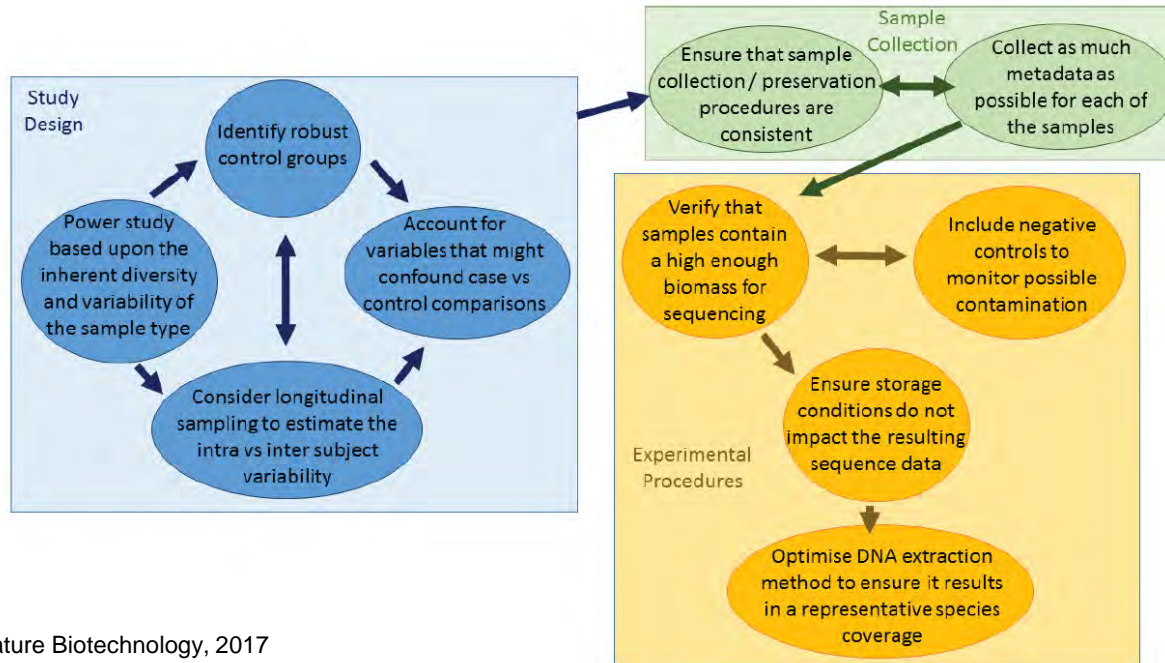


Metagenomics and Metatranscriptomics

Johnny Sena, Ph.D.

Example Workflow to plan a Metagenomics Study



Understanding the potential for confounding factors, and optimization of design, can substantially improve the quality of both metagenomic sequence data, and interpretation

Nature Biotechnology, 2017

Sample collection and DNA extraction

- Sample collection and preservation methods can affect quality and accuracy of metagenomic data
 - Collect sufficient biomass
 - Minimize contamination
 - Enrichment methods where applicable
- DNA extraction methods can affect the composition of downstream sequence data
 - Method must be effective for diverse microbial taxa
 - Mechanical lysis (bead beating) method is considered superior, however, data will be biased for easy-to-lyse microbes
 - Bead beating will result in short DNA fragments and lead to DNA loss during library prep methods.

Sources of contamination

- Kit or lab reagents
- Low biomass samples are vulnerable to contamination as there is less 'real' signal to compete with low levels of contamination
 - Use ultraclean kits
 - Include blank sequencing controls
- Cross- over from previous sequencing runs
- PhiX control DNA
- Human/ host DNA

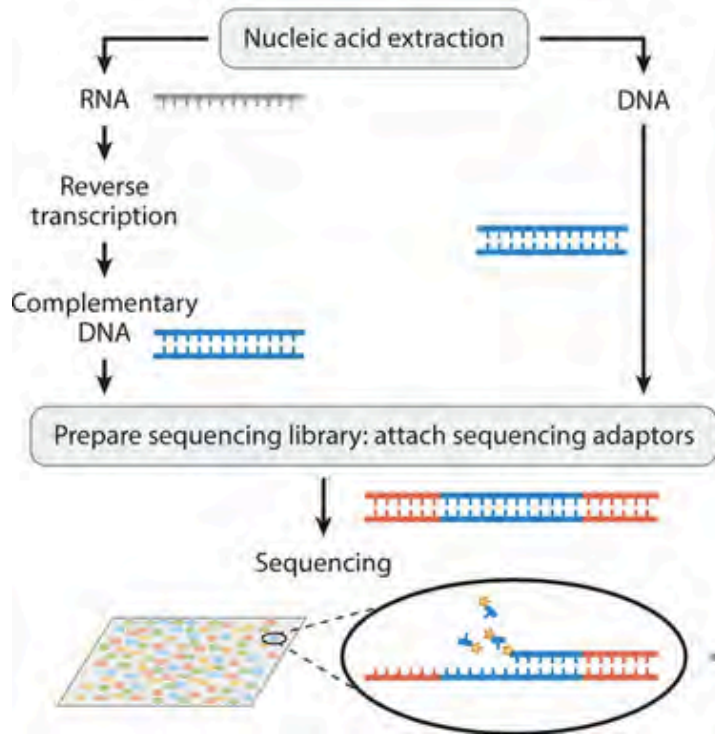
Coverage and Sequencing considerations

- No published guidelines for ‘correct’ amount of coverage for a given environment
 - Choose a system that maximizes output in order to recover sequences from as many low-abundance members of the microbiome as possible
 - HiSeq 2500 or 4000, NextSeq and NovaSeq produce high volume data (120Gb- 1.5 Tb per run) – suited for metagenomics study
 - Multiplexing prudently will enable desired per-sample sequencing depth

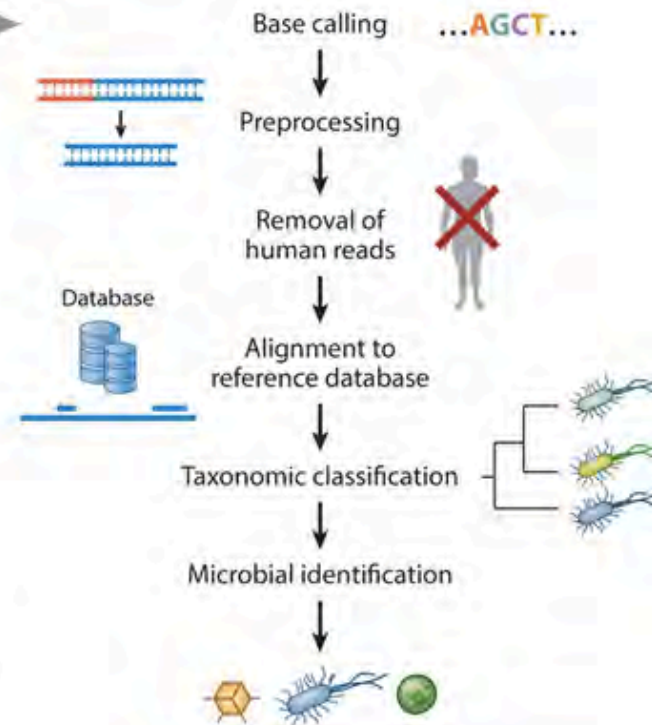
Illumina sequencers and yield

	platform	read config	output
Production scale	HiSeq 2500	2 x 250	180 Gb -1 Tb
	HiSeq 4000	2 x 150	1.5 Tb
	HiSeq X	2 x 150	1.8 Tb
	NovaSeq	2 x 250	6 Tb
benchtop	NextSeq	2 x 150	120 Gb
	MiSeq	2 x 300	15 Gb
	Iseq	2 x 150	1.2 Gb
	MiniSeq	2 x 150	7.5 Gb

Wet lab pipeline



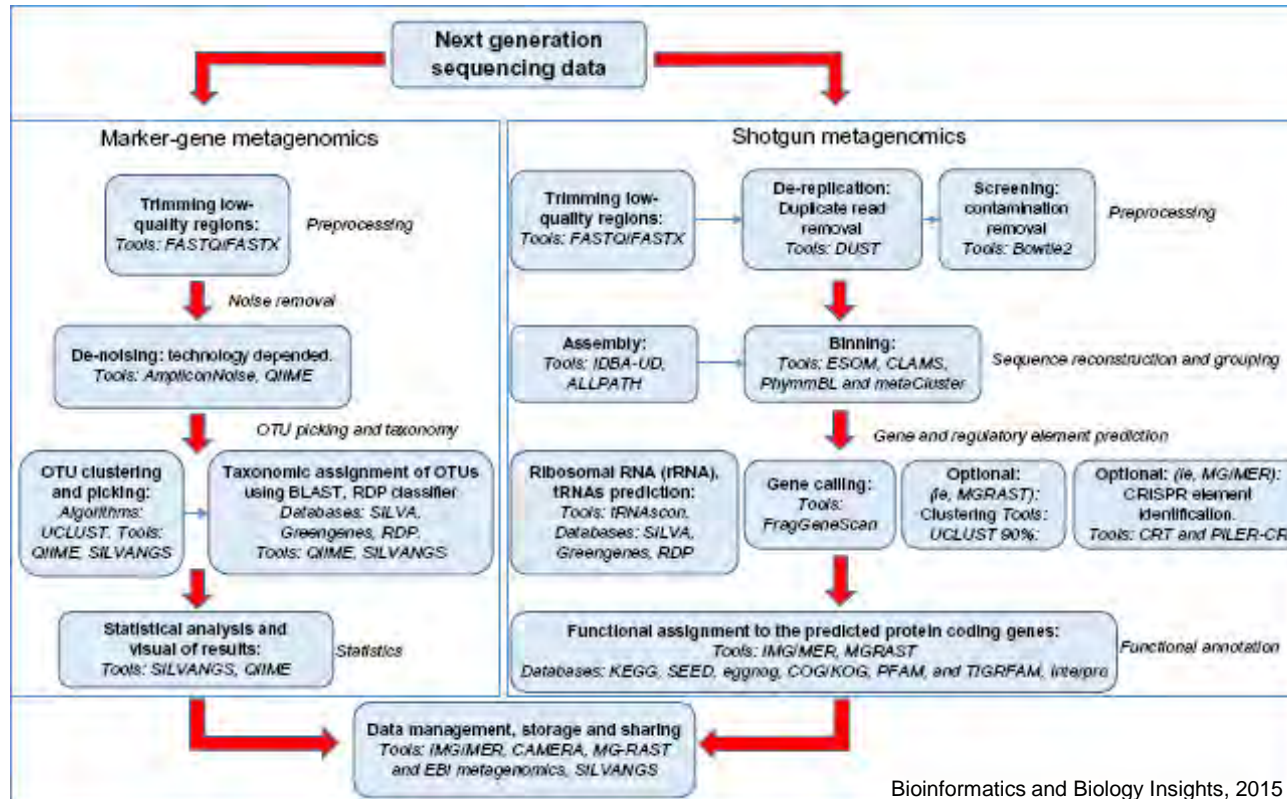
Dry lab (informatics) pipeline



Annu Rev Pathol. 2019

Generalized workflow of metagenomic next-generation sequencing for diagnostic clinical use

Generic Analysis Workflow



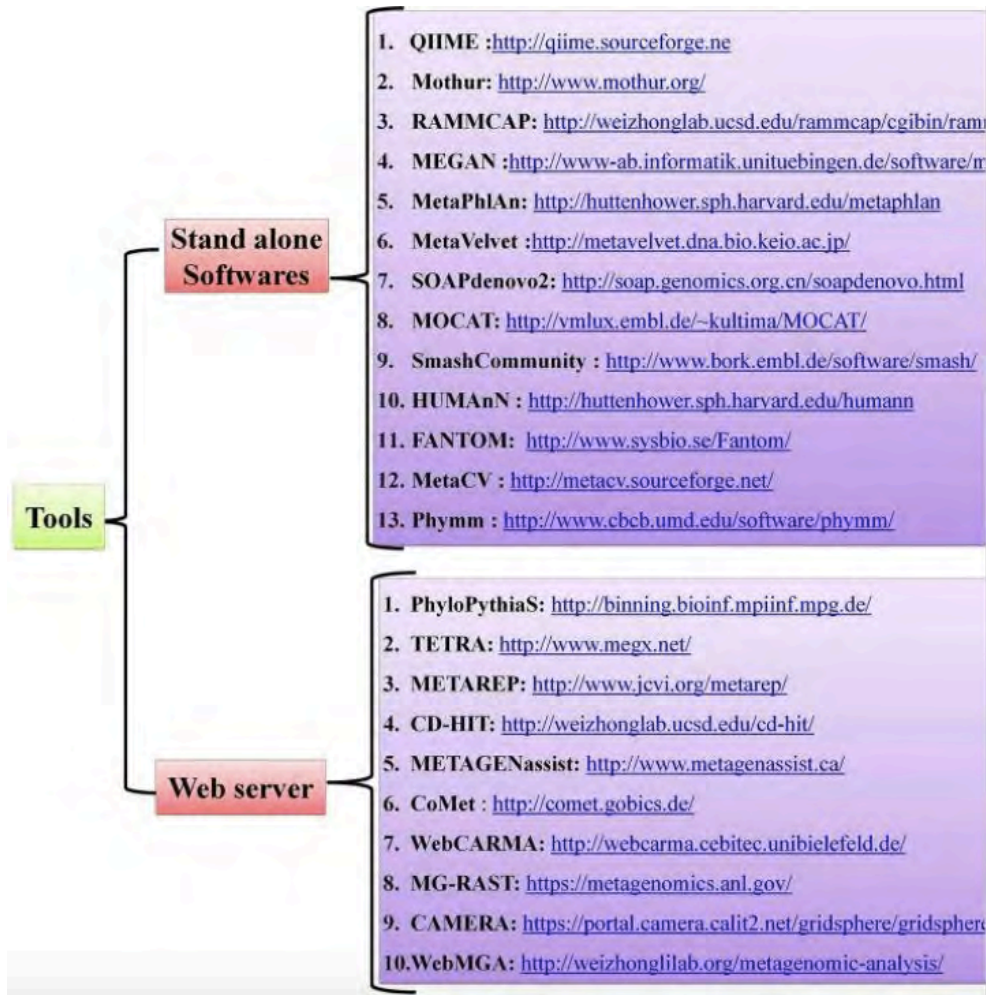
Bioinformatics and Biology Insights, 2015

