2020

Uncovering Host-microbiome Interactions in Global Systems with Collaborative Programming: A Novel Approach Integrating Social and Data Sciences [version 1; peer review: awaiting peer review]

Jenna Oberstaller
University of South Florida, jobersta@usf.edu

Swamy Rakesh Adapa
University of South Florida, swamyrakesh@usf.edu

Guy Dayhoff II
University of South Florida, gdayhoff@usf.edu

Justin Gibbons
University of South Florida, jgibbons1@usf.edu

Gregory S. Herbert
University of South Florida, gherbert@usf.edu

Scholar Commons Citation

Oberstaller, Jenna; Adapa, Swamy Rakesh; Dayhoff, Guy II; Gibbons, Justin; and Herbert, Gregory S., "Uncovering Host-microbiome Interactions in Global Systems with Collaborative Programming: A Novel Approach Integrating Social and Data Sciences [version 1; peer review: awaiting peer review]" (2020). *School of Geosciences Faculty and Staff Publications*. 2312.
https://digitalcommons.usf.edu/geo_facpub/2312

F1000Research

SOFTWARE TOOL ARTICLE

Uncovering host-microbiome interactions in global systems with collaborative programming: a novel approach integrating social and data sciences [version 1; peer review: awaiting peer review]

Jenna Oberstaller[1], Swamy Rakesh Adapa[1], Guy W. Dayhoff II[1], Justin Gibbons[1], Thomas E. Keller[1], Chang Li[1], Jean Lim[1], Minh Pham[1], Anujit Sarkar[1], Ravi Sharma[1], Agaz H. Wani[1], Andrea Vianello[1], Linh M. Duong[1], Chenggi Wang[1], Celine Grace F. Atkinson[1], Madeleine Barrow[1], Nathan W. Van Bibber[1], Jan Dahrendorff[1], David A. E. Dean[1], Omkar Dokur[1], Gloria C. Ferreira[1], Mitchell Hastings[1], Gregory S. Herbert[1], Khandaker Tasnim Huq[1], Youngchul Kim[1,2], Xiangyun Liao[3], XiaoMing Liu[1], Fahad Mansuri[1], Lynn B. Martin[1], Elizabeth M. Miller[1], Ojas Natarajan[1], Jinyong Pang[1], Francesca Prieto[1], Peter W. Radulovic[1], Vyoma Sheth[1], Matthew Sumpter[1], Desirae Sutherland[1], Nisha Vijayakumar[1], Rays H. Y. Jiang[1]

[1]University of South Florida, Tampa, FL, 33612, USA
[2]Moffit Cancer Center, Tampa, FL, 33612, USA
[3]Texas A&M, College Station, TX, USA

v1 **First published:** 17 Dec 2020, **9**:1478
https://doi.org/10.12688/f1000research.26459.1

Latest published: 17 Dec 2020, **9**:1478
https://doi.org/10.12688/f1000research.26459.1

Open Peer Review

Reviewer Status *AWAITING PEER REVIEW*

Any reports and responses or comments on the article can be found at the end of the article.

Abstract

Microbiome data are undergoing exponential growth powered by rapid technological advancement. As the scope and depth of microbiome research increases, cross-disciplinary research is urgently needed for interpreting and harnessing the unprecedented data output. However, conventional research settings pose challenges to much-needed interdisciplinary research efforts due to barriers in scientific terminologies, methodology and research-culture. To breach these barriers, our University of South Florida OneHealth Codeathon was designed to be an interactive, hands-on event that solves real-world data problems. The format brought together students, postdocs, faculty, researchers, and clinicians in a uniquely cross-disciplinary, team-focused setting. Teams were formed to encourage equitable distribution of diverse domain-experts and proficient programmers, with beginners to experts on each team. To unify the intellectual framework, we set the focus on the topics of microbiome interactions at different scales from clinical to environmental sciences, leveraging local expertise in the fields of genetics, genomics, clinical data, and social and geospatial sciences. As a result, teams developed working methods and pipelines to face major challenges in current microbiome research, including data integration, experimental power calculations, geospatial mapping, and machine-learning classifiers. This broad, transdisciplinary and efficient workflow will be an example for future workshops to deliver useful data-science products.

Keywords
hackathon, codeathon, data science, transdisciplinary, gut microbiome, oral microbiome, human migration microbiome, Clinical Informatics, Bioinformatics, Operational Taxonomic Unit (OTU), 16S rRNA, machine learning, Geographic Information Systems (GIS)

This article is included in the Hackathons collection.

Corresponding author: Rays H. Y. Jiang (Jiang2@usf.edu)

Author roles: Oberstaller J: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Adapa SR**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Dayhoff II GW**: Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Gibbons J**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Keller TE**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Li C**: Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Lim J**: Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Pham M**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Sarkar A**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Sharma R**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Wani AH**: Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Vianello A**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Duong LM**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Wang C**: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Atkinson CGF**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Barrow M**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Van Bibber NW**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Dahrendorff J**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Dean DAE**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Dokur O**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Ferreira GC**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Hastings M**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Herbert GS**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Huq KT**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Kim Y**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Liao X**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Liu X**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Mansuri F**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Martin LB**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Miller EM**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Natarajan O**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Pang J**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Prieto F**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Radulovic PW**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Sheth V**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Sumpter M**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Sutherland D**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Vijayakumar N**: Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Jiang RHY**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The OneHealth codeathon is supported by the USF Conference and Event award (Fall 2019 cycle) to RHYJ; USF Genomics program, Institute for the Advanced Study of Culture and the Environment, and USF library systems. RHYJ's microbiome research is supported by COPH interdisciplinary research award (2019 -20 cycle), and USF-Exeter international catalytic research award. *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

How to cite this article: Oberstaller J, Adapa SR, Dayhoff II GW *et al.* **Uncovering host-microbiome interactions in global systems with collaborative programming: a novel approach integrating social and data sciences [version 1; peer review: awaiting peer review]** F1000Research 2020, **9**:1478 https://doi.org/10.12688/f1000research.26459.1

First published: 17 Dec 2020, **9**:1478 https://doi.org/10.12688/f1000research.26459.1

Introduction

OneHealth Codeathon: Genesis of a working model for applied, interdisciplinary problem-solving

The National Institutes of Health National Center for Biotechnology Information (NIH NCBI) model for codeathons—intensely collaborative, time-limited data workshops which encourage teams of participants to produce software prototypes to solve problems related to a common biomedical topic—are an effective avenue for the generation of software prototypes in the biomedical informatics space. Our previous "Iron Hack" event[1], centered on rare iron-related diseases, was a transdisciplinary twist on this NCBI model designed to complement and unite local University of South Florida (USF) research programs, inspiring participation from clinicians, genetic counsellors, and researchers from a diversity of biomedical fields at all different career-stages.

We set out to further expand on the more traditional foundation of codeathons for this year's event, working with the local research-community to select challenges that would encourage and more heavily utilize skillsets less-traditionally drawn to codeathons (e.g. social science researchers), while also supporting emerging USF research initiatives and addressing wider challenges in biomedical data science. This year's event (dubbed the USF OneHealth Codeathon) therefore focused on the fast-evolving field of host-microbiome interactions, with concepts for our team-projects designed around data-centric problems encountered by our interdisciplinary participants in their research and practice. The event took place on USF's Tampa campus over February 26–28, 2020.

As a result of these intense collaborative efforts, teams developed resources that are relevant not only to microbiome studies, but also general bioinformatics problems. The objective of this report is to demonstrate the utility of a codeathon model to rapidly develop tools for human and environmental health research, with the added community-building benefits of (1) providing opportunities for meaningful, long-term, cross-departmental interactions that stimulate collaborations and creative project design, and (2) offering in-depth exposure to applied data-science for members of traditionally less-computational fields.

Critical gaps OneHealth Codeathon projects sought to address

We addressed challenges related to the host microbiome, including the great need for novel genomics tools to handle large, recently generated heterogenous microbiome datasets. We established six OneHealth Codeathon teams to develop six computational-tool prototypes broadly focused on (1) power calculation for microbiome study design, (2) geographical information systems-analysis of microbiome data and associated risk factors, (3) mining archaeological microbiome data, and (4) searching for ecological drivers of earth microbiomes (Figure 1). These team-efforts have led to the convergence of social science, ecology and medical communities with genomics data-science researchers to produce promising computational tools, strengthened through an iterative process of soliciting ideas and feedback from domain experts.

The remainder of this report is organized into subsections by project, beginning with a detailed description for the six projects, the motivations behind them, and the gaps they seek to fill. We next describe the methodologies and implementations of the projects into usable software applications, how to operate the software applications, and results produced using the software applications. Finally, we discuss the pros and cons of this new highly interdisciplinary and community-driven twist on more traditional hackathons.

Team 1 – MicroPower Plus

Project title: Microbiome power-calculation tool for biologists: towards rigorous, reproducible microbiome study-design

Rationale: Measured differences between sample groups can result from any number of experimental artifacts not reflective of actual biology, including differing definitions of what a clinical population signifies within different studies, how samples are prepared, and analytical decisions (e.g., bioinformatic and statistical tool-selection, parameter-settings[2–4]). Statistical power calculations are a key part of quality study-design, informing the sample-size required to have sufficient statistical power to detect differences between experimental groups. The size of this difference between groups—the effect size—should also be taken into account during experimental planning; smaller effects are more sensitive to being obscured by experimental noise. Sufficiently powered studies are critical for robust biological conclusions, and funding agencies increasingly require power and sample-size analyses to consider applications for support.

R-based software packages enabling power analyses modeling relationships between sample-size and detectable effect-size using PERMANOVA-based methods have been developed to estimate required samples for microbiome experimental design[5], given input data from pilot studies. These handy tools are not generally accessible to biologists with limited computational experience and/or a more cursory grasp of statistics. We sought to build on these methods to create a more intuitive calculator/guide for biologists, who often need only a quick point-and-click reference for experimental planning.

Goal: To provide an intuitive power- and effect-size calculator-tool for biologists with limited computational experience.

Methods

Data-sources and processing

Predicted effect sizes detectable at a range of sample sizes and power-levels were precomputed on OTU tables from a variety of human body-site datasets from the Human Microbiome Project (HMP) using the R package micropower (v0.4) (Jaccard distance method)[5,6]. We used these precomputed data as a reference for quick and interactive power calculations for commonly used sample sizes by body-site.

We added additional functionality for calculating the effect size of the experimental intervention given a control group vs. an experimental group using linear modeling. Our tool computes the

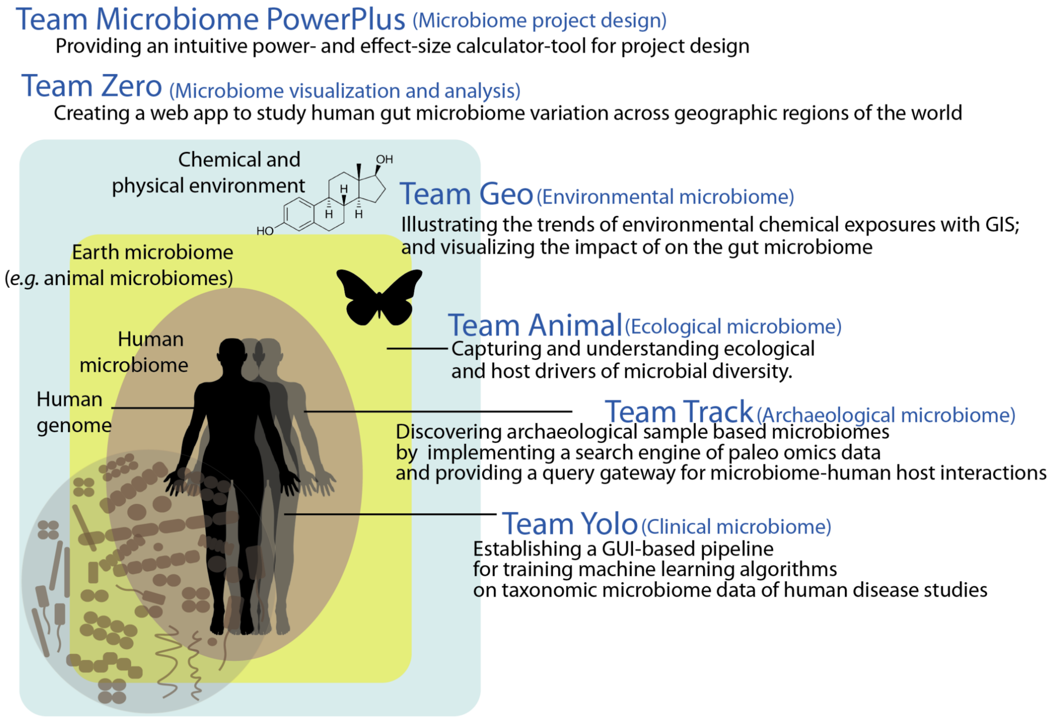

Team Microbiome PowerPlus (Microbiome project design)
Providing an intuitive power- and effect-size calculator-tool for project design
Team Zero (Microbiome visualization and analysis)
Creating a web app to study human gut microbiome variation across geographic regions of the world
Chemical and physical environment
OH
H
H
H
HO
Team Geo (Environmental microbiome)
Illustrating the trends of environmental chemical exposures with GIS; and visualizing the impact of on the gut microbiome
Earth microbiome (e.g. animal microbiomes)
Team Animal (Ecological microbiome)
Capturing and understanding ecological and host drivers of microbial diversity.
Human microbiome
Human genome
Team Track (Archaeological microbiome)
Discovering archaeological sample based microbiomes by implementing a search engine of paleo omics data and providing a query gateway for microbiome-human host interactions
Team Yolo (Clinical microbiome)
Establishing a GUI-based pipeline for training machine learning algorithms on taxonomic microbiome data of human disease studies

Figure 1. Scope of human holobiont interactions with microbiomes in various contexts explored through USF's OneHealth Codeathon. Two teams (Teams MicroPower Plus and Zero) focused on developing practical computational tools for microbiome study-design and data-analysis. Four teams (Teams Geo, Animal, Track and Yolo) focused on exploring different aspects of host-microbiome interactions from environmental consequences to clinical presentations.

Bray-Curtis distance between all samples, then uses the Adonis function from the vegan package (v2.5.6) to calculate the correlation parameter Pearson's R[7].

