Published in final edited form as:

J Appl Stat. 2018 ; 45(14): 2521–2535. doi:10.1080/02664763.2018.1426741.

Statistical correction for functional metagenomic profiling of a microbial community with short NGS reads

Ruofei Du¹ and Zhide Fang^{2,*}

¹Biostatistics Shared Resource, University of New Mexico Comprehensive Cancer Center, Albuquerque, USA

²Biostatistics Program, School of Public Health, Louisiana State University Health Sciences Center, New Orleans, USA

Abstract

By sequence homology search, the list of all the functions found and the counts of reads being aligned to them present the functional profile of a metagenomic sample. However, a significant obstacle has been observed in this approach due to the short read length associated with many next generation sequencing technologies. This includes artificial families, cross-annotations, length bias and conservation bias. The widely applied cutoff methods, such as BLAST E-value, are not able to solve the problems. Following the published successful procedures on the artificial families and the cross-annotation issue, we propose in this paper to use zero-truncated Poisson and Binomial (ZTP-Bin) hierarchical modelling to correct the length bias and the conservation bias. Goodness-of-fit of the modelling and cross-validation for the prediction using a bioinformatic simulated sample show the validity of this approach. Evaluated on an in vitro-simulated data set, the proposed modelling method outperforms other traditional methods. All three steps were then sequentially applied on real-life metagenomic samples to show that the proposed framework will lead to a more accurate functional profile of a short read metagenomic sample.

AMS 2010 SUBJECT CLASSIFICATIONS

Primary 62F10; secondary 62P10

Keywords

metagenomics; functional profiling; short reads; length bias; conservation bias

1. Introduction

Metagenomics is a recently developed, fast growing research area, in which researchers study the genomic diversity in a community of microbes by analysing the genomic sequences generated directly from environmental or clinical samples [13]. Along with the rapid advancement of DNA sequencing technology, metagenomics has opened up

*Corresponding author: zfang@lsuhsc.edu.

possibilities for scientists to investigate the complete composition of a complex microbial community, and to understand how the community members interact and influence each other to carry out the functions as a whole.

1.1. Read count based profiling of a next-generation sequencing metagenomic sample

In order to functionally profile a microbial community with metagenomic approach, a sample is collected from the site where the community reside *in situ*. Since it is culture-independent, this approach is able to include unculturable microorganisms into analysis that generally make up the larger proportion of the living organisms in the community [14, 22]. Genomic DNA is then extracted from the sample and subsequently sequenced. The next-generation sequencing (NGS) technologies are widely used at present. Different from the traditional Sanger sequencing, the cloning step is not needed in NGS technologies and thus the possible cloning-bias will not be introduced [16, 17].

A typical metagenomic functional analysis will compare the sequencing reads to a protein/domain sequence database (e.g. COG/KOG [23, 24], PFAM [2], or TIGRFAM [12]), to find the homologous sequence of a read and assign the read with the function to which its homolog is classified [35]. Together, the list of all the functions found and the counts of reads being assigned to them show the functional diversity of a community and the relative abundances among the functions. With this profiling approach, the research goal is to find the functional categories of coding sequences (CDS) in the extracted DNA and quantify the corresponding relative abundances so that we can reveal the potential functional competencies of the microbial community.

However, due to the limitation of many NGS technologies, the extracted DNA needs to be fragmented into short pieces prior to the sequencing procedure. Therefore, by the generated sequencing reads, we can only obtain a fragmented genomic picture of the microbial community. Figure 1 illustrates the extracted DNA and the sequencing reads generated by the sequencer. In the top panel, the colourful line-shaped rectangles represent the intact genomes obtained from a sample. For a species, its relative abundance in the community is signified by the number of the same coloured rectangles, as compared to the numbers of rectangles with different colours. Within a genome, each white pentagon indicates a CDS with the arrow showing forward or backward coding direction. The spaces between the white pentagons are non-coding regions. The DNA is randomly fragmented before the sequencing step, which results in a sequencing read being either from a CDS region, a non-coding region or their junction. The bottom panel in Figure 1 gives a picture of the NGS metagenomic reads. Here, the small rectangles, in different colours, represent the sequencing reads generated from the intact genomes. The pure white ones are the reads completely from a CDS region; the pure colourful ones are those completely from a non-coding region; the mixing white and colour ones come from the junctions. There may also be certain cases, in which a sequencing read is long enough to cover an entire CDS region. However, since the mean length of CDSs in prokaryotes was estimated 924 base pairs [34], for the considered NGS short reads those cases are negligible. All the sequencing reads are mixed together without any order; thus, we do not have direct knowledge about which genome and what location a read is from in a real sequencing sample.

