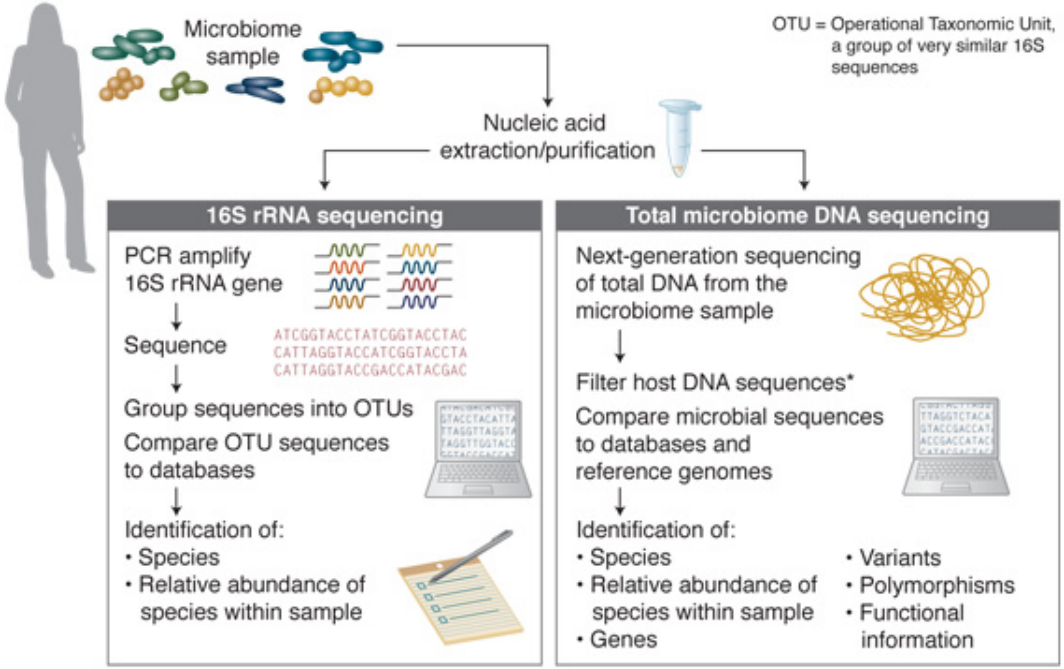
UNIVERSITY OF SOUTH FLORIDA

A PREEMINENT RESEARCH UNIVERSITY



Best practices for microbiome sample-handling and nucleic acid-processing

Swamy Adapa

https://international.neb.com/tools-and-resources/feature-articles/addressing-challenges-in-microbiome-dna-analysis

Human Microbiome

Microbiota is the coevolved ecological community of commensal, symbiotic and pathogenic microorganisms that share our body space.

General health and disease

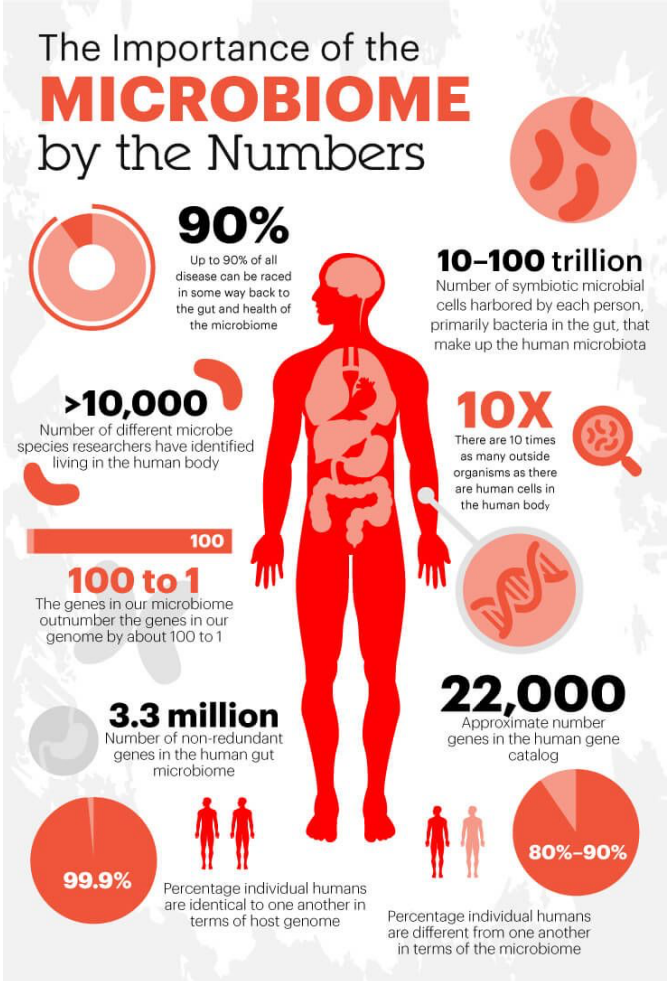
- Obesity (Energy use, Extraction/Absorption of calories)
- Arthritis
- IBD
- Neuropsychiatric illnesses (Autism, Schizophrenia, OCD, ADHD, CFS)

