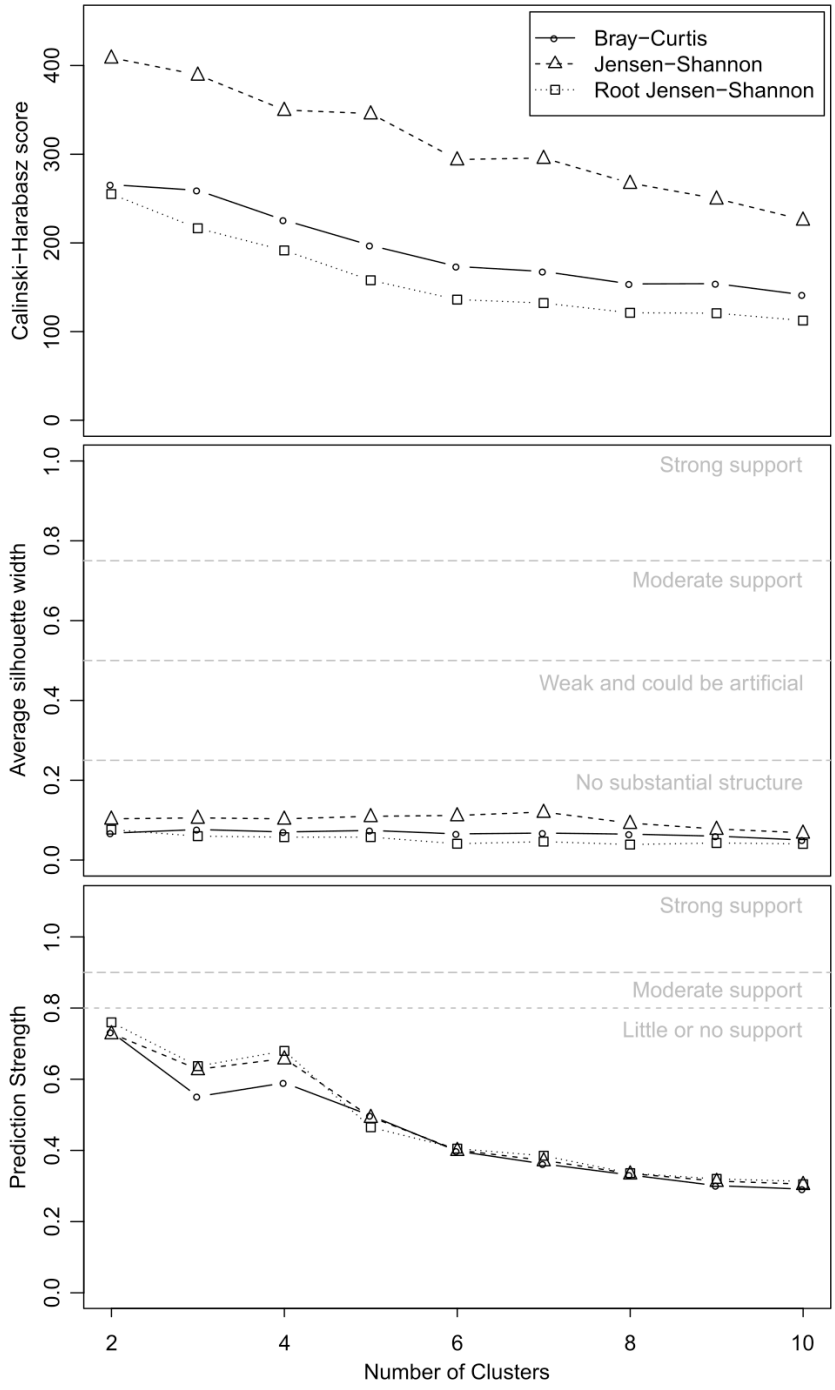
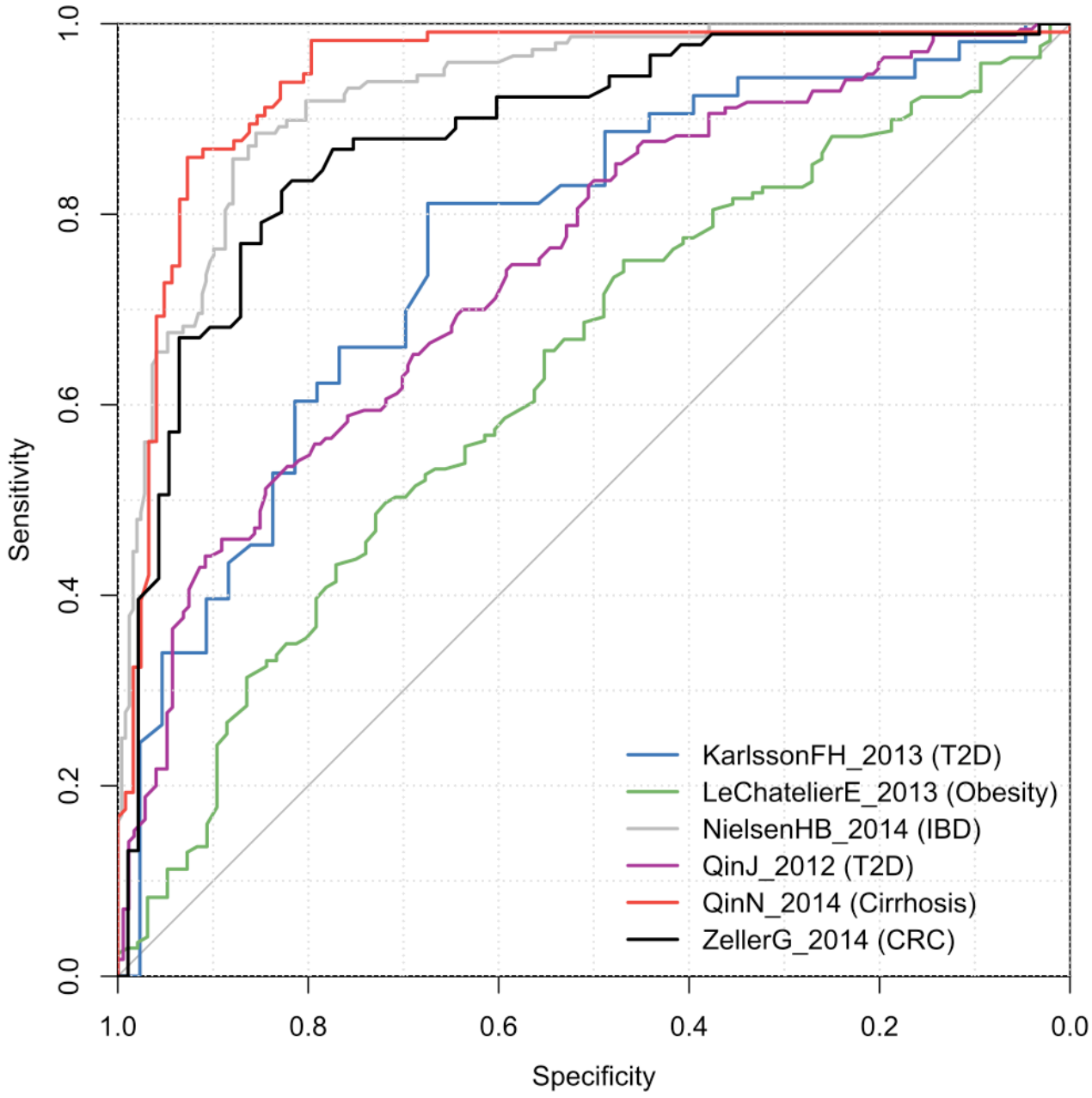


