



## OPEN

# How much metagenomic sequencing is enough to achieve a given goal?

SUBJECT AREAS:  
MOLECULAR ECOLOGY  
MICROBIAL ECOLOGY  
ENVIRONMENTAL  
MICROBIOLOGY  
METAGENOMICS

Jiajia Ni<sup>1,2\*</sup>, Qingyun Yan<sup>1\*</sup> & Yuhe Yu<sup>1</sup>

<sup>1</sup>Key Laboratory of Aquatic Biodiversity and Conservation of Chinese Academy of Sciences, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China, <sup>2</sup>Graduate University of Chinese Academy of Sciences, Beijing 100049, China.

Received  
18 March 2013

Accepted  
23 May 2013

Published  
11 June 2013

Correspondence and  
requests for materials  
should be addressed to  
Y.H.Y. (yhyu@ihb.ac.  
cn)

\* These authors  
contributed equally to  
this work.

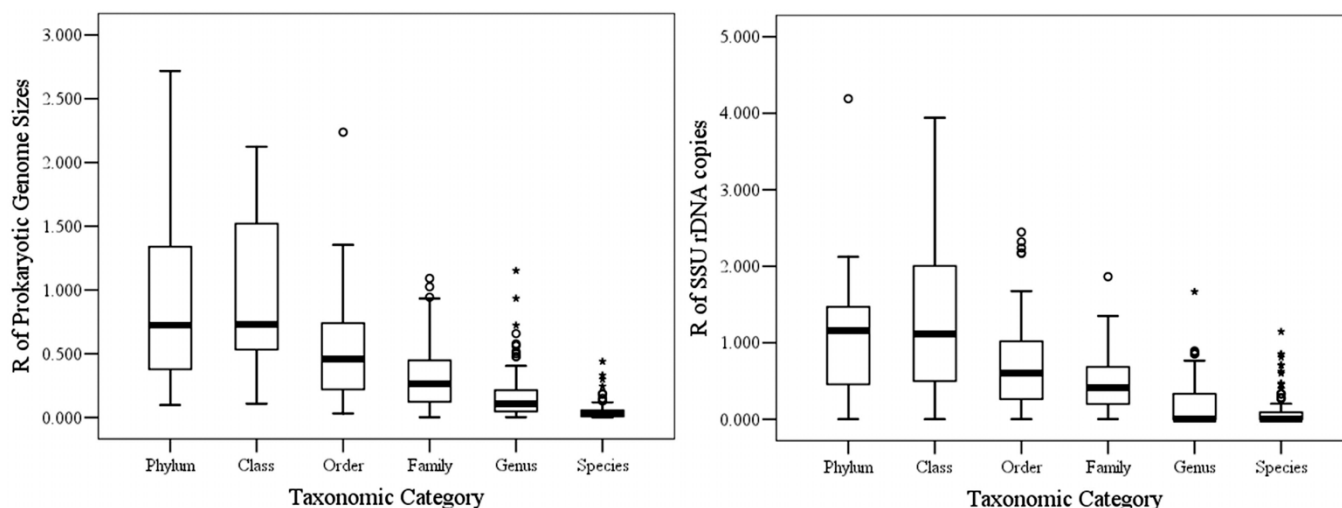
Metagenomic studies have dramatically expanded our knowledge of the microbial world. Furthermore, the amount of sample for sequencing has significantly increased with the development of high-throughput sequencing technologies. However, fully capturing all DNA sequences carried by every microorganism in the environment is still impossible. Therefore, estimating a reasonable and practical amount for sequencing to achieve the objectives is particularly necessary. In the present study, we introduce a novel method for estimating the required minimum amount for metagenomic sequencing for a given goal. We also calculated the genomic proportion of each operational taxonomic unit and the detection efficiency of a specific gene (we have used SSU rRNA gene as an example) based on a given amount for random metagenomic sequencing. The reasonable and practical estimated amount for sequencing in metagenomic studies will provide good reference information when applying high-throughput sequencing for a given goal.

Metagenomic studies have dramatically expanded our knowledge of the microbiota in various environments (e.g., the human intestinal tract, oceans and soil)<sup>1–6</sup>. However, fully enumerating all the DNA sequences carried by each microorganism in a given environment is impossible despite the great increase in sequencing depth in recent years. Therefore, estimating the amount needed for sequencing to fulfill the objectives of a given study is especially important. Many studies have estimated the metagenomic content in a given environment<sup>7–10</sup>. However, most of these studies did not consider the SSU rDNA copies per prokaryotic genome or assumed that all prokaryotes have the same genome size<sup>8,9</sup>. Prokaryotic genomes have 1 to 15 SSU rDNA copies that range from 139 kb to 13,034 kb. These diversities significantly hinder the accurate estimation of the sequencing amount. In this study, we introduce a novel computational approach for estimating the amount required for sequencing according to previously reported data (i.e., SSU rDNA distribution from targeted SSU rDNA sequencing, prokaryotic genome size and SSU rDNA copies of different genomes).

