Metagenomic Visualization and Relevant Microbes Analysis in Human Oral Microbiota Related to Periodontal Disease *

[1]Wei-Ren Lin, [1]Wen-Qing Luo, [1]Wen-Pei Chen and [2]Ming-Li Liou, [1]Yaw-Ling Lin*

[1]Dept. Computer Science and Information Engineering
Providence University, Taichung, Taiwan
[2]Dept. Medical Laboratory Science and Biotechnology
Yuanpei University, Hsin-Chu City, Taiwan
a7826723@yahoo.com.tw, g1050308@pu.edu.tw, wenpei-keynes@gmail.com
d918229@yahoo.com.tw, yllin@pu.edu.tw

Abstract

Metagenomic information provides deeper understanding of the ecological role, metabolism, and evolutionary history of microbes in a given ecosystem by analyzing environmental DNA directly without prior cultivation. In this paper, we propose data visualization methods and implement tools to facilitate the bioinformatics analysis of metagenomic data. The open-source metagenomic sequences data analysis softwares were integrated to construct accessible platforms for metagenomic data analysis. Microbial ecologists can now start digging into the accumulating mountains of metagenomic data to uncover the occurrence of functional genes and their correlations to microbial community members. Limitations and biases in DNA extraction and sequencing technologies impact sequence distributions, and therefore, have to be considered.

1 Introduction

The NGS (next generation sequencing)-based metagenomic data analysis is becoming the mainstream for the study of microbial communities. Metagenomics is the study of genomes of multiple species from environmental samples, such as soil sea water, and the human gut [4, 8, 31, 2, 12]. The link with human body environments generated many studies of microbial community composition designed to assess its role in various metabolic pathway and to determine whether it is involved in inducing and preventing specific pathological conditions. Such investigations could help to clarify the pathogenesis of specific diseases and could also lead to novel disease-markers and to the development of novel therapeutic strategies.

Faced with a large amount of data in metagenomic research, effective data visualization is important for scientists to effectively explore, interpret and manipulate such rich information. The visualization of the metagenomic data, especially multi-sample data, is one of the most critical challenges. Visualization has thus been a prominent aspect of the field, beginning with the analysis package MEGAN [1,2]. Distilling metagenomic data into graphical representations, however, is not a trivial task [28, 22, 33, 16, 17]. The foundation of most metagenomic studies is the assignment of observed nucleic acids to taxonomic or functional hierarchies. The various levels of granularity (e.g. ranks) inherent in these classifications pose a challenge for visualization.

Due to technological improvements in sequencing methods and sample extraction techniques, virtually all the microbes from a given environment can be analyzed in an efficient run, avoiding cultivation steps. In particular, procedures based on 16S rRNA next-generation sequencing, which allow the high throughput microbial identification within a specific metagenome, represent a powerful means to investigate the composition and the biodiversity of microbial communities [18]. The enormous amount of next-generation metagenomic data generated by such procedures necessitates bioinformatic tools and platforms able to analyze them. In fact, an accurate taxonomic assignment of each microbe in a target environment

*Corresponding author. Email:yllin@pu.edu.tw.

is required to evaluate the structure, the biodiversity, the richness and the role of the community resident in a given environment [25, 9].

In this paper, we provide metagenomic analysis platforms constructed by integrating QIIME [6], PEAR [32], UCHIME [11], UPARSE [10], and other open-source tools. The system is integrated with Hadoop cloud platform and provides efficient and reliable solutions. We also introduce feature selection algorithms for SVMs. The method was based on correlation coefficient between microorganism and healthy state associated with periodontal disease. Bioinformatic analysis of human oral metagenomic data are conducted on the platform to identify the microbiome composition. As a result, the characteristics of human oral environment and analysis of the diversity and richness of the microbial community is reported in the paper.

2 Materials and Methods

2.1 16S rRNA Sequence Dataset

We analyze the efficiency of our refined metagenomics analysis by performing several sets of experiment on the dataset which are subgingival plague. The next generation sequencing evaluation of their oral microbial communities was carried out by using *Illumina MiSeq* after performing amplicon sequencing on 16S rRNA V1-V2 region and PCR reaction of 10 to 18 cycles to enrich the adapter-modified DNA fragments. The minimum length = 35 and error probability < 0.05 was adopted as the criteria for quality trim processing.

