Accessible, curated metagenomic data through ExperimentHub

Edoardo Pasolli[*,1], Lucas Schiffer[*,2], Paolo Manghi[*,1], Audrey Renson[2], Valerie Obenchain[3], 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 current release (development version 1.3.7) of the curatedMetagenomicData package. 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]

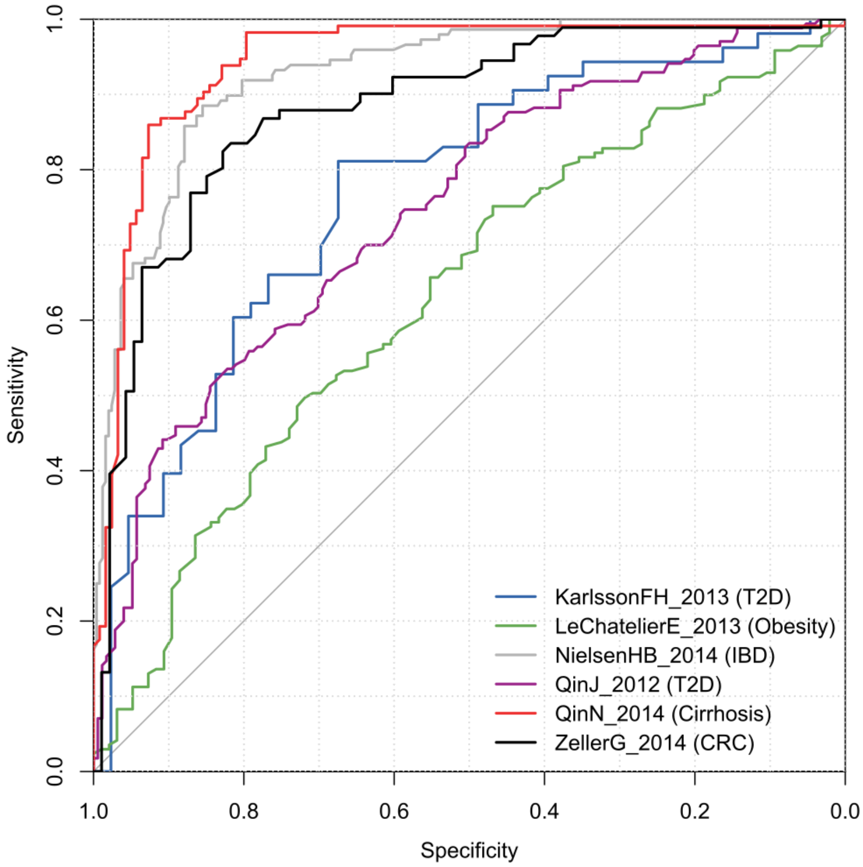
QinN_2014	Stool	Liver cirrhosis	237	123	51.6 (30.9)	3.0	12.2	[21]
RampelliS_2015	Stool	Other condition	38	-	22.3 (19.3)	0.2	0.8	[22]
RaymondF_2016	Stool	Other condition	72	-	135.1 (50.4)	2.7	9.7	[23]
SchirmerM_2016	Stool	None	471	-	30.3 (8.2)	3.1	14.3	[24]
TettAJ_2016	Skin	Psoriasis	97	97	3.0 (5.2)	0.1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