## Results

**Maximum potential error for estimates for prokaryotic genome size and SSU rDNA copies under different taxonomic categories.** To evaluate the maximum potential error for estimates for prokaryotic genome size and SSU rDNA copies under different taxonomic categories (i.e., species, genus, family, order, class and phylum), we introduced a ratio ( $R$ ) calculated according to equation (1). Higher  $R$  values indicate larger maximum potential errors. We calculated the  $R$  values for prokaryotic genome size and SSU rDNA copies for each taxonomic category (Supplementary dataset 1). The results indicate a significant difference among the  $R$  values for genome size among the different taxonomic categories (one-way ANOVA,  $F = 87.037$ ,  $p < 0.001$ ; Figure 1). Both the LSD and the Student–Newman–Keuls post hoc tests showed no significant differences between the  $R$  values at the species and the genus levels and between the class and phylum levels, whereas significant differences were detected among residual levels. Similar results were obtained for the SSU rDNA copies except for that between the genus level and the species level. Therefore, the diversity of both SSU rDNA copies and prokaryotic genome sizes could significantly disturb the accurate estimation of the amount for metagenomic sequencing.

**$C_i$  difference of the human faecal samples among different taxonomic categories.** The abundances of different prokaryotes in a particular environment are generally different. Most individuals in an ecosystem belong to a few numerically dominant species with low abundance. For instance, Sogin *et al.*<sup>11</sup> reported that the relative abundance of different operational taxonomic units (OTUs) in deep sea samples vary by more than three



**Figure 1** | Box plots showing the distribution of  $R$  values for prokaryotic genome size and SSU rDNA copies at different taxonomic categories.

orders of magnitude and the dominant bacterial populations in the diffuse flow and deep water samples from the North Atlantic are only very few. Eckburg *et al.*<sup>12</sup> pointed out that all 1524 archaeal SSU rDNA sequences from six major subdivisions of the human colon and faecal samples belong to a single phylotype (*Methanobrevibacter smithii*), with only 395 bacterial phylotypes assigned to 7 phyla (301 phylotypes belong to Firmicutes). Therefore, we cannot accurately calculate the amount required for sequencing if the abundance of each prokaryote in a given environment is unknown. To demonstrate and verify the feasibility of our estimation method, we retrieved the SSU rDNA sequence datasets of human faecal microbiota from the study by Eckburg *et al.*<sup>12</sup>. Briefly, 2339 SSU rDNA sequences (1060 from specimen A, 617 from specimen B and 662 from specimen C) were retrieved.

