BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 19 2012, pages 2493–2501
doi:10.1093/bioinformatics/bts470

Systems biology

Advance Access publication July 26, 2012

Meta-Storms: efficient search for similar microbial communities based on a novel indexing scheme and similarity score for metagenomic data

Xiaoquan Su, Jian Xu and Kang Ning*

Shandong Key Laboratory of Energy Genetics, CAS Key Laboratory of Biofuels and BioEnergy Genome Center, Qingdao Institute of Bioenergy and Bioprocess Technology, Chinese Academy of Sciences, Qingdao 266101, Shandong Province, People's Republic of China

Associate Editor: Michael Brudno

ABSTRACT

Background: It has long been intriguing scientists to effectively compare different microbial communities (also referred as 'metagenomic samples' here) in a large scale: given a set of unknown samples, find similar metagenomic samples from a large repository and examine how similar these samples are. With the current metagenomic samples accumulated, it is possible to build a database of metagenomic samples of interests. Any metagenomic samples could then be searched against this database to find the most similar metagenomic sample(s). However, on one hand, current databases with a large number of metagenomic samples mostly serve as data repositories that offer few functionalities for analysis; and on the other hand, methods to measure the similarity of metagenomic data work well only for small set of samples by pairwise comparison. It is not yet clear, how to efficiently search for metagenomic samples against a large metagenomic database.

Results: In this study, we have proposed a novel method, Meta-Storms, that could systematically and efficiently organize and search metagenomic data. It includes the following components: (i) creating a database of metagenomic samples based on their taxonomical annotations, (ii) efficient indexing of samples in the database based on a hierarchical taxonomy indexing strategy, (iii) searching for a metagenomic sample against the database by a fast scoring function based on quantitative phylogeny and (iv) managing database by index export, index import, data insertion, data deletion and database merging. We have collected more than 1300 metagenomic data from the public domain and in-house facilities, and tested the Meta-Storms method on these datasets. Our experimental results show that Meta-Storms is capable of database creation and effective searching for a large number of metagenomic samples, and it could achieve similar accuracies compared with the current popular significance testing-based methods.

Conclusion: Meta-Storms method would serve as a suitable database management and search system to quickly identify similar metagenomic samples from a large pool of samples.

Contact: ningkang@qibebt.ac.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 16, 2012; revised on July 9, 2012; accepted on July 21, 2012

1 INTRODUCTION

Microbes are everywhere around us on the planet, and the total number of microbial cells on earth is huge: a rough estimation of their number is 10^{30} (Proctor, 1994). Microbes usually live in communities, and each of these communities has different community structures and functions. As such, microbial communities would serve as the largest reservoir of genes and genetic functions for a vast number of applications in 'bio'-related disciplines, including biomedicine in healthcare, bioenergy, bioremediation and biodefense (National Research Council (U.S.). Committee on Metagenomics: Challenges and Functional Applications. and National Academies Press (U.S.), 2007).

Because >90% of the strains in a microbial community could not be isolated and cultivated (Jurkowski *et al.*, 2007), the metagenomic methods have been used to analyze the microbial community as a whole. Understanding the taxonomical structure of a microbial community (alpha diversity) and the differences in taxa among microbial communities (beta diversity) have been two of the most important problems in metagenomic research. In contrast to alpha diversity, which measures how many kinds of microorganisms are there in a single community, beta diversity measures how community membership varies over time and space, and is especially important for finding the complex relationships among a large numbers of samples. Understanding the beta diversity is critical for studying microbial ecology. For example, Human Microbiome Projects (Turnbaugh *et al.*, 2006) and related efforts to study microbial communities occupying various human body habitats have shown a surprising amount of diversity among individuals in skin (Fierer *et al.*, 2008), gut (Turnbaugh *et al.*, 2009) and mouth ecosystems (Yang *et al.*, 2012). Furthermore, the microbial communities would differ significantly even for those from types of similar environment (Muegge *et al.*, 2011).

Next-generation sequencing techniques have enabled the fast profiling of a large number of metagenomic data. Thus, a rapidly increasing number of metagenomic profiles of microbial communities have been archived in public repositories and research labs around the world. As such, it is becoming more and more important to compare microbial communities in large scale.

*To whom correspondence should be addressed.

1.1 Comparison of microbial communities

A number of methods have been proposed for class discovery and comparison of different metagenomic samples. For comparison of multiple metagenomic samples, current researchers would adopt two different approaches: taxon-based (using overlap in lists of species, genera, OTUs and so on) or phylogenetic (using overlaps on a phylogenetic tree), both of which were based on the taxonomical information from the samples.

In the first approach, many recent pyro-sequencing studies have been developed to compare samples (Huber *et al*., 2007; Roesch *et al*., 2007; Sogin *et al*., 2006). MEGAN (Huson *et al*., 2007) is a metagenomic analysis tool with recent additions for phylogenetic comparisons (Mitra *et al*., 2010) and statistical analyses (Mitra *et al*., 2009). MEGAN, however, can only compare single pairs of metagenomic samples based on taxonomy, as is also the case with STAMP (Parks and Beiko, 2010), which does introduce a concept of 'biological relevance' in the form of confidence intervals. Other methods, such as MG-RAST (Meyer *et al*., 2008), ShotgunFunctionalizeR (Kristiansson *et al*., 2009), mothur (Schloss *et al*., 2009) and METAREP (Goll *et al*., 2010), all process metagenomic data using standard statistical tests (mainly *t*-tests with some modifications). Those analyses are clear and simple, yet would turn out not to be very accurate (Hamady and Knight, 2009). In the second approach, some researchers have used phylogeny-based methods to compare samples (Hamady *et al*., 2010; Lozupone and Knight, 2005). Phylogenetic beta diversity measures, such as UniFrac and Fast UniFrac (Hamady *et al*., 2010; Lozupone and Knight, 2005), are specifically important because, unlike taxon-based measures (Huson *et al*., 2007), they utilize the similarities and differences among species (Graham and Fine, 2008). This additional information makes phylogenetic beta diversity measures more effective at showing ecological patterns than taxon-based methods (Lozupone *et al*., 2008). Therefore, considerable insight has been gained from applying phylogenetic beta diversity methods to microbes in different environments. The number of samples and sample sizes (e.g. currently the Fast UniFrac online upload restriction is 200 samples) was the current limiting factors for the extension of this approach on the rapidly increasing scale of metagenomic experiments, thus rendering large-scale (e.g. more than 1000 samples) comparison of metagenomic samples difficult.

