Efilter: A Tool for Identifying Unexpected and Erroneous Taxids in Sequencing Data

Sharon Bewick (sharon_bewick@hotmail.com)
University of Maryland https://orcid.org/0000-0002-2563-5761
Xianghui Dong
University of Maryland
David Karig
Clemson University
William F. Fagan
University of Maryland

Research

Keywords: taxonomic identification, microbial habitat association, online tool, metagenomics datasets, contamination, bioinformatics errors, novel niches, taxon discovery

Posted Date: May 21st, 2020

DOI: https://doi.org/10.21203/rs.3.rs-29648/v1

Abstract

Background: Multiple types of error can enter metagenomics sample analysis, including contamination during sample collection, library preparation and sequencing, as well as incorrect taxonomic assignment by bioinformatics packages. Often, such errors either go unidentified, or else are removed, *ad hoc*, based on user knowledge of microbial ecology. However, becausedifferent researchers are more or less familiar with the ecologies of the various organisms in their systems, filtering is applied non-uniformly at best, with differences in the degree of filtering between studies and between taxonomic groups within a single study.

Results: In this paper, we present EFILTER – a tool that capitalizes on decades of research in microbial ecology to identify suspicious or spurious taxa in metagenomics samples based on habitat information.

Conclusions: EFILTER allows all microbiome researchers, regardless of background, to examine taxon lists for unusual entries, and to do so in a manner that is systematic and without bias.

Background

The ability to extract, clone and analyze DNA directly from environmental samples has revolutionized disciplines ranging from biomedical science to environmental microbiology[1-3]. Compared to historic, culture-based approaches,sequencing methods offer unprecedented advancements in terms of cost, effort and speed. They also improve detection of full community membership by exposing unculturable organisms [4]. However, sequencing studies are not without their own set of difficulties.One of the greatest challenges lies in interpreting sequencingresults, which can be complicated by issues of contamination [5-7], sequencing error [8, 9], and bioinformatics misidentification[10]. Although improved laboratory and bioinformatics techniques canhelp to alleviate these difficulties, increased focus on detecting rare taxa, along with more specific classification down to species and even strains, means that interpretation of sequencing data will remain a significant challenge for the foreseeable future.

Contamination is perhaps the largest and most difficult to address source of confusion in sequencing data. Whereasculture-based methods only detect organisms that can be reliably and repeatedly grown from an environmental sample, sequencing techniques identify even small pieces of non-viable DNA. While this is one of the strengths of the sequencing approach, and the main reason that it can uncover unculturable organisms, thisis also why sequencing methods arefar more susceptible to issues with contamination[5-7]. Unfortunately, contamination can occur at any step along the processing pipeline, ranging from issues during sample collection and storage [11], though reagent contamination[12-15] to room source contamination andcontaminating bacteria from the mouth and skin of the researcher during sample preparation [16, 17].While it is relatively straightforward to detect contaminants in samples that should contain a single organism, for example microbial DNA in the cow genome [5], or contamination of pure bacterial cultures with other bacterial taxa [18], identifying contamination in complex,mixed microbial communities presents a more difficult, and still open challenge.

Another major issue with sequencing data is imperfect taxonomic assignment. This challenge is also not easy to overcome, since there is an inherent trade-off between finding all taxa within a sample, and confidence/accuracy in taxon prediction [10]. Thus, bioinformatics methods that are highly sensitive are also more likely to report organisms not actually present. Meanwhile, more conservative classification methods can reduce the number of false positives, but do so at a cost of increasing false negatives. Misclassificationcan be more or less problematic, depending on the microbial community involved, the type of sequencing method employed[19-21], the bioinformatics toolkit used[10, 22, 23], the reference databases available [24],and the quality (e.g., sequencing platform statistics, read depth, and read length) of the sequencing data[25-27].As with contamination, even in simple systems with appropriately chosen sequencing and bioinformatics methods, classification errors can have a significant impact on conclusions, for example by massively inflating diversity or the relative importance of rare organisms.

One method for identifying potentially problematic taxIDs in sequencing data is to rely on knowledge of microbial ecology. Finding a common plant pathogen in a human gut microbiome dataset, for example, might raise a red flag. Whether consciously or not, researchers use this approach to clean their data, for instance by deciding whether or not to trust all of the output taxa from a particular experiment [11]. Unfortunately, use of such information requires extensive knowledge of microbial behavior and characteristics – knowledge that is lacking among many, if not most researchers performing metagenomics analysis. Indeed, even for researchers with strong backgrounds in microbial ecology, the vast diversity of microbes from different environments means that it is impossible for any single researcher to be familiar with all common microbial habitats. Consequently, even with extensive knowledge of the ecologies of individual bacterial taxa, data verification based on microbial ecology is likely to be biased and *ad hoc*. Nevertheless, if such verification can be standardized and automated, it remains a promising method for addressing the challenges of spurious taxIDs in microbiome datasets.

In this paper, we present EFILTER (https://efilter.shinyapps.io/EFilter-app/)– a computational tool that performs***E****nvironmental* ***FILTER****ing* by leveraging known relationships between bacterial taxa and specific habitats in order to identify taxIDs that are surprising or unlikely given the source of a metagenomics sample. EFILTER incorporates all habitat information available in*Bergey's Manual of Systematic Bacteriology* and *National Center for Biotechnology Information* (NCBI) records from the *International Journal of Systematic and Evolutionary Microbiology*(IJSEM). As such, EFILTER allows users to assess the likelihood of taxIDs being in their datasets based on known microbial ecology. Further, it does so without requiring an extensive background in microbial ecology, and also without introducing bias that might otherwise arise based on user experience and knowledge of specific taxonomic groups. To demonstrate the strength and performance of EFILTER, we test our tool on 4 datasets,which include both 16S rRNA and shotgun sequencing methods, analyzed with different bioinformatics pipelines and studied at different taxonomic ranks. Our analysis shows that EFILTER can reliably provide lists of unexpected/spurious organisms in metagenomics samples, offering a new,ecologically motivated tool for validation of microbiome datasets.

