A1 Observations and Mathematical Details

A1.1 Observation on w-mer shared by different genomes.

We randomly pick 200 pairs of genomes from different species in the same genus (genus level). For each pair, we calculate the percentage of common *w*-mers in each genome, i.e. the value of N/L where N is the number of common *w*-mers in the two genomes and L is the number of distinct *w*-mers in the two genomes. We repeat the measurement by picking up 200 pairs of genomes of species from different genus but in the same family (family level). The percentage of common *w*-mers is less than 1% in most situations when $w \geq 20$ for genus level and the percentage is even smaller for family level (see Figure 2). In both cases, the percentage decreases with the value of *w*.

If two genomes share a fragment of length x(x>w), then x-w+1 w-mers from this fragment are shared by them. It's obvious that larger w leads to smaller x-w+1 (i.e. less number of shared w-mers). That's why the percentage decreased with the value of w in Figure 2.

If two genomes are from same genus rather than different genera, they are genetic closer. So they are supposed to have more common regions, which mean they will share more w-mers. That's why genomes from genus level share more w-mers than those from family level as shown in Figure 2.

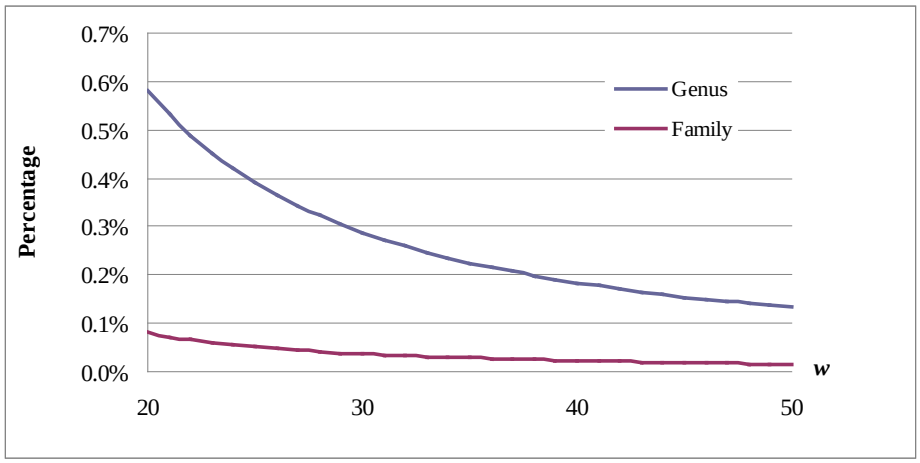

Fig. 1. The percentage of common *w-mer*s for different values of *w* in 200 pairs of randomly picked genomes from the genus level and family level.

A1.2 Upper bound for w in Phase 1.

The binning results produced in the first phase are largely affected by the value of *w* as shown in Figure 2. From the perspective of false-positive merging, it is always better to set a larger *w*. However, we cannot set *w* to be very large; otherwise, very few groups can be merged even when they contain reads from the same genome. This will make the groups contain too few reads for the *q*-mer distribution estimation for Phase 2. In this section, we try to provide a probabilistic measure to find the largest *w* with allowable errors so as to ensure only groups of reads from the same genome are merged.

Although sequencing errors usually occur at the end of a read in real situations, for simplicity we assume that the sequencing errors occur uniformly on a read with error rate *e* in the following calculations. Thus, by setting a relatively large error rate, the calculated probability of false-negative merging is an upper bound on the probability that two groups of reads from the same genome fail to merge.

Let the sequencing depth be *d*, the read length be *l* and the sequencing error rate be *e*. Consider a particular *w*-mer σ in a genome, there are $l - w + 1$ reads can cover σ. Thus the probability that a read covering a particular position of σ, say the starting position of the σ, contains σ as one of its *w*-mers with 0 or 1 error is

$$p = \frac{l-w+1}{l} \cdot [(1-e)^w + we(1-e)^{w-1}] \qquad (1)$$

If there are d reads covering a particular position in a genome, the probability that a w-mer σ being covered correctly by at least two reads is

$$1-(1-p)^d - (1-p)^{d-1}dp \qquad (2)$$

In the first phase, sharing common w-mers is necessary for group merging. So we expect that more than 99% of the w-mers should be covered correctly by at least two reads. For example, assume the read length l is 75, the error rate e is 1% and the sequencing depth d is 15. To achieve at least 99% of w-mers being covered correctly by at least two reads, the value of w should be at most 36.