1.2. False short homologues in profiling of a NGS metagenomic sample

From a protein/domain sequence database, we usually assume there is no homologue to a non-coding sequence (Non-CDS) with significant sequence similarity. However, homologues could be claimed when the short reads fragmented from a Non-CDS being aligned to the database. Panel *a* in Figure 2 provides an example about this scenario. The functional families in the database are denoted as Func. *a*, Func. *b*, Func. *c*, etc. Suppose there is a DNA Non-CDS being fragmented into eight short reads. By alignment, four reads are not functionally annotated (Reads 2, 5, 6 and 8); while the others are annotated (Read 1 to Func. *m*; Read 3 to Func. *n*; Read 4 to Func. *m* and Read 7 to Func. *e*). On the other hand, for the reads fragmented from a CDS which is associated to a specific function, the alignments may annotate some with the same function, some with different functions, and may also find no homologues to the others. Panel *b* in Figure 2 illustrates this situation. There are ten reads fragmented from a CDS which is associated to Func. *b*. Four of them are correctly annotated with Func. *b* (Reads 2, 4, 7 and 10); two of them are annotated with Func. *w* (Reads 6 and 8); one of them is annotated with Func. *p* (Read 1); but the remaining three reads are found no homologues (Reads 3, 5 and 9). Let the Non-CDS or the CDS from which a read is generated be referred as the parent sequence. If an aligned read is assigned to the same function to which its parent sequence associated, the aligned homolog is called a true short homolog (TSH), otherwise a false short homolog (FSH) [10]. There are two types of FSHs being observed: reads from Non-CDSs but being functionally annotated (e.g. Read 1, Read 3, Read 4, Read 7 in Panel *a* of Figure 2), and reads from CDSs but annotated with different functions (e.g. Read 1, Read 6, Read 8 in Panel *b* of Figure 2). For a read generated from the junction of a CDS and a Non-CDS, we think its functional identity follows that of either the CDS or the Non-CDS, depending on which region is covered more by the read. From the illustrations in Figure 2, we can see that in the alignment for the annotation, the similarity being compared is actually between a read and the regional part of the sequence to which the read is aligned. A read will be annotated with the function of the sequence whose regional part shares the highest similarity. That may or may not be the same function to which its parent sequence is associated.

The presence of FSHs causes two major problems in the read-count approach of profiling a microbial community: 1) artificial functional families, which are the functional families assigned by FSHs but do not exist in the metagenomic sample [10]; 2) cross-annotations, in which some reads should be counted to a functional family but instead are miscounted to other families [9]. Considering the mechanism by which the FSHs are produced, it is understandable that the FSHs cannot be excluded by setting a simple cut-off threshold after alignments (e.g., BLAST E-value [1]), which is however a valid approach for long sequencing reads (e.g. Sanger sequencing reads). It has been observed that by setting cutoff at BLAST E-value 10^{-3} or 10^{-5} to filter the alignment output of ~100bp NGS metagenomic reads, far fewer functional families were obtained compared to the functions shown in the alignment output of Sanger sequencing reads [8, 26, 29]. The stringent cutoffs would exclude both functions annotated by FSHs and TSHs. Increasing the average read length (from 100 bases to 400 bases) or improving the read coverage did not resolve the problem either [29].

Zhang et al. provided a simple formula to adjust the read counts in order to alleviate the problem of length bias and conservation bias (detailed below) in functional profiling of short read metagenomes [35] (hereinafter referred to as “Zhang’s method”). The formula can be expressed as,

$$\tilde{y}_i = \frac{y_i \cdot PPV_i}{TPR_i} \cdot LenF_i$$

where $\tilde{y}_i$ and y_i are the adjusted and observed numbers of sequence reads that are assigned to function family i , respectively, PPV_i and TPR_i are the estimated positive predictive value and true positive rate of the annotation for function i , with $LenF_i$ denoting the length factor. For example, if the average length of sequence reads is 100 by a NGS technology and the average length of family i is 500, $LenF_i$ is set equal to 0.2 (=100/500). The values of these parameters were learned from simulated data. However, in our view, the problem is rather more complicated than it can be handled by such straightforward formula. In addition, variations (or uncertainty) associated with read fragment and sequence alignment are not considered in this method.

