Supplementary tables

Tables S1. Comparison of annembed with Trimap for memory and running time.

Dataset	Number of data points	dimension	Peak memory (Trimap)	Peak memory (annembed)	Running time (Trimap)	Running time (annembed)
MNIST FASHION	70,000	748	0.3G	0.9G	3 min 25 s	41.6s
MNIST DIGITS	70,000	748	0.3G	0.3G	3 min 15 s	39.2s
SIFT_1M	1 M	128	4.4G	17G	1 h 03 min	25 min 33s
HIGGS	11 M	20	15.3G	58G	10 h	2 h 39 min

Table S2. LID and hubness estimated by annembed and comparisons with other implementations. 100-NN was used for estimation of LID and hubness based on Euclidean distance.

Dataset	Number of data points	dimension	LID (annembed)[a]	Hubness (annembed)	LID (MLE in Amsaleg et.al.,2015)
MNIST FASHION	70,000	748	22.97 (12.45)	3.28	19.6 (13.9)
MNIST DIGITS	70,000	748	17.5 (7.09)	1.014	15.3 (8.4)
HIGGS	11 M	20	14.9 (5.93)	919.3	--

Table S3. Perplexity Quantile and embedding quality.

Perplexity Quantile	0.05	0.5	0.95	0.99
MNIST FASHION Embedding quality[a]	3.90	5.99	6.00	6.00
GTDB Embedding quality[b]	12	14.9	15	15

[a]Quality was defined by number of true neighbors in embedded space divided by total true neighbors in graph (we use 15, consistent with UMAP).
[b]We use true neihgbors 25 in the genome case because biological database is sparse, and neighbors are not evenly distributed.

Table S4. Running time for t-SNE (single threaded), UWOT (NNS multi-threaded, embedding single threaded) and annembed (fully parallelized) for metagenomic binning, modified in mmgenome2 R package. Tested on a 24-thread machine with R v4.2.1

Algorithm	Running time	NNS Parallelization	Embedding Parallelization
t-SNE	3 min 24s	No	No
UMAP	32.4s	Yes	No
annembed	17.1s	Yes	Yes

Supplementary Figures

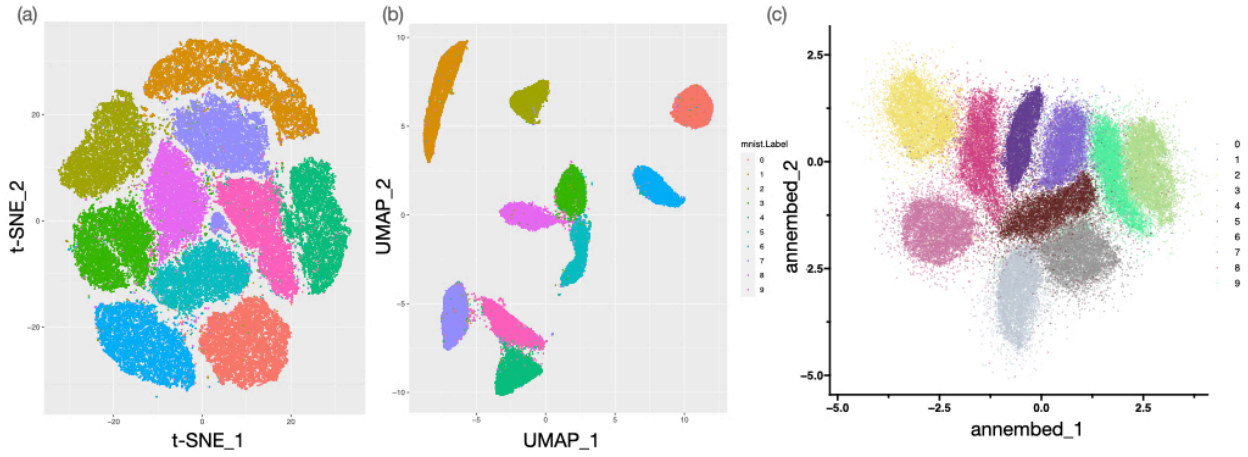

Figure S1. Dimension reduction for t-SNE (a), UMAP (b) and annembed (c) respectively for MNIST-digits dataset. Color legend indicates different labels.