Supplementary Figure 1

Clustering scores for enterotypes in gut WGS samples.

Consistent with Koren *et al.* [5], these plots indicate weak support for any discrete clustering in the data and confirm that the three enterotypes hypothesis is likely an oversimplification that does not hold when considering large set of biogeographycally diverse populations. Thresholds for significance of clustering are presented as dashed lines, and are the same thresholds used by Koren et *al.* [5]. Each plot line represents an analysis that can be accomplished with one line of code using the R packages 'fpc' (prediction strength and Calinski-Harabasz) and 'cluster' (silhouette index), provided in the curatedMetagenomicData package examples.



Supplementary Figure 2

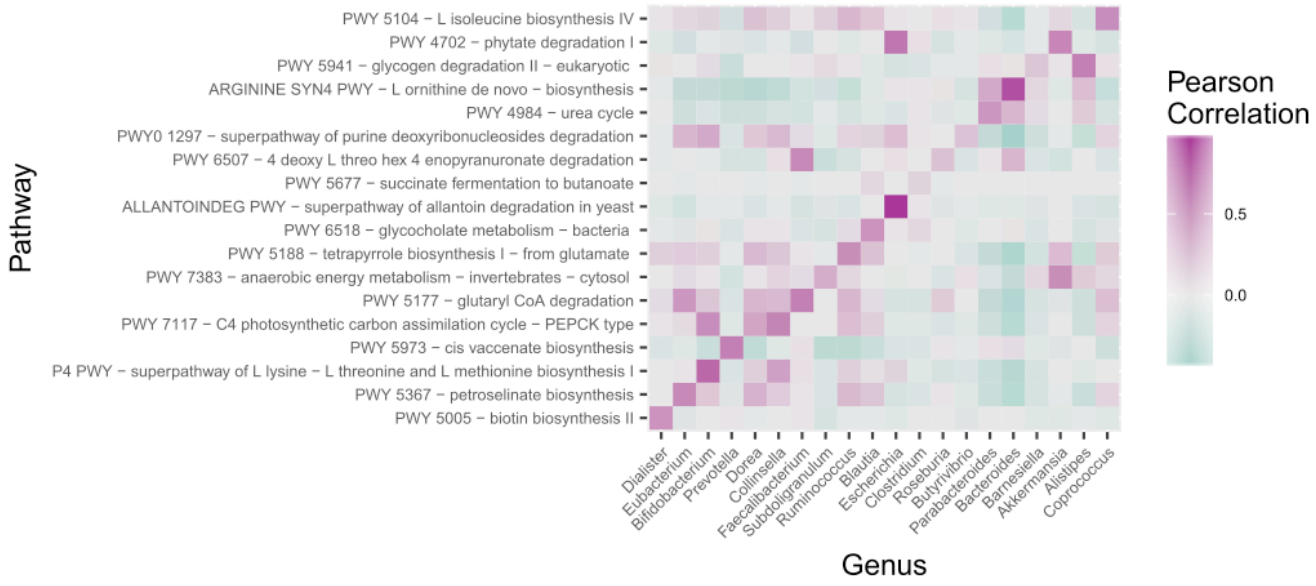
Health status classification from species abundance.

Six different classification problems of health status were attempted using a random forest algorithm and cross-validation to estimate prediction accuracy. Plots show ROC curves by using species abundance as microbiome features, one of the five data types considered in the Example 1 of **Figure 1**. Results are consistent with the meta-analysis conducted in [32].

Supplementary Figure 3

Principal Coordinates Analysis (PCoA) plot of species abundance for gut samples on selected diseases.