A conceptual overview of MicroPower Plus functionality is provided in Figure 2.

Operation and Implementation

The MicroPower Plus[8] workflow is implemented in a user-friendly R-Shiny web application. RStudio and the R packages shiny, plotly and tidyverse are required to operate MicroPower Plus[9–12]. Further documentation and a tutorial are available at the GitHub repository as listed in the code-availability section.

After installation of required packages, all necessary tutorial files can be downloaded from GitHub onto the user's local computer, and MicroPowerPlus can be launched by opening the "app.R" file in RStudio.

Use cases

MicroPower Plus[8] is most useful as a statistical reference-guide for biologists to make quick calculations to aid in experimental design of microbiome studies. We built a user-interface around the human gut microbiome reference dataset that allows the user to visualize the relationship between sample size, effect size and statistical power as a proof of concept using R Shiny[10]. Resulting effect size is reported as a bar graph, with reference to effect sizes reported in the literature for comparisons. We created an additional tool that allows the user to input their own data, calculate the effect size from their experiment and report it as a bar graph. Future iterations of this tool will include interactive visualizations for the pre-computed reference data from other body-sites.

The provided tutorial walks the user through an example power calculation (Figure 3) and effect size calculation (Figure 4) using the pre-computed human gut microbiome datasets.

Team 2 – GEO

Project title: Environmental Chemicals: Impact on Human Microbiomes

Rationale: Environmental exposures to chemicals have been a public health concern due to the ubiquitous nature of its effects on human health and the environment. Industries and manufacturing sectors contribute to chemical exposures by releasing these chemicals into the environment. Chemicals commonly found in commercial products, such as heavy metals and chlorinated hydrocarbon solvents, can persist in the environment for extended periods, increasing the latency of exposure[13].

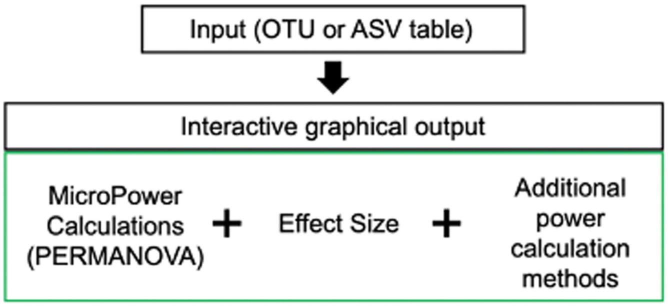
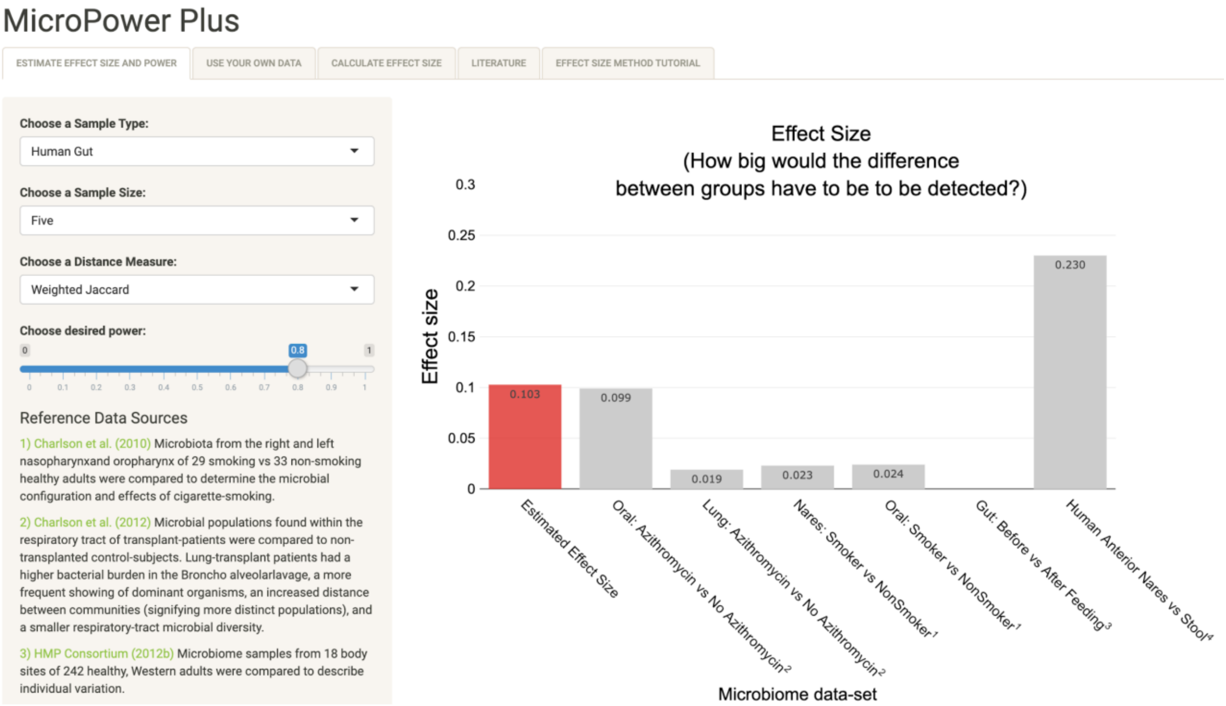

Input (OTU or ASV table)
Interactive graphical output
MicroPower Calculations (PERMANOVA)
+
Effect Size
+
Additional power calculation methods

Figure 2. MicroPower Plus functionality conceptual overview. Input-data are OTU or ASV tables selected from curated, published microbiome studies of various human body-sites from which effect size has been pre-calculated for several common sample-sizes using complementary methodologies. The user can then use the interactive, graphical output to explore the relationships between effect-size, sample size and statistical power to use as a quick reference for their own experimental planning.


MicroPower Plus
ESTIMATE EFFECT SIZE AND POWER
USE YOUR OWN DATA
CALCULATE EFFECT SIZE
LITERATURE
EFFECT SIZE METHOD TUTORIAL
0
0.8
1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Effect Size
(How big would the difference
between groups have to be to be detected?)
Effect size
0.3
0.25
0.2
0.15
0.1
0.05
0
0.103
0.099
0.019
0.023
0.024
0.230
Estimated Effect Size
Oral: Azithromycin vs No Azithromycin2
Lung: Azithromycin vs No Azithromycin2
Nares: Smoker vs NonSmoker1
Oral: Smoker vs NonSmoker1
Gut: Before vs After Feeding3
Human Anterior Nares vs Stool4
Microbiome data-set

Choose a Sample Type: Human Gut

Choose a Sample Size: Five

Choose a Distance Measure: Weighted Jaccard

Choose desired power: 0.8

Reference Data Sources

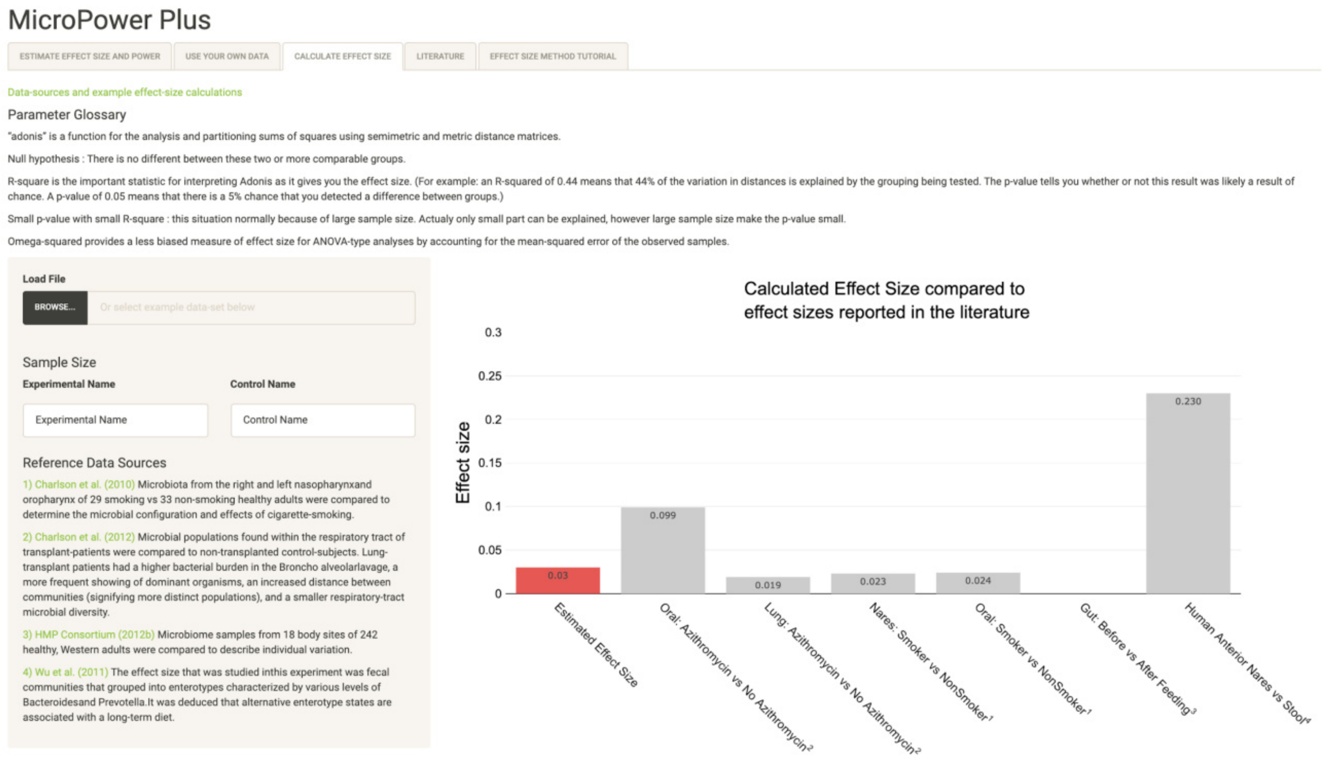
1) Charlson et al. (2010) Microbiota from the right and left nasopharynxand oropharynx of 29 smoking vs 33 non-smoking healthy adults were compared to determine the microbial configuration and effects of cigarette-smoking.

2) Charlson et al. (2012) Microbial populations found within the respiratory tract of transplant-patients were compared to non-transplanted control-subjects. Lung-transplant patients had a higher bacterial burden in the Broncho alveolarlavage, a more frequent showing of dominant organisms, an increased distance between communities (signifying more distinct populations), and a smaller respiratory-tract microbial diversity.

3) HMP Consortium (2012b) Microbiome samples from 18 body sites of 242 healthy, Western adults were compared to describe individual variation.

Figure 3. MicroPower Plus power-calculation interface. The user selects the sample type, the sample size for each group and a distance measure. When the user moves the power slider, the estimated effect-size graph (red) changes to the minimum effect size required to attain the given power level. The gray bars reference effect sizes calculated from the indicated sources. By comparing the estimated effect size to the reference effect sizes, the user can get a sense of how large a difference would have to be between their samples to detect significance using different experimental designs.

A lack of information led to relatively few rules for handling and disposing of chemicals in the first part of the 20th century, which resulted in the random release of these hazardous chemicals and toxins into the environment. Knowledge of toxic waste dumps and their associated human health and environmental health consequences received national attention in the late 1970's[14]. In response to public outcry, Congress created "Superfund" in the 1980's to fund toxic waste clean-up at industrial sites[14,15]. Superfund sites require long-term remediation efforts, and sites are evaluated for eligibility on a point-based system requiring a preliminary assessment and site-inspection (known as the Hazard Ranking System, or HRS)[16]. Reporting from the public or an agency is also considered in assessing a site for the qualification. Superfund sites are prioritized by HRS score onto the National Priority List (NPL)[16]. Currently there are 1335 NPL sites around the U.S., each having specific chemical contaminations.


MicroPower Plus
ESTIMATE EFFECT SIZE AND POWER
USE YOUR OWN DATA
CALCULATE EFFECT SIZE
LITERATURE
EFFECT SIZE METHOD TUTORIAL
Data-sources and example effect-size calculations
Load File
BROWSE...
Or select example data-set below
Sample Size
Experimental Name
Control Name
Experimental Name
Control Name
Calculated Effect Size compared to effect sizes reported in the literature
Effect size
0.3
0.25
0.2
0.15
0.1
0.05
0
0.03
0.099
0.019
0.023
0.024
0.230
Estimated Effect Size
Oral: Azithromycin vs No Azithromycin[2]
Lung: Azithromycin vs No Azithromycin[2]
Nares: Smoker vs NonSmoker[1]
Oral: Smoker vs NonSmoker[1]
Gut: Before vs After Feeding[3]
Human Anterior Nares vs Stool[4]

Parameter Glossary

"adonis" is a function for the analysis and partitioning sums of squares using semimetric and metric distance matrices.

Null hypothesis : There is no different between these two or more comparable groups.

R-square is the important statistic for interpreting Adonis as it gives you the effect size. (For example: an R-squared of 0.44 means that 44% of the variation in distances is explained by the grouping being tested. The p-value tells you whether or not this result was likely a result of chance. A p-value of 0.05 means that there is a 5% chance that you detected a difference between groups.)

Small p-value with small R-square : this situation normally because of large sample size. Actualy only small part can be explained, however large sample size make the p-value small.

Omega-squared provides a less biased measure of effect size for ANOVA-type analyses by accounting for the mean-squared error of the observed samples.

Reference Data Sources

1) Charlson et al. (2010) Microbiota from the right and left nasopharynxand oropharynx of 29 smoking vs 33 non-smoking healthy adults were compared to determine the microbial configuration and effects of cigarette-smoking.

2) Charlson et al. (2012) Microbial populations found within the respiratory tract of transplant-patients were compared to non-transplanted control-subjects. Lung-transplant patients had a higher bacterial burden in the Broncho alveolarlavage, a more frequent showing of dominant organisms, an increased distance between communities (signifying more distinct populations), and a smaller respiratory-tract microbial diversity.

3) HMP Consortium (2012b) Microbiome samples from 18 body sites of 242 healthy, Western adults were compared to describe individual variation.

