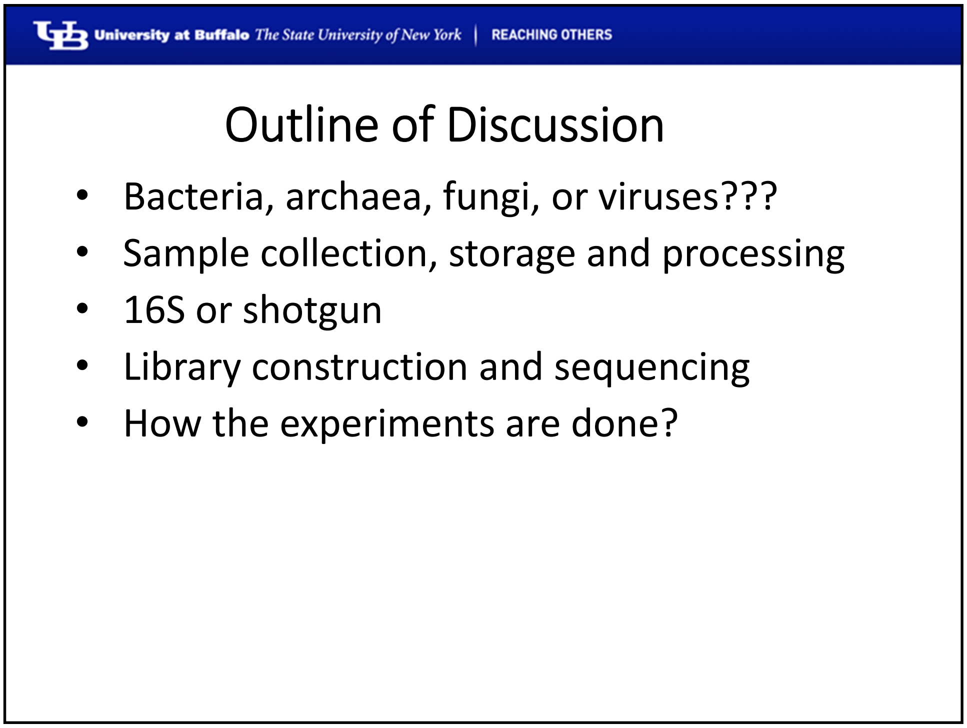
"So you want to do microbiome research…"

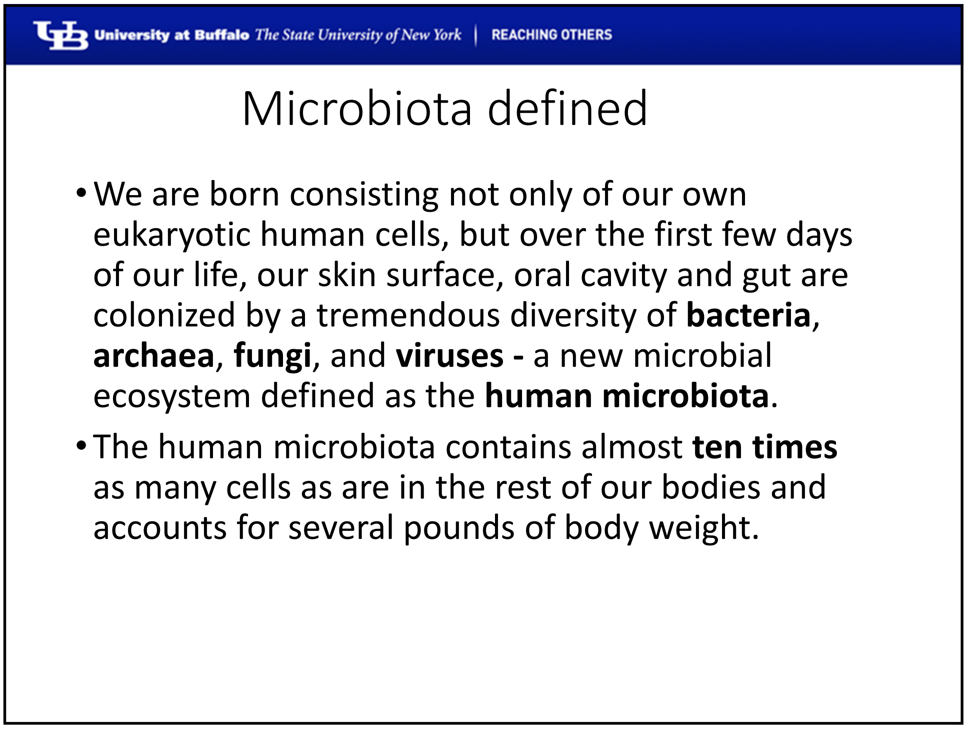
Michael J. Buck, Ph.D.
mjbuck@buffalo.edu

Maria Tsompana, Ph.D.
tsompana@buffalo.edu



Outline of Discussion

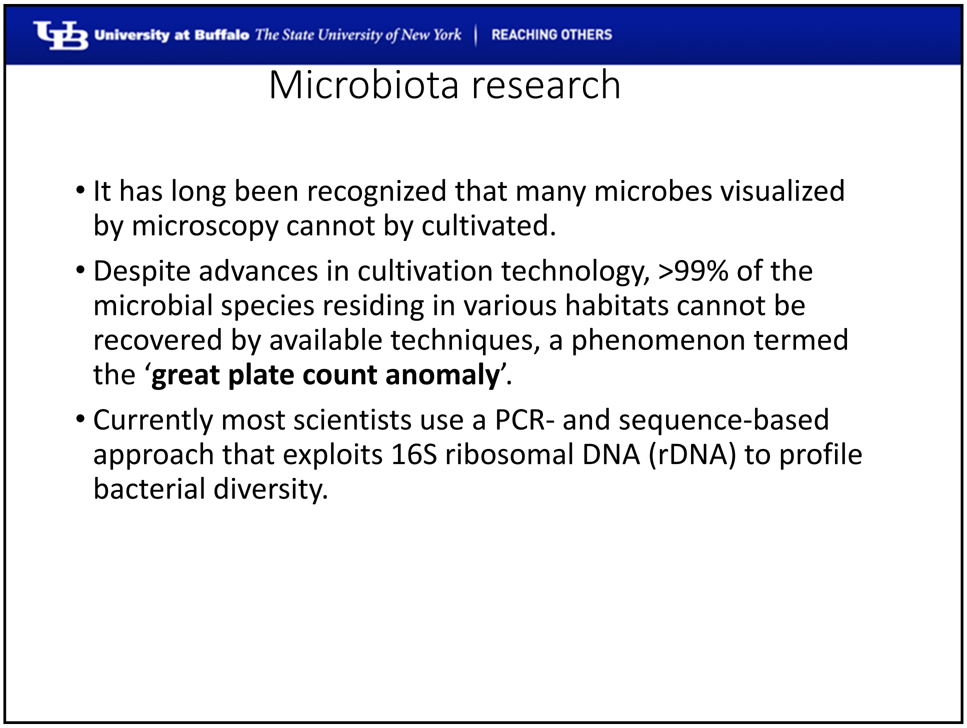
- Bacteria, archaea, fungi, or viruses???
- Sample collection, storage and processing
- 16S or shotgun
- Library construction and sequencing
- How the experiments are done?



University at Buffalo *The State University of New York* | REACHING OTHERS

Microbiota defined

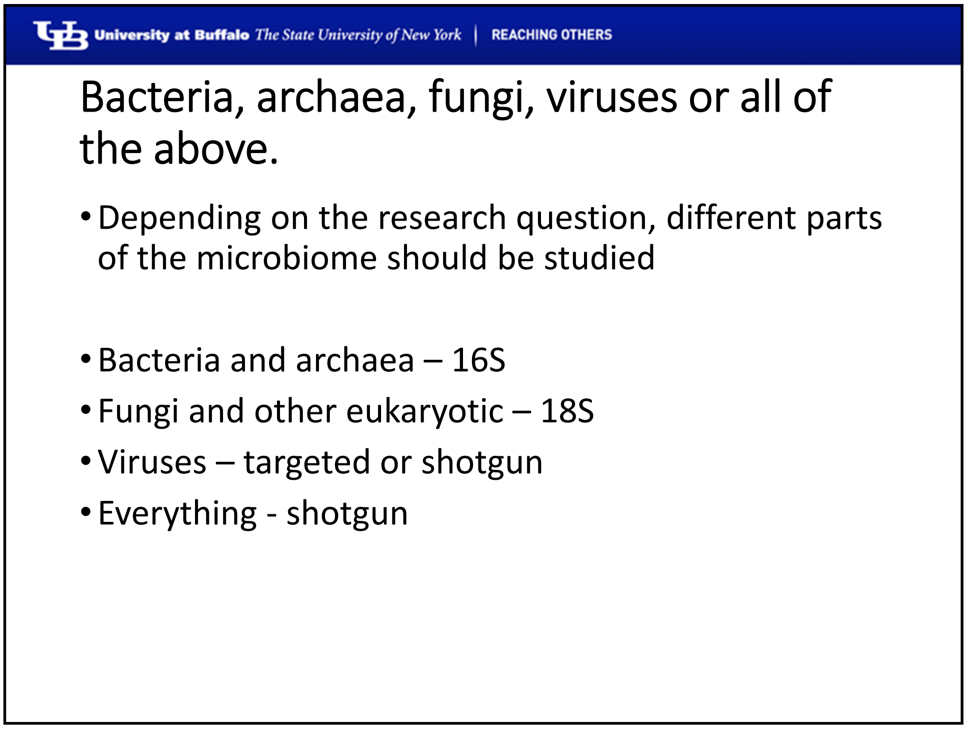
- We are born consisting not only of our own eukaryotic human cells, but over the first few days of our life, our skin surface, oral cavity and gut are colonized by a tremendous diversity of **bacteria**, **archaea**, **fungi**, and **viruses -** a new microbial ecosystem defined as the **human microbiota**.
- The human microbiota contains almost **ten times** as many cells as are in the rest of our bodies and accounts for several pounds of body weight.



