Generative Adversarial Network augmented the gut microbiome-based health index by profoundly improved discrimination power

Yuxue Li [1,$], *Gang Xie* [1,$], *Yuguo Zha* [1,$], *Kang Ning* [1,*]

1 Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-imaging, Center of AI Biology, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China

[$] These authors contributed equally to this work.

* Corresponding author. Tel: +86 27 87793041, e-mail: ningkang@hust.edu.cn (Kang Ning).

Abstract

Summary: Gut microbiome-based health index (GMHI) has been applied with success, while the discrimination powers of GMHI varied for different diseases, limiting its utility on a broad spectrum of diseases. In this work, a Generative Adversarial Network (GAN) model is proposed to improve the discrimination power of GMHI. Built based on the batch corrected data through GAN (https://github.com/HUST-NingKang-Lab/GAN-GMHI), GAN-GMHI has largely reduced the batch effects, and profoundly improved the performance for distinguishing healthy individuals and different diseases. GAN-GMHI has provided results to support the strong association of gut microbiome and diseases, and indicated a more accurate venue towards microbiome-based disease monitoring.

Availability and implementation: GAN-GMHI is publicly available on GitHub: https://github.com/HUST-NingKang-Lab/GAN-GMHI.

Contact: ningkang@hust.edu.cn

Supplementary information: Supplementary data are available at Bioinformatics online.

Introduction

There are important links between many complex chronic diseases and the human gut microbiome (Gupta, et al., 2020). Specific sets of gut microbes could directly or indirectly influence the complex chronic diseases, such as the microbiome dysbiosis in the development of rheumatoid arthritis (Bergot, et al., 2019), thus it is nature that gut microbiome could be utilized for disease

prediction (Gupta, et al., 2020; Shreiner, et al., 2015). However, a general microbiome-based index for prediction of a broad spectrum of diseases is lacking.

A previous work has reported the Gut Microbiome Health Index (GMHI) (Gupta, et al., 2020), a robust index for assessing health status, based on the species-level taxonomic profile of stool metagenomic sequencing samples. GMHI values can be used to classify samples as healthy (GMHI>0), non-healthy (GMHI<0), or neither (GMHI=0), and its results have shown strong reproducibility on the validation datasets. However, GMHI has limited power to distinguish samples from different diseases, largely due to the existence of batch effects: as the stool metagenomes in that study were collected from over 40 published studies, it is nearly impossible to exclude experimental and technical inter-study batch effects (Gupta, et al., 2020). Thus, the overall prediction accuracy of GMHI is still far from perfect: 70.72% for distinguishing healthy individuals and non-healthy individuals.

Therefore, we introduced GAN-GMHI, based on the Generative Adversarial Network (GAN), for improved discrimination power of GMHI. GAN was applied to reduce the batch effects on a large collection of gut microbiome samples from multiple cohorts containing both health and disease individuals. Then GMHI could be applied on the batch corrected data for prediction. Compared with original GMHI, GAN-GMHI makes the distribution of GMHI values within the group more concentrated and the distinction between healthy and non-healthy samples more clearly. The effectiveness of GAN for cross-cohort batch correction has been demonstrated: the prediction accuracy of GAN-GMHI has been improved to 88.70% for distinguishing healthy individuals and non-healthy individuals, compared to the accuracy of 70.95% achieved by GMHI. In summary, batch effect does exist in data sets from different sources, and GMHI can better predict the status of health based on GAN corrected data sets.

Methods

Our GAN-GMHI framework consists of three stages, constructing a dataset containing phenotype and batch information for all samples, and then GAN guiding the batch effect correction of raw data, the corrected datasets are output as the training data set for GMHI prediction (**Supplementary Figure 1**). The batch effect removal method of iMAP (Wang, et al., 2021), a GAN method previously applied on single-cell RNA-Seq data, was adapted for batch effect removal in this study. It is worth noting that the datasets to be batch-corrected by GAN must be classified based on the phenotype first, and the sub-data sets of each phenotype are regrouped according to the batch. To ensure that the unwanted technical variations among different datasets are eliminated, but the biological differences between different phenotypes are not diminished.

Results

We have performed a comprehensive analysis on the integrated dataset of 2,636 healthy and 1,711 diseased (including 12 disease phenotypes) individuals' stool metagenomes from 34 published studies (Gupta, et al., 2020). All of these samples are used as analytical datasets. Additionally, we

have used 679 samples (118 healthy and 561 diseased) as validation datasets (Gupta, et al., 2020). The analytical and validation datasets configuration is the same as in (Gupta, et al., 2020).

