# Project report - Oral microbiome meta-analysis

# Introduction

The human body consists of roughly equal numbers of human cells and microorganisms [1,2]. The microorganisms found in the oral cavity are collectively called the oral microbiome [3]. There are about 1000 species in the Human oral microbiome as identified in the HOMD database [4] it includes both cultivable and not-yet-cultivated species. The dysbiosis of microbiomes causes several oral diseases including caries. Dental caries is the term used for tooth decay and cavities and it is the most prevalent infectious disease in the oral cavity. Caries in children with primary dentition is known as early childhood caries (ECC) and it affects about half of the children worldwide [5,6].

The main factor for caries development is the biochemical transformations caused by acidogenic microbes on the dental surface. There is a hypothesis of relatively lower bacterial diversity in caries lesions due to the constant acidic exposure. This change can be seen by the dominance of aciduric microbes. Streptococcus muta ns are found to be the main caries-causing pathogens, but previous studies have also identified several other microorganisms involved in caries for example Actinomyces, and Lactobacillus [7] [7] [8] [5]. Several acidogenic species have been identified in ECC microbiome for instance, found Scardovia wiggsiae, Lactobacillus salivarius, Streptococcus mutans, and Parascardovia denticolens [9].

Since the arrival of Next Generation Sequencing (NGS), amplicon sequencing using 16s rRNA gene has been a predominant method in identifying and quantifying the bacterial community in complex biological samples [10]. NGS provides a platform to sequence the cultivable or non-culturable strain using short sequence reads to study the oral microbial community. The bacterial species are classified in Operational taxonomic units (OTUs). Although there are site-based differences in microbiome, the most common sites to study the oral microbiome are saliva and supragingival plaque, especially in the context of ECC. However, there is no standard approach yet about which region of 16s rRNA has to be used and this introduces biases among studies and also makes it difficult to resolve it beyond genus level [11]. Such analysis can help in classification of disease and disease-free state and is one of the goals of human microbiome projects [12].

Metagenomic biomarkers have been studied extensively in gut microbiomes for crohn’s disease, colorectal cancer, diabetes, obesity, and inflammatory bowel disease. Given the high dimensionality of the microbiome dataset in any study, more robust models are required to identify the predictive features and to ensure the reproducibility of the analysis. One of the goals of supervised classification methods in microbiomes is to identify the predictive features for a given condition and produce a predictive model. Other than high dimensionality and high sparsity of the microbiome data, one of the key challenges in this direction is very few common species among subjects. Other methods include feature selection to reduce the expected error. Based on the ML methods popularity, a recent review suggested 4 common methods four microbiome data analysis, Random Forest, Support Vector Machines, Logistic Regression, and k-NN [13].

Meta-analysis can help in addressing the discrepancies that arise due to the technical and/or biological inconsistencies between studies [14]. Meta-analysis also enables identifying the universality of the biomarker for a condition for diagnosis or prognosis. On the other hand, it is also important to understand the factors underestimating the effects of these analyses. Different procedures for raw material handling, sequencing methods, data preprocessing methods can lead to discrepancies in datasets. Hence, it is important to address such issues and minimize these factors to the extent possible. Several studies have applied meta-analysis in various diseases like parkinson disease[15], cancer [16,17], Urolithiasis [18]. To our knowledge, no previous study combined the microbiome data for meta-analysis for ECC microbiome. Such analysis would enable us to understand the complexity of microbial profile and site-specific complexity in ECC.

In this report we collected several published studies which have published the raw data based on 16s amplicon sequencing. The studies can be divided based on the site of sample collection: plaque and saliva. We processed the raw data using the same approach wherever possible except when limited by raw data type. This combined study can help in synthesizing the dissimilar conclusion provided by individual studies.

# Methodology

## Inclusion and exclusion of studies

Studies were selected from the publications which included the disease ECC or SECC (severe-ECC) microbiome data with case-control or longitudinal analysis. Among these studies, only the studies which provided the raw sequencing data were selected. The raw data obtained from these studies can be identified of three types: (1) HOMIM microarray (Human Oral Microbe Identification Microarray), (2) 16s rRNA Amplicon sequencing, and (3) shotgun metagenomic data. The 16s rRNA sequencing could be used with different hypervariable regions or with full 16s rRNA sequencing. At this stage we worked only with 16s rRNA Amplicon sequencing data irrespective of the hypervariable region. The two other types of analysis, microarray or shotgun metagenomics, can be included in our meta-analysis in the future. The description of the datasets about the sample size, case-control number, additional metadata given about the samples, and accession number for the raw data included in the analysis is given in Table 2.

## Collection of data and raw data processing

The data was downloaded from NCBI repositories using Fasterq-dump (NCBI SRA-toolkit : <https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software>). Some of the raw sequences were obtained directly from the authors upon request. Analysis was done using the Qiime2 pipeline to obtain the OTU table at the species and genus level. For each dataset, several combinations of left and right trim were tried in DADA2. The performance of the DADA2 trimming was manually checked and the combination which retains a maximum number of reads in the samples with a low number of reads was used for final trimming. HOMD 15.22 was used as a reference database unless otherwise mentioned to classify the OTUs. The reads were extracted from the reference database using the primers mentioned under each study (Table 3). The steps of the Qiime2 pipeline are as follows: Importing, binning of sequences into representative sequences using dada2, training of classifiers using sequencing primers and HOMD reference database, assigning taxonomy to reads using the classifier (Figure Qiime flowchart).

## Preprocessing

The RA or scaled data is first visualized by PCA in Phyloseq to see the batch effects in the different datasets. The following methods were tried to correct the batch effects: Removebatch, ANCOM-BC, Combat, PLSDA, sPLSDA. Finally, sPLSDA method was chosen for further analysis. In some of the results, the imputation was applied using mbImpute methods. For normalization, the three methods were tried, Relative abundance, Scaling to a total of 107  reads in each sample, and Centered Log-Ratio (CLR).

## Analysis

Analysis was done at both genus and species levels with taxa at each level as features for all statistical analysis. The variance of each feature was checked and variance was plotted for variance by study versus variance due to the disease status provided in SIAMCAT. For differential abundance (DA) analysis, a non-parametric Wilcoxon test and DESeq2 were applied with multiple hypotheses testing corrections using the FDR with p < 0.05. To visualize the data, PCA in Phyloseq was used.

## Machine Learning modeling

We tried a number of Machine Learning methods for example Lasso, RandomForest, Xgboost, and Support Vector Machines. For machine learning, 4-fold cross-validation (CV) with 10 repeats was used everywhere. A 4-fold CV gives at least 3 samples in each dataset for small datasets which is not possible for a 5-fold CV. For the meta-analysis, Leave-One-Dataset-Out (LODO) strategy was used. In LODO analysis, all dataset, except one, was used for the training using CV and the left out dataset was then used for the testing to assess the generalizability of the model. The model performance was assessed by AUROC and AUPRC values. We also tried the method Selbal but at present selbal doesnot offer train-test validation. So Selbal was used only with CV by comparing each study individually and by combining 5 studies together. Selbal performance is also assessed by auroc values.

## Machine Learning hyperparameter tuning

The lasso model was used with ‘classif.cvglmnet’ learner from the mlr package with nlambda=100 and alpha=1 which is default under SIAMCAT. For random forest ntree range was set for 100-1000 and mtry = c(round(sqrt.mdim / 2), round(sqrt.mdim), round(sqrt.mdim \* 2)) where mtry and mdim are the number of random variables in each tree and the total number of features, respectively. In Xgboost nrounds, gamma, colsample\_bytree and min\_child\_weight were used with fixed values. For eta, max\_depth and, subsample, a range was provided. For SVM, regularization parameter C and sigma were tuned. Lasso and random forest were used from SIAMCAT package which functions as an interface to the mlr-package for these classifiers. For XGboost and SVM, mikropml package was used which uses caret package from R.

## SPLSDA anlaysis

We also used sPLSDA to generate a classfication model for the identification of ECC and healthy samples. It used PLS approach for the classification of samples. To apply sPLSDA on several studies together we used MINT approach from the R package MixOmics. We considered 2 components in each of the

# Results and discussion

## Preprocessing result

In five plaque datasets, the total number of OTUs ranged from 133 to 342 at the species level with an intersection of 96 species common across all the samples (Figure 1a). While for genus level, the taxa ranged from 59 to 101 with 50 common genera (Figure 1b ). The taxa identified through Qiime2 at genus level and species level varies a lot across all the datasets. Although it sounds reasonable to analyze the data using common OTUs, we lose a lot of OTUs which might be playing a key role in individual study. To circumvent this problem one approach is to do imputation. However, imputation methods have not been explored in the context of microbiome data analysis. Furthermore, there are no practical examples of studies applying imputation for merging data from different studies. We explored a recently published method, mbImpute, for imputation. However, applying imputation on our data still reflects batch effects in the data (Figure 2e) which also decreases the CV performance in all machine learning methods (Figure 10b).