Disease severity

- Autoimmune disorders
- Infectious diseases
- Asthma

Response to medications

- Metabolism, efficacy and toxicity



Current Methods for Microbial Analysis

- Culture
- Gene cloning (Pan 16S rRNA) and Sanger sequencing
- Microarray
- Next-generation sequencing
  - 16S rRNA sequencing
  - Whole genome sequencing
- MALDI
- qPCR–Target dependent
  - 16S rRNA gene
  - Other relevant gene (antibiotic resistance gene, virulence factor gene)

Current Platforms

Next Generation Sequencing
Amplified Single Molecule Sequencing

Illumina

Focused Power

iSeq 100 | MiniSeq | MiSeq

Flexible Power

NextSeq 500/550

Production Power

HiSeq 2500 | HiSeq 3000/4000

Population Power

HiSeq X Five/Ten | NovaSeq

Thermo Fisher Scientific

Ion PGM

Ion Proton

Ion GeneStudio S5

Ion GeneStudio S5 Plus

Ion GeneStudio S5 Prime

Current Platforms

Third Generation Sequencing
Single Molecule Sequencing

PacBio

PacBio RSII

PacBio Sequel

Oxford Nanopore

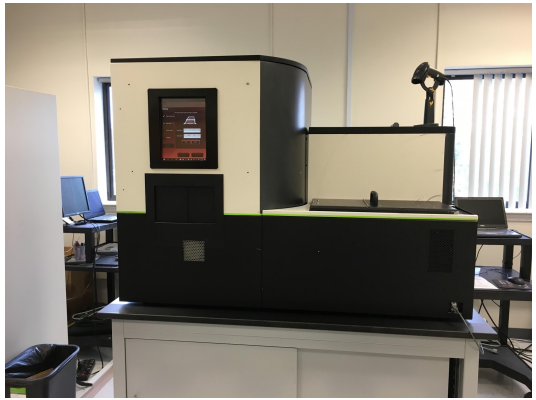
Flongle

MinION

GridION

PromethION

SeqLL



tSMS sequencer

http://omicsomics.blogspot.com/2018/03/a-morning-visit-to-seqll.html

Illumina

	iSeq 100	MiniSeq	MiSeq	NextSeq 550
Read Length	2 x 150 bp	2 x 150 bp	2 x 300 bp	2 x 150 bp
Throughput	1.2 Gb	7.5 Gb	15 Gb	120 Gb
Reads per run	4 million	50 million	50 million	800 million
Accuracy	99,9 % (>80%)	99,9 % (>80%)	99,9 % (>70%)	99,9 % (>80% of the bases)
Run Time	17.5 hours	24 hours	55 hours	29 hours

	HiSeq 2500 / 3000 / 4000	HiSeq X	NovaSeq 6000
Read Length	2 x 125 / 2 x 150 / 2 x 150 bp	2 x 150 bp	2 x 150 bp
Throughput	1000 / 750 / 1500 Gb	1800 Gb	850 – 3000 Gb
Reads per run	4 / 2,5 / 5 billion	6 billion	2.8 – 10 billion
Accuracy	99,9 % (>80% of the bases)	99,9 % (>75%)	99,9 % (>75% of the bases)
Run Time	6 / 3,5 / 3,5 days	< 3 days	36 – 44 hours

Workflow Library preparation -> Bridge amplification -> Reversible termination sequencing

Pair-End Sequencing

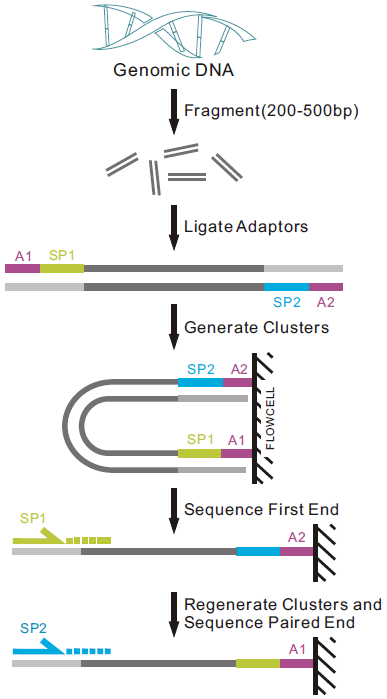


Figure 1-2-1 Pipeline of paired-end sequencing (www.illumina.com)

- The forward and reverse reads are stored in two fastq files.
- The order of pairs and naming is identical, except the designation of forward and reverse.

