

Microbial community analysis in R

CascadiaR Conference

June 3, 2017

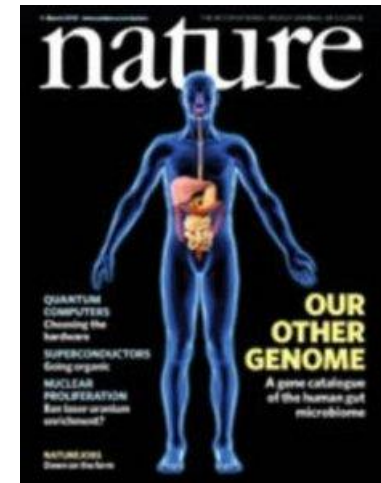
Lisa Karstens, PhD

Microbes play an important role in human health

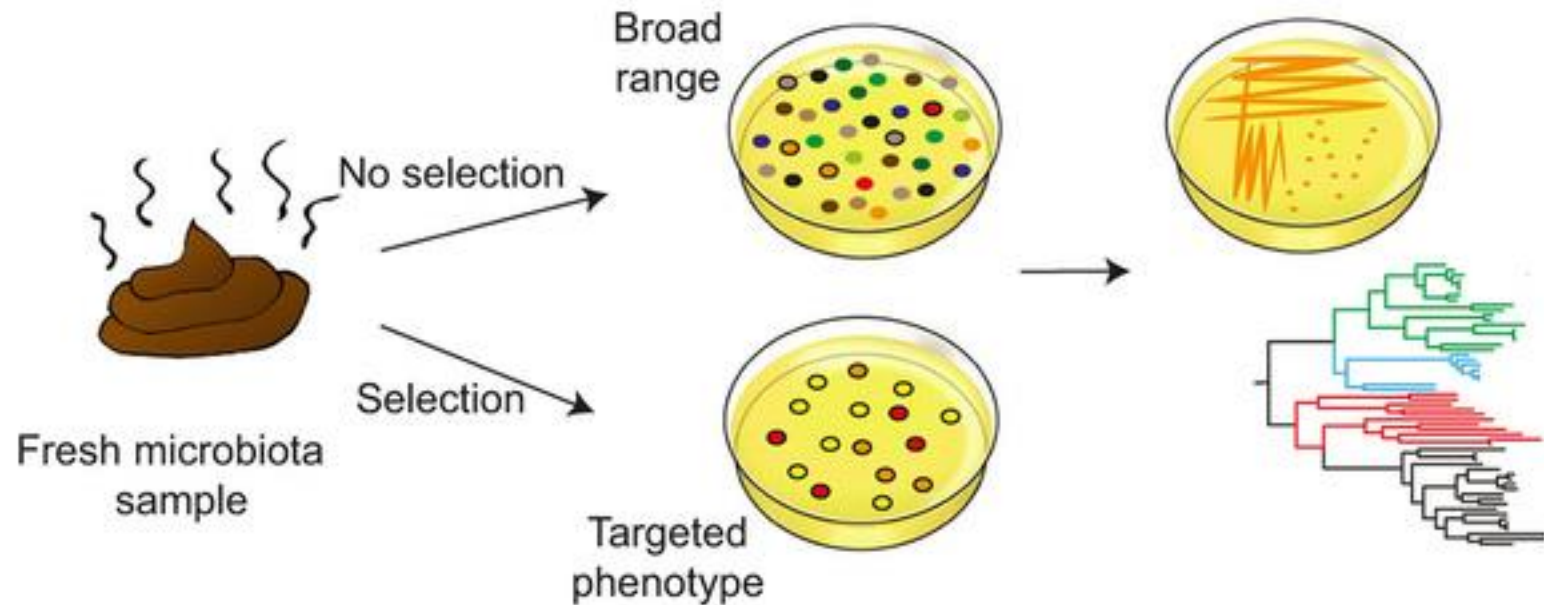
We are not alone...

humans are supraorganisms

- Symbiotic relationship with microbes
- Important for health and disease



Early methods microbial investigation relied on culturing techniques



High-throughput sequencing technology has brought huge growth in microbiome investigations.

