OGUs enable effective, phylogeny-aware analysis of even shallow metagenome community structures

Qiyun Zhu [a,b,#], Shi Huang [b,c], Antonio Gonzalez [b], Imran McGrath [b,d], Daniel McDonald [b], Niina Haiminen [e], George Armstrong [b,c,f], Yoshiki Vázquez-Baeza [c], Julian Yu [a], Justin Kuczynski [g], Gregory D. Sepich-Poore [h], Austin D. Swafford [c], Promi Das [b,i], Justin P. Shaffer [b], Franck Lejzerowicz [b,c], Pedro Belda-Ferre [b], Aki S. Havulinna [j,k], Guillaume Méric [l,m], Teemu Niiranen [j,n,o], Leo Lahti [p], Veikko Salomaa [j], Ho-Cheol Kim [q], Mohit Jain [r,s], Michael Inouye [l,t], Jack A. Gilbert [b,c,i], Rob Knight [b,h,u,#]

[a] School of Life Sciences, Arizona State University, Tempe, Arizona, USA

[b] Department of Pediatrics, School of Medicine, University of California, San Diego, California, USA

[c] Center for Microbiome Innovation, Jacobs School of Engineering, University of California San Diego, La Jolla, California, USA

[d] Division of Biological Sciences, University of California San Diego, La Jolla, California, USA

[e] IBM T. J. Watson Research Center, Yorktown Heights, New York, USA

[f] Bioinformatics and Systems Biology Program, University of California, San Diego, California, USA

[g] Google, Mountain View, CA, USA

[h] Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA

[i] Scripps Institution of Oceanography, University of California San Diego, La Jolla, California, USA

[j] Department of Public Health and Welfare, Finnish Institute for Health and Welfare, Helsinki, Finland

[k] Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki, Finland

[l] Cambridge Baker Systems Genomics Initiative, Baker Heart and Diabetes Institute, Melbourne, Victoria, Australia

[m] Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, Victoria, Australia

[n] Department of Internal Medicine, University of Turku, Turku, Finland

[o] Division of Medicine, Turku University Hospital, Finland

[p] Department of Computing, University of Turku, Turku, Finland

[q] IBM Almaden Research Center, San Jose, California, USA

[r] Department of Medicine, University of California, San Diego, California, USA

[s] Department of Pharmacology, University of California, San Diego, California, USA

[t] Department of Public Health and Primary Care, Cambridge University, Cambridge, UK

[u] Department of Computer Science and Engineering, University of California, San Diego, California, USA

Qiyun Zhu and Shi Huang contributed equally to this work. Author order was determined on the basis of project seniority.

Correspondence: Qiyun Zhu (qiyun.zhu@asu.edu), Rob Knight (robknight@eng.ucsd.edu)

Running title: OGUs for diversity analysis of metagenomic data.

Word count for the abstract: 227.

Word count for the text: 3,307.

Abstract

We introduce Operational Genomic Unit (OGU), a metagenome analysis strategy that directly exploits sequence alignment hits to individual reference genomes as the minimum unit for assessing the diversity of microbial communities and their relevance to environmental factors. This approach is independent from taxonomic classification, granting the possibility of maximal resolution of community composition, and organizes features into an accurate hierarchy using a phylogenomic tree. The outputs are suitable for contemporary analytical protocols for community ecology, differential abundance and supervised learning while supporting phylogenetic methods, such as UniFrac and phylofactorization, that are seldomly applied to shotgun metagenomics despite being prevalent in 16S rRNA gene amplicon studies. As demonstrated in one synthetic and two real-world case studies, the OGU method produces biologically meaningful patterns from microbiome datasets. Such patterns further remain detectable at very low metagenomic sequencing depths. Compared with taxonomic unit-based analyses implemented in currently adopted metagenomics tools, and the analysis of 16S rRNA gene amplicon sequence variants, this method shows superiority in informing biologically relevant insights, including stronger correlation with body environment and host sex on the Human Microbiome Project dataset, and more accurate prediction of human age by the gut microbiomes in the Finnish population. We provide Woltka, a bioinformatics tool to implement this method, with full integration with the QIIME 2 package and the Qiita web platform, to facilitate OGU adoption in future metagenomics studies.

Importance

Shotgun metagenomics is a powerful, yet computationally challenging, technique compared to 16S rRNA gene amplicon sequencing for decoding the composition and structure of microbial communities. However, current analyses of metagenomic data are primarily based on taxonomic classification, which is limited in feature resolution compared to 16S rRNA amplicon sequence variant analysis. To solve these challenges, we introduce Operational Genomic Units (OGUs), which are the individual reference genomes derived from sequence alignment results, without further assigning them taxonomy. The OGU method advances current read-based metagenomics in two dimensions: (i) providing maximal resolution of community composition while (ii) permitting use of phylogeny-aware tools. Our analysis of real-world datasets shows several advantages over currently adopted metagenomic analysis methods and the finest-grained 16S rRNA analysis methods in predicting biological traits. We thus propose the adoption of OGU as standard practice in metagenomic studies.

75 **Keywords**: Operational Genomic Unit, taxonomy independent, reference phylogeny, UniFrac,

76 supervised learning, metagenomics

77

Introduction

The rapidly developing field of shotgun metagenomics has inherited many analytical tools from the more mature field of 16S rRNA gene amplicon studies. For example, diversity analyses provided in platforms such as QIIME 2 (1) can be used for metagenomic analyses. To date, the typical metagenomics workflow starts with taxonomic profiling, which estimates the taxonomic composition of microbial communities by matching sequencing data against a reference database (2). The resulting matches are compiled into an unstructured feature table, with values usually in the form of relative abundances of taxonomic units at a fixed rank (e.g. genus or species level), followed by relevant statistical analyses.

In contrast, the current standard for 16S rRNA analysis involves more advanced feature extraction, including construction of amplicon sequence variants (ASVs), which have replaced operational taxonomic units (OTUs) to deliver the finest-possible resolution from amplicon data (3). Phylogeny-aware algorithms such as UniFrac (4) have been widely-adopted to model community diversity while considering how features interrelate owing to the accessibility of reference phylogenies (5, 6), and the availability of *de novo* and *a priori* phylogenetic inference methods (7). This wisdom should be adopted as well to metagenomics. Thanks to the advances in efficient sequence alignment algorithms, and the expansions of reference genome databases (8, 9) and phylogenomic trees (10, 11), it is now possible and increasingly preferable to develop a fine-resolution, structured data analysis strategy in shotgun metagenomics.

Therefore, we propose an alternative method for constructing metagenomic feature tables, in which features are no longer taxonomic units, but individual reference genomes from a database, and the feature counts are the number of sequences aligned to these genomes. We refer to such features as Operational Genomic Units (OGUs). This term, in an echo of OTU but replacing “taxonomic” with

"genomic", highlights the nature of the genome-based, taxonomy-free analysis. Meanwhile, "operational" indicates that this method does not rely on the direct observation of member genomes of the community, but uses pre-defined reference genomes as a proxy to model the community composition. However, like ASVs, OGUs are exact and do not rely on similarity thresholds as OTUs do.

An OGU table represents the finest-grained resolution of observed genomes in a microbial community relative to the reference database. As such it can be used to quantify the community structure and relationships in correlation with biological traits. It can also work well with cost-efficient "shallow" shotgun metagenomics (12), where limited sequencing depth (even below the previously recommended lower threshold of 500,000 sequences per sample) is adequate for assessing community structure. It further empowers tree-based analyses, such as UniFrac and phylofactorization, which is enhanced by using the "Web of Life" (WoL) reference phylogenetic tree that we recently developed to describe accurate evolutionary relationships among genomes (10).

