Next-gen sequencing technologies and data generation

Miguel Semedo

2024/09/04

Funding

Sponsorship

Support

A long way to the ”Molecular Microbiology Era” (1980’s)

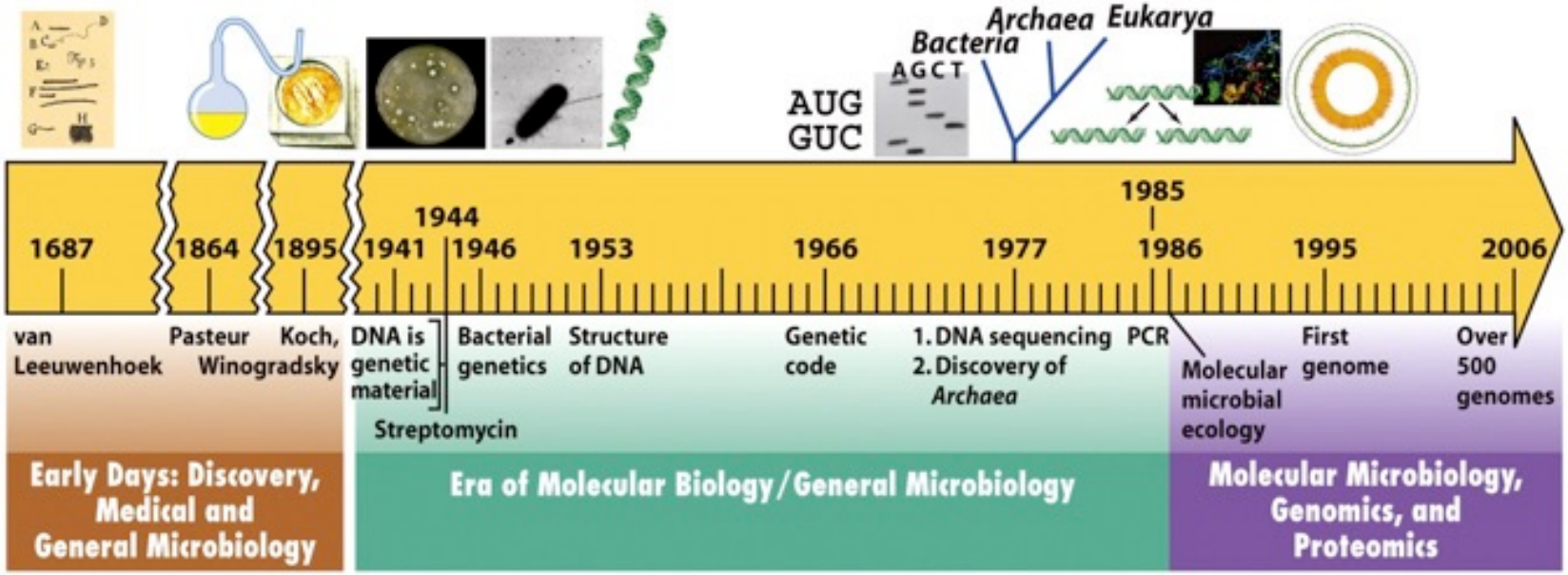


Figure 1-17 Brock Biology of Microorganisms 11/e

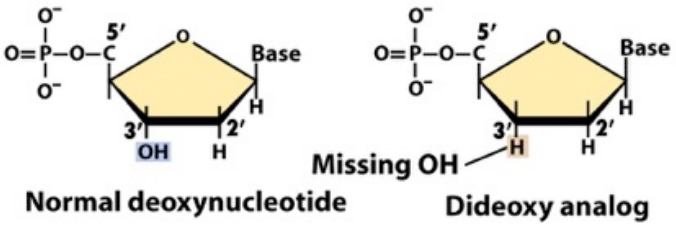
- **1953 (Watson & Crick):** DNA structure (Nobel prize in Medicine and Physiology)
- **1977 (Sanger et al.):** DNA sequencing (Nobel prize in Chemistry)
- **1977 (Woese & Fox):** Ribosomal RNA analysis to recognize a third form of life (Archaea), distinct from Bacteria and Eukaryotes.

Sanger sequencing - 1977

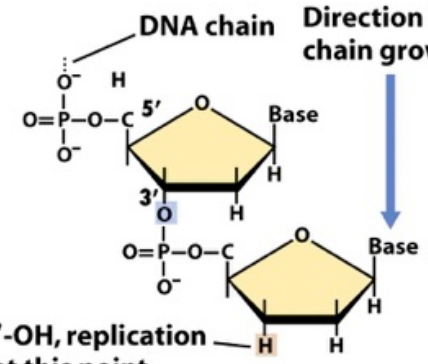
Chain-terminating dideoxynucleotides (ddNTPs) +

DNA Polymerase +

Electrophoresis

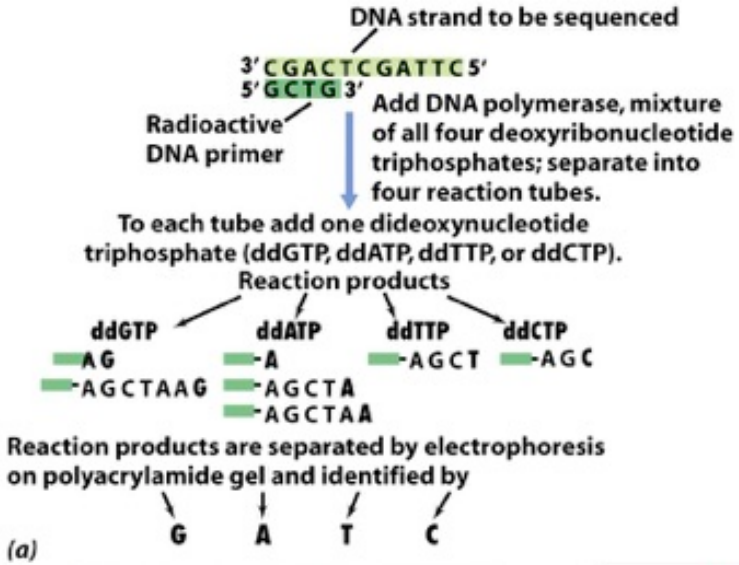


(a)



(b)

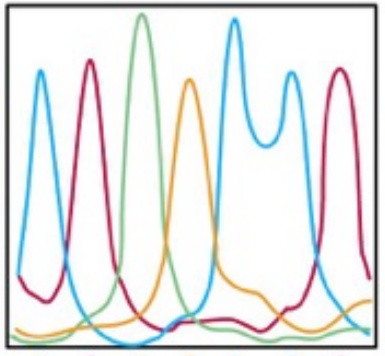
Figure 7-25 Brock Biology of Microorganisms 11/e
© 2006 Pearson Prentice Hall, Inc.



(a)

Sequence reads from bottom of gel as A G C T A A G. Sequence of unknown is 3′ T C G A T T C 5′.

(b)



(c)

Figure 7-26 Brock Biology of Microorganisms 11/e
© 2006 Pearson Prentice Hall, Inc.

