Metagenomics and Metatranscriptomics

Johnny Sena, Ph.D.
Joann Mudge, Ph.D.

Yesterday

Microbes

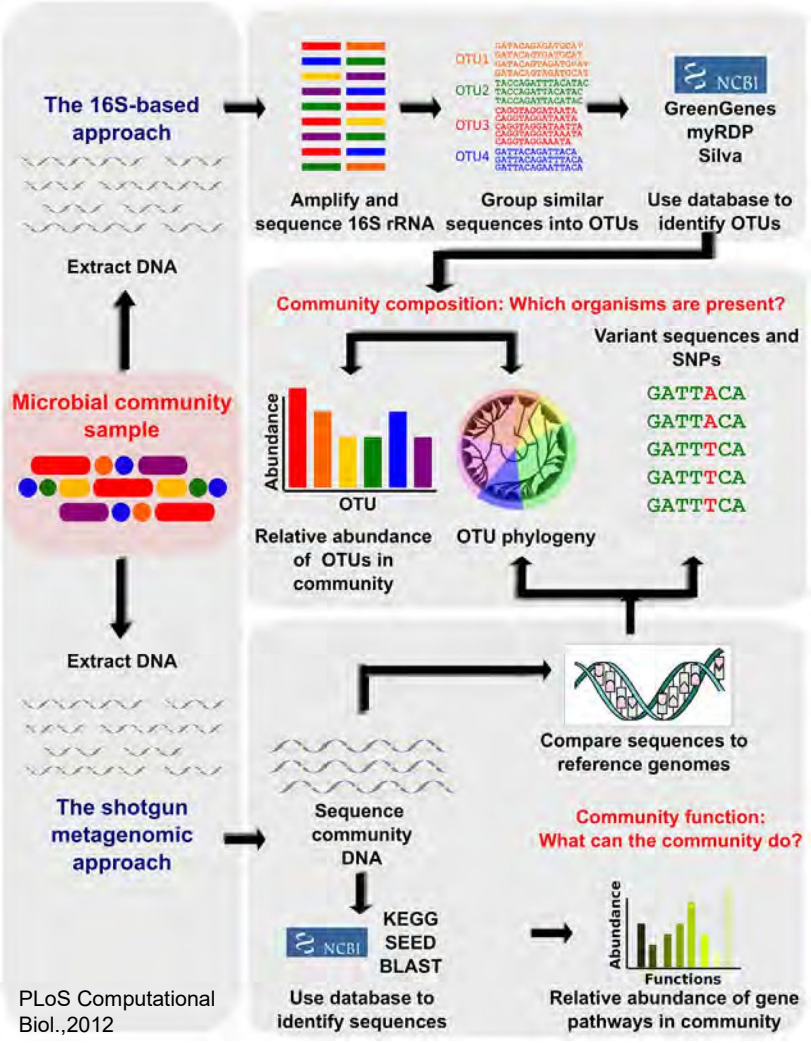
Microbiomes

Community Analysis

Today

Metatranscriptomics

Metagenomics



PLoS Computational Biol.,2012

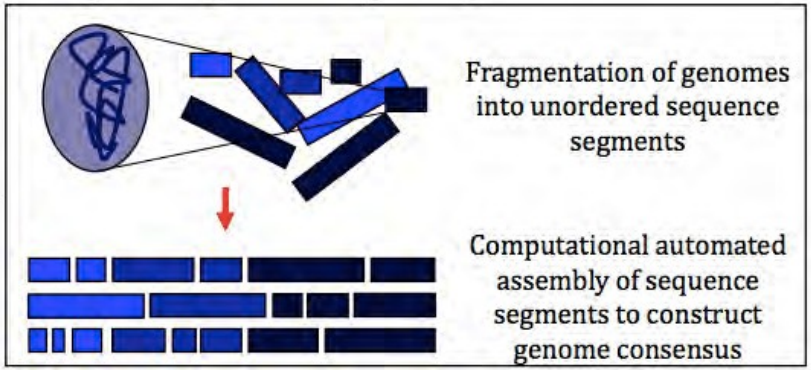
Shotgun Sequencing and Metagenomics

National Center for Genome Resources

Benefits and limitations of whole genome shotgun metagenomics vs community analysis

What do you think?