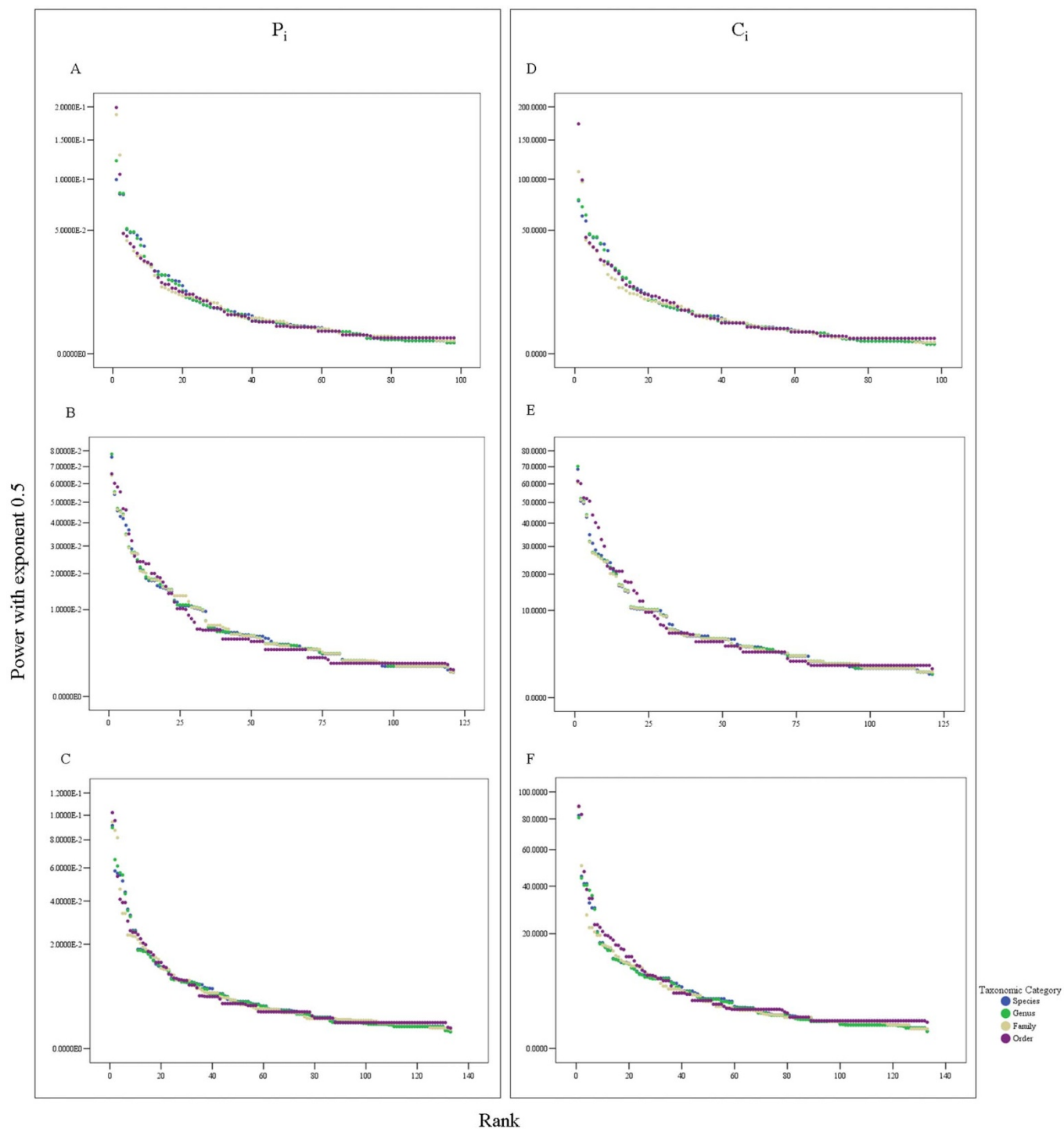
Equations (2) and (3) provide the calculations for the genomic proportion of the  $i^{\text{th}}$  taxonomic group ( $P_i$ ) and the coverage of the  $i^{\text{th}}$  taxonomic group ( $C_i$ ) in the human faecal samples (Supplementary dataset 2). The estimated  $C_i$  among different taxonomic categories were not significantly different (one-way ANOVA,  $P_i$ :  $F_A = 0.000$ ,  $p_A = 1.000$ ;  $F_B = 0.000$ ,  $p_B = 1.000$ ;  $F_C = 0.000$ ,  $p_C = 1.000$ ;  $C_i$ :  $F_A = 0.087$ ,  $p_A = 0.967$ ;  $F_B = 0.135$ ,  $p_B = 0.939$ ;  $F_C = 0.209$ ,  $p_C = 0.890$ ; Figure 2). This result indicates that we could obtain similar results by applying the correction to the different taxonomic categories for estimating the  $P_i$  and the  $C_i$ . Hence, we could estimate  $P_i$  and  $C_i$  using high-throughput sequencing data.

**Genomic coverage estimation of a single species.** The validity of metagenomic datasets is usually assessed using the genomic coverage of a single species<sup>10,13</sup>. The genome of a single species can be accurately assembled from a complex metagenomic dataset when it shows at least  $20\times$  coverage<sup>13</sup>. To assess the genomic coverage of each OTU in a metagenomic study, the amount of DNA and the genomic coverage of each OTU were estimated using the 3G metagenomic data per specimen (Supplementary dataset 2). The ratio of the coverage to the proportion of each OTU among different taxonomic categories was not significantly different in all of the specimens (one-way ANOVA,  $F_A = 1.350$ ,  $p_A = 0.258$ ;  $F_B = 1.528$ ,  $p_B = 0.206$ ;  $F_C = 0.899$ ,  $p_C = 0.442$ ; Figure 3). Therefore, the coverage of each OTU was not overestimated or underestimated in the different taxonomic categories. Less than 15 genomes in the 3G metagenomic data per specimen reached  $20\times$  coverage. This result indicates that the 3G metagenomic sequencing is still insufficient for determining the genetic content of faecal samples. Equation (3) indicated that the genomic coverage could be calculated using different sequencing amounts and the corresponding number of

species with genomes that reached  $20\times$  coverage could be counted. Therefore, the required amount for sequencing could be inferred according to the given number of species with genomes that reached to  $20\times$  coverage.

**Estimation of the required minimum amount of sample for metagenomic sequencing for a given goal.** To provide an example, we estimated the required minimum amount for the metagenomic sequencing of specimen A faecal microbiota from Eckburg *et al.*<sup>12</sup>. Equation (4) indicated that the minimum amount required for metagenomic sequencing could be estimated based on the  $P_i$  calculated in Supplementary datasets 2. Based on the calculations, if the given goal set for ensuring the genomes of species with a relative abundance more than 1% reached  $20\times$  coverage, the estimated amounts for sequencing was 7.15 Gb at the species level, 7.04 Gb at the genus level, 8.14 Gb at the family level and 6.41 Gb at the order level (Supplementary dataset 3). To obtain more reliable estimates, we plotted the regression curves of the estimated amounts for sequencing vs. the rank of the genomic proportion of each taxon and calculated the regression equation (Figure 4). The regression equation indicated that the estimated amounts for sequencing specimen A were 7.00 Gb at the species level, 6.93 Gb at the genus level, 7.10 Gb at the family level and 6.54 Gb at the order level. These results imply that at least 7 Gb is required for sequencing to enumerate the gene contents of prokaryotes with relative abundance of more than 1% in the human faecal microbiota.