Supplemental Table 2: Metadata fields available in curatedMetagenomicData

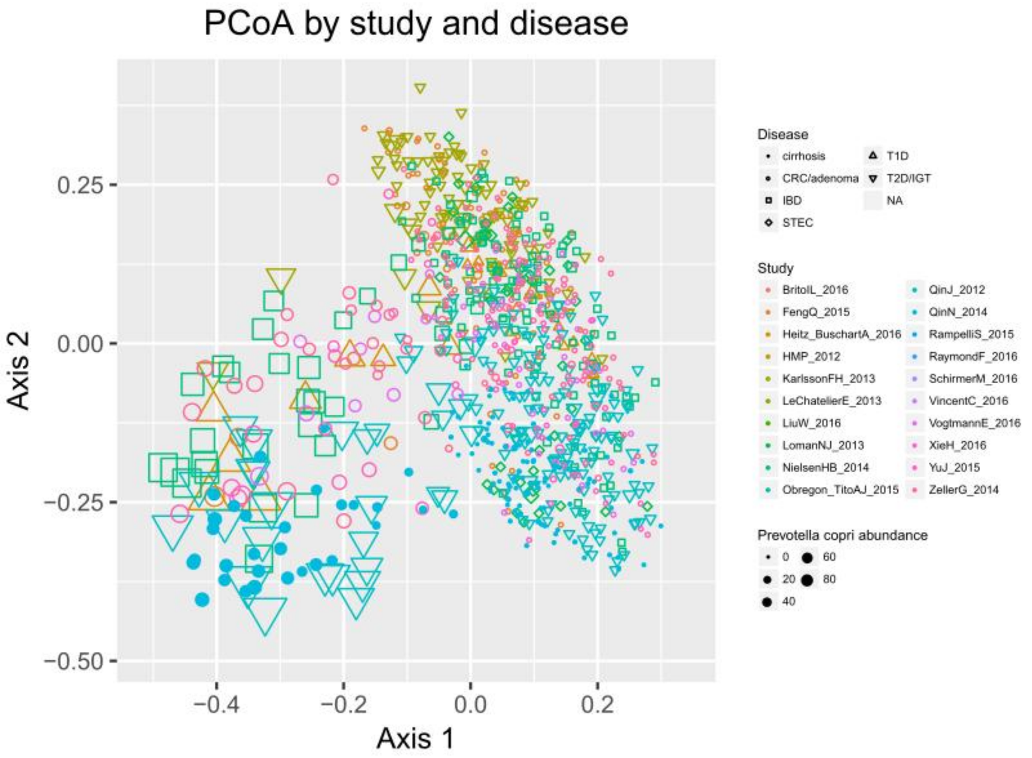
Metadata Field	**Description**
adiponectin	Curators must use mg/l
age	Subject age (years)
age_category	Age category: newborn < 1 year; 1 <= child < 12; 12 <= schoolage < 19; 19 <= adult <= 65; senior > 65
ajcc	AJCC staging for colorectal-cancer
albumine	Albumine level; curators must use g/l
alcohol	Subject 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 (yeah: funny)
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
start_solidfood	First day of solid food introduction (newborns)
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