A1.3 Details of Phase 1

Initially, each pair-end read is considered as one group (if the input is not pair-end data, each single read is treated as a group). We repeatedly merge two groups of reads with a common w-mer if the probability of an incorrect merging is smaller than a threshold p. In order to ensure the merging is correct, we also require that the two reads which induce the merging, besides having a common w-mer, to be similar (within allowable error) throughout the region they overlap, which begins at the end of one read, contains the w-mer and extends to the end of the other read. Note that the smaller the number of occurrences of a w-mer in all reads, the higher the chance that the two reads sharing this w-mer should belong to the same genome. Thus, we perform the merging step in increasing order of the occurrence frequencies of the common w-mers in the dataset, i.e., those groups with lower frequencies of common w-mers are merged first until no more groups can be merged. The probability of incorrect merging also increases with the size of the group (the number of distinct w-mers) and thus the group will stop merging when the size is too large (~8k) and as a consequence each group will be of similar size. Finally, the groups containing only a single pair of reads (or one single read if the input is not pair-end reads), i.e. those that cannot be merged with other groups, will be removed because these reads contain too many errors. From the experiments, less than 0.2% of reads are removed in this phase. Note that, since there are sequencing errors, we do not require the common w-mers to be identical but allow some error e. Based on the calculations in Section A1.2 in the Appendix, we set w = 36 and the threshold p = 0.03 (when we tried a few other similar values, the results were similar) in our experiments. The error e is set to be 3% (i.e. at most one mismatch is allowed in a 36-mer).

A1.4 Probability of false-positive merging.

In the subsection, we show how to estimate the probability of merging two groups of reads incorrectly (false positive), i.e. they are from different genomes. To simplify the analysis, our estimation is based on the following assumptions. The percentage of common w-mers, which are from two different genomes, follows the distribution as given in Figure 2. The abundance ratio of each species is the same, the number of species is known, each w-mer occurs at most once in each genome and each common w-mer occurs in at most two genomes. Although these assumptions might not be realistic, our estimation provides an upper bound on the probability of erroneous false-positive merging.

Let the read length be l, the sequencing depth of each genome be d and the length of each genome be L. Consider two groups G_1 and G_2 of reads containing s_1 and s_2 distinct w-mers from two genomes A and B respectively, and suppose G_1 and G_2 have a common w-mer c which occurs x times among all input reads in the dataset. Let

a. E_{diff} be the event that A ≠ B.
b. E_{common} be the event that G_1 and G_2 have at least one common w-mer c.
c. $E_{\le x}$ be the event that w-mer c occurs at most x times among all input reads in the dataset.

The probability of false-positive merging (merging G_1 and G_2 incorrectly) is

$$\begin{aligned}
&\Pr(E_{diff} \mid E_{common} \wedge E_{\le x}) \\
&= \frac{\Pr(E_{diff} \wedge E_{common} \wedge E_{\le x})}{\Pr(E_{diff} \wedge E_{common} \wedge E_{\le x}) + \Pr(\neg E_{diff} \wedge E_{common} \wedge E_{\le x})} \\
&= \frac{\Pr(E_{diff})\Pr(E_{common} \mid E_{diff})\Pr(E_{\le x} \mid E_{diff} \wedge E_{common})}{\Pr(E_{diff})\Pr(E_{common} \mid E_{diff})\Pr(E_{\le x} \mid E_{diff} \wedge E_{common}) \\ + \Pr(\neg E_{diff})\Pr(E_{common} \mid \neg E_{diff})\Pr(E_{\le x} \mid \neg E_{diff} \wedge E_{common})}
\end{aligned} \qquad (3)$$

The prior probability $\Pr(E_{diff})$ that A ≠ B is $(g-1)/g$.

