Human Microbiome: background, data, and statistical problems

Bressanone July 19-24 workshop

Levi Waldron

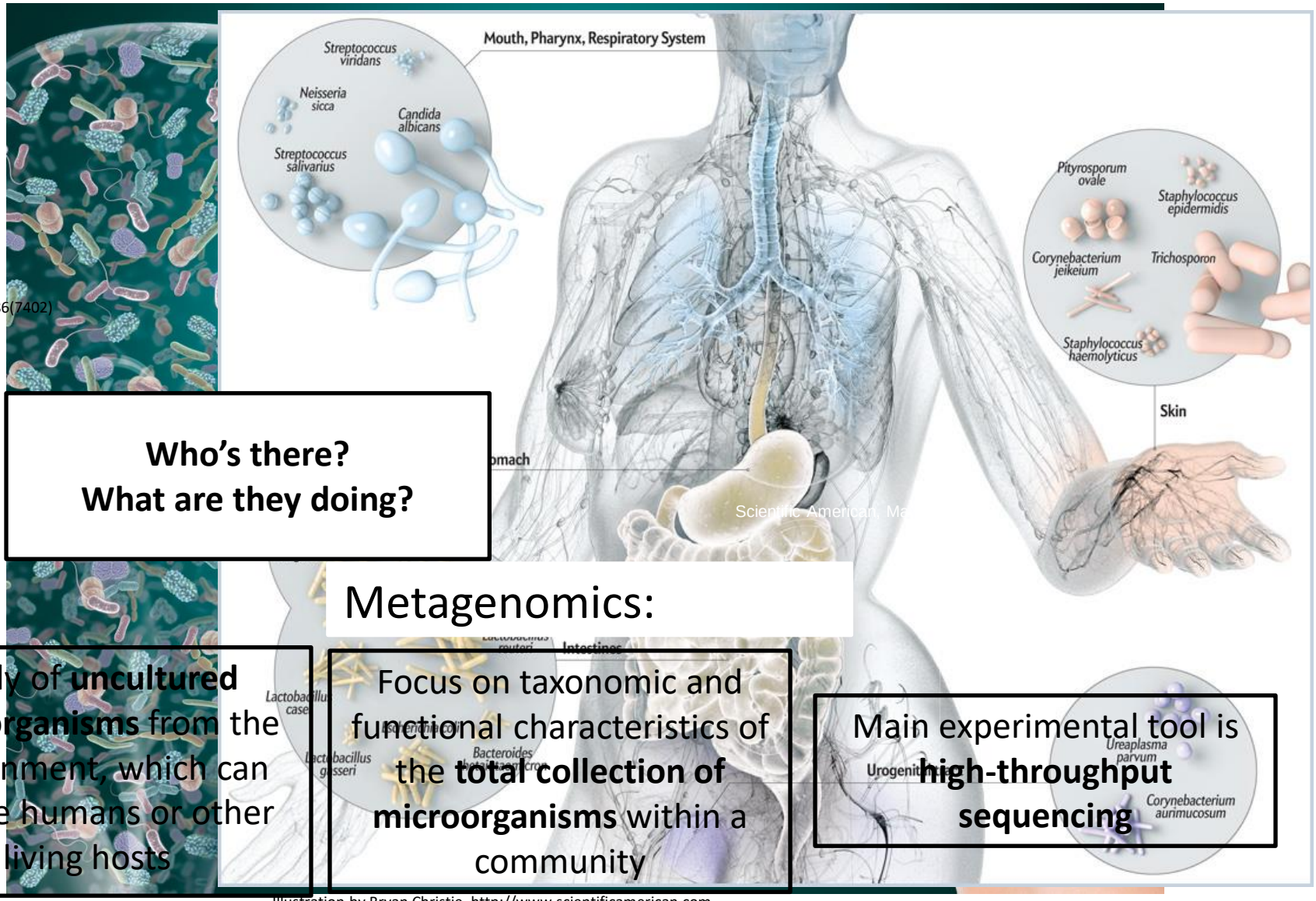
July 19, 2021

Funding: NCI 5R01 CA230551

CUNY SPH GRADUATE SCHOOL OF PUBLIC HEALTH & HEALTH POLICY

ISPH CUNY Institute for Implementation Science in Population Health

Microbial communities: in, on, outside us



Nature 486(7402)

Who's there?
What are they doing?

Metagenomics:

Study of **uncultured microorganisms** from the environment, which can include humans or other living hosts

Focus on taxonomic and functional characteristics of the **total collection of microorganisms** within a community

Main experimental tool is **high-throughput sequencing**

Illustration by Bryan Christie, http://www.scientificamerican.com

Roles of host-associated microbiota

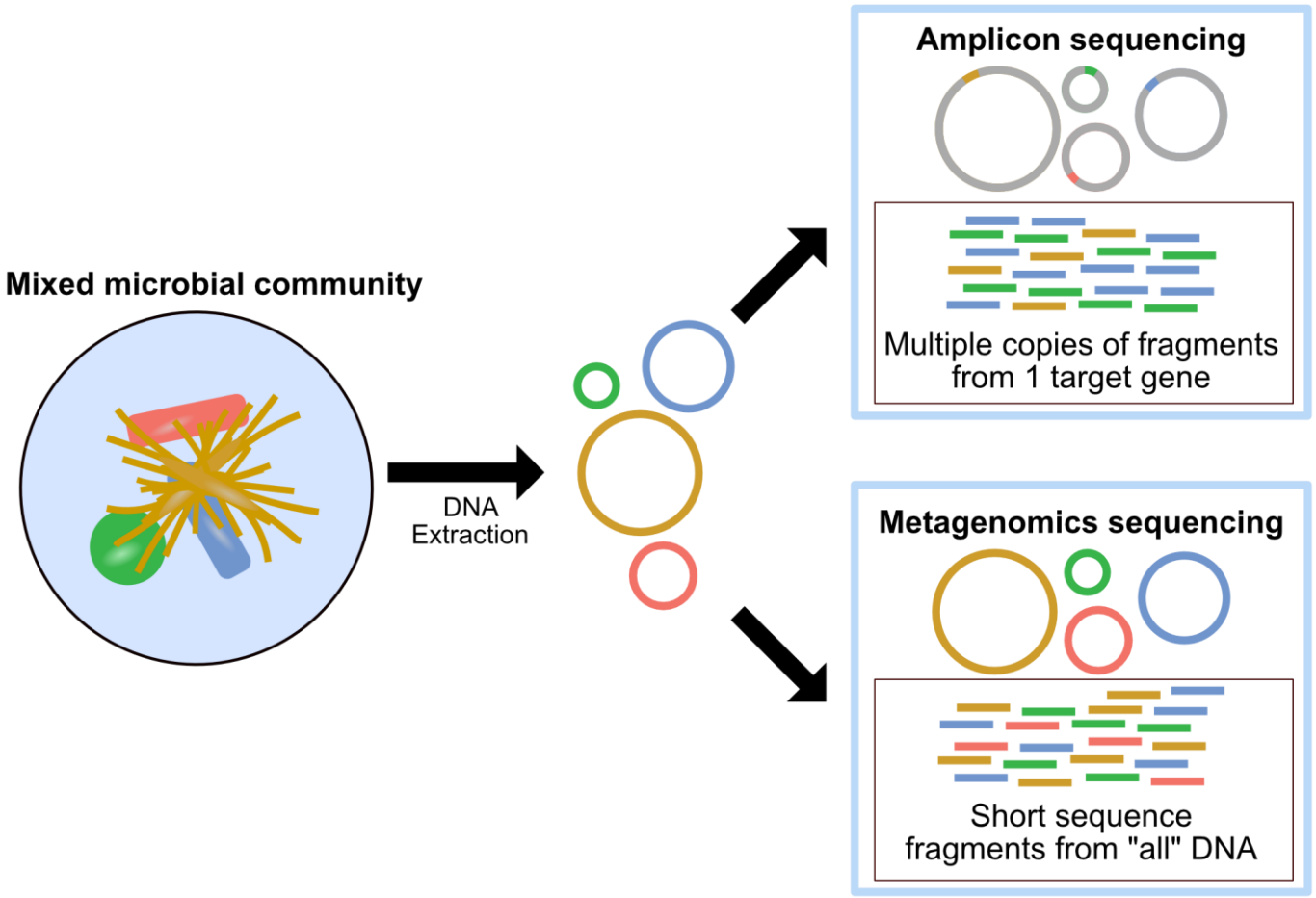
Synthesize	vitamins K and the B complex
Degrade	amino acids
Modify	bile acids and sterols
Hydrogenate	polyunsaturated fatty acids.
Compete	with potential pathogens
Facilitate	normal immune system function

