ARTIFICIAL FUNCTIONAL DIFFERENCE BETWEEN MICROBIAL COMMUNITIES CAUSED BY LENGTH DIFFERENCE OF SEQUENCING READS[*]

QUAN ZHANG
School of Informatics and Computing, Indiana University
150 S. Woodlawn Ave, Bloomington, IN 47405, USA
Email: quzhang@indiana.edu

THOMAS G. DOAK
Biology Department, Indiana University
1001 E. 3rd Street, Bloomington, IN 47405, USA
Email: tdoak@indiana.edu

YUZHEN YE
School of Informatics and Computing, Indiana University
150 S. Woodlawn Ave, Bloomington, IN 47405, USA
Email: yye@indiana.edu

Homology-based approaches are often used for the annotation of microbial communities, providing functional profiles that are used to characterize and compare the content and the functionality of microbial communities. Metagenomic reads are the starting data for these studies, however considerable differences are observed between the functional profiles—built from sequencing reads produced by different sequencing techniques—for even the same microbial community. Using simulation experiments, we show that such functional differences are likely to be caused by the actual difference in read lengths, and are not the results of a sampling bias of the sequencing techniques. Furthermore, the functional differences derived from different sequencing techniques cannot be fully explained by the *read-count bias*, *i.e.* 1) the higher fraction of unannotated shorter reads (*i.e.*, "read length matters"), and 2) the different lengths of proteins in different functional categories. Instead, we show here that specific functional categories are under-annotated, because similarity-search-based functional annotation tools tend to miss more reads from functional categories that contain less conserved genes/proteins. In addition, the accuracy of functional annotation of short reads for different functions varies, further skewing the functional profiles. To address these issues, we present a simple yet efficient method to improve the frequency estimates of different functional categories in the functional profiles of metagenomes, based on the functional annotation of simulated reads from complete microbial genomes.

[*] This work is supported by NSF grant DBI-0845685 and NIH grant 1R01HG004908.

1. Introduction

By enabling the direct analysis of microbial communities (containing many unculturable and often unknown microbes), metagenomics is revolutionizing the field of microbiology, and has excited researchers in the many disciplines that benefit from the study of environmental microbes, including those in ecology, environmental sciences, and biomedicine [1-6]. Functional analysis of a metagenomic dataset provides valuable insights into the functionality of the corresponding environmental microbial community. Also, it is very interesting to compare microbial communities in terms of their functionalities. One way of comparing microbial communities is to build a *functional profile* for each community and then compare the profiles across multiple communities, known as *functional metagenomics* [7, 8]. To predict the function of sequencing reads derived in metagenomics studies, homology-based methods are generally used. Several fast and efficient similarity search tools, such as BLAST [9] and HMMER [10], were introduced in the early 90's, and are still used for this kind of analyses, to classify metagenomic sequences into functional categories, such as COG protein families [11, 12] and KEGG pathways [13]. Subsequently, the "read-count" approach is often used to derive functional profiles of metagenomic datasets, simply equating the frequency of a functional category to the proportion of reads that have this particular functional annotation (*e.g.*, a COG family, or a K number as used in KEGG pathways). However, the reads-count approach has been shown to have a "read-count bias": annotations of short reads will be missed [14], and the reads-count approach favors functional categories (*e.g.* the protein families) composed of longer genes. As a result, the "reads-count" approach sometimes overestimates the frequency of functional categories containing longer genes and underestimates the frequency of functional categories containing shorter genes [15]. It is one of the various kinds of artifacts present in metagenomics data, caused by the limitations of the experimental protocols and/or inadequate data analysis procedures, which can lead to incorrect conclusions about microbial communities [16, 17].

An example of these artifacts is found in the comparison of the distal gut microbiota of genetically obese mice and their lean littermates. Obesity is associated with changes in the relative abundance of Bacteroidetes and Firmicutes species, and Gordon and colleges [3] demonstrated through metagenomic and biochemical analyses that this affected the metabolic potential of the mouse gut microbiota. Interestingly, they sequenced samples from a lean and an obese mouse using two different sequencing techniques—Sanger sequencing and pyrosequencing (which produced reads of ~100 nt at the time, equivalent to current Illumina read lengths; current 454 sequences can produce much longer reads)—and compared the functional profiles of the resulting datasets (Figure 1b in [3]). Their results indicated that different metagenomic datasets sequenced with the same technique share higher functional similarity than do metagenomic datasets sequenced from the same biological sample but using different sequencing techniques. The authors proposed the hypothesis that different sequencing techniques may favor different functional categories, which, if true, would pose a significant challenge for the shotgun next-generation sequencing of metagenomes adopted by most current metagnomics projects.

