VOL9 NO.1 March 2020 - ISSN 2233 – 1859

Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online: http://scjournal.ius.edu.ba

IUS Soft Computing Research Group

Genomic Signal Processing Techniques for Taxonomy Prediction

O. Gürsoy

M. Can

Faculty of Engineering and Natural Sciences,

International University of Sarajevo International University of Sarajevo,

Hrasnicka Cesta 15, Ilidža 71210 Sarajevo,

Bosnia and Herzegovina

ogursoy@ius.edu.ba

mcan@ius.edu.ba

Article Info

Article history:

Article received on 10 January 2020

Received in revised form 1 February 2020

Keywords:

16S rRNA gene, LongestCommonSubsequence, Taxonomic clustering, Snowball

ABSTRACT: To analyze complex biodiversity in microbial communities, 16S rRNA marker gene sequences are often assigned to operational taxonomic units (OTUs). The abundance of methods that have been used to assign 16S rRNA marker gene sequences into OTUs brings discussions in which one is better. Suggestions on having clustering methods should be stable in which generated OTU assignments do not change as additional sequences are added to the dataset is contradicting some other researches contend that the methods should properly present the distances of sequences is more important. We add one more de novo clustering algorithm, Rolling Snowball to existing ones including the single linkage, complete linkage, average linkage, abundance-based greedy clustering, distance-based greedy clustering, and Swarm and the open and closed-reference methods. We use GreenGenes, RDP, and SILVA 16S rRNA gene databases to show the success of the method. The highest accuracy is obtained with SILVA library.

1. INTRODUCTION

Metagenomics is a recently-born and highly popular field that studies the genomic contents of microbial communities living in certain environments and tries to understand the structure and function of these microbial communities by sequencing genomic fragments from environmental samples without the need of cultivating them in a laboratory (Huttenhower et al., 2012; Qin et al., 2010). The microbiome is considered to be the "dark matter of the biological universe" as most of the microorganisms are very difficult to culture and are still unknown (Bernard, Pathmanathan, Lannes, Lopez, & Bapteste, 2018; Kellenberger, 2001; Lok, 2015). Reconstructing the taxonomic composition of a bacterial community has a critical role in understanding that such a community might play an important role in affecting change in that environment and in creating different types of medicinal drugs (Lok, 2015) and determining different types of

Ursell, Parfrey, & Knight, 2012; Oakley, Fiedler, Marrazzo, & Fredricks, 2008; Turnbaugh et al., 2009).

In the early metagenomic studies, the sequencing of a complete 16S rRNA gene was a common approach using the traditional Sanger sequencing methodology (Dethlefsen, Huse, Sogin, & Relman, 2008; Petrosino, Highlander, Luna, Gibbs, & Versalovic, 2009). Although this approach was informative, it is expensive and provides a limited depth of sequencing in discovering the complete bacterial diversity that exists in a complex environment.

With Next-Generation Sequencing (NGS), it has become easy to study the microbial world in their environments without culturing them (Scholz, Lo, & Chain, 2012; Shokralla, Spall, Gibson, & Hajibabaei, 2012). In recent years, rapid development in NGS has made it possible to directly sequence a huge amount of DNA/RNA fragments extracted from environmental samples such as human gut, marine or soil in a reasonable time (Eisen, 2011). It has made sequencing faster and highly economical providing a unique opportunity to study the microbial diversity of many complex environments at a much lower cost (Desai et al., 2013).

The two common sequencing approaches adopted in metagenomic projects are whole-genome shotgun sequencing and target metagenomics methods which are also called amplicon sequencing (Fuhrman, 2012). Shotgun sequencing provides more information to explore the microbial community both functionally and taxonomically. However, it is very expensive and computationally challenging and complex (Bhat, Prabhu, & Balakrishnan, 2019). The shotgun approach involves the sequencing of all genomic fragments and the targeted approach involves sequencing the marker gene such as 16S rRNA.

The 16S rRNA gene is the most commonly used genetic marker since it is conserved in all prokaryotes and consists of highly conservative and highly variable regions (HVRs) (Case et al., 2007; Janda & Abbott, 2007). Thus, NGS has shifted the studies towards sequencing short hypervariable regions of the 16S rRNA gene instead of sequencing the complete gene (Mizrahi-Man, Davenport, & Gilad, 2013).

16S rRNA gene is almost 1600 base pairs long and contains 9 hypervariable regions V1-V9 which are both variable as well as conserved [17]. This approach is highly applicable mostly because the lengths of different HVRs of the 16S rRNA gene are between 100–300 bp which can be easily obtained using short paired-end reads produced by commonly used NGS technologies (Aravindraja, Viszwapriya, & Karutha Pandian, 2013; J. Zhang, Kobert, Flouri, & Stamatakis, 2014).

1.1. OTUs Clustering (OTU Picking)

To simplify the complexity of large datasets generated by NGS technologies, sequences are clustered into meaningful bins. These bins are called operational taxonomic units (OTUs) which are used to study the biodiversity within and between different samples (Schloss & Westcott, 2011). OTUs form the basis for further analysis and comparative studies (Di Bella, Bao, Gloor, Burton, & Reid, 2013). These studies helped researchers to profile the microbiota associated with the human body (Huttenhower et al., 2012), soil (Shade et al., 2013), oceans (Gilbert et al., 2012).

OTUs help us to reduce and eliminate the PCR sequencing errors, merge paralogs and variation between strains of a single species (Robert C Edgar, 2017). Clustering also reduces the run time of subsequent analysis steps. However, poorly clustered OTUs can have a significant impact on downstream analyses.

In general clustering methods can be classified into taxonomy-dependent approaches, where sequences are clustered against a reference database (closed reference or phylotyping), and taxonomy-independent approaches (Mande, Mohammed, & Ghosh, 2012), where sequences are clustered into OTUs based on pairwise similarities without using external reference databases (de novo clustering (Di Bella et al., 2013)). A third approach which is called open reference clustering where sequences are assigned to OTUs using closed-reference clustering and sequences that do not hit the database are then clustered with de novo clustering. (Rideout et al., 2014). There are popular reference databases: Ribosomal Database Project (RDP) (Cole et al., 2009), Greengenes (DeSantis et al., 2006), SILVA (Pruesse et al., 2007), NCBI (Federhen, 2012), Open Tree of Life Taxonomy (OTT) (Hinchliff et al., 2014), and UNITE (Kõljalg et al., 2013).

1.1.1. Closed Reference Approach

This type of clustering is also referred to as phylotyping (Schloss & Westcott, 2011) or closed-reference clustering (Navas-Molina et al., 2013). This approach compares sequence reads to a reference database and then cluster them into the same OTU that is similar to the same reference read. These type of clustering methods can suffer when the reference databases do not sufficiently represent the biodiversity. When a large number of sequences are novel, then they cannot be assigned to an OTU as well. Furthermore, a sequence that represents a piece of the gene may be more than 97% similar to multiple reference sequences.

