PLOS ONE

RESEARCH ARTICLE

A novel approach for predicting microbe-disease associations by bi-random walk on the heterogeneous network

Shuai Zou[☯], **Jingpu Zhang**[☯], **Zuping Zhang***[☯]

School of Information Science and Engineering, Central South University, Changsha, Hunan, China

☯ These authors contributed equally to this work.
* zpzhang@csu.edu.cn

Citation: Zou S, Zhang J, Zhang Z (2017) A novel approach for predicting microbe-disease associations by bi-random walk on the heterogeneous network. PLoS ONE 12(9): e0184394. https://doi.org/10.1371/journal.pone.0184394

Editor: Byung-Jun Yoon, Texas A&M University College Station, UNITED STATES

Received: April 24, 2017

Accepted: August 23, 2017

Published: September 7, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the National Natural Science Foundation of China (61379109, M1321007) (http://www.nsfc.gov.cn/) and Science and Technology Plan of Hunan Province (2014GK2018, 2016JC2011) (http://www.hnst.gov.cn/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Since the microbiome has a significant impact on human health and disease, microbe-disease associations can be utilized as a valuable resource for understanding disease pathogenesis and promoting disease diagnosis and prognosis. Accordingly, it is necessary for researchers to achieve a comprehensive and deep understanding of the associations between microbes and diseases. Nevertheless, to date, little work has been achieved in implementing novel human microbe-disease association prediction models. In this paper, we develop a novel computational model to predict potential microbe-disease associations by bi-random walk on the heterogeneous network (BiRWHMDA). The heterogeneous network was constructed by connecting the microbe similarity network and the disease similarity network via known microbe-disease associations. Microbe similarity and disease similarity were calculated by the Gaussian interaction profile kernel similarity measure; moreover, a logistic function was applied to regulate disease similarity. Additionally, leave-one-out cross validation and 5-fold cross validation were implemented to evaluate the predictive performance of our method; both cross validation methods performed well. The leave-one-out cross validation experiment results illustrate that our method outperforms other previously proposed methods. Furthermore, case studies on asthma and inflammatory bowel disease prove the favorable performance of our method. In conclusion, our method can be considered as an effective computational model for predicting novel microbe-disease associations.

Introduction

There are a large number of microbes in the human body. Research indicates that approximately 90% of the cells in and on the human body are microbial cells [1]. These microbes, including bacteria, eukaryotes, archaea and viruses, reside in and on different body surfaces such as the mouth, skin, vagina and gut, with the vast majority residing in the gastrointestinal tract [2]. These microbes make up an important part of the human body. Recently, due to the impressive advances in metagenomics and metatranscriptomics tools, scientists have begun

Competing interests: The authors have declared that no competing interests exist.

earnestly investigating the human microbiome. For example, the Human Microbiome Project (HMP) was recently launched to explore microbial communities and their relationships with human hosts [1]. The study found that the interaction between human microbiome and cells would affect human health and contribute to the pathogenesis of various diseases [3]. On the one hand, the relationship between humans and the microbiome is symbiotic and mutualistic. For instance, the gut microbiome advances nutrition and energy harvest by fermenting food components that cannot be digested by the host [4]. In addition, the microbiome can help develop the immune system [5, 6], maintain homeostasis [7], and protect against pathogens [8]. On the other hand, there is strong evidence that some microbiomes may lead to various diseases. Recent studies have discovered the associations between body microbiomes and ailments such as cancer [9], diabetes [10, 11], obesity [12–14] and kidney stones [15]. Thus, it is imperative for researchers to achieve a comprehensive understanding of the associations between microbes and diseases, which would not only help determine disease pathogenesis, but also boost disease diagnosis and therapy.

Though some computational methods have recently been proposed to study microorganisms and human diseases [16–18], little work has been undertaken to advance human microbe-disease association prediction models. Until 2016, Ma et al. built the Human Microbe-Disease Association Database (HMDAD) by collecting microbe-disease association data from 61 previous published studies, providing a valuable informational resource for investigating microbe-disease associations. Based on the freely available data, several network based prediction methods have been proposed to achieve microbe-disease association inference. Shen et al. developed RWRHMDA, which applies a random walk with restart algorithm on the heterogeneous network to rank candidate microbes for a specific disease [19]. Chen et al. proposed KATZHMDA to infer potential disease-related microbes by integrating walks of different lengths in the heterogeneous network [20]. Huang et al. introduced PBHMDA to obtain the prediction scores of each candidate microbe-disease pair by evaluating all paths between a microbe and a disease [21]. Meanwhile, during the last few years, the bi-random walk algorithm has been widely used in the field of bioinformatics to address biomedical problems [22–26]. Inspired by its superior performance, we apply bi-random walk algorithm to the study of human microbe-disease associations in the present study. It is a global strategy that explores the missing microbe-disease associations simultaneously, and can predict novel related microbes for diseases without any known associated microbe information.

More specifically, we present a novel computational approach that executes a bi-random walk algorithm on the heterogeneous network to predict potential microbe-disease associations (BiRWHMDA). Based on Gaussian interaction profile kernel similarity and logistic function transformation, we constructed the microbe similarity network and the disease similarity network. Subsequently, the heterogeneous network was constructed by connecting the microbe similarity network and the disease similarity network using the known microbe-disease associations. Then, the bi-random walk algorithm was executed on the heterogeneous network to predict potential microbe-disease associations. Cross validation frameworks are implemented to evaluate the performance of BiRWHMDA. The AUC (the area under of ROC curve) values were 0.8964 and 0.8808 in leave-one-out cross validation (LOOCV) and 5-fold cross validation, respectively. Experiment results of LOOCV demonstrate that our method outperforms other previously proposed methods. Furthermore, case studies of asthma and inflammatory bowel disease (IBD) also demonstrate the favorable performance of our method in predicting novel microbe-disease associations. In summary, BiRWHMDA can be considered as an effective predictive tool for potential microbe-disease associations.

Materials and methods

Dataset

The dataset used in this study (S1 File) was downloaded from the newly built Human Microbe-Disease Association Database (HMDAD, http://www.cuilab.cn/hmdad), which collects human microbe-disease association data from 61 previously published studies. Presently, HMDAD possesses 483 verified microbe-disease association records between 292 microbes and 39 diseases. Here, the microbes are curated at the genus level [27]. However, the set had several duplicate associations; after removing the duplications, we acquired 450 distinct associations and then constructed an adjacency matrix A of the microbe-disease association network. $A(i,j)$ is equal to 1 if there is a known association between disease $d(i)$ and microbe $m(j)$; otherwise, the appropriate coding is 0 [20].

Microbe similarity

To construct the heterogeneous network, the microbe similarity network and the disease similarity network should be separately constructed. Further, we needed to ascertain the similarity between each microbe-microbe pair and each disease-disease pair. In this work, we apply the Gaussian interaction profile kernel similarity measure to determine microbe similarity and disease similarity [28–36].