Materials And Methods

The EFilter Database

We constructed a database of microbe-environment associations using information from *Bergey's Manual of Systematic Bacteriology* and the National Center for Biotechnology Information (NCBI) Nucleotide database. For *Bergey's*, we converted pdf files of Volumes IIB, IIC, III, IV, and V as well as additional chapters on *Aquificae*, *Chlorobi*, *Chloroflexi*, *Chrysiogenetes*, *Crenarchaeota*, *Cyanobacteria*, *Deferribacteres*, *Deinococcus-Thermus*, *Euryarchaeota*, *Nitrospirae*, *Thermodesulfobacteria*, *Thermomicrobia*, and *Thermotogae* to .txt files using the pdfminer functions in Python. The text was then separated into segments for each taxonomic group. Text describing genera and species was mined for environmental information. We did not use text describing higher taxonomic groups, since this information tends to be too broad to be useful. To identify sentences containing environmental information within genus and species descriptions, we used the keywords in Table A.1 (see Appendix A). For the NCBI Nucleotide database, we only used entries with IJSEM[Journal] in the journal field. Our assumption was that all of these entries represent newly defined taxa. Although taxa can be defined in other journals, the *International Journal of Systematic and Evolutionary Microbiology* (IJSEM) represents one of the largest publishers of new taxon descriptions. Most other journals do not specialize in taxonomy, and thus may contain other types of studies where taxonomic specification is not precise/accurate. Our goal was to avoid including habitat information from inaccurate sources, since this is the very problem that EFILTER seeks to correct. Habitat information was taken from the 'isolation source/' field of the NCBI entries. NCBI/IJSEM entries were only used for taxa that did not exist in *Bergey's* (i.e., taxa that were defined since publication of the most recent *Bergey's* volume for that taxonomic group).

Habitat descriptions from *Bergey's Manual of Systematic Bacteriology* and NCBI were stored in tab-delimited .txt files (see Appendix B), along with relevant taxonomic information, forming the basis of our environmental database. For species, all searches are performed on species-level descriptions harvested from either *Bergey's* or NCBI/IJSEM. For genera, searches are performed on genus-level descriptions from *Bergey's*, as well as all species-level descriptions for species within the genus. For higher taxonomic ranks, searches are performed on all genus-level and species-level descriptions within the particular taxonomic group.

User-Defined Search Terms

The entire environmental database can be queried for keywords. In order to minimize search time, EFILTER contains a look-up table linking each unique keyword from the database (see above) to the species/genera in whose descriptions that keyword appears. That is, we built a table of keywords wherein each keyword (column) maps to many rows, with each row being a specific species/genus in the database. All keywords in our look-up table are full, stand-alone words. However, in many cases, users

may be interested in related words (e.g., ‘dog’ and ‘dogs’) as well. For this reason, whenever a keyword is typed into EFILTER, EFILTER automatically generates a list of all larger, stand-alone words (‘**dog**s, **dog**-bite’) containing the inputted keyword. This enables users to select the precise keyword combinations that they want to include. Thus, for the ‘dog’ example, the user may want to include ‘dogs’, ‘dog-bite’, and ’dogbite’ , but not ‘peptidoglycan’, ‘endogenous’, etc. Finally, the environmental database can be queried for phrases (e.g., ‘dogs and cats’). In this case, however, there is no look-up table, and the phrase is searched for directly through the database. Consequently, searches that involve querying for phrases are somewhat slower.

Pre-defined Environments

Although possible, it may be hard and time-consuming for individual users to think of all potential keywords associated with a particular microbial environment. For this reason, we pre-defined keyword combinations for 46 common habitats falling within six broad classes (animal, body sites, food, plant, environment and specialized). To generate keyword combinations, we manually curated all keywords appearing in ³5 species/genus descriptions and then went through these keywords individually to decide whether they or any of their word derivatives fell into our common environmental categories. We believe that our pre-defined environment feature will be the most useful for the majority of users. Lists of keyword combinations used for each pre-defined environment (ranging from 6 words for ‘sand’ to 199 words + derivatives for ‘other mammals’) can be found in Appendix C.

Results

General Usage

The EFILTER database contains 11633 species and 34 phyla (see Table B.1, Appendix B). 72% of species have habitat information specified, while 45% fall into at least one pre-defined environment (see Methods). 100% of phyla have habitat information specified, while 97.1% fall into at least one pre-defined environment. The pre-defined environment with the largest number of taxa is ‘soil’, which includes 17.5% of species. The next largest is ‘human’, which includes 8.4% of species. The pre-defined environments with the fewest number of taxa are ‘amphibian’, ‘reptile’ and ‘symbiont’, with 0.06%, 0.3% and 0.3% of species respectively (see Appendix B, Table B.2).

EFILTER inputs .txt files (including batch upload) in the form of lists of either Latin names or taxIDs. Inputs in the form of taxIDs are mapped onto Latin names in order to infer habitat associations. Only well defined taxa (i.e., taxa with entries in either *Bergeys* or NCBI/IJSEM, see Methods)areincluded in the database. EFILTER allows Latin names or taxIDs referencing any taxonomic rank from phylum to species, including lists with mixed ranks. Taxonomic rank is automatically determined based on name or taxID. Each taxon in an input file is queried against the EFILTER database using a defined environmental filter. Environmental filters can be constructed as user-defined keywords, user-defined phrases, or pre-

defined environments (see Methods). In addition, users can construct any logical combination ('AND', 'OR', 'NOT', for example 'eye AND human NOT other mammal') of single filters. A video tutorial covering filter definition can be found at (https://www.youtube.com/watch?v=UU1rPTUZPTE).

Upon file upload, EFILTER outputs the number of taxa at each taxonomic rank as well as the number of taxa that were not found in the EFILTER database. Missing taxa include taxa that are not well-defined, taxa that were defined more recently than the most recent EFILTER update, or taxa that are not bacterial (for example fungal taxa from the output of shotgun sequencing). These taxa are ignored, meaning that there is no need to curate shotgun sequencing data to include strictly bacteria. EFILTER also automatically reports the pre-defined environmental classes for each unique taxon in an uploaded dataset, starting from the input taxonomic rankup to the taxonomic rank of family.