Coates ME: **Gnotobiotic animals in research: their uses and limitations**. *Lab. Anim.* 1975, **9**:275–282.

Fecal Microbiota Transplantation establishes some roles in health

Antibiotic-resistant *C. difficile* infection	**~90% effective for treating multiple antibiotics-resistant infection** Kassam *et al.* Fecal microbiota transplantation for Clostridium difficile infection: systematic review and meta-analysis. Am. J. Gastroenterol. 108, 500–508 (2013)
Obesity	**Fecal Transplantation from lean and obese humans produce lean and obese mice** Ridaura et al. **Gut microbiota from twins discordant for obesity modulate metabolism in mice**. Science 341, 1241214 (2013)
Resistance to immunotherapy for cancer	Davar *et al.* **Fecal microbiota transplant overcomes resistance to anti-PD-1 therapy in melanoma patients**. Science 371, 595–602 (2021)

Metagenomic sequencing

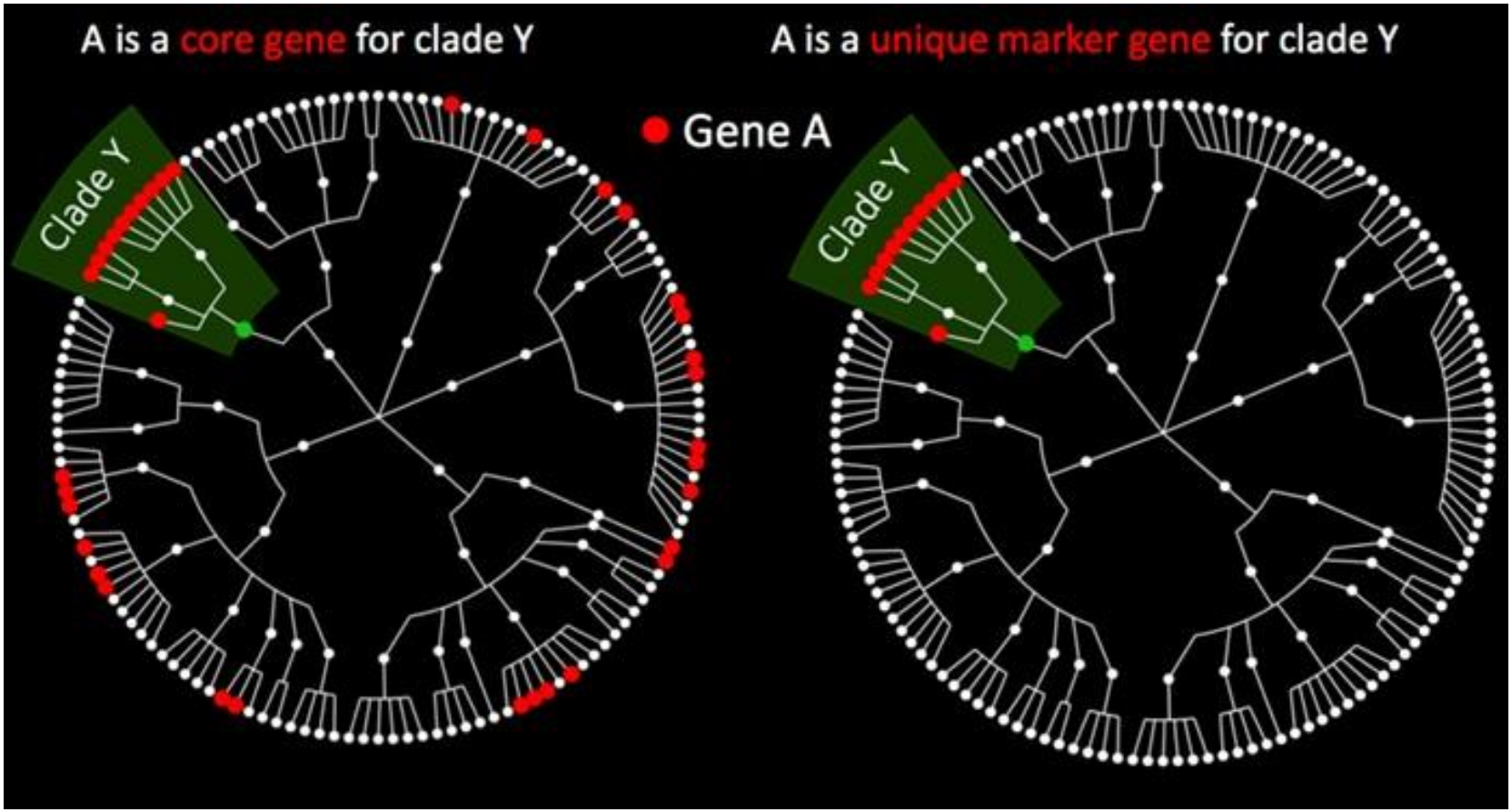


16S
18S
ITS

whole-metagenome "shotgun"

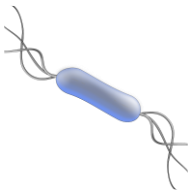
Source: Happy Belly Bioinformatics https://astrobiomike.github.io/misc/amplicon_and_metagen

Unique marker genes for taxonomic classification from shotgun metagenomic sequencing