University at Buffalo *The State University of New York* | REACHING OTHERS

Microbiota research

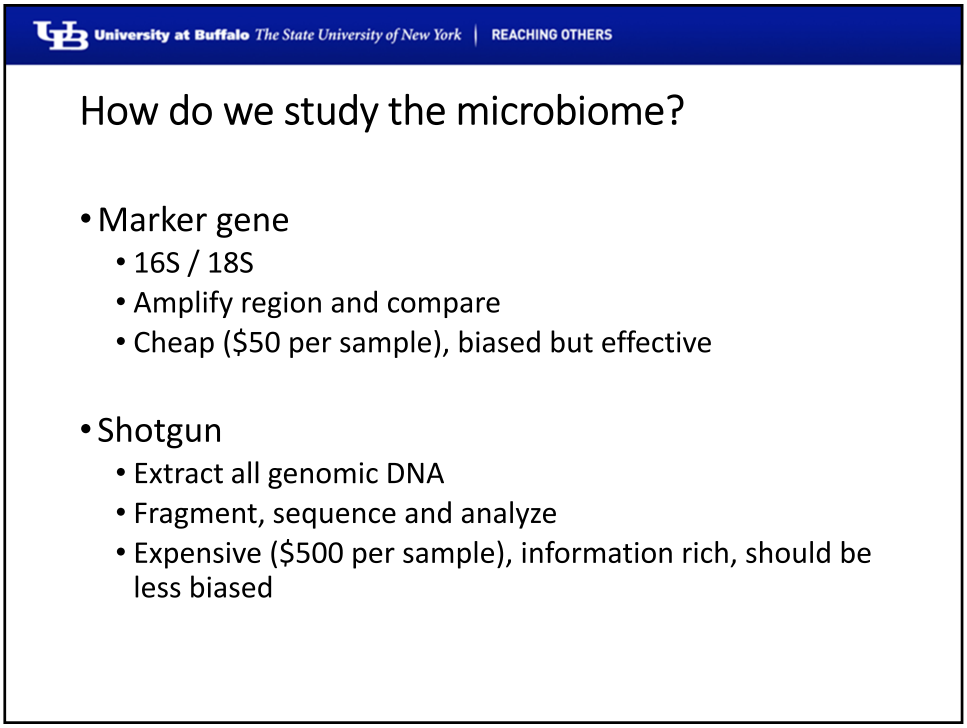
- It has long been recognized that many microbes visualized by microscopy cannot by cultivated.
- Despite advances in cultivation technology, >99% of the microbial species residing in various habitats cannot be recovered by available techniques, a phenomenon termed the '**great plate count anomaly**'.
- Currently most scientists use a PCR- and sequence-based approach that exploits 16S ribosomal DNA (rDNA) to profile bacterial diversity.



Bacteria, archaea, fungi, viruses or all of the above.

- Depending on the research question, different parts of the microbiome should be studied

- Bacteria and archaea – 16S
- Fungi and other eukaryotic – 18S
- Viruses – targeted or shotgun
- Everything - shotgun



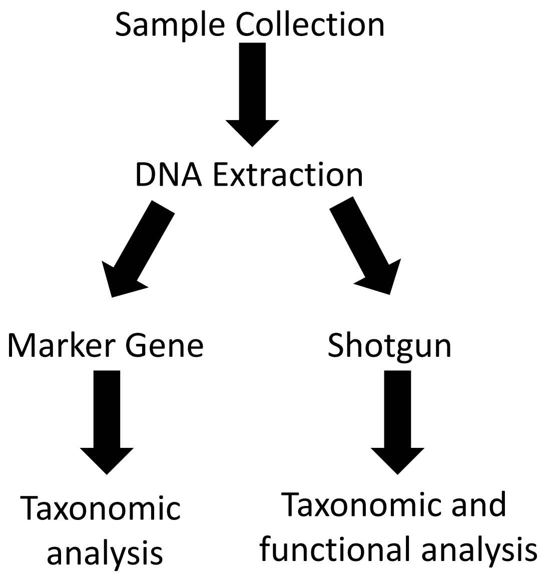
How do we study the microbiome?

- Marker gene
  - 16S / 18S
  - Amplify region and compare
  - Cheap ($50 per sample), biased but effective

- Shotgun
  - Extract all genomic DNA
  - Fragment, sequence and analyze
  - Expensive ($500 per sample), information rich, should be less biased

University at Buffalo *The State University of New York* | REACHING OTHERS

Overview



Sample Collection

DNA Extraction

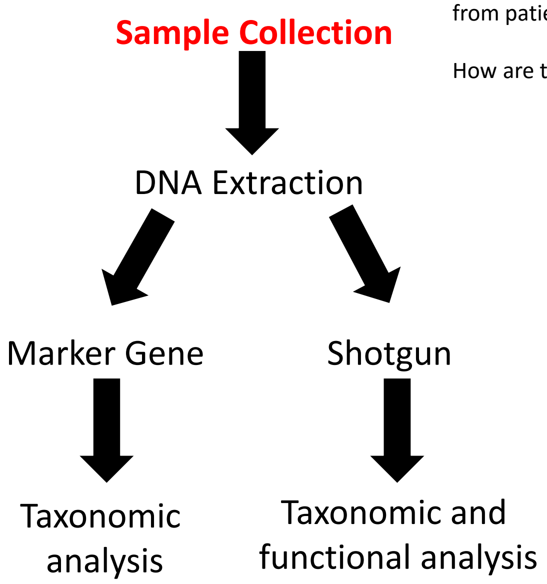
Marker Gene

Shotgun

Taxonomic analysis

Taxonomic and functional analysis

University at Buffalo *The State University of New York* | REACHING OTHERS



Sample Collection

DNA Extraction

Marker Gene

Shotgun

Taxonomic analysis

Taxonomic and functional analysis

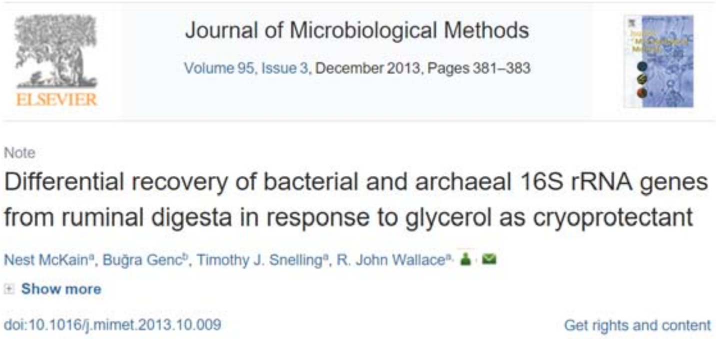
How are samples collected from patients or animals?

How are they stored?

Sample collection, storage and processing

- Bacteria like to grow, *E. coli* doubles every 20-30 minutes!
- Anaerobic versus aerobic bacteria will grow at different rates in sample collection tubes.
  - So if a person collects a sample at home and stores it in in the fridge or even the freezer the population of bacteria will change over time.
- Samples need to be chemically preserved or flash frozen at -80

Sample storage



Journal of Microbiological Methods

Volume 95, Issue 3, December 2013, Pages 381–383

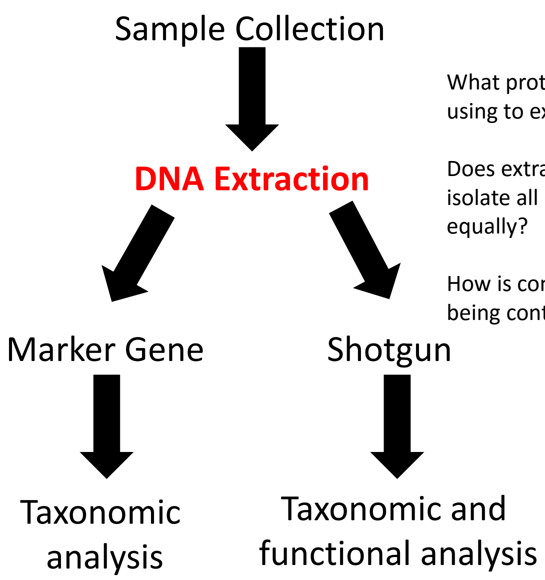
Note

Differential recovery of bacterial and archaeal 16S rRNA genes from ruminal digesta in response to glycerol as cryoprotectant

Nest McKain[a], Buğra Genc[b], Timothy J. Snelling[a], R. John Wallace[a]

doi:10.1016/j.mimet.2013.10.009

"Samples frozen with and without glycerol as cryoprotectant indicated a major loss of *Bacteroidetes* in unprotected samples"



Sample Collection

DNA Extraction

Marker Gene

Shotgun

Taxonomic analysis

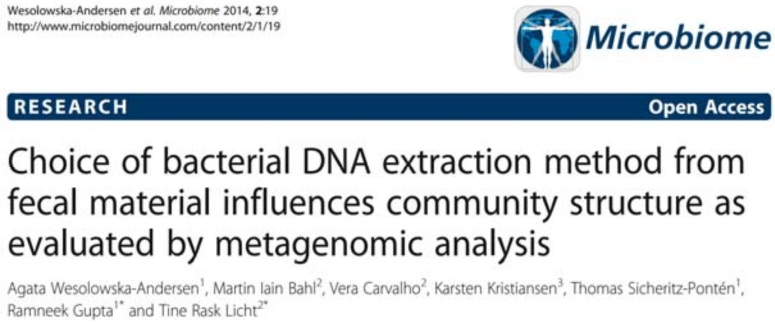
Taxonomic and functional analysis

What protocol are you using to extract DNA?

Does extraction method isolate all organisms equally?

How is contamination being controlled for?

