Enabling technology for microbial source tracking based on transfer learning: From ontology-aware general knowledge to context-aware expert systems

Hui Chong[1], Qingyang Yu[1], Yuguo Zha[1], Guangzhou Xiong[1], Nan Wang[1], Chuqing Sun[1], Sicheng Wu[1], Weihua Chen[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

[*]Correspondence should be addressed to K.N (Email: ningkang@hust.edu.cn)

Abstract

Habitat specific patterns reflected by microbial communities, as well as complex interactions between the community and their environments or hosts' characteristics, have created obstacles for microbial source tracking: diverse and context-dependent applications are asking for quantification of the contributions of different niches (biomes), which have already overwhelmed existing methods. Moreover, existing source tracking methods could not extend well for source tracking samples from understudied biomes, as well as samples from longitudinal studies.

Here, we introduce EXPERT (https://github.com/HUST-NingKang-Lab/EXPERT), an exact and pervasive expert model for source tracking microbial communities based on transfer learning. Built upon the biome ontology information and transfer learning techniques, EXPERT has acquired the context-aware flexibility and could easily expand the supervised model's search scope to include the context-dependent community samples and understudied biomes. While at the same time, it is superior to current

approaches in source tracking accuracy and speed. EXPERT's superiority has been demonstrated on multiple source tracking tasks, including source tracking samples collected at different disease stages and longitudinal samples. For example, when dealing with 635 samples from a recent study of colorectal cancer, EXPERT could achieve an AUROC of 0.977 when predicting the host's phenotypical status. In summary, EXPERT has unleashed the potential of model-based source tracking approaches, enabling source tracking in versatile context-dependent settings, accomplishing pervasive and in-depth knowledge discovery from microbiome.

Introduction

Advances in sequencing technology and informatics are producing an exponential increase in data acquisition and integration, and revolutionizing our understanding of the roles microbes play in health and disease, biogeochemical cycling, etc.[1–4]. Hundreds of thousands of microbial community samples have been accumulated yearly, corresponding to hundreds of niches (biomes) around the globe[5–7], and continuously completing the grand picture about the microbiome world. However, currently annotated biomes only represented the pick of an iceberg of such a grand picture, while a considerable amount of these samples are from understudied and newly discovered biomes, including (1) newly discovered biomes from where microbial community samples were only collected recently, (2) the biomes which were more context-dependent such as those representing the development of gut microbial communities, or those representing different stages of diseases, and (3) the biomes which include the longitudinal samples. These understudied biomes represent a grand pool of unexplored and previously less studied microbiome research sphere, have created huddles for sample comparison and source tracking[8]. The microbial community dark matters, referring to those samples and niches that have been unseen or understudied, have grown exponentially, which urged the context-aware method to source track the biomes. However, there is no context-aware method for such purposes, as such context-dependent investigations requiring both comprehensiveness (serving for characterizing (1)) and scalability (serving for characterizing both (2) and (3)).

Previous approaches for community-wide microbial source tracking (MST) have serious tradeoffs regarding efficiency, accuracy and scalability. Markov Chain Monte Carlo (MCMC)[9] and Expectation-Maximization (EM)[10] methods have a tradeoff between accuracy and efficiency when facing an increasing number of possible sources (referred to **Supplementary Note 1** for a detailed discussion). They can hardly serve for source tracking among thousands of samples, which often requires weeks to years to complete (linearly extrapolated based on results in L. Shenhav et al.[10]). Whereas for Neural Network (NN) method[11], the tradeoff lies in scalability: It can hardly be expanded to context-dependent applications and longitudinal analyses, limiting its utility in current microbiome studies.

We have developed EXPERT, an exact and pervasive expert model for source tracking microbial community samples to address these limitations. EXPERT combines the superior efficiency and accuracy of the Neural Network method, as well as the transfer learning approach's inherent scalability, enabling knowledge transfer into context-dependent settings to better understand the microbial community dark matters. EXPERT is designed to quantitatively assign the contribution of biomes to a specific microbial community, in both ontology-aware and context-aware manners. EXPERT's advantages include its ability to rapidly infer the contributions from multiple sources via ontology-aware forward propagation, and its ability to adapt to emerging research via alterable biome ontology structure (i.e., a hierarchy of biomes involved in the target application). EXPERT has made it possible to look into the dark matter of microbial communities that include millions of samples from hundreds of context-dependent biomes. Thus it has enabled fast, accurate, flexible, and interpretable source tracking at an unprecedented scale towards deeper knowledge discovery from microbial communities.

EXPERT has demonstrated superior performance on a diverse set of source tracking tasks: For source tracking among newly discovered biomes, it has been shown to be able to adapt to emerging samples and biomes with unprecedented AUROC higher than 0.990. For the more context-dependent biomes such as those representing the progression of colorectal cancer, it has achieved an AUROC of 0.977 when predicting the host's phenotypical status, exhibiting its potential in disease diagnosis. It has also enabled the

longitudinal sample source tracking, revealed the compositional shifts of individuals' gut microbial communities along the timeline. In summary, EXPERT has enabled the flexibility of model-based method for context-aware source tracking, expanded the scope of source tracking to understudied biomes, and boosted in-depth knowledge discovery from the microbial communities.

Results

Rationale, modeling and multi-facet applications of EXPERT

EXPERT is an efficient approach for microbial community source tracking employing both neural network modeling and transfer learning techniques[12], enabling knowledge transfer from ontology-aware general knowledge to context-aware expert systems. The data preprocessing pipeline, the limitation of directly use of ontology-aware general knowledge, as well as knowledge transferring procedure of EXPERT are illustrated in **Fig. 1a-c**. Firstly, EXPERT is agnostic to the sequencing data type (16S ribosomal RNA or shotgun sequencing) or the analysis pipeline. And before source tracking, EXPERT applies remapping, normalizing, aggregation (**Supplementary Note 2**), and Z-score standardization for the standardazation of microbial community data (**Fig. 1a**).

Secondly, one of the fundamental models in EXPERT before transfer learning is a neural network model, which is either a general ontology-aware neural network (ONN) model ONN4MST[11] (https://github.com/HUST-NingKang-Lab/ONN4MST) or any other neural network built for the same purpose. This neural network model is also referred to as the ontology-aware general knowledge. The model-based source tracking follows this rationale: For a community sample as query, by assuming that quantifying contributions from a set of independent biomes will also contribute to quantifying contributions from the biomes serving as the superset of these independent biomes (e.g., "Human" is the superset biome of "Human: Oral" and "Human: Fecal"), EXPERT quantifies the contributions of biomes along the ontology layers via a propagation procedure, where messages of higher layers are integrated into those of lower layers, namely ontology-aware forward propagation. The ontology-aware general knowledge modeling has

enabled EXPERT to efficiently source track among up to hundreds of sources biomes (containing sub-millions of source samples). However, the ontology-aware general knowledge comes with poor scalability and can hardly be expanded to context-dependent applications (**Fig. 1b**).

Thirdly, EXPERT can adapt the general knowledge of an existing model to context-dependent applications through three steps, namely transfer, adaptation, and finetuning (**Fig. 1c**). It has achieved this by utilizing both ontology-aware general knowledge, and context-dependent biome ontology, which is essentially a hierarchy of biomes involved in the target application. During the transfer process, EXPERT reuses the existing model's parameters to optimize a context-aware model and encode context-dependent biome ontology into the model through reinitializing context-dependent layers (containing only 5% parameters of the entire model) according to the knowledge of the application-dependent context. The phylogenetic tree from the existing model was also reused (**Supplementary Table S1**). During the adaptation process, EXPERT quickly optimizes only context-dependent layers to enable the model to become suitable for the context. During the finetuning process, EXPERT further optimizes the entire model to thoroughly adapt it to the application. Note that EXPERT can also learn from partially-labeled data by masking the losses of unlabeled layers (**Supplementary Fig. 1, Supplementary Note 3**).

The multi-facets of EXPERT applications are illustrated in **Fig. 1c**. First, we can quantify the contributions of biomes for any microbial community by using EXPERT with superior speed and accuracy (**Fig. 1c i**). Secondly, EXPERT is able to adapt to source tracking among merging microbiome data (also referred to as the "Grafting", **Fig. 1c ii**). Thirdly, we can introduce more detailed biomes in order to investigate the small differences among these closely related biomes (i.e., the biomes representing the infant of different ages, **Fig. 1c iii**), which is also referred to as the "Scattering". Fourthly, by introducing knowledge about diseases (disease ontology), we can build an EXPERT model for characterizing the health status of hosts. Alternatively, by introducing knowledge about disease progression, we can build a model for monitoring disease progression (**Fig. 1c iv**). Finally, we can leverage time points and additional context-

aware knowledge to investigate dynamics of microbial communities along the timeline, such as longitudinal dynamics of communities in accordance with dietary shifts or seasonal shifts (**Fig. 1c v**). All of these context-aware applications were also implemented and assessed in details in the following parts of this work.

Transferring general knowledge into context-dependent applications: systematic assessment

We verified the utility of our transfer scheme by assessing the performance of tranfer learning under two application scenarios: quantification of contributions under general application scenario, as well as quantifying human-associated source contributions under context-dependent application scenario. We introduced two datasets to assess the performances systematically: a combined dataset consists of 118,592 samples collected from 132 biomes (**Supplementary Table S2, Supplementary Table S3**), and a human dataset consists of 52,538 samples collected from 28human-associated biomes (**Supplementary Table S2, Supplementary Table S4**). The general EXPERT model (also referred to as the general model) is generated based on using samples from the combined dataset, and we have first assessed the performance of the general model under general application scenario, namely on source tracking samples in the combined dataset. Through random cross-validation (see **Methods**), we found that the general model was able to quantify source contributions for communities with high AUROC of 0.971 (**Supplementary Fig. 2**).