Strengths and weaknesses of assembly-based and read-based metagenomics analysis

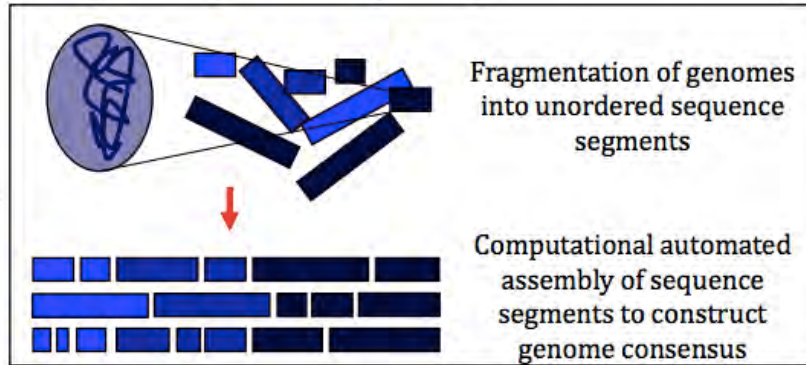
	Assembly-based analysis	Read-based analysis ('mapping')
Comprehensiveness	Can construct multiple whole genomes, but only for organisms with enough coverage to be assembled and binned.	Can provide an aggregate picture of community function or structure, but is based only on the fraction of reads that map effectively to reference databases.
Community complexity	In complex communities, only a fraction of the genomes can be resolved by assembly.	Can deal with communities of arbitrary complexity given sufficient sequencing depth and satisfactory reference database coverage
Novelty	Can resolve genomes of entirely novel organisms with no sequenced relatives.	Cannot resolve organisms for which genomes of close relatives are unknown.
Computational burden	Requires computationally costly assembly, mapping and binning.	Can be performed efficiently, enabling large meta-analyses.
Genome-resolved metabolism	Can link metabolism to phylogeny through completely assembled genomes, even for novel diversity.	Can typically resolve only the aggregate metabolism of the community, and links with phylogeny are only possible in the context of known reference genomes.
Expert manual supervision	Manual curation required for accurate binning and scaffolding and for misassembly detection.	Usually does not require manual curation, but selection of reference genomes to use could involve human supervision.
Integration with microbial genomics	Assemblies can be fed into microbial genomic pipelines designed for analysis of genomes from pure cultured isolates.	Obtained profiles cannot be directly put into the context of genomes derived from pure cultured isolates.

Nature Biotechnol, 2017

Tools for analysis



Benefits and limitations of whole genome metagenomics



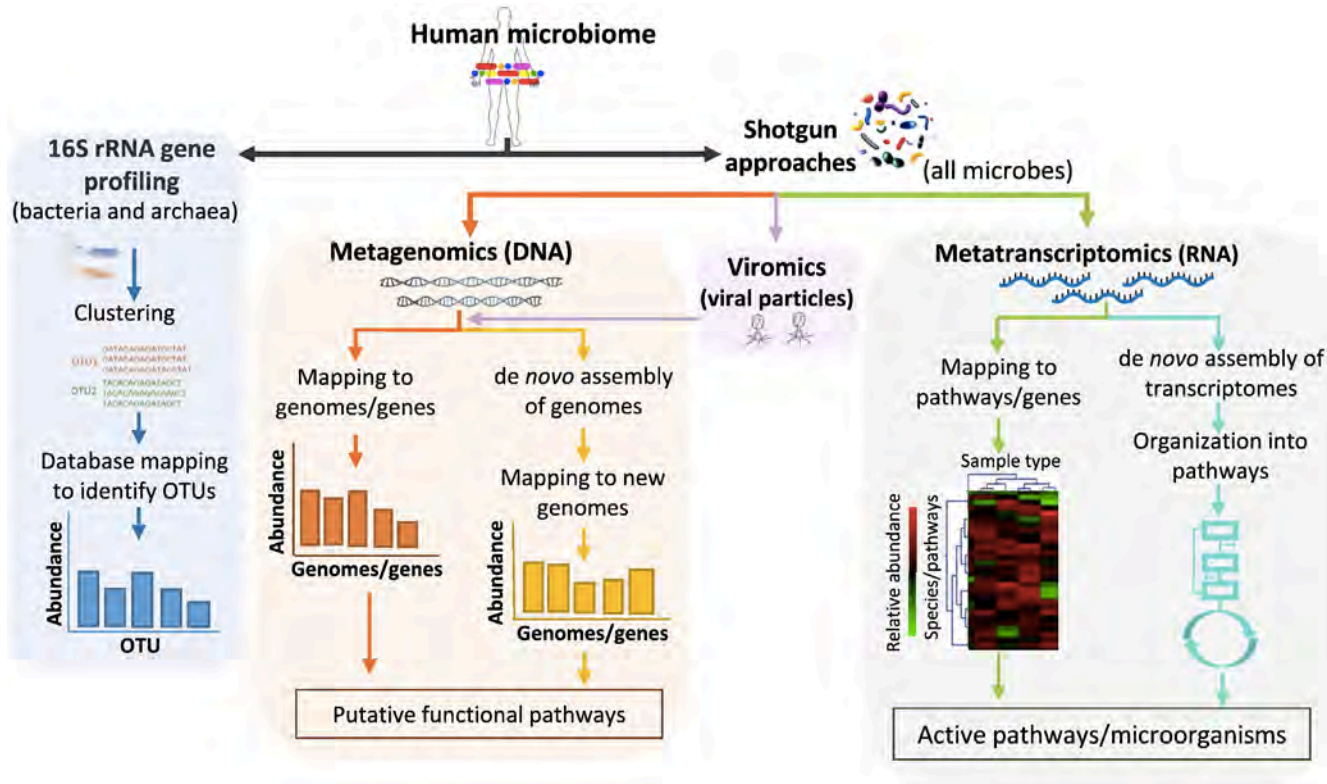
Benefits

- Integrative meta-omics
- Strain-level profiling
- Longitudinal study design
- Capability of sequencing large regions or entire genome
- Identification of organisms in addition to bacteria, archaea
- Increased prediction of genes and functional pathways

Limitations

- Expensive
- Compute intensive
- Incomplete databases
- Biases in functional profiling
- Unvalidated data in the public space
- Live or dead dilemma

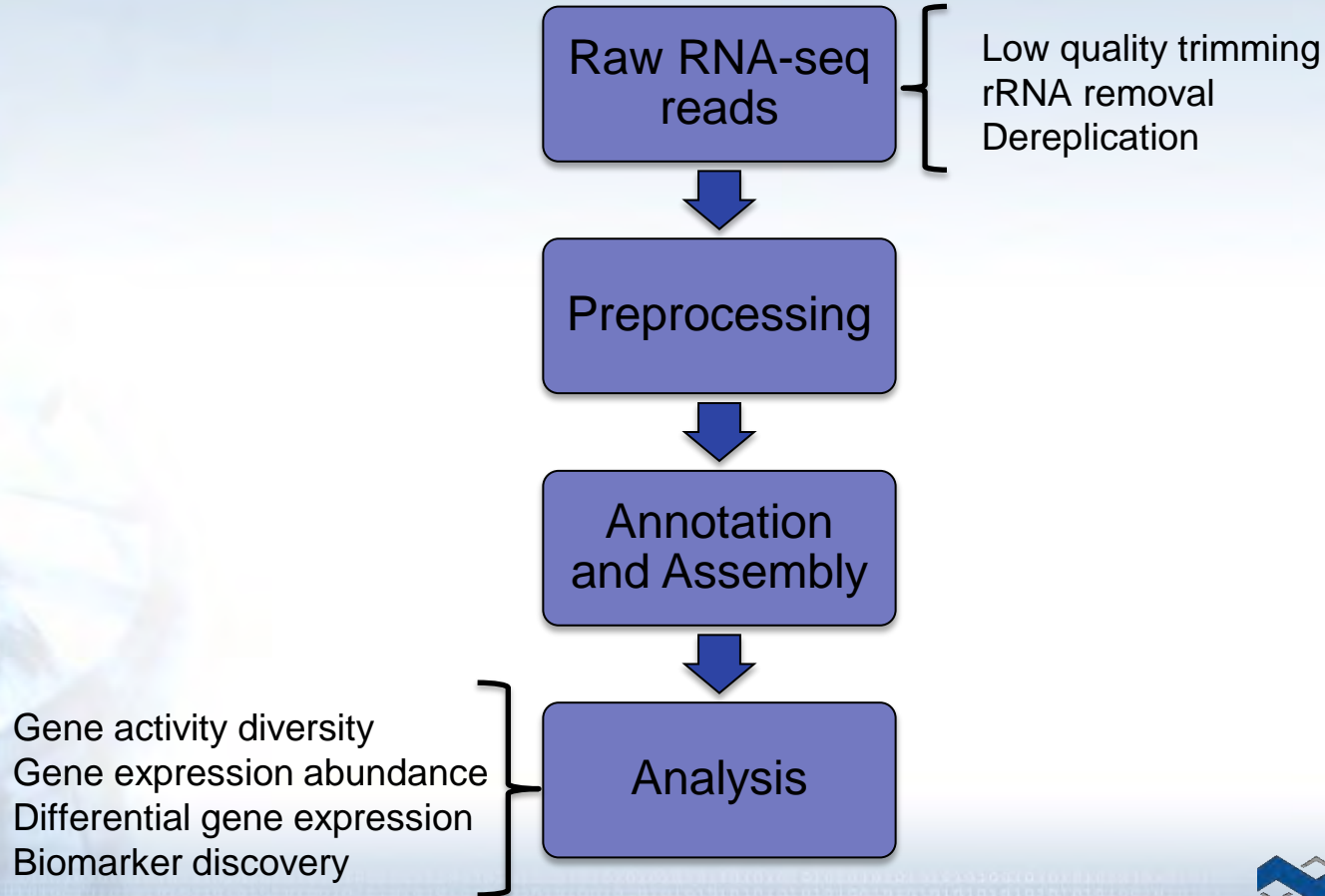
Metatranscriptomics



“What are they doing?”
- Metatranscriptomics

Computational and Structural Biotechnology Journal, 2015

Simplistic Workflow



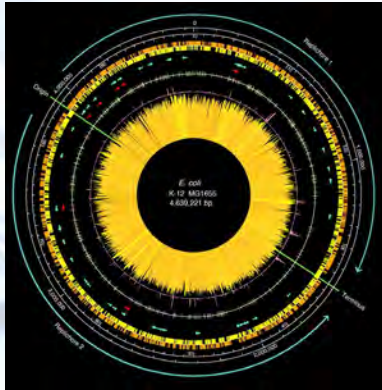
Benefits of Metatranscriptomics

16S Sequencing	Metagenomic analysis	Metatranscriptome analysis
Identifies only a fraction of your gut bacteria; unable to identify nonbacterial microorganisms	cannot identify any RNA viruses or RNA bacteriophages	identifies all microorganisms living in the environment: bacteria, viruses, archaea, yeast, fungi, parasites and bacteriophages
Low resolution (mostly genus or lower)	High resolution (species and strain level), but does not include RNA viruses	High resolution (species and strains) of all microorganisms
Unreliable; sequencing the same sample twice can yield very different results	Minimal variation in results, but partially biased analysis (no RNA data)	Minimal variation in results and unbiased results
Does not measure microbe functions	Does not measure microbe functions	capable of providing functional information
unable to identify microbial metabolites, which are key for maintaining health	unable to identify microbial metabolites, which are key for maintaining health	identifies which metabolites are being produced and which are missing
Sequences DNA, which can come from say food or dead organisms	Sequences DNA, which can come from say food or dead organisms	Sequences RNA, which comes from live microorganisms
low resolution and lack of functional data preclude any actionable recommendations (for therapeutic purposes)	low resolution and lack of functional data preclude any actionable recommendations (for therapeutic purposes)	Allows correlation of microbes and their functions with common chronic conditions, so actionable recommendations can be made