1.3. Three steps to accurately functional profiling a NGS metagenomic sample

In two recent papers [9, 10], we have proposed the methods to address the problems of artificial functional families (Step 1) and cross-annotations (Step 2) shown in a ~100bp NGS metagenomic sample characterized by the COG (Cluster of Orthologous Genes) database.

After Steps 1 and 2, the read counts to the estimated true COGs are given. However, it is inappropriate to use those numbers of reads to directly measure the relative abundances among the COG families. First, there is a length bias observed, by which we describe the fact that a functional family associated with longer CDSs will be assigned a larger number of reads [35]. This can be understood easily, because usually the longer a CDS, the more reads we generate. Second, there exists conservation bias, by which we mean a functional family associated with more conserved CDS tends to be assigned more reads. The reason is that conserved sequences are very alike across species. By sequence similarity search, the reads fragmented from a CDS with higher conservation tend to have higher similarity scores when aligned to a homologous functional family [35]. We propose in this paper to use zero-truncated Poisson and Binomial hierarchical modelling to reduce these two biases, which we name Step 3.

In a summary, the Steps 1, 2 and 3 can be sequentially applied to functional profiling a ~100bp metagenomic sample by the COG database using the BLAST alignments. The combined steps also provide a heuristic approach for profiling a short-read metagenomic sample characterized with different databases.

2. Methods

2.1. Removal of artificial functional families and correction of cross-annotations

A brief summary about Steps 1 and 2 is given here, please refer to Du et al. (2013, 2015) [9, 10] for more details. Two sequence similarity search programs, BLAST and RPS-BLAST [1], are used separately to align a metagenomic sample of ~100bp to the database of COGs. The best-hit alignment is selected for the annotation of a read to a COG family. The read annotation is kept if the best-hit similarity score is greater than the suggested cutoff value (BLAST: 66). Then, the BLAST output is then compared with the RPS-BLAST output to detect the influential artificial COGs. We call these operations as Step 1, which results in a list of estimated true COGs in the filtered (by the score cutoff) BLAST output.

Subsequently, in Step 2, in order to correct the cross-annotations, the Probabilistic Latent Semantic Analysis (PLSA) modelling is used to estimate the proportions of these estimated true COGs. Suppose we now have N COGs, and $y_1, y_2, \dots, y_N$ denote the read counts being assigned to them, respectively. There are $(M-1)$ estimated true COGs with the M^{th} category denoting the Non-CDSs. Let α_{ij} be the conditional probability that a read is annotated to COG i under the condition that it is extracted from COG j ; β_j be the proportion of the reads being from COG j . PLSA modelling is to estimate α_{ij}, β_j for the following decomposition [18, 20]:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \cong \left(\sum_{i=1}^N y_i \right) \begin{pmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1M} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{N1} & \alpha_{N2} & \cdots & \alpha_{NM} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_M \end{pmatrix}.$$

Thus, with the estimated $\hat{\beta}_j$ the read count to the j^{th} family can be estimated by the integer closest to $(\sum_{i=1}^N y_i) \hat{\beta}_j, j=1, 2, \dots, M$.

2.2. Correction of read-count bias

As explained in §1.3, there are the length bias and the conservation bias associated with the read-count approach in functional metagenomics. In this subsection, we propose a zero-truncated Poisson and Binomial hierarchical model (ZTP-Bin) for predicting the number of CDSs in a functional family by using the reads assigned to it. The model is applied to fit a simulated sample; the estimated parameters are then used to correct the biases in functional profiling of a query metagenomic sample.

For a COG family, let U denote the number of reads fragmented from a CDS that is associated to it. We assume $U \sim \text{ZTP}(\lambda)$, that is,

$$P(U = u) = \frac{\lambda^u e^{-\lambda}}{(1 - e^{-\lambda})u!}, u \in \mathbb{N}^+,$$

where λ is the unknown parameter, representing the average count of reads fragmented from the CDS. The fact that the number of reads fragmented from a CDS is always a positive integer, is the motivation of using this zero-truncated Poisson model.