ID_R1_001.fastq

```
@HISEQ:100:C3MG8ACXX:5:1101:1160:2
197 1:N:0:ATCACG
CAGTTGCGATGAGAGCGTTGAGAAGTATAATAGG
AGTTAAACTGAGTAACAGGATAAGAAATAGTGAG
ATATGGAAACGTTGTGGTCTGAAAGAAGATGT
+
B@CFFFFFHHHHHGJJJJJJJJJJJFHHIIIIJJ
JIHGIIJJJJIJIJIJJJJIIJJJJJIIEIHHIJ
HGHHHHHDFFFEDDDDDCDDDCDDDDDDDCDC
```

ID_R2_001.fastq

```
@HISEQ:100:C3MG8ACXX:5:1101:1160:
2197 2:N:0:ATCACG
CTTCGTCCACTTTCATTATTCCTTTCATACATG
CTCTCCGGTTTAGGGTACTCTTGACCTGGCCTT
TTTTCAAGACGTCCCTGACTTGATCTTGAAACG
+
CCCFFFFFHHHHHJJJJIJJJJJJJJJJJJJJ
JJJJJJJIJIJGIJHBGHHIIIJIJJJJJJJJI
JJJHFFFFFFDDDDDDDDDDDDDDDEDCCDDDD
```

Sequencing depth

How many base pairs of sequence data will I get?

- Number of reads being sequenced
- Read length (if paired consider then as individuals)
- Number of samples being sequenced
- Expected percentage of usable data

$$bpPerSample = \frac{readLength * readCount}{sampleCount} * 0.8$$

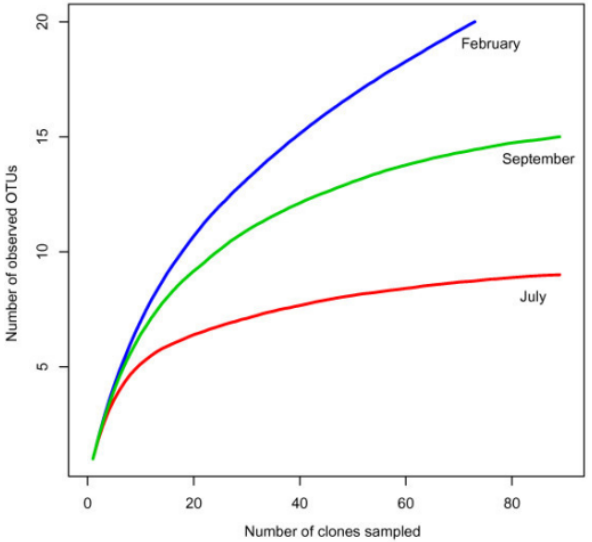
- Reads and read length data - manufacturer's website (search for specifications)
- Always use the lower end of the estimate.

For Amplicon Sequencing

- Number of reads being sequenced
- Proportion that is diversity sample (e.g. PhiX)
- Number of samples being pooled in the run

$$\frac{reads}{sample} = \frac{reads\_sequenced * (1 - diversity\_sample)}{num\_samples}$$

- Illumina 'recommends' 100K per sample
  - For 16S – 50k minimum.
- After Pilot Seq - Plot rarefactions curves to determine if saturation is achieved.



Cost Estimation

- Sample Prep
  - DNA/RNA extraction and QA/QC (Bioanalyzer/Gels)
  - Library QA/QC (Bioanalyzer and Qubit/microplate reader)
  - Pooling
- Sequencing (Number of Lanes / runs)
- Bioinformatics (General rule is to estimate the same amount as data generation, i.e. double your budget)

Sequencing Costs

- Metagenomics
  - Shotgun libraries intended to sequence random genomic sequences from the entire bacterial community.
  - Can be costly per sample ($500 to multi thousands per sample)
  - Better resolution and sensitivity to characterize the sample
  - Due to cost, can only do relatively few samples
- Amplicon Sequencing
  - Sequence only one regions of one gene (e.g. 16s, ITS, LSU)
  - Cheap per sample (at scale, down to $20/sample)
  - Due to cost, can do many hundreds of samples make more global inferences

Bioinformatics Costs

Bioinformatics includes:

- Storage of data, metadata maintenance
- Access and use of computational resources and software
- System Administration time
- Bioinformatics Data Analysis time
- Back and forth consultation/analysis to extract biological meaning

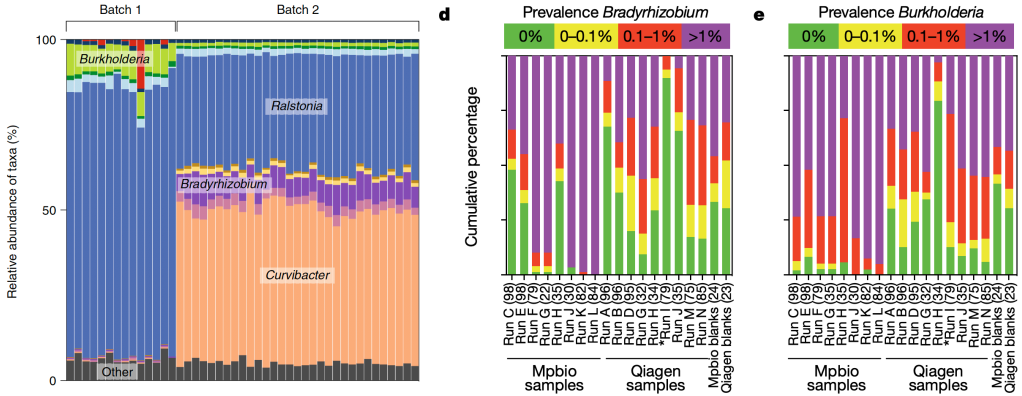
Rule of thumb:

Bioinformatics can and should cost as much (sometimes more) as the cost of data generation.

