Contamination detection and microbiome exploration with GRIMER

Vitor C. Piro vitor.cedranpiro@hpi.de[1] and Bernhard Y. Renard bernhard.renard@hpi.de[1]

[1]Data Analytics and Computational Statistics, Hasso Plattner Insititute, Digital Engineering Faculty, University of Potsdam, 14482 Potsdam, Germany

Abstract

Exploring microbiome data is a time-consuming task that can be only partially automated due to the specific requirements and goals of each project. Visualizations and analysis platforms are crucial to better guide this step. Best practices in the field are constantly evolving and many pitfalls can lead to biased outcomes. Compositionality of data and sample contamination are two important points that should be carefully considered in early stages of microbiome studies. Detecting contamination can be a challenging task, especially in low-biomass samples or in studies lacking proper controls by design. However, external evidences and commonly identified contaminant taxa can be used to discover and mitigate contamination. We propose GRIMER, a tool that automates analysis, generates plots and runs external tools to create a portable dashboard integrating annotation, taxonomy and metadata. It unifies several sources of evidence towards contamination detection. GRIMER is independent of quantification methods and directly analyses contingency tables to create an interactive and offline report. GRIMER reports can be created in seconds and are accessible for non-specialists, providing an intuitive set of charts to explore data distribution among observations and samples and its connections with external sources. Further, we compiled an extensive list of common contaminants and possible external contaminant taxa reported in the literature and use it to annotate data. GRIMER is open-source and available at: `https://gitlab.com/dacs-hpi/grimer`.

Introduction

Microbiome studies enable, via high-throughput sequencing, the investigation of the composition of complex microbial communities from diverse environments. Those studies usually yield large amounts of raw DNA sequences for several samples that can be analyzed with an increasing number of computational methods and databases. Standards and best practices for designing and performing a microbiome study are improving and changing over the years [1, 2] and the field is in constant evolution due to higher availability and reduced costs of sequencing runs as well as with the increase in number of publicly available reference sequences and computational methods.

In early stages of a standard microbiome analysis, raw or quality-filtered DNA sequences are classified or clustered into specific groups and quantified to generate a "profile" for a given environmental sample. Marker gene, whole metagenome or metatranscriptome analysis have their own set of tools and standards which should be carefully chosen to generate a reliable measurements for each sample in the study [3]. This step can be computationally intensive but reduces the large amount of information into a concise table of measurements.

After measurements are obtained, visualizations and statistical analysis are performed. This step is mostly exploratory and specific to the research question pursued and analysis can hardly be fully automated. Compositionality of data should be should be taken in consideration when working with microbiome [4]. Visualization methods are extremely helpful in this stage to understand the data distribution and its properties and to guide the statistical analysis to follow.

At this stage of a study, contamination detection should be considered. Contamination side-effects have gained attention in recent years due to the controversial detection of a placental microbiome [5, 6, 7]. Contamination is characterized by exogenous DNA in a given sample introduced externally or internally. External contamination can come from diverse sources: DNA extraction kits, laboratory reagents, surfaces and equipment, ultra-pure water, residuals from previous sequencing runs as well as technicians body [2, 8, 9]. Internal contamination can be defined as a undesired exchange of genetic material between samples and it is usually referred as well-to-well contamination, cross-contamination, or sample "bleeding" as well as index switching in multiplexed sequencing libraries [10].

Contamination may affect most sequencing project to some degree, but affects largely environmental low-biomass samples [11]. The composition of an environmental sample is mostly unknown before sequencing,

increasing the complexity of detecting contamination when comparing to a defined isolate and targeted sequencing project. Low-biomass samples (e.g. meconium, blood, human tissues) yield low amounts of DNA to be amplified and sequenced, ideal scenario for an exogenous contaminants to out-compete and dominate the biological signal.

It is important that contamination is discovered at the earliest stage of a study prior to statistical analysis, not to bias measurements and not to further propagate itself into databases [12, 13]. Inclusion of negative and positive control samples is the recommended way to detect and mitigate contamination [2, 11, 14]. Negative controls should be included in the study design for every sample, extraction or amplification batch. Once provided, controls should be carefully analyzed *in-silico* and results obtained should be applied to biological samples in terms of prevalence (e.g. observations in negative controls) but also frequency in relation to DNA concentration.

However, due to the complexity and diverse sources of contamination, detection and mitigation are not a trivial tasks. Several approaches to identify and exclude background contamination in microbial datasets were proposed [15, 16], each one with strengths and weaknesses based on the study design, data type and control availability. Further, many studies do not include or have limited number of control samples due to increase in costs. [14] reported that based on publications from the 2018 issues of Microbiome and the ISME journals, only 30% cited the use of negative controls and only 10% positive controls. Additionally, the detection of re-occurring contaminants in extraction kits and reagents (also called "kitome" [17]) is underexplored, mainly for not being properly cataloged, centralized or automated.

To overcome those challenges we propose GRIMER, a tool that analyzes and generates an offline and interactive dashboard to visualize, explore microbiome studies outcomes with focus on contamination detection. Based on a single table of observations per sample, GRIMER automates data analysis, transformations and plots and generates a set of charts integrating evidence for better understanding of analysis and decision making for contamination detection. We also compiled a list of common contaminants reported on the literature as a source of annotation. GRIMER aims to be effortless once quantification is done, turning measurement tables into a interactive and dynamic report in seconds. GRIMER is open-source and available at: `https://github.com/pirovc/grimer`. The tool is independent of analysis methods, do not rely on web or local servers and generates standalone and shareable interactive dashboards.

Methods

GRIMER analyzes and annotates multi-sample studies based on count tables and generates a report with several interactive plots to better explore the data and to facilitate contamination detection. GRIMER integrates several sources, references, analysis and external tools and brings them together in one concise dashboard.

The output of GRIMER is a self-contained HTML file that can be visualized in any modern web-browser. It works independently from any actively running server or web-service. Once generated, it can be used and shared as an offline document. It has the advantages of a static report and a complex dashboard being portable and interactive. This feature makes it very convenient to distribute (e.g. as an e-mail attachment), keep track of changes in analytical pipelines and reproduce analysis in different environments.

GRIMER is independent of any quantification method and requires a contingency table with raw counts of observations/components for each samples/compositions in the study. Observations are usually, but not limited to, taxonomic entries (e.g. genus, species, strains), OTUs, ASVs or sequence features. A count of unclassified or unassigned observations is also supported to generate normalized values. Additional files and information can be provided to expand GRIMER reports: study metadata, taxonomy database, multiple control samples, DNA concentration, custom contaminants and reference groups of interest. The more information provided, the more complete and interactive the final report will be.

Annotation

GRIMER annotates observations and samples linking data with external sources.

Sample annotation is based on the study metadata, where each sample is described in one ore more fields and variables. Those fields can contain either numeric or categorical values and are useful for grouping and clustering analysis as well as detection of batches and control/treatment effects.