The primary goal of EFILTER is to identify organisms that are consistent/inconsistent with the sampled environment (i.e., the user selected filter). Thus, when a taxon list is queried against a chosen filter, EFILTER outputs the number of taxa with at least one habitat in the database that matches (passes) the filter.Taxa without at least one matching habitat are identified as failing the filter. In addition, for each input taxon, EFILTER reports whether the taxon passed or failed at taxonomic ranks higher than the inputted rank. Finally, EFILTER outputs the names of taxa that passed the filter at taxonomic ranks lower than the inputted rank. This latter feature can be useful, for example, to make an educated guess as to which species might be present given a list of genera from a 16S rRNA sequencing run.

Samples versus Controls

As a first demonstration of EFILTER performance, we used our own data, which consisted of 16S rRNA sequencing of 375skin microbiome samples from the foreheads of 50 individuals, along with 100paired controls (blank swabs collected before and/or after each person was sampled). Figure 1 shows the percentage of genera passing four different filters averaged over the samples and paired controls for each person. For the 'skin' filter (A), more organisms passed in the samples relative to the controls for the majority (48/50 or 96%) of people. By contrast, for the 'built' filter (C) more organisms passed in the controls relative to the samples for the majority (44/50 or 88%) of people. As expected, for logical OR combinations of filters, the percentages of passing taxa were greater overall (compare, for example Figures 1A and 1B or Figures 1C and 1D). However, once again, for the 'human/mammal/body-site' filter (C) more taxa passed in the samples relative to the controls for the majority (40/50 or 80%) of people, whereas the reverse was true for the 'built/water' filter (D) for the majority (30/50 or 60%) of people. In general, one would expect that a skin/human/mammal/body-site filter would perform better on human samples than on non-human (e.g., control) samples, which is what we see. Likewise, assuming that many contaminants come from room air or sequencing preparation steps, one would assume that a built/water filter would perform better on control samples than on human samples, which is again what we see.

Samples from Different Sources

As a second demonstration of EFILTER performance, we comparedvarious EFILTER pre-defined environmental filters across a variety of environmental samples. Datasets were downloaded as lists of genera directly from the MG-RAST website (https://www.mg-rast.org/). Specifically, we considered hot springs (mgp5265, mgp6907, mgp5356, mgp5355, mgp5270, mgp5269, mgp5268, mgp5266), tropical soil (mgp4362), a freshwater lake (mgp19525), oceans (mgp20413), cheese (mgp14606), human nares (mgp385), and horses guts (mgp7746). All studies involved shotgun sequencing. We used the same number samples from each environment (8,limited by the availability of hot springs samples; in all cases we selected the 8 samples with the lowest MG-RAST ID numbers), and restricted our analysis to the top 25 most abundant taxa in each sample. Figure 2 shows confusion matrices for samples versus environmental filters.Focusing on the averaged confusion matrix (right panel), we see strong EFILTER performance, with red/orange (high percentage of passing taxa) concentrated along the diagonal and blue/green (low percentage of passing taxa) off-diagonal. Overall, across the seven paired environments and samples, an average of 72% of taxa passed their expected filters. Passing rates were highest (88.5%) for the anterior nares, and lowest (62.5%) for the lake.

Filter fidelity to sample was quite high. For five out of seven samples (hot springs, soil, oceans, cheese and anterior nares), the expected filter (extreme, soil, salt water, dairy, human) performed best. Even when this was not the case, the expected filter still performed well. For the lake study, for instance, both the salt water and soil filters outperformed the fresh water filter (note that all three are in the‘environment’ class, see Methods), though only marginally (63.5% and 62.8% of taxa passing for the salt water and soil filters, relative to 62.5% for the fresh water filter). Likewise, for the horse gut study, the human filter outperformed the other mammal filter (note that both are in the ‘animal’ class), although once again, the other mammal filter still performed well (79.4% passing for the human filter relative to 75.9% for the other mammal filter). Sample fidelity to filter was not as good. For three out of seven filters (extreme, soil and human) the highest percentages of passing taxa were identified in matched samples (hot springs, soil, and anterior nares respectively). For the remaining four filters, however, the highest percentages of passing taxa were identifiedin non-matched samples (the fresh and salt water filters identified the highest percentage of passing taxa in the soil sample, while the dairy and other mammal filtersidentified the highest percentage of passing taxa in the anterior nares).

Samples from Different Body Sites

As a third demonstration of EFILTER performance, we considered a shot-gun sequencing dataset from the Human Microbiome Project (HMP), including 690 samples from 15 body-sites. Importantly, in this dataset all samples fall under a single filter class (‘body-site’, see Methods). As in the previous section, we only considered the top 25 most abundant taxa in each sample. Figure 3 shows confusion matrices for species- and genus-level data for each sample individually (A), and as sample averages over

each body region (B; notice that, unlike Figure 2, there are different numbers of samples contributing to each body region).

EFILTER performance on the HMP dataset was even higher than on the environmental dataset, with an average of 40.9% of species and 76.0% of genera passing the appropriately matched filter. At the same time, however, filter fidelity to sample was lower (compare the right panel in Figure 3B to Figure 2). That is, at the rank of both genus and species, the expected filter (gut/digestive system, oral) performed best for only two (St, To+Kg+Bm+Sa+Sb+Sp+Td) out of five body regions. For the remaining three regions, the gut/digestive system and/or oral filters performed better. The gut/digestive system filter likely performed well because it has the largest number of taxa (see Appendix B). It is less clear why the oral filter out-performed the other filters, since it has fewer taxa than the ear/nose/throat filter at the ranks of both genus and species and has approximately the same number of taxa as the skin filter, atthe rank of genus.

