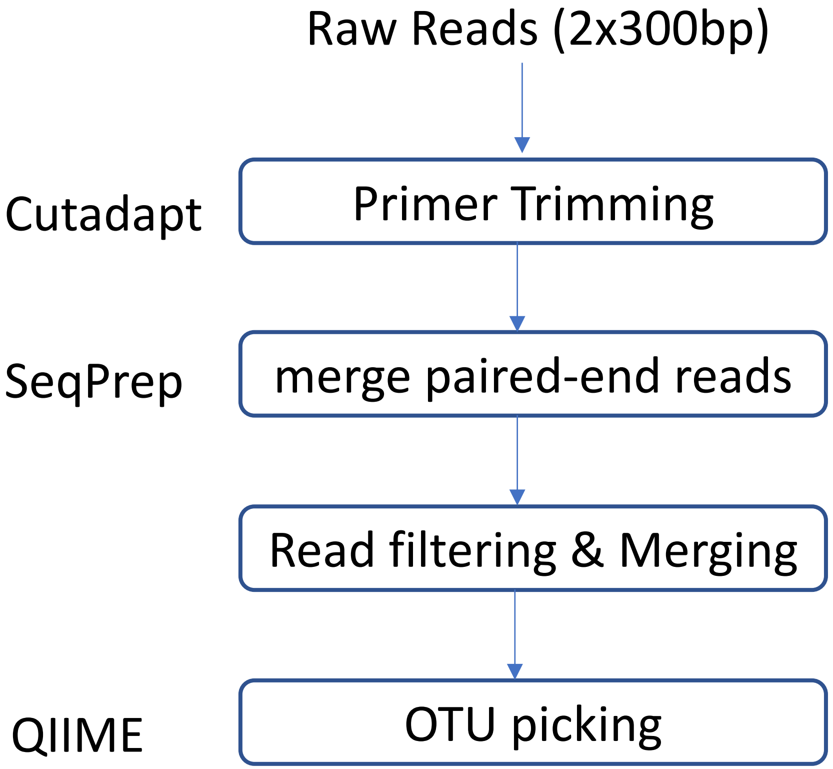
Microbiome pipeline

Di Wu

Pipeline for Microbiome Amplicon Sequencing



Raw Reads (2x300bp)

Cutadapt — Primer Trimming

SeqPrep — merge paired-end reads

Read filtering & Merging

QIIME — OTU picking

Microbiome pipeline

- Primer trimming

  Linked Adapter trimming (ADAPTER1…ADAPTER2)
  Keep reads containing both primers (trimmed_reads)

- Paired-end reads merging

  If R1 and R2 are overlapped, then merged into a longer read
  If not overlapped, only keep R1

