Bioinformatics Advance Access published October 28, 2010

SPHINX - An algorithm for taxonomic binning of metagenomic sequences

Monzoorul Haque M, Tarini Shankar Ghosh, Nitin Kumar Singh, and Sharmila S Mande*

Bio-sciences R&D Division, TCS Innovation Labs, Tata Consultancy Services Limited, 1 Software Units Layout, Madhapur, Hyderabad – 500081, Andhra Pradesh, India

Associate Editor: Prof. Martin Bishop

ABSTRACT

Motivation: Compared to composition based binning algorithms, the binning accuracy and specificity of alignment based binning algorithms is significantly higher. However, being alignment-based, the latter class of algorithms require enormous amount of time and computing resources for binning huge metagenomic data sets. The motivation was to develop a binning approach that can analyze metagenomic data sets as rapidly as composition based approaches, but nevertheless has the accuracy and specificity of alignment based algorithms. This paper describes a hybrid binning approach (SPHINX) that achieves high binning efficiency by utilizing the principles of both 'composition' and 'alignment' based binning algorithms.

Results: Validation results with simulated sequence data sets indicate that SPHINX is able to analyze metagenomic sequences as rapidly as composition based algorithms. Furthermore, the binning efficiency (in terms of accuracy and specificity of assignments) of SPHINX is observed to be comparable to results obtained using alignment based algorithms.

Availability: A web-server for the SPHINX algorithm is available at http://metagenomics.atc.tcs.com/SPHINX/

Contact: sharmila@atc.tcs.com

1 INTRODUCTION

The enormous microbial diversity prevalent in natural ecosystems represents a rich resource for discovery of hitherto unknown microbes and the novel genes/proteins they encompass. Estimates reveal that 99% of these microbes cannot be easily cultured in the laboratory (Amann *et al*., 1995). The rapidly growing field of metagenomics directly investigates this microbial diversity by obtaining and sequencing the entire genomic content present in any given environmental sample (Hugenholtz, 2002; Rappe and Giovannoni, 2003). Since majority of these organisms in environmental samples belong to hitherto unknown taxonomic groups, one of the biggest challenges for computational biologists is not just to catalog the known organisms, but also to identify and characterize new organisms belonging to known or unknown taxonomic groups. These organisms could belong to an entirely new species or genus or family or order or class or even a new phylum.

Researchers typically catalog taxonomic diversity (a process referred to as 'binning') by using computational methods that identify the taxonomic affiliation of all the sequences obtained from a given environmental sample. Sequences sharing the same taxonomic label are subsequently grouped (binned) together. Accurate binning of metagenomic sequences is a crucial step in any metagenomics project since wrongly binned sequences will affect/hinder in the downstream analysis of many subsequent steps, such as sequence assembly, gene prediction and functional annotation, etc. Existing binning algorithms assign a sequence to a particular taxon/clade if 'features' of the sequence are 'similar' to sequence(s) belonging to that taxon/clade. The extent to which these 'features' are 'similar' determines not only the accuracy, but also the specificity of assignment (the taxonomic level at which the sequence is assigned). Besides, binning methods are challenged by sequences generated by current sequencing technologies, such as 454 (Margulies *et al*., 2005) and Illumina. The short lengths of sequences (35-100 base pairs) produced using these sequencing technologies make it difficult to identify 'features' which are similar between sequences from closely related organisms and are distinct from those from distant organisms.

One class of 'features' used by binning algorithms is based on the similarity of the 'compositional' characteristics of the query and the target sequences. A few examples of 'composition based' binning algorithms published in recent years include Phylopythia (McHardy *et al*., 2007), TACOA (Diaz *et al*., 2009) and PhymmBL (Brady *et al*., 2009). Phylopythia (McHardy *et al*., 2007) utilizes Support Vector Machines and uses oligonucleotide frequencies as training features to initially build organism/clade specific classifiers. These classifiers are subsequently used for assigning a query sequence to an organism/clade whose genome(s) is/are most similar to the query (with respect to oligonucleotide usage). However, SVM based classifiers used by Phylopythia are not robust enough to predict the taxonomic labels of 'short' query sequences having lengths less than 1,000 base pairs. As a result, in addition to most of the 'short' sequences remaining unclassified,

this method has a high mis-classification rate. Another composition based binning algorithm, namely TACOA, also uses oligonucleotide frequencies as features for building organism specific models (Diaz *et al*., 2009). Based on the GC content of a genome, TACOA builds a genome model, represented in the form of a vector. The elements of these vectors contain the ratio of observed and the expected frequencies of all possible nucleotide oligomers (of k-mer sizes 4 and 5). TACOA employs a modified k-Nearest Neighbor approach (k-NN) during the prediction phase, wherein the vector corresponding to the query sequence is scored not only against the neighboring vectors of the highest scoring model, but also against the complete set of (precomputed) genome models. The final assignment of a query sequence to a taxa/clade is based on the pattern of the obtained scores. Although TACOA outperforms Phylopythia, majority of the 'short' query sequences are assigned at non-specific taxonomic levels such as super-kingdom by this method (Diaz *et al*, 2009). Thus, this method has limited utility since it prevents end-users from assessing the taxonomic diversity of samples at finer taxonomic levels. For example, sequences from two samples may be very similar at a higher taxonomic level e.g phylum, but may have sequences belonging to two entirely different classes within that phylum. The recently published Phymm binning algorithm (Brady *et al*., 2009) computes variable length oligonucleotide frequencies and generates Interpolated Markov Models (IMMs) for each genome. Subsequently, it scores a given query sequence against all precomputed organism specific models, and assigns the query to an organism which corresponds to the highest scoring model. Phymm assigns all query sequences at the taxonomic level of strain, irrespective of the value of the obtained (highest) score. The end-users then need to interpret the score and appropriately reduce individual query assignments to corresponding higher taxonomic levels. However, the absence of a linear correlation between the score and the correct taxonomic level of the predicted assignment makes this task difficult. In principle, reducing the taxonomic level of all predicted assignments to higher taxonomic levels (such as phylum or super-kingdom) would result in high accuracy. However, the utility of assignments at such non-specific taxonomic levels is limited.

