

## Overcoming regional limitations: transfer learning for cross-regional microbial-based diagnosis of diseases

With great interest, we have read the article by Clooney *et al*, which highlighted the regional effects on the heterogeneity of the gut microbiota among populations with inflammatory bowel disease (IBD).<sup>1</sup> As a result, regional effects would largely limit the microbial-based diagnosis of diseases across regions. Although current machine learning methods based on microbial features have been applied to diagnosis of diseases such as IBD<sup>2</sup> and type 2 diabetes,<sup>3</sup> these methods are unable to mitigate the regional effects and meet the demand of microbial-based cross-regional diagnosis of diseases.

Here, we proposed a machine learning framework (online supplemental figure S1, accessible at: <https://github.com/HUST-NingKang-Lab/EXPERT-Disease-GGMP>), which integrated the neural network and transfer learning, to effectively reduce regional effects for microbial-based cross-regional diagnosis. Importantly, transfer learning can 'borrow' the mature knowledge about diseases from a source city to assist the disease diagnosis for a target city, especially when there is little knowledge about microbiota patterns in the target city.<sup>4</sup>

To assess the framework, we obtained genus-level taxonomy profiles from the Guangdong Gut Microbiome Project.<sup>5</sup> These samples were collected from 14 cities and seven representative diseases were selected for assessment (figure 1A and online supplemental table S1). We randomly divided samples of each city into the training subset and the testing subset (80%:20% by default), then performed assessments for three models: (1) Independent disease neural network (DNN) model: *ab initio* training and testing the DNN model on the training subset and the testing subset of each city, respectively. (2) Regional DNN model: *ab initio* training the DNN model using the training subset of one city A (source city) and testing it on the testing subset of another city B (target city). (3) Transfer DNN model: *ab initio* training the DNN model using training subset of one city A, followed by applying transfer learning to a

certain proportion (from 20% to 80%) of samples from city B, and then testing the transfer DNN model on the testing subset of city B (figure 1B and online supplemental figure S1).

We found that the regional DNN model across cities presented a low average accuracy of 0.506 compared with the independent DNN model with an average accuracy of 0.743 ( $p_{\text{Wilcoxon}} = 2.22 \times 10^{-16}$ ; figure 1C and online supplemental figure S2). It suggested that regional factors largely limited the cross-regional diagnosis, as also indicated in previous studies.<sup>5</sup> However, the transfer DNN model profoundly increased prediction accuracy across cities with an average accuracy of 0.829 ( $p_{\text{Wilcoxon}} = 2.22 \times 10^{-16}$ , compared with the independent DNN model; figure 1C and online supplemental figure S2). Intriguingly, once the proportion of samples used in the target city exceeded 50% for transfer learning, the transfer DNN model could even present higher prediction accuracy than that of the independent DNN model (figure 1D). Furthermore, the transfer DNN models also had good performance when we have applied this approach on two inter-continental cohorts (online supplemental figures S3 and S4).

Moreover, our machine learning framework is advantageous in identification of region-specific microbes, as well as microbes shared across all regions. We used the 'leave-one-feature-out' method to discover certain microbes which were strongly affected by regions, such as Enterobacteriaceae and *Clostridium*, while others were less affected by regions, such as *Parabacteroides* and *Faecalibacterium* (online supplemental table S2). We speculated that the region-specific microbes may contribute to the effectiveness of the transfer DNN model in the cross-regional diagnosis of diseases.

Collectively, our study demonstrates that transfer learning can realise microbial-based cross-regional diagnosis of diseases with high accuracy and robustness, by using knowledge about microbial features across regions. This study provides a new venue to exceed the regional limitation, and facilitate microbial-based cross-regional diagnosis of diseases in clinical trials by artificial intelligence techniques.

Data accession: metagenomic sequencing samples are available in the European Bioinformatics Institute (EBI) database of European Molecular Biology Laboratory (EBI accession number

PRJEB18535) at <https://www.ebi.ac.uk/ena/browser/view/PRJEB18535>.