We have implemented the method for generating OGU tables in the open-source bioinformatics tool, Woltka (https://github.com/qiyunzhu/woltka). This program serves as a versatile interface connecting choices of upstream sequence aligners (such as Bowtie2 and BLAST) and downstream microbiome analysis pipelines (such as QIIME 2). In addition to the standalone program, the package ships with a QIIME 2 (1) plugin to facilitate adoption and integration into existing protocols. We have also made this method available through the Qiita web analysis platform (13) as part of the standard operating procedure for shotgun metagenomic data analysis, thereby enabling massive reprocessing and subsequent meta-analysis of metagenome datasets with OGUs. Thus far, we have applied the OGU method to re-analyze all public and private metagenomic datasets hosted on Qiita, totaling 143 studies and 57,063 samples, as of Mar 3, 2021.

Our team and collaborators have applied prototypes of the OGU method in multiple microbiome and multiomics studies and have obtained biologically relevant results (e.g., (14–16)). In this article, we systematically introduce the principles and practices of the OGU method, demonstrate its efficacy in one synthetic and two real-world microbiome datasets, and compare it with state-of-the-art metagenome analysis approaches and the alternative data type (16S rRNA gene amplicons). Given our findings, we propose the adoption of OGUs as a good practice in metagenomic analyses.

Results

OGUs maximize resolution of community structures

The rationale and benefits of the OGU method are demonstrated with a synthetic case study illustrated in Fig. 1, with the underlying feature tables provided in Table S1. In this simple case, three metagenomes with 12 sequences each were aligned to 10 reference genomes, which were hierarchically organized by taxonomy (left) or by phylogeny (right) (Fig. 1A). Beta diversity was calculated on feature tables at different levels: either on taxonomic units at the rank of genus or species, or directly on reference genomes (i.e., OGUs) without the need for giving them taxonomic labels.

As demonstrated (Fig. 1B), the genus-level analysis, which had the lowest resolution (three genera), yielded spurious proximity between samples B and C, as relative to sample A, largely determined by the differential abundance of genera G1 and G2. The species-level analysis with moderately higher resolution (five species) was able to bring A closer to B and C, mainly contributed by the identical frequencies of species S1, which could not be revealed at the genus level. The OGU-level analysis, having the highest resolution (10 features), revealed the separation between B and C due to distinct OGU composition, despite similar species counts (e.g., O5 and O7 have different counts within S3), and the proximity between A and B due to shared OGUs (O6 and O9). Additional structure was revealed by

using the UniFrac metric, which considers the hierarchical relationships among features, hence further joining samples (here A and B) sharing longer branches in the phylogenetic tree (even by different OGUs, such as O1 and O2) and separating those sharing shorter ones. Taxonomy may serve as a replacement of phylogeny, but it has a lower resolution than phylogeny (e.g., O1 and O2 are evolutionarily closer to each other relative to O3 but taxonomy cannot reveal this), and sometimes does not reflect the true evolutionary relationships among organisms (e.g., O4 and O5 are here placed in different genera), which can impact the accurate modeling of community structures.

In summary, this example illustrates the need for increasing resolution in order to better understand the diversity of microbial communities. This "resolution" has two dimensions of meaning: first, the quantity of features representing individual microbiomes; second, the granularity and accuracy of the hierarchy—if any—that defines the relationships among individual features.

OGUs accurately represent body environment and host sex associated microbiome patterns

We demonstrated the typical use of the OGU method on the classic Human Microbiome Project (HMP) shotgun metagenomic dataset (17), which contains 210 metagenomes sampled from seven body sites of male and female human subjects. We subsampled each metagenome to one million paired-end reads—a sampling depth close to the recommended lower threshold (500k reads) for "shallow" shotgun sequencing (12). The sequences were aligned to the WoL reference genome database (totaling 10,575 bacterial and archaeal genomes) and the alignments were processed using Woltka, resulting in an OGU table with 6,220 features (reference genomes) (Fig. S1A). Beta diversity analysis using the weighted UniFrac metric with the WoL reference phylogeny was performed on the OGU table (Fig. 2). For comparison, we analyzed the dataset using the currently adopted method (CAM) (e.g., (17)): using Bray-Curtis on a species-level taxonomic profile. We exemplified the CAM by using the profile inferred by

Bracken (18) on the same WoL database (Fig. 2), but also tested and reported the results of SHOGUN (19), Centrifuge (20), and MetaPhlAn (21) (Fig. S1).

Principal Coordinates Analysis (PCoA) of OGUs (Figs. 2A and S2A), with the first three axes explaining 71.01% of community structure variance (Figs. 2C and S1B), revealed that microbiomes were clustered mainly by the body site from which they were sampled, which overshadowed clustering by host sex, if any. This pattern is largely consistent with the previous report (17). The PCoA plot by CAM (Figs. 2B and S2B, also see S3), although with less explained variance (46.30%) (Figs. 2C and S1B), also displayed a clustering-by-site pattern. However, it is notable from the plot that sample clusters are aligned diagonally—a typical pattern indicating the saturation of distances caused by the inadequacy of shared features (species) among body sites (22) (Figs. 2B and S2B). This characteristic limits the power of resolving community diversity.

Permutational multivariate analysis of variance (PERMANOVA) of the beta diversity distance matrices suggested that all methods were able to clearly differentiate samples by body site (p=0.001), with OGU generating the strongest statistic (Figs. 2E and S1C) (OGU: F=77.82; CAM: F=42.36). The distinction by host sex was less obvious. Only OGU was able to distinguish microbiome by sex (F=3.011, p=0.013), whereas CAM failed to distinguish sex with statistical significance (F=1.692, p=0.086) (Figs. 2F and S1E-F). This demonstrated the power of the OGU method in capturing subtle but relevant trends, even when another primary factor (body site) is driving most of the community diversity. Three of the seven body sites are located in the oral environment: tongue, teeth and buccal mucosa (Fig. 1A, B). They together indicate weaker differentiation by sex (OGU: F=1.905, p=0.099; CAM: F=1.610, p=0.130) (Figs. 2F and S1G-H). In parallel, we reason that sites sharing the same environment likely have higher microbial connections. To test this effect, we calculated the relative distance between the three oral sites versus oral sites to non-oral sites. This distance is significantly smaller with OGU (0.699

± 0.098, mean and std. dev., same below) than with CAM (0.808 ± 0.051) (two-tailed paired *t*=-14.398, *p*=2.57e-26) (Figs. 2D and S1D), suggesting that OGU is more effective at relating subgroups of samples with shared properties.

The OGU table plus the WoL tree further enabled differential abundance analysis using the phylogenetic factorization method (23) (Figs. S4-5). The result was visualized and analyzed using the recently released massive tree visualizer EMPress (24) (Fig. 2G). It revealed that the phylogenetic clade separated by Factor 1 represents the genus *Lactobacillus*, contained in predominantly posterior fornix samples from female hosts, which is expected (25). Meanwhile, Factor 2 (genus *Neisseria*), Factor 3 (genus *Capnocytophaga*) and Factor 4 (species *Leptotrichia buccalis*) are more frequently observed in the oral sites of male hosts. For comparison, we applied the tree-free method ANCOM (26) on the taxonomic profiles generated by alternative methods (Table S2). At genus level, all four methods were able to capture only *Lactobacillus*, consistent with our Factor 1. However, at species and OGU levels, results were discordant between methods and no method reported any *Lactobacillus* sp., again showing the limitations of confining analyses to taxonomic ranks without phylogenetic information.