Benefits and limitations of whole genome metagenomics



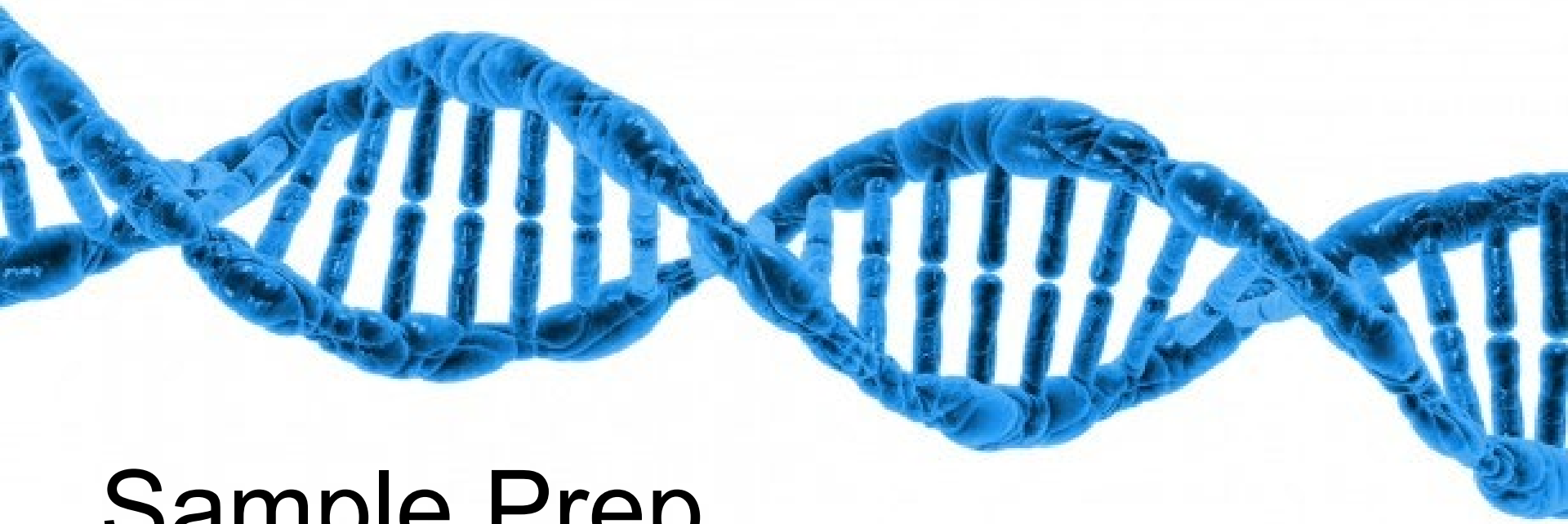
Benefits

- 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

NCGR
National Center for Genome Resources



Sample Prep

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

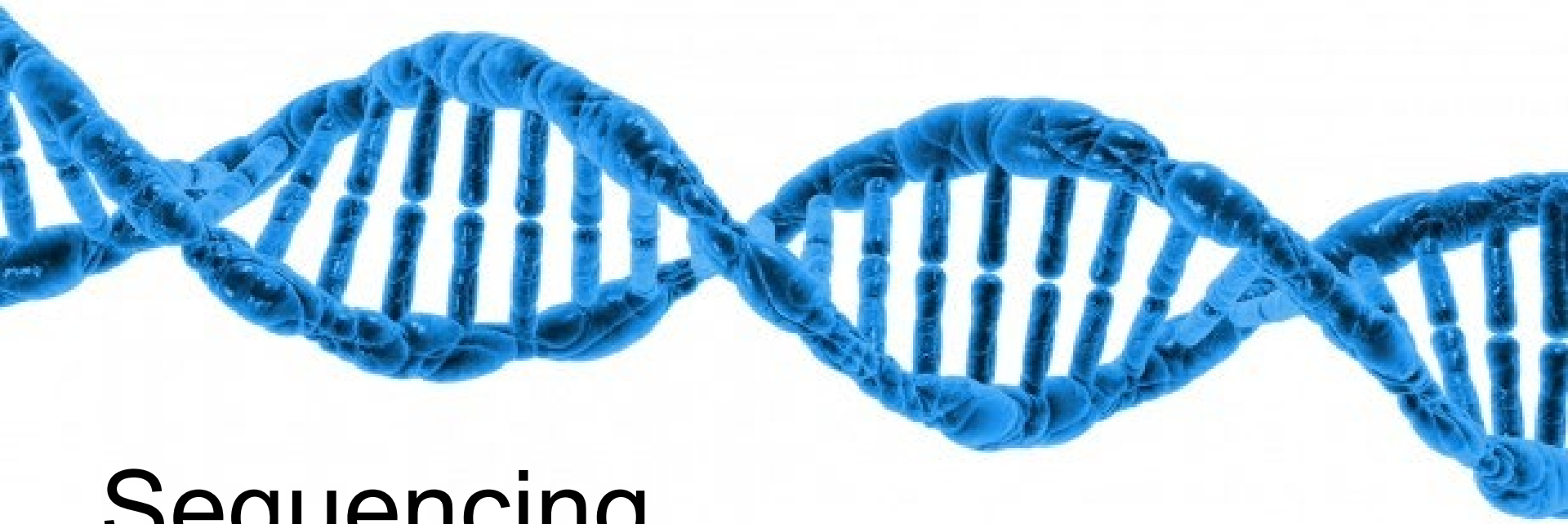
What do you think?

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

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)



Sequencing

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

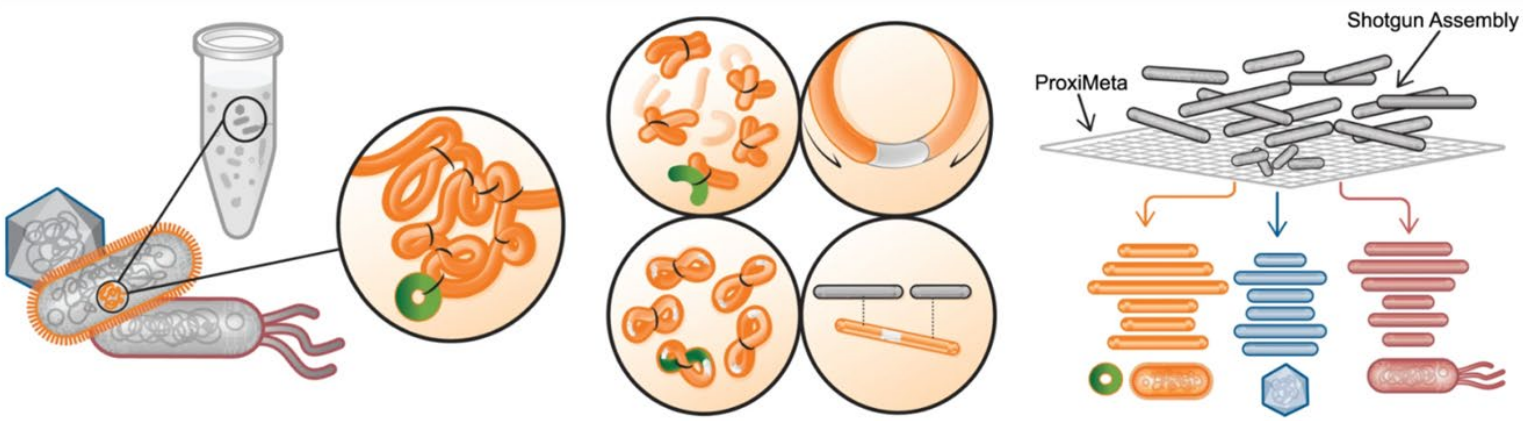
Long reads

- PacBio
  - Increased throughput and lower cost
  - HiFi (>99% accuracy) versus CLR
- Nanopore
  - Longer (up to 2 MB)
  - Recent improvements in accuracy

ProxiMeta (Hi-C) from Phase Genomics

Master the Microbiome

The ProxiMeta Metagenome Deconvolution Platform combines cost-effective proximity ligation data (generated with our optimized kits) with shotgun sequencing data, to assemble high-quality metagenomes and associate mobile genetic elements with their hosts. Capture strain-resolution insights without relying on 16S-based techniques, binning or culturing.

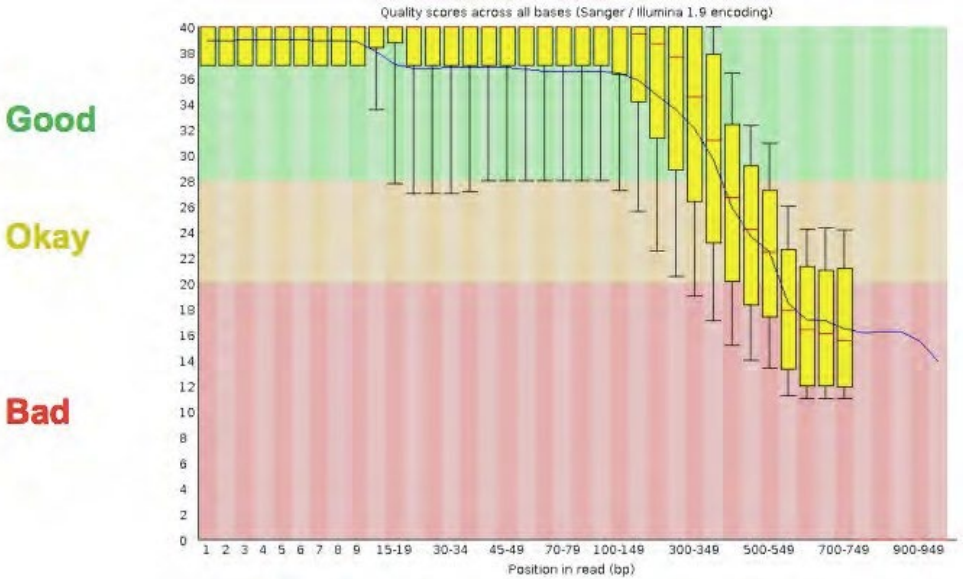


Proximity ligation (Hi-C) libraries are generated from a single mixed microbial sample. Interactions are captured by crosslinking, digesting, and creating chimeric junctions that are sequenced and analyzed with a shotgun assembly to deconvolve chromosomes and plasmids into complete genomes.