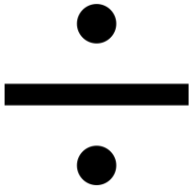


Estimation of taxonomic relative abundance

Map DNA sequence reads to marker genes



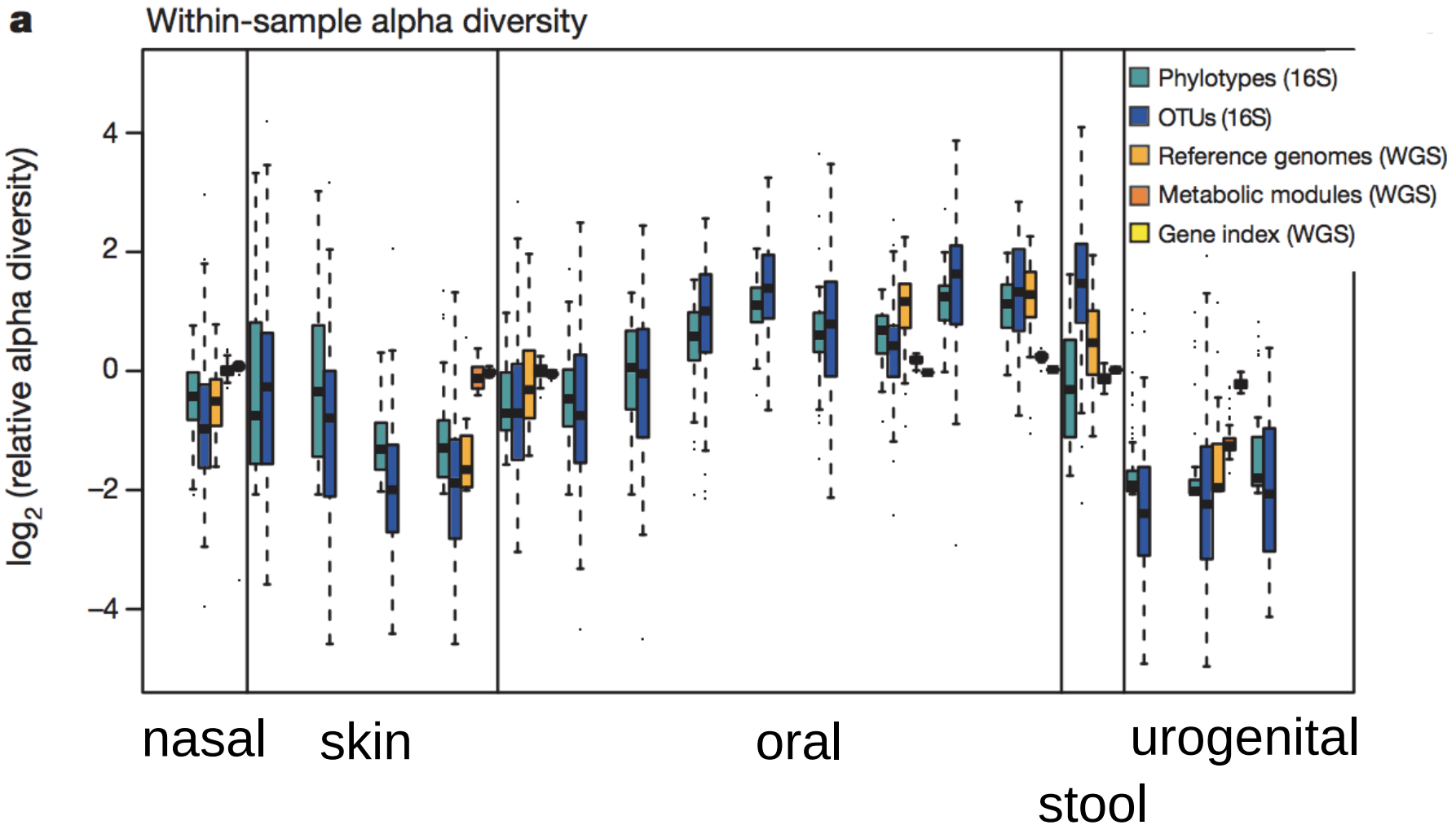
Count sequence reads per species using reference database



Normalize by the number and length of marker genes per species

- Result is a *relative abundance* (%) estimate for each species
- Multiplying by total number of sequence reads can give very approximate number of reads associated with that species

Properties of the data: body site-specific diversity, high individuality



C Huttenhower *et al. Nature* **486**, 207-214 (2012) doi:10.1038/nature11234

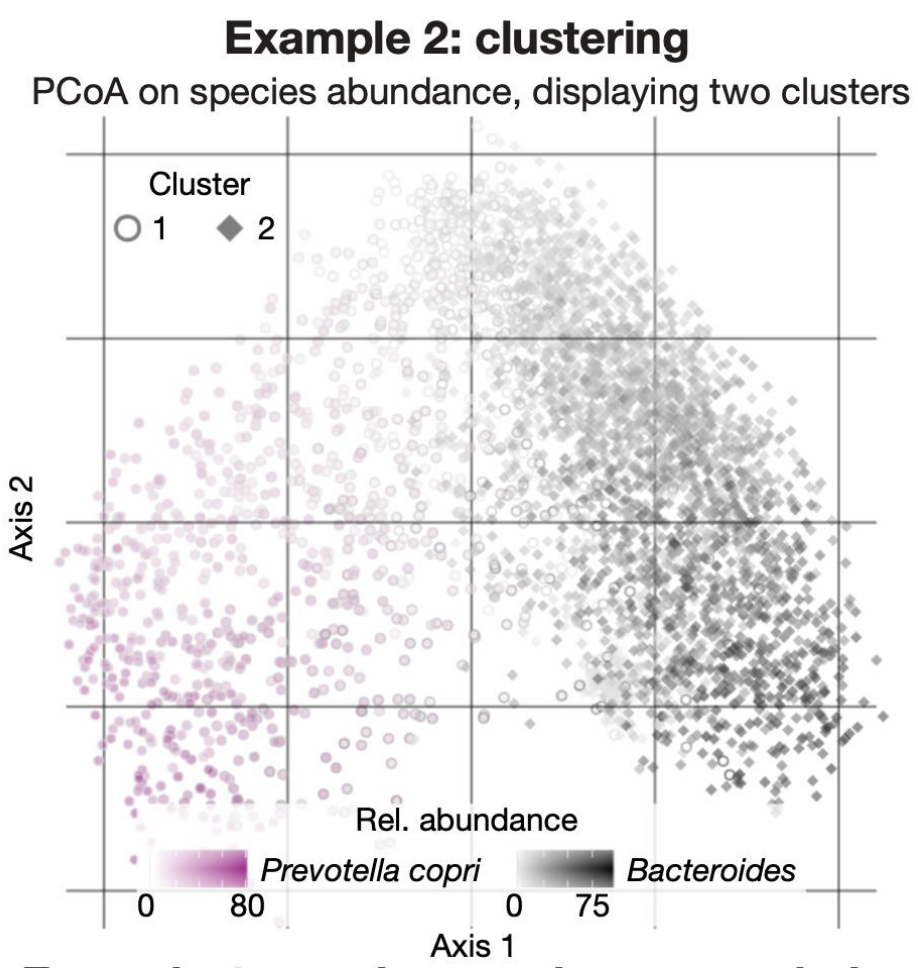
Properties of the data: distribution

- Between 0 and 100%
- Negative Binomial or Zero-inflated Negative Binomial distribution fit well (When multiplied by # sequence reads) (Calgaro *et al* 2020)
  - Some variables have low mean therefore mostly zeroes

[1] Calgaro M, Romualdi C, Waldron L, Risso D, Vitulo N. Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. Genome Biol. 2020;21:191. doi:10.1186/s13059-020-02104-1.

Properties of the data: compositionality

- Relative abundance between 0 and 100%
- “Horseshoe” effect in PCA



Pasolli E, Schiffer L, *et al*. **Accessible, curated metagenomic data through ExperimentHub**. Nat Methods. 2017;14:1023–4. doi:10.1038/nmeth.4468.