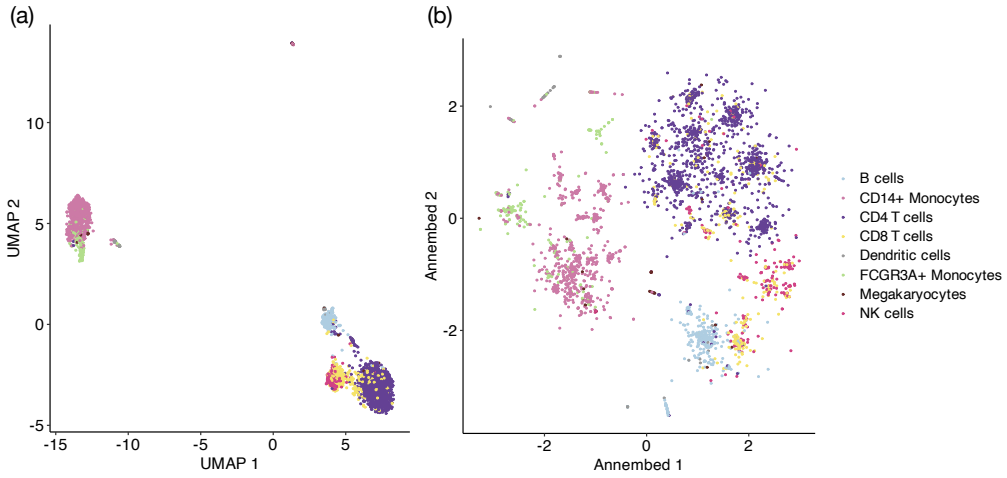

Figure S2. Comparisons between UMAP and annembed for single cell RNA sequencing dataset PBMC (Peripheral Blood Mononuclear Cells). Note that the distances between clusters or cell types do not have meaningful interpretations.

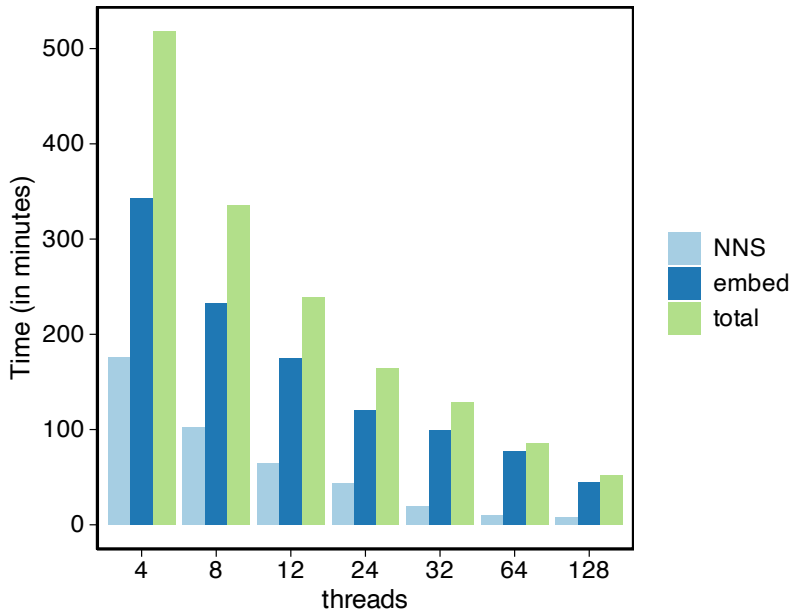

Figure S3. Scalability of annembed with respect to the number of NUMA threads for the HIGGS dataset. Experiments were performed on a 128-threads AMD EPYC 7713 processor (2 64-thread NUMA node). Note that for 128 threads, NNS time does not scale linearly compared to 64 threads because HNSW graph construction is NUMA sensitive.