From: **Stadler, T. et al. The ISME Journal 2019; 13: 2437 – 2446.**

Data Preprocessing

- FastQC



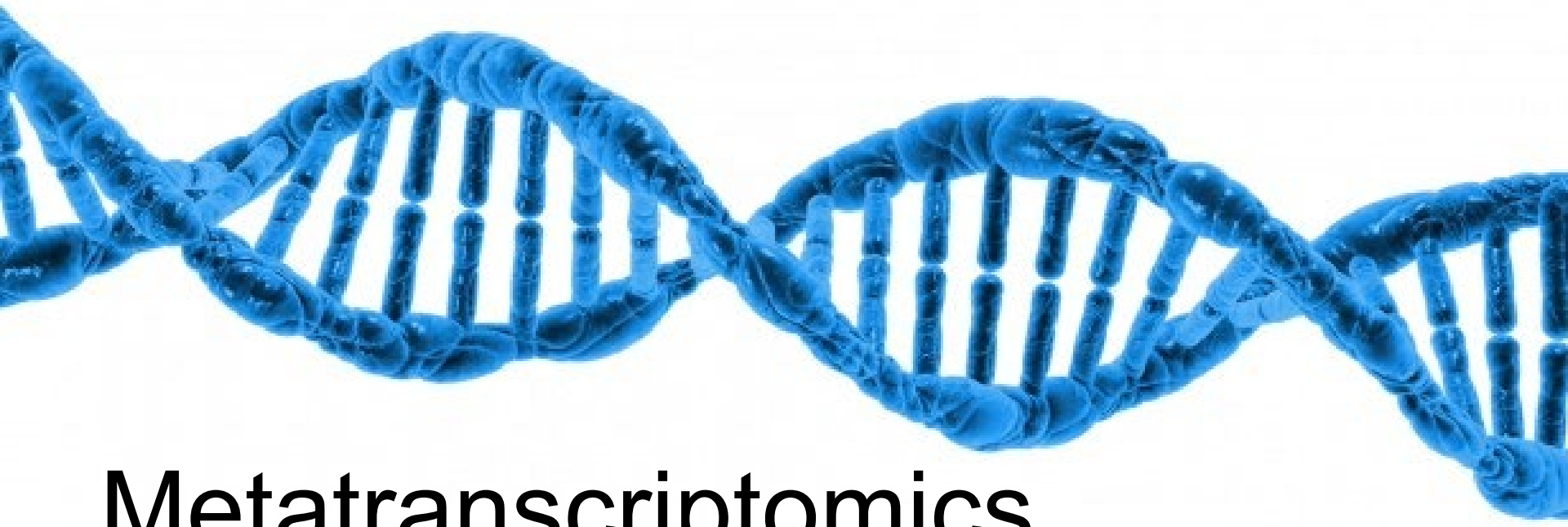
- Many tools/options to filter and trim data
- Trimming does not always improve things as valuable information can be lost!
- Removal of adapters is critical for downstream analysis

Let’s go through results

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

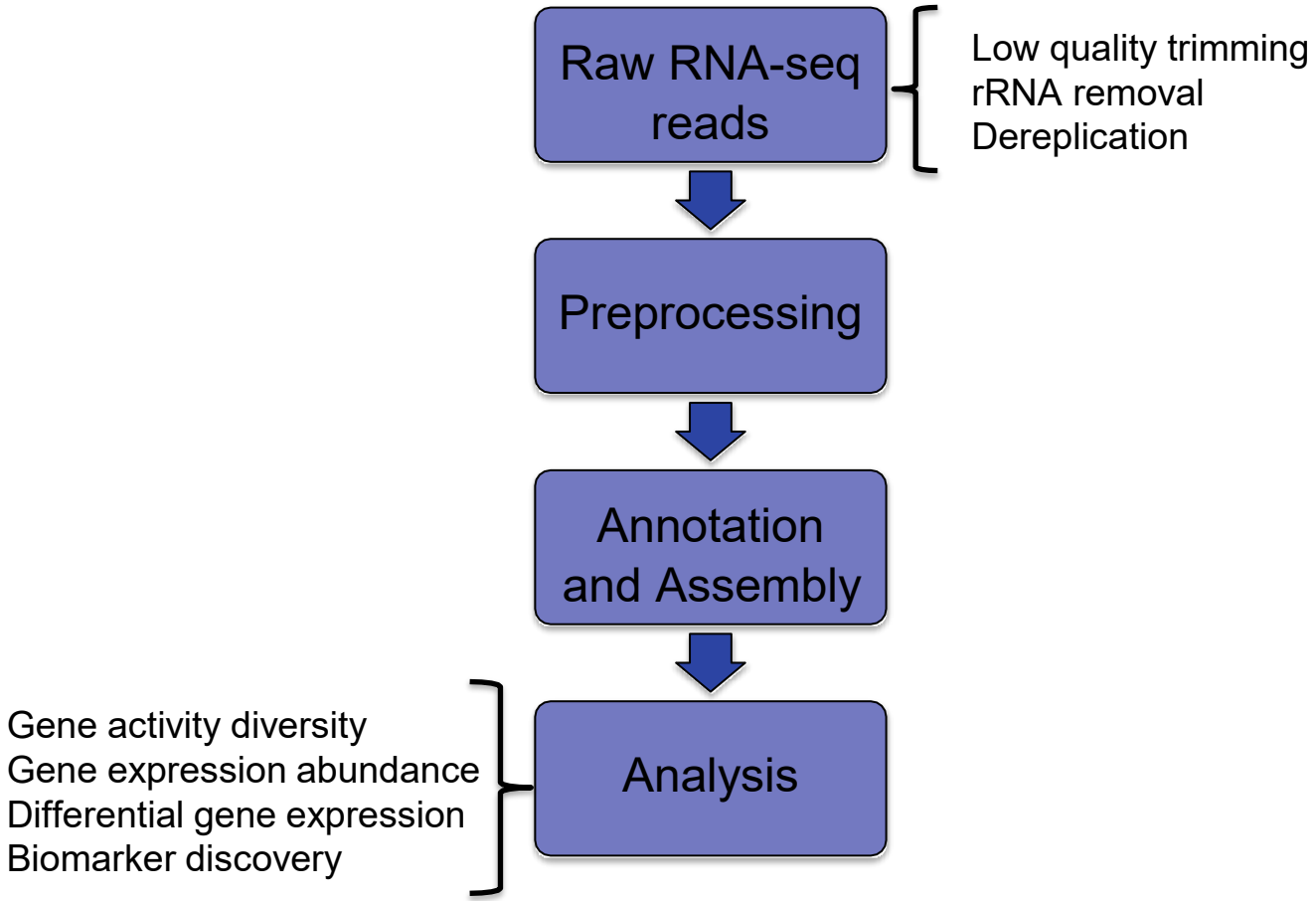
https://www.bioinformatics.babraham.ac.uk/projects/fastq_c/