We have first assessed and compared prediction results based on the discovery cohort (training data). By comparison of the species-level GMHI before (RAW) and after (GAN) batch correction (**Figure 1 (A)**) for distinguishing samples from healthy and non-healthy individuals, we observed that the accuracy for prediction of the healthy and diseased groups after correction is 87.03% and 91.29% (with overall accuracy of 88.70%), respectively, compared with 75.61% and 63.76% before correction (overall accuracy of 70.95%), which has proven the advantage of GAN-GMHI over GMHI (**Supplementary Table 1**). Additionally, we compared the abilities of GAN-GMHI and Shannon diversity indicators to differentiate the gut microbiome of healthy and non-healthy individuals. The results demonstrated that GAN-GMHI could yield clearer separation compared with Shannon's diversity in differentiating healthy and non-healthy individuals (**Figure 1 (B)**). Furthermore, results on comparison among the healthy group and the 12 non-healthy phenotypes have showed that: when GMHI was applied, the GMHI values were dispersed over a wide range, and GMHI values for healthy samples were slightly higher than those for non-healthy samples except for SA. On the other hand, when GAN-GMHI was applied, the GMHI values were concentrated for each group, and the healthy group was significantly higher than the 12 disease phenotypes (p-value < 0.05 for all disease groups), and the third quartile of GMHI was lower than 0 for all disease phenotypes (**Figure 1 (C)**). Moreover, GMHI's results are easier for clinical interpretation. For example, on Type 2 diabetes (T2D), GAN-GMHI has captured Lactobacillus as biomarkers, which are well founded by published works (Wang, et al.).

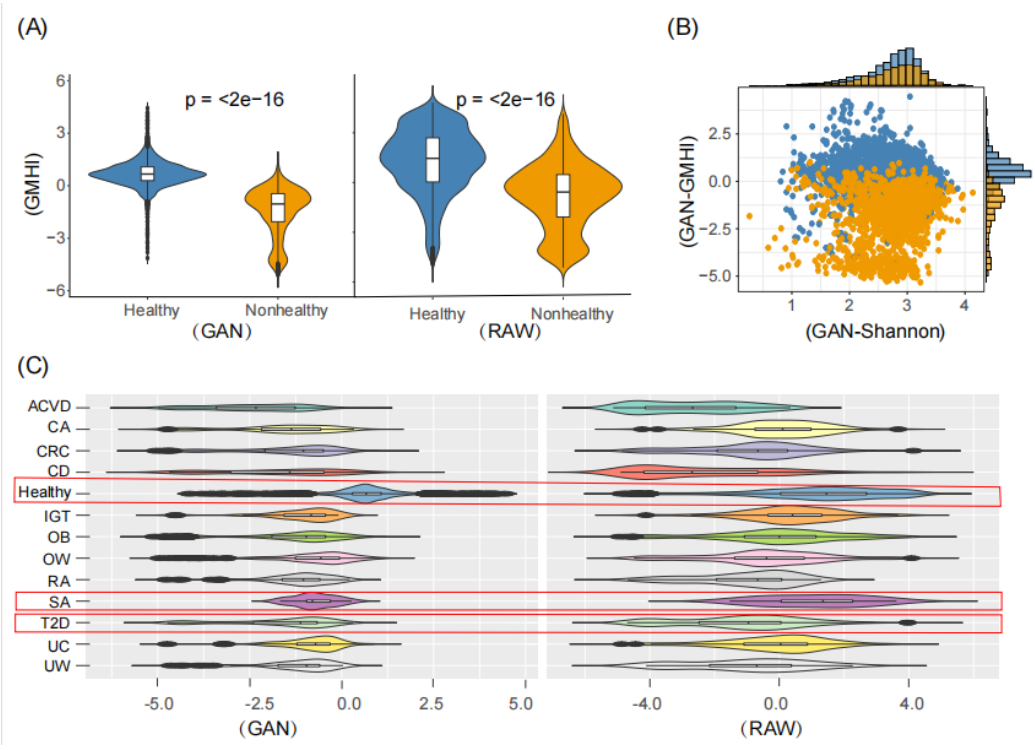


Figure 1. Comparison of GAN-GMHI with other methods under different settings. (A) Violin plots of GMHI for the healthy and non-healthy groups before (left) and after (right) batch

correction by GAN. (B) the distribution of corrected GMHI and Shannon diversity. (C) Violin plots of GMHI index for the healthy and 12 disease phenotypes before (right) and after (left) batch correction by GAN. ACVD: Arteriosclerosis Cardiovascular Disease, CA: colorectal adenoma, CC: colorectal cancer, CD: Crohn's disease, IGT: Impaired glucose tolerance, OB: obesity, OW: overweight, RA: rheumatoid arthritis, SA: Symptomatic arteriosclerosis, T2D: Type 2 Diabetes, UC: ulcerative colitis, UW: underweight.

