Published online 16 June 2023

Nucleic Acids Research, 2023, Vol. 51, No. 13 **6593–6608**
https://doi.org/10.1093/nar/gkad527

Achieving pan-microbiome biological insights via the dbBact knowledge base

Amnon Amir [1,*,†], Eitan Ozel[2], Yael Haberman [3] and Noam Shental [2,*,†]

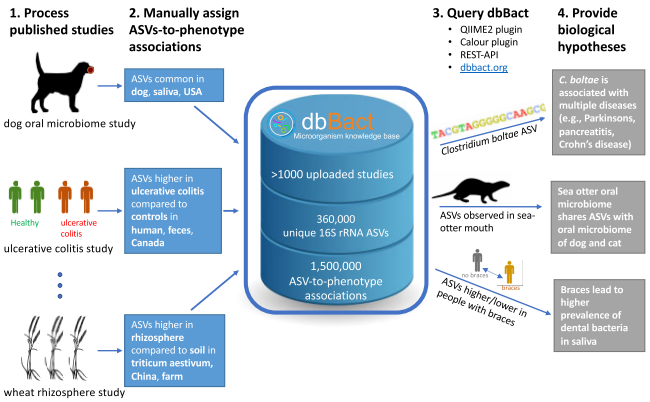
[1]Microbiome center, Sheba Medical Center, Israel, [2]Dept. of Computer Science, The Open University of Israel, Israel and [3]Pediatric Gastroenterology, Hepatology and Nutrition Unit, Sheba Medical Center, Israel

Received February 21, 2023; Revised May 26, 2023; Editorial Decision May 27, 2023; Accepted June 08, 2023

ABSTRACT

16S rRNA amplicon sequencing provides a relatively inexpensive culture-independent method for studying microbial communities. Although thousands of such studies have examined diverse habitats, it is difficult for researchers to use this vast trove of experiments when interpreting their own findings in a broader context. To bridge this gap, we introduce dbBact – a novel pan-microbiome resource. dbBact combines manually curated information from studies across diverse habitats, creating a collaborative central repository of 16S rRNA amplicon sequence variants (ASVs), which are assigned multiple ontology-based terms. To date dbBact contains information from more than 1000 studies, which include 1500000 associations between 360000 ASVs and 6500 ontology terms. Importantly, dbBact offers a set of computational tools allowing users to easily query their own datasets against the database. To demonstrate how dbBact augments standard microbiome analysis we selected 16 published papers, and reanalyzed their data via dbBact. We uncovered novel inter-host similarities, potential intra-host sources of bacteria, commonalities across different diseases and lower host-specificity in disease-associated bacteria. We also demonstrate the ability to detect environmental sources, reagent-borne contaminants, and identify potential cross-sample contaminations. These analyses demonstrate how combining information across multiple studies and over diverse habitats leads to better understanding of underlying biological processes.

GRAPHICAL ABSTRACT



INTRODUCTION

Bacteria play an important role in the Earth's ecosystem, having a total biomass higher than that of all vertebrates and fish, second only to plants (1). The introduction of 16S rRNA amplicon sequencing provides relatively cheap and accurate microbial profiling, which has been massively used to examine microbial populations in oceans (2), soil (3), plants (4), animals (5), and for performing cross-sectional human studies (6–8).

These large and diverse studies, in conjunction with reporting guidelines (9), may potentially allow us, for the first time, to conduct comparisons between microbial ecosystems across multiple studies, and also over different habitats. However, such comparisons require a designated infrastructure that is currently unavailable. The complexity of bacterial populations ranges from tens of different bacteria in a saliva sample (10), to thousands in a soil sample (11). Comparisons among bacterial populations across studies cannot rely on textual search engines, as the number of taxonomic names is much smaller than the number of bacteria (e.g. there are ∼300000 unique 16S rRNA sequences of length 90nt in the Earth Microbiome Project (11) compared to ∼4000 unique genera and 20000 unique species in NCBITAX (12) and in the Encyclopedia of Life

*To whom correspondence should be addressed. Email: amnonim@gmail.com
Correspondence may also be addressed to Noam Shental. Email: shental@openu.ac.il
†Joint Authors.

(13)). Moreover, grouping bacteria in higher taxonomic levels may not always maintain the basic habitat properties (11). Therefore, cross-study comparisons should directly rely on 16S rRNA amplicon sequences. Recently, several denoising methods (14–16) have been used to derive amplicon sequence variants (ASVs), which are database independent, and therefore provide a useful vocabulary for comparing bacteria across studies. Hence, a single bacterial species present in different studies will result in the same ASV, even when processed separately and denoised using different methods (14,17).

To facilitate cross-sectional comparisons of microbial communities across studies and habitats, we introduce dbBact - a wiki-like, manually curated microbial resource, which currently includes over 1000 individual studies, from which genotype-phenotype associations have been extracted, thus directly linking ASVs to various phenotypes, using ontology terms of interest (e.g. 'the relative abundance of ASV 'ACTGGA…' was higher in fecal samples of children with inflammatory bowel disease compared to healthy controls in California'; see example in Figure 1). To date, dbBact includes 1500000 such associations between 360000 ASVs and 6500 ontology terms. dbBact supports two main queries: (a) searching for ontology terms enriched for a given set of ASVs - allowing users to gain insights regarding the biology associated with their ASVs of interest (Figure 2A); and (b) identification of enriched ontology terms when contrasting two sets of ASVs (Figure 2B), analogous to gene enrichment analysis (18,19).

The idea of collecting information about bacteria from multiple experiments is not new, and several microbial repositories and databases currently exist, e.g. Qiita (20), redbiom (21), MGnify (22), GutMDisorder (23), BugSigDB (24) and Disbiome database (25). However, dbBact provides a combination of several key aspects that to the best of our knowledge are not available together in any other resource: (a) *Manual annotation*: genotype-phenotype associations for each study are manually curated by human experts that understand the experimental setting and therefore can identify bacteria related to the different phenotypic groups. In contrast, other microbial data repositories (8,20–22,26–28) merely provide raw experimental data and metadata. (b) *Unlimited scope*: dbBact includes data from multiple habitats, unlike databases that are highly limited in scope (23–25,29,30), or are context-specific (31,32). (c) *Size and extendibility*: With over 1000 manually curated studies, the number of studies in dbBact is currently ∼40% higher than in Qiita, and the number of ASVs is comparable to that in the Earth Microbiome Project. dbBact continues to grow via addition of new studies by the dbBact team as well as contributions by researchers, using a simple user-friendly interface for uploading new studies into the database. (d) *Structured genotype/phenotype search*: observations are uploaded at the ASV level, allowing queries of specific subsequences. In addition, as phenotypes are derived from multiple hierarchical ontologies, querying allows for 'cross-sectioning' of the data. For example, ASVs associated with Crohn's disease and ulcerative colitis will be recalled when querying their 'parent' *term,* 'inflammatory bowel disease.' (e) *Harmonizing studies performed using different variable regions*: as uploaded studies may be sequenced using different 16S rRNA variable regions, cross-region queries are facilitated by 'linking' the ASVs through the SILVA database of full-length 16S rRNA genes (33). (f) *Data analysis tools*: dbBact provides a set of statistical tools for analyzing new datasets and for generating novel biological hypotheses.

In the following sections we demonstrate the usability of dbBact by re-analyzing data from published papers via dbBact and present novel, and sometimes unexpected, biological insights regarding individual bacteria and similarities among bacterial communities. The dbBact resource is publicly available at http://dbbact.org and can also be accessed via QIIME2 (34), Calour (35), and the dbBact REST-API interface.

MATERIAL AND METHODS

As dbBact is constantly growing in scope, and to facilitate reproducibility, the dbBact infrastructure described in this section and all analyses presented in the paper carried out using dbBact release 2022.07.01, available for download as part of the weekly snapshots at https://dbbact.org/download.

Nomenclature

Throughout the paper, specific italicized words are used in a dbBact context as follows:

Experiment: is a single 16S rRNA amplicon study that was uploaded to dbBact. Each *experiment* is linked to its corresponding paper and accession number (e.g. in SRA or Qiita).

Sequence: is an amplicon sequence variant (ASV) that has been uploaded to dbBact. The ASV originates from one of the supported dbBact 16S rRNA variable regions, having a minimal length of 100nt.

Term: is an ontology-derived phenotype used to describe a set of samples (e.g. 'feces', 'ulcerative colitis', 'horse').

Annotation: is a biological observation linking *sequences* with a set of *terms* in a given *experiment*. For example, an *annotation* may associate a set of *sequences* whose relative abundance is higher in samples from horses suffering from ulcerative colitis compared to healthy controls in a specific *experiment*.

Implementation

Database. The dbBact database is stored as a SQL relational database (PostgreSQL 9.5.10). The database schema and detailed table descriptions are provided in Supplementary File 1 and Figures S22, S23.

Ontologies. Table S1 presents ontologies available in dbBact release 2022.07.01. dbBact supports the addition of ontologies to allow more accurate annotations. When users provide *terms* that do not appear in any of these ontologies, a new *term* is automatically added to the generic dbBact ontology.

ASV sequences

Primers and trimming. dbBact uses exact prefix search for *sequence* identification, and therefore all *sequences*

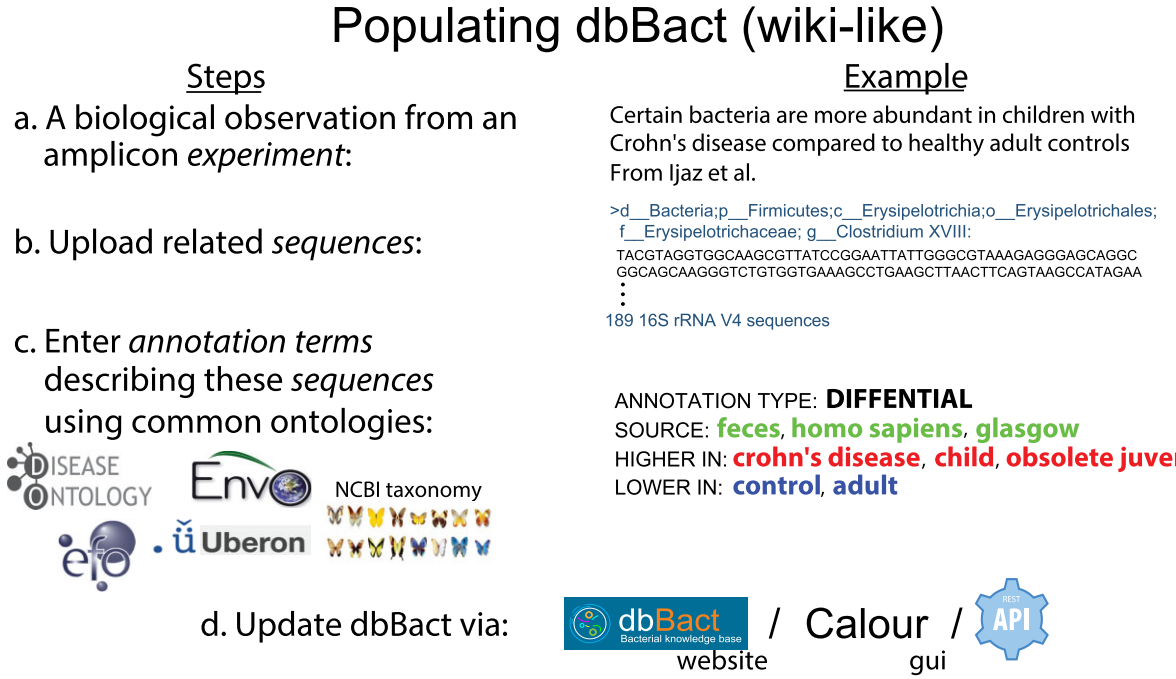


Figure 1. Steps in adding entries to dbBact. Users add new entries in a wiki-like way, by uploading study results. **a.** For example, analyzing data from Ijaz et al. (36), we identified 189 ASVs that are more abundant in fecal samples of Scottish children with Crohn's disease compared to healthy controls (see Methods section). **b.** These ASVs are uploaded as a FASTA file. **c.** Associations between ASVs and phenotypes are called *annotations*, which are created by assigning a set of ontology *terms* and predicates that characterize the context. The 189 ASVs were *annotated* as 'DIFFERENTIAL,' i.e. having a higher relative abundance in children with Crohn's disease ('HIGHER IN' *terms*), compared to healthy controls ('LOWER IN' *terms*). The general background *terms* common to both groups, i.e. 'homo sapiens,' 'feces' and 'glasgow' are designated by 'SOURCE.' *Terms* may be selected from several ontologies (e.g. DOID (37), ENVO (38,39), GAZ (40), UBERON (41), EFO (42), and NCBI Taxonomy (12)), allowing easy and precise *annotations*. **d.** Uploading *annotations* may be performed either through the dbBact website, dedicated clients (i.e. Calour (43)) or by REST-API. For clarity, the following nomenclature holds throughout the manuscript where 'reserved' words appear in italics (e.g. *experiment*, *sequence*, *annotation, term*), predicates appear in all caps (e.g. 'HIGHER IN,' 'LOWER IN,' 'SOURCE'), and specific *term* names follow the ontology convention of being lower case (e.g. 'homo sapiens').

in dbBact are primer trimmed and originate from one of the supported 16S rRNA forward primers. For dbBact release 2022.07.01, the supported forward primers are V1-27F (AGAGTTTGATCMTGGCTCAGxxx), V3-341F (CCTACGGGNGGCWGCAGxxx), V4-515F (GTGCCAGCMGCCGCGGTAAxxx), where 'xxx' denotes the beginning of the ASV sequence stored in dbBact. Although additional primers can be added to dbBact, the vast majority of 16S rRNA studies uses one of the three primers described. The minimum length of *sequences* uploaded to dbBact is 100nt. Upon upload, *sequences* are stored at their full length rather than being truncated to a fixed length. Note that reads are not limited to specific lengths, and while the minimal length is 100nt, an overwhelming majority (>98%) of current reads are of length 150nt. When submitting a query *sequence*, exact sequence matches are searched using $length = min(query\_sequence\_length, database\_sequence\_length)$.