Another limitation when doing data comparison from multiple studies is batch effects. Again, very few studies doing meta-analysis performed batch correction (Table 4). Most of the time this is done with the help of either taking relative abundance or rarifying the data to a certain number of reads. In our study, we tried a number of approaches for batch correction as mentioned in the method section. We finally selected the sPLSDA based on its performance as observed through redundancy analysis (rda in vegan package in R) and PCA plots (Figure 3).

## Differential abundance analysis

For differential abundance analysis, other than Vivianne\_2020 data other data show very few or no significant OTUs both at species and genus level when analyzed using the Wilcoxon test. In the further analysis using DESeq2, we identified that this approach identifies more OTUs as differentially abundant features (Table 1). In the future, we would like to explore the reason for less number of significant species in DA analysis in some of the the datasets. One hypothesis is the power of the statistical test. As there are more samples in the Vivianne dataset (Table 2) we see more taxa identified as DA. Another reason could be the fewer number of reads in some of the samples as the suggested number of reads is ~5000 in each sample. Though most of the samples in our study have more than the specified number of reads, however, sometimes samples with less number of reads were also included due to the small sample size in some of the datasets.

## Machine Learning

We used multiple machine learning methods for meta-analysis. We tried some commonly used ML methods in metagenomic studies which are Lasso, RandomForest, Xgboost and Support Vector Machines. The CV performance was best with Xgboost and Lasso while random forest results were very close to these two methods. Here we used the term cross-validation for the performance within the same study and cross-study validation for applying a model trained on one study and tested on another study. Machine learning classifiers work best when used with all OTUs for cross-validation. Cross-validation results were relatively better when all the OTUs, identified in Qiime2, were used instead of only common OTUs. In CV analysis studies with more samples produce better results than the studies with fewer samples. Vivianne, Agnello in plaque, and Grier in saliva samples always generated better results within their group.

Selbal method gave comparable or better classification performance even with a smaller number of taxa. But the lack of model transferability in selbal method does not make it an ideal choice in our analysis.

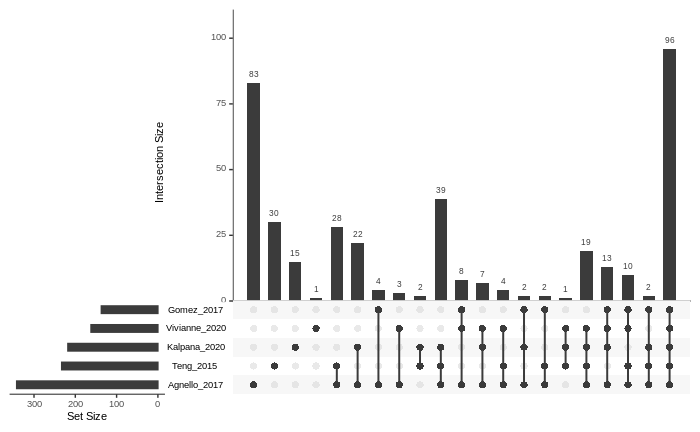
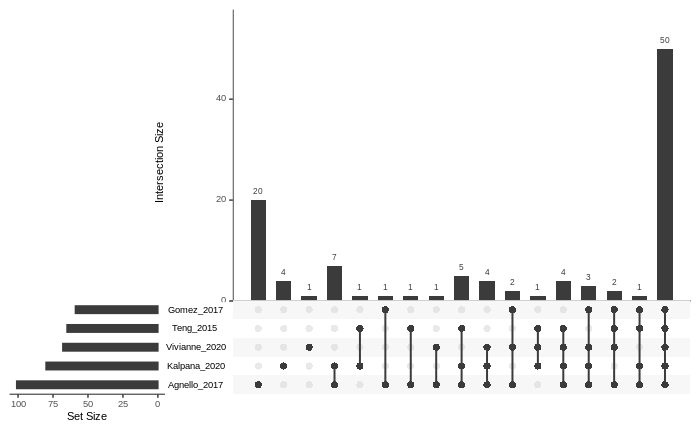
## Meta-analysis

For the meta-analysis, we wanted to see the generalizability of the models trained on multiple datasets and then test on the holdout dataset (LODO). We also examined the performance of ML models by training it on one dataset and testing on other datasets one by one (Figure 6,7,8,9,10). Results for LODO analysis are not very promising and lack consistency. Another conclusion that can be made from the graphs is that genus-level analysis provides better results than species level. This is also consistent with the batch effect results as batch effects are lower in the case of genus than species. Surprisingly, these results are better with all OTUs while in this case some of the OTUs are not consistently present in all the studies. This might be explained by the presence of some OTUs in multiple studies even if not in all the studies and the classifier picks the important features upon merging the datasets in LODO analysis. We also compared the features identified in each study for CV at the genus level for plaque samples and we observed that there were very few common features among studies and no single feature was common across all five datasets (Figure 11, 12).

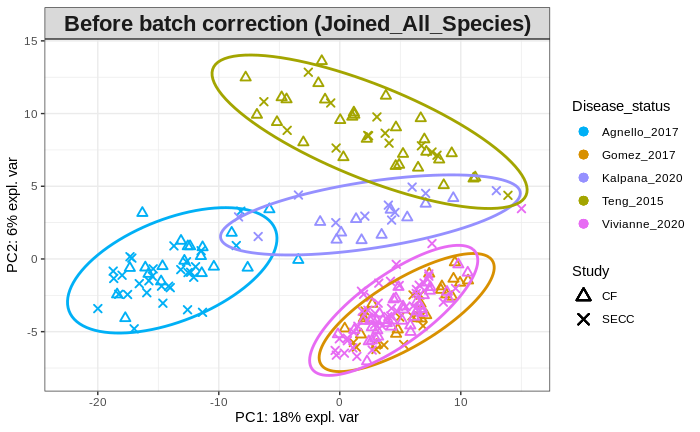
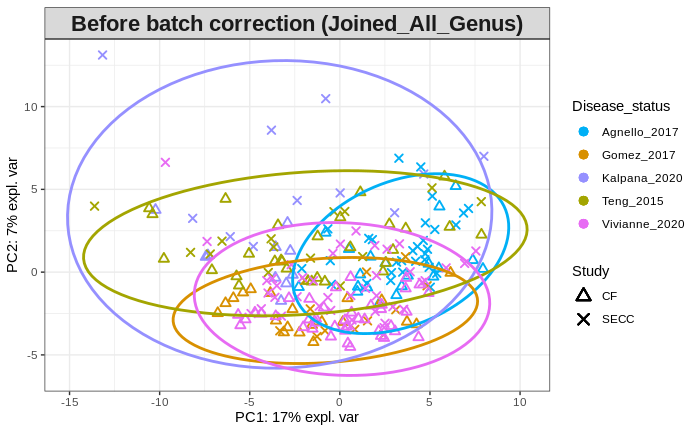
The absence of common features explains the poor performance of LODO analysis. Meta-analysis in these studies seems difficult because of very few common features identified by Machine Learning models. At this stage, the reason for the large variation in features is unknown. Some possible explanations could be as follows: the number of reads in each sample varies greatly in between samples and Each study uses a particular variable region or combination of the variable regions of 16s sequencing for the amplicon sequencing to identify the bacterial community [19]The differences in the datasets might be due to the different regions selected during the sequencing as shown in column3 of Table 2. [19].

