Accessible, curated metagenomic data through ExperimentHub

Edoardo Pasolli[*,1], Lucas Schiffer[*,2], Audrey Renson[2], Valerie Obenchain[3], Paolo Manghi[1], Duy Tin Truong[1], Francesco Beghini[1], Faizan Malik[2], Marcel Ramos[2], Jennifer B. Dowd[2,4], Curtis Huttenhower[5,6], Martin Morgan[3], Nicola Segata[^,1], Levi Waldron[^,2]

Affiliations:
[1] Centre for Integrative Biology, University of Trento, Trento, Italy
[2] Institute for Implementation Science and Population Health, City University of New York School of Public Health, New York, New York, United States of America
[3] Roswell Park Cancer Institute, University of Buffalo, Buffalo, New York, United States of America
[4] Department of Global Health and Social Medicine, King's College London
[5] Biostatistics Department, Harvard School of Public Health, Boston, Massachusetts, United States of America
[6] The Broad Institute, Cambridge, Massachusetts, United States of America

[*] Equal contribution
[^] Corresponding authors: levi.waldron@sph.cuny.edu; nicola.segata@unitn.it

Supplemental Tale 1: Study characteristics for the first release of the curatedMetagenomicData package. Additional details on the datasets are available in the Methods.

Dataset Name	Body Site	Disease	# Total Samples	# Case Samples	Average Reads per Sample (std)	Size (Tb)	# Reads (G)	Reference
HMP_2012	Several	None	749	-	51.5M (44.8 M)	9.4	38.6	4
KarlssonFH_2013	Gut	Type 2 diabetes	145	53	31.0 M (17.6 M)	1.4	4.5	7
LeChatelierE_2013	Gut	Obesity	292	169	69.0 M (23.2 M)	4.0	20.1	8
LomanNJ_2013	Gut	Shiga-toxigenic *E. coli*	53	53	8.3 M (11.2 M)	0.15	0.4	9
NielsenHB_2014	Gut	Inflammatory bowel diseases	396	148	53.9 M (20.2 M)	3.5	21.4	10
Obregon-TitoAJ_2015	Gut	None	58	-	47.1 M (20.9 M)	0.6	2.7	11
OhJ_2014	Skin	None	291	-	24.7 M (38.1 M)	2.2	7.2	12
QinJ_2012	Gut	Type 2 diabetes	363	170	40.2 M (11.8 M)	4.0	14.6	13
QinN_2014	Gut	Liver cirrhosis	237	123	51.6 M (30.9 M)	3.0	12.2	14
RampelliS_2015	Gut	None	38	-	22.3 M (19.3 M)	0.23	0.8	15
TettAJ_2016	Skin	Psoriasis	97	97	3.0 M (5.2 M)	0.07	0.3	-
ZellerG_2014	Gut	Colorectal cancer	156	53	60.0 M (25.5 M)	1.8	9.4	16
TOTAL	-	-	2875	866	46.0 M (34.4 M)	30.3	132.3	-