Cost Estimation

- Amplicons
  - 384 Samples
  - Amplicon generation (~$20/sample)= $7,680
  - Sequencing PE300, target 50K reads per sample
  - Bioinformatics
- Metagenome
  - 12 samples (DNA) = $400/sample = $4,800
  - Expectations: Host Proportion < 50%, use average genome size of EColi, Target the 1% and coverage of 20
  - Sequencing PE150
  - Bioinformatics

Kit’ome - Reagent microbiome



ARTICLE

Corrected: Author Correction

https://doi.org/10.1038/s41586-019-1451-5

Human placenta has no microbiome but can contain potential pathogens

Marcus C. de Goffau[1,2,8], Susanne Lager[3,4,5,8], Ulla Sovio[3,4], Francesca Gaccioli[3,4], Emma Cook[3], Sharon J. Peacock[1,6,7], Julian Parkhill[1,2*], D. Stephen Charnock-Jones[3,4,9] & Gordon C. S. Smith[3,4,9*]

Recognizing the reagent microbiome

A noticeable part of the microbiome literature, especially that working with low-biomass samples, is plagued by reagent contamination. Here we describe visual, statistical, methodical and ecological techniques to facilitate recognition of signals that represent contamination.

Marcus C. de Goffau, Susanne Lager, Susannah J. Salter, Josef Wagner, Andreas Kronbichler, D. Stephen Charnock-Jones, Sharon J. Peacock, Gordon C. S. Smith and Julian Parkhill

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Aug. 1998, p. 3110–3113
0099-2240/98/$04.00+0

Vol. 64, No. 8

Specific Ribosomal DNA Sequences from Diverse Environmental Settings Correlate with Experimental Contaminants

MICHAEL A. TANNER, BRETT M. GOEBEL,† MICHAEL A. DOJKA, AND NORMAN R. PACE*

Departments of Plant and Microbial Biology and Molecular and Cell Biology, University of California, Berkeley, California 94720-3102

Received 23 April 1998/Accepted 10 June 1998

Kit'ome - Reagent microbiome

- Breast tissue from cancer patients and normal controls
  - Different abundances of Methylobacterium radiotolerans and Sphingomonas yanoikuyae.
  - Methylobacteriaceae and Shingomonadaceae popped up in some of the DNA extraction kits.
- Bradyrhizobium enterica, in patients with a colitis syndrome.
  - Bradyrhizobium were also detected in DNA extraction kits

OPEN ACCESS Freely available online

PLOS ONE

Microbial Dysbiosis Is Associated with Human Breast Cancer

Caiyun Xuan[1], Jaime M. Shamonki[2], Alice Chung[3], Maggie L. DiNome[4], Maureen Chung[4], Peter A. Sieling[1], Delphine J. Lee[1]*

1 Dirks/Dougherty Laboratory for Cancer Research, Department of Translational Immunology, John Wayne Cancer Institute, Santa Monica, California, United States of America, **2** Pathology Department, Saint John's Health Center, Santa Monica, California, United States of America, **3** Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, California, United States of America, **4** Margie Petersen Breast Center, Saint John's Health Center, Santa Monica, California, United States of America