We then evaluated the transfer learning approach in a context-dependent application. We examined the application of knowledge transfer and finetuning (refer to **Methods** for details) on quantifying human-associated source contributions under context-dependent application scenario, in three aspects: efficiency, accuracy, . In this context, the transfer models built based on transferring knowledge from the general model with and without finetune were referred to as Transfer (GM) and Transfer (GM0), respectively (**Supplementary Table S5**). These transfer models were compared with independent model, which was generated by general ontology-aware neural network approach[11] based on the samples and biomes in the context-dependent application (**Supplementary Table**

S5). The results have shown that the knowledge transfer scheme with finetuning enabled more accurate quantification of source contributions for query samples (average AUROC of 0.960, **Fig. 2c**) compared to the independent model. We have also found that 95.7% of the parameters from the original general model has been transferred to Transfer (GM), representing 99.3% parameters of Transfer (GM), confirming that the high source tracking accuracy actually come from the transfer learning process (**Supplementary Table S5**). Notably, the finetune optimization process comes up with a cost: Three times as much time was spent on performing this optimization (**Fig. 2d**), primarily due to the low learning rate (1x10-5) utilized by finetuning. Nevertheless, considering the advantages in accuracy (**Fig. 2c**), we determined finetuning as a default setting in the following sections.

We further compared the performance of EXPERT with FEAST[10] in a flat setting, where the potential sources (only the bottom layer of the human ontology) are considered independent of each other (**Fig. 2a**), as FEAST cannot recognize the hierarchical relationships among biomes (**Supplementary Note 4**). Results have shown that EXPERT could perform source tracking with excellent performance (F-max of 0.914) and ultrahigh search speed (over 200 samples per second). However, though FEAST has been proven to be much faster than SourceTracker[9] (**Fig. 2e**, **Supplementary Note 1**), we have noticed a severe tradeoff in FEAST between accuracy and running time, which is heavily depend on the number of sources: when the numbers of samples as sources for FEAST increase, the accuracies could reach those similar with EXPERT, but at the cost of magnitude more time (**Fig. 2e**, **Supplementary Table S6**).

EXPERT enables adaptation for emerging microbiome data

In this context, we aim to adapt the general model to emerging microbiome data from understudied or newly discovered biomes, which is also referred to as ”Grafting”. An increasing number of biomes are being studied in metagenomics, accompanied by tens of thousands of communities deposited into public databases. Many of these biomes are understudies, and a profound number of biomes are newly discovered, such as root: Host-associated: Fish and root: Host-associated: Birds, just in year 2020 (**Fig. 3a**). To assess

the capability of EXPERT on such emerging microbiome data, we collected 34,209 samples, which were collected and analyzed by MGnify[6] in 2020. Among them, there are 3,419 samples originated from 8 newly discovered biomes (**Supplementary Table S2, Supplementary Table S7**). We used the hierarchical biome organization of these samples to construct a new biome ontology (Biome ontology (2020)), which included 36 biomes in total (**Fig. 3a**). When comparing the accuracy between the independent model (AUROC = 0.993, F-max = 0.986) and the Transfer (GM) model (AUROC = 0.991, F-max = 0.978), we confirmed that EXPERT is able to source track accurately among biomes in the Biome ontology (2020) (**Fig. 3b**). Again, we confirm that the high source tracking accuracy actually come from the transfer learning process (**Supplementary Table S5**). The average search time of the transfer model is also less than that of the independent model (**Fig. 3c**). Two understudied biomes (root: Host-associated: Fish and root: Host-associated: Birds: Digestive System: Ceca) were chosen to illustrate the source tracking accuracy of EXPERT at specific layers of the Biome ontology (2020). We noticed that the contribution of the correct biome (Fish and Birds) has a high value on multiple layers of the Biome ontology (2020) (**Fig. 3d**). In a nutshell, EXPERT could accurately source track samples from understudies or newly discovered biomes, enabling lifelong adaptation for the emerging microbiome data.

EXPERT for monitoring the succession of infant gut microbial communities

In this context, we aim to utilize knowledge transfer to source track among more detailed biomes, which is also referred to as ”Scattering”. Under this circumstance, we could explore the dynamic patterns of gut microbiota from a specific period of life. For instance, if infant samples from multiple time points and sources are offered, EXPERT could estimate how much of microbial community in the infant's gut is originated from birth and subsequent time points. To confirm this capability, we used longitudinal data from Backhed et al.[13], which consists of fecal samples from 98 infants and their mothers, delivered by vaginal delivery or cesarean section (**Fig. 4a, Supplementary Table S2, Supplementary Table S8**). Except for the general EXPERT model, here we introduced the human EXPERT model (also referred to as the human model), which is generated based on using samples from the human dataset consists of 53,553 samples collected

from 27 human-associated biomes (**Supplementary Table S2, Supplementary Table S4**). The context-dependent ontology is organized by dividing samples by development stage first, followed by birth mode (**Fig. 4a**).

In this part of the study, we considered samples from infants at 12 months of age as queries, and samples from early time points or mothers were treated as sources. Firstly, there is no significant difference among the samples from mothers (Wilcoxon test, p = 0.388 for Transfer (GM), p = 0.929 for Transfer (HM), **Fig. 4b**), which is consistent with the results of J. Stokholm et al.[14]. This indistinguishable pattern is further supported by Principal Coordination Analysis (PCoA) using distance metric either in weighted-Unifrac[15] or Jensen Shannon divergence[16] (**Fig. 4c** and **Supplementary Fig. 3**), in which samples from mothers and infants at 12 months are indistinguishable.

We then assessed the performances of different source tracking models via cross-validation. For infant gut microbial communities at 12 months of age, even samples were collected from hosts of different delivery modes, the maternal contribution is dominant (**Fig. 4b**). When comparing with other methods, we found that the transfer model built from the human model (Transfer (HM)) has the best performance (AUROC = 0.773) compared to the independent model ((AUROC = 0.738) and the Transfer (GM) model (AUROC = 0.720), indicating the outstanding performance of EXPERT for source tracking gut microbial communities for infants at different stages (**Fig. 4d**). Again, we confirm that the high source tracking accuracy actually come from the transfer learning process (**Supplementary Table S5**). When examining the results of source tracking on individual samples, we also observe that for quite a few samples the predicted biome contributions by independent model were not consistent with ground truth, while the predictions by Transfer (HM) were in agreement with ground truth, confirming the advantage of transfer learning in this context, especially when samples are difficult to distinguish (**Fig. 4e**). Furthermore, we made attempt by changed the context-dependent setting: we have used the infant gut microbial community sample at birth as queries (**Supplementary Fig. 4**), with results showing that samples from infant at birth have most contributions from infant of 4 months, consistent with the results of J. Stokholm et al.[14]. We have also changed the biome ontology, by means of dividing samples by birth

mode first, followed by development stage (**Supplementary Fig. 5**), and the source tracking results of Transfer (HM) is still superior than other models.

EXPERT for disease diagnosis and monitoring

It has been reported that the human gut microbiota is potentially strongly associated with a vast array of complex, chronic diseases[17–21]. Considering these potential associations between the human gut microbial community and diseases, we also attempted to characterize disease phenotypes (**Fig. 5a**). We collected 13,642 microbial community samples from the GMrepo database[7] and retrieved their corresponding disease classification from the NCBI MeSH database[22] (**Supplementary Table S2, Supplementary Table S9**, and **Fig. 5b**). Except for the general EXPERT model and the human EXPERT model, here we introduced the disease EXPERT model (also referred to as the disease model), which is generated based on using samples from these 13,642 samples collected from 20 disease-assiciated biomes. Built based on the disease-related biome ontology, EXPERT can precisely distinguish 20 phenotypes, including health and 19 gut-associated hematologic diseases, liver diseases, intestinal diseases, and bacterial infections (over 0.8 AUROC for most phenotypes, **Fig. 5c**). Notably, a high AUROC (over 0.8) was also observed when the model was rebuilt based on the human model (Transfer (HM), **Fig. 5d**), suggesting the potential of EXPERT in host health status prediction. To the best of our knowledge, this is the most comprehensive study in terms of multiple disease diagnosis (13,642 samples, 20 health-associated phenotypes) with superior accuracy[23]. Again, we confirm that the high source tracking accuracy actually come from the transfer learning process (**Supplementary Table S5**). These results also suggests the potential role of the gut microbes in the alteration of host health status[24,25].

One confirmed association is that the gut microbiota is involved in colorectal cancer (CRC) progression[26]. It holds great potential to investigate whether CRC progression could be monitored through gut microbiota. We assessed such applicability of EXPERT by introducing CRC knowledge from Zeller, G. et al.[23]: we considered five stages in the progression of CRC: 0 (Healthy control) I, II, III, and IV according to the study of Zeller

G. et al.[23] (**Fig. 5f, Supplementary Table S2, Supplementary Table S10**). We first found that the compositional shifts of the human gut within such progression are indistinguishable by traditional methods, exemplified by Principle Coordination Analysis (PCoA) using distance metric either in weighted-Unifrac[15] or Jensen Shannon divergence[16] (**Fig. 5g, Supplementary Fig. 6**). Then we have compared the performance of the independent model and transfer models built from the human model (Transfer (DM)) and that from the disease mode (Transfer (DM)). Results have shown that Transfer (DM) achieved a better performance (AUROC over 0.95, **Fig. 5h, i**) among these three models, proving the superior applicability of EXPERT as a method for early detection of the occurrence of colorectal cancers. We also notice that the samples used for this analysis is not from a prospective study, and we expected that for an actual prospective study on the development of cancers, more insights could be gained through such a knowledge transfer approach.