Different Analysis Modules

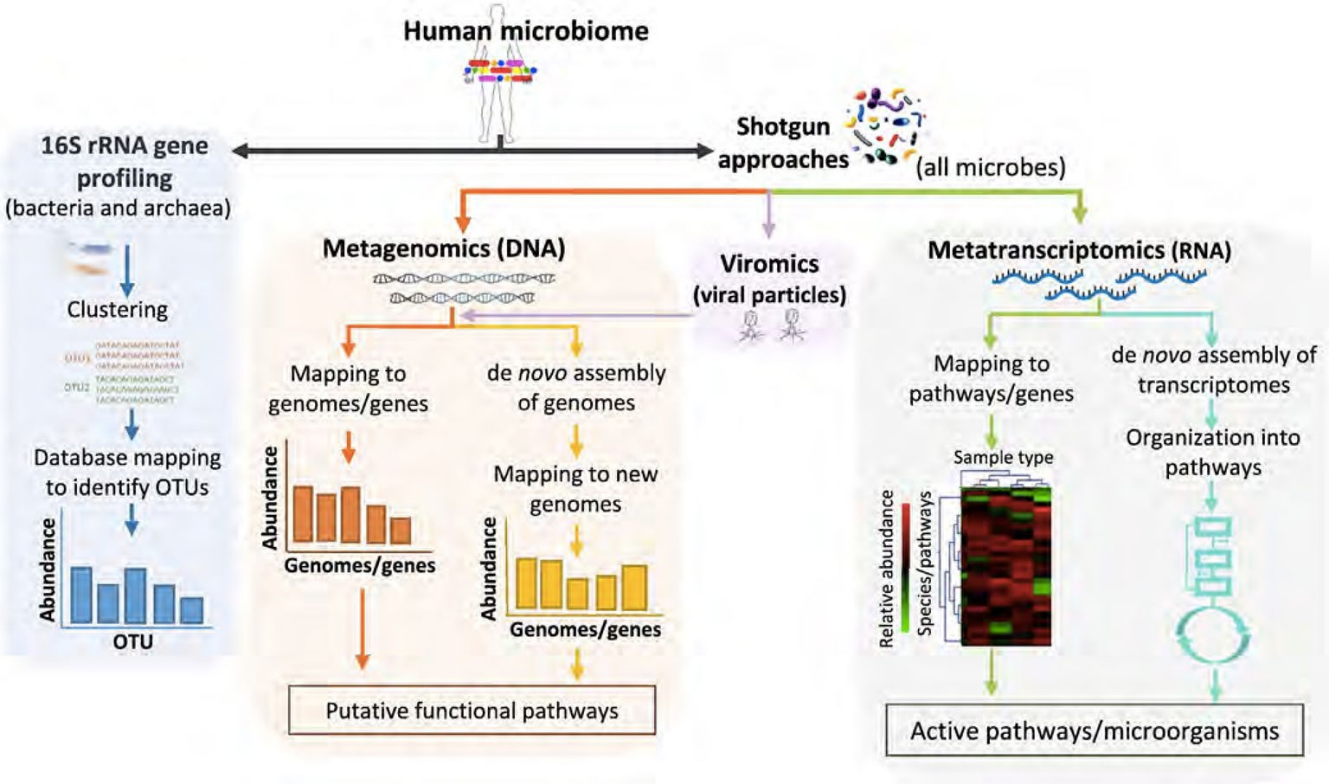


Metatranscriptomics

Simplistic Workflow



Metatranscriptomics



"What are they doing?"
- Metatranscriptomics

Computational and Structural Biotechnology Journal, 2015

Benefits of Metatranscriptomics

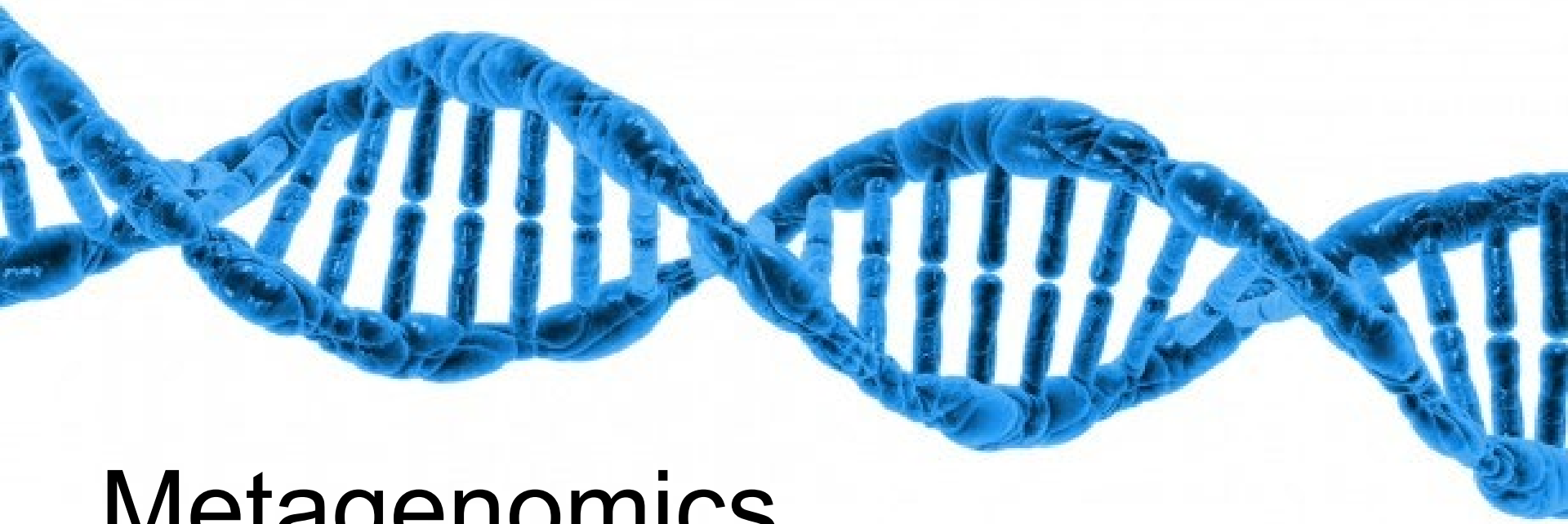
vs community analysis or whole genome metagenomics

What do you think?