Based on the assumption that similar microbes are more likely to show a similar interaction and non-interaction pattern with diseases, Gaussian interaction profile kernel similarity for microbes can be calculated from the known microbe-disease association network [33]. The microbe interaction profile $m(i)$ is a binary vector encoding the presence or absence of the associations with each disease in the known microbe-disease association network, defined as the *ith* column of the adjacency matrix A of the microbe-disease association network constructed above. Then, the Gaussian interaction profile kernel similarity between microbe $m(i)$ and $m(j)$ is calculated from their interaction profiles as follows:

$$SM(m(i), m(j)) = \exp(-\gamma_m \|m(i) - m(j)\|^2) \quad (1)$$

The parameter γ_m denotes the normalized kernel bandwidth, which is calculated based on the new kernel bandwidth parameter γ'_m as follows:

$$\gamma_m = \gamma'_m / \left(\frac{1}{n_m} \sum_{k=1}^{n_m} \|m(i)\|^2\right) \quad (2)$$

Here, n_m is the number of microbes and γ'_m is simply set to 1 [20].

Disease similarity

Similar to microbes, Gaussian interaction profile kernel similarity between disease $d(i)$ and $d(j)$ can be defined as follows:

$$KSD(d(i), d(j)) = \exp(-\gamma_d \|d(i) - d(j)\|^2) \quad (3)$$

$$\gamma_d = \gamma'_d / \left(\frac{1}{n_d} \sum_{k=1}^{n_d} \|d(i)\|^2\right) \quad (4)$$

where n_d is the number of diseases and γ'_d is also set to 1.

According to a prior study [37], similarity value ranges in [0, 0.3] are not informative, while similarity value ranges in [0.6, 1] are informative. To improve predictive accuracy, we regulate

disease similarity by applying logistic function transformation. The function is defined as follows:

$$SD(d(i), d(j)) = \frac{1}{1 + e^{c \cdot KSD(d(i),d(j)) + d}} \quad (5)$$

where $KSD(d(i),d(j))$ is the Gaussian interaction profile kernel similarity between diseases, and c and d are parameters that control the adjustment effect. For $KSD(d(i),d(j)) \in [0,0.3]$, $SD(d(i),d(j)) \approx 0$; and for $KSD(d(i),d(j)) \in [0.6,1]$, $SD(d(i),d(j)) \approx 1$. When $KSD(d(i),d(j)) = 0$, we set $SD(d(i),d(j)) = 0.0001$, which set d as log(9999). Vanunu et al. tune the parameter using cross validation and set c = -15 [37]. In the present study, we used the adjusted result, SD, to represent the final disease similarity.

Construction of the heterogeneous network

Based on the microbe similarity and disease similarity calculated above, both the microbe similarity network and the disease similarity network can be constructed. In the microbe similarity network, let $M = \{m(1), m(2), \ldots, m(nm)\}$ denote the node set of nm microbes; the edge between two microbes is weighted by the similarity value of these two microbes. Likewise, in the disease similarity network, let $D = \{d(1), d(2), \ldots, d(nd)\}$ denote the node set of nd diseases, while the edge between two diseases is weighted by the similarity value of these two diseases. We further analyze the frequency distribution of edge weights in each network. Fig 1 indicates that the distribution of edge weights in the disease similarity network is more concentrated after logistic function transformation.

Besides, the microbe-disease association network can be modeled as a bipartite graph [38]. In the bipartite graph, the heterogeneous nodes correspond to either microbes or diseases, and edges denote the presence or absence of the associations between them. If there is a known association between disease $d(i)$ and microbe $m(j)$, the weight of the edge is equal to 1; otherwise 0. To get a comprehensive view of the bipartite graph, we analyze the degree distribution of the microbes and diseases in the microbe-disease association network (Fig 2). It shows the activeness of all nodes in the entire network. On average, each microbe is associated with 1.54 diseases and each disease is associated with 11.54 microbes.

The global heterogeneous network contains above-mentioned two types of nodes (microbes and diseases) and three types of edges between them, which can be constructed by connecting the microbe similarity network and the disease similarity network via the known microbe-disease associations.

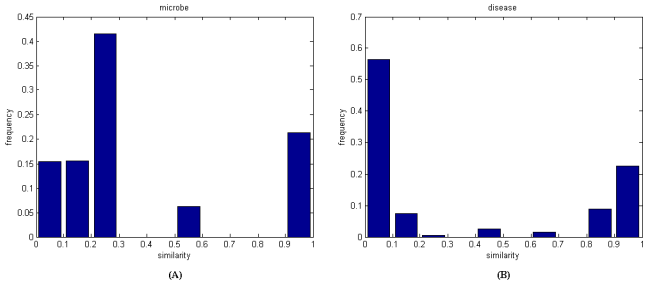


Fig 1. Frequency distribution of microbe similarity and disease similarity. (A) Frequency distribution of microbe similarity. (B) Frequency distribution of disease similarity.

https://doi.org/10.1371/journal.pone.0184394.g001

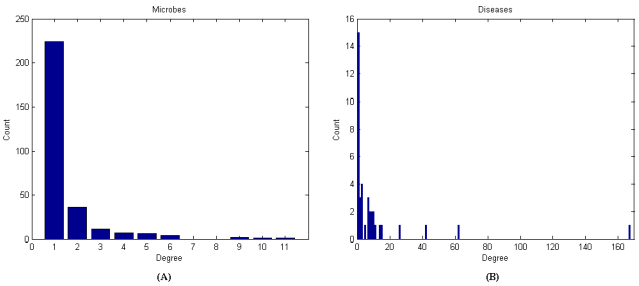
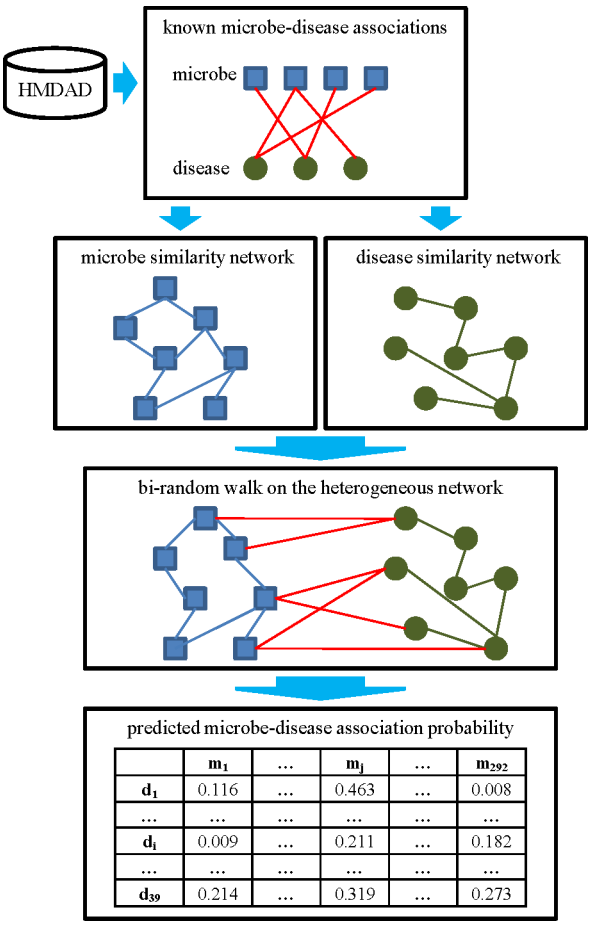