4) Wu et al. (2011) The effect size that was studied inthis experiment was fecal communities that grouped into enterotypes characterized by various levels of Bacteroidesand Prevotella.It was deduced that alternative enterotype states are associated with a long-term diet.

Figure 4. MicroPower Plus effect-size calculation (concept). The user uploads a matrix of their microbiome measurements, enters the names of the groups that can be used to distinguish the sample columns by group. MicroPower Plus then calculates the effect size for the experiment (red bar). The gray bars are effect sizes calculated from the indicated literature. Comparing the red bar to the gray bars allows the user to get a sense of the magnitude of their experimental effect.

Human exposure to toxic chemicals has been shown to elicit different effects depending on the host's immune response, with long-term exposures associating with a range of serious maladies varying from cancers acting on various bodily tissues to neurological effects[17]. The gut microbiome is hypothesized to have a unique role in enhancing and maintaining host health through the microbiome-gut-brain axis and can impact endocrine, immunological and nutrient signals[18]. Microbiome dysbiosis can occur with exposure to toxic environmental contaminants via ingestion or inhalation and can lead to several chronic conditions. Due to its diverse functions in the body, the gut microbiome acts as an indicator for health, and there is a growing body of literature exploring the interactions of environmental contaminants with the host microbiome[13,17,18].

Environmental contaminants present in Superfund sites around the U.S. can significantly affect the health of the population in the surrounding areas. To illustrate this effect, we created a tool for visualizing the impact of environmental toxicants on the gut microbiome.

Goals: 1) To illustrate the trends of environmental chemical exposures from U.S. Superfund sites over time. 2) to create a tool for visualizing the impact of exposure to environmental chemicals on the gut microbiome around the U.S.

Methods

Implementation and Operation

Data-sources and processing: We processed and combined datasets from the American Gut Project (AGP), census data, and EPA Superfund data to search for informative patterns using the R package phyloseq 1.30.0[19]. We identified most abundant taxa by Superfund site/geographic location. We then performed basic association analyses to assess relationships between abundant/rare taxa, various Superfund sites and contaminants. Archived code are available, see *Software availability*[20].

1) American Gut Project data: The American Gut Project (AGP) is a large-scale, crowdsourced project (n =29778) of microbial sequence data with the aim of characterizing the human gut microbiome including associated mitigating factors ranging from diet, lifestyle, overall health, and the broader environment. The metadata file obtained from AGP sample information (file 04-meta). was reduced to responses from participants within the United States only. Important variables that have been previously found to be associated with differential phenotypes mediated by air pollution in microbial communities in published studies were also selected and included in subsequent testing for associations with Superfund-site proximity.

2) Superfund data: Superfund sites and associated contamination data for current NPL sites were retrieved from EPA data using the R superfundr 0.0.0.9000 package. The data were prepared and transformed using Statistical Analysis Software (SAS v 9.4, Cary, NC). We focused on 10 priority chemicals listed by the EPA.

3) Census data: Select data from the American Community Survey (ACS) were downloaded from the U.S. Census Bureau: American FactFinder website via the download center (U.S. Census Bureau, 2020). This population-based data source provides descriptive socio-demographic data (e.g., sex, race, ethnicity, economic indicators, etc.) by zip code across the nation. Once all datasets were downloaded for each variable, all variables were then merged by a linking variable (i.e., zip code) that each dataset had in common. After data-cleaning, percentages were calculated for each variable. All data-cleaning was conducted using Statistical Analysis Software (SAS v 9.4).

Loading and filtering OTU tables was memory-intensive, as the initial dataset is very large. Initial attempts for loading the OTU table with a 16 GB laptop were insufficient. To solve the problem, we performed this filtering on a high-performance computation cluster with 180 GB of memory.

Merging data across disparate datasets: Several distinct datasets across the AGP, Superfund, and ACS provided unique information connected only by geographic location and could be merged by an appropriate linking variable (e.g., zip code). Data from all three sources were combined for a total of ~1000 samples. We further reduced the dataset to only samples that were directly related to the gut for downstream prediction using machine learning approaches.

ArcMap version 10.7 (2020) was used to create choropleth maps from the combined ACS and Superfund datasets to evaluate the association of chemicals found at EPA Superfund sites with select population-based socio-demographic data by zip code overtime. An open source software can be used for the same work is QGIS Geographic Information System, at Open Source Geospatial Foundation Project (http://qgis.org).

Machine learning analysis on data collected from individuals near Superfund sites: We selected individuals that were self-identified to be within 5 km of Superfund sites from the final combined dataset. We next performed a classification analysis using random forests implemented via the R package ranger[21]. For each contaminant, we classified each individual as exposed or unexposed based on their proximity to a Superfund site with that contaminant. We then performed 10-fold cross-fold validation and reported the accuracy of the most and least informative contaminants in regard to the microbiome.

Results

Geographic distribution of select Superfund-site contaminants and abundance of *Bacteriodetes* OTUs are shown in Figure 5. We next explored a potential relationship between abundance of this bacterial phylum and individual contaminants, and further possible predictive efficacy of contaminants for certain OTUs, using proof-of-concept modeling. We restricted samples to those within 5 km of a Superfund site for these analyses. We constructed a random forest using each contaminant as a binary predictor-variable. We found a strong relationship between several contaminants and microbial composition. The two most predictive contaminants were polycyclic aromatic hydrocarbons and polychlorinated biphenyls (PAH, 94% and PCB, 81%, respectively). The contaminant with the lowest accuracy was lead (60%).

It is worth noting that PAH are known to bio-amplify as they go through food-webs. Other health outcomes linked to PAH exposure are various forms of cancer, as well as developmental impacts. PCB have been banned in the manufacturing process since 1979, yet they do not readily break down and remain a hazard over long periods of time. Because of these properties, they are commonly listed as Superfund contaminants of high concern. In conclusion, we found that for several contaminants the microbial composition varied significantly among individuals living near Superfund sites with high or low levels of PAH and PCB, respectively.

Team 3 - ZERO

Project title: Creating a web app to study human gut microbiome variation across geographic regions of the world

Project Rationales, Descriptions and Goals

Rationale: The human gut microbiome is one of the most densely populated sites by bacteria in the human body. It performs numerous functions, and its dysbiosis has been associated with several diseases. A major goal of microbiome researchers has been to understand the diversity of the gut microbiome across human populations. Although several studies have been undertaken for this purpose, these studies are limited in scope and comparative ability. Therefore, the rationale of the present work was to create a web tool which will be equipped with reference databases, populations and necessary scripts for the users to upload, analyze and visualize their own microbiome data at the server, with additional options to compare with the reference populations. Results can subsequently be downloaded by the user. Finally, all the reference population data is to be made available for download, along with necessary scripts to enable the user to run the program on their local computers, without the need to upload their raw data. Such a tool will be extremely useful to any interdisciplinary researchers who may have microbiome-related research questions but are not experienced in writing code, handling large microbiome datasets or who do not have access to advanced computational facilities. The codes, instructions and guidelines are available through a GitHub repository. The flowchart summarizing the approach is provided in Figure 6.

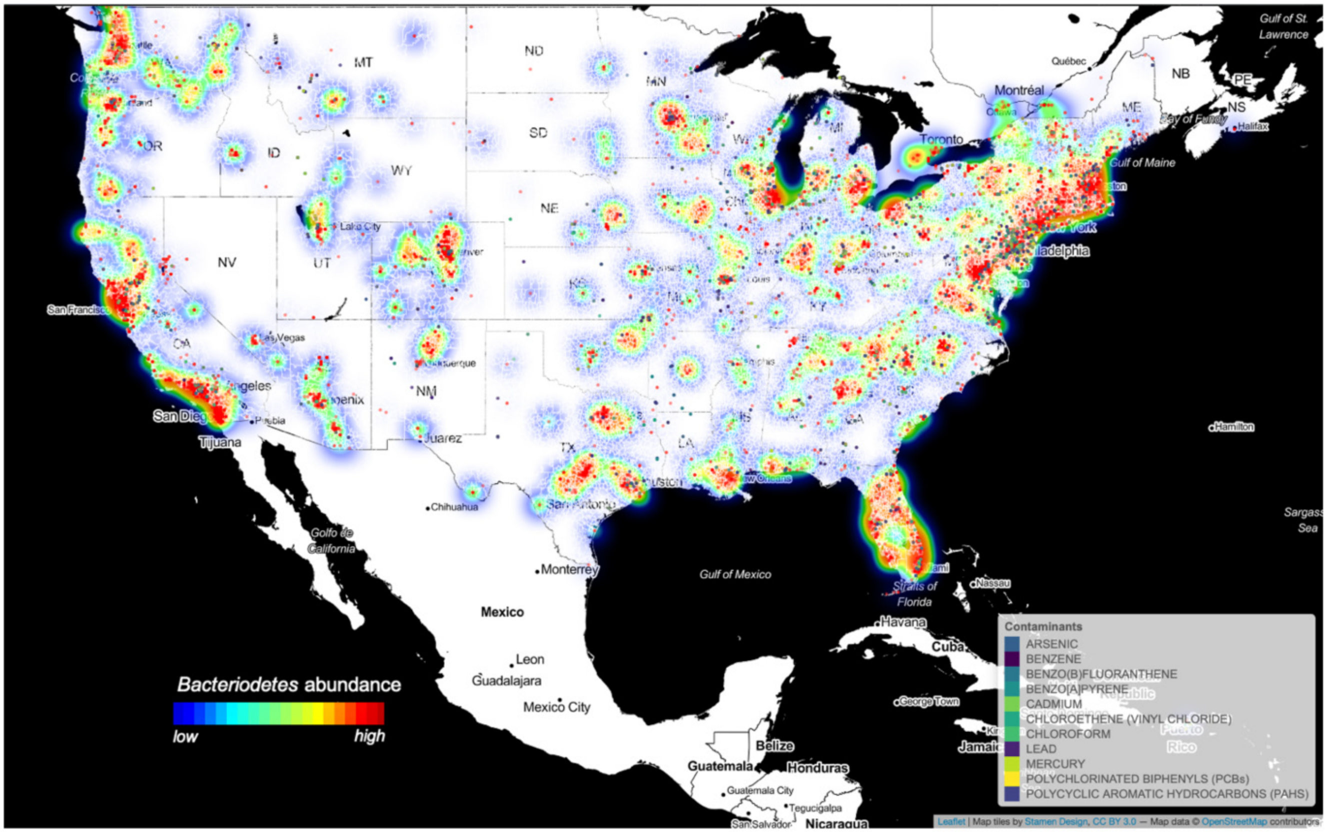


Figure 5. Contaminant associations with the most dominant bacterial phylum. Geographic distribution of select Superfund-site contaminants (circles color-coded by contaminant) and abundance of Bacteriodetes OTUs (underlying heatmap) from samples collected within 5km of a Superfund site. We found a strong relationship between several contaminants and microbial composition.

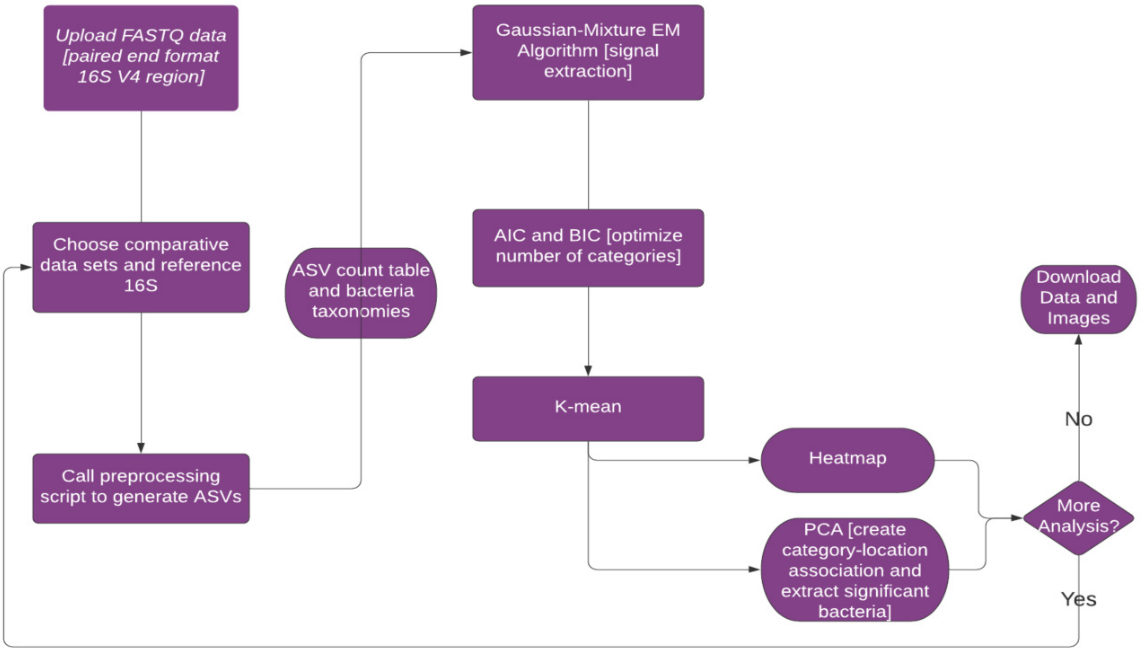

Upload FASTQ data [paired end format 16S V4 region]
Choose comparative data sets and reference 16S
Call preprocessing script to generate ASVs
ASV count table and bacteria taxonomies
Gaussian-Mixture EM Algorithm [signal extraction]
AIC and BIC [optimize number of categories]
K-mean
Heatmap
PCA [create category-location association and extract significant bacteria]
More Analysis?
No
Yes
Download Data and Images

Figure 6. Proposed Team Zero web-app workflow. Users will be able to upload fastq files for analyses and choose reference-datasets for comparison. The in-built pipeline will then generate the Amplicon Sequence Variants (ASVs) from which the most informative for differentiating populations will be chosen using a Gaussian-Mixture EM algorithm followed by unsupervised K-means clustering. Heatmaps and PCA-plots describing the data will be generated and made available for download.

Goals: 1) To download raw microbiome data (V4 region of 16S rRNA gene) from various world populations and generate amplicon sequence variant (ASV) table for comparison purposes. 2) Construct simple, but informative plots such as heatmaps and principle component analysis (PCA) plots to visualize relationships/patterns in the data through the proposed web app. 3) Provide all raw sequencing data, bash scripts and R scripts to run all steps of the analyses, as well as appropriate documentation and guidelines for an easy and error-free run of the pipeline on the user's local computer.