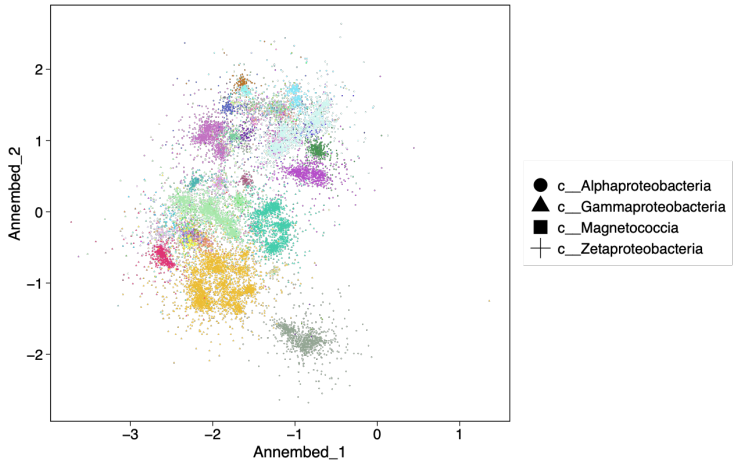

Figure S4. Detailed annembed plot of phylum Proteobacteria (Green in Figure 4). Each shape represents a class while different colors in each class represents different orders.

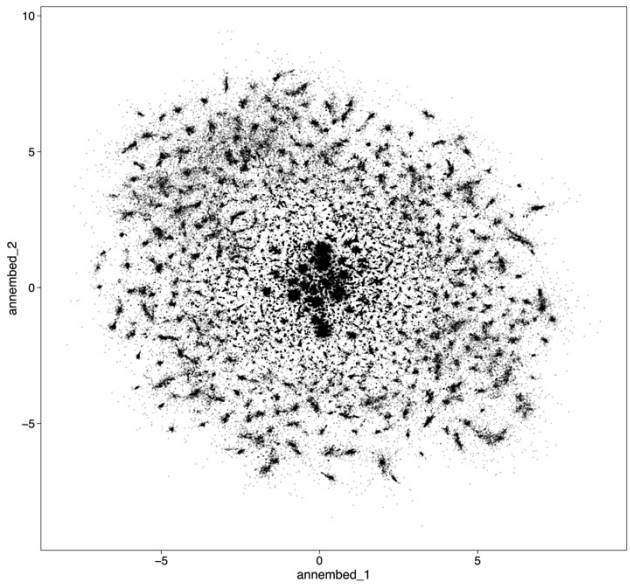

Figure S5. Annembed plot of all NCBI/RefSeq prokaryotic genomes (~318K) using nucleotide genome sequences. Note that each cluster is well separated, representing species to genus level clusters (80%~95% ANI)