**Detection efficiency of a particular gene.** Comparing the compositions of gene families or orthologs to analyze the functional differences of microbiota is a crucial component of microbial ecology. Metagenomic sequencing provides an effective way to enumerate the compositions of gene families or orthologs. Assessing the detection efficiency of each gene is a basic issue in this case, especially for those rare gene families or orthologs. To estimate the detection efficiency of a particular gene, we used the SSU rDNA as the target gene for the example. We assumed that any SSU rDNA read could confirm the presence of a particular gene and the  $L_{\text{read}}$  was 100 bp<sup>13</sup>. The estimated efficiency of SSU rDNA from the selected OTUs with different OTU proportions in faecal microbiota was calculated according to equation (6). Among the different taxonomic categories, the ratio of detected efficiency and proportion of each OTU was significantly different in specimens B and C but not in specimen A (one-way ANOVA,  $F_A = 0.899$ ,  $p_A = 0.450$ ;  $F_B = 3.020$ ,  $p_B = 0.039$ ;  $F_C = 5.404$ ,  $p_C = 0.002$ ; Figure 5). The LDS and the Student–Newman–Keuls post hoc tests also show that the efficiency at the

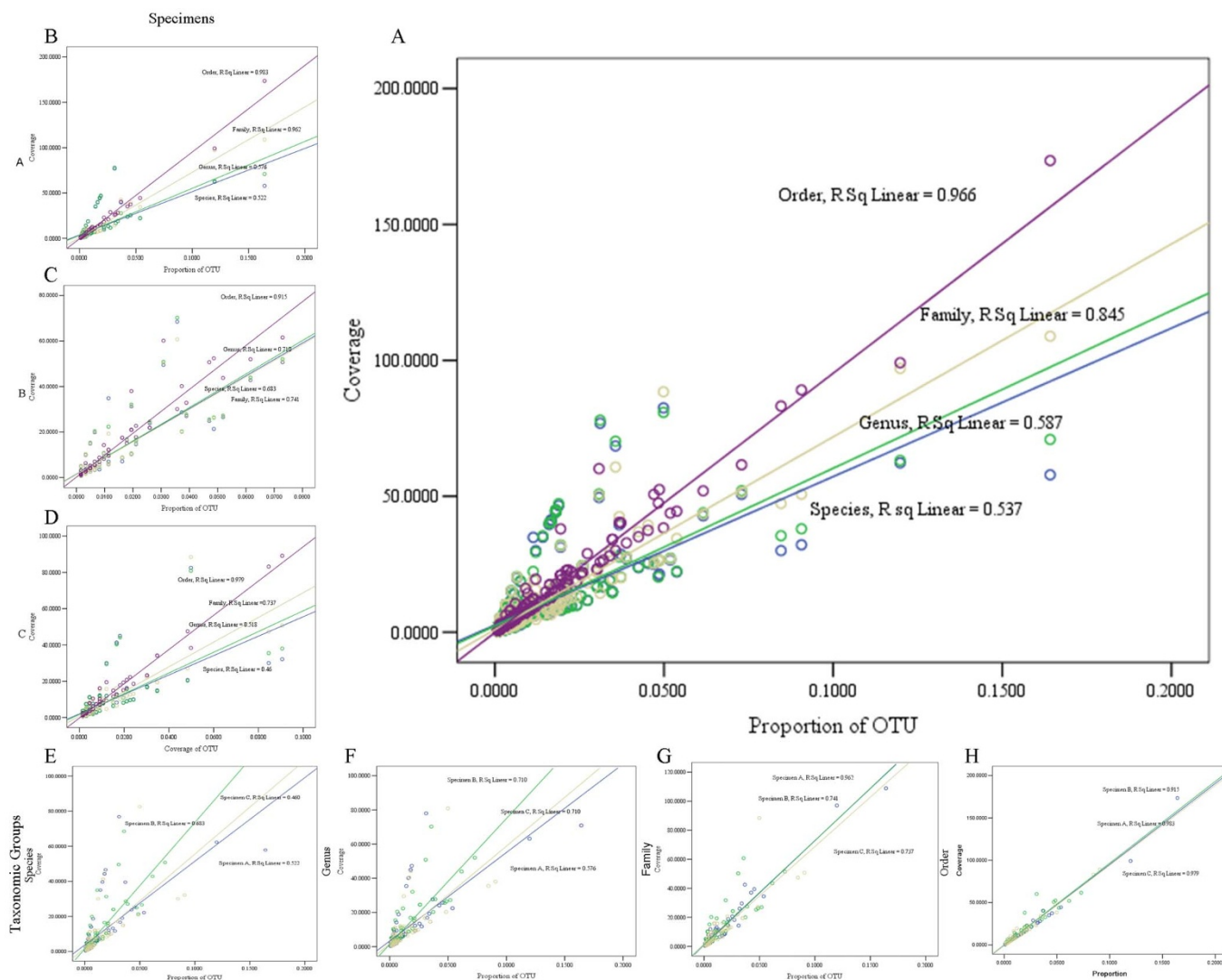


**Figure 2** | Rank–proportion curves (A, B, C) and rank–coverage curves (D, E, F) of human faecal samples with 3G metagenomic data per specimen. (A, D) specimen A; (B, E) specimen B; (C, F) specimen C. Species indicates the taxonomic categories corrected to the species level. The genomic sizes and SSU rDNA copies were replaced with an attributively higher taxonomic group when a taxonomic group does not have a finished genomic sequence in the Integrated Microbial Genomes (IMG) system (<http://img.jgi.doe.gov/cgi-bin/w/main.cgi>). Genus, family and order indicate the taxonomic categories corrected to the corresponding taxon level.

order level in specimens B and C were significantly different from those at other levels. These results indicate that the detected efficiency of a specific gene was possibly overestimated at the order level. The inaccuracy could be caused by differences in prokaryotic genome size. Therefore, a target gene should be affiliated to family level or a lower taxonomic category when estimating the detection efficiency of a particular gene.