Methods

Data sources and processing

We first mined microbiome data from various world populations by geographical region. We narrowed our focus to studies on the human gut microbiome involving the V4 region of the 16S rRNA gene. A total of 1428 samples spanning populations from China, the Indian subcontinent (Himalayan region), Brazil and Europe meeting these criteria were incorporated. Raw data were downloaded from the European Nucleotide Archive (Accessions: China, PRJNA396815; Indian subcontinent, PRJEB29137; Europe, PRJNA497734; Brazil, PRJEB19103) (Table 1).

Despite this initial filtration step, analysis-time was still estimated to be too high to move forward under Codeathon time-restrictions. Thus, in a second step to reduce data volume, 5000 sequences were subsampled using Seqtk 1.3-r115-dirty[22] from each of the forward and reverse fastq files for each of the samples. All the downstream analyses were based on the subsampled reads. The fastq files were analyzed using the standard DADA2 1.14.1 pipeline[23] to generate the distribution of ASVs observed in this dataset. The corresponding classification of each ASV was obtained using the Silva database (v132)[24]. The bacterial count table was further utilized for downstream analysis.

The resulting ASV table contained 1,428 samples with 2,655 bacterial taxa. Considering the very sparse data in the ASV table (only 1.231% of ASV elements exhibit reads numbers > 0), we used a Gaussian-mixture model to remove the bacteria with lower reads-coverage. A total of 1,783 taxa were removed and the remaining ASV table was normalized for each sample by the proportion of reads in each taxon using orders-of-magnitude multipliers (1-e^8). The distribution of standard deviation in reads-number was calculated, and taxa at the tail-ends of the distribution were eliminated, leaving 237 taxa. Similarly, individual samples at the extreme low-end of the reads-number distribution (365 samples) were also removed using the Gaussian-mixture model. Unfortunately, all Chinese-population samples were eliminated during this step, and all downstream analyses were performed only on the populations from Europe, Brazil and the Indian subcontinent.

Table 1. **Team Zero web-app data-sources by population.**

Population	ENA study accession No.	No. of samples
China	PRJNA396815	200
Indian subcontinent	PRJEB29137	77
Europe	PRJNA497734	1000
Brazil	PRJEB19103	150

Modeling relationships between population and bacterial taxa

We used the resulting filtered dataset to perform K-means clustering to determine the optimal number of categories, finding k=18 to be most informative for the data. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) were utilized to measure model robustness.

Operation and implementation

We incorporated a set of unsupervised machine learning back-end computational methods to investigate the datasets for encoded geographical information. We used python v3.6.9 along with the django web framework and conda 4.7.12 to build our workflow. The machine learning components of the workflow to identify ASVs distinguishing populations by geography are performed using TensorFlow2[25]. Data preprocessing and data visualization are mediated through R scripts (see Implementation and Software Availability).

Herein we implemented a web-based application[26] for the deposition and rapid analysis of microbiome data. Importantly, users are able to (1) download a prepared database along with the server source code, or (2) construct their own database for analysis. The web-based application source code, the preprocessing and data visualization scripts, and instructions for their usage are available online as listed in the *Software availability* section.

Results

The unsupervised classification algorithm indicates strong bacterial association with geographic populations

Our k-means parameter-exploration indicated 18 classes within the sample ASV data. The result indicates at least one or two bacterial groups are enriched for each class (Figure 7A). Classification further indicated differences in community composition by geographical location (Figure 7B). We performed a PCA to further characterize the relationship between sample categories detected via clustering. We found that the samples from classes 1, 6, 9 and 14 form clearly distinct clusters from each other (Figure 7C), further indicative of underlying geographic patterns. We identified important bacterial taxa contributing to sample classification (Figure 8) and plotted relative contribution of each ASV (classified up to genus-level) driving ordination (Figure 9). Differential relative abundance of these ASVs across all geographic populations indicated distinct geographical patterns, with several ASVs strongly associating with Indian, Brazilian, or European (to a lesser extent) populations (Figure 9). The classification of the ASVs corresponding to Figure 9 are provided in Table 2.

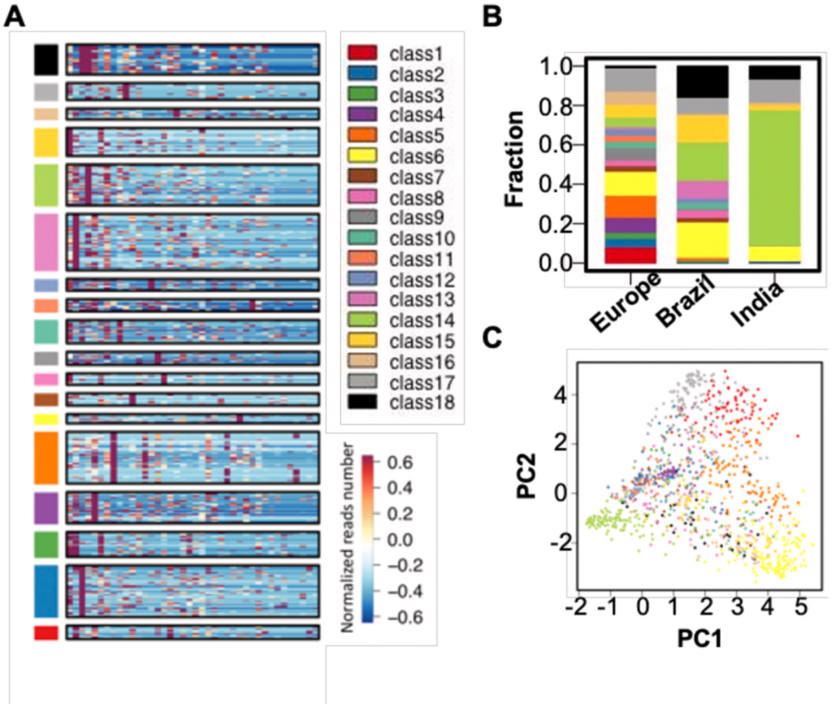


Figure 7. Samples cluster distinctly by OTU composition and geographic population. The color scales indicate the 18 categories used for the classification and normalized reads-number for the studied samples. (**A**) The heatmap indicates enrichment for at least one or two bacterial OTUs in each cluster. (**B**) Enrichment of K-means category by geographic location. The 18 classes showed maximum differential abundance across the three studied populations. (**C**) The PCA plot shows the sample affinities for the classes 1, 6, 9 and 14 which showed the greatest geographical pattern.

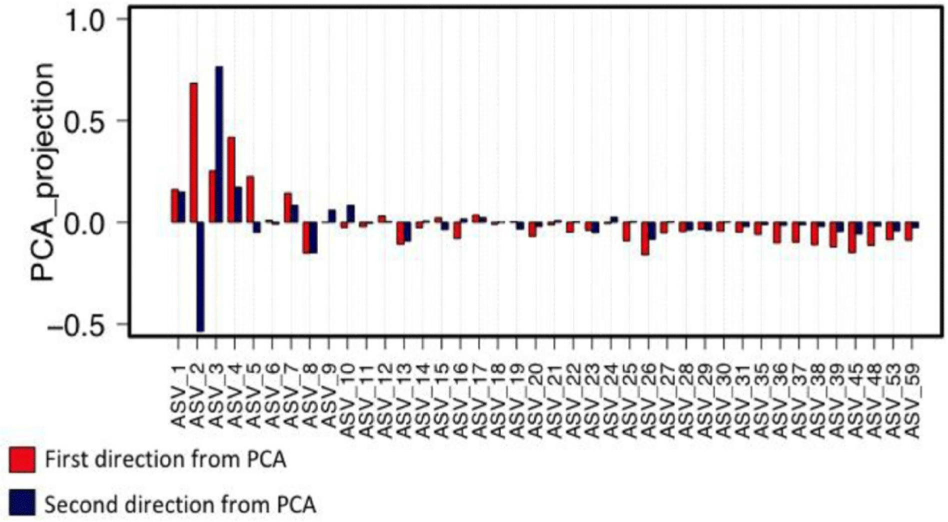


Figure 8. Bacteria driving sample classification. The X-axis shows the major ASVs, and their relative contribution to the PCA (Figure 7B) is shown on the Y-axis.

Conclusions

Our work was aimed at creating a web app to study the geographical patterns of the human microbiome and selecting features which could be useful to distinguish the populations. Using publicly available resources, we were able to include different geographical populations and select features to identify differences across groups. The resources for our study are deposited in our GitHub repository (see *Software availability*). Limitations of

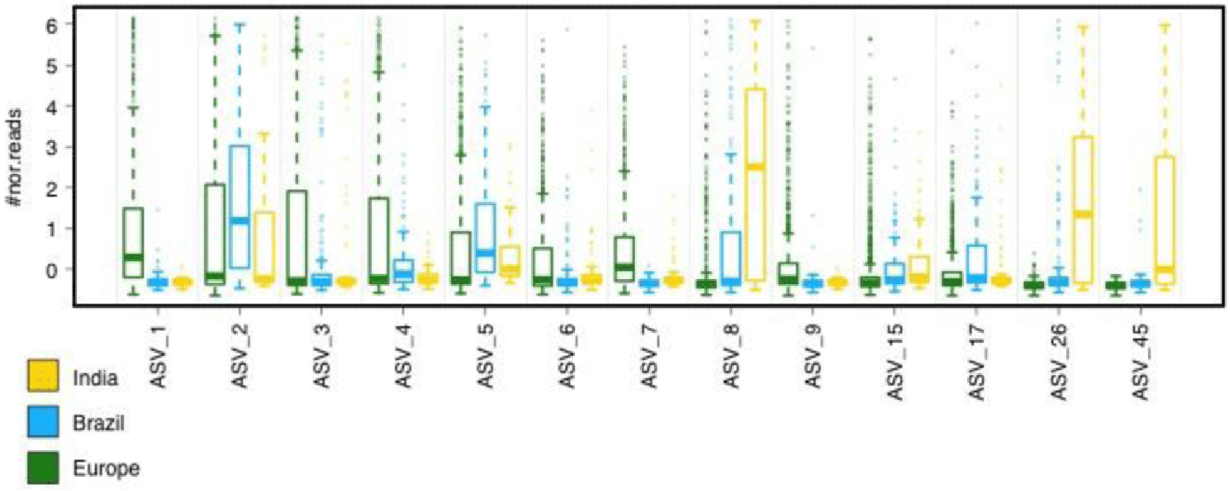


Figure 9. The top 13 bacterial taxa driving sample-classification have strong population associations. The color of the boxplot indicates geographic affiliations. The X-axis indicates the top 13 ASVs and the Y-axis shows the corresponding number of normalized reads.

Table 2. Classification of ASVs displaying highest geographical patterns as shown in Figure 9. Classification only up to genus level were obtained since the studied region was limited to V4 of the 16S rRNA gene. When two ASVs were affiliated with the same genus, they were distinguished by adding a serial number as suffix. For example, Bacteroides_1 and Bacteroides_2 belong to the same genus.

ASV	Phylum	Class	Order	Family	Genus
ASV_1	Actinobacteria	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium
ASV_2	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides_1
ASV_3	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides_2
ASV_4	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Faecalibacterium_1
ASV_5	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Faecalibacterium_2
ASV_6	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
ASV_7	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Blautia
ASV_8	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella_9
ASV_9	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	NA
ASV_15	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Agathobacter
ASV_17	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides_3
ASV_26	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella_10
ASV_45	Proteobacteria	Gammaproteobacteria	Aeromonadales	Succinivibrionaceae	Succinivibrio

this study include that factors such as age, gender and other participant phenotypes which could be contributing to geographical patterns were not included in these analyses. However, we were able to create a web-app prototype for identifying features from the composition of the human gut microbiome related to geographical population. In the future, this work can be extended to include other variable regions of the 16S rRNA gene, as well as including other body sites such as the oral cavity, skin, etc. Similarly, batch-effect correction-tools need to be incorporated for unbiased comparison across different studies.

Team 4 - YOLO

Project title: A web-based machine learning pipeline for disease prediction using microbial data

Project Rationales, Descriptions and Goals

Rationale: High-throughput sequencing technologies have resulted in the generation of an increasing amount of microbial data, such as microbiome data. Using these data, machine learning methods are powerful in identifying functionally active microbes and predicting disease status. Even though machine learning algorithms are popular approaches to investigate microbiome, to adopt these methods effectively usually requires specialized training. In addition, model selection and hyper-parameter tuning can be time-consuming even for experienced practitioners. Thus, our project focused on the efficiency of AI in solving big-data problems and facilitating humans to perform other cognition-demanding tasks by developing a GUI-based pipeline for training machine learning algorithms on taxonomic microbiome data. Our pipeline expands access of computational tools to researchers in non-computational disciplines to improve cross-disciplinary study. As a proof of concept, we successfully utilized our pipeline to train a predictive algorithm for obesity rates based upon orthogonal taxonomic units which may be applied toward generating health-related features from clinical, historical, or forensic samples. Our code utilizes three methods: K-nearest neighbors (KNN), support vector machine (SVM), and adaptive boosting (AdaBoost) to achieve respective accuracies near eighty-four, ninety-one, and eighty-six percent. Both KNN and SVM utilized a 10-fold cross-validations to prevent overtraining. Under this method, training was achieved near instantaneously on a 16 GB MacBook to demonstrate feasibility. Outputs are processed into interactive graphical visualizations to improve ease-of-use. Although previous projects have utilized these computational techniques toward processing microbiome data, our pipeline removes barriers to use for researchers without coding backgrounds while streamlining efficiency for all users.

Studies have revealed significant diversity in the gut microbiome composition related to various phenotypes. Obesity has been associated with changes in the microbiota at phylum-level, reduction in bacterial diversity, and different representations of bacterial genes. For example, studies of lean and obese mice suggest a strong relationship between gut microbiome and obesity. Phylogenetic marker genes uncovered by 16S rRNA gene amplicon sequencing have revolutionized the field of microbial ecology. This PCR-based method has the advantage of identifying difficult to culture bacterial organisms. Various bioinformatic pipelines can then group these sequences into clusters called OTUs. OTUs are based on their sequence similarity to each other rather than a reference taxonomic dataset which may be biased towards existing taxonomic classification[27].