We used simulation studies to explore the possible explanations for the observed functional differences between annotations of the same microbial communities built from different

sequencing techniques. In our first study, we simulated short reads (of ~100 nt) from the Sanger reads for the mice microbiomes [3]. We then applied the same functional annotation pipeline used for annotating the Sanger reads to the annotation of the simulated short reads, to examine and compare the functional profile of the simulated dataset with that of the original Sanger dataset. In the second study, we simulated short reads from 7 complete genomes (so we know what functions we expect in the short reads) to compare the real annotations of short reads with the expectation. Our simulation studies revealed that there are two confounding factors that bias the functional profiles of a metagenome: 1) an unequal fraction of reads tend to be missed in functional annotation for different functional categories (*i.e.* protein families); and 2) different functional categories tend to have different proportions of wrong annotations. We report below the details of our simulation studies and the results. Based on these results, we present a simple yet efficient method to improve the frequency estimates of different functional categories in metagenomes, utilizing the functional annotations of simulated short reads from complete genomes.

2. Methods

2.1. *Datasets*

2.1.1. *Obese- and lean-mouse metagenomic sequences*

The Sanger and pyrosequences (of ~100 nt) of obese mice and lean mice [3] were obtained from the MG-RAST server (http://metagenomics.anl.gov/).

2.1.2. *Complete genomes*

Table 1. A list of complete genomes for short reads simulation.

Genome id	Species name
NC_002737	*Streptococcus pyogenes* M1 GAS
NC_002927	*Bordetella bronchiseptica* RB50
NC_002937	*Desulfovibrio vulgaris* subsp. vulgaris str. Hildenborough
NC_003902	*Xanthomonas campestris* pv. campestris str. ATCC 33913
NC_006905	*Salmonella enterica* subsp. enterica serovar Choleraesuis str. SC-B67
NC_006932	*Brucella abortus* biovar 1 str. 9-941 chromosome I
NC_007795	*Staphylococcus aureus* subsp. aureus NCTC 8325

We selected 7 genomes (see Table 1) to simulate short reads from, to estimate the annotation accuracy of different protein families. The genomes were downloaded from the NCBI ftp site and the NCBI annotations of the protein-coding genes were used in this study.

2.1.3. *Simulated short reads*

1. Two datasets of short reads of ~100 nt (denoted as "Short-simu", short for "short simulated reads") were simulated from the lean-mouse and obese-mouse Sanger reads, respectively. See details of the simulated datasets in Table 2.

Table 2. Statistics of the real and simulated metagenomic datasets.

Reads type	Lean mouse		Obese mouse	
	Average length of reads (bp)	Total number of reads	Average length of reads (bp)	Total number of reads
Sanger	780	11,321	766	11,381
Short	108	1,045,701	110	675,880
Short-simu	107	1,000,000	110	670,000

"Sanger" is for Sanger sequences; "Short" for pyrosequences; and "Short-simu" for the short reads simulated from the Sanger dataset.

2. Short reads of ~100 nt were simulated from the 7 selected complete genomes (at 10X coverage) using MetaSim [18] (Version 0.9.1). For simplicity, no sequencing errors were introduced in the simulation.

2.2. *Functional Profiling of Metagenomes*

Here we used COG categories to create functional profiles for metagenomes: the extended COG definitions [11, 12] consist of 4873 COG families, which can be grouped into 25 broad functional categories. The functional profiles of metagenomes can then be built at the COG family level (*i.e.* a functional profile is represented by a vector of 4873 dimensions), or at the level of broad functional categories (*i.e.* a functional profile is represented by a vector of 25 dimensions). To compute the functional profile of a metagenomic dataset using the broad functional categories, the frequency of a category is estimated by the total reads that can be assigned to the families belonging to that category, normalized by the total number of sequencing reads in the dataset.

2.3. *Functional Annotation*

We used similarity-search-based methods for the functional annotation of long reads, or the simulated short reads. We tested both BLAST and HMMER searches (HMMER3) [19]. For BLAST searches, we used an E-value cutoff of 10^{-3} (a typical threshold used in functional annotations of metagenomes [3, 20]), and for HMMER-search-based annotation (by hmmscan from the HMMER3 package), we used an E-value cutoff of 10^{-2}. For both methods, we retained the best hit from non-overlapping regions, so that if a gene (or a read) contains multiple domains each with a distinct function, all the functions will be reported.

For HMMER searches, we built Hidden Markov Models (HMMs) for all COG families, using the sequences annotated in the eggNOG database (version 1.0, http://eggnog.embl.de/) [21]. MUSCLE [22] was used to generate a multiple alignment for each protein family, and the HMM builder from the HMMER3 package was then applied to build a HMM for each COG.