Fig 2. Degree distribution for microbes and diseases in the microbe-disease association network. (A) Degree distribution of microbes. (B) Degree distribution of diseases.

https://doi.org/10.1371/journal.pone.0184394.g002

BiRWHMDA

In this study, we developed a novel computational method of BiRWHMDA to predict human microbe-disease associations. Fig 3 shows the flowchart of BiRWHMDA. Firstly, microbe similarity and disease similarity could be calculated based on the known microbe-disease associations originated from HMDAD. Secondly, the global heterogeneous network was built by combining the microbe similarity network, the disease similarity network and the microbe-



	m_1	...	m_j	...	m_{292}
d_1	0.116	...	0.463	...	0.008
...	...	...	...	...	...
d_i	0.009	...	0.211	...	0.182
...	...	...	...	...	...
d_{39}	0.214	...	0.319	...	0.273

Fig 3. The flowchart of BiRWHMDA.

https://doi.org/10.1371/journal.pone.0184394.g003

disease association network. Finally, the bi-random walk algorithm was performed on the heterogeneous network to obtain the association probability scores between microbes and diseases. The source code for BiRWHMDA is available in S2 File. In the following, we focus on the bi-random walk algorithm for microbe-disease association prediction.

To have a deep understanding of this algorithm, we first introduce the concept of circular bigraph (CBG), which plays an important role in the procedure. A CBG is defined as a subgraph consisting of a microbe path $\{m_1, m_2, \ldots, m_m\}$and a disease path $\{d_1, d_2, \ldots, d_n\}$, with two ends linked by two known microbe-disease associations (m_1, d_1) and (m_m, d_n). The length of a CBG is defined as the length of the longer path of the two paths (Fig 4). A CBG describes a vicinity relation between the associations (m_1, d_1) and (m_m, d_n). Accordingly, a potential microbe-disease association is evaluated by its distance to other associations in the microbe similarity network and the disease similarity network [24].

Bi-random walk explores the CBG patterns by iteratively performing random walk on the microbe similarity network and the disease similarity network simultaneously, to infer novel microbe-disease associations [22]. The CBGs are weighted by a decay factor α, which ranges from 0 to 1; the importance of a CBG is decreased when the path length becomes longer. Nevertheless, the microbe similarity network and the disease similarity network contain diverse topologies and structures, which would generate disparate optimal amounts of random walk steps. To solve this problem, two parameters, l and r, are introduced to restrict steps on the two sides [26]. The iterative process is described as follows:

$$\text{random walk on the microbe similarity network}: \quad R_m = \alpha MD_{t-1} \cdot SM + (1-\alpha)A \tag{6}$$

$$\text{random walk on the disease similarity network}: \quad R_d = \alpha SD \cdot MD_{t-1} + (1-\alpha)A \tag{7}$$

Here, α is the decay factor. $R_d(i,j)$ and $R_m(i,j)$ denote the probability that disease $d(i)$ associates with microbe $m(j)$. The algorithm is detailed in Fig 5:

At the end of the process, the matrix, MD, is acquired as the final prediction result, illustrating the association probability between each microbe and disease pair. For each disease, the potential associated microbes can be ranked according to the prediction probability scores. The top ranked microbes indicate the most relevant associations, potentially providing valuable information for further microbe-disease association research.

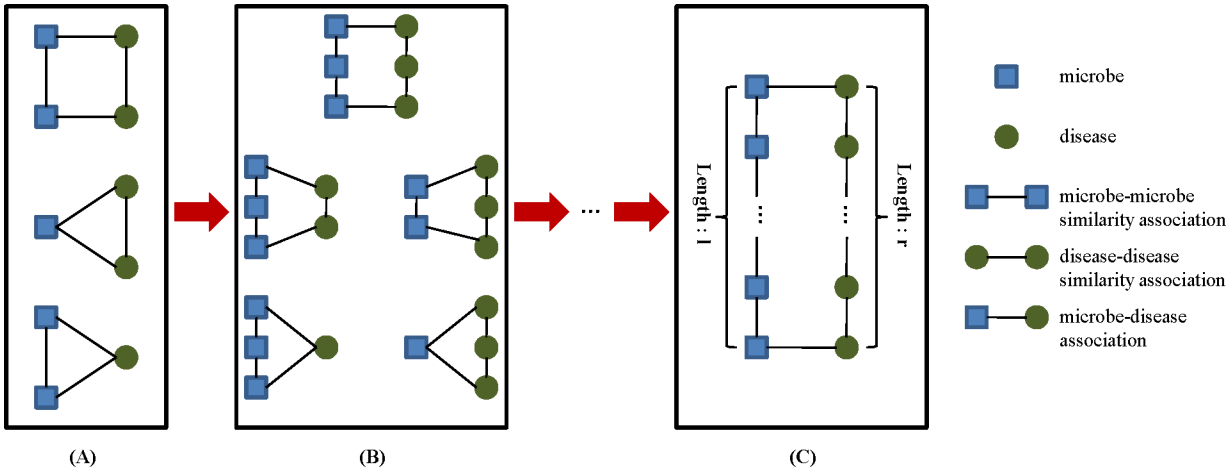


Fig 4. CBGs in the microbe-disease association network. (A) CBG of length 1. (B) CBG of length 2. (C) CBG of length max (l, r).

https://doi.org/10.1371/journal.pone.0184394.g004

Algorithm Bi-random Walk

Input: disease set D and microbe set M, disease-microbe association adjacency matrix A, parameter α, l and r

Output: predicted microbe-disease association matrix MD

Bi-random Walk (D, M, A, α , l, r)

