reads and the largest sample has about 465,000 reads. The average read length is 425 ± 117 bp.

The two seawater datasets were retrieved from MEGAN database (http://www.megan-db.org/megan-db/) and were studied in [20]. Each dataset consists of 10,000 reads and they are part of the Sargasso Sea Samples studied in [26]. The reads are about 800 bp long in both seawater datasets.

Results

Results for Simulation Study 1

Using the same abundance setup as in [20], 150,000 reads are generated for each of the three complexity datasets, simLC, simMC, and simHC, with average length of 100 bp. For the simSC dataset, 100 genomes with the same abundance are randomly selected and 150,000 reads are generated. The characteristics of the datasets are listed in Table S1. For this simulation study, we compare TAMER with MEGAN. The proportions of reads correctly (TP) and incorrectly (FP) assigned at different taxonomy ranks are reported in Table 1. Here TP = number of correctly assigned reads / total number of reads $\times 100$, and FP = number of incorrectly assigned reads/ total number of reads $\times 100$. For instance, for the simLC data, 146,880 reads are assigned to the corresponding species correctly, and 30 reads are assigned incorrectly, then $TP = 146{,}880/150{,}000 \times 100 = 97.92$ and $FP = 30/150{,}000 \times 100 = 0.02$. Note that the sum of TP and FP is not 100 as some reads do not have hits in the reference database.

The simLC dataset consists of 25,926 reads generated from *E. coli str. K-12 substr. MG1655* and 124,074 reads generated from *Methanoculleus marisnigri JR1*. Totally there are about 160 million base pairs and the simulated error rate is 0.027. The estimated probability of observing a mismatched base pair is 0.025 by TAMER. Using MegaBLAST, hits are found for 97.94% of the 150,000 reads in 4,407 unique taxa. At rank Species, TAMER accurately assigns 25,221 reads to species *Escherichia coli* which is close to the true value of 25,926 reads, while MEGAN only assigns 5,583 reads to this taxon (Figure 1 (a)). At rank Genus, MEGAN

Table 1. Results for simulation study 1 with average read length o1 100 bp.

	simLC				simMC				simHC				simSC			
	TAMER		MEGAN		TAMER		MEGAN		TAMER		MEGAN		TAMER		MEGAN	
	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
Species	97.92	0.02	84.74	0.01	96.14	0.86	67.57	0.83	96.97	0.15	91.97	0.00	95.70	1.20	86.05	0.11
Genus	97.92	0.00	85.02	0.01	96.17	0.76	72.95	0.86	97.00	0.03	93.59	0.00	95.87	0.70	91.67	0.00
Family	97.93	0.00	96.68	0.01	96.91	0.02	95.13	0.01	97.01	0.02	96.59	0.00	94.71	0.02	93.63	0.00
Order	97.93	0.00	96.69	0.01	96.91	0.02	95.20	0.01	97.01	0.02	96.74	0.00	95.36	0.01	94.37	0.00
Class	97.93	0.00	96.79	0.01	96.93	0.00	95.44	0.01	82.85	0.00	82.64	0.00	92.10	0.00	91.21	0.00
Phylum	97.93	0.00	96.87	0.01	96.93	0.00	95.51	0.01	97.03	0.00	96.92	0.00	96.16	0.00	95.35	0.00
Kingdom	97.94	0.01	96.98	0.01	96.99	0.00	95.66	0.01	97.11	0.00	96.97	0.00	96.90	0.00	96.19	0.00

The percentage of correctly (TP) and incorrectly (FP) assigned reads out of total 150,000 reads with average read length of 100 bp at different taxonomic ranks using TAMER and MEGAN for the simLC, simMC, simHC and simSC datasets in simulation study 1.

doi:10.1371/journal.pone.0046450.t001

assigns 5,974 reads to *Escherichia* which is only about 23% of the true value and the number of reads assigned by TAMER to that genus (Figure 1 (b)). Considering the low proportion of incorrect assignment (Table 1), TAMER accurately identifies and quantifies the different genomes at low taxonomic ranks.