Properties of the data: compositionality

- Relative abundance between 0 and 100%
- Popular methodological area with statisticians

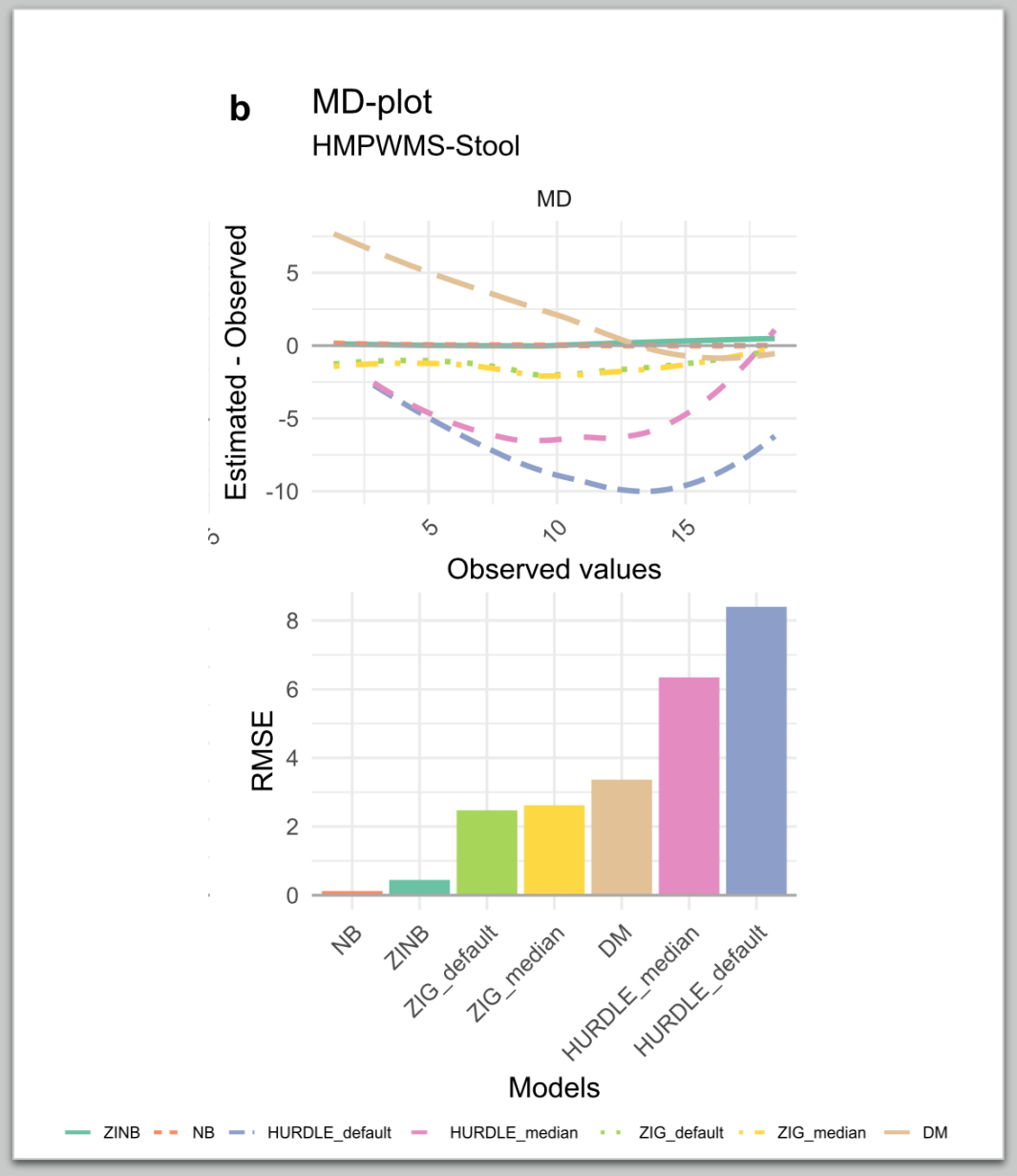
"Because the relative abundance of taxa sum to 1, it is not appropriate to use standard statistical methods such as the t-test, ANOVA, and so on directly on the relative abundances, because the standard methods implicitly assume that there are no such restrictions on the data (9)."

Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD. **Analysis of composition of microbiomes: a novel method for studying microbial composition**. Microb Ecol Health Dis. 2015;26:27663. doi:10.3402/mehd.v26.27663.

(9) Aitchison J. **The statistical analysis of compositional data**. J R Stat Soc Series B (Methodological) 1982; 44: 139□77.

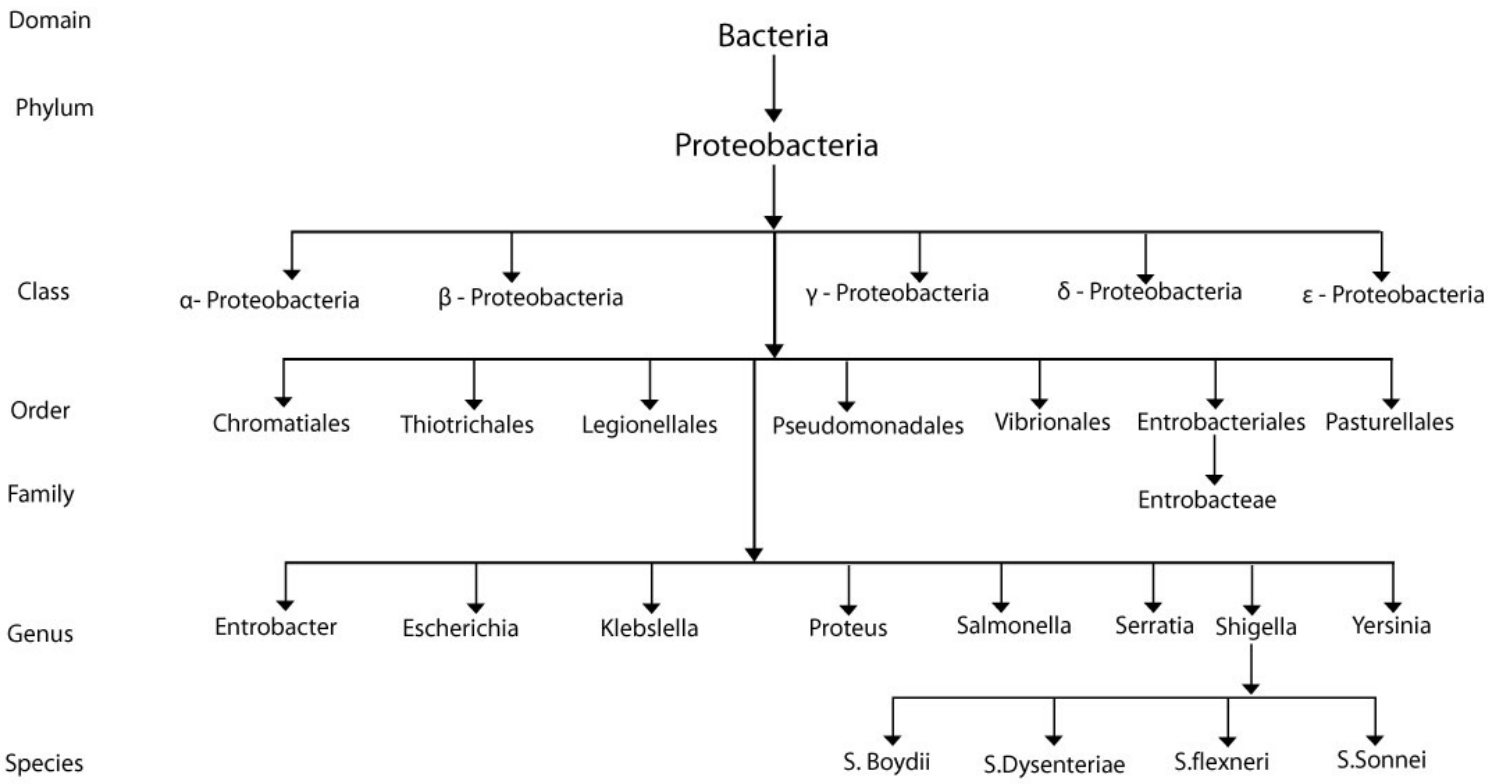
Properties of the data: compositionality

- Relative abundance between 0 and 100%
- Other statisticians find non-compositional methods work well anyways



Calgaro M, Romualdi C, Waldron L, Risso D, Vitulo N. **Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data.** Genome Biol. 2020;21:191. doi:10.1186/s13059-020-02104-1.