*biggest challenge metatranscriptomics faces: removal of ribosomal RNA



From Genomes to Metagenomes



E. coli, 1997, Science

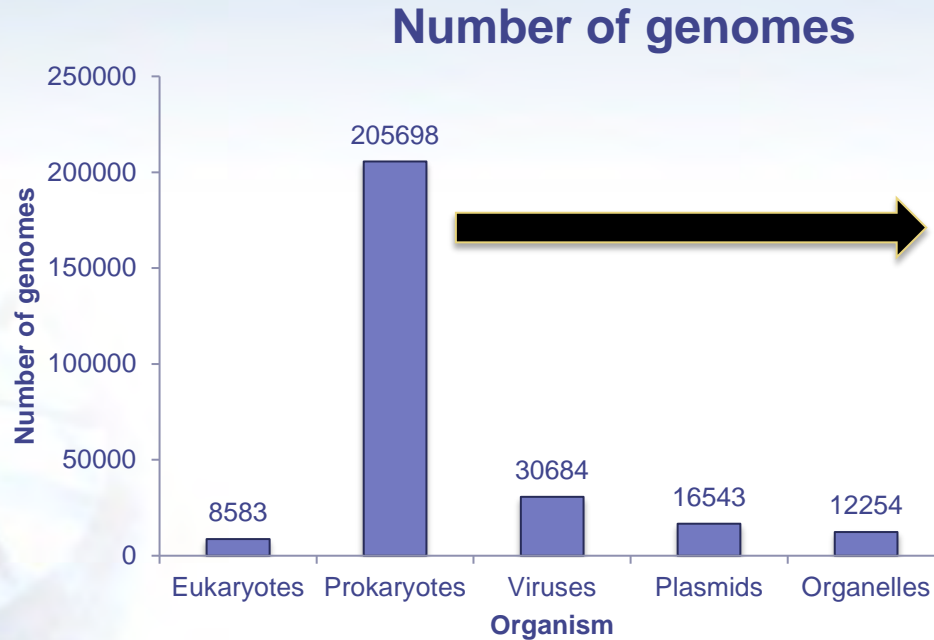


Human, 2001, Nature/Science,
project completed 2003



100,000th human genome
at Broad, Apr 2018

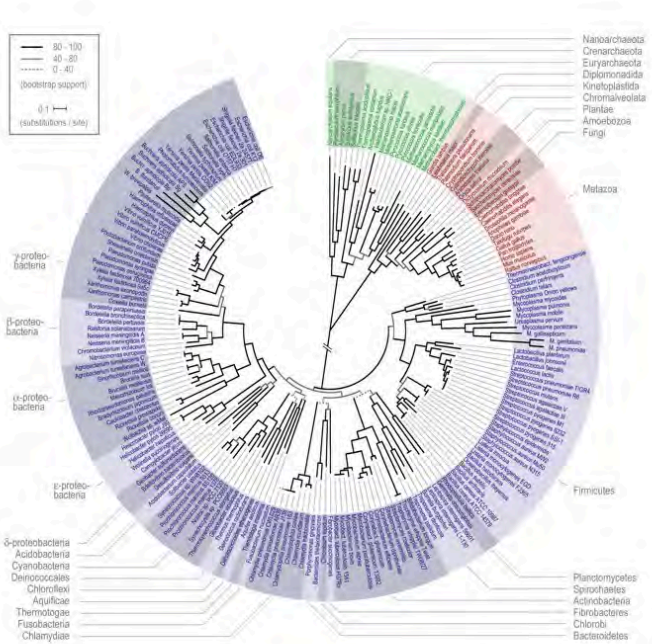
Total number of genomes at NCBI



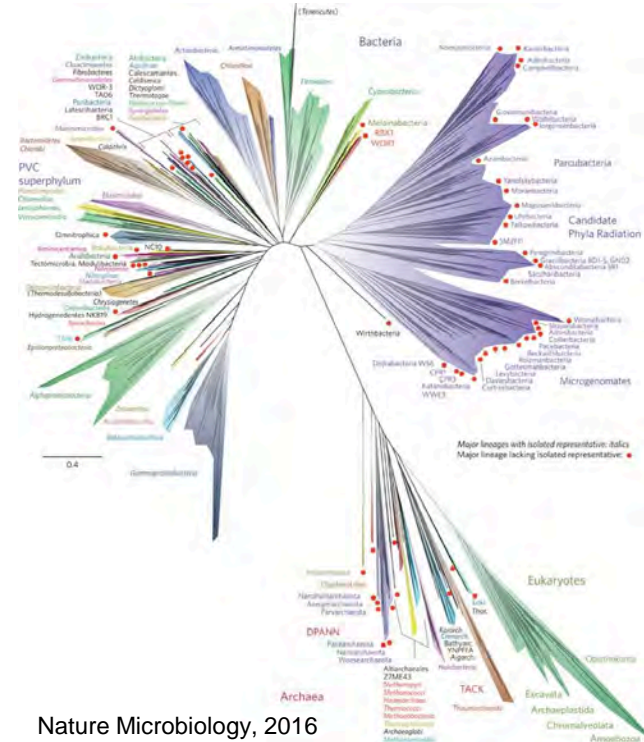
- Haploid genome
- Single circular chromosome, plasmids
- Metabolic diversity
- Genetic malleability
- No nucleus
- Easy interspecies gene transfer

<https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/>

Tree of Life.....just got weird!



Science, 2006



Nature Microbiology, 2016

“Visible organisms represent the smallest sliver of life’s diversity. Bacteria are the true lords of the world. They have been on this planet for billions of years and have irrevocably changed it, while diversifying into endless forms most wonderful and most beautiful.”(The Atlantic)

Microbiomes and their significance

- Microbes do not work or function as a single entity
- Most microbial activities are performed by complex communities of microorganisms - microbiome

What is a microbiome

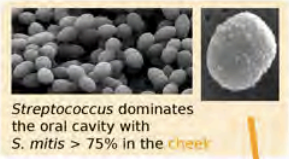
- Totality of microbes in a defined environment, and their intricate interactions with each other and the surrounding environment
 - A population of a single species is a culture(monoculture), extremely rare outside of lab and in some infections
 - A microbiome is a mixed population of different microbial species
 - MIXED COMMUNITY IS THE NORM!

Why Study Microbiomes

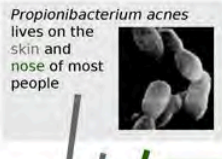
- Microbes modulate and maintain the atmosphere
 - Critical elemental cycles (carbon, nitrogen, sulfur, iron,...)
 - Pollution control, clean up fuel leaks
- Microbes keep us healthy
 - Protection from pathogens
 - Absorption/production of nutrients in the gut
 - Role in chronic diseases (obesity, Crohn's/IBD, arthritis...)
- Microbes support plant growth and suppress plant disease
 - Most complex communities reside in soil
 - Crop productivity

A map of diversity in the human microbiome

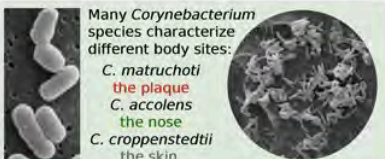
Streptococcus dominates the oral cavity with *S. mitis* > 75% in the **cheek**



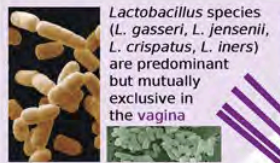
Propionibacterium acnes lives on the skin and nose of most people



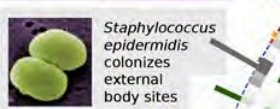
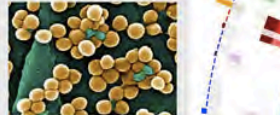
Many *Corynebacterium* species characterize different body sites:
C. matruchoti the **plaque**
C. accolens the **nose**
C. croppenstedtii the **skin**



Lactobacillus species (*L. gasseri*, *L. jensenii*, *L. crispatus*, *L. iners*) are predominant but mutually exclusive in the **vagina**



Staphylococcus epidermidis colonizes external body sites

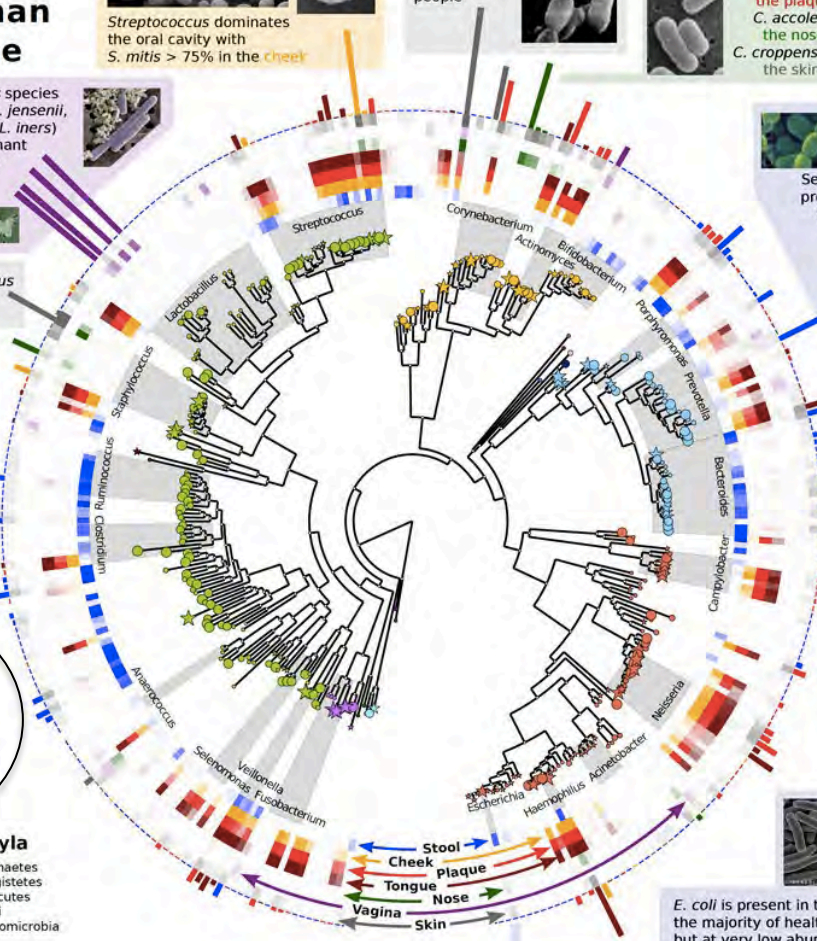
○ Commensal microbes
 ☆ Potential pathogens

The four most abundant phyla

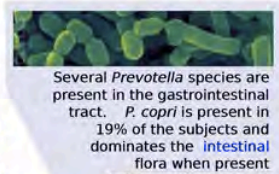
- Actinobacteria
- Bacteroidetes
- Firmicutes
- Proteobacteria

Low abundance phyla

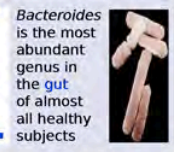
- Chloroflexi
- Cyanobacteria
- Euryarchaeota
- Fusobacteria
- Lentisphaerae
- Spirochaetes
- Synergistetes
- Tenericutes
- Thermi
- Verrucomicrobia



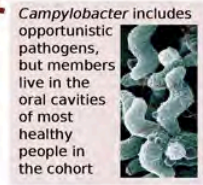
Several *Prevotella* species are present in the gastrointestinal tract. *P. copri* is present in 19% of the subjects and dominates the **intestinal** flora when present




Bacteroides is the most abundant genus in the **gut** of almost all healthy subjects



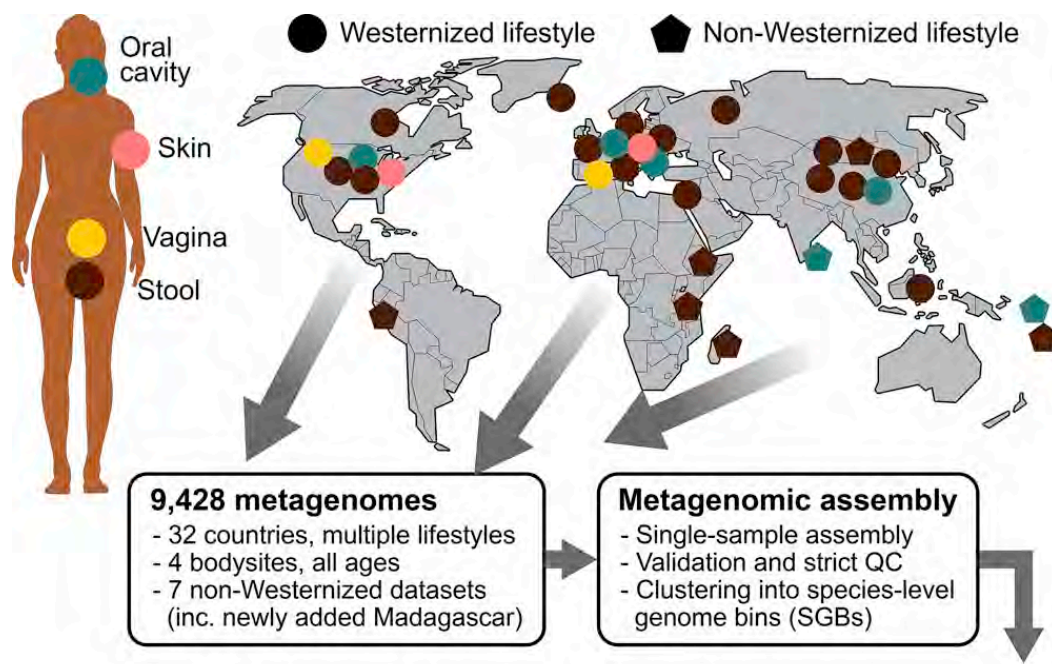
Campylobacter includes opportunistic pathogens, but members live in the oral cavities of most healthy people in the cohort