The other class of binning algorithms uses 'sequence similarity' as a 'feature' to assign a query sequence to an organism/clade (Huson *et al*., 2007; Krause *et al*., 2008, Monzoorul *et al*., 2009). These algorithms, referred to as 'alignment based' binning algorithms, follow a two-phase approach. In the first phase, an alignment algorithm such as BLAST (Altschul *et al*., 1990), is used for comparing each query sequence in a given metagenomic sample with all target sequences present in a reference database. Subsequently, in the second phase, the pattern and quality of the generated BLAST hits are analyzed, and information from this analysis is utilized for finally assigning each query sequence to an organism/clade. Given the robustness of the existing alignment algorithms (BLAST), alignment based binning algorithms are observed to have greater binning accuracy and specificity as compared to existing composition based methods. However, the first phase of alignment based binning algorithms represents a major bottleneck since enormous amount of time and computing resources are needed for generating alignments of each of the sequences in a given metagenome with sequences present in the existing reference database, such as nr (presently containing more than 9 million sequences). Thus, one of the challenges is to develop binning approaches whose performance efficiency in terms of taxonomic assignments is not only comparable to that by alignment based approaches, but also have improved efficiency in terms of computational time required for taxonomic assignment.

In this paper, we present SPHINX, a hybrid binning approach which aims at reducing the overall time taken by alignment based binning approaches by approximately an order of magnitude. The approach is termed as 'hybrid' since it utilizes both 'compositional' and 'similarity' features of the query sequence during the binning process. Binning of sequences by SPHINX involves three steps. In the first step, 'compositional' features of the query sequence are utilized to identify a small subset of sequences (in the reference database) which are similar in composition to the query sequence. The second step uses BLAST to perform a 'similarity' search of the query sequence against this small subset of database sequences. In the final step, the output of these BLAST searches is analyzed, and the query is assigned to a taxon/clade with which 'significant' hits were obtained.

2 METHODS

2.1 Clustering reference database sequences: Clustering sequences present in the reference database based on the compositional characteristics is the only pre-processing step in SPHINX. For this purpose, ffn files (which contain only protein coding gene sequences) corresponding to 952 completely sequenced prokaryotic genomes were downloaded from NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Frequencies of all possible tetra-nucleotides in each coding sequence were computed and stored as a 256 dimensional vector. Tetra-nucleotides were used since previous observations had demonstrated that tetra-nucleotide frequencies carry an inherent taxonomic signal (Pride *et al*., 2003). Using k-means clustering approach (Hartigan *et al.*, 1979), more than 1.9 million vectors corresponding to the coding sequences were clustered. The Manhattan distance (L1 norm) between individual vectors was used as the similarity criterion for clustering. Once the clustering process was completed, individual cluster centroids were computed and stored. Subsequently, coding sequences tagged to each cluster were translated into protein sequences (using the appropriate genetic code). These translated sequences resulting from each cluster were stored as separate databases. A schematic work-flow depicting the pre-processing steps (described above) is given in Supplementary Figure 1.

2.2 Steps for taxonomic assignment: The following three steps were followed by SPHINX for identifying the taxon/clade to which a query sequence belongs. A schematic work-flow depicting these three steps is given in Supplementary figure 2.

Step a – Reduction of search space: A vector representing the frequencies of all 256 tetra-nucleotides in the query sequence was generated. The distance of this query vector to each 'cluster centroid' was computed and the cluster having the least distance to the query vector was identified.

Step b – Identification of similar sequence from the reduced space: A BLASTx search of the query sequence was performed against all translated sequences present in the identified cluster.

Step c – Taxonomic assignment of sequences: Alignment parameters (bit-score, alignment length, identities, positives), organism name and the alignment sub-sequences from each hit were parsed from the BLASTx output obtained in Step b. Using an approach similar to SOrt-ITEMS (Monzoorul *et al*., 2009), each query sequence was finally assigned to a taxon/clade based on the quality of alignment and the degree of orthology observed between the query and the hit sequences.

2.3 Validation of the SPHINX approach on simulated metagenomic data sets: The performance of binning methods is usually validated using a "leave one (species) out" strategy, wherein sequences belonging to only a single species are removed from the reference database. Subsequently, test sequences derived from the removed species are queried against this modified database. This strategy is intended to simulate a scenario of only one 'unknown' (new) species being present in the metagenomic sample. However, in this study, a "leave one clade out" strategy was adopted, wherein sequences belonging to an entire clade (genus, family, order, class and phylum) were removed from the reference database. Subsequently, test sequences originating from genomes belonging to the removed clades were queried against this modified database. This strategy was adopted to closely mimic sequences derived from a real metagenomic scenario, wherein majority of sequences belong to entirely hitherto unknown (new) clades at various taxonomic levels.

2.3.1 Generation of a simulated reference database: For creating a simulated database scenario, coding sequences corresponding to 300 randomly selected genomes (out of 952) were removed before creating clusters for the reference database. Tracing the taxonomic lineage of these 300 genomes revealed the complete removal of certain clades (species, genera, families, orders, classes and phyla) from the reference database. Supplementary Table 1 lists the 300 genomes which were not considered while clustering the reference database.

2.3.2 Test data sets used for validation: Validation of the SPHINX approach was performed using 1,40,000 test sequences of varying lengths. These test sequences were generated using MetaSim (Daniel *et al.*, 2008). Supplementary Table 2 lists the genomes from which these test sequences were derived. To mimic a typical metagenomic scenario, 1,32,000 (94.2%) of these test sequences were derived from genomes belonging to those clades which were removed while creating the reference database. Based on the lengths of the read sequences, these test sequences were divided into four validation data sets, termed as Sanger data set, 454-400 data set, 454-250 data set, 454-100 data set, each containing 35,000 sequences. Test sequences constituting these four data sets simulated the typical sequence lengths and error~~s~~ models obtained from the four commonly used sequencing technologies, namely Sanger (read length centered around 800 bp), 454-GS-FLX-Titanium (400 bp), 454-GS-FLX-Standard (250 bp), and Roche-454-GS20 (100 bp), respectively. Taxonomic assignments of sequences in all four data sets were carried out following the steps described in section 2.2.

2.3.3 Categorization of taxonomic assignments: Taxonomic assignments obtained with SPHINX were categorized as 'correct/incorrect' using the following approach. Taxonomic assignment of query sequences to either the taxa corresponding to their source organism, or to taxa that lie in the path between the root to the taxa corresponding to the source organism of the query sequence were categorized as 'correct'. Assignments of sequences to taxa which did not lie in the path from the root to the taxon corresponding to the respective source organism of the query sequence were categorized as 'wrong'. For example, query sequences from *Burkholderia ambifaria* AMMD which were not assigned to any of the taxa mentioned below were categorized as 'wrong'.

root; cellular organisms; Bacteria; Proteobacteria; Betaproteobacteria; Burkholderiales; Burkholderiaceae; Burkholderia; Burkholderia cepacia complex; *Burkholderia ambifaria*; *Burkholderia ambifaria* AMMD

To evaluate the performance of SPHINX in terms of specificity of assignments, the percentage of query sequences 'correctly' assigned at the taxonomic levels of phylum or below were considered as 'specific' assignments. A binning algorithm with high specificity is expected to assign more number of sequences at 'specific' levels.