Properties of the data: taxonomical hierarchy



Cirrhosis dataset

```
[1] "Next study condition: cirrhosis  /// Body site:  stool"
> se
class: TreeSummarizedExperiment
dim: 693 282
metadata(0):
assays(1): relative_abundance
rownames(693):
  k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Bacteroidaceae|g__Bacteroides|s__Bacteroides_vulgatus
  k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Bacteroidaceae|g__Bacteroides|s__Bacteroides_uniformis
  ...
  k__Bacteria|p__Firmicutes|c__Bacilli|o__Bacillales|f__Staphylococcaceae|g__Staphylococcus|s__Staphylococcus_argenteus
  k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bacteroidales|f__Prevotellaceae|g__Prevotella|s__Prevotella_marshii
rowData names(7): Kingdom Phylum ... Genus Species
colnames(282): SID0002_gti SID0004_cle ... LV-8 LV-9
colData names(31): study_name subject_id ... inr ctp
reducedDimNames(0):
mainExpName: NULL
altExpNames(0):
rowLinks: a LinkDataFrame (693 rows)
rowTree: 1 phylo tree(s) (10430 leaves)
colLinks: NULL
colTree: NULL
>
```

https://github.com/waldronlab/curatedMetagenomicAnalyses
vignettes/MLdatasets.Rmd

```
> View(data.frame(rowData(se)))
```

MLdatasets.Rmd | data.frame(rowData(se))

Filter

	Kingdom	Phylum	Class	Order	Family	Genus	Species
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bact...	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	Bacteroides vulgatus
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bact...	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	Bacteroides uniformis
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bact...	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	Bacteroides plebeius
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostridi...	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Lachnospiraceae ...	Eubacterium rectale
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bact...	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes	Alistipes finegoldii
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bact...	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	Bacteroides caccae
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bact...	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Tannerellaceae	Parabacteroides	Parabacteroides distasonis
k__Bacteria|p__Firmicutes|c__Bacilli|o__Lactobacillal...	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	Streptococcus salivarius

Showing 1 to 7 of 693 entries, 7 total columns

```
> View(assay(se))
```

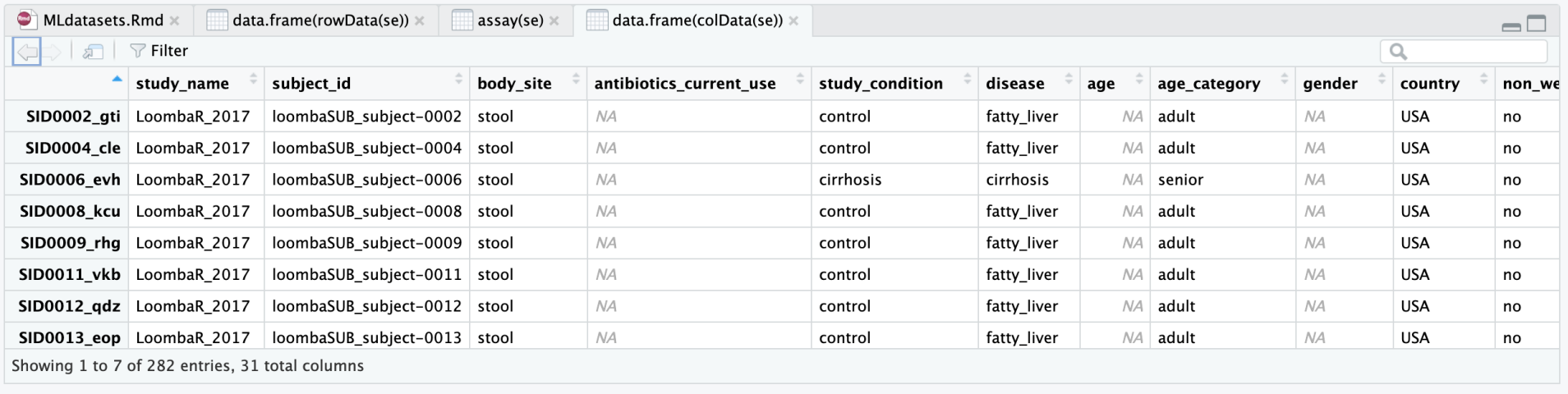
MLdatasets.Rmd | data.frame(rowData(se)) | assay(se)

Filter | Cols: 0 – 50

	SID0002_gti	SID0004_cle	SID0006_evh	SID0008_kcu	SID0009_rhg	SID0011_vkb	SID0012_qdz	SID0013_eop
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bac...	16.29859	14.89908	20.18926	0.31815	20.13217	2.78693	3.83645	0.86755
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bac...	10.40806	19.41034	2.92210	0.27528	4.07602	3.44081	4.37935	0.27649
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bac...	9.71026	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
k__Bacteria|p__Firmicutes|c__Clostridia|o__Clostrid...	8.05587	0.02132	2.96656	16.82329	0.12402	3.24912	4.82069	20.88263
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bac...	6.78360	0.00000	0.00000	0.01673	0.00000	0.00000	0.00000	0.00374
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bac...	6.53608	0.00000	2.53602	0.00000	0.00000	0.00000	0.54046	0.01842
k__Bacteria|p__Bacteroidetes|c__Bacteroidia|o__Bac...	5.59468	5.35510	2.10484	1.05405	12.02912	0.00000	3.85517	0.00000
k__Bacteria|p__Firmicutes|c__Bacilli|o__Lactobacill...	3.92924	0.00663	3.04440	1.17655	0.01676	0.07167	0.00000	0.61688

Showing 1 to 7 of 693 entries, 282 total columns

> View(data.frame(rowData(se)))


MLdatasets.Rmd | data.frame(rowData(se)) | assay(se) | data.frame(colData(se))

Filter

	study_name	subject_id	body_site	antibiotics_current_use	study_condition	disease	age	age_category	gender	country	non_we
SID0002_gti	LoombaR_2017	loombaSUB_subject-0002	stool	NA	control	fatty_liver	NA	adult	NA	USA	no
SID0004_cle	LoombaR_2017	loombaSUB_subject-0004	stool	NA	control	fatty_liver	NA	adult	NA	USA	no
SID0006_evh	LoombaR_2017	loombaSUB_subject-0006	stool	NA	cirrhosis	cirrhosis	NA	senior	NA	USA	no
SID0008_kcu	LoombaR_2017	loombaSUB_subject-0008	stool	NA	control	fatty_liver	NA	adult	NA	USA	no
SID0009_rhg	LoombaR_2017	loombaSUB_subject-0009	stool	NA	control	fatty_liver	NA	adult	NA	USA	no
SID0011_vkb	LoombaR_2017	loombaSUB_subject-0011	stool	NA	control	fatty_liver	NA	adult	NA	USA	no
SID0012_qdz	LoombaR_2017	loombaSUB_subject-0012	stool	NA	control	fatty_liver	NA	adult	NA	USA	no
SID0013_eop	LoombaR_2017	loombaSUB_subject-0013	stool	NA	control	fatty_liver	NA	adult	NA	USA	no