DNA extraction techniques can introduce bias



Wesolowska-Andersen *et al. Microbiome* 2014, **2**:19
http://www.microbiomejournal.com/content/2/1/19

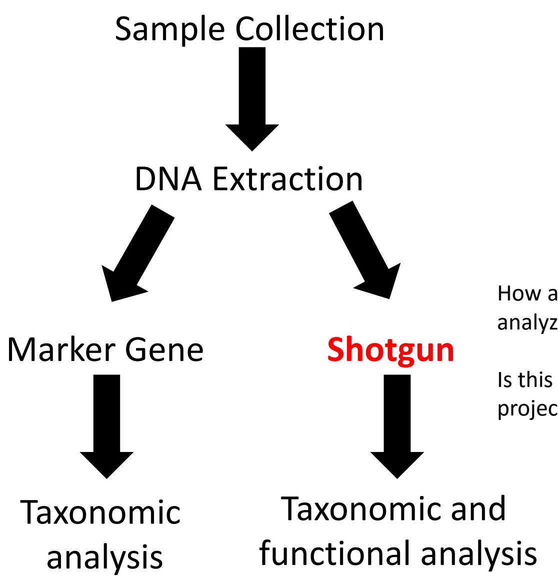
Microbiome

RESEARCH **Open Access**

Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis

Agata Wesolowska-Andersen[1], Martin Iain Bahl[2], Vera Carvalho[2], Karsten Kristiansen[3], Thomas Sicheritz-Pontén[1], Ramneek Gupta[1*] and Tine Rask Licht[2*]

“We observed significant differences in distribution of bacterial taxa depending on the method”

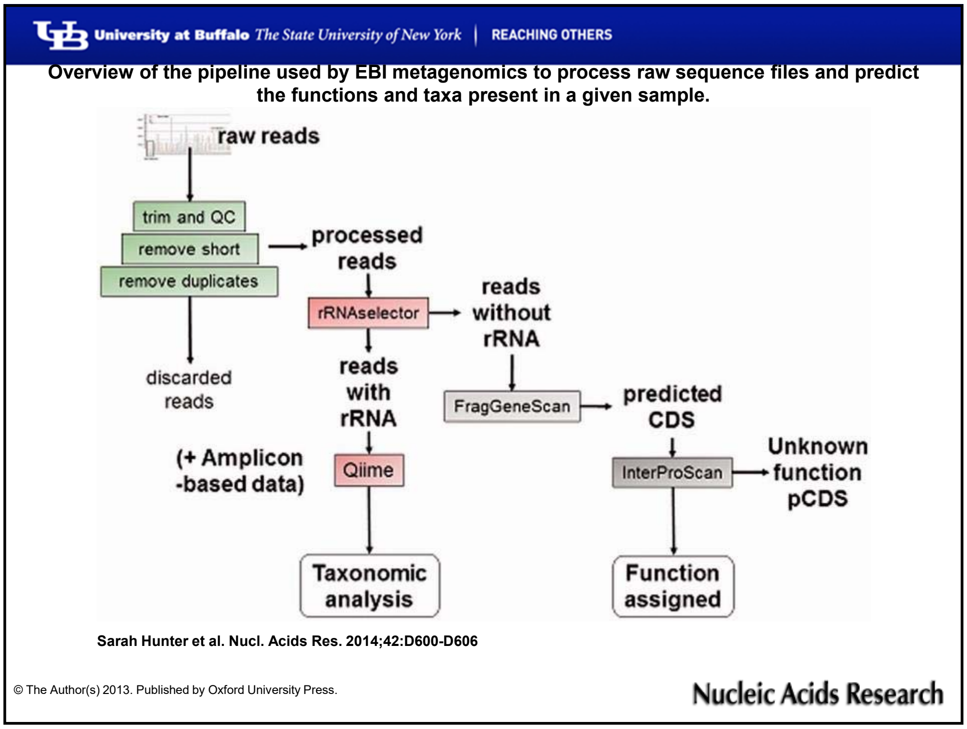


How are you going to analyze the data?

Is this overkill for your project?

Shotgun sequencing

- Isolate all DNA from a sample and sequence it
- Should be less biased compared to “marker studies”



University at Buffalo The State University of New York | REACHING OTHERS

Overview of the pipeline used by EBI metagenomics to process raw sequence files and predict the functions and taxa present in a given sample.

raw reads

trim and QC

remove short

remove duplicates

processed reads

discarded reads

rRNAselector

reads without rRNA

reads with rRNA

(+ Amplicon -based data)

Qiime

FragGeneScan

predicted CDS

InterProScan

Unknown function pCDS

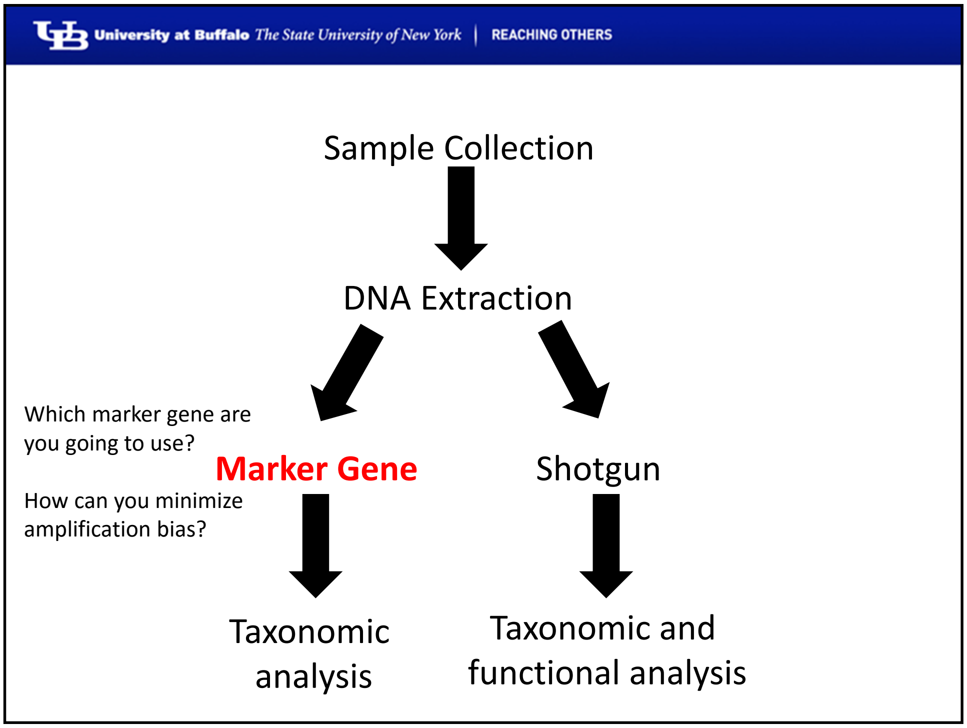
Taxonomic analysis

Function assigned

Sarah Hunter et al. Nucl. Acids Res. 2014;42:D600-D606

© The Author(s) 2013. Published by Oxford University Press.

Nucleic Acids Research



University at Buffalo The State University of New York | REACHING OTHERS

Sample Collection

DNA Extraction

Which marker gene are you going to use?

How can you minimize amplification bias?

Marker Gene

Shotgun

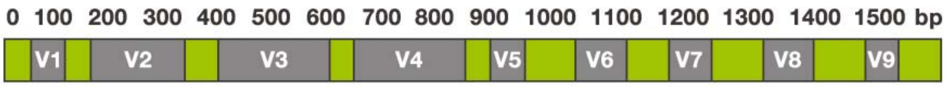
Taxonomic analysis

Taxonomic and functional analysis

16S rDNA

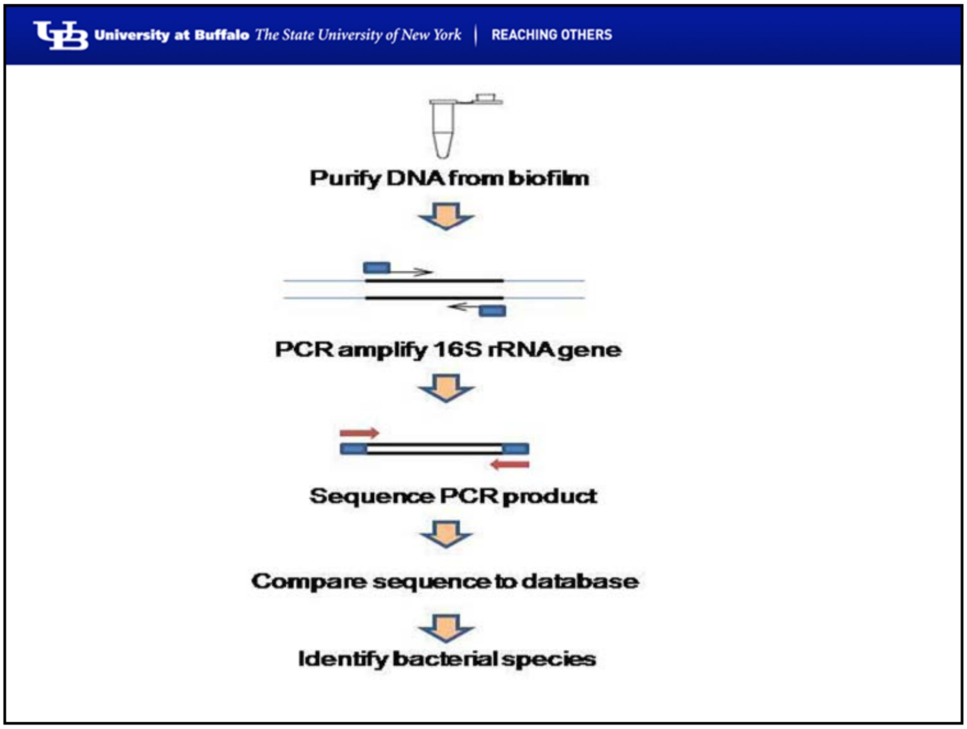
- 16S rDNA: is a component of the 30S small subunit of prokaryotic ribosomes.
- 16S rDNA: it satisfies the criteria of a marker by containing both highly conserved, ubiquitous sequences and regions that vary with greater or lesser frequency over evolutionary time.
- The products of the rRNA genes can fold into a complex, stable secondary structure, consisting of stems and loops. The sequences of some of the loops are conserved across nearly all bacterial species because of the essential functions involved, whereas the features of the structural parts are largely variant and specific to one or more classes.