Sanger sequencing – since 1977

- Read length: 400 – 900 bp
- Reads per run: ~ 400
- Total bases per run: < 360 kb
- Time per run : ~3 hrs
- Cost per 1M bases: ~$500/Mb
- Qualities: high-quality (99.9999%), long reads
- Limitations: Slow, expensive, laborious, sequencing depth
- But, for ~30 years was virtually the only sequencing method
- **Human Genome Project (1990-2003): more than 10 years, hundreds of labs, for one species genome. What about nowadays?**

“Molecular Microbiology Era”

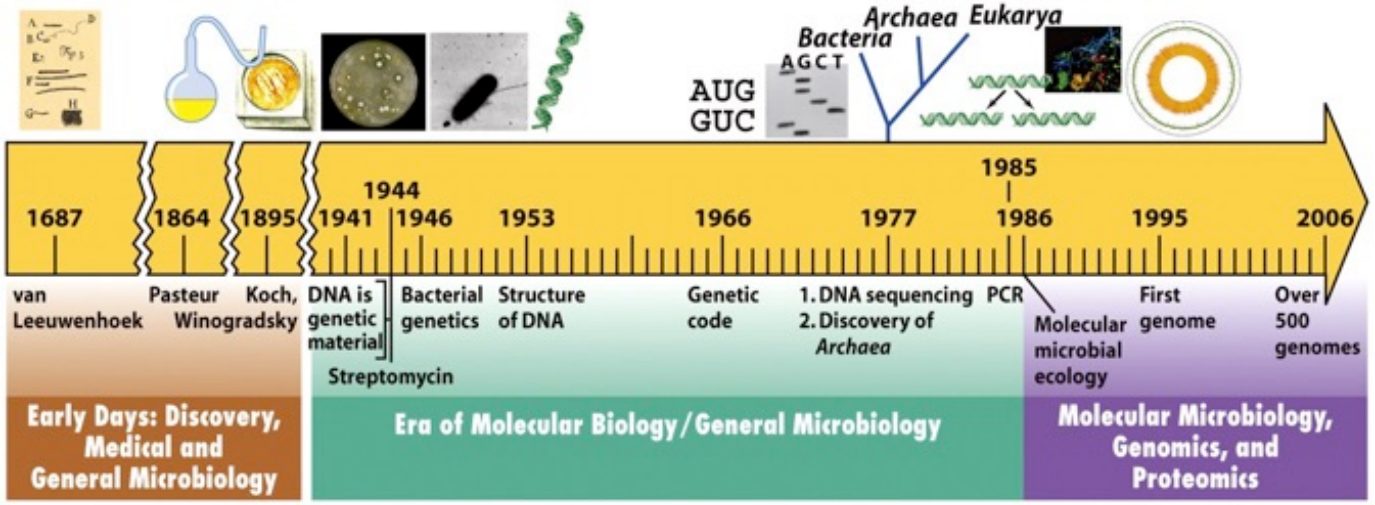


Figure 1-17 Brock Biology of Microorganisms 11/e

- **1985 (Kary Mullis):** Polymerase chain reaction - PCR (Nobel in Chemistry).
- **1986 (Norman Pace):** Investigating uncultivated microorganisms in environmental samples. DNA/RNA extraction from the environment.
- **1995 (Venter and Smith):** Completed first genome sequence of a bacteria
- **2004 (Venter et al.):** large scale of environmental genome sequences from the Sargasso Sea (shotgun sequencing)

Since then, various sequencing technologies and molecular methods were developed to determine microbial diversity in various environments

“Molecular Microbiology Era” (Next-gen sequencing)

NEXT-GENERATION SEQUENCING: high-throughput sequencing (generating hundreds of millions of sequences in a single run). Also known as “massive parallel sequencing”, ”2nd generation sequencing”, “3rd generation sequencing”, NGS.

Early days (2005 – 2020)

454/Roche

Pyrosequencing

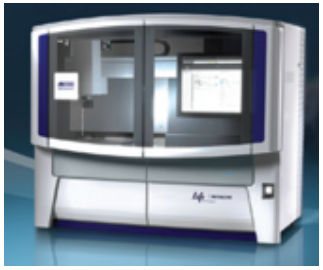
- Read length: 750 – 1000 bp
- Reads per run: ~ 1,000,000
- **Total bases per run: < 1 Gb**
- Time per run : ~24 hrs
- Cost per 1M bases: ~$10
- long reads
- Expensive

Ion Torrent PGM/Life Technologies

Semicondutor (pH change)

- Read length: 35 – 400 bp
- Reads per run: ~ 12,000,000
- **Total bases per run: < 4 Gb**
- Time per run : ~2 hrs
- Cost per 1M bases: ~$1
- low-cost, fast
- Error rate (2%)

SOLiD/Applied Biosystems



Ligation

- Read length: 75 bp
- Reads per run: ~ 400,000,000
- **Total bases per run: < 30 Gb**
- Time per run : ~10 days
- Cost per 1M bases: ~$0.13
- accuracy (99.9%), low-cost
- Short reads

Metagenome sequencing technologies evolution (JGI)