Showing 1 to 7 of 282 entries, 31 total columns

rowTree(se)

```
> rowTree(se)

Phylogenetic tree with 10430 tips and 10429 internal nodes.

Tip labels:
  k__Archaea|p__Candidatus_Micrarchaeota|c__Candidatus_Micrarchaeota_unclassified|o__Candidatus_Micrarchaeota_unclassified|f__Candidatus_Micrarchaeota_unclassified|g__Candidatus_Micrarchaeota_unclassified|s__Candidatus_Micrarchaeota_archaeon_CG1_02_55_22, k__Archaea|p__Archaea_unclassified|c__Archaea_unclassified|o__Archaea_unclassified|f__Archaea_unclassified|g__Archaea_unclassified|s__archaeon_GW2011_AR15, k__Archaea|p__Candidatus_Diapherotrites|c__Candidatus_Diapherotrites_unclassified|o__Candidatus_Diapherotrites_unclassified|f__Candidatus_Diapherotrites_unclassified|g__Candidatus_Diapherotrites_unclassified|s__Candidatus_Diapherotrites_archaeon_CG08_land_8_20_14_0_20_34_12, k__Archaea|p__Archaea_unclassified|c__Archaea_unclassified|o__Archaea_unclassified|f__Archaea_unclassified|g__Archaea_unclassified|s__archaeon_GW2011_AR10, k__Archaea|p__Candidatus_Diapherotrites|c__Candidatus_Diapherotrites_unclassified|o__Candidatus_Diapherotrites_unclassified|f__Candidatus_Diapherotrites_unclassified|g__Candidatus_Diapherotrites_unclassified|s__Candidatus_Diapherotrites_archaeon_CG11_big_fil_rev_8_21_14_0_20_37_9, k__Archaea|p__Euryarchaeota|c__Euryarchaeota_unclassified|o__Euryarchaeota_unclassified|f__Euryarchaeota_unclassified|g__Euryarchaeota_unclassified|s__Euryarchaeota_archaeon_TMED173, ...

Rooted; includes branch lengths.
```

Introducing a project idea

BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 21 2014, pages 3062–3069
doi:10.1093/bioinformatics/btu488

Gene expression

Advance Access publication July 23, 2014

Más-o-menos: a simple sign averaging method for discrimination in genomic data analysis

Sihai Dave Zhao[1,*], Giovanni Parmigiani[2,3], Curtis Huttenhower[2] and Levi Waldron[4]

[1]Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, [2]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, [3]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02115, [4]City University of New York School of Public Health, Hunter College, New York, NY 10035, USA

Associate Editor: Ziv Bar-Joseph

(1) Standardize the covariates such that $(n-1)^{-1}\sum(X_{ij}-\overline{X}_j)^2=1, j=1,\ldots,p$, where $\overline{X}_j = n^{-1}\sum_i X_{ij}$.

(2) Perform univariate regressions of the outcome on each gene to obtain marginal estimates of the regression coefficient $\hat{\alpha}_j$.

(3) Let $\hat{v}_j = \mathrm{sgn}(\hat{\alpha}_j)/p^{1/2}$, where $\mathrm{sgn}(c) = 2I(c>0) - 1$ for $c \neq 0$ and $\mathrm{sgn}(c)=0$ for $c = 0$.

(4) The risk score for the i^{th} patient is calculated as $\mathbf{X}_i^T\hat{\mathbf{v}}$, where $\hat{\mathbf{v}} = (\hat{v}_1,\ldots,\hat{v}_p)^T$.