- Length filtering and merging

  Discard reads <100bp
  Get one FASTA file for each target region

Microbiome pipeline

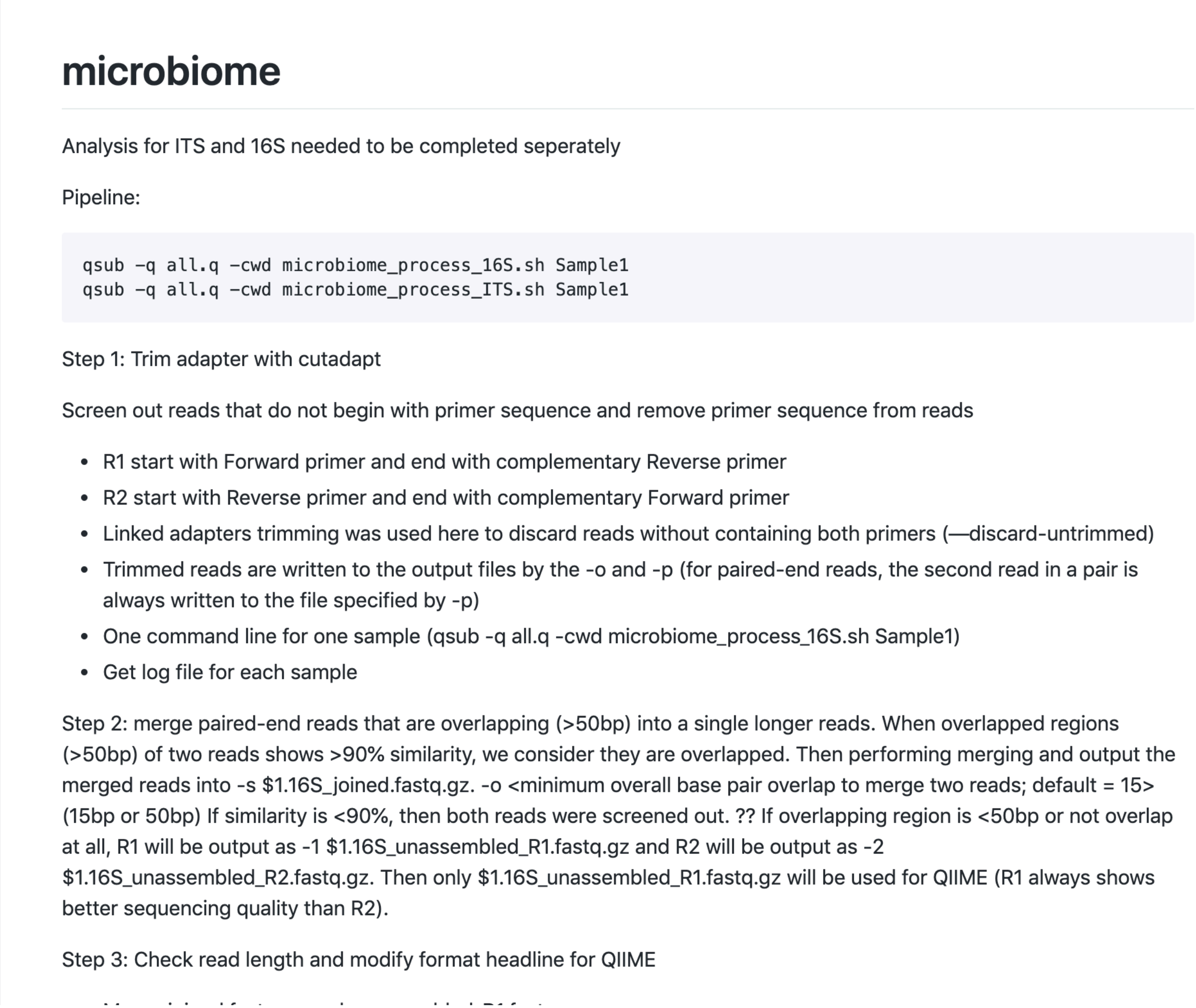
- OTU picking

  BLAST 16S.fasta/ITS.fasta against reference
  Generating BIOM (Biological Observation Matrix) table

- QC checking

  % mapped reads: >70% (16S)
                             >60% (ITS)
  aligned reads: >5000 per sample

Backup on Github



microbiome

Analysis for ITS and 16S needed to be completed seperately

Pipeline:

```
qsub -q all.q -cwd microbiome_process_16S.sh Sample1
qsub -q all.q -cwd microbiome_process_ITS.sh Sample1
```

Step 1: Trim adapter with cutadapt

Screen out reads that do not begin with primer sequence and remove primer sequence from reads

- R1 start with Forward primer and end with complementary Reverse primer
- R2 start with Reverse primer and end with complementary Forward primer
- Linked adapters trimming was used here to discard reads without containing both primers (—discard-untrimmed)
- Trimmed reads are written to the output files by the -o and -p (for paired-end reads, the second read in a pair is always written to the file specified by -p)
- One command line for one sample (qsub -q all.q -cwd microbiome_process_16S.sh Sample1)
- Get log file for each sample

Step 2: merge paired-end reads that are overlapping (>50bp) into a single longer reads. When overlapped regions (>50bp) of two reads shows >90% similarity, we consider they are overlapped. Then performing merging and output the merged reads into -s $1.16S_joined.fastq.gz. -o <minimum overall base pair overlap to merge two reads; default = 15> (15bp or 50bp) If similarity is <90%, then both reads were screened out. ?? If overlapping region is <50bp or not overlap at all, R1 will be output as -1 $1.16S_unassembled_R1.fastq.gz and R2 will be output as -2 $1.16S_unassembled_R2.fastq.gz. Then only $1.16S_unassembled_R1.fastq.gz will be used for QIIME (R1 always shows better sequencing quality than R2).

