# REANALYZE UNASSIGNED READS IN METAGENOMIC DATA USING CONSERVED GENE ADJACENCY

Daryi Wang* and Francis Cheng-Hsuan Weng

*Biodiversity Research Center, Academia Sinica,*
*Taipei, 115, Taiwan*
**Email: dywang@gate.sinica.edu.tw*
*Email: manpower@gate.sinica.edu.tw*

Huai-Kuang Tsai*, Chien-Hao Su, Ming-Tsung Hsu, and Tse-Yi Wang

*Institute of Information Science, Academia Sinica,*
*Taipei, 115, Taiwan*
**Email: hktsai@iis.sinica.edu.tw*
*Email: d95922033@ntu.edu.tw, mingtsung@iis.sinica.edu.tw, tseyiwang@gmail.com*

Investigation of metagenomes provides greater insight into uncultured microbial communities. The improvement in high-throughput sequencing technology, which yields a large amount of sequence data, has led to major breakthroughs in the field. However, at present, taxonomic binning tools for metagenomes discard 30-40% of the data due to the stringency of BLAST cut-offs. In an attempt to provide a comprehensive overview of metagenomic data, we re-analyzed the discarded metagenomes by using less stringent cut-offs; however, we added a new criterion, namely, the evolutionary conservation of adjacency between neighboring genes. To validate the feasibility of our approach, we re-analyzed discarded contigs and singletons from several environments with different levels of complexity. We also compared the consistency between our taxonomic binning and that reported in the original studies. Among the discarded data, we found that 20.8±3.9% of singletons and 11.1±1.0% of contigs could be assigned to taxa. The recovery rates for singletons were higher than those for contigs. Using *Pearson* correlation coefficient revealed a high degree of similarity (0.94±0.03 at the phylum rank and 0.85±0.06 at the family rank) between the proposed taxonomic binning approach and those reported in previous studies. In addition, an evaluation using simulated data demonstrated the reliability of the proposed approach. Our findings suggest that taking account of conserved neighboring gene adjacency can improve taxa assignment when analyzing metagenomes. In other words, utilizing the conserved gene order as a criterion can help reduce the amount of data discarded when analyzing metagenomes.

## 1. INTRODUCTION

The investigation of metagenomes, which sequences DNA from mixed environmental samples directly, has provided insights into microbial communities and is now widely used to study living microorganisms as a system [1-4]. The major goal of metagenomic studies is to determine the systemic properties of a microbial community, including the genetic, metabolic, ecological, physiological and behavioral aspects of all community members [5-8]. Recent investigations based on the reference database of known microbial genomes have revealed enormous variations in the microbiomes of diverse environments, such as human intestinal and salivary microbiota [9-11], microbial communities growing on sunken whale skeletons [12], and open ocean communities [13, 14].

The current trend in metagenomic analysis follows the so-called gene-centric approach, which assumes that genes that appear more frequently in one community than in others endow a beneficial function on that community [2]. The taxonomic assignment of scaffolds and contigs is performed using BLAST [15] or other homology search tools [16] with the sequence databases. However, for sequencing metagenomes, whole genome shotgun sequencing (WGS) technique [17] only yields sequences (reads) of ~1,000 base pairs in length. Since the majority of the reads only contain partial coding regions, they usually fail to be identified because of the limited match length. It is estimated that existing analytical methods discard approximately 30-40% of metagenomic data [9, 10, 12, 13, 18, 19].

To overcome the limitations of current binning approaches that rely heavily on the BLAST hit score, we propose a method for assigning reads discarded by the original studies. The new approach combins the BLAST search scores (two or more putative coding

* Corresponding author.

sequences (CDSs) in a read) and the concept of conserved gene adjacency. The rationale is based on the theory that genomes are shuffled, so local gene-order conservation reflects the specificity of microbial organisms [20]. For example, the conservation of the gene order in prokaryotes is known to be an important feature; hence, it has been used in function inference [21, 22]. Since gene order conservation is a genomic feature that is extensively conserved between closely related species [23, 24], the trend should be universal in prokaryotic genomes [25]. Furthermore, it is known that overlapping gene pairs are frequently observed in microbial chromosomes [26] and conserved across species [27]. Therefore, we argue that, if a genomic fragment contains two or more adjacent CDSs that can be identified by BLAST, it is reasonable to assign the sequence by using the proposed strategy, which combines two BLAST hit scores and the adjacency of the two genes.