Additionally, we have compared GAN-GMHI and GMHI on the validation datasets. Cross-cohort batch correction by GAN profoundly improved the performance for distinguishing healthy individuals and different diseases. The prediction accuracy of GAN-GMHI has been significantly improved to 88.70% for distinguishing healthy individuals and non-healthy individuals, compared to the accuracy of 70.95% achieved by GMHI, and GAN-GMHI still outperforms GMHI on the independent validation cohort (**Supplementary Table 1**).

Moreover, GAN is not only applicable for GMHI disease prediction model, but could also be easily adapted to other models, such as Random Forest (RF). It has already been observed that GMHI and RF have similar performances on the validation datasets, while GMHI's results are easier for clinical interpretation (**Supplementary Table 1**). We emphasize that although the results of GAN-GMHI and GAN-RF also have similar accuracies on the validation datasets, GAN-GMHI has inherited the interpretability of the GMHI method, and thus is more suitable for clinical interpretation. For example, on Type 2 diabetes (T2D), GAN-GMHI has captured Lactobacillus as biomarkers, which are well founded by published works (Wang, et al.).

Conclusion

The association of gut microbiome and diseases has been proven for many diseases, while transformation of such association to a robust and universal disease prediction model has remained illusive, largely due to the batch effects presents in multiple microbiome cohorts. GAN-GMHI is a novel method built based on the batch corrected data through GAN, as well as GMHI for prediction of a broad spectrum of diseases. Results have shown that it has largely reduced the batch effects, and profoundly improved the performance for distinguishing disease and healthy individuals. In summary, Generative Adversarial Network augmented the gut microbiome-based health index, and GAN-GMHI has indicated a more accurate venue towards microbiome-based disease monitoring.

Acknowledgments

The authors are grateful to Mingyue Cheng and Hui Chong for insightful discussions.

Funding

This work was partially supported by National Science Foundation of China grant 81774008, 81573702, 31871334, and the Ministry of Science and Technology's national key research and development program grant (No. 2018YFC0910502).

Conflict of Interest: none declared.

References

Bergot, A.-S., Giri, R. and Thomas, R. The microbiome and rheumatoid arthritis. *Best Practice & Research Clinical Rheumatology* 2019;33(6):101497.

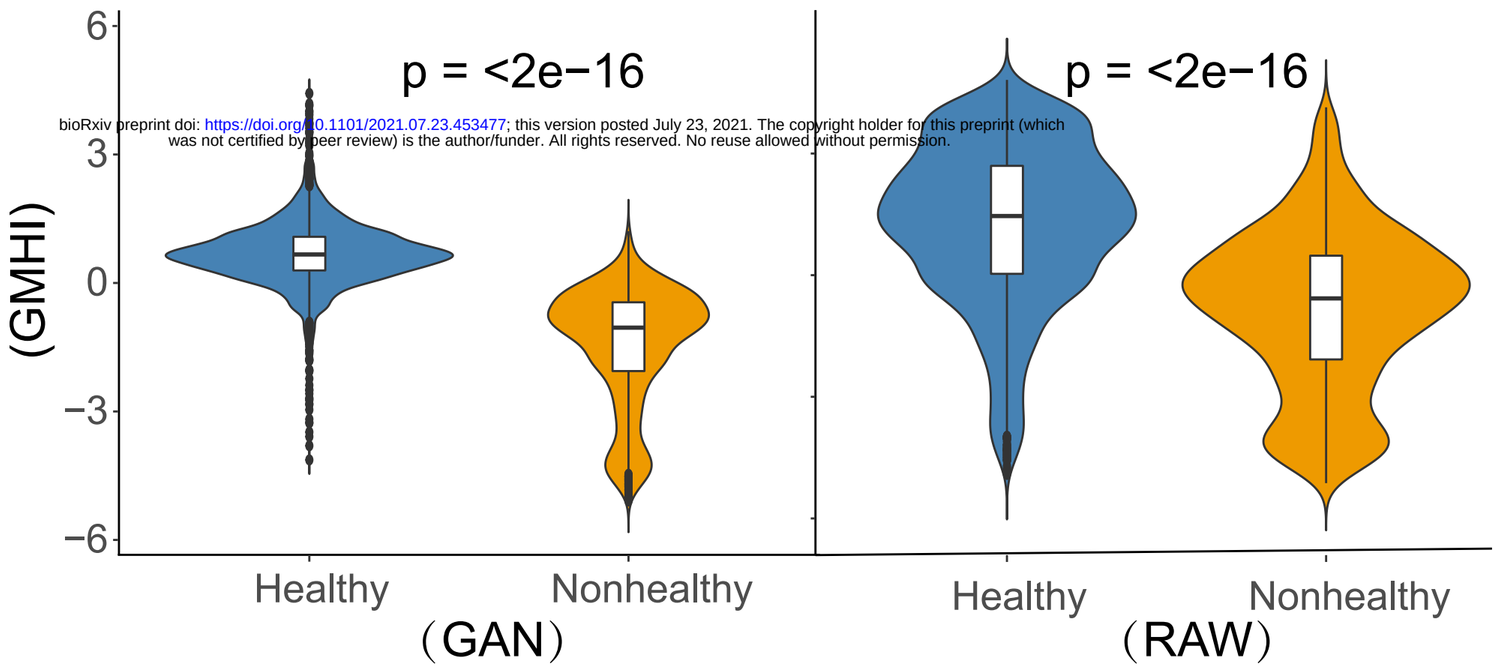
Gupta, V.K., *et al.* A predictive index for health status using species-level gut microbiome profiling. *Nature Communications* 2020;11(1):4635.