## Meta-analysis using sPLSDA

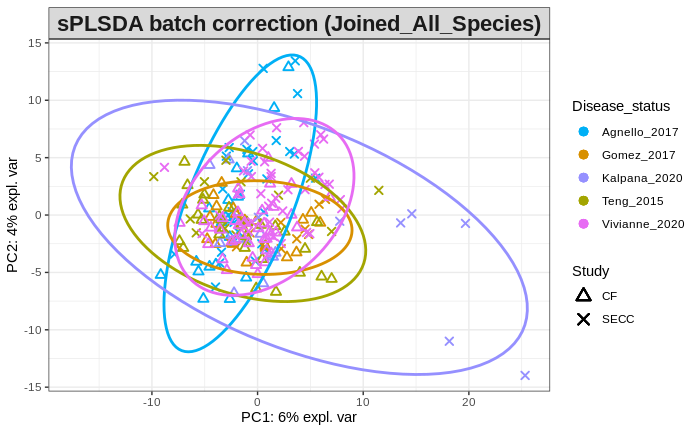
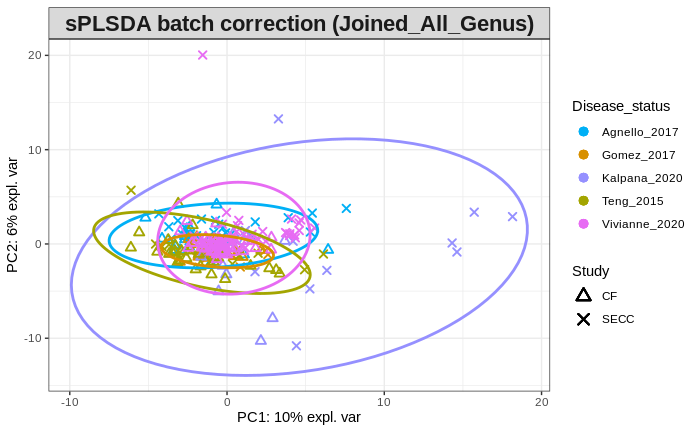
## Selbal results



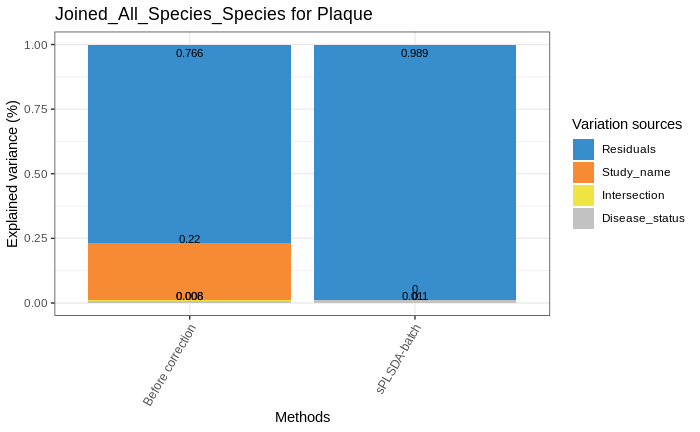
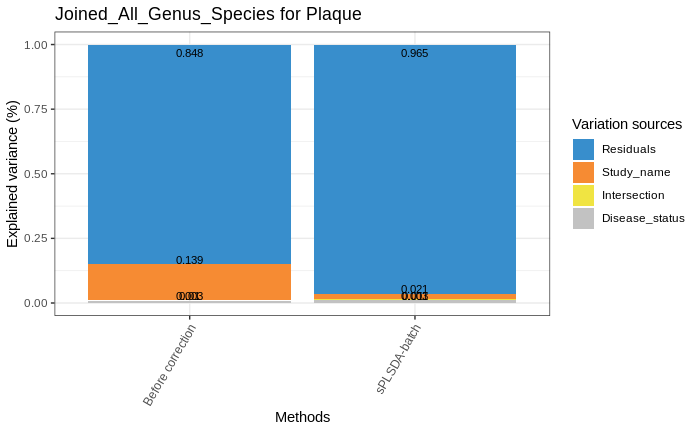
*Figure 1: (Plaque data) Shared and unique OTUs across different datasets (a) Genus level (b) Species level*



*Figure 2: (Plaque data) PCA analysis for batch effects in raw data (a) Genus level (b) Species level (Prashen commented we need to analyze at species level)*



*Figure : (Plaque data) PCA analysis for batch effects after CLR normalization and sPLSDA batch correction (a) Genus level (b) Species level (batch corrected data, based on all features with replacing missing with 0-problem using 0?)*

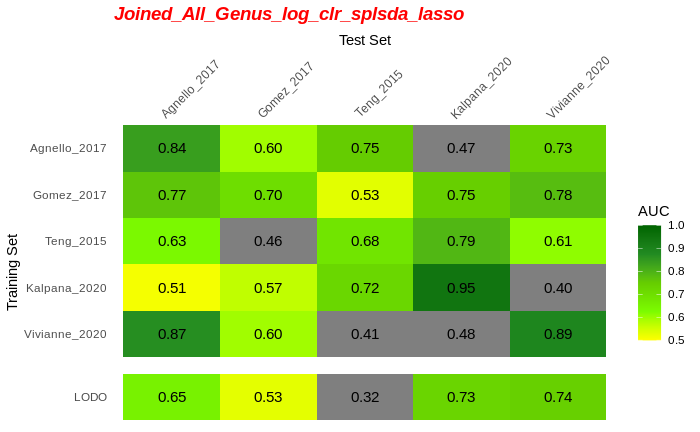


*Figure : (Plaque data) Source of variation before and after batch correction (a) Genus level (b) Species level*

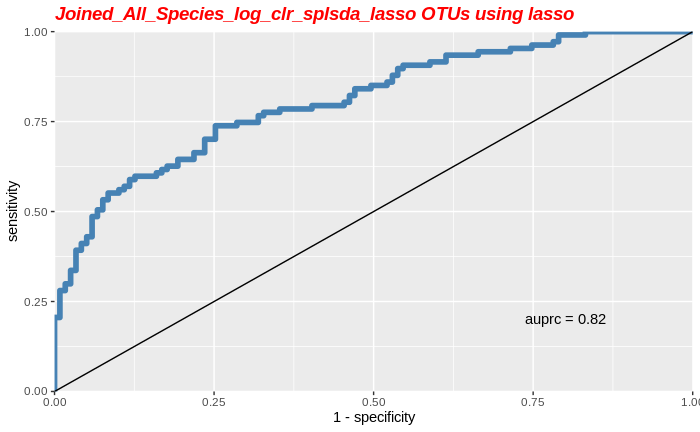
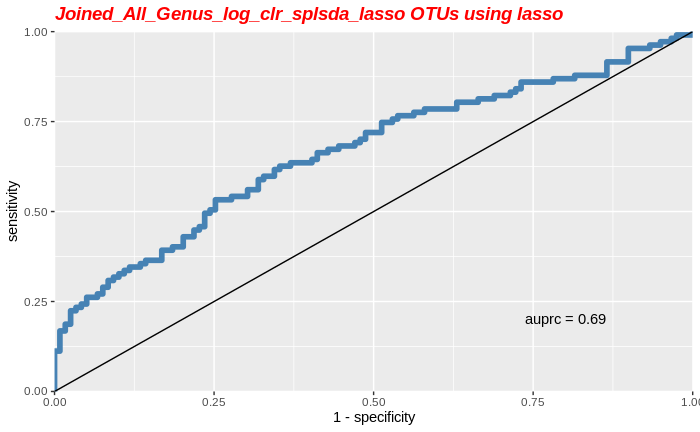
A picture containing chart

Description automatically generated

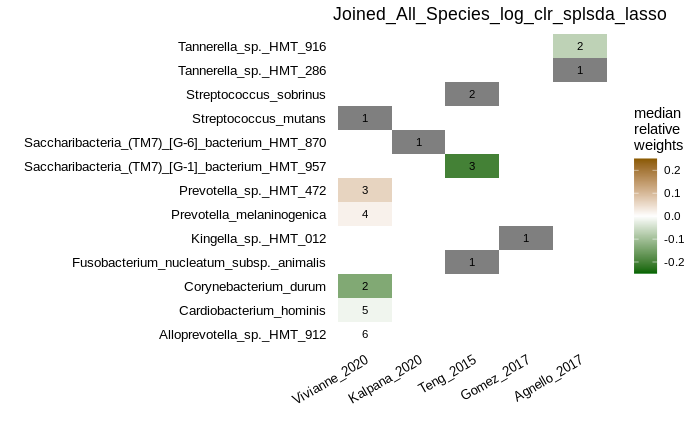
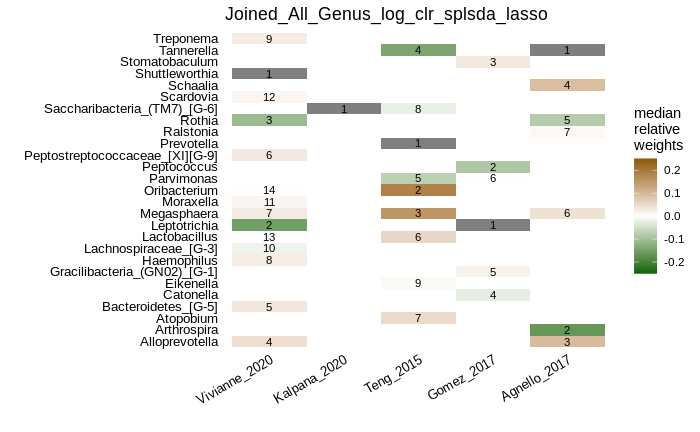
*Figure : (Plaque data) AUROC values for CV, study-to-study transfer and LODO validation using xGBoost method for classification of SECC vs CF (a) Genus level (b) Species level*