Observation annotation is based on external lists of taxonomic entries, which are divided in two main categories: contamination sources and reference sources. Contamination sources contains entries which are possible contaminants and reference sources are general lists of references of interest to be annotated. Custom sources can be easily provided by the user in a simple list of names of taxonomic identifiers or in a formatted and annotated file (more information can be found in the GRIMER repository). Moreover, observations are annotated with a summary generated from MGnify [18], with counts of occurrences for each observation in thousands of microbiome studies, grouped by biome.

Organism group	Genus	Species	Reference
Bacteria	6	0	1998 Tanner, M.A. et al. [20]
Bacteria	4	0	2003 Grahn, N. et al. [21]
Bacteria	16	0	2006 Barton, H.A. et al. [22]
Bacteria	11	1	2014 Laurence, M. et al. [23]
Bacteria	92	0	2014 Salter, S.J. et al. [8]
Bacteria	7	0	2015 Jervis-Bardy, J. et al. [24]
Bacteria	28	0	2015 Jousselin, E. et al. [25]
Bacteria	23	0	2016 Lauder, A.P. et al. [26]
Bacteria	6	0	2016 Lazarevic, V. et al. [27]
Bacteria	77	127	2016 Glassing, A. et al. [9]
Bacteria	62	0	2017 Salter, S.J. et al. [28]
Bacteria	0	122	2018 Kirstahler, P. et al. [29]
Bacteria	8	26	2019 de Goffau, M.C. et al. [30]
Bacteria	52	2	2019 Weyrich, L.S. et al. [31]
Bacteria	15	93	2020 Nejman D. et al. [32]
Viruses	0	1	2015 Mukherjee, S. et al. [33]
Viruses	0	1	2015 Kjartansdóttir, K.R. et al. [34]
Viruses	0	301	2019 Asplund, M. et al. [35]
Total (unique)	201	625	-

Table 1: Summary of common contaminants extracted from the literature. List of taxa listed can be found in the GRIMER repository (`https://github.com/pirovc/grimer/tree/main/sources/contaminants`)

Contamination sources

We compiled an extensive list of possible contaminant taxa reported in several studies (Table 1). The studies selected were obtained from cross-references in review articles [11] and individual selected findings in the literature. Articles were manually curated and more studies could potentially be added to the list. The studies selected are very diverse in terms of sequencing technology, methodology used and environment of origin. Contamination in those studies can be originated from sequencing kits as well as the lab environment or other unknown sources. The idea behind compiling this list is to detect which taxa is the most recurrently identified as contaminant in diverse conditions, providing a guideline and consensus for further studies. Entries on this list are not strictly considered a contaminant. However, the list serves a good evidence towards it, specially if entries are highly recurrent. Those contaminants were reported mainly at genus or species level in different formats, names and taxonomies. We manually curated and converted them into a the NCBI taxonomy [19] nomenclature for standardized usage.

Reference sources

Reference annotations give another level of evidence while visualizing the outcome of a study. We provided a list of common organisms found in probable external contamination sources: taxa commonly occurring in human skin, oral and nasal cavities as well as face and other human limbs. Those were reported as possible sources of contamination [11]. Reference organisms were obtained from BacDive [36], eHOMD [37] and further publications [38]. Additionally, any custom set of references can be used to connect analysis outcomes with known environments or biome to further explore the relations of observations in a study.

MGnify

MGnify [18] is a resource to analyze microbiome data in an automated and standardized way. Thousands of analyzed studies are publicly available with related metadata. We mined this repository with the open API (`https://www.ebi.ac.uk/metagenomics/api/v1/`) and collected all taxonomic classifications available for every study. For each study, we collected the latest taxonomic classification based on the highest pipeline version available. If multiple classification were present (SSU, LSU, ITS), we selected the largest one by file size. For each study output, the top 10 top most abundant organisms were linked to the study respective biome(s) definition and a final count of top organisms by biome is generated. GRIMER uses this resource to annotate observations and links how many times each identified taxa were present in other biomes. This gives another level of evidence for the possible origin of certain taxa in a study, compared to thousands other microbiome studies. For example, in the current version, the genus *Ralstonia*, a commonly reported contaminant, appeared in 30 Environmental Aquatic biome studies and 14 Engineered Bioreactor studies (out of a total of 79 studies) while the human-related bacterial genus *Prevotella* appears mostly in host-associated biomes (89% of occurences).

All 5 levels of biome classification are available for each taxonomic entry.

GRIMER

GRIMER requires only a contingency table to generate the full report, either in a text/tabular format (observations and samples either in rows or columns with a header) or a BIOM file [39]. Further data can be provided to improve the report:

- Metadata: to annotate samples and give further technical information. The file is tabular and accepts categorical and numerical fields. The QIIME 2 metadata header standards are accepted (`https://docs.qiime2.org/2021.2/tutorials/metadata/`), taking advantage of a large set of tools and validations already available for this format.
- Taxonomy database: automatically summarizes data in many taxonomic levels and generate plots accordingly. Taxonomy is fully automated and match observations based on a taxonomic identifier (taxid) or taxonomic name for several commonly used taxonomies (NCBI, GTDB, SILVA, GreenGenes, OTT).
- Controls: one or more groups of controls can be provided in a simple text file. Those samples will be further used to summarize data and annotate plots.
- Sources: custom sources of contamination/references can be provided in addition to the pre-compiled ones described above.

GRIMER will parse and process the data provided and run a set of analysis:

- Filtering and transformation: observations and samples can be filtered to reduce noise or small counts. Transformations are applied (log, clr, normalization) to account for the composionality of the data and improve some visualizations.
- Hierarchical clustering: one or more metrics and methods can be used to perform the clustering. The combination of all of them are executed and available in the report. For this analysis, zeros are replaced by small counts defined by the user.
- Correlation: Spearman correlation coefficient is calculated for top abundant observations in the study
- DECONTAM [40]: R package with a simple method to detect contaminating taxa/observations based on two main assumptions: frequency of contaminant taxa inversely correlate with DNA concentrations and contaminant taxa is more prevalent in control samples than in biological samples. DECONTAM uses linear models based on the assumptions and frequencies of the data and outputs a score for each observation to define contamination. If DNA concentration is not provided, total counts are used instead as an indirect concentration value replacement.