HMP sample fidelity to filter, on the other hand, was strong, at least at the rank of species. In particular, four out of five filters (skin, oral, gut/digestive system, female reproductive system) identified the highest percentage of passing taxa in their matched samples (Rc, To+Kg+Bm+Sa+Sb+Sp+Td, St, Va). Sample fidelity to filter was slightly weaker at the rank of genus, where only two out of five filters (oral, gut/digestive system) identified the highest percentage of passing taxa in their matched samples. This fidelity, was similar to the fidelity observed in our environmental dataset (see Figure 2).One particularly striking feature in Figure 3 is the poor performance of all but the gut/digestive system filter on stool samples. Whereas >56% of genera from every other body region passed every body-site filter, only 17-35% of stool genera passednon-gut/digestive system filters. At the same time, however, 95% of stool genera passed the gut/digestive system filter. This suggests that stool taxa are unique to the gut environment, whereas taxa from other body sites exhibit broad body distributions.

Samples from Different Bioinformatics Pipelines

As a final demonstration of EFILTER performance, we compared EFILTER output for different bioinformatics pipelines. A common problem with microbiome studies is that different pipelines can give different taxon lists, including substitutions among distantly related organisms. EFILTER can be used to gauge performance of different bioinformatics methods and to identify potentially spurious taxa introduced through taxonomic assignment steps. To illustrate this, we used a shotgun sequencing dataset of the human skin microbiome [28], processed using both Kraken [29] and MetaPhlAn [30], and thresholded at various different read percentages. Results are shown in Figure 4, which also serves as an example of the EFILTER graphical output.

With larger thresholds, the pass:fail ratio increases for both bioinformatics pipelines. In other words, a larger fraction of the rare tail fails, regardless of the bioinformatics method used. This should not be surprising. Everything from contaminants and transient taxa to inherently rare/understudied organisms and bioinformatics errors are expected to contribute to the rare tail. In keeping with the consensus that

Kraken has a tendency to over classify [10, 31], the pass:fail ratio for Kraken is significantly lower than it is for MetaPhlAn. This is particularly true when there is no threshold, with Kraken givingpass rates of 42% and35% for genera and species respectively relative to MetaPhlAn's 67% and 61% at the same ranks. With a 1% threshold, both pipelines givesimilar results for species (72% passing), while Kraken actually shows a higher pass:fail ratio for genera (92% versus 88%). Surprisingly, though, despite Kraken's lower pass:fail ratio for most scenarios, MetaPhlAn usuallyidentifies a greater absolute number of passing taxa. Thus, while it is likely that some of the failing taxa identified by Kraken (and MetaPhlAn) are truly present on skin,based on overall performance as judged by environmental consistency of assigned taxa, MetaPhlAn appears to be the better pipeline for this dataset.

Discussion

A wealth of microbiome research has emerged over the past decade. Most of it, however, has been plagued by contamination, sequencing errors and taxonomic misidentification[6-10]. This has led to a lack of reproducibility across pipelines and amongst labs. Recently, there has been a call for improved standards and validation of microbiome datasets, including efforts like the MicroBiome Quality Control (MBQC) project [15] and the Critical Assessment of Metagenome Interpretation (CAMI) initiative (http://microbiome-cosi.org/cami), as well as standards development by the National Institute of Standards and Technology (NIST) [32]. In this paper, we introduce EFILTER as a new tool for helping to assess the quality of microbiome datasets and for identifying suspicious taxIDs within them. EFILTER is unique amongst microbiome informatics approaches in that it leverages ecological information to target organisms that are unexpected based on sample source.

Ultimately, EFILTER provides lists of suspicious taxIDs. The question, of course, is what to do with this information. In general, we advocate against indiscriminant dismissal of taxa that fail a chosen filter. Rather, two approaches are possible. First, depending on the analysis being performed, analysis can be run with and without discarding the problematic taxa in order to determine whether any conclusions change and, if they do, which and how many. In Figure 5, for example, we show how conclusions about relative sample diversity differ depending on whether analysis includes all taxa or only those taxa passing the salt water OR fresh water OR general water filter for the tara ocean dataset from MG-RAST (mgp20413). Notably, 93.6% of diversity orderings remain unchanged when failing taxa are ignored, suggesting that suspicious taxa are not overly problematic for the conclusions of thisparticular analysis with this particular dataset.

A second approach for using EFILTER data is to carefully examine taxa that fail and to make decisions about whether to include them based on additional knowledge or further experimentation. Depending on the system, the list of failing taxa can still be large. Thus, we recommend using broad filters and focusing on particularly egregious failures. That is, either failures that involve very abundant taxa, or else failures that extend to higher taxonomic ranks. We illustrate such an approach in Table C.1 of Appendix C for our own skin microbiome dataset. Notably, although we cannot be certain about the source of any taxon failures, we can make educated guesses about a sizeable fraction. Some taxa, for example

*Geodermatophilus*and *Methylobacterium*, are known contaminants of blanks from other sequencing studies, suggesting a contamination origin. Others, for example *Rhodothermus* and *Rheinheimera* are present in samples and controls taken at the same time or are ubiquitous in controls, again suggesting a contamination origin. Our lab has cultured certain suspicious taxa, for example *Enhydrobacter*, directly from skin, suggesting an unreported habitat for this organism. Finally, some taxa, including*Modestobacter* and *Hymenobacter*, are found in a range of samples and in no controls, hinting that they may be real, as yet undiscovered taxa from human skin. Supporting this claim, both *Modestobacter* and *Hymenobacter* have been found in other human microbiome datasets. *Modestobacter*, in particular, would be interesting to explore further. Although it is currently known from extreme (desert), rock/stone and soil environments it showed up in both 16S amplicon sequencing from our skin samples and in shotgun sequencing from a separate skin study performed in a different lab using an entirely different sequencing and bioinformatics pipeline[28]. This suggests that the source could be an as yet unknown species of *Modestobacter* that resides on human skin.