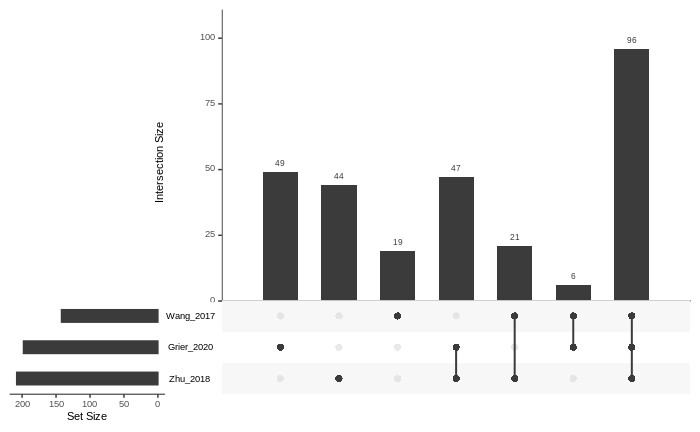
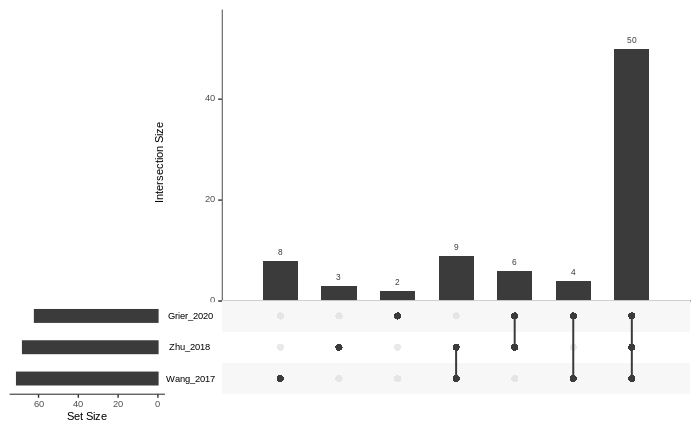
*Figure : (Plaque data) AUROC values for CV, study-to-study transfer and LODO validation using Lasso method for classification of SECC vs CF (a) Genus level (b) Species level*



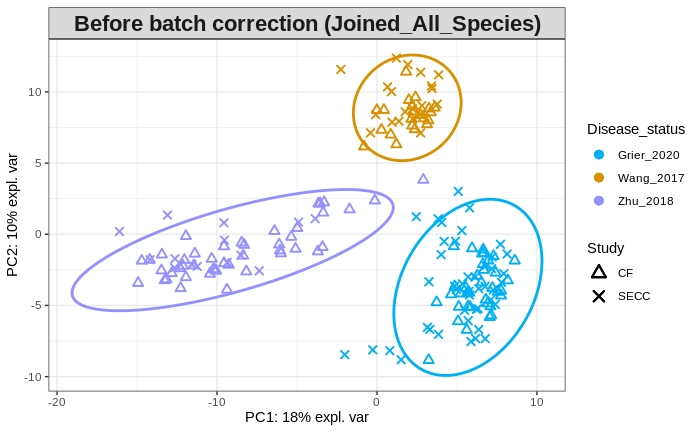
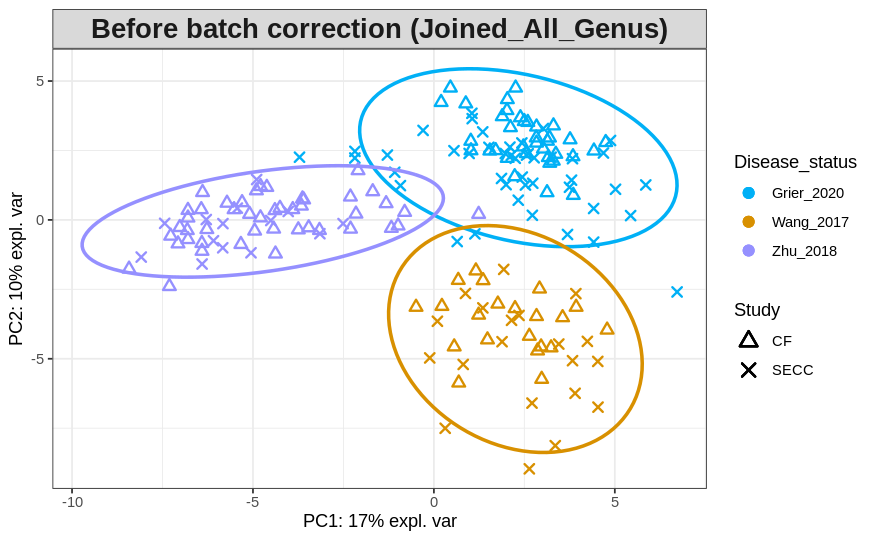
*Figure : (Plaque data) AUROC values for CV for classification of SECC vs CF by combing all studies as one dataset (a) Genus level (b) Species level*



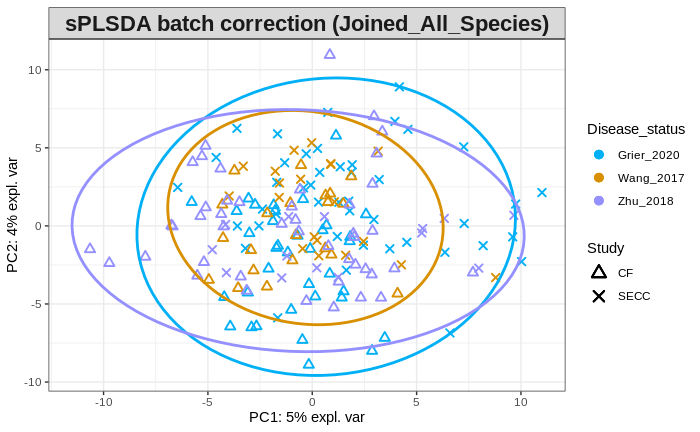
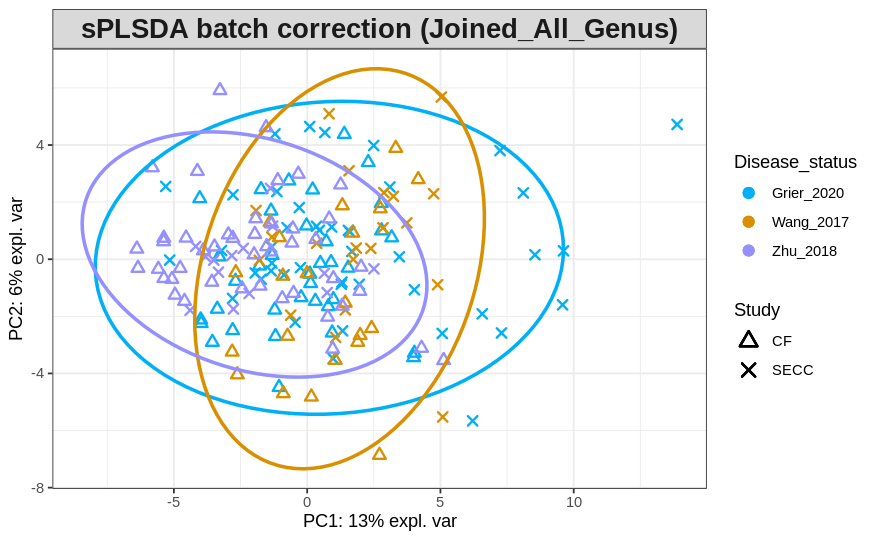
*Figure : (Plaque data) Feature importance for Lasso method performed on individual study (a) Genus level (b) Species level*