The steps to find the maximum likelihood estimate (MLE) of λ are listed as follows. Let c be the number of CDSs associated to the COG and assume that the U_i 's are independent and identically-distributed random variables with $U_i \sim ZTP(\lambda)$, $i=1,2,\dots,c$. Then the likelihood function is,

$$L = \prod_{i=1}^c P(U_i = u_i) = \prod_{i=1}^c \frac{\lambda^{u_i} e^{-\lambda}}{(1 - e^{-\lambda})^{u_i}} = \frac{e^{-c\lambda} \lambda^{\sum_{i=1}^c u_i}}{(1 - e^{-\lambda})^{\sum_{i=1}^c u_i}}.$$

By setting the first-order derivative, with respect to λ , of the log-likelihood function equal to zero, we have the following equation for the MLE of λ ,

$$l(\lambda) = \lambda - \bar{u}(1 - e^{-\lambda}) = 0,$$

where $\bar{u} = \sum_{i=1}^c u_i / c$. We then adopt the Newton-Raphson iteration method to find the solution. Since $dl/d\lambda = 1 - e^{-\lambda}$, the updated solution in the k^{th} iteration is

$$\lambda_k = \lambda_{k-1} - \frac{\lambda_{k-1} - \bar{u}(1 - e^{-\lambda_{k-1}})}{1 - e^{-\lambda_{k-1}}}.$$

The iteration process stops when the absolute difference between two consecutive solutions is less than 10^{-6} .

Now let Z be the read count being annotated to the COG, out of the U reads fragmented from the CDS, after the BLAST alignment output filtered by the score cutoff. We assume that $Z|U \sim \text{Bin}(p)$, that is,

$$P(Z = z | U = u) = \frac{u!}{z!(u-z)!} p^z (1-p)^{u-z}, z = 0, 1, 2, \dots, u.$$

Given the observations $(u_i, z_i)_{i=1}^c$, it is easy to show that the MLE of p is the ratio of $\sum_{i=1}^c z_i$ to $\sum_{i=1}^c u_i$.

The above steps show how we can obtain the MLEs of λ and p for a COG family from a learning data set. Next, we explain how to predicate the number of CDSs related to a COG family using the count of short reads that are annotated to the COG in a query metagenomic

sample. Note that in practice, Steps 1 and 2 described in §2.1 need to be performed first. For prediction, we obtain the marginal distribution, $P(Z = z)$, as follows,

$$\begin{aligned}
 P(Z = z) &= \sum_{u=1}^{\infty} P(Z = z \mid U = u)P(U = u) \\
 &= \sum_{u=z}^{\infty} \frac{u!}{z!(u-z)!} p^z (1-p)^{u-z} \frac{\lambda^u e^{-\lambda}}{(1-e^{-\lambda})u!} \\
 &= \frac{1}{(1-e^{-\lambda})} \frac{e^{-\lambda} \left(\frac{p}{1-p}\right)^z}{z!} \sum_{u=z}^{\infty} \frac{[\lambda(1-p)]^u}{(u-z)!} \\
 &= \frac{1}{(1-e^{-\lambda})} \frac{(\lambda p)^z e^{-\lambda p}}{z!}, z \in \mathbb{N}.
 \end{aligned}$$

The expectation of Z is

$$\begin{aligned}
 E(Z) &= \sum_{z=0}^{\infty} z \frac{1}{(1-e^{-\lambda})} \frac{(\lambda p)^z e^{-\lambda p}}{z!} \\
 &= \frac{1}{(1-e^{-\lambda})} \sum_{z=0}^{\infty} z \frac{(\lambda p)^z e^{-\lambda p}}{z!} \\
 &= \frac{\lambda p}{(1-e^{-\lambda})}.
 \end{aligned}$$

Define X as the sum of the Z -variables, $\sum_{i=1}^c Z_i$. Then the realization, x , of X is the read count that are annotated to the COG. By the Method of Moments, we have

$$E(Z) = \bar{z} \Rightarrow \frac{\lambda p}{1-e^{-\lambda}} = \frac{x}{c}.$$

With $\hat{\lambda}$ and $\hat{p}$, the MLEs obtained from a learning (simulated) sample, we can estimate c as

$$\hat{c} = \frac{x(1-e^{-\hat{\lambda}})}{\hat{\lambda}\hat{p}}.$$

This is the proposed estimate of the count of CDSs that are associated to the COG, after removing artificial families and correcting cross-annotations and read-count bias, and will be used for the construction of the functional profile of the metagenomic sample.