1.2 Databases of microbial communities

As metagenomic samples have been rapidly accumulated, it is natural to facilitate the comparison of microbial communities: create a database system with large number of metagenomic samples, and then given a set of samples as queries, find if there are similar metagenomic samples in the database. In other words, 'what does my sample like, how is it similar and different from any known samples?'

However, current metagenomic databases, such as MG-RAST (Meyer *et al*., 2008) and CAMERA2 (http://camera.calit2.net/), mainly serve as the data repositories, with only some modules for basic alpha and beta diversity analysis but neither full support of comparison nor search functions. Recently, a new database, MeganDB (http://www.megan-db.org/megan-db/home/), has been developed with public metagenomic samples pre-processed and archived, against which query metagenomic samples could be searched. This is an increasingly viable approach for searching for similar metagenomic samples, yet the current approach based on pairwise taxonomical comparison of metagenomic samples still lacks of index supporting for high efficient query.

1.3 Efficient search for similar metagenomic communities in a database

To facilitate the metagenomic sample comparison and search, it is crucial to design an effective metagenomic database system, primarily based on database creation and search schemes. Such scheme would include key advantages such as efficient indexing of database and accurate search scores, based on which, (i) the metagenomic samples are not only the accumulation of raw data (Supplementary Fig. S1A), but a set of well-organized and indexed datasets (Supplementary Fig. S1C) and (ii) the focus of the 'comparison of different microbial communities' is not the traditional pair-wise comparison for all possible pairs (Supplementary Fig. S1B), but more of the query-based similarity analysis (Supplementary Fig. S1C).

In this work, we have designed a novel database indexing and searching method, Meta-Storms, based on taxonomical annotations and phylogenetic structure of metagenomic samples. The overall scheme of Meta-Storms is illustrated in Figure 1.

In the database creation part (Fig. 1A), all metagenomic samples are pre-computed to parse out their taxonomical structures (community structure), based on which Meta-Storms then generates the index of the database by four alternative methods: scanning sample files to automatically add each sample, inserting single index entry manually, merged with other databases and merged by indices of other databases. In the database query part (Fig. 1B), the query sample's taxonomical structure is also parsed out, which is then searched against the samples in the database by very fast indexing to retrieve candidate samples. After that, a quantitative phylogenetic-based similarity scoring function is used to compute the similarity between query and each of these candidates, and candidates with matching scores above thresholds are considered as matches to query. Meta-Storms also has database management part (Fig. 1 C) to efficiently

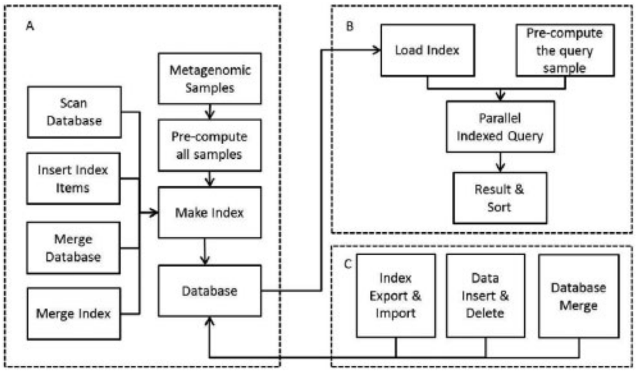


Fig. 1. Overall scheme of the Meta-Storms method. (**A**) Database and index creation part. (**B**) metagenomic sample query part. (**C**): Database management part.

manage the database, which is specifically useful when the database is huge.

2 METHODS

We have used C/C++ to implement the Meta-Storms. This software package includes database and index creation, metagenomic data query, database management and other functions which, together, could be beneficial for the maintenance and usage of metagenomic sample management, comparison and query.

2.1 Database and index creation

2.1.1 Database creation by taxonomical structure The taxonomical structure of a metagenomic sample is directly associated with the community structure of the corresponding microbial community. The database is composed of metagenomic datasets, all of which are first analyzed by a highly efficient taxonomy- and phylogeny-based metagenomic analysis pipeline, Parallel-META [(Su *et al*., 2012), version 1.3], to obtain the taxonomical structure. Then, a unique ID is assigned to each sample in the database.

2.1.2 Database indexing An index is created to organize the whole metagenomic database into hierarchical structure, mainly for the purpose of fast look-up for query samples. It is noticed that for many microbial communities (though might not be true for all), an intricate nature of its structure is that each microbial community is dominated by a few genomes (Hugenholtz and Tyson, 2008). Based on this general fact, we selected a set of phyla of each sample with high abundance (in the lexicographical order) as the 'index key'. To avoid arbitrary choice of indexing, each selected phylum should have significantly large proportion of reads. Based on previous experience, the abundance threshold of phyla for index key selection was decided to be 15%. We then built a Trie structure for high efficient indexing of the corresponding subdatasets of the samples, which have the same index key.

Figure 2A illustrates the hierarchical structure of the index with an example of a Trie. In the index, each character above the branches represents one phylum name in the index key, and the subdatasets are represented by the nodes, such as index key 'BC' is for samples in subdataset 8 (node 8 in Fig. 2A). Meta-Storms initialized the index by inserting all samples of the database into an empty Trie. Once the index is built, adding new samples in the database is simple: just insert them into the proper position of the Trie by their index keys. For instance, samples with index key 'BCE' (the selected phyla are B, C and E) should be inserted into subdataset 15 (node 15 of Fig. 2A) and sample with index key 'CDH' should be added to a new subdataset (Fig. 2B, node 19 in dashed line). Samples without index keys (i.e. without phylum of high abundance) are stored into the same subdataset named 'others' (node 4 of Fig. 2A).