Goals: We were interested in finding out if there is an association between gut microbiome OTUs and obesity. Additionally, we wanted to be able to use this data to distinguish between lean, overweight, and obese phenotypes in humans. We were able to successfully develop a machine-learning based pipeline that shows the association between gut microbiome OTUs and obesity with high accuracy. Furthermore, this pipeline can predict whether sample OTU data comes from a lean, overweight, or obese human phenotypes. Our work is significant because a heavy coding background is not required for use of high-accuracy machine learning tools.

Methods

Data preprocessing

To develop our pipeline[28], sample microbiome data was retrieved from 29. First, we cleaned the data by removing duplicate entries which leaves us with 151 samples. Second, to deal with the sparsity of OTU count data, we added a random small positive number to all 0 entries. Third, data was normalized using the centered log-ratio (CLR) transformation[30]. Then, the dimensionality reduction was performed. We chose to use the Max-min Markov Blanket method to recursively select a small subset of features that are important to the outcome of interest (Obesity or lean in this case). A total of 10 highly informative OTUs were selected during this process and various machine learning methods were explored based on a recent review article[31].

Data transformations and machine learning methods

Principal component analysis (PCA) is an unsupervised dimensionality reduction technique that finds relationships in the dataset, then transforms and reduces them into principal components (i.e. uncorrelated features that embody the information contained within the dataset) that do not have redundant information.

Random forest describes a supervised machine learning strategy that splits samples into successively smaller groups based on specific features and associated threshold values. This method is in the planning phase for future versions.

SVM is a method of supervised machine learning that is useful for classification, regression, and detection of outliers. SVMs are effective in higher dimensions where the dimensions are greater than the numbers of samples. Linear Support Vector Machine (SVM) classifier was used to project samples into a higher dimensional space so that they are linearly separable. Linear SVM was performed using 10-fold cross-validation with 3 repeats.

KNN is a machine learning algorithm that can be used for classification and regression. In our pipeline, KNN classifier was used for the classification of disease-status, with classification determined by majority-vote of close-by data points ($n = K$).

AdaBoost is a machine learning meta-algorithm that can be used to improve performance of other machine learning algorithms. AdaBoost classifier was used to train multiple tree classifiers (where each tree has a subset of available features) to iteratively add more weight to those misclassified samples in the next training loop. GitHub readme and description are available in the software accessibility section.

Operation and implementation

We implemented various machine learning models, namely k Nearest Neighbor, AdaBoost, and Support Vector Machines, to predict disease from the microbiome pre-processed data. It includes three main steps. 1) Users can prepare the biome OTU table to perform downstream analysis, such as PCA and machine learning. 2) In the next step, the processed data can be used to perform PCA for exploratory analysis. 3) The data is fed into machine learning models to select the highly predictive features and for the final prediction of disease-status.

Feature selection and machine learning were implemented using MXM 1.4.5[32] and caret 6.0-85 R packages[33], respectively, in R version 3.6. To make it easy for others to use this implementation, we designed a shiny application with an intuitive graphical user interface (GUI). Users can plot, visualize, and download their results generated through the app.

Results

We show that machine learning can be used to differentiate disease from the normal states using OTU information. We used pre-processed data from a twin study with 281 samples and 5462 OTUs[29]. For exploratory analysis we performed PCA (Figure 10 and Figure 11; analyses and plots generated using our Shiny app) as shown in Figure 10 and Figure 11. This analysis and plots are generated using the Shiny app. We performed feature selection to select the highly significant features, shown in Table 3. Abundance of significant OTUs is shown in Figure 12. By using a set of predefined hyperparameters for each model, we achieved 10-fold cross validated accuracy of 0.936 using a linear support vector machine (Figure 13). Additionally, 10 OTUs we identified as important to obesity-status are provided in Table 3. While we do not have assigned significant functional annotations for them in the current development, future studies could use them as candidate functional groups to aid experimental design for validating clinical and public health microbiome findings.

Team 5 - TRACK

Project title: Tracking ancient global epidemics

Project Rationales, Descriptions and Goals

Rationale: As the collection of human microbiome data grows, developing user-friendly tools to search proteomics databases has become critical. Bridging the gap between computer science and biological science expertise will facilitate microbiome analysis for both explanatory and predictive purposes, making significant additions to general knowledge in this field. Such effective and convenient methods of sifting through vast datasets would be well-suited to the investigation of not only modern-day microbiome samples, but also preserved historical microbial and proteomic data retrieved from ancient populations at archaeological sites worldwide. Proteomic determination of the microbes of deceased individuals would provide another dimension to forensic analysis by uncovering the pathogens that might have been responsible for their death. The significance of this determination goes beyond simply detecting the presence of bacterial peptides, also extending to tracking the migration and virulence of diseases over time in human populations.

Exploring ancient or paleolithic host-microbiome interactions is an emerging approach to explore widespread microbial infectious diseases, and even pandemics, by identifying pathogen-expressed proteins in human dental calculus. This approach is supplemented by data from metabolomic analyses, anthropological and paleopathological data from the skeletal material, archaeological contexts, and archival data. Examining protein content of dental calculus has typically given insight into diet and oral health of communities of past generations[34–36].

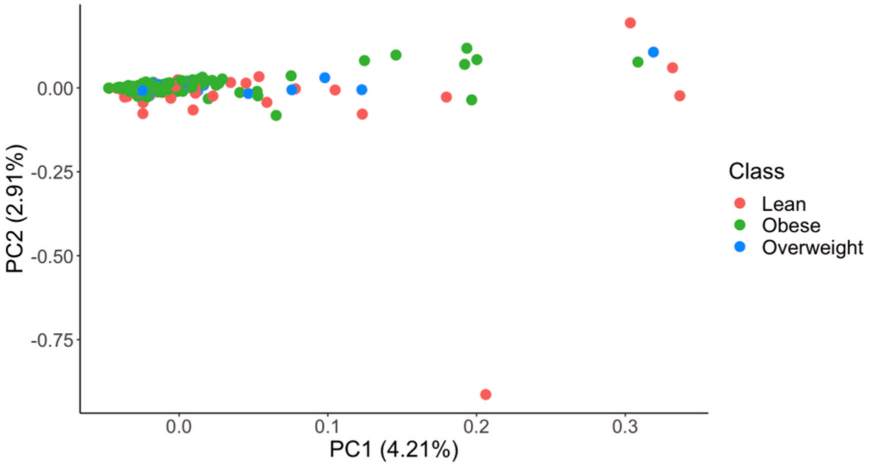


Figure 10. A principal component analysis of microbiome data from over 5400 OTUs involving 281 individuals by disease-class. PCA plot tries to identify linear combinations of different OTUs (features) corresponding to microbiome composition discriminating by disease class. PC1 and PC2 explain only a small amount of the variance in OTUs observed across different disease classes.

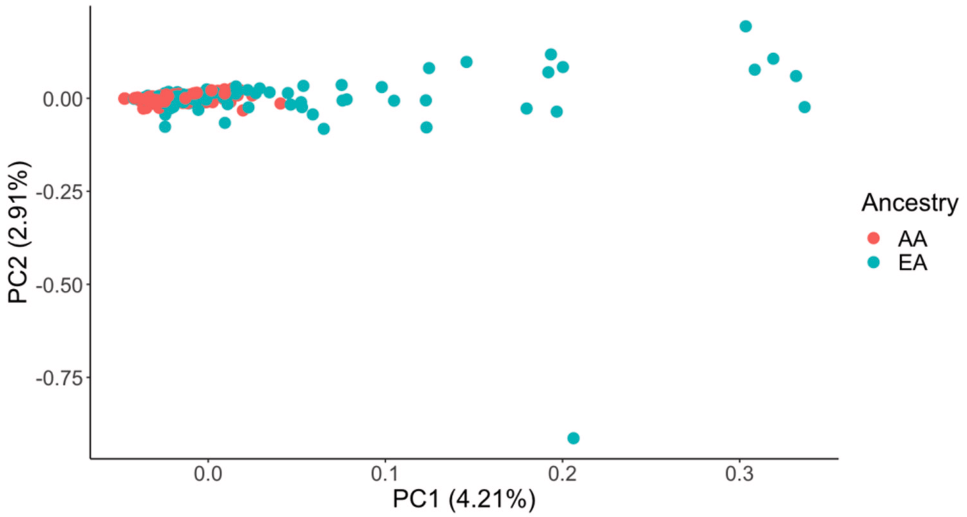


Figure 11. PCA plot explaining variation for ancestry between African Americans (AA) and Europeans (EA). The same number of OTUs and individuals are used as in Figure 10 for different classes. This PCA plot shows more separation in the OTU clusters based on ancestry than by different disease classes (shown in Figure 10).

Table 3. Informative OTUs identified by the feature selection process. These 10 OTUs are all bacteria which come from 2 distinct phyla. Most of the OTUs identified are at genus-level.

Phylum	Class	Order	Family	Genus	Species
Firmicutes	Erysipelotrichia	Erysipelotrichales	Erysipelotrichaceae	*Eubacterium*	*Biforme*
Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	*Faecalibacterium*	*Prausnitzii*
Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	*Roseburia*	NA
Firmicutes	Clostridia	Clostridiales	Veillonellaceae	*Phascolarctobacterium*	NA
Firmicutes	Clostridia	Clostridiales	NA	NA	NA
Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	NA	NA
Firmicutes	Clostridia	Clostridiales	Veillonellaceae	*Megasphaera*	NA
Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	NA	NA
Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	*Prevotella*	*Copri*
Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	NA	NA

Since dental calculus is formed as a result of bacterial plaque accumulation around the gingiva, dental calculus consists primarily of bacteria. Thus, dental calculus lends itself well to oral microbiome analysis. For example, it was found in a medieval sample that 85–95% of the calculus was composed of bacterial proteins[36]. This indicates a novel method of examining the constituents of the oral microbiome and its variation across cultures, geographies, and various historical periods.

The availability of a unique set of data from the first quarantine in the world will enable substantial focus on infectious diseases and the modeling of ancient epidemics (Figure 14). All of the approximately 1500 individuals for this project died of an infectious disease, we know this from archival records. The addition of body responses to the environment and diseases (metabolites), as well as dietary data (stable isotopes to detect malnutrition), will be trialed, providing the best chance to recognize the pathogen responsible and its overall effects. In genetics and medicine, the combination of code, workflow, logic and available data will provide over 300 years of data on epidemics (especially bubonic plague) including the first influenza pandemic, dated 1580, and outbreaks of typhus and measles. It will be possible to reach ca. 600 years of data at one location using historical and medical records. The plague and other similar illnesses provoking fever are replaced by smallpox, measles and flu in later times, as medicine provides therapies, mobility increases

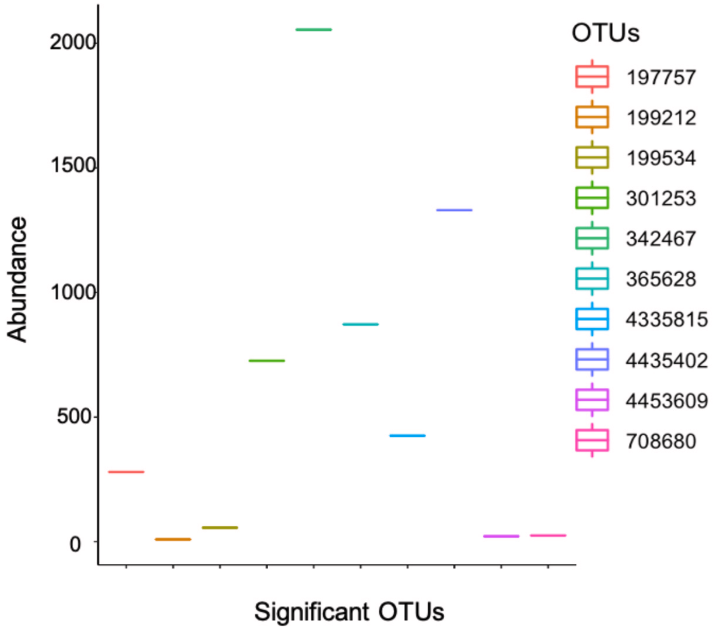


Figure 12. Abundance of significant OTUs selected by machine learning These OTUs are highly predictive for the classification of disease vs. normal class.

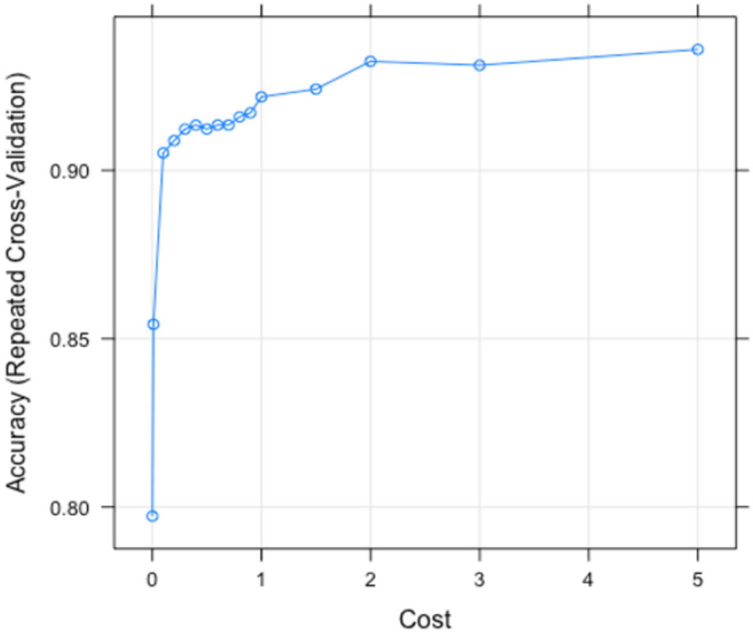


Figure 13. Cross-validation using the support vector regression approach. The model showed the best cross-validation score with cost=5 (accuracy = 0.936).

and diet changes with many plants cultivated in different continents from where they originated. Our TRACK prototypes will enable investigations related to pathogen evolution, microbiome adaptations and human immunity responses changes.