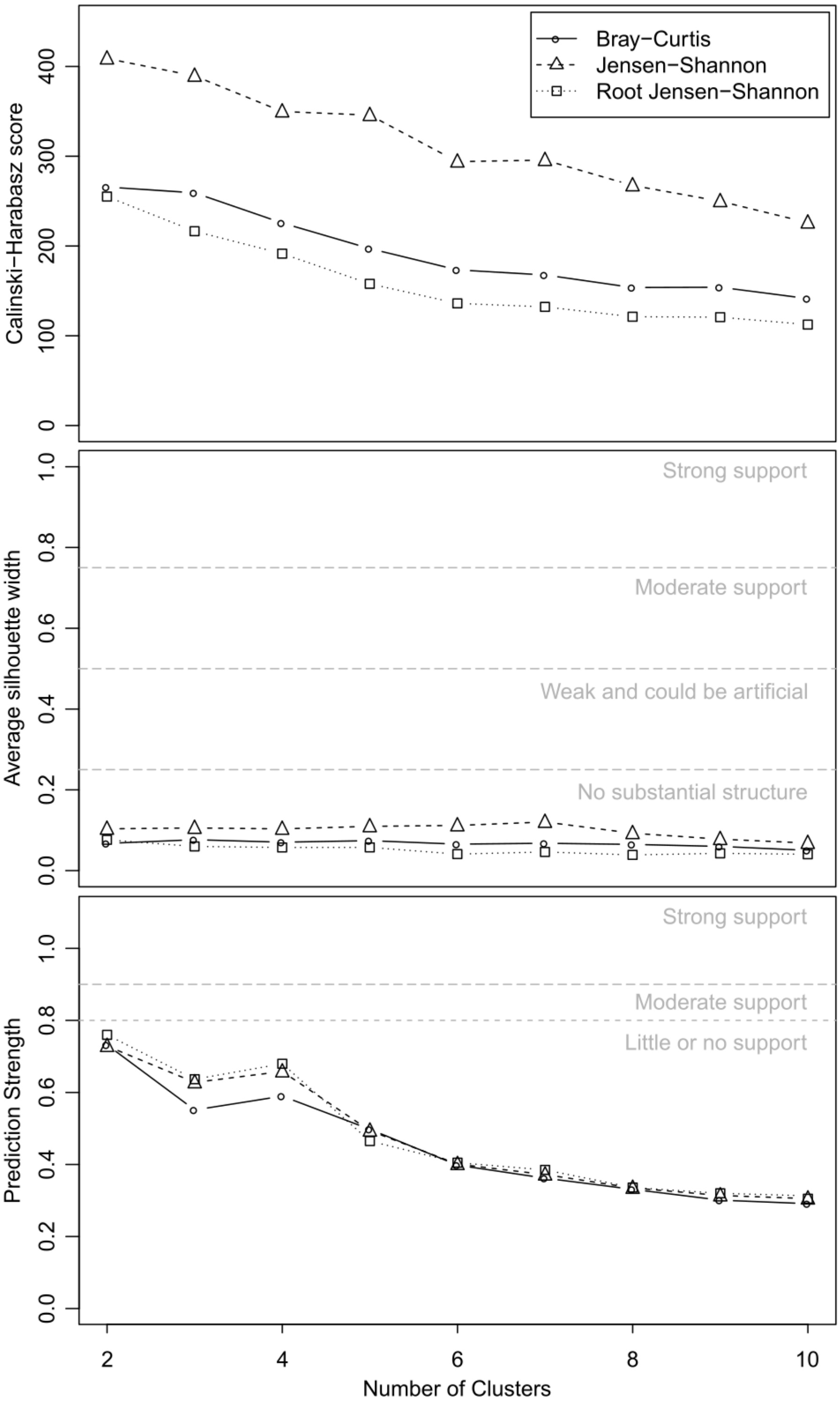
Supplemental Figure 1: **Health status classification from species abundance.** Six different classification problems of health status were attempted using a random forest algorithm and cross-validation to estimate prediction accuracy. Plots show ROC curves by using species abundance as microbiome features, one of the five data types considered in the Example 1 of **Figure 1**. Results are consistent with the meta-analysis conducted in [31].



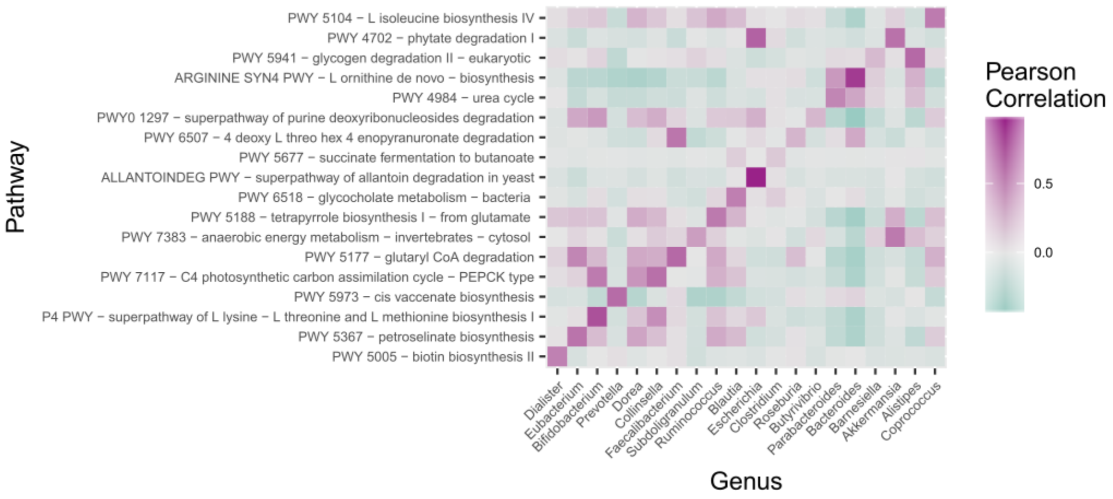
Supplemental Figure 2: Principal Coordinates Analysis (PCoA) plot of species abundance for stool samples on selected diseases. Specimens are annotated by disease state (shape), study name (color), and abundance of *Prevotella copri* (size).