3. Results

3.1. ZTP-Bin model fitting on a bioinformatic simulated metagenomic sample

One hundred NCBI bacterial genome accession numbers were randomly selected. With 57 genomes being identified for plasmid DNA sequences and excluded, the reference genome sequences of the other 43 bacterial species were obtained from NCBI ftp site (<ftp://ftp.ncbi.nih.gov/genomes/refseq/bacteria>). The COG-annotation files, containing the locations of CDSs coding for the COG proteins, were also obtained. MetaSim [21], a bioinformatic simulator, was used to generate the sequencing reads of ~100bp. The information of the 43 genomes is available online at [https://github.com/rdu2017/ZTP-Bin/tree/master/data/Appendix_genomes_\(file:simu.docx\)](https://github.com/rdu2017/ZTP-Bin/tree/master/data/Appendix_genomes_(file:simu.docx)). A read is assumed to have the function of a COG if it overlaps a CDS associated with the COG by at least 60 bases [35]. This simulated metagenomic sample, referred to as *Simu*, was aligned by the BLAST against the COG database and underwent the Steps 1 and 2.

The estimates of parameters, $\hat{\lambda}$ and $\hat{p}$, are obtained by the ZTP-Bin model fitting algorithm as above; and the number of CDSs associated to each of detected COGs is further estimated (i.e. $\hat{c}$). The Pearson's chi-square test for goodness-of-fit of $\hat{c}$'s to c 's gives a p-value of 1. Figure 3 shows that the differences between the fitted values and the observed values, for the top 20 abundant COGs in the sample, are negligible. We will then use $\hat{\lambda}$ and $\hat{p}$ to correct the read-count biases for a future sample.

We further examine if a systematic bias exists when the proposed model is used for prediction. One hundred times of 2-fold cross-validations are preformed: the CDSs associated to a COG are randomly split into two sets with the same size. The ZTP-Bin model is fitted on one set; the learned parameters are then used to predict the number of CDSs in the other set. Figure 4 compares the predicted numbers vs. the true numbers of CDSs when the true numbers = 50. The Pearson correlation is 0.97. There are no systematic bias observed visually; the average values of the predicted numbers spread above or lower than the diagonal line randomly. Furthermore, there is no specific variation pattern being identified; for example, the variance of prediction does not become larger (or smaller) as the true number goes larger. The predictive performance of ZTP-Bin is evaluated and compared with the other methods on the *in vitro*-simulated dataset below.

3.2. Result from an in vitro-simulated metagenomic sample

The three *in vitro*-simulated 4X read-coverage data sets were combined as one simulated metagenomic sample [6], called MC_4X. A 454 GS20 sequence platform was used to generate the short reads with about 100bp from the sample. The BLAST program was then applied to align the sequencing reads to the database of COGs. Annotation of the short reads was according to the best-hit homologs, and a read was kept if the similarity score is greater than 66. The cross-annotations were corrected by the PLSA modelling; we then obtained the corrected read count to each of the COGs in the similarity score 66 above region. Last, the number of CDSs related to a COG was estimated by the algorithm given in Methods (§2.2).

Note that since the species in the sample were predefined, we were able to download the COG-annotation files of the metagenomes to identify the artificial COGs in the annotations,

and to get the number of CDSs to each of the COGs appearing in the sample. The true proportions of COGs in the sample are calculated. Figure 5 exhibits the proportions of CDSs related to the 20 most abundant truly existing COGs in the left panel, and compares the estimation errors of the complete functional profiles, calculated via four quantities [32] in the right panel. Those include Root relative means square error (RRMSE):

$$\sqrt{\frac{1}{L-1} \sum_{s=1}^{L-1} \left(\frac{\hat{\theta}_s - \theta_s}{\theta_s} \right)^2};$$

Average relative error (AVGRE):

$$\frac{1}{L-1} \sum_{s=1}^{L-1} \frac{|\hat{\theta}_s - \theta_s|}{\theta_s};$$

Maximum relative error (MAXRE):

$$\max_s \left\{ \frac{|\hat{\theta}_s - \theta_s|}{\theta_s} \right\};$$

and Total variation distance (DTV):

$$\frac{1}{2} \sum_{s=1}^{L-1} |\hat{\theta}_s - \theta_s|,$$

where θ_s denotes the true proportion of the CDSs to a COG family, with $\hat{\theta}_s$ being its estimate, and L is the count of the COGs truly existed in the sample. It is clear from Figure 5 that our method provides more accurate (lower error) estimations of the proportions of CDSs associated to COGs than those proposed in Zhang et al. [35] or based on cutoffs at BLAST E-values 1e-3, 1e-5. The mechanism of the DNA fragmentation and the occurrence of the sequencing errors are expected not exactly the same between a bioinformatic simulation and a real sequencing process. The comparison also shows that our method is robust against these factors.