Sanger Sequencing to Next-Gen Sequencing by Synthesis

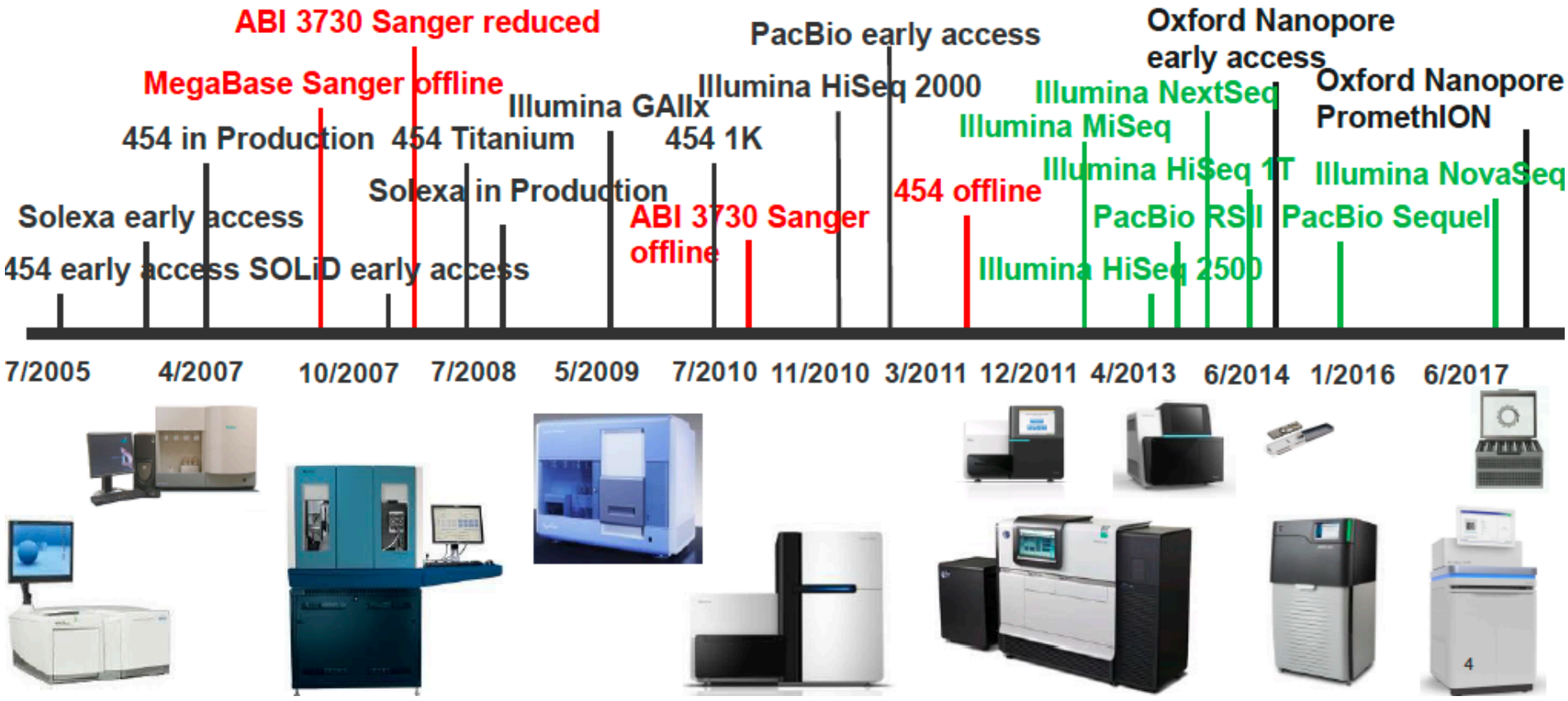


Image courtesy of JGI.

“Molecular Microbiology Era” (Next-gen sequencing)

The current “Big-three” (since 2010)

MiSeq-HiSeq-NovaSeq/Illumina

Sequencing by Synthesis

- Read length: 2 x 250 or 150 bp
- Reads per run: 17M – 6000M
- **7 Gb (MiSeq) – <u>1,200 Gb (HiSeq)/run</u>**
- Time per run : ~ 1 – 10 days
- Cost per 1M bases: ~$0.05-0.15
- Extremely high-throughput (most commonly used)
- Expensive instrument

RSII-Sequel/PacBio

SMRTS + HiFi

- Read length: 5000 – 20,000 bp
- Reads per run: ~ 50,000
- **Total bases per run: < 1 Gb**
- Time per run : ~ 1 hr.
- Cost per 1M bases: ~$0.10
- No amplification needed. Long reads.
- Expensive. Low yield.

PromethION-MinION/Oxford Nanopore

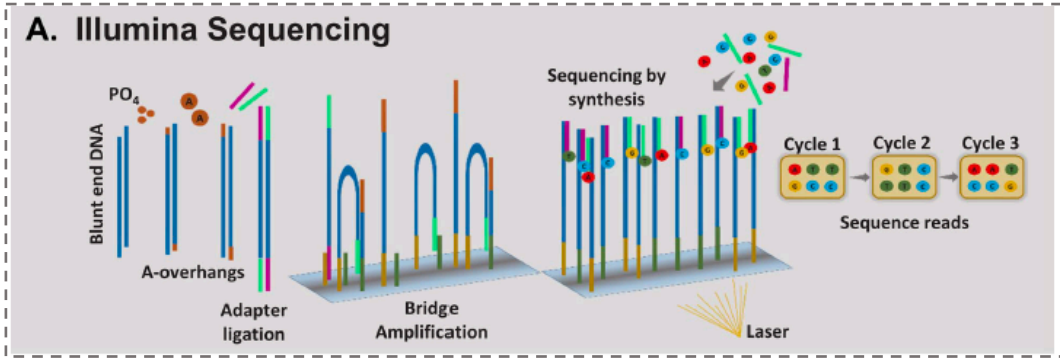
Minion

MinION

Current change through pores

- Read length: 300,000 – 4M bp
- Reads per run: ~ 1M
- **Total bases per run: 3 Gb – 15 Tb**
- Time per run : ~48 hrs
- Cost per 1M bases: 0.03
- Long reads, portable. Affordable. No amplification needed
- Low accuracy

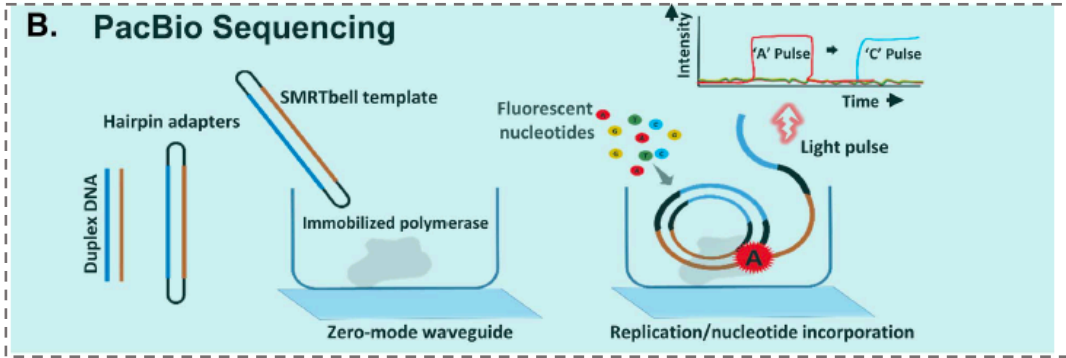
Metagenome sequencing technologies



Short reads (sequencing depth)

Single or Paired reads generated (e.g. 2 x 150). Useful overlap.

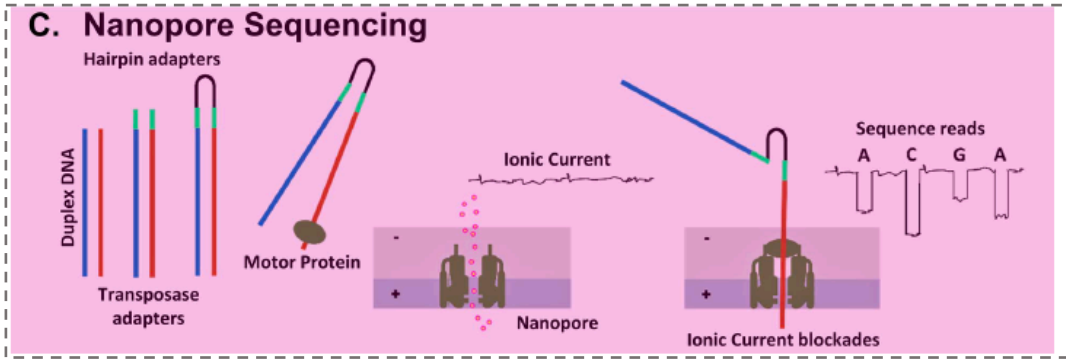
Amplification needed (bias)



Long reads (sequencing length)

No amplification needed

Sample quality limitation



Long reads (sequencing length)

No amplification needed

Sample quality limitation

Bharti and Grimm, 2021

Generated Data

Main output from any sequencing technology is a **FASTQ file**

- Text files that store biological information in the form of strings of nucleotides
- Formally defined in 2010
- Gold standard file format for storing and representing biological sequence data

@ and sequence identifier — @M04577:12:000000000-D2N7B:1:1101:15525:1827 1:N:0:23

raw sequence — TCTTTCGCTGGCGCGTAACAGCAAAAGCAGGTACTGATTCAAATTGTTTGTCTTTGTTCTTCTTTGCTTCATTCCTTTTTTCTTCTTTC
TTTCTTTTTTTTCTTTTTTTTTTTTTTTTTTTTTTTCCTTTATTCTTTTCTTTCTTTTTCTCTTTTTTTTTTCTCTCTTTTTTTTTTCTTTTTCTT
TTTTTCTTTTTTCCTTTTCTTCTTTTTTTTTTTTTTTTCTTTTTTTTTTTTTTCTTTTTTTCCTTTTT

metadata — +

Quality score per base call — 1AAAAFAAAADAGGGGGGEBBC00100010/B1AB22DD21222B1220002DDB1B0121221A11DD112D2222B11>01@22112BB1221111///0
11B11/////>/>/-------:/000000;000000;000000;9/00;0000:-9-///;//:;9/----9/://9B//:/:/;9////:9@///:;B//;///;//------9---9///9//;9-9----;/;//:/-9/////:/

What is a quality core?