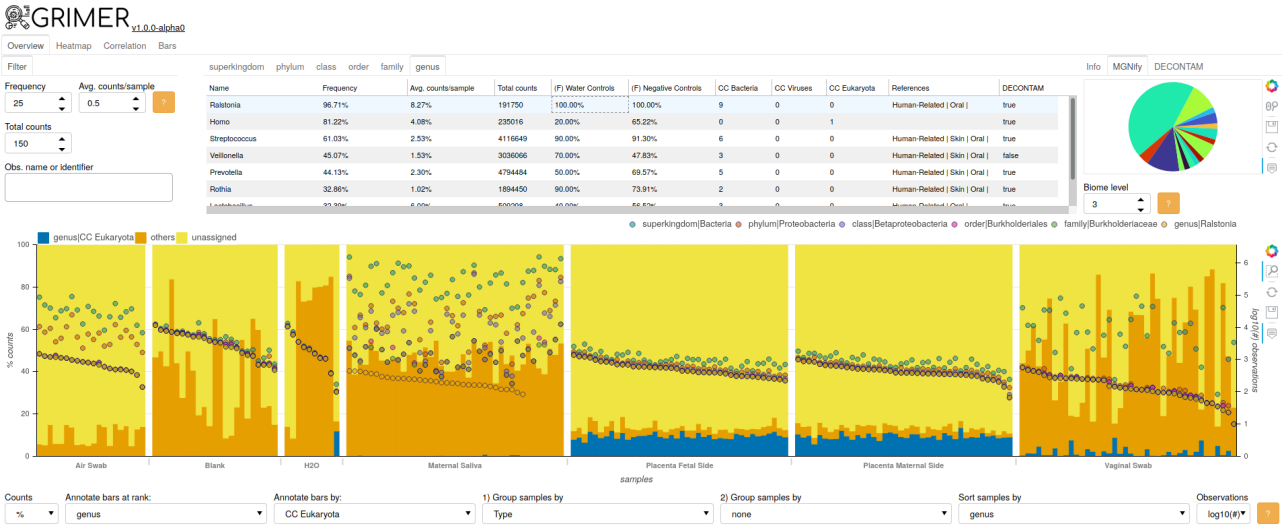
GRIMER Report

GRIMER will generate a report/dashboard with visualizations to better understand the distribution of observation counts among samples and the connection with external annotations, metadata and taxonomy. Currently GRIMER reports contain 4 main panels: Overview (Figure 1), Heatmap (Figure 2), Correlation and Bars (Figure 3). Some of them were previously suggested to be adequate for contamination detection [16] and are commonly used in microbiome analysis.

An individual summary for each observation and its relation to annotations and distribution among samples can be found in the Overview panel (Figure 1). Here all evidences related to a specific observation are integrated for further examination. Each observation provided in the study is listed and summarized in a tabular format. Once selected, the distribution of counts for the specific observation for each sample can be observed in the bottom bar plot. Information of annotations, MGnify biomes and DECONTAM output are also available in the same interface for each individual observation. The DECONTAM output indicates if the observation is classified as a contaminant with a score and a plot showing the frequency of the selected observation against the DNA concentration for all samples. Linear models showing the expected values for contamination and non-contamination values are also plotted. If provided, taxonomic lineages are integrated in the table and plots and observation are decomposed into taxonomic levels. The Overview panel also summarizes samples in the bar plot (Figure 1). Those bars can be transformed, annotated, grouped and sorted to connect specific observation values to overall sample distribution.

Several transformations can be applied to the data (normalization, log, center log ratio) to be further visualized in the Heatmap panel (Figure 2). Hierarchical clustering and sorting options can be independently selected for samples and observations to enable pattern detection (e.g. batch effects, treatment effects, etc).

Figure 1: Overview panel. Table summarizes observations at each taxonomic level. Selecting observations will trigger changes to the related charts and panels (Info, MGnify, DECONTAM) and to the bottom bar plot. The bar plot contains general and summarized counts (left y-axis) and specific observation counts (right y-axis) for each sample (x-axis). Bars can be annotated by different sources at different taxonomic levels, grouped and sorted by metadata and y-axis values can be log transformed or normalized.



Dendrograms are also available when showing clustering options. Annotation bars are plotted around the heatmap showing dynamically selection of sample annotations (metadata) and observation annotations (contamination sources, reference sources, controls and DECONTAM output). Metadata is automatically colored to reflect categories (distinct colors) and numeric (sequential colors) fields. Multiple metadata fields can be select interactively. One heatmap is generated for each taxonomic level.

Correlation between observations are plotted as a matrix (Figure 3). Positive or negative correlations among observations can point towards meaningful signals in the microbiome analysis. Observations present in multiple samples in similar ratios are positively correlated and the opposite configures negative correlation. Once a signal is observed, the correlation matrix can indicate co-occurrence of observations and help to identify further candidates (e.g. cluster of co-occurring contaminants). Normalized distribution of top observations for each sample can be visualized in the Bars panel (Figure 3) to easily compare overall distribution of observations among samples, with options for grouping and sorting by metadata.

