JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 19, Number 2, 2012

Pp. 241–249
DOI: 10.1089/cmb.2011.0276

MetaCluster 4.0: A Novel Binning Algorithm for NGS Reads and Huge Number of Species

YI WANG, HENRY C.M. LEUNG, S.M. YIU, and FRANCIS Y.L. CHIN

ABSTRACT

Next-generation sequencing (NGS) technologies allow the sequencing of microbial communities directly from the environment without prior culturing. The output of environmental DNA sequencing consists of many reads from genomes of different unknown species, making the clustering together reads from the same (or similar) species (also known as *binning*) a crucial step. The difficulties of the binning problem are due to the following four factors: (1) the lack of reference genomes; (2) uneven abundance ratio of species; (3) short NGS reads; and (4) a large number of species (can be more than a hundred). None of the existing binning tools can handle all four factors. No tools, including both AbundanceBin and MetaCluster 3.0, have demonstrated reasonable performance on a sample with more than 20 species. In this article, we introduce MetaCluster 4.0, an unsupervised binning algorithm that can accurately (with about 80% precision and sensitivity in all cases and at least 90% in some cases) and efficiently bin short reads with varying abundance ratios and is able to handle datasets with 100 species. The novelty of MetaCluster 4.0 stems from solving a few important problems: how to divide reads into groups by a probabilistic approach, how to estimate the 4-mer distribution of each group, how to estimate the number of species, and how to modify MetaCluster 3.0 to handle a large number of species. We show that Meta Cluster 4.0 is effective for both simulated and real datasets. Supplementary Material is available at www.liebertonline.com/cmb.

Key words: binning, environmental genomics, metagenomics.

1. INTRODUCTION

Analysis of the collective genomes of all microorganisms from an environmental sample (also known as *metagenomics*, *environmental genomics*, or *community genomics*) is an important research area. For example, the diversity of microbes in humans is found to be associated with some common diseases such as gastrointestinal disturbance (Khachatryan et al., 2008) and inflammatory bowel disease (IBD) (Qin et al., 2010). High-throughput next-generation sequencing (NGS) techniques enable researchers to directly sequence the genomes of multiple species obtained from such an environmental sample for analysis. There are numerous successful metagenomic projects based on NGS technologies (Costello et al., 2009; Grice et al., 2009; Hamady and Knight, 2009; Qin et al., 2010; Rusch et al., 2007; Tyson et al., 2004). An important step in

Department of Computer Science, The University of Hong Kong, Hong Kong, China.

metagenomic analysis is to group DNA fragments (or *reads*) from similar species together (known as *binning* or *clustering*). However, there are several factors that make this binning problem difficult.

Lack of reference genomes

Most bacteria (up to 99%) found in environmental samples are unknown (Eisen, 2007) and cannot be cultured and separated in the laboratory (Amann et al., 1990). Traditional binning methods align reads against reference genomes and assign reads aligned to similar genomes in a group. The process is time consuming and many reads cannot be aligned to known genomes, as $<1\%$ sequences of microorganisms are known (Koski and Golding, 2001).

Some algorithms (Brady and Salzberg, 2009; McHardy et al., 2006) group DNA fragments using taxo nomic markers (e.g., 16S rRNA [Cole et al., 2005], recA, and rpoB) to classify fragments into different classes (Navlakha et al., 2009; Wu and Eisen, 2008) or as constraints for semi-supervised clustering or classification. However, these generic features pertain to only a small percentage ($<1\%$) of the reads (Garcia Martin et al., 2008). It has also been reported (Case et al., 2007) that multiple markers may be shared by some species and multiple markers may exist in the same species. Thus, the reliability of these features is in doubt.

Recent binning tools (Chatterji et al., 2008; Kelley and Salzberg, 2010; Prabhakara and Acharya, 2011; Teeling et al., 2004a,b; Wu and Ye, 2011; Yang et al., 2010a,b) are based on the observation that the q-mer (length-q substrings of a fragment) distributions of the DNA fragments from the same genome are more similar than those from different genomes. Thus, without using any reference genomes (i.e., *unsupervised*), one can determine if two fragments are from genomes of similar species based on their q-mer distributions. Unfortunately, none of these binning tools can solve all the following issues.