## Discussion

The maximum potential error for estimating prokaryotic genome sizes and SSU rDNA copies at higher taxonomic levels was significantly higher than that at lower levels. Fortunately, prokaryotes can be classified into different species based on their SSU rDNA sequences<sup>14,12</sup>, and the taxonomic affiliations of the SSU rDNA sequences generated using Illumina and pyrophosphate sequencing



**Figure 3 | Correlation between the proportion of OTU and prokaryotic genomic coverage in the 3G metagenomic data per specimen.** (A) Distribution of the estimated coverage with the proportion of OTU at different taxonomic levels; (B, C, D) distribution diversity of the estimated coverage at different taxonomic levels with the proportion of OTU from specimens A, B and C; (E, F, G, H) distribution diversity of the estimated coverage of the three specimens with the proportion of OTU at species, genus, family and order levels.

platform can reach the genus level<sup>15–18</sup>. The taxonomic affiliations of sequences generated from these high-throughput sequencing platform premised to reach a relatively lower taxonomic category (i.e., species level) following the optimisation of high-throughput sequencing technologies, especially with the increasing length of reads.

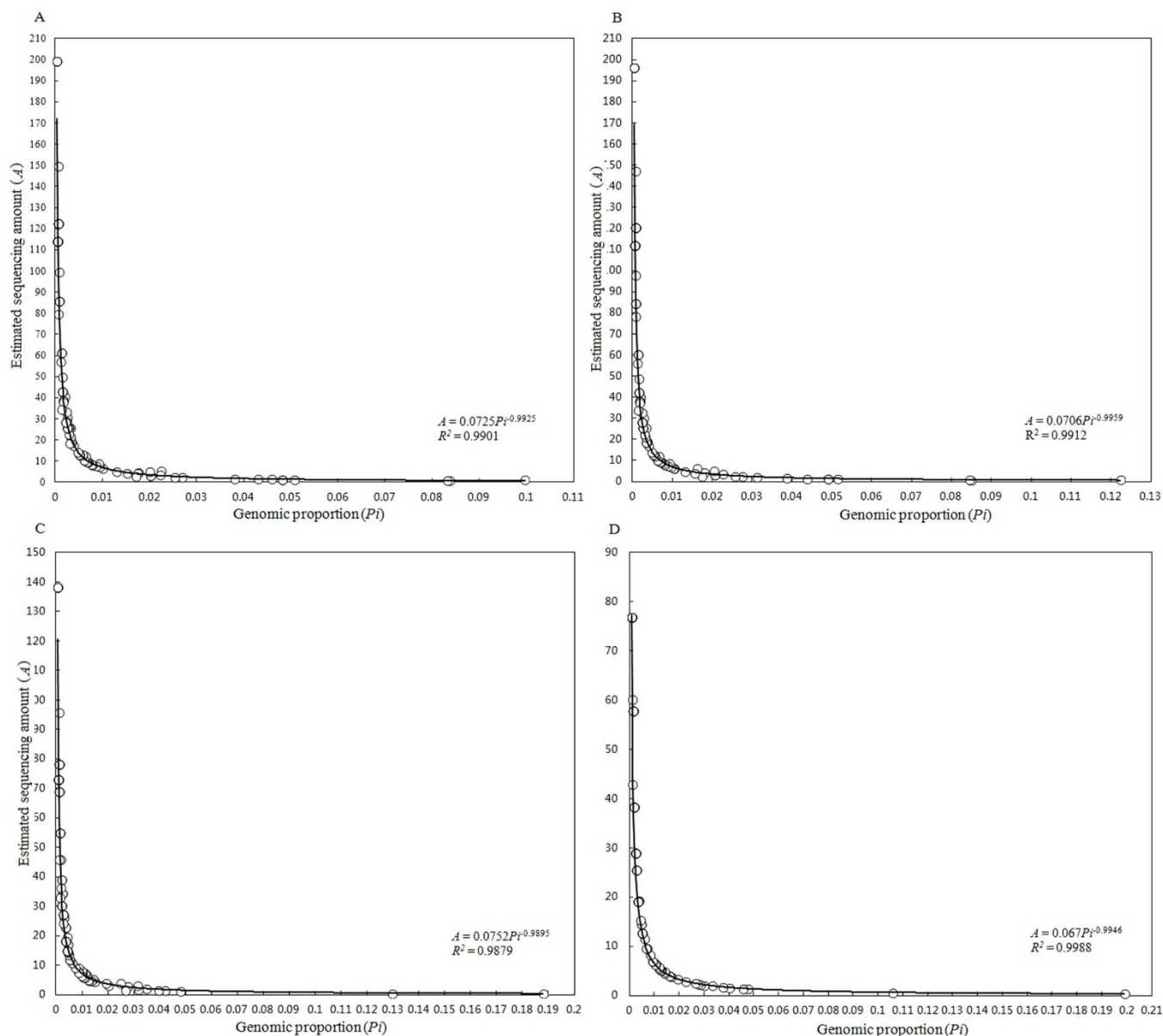
Although we tried to perform an accurate estimation, two main factors influenced the accuracy. First, only the average genome size and average SSU rDNA copies of higher taxonomic category can be used for many prokaryotic genomes that are absent in the public databases (e.g., *Barnesiella*, *Butyricoccus*, *Butyricimonas*, *Oscillibacter*, *Peptococcus* and *Succiniclasticum*). Obviously, this replacement will overestimate the *R* value. However, this limitation will improve with the increasing prokaryotic genomes that will be sequenced in the future. Second, some of the SSU rDNA sequences were not accurately affiliated with a genus or a species. Consequently, we had to apply a higher taxonomic category to estimate the amount required for sequencing. Although sequencing the SSU rDNA of all prokaryotic species is impossible, supplementing SSU rDNA sequences, especially novel species SSU rDNA sequences, will partially overcome the shortage in the future.