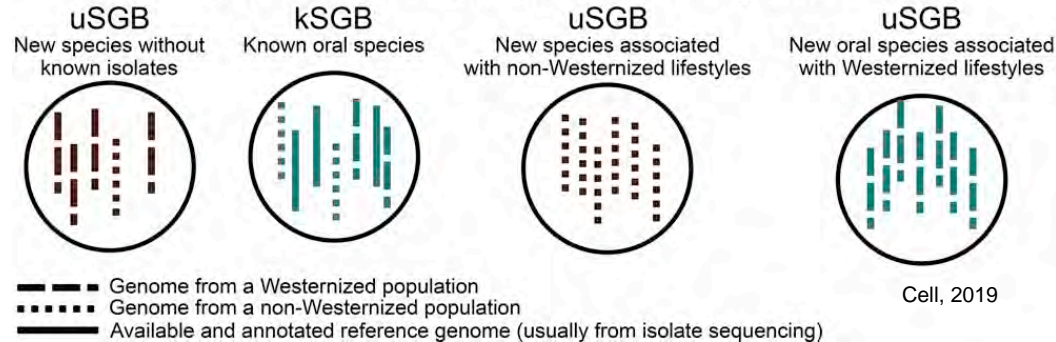
E. coli is present in the **gut** of the majority of healthy subjects but at very low abundance



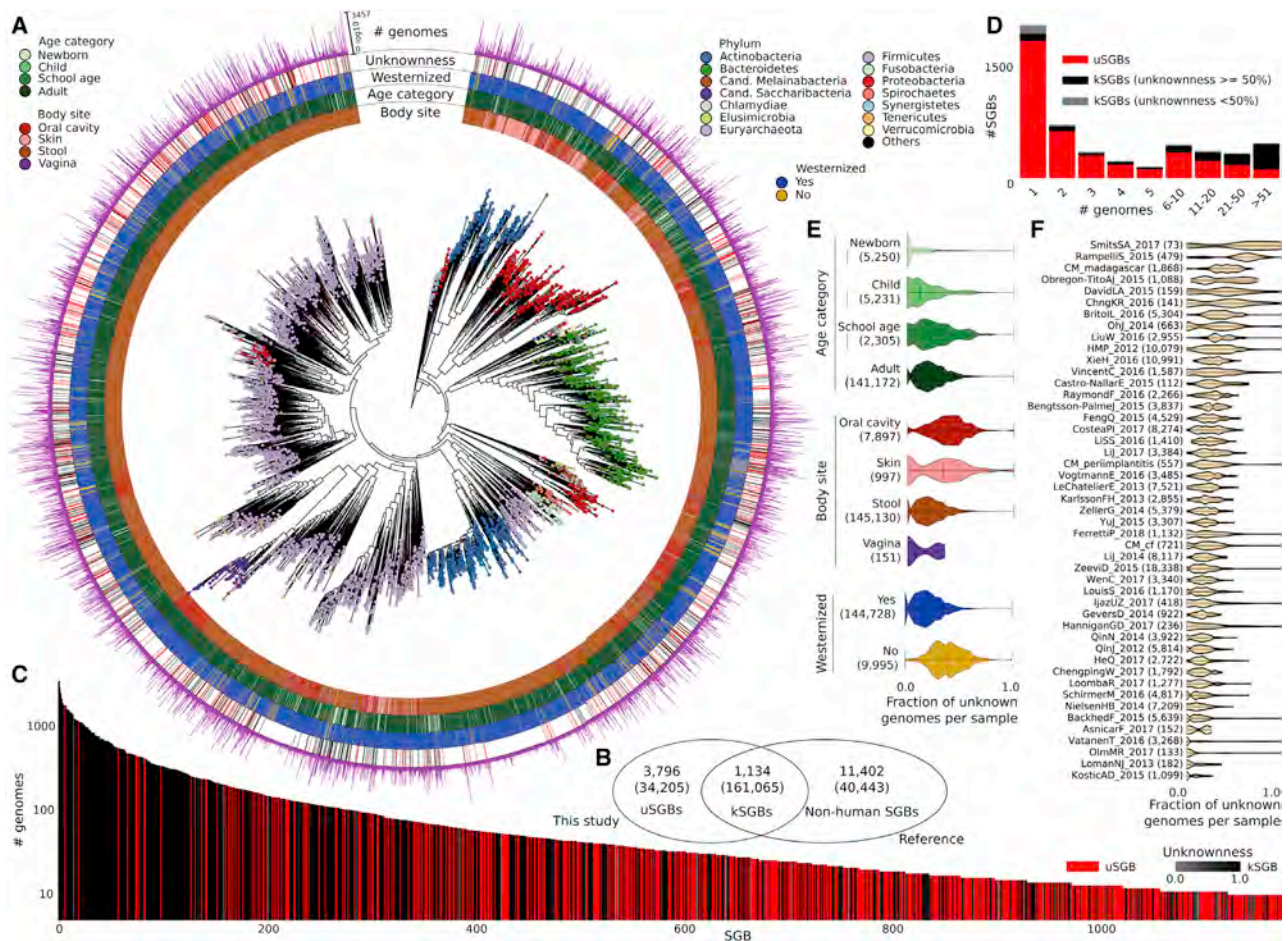
	Human (isolated)	Microbiota
weight	~ 50-100 kg	~ 2 kg
species	1	1000-5000
cells	~ 10 ¹²	10 ¹³ - 10 ¹⁴
genes	25.000	>4.000.000



154,723 microbial genomes from metagenomes



Cell, 2019



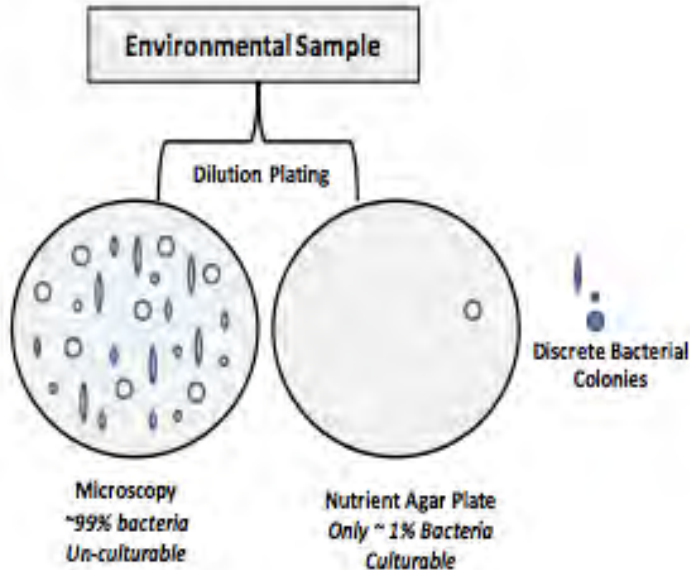
Recovered over 150,000 microbial genomes from ~10,000 metagenomes

70,178 genomes assembled with higher than 90% completeness

3,796 SGBs (species-level genome bins) identified -77% of the total representing species without any publicly available genomes

- American Gut Project
- Earth microbiome Project
- Human Oral Microbiome Database
- CardioBiome
- Human Microbiome Studies – JCVI
- MetaSub – Metagenomics and metadesign of Subways and Urban Biomes
- Gut microbiota for Health
- NASA: Study of the impact of long term space travel in the Astronaut's microbiome
- Michigan microbiome project
- Coral microbiome project
- Seagrass microbiome project
- Brazilian microbiome project
- Home microbiome study

The great “plate count” anomaly



- Cultivation based cell counts are orders of magnitude lower than direct microscopic observation
- As microbiologists, we are able to cultivate only a small minority of naturally occurring microbes
- Our nucleic acid derived understanding of microbial diversity has rapidly outpaced our ability to culture new microbes

IJSR, Sept 2013

Why is microbiome research new?

- Bias for microbes (especially pathogens) that are cultivable
 - Culture-based methods do not detect majority of microbes
 - Only pathogens are easily detected
 - And most microbes are not pathogens
- Availability of novel tools
 - Discovery of culture independent techniques
 - Amplicon sequencing and DNA sequencing

Roadmap to Culture Independent Techniques

1977: rRNA as an evolutionary marker (Woese and Fox, PNAS)

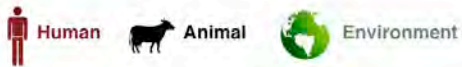
1985: Polymerase Chain Reaction (K. Mullis, Science)

1985: “Universal Primers” for rRNA sequencing (N. Pace, PNAS)

1989: PCR amplification of 16S rRNA gene (Bottger, FEMS Microbiol)