Goals: To achieve the transdisciplinary goals inherent to the nature of this paleo-omics project, a central database able to contain different data types is required. Towards this objective, we created and implemented a paleo-omics workflow consisting

Figure 14. Mask worn by doctors visiting people in quarantine in Venice to protect themselves during the 17th century. Left: Masque porté vers 1630 par les médecins visitant les pestiférés from R. Blanchard, in *Archives de parasitologie*, 1900. Pl. V. **Right**: drawing of a doctor wearing the mask. From Thomas Bartholin, Plague doctor, *Thomæ Bartholini Historiarum anatomicarum rariorum*, Hafniae: Sumptibus P. Hauboldt, 1654, p. 143

of: 1) a search engine to query the multi-data database, 2) a retrieving pipeline for paleo proteins, and 3) a query gateway for microbiome-human host interactions (Figure 15).

While mass spectrometry (MS), shotgun sequencing, and 16S rRNA sequencing data can be employed in paleo-omics, we focused on an MS-based meta-proteomics approach for proof-of-concept demonstration of our prototype within the time constraints of the Codeathon, which we applied to data derived from human dental calculus protein-samples taken from archeological sites.

Methods

Data sources and processing

MS data and shotgun sequencing data obtained from ancient human dental calculus samples were used for these analyses[36,37].

(1) MS data: peptides were identified from raw data files by comparing spectra from the second spectrometer of a tandem-MS (MS2) to reference spectra available in protein databases. Many existing proteomics software packages, such as MaxQuant, have been designed for analyzing large MS data sets, such as the MaxQB database, and thus can perform this task[38].

(2) Shotgun sequencing data: the resulting short reads in FASTQ data format have been initially verified if they correspond to human DNA sequences, sequence reads were aligned to a human reference genome (Genome Reference Consortium Human Build 38) to verify human sequences using the Bowtie version 1.3.0 and BWA programs version 0.7.17[39,40]. Reads not aligning to the human reference genome were characterized as non-human.

All processed data were stored in a high-performance database for future analysis. A web user interface and a search/analysis engine[41] were developed to access these data.

Assessing presence of select pathogens

We performed targeted pathogen searches for sequences of oral pathogenic microbes and other human pathogens, including the major human malaria parasite *Plasmodium falciparum*. We identified pathogenic oral microbes similar to previously published results, but no significant hits to *P. falciparum* from these two test-sets were identified. We additionally searched for marker oral microbiome species for other human infectious diseases as reported in detail in the results section.

Operation and Implementation

Source-code for our prototype is available through our GitHub repository (see *Software availability* section). This implementation requires the following software packages to reproduce: Python version 3.6.0; Flask version 1.1; R version 3.4.4; Perl version 5.26.1; BLAST version 2.10.0.

Results

To test our prototype[41], we searched for pathogen sequences against the two archaeological samples in the database, one from Denmark 1100-1450 AD[36] and one from the United Kingdom 1770-1855 AD[34]. The medieval Danish samples were used with a complete set of dental pathology characterization and individual

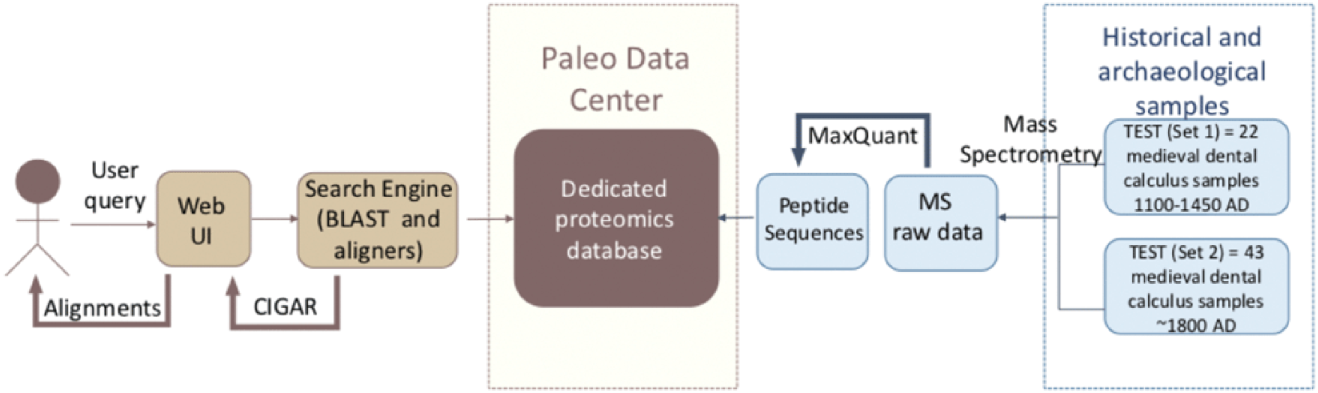

User query
Web UI
Search Engine (BLAST and aligners)
Alignments
CIGAR
Paleo Data Center
Dedicated proteomics database
MaxQuant
Peptide Sequences
MS raw data
Mass Spectrometry
Historical and archaeological samples
TEST (Set 1) = 22 medieval dental calculus samples 1100-1450 AD
TEST (Set 2) = 43 medieval dental calculus samples ~1800 AD

Figure 15. Prototype paleo-data center workflow. Data derived from laboratory-based analyses of biopaleological samples are processed and analyzed by established analytical software. Results from these analyses are then compared to existing databases, such as RefSeq, and both the known and unknown information are stored in a centralized Paleo-pathology database. A search engine and a web user interface (UI) then provides users access to this centralized Paleo-pathology database. The dedicated proteomics database can be expanded and rebuilt by data scientists with new data sets and novel data structures. Abbreviations: *BLAST* (Basic Local Alignment Tool): a popular algorithm for comparing biological sequence information, such as the amino-acid sequences of proteins or the nucleotides of DNA RNA sequences[42]. *CIGAR* (Concise Idiosyncratic Gapped Alignment Report): a string format used to represent information such as which bases align (either a match/mismatch) with the reference, are deleted from the reference, and are insertions that are not in the reference. *MaxQuant*: a quantitative proteomics software package designed for analyzing large mass-spectrometric data sets.

data. Consistent with the reported results[36], there are oral disease pathology and bacteria normally found in the oral microbiome that can be recovered (Figure 16). For instance, the species *Porphyromonas gingivalis* is frequently present in individuals with orthodontic diseases, while *Streptococcus sanguinis* is present in both medieval and contemporary individuals with satisfactory oral health.

This approach can also be used to discover other bacteria linked to health and possibly reveal other correlations between microbiome bacteria and health status as well as recent evolutionary changes. In archaeology, the current focus is on revealing specific pathogens and there is no established reference material to investigate the past microbiome or its effects on health. Even in recent studies, any conclusions on medieval or older individuals is based on direct comparison with the contemporary microbiome. By using archaeological methods (chronological seriation) together with software developed from our code, it will be possible to investigate any correlation between microbiome and health searching individuals dating to older periods. Such work could provide a reference standard for archaeologists, and evolutionary data to health professionals. For example, using the existing data, we found the opportunistic respiratory pathogen *Haemophilus parainfluenzae*[43,44] present less frequently in this set of medieval samples (Wilcoxen test, $p < 0.05$), raising interesting questions about human society transition and infectious diseases. This group appears in Neolithic agrarian human oral microbiomes (7440 BCE)[45], but is at low levels in human groups practicing hunting and gathering (2000 BCE, modern day South Africa). Questions of interest to both health professionals and archaeologists that could be answered by employing our code may be when this pathogen became more frequent and why.

Understanding the origins and evolution of pathogens is very important to prepare for future pandemics. The only successful work attempted on combining archaeology with genetics and health studies to investigate past pathogens, the reconstruction of the 1918 flu pathogen[46] proved to be both technically challenging and costly even though fewer than a hundred years had passed since the pandemic because that work tried to reproduce an active virus now extinct. It was also very useful to demonstrate that the strong virulence reported in historical sources, but unconfirmed in medicine, was real. Since 1919, only COVID-19 has demonstrated a similar virulence, proving that data from historical record can be critical in addressing new types of known viruses and pathogens, which can regain traits unseen for a century or more within that category of pathogens (respiratory viruses with flu-like symptoms in this case). That work has shown also how the choice of suitable burial grounds is essential for such work. Our work uses new -omics analyses that are providing new sources of data and could prove equally valuable, revealing the history of recent pathogens, characteristics that may have been present only occasionally, and their successes and failures. Future pathogens might reuse and re-combine successful traits (symptoms, virulence) from past epidemics and therefore our preparedness depends on knowing what to expect, on learning from the past.

The results of our work are therefore limited to making possible future interdisciplinary research and open up a path to answer new questions. Sequencing proteomic and metabolomic data from pre-modern individuals is still rare and there is no existing database, besides data from a few academic papers, that our software code could search. Yet, making possible new studies through a working proof-of-concept will accelerate the production of databases for ancient individuals. Existing

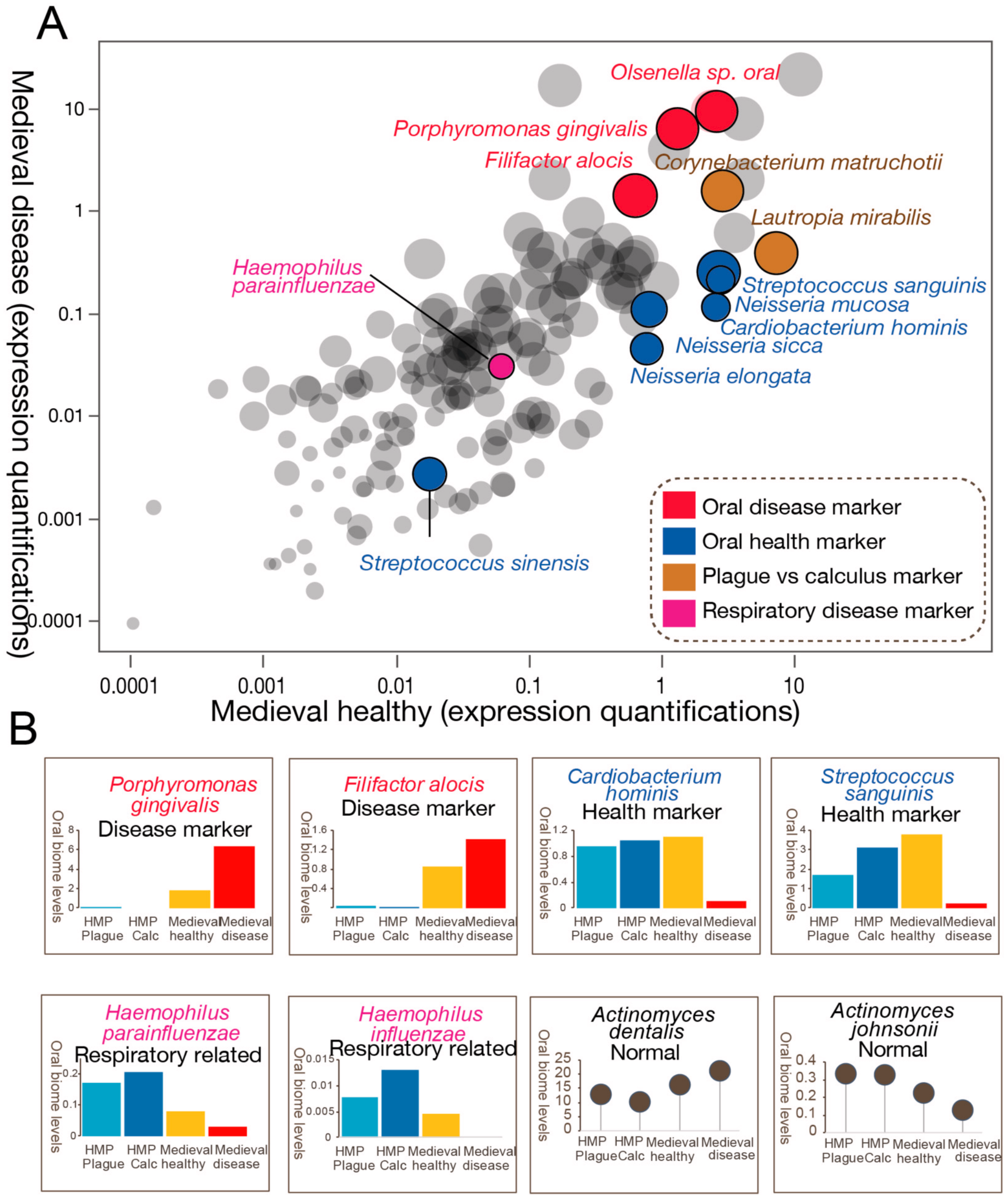


Figure 16. Medieval oral microbiome with bacterial species as markers for oral diseases. A. A total of over 200 bacterial species have been recovered from a metaproteomics study using medieval dental calculus[36]. The Label- free protein quantitation (LFQ) was used to quantify all samples and conduct comparative analysis. The taxa abundance levels were normalized on a scale from 0 to 10; and the circle sizes indicate the frequency of taxa occurrences in the study **B**. Representative species of oral diseases (e.g. *Porphyromonas gingivalis* and *Filifactor alocis*), oral health (*Cardiobacterium hominis* and *Streptococcus sanguinis*), and potential respiratory disease markers (*Haemophilus spp.*)[43,44]. Modern day oral microbiomes from dental plagues and calculus are from the HMP database.

archaeological studies have borne out of early full sequencing of genomes and have been severely limited by such approaches. The benefits deriving from new -omics analyses combined with our approach can provide valuable information on older pathogens. Future work may focus on epidemics initially, but with a potential also for revealing and understanding more subtle and complex relationships between human microbiome and health.

Team 6 - Animal

Project title: Capturing ecological and host drivers of microbiomes

Project Rationales, Descriptions and Goals

Rationale: One primary goal of host-microbiome studies is to capture and understand ecological and host drivers of microbial diversity. Research on host-microbiome associations across host species has been facilitated by the increasing accessibility of high-throughput sequencing techniques and the availability of integrated microbiome datasets, such as the Earth Microbiome Project dataset[47]. These have yielded useful insights on how host-microbiome associations are impacted by host diet[48], host taxonomy or phylogeny[49], host immune system[50], and environmental factors[51]. However, host species traits vary immensely across species and such diversity has been under sampled in microbiome studies. As a result, the effects of other host factors, including body mass and life history, in relation to previously characterized host and environmental effects, on host-microbiome associations have been understudied.