Defining OTUs in this clustering approach can be problematic because two sequences could be similar to the same reference sequence at a certain level but not similar to each other at the same level. Otherway around, a sequence may be equally similar to two or more reference sequence reads. In order to overcome these obstacles, a classifier could be used to assign taxonomy to each sequence read so that they can be clustered at a certain level.

The advantages of the closed-reference clustering methods are that they are fast, highly parallelizable and resulting

OTU assignments can be comparable between different studies (Westcott & Schloss, 2015).

1.1.2. De novo Approach

De novo clustering (Navas-Molina et al., 2013) which is also referred to as distance-based (Schloss & Westcott, 2011) clustering, the distance between sequences is used to bin sequences into OTUs rather than using a reference database to calculate distances. The computational cost of this type of clustering method scales quadratically with the number of unique sequence reads. Sequencing errors increases the number of unique sequences requiring large amounts of memory and time for clustering. Input order of the sequences in De novo clustering for OTU assignments are highly sensitive (He et al., 2015; Mahé, Rognes, Quince, de Vargas, & Dunthorn, 2014).

The power of de novo clustering is its independence of reference databases for clustering. Hence it has been preferred in most studies where novel sequence reads are expected.

1.1.3. Open Reference Approach

Open-reference clustering is a combination of the closed-reference clustering and de novo clustering (Navas-Molina et al., 2013; Rideout et al., 2014). This type of clustering performs closed-reference clustering first and then continues with the de novo clustering for those sequences whose similars are not found in the reference database. Theoretically, this approach has the potential power of both closed-reference and de novo clustering but the different OTU definitions used by these clustering approaches have a possible problem when the approaches are combined. An alternative way to this method is to classify sequences to a bacterial family or genus level and then bin those sequences into OTUs within the chosen taxonomic groups using the average linkage method (Schloss & Westcott, 2011).

An advantage of the open reference clustering is that it is highly parallelizable since each taxonomic group can be processed separately. However, it is still subject to the problems associated with reference database quality and the classification error.

1.2. Clustering Methods

Regardless of the reference database used, in general, there are two types of clustering in use: greedy heuristics-based clustering and hierarchical clustering. A few methods also use the model-based clustering.

1.2.1. Greedy Heuristic Clustering

Greedy heuristic-based methods select a sequence read as a seed and apply either de novo approach or closed reference approach using a particular threshold value which is generally 97%. The reads which do not match the selected seed are treated as a new seed. The main algorithms which apply greedy heuristic approach are CDHIT (Weizhong & Adam, 2006), USEARCH (Robert C. Edgar, 2010), UCLUST (Robert C. Edgar, 2010), VSEARCH (Rognes, Flouri, Nichols, Quince, & Mahé, 2016), SUMACLUST (Mercier, Boyer, Bonin, & Coissac, 2013), OTUCLUST (Albanese, Fontana, De Filippo, Cavalieri, & Donati, 2015), GramCluster (Russell, Way, Benson, & Sayood, 2010), and DNACLUST (Ghodsi, Liu, & Pop, 2011).

CDHIT sorts all the sequence reads before selecting any seed and picks the longest one as the initial seed and then clusters the sequence reads which are similar to the selected one at some threshold. VSEARCH applies optimal global aligner in parallel with multiple threading to perform alignments at a high speed. UCLUST works similar to CDHIT but it does not select the longest read as the initial seed. It has both de novo and closed reference clustering approaches. It takes experimental sequence reads as input for cluster centroids (de novo). A reference database of 16S rRNA sequences is used to generate the cluster centroids for assigning experimental sequencing reads for the referenced centroids (closed reference). USEARCH extends aligner for the alignment search with a heuristic seed. It calculates kmer based heuristic distances for generating the cluster centroids. VSEARCH is an open-sourced version of the commercial USEARCH. It uses Needleman-Wunsch (Needleman & Wunsch, 1970) dynamic programming to global alignment distances. SUMACLUST and OTUCLUST utilize exact sequence alignment and clusters are formed incrementally by checking an abundance-ordered list of input reads against the already selected representative set sequences. Kmer based searching is used to measure an identity distance that is calculated by the length of the Longest Common Subsequence divided by the shortest alignment. GramCluster uses a grammar-based distance metric where new sequences are compared with cluster-representative sequences to determine membership. DNACLUST uses a novel k-mer filtering approach without a pairwise alignment process. Most of these greedy methods perform in O (n) time complexity but cluster quality is not always as good as hierarchical methods.

1.2.2. Hierarchical Clustering

Hierarchical methods use a genetic distance matrix which is calculated by pairwise comparison of all reads in an agglomerative way. Most of these algorithms have O (n2) time complexity which is a bottleneck for processing the big data. The main algorithms using hierarchical methods are MOTHUR (Schloss et al., 2009) and ESPRIT (Sun et al., 2009). MOTHUR has the option to calculate the hierarchical distance with the nearest neighbor, average neighbor, and furthest neighbor. It uses the multiple sequence alignment tool MUSCLE (R. C. Edgar, 2004) to calculate the pairwise distances. ESPRIT uses pairwise global alignment (Bhat et al., 2019).

1.2.3. Model-based Clustering

Model-based clustering approaches employ probabilistic methods like the Gaussian mixture model and machine learning techniques. SWARM (Mahé et al., 2014) clusters identical reads iteratively and uses the abundance and internal structure of each cluster to optimize the results. CROP (Hao, Jiang, & Chen, 2011) uses the unsupervised Bayesian clustering method without any threshold (Bhat et al., 2019).

There are also OTU quality metrics that measure the accuracy of clustered OTUs including richness (Sun et al., 2009), normalized mutual information (Cai & Sun, 2011; Zheng, Kramer, & Schmidt, 2012) and Matthews' Correlation Coefficient which measures the correlation between predicted and known values as accuracy value (Schloss & Westcott, 2011). Normalized mutual information is an information theory measuring the mutual dependence of two frequency distributions (Cover & Thomas, 1991).

Once OTUs are constructed, further analysis and taxonomic annotation are done. It is also possible to skip the clustering step and to use a reference database to identify each sequence, binning together those that have the same taxonomy (Aho et al., 2015) based on the number of insufficient sample size.

1.3. Denoising

Due to its unique structure containing conserved and variable regions and its presence in all prokaryotes, the 16S rRNA gene is used as a marker gene. This approach is often preferred over shotgun sequencing due to the high cost. However, 16S rRNA gene sequencing errors also complicate distinguishing real nucleotide differences from the artifacts. To overcome this problem, sequence reads are often clustered into OTUs at 97% identity threshold but then again it has a big effect on taxonomic resolution. In order to increase the taxonomic resolution, some new sequence denoising pipelines have been introduced to correct sequencing errors and improve the taxonomic resolution.

Denoising approaches can improve the taxonomic resolution and free us from choosing one of various OTU strategies which may give different results (Robert C. Edgar, 2017). Furthermore, amplicon sequence variants can be identified by their unique sequences which allows comparison of different studies with different datasets (Callahan, McMurdie, & Holmes, 2017).