A recent study showed that the average gene density in prokaryotic genomes is one gene per 1,000 nucleotides [28], which is close to the sequence length yielded by WGS. Thus, in this study, we re-analyzed the fragments that conventional approaches had discarded from two types of metagenomic data, namely, 13 healthy Japanese individuals [10] and the skeletons of whale carcasses (whale fall) [12]. Two types of genomic fragments, assembled contigs and raw single reads (singletons), were analyzed separately. The results showed that between 9.9% and 28.9% of the discarded data could be assigned to taxa. Furthermore, the microbial compositions using discarded data and those reported in previous studies [10, 12] were highly consistent in the family and phylum ranks. Therefore, we conclude that the proposed metagenomic sequencing approach can provide a more comprehensive overview of the functional and taxonomic content of a microbiome.

## 2. MATERIALS AND METHODS

Figure 1 shows an overview of our methodology. We analyzed two types of discarded genomic fragments from sunken whale skeletons [12] and human distal guts [10] (Table 1). To incorporate the conservation of gene order into the taxonomic classification, each discarded genomic fragment was screened for protein encoding genes via a BLASTX search against the NCBI ENTREZ Genome Project database. An expected cut-off value ($E$) of $10^{-5}$ was used to select the top 250 potential coding elements.

**Table 1.** Summary of collected metagenomic fragments.

| Data Type I - contigs - | Total contigs | Un-assigned | |
|---|---|---|---|
| | | Contigs[a] | Average length (bp) |
| whale fall sub. 1 | 35975 | 7039 | 1167 |
| whale fall sub. 2 | 32459 | 7660 | 1199 |
| whale fall sub. 3 | 27130 | 4990 | 1357 |
| Data Type I - contigs - | Total reads | Un-assigned | |
| | | Singletons | Average length (bp) |
| Japanese In-A | 76434 | 13399 | 1057 |
| Japanese In-B | 80617 | 7078 | 1058 |
| Japanese In-D | 84237 | 28244 | 1034 |
| Japanese In-E | 80852 | 10838 | 1124 |
| Japanese In-M | 89340 | 8456 | 1008 |
| Japanese In-R | 85787 | 21661 | 998 |
| Japanese F1-S | 78452 | 15378 | 1005 |
| Japanese F1-T | 81348 | 21780 | 958 |
| Japanese F1-U | 82525 | 11791 | 969 |
| Japanese F2-V | 80772 | 19733 | 1006 |
| Japanese F2-W | 79163 | 16961 | 1039 |
| Japanese F2-X | 80858 | 19351 | 1040 |
| Japanese F2-Y | 79754 | 20061 | 990 |

[a] Genes with best hits less than 30% identity in Archaea and Bacteria kingdoms from JGI.

Normally, the best hits are selected from BLAST results, but best hits do not provide information on adjacent genes. Therefore, the top 250 hits were selected instead. In our strategies, adjacent gene pair is a pair of genes that are directly next to each other in a given chromosome. Thus, each hit was grouped with its corresponding species. These hits were then compared in a pair-wise fashion in order to identify adjacent CDSs. The transcriptional direction (unidirectional (→→), convergent (→←), and divergent (←→)) of all identified adjacent CDSs should be consistent with the genomic arrangement of reference genomes. Next, the pairs with inconsistent genomic arrangement were removed. Subsequently, among the remaining pairs, the one with the minimum *E*-value product was selected, and its corresponding species was assigned to the taxa. In cases where the minimum *E*-value products of two or multiple pairs are the same or equaling to zero, the lowest common ancestor (LCA) [29] was used.

We also used singletons obtained from simulated datasets, a low-complexity community (simLC), to evaluate the performance of our taxonomic assignment