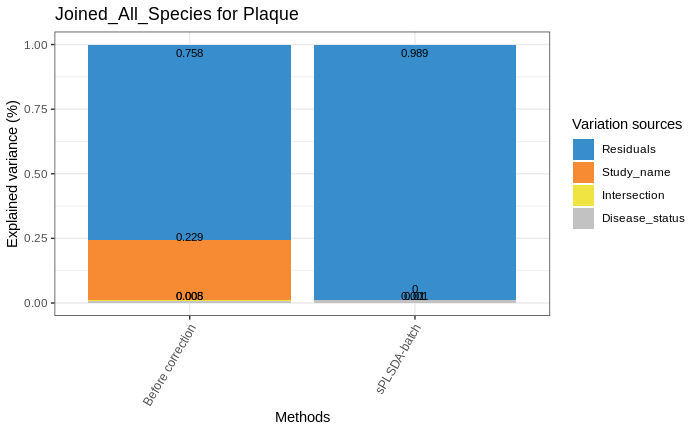
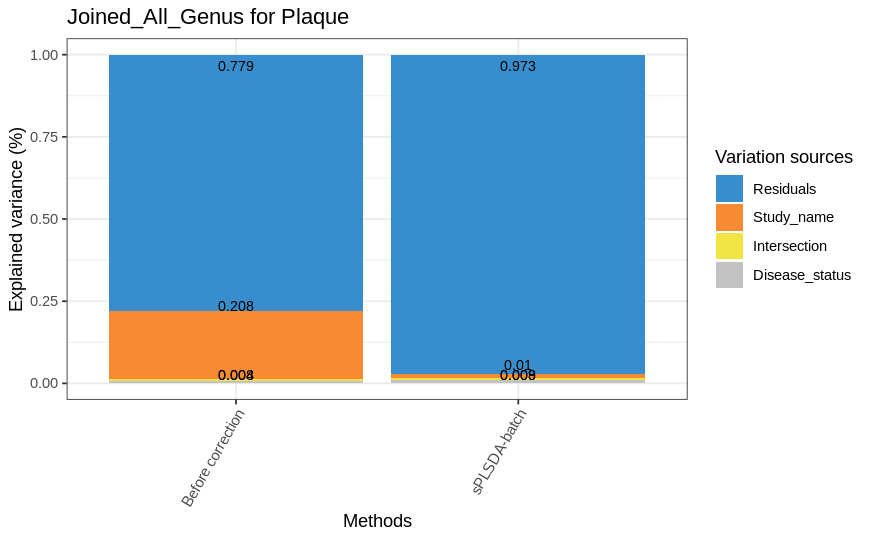
*Figure 1: (Saliva data) Shared and unique OTUs across different datasets (a) Genus level (b) Species level*



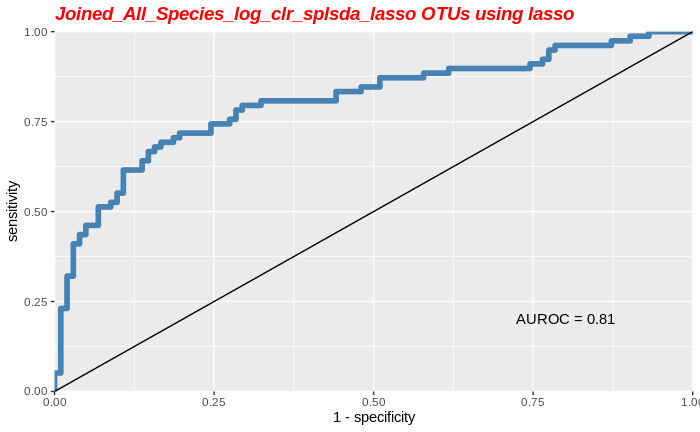
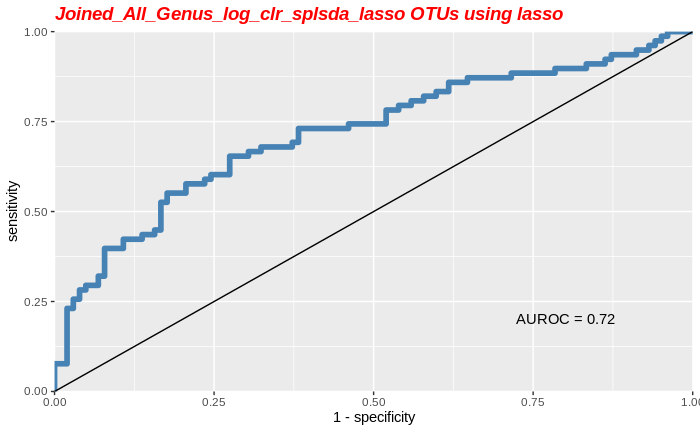
*Figure 2: (Saliva data) PCA analysis for batch effects in raw data (a) Genus level (b) Species level*



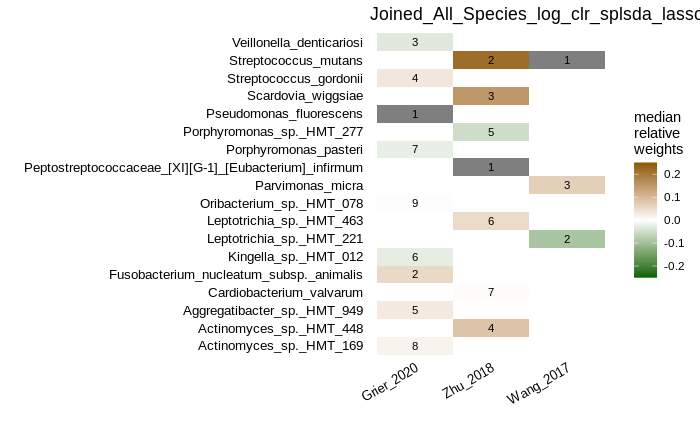
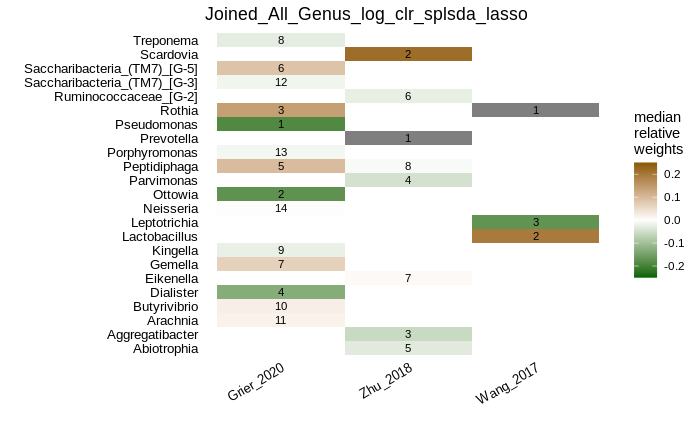
*Figure : (Saliva data) PCA analysis for batch effects after CLR normalization and sPLSDA batch correction (a) Genus level (b) Species level*



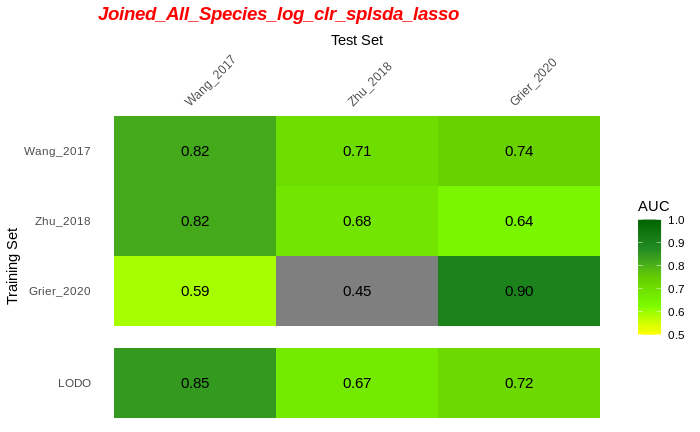
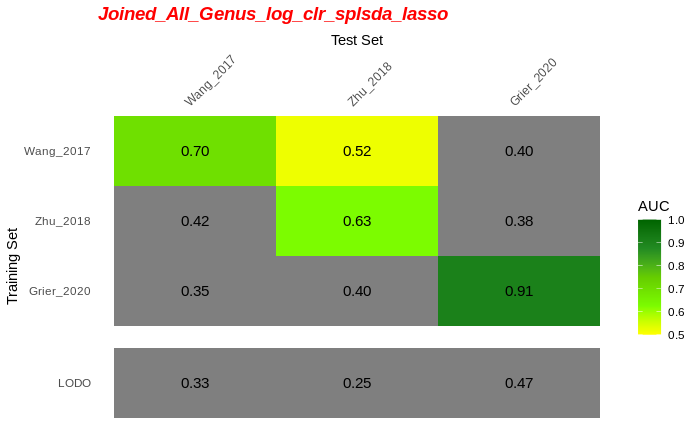
*Figure : (Saliva data) Source of variation before and after batch correction (a) Genus level (b) Species level*