Early 1990's: Curation and hosting of RDP (rRNA database) FTP

2001: Term ‘microbiome’ coined by Lederberg and McCray

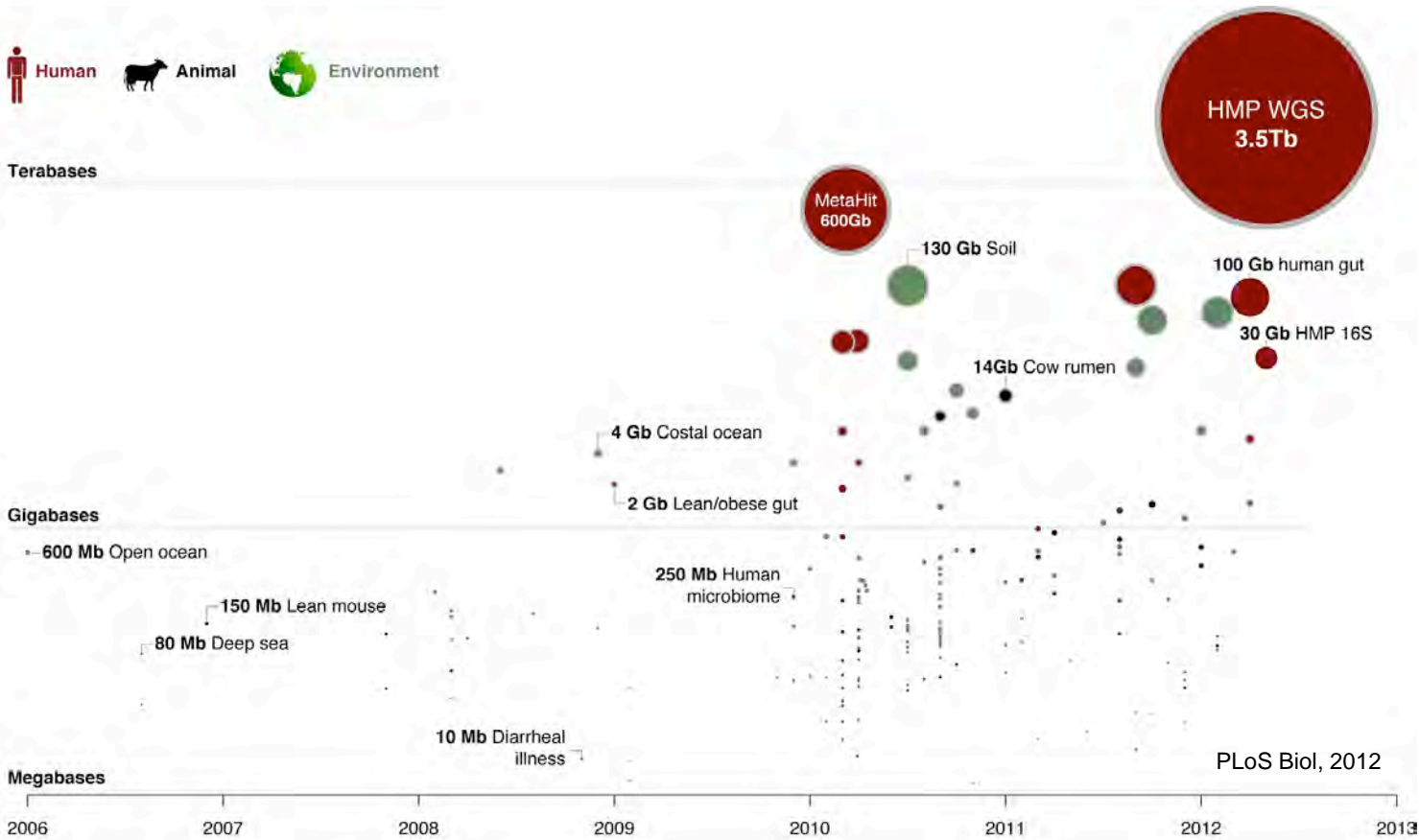


Terabases

Gigabases

Megabases

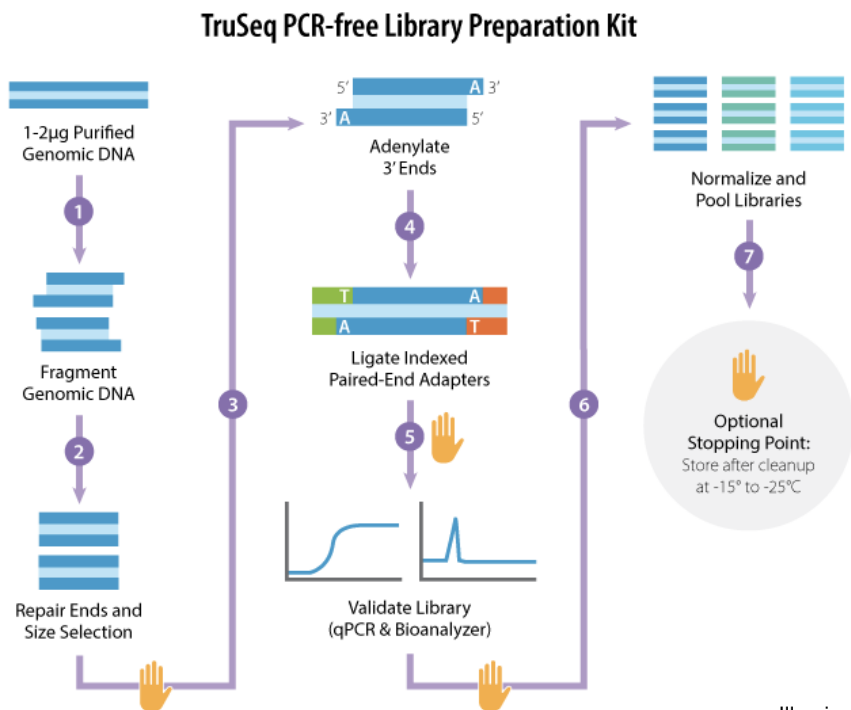
2006 2007 2008 2009 2010 2011 2012 2013



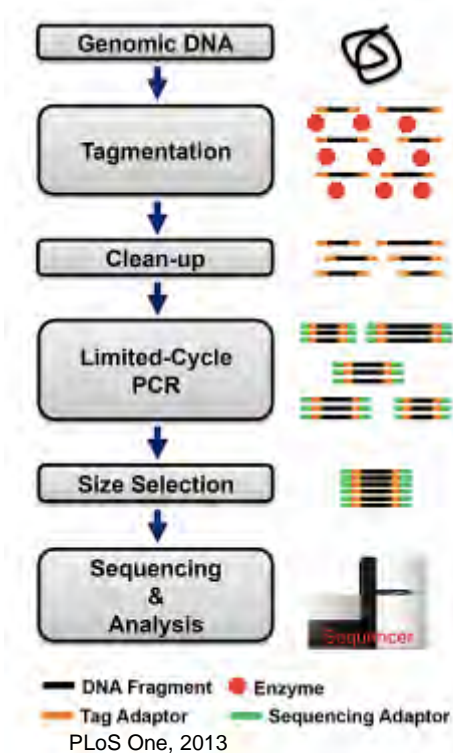
Timeline of microbial community studies using high-throughput sequencing

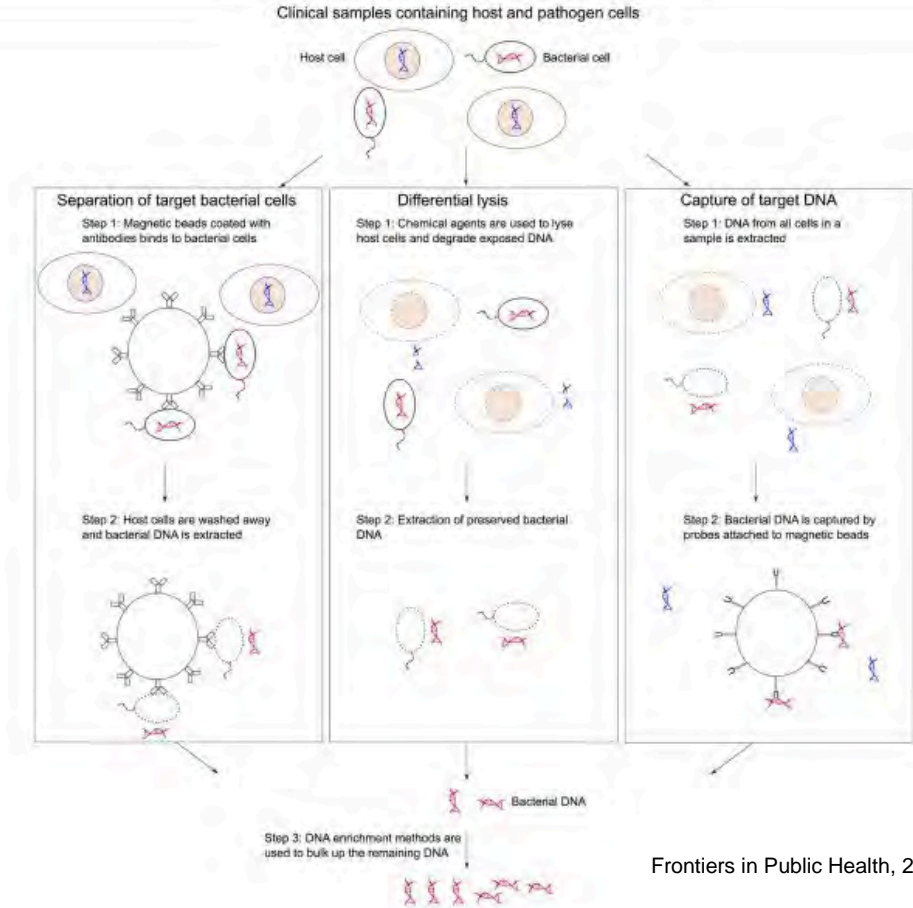
Library preparation workflow

Illumina TruSeq DNA PCR-free, Illumina Nextera XT, Kapa Hyperprep



Illumina.com

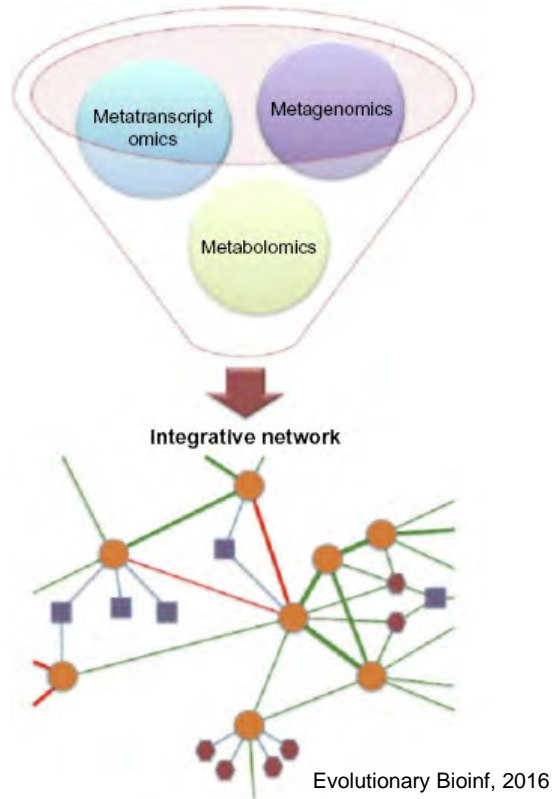




Pretreatment methods for metagenomics:

1. Microbial separation
2. Depletion of host nucleic acid
3. Targeted enrichment of pathogen DNA after extraction

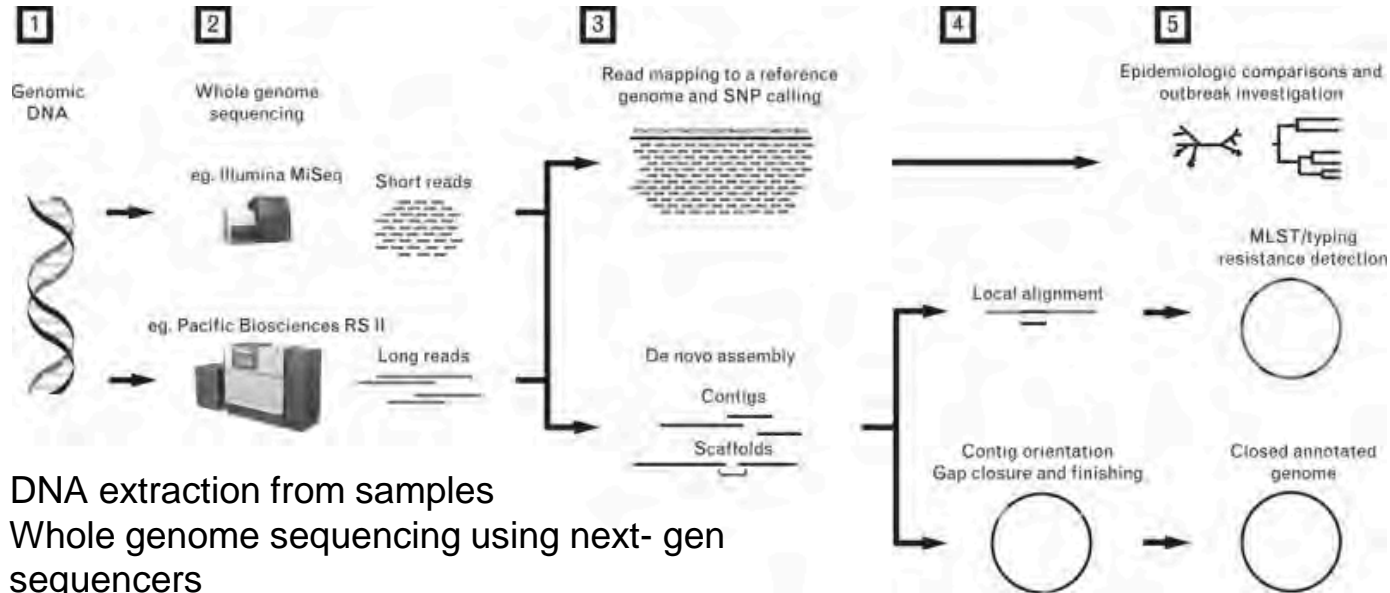
Where we are headed!



Integrated networks for multi omics data

NGS and pathogen detection

Whole genome sequencing workflow



Pathology, 2015

1. DNA extraction from samples
2. Whole genome sequencing using next- gen sequencers
3. Reference based SNP calling to perform phylogenetic analysis to assist with epidemiological outbreak
4. Resulting assembly used for typing and resistance detection
5. Closed genome used for further analysis

MetaG Workshop Instructional Plan

Pathogen detection (PH labs) – what shouldn't belong

- Amplicon/targeted sequencing- 16S (Sanger, Illumina)
- This will tell you what doesn't belong

PHASE I



Biotinylated magnetic bead method for sample enrichment once suspected pathogen is identified by previous method

- Using enriched DNA from this protocol, libraries will be made and sequenced on MiSeq.
- From the reads (BLAST reads or assemble reads and BLAST) the AMR profile and virulence factors will be determined

PHASE II



Metagenomics-pathogen detection by mid-throughput sequencing

- Novel methods/pipelines for metagenomic pathogen identification

PHASE III

Read Processing

- Demultiplex by index or barcode
- Remove adapter sequences
- Trim reads by quality

- Analysis of sequence quality, GC content, presence of adapters, overrepresented k-mers, duplicated reads, etc

- Go through format of fastq (paired and single end reads)

Fastq format

```
@HWI-ST911:111:C0N4WACXX:5:1101:2249:2216 1:N:0:TTAGGC CGATC:@@FF
NATGGCACCATTAAAAAGAATGTTTTATATGGTGTGAGAAGGACAAAGCTGAAGAAGAAATTTAGTCTGCACTTGATGTTGCAAATGCAAAGAAA
+
#2A2<CCFHIIIIIIIIIIIGCCCHIIIGIIIFFHIIIDGHIGIIIIIIICHGIIIGGCECEGICFHCECDEFFFFFDEEEEEDDDDDCDDCDDDDBC
@HWI-ST911:111:C0N4WACXX:5:1101:2509:2197 1:N:0:TTAGGC CGATC:1+4=B
NATGAGATAAATCAATTGTCTTTAATGAAGTACAGTCTTTGAATAATGAGTTTTGAACTCTTCTGCAACTTTTTGGAAACTTTAAAGTTTGAATG
+
#4A2<AADHIIIIIIIIIIHHIIIIIIIIIFGIII@GIIFIIIGIIIIIEIDHEHIIHIIIIIIIIIIICHIIIHHEEDFFFFFEEEEEADDFC
@HWI-ST911:111:C0N4WACXX:5:1101:3746:2179 1:N:0:TTAGGC CCATC:+11+A
NATGTCATCCATCTTTTCTATCTAAAAAGAATCAAAAAAGGGATAGTACAGAGGGAAAGTTCAATCCAGAGGACGATGAAACACTGATTGATGG
+
```

A FASTQ file normally uses four lines per sequence.

- Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).
- Line 2 is the raw sequence letters.
- Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.
- Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

FastQC- Quality Control for high throughput sequence data

- Exercise: Run FastQC (Refer to pdf)

Invoke/activate the environment where all bioinformatics programs are installed on logrus

```
$source activate bio
```

(Bio is the environment created on logrus where Conda has been installed. Conda is an open source management system and an environment management system that runs on Mac, Windows and Linux. It installs, runs and updates bioinformatics packages)

Let's go through results

- ✓ [Basic Statistics](#)
- ✓ [Per base sequence quality](#)
- ✓ [Per tile sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ✓ [Per base sequence content](#)
- ✓ [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ✓ [Sequence Length Distribution](#)
- ✓ [Sequence Duplication Levels](#)
- ✓ [Overrepresented sequences](#)

Different Analysis Modules

- <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/>
- <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

Documentation

A [copy of the FastQC](#) documentation is available for you to try before you buy (well download..).

