



OPEN

# Alignment-free Visualization of Metagenomic Data by Nonlinear Dimension Reduction

SUBJECT AREAS:  
METAGENOMICS  
MACHINE LEARNINGCedric C. Laczny<sup>1</sup>, Nicolás Pinel<sup>1,2</sup>, Nikos Vlassis<sup>1</sup> & Paul Wilmes<sup>1</sup>Received  
10 July 2013Accepted  
13 March 2014Published  
31 March 2014

Correspondence and requests for materials should be addressed to N.V. (nikos.vlassis@uni.lu) or P.W. (paul.wilmes@uni.lu)

<sup>1</sup>Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg, <sup>2</sup>Institute for Systems Biology, Seattle, Washington, USA.

The visualization of metagenomic data, especially without prior taxonomic identification of reconstructed genomic fragments, is a challenging problem in computational biology. An ideal visualization method should, among others, enable clear distinction of congruent groups of sequences of closely related taxa, be applicable to fragments of lengths typically achievable following assembly, and allow the efficient analysis of the growing amounts of community genomic sequence data. Here, we report a scalable approach for the visualization of metagenomic data that is based on nonlinear dimension reduction via Barnes-Hut Stochastic Neighbor Embedding of centered log-ratio transformed oligonucleotide signatures extracted from assembled genomic sequence fragments. The approach allows for alignment-free assessment of the data-inherent taxonomic structure, and it can potentially facilitate the downstream binning of genomic fragments into uniform clusters reflecting organismal origin. We demonstrate the performance of our approach by visualizing community genomic sequence data from simulated as well as groundwater, human-derived and marine microbial communities.

Mixed microbial communities are ubiquitous and contribute essential functionalities to all ecosystems. High-resolution molecular approaches, primarily community genomics or metagenomics, are dramatically altering our view of the composition and functional potential of microbial consortia *in situ*. Although metagenomics is being applied to an ever growing sample set<sup>1–3</sup>, major methodological challenges in data analysis remain. One such challenge is the efficient visualization of congruent groups of genomic sequence fragments of closely related taxa.

Observations dating back to the 1960's suggest the presence of species-specific oligonucleotide signatures in genomic sequences<sup>4</sup>, which have been lately corroborated by analyses of complete microbial genomes<sup>2,5–7</sup>. These signatures are normalized frequencies of *k*mers of a particular size (typically  $k = 4$ ). They can be represented as vectors in a high-dimensional Euclidean space, for instance, 136-dimensional in the case of 4mers (by accounting for reverse complements and palindromes<sup>8</sup>). Transformation of the high-dimensional data into a low-dimensional space (typically two dimensions) enables human interpretation. Importantly, to allow faithful representation of the information contained within the metagenomic sequence fragments, a transformation/visualization method should ideally preserve the data-inherent taxonomic structure. In other words, groups of sequences of closely related taxa (e.g. Operational Taxonomic Units) should be proximal.

The state of the art in alignment-free metagenomic data visualization is based on Emergent Self-Organising Maps (ESOMs)<sup>2,6,7,9–11</sup>, which are models for topographic mapping. Recently, ESOM-based visualization, combined with clustering (“binning”) in the low-dimensional space, has allowed discovery of a novel bacterial lineage and metabolic details of other uncharacterized lineages<sup>2</sup>. However, the absence of any apparent continuous manifold structure in metagenomic data makes the use of topographic mapping (and hence ESOM) for the visualization of metagenomic data questionable.

We report a new approach to alignment-free metagenomic data visualization based on Barnes-Hut Stochastic Neighbor Embedding (BH-SNE)<sup>12</sup> of centered log-ratio (CLR)-transformed<sup>13,14</sup> oligonucleotide signatures of assembled sequence fragments. CLR-transformation is a common data normalization step when dealing with compositional data such as oligonucleotide signatures. BH-SNE is an efficient (runtime of  $O(n \log n)$  where  $n$  is the number of genomic fragments) nonlinear dimension reduction method for data visualization in two dimensions, which is designed to reliably capture the neighborhood structure of the data in the original space. We apply our approach to simulated as well as real-world metagenomic datasets, demonstrating its potential to provide



representative, alignment-free visualizations of data-inherent taxonomic structure. Optionally followed by two-dimensional clustering combined with human input<sup>15,16</sup>, our approach can facilitate downstream analysis by enabling high-quality, human-augmented binning of metagenomic data.

## Results

**Simulated datasets.** *Characteristics of the simulated datasets.* In order to assess the ability of BH-SNE to visualize and provide appropriate representations of genomic data from mixtures of microorganisms, we first generated simulated metagenomic datasets which allowed objective assessment of the performance of our method, in particular when compared to the state-of-the-art ESOM-based approach. The simulated datasets comprise genomic fragments (1,000 nt in length) from ten microbial isolate genomes each, and represent benchmarking ground truths. These datasets were assembled to reflect the three major characteristics of microbial communities, namely that (i) they comprise mixtures of different taxa (species; EqualSet01), (ii) the different taxa follow a typical rank abundance distribution (LogSet01), (iii) they may comprise taxa which are closely related (EqualSet02). EqualSet01 and EqualSet02 exhibit uniform sequence fragment coverages per genome, whereas in LogSet01 the distribution of genomic fragment abundances follows a logarithmic rank-abundance curve. EqualSet01 and LogSet01 are characterized by genomes from organisms of nine distinct genera. EqualSet02 contains genomic fragments from three distinct genera and thereby allows for the assessment of separation performance of genomic data from closely related microorganisms (see Supplementary Fig. S1 online). Additional information on the simulated datasets can be found in Supplementary Note 1 (online) and associated Supplementary Tables S1–S2 (online).

*Application of the method to an even community.* The application of our proposed combination of CLR-transformed pentanucleotide (5mers) frequencies and BH-SNE to EqualSet01 allows clear discrimination of sequence clusters, reflecting organismal origin, with a sequence length of 1,000 nt (Figure 1). Sequences from the two *E. coli* strains produce very similar two-dimensional mappings and therefore overlap in the scatter plot (Figure 1). The use of CLR-transformed pentanucleotide-based signatures is motivated by the improved regularity and congruency of resulting sequence clusters in the BH-SNE-based visualization (see Supplementary Fig. S2a online), compared to untransformed 4mer-based signatures (see Supplementary Fig. S2b online) and CLR-transformed 4mer-based signatures (see Supplementary Fig. S2c online).

We also investigated how the apparent separation depends on sequence fragment length (see Supplementary Fig. S3 online) or sequence error rate (see Supplementary Fig. S4 online). While a length of 1,000 nt shows the highest degree of separation, a length of 600 nt might also be considered for certain applications although sequences from moderately closely related species may no longer be separable. Below 600 nt, the quality of separation and, thus, sequence fragment clustering in two dimensions drops markedly (see Supplementary Fig. S3c online). The separation of originally distinct sequence fragment clusters is preserved for sequence error rates of up to 5% (see Supplementary Fig. S4 online). Such high sequencing error rates would not be expected from data generated using current sequencing technologies.