*Figure : (Saliva data) AUROC values for CV for classification of SECC vs CF by combing all studies as one dataset (a) Genus level (b) Species level*



*Figure : (Saliva data) Feature importance for Lasso method performed on individual study (a) Genus level (b) Species level*



*Figure : (Saliva data) AUROC values for CV, study-to-study transfer and LODO validation using Lasso method for classification of SECC vs CF (a) Genus level (b) Species level*



*Figure : (Saliva data) AUROC values for CV, study-to-study transfer and LODO validation using xGBoost method for classification of SECC vs CF (a) Genus level (b) Species level*

##### **Table 2:** List of published ECC and Severe-ECC (SECC) studies that have provided the raw data.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **-Saliva (only 3 studies)** | | | | | | |
| **Study** | **Cases/**  **Controls** | **Raw Data Format and Accession no.** | **Metadata** | **Location** | **Current Status** | **Additional Remarks** |
| **Ma\_2015** [24] | 20 SECC  20 CF | OTU table (HOMIM 16S rRNA gene | Age, Sex, dmfs | Beijing, China | Processed and got HOMD based OTU table | This data is obtained through microarray technology. |
| **Teng\_2015** [5] | 28 SECC  17 CF (with repeated measurement) | Fastq -16S rRNA  SRP040945 and SRP040947  V1-V3 | dmfs | Guangzhou, China | Processed and got HOMD based OTU table |  |
| **Wang\_2017** [25] | 21 SECC  20 CF | Fastq-16S rRNA  SRP108162  V1–V9 | None | Hangzhou, China | Processed and got HOMD based OTU table |  |
| **Zhu\_2018** [26] | 14 ECC  45 CF | Fastq-16S rRNA  SRP162752  V3-V4 | Age, Sex,  dmfs | Beijing, China | Processed and got HOMD based OTU table |  |
| **Grier\_2020** [27] | 36 ECC  20 CF | Fastq-16S rRNA  PRJNA622300  V1-V3 | Age | NewYork, USA | Processed and got HOMD based OTU table | 8F - 534R |
| **Plaque Samples (5 studies)** | | | | | |  |
| **Ma\_2015** [24] | 20 SECC  20 CF | OTU table (HOMIM 16S rRNA gene | Age  Sex  DMFS | Beijing, China | Processed and got HOMD based OTU table | This data is obtained through microarray technology. |
| **Teng\_2015** [5] | 28 SECC  17 CF (with repeated measurement) | Fastq -16S rRNA  SRP040945 and SRP040947  V1-V3 | Age, dmfs | Guangzhou, China | Processed and got HOMD based OTU table | Dmfs value for each participant is provided. Only those samples which qualify SECC status are selected at this stage |
| **Agnello\_2017** [28] | 30 SECC  20 CF | 16S Data  V3-V4 | Age, Sex, etc | Winnipeg, Canada. | Obtained from Dr Schroth | - |
| **Gomez\_2017** [29] | 12 ECC  20 CF | Fastq-16S rRNA  PRJNA383868  V4 | Age, Sex | Australia | Processed and got HOMD based OTU table | Only the term CARIES is used in the paper and dmfs value is not given |
| **Vivianne\_2020** [30] | 40 SECC  40 CF | Fastq-16S rRNA  V4 | Age, Sex, etc | Winnipeg, Canada | Processed and got HOMD based OTU table | Only SECC samples were collected.  Dmfs values not provided |
| **Kalpana\_2020** [31] | 20 SECC  20 Recurrent ECC  15 CF | 16s rRNA raw reads  PRJNA454811  V3-V4 | None | Tiruchengode, India | Processed and got HOMD based OTU table | Raw files are provided for only 31 participants |
| **Buccal Swab** | | | | | |  |
| **Vivianne\_2020** [32] | 40 SECC  40 CF | Fastq-16S rRNA  V4 | Age, Sex etc | Winnipeg, Canada | Processed and got HOMD based OTU table |  |

##### 

#### References

1. Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. PLoS Biol. 2016;14: e1002533.

2. R. Rizkallah M, Gamal-Eldin S, Saad R, K. Aziz R. The PharmacoMicrobiomics Portal: A Database for Drug-Microbiome Interactions. Curr Pharmacogenomics Person Med. 2012;10: 195–203.

3. Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner ACR, Yu W-H, et al. The human oral microbiome. J Bacteriol. 2010;192: 5002–5017.

4. Escapa IF, Chen T, Huang Y, Gajare P, Dewhirst FE, Lemon KP. New Insights into Human Nostril Microbiome from the Expanded Human Oral Microbiome Database (eHOMD): a Resource for the Microbiome of the Human Aerodigestive Tract. mSystems. 2018;3. doi:10.1128/mSystems.00187-18

5. Teng F, Yang F, Huang S, Bo C, Xu ZZ, Amir A, et al. Prediction of Early Childhood Caries via Spatial-Temporal Variations of Oral Microbiota. Cell Host Microbe. 2015;18: 296–306.

6. Casamassimo PS, Thikkurissy S, Edelstein BL, Maiorini E. Beyond the dmft: the human and economic cost of early childhood caries. J Am Dent Assoc. 2009;140: 650–657.

7. Li Y, Ge Y, Saxena D, Caufield PW. Genetic profiling of the oral microbiota associated with severe early-childhood caries. J Clin Microbiol. 2007;45: 81–87.

8. Tanner ACR, Kent RL Jr, Holgerson PL, Hughes CV, Loo CY, Kanasi E, et al. Microbiota of severe early childhood caries before and after therapy. J Dent Res. 2011;90: 1298–1305.

9. Richards VP, Alvarez AJ, Luce AR, Bedenbaugh M, Mitchell ML, Burne RA, et al. Microbiomes of Site-Specific Dental Plaques from Children with Different Caries Status. Infect Immun. 2017;85. doi:10.1128/IAI.00106-17

10. Lozupone CA, Stombaugh J, Gonzalez A, Ackermann G, Wendel D, Vázquez-Baeza Y, et al. Meta-analyses of studies of the human microbiota. Genome Res. 2013;23: 1704–1714.

11. Baker JL, Morton JT, Dinis M, Alvarez R, Tran NC, Knight R, et al. Deep metagenomics examines the oral microbiome during dental caries, revealing novel taxa and co-occurrences with host molecules. Genome Res. 2021;31: 64–74.

12. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486: 207–214.

13. Marcos-Zambrano LJ, Karaduzovic-Hadziabdic K, Loncar Turukalo T, Przymus P, Trajkovik V, Aasmets O, et al. Applications of Machine Learning in Human Microbiome Studies: A Review on Feature Selection, Biomarker Identification, Disease Prediction and Treatment. Front Microbiol. 2021;12: 634511.

14. Bisanz JE, Upadhyay V, Turnbaugh JA, Ly K, Turnbaugh PJ. Meta-Analysis Reveals Reproducible Gut Microbiome Alterations in Response to a High-Fat Diet. Cell Host Microbe. 2019;26: 265-272.e4.

15. Romano S, Savva GM, Bedarf JR, Charles IG, Hildebrand F, Narbad A. Meta-analysis of the gut microbiome of Parkinson’s disease patients suggests alterations linked to intestinal inflammation. bioRxiv. medRxiv; 2020. doi:10.1101/2020.08.10.20171397

16. Limeta A, Ji B, Levin M, Gatto F, Nielsen J. Meta-analysis of the gut microbiota in predicting response to cancer immunotherapy in metastatic melanoma. JCI Insight. 2020;5. doi:10.1172/jci.insight.140940

17. Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, Zolfo M, et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med. 2019;25: 667–678.

18. Kachroo N, Lange D, Penniston KL, Stern J, Tasian G, Bajic P, et al. Meta-analysis of clinical microbiome studies in urolithiasis reveal age, stone composition, and study location as the predominant factors in urolithiasis-associated microbiome composition. MBio. 2021;12: e0200721.