The simMC dataset consists of nine microbial organisms from phylum *Proteobacteria* with diverse relative abundance. Hits are found for 97.00% of the 150,000 reads in 9,925 unique taxa. TAMER is able to dramatically reduce the huge number of taxa, and accurately identifies the nine organisms and assigns the reads to the corresponding originated organisms. TAMER assigns 96.14% of the reads correctly at rank Species, while MEGAN only assigns 67.57% of reads (Table 1). At rank Genus, the proportion of assigned reads by MEGAN is increased to 72.95%, however it is 23% less than that by TAMER. The percentage of incorrectly assigned reads is about 0.8% for both TAMER and MEGAN at both ranks of Species and Genus. It is evident that the number of reads assigned to different taxa by TAMER is very close to the true value, while MEGAN assigns 6,643 less reads to *Francisella tularensis* and 39,184 less reads to *Shigella dysenteriae* than TAMER does (Figure 2 (a)). At rank Genus, TAMER assigns 39,191 reads to *Shigella* which is close to the true value and is about eight times as many as MEGAN does (Figure 2 (b)).

The simHC dataset consists of 11 microbial organisms. Using MegaBLAST, hits are found for 97.11% of 150,000 reads in 2,511 unique taxa. TAMER identifies all 11 genomes and assigns the reads accurately to the original organisms. For these 11 distantly related organisms, MEGAN also does a satisfactory work by assigning about 92% of reads at rank Species which is 5% less than TAMER does (Table 1). Population distributions of reads at rank Species (Figure S1) and Genus (Figure S2) show that the assignments of reads by both methods are similarly accurate.

The simSC dataset is generated from 100 microbial organisms. About 96.90% of 150,000 reads have matches in 14,205 unique taxa. TAMER identifies 149 genomes with 103 of them having at least 5 assigned reads. Summarizing the results at different taxonomic ranks, TAMER assigns about 8% more reads than MEGAN at rank Species, and TAMER and MEGAN are comparable at higher taxonomic ranks (Table 1).

For simMC and simHC, we also perform a simulation study using 10,000 reads with average read length of 400 bp. With longer read length, the proportion of correctly assigned reads at low taxonomic ranks is improved for both methods. This further confirms the very well-known fact that longer reads are more sensitive in estimating the relative abundance of the multiple species. For simMC data, TAMER and MEGAN assign about 99.9% and 71.4% of reads correctly at rank Species, respectively, while the proportion of incorrectly assigned reads only increases about 0.1% for TAMER (Table S2). At rank Genus, TAMER assigns about 23% more reads correctly than MEGAN (99.91% for TAMER and 76.53% for MEGAN) while the false positive rate only increases about 0.08%. For simHC simulation study, the results of TAMER and MEGAN are highly comparable.

Results for Simulation Study 2

For the CARMA3 evaluation dataset, the results based on TAMER and MEGAN are listed in Table 2, where we also list the results of CARMA3 which are reported in the original paper [10]. At rank Species, the percentage of correctly assigned reads is 99.24% for TAMER, 81.45% for MEGAN, and 4.57% for CARMA3 (Table 2). At rank Genus, the proportion of correctly assigned reads by TAMER (99.26%) is 7% and 35% more than MEGAN (91.52%) and CARMA3 (64.10%), respectively.

Consistent with the conclusions from simulation study 1, the numbers of reads assigned by TAMER are very close to the true values, the true positive rate is high, and the false positive rate is very low. TAMER gives more accurate assignments than MEGAN and CARMA3 at rank Genus (Figure 3). For example, it assigns about 14 times as many as reads to *Shigella* than MEGAN and CARMA3.

Results for Real Data Analysis

Oral data. Identifying and quantifying bacterial species in the normal and diseased samples will help understand the development of dental caries. About 46% of the 2 million reads have hits and could be assigned to taxonomic ranks by TAMER. The number of identified species varies from about 700 to 1,400 across the eight samples. Totally 2,500 unique species are detected from this study, about 1,300 of them have at least 5 assigned reads, and about 400 species are shared by all samples.

Estimated proportions of reads for the dominant classes based on TAMER are shown in Figure 4 (a). Generally, normal sample contains more *Bacilli* and *Gammaproteobacteria* but less *Bacteroidia* than the diseased sample, which agrees with taxonomic assignment using MEGAN approach [25] (Figure S3). We also observe a large variation among the individual samples although the eight samples were selected with homogeneous clinical features. For