2.2 Bioinformatics Analysis

2.2.1 Pre-analysis Step

The 16S rRNA next-generation sequencing run enviroment and oral microbiome to produce biom file in the pre-processing step which called bioinformatic analysis in Figure 1. The pre-analysis step includes paired-end reads assembly, barcodes filtering and trimming, and chimeras removing. The goal of this step is to filtering out noise sequences; and then, once denoising and additional quality control processes are completed, chimeric sequences should be removed from the dataset. The following parameters were set for our experiments: (1) a minimum average quality Phred score of 25 allowed in reads ; (2) 10 bases minimum overlap required in assembly processing; (3) a minimum and maximum sequence length in the range of 50-1000 bases; and (4) a maximum number of ambiguous bases and length of homopolymers equal to 6. In addition, to be as stringent as possible, no any primer mismatches was allowed in our experiments and only a 1.5 maximum number of errors in barcodes was allowed. The "Gold" database which is a FASTA file containing the ChimeraSlayer reference database in the Broad Microbiome Utilities [14] (http://microbiomeutil.sourceforge.net/) was used for chimeras detection and removing.

2.2.2 16S rRNAs Detection, Clustering, and Identification

The method of UPARSE-OTU is used to generate clusters from NGS. OTU clustering is an indispensable process in classification analysis. Dereplication, abundance sort, discard singletons and OTU clustering compose the OTU picking procedure. Reads that are singletons after quality filtering and global trimming are discarded after the removal of duplicated sequences. Then, reads with abundances of two or more are sorted by decreasing abundance and are used as input for OTU clustering. In OTU clustering precess, reads are assigned to OTUs by clustering the reads that match the OTU with $\geq 97\%$ identity. A sequence is taken in a sequence collection that represents the presence of a taxonomic unit when it shows a similarity level above the required threshold (97% identity). After the OTU picking step, the representative sequence for each OTU, namely, the most abundant sequences in that OTU, is chosen for subsequent analyses in order to reduce the computational power and the analysis time, without losing the frequency information.

2.2.3 Taxonomic Classification

QIIME can perform the taxonomy assignment using different methods such as BLAST, RDP [30], UCLUST, Rtax and Mothur [26]. We adopted the BLAST against the Human Oral Microbiome Database [7] (available at http://www.homd.org/) and Greengenes Database, setting the Maximum e-Value Cutoff to 0.001 in this study. Reads assigned to the Bacteria root but not attaining the threshold at the chosen taxonomic level fell in the category "Unclassified", while sequences not assigned to the Bacteria root were classified as "No Hits". After taxonomic assignment, QIIME generates a Biological Observation Matrix (BIOM) file useful to transfer the obtained data to other tools for analysis purposes.

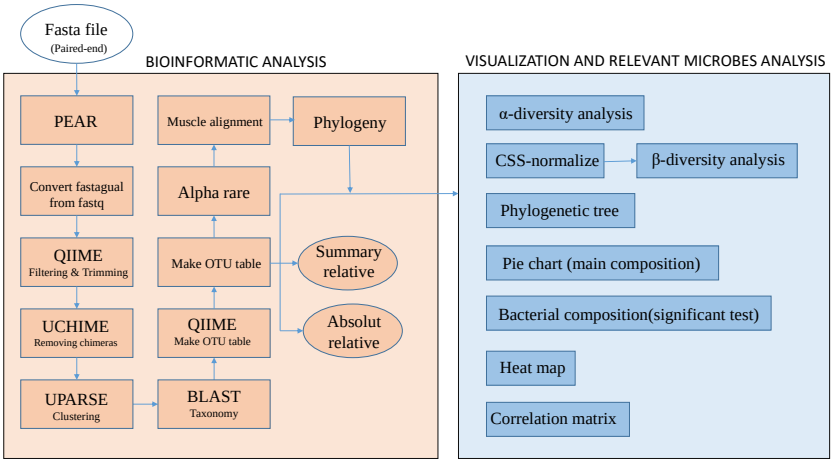

Figure 1: The diversity and abundance analysis of Oral microbiome with two types in phynotype.

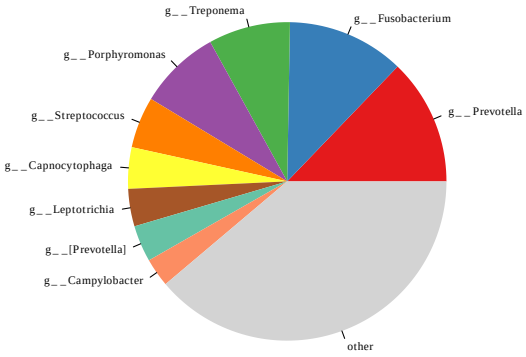

Figure 2: The pie chart of top 10 taxonomic composition at genus level.