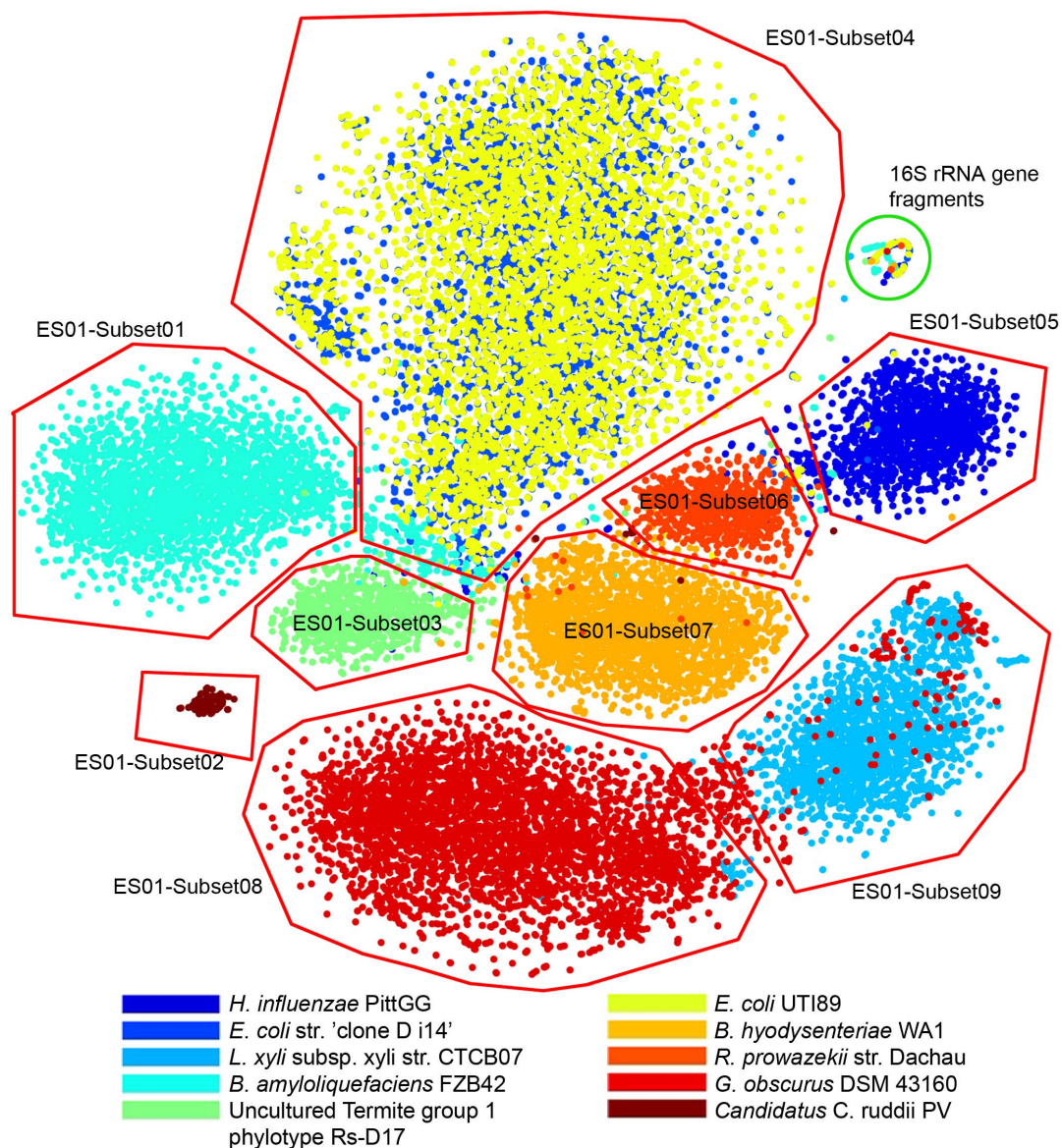
Using EqualSet01, we also examined the placement of common taxonomic markers, such as the 16 S rRNA gene, *recA*, and *rpoB* on the resulting two-dimensional representations (see Supplementary Fig. S5 online). From the BH-SNE-based visualizations, it becomes apparent that the 16 S rRNA gene sequences tend to form a dedicated, distinct cluster outside of the organism-specific clusters. Yet, the sequences of the marker genes *recA* and *rpoB* tend to be located within the cluster of the corresponding organisms.

Visualization of the simulated data allows rapid evaluation of the cluster structure. To assess the sensitivity (completeness), specificity (homogeneity), and precision (fraction of fragments in a cluster belonging to the correct genome) of the apparent clusters (Methods), we performed a human-augmented polygonal selection (Methods). Based on the ground truth data, the average values for sensitivity, specificity, and precision were found to be above 94%, above 99%, and above 97%, respectively (Table 1). As a (semi-)automated alternative to the polygonal selection-approach, we applied a human-augmented clustering by fitting a multi-component Gaussian Mixture model to the data in two dimensions<sup>17</sup>. This method relies on manual placement of initial means, thereby serving as the starting points for the Expectation-Maximization-based optimization of the model parameters (mean, variance, etc.; Methods). With the exception of two slightly different clusters (see Cluster03 and Cluster04 in Supplementary Fig. S6 and Supplementary Table S3 online), the sensitivity and specificity values are nearly identical to those obtained by the human-augmented polygonal selection, thus resulting in average values above 94% and above 99%, respectively (see Supplementary Table S3 online). The precision values are also very similar with the exception of Cluster03 (see Supplementary Fig. S6 and Supplementary Table S3 online) which has a precision value of 51.33%. This single comparatively low precision value may be explained by the increased number of false positives in Cluster03 incurred by the (semi-)automated clustering, in contrast to the human-augmented polygonal selection for which the precision value of the corresponding cluster (ES01-Subset03) is 97.57% (Figure 1; Table 1). The average precision value of the (semi-)automated clustering is 91.41%.

In order to further compare the performance of our method to the state-of-the-art ESOM-based approach, we also applied the ESOM-based method to EqualSet01. Application of the ESOM-based approach results in topographically separated areas (Figure 2a). The application of the “floodfill” algorithm to this simulated dataset, using the default threshold parameter value, leads to the clustering of three distinct groups (Figure 2b) when only one group is expected (Figure 2c). Using a more stringent threshold leads to the creation of numerous artificial subgroups or subterritories (Figure 2d) and a more loose threshold fills almost the entire topological map with two colours (Figure 2e), basically separating the ten organism-community into only two distinct groups. In contrast, the application of our method resulted in the clear delineation of nine out of ten groups with one group comprising genomic sequences from strains of the same species.

*Application of the method to an uneven community.* LogSet01 models a mixture of unevenly distributed taxa (as expected for metagenomic data; Figure 3a). The BH-SNE-based two-dimensional representation of the centered log-ratio-transformed pentanucleotide signatures demonstrates that genomic fragments of the same or of closely related organisms (i.e., of the two *E. coli* strains) are grouped together whereas fragments of distinct taxa are separate from each other (Figure 3b). Despite the fact that the *E. coli* strains are represented here in different abundances (Figure 3a), the respective fragments form a congruent cluster (Figure 3b). Moreover, the clusters of genomic fragments from *Geodermatophilus obscurus* and *Leifsonia xyli* are spatially separated. The same is observed for *Brachyspira hyodysenteriae* and *Candidatus Carsonella ruddii*.

The results of the application of the ESOM-based approach to the robust z-transformed<sup>7</sup> tetranucleotide frequencies of genomic fragments in LogSet01 are shown in Supplementary Figure S7 (online). Similar to above, the default parameters of the Databionics ESOM suite were used (see Supplementary Fig. S7a online). For this dataset with logarithmic abundances, performing a “floodfill” with the default threshold leads to the flooding of the entire map (see Supplementary Fig. S7b online) and a more stringent threshold allows for the definition of three major areas (see Supplementary



**Figure 1** | BH-SNE-based visualization and human-augmented binning of an evenly distributed simulated microbial community (EqualSet01).

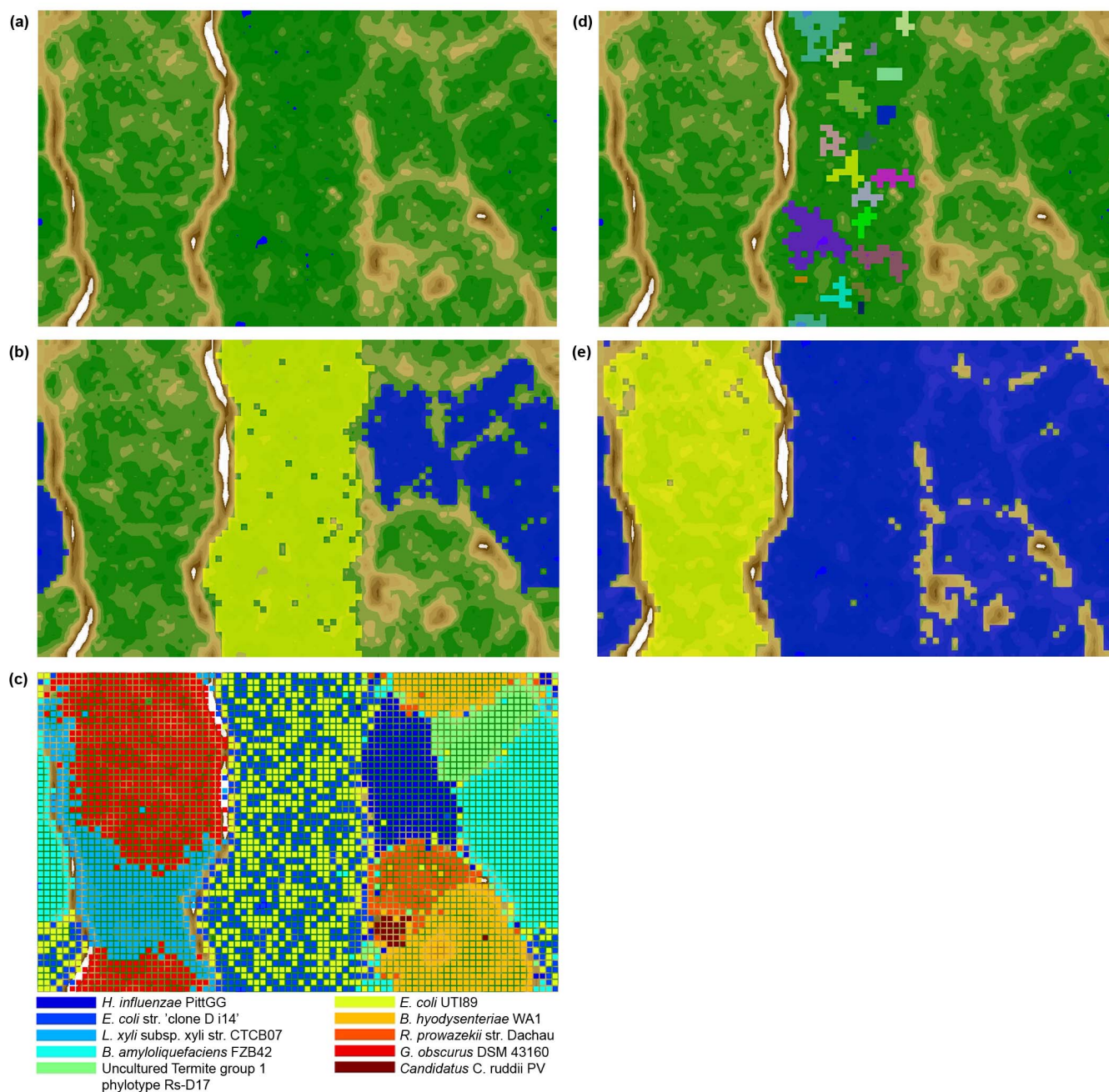
Red polygons contain mappings of genomic fragment signatures and represent congruent subsets of interest. These subsets were used for the computation of the sensitivity, specificity, and precision values. The individual colors (representing distinct microbial genomes of origin) were not present during the selection process but have been added afterwards. The green circle highlights the mappings of 16S rRNA gene fragments. Genomic fragments have a size of 1,000 nt.

