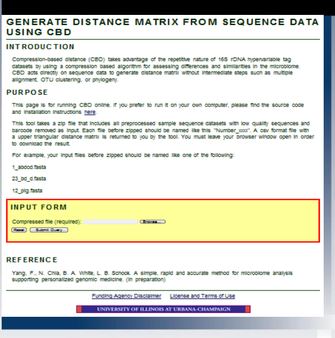
ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Laboratory of

Comparative Genomics



A Simple, Rapid and Accurate Method for Microbiome Analysis Supporting Personalized Genomic Medicine

Fang Yang[1,2], Nicholas Chia[2], Bryan A. White[1,2,3], Lawrence B. Schook[1,2,3]

[1]Division of Nutritional Sciences, [2]Institute for Genomic Biology,

[3]Department of Animal Sciences, University of Illinois at Urbana-Champaign

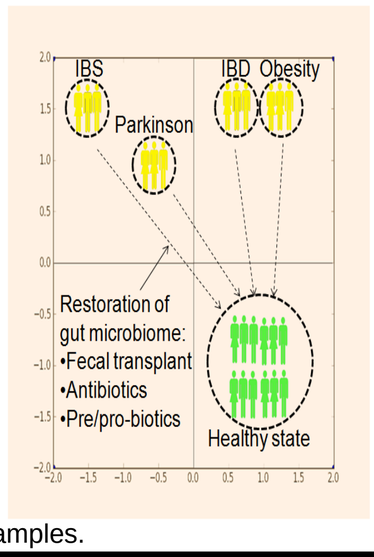
Abstract

Microbial community structure is associated with a variety of diseases, many of which have been treated using regimens that alter the gut microbiome toward that of a healthy individual. High-throughput sequencing provides a lens for assessing microbially-based personalized genomic medicine. However, analyzing large sequencing datasets for clinically relevant information currently requires exceptional expertise and computational resources. Thus, the widespread clinical application of microbiome diagnoses requires, a simple, rapid, and accurate means of comparing microbial communities. Our metric, compression-based distance (CBD), quantifies the degree of similarity between microbial communities. CBD takes advantage of the repetitive nature of 16S rDNA and compression algorithms to assess similarities between microbial communities according to the degree their concatenated datasets compress. Three published microbiome datasets were used as test cases for CBD as a clinically applicable tool. These datasets concerned the human gut microbiome, humanized mouse gut microbiome and human mucosa-associated microbiome. Our study revealed that CBD recaptured 100% of the statistically significant conclusions reported in the previous studies, achieved a decrease in computational time needed when compared to similar tools without expert user intervention. CBD provides a simple, rapid and accurate method for accessing gut microbiome composition for clinical applications related to personalized medicine.

Problem, rationale and goal

Problem: High-throughput sequencing provides a lens for assessing microbially-based personalized genomic medicine. However, analyzing large sequencing datasets for clinically relevant information currently requires exceptional expertise and computational resources.

Rationale: Many diseases ranging from neurological disorders, such as Parkinson, to GIT-related diseases, such as obesity, inflammatory bowel disease (IBD) and irritable bowel syndrome (IBS), are correlated with disturbed microbiomes that differ from those of healthy individuals. The alleviation of symptoms has been achieved using treatments that alter the gut microbiome such as fecal transplants, antibiotics and pro/prebiotics toward that of a healthy individual.



Goal: In order to efficiently access differences in GIT microbiome composition between samples enabling personalized genomic medicine to be used in clinical settings, we developed a simple, rapid, and accurate method called compression-based distance (CBD) to quantitatively analyze similarities between microbiome samples.

Compression-based distance

The similarity between microbial communities can be characterized by the amount of repetition or overlap between them. Compression algorithms make use of repetitive data for more efficient data storage. CBD used the relative compression of combined and individual datasets to produce a distance score to quantify overlaps between two microbial communities (Fig. 1).