The NEW ENGLAND JOURNAL of MEDICINE

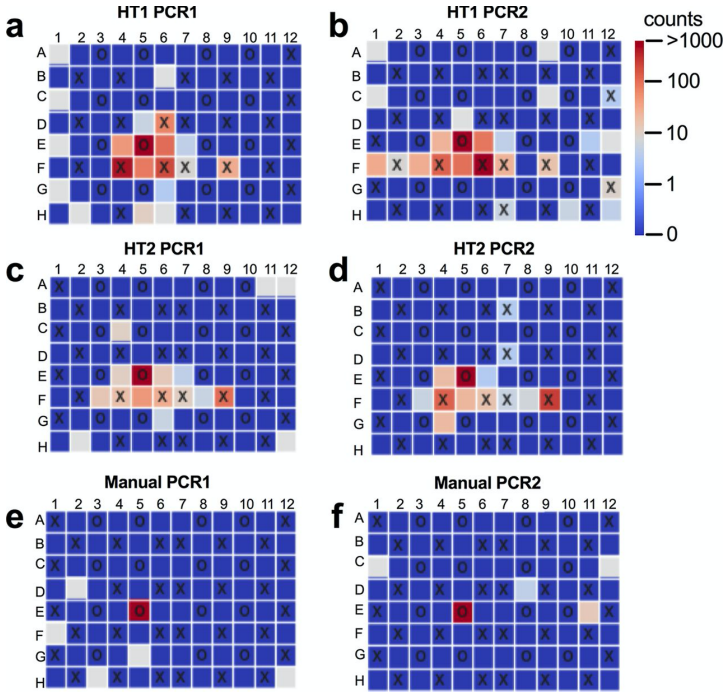
ORIGINAL ARTICLE

Sequence-Based Discovery of *Bradyrhizobium enterica* in Cord Colitis Syndrome

Ami S. Bhatt, M.D., Ph.D., Samuel S. Freeman, B.S.E., Alex F. Herrera, M.D., Chandra Sekhar Pedamallu, Ph.D., Dirk Gevers, Ph.D., Fujiko Duke, B.S., Joonil Jung, Ph.D., Monia Michaud, M.Sc., Bruce J. Walker, B.S., Sarah Young, Ph.D., Ashlee M. Earl, Ph.D., Aleksander D. Kostic, Ph.D., Akinyemi I. Ojesina, M.D., Ph.D., Robert Hasserjian, M.D., Karen K. Ballen, M.D., Yi-Bin Chen, M.D., Gabriela Hobbs, M.D., Joseph H. Antin, M.D., Robert J. Soiffer, M.D., Lindsey R. Baden, M.D., Wendy S. Garrett, M.D., Ph.D., Jason L. Hornick, M.D., Ph.D., Francisco M. Marty, M.D., and Matthew Meyerson, M.D., Ph.D.

Well-to-Well Contamination

- Decreased cost of 16S rRNA -> Automation
- Outsourced samples are always processed using High-throughput platforms.
- Between-sample contamination is often neglected.
- Study found
  - DNA extraction primary contamination.
  - Lesser extent during library preparation
  - Barcode leakage was negligible.
- Recommendations:
  - Samples should be randomized across plates
  - Samples of similar biomasses should be processed together
  - Manual single-tube extractions or hybrid plate-based cleanups should be employed.



J. J. Minich *et al.*, Quantifying and Understanding Well-to-Well Contamination in Microbiome Research. *mSystems* **4**, e00186-00119 (2019).

Need for Controls – Blanks & Mock

- Positive or Mock Community control
  - To assess sequencing error rate
- Negative or water extraction sample.
  - Essential for low biomass samples
  - Probably a good idea for high biomass samples
- Airborne contaminants.
- Autoclave does not kill DNA.

Avoid simplistic removals of taxa or operational taxonomic units (OTUs) appearing in negative controls – The contaminant could be real.

Batch Effect – Run Preparation

- Do not just sequence samples as they come.

- Human Microbiome Project
  - Largest signal in that study was where the DNA extractions were done - St. Louis or Houston (Regional effect on the microbiome?)
  - Geographic region was perfectly confounded with who did the sampling, extractions, amplifications, and sequencing.
- Salter et. al, found that the differences they observed were due to differences in the contaminants that were present in the DNA extraction kits that they used.
  - Two batches of extractions used different lots of extraction kits
  - Treatment groups were confounded with the DNA extraction kit lots

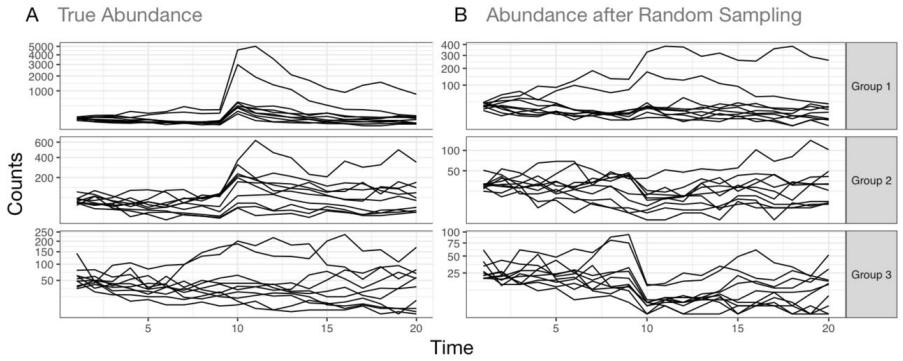
- Wait until you have all of your samples.
- Either randomize the samples or thoughtfully block your samples.
  - So that treatment groups are evenly represented across your extraction sets and sequencing runs.

Turner P, Turner C, Jankhot A, Helen N, Lee SJ, et al. (2012) A Longitudinal Study of Streptococcus pneumoniae Carriage in a Cohort of Infants and Their Mothers on the Thailand-Myanmar Border. PLOS ONE 7(5): e38271.

Salter, S.J., Cox, M.J., Turek, E.M. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* **12,** 87

Random sampling can lead to spurious conclusions in longitudinal studies.

- Measurements of microbiota can differ substantially from the underlying true community structure.
- DNA of samples are randomly selected, processed, and then included in the subsequent step.
- The randomness of this data generation process introduces uncertainty into microbiome measurements.
- Technical variation from sample processing can influence ex vivo microbiota dynamics, accounting for 76% of community variation on hourly timescales.