Although the goal of EFILTER is to identify suspicious taxa in metagenomics samples, it is worth pointing out that EFILTER also works reasonably well for determining sample source origin, at least when the possible sources are quite distinct. In Figure 2, for instance, it is possible to discriminate soil/water sources, animal sources and extreme sources based on the percentage of abundant (top 25) genera passing the associated filters. Such discrimination is more difficult for closely related sources, for example for sources from different mammals, sources from different types of water, or sources from different human body regions (see Figure 3). A future goal would be to extend EFILTER such that source discrimination is improved, possibly by leveraging additional habitat information from sequencing studies.

On a similar note, one of the downsides of the existing EFILTER database is that it only uses habitat information from validly published taxa descriptions. Although this ensures that contaminated or otherwise compromised metagenomics samples do not influence the EFILTER database, it means that EFILTER does not capitalize on the widely available, though imperfect, sequencing data currently in the public domain. A future direction would be to extend EFILTER to include these data, but to do so in a manner that incorporates confidence in the data source. This could result in higher percentages of taxa passing the appropriate filters, and also improved source discrimination.

Conclusions

Contamination, sequencing error and bioinformatics misidentifications will continue to plague microbiome research for the foreseeable future. Historically, many questionable results from sequencing studies have been addressed *ad hoc* by researchers who realize that particular organisms should not be present in particular samples. With EFILTER, we extend this capability to anyone involved in microbiome research, and do so in a way that allows rigorous, systematic, and repeatabledataset cleaning and validation based on ecologically relevant habitat information.

Declarations

Ethics approval and consent to participate:

Not applicable

Consent for publication:

Not applicable

Availability of data and material:

All data generated or analysed during this study are included in this published article [and its supplementary information files].

Competing Interests:

There are no competing interests.

Funding:

This material is based upon work supported by, or in part by, the U. S. Army Research Laboratory and the U. S. Army Research Office under contract/grant number #W911NF-14-1-0490.

Authors' contributions: SB conceived of the idea, built the database, and ran analyses on test datasets; XD developed the online platform in RShiny; SB, XD, DK and WFF contributed to idea development, results interpretation, and manuscript writing.

Acknowledgements:

Not applicable

References

1. Stahl DA, Lane DJ, Olsen G, Pace NR: **Analysis of hydrothermal vent-associated symbionts by ribosomal RNA sequences.***Science* 1984, **224:**409-412.
2. Amann RI, Ludwig W, Schleifer K-H: **Phylogenetic identification and in situ detection of individual microbial cells without cultivation.***Microbiological reviews* 1995, **59:**143-169.

3. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W: **Environmental genome shotgun sequencing of the Sargasso Sea.***science* 2004, **304:**66-74.
4. Riesenfeld CS, Schloss PD, Handelsman J: **Metagenomics: genomic analysis of microbial communities.***Annu Rev Genet* 2004, **38:**525-552.
5. Merchant S, Wood DE, Salzberg SL: **Unexpected cross-species contamination in genome sequencing projects.***PeerJ* 2014, **2:**e675.
6. Strong MJ, Xu G, Morici L, Bon-Durant SS, Baddoo M, Lin Z, Fewell C, Taylor CM, Flemington EK: **Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples.***PLoS Pathog* 2014, **10:**e1004437.
7. Lusk RW: **Diverse and widespread contamination evident in the unmapped depths of high throughput sequencing data.***PloS one* 2014, **9:**e110808.
8. Kunin V, Engelbrektson A, Ochman H, Hugenholtz P: **Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates.***Environmental microbiology* 2010, **12:**118-123.
9. Lee CK, Herbold CW, Polson SW, Wommack KE, Williamson SJ, McDonald IR, Cary SC: **Groundtruthing next-gen sequencing for microbial ecology–biases and errors in community structure estimates from PCR amplicon pyrosequencing.***PloS one* 2012, **7:**e44224.
10. Peabody MA, Van Rossum T, Lo R, Brinkman FS: **Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities.***BMC bioinformatics* 2015, **16:**362.
11. DeLong EF: **Microbial community genomics in the ocean.***Nature Reviews Microbiology* 2005, **3:**459-469.
12. Jervis-Bardy J, Leong LE, Marri S, Smith RJ, Choo JM, Smith-Vaughan HC, Nosworthy E, Morris PS, O'Leary S, Rogers GB: **Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of Illumina MiSeq data.***Microbiome* 2015, **3:**19.
13. Woyke T, Sczyrba A, Lee J, Rinke C, Tighe D, Clingenpeel S, Malmstrom R, Stepanauskas R, Cheng J-F: **Decontamination of MDA reagents for single cell whole genome amplification.***PloS one* 2011, **6:**e26161.
14. Motley ST, Picuri JM, Crowder CD, Minich JJ, Hofstadler SA, Eshoo MW: **Improved multiple displacement amplification (iMDA) and ultraclean reagents.***BMC genomics* 2014, **15:**443.
15. Sinha R, Abnet CC, White O, Knight R, Huttenhower C: **The microbiome quality control project: baseline study design and future directions.***Genome biology* 2015, **16:**276.
16. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW: **Reagent and laboratory contamination can critically impact sequence-based microbiome analyses.***BMC biology* 2014, **12:**87.
17. Weiss S, Amir A, Hyde ER, Metcalf JL, Song SJ, Knight R: **Tracking down the sources of experimental contamination in microbiome studies.***Genome biology* 2014, **15:**564.