Step 3: Check read length and modify format headline for QIIME

- Merge joined fastq and unassembled R1 fast

Microbiome pipeline summary

- Deliverables

  FASTQ
  QC table (raw reads; reads with primers and %; assembled reads and %; mapped reads and %;)
  OTU table in both biom and txt formats

- Worked on real data (Shiao KK-6764—04—18—2019.xlsx)

Microbiome Introduction

Jie Tang

Introduction

- Microbiome (microbiota): the collective microorganisms that reside in our bodies

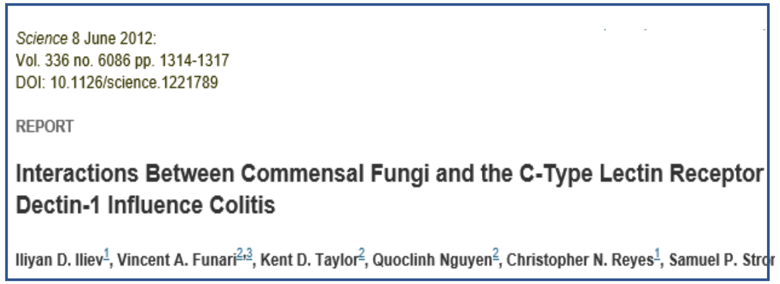
- Human symbiotic commensal microbiome comprises 100 trillion cells
  - ~1kg of adult body weight
  - >10 fold more cells than host cells
  - Carry ~150 fold more genes than host
  - Express >10 fold more unique genes than host

- The majority are found in the human gastrointestinal tract
  - Bacteria (>90% of total microbes)
  - Fungi (<10%)
  - Archaea (~1%)
  - Viruses (<2%)

Introduction

- Microbiome is plastic and contextual
  - Age
  - Diet
  - State of immune system
  - Antibiotic/Prebiotic/Xenobiotic

- Role of gut microbiome in diseases
  - Invading pathogen in infection
  - Cancer risk: stomach cancer, colon cancer
  - Autoimmune: rheumatoid arthritis, inflammatory bowel disease
  - Metabolism syndrome: type II diabetes
  - Neurodevelopmental disorder: autism

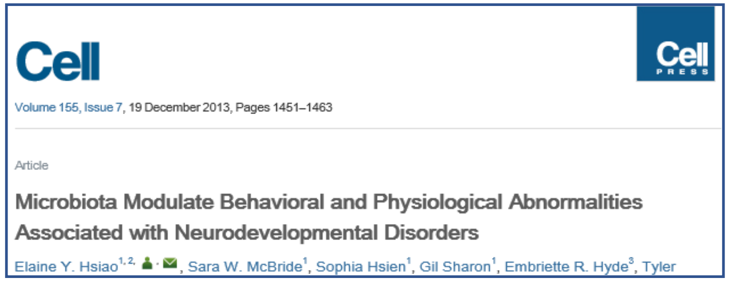


Science 8 June 2012:
Vol. 336 no. 6086 pp. 1314-1317
DOI: 10.1126/science.1221789

REPORT

Interactions Between Commensal Fungi and the C-Type Lectin Receptor Dectin-1 Influence Colitis

Iliyan D. Iliev[1], Vincent A. Funari[2,3], Kent D. Taylor[2], Quoclinh Nguyen[2], Christopher N. Reyes[1], Samuel P. Stro



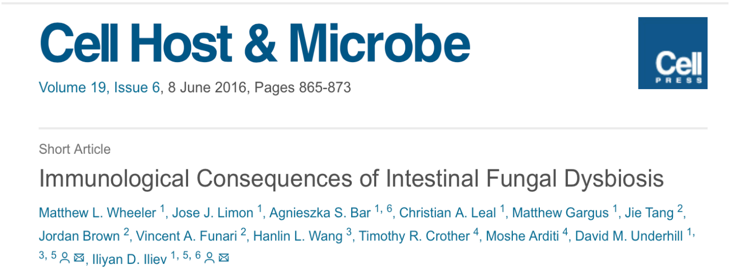
Cell

Volume 155, Issue 7, 19 December 2013, Pages 1451–1463

Article

Microbiota Modulate Behavioral and Physiological Abnormalities Associated with Neurodevelopmental Disorders