2.3 Visualization and Relevant Microbes Analysis

Visualization is an intuitive way to analyze large-scale alignment data in genomic studies. There are many visualization tools available. Some are web browser-based such as UCSC genome browser [13], LookSeq [19] and JBrowse [27]. Some are standalone programs such as Tablet [20], GenomeView [1], MapView [5], IGB [21], IGV [29], SamScope [24] and so on.

2.3.1 Taxonomic Composition Analysis

Figure 2 shows dominant composition at genus level with Oral microbiome.

2.3.2 Alpha-diversity Analysis

Alpha-diversity estimates are methods for describing of the number of types of organisms in a single sample. These measures can also take into account the evenness of taxa in a sample. Alpha-diversity analyses are useful for examining patterns of dominance, rarity and community complexity. We use open-source software phyloseq package in R to perform alpha diversity analysis according to (Case, Control) and (Health-HH, moderate patients-MP, severe patients-SP) paired in phenotype with the observed species, ACE, and Chao 1 metric [15].

Figure 3 and Figure 4 show difference in bacterial abundance and diversity.

2.3.3 Beta-diversity Analysis

Beta-diversity approaches provide a way for comparing the microbial community composition between two samples. With these methods we are able to simultaneously compare changes in the presence/absence or abundance of thousands of taxa in a microbiome dataset and summarize these into how similar or dissimilar two samples are. We use cumulative sum scaling (CSS) normalization [23], which corrects the bias in the assessment of differential abundance introduced by total-sum normalization (TSS) [23]. Figure 5 shows original BIOM compared to normalized BIOM by CSS in CCA, NMDS, PCoA, RDA scaling methods, where CSS normalization is able to best separate samples based on different phynotypes while controlling within-group variance. How-

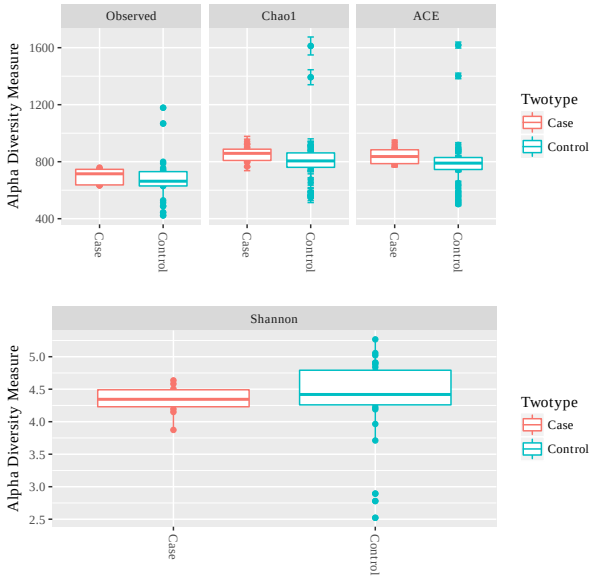

Figure 3: The diversity and abundance analysis of Oral microbiome with two types in phenotype.

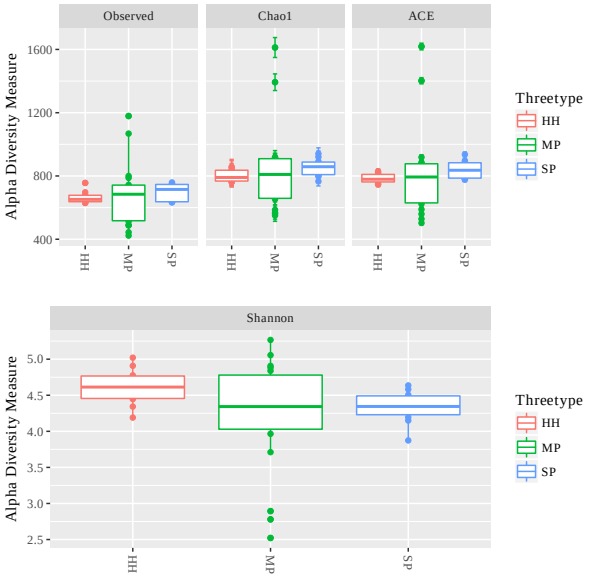

Figure 4: The diversity and abundance analysis of Oral microbiome with three types in phynotype.