**Table 1** | Sensitivity, specificity & precision values for simulated & real-world datasets

Subset*	Sensitivity (%)	Specificity (%)	Precision (%)	Organism†
ES01-Subset01	90.06	99.99	99.94	<i>B. amyloliquefaciens</i>
ES01-Subset02	91.25	100.00	100	<i>Candidatus C. ruddii</i>
ES01-Subset03	95.42	99.90	97.57	Uncultured Termite group 1 bacterium
ES01-Subset04	98.60	98.23	96.67	<i>E. coli</i>
ES01-Subset05	92.76	99.93	98.86	<i>H. influenzae</i>
ES01-Subset06	97.09	99.69	92.39	<i>R. prowazekii</i>
ES01-Subset07	96.97	99.88	98.95	<i>B. hyodysenteriae</i>
ES01-Subset08	95.02	99.13	96.09	<i>G. obscurus</i>
ES01-Subset09	96.82	99.22	92.66	<i>L. xyli</i>
HM-Subset01	97.67	99.75	92.45	<i>E. coli</i>
HM-Subset02	92.73	99.29	78.95	<i>E. lenta</i>

\*: ES01: Simulated data - EqualSet01; HM: Human microbiome.

†: Identity based on ground truth or NCBI online BLAST search.



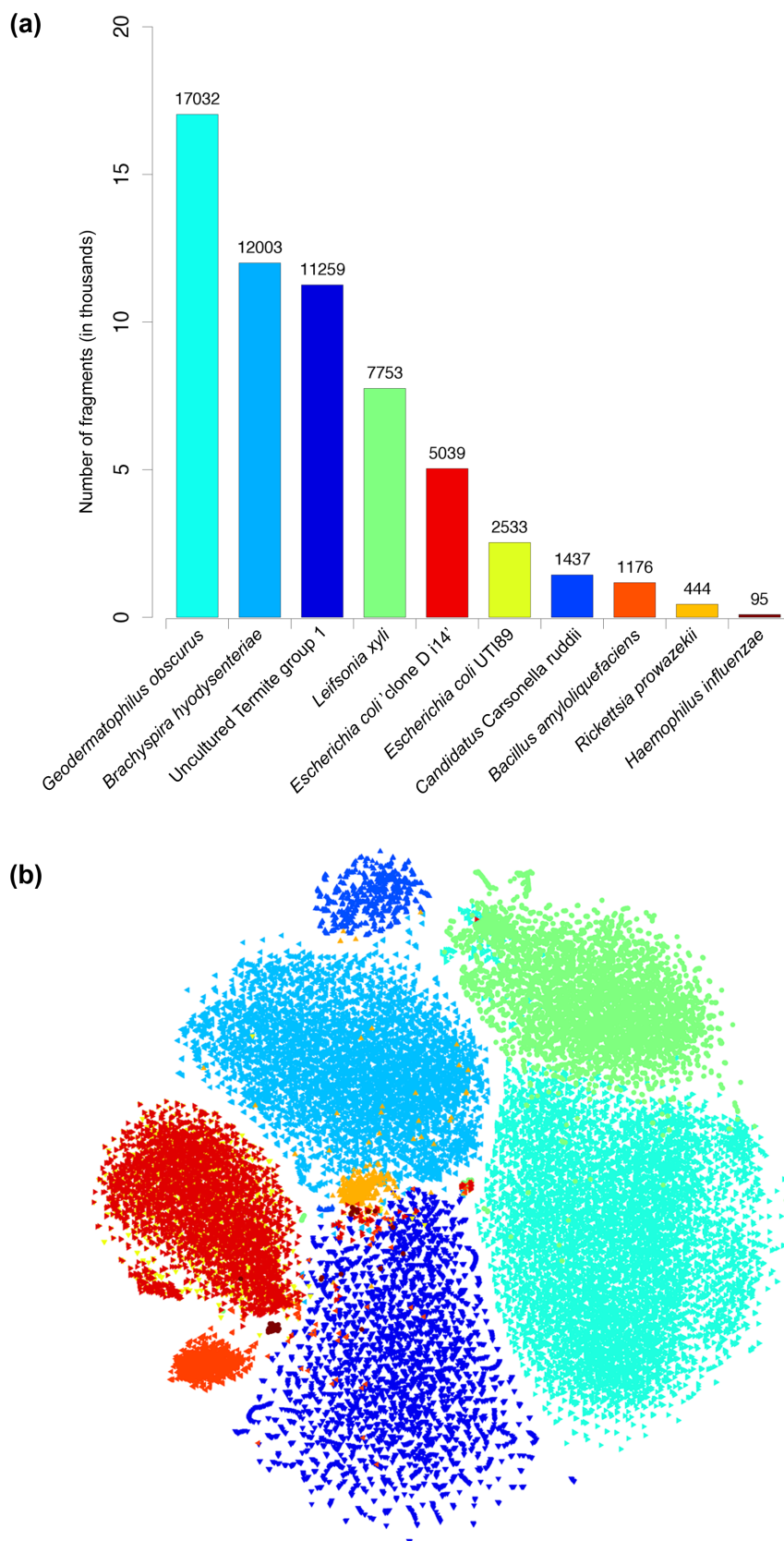
**Figure 2 | ESOM-based U-Matrix visualization of genomic fragment signatures for the evenly distributed simulated microbial community (EqualSet01).** (a) The topological map obtained via the computation of the U-Matrix based on the ESOM training on the genomic fragment signatures for the selected microorganisms. (b) Floodfill with default threshold of 0.2. Two areas were selected (yellow and blue). (c) Overlay of known information on learned topological representation. (d) Floodfill with more stringent threshold of 0.1. Numerous artificial subgroups are apparent. (e) Floodfill with more loose threshold of 0.3. The entire map is filled with two subgroups. The color coding is identical to that in Figure 1.

Fig. S7c online). The congruency of the fills was then compared against the known ground truth (see Supplementary Fig. S7d online), revealing that two of the three major areas comprise at least two organisms each (*Geodermatophilus obscurus* with *Leifsonia xyli* and *Brachyspira hyodysenteriae* with *Candidatus Carsonella ruddii*, respectively; see Supplementary Fig. S7c online). This highlights the lack of separation of the ESOM-based approach for sequence fragments if these follow a typical rank abundance distribution, in contrast to the results obtained using BH-SNE.

*Application of the method to an even community of closely related organisms.* In order to assess the performance of the methods

(BH-SNE- or ESOM-based) on genomic fragments from closely related organisms, we applied them to EqualSet02 which contains genomic fragments from three distinct genera. The application of our BH-SNE-based approach to EqualSet02 results in a representation depicting distinct, congruent clusters (see Supplementary Fig. S8 online). In the scatter plot, closely related organisms, such as *Shewanella baltica* and *Shewanella putrefaciens* cluster together. However, *Shewanella amazonensis* forms a distinct cluster, clearly separated from the fragment clusters of the other *Shewanella* species. A similar situation can be observed for the *Lactobacilli*.