3.3. Results from real metagenomic samples

At the Hawaii Ocean Time Series (HOT) station, metagenomic samples were collected at 25m-depth, 75m-depth and 500m-depth seawater separately. The sequencing reads for these samples were all generated on a 454 GS20 platform with read length of ~100 bases [7, 11]. We name the three data sets HOT25, HOT75 and HOT500, respectively. Since we do not know which COG families exist in a real query dataset, we are not able to examine the accuracy of an estimated profile as compared to the true one. However, for each of three HOT datasets, we do observe a large discrepancy of the estimation between different methods, for the 20 most abundant COGs and for the complete functional profile as well (figure not shown).

We use three COG families, COG1028, COG3349 and COG0477, as examples to demonstrate that the proposed method performs well in that the resulted profiles are consistent with the findings in literature. These COG families along with their proportions and ranks of abundances by the four approaches are presented in Table 1, respectively. Our method lists COG1028, a member of the COG category I (controlling lipid transport and metabolism), as the No. 1 (the most abundant) family consensually in three datasets. This is in accord with the conclusions from previous studies [5]. But different ranks are given by the other methods. The Conserved Protein Domain family, COG3349, includes NAD-binding domain and a Fe-S cluster. A recent study revealed that it is enriched on the surface level of the South China Sea (10m), but not found in the 100m, 100m and 300m seawater [25]. The finding agrees with our analysis result in that COG3349 is found to be the 4th and the 3rd abundant family in HOT25 and HOT75 samples respectively, but disappeared in the HOT500 sample. Zhang's method missed this COG family consistently in the estimated profiles. For COG0477, the metagenomics studies in Mou [19] show that this family has high abundance in the samples of short sequence reads (average length of 97bp), however, interestingly the PCR-based 16SrDNA analysis on the same samples reveals that no genes from the detected genomes are related to COG0477. From the simulations we performed, it's almost sure that COG0477 is an artifact due to the alignment using the short reads. As shown in Table 1, both the E-value cutoff methods were not able to rule it out.

4. Summary and discussion

In this paper, we continue to draw attention about problems emerged when functionally profiling a metagenomic sample by the read-count approach using reads of short length, including artificial families, cross-annotations, length bias and conservation bias. How these problems are introduced during the sequencing and alignment are explained. Following the preceding successful procedures (Steps 1 and 2) for the first two problems, we propose to use ZTP-Bin modelling to correct the length bias and the conservation bias. Goodness-of-fit of the modelling and cross-validation for the prediction using a bioinformatic simulated sample show the validity of this approach. Evaluated on the *in vitro*-simulated dataset MC_4X, the approach performs superiorly comparing to Zhang's method and E-value cutoff based methods in terms of the accuracy of estimations of proportions of CDSs to COGs in the sample. Applying these three steps sequentially on the real-life metagenomic samples, we obtain functional profiles in which many estimates are consistent with the conclusions in literature.

The proposed approach works with the raw reads directly. There might be an argument that assembling metagenomic short reads into contigs may overcome the problems discussed above. However, such assembly is very difficult owing to the fact that for metagenomic studies the genetic materials are extracted from multiple organisms. The study in [4] has showed that we may have poor assembled results if there are many closely related strains in samples. In addition, there appears a considerable proportion of chimeric contigs after read assembly [27]. In our opinion, to handle this phenomenon is even more challenging. Analysis directly with raw sequencing reads is one of the methods that have been applied to study microbial communities [27, 30–32].

One may argue that, as the DNA sequencing technologies are advancing very fast, the problems due to the short read length do not bother us any longer. However, there are still metagenomics experiments currently performed using the ‘old’ sequencers that produce short reads [3, 33]; secondly, it is of scientific importance to analyse the ‘old’ sequencing data with improved methods to draw more biologically meaningful conclusions [15, 28]. There are very large volume of such short reads generated in last decade. Thirdly, as showed by our simulations for reads of ~200 to ~400 bases, which is the read length region of many currently used NGS sequencers, the problems being demonstrated still remain although much reduced. Our method can serve as a good heuristic tool for these situations.