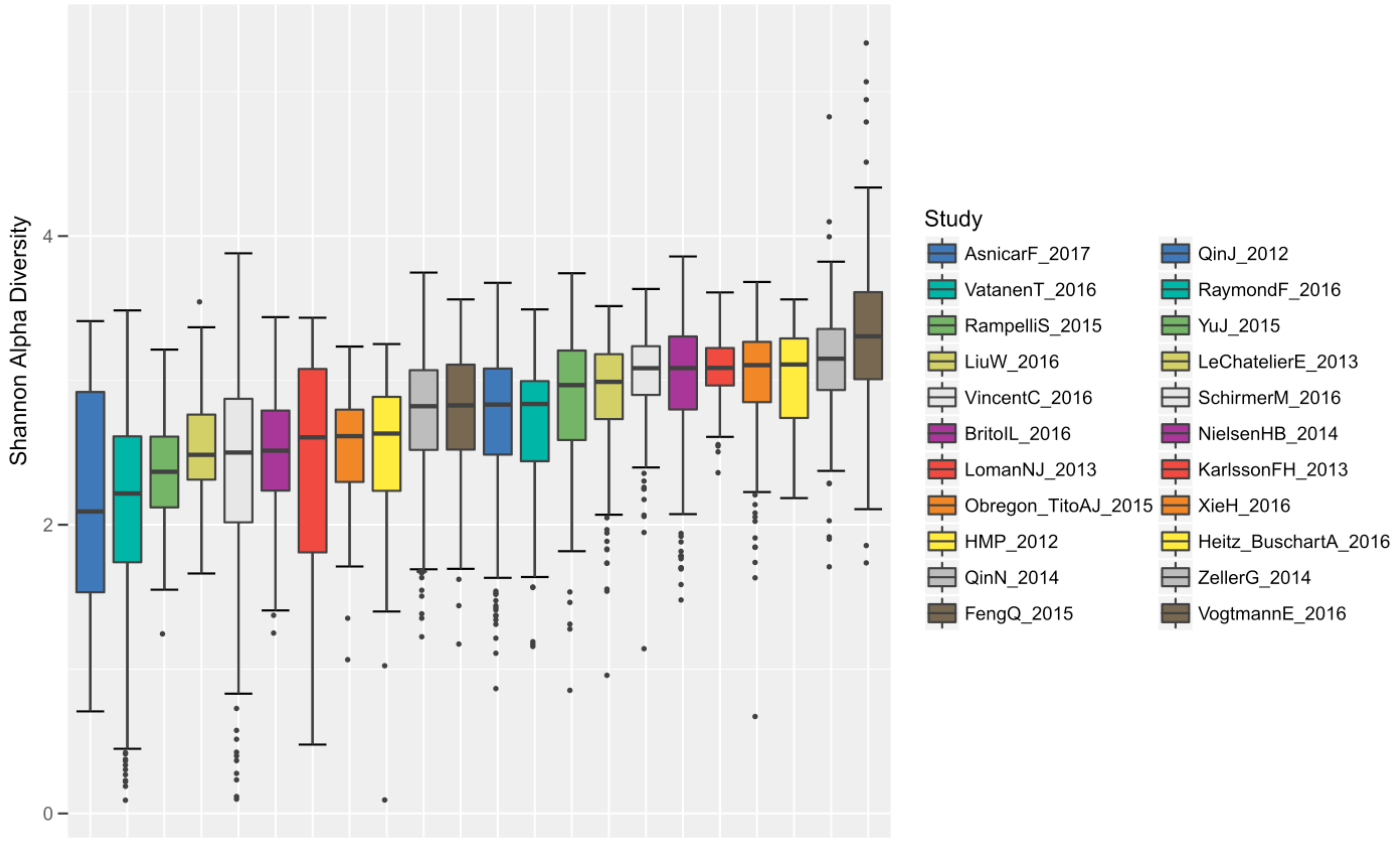
Specimens are annotated by disease state (shape), study name (color), and abundance of *Prevotella copri* (size).



Supplementary Figure 4

Top correlations between metabolic pathways and genera.

Pearson correlation was calculated between each individual pathway (HUMAnN2 pathways from the full UniRef90 database) and each of the top 20 most abundant microbial genera, in a combined dataset obtained from merging 20 studies of gut specimens. The top correlations are 1) Ornithine de novo biosynthesis: Bacteroides (r = 0.86), activity that has been confirmed in cultures of this organism [33], and 2) superpathway of allantoin degradation in yeast: Escherichia (r =0.95). Although this superpathway has been associated with yeast, it includes subpathways (such as allantoin degradation to glyoxylate I and allantoin degradation to ureidoglycolate I) that are common in Escherichia, which is known to be an allantoin utilizier under anaerobic conditions [34]. Of note, the top 100 correlations have adjusted $p < 0.001$.



Supplementary Figure 5

Alpha diversity of taxa from 22 studies of the gut microbiome.

Shannon Alpha Diversity was calculated for each individual sample within each human gut microbiome study. The median diversity varies by a maximum factor of 1.5 between studies, however the variability within studies as measured by interquartile range varies by more than 3-fold.

Supplementary Methods

Available datasets

To date (development version 1.7.5), we have curated metadata for and packaged a total of 5,716 publicly available shotgun metagenomic samples from 26 large-scale studies (see **Supplementary Tables 1-2**). Accessing the most recently added datasets requires installing the development version of Bioconductor. All these metagenomes have been sequenced on the Illumina platform at an average depth of 45 M reads.