2.4. *"Perfect" Annotation of Simulated Short Reads*

For simulated short reads, we "assign" their functional annotations based on the functional annotation of the longer reads, or the complete genes from which the short reads were sampled

(*i.e.* the *parent* sequence). Considering that similarity searches will miss annotations of extremely short reads, we only transfer functions of the parent sequence to the simulated reads if the simulated reads and the parent sequence overlap by at least 60 nt (*i.e.* encoding 20 aa). In this way, the simulated short reads "inherit" functions from their parent long reads or complete genes, achieving 'perfect' annotations, which could not in reality be achieved by similarity-search-based methods.

3. Results

3.1. *Read Length Differences can Cause Artificial Functional Differences between Metagenomes without Real Functional Differences*

The first question we asked is *if* different read lengths can cause artificial functional differences between metagenomes. To address this question, we simulated short reads (of ~100 nt) from actual experimental Sanger reads (long reads): we do not expect to observe real biological difference between the simulated short-reads dataset and the original Sanger dataset (referred as the *parent* datasets). We then built functional profiles (based on the broad COG categories) based on HMMER searches against COG proteins, for the simulated datasets and for the parent Sanger datasets for further comparison.

We show that if the functional annotations were perfect (by simply transferring the function annotations from the Sanger reads to simulated short reads), there would be no significant difference between the functional profiles of the simulated dataset and its parent dataset, although small differences are expected due to the randomness of read sampling (see Figure 1A). However, when we applied actual annotation to the simulated reads, we observed a significant difference between the functional profile of the simulated short-read dataset and that of the parent Sanger dataset (Figure 1B), similar to the difference observed between the profiles of the real pyrosequence dataset and the Sanger dataset (see Figure 3B below). As a result, the profiles of the datasets appear to be more similar if the reads are of similar length, regardless of the sample resources of the datasets (*e.g.*, obese versus lean mice), similar to what had been reported in [3].

To examine the effect of different similarity search tools on building the functional profiles, we also tried BLAST searches (similar methods based on BLAST searches were used for the annotation in [3]), and observed similar results: the functional profiles of the datasets appear to be more similar if the reads are of similar length, regardless of the sample resources of the datasets (data not shown). Also, correcting for the protein lengths of different families did not reduce the differences (data not shown). Taking advantage of the fast speed of HMMER3 and the high sensitivity and significance evaluation provided, we used HMMER searches as the functional annotation method in our subsequent studies.

In summary, our simulation experiments indicated that the observed functional difference reported in [3] could be simply caused by the length difference of the reads, even when there was no real biological difference involved. We show that differences in read lengths (not sequencing method) result in biased functional profiles of metagenomic samples, if the length difference is not considered properly.