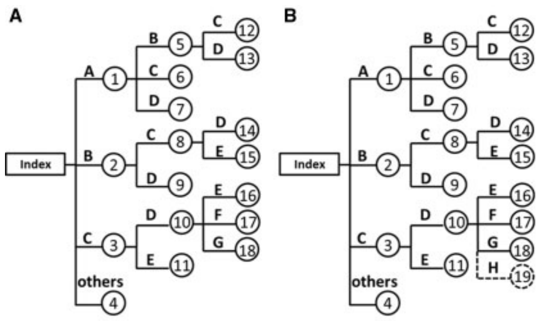


Fig. 2. Index structure based on Trie. (**A**) Characters above branches represent phylum names in index keys, and nodes represent the corresponding subdatasets. (**B**) is updated from **A** by adding a new subdataset (node 19) with index key 'CDH'.

2.2 Database query

2.2.1 Indexed query and parallel indexed query The 'query key' of the query metagenomic sample is parsed out by the same method as for the index key of samples in the database. Meta-Storms maps the query key in the Trie to the deepest (right-most) level, and fetches the subdataset in which samples have high abundance phyla as candidates, such as query key 'ABD' should be mapped to subdataset 13 (node 13 of Fig. 2A). When the query key cannot be fully mapped in the Trie, the least abundant phylum is removed from the query key, and this updated query key is tried again. If successful, the indexed query compete, otherwise this query key updating and mapping procedure continues until there is a hit, or all phyla are removed from the query key, which means that there is no hit for indexed query. For example, query key 'AEFD' (with abundance A > D > F > E) could be mapping to subdataset 7 (node 7 of Fig. 2A) by removing the two least abundance phyla F and E, and query key 'AGHI' (with abundance G > H > I > A) could be considered as 'not hit'.

Then, a quantitative phylogenetic-based scoring function (refer to Section 2.2.2 for details) is used to accurately search for the query sample by pairwise comparison with the candidate samples. As each candidate sample in the subdataset is independent, this comparison can be parallelized based on parallel programing. In this work, we map each pair of comparison with a thread, and on multi-core CPUs, these threads can be processed at the same time. All threads store their results in a shared-memory space which can be accessed by other threads.

2.2.2 Scoring function The scoring function is to compare the microbial communities' structure similarity by quantitative (i.e. the relative abundance of each species in a sample) and phylogenetic (i.e. the evolutionary distances between each species and its parent) (Hamady and Knight, 2009) calculation of their common component's proportion based on the phylogenetic relationship of species in two microbial communities.

(1) Preprocessing: initially, a common 'binary' phylogenetic tree of the two metagenomic samples with branch lengths and node weights is built: The branch length between each species and its ancestor represents the number of different nucleotides in every 100 nucleotides between their genome sequences, which reflects their degree of difference and the node weights are the proportion of reads for species of each sample that can be obtained by taxonomical structure analysis.

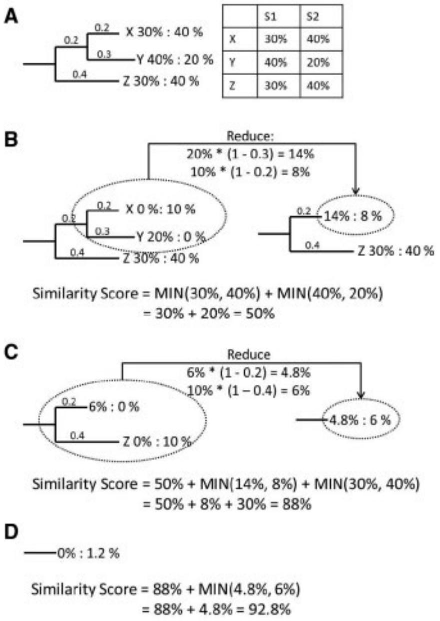
(2) Recursive scoring function: from each leaf node to the root, the similarity score for every single branch in the common phylogenetic tree is calculated by the recursive functions described as below:

Suppose for a single species X (represented by a leaf node in the phylogenetic tree), the proportions of this species in the two samples are X.P1 and X.P2, respectively. For example, the proportions of species X are 30% in sample S1 and 40% in sample S2 (node X in Fig. 3A). We define MIN(X) as the similarity score for a single species:

$$\mathrm{MIN}(X) = (X.P1 \leq X.P2)\ ?\ X.P1 : X.P2; \tag{1}$$

That is, the similarity score of one single species can be interpreted as the common proportion on this species of two samples, such as at node X, the score is MIN(30%, 40%) = 30%.

After extracting the common proportion, sample with larger proportion is remained of which the new proportion is $|X.P1 - X.P2|$ (node X in Fig. 3B). Samples at different nodes in the phylogenetic tree (node X and nod Y in Fig. 3B) cannot be compared directly, but should be reduced to their common ancestor X. Ancestor (if X is not root) by being multiplied



	S1	S2
X	30%	40%
Y	40%	20%
Z	30%	40%

Fig. 3. An example of how scoring function works. Here, we have two metagenomic samples S1 and S2, their common binary phylogenetic tree with three species X, Y, Z and the proportion of each species (**A**). The first step is to get the similarity score from leaf node of X and Y. Then the remaining component of X and Y are reduced to their common ancestor by being multiplied by 1-Dist and continue the comparison at both the ancestor node and leaf node Z (**B**). The remaining component of ancestor node of X and Y and leaf node Z are recursively reduced to the root node and (**C**) the overall similarity are got after the comparison at root. Finally, (**D**) the overall similarity of these two samples is 92.8%.

by the factor of 1-Dist ('Dist' is the phylogenetic distance between this species and its ancestor), which indicates how much component they shared with their ancestor. This step is referred to as the 'Reduce' function.

$$
\begin{aligned}
&\mathrm{Reduce(X)}\{\\
&\qquad \mathrm{M = MIN(X);}\\
&\qquad \mathrm{X.Ancestor.P1{+}= (X.P1 - M) * (1 - Dist);}\\
&\qquad \mathrm{X.Ancestor.P2{+}= (X.P2 - M) * (1 - Dist);}\\
&\qquad \}
\end{aligned}
\tag{2}
$$

We further define that for an internal node X, its two child nodes are X.Left and X.Right. Then the overall similarly score of one whole branch in the phylogenetic tree can be calculated recursively by this function:

$$
\mathrm{GetSimilarity(X)} =
\begin{cases}
\mathrm{MIN(X);}\ \mathrm{Reduce(X);} & \textit{If X is a Leaf Node}\\[1em]
\mathrm{GetSimilarity(X.Left)}\\
\quad +\mathrm{GetSimilarity(X.Right)}\\
\quad +\mathrm{MIN(X);}\ \mathrm{Reduce(X);} & \textit{If X is an Internal Node}
\end{cases}
\tag{3}
$$

All branches are calculated in the same way and at the root of the phylogenetic tree, we can get the overall similarity score of two metagenomic samples. For example, the overall similarity score for comparing S1 and S2 is 92.8% (Fig. 3D).

