Alignment-free Visualization of Metagenomic Data by Nonlinear Dimension Reduction

Cedric C. Laczny[1], Nicolás Pinel[1,2], Nikos Vlassis[1,*] Paul Wilmes[1,*]

1 Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg

2 Institute for Systems Biology, Seattle, Washington, USA

∗ **E-mail: Correspondence to be addressed to nikos.vlassis@uni.lu or paul.wilmes@uni.lu**

Contents

Supplementary Note 1 - Simulated data 2
Supplementary Note 2 - Human microbiome 2
Supplementary Note 3 - Marine microbial community 2
Supplementary Note 4 - Two-component EMGM 3

List of Supplementary Figures

S1 Phylogenetic tree - EqualSet02 4
S2 BH-SNE-based visualization - EqualSet01 5
S3 BH-SNE-based visualization - EqualSet01 with varying fragment length 6
S4 BH-SNE-based visualization - EqualSet01 with varying error rates 7
S5 BH-SNE-based visualization including different phylogenetic marker genes - EqualSet01 . 8
S6 Multi-component EMGM-based clustering - EqualSet01 9
S7 ESOM-based U-Matrix visualization - LogSet01 10
S8 BH-SNE-based visualizations - EqualSet02 11
S9 ESOM-based U-Matrix visualization - EqualSet02 (4mers) 12
S10 ESOM-based U-Matrix visualization - EqualSet02 (5mers) 13
S11 Percentage of GC content vs. genomic fragment length - Groundwater metagenomic dataset 14
S12 BH-SNE-based visualization - Marine microbial community genomics dataset 15

List of Supplementary Tables

S1 Characteristics of the isolate genomes for EqualSet01 and LogSet01 16
S2 Characteristics of the isolate genomes for EqualSet02 16
S3 Sensitivity, specificity & precision values of multi-component EMGM - EqualSet01 16
S4 Overlap - Polygonal vs. (semi-)automated two-component clustering 17
S5 Assembly metrics - Groundwater dataset 17
S6 Runtimes - Simulated datasets 17

Supplementary Notes

Supplementary Note 1 - Simulated data

We built two genome sets, each constisting of ten microbial isolate genomes selected at random from the NCBI microbial genome database. The first genome set comprises genomic information from distinct genera, with the exception of the presence of two *E. coli* strains, see Supplementary Table S1. Based on this genome set, we generated two simulated community genomic datasets (metagenomic datasets): one where generated genome fragments had a one-fold coverage of the respective genomes and one with fragment abundances that followed a logarithmic rank abundance distribution of mixed microbial communities (Figure 3a). These simulated datasets are refered to throughout this work as EqualSet01 and LogSet01, respectively. The other community is composed of ten organisms from three distinct genera (Supplementary Table S2). For this dataset, referred to as EqualSet02, we used a one-fold genome coverage and the rpoB-based phylogenetic relationships of the members of the microbial community in EqualSet02 are shown in Supplementary Figure S1.

The abundances of the organisms as represented in the individual datasets are defined by the number of genomic fragments derived from the individual organisms' genomes. As such, high abundance organisms are characterized by large numbers of fragments and vice versa for low abundance organisms. These three datasets served as ground truths to objectively asses the individual performances of our approach as well as the ESOM-based approach. In particular, the datasets were used to compare the individual performances in allowing discrimination between discrete sequence clusters and the individual required runtimes.

Supplementary Note 2 - Human microbiome

In addition to the results in the main text, we report here more individual details on the analyses of the BH-SNE-based selections HM-Subset01 and HM-Subset02 (Figure 5).

For HM-Subset01, around 94% of contigs (300/318) in this group exhibit significant alignments to *Escherichia coli* genomes as the top-hit or among the ten top-hits, with similar or marginally decreased scores. 294 of these contigs had a query coverage $> 95\%$ while six of these contigs aligned with significant low E-values but with a query coverage of $< 95\%$ (Methods). Aligning all available contigs (9,911) against all currently available *E.coli* complete genomes revealed 301 contigs which resulted in significant alignments and a query coverage $> 95\%$. Sensitivity, specificity, and precision were 97.67%, 99.75%, and 92.45%, respectively.

For HM-Subset02, around 85% (276/323) of contigs report a significant top hit against *Eggerthella lenta*, with 21 contigs with a query coverage $< 95\%$. Aligning all available contigs (9,911) against all currently available *Eggerthella* complete genomes revealed 275 contigs to report significant alignments and a query coverage $> 95\%$ (Methods). Sensitivity, specificity, and precision were 92.73%, 99.29%, and 78.95%, respectively.

Supplementary Note 3 - Marine microbial community

For this environmental dataset, we focused on minimum sequence lengths of 1,500nt and 2,000nt. As described in the main text, the rationale for choosing a minimum contig length of 2,000nt is that smaller contig sizes lead to the emergence of a limited amount of discrete clusters (Supplementary Figure S12a) when compared to the other datasets which are reported in this work. In addition to the results for this dataset, we report more individual details on the analyses of the BH-SNE-based selections for DS-Subset01 and DS-Subset02 (Supplementary Figure S12b).

DS-Subset01 contained 125 contigs and only 13 reported significant alignments against the reference database. Out of these 13, 12 were reporting the top hit against Uncultured marine microorganism

HF4000-related fosmid sequences from Konstantinidis *et al.*[1]. 78 contigs did not result in any hit to available reference sequences in NCBI's Genbank non-redundant nucleotide database and the remaining 34 contigs either have an E-value or query coverage smaller than our thresholds.

DS-Subset02, which consists of 322 contigs, includes 21 contigs with no significant hit and 112 contigs with either E-value or query coverage smaller than our thresholds (Online Methods). More than 50% (169) of all the contigs aligned to fosmid sequences of Uncultured Group I marine crenarchaea HF4000 from Konstantinidis *et al.*[1]. Other significant alignments were against Uncultured marine crenarchaeote sequences (9), *Phakopsora pachyrhizi* clone JGIAFNA-829C10 (7), Uncultured *Alteromonas* sp. AD1000-G12-6 (2), *Candidatus* Nitrosopumilus sp. AR2 (1), and *Phakopsora pachyrhizi* clone JGIAFNA-2193J23 (1), respectively.