$Q_{score} = -10 \log_{10} p$

p = probability that some base was called incorrectly

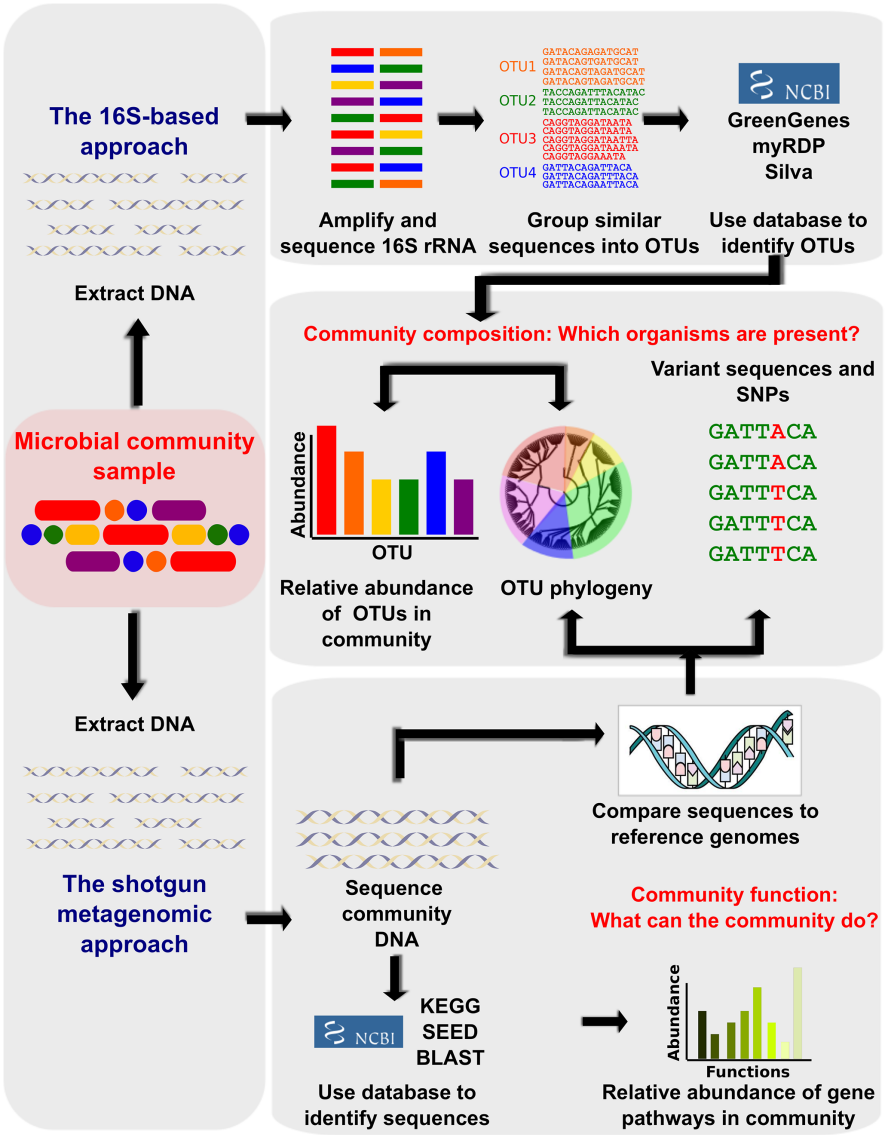
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Table source: wikipedia.org

General Applications of NGS

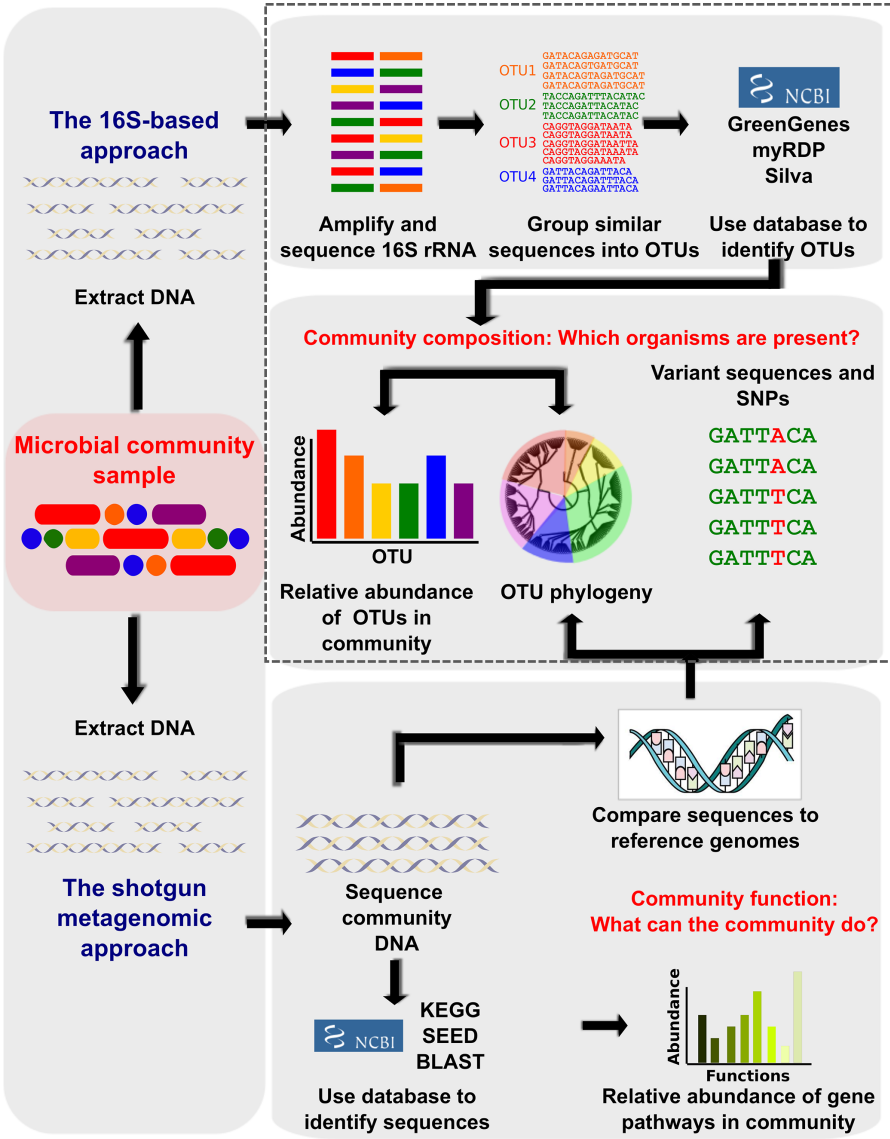
- Genome sequencing
- RNA-Seq (gene expression, exon-intron structure)
- CHIP-Seq (protein-DNA interactions)
- Single nucleotide polymorphisms
- Epigenetics
- Metagenomics: "*genomic analysis of microorganisms by direct extraction ~~and cloning~~ of DNA from an assemblage of microorganisms*" (Handelsman et al., 2004). Aka as environmental and community genomics.