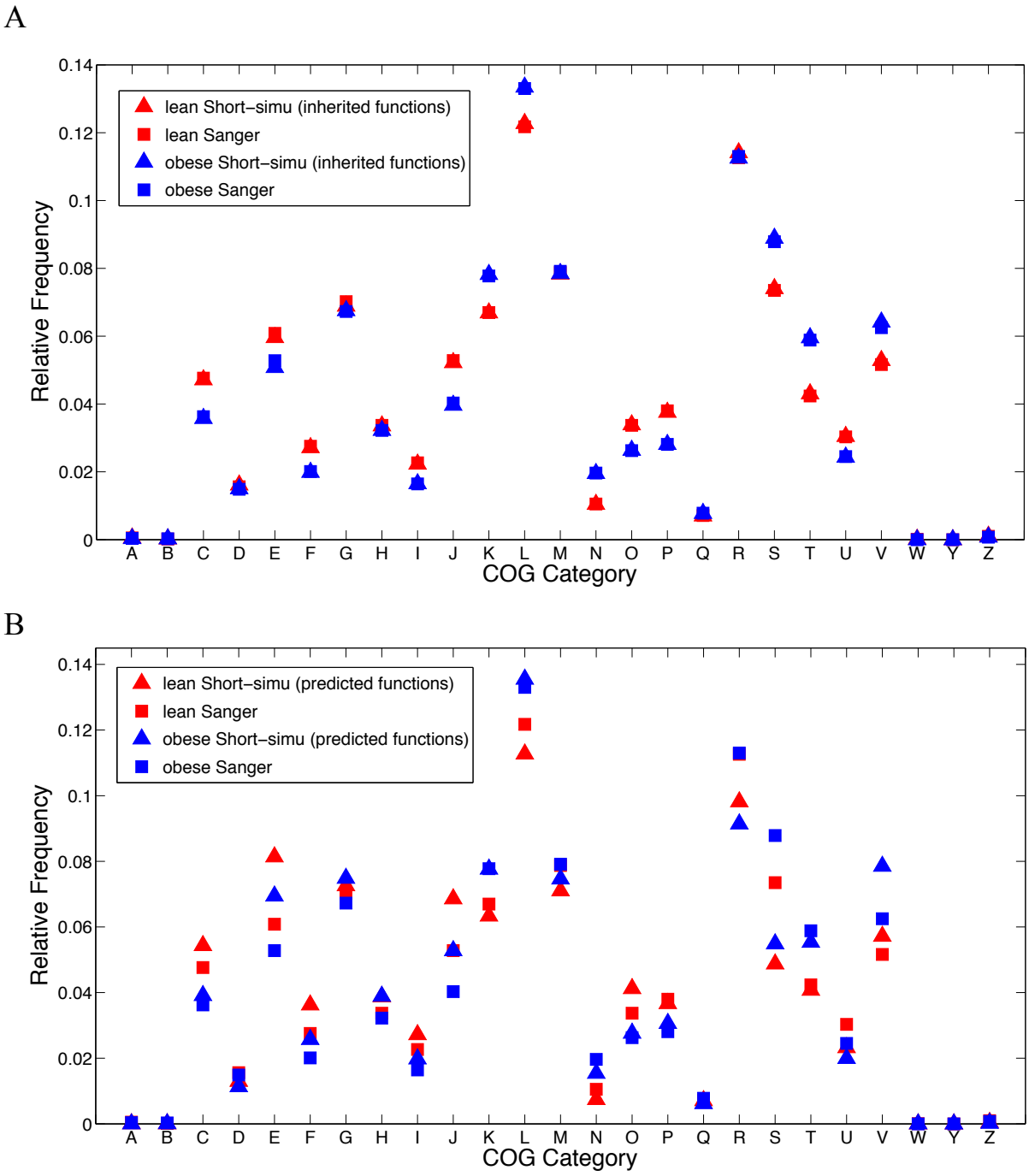


Fig. 1. Comparison of functional profiles of lean- (shown in red) and obese-mouse fecal microbiomes (shown in blue) built from Sanger reads (shown in squares) and simulated short reads (triangles), (A) providing 'perfect' annotation for the short reads, and (B) applying real annotations for the short reads based on HMMER search results.

3.2. *Read Length Matters Differentially to Different Protein Families*

We then asked *how* different read lengths cause artificial functional differences. To address this question, we simulated short reads from 7 selected complete genomes (see Table 1), so that we can compare annotations based on the short reads to the annotations of the complete genes. It has been shown that read length matters and for all protein families, there is a significant sensitivity loss in the similarity searches for short reads of < 200 nt (homology finding is limited by the read length) [14]; we further show that 1) unequal fractions of reads tend to be missed in annotations for different functional categories (*i.e.* protein families); and 2) different functional categories tend to have different proportions of wrong annotations. These two confounding factors together can bias functional profiling of metagenomes, causing artificial functional differences between metagenomes—if the lengths of reads are not carefully controlled for.