Benefits of Metatranscriptomics

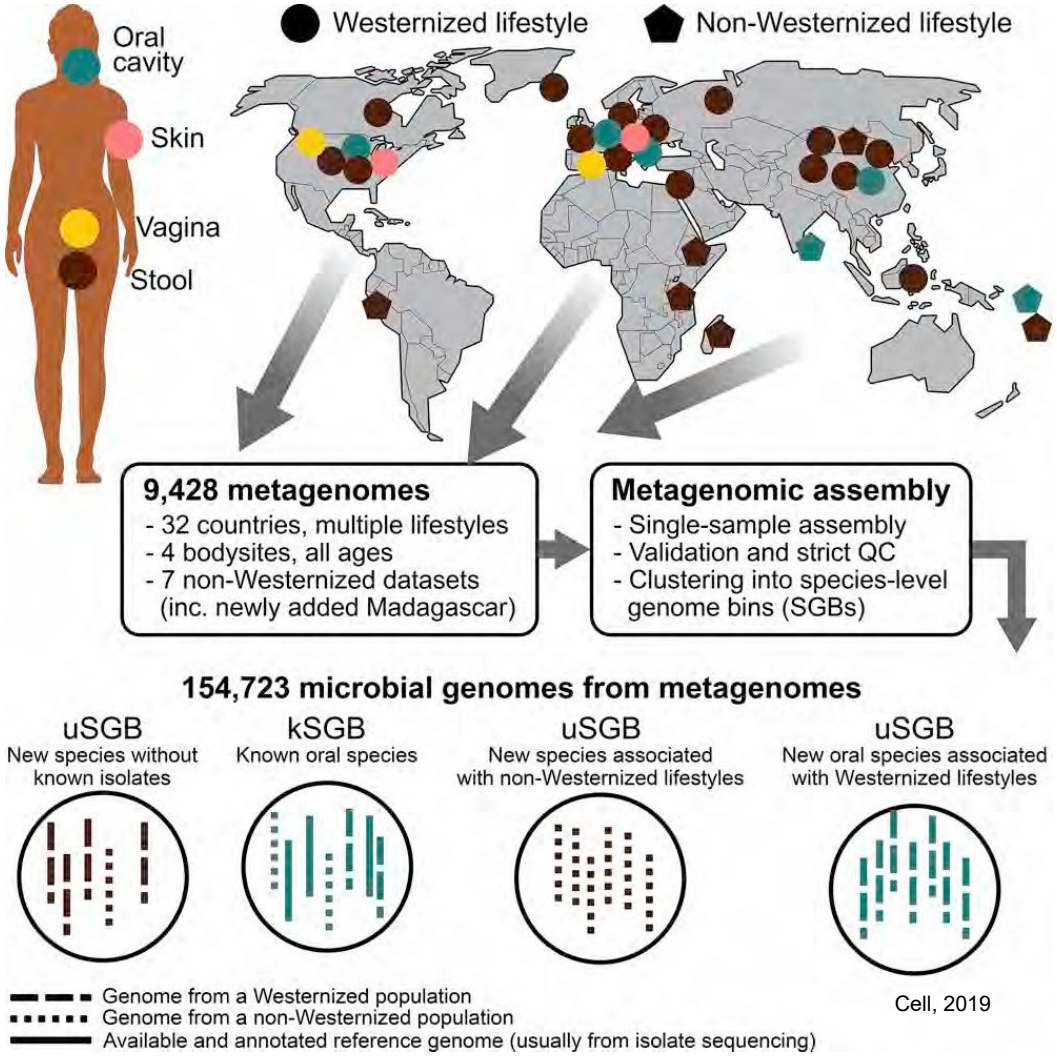
16S Sequencing	Metagenomic analysis	Metatranscriptome analysis
Identifies only a fraction of your gut bacteria; unable to identify nonbacterial microorganisms	cannot identify any RNA viruses or RNA bacteriophages	identifies all microorganisms living in the environment: bacteria, viruses, archaea, yeast, 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



