Contamination detection and microbiome exploration with GRIMER

Vitor C. Piro[1,2,*] and Bernhard Y. Renard[1]

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

[2]Department of Mathematics and Computer Science, Freie Universität Berlin, Takustr. 9, 14195 Berlin, Germany

[*]Corresponding author: vitor.piro@fu-berlin.de

Abstract

Background: Contamination detection is a important step that should be carefully considered in early stages when designing and performing microbiome studies to avoid biased outcomes. Detecting and removing true contaminants is challenging, especially in low-biomass samples or in studies lacking proper controls. Interactive visualizations and analysis platforms are crucial to better guide this step, to help to identify and detect noisy patterns that could potentially be contamination. Additionally, external evidence, like aggregation of several contamination detection methods and the use of common contaminants reported in the literature could help to discover and mitigate contamination. **Results:** We propose GRIMER, a tool that performs automated analyses and generates a portable and interactive dashboard integrating annotation, taxonomy and metadata. It unifies several sources of evidence to help detect contamination. GRIMER is independent of quantification methods and directly analyses contingency tables to create an interactive and offline report. 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 and used an extensive list of possible external contaminant taxa and common contaminants with 210 genera and 627 species reported in 22 published articles. **Conclusion:** GRIMER enables visual data exploration and analysis, supporting contamination detection in microbiome studies. The tool and data presented are 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. Microbiome studies usually yield large amounts of raw

Name	Website	Reference
MG-RAST	https://www.mg-rast.org/	[10]
MGnify	https://www.ebi.ac.uk/metagenomics/	[11]
MicrobiomeDB	https://microbiomedb.org/	[12]
Nephele	https://nephele.niaid.nih.gov/	[13]
Qiita	https://qiita.ucsd.edu/	[14]

Table 1: Web resources to process, analyze and visualize microbiome data

sequences for several samples that can be analyzed with an increasing number of computational methods and databases. Standards, protocols, and best practices for designing and performing a microbiome study have been 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 *in silico* microbiome analysis, raw or quality-filtered 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 reliable measurements for each sample in the study [3]. This step can be computationally intensive but reduces the large amount of data into a concise table of measurements. Alternatively, genome assembly can be performed for metagenomics samples, allowing genome-resolved analysis. Although still a complex task, gene prediction, taxonomic and functional analysis are improved with metagenome-assembled genomes, resulting in overall better measurements [4].

After measurements are obtained, hypotheses are validated through data mining and statistical analysis. This step is mostly exploratory and specific to the hypotheses and research questions pursued and the required analyses are difficult to be fully automatized. It is also very important to take in consideration the compositionality of data at this stage when working with the microbiome [5]. Several comprehensive and generalized analytical packages [6, 7, 8] and web platforms (Table 1) are available to perform a large number of microbiome analysis: basic data summaries, diversity and functional analysis, microbial interactions, differential abundance among others. Additionally, interactive tools for analytical and visual exploration are extremely helpful in this stage to better understand the data distribution and its properties and to guide further investigations to follow. In the last decade, several applications were developed with focus on visualization of microbiome data (Table 2). A comparison among many of those methods and their functionalities can be found in a recent review [9].

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 [31, 32, 33]. However, the issue is not new and contamination has been known and reported for decades in the literature [34]. 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 microbes from laboratory technicians [2, 35, 36]. Internal contamination can be defined as a undesired exchange of genetic material between

Name	Focus	Plataform	Website	Year	Reference
METAGENassist	Comparative metagenomics	Web	http://www.metagenassist.ca/METAGENassist/	2012	[15]
VAMPS	Microbial population structures	Web	https://vamps2.mbl.edu/	2014	[16]
Shiny-phyloseq	Microbiome analysis	Locally hosted (R)	https://joey711.github.io/shiny-phyloseq/	2015	[17]
MetaCoMET	Microbiome analysis	Web	https://probes.pw.usda.gov/MetaCoMET/	2016	[18]
BusyBee Web	Metagenomics binning and analysis	Web	https://ccb-microbe.cs.uni-saarland.de/busybee	2017	[19]
MicrobiomeAnalyst	Microbiome analysis	Web	https://www.microbiomeanalyst.ca	2017	[20]
Burrito	Taxonomy and function analysis	Web	http://elbo-spice.cs.tau.ac.il/shiny/burrito/	2018	[21]
Pavian	Metagenomics analysis	Locally hosted (R)	https://github.com/fbreitwieser/pavian	2019	[22]
GenePiper	Microbiome analysis	Locally hosted (R)	https://github.com/raytonghk/genepiper	2020	[23]
animalcules	Microbiome analysis	Locally hosted (R)	https://github.com/compbiomed/animalcules	2021	[24]
MicrobiomeExplorer	Microbiome analysis	Locally hosted (R)	https://github.com/zoecastillo/microbiomeExplorer	2021	[25]
microViz	Microbiome analysis	Locally hosted (R)	https://github.com/david-barnett/microViz/	2021	[26]
Namco	Microbiome analysis	Web	https://exbio.wzw.tum.de/namco/	2021	[27]
OpenContami	Contaminant detection	Web	https://openlooper.hgc.jp/opencontami/	2021	[28]
wiSDOM	Microbiome analysis	Web or Locally hosted (R)	https://github.com/lunching/wiSDOM	2021	[29]
Mian	Microbiome analysis	Web	https://miandata.org/	2022	[30]
GRIMER	Contaminant detection	CLI + standalone file	https://github.com/pirovc/grimer	2022	this work

Table 2: Interactive analysis and visualization tools for microbiome data published in the last 10 years

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 [37].

Contamination may affect most sequencing projects to some degree, especially low-biomass samples [38]. The composition of an environmental sample is mostly unknown before sequencing, increasing the complexity of detecting contamination when compared to a defined isolate genome and targeted sequencing project. Low-biomass samples (e.g. meconium, blood, human tissues) yield little to no DNA to be amplified and sequenced, an ideal scenario for exogenous contaminants to out-compete and dominate the biological signal.

It is important that contamination is acknowledged, accounted for and discovered at the earliest stage of a study prior to statistical analysis, to not bias measurements and to ensure that bias is not propagated into databases [39, 40]. Inclusion of negative and positive control samples is the recommended way to measure, detect, and mitigate contamination [2, 38, 41]. 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 base on the frequency in relation to DNA concentration [42, 43].

However, due to the complexity and diverse possible sources of contamination, detection and mitigation are not a trivial tasks. Several approaches to identify and exclude background contamination in microbial studies have been proposed. These are based on exclusion of organisms detected in negative controls, use of replicates to find possible contaminants, removal of low abundant signals, negative correlation between organism abundance and bacterial load, clustering analysis, and others [44, 45]. Each approach has 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 the required increase in costs. [41] reported that based on publications from the 2018 issues of Microbiome and The ISME Journal, only 30% cited the use of negative controls and only 10% positive controls. Moreover, Harrison et al. [46] reported that out of 50 selected publications from 2019 and 2020, only 15 used some type of negative control and 10 of positive control to account for reagent contamination. There was also no observed increase in positive or negative controls usage in the literature from 2015 to 2020, based on selected publications. Additionally, the detection of re-occurring contaminants in extraction kits and reagents (also called "kitome") is known to be an issue [47] but remains under-explored, mainly for not being

properly cataloged, centralized or automated.

To overcome some of those challenges we propose GRIMER, a tool to analyze, visualize and explore microbiome studies with a focus on contamination detection. Based on a table of observations per sample, GRIMER generates an offline and interactive dashboard to automate data analysis, transformations and plots and generates a set of charts integrating evidence for better decision making and contamination detection. Additionally, we compiled an extensive list of common contaminants containing 210 genera and 627 species reported in 22 published articles. This data is integrated into the report. GRIMER is an effortless step 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`. Installation and usage instructions as well as an user manual are available in the repository. The tool is independent of analysis methods, does 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, analyses as well as 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 analyses in different environments.