Supplemental Table 2: Metadata fields available in curatedMetagenomicData

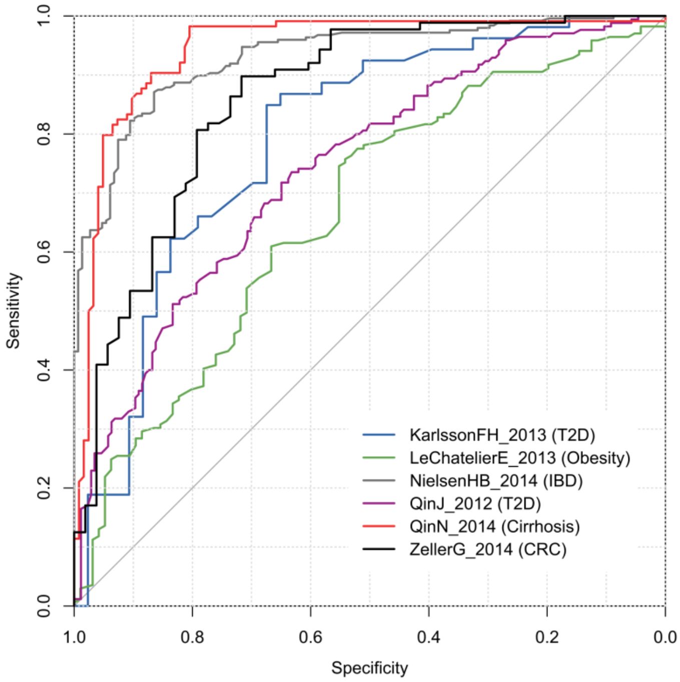
Metadata Field	**Description**
16s_rrna	16S rRNA analysis performed in the study
adiponectin	Adiponectin (mg/L)
affected	Affected syte
age	Subject age (years)
age_of_onset	Age of disease onset
age_range	Subject age range (years)
ajcc_stage	AJCC stage of the tumor (na: no classification for healthy controls or adenomas)
alb	Alb (g/L)
alcohol_related	Cirrhosis related to alcohol
antibiotic_usage	Has the subject used antibiotics
antivirus	Antivirus
arthritis	Has the subject arthritis
ascites	Ascites
beta-blocker	Beta-blocker
bmi	Body mass index (kg/m2)
bmi_class	Body mass index class
bodysite	Bodysite of acquisition
bsa	Body surface area (BSA)
c_difficile_frequency	Prediceted abundance of Clostridium difficile relative to other bacterial species detected in the sample in the MetaPhlAn analysis
c-peptide	C-peptide (nmol/L)
camp	Camp name
cd163	Cluster of differentiation 163 (ng/ml)
cholesterol	Cholesterol (mmol/L)
cirrhotic	Is the subject cirrhotic
classification	Classification
country	Country of acquisition
crea	Crea (umol/L)
creatinine	Creatinine (?mol/L)
ctp	CTP
daysafteronset	Days after onset of diarrhea
dbp	Diastolic blood pressure (mm Hg)
designation	Sample designation
dfmp	Known consumers of a defined fermented milk product (DFMP)
diabetic	Is the subject diabetic
disease	Disease presence and type
estimated_median_insert_size	Estimated median insert size
ethnicity	Subject ethnicity
fasting_glucose	Fasting glucose (mmol/L)
fasting_insulin	Fasting insulin (mU/L)
fbg	Fasting blood glucose (mmol/L)

fcp	Fasting serum C-peptide (ng/ml)
fgf-19	Fibroblast growth factor 19 (pg/ml)
fins	Fasting serum insulin (mU/L)
first	Identifier associated with the sampleID
fobt	Result of the fecal occult blood test (FOBT)
gad-antibodies	Glutamic acid decarboxylase antibodies (for units see [17])
gender	Subject gender
gene_count_class	Gene count class
gene_number	Gene number
gene_number_for_11m_uniquely_matched_reads	Gene number for 11 M uniquely matched reads
glp-1	Glucagon-like peptide 1 (pmol/L)
group	Sample group (control: healthy controls and patients with small adenomas; crc: patients with CRC; na: patients with large adenoma not included)
hba1c	Glycosylated hemoglobin A1c (mmol/mol)
hbalc	Glycosylated hemoglobin HbAlc (%)
hbv_related	Cirrhosis related to HBV
hdl	High-density lipoprotein (mmol/L)
he	HE
height	Subject height (cm)
hitchip_probe_class	HITChip Probe class
hitchip_probe_number	HITChip probe number
hscrp	High-sensitivity C-reactive protein (mg/L)
hus	Hemolytic-uremic syndrome
il-1	Interleukin 1 (pg/ml)
inr	INR
insulin	Insulin
ldl	Low-density lipoprotein (mmol/L)
leptin	Leptin (?g/L)
localization	Localization of the tumor/adenoma (rc: right colon; lc: left colon; lc/rc; multiple localizations; sigma: sigma; rectum: rectum)
matched_reads	Number of matched reads
meld	MELD
method	Acquisition method
mgs_profile_matched_sample_pairs	MGS profile matched sample pairs
mgs_richness	MGS richness
non_uniquely_align_to_human_with_0_2_mismatches	Number of reads non-uniquely aligned to human with 0-2 mismatches
nonhuman	Percentage of sequenced reads that did not align against the humane refence genome and thus were used in futher analysis
number_reads	Number of final reads
oral_anti-diabetic_medication	Oral anti-diabetic medication (meth: metformin; sulph: sulphonylurea)
other_causes_related	Cirrhosis related to other causes
paired_end_insert_size	Paired-end insert size (bp)