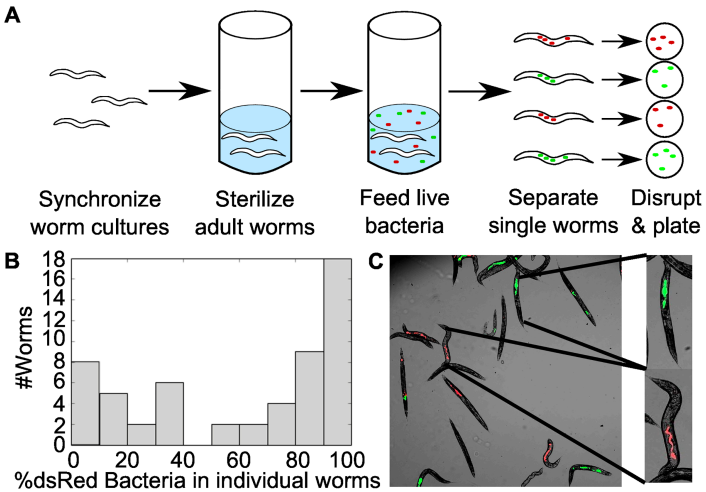


https://drive.google.com/file/d/1U36NyMinKAm1R2rwgRb0fJUFVHbeYS9J/view

Silverman, J.D., Durand, H.K., Bloom, R.J. *et al.* Dynamic linear models guide design and analysis of microbiota studies within artificial human guts. *Microbiome* **6,** 202 (2018)

Microbiome diversity is influenced by chance encounters

- Genetically identical worms were put into identical environments and fed them the same diet.
- The worms developed very different populations of bacteria in their gut, depending on which bacteria happened to make it there first.
- When the experiment began, the worms had no bacteria in their digestive tracts, but after being exposed to the identical bacterial diet, the worms each generated very different microbe populations in their guts.



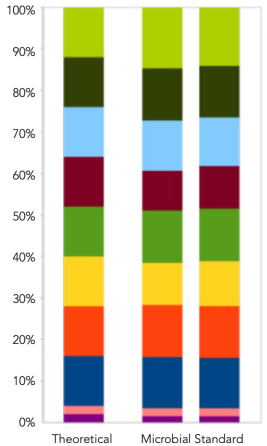
Vega NM, Gore J (2017) Stochastic assembly produces heterogeneous communities in the *Caenorhabditis elegans* intestine. PLOS Biology 15(3): e2000633.

- Studies involving multiple individuals over time should design the experiment in a way that will subsequently be utilized by the the data modeling scheme.
- Seasonality effects can also be important to consider in experimental design.
- While perhaps most obvious in environmental studies, dietary changes surrounding holidays or weekends, or even natural circadian variation can also be important factors to consider.

Silverman, J.D., Durand, H.K., Bloom, R.J. *et al.* Dynamic linear models guide design and analysis of microbiota studies within artificial human guts. *Microbiome* **6,** 202 (2018)

Microbiome Standards - Mock Communities

- Mock community control composed of predetermined ratios of Microbes/DNA from a mixture of bacterial species.
  - Quantification of sequencing error
  - Identify bias introduced during the sampling and library preparation processes.
- Pre-prepared bacterial communities are available in two different formats: DNA mock communities and whole-cell mock communities.
- Whole-cell mock communities are useful for establishing the efficiency of the DNA extraction step.
- DNA mock communities will only assess the efficiency of PCR, clean-up, sequencing, and analysis steps.
- Mock communities are available from the American Type Culture Collection (ATCC - BIE) and Zymo Research.