19. Abellan-Schneyder I, Matchado MS, Reitmeier S, Sommer A, Sewald Z, Baumbach J, et al. Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing. mSphere. 2021;6. doi:10.1128/mSphere.01202-20

20. Montassier E, Al-Ghalith GA, Hillmann B, Viskocil K, Kabage AJ, McKinlay CE, et al. CLOUD: a non-parametric detection test for microbiome outliers. Microbiome. 2018;6: 137.

21. Leigh RJ, Murphy RA, Walsh F. uniForest: an unsupervised machine learning technique to detect outliers and restrict variance in microbiome studies. bioRxiv. 2021. p. 2021.05.17.444491. doi:10.1101/2021.05.17.444491

22. Goren E, Wang C, He Z, Sheflin AM, Chiniquy D, Prenni JE, et al. Feature selection and causal analysis for microbiome studies in the presence of confounding using standardization. BMC Bioinformatics. 2021;22: 362.

23. Zhou Y-H, Gallins P. A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction. Front Genet. 2019;10: 579.

24. Ma C, Chen F, Zhang Y, Sun X, Tong P, Si Y, et al. Comparison of oral microbial profiles between children with severe early childhood caries and caries-free children using the human oral microbe identification microarray. PLoS One. 2015;10: e0122075.

25. Wang Y, Zhang J, Chen X, Jiang W, Wang S, Xu L, et al. Profiling of Oral Microbiota in Early Childhood Caries Using Single-Molecule Real-Time Sequencing. Front Microbiol. 2017;8: 2244.

26. Zhu C, Yuan C, Ao S, Shi X, Chen F, Sun X, et al. The Predictive Potentiality of Salivary Microbiome for the Recurrence of Early Childhood Caries. Front Cell Infect Microbiol. 2018;8: 423.

27. Grier A, Myers JA, O’Connor TG, Quivey RG, Gill SR, Kopycka-Kedzierawski DT. Oral Microbiota Composition Predicts Early Childhood Caries Onset. J Dent Res. 2020; 22034520979926.

28. Agnello M, Marques J, Cen L, Mittermuller B, Huang A, Chaichanasakul Tran N, et al. Microbiome Associated with Severe Caries in Canadian First Nations Children. J Dent Res. 2017;96: 1378–1385.

29. Gomez A, Espinoza JL, Harkins DM, Leong P, Saffery R, Bockmann M, et al. Host Genetic Control of the Oral Microbiome in Health and Disease. Cell Host Microbe. 2017;22: 269-278.e3.

30. de Jesus VC, Shikder R, Oryniak D, Mann K, Alamri A, Mittermuller B, et al. Sex-Based Diverse Plaque Microbiota in Children with Severe Caries. J Dent Res. 2020;99: 703–712.

31. Kalpana B, Prabhu P, Bhat AH, Senthilkumar A, Arun RP, Asokan S, et al. Bacterial diversity and functional analysis of severe early childhood caries and recurrence in India. Sci Rep. 2020;10: 21248.

32. de Jesus VC, Khan MW, Mittermuller B-A, Duan K, Hu P, Schroth RJ, et al. Characterization of Supragingival Plaque and Oral Swab Microbiomes in Children With Severe Early Childhood Caries. Front Microbiol. 2021;12: 1517.

33. Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. Nat Commun. 2017;8: 1784.

34. Rocca JD, Simonin M, Blaszczak JR, Ernakovich JG, Gibbons SM, Midani FS, et al. The Microbiome Stress Project: Toward a Global Meta-Analysis of Environmental Stressors and Their Effects on Microbial Communities. Front Microbiol. 2018;9: 3272.

35. Greathouse KL, White JR, Padgett RN, Perrotta BG, Jenkins GD, Chia N, et al. Gut microbiome meta-analysis reveals dysbiosis is independent of body mass index in predicting risk of obesity-associated CRC. BMJ Open Gastroenterol. 2019;6: e000247.

36. Kosti I, Lyalina S, Pollard KS, Butte AJ, Sirota M. Meta-Analysis of Vaginal Microbiome Data Provides New Insights Into Preterm Birth. Front Microbiol. 2020;11: 476.

37. Malacrinò A, Sadowski VA, Martin TK, Cavichiolli de Oliveira N, Brackett IJ, Feller JD, et al. Biological invasions alter environmental microbiomes: A meta-analysis. PLoS One. 2020;15: e0240996.

38. Chica Cardenas LA, Clavijo V, Vives M, Reyes A. Bacterial meta-analysis of chicken cecal microbiota. PeerJ. 2021;9: e10571.

39. Wirbel J, Zych K, Essex M, Karcher N, Kartal E, Salazar G, et al. Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. Genome Biol. 2021;22: 93.

# Supplementary

##### **Table 3**: Dataset specific processing

|  |  |
| --- | --- |
| **-Saliva** | |
| **Study** | **Data processing** |
| **Ma\_2015**  [24] | * Download the supplementary data <https://doi.org/10.1371/journal.pone.0122075.s001> for OTUs * HOT identify was replaced as HMT identity * HMT numbers were matched with Homd 15.22 HMT-id as provided in HOMD database * An updated table with new taxonomy names was obtained. |
| **Teng\_2015**  [5] | * Download the two multiplexed files   + SRP040945 and SRP040947 * Demultiplex using unique headers in the fastq files and obtain 10 files - Using demuxbyname.sh option from BBTools * Further demultiplex these files using the barcodes file provided by the authors through email- using fastx\_barcode\_splitter.pl from FASTX-Toolkit * Remove barcodes from the demultiplexed files - using fastx\_trimmer * Select samples with dmfs values >5 (SECC) and dmfs = 0 (caries-free) using the table given in the paper * For DADA2 (denoise-pyro) trim-left = 20 and trunc-len = 490 and max-ee 5 was used for final analysis * The following primers were used to extract reference sequences in Qiime2   + f-primer GAGTTTGATCCTGGCTCAG   + r-primer TACCGCGGCTGCTGGCAC   + This is slightly different than the sequences mentioned in the paper. The forward primer in the paper starts at position 5 but our reference starts at position 9. So, necessary adjustments were made. |
| **Wang\_2017**  [25] | * Fastq reads were downloaded from NCBI SRP108162 * Reads were trimmed to extract V4 region because the real quality at the end was not good. to match it with Vivianne data was obtained through V4 seqencing. * Cutadapt was using with the following option   + F515 = "GTGCCAGCMGCCGCGGTAA"   + R806 = "GGACTACHVGGGTWTCTAAT"   + --trimmed-only --rc --match-read-wildcards -j 4 -M 265 * Used DADA2 denoise-single with no left and right truncation |
| **Zhu\_2018**  [26] | * Download the sequence SRP162752 * Primers used in sequencing   + f-primer GTACTCCTACGGGAGGCAGCA   + r-primer GTGGACTACHVGGGTWTCTAAT * Remove primer sequences from the reads to improve detection in DADA2 using cutadapt with the following option   + --match-read-wildcards -j 4 * The options for DADA2 denoise-paired   + --p-trunc-len-f 240   + --p-trunc-len-r 205 |
| **Grier\_2020**  [27] | * Download the sequence from PRJNA622300 * Primers   + f-primer GAGTTTGATCCTGGCTCAG   + r-primer ATTACCGCGGCTGCTGG * Cutadapt to remove the primers   + --match-read-wildcards -j 4 * The options for DADA2 denoise-paired   + --p-trunc-len-f 290   + --p-trunc-len-r 270 |
| **Plaque Samples** | |
| **Ma\_2015**  [24] | * Download the supplementary data <https://doi.org/10.1371/journal.pone.0122075.s001> for OTUs * HOT identify was replaced as HMT identity * HMT numbers were matched with HOMD 15.22 HMT-id as provided in the HOMD database * An updated table with new taxonomy names was obtained. |
| **Teng\_2015**  [5] | * Download the two multiplexed files   + SRP040945 and SRP040947 * Demultiplex using unique headers in the fastq files and obtain 10 files - Using demuxbyname.sh option from BBTools * Further demultiplex these files using the barcodes file provided by the authors through email- using fastx\_barcode\_splitter.pl from FASTX-Toolkit * Remove barcodes from the demultiplexed files - using fastx\_trimmer * Select samples with dmfs values >5 (SECC) and dmfs = 0 (caries-free) using the table given in the paper * For DADA2 (denoise-pyro) trim-left = 20 and trunc-len = 490 and max-ee 5 was used for final analysis * The following primers were used to extract reference sequences in Qiime2   + f-primer GAGTTTGATCCTGGCTCAG   + r-primer TACCGCGGCTGCTGGCAC   + This is slightly different than the sequences mentioned in the paper because in the paper, it starts at position 5 but our reference starts at position 9. So, necessary adjustments were made. |
| **Agnello\_2017**  [28] | * The fasta files (NOT fastq) were provided by Dr Schroth * Primer sequence   + f-primer CCTACGGGNGGCWGCAG   + r-primer GACTACHVGGGTATCTAATCC * Because DADA2 works only of fastq files so in this case, to generate the representative sequences from the reads vsearch option was used in qiime2 with the following option   + dereplicate-sequences   + cluster-features-de-novo with perc-identity 0.99 |
| **Gomez\_2017**  [29] | * The data were downloaded from PRJNA383868 * Primers   + f-primer GTGCCAGCMGCCGCGGTAA   + r-primer GGACTACHVGGGTWTCTAAT * DADA2 denoise-paired was used with the following options   + --p-trunc-len-f 200   + --p-trunc-len-r 150   + --p-trim-left-f 15   + --p-trim-left-r 15 |
| **Vivianne\_2020**  [30] | * Raw Fastq sequences were provided by Vivianne * Primers used for these sequences were   + f-primer GTGCCAGCMGCCGCGGTAA   + r-primer GGACTACHVGGGTWTCTAAT * Options for DADA2 denoise-paired (Values suggested by Vivianne as used in [30] paper )   + --p-trunc-len-f 210   + --p-trunc-len-r 210   + --p-trim-left-f 15   + --p-trim-left-r 15 |
| **Kalpana\_2020**  [31] | * The sequence were downloaded from PRJNA454811 * Primers   + f-primer CCTACGGGNBGCASCAG   + r-primer GACTACNVGGGTATCTAATCC * DADA2 denoise-paired was used with the following options   + --p-trunc-len-f 250   + --p-trunc-len-r 230   + --p-trim-left-f 5   + --p-trim-left-r 5 |
| **Buccal Swab** | |
| **Vivianne\_2020** [32] | * Raw Fastq sequences were provided by Vivianne * Primers used for these sequences were   + f-primer GTGCCAGCMGCCGCGGTAA   + r-primer GGACTACHVGGGTWTCTAAT * Options for DADA2 denoise-paired (Values suggested by Vivianne as used in [30] paper )   + --p-trunc-len-f 210   + --p-trunc-len-r 210   + --p-trim-left-f 15   + --p-trim-left-r 15 |