GRIMER is independent of any quantification method and only requires a contingency table with raw counts of observations/components for each sample/composition in the study. Observations are usually, but not limited to, taxonomic entries (e.g. genus, species, strains), operational taxonomic units (OTUs), amplicon sequence variants (ASVs), or sequence features. A count of unclassified or unassigned observations is also supported to generate normalized values. Additional files and data can be provided to expand GRIMER reports: study metadata, a taxonomy database, multiple control samples, the 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 data sources.

Sample annotations are based on a user-provided 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 analyses as well as detection of batches and control/treatment effects.

Observation annotations are based on external lists of taxonomic entries, which can be used, for example,

Organism group	Genus	Species	Reference
Bacteria	6	0	1998 Tanner, M.A. et al. [49]
Bacteria	0	10	2002 Kulakov, L.A. et al. [50]
Bacteria	4	0	2003 Grahn, N. et al. [51]
Bacteria	16	0	2006 Barton, H.A. et al. [52]
Bacteria	11	1	2014 Laurence, M. et al.[53]
Bacteria	92	0	2014 Salter, S.J. et al. [35]
Bacteria	7	0	2015 Jervis-Bardy, J. et al. [42]
Bacteria	28	0	2015 Jousselin, E. et al. [54]
Bacteria	77	127	2016 Glassing, A. et al.[36]
Bacteria	23	0	2016 Lauder, A.P. et al. [55]
Bacteria	6	0	2016 Lazarevic, V. et al. [56]
Bacteria	62	0	2017 Salter, S.J. et al. [57]
Bacteria	0	122	2018 Kirstahler, P. et al. [58]
Bacteria	34	0	2018 Stinson, L.F. et al. [59]
Bacteria	18	0	2019 Stinson, L.F. et al. [60]
Bacteria	52	2	2019 Weyrich, L.S. et al. [61]
Bacteria	8	26	2019 de Goffau, M.C. et al. [62]
Bacteria	15	93	2020 Nejman D. et al. [63]
Viruses	0	1	2015 Kjartansdóttir, K.R. et al. [64]
Viruses	0	1	2015 Mukherjee, S. et al. [65]
Viruses	0	291	2019 Asplund, M. et al. [66]
Eukaryota	0	3	2016 Czurda, S. et al. [67]
Eukaryota	0	1	PRJNA168
Total (unique)	210	627	-

Table 3: Summary of common contaminants taxa extracted from the literature. The complete list of taxa per study can be found in the GRIMER repository (`https://github.com/pirovc/grimer/`)

to link findings to common contaminants or connect analyses outcomes with known environments or biomes. Those entries can be easily provided by the user in a simple list of names or taxonomic identifiers in a formatted and annotated file (more information can be found in the GRIMER repository).

Contamination references

We compiled an extensive list of possible contaminant taxa reported in several studies (Table 3). The studies selected were obtained from cross-references in review articles [38] and individual selected findings in the literature, usually focusing on contamination detection or mitigation. Articles were manually curated and more studies can potentially be added to the list, which is dynamically maintained. Contributions are welcome through the GRIMER repository (https://github.com/pirovc/grimer/). The studies selected are very diverse in terms of sequencing technology, methodology used and environment studied. Contamination in those studies can originate from diverse sequencing kits and reagents 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 and should not be used alone to define contamination in a study. However, it serves as an additional evidence supporting it, especially if entries are highly recurrent (Table 4) and corroborate with additional lines of evidence. Those contaminants were reported mainly at genus or species level in different formats, names and taxonomies. We manually curated and converted them into the NCBI taxonomy [48] nomenclature for standardized usage.

Genus	# reported	Species	# reported
Pseudomonas	13	*Cutibacterium acnes*	4
Stenotrophomonas	13	*Pseudomonas fluorescens*	4
Ralstonia	12	*Stenotrophomonas maltophilia*	4
Bradyrhizobium	11	*Acinetobacter baumannii*	3
Methylobacterium	11	*Bradyrhizobium elkanii*	3
Acinetobacter	10	*Corynebacterium tuberculostearicum*	3
Corynebacterium	10	*Rhodococcus fascians*	3
Sphingomonas	10	*Streptococcus mitis*	3

Table 4: Top 8 most reported taxa from Table 3 at genus and species level. If multiple child nodes of organisms are reported in the same study, they are counted here just once.

Additionally, we compiled another 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 [38]. Reference organisms names were obtained from BacDive [68], eHOMD [69] and further publications [70].

MGnify

Additionally to the contamination references, a summary generated from MGnify repository [11] is provided with counts of occurrences for each observation in thousands of microbiome studies, grouped by biome. MGnify 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 provided 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 classifications from different sources were present, 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 of 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 occurrences). All five levels of biome classification are available for each taxonomic entry.

Input data

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 [71]. Further data can be provided to extend the report:

- Metadata: annotate samples and give further technical information. The metadata should be tabular and categorical and numerical fields are supported.

- Taxonomy: GRIMER will automatically parse a given taxonomic annotation or generate one based on the provided observations. Data will be summarized in many taxonomic levels and plots will be created accordingly. Taxonomy is fully automated for several commonly used taxonomies (NCBI, GTDB, SILVA, GreenGenes, OTT).

- Controls: one or more groups of control samples can be provided in a simple text file. Those samples will be further used to summarize data and annotate plots.

- References: custom sources of contamination or any 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 analyses:

- General data summary by observation and samples, linking references, taxonomy and metadata

- Filtering and transformation: observations and samples can be filtered to reduce noise or small counts. Transformations are applied (log, centered log-ratio, 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: Symmetric proportionality coefficient (rho correlation) [72, 73] is calculated for top abundant observations in the study

- DECONTAM [43]: 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 are 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.

- MGnify: Each taxa reported will be linked to the respective MGnify entry, reporting most common biome occurrences.

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 6), Samples (Figure 1), Heatmap (Figure 4), and Correlation (Figure 5). Some of them were previously suggested to be adequate for contamination detection

[45] and are commonly used in standard microbiome analysis. Every panel has one or more visualization and widgets to select, filter, group, and modify its contents. Panels can be reported independently.

An individual summary for each observation and its relation to annotations and distribution among samples can be found in the Overview panel (Figure 6). Here, all evidence 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 a bar plot. Information of annotations, MGnify biomes, and DECONTAM output are also available in the same interface. 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 containing that observation. 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 and summarized into taxonomic levels. The Overview panel also roughly summarizes samples contents in the bar plot, with general classification metrics. Those bars can be transformed, annotated, grouped and sorted to connect specific observation values to overall sample distribution.

In depth evaluation of individual samples can be performed in the Samples panel (Figure 1). Normalized distribution of top observations for each sample can be visualized in the bar plot to easily compare the overall distribution of observations among samples, with options for grouping and sorting by metadata. Automated selection of groups of samples is also possible by counts and metadata.

Several transformations can be applied to the data (normalization, log, center log ratio) to be further visualized in the Heatmap panel (Figure 4). Hierarchical clustering, grouping and sorting options can be independently selected for samples and observations to enable pattern detection (e.g. batch effects, treatment effects etc). Dendrograms are plotted when clustering options are selected. Annotation bars are plotted around the heatmap showing dynamically selection of sample annotations (metadata) and observation annotations (references, 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. Observation annotation values are normalized and plotted in the same color scale for easier interpretation. One heatmap is generated for each taxonomic level.

Correlation between observations are plotted as a matrix (Figure 5). Positive or negative correlations among observations can point towards concurrent 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 at similar ratios).

Implementation

GRIMER is written in Python and Javascript and outputs a report file in 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 tool-tips, showing more information about the data under the mouse cursor. Help buttons are also included, explaining the plots and options.

Further libraries were used to analyze samples and generate the report: pandas [74] for general parsing and data structures, scipy [75] for hierarchical clustering, 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 parse given taxonomies or download and convert any taxonomic id or name internally and decompose results in taxonomic ranks. Currently supported taxonomies are: NCBI, GTDB, Silva, GreenGenes, OpenTree Taxonomy. Reference lists are currently only available based on the NCBI Taxonomy.