2.3.4 Comparison with other methods: The results obtained with SPHINX were compared with those obtained with two alignment based methods, namely SOrt-ITEMS (Monzoorul *et al.*, 2009) and MEGAN (Huson *et al.*, 2007). Results were also compared with those obtained with a composition based method, namely TACOA (Diaz *et al*., 2009). All three programs were run with default parameters. In order to accurately compare the effect of database clustering on binning time and binning efficiency, results of SOrt-ITEMS and MEGAN were obtained by performing similarity searches against the same (translated) ffn database (used for validating SPHINX), but in an un-clustered format. Similarly, taxonomic assignments were obtained with TACOA using a modified database containing only those genomes which were present in the databases used by SPHINX, SOrt-ITEMS and MEGAN during their validation process.

For various assignment categories, the overlap between the sequence assignments by SPHINX, SOrt-ITEMS, MEGAN and TACOA were analyzed.

2.4 Validation of the SPHINX approach on a real metagenomic data set: The performance of SPHINX was validated on sequences belonging to the hyper-saline saltern metagenome (Dinsdale *et al*., 2008). This relatively small metagenomic data set (with only 35,446 sequences) was chosen since it enabled a quick comparison of taxonomic assignment patterns obtained using various binning methods. Results obtained with this data set, would also indicate the suitability of SPHINX for use with much bigger metagenomic data sets.

Taxonomic assignments for the 35,446 sequences constituting this data set were first obtained using SPHINX. A database (in clustered format) consisting of coding sequences belonging to completely sequenced prokaryotic genomes was used during the binning process. Taxonomic assignments of all sequences were also obtained using SOrt-ITEMS and MEGAN. Assignments obtained by all three methods were then compared at the level of phylum, class and order. For this purpose, assignments obtained at or below the desired taxonomic level of comparison were first collapsed to that level and subsequently enumerated.

2.5 Testing of the SPHINX approach on mouse gut metagenomic sequences: The performance of SPHINX was tested on the metagenome data sets corresponding to the gut from lean and obese mice which were previously analyzed by Turnbaugh *et al*., (2006). In contrast to the hyper-saline saltern metagenome, the two data sets of the mouse gut metagenome represented a typical metagenomic data set in terms of sheer volume and taxonomic complexity. Analysis of such a huge and complex data set would

help in evaluating the performance of SPHINX in a real-world data scale-up scenario.

The two data sets (hereafter referred to as lean data set and obese data set), consisting of 10,57,022 and 6,87,244 sequences from the lean and the obese mouse gut respectively, were analyzed using SPHINX. Taxonomic assignments by SPHINX for both data sets were compared at the level of phylum, class, order and family in the following manner. All assignments obtained at or below the desired taxonomic level of comparison were first collapsed to that level and subsequently enumerated. For each desired taxonomic level, a 2x2 contingency matrix (for that taxon) was then generated using the approach described by White *et. al.,* (2009). The cumulative number of assignments obtained for that taxa in both data sets and the total number of assignments at or below that taxonomic level in both data sets were used for generating the matrix. A chi-square test was subsequently performed on this contingency matrix. Based on the results of the test, taxa (having more than 1,000 assigned sequences) showing a statistically significant difference were identified.

2.6 Assessing the impact of non-coding regions on binning efficiency: The efficiency of alignment-based binning algorithms is expected to differ, when test sequences are queried against a reference database which contains sequences originating from not only coding regions but also sequences from the non-coding regions of known genomes. In order to study the impact/contribution of non-coding regions on binning efficiency of SPHINX, a simulated database was created in the following manner. All completely sequenced genomes from NCBI were downloaded, and each genome was dissected into 1kb fragments. Using the same steps described in section 2.1, these fragments were clustered using k-means approach, followed by computation of cluster centroids and finally storing of translated sequences belonging to each cluster into separate databases. For the purpose of comparing the results of this approach with other methods, sequence fragments obtained from the same 300 randomly selected genomes were not considered while clustering and subsequently creating the database. The performance of SPHINX was evaluated with this simulated database variant.

3 RESULTS

The summarized binning results obtained with SPHINX, SOrt-ITEMS, MEGAN and TACOA (for all four validation data sets) are given in Supplementary Table 3. A graphical representation of these results is illustrated in Figure 1. As mentioned previously, these results were generated using the ffn database in clustered (SPHINX) and un-clustered formats (SOrt-ITEMS, MEGAN and TACOA). The results of SPHINX, SOrt-ITEMS and MEGAN obtained with ffn database are referred to as results with 'SPHINX-FFN', 'SOrt-ITEMS-FFN' and 'MEGAN-FFN', respectively. Results of SPHINX obtained using a reference database containing both coding and non-coding regions are referred to as results with 'SPHINX-FNA'.

3.1 Impact of database clustering on binning time and binning efficiency: Supplementary Table 3 shows a comparison of results obtained with SPHINX, SOrt-ITEMS, MEGAN and TACOA. These results are discussed below