18. Olson ND, Zook JM, Morrow JB, Lin NJ: **Using metagenomic methods to detect organismal contaminants in microbial materials.** PeerJ Preprints; 2017.
19. Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL: **Analysis of the microbiome: advantages of whole genome shotgun versus 16S amplicon sequencing.***Biochemical and biophysical research communications* 2016, **469:**967-977.
20. Qunfeng D, Claudia V: **Evaluation of the RDP classifier accuracy using 16S rRNA gene variable regions.***Metagenomics* 2012, **2012**.
21. Claesson MJ, Wang Q, O'sullivan O, Greene-Diniz R, Cole JR, Ross RP, O'toole PW: **Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions.***Nucleic acids research* 2010**:**gkq873.
22. Ricke DO, Shcherbina A, Chiu N: **Evaluating performance of metagenomic characterization algorithms using in silico datasets generated with FASTQSim.***bioRxiv* 2016**:**046532.
23. Hall RJ, Draper JL, Nielsen FG, Dutilh BE: **Beyond research: a primer for considerations on using viral metagenomics in the field and clinic.***Frontiers in microbiology* 2015, **6**.
24. Golob JL, Margolis E, Hoffman NG, Fredricks DN: **Evaluating the accuracy of amplicon-based microbiome computational pipelines on simulated human gut microbial communities.***BMC bioinformatics* 2017, **18:**283.
25. Schloss PD: **The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies.***PLoS Comput Biol* 2010, **6:**e1000844.
26. Wommack KE, Bhavsar J, Ravel J: **Metagenomics: read length matters.***Applied and environmental microbiology* 2008, **74:**1453-1463.
27. Mosher JJ, Bowman B, Bernberg EL, Shevchenko O, Kan J, Korlach J, Kaplan LA: **Improved performance of the PacBio SMRT technology for 16S rDNA sequencing.***Journal of microbiological methods* 2014, **104:**59-60.
28. Oh J, Byrd AL, Deming C, Conlan S, Barnabas B, Blakesley R, Bouffard G, Brooks S, Coleman H, Dekhtyar M: **Biogeography and individuality shape function in the human skin metagenome.***Nature* 2014, **514:**59.
29. Wood DE, Salzberg SL: **Kraken: ultrafast metagenomic sequence classification using exact alignments.***Genome biology* 2014, **15:**R46.
30. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C: **Metagenomic microbial community profiling using unique clade-specific marker genes.***Nature methods* 2012, **9:**811.
31. Clooney AG, Fouhy F, Sleator RD, O'Driscoll A, Stanton C, Cotter PD, Claesson MJ: **Comparing apples and oranges?: next generation sequencing and its impact on microbiome analysis.***PLoS One* 2016, **11:**e0148028.
32. Stulberg E, Fravel D, Proctor LM, Murray DM, LoTempio J, Chrisey L, Garland J, Goodwin K, Graber J, Harris MC: **An assessment of US microbiome research.***Nature microbiology* 2016, **1:**15015.

Figures

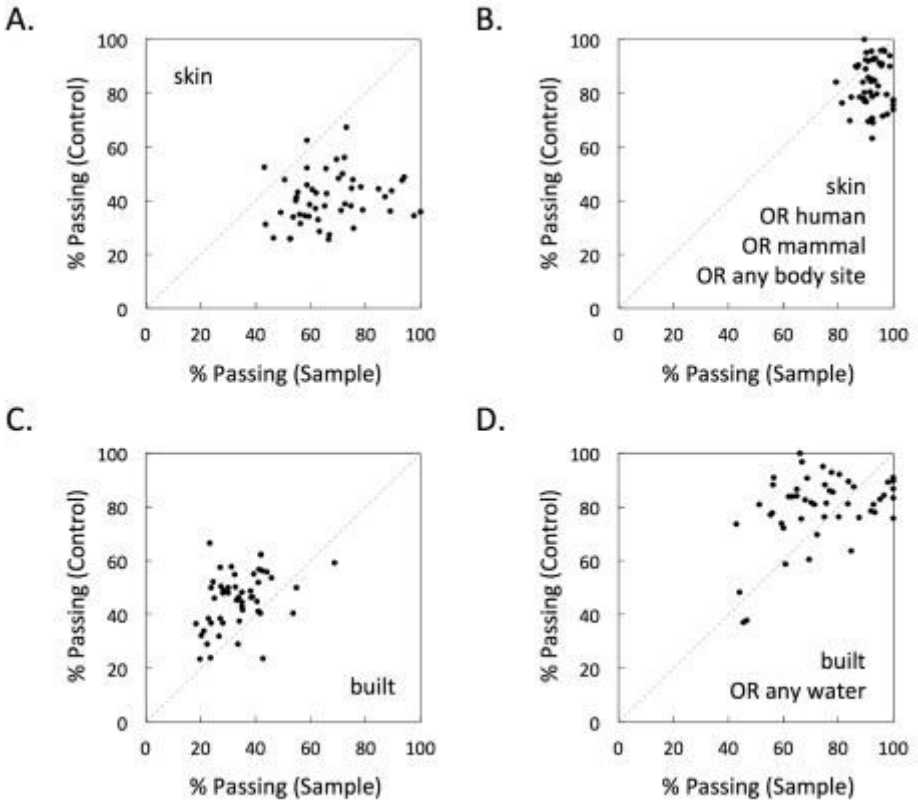

Figure 1

Average percentage of taxa passing the pre-defined (A) skin, (B) skin OR human OR other mammal OR eye OR oral OR ear/nose/throat OR female reproductive systems OR male reproductive systems OR general reproduction OR lymph circulatory system OR nervous system OR bone/muscle OR liver/urinary tract/pancreas OR gut/digestive system (C) built and (D) built OR fresh water OR salt water OR general water filters for each of our 50 sets of samples (x-axis) and paired controls (y-axis). Samples consisted of 5-11 forensic swabs rubbed against the skin at 2 cm intervals across the entire width of each person's forehead. Paired controls consisted of forensic swabs exposed to room air before and after sampling from each person. Samples were analyzed using 16S rRNA sequencing of the V3-V4 region followed by QIIME based on 97% sequence similarity. Lists of taxa were generated by including any genus that was present at >0.1% of reads in any sample/control. A greater fraction of sample (control) taxa pass the filter for any point that lies below (above) the dashed grey line in each panel.

