Quantitative Biology 2015, 3(1): 4–18
DOI 10.1007/s40484-015-0040-3

RESEARCH ARTICLE

Application of Meta-Mesh on the analysis of microbial communities from human associated-habitats

Xiaoquan Su†, Xiaojun Wang†, Gongchao Jing, Shi Huang, Jian Xu and Kang Ning*

Bioinformatics Group of Single cell Center, Shandong Key Laboratory of Energy Genetics and CAS Key Laboratory of Biofuels, Qingdao Institute of Bioenergy and Bioprocess Technology, Chinese Academy of Sciences, Qingdao 266101, China
* Correspondence: ningkang@qibebt.ac.cn

Received December 2, 2014; Revised February 5, 2015; Accepted February 11, 2015

With the current fast accumulation of microbial community samples and related metagenomic sequencing data, data integration and analysis system is urgently needed for in-depth analysis of large number of metagenomic samples (also referred to as "microbial communities") of interest. Although several existing databases have collected a large number of metagenomic samples, they mostly serve as data repositories with crude annotations, and offer limited functionality for analysis. Moreover, the few available tools for comparative analysis in the literature could only support the comparison of a few pre-defined set of metagenomic samples. To facilitate comprehensive comparative analysis on large amount of diverse microbial community samples, we have designed a Meta-Mesh system for a variety of analyses including quantitative analysis of similarities among microbial communities and computation of the correlation between the meta-information of these samples. We have used Meta-Mesh for systematically and efficiently analyses on diverse sets of human associate-habitat microbial community samples. Results have shown that Meta-Mesh could serve well as an efficient data analysis platform for discovery of clusters, biomarker and other valuable biological information from a large pool of human microbial samples.

Keywords: metagenome; microbial community; data mining

INTRODUCTION

Microbes are ubiquitous on our planet, and it is well-known that the total number of microbial cells on earth is huge [1]. These organisms usually live in communities, and each of these communities has a different structure. 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, bioenergy, bioremediation, and biodefense [2].

Since over 90% of strains in a microbial community could not be isolated or cultivated [3], metagenomic methods have become popular for analysis of a microbial community as a whole. Such an approach enables the exploration of the relationships among microbes, their communities and habitats at the most fundamental genomic level. 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 [4], in which understanding beta diversity is especially critical for studying microbial communities' heterogenity. For example, the Human Microbiome Project [5] and related efforts to study microbial communities occupying various human body habitats have shown a surprising amount of diversity among individuals in skin [6,7], gut [8,9], and mouth ecosystems [10,11]. Furthermore, even microbial communities from similar types of environment might differ significantly [12].

Next-generation sequencing techniques have enabled fast profiling of large volumes of metagenomic samples. As a result, a rapidly increasing number of metagenomic

† These authors contributed equally to this work

profiles of microbial communities have been archived in public repositories and research labs around the world. Therefore, it is becoming more and more important to perform in-depth analysis for the valuable biological information hidden in large number of samples. Hence, a system that provides functionalities for data analysis and mining would be of significant value to a worldwide user-base from multiple disciplines.

Current metagenomic databases, such as MG-RAST [13] and CAMERA2 [14], serve mainly as data repositories, with neither comprehensive tools for comparative analysis nor capabilities for extensive search. A number of methods have been proposed for the comparison and clustering of different metagenomic samples. MEGAN [15] is a metagenomic analysis tool for taxonomical comparisons [16] and statistical analyses [17], which can only compare single pairs of metagenomic samples based on taxonomical annotations, as is also the case with STAMP [18]. ShotgunFunctionalizeR [19], Mothur [20] and METAREP [21] identify the differences between samples using standard statistical tests (mainly t-tests with some modifications). UniFrac [22] and Fast UniFrac [23] examine the similarities among species based on their overlaps in phylogenetic tree to discover ecological patterns.

In this work we introduce Meta-Mesh, an integrated metagenomic analysis system that includes a set of highly efficient tools, and verified the reliability and correctness based on simulated datasets constructed by known genomes. Then we applied Meta-Mesh on two studies with human associated-habitat samples from different body sites and health status: one study including gut samples from feces, oral samples from tongue, skin samples from left palm and right palm; another study with saliva samples from naturally occurring gingivitis, healthy gingivae and experimental gingivitis. Meta-Mesh quantitatively evaluated the similarity among samples, distinguished samples from different conditions by the taxonomical distributions and phylogenetic relationships, elucidated the key taxa led to the structure difference by biomarker analysis, and further calculated the correlation between the taxa distribution and the environmental meta-data (e.g., hosts, habitats, healthy conditions). Results have shown that Meta-Mesh would serve as an effective data analysis platform to quickly discovery the valuable biological information from a large pool of metagenomic samples.

RESULTS

In this work we have applied the Meta-Mesh in two studied of human-associated habitat microbial communities for data comparison and biomarker discovery.

Case study 1: Study on simulated microbial community samples for accuracy assessment and correctness verification

In this case we constructed two simulated metagenomic datasets with human oral related genomes and human gut related genomes, respectively. Each dataset contained 50 samples. The oral related genomes mainly included *Actinomyces odontolyticus*, *Campylobacter rectus*, *Fusobacterium periodonticum*, *Neisseria subflava*, *Porphyromonas gingivalis* and *Prevotella denticola* , which were all from HOMD database [24], and mixed with some other randomly selected oral genomes in very low ratio. The gut related genomes mainly included *Bacillus cereus*, *Clostridium beijerinckii, Deinococcus radiodurans*, *Escherichia coli* and *Helicobacter pylori*, which were all from the NCBI database (http://www.ncbi.nlm.nih.gov/), and also mixed with some other randomly selected gut genomes in very low ratio.

We parsed out the taxonomical structure of all samples by metagenomic structure analysis, and evaluated the error rate of the analysis by the Euler distance of the relative abundance between the calculated values and real values of all taxa. Suppose that N taxa (T_i, i = 0 to N-1) existed in a simulated sample, V_i was the calculated relative abundance value and V'_i was the real relative abundance value for T_i, then the error was evaluated by formula (1). Results in Figure 1A showed the average error rate of all samples was lower than 4% (2.86% and 3.25% for oral sample on phylum and genus level, 2.10% and 3.59% for gut sample on phylum and genus level), which indicated the high reliability of the metagenomic structure analysis in Meta-Mesh system.

$$\mathrm{E} = \sqrt{\frac{\sum_{i=0}^{N-1} (V_i - V'_i)^2}{N}}. \quad (1)$$

Then we computed the pair-wised similarity matrix for oral samples and gut sample separately by Meta-Mesh scoring function, and verified the results by the UniFrac [23] that had already been widely used for comparison among microbial communities. The strong correlation between results by Meta-Mesh scorning function and weighted UniFrac (Pearson coefficient R = 0.9928 for oral samples and R = 0.9731 for gut samples, Figure 1B and Figure 1C) illustrated that Meta-Mesh could precisely assess the quantitative difference among metagenomic samples. Based on the pair-wised similarity matrix we performed the clustering analysis by PCoA. It was obvious that samples from two datasets could be differentiate based on their similarity scores (Figure 1D), and the following biomarker analysis also explained the reason by the dissymmetrical distribution of the relative abundance on abundant taxa between two