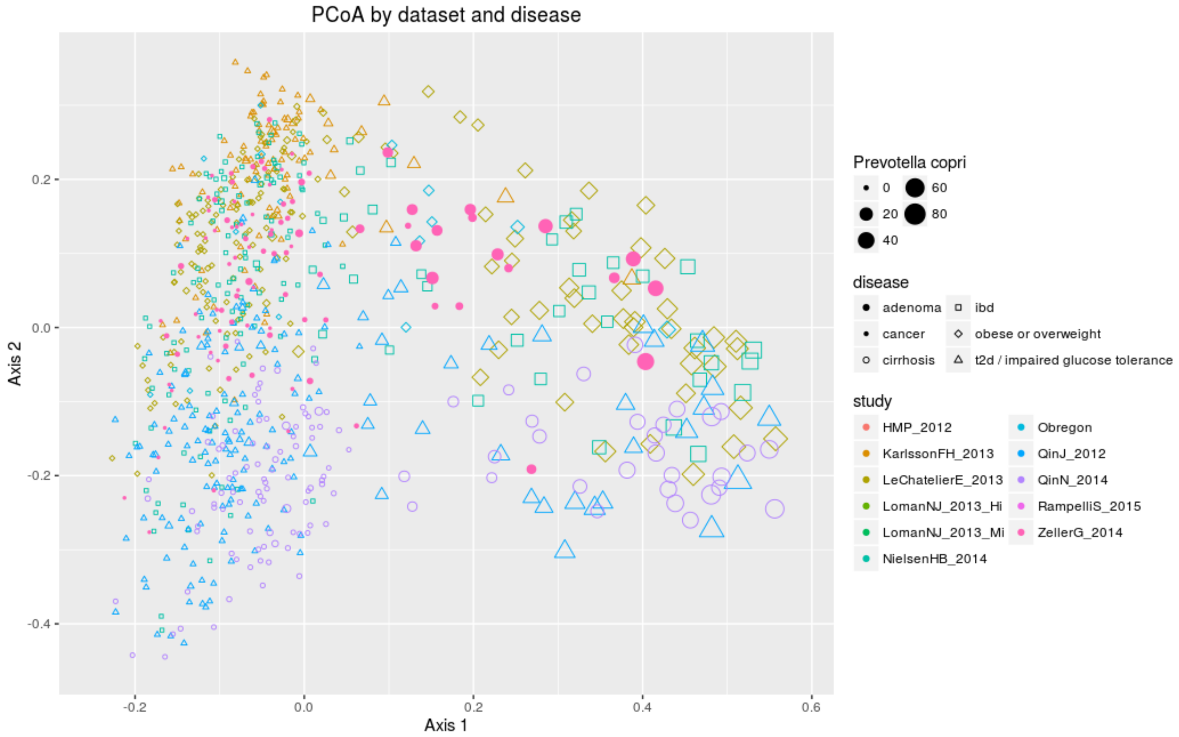
pasi	Psoriasis Area and Severity Index (PASI)
population	Subject population
pt	PT (S)
pubmedid	Identifier of the main publication in PubMed
quality_control	Number of reads after quality control
read_length	Read length (bp)
reads_removed_because_of_read_pair_trimming_discrepancy	Number of reads removed because of read pair trimming discrepancy
readsmillions	Number of original reads (millions)
repeat	Samples with the same repeat number were acquired from the same subject
reported_as_failed_qc	Number of reads reported as failed QC
sampleID	Sample identifier
sampling_day	Sampling day (relative to September 20th 2007)
sbp	Systolic blood pressure (mm Hg)
sequencing_technology	Sequencing technology
shigatoxin2elisa	Shiga-toxin 2 enzyme-linked immunosorbent assay
shotgun_metagenome	Shotgun metagenomic analysis performed in the study
site_characteristic	Syte characteristic
site_symmetry	Syte and symmetry of sample acquisition
snprnt	SNPRNT
stage	Acquisition stage/phase
statins	Statins
stec_count	Colony counts of STEC from samples (low < 10^4; moderate 10^4 to 10^6; high > 10^6 colony-forming units/mL)
stec_coverage	Average coverage of the chromosome of the STEC O104:H4 reference genome
stooltexture	Stool texture
stx_ratio	Ratio of reads mapping to the Shiga-toxin genes to the reads mapping to STEC chromosomal loci
stxab_detected	Shiga-toxin gene detected
subjectID	Subject identifier
tb	TB (umol/L)
tcho	Total cholesterol (mmol/L)
tg	Triglyceride (mmol/L)
tnfa	Tumor necrosis factor ? (ng/L)
tnm_stage	TNM stage of the tumor
too_short_after_quality_trimming(<50bp)	Number of reads too short after quality trimming (<50bp)
total_initial_reads	Number of initial reads
triglycerides	Triglycerides (mmol/L)
type	Psoriasis type
typingdata	Whether information on the serotpye (H4) and the multilocus sequence type for the outbreak strain could be recovered from the sample sequences
uniquely_align_to_human	Number of reads uniquely aligned to human
uniquely_matched_reads	Number of uniquely matched reads (two paired end