There are already several bioinformatic comparisons of OTU-based methods (Allali et al., 2017). A comprehensive comparison of the above-mentioned denoising methods along with an open-reference 97% OTU-based approach (Rognes et al., 2016) shows that even though all denoising methods give similar community structure, the number of ASVs/OTUs and resulting alpha-diversity metrics varies in the mock community analyses and it is recommended to be considered when attempting to identify rare organisms from possible noise (Nearing, Douglas, Comeau, & Langille, 2018).

1.4. Taxonomy Prediction

One of the fundamental tasks in microbiology is the prediction of taxonomy for marker gene sequences where a reference database is used with taxonomy annotations. Sequence reads can be studied as bins of similar sequences (operational taxonomic units: OTUs), or as raw reads. In any case, the taxonomic prediction of these sequence reads characterizes the microbiota composition.

There are two approaches for carrying out a taxonomic prediction: homology-based (alignment) and prediction-based (k-mer) approach (Chaudhary, Sharma, Agarwal, Gupta, & Sharma, 2015). The homology-based approach requires the alignment of a query sequence with all available sequences within the reference database used for prediction. Sequences are identified by similarities and differences requiring the comparison of each nucleotide residue. Hence, the quality of the reference database used in the taxonomy prediction is also important (Gupta, Kapil, Dhakan, & Sharma, 2014).

Some popular tools designed for taxonomy prediction are given in Section 4.

2. PROBLEM STATEMENT

The main problem with existing methods in taxonomy prediction, OTU clustering, and denoising is the tradeoff between computational time and accuracy. The length of short reads has a huge impact on this challenge. Furthermore, the best performing tools often may not be open-sourced and free (Nearing et al., 2018).

NGS technologies provide short reads and huge sequencing depth at a much lower cost. Hence, recent metagenomic projects shift to focus on the sequencing of only a single or combination of two or more hypervariable regions. Therefore, specialized tools are needed for highly accurate taxonomic classification of species using these short length sequences.

The amount of genetic data produced by NGS technologies is growing from tens of thousands to several million reads, faster than the rate at which it can be analyzed (Caporaso et al., 2010a). The latest Illumina HiSeq 2500 platform can produce approximately 600 million sequences of 300bp in around 40 hours. The rapid accumulation of these genomic information provides a valuable source for biological knowledge. However, it introduces a serious challenge for data analysis (Cai et al., 2017). Computational methods for analyzing these large collections of sequences are limited.

The 16S rRNA gene has limitations in specificity such that two different species may have identical marker genes but it is still highly sensitive and one single nucleotide difference

can detect important genomic variation (Thompson et al., 2005; Ward, Ferris, Nold, & Bateson, 1998).

Taxonomy prediction and OTUs clustering both have some challenges. Taxonomy predictions with the existing tools suffer from the short-read sequences (Claesson et al., 2010; Wang, Garrity, Tiedje, & Cole, 2007a). Furthermore, when novel taxa that are not present in the reference database used, the taxonomy prediction tool should identify it with the closest taxonomic lineage and should not go further (Bokulich et al., 2018). Analyses that use similarity comparison to taxonomic reference databases may also provide poorly resolved results especially for samples that have high diversity (Eren et al., 2013).

On the other hand, one of the main problems with denoising ASVs methods is the discrimination between PCR sequencing errors and biological variation. Due to divergence in rRNA operons, it is possible that the same bacterial genome may have 16S rRNA genes that are different by more than 40 base pairs which could lead to multiple ASVs and diversity which can complicate the downstream analysis and to identify specific taxa (Fierer, Brewer, & Choudoir, 2017).

Current popular tools for denoising are DADA2, Deblur, and UNOISE3. DADA2 and Deblur are open-sourced and free. UNOISE3 is closed-source and offers a free 32bit academic version which is limited by supporting only up to 4GB of available memory. The Run time of these tools on the same dataset is significantly different and the fastest one is UNOISE3 (Nearing et al., 2018).

3.BACKGROUND

Many algorithms have been developed for taxonomy prediction such as RDP Naive Bayesian Classifier (NBC) (Wang, Garrity, Tiedje, & Cole, 2007b), GAST (Huse et al., 2008), 16Sclassifier (Chaudhary et al., 2015), SPINGO (Allard, Ryan, Jeffery, & Claesson, 2015), Metaxa2 (Bengtsson-Palme et al., 2015), SINTAX (R. Edgar, 2016), PROTAX (Somervuo, Koskela, Pennanen, Henrik Nilsson, & Ovaskainen, 2016), microclass (Liland, Vinje, & Snipen, 2017). There are also implemented methods in MOTHUR, QIIME v1 (Caporaso et al., 2010b) and QIIME v2 (Bolyen et al., 2019).

RDP-Classifier which uses a Naive Bayesian Classifier (Claesson et al., 2010; Wang et al., 2007a) is one of the most commonly used tools. It is highly accurate on complete 16S rRNA sequences but suffers in accuracy for targeted HVRs which are short in length (Vilo & Dong, 2012).

In order to validate taxonomy prediction methods, some benchmark techniques are proposed. The leave-none-out technique uses both test set and training set from a complete reference database (Werner et al., 2012). The leave-one-out technique uses each sequence from the reference database as for query while the remaining sequences are used as a training set (Deshpande et al., 2016; R. Edgar, 2016; Wang et al., 2007a). Leave-clade-out is also a cross-validation technique (Brady & Salzberg, 2009) where test and training sets are selected in a way that each taxonomy at a given rank is included in the test or training set but not both. In the k-fold cross-validation technique, the reference database is randomly split into test sets and training sets of relative sizes. This technique provides performance evaluation on novel query sequences (Lan, Wang, Cole, & Rosen, 2012). There are also mock communities (artificial) which contain known strains have also been used for validation (Allard et al., 2015; Bokulich, 2017; Robert C. Edgar, 2017). Cross-validation by identity technique creates a model using the dissimilarity between the query and reference sequences (Robert C. Edgar, 2018).

Many OTU clustering methods exist (Robert C. Edgar, 2013; Rideout et al., 2014; Schloss & Handelsman, 2005; Schloss et al., 2009; Seguritan & Rohwer, 2001; Ye, 2010) most of which apply a threshold of 97% sequence similarity following the general wisdom that 97% corresponds approximately to species (Schloss & Handelsman, 2005; Seguritan & Rohwer, 2001; Westcott & Schloss, 2017). This threshold was proposed in 1994 (STACKEBRANDT & GOEBEL, 1994) when only a few 16S rRNA sequences were available.

There are discussions on replacing OTUs with ASVs in the marker-gene analysis (Brandt et al., 2019; Callahan et al., 2017). ASV methods have shown sensitivity and specificity similar or better than OTU methods and they are also better in distinguishing the patterns (Callahan et al., 2016; Eren et al., 2013, 2015; Needham, Sachdeva, & Fuhrman, 2017). Amplicon sequence variants (ASVs) are obtained by a de novo like approach where sequences are distinguished from errors with the assumption indicating that the sequences are more likely to be repeatedly observed than those with an error. Hence, obtaining ASV cannot be performed independently and the smallest unit of data in which ASVs can be obtained must be a sample. Unlike OTUs, ASVs are consistent labels and represent a biological reality that exists outside of the analyzed data. Therefore ASVs that are obtained independently from different samples or different studies are comparable (Callahan et al., 2017).