Taxonomy assignment. Species-level taxonomy is assigned a 100% identity match to the SILVA database. In case a *sequence* matches several SILVA sequences, all these taxonomies will be retrieved. To account for *sequences* that do not match SILVA entries, we also assign lower level taxonomy using RDP version 2.12 (47). A python script runs daily to add RDP-based taxonomy assignments to uploaded dbBact ASV *sequences*. Although taxonomy is not used in dbBact for analysis, it may be used for querying dbBact (e.g. retrieving *annotations* associated with bacteria of the genus *Streptococcus*).

Inter-region querying. dbBact supports the harmonization of microbiome studies performed using different protocols by inter-region linking. When submitting a *sequence*, dbBact uses the SILVA database of full length 16S rRNA genes (SILVA version 132) to identify *sequences* whose 'footprint' in other variable regions matches the query. First, the SILVA sequences containing the query *sequence* are detected. Second, all dbBact *sequences* that match these SILVA sequences, in any region, are retrieved (i.e. $\text{Query(S)} = \{T : T \in dbBact, \exists R \in SILVA \; so \; that \; ss(S, R) \; and \; ss(T, R)\}$ where *ss* stands for 'subsequence'). Querying is performed using the 'wholeSeqIDsTable' table in the dbBact implementation. To enable fast queries, a daily script is run on new dbBact *sequences*, linking all *sequences* sharing a SILVA sequence. Such linking is performed only when querying dbBact, therefore new versions of SILVA or other full length 16S rRNA databases may be seamlessly applied. Currently, dbBact supports linking the V1, V3, and V4 forward primer reads; additional primers may be incorporated if needed.

Queries of different sequence length and inexact matches. *Sequences* uploaded to dbBact may vary in length

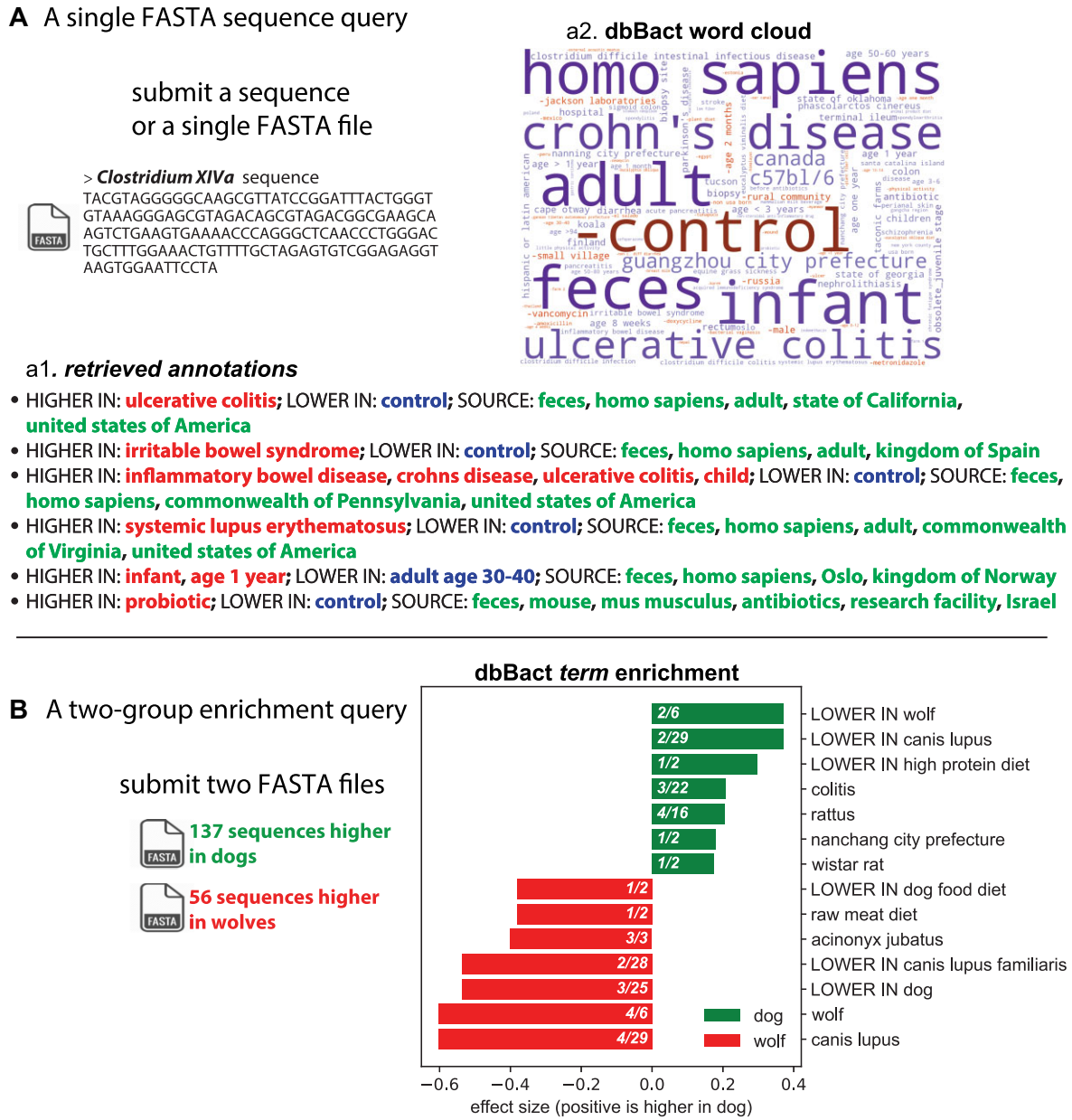


Figure 2. dbBact provides two basic query types: **A.** Uploading a FASTA file of ASVs results in a list of the most relevant *annotations* containing these *sequences*, and a 'word cloud' of best matching *terms*. In this example, a V4 ASV of *Clostridium XIVa*, which is highly abundant in fecal samples of chronic fatigue syndrome patients (CSF), as detected by Giloteaux et al. (44), was submitted. Panel **a1** provides representative *annotations* containing the query ASV (the full list of ∼150 *annotations* appears in Supplementary File 3). dbBact found this ASV to have higher relative abundance in the disease group than in healthy controls in several studies (ulcerative colitis, irritable bowel disease, and lupus), and in antibiotic-treated mice supplemented with probiotics (last *annotation* arising from (45)). Panel **a2** displays the word cloud summarizing the *terms* associated with the query ASV. The size corresponds to a *term*'s F_1 score, while color designates the associated predicate, i.e. blue for 'SOURCE'/'HIGHER IN' *terms*, and red color preceded by a minus sign corresponds to 'LOWER IN' *terms*. Intensity corresponds to reliability, where the lighter the color the less *annotations* are associated with the *term*. Hence, this *Clostridium XIVa* query ASV is associated with human feces in dysbiosis states of 'crohn's disease,' 'ulcerative colitis,' 'diarrhea,' and 'c. difficile infection' (a full list of F_1 scores per term appears in Supplementary File 7). **B.** By contrasting two groups of ASVs, dbBact identifies enriched *terms* characterizing each group. For example, 137 and 56 ASVs were submitted, corresponding to differentially abundant ASVs with higher relative abundance in fecal samples from domestic dogs and wolves living in zoos, respectively (data from (46)). Bar lengths show the normalized rank-mean difference for the top significantly enriched *terms* in the dog and wolf ASVs (green and red bars, respectively). *Term* enrichment is based on a non-parametric rank mean test with FDR < 0.1 using dsFDR (see the *term* enrichment analysis section in Methods). The numbers in the bar of each *term* correspond to the number of dbBact *experiments* in which the *term* differs significantly between the two ASV groups (numerator) and the total of dbBact *experiments* containing the *term* (denominator). *Sequences* that were more abundant in the wolf group are enriched in *terms* related to wolf, meat diet, and cheetah (*Acinonyx jubatus*).

depending on the sequencing platform and sequenced region. When adding new *annotations*, dbBact stores the full-length sequence of each ASV. For example, when two *experiments* provide information about the same bacterium using 150nt and 200nt reads, respectively, dbBact stores these *sequences* as separate entries and links each *annotation* to the corresponding *sequence*. Yet, when submitting a query using either *sequence,* dbBact retrieves *annotations* using exact match on the shortest common *sequence*, hence also retrieving *annotations* related to the other *sequence*. For example, if a given *annotation* is associated with a 150nt *sequence*, then submitting a query ASV of length 100nt that match the first 100nt out of the 150nt would retrieve this *annotation*. Similarly, if the query ASV is of length 200nt, the above-mentioned *annotation* will be retrieved if it matches its first 150nt. In addition to exact matches, dbBact searches for sequences with up to 2 mismatches using a specific REST-API endpoint.

Version freezing. As the number of studies uploaded to dbBact is constantly growing, and in order to facilitate reproducible analysis based on dbBact, we have implemented a versioning option for dbBact. All REST-API queries enable limiting results to a maximal *annotation* ID, hence disabling the effect of future *annotations* that were not present when the analysis was performed. In addition, weekly database snapshots are available for recreating dbBact versions on a local server.

dbBact interfaces.

REST-API server. The dbBact REST-API server (http://api.dbbact.org) is implemented in Python 3.6, using Flask version 0.12/Gunicorn v19.9 to handle web queries, and psycopg2 version 2.7.1 for handling Postgres data queries. Full API documentation is available at http://api.dbbact.org/docs. Examples using the REST-API for querying are available at: https://github.com/amnona/dbbact-examples. The REST-API enables access to all dbBact functions. Querying dbBact or adding anonymous *annotations* does not require registration. Registration by username/password enables editing *annotations* submitted by the same user.

dbBact website. The dbBact website (http://dbbact.org) enables dbBact *annotation* retrieval based on ASVs, taxonomy, or ontology *terms*. Additionally, the website provides word-cloud generation and *term* enrichment analysis. The source code for the website as well as deployment instructions are available on the dbBact-website github page (https://github.com/amnona/dbbact-website).

dbBact-calour interface. dbBact is integrated into the Calour microbiome analysis program (https://github.com/biocore/ calour), using the dbBact-Calour module (https://github.com/amnona/dbbact-calour). Using this interface, users can both query dbBact regarding bacterial sequences, and add new *annotations.* The dbBact-Calour module provides dbBact *annotation* retrieval from the interactive Calour heatmap display, showing all *annotations* associated with the selected *sequence*. Additionally, the module enables GUI-based creation of new dbBact *annotations* for selected *sequences*, and performs *term* enrichment analysis, *term*-based principal coordinate analysis (PCA) and word cloud generation. A Jupyter notebook tutorial is available at: http://biocore.github.io/calour/notebooks/microbiome_databases.html.