Results

We re-analyzed publicly available studies to demonstrate the use of GRIMER reports in real case scenarios and what types of analyses are possible. In some examples, we try to reproduce analyses and in other cases point to new evidence that may have been overlooked. We encourage the readers to open the GRIMER reports (`https://doi.org/10.5281/zenodo.7525798` or `https://pirovc.github.io/grimer-reports/`) and interactively visualize the results being described to fully understand the capabilities of the report. All reports presented below were generated using GRIMER version 1.1.0.

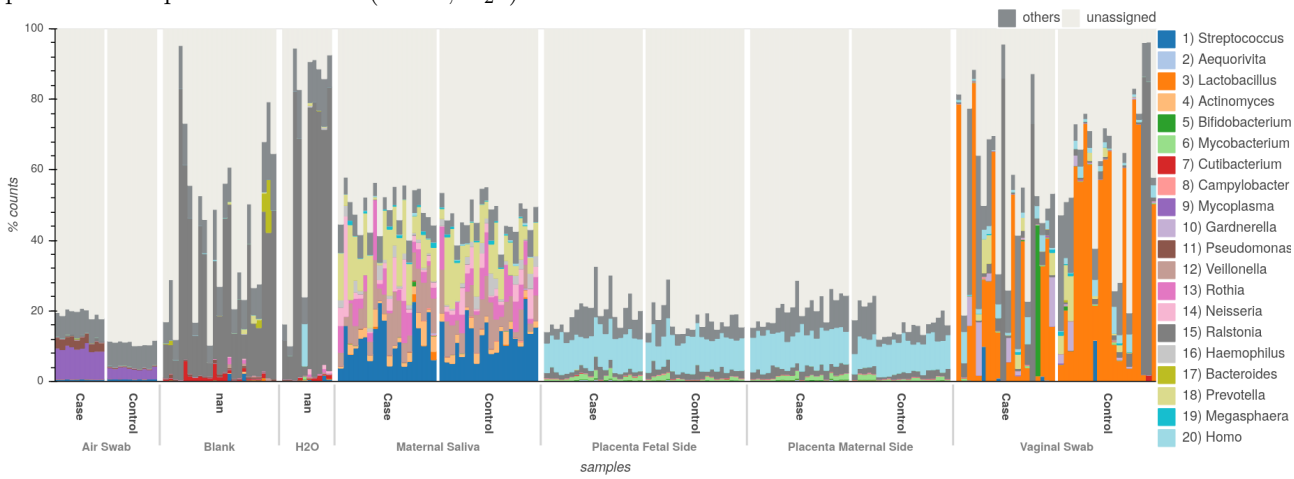
Detecting contamination

The attempt to detect and describe a possible human placental microbiome has motivated several studies and investigations [76, 47, 62, 77]. Leiby et al. [78] published a detailed and well designed study contributing to the subject. Placental samples for term (control) and preterm (case) newborns were collected for the maternal and fetal sides. 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 metagenomics (MGS). The authors could not distinguished a unique placental microbiome which differs from the contamination background. We re-analyzed the samples in a standard pipeline with QIIME2 [6] for amplicon data and ganon for MGS data [79], generated a GRIMER report for both and searched for the previously detected contamination.

In the MGS report, the bar plot (Figure 1) shows a stark difference in signal between sample types but a smaller difference in case and control groups. The *Ralstonia* genus is present in 96% of the all samples with an average abundance of 8.24%. Reads assigned for this genus were found in all negative control samples and H_20 samples. *Ralstonia* was also reported in 12 studies as a common contaminant, based on our compiled contaminant list (Table 4) and it was classified as a contaminant by the DECONTAM method, based on the correlation of frequencies and the total number of reads per 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 depicted in the Figure 2. Those results are inline with the ones reported in the original publication [78], even though the data was re-analyzed with a different set of tools, parameters and reference databases. All evidence describe pointing to *Ralstonia* as a contaminant was automatically generated by GRIMER and can be directly extracted from the Overview panel from the report. Besides human reads, *Ralstonia insidiosa* is the most prevalent species in this study. For the amplicon data, a similar pattern can be detected for the *Ralstonia* genus based on amplicon sequence variants (Figure 2).

Figure 1: Bar plot with relative abundance of top 20 genera for the placenta study. Bars are grouped by sample type and case/control. There is a stark difference in composition between air, vaginal and saliva samples to placental samples and controls (Blank, H_20).



Further, all other taxa present can easily be verified for the same patterns. *Pseudonomas* show similar distribution and was also reported originally as probable contaminants in the placental samples. *Corynebacterium*, *Cutibacterium* and *Mycobacterium*, although less prevalent, are further taxa with very similar patterns that could be potential contaminants and were not reported in the original publication.

Multiple microbiome studies exploration

Definitive and robust conclusions from low-biomass studied environment are only possible with a set of controls and protocols to deal with contamination. KatharoSeq [80] is a well designed protocol to better handle contamination in high-throughput low-biomass DNA microbial studies for amplicon sequencing or shotgun metagenomics. The protocol has guidelines for positive and negative controls implementation at the DNA extraction and library construction steps as well as computational approaches to define and exclude samples that did not achieve minimal amount of signal to be used. In their publication [80], the authors validate the protocol sequencing and analyzing with 3 low-biomass environments: the Jet Propulsion Laboratory spacecraft assembly facility (SAF), rooms of a neonatal intensive care unit (NICU) and an endangered-abalone-rearing facility (abalone). A set of negative controls to compare extraction kits is also included in the study (LBM).

We downloaded the OTU table and metadata from KatharoSeq evaluations for the 16S rRNA analyses available in Qiita [14] in the following configuration: reads trimmed at 150bp and classified using closed-

reference OTUs clustered at 97% similarity annotated with the greengenes taxonomy. A GRIMER report was generated for the raw table with all samples without any filtration. The heatmap generated for the annotated species level (Figure 3) shows a distinct and clear pattern between environments and the LBM. As reported in the publication, abalone samples have a higher richness (here as species annotated OTUs) as well as the highest average number of reads per sample. It is possible to identify potential contaminants in the study by looking for observations prevalent across environments and the relation to its annotations. Using this analysis we detected *Cutibacterium acnes* which is reported as a common contaminant and human-related species, present among all 4 environments studied as well as highly frequent in negative and positive controls. Even though DECONTAM did not identify this taxa as a contaminant, related data still hold strong evidence for contaminant of *C. acnes* in this study. Furthermore, *Staphylococcus aureus* and *Staphylococcus epidermidis*, known as human-related bacteria, were detected in high abundances in both NICU and SAF environments - areas with low and high human exposure, respectively. However, both species were also relatively highly present in negative controls, the abalone environment and LBM samples. Additionally, both were positively classified as contamination by DECONTAM, indicating that besides human exposure, those organisms could be driven by an external source of contamination.

Species identification based on 16S rRNA is limited due to its low resolution: approximately 15% of the OTUs are annotated at species level and 69% at genus level in this study. The same analysis visualized at genus level gives an increased perception of the distribution of the data in this study. With a higher signal it is possible to visualize how several clusters are formed and in many cases agree in multiple levels of evidence supporting the possibility of contamination (Figure 4).

Looking at the correlation between top observations reported (Figure 5), a matrix of highly correlated genera can be detected. Such a pattern was previously reported to be an indication of contamination from reagent-derived sources since they are invariably present within samples in similar ratios [45]. Further inspection of those genera (*Glaciecola*, *Leucothrix*, *Mycoplasma*, *Oleibacter*, *Polaribacter*, *Pseudoalteromonas*, *Psychrilyobacter*, *Psychromonas*, *Shewanella*) shows that they are mainly from Aquatic/Marine biomes with help of the matching results with the MGnify database (Figure 6 B). Further, they are more prevalent in negative controls (Figure 6 A), an evidence of DNA extraction kit or sample processing contaminants. Those organisms are highly frequent in the abalone study, which is a Marine environment and some of them were also described in the original publication. Although in very low amounts, those groups were also reported present in NICU samples (Figure 6 C), pointing to possible well-to-well contamination.