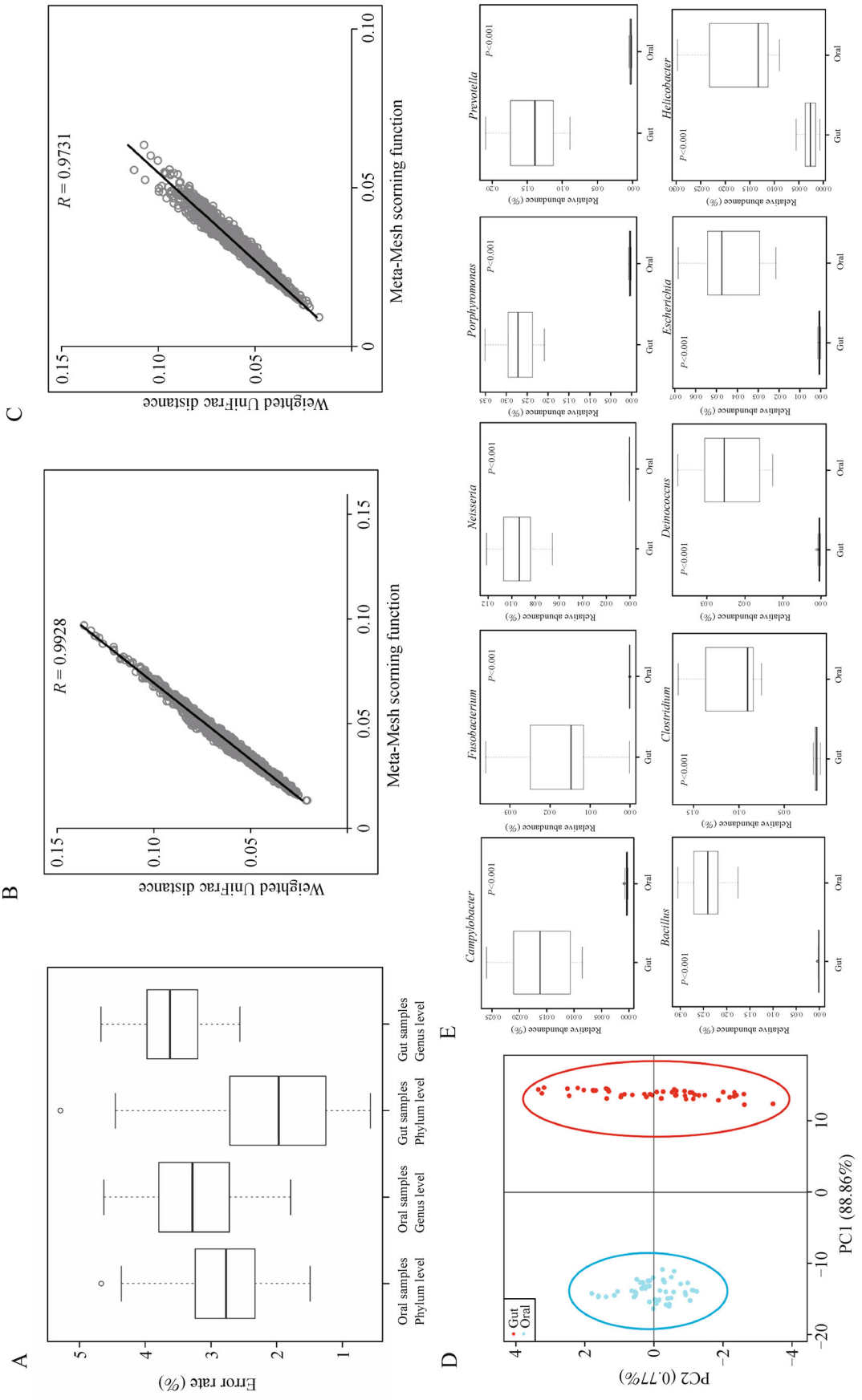

Figure 1. The accuracy assessment of Meta-Mesh based on simulated datasets. (A) Error rate of metagenomic structure analysis on phylum and genus level. (B) Correlation of Meta-Mesh scorming function and weight UniFrac with simulated oral samples. (C) Correlation of Meta-Mesh scorming function and weight UniFrac with simulated gut samples. (D) PCoA analysis results of simulated datasets. (E) Biomarker analysis results of simulated datasets.

datasets (Figure 1E), which was coincident to the different genomes used for simulation.

Case study 2: Study on microbial community samples from different human-habitats

In this case we used 1,758 human-associated habitat microbial community samples from four different body sites [25] (gut samples from feces, oral samples from tongue, skin samples from left palm and right palm, Table 1) of two individuals of opposite genders from the same family.

We firstly generated the pair-wised similarity matrix (Figure 2A) of all sample pairs among the 1,758 samples, and performed the clustering and PCoA analysis (Figure 2B) based on the similarity matrix. Results in Figure 2 have shown that samples from the same body site type could be clustered together. In Figure 2A we found that samples from the same site (skin, oral and gut) have higher similarity values, compared with those from different sites. Among them, the taxonomical structures oral samples were very similar with those for skin samples, while the taxonomical structure of gut samples were quite different from others. The PCoA analysis results (Figure 2B) supported the similarity patterns in Figure 2A. In Figure 2B, gut samples and oral samples were divided by PC1 very well, and clear separation of samples from oral and gut could also been observed by PC2, yet samples from left palm and right palm were mixed.

The following biomarker analysis (Figure 3) on the taxonomical structure analysis also elucidated the reasons that taxa *Bacteroidaceae* and *Clostridiaceae* were dominated in gut samples, among which *Bacteroidaceae* predominantly habitats in the alimentary canal or on mucous surfaces of warm-blooded animals [25], which could serve well as biomarkers for gut microbiota. And *Neisseriaceae* and *Prevotellaceae* were dominated in oral samples, among which species *Prevotellaceae* is among the most abundant microbes cultivable from the rumen and hind gut of cattle and sheep as well as human oral environment, where they help the breakdown of protein and carbohydrate in foods [25]. *Corynebacterineae* were dominated in skin samples.

Table 1. Information of human associated-habitat samples.

Type	Habitat	Sex	# of Samples
Gut	Feces	Female	130
Gut	Feces	Male	331
Oral	Tongue	Female	131
Oral	Tongue	Male	365
Skin	Left palm	Female	123
Skin	Left palm	Male	276
Skin	Right palm	Female	128
Skin	Right palm	Male	274

We further compared samples of each body site separately, and found that the microbial community structures in gut and oral environment of the two hosts had difference distribution in PCoA results (Figure 4A and Figure 4B). The reason was that the high abundance taxa in gut (*Clostridiaceae* and *Prophyromonadaceae*, Figure 5A) and oral (*Neisseriaceae* and *Streptococcaceae*, Figure 5B) showed significant different with Wilcoxon Rank-sum Test P-value < 0.001. However, skin samples from different hosts and palms had higher similarity to each other reflected in the PCoA results (Figure 4C and Figure 4D), which could also been explained by the biomarker analysis that their abundant taxa (*Corynebacterineae* and *Streptococcaceae,* Figure 5C and Figure 5D) had similar pattern between hosts and palms.

As regard to efficiency, we have evaluated Meta-Mesh for it running time on pair-wise similarity analysis, PCoA and biomarker analysis. Meta-Mesh would be able to compute (based on GPU) the pair-wise similarity values for 1,758 human-associated habitat microbial community samples within six minutes and following PCoA and biomarker analysis (based on all samples) within nine minutes (Figure 6), which in theory have outpaced the speed of metagenomic data generation by several folds. Therefore, the efficient computational engine in Meta-Mesh have enabled near real-time process of microbial communities.

Case study 3: Study on oral microbial communities of different health status

In this case study we applied the Meta-Mesh in the study of oral samples to discover the microbial community structure variation among different health status. 150 samples were collected from 50 hosts' saliva in 3 stages: naturally occurring gingivitis (I), healthy gingivae (B) and experimental gingivitis (E) by Huang, et al, 2014 [26]. The Mazza Gingival Index (MGI) of each stage for the host was recorded to reflect the gingival conditions from medical aspect.

All samples were compared to construct the similarity matrix (Figure 7A). Then we performed the clustering (Figure 7A) and PCoA analysis (Figure 7B). From the results we can observe that samples could be divided by the gingival status: healthy and gingivitis (both naturally and experimental occurred).