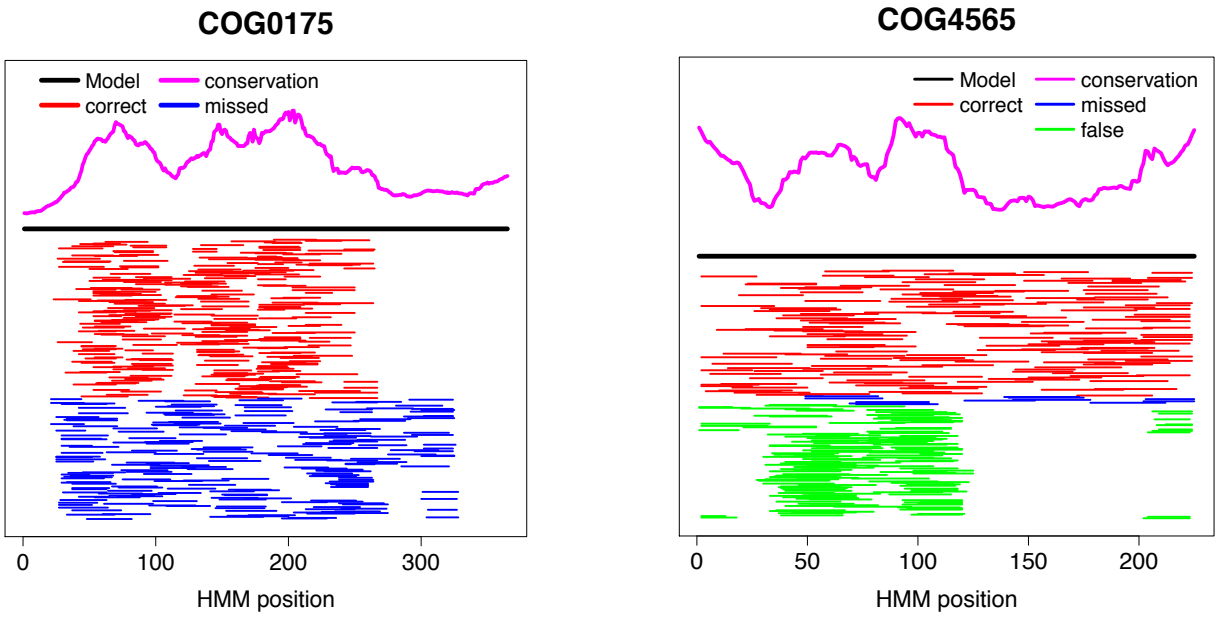


Fig. 2 Demonstration of functional annotation of short reads assigned to COG0175 (A) and COG4565 (B). The black line represents the HMM of the corresponding COG family. Red lines represent the reads that have the correct annotation. Blue lines represent the reads that are of that function, but are missed by the annotation pipeline. Green lines represent the reads that are incorrectly assigned to the family. The magenta curve above the HMM line shows the conservation score of the family. The higher the value is, the more conserved the position is.

For each family, we calculated the *precision* of the annotations, defined as the proportion of correctly annotated reads among all the reads assigned to the family, and the *recall*, defined as the number of reads correctly annotated over the actual number of reads with the function. Note that in the simulation experiment, we can assign any read based on the functional annotation of the complete gene, and thus the number of reads within a functional category can be precisely counted. Both the precision and recall of annotations differ among different protein families. Among the 4077 COG families present in the 7 genomes, 2146 families have a precision >90%,

and 674 families have a precision <= 20%; 880 families have a recall > 90% and 804 families have a recall <=20%. Figure 2 shows two examples of COG families with different levels of false annotations and missed annotations: COG0175 has significant missed annotations (293 reads are correctly assigned to this function, while 223 reads of this function are missed); by contrast, COG4565 has a large number of false annotations (158 reads are incorrectly assigned to this function) but very few missed annotations (12 reads). In the figure, we indicate the positional conservation for each COG (estimated based on the "hitting probability" [23]), which helps us understand the patterns of missed and false annotations of each COG family (see Discussion).

Some COG families share sequence similarity, and we observed incorrect functional assignments of short reads among these families. For example, COG3829 (transcriptional regulator containing PAS, AAA-type ATPase, and DNA-binding domains) and COG2204 (response regulator containing CheY-like receiver, AAA-type ATPase, and DNA-binding domains) both belong to the functional category T (signal transduction mechanisms); COG3829 also belongs to category K (transcription). 85 reads that are sampled from genes of function COG3829 were incorrectly assigned to COG2204, while 520 reads that are of function COG2204 were assigned to COG3829. Another pair is COG4664 (TRAP-type mannitol/chloroaromatic compound transport system, large permease component) and COG1593 (TRAP-type C4-dicarboxylate transport system, large permease component); these two families belong to different functional categories: COG4664 belongs to Q (secondary metabolites biosynthesis, transport and catabolism) and COG1593 belongs to G (carbohydrate transport and metabolism). 480 reads that are sampled from genes of function COG1593 were incorrectly assigned to COG4664, while no reads that are of function COG4664 were assigned to COG1593. An extreme case is function COG4565 (response regulator of citrate/malate metabolism): short reads simulated from genes of seven different functions (including COG3290, COG0745, COG2197, COG2204, COG3279, COG1629, COG0784) are assigned to this function.

Table 3. Annotations of the reads simulated from gene YP_012234.1.

Gene/reads	Positions	COG3604	COG2197	COG3437	COG3829	COG4565	COG2204
YP_012234.1	1–1419	6.80e-114	2.90e-19	4.80e-27	3.50e-115	1.70e-23	***6.40e-145***
r231087.1	49–162	-	-	***7.90e-05***	-	-	-
r173226.1	57–167	-	-	***9.80e-06***	-	0.036	0.016
r4681.1	140–240	-	7.40e-05	4.60e-06	-	***2.90e-07***	6.00e-05
r122743.1	160–268	-	0.00012	0.0021	-	***2.30e-05***	0.00021
r111158.1	178–286	-	***0.00016***	0.0053	-	0.013	0.0032
r191730.1	470–585	1.10e-15	-	-	***2.10e-16***	-	3.80e-14
r196288.1	547–662	***8.10e-18***	-	-	4.80e-16	-	1.40e-15

This table lists the E-values of selected reads sampled from gene YP_012234.1 (with sampled regions shown in the 'Positions' column) that have different best matches than the complete gene. The best COG family assignment (with lowest E-value) for each read was highlighted with the E-value shown in bold and italic, with '-' indicating an insignificant match between the read and corresponding COG family.