Nan Wang , Mingyue Cheng , Kang Ning 

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

**Correspondence to** Professor Kang Ning, Department of Bioinformatics and Systems Biology, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China; [ningkang@hust.edu.cn](mailto:ningkang@hust.edu.cn)

**Contributors** NW and KN designed the study, conceived and proposed the idea. NW performed the experiments and analysed the data. NW visualised the data. NW, MC and KN contributed to editing and proofreading the manuscript. All authors read and approved the final manuscript.

**Funding** This work was partially supported by the National Natural Science Foundation of China (Grant Nos. 32071465, 31871334, and 31671374), and the National Key R&D Programme of China (Grant No. 2018YFC0910502).

**Competing interests** None declared.

**Patient consent for publication** Not applicable.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Supplemental material** This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.



### OPEN ACCESS

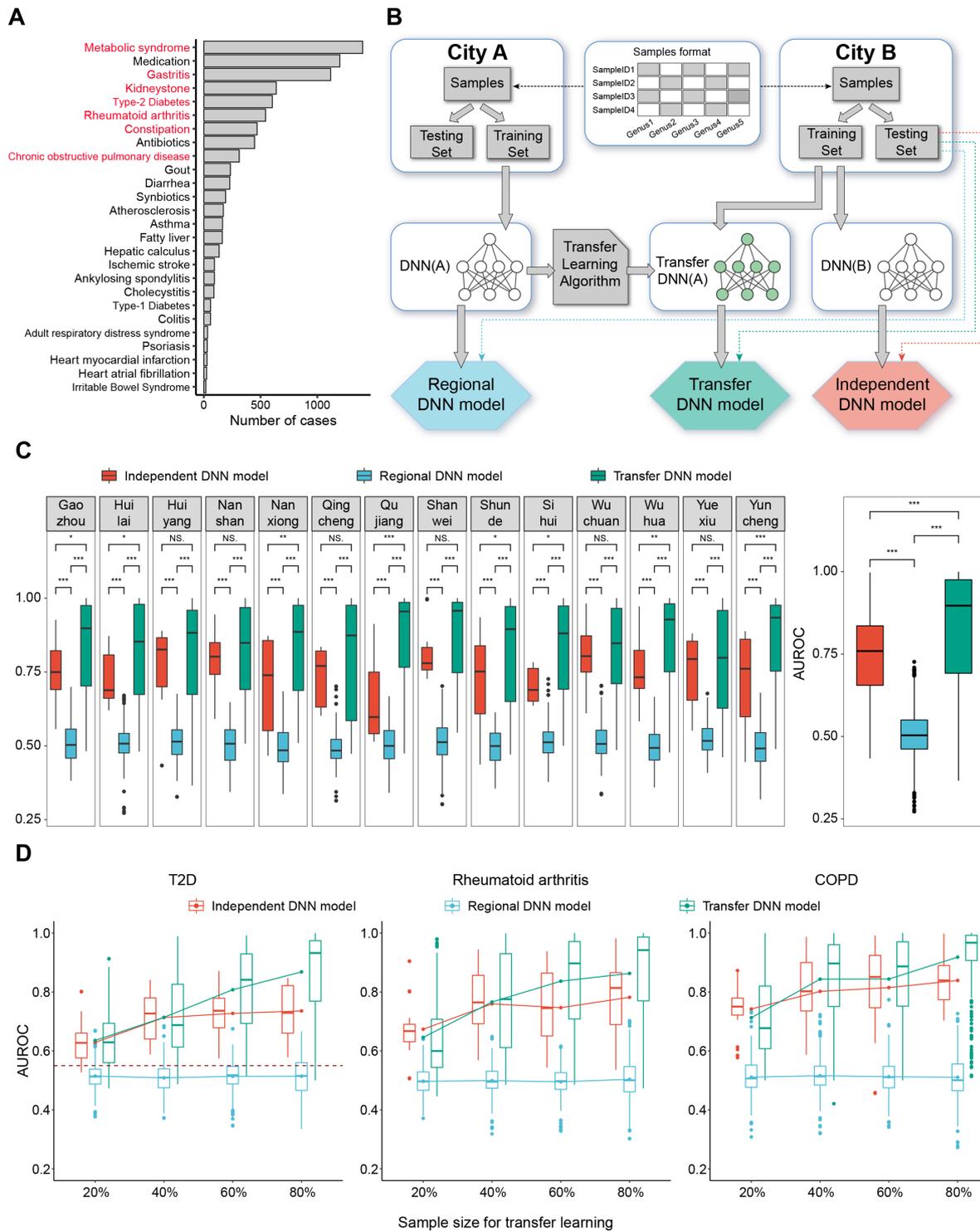
**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

© Author(s) (or their employer(s)) 2022. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

► Additional supplemental material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/gutjnl-2022-328216>).



