Supplemental information

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

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 AI Biology, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China.

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

Analysis of the intercontinental cohort

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

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

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

Leave-one-feature-out analysis

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

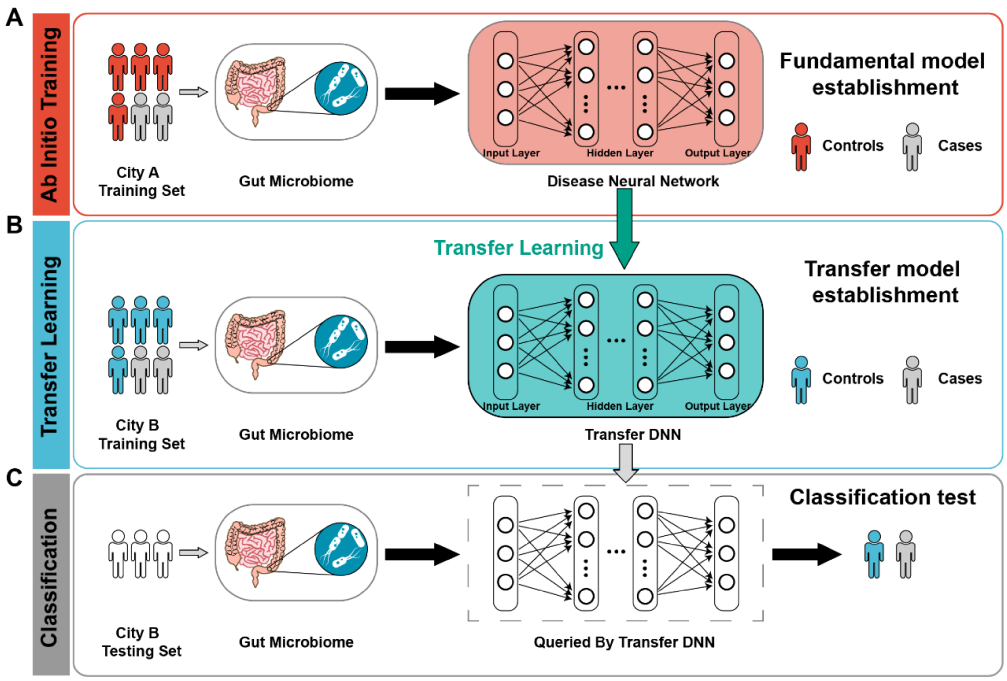
Statistical analyses

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

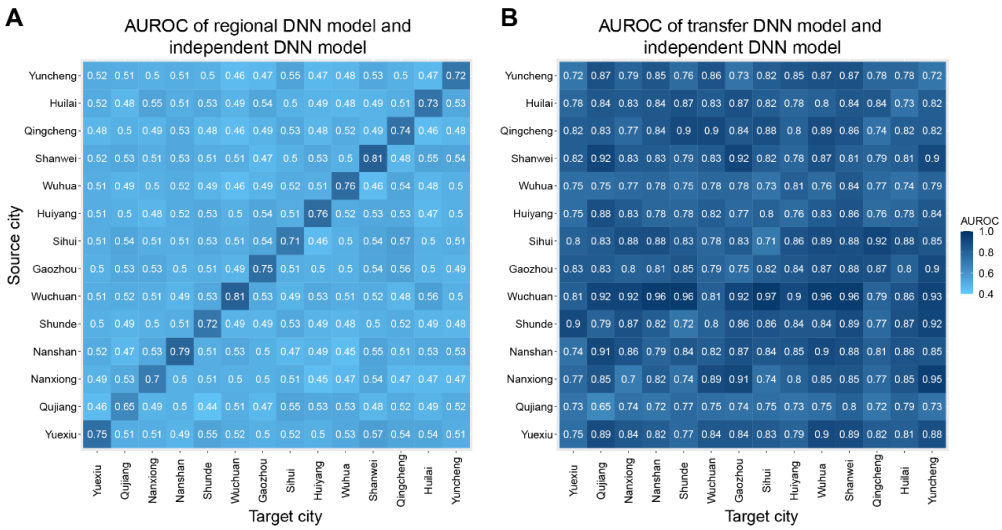
Codes and models availability

The pre-trained models and source codes of scripts for training, querying and transfer learning are made publicly available at https://github.com/HUST-NingKang-Lab/EXPERT-Disease-GGMP. The program "EXPERT" is available at https://github.com/HUST-NingKang-Lab/EXPERT.