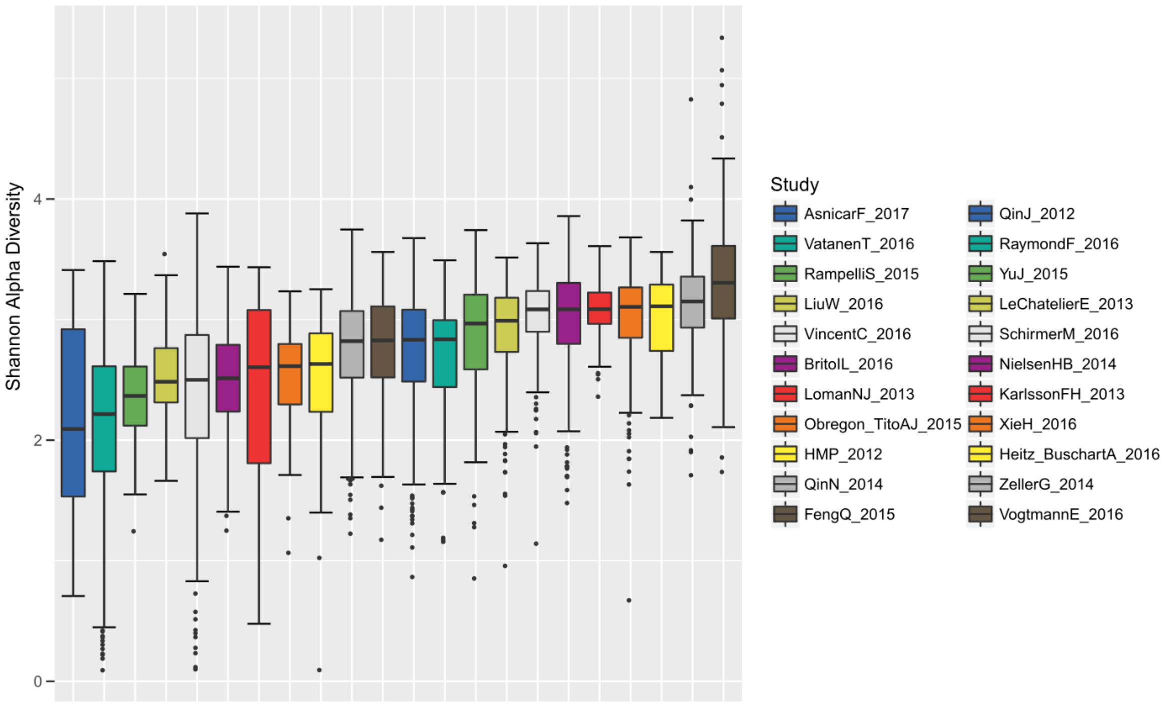
Supplemental Figure 3. Clustering scores for enterotypes in gut WGS samples. Consistent with Koren *et al.* [5], these plots indicate weak support for any discrete clustering in the data and confirm that the three enterotypes hypothesis is likely an oversimplification that does not hold when considering 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 20 studies of stool specimens. The top correlations are 1) Ornithine de novo biosynthesis: Bacteroides (r = 0.86), activity that has been confirmed in cultures of this organism[32], 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 [33]. Of note, the top 100 correlations have adjusted $p < 0.001$.



Supplemental 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.


Shannon Alpha Diversity
Study
AsnicarF_2017
VatanenT_2016
RampelliS_2015
LiuW_2016
VincentC_2016
BritoIL_2016
LomanNJ_2013
Obregon_TitoAJ_2015
HMP_2012
QinN_2014
FengQ_2015
QinJ_2012
RaymondF_2016
YuJ_2015
LeChatelierE_2013
SchirmerM_2016
NielsenHB_2014
KarlssonFH_2013
XieH_2016
Heitz_BuschartA_2016
ZellerG_2014
VogtmannE_2016