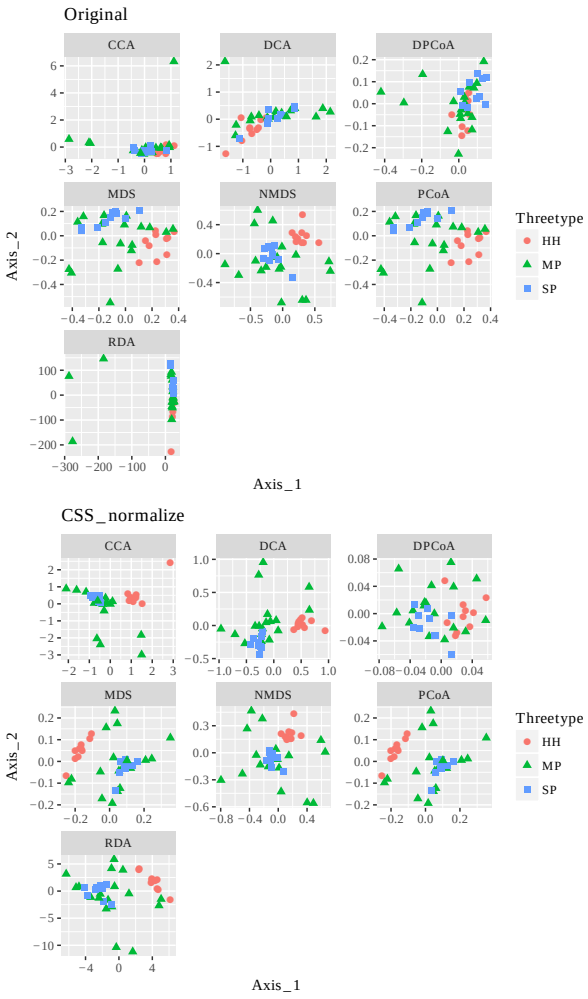

Figure 5: The comparison of original and normalized BIOMs through CSS in different scaling methods.

ever, we propose scale_normalize approach about PCA through selected significant taxonomy with Kruskal-Willis test and scale distance matrix. Figure 6 shows scale_normalization is better than CSS and original scaling.

2.3.4 Phylogenetic Tree

GraPhlAn [3] is a software tool for producing high-quality circular representations of taxonomic and phylogenetic trees. GraPhlAn focuses on concise, integrative, informative, and publication-ready representations of phylogenetically- and taxonomically-driven investigation. We use open-source GraPhlAn to perform high-quality phylogenetic tree with Oral microbiome. Figure 7 shows annotations of phylum and Class at left legend and

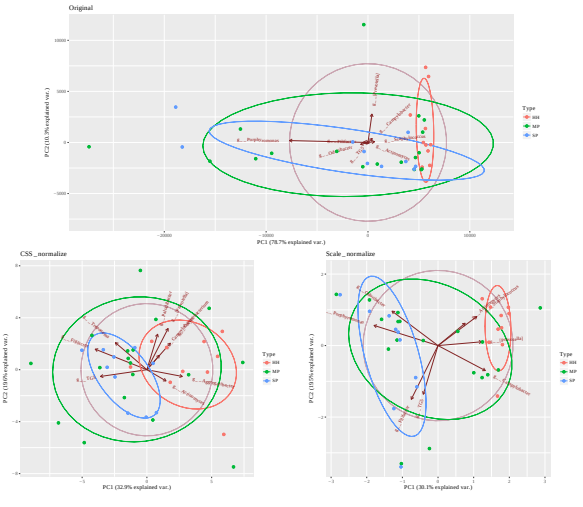

Figure 6: The PCA of Original, CSS_normalize and Scale_normalize approach.

Taxonomy	*p*-value
Staphylococcus	0.0000239
Filifactor	0.000125
Campylobacter	0.000513
Odoribacter	0.00425
[Prevotella]	0.00504
Actinomyces	0.00618
Porphyromonas	0.0064
TG5	0.007

Table 1: Greater than 0.5% abundance taxanomy with *p*-value less than 0.01 at genus level.

genus and species in phylogenetic tree.

2.3.5 Bacterial Community Composition

We select taxonomy greater than 0.5% abundance in each sample at genus level and find out significant taxonomy through *p*-value less than 0.01 with Kruskal-Willis test. Figure 8 shows bacteria composition of different phynotype and significant taxonomy among HH, MP, and SP.