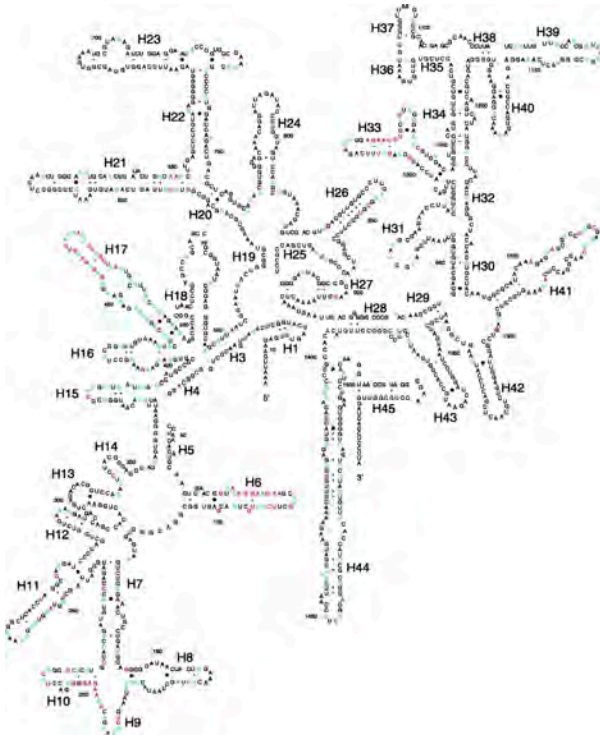
Example Reports

- [Good Illumina Data](#)
- [Bad Illumina Data](#)
- [Adapter dimer contaminated run](#)
- [Small RNA with read-through adapter](#)
- [Reduced Representation BS-Seq](#)
- [PacBio](#)
- 454

FastQC for 16S rRNA dataset

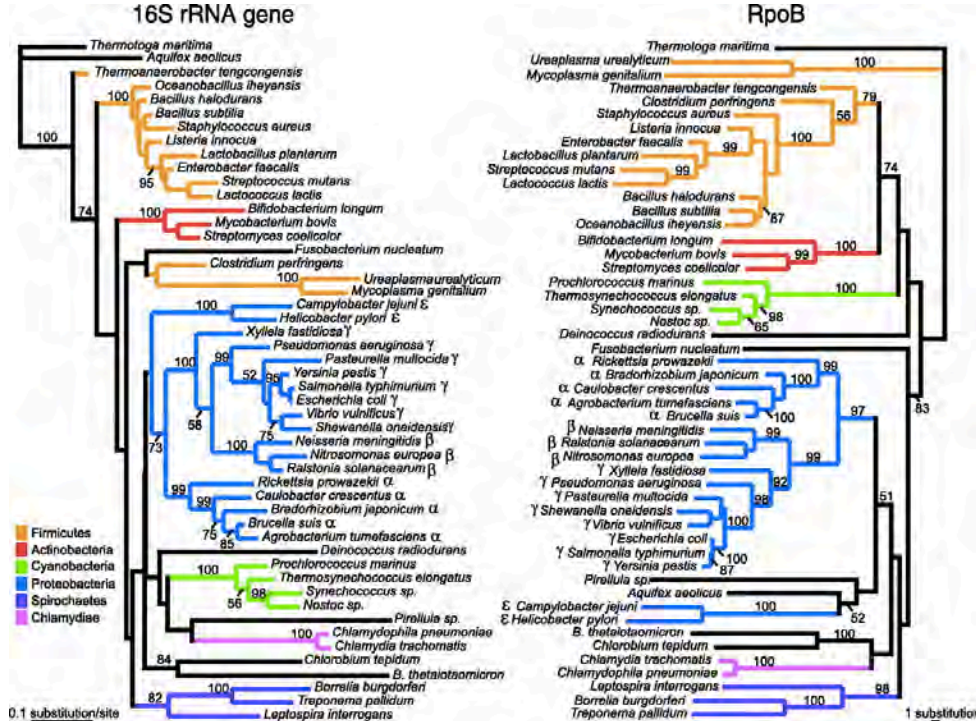
- Extremely biased per base sequence content
- Extremely narrow distribution of GC content
- Very high sequence duplication levels
- Abundance of overrepresented sequences
- In cases where the PCR target is shorter than the read length, the sequence will read through into adapters

16S rRNA as an evolutionary chronometer



- Ubiquitous – present in all known life (excluding viruses)
- Functionally constant wrt translation and secondary structure
- Evolves very slowly – mutations are extremely rare
- Large enough to extract information for evolutionary inference
- Limited exchange – limited examples of rRNA gene sharing between organisms

16S rRNA vs *rpoB* (RNA polymerase β subunit gene)



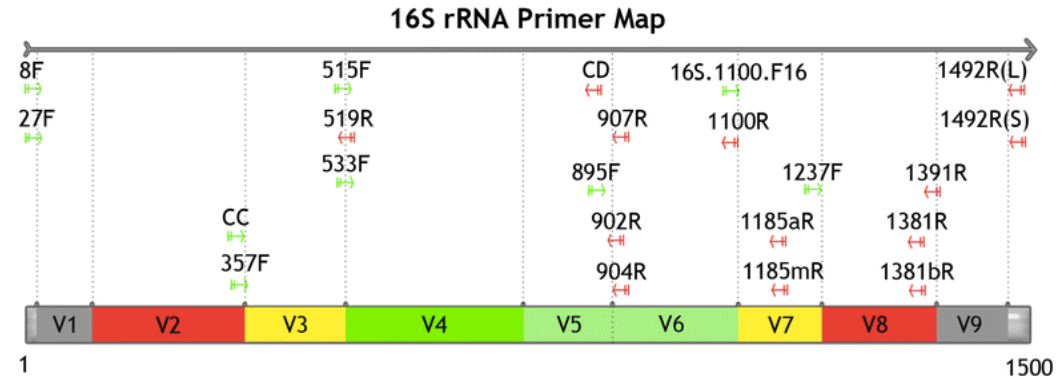
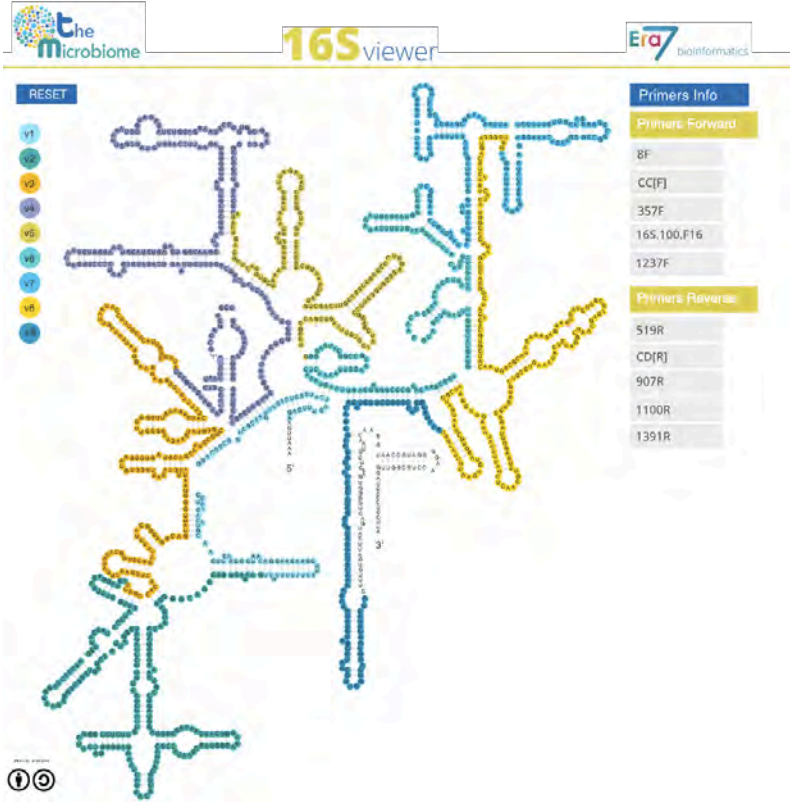
rpoB

- Universal presence
- Slow evolving
- Housekeeping function
- Large enough to contain phylogenetic information

** FUNDAMENTAL PROPERTY OF PROTEIN CODING GENES: SATURATION OF ALL THIRD CODON POSITIONS OVER A LONG EVOLUTIONARY SCALE

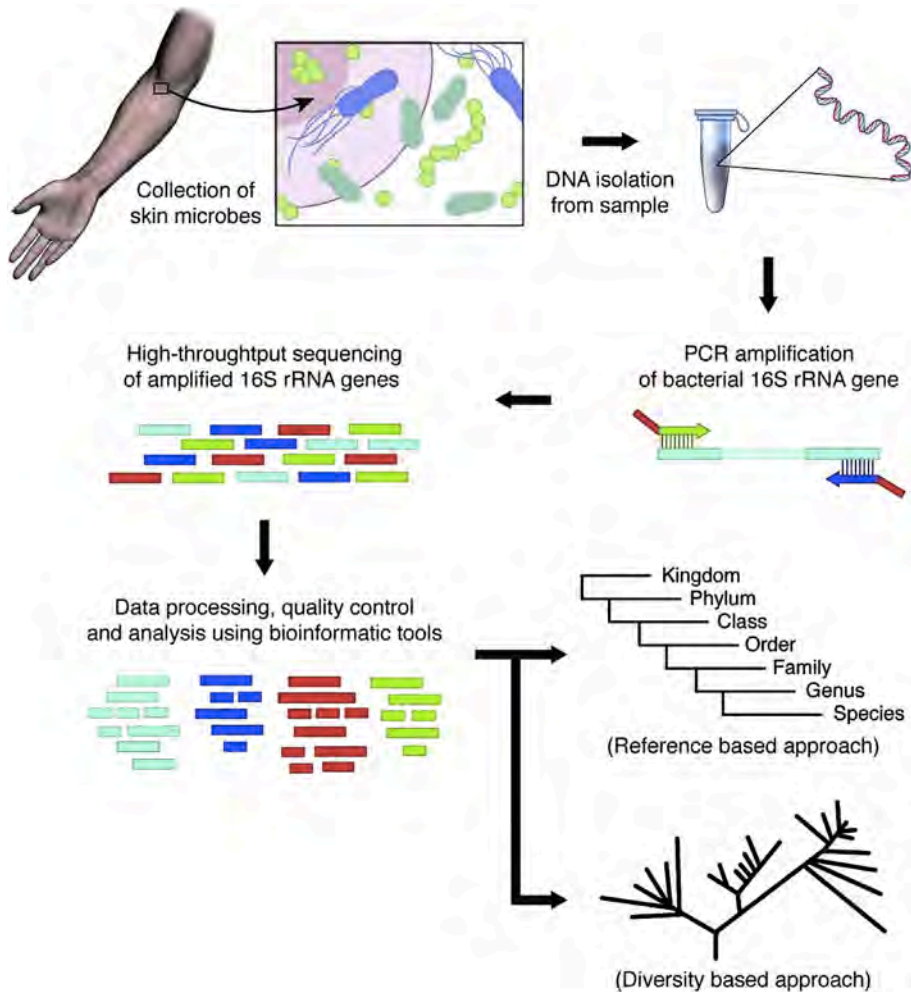
Kjellerberg et al, Microbial Ecology, 2007

16S rRNA hypervariable regions



BMC Bioinf, 2016

Illustration of different hypervariable regions of 16S rRNA



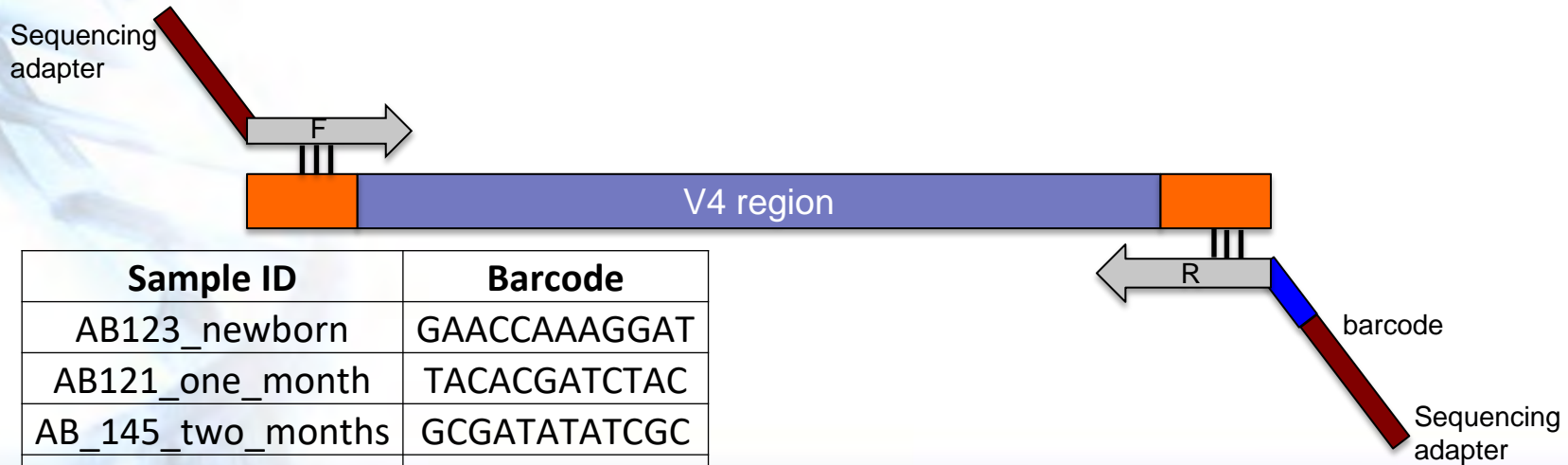
Basic Workflow for 16S Gene Based Sequencing