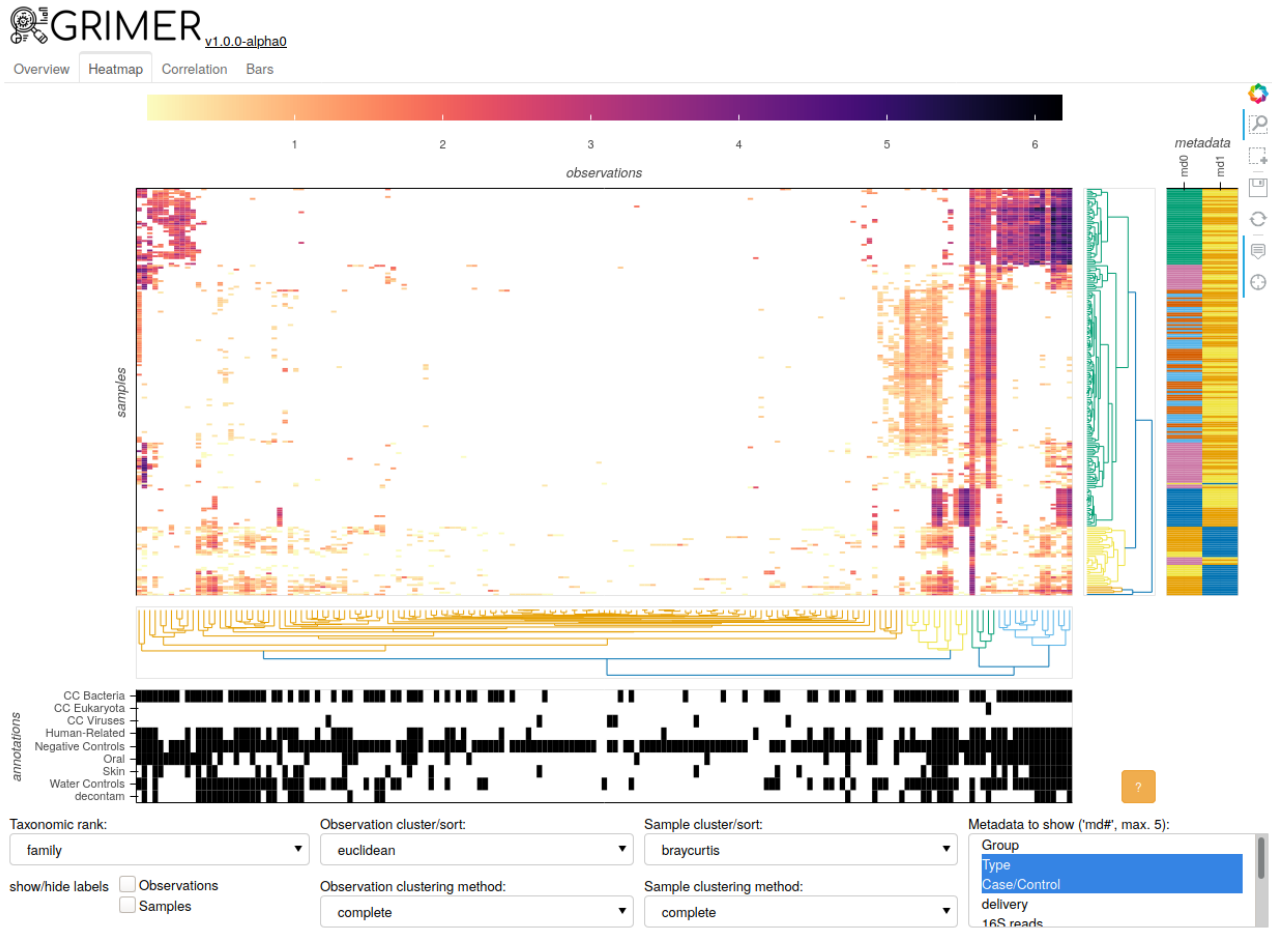
Implementation

GRIMER is written in Python and Javascript and outputs a report file the HTML format. All visualizations and layouts are created with the the Bokeh library (`https://bokeh.org`). Bokeh plots, tables and charts automatically provide a set of tools for interaction (e.g. zoom, selection) with an option to export the current selection to an image file. Many plots have interactive tooltips, showing more information about the data under the mouse cursor. Help buttons are also included, explaining basics about the plots and information displayed.

Further libraries were used to analyze samples and generate the report: pandas [41] for general parsing and data structures, scipy [42] for hierarchical clustering and correlation analysis, scikit-bio (`http://scikit-bio.org/`) for transformations. Scripts to download and generate MGnify annotations and update reference sources are provided in the GRIMER repository (`https://github.com/pirovc/grimer`).

GRIMER automatically handles taxonomic entries using MultiTax (`https://github.com/pirovc/multitax`). GRIMER will automatically download and convert any taxonomic id or name internally and decompose results in taxonomic ranks. Supported taxonomies are: NCBI, GTDB, Silva, GreenGenes, OpenTree Taxonomy. Reference and annotation sources are currently only available for NCBI Taxonomy.

Figure 2: Heatmap panel. Values from input table are transformed and plotted with samples in the y-axis and observations in the x-axis. One heatmap is generated for each taxonomic rank available. Data can be sorted by annotations sources, metadata and also clustered by multiple methods and metrics. If clustering is selected, dendrograms are automatically plotted in the side panels. The bottom panel show annotation for each observation and the right panel shows colored-based metadata for each sample. Metadata can be dynamically selected. User can visualize values for heatmap entries and metadata with the mouse cursor over the plots.



Results

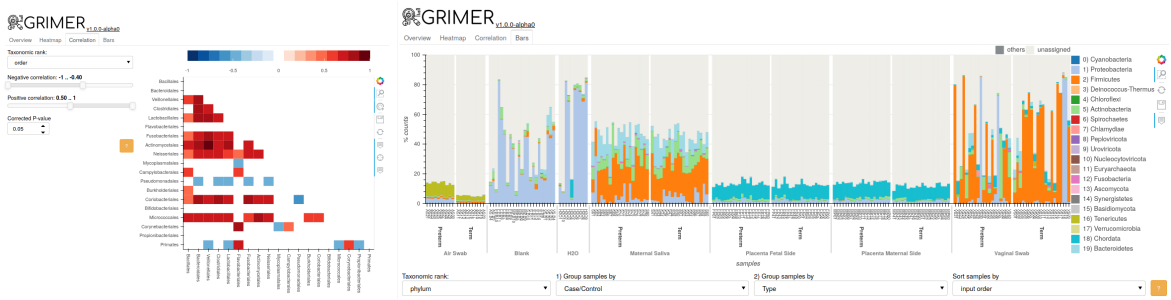
We re-analyzed publicly available studies to demonstrate the use of GRIMER reports in real case scenarios and what types of analysis are possible. In some examples we try to reproduce analysis and in other cases point to new evidences that may be overlooked. We encourage the readers to open the GRIMER reports provided in the Supplementary Materials and interactively visualize the results being described to fully understand the capabilities of the report.

Detecting contamination

The attempt to detect and describe a possible human placental microbiome has generated several studies and investigations [43, 17, 30, 44]. Leiby et al. [45] published a detailed and well designed study contributing to the subject. Placental samples for term and preterm newborns were collected for the maternal and fetal side. Additionally, positive control samples were obtained from the mothers (saliva and cervicovaginal fluid) as well as negative control samples (air from the sample processing room, empty tubes and PCR grade water). The study was performed in both marker gene sequencing (amplicon) and with whole (meta)genome sequencing (WGS). The authors could not distinguished a unique placental microbiome which differs from the contamination background. We re-analyzed the samples in a standard analysis pipeline, generated a GRIMER report for WGS and amplicon data (Supplementary Materials) and searched for the previously detected contamination.

In the WGS analysis outcome, the *Ralstonia* genus is present in 96% of the all samples with an average abundance of 8.27%. Reads assigned for this genus were found in all negative control samples. *Ralstonia* was also reported in 9 studies as a common contaminant, based on our compiled contaminant list and it was classified

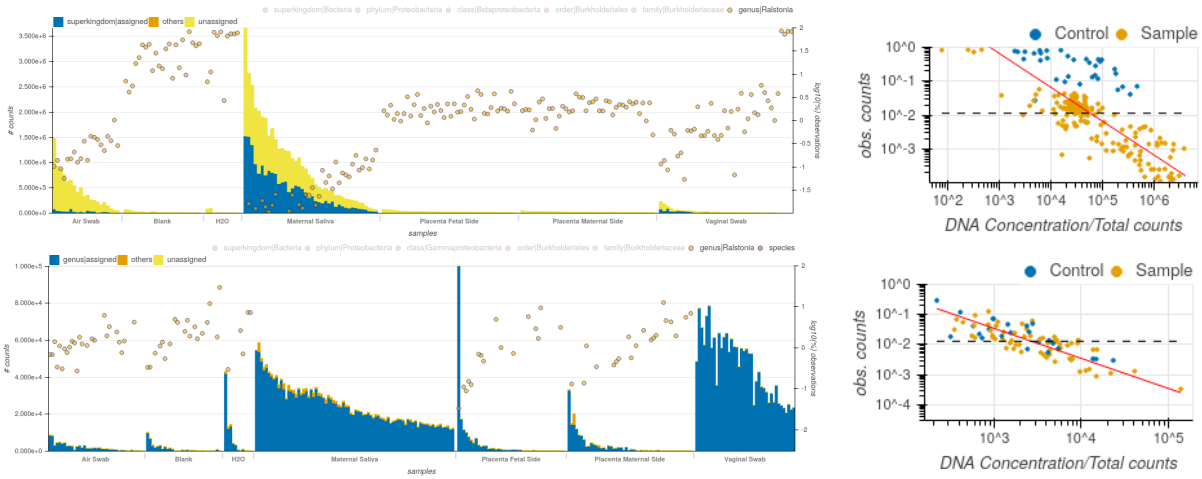
Figure 3: Correlation (left) and Bars (right) panels. Correlation plot shows positive or negative correlation between top taxa for each taxonomic rank with option to filter value ranges and p-value. Bars shows normalized values (y-axis) for each sample (x-axis) at each taxonomic rank available. Samples can be grouped and sorted by metadata.



as a contaminant by the DECONTAM method, based on the correlation of frequencies and the total number of reads/sample. Further, the abundance of this genus is higher in negative controls and placental samples as well as in samples with low number of reads, probably related to their low biomass as shown in the Figure 4. Those results are inline with the ones reported in the original publication [45] even though the data was re-analyzed with a different set of tools, parameters and database version. All evidences describe pointing to *Ralstonia* as a contaminant was automatically generated by GRIMER and can be directly extracted from the Overview panel from the report. For the amplicon data, a similar pattern can be detected for the *Ralstonia* genus based on amplicon sequence variants (Figure 4).

