Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Emerging Trends and Innovations in Electronics and Communication Engineering - ETIECE-2017

Volume 5, Special Issue 01, Jan.-2018 (UGC Approved)

OTU Clustering: A window to analyze uncultured microbial world

Ashaq Hussain Bhat[1], Puniethaa Prabhu[2]

[1,2]*Department of Biotechnology K. S. Rangasamy College of Technology, Tiruchengode, Tamil Nadu*

Abstract – *Clustering is the technique used to deal with higher amounts of data by partitioning that data in some groups based on some attributes. Clustering technique has many applications in different fields of science and technology. It is an important tool in metagenomics which performs taxonomic profiling of microbial world by grouping 16S rDNA amplicon reads into clusters called as Operational Taxonomic Units (OTUs). Taxonomic profiling of 16S rDNA is an important step in Metagenomic pipeline analysis. There are several OTU clustering algorithms whichclusters the amplicon reads of 16S rDNA into OTUs, each algorithm use a specific type of clustering technique. Some of the algorithms are Uclust, swarm, SUMACLUST, SortMeRNA, USEARCH, etc. So in this paper we are first going to give a view of clustering, its types and then OTU clustering algorithms and classifications.*

Keywords – *16S rDNA;OTUs; Uclust; SUMACLUST; SortMeRNA; USEARCH; taxonomic profiling*

I. INTRODUCTION

Clustering is also called as unsupervised learning is a meta-learning tool, which deals with the finding of natural structures based on some metric in a pool of unlabeled data. A cluster is therefore a group of objects which showsimilar patternsamongthemselves within a clusterbutdissimilar patterns to the objects belonging to other clusters. In different fields,clustering is referred with different names like cluster analysis, automatic classification, numerical taxonomy, typological analysis, etc.A good clustering means high quality clusters in which intra-cluster similarity is high and inter class similarity is low.The applications of clustering are very wide, it has lot of importance in different fields like it has been widely used in biological systems to gain insights in large-scale biological data, such as gene expression data [1], microbiome to study microorganisms in different environments, histone modifications[2], it has wide application in big data analytics [3], and nucleosome positioning [4], [5].Next Generation Sequencing (NGS) has changed the way of thinking towards the microbial communities. Metagenomics, the study of uncultured microbes from their environment, has evolved so much with the help of Pyrosequencing so that it's now racing in parallel with other big data sciences. Taxonomic profiling, using hyper-variable regions of 16S rDNA, is one of the important part in metagenomics. And Operational Taxonomic Units (OTUs) Clustering algorithms are the important tools to perform taxonomic profiling by grouping 16S rDNA reads into OTU clusters. There are several OTU clustering algorithms which clusters the amplicon reads of 16S rDNA into OTUs. Existing OTU clustering tools can be grouped into three approaches: closed-reference approach, de novo approach and open-reference approach. The closed approach matches input sequences against a reference database to perform OTU clustering. De novo approach clusters without using a reference database but instead take a sequence as seed, searches it against other remaining sequences and open-referencing is a hybrid of closed and denovo, it first uses the closed approach and after that denovo approach for those sequences which do not hit with reference sequences.Remaining part of the paper are organised in the following way. Section II discusses thoroughly various categories of clustering. Section III discusses OTU Clustering approaches. Section IV discusses various OTU Clustering algorithms. Section V shows the classification of OTU Clustering Algorithms. Finally section VI concludes the paper.

II. CLUSTERING CATEGORIES

In this section different types of clustering methods are discussed. Actually there is not a standard scale which can differentiate the various algorithms of clustering properly, because the different classes of algorithms overlap at sometimes. All types of algorithms are dividing the data in the clusters based on some characteristic threshold. In general clustering algorithms can be broadly classified as follows

2.1. Hierarchical-based Clustering: In hierarchical-based clustering algorithms, data are organized in a hierarchical manner by combining data into clusters and these clusters in bigger clusters, and so on. In this way it's creating a hierarchical like structure called as dendrogram. The dendrogram represents the whole dataset, where individual objects are the leaves of the tree, each leaf node represents the individual data item and interior nodes are nonempty clusters. There are two types of Hierarchical clustering methods agglomerative or bottom-up approach and divisive or top-downapproach. An agglomerative clustering is a bottom up approach and which starts with one object for each cluster and the recursively merges mostappropriate two or more clusters. On the other hand divisive clustering is top-down approach which starts with the whole dataset as one cluster and then splits in a recursive to the most appropriate clusters. The process continues until a threshold condition is satisfied (i.e. k number of clusters). The issue

with hierarchical clustering approach is that once a step (merge or split) is performed, this cannot be not be done again. The main examples in this method are BIRCH, CURE, ROCK and Chameleon.