Given A ≠ B, the probability that a *w*-mer in genome A also occurs in genome B is N/L (which can be estimated from Figure 2), where N is the number of common *w*-mers in genomes A and B. The expected number of *w*-mers in G_1 also occurring in genome B is s_1N/L. Since the probability that each of these s_1N/L *w*-mers occurs in G_2 is s_2/L, the expected number of common *w*-mers in G_1 and G_2 is s_1s_2N/L^2. By the Markov inequality $\Pr[Y \geq \alpha] \leq E[Y]/\alpha$, we have

$$\Pr(E_{common} \mid E_{diff}) \leq s_1 s_2 \frac{N}{L^2}$$

Since the sequencing depth is d, dL/l reads are sampled from each genome and each read contains $(l - w + 1)$ *w*-mers, the total number of *w*-mers sampled from the two genomes is $2dL(l - w + 1)/l$. Since the probability that the *w*-mer c being sampled is $1/L$, the probability that c occurs in G_1 and G_2 at most x times in the dataset can be approximated by the binomial distribution

$$\Pr(E_{\leq x} \mid E_{diff} \wedge E_{common})$$

$$= \sum_{i=2}^{x} \binom{2dL(l-w+1)/l}{i} \left(\frac{1}{L}\right)^i \left(1-\frac{1}{L}\right)^{2dL(l-w+1)/l-i} = F'(x;2N,\frac{1}{L})$$

where $N = dL(l\text{-}w+1)/l$ and $F'(x;n,p)$ is the Binomial cumulative distribution from 2 to x, i.e $F'(x;n,p) = F(x;n,p) - (1-p)^n - np(1-p)^{n-1}$ where $F(x;n,p)$ is the Binomial cumulative distribution.

Similarly, the prior probability that G_1 and G_2 contain reads from the same genome $\Pr(\neg E_{diff})$ is $1/g$. Given G_1 and G_2 contain reads from the same genomes, the probability that G_1 and G_2 have at least one common *w*-mer is

$$\Pr(E_{common} \mid \neg E_{diff}) = 1-\left(1-\frac{s_2}{L}\right)^{s_1} \geq \frac{s_1 s_2}{L} - \binom{s_2}{2}\left(\frac{s_2}{L}\right)^2$$

The probability of a unique *w*-mer c occurring in one genome sample at most x times can be approximated by the binomial distribution

$$\Pr(E_{\leq x} \mid \neg E_{diff} \wedge E_{common}) = \sum_{i=2}^{x} \binom{dL(l-w+1)/l}{i} \left(\frac{1}{L}\right)^i \left(1-\frac{1}{L}\right)^{dL(l-w+1)/l-i} = F'(x;N,\frac{1}{L})$$

By substituting the above probability in Equation (3) and simplifying, we have

$$\Pr(\text{false positive})$$

$$\leq \frac{(g-1)s_1 s_2 \frac{N}{L^2} \cdot F'(x;2N,\frac{1}{L})}{(g-1)s_1 s_2 \frac{N}{L^2} \cdot F'(x;2N,\frac{1}{L}) + \left[\frac{s_1 s_2}{L} - \binom{s_1}{2}\left(\frac{s_2}{L}\right)^2\right] \cdot F'(x;N,\frac{1}{L})}$$

$$\leq \frac{(g-1)\frac{N}{L} \cdot F'(x;2N,\frac{1}{L})}{(g-1)\frac{N}{L} \cdot F'(x;2N,\frac{1}{L}) + \left[\frac{1}{L} - \frac{(s_1-1)s_2}{2}\right] F'(x;N,\frac{1}{L})} \qquad (4)$$

The effect of each parameter is summarized in Table 7. By setting a reasonably large g (say g = 100), small d (d = 15) and large w (w = 36), we can calculate an upper bound of false positives for different x and group sizes s_1 and s_2 (when s_1 and s_2 are large). In our experiments, we allow MetaCluster 4.0 to merge two groups if the probability of false positive ≤ 3%.

Table 7. Effects of each parameter on probability of false positive merging

Parameter increases	Pr(FP)	Reasons
Number of species g	increases	Prior probability (g – 1)/g increases
Group sizes s_1 and s_2	increases	Denominator decreases as $[1/L - (s_1 - 1)\, s_2/2]$ decreases
Number of occurrence x	increases	$F'(x; 2N, 1/L)$ increases faster than $F'(x; N, 1/L)$ with x
Sequencing depth d	decreases	$F'(x; 2N, 1/L)$ decreases faster than $F'(x; N, 1/L)$ with $N = dL(l\text{-}w+1)/l$
Length of *w*-mer w	decreases	N/L decreases with w (Figure 2)

A1.4 Computing EFN + EFP for Phase 2.