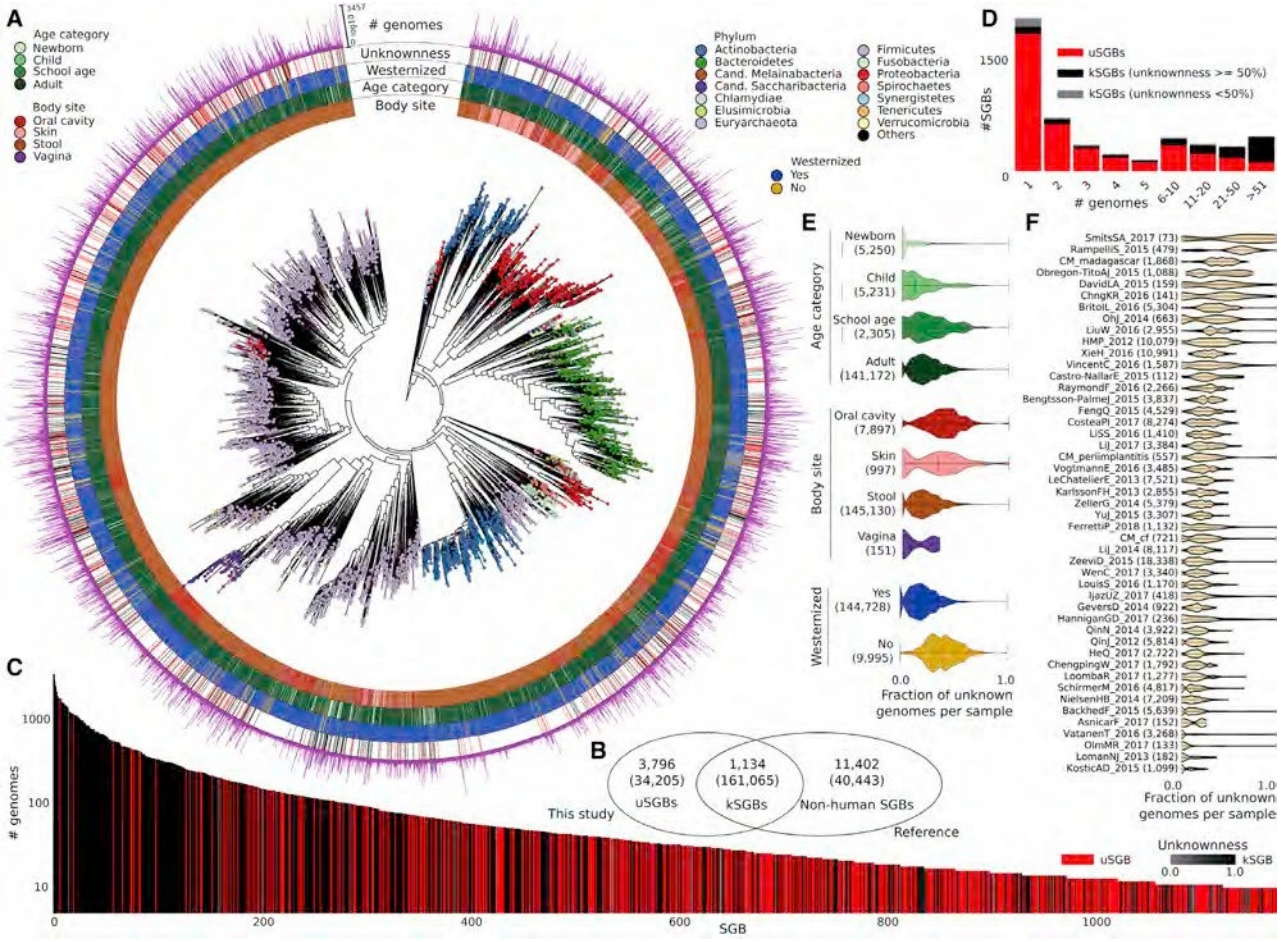
Metagenomics

- 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



Cell, 2019

National Center for Genome Resources



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

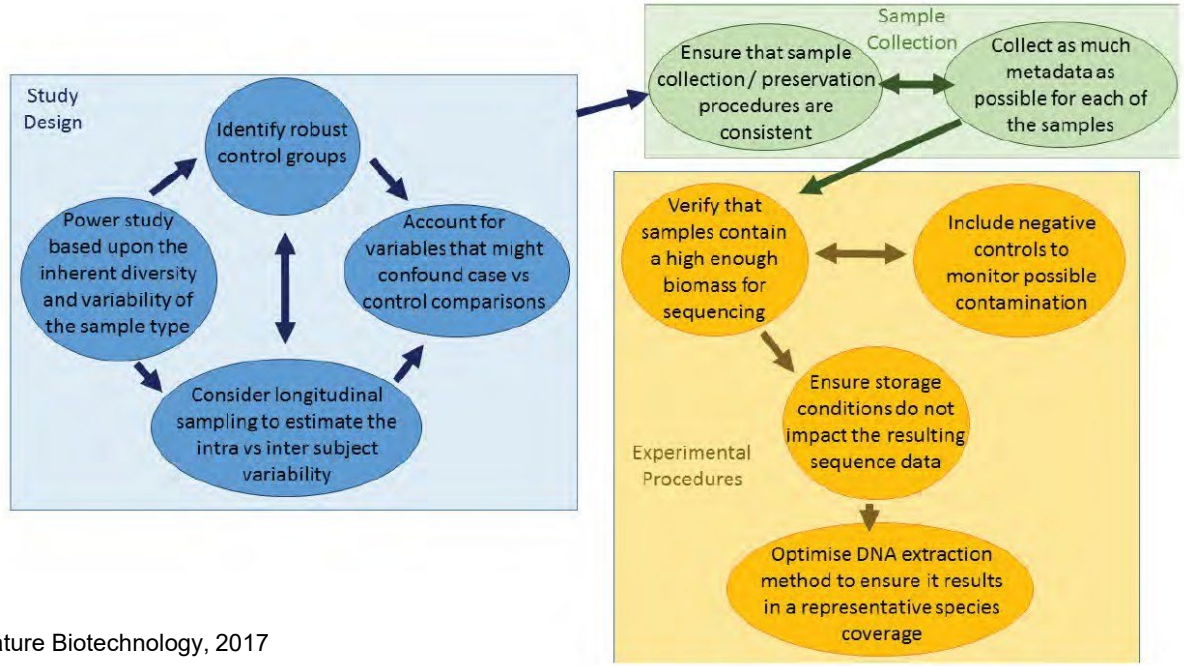
70,178 genomes assembled with higher than 90% completeness