Goal: In this project, we aim to investigate the effects of various host traits, including diet, host taxonomy, body mass, and longevity, in relation to environmental factors, on the intestine, fecal, foregut, and stomach microbiomes of Metazoan (animal) species. We first mined available microbiome and metadata datasets, then applied unsupervised learning directly on rarefied OTU abundance data to uncover clusters of microbial community similarity among animals.

Methods

Data sources and processing

Rarefied OTU table (1000 reads per sample) and metadata of internal animal microbiomes from the Earth Microbiome Project[47] was obtained from Woodhams *et al.*[52]. The OTU table was filtered to remove plant samples (Kingdom Plantae), OTUs with <10 total counts across samples, and OTUs occurring in <2 samples.

Metadata collection

For each sequenced species in our dataset, we added metadata for body mass and maximum longevity, if available. Body mass data was collected from the Pantheria archives[53], the Caviede Vidal dataset[54], and the Encyclopedia of Life. Body mass data was categorized to create three equally sized groups (excluding *Homo Sapiens*): big (> 58.7 kg), medium (>19.57 kg, ≤ 58.7 kg), and small (≤ 19.57 kg). Maximum longevity data was obtained from AnAge[55].

Unsupervised learning analysis

To explore distinct microbial composition structures across samples, an unsupervised cluster analysis was performed on the processed OTU table. OTUs present in less than 5% samples were discarded to obtain robust clusters. Sample-wise distance matrix was then computed using Jensen-Shannon distance on the OTU table of relative abundance. The PAM (partition around medoids) clustering analysis was completed using the cluster version 2.1.0 package in R software version 3.6.1[56]. The optimal number of clusters was determined to maximize the Silhouette coefficient[57]. To visualize results of the cluster analysis, principal component analysis was completed using ade4 version 1.7-13 package in R software. Individual samples were depicted on the space of top two principal components.

ANOVA F-test and correlation analysis

For feature selection, ANOVA F-tests were implemented in python to identify quantitative metadata variables with significant means variance differences between clusters. Pearson correlation analysis was also performed in python to evaluate linear relationships between metadata variables.

Operation and implementation

The analyses can be performed on a local computer or server with R and Python installed. A step-by-step tutorial of the unsupervised clustering approach is available at https://enterotype.embl.de/enterotypes.html. R markdown and Python codes used for analyses are also available as listed in the *Software availability* section[58].

Results

We analyzed 726 samples spanning 199 terrestrial and freshwater Metazoan species within seven classes (Figure 17). Our unsupervised learning approach generated three sample clusters (Figure 18). The largest and most diverse cluster (cluster 1) comprised ~92% of all samples (n=667) from 21 Metazoan orders. These included lepidoptera (butterflies and moths; n=165), primates (n=85), anura (n=79), chiroptera (bats; n=44), carnivora (n=41), passeriformes (perching birds; n=37), hymenoptera (n=27), artiodactyla (n=26), diprotodontia (n=24), rodentia (n=23), lagomorpha (n=19), columbiformes (n=18), cypriniformes (n=18), squamata (n=17), anseriformes (n=9), gasterosteiformes (n=9), coleoptera (n=7), pilosa (n=7), cingulata (n=6), casuariiformes (n=5), and hemiptera (n=1). Cluster 2 comprised 34 samples from bats (n=16), butterflies and moths (n=10), perching birds (n=6), the dung beetle *Teuchestes fossor* (n=1), and the giant anteater *Myrmecophaga tridactyla* (n=1). Cluster 3 was the smallest (n=25) and exclusively comprised butterfly and moth samples belonging to seven species. These included *Maculinea alcon* (n=9), *Durbania amakosa* (n=5), *Spalgis epeus* (n=5), *Lycaena clarki* (n=2), *Surendra vivarna* (n=2), *Anthene usamba* (n=1), and *Rapla iarbus* (n=1).

ANOVA analysis indicated that clusters had the most significant mean differences in microbial alpha diversity, Simpson diversity, Shannon diversity, Faith's phylogenetic diversity, and Chao 1 diversity (Table 4). Digestive habitat type, host taxonomy/phylogeny, immune complexity, and life stage, were also significantly different between clusters, along with DNA extraction methods and environmental variables. Notably, body mass and maximum longevity were also significantly different between clusters.

Cluster-specific correlation analyses showed that alpha diversity in clusters 1 and 2 was consistently positively correlated with host taxonomy, immune complexity, diet, maximum longevity and latitude. Body mass, vegetation index, terrain complexity, mean temperature of the driest quarter and precipitation of the warmest and coldest quarters showed positive correlations with alpha diversity in cluster 1, but not cluster 2. Latitude and country were positively correlated with alpha diversity in cluster 2, but not

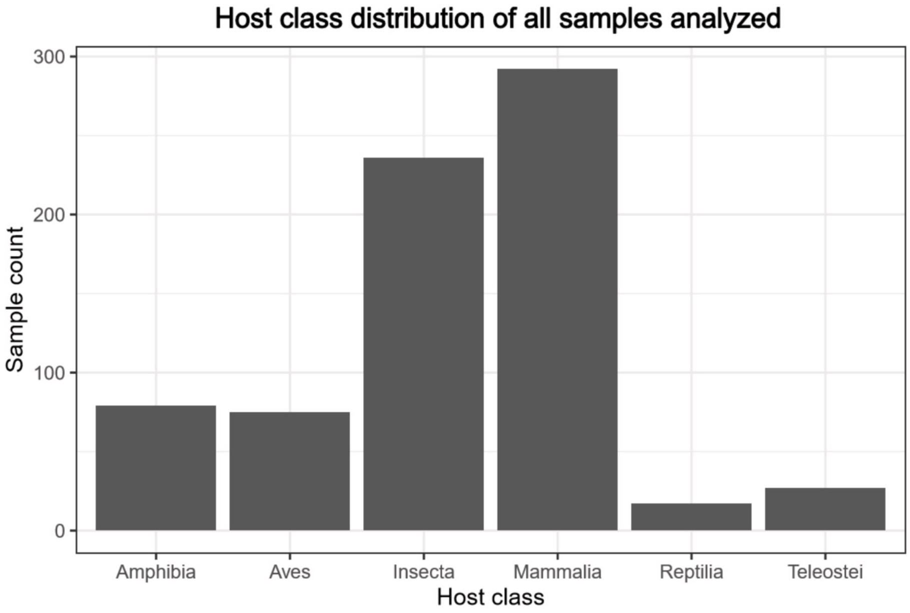


Figure 17. **Number of samples (y-axis) analyzed for each host class (x-axis) in this study.**

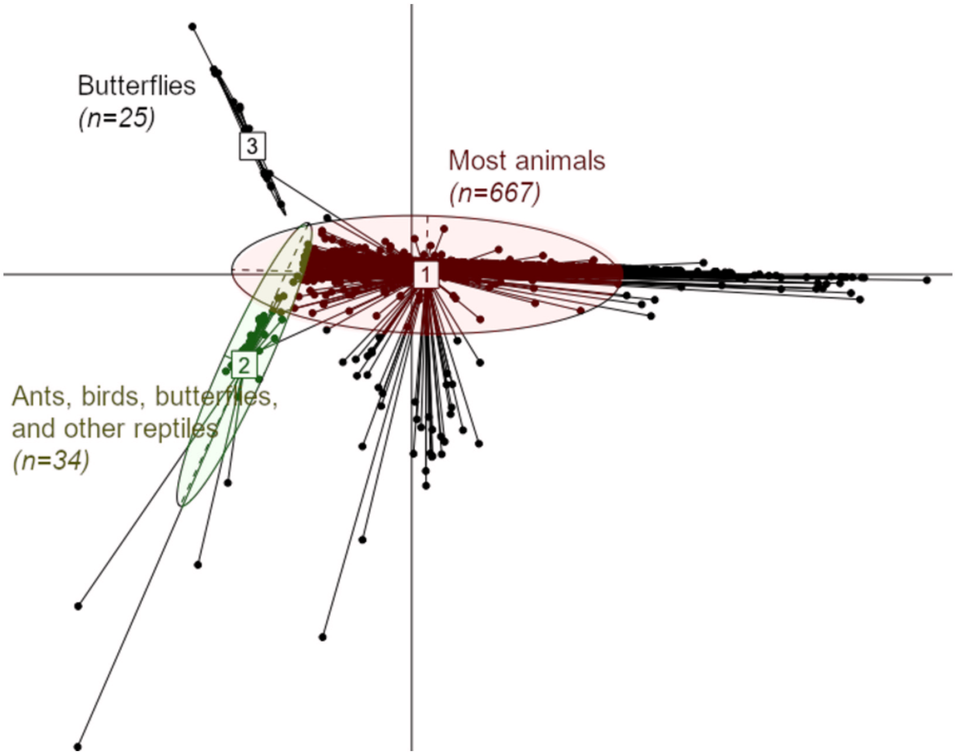


Figure 18. **Principal component analysis (PCA) plot showing the three animal clusters.** The data clusters were generated by the Partitioning Around Medoids (PAM) clustering algorithm on Jensen-Shannon divergence calculated from OTU relative abundances. Each point on the plot represents a sample, and each cluster was labelled with its general taxonomic composition and sample sizes.

cluster 1. Alpha diversity in cluster 3, which comprised butterflies and moths, was positively correlated with environmental variables (terrain complexity, mean diurnal temperature range, precipitation seasonality, elevation) and host factors (digestive habitat type and diet).

The results support our premise that host traits, including but not limited to body mass and maximum longevity, are under sampled in microbial diversity studies. Understudied host traits could also shape animal internal microbiomes together with well-characterized host traits and environmental variables. Based on

Table 4. **PERMANOVA F scores and p-values of metadata variables significantly associated (p<0.05) with cluster groupings.**

Metadata Variable	F Score	p value
Simpson diversity	135.93	7.71E-51
Shannon diversity	85.60	4.30E-34
Digestive habitat type (intestine,fecal,foregut,stomach)	43.29	1.75E-18
Faith's phylogenetic diversity	33.94	8.14E-15
Host phylum	28.96	7.99E-13
Host phylogeny (nDMS proxy)	28.95	8.03E-13
Immune complexity (ordinal score)	27.31	3.67E-12
Observed OTUs	26.29	9.49E-12
Lifestage (larvae,juvenile/pupae, infant, adult)	20.40	2.40E-09
Chao1 diversity	20.26	2.76E-09
Preservation method (ethanol, freezing, RNAlater, others)	17.91	2.55E-08
Maximum temperature of the warmest month	17.38	4.25E-08
Host family	14.50	6.68E-07
Longitude	11.64	1.06E-05
Body mass	10.50	3.19E-05
Mean diurnal temperature range	10.46	3.33E-05
Surrounding habitat (freshwater, terrestrial)	5.55	0.004051
Host order	5.24	0.005518
Mean temperature of the driest quarter	5.08	0.006419
Maximum longevity	3.73	0.024552
Precipitation seasonality	3.39	0.034152
DNA extraction method (DNeasy Powersoil, EZna Stool Dna Kit, PowerFecal, QIAamp DNA Stool Mini Kit, ZR Fecal DNA Miniprep Kit)	3.16	0.042965
Vegetation index (NDVI MODIS)	3.14	0.043888

our results, we propose comprehensive sampling of host traits in future microbiome studies, which may yield new and unexpected patterns of microbial community organization serving as a baseline for deeper investigations.

Lessons learned

Throughout this process we identified several areas where improvements could be made for future disease-focused hackathons. A few of these are described below.

Collaboration across domains requires extensive communications with minimum use of jargons, and active learning from diverse backgrounds. We aimed to further expand on the traditional foundation of codeathons, and we generated novel tools by leveraging research strengths of the local community. However, there has been some challenges in the six teams to efficiently work together, with barriers in communicating the feasibility and significance of particular problems. In-depth and succinct explanation of the technical problems are critical for the successful operations.

Scalability of R has been called into question during the prototype development. For large dataset computations, more efficient implementation can be developed once the prototype has proven to be useful for the community. However, the granularity of solutions available in R make it the preferred tool for designing and experimenting with different solutions.

Meticulous documentation of each analysis step remains crucial for effective dissemination of our approach and results. These necessary components of any project are also excellent opportunities to apply the skillsets of non-coders, as well as to heighten engagement of trainees by reinforcing project rationale. Good documentation, including simple flowcharts, are very useful tools for keeping focus. Non-coding participants who want to gain some experience can often quickly learn markdown language and be vital contributors to repositories.

Conclusion and next steps

Interdisciplinary collaborations have proven to be very productive as shown by our six working prototypes addressing broad microbiome related challenges, ranging from power calculations, AI classifiers, GIS integration and large data set visualizations. Although working across fields has been a challenging task, we found that parsing a complex question into distinct parts can help different domain-experts to work together and accomplish tasks none of the individuals could accomplish in isolation. The codeathon workflow is thus a useful research model for many urgent societal problems that suffer from knowledge-transfer and communication issues. We have made all data and code publicly available for further exploration of these tools. Importantly, we are developing impactful projects to further pursue intersectional research spurred by this event, including microbiome-related machine learning, and data mining across archaeological time and geography.

Data availability

All data underlying the results are available as part of the article and no additional source data are required.

Software availability

Team 1

Source code available from: https://github.com/USFOneHealth-Codeathon2020/Team1_MicroPowerPlus.

Archived source code at time of publication: https://doi.org/10.5281/zenodo.4031770[8].

License: GNU General Public License 3.0.

Team 2

Source code available from: https://github.com/USFOneHealth-Codeathon2020/Team2_GEO

Archived source code at time of publication: https://doi.org/10.5281/zenodo.4034466[20].

License: GNU General Public License 3.0.

Team 3

Source code available from: https://github.com/USFOneHealth-Codeathon2020/projectZer0.

Archived source code at time of publication: https://doi.org/10.5281/zenodo.4031780[26]

License: GNU General Public License 3.0.

Team 4

Source code available from: https://github.com/USFOneHealth-Codeathon2020/Team-YOLO

Archived source code at time of publication: https://doi.org/10.5281/zenodo.4031776[28]

License: GNU General Public License 3.0.