Finally, we assessed the efficacy of OGUs along a gradient of decreasing sampling depths. The correlation between the original OGU table (from one million paired-end reads) and each of the subsampled OGU tables was consistently high. A Pearson's *r* of 0.961 ± 0.0726 (mean and std. dev., same below) was retained even at the sampling depth of 200 (Fig. S6A). The PCoA clustering pattern largely remained the same at all sampling depths (Fig. S7). The oral-vs-other relative distance (see above) retained a Pearson's *r* of 0.971 ± 0.00613 when sampling depth was 200 (Fig. S6B). The PERMANOVA *F*-statistics calculated based on 10 replicates of random subsampling were close to the original statistic and largely stable down to very low sampling depths. The mean difference from the original statistic was still within 5% at the sampling depth of 1,000 for body site (3.349 ± 1.361, unit:

213 percentage of the original statistic, same below), or 500 for host sex (2.680 ± 5.473) (Fig. S6C-D).
214 These findings suggest that the OGU method remains valid even on very shallow metagenomic samples,
215 including those that would otherwise be considered unusable for typical metagenomic analyses.

216 **OGUs improve prediction of host age from the gut microbiome**

217 We next analyzed 6,430 stool samples collected through a random sampling of the Finnish population
218 using both 16S rRNA gene amplicon sequencing and shallow shotgun metagenomic sequencing. This
219 "FINRISK" study (27) provides an opportunity to explore the dependency of feature sets (e.g.
220 taxonomic levels and data source: 16S rRNA amplicon vs. shotgun metagenomic data) on the prediction
221 accuracy of a machine learning model on the targeted phenotype (e.g., age). We quantitatively examined
222 the impact of taxonomic level of microbiome features on the empirical error (mean absolute error, or
223 MAE) in predicting human chronological age using a Random Forests regressor (28), constructed using
224 5-fold cross-validation.

225 Our results (Fig. 3A) showed the prediction accuracy continued to improve, resulting in lower absolute
226 errors with finer microbial feature classification levels. Shotgun data outperformed 16S data at all levels,
227 and was able to reduce MAE to less than 10 years at the genus level or below. At the lower limit of both
228 16S and shotgun data, we achieved an MAE of 9.581 ± 0.116 years (mean and std. dev., same below)
229 with OGUs (Fig. 3B), whereas ASVs, the highest possible resolution allowed by 16S data, resulted in a
230 higher MAE of 10.110 ± 0.103 years (two-tailed t=-7.25, p=8.81e-5). Meanwhile, using the species-
231 level profile inferred by Bracken, we also obtained a higher MAE of 10.273 ± 0.089 years (vs. OGU:
232 two-tailed t=-10.59, p=5.53e-6) (Fig. S8). Decreasing sequencing depth did not reduce the age
233 prediction accuracy for individual samples (Fig. S9). For example, samples with 320-366k metagenomic
234 sequences (2nd bin from low end in the figure) had an MAE of 9.290 ± 6.378 years, whereas samples
235 with 1,386-1,931k sequences (2nd bin from high end) had an MAE of 10.118 ± 6.086 years, which were

not significantly different (two-tailed t=-1.37, p=0.170). We then explored which OGUs contributed to the superior performance in age prediction as compared to 16S rRNA ASVs. Therefore, we identified a reduced set (n=128) of the most important OGUs that can maximize the prediction accuracy via a recursive feature elimination approach (Fig. S10). Among these important features, a few gut microbial strains increased in abundance with aging, such as multiple strains from *Streptococcus mutans*, *Eubacterium sp.* (Figs. 3C, S11-12). Remarkably, those *Streptococcus* spp. are typically located in the oral cavity yet can be over-represented in the gut of elderly individuals, suggesting potential microbial transmissions between oral and gut microbiomes related to typical aging in a large population (29, 30). Next, we also identified a few microbial OGUs that were under-represented in the elderly, such as *Anaerostipes hadrus* DSM 3319 and members of *Bifidobacterium*, including *B. longum* NCC2705 and *B. saguini* DSM 23967 Bifsag. Many of these important taxonomic features were not identified in the 16S data, putatively because the partial sequences of a 16S rRNA gene cannot provide sufficient resolution to distinguish species or strains. For example, a few 16S rRNA ASVs annotated with *Lachnospiraceae* have been associated with aging and were identified in either this or past studies (31), whereas our method identified several OGUs (*Anaerostipes hadrus* DSM 3319) within the family of *Lachnospiraceae* that exhibited strong predictive powers for discriminating aging.

Discussion

The OGU method introduced in this article provides a way to maximize the resolution of feature tables by directly considering reference genomes without the reliance on taxonomic classification in shotgun metagenomics studies. Although the strategy of taxonomy-free community structure analysis has been widely adopted in 16S data analysis (e.g., ASV or *de novo* OTU clustering), it remains underexplored in metagenomics, largely due to the difficulties in defining and quantifying “features” without using an *a*

priori classification system. Our study shows that sequence alignment hits to individual reference genomes can be used as the minimum unit for features, referred to as OGUs.

Through comparative analysis of OGU and alternative methods using a synthetic case study and two real-world microbiome studies, we demonstrated that classical high-dimensional statistics and machine learning methods developed and matured in the field of 16S rRNA gene amplicon analysis can be directly applied to OGUs to provide biologically relevant insights. The OGU results often are superior to currently adopted metagenomic classification methods and ASV analysis of the 16S rRNA data. Meanwhile, we showed that the use of taxonomic units as features, as many researchers have been practicing to date, has conceptual and performance limitations compared with the OGU method, particularly at higher taxonomic ranks due to the loss of resolution.

The independence from taxonomy further enables the utilization of explicit phylogenetic trees. A researcher can choose from pre-computed reference phylogenies, such as the one we introduced in the "Web of Life" (WoL) project (10), or custom phylogenomic trees computed from *de novo* construction or placement, through tools such as PhyloPhlAn3 (32) and DEPP (33), which are scalable to large numbers of genomes. This connects evolutionary biologists' efforts in updating the tree of life (e.g., (10, 11, 34)), computational biologists' efforts in forging phylogeny-aware methods (e.g., UniFrac and PhyloFactor), and microbiome scientists' pursuits of relating high-dimensional microbiome data with biology.

Taxonomy, despite being relatively coarse-grained and error-prone as a classification system, may serve as an implicit replacement of phylogeny if the latter is not available. We tested this idea by applying UniFrac to an artificial taxonomic tree with constant branch lengths between ranks (analogous to (35)). Although this treatment is controversial, because taxonomic ranks do not directly indicate evolutionary distances, we did observe improvement compared to not using a tree (Fig. S13). Although there have

been remarkable efforts for curating taxonomy using phylogenetics, however, the number of taxonomic ranks is limited (typically 7 to 8), and can constrain the topology for an ever-growing number of sequenced genomes. For example, the current release (R95) of GTDB (36) has 31,910 species clusters, constituting a taxonomy tree of 45,502 vertices, whereas NCBI RefSeq and GenBank host 977,729 unique genomes as of March 30, 2021, and a fully resolved phylogenetic tree of them can theoretically have 1,955,456 vertices. The history of 16S rRNA studies (7) is repeating itself in whole-genome studies, such that building a phylogeny is not only advantageous but often more feasible than defining taxonomy, and the OGU method powerfully provides an analogous extension to shotgun sequencing studies. As a new notion to microbiome research, OGU's properties in statistical analyses has yet to be characterized in a large number of studies, as was done for 16S rRNA ASVs. Unique challenges in shotgun metagenomics may impact analyses that were designed for 16S rRNA data. For example, very-low-abundance false positive assignments, which are prevalent from typical metagenomic classifiers, may impair the accuracy of the recovered community composition (37). A typical treatment is to only consider features with relative abundance above a given threshold in each sample (37). While we provide this function in Woltka to facilitate user's preferences, our tests suggested that the result of an OGU analysis is highly stable against a wide range of filtering thresholds when using abundance-based metrics (weighted UniFrac and Bray-Curtis), as compared with presence/absence-based metrics (unweighted UniFrac and Jaccard) (Fig. S14). This observation implies the OGU method is robust to noise commonly introduced into metagenomic datasets from many low abundance observations.