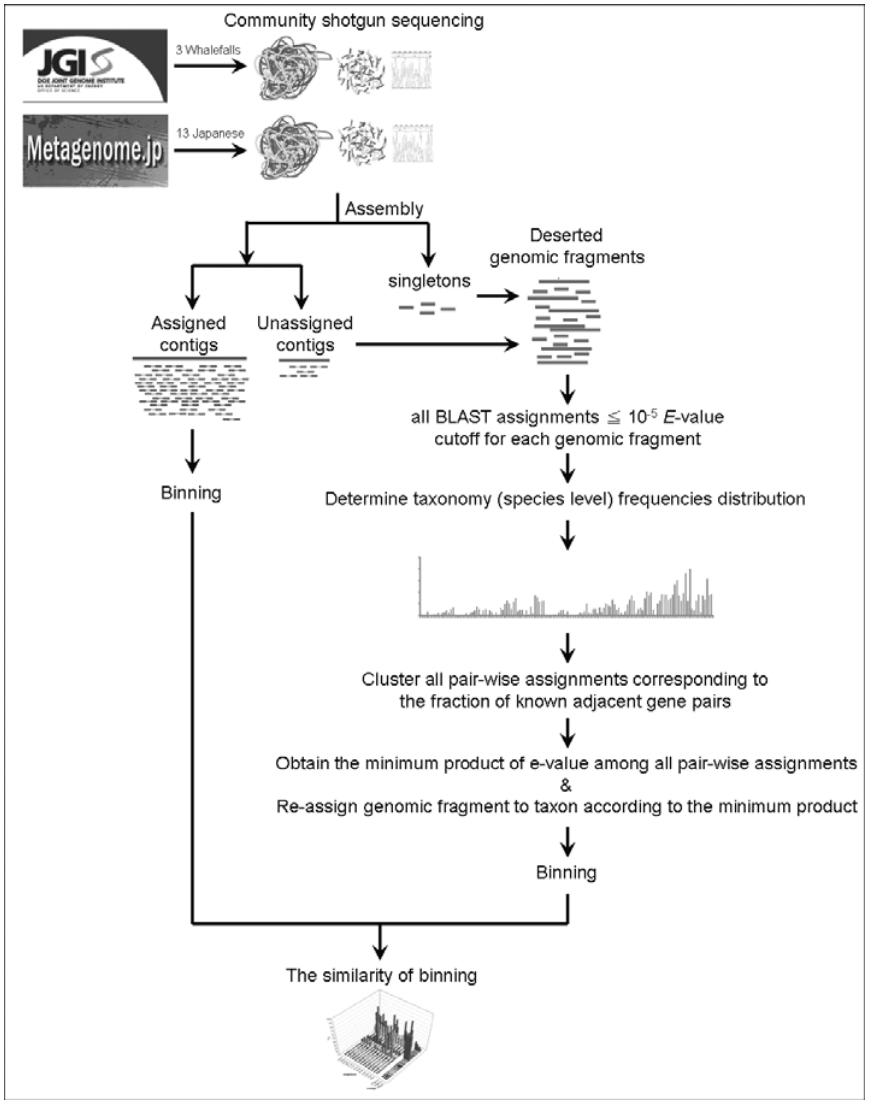


**Fig. 1.** Overview of the proposed approach

method. Taxonomic reassignment for simMC singletons was evaluated by comparing the assignments made by our method to those of the real corresponding taxa in different taxonomic ranks (i.e., species, genus, family, order, class, phylum and superkingdom). The capability of correct assignment was taken as a measure of sensitivity [30, 31]:

$$sens_{rank} = \frac{TP_{rank}}{TP_{rank} + FN_{rank}}. \quad (1)$$

where $TP_{rank}$ (true positive) denotes correct matches (both adjacent CDSs of a singleton are correctly identified) and $FN_{rank}$ (false negative) indicates cases of overlooked singletons.

In addition, the the reliability of taxonomic assignment was taken as a measure of specificity:

$$spec_{rank} = \frac{TP_{rank}}{TP_{rank} + FP_{rank}}. \quad (2)$$

$FP_{rank}$ (false positive) was measured as follows:

$$FP_{rank} = FP_{CDSs} + FP_{noCDSs}. \quad (3)$$

where $FP_{CDSs}$ denotes incorrect assignment of a singleton that contains CDSs, and $FP_{noCDSs}$ denotes incorrect assignment of a singleton that contains one CDS or none.

## 3. RESULTS AND DISCUSSIONS

Current taxonomic binning methods have to discard a large number of sequences due to low homology scores. To address this problem, we developed a method that assigns discarded genomic fragments by combining the BLAST search scores and the criterion of gene adjacency. As shown in Table 2, between 9.9% and 11.8% of the discarded contigs in the whale fall samples could be assigned to taxa under the proposed approach. In the group of Japanese individuals, we were able to assigne between 16.0% and 28.9% of the discarded singletons to taxa. We observed that the assignment rate for singletons was higher than for contigs (11.1±1.0% for discarded contigs and 20.8±3.9% for discarded singletons). We reasoned that the contig sequences belonged to sequences with good sequencing quality and depth of coverage; thus, they had a better chance of being assigned.