Mas-o-menos in simulated data

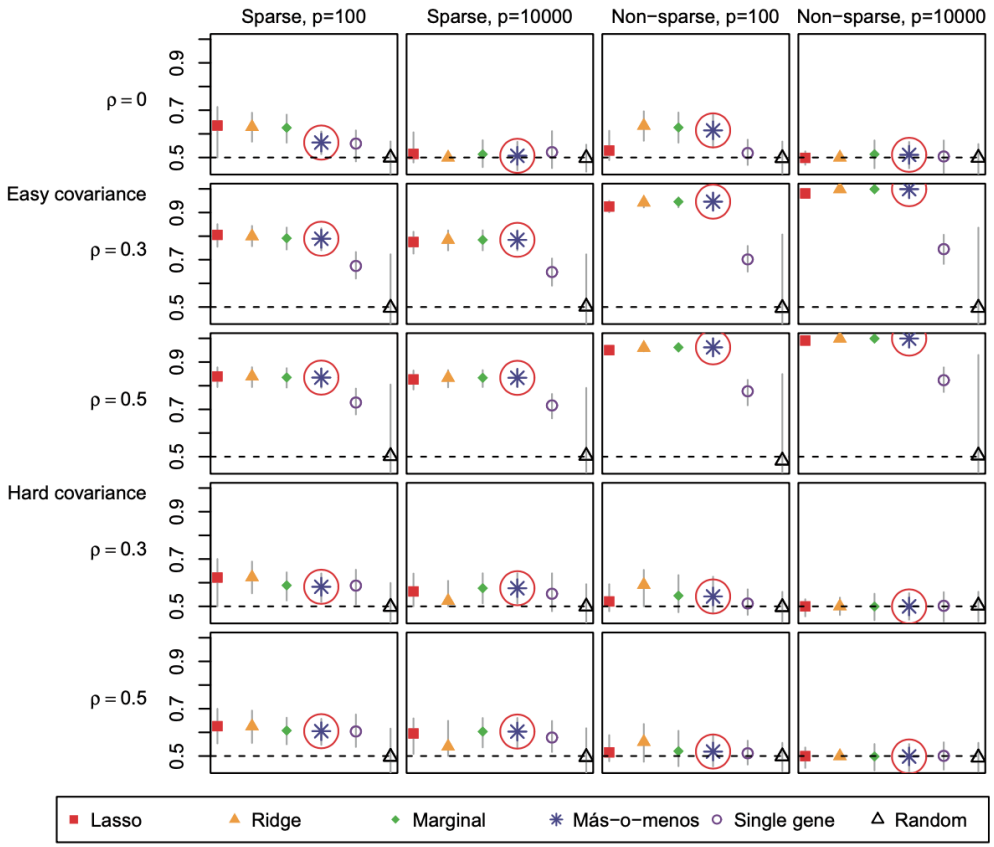
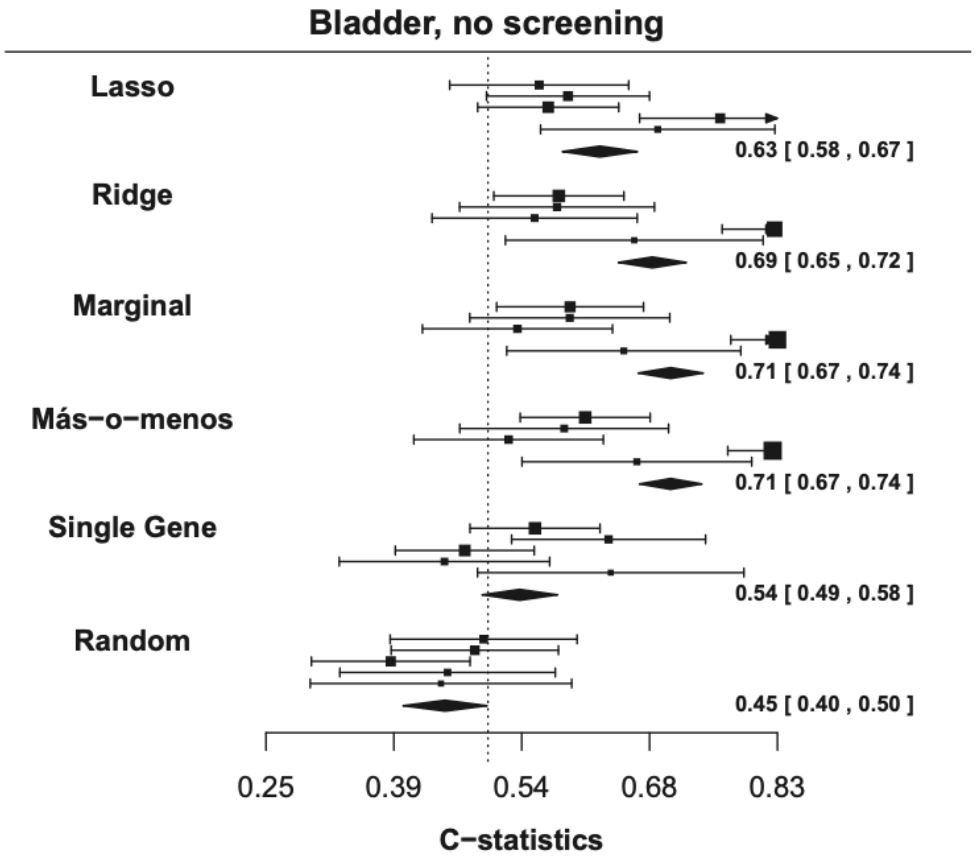
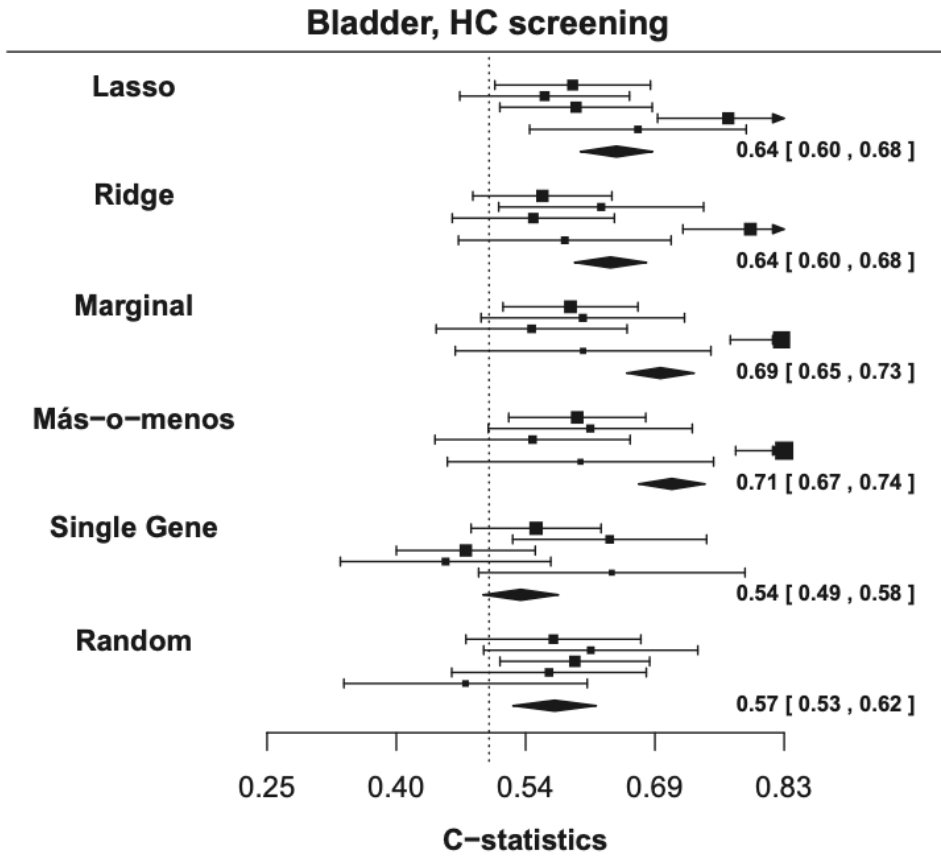


Fig. 1. Average validation C-statistics of different discrimination methods in simulated data. Más-o-menos results highlighted by circle. Vertical bars represent confidence intervals

Mas-o-menos in real data





Also performed well in cross-validation and in ovarian, breast cancer

Current mas-o-menos use

- Goes by many other names
- Popular with non-statisticians
  - Easy to train, easy to explain, easy to apply
- More popular with statisticians when forced to transfer a model to different technologies or to semi-related data
  - e.g. gene expression to protein presence
  - Any time where scale of variables is clearly meaningless