The module also works with EZCalour, the full GUI version of Calour, (https://github.com/amnona/EZCalour). A tutorial for dbBact enrichment analysis using EZCalour is available at: https://github.com/amnona/EZCalour/blob/master/using-ezcalour.pdf.

QIIME2 plugin. The q2-dbBact plugin (https://github.com/amnona/q2-dbbact) enables dbBact *annotation*-based analysis using the QIIME2 framework (34). The interface provides *term* enrichment analysis for the output of various QIIME2 differential abundance plugins (ANCOM (48), Songbird (49), ALDEx2 (50), DACOMP (51), or a rank-mean method). Additionally, the plugin supports dbBact *term* word cloud and interactive heatmap generation.

Standard analysis: default dbBact preprocessing of an *experiment*

Although dbBact is a wiki-style knowledge base, the vast majority of *annotations* in dbBact release 2022.07.01 was added by the dbBact team. Studies were selected from published microbiome papers, and *annotations* were added following the re-processing of the experimental data, using a 'standard' manual analysis pipeline as follows:

The raw data of each scientific paper (i.e. per-sample FASTA files and corresponding metadata) were downloaded using the provided accession (e.g. by SRA/ENA accession or Qiita (20) study ID). When data or metadata were not available, the authors were contacted and provided the missing data directly. When primer sequences were part of the reads, they were removed using a custom script (https://github.com/amnona/GetData). Subsequently, the Deblur pipeline (14) was applied to the reads of each sample (Deblur script version 1.1.0, using default parameters, https://github.com/biocore/deblur), resulting in a denoised biom table. Since dbBact annotations focus on biologically relevant *sequences*, rare ASVs have a minor effect. Hence, similar results would be obtained using DADA2 (15) instead of Deblur (Supplementary Figure S19).

This biom table, together with the per-sample metadata, were manually re-analyzed using Calour (43), to add *annotations* capturing biological conclusions arising from the study (see 'Calour implementation' section in Supplementary Methods). Initially, samples having less than 1000 reads were removed, and samples were normalized using total sum scaling (i.e. dividing ASV counts in each sample by the sample total reads (52)) to obtain the relative abundance of each ASV in the given sample. Three types of predicates were sought:

(i) 'DIFFERENTIAL:' To detect sets of *sequences* associated with relevant conditions (e.g. sick vs. healthy), *sequences* significantly enriched between two conditions were identified using the Calour diff_abundance() method, a non-parameteric permutation based

rank-mean test. Briefly, diff_abundance() first ranks each feature (sequence) across the samples (using the per-sample relative abundance of the feature), and then calculates the difference between the mean of the ranks across the two sample groups. This value is compared to the analogous values following 1000 random permutations of the sample group labels, and a corresponding p-value is calculated. Following the per-sequence p-value calculation, a multiple hypothesis correction is performed across all sequences using dsFDR (53), controlling the false positive fraction to less than 0.1. dsFDR is a permutation-based variant of the Benjamini-Hochberg (BH) approach (54) adapted to microbiome data, which are typically sparse and originate from a limited number of samples (see Supplementary Methods section 'dsFDR multiple hypothesis correction'). Correlations with continuous metadata fields (e.g. body mass index - BMI) were detected with a permutation-based Spearman test with dsFDR correction using calour.correlation(), which is implemented similarly to the diff_abundance() function, but using the Spearman correlation as the test statistic. In both cases, the set of *sequences* of higher or lower relative abundance in one condition than in the other were then annotated as 'DIFFERENTIAL,' i.e. 'HIGHER IN' condition 1 and 'LOWER IN' condition 2. Other differential abundance methods, such as compositionality-aware tests, can also be used for differential *annotations*. The specific method applied is stored within each *annotation*.

(ii) 'COMMON'/'DOMINANT:' For each study, *sequences* present in more than half of the relevant samples were *annotated* as 'COMMON.' 'DOMINANT' *sequences* were identified as *sequences* whose mean relative abundance in the relevant samples was higher than 0.01. In studies containing samples from multiple sources (i.e. fecal and saliva samples, or samples from individuals from several countries or disease vs. healthy), a 'COMMON'/'DOMINANT' *annotation* was added separately to each source subset.

(iii) 'CONTAMINANT:' Indicates candidate reagent-related contaminants, especially in low biomass samples. For example, when a study contained a set of negative control (blank) samples, ASVs showing higher relative abundances in these controls than in the non-blank samples were manually annotated as a possible 'CONTAMINANT'.

Examples of the different predicates appear in Table S3.

Remark: Using the abovementioned pipeline is not a prerequisite for adding new *annotations*, and any denoising method followed by statistical analyses can be applied by users contributing to dbBact.

Statistical analysis in dbBact

Word cloud generation. Calculating a term's F_1*score*: Given a set of query *sequences S*, each dbBact *termt* is assigned an F_1 score, corresponding to the harmonic mean of precision and recall. For each sequence s in S, we calculate the fraction of s's *annotations* that contain the *termt*. The average of these values across S provides the precision of t on S. Similarly, for a given *sequences* and a *termt*, recall is calculated as the fraction of t's *annotations* that contain s. To suppress terms that appear in a small number of *experiments*, the total number of t's *annotations* (i.e. the denominator) is artificially increased by 1. The average of these values across S provides the recall of t on S.

Displaying a word cloud. The word cloud size of each *term* is proportional to its F_1 score. If the term appears in 'LOWER IN' annotations, its color is orange, otherwise it is blue. The brightness of each *term* represents the number of *experiments* containing the *term,* indicating the reliability of the term (white for a single *experiment*; ranging to dark blue/orange for $>= 10$ *experiments*).

Word clouds were generated with the dbBact-Calour module for Calour (https://github.com/amnona/dbbact-calour) using the draw_wordcloud() function.

Term enrichment analysis. Given two sets of *sequences*, $S1$ and $S2$, we search for *terms* significantly enriched in either group using the following steps:

(a) *Calculating an annotation-based score per sequence and term*. Each *annotationa* in dbBact is assigned a 'weight' $w(a)$ according to its predicate. The predicates 'COMMON,' 'CONTAMINATION,' and 'DIFFERENTIAL' are assigned a weight of 1, and the predicates 'DOMINANT' and 'OTHER' are assigned a weight of 2 and 0.5, respectively. These weights are applied to calculate a score $Score(s, t)$ for each *termt* in dbBact and a *sequences* in either $S1$ or $S2$. The score sums the weights of all *annotations* involving s and t. When t appears as 'LOWER IN' in the predicate 'DIFFERENTIAL,' a new *term* 'not t' is created and assigned a weight of 1.

(b) *Calculating effect size of a term*. For a *termt*, $Score(s, t)$ is calculated for all *sequences* in $S1$ and $S2$, and the effect size of t is defined as $e(t) = |(\frac{\sum_{s \in S1} Score(s,t)}{|S1|} - \frac{\sum_{s \in S2} Score(s,t)}{|S2|})|$ where $|S|$ corresponds to the number of *sequences* in the set S. Throughout this paper, and also in the dbBact plugins, the scores for each *term* are rank transformed (across *sequences*) prior to calculating the effect size, and therefore an additional normalization is applied: $e(t) = 2|(\frac{\sum_{s \in S1} rank(Score(s,t))}{|S1|} - \frac{\sum_{s \in S2} rank(Score(s,t))}{|S2|})|/(|S1| + |S2|)$.

(c) *Finding significant terms*. Each *term* is assigned a p-value by comparing its scores over 1000 random permutations of the combined $S1$ and $S2$*sequences* to sets of size $|S1|$ and $|S2|$ using the calour.diff_abundance() method described above, including a dsFDR multiple hypothesis correction (with a threshold of 0.1), to detect significant *terms*.

(d) *Calculating the significance of a term across experiments.* Until this stage, we measured the enrichment of a *term* based on all dbBact *experiments* combined. To estimate whether such significance occurs across multiple *experiments,* or whether it is driven by a single or a few *experiments*, steps (a)-(c) were also repeated using each individual *experiment* that contains t. The total number of *experiments* containing each *term* and the fraction in which the *term* was significant appear in each figure.

The abovementioned analysis is performed using the dbbact-calour module enrichment() function.

An alternative *term* enrichment method that uses all ASVs in each group of samples, weighted by their relative abundances in each sample, is described in the Supplementary Methods (Supplementary Figure S21).

Venn diagrams. Given two sets of sequences, $S1$ and $S2$, and a *termt*, we plot a Venn diagram indicating the number of *sequences* associated with *t* across all dbBact *annotations*, and the overlap of these *sequences* with $S1$ and $S2$.

Venn diagrams were generated using the dbbact-calour module plot_term_venn_all() function.

dbBact term-*based principal component analysis.* Each sample x is represented by a vector of *sequences'* relative abundances (abbreviated 'sf'), $x(S) = (sf_1, sf_2, \ldots, sf_n)$ across the *nsequences* that appear in all study samples (sf equals zero in case a *sequence* does not appear in a specific sample). We then transform $x(S)$ into a *term*-based representation of x, i.e. $x(T) = (ts_1, ts_2, \ldots, ts_m)$, where each ts is a *term*-score described below, calculated across all m dbBact *terms*. The ts score for the *termt* is given by $ts = \sum_{i=1}^{n} sf_i \cdot precision(i, t)$, where $precision(i, t)$ is the fraction of *annotations* associated with *sequencei* that also contain the *termt*. Finally, once $x(T)$ is calculated over all samples in a study, we perform principal component analysis of this space. Each principal axis is defined by its weights, where the highest weights (in absolute value) are used for providing biological meaning.

The abovementioned analysis is performed using the dbbact-calour module plot_term_pcoa() function.

Processing of datasets

All datasets discussed in the paper were processed using the following pipeline: raw reads were downloaded and denoised using the Deblur pipeline (14) with default parameters; the resulting denoised biom table was loaded into Calour (43), and differentially abundant bacteria were identified using a permutation-based non-parametric rank mean test with dsFDR multiple hypothesis correction (53) set to 0.1 (using the calour diff_abundance() method). In the case of the American Gut Project dataset, multiple samples originating from the same individual were aggregated to a single sample using mean relative abundance for each ASV. The groups of high and low fruit consumption were controlled for confounders by stratifying samples in both groups based on the AGP metadata fields: age category ('AGE_CAT'), sex and BMI category ('BMI_CAT'), and randomly dropping samples to equalize the number of samples from each stratum prior to differential abundance testing. dbBact *term* word clouds were generated by applying the above-described word cloud approach using the dbbact-calour module draw_wordcloud() method on all sequences present in at least 30% of the samples. *Term* enrichment was performed using the above-described term enrichment approach using the dbbact-calour module enrichment() method with default parameters (dsFDR = 0.1). dbBact term PCAs were generated using the dbbact-calour module plot_term_pcoa() function. Accession numbers for each dataset used are available in Supplementary File 2. Jupyter notebooks used for the creation of each figure are available at https://github.com/amnona/dbbact-paper.

RESULTS

We first present general statistics of dbBact and display its comprehensiveness. We then demonstrate the importance of ASV-based associations and provide a detailed example of a dbBact-based analysis. Finally, we summarize biological insights across various habitats derived by dbBact analysis of 16 published papers. As dbBact is continuously growing, and to allow readers to recreate the figures presented in this paper, all results presented in this section are based on a 'freezed' dbBact release 2022.07.01 that contained 913 unique *experiments*.

dbBact: scope and comprehensiveness

dbBact release 2022.07.01 contains approximately 345000 unique bacterial amplicon *sequences*, an amount that is on par with the 300000 sequences observed by the Earth Microbiome Project (11). *Sequences* arise from over 900 unique *experiments*, i.e. studies from which observations were added (Figure 3A). Over 7000 dbBact *annotations* associate these *sequences* with various phenotypes using ontology derived *terms*. As each annotation typically includes many *sequences*, this results in over 1400000 unique genotype-phenotype associations.