Targeted and Untargeted Metagenomics



Morgan and Huttenhower, 2012. PLoS Comput Biol 8(12):e1002808.

Targeted and Untargeted Metagenomics



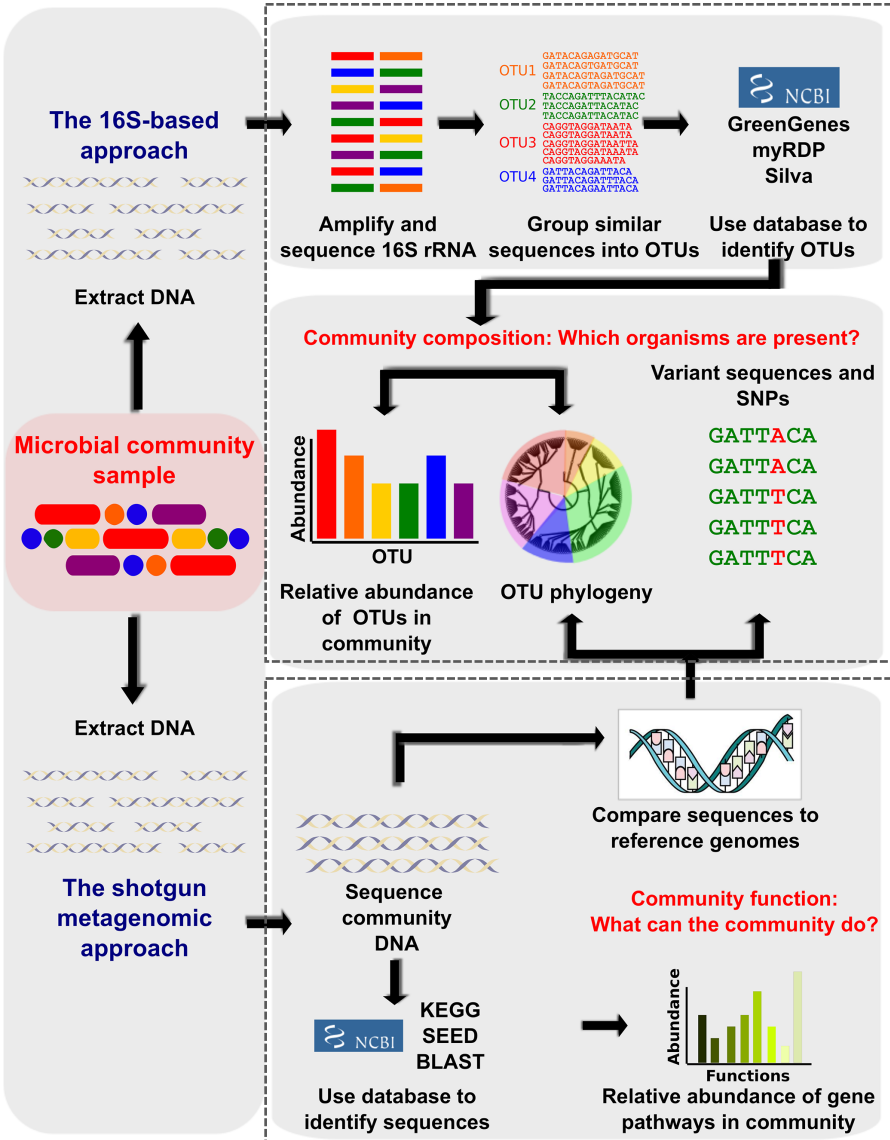
Targeted Metagenomics / Metabarcoding / Amplicon Sequencing

Marker genes amplification (e.g. 16S rRNA, others)

- Who's there (in general)?
  - With taxonomic markers: What can they possibly do (functional inference)?
  - With functional genes: Who's doing this specific function (potentially)?

Morgan and Huttenhower, 2012. PLoS Comput Biol 8(12):e1002808.

Targeted and Untargeted Metagenomics



Targeted Metagenomics / Metabarcoding / Amplicon Sequencing

Marker genes amplification (e.g. 16S rRNA, others)

- Who's there (in general)?
  - With taxonomic markers: What can they possibly do (functional inference)?
  - With functional genes: Who's doing this specific function (potentially)?

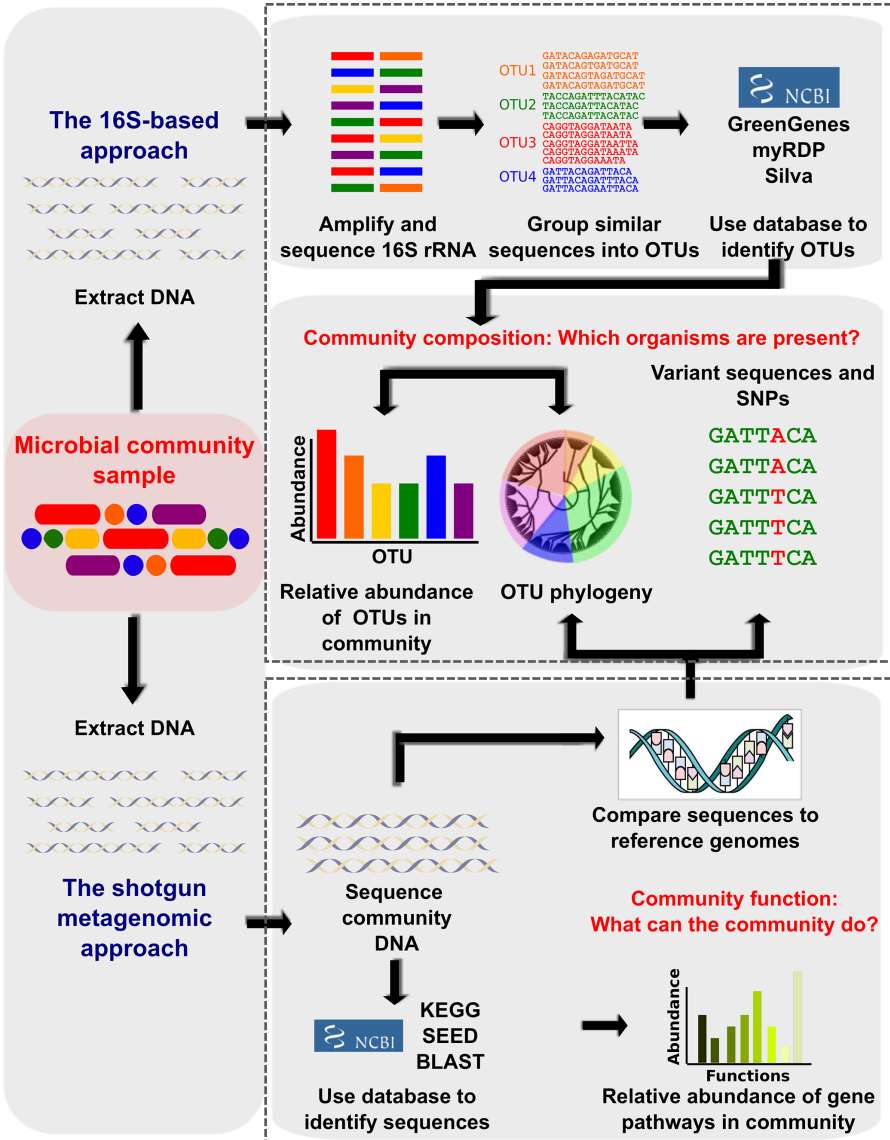
Untargeted Metagenomics / Shotgun Sequencing / Whole Genome Sequencing

All gene fragments (no PCR amplification)

- Who's there (in general)?
- What can they do (potentially)? Across infinite functions. **FUNCTIONS UNCHAINED**!