Twelve of these studies were performed to assess the association of the human gut microbiome with different diseases. In particular, four studies were devoted to the characterization of the human microbiome in colorectal cancer patients: FengQ_2015[11] (154 samples, 93 cases), VogtmannE_2016[28] (110 samples, 52 cases), YuJ_2015[30] (128 samples, 75 cases), and ZellerG_2014[31] (199 samples, 133 cases). Heintz-BuschartA_2016[12] includes a total of 53 samples, 27 of which are associated with type 1 diabetes (T1D). KarlssonFH_2013[13] sampled on European women and includes 53 type 2 diabetes (T2D) patients, 49 impaired glucose tolerance individuals and 43 normal glucose tolerance individuals. QinJ_2012[20] sampled an additional T2D dataset and is composed by 170 Chinese T2D patients and 193 non-diabetic controls. LeChatelierE_2013[14] includes 123 non-obese and 169 obese individuals. LomanNJ_2013[16] includes 43 samples from patients with life-threatening diarrhea during the 2011 outbreak of Shiga-toxigenic Escherichia coli (STEC) O104:H4 in Germany. NielsenHB_2014[17] focuses on inflammatory bowel disease (IBD) and comprises a total of 396 samples, 21 of which are from Crohn's disease patients and 127 from ulcerative colitis patients. QinN_2014[21] includes 123 patients affected by liver cirrhosis and 114 healthy controls. VincentC_2016[27] focused on microbiota dynamics in response to hospital exposures and Clostridium difficile colonization infection in a total of 229 samples.

We included also four datasets that investigated gut configuration in hunter-gatherer or non-westernized populations. BritoIL_20168 considered agrarian Fiji islanders for a total of 312 samples, including also some samples from the oral cavity. LiuW_2016[15] investigated 110 Mongolian adults. Obregon-TitoAJ_2015[18] sequenced 58 samples, which include hunter-gatherer and traditional agriculturalist communities in Peru. RampelliS_2015[22] comprises 38 samples, part of which were collected from Hadza hunter-gatherers of Tanzania.

Additional datasets were acquired entirely from healthy subjects. AsnicarF_20177 collected 24 samples for studying vertical microbiome transmission from mothers to infants. RaymondF_2016[23] acquired 72 samples to evaluate effects of a standard antibiotic treatment on the microbiome. SchirmerM_2016[24] investigated 471 samples to link the microbiome to inflammatory cytokine production capacity. VatanenT_2016[26] considered 222 infants in Northern Europe from birth until age three for a total of 785 samples. XieH_2016[29] investigated 250 adult twins to evaluate genetic and environmental impacts on the microbiome.

Some datasets not strictly related to the gut microbiome are also taken into account. Castro-NallarE_20159 collected 32 samples from the oral cavity to investigate the oropharyngeal microbiome in individuals with schizophrenia. HMP4 includes 749 samples collected for the Human Microbiome Project from five major body sites (i.e., gastrointestinal tract, nasal cavity,

oral cavity, skin, and urogenital tract). Finally, three datasets focused on the skin microbiome. OhJ_2014[19] is composed by 291 samples collected from several different skin sites in healthy conditions. Skin samples but from patients affected by atopic dermatitis and psoriasis were acquired in ChngKR_2016[10] (78 samples) and TettAJ_2016[25] (97 samples), respectively.

Raw data pre-processing
Approximately 63 TB of raw sequencing data were downloaded from public repositories. All samples were subject to standard pre-processing as described in the SOP of the Human Microbiome Project4, without however the step of human DNA removal as these publicly available metagenomes were deposited free of reads from human DNA contamination.

MetaPhlAn2 profiling and data products
MetaPhlAn22 (v2.0) was ran on the pre-processed reads with default parameters to generate microbial community profiles (from kingdom- to species-level) including Bacteria, Archaea, microbial Eukaryotes and Viruses. These profiles were generated from ~1 M unique clade-specific marker genes identified from ~17,000 reference genomes (~13,500 bacterial and archaeal, ~3,500 viral, and ~110 eukaryotic). MetaPhlAn2 has the capability of characterizing organisms at a finer resolution using non-aggregated marker information ("-t marker_pres_table" and "-t marker_ab_table" mode). Single marker-level profiles were then merged in samples versus markers tables removing markers there were never detected in any samples.

Such processing resulted in three data products: i) species-level relative abundance (denoted as "metaphlan_bugs_list" in the package); ii) marker presence ("marker_presence"); and iii) marker abundance ("marker_abundance"). Species abundance is expressed in percentage and sum up to hundred within each sample when selecting a single taxonomic level. Marker presence and marker abundance assume binary and real values, respectively.

HUMAnN2 profiling and data products
HUMAnN23 (v0.7.1) was run on the pre-processed reads with default parameters for profiling the presence/absence and abundance of microbial pathways in the community. The mapping was done using the full UniRef90 database (~11 GB), which enabled identifying also protein families without functional annotations. Three main outputs were generated: gene family abundance, pathway abundance, and pathway coverage. The two abundance output files were normalized in terms of relative abundance through the "humann2_renorm_table" ("--units relab" mode).

In this way, three additional data products were produced: i) normalized gene family abundance (denoted as "genefamilies_relab" in the package); ii) normalized pathway abundance ("pathabundance_relab"); and iii) pathway coverage ("pathcoverage"). Features assume values in the range [0, 1], where the two normalized abundance profiles sum up to 1 when excluding species-specific contributions.