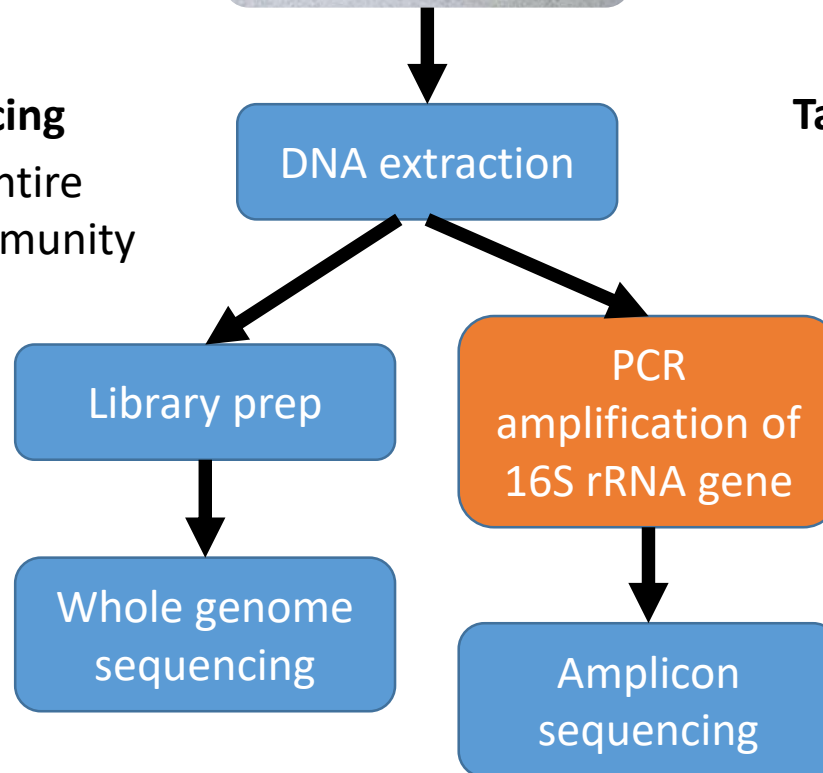


Metagenome sequencing

Metagenome – entire genome of a community

Targeted gene sequencing

16S rRNA gene – marker gene for prokaryotes



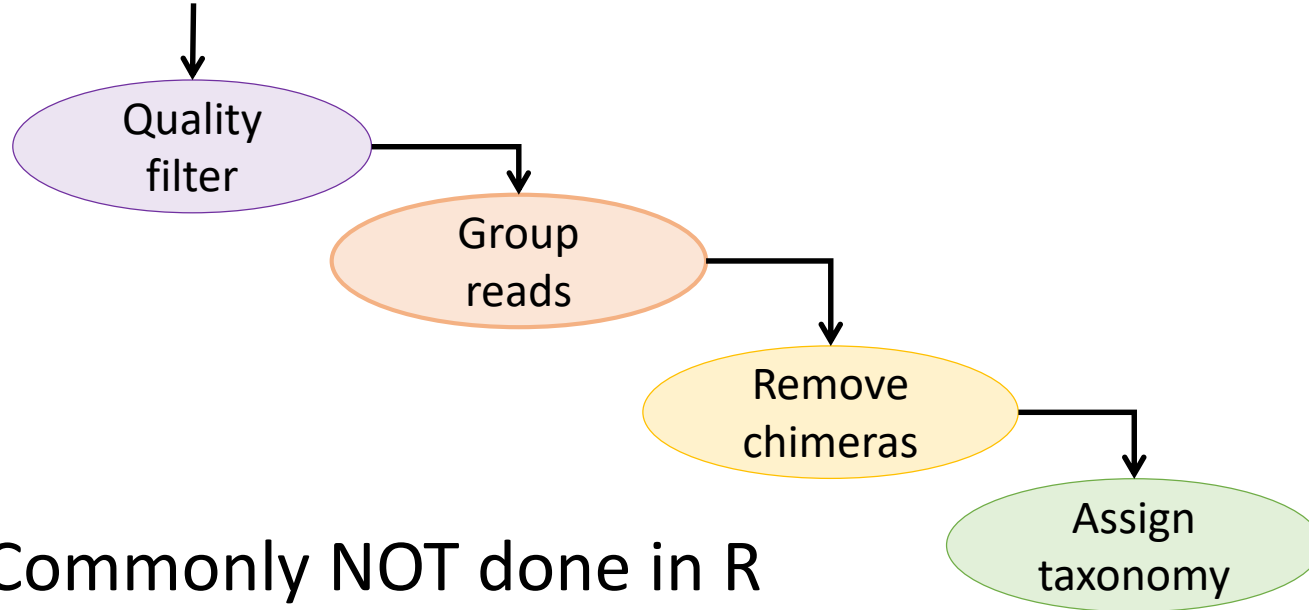
16S sequencing studies generate a lot of data!