Supplementary Note 4 - Two-component Expectation-Maximization Gaussian Mixture model-based clustering

The two-dimensional visualization of CLR-transformed oligonucleotide signatures using our BH-SNE-based approach reveals clusters that exhibit characteristics akin to two-dimensional Gaussian probability density functions. Motivated by this observation, we used Gaussian Mixtures models for cluster delineation. In particular, to capture an individual cluster, we postulate a two-component Gaussian Mixture model, whereby one component represents the cluster of interest ("foreground"), and the other component captures everything else ("background"). We fit this two-component mixture model with the Expectation-Maximization (EM) algorithm for Gaussian Mixture models[2], whereby the mean of the "foreground" component is manually initialized by the user (by simply clicking on the scatterplot) and its covariance matrix is set to a small multiple of the identity matrix (i.e., spherical Gaussian). The "background" component is initialized by the mean and the covariance matrix of all points in the scatterplot. The EM algorithm iteratively optimizes the means and covariance matrices of the two components, as well as the mixing weight (which reflects the relative abundance of the local cluster against all other points). Each step of EM is guaranteed to improve (or leave unchanged, if a local optimum has been reached) the likelihood of the two-dimensional points under the postulated two-component mixture model. Note that the above procedure is semi-automatic, as it involves an initialization by the user of some of the parameters (namely, the mean of the first component). It is known that automated clustering approaches can strongly benefit from human-augmented input[3,4], and it is the two-dimensional nature of the projected signatures that allows leveraging the innate capacity of the human eye-brain system for quick and accurate pattern recognition in two dimensions[5].

Supplementary References

1. Konstantinidis, K. T. & DeLong, E. F. Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *ISME J* **2**, 1052–65 (2008).

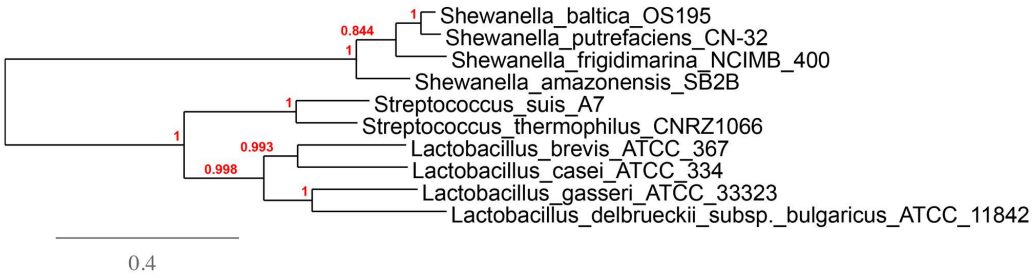
2. Redner, R. & Walker, H. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev* **26**, 195–239 (1984).

3. Zou, J. & Nagy, G. Human Computer Interaction for Complex Pattern Recognition Problems. In Basu, Mitra and Ho, T. (ed.) *Data Complex Pattern Recognit*, 271–286 (Springer London, 2006).

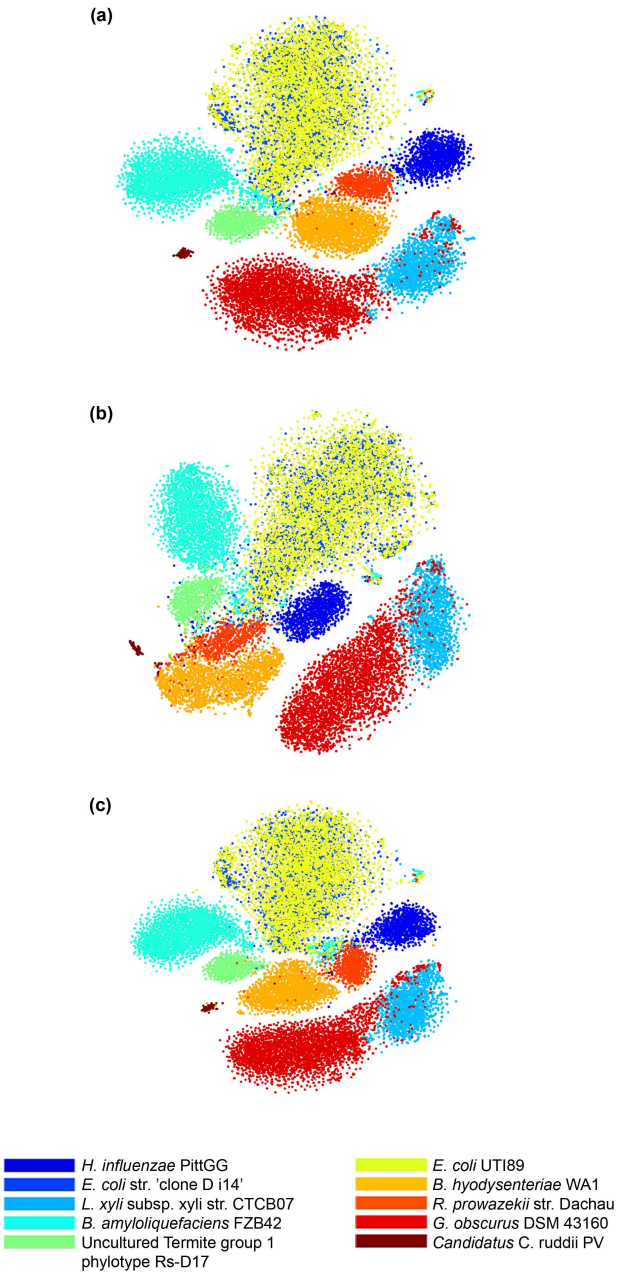
4. Zhu, Y. *et al.* caBIG VISDA: modeling, visualization, and discovery for cluster analysis of genomic data. *BMC Bioinformatics* **9**, 383 (2008).