To validate our approach, we compared the proposed taxonomic binning strategy using discarded

**Table 2.** Summary of reassignments using discarded metagenomic data. The consistency between binning with discarded fragments and that in the original studies was tested by the *Pearson* correlation coefficient (r).

| Data Type I - contigs - | Un-assigned | | r (phylum) | r (family) |
|---|---|---|---|---|
| | Contigs[a] | Rate (%) | | |
| whale fall sub. 1 | 809 | 11.5 | 0.99 | 0.88 |
| whale fall sub. 2 | 761 | 9.9 | 0.98 | 0.79 |
| whale fall sub. 3 | 590 | 11.8 | 0.99 | 0.78 |
| Data Type I - contigs - | Un-assigned | | r (phylum) | r (family) |
| | Singletons | Rate (%) | | |
| Japanese In-A | 3125 | 23.3 | 0.97 | 0.86 |
| Japanese In-B | 1851 | 26.2 | 0.98 | 0.87 |
| Japanese In-D | 4966 | 16.9 | 0.87 | 0.81 |
| Japanese In-E | 2555 | 23.6 | 0.96 | 0.94 |
| Japanese In-M | 1936 | 22.9 | 0.98 | 0.88 |
| Japanese In-R | 3470 | 16.0 | 0.92 | 0.85 |
| Japanese F1-S | 2721 | 17.7 | 0.91 | 0.85 |
| Japanese F1-T | 3784 | 17.4 | 0.94 | 0.74 |
| Japanese F1-U | 3402 | 28.9 | 0.99 | 0.99 |
| Japanese F2-V | 3365 | 17.1 | 0.94 | 0.73 |
| Japanese F2-W | 3493 | 20.6 | 0.88 | 0.85 |
| Japanese F2-X | 3817 | 19.7 | 0.92 | 0.80 |
| Japanese F2-Y | 4141 | 20.6 | 0.96 | 0.90 |

datasets with the strategies in previous studies [9, 10, 12]. We used *Pearson* correlation coefficient to evaluate the similarity of the two groups. We found that the results derived by our taxonomic binning strategy and those reported in previous studies were consistent; the correlation coefficients were 0.94±0.03 in the phylum rank and 0.85±0.06 in the family rank (Table 2). The consistency between the two datasets indicates that taxonomic binning using discarded data is as representative as the binning strategies used in previous studies.

To further evaluate our approach, we used 10,000 simulated singletons (simMC) for taxonomic binning. As shown in Table 3, the singletons were correctly assigned with sensitivity between 63.1-59.4% and specificity between 77.3-66.1% from superkingdom to species. As expected, these two indexes declined from the superkingdom to the species rank. The results indicate that our system's performance should be reliable.

**Table 3.** Summary of system performance. The sensitivity and specificity of taxonomic assignments were measured using 10,000 simulated singletons randomly selected from the simMC dataset.

| Taxonomic rank | Number of singletons | Sensitivity (%) | Specificity (%) |
|---|---|---|---|
| Species | | 59.4 | 66.1 |
| Genus | | 60.4 | 69.2 |
| Family | | 60.5 | 69.4 |
| Order | 10,000 | 60.8 | 70.1 |
| Class | | 60.5 | 69.2 |
| Phylum | | 62.2 | 74.5 |
| Superkingdom | | 63.1 | 77.3 |

## 4. CONCLUSIONS

Because a large amount of metagenomic data fails to satisfy the cut-off for taxonomic binning, we introduce a criterion based on a genomic feature, namely, the conservation of gene adjacency between prokaryotes. Our analysis suggests that considering the conserved neighboring gene adjacency could help reduce the amount of data discarded by current methods when analyzing metagenomes.


## Acknowledgments

This work was supported by the National Science Council of Taiwan under grants NSC97-2923-B-001-001-MY2, NSC96-2621-B-001-008-MY3 to DW, and grant NSC98-2221-E-001-015, NSC98-2627-B-001-003 to HKT.