Elaine Y. Hsiao[1, 2,] 👤 · ✉, Sara W. McBride[1], Sophia Hsien[1], Gil Sharon[1], Embriette R. Hyde[3], Tyler



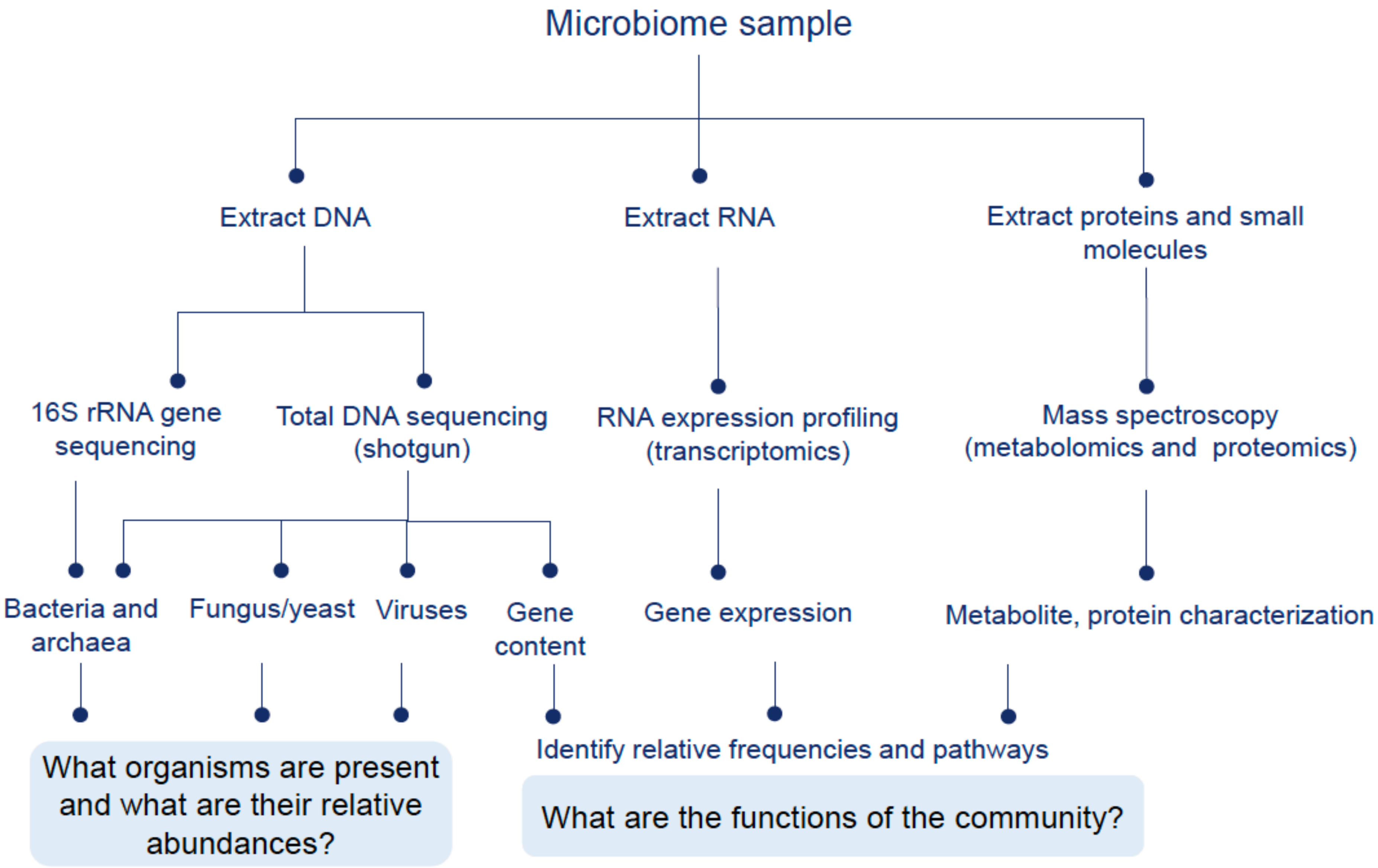
Cell Host & Microbe

Volume 19, Issue 6, 8 June 2016, Pages 865-873

Short Article

Immunological Consequences of Intestinal Fungal Dysbiosis

Matthew L. Wheeler [1], Jose J. Limon [1], Agnieszka S. Bar [1, 6], Christian A. Leal [1], Matthew Gargus [1], Jie Tang [2], Jordan Brown [2], Vincent A. Funari [2], Hanlin L. Wang [3], Timothy R. Crother [4], Moshe Arditi [4], David M. Underhill [1, 3, 5] 👤 ✉, Iliyan D. Iliev [1, 5, 6] 👤 ✉



Microbiome sample

Extract DNA

Extract RNA

Extract proteins and small molecules

16S rRNA gene sequencing

Total DNA sequencing (shotgun)

RNA expression profiling (transcriptomics)

Mass spectroscopy (metabolomics and proteomics)

Bacteria and archaea

Fungus/yeast

Viruses

Gene content

Gene expression

Metabolite, protein characterization

What organisms are present and what are their relative abundances?

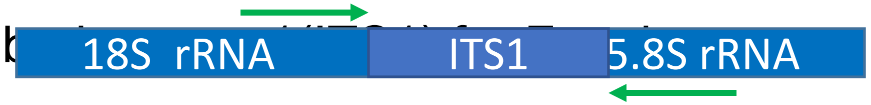
Identify relative frequencies and pathways

What are the functions of the community?

Methods

- Metagenomic sequencing
  - Expensive ($200-500/sample)
  - Stool samples only (because of host contamination)
  - Taxonomic and functional profile
  - Comprehensive and no primer bias

- Targeted amplicon sequencing
  - Cost-effective (<$50/sample)
  - All over the body
  - Taxonomic profile only (functional inference is possible)
  - Primer bias
    - 16S ribosomal DNA for Bacteria and Archaea
    - Internal transcrib