Creation of curatedMetagenomicData

To create the curatedMetagenomicData package, processed data, in the form of tab-delimited files, from the MetaPhlAn2 and HUMAnN2 pipelines and patient-level metadata are compressed into a single archive file per dataset. Then from within the R/Bioconductor environment a single function is used to process the compressed archive, create documentation, and add to curatedMetagenomicData, with internal intermediate steps as follows. First, patient-specific metadata is read in using the readr package (https://CRAN.R-project.org/package=readr), filtered using the dplyr (https://CRAN.R-project.org/package=dplyr) and magrittr (https://CRAN.R-project.org/package=magrittr) packages, and coerced to the appropriate format. Study-level metadata is then created by querying PubMed using the RISmed package (https://CRAN.R-project.org/package=RISmed), which collects citation information of published studies that can then be coerced to the appropriate format. Finally, patient-level sample data is read in (again using the readr package), merged, standardized, and used to create Bioconductor ExpressionSet objects featuring the patient and study-level metadata. Within each study, processed data is separated into six data products, as highlighted above, and further separated by bodysite so as to allow for efficient search and data transfer.

Once data from the MetaPhlAn2 and HUMAnN2 pipelines have been processed into Bioconductor ExpressionSet objects, documentation, package metadata, and upload to ExperimentHub are accomplished using developer functions available in curatedMetagenomicData. Documentation is automatically produced from the ExpressionSet objects using roxygen2 (https://CRAN.R-project.org/package=roxygen2), although this may change in the future. Package metadata is also produced from the ExpressionSet objects and used in the creation of ExperimentHub records, with further details concerning ExperimentHub below. Finally, a convenience function is provided to write a shell script to upload all data to ExperimentHub, such that the error-prone process of working with Amazon Web Services (AWS) Command Line Interface (CLI) is trivial.

Bioconductor object classes

curatedMetagenomicData data objects are represented using the Bioconductor ExpressionSet S4 class. This class links numeric microbiome data with subject information and whole-experiment level data, while maintaining correct alignment between numeric microbiome data subject data during subset operations. The following ExpressionSet slots are populated in each data product:

- *experimentData*: "MIAME" class object providing study-level information - Pubmed ID, authors, title, abstract, sequencing technology, etc. Extracted using experimentData(object).
- *phenoData*: "AnnotatedDataFrame" class object providing specimen-level information - subject IDs, disease, body site, number of reads, etc. Extracted using pData(object) or phenoData(object).
- *assayData*: matrix class object providing taxonomic or pathway abundances. Extracted using exprs(object).

ExpressionSet objects can be analyzed for differential abundance using popular Bioconductor packages for RNA-seq such as *limma, edgeR, and DESeq2.* For MetaPhlAn2 abundances, however, it is more convenient to convert these to *phyloseq* objects for analysis with the

phyloseq Bioconductor package for phylogenetics, using the *ExpressionSet2phyloseq* function from *curatedMetagenomicData*. Phyloseq objects additionally represent taxonomy and phylogenetic distances, and enable straightforward calculation of alpha and beta diversity measures, ordination plots, and other phylogenetic-specific analyses.

ExperimentHub

curatedMetagenomicData datasets are distributed through ExperimentHub, a new Bioconductor software package we developed to provide programmatic access to experimental data files stored in the Amazon Web Services (AWS) cloud. All data (referred to as "resources") in ExperimentHub have undergone some level of curation and are provided as R/Bioconductor data structures instead of in raw format. Data sets are generally a collation of different sources combined by disease or cohort or data used in a published experiment or short courses.

The two primary components of ExperimentHub are the data files and the metadata describing them. Files are stored in AWS S3 buckets and the metadata in a database on the ExperimentHub server. The database version is reflected in the "snapshot date" which is updated whenever the database is modified. Users interacting with ExperimentHub can select a specific snapshot date which, along with the version of R / Bioconductor, modifies which resources are exposed.

ExperimentHub resources are accessed by invoking ExperimentHub() to create an 'ExperimentHub' object, e.g., hub <- ExperimentHub(). This call downloads the database of metadata from the ExperimentHub server and caches it locally. The 'hub' of metadata can be searched with the query() function and subset by numerical index or 'EH' identifier. Once a resource is identified, the double-bracket method ('[[') will initiate the download. Downloaded resources are cached locally enabling fast repeated access to the data. When a resource is loaded in an R session, the accompanying software package is also loaded ensuring all documentation and helper functions are readily available. A second option for accessing the data is to invoke the resource name as a function, e.g., data123(). In this approach, the creation and searching of the 'hub' is not exposed to the user and does not require knowledge of ExperimentHub objects.

Resources are added to ExperimentHub by creating a software package according to the guidelines in the ExperimentHubData vignette (https://bioconductor.org/packages/release/bioc/vignettes/ExperimentHubData/inst/doc/ExperimentHubData.html). The software package includes man pages and a vignette documenting expected use as well as functions to create the resource metadata. If desired, the author may include additional functions for resource discovery and manipulation. Data are stored separately in AWS and are not part of the software package; this separation enables lightweight installation of the package regardless of the size of the data.

Accessing curatedMetagenomicData objects in R

Within the R/Bioconductor environment there are two distinct methods for accessing data, depending on the needs of the end-user. In the case that a specific dataset is desired and its name is known, then convenience functions have been provided for all datasets and calling the function will retrieve the dataset from ExperimentHub. Otherwise, if no specific dataset

is desired, it is possible to search through all datasets and return those matching a pattern (e.g., all datasets from the stool bodysite). This method also features wildcard search to allow for powerful selection and can return either a list of references to the datasets or download the datasets from ExperimentHub. The later search method is of particular use in conducting cross validation studies using curatedMetagenomicData, as it provides for highly specific filtering conditions.

Accessing curatedMetagenomicData from the command line
A convenience command-line interface is provided for users who do not want to use the R or Bioconductor framework for the analysis. The command-line program is invoked with the names of one or more datasets with optional wildcard expansion, and provides flags for including specimen information in addition to microbiome data, and for returning relative abundances or counts. Datasets are written to disk as tab-separated value plain text files.