Then the in-depth biomarker analysis focused on discovering the taxa that caused the differences. From the results (Figure 8) we found that the distribution of taxa *Streptococcaceae*, *Actinomycineae*, *Micrococcineae* and *Prevotellaceae* had significant different between gingival

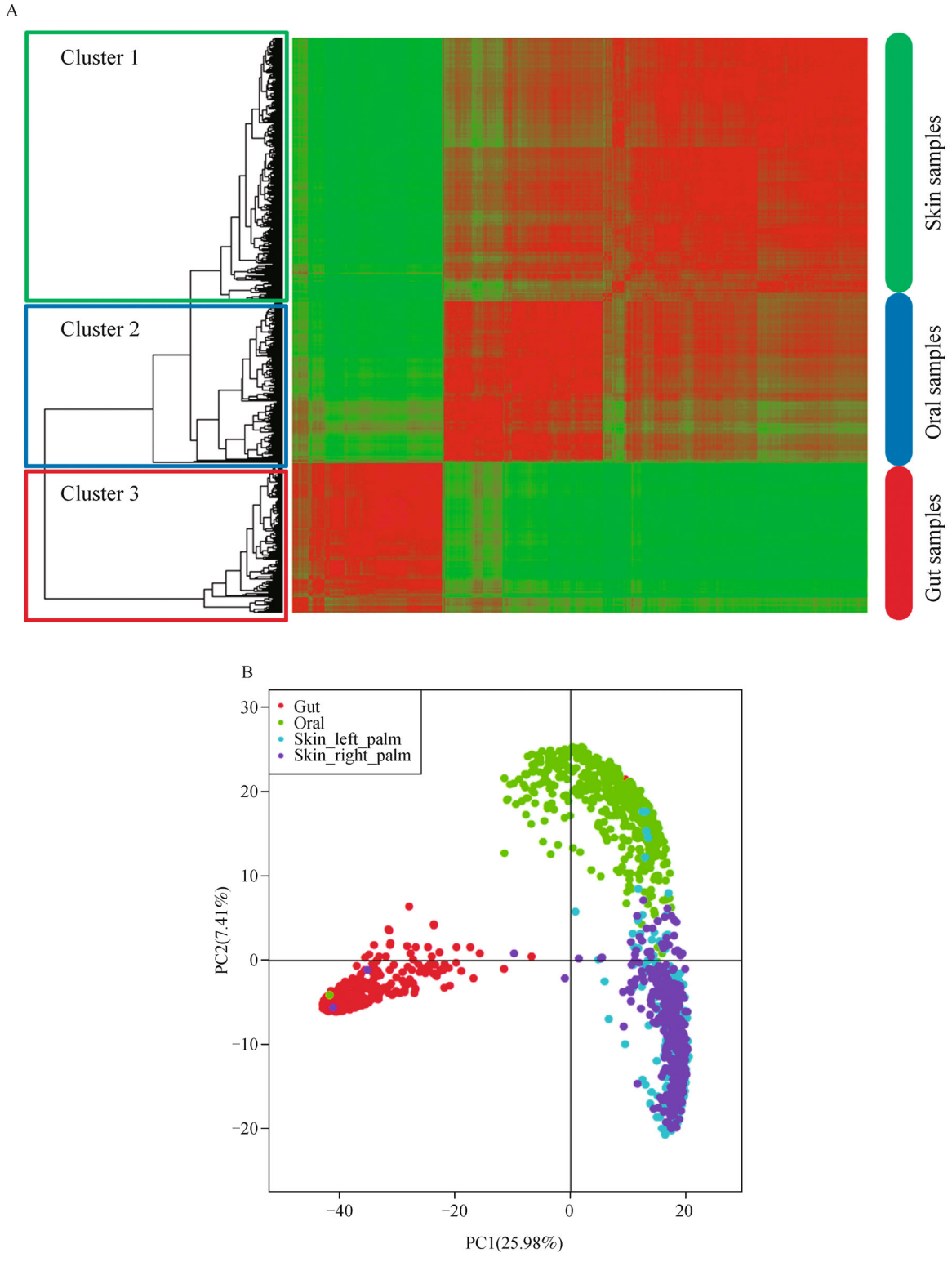

Figure 2. The clustering analysis results of human associated-habitat samples from different body sites based on the similarity matrix. (A) Similarity matrix and hierarchical clustering results. (B) PCoA analysis results of body sites.

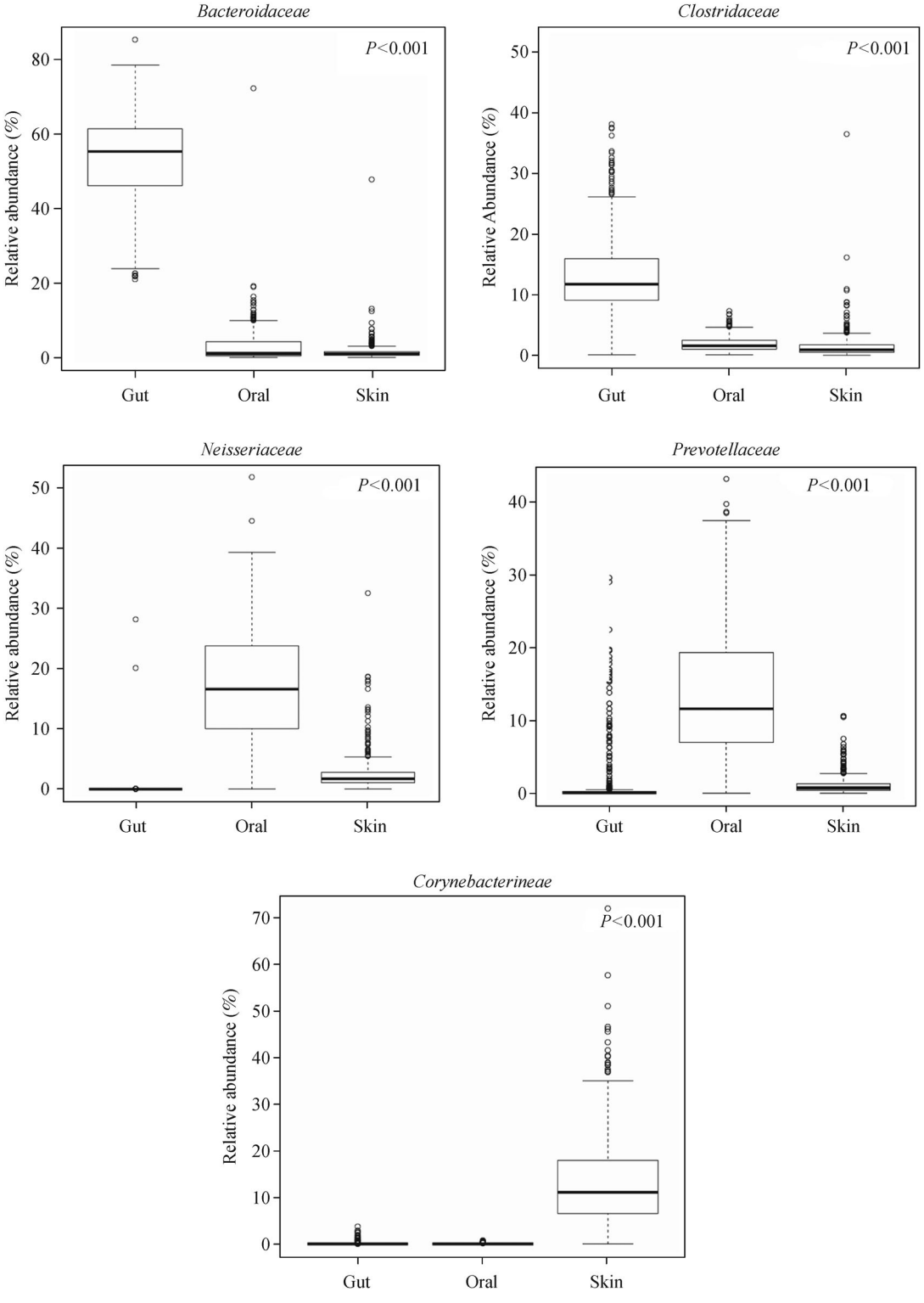

Figure 3. The biomarker analysis results of of human associated-habitat samples from different body sites. Boxplots have shown the relative abundance of each biomarker taxa in different body site types, and the rank-sun test *P*-value indicated the significance of their difference.