	reads that matched the same gene were counted as one read)
uniquely_matching_reads	Number of uniquely matching reads
visit_number	Visit number
wc	Waist circumference (cm)
weight	Subject weight (km)
whr	Waist-to-hip ratio (cm/cm)
wif-1_gene_methylation_test	Result of the wif-1 gene methylation test
wmsphase	Acquisition stage/phase
y-gt	?-glutamyltransferase (?kat/L)
years_in_sweden	Years in Sweden

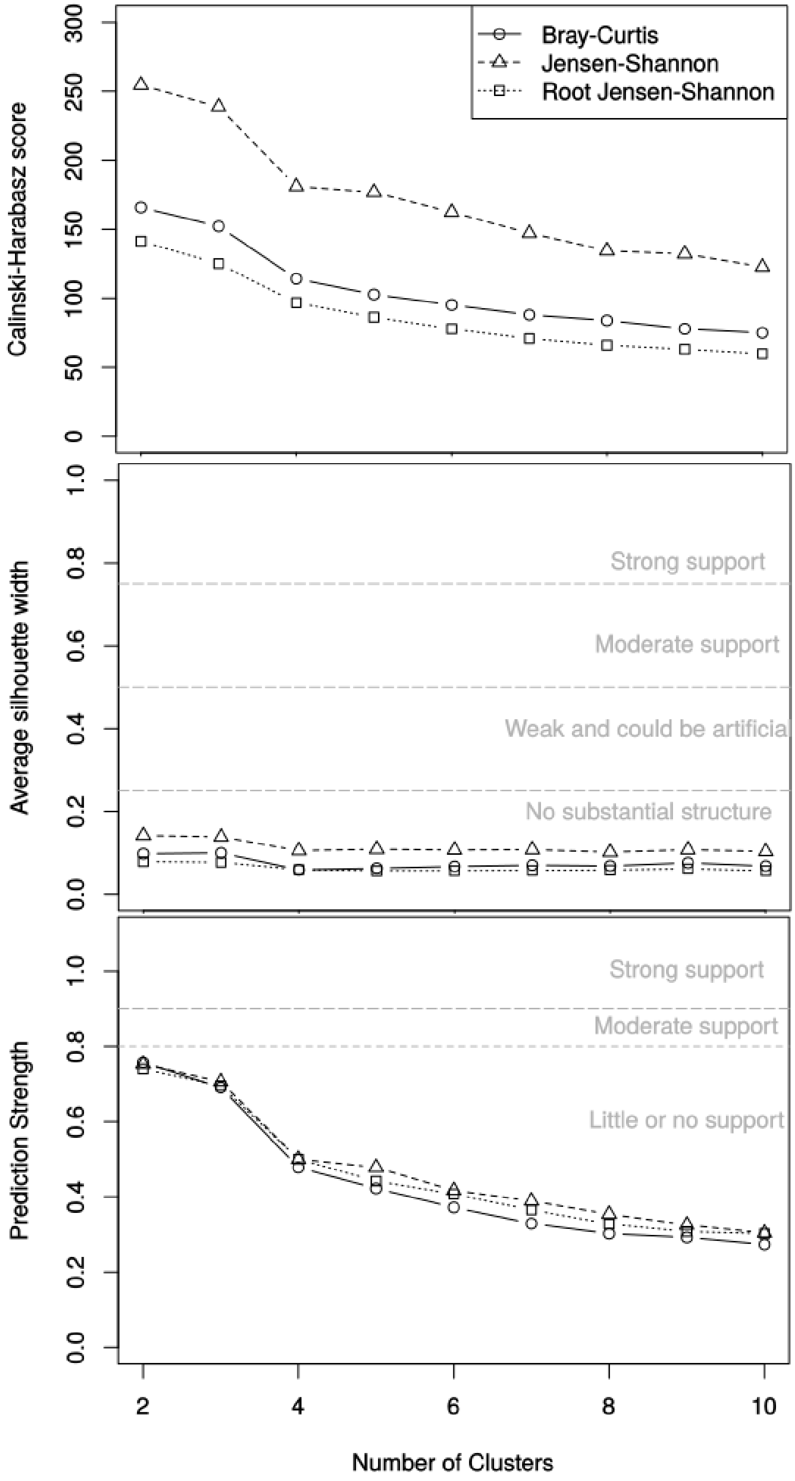
Supplemental Figure 1: **Health status classification from species abundance.