EXPERT implicates dynamic patterns of microbial communities in longitudinal studies

Using EXPERT for time-series analysis offers a quantitative way to characterize time-related microbial compositional shifts, such as the dynamics of gut microbial communities during international travel. In this context, by leveraging additional metadata such as geographic regions and time points, we can characterize the community's compositional shifts associated with exceptional events. To demonstrate this capability, we used longitudinal samples from Liu H. et al.[27], including samples from ten Chinese travelers (MT1-9 normally returned, while MT10 returned with an exceptional early-back, **Fig. 6a,b**) who had a six-month long-stay from China to Trinidad and Tobago, as well as both China and Trinidad and Tobago native persons (**Supplementary Table S2** and **Supplementary Table S11**). Specifically, we performed an individual (people)-level leave-one-out experiment for these ten travelers: each time considering all samples from a specific traveler as queries and the rest samples (except for samples from MT10) as sources (referred to **Methods** for details), ten times. And we generated a transfer model based on transferring knowledge from the disease model (containing samples all from

human gut) for source tracking in this context. In these settings, our model can reveal the compositional shifts of Chinese travelers gut microbial communities over time and accurately pinpoint the early return timepoint of MT10 (**Fig. 6c**). These results are consistent with their travel trajectory within ten months (from Dec 2015 to Sep 2016)[27]. And the significant difference (p-value 1.9×10^{-6}, 2.4×10^{-4}, and 3.9×10^{-5} for three stages, Wilcoxon test) in source contributions revealed by our model also demonstrated predictive power (for such compositional shifts) of EXPERT.

We further explored whether we could observe such time-related compositional shifts in another cohort, about the seasonal changes of the Hadza persons' gut microbial communities, by using longitudinal samples from Smith S. A. et al.[28] (**Supplementary Table S2** and **Supplementary Table S12**). By dividing samples into "Dry" and "Wet" catagories (**Fig. 6d**), an average AUROC of 0.82 was achieved for distinguishing seasonal patterns ("Wet" or "Dry") among gut microbial communities from Hadza hunter-gatherers (**Fig. 6e**). These results are consistent with the results by the original study of Smith S. A. et al.[28], confirmed the capability of EXPERT in longitudinal microbiome analysis.

Discussion

This work presents an enabling technology for microbial source tracking based on transfer learning, EXPERT, for source tracking microbial community samples in different context-dependent applications, including prediction and monitoring the development of diseases and longitudinal studies. Based on transfer learning techniques, it provides a fast, accurate, flexible, and interpretable computational approach that could quickly adapt the supervised model to source tracking samples from understudied and newly discovered biomes. According to the needs of different applications, EXPERT can quantitatively assign the contribution of biomes to a specific microbial community, in both ontology-aware and context-aware manners, providing a method that could potentially illuminate the microbial dark matters.

The utility of EXPERT is established in three contexts. First, we used EXPERT as it was

originally intended—to quantify the contribution of different source biomes. In this context, we were able to address questions about more detailed biomes, including microbial communities from infants of different stages. Specifically, using EXPERT, we quantitatively reaffirmed the findings of J. Stokholm et al.[14], suggesting that for 12 months infant, either delivered by cesarean section or vaginal delivery, there is no significant difference in their gut microbial communities with those of the adults. Second, we used EXPERT for disease diagnosis. In this context, EXPERT can serve for the early detection of host health status and the scrutinizing of progression of cancer. Third, EXPERT could be used for longitudinal analysis to better understand the dynamics of the microbial communities. In this context, EXPERT can identify important events such as "turning point" along the timeline, which might be linked to many aspects of human physiology and health, including obesity[18], inflammatory diseases[24], cancer[26], metabolic diseases[18], aging, etc. These context-dependent applications have highlighted the necessity of the EXPERT model for microbiome studies, especially when facing the understudied or newly discovered microbiome data.

The current approaches, including EXPERT, inevitably come up with multi-facet challenges[8] of the microbial dark matters . One is about the ontology structure, for which a better representation might be a graph neural network[29] rather than a tree-like hierarchical structure, for more precise quantification of contributions from biomes. Another is when facing countless context-dependent applications, a collection of transfer models (like Transfer (GM), Transfer (HM), Transfer (DM), etc.) might also be needed to quickly adapt for applications either for environmental source tracking, large-scale time-series analysis, or global scale pattern discovery.

Taken together, EXPERT is an ontology-aware neural network method based on biome ontology information and transfer learning techniques, which may contribute to obtaining biological insights from understudied or emerging microbiome dark matter. Combining the supervised model-based efficiency and accuracy, together with the flexibility of the transfer learning approach[12], EXPERT has enabled a broad spectrum of context-aware applications, including adaptation to emerging data and host health monitoring as well as longitudinal monitoring of microbial communities.

References

1. Gilbert, J. A., Jansson, J. K. & Knight, R. The Earth Microbiome project: successes and aspirations. *BMC Biol.* **12**, 69 (2014).

2. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* vol. 551 457–463 (2017).

3. Turnbaugh, P. J. *et al.* The human microbiome project. *Nature* vol. 449 804–10 (2007).

4. Integrative, H. M. P. R. N. C. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe* vol. 16 276–89 (2014).

5. Gonzalez, A. *et al.* Qiita: rapid, web-enabled microbiome meta-analysis. *Nature methods* vol. 15 796–798 (2018).

6. Mitchell, A. L. *et al.* MGnify: the microbiome analysis resource in 2020. *Nucleic acids research* vol. 48 D570–D578 (2020).

7. Wu, S. *et al.* GMrepo: a database of curated and consistently annotated human gut metagenomes. *Nucleic Acids Res* vol. 48 D545–D553 (2020).

8. Bernard, G., Pathmanathan, J. S., Lannes, R., Lopez, P. & Bapteste, E. Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery. *Genome Biol Evol* 9 (2018).

9. Knights, D. *et al.* Bayesian community-wide culture-independent microbial source tracking. *Nature methods* vol. 8 761–763 (2011).

10. Shenhav, L. *et al.* FEAST: fast expectation-maximization for microbial source tracking. *Nature Methods* vol. 16 627 (2019).

11. Zha, Y. *et al.* Ontology-Aware Deep Learning Enables Ultrafast, Accurate and Interpretable Source Tracking among Sub-Million Microbial Community Samples from Hundreds of Niches. *bioRxiv* 2020.11.01.364208 (2020) doi:10.1101/2020.11.01.364208.

12. Pan, S. J. & Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. DATA Eng.* **22**, 15 (2010).

13. Bäckhed, F. *et al.* Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe* vol. 17 690–703 (2015).

14. Stokholm, J. *et al.* Delivery mode and gut microbial changes correlate with an increased risk of childhood asthma. *Sci. Transl. Med.* **12**, (2020).

15. Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology* vol. 71 8228–8235 (2005).

16. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory* vol. 37 145–151 (1991).

17. Gupta, V. K. *et al.* A predictive index for health status using species-level gut microbiome profiling. *Nat Commun* vol. 11 4635 (2020).

18. Chen, X. & Devaraj, S. Gut Microbiome in Obesity, Metabolic Syndrome, and Diabetes. *Curr. Diab. Rep.* **18**, 129 (2018).

19. Shreiner, A. B., Kao, J. Y. & Young, V. B. The gut microbiome in health and in disease. *Curr. Opin. Gastroenterol.* **31**, 69–75 (2015).

20. Ni, J., Wu, G. D., Albenberg, L. & Tomov, V. T. Gut microbiota and IBD: causation or correlation? *Nat. Rev. Gastroenterol. Hepatol.* **14**, 573–584 (2017).

21. Theriot, C. M. & Young, V. B. Interactions Between the Gastrointestinal Microbiome and Clostridium difficile. *Annu. Rev. Microbiol.* **69**, 445–461 (2015).

22. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **48**, D9–D16 (2020).

23. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol* vol. 10 766 (2014).

24. Balato, A. *et al.* Human Microbiome: Composition and Role in Inflammatory Skin Diseases. *Arch. Immunol. Ther. Exp. (Warsz.)* **67**, 1–18 (2019).

25. Sirisinha, S. The potential impact of gut microbiota on your health:Current status and future challenges. *Asian Pac. J. Allergy Immunol.* **34**, 249–264 (2016).

26. Zhu, Q., Gao, R., Wu, W. & Qin, H. The role of gut microbiota in the pathogenesis of colorectal cancer. *Tumor Biol.* **34**, 1285–1300 (2013).

27. Liu, H. *et al.* Resilience of human gut microbial communities for the long stay with multiple dietary shifts. *Gut* vol. 68 2254–2255 (2019).

28. Smits, S. A. *et al.* Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. *Science* vol. 357 802–806 (2017).

447 29. Wu, Z. *et al.* A Comprehensive Survey on Graph Neural Networks. *IEEE Trans.*
448 *Neural Netw. Learn. Syst.* **32**, 4–24 (2021).
449 30. Leinonen, R. *et al.* The European Nucleotide Archive. *Nucleic Acids Res.* **39**,
450 D28–D31 (2011).
451 31. Kuczynski, J. *et al.* Using QIIME to analyze 16S rRNA gene sequences from
452 Microbial Communities. *Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al*
453 **CHAPTER**, Unit10.7 (2011).
454 32. Kebschull, J. M. *et al.* High-Throughput Mapping of Single-Neuron Projections
455 by Sequencing of Barcoded RNA. *Neuron* **91**, 975–987 (2016).
456 33. Mukherjee, S. *et al.* Genomes OnLine Database (GOLD) v.8: overview and
457 updates. *Nucleic Acids Res.* **49**, D723–D733 (2021).
458 34. Schriml, L. M. *et al.* Human Disease Ontology 2018 update: classification,
459 content and workflow expansion. *Nucleic Acids Res* vol. 47 D955–D962 (2019).
460 35. Caruana, R. Multitask learning. *Machine learning* vol. 28 41–75 (1997).
461 36. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by
462 back-propagating errors. *nature* **323**, 533–536 (1986).
463 37. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv*
464 *preprint arXiv:1412.6980* (2014).
465
466

Methods

Datasets

We used eight datasets to evaluate the performance of EXPERT.

For systematic assessment of our general model, the dataset was obtained from MGnify, which consists of 118,592 communities collected from 132 biomes. Among them, 52,537 samples originated from human biomes, 14,045 samples originated from mammal biomes, 7,189 samples originated from terrestrial biomes, 27,667 samples originated from aquatic biome. These samples were analyzed by MGnify[6] before February 2020 (**Supplementary Table S2 Supplementary Table S3**).