Uneven abundance ratio

The proportion in which a species exists in a sample is called *abundance ratio*. Most of the tools can only handle species with even abundance ratios, and their binning performances degrade significantly in real situations when the abundance ratios of the species are different. AbundanceBin (Wu and Ye, 2011) bins reads based on the coverage of their q-mers and works for very different abundance ratios. But problems arise when some species have similar abundance ratios. MetaCluster 3.0 (Leung et al., 2011) tries to group the reads into many small clusters so that reads from minority species (with low abundance ratios) could exist as isolated clusters. Then it merges clusters that are likely to be from the same species. However, it requires the fragment length to be long and cannot work directly on short NGS reads.

Short read length

NGS technologies usually produce short reads of length 50–150 bp. Algorithms (Kelley and Salzberg, 2010; McHardy et al., 2006; Yang et al., 2010b) that make use of q-mer distribution require fragments to be a lot longer (at least 400 bp) to have a statistically stable distribution; thus, they cannot work directly on short reads.

A large number of species

No existing binning algorithm has demonstrated reasonable performance for a dataset with more than 20 species. The accuracy of these tools drops significantly when the number of species in the sample is large. AbundanceBin does not work well when the number of species in the sample is larger than 4, especially when there exist species with similar abundance ratios (Wu and Ye, 2011). MetaCluster 3.0 (Leung et al., 2011) starts to deteriorate when the number of species in the sample is larger than 10.

Our contribution

We introduce the first binning algorithm, MetaCluster 4.0, which can handle all the above issues and can accurately (up to 80% precision and 85% sensitivity) bin short reads (75 bp) from a sample of up to 100 species of varying abundance ratios unsupervised. MetaCluster 4.0 (1) first, forms groups of short reads likely from the same genome; (2) next, estimates the q-mer distribution of each group; and (3) finally, employs a modified version of MetaCluster 3.0 to bin the groups based on their q-mer distributions. The novelty of our approach stems from a few non-trivial ideas behind the algorithm's technical components.

In the first phase, the grouping is based on long common w-mers (substring of length w) that exist in two different groups and guided by a probabilistic measure. Species with different abundance ratios will end up with similar numbers of groups. Intuitively, majority species (those with higher abundance ratios) will have more reads in each group but will only have slightly more groups than minority species. Hence, there is an effect of balancing the uneven abundance ratios of species.

Secondly, we cannot simply count and sum the q-mers in each read to obtain the q-mer distribution of a group since reads may overlap resulting in double-counting. We introduce a method of estimating the q-mer distribution of a group without double-counting.

Thirdly, MetaCluster 3.0 can perform well when the number of species is known and the abundance ratios of species are balanced. Thus, we use MetaCluster 3.0 iteratively as a tool to accurately estimate the number of species. Then, making use of the estimated number of species, and the balanced number of groups per species we have formed, we can use the MetaCluster 3.0 to bin the groups (of reads) based on their q-mer distributions.

2. METHODS

MetaCluster 4.0 consists of three phases: (1) probabilistic grouping of short reads; (2) q-mer distribution estimation for each group; and (3) binning. Figure 1 shows the overall framework. The details of each phase are described below.