1. Construct disease similarity matrix $SimD$ and microbe similarity matrix $SimM$;
2. $SD= SimD^{-1/2}*D_{SimD}*SimD^{-1/2}$ // Laplacian normalization. $D_{SimD}(i, i)$ is the sum of the ith row of $SimD$
3. $SM= SimM^{-1/2}*D_{SimM}*SimM^{-1/2}$ // Laplacian normalization. $D_{SimM}(i, i)$ is the sum of the ith row of $SimM$
4. $MD_0=A=A/sum(A)$ // MD_0 is the initial probability
5. for t=1 to $max(l, r)$
6. m=n=0;
7. if($t<=l$)
8. $Mm=\alpha*MD_{t-1}*SM+(1-\alpha)*A$ // random walk on the microbe similarity network
9. m=1
10. end if
11. if($t<=r$)
12. $Md=\alpha*SD*MD_{t-1}+(1-\alpha)*A$ // random walk on the disease similarity network
13. n=1
14. end if
15. $MD_t=(m*Mm+n*Md)/(m+n)$ // combination of the results
16. end for
17. return (MD)

Fig 5. Description of algorithm bi-random walk.

https://doi.org/10.1371/journal.pone.0184394.g005

Experiments and results

Performance evaluation

To evaluate the prediction performance of the model we proposed, LOOCV and 5-fold cross validation were implemented on the 450 known microbe-disease associations. In each round of LOOCV, every known microbe-disease association was taken as the test sample, and the other known associations were taken as the training samples [39]. In addition, the microbe similarity and the disease similarity were recalculated at every turn. The predictive performance was evaluated by the rank of the test sample in the candidate samples (all unverified microbe-disease associations) based on their prediction scores. In 5-fold cross validation, the 450 known microbe-disease associations were randomly divided into five subsets. For each trial, one subset is processed as test samples and the other four subsets are processed as training samples; the unverified microbe-disease associations are regarded as candidate samples [40, 41]. Moreover, to reduce potential sample division bias, we performed random divisions 100 times.

A receiver-operating characteristic (ROC) curve, which plots the relationship between the true positive rate (TPR, sensitivity) and the false positive rate (FPR, 1-specificity) by setting different thresholds, was applied to determine the prediction performance. Sensitivity represents the percentage of the test samples that rank higher than the given threshold, while specificity represents the opposite. AUC was also calculated, such that an AUC value of 1 denotes perfect performance, and an AUC value of 0.5 indicates random performance [42–44]. As a result, our model achieves AUC values of 0.8964 and 0.8808 in the LOOCV and 5-fold cross validation frameworks, respectively (Fig 6). While 0.8808 is the average AUC value of 100 operations in 5-fold cross validation, we further obtain the standard deviation of 0.0029. Ultimately, these results confirm the superior performance of this method.

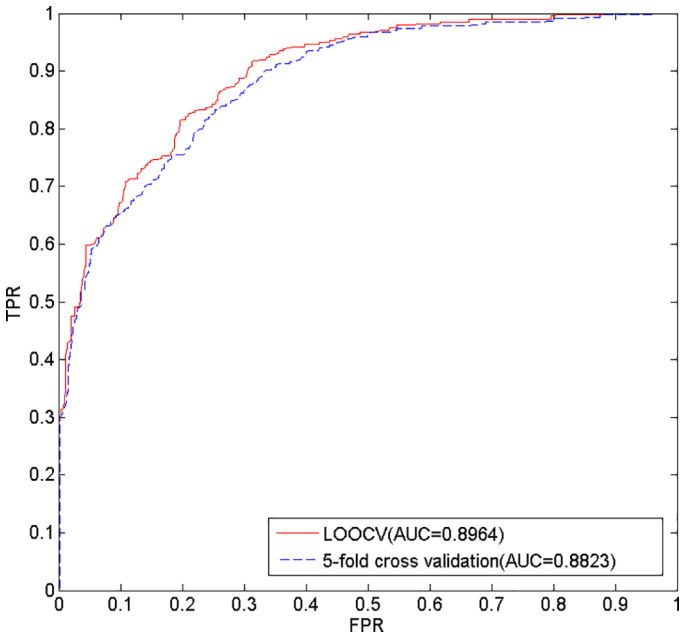


Fig 6. The ROC curve and AUC values of our method.

https://doi.org/10.1371/journal.pone.0184394.g006

Effect of parameters

There are three parameters in our model. The parameter α is the decay factor, which is used to down-weights the importance of a CBG when its path becomes longer. The parameters l and r are introduced to limit the number of random walk steps in the microbe and disease similarity network, respectively. To investigate the effects of the three parameters, we set various values for them and then calculated the AUC values by LOOCV. The details can be seen in Table 1. The experimental results illustrate that BiRWHMDA achieves satisfactory performance when parameter l is equal to r. Taking various parameter combinations into account, we set the three parameters as $\alpha = 0.4$, $l = 2$ and $r = 2$ in our experiment.

Comparison with other methods

To our knowledge, RWRHMDA, KATZHMD and PBHMDA are state-of-the-art computational methods for predicting microbe-disease associations. In considering important differences, RWRHMDA is based on a stochastic process that aims to predict candidate microbes for a disease by calculating the probability of the random walker reaching them [19]; KATZHMDA is based on the KATZ measure that calculates nodes' similarity in the heterogeneous network to solve the problem of link prediction [20]; PBHMDA is a path-based method that utilizes a special depth-first search algorithm in the heterogeneous interlinked network to infer potential microbe-disease associations [21]. These methods are similar in that they are all accomplished based on a heterogeneous network which is constructed by connecting the microbe similarity network and the disease similarity network via the known microbe-disease associations.

Our method, BiRWHMDA, aims to predict novel microbe-disease associations by capturing CBG patterns on the global heterogeneous network. It is a multi-task learning method, which explores the missing microbe-disease associations simultaneously, instead of prioritizing candidate

Table 1. Effect of parameters α, l and r in the results.

$\alpha = 0.2$				
	$r = 1$	$r = 2$	$r = 3$	$r = 4$
$l = 1$	0.8944	0.8631	0.7892	0.7275
$l = 2$	0.8612	0.8952	0.8656	0.7911
$l = 3$	0.8530	0.8610	0.8954	0.8658
$l = 4$	0.8527	0.8529	0.8610	0.8954
$\alpha = 0.4$				
	$r = 1$	$r = 2$	$r = 3$	$r = 4$
$l = 1$	0.8944	0.8807	0.8424	0.7930
$l = 2$	0.8669	0.8964	0.8820	0.8480
$l = 3$	0.8513	0.8653	0.8960	0.8819
$l = 4$	0.8503	0.8511	0.8647	0.8916
$\alpha = 0.6$				
	$r = 1$	$r = 2$	$r = 3$	$r = 4$
$l = 1$	0.8944	0.8880	0.8676	0.8416
$l = 2$	0.8700	0.8966	0.8895	0.8660
$l = 3$	0.8492	0.8670	0.8965	0.8885
$l = 4$	0.8478	0.8483	0.8648	0.8960
$\alpha = 0.8$				
	$r = 1$	$r = 2$	$r = 3$	$r = 4$
$l = 1$	0.8944	0.8930	0.8805	0.8623
$l = 2$	0.8727	0.8969	0.8917	0.8747
$l = 3$	0.8467	0.8667	0.8956	0.8817
$l = 4$	0.8425	0.8428	0.8580	0.8636

https://doi.org/10.1371/journal.pone.0184394.t001

microbes for a specific disease [45–47]. Additionally, BiRWHMDA can predict novel microbes for diseases without any known associated microbe information.