General statistics of dbBact

The dbBact *experiments* cover a wide range of habitats (Supplementary Figure S1a), geographic regions (Supplementary Figure S1b), plant and animal hosts (Supplementary Figure S1c), human body sites (Supplementary Figure S1d), and human diseases (Figure 3B). For example, 149 *experiments* cover diseases, of which 86 are of an anatomical entity (e.g. Crohn's disease or ulcerative colitis), and seven more are defined as metabolic diseases. The most abundant dbBact *terms* are 'united states of america,' 'homo sapiens,' and 'feces,' each appearing in over 1000 *annotations* arising from more than 150 different *experiments.* Most of the other *terms* appear in less than twenty *experiments* (Figure 3C). The most prevalent bacterial *sequence* is *E. coli,* appearing in over 900 *annotations* from over 300 *experiments* (Figure 3D). Although this could reflect the universality of *E. coli* in various habitats, it may also be due to potential contaminations (55), a reason that may also explain the high prevalence of *Staphylococcus* (appearing in over 200 dbBact *experiments*). The number of *experiments* per *sequence* follows a power law distribution, with a majority of *sequences* appearing in a single *experiment*, yet over 80000 *sequences* were observed in more than one *experiment* and 7000 *sequences* appeared in at least ten *experiments* (Figure 3D).

dbBact allows the upload of *sequences* from several commonly used regions (V1-V2, V3-V4 or V4; see Table S2 for a list of primers). Upon upload, *sequences* from different regions are 'linked' through their full-length 16S rRNA sequence in the SILVA database (see Inter-region querying section in Methods). When submitting query *sequences*

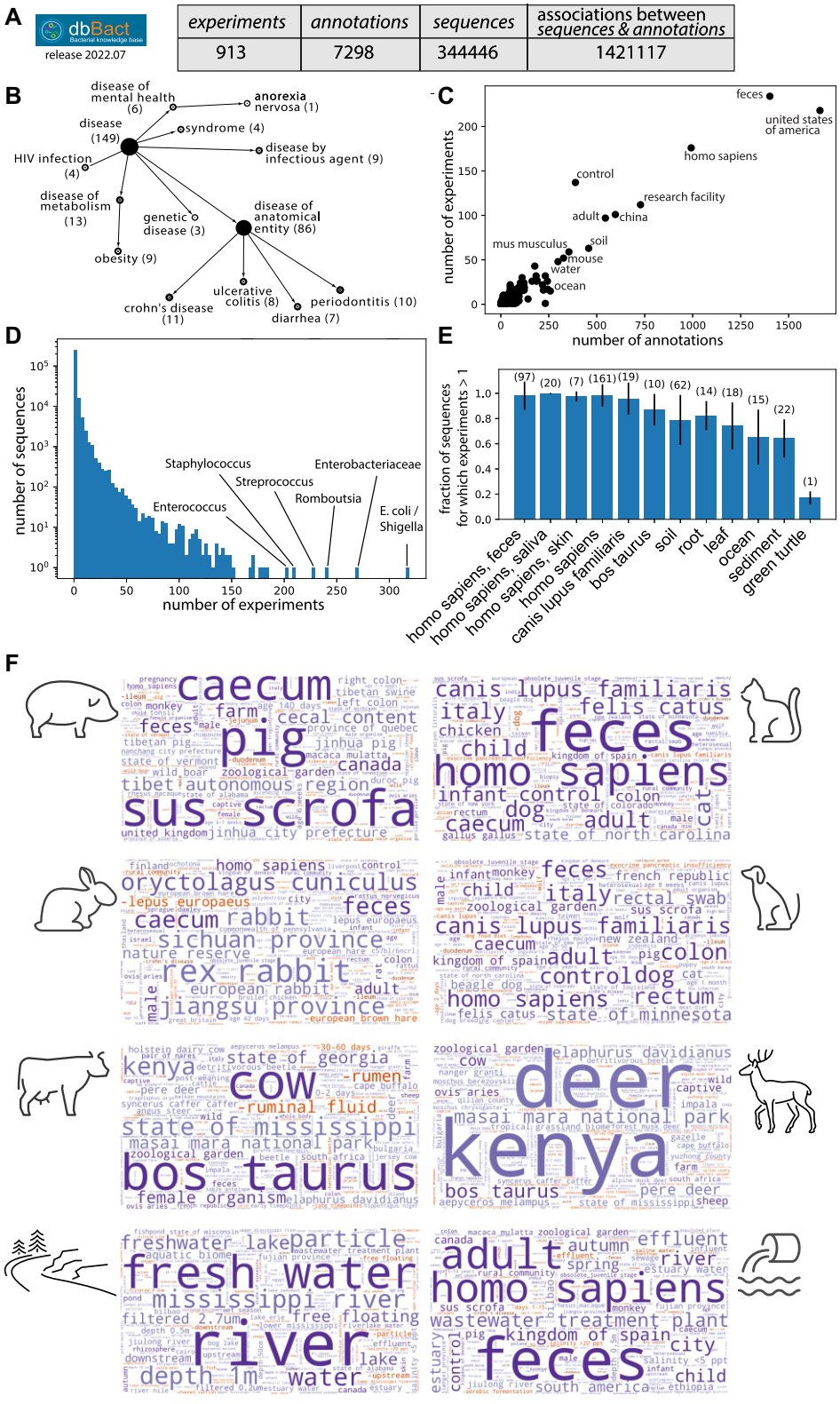


Figure 3. The scope and comprehensiveness of dbBact. **A.** Scope of dbBact release 2022.07 (used for the analysis presented in this paper). **B.** The number of *experiments* for representative disease categories based on the DOID ontology. **C.** Scatter plot of the total number of *annotations* and *experiments* in which each dbBact *term* appears. **D.** Histogram of the number of *experiments* in which each dbBact *sequence* appears. **E.** Knowledge base comprehensiveness. The fraction of 'COMMON' *sequences* from each *experiment* that have been annotated in additional *experiments* is shown for various *terms*. The number of experiments containing the *term* is designated above each bar. **F.** Comprehensiveness in a source tracking task. *Sequences* from eight sample types from Hägglund et al. were blindly submitted to dbBact. Their word clouds clearly display the sources of the samples (shown by the matched cartoon). *Term* sizes correspond to the F_1 score of each *term,* combined for all *sequences* present in > 0.3 of the samples (for each sample type).

from one region, dbBact retrieves all *annotations* containing the corresponding *sequences* across all regions (including, naturally, the region from which the query was provided). To demonstrate the usefulness of such 'linking,' Supplementary Figure S2 provides several examples of V1-V2 and V3-V4 *sequence* queries that are successfully characterized based solely on 'linked' V4 *sequences*. The figure also benchmarks 'linking' using sets of fecal and ocean water samples each sequenced by V1-V2, V4 and V3-V4.

Comprehensiveness of dbBact

Intra-dbBact estimates. To estimate the comprehensiveness of dbBact, we tested how many bacterial *sequences* typical of a specific environment (e.g. human feces) have *annotations* arising from more than one *experiment*. We selected *sequences* having an *annotation* of type 'COMMON' (i.e. present in more than half of the samples in an *experiment*) for each of several *terms*, and measured the fraction of these *sequences* that have *annotations* from another dbBact *experiment* (Figure 3E). For example, there are 97 *experiments* having a 'COMMON IN homo sapiens feces' *annotation*. Iterating over each of these *annotations*, about 98% of the associated *sequences* appear in more than one *experiment*. Hence, fecal bacteria are already well covered by dbBact. A similar level of 'coverage' occurs for several other human-related *terms* and for dogs, where almost all *sequences* were observed in more than a single *experiment*. Regarding the *terms* 'cow,' 'soil,' 'root,' and 'leaf,' about 80% of the *sequences* appear in more than a single *experiment*, whereas the 'coverage' of 'green turtle' is much lower, indicating that additional experiments are required to capture its full bacterial diversity.

Out-of-sample comprehensiveness. As another example of comprehensiveness, dbBact was tested in a source tracking task, i.e. identifying the host or niche of a sample based on its bacterial composition. Hägglund et al. collected samples from either sewage influent or from freshwater, as well as feces sampled from several animals (rabbit, cat, wild boar, dog, cow and deer), aiming to find unique bacterial footprints of each source (56). We used all *sequences* present in more than 1/3 of the samples from each group as queries to dbBact resulting in word clouds describing each sample group (Figure 3F). In almost all cases, the notable *terms* in each word cloud were indicative of the sources of the samples, e.g. *sus scrofa* for the wild boar fecal sample, allowing accurate source tracking. The only exception was cat fecal samples, which were detected as a combination of cat, dog, and human, probably because of the small number of cat fecal samples present in the current dbBact release.

The advantage of sequence-based associations

Results of 16S rRNA profiling experiments comprise a list of ASVs found in each sample and their abundances. Corroboration of these results with other microbiome studies is typically performed by searching published studies mentioning the taxonomy of these sequences. In many cases, however, such text-based mining may be limited because of constraints in taxonomic assignment. First, taxonomy is far from being full, e.g. species-level assignment is missing for about 80% of 16S rRNA sequences in Greengenes (57), and about 35% of the Greengenes sequences lack a genus assignment (58). Second, in many cases the same assigned taxonomy may be associated with vastly different phenotypes. As observed by the Earth Microbiome Project, bacteria of the same genus may be present in vastly different habitats, whereas specific sequences are associated with a certain habitat (11). This phenomenon underscores the importance of sequence-based association as provided by dbBact. For example, both *sequences* in Figure 4A belong to the genus *Blautia*, hence taxonomy-based associations may conclude that they play similar 'roles' and are associated with the same phenotype. But querying dbBact with each of these two *sequences* results in a strikingly different picture, which we refer to as a 'good' and 'bad' *Blautia*. The 'good' *Blautia* is more abundant in healthy controls than in patients of type 1 diabetes (T1D), Crohn's disease (CD), inflammatory bowel disease (IBD), diarrhea, and kidney stones (Figure 4A), whereas the 'bad' *Blautia* is more prevalent in patients suffering from IBD, CD and ulcerative colitis (Figure 4B).

Collecting all 'disease' related dbBact annotations shows that the 'bad' *Blautia* is 'HIGHER IN' in the disease group (compared to controls) in 8/9 disease *annotations* associated with it, whereas the 'good' *Blautia* is 'LOWER IN' in the disease group (compared to controls) in 22/24 disease *annotations* (Figure 4C). Therefore, sequence-based analysis provides a solid genotype-to-phenotype association compared to taxonomy-based associations.

dbBact provides a pan-microbiome view: detailed example

dbBact may add another layer to data analysis in microbiome studies by identifying commonalities between different conditions and diseases, generating novel biological hypotheses. To demonstrate such a pan-microbiome analysis, we use data from a study comparing subjects consuming a calorie restricted diet to those without dietary restrictions (59), and demonstrate the use of dbBact *term* enrichment.

Fecal samples from two groups of lean individuals (body mass index, BMI < 25) who either followed a caloric restriction diet (CR) or did not have dietary restrictions (AMER) were selected. Standard analysis with FDR set to 0.1 (see 'standard analysis' section in Methods) identified 28 and 141 bacterial *sequences* significantly more abundant in AMER and CR cohorts, respectively (Figure 5A). For clarity, we refer to these groups of *sequences* as S-AMER and S-CR, respectively. Figure 5B shows the internal 'transformation' performed by dbBact from a heatmap of bacterial relative abundances to a heatmap of association scores of *terms* for each *sequence* (columns in Figure 5A-B are aligned and correspond to the same *sequences*). For example, the *term* 'high BMI,' appears in almost all S-AMER *sequences*, while it is almost absent in S-CR *sequences*. These association scores in Figure 5B are then used as input to a non-parametric differential abundance test (see Methods section 'statistical analysis in dbBact'), identifying *terms* significantly enriched in each of the two sequence groups (Figure 5C). Results indicate that *sequences* in the S-CR group are associated with *terms* related to low BMI ('low bmi,' 'LOWER IN high bmi') and with rural/undeveloped habitats ('LOWER IN united states of america,' 'small village')