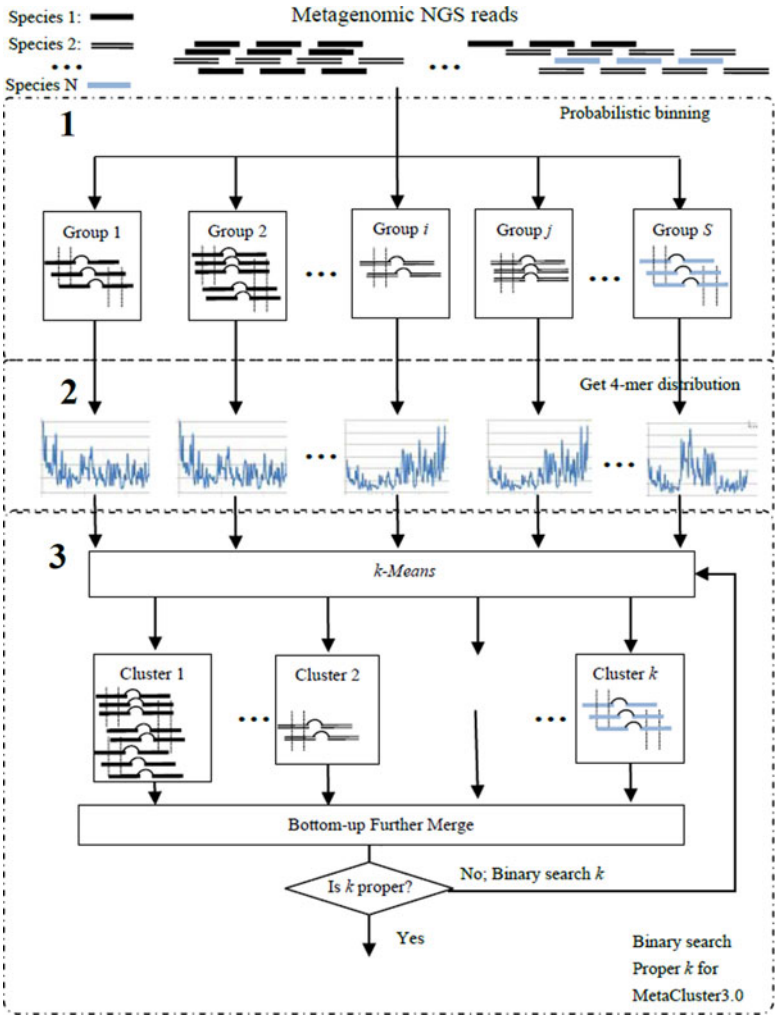

FIG. 1. The pipeline of MetaCluster 4.0.

2.1. Phase 1: probabilistic grouping

In Phase 1, each read is considered as a group and are progressively merged as long as a common w-mer (substring of length w) occurs in both groups and the likelihood of false-positive merging (merging reads from different genomes) is below a threshold p. The idea of grouping reads before binning is important so as to compensate for the short length of the reads. The probability that two reads from different genomes share a common w-mer is very small when $w \geq 35$ ($< 0.03\%$ and $< 0.22\%$ if the genomes are from the same family or genus levels, respectively). The smaller w is and the taxonomically closer the genomes are, the higher is this probability. As w increases, the probability that two reads with a common w-mer are from the same genome increases. However, due to sequencing errors, setting w too large may prevent groups, with reads from the same genome, from being merged together. Based on the error rate and coverage, the largest w that would ensure the majority of w-mers would be correctly covered by at least two reads can be computed. The details are given in the Supplementary Material (which is available at www.liebertonline.com/cmb).

Diminishing the effect of uneven abundance ratios. After the probabilistic grouping, each group is regarded as a set of ''virtual'' contigs (although it may not be easy or possible to construct them for each group). Since the probability of having false-positive merging increases with the size of the group, a group will not be merged further if it grows to a certain size. Intuitively, if a species has a higher abundance ratio, there would only be more reads in its group, but there would not be more groups for that species. Thus, each species should have similar number of groups, regardless of its abundance ratio. This observation is verified through analysis (in the Supplementary Material, which is available at www.liebertonline.com/cmb) as well as by the experimental results in Section 3. This observation is critical for handling a large number of species with varying abundance ratios. Since the effect of varying abundance ratios diminished, once the number of species is estimated, no merging step is needed in MetaCluster 3.0.

2.2. Phase 2: 4-mer distribution estimation

After Phase 1, each group represents some DNA regions (a set of virtual contigs) in the genome. The q-mer distribution for each group is estimated in this phase. Following the other studies (Yang et al., 2010b), $q = 4$ is picked in this study. Since the reads in a group may overlap and contain errors, the q-mer distribution cannot be computed trivially by counting the occurrence frequency of each q-mer in the reads directly. A set of ''correct'' r-mers (with large enough r) that are likely to be on the genome is extracted from the reads. Based on the error rate and coverage, a correct r-mer should occur at least a certain number of times, say t, in G. Then, the q-mer distribution of G is estimated by adding up the q-mer distributions of these correct r-mers.

Similar to the value of w in Section 2.1, the accuracy of this estimation depends on the value of r. When r is small, many r-mers introduced by sequencing error may occur t times and are treated as correct (false positive). When r is large, correct r-mers may occur fewer than t times and are treated as incorrect (false negative). A suitable value of r is selected such that the total number of errors (false positive and false negative) is minimized. Details are given in the Supplementary Material (which is available at www.liebertonline.com/cmb).