The robustness of an OGU analysis is only limited by the comprehensiveness of the reference. Despite that available genomic data have grown to an enormous volume, the size of a reference genome database that can be realistically used in a metagenomic analysis with typical computing facilities is circumscribed, limiting the increase of resolution beyond sub-species levels. Balancing alignment accuracy and database content is therefore an important consideration in designing the analytical

strategy. The algorithm we previously designed and used in the WoL database to maximize the covered biodiversity given a fixed number of genomes (10) may be beneficial in this situation, but its efficacy needs to be further tested in the background of various biospecimens and biological questions. Leaderboard sequencing may also be a useful strategy for iteratively augmenting the reference database with the common genomes in each sample (38). In the long run, efforts to improve algorithms, increase database coverage, and improve computing efficiency are all needed to facilitate effective advances in the field of metagenomics, and the OGU method provides an important step forward in that direction.

Materials and Methods

Protocol details

The OGU method is flexible to the type of sequence alignment. The recommended protocol, which is also the protocol demonstrated and benchmarked in this article, is as follows: Shotgun metagenomic sequencing data were aligned against the WoL reference genome database using SHOGUN v1.0.8 (19), with Bowtie2 v2.4.1 (39) as the backend. This process is equivalent to a Bowtie2 run with the following parameters:

```
--very-sensitive -k 16 --np 1 --mp "1,1" --rdg "0,1" --rfg "0,1" --score-min "L,0,-0.05"
```

The sequence alignment is treated as a mapping from queries (sequencing data) to subjects (reference genomes). It is possible that one sequence is mapped to multiple genomes (up to 16 using the aforementioned Bowtie2 command). In this scenario, each genome is counted $1 / k$ times (k is the number of genomes to which this sequence is mapped. The frequencies of individual genomes were summed after the entire alignment was processed, and rounded to the nearest even integer. Therefore, the sum of OGU frequencies per sample is nearly (considering rounding) equal to the number of aligned sequences in the dataset. The output feature table has columns as sample IDs, rows as feature IDs (OGUs), and cell values as the frequency of each OGU in each sample. This table is ready to be analyzed using software packages such as QIIME 2 (1).

Implementation

The OGU method is implemented in the bioinformatics tool Woltka (Web of Life Toolkit App), under the BSD-3-Clause open-source license. The program is written in Python 3, following high-quality software engineering standards. Its unit test coverage is 100%. The source code is hosted in the GitHub

repository: https://github.com/qiyunzhu/woltka, together with instructions, tutorials, command-line references, and test datasets. The program has been included in the Python Package Index (PyPI). In addition to the standalone Woltka program, a QIIME 2 (1) plugin is included in the software package.

Woltka automatically recognizes and parses multiplexed or per-sample sequence alignment files, either original or compressed using Gzip, Bzip2 or LZMA algorithms. It supports three alignment file formats: 1) SAM (Sequence Alignment Map) (40), which is supported by multiple short read alignment programs, such as Bowtie2 (39), BWA (41) and Minimap2 (42); 2) the standard BLAST (43) tabular output format ("-outfmt 6"), which is supported by multiple sequence alignment programs, such as BLAST, VSEARCH (44) and DIAMOND (45); 3) A plain mapping of query sequences to subject genomes, which is customizable to adopt other tools and pipelines.

In addition to OGU table generation, Woltka supports summarizing features into higher-level groups. This enables taxonomic classification, for comparison purposes. The output of Woltka's classification function and that of SHOGUN's "assign_taxonomy" function are identical. Woltka supports three formats of classification systems: 1) the Greengenes-style lineage strings (supported by programs such as QIIME 2 (1), MetaPhlAn (21) and GTDB-tk (46)); 2) The NCBI-style taxonomy database (47) (a.k.a. "taxdump", supported by programs such as Kraken 2 (48), Centrifuge (20) and DIAMOND (45)); 3) One or multiple plain mappings of child-to-parent classification units.

Deployment

The Woltka program has been incorporated in the Qiita web analysis platform (https://qiita.ucsd.edu/) (13), as part of the standard operating procedure for analyzing shotgun metagenomic data (qp-woltka, code hosted at: https://github.com/qiita-spots/qp-woltka). It can be directly launched from the graphic user interface. A job array system is used to parallelize analyses on a per-sample base to maximize

processing speed. Each process uses eight cores of an Intel E5-2640 v3 CPU and 90 GB DDR4 memory. Two reference genome databases are available for user choice: 1) The "Web of Life" (WoL) database (10), with 10,575 bacterial and archaeal genomes that were evenly sampled through an algorithm. 2) The reference and representative genomes of microbes defined in NCBI RefSeq release 200 (8). The subsequent community ecology analyses based on the OGU table are also available from Qiita. The WoL reference phylogeny is available for choice for phylogenetic analyses (such as UniFrac (4)).

This system allowed us to re-analyze all metagenomic datasets hosted on Qiita (totaling 143 studies and 57,063 samples, as of Mar 3, 2021) to generate OGU tables as well as tables at multiple taxonomic ranks, which are ready for subsequent meta-analysis by Qiita users. Although runtime varies by sample size, the average wall clock time for analyzing one metagenomic sample (including sequence alignment against WoL using Bowtie2 and feature table generation using Woltka) was 13.8 minutes in this large effort.

The HMP dataset

The Human Microbiome Project (HMP) (17) dataset was downloaded from the official website (https://www.hmpdacc.org/hmp/). It contains 241 samples of 100 bp paired-end whole genome sequencing (WGS) reads. The sequencing data were already processed to remove human contamination and low-quality regions. We dropped samples with less than 1M paired-end reads, leaving 210 samples. They were randomly subsampled to 1M paired-end reads per sample. These samples represent both male (n=138) and female (n=72) human subjects. They represent seven body sites: stool (n=78), tongue dorsum (n=42), supragingival plaque (n=33), buccal mucosa (n=28), retroauricular crease (n=13), posterior fornix (n=10), and anterior nares (n=6).

Taxonomic profiling

In comparison with the OGU method, we performed taxonomic profiling on the shotgun metagenomic data using four existing methods, specified as below. The default parameters were used for all programs. To maximize comparability, we used the WoL reference genome database (10) for all methods, except for MetaPhlAn (because it uses a special marker gene database which is difficult to customize).

1. SHOGUN: SHOGUN v1.0.8 (19), which calls Bowtie2 v2.4.1 to perform sequence alignment.
2. Bracken: Bracken v2.5 (18) on the results of Kraken v2.0.8 (48).
3. Centrifuge: Centrifuge v1.0.3 (20).
4. MetaPhlAn: MetaPhlAn v2.6.0 (21) with its database (mpa_v20_m200). Results (relative abundances) were normalized to counts per million sequences.

Beta diversity analysis

Beta diversity analysis of the HMP dataset was performed using QIIME 2 (1), following recommended protocols (49). Specifically, beta diversity distance matrices were constructed using the "`qiime diversity beta`" command with Jaccard and Bray-Curtis metrics, and using the "`qiime diversity beta-phylogenetic`" command (50) with unweighted UniFrac and weighted UniFrac metrics, based on the WoL reference phylogeny. Principal coordinates analysis (PCoA) was performed using the "`qiime diversity pcoa`" command. The correlation between biological factors (body site and host sex) and beta diversity was assessed using the PERMANOVA test, through the command "`qiime diversity adonis`", with 999 permutations (the default setting).