5. Sinha, P. Recognizing complex patterns. *Nat Neurosci* **5 Suppl**, 1093–7 (2002).

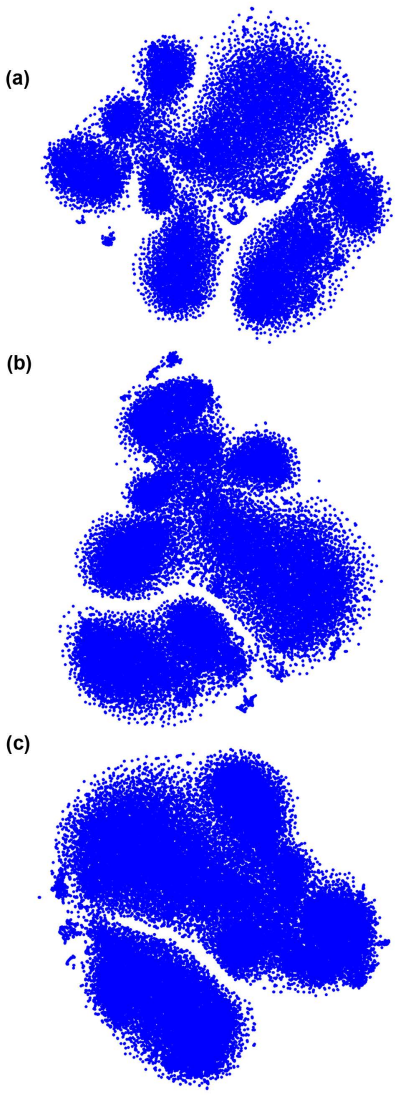
Supplementary Figures



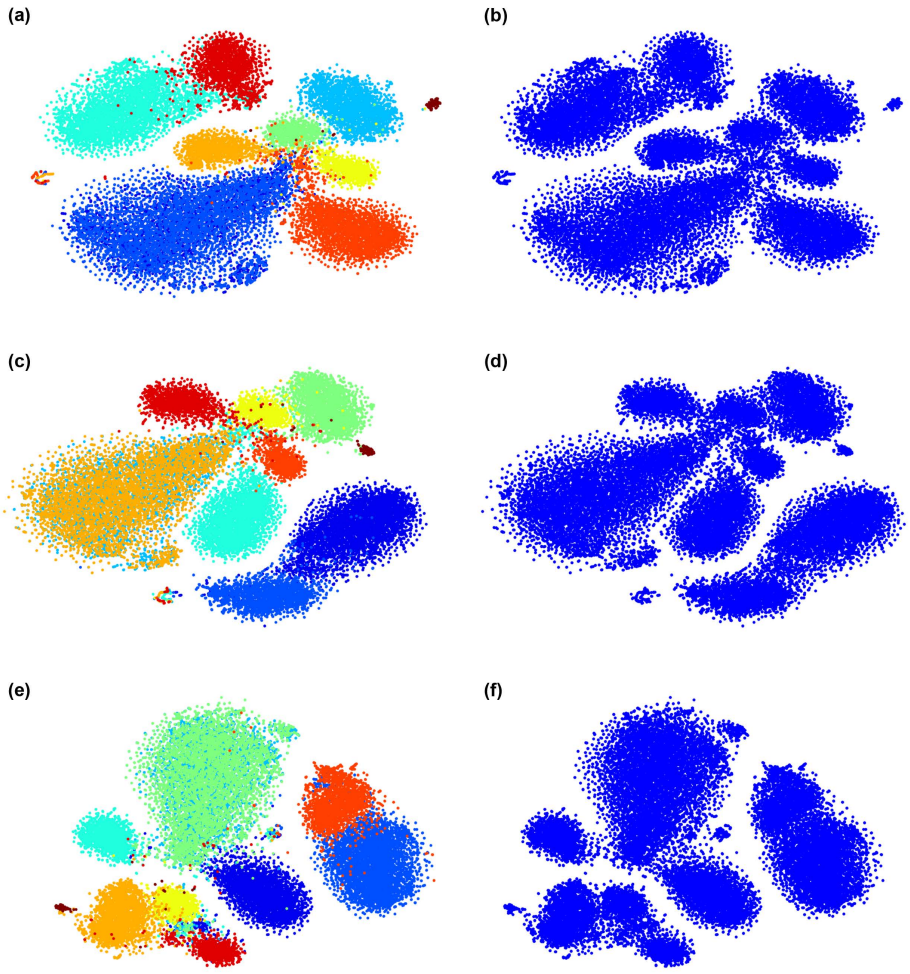
Supplementary Figure S1: Phylogenetic tree based on rpoB for the simulated dataset of closely related organisms (EqualSet02). The branch length is proportional to the number of substitutions per site, with the substitution rate indicated in the plot. Red values represent "branch support values".



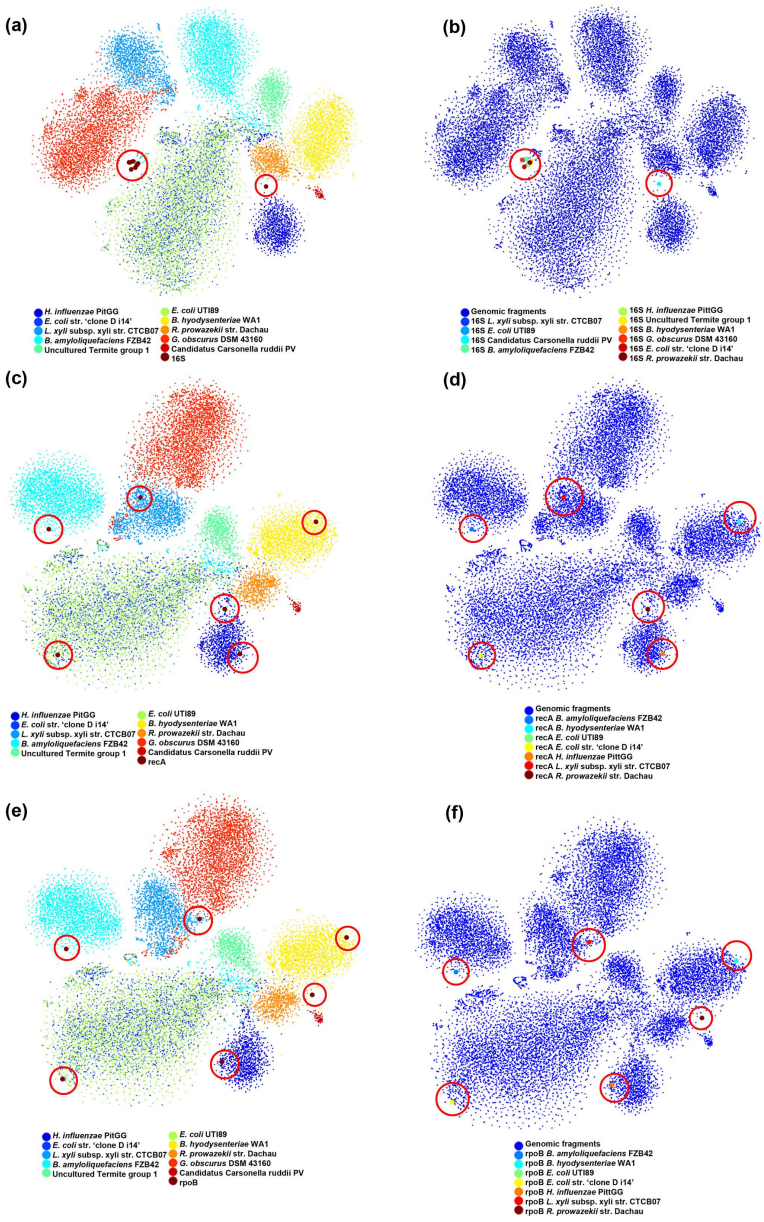
Supplementary Figure S2: BH-SNE-based visualization of genomic fragment signatures for the evenly distributed simulated dataset (EqualSet01). (a) Based on CLR-transformed pentanucleotides (5mers). (b) Based on untransformed tetranucleotides (4mers). (c) Based on CLR-transformed tetranucleotides (4mers). (a–c) Genomic fragments have a length of 1,000nt. The color coding reflects the organismal origin of the represented genomic fragments. Colors have been added for demonstration purposes only.