2.3. Binning: modified MetaCluster 3.0

MetaCluster 3.0 (Leung et al., 2011) is a binning tool that can determine the number of genomes automatically. It is based on an observation that under the Spearman distance definition, 4-mer distributions of fragments from the same genome are more similar than those from different genomes (Yang et al., 2010a). MetaCluster 3.0 consists of two stages: (a) Top-down separation stage groups the reads based on the k-mean clustering algorithm. After the top-down separation, the number of clusters constructed is much larger than the number of species in the dataset in order to handle the uneven abundance ratio problem. (b) Bottom-up merging stage uses two normal distributions to model the Spearman distance between two contigs from the same genome and that from different genomes. Based on the expected distances of these two distributions, MetaCluster 3.0 can determine whether two groups of reads should be merged.

MetaCluster 3.0 guarantees the expected number of errors introduced in each merging step is small. However, when the number of species increases, a large amount of errors will still be accumulated after a

series of merges and MetaCluster 3.0 might fail. The improvement of MetaCluster 4.0 is based on two major observations. Firstly, after Phase 1, MetaCluster 4.0 has diminished the effect of uneven abundance ratios by constructing similar number of groups per species. Thus, the number of merging steps is reduced by constructing smaller number of groups. Secondly, MetaCluster 3.0 performs bottom-up merging when the estimated number of clusters k in the top-down separation stage of MetaCluster 3.0 is larger than the actual number of species. Thus, it can be used to predict the number of species based on a binary search approach. Combining these two observations, MetaCluster 4.0 is able to handle a large number of species with varying abundance ratios effectively.

3. EXPERIMENTS AND RESULTS

In this section, we evaluate the effectiveness of MetaCluster 4.0 on both simulated and real datasets (Qin et al., 2010). Since AbundanceBin is the only available tool that can work on short reads, we compare MetaCluster 4.0 with AbundanceBin. All simulated data are generated based on the genomes in the NCBI database (ftp.ncbi.nih.gov/genomes/). All the experiments are run on a machine with 140-G memory and 16 CPU of Intel Xeon E5620 at 2.40 GHz. Our software is available online (http://i.cs.hku.hk/∼alse/MetaCluster/).

3.1. Experiments on simulated data

Given an abundance ratio for a species with known genome in NCBI, we randomly picked a set of length-75 pair-end reads from the genome with 1% sequencing error and 250 ± 50 bp insert distance with coverage 15× abundance ratio. Minimum coverage of 15 is assumed in practice, as genome with coverage of <15 might be difficult to detect. The performance of MetaCluster 4.0 and AbundanceBin are evaluated based on precision and sensitivity. Assume there are N genomes in the dataset and a binning algorithm outputs M clusters C_i $(1 \le i \le M)$. The overall precision and sensitivity is given as:

$$\text{precision} = \frac{\sum_{i=1}^{M} \max_j R_{ij}}{\sum_{i=1}^{M} \sum_{j=1}^{N} R_{ij}}$$

$$\text{sensitivity} = \frac{\sum_{j=1}^{N} \max_i R_{ij}}{\sum_{i=1}^{M} \sum_{j=1}^{N} R_{ij} + \text{number of unclassified reads}}$$

Note that if $M > N$, the majority of reads in each cluster probably belongs to a single genome and thus precision would be high. However, sensitivity would be low as some genomes are represented by more than one cluster. On the other hand, if $M < N$, some clusters would contain reads from more than one genome and precision would be poor. Thus, precision increases with the number of predicted clusters while sensitivity decreases with the number of predicted clusters.

We generated six datasets each with 20 species, with three for the family-genus level (FG) and three for the genus-species level (GS). A number of random families (or genera for GS level) are selected, and then a random number of species in each will be selected. The detailed information of the datasets is given in Table 1. We compare the performance of MetaCluster 4.0 with AbundanceBin.