For systematic assessment of our human model, the dataset was a part of the first dataset, in which 52,537 communities from 28 human biomes were selected (**Supplementary Table S2**, **Supplementary Table S4**).

We also used emerging data in 2020 from MGnify[6]. Which consists of 34,209 communities collected from 35 biomes. Throughout the dataset, 3,421 samples belong to 8 biomes were newly added by MGnify[6] after January 2020 (**Supplementary Table S2**, **Supplementary Table S7**).

For source tracking the succession of infant gut microbiome, the dataset was obtained from MGnify[6] which consists of 392 fecal samples collected from 98 infant and their biological mothers. Among them, 85 infants were born by vaginal delivery and 13 infants were born by cesarean section. The infant samples were collected at three time points including birth, fourth-month and twelfth-month. The maternal samples were collected during the first week after delivery (**Supplementary Table S2**, **Supplementary Table S8**).

For disease modeling, the dataset was obtained from GMrepo[7], including 13,642 communities collected from feces of hosts diagnosed with 20 different phenotypical status (different kinds of diseases, plus healthy controls, **Supplementary Table S2**, **Supplementary Table S9**).

For cancer monitoring, the dataset was obtained from GMrepo[7], which consists of 16, 93,

126, 196, and 204 communities respectively collected at CRC stage 0, I, II, III, and IV (**Supplementary Table S2**, **Supplementary Table S10**).

For longitudinal dietary shifts analysis, the dataset was obtained from ENA[30] and anlyzed mainly using QIIME (version 1.9)[27,31], including 280 samples collected from travelers and other native persons, (**Supplementary Table S2**, **Supplementary Table S11**).

For investigation of seasonal patterns within Hadza hunter-gatherers' gut microbiome, The dataset used was obtained from ENA and analyzed mainly by using MapSEQ[32], including 203 samples collected from the Hadza hunter-gatherers' gut (referred to Wu, S. et al.[7] for details, **Supplementary Table S2**, **Supplementary Table S12**).

Data preprocess

Relative abundance calculation according to reference database

In order to bypass the impact from the sequencing depth of the rich-sourced data, we firstly regularized the abundance data by calculating relative abundances according to only the taxa mapped to our phylogenetic tree (**Supplementary Table S1**), which is a part of taxonomical classification tree in NCBI taxonomy database[22] and reflects the input feature set for EXPERT model (as explained below).

Universal feature set for heterogeneous transfer learning

In order to realize knowledge transfer among mulit-faceted microbial source tracking applications, we utilized a uniform collection of features in all these applications. Such universal feature set was established according to only the variance of relative abundances at genes level: Among the first dataset (for which the most comprehensive), 6,006 genura with variance of relative abundances above the threshold average variance x 10-3. We constructed a phylogenetic tree based on the classification of these 6,006 genera (three taxon ranks were included: "phylum", "order", "genus"), and served the tree for abundance mapping (**Supplementary Note 2**).

Z-score standardization of relative abundances

In order to speed up the optimization of EXPERT model, we standardized the relative

abundances before feeding them into the model, by applying z-score standardization on normalized relative abundances.

Establishment of the ontology

Throughout the paper, the reference information utilized for biome ontology constructions are: (1) hierarchical biome classification from MGnify database[6] and ecosystem classification paths from GOLD database[33] for assessing the generalizing performance of general knowledge (**Supplementary Table S13**), assessing the performance in context-dependent settings (**Fig. 2a**), and assessing the performance on emerging data (**Fig. 3a**); (2) sampling time of infant gut[13] and delivery modes of infants[13] for source tracking the succession of infant gut microbiome (**Fig. 4a, Supplementary Fig. 4**); (3) disease classification from NCBI MeSH database[22] and Human Disease Ontology[34] for disease diagnosis and monitoring (**Fig. 5a**); (4) geographic origins of samples[27] and sampling time[28] for longitudinal data analyses (**Fig. 6a,b**).

The EXPERT model

The probabilistic model of EXPERT

Considering a query sample q represented by its community structure, as well as its potential sources represented by a biome ontology O, to quantify contributions $\hat{y}_q$ from ontologically organized biome sources to q, we employed a multi-task neural network to learn a mapping M from a series of source samples $s \in D_S$ to their biome sources, $y_s = (y_s^2, \ldots, y_s^I)$ (where y_s^2 is biome source for source sample s in the second layer of the biome ontology), and then apply M on q to determine the contributions from biome sources.

$$\hat{y}_q = \left(\hat{y}_q^i\right)_{0<i\le I_O} = M(q)$$

Fast inference via forward propagation

We assume that the learning in the higher ontology layers, such as the distinguishing

between "Human" and "Environmental", are helpful to the learning in the lower ontology layers, such as the distinguishing between "Human: Digestive System", "Environmental: Aquatic" and "Environmental: Terrestrial"[35]. EXPERT integrates the representation of the lower layer (which is calculated by its layer-specific modules M_{inter}, into higher layer), by employing several integrator module M_{integ}. Therefore, together with layer-specific output module M_{output}, the representation of the contributions is given by

$$M(q) = (M^i_{output}(R^i_{integ}))_{0<i\le I_O}$$

Where

$$R^i_{integ}(q) = \begin{cases} M^i_{integ}(M^i_{inter}(M_{base}(q)), 0) if \quad i = 1 \\ M^i_{integ}(M^i_{inter}(M_{base}(q)), R^{i-1}_{integ}) otherwise \end{cases}$$

Robust optimization via backward propagation and transfer learning

We also assume that the quantification of contributions from a biome ontology is helpful to the quantification of the contributions from a series of associated biome ontologies[12]. Considering $\tilde{M}_{base}$ of a source model as a static mapping, the parameters of the rest modules $\hat{w}.$ can be solved using gradient descent as well as backpropagation algorithm[36,37]

$$\hat{w} = \underset{\hat{w}}{\operatorname{argmin}} \sum_{i=0}^{I_O} \alpha\,(B^i_O)\, L\left(\hat{y}^i_s(\hat{w}), y^i_s\right)$$

Where

$$\alpha(B^i_O) = \frac{B^i_O}{B_O f}$$

$$L\left(\hat{y}^i_s, y^i_s\right) = \sum_{b\in O^i} \left(CrossEntropy\left(\hat{y}^i_s(b), y^i_s(b)\right)\right)$$

$$b \in O^i$$

B^i_O stand for the number of biomes contained in the i-th layer of the biome ontology

569 O^i,O^i stand for the i-th layer of the biome ontology O^i

570 Then, optimizing the parameters of the entire model (including $\tilde{M}_{base}$), the parameters of
571 the rest modules w can be solved by using gradient descent as well as backpropagation
572 algorithm[36,37]

573 $$w = \underset{w}{\operatorname{argmin}} \sum_{i=0}^{I_O} \alpha\,(B_O^i)\, L\left(\hat{y}_s^i(w), y_s^i\right)$$

574 For independent optimization (optimization based on completely random initialization),
575 EXPERT straightforwardly optimizes the entire model. See **Supplementary Note 5** for
576 detailed description for optimization.

577

578 **Cross-validation**

579 We evaluated performances of EXPERT utilizing multiple cross-validation settings.
580 When assessing the performances of general model (including adapted model on
581 emerging data), human model, and disease model, we repeatedly performed random
582 cross-validation for five times--each time randomly select 10% of the dataset as queries
583 and the rest as sources.

584 When source tracking the succession of infant gut microbiome, we performed
585 proportional sampling cross-validation by randomly select 10% of the samples from
586 mother, infants at birth, 4 months and 12 months of age as query, respectively. This
587 process was also repeated for five times.

588 When analyzing the compositional shifts of ten Chinese individuals during the travel to
589 Trinidad and Tobago, we performed individual-level leave one out cross-validation--each
590 time consider all samples of an individual as queries and all samples from the other
591 individuals as sources. Notably, we didn't consider samples from MT10 as sources as he
592 returned with an exception.

593 When analyzing the compositional shifts of Hadza hunter-gatherers, we performed

proportional sampling cross validation by randomly select 50% of the samples from from young adults (18 <= age < 50) in each season ("2013-Late Dry", "2014-Early Wet", "2014-Late Wet", "2014-Early Dry", and "2014-Late Dry") as queries and the rest as sources. This process was also repeated for five times.

Performance measures

To benchmark and compare the performance of EXPERT model based on knowledge transfer (Tranfer (GM), Tranfer (HM), Tranfer (DM),) and independent model, as well as other methods, we used these measures:

$$TP_b(t) = \sum_s I(\hat{y}_s(b) > t \wedge b \in y_s)$$

$$TN_b(t) = \sum_s I(\hat{y}_s(b) < t \wedge b \notin y_s)$$

$$FP_b(t) = \sum_s I(\hat{y}_s(b) > t \wedge b \notin y_s)$$

$$FN_b(t) = \sum_s I(\hat{y}_s(b) < t \wedge b \in y_s)$$

$$\begin{aligned} TPR_b(t) &= \frac{TP_b(t)}{TP_b(t) + FN_b(t)} \\ FPR_b(t) &= \frac{FP_b(t)}{FP_b(t) + TN_b(t)} \\ Recall_b(t) &= \frac{TP_b(t)}{TP_b(t) + FN_b(t)} \\ Precition_b(t) &= \frac{TP_b(t)}{TP_b(t) + FP_b(t)} \end{aligned}$$

Where TP is true positive, TN is true negative, FP is false positive, FN is false negative, \hat{y}_s(b) is the quantified contribution from a biome source b for a microbial community sample s, threshold t \in [0,1] with a step size of 0.01, y_s is a set of actual biomes for a sample s, and I is a logical operation function, the value of I is 1 when the result of logical operation is TRUE, else 0.