Site clustering by environment

In the HMP study, we quantified the proximity of the three oral sites (tongue dorsum, supragingival plaque, and buccal mucosa) as compared with the four non-oral sites (stool, retroauricular crease,

399 posterior fornix, and anterior nares) as follows: For each sample in the three oral sites, we calculated the
400 beta diversity distance to all samples in all but the current site. We then separated these distances into
401 oral (i.e., the two oral sites other than the current one) and non-oral (i.e., the four non-oral sites). We
402 calculated the ratio of the mean distance of the former versus the latter. Finally we reported the
403 distribution of the mean ratios of all oral samples.

404 **Phylogenetic factorization**

405 We performed phylogenetic factorization as implemented in Phylofactor v0.0.1 to infer phylogenetic
406 clades ("factors") that are differentially abundant between male and female subjects. Two samples with
407 less than 100,000 OGU counts were excluded from the analysis. OGUs with relative abundance below
408 0.01% were dropped from each sample, and OGUs present in fewer than two samples were also
409 excluded. We built an explained variance-maximizing (the choice parameter was set to "var")
410 Phylofactor model using the OGU table and the WoL phylogeny. We specified the model to return 20
411 factors. They were labeled by the taxonomic annotation of the corresponding phylogenetic clades as
412 provided in the WoL database. The results were visualized with EMPress. In each factor, we tested the
413 differences in male vs female subjects by comparing the ILR-transformed vectors corresponding to each
414 sample group using a two-tailed independent samples *t*-test.

415 **Subsampling of OGU tables**

416 To assess the impact of sampling depth on analysis results, we randomly subsampled the OGU tables to
417 lower depths (sum of OGU frequencies per sample). This process mimicked lower sequencing depths in
418 the original data, because the sum of OGU frequencies is nearly equal to the number of aligned
419 sequences (see above). This process further considered the unaligned part of the sequencing data. For
420 example, if *m* out of *n* sequences in a sample were aligned to at least one reference genome (therefore

421 the sum of OGU frequencies was *m*), we added an extra "unaligned" feature of frequency of *n* - *m* to the
422 OGU table, prior to random subsampling, and removed this feature after sampling.

423 **The FINRISK 2002 datasets**

424 The FINRISK 2002 is a large, well-phenotyped, and representative cohort based on a stratified random
425 sample of the population aged 25 to 74 years from specific geographical areas of Finland (27). All
426 volunteer participants took a self-administered questionnaire, physical measurements and collection of
427 blood and stool samples. The microbiome data and metadata that support the findings of this study are
428 available from the THL Biobank based on a written application and following relevant Finnish
429 legislation. Details of the application process are described in the website of the Biobank:
430 https://thl.fi/en/web/thl-biobank/for-researchers.

431 Paired 16S rRNA gene amplicon sequencing data and shotgun metagenomic sequencing data are
432 available for 6,430 stool samples. The 16S rRNA data were demultiplexed, quality filtered, and denoised
433 with deblur v1.1.0 (51), resulting in an average ASV frequency of 8,787 per sample, followed by
434 normalization to 10,000 per sample. Taxonomic classification was performed using a pre-trained Naive
435 Bayes classifier against the Greengenes 13_8 database at an OTU clustering level of 99%. Feature tables
436 were rarefied to a sampling depth of 10,000. The shotgun metagenomic data were trimmed and quality
437 filtered using Atropos v1.1.25 (52), resulting in an average of 1.07 million paired-end sequences per
438 sample. They were aligned to the WoL database using SHOGUN v1.0.8. An OGU table was generated
439 using the current approach. As a comparison, Bracken v2.5 with Kraken v2.0.8 were used to infer
440 taxonomic profiles using the same WoL database. These analyses were the same as the corresponding
441 analyses of the HMP shotgun metagenomic dataset, as described above.

442 **Supervised regression for age prediction**

443 We performed machine learning analysis of microbial profiles derived from both 16S amplicon
444 sequencing and shotgun metagenomics sequencing, at distinct levels of resolution. These included
445 taxonomic ranks (phylum, class, order, family, genus and species) for both 16S rRNA and shotgun
446 metagenomic data (the latter of which were inferred by either SHOGUN or Bracken), ASV for 16S
447 rRNA data, and OGU for shotgun metagenomic data (inferred by SHOGUN with Woltka). In each
448 profile, features with a study-wide prevalence less than 0.001 were excluded. Random Forest regressors
449 for predicting chronological age were trained based on each profile with tuned hyperparameters with a
450 stratified 5-fold cross-validation approach using R package ranger v0.12.1 (53). Each dataset was split
451 into five groups with similar age distributions, and we trained the classifier on 80% of the data, and
452 made predictions on the remaining 20% of the data in each fold iteration. We next evaluated the
453 performance of age prediction using mean absolute error (MAE), which calculated as $\mathrm{MAE}=\frac{\sum_{i=1}^{n}|y_i-x_i|}{n}$,
454 where y denotes the predicted age, x denotes the chronological age, and n is the total number of samples.
455 Based on the MAE evaluation, we next determined the most predictive taxonomic levels derived from
456 both 16S and shotgun metagenomics.

457 To identify the most important taxonomic features that contributed to the age prediction, we visualized
458 the top-128 ranked important features by built-in Random Forest importance scores and their
459 phylogenetic relationships using EMPress (54). We next performed the feature selection analysis to
460 identify a set of important microbial features that can maximize the model performance. We built age
461 regressors using a series of reduced sets (n = 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, and the number of
462 all features) of the most predictive taxonomic features (namely, OGU) and compared their performance.
463 The rationale is to observe the trough in MAE when additional features are added into the regression
464 model.

Statistics statement

All data analysis was performed using QIIME 2 release 2020.6. PERMANOVA was performed using the "adonis" command (which wraps the "adonis" function in vegan v2.5-6). Paired *t*-test was performed using the "ttest_rel" function in SciPy v1.4.1.

Acknowledgements

We are grateful to Gabriel Al-Ghalith, Zachary Burcham, Jeff DeReus, Marcus Fedarko, Shalisa Hansen, Stefan Janssen, Emily Kobayashi, Evguenia Kopylova, Tomasz Kosciolek, Holly Lutz, Cameron Martino, Siavash Mirarab, James Morton, Oriane Moyne, Wayne Pfeiffer, Daniel Roush and Se Jin Song for valuable testing of the methodology, insightful discussions on this study and additional assistance.

This work is supported in part by an Arizona State University start-up grant (to Q.Z.), Sloan Foundation G-2017-9838, IBM Artificial Intelligence for Healthy Living A1770534, DARPA JUMP/CRISP, NIH P30DK120515, DP1AT010885, U19AG063744, U24CA248454, Emerald Foundation Distinguished Investigator Award, Crohn's and Colitis Foundation 675191, NSF RAPID 2038509, IBM Research AI through the AI Horizons Network and the UC San Diego Center for Microbiome Innovation (to S.H., I.M., Y.V.-B., and R.K.). G.D.S.-P. is supported by a fellowship from the National Institutes of Health (F30 CA243480). T.N. was funded by the Emil Aaltonen Foundation, the Finnish Medical Foundation, the Finnish Foundation for Cardiovascular Disease, and the Academy of Finland (grant 321351). V.S. was supported by the Finnish Foundation for Cardiovascular Research. This work used the Comet supercomputer at the San Diego Supercomputer Center through allocation BIO150043 through the Extreme Science and Engineering Discovery Environment (XSEDE).