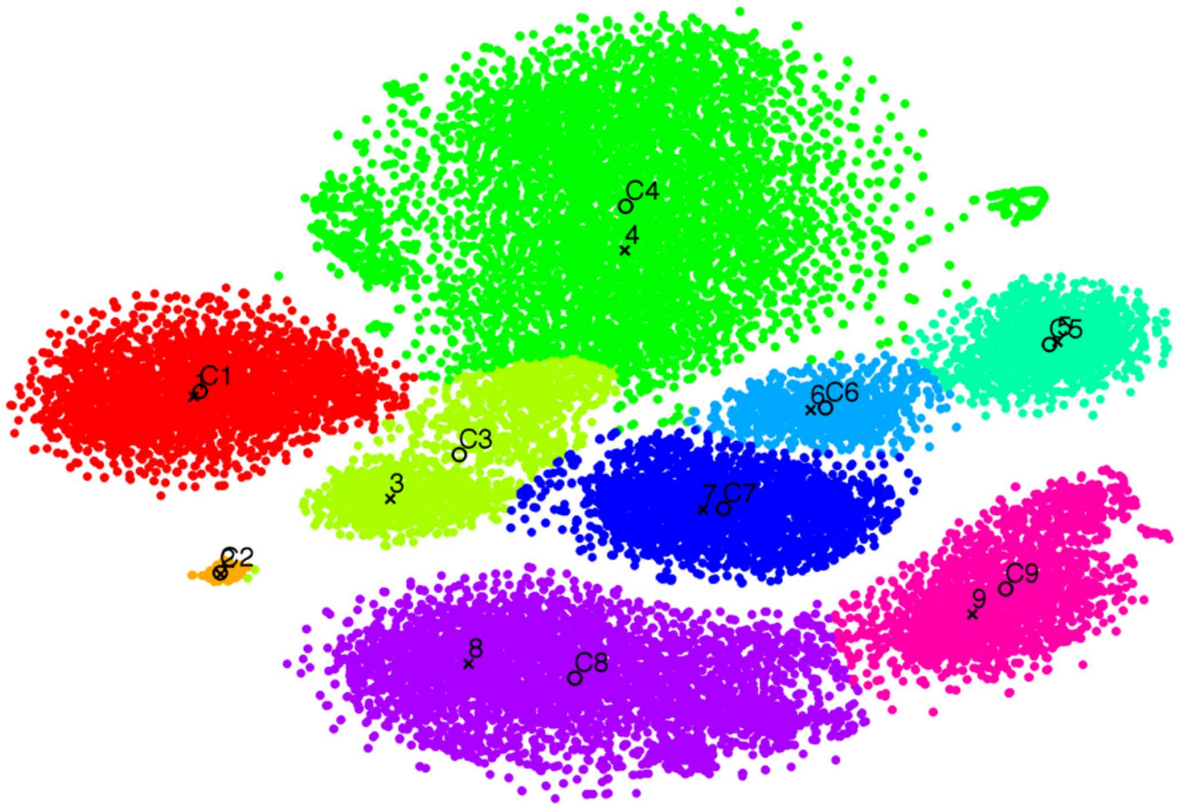
Supplementary Figure S3: BH-SNE-based visualization of genomic fragment signatures for EqualSet01 (even community, overall reflecting distant taxonomic relatedness) with varying fragment lengths. (a) 800nt. (b) 600nt. (c) 400nt.