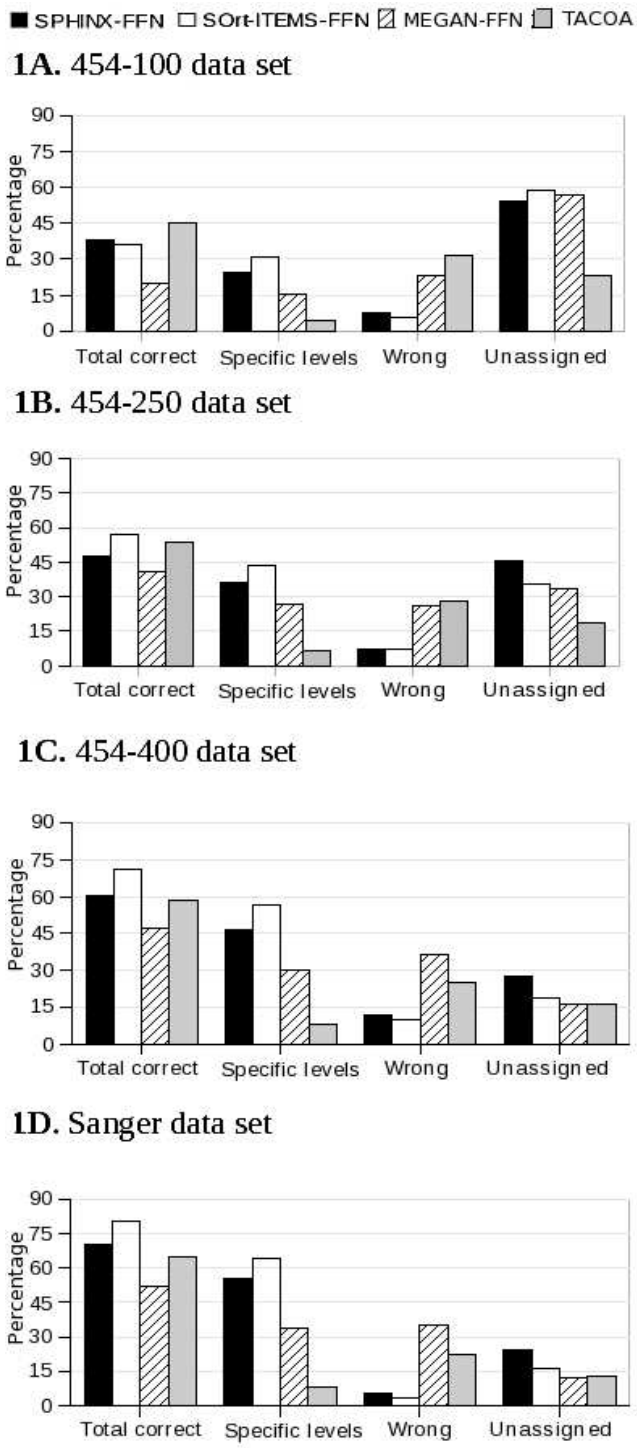

Figure 1: Graphical representation of results obtained with SPHINX-FFN, SOrt-ITEMS-FFN, MEGAN-FFN and TACOA. While assignment categories are given on the x-axis, the scale on the y-axis denotes the percentage of sequences assigned. (Database variant FFN contains coding sequences from completely sequenced genomes)

Correct Assignments: As expected, the percentage of correct assignments (by all four methods) shows a progressive increase as the length of the query sequence increases from 100 to 800 bp (Figure 1). The percentage of correct assignments by SPHINX-FFN is slightly less (by 2-10%) as compared to SOrt-ITEMS-FFN for all the data sets. This indicates that the 'reduced search-space' (database clustering) strategy used by SPHINX does not have a significant impact on binning accuracy. At the same time, this reduced search space strategy of SPHINX helps in achieving a 15 to 20 fold reduction in the overall time taken for binning (Table 1).

Furthermore, it is also observed that the number of sequences correctly classified by SPHINX-FFN is 1.2 to 1.9 times higher as compared to that by MEGAN-FFN. This is expected, since MEGAN uses only a single alignment parameter (bit-score) to judge the quality of hits in BLASTx outputs and subsequently assigns sequences using the Least Common Ancestor (LCA) approach. Adopting this approach was earlier shown to be associated with a high misclassification rate, especially in a metagenomic scenario, wherein majority of sequences originate from hitherto unknown organisms/taxonomic clades (Monzoorul *et al.*, 2009).

Though the percentage of correct assignments by SPHINX-FFN is seen to be comparable with a composition based binning method like TACOA, it is observed that a majority of the correct assignments (85-90%) by TACOA are at the level of super-kingdom. The assignments made at such non-specific levels have limited utility.

Specific assignments: As in the case of correct assignments, the percentage of correct assignments at specific levels (i.e assignments at the taxonomic level of phylum or below) increases with increasing length of the query sequences (Figure 1). It is seen that SPHINX-FFN, SOrt-ITEMS-FFN and MEGAN-FFN assign 76-80% of the correct assignments at specific levels, indicating that the reduced space strategy does not significantly affect the binning specificity. The only exception to this pattern is observed in the 454-100 data set, where SPHINX-FFN is able to assign 64% of correctly assigned sequences at specific levels. On the other hand, it is observed that less than 14% of query sequences are assigned by TACOA at specific taxonomic levels. This demonstrates (and confirms) the high binning specificity of alignment based binning approaches.

Supplementary Table 3 gives a breakup of specific assignments by all four methods at various taxonomic levels, namely, phylum, class, order, family, genus and below. These results indicate that, in most data sets, SPHINX assigns a relatively higher percentage of sequences at phylum level as compared to that by SOrt-ITEMS, MEGAN and TACOA. No clear pattern can be established between the results of SPHINX, MEGAN and TACOA at the class level. However, it is observed that SPHINX assigns a relatively lower percentage of sequences at class level compared to SOrt-ITEMS. Though it is observed that SPHINX and SOrt-ITEMS do not assign any sequences at order level, very few

Table 1: Average time (minutes) taken for binning 10,000 sequences using SPHINX-FFN, SOrt-ITEMS-FFN, MEGAN-FFN, and SPHINX-FNA. All values were estimated using a desktop with the following specifications - Intel Xeon quad core processor and 4 GB RAM.

FRAGMENT LENGTH	DATABASE VARIANT			
	SPHINX-FFN[1]	SOrt-ITEMS-FFN[1]	MEGAN-FFN[1]	SPHINX-FNA[2]
800 bp	21	381	366	97
400 bp	17	338	323	69
250 bp	13	340	305	51
100 bp	10	270	265	32

1 Database variant FFN contains coding sequences from completely sequenced genomes.

2 Database variant FNA contains both coding and non-coding sequences from completely sequenced genomes

sequences (0.4 to 4%) are assigned by MEGAN and TACOA at this level. Furthermore, the percentage of sequences assigned by SPHINX at family level is consistently higher than that by MEGAN and TACOA at this level. Furthermore, the percentage of sequences assigned by SPHINX at family level is consistently higher than that by MEGAN and TACOA. However, the percentage of sequences assigned by SPHINX at this level is relatively lower (0.4 to 7%) than that by SOrt-ITEMS. At the taxonomic level of genus and below, it is seen that the percentage of assignments by SPHINX is comparable to that by SOrt-ITEMS and MEGAN.

However, SPHINX assigns a significantly higher percentage of sequences at this level as compared to TACOA (in all four data sets). In addition to the above observations, the percentage of assignments by SPHINX at non-specific levels (above the level of phylum) is observed to be consistently lower than that by SOrt-ITEMS and MEGAN (in three out of four data sets) and TACOA (in all four data sets). This further indicates that the reduced search space approach of SPHINX does not impact the specificity of assignments.