The proposed method has been implemented with R functions to facilitate statisticians to utilize the method for analysing short read metagenomic data. The R functions and examples can be found online at <https://github.com/rdu2017/ZTP-Bin>. One may take advantage of this availability to apply the method not limited to a specific NGS technology or a functional database.

Acknowledgments

We are grateful for the helpful suggestions of the Editor and the reviewers. ZF’s research was supported by 2 U54 GM104940-02 from the National Institute of General Medical Sciences of the National Institutes of Health which funds the Louisiana Clinical and Translational Science Center of Pennington Biomedical Research Center.

References

1. Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*. 1997; 25:3389–3402. [PubMed: 9254694]
2. Bateman A, et al. The Pfam protein families database. *Nucleic acids research*. 2004; 32:D138–D141. [PubMed: 14681378]
3. Bell KL, et al. Review and future prospects for DNA barcoding methods in forensic palynology. *Forensic Science International: Genetics*. 2016; 21:110–116. [PubMed: 26751251]
4. Charuvaka A, Rangwala H. Evaluation of short read metagenomic assembly. *BMC genomics*. 2011; 12:S8.
5. Cottrell MT, Kirchman DL. Virus genes in Arctic marine bacteria identified by metagenomic analysis. *Aquatic Microbial Ecology*. 2012; 66:107–116.
6. Dalevi D, et al. Annotation of metagenome short reads using proxygenes. *Bioinformatics*. 2008; 24:i7–i13. [PubMed: 18689842]
7. DeLong EF, et al. Community genomics among stratified microbial assemblages in the ocean’s interior. *Science*. 2006; 311:496–503. [PubMed: 16439655]
8. Dinsdale EA, et al. Functional metagenomic profiling of nine biomes. *Nature*. 2008; 452:629–632. [PubMed: 18337718]
9. Du R, et al. A Statistical Approach to Correcting Cross-Annotations in a Metagenomic Functional Profile Generated by Short Reads. *Journal of Biometrics & Biostatistics*. 2015; 2014
10. Du R, Mercante D, Fang Z. An artificial functional family filter in homolog searching in next-generation sequencing metagenomics. *PloS one*. 2013; 8:e58669. [PubMed: 23516532]
11. Frias-Lopez J, et al. Microbial community gene expression in ocean surface waters. *Proceedings of the National Academy of Sciences*. 2008; 105:3805–3810.
12. Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. *Nucleic acids research*. 2003; 31:371–373. [PubMed: 12520025]
13. Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and molecular biology reviews*. 2004; 68:669–685. [PubMed: 15590779]