**To cite** Wang N, Cheng M, Ning K. Gut Epub ahead of print: [please include Day Month Year]. doi:10.1136/gutjnl-2022-328216



**Figure 1** Data distribution, assessment workflow and framework evaluation. (A) data distribution. The number of samples of different diseases. The seven diseases marked in red were assessed, including metabolic syndrome, gastritis, kidney stones, T2D, rheumatoid arthritis, constipation and COPD. (B) The workflow of assessment. The genera abundance profiles of samples from each city were randomly divided into the training subset (80%) and the testing subset (20%). Three assessment workflows for each model were marked by three different colours. The testing subset of city B was used to test all the three models. (C) Framework evaluation: comparison of the AUROC of three models. Boxplots in the left panel show the AUROC of the three models for diagnosing seven diseases using samples in each of city, and the right panel shows these values collectively. \*,  $p < 0.05$ ; \*\*,  $p < 0.01$ ; \*\*\*,  $p < 0.005$ ; Mann-Whitney-Wilcoxon test. (D) The relationship between sample size and the performance of three models. Boxplots show AUROC of three models for diagnosing three diseases (COPD, rheumatoid arthritis and T2D). The lines show the change in average AUROC of three models with sample size increasing. The dashed line shows the average AUROC of cross-regional diagnosis of T2D using random forest model.<sup>5</sup> For all the boxplots, boxes represent the IQR between the first and third quartiles and the line inside represents the median. Whiskers denote the lowest and highest values within the  $1.5 \times IQR$  from the first and third quartiles, respectively. AUROC, area under the receiver operating characteristic; COPD, chronic obstructive pulmonary disease; T2D, type 2 diabetes.

Received 5 July 2022  
Accepted 10 October 2022

*Gut* 2022;**0**:1–3. doi:10.1136/gutjnl-2022-328216

#### ORCID iDs

Nan Wang <http://orcid.org/0000-0001-8671-6613>

Mingyue Cheng <http://orcid.org/0000-0003-1243-5039>

Kang Ning <http://orcid.org/0000-0003-3325-5387>

#### REFERENCES

- 1 Clooney AG, Eckenberger J, Laserna-Mendieta E, *et al*. Ranking microbiome variance in inflammatory bowel disease: a large longitudinal intercontinental study. *Gut* 2021;**70**:499–510.
- 2 Weng F, Meng Y, Lu F, *et al*. Differentiation of intestinal tuberculosis and crohn's disease through an explainable machine learning method. *Sci Rep* 2022;**12**:1714.
- 3 Gou W, Ling C-W, He Y, *et al*. Interpretable machine learning framework reveals robust gut microbiome features associated with type 2 diabetes. *Diabetes Care* 2021;**44**:358–66.
- 4 Sun Q, Liu Y, Chua T. Meta-transfer learning for few-shot learning. 2019 IEEE Conf Comput Vis Pattern Recognit, 2019:403–12.
- 5 He Y, Wu W, Zheng H-M, *et al*. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat Med* 2018;**24**:1532–5.

1 **Supplemental information**

2

3 **Overcoming regional limitations: Transfer learning for cross-regional**  
4 **microbial-based diagnosis of diseases**

5

6 Nan Wang, Mingyue Cheng, Kang Ning\*

7

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

13

14 \*Correspondence to Dr Kang Ning, E-mail: ningkang@hust.edu.cn.