16S rDNA



CONSERVED REGIONS: unspecific applications

VARIABLE REGIONS: group or species-specific applications



University at Buffalo *The State University of New York* | REACHING OTHERS

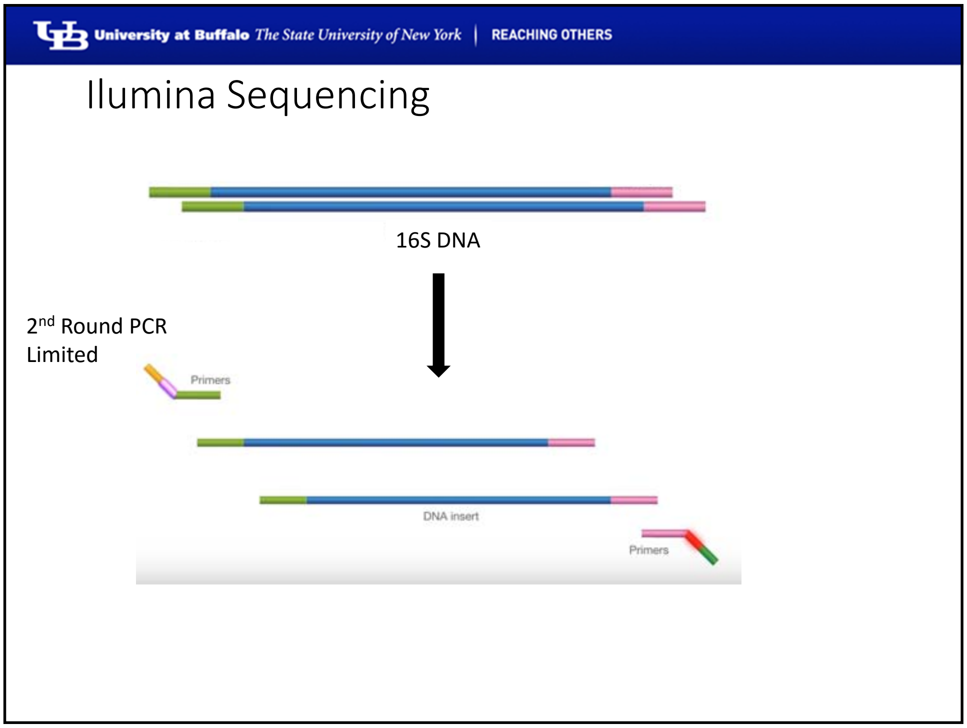
Purify DNA from biofilm

PCR amplify 16S rRNA gene

Sequence PCR product

Compare sequence to database

Identify bacterial species

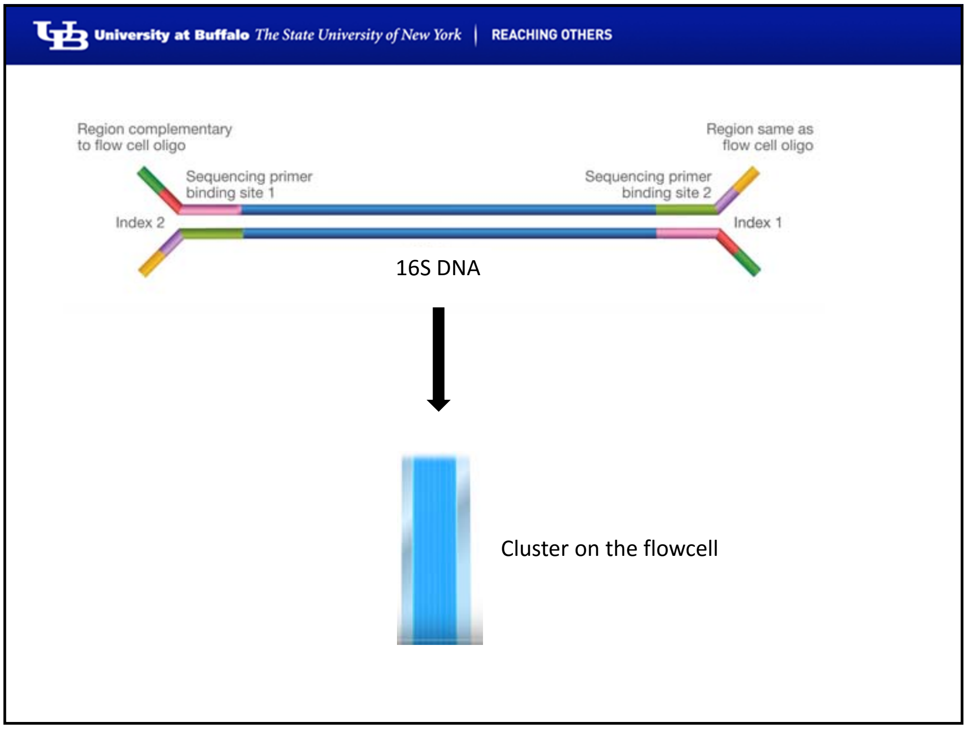


University at Buffalo *The State University of New York* | REACHING OTHERS

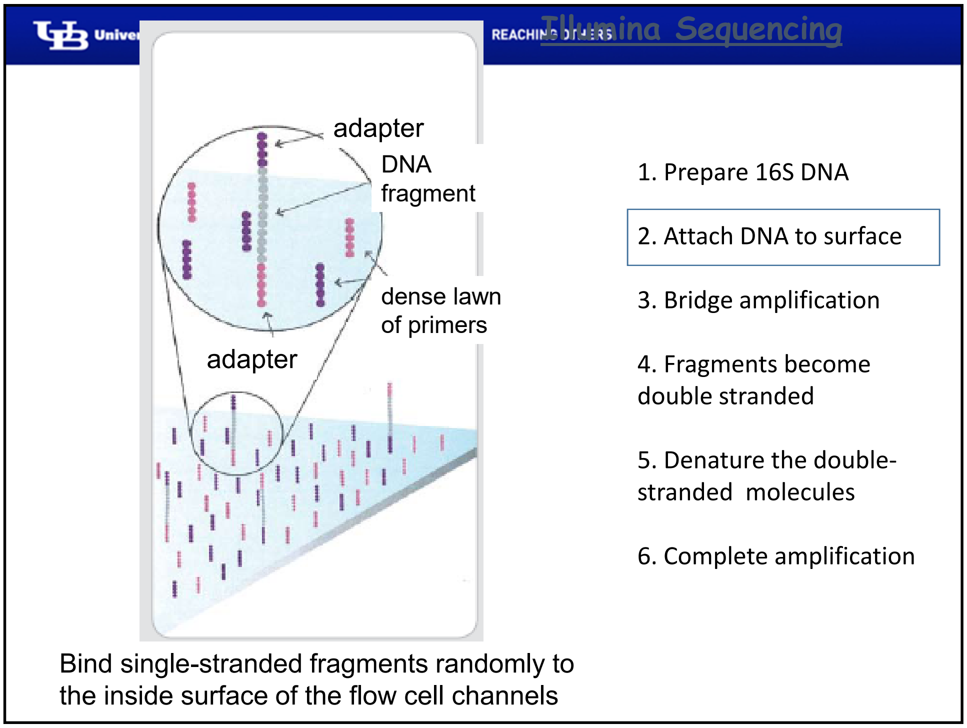
Ilumina Sequencing

16S DNA

2nd Round PCR
Limited


University at Buffalo The State University of New York | REACHING OTHERS
Region complementary to flow cell oligo
Sequencing primer binding site 1
Index 2
Region same as flow cell oligo
Sequencing primer binding site 2
Index 1
16S DNA

Cluster on the flowcell



Illumina Sequencing

1. Prepare 16S DNA

2. Attach DNA to surface

3. Bridge amplification

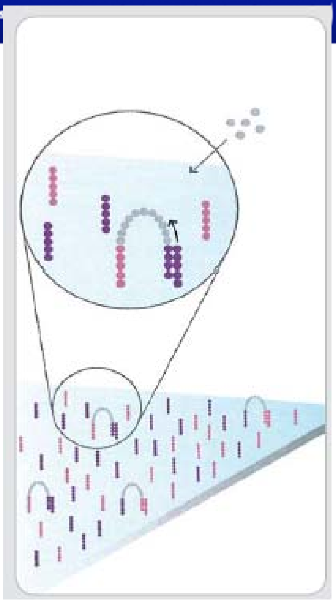
4. Fragments become double stranded

5. Denature the double-stranded molecules

6. Complete amplification

Bind single-stranded fragments randomly to the inside surface of the flow cell channels

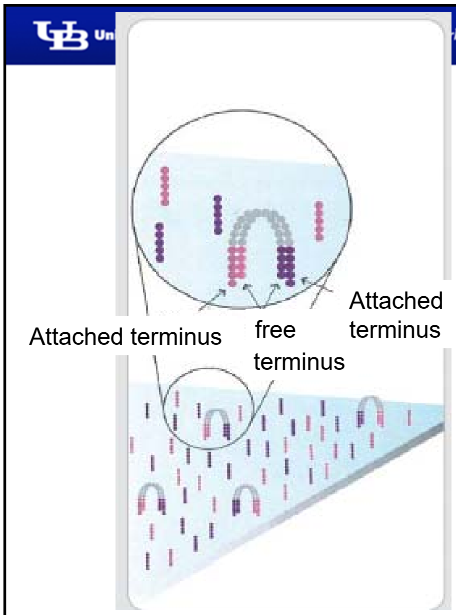
Illumina Sequencing



1. Prepare genomic DNA
2. Attach DNA to surface
3. Bridge amplification
4. Fragments become double stranded
5. Denature the double-stranded molecules
6. Complete amplification

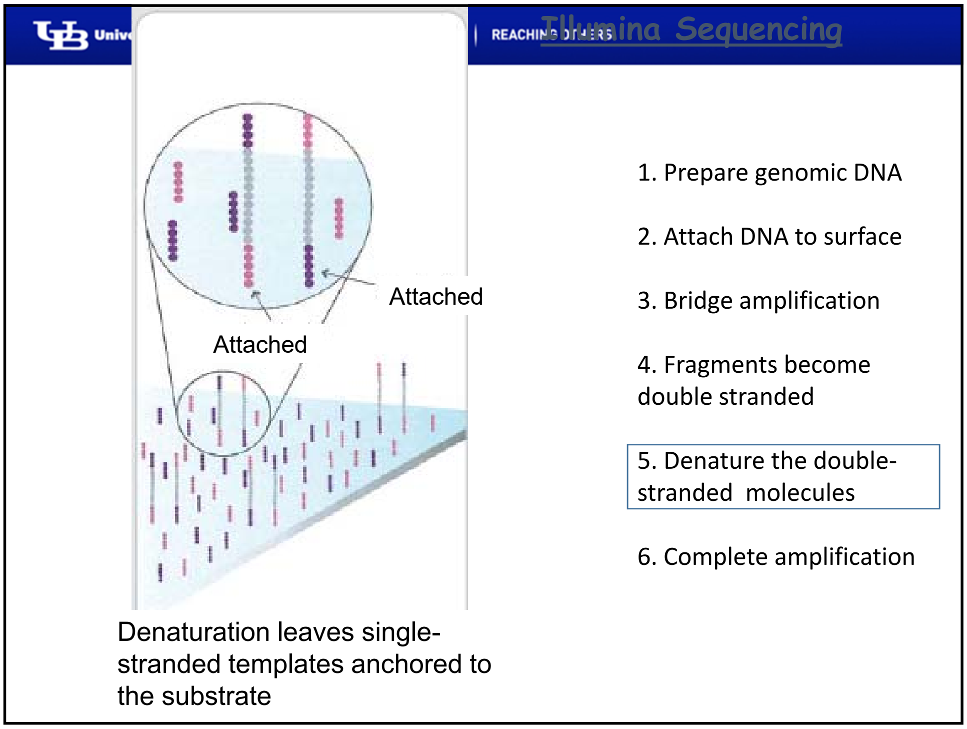
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification

Illumina Sequencing



1. Prepare genomic DNA
2. Attach DNA to surface
3. Bridge amplification
4. Fragments become double stranded
5. Denature the double-stranded molecules
6. Complete amplification

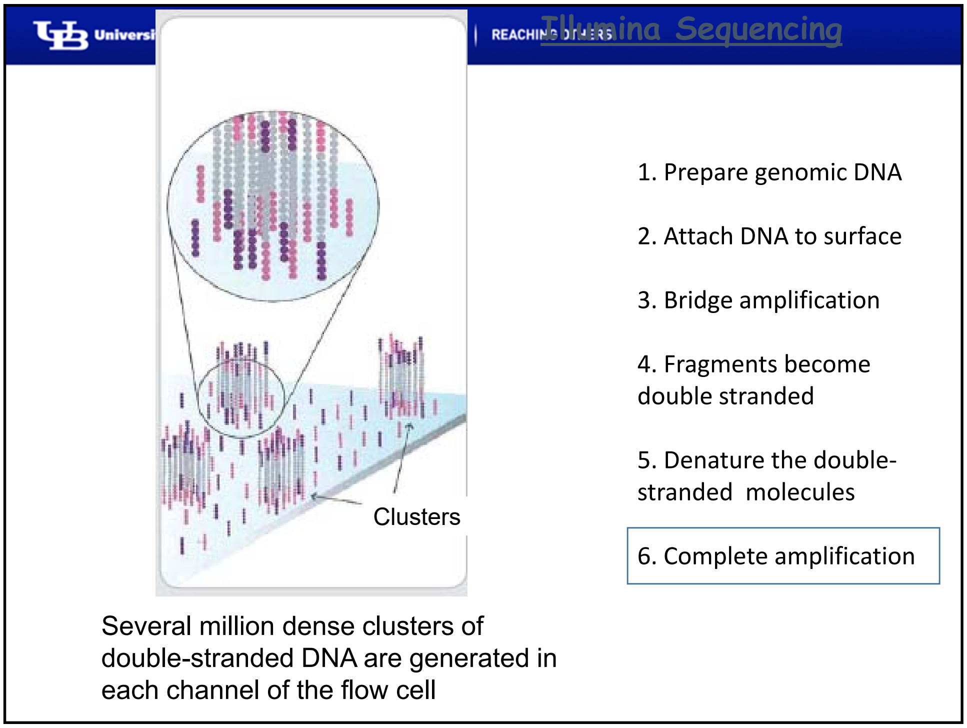
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate



Illumina Sequencing

Denaturation leaves single-stranded templates anchored to the substrate

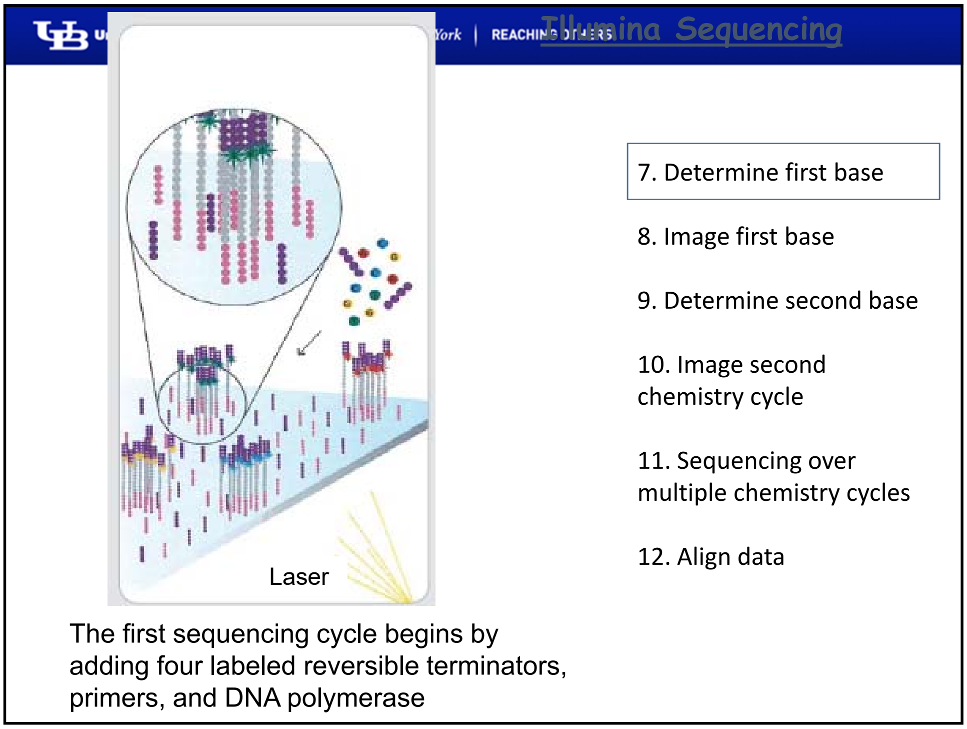
1. Prepare genomic DNA
2. Attach DNA to surface
3. Bridge amplification
4. Fragments become double stranded
5. Denature the double-stranded molecules
6. Complete amplification



Illumina Sequencing

Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell

1. Prepare genomic DNA
2. Attach DNA to surface
3. Bridge amplification
4. Fragments become double stranded
5. Denature the double-stranded molecules
6. Complete amplification