(3) The properties of scoring function: the scoring function calculates the overall proportion of common components of two metagenomic samples in all phylogenetic level by bottom–up analysis of the weighted common phylogenetic tree. This calculation based on weighted sum take into consideration both (a) the evolutionary distances of the species in each sample that are between 0 and 1 and (b) the different normalized abundances of these species in different samples represented by leaf nodes in the common phylogenetic tree, which also make the similarity score of two samples is always between 0% and 100%. Therefore, it realistically reflects the (weighted) structure differences between two different metagenomic samples. As such, the query results could be ranked directly by their similarity values compared with query sample. Moreover, since it depends on only a hierarchical annotation structure, such a scoring function could measure any kind of beta diversity for two samples.

(4) Data structure and time complexity: the common binary phylogenetic tree is stored in Newick format (Cayley *et al.*, 2000) as a character string, and the proportion of read for species in two samples are stored in a hash table for constant-time access. As the similarity score calculation of two metagenomic samples can be completed by scanning the character string of the Newick-formatted phylogenetic tree only once, the algorithm of scoring function is in O(n) time, where n is the total number of nodes in the common phylogenetic tree. In our experiments, more than 10 000 times of execution of scoring function could be finished within minute (refer to Fig. 7 in Section 3.6 for details).

(5) P-value calculation: to assess the significance of the similarity between two samples, we further defined a statistically meaningful *P*-value between two samples: the proportions of species in each sample are randomly permuted for 1000 times, and random similarity scores based on new abundance are calculated by the scoring function, then *P*-value is the rank of the original real similarity score among the permutated scores. Since two samples are significantly similar, if the real similarity score is significantly higher than would be expected when the sequences were randomly distributed between the two samples (Lozupone and Knight, 2005), if the *P*-value is equal to or lower than an experimental threshold of 10% (refer to Section 3.5 for details), then the two metagenomic samples can be considered as significantly similar.

2.2.3 *Exhaustive search*

Meta-Storms also provide the exhaustive search, which is based on comparing the query sample to all samples in the database by our scoring function to get the best match scores (considered as 'golden standard'). As all samples are computed by this method in the database, exhaustive search is much slower than indexed query (refer to Section 3.3 for details).

2.2.4 *Index consistency*

The index consistency between indexed query and exhaustive search is calculated by checking the probability that results of these two methods to be equal, which can reflect the reliability that whether the indexed query can find out the same best-match result as the 'golden standard'.

We can give the Dominant Genomes Lemma to predict the index consistency. In the database, a sample is considered to have dominant phyla if and only if the total proportion of the phyla which are selected in the index key is >60%. Then, the Dominant Genomes Lemma is described as the following:

Dominant Genomes Lemma:
If D% samples in the database have dominant phyla, then the index consistency is at least D%.

2.3 Database management

Meta-Storms also has modules to efficiently manage the database, including index import and export, data insertion and deletion, database merging and so on, which would be especially useful when the database is huge. We have implemented database operations that include (Fig. 1 C): (i) index export, to export an index of a database for query after database creation; (ii) index import, to use an index on a database for efficient query; (iii) data insertion, to insert new metagenomic sample(s) into a database; (iv) data deletion, to remove metagenomic sample(s) from a database and (v) database merging, to merge two different metagenomic databases into one.

3 RESULTS AND DISCUSSIONS

3.1 Experiment settings

In this work, we performed seven experiments with five datasets to evaluate the index consistency, index efficiency, scoring function consistency, scoring function efficiency and so on of Meta-Storms. Tables 1 and 2 (refer to 'Dataset details' in Supplementary Material for more information and availability) show the information of our experiments and datasets. We also built a 'test database' with all 1363 metagenomic samples of dataset Database 1 (Table 2) for the first two experiments on indexing.

3.2 Indexing consistency evaluation

In this experiment, we used dataset Query 1 (Table 2, 80 samples in total) as query samples to assess the consistency of indexing in the 'test database'. With 1179 of 1383 samples having dominant phyla, the index consistency of the test database would be >86.50%, as predicted by the 'Dominant Genomes Lemma' (refer to Section 2.2.4). We also tracked the 184 samples that do not have dominant phyla. Among them, 99 samples were from 'Dog Fleas' and 27 samples were from 'Snake Gut' of which the unclassified reads took a percentage of >20%. This made the average of their index keys to be 45.78%. For the other 58 samples, the average proportion of their index keys was 54.82%, which is very close to 60%. Therefore, it is conjectured that the major reason for samples without dominant phyla is largely due to the incompleteness of the taxonomy annotation of these samples, based on current taxa database.

In this experiment, as 72 indexed query results of 80 query samples (Fig. 4A) were the same as those based on exhaustive search, the index consistency was considered to be 90.0%, which verified the Dominant Genomes Lemma. We also checked the rank of indexed query results in the top five best-match results of exhaustive search (Fig. 4A). Results have shown that all samples were in top three matches. Such high ranks of indexed query

Table 1. Description of experiment setting

Experiment	Description
Index consistency	Index consistency comparison with exhaustive search based on dataset Database 1 and Query 1.
Index efficiency	Index efficiency in comparison with exhaustive search based on dataset Database 1 and Query 1.
Scoring function Consistency	Scoring function consistency with weighted Fast UniFrac based on 200 samples (dataset Select)
Scoring Function *P*-value Analysis	Scoring function p-value based on permutation with 200 samples (dataset **Select**)
Scoring function efficiency	Scoring function speed test with 200 samples (dataset Select)
Application of Meta-Storms on human microbiome	Application of Meta-Storms on human microbiome with 60 database samples (dataset Database 2) and 120 query samples (dataset Query 2).
Functional comparison	Functional comparison of metagenomic samples with 36 samples from mouse gut.