** Six different classification problems of health status were attempted using a random forest algorithm and cross-validation to estimate prediction accuracy. Plots show ROC curves by using species abundance as microbiome features, one of the five data types considered in the Example 1 of **Figure 1**. Results are consistent with the meta-analysis conducted in [18].



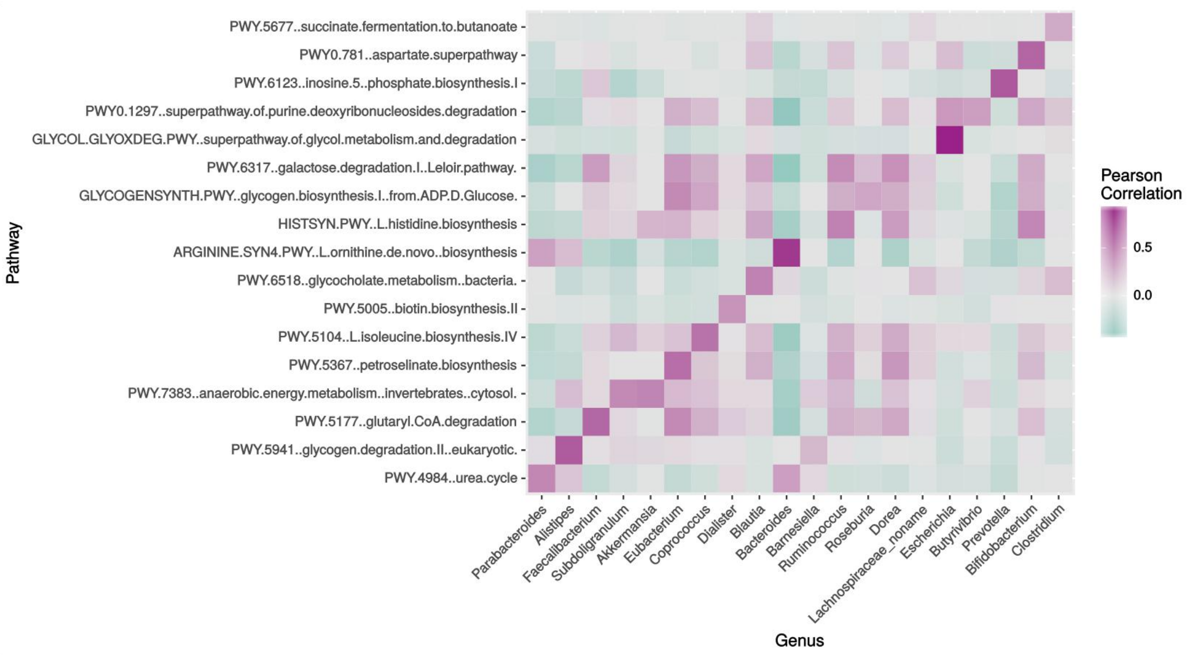
Supplemental Figure 2: Principal Coordinates Analysis (PCoA) plot of species abundance for all available stool samples. Specimens are annotated by dataset name (color), disease state (shape), and abundance of *Prevotella copri* (size).