Discussion

GRIMER is an easy-to-use and accessible tool for specialists and non-specialists that generates a concise interactive offline dashboard with a set of analyses, 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 reports are a valuable resource for investigating contamination, a

problem that affects every microbiome study to some degree.

All the conclusion and visualizations presented in this work in the results section were solely based on GRIMER reports, showing that microbiome analysis, contamination investigation and detection are possible with the methodology proposed. The use of multiple sources of evidence to annotate observations improves the ability to better detect clear contaminants in microbiome studies as well as to point to probable groups of candidate contaminants.

In addition to the GRIMER software, we compiled and provided in this work a list of common taxa contaminants based on 22 publications (Table 3). Many of the reported contaminants are recurrent in diverse studies, pointing to a consensus for some taxa (Table 4) as a probable contaminant. Taxa in this list cannot be strictly considered a contaminant by itself. However, they can corroborate suspicious contamination discovered via several other lines of evidence without the extra effort of researching the literature. The presented list is not comprehensive but a 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. The information of common contaminants is a valuable resource to aid contamination detecting and we are willing to keep and extend it. Improvements to the list and suggestions of further candidate taxa can be provided via the GRIMER repository at `https://github.com/pirovc/grimer/`. As a future work, the list can be associated with study details as biome, extraction kit and methodology to be further queried and integrated in more details.

Additionally to the aforementioned common contaminants, GRIMER can also use general lists of custom organisms to annotate samples. In this work and by default, human-related organisms commonly occurring in human skin, oral and nasal cavities as well as face and other human limbs are used since they can be external sources of contamination. Those lists can be easily provided as taxonomic identifiers or names to GRIMER. If the target study conflicts with any of those environments (e.g. study of human skin), one could simply remove the related entries from the configuration files. More details and examples on how to perform this can be found in the online documentation.

GRIMER works out-of-the-box with as little data as possible but can incrementally expand the reports when more data is provided and can be adapted for user necessities. GRIMER is fast and generate reports in a matter of seconds on a standard notebook. The outcome dashboard is lightweight and can handle hundreds to thousands of samples and observations. Report sizes usually vary from 1-10MB and are highly compressible, since they are text-based HTML files. GRIMER reports with higher number of samples (thousands) can grow significantly in size (10-100MB) but still run normally. If report size is a limitation, many options can be adjusted to generated more compact files: reducing number of taxonomy ranks displayed, less combinations of analyses, filtering very low abundant observations, among others.

One of GRIMER core strengths is the taxonomy automation. It accepts taxonomic identifiers from several different taxonomies, but also parses names and converts them to their respective identifiers. If only one taxonomic level is provided (e.g. species level), GRIMER can decompose and summarize the data in higher ranks. That means that users do not have to handle taxonomy and everything will work automatically. GRIMER

Name	Reason	Year	Reference
Community-analyzer	website offline	2013	[81]
calypso	website offline	2017	[82]
Metaviz	web tool not responsive	2018	[83]
iMAP	no longer supported due to funding	2019	[84]
biomminer	page not found	2020	[85]

Table 5: Tools and web resources no longer available, supported or inaccessible (as of 2022-02-28)

was developed in a way that new visualizations can be included with little effort.

We listed and summarized a list of similar currently available methods published in the last 10 years (Table 2) as well as web-plataforms for complete analyses of microbiome data (Table 1). A list of functionalities between similar available tools is provided in [9] but a detailed comparison with GRIMER is out of the scope of this work. Most methods share some basic functions (e.g. taxonomic abundance analysis) but are diverse in many other aspects and were sometimes developed with specific goals (e.g. function analysis, biomarker identification). However, there is no comprehensive method that can provide a complete solution for the many possible analyses in a microbiome study. We believe that many of those tools, besides their overlapping functions, are complementary and can be used concurrently. GRIMER mainly shares features with pavian [22] in terms of general microbiome exploration and support to metagenomics data and with OpenContami [28] regarding contaminant detection. GRIMER, however, is unique in its output format. The vast majority of the currently available tools are web-based, hosted in a remote server or rely on a local hosted web-server to properly work (Table 2). This may be impractical for many non-specialists and for long term storage and reproducibility. GRIMER reports are portable and fully functional offline. This allows analysis to be accessible by many researchers with different backgrounds working together in the same study, increasing direct interaction with data. The portability also enables better documentation of results, reproducibility and shareability. Further, web-based tools may disappear after some years of inactivity or lack of funding and analysis may be lost, as it is the case for for some methods (Table 5). GRIMER reports are completely offline and will work as long as the report file is safely stored.

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

Declarations

Availability of data and materials

The datasets and metadata for the placenta study were obtained from:

https://www.ebi.ac.uk/ena/browser/view/PRJNA451186

The datasets and metadata for the KatharoSeq study were obtained from (log-in required):

https://qiita.ucsd.edu/study/description/10934

Competing interests

The authors declare no competing interests.

Funding

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

Authors' contributions

V.C.P and B.Y.R. conceptualized and developed the idea. V.C.P. wrote the software and main manuscript text, including all analyses. B.Y.R. reviewed and contributed to the manuscript text. All authors reviewed the manuscript.

Acknowledgements

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

Availability and requirements

Project name: GRIMER

Project home page: https://github.com/pirovc/grimer

Operating system(s): Platform independent

Programming language: Python 3.5 or higher

Other requirements: bokeh 2.2.3 or higher

License: MIT License

Any restrictions to use by non-academics: use based on MIT licence

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 Baldassano, 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] Masood ur Rehman Kayani, Wanqiu Huang, Ru Feng, and Lei Chen. Genome-resolved metagenomics using environmental and clinical samples. *Briefings in Bioinformatics*, (bbab030), March 2021.

[5] 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.

[6] Evan Bolyen, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian C. Abnet, Gabriel A. Al-Ghalith, Harriet Alexander, Eric J. Alm, Manimozhiyan Arumugam, Francesco Asnicar, Yang Bai, Jordan E. Bisanz, Kyle Bittinger, Asker Brejnrod, Colin J. Brislawn, C. Titus Brown, Benjamin J. Callahan, Andrés Mauricio Caraballo-Rodríguez, John Chase, Emily K. Cope, Ricardo Da Silva, Christian Diener, Pieter C. Dorrestein, Gavin M. Douglas, Daniel M. Durall, Claire Duvallet, Christian F. Edwardson, Madeleine Ernst, Mehrbod Estaki, Jennifer Fouquier, Julia M. Gauglitz, Sean M. Gibbons, Deanna L. Gibson, Antonio Gonzalez, Kestrel Gorlick, Jiarong Guo, Benjamin Hillmann, Susan Holmes, Hannes Holste, Curtis Huttenhower, Gavin A. Huttley, Stefan Janssen, Alan K. Jarmusch, Lingjing Jiang, Benjamin D. Kaehler, Kyo Bin Kang, Christopher R. Keefe, Paul Keim, Scott T. Kelley, Dan Knights, Irina Koester, Tomasz Kosciolek, Jorden Kreps, Morgan G. I. Langille, Joslynn Lee, Ruth Ley, Yong-Xin Liu, Erikka Loftfield, Catherine Lozupone, Massoud Maher, Clarisse Marotz, Bryan D. Martin, Daniel McDonald, Lauren J. McIver, Alexey V. Melnik, Jessica L. Metcalf, Sydney C. Morgan, Jamie T. Morton, Ahmad Turan Naimey, Jose A. Navas-Molina, Louis Felix Nothias, Stephanie B. Orchanian, Talima Pearson, Samuel L. Peoples, Daniel Petras, Mary Lai Preuss, Elmar Pruesse, Lasse Buur Rasmussen, Adam Rivers, Michael S. Robeson, Patrick Rosenthal, Nicola Segata, Michael Shaffer, Arron Shiffer, Rashmi Sinha, Se Jin Song, John R. Spear, Austin D. Swafford, Luke R. Thompson, Pedro J. Torres, Pauline Trinh, Anupriya Tripathi, Peter J. Turnbaugh, Sabah Ul-Hasan, Justin J. J. van der Hooft, Fernando Vargas, Yoshiki Vázquez-Baeza, Emily Vogtmann, Max von Hippel, William Walters, Yunhu Wan, Mingxun Wang, Jonathan Warren, Kyle C. Weber, Charles H. D. Williamson, Amy D. Willis, Zhenjiang Zech Xu, Jesse R. Zaneveld, Yilong Zhang, Qiyun Zhu, Rob Knight, and J. Gregory Caporaso. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, 37(8):852–857, August 2019.