One sequencing run can have over 1 million reads from hundreds of samples!



Example workflow processing sequence reads

```
GGACGGCAAGCTGGACGGCAACTTTCCA
GGACGGCAAGCTGGACGGCAACTTTCCA
GGACGGCAAGCTGGACGGCAACTTTCCA
CTGGAATGATCTGGAATGAAGGGTTCCA
CTGGAATGATCTGGAATGAAGGGTTCCA
```



- Commonly NOT done in R
- Each step is an active area of research
- Many options exist

	S1	S2	S...
Bacteria 1	4	0	2
Bacteria 2	43	49	24
Bacteria 3	56	65	43
...			

Data analysis

R offers an ideal environment for processing and analysis

- Reproducible, Organized, Sharable
 - R Studio
 - Interactive, friendly environment
 - R Markdown
 - Documents processing steps
 - RData files
 - Sequence data, and sample data, results in one file

Data table

	S1	S2	S...
Bacteria 1	4	55	78
Bacteria 2	50	32	32
Bacteria 3	20	2	4
...			

Sample data table

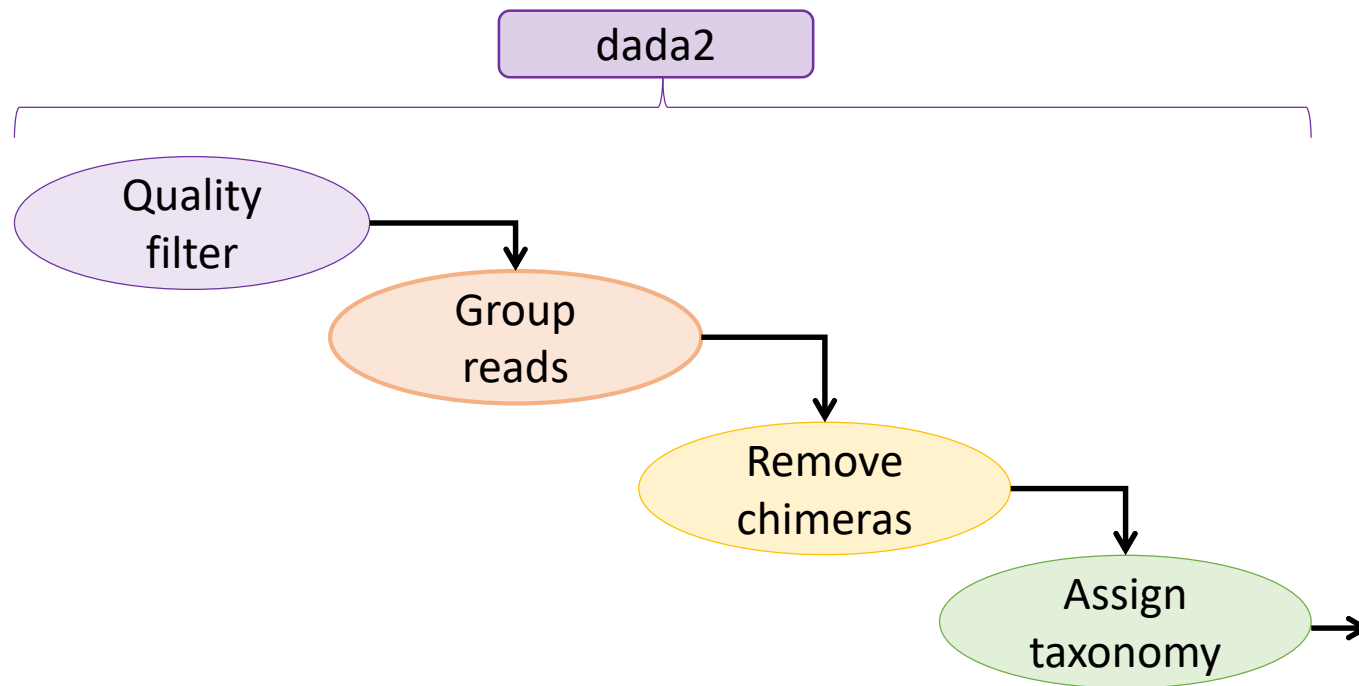
	S1	S2	S...
Diagnosis	D	C	D
BMI	28	25	23
Genotype	0	1	0
...			