## 15 **Analysis of the intercontinental cohort**

16 To validate the advantage of our proposed transfer learning approach, we also applied  
17 it on the dataset used by Clooney *et al*<sup>1</sup> (EBI accession number PRJNA414072). The  
18 dataset contained samples from intercontinental regions: Ireland (Irish, 118 Crohn's  
19 disease patients, 84 Ulcerative colitis patients and 84 controls) and Canada (Canadian,  
20 197 CD patients, 167 UD patients and 79 controls). Samples from each country were  
21 randomly divided into training subset (70%) and testing subset (30%). As shown in the  
22 Supplemental Figure S3, the independent DNN model has shown the AUROC of 0.872  
23 (Irish), 0.822 (Canadian), respectively. The AUROC values of the regional DNN  
24 models were 0.568 (Irish to Canadian) and 0.551 (Canadian to Irish), respectively (as  
25 shown in **Supplemental Figure S3A**). Notably, the AUROC values of the transfer DNN  
26 models were increased to 0.867 (Irish to Canadian) and 0.823 (Canadian to Irish),  
27 respectively (as shown in **Supplemental Figure S3B**).

28

29 We have also applied it on another intercontinental cohort for cross-regional disease  
30 diagnosis. We chose an irritable bowel syndrome cohort including samples from three  
31 continents obtained from the American Gut Project<sup>2</sup> (EBI accession number  
32 PRJEB11419). This dataset includes samples from United States of America (USA,  
33 with 464 controls and 245 IBS patients), United Kingdom (UK, with 120 controls and  
34 118 IBS patients) and Australia (AUS, with 23 controls and 17 IBS patients). Samples  
35 from each country were randomly divided into training subset (70%) and testing subset  
36 (30%). As shown in the Supplemental Figure S4, the independent DNN model of each  
37 country was constructed on its training subset and tested on the testing subset, showing  
38 the AUROC of 0.725 (USA), 0.559 (UK) and 0.978 (AUS). We then used testing subset  
39 of the countries with a smaller dataset to test the transfer DNN model. The transfer  
40 DNN models showed AUROC of 0.721 (USA to UK), 0.836 (USA to AUS) and 0.697  
41 (UK to AUS), higher than that of the regional DNN models with 0.584 (USA to UK),  
42 0.766 (USA to AUS) and 0.234 (UK to AUS), respectively.

43

44 These results showed that even if applied to intercontinental cohorts, transfer learning  
45 could improve the prediction ability across regions, thereby exceeding the limitation on  
46 microbial-based diagnosis of diseases across regions.

47

#### 48 **Leave-one-feature-out analysis**

49 Each of microbial features was removed in turn when constructing the transfer DNN  
50 model, then the change of the AUROC of the transfer DNN model was calculated.  
51 Microbes with the strongest change of the AUROC (top 2) would be designated as  
52 “region-specific”. On the contrary, the microbes with the least change of the AUROC  
53 (top 2) would be designated as “shared across all regions”.

54

#### 55 **Statistical analyses**

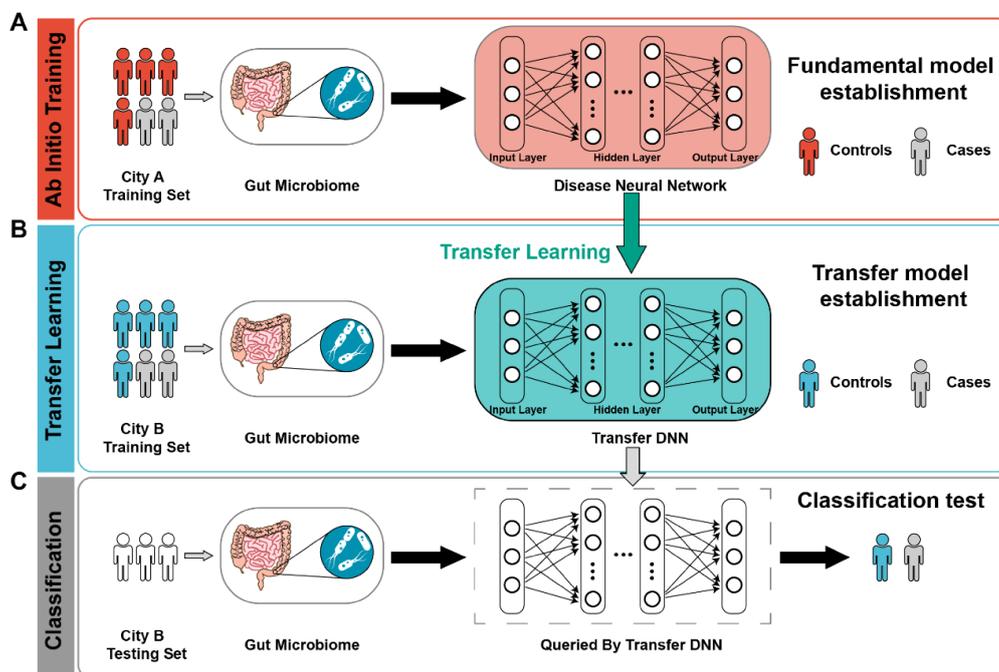
56 Mann-Whitney-Wilcoxon test was used to test the statistical significance of the  
57 differences in AUROC among independent DNN model, the regional DNN model and  
58 the transfer DNN model.