Wrong assignments: Only 5-12% of query sequences are incorrectly assigned by SPHINX-FFN (Figure 1). This is comparable to the percentage of sequences (4-10%) incorrectly assigned by SOrt-ITEMS-FFN. This further indicates that the reduced space strategy adopted by SPHINX does not significantly affect the accuracy of binning. In comparison, a much higher percentage of sequences are misclassified by TACOA (22-31%) and MEGAN (23-35%).

Unassigned: As seen from Figure 1, the percentage of query sequences categorized as unassigned increases with decreasing length of the query sequences. Though, the number of unassigned sequences is slightly higher for SPHINX-FFN as compared to SOrt-ITEMS-FFN and MEGAN-FFN in the case of 454-250, 454-400 and Sanger data sets, it is interesting to note that in the

case of 454-100 validation data set, the number of sequences categorized as 'unassigned' by SPHINX-FFN is significantly lower than that of SOrt-ITEMS-FFN and MEGAN-FFN. The number of sequences categorized as unassigned by TACOA is significantly lower than other binning methods. However, as observed above, the assignments made by TACOA are of limited utility, since most of these assignments are at super-kingdom level.

Pattern of overlap between sequence assignments by all four binning methods

Venn diagrams depicting the overlap between the sequence assignments by SPHINX, SOrt-ITEMS, MEGAN and TACOA (for various assignment categories) are given in Supplementary Figure 4. Except for the 454-100 data set, the pattern of overlap for correctly assigned sequences indicates that assignments by SPHINX and MEGAN form a subset of the assignments by SOrt-ITEMS. For the 454-100 data set, the correctly assigned sequences by SOrt-ITEMS and MEGAN form a subset of the correct assignments by SPHINX. Furthermore, the correct assignments by TACOA are seen to have a partial overlap with the correct assignments by other three methods. However, all the additional sequences correctly classified by TACOA (in comparison to the other three methods) are observed to be classified at non-specific taxonomic levels. The overlap between the non-specific sequence assignments by all four methods, also indicates that a high proportion of correct assignments by TACOA are classified at non-specific levels.

For all data sets, a clear pattern of overlap is observed between the assignments at 'specific levels' by all four methods (Supplementary Figure 4). SOrt-ITEMS is seen to have the highest set of specific assignments. The specific assignments of SPHINX, MEGAN and TACOA are seen to form progressively smaller subsets within the set of specific sequences of SOrt-ITEMS. In contrast, for the wrong assignments category, the set of sequences misclassified by SPHINX and SOrt-ITEMS are seen to form small subsets within the set of sequences misclassified by MEGAN and TACOA.

The pattern of overlap seen for the set of unassigned sequences by all four methods (Supplementary Figure 4) indicates the following. For the 454-100 data set, the set of unassigned sequences by SPHINX and TACOA are seen to form a subset of the unassigned sequences by SOrt-ITEMS. In contrast, for the other three data sets, sequences categorized as 'unassigned' by SOrt-ITEMS, MEGAN and TACOA are observed to form progressively smaller subsets within the set of sequences categorized by SPHINX as 'unassigned'.

3.2 Impact of non-coding regions on binning efficiency

The results obtained with both variants of SPHINX (FFN and the FNA variants) are graphically illustrated in Figure 2. Supplementary Table 3 gives a comparison of results (at various taxonomic levels) obtained by the two variants of SPHINX with those obtained using SOrt-ITEMS-FFN, MEGAN-FFN and TACOA. Results indicate that 54-74% of sequences in test data

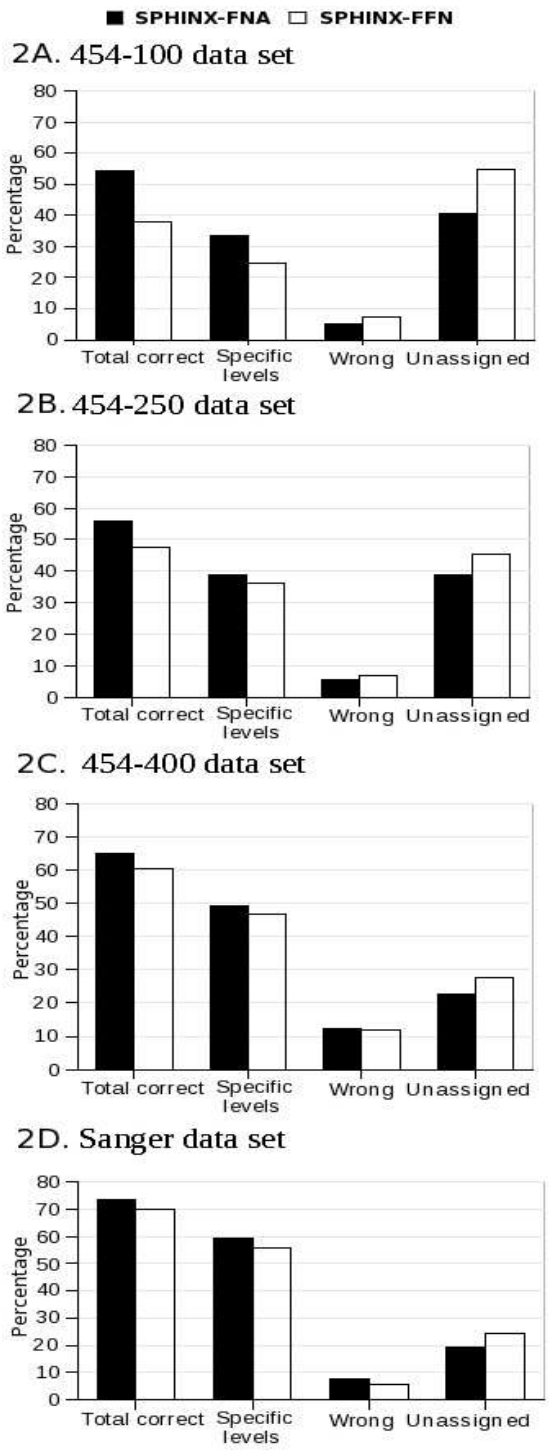

Figure 2: Graphical representation of results obtained with SPHINX-FFN and SPHINX-FNA. While assignment categories are given on the x-axis, the scale on the y-axis denotes the percentage of sequences assigned. (Database variant FFN contains coding sequences from completely sequenced genomes. Database variant FNA contains both coding and non-coding sequences from completely sequenced genomes)

sets are correctly assigned by SPHINX-FNA as compared to 37-70% by SPHINX-FFN.