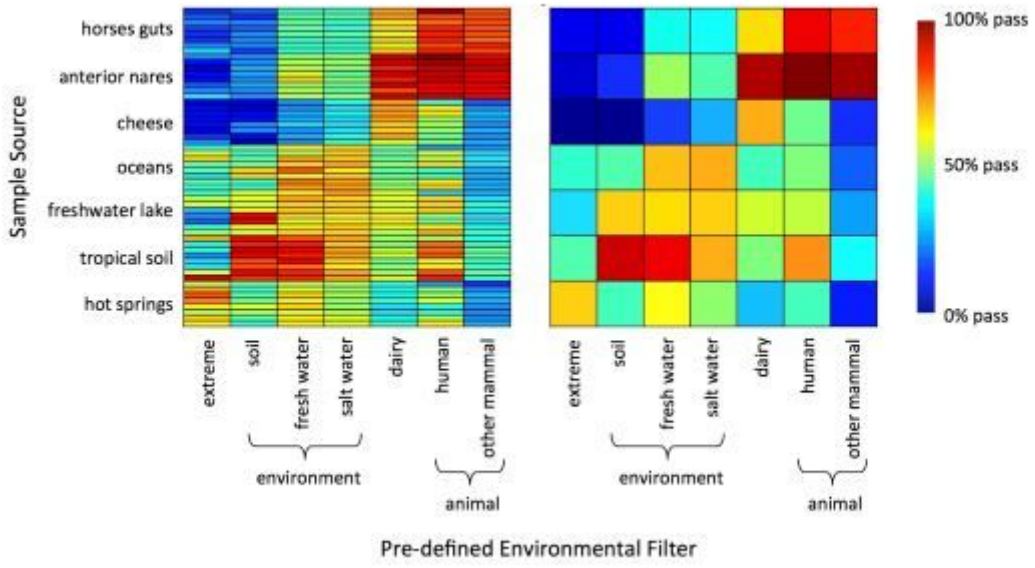

Figure 2

Confusion matrices showing the percentage of taxa passing each pre-defined environmental filter (x-axis) against each sample (y-axis) considering samples individually (left panel), or averaged over studies/sources (right panel). Red indicates that the majority of taxa passed the filter; blue indicates that the majority of taxa did not pass the filter. Filters are ordered such that those coming from the same broad class (e.g., environment, animal) are next to each other. Strong EFILTER performance is indicated by warm colors along the diagonal and cool colors in off-diagonal regions.

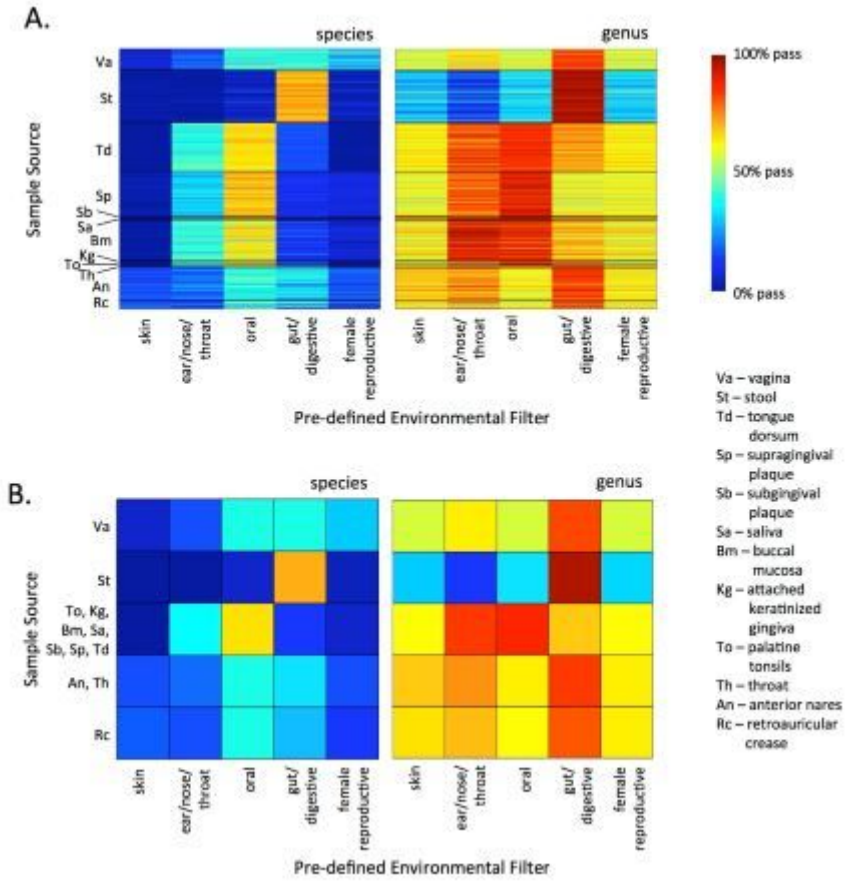

Figure 3

Confusion matrices showing the percentage of species (left panels) and genera (right panels) passing each pre-defined environmental filter (x-axis) against each sample (y-axis) considering samples individually (A), or averaged over body regions (B). Red indicates that the majority of taxa passed the filter; blue indicates that the majority of taxa did not pass the filter. For this dataset, FASTA files were downloaded directly from the HMP ftp server on 07/10/2016 (ftp://public-ftp.ihmpdcc.org/HMGI/). At the time of download, 690 samples were available from 15 body sites. Because of the small number of samples from mid-vagina and vaginal introitus, we did not consider vaginal sites separately. Likewise, we pooled left and right retroauricular creases, leaving 12 distinct body sites. FASTA files were analyzed for taxonomy using the default settings in MetaPhlAn2.