Due to the close relationships which are expected in EqualSet02 (see Supplementary Fig. S1 online), we decided to perform the



**Figure 3 | Visualization of an unevenly distributed simulated microbial community genomic dataset (LogSet01).** (a) Simulated logarithmic rank abundances of genomic fragments per taxon. (b) BH-SNE scatter plot of log-ratio transformed pentanucleotide signatures of genomic fragments which have a length of 1,000 nt. The color coding reflects the organismal origin of the represented genomic fragments as colored in (a). Colors are added for visualization purposes only. Details on the ten organisms are provided in Supplementary Table 1.



computation of the ESOM and U-Matrix in several different configurations. First, we used the parameters as reported by Dick *et al.*<sup>7</sup> for the grid-resolution on the robust z-transformed tetranucleotide signatures of the genomic fragments. This results in around 17,000 neurons (104 by 170) and represents a marked increase in resolution of the ESOM compared to the default resolution of around 4,000 neurons (50 by 82). We ran the training for 100 epochs (see Supplementary Fig. S9 online). Following this, in order to examine potential differences in the topological maps when using pentanucleotide signatures, we used as input the same CLR-transformed pentanucleotide frequencies as used for our BH-SNE-based visualization (Methods) and used the default parameters (around 4,000 neurons) or the reported parameters (around 17,000 neurons) with the default 20 training epochs (see Supplementary Fig. S10 online). Both of these configurations result in maps with similar topologies. However, no representative congruent clusters are visible any longer once the labels are removed (see Supplementary Fig. S10a–d online). Notably, as these results show, the map representation is dominated by ridges that create a high number of individual, very small valleys for all of the herein chosen configurations. Finally, performing a “floodfill” on these maps yields no representative selection of congruent areas (see Supplementary Fig. S10e–h online).

**Real-world datasets.** *Characteristics of the real-world datasets.* If our approach is to be broadly applicable to metagenomic data, it must enable the visualization of real-world data from mixed microbial communities. To investigate this, we first used metagenomic sequencing data from a microbial community derived from an acetate-amended aquifer. Recently described by Wrighton *et al.*<sup>2</sup>, this arguably represents the most well curated and validated population-resolved metagenomic dataset to date. As this dataset was originally characterized using an ESOM-based approach, it allows for an independent comparison with our BH-SNE-based approach. Moreover, due to the recently emerging interest in human host-associated microbiota, we applied our approach to metagenomic data by Arumugam *et al.*<sup>3</sup>, for an arbitrarily chosen sample (SampleA; Supplementary Note 2 online). Both of these datasets were analyzed based on genomic fragments of a minimum size of 1,000 nt. Finally, we applied our approach to a ubiquitous and presumably highly complex microbial ecosystem, the deep sea as characterized by Konstantinidis *et al.*<sup>1</sup> (see Supplementary Note 3 online). Here, we focused on minimum sequence lengths of 1,500 nt and 2,000 nt.

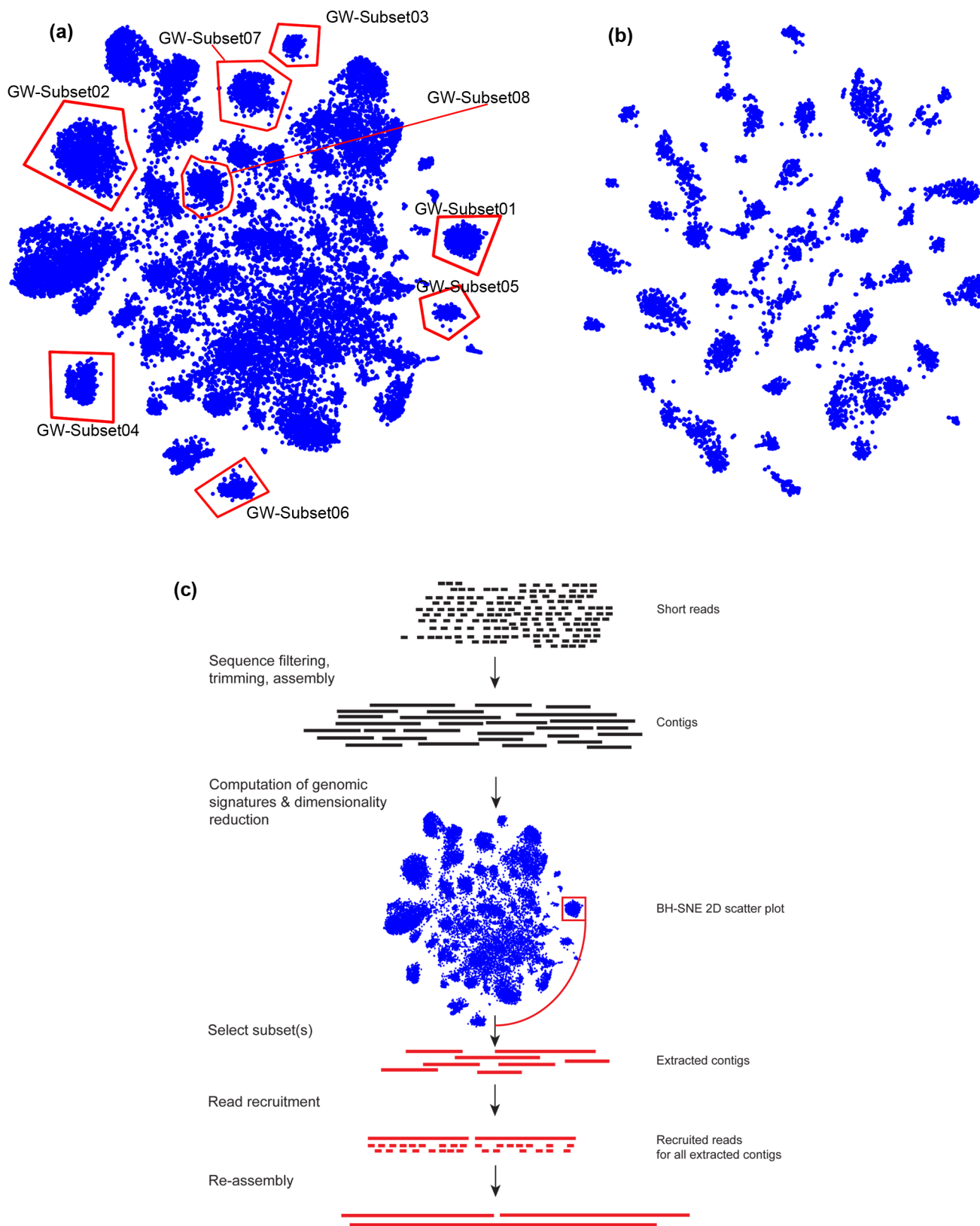
For each dataset, cluster delineation was carried out using human-augmented polygonal selection. Independently, we applied a (semi-)automated approach that models the cluster-of-interest as one component (“foreground”) and the remaining points as the second component (“background”) of a two-component Gaussian Mixture model (Methods). This clustering approach relies on the manual placement of a single initial mean of the “foreground” component. Contrary to the multi-component model used for the simulated data, for which the number of components (i.e., number of organisms in the community) is known *a priori*, this two-component approach is designed for the selection of an individual, local cluster of interest. The final parameters (mean, covariance, etc.) of the two components (i.e., of the “foreground” and of the “background”) are estimated/learned using optimization by Expectation-Maximization (Methods). Generally, the thus-selected points showed a high overlap with the polygonally selected points (see Supplementary Table S4 online). Accordingly, we report the results in the following based on the human-augmented, polygonal selection. Since the identities of genomic fragments with respect to their originating organisms are generally not known for real-world datasets, we computed the sensitivity, specificity, and precision values of the selected subsets based on sequence homology to previously characterized reference genomes, wherever available. Because the genomic complements of

populations resolvable through metagenomics will differ from the genomic sequences of their isolated counterparts (in particular, in terms of sequence identity), these values are provided to allow general assessment of the congruence of the clusters.