Further, all other taxa present can easily be verified for the same patterns. *Pseudonomas* show similar patterns and were also reported originally as probable contaminants in the placental samples. The correlation plot can be used to detect further other organisms with similar patterns of the ones already identified. *Cutibacterium*, another common contaminant not detected in the original publication, has a positive correlation value of 0.451 and 0.409 for WGS and amplicon analysis, respectively, with *Ralstonia* and shows similar patterns as previously detected contaminants. This make *Cutibacterium* a probable contamination source not mentioned in the original study.

Figure 4: LEFT: Bar plots as described in Figure 1. Right y-axis shows normalized abundance of genus *Ralstonia* in log scale for each sample in the WGS (TOP) and amplicon (BOTTOM) data. Samples are grouped by sample type and sorted by number of total reads. RIGHT: DECONTAM plot for *Ralstonia* for the WGS (TOP) and amplicon (BOTTOM). The bar plots show higher abundance of the genus in the negative control samples (Blank and H2O) as well as increased abundance with fewer reads in all sample types. DECONTAM plot show that biological samples (orange dots) follow the expected distribution for contamination (red line). Both evidences point towards *Ralstonia* as a contaminant taxa.



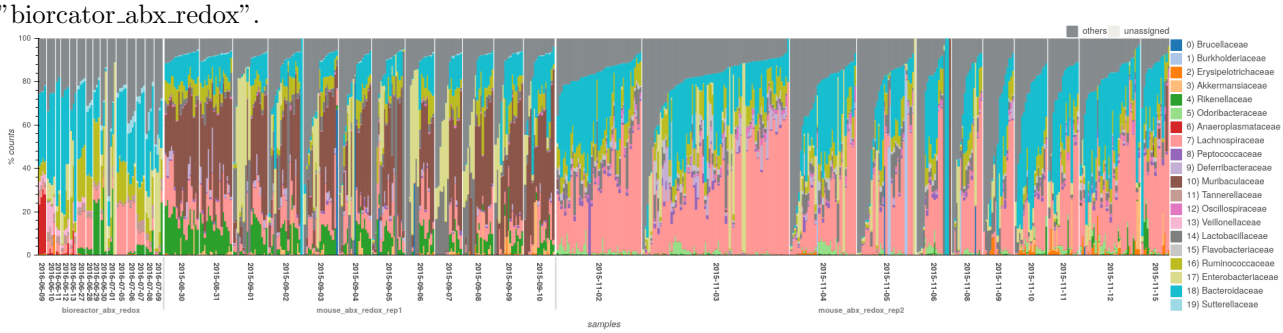
Microbiome study exploration

MGnify

Besides utilizing the complete MGnify resource as a source for annotation, GRIMER can also analyze any single study publicly available in the website automatically. Although MGnify provides a rich platform and data, analysis by study is limited and only individual runs and samples can be visualized online. A script to download taxonomic profiles and metadata directly from EMBL-EBI servers is provided. With only a study accession (e.g. MGYS00001234) is possible to generate a complete GRIMER report for any study deposited in MGnify.

We randomly selected a study to demonstrate the possibilities of analysis and visualizations. The study entitled "Antibiotic induced changes in the microbiota disrupt redox dynamics in the gut" (`https://www.ebi.ac.uk/metagenomics/studies/MGYS00005180`) investigated mouse gut microbial responses to broad spectrum antibiotic treatments [46]. GRIMER can easily generate bar plots at any taxonomic level with dynamic grouping, as shown in Figure 5 where patterns to specific groups and collections dates are detectable in the sample bars.

Figure 5: Bar plot for the top 20 family level taxonomic entries grouped by "project name" and "collection date". It is possible to detect a shift in the microbiome composition based on the date collected for the "biorcator_abx_redox".



MicrobiomeDB

MicrobiomeDB is a data-mining platform for interrogating microbiome experiments [47]. Although smaller than MGnify, it contains well curated studies with extensive metadata information. GRIMER can easily expand the possibilities of visualizations directly from the BIOM files provided in their website. As an example, we generated a GRIMER report for "The Early Childhood Antibiotics and the Microbiome" (ECAM) study where effects of antibiotics, delivery mode and diet were evaluated on the gut microbiome. A picture of the interactive heatmap is show in Figure 6. Clustering methods and metadata were dynamically selected in the report. With very little effort GRIMER report enables exploration and discovery of complex relations between data and metadata. Different clusters of data can be explained by different metadata fields, showing a clear pattern (Figure 6).

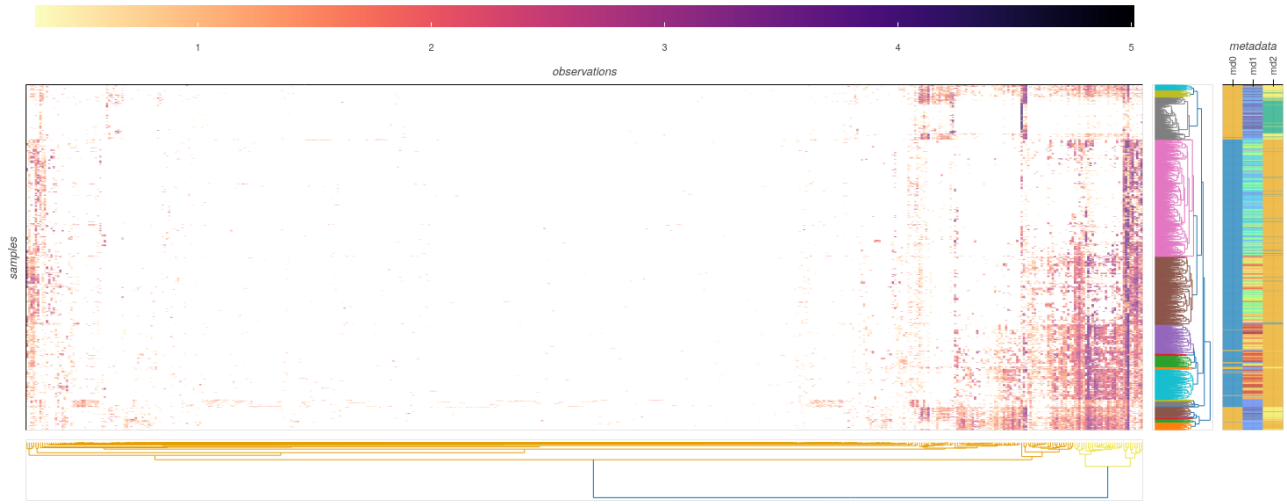
Discussion

GRIMER is an easy-to-use and accessible tool for specialists and non-specialists that generates a concise interactive report with a set of analysis, visualizations, and data connections from a simple table of counts. It automatically summarizes several levels of evidence to better understand the relation between observations, samples, metadata and taxonomy. GRIMER is a valuable tool for investigating contamination, a problem that affects every microbiome study to some degree.