Then, three evaluation metrics (Accuracy, F-max, AUROC) was introduced. These

611 evaluation metrics were calculated with the following formulas:

$$Accuracy_b(t) = \frac{TP_b(t) + TN_b(t)}{TP_b(t) + TN_b(t) + FP_b(t) + FN_b(t)}$$

612 $$F\text{-}max_b(t) = \max_t \frac{2Precision_b(t)Recall_b(t)}{Precision_b(t) + Recall_b(t)}$$

$$AUROC_b(t) = \sum_{t=0}^{1} \frac{(TPR_b(t) + TPR_b(t+0.01))(FPR_b(t+0.01) - FPR_b(t))}{2}$$

613 Then, we treated the average performance across all biomes in a specific ontology layer
614 as the performance of the entire model.

615

616 **Statistical analysis**

617 Statistical analyses of the contributions have been performed utilizing Wilcoxon test, at
618 the significance level $\alpha = 0.05$. For all the tests, when the p-value associated is lower
619 than the significance level, one should reject the null hypothesis H0, and accept the
620 alternative hypothesis Ha.

621

622 **Data distribution**

623 Throughout the paper, the box-plot elements are: center line, median; box limits, upper
624 and lower quartiles; whiskers, 1.5× interquartile range (IQR); points and outliers. The
625 Violin plot is also used for data distribution analysis, mainly for comparison.

626

627 ## Data availability

628 The collected samples from MGnify/GMrepo were annotated with their associated
629 biomes/phenotypes in **Supplementary Table S3-4** and **Supplementary Table S7-12**. All
630 the processed data are uploaded and hosted at https://github.com/HUST-NingKang-
631 Lab/EXPERT-use-cases.

Code availability

All source codes have been uploaded to the website at: https://github.com/HUST-NingKang-Lab/EXPERT. Detailed parameters of software and package used in this study are provided in Supplementary Table S2.

Acknowledgments

This work was partially supported by National Science Foundation of China grant 32071465, 31871334 and 31671374, and the Ministry of Science and Technology's national key research and development program grant (No. 2018YFC0910502).

Author contributions

KN and HC conceived of and proposed the idea, and designed the study. HC, QY, GX, NW, CS, and SW performed the experiments and analyzed the data. HC, QY, YZ, WC, and KN contributed to editing and proof-reading the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable.

Figures

Figure 1

Figure 1: Illustration of EXPERT's data processing, knowledge transfer process, and multi-faceted applications. a. The data preprocessing workflow of EXPERT. For each sample, we mapped the taxonomical abundances to the phylogenetic tree, which is compatible with NCBI phylogenetic tree, to obtain a regular abundance matrix. And then, the matrix is normalized and standardized in order to obtain a standard abundance matrix. Blue and White boxes indicate entities of data and operations, respectively. The final input data is a matrix, in which each row represents abundances for a sample, and each column represents abundances for a genus (in total, 6,006 genera were used in this study). **b.** The general ontology-aware neural network (ONN) model (built based on a biome ontology with more than a hundreds biomes, referred to https://github.com/HUST-NingKang-Lab/ONN4MST for details) has a fixed structure and poor scalability, thus cannot perform well on source tracking for context-dependent applications. **c.** EXPERT can adapt the general knowledge of an existing model to such applications through three steps: transfer (reuse parameters of an existing model and reinitialize context-dependent layers according to prior pieces of knowledge about the context, green arrows), adaptation (quickly optimize only the context-dependent layers using iterative forward-backward propagation, green circular arrows), and finetuning (further optimize the entire model using the iterative forward-backward propagation). The existing model is a general ONN model (either in GM, HM, or DM model used in this work, refer to **Supplementary Table S2** for details) to be adapted, with two fully connected layers (relatively independent to context) and a series of context-dependent layers (highly specified to a context, **Supplementary Fig. 1**). Red arrows indicate the three steps of the transfer process. Different background colors of the model indicate the suitability of the layers to the context-dependent application. This model could be applied to community samples collected from new biome ontology (representing context-dependent knowledge) that cannot be source-tracked by the general ONN model directly. The transferred model can serve a broad spectrum of source tracking applications (based on research purposes, further illustrated in **Fig. 1d**). **d**. The applications of knowledge transfer utilizing EXPERT. **i.** EXPERT quantifies the source biome contributions for query communities. **ii.** EXPERT enables adaptation of emerging microbiome data from understudied biomes (also referred to as Grafting). **iii.** Knowledge transfer can source track among more detailed biomes (also referred to as Scattering), such as quantifying gut microbial communities' contribution from mothers to their developing infants. **iv.** EXPERT enables disease diagnosis, as well as disease progression monitoring. **v.** EXPERT enables investigation of dynamics of microbial communities along the timeline.

Figure 2

Figure 2. Assessment of the transfer learning model on context-dependent application. a. Biome ontology of human-associated biomes, used for assessing the generalizing performance of EXPERT. The ontology is constructed based on 52,537 samples, with four layers in total (from top down). The biomes with orange color indicates biomes with enough samples (> 100) within each biome. The biomes with orange color and in dashed line box with yellow background indicates biomes for comparison with FEAST, which are selected as they represent most detailed biomes (comply with the culture-independent assumption of SourceTracker and FEAST,), and each of then has enough samples (> 100). **b.** The performance (AUROC and F-max, X-axis) of Transfer (GM) for each biome (Y-axis) by using repetitive cross-validation (5 times, 90% for training, 10% for testing, see **Methods**). **c.** The performances (AUROC and F-max, Y-axis) of three models (X-axis): Transfer (GM) model, Transfer (GM0) model, and Independent model (built based on independent training). **d.** The training time and query time (Y-axis) across different optimization schemes (Independent, Transfer, Transfer (0), X-axis). **e.** Comparison of Transfer (GM) with FEAST in source tracking accuracy (first Y-axis) and efficiency (second Y-axis). FEAST10, FEAST20, FEAST30 represent FEAST based on using 10, 20 and 30 samples per biomes for all biomes with enough samples (> 100) in the human dataset as source samples, respectively. Transfer (GM) and Transfer (GM0) refer to models built based on transferring knowledge from the general EXPERT model with and without finetune, respectively.

Figure 3

Figure 3. EXPERT enables adaptation for emerging microbiome data. a. Biome ontology (2020) constructed based on 34,209 samples collected and analyzed by MGnify after January 2020. These samples are from existing biomes and the newly discovered biomes, hierarchically organized in four layers (indicated by the black interval on the left, from top down). The yellow nodes indicate newly discovered biomes in the MGnify database, which have hardly ever been analyzed before January 2020. **b.** Generalizing performance (AUROC and F-max, Y-axis) of models optimized using different training schemes (Independent model based on 34,209 samples used in this application, and Transfer (GM) model built based on transferring knowledge from the general model with finetune, X-axis). **c.** Searching time per sample by different models. **d.** Estimated biome contributions for query samples, exemplified by samples from two newly discovered biomes ("Fish" and "Birds") on different layers. Notice that on the third layer, the contribution of Fish is 88.29% for query samples from Fish, and the contribution of Birds is 73.78% for query samples from Birds; and on the fourth layer, the contribution of Digest system (birds) is 72.22% for query samples from Birds; while on the fifth layer, the contribution of Ceca is 56.35% for query samples from Birds.