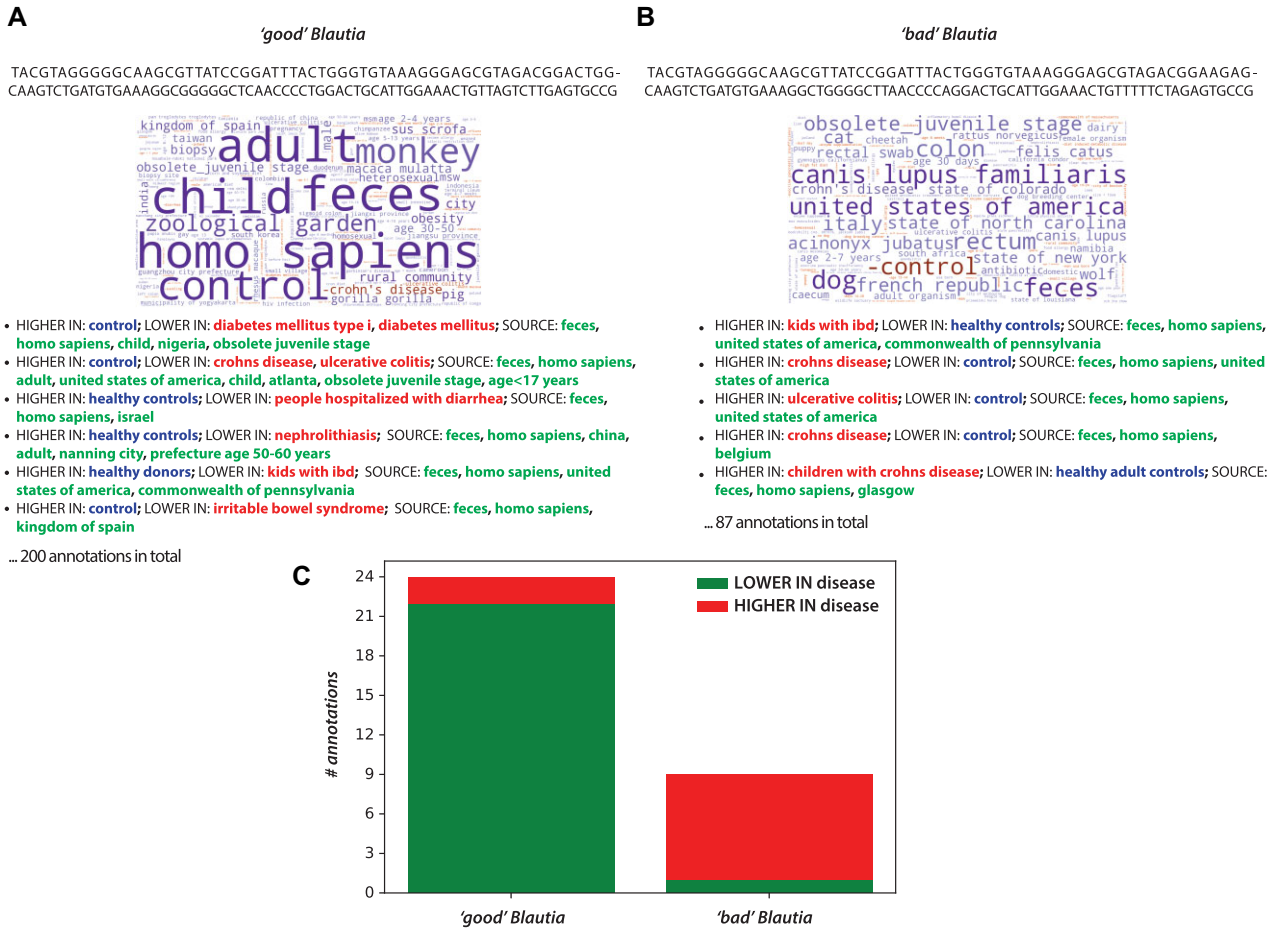


Figure 4. Sequence-based analysis provides more accurate genotype-to-phenotype associations compared to taxonomy-based associations. **A,B.** Two ASVs of the genus *Blautia*, that differ by nine bases over the 150nt Illumina read of the 16S rRNA V4 region are associated with opposite phenotypes, as discovered by dbBact. The two word clouds and annotations for each ASV, display 'opposite' associations with disease. The left ASV (**A**) is more prevalent in healthy subjects ('good' *Blautia*), whereas the other (**B**) is highly abundant in a series of disease-related *annotations*. Such differences can be traced through dbBact, but are completely missed by a taxonomy-based analysis. **C.** The number of disease-related *annotations* for the two *Blautia* ASVs across dbBact displays an opposite trend of being low and high in disease, for the 'good' and 'bad' *Blautia*, respectively. The total number of *annotations* in dbBact 2022.07.01 associated with the 'good' and 'bad' *Blautia sequences* is 377 and 124, respectively.

(see Supplementary File 4 for the full list of enriched terms). By contrast, bacteria from the S-AMER group have a significantly higher number of *annotations* related to high BMI ('LOWER IN low bmi,' 'high bmi') and urban/modernized habitats ('state of oklahoma,' 'LOWER IN rural community,' 'LOWER IN small village').

Thus, although participants from both diet groups were lean, certain aspects of the underlying microbiome were associated with high and low BMI bacteria, for AMER and CR, respectively. Additionally, bacteria enriched in CR vs. AMER tend to be associated with rural/undeveloped habitats, which may indicate an adaptation of some bacteria found in rural communities to a low-calorie/higher vegetable diet content.

To further confirm the relationship between diet and BMI, we collected all *sequences* across dbBact having a 'high bmi' annotation, resulting in 319 *sequences*. The overlap between these *sequences* and the S-AMER and S-CR groups is shown in Figure 5D (left). Although 93% (26/28) of S-AMER *sequences* overlap with 'high bmi'-associated *sequences*, the overlap of S-CR *sequences* is 11% (15/141), i.e. a much larger fraction of S-AMER *sequences* is associated with high BMI. An analogous Venn diagram for the *term* 'low bmi' displays an overlap of 54% (15/28) and 78% (110/141) of S-AMER and S-CR bacteria with 'low bmi'-associated bacteria in dbBact, respectively (Figure 5D left). As participants from both CR and AMER groups were lean (BMI < 25), one may hypothesize that the effect of BMI on the microbial composition, observed in various studies, is due to dietary differences rather than the high BMI phenotype.

Remark regarding spurious/irrelevant terms: dbBact release 2022.07.01 contains *annotations* of approximately 6500 unique *terms*, some appearing only in a few *experiments*. As a result, word clouds and bar plots may often include seemingly odd *terms*. For example, the *term* 'state of oklahoma' in Figure 5D is significantly enriched in S-AMER, a fact that seems implausible. This *term* appears only in two dbBact *experiments*, one of which compared a rural community in Peru to an urban community in

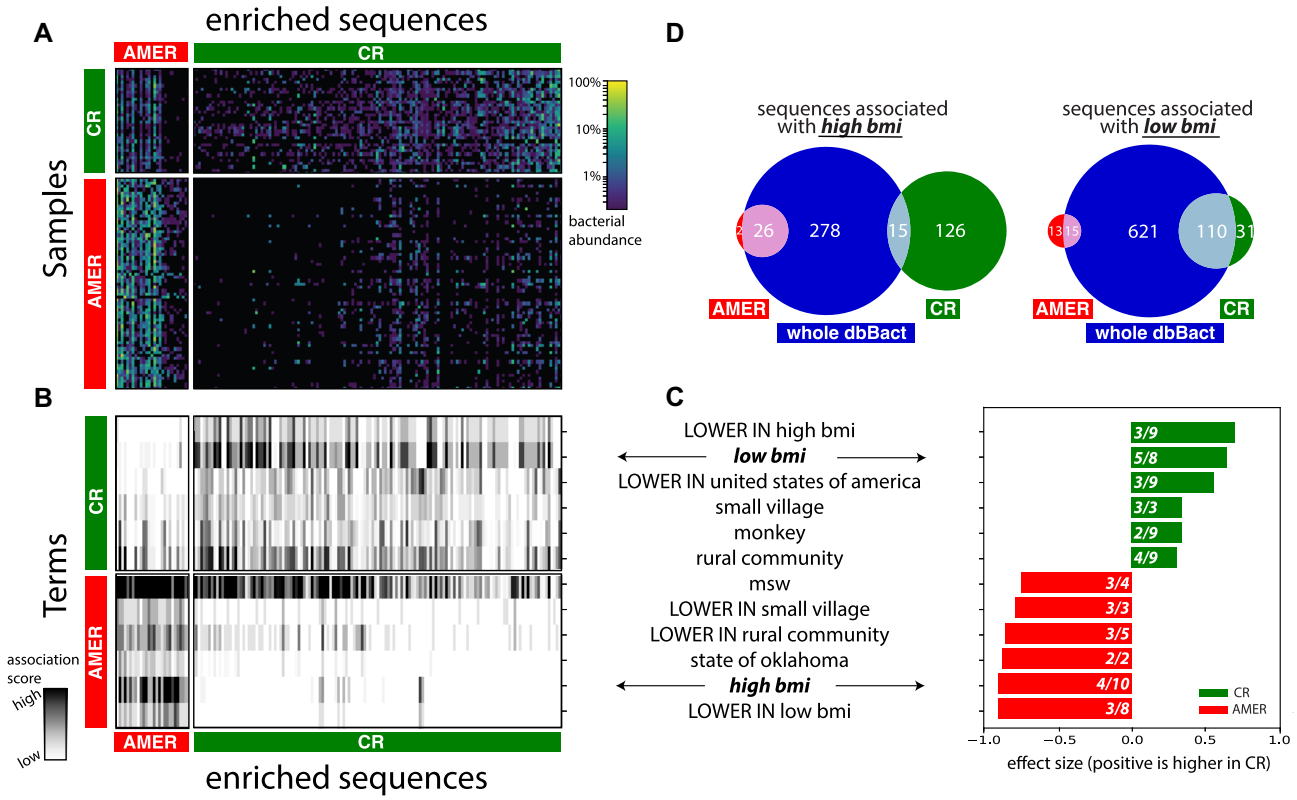


Figure 5. dbBact links caloric restriction associated bacteria to other phenotypes. **A**. Heatmap displaying bacterial relative abundances across fecal samples (rows) of low BMI individuals (BMI < 25) practicing either a caloric restriction diet (CR, n = 33) or without dietary restrictions (AMER, n = 66), over a set of *sequences* (columns) that are significantly higher in either group. A differential abundance test (rank-mean test with dsFDR = 0.1 multiple hypothesis correction) identified 136 bacteria whose relative abundance was higher in the CR group (S-CR) and 27 bacteria higher in the AMER group (S-AMER). **B**. dbBact *terms* (rows) enriched in the *sequences* appearing in panel **A** (columns in panels **A** and **B** are aligned). Heatmap values indicate the *term* score for each bacterium. *Terms* were identified using a non-parametric rank mean difference test with dsFDR = 0.1 (top 6 terms for each direction are shown; see Supplementary File 4 for full list of enriched *terms*). **C**. Summary of the top enriched *terms* in the CR and AMER diets (green and red bars, respectively). Bar length and numbers are as in Figure 2. **D**. Venn diagrams of dbBact *annotations* related to the *terms* 'low bmi' (right) and 'high bmi' (left). Green and red circles indicate the number of *sequences* associated with the *term* in the CR and AMER diets, respectively; the blue circle indicates the number of such *sequences* across dbBact as a whole. The intersections of 'low bmi' bacteria with the CR group are significantly higher ($p = 7E\text{-}5$, using two-sided Fisher's exact test), confirming the association. Similarly, the intersection of 'high bmi' annotated *sequences* across dbBact with the AMER group is significantly higher than that with the CR group ($p = 3E\text{-}17$, using two-sided Fisher's exact test).

Oklahoma (60). Hence, *annotations* from this *experiment* mentioned the *term* 'state of oklahoma' together with more relevant *terms* (e.g. 'rural community') which, in turn, caused its inclusion. As dbBact continues to grow, such 'transient' irrelevant inductions are expected to diminish.

Integrating dbBact into microbiome analysis pipelines allows generating novel biological hypotheses

To demonstrate how dbBact may be incorporated into microbiome analysis pipelines, we chose 16 dbBact *experiments*, excluded them from dbBact, then analyzed their results in the same way researchers would look at their own studies. In each of these studies, dbBact provided novel hypotheses that did not appear in the original paper and could not be formulated by standard methods.

The *experiments* presented here were chosen almost arbitrarily to provide examples from different habitats and niches: the human host, animals, and environmental samples (Figure 6A). dbBact-derived hypotheses may be divided into several 'types,' as follows.