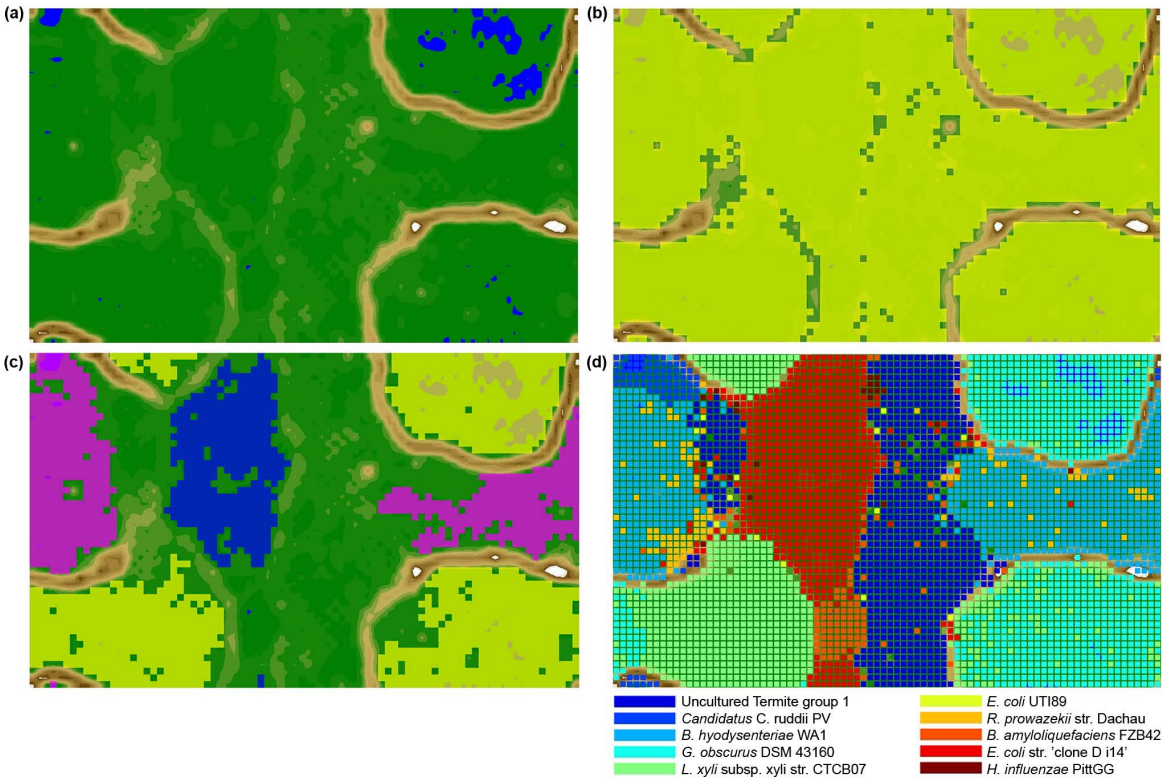
Supplementary Figure S4: BH-SNE-based visualization of genomic fragment signatures for EqualSet01 (even community, overall reflecting distant taxonomic relatedness) with varying error rates. (a, b) 1% error. (c, d) 3% error. (e, f) 5% error. Genomic fragment length is 1,000nt for all error rates. The color coding reflects the organismal origin of the represented genomic fragments. Colors in panels a, c and e have been added for demonstration purposes only.



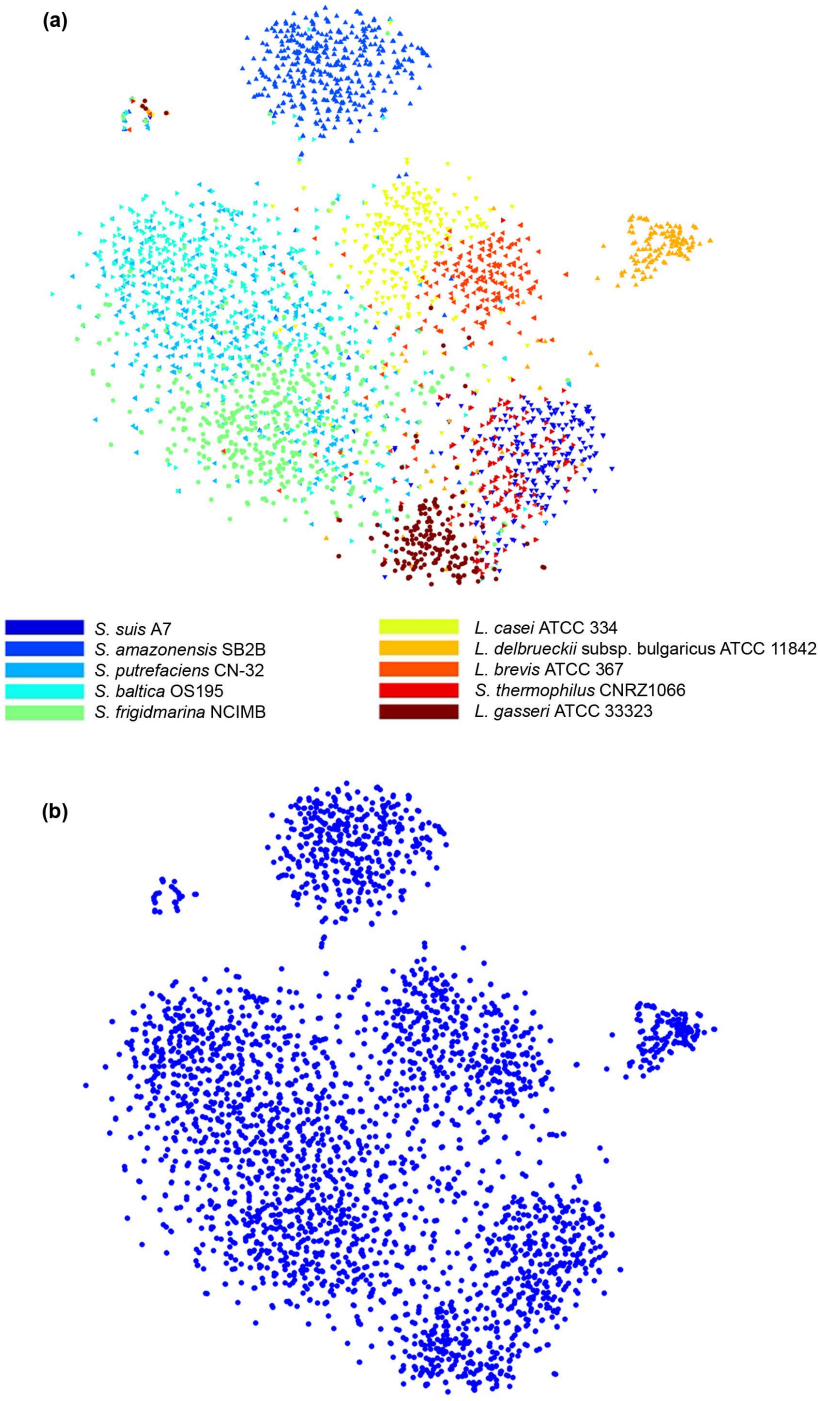
Supplementary Figure S5: BH-SNE-based visualization of the evenly distributed simulated microbial community (EqualSet01) with highlighted positions of different phylogenetic marker genes. Datapoints corresponding to (a, b) the 16S rRNA gene sequences, (c, d) the recA genes sequences and (e, f) the rpoB genes sequences are highlighted with a larger pointsize and red circles. (a, c, e) Visualized contig signatures colored according to the taxonomic affiliation of the contigs, (b, d, f) uncolored. Positions in the left panels are the same as in the right panels. Individual legends are provided to highlight the genomic origin of the genomic fragments or of the marker gene sequences.