Examples of enabled downstream tasks: supervised classification analysis
We considered six different classification problems of health status to evaluate capabilities of disease classification from gut microbial profiling (see Example 1 of **Figure 1** and **Supplementary Figure 2**). In KarlssonFH_2013, we discriminated between "healthy" and "T2D" subjects. We took into account 96 samples after excluding impaired glucose tolerance individuals. In LeChatlierE_2013, we discriminated between "lean" (BMI ≤ 25 kg m^{-2}) and "obese" (BMI ≥ 30 kg m^{-2}) subjects for a total of 265 samples. Individuals having an intermediate BMI (i.e., > 25 and < 30 kg m^{-2}) were excluded. NielsenHB_2014 was composed by a total of 396 samples, in which the "diseased" class included inflammatory bowel disease (IBD) patients affected by both "Crohn's disease" and "ulcerative colitis". In QinJ_2012 we considered a total of 344 samples and discriminated between "healthy" and "T2D" individuals. In QinN_2014, all the 237 available samples (subdivided into "healthy" and affected by "liver cirrhosis" subjects) were taken into account. Finally, in ZellerG_2014 we removed the individuals affected by "large adenoma", which resulted in a total of 184 samples. "Cancer" patients were discriminated from "healthy" subjects, which included also persons affected by "small adenoma".

We compared five different data products, three taxonomic (i.e., relative abundance, marker presence, and marker abundance) and two functional (i.e., normalized pathway abundance and pathway coverage). We subset relative abundance profiles to consider only species-level features, while the whole set of available features were taken into account for the other four data products.

The classification problems were attempted using the random forest algorithm through the R packages "randomForest" and "caret". Original features were preprocessed ("preProc") by centering ("center"), scaling ("scale") and removal of zero-variance predictors ("zv") procedures. Prediction accuracies were estimated using a 10-fold cross-validation approach ("method=repeatedcv" and "number=10" in the "trainControl" function). The two main parameters of the classifier were set in this way: i) the number of trees ("ntree") was set to 500; ii) the number of variables randomly sampled as candidates at each split ("mtry") were estimated through grid search. Area under the curve (AUC) values were computed through the "auc" function in the R package "pROC". The scatterplot matrix of AUC values (Example 1 of **Figure 1**) was generated through the R package "gclus", which provided possibility to i)

rearrange the variables so that those with higher correlations are closer to the principal diagonal and ii) color the cells to reflect the value of the correlations. The "pROC" package was also adopted to plot the receiver operating characteristic (ROC) curves (**Supplementary Figure 2**) using the "roc" function.

Examples of enabled downstream tasks: unsupervised clustering analysis

To assess the presence of discrete clustering in the data (see Example 2 of **Figure 1** and **Supplementary Figure 1**), we merged taxonomic abundance data from all gut samples (excluding newborns), on which we calculated three distance measures using the R package "phyloseq": the Bray-Curtis distance metric, the Jenson-Shannon divergence (JSD), and the square root of the Jenson-Shannon divergence (root-JSD). We then performed clustering against each of the three distance measures by partitioning around medoids using the R package "cluster". We determined the optimal number of clusters based on the prediction strength (PS) using the R package "fpc", and silhouette index (SI) using the R package "cluster". We used a threshold of ≥ 0.90 for PS, and ≥ 0.75 for SI, to indicate strong clustering5. We additionally calculated the Calinski-Harabasz (CH) statistic for comparison to PS and SI, using the R package "fpc".

Package maintenance

We set up the curatedMetagenomicData to be scalable to the growing size of metagenomic datasets being produced and we plan to expand to over 10K total samples by the end of 2017, with dedicated personnel for the addition of processed metagenomic datasets. The curatedMetagenomicData pipeline directly uses output of the publicly available MetaPhlAn2 and HUMAnN2 packages, in a documented subdirectory structure for data "handoff" to our pipeline for incremental dataset addition to curatedMetagenomicData in ExperimentHub (https://github.com/waldronlab/curatedMetagenomicData/wiki).

Authors welcome the addition of new datasets provided they can be or already have been run through the MetaPhlAn2 and HUMAnN2 pipelines. Please contact the maintainer if you have a shotgun metagenomic dataset that would be of interest to the Bioconductor community.

Availability and support

The curatedMetagenomicData package can be installed with a single command from an R installation with the current Bioconductor release or development version installed (`BiocInstaller::biocLite("curatedMetagenomicData")`). The package is described at https://waldronlab.github.io/curatedMetagenomicData/, including information on installation, datasets to be added in the near future, and example analyses. Requests for help should be raised at https://support.bioconductor.org with the tag *curatedMetagenomicData*. Bugs in code or curation should be reported using the issue tracker at https://github.com/waldronlab/curatedMetagenomicData/issues. Instructions for adding datasets or re-using parts or all of the pipeline for other purposes are provided on the wiki at https://github.com/waldronlab/curatedMetagenomicData/wiki.

Reproducible analysis

All analyses presented in this manuscript are reproducible by the script PaperFigures.Rmd at https://github.com/waldronlab/curatedMetagenomicData/tree/master/vignettes/extras.

Licensing

The curatedMetagenomicData package is licensed under the permissive Artistic 2.0 license.