2.2. Partitioning Relocation Clustering: The partition based algorithms divide the data objects into a number of partitions, where each partition represents a cluster.Iterative optimization is used to relocate the data items between the clusters to improve the cluster quality unlike the hierarchical method where once cluster is created it's not revisited. The main thing is that each group should contain at least one data item, andeach data item must belong to exactly one group. The main classes of this type are:

2.2.1. Probabilistic clustering: In this approach, the dataset is assumed as sample independently drawn from mixture model of several probability distributions. Let the randomly picked model j has probabilityt_j, j=1: k, and point x is drawn from corresponding probability. Point x is believed to belong only one cluster, to estimate the probability of point x:

$$\Pr(X|C) = \prod_{i=1:N} . \sum_{j=1:k} t_j \Pr(x_i | C_j$$

2.2.2. K-Medoids Methods: In K-medoids algorithm objects which are near the centre represent the clusters. In other words we can say that cluster is represented by one central point. Medoids are not sensitive to outliers because they do not affect them. PAM (Point around Medoids), CLARA (Clustering LARge Applications) and CLARANS (Clustering Large Applications based upon RANdomized Search) are important representatives of K-Medoids methods.

2.2.3. K-Means Methods: It is the simplest and most used clustering algorithm in which the centre is the average of all points and coordinates representing the arithmetic mean. The objective function used here is the sum of distances between elements of cluster and its centroid expressed through an appropriate distance function.

2.3. Density-based Methods: In Density-based clustering methodsdensity, connectivity and boundary are used to separate the data items into clusters based on their regions.The conceptis closely related to point-nearest neighbours.Depending upon the density a cluster can grow in any direction that density leads to. This type of methods locates the regions with high density which are separated from the regions with low density.For this reason density based algorithms can also form clusters of different or irregular shapes, and this provides anatural protection against outliers. The well known examples of density-based algorithms which are used to filter out noise (outliers) and discover clusters of arbitrary shape are DBSCAN,OPTICS, DBCLASD and DENCLUE.

2.4. Grid-based Methods: The methods that partition the space are frequently called as grid based methods, the space of the data items is dividedinto grids and each grid is called as a cluster. Grid-based methods have fast processing time, because such approaches go through the whole datasetonce to compute the statistical values for the grids and areindependent of the number of data items thatemploy a uniform grid to collect regional statistical data, and finally performs the clustering on the grid, instead to the database. The performance depends on the size of the grid andsize of the grids is less than the size of the database. The grid based methods contain both partitioning and hierarchical algorithms.Important examples of grid-based clustering are DENCLUE, CLIQUE,Wave-Clusterand STING.

2.5. Other Clustering Techniques: A large number of clustering techniques have been developed from time to time, each for a particular type of problem and each having its specific methodology. Some of them are as:

2.5.1. Constraint-Based Clustering: To find the clusters with certain satisfying limitations is the main field of current research. The constraints in constraint based clustering includes like constraints on individual objects, the parameter constraints, constraints in terms of bounds on aggregate functions, etc. Application of this type of methods is clustering two dimensional spatial data in presence of obstacles, like COD.

2.5.2. Graph-Based Partitioning: Graph based clustering is done by just simply deleting some of the edges from the main graph to get sub partitions. It's desirable to cut minimum edges but it is producing unbalanced clusters. Exact optimization of minimum cut leads to NP-hard. Some approaches of graph partitioning uses the idea of graph flows. The important application of graph partitioning is VLSI.