Morgan and Huttenhower, 2012. PLoS Comput Biol 8(12):e1002808.

Targeted and Untargeted Metagenomics



Targeted Metagenomics / Metabarcoding / Amplicon Sequencing

MAIN LIMITATIONS:

- PCR-biased.
- Functional discovery limited to taxonomic classification (for 16S).

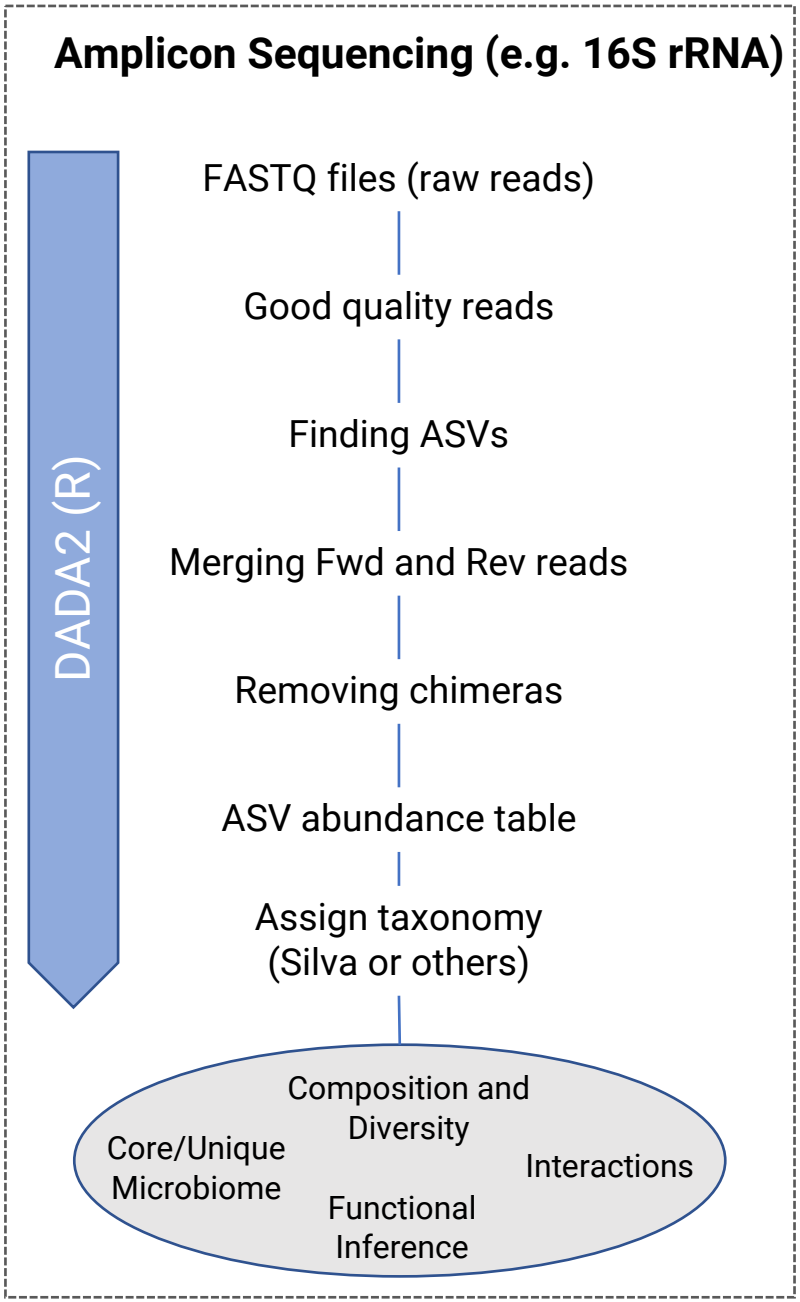
Untargeted Metagenomics / Shotgun Sequencing / Whole Genome Sequencing

MAIN LIMITATIONS:

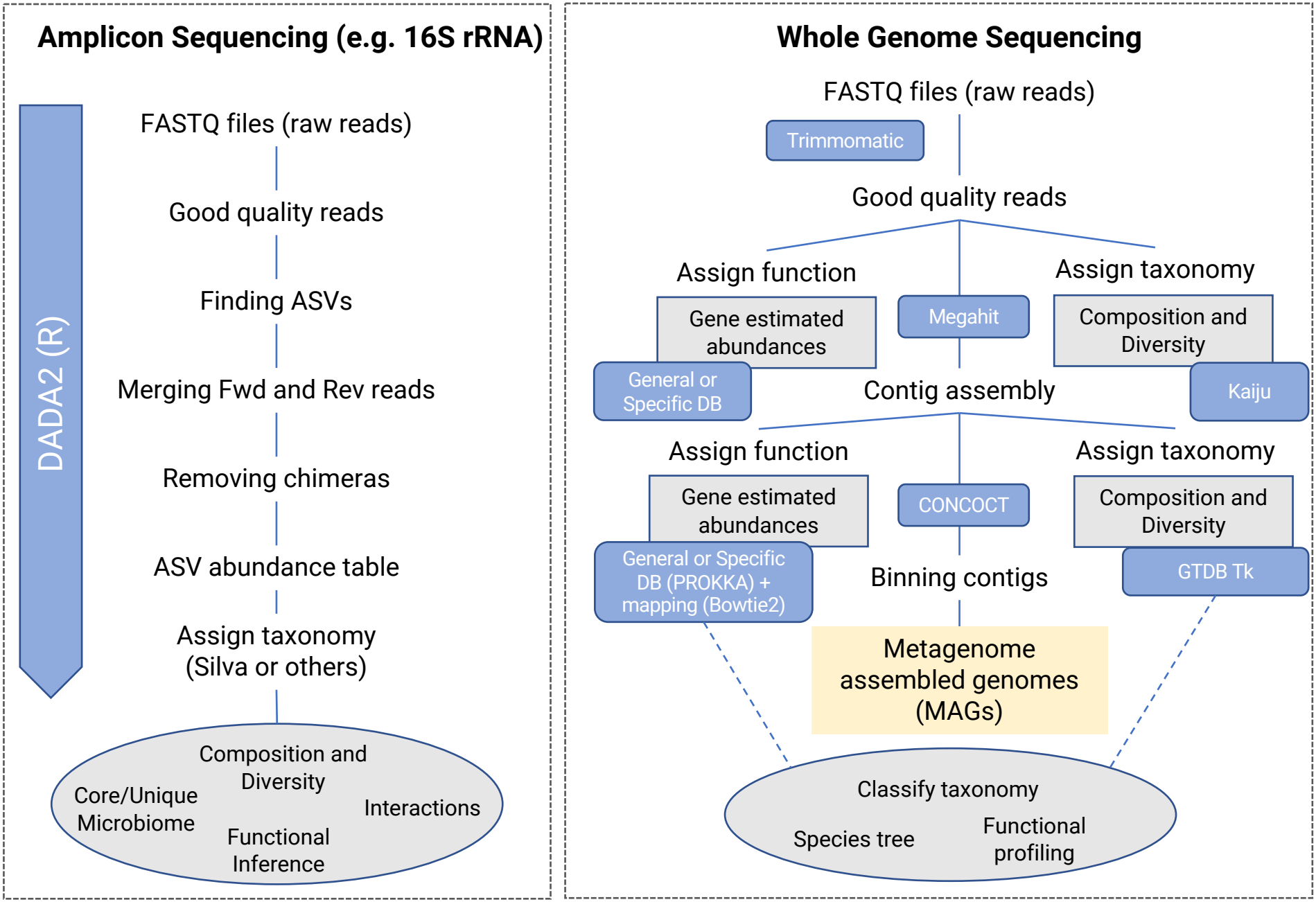
- Sequencing depth.
- Cost (€).