Refer to **Table 2** for details of these datasets.

Table 2. Description of datasets

Name	Description	Availability
Database 1	1363 samples randomly selected from 18 projects of MG-RAST, CAMERA2 and in-house (QIBEBT CAS). The test database is built based on all samples of Database 1.	Database 1 in Supplementary Materials
Query 1	80 samples randomly selected from the same source as dataset Database1 but all different from Database1	Query 1 in Supplementary materials
Select	200 samples randomly select from dataset Database 1	Select in Supplementary Materials
Database 2	60 samples randomly selected from three different sites in human gut, palm skin and mouth of two individuals with different genders (Caporaso *et al.*, 2011)	Database 2 in Supplementary Materials
Query 2	120 samples randomly selected from the same source as dataset Database 2 but all different from Database 2	Query 2 in Supplementary Materials

All of these experiments were performed on a desktop workstation with a CPU of dual Intel Xeon X5650 with 12 cores and clock 2.66 GHz, 72 GB DDR3 RDIMM and 4 TB RAID 0 Disk.

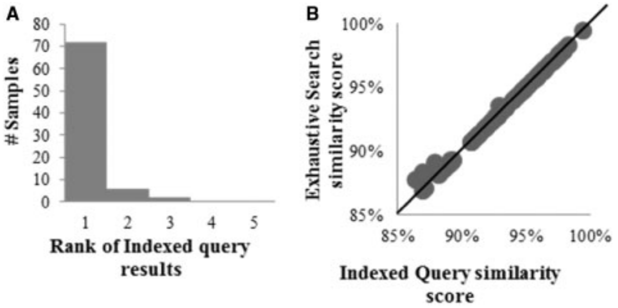


Fig. 4. Index consistence. (**A**) The rank of indexed query results ordered by exhaustive search similarity scores. (**B**) The comparison of best-match results similarity scores between indexed query and exhaustive search.

results compared with 'golden standard' partially indicated that the results between the two approaches (indexed query and exhaustive search) would be significantly similar. This is also verified by direct comparison: the average difference of best-match results between them are only 0.19% (Fig. 4B). Here 'average difference' was calculated by the formula $(\sum_{i=1}^{80} Ei - Ii)/80$. In which, Ei represents the similarity score of query sample i and its best-match sample by exhaustive search and Ii represents the similarity score of query sample i and its best-match sample by indexed query.

3.3 Indexing efficiency evaluation

To evaluate the efficiency of indexing, we also compared the running time of indexed query and exhaustive search with the test query samples in the last experiment (80 samples of dataset Query 1 as queries against the test database). This is performed for four times to get the average performance.

For indexed query, as all samples in the subdataset represented by the query key of query sample are scanned to compute the similarity score with query sample, the running time of indexed query mostly depends on the size of the corresponding subdataset (number of samples, Supplementary Fig. S2A) of the database. Additionally, there is no apparent correlation between query sample size (number of reads) and the processing speed by either indexing or exhaustive search (Supplementary Fig. S2B). In this experiment, the indexed query had achieved an average speed of 12.02 compared with exhaustive search.

3.4 Scoring function analysis

The scoring function module is one of the key components of the Meta-Storms method. To test the reliability of scoring function, we calculated the pairwise similarity score of samples in our database and compared with weighted Fast UniFrac (Hamady et al., 2010) distance, which has different principles for calculation of pairwise score for metagenomic samples. Limited by the Fast UniFrac online upload restriction of 200 samples, we randomly selected 200 samples from six different kinds of source (dataset Select, Table 2) from 1363 samples of dataset Database 1.

For 200 samples, there were $200 * 199/2 = 19900$ sample pairs, each has a pairwise distance. In Figure 5, distance values (dots in Fig. 5) were distributing in the diagonal area, meaning that results of the two methods (our scoring function distance and Fast UniFrac distance) are very consistent (Spearman correlation of 0.915) in measuring pairwise similarity (distance) between samples.

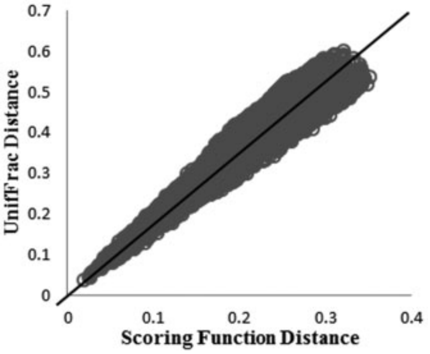


Fig. 5. Comparison of sample distance of Scoring Function and distance of weighted Fast UniFrac. Scoring Function Distance was calculated by (1- Similarity Score).

Having high correlation of pairwise distances with weighted Fast UniFrac, Meta-Storms has three major advantages/differences compared with other methods: (i) Meta-Storms is fully open-source, which could be installed locally as a stand-alone version with no sample number restriction (advantage in software configuration); (ii) Meta-Storms can take input of the 16 s rRNA pyro-sequencing data or Whole Genome Sequencing (WGS) data, so it is independent of any other method to retrieve the community structure of the metagenomic samples (advantage in system integrity) and (iii) Meta-Storms system is a metagenomic data management and search system, rather than a multiple sample comparison method (difference in design principle). Thus, Meta-Storms' scoring function serves as one of the several important components for its advantages and uniqueness.

3.5 *P*-value analysis

In this part, to test *P*-values, we have used the *P*-value of 19 900 sample pairs of 200 samples of dataset 'Select' (Table 2). To avoid inaccuracy caused by random sampling, permutation of each sample pair was repeated 1000 times.