2.5.3. Artificial Neural Networks: The neural network approach uses a set of connected input/output units, where each connection has a weight associated with it. Neural networks have several properties that make them popular for clustering, like they are parallel and distributed processing architectures. And also neural networks get training by learning from their interconnection weights so as to best fit for the data. Neural networks process numerical vectors and require object patterns to be represented by quantitative features only. Many clustering tasks handle only numerical data or can transform their data into quantitative features if needed. The neural network approach to clustering tends to represent each cluster as an exemplar. An exemplar acts as a prototype of the cluster and does not necessarily have to

correspond to a particular object. New objects can be assigned to the cluster whose exemplar is the most similar, based on some distance measure.

2.5.4. Evolutionary Methods: The two important concepts used in evolutions methods include simulated annealing and genetic algorithms. The perturbation operator in simulated annealing techniques is used to relocate the points from the current to new randomly chosen cluster. These methods are mostly used in surveillance monitoring. Genetic Algorithms are used, for cluster analysis like for fuzzy and hard k-means, andclustering of categorical data. The limitation of evolutionary methods is that they have high computational cost hence are rarely used in data mining.

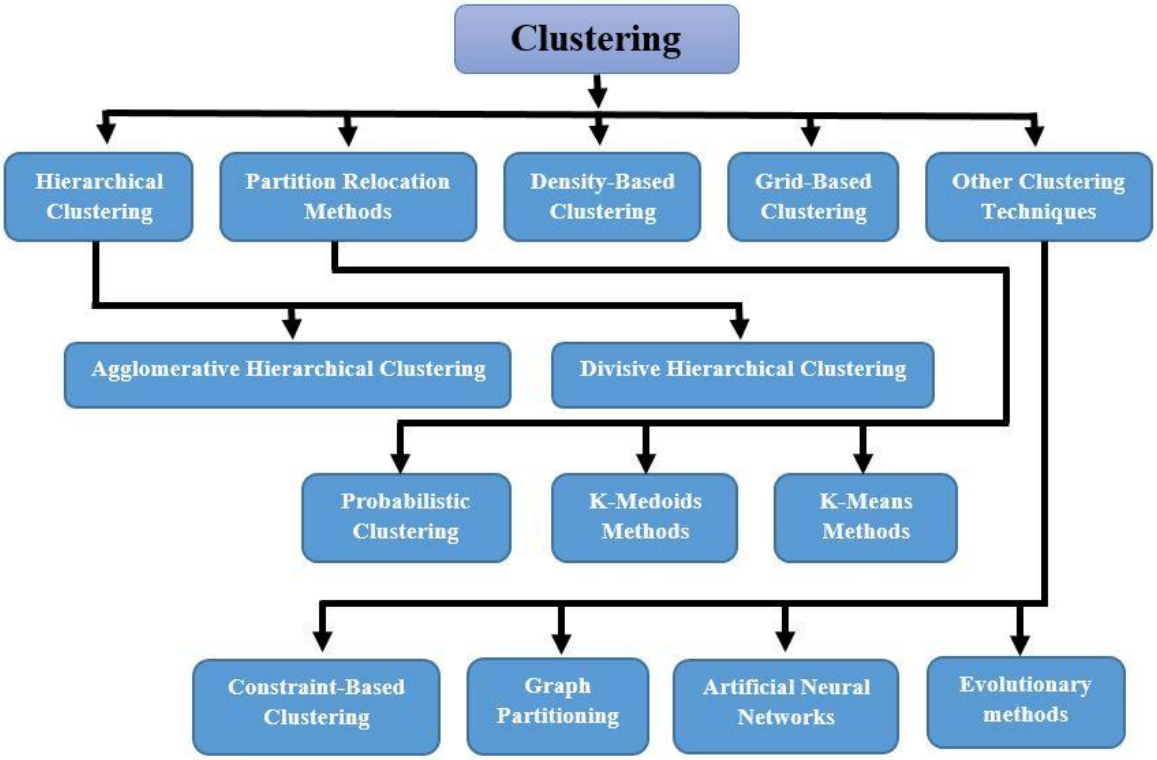

Figure 1.An overview of various clustering methods

III. OTU CLUSTERING

Next generation sequencing (NGS) includes the sequencing tools producing tremendous amount of data in less time. After sequencing the data, it is pre-processed before its going for clustering process. Clustering tools group the 16S rDNA sequences into clusters called Operational Taxonomic Units (OTUs). There are different types of OTU clustering tools or algorithms. These OTU clustering algorithms can be grouped in to three approaches: closed reference approach, de novo approach and open reference approach (hybrid of closed and de novo approach). In closed-reference approach the input dataset sequences are searched against a reference database like Greengenes to know the known microbes present in the data set. Although known microbes can beefficiently classified but this approach lacks the ability to find the novel species. According to the 'rare biosphere' theory [6], [7], there are still many microbes which have not been identified in existingreference databases. Therefore, grouping unknown microbes is an important task, for which the de novo approach is used. The de novo approach performs microbial profilingby grouping the 16S rDNA sequences of input dataset into OTU clusters. The open-reference approach is just combination of closed reference approach and de novoapproach or we can say it is a hybrid approach in which input dataset sequences are first searched against database i.e. closed-referencing and the rest of sequences which fail to cluster in closed referencing are given to de novo algorithm for clustering. Most existing studies and tools use threshold values of 97 and 95 percent for grouping at the species and at the genus level respectively. Depending on the way of forming clusters, most existing algorithms for the de novo approach can be further divided into two categories: greedy heuristic clustering(GHC) and agglomerative hierarchical clustering (AHC).