Interestingly, for data sets containing short query sequences (100 bp), using the SPHINX-FNA variant (that performs searches against a database that contains fragments originating from both coding and non-coding regions), results in a noticeably higher percentage of correct assignments (54%) as compared to the SPHINX-FFN variant (37%). This is probably due to the fact that some of the assigned query sequences in the 454-100 data set may have originated partially/entirely from non-coding regions. However, as the length of the query sequences increases, it is observed that the difference between the percentage of correct assignments by both variants of SPHINX becomes smaller.

For both variants of SPHINX, (except for the 454-100 data set), it is observed that the percentage of specific assignments out of the correctly assigned sequences ranges between 69-80%. Furthermore, the difference between the percentage of wrong assignments by both variants of SPHINX ranges between 0.33-2.25%. This indicates that for data sets containing sequences having lengths greater than equal to 250 bp, the inclusion of the non-coding regions in the reference database does not significantly alter the accuracy and specificity of binning. It is also seen that the inclusion of the non-coding regions in the database increases the overall time taken for binning by three to four times (Table 1). The increase in time for FFN data sets is due to the added volume of the non-coding sequences in the database. Based on the above observations, it can be concluded that it is optimal to use the SPHINX-FNA variant only in cases of data sets containing short sequences . It is also important to note that the binning time of both variants of SPHINX increases linearly with increase in the size of the data sets (Supplementary Table 5).

3.3 Validation results of SPHINX for hyper-saline saltern metagenomic data set

A comparison of the total numbers and the cumulative percentage of sequences (in the hyper-saline saltern metagenomic data set) assigned at phylum, class and order levels by SPHINX-FFN, SOrt-ITEMS-FFN and MEGAN-FFN is shown in Table 2.

Results in Table 2 indicate that SPHINX is able to assign 48.6% of the sequences in the hyper-saline saltern metagenomic data set. This is roughly similar to the percentage of sequences assigned by SOrt-ITEMS-FFN (49.9%) and MEGAN-FFN (50.5%). For SPHINX, it is also observed that 15,773 (i.e 91.6%) of the assigned sequences (17,227) are at taxonomic levels of phylum or below. A similar pattern is also seen for SOrt-ITEMS (93.8%) and MEGAN (89.7%). Furthermore, it is observed that a majority (98.3-99.64%) of assignments (by all three alignment-based methods) are to microbial species belonging to two orders, namely Halobacteriales and Sphingobacteriales. This observation is in conformance with earlier reports which indicated the over abundance of these two orders in hyper-saline environments (Dinsdale *et al.*, 2008; Willner *et al.*, 2009). Further, all three methods show a similar pattern of taxonomic assignments at all taxonomic levels (Table 2). These observations reaffirm that the

Table 2: The cumulative percentage of sequences (in the hyper-saline saltern metagenomic data set) assigned by SPHINX-FFN, SOrt-ITEMS-FFN and MEGAN-FFN at order, class and phylum levels.

Binning method	Total number of sequences	Total number of sequences assigned	Cumulative number of sequences assigned at different taxonomic levels		
			PHYLUM	CLASS	ORDER
SPHINX	35446	17227 (48.6%)	15773	15632	15632
SOrt-ITEMS	35446	17688 (49.9%)	16589	16518	16518
MEGAN	35446	17901 (50.5%)	16057	15703	15703

TAXON NAME	Relative percentage of sequence assignments to various taxa at 'order' level		
	SPHINX	SOrt-ITEMS	MEGAN
Halobacteriales	62.76	62.14	64.67
Bacteroidales	0	0	0.04
Chromatiales	0	0	0.08
Sphingobacteriales	35.54	37.5	33.8
Burkhoderiales	0.05	0.08	0.27
Flavobacteriales	0.08	0.08	0.04
Actinomycetales	0.78	0.07	0.76
Myxococcales	0.58	0	0.3
Rhodobacterales	0.01	0.06	0.04
Clostridiales	0.2	0	0

TAXON NAME	Relative percentage of sequence assignments to various taxa at 'class' level		
	SPHINX	SOrt-ITEMS	MEGAN
Deltaproteobacteria	0.61	0	0.3
Alphaproteobacteria	0.01	0.06	0.08
Gammaproteobacteria	0.05	0.09	0.12
Halobacteria	62.88	62.11	64.61
Sphingobacteria	35.1	37.48	33.77
Bacteroidia	0	0	0.04
Betaproteobacteria	0.05	0.08	0.27
Flavobacteria	0.08	0.08	0.04
Actinobacteria (class)	0.78	0.07	0.76
Clostridia	0.44	0	0

TAXON NAME	Relative percentage of sequence assignments to various taxa at 'phylum' level		
	SPHINX	SOrt-ITEMS	MEGAN
Bacteroidetes	35.23	33.14	33.47
Proteobacteria	0.67	2.14	1.22
Actinobacteria	0.78	1.35	0.75
Euryarchaeota	62.88	62.97	64.6
Firmicutes	0.44	0.22	0
Chloroflexi	0	0.08	0
Cyanobacteria	0	0.07	0
Crenarchaeota	0	0.03	0

'reduced search-space' strategy used by SPHINX does not have a significant impact on binning accuracy and specificity. The binning efficiency of SPHINX is further highlighted by the fact that SPHINX takes only 88 minutes to analyze this data set, as compared to more than 30 hours taken by SOrt-ITEMS and MEGAN. For all the three methods, the execution time indicated above includes the time taken for the BLASTx comparisons.

3.4 Testing of SPHINX approach on mouse gut data sets

Supplementary Table 5 shows the summarized results obtained with SPHINX for the lean and obese mouse gut data sets. SPHINX is able to assign 54% (5,71,921 out of 10,57,022) and 41% (2,86,945 out of 6,87,244) of the sequences in the lean and obese data sets, respectively (Supplementary Table 5A). Besides, in both data sets, an examination of the BLAST hits revealed that alignments obtained for approximately two-thirds of the assigned sequences were of poor quality. Consequently, these sequences were assigned by SPHINX at the taxonomic level of phylum. In contrast, only 0.2% of sequences generated BLAST hits with alignments having identity values of greater than 90% and were subsequently assigned at the level of genus or below. These observations indicate that the majority of sequences in the mouse gut metagenome have originated from genomes of hitherto unknown organisms. Table 3 provides a qualitative analysis of the cumulative assignments obtained at the taxonomic level of phylum. Qualitative analysis of cumulative assignments at other taxonomic levels, namely class, order and family, are given in Supplementary Tables 5B-D.