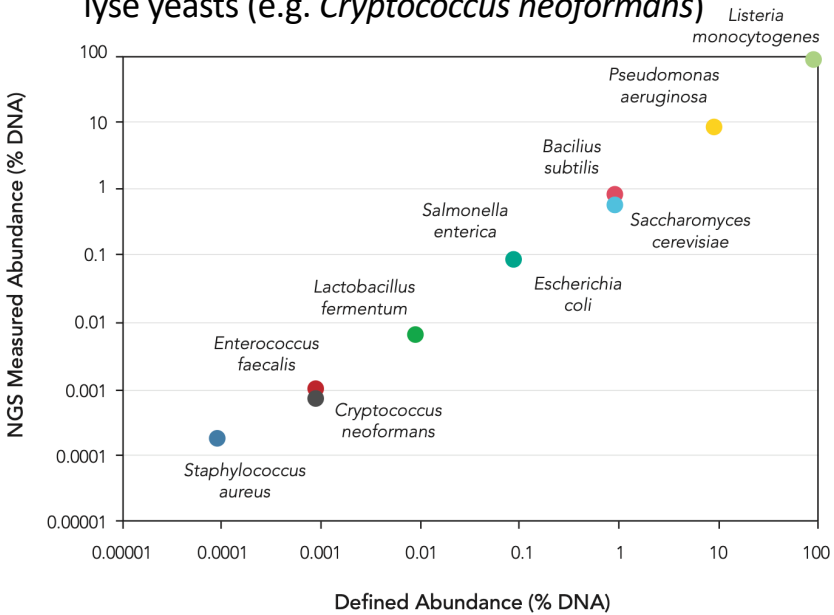


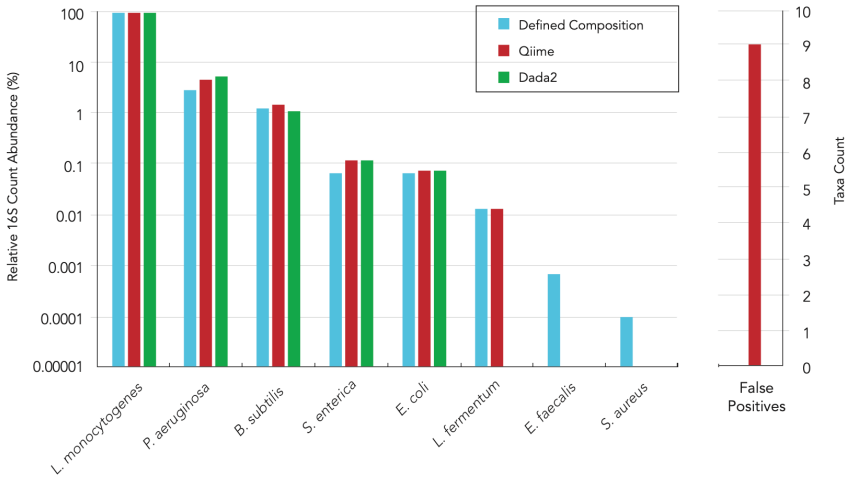
Species	Avg. GC (%)	Gram Stain	gDNA Abun. (%)
Pseudomonas aeruginosa	66.2	-	12
Escherichia coli	56.8	-	12
Salmonella enterica	52.2	-	12
Lactobacillus fermentum	52.8	+	12
Enterococcus faecalis	37.5	+	12
Staphylococcus aureus	32.7	+	12
Listeria monocytogenes	38.0	+	12
Bacillus subtilis	43.8	+	12
Saccharomyces cerevisiae	38.4	Yeast	2
Cryptococcus neoformans	48.2	Yeast	2

https://www.zymoresearch.com/products/zymobiomics-microbial-community-standard

Staggered Mock Communities

Zymo Research's Microbial Community Standard includes three easy-to-lyse Gram-negative bacteria (e.g. *Escherichia coli*), five tough-to-lyse Gram-positive bacteria (e.g. *Listeria monocytogenes*), and two tough-to-lyse yeasts (e.g. *Cryptococcus neoformans*)





https://www.zymoresearch.com/products/zymobiomics-microbial-community-standard-ii-log-distribution