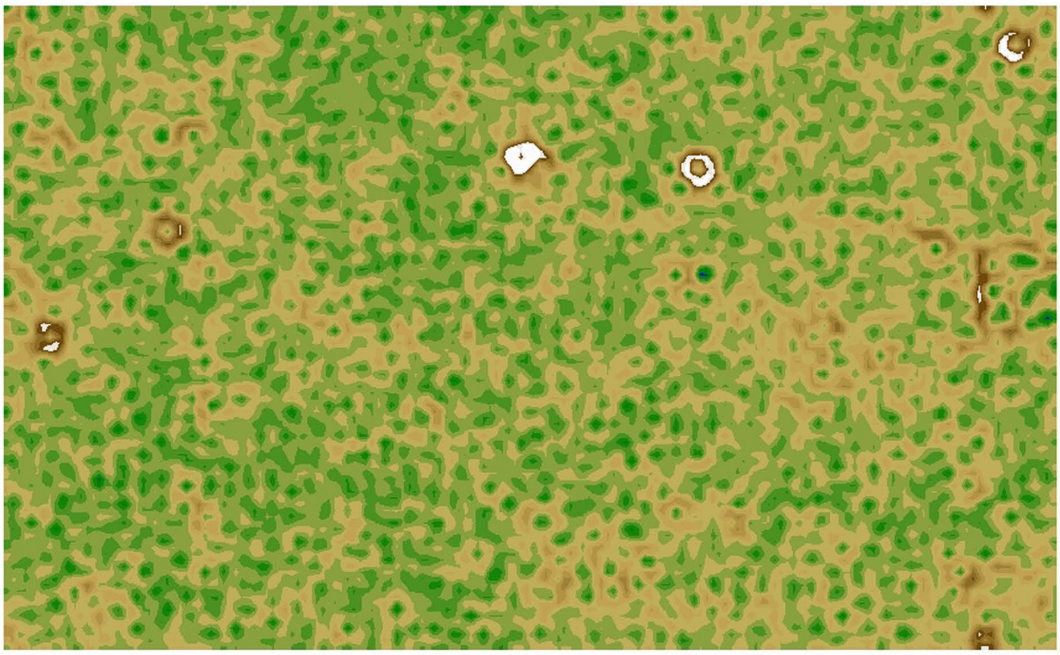
Supplementary Figure S6: (Semi-)automated clustering based on a multi-component Expectation-Maximization Gaussian Mixture model of the evenly distributed simulated microbial community (EqualSet01). Distinct colors represent distinct cluster assignments. 'x' denote the manually placed initial means and 'o' represent the learned means. Respective cluster numbers are shown next to the 'x' and 'o', respectively. To further highlight the learned means, the respective associated cluster numbers are prefixed by a 'C'.



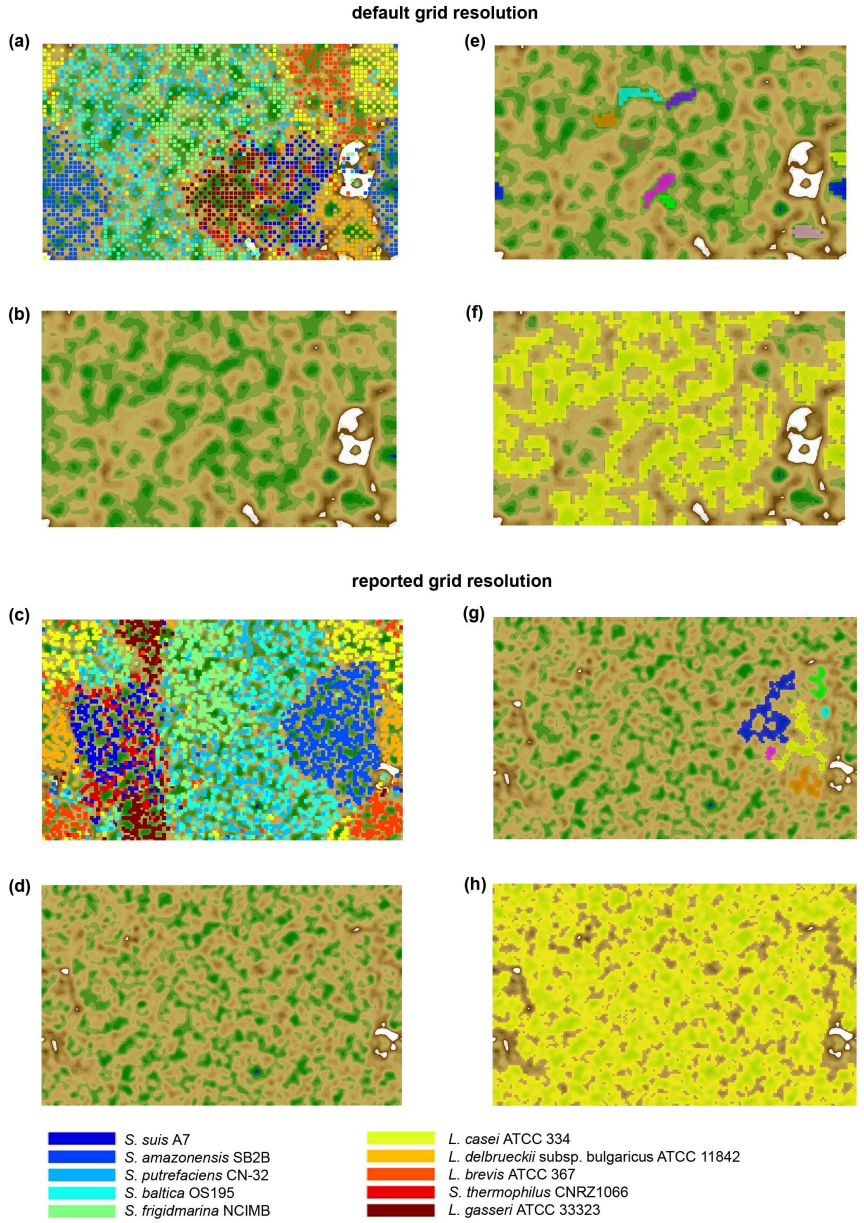
Supplementary Figure S7: ESOM-based U-Matrix visualization of genomic fragment signatures for the unevenly distributed simulated dataset (LogSet01). (a) The topological map obtained via the computation of the U-Matrix based on the ESOM trained on the genomic fragment signatures for the selected microorganisms. (b) Floodfill with default threshold of 0.2. (c) Floodfill with stringent threshold of 0.1. (d) Overlay of known information on the learned topological representation.