Illumina Sequencing

7. Determine first base

8. Image first base

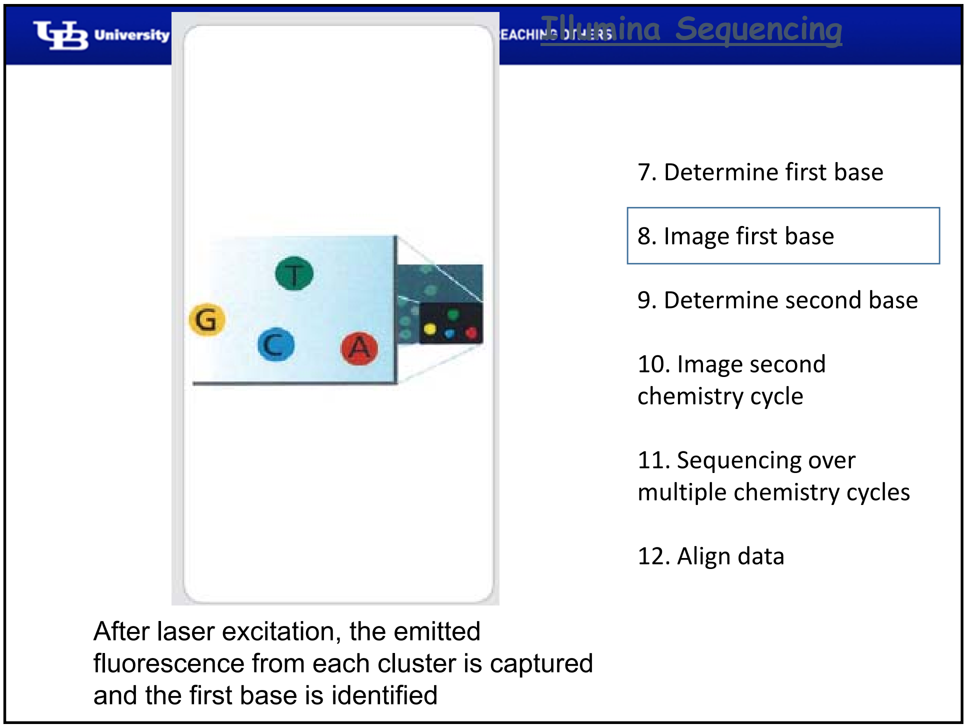
9. Determine second base

10. Image second chemistry cycle

11. Sequencing over multiple chemistry cycles

12. Align data

The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase



Illumina Sequencing

7. Determine first base

8. Image first base

9. Determine second base

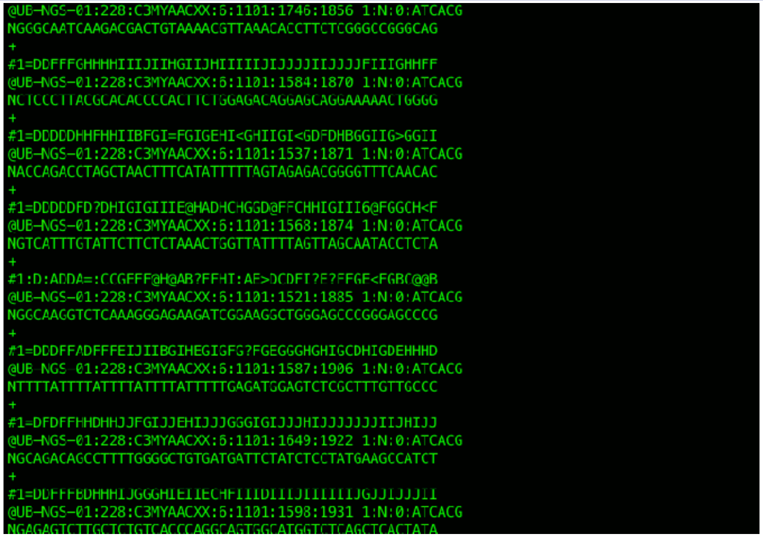
10. Image second chemistry cycle

11. Sequencing over multiple chemistry cycles

12. Align data

After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified

University at Buffalo *The State University of New York* | REACHING OTHERS



```
@UB-NGS-01:228:C3MYAACXX:6:1101:1746:1856 1:N:0:ATCACG
NGGGCAATCAAGACGACTGTAAAACGTTAAACACCTTCTCGGGCCGGGCAG
+
#1=DDFFFGHHHHIIIJIIHGIIJHIIIIIJIJJJJIIJJJJFIIIGHHFF
@UB-NGS-01:228:C3MYAACXX:6:1101:1584:1870 1:N:0:ATCACG
NCTCCCTTACGCACACCCCACTTCTGGAGACAGGAGCAGGAAAAACTGGGG
+
#1=DDDDDHHFHHIIBFGI=FGIGEHI<GHIIGI<GDFDHBGGIIG>GGII
@UB-NGS-01:228:C3MYAACXX:6:1101:1537:1871 1:N:0:ATCACG
NACCAGACCTAGCTAACTTTCATATTTTTAGTAGAGACGGGGTTTCAACAC
+
#1=DDDDDFD?DHIGIGIIIE@HADHCHGGD@FFCHHIGIII6@FGGCH<F
@UB-NGS-01:228:C3MYAACXX:6:1101:1568:1874 1:N:0:ATCACG
NGTCATTTGTATTCTTCTCTAAACTGGTTATTTTAGTTAGCAATACCTCTA
+
#1:D:ADDA=:CCGFFF@H@AB?FFHI:AE>DCDFI?E?FFGE<FGBC@@B
@UB-NGS-01:228:C3MYAACXX:6:1101:1521:1885 1:N:0:ATCACG
NGGCAAGGTCTCAAAGGGAGAAGATCGGAAGGCTGGGAGCCCGGGAGCCCG
+
#1=DDDFFADFFFEIJIIBGIHEGIGFG?FGEGGGHGHIGCDHIGDEHHHD
@UB-NGS-01:228:C3MYAACXX:6:1101:1587:1906 1:N:0:ATCACG
NTTTTATTTTATTTTATTTTATTTTTGAGATGGAGTCTCGCTTTGTTGCCC
+
#1=DFDFFHHDHHJJFGIJJEHIJJJGGGIGIJJJHIJJJJJJJJIIJHIJJ
@UB-NGS-01:228:C3MYAACXX:6:1101:1649:1922 1:N:0:ATCACG
NGCAGACAGCCTTTTGGGGCTGTGATGATTCTATCTCCTATGAAGCCATCT
+
#1=DDFFFEDHHHIJGGGHIEIIECHFIIIDIIIJIIIIIIIJGJJIJJJII
@UB-NGS-01:228:C3MYAACXX:6:1101:1598:1931 1:N:0:ATCACG
NGAGAGTCTTGCTCTGTCACCCAGGCAGTGGCATGGTCTCAGCTCACTATA
```

University at Buffalo *The State University of New York* | REACHING OTHERS

UB NSG Core

Illumina HiSeq 2500

8 flow cell lanes with 2 flow cells
500-1000 Gb
2 billion reads per flow cell
2 x 250 bp max read length

MiSeq

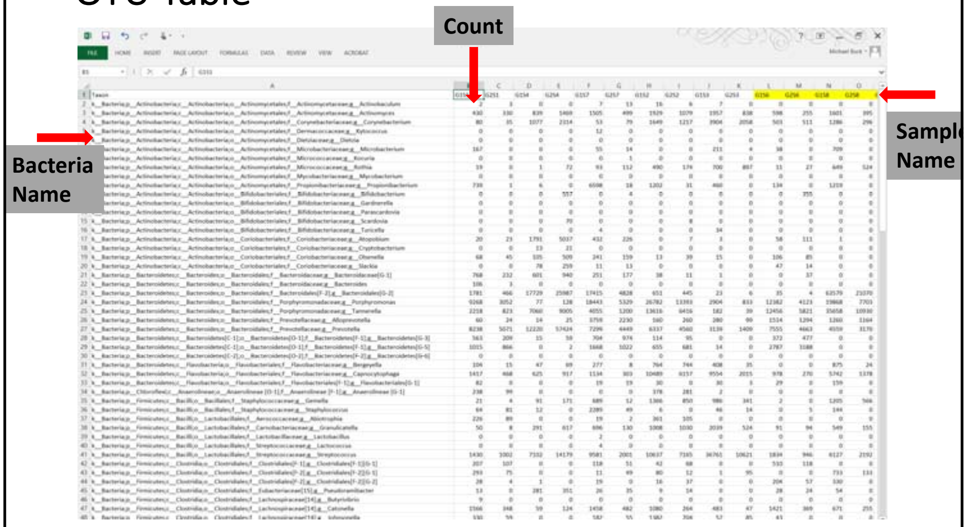
1 flow cell lane
0.3 – 15 Gb
25 million reads
2 x 300 max read length
- 1 billion 35-100 bp

Sample analysis

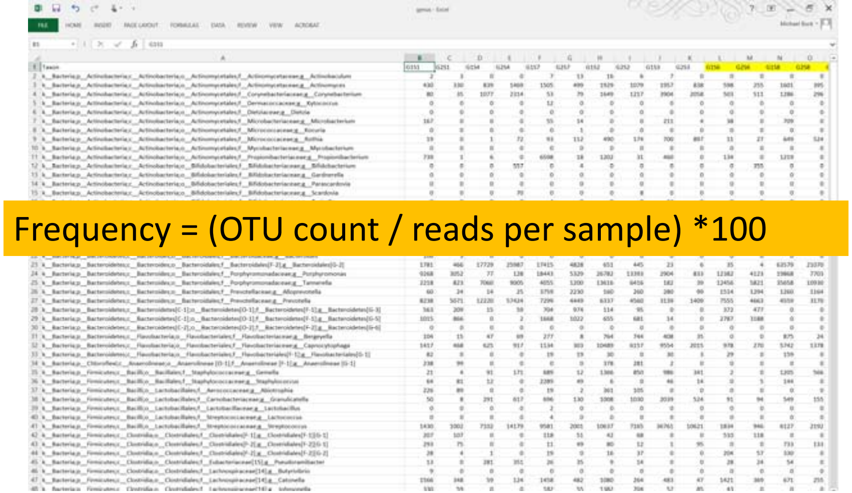
- Quality filter -- Was the sequencing good
- Paired-end sequence joining
- OTU calling
    - Reference-based
    - Non-reference based

- **Reference-based analysis will change over time as databases are updated.**

OTU Table



Counts need to converted into relative frequency



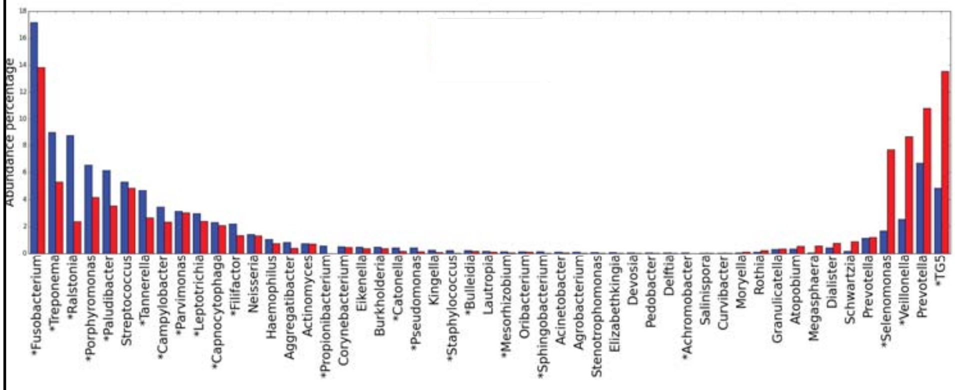
Frequency = (OTU count / reads per sample) *100

Total # of Reads per sample

Number of reads per sample will vary due to sequencing

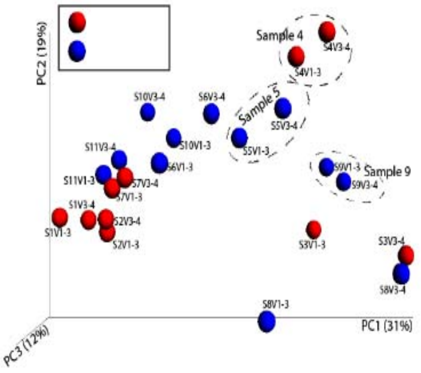
>>>> Each sample needs to be normalized to each other <<<<<