Results table

	Change	p
Bacteria 1	2.5	.005
Bacteria 2	5	.05
Bacteria 3	1.3	.96
...		

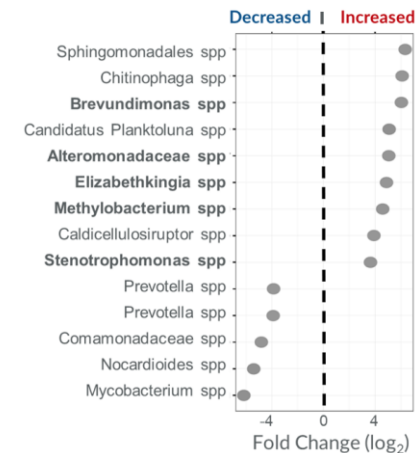
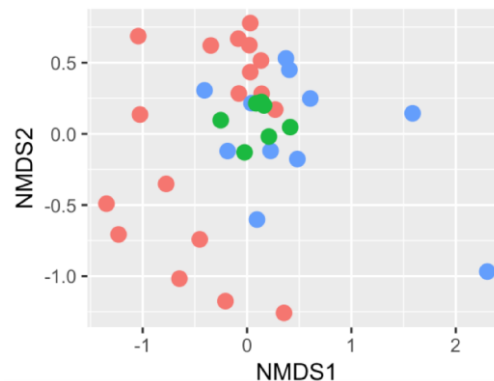
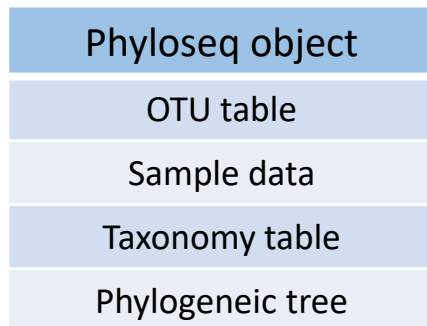
Microbiome specific workflows in R

- **dada2**
 - A package to model and correct 16S sequence errors
 - Replaces clustering algorithms for grouping reads



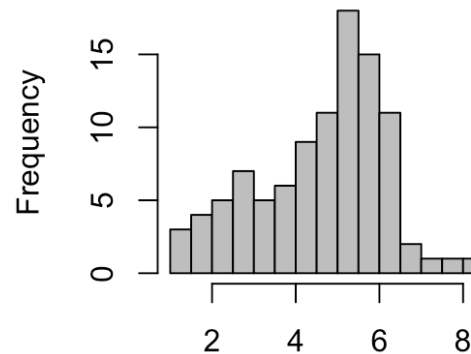
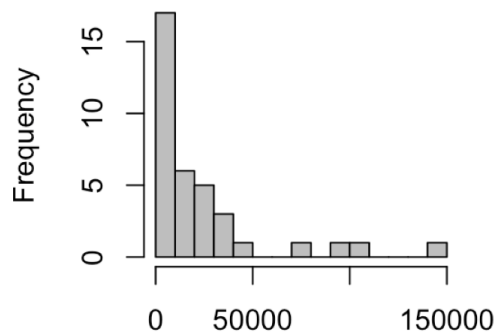
Microbiome specific workflows in R

- Phyloseq
 - Tool to import, store, analyze, and graphically display complex phylogenetic sequencing data
 - Many excellent tutorials available online



R offers an ideal environment for processing and analysis

- Exploratory Data Analysis
 - Data structure
 - Understand variables
 - Identify (and fix) potential problems
 - Decision making
 - Identify and evaluate transformations, filtering steps
 - Identify data distributions for appropriate analysis



R offers flexibility for data analysis



- Hypothesis testing
 - DeSEQ2
 - MetagenomeSeq
- Multivariate methods
 - Vegan
 - Mixomics

Challenges in microbiome workflows in R



- Requires user to know how to use R
- Appropriate application of packages can be challenging
- Storing raw sequencing data in R is not ideal
- Common processing algorithms are not available



Thank You

Acknowledgements

- Students:
 - Vincent Caruso
 - Eric Leung
 - Mark Klick
- Mentors and Collaborators
 - Shannon McWeeney, PhD
 - Jim Rosenbaum, MD
 - Mark Asquith, PhD
 - Damien A. Fair, PA-C, Ph.D
 - Rahel Nardos, MD, MCR
 - Tom Gregory, MD
 - Rosenbaum Lab
 - Fair Neuroimaging Lab

Funding



K12 HD043488

**Collins Medical
Trust**