Table 1. Datasets of Simulated Data

Dataset	*No. of groups*	*Total no. of species*	*No. of species in each group*	*Abundance ratio*
1a	6 families	20	1,3,3,3,4,6	1:1:1:1:1:1:1:1:1:1:2:2:2:3:3:3:4:4:4
1b	5 families	20	1,2,3,4,10	1:1:1:1:1:1:1:1:1:1:2:2:2:3:3:3:4:4:4
1c	5 families	20	1,5,4,4,6	1 × 4:4 × 4:6 × 4:8 × 4:10 × 4
2a	4 genera	20	5,5,7,3	1:1:1:1:1:1:1:1:1:1:2:2:2:3:3:3:4:4:4
2b	4 genera	20	4,6,4,6	1:1:1:1:1:1:1:1:1:1:2:2:2:3:3:3:4:4:4
2c	4 genera	20	2,6,3,9	1 × 4:4 × 4:6 × 4:8 × 4:10 × 4

We first consider the family-genus level (FG) (i.e., Dataset 1a–1c). Recall that there are three phases in MetaCluster 4.0 and groups are merged in both Phase 1 and 3. The performances of MetaCluster 4.0 after Phases 1 and 3 are shown in Tables 2 and 3. The accuracy of Phase 1 is quite high, and the total error is less than 3% as predicted (in the Supplementary Material, which is available at www.liebertonline.com). The number of groups is larger than the number of genomes in the dataset, as we aimed at reducing the number of false positives with the trade-off of having more groups to be handled in Phase 3. Hence, it is not so meaningful to show the sensitivity here. Recall that groups of size 1 will be removed after Phase 1. From our experiments, the number of removed reads usually is very small ($<0.7\%$). In Phase 3, the number of species is determined by running MetaCluster 3.0 with binary search on k. MetaCluster 4.0 can predict the number of species quite accurately, with the precision and sensitivity higher than 85% and 78%, respectively.

Although AbundanceBin can estimate the number of species in the dataset, its estimation is far from satisfactory, and it takes a very long time to run (over 1 day). Thus, we provide the numbers of species to AbundanceBin in advance and check the performance of AbundanceBin. From Table 3, the accuracy of the binning results by AbundanceBin is quite low, even when the exact number of species in the datasets is given. It mixed reads of species with similar abundance ratios (Dataset 1a and 1b). And it cannot finish binning Dataset 1c in 2 days (Table 4).

We also show the performance of MetaCluster 4.0 on genus-species level (Tables 5 and 6) where the species are more similar with each other when compared with the three datasets above. Since the probability of two reads from different genomes having common k-mer increases when two genomes are in lower taxonomic level, both the precision and sensitivity of MetaCluster 4.0 decrease after Phase 1 and Phase 3. However, the performance is still reasonably good.

Table 2. Performance of MetaCluster 4.0 after Phase 1 for Family-Genus Level

Dataset	*No. of groups*	*Precision*	*No. of removed reads*
Dataset 1a	3,452	98.95%	0.1406%
Dataset 1b	3,758	99.03%	0.1581%
Dataset 1c	8,303	99.29%	0.6547%

Table 3. Comparison of MetaCluster 4.0 and AbundanceBin for Family-Genus Level

Dataset	*No. of groups*	*Precision*	*Sensitivity*	*Precision of AbundanceBin*	*Sensitivity of AbundanceBin*
Dataset 1a	31	92.24%	80.91%	53.29%	17.77%
Dataset 1b	21	90.84%	90.76%	64.42%	15.20%
Dataset 1c	25	87.22%	77.73%	Cannot finish within 48 h	

Table 4. Running Time and Memory Requirement

	Memory requirement		*Running time*	
Dataset	*MetaCluster 4.0*	*AbundanceBin*	*MetaCluster 4.0*	*AbundanceBin*
Dataset 1a	28.9 GB	25 GB	1 h 2 min	18 h 20 min
Dataset 1b	25.6 GB	16.6 GB	1 h 17 min	9 h 40 min
Dataset 1c	55 GB	50 GB	3 h 17 min	>48 h

Table 5. Performance of MetaCluster 4.0 after Phase 1 for Genus-Species Level

Dataset	*No. of groups*	*Precision*	*No. of removed reads*
Dataset 2a	3,175	97.45%	0.1576%
Dataset 2b	3,494	97.32%	0.1728%
Dataset 2c	10,993	98.96%	0.9712%

Furthermore, we show the performance on 100 genomes on genus-species level (Table 7). Dataset 3a consists of 100 different species from 18 different genera. The precision of the merged groups after Phase 1 (Table 8) is slightly lower when compared with Table 5. It is because the probability of false positive merging increases with the number of species in the sample. However, we show that MetaCluster 4.0 can still maintain a precision of 94% after Phase 1 even when the number of species is 100. The number of groups formed is in proportion to the six datasets above, as the probability of false negative merging does not depend much on the number of species in the sample. With 100 species in the sample, MetaCluster 4.0 can still obtain an overall precision of 80.02% and sensitivity of 85.66% for the final output (Table 9). Additional experimental results such as for extreme abundance ratios and different number of species with different abundance ratios can be found in the Supplementary Material (which is available at www .liebertonline.com/cmb).