59

#### 60 **Codes and models availability**

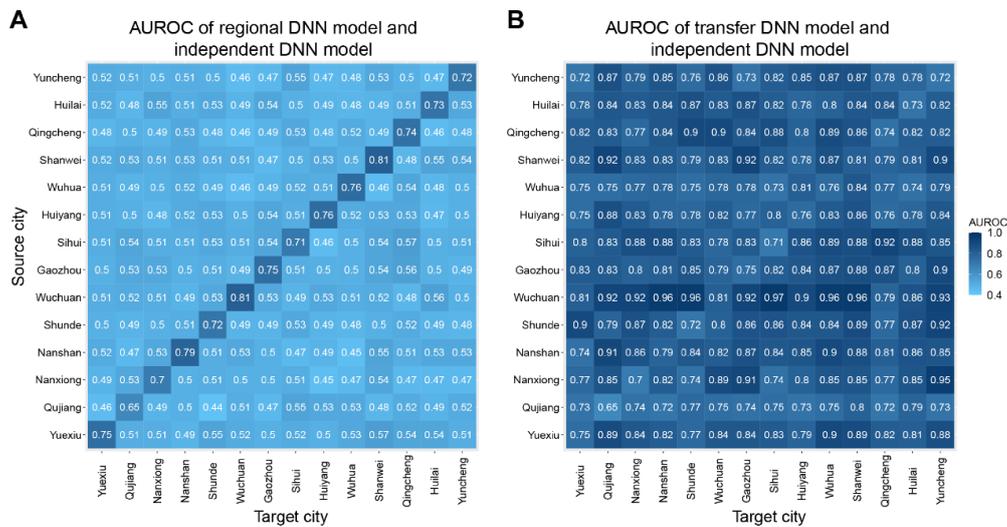
61 The pre-trained models and source codes of scripts for training, querying and transfer  
62 learning are made publicly available at [https://github.com/HUST-NingKang-](https://github.com/HUST-NingKang-Lab/EXPERT-Disease-GGMP)  
63 [Lab/EXPERT-Disease-GGMP](https://github.com/HUST-NingKang-Lab/EXPERT-Disease-GGMP). The program “EXPERT” is available at  
64 <https://github.com/HUST-NingKang-Lab/EXPERT>.