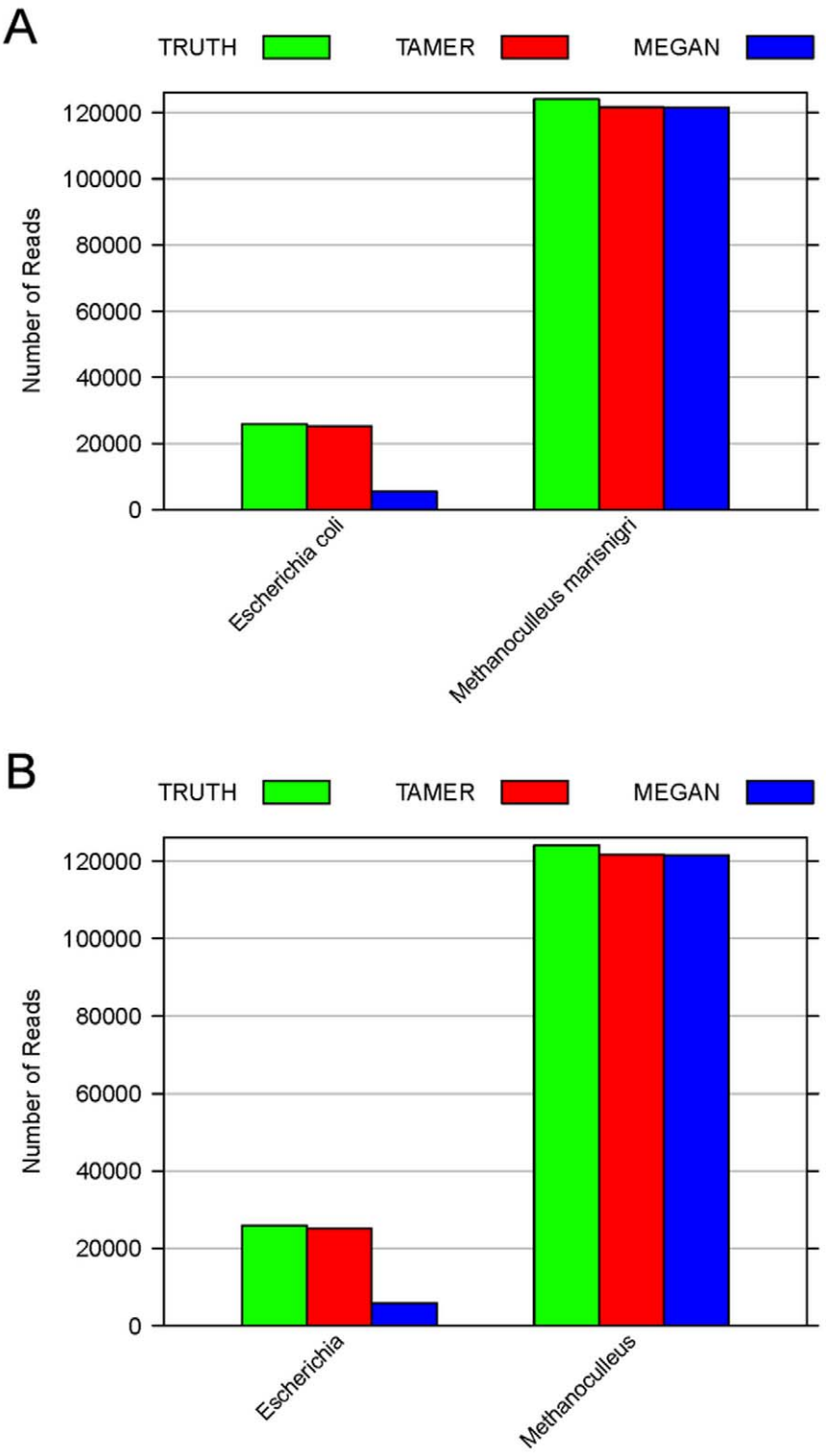

Figure 1. Reads assignment at rank Species and Genus for simLC dataset. Numbers of reads assigned to rank (A) Species and (B) Genus using TAMER and MEGAN are compared with the true values (TRUTH) for the simLC dataset with 150,000 reads and average read length of 100 bp in simulation study 1.
doi:10.1371/journal.pone.0046450.g001

instance, *Actinobacteria* is abundant in the two control samples, and depleted in the remaining samples except for one sample from within cavities where it shows high proportion. *Betaproteobacteria* is high in one of the two controls and one of the samples with treated cavities, but low in the remaining six samples. Examining the population distribution at the genus level (Figure 4 (b)), *Streptococcus* is enriched in the normal samples, *Prevotella* and *Veillonella* are associated with the disease, and *Fusobacterium* is not abundant in the