2.3.6 Heatmap

We select taxonomy greater than 0.5% abundance in each sample at genus level and find out significant taxonomy through *p*-value less than 0.01 with Kruskal-Willis test and cluster samples and genera to generate phylogenetic. Figure 9 is the heat map of taxonomic composition at genus level, *Porphyromonas* is the most rich genus in severe samples, *Campylobacter*and *[Provotella]* are the most rich

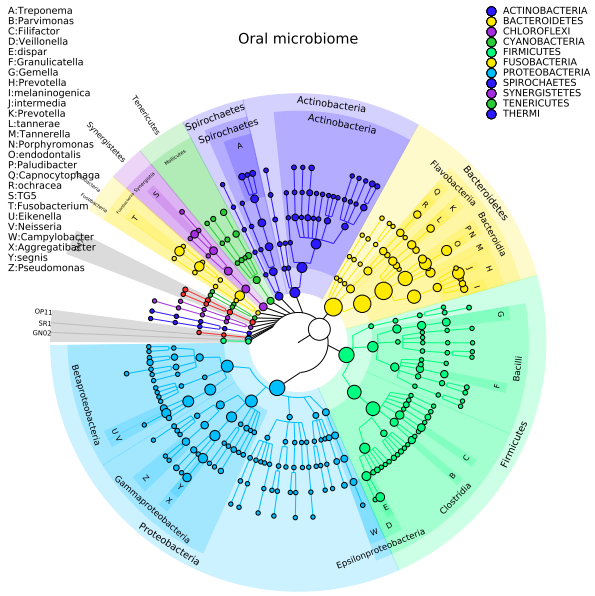

Figure 7: The integrated phylogenetic tree of oral microbiome with phylum, order, genus, species annotations.

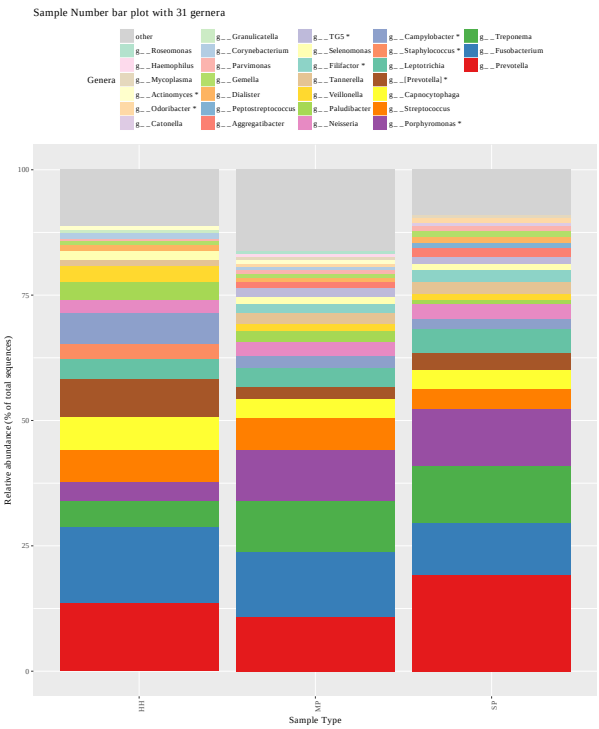

Figure 8: The taxonomic composition with taxonomy greater than 0.5% abundance at genus level in each type, *: Kruskal-Willis test *p*-value less than 0.01.

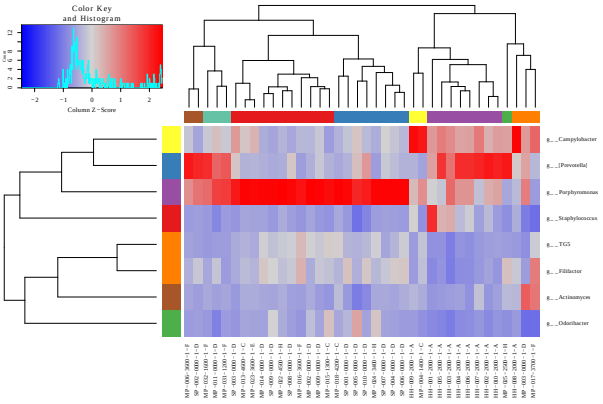

Figure 9: The heatmap of taxonomic composition at genus level.

genus in health samples. Through automatically cluster samples in x-axis with eight bacteria at genus level in Table 1, health sample and severe sample are seperated to different group obviously, moreover the ambiguous moderate samples are assigned to closely severe or health sample clearly.