3.1. Greedy heuristic clustering is a partitioned based clustering method that works at a specific distance level at a time. Greedy clustering works by first choosing an input sequence as a seed and then each subsequent input read is compared against the existing set of seeds. If this sequence matches one of the seeds within a predefined level of 97 percent sequence similarity, it will be added to the cluster represented by that seed. Otherwise,it will be taken as a new seed. Examples in this category are UCLUST [9], USEARCH6, UPARSE [10], CD-HIT-OTU [17], and

QIIME'spick_otus [8]. UCLUST selects the seed of the cluster based on the percentage identity between a sequence and a seed.USEARCH and UPARSE perform a similar seed choice asUCLUST with additional filtering of clusters with low abundance i.e., small cluster sizes. CD-HIT-OTU groups similar sequences above 97 percent identity threshold and keepsthe longest sequence as seeds. QIIME's pick_otusimplements many reference-based and denovo OTU algorithms, but the UCLUST algorithm is the default method in QIIME. All GHC methods have linear time andspace complexities.

3.2. Agglomerative hierarchical clustering (AHC) is a clustering method works by computing on a pairwise genetic distance matrix derived from an all-against-all read comparison in a bottom-up manner. Examples in this category include Mothur [11], ESPRIT [12] and ESPRIT-Tree [13]. ESPRIT employs the traditional hierarchical approach of first computing an alignment-based all-against-all distancematrix and then performs either average-linkage or complete-linkage clustering on that matrix. ESPRIT reduces computational complexity by generating only the lower part of a dendrogram. The approach of Mothur and ESPRIT is similar but instead of pairwise global alignment used by ESPRIT,Mothur uses multiple sequence alignment tool such asMUSCLE [14] to compute the pairwise distance matrix. It has been seen that pairwise alignment produces better clustering outcomes than multiple sequence alignments [7], [15]. Different from ESPRIT and Mothur, ESPRIT-Tree uses both greedy and hierarchical strategies. Instead of seeds, it uses "probabilistic sequences" to present a group of similar sequences and then applies a BIRCH-like [16] clustering method to build and refine a "pseudo-metric based partition tree" of probabilistic sequences. ESPRIT-Tree has quasilinear space and time complexity [13]. In general the GHC approachesare often faster than the AHC approaches, but on the other hand AHC tools produce higher quality clusters than GHC tools [15]. The main drawback of the AHC approach is its high computational complexity and hence it is not suited for large datasets. Most existing OTU clustering methodsuse the threshold cutoff value of 97 percent sequence similarity. This de facto choice is based on the assumption that the pairwise genetic distance between a pair of 16S rDNA short reads from the same full-length 16S rDNA (hence from the same species) is less than 0.03. This assumption holds and hence is only applicable for datasets in which the pairwise distances between reads from the same species are less than 0.03 and the distances between reads from different species are larger than 0.03. When the distance distribution does not follow this assumption, a more flexible approach to determine the final OTU grouping is preferred.