Machine learning in metagenomics

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights

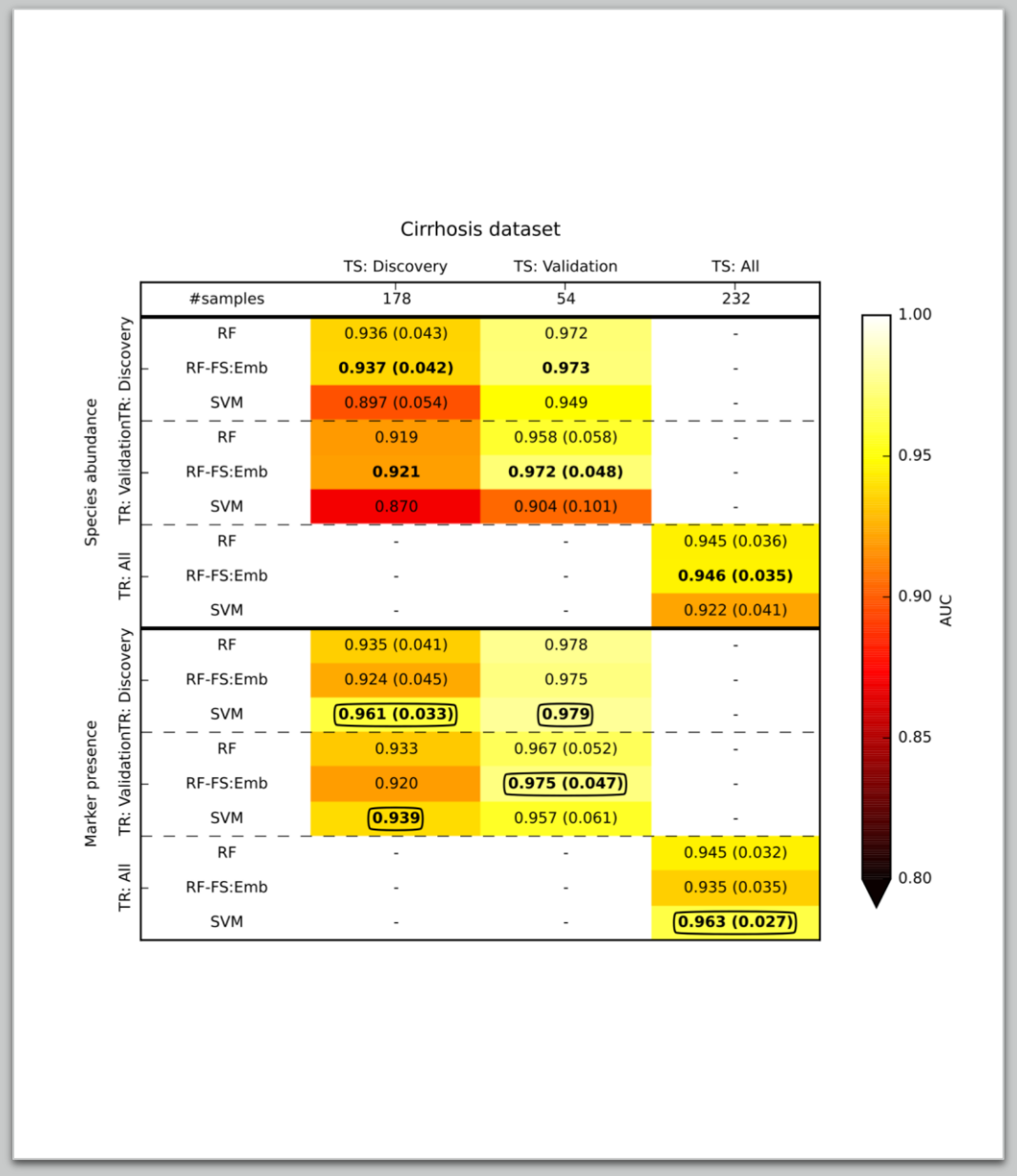
Edoardo Pasolli[1], Duy Tin Truong[1], Faizan Malik[2], Levi Waldron[2], Nicola Segata[1]*

1 Centre for Integrative Biology, University of Trento, Trento, Italy, **2** Graduate School of Public Health and Health Policy, City University of New York, New York, New York, United States of America

Pasolli E, Truong DT, Malik F, Waldron L, Segata N (2016) Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. PLoS Comput Biol 12(7): e1004977. doi:10.1371/journal.pcbi.1004977

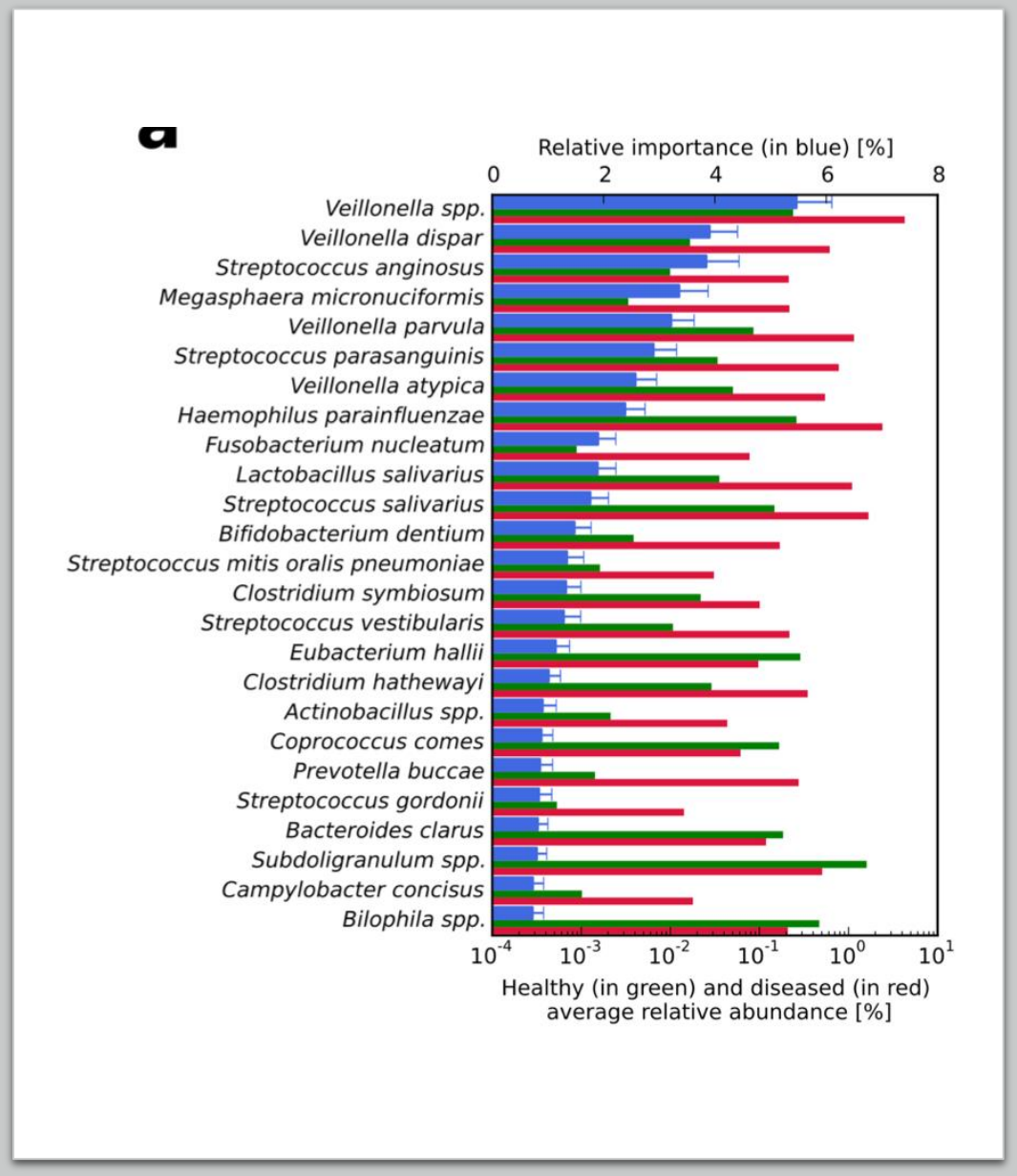
Main findings

- Random Forest (RF) and SVM out-performed Elastic Net and Lasso
- Cross-study validation is harder than cross-validation (no surprise)
- Warnings about over-fitting in feature selection + cross-validation



Main findings

- RF provides useful ”relative importance” measures of features
- RF classifiers have come to dominate machine learning for microbiome studies



Questions

- How does a simple mas-o-menos classifier compare to Random Forest?
  - With and without feature screening
- Do the Pasolli *et al.* results hold up in your expanded datasets?

Resources

Machine learning packages in R

1. Kuhn M. **caret: Classification and Regression Training. Astrophysics Source Code Library**. 2015;:ascl:1505.003.
2. Wirbel J, Zych K, Essex M, Karcher N, Kartal E, Salazar G, et al. **Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox**. Genome Biol. 2021;22:93. doi:10.1186/s13059-021-02306-1.
3. **MLinterfaces**: http://bioconductor.org/packages/MLInterfaces/

Data management

- Bioconductor SummarizedExperiment package: http://www.bioconductor.org/packages/SummarizedExperiment/
- Bioconductor biobroom package: http://www.bioconductor.org/packages/biobroom/