In this subsection, we analyze how to set the value of r and t in Phase 2. MetaCluster 4.0 determines the suitable values of r and t for each group of reads such that an r-mer occurring at least t times in the group is considered as a correct r-mer. The correct r-mers are used to estimate the q-mer distribution. If too many reads covering a particular r-mer have sequencing errors at those positions corresponding to that r-mer, we might not have t copies of that r-mer in a group G and that r-mer will be missed (false negative E_{FN}). On the other hand, if the same sequencing errors occur at the same position of reads covering the same r-mer, an incorrect r-mer (r-mer does not occur in the virtual contig) may have t copies in G and is considered as a correct r-mer (false positive E_{FP}). In this subsection, we calculate the expected number of false negatives and false positives for different values of r and t. Based on this calculation, MetaCluster 4.0 can determine the values of r and t that minimizes the expected total error ($E_{FN} + E_{FP}$).

We assume that sequencing errors occur uniformly in a read and the Hamming distance between any r-mers in the virtual contigs of a group of reads is at least 2. Note that since the length of virtual contigs is about 8000 bp and the length of r-mer is about 16 bp, this assumption is correct in most real data. Let l be the read length and e be the sequencing error rate. The probability that a particular r-mer covered by a read without error is

$$p' = (1-e)^r$$

Given a particular r-mer covered by f reads, the probability that the number of copies of a correct r-mer < t (false negative) is

$$P_{FN} = \sum_{\alpha=0}^{t-1}\left[\binom{f}{\alpha} p'^{\alpha} (1-p')^{f-\alpha}\right] \qquad (5)$$

Thus, the expected number of false negatives per correct r-mer is $E_{FN} = P_{FN}$.

On the other hand, there may be t copies of some incorrect r-mers in G because of sequencing errors. Consider a correct r-mer π in a virtual contig. The probability that a read containing π has another r-mer π_x differed from π at α positions is

$$p(r,\alpha) = (1-e)^{r-\alpha}(e/3)^{\alpha} \qquad (6)$$

Since the sequencing error e is small (<1%), $p(r, \alpha)$ is very small when $\alpha \geq 2$, e.g. when $r = 16$, $p_2 = 9.65 \times 10^{-6}$. Thus, we can assume that if an incorrect r-mer π_α occurs more than once, all its occurrences should be derived from the correct r-mer π with one sequencing error. Thus, the probability that a particular incorrect r-mer π_x occurs at least t times in a group of reads is

$$P_\alpha = \sum_{y=t}^{f}\left[\binom{f}{y} p(r,\alpha)^y (1-p(r,\alpha))^{f-y}\right] \qquad (7)$$

where f is the number of reads covering the corresponding the correct r-mer π. By considering all possible incorrect r-mers, the expected number of false positives per correct r-mer is

$$E_{FP} = \sum_{\alpha=1}^{r}\left(\binom{r}{\alpha} 3^{\alpha} P_\alpha\right) \approx 3rP_1 \qquad (8)$$

For each group G, we assume that every correct r-mer in G has the same read coverage f, which is estimated from the average occurrence frequencies of r-mers except those r-mers with occurrence frequencies in the lowest 5%. We ignore those r-mers with low occurrence frequencies because they may represent incorrect r-mers introduced by sequencing errors and should not be used to estimate the read coverage.

A1.5 Robustness on varying abundance ratio.

In this subsection, we want to show that even if two genomes have quite different abundance ratios, after the probabilistic grouping (Phase 1), the number of groups for each genome will not differ a lot, thus diminishing the effect of uneven abundance ratios. Consider a length-L genome with sequencing depth d. There are $L - w + 1$ w-mer in the genome (assume w is large, say $w = 36$, such that each w-mer uniquely occurs in the genome). However, due to sequencing error, the total number of distinct w-mers $W(d, e, L)$ sampled from the genome should be larger than $L - w + 1$ and the number of distinct w-mers increases with the sequencing depth d and sequencing error rate e. As the groups of reads from a genome usually stop merging when their numbers of distinct w-mers are large (s_1 and s_2 are so large that the probability of false positive is larger than 4% in Equation (2)), the number of distinct w-mers in each group would be similar and the number of groups per genome depends on the value of $W(d, e, L)$.