## References


1. Vieites, J.M., et al., *Metagenomics approaches in systems microbiology.* FEMS Microbiol Rev, 2009. **33**(1): p. 236-55.
2. Hugenholtz, P. and G.W. Tyson, *Microbiology: Metagenomics.* Nature, 2008. **455**(7212): p. 481-483.
3. Pignatelli, M., et al., *Metagenomics reveals our incomplete knowledge of global diversity.* Bioinformatics, 2008. **24**(18): p. 2124-2125.
4. Tringe, S.G. and E.M. Rubin, *Metagenomics: DNA sequencing of environmental samples.* Nat Rev Genet, 2005. **6**(11): p. 805-814.
5. Biddle, J.F., et al., *Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment.* Proceedings of the National Academy of Sciences, 2008. **105**(30): p. 10583-10588.
6. Hooper, S.D., et al., *A Molecular Study of Microbe Transfer between Distant Environments.* PLoS ONE, 2008. **3**(7): p. e2607.
7. Turnbaugh, P.J., et al., *The Human Microbiome Project.* Nature, 2007. **449**(7164): p. 804-810.
8. Fraser, C., et al., *The Bacterial Species Challenge: Making Sense of Genetic and Ecological Diversity.* Science, 2009. **323**(5915): p. 741-746.
9. Gill, S.R., et al., *Metagenomic Analysis of the Human Distal Gut Microbiome.* Science, 2006. **312**(5778): p. 1355-1359.
10. Kurokawa, K., et al., *Comparative Metagenomics Revealed Commonly Enriched Gene Sets in Human Gut Microbiomes.* DNA Res, 2007. **14**(4): p. 169-181.
11. Nasidze, I., et al., *Global diversity in the human salivary microbiome.* Genome Res, 2009. **19**(4): p. 636-43.
12. Tringe, S.G., et al., *Comparative Metagenomics of Microbial Communities.* Science, 2005. **308**(5721): p. 554-557.

13. Venter, J.C., et al., *Environmental Genome Shotgun Sequencing of the Sargasso Sea.* Science, 2004. **304**(5667): p. 66-74.
14. Yooseph, S., et al., *The Sorcerer II Global Ocean Sampling Expedition: Expanding the Universe of Protein Families.* PLoS Biol, 2007. **5**(3): p. e16.
15. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
16. Bork, P. and E.V. Koonin, *Predicting functions from protein sequences--where are the bottlenecks?* Nat Genet, 1998. **18**(4): p. 313-8.
17. Chen, K. and L. Pachter, *Bioinformatics for whole-genome shotgun sequencing of microbial communities.* PLoS Comput Biol, 2005. **1**(2): p. 106-12.
18. Garcia Martin, H., et al., *Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities.* Nat Biotechnol, 2006. **24**(10): p. 1263-9.
19. Krause, L., et al., *Phylogenetic classification of short environmental DNA fragments.* Nucl. Acids Res., 2008. **36**(7): p. 2230-2239.
20. Huynen, M.A. and P. Bork, *Measuring genome evolution.* Proc Natl Acad Sci U S A, 1998. **95**(11): p. 5849-56.
21. Overbeek, R., et al., *The use of gene clusters to infer functional coupling.* Proceedings of the National Academy of Sciences of the United States of America, 1999. **96**(6): p. 2896-2901.
22. Dandekar, T., et al., *Conservation of gene order: a fingerprint of proteins that physically interact.* Trends in Biochemical Sciences, 1998. **23**(9): p. 324-328.
23. Tamames, J., et al., *Conserved clusters of functionally related genes in two bacterial genomes.* J Mol Evol, 1997. **44**(1): p. 66-73.
24. Tamames, J., *Evolution of gene order conservation in prokaryotes.* Genome Biology, 2001. **2**(6): p. research0020.1 - research0020.11.
25. Mushegian, A.R. and E.V. Koonin, *Gene order is not conserved in bacterial evolution.* Trends in Genetics, 1996. **12**(8): p. 289-290.
26. Palleja, A., E. Harrington, and P. Bork, *Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions?* BMC Genomics, 2008. **9**(1): p. 335.
27. Fukuda, Y., Y. Nakayama, and M. Tomita, *On dynamics of overlapping genes in bacterial genomes.* Gene, 2003. **323**: p. 181-7.
28. Rogozin, I.B., et al., *Congruent evolution of different classes of non-coding DNA in prokaryotic genomes.* Nucleic Acids Res, 2002. **30**(19): p. 4264-71.
29. Huson, D.H., et al., *MEGAN analysis of metagenomic data.* Genome Res, 2007. **17**(3): p. 377-86.
30. Hoff, K., et al., *Gene prediction in metagenomic fragments: A large scale machine learning approach.* BMC Bioinformatics, 2008. **9**(1): p. 217.
31. Diaz, N., et al., *TACOA - Taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach.* BMC Bioinformatics, 2009. **10**(1): p. 56.