In addition, we provided a compiled list of common taxa contaminants based on several publications with many of them being recurrent in diverse studies, pointing to a consensus. The list is the first step to centralize and standardize re-occurring contaminants described in the literature. We expect this list to incrementally grow overtime as more evidence of kit and laboratory contamination becomes available. We plan to associate each contaminant in the list with its study metadata so they could be selected by several metrics like study type, technology, kit used, etc. We accept data donations and improvements into the current list currently hosted at `https://github.com/pirovc/grimer/tree/main/sources`. We plan to provide GRIMER as a web-application where donation of control samples could be automatized with the user consent.

GRIMER works out-of-the-box with as little data as possible but incrementally expands when more data is provided and can be adapted for user necessities. GRIMER is fast and run in a matter of seconds in a standard notebook. Report sizes usually vary from 1-10MB and are highly compressible, since they are text-based HTML

Figure 6: Heatmap visualization of genus level log transformed counts for the ECAM study obtained from MicrobiomeDB. Samples are clustered with average method and braycurtis metric and observations clustered with average method and euclidean metric. Metadata fields are md0: Mother (orange) or child (blue), md1: Age in days (darker red = older) and md2: Sample type (Colon - Orange , Rectum - Yellow, Vagina - Green). Patterns and clusteres based on the microbial composition can be linked to metadata groups.



files. GRIMER reports with higher number of samples (> 1000) can grow in size (10-100MB) but still work normally. If report size is a limitation, many options can be changed to generated more compact files: reducing number of taxonomy ranks displayed, less combinations of analysis, filtering very low abundant observations, among others.

One of GRIMER core strengths is the taxonomy automation. It accepts taxonomic ids from several different taxonomies, but also parse names and converts them to their respective identifiers. That way it is possible to automatically summarize data in different ranks. That means that users do not have to handle any taxonomic classification and everything will work automatically. GRIMER was developed in a way where new visualizations can be included with little effort. Other analysis and their plots are already being developed and will be available in next releases including alpha-diversity metrics, beta-diversity metrics, differential abundance analysis.

GRIMER reports are portable and fully functional offline. This allows analysis to be performed by many researchers with different backgrounds working together in the same study, increasing direct interaction with data also for non-bioinformaticins. The portability also enable better documentation of results, reproducibility and shareability.

Overall we believe that GRIMER is a valuable contribution to the field of microbiome and can facilitate data exploration, analysis and contamination detection.

Acknowledgements

We thank all partners in the ZooSeq project for helpful discussions and support in this project.

Funding

This work was financially supported by the German Ministry for Education and Research (Bundesministerium für Bildung und Forschung - BMBF) grant number 01KI1905D.

References

[1] Jolinda Pollock, Laura Glendinning, Trong Wisedchanwet, and Mick Watson. The madness of microbiome: Attempting to find consensus "best practice" for 16S microbiome studies. *Applied and Environmental Microbiology*, (February):AEM.02627–17, 2018.

[2] Dorothy Kim, Casey E. Hofstaedter, Chunyu Zhao, Lisa Mattei, Ceylan Tanes, Erik Clarke, Abigail Lauder, Scott Sherrill-Mix, Christel Chehoud, Judith Kelsen, Máire Conrad, Ronald G. Collman, Robert Baldas-

sano, Frederic D. Bushman, and Kyle Bittinger. Optimizing methods and dodging pitfalls in microbiome research. *Microbiome*, 5(1):52, 2017. Publisher: Microbiome ISBN: 4016801702.

[3] Rob Knight, Alison Vrbanac, Bryn C. Taylor, Alexander Aksenov, Chris Callewaert, Justine Debelius, Antonio Gonzalez, Tomasz Kosciolek, Laura-Isobel McCall, Daniel McDonald, Alexey V. Melnik, James T. Morton, Jose Navas, Robert A. Quinn, Jon G. Sanders, Austin D. Swafford, Luke R. Thompson, Anupriya Tripathi, Zhenjiang Z. Xu, Jesse R. Zaneveld, Qiyun Zhu, J. Gregory Caporaso, and Pieter C. Dorrestein. Best practices for analysing microbiomes. *Nature Reviews Microbiology*, 16(7):410–422, July 2018.

[4] Gregory B. Gloor, Jean M. Macklaim, Vera Pawlowsky-Glahn, and Juan J. Egozcue. Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, 8(November):1–6, November 2017.

[5] W. Florian Fricke and Jacques Ravel. Microbiome or no microbiome: are we looking at the prenatal environment through the right lens? *Microbiome*, 9(1):9, January 2021.

[6] Martin J. Blaser, Suzanne Devkota, Kathy D. McCoy, David A. Relman, Moran Yassour, and Vincent B. Young. Lessons learned from the prenatal microbiome controversy. *Microbiome*, 9(1):8, January 2021.

[7] Jens Walter and Mathias W. Hornef. A philosophical perspective on the prenatal in utero microbiome debate. *Microbiome*, 9(1):5, January 2021.

[8] Susannah J. Salter, Michael J. Cox, Elena M. Turek, Szymon T. Calus, William O. Cookson, Miriam F. Moffatt, Paul Turner, Julian Parkhill, Nicholas J. Loman, and Alan W. Walker. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, 12(1):87, November 2014.

[9] Angela Glassing, Scot E. Dowd, Susan Galandiuk, Brian Davis, and Rodrick J. Chiodini. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathogens*, 8(1):24, May 2016.

[10] Jeremiah J. Minich, Jon G. Sanders, Amnon Amir, Greg Humphrey, Jack A. Gilbert, and Rob Knight. Quantifying and Understanding Well-to-Well Contamination in Microbiome Research. *mSystems*, 4(4), August 2019. Publisher: American Society for Microbiology Journals Section: Research Article.

[11] Raphael Eisenhofer, Jeremiah J. Minich, Clarisse Marotz, Alan Cooper, Rob Knight, and Laura S. Weyrich. Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends in Microbiology*, 27(2):105–117, February 2019.

[12] Karl Gruber. Here, there, and everywhere. *EMBO reports*, 16(8):898–901, August 2015. Publisher: John Wiley & Sons, Ltd.