|  |
| --- |
| **Table 4: Meta-analysis in microbiome tools and results** |
| This paper by Lozupone et al. 2013 [10] also used many published metagenomic studies with different hypervariable regions to show that the differences in the microbiome are largely due to the site rather than studies. This is also a reason for us to select several studies irrespective of the hypervariable regions used for the 16s sequencing.  But in our analysis, we found a higher variance due to the study rather than ECC status. [10] paper only did PCA analysis. There is no such published study for the oral microbiome. So at this stage, its inconclusive that the variations in our analysis arise due to the microbiome differences in the population or these are batch effects. |
| Duvallet et al. 2017 [33] combined 28 published studies to create the MicrobiomeHD database. Performed a cross-disease meta-analysis and concluded “ many associations found in case–control studies are likely not disease-specific but rather part of a non-specific, shared response to health and disease”. To assess the generalizability of the results they performed leave-one-dataset-out and leave-one-disease-out cross-validation using Random Forest classifier and suggested that tuning did not alter the results significantly.  On the other hand, to perform the preprocessing of the raw data in 9 studies Rocca et al. 2018 [34] selected studies with the common hypervariable region of 16s sequencing (V4) and verified the data within the studies. |
| Bisanz et al. 2019 [14] study performed meta-analysis combining studies with raw data from various technologies (fig 1, [14] ) and used normlization with rarification and CLR transformation within the study. DA was calculated by Welch’s t test. For randomforest classification, Low power studies were excluded from plotting in the AUROC heatmap (Figure 4F), due to insufficient samples. Their result shows that the trained models are generalizable across studies. |
| Greathouse et al. 2019 [35] also combines data from v3, v4 and wgs sequencing. They discussed that their model did not identify any universal BMI-associated microbial biomarkers of CRC. They did not mention how the read counts were adjusted across studies but it appears that they worked with a relative abundance and no other adjustments. |
| Kosti et al. 2020 [36] took raw samples from five studies and performed batach correction using Combat on log-transformed values. In subsequent steps, bottom 30% otus were filtered out to fix the convergence problem. They showed that it does not affect the type I error and helps in removing OTUs generated through biases and misidentification. |
| Limeta et al. 2020 [16] worked on meta-anlaysis of 4 studies with a total of 130 samples, using metagenomic sequences. For normalization, log normalization was performed on relative abundance. For classification, RF was used only on 17 differentially abundant features with 100,000 trees. Performance of the classifier on an independent dataset was 0.625(95% CI: 0.348–0.899)  All codes are available at: <https://github.com/angelolimeta/Gut-microbiome-immunotherapy> . |
| Malacrinò et al. 2020 [37] 5 studies 335 samples used V4 region in all the studies. 61776.92 reads per sample (That is quite high, only Vivianne and Agnello data have a comparable number of reads). Read counts were normalized using DESeq2. Only performed PERMANOVA for samples differences between cases and control. |
| Romano et al. 2020 [15] looking for parkinsons disease(PD) marker in gut microbiome.10 studies with 1211 samples.Worked on functional group abundance rather than taxonomic abundance. Used different normalization approaches for different DA test: (1)Total sum scaling (TSS) and non-parametric tests (DA using a two-sided WMW test), (2) Variance stabilizing transformation (VST) and DESeq2 analyses, (3) Centered log-ratios (CLR) and ANCOM. No ML method was used. |
| Chica Cardenas et al. 2021 [38] used meta-analysis to define core microbiome (Genus present in 80% of the samples). A total of 9 studies with 324 sample. Three different regions (V3, V4, V3V4). Qiime analysis for same hypervariable region was performed together. After merging the ASVs, 109 genera and 11 phyla were identified. All further analyses were done within Qiime2. Number of genus in core microbiome are very few (1-5) |
| Kachroo et al. 2021 [18] 6 studies 201 patient + 136 control samples. Normalized with a DESeq2 normalization protocol which corrected for sequencing depth and composition bias across samples. Only alpha diversity, beta diversity and DA analysis was done. Used R package metamicrobiomeR. |
| Wirbel et al. 2021 [39] Most of the data was obtained from the ‘curatedMetagenomicsData R package’. Only RA and no batch correction. Extensive use of machine learning with LODO validation to identify universal markers. |

##### 

#### Future work ideas

1. Can selbal method be modified to be used as train-test type validation.
2. In sPLS-DA method, use error rate instead of auroc as given in this tutorial: <http://mixomics.org/mixmint/mint-stem-cells-case-study/>
3. When taking common OTUs only remove the OTUs which are present only in one dataset. i.e unique to a dataset. This way we can retain the OTUs which are not present in all the datasets but present in more than one dataset
4. Find the feature importance through the Selbal method. It uses an approach of using balances for machine learning.
5. Removing outlier samples in each dataset for better predictive modeling. Some of the methods are CLOUD [20], uniForest [21], and [Outlier Detection with Isolation Forest | by Eryk Lewinson](https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e).
6. Removing features showing high variance with study and not by Disease status.
7. Perform DA with different normalization and also look for a robust DA method for microbiome data.
8. Define core microbiome in each study for healthy vs ECC and compare between 5 studies. Try different cut off to define core microbiome eg 80%, 70% etc.
9. Functional analysis using PICRUSt or Tax4fun. As we see that we find different taxa in different studies so we would like to compare if the functional profiles have something common in those different taxa. The hypothesis is that the functional profiles of different taxa can be similar.
10. As Agnello\_2017 (plaque) data was collected from the first nation population, can this be tested on first nation samples in Vivianne\_2020?
11. Enterotyping analysis <https://enterotype.embl.de/enterotypes.html>
12. In plaque samples, what if we remove the samples which provide low CV performance and combine the samples with high CV values e.g. Vivianne, Agnello, Kalpana. And later, perform the feature selection.
13. Should we limit the selection of data based on common hypervariable regions, like v4, and exclude studies that do not have V4 region
14. Increasing the number of repeats in cross-validation because the data size is small and CV results vary a lot with every repeat. Can auroc range be calculated?
15. DA analysis through LEfSe approach using LEfSeR method in R.
16. Applying feature selection strategy as mentioned in [22] [23]
17. Try random Forest with feature selection