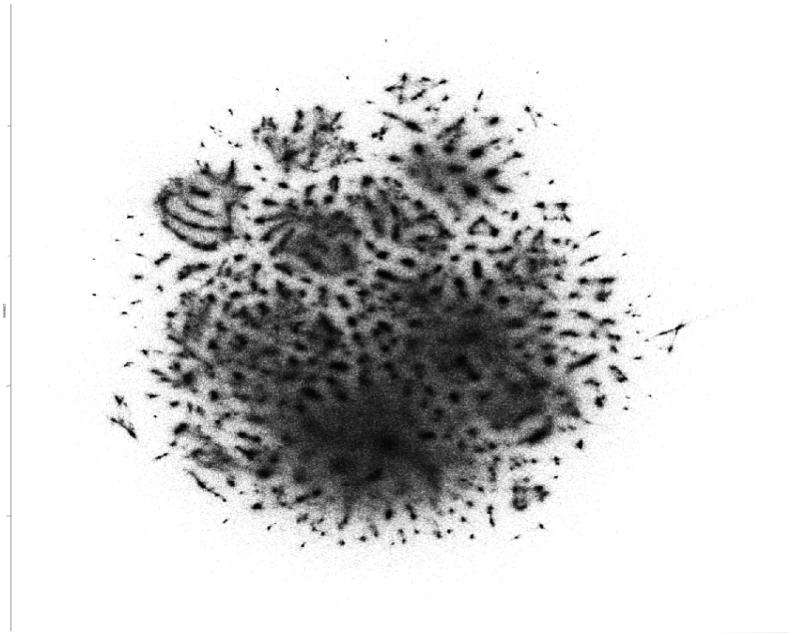

Figure S6. Annembed plot of IMG/VR v4 (total ~3 million genomes). Each cluster represents a phylogenetic group (family or class level) of viral genomes. ProbMinHash distances are based on amino acid sequenced predicted from the genome sequences.

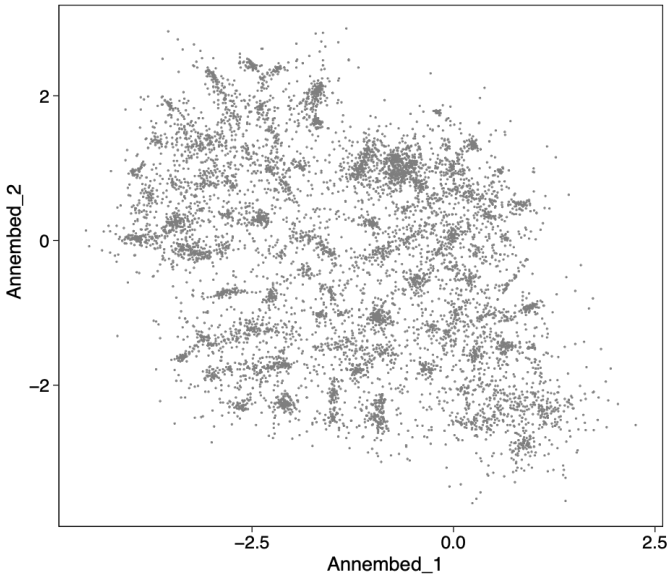

Figure S7. Annembed plot of MyCOcosm database (10,000 fungal genomes). ProbMinHash distances are based on nucleotide Jaccard distance.

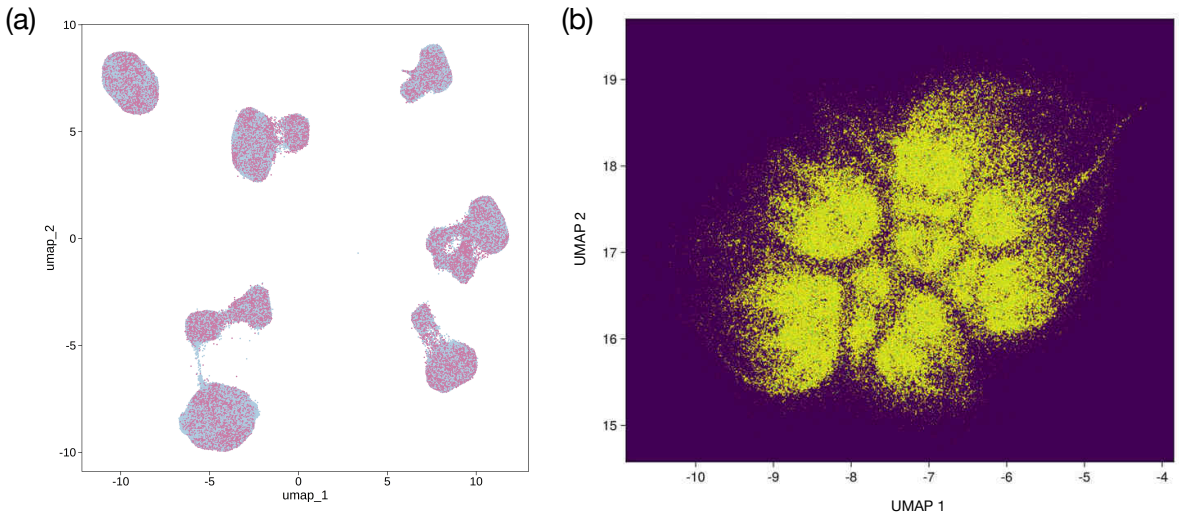

Figure S8. Visualization of embedding results for the HIGGS dataset using UMAP (a) and annembed (b). Annembed default parameters were used for maintaining the quality of the embedding.