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

Cell, 2019

NCGR
National Center for Genome Resources

Example Workflow to plan a Metagenomics Study

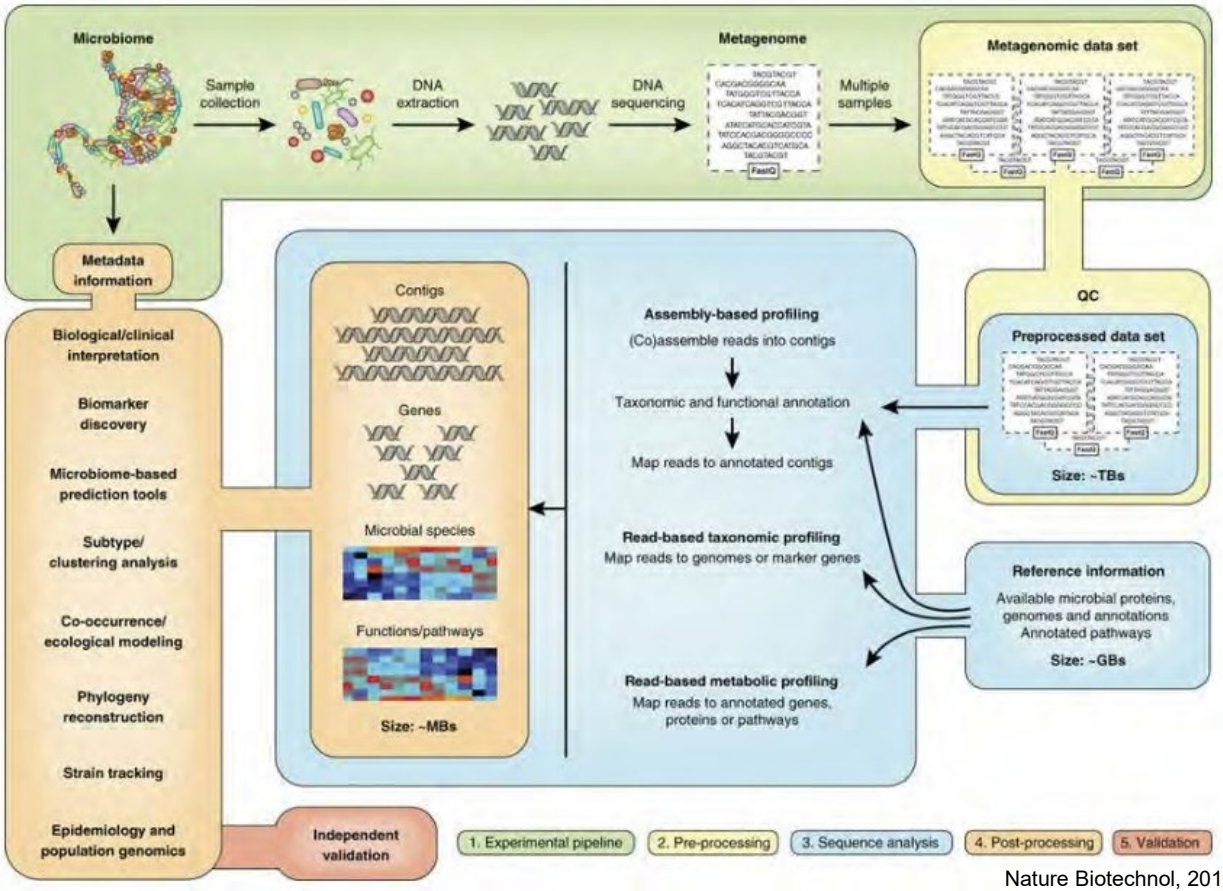


Nature Biotechnology, 2017

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

NCGR
National Center for Genome Resources

Shotgun Metagenomics



Nature Biotechnol, 2017

Metagenomics: Untargeted sequencing of all microbial genomes present in a sample

- Study design and experimental protocol
- Computational pre-processing
- Sequence analysis
- Post-processing
- Validation

NCGR
National Center for Genome Resources

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

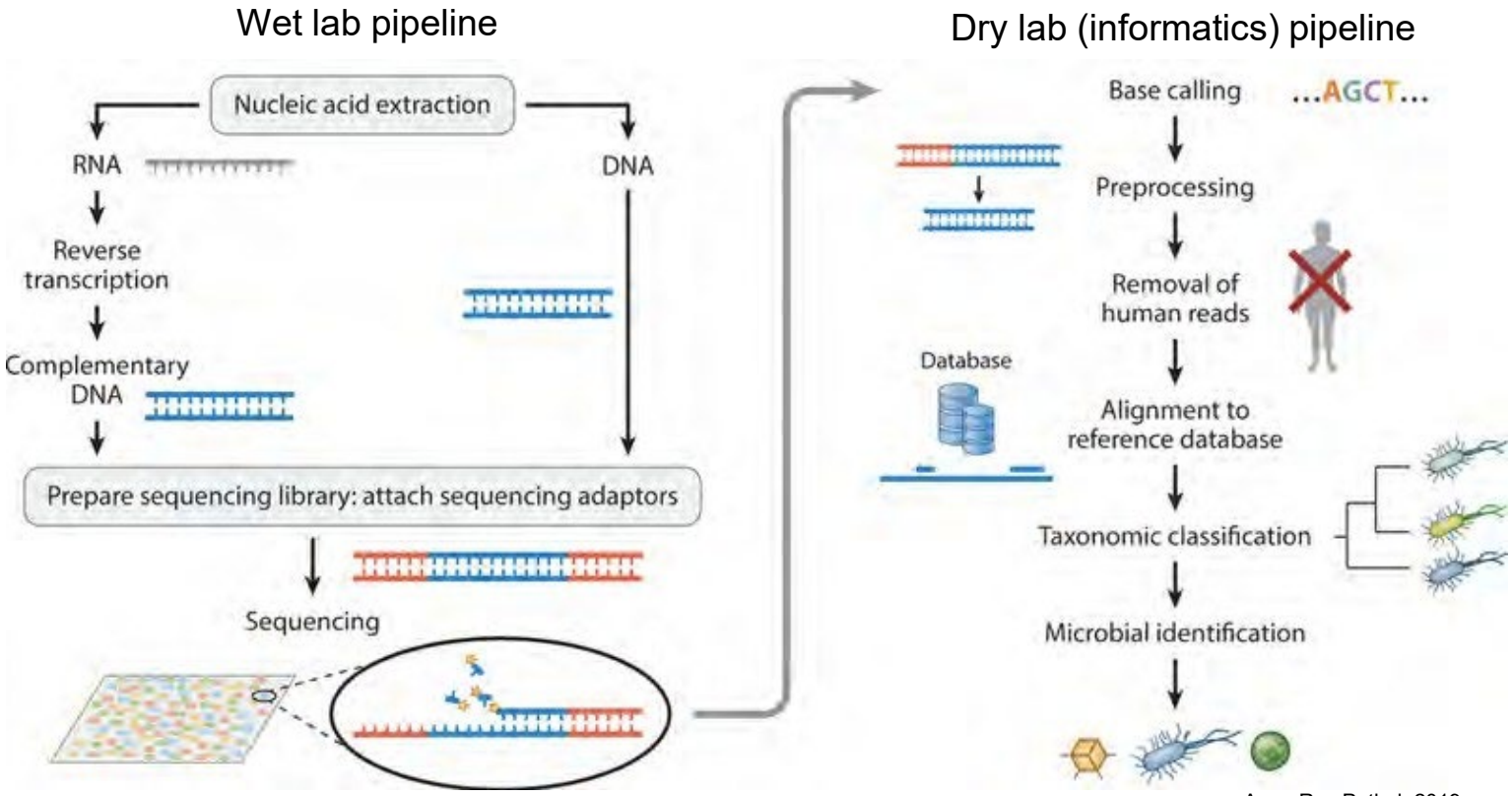
What do you think?