[7] Patrick D. Schloss, Sarah L. Westcott, Thomas Ryabin, Justine R. Hall, Martin Hartmann, Emily B. Hollister, Ryan A. Lesniewski, Brian B. Oakley, Donovan H. Parks, Courtney J. Robinson, Jason W. Sahl, Blaz Stres, Gerhard G. Thallinger, David J. Van Horn, and Carolyn F. Weber. Introducing mothur: Open-

Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*, 75(23):7537–7541, December 2009.

[8] Paul J. McMurdie and Susan Holmes. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLOS ONE*, 8(4):e61217, April 2013. Publisher: Public Library of Science.

[9] Jannes Peeters, Olivier Thas, Ziv Shkedy, Leyla Kodalci, Connie Musisi, Olajumoke Evangelina Owokotomo, Aleksandra Dyczko, Ibrahim Hamad, Jaco Vangronsveld, Markus Kleinewietfeld, Sofie Thijs, and Jan Aerts. Exploring the Microbiome Analysis and Visualization Landscape. *Frontiers in Bioinformatics*, 1, 2021.

[10] Folker Meyer, Saurabh Bagchi, Somali Chaterji, Wolfgang Gerlach, Ananth Grama, Travis Harrison, Tobias Paczian, William L Trimble, and Andreas Wilke. MG-RAST version 4—lessons learned from a decade of low-budget ultra-high-throughput metagenome analysis. *Briefings in Bioinformatics*, 20(4):1151–1159, July 2019.

[11] 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.

[12] 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.

[13] Nick Weber, David Liou, Jennifer Dommer, Philip MacMenamin, Mariam Quiñones, Ian Misner, Andrew J Oler, Joe Wan, Lewis Kim, Meghan Coakley McCarthy, Samuel Ezeji, Karlynn Noble, and Darrell E Hurt. Nephele: a cloud platform for simplified, standardized and reproducible microbiome data analysis. *Bioinformatics*, 34(8):1411–1413, April 2018.

[14] Antonio Gonzalez, Jose A. Navas-Molina, Tomasz Kosciolek, Daniel McDonald, Yoshiki Vázquez-Baeza, Gail Ackermann, Jeff DeReus, Stefan Janssen, Austin D. Swafford, Stephanie B. Orchanian, Jon G. Sanders, Joshua Shorenstein, Hannes Holste, Semar Petrus, Adam Robbins-Pianka, Colin J. Brislawn, Mingxun Wang, Jai Ram Rideout, Evan Bolyen, Matthew Dillon, J. Gregory Caporaso, Pieter C. Dorrestein, and Rob Knight. Qiita: rapid, web-enabled microbiome meta-analysis. *Nature Methods*, 15(10):796–798, October 2018. Number: 10 Publisher: Nature Publishing Group.

[15] David Arndt, Jianguo Xia, Yifeng Liu, You Zhou, An Chi Guo, Joseph A. Cruz, Igor Sinelnikov, Karen Budwill, Camilla L. Nesbø, and David S. Wishart. METAGENassist: a comprehensive web server for comparative metagenomics. *Nucleic Acids Research*, 40(W1):W88–W95, July 2012.

[16] Susan M. Huse, David B. Mark Welch, Andy Voorhis, Anna Shipunova, Hilary G. Morrison, A. Murat Eren, and Mitchell L. Sogin. VAMPS: a website for visualization and analysis of microbial population structures. *BMC Bioinformatics*, 15(1):41, February 2014.

[17] Paul J. McMurdie and Susan Holmes. Shiny-phyloseq: Web application for interactive microbiome analysis with provenance tracking. *Bioinformatics*, 31(2):282–283, January 2015.

[18] Yi Wang, Ling Xu, Yong Q. Gu, and Devin Coleman-Derr. MetaCoMET: a web platform for discovery and visualization of the core microbiome. *Bioinformatics*, 32(22):3469–3470, November 2016.

[19] Cedric C. Laczny, Christina Kiefer, Valentina Galata, Tobias Fehlmann, Christina Backes, and Andreas Keller. BusyBee Web: metagenomic data analysis by bootstrapped supervised binning and annotation. *Nucleic Acids Research*, pages 1–9, 2017.

[20] Achal Dhariwal, Jasmine Chong, Salam Habib, Irah L. King, Luis B. Agellon, and Jianguo Xia. MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Research*, 45(W1):W180–W188, July 2017.

[21] Colin P. McNally, Alexander Eng, Cecilia Noecker, William C. Gagne-Maynard, and Elhanan Borenstein. BURRITO: An Interactive Multi-Omic Tool for Visualizing Taxa–Function Relationships in Microbiome Data. *Frontiers in Microbiology*, 9, 2018.

[22] Florian P. Breitwieser and Steven L. Salzberg. Pavian: interactive analysis of metagenomics data for microbiome studies and pathogen identification. *Bioinformatics*, 2019.

[23] W. M. Tong and Yuki Chan. GenePiper, a Graphical User Interface Tool for Microbiome Sequence Data Mining. *Microbiology Resource Announcements*, January 2020. Publisher: American Society for Microbiology 1752 N St., N.W., Washington, DC.

[24] Yue Zhao, Anthony Federico, Tyler Faits, Solaiappan Manimaran, Daniel Segrè, Stefano Monti, and W. Evan Johnson. animalcules: interactive microbiome analytics and visualization in R. *Microbiome*, 9(1):76, March 2021.

[25] Janina Reeder, Mo Huang, Joshua S. Kaminker, and Joseph N. Paulson. MicrobiomeExplorer: an R package for the analysis and visualization of microbial communities. *Bioinformatics (Oxford, England)*, 37(9):1317–1318, June 2021.

[26] David J. m Barnett, Ilja C. w Arts, and John Penders. microViz: an R package for microbiome data visualization and statistics. *Journal of Open Source Software*, 6(63):3201, July 2021.

[27] Alexander Dietrich, Monica Steffi Matchado, Maximilian Zwiebel, Benjamin Ölke, Michael Lauber, Ilias Lagkouvardos, Jan Baumbach, Dirk Haller, Beate Brandl, Thomas Skurk, Hans Hauner, Sandra Reitmeier, and Markus List. Namco: A microbiome explorer. *bioRxiv*, page 2021.12.15.471754, December 2021. Section: New Results Type: article.

[28] Sung-Joon Park and Kenta Nakai. OpenContami: A web-based application for detecting microbial contaminants in next-generation sequencing data. *Bioinformatics*, (btab101), February 2021.

[29] Shih-Chi Su, James E. Galvin, Shun-Fa Yang, Wen-Hung Chung, and Lun-Ching Chang. wiSDOM: a visual and statistical analytics for interrogating microbiome. *Bioinformatics*, 37(17):2795–2797, September 2021.

[30] Boyang Tom Jin, Feng Xu, Raymond T Ng, and James C Hogg. Mian: interactive web-based microbiome data table visualization and machine learning platform. *Bioinformatics*, 38(4):1176–1178, February 2022.

[31] 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.

[32] 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.

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

[34] Kenneth H. Rand and Herbert Houck. Taq polymerase contains bacterial DNA of unknown origin. *Molecular and Cellular Probes*, 4(6):445–450, December 1990.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

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

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] Joshua G. Harrison, Gregory D. Randolph, and C. Alex Buerkle. Characterizing Microbiomes via Sequencing of Marker Loci: Techniques To Improve Throughput, Account for Cross-Contamination, and Reduce Cost. *mSystems*, 0(0):e00294–21, 2021. Publisher: American Society for Microbiology.

[47] 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.

[48] 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.

[49] 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.