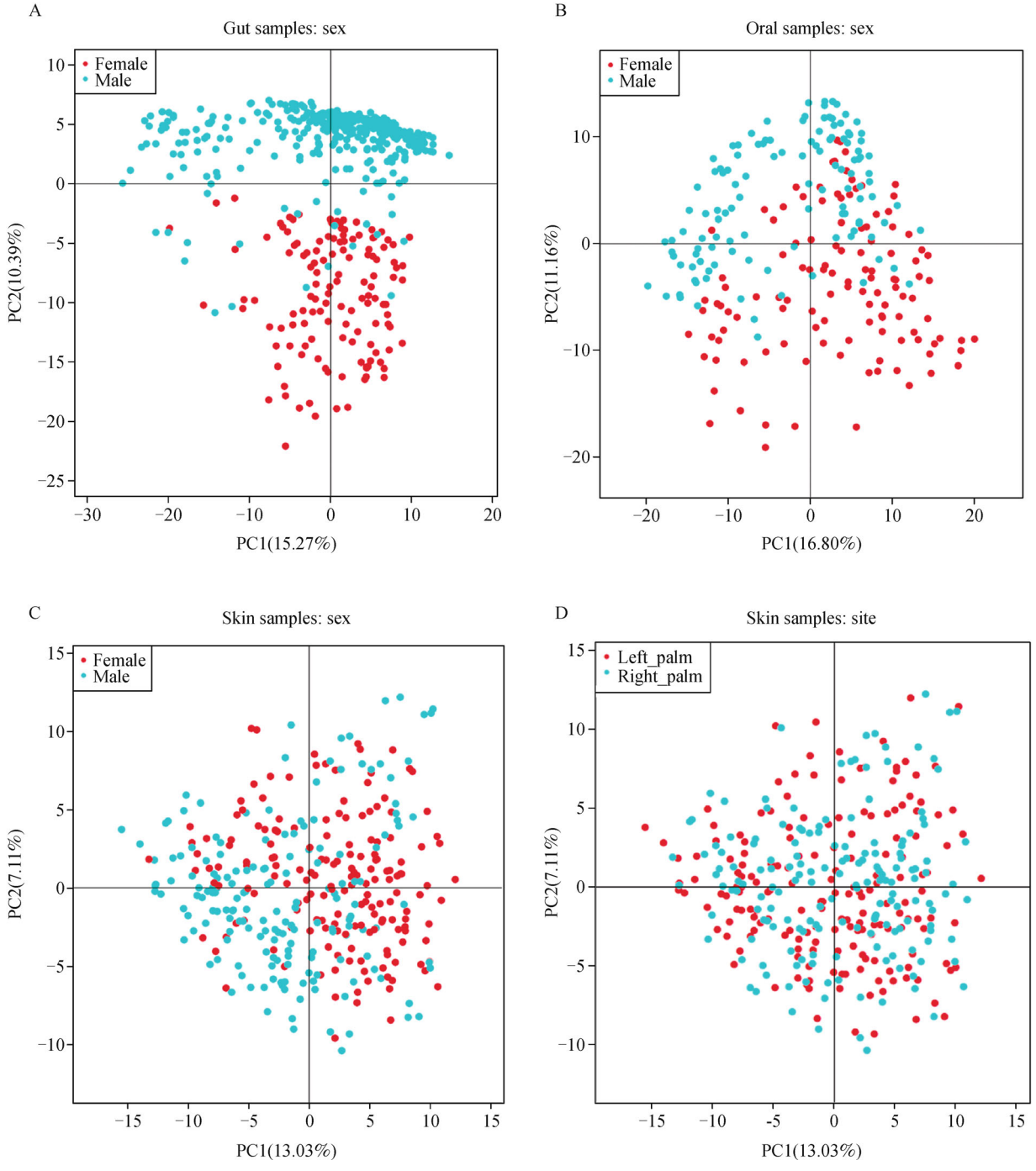

Figure 4. The PCoA analysis results of samples from different body sites. (A) Gut samples from different hosts. (B) Oral samples from different hosts. (C) Skin samples from different hosts and (D) Skin samples from different palms.

status (Wilcoxon Rank-sum Test P-value < 0.001). Among them, *Prevotellaceae* was abundant in gingivitis samples (I and E), while other three taxa were abundant in healthy samples (B), which has also been verified by the previous works in Huang, et al, 2014 [26].

We also measured the correlation between the taxa

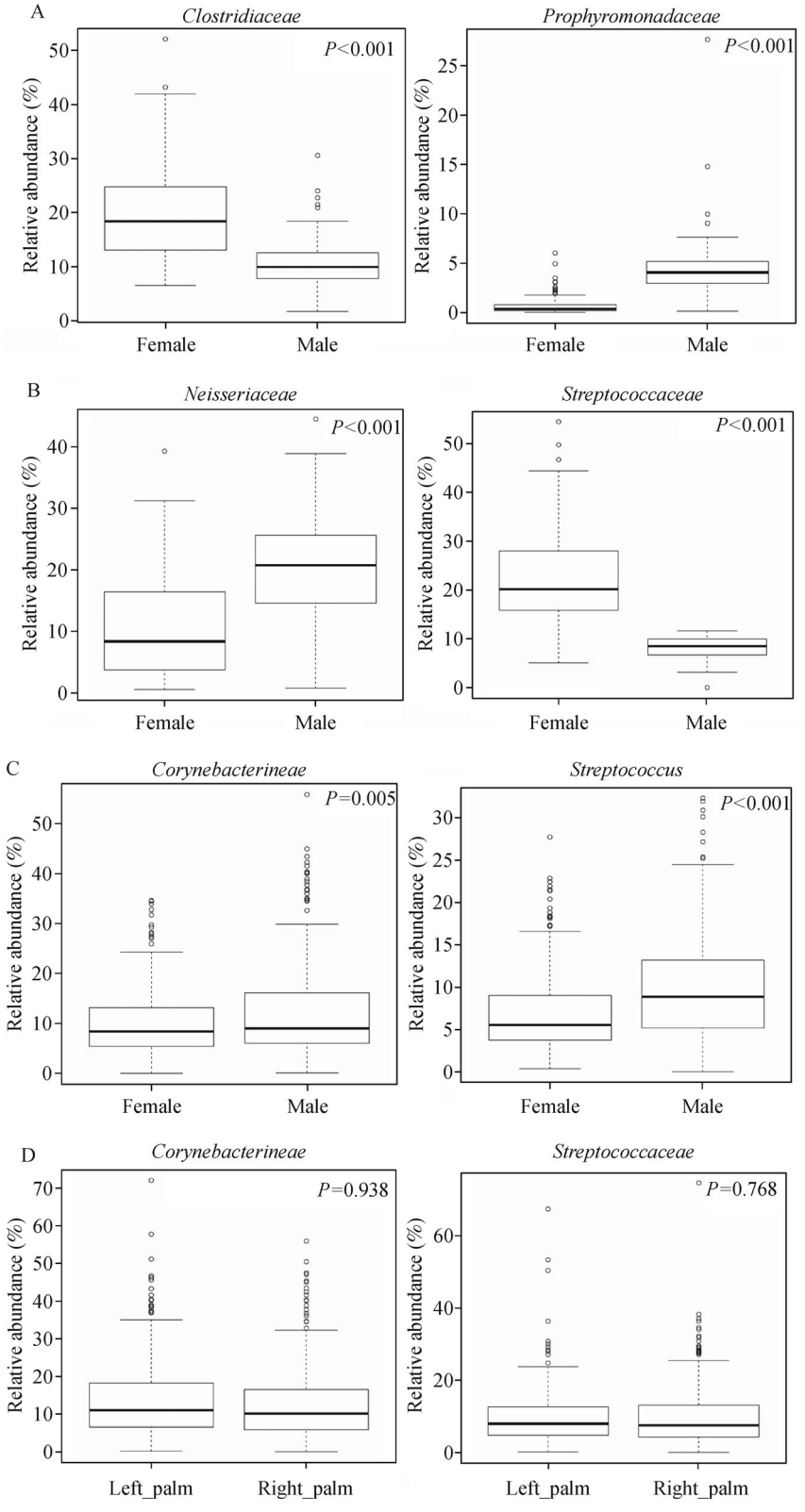

Figure 5. Biomarker analysis results of samples from each body site. (A) Gut samples from different hosts. (B) Oral samples from different hosts. (C) Skin samples from different hosts and (D) Skin samples from different palms.