In this study, we implemented these three methods using the same datasets as BiRWHMDA, and then compared their performance by the LOOCV method. Consequently, BiRWHMDA achieves the best performance among all the methods with an AUC value of 0.8964, while RWRHMDA, KATZHMDA and PBHMDA yield AUC values of 0.7254, 0.8382 and 0.8760, respectively (Fig 7). The results demonstrate that BiRWHMDA works better than the other methods, and the predictive performance of BiRWHMDA increases nearly two percentage points higher than the latest method, PBHMDA.

Case studies

We also implemented case studies involving asthma and inflammatory bowel disease (IBD) to further evaluate the ability of our method to predict novel microbe-disease associations. Here, novel associations refer to the microbe-disease pairs that are not known to be associated in the dataset. For each disease, the candidate associated microbes are ranked according to the prediction association scores obtained from BiRWHMDA. We observed microbes from the top 10 candidate microbes confirmed by current research. Furthermore, we compare the results of BiRWHMDA with the latest method, PBHMDA. In this study, we assume that if a microbe is associated with one disease, the genus that the microorganism belongs to is also associated with the disease.

Asthma is a common long-term inflammatory disease of the lung airways. In BiRWHMDA, a total of eight of the predicted microbes in the top 10 candidate microbes have been validated

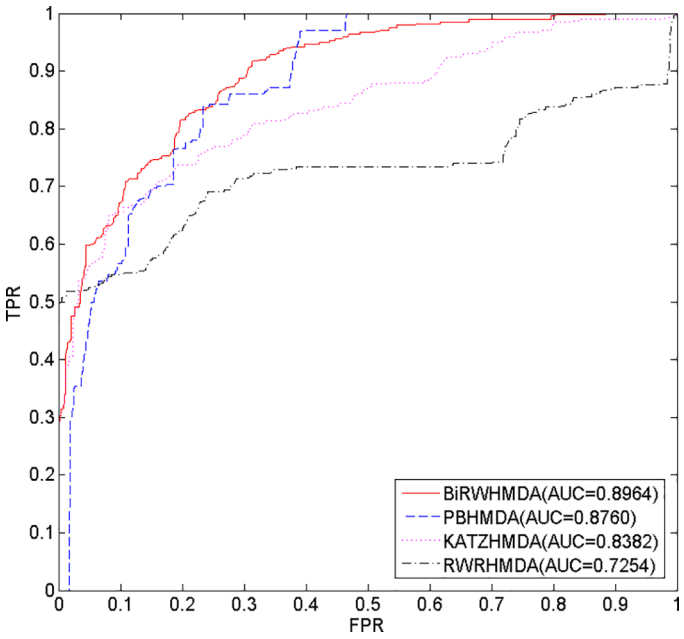


Fig 7. The ROC curve and AUC values of different methods.

https://doi.org/10.1371/journal.pone.0184394.g007

(Table 2). Pseudomonas aeruginosa could cause asthma, which has already been diagnosed by bronchoscopic examination [48]. Lactobacillus rhamnosus is associated with asthma prevention [49]. Colonization by clostridium difficile at 1 month of age is associated with the incidence of asthma between ages 6 and 7 [50]. Firmicutes and actinobacteria are present in lower proportions in asthmatic patients [51]. Clostridium coccoides XIVa species is significantly associated with a positive Asthma Predictive Index (API) [52]. Propionibacterium acnes is more prevalent in asthma patients; therefore, Propionibacterium is also considered to be associated with asthma [53]. Only Burkholderia and Oxalobacter formigenes have not been validated to date. The top 10 candidate microbes of asthma obtained from PBHMDA are also listed in Table 2; nine of these microbes have been previously confirmed [49, 51, 54–59].

IBD is a group of inflammatory conditions of the colon and small intestine. In BiRWHMDA, each of the microbes in the top 10 has been validated (Table 3). There is an

Table 2. Prediction results of associated microbes for disease asthma.

Rank	BiRWHMDA		PBHMDA	
	Microbe	Evidence	Microbe	Evidence
1	Pseudomonas	PMID:13268970	Firmicutes	PMID:23265859
2	Lactobacillus	PMID:20592920	Lactobacillus	PMID:20592920
3	Burkholderia	Unconfirmed	Lachnospiraceae	Lee et al., 2014
4	Clostridium difficile	PMID:21872915	Veillonella	PMID:25329665
5	Firmicutes	PMID:23265859	Bacteroides	PMID:18822123
6	Actinobacteria	PMID:23265859	Bacteroidaceae	Qiu et al., 2013
7	Clostridium coccoides	PMID:21477358	Streptococcus	PMID:17950502
8	Propionibacterium	PMID:27433177	Fusobacterium	Dang et al., 2013
9	Propionibacterium acnes	PMID:27433177	Actinobacteria	PMID:23265859
10	Oxalobacter formigenes	Unconfirmed	Eubacterium	unconfirmed

https://doi.org/10.1371/journal.pone.0184394.t002

Table 3. Prediction results of associated microbes for disease IBD.

Rank	BiRWHMDA		PBHMDA	
	Microbe	Evidence	Microbe	Evidence
1	Helicobacter pylori	PMID:22221289	Bacteroidetes	PMID:25307765
2	Clostridium difficile	Azimirad et al.,2012	Firmicutes	PMID:25307765
3	Clostridium coccoides	PMID:19235886	Veillonella	unconfirmed
4	Bacteroidetes	PMID:25307765	Prevotella	PMID:25307765
5	Firmicutes	PMID:25307765	Haemophilus	unconfirmed
6	Prevotella	PMID:25307765	Bacteroidaceae	Maukonen et al.,2009
7	Staphylococcus aureus	Azimirad et al.,2012	Lactobacillus	PMID:26340825 26340825
8	Bifidobacterium	PMID:24478468	Bacteroides	Maukonen et al.,2009
9	Staphylococcus	Azimirad et al.,2012	Clostridium coccoides	PMID:19235886
10	Clostridia	PMID:25307765	Streptococcus	PMID:23679203

https://doi.org/10.1371/journal.pone.0184394.t003

inverse association between helicobacter pylori and IBD [60]. Research shows a significant relationship between the simultaneous presence of toxigenic strains of staphylococcus aureus and clostridium difficile in IBD patients; staphylococcus is thus validated [14]. Clostridium coccoides are less represented in A-IBD patients [61]. Bacteroidetes, firmicutes, Prevotella and clostridia have been shown to be associated with IBD via the Kruskal-Wallis test [62]. Bifidobacterium shows an increased proportion in IBD [63]. The top 10 candidate microbes for IBD obtained from PBHMDA are also listed in Table 3; of these, eight microbes have been previously validated [61, 62, 64–66].