Morgan and Huttenhower, 2012. PLoS Comput Biol 8(12):e1002808.

Targeted and Untargeted Metagenomics

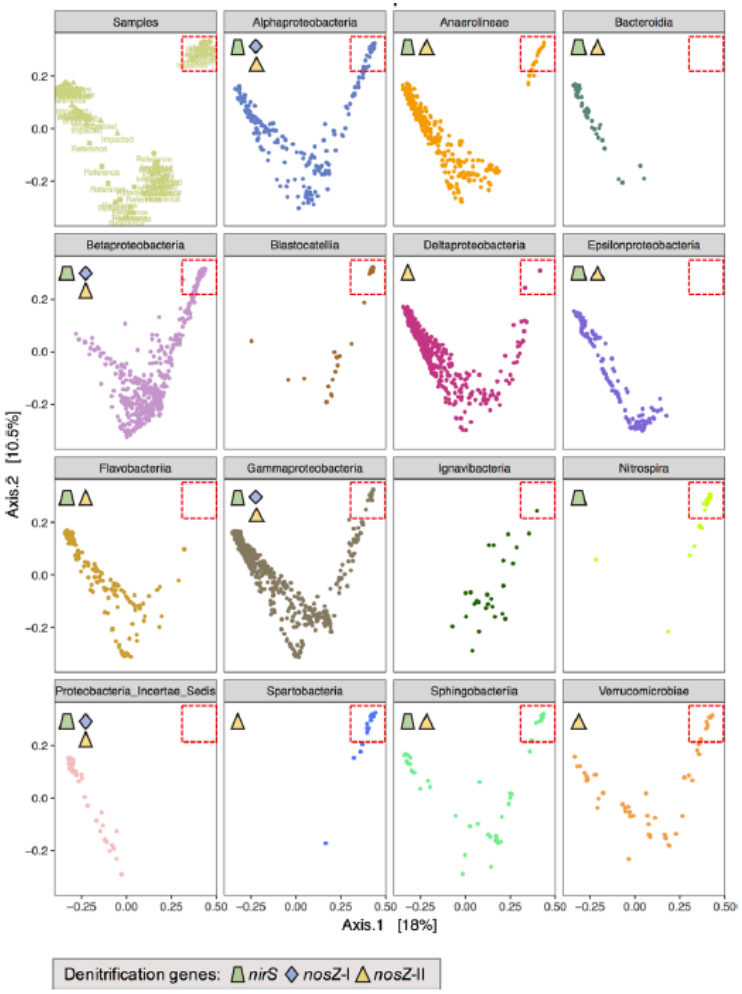


Targeted and Untargeted Metagenomics



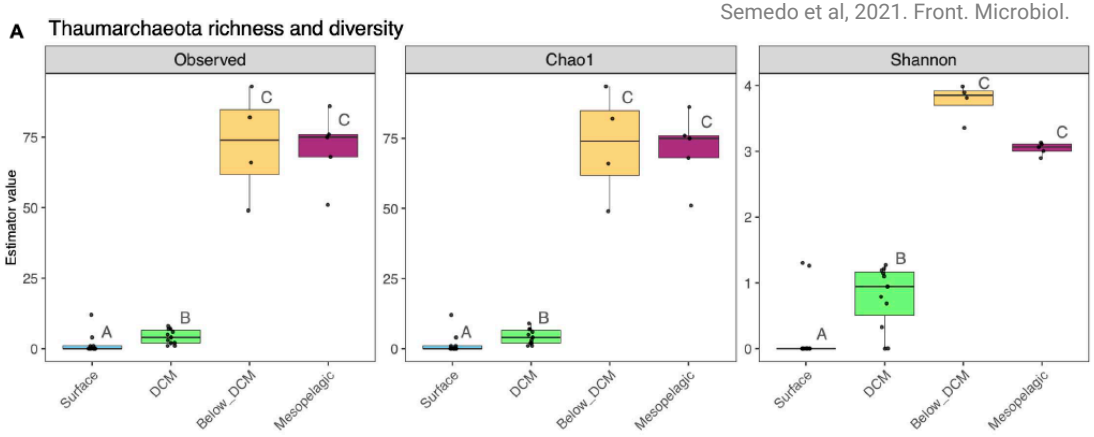
What to do with amplicon-based sequences

Composition and Diversity

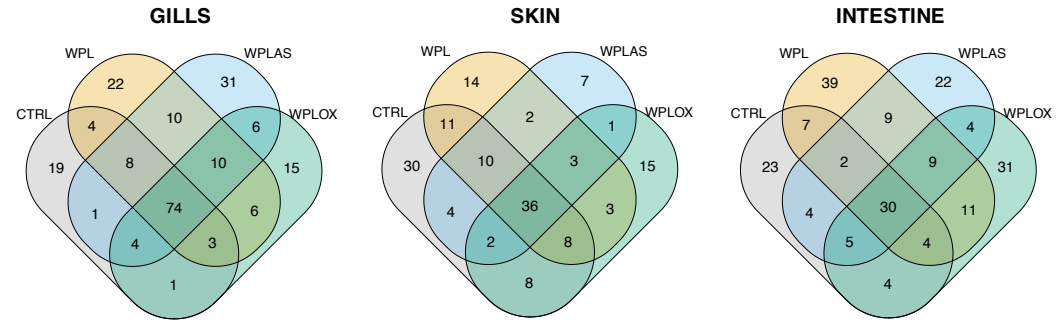


Semedo and Song, 2020. ES&T

Semedo et al, 2021. Front. Microbiol.