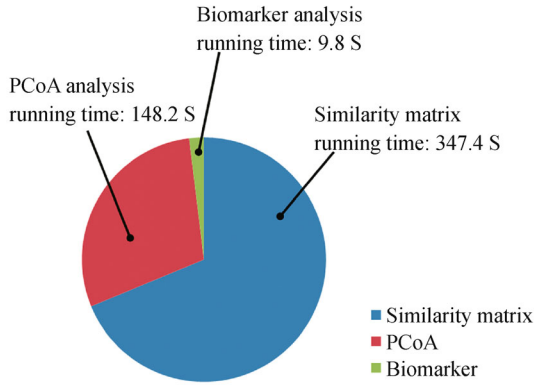

Figure 6. Efficiency analysis for processing speed of Meta-Mesh for 1,758 human-associated habitat microbial community samples.

pattern and the MGI and calculated the Pearson correlation coefficient (R). Results in Figure 9 illustrate the strong correlation of the abundant taxa to the MGI: the pattern of *Streptococcaceae* ($R = -0.806$), *Actinomycineae* ($R = -0.755$), *Micrococcineae* ($R = -0.879$) were negatively correlated to the MGI, while *Prevotellaceae* was positively correlated ($R = 0.810$) to the MGI. Interesting, the abundant yet dynamic species *Prevotellaceae,* which can server as biomarker of oral samples from gut and skin samples, could again serve as biomarkers to differentiate healthy and gingivitis samples within oral environment, indicating that it could serve as the "local weather forecaster" for oral health status.

For processing speed, again we observed that Meta-Mesh could analyze 150 oral microbial community samples, including pair-wise similarity analysis, PCoA and biomarker analysis (based on all samples), within only 30 seconds. This also showed the capability of Meta-Mesh for facilitating near real-time oral microbial community monitoring.

DISCUSSION AND CONCLUSIONS

With the rapid accumulation of metagenomic samples and sequencing data, methods for efficient comparison and database search for metagenomic samples are becoming increasingly important. However, a good integrated system for organizing metagenomic samples, providing large-scale comparison, as well as data-mining is still lacking and thus urgently needed. Current metagenomic sample comparison methods are generally based on pair-wise comparisons (which make them difficult for large scale analysis) without efficient data indexing that could support large-scale comparison and data-mining. Overcoming of this drawback has been attempted by our Meta-Mesh system, an integrated platform that could be used for sample taxonomical structure analysis, sample comparison and similarity matrix generation, sample clustering and biomarker identification, all based on an efficient computational engine.

The application of Meta-Mesh on comparison of different metagenomic samples has shown that it is able to accurately and efficiently cluster similar microbial communities and identify the biomarkers for these communities, as well as shedding new light on the functional diversity of the microbial communities. For example from thousands of oral microbial communities, Meta-Mesh could quickly differentiate communities from hosts with different health status with very high accuracy, and identify biomarkers that might contribute for such differences. Therefore, Meta-Mesh has proven its ability for efficient data-mining for large-scale metagenomic datasets.

With the advancements in Whole Genome Sequencing (WGS) of human related metagenomic samples, it is anticipated that the profiling and comparison of a large number of metagenomic samples would become more and more important, for which an integrated comparison system would be of great help. Currently bioinformatics analysis of metagenomic data has also entered the era of "big-data". Having this in its development philosophy, Meta-Mesh is well-positioned to provide key methods for human microbiome projects to facilitate research in metagenomics, and thus would be suitable for data-driven and in-depth data-mining for a variety of human microbial community datasets, including those from Human Microbiome Project, etc.

The Meta-Mesh system that we presented in this work will be continuously updated (http://www.meta-mesh.org/). Firstly, for the metagenomic sample database, automatic and manual curation will continuously provide high-quality metagenomic data, ensuring that users have access to accurate and consistently annotated experimental information coupled with manually verified sequence analysis results. Secondly, some text-mining techniques to further improve the efficiency in data curation would also be attempted for more accurate annotation. Thirdly, the current Meta-Mesh database is optimized for taxonomical annotation; with the fast accumulation of human related metagenomic data, a functional annotation-based system would be developed in the future.

METHODS

Meta-Mesh is an open-source system designed for efficient and integrated analysis of metagenomic samples that offer functions such as community structure analysis, sample comparison and metagenomic data mining for microbial community researches. The overall scheme of Meta-Mesh database is illustrated in Figure 10.

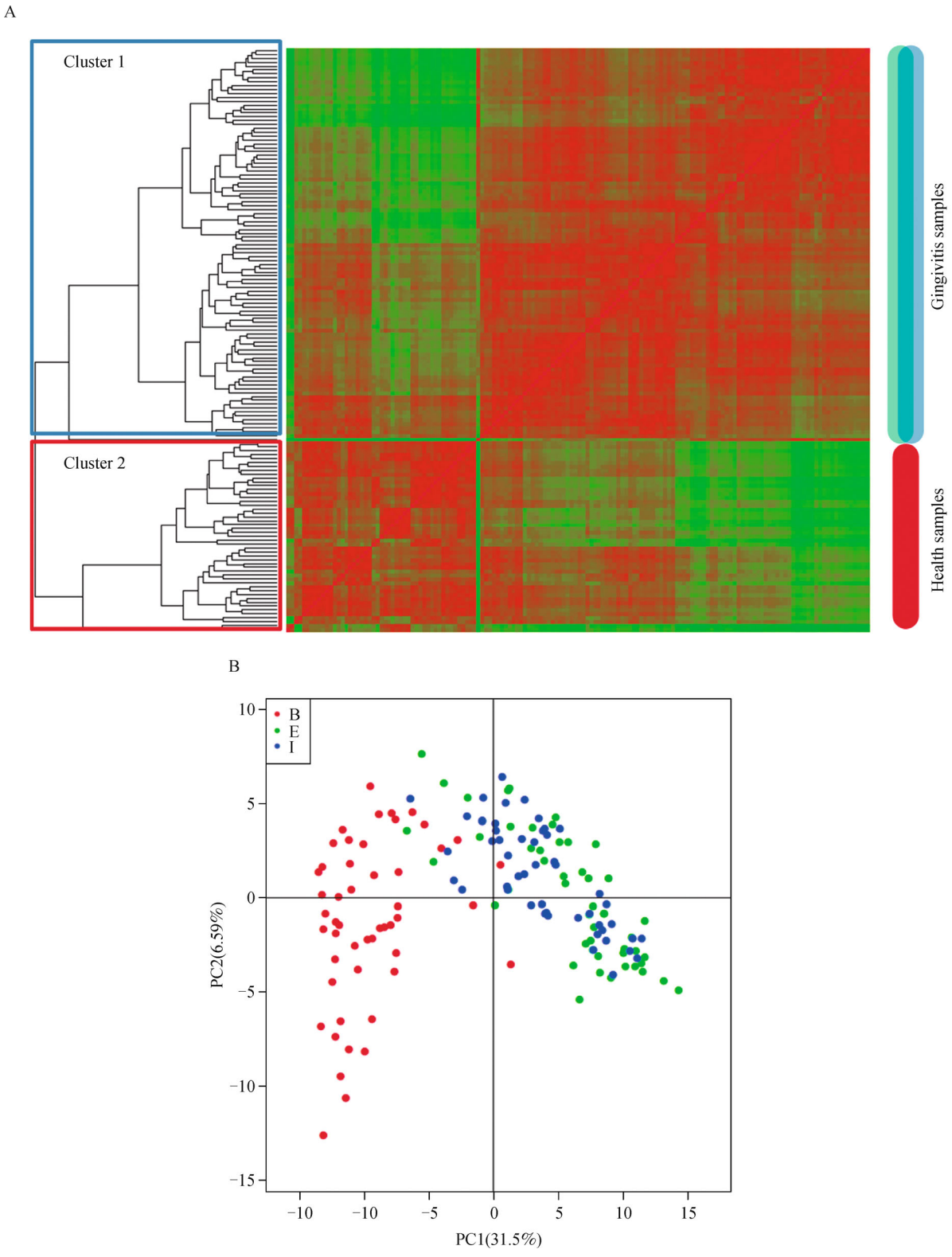

Figure 7. The clustering analysis results of human associated-habitat samples from different body sites based on the similarity matrix. (A) Similarity matrix and hierarchical clustering results. (B) PCoA analysis results by different healthy status. Saliva samples could be categorized into 3 groups: naturally occurring gingivitis (I), healthy gingivae (B) and experimental gingivitis (E).