Mock communities help identify aberrant sequencing run

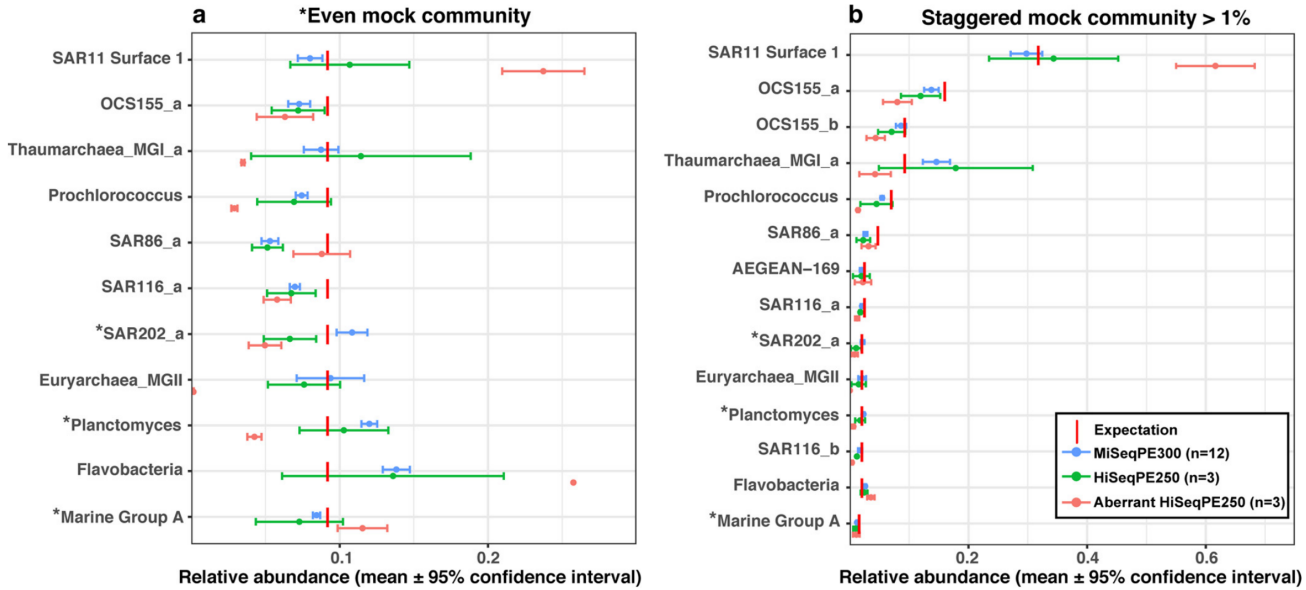


FIG 1 Comparisons of "even" mock communities (a) and "staggered" mock communities (b) sequenced by MiSeqPE300 and HiSeqPE250. Values that are significantly different for a clone by MiSeqPE300 versus HiSeqPE250 are indicated with an asterisk before the clone name ($P < 0.05$ by Wilcoxon rank sum test). Significant differences in the whole-community composition by MiSeqPE300 and HiSeqPE250 were found only in the even mock community ($P < 0.05$ by ANOSIM test).

Y.-C. Yeh, D. M. Needham, E. T. Sieradzki, J. A. Fuhrman, Taxon Disappearance from Microbiome Analysis Reinforces the Value of Mock Communities as a Standard in Every Sequencing Run. *mSystems* **3**, e00023-00018 (2018).

Microbiome Standards - Spike-in Controls

- Unlike mock communities, these standards are added directly to samples.
- Quality control can be performed on a per-sample basis.
- Care must be taken to select bacteria which are highly unlikely to occur in the samples of interest.

Example:

Imtechella halotolerans(Gram-negative) and *Allobacillus halotolerans* (Gram-positive) – which are alien to human microbiome.

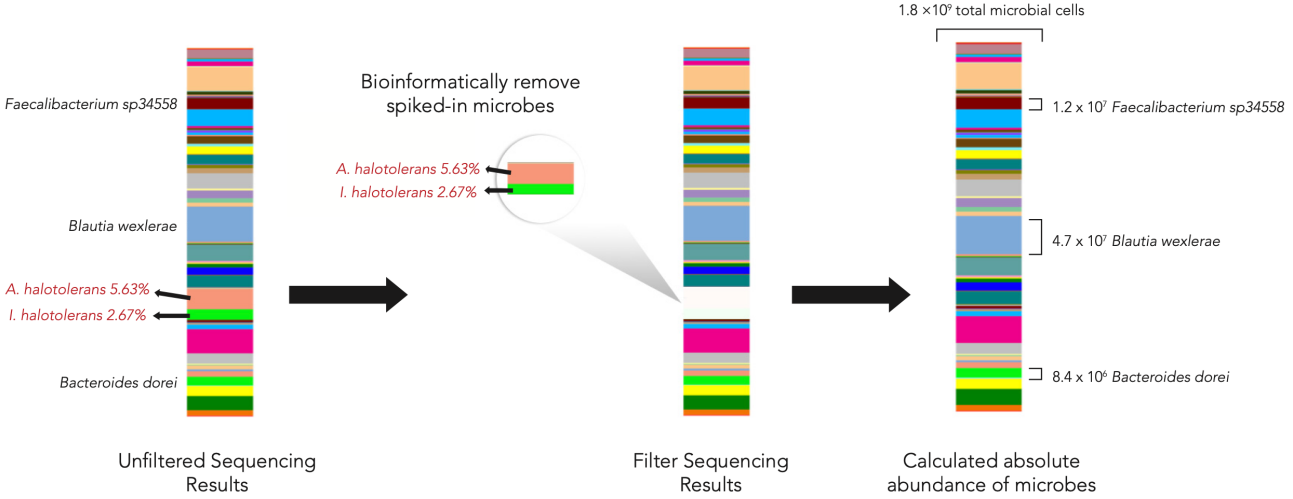
Spike-in controls

- Microbial Composition

Species	Per Prep (20 µl)		
	Cells	16S Copies[1]	Total DNA[2] (ng)
Imtechella halotolerans	**2 x 10^7**	6.0 x 10^7	67.2
Allobacillus halotolerans	**2 x 10^7**	1.4 x 10^8	58.2

[1] 16S copies = cells × 16S copy number per cell/genome.
[2] Total genomic DNA (ng) = cells × genome size (bp/genome) × DNA unit conversion constant (ng/bp).
DNA unit conversion constant (ng/bp) = 1.079 x 10^{-12}.

- Differences in abundance ratio indicate potential workflow bias.
- If the abundance of *I. halotolerans* is much higher than that of *A. halotolerans.*
  - Bias during DNA extraction
  - *A. halotolerans* is Gram-positive, tougher to lyse than *I. halotolerans*.



https://www.zymoresearch.com/collections/zymobiomics-microbial-community-standards/products/zymobiomics-spike-in-control-i-high-microbial-load

UNIVERSITY OF SOUTH FLORIDA

A PREEMINENT RESEARCH UNIVERSITY