Addressing the 'fine print' while generating 16S rRNA gene amplicon libraries

- Sample Collection
 - Sample collection significantly influences the microbiome profile after sequencing
 - Sample storage
- DNA isolation
 - Template concentration
 - Template extraction protocol
- PCR amplification
 - PCR bias and inhibitors
 - Amplification of contaminants

Steps Involved

- Experimental Design: How many samples can be included in the sequencing run?
 - By using barcoded primers, numerous samples can be sequenced simultaneously (multiplexing)



Sample ID	Barcode
AB123_newborn	GAACCAAAGGAT
AB121_one_month	TACACGATCTAC
AB_145_two_months	GCGATATATCGC
AB_134_birth	CAGTGCATATGC
AB_189_birth	TCCAAAGTGTTT
AB_170_birth	GGCCACGTAGTA

Samples

- More the number of samples, more cost effective the run (sequencing depth will be compromised)

Comparison of multiplexing capacity by sequencing system

Sequencing system	Multiplexing capacity ^a			Run data quality	
	PF paired reads ^b	15K reads per sample	100K reads per sample	% PhiX	% Reads ≥ Q30 ^c
iSeq 100 System ^d 2 × 150 bp	4M	267	40	5	94.1
MiSeq System ^e 2 × 300 bp	25M	1667	250	10-25	74.9

a. Based on recommended 15K-100K reads per sample for analysis with 16S Metagenomics BaseSpace App.

b. Based on published instrument specifications.

c. Average of Read 1 and Read 2 data.

d. iSeq 100 System: v1 > 80% bases higher than Q30 at 2 × 150 bp.

e. MiSeq System: v3 > 70% bases higher than Q30 at 2 × 300 bp.

illumina.com

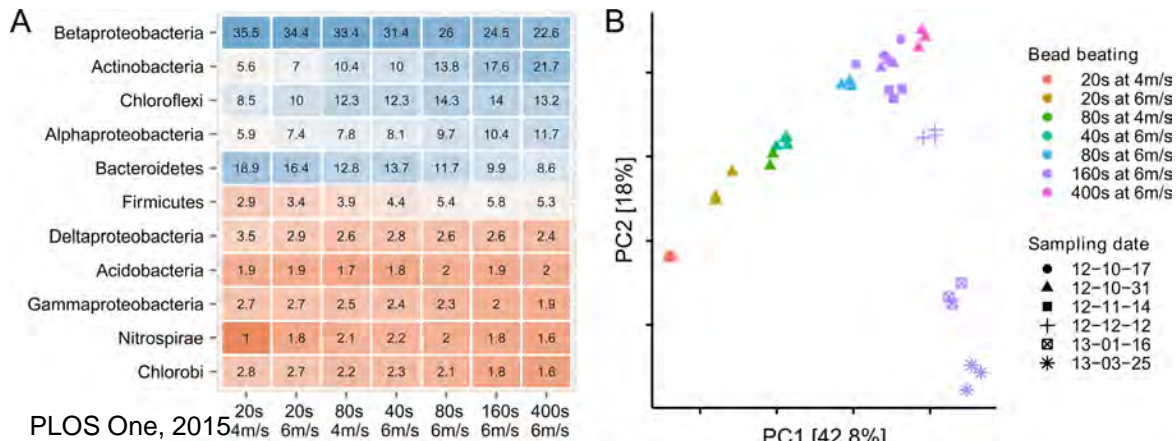
- It is critical to have a 'library prep manifest' to document the position of each sample with its associated barcode along with additional metadata information

Include Controls

- Between run repeat (process any sample in duplicate per run to measure reproducibility across runs)
- Within run repeat (process any sample in duplicate per plate to measure reproducibility)
- Water used during PCR (water blank- to determine if any contaminant was introduced during PCR reaction)
- Water spiked with known bacterial DNA (mock bacterial communities- enables quantification of sequencing errors, minimizes bias during sampling and library preparation)

DNA extraction protocol

- Effect of mechanical lysis methods for extraction
- Presence of inhibitors such as organic matter, humic acid, bile salts, polysaccharides
- DNA yield post extraction and reproducibility



Effect of bead beating was larger than sampling time over 5 months

- A. Percentage read abundance of the 11 most abundant phyla as a result of bead beating intensity
- B. PCA of samples with different bead beating intensities vs. samples taken at different dates

Selection of primers and region of 16S gene influence microbial profile

PLOS ONE | A Peer-Reviewed, Open Access Journal
View this Article | Submit to PLOS | Get E-Mail Alerts | Contact Us

PLoS One. 2016; 11(2): e0148047. PMID: PMC4734828
Published online 2016 Feb 1. doi: [10.1371/journal.pone.0148047](https://doi.org/10.1371/journal.pone.0148047) PMID: [26829716](https://pubmed.ncbi.nlm.nih.gov/26829716/)

Development of an Analysis Pipeline Characterizing Multiple Hypervariable Regions of 16S rRNA Using Mock Samples

Jennifer J. Barb,^{1*} Andrew J. Oler,² Hyung-Suk Kim,³ Natalia Chalmers,⁴ Gwenyth R. Wallen,⁵ Ann Cashion,³ Peter J. Munson,¹ and Nancy J. Ames⁵

Kostas Bourtzis, Editor

[Author Information](#) · [Article notes](#) · [Copyright and License Information](#) [Disclaimer](#)

This article has been [cited by](#) other articles in PMC.

Associated Data

[Supplementary Materials](#)

[Data Availability Statement](#)

Abstract

Go to:

Objectives

There is much speculation on which hypervariable region provides the highest bacterial specificity in 16S rRNA sequencing. The optimum solution to prevent bias and to obtain a comprehensive view of complex bacterial communities would be to sequence the entire 16S rRNA gene; however, this is not possible with second generation standard library design and short-read next-generation sequencing technology.

Methods

V2, V4, V6-V7 regions produced consistent results

[J Microbiol Methods](#). Author manuscript; available in PMC 2008 Oct 7.
Published in final edited form as:
[J Microbiol Methods](#). 2007 May; 69(2): 330–339.
Published online 2007 Feb 22. doi: [10.1016/j.mimet.2007.02.005](https://doi.org/10.1016/j.mimet.2007.02.005)

PMCID: PMC2562909
NIHMSID: NIHMS22082
PMID: [17391789](https://pubmed.ncbi.nlm.nih.gov/17391789/)

A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria.

Sourmitesh Chakravorty,¹ Danica Helb,¹ Michele Burday,² Nancy Connell,¹ and David Alland^{1,*}

► Author information ► Copyright and License information [Disclaimer](#)

The publisher's final edited version of this article is available at [J Microbiol Methods](#).
See other articles in PMC that [cite](#) the published article.

Associated Data

► Supplementary Materials

Abstract

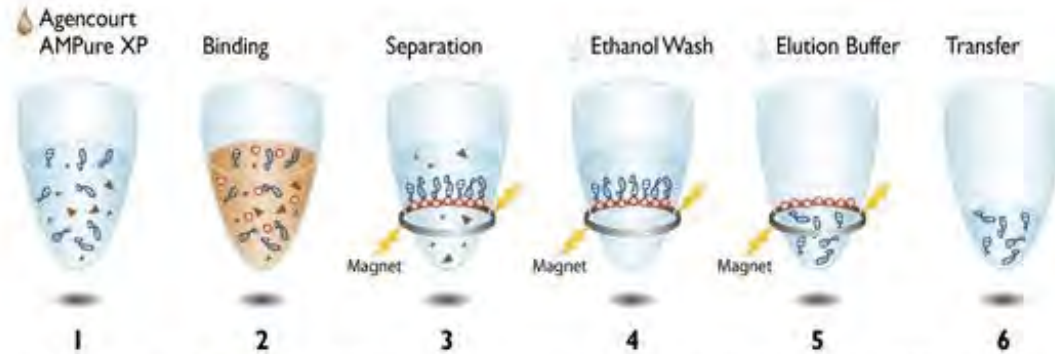
Go to: 

Bacterial 16S ribosomal RNA (rRNA) genes contain nine “hypervariable regions” (V1 – V9) that demonstrate considerable sequence diversity among different bacteria. Species-specific sequences within a given hypervariable region constitute useful targets for diagnostic assays and other scientific investigations. No single region can differentiate among all bacteria; therefore, systematic studies that compare the relative advantage of each region for specific diagnostic goals are needed. We characterized V1 - V8 in 110 different bacterial species including common blood borne pathogens, CDC-defined select agents and

- V2, V3 and V6 contain maximum nucleotide heterogeneity
- V6 is the shortest hypervariable region with the maximum sequence heterogeneity
- V1 is best target for distinguishing pathogenic *S aureus*
- V2 and V3 are excellent targets for speciation among *Staph* and *Strep* pathogens as well as *Clostridium* and *Neisseria* species
- V2 especially useful for speciation of *Mycobacterium sp.* and detection of *E coli O157:H7*
- V3 useful for speciation of *Haemophilus sp*
- V6 best target for probe based PCR assays to identify CDC select agents (bio-terrorism agents)

Purification of amplicons

- After one –step or two-step PCR, products are cleaned up using AMPure beads



1. PCR reaction 2. Binding of PCR amplicons to magnetic beads 3. Separation of PCR amplicons bound to magnetic beads from contaminants 4. Washing of PCR amplicons with Ethanol 5. Elution of PCR amplicons from the magnetic particles 6. Transfer away from the beads into a new plate

- Gel Electrophoresis and quantification of cleaned amplicon products
 - Qubit
- Sample pooling – equimolar concentrations (how many samples do you want to pool? How many reads per sample?)
- Gel extraction of pooled product
- Final clean up (Qiagen kit) and QC

Step 1

Design your experiment:
Controls
DNA extraction protocol
Based on your research question

Step 2

PCR amplification

Step 3

Cleaning of amplicon products

Step 4

Gel electrophoresis and quantification of cleaned amplicon product

Step 5

Agarose gel extraction of pooled product

Step 6

Cleaning of extracted amplicon library and final QC steps

Step 7

Sequencing of the final library

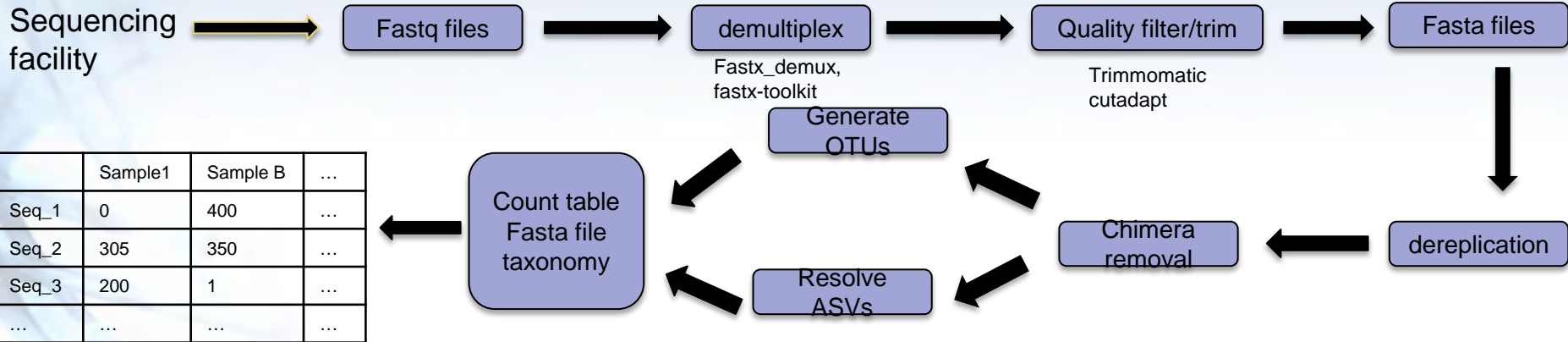
Step 8

Bio-informatics steps

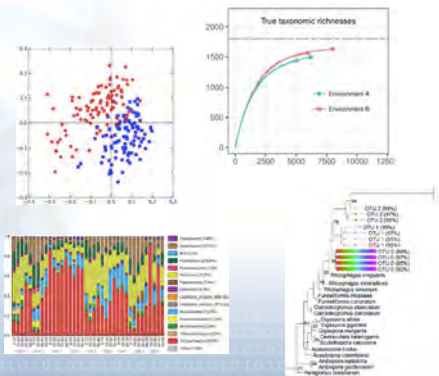
● Amplicon Sequencing Library Prep- PacBio

<https://www.pacb.com/wp-content/uploads/Procedure-Checklist-Amplicon-Template-Preparation-and-Sequencing.pdf>