14. Handelsman J, et al. The new science of metagenomics: revealing the secrets of our microbial planet. *Nat Res Council Report*. 2007; 13:47–84.
15. Knudsen LRV, et al. Revisiting bovine pyometra—New insights into the disease using a culture-independent deep sequencing approach. *Veterinary microbiology*. 2015; 175:319–324. [PubMed: 25550285]
16. Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends in genetics*. 2008; 24:133–141. [PubMed: 18262675]
17. Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*. 2008; 9:387–402. [PubMed: 18576944]
18. Monay F, Gatica-Perez D. PLSA-based image auto-annotation: constraining the latent space. *Proceedings of the 12th annual ACM international conference on Multimedia*; ACM; 2004. 348–351.
19. Mou X. Culture-independent characterization of DOC-transforming bacterioplankton in coastal seawater. *uga*; 2006.
20. Mrva D, Woodland PC. INTERSPEECH. Citeseer; 2004. A PLSA-based language model for conversational telephone speech.
21. Richter DC, et al. MetaSim—a sequencing simulator for genomics and metagenomics. *PloS one*. 2008; 3:e3373. [PubMed: 18841204]
22. Schloss PD, Handelsman J. Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome biology*. 2005; 6:1.
23. Tatusov RL, et al. The COG database: an updated version includes eukaryotes. *BMC bioinformatics*. 2003; 4:1. [PubMed: 12513700]
24. Tatusov RL, et al. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic acids research*. 2000; 28:33–36. [PubMed: 10592175]
25. Tseng CH, et al. Prokaryotic assemblages and metagenomes in pelagic zones of the South China Sea. *BMC genomics*. 2015; 16:219. [PubMed: 25879764]
26. Turnbaugh PJ, et al. An obesity-associated gut microbiome with increased capacity for energy harvest. *nature*. 2006; 444:1027–131. [PubMed: 17183312]
27. Vázquez-Castellanos JF, et al. Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. *BMC genomics*. 2014; 15:37. [PubMed: 24438450]
28. Weng FC, et al. Reanalyze unassigned reads in Sanger based metagenomic data using conserved gene adjacency. *BMC bioinformatics*. 2010; 11:565. [PubMed: 21083935]
29. Wommack KE, Bhavsar J, Ravel J. Metagenomics: read length matters. *Applied and environmental microbiology*. 2008; 74:1453–1463. [PubMed: 18192407]
30. Wooley JC, Ye Y. Metagenomics: facts and artifacts, and computational challenges. *Journal of computer science and technology*. 2010; 25:71–81.
31. Wu J, et al. Integrated metagenomic and metatranscriptomic analyses of microbial communities in the meso-and bathypelagic realm of North Pacific Ocean. *Marine drugs*. 2013; 11:3777–3801. [PubMed: 24152557]
32. Xia LC, et al. Accurate genome relative abundance estimation based on shotgun metagenomic reads. *PloS one*. 2011; 6:e27992. [PubMed: 22162995]
33. Xie G, et al. Genome Sequence of *Porphyromonas gingivalis* Strain A7A1-28. *Genome Announcements*. 2017; 5:e00021–17. [PubMed: 28280013]
34. Xu L, et al. Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. *Molecular biology and evolution*. 2006; 23:1107–1108. [PubMed: 16611645]
35. Zhang Q, Doak TG, Ye Y. Artificial functional difference between microbial communities caused by length difference of sequencing reads. *Pac Symp Biocomput*. 2012:259–270. [PubMed: 22174281]

Figure 1.

The extracted intact metagenomic DNA and the sequencing reads from these DNA. In the top panel, the colorful line-shaped rectangles are different genomes; and the white pentagons in each genome indicate the CDSs. In the bottom panel, the small rectangles indicate the sequencing reads.

Figure 2.

The alignments of fragmented reads. Panel *a.*, the alignments of reads fragmented from a Non-CDS: FSHs (aligned homologues for Reads 1, 3, 4, and 7). Panel *b.*, the alignments of reads fragmented from a CDS: TSHs (aligned homologues for Reads 2, 4, 7 and 10), FSHs (aligned homologues for Reads 1, 6 and 8).

Figure 3.
The observed and corresponding fitted numbers of CDSs to each of the top 20 abundant COGs in the *Simu* sample.

Figure 4.

Plot of predicted numbers, by the 2-fold cross-validation, versus the true numbers of CDSs using the *Simu* dataset, when the true numbers of CDSs to a COG are greater than 50.

Figure 5. Comparison of four approaches for estimating the proportions of CDSs associated to COG functions in MC_4X. Left: the proportions of CDSs to the 20 most abundant COGs, estimated by our method, Zhang's method, the methods based on E-value cutoffs 1e-3 and 1e-5; Right: estimation errors of four methods: root relative means square error (RRMSE), average relative error (AVGRE), maximum relative error (MAXRE), and total variation distance (DTV).

Table 1

The proportions and ranks of abundances of three COGs in HOT samples by four approaches, respectively.

COG	Dataset	Our method		Zhang's		E-value 1e-3		E-value 1e5	
		Prop.	Rank	Prop.	Rank	Prop.	Rank	Prop.	Rank
COG1028 (Dehydrogenases with different specificities)	HOT25	0.0078	1	0.01	3	0.0059	3	0.0049	13
	HOT75	0.0078	1	0.01	4	0.0057	4	0.0045	15
	HOT500	0.0117	1	0.0168	2	0.0096	1	0.0075	4
COG3349 (Uncharacterized conserved protein)	HOT25	0.0052	4	-	-	0.0021	82	0.0026	63
	HOT75	0.0068	3	-	-	0.0028	40	0.0033	36
	HOT500	-	-	-	-	<.0001	2111	-	-
COG0477 (Permeases of the major facilitator superfamily)	HOT25	-	-	-	-	0.0015	159	0.0011	273
	HOT75	-	-	-	-	0.0016	147	0.0012	256
	HOT500	-	-	-	-	0.0018	107	0.0011	252

Note: Prop. is short for Proportion; - indicates the COG was not detected by the method.