One of the leading popular denoising tools DADA2 generates an error model that is trained on the sequence reads and then uses that model to correct errors. Sequences are then collapsed into ASVs (Callahan et al., 2016). ASVs are also called sub-OTUs, or zero-radius OTUs. Deblur uses error profiles to obtain error-free sequences and can perform on Illumina MiSeq and HiSeq sequencing data (Amir et al., 2017). Sequences are aligned together into sub-OTUs and predicted error-derived reads are removed from neighboring sequences. It calculates the pairwise Hamming distances in each sample separately which is efficient for both memory and computational power. UNOISE3 (Robert C Edgar, 2016) uses a one-pass clustering method that does not require quality scores. A cluster is formed with a centroid sequence which has a higher abundance and similar member sequences with lower abundances. Two parameters with pre-set values are curated to generate zero-radius

OTUs. The one-pass clustering method has an advantage in computational time.

In genomic signal processing, representation of a DNA sequence in a discrete numerical sequence is essential for digital signal processing based analysis (Anastassiou, 2001; Berger, Mitra, Carli, & Neri, 2002; Cheever, Searls, Karunaratne, & Overton, n.d.). This representation is often called a mapping scheme. Each mapping scheme has a different set of discrete numeric representation of nucleotides.

Many mapping scheme have been developed such as Voss representation (Voss, 1992), the real number (Chakravarthy, Spanias, Iasemidis, & Tsakalis, 2004), the integer number (Paul Dan Cristea, 2002), tetrahedron (Silverman & Linsker, 1986), the complex number (Paul D. Cristea, 2002), the quaternion, the paired numeric (Akhtar, Epps, & Ambikairajah, 2007), the electron-ion interaction potentials (EIIP) (Nair & Sreenadhan, 2006), the atomic number (Holden et al., 2007), Z-curve (R. Zhang & Zhang, 1994), the DNA walk (Berger, Mitra, Carli, & Neri, 2004). Table 1. shows a comprehensive list of the existing mapping schemes (Ning Yu, Zhihua Li, 2018).

Table 1 Some Encoding Schemes (Ning Yu, Zhihua Li, 2018)

Scheme Name	Discrete numeric values
Atomic Number	C=58, T=66,A=70, G=78
EIIP	C=0.1340, T=0.1335, A=0.1260,G=0.0806
Molecular Mass	C=111.1, T=112.1, A=135.13,G=151.13 or C=110, T=125, A=134, G=150
Thermodynamics	TC=5.6, GA=5.6, CA=5.8, TG=5.8, TA=6.0, AC=6.5, GT=6.5, CT=7.8,AG=7.8, AT=8.6, TT=9.1, AA=9.1, CC=11.0, GG=11.0, GC=11.1, CG=11.9
Three-group	(1) R={A, G}, Y={C, T} , (2) M={A, C}, K={G, T}, (3) W={A, T}, S={G, C}
Dinucleotide	Sixteen dinucleotides are mapped to a unit circle.
Ring Structure	AG: (0, 1.5), CT: (0, -1.5), CA:(1, 1), TG: (-1, -1), CG: (1, -1), TA: (-1, 1), GA: (1,0), GT(0.5, -1.25), GC: (-0.5, -1.25), TC:(-1, 0), AC: (-0.5, 1.25), AT: (0.5, 1.25),AA: (0, 1), TT: (0.5, 0), GG: (0, -1), CC:(-0.5, 0).

4. MATERIALS AND METHODS

Existing 16S rRNA Reference databases Greengenes, SILVA, and RDP are used.

Sequences are converted to genomic signals with complex numbers encoding scheme [i,-i,1,-1] and randomly selected 50 taxa each having 50 sequences from the genus level are used to compute in-class and inter-class similarities. The average similarities are 69.94% and 25.37% respectively. The difference between in-class / inter-class similarities is very promising, and such a similarity measure results in good taxonomy prediction accuracy and specificity in OTUs clustering.

Preliminary results for the SILVA database at the genus level, show distinguishable in-class, inter-class similarities

4.1. Data

Gene databases independently get updated and have a different approach to taxonomy construction. Taxonomy is ranked as kingdom/domain, phylum, class, order, family, genus, and species levels. RDP has no species level and has additional subclass and suborder levels.

Taxonomy predictions can be based on manual and computational analyses after multiple alignments (McDonald et al., 2012; Yilmaz et al., 2014). Greengenes database contains Archae and Bacteria. It uses rank mapping mainly from NCBI and other sources and De novo tree construction to make classifications (Balvočiute & Huson, 2017). Since the last release in 2013, Greengenes is not updated.

SILVA database taxonomic ranking assignments are manually curated [113]. SILVA and RDP reference databases contain Archae, Bacteria, and Eukarya (Fungi). RDP database contains sequences from the International Nucleotide Sequence Database Collaboration (Cochrane, Karsch-Mizrachi, & Takagi, 2016).

SILVA, UNITE, and Greengenes have environmental sequences. OTT contains a synthesis of phylogenetic trees that are ranked and merge together. In Greengenes, RDP and SILVA there is no attempt to classify unnamed groups (R. Edgar, 2016).

Besides the above-mentioned taxonomy reference databases, this research will also make use of the datasets including mock community datasets provided by the early leading studies in taxonomy prediction, OTUs clustering and denoising for benchmarking and comparison purposes.

4.2. Signal Similarity

Current reference databases and available data sets provided by similar studies will be converted to genomic signals through available encoding schemes. Conversion of DNA sequences into the genomic signals offers the possibility to apply various types of signal processing methods that can identify hidden features. Genomic signal processing methods can provide different types of similarities that can be used in taxonomy prediction, OTUs clustering, and denoising.

Each nucleotide base is converted to a number according to one of the encoding schemes in Table 1. Using for instance the encoding scheme [i,-i,1,-1] the sequences "AATACGCG" and "CAG" are converted to two signals : [i,i,-i,i,-1,1,-1,1] and, [-1,i,1]. Then, a new vector is generated by cross-correlation. If the real part of the new

vector has a positive peak then it is considered as a similarity between these two sequences and a negative peak is considered as a complementary similarity between them (Rockwood, Crockett, Oliphant, & Elenitoba-Johnson, 2005)(Paul Dan Cristea, 2002).

Cross Correlation in Statistics

In statistics, a cross-correlation function is a measure of association. For example, the most common correlation coefficient, the Pearson product-moment correlation coefficient (PPMC), is a normalized version of a cross-correlation.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2]\,[n(\sum y^2) - (\sum y)^2]}}$$

The PPMC gives a measure of temporal similarity for two time series.