Table 6. Performance of MetaCluster 4.0 after Phase 3 for Genus-Species Level

Dataset	*No. of groups*	*Precision*	*Sensitivity*	*Space cost*	*Time cost*
Dataset 2a	26	86.67%	79.83%	19.9 GB	1 h 8 min
Dataset 2b	23	90.87%	88.26%	23.0 GB	1 h 10 min
Dataset 2c	19	80.71%	86.48%	53.0 GB	2 h 51 min

Table 7. Datasets of 100 Species

Dataset	*Groups*	*Species*	*No. of species in each group*	*Abundance ratio*
3a	18 genera	100	$3,4 \times 7,5 \times 3,6 \times 2,7 \times 2,9 \times 2,10$	Equal abundance ratio

Table 8. Performance of MetaCluster 4.0 after Phase 1 for 100 Species

Dataset	*No. of groups*	*Precision*	*No. of removed reads*
Dataset 3a	11,883	94.23%	0.2919%

Table 9. Performance of MetaCluster 4.0 after Phase 3 for 100 Species

Dataset	*No. of groups*	*Precision*	*Sensitivity*	*Space cost*	*Time cost*
Dataset 3a	113	80.02%	85.41%	66 GB	4 h 16 min

Table 10. Precision on Real Data

Groups	*Precision (%)*	*No. of reads(M)*	*Major species*	*No. of reads from the same genus*
group1	97	2.98	*Bacteroides* sp. 2_1_7	99.1
group2	77	2.84	*Bacteroides uniformis*	97.6
group3	45	0.54	*Bacteroides uniformis*	57.8
group4	76	0.76	*Ruminococcus bromii* L2-63	88.3
group5	78	0.58	*Clostridium* sp. SS2-1	77.6
group6	68	1.06	*Bacteroides thetaiotaomicron* VPI-5482	99.4
group7	68	1.56	*Parabacteroides merdae*	68.5
group8	98	0.74	*Alistipes putredinis*	98.1
group9	90	0.92	*Alistipes putredinis*	90.5
group10	68	0.76	*Parabacteroides merdae*	68.1
group11	66	10.2	*Bacteroides vulgatus* ATCC 8482	97.6

3.2. Experiments on real biological data

Qin et al. (2010) performed a deep sequencing on samples obtained from the feces of 124 European adults using Illumina Genome Analyzer technology. We picked three samples from Denmark with pair-end reads with length around 75 bp and combine them together to construct a dataset with 23M pair-end reads. Qin's article provides the references of 57 frequent microbial genomes in the 124 datasets. As species with low coverage are difficult to detect, we filter out reads not from the most abundant 10 species. We use the software BLAT to map the reads to the 10 reference genomes and allow 5% mismatch (i.e., reads with more than 5% mismatch are filtered out). Then we performed MetaCluster4.0 on the data after filtering. The performance is shown in Table 10.

Without using any reference genome information, MetaCluster 4.0 reported that there are 11 groups in the samples. Among the 10 reference species, MetaCluster4.0 cannot detect two species (which do not appear in the Major Species column in Table 10) because the coverage of these two species is low (average coverages are 9.6 and 10.5, respectively). Nevertheless, we still can get an average precision higher than 70% and very high precision for groups 1 and 8. It seems the precision of some groups are not high (e.g., the largest group, 11). However, we found that 97.6% of reads in group 11 are from the same genus. So the precision in genus level is still high for group 11. We did the same analysis for other groups, which are shown in the fifth column in Table 10. Reads from highly related genomes are easier to be mixed together, which is why we can obtain much higher precision in genus level. For those who want to study species from a certain genus, our method provides promising binning results.