In summary, these case studies further demonstrate that the approach we proposed is powerful in predicting novel microbe-disease associations. The predictions for all the 39 diseases are listed in S3 File.

Conclusion

A growing body of research suggests that the microbiome plays a vital role in human health and disease. Microbe-disease associations can not only reveal disease pathogenesis but also contribute to disease diagnosis and prognosis [67]. Nevertheless, due to the limited research on existing microbe-disease association data, only a few methods have been developed to address the gap.

In the present study, we proposed a novel approach based on bi-random walk on the heterogeneous network to predict novel microbe-disease associations. The heterogeneous network is constructed by connecting the microbe similarity network and the disease similarity network via the known disease-microbe associations. The measure we utilized to calculate microbe similarity and disease similarity was the Gaussian interaction profile kernel similarity measure. In addition, a logistic function was applied to adjust disease similarity. We sought to obtain the predictive association scores between each microbe and disease pair through BiRWHMDA. For each disease, the top ranked microbes are considered the most probable associated microbes. Cross validation frameworks, including LOOCV and 5-fold cross validation, were also implemented to evaluate predictive performance of our approach. Moreover, the approach was compared with three other state-of-the-art methods by using LOOCV. Ultimately, our method obtained better performance than these competing methods. Additionally, we implemented case studies for asthma and IBD to evaluate the predictive performance of BiRWHMDA. In total, eight and ten of the predicted microbes in the top 10 microbe candidates have been confirmed by recent studies. Our method demonstrated favorable utility in predicting novel microbe-disease associations.

Despite the current success, there are still some limitations that can be improved in future studies. First, only one database exists: the HMDAD, which contains 483 verified microbe-disease association records. Therefore, predictive performance will be certainly limited due to the lack of available experimental data. This could be solved through an increase in microbe-disease associations discovered in the future. In addition, microbe and disease similarity are calculated based solely on known microbe-disease associations, which could cause bias for microbes and diseases already extant in the database. Data from different sources should be integrated to improve the completeness and quality of the experimental data, which would ultimately be conductive to improving predictive performance.

Supporting information

S1 File. The dataset explored in this work.
(ZIP)

S2 File. The source code for BiRWHMDA.
(ZIP)

S3 File. The prediction results for each disease.
(ZIP)

Acknowledgments

We would like to thank the anonymous referees for their constructive comments and suggestions.

Author Contributions

Conceptualization: Jingpu Zhang.

Data curation: Jingpu Zhang.

Formal analysis: Shuai Zou, Jingpu Zhang.

Funding acquisition: Zuping Zhang.

Investigation: Shuai Zou.

Methodology: Jingpu Zhang.

Project administration: Zuping Zhang.

Resources: Zuping Zhang.

Software: Shuai Zou, Jingpu Zhang.

Supervision: Zuping Zhang.

Validation: Shuai Zou, Jingpu Zhang.

Visualization: Shuai Zou.

Writing – original draft: Shuai Zou.

Writing – review & editing: Shuai Zou, Jingpu Zhang, Zuping Zhang.

References

1. Althani A, Marei HE, Hamdi WS, Nasrallah GK, El Zowalaty ME, Al KS, et al. Human Microbiome and Its Association With Health and Diseases. Journal of Cellular Physiology. 2015; 231(8):1688–94.

2. Holmes E, Wijeyesekera A, Taylorrobinson SD, Nicholson JK. The promise of metabolic phenotyping in gastroenterology and hepatology. Nature Reviews Gastroenterology & Hepatology. 2015; 12(8):458–71.
3. Methé BA, Nelson KE, Pop M, Creasy HH, Giglio MG, Huttenhower C, et al. A framework for human microbiome research. Nature. 2012; 486(7402):215–21. https://doi.org/10.1038/nature11209 PMID: 22699610
4. Kau AL, Ahern PP, Griffin NW, Goodman AL, Gordon JI. Human nutrition, the gut microbiome and the immune system. Nature. 2011; 474(7351):327–36. https://doi.org/10.1038/nature10213 PMID: 21677749
5. Round JL, Mazmanian SK. Inducible Foxp3+ regulatory T-cell development by a commensal bacterium of the intestinal microbiota. Proceedings of the National Academy of Sciences of the United States of America. 2010; 107(27):12204–9. https://doi.org/10.1073/pnas.0909122107 PMID: 20566854
6. Gollwitzer ES, Saglani S, Trompette A, Yadava K, Sherburn R, McCoy KD, et al. Lung microbiota promotes tolerance to allergens in neonates via PD-L1. Nature Medicine. 2014; 20(6):642–7. https://doi.org/10.1038/nm.3568 PMID: 24813249
7. Bouskra D, Brézillon C, Bérard M, Werts C, Varona R, Boneca IG, et al. Lymphoid tissue genesis induced by commensals through NOD1 regulates intestinal homeostasis. Nature. 2008; 456 (7221):507–10. https://doi.org/10.1038/nature07450 PMID: 18987631
8. Kreth J, Zhang Y, Herzberg MC. Streptococcal antagonism in oral biofilms: Streptococcus sanguinis and Streptococcus gordonii interference with Streptococcus mutans. Journal of Bacteriology. 2008; 190 (13):4632–40. https://doi.org/10.1128/JB.00276-08 PMID: 18441055
9. Moore WE, Moore LH. Intestinal floras of populations that have a high risk of colon cancer. Applied & Environmental Microbiology. 1995; 61(9):3202–7.
10. Brown CT, Davisrichardson AG, Giongo A, Gano KA, Crabb DB, Mukherjee N, et al. Gut Microbiome Metagenomics Analysis Suggests a Functional Model for the Development of Autoimmunity for Type 1 Diabetes. Plos One. 2011; 6(10):e25792. https://doi.org/10.1371/journal.pone.0025792 PMID: 22043294
11. Giongo A, Gano KA, Crabb DB, Mukherjee N, Novelo LL, Casella G, et al. Toward defining the autoimmune microbiome for type 1 diabetes. Isme Journal Multidisciplinary Journal of Microbial Ecology. 2011; 5(1):82–91.
12. Zhang H, Dibaise JK, Zuccolo A, Kudrna D, Braidotti M, Yu Y, et al. Human gut microbiota in obesity and after gastric bypass. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106(7):2365–70. https://doi.org/10.1073/pnas.0812600106 PMID: 19164560
13. Ley RE, Bäckhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI. Obesity alters gut microbial ecology. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102 (31):11070–5. https://doi.org/10.1073/pnas.0504978102 PMID: 16033867
14. Azimirad M, Bahreiny R, Hasani Z, Molaei M, Rashidan M, Zali M, et al. Prevalence of superantigenic Staphylococcus aureus and toxigenic Clostridium difficile in patients with IBD. 2012.
15. Hoppe B, Groothoff JW, Hulton SA, Cochat P, Niaudet P, Kemper MJ, et al. Efficacy and safety of Oxalobacter formigenes to reduce urinary oxalate in primary hyperoxaluria. Nephrology, dialysis, transplantation: official publication of the European Dialysis and Transplant Association—European Renal Association. 2011; 26(11):3609–15.
16. Edgar D C, AndrÉ M S, Joel P A, JosÉ LuÍS O. Computational methodology for predicting the landscape of the human-microbial interactome region level influence. Journal of Bioinformatics & Computational Biology. 2015; 13(5):1550023.
17. Nayfach S, Fischbach MA, Pollard KS. MetaQuery: a web server for rapid annotation and quantitative analysis of specific genes in the human gut microbiome. Bioinformatics. 2015; 31(20):3368–70. https://doi.org/10.1093/bioinformatics/btv382 PMID: 26104745
18. Cao Y, Zheng X, Li F, Bo X. mmnet: An R Package for Metagenomics Systems Biology Analysis. Biomed Research International. 2015; 2015(7402):1–5.
19. Shen X, Chen Y, Jiang X, Hu X, He T, Yang J, editors. Predicting disease-microbe association by random walking on the heterogeneous network. IEEE International Conference on Bioinformatics and Biomedicine; 2016.
20. Chen X, Huang YA, You ZH, Yan GY, Wang XS. A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. Bioinformatics. 2017; 33(5):733–9. https://doi.org/10.1093/bioinformatics/btw715 PMID: 28025197
21. Huang ZA, Chen X, Zhu Z, Liu H, Yan GY, You ZH, et al. PBHMDA: Path-Based Human Microbe-Disease Association Prediction. Frontiers in Microbiology. 2017; 8(2):233.