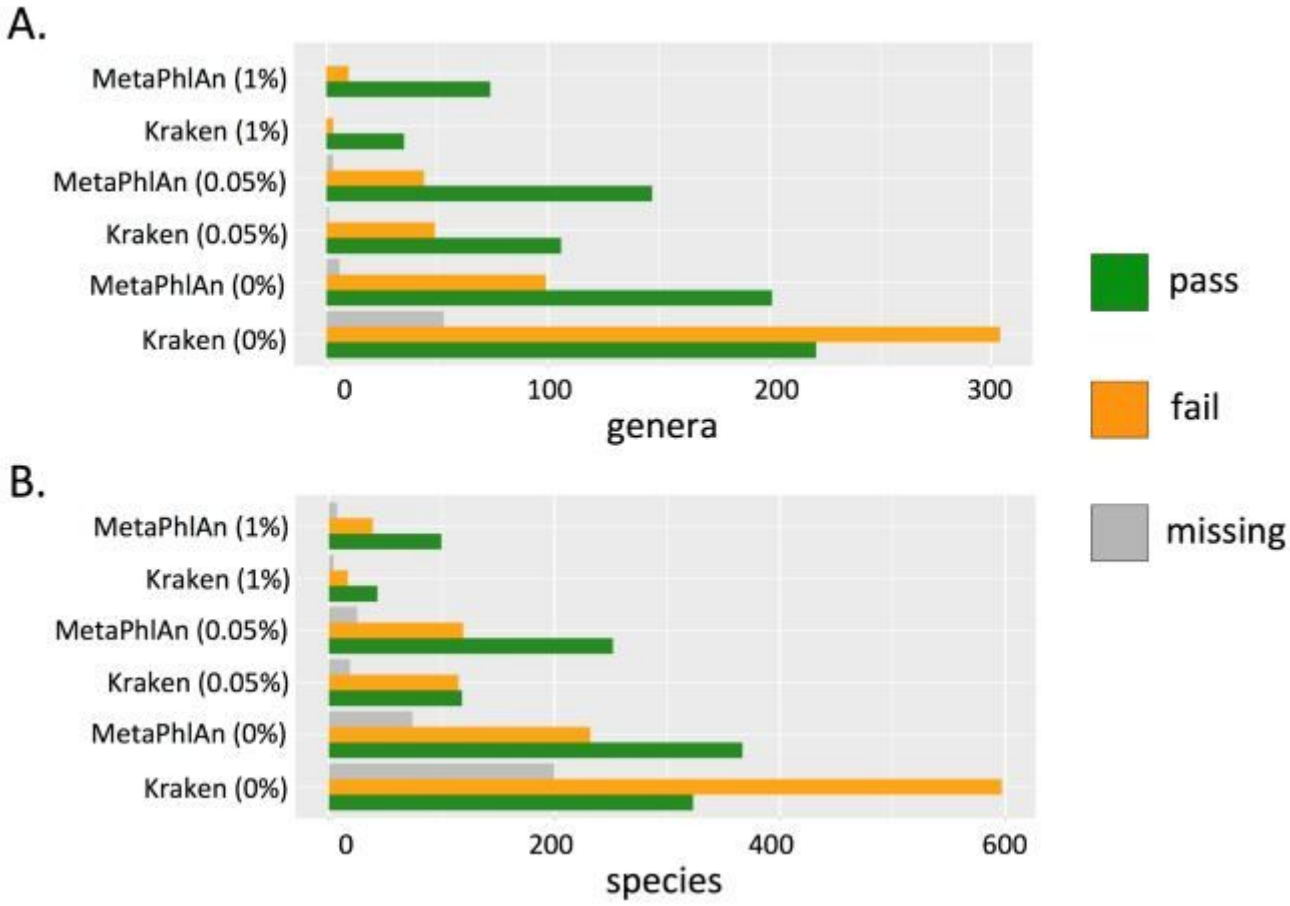

Figure 4

Comparison of Kraken and MetaPhlAn taxon lists generated from the FASTA files in [28] for both genera (A) and species (B). For this analysis, we only considered taxa that were present in at least one sample at levels above the defined threshold percentage of reads (in brackets). Full taxon lists were generated by pooling the taxa in all samples. For Kraken, a reference database was constructed using the complete genomes in RefSeq for the bacterial (2,199 taxonomic IDs), archaeal (165 taxonomic IDs), and viral (4,011 taxonomic IDs) domains, as well as eight representative fungal taxonomic IDs, the Plasmodium falciparum 3D7 genome, the human genome, and the UniVec Core database (ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec). Low complexity regions of the microbial reference sequences were masked using the dustmasker program with a DUST level of 20 [http://www.ncbi.nlm.nih.gov/pubmed/16796549]. After masking, every 31-mer nucleotide sequence present in the collection of reference FASTA sequences is stored at the taxonomic ID of the lowest common ancestor among the leaf nodes that share that 31-mer. MetaPhlAn was run using the default settings in metaphlan2. Results are shown for the EFILTER pre-defined logical combination human OR other mammal OR eye OR oral OR ear/nose/throat OR female reproductive systems OR male reproductive systems OR general reproduction OR lymph circulatory system OR nervous system OR bone/muscle OR liver/urinary tract/pancreas OR gut/digestive system.

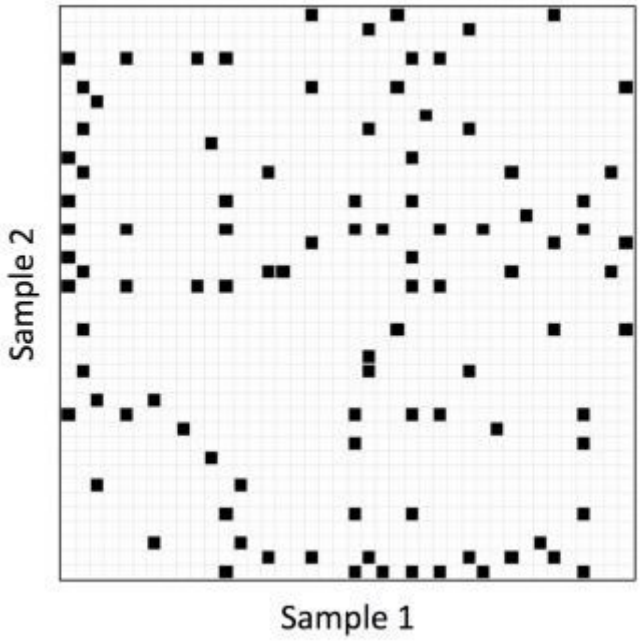

Figure 5

Pairwise combinations of samples from the tara oceans dataset, where black indicates that the sample with the highest diversity depends on whether taxa failing the salt water/fresh water/general water filter are included, and white indicates that it does not. Data were downloaded as lists of genera directly from the MG-RAST website.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFile1.docx
- BewicketalCL.docx