## 65 Supplemental figures



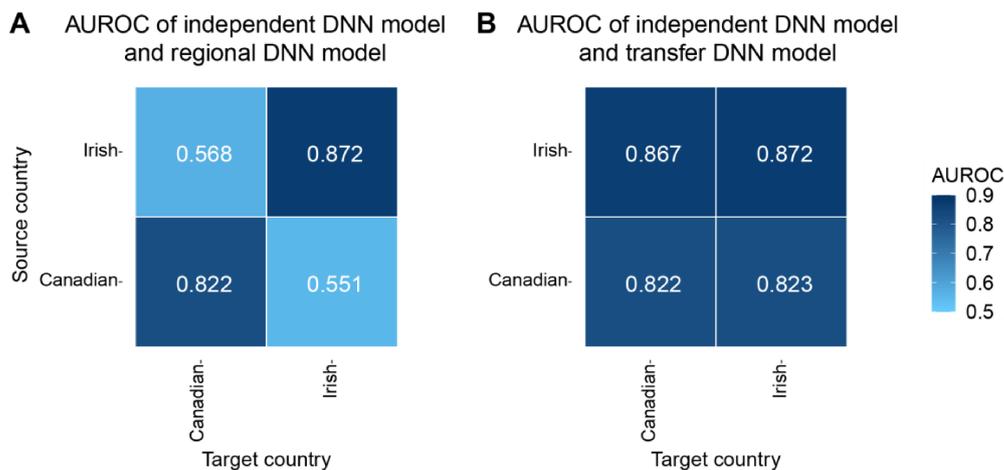
66

67 **Supplemental Figure S1** The transfer learning workflow of microbial-based cross-  
 68 regional diagnosis of diseases. (A) Fundamental model establishment: *ab initio* training  
 69 the DNN model on the training subset of city A. (B) Transfer model establishment:  
 70 implement transfer learning to generate a transfer DNN model for another city B by  
 71 using part of the training subset of city B. (C) Classification test: test the transfer DNN  
 72 model on the testing subset of city B. DNN, disease neural network.



73

74 **Supplemental Figure S2** Average AUROC of three DNN models for cross-regional  
 75 disease diagnosis. In both heatmaps, the average AUROC of the independent DNN  
 76 model on seven diseases across 14 cities was shown in the diagonal (A and B). The  
 77 average AUROC of the regional DNN model on seven diseases across 14 cities was  
 78 shown in the area outside of the diagonal. The average AUROC of the transfer DNN  
 79 model on seven diseases across 14 cities was shown in the area outside of the diagonal  
 80 (B). AUROC, area under the receiver operating characteristic curve; DNN, disease  
 81 neural network.



82

83 **Supplemental Figure S3.** Results of disease diagnosis across intercontinental regions.

84 In both heatmaps, the AUROC values of the independent DNN model on IBD across

85 two countries were shown in the diagonal (A and B). The AUROC values of the regional

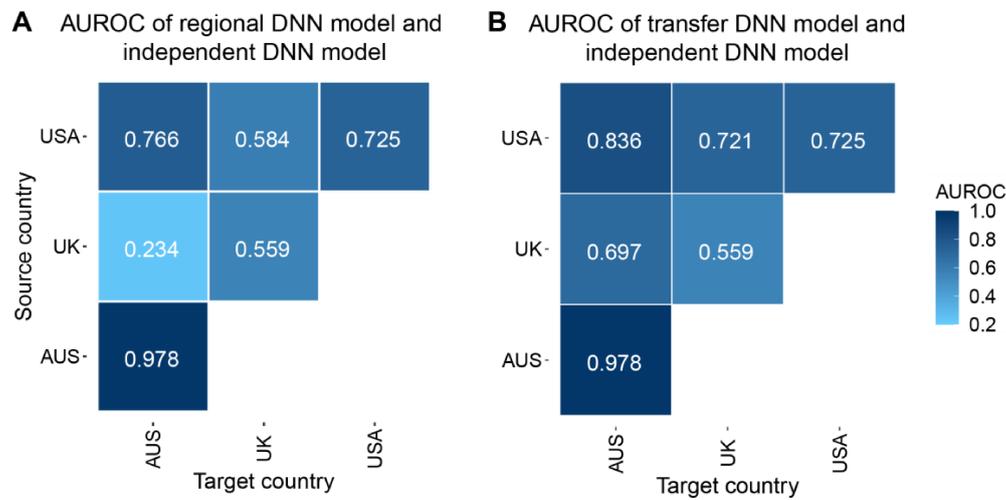
86 DNN model on IBD across two countries were shown in the area outside of the diagonal

87 (A). The AUROC values of the transfer DNN model on IBD across two countries were

88 shown in the area outside of the diagonal (B). AUROC, area under the receiver

89 operating characteristic curve; IBD: inflammatory bowel disease; DNN, disease neural

90 network.