**Groundwater microbial community.** From the BH-SNE-based two-dimensional embedding of the groundwater microbial community, distinct groups are apparent (Figure 4a). To examine the congruency of the groups, we selected eight clusters from the resulting two-dimensional scatter plot and inspected how the grouped contigs map onto the published complete, near-complete, or at least partially reconstructed genomes<sup>2</sup>. All of the contigs that were mapped to these reconstructed genomes aligned in full-length and with 100% identity. For seven of the eight selected groups, 97% of the contigs could be mapped back to their respective reconstructed reference genomes. The remaining subset (GW-Subset01; Figure 4a) mapped against two genomes: 90.1% of the contigs mapped to genomic fragments associated with the organismal group called ACD47, whereas 9.9% mapped to ACD59. Clearly, ACD47 appears to be the dominant genome, but the comparably large proportion of contigs mapping to ACD59 led us to also further inspect this cluster. Examination of the %GC-content of the 326 scaffolds (contigs linked by paired-end/mate-pair information) of the reconstructed ACD47 genome reveals a uniform group. In contrast, the %GC content of the scaffolds of ACD59 shows a bimodal distribution with 28 of 75 scaffolds having an unexpectedly low %GC (see Supplementary Fig. S11 online). All the contigs recovered as GW-Subset01 grouped with the higher %GC ACD47/59 scaffolds. In order to allow a better comparison with the ESOM available from the authors of the publication of the groundwater dataset (<http://genegrabber.berkeley.edu/>), a BH-SNE-based visualization was carried out on contigs  $\geq 5,000$  nt. The superior separation obtainable with the BH-SNE-based approach is apparent (Figure 4b). With such long sequences, the groups are strongly separated with only very minor overlaps in a few cases. Re-assembly of the reads mapping to the contigs recovered from BH-SNE as GW-Subset01 (Methods) resulted in an increase in all standard assembly metrics relative to the extracted contigs without the need for extensive manual curation (see Supplementary Table S5 online). The workflow shown in Figure 4c exemplifies this iterative re-assembly approach for community genomic fragments based on the initial two-dimensional mapping obtained by BH-SNE.

Due to the lack of independently characterized reference genome sequences, we inspected the quality of the eight BH-SNE-based subsets and their corresponding ESOM-based clusters using previously described sets of single-copy marker genes<sup>18–20</sup>. These sets of conserved genes allow estimation of the completeness of a genome as well as the integrity of the set of genomic fragments by assessing the number of single or multiple copies within a given cluster. Concretely, we used the more recently applied 107 universal, single-copy genes, referred to in the following as “essential genes”<sup>19</sup>. Five of the eight selected BH-SNE-based clusters are over 50% complete, with four of them over 80% complete (Table 2). Overall, our approach identified clusters with lower amounts of multiple essential gene copies and higher amounts of single essential gene copies compared to the bins defined in the original publication<sup>2</sup> (Table 2).

**Human microbiome.** To further prove the broader applicability of the method, we applied it to metagenomic data from the human gastrointestinal tract. As for the groundwater data, BH-SNE-based visualization resulted in distinct clusters which we further evaluated by focussing on specific subsets (Figure 5; Supplementary Note 2 online). For this, we chose three groups consisting of 318 (HM-Subset01), 323 (HM-Subset02), and 282 (HM-Subset03) contigs, respectively. For HM-Subset01 and HM-Subset02 at least 85% of the contigs resulted in significant alignments against isolate genomes from *Escherichia coli* or *Eggerthella lenta*, respectively. Accordingly, the sensitivity (completeness) for the selected HM-Subset01 and



**Figure 4 |** BH-SNE-based visualization, human-augmented binning and population-level assembly of groundwater metagenomic data. (a) Genomic fragments have a minimum size of 1,000 nt. Red polygons contain mappings of genomic fragment signatures and represent congruent subsets of interest. These subsets were further analyzed as described in the text. (b) Genomic fragments have a minimum size of 5,000 nt. (c) Exemplified workflow for the iterative population-level assembly based on subsets selected from the BH-SNE scatter plot.



**Table 2 | Results of the “essential genes” (in single or multiple copies) analyses<sup>19,20</sup> on the groundwater dataset**

Original <sup>†</sup>	Single	Multiple	BH-SNE-based	# contigs	Single	Multiple
ACD20	95	11	GW-Subset05	225	99	5
ACD39	20	3	GW-Subset02	1375	18	5
ACD47	40	1	GW-Subset01	544	43	2
ACD59	5	0	GW-Subset01	544	43	2
ACD62	91	0	GW-Subset04	513	92	0
ACD64	48	0	GW-Subset03	214	47	0
ACD6	65	22	GW-Subset07	601	80	9
ACD73	55	0	GW-Subset08	541	55	1
ACD77	85	6	GW-Subset06	309	85	0

<sup>†</sup>: Identification based on a majority vote rule with respect to the originally reconstructed genomes.

HM-Subset02 is at least 92%, and the specificity (homogeneity) is above 99%, respectively (Table 1; Supplementary Note 2 online). The precision (fraction of fragments in a cluster belonging to the correct genome) is 92.45% for HM-Subset01 and 78.95% for HM-Subset02. The lower precision for HM-Subset02 is due to our stringent criteria for query length covered ( $\geq 95\%$ ; Methods). Given that the genomic complements of populations resolvable through metagenomics will differ from the genome sequences of their isolated counterparts, this observation is expected. Using a threshold of at least 85% of the query length covered results in a precision value of 85.45% (HM-Subset02). Finally, for HM-Subset03, none of the contigs reported a significant hit when using a query coverage above our threshold (Methods). In fact, 194 of 282 did not report any hit to the latest version of NCBI’s Genbank non-redundant nucleotide database and, thus, tentatively represent genomic sequences from an *hitherto* uncharacterized organismal lineage. Due to this, the genomic fragments of this population could not be linked to a cultured representative and, accordingly, no sensitivity, specificity, or precision values could be computed for this subset. In-depth characterization of this subset of sequences and the respective microbial genome/population is of pronounced scientific interest but goes beyond the scope of the present work.

**Marine microbial community.** Finally, in order to assess the performance of the method on a very diverse microbial community dataset, we chose a publicly available high-quality metagenomic dataset which had been generated on microbial communities from the deep-sea<sup>21</sup>. Here, a minimum cut-off of 2,000 nt had to be set to result in clear distinction of clusters and to resolve distinct sequence groups using our BH-SNE-based method (see Supplementary Fig. S12 online). In order to assess the quality of the clustering, two selected subsets (DS-Subset01 and DS-Subset02; Supplementary Note 3 online) were further investigated. For DS-Subset01, around 10% of the contigs reported a significant alignment against uncultured marine microorganism HF4000-related fosmid sequences from Konstantinidis *et al.*<sup>21</sup>. The remaining contigs did either not result in any significant hit at all or the results were below relevant thresholds (Methods). Aligning the contigs of DS-Subset02 revealed more than 50% of all the contigs to map to fosmid sequences of Uncultured Group I marine crenarchaea HF4000<sup>21</sup>. Both organismal groups represent organisms not fully characterized so far. As highlighted for HM-Subset03, the detailed further characterization of these congruent sequence clusters, in particular DS-Subset01, should be carried out but this goes beyond the scope of the present work. Therefore, for this dataset, no sensitivity or specificity values could be computed for any of the deep-sea subsets as the numbers of contigs with significant and high-quality alignments were limited.

**Runtimes.** For all the simulated datasets analyzed in this work, our BH-SNE-based approach was markedly faster than the ESOM-based approach (see Supplementary Table S6 online). More specifically, reported speedups range from around 3.8-fold to 50.4-fold. The

runtime of the ESOM-based approach is largely dependent on two parameters: the resolution of the grid (number of neurons;  $m$ ) and the number of datapoints ( $n$ ), here number of genomic fragments. Thus, the runtime complexity is  $O(mn)$  when both parameters are varied. The grid-resolution ( $m$ ) may be kept constant, but as suggested by Dick *et al.*<sup>7</sup>, it should be a linear function of the number of datapoints, resulting in overall runtime complexity  $O(n^2)$ . BH-SNE, in contrast, is robust with respect to the parameters. We have therefore used the default parameters (Methods). The runtime complexity of BH-SNE is  $O(n \log n)$ . Accordingly, greater speedups for our BH-SNE-based approach are expected when the number of datapoints ( $n$ ) is increased. Concretely, with the previously reported parameters<sup>7</sup>, the estimated runtime of the ESOM-based approach on EqualSet01 would be almost 11 h, when assuming a perfectly linear runtime-increase with respect to the grid-resolution ( $m$ ). The runtime of 225 seconds of our approach represents a speedup of around 174-fold. A summary of runtimes using our BH-SNE-based approach is provided in Table 3 for all herein analyzed simulated and real-world datasets.