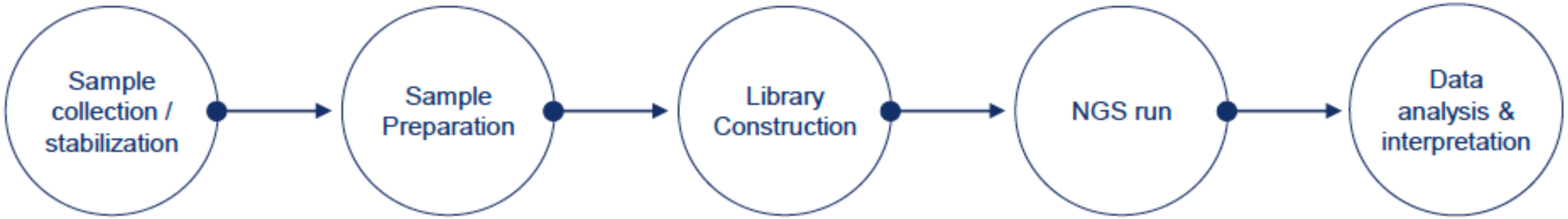


Challenges

- Sampling sources & Replicates
- Limit of “Snapshot” sampling

- PCR bias
- Variable target length

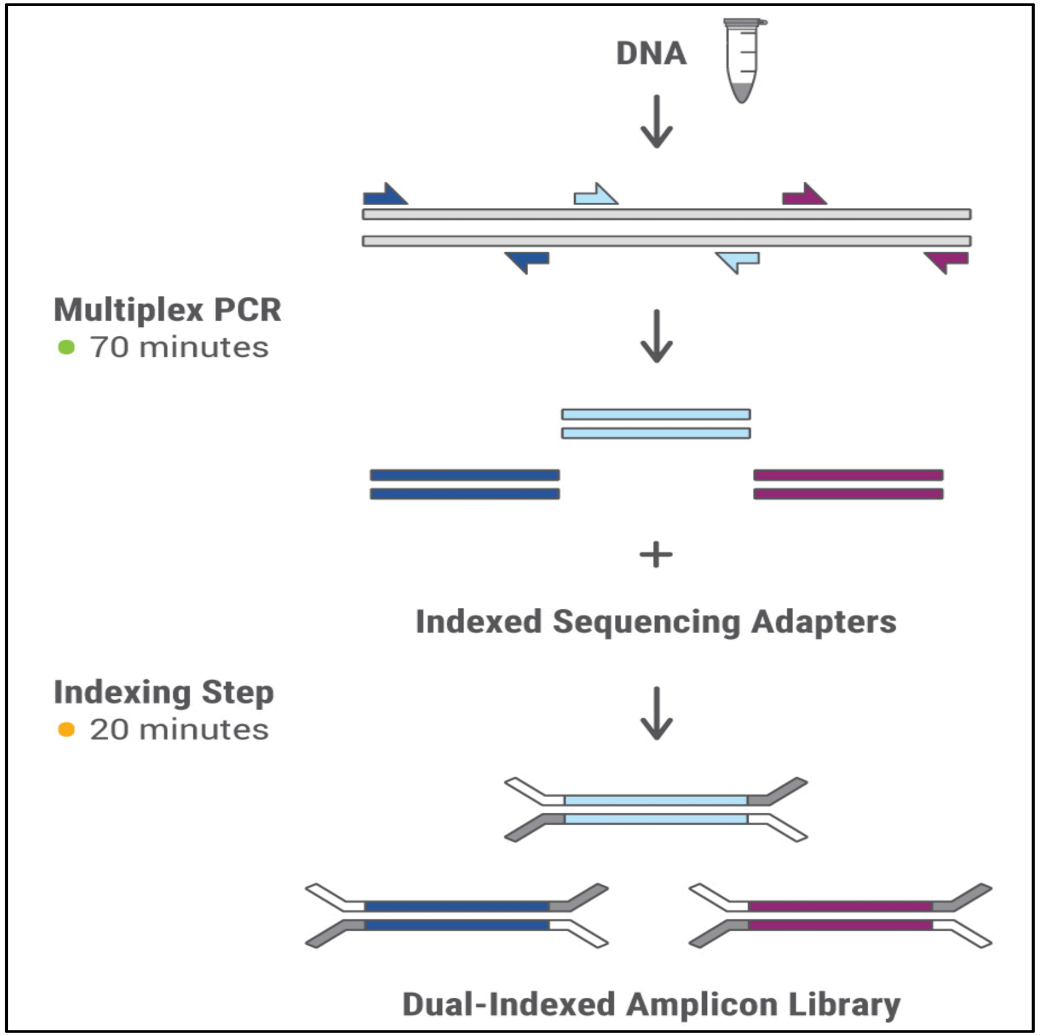
- Filtering strategy
- Normalization
- Annotation
- Complex interactions in composition data
- Multi-variate analysis with unknown factors


Sample collection / stabilization
Sample Preparation
Library Construction
NGS run
Data analysis & interpretation