We show that false annotations of short reads among related families may be explained by the local sequence variation of the involved families and the fact that different families may share local similarities [24, 25]. Table 3 lists the annotations of short reads sampled from gene YP_012234.1. Gene YP_012234.1 can be assigned to function COG2204, although it also shares significant similarity with several other COG families, including COG3604 and COG3284. 66% of the short reads sampled from this gene can be assigned to a function (with E-value smaller than the cutoff), among which a mere 23% are assigned to the same function (COG2204) as the complete gene, using the best hit strategy. Among the short reads that were assigned to different COGs, read 'r173226.1' (which is sampled between 57 and 167 bp of this gene) can be assigned to COG3437 with an E-value of 9.80e-06; by contrast, the match between this read and the "correct" function COG2204 is much less significant, with an E-value of 0.016 (see more examples in Table 3).

3.3. *Artificial Bias of Functional Profiles can be Adjusted by Considering Read Lengths*

Having shown that different read lengths can cause artificial functional differences because of the differential impact of short reads on different protein family, we then asked: *can* the frequency estimates be adjusted, to allow functional profiling of a metagenome independent of read length? We proposed to adjust the functional profile of a metagenome—built with reads of a certain length—using the recall and the precision of functional annotations of different families we learned from simulated reads of the same length:

$$Hit(f) = \frac{Hit_0(f)Precision(f)}{Recall(f)Len(f)} \quad (1)$$

where $Hit_0(f)$ and $Hit(f)$ are the original and adjusted number of query sequencing reads that are assigned to function f, respectively. $Precision(f)$ and $Recall(f)$ are the precision and recall of functional annotation for functional category f based on similarity searches of simulated reads of the same length as short reads. $Len(f)$ is the average length of the protein sequences that are of the function f (calculated from the protein sequences that are assigned to functional category f collected in the eggNOG database). These family-specific values of precision and recall can be applied to any metagenomic project, once calculated for applicable read-lengths.

We used Eq. (1) to adjust the frequency of different COG families for the simulated twin-mouse short-read datasets, resulting in improvements of the frequency estimates (comparing to those estimated from the Sanger reads) for many functional categories, including E (amino acid transport and metabolism), F (nucleotide transport and metabolism), J (translation, ribosomal structure and biogenesis), N (cell motility), and S (function unknown) (see Figure 3A). Here we show the details of the comparison for functional category E. Its frequencies estimated from both obese and lean datasets (shown in red and blue unfilled triangles in Figure 3A, respectively) were significantly overestimated; as a result, they differed greatly from those estimated from the Sanger datasets (shown in squares in Figure 3A), even though the short reads were simulated from the Sanger datasets. The adjustment decreased the artificial functional difference caused by the read length difference, revealing a clearer abundance difference of this functional category between the

obese and lean metagenomes. We believe that such adjustment is important for interpreting real metagenomes derived from different sequencing techniques, as shown in the twin-mouse datasets (see Figure 3B). For most functional categories (except for functional category V, defense mechanisms), the frequency adjustment helped to decrease the differences between the frequencies estimated from the short-read datasets and the Sanger datasets of the same microbial community, revealing consistent functional differences (*e.g.*, for functional category L, replication, recombination and repair) between obese and lean metagenomes (functional category L is more abundant in the obese metagenomes).

4. Discussion

Using simulation studies, we show that differences in read lengths can result in biased functional profiles of metagenomic samples, if the length difference is not treated properly. This is an important observation especially for metagenomic research, which utilize different kinds and generations of next-generation sequencing (NGS) techniques that produce reads of various lengths. It indicates that conclusions need to be drawn carefully when 1) characterizing the functional content (capacity) of a metagenome derived by NGS (as the functional profile built from short reads may not faithfully reflect the actual functional content of the microbiome), and when 2) comparing different metagenomes that are produced by different NGS techniques (*e.g.*, 454 versus Illumina). While Gordon *et al.* [3] suggest that different sequencing techniques will produce reads in a biased way (*e.g.,* pyrosequencing may under-sample reads from functional category A) to explain the functional difference observed between metagenomes containing reads of various lengths, our conclusion is that biased-annotation of short reads can explain the observed difference, indicating sequencing is not biased, rather annotation is biased. We expect that such functional annotation bias due to read length differences exists in other projects as well, as long as the projects involve short reads (such as those from current Illumina sequencers).

It is important to see how read length impacts other functional classifications: we plan to extend our study to include more functional categories, including the FIG families [26] and the KEGG pathways. We will try more sophisticated methods for functional profile adjustments, aiming to achieve better improvements. And more systematic tests are needed. One direction is to analyze the individual families (*e.g.*, their overall evolutionary conservation and positional-conservation patterns, as shown in Figure 2) and their relationships (*e.g.*, the local similarity between different families) in more details, and utilize this information when constructing functional profiles. Another direction is to consider the read length explicitly when adjusting the functional profiles, so that we do not need a specific learning procedure for each read length.