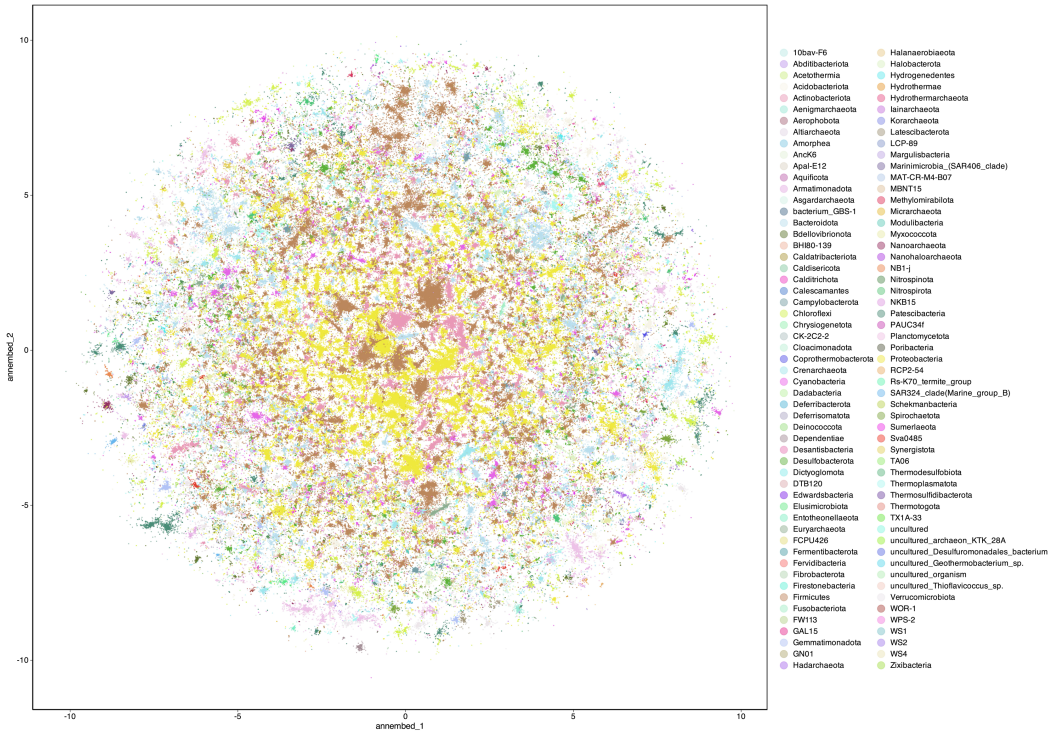

Figure S9. Visualization of 1.6 million SILVA 16S rRNA sequence database (prokaryotes only). Order MinHash is used for approximating Edit distance. Each color indicates a phylum.