- Lyse microbes
- Variable Biomass
- Ambient contamination

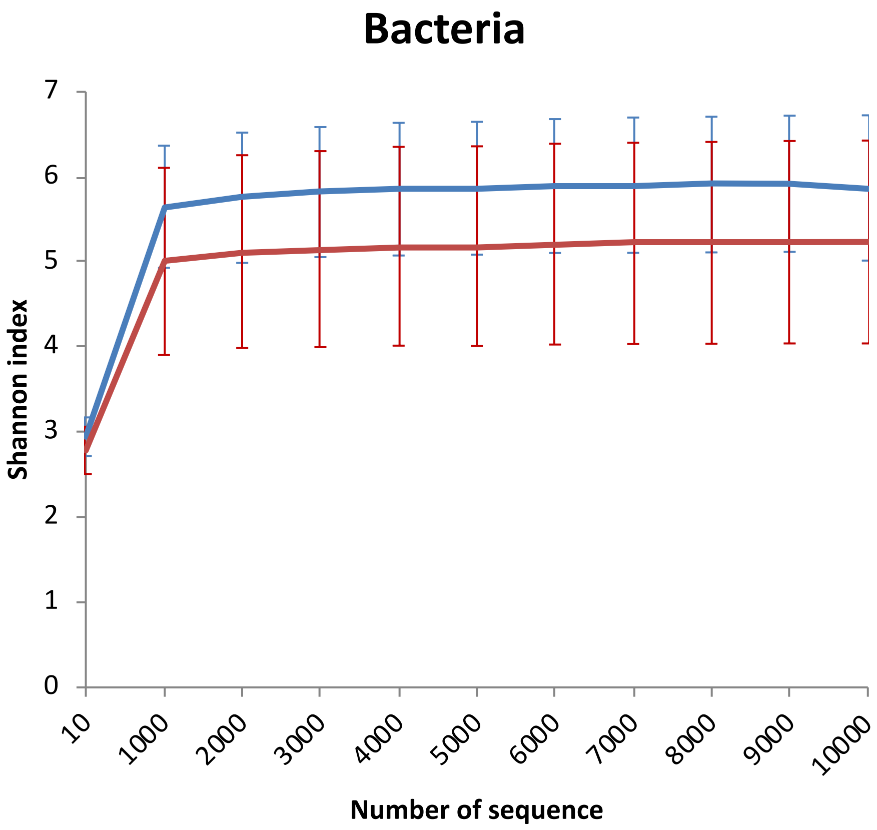
- Sequencing errors
- Sequencing depth

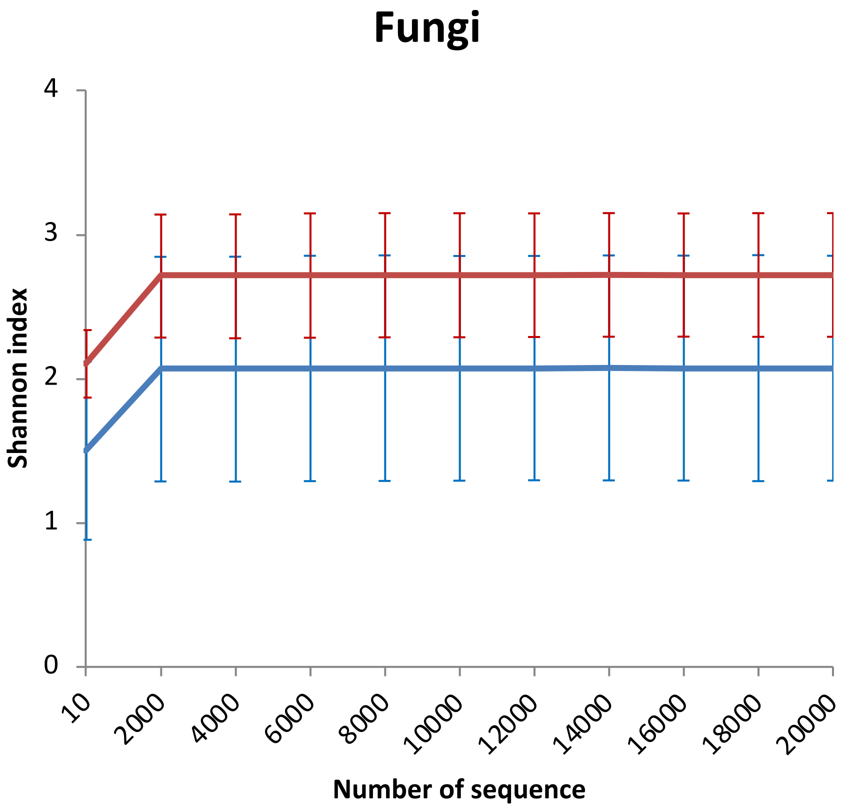
Recent updates



Leverages multiplexed primers covering all variable regions of 16S rRNA, ITS1, ITS2, and customizable region (e.g. add virulence genes, biocide resistance genes) all in one PCR reaction