Detecting inter-host similarities: dbBact can identify unexpected similarities in microbial populations across hosts. (*i*) For example, when examining the oral microbiome of wild sea otters (61), dbBact indicates a high similarity to the microbiome of the oral cavity of dogs and cats (Figure 6B and Supplementary Figure S3). (*ii*) In another example, fecal bacteria of Himalayan Griffons (62) are found to be similar to those of another carrion feeder, the California Condor (Supplementary Figure S4). (*iii*) Such inter-host similarities are also observed for disease-related bacteria. Examining bacteria in colitis in horses (63), dbBact detects an enrichment of human-associated bacteria, indicating a possible colonization by bacteria that are less host-specific (Supplementary Figure S5). (*iv*) Another recent meta-analysis of various human diseases identified shared disease-related bacteria in multiple diseases (64). When examining non-human-related *annotations*, dbBact finds these bacteria to be enriched in non-primate, homeothermic animals (mouse, horse, rat, chicken). By contrast, health-related bacteria found in this study are enriched in monkey-associated *terms* (Supplementary Figure S6). This may indicate the disap-

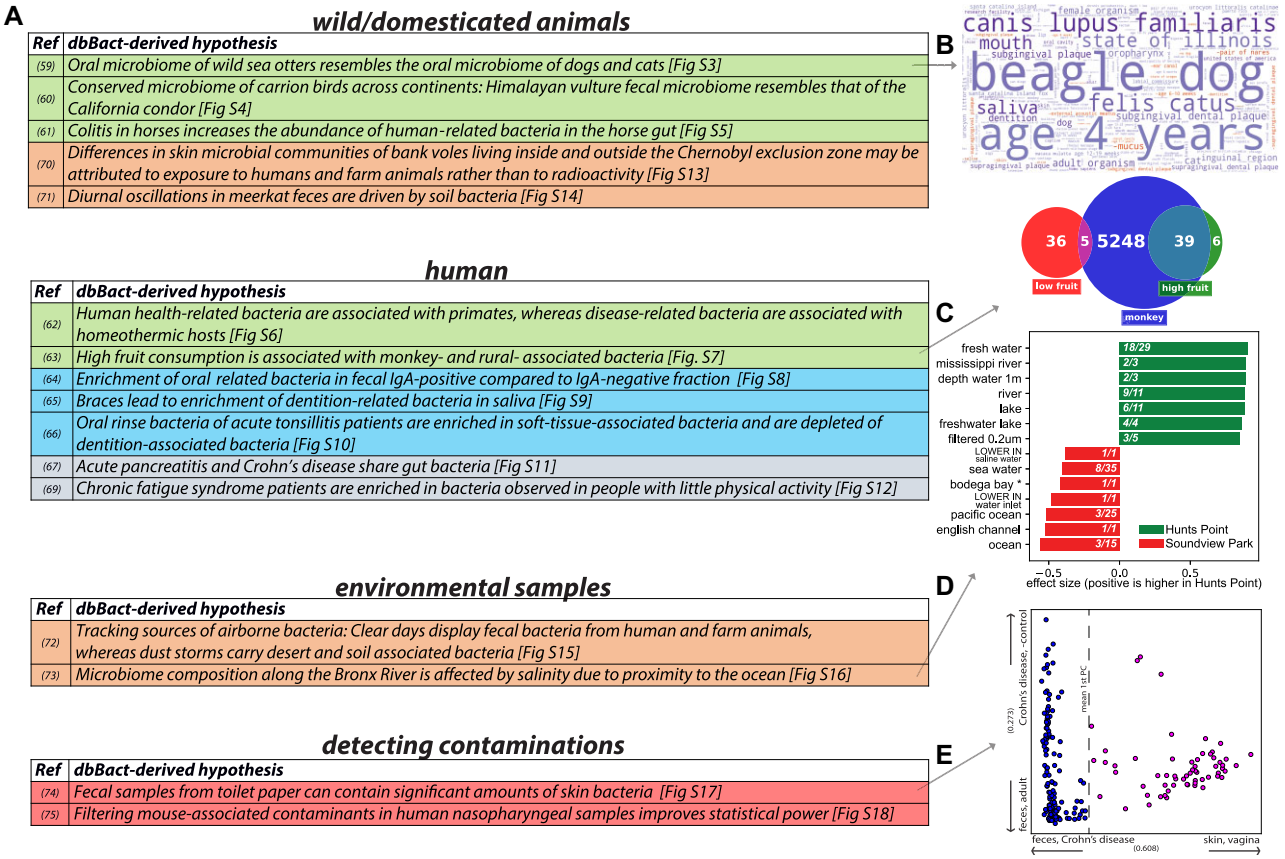


Figure 6. dbBact leads to novel biological hypotheses. **A.** Summary of biological hypotheses derived from dbBact-based analysis of published studies. Details of each analysis are given in the corresponding Supplementary Results section. Row colors correspond to hypothesis 'type' (inter-host similarities – green; intra-host similarities – blue; inter-disease similarities – gray; environmental sources – brown; contamination detection – red). **B-E.** Analysis results related to conclusions shown in panel a. **B.** dbBact *term* word cloud for *sequences* found in sea otter oral samples shows resemblance to dogs' and cats' samples. **C.** Venn diagram showing number of dbBact *sequences* associated with the *term* 'monkey' across dbBact (blue), and their intersection with *sequences* found in individuals from the American Gut study, who consume a high (green) and low (red) number of fruits per week. *Sequences* in the high-fruit consumption group are significantly more associated with the *term* 'monkey' (Fisher's exact test p-value < 0.00001). **D.** dbBact *term* enrichment comparing water samples collected in Hunts Point and Soundview Park, along the Bronx River in New York. *Sequences* whose relative abundance was higher in Hunts Point (green, located upstream) show significant fresh-water-related *term* enrichment (dsFDR $= 0.1$). **E.** *Term*-based principal coordinate analysis (PCA) of fecal samples of one individual collected daily for one year. The first principal component is the 'feces-skin' axis, where higher values correspond to 'skin' (see Methods for details). The values of a subset of samples, shown in magenta, is high, indicating possible skin-derived contamination in these fecal samples.

pearance of host-specific bacteria in multiple diseases, together with the appearance of more generalist bacteria. (*v*) A similar enrichment in monkey-associated bacteria and rural-community related *terms* is observed in individuals from the American gut project (7) who report high consumption of fruits, compared to those reporting low consumption (Figure 6C and Supplementary Figure S7).

Detecting intra-host similarities: dbBact can identify similarities within hosts. (*i*) For example, Scheithauer et al. (65) profiled the bacteria detected in the IgA-positive and IgA-negative fractions of fecal samples. dbBact-based analysis shows that the IgA-positive fraction is enriched in oral related *terms*, indicating a possible contribution of oral IgA to bacterial antibody coating (Supplementary Figure S8). (*ii*) In another study (66), dbBact finds an enrichment in dentition-related *terms* in an oral rinse of adolescents with braces compared to an enrichment in soft-tissue-associated bacteria in those that do not wear braces (Supplementary Figure S9). (*iii*) Such soft-tissue-associated bacteria are also observed when analyzing Yeoh et al. data (67) of tonsilitis patients (Supplementary Figure S10).

Detecting inter-disease similarity: (*i*) Zhu et al. compared the fecal microbiome of acute pancreatitis patients with that of healthy controls (68). dbBact-based analysis hints at a common gut response between pancreatitis and diarrhea, and Crohn's disease, i.e. a phenomenon of general dysbiosis formerly suggested by Duvallet et al. (69) (Figure S11, Supplementary Files 5,6). (*ii*) Giloteaux et al. (44) compared fecal samples of chronic fatigue syndrome patients with those of healthy controls. dbBact finds shared *sequences* between these patients and individuals who do little physical activity (Supplementary Figure S12).

Detecting environmental sources: dbBact can detect the sources of bacterial communities. (*i*) For example, Lavrinienko et al. collected skin swabs of bank voles inside the uninhabited Chornobyl exclusion zone and outside the contaminated region in the outskirts of Kyiv (70). dbBact-based analysis shows an overrepresentation of soil- and

plant-related bacteria inside the exclusion zone, while skin bacteria of bank voles near Kyiv were enriched in human and farm animal *terms*. This leads to the hypothesis that the difference between the two sample groups is due to contact with humans and farm animals rather than to exposure to radioactivity (Supplementary Figure S13). (*ii*) Similarly, Risely et al. (71) observed strong diurnal oscillations in the microbiome composition of South African wild meerkats' fecal samples. dbBact-based analysis indicates that this effect is driven by a large number of soil/rhizosphere-related bacteria appearing in the afternoon fecal samples (Supplementary Figure S14). (*iii*) dbBact analysis of air samples taken by Gat et al. (72) during clear days in Israel shows human-associated and farm-associated *terms* as a source of air bacteria, compared to samples taken during a dust storm, which display desert and soil-associated bacteria (Supplementary Figure S15). Hence, fecal bacteria from human and farm animals are airborne during ambient weather conditions, whereas dust storms bring over desert and soil associated bacteria. (*iv*) Finally, analysis of river water samples in two locations near the Bronx River estuary (73) shows that the difference in bacterial communities in these two locations is partially explained by ocean vs. freshwater bacteria, probably related to the salinity levels in the two sample locations (Figure 6D and Supplementary Figure S16).

Detecting potential cross-sample contaminations: dbBact allows the straightforward detection of potential contaminants. Each bacterium in a study may be assigned its best fitting dbBact *term*, thus 'awkward' bacteria may be detected and discarded from downstream analysis. (*i*) Caporaso et al. (74) followed the oral, skin, and fecal microbiome of an individual using daily samples for a year. dbBact-based analysis detected a group of skin-associated *sequences* in a subset of fecal samples, indicating a potential contamination (Figure 6E and Supplementary Figure S17). (*ii*) Similarly, in a dataset of infant nasopharyngeal samples (75), we observed a cluster of mouse-associated *sequences* (Supplementary Figure S18), which may be attributed to a contamination or to low biomass kit-related bacteria. As these mouse-associated sequences are evenly spread across the sample types, they did not introduce a systemic bias in the authors' results. But removal of the *sequences* before downstream analysis reduces inter-sample noise and increases the statistical power (Supplementary Figure S18c,d). In addition, dbBact can detect reagent-borne contaminants using *sequences* flagged as 'CONTAMINANT' (Supplementary Figure S20). Identifying ASVs as reagent-borne or potentially originating from cross-sample contamination may be context-dependent, and therefore should not be performed automatically (Supplementary Figure S20).

DISCUSSION

dbBact integrates 16S rRNA microbiome studies into a collaborative, coherent body of knowledge that facilitates pan-microbiome analysis of new studies using a rigorous statistical and algorithmic framework.

An important advantage of dbBact, compared to standard meta-analysis methods, is that the latter may suffer from the 'streetlight effect' (76). For example, when examining the effect of fruit consumption in the American Gut Project (Figure 6A), one might consider including other diet-related studies in the meta-analysis. But this would miss the link between high fruit consumption and primate-associated bacteria. dbBact retrieves *annotations* from a wide range of sample types and habitats, providing additional and potentially unexpected insights into the biological contexts in which specific bacteria appear.

Terms in dbBact *annotations* are based on ontologies, providing a common language for phenotype description. The tree structure of ontologies facilitates the discovery of commonalities between bacteria in studies conducted under similar, albeit not identical, conditions. For example, data from Crohn's disease and ulcerative colitis *experiments* may be combined based on their ontological 'parent' *term* 'inflammatory bowel disease.' Moreover, many 'cross-sectional' questions may be asked and possibly answered using dbBact. For instance, what *terms* are similar with respect to their bacteria (e.g. are dogs more similar to cats or to wolves?), or are there connections between phylogeny and specific phenotypes (e.g. does genus X appear only in host Y or in geographic location Z?).

Apart from putting forth novel hypotheses, dbBact makes possible the detection of sources of bacterial groups. We recommend querying dbBact as a first step in any microbiome analysis (e.g. using the interactive heatmap of the dbBact-Calour module). Identifying relevant bacterial groups and their dbBact *annotations* fosters an initial understanding of biological processes, supporting better downstream analysis. dbBact also enables detecting potential cross-sample contaminations. We have encountered numerous cases where examining ASVs in a study identified bacteria of another origin (e.g. 'mus musculus' *annotations* in human samples). Removing these sequences prior to downstream analysis can remove biases and increase the statistical power. In addition, using the 'CONTAMINANT' *annotation* dbBact may detect reagent-borne contaminants. However, we emphasize that identifying bacteria as reagent-borne or potentially originating from cross-sample contamination may be context-dependent, and therefore should not be performed automatically (Supplementary Figure S20).