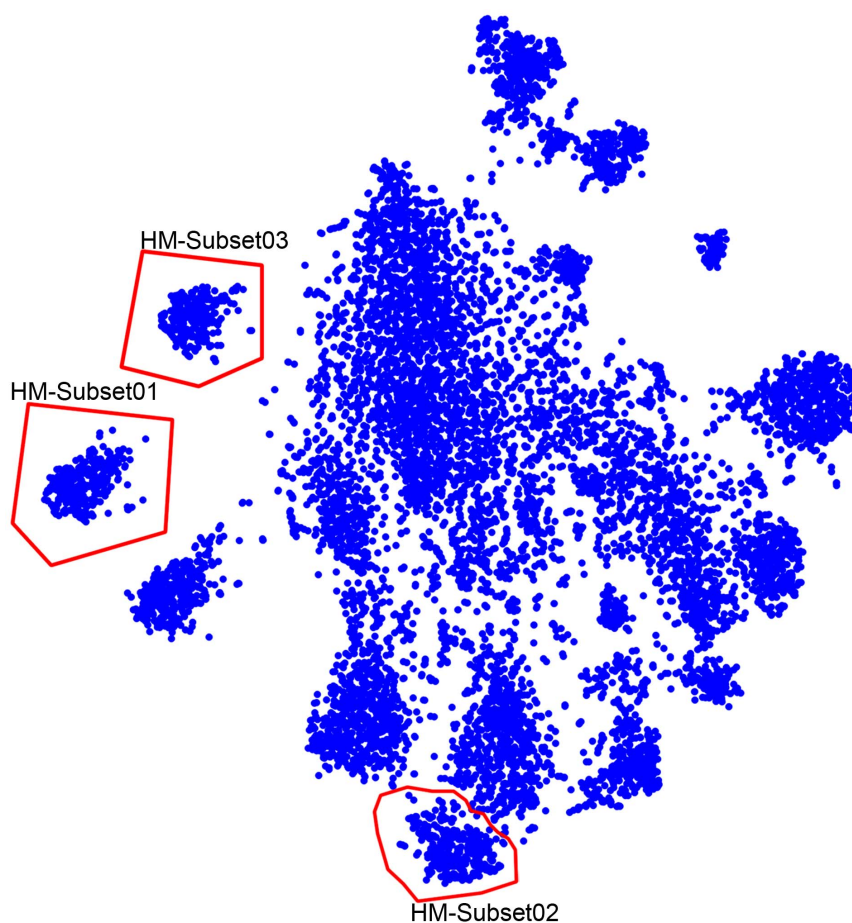
## Discussion

We report an approach for the efficient alignment-free visualization of metagenomic data based on the combination of centered log-ratio transformed oligonucleotide frequencies with BH-SNE. Our approach allows for the faithful two-dimensional representation of the data-inherent taxonomic structure as illustrated by the high degree of differentiation of genomic fragments (Figure 1), even those of closely related organisms (see Supplementary Fig. S8 online). The sensitivity, specificity, precision and essential gene analysis results reported herein (Table 1; Table 2) objectively attest to the potential of our visualization approach followed by low-dimensional clustering of metagenomic data. Thus, the described method holds great potential for its application to low to medium complexity microbial communities.

As demonstrated by the results reported for the simulated datasets, our approach enables the visualization, and subsequent binning of genomic fragments at least at the genus level and possibly at the species level (Figures 1, 3 and Supplementary Fig. S8). Moreover, the reported two-dimensional visualizations illustrate that only minor amounts of genomic fragments may fall into clusters of disparate organismal origin, a claim which is supported quantitatively by the reported sensitivity, specificity, and precision values (Table 1). These misplaced sequences may potentially represent recently transferred genomic elements which therefore may still exhibit the oligonucleotide signature of the donor genome. Furthermore, the ground truth datasets enable us to observe that, while commonly used marker genes (*recA* or *rpoB*) exhibit an association to their respective genomes of origin (see Supplementary Fig. S5c–f online), the 16 S rRNA gene sequences of these organisms do not exhibit such a behavior but rather tend to form distinct clusters (see Supplementary Fig. S5a and b online), likely due to an overall higher amount of sequence conservation and accordingly similar pentanucleotide frequencies. However, it should be noted that 16 S rRNA gene sequences might be correctly placed in real-world datasets if the signatures of flanking sequences (up- and downstream of the 16 S rRNA genes themselves) adequately represent the genomic signature of the organism of origin.

Apart from applying the described method to simulated datasets, we also assessed its performance on a collection of real-world datasets. Comparison of our results on the groundwater dataset (Wrighton *et al.*, 2012) to the original results obtained by the state-of-the-art ESOM-based approach demonstrates a general agreement as well as an interesting difference in the resulting clusters. Specifically, our results suggest a heterogenous genomic composition for the reconstructed ACD59 genome. This observation is supported by the small amount of essential genes identified in the genomic





**Figure 5** | BH-SNE-based visualization of genomic fragment signatures for the contigs of the human microbiome dataset. Red polygons represent subsets of interest that were further characterized as detailed in the text. Contigs have a minimum length of 1,000 nt.

fragments of this cluster (Table 2) and these genes formed part of GW-Subset01 along with the other essential genes of ACD47, thereby completing the population-level genome of ACD47. Apart from having identified these misassigned genomic fragments, our method also resulted in more complete (higher number of single essential gene copies) and more homogeneous (lower number of multiple essential gene copies) clusters compared to the original ESOM-based results (Table 2).

While this manuscript was under review, Gisbrecht *et al.* reported similar work on cluster identification in metagenomic samples using genomic signatures and nonlinear dimension reduction<sup>22</sup>. Using data consisting of genomic fragments of length 10,000 nt from 21 known bacterial genomes, represented as points in 136D (they used tetramer-based genomic signatures), they applied three dimension reduction methods, principal component analysis (PCA), generative topographic mapping (GTM; the probabilistic counterpart of self-organizing maps), and t-distributed SNE (t-SNE), followed by

clustering of the two-dimensional points by affinity propagation. Similar to our findings, Gisbrecht *et al.* reported that the combination of t-SNE with clustering in two dimensions yields the best results (in terms of increased accuracy and higher Dunn index), as compared to other dimension reduction methods or direct clustering of the data in the signature (136D) space. This suggests that binning of metagenomic data, either by a manual approach or by (semi-)automated clustering, can strongly benefit from prior reduction of the dimensionality by SNE-based approaches, followed by human-augmented clustering of the data. In our work we have employed the BH-SNE algorithm on CLR-transformed pentanucleotide signatures instead of t-SNE on raw tetranucleotide signatures. The CLR-transformation improves the regularity and congruency of resulting sequence clusters in the two-dimensional scatter plot (see Supplementary Fig. S2 online). Additionally, based on the reported results (see Supplementary Fig. S2 online), the use of pentanucleotide signatures results in improved separation of clusters following

**Table 3** | Sequence characteristics of the simulated and real-world datasets and corresponding BH-SNE-based runtimes

Type	# sequences	Total size (nt)	BH-SNE runtime (s)
EqualSet01 (1,000 nt)	29,212	29,212,000	225
LogSet01 (1,000 nt)	58,771	58,771,000	479
EqualSet02 (1,000 nt)	3,194	3,194,000	15
Groundwater ( $\geq 1,000$ nt)	25,278	97,849,086	148
Groundwater ( $\geq 5,000$ nt)	4,827	49,909,353	20
Human-derived ( $\geq 1,000$ nt)	9,911	18,291,540	58
Marine ( $\geq 1,500$ nt)	19,639	39,528,304	137
Marine ( $\geq 2,000$ nt)	5,743	16,154,498	27



nonlinear dimension reduction via BH-SNE. The improved computational efficiency of BH-SNE ( $O(n \log n)$  where  $n$  is the number of genomic fragments) compared to t-SNE makes the former more appropriate in real-world applications due to the generally large amount (in the order of tens of thousands) of genomic fragments.

Microbial communities of high complexity, such as those found along the human gastrointestinal tract, impose particular challenges to the comprehensive, low-dimensional visualization via BH-SNE (Figure 5). Nevertheless, using our method, we were able to recover two clusters of genomic fragments with significant homology to existing reference genomes (HM-Subset01 & HM-Subset02) as well as a cluster of genomic fragments (HM-Subset03) which appears to represent a dominant but previously uncharacterized microbial population from the human gastrointestinal tract. As recently reported by Sunagawa *et al.* (2013), the majority (i.e., on average 58% of the microbial species richness) of human-microbiome derived microbial species are currently uncharacterized<sup>23</sup>. To fully resolve communities of high complexity, an iterative binning and re-assembly approach involving repeatedly removing the most prominent clusters can be applied, analogous to the workflow applied to GW-Subset01 and detailed in Figure 4c. However, such an approach may ultimately be limited by the size and quality of contigs which are obtainable via assembly from metagenomic read data. As highlighted, the current minimum contig length for our method to result in optimal visualizations is 1,000 nt. Using currently available sequencing techniques and assembly algorithms, such contig lengths are likely unobtainable from low abundance organisms. Nevertheless, advances in DNA sequencing technologies, yielding reads of 1 kbp, will allow the direct application of our method to raw reads thereby allowing dedicated genome assemblies following read-based binning. While such data was unavailable to us at the time of writing, the reported improved assembly results (e.g. longer overall length with smaller number of contigs) on the short read-based groundwater dataset demonstrate the potential of such iterative assembly strategies (see Supplementary Table S5 online). Alignment-free visualization and subsequent clustering of *hitherto* prohibitive amounts of metagenomic sequence data will be essential for high-resolution population-level genomic reconstructions, which in turn will be prerequisites for the proper interpretation of functional high-resolution omic data<sup>24</sup>. In line with this, our approach demonstrates marked superior performance in terms of reduced input sequence length, discriminatory power, and computational cost over state-of-the-art ESOM-based approaches. Given the demonstrated power of nonlinear dimension reduction methods for visualization of metagenomic data, optionally followed by binning, further development of such methods is necessary to allow them to be applied to the ever-growing metagenomic datasets in the future.