Alpha-Diversity Analysis

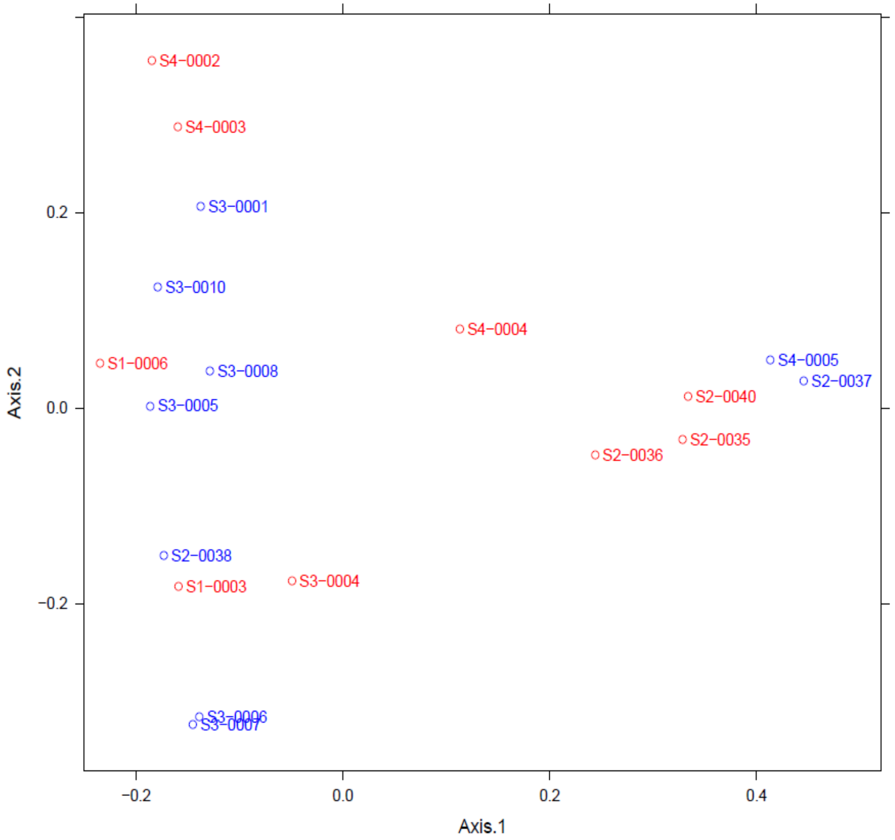




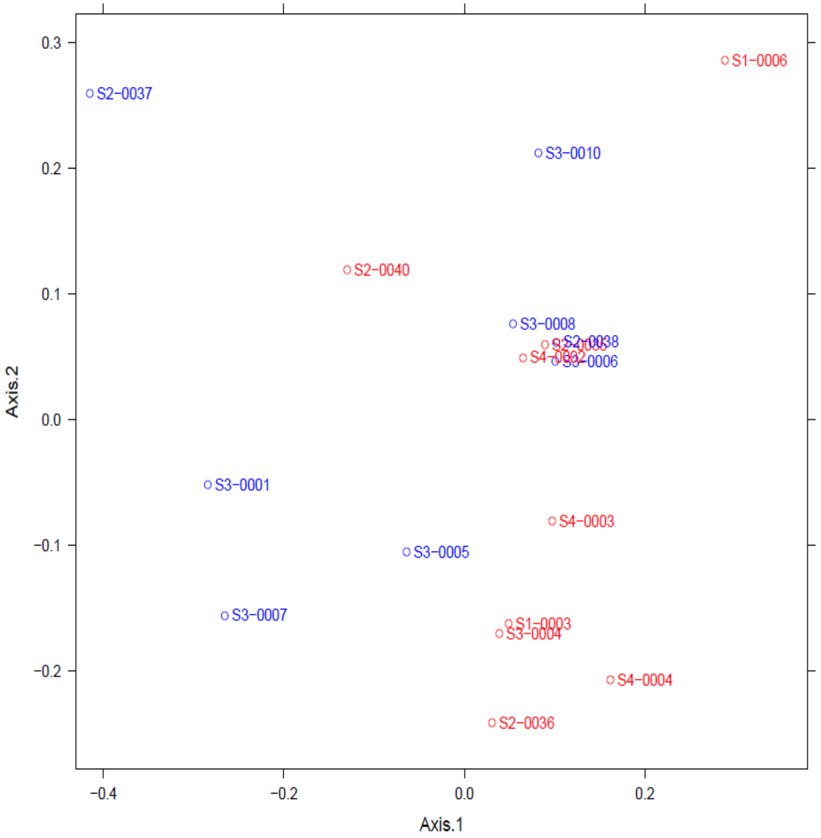
--HSCR -- HAEC

Beta-Diversity: Principle Coordinate Analysis

Bacteria : separated by age



Fungus: separated by conditions



--HSCR -- HAEC

Jaccard distance