Statistical tests can then be performed



University at Buffalo The State University of New York | REACHING OTHERS

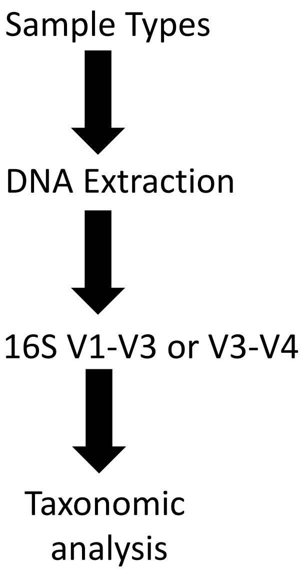
Advanced analysis



Principal components analysis of V1-V3 and V3-V4 amplicons

University at Buffalo The State University of New York | REACHING OTHERS

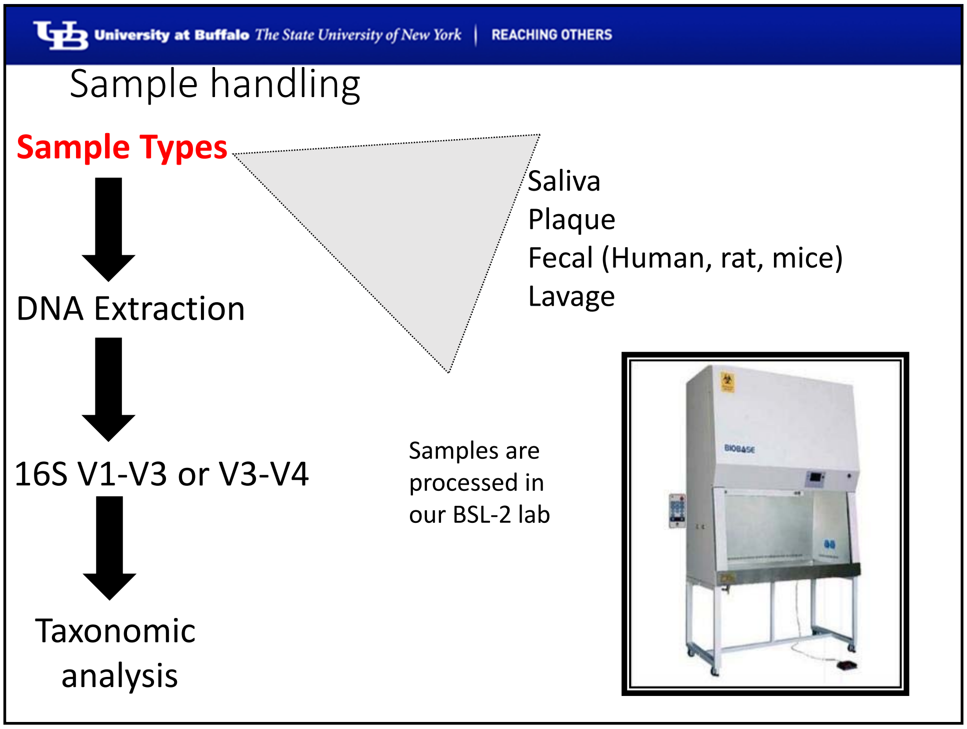
How we do it in the NGS core


Sample Types

DNA Extraction

16S V1-V3 or V3-V4

Taxonomic analysis



University at Buffalo The State University of New York | REACHING OTHERS

Sample handling

Sample Types

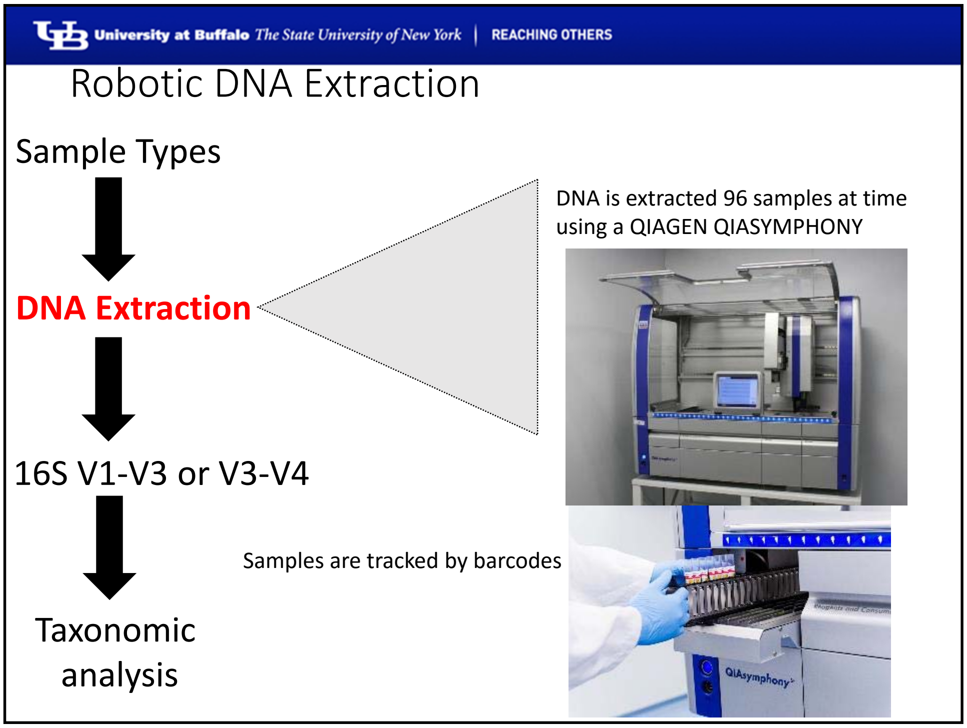
DNA Extraction

16S V1-V3 or V3-V4

Taxonomic analysis

Saliva
Plaque
Fecal (Human, rat, mice)
Lavage

Samples are processed in our BSL-2 lab



University at Buffalo The State University of New York | REACHING OTHERS

Robotic DNA Extraction

Sample Types

DNA Extraction

16S V1-V3 or V3-V4

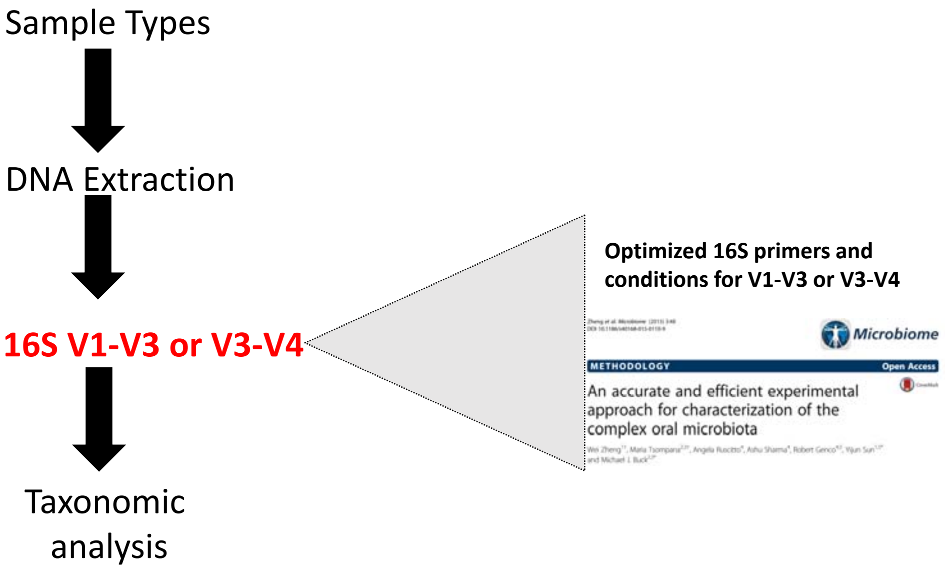
Taxonomic analysis

DNA is extracted 96 samples at time using a QIAGEN QIASYMPHONY

Samples are tracked by barcodes

University at Buffalo *The State University of New York* | REACHING OTHERS

16S amplification



Sample Types

DNA Extraction

16S V1-V3 or V3-V4

Taxonomic analysis

Optimized 16S primers and conditions for V1-V3 or V3-V4

Zheng et al. Microbiome (2015) 3:48
DOI 10.1186/s40168-015-0110-9

Microbiome

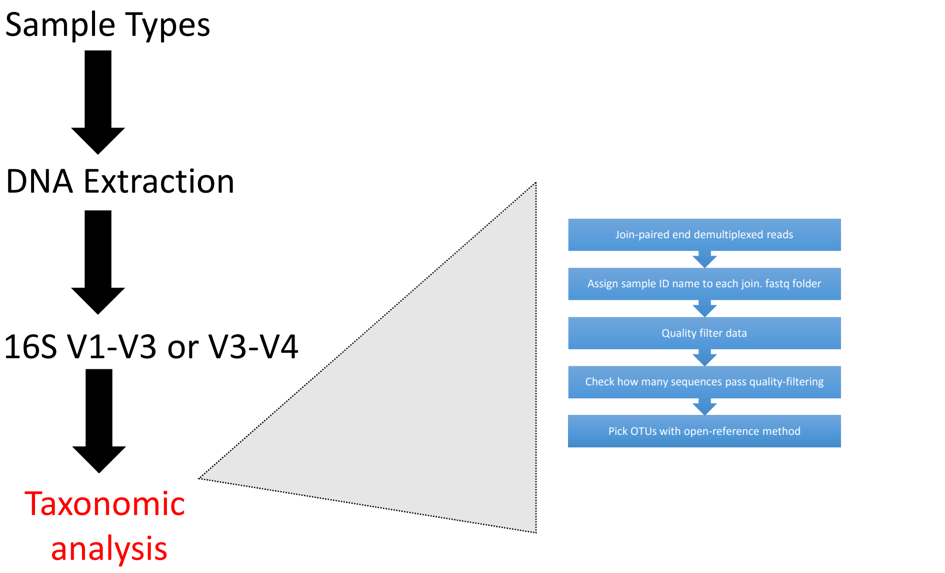
METHODOLOGY — Open Access

An accurate and efficient experimental approach for characterization of the complex oral microbiota

Wei Zheng[1†], Maria Tsompana[2,3†], Angela Ruscitto[4], Ashu Sharma[4], Robert Genco[4,5], Yijun Sun[1,5*] and Michael J. Buck[2,3*]