Figure 4

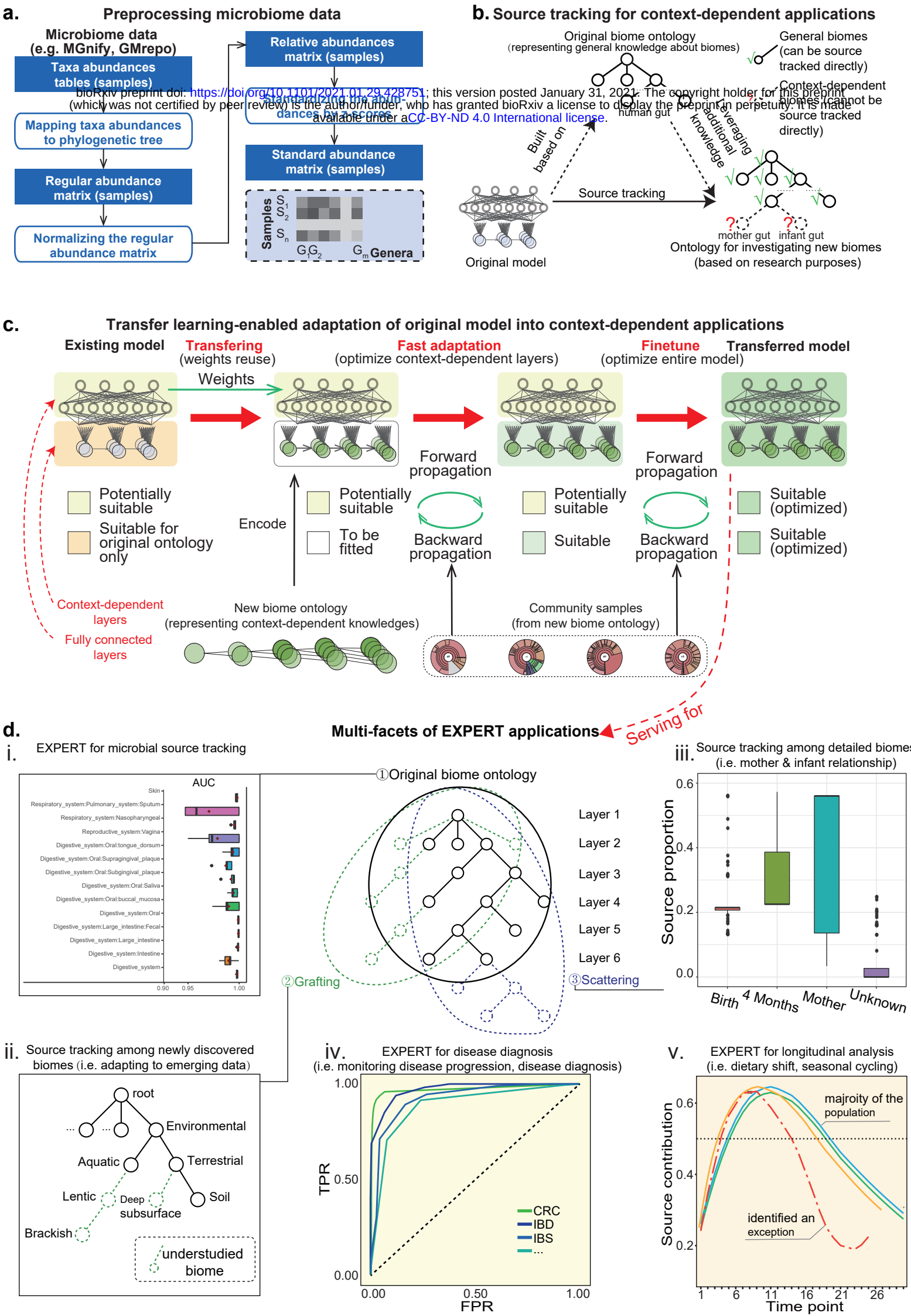
Figure 4. EXPERT's performance in characterizing gut microbial community development over time for infants. a. The biome ontology, corresponding to infant samples collected from the ENA database. Layer 2 is based on sampling time, while layer three is based on delivery modes (Backhed et al.). For this part of the study, sources include gut microbiome of the mother, infant at birth, and four months, queries include the gut microbiome of the infant at 12 months. **b.** Estimated contributions of biomes by different models of EXPERT, separated by two delivery modes and transferred from the general model (Transfer (GM), above) and human model (Transfer (HM), below). **c.** Distribution of infant gut microbial communities during their first year, using principal-coordinates analysis (PCoA) and distance metric of Jensen Shannon divergence. Infant gut given by different delivery modes developed in an adult-like pattern over time, and the compositional shifts of infant's gut at the age of 12 months are indistinguishable. Dotted line refers to samples delivered by vaginal delivery, and the full line refers to samples delivered by cesarean section. The baby of 4 month is abbreviated to baby 4M., the baby of 12 month is abbreviated to baby 12M. "C" represents cesarean section, "V" represents vaginal delivery. Top panel: samples from infant's gut are plotted according to their source and collection date on the Y-axis, and position on the X-axis is plotted according to their first principal coordinate in the PCoA. **d.** The overall performance of models generated based on different models, in which Independent model based on the samples and biomes used in this application, Transfer (GM) and Transfer (HM) refer to models built based on transferring knowledge from the general model and human model with finetune, respectively. Violin plot: Red represents AUROC, and blue represents F-max. **e.** Samples whose predicted source contributions were not consistent with the ground truth by the independent model, but were consistent with the ground truth by the Transfer (HM) model. Different colors refer to different biomes: purple represents unknown sources, blue represents the baby of 12th month, green represents the baby of 4th month, red represents the baby of birth. Baby of 4 month is abbreviated to baby 4M., the baby of 12 month is abbreviated to baby 12M.

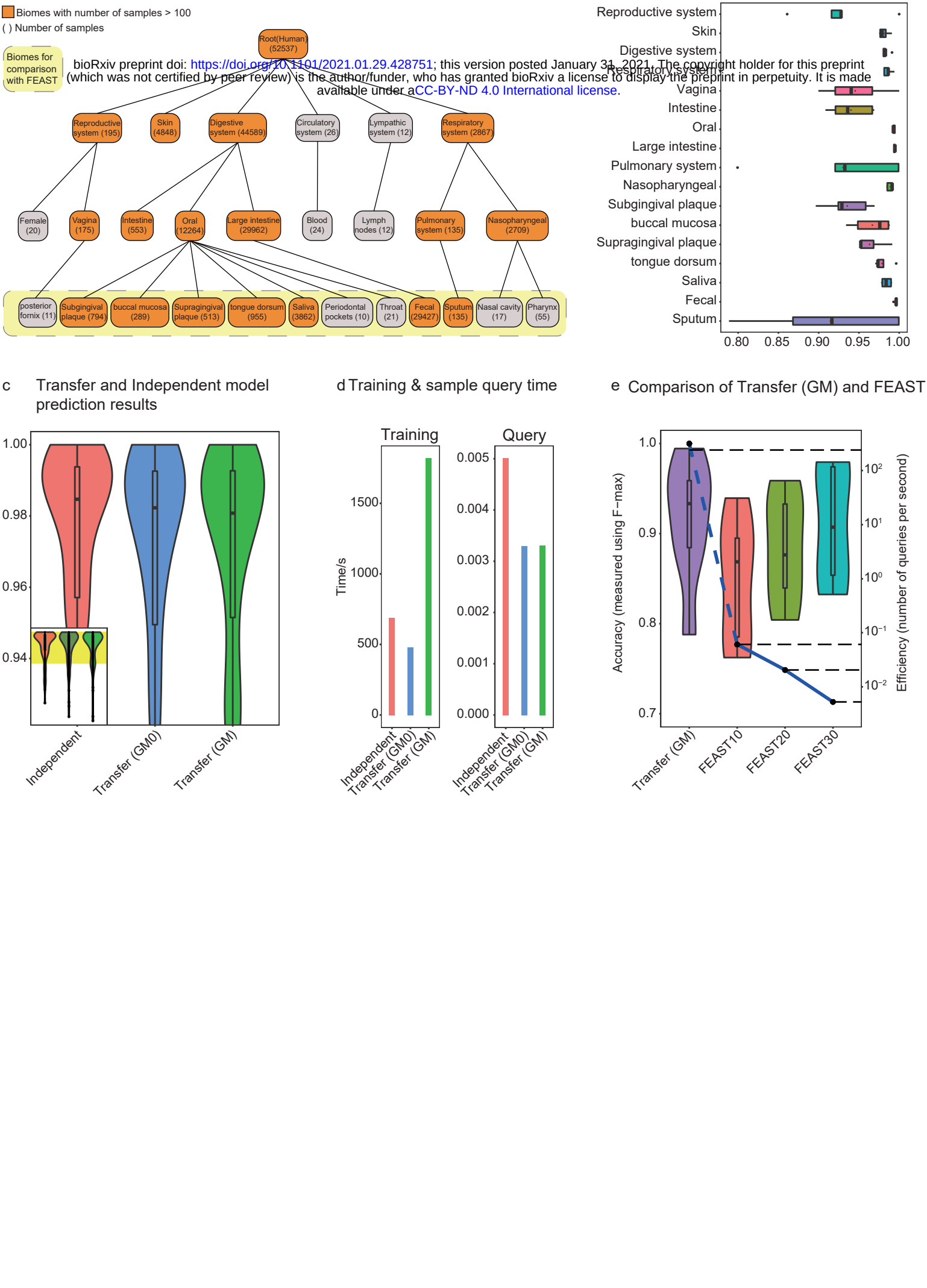
Figure 5

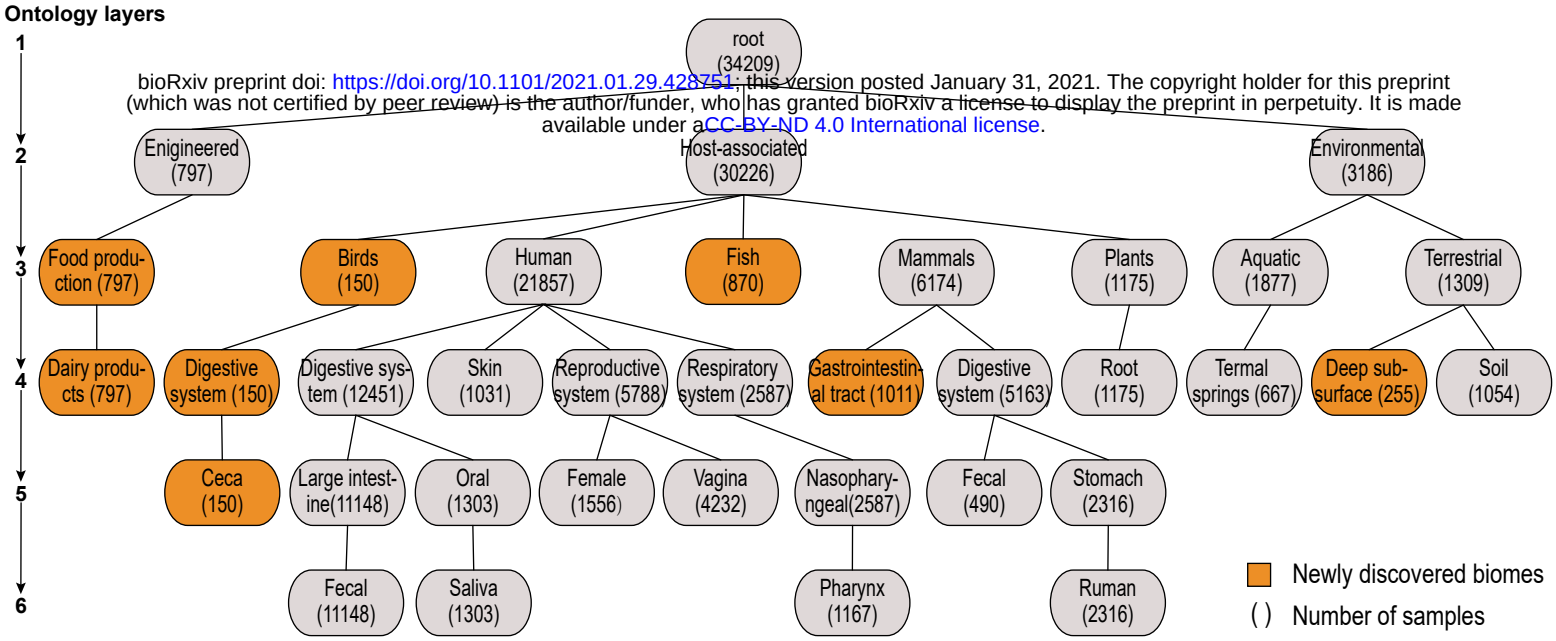
Figure 5. EXPERT for disease diagnosis and monitoring. a. Illustration of knowledge transfer utilized for disease phenotype differentiating. The knowledge transfer between models trained using different datasets (named Source dataset and Target dataset) are illustrated using different colors (white for human, yellow for disease, and red for colorectal cancer). In health status prediction, the knowledge from human model containing 53,553 samples and 28 biomes, were tranferred to disease model containing 13,642 samples and 20 biomes. **b.** the disease ontology constructed based on host phenotype and considering NCBI MeSH database (https://www.ncbi.nlm.nih.gov/mesh) and Human Disease Ontology (https://www.ebi.ac.uk/ols/ontologies/doid) as references of disease classification. The disease ontology includes 20 phenotypes (19 different disease and infections, plus healthy control) distributed in seven different layers (X-axis). **c.** The diagnostic model's performance on each phenotype, evaluated based on repetitive cross-validation (5 times, 90% for training, and 10% for testing) and biome-specific evaluation (see **Methods**). A dashed line indicates AUROC of 0.8. **d.** The Transfer (HM) model's overall performances on differentiation of diseases. **e.** Illustration of knowledge transfer utilized for differentiating disease phenotypes. We generated two transferred models: Transfer (HM) and Transfer (DM) refer to models built based on transferring knowledge from the human model and disease model with finetune, respectively. In CRC progression monitoring, the knowledge from Transfer (HM) containing 53,553 samples and 28 biomes, as well as Transfer (DM)containing 13,642 samples and 20 biomes, were tranferred to CRC applications in which there are 5 stages. **f.** The five stages of colorectal cancer progression, and the number of samples for each stage. Among them, stage 0 stand for healthy control. **g.** The distribution of gut microbiome, visualized by PCoA (utilizing distance metric of weighted-Unifrac). **h.** The performances (AUROC and F-max) of different models (Independent model based on independent training, Transfer (HM) based on the human model, and Transfer (DM) based on disease model). **i.** The stage-specific performances (AUROC, see **Methods**) of EXPERT on different CRC stages, evaluated based on repetitive cross-validation (5 times, 90% for training, and 10% for testing).

