**Supplemental Tables and Figures**



**Supplementary Figure 1. The bar plots show the relative abundance of bacterial taxa between healthy and caries-active children in the Qingdao city of China.** (A)Relative abundances of microbial at the phylum, genus, species level, respectively. (B) Relative abundances of microbial at the genus level between H and C group. (C) Relative abundances of microbial at the species level between H and C group. Significance determined by *p*<0.05(\*), *p*≤0.01 (\*\*), *p*≤0.001 (\*\*\*).



**Supplementary Figure 2. Geography-associated oral microbiome.** (**A**)Boxplots indicate the relative abundance of the 8 species between Qingdao and Guangzhou group. (**B**)A classification model classified the city of origin based on 8 key microbial drivers from hosts in the C group. (**C**)Box plot shows the prediction probability of Guangzhou city in caries samples. The probability of Guangzhou city was significantly higher in the Guangzhou samples than in the Qingdao samples from the C group.



**Supplementary Figure 3. Alpha and beta diversity of all samples in term of disease status.** (**A**) Caries samples featured a reduction in alpha diversity. (**B**) The beta diversity measured by Jensen-Shannon distance was distinctive between caries and healthy microbiota.



**Supplementary Figure 4.** Classification models based on taxa at the phylum (**A**), genus (**B**), and species (**C**) levels to discriminate between healthy and caries samples from all geographic locations. The prediction accuracy (AUC) was maximized at the species level. The AUC of all samples-based Random Forests classifier of caries diagnosis is 64.11,77.61 and 88.56 respectively.



**Supplementary Figure 5.** Theperformance of cross-city prediction using each city-specific caries diagnostic models. The blue lines indicate the ROC curves evaluating the caries classification model self-validated in each city. The red lines indicate the ROC curves evaluating the application of a caries classification model from one city to another. The performance of the models and their applications to the other dataset was scored by AUC in the ROC analysis.



**Supplementary Figure 6. Contributions of the eight most discriminant species to microbiota-based classification of host status and geography.** The most discriminatory taxa (N=8) of geography do not show correlated with caries state. The scatterplot shows the relative rank of microbial markers in both Random Forest models for classifying disease status and geographic locations. Dots on the reference line which slope=1 suggests a taxon is equally important to both models.



**Supplementary Figure 7. Caries diagnosis model based on the one species of *Streptococcus mutans* after ruling out eight geography-specific signatures.** Performance of status prediction using only *Streptococcus mutans* as a predictor, the simplified Random Forest model led to lower, yet still decent and meaning performance model (AUC=81.62%).



**Supplementary Figure 8. The frequency of fourteen species occurrence in healthy sample and caries samples**