Cross Correlation in Signal Processing

In signal processing, cross correlation is where you take two signals and produce a third signal. The method, which is basically a generalized form of "regular" linear correlation, is a way to objectively compare different time series and allows you to see how two signals match and where the best match occurs. It can be used to create plots that may reveal hidden sequences.

The basic process involves:
1) Calculate a correlation coefficient. The coefficient is a measure of how well one series predicts the other.
2) Shift the series, creating a lag. Repeat the calculations for the correlation coefficient.
3) Repeat steps 1 and 2.
How many times you repeat the process will depend on your data, but as the lag increases the potential matches will decrease.
4) Identify the lag with the highest correlation coefficient. The lag with the highest correlation coefficient is where the two series match the best.

4.2.1. In-Class / Inter-Class Similarities Using Signals

Preliminary results for the SILVA database at the genus level, show distinguishable in-class, inter-class similarities.

Sequences are converted to genomic signals with complex numbers encoding scheme [i,-i,1,-1] and randomly selected 50 taxa each having 50 sequences from the genus level are used to compute in-class and inter-class similarities.

The average in class similarities and interclass averages are computed for all taxon levels in the three databases Greengenes, SILVA, and RDP. The results are shown in Table 2.

The difference between in-class / inter-class similarities is very promising, and such a similarity measure results in good taxonomy prediction accuracy and specificity in OTUs clustering.

Table 2 In-class / Inter-class similarities for all taxon levels using Signals

	Databases	In-Class	Inter-Class
Phylum	Greengenes	39.79%	19.45%
	SILVA	44.85%	21.68%
	RDP	55.39%	25.79%
	Mean	46.68%	22.31%
Class	Greengenes	44.09%	18.85%
	SILVA	50.01%	22.01%
	RDP	55.97%	28.17%
	Mean	50.02%	23.01%
Order	Greengenes	49.87%	20.52%
	SILVA	51.10%	21.95%
	RDP	60.97%	30.34%
	Mean	53.98%	24.27%
Family	Greengenes	55.11%	23.17%
	SILVA	58.14%	23.22%
	RDP	63.05%	38.55%
	Mean	58.77%	28.31%
Genus	Greengenes	50.16%	21.16%
	SILVA	69.94%	25.37%
	RDP	67.68%	44.98%
	Mean	62.59%	30.50%
Species	Greengenes	59.65%	22.35%
	SILVA	73.41%	28.74%
	Mean	66.53%	25.55%
	Overall	55.83%	25.66%

REFERENCES

Aho, V. T. E., Pereira, P. A. B., Haahtela, T., Pawankar, R., Auvinen, P., & Koskinen, K. (2015). The microbiome of the human lower airways: A next generation sequencing perspective. *World Allergy Organization Journal*, *8*(1). https://doi.org/10.1186/s40413-015-0074-z

Akhtar, M., Epps, J., & Ambikairajah, E. (2007). On DNA Numerical Representations for Period-3 Based Exon Prediction. In *2007 IEEE International Workshop on Genomic Signal Processing and Statistics* (pp. 1–4). IEEE. https://doi.org/10.1109/GENSIPS.2007.4365821

Albanese, D., Fontana, P., De Filippo, C., Cavalieri, D., & Donati, C. (2015). MICCA: A complete and accurate software for taxonomic profiling of metagenomic data. *Scientific Reports*, *5*. https://doi.org/10.1038/srep09743

Allali, I., Arnold, J. W., Roach, J., Cadenas, M. B., Butz, N., Hassan, H. M., … Azcarate-Peril, M. A. (2017). A comparison of sequencing platforms and bioinformatics pipelines for compositional analysis of the gut microbiome. *BMC Microbiology*, *17*(1). https://doi.org/10.1186/s12866-017-1101-8

Allard, G., Ryan, F. J., Jeffery, I. B., & Claesson, M. J. (2015). SPINGO: A rapid species-classifier for microbial amplicon sequences. *BMC Bioinformatics*, *16*(1). https://doi.org/10.1186/s12859-015-0747-1

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., … Knight, R. (2017). Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *MSystems*, *2*(2). https://doi.org/10.1128/msystems.00191-16

Anastassiou, D. (2001). Genomic signal processing. *IEEE Signal Processing Magazine*, *18*(4), 8–20. https://doi.org/10.1109/79.939833

Aravindraja, C., Viszwapriya, D., & Karutha Pandian, S. (2013). Ultradeep 16S rRNA Sequencing Analysis of Geographically Similar but Diverse Unexplored Marine Samples Reveal Varied Bacterial Community Composition. *PLoS ONE*, *8*(10). https://doi.org/10.1371/journal.pone.0076724

Balvočiute, M., & Huson, D. H. (2017). SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare? *BMC Genomics*, *18*(S2), 114. https://doi.org/10.1186/s12864-017-3501-4

Bengtsson-Palme, J., Hartmann, M., Eriksson, K. M., Pal, C., Thorell, K., Larsson, D. G. J., & Nilsson, R. H. (2015). metaxa2: Improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Molecular Ecology Resources*, *15*(6), 1403–1414. https://doi.org/10.1111/1755-0998.12399

Berger, J. A., Mitra, S. K., Carli, M., & Neri, A. (2002). New Approaches To Genome Sequence Analysis Based on Digital Signal Processing. *Genomic Signal Processing and Statics GENSiPS IEEE International Workshop*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.12.9353&rep=rep1&type=pdf

Berger, J. A., Mitra, S. K., Carli, M., & Neri, A. (2004). Visualization and analysis of DNA sequences using DNA walks. *Journal of the Franklin Institute*, *341*(1–2), 37–53. https://doi.org/10.1016/j.jfranklin.2003.12.002

Bernard, G., Pathmanathan, J. S., Lannes, R., Lopez, P., & Bapteste, E. (2018). Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery. *Genome Biology and Evolution*, *10*(3), 707–715. https://doi.org/10.1093/gbe/evy031

Bhat, A. H., Prabhu, P., & Balakrishnan, K. (2019). A critical analysis of state-of-the-art metagenomics OTU clustering algorithms. *Journal of Biosciences*, *44*(6). https://doi.org/10.1007/s12038-019-9964-5

Bokulich, N. A. (2017). Optimizing taxonomic classification of marker gene sequences. *PeerJ PrePrints*. https://doi.org/10.7287/peerj.preprints.3208v1

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., … Gregory Caporaso, J. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*, *6*(1), 90. https://doi.org/10.1186/s40168-018-0470-z

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., … Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, *37*(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

Brady, A., & Salzberg, S. L. (2009). Phymm and PhymmBL: Metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods*, *6*(9), 673–676. https://doi.org/10.1038/nmeth.1358

Brandt, M. I., Trouche, B., Quintric, L., Wincker, P., Poulain, J., & Arnaud-Haond, S. (2019). A flexible pipeline combining bioinformatic correction tools for prokaryotic and eukaryotic metabarcoding. *BioRxiv [Preprint]*. https://doi.org/10.1101/717355