[13] Florian P. Breitwieser, Mihaela Pertea, Aleksey Zimin, and Steven L. Salzberg. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Research*, page gr.245373.118, May 2019.

[14] Bastian V. H. Hornung, Romy D. Zwittink, and Ed J. Kuijper. Issues and current standards of controls in microbiome research. *FEMS Microbiology Ecology*, 95(5), May 2019. Publisher: Oxford Academic.

[15] Robyn L. Marsh, Maria T. Nelson, Chris E. Pope, Amanda J. Leach, Lucas R. Hoffman, Anne B. Chang, and Heidi C. Smith-Vaughan. How low can we go? The implications of low bacterial load in respiratory microbiota studies. *Pneumonia*, 10(1):7, July 2018.

[16] Marcus C. de Goffau, Susanne Lager, Susannah J. Salter, Josef Wagner, Andreas Kronbichler, D. Stephen Charnock-Jones, Sharon J. Peacock, Gordon C. S. Smith, and Julian Parkhill. Recognizing the reagent microbiome. *Nature Microbiology*, 3(8):851–853, August 2018.

[17] Isoken Nicholas Olomu, Luis Carlos Pena-Cortes, Robert A. Long, Arpita Vyas, Olha Krichevskiy, Ryan Luellwitz, Pallavi Singh, and Martha H. Mulks. Elimination of "kitome" and "splashome" contamination results in lack of detection of a unique placental microbiome. *BMC microbiology*, 20(1):157, June 2020.

[18] Alex L. Mitchell, Alexandre Almeida, Martin Beracochea, Miguel Boland, Josephine Burgin, Guy Cochrane, Michael R. Crusoe, Varsha Kale, Simon C. Potter, Lorna J. Richardson, Ekaterina Sakharova, Maxim Scheremetjew, Anton Korobeynikov, Alex Shlemov, Olga Kunyavskaya, Alla Lapidus, and Robert D. Finn. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research*, 48(D1):D570–D578, January 2020.

[19] Conrad L Schoch, Stacy Ciufo, Mikhail Domrachev, Carol L Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, Richard Mcveigh, Kathleen O'Neill, Barbara Robbertse, Shobha Sharma, Vladimir Soussov, John P Sullivan, Lu Sun, Seán Turner, and Ilene Karsch-Mizrachi. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database*, 2020(baaa062), January 2020.

[20] Michael A. Tanner, Brett M. Goebel, Michael A. Dojka, and Norman R. Pace. Specific Ribosomal DNA Sequences from Diverse Environmental Settings Correlate with Experimental Contaminants. *Applied and Environmental Microbiology*, 64(8):3110–3113, August 1998. Publisher: American Society for Microbiology Section: GENERAL MICROBIAL ECOLOGY.

[21] Niclas Grahn, Margaretha Olofsson, Katarina Ellnebo-Svedlund, Hans Jürg Monstein, and Jon Jonasson. Identification of mixed bacterial DNA contamination in broad-range PCR amplification of 16S rDNA V1 and V3 variable regions by pyrosequencing of cloned amplicons. *FEMS microbiology letters*, 219(1):87–91, February 2003.

[22] H. A. Barton, N. M. Taylor, B. R. Lubbers, and A. C. Pemberton. DNA extraction from low-biomass carbonate rock: an improved method with reduced contamination and the low-biomass contaminant database. *Journal of Microbiological Methods*, 66(1):21–31, July 2006.

[23] Martin Laurence, Christos Hatzis, and Douglas E. Brash. Common Contaminants in Next-Generation Sequencing That Hinder Discovery of Low-Abundance Microbes. *PLOS ONE*, 9(5):e97876, May 2014. Publisher: Public Library of Science.

[24] Jake Jervis-Bardy, Lex E. X. Leong, Shashikanth Marri, Renee J. Smith, Jocelyn M. Choo, Heidi C. Smith-Vaughan, Elizabeth Nosworthy, Peter S. Morris, Stephen O'Leary, Geraint B. Rogers, and Robyn L. Marsh. Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of Illumina MiSeq data. *Microbiome*, 3(1):19, May 2015.

[25] E. Jousselin, A.-L. Clamens, M. Galan, M. Bernard, S. Maman, B. Gschloessl, G. Duport, A. S. Meseguer, F. Calevro, and A. Coeur D'acier. Assessment of a 16S rRNA amplicon Illumina sequencing procedure for studying the microbiome of a symbiont-rich aphid genus. *Molecular Ecology Resources*, 16(3):628–640, 2015. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.12478.

[26] Abigail P. Lauder, Aoife M. Roche, Scott Sherrill-Mix, Aubrey Bailey, Alice L. Laughlin, Kyle Bittinger, Rita Leite, Michal A. Elovitz, Samuel Parry, and Frederic D. Bushman. Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. *Microbiome*, 4, June 2016.

[27] Vladimir Lazarevic, Nadia Gaïa, Myriam Girard, and Jacques Schrenzel. Decontamination of 16S rRNA gene amplicon sequence datasets based on bacterial load assessment by qPCR. *BMC Microbiology*, 16(1):73, April 2016.

[28] Susannah J. Salter, Claudia Turner, Wanitda Watthanaworawit, Marcus C. de Goffau, Josef Wagner, Julian Parkhill, Stephen D. Bentley, David Goldblatt, Francois Nosten, and Paul Turner. A longitudinal study of the infant nasopharyngeal microbiota: The effects of age, illness and antibiotic use in a cohort of South East Asian children. *PLoS Neglected Tropical Diseases*, 11(10), October 2017.

[29] Philipp Kirstahler, Søren Solborg Bjerrum, Alice Friis-Møller, Morten la Cour, Frank M. Aarestrup, Henrik Westh, and Sünje Johanna Pamp. Genomics-Based Identification of Microorganisms in Human Ocular Body Fluid. *Scientific Reports*, 8(1):4126, March 2018. Number: 1 Publisher: Nature Publishing Group.

[30] Marcus C. de Goffau, Susanne Lager, Ulla Sovio, Francesca Gaccioli, Emma Cook, Sharon J. Peacock, Julian Parkhill, D. Stephen Charnock-Jones, and Gordon C. S. Smith. Human placenta has no microbiome but can contain potential pathogens. *Nature*, 572(7769):329–334, August 2019. Number: 7769 Publisher: Nature Publishing Group.

[31] Laura S. Weyrich, Andrew G. Farrer, Raphael Eisenhofer, Luis A. Arriola, Jennifer Young, Caitlin A. Selway, Matilda Handsley-Davis, Christina J. Adler, James Breen, and Alan Cooper. Laboratory contamination over time during low-biomass sample analysis. *Molecular Ecology Resources*, 19(4):982–996, July 2019. Publisher: John Wiley & Sons, Ltd.