IV. OTU CLUSTERING ALGORITHMS

There are different OTU clustering algorithms, some are closed source some are open and some works separately and some are embedded in different metagenomic sequence pipelines. QIIME is one the metagenomic software pipeline which employs many OTU clustering algorithms, but its default OTU clustering algorithm is UCLUST. The various OTU clustering algorithms mostly embedded in QIIME software pipeline are as:

4.1. Swarm [20],[21]is a *de novo* clustering algorithm which uses an unsupervised agglomerative hierarchical single-linkage clustering method.There are two steps in Swarm: (i) first the set of OTUs is constructed based on similarity of sequence reads by agglomerative clustering method (ii) Second the abundance value is calculated and which is then used to divide the OTUs into sub-OTUs if needed.

4.2. OTUCLUST[19]and **SUMACLUST,** use *de novo* clustering approach so no need of reference database. Both the algorithmsuse a greedy heuristic strategy which compares abundance-ordered list of input sequences against the representative set of already-chosen sequences which are initially empty and the clusters are made by increments [24].

4.3. UCLUST and **CD-HIT** also functions like that of OTUCLUST and SUMACLUST. But CD-HIT performs exact sequence alignment, rather than depending on fast heuristics. OTUCLUST is the default clustering algorithm of QIIME and also it performs its own sequence dereplication and chimera removal with the help of UCHIME [25]. And also UCLUST is used in all the 3 approaches i.e. closed reference, denovo and open referencing.

4.4. Mothur is also*de novo approach*and do not use any reference data base. It uses three methods for clustering of OTUs viz. single linkage, average linkage and complete linkage (nearest neighbour, average neighbour and farthest neighbour). All of these usegenomic distance for clustering the sequences. In single linkage a sequence is linked to an OTU if it is similar to any other sequence in that OTU. In the complete linkagea sequence is linked to an OTU if it is similar to all other sequences in that OTU. And in average linkage sequence is linked to an OTU if it is similar to the averaged similarity between all other sequences in that OTU [18], [25].

4.5. SortMeRNA [22] is closed-reference OTU clustering method and it needs a referencing database for clustering. Sequences from dataset are searched against the sequences in the database for matching and an *E* value threshold is applied to evaluate the quality of resultingalignments.The run time of SortMeRNA is not affected by reducing these thresholds as in UCLUST (e.g., clustering at 60% identity).

4.6. USEARCH and **UCLUS**T, both tools can perform all the three approaches that is de *novo*, closed-reference, and open-reference clustering. USEARCH uses UPARSE [10]which is a*de novo* amplicon analysispipeline. UPARSE has in buildstringent quality filtering, length trimming to remove erroneous reads, parallel chimera removal and also implements a novel greedy algorithm that performs OTU clustering.

V. CLASSIFICATION OF OTU ALGORITHMS

There are different and many OTU clustering algorithms, some implements hierarchical clustering technique and some use partitioned greedy heuristic based approach. So these algorithms can be classified in different ways. Here these are classified on the bases of whether they are open source or closed source (or proprietary software). The classification is given as under in figure 2:

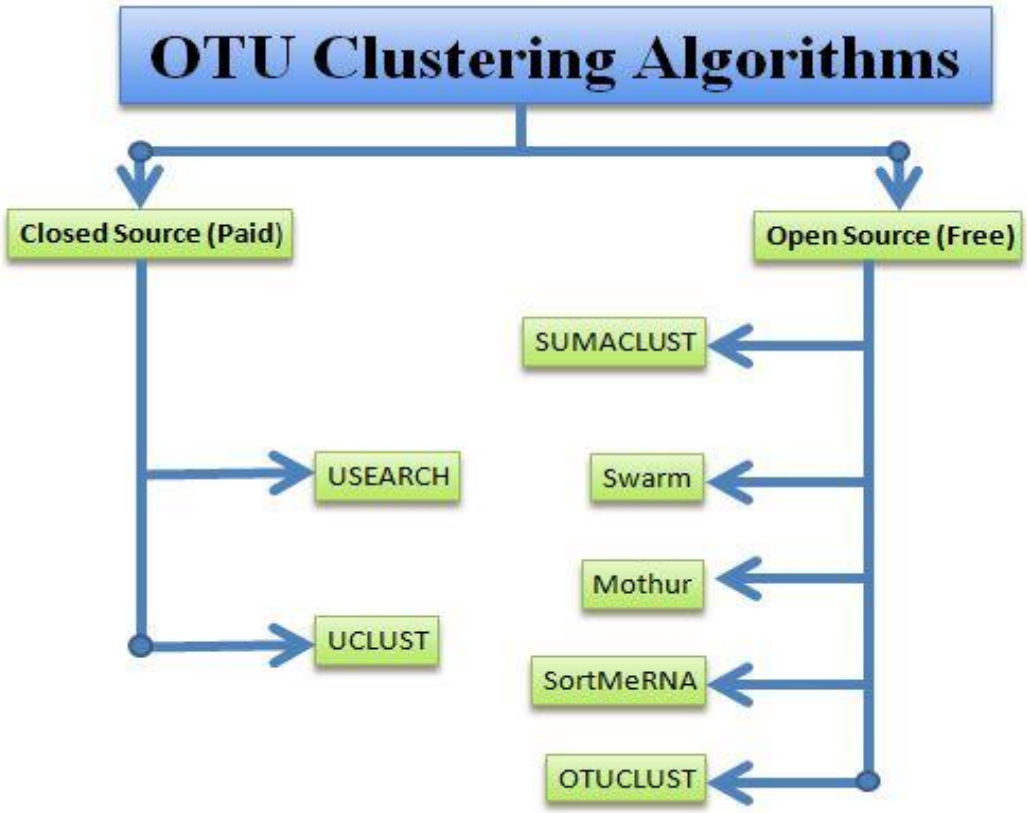

Figure 2.Classification of various OTU Clustering Algorithms

VI. CONCLUSION

Clustering is one of the essential tasks in data mining and needs improvement nowadays more than before to assist data analysts to extract knowledge from terabytes and petabytes of data in one time. The paper provides a detailed view of various clustering approaches and algorithms.The main focus of the paper is the application of clustering in Metagenomics and Genomicsfields of biology. A comprehensive details of various OTU clustering algorithms and their classification has been given so that researchers from the fields of computational sciences can get the idea of use of OTU clustering algorithm in their research work.As a future work we have to analyse properly all these algorithms and to see their advantages and disadvantages. So that we would develop a new algorithms, that will be efficient, scalable and can handle the massive amounts of data coming from Next Generation Sequencing (NGS) platforms.

REFERENCES

[1] P. D'haeseleer, "How does gene expression clustering work?" Nat. Biotechnol., vol. 23, pp. 1499–501, 2005.
[2] N. D. Heintzman, G. C. Hon, R. D. Hawkins, P. Kheradpour, A. Stark, L. F. Harp, Z. Ye, L. K. Lee, R. K. Stuart, and C. W. Ching,"Histone modifications at human enhancers reflect global celltype- specific gene expression," Nature, vol. 459, no. 7243, pp. 108–112, 2009.
[3] R. K. Chodavarapu, S. Feng, Y. V. Bernatavichute, P.-Y. Chen, H. Stroud, Y. Yu, J. a. Hetzel, F. Kuo, J. Kim, S. J. Cokus, D. Casero,M. Bernal, P. Huijser, A. T. Clark, U. Kramer, S. S. Merchant, X. Zhang, S. E. Jacobsen, and M. Pellegrini, "Relationship betweennucleosome positioning and DNA methylation," Nature, vol. 466, pp. 388–92, 2010.
[4] X. Wang, G. O. Bryant, M. Floer, D. Spagna, and M. Ptashne, "An effect of DNA sequence on nucleosome occupancy and removal,"Nat. Publishing Group, vol. 18, pp. 507–509, 2011.
[5] A. S. Shirkhorshidi, S. Aghabozorgi, T. Y. Wah, T. Herawan, "Big Data Clustering: A Review" Computational Science and Its Applications – ICCSA 2014Volume 8583 of the series Lecture Notes in Computer Science pp 707-720.