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cai, Y., & Sun, Y. (2011). ESPRIT-Tree: Hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. *Nucleic Acids Research*, *39*(14). https://doi.org/10.1093/nar/gkr349

Cai, Y., Zheng, W., Yao, J., Yang, Y., Mai, V., Mao, Q., & Sun, Y. (2017). ESPRIT-Forest: Parallel clustering of massive amplicon sequence data in subquadratic time. *PLoS Computational Biology*, *13*(4), 1–16. https://doi.org/10.1371/journal.pcbi.1005518

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME Journal*, *11*(12), 2639–2643. https://doi.org/10.1038/ismej.2017.119

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Can, M. (2019). Annotation of Bacteria by Greengenes Classifier Using 16S rRNA Gene Hyper Variable Regions. *Southeast Europe Journal of Soft Computing*, *8*(2). https://doi.org/10.21533/scjournal.v8i2.181

Can, M., & Gürsoy, O. (2019). Taxonomic Classification of Bacteria Using Common Substrings. *Southeast Europe Journal of Soft Computing*, *8*(1). https://doi.org/10.21533/scjournal.v8i1.167

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., … Knight, R. (2010a). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336. https://doi.org/10.1038/nmeth.f.303

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., … Knight, R. (2010b, May). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*. https://doi.org/10.1038/nmeth.f.303

Case, R. J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W. F., & Kjelleberg, S. (2007). Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Applied and Environmental Microbiology*, *73*(1), 278–288. https://doi.org/10.1128/AEM.01177-06

Chakravarthy, N., Spanias, A., Iasemidis, L. D., & Tsakalis, K. (2004). Autoregressive Modeling and Feature Analysis of DNA Sequences. *EURASIP Journal on Advances in Signal Processing*, *2004*(1), 952689. https://doi.org/10.1155/S111086570430925X

Chaudhary, N., Sharma, A. K., Agarwal, P., Gupta, A., & Sharma, V. K. (2015). 16S classifier: A tool for fast and accurate taxonomic classification of 16S rRNA hypervariable regions in metagenomic datasets. *PLoS ONE*, *10*(2), 1–13. https://doi.org/10.1371/journal.pone.0116106

Cheever, E. A., Searls, D. B., Karunaratne, W., & Overton, G. C. (n.d.). Using signal processing techniques for DNA sequence comparison. In *Proceedings of the Fifteenth Annual Northeast Bioengineering Conference* (pp. 173–174). IEEE. https://doi.org/10.1109/NEBC.1989.36756

Claesson, M. J., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, J. R., Ross, R. P., & O'Toole, P. W. (2010). Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Research*, *38*(22). https://doi.org/10.1093/nar/gkq873

Clemente, J. C., Ursell, L. K., Parfrey, L. W., & Knight, R. (2012). The impact of the gut microbiota on human health: An integrative view. *Cell*, *148*(6), 1258–1270. https://doi.org/10.1016/j.cell.2012.01.035

Cochrane, G., Karsch-Mizrachi, I., & Takagi, T. (2016). The international nucleotide sequence database collaboration. *Nucleic Acids Research*, *44*(D1), D48–D50. https://doi.org/10.1093/nar/gkv1323

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., … Tiedje, J. M. (2009). The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, *37*(SUPPL. 1). https://doi.org/10.1093/nar/gkn879

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory*. *Elements of Information Theory*. New York, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/0471200611

Cristea, Paul D. (2002). Genetic signal representation and analysis. In M. D. Kessler & G. J. Mueller (Eds.), *Proc. SPIE COnf., Int. Biomedical Optics Symp. (BIOS02)* (pp. 77–84). https://doi.org/10.1117/12.491244

Cristea, Paul Dan. (2002). Conversion of nucleotides sequences into genomic signals. *Journal of Cellular and Molecular Medicine*, *6*(2), 279–303. https://doi.org/10.1111/j.1582-4934.2002.tb00196.x

Desai, A., Marwah, V. S., Yadav, A., Jha, V., Dhaygude, K., Bangar, U., … Jere, A. (2013). Identification of Optimum Sequencing Depth Especially for De Novo Genome Assembly of Small Genomes Using Next Generation Sequencing Data. *PLoS ONE*, *8*(4). https://doi.org/10.1371/journal.pone.0060204

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., … Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, *72*(7), 5069–5072. https://doi.org/10.1128/AEM.03006-05

Deshpande, V., Wang, Q., Greenfield, P., Charleston, M., Porras-Alfaro, A., Kuske, C. R., … Tran-Dinh, N. (2016). Fungal identification using a Bayesian classifier and the Warcup training set of internal transcribed spacer sequences. *Mycologia*, *108*(1), 1–5. https://doi.org/10.3852/14-293

Dethlefsen, L., Huse, S., Sogin, M. L., & Relman, D. A. (2008). The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16s rRNA sequencing. *PLoS Biology*, *6*(11), 2383–2400. https://doi.org/10.1371/journal.pbio.0060280

Di Bella, J. M., Bao, Y., Gloor, G. B., Burton, J. P., & Reid, G. (2013). High throughput sequencing methods and analysis for microbiome research. *Journal of Microbiological Methods*, *95*(3), 401–414. https://doi.org/10.1016/j.mimet.2013.08.011

Edgar, R. (2016). SINTAX: a simple non-Bayesian

taxonomy classifier for 16S and ITS sequences. *BioRxiv*, 074161. https://doi.org/10.1101/074161

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Edgar, Robert C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Edgar, Robert C. (2013). UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, *10*(10), 996–998. https://doi.org/10.1038/nmeth.2604

Edgar, Robert C. (2017). Accuracy of microbial community diversity estimated by closed- and open-reference OTUs. *PeerJ*, *2017*(10). https://doi.org/10.7717/peerj.3889

Edgar, Robert C. (2018). Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ*, *2018*(4), 1–29. https://doi.org/10.7717/peerj.4652

Edgar, Robert C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *BioRxiv*, 081257. https://doi.org/10.1101/081257

Edgar, Robert C. (2017). SEARCH_16S: A new algorithm for identifying 16S ribosomal RNA genes in contigs and chromosomes. *BioRxiv*, 124131. https://doi.org/10.1101/124131

Eisen, J. A. (2011). Environmental Shotgun Sequencing: Its Potential and Challenges for Studying the Hidden World of Microbes. *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches*, 157–162. https://doi.org/10.1002/9781118010518.ch20

Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., & Sogin, M. L. (2013). Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods in Ecology and Evolution*, *4*(12), 1111–1119. https://doi.org/10.1111/2041-210X.12114

Eren, A. M., Morrison, H. G., Lescault, P. J., Reveillaud, J., Vineis, J. H., & Sogin, M. L. (2015). Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME Journal*, *9*, 968–979. https://doi.org/10.1038/ismej.2014.195

Federhen, S. (2012). The NCBI Taxonomy database. *Nucleic Acids Research*, *40*(D1), D136–D143. https://doi.org/10.1093/nar/gkr1178

Fierer, B. N., Brewer, T., & Choudoir, M. (2017). Fierer Lab Lumping versus splitting – is it time for microbial ecologists to abandon OTUs ?, 5–7.