In the present study, the reference data for prokaryotic genome size and SSU rDNA copies were not selected based on habitat.

However, the prokaryotes in different environments may encounter different selective stresses; therefore, the genome sizes and gene compositions may differ. For instance, Frese *et al.*<sup>19</sup> found that *Lactobacillus reuteri* has host-specific genomic features. They found that *L. reuteri* F275 isolated from a healthy human adult is unable to colonise *Lactobacillus*-free mice and it has a genome approximately 279 kb smaller than that of rodent isolate *L. reuteri* 100–23. Therefore, the maximum potential error should be reduced when the reference data sets are constructed based on the habitat-selected prokaryotic genomes and the estimated amount required for sequencing should be more accurate.

Tamanes *et al.*<sup>10</sup> argued that the number of sequence reads belonging to unobserved OTUs should be subtracted from the total number of reads when estimating the coverage of a species using the composition of SSU rDNA. Although the unobserved OTUs significantly affect the alpha-diversity of a given metagenome and the genomic coverage of rare species, our results indicate that they did not affect the genomic coverage of dominant and common species, as shown by the unchanged proportion of dominant and common species with the increasing total number of reads (unpublished data by Ni *et al.*).

Considering plasmid DNA, mitochondrial DNA and eukaryotic DNA are also present in environmental metagenomes<sup>7</sup>, the



**Figure 4 | Correlation between the estimated sequencing amounts and the genomic proportion of each taxon. (A)** Corrected to the species level; **(B)** corrected to the genus level; **(C)** corrected to the family level; and **(D)** corrected to the order level.

proportion of plasmid DNA and eukaryotic DNA in metagenomes should also be considered to estimate more accurately the amount for sequencing. We did not consider the composition of archaea genomes; thus, the coverage of the bacterial genome may have been overestimated. Considering the proportion of plasmid, mitochondrial, archaea and eukaryotic DNAs in the entire metagenome, a more accurate estimation could be conducted using a corrected equation that replaces  $\sum_{i=1}^n (AX_{genome})_i$  in equation (2) with the total metagenomic DNA.

In conclusion, the required minimum amount for metagenomic sequencing could be estimated using the proposed method when the abundances of different prokaryotes in particular samples are known.