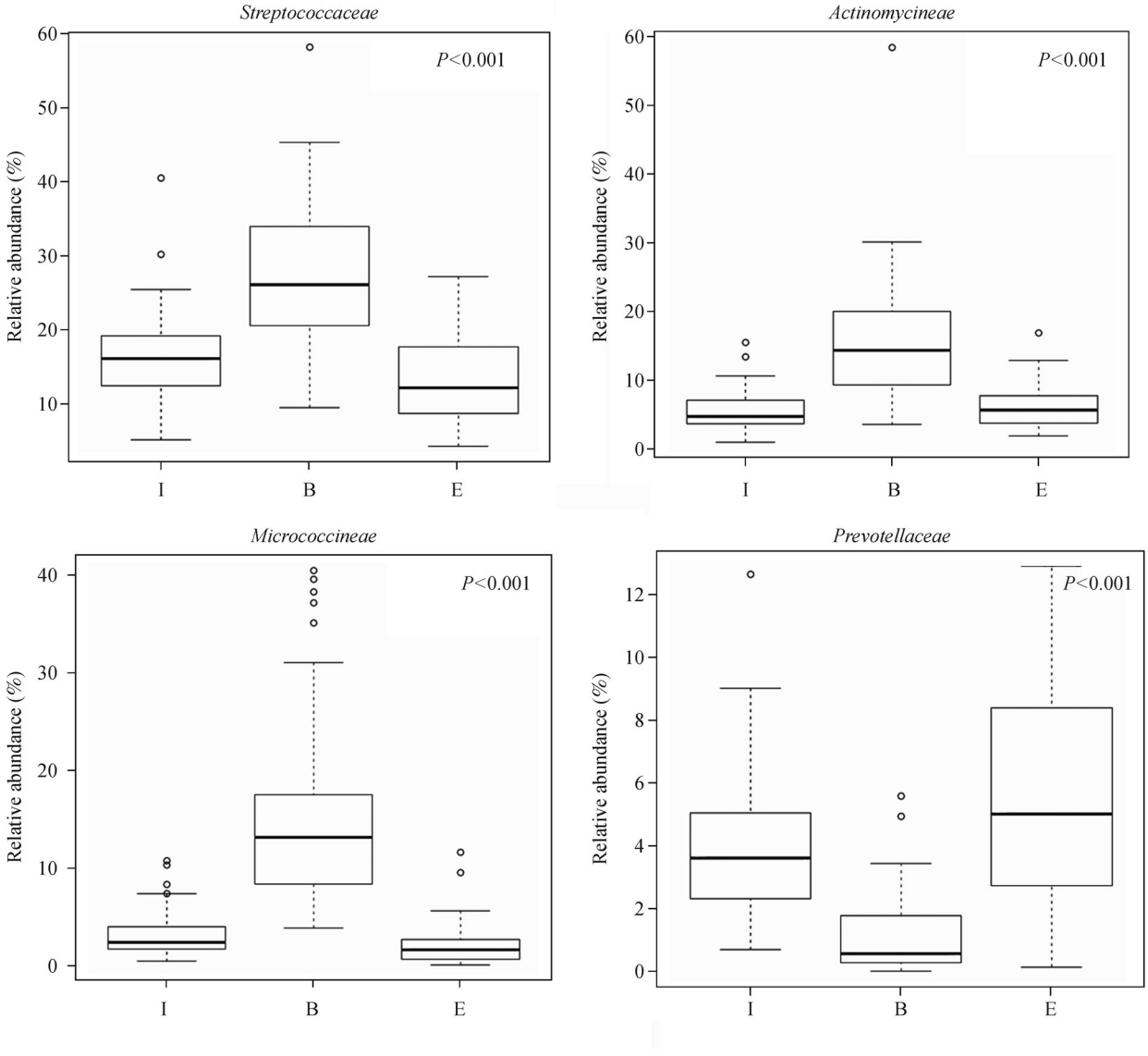

Figure 8. Biomarker analysis results of samples in different healthy status. Saliva samples could be categorized into three groups: naturally occurring gingivitis (I), healthy gingivae (B) and experimental gingivitis (E).

Metagenomic structure analysis

Meta-Mesh parses taxonomical and phylogenetic structure of metagenomic samples based on the Parallel-META software [27,28] with configurable parameters. Parallel-META extracts 16S rRNA or 18S rRNA fragments which are considered as bio-markers by the HMM algorithm [29] with the model built by Sliva database [30], and maps these extracted rRNA sequences to GreenGenes [31], RDP [32], Sliva [30] or Oral Core [33] database using megablast for high accuracy species identification, taxonomical annotation and phylogenetic analysis.

Quantitative phylogenetic-based sample comparison and similarity matrix

Meta-Mesh evaluated the similarity value (0%–100%) between each sample pair based on their quantitative comparison of common phylogenetic tree using the Meta-Storms scoring function [34]. This scoring function compares two microbial community samples' structures by bottom-up recursive traversal to their common weighted phylogenetic tree, in which edge weight represents the phylogenetic distance and node weight represents the species abundance. The similarity value of each leaf node (species) is the common relative