Q.Z. and R.K. conceived the project. Q.Z. led the development of the methodology and software. S.H. and Q.Z. led the analysis and interpretation of the datasets presented in this article. S.H., A.G., D.M. and Y.V.-B. contributed to the design of the method. A.G., D.M. and G.A. contributed to the development of the software. G.D.S.-P., A.D.S., P.D., F.L. contributed to the test of the method. P.B.-F., A.S.H., G.M., T.N., L.L., V.S. and M.J. contributed to data curation. A.G., I.M., J.Y., Y.V-B. and J.K. contributed to data analysis. N.H., G.D.S.-P., A.S.H., G.M., T.N., L.L., V.S., H.-C.K., M.J., M.I., J.A.G. and R.K. contributed to result interpretation. R.K. and Q.Z. managed the project. All the authors contributed to the composition and discussion of the manuscript.

We declare that we have no competing interests.

References

1. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS 2nd, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37:852–857.

2. Breitwieser FP, Lu J, Salzberg SL. 2019. A review of methods and databases for metagenomic classification and assembly. Brief Bioinform 20:1125–1136.

3. Callahan BJ, McMurdie PJ, Holmes SP. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J 11:2639–2643.

4. Lozupone C, Knight R. 2005. UniFrac: a new phylogenetic method for comparing microbial

518 communities. Appl Environ Microbiol 71:8228–8235.

519 5. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight
520 R, Hugenholtz P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and
521 evolutionary analyses of bacteria and archaea. ISME J 6:610–618.

522 6. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The
523 SILVA ribosomal RNA gene database project: improved data processing and web-based tools.
524 Nucleic Acids Res 41:D590–6.

525 7. Janssen S, McDonald D, Gonzalez A, Navas-Molina JA, Jiang L, Xu ZZ, Winker K, Kado DM,
526 Orwoll E, Manary M, Mirarab S, Knight R. 2018. Phylogenetic Placement of Exact Amplicon
527 Sequences Improves Associations with Clinical Information. mSystems 3.

528 8. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B,
529 Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin
530 V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W,
531 Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K,
532 Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-
533 Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A,
534 Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq)
535 database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids
536 Res 44:D733–45.

537 9. Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, Pollard KS, Sakharova E,
538 Parks DH, Hugenholtz P, Segata N, Kyrpides NC, Finn RD. 2021. A unified catalog of 204,938
539 reference genomes from the human gut microbiome. Nat Biotechnol 39:105–114.

10. Zhu Q, Mai U, Pfeiffer W, Janssen S, Asnicar F, Sanders JG, Belda-Ferre P, Al-Ghalith GA, Kopylova E, McDonald D, Kosciolek T, Yin JB, Huang S, Salam N, Jiao J-Y, Wu Z, Xu ZZ, Cantrell K, Yang Y, Sayyari E, Rabiee M, Morton JT, Podell S, Knights D, Li W-J, Huttenhower C, Segata N, Smarr L, Mirarab S, Knight R. 2019. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. Nat Commun 10:5477.

11. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol 36:996–1004.

12. Hillmann B, Al-Ghalith GA, Shields-Cutler RR, Zhu Q, Gohl DM, Beckman KB, Knight R, Knights D. 2018. Evaluating the Information Content of Shallow Shotgun Metagenomics. mSystems 3.

13. Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vázquez-Baeza Y, Ackermann G, DeReus J, Janssen S, Swafford AD, Orchanian SB, Sanders JG, Shorenstein J, Holste H, Petrus S, Robbins-Pianka A, Brislawn CJ, Wang M, Rideout JR, Bolyen E, Dillon M, Caporaso JG, Dorrestein PC, Knight R. 2018. Qiita: rapid, web-enabled microbiome meta-analysis. Nat Methods 15:796–798.

14. Poore GD, Kopylova E, Zhu Q, Carpenter C, Fraraccio S, Wandro S, Kosciolek T, Janssen S, Metcalf J, Song SJ, Kanbar J, Miller-Montgomery S, Heaton R, Mckay R, Patel SP, Swafford AD, Knight R. 2020. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature 579:567–574.

15. Gauglitz JM, Morton JT, Tripathi A, Hansen S, Gaffney M, Carpenter C, Weldon KC, Shah R, Parampil A, Fidgett AL, Swafford AD, Knight R, Dorrestein PC. 2020. Metabolome-Informed

562 Microbiome Analysis Refines Metadata Classifications and Reveals Unexpected Medication
563 Transfer in Captive Cheetahs. mSystems 5.

564 16. Ha CWY, Martin A, Sepich-Poore GD, Shi B, Wang Y, Gouin K, Humphrey G, Sanders K,
565 Ratnayake Y, Chan KSL, Hendrick G, Caldera JR, Arias C, Moskowitz JE, Ho Sui SJ, Yang S,
566 Underhill D, Brady MJ, Knott S, Kaihara K, Steinbaugh MJ, Li H, McGovern DPB, Knight R,
567 Fleshner P, Devkota S. 2020. Translocation of Viable Gut Microbiota to Mesenteric Adipose Drives
568 Formation of Creeping Fat in Humans. Cell 183:666–683.e17.

569 17. Turnbaugh PJ, Qin J, D N Fredricks T L Fiedler, Costello EK, Grice EA, Ravel J, Segata N,
570 Gillespie JJ, Sharpton TJ, Sokol H, JA. Aas, BJ. Paster, LN. Stokes, I. Olsen, FE. Dewhirst, Medini
571 D, S K Mazmanian J L Round, Goodman AL, Kuehnert MJ, Caporaso JG, M. Kanehisa, S. Goto,
572 M. Furumichi, M. Tanabe, M. Hirakawa, H. Li RD, Giannoukos G, MG. Langille FSB. 2012.
573 Structure, function and diversity of the healthy human microbiome. Nature 486:207–214.

574 18. Lu J, Breitwieser FP, Thielen P, Salzberg SL. 2017. Bracken: estimating species abundance in
575 metagenomics data. PeerJ Comput Sci 3:e104.

576 19. Hillmann B, Al-Ghalith GA, Shields-Cutler RR, Zhu Q, Knight R, Knights D. 2020. SHOGUN: a
577 modular, accurate and scalable framework for microbiome quantification. Bioinformatics 36:4088–
578 4090.

579 20. Kim D, Song L, Breitwieser FP, Salzberg SL. 2016. Centrifuge: rapid and sensitive classification of
580 metagenomic sequences. Genome Res 26:1721–1729.

581 21. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, Tett A, Huttenhower C,
582 Segata N. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods

12:902–903.

22. Morton JT, Toran L, Edlund A, Metcalf JL, Lauber C, Knight R. 2017. Uncovering the Horseshoe Effect in Microbial Analyses. mSystems 2.

23. Washburne AD, Silverman JD, Leff JW, Bennett DJ, Darcy JL, Mukherjee S, Fierer N, David LA. 2017. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. PeerJ 5:e2969.

24. Cantrell K, Fedarko MW, Rahman G, McDonald D, Yang Y, Zaw T, Gonzalez A, Janssen S, Estaki M, Haiminen N, Beck KL, Zhu Q, Sayyari E, Morton JT, Armstrong G, Tripathi A, Gauglitz JM, Marotz C, Matteson NL, Martino C, Sanders JG, Carrieri AP, Song SJ, Swafford AD, Dorrestein PC, Andersen KG, Parida L, Kim H-C, Vázquez-Baeza Y, Knight R. 2021. EMPress Enables Tree-Guided, Interactive, and Exploratory Analyses of Multi-omic Data Sets. mSystems 6.