Overview of generic amplicon workflow



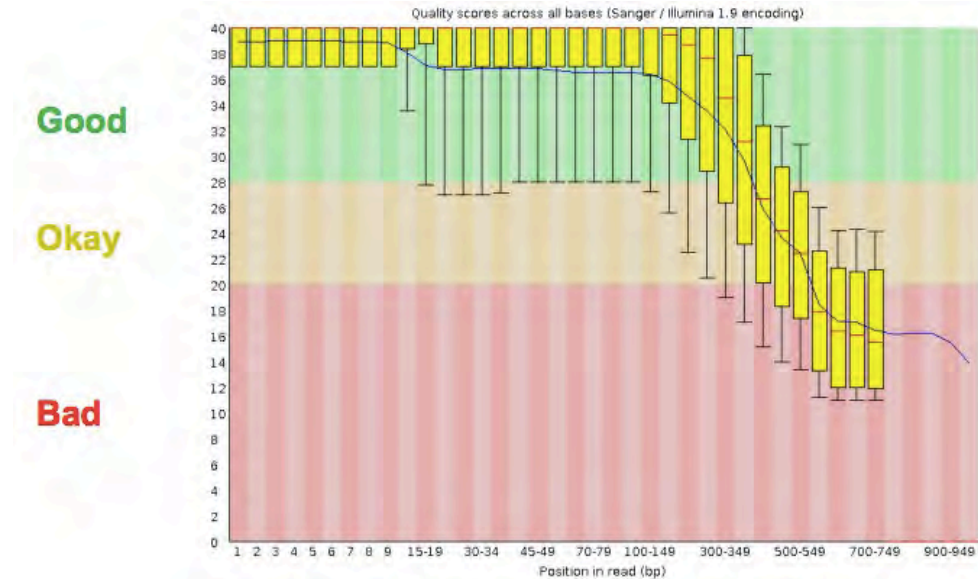
Analysis



For this workshop,
 QIIME2
 MOTHUR

Data Preprocessing

■ FastQC



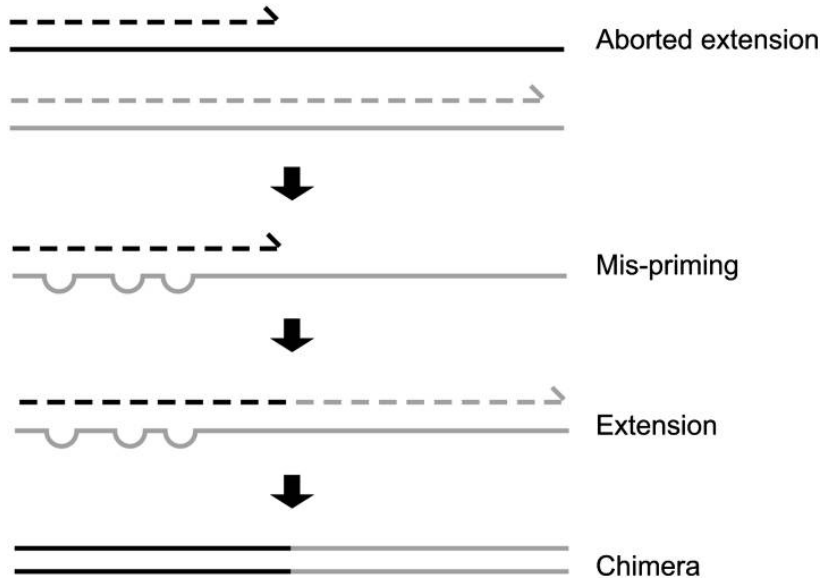
- Many tools/options to filter and trim data
- Trimming does not always improve things as valuable information can be lost!
- Removal of adapters is critical for downstream analysis

Dereplication

- In this process all the quality-filtered sequences are collapsed into a set of unique reads, which are then clustered into OTUs
- Dereplication step significantly reduces computation time by eliminating redundant sequences

Chimera detection and removal of non-bacterial sequences

- Chimeras as artifact sequences formed by two or more biological sequences incorrectly joined together



Genome Research, 2011

Incomplete extensions during PCR allow subsequent PCR cycles to use a partially extended strand to bind to the template of a different, but similar, sequence. This partially extended strand then acts as a primer to extend and form a chimeric sequence.

Clustering

- Analysis of 16S rRNA relies on clustering of related sequences at a particular level of identity and counting the representatives of each cluster



- Some level of sequence divergence should be allowed – 95% (genus-level, partial 16S gene), 97% (species-level) or 99% typical similarity cutoffs used in practice and the resulting cluster of nearly identical tags (assumedly identical genomes) is referred to as an OTU (Operational Taxonomic Unit)

Create OTU tables

- OTU table is a matrix that gives the number of reads per sample per OTU

#OTU ID	F3D0	F3D141	F3D142	F3D143	F3D144	F3D145	F3D146	F3D147
OTU_6	749	535	313	372	607	849	493	2025
OTU_25	29	57	14	2	14	22	16	127
OTU_1	613	497	312	247	472	719	349	1720
OTU_8	426	378	255	237	382	627	330	1417
OTU_31	149	38	10	19	25	21	43	31
OTU_2	366	392	327	185	313	542	248	1367
OTU_7	196	370	92	107	48	155	74	105
OTU_10	46	169	87	109	171	209	120	864
OTU_80	26	6	0	1	4	8	18	11

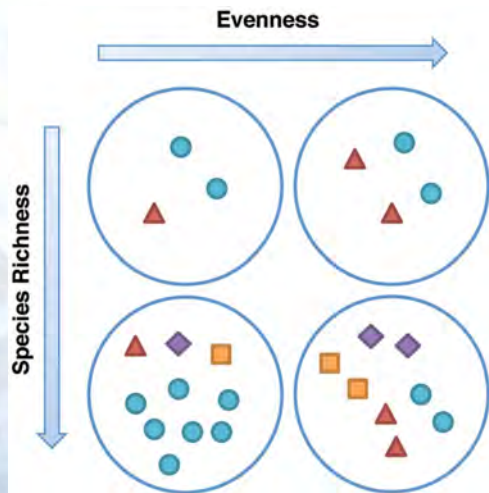
Bin OTUs into Taxonomy (assign taxonomy)

- Accuracy of assigning taxonomy depends on the reference database chosen
 - Ribosomal Database Project
 - GreenGenes
 - SILVA
- Accuracy depends on the completeness of databases

	A	B	C	D	E	F	G	H
1	OTU	Reads	Taxonomy					
2	Otu0001	342	Bacteria	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus
3	Otu0002	265	Bacteria	Firmicutes	Bacilli	Bacillales	Listeriaceae	Listeria
4	Otu0003	222	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus
5	Otu0004	191	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus
6	Otu0005	184	Bacteria	Firmicutes	Bacilli	Bacillales	Bacillaceae	Bacillus
7	Otu0006	170	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae	Clostridium
8	Otu0007	157	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	unclassified
9	Otu0008	152	Bacteria	Actinobacteria	Actinobacteria	Propionibacteriales	Propionibacteriaceae	Propionibacterium
10	Otu0009	144	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
11	Otu0010	143	Bacteria	Proteobacteria	Betaproteobacteria	Neisseriales	Neisseriaceae	Neisseria
12	Otu0011	139	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia-Shigella
13	Otu0012	125	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus
14	Otu0013	112	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
15	Otu0014	94	Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter
16	Otu0015	77	Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Rhodobacter

Population diversity – alpha diversity

- Assessment of diversity involves two aspects
 - Species richness (# of species present in a sample)
 - Species evenness (distribution of relative abundance of species)



Human Mol. Genet., 2013

Total community diversity of a single sample/environment is given by alpha-diversity and represented using rarefaction curves

Quantitative methods such as Shannon or Simpson indices measure evenness of the alpha-diversity

Beta-diversity

- Beta-diversity measures community structure differences (taxon composition and relative abundance) between two or more samples
 - For example, beta-diversity indices can compare similarities and differences in microbial communities in healthy and diseases states
- Many qualitative (presence/absence taxa) and quantitative (taxon abundance) measures of community distance are available using several tools
 - LIBHUFF, TreeClimber, DPCoA, UniFrac (QIIME)

Measuring Population Diversity – alpha and beta diversity

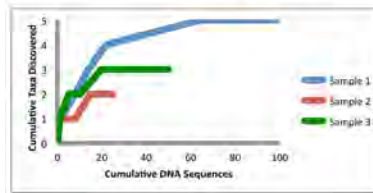
A) Sequence Abundance

OTU	Sample 1	Sample 2	Sample 3
A	60	0	35
B	24	5	5
C	10	0	0
D	5	0	0
E	1	0	0
F	0	20	10
Total	100	25	50

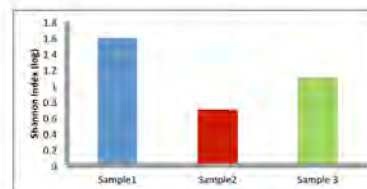
B) Sequence Relative Abundance

OTU	Sample 1	Sample 2	Sample 3
A	0.60	0	0.70
B	0.24	0.20	0.10
C	0.10	0	0
D	0.05	0	0
E	0.01	0	0
F	0	0.80	0.20
Total	1.0	1.0	1.0

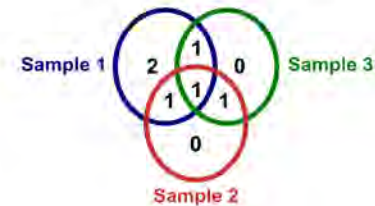
C) Collector's Curve of Sample Richness



D) Within-Sample Alpha Diversity

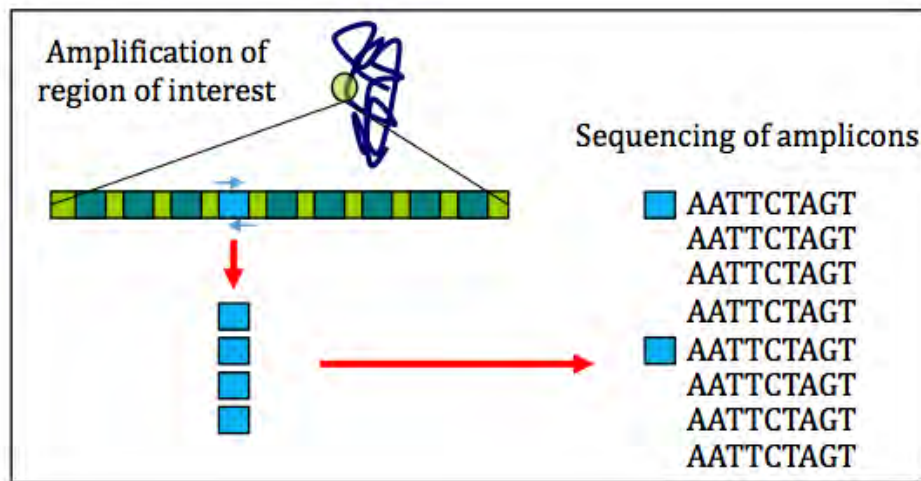


E) Between-Sample Beta Diversity



PLoS Computational Biol., 2012

16S rRNA sequencing – benefits and limitations

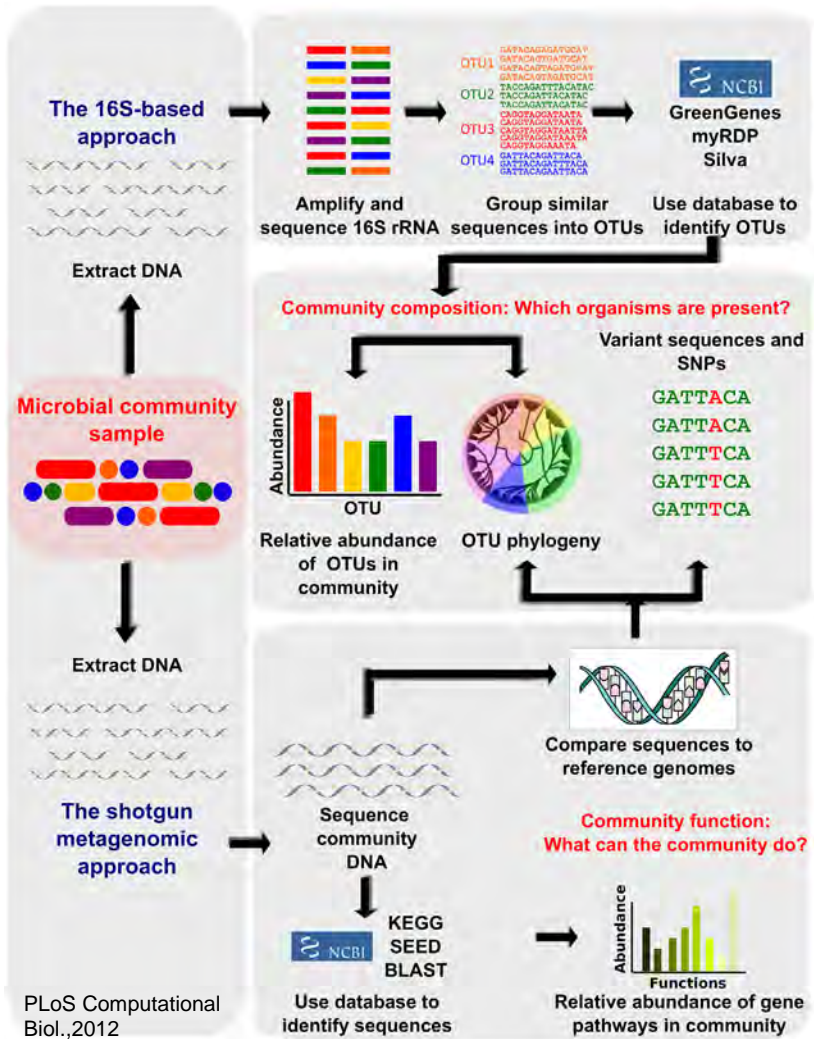


Benefits

- Cost effective
- Data analysis can be performed by established pipelines
- Large body of archived data is available for reference

Limitations

- Sequences only a single region of the genome
- Classifications often lack accuracy at the species level
- Copy number per genome can vary. While they tend to be taxon specific, variation among strains is possible
- Relative abundance measurements are unreliable because of amplification biases
- Diversity of the gene tends to overinflate diversity estimates



Shotgun Sequencing and Metagenomics