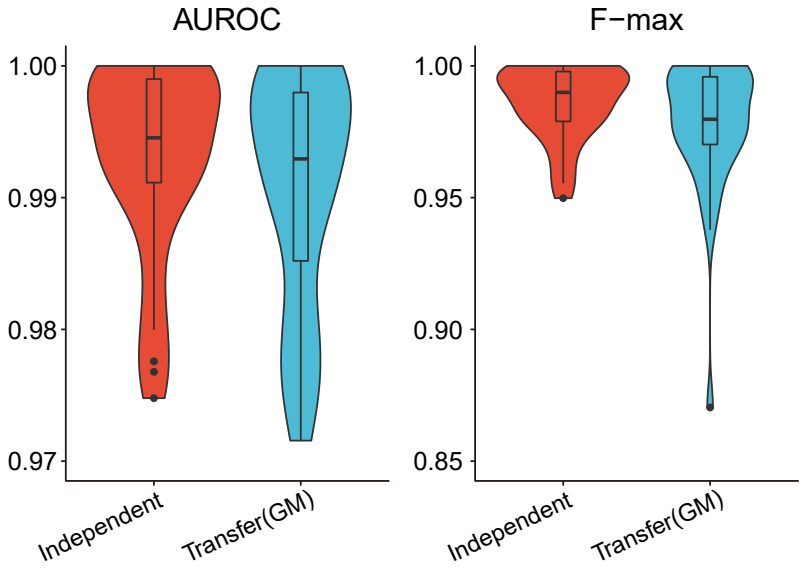
Figure 6

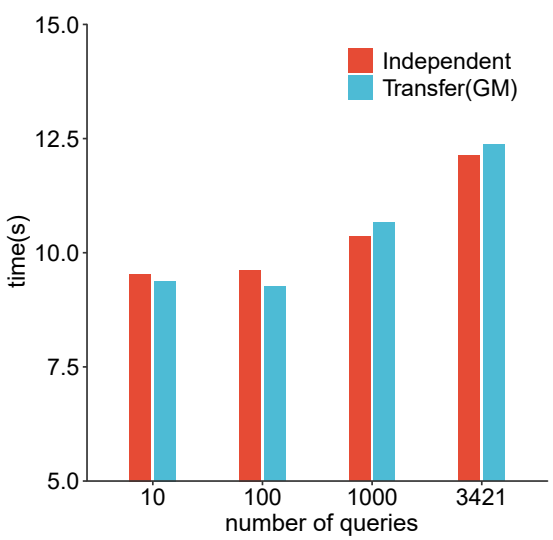
Figure 6. The EXPERT model could reveal compositional shifts within hosts' gut microbiome over time. a. Biome ontology used for analyzing dynamic patterns among gut microbiome during international travel. Two countries (China and Trinidad and Tobago, black nodes) leveraged in the travel are modeled into the biome ontology. Additional information (dotted green nodes) can be inferred based on the model's prediction and additional metadata. **b.** Sampling time points (30 time points in total, X-axis) of metagenomic samples for each individual (Y-axis). Three stages are included: before travel (in China), during travel (in Trinidad and Tobago), and after travel (returned to China), and transitions between these stages (flights) are marked on the top. Different colors indicate samples from different individuals. **c.** Estimated source contributions (Y-axis) by EXPERT over the timeline (also 30 time points in total, X-axis), using individual-level leave one out and considering additional samples from native people as potential sources (see **Methods**). **d.** Biome ontology used for analyzing seasonal patterns among Hadza hunter-gatherers' gut microbial communities. Samples are divided into "Dry" and "Wet" biomes in the biome ontology, Additional information (dotted green nodes) can be inferred based on the model's prediction and additional metadata. **e.** Estimated source contributions (Y-axis) by EXPERT over the timeline (15 time points in total, X-axis), by using samples from the "Dry" season as sources. Result are evaluated based on repetitive proportional sampling cross-validation (see **Methods**).