[50] Leonid A. Kulakov, Morven B. McAlister, Kimberly L. Ogden, Michael J. Larkin, and John F. O'Hanlon. Analysis of Bacteria Contaminating Ultrapure Water in Industrial Systems. *Applied and Environmental Microbiology*, 68(4):1548–1555, April 2002. Publisher: American Society for Microbiology Section: MICROBIAL ECOLOGY.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] Lisa F. Stinson, Jeffrey A. Keelan, and Matthew S. Payne. Comparison of Meconium DNA Extraction Methods for Use in Microbiome Studies. *Frontiers in Microbiology*, 9:270, February 2018.

[60] L. F. Stinson, J. A. Keelan, and M. S. Payne. Identification and removal of contaminating microbial DNA from PCR reagents: impact on low-biomass microbiome analyses. *Letters in Applied Microbiology*, 68(1):2–8, January 2019.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] S. Czurda, S. Smelik, S. Preuner-Stix, F. Nogueira, and T. Lion. Occurrence of Fungal DNA Contamination in PCR Reagents: Approaches to Control and Decontamination. *Journal of Clinical Microbiology*, 54(1):148–152, January 2016.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] David Lovell, Vera Pawlowsky-Glahn, Juan José Egozcue, Samuel Marguerat, and Jürg Bähler. Proportionality: A Valid Alternative to Correlation for Relative Data. *PLOS Computational Biology*, 11(3):e1004075, March 2015. Publisher: Public Library of Science.

[73] Ionas Erb and Cedric Notredame. How should we measure proportionality on relative gene expression data? *Theory in Biosciences*, 135(1-2):21–36, June 2016.

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

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] Vitor C Piro, Temesgen H Dadi, Enrico Seiler, Knut Reinert, and Bernhard Y Renard. ganon: precise metagenomics classification against large and up-to-date sets of reference sequences. *Bioinformatics*, 36(Supplement_1):i12–i20, July 2020.

[80] Jeremiah J. Minich, Qiyun Zhu, Stefan Janssen, Ryan Hendrickson, Amnon Amir, Russ Vetter, John Hyde, Megan M. Doty, Kristina Stillwell, James Benardini, Jae H. Kim, Eric E. Allen, Kasthuri Venkateswaran, and Rob Knight. KatharoSeq Enables High-Throughput Microbiome Analysis from Low-Biomass Samples. *mSystems*, 3(3), June 2018. Publisher: American Society for Microbiology Journals Section: Research Article.

[81] Bhusan K. Kuntal, Tarini Shankar Ghosh, and Sharmila S. Mande. Community-analyzer: a platform for visualizing and comparing microbial community structure across microbiomes. *Genomics*, 102(4):409–418, October 2013.

[82] Martha Zakrzewski, Carla Proietti, Jonathan J Ellis, Shihab Hasan, Marie-Jo Brion, Bernard Berger, and Lutz Krause. Calypso: a user-friendly web-server for mining and visualizing microbiome–environment interactions. *Bioinformatics*, 33(5):782–783, March 2017.

[83] Justin Wagner, Florin Chelaru, Jayaram Kancherla, Joseph N Paulson, Alexander Zhang, Victor Felix, Anup Mahurkar, Niklas Elmqvist, and Héctor Corrada Bravo. Metaviz: interactive statistical and visual analysis of metagenomic data. *Nucleic Acids Research*, 46(6):2777–2787, April 2018.

[84] Teresia M. Buza, Triza Tonui, Francesca Stomeo, Christian Tiambo, Robab Katani, Megan Schilling, Beatus Lyimo, Paul Gwakisa, Isabella M. Cattadori, Joram Buza, and Vivek Kapur. iMAP: an integrated

bioinformatics and visualization pipeline for microbiome data analysis. *BMC Bioinformatics*, 20(1):374, July 2019.

[85] Amirhossein Shamsaddini, Kimia Dadkhah, and Patrick M. Gillevet. BiomMiner: An advanced exploratory microbiome analysis and visualization pipeline. *PLOS ONE*, 15(6):e0234860, June 2020. Publisher: Public Library of Science.

Figure 2: Evidence supporting *Ralstonia* as a contaminant taxa in the placenta study. a-b) Right y-axis shows normalized abundance of genus *Ralstonia* in log scale for each sample in the MGS (a) and amplicon (b) data. Bars (left y-axis) summarized counts at genus level for each sample. Samples are grouped by sample type and sorted by number of reads (x-axis). The yellow circles show abundance of *Ralstonia*, which is higher in the control samples (Blank and H_20) as well as increased in real samples with low read count. c-d) DECONTAM plots for *Ralstonia* genus for the MGS (c) and amplicon (d) data. DECONTAM plots show that taxa counts follow the expected distribution for contamination based on the number of reads per sample (red line).

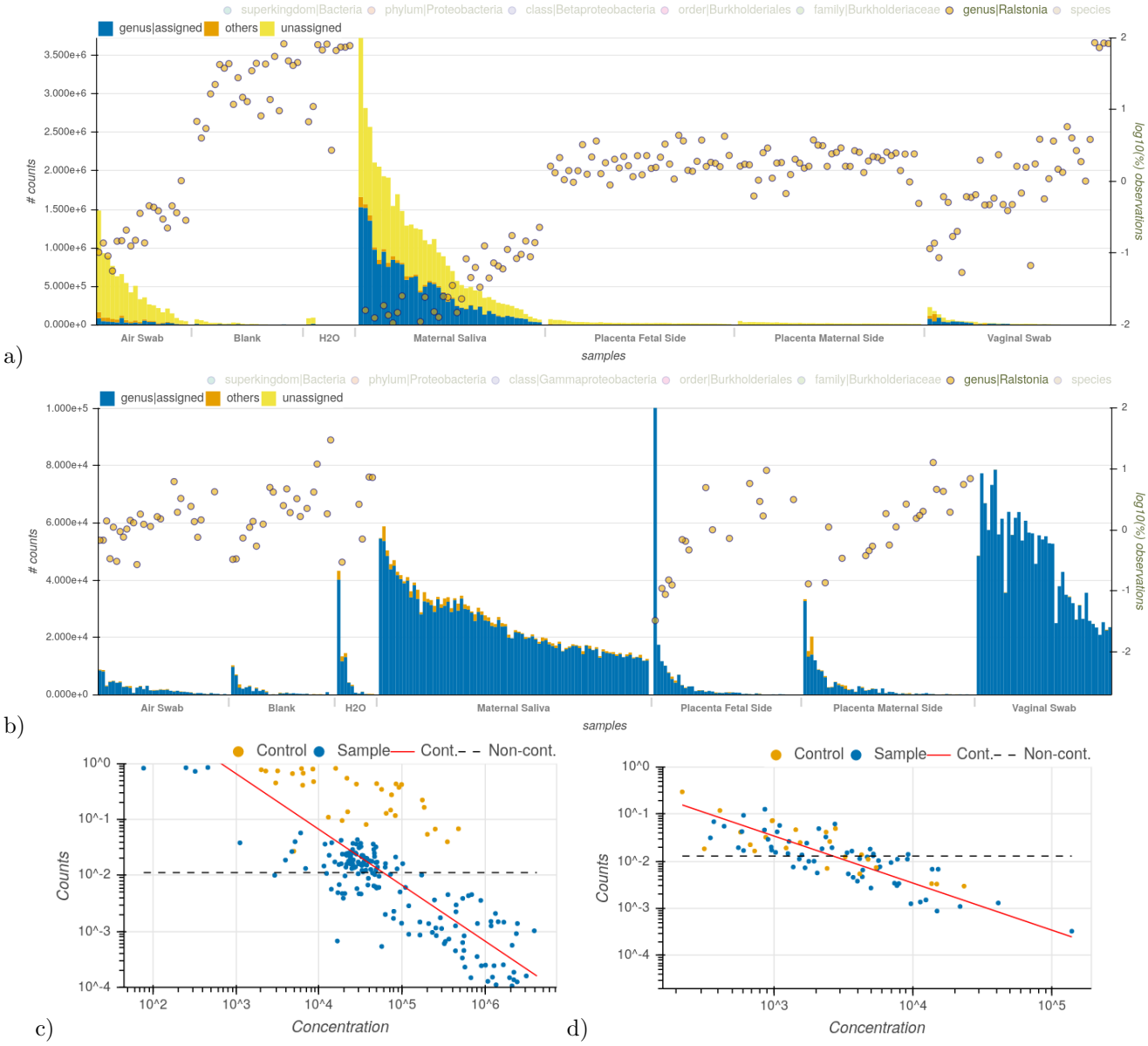


Figure 3: Heatmap visualization at species level for the KatharoSeq data. Samples are grouped by study type (y-axis) and clustered by observations (x-axis, euclidean distance metric, complete method). Data in the heatmap is center-log ratio transformed. Bottom panel show annotation related to the observations. "Contaminants" and "Human-related" annotations are normalized counts against pre-compiled list of references described in this paper. "decontam" is the normalized DECONTAM p-score. All "control" annotations show the proportion of the observation in the indicated group of control samples.