Analysis of assignments at the level of phylum indicates the presence of organisms belonging to 24 and 23 phyla in the lean and obese data sets, respectively (Table 3). Results with SPHINX indicate that a significantly higher percent of sequences in the obese data set are assigned to the phylum Firmicutes (23.7%) and Euryarchaeota (6.6%), as compared to percentage of sequences assigned to these phyla in the lean data set (14.5% and 5.4%, respectively). In contrast, the phylum Bacteroidetes is present in a significantly higher proportion in the lean data set (6.9%) in comparison with that in the obese data set (4.8%). The above observations are in line with the previous findings by Turnbaugh *et al.* (2006). In addition, results of SPHINX indicate the new finding that most of the sequences from the obese and the lean data sets are assigned to the phyla Proteobacteria (49.4%, 43%), Actinobacteria (8.5%, 5.2%) and Cyanobacteria (5.5%, 5.9%).

An analysis of assignments to various classes under each identified phyla revealed interesting patterns of taxonomic diversity in lean and obese data sets (Supplementary Table 5B). In the lean data set, 5584 sequences are assigned to the class Actinobacteria, 1936 sequences to class Alphaproteobacteria, 1926 sequences to class Betaproteobacteria, and approximately 600-700 sequences each to classes Chloroflexi, Spirochaetes and Methanomicrobia. In contrast, none of the sequences in the obese data set are assigned to any of these classes. This observation could either be due to the complete absence of organisms belonging to these classes in the obese data set or due to the presence of hitherto unknown organisms belonging to these classes, the genomes of which have significantly diverged from the known organisms within these classes. Since sequences from such diverged genomes will generate BLAST hits having extremely weak alignment parameters, they are assigned by SPHINX either at higher taxonomic levels (phylum or above) or are categorized as unassigned. Similar to the pattern observed at the level of phylum, while a significantly higher proportion of sequences of the lean data set are assigned by SPHINX to class Bacteroidia (belonging to phylum Bacteroidetes), the classes Clostridia and Bacilli are present in significantly higher proportion in the obese data set. Many sequences from both the lean and obese data sets are assigned to classes Gammaproteobacteria, Deltaproteobacteria and Chlorobia. However, differences in the relative proportion of sequences assigned to the latter classes in both data sets are not statistically significant.

Table 3: Taxonomic assignments obtained using SPHINX-FNA at the level of phylum in the lean and obese mouse gut data sets

PHYLUM	Lean	Obese	CHI SQUARE	PREVALENCE
Crenarchaeota	6063	6550	19	High in Obese
Aquificae	1142	1928	202	High in Obese
candidate division TG1	190	304	26	High in Obese
Euryarchaeota	20948	25438	462	High in Obese
Nitrospirae	292	497	53	High in Obese
Deinococcus-Thermus	1413	1667	21	High in Obese
Verrucomicrobia	2349	3334	171	High in Obese
Spirochaetes	3327	4778	262	High in Obese
Firmicutes	56247	91769	10541	High in Obese
Tenericutes	855	1532	192	High in Obese
Cyanobacteria	21418	22801	45	High in Obese
Chlamydiae	661	1166	139	High in Obese
Dictyoglomi	305	553	71	High in Obese
Thermotogae	2537	3521	161	High in Obese
Fusobacteria	141	281	46	High in Obese
Korarchaeota	249	0	249	High in Lean
Chloroflexi	5739	5204	26	High in Lean
Bacteroidetes	26830	18674	1553	High in Lean
Chlorobi	8645	8332	5	High in Lean
Gemmatimonadetes	564	321	66	High in Lean
Proteobacteria	191006	166353	3160	High in Lean
Actinobacteria	32867	20150	3274	High in Lean
Planctomycetes	1024	500	180	High in Lean
Acidobacteria	2060	1209	222	High in Lean

Assignments at the levels of order and family also show assignment patterns similar to those seen at the level of class. The obese data set is characterized by the conspicuous absence of many orders/families which are present in the lean data set (Supplementary Table 5C-D).

It is to be noted that all inferences mentioned above were based on similarity searches against a clustered sequence database created using genomic fragments from 952 completely sequenced prokaryotic genomes (available in the NCBI database at the time of analysis). Though the overall pattern of taxonomic assignments confirm to those obtained in earlier studies (Turnbaugh *et al.*, 2006), these results are expected to improve as sequence information of more and more genomes becomes available.

4 DISCUSSION

Current alignment based binning methods depend on exhaustive database searches. Performing such exhaustive database searches for millions of sequences (present in typical metagenomic samples) would take enormous amounts of time. This hinders research groups having modest computational resources from using similarity-based binning methods. The hybrid approach - SPHINX presented in this paper addresses this issue by adopting a 'reduced search-space' approach. Using this approach, SPHINX achieves a 15 to 20 fold reduction in the time taken for binning, compared to other binning approaches which depend on exhaustive database searches (Table 1). On a single desktop with modest specifications (Intel Xeon quad core, 4GB RAM), SPHINX-FNA takes just about 88 hours for binning 1.6 million sequences of the mouse gut metagenome. Binning the same set of sequences using SPHINX-FFN would have further reduced the time to an estimated 20-24 hours without significant loss in accuracy and specificity. In contrast, exhaustive search approaches (against the same database) would require an estimated six weeks for completing the same task. Furthermore, validation results of SPHINX obtained with simulated data sets indicate that the binning efficiency of SPHINX matches those obtained with alignment based binning algorithms which depend on the output of exhaustive database searches (e.g. SOrt-ITEMS).

An examination of the results obtained for the same set of sequences with different variants of databases indicate highest binning efficiency (in terms of accuracy and specificity) with the SPHINX-FNA database variant since it contains fragments originating from both coding and non-coding regions (including fragments spanning both these regions) of completely sequenced genomes. However, results of SPHINX obtained with a database variant devoid of fragments spanning coding and non-coding regions (referred to as SPHINX-C-NC in Supplementary Table 3) indicate that using such a database variant results in an increase in taxonomic assignments at non-specific levels, as well as, an increase in the percentage of wrong assignments (as compared to the SPHINX-FNA variant). These results show that the most optimal results are obtained using the SPHINX-FNA database variant.

Since our observations with clustered and non-clustered variants of the FFN database indicate no significant loss in binning efficiency using a reduced search-space strategy, it will be beneficial to adopt the latter strategy even for the nr database (which is more comprehensive in terms of having sequences even from partially characterized genomes in addition to sequences from completely sequenced genomes). This will help in significantly reducing the overall time taken by alignment based binning approaches.