The resultant figure (Fig. 6) was divided into four areas by horizontal reference line indicating *P*-value 10% and vertical reference line indicating similarity score 85%. Sample pairs in area 1 are very few, and they have low similarity scores; Sample pairs in area 2 could be considered as not similar for high *P*-value but low similarity scores; In area 3, all sample pairs have pairwise similarity ≥85%, and *P*-value ≤10% (significant similar according to *P*-value). In area 4, although the *P*-values of sample pairs were >10%, they were still considered to be significant similar because, in-depth analyses have shown that for most of the sample pairs, the leaf nodes of each common phylogenetic tree were quite similar. For 1016 (Fig. 6 sample pairs marked by light-colored circles) of 1090 sample pairs in area 4, two samples in the pair were from the same source such as soil, ocean and human, meaning that the structure of two samples were similar, based on which most permutations of the proportion led to high similarity scores.

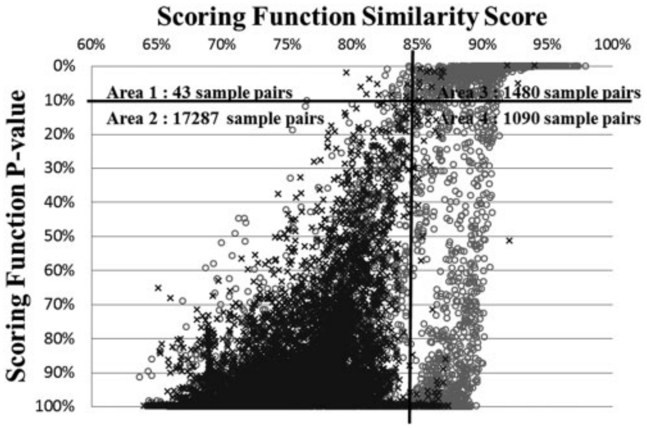
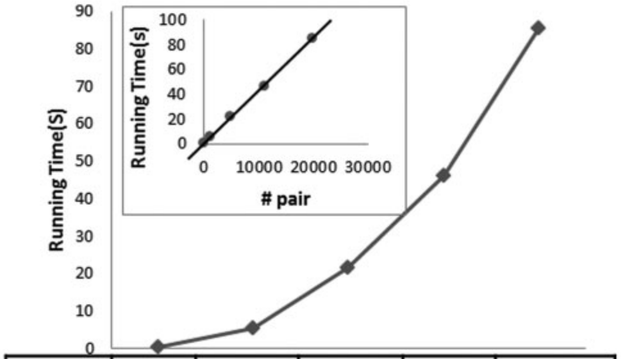


Fig. 6. Relationship between similarity score by scoring function and *P*-value by statistical analysis. Points marked by light-colored circles represent sample pairs with two samples from same sources, while others are marked by deep-colored crosses.



# Sample	10	50	100	150	200
# Pair	45	1225	4950	11175	19900
Runtime	0.324s	5.559s	21.610s	45.972s	1m25.426s

Fig. 7. Runtime of scoring function of computing distance matrices. A distance matrix with N samples contains $\mathrm{C}(N,\ 2)=N*(N-1)/2$ pairs, which means scoring function was executed $N*(N-1)/2$ times. Insert: This insert shows the relationship of running time and no. of pairs.

Based on these results, we could provide a recommended threshold of similarity score of 85%, which means that for two samples, if the similarity score is ≥85%, they could be considered as significantly similar. From the results in Figure 6, we can observe that in this test, 2570 sample pairs were considered significantly similar (1480 sample pairs in area 3 and 1090 sample pairs in area 4). Among them, 96.61% (1467 sample pairs in area 3 and 1016 sample pairs in area 4, refer to Supplementary Table S3 for details) were with two samples from the same source. On the other hand, there were 4404 sample pairs with two same-source samples in total but only 56.38% of them were considered as significantly similar. Therefore, if two samples are significantly similar, they tend to be from the same source, but samples from same source might not be significantly similar.

3.6 Scoring function efficiency evaluation

As stated previously, each similarity score for a pair of metagenomic samples could be calculated in a liner time. Here, we computed distance matrices of 10, 50, 100, 150 and 200 samples randomly selected from the dataset ‘Database 1’ (Table 2) by scoring function to test its speed. Results (sub-graph on left top of Fig. 7) have shown that the running time had a linear relationship with the number of sample pairs. For 200 samples, the scoring function had been executed for 19 900 times and could be finished in 1.5 min (Fig. 7). Thus, the running time is quite fast, given the current scale of metagenomic samples. We have also tested Meta-Storms system on a collection of more than 10 000 metagenomic samples (data not shown), and the distance matrix could be computed within several hours. As far as we know, Meta-Storms is one of the first systems that could handle such a large number of samples within a day’s time.

3.7 An application of Meta-Storms on human microbiome samples

In this section for application test, Meta-Storms is applied on the human microbiome from different sites and individuals. We used all 60 samples of dataset Database 2 (Table 2) from three different sites in human gut, palm skin and mouth of two individuals with different genders (Caporaso *et al.*, 2011) to build a metagenomic database. In this database, for each site on host, there are 10 different samples. For these 60 samples in the database, the similarity score of 1770 different sample pairs [C(60,2) = 1770] and the pairwise distance matrix are generated using our scoring function (Fig. 8A).

Based on this distance matrix, a phylogeny tree of samples from each body site in the database has been created using Phylip-fitch (Makarenkov, 2001) (Fig. 8B, right part). From which we could observe that samples were more diverse among different sites on same host, rather than between different individuals. This result is consistent with the research conclusion of Caporaso *et al.* (2011). It is also interesting to find out that the oral microbial community structures are more similar to those on palm skin than those in gut, probably due to the fact that *Staphylococcus aurous* is one of the dominant species both in oral conditions and on skin (Hamady and Knight, 2009; Kong, 2011).

Then, we used all 120 samples of dataset Query 2 (Table 2), which were collected from the same three sites of these two people but not in the database, as queries to find the best matches in the database (Fig. 8B, left part). Results had correctly identified the location from where the microbial community samples were collected, yet it was not very sensitive to host (Fig. 8B, left part).Among the queries, 99.2% (119 out of 120) were matched to the right sites, and 75.8% (91 out of 120) were perfectly matched to the right sites of the right host. This proved that the search accuracy of the Meta-Storms method is very high. As regard to speed, the whole analysis (from database creation to sample query) has taken <5 min.