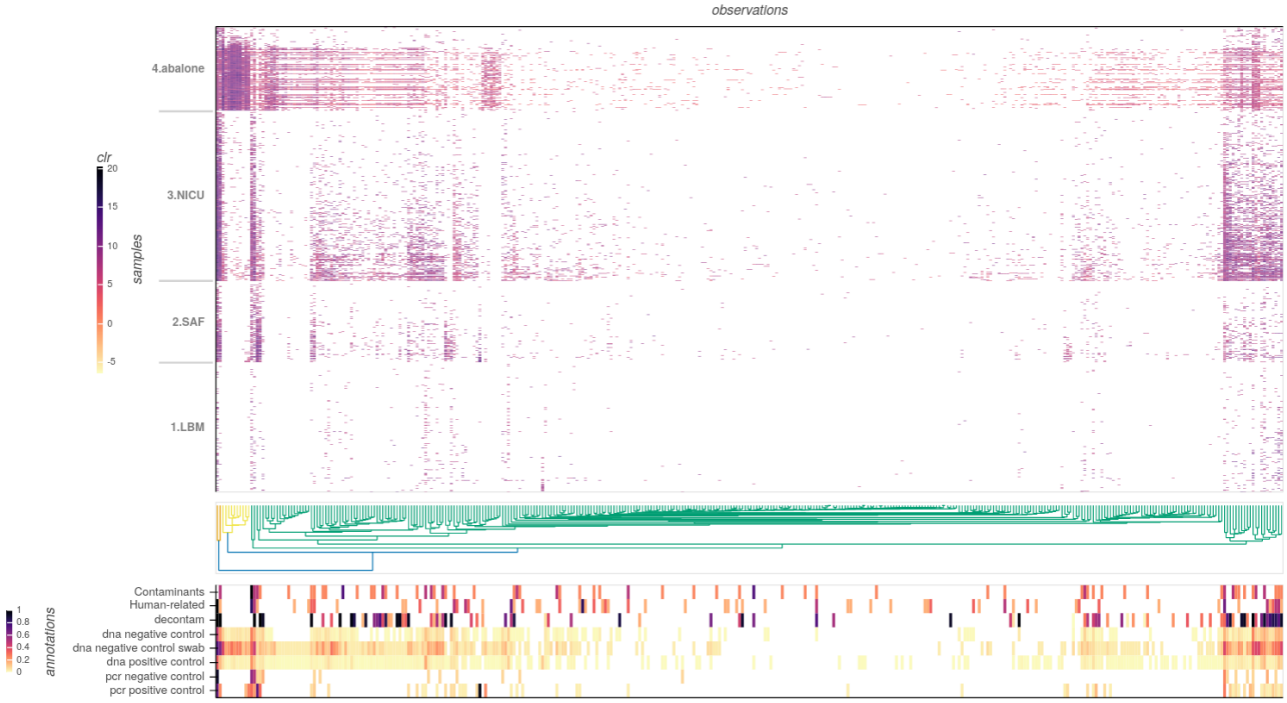


Figure 4: Heatmap visualization at genus level for the KatharoSeq data. Samples and observations axis are clustered and sorted based on the euclidean distance metric, complete method. Data in the heatmap is center-log ratio transformed. Bottom panel show annotation related to the observations. "Contaminants" and "Human-related" annotations are normalized counts against pre-compiled list of references described in this paper. "decontam" is the normalized DECONTAM p-score. All "control" annotations show the proportion of the observation in the indicated group of control samples. Metadata panel show color-coded sample information on study (md_title) and type of sample (md_control_verbose). The annotation panel shows higher values on multiple sources of evidence for contamination relative to data clusters of the heatmap. Metadata panel shows how samples show independent patterns based on the environment (md_title) and difference from controls (md_control_verbose).

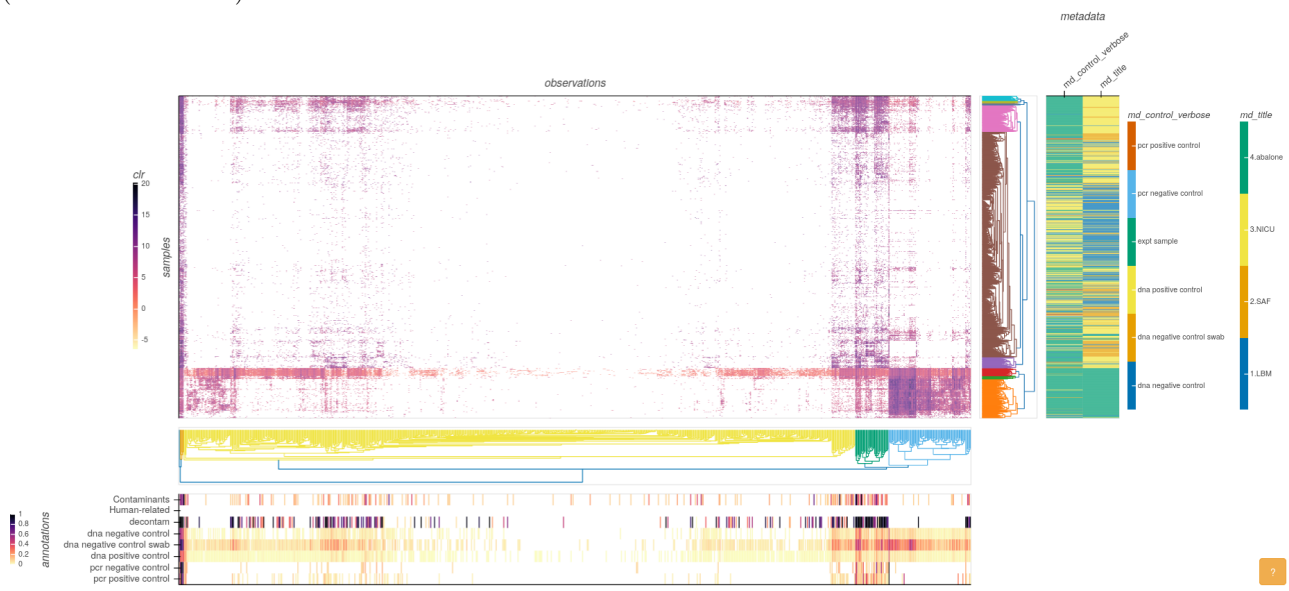


Figure 5: Symmetric proportionality coefficient (rho correlation) between top 50 most abundant genera in the KatharoSeq data. Positive correlation values (between 0 and 1) are displayed in red. Negative correlation values (between -1 and 0) are displayed in blue. Highly correlated matrix among 9 genera (dark red) points to reagent-derived contamination, when considered with other lines of evidence (Figure 6)