University at Buffalo *The State University of New York* | REACHING OTHERS

OTU Calling



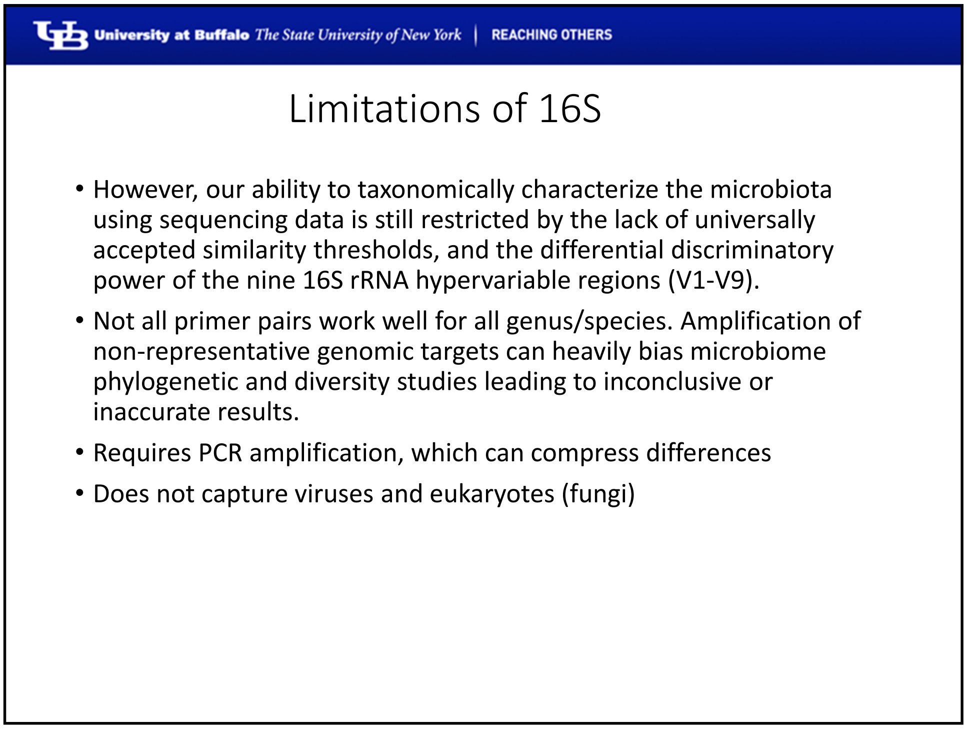
Sample Types

DNA Extraction

16S V1-V3 or V3-V4

Taxonomic analysis

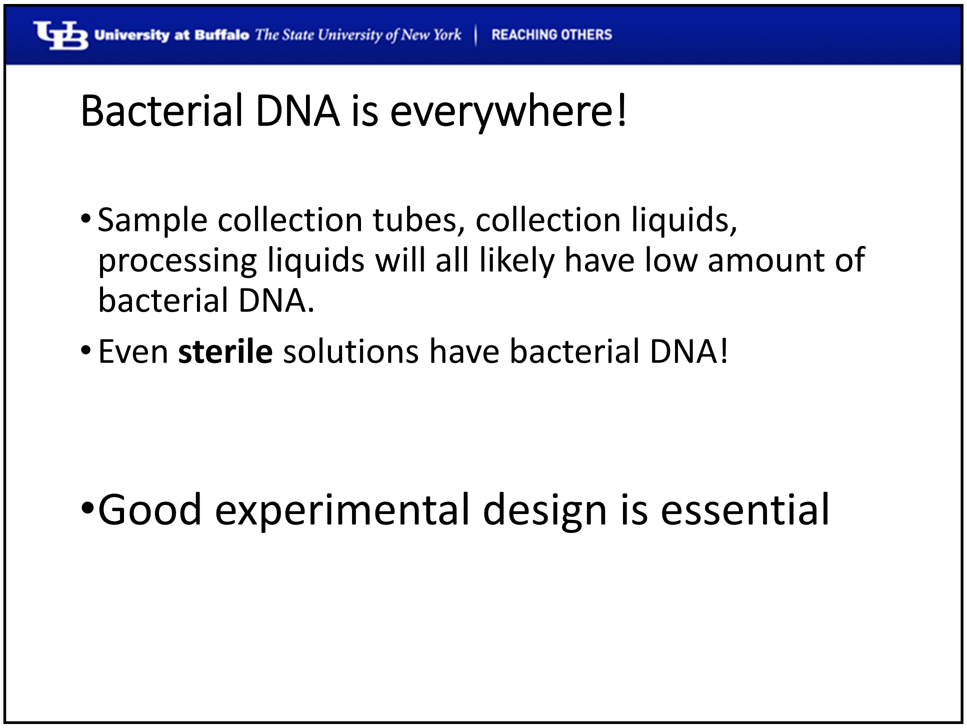
1. Join-paired end demultiplexed reads
2. Assign sample ID name to each join. fastq folder
3. Quality filter data
4. Check how many sequences pass quality-filtering
5. Pick OTUs with open-reference method



University at Buffalo *The State University of New York* | REACHING OTHERS

Limitations of 16S

- However, our ability to taxonomically characterize the microbiota using sequencing data is still restricted by the lack of universally accepted similarity thresholds, and the differential discriminatory power of the nine 16S rRNA hypervariable regions (V1-V9).
- Not all primer pairs work well for all genus/species. Amplification of non-representative genomic targets can heavily bias microbiome phylogenetic and diversity studies leading to inconclusive or inaccurate results.
- Requires PCR amplification, which can compress differences
- Does not capture viruses and eukaryotes (fungi)

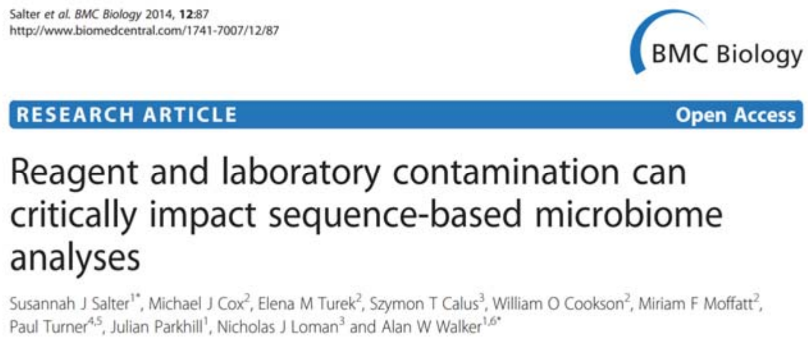


University at Buffalo *The State University of New York* | REACHING OTHERS

Bacterial DNA is everywhere!

- Sample collection tubes, collection liquids, processing liquids will all likely have low amount of bacterial DNA.
- Even **sterile** solutions have bacterial DNA!

- Good experimental design is essential



Salter *et al. BMC Biology* 2014, **12**:87
http://www.biomedcentral.com/1741-7007/12/87

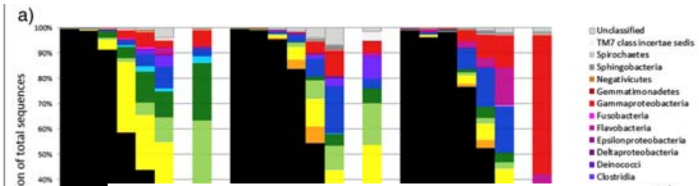
BMC Biology

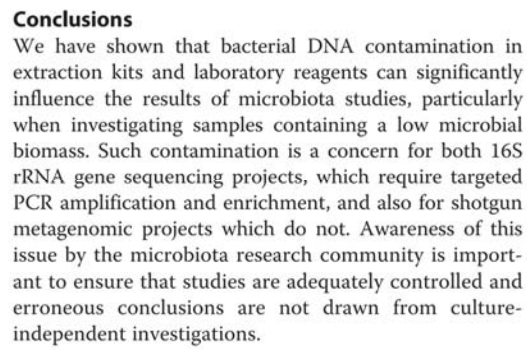
RESEARCH ARTICLE Open Access

Reagent and laboratory contamination can critically impact sequence-based microbiome analyses

Susannah J Salter[1*], Michael J Cox[2], Elena M Turek[2], Szymon T Calus[3], William O Cookson[2], Miriam F Moffatt[2], Paul Turner[4,5], Julian Parkhill[1], Nicholas J Loman[3] and Alan W Walker[1,6*]

- Sequence a pure culture of *Salmonella bongori*
- Extracted DNA using different kits
- Did serial dilutions of the pure culture to assess impact of contaminating species





Conclusions

We have shown that bacterial DNA contamination in extraction kits and laboratory reagents can significantly influence the results of microbiota studies, particularly when investigating samples containing a low microbial biomass. Such contamination is a concern for both 16S rRNA gene sequencing projects, which require targeted PCR amplification and enrichment, and also for shotgun metagenomic projects which do not. Awareness of this issue by the microbiota research community is important to ensure that studies are adequately controlled and erroneous conclusions are not drawn from culture-independent investigations.

University at Buffalo *The State University of New York* | REACHING OTHERS

Acknowledgments

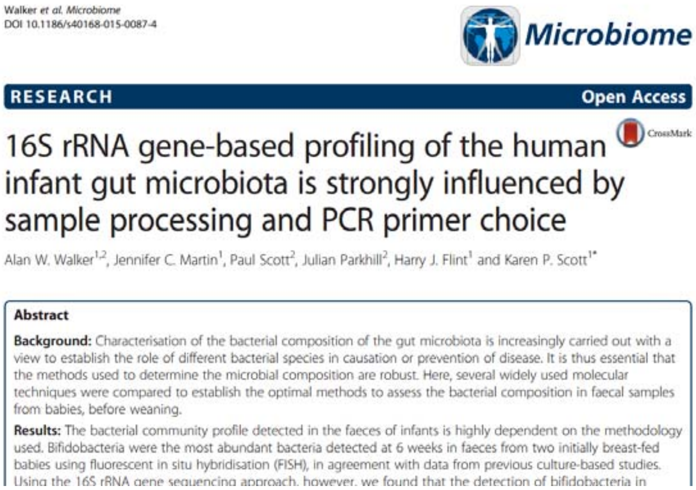
UB Genomics & Bioinformatics Facility

Norma Nowak
Maria Tsompana
Sujith Valiyaparambil
Natalie Waldron
Jonathan Bard
Brandon Marzullo

WHI Microbiome Team

Jean Wactawski-Wende
Robert Genco
Mike Lamonte
Amy Millen
Chris Andrews
Jo Freudenheim
Yijun Sun
Karen Falkner
Kathy Hovey
Wei Zheng
Xiaodan Mai

University at Buffalo *The State University of New York* | REACHING OTHERS



Walker *et al. Microbiome*
DOI 10.1186/s40168-015-0087-4

Microbiome

RESEARCH Open Access

16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice

Alan W. Walker[1,2], Jennifer C. Martin[1], Paul Scott[2], Julian Parkhill[2], Harry J. Flint[1] and Karen P. Scott[1*]

Abstract

Background: Characterisation of the bacterial composition of the gut microbiota is increasingly carried out with a view to establish the role of different bacterial species in causation or prevention of disease. It is thus essential that the methods used to determine the microbial composition are robust. Here, several widely used molecular techniques were compared to establish the optimal methods to assess the bacterial composition in faecal samples from babies, before weaning.

Results: The bacterial community profile detected in the faeces of infants is highly dependent on the methodology used. Bifidobacteria were the most abundant bacteria detected at 6 weeks in faeces from two initially breast-fed babies using fluorescent in situ hybridisation (FISH), in agreement with data from previous culture-based studies. Using the 16S rRNA gene sequencing approach, however, we found that the detection of bifidobacteria in

University at Buffalo *The State University of New York* | REACHING OTHERS

He *et al. Microbiome* (2015) 3:20
DOI 10.1186/s40168-015-0081-x

RESEARCH Open Access

Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity

Yan He[1], J Gregory Caporaso[2,3], Xiao-Tao Jiang[1], Hua-Fang Sheng[1], Susan M Huse[4], Jai Ram Rideout[3], Robert C Edgar[5], Evguenia Kopylova[6], William A Walters[7], Rob Knight[6,8] and Hong-Wei Zhou[1*]

Abstract

Background: The operational taxonomic unit (OTU) is widely used in microbial ecology. Reproducibility in microbial ecology research depends on the reliability of OTU-based 16S ribosomal subunit RNA (rRNA) analyses.

Results: Here, we report that many hierarchical and greedy clustering methods produce unstable OTUs, with membership that depends on the number of sequences clustered. If OTUs are regenerated with additional sequences or samples, sequences originally assigned to a given OTU can be split into different OTUs. Alternatively, sequences assigned to different OTUs can be merged into a single OTU. This OTU instability affects alpha-diversity analyses such as rarefaction curves, beta-diversity analyses such as distance-based ordination (for example, Principal

University at Buffalo *The State University of New York* | REACHING OTHERS