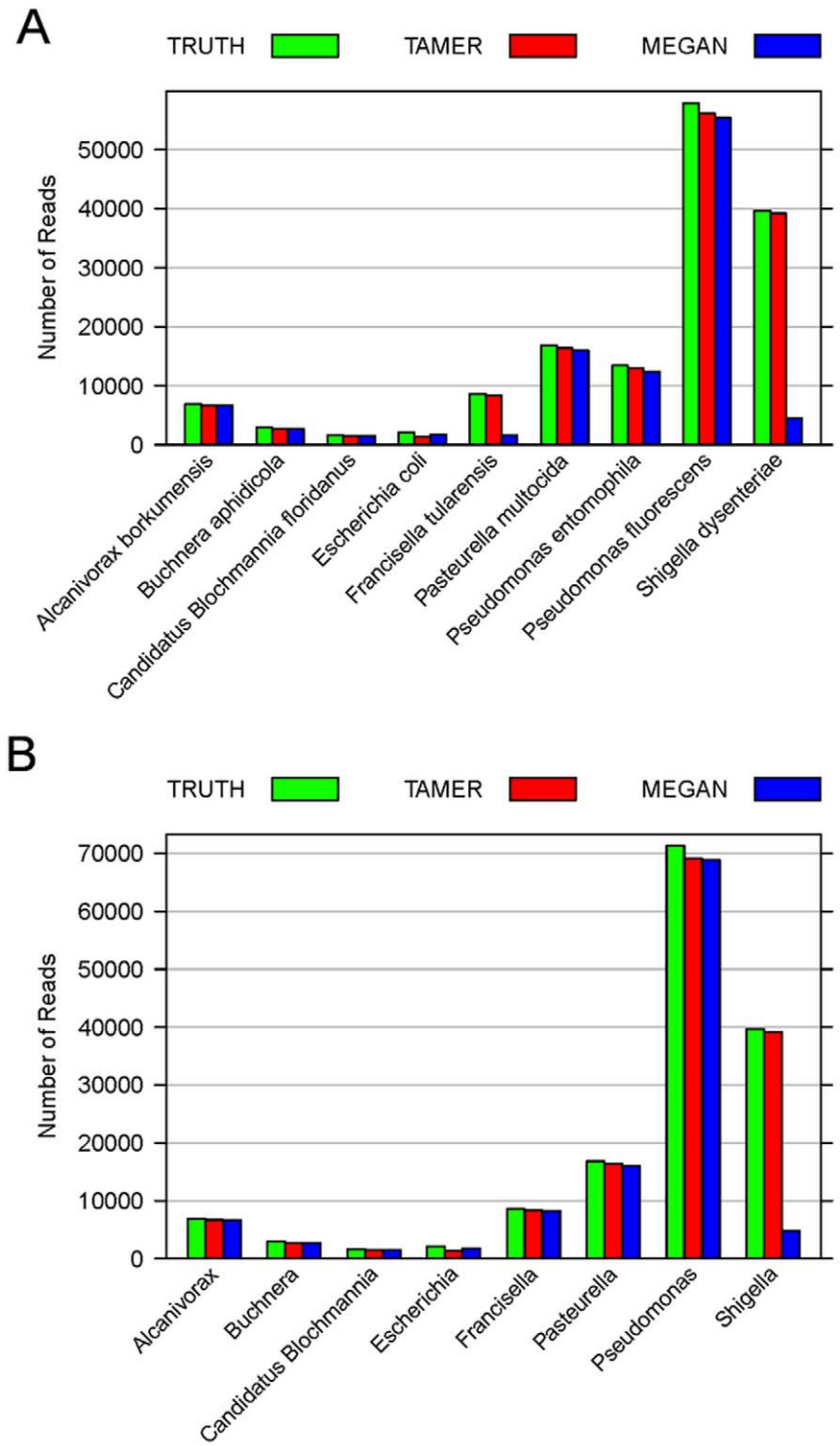

Figure 2. Reads assignment at rank Species and Genus for simMC dataset. Numbers of reads assigned to rank (A) Species and (B) Genus using TAMER and MEGAN are compared with the true values (TRUTH) for the simMC dataset with 150,000 reads and average read length of 100 bp in simulation study 1.
doi:10.1371/journal.pone.0046450.g002

disease samples. Our findings about these genera are also reported in a recent study [27] which hence further verified our results.

