**Proposed Approach/Methodology**

Following the proposed solution of 16S sequencing, we will execute the analyses using the methodology described below:

### **Amplicon generation and library preparation for 16S sequencing and generating raw fastq data**

For 16S sequencing, we will start with PCR amplification. Amplicons will be generated using the primers for V3-V4 regions of bacterial 16S rRNA gene. These primers are designed to be universal for bacteria, allowing them to capture a wide range of bacterial taxa. This is crucial for skin microbiome studies, as the skin hosts a diverse bacterial community, including common genera such as *Staphylococcus*, *Corynebacterium*, and *Propionibacterium*. Additionally, the V3-V4 region provides sufficient taxonomic resolution to differentiate between closely related bacterial species and strains. This is important for skin microbiome analysis, where slight differences in microbial composition can be associated with different skin conditions, health states, or body sites.

For V3-V4 amplification, all PCR reactions will be performed with 15 μL of Q5® High-Fidelity 2 Master Mix(New England Biolabs, USA), 0.2 μM of forward and reverse primers, and around 10 ng template DNA. For thermal cycling, initial denaturation at 98 °C for 30s will be followed by 32 cycles of denaturation at 98 °C for 10 s, annealing at 55 °C for 30 s, and elongation at 72 °C for 20 s. Final elongation will be carried out for 2 min at 72 °C. NEBNext® Ultra™ II DNA Library Prep Kit for Illumina®(New England Biolabs, USA) will be used as per the manufacturer’s protocol for sequencing library preparation, and index codes were appended. The quality of the library will be evaluated by employing the Qubit@2.0 Fluorometer (Thermo Scientific) and the Agilent Bioanalyzer 2100 system. Finally, the 16S libraries will be sequenced on an Illumina NextSeq2000 platform at CHRF, resulting in 300-bp paired-end reads. All sequence data will be saved in fastq data format.

### **Data pre-processing and quality control**

### The paired-end sequences will be converted to the QIIME2 format for use using the QIIME2 platform for data preprocessing, quality control, taxonomic assignment, differential abundance identification, and functional analysis [1,2]. More precisely, the DADA2 plug-in in QIIME2 will be used to do the data pre-processing of paired-end sequences [3]. DADA2 will combine de-noised paired-end reads, eliminate duplicates and singletons, filter noisy reads, correct errors in marginal sequences, and de-replicate the filtered reads. Variants in the amplicon sequence will be the features generated by DADA2.

### **Taxonomic assignment**

We will apply a pre-trained classifier based on the Naive Bayes machine-learning model for taxonomy assignment. Training this model will be done on Greengenes 13\_8 data with 99% sequence similarity. After that, this classifier will be used to assign amplicon sequence variations to taxonomies.

### **Diversity analysis**

Several diversity metrics in QIIME2 will require a rooted phylogenetic tree generated from the amplicon sequence variants of the sampled data . A reference-based fragment insertion method, using the q2-fragment-insertion tool, will be applied to construct the rooted tree for this purpose. Greengenes 13\_8 data will be used as a reference database in the q2-fragment-insertion tool. The sequencing depth of the samples will be set to 3525 to observe the richness. This phenomenon will be checked with the alpha rarefaction curve generated by the q2-diversity tool. The microbiome within and between samples will be calculated by the core-metric-phylogenetic method of the q2-diversity tool. This method will compute several alpha (Observed features, Shannon diversity, Faith’s phylogenetic diversity, Pielou evenness) and beta (Jaccard distance, Bray–Curtis distance, unweighted UniFrac distance, and weighted UniFrac distance) diversity metrics altogether. Based on each beta diversity metric, this command will also perform principal coordinates analysis (PCoA). To visualize the PCoA plots for every beta diversity metric, the EMPeror visualization tool will be utilized to generate the figures. Several statistical tests will be conducted during diversity analysis such as Kruskal–Wallis H test, two-way ANOVA, paired t-test, and PERMANOVA test. The *boxplot()* function in R will be used to draw boxplots based on alpha diversity values. This function will follow the 1.5 IQR method for detecting outliers which will be placed above and below the whiskers on the boxplots.

### **Differential abundance test**

To classify the features that will be differentially abundant across various sample groups, the analysis of the composition of microbiomes (ANCOM) method will be applied by the q2-composition tool [2]. This statistical framework will be deployed at the genus level. For linear discriminant analysis by the LefSe [4], the feature table will be collapsed at the genus level. This tool will first perform the non-parametric Kruskal-Wallis (KW) sum-rank test to identify the features which have significant differential abundance across different metadata categories. Finally, LEfSe will apply linear discriminant analysis to compute the effect size of each differentially abundant feature and plot the linear discriminant analysis score in the log10 scale. Results of both ANCOM and LEfSe analysis were conducted for all samples.

Functional analysis

BURRITO is an interactive visualization web server and will be utilized to explore the taxa-function relationship within the samples of the study [5]. To acquire gene contents and functional annotations, this tool will adopt the PICRUSt and KEGG Orthology databases, respectively. At first, features with a sample size will be filtered from the original feature table to remove very low abundant taxa. The q2-search tool will be employed for closed-reference clustering of retrieved features at 97% identity based on Greengenes 97% OTU IDs as reference. Thus, the acquired OTU table will then be converted to the appropriate table format for BURRITO input.

**Reference:**

1. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 2019;37:852–7.

2. Lin H, Peddada SD. Analysis of compositions of microbiomes with bias correction. Nat Commun. 2020;11:3514.

3. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13:581–3.

4. Xu G, Xing W, Li T, Xue M, Ma Z, Jiang N, et al. Comparative study on the effects of different feeding habits and diets on intestinal microbiota in Acipenser baeri Brandt and Huso huso. BMC Microbiol. 2019;19:297.

5. McNally CP, Eng A, Noecker C, Gagne-Maynard WC, Borenstein E. BURRITO: An Interactive Multi-Omic Tool for Visualizing Taxa–Function Relationships in Microbiome Data. Front Microbiol [Internet]. 2018 [cited 2024 May 13];9. Available from: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2018.00365/full