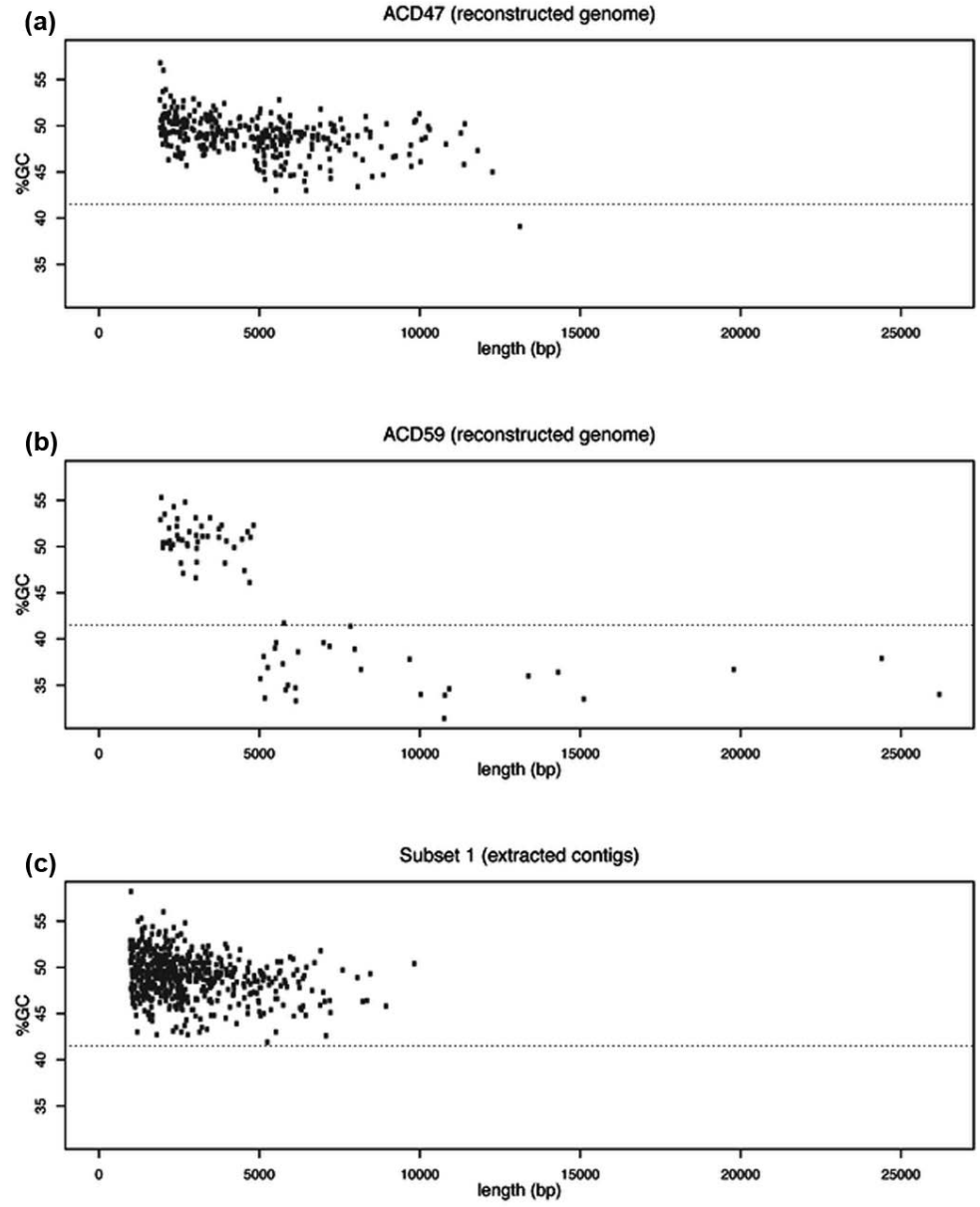
Supplementary Figure S8: BH-SNE-based visualizations for EqualSet02: simulated microbial community comprising closely related organisms. (a) Color-coding of the points according to known information (see legend). (b) No labeling information provided.



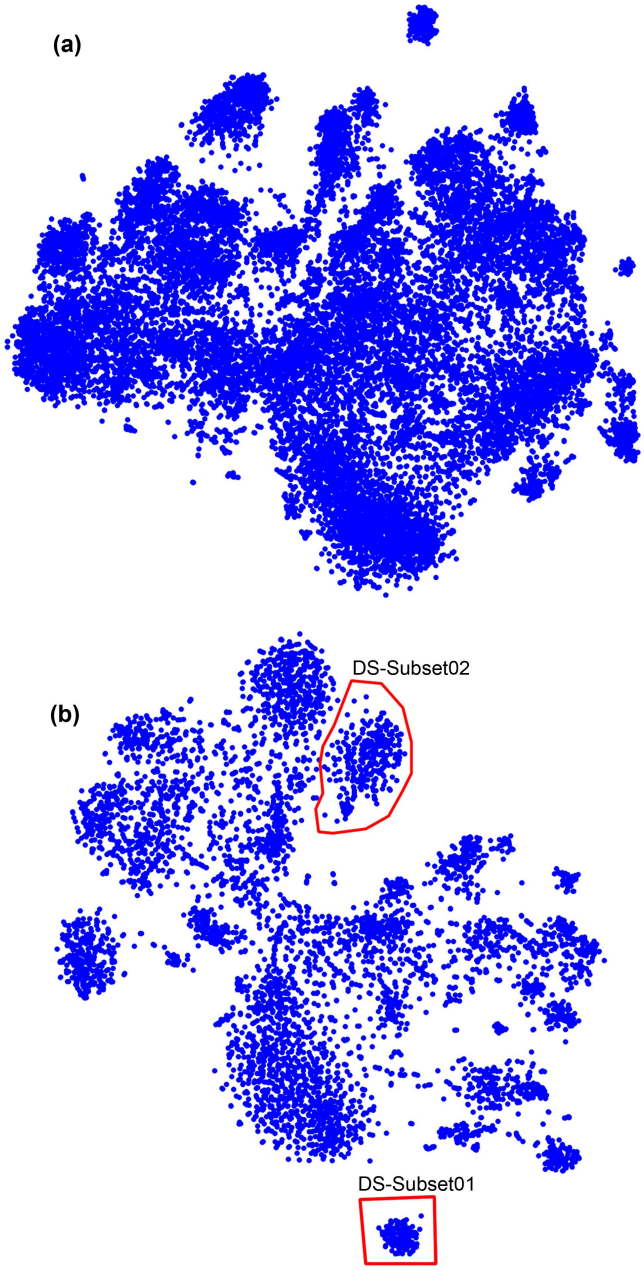
Supplementary Figure S9: ESOM-based U-Matrix visualization of genomic fragment signatures for EqualSet02 (4mers): simulated microbial community comprising closely related organisms. The number of training epochs was 100 with a grid-resolution of around 17,000 neurons (104 by 170).