An interesting question is whether the distances between clusters in the two-dimensional scatter plot may match the phylogenetic distances between corresponding organismal groups. Given that BH-SNE is a general nonlinear dimension reduction method which aims to represent the neighborhood structure of multidimensional data, the relative positions of the clusters in two dimensions need not *a priori* reflect phylogenetic distances between microbial populations. In order for the clusters to explicitly preserve such distances, an adaptation of the BH-SNE model and algorithm to incorporate phylogenetic distances is deemed necessary but goes beyond the scope of the present work.

## Methods

All the operations described in this work were performed on a standard laptop computer (MacBook Pro (Late 2011) with a 2.8 GHz Intel Core i7 and 8 GB 1333 MHz DDR3 memory and Mac OS X Lion 10.7.5 (11G63b)).

**Simulated data.** Complete microbial genome sequences were retrieved from the NCBI microbial genomes database (<ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/>). Ten bacterial species were selected at random. An overview of the organisms, the respective genome size and %GC of each genome can be found in Supplementary

Tables S1–S2 (online). Sequence fragments of specific lengths were generated using Grinder<sup>25</sup>. The following parameters were used: -rd 1000 normal 0 -cf 1, while also applying different lengths for the sequences, e.g. -rd 800, to assess the separation of groups based on the sequence length. The coverage parameter was specified according to the desired abundances (equal or logarithmic). Uniformly distributed sequencing errors of 1%, 3%, and 5% were introduced (e.g., -md uniform 1 for a 1% error) to assess the separation of groups given different error rates.

**Groundwater data.** Sequence data for the groundwater metagenome<sup>2</sup> were obtained via the Whole-Genome Shotgun repository service of NCBI GenBank (contigs for record AMFJ01) as well as from the ggKbase repository for the reconstructed genomes of organisms from the Rifle site (<http://ggkbase.berkeley.edu/rifle/organisms/>)<sup>2</sup>. Contigs  $\geq 1,000$  nt were used as provided. The sequence fragments were not split into uniform fragment sizes. Overall, this resulted in 25,278 sequences with a total length of 97,849,086 nt.

**Human microbiome-derived data.** Sequence data for the human-derived microbial community<sup>3</sup> was obtained from the project's website ([http://www.bork.embl.de/Docu/Arumugam\\_et\\_al\\_2011/downloads.html](http://www.bork.embl.de/Docu/Arumugam_et_al_2011/downloads.html)). The assembled contigs for the randomly chosen individual sample (A) were downloaded (A.contigs.fa.gz). Contigs  $\geq 1000$  nt were used as provided, without splitting them into uniform fragment sizes, overall resulting in 9,911 sequences with a total length of 18,291,540 nt.

**Marine data.** Assembled sequence data for a deep-sea (4,000 m) marine microbial community<sup>4</sup> was obtained from the Whole-Genome Shotgun repository service of NCBI GenBank (contigs for record ABEF01). Contigs were used as provided, without splitting them into uniform fragment sizes. Contigs  $\geq 1,500$  nt yielded 19,639 sequences with a total length of 39,528,304 nt and contigs  $\geq 2,000$  nt yielded 5,743 sequences with a total length of 16,154,498 nt.

**Marker gene sequences.** Three genes were used as markers: 16 S rRNA, *recA*, and *rpoB*. The sequences for 16 S, when available, were retrieved from the website of the Ribosomal Database Project<sup>26</sup>. The sequences for *rpoB* and *recA*, when available, were retrieved from the NCBI Nucleotide database.

**Cluster selection based on Gaussian Mixtures. Multi-component.** We used the *gmdistribution.fit* and *cluster* functions from the Statistics Toolbox in MATLAB Release 2013a (The MathWorks, Inc., Natick, Massachusetts, United States) for the two-dimensional multi-component Expectation-Maximization Gaussian Mixture model-based clustering. The initial means of the components were defined by manual placement on the BH-SNE-based visualization.

**Two-component.** We implemented the two-component Expectation-Maximization Gaussian Mixture model-based clustering in MATLAB. The code is available from the authors upon request. The mean of the “foreground” component was initialized by manual placement on the BH-SNE-based visualization. We refer the interested reader to Supplementary Note 4 (online) for more details.

**Computation of phylogenetic tree.** The webservice of phylogeny.fr was used for the generation of the phylogenetic tree<sup>27</sup>. More specifically, The “One Click” mode was used that requires the sequences of interest to be uploaded and returns the phylogenetic tree without intervention by the user. The following tools were used along this process: alignment by MUSCLE, curation by Gblocks, phylogeny by PhyML, and treeRendering by TreeDyn.

**Analysis of “essential” single-copy genes.** We followed the steps described in the Supplementary Information of Albertsen *et al.*<sup>20</sup> for the analysis of completeness and integrity of the genomic fragments of each selected subset (bin). This included the prediction of open reading frames in the genomic fragments using the metagenome version of Prodigal (-p meta)<sup>28</sup>. The predicted open reading frames were then searched against a set of 107 Hidden Markov Models (HMMs) of essential single-copy genes representing 107 proteins found to be conserved in 95% of all sequenced bacteria<sup>19</sup>. The *hmmsearch* tool from HMMER3 (<http://hmmer.janelia.org/>) was used for searching with the default parameters, except that the trusted cutoff was used (-cut\_tc).

**Calculation of genomic signatures.** Genomic signatures were computed as follows: First, the count of occurrences for every *kmer* (e.g. 5mer) along an input sequence was calculated. Pairs of reverse complementary *kmers* were summed up, as described in<sup>10</sup>. This symmetrized signature has been shown to individually perform best when compared to other similar signatures<sup>8</sup>. Overall signature counts were incremented by a single unit to remove zero-counts. The resulting pseudocounts were normalized by dividing by the total sum. Elimination of zero-counts is in particular a requirement for the subsequent application of the centered log-ratio (CLR) transformation<sup>14</sup> on individual signatures to account for the constant-sum constraint. The implementation of BH-SNE used in this work performs an initial dimension reduction to a default of 50 dimensions using Principal Component Analysis (PCA). Accordingly, the high-dimensional genomic signatures computed for 5mers (or other  $k \geq 4$ ) are reduced to the first 50 principal components, thus improving the computational efficiency by creating smaller matrices for storage and processing.

**Visualization using BH-SNE.** The visualization based on BH-SNE results in two-dimensional scatter plots. Similar points in the high-dimensional space are depicted in close proximity to each other in the low-dimensional space, thus representing clusters. The following default parameters were used for BH-SNE: perplexity of 30, initial dimensions of 50, theta of 0.5.



**ESOM-based analysis.** The low-dimensional visualization of high-dimensional compositional data, such as oligonucleotide frequencies, based on the U-Matrix<sup>9</sup> results in a topological map, where dense regions regrouping closely related datapoints are shown in colors reflecting valleys (green) or lakes (blue). Accordingly, topographically separated areas should contain sequences that are distantly related. In this work, if not stated otherwise, robust z-transformed tetranucleotide frequencies, as reported by Dick *et al.*<sup>7</sup>, were used for computation of the ESOMs, with default parameters of the Databionics ESOM suite. ESOM-based cluster selection was achieved by applying the “floodfill” algorithm provided by the Databionics ESOM suite. This allowed for the selection of areas comprising closely related datapoints according to specified threshold parameter values.

**Mapping of contigs to reconstructed genomes, read recruitment, and genome re-assembly.** The sequences for contigs forming coherent groups in BH-SNE scatter plots were recovered, and searched against a BLAST database containing the 100 Rifle reconstructed genomes (from the groundwater metagenomic data<sup>2</sup>). Top hits representing a minimum sequence identity of 95% over a minimum 90% of aligned query sequence were used to classify the recovered sequences, and to confirm the homogeneity of the groups. Raw sequence reads for the Rifle datasets were obtained from the NCBI Sequence Read Archive, under accession number SRA050978.1. Reads from all three Rifle datasets (Arthur, Cristobol, and Dolly) were trimmed and filtered using the trim-fastq.pl script from the PoPoolation suite (with default parameters<sup>29</sup>), and recruited against the BH-SNE-recovered contigs using the custom mode ‘very-sensitive-local’ of Bowtie2<sup>30</sup>. Mapped reads were recovered from the quality-filtered read files, and used for a round of *de novo* assembly using idba\_ud v.1.1.0<sup>31</sup>. Complete and detailed information on read processing, assembly and genome scaffolding is lacking from<sup>2</sup>. Therefore, it was not possible to entirely replicate the assembly process implemented by Wrighton *et al.* and our re-assembly exercise does not allow nor attempt a direct comparison on assembly techniques.