Supplementary References

1. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12,** 115–121 (2015).
2. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomics taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
3. Abubucker, S. *et al.* Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* **8,** e1002358 (2012).
4. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214 (2012).
5. Koren, O. *et al.* A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput. Biol.* **9,** e1002863 (2013).
6. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).
7. Asnicar, F. *et al.* Studying vertical microbiome transmission from mothers to infants by strain-level metagenomic profiling. *mSystems* **2**, e00164-16 (2017).
8. Brito, I. L. *et al.* Mobile genes in the human microbiome are structured from global to individual scales. *Nature* **535**, 435–439 (2016).
9. Castro-Nallar, E. *et al.* Composition, taxonomy and functional diversity of the oropharynx microbiome in individuals with schizophrenia and controls. *PeerJ* **3**, e1140 (2016).
10. Chng, K. R. *et al.* Whole metagenome profiling reveals skin microbiome-dependent susceptibility to atopic dermatitis flare. *Nat. Microbiol.* **1**, 16106 (2016).
11. Feng, Q. *et al.* Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat. Commun.* **6**, 6528 (2015).
12. Heintz-Buschart A. *et al.* Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat Microbiol.* **2**, 16180 (2016).
13. Karlsson, F. H. *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**, 99–103 (2013).
14. Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500,** 541–546 (2013).
15. Liu, W. *et al.* Unique features of ethnic mongolian gut microbiome revealed by metagenomic analysis. *Sci. Rep.* **6**, (2016).
16. Loman, N. J. *et al.* A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104:H4. *JAMA* **309**, 1502–1510 (2013).
17. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32,** 822–828 (2014).
18. Obregon-Tito, A. J. *et al.* Subsistence strategies in traditional societies distinguish gut microbiomes. *Nat. Commun.* **6**, 6505 (2015).
19. Oh, J. *et al.* Biogeography and individuality shape function in the human skin metagenome. *Nature* **514**, 59–64 (2014).
20. Qin, J. *et al.* A metagenome-wide association study of the gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).
21. Qin, N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59–

64 (2014).
22. Rampelli, S. *et al.* Metagenome sequencing of the Hadza hunter-gatherer gut microbiota. *Curr. Biol.* **25,** 1682–1693 (2015).
23. Raymond, F. *et al.* The initial state of the human gut microbiome determines its reshaping by antibiotics. *ISME J.* **10**, 707 (2016).
24. Schirmer, M. *et al.* Linking the human gut microbiome to inflammatory cytokine production capacity. *Cell* **167**, 1125–1136 (2016).
25. Tett, A. J. et al. Unexplored diversity and strain-level structure of the skin microbiome associated with psoriasis. *NPJ Biofilms Microbiomes* **3**, 1 (2017).
26. Vatanen, T. *et al.* Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. *Cell* **165**, 842–853 (2016).
27. Vincent, C. *et al.* Bloom and bust: intestinal microbiota dynamics in response to hospital exposures and Clostridium difficile colonization or infection. *Microbiome* **4**, 12 (2016).
28. Vogtmann, E. *et al.* Colorectal cancer and the human gut microbiome: reproducibility with whole-genome shotgun sequencing. *PLOS One* **11**, e0155362 (2016).
29. Xie, H. *et al.* Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. *Cell Syst.* **3**, 572–584 (2016).
30. Yu, J. *et al.* Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* **66**, 7078 (2017).
31. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10,** 766 (2014).
32. Pasolli, E. *et al.* Machine learning meta-analysis of large metagenomics datasets: tools and biological insights. *PLoS Comput. Biol.* **12,** e1004977 (2016).
33. Shi, D. *et al.* Structure and catalytic mechanism of a novel N-succinyl-L-ornithine transcarbamylase in arginine biosynthesis of Bacteroides fragilis. *J. Biol. Chem.* **281,** 20623–20631 (2006).
34. Cusa, E. *et al.* Genetic analysis of a chromosomal region containing genes required for assimilation of allantoin nitrogen and linked glyoxylate metabolism in Escherichia coli. *J. Bacteriol.* **181**, 7479–7484 (1999).

Supplementary Table 1: Metadata fields available in curatedMetagenomicData. These fields are continuously and formally checked for syntax in all datasets, at https://github.com/waldronlab/curatedMetagenomicDataCuration.

Metadata Field	Description
adiponectin	Curators must use mg/l
age	Subject age (years)
age_category	Age category: newborn < 1 year; 1 <= child < 12; 12 <= schoolage < 19; 19 <= adult <= 65; senior > 65
ajcc	AJCC staging for colorectal-cancer
albumine	Albumine level; curators must use g/l
alcohol	Subject is reported as a drinker
antibiotics_current_use	Subject is currently taking antibiotics
antibiotics_family	Family of antibiotics currently used; Semicolon-separated
bilubirin	Bilubirin; curators must use mg/dl
birth_control_pil	Use of the birth-control-pils at the sampling time (men: no)
BMI	Body mass index (kg/m2)
body_site	Bodysite of acquisition
body_subsite	Subsite of body site of acquisition
cd163	Curators must use ng/ml
cholesterol	Curators must use mg/dl
country	Country of acquisition using ISO3 code from http://www.fao.org/countryprofiles/iso3list/en/
c_peptide	Curators must use ng/ml
creatine	Curators must use micro-mol/l
creatinine	Curators must use micro-mol/l
ctp	Cytidine triphosphate level
days_after_onset	Days from the onset of the disease
days_from_first_collection	Used for time series studies
disease	Semicolon-delimited vector of conditions; Use healthy only if subject is known to be healthy; CRC=colorectal cancer
disease_subtype	Disease subtype; CD=Chrohn's Disease
DNA_extraction_kit	DNA extraction kit
dyastolic_p	Measured in mm/Hg
ever_smoker	Ever been a smoker
family	A number identifying the family subjects belong; not corrected for meta-analyses
fasting_insulin	Curators must use micro-units/ml
ferm_milk_prod_consumer	Dfmp means yes (defined milk product)
fgf_19	Curators must use pg/ml
flg_genotype	Any term for filaggrin-protein genotype
fobt	Fecal occult blood test
gender	Subject gender
glp_1	Curators must use pmol/l
glucose	Curators must use mg/dl
glutamate_decarboxylase_2_antibody	Glutamic acid decarboxylase (GAD65) antibody assay
hba1c	Curators must use %
hdl	Curators must use mg/l