Supplementary Figure S10: ESOM-based U-Matrix visualization of genomic fragment signatures for EqualSet02 (5mers): simulated microbial community comprising closely related organisms. (a, b, e, f) The representations obtained when using a default grid resolution of around 4,000 neurons. (c, d, g, h) The representations obtained when using the reported grid resolution of around 17,000 neurons. Panels (a) and (c) provide color-coding of the points according to the legend at the bottom of the figure. The points in panel (c) are plotted bigger than in panel (a) only to make them better visible for the increased grid resolution. Panels (e)–(h) on the right show the results of applying the "floodfill" algorithm with different thresholds: (e) - 0.2, (f) - 0.3, (g) - 0.3, (h) - 0.4. Colors in the panels on the right refer to the groups of neurons selected by the "floodfill" algorithm with the respective parameter and do not reflect the colors in the legend.



Supplementary Figure S11: Percentage of GC content versus genomic fragment length in the groundwater metagenomic dataset for the scaffolds of the reconstructed reference genomes and contigs of GW-Subset01. (a) Organismal group ACD47 (original publication), (b) Organismal group ACD59 (original publication) and (c) contigs of the BH-SNE-based selection of GW-Subset01.



Supplementary Figure S12: BH-SNE-based visualization of genomic fragment signatures for the marine microbial community genomics dataset. (a) Minimum genomic fragment length of 1,500nt. (b) Minimum genomic fragment length of 2,000nt. Red polygons depict subsets of interest that were further characterized as detailed in the text.

Supplementary Tables

Supplementary Table S1: Characteristics of the isolate genomes for EqualSet01 and LogSet01.

Organism	Genome size (nt)	%GC
Leifsonia xyli subsp. xyli str. CTCB07	2,584,158	67.7
Escherichia coli UTI89	5,065,741	50.6
Candidatus Carsonella ruddii PV	159,662	16.6
Haemophilus influenzae PittGG	1,887,192	38.0
Bacillus amyloliquefaciens FZB42	3,918,589	46.5
Brachyspira hyodysenteriae WA1	3,000,694	27.1
Geodermatophilus obscurus DSM 43160	5,322,497	74.0
Rickettsia prowazekii str. Dachau	1,109,051	29.0
Escherichia coli str. 'clone D i14'	5,038,386	50.6
Uncultured Termite group 1 bacterium phylotype Rs-D17	1,125,857	35.2

Supplementary Table S2: Characteristics of the isolate genomes for EqualSet02.

Organism	Genome size (nt)	%GC
Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842	1,864,998	49.7
Lactobacillus brevis ATCC 367	2,291,220	46.2
Lactobacillus casei ATCC 334	2,895,264	46.6
Lactobacillus gasseri ATCC 33323	1,894,360	35.3
Shewanella amazonensis SB2B	4,306,142	53.6
Shewanella putrefaciens CN-32	4,659,220	44.5
Shewanella baltica OS195	5,347,283	46.3
Shewanella frigidimarina NCIMB 400	4,845,257	41.6
Streptococcus suis A7	2,038,409	41.2
Streptococcus thermophilus CNRZ1066 chromosome	1,796,226	39.1

Supplementary Table S3: Sensitivity, specificity & precision values calculated for the clusters resulting from the application of multi-component Expectation-Maximization Gaussian Mixture model-based clustering on EqualSet01.

Cluster†	Sensitivity (%)	Specificity (%)	Precision (%)	Originating organism
Cluster01	90.17	99.98	99.89	*Bacillus amyloliquefaciens*
Cluster02	88.75	100.00	100.00	*Candidatus* Carsonella ruddii
Cluster03	94.83	96.22	51.33	Uncultured Termite group1 bacterium
Cluster04	91.70	99.06	98.07	*Escherichia coli*
Cluster05	93.08	99.93	98.97	*Haemophilus influenzae*
Cluster06	97.18	99.40	86.43	*Rickettsia prowazekii*
Cluster07	98.07	99.63	96.81	*Brachyspira hyodysenteriae*
Cluster08	96.22	99.70	98.65	*Geodermatophilus obscurus*
Cluster09	96.70	99.24	92.58	*Leifsonia xyli*

† : Cluster numbers follow the numbering in Supplementary Figure S6

Supplementary Table S4: Overlap of polygonal vs. (semi-)automated two-component Expectation-Maximization Gaussian Mixture model-based clustering on real-world datasets.

Dataset	**Average overlap (%)**
Groundwater	96.39
Human microbiome	97.82
Marine	95.41

Supplementary Table S5: Assembly metrics of the groundwater dataset. Reads were recruited against contigs selected as GW-Subset01 using BH-SNE.

Metric	**Original assembly**	**Re-assembled ($\geq$ 1knt)**	**Re-assembled ($<$ 1knt)**
# contigs	544	519	355
seq. length (Mbp)	1.532	1.687	1.646
N50	3,305	3,775	454
avg. seq. length	2,818	3,251	464
max. seq. length	9,853	19,879	994

Supplementary Table S6: Runtimes for the ESOM-based approach and our BH-SNE-based approach on the simulated datasets.

Type	**# sequences**	**ESOM - 4k†(s)**	**ESOM - 17k‡(s)**	**BH-SNE (s)**
EqualSet01 (1,000nt)	29,212	1,000	n.d.*	225
LogSet01 (1,000nt)	58,771	1,820	n.d.*	479
EqualSet02 (1,000nt)	3,194	174	756	15

† : Resolution: 50 by 82
‡ : Resolution: 104 by 170
* : n.d.: not determined