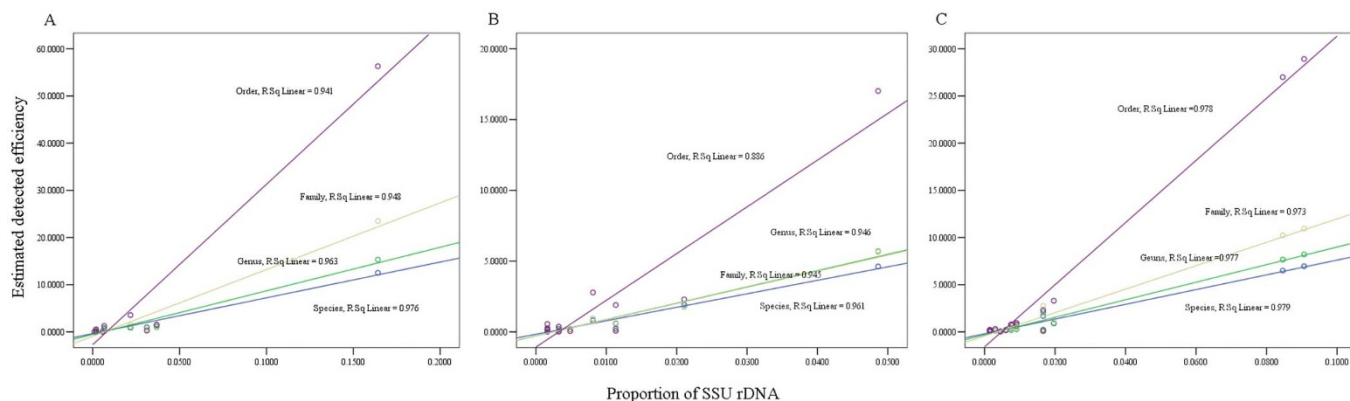
## Methods

In all microbial diversity studies using polymerase chain reaction (PCR)-based methods, the targeted genes (e.g., SSU rDNA) are assumed to have equal amplification efficiency according to the following classical PCR equation:  $N_{Cq} = N_0 \cdot (1 + E)^{Cq}$ , where  $N_0$  is the amount of the target sequence before PCR,  $N_{Cq}$  is the amount of PCR

products after  $C_q$  rounds of amplification and  $E$  is the assumed constant until the onset of PCR exhaustion<sup>20</sup>. The proportion of amplified DNA sequences from different microorganisms in the PCR products was assumed equal to those in the original sample. Therefore, we could infer the proportion of bacteria in an environmental sample according to the bacterial SSU rDNA sequences, which can be acquired through targeted SSU rDNA sequencing. Furthermore, the required sequencing amount could be inferred when the genome sizes of each bacterium are known. Thus, the structure and function of bacteria in a particular environment can be investigated using high-throughput sequencing technologies. The detection limit of a specific gene could be estimated when the size of a specific gene and its copies are given.

To obtain the reference data for the prokaryotic genome sizes and SSU rDNA copies, the genomic data of 1627 sequenced and annotated prokaryotes, including 111 archaea and 1516 bacteria, were retrieved from the Integrated Microbial Genomes system<sup>21</sup>. These prokaryotic genomes belong to 620 known genera and 2 unclassified genera (Supplementary dataset 4). The prokaryotic genome sizes range from 138,927 bp to 13,033,779 bp ( $3,497,775 \pm 46,278.101$  bp, s.e.m), and the SSU rDNA copies range from 1 to 15 ( $3.94 \pm 0.068$ , s.e.m) (Supplementary dataset 4 and Figure S1).

To evaluate the maximum potential error in the estimation of the prokaryotic genome sizes and SSU rDNA copies at different taxonomic categories (i.e., species, genus, family, order, class and phylum), we introduced a ratio ( $R$ ) calculated according to the following equation:



**Figure 5** | Correlation between the detection efficiency of SSU rDNA and the proportion of OTUs. (A) specimen A; (B) specimen B; (C) specimen C.

$$R = \max\{E' | E' = |E - AX|\} / AX \quad (1)$$

where  $E$  is the extreme value of genome sizes or SSU rDNA copies,  $AX$  is the mean of the corresponding genome size or SSU rDNA copy and  $E'$  is the margin between  $E$  and  $AX$ . Bigger  $R$  values indicate larger maximum potential errors.

The genomic proportion of the  $i^{\text{th}}$  taxonomic group ( $P_i$ ) was calculated using the following equation:

$$P_i = ((N_i / (AX_{ssu})_i) \cdot (AX_{genome})_i) / (\sum_{i=1}^n (N_i / (AX_{ssu})_i) \cdot (AX_{genome})_i) \quad (2)$$

where  $N_i$  is the abundance of the SSU rDNA copies of the  $i^{\text{th}}$  taxonomic group,  $(AX_{ssu})_i$  and  $(AX_{genome})_i$  are the average value of SSU rDNA copies and the mean genome size of the  $i^{\text{th}}$  taxonomic group, respectively, and  $N_i / (AX_{ssu})_i$  is the individual number of the  $i^{\text{th}}$  taxonomic group. The introduction of  $N_i / (AX_{ssu})_i$  into equation (2) eliminates the disturbance of SSU rDNA copies on  $P_i$ . Equation (2) also eliminates the disturbance of genome size differences on  $P_i$ .