2.3.7 Correlation Matrix

In order to have deeper understanding of relationship among significant bacteria, we calculated the correlation coefficients with Pearson. Here, positive correlation is 1, negative correlation is -1 , and no correlation is 0 between two bacteria. Table 1 shows the top eight features with lower p-value. The correlation coefficient between this features were analyzed, Figure 10 shows the correlation between the top 8 informative features. It can find that, *Porphyromonas*, *Fillifactor* and *TG5* have more symbiotic relationship.

3 Experimental Results

In this experiment, alpha-diversity analysis of data mining in Figure 1, Figure 3, and Figure 4 shows that diversity of SP is greater than HH, but abundance of HH is greater than SP. Beta-diversity analysis can use CSS normalize approach to optimal scaling for filtering biases in raw data. Furthermore, we propose scale approach in PCA can also optimal scaling among significant bacteria. GraphlAn provide high-quality phylogenetic tree annotation. We draw pie chart to find out dominant composition of bacteria community. Bacterial composition shows taxonomy composi-

Figure 10: The correlation between the 8 information feature with p-value less than 0.01.

tion of each sample and we use Kruskal-Willis test for performing significant taxonomy. The heatmap shown in Figure 9 reveals the following significant genera *Campylobacter*, *[Prevotella]*, *Porphyromonas*, *Staphylococcus*, *TG5*, *Fillifactor*, *Actinomyces* and *Odoribacter*; the assoicated p-values of these genera is summarized in Table 1. For deeper understanding of relationship among bacteria, we can calculate correlation coefficient between significant bacteria .

4 Future Works and Conclusions

For the past few years, metagenomics data have been growing explosively. The problem is how to find clue in these datasets. In this paper, we integrate many open source software system of metagenomic analysis and use visual analysis tool for data mining, we use alpha-diversity analysis, beta-diversity analysis, phylogenetic tree, dominant composition, significant test, heat map, correlatiom matrix. Moreover, in order to filter biases from analysis data sets, we use CSS normalize approach and propose our approach to analyze beta-diversity. Furthermore, we construct several tailored VM (virtual machine) images by OpenStack hypervisor [15] that can be download from our web site service to provide services for metagenomics analysis researchers and interested biologists.

5 Acknowledgment

This research was partially supported by the Ministry of Science and Technology under the Grants MOST 103-2632-E-126-001-MY3. The 16S rRNA sequence dataset used in this paper was offered by the project which sponsored by the Ministry of Science and Technology under the Grants MOST 103-2622-E-126-002-CC1.

References

[1] T. Abeel, T. Van Parys, Y. Saeys, J. Galagan, and Y. Van de Peer. GenomeView: a next-generation genome browser. *Nucleic acids research*, 40(2):e12–e12, 2012.

[2] L. D. Alcaraz, P. Belda-Ferre, R. Cabrera-Rubio, H. Romero, . Simn-Soro, M. Pignatelli, and A. Mira. Identifying a healthy oral microbiome through metagenomics. *Clinical Microbiology and Infection*, 18:54–57, 2012.

[3] B. J. Baker, C. S. Sheik, C. A. Taylor, S. Jain, A. Bhasi, J. D. Cavalcoli, and G. J. Dick. Community transcriptomic assembly reveals microbes that contribute to deep-sea carbon and nitrogen cycling. *The ISME Journal*, 7(10):1962–1973, 2013.

[4] H. Bao, H. Guo, J. Wang, R. Zhou, X. Lu, and S. Shi. MapView: visualization of short reads alignment on a desktop computer. *Bioinformatics*, 25(12):1554–1555, 2009.

[5] J. G. Caporaso, J. Kuczynski, and J. Stombaugh. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336, 2010.

[6] T. Chen, W.-H. Yu, J. Izard, O. V. Baranova, A. Lakshmanan, and F. E. Dewhirst. The human oral microbiome database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database*, 2010.

[7] I. Cho and M. J. Blaser. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics*, 13(4):260–270, 2012.

[8] C. De Filippo, M. Ramazzotti, P. Fontana, and D. Cavalieri. Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. *Briefings in Bioinformatics*, 13(6):696–710, 2012.

[9] R. C. Edgar. Uparse: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, 10(10):996–998, 2013.