Shreiner, A.B., Kao, J.Y. and Young, V.B. The gut microbiome in health and in disease. *Current Opinion in Gastroenterology* 2015;31(1).

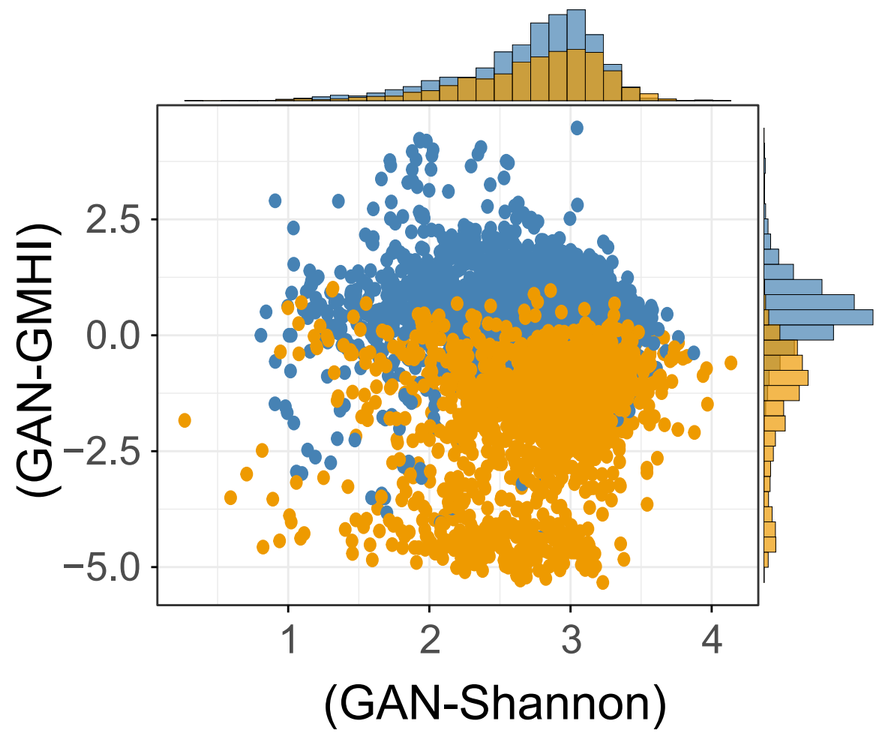
Wang, D., *et al.* iMAP: integration of multiple single-cell datasets by adversarial paired transfer networks. *Genome Biology* 2021;22(1):63.

Wang, Y.A.-O.X., *et al.* Phocea, Pseudoflavonifractor and Lactobacillus intestinalis: Three Potential Biomarkers of Gut Microbiota That Affect Progression and Complications of Obesity-Induced Type 2 Diabetes Mellitus. (1178-7007 (Print)).

(A)



(B)



(C)