Seawater data. Using BLAST, about 97% of reads in sample 1 and 94% of reads in sample 2 have hits in the nt database and could be assigned to taxonomic ranks by TAMER. There are about 900 and 1,400 species detected in sample 1 and 2, respectively. TAMER assigns more reads than MEGAN and CARMA3 at different taxonomic ranks (Table 3). At rank Species, TAMER assigns about 50% more reads than MEGAN and about 90% more reads than CARMA3 for sample 1. *Candidatus*

Table 2. Results for CARMA3 evaluation dataset.

	TAMER		MEGAN		CARMA3	
	TP	FP	TP	FP	TP	FP
Species	99.24	0.73	81.45	0.02	4.57	0.12
Genus	99.26	0.68	91.52	0.03	64.10	0.43
Family	89.39	0.00	88.55	0.00	73.20	0.10
Order	97.22	0.00	96.40	0.00	83.48	0.12
Class	92.11	0.00	91.42	0.00	82.34	0.10
Phylum	99.94	0.00	99.31	0.00	90.50	0.07
Kingdom	99.97	0.00	99.42	0.00	90.90	0.12

The percentage of correctly (TP) and incorrectly (FP) assigned reads out of total 25,000 reads at different taxonomic ranks using TAMER, MEGAN and CARMA3 for the CARMA3 evaluation dataset in simulation study 2.
doi:10.1371/journal.pone.0046450.t002

Pelagibacter ubique is dominant in both samples (Figures S4). In fact this organism is highly dominant in both salt and fresh water worldwide [28]. At rank Genus, the differences among the number of assigned reads using different methods become smaller. However TAMER still assigns about 18% more reads than MEGAN and about 37% more reads than CARMA3 for sample 1. The two seawater samples are characterized as differing from each other based on relative frequency with sample 1 containing more *Shewanella* and *Burkholderia* than sample 2 (Figures S5), which is consistent with previous conclusions [7,26].

Discussion

The term metagenomics, first appeared in publication about 10 years ago [29]. To date, many metagenomic projects have undertaken characterization of microbiomes in samples from different environments including human gut [30], seawater [26], and soil [31], due to the next generation sequencing technologies. Therefore metagenomics has a broad impact across many diverse areas including human health, ecology, environmental remediation, and agriculture. Tens of millions of sequence reads can be obtained from sequencing one sample. An enormous challenge is attaining efficient and accurate data capture and storage coupled with computational and statistical methods to mine information from these massive datasets.

In this paper, we propose a rigorous statistical model to accurately identify and quantify genomes contained in a metagenomic sample by taking into account both sequence alignment scores and relative proportion of reads generated by the genomes. Identification of multiple genomes is an important goal in metagenomic studies. When a read is assigned to the high rank of the taxonomy tree, it is difficult to differentiate what genus or species actually are, or are not contained in the sample, as a high rank of taxonomy tree usually contains many genera and species. The proposed method, TAMER, can be applied to unassembled reads directly. The uniqueness of TAMER is that it assigns reads among the candidate genomes to which the sequence