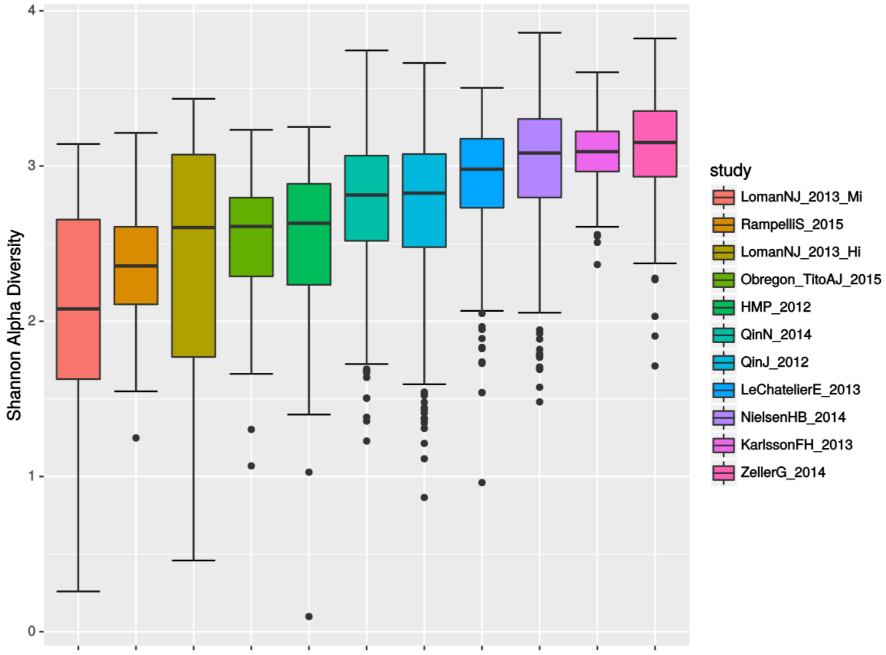
Supplemental Figure 3. Clustering scores for enterotypes in gut WGS samples. Consistent with Koren *et al.* [5], these plots indicate weak support for any discrete clustering in the data and confirm that the three enterotypes hypothesis is likely an oversimplification that does not hold when considering 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.



Supplemental 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 11 studies of stool specimens. The top correlations are 1) superpathway of glycol metabolism and degradation: Escherichia (r =0.96), and 2) Ornithine de novo biosynthesis: Bacteroides (r = 0.86), activities that have been confirmed in cultures of these organisms[19, 20]. Of note, the top 100 correlations have adjusted $p < 0.001$.



Supplemental Figure 5: **Alpha diversity of taxa from 11 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.


Shannon Alpha Diversity
study
LomanNJ_2013_Mi
RampelliS_2015
LomanNJ_2013_Hi
Obregon_TitoAJ_2015
HMP_2012
QinN_2014
QinJ_2012
LeChatelierE_2013
NielsenHB_2014
KarlssonFH_2013
ZellerG_2014