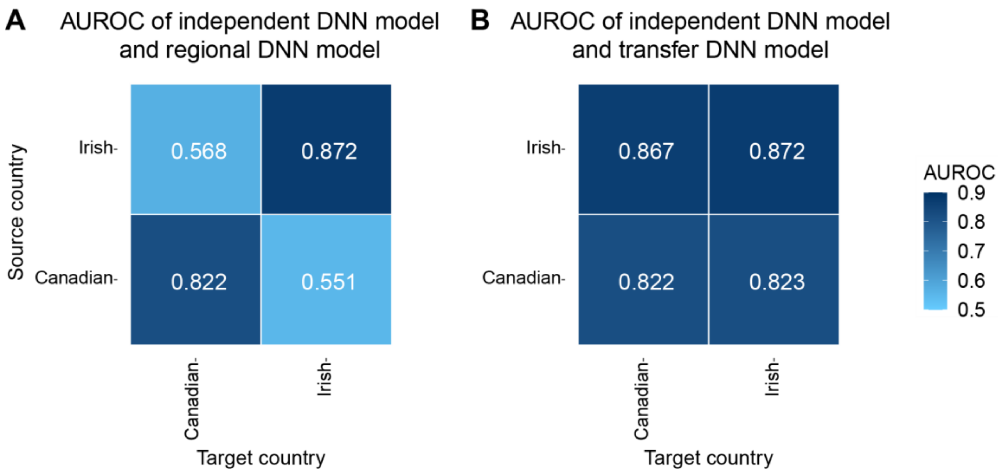
Supplemental figures



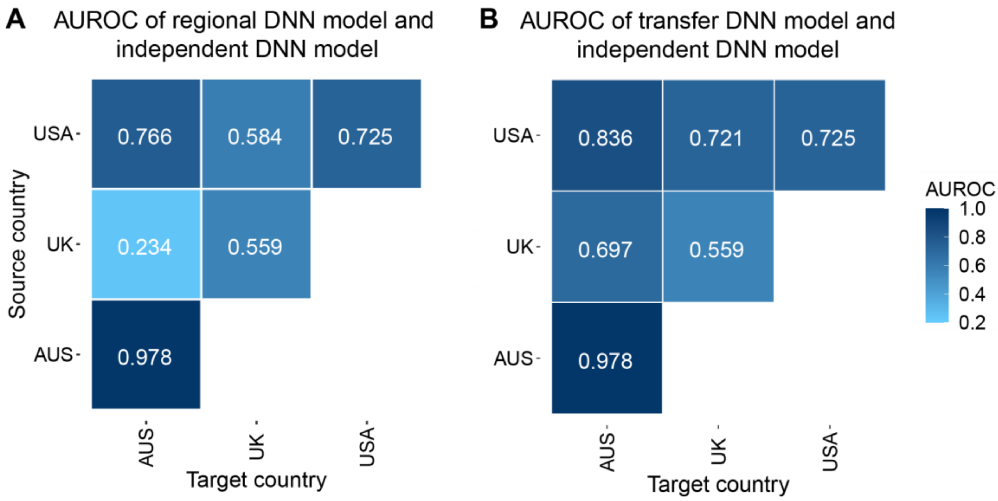
Supplemental Figure S1 The transfer learning workflow of microbial-based cross-regional diagnosis of diseases. (A) Fundamental model establishment: *ab initio* training the DNN model on the training subset of city A. (B) Transfer model establishment: implement transfer learning to generate a transfer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



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



Supplemental Figure S3. Results of disease diagnosis across intercontinental regions. In both heatmaps, the AUROC values of the independent DNN model on IBD across two countries were shown in the diagonal (A and B). The AUROC values of the regional DNN model on IBD across two countries were shown in the area outside of the diagonal (A). The AUROC values of the transfer DNN model on IBD across two countries were shown in the area outside of the diagonal (B). AUROC, area under the receiver operating characteristic curve; IBD: inflammatory bowel disease; DNN, disease neural network.



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

Taxonomy	Constipation	COPD	Gastritis	Kidneystone	MS	RA	T2D	Variance
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.

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(3):499–510.

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