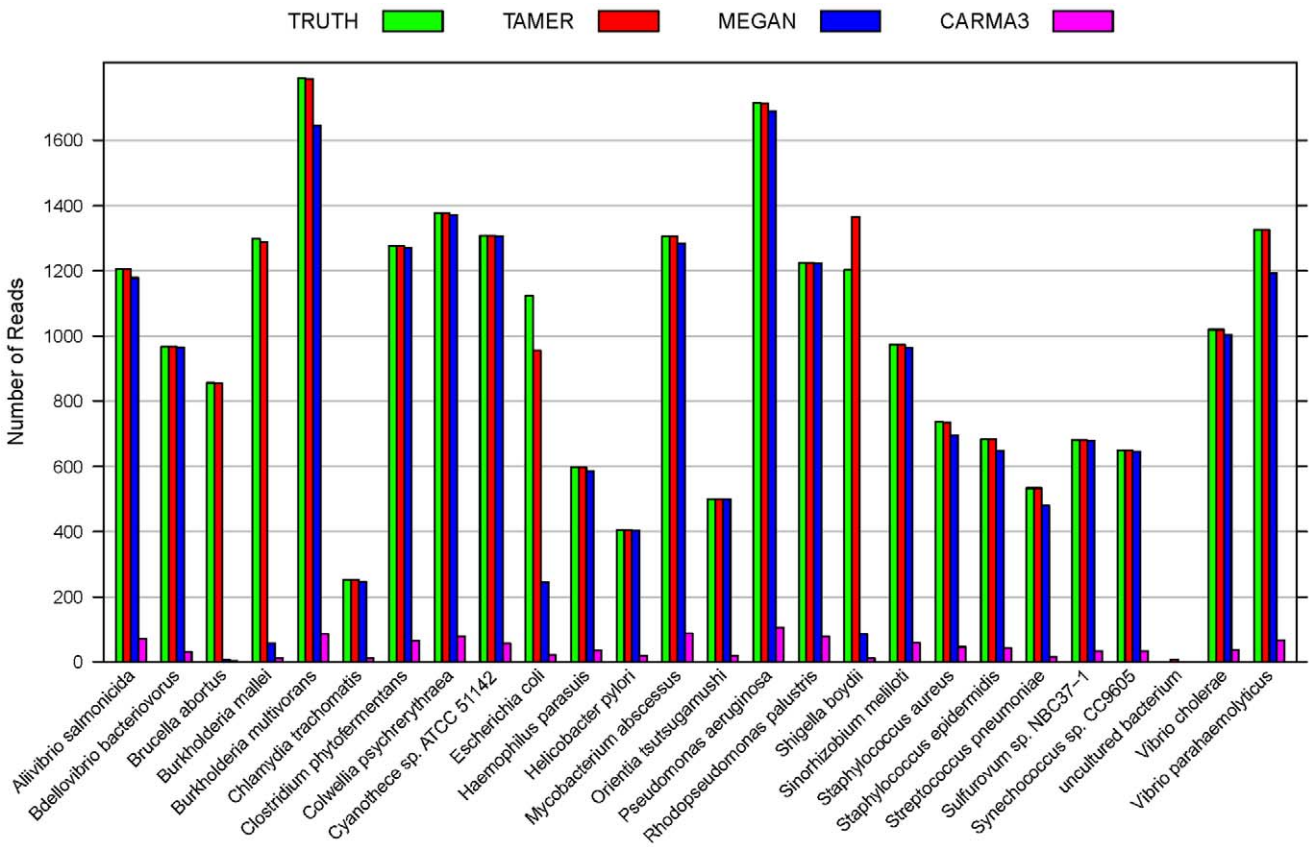

Figure 3. Reads assignment at rank Genus for CARMA3 dataset. Numbers of reads assigned to rank Genus using TAMER, MEGAN, and CARMA3 are compared with the true values (TRUTH) for the CARMA3 evaluation dataset in simulation study 2.
doi:10.1371/journal.pone.0046450.g003

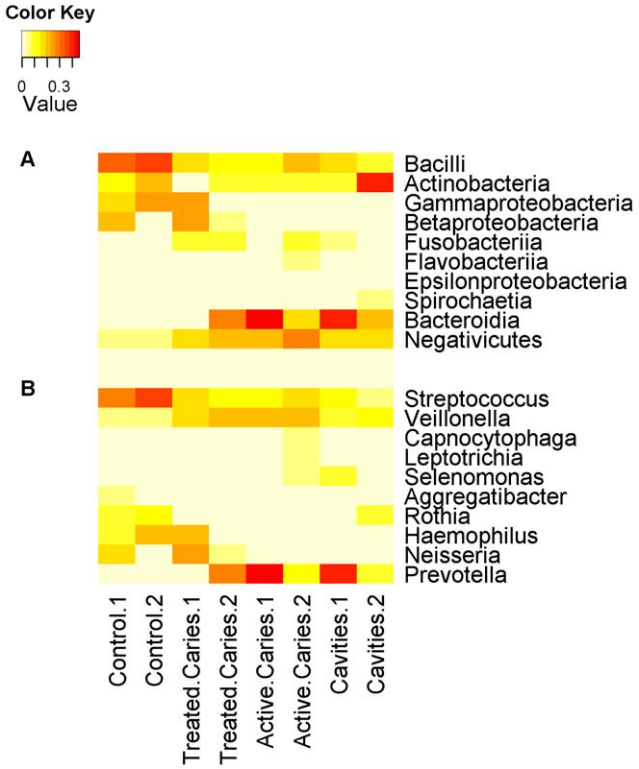

Figure 4. Heatmaps for oral samples. Heatmaps for the abundant (A) classes and (B) genera represent the estimated proportion of reads assigned to each of the eight samples based on TAMER.
doi:10.1371/journal.pone.0046450.g004

reads have hits. It does not assign the reads one by one to the taxonomy tree. On the contrary TAMER fully utilizes the information available from all reads by employing the mixture model. Roughly speaking, the taxonomic assignment of a read not only depends on its matching score against a genome but also borrows strength or information from other reads in the dataset. If a read achieved a high score in only one genome, then this read would be considered genome-specific and will be assigned to the corresponding genome. After assigning reads to multiple genomes, we sum up the assigned reads at different taxonomic ranks. Different from other mixture models [18,32], the sequencing error is considered and estimated in our proposed method. The comprehensive simulation studies demonstrate that TAMER is comparable with MEGAN at high taxonomic ranks, but TAMER assigns reads more accurately than MEGAN at Genus level or even Species level.