Consider a w-mer σ sampled in a genome, the probability that there are α specific sequencing errors in σ is $p(w, \alpha)$ (Equation (6) in Section A1.4) and there are $\binom{w}{\alpha}$ w-mers with exactly α mismatches with σ. Thus, the probability that when a read covering σ is sampled and we get another w-mer with α mismatches is $\binom{w}{\alpha} p(w, \alpha)$. By consider all possible w-mers, the probability that another w-mer (instead of σ) being got is

$$\sum_{\alpha=1}^{w} \binom{w}{\alpha} p(w,\alpha) \qquad (9)$$

When the sequencing depth is d, dL/l reads are sampled each contains $(l - w + 1)$ w-mers. The expected number of extra w-mers sampled is at most

$$\frac{dL(l-w+1)}{l} \sum_{\alpha=1}^{w} \binom{w}{\alpha} p(w,\alpha) \qquad (10)$$

This is an approximation because some of the extra w-mers sampled due to sequencing error may also occur in the genome. Thus, the expected value of distinct w-mers sampled from a genome is

$$E(W(d,e,L)) \le L - w + 1 + \frac{dL(l-w+1)}{l} \sum_{\alpha=1}^{w} \binom{w}{\alpha} p(w,\alpha) \qquad (11)$$

Consider two genomes A and B with sequencing depth d_A and d_B respectively, the number of merged groups of reads from genomes A and B is directly proportional to E(W(d_A, e, L)) and E(W(d_B, e, L)). Assume genomes A and B are of length 3×10^6 bp, read length 75, length of w-mer 36 and sequencing error rate 1%, if the sequencing depth of genomes A and B are 1 and 64 respectively, the ratio of E(W(d_B, e, L)) and E(W(d_A, e, L)) is 3.87 only. Thus, the number of merged groups for a genome is not that sensitive to the sequencing depth.

A2 Experiments on estimating *q*-mer distribution

We verify if the estimated q-mer (q = 4) distributions in phase 2 give the same power for distinguishing groups of reads from different genomes. Given a pair of genomes A and B from the same genus, we randomly picked two non-overlapped contigs from A and one contig from B of the same length. We then calculated the 4-mer distributions of these 3 contigs and determine if the contigs from genome A are closer to each other than the contig from genome B in Spearman distance (Figure 3). As mentioned in [26], the accuracy increases with the length of contigs.

In MetaCluster 4.0, we have a group of reads instead of contigs. Thus we randomly select reads with different sequencing depths from the three contigs with 1% error rate and get three groups of reads. We can estimate the 4-mer distributions of these three groups using the calculation in Section A1.3 and determine if the groups from genome A are closer to each other than to the groups from genome B in Spearman distance. We have repeated this experiment for 200 pairs of genomes with different lengths of contig and sequencing depths. Figure 3 shows the results.

The power for distinguishing groups of reads from different genomes increases with the contig length and sequencing depth. Even when the sequencing depth is as low as 15, the accuracy is over 80% when the contig length is longer than 8k bp. Thus, the 4-mer estimation in MetaCluster4.0 provides similar information compared to the 4-mer distribution calculated directly from the contigs and the accuracy is high for merging groups of reads from the same species.

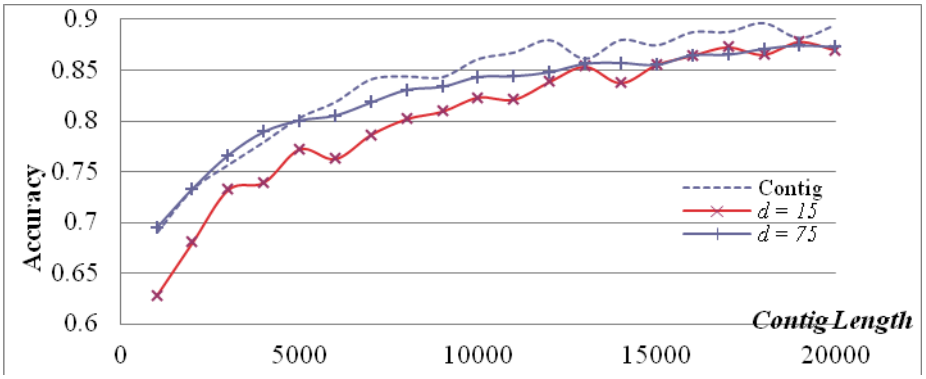

Fig. 2. The accuracy of distinguishing contigs and groups of reads from different genomes using 4-mer distribution and estimated 4-mer distribution, respectively.

A3 Details of datasets of simulated data.