[10] R. C. Edgar, B. J. Haas, J. C. Clemente, C. Quince, and R. Knight. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, 27(16):2194–2200, 2011.

[11] S. Fang and R. M. Evans. Microbiology: Wealth management in the gut. *Nature*, 500(7464):538–539, 2013.

[12] P. A. Fujita, B. Rhead, A. S. Zweig, A. S. Hinrichs, D. Karolchik, M. S. Cline, M. Goldman, G. P. Barber, H. Clawson, A. Coelho, et al. The UCSC genome browser database: update 2011. *Nucleic acids research*, page gkq963, 2010.

[13] B. J. Haas, D. Gevers, A. M. Earl, and M. Feldgarden. Chimeric 16s rRNA sequence formation and detection in sanger and 454-pyrosequenced pcr amplicons. *Genome Research*, 21(3):494–504, 2011.

[14] T. C. Hill, K. A. Walsh, and J. A. Harris. Using ecological diversity measures with bacterial communities. *FEMS Microbiology Ecology*, 43(1):1–11, 2003.

[15] S. Hunter, M. Corbett, H. Denise, M. Fraser, A. Gonzalez-Beltran, C. Hunter, P. Jones, R. Leinonen, C. McAnulla, E. Maguire, et al. EBI metagenomics–a new resource for the analysis and archiving of metagenomic data. *Nucleic acids research*, 42(D1):D600–D606, 2014.

[16] D. H. Huson and N. Weber. Microbial community analysis using megan. *Methods in enzymology*, 531:465–485, 2012.

[17] B.-S. Kim, Y.-S. Jeon, and J. Chun. Current status and future promise of the human microbiome. *Pediatric Gastroenterology, Hepatology & Nutrition*, 16(2):71–79, 2013.

[18] H. M. Manske and D. P. Kwiatkowski. LookSeq: a browser-based viewer for deep sequencing data. *Genome research*, 19(11):2125–2132, 2009.

[19] I. Milne, M. Bayer, L. Cardle, P. Shaw, G. Stephen, F. Wright, and D. Marshall. Tabletnext generation sequence assembly visualization. *Bioinformatics*, 26(3):401–402, 2010.

[20] J. W. Nicol, G. A. Helt, S. G. Blanchard, A. Raja, and A. E. Loraine. The integrated genome browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics*, 25(20):2730–2731, 2009.

[21] B. D. Ondov, N. H. Bergman, and A. M. Phillippy. Interactive metagenomic visualization in a web browser. *BMC bioinformatics*, 12(1):385, 2011.

[22] J. N. Paulson, O. C. Stine, H. C. Bravo, and M. Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature methods*, 10(12):1200–1202, 2013.

[23] K. Popendorf and Y. Sakakibara. SAMSCOPE: an OpenGL-based real-time interactive scale-free sam viewer. *Bioinformatics*, 28(9):1276–1277, 2012.

[24] P. Ribeca and G. Valiente. Computational challenges of sequence classification in microbiomic data. *Briefings in Bioinformatics*, 2011.

[25] P. D. Schloss, S. L. Westcott, T. Ryabin, and J. R. Hall. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75(23):7537–7541, 2009.

[26] M. E. Skinner, A. V. Uzilov, L. D. Stein, C. J. Mungall, and I. H. Holmes. JBrowse: a next-generation genome browser. *Genome research*, 19(9):1630–1638, 2009.

[27] B. Song, X. Su, J. Xu, and K. Ning. MetaSee: an interactive and extendable visualization toolbox for metagenomic sample analysis and comparison. *PLoS One*, 7(11):e48998, 2012.

[28] H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2):178–192, 2013.

[29] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole. Nave bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73(16):5261–5267, 2007.

[30] P. L. Zeeuwen, M. Kleerebezem, H. M. Timmerman, and J. Schalkwijk. Microbiome and skin diseases. *Current Opinion in Allergy and Clinical Immunology*, 13(5), 2013.

[31] J. Zhang, K. Kobert, T. Flouri, and A. Stamatakis. PEAR: a fast and accurate illumina paired-end read merger. *Bioinformatics*, 30(5):614–620, 2014.

[32] Z. Zhu, B. Niu, J. Chen, S. Wu, S. Sun, and W. Li. MGAviewer: a desktop visualization tool for analysis of metagenomics alignment data. *Bioinformatics*, 29(1):122–123, 2013.