The coverage of the  $i^{\text{th}}$  taxonomic group ( $C_i$ ) was calculated using the following equation:

$$C_i = (P_i \cdot A) / (AX_{genome})_i \quad (3)$$

where  $A$  is a given sequencing amount of the microbial metagenome. Considering  $C_i$  is given, the sequencing amount of a microbial metagenome was calculated using the following conversion equation:

$$A = C_i \cdot (AX_{genome})_i / P_i \quad (4)$$

The detected probability of a given gene ( $P$ ) was calculated using the following equation:

$$P = P_i \cdot (L \cdot n) / (AX_{genome})_i \quad (5)$$

where  $L$  is the gene length,  $n$  is the gene copies and  $P_i$  is the proportion of the taxonomic group(s) that contain the gene in the metagenome.  $P_i$  was calculated using equation (2). If the presence of one of the reads of the special gene could be confirmed, equation (5) should be corrected using the following equation:

$$P = (P_i \cdot (L \cdot n) / (AX_{genome})_i) \cdot (L / L_{read}) \quad (6)$$

where  $L_{read}$  is the length of a single read of the high-throughput sequencing.

- Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* **312**, 1355–1359 (2006).
- Biddle, J. F., Fitz-Gibbon, S., Schuster, S. C., Brenchley, J. E. & House, C. H. Metagenomic signature of the Peru Margin seafloor biosphere show a genetically distinct environment. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 10583–10588 (2008).
- Dinsdale, E. A. *et al.* Functional metagenomic profiling of nine biomes. *Nature* **452**, 629–632 (2008).
- Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480–484 (2009).
- Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).
- Xie, W. *et al.* Comparative Metagenomics of microbial communities inhabiting deep-sea hydrothermal vent chimneys with contrasting chemistries. *ISME J.* **5**, 414–426 (2011).

7. Raes, J., Korb, J. O., Lercher, M. J., von Mering, C. & Bork, P. Prediction of effective genome size in metagenomic samples. *Genome Biol.* **8**, R10 (2007).
8. Quince, C., Curtis, T. P. & Sloan, W. T. The rational exploration of microbial diversity. *ISME J.* **2**, 997–1006 (2008).
9. Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K. & Hugenholtz, P. A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev.* **72**, 557–578 (2008).
10. Tamames, J., de la Peña, S. & de Lorenzo, V. COVER: *a priori* estimation of coverage for metagenomic sequencing. *Environ. Microbiol. Rep.* **4**, 335–341 (2012).
11. Sogin, M. L. *et al.* Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 12115–12120 (2006).
12. Eckburg, P. B. *et al.* Diversity of the human intestinal microbial flora. *Science* **308**, 1635–1638 (2005).
13. Luo, C., Tsementzi, D., Kyrpides, N. C. & Constantinidis, K. T. Individual genome assembly from complex community short-read metagenomic datasets. *ISME J.* **6**, 898–901 (2012).
14. Bosshard, P. P., Abels, S., Zbinden, R., Bottger, E. C. & Altwegg, M. Ribosomal DNA sequencing for identification of aerobic gram-positive rods in the clinical laboratory (an 18-month evaluation). *J. Clin. Microbiol.* **41**, 4134–4140 (2003).
15. Huber, J. A. *et al.* Microbial population structures in the deep marine biosphere. *Science* **318**, 97–100 (2007).
16. Liu, Z., DeSantis, T. Z., Andersen, G. L. & Knight, R. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res.* **36**, e120 (2008).
17. Bartram, A. K., Lynch, M. D. J., Stearns, J. C., Moreno-Hagelsieb, G. & Neufeld, J. D. Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. *Appl. Environ. Microb.* **77**, 3846–3852 (2011).
18. Zhou, H. W. *et al.* BIPES, a cost-effective high-throughput method for assessing microbial diversity. *ISME J.* **5**, 741–749 (2011).
19. Frese, S. A. *et al.* The evolution of host specialization in the vertebrate gut symbiont *Lactobacillus reuteri*. *PLoS Genet.* **7**, e1001314 (2011).
20. Bernth Jensen, J. M., Petersen, M. S., Stegger, M., Østergaard, L. J. & Møller, B. K. Real-time relative qPCR without reference to control samples and estimation of run-specific PCR parameters from run-internal mini-standard curves. *PLoS ONE* **5**, e11723 (2010).
21. Markowitz, V. M. *et al.* IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.* **40**, D115–D122 (2012).

## Acknowledgments

We thank Elisabeth M. Bik and David A. Relman from the Stanford University School of Medicine for providing us the SSU rDNA sequence datasets for human faecal microbiota and permitting us to use the datasets as examples in the manuscript. This work was supported by the Major State Basic Research Development Program of China (Grant number 2009CB118705) and the National Natural Science Foundation of China (Grant numbers 30970358 and 31172084).

## Author contributions

Y.Y. and J.N. designed the research. J.N. and Q.Y. analysed the data and wrote the paper.

## Additional information

Supplementary information accompanies this paper at <http://www.nature.com/scientificreports>



**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Ni, J., Yan, Q. & Yu, Y. How much metagenomic sequencing is enough to achieve a given goal? *Sci. Rep.* 3, 1968; DOI:10.1038/srep01968 (2013).



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs Works 3.0 Unported license. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/3.0>