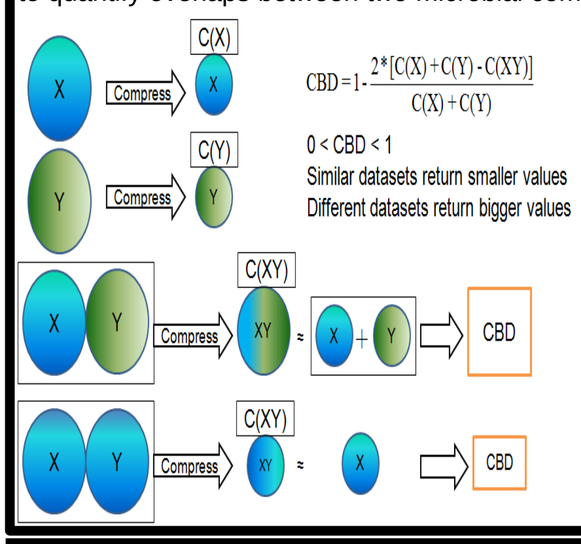


Figure 1. Schematic of how CBD measures distance between two microbiomes. C(X) indicates the compressed size of data X and C(XY) denotes the compressed size of the concatenation of data X and Y.

Test of CBD on artificial datasets

A reliable metric of community distance should return greater values for communities that are more distant and very small values for communities that are virtually the same. In order to test for this, CBD metric was used on artificial datasets which were generated by sampling different proportions of sequences obtained from two individuals (Fig. 2). All artificial datasets were pairwise compared with data 1 using CBD to verify if CBD could reliably assess microbiome similarity (Fig. 3).

Test of CBD on artificial datasets

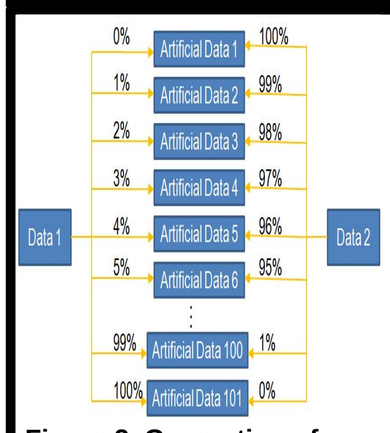


Figure 2. Generation of artificial datasets.

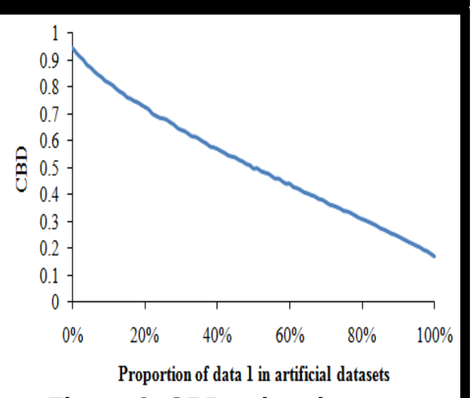


Figure 3. CBD value decreases with the increasing proportion of data 1 in artificial datasets.

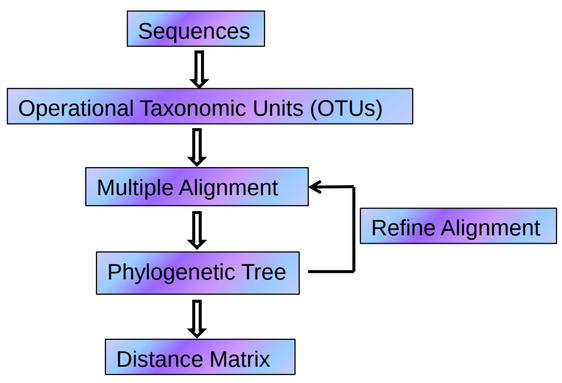
Datasets used in this analysis

1) V2 16S rDNA datasets from a recent study that focused on the GIT microbiome of lean and obese twin pairs and their mothers;[1] 2) V2 16S rDNA datasets from an analysis of the effect of diet switch from low-fat diet to high-fat diet affected humanized gut microbiome composition in mice;[2] and 3) full-length 16S rDNA datasets from mucosa-associated microbiome from inflamed and non-inflamed sites of Crohn's disease (CD) and ulcerative colitis (UC) patients in the colon as well as that from healthy controls[3] were used to test if CBD could successfully recapture the conclusions of previous clinical studies. The first human gut microbiome data was also used to assess the speed of CBD.

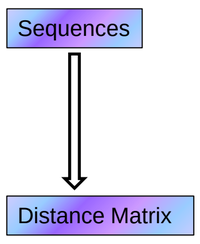
Results

1. **Simple**. CBD directly utilized sequence data, rather than requiring multiple alignments and phylogeny. The need for expert intervention in assigning similar sequences to operational taxonomic units (OTUs) is omitted as well as aligning sequence reads, generating phylogenetic trees, realigning sequence reads and the choosing of proper software and parameters.