4. CONCLUSION

Binning metagenomics reads remains a crucial step in metagenomics analysis. In this article, we introduce MetaCluster 4.0, an unsupervised binning algorithm for short reads. Our approach can (1) deal with short reads without any prior knowledge; (2) handle genomes of similar abundance ratios as well as extreme ratios; (3) determine the number of genomes automatically; and (4) perform well even when there are much more than 20 genomes in the dataset. We also do not assume any knowledge about the phylogenetic levels of the genomes in the sample. We show that MetaCluster 4.0 works well in both the genus-species and the family-genus levels.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Amann, R.I., Binder, B.J., Olson, R.J., et al. 1990. Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Appl. Environ. Microbiol.* 56, 1919.

Brady, A., and Salzberg, S.L. 2009. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat. Methods* 6, 673–676.

Case, R.J., Boucher, Y., Dahllof, I., et al. 2007. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl. Environ. Microbiol.* 73, 278.

Chatterji, S., Yamazaki, I., Bai, Z., et al. 2008. CompostBin: a DNA composition-based algorithm for binning environmental shotgun reads. *Lect. Notes Bioinform.* 4955, 17–28.

Cole, J., Chai, B., Farris, R., et al. 2005. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res.* 33, D294.

Costello, E.K., Lauber, C.L., Hamady, M., et al. 2009. Bacterial community variation in human body habitats across space and time. *Science* 326, 1694.

Eisen, J.A. 2007. Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. *PLoS Biol.* 5, e82.

Garcia Martín, H., Ivanova, N., Kunin, V., et al. 2006. Metagenomic analysis of phosphorus removing sludge communities. *Nature Biotechnology* 24, 1263–1269.

Grice, E.A., Kong, H.H., Conlan, S., et al. 2009. Topographical and temporal diversity of the human skin microbiome. *Science* 324, 1190.

Hamady, M., and Knight, R. 2009. Microbial community profiling for human microbiome projects: tools, techniques, and challenges. *Genome Res.* 19, 1141.

Kelley, D.R., and Salzberg, S.L. 2010. Clustering metagenomic sequences with interpolated Markov models. *BMC Bioinform.* 11, 544.

Khachatryan, Z.A., Ktsoyan, Z.A., Manukyan, G.P., et al. 2008. Predominant role of host genetics in controlling the composition of gut microbiota. *PLoS ONE* 3, e3064.

Koski, L.B., and Golding, G.B. 2001. The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.* 52, 540–542.

Leung, H., Yiu, S., Yang, B., et al. 2011. A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. *Bioinformatics* 27, 1489.

McHardy, A.C., Martín, H.G., Tsirigos, A., et al. 2006. Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods* 4, 63–72.

Navlakha, S., White, J., Nagarajan, N., et al. 2009. Finding biologically accurate clusterings in hierarchical tree decompositions using the variation of information. *Lect. Notes Comput. Sci.* 5541, 400–417.

Prabhakara, S., and Acharya, R. 2011. A two-way multi-dimensional mixture model for clustering metagenomic sequences. Presented at the ACM Conference on Bioinformatics, Computational Biology and Biomedicine 2011, Chicago.

Qin, J., Li, R., Raes, J., et al. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65.

Rusch, D.B., Halpern, A.L., Sutton, G., et al. 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* 5, e77.

Teeling, H., Meyerdierks, A., Bauer, M., et al. 2004a. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.* 6, 938–947.

Teeling, H., Waldmann, J., Lombardot, T., et al. 2004b. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinform.* 5, 163.

Tyson, G.W., Chapman, J., Hugenholtz, P., et al. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428, 37–43.

Wu, M., and Eisen, J.A. 2008. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.* 9, 1–11.

Wu, Y.W., and Ye, Y. 2011. A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *J. Comput. Biol.* 18, 523–534.

Yang, B., Peng, Y., Leung, H., et al. 2010a. MetaCluster: unsupervised binning of environmental genomic fragments and taxonomic annotation. *Proc. First ACM Int. Conf. Bioinform. Comput. Biol.* 170–179.

Yang, B., Peng, Y., Leung, H., et al. 2010b. Unsupervised binning of environmental genomic fragments based on an error robust selection of l-mers. *BMC Bioinform.* 11, S5.

Address correspondence to:
Dr. Francis Y.L. Chin
CB301A, Chow Yei Ching Building
The University of Hong Kong
Pokfulam Road
Hong Kong

E-mail: chin@cs.hku.hk