We have presented a simple method to adjust functional profiles, based on the read length, and we demonstrate that it is possible to alleviate the "read-length-bias" problem. We feel that this approach is immediately applicable to many past and on-going metagenomic studies, including the Human Microbiome Project [27], impacting not only the functional annotation of metagenomes (as shown in this paper), but also other aspects, such as the enterotype-study of human gut microbiomes [1], which is based on read mapping to reference genomes for quantification of the abundances of different genera, and where obvious differences can be observed in the results drawn from metagenomes sequenced by Illumina or Sanger sequencing [1].

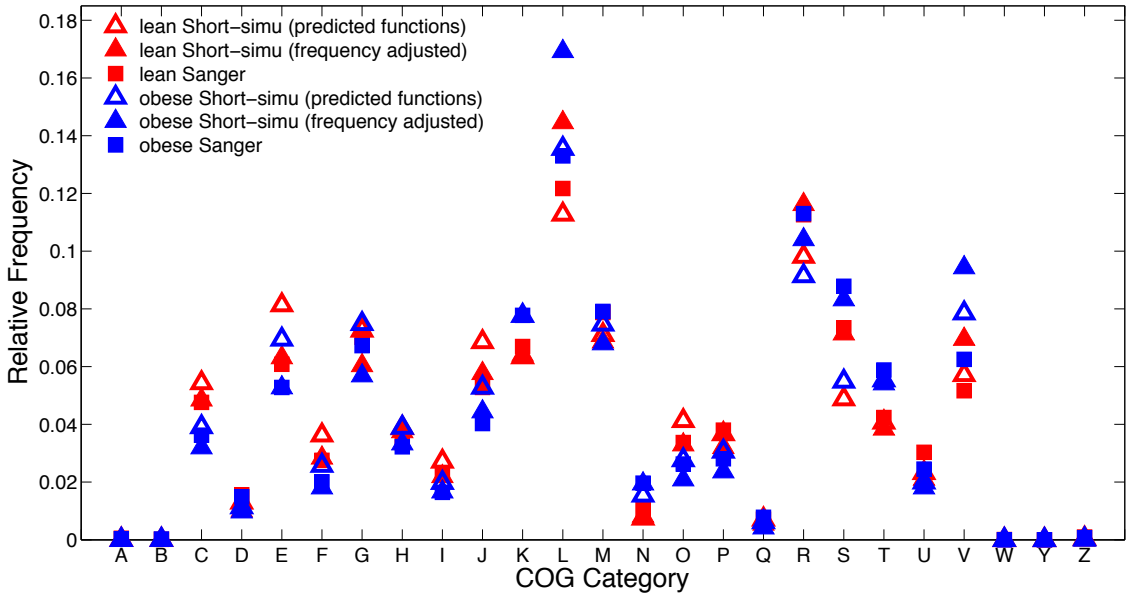
5. Acknowledgements

The authors thank Drs. Mina Rho and Haixu Tang for helpful discussions.