Tree-based metric:



CBD:



2. **Rapid.** CBD was shown to be dramatically faster for different tested sizes than the fastest alternative microbiome comparisons suite, QIIME (Fig. 4). Furthermore, the speed advantages of CBD were highlighted as the size of the input files increased (Fig. 4).

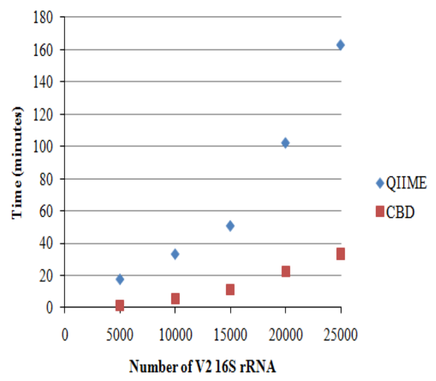


Figure 4. Speed comparisons between CBD and QIIME using GIT microbiome of lean and obese twins.[1]

3. **Accurate.** Comparisons between the results obtained by using CBD captured the analyses of previous work[1-3] in 100% of all cases (Fig. 5-7). The conclusions obtained by CBD using three published gut microbiome datasets were consistent with those from published studies (Fig. 5-7).[1-3]

Unifrac **CBD on all sequences**

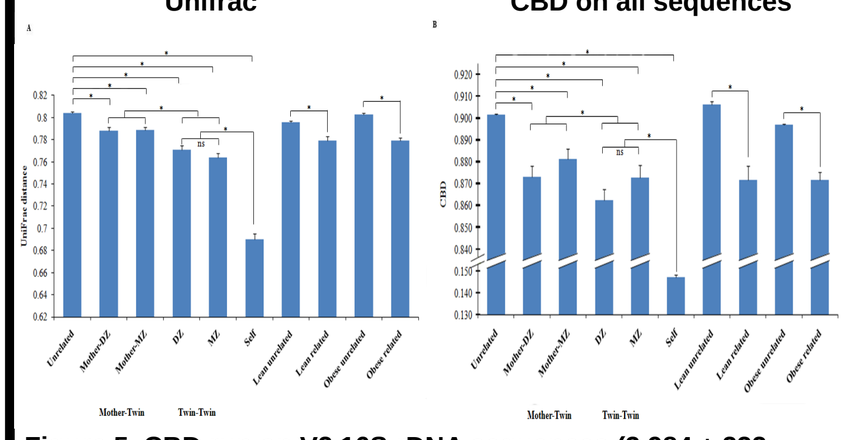


Figure 5. CBD run on V2 16S rDNA sequences (3,984 ± 232 sequences per sample) demonstrated agreement with UniFrac analysis.[1] Analyses on V2 16S rDNA datasets showed that CBD performed well on computing similarities among multiple microbiome categories and in addition could analyze even modest sequence read numbers.

Results

Humanization with fresh feces **Humanization with frozen feces**

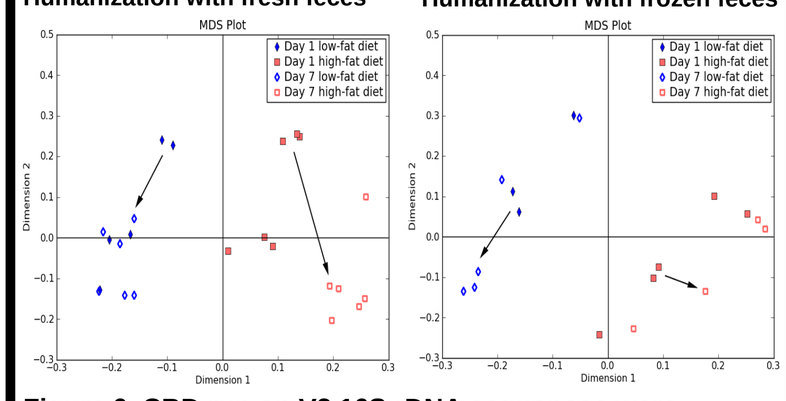
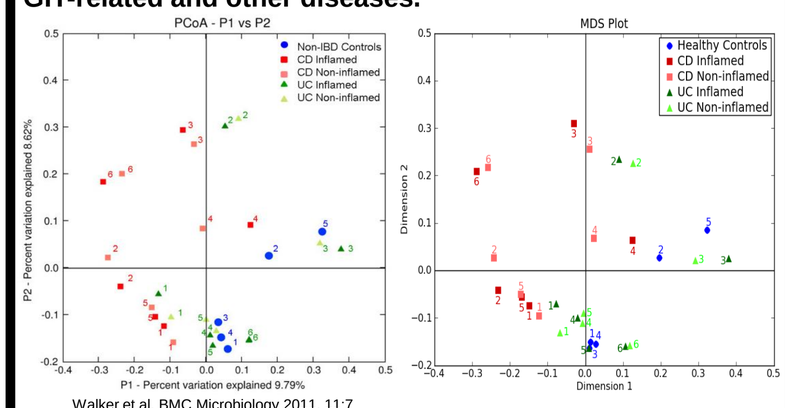