Fuhrman, J. A. (2012). Metagenomics and its connection to microbial community organization. *F1000 Biology Reports*, *4*(1). https://doi.org/10.3410/B4-15

Ghodsi, M., Liu, B., & Pop, M. (2011). DNACLUST: Accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinformatics*, *12*. https://doi.org/10.1186/1471-2105-12-271

Gilbert, J. A., Steele, J. A., Caporaso, J. G., Steinbrück, L., Reeder, J., Temperton, B., … Field, D. (2012). Defining seasonal marine microbial community dynamics. *ISME Journal*, *6*(2), 298–308. https://doi.org/10.1038/ismej.2011.107

Gupta, A., Kapil, R., Dhakan, D. B., & Sharma, V. K. (2014). MP3: A software tool for the prediction of pathogenic proteins in genomic and metagenomic data. *PLoS ONE*, *9*(4). https://doi.org/10.1371/journal.pone.0093907

Hao, X., Jiang, R., & Chen, T. (2011). Clustering 16S rRNA for OTU prediction: A method of unsupervised Bayesian clustering. *Bioinformatics*, *27*(5), 611–618. https://doi.org/10.1093/bioinformatics/btq725

He, Y., Caporaso, J. G., Jiang, X.-T., Sheng, H.-F., Huse, S. M., Rideout, J. R., … Zhou, H.-W. (2015). Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. *Microbiome*, *3*(1). https://doi.org/10.1186/s40168-015-0081-x

Hinchliff, C., Smith, S., Allman, J., Burleigh, G., Chaudhary, R., Cognill, L., … Cranston, K. A. (2014). Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Synthesis of Phylogeny and Taxonomy into a Comprehensive Tree of Life*, 012260. https://doi.org/10.1101/012260

Holden, T., Subramaniam, R., Sullivan, R., Cheung, E., Schneider, C., Tremberger, Jr., G., … Cheung, T. D. (2007). ATCG nucleotide fluctuation of Deinococcus radiodurans radiation genes. In R. B. Hoover, G. V. Levin, A. Y. Rozanov, & P. C. W. Davies (Eds.) (p. 669417). https://doi.org/10.1117/12.732283

Huse, S. M., Dethlefsen, L., Huber, J. A., Welch, D. M., Relman, D. A., & Sogin, M. L. (2008). Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genetics*, *4*(11). https://doi.org/10.1371/journal.pgen.1000255

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., … White, O. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207–214. https://doi.org/10.1038/nature11234

Janda, J. M., & Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. *Journal of Clinical Microbiology*, *45*(9), 2761–2764. https://doi.org/10.1128/JCM.01228-07

Kellenberger, E. (2001). Exploring the unknown. *EMBO Reports*, *2*(1), 5–7. https://doi.org/10.1093/embo-reports/kve014

Kõljalg, U., Nilsson, R. H., Abarenkov, K., Tedersoo, L., Taylor, A. F. S., Bahram, M., … Larsson, K. H. (2013). Towards a unified paradigm for sequence-based identification of fungi. *Molecular Ecology*, *22*(21), 5271–5277. https://doi.org/10.1111/mec.12481

Lan, Y., Wang, Q., Cole, J. R., & Rosen, G. L. (2012). Using the RDP classifier to predict taxonomic novelty and reduce the search space for finding novel organisms. *PLoS ONE*, *7*(3). https://doi.org/10.1371/journal.pone.0032491

Liland, K. H., Vinje, H., & Snipen, L. (2017). microclass: An R-package for 16S taxonomy classification. *BMC Bioinformatics*, *18*(1). https://doi.org/10.1186/s12859-017-1583-2

Lok, C. (2015). Mining the microbial dark matter. *Nature*, *522*(7556), 270–273. https://doi.org/10.1038/522270a

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2014). Swarm: Robust and fast clustering method for amplicon-based studies. *PeerJ*, *2014*(1). https://doi.org/10.7717/peerj.593

Mande, S. S., Mohammed, M. H., & Ghosh, T. S. (2012). Classification of metagenomic sequences: Methods and challenges. *Briefings in Bioinformatics*, *13*(6), 669–681. https://doi.org/10.1093/bib/bbs054

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., Desantis, T. Z., Probst, A., … Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME Journal*, *6*(3), 610–618. https://doi.org/10.1038/ismej.2011.139

Mercier, C., Boyer, F., Bonin, A., & Coissac, E. (2013). SUMATRA and SUMACLUST: fast and exact comparison and clustering of sequences. *Programs and Abstracts of the SeqBio 2013 Workshop. Abstract*, 27–29. https://doi.org/10.1109/RCIS.2013.6577673

Mizrahi-Man, O., Davenport, E. R., & Gilad, Y. (2013). Taxonomic Classification of Bacterial 16S rRNA Genes Using Short Sequencing Reads: Evaluation of Effective Study Designs. *PLoS ONE*, *8*(1). https://doi.org/10.1371/journal.pone.0053608

Nair, A. S., & Sreenadhan, S. P. (2006). A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). *Bioinformation*, *1*(6), 197–202. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17597888%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1891688

Navas-Molina, J. A., Peralta-Sánchez, J. M., González, A., McMurdie, P. J., Vázquez-Baeza, Y., Xu, Z., … Knight, R. (2013). Advancing our understanding of the human microbiome using QIIME. In *Methods in Enzymology* (Vol. 531, pp. 371–444). Academic Press Inc. https://doi.org/10.1016/B978-0-12-407863-5.00019-8

Nearing, J. T., Douglas, G. M., Comeau, A. M., & Langille, M. G. I. (2018). Denoising the Denoisers: An independent evaluation of microbiome sequence error- correction approaches. *PeerJ*, *2018*(8). https://doi.org/10.7717/peerj.5364

Needham, D. M., Sachdeva, R., & Fuhrman, J. A. (2017). Ecological dynamics and co-occurrence among marine phytoplankton, bacteria and myoviruses shows microdiversity matters. *ISME Journal*, *11*(7), 1614–1629. https://doi.org/10.1038/ismej.2017.29

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Ning Yu, Zhihua Li, and Z. Y. (2018). Survey on encoding schemes for genomic data representation and feature learning—from signal processing to machine learning. *Big Data Mining and Analytics*, *1*(3), 191–210. https://doi.org/10.26599/BDMA.2018.9020018

Oakley, B. B., Fiedler, T. L., Marrazzo, J. M., & Fredricks, D. N. (2008). Diversity of human vaginal bacterial communities and associations with clinically defined bacterial vaginosis. *Applied and Environmental Microbiology*, *74*(15), 4898–4909. https://doi.org/10.1128/AEM.02884-07

Petrosino, J. F., Highlander, S., Luna, R. A., Gibbs, R. A., & Versalovic, J. (2009). Metagenomic pyrosequencing and microbial identification. *Clinical Chemistry*, *55*(5), 856–866. https://doi.org/10.1373/clinchem.2008.107565

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glöckner, F. O. (2007). SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, *35*(21), 7188–7196. https://doi.org/10.1093/nar/gkm864

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., … Zoetendal, E. (2010). A human gut microbial gene catalogue established by

metagenomic sequencing. *Nature*, *464*(7285), 59–65. https://doi.org/10.1038/nature08821

Rideout, J. R., He, Y., Navas-Molina, J. A., Walters, W. A., Ursell, L. K., Gibbons, S. M., … Gregory Caporaso, J. (2014). Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. *PeerJ*, *2014*(1). https://doi.org/10.7717/peerj.545

Rockwood, A. L., Crockett, D. K., Oliphant, J. R., & Elenitoba-Johnson, K. S. J. (2005). Sequence alignment by cross-correlation. *Journal of Biomolecular Techniques*, *16*(4), 453–458.