Team 5

Source code available from: https://github.com/USFOneHealth-Codeathon2020/Team5_MinhRays

Archived source code at time of publication: https://doi.org/10.5281/zenodo.4031785[41].

License: GNU General Public License 3.0.

Team 6

Source code available from: https://github.com/USFOneHealth-Codeathon2020/Team6_LimSharma

Archived source code at time of publication: https://doi.org/10.5281/zenodo.4031778[58].

License: GNU General Public License 3.0.

Acknowledgements

We thank Paige Hunt for coordinating and helping organizing the event. We also thank John Adams, Derek Wildman, Monica Uddin from USF genomics program for their support. We thank Ming Ji, Derek Wildman, Christopher Collazo for forming the evaluation panel. We acknowledge the support from the USF Genomics Program, USF Omics Hub, USF Library, and Institute for the Advanced Study of Culture and the Environment for extensive logistical help in organizing the OneHealth Host-Microbiome Interactions Codeathon.

Author contributions

Contributor Role	Role Definition
Conceptualization	RHYJ, JO, JG, CW, TEK, AS, SA contributed to forming Ideas and formulation or evolution of overarching research goals and aims.
Data Curation	RHYJ, JO, JG, SA, TEK, GD, AS, MP, AW, CL, JL contributed to the management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later reuse.
Formal Analysis	All authors participated in application of statistical, mathematical, computational, or other formal techniques to analyze or synthesize study data.
Funding Acquisition	RHYJ contributed to acquisition of the financial support for the project leading to this publication.
Investigation	All authors contributed to conducting a research and investigation process, specifically performing the experiments, or data/evidence collection.
Methodology	All authors contributed to development or design of methodology; creation of models.
Project Administration	RHYJ, JO, JG, SA, TEK, GD, AS, MP, AW, CL,JL contributed to management and coordination responsibility for the research activity planning and execution.
Resources	RHYJ, JO, JG, SA, TEK, GD, AS, MP, AW, CL,JL contributed to provision of study materials, and computing resources, or other analysis tools.
Software	All authors contributed to programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components.
Supervision	RHYJ, JO, JG, SA, TEK, GD, AS, MP, AW, CL,JL provided oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team.

Contributor Role	Role Definition
Validation	All authors contributed to the verification, whether as a part of the activity or separate, of the overall replication/reproducibility of results/experiments and other research outputs.
Visualization	All authors contributed to the preparation, creation and/or presentation of the published work, specifically visualization/data presentation.
Writing – Original Draft Preparation	All authors contributed to the creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation).
Writing – Review & Editing	All authors contributed to the reparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision – including pre- or post-publication stages.

References

1. Ferreira GC, Oberstaller J, Fonseca R, *et al.*: **Iron Hack - A symposium/hackathon focused on porphyrias, Friedreich's ataxia, and other rare iron-related diseases [version 1; peer review: 2 approved].** *F1000Res.* 2019; **8**: 1135.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Debelius J, Song SJ, Vazquez-Baeza Y, *et al.*: **Tiny microbes, enormous impacts: what matters in gut microbiome studies?** *Genome Biol.* 2016; **17**(1): 217.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Debelius JW, Vázquez-Baeza Y, McDonald D, *et al.*: **Turning Participatory Microbiome Research into Usable Data: Lessons from the American Gut Project.** *J Microbiol Biol Educ.* 2016; **17**(1): 46–50.
PubMed Abstract | Publisher Full Text | Free Full Text
4. NIH Human Microbiome Portfolio Analysis Team: **A review of 10 years of human microbiome research activities at the US National Institutes of Health, Fiscal Years 2007-2016.** *Microbiome.* 2019; **7**(1): 31.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Kelly BJ, Gross R, Bittinger K, *et al.*: **Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA.** *Bioinformatics.* 2015; **31**(15): 2461–8.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Human Microbiome Project Consortium: **A framework for human microbiome research.** *Nature.* 2012; **486**(7402): 215–21.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Oksanen J, Blanchet FG, Friendly M, *et al.*: **vegan: Community Ecology Package. R package version 2.5-6**. 2019.
Reference Source
8. Oberstaller J, DokurOmkar V, Gibbons J, *et al.*: **USFOneHealthCodeathon2020/Team1_MicroPowerPlus: v1.0.0 (Version v1.0.0).** *Zenodo.* 2020.
http://www.doi.org/10.5281/zenodo.4031770
9. RStudio Team: **RStudio: Integrated Development Environment for R**. RStudio, PBC: Boston, MA. 2020.
10. Chang W, Cheng J, Allaire JJ, *et al.*: **shiny: Web Application Framework for R**. R package version 1.5.0. 2020.
Reference Source
11. Inc., P.T: **Collaborative data science**. 2015.
12. Wickham H, Averick M, Bryan J, *et al.*: **Welcome to the tidyverse.** *J Open Source Softw.* 2019; **4**(43): 1686.
Publisher Full Text
13. Laue HE, Brennan KJM, Gillet V, *et al.*: **Associations of prenatal exposure to polybrominated diphenyl ethers and polychlorinated biphenyls with long-term gut microbiome structure: a pilot study.** *Environ Epidemiol.* 2019; **3**(1): e039.
PubMed Abstract | Publisher Full Text | Free Full Text
14. U.S. Environmental Protection Agency: **What is Superfund**? 2018.
Reference Source
15. Friis RH: **Essentials of Environmental Health**. Jones & Bartlett Learning. 2019.
Reference Source
16. U.S. Environmental Protection Agency: **Hazard Ranking System Guidance Manual**. 2020.
Reference Source
17. Jin Y, Wu S, Zeng Z, *et al.*: **Effects of environmental pollutants on gut microbiota.** *Environ Pollut.* 2017; **222**: 1–9.
PubMed Abstract | Publisher Full Text
18. Capellini FM, Vencia W, Amadori M, *et al.*: **Characterization of MDCK cells and evaluation of their ability to respond to infectious and non-infectious stressors.** *Cytotechnology.* 2020; **72**(1): 97–109.
PubMed Abstract | Publisher Full Text | Free Full Text
19. McMurdie PJ, Holmes S: **phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data.** *PLoS One.* 2013; **8**(4): e61217.
PubMed Abstract | Publisher Full Text | Free Full Text
20. CancerGenetics007, Keller T, dahrendorff A, Oberstaller J: **USFOneHealthCodeathon2020/Team2_GEO: v1.0.0 (Version v1.0.0).** *Zenodo.* 2020.
http://www.doi.org/10.5281/zenodo.4034466
21. Wright MN, Ziegler A: **ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R.** *J Stat Softw.* 2017; **77**(1).
Publisher Full Text
22. Li H: **Seqtk: a fast and lightweight tool for processing FASTA or FASTQ sequences**. 2013.
Reference Source
23. Callahan BJ, McMurdie PJ, Rosen MJ, *et al.*: **DADA2: High-resolution sample inference from Illumina amplicon data.** *Nat Methods.* 2016; **13**(7): 581–3.
PubMed Abstract | Publisher Full Text | Free Full Text
24. Quast C, Pruesse E, Yilmaz P, *et al.*: **The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.** *Nucleic Acids Res.* 2013; **41**(Database issue): D590–6.
PubMed Abstract | Publisher Full Text | Free Full Text
25. Abadi M, *et al.*: **TensorFlow: Large-scale machine learning on heterogeneous systems**. 2015.
26. Anujit-sarkar W, Oberstaller J: **USFOneHealthCodeathon2020/projectZer0: v1.0.0 (Version v1.0.0).** *Zenodo.* 2020.
http://www.doi.org/10.5281/zenodo.4031780
27. Mysara M, Vandamme P, Props R, *et al.*: **Reconciliation between operational taxonomic units and species boundaries.** *FEMS Microbiol Ecol.* 2017; **93**(4): fix029.
PubMed Abstract | Publisher Full Text | Free Full Text
28. Agaz NW, Bibber V, Dean A, *et al.*: **USFOneHealthCodeathon2020/Team-YOLO: v1.0.0 (Version v1.0.0).** *Zenodo.* 2020.
http://www.doi.org/10.5281/zenodo.4031776
29. Turnbaugh PJ, Hamady M, Yatsunenko T, *et al.*: **A core gut microbiome in obese and lean twins.** *Nature.* 2009; **457**(7228): 480–4.
PubMed Abstract | Publisher Full Text | Free Full Text
30. Randolph TW, Zhao S, Copeland W, *et al.*: **Kernel-Penalized Regression for Analysis of Microbiome Data.** *Ann Appl Stat.* 2018; **12**(1): 540–566.
PubMed Abstract | Publisher Full Text | Free Full Text
31. Zhou YH, Gallins P: **A Review and Tutorial of Machine Learning Methods for Microbiome Host Trait Prediction.** *Front Genet.* 2019; **10**: 579.
PubMed Abstract | Publisher Full Text | Free Full Text
32. Tsagris M, Tsamardinos I: **Feature selection with the R package *MXM* [version 2; peer review: 2 approved].** *F1000Res.* 2018; **7**: 1505.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Kuhn M: **Building Predictive Models in R Using the caret Package.** *J Stat Softw.* 2008; **28**(5).
Publisher Full Text

34. Hendy J, Warinner C, Bouwman A, *et al.*: **Proteomic evidence of dietary sources in ancient dental calculus.** *Proc Biol Sci.* 2018; **285**(1883).
Publisher Full Text

35. Hendy J, Welker F, Demarchi B, *et al.*: **A guide to ancient protein studies.** *Nat Ecol Evol.* 2018; **2**(5): 791–799.
PubMed Abstract | **Publisher Full Text**

36. Jersie-Christensen RR, Lanigan LT, Lyon D, *et al.*: **Quantitative metaproteomics of medieval dental calculus reveals individual oral health status.** *Nat Commun.* 2018; **9**(1): 4744.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

37. Velsko IM, Yates JAF, Aron F, *et al.*: **Microbial differences between dental plaque and historic dental calculus are related to oral biofilm maturation stage.** *Microbiome.* 2019; **7**(1): 102.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

38. Tyanova S, Temu T, Cox J: **The MaxQuant computational platform for mass spectrometry-based shotgun proteomics.** *Nat Protoc.* 2016; **11**(12): 2301–2319.
PubMed Abstract | **Publisher Full Text**

39. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods.* 2012; **9**(4): 357–9.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

40. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform.** *Bioinformatics.* 2010; **26**(5): 589–95.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

41. Pham M, Rays Jiang A, Nguyen D, *et al.*: **USFOneHealthCodeathon2020/Team5_MinhRays: v1.0.0 (Version v1.0.0).** *Zenodo.* 2020.
http://www.doi.org/10.5281/zenodo.4031785

42. Altschul SF, Gish W, Miller W, *et al.*: **Basic local alignment search tool.** *J Mol Biol.* 1990; **215**(3): 403–10.
PubMed Abstract | **Publisher Full Text**

43. Hofstra JJ, Matamoros S, van de Pol MA, *et al.*: **Changes in microbiota during experimental human Rhinovirus infection.** *BMC Infect Dis.* 2015; **15**: 336.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

44. Kosikowska U, Biernasiuk A, Rybojad P, *et al.*: **Haemophilus parainfluenzae as a marker of the upper respiratory tract microbiota changes under the influence of preoperative prophylaxis with or without postoperative treatment in patients with lung cancer.** *BMC Microbiol.* 2016; **16**: 62.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

45. Adler CJ, Dobney K, Weyrich LS, *et al.*: **Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions.** *Nat Genet.* 2013; **45**(4): 450–5, 455e1.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

46. Jordan D: **The Deadliest Flu: The Complete Story of the Discovery and Reconstruction of the 1918 Pandemic Virus.** 2019.
Reference Source

47. Thompson LR, Sanders JG, McDonald D, *et al.*: **A communal catalogue reveals Earth's multiscale microbial diversity.** *Nature.* 2017; **551**(7681): 457–463.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

48. Ley RE, Hamady M, Lozupone C, *et al.*: **Evolution of mammals and their gut microbes.** *Science.* 2008; **320**(5883): 1647–51.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

49. Lim SJ, Bordenstein SR: **An introduction to phylosymbiosis.** *Proc Biol Sci.* 2020; **287**(1922): 20192900.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

50. Franzenburg S, Walter J, Künzel S, *et al.*: **Distinct antimicrobial peptide expression determines host species-specific bacterial associations.** *Proc Natl Acad Sci U S A.* 2013; **110**(39): E3730–8.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

51. Kohl KD, Brun A, Magallanes M, *et al.*: **Gut microbial ecology of lizards: insights into diversity in the wild, effects of captivity, variation across gut regions and transmission.** *Mol Ecol.* 2017; **26**(4): 1175–1189.
PubMed Abstract | **Publisher Full Text**

52. Woodhams DC, Bletz MC, Becker CG, *et al.*: **Host-associated microbiomes are predicted by immune system complexity and climate.** *Genome Biol.* 2020; **21**(1): 23.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

53. Collen B, McRae L, Deinet S, *et al.*: **Predicting how populations decline to extinction.** *Philos Trans R Soc Lond B Biol Sci.* 2011; **366**(1577): 2577–86.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

54. Caviedes-Vidal E, McWhorter TJ, Lavin SR, *et al.*: **The digestive adaptation of flying vertebrates: high intestinal paracellular absorption compensates for smaller guts.** *Proc Natl Acad Sci U S A.* 2007; **104**(48): 19132–7.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

55. Tacutu R, Thornton D, Johnson E, *et al.*: **Human Ageing Genomic Resources: new and updated databases.** *Nucleic Acids Res.* 2018; **46**(D1): D1083–D1090.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

56. Arumugam M, Raes J, Pelletier E, *et al.*: **Enterotypes of the human gut microbiome.** *Nature.* 2011; **473**(7346): 174–80.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

57. Lovmar L, Ahlford A, Jonsson M, *et al.*: **Silhouette scores for assessment of SNP genotype clusters.** *BMC Genomics.* 2005; **6**: 35.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

58. Swadtasnim SO, Sumpter M, Kim Y, *et al.*: **USFOneHealthCodeathon2020/Team6_LimSharma: v1.0.0 (Version v1.0.0).** *Zenodo.* 2020.
http://www.doi.org/10.5281/zenodo.4031778