91

92 **Supplemental Figure S4.** Results of disease diagnosis across intercontinental regions.  
93 In both heatmaps, the AUROC values of the independent DNN model on IBS across  
94 three countries were shown in the diagonal (A and B). The AUROC values of the  
95 regional DNN model on IBS across three countries were shown in the area outside of  
96 the diagonal (A). The AUROC values of the transfer DNN model on IBS across three  
97 countries were shown in the area outside of the diagonal (B). AUROC, area under the  
98 receiver operating characteristic curve; IBS: irritable bowel syndrome; DNN, disease  
99 neural network.

## Supplemental tables

**Supplemental Table S1. The number of cases and controls of each city and each disease.**

City	Control	Constipation	COPD	Gastritis	Kidneystone	MS	RA	T2D	Total number of participants
Yuexiu	261	51	41	74	50	158	38	92	448
Qujiang	214	54	31	154	130	207	166	112	488
Nanxiong	306	51	60	173	130	176	35	89	590
Nanshan	295	104	50	177	127	269	58	124	618
Shunde	189	45	36	117	53	122	55	49	362
Wuchuan	333	83	38	77	35	79	36	63	533
Gaozhou	254	41	57	173	73	142	48	109	509
Sihui	256	39	28	79	66	129	91	55	453
Huiyang	232	49	43	93	105	179	102	99	444
Wuhua	207	51	37	258	184	246	111	160	500
Shanwei	227	80	24	119	40	148	69	64	415
Qingcheng	308	60	32	129	108	191	41	73	577
Huilai	247	80	63	171	8	220	172	113	506
Yuncheng	285	69	49	148	159	215	89	117	555

*Note* : COPD, Chronic obstructive pulmonary disease; MS, Metabolic syndrome; RA, Rheumatoid arthritis; T2D, Type 2 diabetes.

**Supplemental Table S2. Transfer learning identifies the region-specific microbes, as well as microbes shared across all regions.**

<b>Taxonomy</b>	<b>Constipation</b>	<b>COPD</b>	<b>Gastritis</b>	<b>Kidneystone</b>	<b>MS</b>	<b>RA</b>	<b>T2D</b>	<b>Variance</b>
Roseburia	0.1079	-0.0215	-0.0885	0.1647	-0.0571	0.1437	0.1037	0.0107
Faecalibacterium	0.0587	-0.0052	0.0244	0.0744	-0.0419	0.0746	0.0332	0.0019
Enterobacteriaceae	0.2183	0.1223	-0.0316	0.1898	-0.2073	0.0835	0.3342	0.0316
Bacteroides	-0.0299	0.1740	-0.0886	0.1092	-0.0308	0.2233	0.0645	0.0134
Lachnospiraceae	0.0488	0.0245	-0.0287	0.1413	-0.0982	0.1019	0.1056	0.0071
Oscillospira	0.0670	0.0081	0.0127	-0.0468	-0.0114	0.1329	0.1409	0.0052
Parabacteroides	0.0094	0.1030	0.0045	0.0706	-0.0201	0.0588	0.0671	0.0020
Prevotella	0.1656	0.0089	0.0825	0.1081	-0.0243	0.0231	0.2330	0.0084
Clostridium	0.0227	-0.0439	-0.0829	-0.0499	0.0005	-0.0183	0.3116	0.0177
Ruminococcus	0.1347	-0.0094	-0.0082	0.1081	0.0167	0.1460	0.1002	0.0046

10 region-specific microbes, as well as microbes shared across all regions were filtered out by using the "leave-one-out" analysis. The values represent the change of AUROC of the transfer DNN model used for the diagnosis of the seven diseases after removing the microbe. The positive value indicates that the microbe has a positive effect on the prediction, conversely, the microbe has a negative effect. The variance of these values are also displayed. DNN, disease neural network.

*Note* : COPD, Chronic obstructive pulmonary disease; MS, Metabolic syndrome; RA, Rheumatoid arthritis; T2D, Type 2 diabetes.

100 **References**

101 1. Clooney AG, Eckenberger J, Laserna-Mendieta E, et al. Ranking microbiome  
102 variance in inflammatory bowel disease: a large longitudinal intercontinental  
103 study. *Gut* 2021;70(3):499–510.

104 2. McDonald D, Hyde E, Debelius Justine W, et al. American Gut: an Open Platform  
105 for Citizen Science Microbiome Research. *mSystems* 2018;3(3):e00031–18.

106