[32] Deborah Nejman, Ilana Livyatan, Garold Fuks, Nancy Gavert, Yaara Zwang, Leore T. Geller, Aviva Rotter-Maskowitz, Roi Weiser, Giuseppe Mallel, Elinor Gigi, Arnon Meltser, Gavin M. Douglas, Iris Kamer, Vancheswaran Gopalakrishnan, Tali Dadosh, Smadar Levin-Zaidman, Sofia Avnet, Tehila Atlan, Zachary A. Cooper, Reetakshi Arora, Alexandria P. Cogdill, Md Abdul Wadud Khan, Gabriel Ologun,

Yuval Bussi, Adina Weinberger, Maya Lotan-Pompan, Ofra Golani, Gili Perry, Merav Rokah, Keren Bahar-Shany, Elisa A. Rozeman, Christian U. Blank, Anat Ronai, Ron Shaoul, Amnon Amit, Tatiana Dorfman, Ran Kremer, Zvi R. Cohen, Sagi Harnof, Tali Siegal, Einav Yehuda-Shnaidman, Einav Nili Gal-Yam, Hagit Shapira, Nicola Baldini, Morgan G. I. Langille, Alon Ben-Nun, Bella Kaufman, Aviram Nissan, Talia Golan, Maya Dadiani, Keren Levanon, Jair Bar, Shlomit Yust-Katz, Iris Barshack, Daniel S. Peeper, Dan J. Raz, Eran Segal, Jennifer A. Wargo, Judith Sandbank, Noam Shental, and Ravid Straussman. The human tumor microbiome is composed of tumor type–specific intracellular bacteria. *Science*, 368(6494):973–980, May 2020. Publisher: American Association for the Advancement of Science Section: Research Article.

[33] Supratim Mukherjee, Marcel Huntemann, Natalia Ivanova, Nikos C. Kyrpides, and Amrita Pati. Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Standards in Genomic Sciences*, 10(1):18, March 2015.

[34] Kristín Rós Kjartansdóttir, Jens Friis-Nielsen, Maria Asplund, Sarah Mollerup, Tobias Mourier, Randi Holm Jensen, Thomas Arn Hansen, Alba Rey-Iglesia, Stine Raith Richter, David E. Alquezar-Planas, Pernille V. S. Olsen, Lasse Vinner, Helena Fridholm, Thomas Sicheritz-Pontén, Lars Peter Nielsen, Søren Brunak, Eske Willerslev, Jose M. G. Izarzugaza, and Anders Johannes Hansen. Traces of ATCV-1 associated with laboratory component contamination. *Proceedings of the National Academy of Sciences*, 112(9):E925–E926, March 2015. Publisher: National Academy of Sciences Section: Letter.

[35] M. Asplund, K.R. Kjartansdóttir, S. Mollerup, L. Vinner, H. Fridholm, J.A.R. Herrera, J. Friis-Nielsen, T.A. Hansen, R.H. Jensen, I.B. Nielsen, S.R. Richter, A. Rey-Iglesia, M.L. Matey-Hernandez, D.E. Alquezar-Planas, P.V.S. Olsen, T. Sicheritz-Pontén, E. Willerslev, O. Lund, S. Brunak, T. Mourier, L.P. Nielsen, J.M.G. Izarzugaza, and A.J. Hansen. Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries. *Clinical Microbiology and Infection*, 25(10):1277–1285, October 2019.

[36] Lorenz Christian Reimer, Anna Vetcininova, Joaquim Sardà Carbasse, Carola Söhngen, Dorothea Gleim, Christian Ebeling, and Jörg Overmann. BacDive in 2019: bacterial phenotypic data for High-throughput biodiversity analysis. *Nucleic Acids Research*, 47(D1):D631–D636, January 2019.

[37] Isabel F. Escapa, Tsute Chen, Yanmei Huang, Prasad Gajare, Floyd E. Dewhirst, and Katherine P. Lemon. New Insights into Human Nostril Microbiome from the Expanded Human Oral Microbiome Database (eHOMD): a Resource for the Microbiome of the Human Aerodigestive Tract. *mSystems*, 3(6), December 2018. Publisher: American Society for Microbiology Journals Section: Resource Report.

[38] Allyson L. Byrd, Yasmine Belkaid, and Julia A. Segre. The human skin microbiome. *Nature Reviews Microbiology*, 16(3):143–155, March 2018. Number: 3 Publisher: Nature Publishing Group.

[39] Daniel McDonald, Jose C Clemente, Justin Kuczynski, Jai Ram Rideout, Jesse Stombaugh, Doug Wendel, Andreas Wilke, Susan Huse, John Hufnagle, Folker Meyer, Rob Knight, and J Gregory Caporaso. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaScience*, 1(2047-217X-1-7), December 2012.

[40] Nicole M. Davis, Diana M. Proctor, Susan P. Holmes, David A. Relman, and Benjamin J. Callahan. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*, 6(1):226, December 2018.

[41] Wes McKinney. Data Structures for Statistical Computing in Python. pages 56–61, Austin, Texas, 2010.

[42] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, and Paul van Mulbregt. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, March 2020. Number: 3 Publisher: Nature Publishing Group.

[43] Rachel B. Silverstein and Indira U. Mysorekar. Group therapy on in utero colonization: seeking common truths and a way forward. *Microbiome*, 9(1):7, January 2021.

[44] Irene Sterpu, Emma Fransson, Luisa W. Hugerth, Juan Du, Marcela Pereira, Liqin Cheng, Sebastian Alexandru Radu, Lorena Calderón-Pérez, Yinghua Zha, Pia Angelidou, Alexandra Pennhag, Fredrik Boulund, Annika Scheynius, Lars Engstrand, Eva Wiberg-Itzel, and Ina Schuppe-Koistinen. No evidence

for a placental microbiome in human pregnancies at term. *American Journal of Obstetrics and Gynecology*, 224(3):296.e1–296.e23, March 2021.

[45] Jacob S. Leiby, Kevin McCormick, Scott Sherrill-Mix, Erik L. Clarke, Lyanna R. Kessler, Louis J. Taylor, Casey E. Hofstaedter, Aoife M. Roche, Lisa M. Mattei, Kyle Bittinger, Michal A. Elovitz, Rita Leite, Samuel Parry, and Frederic D. Bushman. Lack of detection of a human placenta microbiome in samples from preterm and term deliveries. *Microbiome*, 6(1):196, October 2018.

[46] Aspen T Reese, Eugenia H Cho, Bruce Klitzman, Scott P Nichols, Natalie A Wisniewski, Max M Villa, Heather K Durand, Sharon Jiang, Firas S Midani, Sai N Nimmagadda, Thomas M O'Connell, Justin P Wright, Marc A Deshusses, and Lawrence A David. Antibiotic-induced changes in the microbiota disrupt redox dynamics in the gut. *eLife*, 7:e35987, June 2018. Publisher: eLife Sciences Publications, Ltd.

[47] Francislon S. Oliveira, John Brestelli, Shon Cade, Jie Zheng, John Iodice, Steve Fischer, Cristina Aurrecoechea, Jessica C. Kissinger, Brian P. Brunk, Christian J. Stoeckert, Gabriel R. Fernandes, David S. Roos, and Daniel P. Beiting. MicrobiomeDB: a systems biology platform for integrating, mining and analyzing microbiome experiments. *Nucleic Acids Research*, (November):1–8, November 2017.