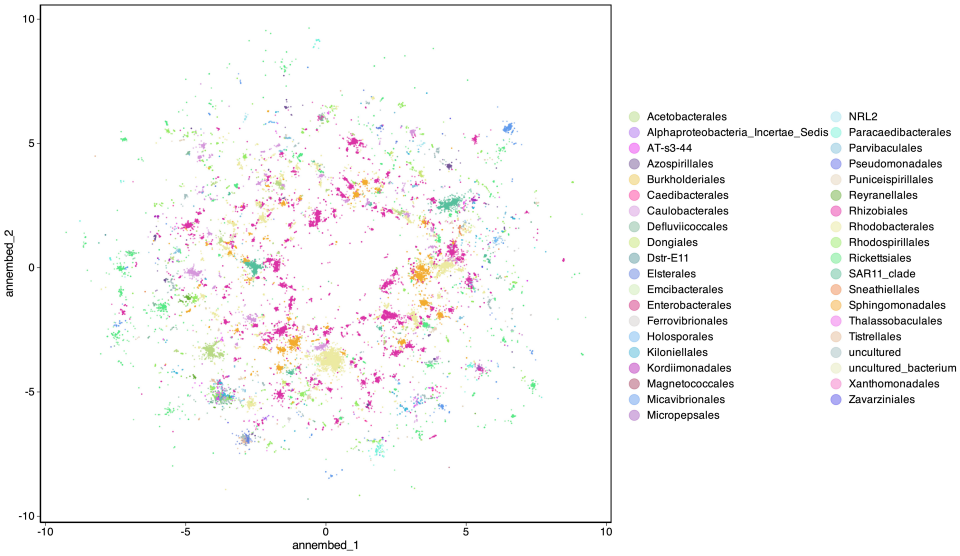

Figure S10. Visualization of the class alphaproteobacteria in the phylum proteobacteria (yellow points in Figure S9 above). Each color indicates an order within the class.

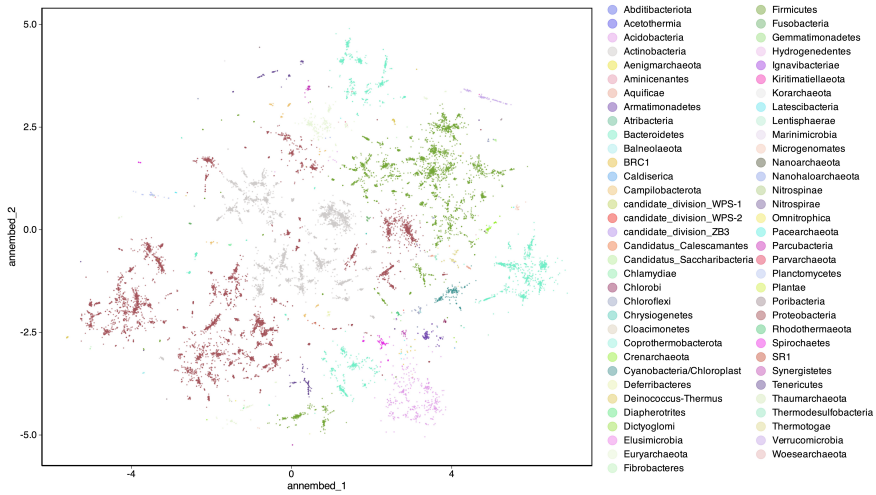

Figure S11. Visualization of RDB v18 ribosomal RNA database (~20K). Each color indicates a phylum.

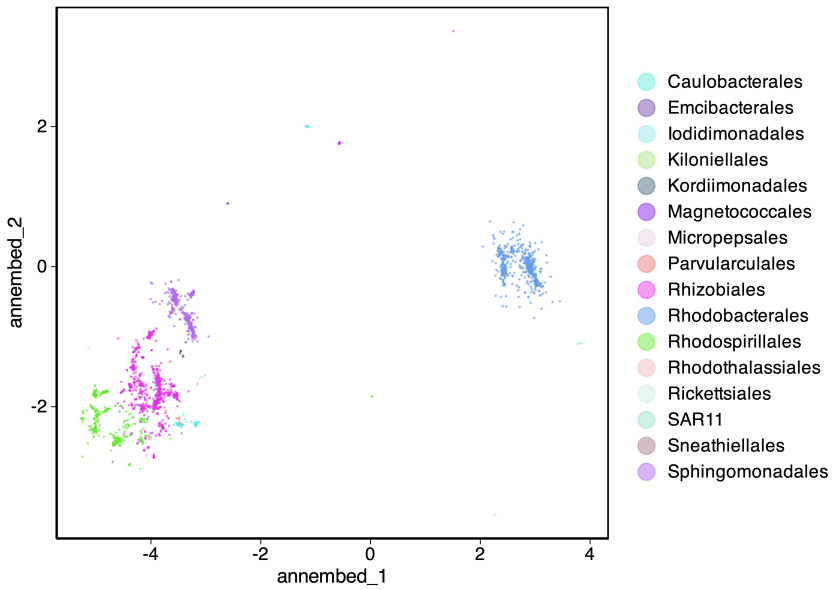

Figure S12. Visualization of genome clusters of the *Alphaproteobacteria* class of the *Proteobacteria* phylum (Brown points in Figure S11 above) in RDB v18 ribosomal RNA database (~20K). Each color indicates an order within the class. Also note that *Proteobacteria* have been recently renamed as *Pseudomonadota*.