hitchip_probe_class	High/Low species content on the HIT-chip probe
hitchip_probe_number	HIT-chip probe score
hla_dbq11	Hla_dbq11 allele
hla_dbq12	Hla_dbq12 allele
hla_dqa11	Hla_dqa11 allele
hla_dqa12	Hla_dqa12 allele
hla_drb11	Hla_drb11 allele
hla_drb12	Hla_drb12 allele
hscrp	High-sensitivity C-reactive protein test result
il_1	Curators must use pg/ml
infant_age	Infant age (days); should be used for infants < 2 years old
inr	International normalized ratio
insulin_cat	Insulin intake as a boolean
lactating	Lactating subjects (men: no)
ldl	Curators must use mg/l
leptin	Curators must use micrograms/l
location	Free-form additional location information
median_read_length	Median read length - calculated from raw data
mgs_richness	Metagenomic species richness
minimum_read_length	Minimum read length - calculated from raw data
momeducat	Years of education of the mother of the subject
mumps	Subject has been through mumps in life
NCBI_accession	Semicolon-separated vector of NCBI accessions
non_westernized	Subject belongs to a non-westernized community
number_bases	Total number of bases sequenced in the sample
number_reads	Number of final reads - calculated from raw data
PMID	Identifier of the main publication in PubMed
pregnant	Pregnancy of the subject (men: no)
protein_intake	Indication about the protein intake in the Mongolians diet
prothrombin_time	Prothrombin time in seconds
sampleID	Sample identifier
sequencing_platform	This will be modified as new sequencing platforms are added to the database
shigatoxin_2_elisa	Enzyme-linked immunosorbent assay for Shiga-toxigenic E.coli
smoker	Currently a smoker at sampling
stec_count	Amount of STEC colonies detected
stool_texture	Texture of the stool at sampling time
study_condition	The main disease or condition under study; control for controls
subjectID	Subject identifier
systolic_p	Measured in mm/Hg
tnm	TNM classification for colorectal-cancer
triglycerides	Curators must use mg/l
visit_number	Visit number for studies with repeated visits

Supplementary Table 2: Study characteristics for the current development version (1.7.5) of the curatedMetagenomicData package. Note that accessing the most recently added datasets requires installing the development version of Bioconductor. Additional details on the datasets are available in the Supplementary Methods.

Dataset Name	Body Site	Disease	# Total Samples	# Case Samples	Average Reads per Sample (std) (M)	Size (Tb)	# Reads (G)	Reference
AsnicarF_2017	Stool, milk	None	26	-	21.4 (19.8)	0.2	0.5	7
BritoIL_2016	Stool, oral	Other condition	312	-	67.4 (51.8)	5.6	21.0	8
Castro-NallarE_2015	Oral	Schizophrenia	32	16	61.0 (25.2)	0.5	2.0	9
ChngKR_2016	Skin	Atopic dermatitis	78	38	15.8 (7.5)	0.3	1.2	10
FengQ_2015	Stool	Colorectal cancer	154	93	53.8 (8.5)	2.3	8.3	11
Heitz-BuschartA_2016	Stool	Type 1 diabetes	53	27	44.5 (0.9)	0.5	2.4	12
HMP_2012	Several	None	749	-	51.5 (44.8)	9.4	38.6	4
KarlssonFH_2013	Stool	Type 2 diabetes	145	53	31.0 (17.6)	1.4	4.5	13
LeChatelierE_2013	Stool	Obesity	292	169	69.0 (23.2)	4.0	20.1	14
LiuW_2016	Stool	Other condition	110	-	58.3 (26.8)	1.8	6.4	15
LomanNJ_2013	Stool	Shiga-toxigenic *E. coli*	43	43	9.2 (12.1)	0.1	0.4	16
NielsenHB_2014	Stool	Inflammatory bowel diseases	396	148	53.9 (20.2)	3.5	21.4	17
Obregon-TitoAJ_2015	Stool	Other condition	58	-	47.1 (20.9)	0.6	2.7	18
OhJ_2014	Skin	None	291	-	24.7 (38.1)	2.2	7.2	19

QinJ_2012	Stool	Type 2 diabetes	363	170	40.2 (11.8)	4.0	14.6	20
QinN_2014	Stool	Liver cirrhosis	237	123	51.6 (30.9)	3.0	12.2	21
RampelliS_2015	Stool	Other condition	38	-	22.3 (19.3)	0.2	0.8	22
RaymondF_2016	Stool	Other condition	72	-	135.1 (50.4)	2.7	9.7	23
SchirmerM_2016	Stool	None	471	-	30.3 (8.2)	3.1	14.3	24
TettAJ_2016	Skin	Psoriasis	97	97	3.0 (5.2)	0.1	0.3	25
VatanenT_2016	Stool	Other condition	785	171	21.0 (11.1)	4.4	16.4	26
VincentC_2016	Stool	CDI	229	33	17.4 (12.7)	1.6	4.0	27
VogtmannE_2016	Stool	Colorectal cancer	110	52	66.4 (15.6)	1.6	7.3	28
XieH_2016	Stool	Other condition	250	-	72.9 (9.1)	5.2	18.2	29
YuJ_2015	Stool	Colorectal cancer	128	75	56.3 (10.0)	2.1	7.2	30
ZellerG_2014	Stool	Colorectal cancer	199	133	63.5 (26.9)	2.9	12.6	31
TOTAL	-	-	5718	1441	44.5	63.3	254.3	-