One limitation of the proposed method is that it is based on homology searches of the sequence reads in the reference databases. For reads generated from new genomes, they would not be included in the model if matches are not found. De novo assembly methods and deep sequencing are needed to discover new genomes.

As of future work, we can further assess the accuracy and uncertainty of the proportion of assigned reads along the taxonomy tree. The bootstrap method [33] by resampling the original sequence reads (i.e., sampling rows of the scoring matrix) with replacement can be used for the statistical inference. Subsequently, the parameters are estimated using the described EM algorithm for the bootstrap sample. By replicating this procedure, i.e., resampling and estimating a large number of times, (e.g., $B = 1000$ bootstraps), we are able to obtain the confidence interval for each parameter of interest. Since we construct the confidence intervals for the K parameters $R_i (i = 1,2,\cdots,K)$ simultaneously, a multiple correction method [34] such as Bonferroni correction can be used to guarantee a pre-specified (1-α)*100% family confidence level.

Table 3. Results for the two seawater samples.

	Sample 1			Sample 2		
	TAMER	MEGAN	CARMA3	TAMER	MEGAN	CARMA3
Species	93.92	45.82	3.92	86.92	35.36	0.54
Genus	91.77	70.98	60.20	81.84	34.02	15.63
Family	75.86	59.82	54.16	57.77	18.30	5.00
Order	91.62	77.12	57.88	80.69	41.74	8.44
Class	85.57	74.40	73.71	74.10	42.81	37.66
Phylum	92.84	81.36	83.92	83.53	50.12	53.65
Kingdom	96.50	87.32	90.27	93.86	68.85	73.87

The percentage of reads out of total 10,000 reads assigned at different taxonomic ranks using TAMER, MEGAN and CARMA3, for each of the two seawater samples.
doi:10.1371/journal.pone.0046450.t003

Supporting Information

Figure S1 Barplot of the number of assigned reads by TAMER and MEGAN at rank Species for simHC data. Numbers of reads assigned to rank Species using TAMER and MEGAN are compared with the true values (TRUTH) for the simHC data set of 150,000 reads with average read length of 100 bp.
(TIFF)

Figure S2 Barplot of the number of assigned reads by TAMER and MEGAN at rank Genus for simHC data. Numbers of reads assigned to rank Genus using TAMER and MEGAN are compared with the true values (TRUTH) for the simHC data set of 150,000 reads with average read length of 100 bp.
(TIFF)

Figure S3 Scatter plot of estimated proportions byTAMER and MEGAN at different taxonomic ranks for the oral data. Scatter plots of estimated abundance (proportion of reads) at different taxonomic ranks by MEGAN and TAMER for all eight samples.
(TIF)

Figure S4 Population distribution of sea water samples at rank Species. Proportions of reads assigned to the taxa at rank Species using TAMER, MEGAN and CARMA3 are compared for the sea water datasets.
(TIFF)

Figure S5 Population distribution of sea water samples at rank Genus. Proportions of reads assigned to the taxa at rank Genus using TAMER, MEGAN and CARMA3 are compared for the sea water datasets.
(TIFF)

Table S1 Characteristics of data sets for simulation study 1. Number of reads generated from each organism is listed for the simLC, simMC, simHC, and simSC datasets.
(XLS)

Table S2 Results for simulation study 1 with average read length of 400 bp. The percentage of correctly (TP) and incorrectly (FP) assigned reads out of total 10,000 reads with average read length of 400 bp at different taxonomic ranks using TAMER and MEGAN for simMC and simHC datasets.
(DOC)

Acknowledgments

The authors would like to thank Dr. Ingrid Glurich for editorial assistance and Fei Peng for computational assistance.

Author Contributions

Conceived and designed the experiments: HJ. Performed the experiments: HJ LA. Analyzed the data: HJ LA YQ. Contributed reagents/materials/analysis tools: SL GF. Wrote the paper: HJ.