The quality of reference database clusters is an important factor that determines the binning specificity and accuracy of the SPHINX algorithm. Though this study adopted the k-means clustering approach, in principle, any other clustering method that can efficiently segregate DNA sequences (of varying oligonucleotide composition) in feature vector space can be applied to the current work-flow of SPHINX. The k-means clustering approach adopted in the present study takes only about 4 hours on a simple desktop to cluster approximately 1.9 million vectors corresponding to coding sequences of 952 prokaryotic genomes. This includes the time taken for generating all the vectors, as well as, for obtaining the vectors corresponding to the cluster centroids.

The quality of the clustered reference database is also dependent on the k-mer size. A typical metagenomic sequence (obtained using existing sequencing technologies) has a length less than 800-1000 base pairs. For such short sequences, the taxonomic discrimination capability achieved using lower k-mer sizes is expected to be poor. On the other hand, the frequencies of various oligo-nucleotides obtained using large k-mer sizes will be extremely low and statistically insignificant. A k-mer size of 4 was used in the current study since an earlier study had demonstrated that tetra-nucleotide frequencies carry an inherent taxonomic signal (Pride *et al*., 2003). However, using k-mer size of 4 limits the applicability of the current method for obtaining reliable taxonomic assignments for very short sequences, especially those obtained from sequencing technologies like Solexa and Illumina.

The binning efficiency of SPHINX also depends on the taxonomic coverage of the clustered reference database. The clustered database used in the current study contained sequences originating from 952 completely sequenced prokaryotic genomes. As more and more genomes are being sequenced, updating the clustered database with sequences from these genomes will help in improving the quality of taxonomic assignments. Updating the clustered reference database with sequences from these new genomes takes only a few minutes, since the process only involves mapping the new sequences to existing cluster centroids, and subsequently recomputing the cluster centroids.

The clustered reference database used in the current study contained sequences originating from prokaryotic genomes. In principle, this database can be constructed using sequences from other taxonomic domains, such as viruses or eukaryotes. However, given that the information content in sequences originating from non-coding regions (which form major portions of eukaryotic genomes) does not conform to known patterns of taxonomic hierarchy, extending the SPHINX approach to such sequences will be a challenging task. It is also difficult to bin sequences originating from repetitive stretches (characteristic of eukaryotic genomes) at specific taxonomic levels. This is due to the following reason. Sequences originating from repetitive stretches will generate hits with multiple sequences originating from taxa belonging to diverse taxonomic clades. Since, it is difficult to identify the exact genome from which the query sequence has originated, such sequences can be assigned only at high (non-specific) taxonomic levels. An additional challenge will be to taxonomically characterize metagenomic fragments containing overlapping reading frames, known to be prevalent in viral and eukaryotic genomes.

In spite of the high accuracy and specificity of alignment based approaches, the percentage of metagenomic sequences classified (by these approaches) either as 'unassigned' or at non-specific taxonomic levels is still very high. This is due to the fact that many of these sequences either do not generate alignments or generate

poor quality alignments with sequences in existing reference databases. This is typical of sequences originating from organisms belonging to novel (hitherto unknown) taxonomic clades. It will be desirable if alignment based approaches are complemented with methods that facilitates meaningful grouping of such sequences.

5 CONCLUSION

Alignment based binning algorithms are observed to have greater binning accuracy and specificity as compared to existing composition based methods. In this paper, we demonstrate that adopting a reduced search space strategy enables one to drastically reduce the overall time taken for alignment based binning approaches, with no significant loss of binning efficiency. This indicates the immense applicability of the proposed algorithm in rapidly mapping the taxonomic diversity of large metagenomic samples with high accuracy and specificity.

ACKNOWLEDGMENTS

We thank Stephan Schuster and Daniel Huson for allowing us to use MEGAN and MetaSim software for this work. We thank Sudha Chadaram and Dinakar Komanduri for their help in testing SPHINX and for preparing a web-server for the same.

REFERENCES

Altschul,S., Madden,T., Shaffer,A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D. (1997) Gapped Blast and PSIBlast: a new generation of protein database search programs. *Nucleic Acids Res*.,25., 3389–3402.

Amann RI, Ludwig W, Schleifer KH.(1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* 59: 143-69

Brady A, Salzberg SL (2009) Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods* 6: 673–676

Diaz NN, Krause L, Goesmann A, Niehaus K, Nattkemper TW (2009) TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics* 10:56

Dinsdale EA, et al (2008) Functional metagenomic profiling of nine biomes. *Nature* 452:629–632. doi:10.1038/nature06810.

Hartigan, JA. and Wong, MA. A K-Means Clustering Algorithm. *Applied Statistics*. 1979;28:100–108. doi: 10.2307/2346830.

Hugenholtz P. (2002) Exploring prokaryotic diversity in the genomic era. *Genome Biol*. 3. REVIEWS0003.

Huson DH, Auch AF, Qi J, Schuster SC. (2007) MEGAN analysis of metagenomic data. *Genome Res*. 17:377–386.

Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, et al. Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res.* 2008;36(7):2230–2239

Konstantinidis, KT. and Tiedje, JM. (2005) Towards a Genome-Based Taxonomy for Prokaryortes. *J. Bacteriol*., 187(18): 6258 – 6264.

Margulies, M., M. Egholm, et al. (2005). Genome sequencing in micro-fabricated high-density pico-litre reactors. *Nature* 437(7057): 376-80.

Mavromatis K, et al. (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat. Methods* 4:495–500.

McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I. (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods* 4:63–72.

Monzoorul HM, Tarini SG, Dinakar K, Sharmila SM (2009) SOrt-ITEMS: sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics* 25:1722–1730

Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ: Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* 2003, **13:**145-158.

Rappe, M. and Giovannoni, S. (2003) The uncultured microbial majority. *Annu. Rev. Microbiol*, . 57, 369–394

Richter DC, Ott F, Auch AF, Schmid R, Huson DH (2008). MetaSim - A Sequencing Simulator for Genomics and Metagenomics. *PLoS ONE* 3(10): e3373.

Sanger, F., et al. (1977) The nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 265, 687–695

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin E, Rokhsar DS, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428:37–43.

Teeling H, Waldmann J, Lombardot T, Bauer M, Glockner FO. (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* 5:163.

Venter JC, et al, Environmental genome shotgun sequencing of the Sargasso Sea. *Science* (2004) 11:66–74.

White JR, Nagarajan N, Pop M (2009) Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. *PLoS Comput Biol* 5(4): e1000352. doi:10.1371/journal.pcbi.1000352

Willner, D, R Vega Thurber, F Rohwer (2009) Metagenomic signatures of 86 microbial and viral metagenomes. *Env Microbiol*. 11: 1752–1766