1 Supplementary material

1.1 Evaluation metric

We use precision, recall, F-score, ARI, and NMI to evaluate the binning performance for the contigs with labels.

- Precision: Evaluate the purity of the bins.
- Recall: Evaluate the completeness of the genome bins.
- F-score: The harmonic mean of precision and recall.
- Adjusted Rand Index (ARI): Evaluate whether contigs belonging to the same genomes are clustered together.
- Normalized Mutual Information (NMI): Using mutual information to measure the overall level of binning performance.

The specific calculation method is as follows. Adopting the notations of COCACOLA (**?**), the classification of pairs of contigs belongs one of the four cases: TP (True Positive) represents the number of pairs of contigs belonging to the same genomes that are clustered into the same clusters, FP (False Positive) represents the number of pairs of contigs belonging to the same genomes that are clustered into different clusters; FN (False Negative) and TN (True Negative) represent the number of pairs of contigs belonging to different genomes that are clustered into the same clusters and different clusters, respectively. If there are N contigs from S genomes being clustered into K clusters, a matrix of $K \times S$ dimensions can be constructed $A = a_{ks}$, a_{ks} represents the number of contigs shared by the k-th cluster and the sth genome. Total number of contig pairs:

$$\binom{N}{2} = \frac{N(N-1)}{2} \tag{1}$$

$$TP = \sum_{k,s} \binom{a_{ks}}{2} \tag{2}$$

$$FP = \sum_{k} \binom{a_{k\cdot}}{2} - \sum_{k,s} \binom{a_{ks}}{2} \tag{3}$$

$$FN = \sum_{s} \binom{a_{\cdot s}}{2} - \sum_{k,s} \binom{a_{ks}}{2} \tag{4}$$

$$\begin{aligned} TN &= \binom{N}{2} - TP - FP - FN \\ &= \binom{N}{2} + \sum_{k,s} \binom{a_{ks}}{2} - \sum_{k} \binom{a_{k\cdot}}{2} - \sum_{s} \binom{a_{\cdot s}}{2} \end{aligned} \tag{5}$$

$$Precision = \frac{1}{N} \sum_{k} \max_{s} \{a_{ks}\} \tag{6}$$

$$Recall = \frac{1}{N} \sum_{s} \max_{k} \{a_{ks}\} \tag{7}$$

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{8}$$

$$ARI = \frac{\sum_{k,s} \binom{a_{ks}}{2} - \frac{\sum_s \binom{a_{\cdot s}}{2} \sum_k \binom{a_{k\cdot}}{2}}{\binom{N}{2}}}{\frac{1}{2}(\sum_s \binom{a_{\cdot s}}{2} + \sum_k \binom{a_{k\cdot}}{2}) - \frac{\sum_s \binom{a_{\cdot s}}{2} \sum_k \binom{a_{k\cdot}}{2}}{\binom{N}{2}}} \tag{9}$$

$$NMI = \frac{2 \times \sum_{k,s} a_{ks} log(\frac{N \times a_{ks}}{a_{k\cdot} \times a_{\cdot s}})}{\sum_s a_{\cdot s} log \frac{a_{\cdot s}}{N} + \sum_k a_{k\cdot}.log \frac{a_{k\cdot}}{N}} \tag{10}$$

1.2 Taxonomy assignment

We used TAXAassign[1] to assign taxonomy to contigs (with ’-p -c 48 -m 98 -q 98 -t 95 -a "60,70,80,95,95,98" -f contig.fasta’ options) on a machine with 4-way 6-core 1.87 GHz Intel Xeon CPUs and 1T memory, where ’-p’ means parallel processing, ’-c’ means the number of the cores, ’-m’ means percentage identity minimum in blastn, ’-q’ means query coverage minimum in blastn, ’t’ means ’consensus threshold’ and ’-a’ means the percentage identity minimum argument at different taxonomic levels. The running times of the TAXAassign on the different datasets are shown in Table S1. Users can apply other suitable taxonomic binning tools to assign taxonomy to contigs and generate must-link constraints according to the assignments for SolidBin-coalign.

Table S1. The running time of the TAXAassign on the different datasets.

Datasets	Time
SpeciesMock	261m10s
StrainMock	83m24s
SimHC	236m0s
Sharon	53m22s
MetaHIT	72m44s

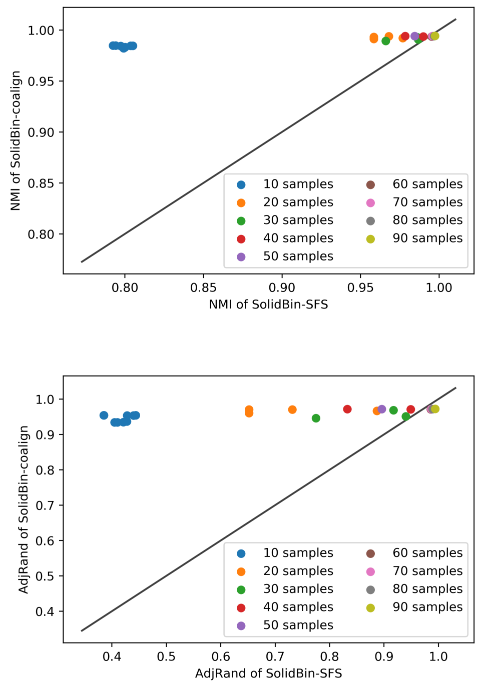

Fig. S1. Evaluation of the results of SolidBin-coalign and SolidBin-SFS on sub-samples of ’SpeciesMock’ dataset.

1.3 Supplementary figures

The results of SolidBin-coalign and SolidBin-SFS on sub-samples of ’SpeciesMock’ dataset are as shown in Figure S1.

[1] https://github.com/umerijaz/taxaassign