22. Luo H, Wang J, Li M, Luo J, Peng X, Wu FX, et al. Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. Bioinformatics. 2016; 32(17):2664–71. https://doi.org/10.1093/bioinformatics/btw228 PMID: 27153662

23. Chen X. miREFRWR: a novel disease-related microRNA-environmental factor interactions prediction method. Molecular Biosystems. 2016; 12(2):624–33. https://doi.org/10.1039/c5mb00697j PMID: 26689259

24. Xie MQ, Xu YJ, Zhang YG, Hwang TH, Kuang R. Network-based Phenome-Genome Association Prediction by Bi-Random Walk. Plos One. 2015; 10(5):e0125138. https://doi.org/10.1371/journal.pone.0125138 PMID: 25933025

25. Luo J, Xiao Q. A novel approach for predicting microRNA-disease associations by unbalanced bi-random walk on heterogeneous network. Journal of Biomedical Informatics. 2017; 66:194–203. https://doi.org/10.1016/j.jbi.2017.01.008 PMID: 28104458

26. Xie M, Hwang T, Kuang R, editors. Prioritizing Disease Genes by Bi-Random Walk. Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining; 2012.

27. Ma W, Zhang L, Zeng P, Huang C, Li J, Geng B, et al. An analysis of human microbe-disease associations. Briefings in Bioinformatics. 2017; 18(1):85–97. https://doi.org/10.1093/bib/bbw005 PMID: 26883326

28. Van LT, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug-target interaction. Bioinformatics. 2011; 27(21):3036–43. https://doi.org/10.1093/bioinformatics/btr500 PMID: 21893517

29. Chen X, Yan CC, Zhang X, You ZH, Huang YA, Yan GY. HGIMDA: Heterogeneous graph inference for miRNA-disease association prediction. Oncotarget. 2016; 7(40):65257–69. https://doi.org/10.18632/oncotarget.11251 PMID: 27533456

30. Chen X, You ZH, Yan GY, Gong DW. IRWRLDA: improved random walk with restart for lncRNA-disease association prediction. Oncotarget. 2016; 7(36):57919–31. https://doi.org/10.18632/oncotarget.11141 PMID: 27517318

31. Chen X. KATZLDA: KATZ measure for the lncRNA-disease association prediction. Scientific Reports. 2015; 5:16840. https://doi.org/10.1038/srep16840 PMID: 26577439

32. Chen X, Jiang ZC, Xie D, Huang DS, Zhao Q, Yan GY, et al. A novel computational model based on super-disease and miRNA for potential miRNA-disease association prediction. Molecular Biosystems. 2017; 13(6):1202–12. https://doi.org/10.1039/c6mb00853d PMID: 28470244

33. Chen X, Yan GY. Novel human lncRNA-disease association inference based on lncRNA expression profiles. Bioinformatics. 2013; 29(20):2617–24. https://doi.org/10.1093/bioinformatics/btt426 PMID: 24002109

34. You ZH, Huang ZA, Zhu Z, Yan GY, Li ZW, Wen Z, et al. PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction. Plos Computational Biology. 2017; 13(3):e1005455. https://doi.org/10.1371/journal.pcbi.1005455 PMID: 28339468

35. Chen X, Wu QF, Yan GY. RKNNMDA: Ranking-based KNN for MiRNA-Disease Association prediction. RNA Biology. 2017; 14(7):952–62. https://doi.org/10.1080/15476286.2017.1312226 PMID: 28421868

36. Chen X, Yan CC, Zhang X, You ZH, Deng L, Liu Y, et al. WBSMDA: Within and Between Score for MiRNA-Disease Association prediction. Scientific Reports. 2016; 6:21106. https://doi.org/10.1038/srep21106 PMID: 26880032

37. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating Genes and Protein Complexes with Disease via Network Propagation. Plos Computational Biology. 2010; 6(1):e1000641. https://doi.org/10.1371/journal.pcbi.1000641 PMID: 20090828

38. Ju Y, Zhang S, Ding N, Zeng X, Zhang X. Complex Network Clustering by a Multi-objective Evolutionary Algorithm Based on Decomposition and Membrane Structure. Scientific Reports. 2016; 6:33870. https://doi.org/10.1038/srep33870 PMID: 27670156

39. Zou Q, Li J, Wang C, Zeng X. Approaches for Recognizing Disease Genes Based on Network. Biomed Research International. 2014; 2014(5013):416323.

40. Li JQ, Rong ZH, Chen X, Yan GY, You ZH. MCMDA: Matrix completion for MiRNA-disease association prediction. Oncotarget. 2017; 8(13):21187–99. https://doi.org/10.18632/oncotarget.15061 PMID: 28177900

41. Zeng X, Zhang X, Liao Y, Pan L. Prediction and validation of association between microRNAs and diseases by multipath methods. Biochimica et biophysica acta. 2016; 1860(11):2735–9.