Table 8 Datasets of Simulated Data

Dataset	group #	species #	species # in each group	Abundance Ratio
1a	6 families	20	1,3,3,3,4,6	1×11:2×3:3×3:4×3
1b	5 families	20	1,2,3,4,10	1×11:2×3:3×3:4×3
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	1×11:2×3:3×3:4×3
2b	4 genera	20	4,6,4,6	1×11:2×3:3×3:4×3
2c	4 genera	20	2,6,3,9	1×4:4×4:6×4:8×4:10×4
3a	18 genera	100	3,4×7,5×3,6×2,7×2,9×2,10	Equal abundance ratio
3b	1 genus	7	7	1:2:4:8:16:32:64
3c	7 genera	50	5,6,6,7,7,9,10	1×10:1.5×10:2×10:2.5×10:3×10

A4 Performance of MetaCluster 4.0 on different number of species under different situation.

In this section, we evaluate the performance of MetaCluster 4.0 on different number of species (10, 15, 20, 25, 30, 35, 40) on both FG and GS levels under three different types of abundance ratios (even, random, and uneven). The species are selected as follows. Say, for the GS level, for 10 species, we randomly select 5 genera from NCBI and randomly pick 2 species from each genus. For 15 species, we randomly select 5 genera and 3 species from each genus. For 20, 25, 30, 35, and 40 species, we select 4, 5, 6, 7, and 8 genera, respectively. For FG level, we select families instead of genera. For even abundance, we use the same coverage for all species. For random abundance, we assign coverage to each species randomly in the range of 15 to 75 under uniform distribution. For uneven abundance, we assign the abundance ratio of 5 species from each genus as 1:2:3:4:5 (i.e. the coverage is 15, 30, 45, 60, 75).

As shown in the above, MetaCluster4.0 can achieve precision and sensitivity higher than 80% for all the 42 datasets. Similar as in Section 3.1.1, the precision and sensitivity of FG level datasets are usually higher than the GS level datasets because the species are more similar with each other in the GS level. There is a tendency that the precision and sensitivity decrease slightly when the number of species increases. However, this drop is small because 1) there are limited number of families with at least 8 known genomes from different genera in the NCBI database such that we cannot repeat the experiments many times with different genomes. 2) Different from other binning algorithms, the performance of MetaCluster4.0 does not drop sharply with the number of species. We show the number of predicted groups for each dataset in Table 10. As we can see that the predicted numbers of species are quite close to the actual number of species in most cases. Also, from our experiments, it seems that MetaCluster 4.0 is quite robust for different abundance ratios.