Nature Biotechnol, 2017

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

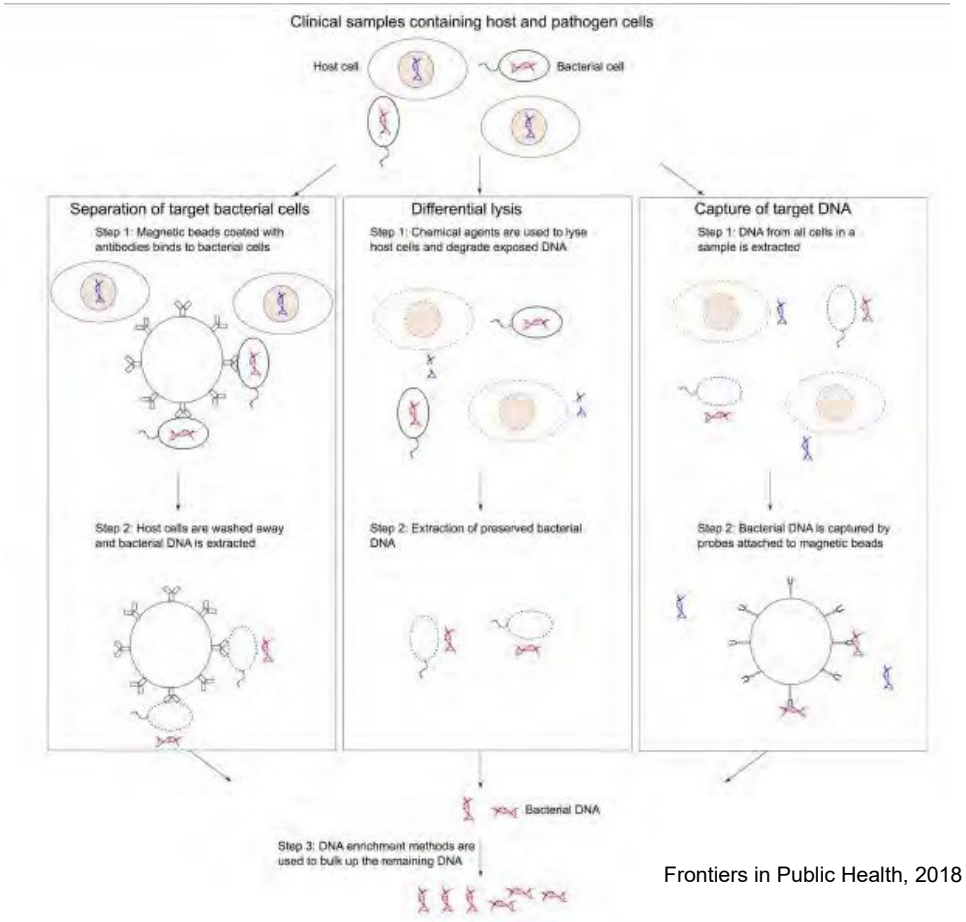
	Assembly-based analysis	Read-based analysis ('mapping')
Comprehensiveness	Can construct multiple whole genomes, but only for organisms with enough coverage to be assembled and binned.	Can provide an aggregate picture of community function or structure, but is based only on the fraction of reads that 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



Annu Rev Pathol. 2019

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



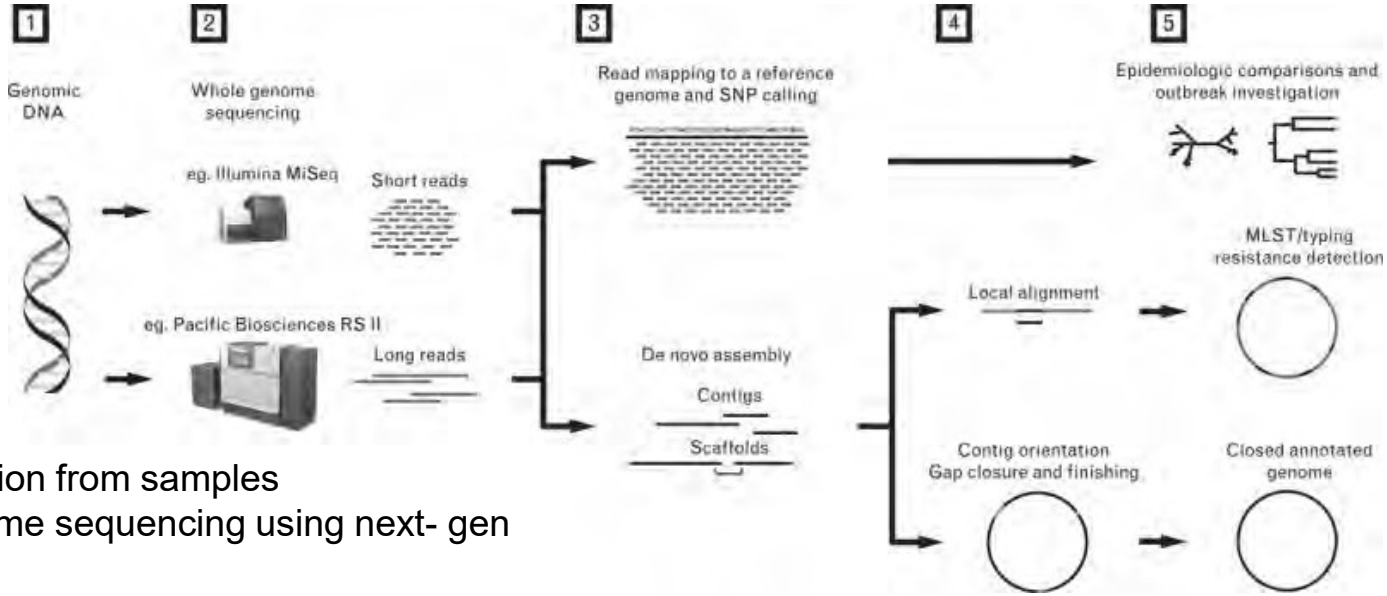
Frontiers in Public Health, 2018

Pretreatment methods for metagenomics:

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

NGS and pathogen detection

Whole genome sequencing workflow

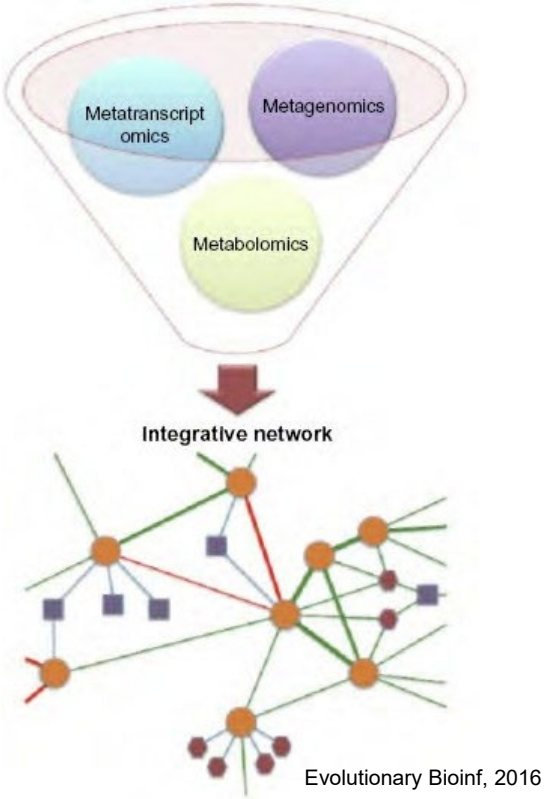


1. DNA extraction from samples
2. Whole genome sequencing using next- gen sequencers
3. Reference based SNP calling to perform phylogenetic analysis to assist with epidemiological outbreak
4. Resulting assembly used for typing and resistance detection
5. Closed genome used for further analysis

Pathology, 2015

National Center for Genome Resources

Where we are headed!



Evolutionary Bioinf, 2016

Integrated networks for multi omics data

NCGR
National Center for Genome Resources

Published studies

1. Whole genome metagenomic or metatranscriptomic?
2. What are the samples?
3. How many samples?
4. How many replicates?
5. What sequencing technologies?
6. How much sequencing coverage?
7. Sample complexity?
8. Community structure?
9. Assembly?
10. Functional?
11. Conclusions?