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, *2016*(10). https://doi.org/10.7717/peerj.2584

Russell, D. J., Way, S. F., Benson, A. K., & Sayood, K. (2010). A grammar-based distance metric enables fast and accurate clustering of large sets of 16S sequences. *BMC Bioinformatics*, *11*. https://doi.org/10.1186/1471-2105-11-601

Schloss, P. D., & Handelsman, J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Applied and Environmental Microbiology*, *71*(3), 1501–1506. https://doi.org/10.1128/AEM.71.3.1501-1506.2005

Schloss, P. D., & Westcott, S. L. (2011). Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Applied and Environmental Microbiology*, *77*(10), 3219–3226. https://doi.org/10.1128/AEM.02810-10

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., … Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09

Scholz, M. B., Lo, C. C., & Chain, P. S. G. (2012). Next generation sequencing and bioinformatic bottlenecks: The current state of metagenomic data analysis. *Current Opinion in Biotechnology*, *23*(1), 9–15. https://doi.org/10.1016/j.copbio.2011.11.013

Seguritan, V., & Rohwer, F. (2001). FastGroup: A program to dereplicate libraries of 16S rDNA sequences. *BMC Bioinformatics*, *2*. https://doi.org/10.1186/1471-2105-2-9

Shade, A., Klimowicz, A. K., Spear, R. N., Linske, M., Donato, J. J., Hogan, C. S., … Handelsman, J. (2013). Streptomycin application has no detectable effect on bacterial community structure in apple orchard soil. *Applied and Environmental Microbiology*, *79*(21), 6617–6625. https://doi.org/10.1128/AEM.02017-13

Shokralla, S., Spall, J. L., Gibson, J. F., & Hajibabaei, M. (2012). Next-generation sequencing technologies for environmental DNA research. *Molecular Ecology*, *21*(8), 1794–1805. https://doi.org/10.1111/j.1365-294X.2012.05538.x

Silverman, B. D., & Linsker, R. (1986). A measure of DNA periodicity. *Journal of Theoretical Biology*, *118*(3), 295–300. https://doi.org/10.1016/S0022-5193(86)80060-1

Somervuo, P., Koskela, S., Pennanen, J., Henrik Nilsson, R., & Ovaskainen, O. (2016). Unbiased probabilistic taxonomic classification for DNA barcoding. *Bioinformatics*, *32*(19), 2920–2927. https://doi.org/10.1093/bioinformatics/btw346

STACKEBRANDT, E., & GOEBEL, B. M. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *International Journal of Systematic and Evolutionary Microbiology*, *44*(4), 846–849. https://doi.org/10.1099/00207713-44-4-846

Sun, Y., Cai, Y., Liu, L., Yu, F., Farrell, M. L., Mckendree, W., & Farmerie, W. (2009). ESPRIT: Estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Research*, *37*(10). https://doi.org/10.1093/nar/gkp285

Thompson, J. R., Pacocha, S., Pharino, C., Klepac-Ceraj, V., Hunt, D. E., Benoit, J., … Polz, M. F. (2005). Genotypic diversity within a natural coastal bacterioplankton population. *Science*, *307*(5713), 1311–1313. https://doi.org/10.1126/science.1106028

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., … Gordon, J. I. (2009). A core gut microbiome in obese and lean twins. *Nature*, *457*(7228), 480–484. https://doi.org/10.1038/nature07540

Van Rossum, G. (n.d.). Python tutorial, Technical Report {CS}-R9526, Centrum voor Wiskunde en Informatica ({CWI}), Amsterdam, May 1995. Retrieved from https://www.python.it/

Vilo, C., & Dong, Q. (2012). Evaluation of the RDP Classifier Accuracy Using 16S rRNA Gene Variable Regions. *Metagenomics*, *1*, 1–5. https://doi.org/10.4303/mg/235551

Voss, R. F. (1992). Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Physical Review Letters*, *68*(25), 3805–3808. https://doi.org/10.1103/PhysRevLett.68.3805

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R.

(2007a). Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261–5267. https://doi.org/10.1128/AEM.00062-07

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007b). Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261–5267. https://doi.org/10.1128/AEM.00062-07

Ward, D. M., Ferris, M. J., Nold, S. C., & Bateson, M. M. (1998). A Natural View of Microbial Biodiversity within Hot Spring Cyanobacterial Mat Communities. *Microbiology and Molecular Biology Reviews*, *62*(4), 1353–1370. https://doi.org/10.1128/mmbr.62.4.1353-1370.1998

Weizhong, L., & Adam, G. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659.

Werner, J. J., Koren, O., Hugenholtz, P., Desantis, T. Z., Walters, W. A., Caporaso, J. G., … Ley, R. E. (2012). Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. *ISME Journal*, *6*(1), 94–103. https://doi.org/10.1038/ismej.2011.82

Westcott, S. L., & Schloss, P. D. (2015). De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ*, *2015*(12). https://doi.org/10.7717/peerj.1487

Westcott, S. L., & Schloss, P. D. (2017). OptiClust, an Improved Method for Assigning Amplicon-Based Sequence Data to Operational Taxonomic Units. *MSphere*, *2*(2). https://doi.org/10.1128/mspheredirect.00073-17

Ye, Y. (2010). Identification and quantification of abundant species from pyrosequences of 16S rRNA by consensus alignment. *Proceedings - 2010 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2010*, 153–157. https://doi.org/10.1109/BIBM.2010.5706555

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., … Glöckner, F. O. (2014). The SILVA and "all-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Research*, *42*(D1). https://doi.org/10.1093/nar/gkt1209

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, *30*(5), 614–620. https://doi.org/10.1093/bioinformatics/btt593

Zhang, R., & Zhang, C.-T. (1994). Z Curves, An Intutive Tool for Visualizing and Analyzing the DNA Sequences. *Journal of Biomolecular Structure and Dynamics*, *11*(4), 767–782. https://doi.org/10.1080/07391102.1994.10508031

Zheng, Z., Kramer, S., & Schmidt, B. (2012). DySC: Software for greedy clustering of 16S rRNA reads. *Bioinformatics*, *28*(16), 2182–2183. https://doi.org/10.1093/bioinformatics/bts355