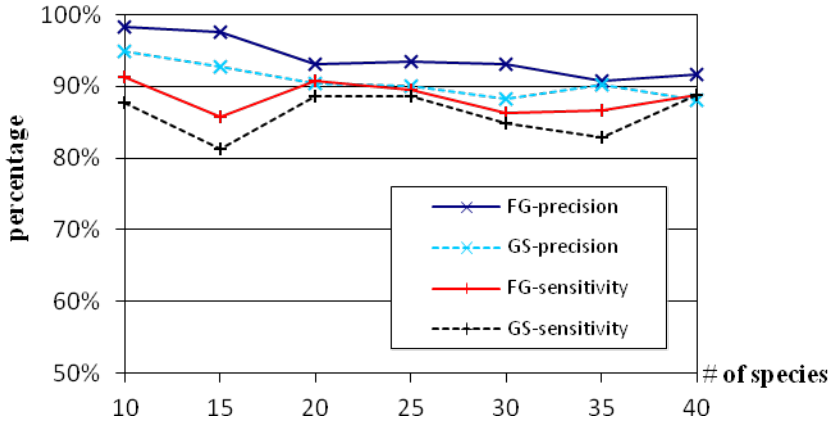

Fig. 3. Performance on datasets with even abundance ratio

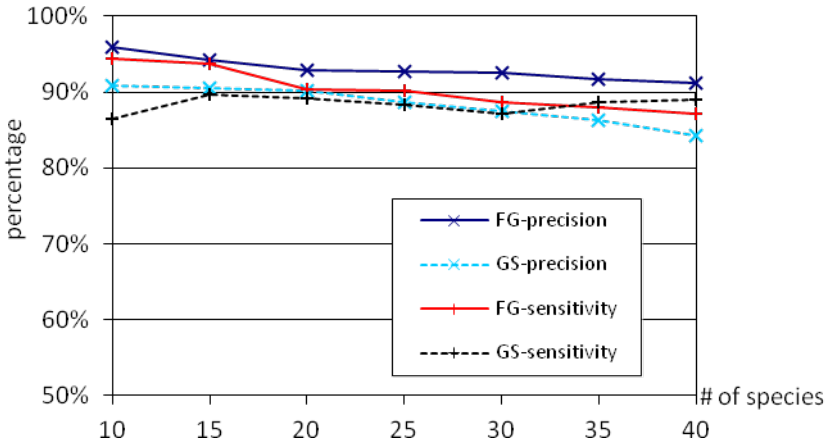

Fig. 4. Performance on datasets with random abundance ratio

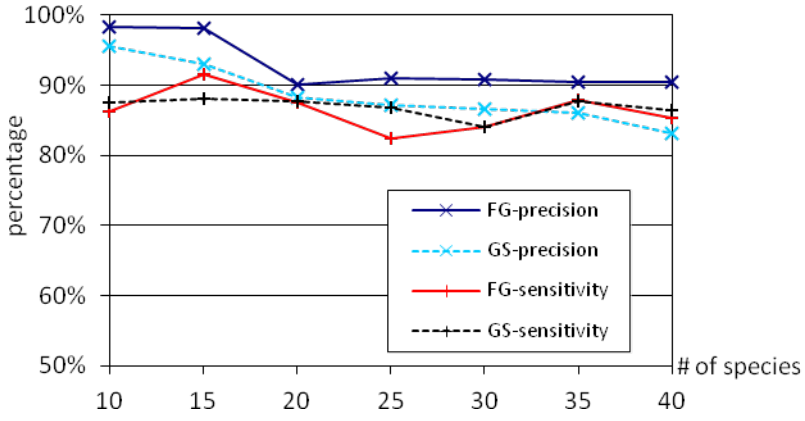

Fig. 5. Performance on datasets with uneven abundance ratio

Table 10 Predicted number of species

species #	Family-Genus level			Genus-Species level		
	Even	Uneven	Random	Even	Uneven	Random
10	16	14	12	16	17	11
15	21	20	18	18	22	16
20	26	21	21	28	24	26
25	25	26	25	37	36	25
30	37	36	34	37	39	38
35	42	38	36	46	48	46
40	43	45	41	47	49	49

A5 More experiments on extreme cases

Table 8 Datasets of extreme cases

Dataset	# of groups	# of species	#. of species in each group	Abundance Ratio
3b	1 genus	7	7	1:2:4:8:16:32:64
3c	7 genera	50	5,6,6,7,7,9,10	1×10:1.5×10:2×10:2.5×10:3×10

Table 9 Performance of MetaCluster 4.0 after Phase 1 for family-genus level

Dataset	#. of groups	Precision	Number of removed reads
Dataset 3b	3077	97.60%	2.4273%
Dataset 3c	11441	97.77%	2.0904%

Table 10 Performance of MetaCluster 4.0 after Phase 3 for genus-species level

Dataset	#. of groups	Precision	Sensitivity	Space cost	Time cost
Dataset 3b	9	91.08%	84.73%	54GB	30h
Dataset 3c	51	80.29%	85.79%	50GB	3h51min

As shown above, the performance of MetaCluster 4.0 on extreme abundance ratios is also good. Dataset 3b contains 7 species from the same genus with extreme abundance ratios 1:2:4:8:16:32:64. On this dataset, we can get precision of 91.08% and sensitivity of 86.8%. From the method section, we can see that the *w*-mer occurrence frequency within a group does not make much difference in Phase 1 as long as the coverage is not very low, say at least 15. Even if the sequencing depths of two genomes are of extreme ratios, the difference in the number of groups formed for each species after Phase 1 will not be very large while the group sizes may differ a lot. For example, with respect to the two species with extreme ratio 1:64 in dataset 3b, the ratio of the number of groups formed after Phase 1 is only about 1:3.3 (comparable to 1:3.8 calculated in Section A1.5) while their genome lengths are about the same ($\sim 3\times10^6$ bp).

Dataset 3c is introduced to show that our method provides acceptable results when the number of species is large and their abundance ratios are uneven. It contains 50 species from 7 genera. As shown in Table 8, we can still obtain precision of 80.29% and sensitivity of 87.58%. Besides the precision and sensitivity, MetaCluster 4.0 can predict the number of species in the sample quite accurately in all experiments.