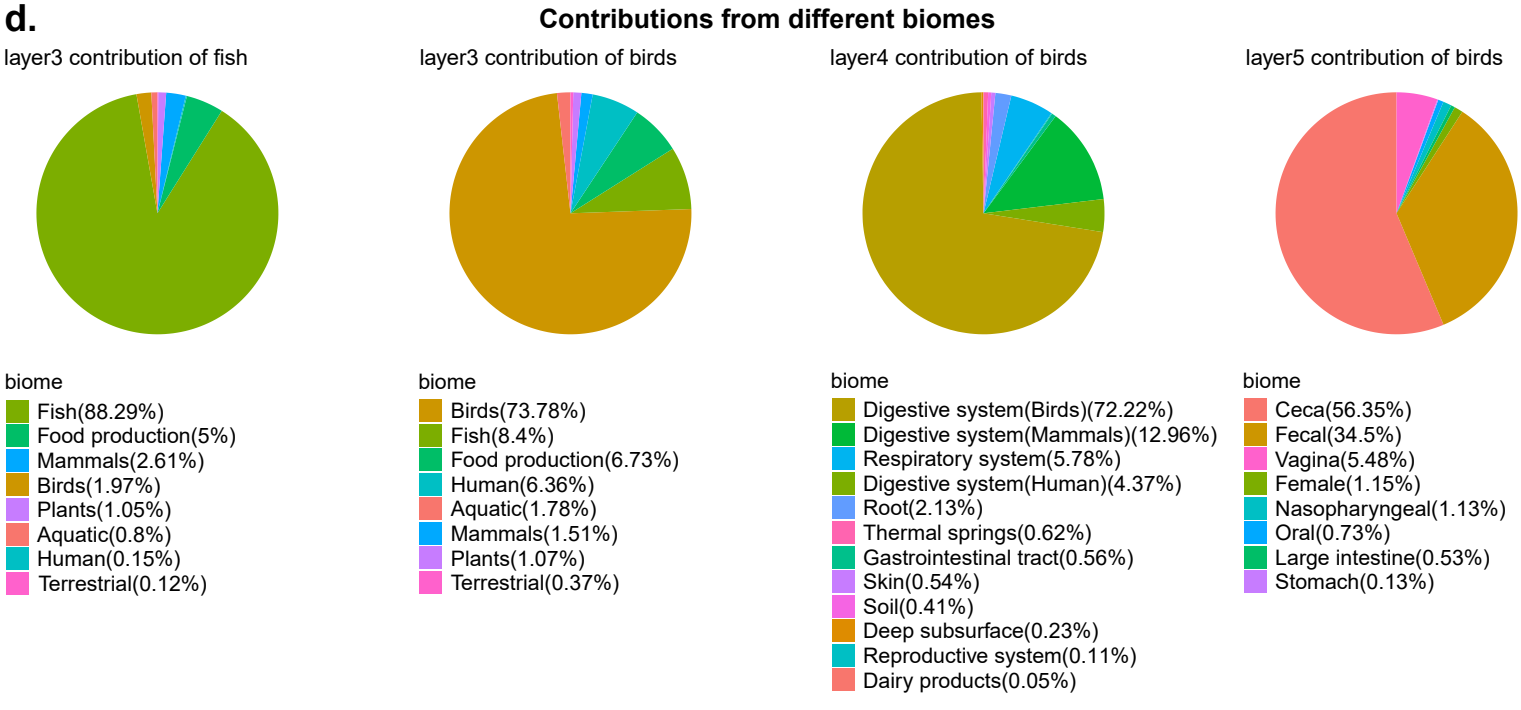


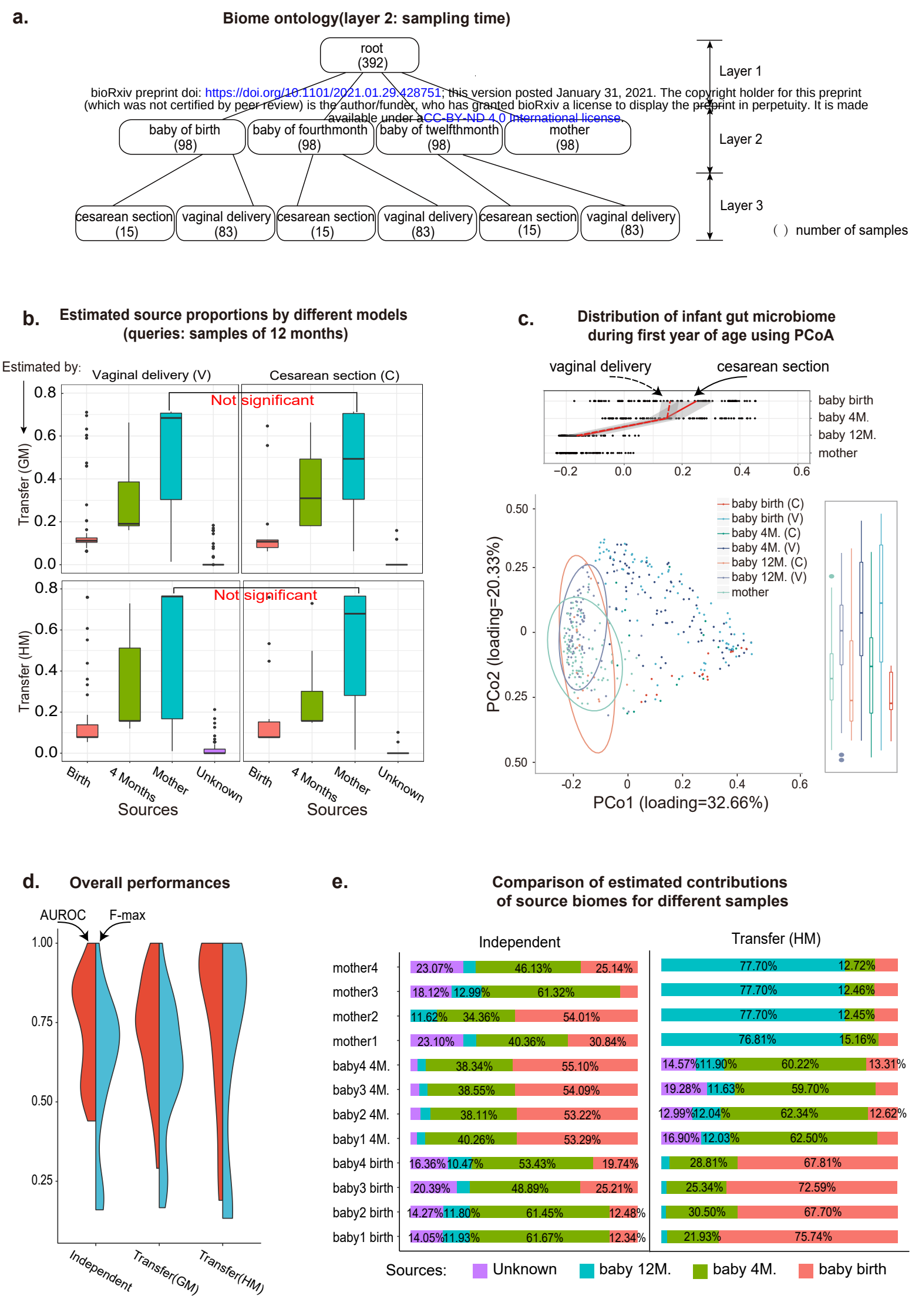


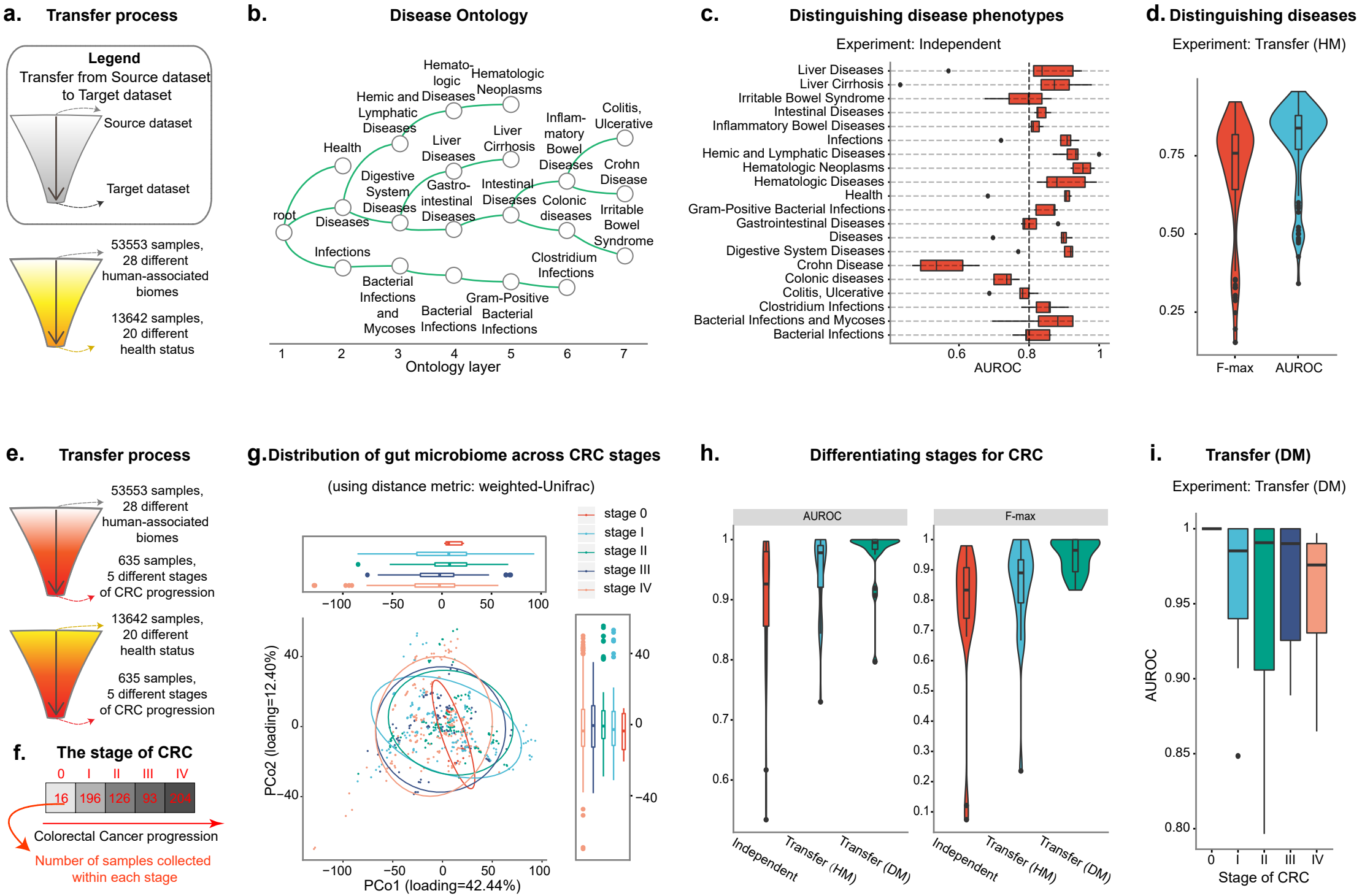











a.
Transfer process
Legend
Transfer from Source dataset to Target dataset
Source dataset
Target dataset
53553 samples, 28 different human-associated biomes
13642 samples, 20 different health status
b.
Disease Ontology
root
Diseases
Infections
Health
Digestive System Diseases
Hemic and Lymphatic Diseases
Hemato-logic Diseases
Hematologic Neoplasms
Liver Diseases
Liver Cirrhosis
Gastro-intestinal Diseases
Intestinal Diseases
Bacterial Infections and Mycoses
Bacterial Infections
Gram-Positive Bacterial Infections
Inflam-matory Bowel Diseases
Colonic diseases
Colitis, Ulcerative
Crohn Disease
Irritable Bowel Syndrome
Clostridium Infections
1 2 3 4 5 6 7
Ontology layer
c.
Distinguishing disease phenotypes
Experiment: Independent
Liver Diseases
Liver Cirrhosis
Irritable Bowel Syndrome
Intestinal Diseases
Inflammatory Bowel Diseases
Infections
Hemic and Lymphatic Diseases
Hematologic Neoplasms
Hematologic Diseases
Health
Gram-Positive Bacterial Infections
Gastrointestinal Diseases
Diseases
Digestive System Diseases
Crohn Disease
Colonic diseases
Colitis, Ulcerative
Clostridium Infections
Bacterial Infections and Mycoses
Bacterial Infections
0.6 0.8 1
AUROC
d.
Distinguishing diseases
Experiment: Transfer (HM)
0.75
0.50
0.25
F-max
AUROC
e.
Transfer process
53553 samples, 28 different human-associated biomes
635 samples, 5 different stages of CRC progression
13642 samples, 20 different health status
635 samples, 5 different stages of CRC progression
f.
The stage of CRC
0 I II III IV
16 196 126 93 204
Colorectal Cancer progression
Number of samples collected within each stage
g.
Distribution of gut microbiome across CRC stages
(using distance metric: weighted-Unifrac)
stage 0
stage I
stage II
stage III
stage IV
−100 −50 0 50 100
PCo1 (loading=42.44%)
PCo2 (loading=12.40%)
40
0
−40
h.
Differentiating stages for CRC
AUROC
F-max
Independent
Transfer (HM)
Transfer (DM)
i.
Transfer (DM)
Experiment: Transfer (DM)
AUROC
1
0.95
0.90
0.85
0.80
0 I II III IV
Stage of CRC