The current coverage of dbBact is high in a large number of habitats, but many other habitats are still poorly covered (e.g. Figure 3E). Therefore, *terms* appearing in a small number of *annotations* may lead to dubious conclusions. For example, dbBact contains a single *experiment* originating from a scrubland environment. This *experiment* profiled the leaf microbiome of ivy plants, hence, querying a set of ivy leaf-related bacteria may result in the enrichment of both 'ivy' and 'scrubland' *terms*. Therefore, to avoid incorrect conclusions, users are advised to further examine the set of *experiments* associated with each enriched *term*. As the number and diversity of *experiments* in dbBact increase, such spurious *terms* are expected to be suppressed. Tens of microbiome studies are published weekly, but the dbBact team can process only a fraction. We believe that researchers may benefit from uploading their studies to dbBact, as contributing *annotations* increases the impact of a study by enabling other researchers to use its results. We, therefore, hope the microbiome community will contribute to dbBact and help increase the number and diversity of

uploaded studies. We intend to encourage such efforts and are committed to providing proper training and support to the community. As the number of contributors grows, *annotations*' quality may vary, and hence dbBact curators would review new *annotations*. In addition, dbBact enables users to flag potentially incorrect *annotations*, for later inspection by the dbBact team.

dbBact may also be used in shotgun metagenomics studies. Whenever 16S rRNA sequences are inferred from shotgun data they may be submitted as queries or uploaded to dbBact. The 'linking' mechanism for harmonizing studies from different variable regions enables shotgun and amplicon studies to be integrated into one coherent knowledge base. Similarly, studies using long read technologies (or synthetic long reads) also provide full-length 16S rRNA sequences, and thus can be integrated into dbBact in the same manner.

In conclusion, dbBact augments current microbiome analysis pipelines by systemically and comprehensively examining the contexts associated with each ASV. Examining the key *terms* associated with a study's ASVs, by word clouds or *term*-based PCA, unravels the general factors shaping microbial communities, sometimes pointing to unexpected cross-habitat similarities. *Term* enrichment across different sample groups can be statistically analyzed, potentially leading to better understanding of the underlying processes affecting the differences between these groups. Also, contamination-associated ASVs can be identified and potentially excluded from downstream analysis.

In sum, dbBact introduces a new 'layer' of data analysis in microbiome studies. We believe that the large scope, ontology-based structure and associated statistical methods of dbBact provide new means for studying core factors affecting bacterial communities, possibly answering questions that could not have otherwise been asked.

DATA AVAILABILITY

dbBact website http://dbbact.org

dbBact website source code https://github.com/amnona/dbbact-website (DOI: 10.5281/zenodo.7961853)

dbBact REST-API server api.dbbact.org

Documentation for the REST-API api.dbbact.org/docs

Examples for using the REST-API interface https://github.com/amnona/dbbact-examples (DOI: 10.5281/zenodo.7958467)

Source code for the REST-API server https://github.com/amnona/dbbact-server (DOI: 10.5281/zenodo.7961828)

Source code for the REST-API sequence translation server https://github.com/amnona/dbbact-sequence-translator (DOI: 10.5281/zenodo.7961754)

Jupyter notebooks for figures presented in the paper https://github.com/amnona/dbbact-paper (DOI: 10.5281/zenodo.7958481)

Weekly dump of the complete dbBact database (excluding user details) https://dbbact.org/download

dbBact v2022.07.01 database dump (used for the examples in the paper) (DOI: 10.5281/zenodo.7961961)

dbBact-calour plugin https://github.com/amnona/dbbact-calour (DOI: 10.5281/zenodo.7961875)

Qiime2 dbBact plugin https://github.com/amnona/q2-dbbact (DOI: 10.5281/zenodo.7961750)

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We wish to thank Tzipi Brown and Rotem Hadar from the Sheba Microbiome Center, Zhenjiang (Zech) Xu, Jon Sanders, Qiyun Zhu, Tomasz Kosciolek, Stefan Janssen and Jeremiah Minich for fruitful discussions, feedback, and suggestions.

Author contributions: Conceptualization, A.A. and N.S.; Methodology, A.A., N.S. and E.O.; Software, A.A. and E.O.; Validation, A.A. and N.S.; Formal Analysis, A.A. and N.S.; Resources, A.A., N.S. and Y.H.; Data Curation, A.A.; Writing – Original Draft, A.A. and N.S.; Writing – Review & Editing, A.A., N.S. and Y.H.; Visualization, A.A. and N.S.; Funding Acquisition, N.S. All authors read and approved the final manuscript.

FUNDING

N.S. is funded by the Ministry of Science, Technology & Space, Israel [3-16033]. Funding for open access charge: Open university internal fund.

Conflict of interest statement. None declared.

REFERENCES

1. Smil,V. (2003) In: *The Earth's Biosphere: Evolution, Dynamics, and Change*. MIT Press.
2. Herlemann,D.P.R., Labrenz,M., Jürgens,K., Bertilsson,S., Waniek,J.J. and Andersson,A.F. (2011) Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *ISME J.*, **5**, 1571–1579.
3. Bahram,M., Hildebrand,F., Forslund,S.K., Anderson,J.L., Soudzilovskaia,N.A., Bodegom,P.M., Bengtsson-Palme,J., Anslan,S., Coelho,L.P., Harend,H. *et al.* (2018) Structure and function of the global topsoil microbiome. *Nature*, **560**, 233–237.
4. Peiffer,J.A., Spor,A., Koren,O., Jin,Z., Tringe,S.G., Dangl,J.L., Buckler,E.S. and Ley,R.E. (2013) Diversity and heritability of the maize rhizosphere microbiome under field conditions. *Proc. Natl. Acad. Sci. U.S.A.* , **110**, 6548–6553.
5. Muegge,B.D., Kuczynski,J., Knights,D., Clemente,J.C., González,A., Fontana,L., Henrissat,B., Knight,R. and Gordon,J.I. (2011) Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science*, **332**, 970–974.
6. Goodrich,J.K., Waters,J.L., Poole,A.C., Sutter,J.L., Koren,O., Blekhman,R., Beaumont,M., Van Treuren,W., Knight,R., Bell,J.T. *et al.* (2014) Human genetics shape the gut microbiome. *Cell*, **159**, 789–799.
7. Mcdonald,D., Hyde,E., Debelius,J.W., Morton,J.T., Gonzalez,A., Ackermann,G., Aksenov,A.A., Behsaz,B., Brennan,C., Chen,Y. *et al.* (2018) American gut: an open platform for citizen science microbiome research. *Msystems*, **3**, e00031-18.
8. Huttenhower,C., Gevers,D., Knight,R., Abubucker,S., Badger,J.H., Chinwalla,A.T., Creasy,H.H., Earl,A.M., Fitzgerald,M.G., Fulton,R.S. *et al.* (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.
9. Mirzayi,C., Renson,A., Furlanello,C., Sansone,S.-A., Zohra,F., Elsafoury,S., Geistlinger,L., Kasselman,L.J., Eckenrode,K., van de Wijgert,J. *et al.* (2021) Reporting guidelines for human microbiome research: the STORMS checklist. *Nat. Med.*, **27**, 1885–1892.

10. Burcham,Z.M., Garneau,N.L., Comstock,S.S., Tucker,R.M., Knight,R., Metcalf,J.L., Miranda,A., Reinhart,B., Meyers,D., Woltkamp,D. *et al.* (2020) Patterns of oral microbiota diversity in adults and children: a crowdsourced population study. *Sci. Rep.*, **10**, 2133.
11. Thompson,L.R., Sanders,J.G., Mcdonald,D., Amir,A., Ladau,J., Locey,K.J., Prill,R.J., Tripathi,A., Gibbons,S.M., Ackermann,G. *et al.* (2017) A communal catalogue reveals Earth's multiscale microbial diversity. Meta-Analysis, **551**, 457–463.
12. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D1086–D1098.
13. Parr,C.S., Wilson,N., Leary,P., Schulz,K.S., Lans,K., Walley,L., Hammock,J.A., Goddard,A., Rice,J., Studer,M. *et al.* (2014) The Encyclopedia of Life v2: providing global access to knowledge about life on earth. *Biodiversity Data J.*, **2**, e1079.
14. Amir,A., McDonald,D., Navas-Molina,J.A., Kopylova,E., Morton,J.T., Xu,Z.Z., Knightly,E.P., Thompson,L.R., Hyde,E.R., Gonzalez,A. *et al.* (2017) Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems*, **2**, e00191-16.
15. Callahan,B.J., McMurdie,P.J., Rosen,M.J., Han,A.W., Johnson,A.J.A. and Holmes,S.P. (2016) DADA2: high resolution sample inference from Illumina amplicon data. *Nat. Methods*, **13**, 581–583.
16. Edgar,R. (2016) UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv doi: https://doi.org/10.1101/081257, 15 October 2016, preprint: not peer reviewed.
17. Prodan,A., Tremaroli,V., Brolin,H., Zwinderman,A.H., Nieuwdorp,M. and Levin,E. (2020) Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLoS One*, **15**, e0227434.
18. Dennis,G., Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C. and Lempicki,R.A. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.*, **4**, R60.
19. Raudvere,U., Kolberg,L., Kuzmin,I., Arak,T., Adler,P., Peterson,H. and Vilo,J. (2019) G:profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.*, **47**, W191–W198.
20. Gonzalez,A., Navas-Molina,J.A., Kosciolek,T., McDonald,D., Vázquez-Baeza,Y., Ackermann,G., DeReus,J., Janssen,S., Swafford,A.D., Orchanian,S.B. *et al.* (2018) Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods*, **15**, 796–798.
21. McDonald,D., Kaehler,B., Gonzalez,A., DeReus,J., Ackermann,G., Marotz,C., Huttley,G. and Knight,R. (2019) redbiom: a Rapid Sample Discovery and Feature Characterization System. *Msystems*, **4**, e00215-19.
22. Mitchell,A.L., Almeida,A., Beracochea,M., Boland,M., Burgin,J., Cochrane,G., Crusoe,M.R., Kale,V., Potter,S.C., Richardson,L.J. *et al.* (2020) MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.*, **48**, D570–D578.
23. Cheng,L., Qi,C., Zhuang,H., Fu,T. and Zhang,X. (2020) GutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Res.*, **48**, D554–D560.
24. Geistlinger,L., Mirzayi,C., Zohra,F., Azhar,R., Elsafoury,S., Grieve,C., Wokaty,J., Gamboa-Tuz,S.D., Sengupta,P., Hecht,I. *et al.* (2022) BugSigDB: accelerating microbiome research through systematic comparison to published microbial signatures. medRxiv doi: https://doi.org/10.1101/2022.10.24.22281483, 01 November 2022, preprint: not peer reviewed.
25. Janssens,Y., Nielandt,J., Bronselaer,A., Debunne,N., Verbeke,F., Wynendaele,E., Van Immerseel,F., Vandewynckel,Y.P., De Tré,G. and De Spiegeleer,B. (2018) Disbiome database: linking the microbiome to disease. *BMC Microbiol.*, **18**, 4–9.
26. Leinonen,R., Sugawara,H. and Shumway,M. (2011) The sequence read archive. *Nucleic. Acids. Res.*, **39**, D19–D21.
27. Proctor,L.M., Creasy,H.H., Fettweis,J.M., Lloyd-Price,J., Mahurkar,A., Zhou,W., Buck,G.A., Snyder,M.P., Strauss,J.F., Weinstock,G.M. *et al.* (2019) The Integrative Human Microbiome Project. *Nature*, **569**, 641–648.
28. Parente,E., De Filippis,F., Ercolini,D., Ricciardi,A. and Zotta,T. (2019) Advancing integration of data on food microbiome studies: foodMicrobionet 3.1, a major upgrade of the FoodMicrobionet database. *Int. J. Food Microbiol.*, **305**, 108249.
29. Yao,G., Zhang,W., Yang,M., Yang,H., Wang,J., Zhang,H., Wei,L., Xie,Z. and Li,W. (2020) MicroPhenoDB associates metagenomic data with pathogenic microbes, microbial core genes, and human disease phenotypes. *Genomics*, **18**, 760–772.
30. Skoufos,G., Kardaras,F.S., Alexiou,A., Kavakiotis,I., Lambropoulou,A., Kotsira,V., Tastsoglou,S. and Hatzigeorgiou,A.G. (2021) Peryton: a manual collection of experimentally supported microbe-disease associations. *Nucleic Acids Res.*, **49**, D1328–D1333.
31. Yang,J., Park,J., Park,S., Baek,I. and Chun,J. (2019) Introducing murine microbiome database (MMDB): a curated database with taxonomic profiling of the healthy mouse gastrointestinal microbiome. *Microorganisms*, **7**, 480.
32. Moitinho-Silva,L., Nielsen,S., Amir,A., Gonzalez,A., Ackermann,G.L., Cerrano,C., Astudillo-Garcia,C., Easson,C., Sipkema,D., Liu,F. *et al.* (2017) The sponge microbiome project. *GigaScience*, **6**, gix077.
33. Quast,C., Pruesse,E., Yilmaz,P., Gerken,J., Schweer,T., Yarza,P., Peplies,J. and Glöckner,F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–D596.
34. Bolyen,E., Rideout,J.R., Dillon,M.R., Bokulich,N.A., Abnet,C.C., Al-Ghalith,G.A., Alexander,H., Alm,E.J., Arumugam,M., Asnicar,F. *et al.* (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.*, **37**, 852–857.
35. Xu,Z.Z., Amir,A., Sanders,J.G., Zhu,Q., Morton,J.T., Bletz,M.C., Tripathi,A., Huang,S., McDonald,D., Jiang,L. *et al.* (2019) Calour: an Interactive, Microbe-Centric Analysis Tool. **4**, e00269-18.
36. Ijaz,U.Z., Quince,C., Hanske,L., Loman,N., Calus,S.T., Bertz,M., Edwards,C.A., Gaya,D.R., Hansen,R., McGrogan,P. *et al.* (2017) The distinct features of microbial 'dysbiosis' of Crohn's disease do not occur to the same extent in their unaffected, genetically-linked kindred. *PLoS One*, **12**, e0172605.
37. Schriml,L.M., Mitraka,E., Munro,J., Tauber,B., Schor,M., Nickle,L., Felix,V., Jeng,L., Bearer,C., Lichenstein,R. *et al.* (2019) Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.*, **47**, D955–D962.
38. Buttigieg,P.L., Pafilis,E., Lewis,S.E., Schildhauer,M.P., Walls,R.L. and Mungall,C.J. (2016) The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *J. Biomed. Semantics.*, **7**, 57.
39. Buttigieg,P.L., Morrison,N., Smith,B., Mungall,C.J. and Lewis,S.E. (2013) The environment ontology: contextualising biological and biomedical entities. *J. Biomed. Semantics*, **4**, 43.
40. Whetzel,P.L., Noy,N.F., Shah,N.H., Alexander,P.R., Nyulas,C., Tudorache,T. and Musen,M.A. (2011) BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res.*, **39**, W541–W545.
41. Mungall,C.J., Torniai,C., Gkoutos,G.V., Lewis,S.E. and Haendel,M.A. (2012) Uberon, an integrative multi-species anatomy ontology. *Genome Biol.*, **13**, R5.
42. Malone,J., Holloway,E., Adamusiak,T., Kapushesky,M., Zheng,J., Kolesnikov,N., Zhukova,A., Brazma,A. and Parkinson,H. (2010) Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, **26**, 1112–1118.
43. Xu,Z.Z., Amir,A., Sanders,J.G., Zhu,Q., Morton,J.T., Bletz,M.C., Tripathi,A., Huang,S., McDonald,D., Jiang,L. *et al.* (2019) Calour: an Interactive, Microbe-Centric Analysis Tool. **4**, e00269-18.
44. Giloteaux,L., Goodrich,J.K., Walters,W.A., Levine,S.M., Ley,R.E. and Hanson,M.R. (2016) Reduced diversity and altered composition of the gut microbiome in individuals with myalgic encephalomyelitis/chronic fatigue syndrome. *Microbiome*, **4**, 30 .
45. Zmora,N., Zilberman-Schapira,G., Suez,J., Mor,U., Dori-Bachash,M., Bashiardes,S., Kotler,E., Zur,M., Regev-Lehavi,D., Brik,R.B.Z. *et al.* (2018) Personalized gut mucosal colonization resistance to empiric probiotics is associated with unique host and microbiome features. *Cell*, **174**, 1388–1405.
46. Liu,Y., Liu,B., Liu,C., Hu,Y., Liu,C., Li,X., Li,X., Zhang,X., Irwin,D.M., Wu,Z. *et al.* (2021) Differences in the gut microbiomes of dogs and wolves: roles of antibiotics and starch. *BMC Vet. Res.*, **17**, 112.