**Mapping of contigs to reference genomes and evaluating the sensitivity, specificity, and precision of groups.** Similar to above, the sequences for contigs forming coherent groups (subsets) in BH-SNE scatter plots were recovered, and aligned using the BLAST webservice<sup>32</sup>, taking into account all reference sequences currently available through this service. Parameters were left default, except for the “Max target sequences”, which were set to 10 and sorted by “Max score”. Characterization of the output was based on the associated XML-file and a custom script was used to filter according to an *E*-value threshold of  $<10^{-30}$  and query length covered  $\geq 95\%$  (script available upon request from the authors). Each contig was assigned to the genome that reported the alignment (more specifically the Highest-Scoring Pair) with the highest “Max score”. In the case of multiple top hits with the same “Max score”, the first reported genome was chosen per contig. The taxonomic affiliation of the top hits conferred the taxonomic identity at the species and/or genus level of the subset-specific dominant population using a majority vote rule. All available genomes of the respective genus/species were then used to create a subset-specific, dominant population-based reference database which was used to define the gold-standard by aligning all the contigs used in the visualization to the respective database.

Given the availability of isolate reference genomes with high levels of identity to the population-level genomic complements resolved using our method, we computed the sensitivity, specificity, and precision based on the following definitions of true positives (TP), false positives (FP), false negatives (FN) or true negatives (TN):

TP Contigs in a subset aligning to the correct (dominant) reference genome(s)

FP Contigs in a subset not aligning to the correct (dominant) reference genome(s)

FN Contigs not in a subset but aligning to the correct (dominant) reference genome(s)

TN Contigs being neither TP, FP, nor FN.

The alignment parameters used are reported above. For the sensitivity, specificity, and precision analysis of the ground truth data, the respective originating microbial genome was known for every contig, hence no alignments were computed. Based on the above definitions, the sensitivity was computed as follows:

$$\text{Sensitivity} = \frac{|TP|}{|TP| + |FN|}$$

the specificity according to:

$$\text{Specificity} = \frac{|TN|}{|FP| + |TN|}$$

and the precision as follows:

$$\text{Precision} = \frac{|TP|}{|TP| + |FP|}$$

1. Konstantinidis, K. T., Braff, J., Karl, D. M. & DeLong, E. F. Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific subtropical gyre. *Appl Environ Microbiol* **75**, 5345–55 (2009).

2. Wrighton, K. C. *et al.* Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* **337**, 1661–5 (2012).
3. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174–80 (2011).
4. Chenc, T. Y. & Sueoka, N. Heterogeneity of DNA in density and base composition. *Science* **141**, 1194–6 (1963).
5. Teeling, H., Meyerdieck, A., Bauer, M., Amann, R. & Glöckner, F. O. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol* **6**, 938–47 (2004).
6. Wilmes, P. *et al.* Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. *ISME J* **2**, 853–64 (2008).
7. Dick, G. J. *et al.* Community-wide analysis of microbial genome sequence signatures. *Genome Biol* **10**, R85 (2009).
8. Gori, F., Mavroedis, D., Jetten, M. S. & Marchiori, E. Genomic signatures for metagenomic data analysis: Exploiting the reverse complementarity of tetranucleotides. In *2011 IEEE Int Conf Syst Biol*, 149–154 (IEEE, 2011).
9. Ultsch, A. & Mörchen, F. ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM. Tech. Rep., Data Bionics Research Group, University of Marburg (2005). URL <http://www.cbs.dtu.dk/courses/27618.chemo/ESOM.pdf>. Last access: 03 February 2014.
10. Abe, T., Sugawara, H., Kanaya, S., Kinouchi, M. & Ikemura, T. Self-Organizing Map (SOM) unveils and visualizes hidden sequence characteristics of a wide range of eukaryote genomes. *Gene* **365**, 27–34 (2006).
11. Sharon, I. *et al.* Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res* **23**, 111–20 (2013).
12. Van Der Maaten, L. Barnes-Hut-SNE arXiv:1301.3342v1 [cs.LG] (2013).
13. Kucera, M. & Malmgren, B. A. Logratio transformation of compositional data - a resolution of the constant sum constraint. *Mar Micropaleontol* **34**, 117–120 (1998).
14. Aitchison, J. *The Statistical Analysis of Compositional Data* (Blackburn Press, 2003).
15. 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).
16. Zhu, Y. *et al.* caBIG VISDA: modeling, visualization, and discovery for cluster analysis of genomic data. *BMC Bioinformatics* **9**, 383 (2008).
17. Redner, R. & Walker, H. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev* **26**, 195–239 (1984).
18. Hess, M. *et al.* Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* (New York, N.Y.) **331**, 463–7 (2011).
19. Dupont, C. L. *et al.* Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J* **6**, 1186–99 (2012).
20. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* **31**, 533–8 (2013).
21. Konstantinidis, K. T. & DeLong, E. F. Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *ISME J* **2**, 1052–65 (2008).
22. Gisbrecht, A., Hammer, B., Mokbel, B. & Szczyrba, A. Nonlinear dimensionality reduction for cluster identification in metagenomic samples. *17th Int Conf Inf Vis* (2013).
23. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nat Methods* **10**, 1196–9 (2013).
24. Muller, E. E. L., Glaab, E., May, P., Vlassis, N. & Wilmes, P. Condensing the omics fog of microbial communities. *Trends Microbiol* **21**, 325–33 (2013).
25. Angly, F. E., Willner, D., Rohwer, F., Hugenholtz, P. & Tyson, G. W. Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res* **40**, e94 (2012).
26. Cole, J. R. *et al.* The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37**, D141–5 (2009).
27. Dereeper, A. *et al.* Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res* **36**, W465–9 (2008).
28. Hyatt, D., LoCascio, P. F., Hauser, L. J. & Uberbacher, E. C. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **28**, 2223–30 (2012).
29. Kofler, R. *et al.* PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One* **6**, e15925 (2011).
30. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–9 (2012).
31. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–8 (2012).
32. Zhang, Z., Schwartz, S., Wagner, L. & Miller, W. A greedy algorithm for aligning DNA sequences. *J Comput Biol a J Comput Mol cell Biol* **7**, 203–14 (2000).

## Acknowledgments

We would like to thank Patrick May and Dilmurat Yusuf from the Luxembourg Centre for Systems Biomedicine, for their assistance and support. Furthermore, we would like to thank



Anders Andersson at the Science for Life Laboratory in Stockholm for fruitful discussions, assistance and support. The present work was supported by an ATTRACT programme grant (A09/03) and a European Union Joint Programming in Neurodegenerative Diseases grant (INTER/JPND/12/01) to PW and an Aide à la Formation Recherche grant (AFR PHD/4964712) to CCL all funded by the Luxembourg National Research Fund (FNR). We also thank the Luxembourg Centre for Systems Biomedicine and the University of Luxembourg for support of NP. This research includes results from GeneGrabber/ACD developed by the Banfield Laboratory at UC Berkeley with funding provided by the Subsurface Biogeochemistry and Genomic Sciences Programs, Biological and Environmental Research (BER), Office of Science, U.S. Department of Energy. The ACD metagenome was collected and developed with the support of the Integrated Field Research Challenge (IFRC) site at Rifle, Colorado. The Rifle IFRC Project is a multidisciplinary, multi-institutional project managed by Lawrence Berkeley National Laboratory, Berkeley, California for the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

### Author contributions

P.W., N.V. and C.C.L. designed the research. C.C.L. carried out the analyses. C.C.L., N.P. and P.W. analyzed the data. All authors participated in writing the manuscript. All authors reviewed the manuscript.

### Additional information

**Supplementary information** accompanies this paper at <http://www.nature.com/scientificreports>

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Laczny, C.C., Pinel, N., Vlassis, N. & Wilmes, P. Alignment-free Visualization of Metagenomic Data by Nonlinear Dimension Reduction. *Sci. Rep.* **4**, 4516; DOI:10.1038/srep04516 (2014).



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. The images in this article are included in the article's Creative Commons license, unless indicated otherwise in the image credit; if the image is not included under the Creative Commons license, users will need to obtain permission from the license holder in order to reproduce the image. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/3.0/>