25. Ma B, Forney LJ, Ravel J. 2012. Vaginal microbiome: rethinking health and disease. Annu Rev Microbiol 66:371–389.

26. Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD. 2015. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb Ecol Health Dis 26:27663.

27. Borodulin K, Tolonen H, Jousilahti P, Jula A, Juolevi A, Koskinen S, Kuulasmaa K, Laatikainen T, Männistö S, Peltonen M, Perola M, Puska P, Salomaa V, Sundvall J, Virtanen SM, Vartiainen E. 2018. Cohort Profile: The National FINRISK Study. Int J Epidemiol 47:696–696i.

28. Breiman L. 2001. Random Forests. Mach Learn 45:5–32.

29. Zhang X, Zhong H, Li Y, Shi Z, Ren H, Zhang Z, Zhou X, Tang S, Han X, Lin Y, Yang F, Wang

D, Fang C, Fu Z, Wang L, Zhu S, Hou Y, Xu X, Yang H, Wang J, Kristiansen K, Li J, Ji L. 2021. Sex- and age-related trajectories of the adult human gut microbiota shared across populations of different ethnicities. Nature Aging 1:87–100.

30. Schmidt TS, Hayward MR, Coelho LP, Li SS, Costea PI, Voigt AY, Wirbel J, Maistrenko OM, Alves RJ, Bergsten E, de Beaufort C, Sobhani I, Heintz-Buschart A, Sunagawa S, Zeller G, Wilmes P, Bork P. 2019. Extensive transmission of microbes along the gastrointestinal tract. Elife 8.

31. Huang S, Haiminen N, Carrieri A-P, Hu R, Jiang L, Parida L, Russell B, Allaband C, Zarrinpar A, Vázquez-Baeza Y, Belda-Ferre P, Zhou H, Kim H-C, Swafford AD, Knight R, Xu ZZ. 2020. Human Skin, Oral, and Gut Microbiomes Predict Chronological Age. mSystems 5.

32. Asnicar F, Thomas AM, Beghini F, Mengoni C, Manara S, Manghi P, Zhu Q, Bolzan M, Cumbo F, May U, Sanders JG, Zolfo M, Kopylova E, Pasolli E, Knight R, Mirarab S, Huttenhower C, Segata N. 2020. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. Nat Commun 11:2500.

33. Jiang Y, Balaban M, Zhu Q, Mirarab S. 2021. DEPP: Deep Learning Enables Extending Species Trees using Single Genes. Cold Spring Harbor Laboratory.

34. Castelle CJ, Banfield JF. 2018. Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life. Cell 172:1181–1197.

35. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, Gregor I, Majda S, Fiedler J, Dahms E, Bremges A, Fritz A, Garrido-Oter R, Jørgensen TS, Shapiro N, Blood PD, Gurevich A, Bai Y, Turaev D, DeMaere MZ, Chikhi R, Nagarajan N, Quince C, Meyer F, Balvočiūtė M, Hansen LH, Sørensen SJ, Chia BKH, Denis B, Froula JL, Wang Z, Egan R, Don Kang D, Cook JJ, Deltel

C, Beckstette M, Lemaitre C, Peterlongo P, Rizk G, Lavenier D, Wu Y-W, Singer SW, Jain C, Strous M, Klingenberg H, Meinicke P, Barton MD, Lingner T, Lin H-H, Liao Y-C, Silva GGZ, Cuevas DA, Edwards RA, Saha S, Piro VC, Renard BY, Pop M, Klenk H-P, Göker M, Kyrpides NC, Woyke T, Vorholt JA, Schulze-Lefert P, Rubin EM, Darling AE, Rattei T, McHardy AC. 2017. Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. Nat Methods 14:1063–1071.

36. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. 2020. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol 38:1079–1086.

37. Ye SH, Siddle KJ, Park DJ, Sabeti PC. 2019. Benchmarking Metagenomics Tools for Taxonomic Classification. Cell 178:779–794.

38. Sanders JG, Nurk S, Salido RA, Minich J, Xu ZZ, Zhu Q, Martino C, Fedarko M, Arthur TD, Chen F, Boland BS, Humphrey GC, Brennan C, Sanders K, Gaffney J, Jepsen K, Khosroheidari M, Green C, Liyanage M, Dang JW, Phelan VV, Quinn RA, Bankevich A, Chang JT, Rana TM, Conrad DJ, Sandborn WJ, Smarr L, Dorrestein PC, Pevzner PA, Knight R. 2019. Optimizing sequencing protocols for leaderboard metagenomics by combining long and short reads. Genome Biol 20:226.

39. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.

40. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

41. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.

646 Bioinformatics 25:1754–1760.

647 42. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100.

648 43. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool.
649 Journal of Molecular Biology.

650 44. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool
651 for metagenomics. PeerJ 4:e2584.

652 45. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat
653 Methods 12:59–60.

654 46. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2019. GTDB-Tk: a toolkit to classify genomes
655 with the Genome Taxonomy Database. Bioinformatics
656 https://doi.org/10.1093/bioinformatics/btz848.

657 47. Federhen S. 2011. The NCBI Taxonomy database. Nucleic Acids Res 40:D136–D143.

658 48. Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. Genome Biol
659 20:257.

660 49. Estaki M, Jiang L, Bokulich NA, McDonald D, González A, Kosciolek T, Martino C, Zhu Q,
661 Birmingham A, Vázquez-Baeza Y, Dillon MR, Bolyen E, Gregory Caporaso J, Knight R. 2020.
662 QIIME 2 Enables Comprehensive End-to-End Analysis of Diverse Microbiome Data and
663 Comparative Studies with Publicly Available Data. Current Protocols in Bioinformatics.

664 50. McDonald D, Vázquez-Baeza Y, Koslicki D, McClelland J, Reeve N, Xu Z, Gonzalez A, Knight R.
665 2018. Striped UniFrac: enabling microbiome analysis at unprecedented scale. Nat Methods 15:847–

848.

51. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. 2017. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. mSystems 2.

52. Didion JP, Martin M, Collins FS. 2017. Atropos: specific, sensitive, and speedy trimming of sequencing reads. PeerJ 5:e3720.

53. Wright MN, Ziegler A. 2017. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Softw 77:1–17.

54. Cantrell K, Fedarko MW, Rahman G, McDonald D, Yang Y, Zaw T, Gonzalez A, Janssen S, Estaki M, Haiminen N, Beck KL, Zhu Q, Sayyari E, Morton J, Tripathi A, Gauglitz JM, Marotz C, Matteson NL, Martino C, Sanders JG, Carrieri AP, Song SJ, Swafford AD, Dorrestein PC, Andersen KG, Parida L, Kim H-C, Vázquez-Baeza Y, Knight R. 2020. EMPress enables tree-guided, interactive, and exploratory analyses of multi-omic datasets. Cold Spring Harbor Laboratory.