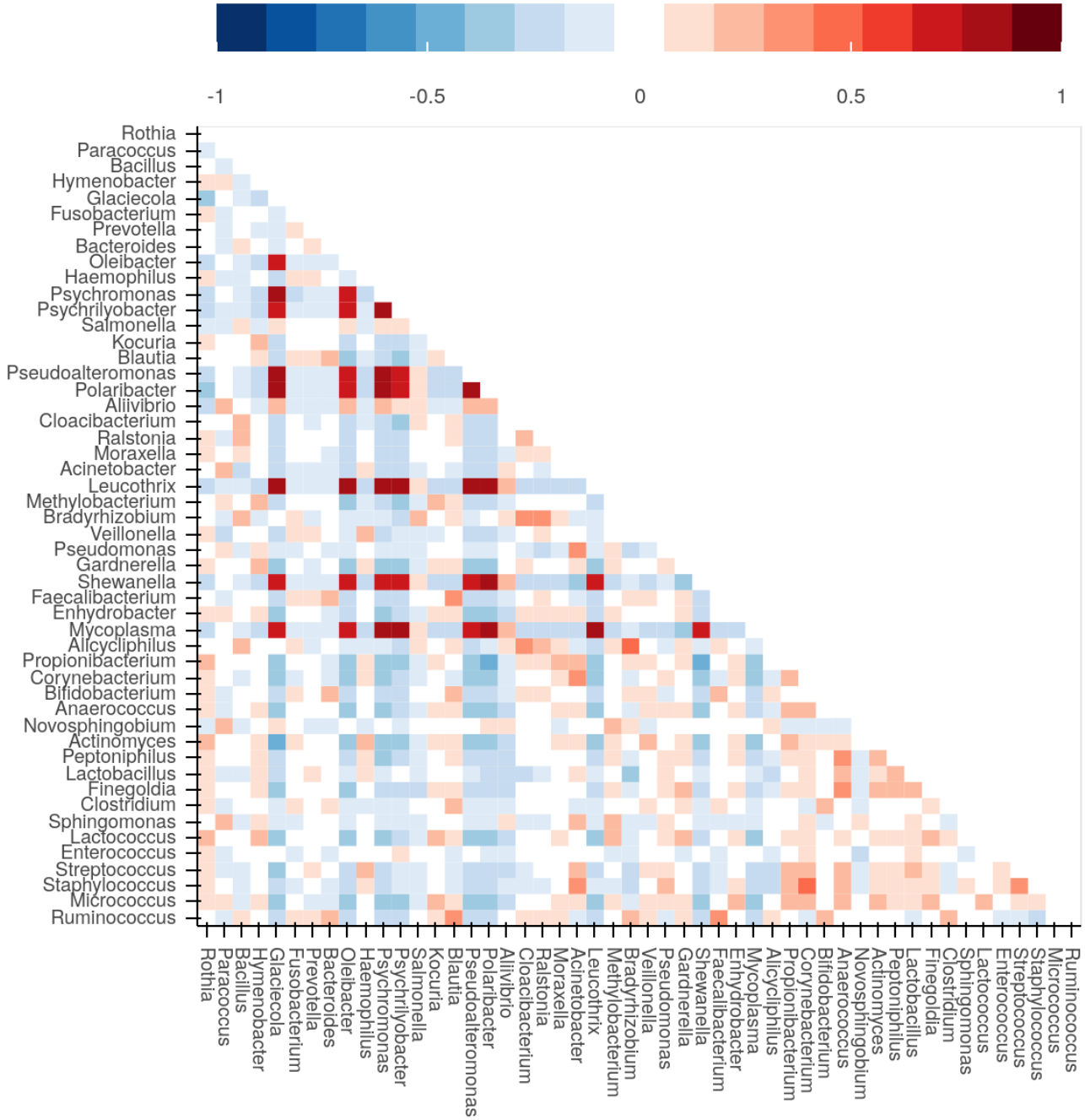
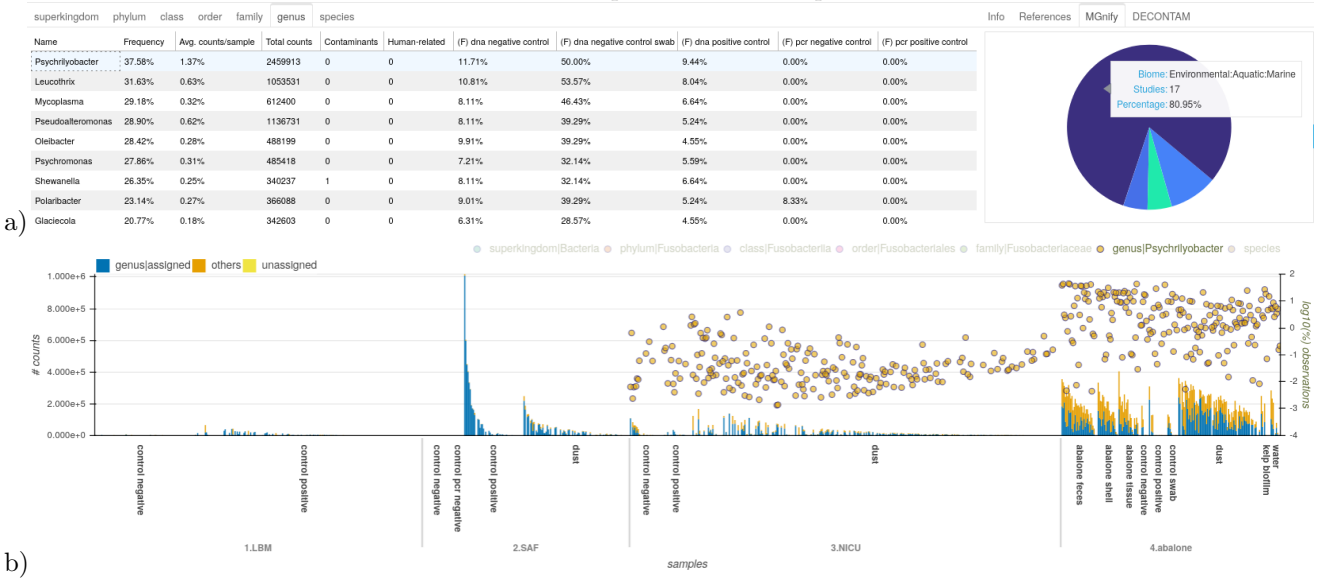


Figure 6: GRIMER Overview panel plots a) listing of 9 highly correlated genera detected in Figure 5. Samples have high incidence in DNA negative controls. MGnify plot showing proportion of biomes related to *Psychrilyobacter* in the whole MGnify database b) bar plot listing samples (x-axis), grouped by study and sample type and sorted by total number of reads. Bars represent the total number of counts for each sample and are annotated with the proportion assigned to genus level (left y-axis). Log-transformed abundance of *Psychrilyobacter* is displayed in yellow circles (right y-axis). This taxon is abundant in the abalone samples but has some signal in the NICU samples that are inversely correlated to the total amount of reads, pointing to potential contamination. The other 8 taxa from show similar patterns in the report.



Name	Frequency	Avg. counts/sample	Total counts	Contaminants	Human-related	(F) dna negative control	(F) dna negative control swab	(F) dna positive control	(F) pcr negative control	(F) pcr positive control
Psychrilyobacter	37.58%	1.37%	2459913	0	0	11.71%	50.00%	9.44%	0.00%	0.00%
Leucothrix	31.63%	0.63%	1053531	0	0	10.81%	53.57%	8.04%	0.00%	0.00%
Mycoplasma	29.18%	0.32%	612400	0	0	8.11%	46.43%	6.64%	0.00%	0.00%
Pseudoalteromonas	28.90%	0.62%	1136731	0	0	8.11%	39.29%	5.24%	0.00%	0.00%
Oleibacter	28.42%	0.28%	488199	0	0	9.91%	39.29%	4.55%	0.00%	0.00%
Psychromonas	27.86%	0.31%	485418	0	0	7.21%	32.14%	5.59%	0.00%	0.00%
Shewanella	26.35%	0.25%	340237	1	0	8.11%	32.14%	6.64%	0.00%	0.00%
Polaribacter	23.14%	0.27%	366088	0	0	9.01%	39.29%	5.24%	8.33%	0.00%
Glaciecola	20.77%	0.18%	342603	0	0	6.31%	28.57%	4.55%	0.00%	0.00%