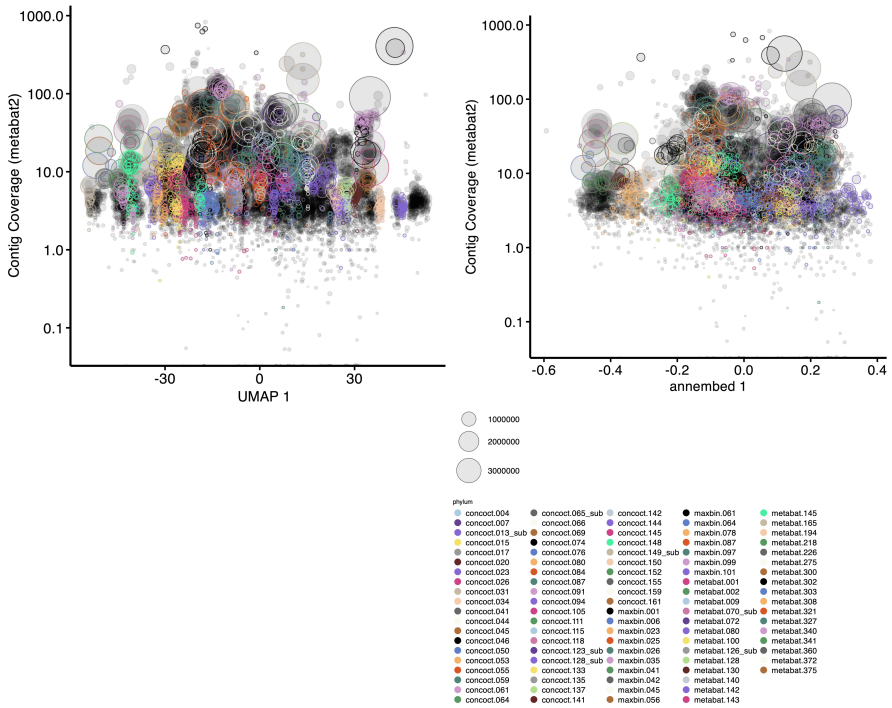

Figure S13. UMAP and annembed visualization of mmgenome2 genome binning results. X axis is UMAP firs dimension and annembed first dimension of contig 4-mer composition based on Euclidean distance while Y is contig coverage based on metabat2 definition. Each color indicates a bin or MAG that consists of those contigs (the circle) with similar coverage and kmer composition. The size of the circle indicates contig length. See Methods & Material for how the data were prepared.

Supplementary Methods

The true loss function that should be optimized in UMAP is:

$$\begin{aligned}\mathcal{L}(\{e_i\}|\{\mu_{ij}\}) &= -2 \sum_{1\le i<j\le n} \mu_{ij} \log(\nu_{ij}) \qquad\quad +(1-\mu_{ij}) \log(1-\nu_{ij}) \\ &= -2 \sum_{1\le i<j\le n} \mu_{ij} \underbrace{\log(\phi(e_i,e_j))}_{-\mathcal{L}^a_{ij}} \quad +(1-\mu_{ij}) \underbrace{\log(1-\phi(e_i,e_j))}_{-\mathcal{L}^r_{ij}}.\end{aligned}$$

However, UMAP implementation did not multiply by 2 in the first term, which leads to more attractive force (Damrich and Hamprecht, 2021). In annembed, we use the default loss function above.

Damrich, S., and Hamprecht, F.A. (2021) On UMAP's true loss function. *Advances in Neural Information Processing Systems* **34**: 5798-5809.