References

1. Wooley J, Godzik A, Friedberg I (2010) A Primer on Metagenomics. PLoS Comput Biol 6.
2. Sanger F, Nicklen S, Coulson AR (1977) DNA Sequencing with Chain-Terminating Inhibitors. Proc Natl Acad Sci U S A 74: 5463–5467.
3. Mardis ER (2008) Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics 9: 387–402.
4. Ansorge WJ (2009) Next-generation DNA sequencing techniques. New Biotechnology 25: 195–203.
5. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389–3402.
6. Angly FE, Willner D, Prieto-Davo A, Edwards RA, Schmieder R, et al. (2009) The GAAS Metagenomic Tool and Its Estimations of Viral and Microbial Average Genome Size in Four Major Biomes. PLoS Comput Biol 5.
7. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res 17: 377–386.
8. Clemente JC, Jansson J, Valiente G (2010) Accurate taxonomic assignment of short pyrosequencing reads. Pac Symp Biocomput: 3–9.
9. Gori F, Folino G, Jetten MSM, Marchiori E (2011) MTR: taxonomic annotation of short metagenomic reads using clustering at multiple taxonomic ranks. Bioinformatics 27: 196–203.
10. Gerlach W, Stoye J (2011) Taxonomic classification of metagenomic shotgun sequences with CARMA3. Nucleic Acids Research 39: e91.
11. Meinicke P, Asshauer KP, Lingner T (2011) Mixture models for analysis of the taxonomic composition of metagenomes. Bioinformatics 27: 1618–1624.
12. Patil KR, Haider P, Pope PB, Turnbaugh PJ, Morrison M, et al. (2011) Taxonomic metagenome sequence assignment with structured output models. Nat Methods 8: 191–192.
13. Monzoorul Haque M, Ghosh TS, Komanduri D, Mande SS (2009) SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. Bioinformatics 25: 1722–1730.
14. Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Res 12: 656–664.
15. Sharma VK, Kumar N, Prakash T, Taylor TD (2012) Fast and Accurate Taxonomic Assignments of Metagenomic Sequences Using MetaBin. Plos One 7.
16. Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. Journal of Computational Biology 7: 203–214.
17. Dempster APea (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society Series B (Methodological) 39: 38.
18. Xia LC, Cram JA, Chen T, Fuhrman JA, Sun F (2011) Accurate genome relative abundance estimation based on shotgun metagenomic reads. Plos One 6: e27992.
19. Team RDC (2010) R: A Language and Environment for Statistical Computing: R Foundation for Statistical Computing Vienna Austria.
20. Richter DC, Ott F, Auch AF, Schmid R, Huson DH (2008) MetaSim-A Sequencing Simulator for Genomics and Metagenomics. Plos One 3.
21. Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, et al. (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nat Methods 4: 495–500.
22. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, et al. (2012) Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 40: D13–25.
23. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M (2007) CAMERA: A community resource for metagenomics. Plos Biology 5: 394–397.
24. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9.
25. Belda-Ferre P, Alcaraz LD, Cabrera-Rubio R, Romero H, Simon-Soro A, et al. (2011) The oral metagenome in health and disease. ISME J.
26. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66–74.
27. Liu B, Faller LL, Klitgord N, Mazumdar V, Ghodsi M, et al. (2012) Deep Sequencing of the Oral Microbiome Reveals Signatures of Periodontal Disease. Plos One 7.
28. Morris RM, Rappe MS, Connon SA, Vergin KL, Siebold WA, et al. (2002) SAR11 clade dominates ocean surface bacterioplankton communities. Nature 420: 806–810.
29. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol 5: R245–249.
30. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, et al. (2011) Enterotypes of the human gut microbiome. Nature 473: 174–180.
31. Brennerova MV, Josefiova J, Brenner V, Pieper DH, Junca H (2009) Metagenomics reveals diversity and abundance of meta-cleavage pathways in microbial communities from soil highly contaminated with jet fuel under air-sparging bioremediation. Environ Microbiol 11: 2216–2227.
32. Meinicke P, Asshauer KP, Lingner T (2011) Mixture models for analysis of the taxonomic composition of metagenomes. Bioinformatics 27: 1618–1624.
33. Miller RGJ (1981) Simultaneous Statistical Inference. New York: Springer.
34. Efron B (1982) The jackknife, the bootstrap, and other resampling plans.