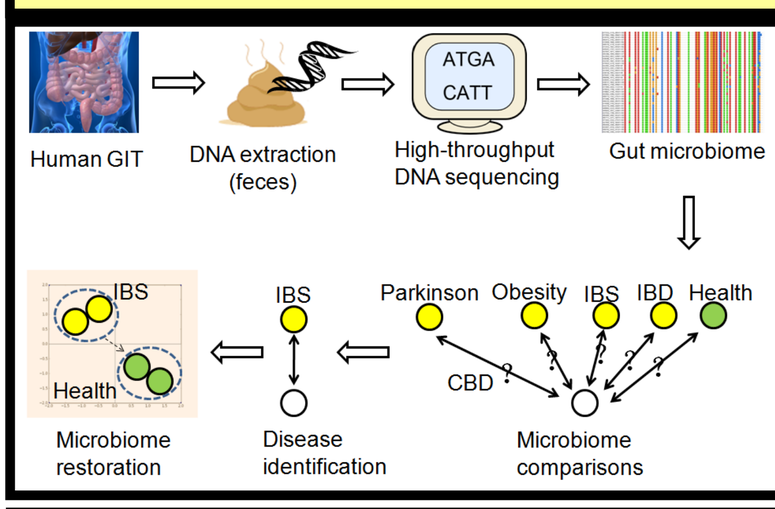
Figure 6. CBD run on V2 16S rDNA sequences were consistent with the analyses using UniFrac.[2] Significant shifts in gut microbiome composition on humanized mice were observed within one day after diet switch. Shifts were more significant on day 7. CBD reliably measures the effect of dietary manipulations on microbiome facilitating the exploration of new treatment strategies for diagnosing GIT-related and other diseases.



Walker et al. BMC Microbiology 2011, 11:7

Figure 7. CBD analyses using full-length 16S rDNA sequences were consistent with the Fast-UniFrac analyses.[3] Inflamed and non-inflamed mucosa-associated microbiome from the same individual clustered together. There was no significant difference in mucosa-associated microbiome between IBD and healthy samples.

Personalized medicine



Conclusions

CBD provides a simple, rapid and accurate method for accessing gut microbiome composition for clinical applications related to personalized medicine. CBD recaptured 100% of the statistically significant conclusions reported in the previous studies, achieved a decrease in computational time needed when compared to similar tools without expert user intervention.

CBD is web-based and freely accessible at http://tornado.igb.uiuc.edu/CBD.html. CBD is copyrighted by the board of trustees of the University of Illinois.

Acknowledgements

Supported by grants AG2008-34480-19328 and 538AG58-5438-7-317l (to Dr. Schook) from United States Department of Agriculture and Agricultural Research Service as well as by the Institute for Genomic Biology (to Dr. Chia).

References

1. Turnbaugh PJ, Hamady M, Yatsunenko T, et al. A core gut microbiome in obese and lean twins. Nature 2009;457:480-4.
2. Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R, Gordon JI. The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. Sci Transl Med 2009;1:6ra14.
3. Walker AW, Sanderson JD, Churcher C, et al. High-throughput clone library analysis of the mucosa-associated microbiota reveals dysbiosis and differences between inflamed and non-inflamed regions of the intestine in inflammatory bowel disease. BMC Microbiol 2011;11:7.