[6] M. L. Sogin, H. G. Morrison, J. A. Huber, D. Mark Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl, "Microbial diversity in the deep sea and the underexplored 'rare biosphere'", Proc. Nat. Acad. Sci. USA, vol. 103, no. 32, pp. 12115– 12120, 2006.
[7] S. M. Huse, D. M. Welch, H. G. Morrison, and M. L. Sogin. (2010).Ironing out the wrinkles in the rare biosphere through improved OTU clustering," Environmental Microbiol., vol. 12, no. 7, pp. 1889–1898.
[8] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D.Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I.Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E.Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J.Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann,T. Yatsunenko, J. Zaneveld, and R. Knight, "QIIME allowsanalysis of high-throughput community sequencing data," NatureMethods, vol. 7, no. 5, pp. 335–336, May 2010.
[9] R. C. Edgar. (2010). "Search and clustering orders of magnitudefaster than BLAST" Bioinformatics, vol. 26, no. 19, pp. 2460–2461.
[10] R. C. Edgar, "UPARSE: highly accurate OTU sequences frommicrobial amplicon reads," Nat. Methods, vol. 10, no. 10, pp. 996–8, Oct. 2013.
[11] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E.B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson,J. W. Sahl, B. Stres, G. G. Thallinger, D. J. V. Horn, and C. F.Weber,"Introducing mothur: Open-source platform-independentcommunity supported software for describing and comparingmicrobial communities", Appl. Envir. Microbiol., vol. 75,no. 23, pp. 7537–7541, 2009.
[12] Y. Sun, Y. Cai, L. Liu, F. Yu, M. L. Farrell, W. McKendree, and W.Farmerie,"ESPRIT: Estimating species richness using largecollections of 16S rRNA pyrosequences", Nucleic Acids Res., vol. 37,no. 10, p. e76, 2009.
[13] Y. Cai and Y. Sun., "ESPRIT-Tree: Hierarchical clusteringanalysis of millions of 16S rRNA pyrosequences in quasilinearcomputational time" Nucleic Acids Res., vol. 39, no. 14, p. e95, 2011.
[14] R. C. Edgar., "MUSCLE: Multiple sequence alignment withhigh accuracy and high throughput", Nucleic Acids Res., vol. 32, no. 5,pp. 1792–1797, 2004.
[15] Y. Sun, Y. Cai, S. M. Huse, et al., "A large-scale benchmarkstudy of existing algorithms for taxonomy-independent microbialcommunity analysis," Briefings in Bioinformatics, vol. 13, no. 1,pp. 107–121, 2011.
[16] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: Anew data clustering algorithm and its applications", Data MiningKnowl. Discovery, vol. 1, no. 2, pp. 141–182, 1997.
[17] W. Li and A. Godzik., "Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences", Bioinformatics, vol. 22, no. 13, pp. 1658–1659, 2006.
[18] Schloss PD, Handelsman J., "Introducing DOTUR, a computerprogram for defining operational taxonomic units and estimating speciesrichness"Appl Environ Microbiol 71:1501–1506. http://dx.doi.org/10.1128/AEM.71.3.1501, 2005.
[19] Albanese D, Fontana P, De Filippo C, Cavalieri D, Donati C., "Micca:a complete and accurate software for taxonomic profiling of metagenomicdata", Sci Rep 5:9743,http://dx.doi.org/10.1038/srep09743, 2015.
[20] Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M., "Swarm:robust and fast clustering method for amplicon-based studies",PeerJ2:e593,http://dx.doi.org/10.7717/peerj.593, 2014.
[21] Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M., "Swarmv2: highly-scalable and high-resolutionamplicon clustering",PeerJ3:e1420,http://dx.doi.org/10.7717/peerj.1420, 2015.
[22] Kopylova E, Noé L, Touzet H., "SortMeRNA: fast and accuratefiltering of ribosomal RNAs in metatranscriptomic data", Bioinformatics28:3211–3217. http://dx.doi.org/10.1093/bioinformatics/bts611, 2012.
[23] Hobohm U, Scharf M, Schneider R, Sander C., "Selection ofrepresentative protein data sets" Protein Sci 1, 409 – 417, http://dx.doi.org/10.1002/pro.5560010313, 1992.
[24] Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R., "UCHIMEimproves sensitivity and speed of chimera detection" BioInformatics27, 2194–2200, http://dx.doi.org/10.1093/bioinformatics/btr381, 2011.
[25] Legendre P, Legendre L., "Numerical ecology", 2nd ed, Developmentsin environmental modelling, vol 20, p . Elsevier Science, Amsterdam,The Netherlands, 1998.