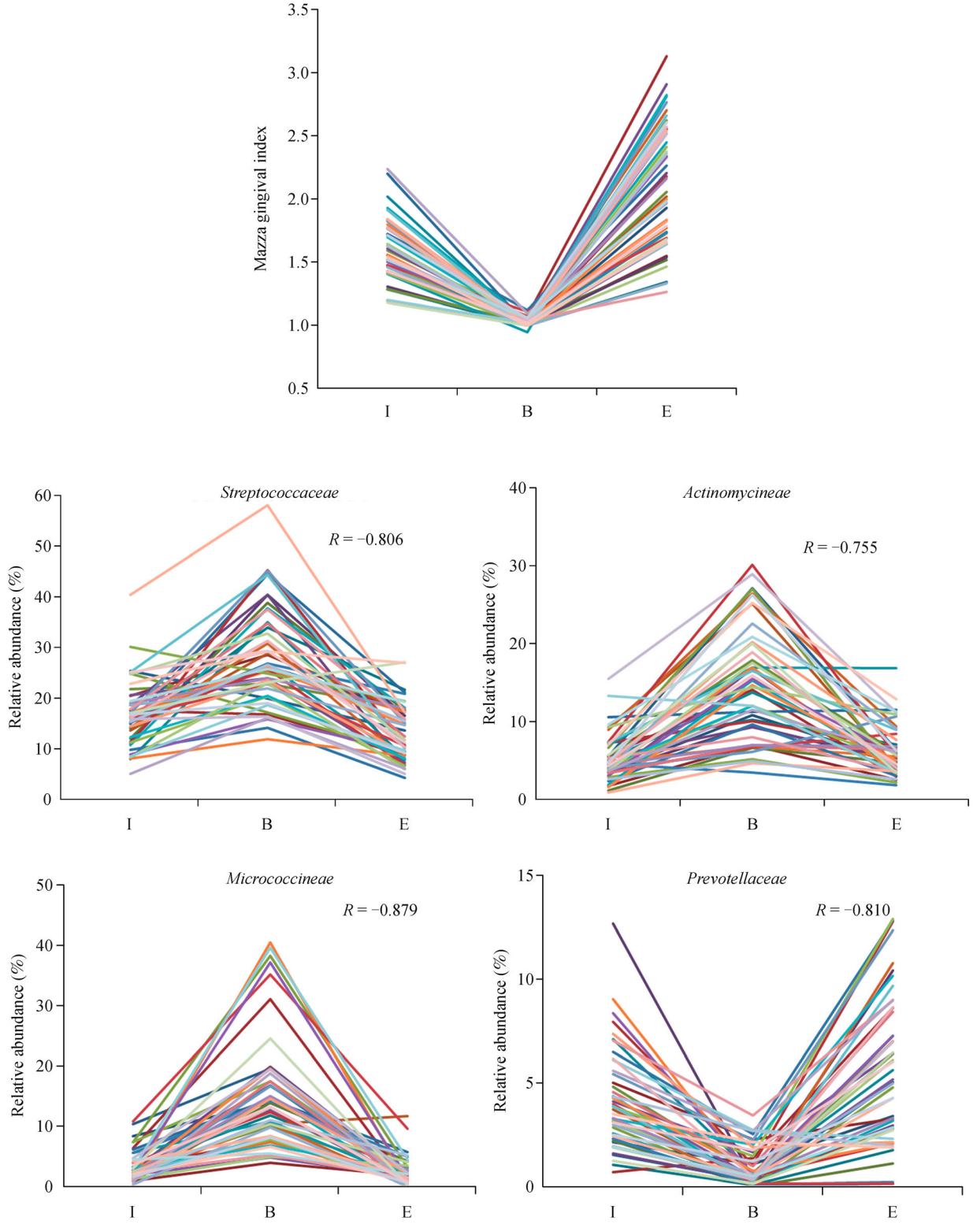

Figure 9. Correlation analysis results of the biomarker taxa of samples in different healthy status gauged by MGI.

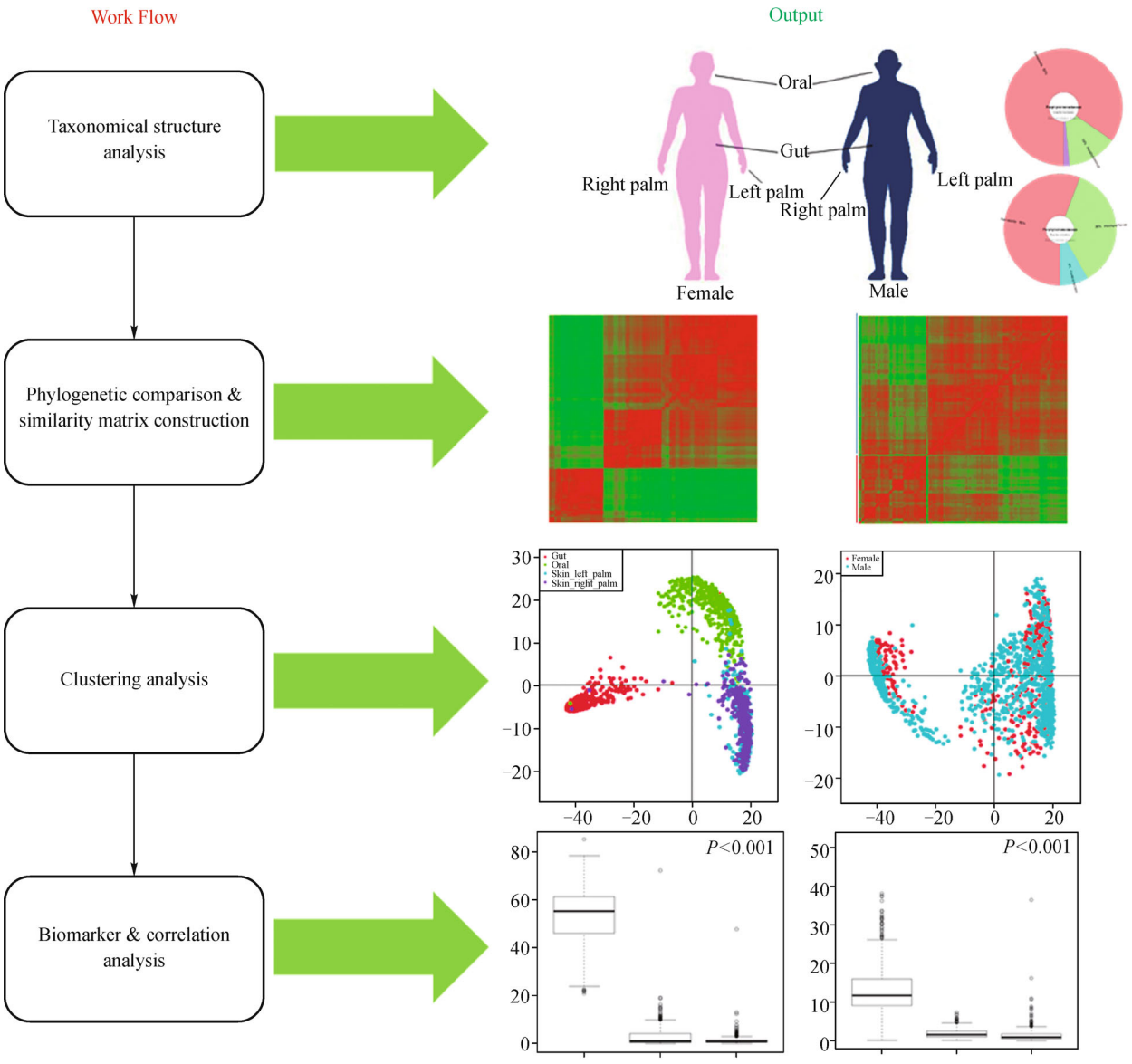

Figure 10. The overall scheme of Meta-Mesh system.

abundance of the two samples. Then the value of their abundance difference is reduced to the ancestor node by multiplying the phylogenetic distance for the calculation on upper level in the phylogenetic tree as the leaf nodes. Finally the sum of similarity values on all nodes is the overall similarity score (between 0% and 100%) of these two samples (for more details please refer to Su, et al., 2012 [34]). Then the pair-wised similarity values of all given sample pairs construct a similarity matrix. Meta-Mesh produces a visualization of the matrix where each tile represents a similarity value between two samples from a color gradient between red and green: red color indicates higher similarity value and green color indicates lower similarity value, with red/green shades in between indicating intermediate values.

Clustering analysis based on similarity matrix

The hierarchical clustering analysis measures the relationships among the microbial community samples based on the similarity matrix. This method is implemented by "HClust" function of CRAN R [35], and results are visualized by MetaSee software [36] and "gplots" package (Gregory R., et al., gplots: Various R program-

ming tools for plotting data. http://CRAN.R-project.org/package = gplots) of CRAN R. The Principal Coordinates Analysis (PCoA) are used to elucidate the distribution of similarity among samples by a give meta-data (e.g., host, body site) which is implemented by "vegan" package (ari Oksanen, et al., vegan: Community Ecology Package. http://CRAN.R-project.org/package = vegan) of CRAN R.

Biomarker and correlation analysis

In biomarker analysis the Meta-Mesh select abundant taxa which are considered as key factor led to the variations among sample clusters and groups. The significance of difference in abundance distributions is measured by the Wilcoxon and Kruskal rank-sum test, then Meta-Mesh select the taxa that have *P*-value smaller or equal to 0.001. Meta-Mesh also calculates the Pearson correlation coefficient (*R*) between abundance values of biomarker taxa their meta-data, and select the taxa with absolute *R* value equal to or larger than 0.7 which indicate the strong correlated varying trend between the abundance values and meta-data.