3.8 Functional comparison of metagenomic samples

With the fast development of next-generation sequencing (NGS) techniques, some of the whole genome sequencing-based metagenomic sequencing projects had been conducted (Muegge

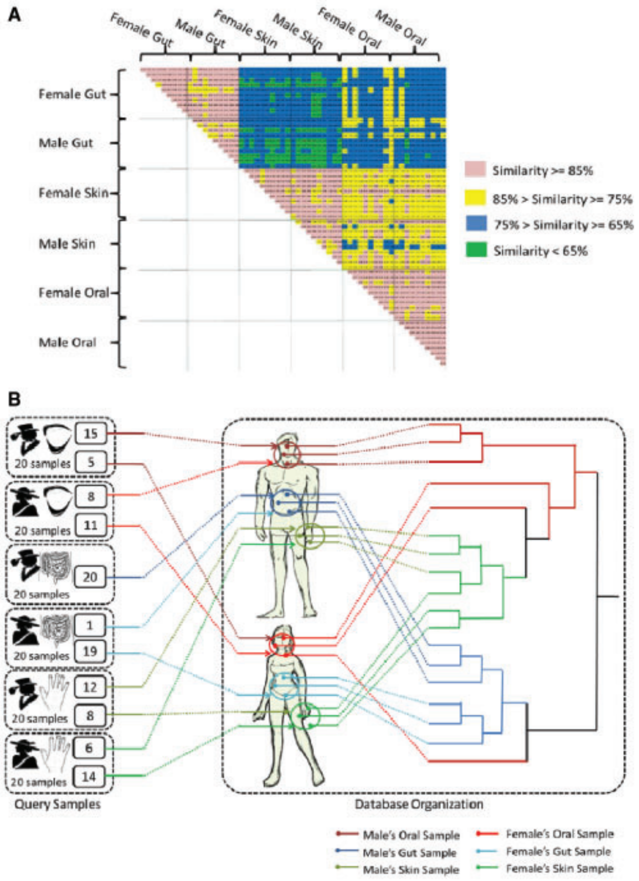


Fig. 8. Results of Meta-Storms on human microbiome samples. (**A**) Distance matrix of sample pairs in the human microbiome database. (**B**) Phylogenetic tree of human microbiome samples and query results, which were matched to the right body sites. Numbers on left part are samples that were matched to the body sites indicated by the corresponding arrows.

et al.,2011; Yang *et al.*, 2012). With these data available, it was now possible to profile all functions in a metagenomic sample. In such context, another measure similar to taxonomy-based beta diversity could be measured: hierarchical functional diversity [such as subsystems (Cayley and Holt, 1997)] of different samples. This measure could answer the fundamental question regarding the functional diversities among samples.

Previously, we have already mentioned the applicability of Meta-Storms on comparing functional diversity of metagenomic samples. In this part, we have tested Meta-Storms' ability to perform similarity search, which was conducted similar to the taxonomy-based search, except that indexing and searching were both based on functional hierarchy. We have selected functional annotation data from metagenomic samples obtained from a recent study on mouse gut microbiota (Faith *et al.*, 2011). We have selected the Escherichia coli samples from 36 mouse/diet period in total based on 13 mice [refer to (Faith *et al.*, 2011) for details]. Results have shown that the microbial communities' functional structures were not sensitive to different hosts, as samples from different hosts could be clustered together (refer to Supplementary Fig. S3 for more details). Yet these functional structures were quite sensitive to different diet periods: samples of third diet period were more conserved, whereas samples of fourth diet period were more diverse (refer to 'The functional diversity of different metagenomic samples' in Supplementary Material for more details). This might be explained by the microbial communities' exploitation or adaptation for the diet over time.

As shown in these applications on real metagenomic samples, because the WGS-based metagenomic studies have been more and more commonly used, it is now possible to have a metagenomic database with both taxonomical and functional profiles of metagenomic samples. The application of sample search and comparison methods such as Meta-Storms could help to discover the intricate relationship among samples, which would lead to great enhancement of metagenomic data analysis and data mining for a wider range of studies of microbial ecology.

4 CONCLUSION

With the fast accumulation of metagenomic samples and sequencing data, their comparison becomes very important to illustrate their similarities and differences. As such, the database search for metagenomic samples is becoming more and more important. However, current metagenomic data are not organized well (thus they are only data repositories, but not yet modernized and manageable database), and current metagenomic sample comparison methods are generally based on pairwise comparisons (thus, only comparable in small-scale, and difficult with large-scale analysis) without efficient index supporting. Therefore, we have designed the Meta-Storms system as a system for database creation, indexing and searching for similar metagenomic samples. Meta-Storms is not only a database builder and searcher, but a search-engine-based metagenomic sample comparison system that could organize the database well, and could perform quick and accurate search.

Although it is one of the first systems of its kind, Meta-Storms has already been proved to perform well in similarity search for a large numbers of samples. Its main advantages include: ability to handle large number of samples in an integrated system, fast indexing and candidate retrieval, accurate scoring function for comparison, etc.

First, Meta-Storms has advantages in system integrity: it can take input of the 16 s rRNA pyro-sequencing data or WGS data, thus is independent of any other method to retrieve the community structure of the metagenomic samples. Additionally, the database could be built based on more than 10 000 metagenomic samples, making it one of the first systems that could handle such a large number of samples.

Second, its index is simple, yet quite fast and accurate in clustering different samples. The fast speed comes out of the simple categorization by only using phyla with high abundance as the index key. Therefore, such indexing scheme would be useful, especially when there are many samples, for organizing samples in a metagenomic database.

Third, for sample comparison, it was noted that significance tests such as the *P*-test (Martin, 2002) and the Fast UniFrac replicated permutation become decreasingly useful as the depth of coverage and the number of samples increase (Hamady *et al.*, 2010). However, although our methods based on scoring function does not need complex computation, it is as accurate as those based on significance tests with fast speed. Therefore, our scoring function might be a better alternative to the traditional time-consuming significance test.