47. Wang,Q., Garrity,G.M., Tiedje,J.M. and Cole,J.R. (2007) Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, **73**, 5261–5267.
48. Mandal,S., Van Treuren,W., White,R.A., Eggesbø,M., Knight,R. and Peddada,S.D. (2015) Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb.Ecol. Health Dis.*, **26**, 27663.
49. Morton,J.T., Marotz,C., Washburne,A., Silverman,J., Zaramela,L.S., Edlund,A., Zengler,K. and Knight,R. (2019) Establishing microbial composition measurement standards with reference frames. *Nat. Commun.*, **10**, 2719.
50. Fernandes,D.,A., Reid,J., Macklaim,M.,J., McMurrough,T.A. and Edgell,D.R. (2014) Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, **2**, 15.
51. Brill,B., Amir,A. and Heller,R. (2020) Testing for differential abundance in compositional counts data, with application to microbiome studies. arXiv doi: https://arxiv.org/abs/1904.08937, 30 march 2020, preprint: not peer reviewed.
52. Lin,H. and Peddada,S.D. (2020) Analysis of microbial compositions: a review of normalization and differential abundance analysis. *NPJ Biofilms Microbiomes*, **6**, 60.
53. Jiang,L., Amir,A., Morton,J.T., Heller,R., Arias-Castro,E. and Knight,R. (2017) Discrete false-discovery rate improves identification of differentially abundant microbes. *Msystems*, **2**, e00092-17.
54. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Ser. B (Methodological)*, **57**, 289–300.
55. Amir,A., McDonald,D., Navas-Molina,J.A., Debelius,J., Morton,J.T., Hyde,E. and Robbins-pianka,A. (2017) Correcting for microbial blooms in fecal samples during room temperature shipping amnon. *Msystems*, **2**, e00199-16.
56. Hägglund,M., Bäckman,S., Macellaro,A., Lindgren,P., Borgmästars,E., Jacobsson,K., Dryselius,R., Stenberg,P., Sjödin,A., Forsman,M. *et al.* (2018) Accounting for bacterial overlap between raw water communities and contaminating sources improves the accuracy of signature-based microbial source tracking. *Front.Microbiol.*,**9**, 2364.
57. DeSantis,T.Z., Hugenholtz,P., Larsen,N., Rojas,M., Brodie,E.L., Keller,K., Huber,T., Dalevi,D., Hu,P. and Andersen,G.L. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, **72**, 5069–5072.
58. Nejman,D., Livyatan,I., Fuks,G., Gavert,N., Zwang,Y., Geller,L.T., Rotter-Maskowitz,A., Weiser,R., Mallel,G., Gigi,E. *et al.* (2020) The human tumor microbiome is composed of tumor type–specific intracellular bacteria. *Science*, **368**, 973–980.
59. Griffin,N.W., Ahern,P.P., Cheng,J., Heath,A.C., Ilkayeva,O., Newgard,C.B., Fontana,L. and Gordon,J.I. (2017) Prior dietary practices and connections to a human gut microbial metacommunity alter responses to diet interventions. *Cell Host Microbe*, **21**, 84–96.
60. Obregon-Tito,A.J., Tito,R.Y., Metcalf,J., Sankaranarayanan,K., Clemente,J.C., Ursell,L.K., Zech Xu,Z., Van Treuren,W., Knight,R., Gaffney,P.M. *et al.* (2015) Subsistence strategies in traditional societies distinguish gut microbiomes. *Nat. Commun.*, **6**, 6505.
61. Dudek,N.K., Switzer,A.D., Costello,E.K., Murray,M.J., Tomoleoni,J.A., Staedler,M.M., Tinker,M.T. and Relman,D.A. (2022) Characterizing the oral and distal gut microbiota of the threatened southern sea otter (Enhydra lutris nereis) to enhance conservation practice. *Conserv. Sci. Pract.*, **4**, e12640.
62. Wang,W., Gao,X., Zheng,S., Lancuo,Z., Li,Y., Zhu,L., Hou,J., Hai,J., Long,X., Chen,H. *et al.* (2021) The gut microbiome and metabolome of Himalayan Griffons (Gyps himalayensis): insights into the adaptation to carrion-feeding habits in avian scavengers. *Avian Res.*, **12**, 52,
63. Arnold,C.E., Pilla,R., Chaffin,M.K., Leatherwood,J.L., Wickersham,T.A., Callaway,T.R., Lawhon,S.D., Lidbury,J.A., Steiner,J.M. and Suchodolski,J.S. (2021) The effects of signalment, diet, geographic location, season, and colitis associated with antimicrobial use or Salmonella infection on the fecal microbiome of horses. *J. Vet. Intern. Med.*, **35**, 2437–2448.
64. Abbas-Egbariya,H., Haberman,Y., Braun,T., Hadar,R., Denson,L., Gal-Mor,O. and Amir,A. (2022) Meta-analysis defines predominant shared microbial responses in various diseases and a specific inflammatory bowel disease signal. *Genome Biol.*, **23**, 61.
65. Scheithauer,T.P.M., Davids,M., Winkelmeijer,M., Verdoes,X., Aydin,Ö., de Brauw,M., van de Laar,A., Meijnikman,A.S., Gerdes,V.E.A., van Raalte,D. *et al.* (2022) Compensatory intestinal antibody response against pro-inflammatory microbiota after bariatric surgery. *Gut. Microbes.*, **14**, 2031696.
66. Willis,J.R., González-Torres,P., Pittis,A.A., Bejarano,L.A., Cozzuto,L., Andreu-Somavilla,N., Alloza-Trabado,M., Valentín,A., Ksiezopolska,E., Company,C. *et al.* (2018) Citizen science charts two major 'stomatotypes' in the oral microbiome of adolescents and reveals links with habits and drinking water composition. *Microbiome*, **6**, 218.
67. Yeoh,Y.K., Chan,M.H., Chen,Z., Lam,E.W.H., Wong,P.Y., Ngai,C.M., Chan,P.K.S. and Hui,M. (2019) The human oral cavity microbiota composition during acute tonsillitis: a cross-sectional survey. *BMC Oral Health*, **19**, 275.
68. Zhu,Y., He,C., Li,X., Cai,Y., Hu,J., Liao,Y., Zhao,J., Xia,L., He,W., Liu,L. *et al.* (2018) Gut microbiota dysbiosis worsens the severity of acute pancreatitis in patients and mice. *J. Gastroenterol.*, **54**, 347–358.
69. Duvallet,C., Gibbons,S.M., Gurry,T., Irizarry,R.A. and Alm,E.J. (2017) Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.*, **8**, 1784.
70. Lavrinienko,A., Tukalenko,E., Mappes,T. and Watts,P.C. (2018) Skin and gut microbiomes of a wild mammal respond to different environmental cues. *Microbiome*, **6**, 209–209.
71. Risely,A., Wilhelm,K., Clutton-Brock,T., Manser,M.B. and Sommer,S. (2021) Diurnal oscillations in gut bacterial load and composition eclipse seasonal and lifetime dynamics in wild meerkats. *Nat. Commun.*, **12**, 6017.
72. Gat,D., Mazar,Y., Cytryn,E. and Rudich,Y. (2017) Origin-dependent variations in the atmospheric microbiome community in eastern Mediterranean dust storms. *Environ. Sci. Technol.*, **51**, 6709–6718.
73. Naro-Maciel,E., Ingala,M.R., Werner,I.E. and Fitzgerald,A.M. (2020) 16S rRNA amplicon sequencing of urban prokaryotic communities in the South Bronx River Estuary. *Microbiol. Resour. Announc.*, **9**, e00182-20.
74. Caporaso,J.G., Lauber,C.L., Costello,E.K., Berg-Lyons,D., Gonzalez,A., Stombaugh,J., Knights,D., Gajer,P., Ravel,J., Fierer,N. *et al.* (2011) Moving pictures of the human microbiome. *Genome Biol.*, **12**, R50.
75. Xu,L., Earl,J. and Pichichero,M.E. (2021) Nasopharyngeal microbiome composition associated with Streptococcus pneumoniae colonization suggests a protective role of Corynebacterium in young children. *PLoS One*, **16**, e0257207.
76. Freedman,D.H. (2010) Why scientific studies are so often wrong: the streetlight fffect. *Discover*, **31**, 55–57.