References

1. M. Arumugam, J. Raes, E. Pelletier *et al*, *Nature*, **473**:174-180 (2011).
2. J. Qin, R. Li, J. Raes *et al*, *Nature*, **464**(7285):59-65 (2010).
3. P. J. Turnbaugh, R. E. Ley, M. A. Mahowald *et al*, *Nature*, **444**(7122):1027-1031 (2006).
4. G. W. Tyson, J. Chapman, P. Hugenholtz *et al*, *Nature*, **428**(6978):37-43 (2004).
5. J. C. Venter, K. Remington, J. F. Heidelberg *et al*, *Science*, **304**(5667):66-74 (2004).
6. C. S. Riesenfeld, P. D. Schloss and J. Handelsman, *Annu Rev Genet*, **38**:525-552 (2004).
7. S. G. Tringe, C. von Mering, A. Kobayashi *et al*, *Science*, **308**(5721):554-557 (2005).
8. W. Xie, F. Wang, L. Guo *et al*, *ISME J*, **5**(3):414-426 (2011).
9. S. F. Altschul, W. Gish, W. Miller *et al*, *J Mol Biol*, **215**(3):403-410 (1990).
10. A. Krogh, M. Brown, I. S. Mian *et al*, *J Mol Biol*, **235**(5):1501-1531 (1994).
11. R. L. Tatusov, N. D. Fedorova, J. D. Jackson *et al*, *BMC Bioinformatics*, **4**:41 (2003).
12. R. L. Tatusov, M. Y. Galperin, D. A. Natale *et al*, *Nucleic Acids Res*, **28**(1):33-36 (2000).
13. M. Kanehisa, M. Araki, S. Goto *et al*, *Nucleic Acids Res*, **36**(Database issue):D480-484 (2008).
14. K. E. Wommack, J. Bhavsar and J. Ravel, *Appl Environ Microbiol*, **74**(5):1453-1463 (2008).
15. I. Sharon, A. Pati, V. M. Markowitz *et al*, *A statistical framework for the functional analysis of metagenomes*, in *RECOMB 2009; (*Tucson, AZ, 2009), pp 496-511.
16. J. C. Wooley and Y. Ye, *J Comput Sci Technol*, **25**(1):71-81 (2009).
17. N. Shah, H. Tang, T. G. Doak *et al*, *Comparing bacterial communities inferred from 16S rRNA gene sequencing and shorgun metagenomics*, in *Pac Symp Biocomput; (*The Big Island of Hawaii, 2011), pp 165-176.
18. D. C. Richter, F. Ott, A. F. Auch *et al*, *PLoS One*, **3**(10):e3373 (2008).
19. S. R. Eddy, *Genome Inform*, **23**(1):205-211 (2009).
20. E. A. Dinsdale, R. A. Edwards, D. Hall *et al*, *Nature*, **452**(7187):629-632 (2008).
21. L. J. Jensen, P. Julien, M. Kuhn *et al*, *Nucleic Acids Res*, **36**(Database issue):D250-254 (2008).
22. R. C. Edgar, *Nucleic Acids Res*, **32**(5):1792-1797 (2004).
23. B. Schuster-Bockler, J. Schultz and S. Rahmann, *BMC Bioinformatics*, **5**:7 (2004).
24. G. Agarwal, S. Mahajan, N. Srinivasan *et al*, *PLoS One*, **6**(3):e17826 (2011).
25. B. Morgenstern, K. Frech, A. Dress *et al*, *Bioinformatics*, **14**(3):290-294 (1998).
26. R. Overbeek, T. Begley, R. M. Butler *et al*, *Nucleic Acids Res*, **33**(17):5691-5702 (2005).
27. J. Peterson, S. Garges, M. Giovanni *et al*, *Genome Res*, **19**(12):2317-2323 (2009).

A



B

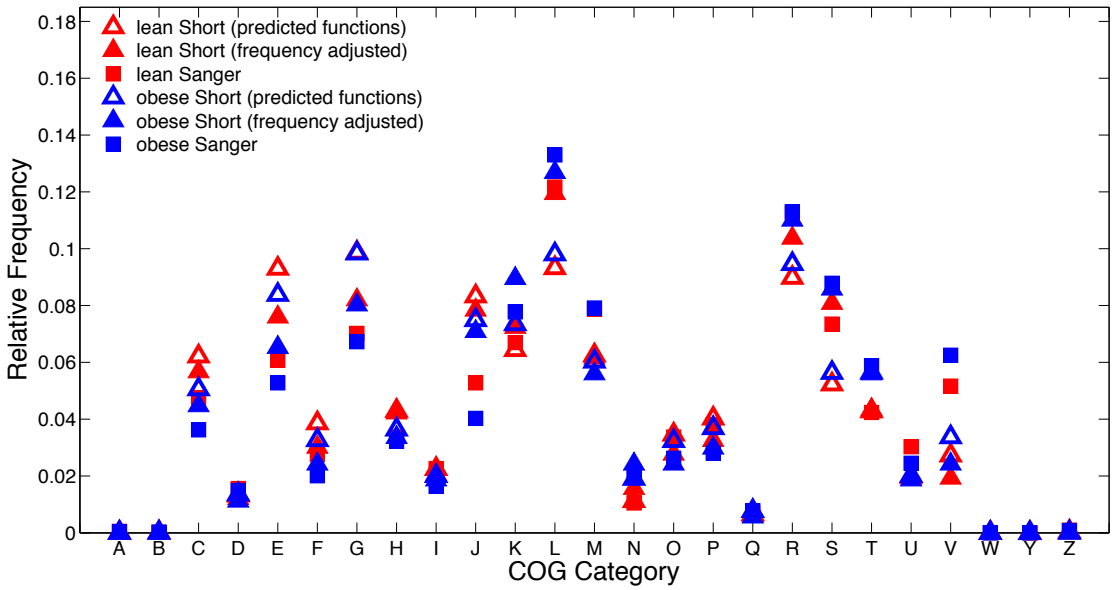


Fig. 3 Comparison of functional profiles of lean- (shown in red) and obese-mouse fecal microbiomes (shown in blue) built from Sanger reads (shown in squares) and short reads (triangles). Both original functional profiles (shown in unfilled triangles) and adjusted ones by applying Eq. 1 (shown in filled triangles) are shown in the figures: (A) for simulated short reads, and (B) for actual pyrosequence datasets.