Functional comparison of different metagenomic samples have also shown the ability of Meta-Storms to accurately and efficiently identify microbial communities of similar functions, and also help to shed new light onto the functional diversity of the microbial communities. With the advance of WGS of metagenomic samples, it is anticipated that the functional profiling and comparison of metagenomic samples would become more and more important, for which Meta-Storms would be of greater importance.

In conclusion, Meta-Storms would provide key methods for metagenomic projects to facilitate the research in metagenomics, and more broadly, microbial communities, including Human Microbiome Projects, Earthmicrobiome Project, etc.

Current Meta-Storms method could be updated in index efficiency and consistence. First, current indexing strategy is suitable for searching samples with dominant genomes (which are always the case) against large database. For query samples without dominant phyla, or when samples in the database are quite similar to each other (e.g. from the same source), we recommend the exhaustive search for high reliability rather than indexed query. Second, the statistical model for the database search could be improved by integrating a mixture model. By these refinements, it is expected that the overall performance of Meta-Storms could improve further. All of these issues are expected to be addressed in our future work.

5 SOFTWARE AND DATA AVAILABILITY

The Meta-Storms software and metagenomic sample data mentioned in this article could be downloaded at http://www.computationalbioenergy.org/meta-storms.html.

ACKNOWLEDGEMENTS

We thank Professor Tao Jiang and Xingping Cui from University of California, Riverside (UCR) for insightful comments and discussions.

Funding: Chinese Academy of Sciences' e-Science grant (INFO-115-D01-Z006); Ministry of Science and Technology's 40 high-tech (863) grant (2009AA02Z310) and National Natural Science Foundation of China (NSFC) grant (61103167).

Conflict of interest: None declared.

REFERENCES

Caporaso,J.G. *et al.* (2011) Moving pictures of the human microbiome. *Genome Biol.*, **12**, R50.

Cayley,A.S. and Holt,R.D. (1997) The influence of audit on the diagnosis of occlusal caries. *Caries Res.*, **31**, 97–102.

Cayley,A.S. *et al.* (2000) Electropalatographic and cephalometric assessment of tongue function in open bite and non-open bite subjects. *Eur. J. Orthod.*, **22**, 463–474.

Faith,J.J. *et al.* (2011) Predicting a human gut microbiota's response to diet in gnotobiotic mice. *Science*, **333**, 101–104.

Fierer,N. *et al.* (2008) The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proc. Natl Acad. Sci. USA*, **105**, 17994–17999.

Goll,J. *et al.* (2010) METAREP: JCVI metagenomics reports–an open source tool for high-performance comparative metagenomics. *Bioinformatics*, **26**, 2631–2632.

Graham,C.H. and Fine,P.V. (2008) Phylogenetic beta diversity: linking ecological and evolutionary processes across space in time. *Ecol. Lett.*, **11**, 1265–1277.

Hamady,M. and Knight,R. (2009) Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res.*, **19**, 1141–1152.

Hamady,M. *et al.* (2010) Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J.*, **4**, 17–27.

Huber,J.A. *et al.* (2007) Microbial population structures in the deep marine biosphere. *Science*, **318**, 97–100.

Hugenholtz,P. and Tyson,G.W. (2008) Microbiology - Metagenomics. *Nature*, **455**, 481–483.

Huson,D.H. *et al.* (2007) MEGAN analysis of metagenomic data. *Genome Res.*, **17**, 377–386.

Jurkowski,A. *et al.* (2007) Metagenomics: a call for bringing a new science into the classroom (while it's still new). *CBE Life Sci. Educ.*, **6**, 260–265.

Kong,H.H. (2011) Skin microbiome: genomics-based insights into the diversity and role of skin microbes. *Trends Mol. Med.*, **17**, 320–328.

Kristiansson,E. *et al.* (2009) ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics*, **25**, 2737–2738.

Lozupone,C. and Knight,R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, **71**, 8228–8235.

Lozupone,C.A. *et al.* (2008) The convergence of carbohydrate active gene repertoires in human gut microbes. *Proc. Natl Acad. Sci. USA*, **105**, 15076–15081.

Makarenkov,V. (2001) T-REX: reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics*, **17**, 664–668.

Martin,A.P. (2002) Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Appl. Environ. Microbiol.*, **68**, 3673–3682.

Meyer,F. *et al.* (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.

Mitra,S. *et al.* (2010) Comparison of multiple metagenomes using phylogenetic networks based on ecological indices. *ISME J.*, **4**, 1236–1242.

Mitra,S. *et al.* (2009) Visual and statistical comparison of metagenomes. *Bioinformatics*, **25**, 1849–1855.

Muegge,B.D. *et al.* (2011) Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science*, **332**, 970–974.

National Research Council (U.S.); Committee on Metagenomics: Challenges and Functional Applications. and National Academies Press (U.S.). (2007). *The New Science of Metagenomics: Revealing the Secrets of our Microbial Planet*. National Academies Press, Washington, DC.

Parks,D.H. and Beiko,R.G. (2010) Identifying biologically relevant differences between metagenomic communities. *Bioinformatics*, **26**, 715–721.

Proctor,G.N. (1994) Mathematics of microbial plasmid instability and subsequent differential growth of plasmid-free and plasmid-containing cells, relevant to the analysis of experimental colony number data. *Plasmid*, **32**, 101–130.

Roesch,L.F. *et al.* (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J.*, **1**, 283–290.

Schloss,P.D. *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**, 7537–7541.

Sogin,M.L. *et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc. Natl Acad. Sci. USA*, **103**, 12115–12120.

Su,X. *et al.* (2012) Parallel-META: efficient metagenomic data analysis based on high-performance computation. *BMC Systems Biology*, **6**, S16.

Turnbaugh,P.J. *et al.* (2009) A core gut microbiome in obese and lean twins. *Nature*, **457**, 480–484.

Turnbaugh,P.J. *et al.* (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, **444**, 1027–1031.

Yang,F. *et al.* (2012) Saliva microbiomes distinguish caries-active from healthy human populations. *ISME J.*, **6**, 1–10.