42. Sun D, Li A, Feng H, Wang M. NTSMDA: prediction of miRNA-disease associations by integrating network topological similarity. Molecular Biosystems. 2016; 12(7):2224–32. https://doi.org/10.1039/c6mb00049e PMID: 27153230

43. Zeng X, Liao Y, Liu Y, Zou Q. Prediction and validation of disease genes using HeteSim Scores. IEEE/ACM Transactions on Computational Biology & Bioinformatics. 2016;PP(99):1.
44. Zou Q, Wan S, Ju Y, Tang J, Zeng X. Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. BMC Systems Biology. 2016; 10(4):114.
45. Chen X, Liu MX, Yan GY. Drug-target interaction prediction by random walk on the heterogeneous network. Molecular Biosystems. 2012; 8(7):1970–8. https://doi.org/10.1039/c2mb00002d PMID: 22538619
46. Chen X, Liu MX, Yan GY. RWRMDA: predicting novel human microRNA-disease associations. Molecular Biosystems. 2012; 8(10):2792–8. https://doi.org/10.1039/c2mb25180a PMID: 22875290
47. Zou Q, Li J, Song L, Zeng X, Wang G. Similarity computation strategies in the microRNA-disease network: a survey. Briefings in Functional Genomics. 2016; 15(1):55–64. https://doi.org/10.1093/bfgp/elv024 PMID: 26134276
48. Fein BT. Bronchial asthma caused by Pseudomonas aeruginosa diagnosed by bronchoscopic examination. Annals of Allergy. 1955; 13(6):639–41. PMID: 13268970
49. Yu JH, Seongok J, Byoungju K, Song YH, Jiwon K, Kang MJ, et al. The effects of Lactobacillus rhamnosus on the prevention of asthma in a murine model. Allergy Asthma & Immunology Research. 2010; 2 (3):199–205.
50. van Nimwegen FA, Penders J, Stobberingh EE, Postma DS, Koppelman GH, Kerkhof M, et al. Mode and place of delivery, gastrointestinal microbiota, and their influence on asthma and atopy. Journal of Allergy & Clinical Immunology. 2011; 128(5):948–55.
51. Marri PR, Stern DA, Wright AL, Billheimer D, Martinez FD. Asthma-associated differences in microbial composition of induced sputum. Journal of Allergy & Clinical Immunology. 2013; 131(2):346–52.
52. Vael C, Vanheirstraeten L, Desager KN, Goossens H. Denaturing gradient gel electrophoresis of neonatal intestinal microbiota in relation to the development of asthma. BMC Microbiology. 2011; 11(1):68.
53. Jae-Woo J, Jae-Chol C, Jong-Wook S, Jae-Yeol K, In-Won P, Whui CB, et al. Lung Microbiome Analysis in Steroid-Naïve Asthma Patients by Using Whole Sputum. Tuberculosis & Respiratory Diseases. 2016; 79(3):165–78.
54. Dang HT, Song AK, Park HK, Shin JW, Park SG, Kim W. Analysis of Oropharyngeal Microbiota between the Patients with Bronchial Asthma and the Non-Asthmatic Persons. Journal of Bacteriology & Virology. 2013; 43(4):270–8.
55. Qiu R, Li N, Yang Z, He M, Xin F, Li J, et al. Analysis of the Sputum Microbiome in the Severe Asthma. Chest. 2016; 149(4):A14.
56. Vael C, Nelen V, Verhulst SL, Goossens H, Desager KN. Early intestinal Bacteroides fragilis colonisation and development of asthma. BMC Pulmonary Medicine. 2008; 8(1):19.
57. Lee E, Hong SA, Yang SI, Kim KW, Shin YH, Kang MA, et al. The Home Microbiome and Childhood Asthma. Retour Au Numéro. 2014; 133(2):AB70.
58. Preston JA, Essilfie AT, Horvat JC, Wade MA, Beagley KW, Gibson PG, et al. Inhibition of allergic airways disease by immunomodulatory therapy with whole killed Streptococcus pneumoniae. Vaccine. 2007; 25(48):8154–62. https://doi.org/10.1016/j.vaccine.2007.09.034 PMID: 17950502
59. Park HK, Shin JW, Park SG, Kim W. Microbial Communities in the Upper Respiratory Tract of Patients with Asthma and Chronic Obstructive Pulmonary Disease. Plos One. 2014; 9(9):e109710.
60. Sonnenberg A, Genta RM. Low prevalence of Helicobacter pylori infection among patients with inflammatory bowel disease. Alimentary Pharmacology & Therapeutics. 2012; 35(4):469–76.
61. Sokol H, Seksik P, Furet JP, Firmesse O, Nionlarmurier I, Beaugerie L, et al. Low counts of Faecalibacterium prausnitzii in colitis microbiota. Inflammatory Bowel Diseases. 2009; 15(8):1183–9. https://doi.org/10.1002/ibd.20903 PMID: 19235886
62. Walters WA, Xu Z, Knight R. Meta-analyses of human gut microbes associated with obesity and IBD. Febs Letters. 2014; 588(22):4223–33. https://doi.org/10.1016/j.febslet.2014.09.039 PMID: 25307765
63. Wang W, Chen L, Zhou R, Wang X, Song L, Huang S, et al. Increased proportions of Bifidobacterium and the Lactobacillus group and loss of butyrate-producing bacteria in inflammatory bowel disease. Journal of Clinical Microbiology. 2014; 52(2):398–406. https://doi.org/10.1128/JCM.01500-13 PMID: 24478468
64. Kojima A, Nomura R, Ooshima T, Nakano K. Aggravation of Inflammatory Bowel Diseases by Streptococcus sanguinis. Oral Diseases. 2014; 20(4):359–66. https://doi.org/10.1111/odi.12125 PMID: 23679203
65. Thomas M, Langella P, Neyrolles O. Lactobacillus acidophilus: a promising tool for the treatment of inflammatory bowel diseases. Medecine Sciences M/s. 2015; 31(8–9):715–7. https://doi.org/10.1051/medsci/20153108004 PMID: 26340825

66. Maukonen J, Klemetti P, Vaarala O, Saarela M. Paediatric patients with inflammatory bowel disease have significantly reduced diversity in Bacteroides flagilis group, Clostridium leptum group, and bifidobacteria as compared to healthy children.

67. Szafranski SP, Wos-Oxley ML, Vilchez-Vargas R, Jáuregui R, Plumeier I, Klawonn F, et al. High-resolution taxonomic profiling of the subgingival microbiome for biomarker discovery and periodontitis diagnosis. Applied & Environmental Microbiology. 2015; 81(3):1047–58.