High-performance computation techniques for Meta-Mesh

In Meta-Mesh system, all steps are supported by high-performance computation techniques including multi-thread computation and GPU computation.

In metagenomic structure analysis we implemented CPU based parallel computing to improve the analysis efficiency. In Parallel-META software we developed the OpenMP technology based parallel 16S/18S rRNA sequence mapping on multi-core CPU. This strategy could achieved much higher efficiency based on multi-core CPU and OpemMP technology than single thread serial computing [27,28].

In sample comparison analysis we used GPU computing for calculation of similarity matrix with large number of samples. Benefited by the many-core architecture of GPU, scoring function can be invocated in parallel by multi-threading to compute the similarities among massive amount of samples. To calculate the pair-wise similarity matrix of *N* samples, $N\times(N\text{-}1)/2$ threads are launched in GPU to make each similarity value in the matrix processed by one independent thread. In our previous experiment the running time with GPU would be able to compute the pair-wise similarity values for 10, 240 samples within 10 min, which gained a speed-up of $>17{,}000$ times compared with single-core CPU, and $>2{,}600$ times compared with 16-core CPU [37].

Therefore, the high-performance computation backbone for Meta-Mesh could significantly accelerate the comparison analysis, thus facilitate in-depth data mining among massive microbial community samples.

ACKNOWLEDGMENTS

This work is supported in part by Chinese Academy of Sciences' e-Science grant (NO. INFO-115-D01-Z006), Ministry of Science and Technology's high-tech (863) grant (NO. 2009AA02Z310 and NO. 2014AA21502), National Science Foundation of China Grant (NO. 61103167, NO. 31271410 and NO. 61303161), and The Open Fund of Key Laboratory of Marine Ecology and Environmental Science, Institute of Oceanology, Chinese Academy of Sciences (NO. KLMEES201304).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Xiaoquan Su, Xiaojun Wang, Gongchao Jing, Shi Huang, Jian Xu and Kang Ning declare that they have no conflict of interest.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. 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
2. National Research Council. (2007) The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet. Washington: the National Academies Press
3. Jurkowski, A., Reid, A. H. and Labov, J. B. (2007) Metagenomics: a call for bringing a new science into the classroom (while it's still new). CBE Life Sci. Educ., 6, 260–265
4. Ley, R. E., Lozupone, C. A., Hamady, M., Knight, R. and Gordon, J. I. (2008) Worlds within worlds: evolution of the vertebrate gut microbiota. Nat. Rev. Microbiol., 6, 776–788
5. Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R. and Gordon, J. I. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature, 444, 1027–1031
6. Grice, E. A. and Segre, J. A. (2011) The skin microbiome. Nat. Rev. Microbiol., 9, 244–253
7. Kong, H. H. and Segre, J. A. (2012) Skin microbiome: looking back to move forward. J. Invest. Dermatol., 132, 933–939
8. Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. A., Affourtit, J. P., et al. (2009) A core gut microbiome in obese and lean twins. Nature, 457, 480–484
9. Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., et al. (2012) Human gut microbiome viewed across age and geography. Nature, 486, 222–227
10. Yang, F., Zeng, X., Ning, K., Liu, K. L., Lo, C. C., Wang, W., Chen, J., Wang, D., Huang, R., Chang, X., et al. (2012) Saliva microbiomes distinguish caries-active from healthy human populations. ISME J., 6, 1–10
11. Nasidze, I., Li, J., Schroeder, R., Creasey, J. L., Li, M. and Stoneking, M. (2011) High diversity of the saliva microbiome in Batwa Pygmies. PLoS One, 6, e23352
12. Muegge, B. D., Kuczynski, J., Knights, D., Clemente, J. C., González, A., Fontana, L., Henrissat, B., Knight, R. and Gordon, J. I. (2011) Diet

drives convergence in gut microbiome functions across mammalian phylogeny and within humans. Science, 332, 970–974

13. Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., et al. (2008) The metagenomics RAST server— a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics, 9, 386
14. Seshadri, R., Kravitz, S. A., Smarr, L., Gilna, P. and Frazier, M. (2007) CAMERA: a community resource for metagenomics. PLoS Biol., 5, e75
15. Huson, D. H., Auch, A. F., Qi, J. and Schuster, S. C. (2007) MEGAN analysis of metagenomic data. Genome Res., 17, 377–386
16. Mitra, S., Gilbert, J. A., Field, D. and Huson, D. H. (2010) Comparison of multiple metagenomes using phylogenetic networks based on ecological indices. ISME J., 4, 1236–1242
17. Mitra, S., Klar, B. and Huson, D. H. (2009) Visual and statistical comparison of metagenomes. Bioinformatics, 25, 1849–1855
18. Parks, D. H. and Beiko, R. G. (2010) Identifying biologically relevant differences between metagenomic communities. Bioinformatics, 26, 715–721
19. Kristiansson, E., Hugenholtz, P. and Dalevi, D. (2009) ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. Bioinformatics, 25, 2737–2738
20. Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., et al. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol., 75, 7537–7541
21. Goll, J., Rusch, D. B., Tanenbaum, D. M., Thiagarajan, M., Li, K., Methé, B. A. and Yooseph, S. (2010) METAREP: JCVI metagenomics reports—an open source tool for high-performance comparative metagenomics. Bioinformatics, 26, 2631–2632
22. Lozupone, C. and Knight, R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities. Appl. Environ. Microbiol., 71, 8228–8235
23. Hamady, M., Lozupone, C. and Knight, R. (2010) Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. ISME J., 4, 17–27
24. Chen, T., Yu, W. H., Izard, J., Baranova, O. V., Lakshmanan, A. and Dewhirst, F. E. (2010) The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. Database, 2010, baq013
25. Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., Knights, D., Gajer, P., Ravel, J., Fierer, N., et al. (2011) Moving pictures of the human microbiome. Genome Biol., 12, R50
26. Huang, S., Li, R., Zeng, X., He, T., Zhao, H., Chang, A., Bo, C., Chen, J., Yang, F., Knight, R., et al. (2014) Predictive modeling of gingivitis severity and susceptibility via oral microbiota. ISME J., 8, 1768–1780
27. Su, X., Xu, J. and Ning, K. (2012) Parallel-META: efficient metagenomic data analysis based on high-performance computation. BMC Syst. Biol., 6, suppl 1, S16
28. Su, X., Pan, W., Song, B., Xu, J. and Ning, K. (2014) Parallel-META 2.0: enhanced metagenomic data analysis with functional annotation, high performance computing and advanced visualization. PLoS One, 9, e89323
29. Rabiner, L. R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE, 77, 257–286.
30. Pruesse,E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J. and Glöckner, F. O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res., 35, 7188–7196
31. DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P. and Andersen, G. L. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol., 72, 5069–5072
32. Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., Marsh, T., Garrity, G. M., et al. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res., 37, D141–D145
33. Griffen, A. L., Beall, C. J., Firestone, N. D., Gross, E. L., Difranco, J. M., Hardman, J. H., Vriesendorp, B., Faust, R. A., Janies, D. A. and Leys, E. J. (2011) CORE: a phylogenetically-curated 16S rDNA database of the core oral microbiome. PLoS One, 6, e19051
34. Su, X., Xu, J. and Ning, K. (2012) Meta-Storms: efficient search for similar microbial communities based on a novel indexing scheme and similarity score for metagenomic data. Bioinformatics, 28, 2493–2501
35. Dessau, R. B. and Pipper, C. B. (2008) "R"—project for statistical computing. Ugeskr Laeg, 170, 328–330
36. Song, B., Su, X., Xu, J. and Ning, K. (2012) MetaSee: an interactive and extendable visualization toolbox for metagenomic sample analysis and comparison. PLoS One, 7, e48998
37. Su, X., Wang, X., Jing, G. and Ning, K. (2014) GPU-Meta-Storms: Computing the structure similarities among massive amount of microbial community samples using GPU. Bioinformatics, 1031–1033