Figure Legends

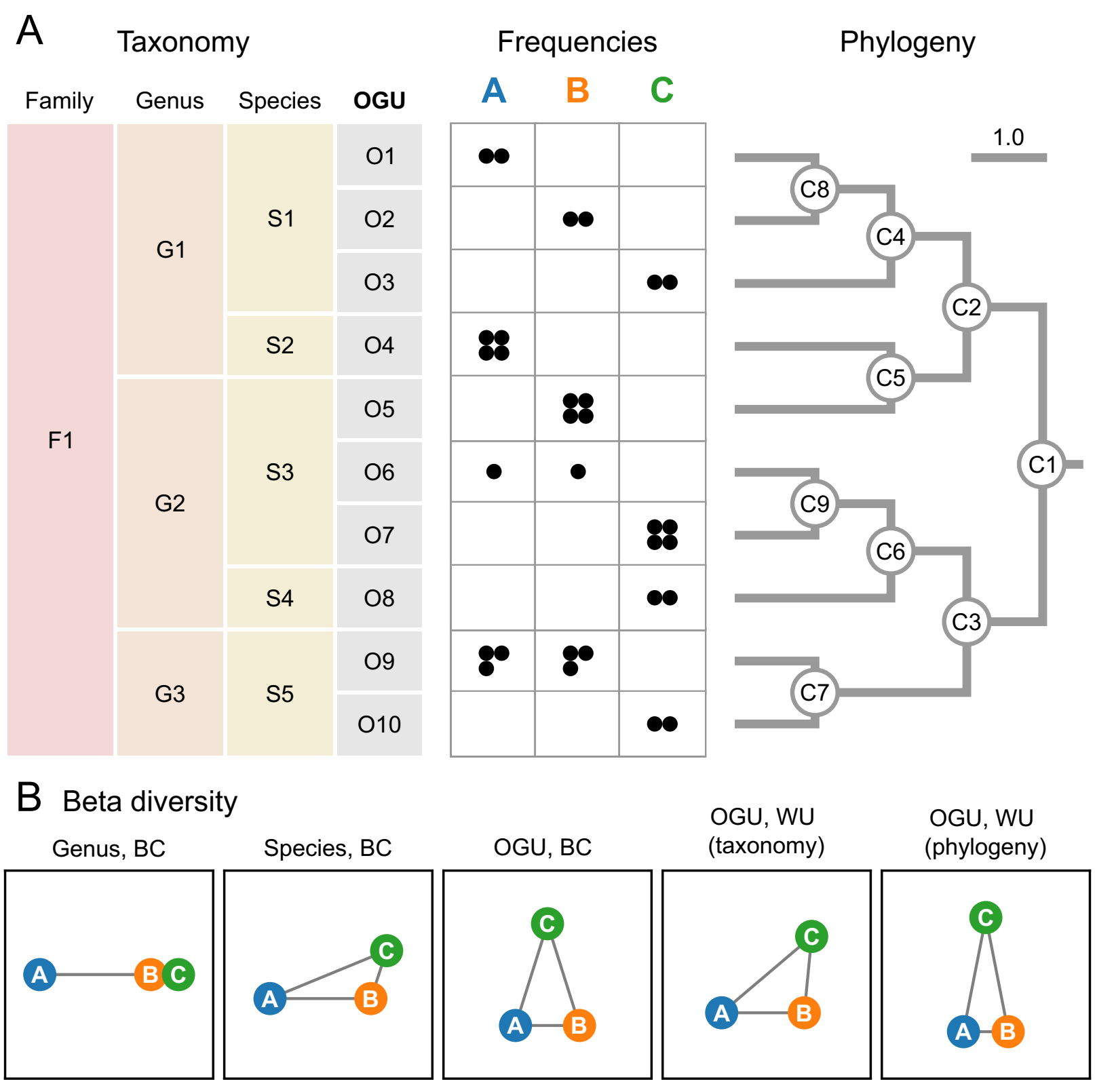
Figure 1. Feature resolution impacts community structure analysis even in small conceptual examples. **A**. A synthetic dataset involving three microbial communities, each of which having 12 unique read hits, as represented by black circles in the frequency table, to a total of 10 reference genomes (OGUs), classified under five species, three genera and one family, as noted to the left. A phylogenetic tree of the 10 genomes is shown on the right. In this simplified case, the phylogeny is not much more complex than the taxonomy (with three more edges); however, the taxonomic assignment and the phylogenetic placement of genome O5 are not consistent. **B**. Beta diversity of the dataset. The three samples (circles) are connected by edges representing the pairwise distances calculated by Bray-Curtis (BC) or weighted UniFrac (WU) on the frequency table. For the latter measure, either the taxonomy or the phylogeny was used to quantify the hierarchical relationships among OGUs, as noted in the parentheses. The edge lengths were normalized so that their sum is equal in each graph. This synthetic case study demonstrates that different resolutions of features and feature structures can lead to very different conclusions regarding sample relationships.

Figure 2. Analysis of the HMP metagenomes reveals clustering by body environment and differentiation by host sex. Beta diversity analysis was performed on 210 samples subsampled to one million paired-end shotgun reads each. **A**. PCoA by the method proposed in this study (OGU): weighted UniFrac metric calculated with the WoL reference phylogeny based on the OGU table. Samples (dots) are colored by body site and shaped by host sex. **B**. PCoA using the current adopted method (CAM): Bray-Curtis calculated on species-level taxonomic units identified by Bracken, which shows a diagonal pattern that aligns all samples of the four non-oral body sites in one plane (also see Figs. S2B and S3). **C**. Proportions of community structure variance explained by the first three axes of PCoA. **D**. Mean

ratio of the beta diversity distances from any oral sample to a sample of the two other oral sites versus to that of non-oral body sites. The lower the mean ratio is, the more similar communities of the three oral sites are to each other in the background of multiple body environments. The bold line in each box represents the median. The whiskers represent 1.5 IQR. **E** and **F**. PERMANOVA pseudo-*F* statistics indicating the differentiation of community structures by body site (**E**) and by host sex (**F**). The larger *F* is, the more distinct the community structures are between groups versus within groups. The *y*-axis is aligned to *F*=1.0 which indicates no difference. For **E**, all statistics have a *p*-value of 0.001. For **F**, an asterisk (*) indicates *p*-value ≤ 0.05. **G**. Differentially abundant phylogenetic clades by host sex inferred using PhyloFactor and visualized using EMPress on the WoL reference phylogeny. The tree was subsetted to only include OGUs detected in the dataset. The top 20 clades by effect size are colored (full details provided in Figs. S4-5). The top five clades are numbered 1 through 5 by decreasing effect size, circled, and labeled with corresponding taxonomic annotations. The small color ring represents phylum-level annotations. The inner and outer barplot rings indicate the OGU counts split by body site (using the same color scheme as in A and B) and by host sex, respectively.

Figure 3. Analysis of the FINRISK metagenomes showing superior prediction accuracy over taxonomic units and 16S rRNA data. **A.** The empirical error (mean absolute error, MAE) in predicting host chronological age using microbiome features at distinct taxonomic ranks in paired 16S rRNA amplicon and shotgun metagenomics data with a Random Forests regressor. “None” represents the taxonomy-free, finest-possible level (ASV for 16S, OGU for shotgun). Small circles indicate MAEs in all iterations of five-fold cross validation. Large circles and error bars indicate the mean and standard deviations of the five MAEs. **B.** Scatter plot of the actual age vs. the predicted age by the best-performing model with OGU features in the five-fold cross-validation. The black line was generated using ggplot2’s local polynomial regression fitting. **C**. Phylogenomic tree of 169 OGUs with importance

score ≥ 0.1 in the prediction model. The tree was subsampled based on the WoL reference phylogeny, and drawn to scale (branch lengths represent mutations per site). Branch colors indicate the mean importance score of all descendants of the clade. Taxonomic labels are displayed where needed. Circles and lines with stops are displayed where needed to assist location of taxonomic labels to target branches or clades.


A
Taxonomy
Frequencies
Phylogeny
Family
Genus
Species
OGU
A
B
C
F1
G1
G2
G3
S1
S2
S3
S4
S5
O1
O2
O3
O4
O5
O6
O7
O8
O9
O10
1.0
C8
C4
C2
C5
C1
C9
C6
C3
C7
B Beta diversity
Genus, BC
Species, BC
OGU, BC
OGU, WU (taxonomy)
OGU, WU (phylogeny)