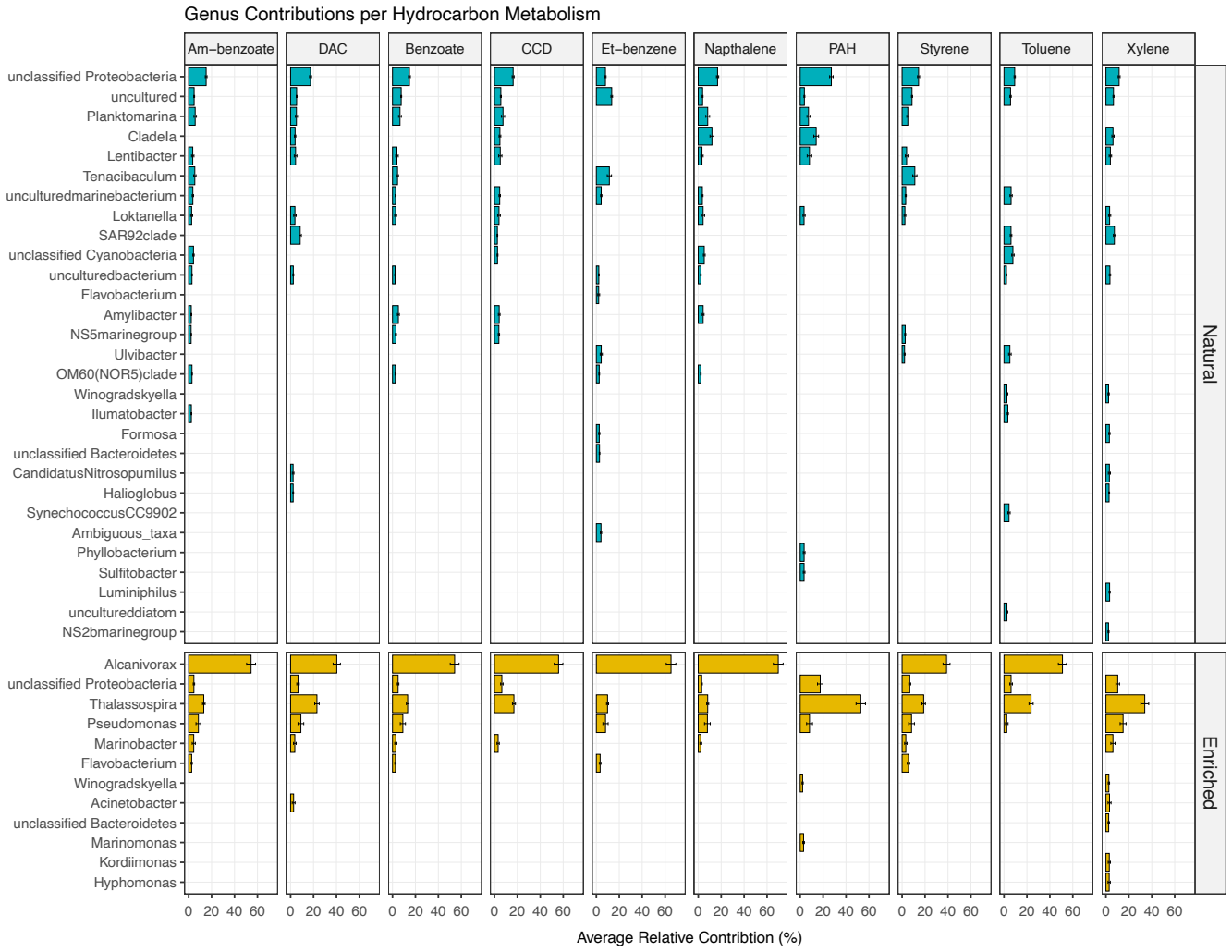
Core/Unique Microbiome



Semedo et al., 2024, in prep

What to do with amplicon-based sequences

Functional inference

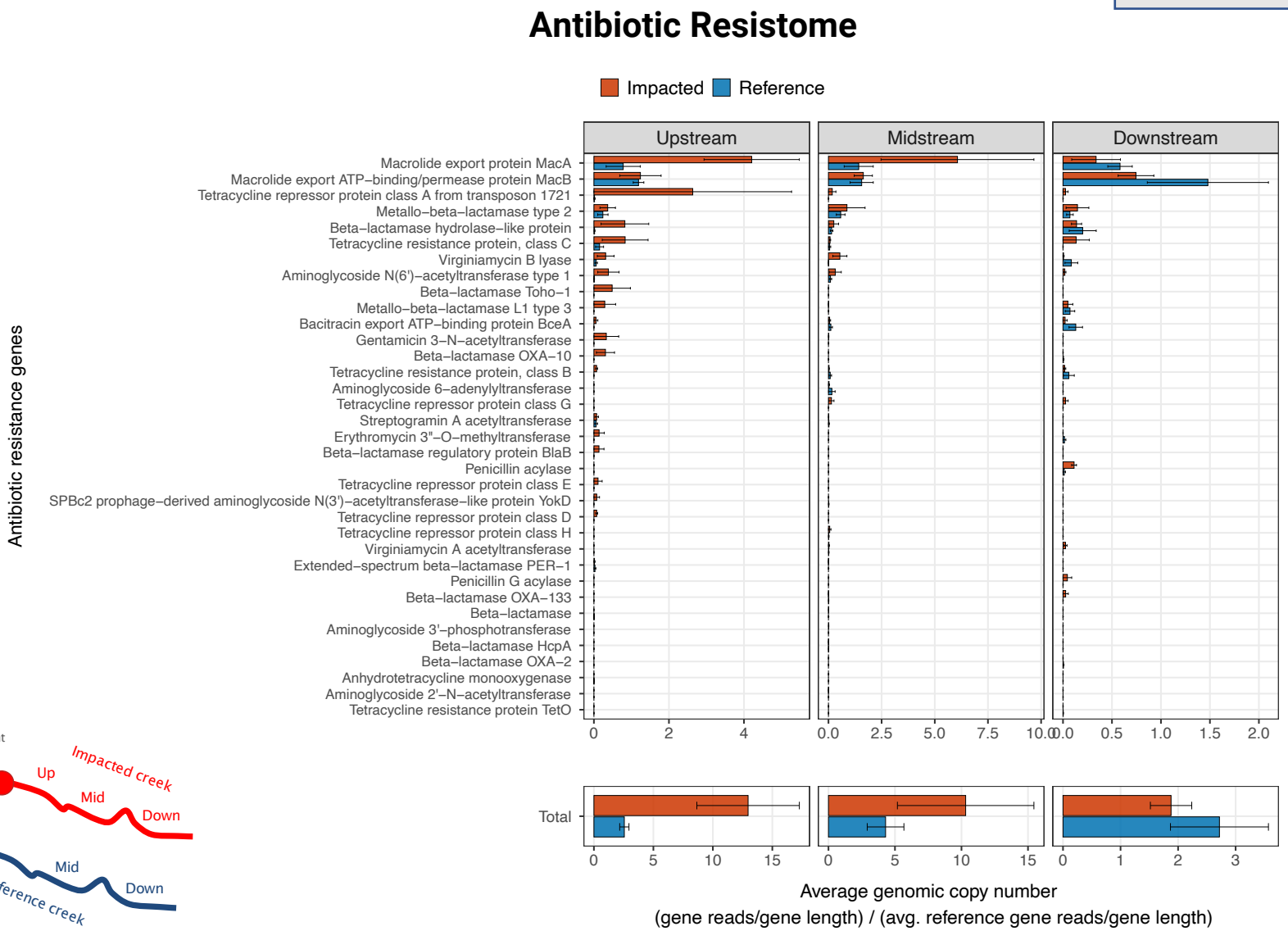


Cautiously

Bôto et al., 2021, Front. Microbiol.

What to do with metagenome sequences

Gene estimated abundances

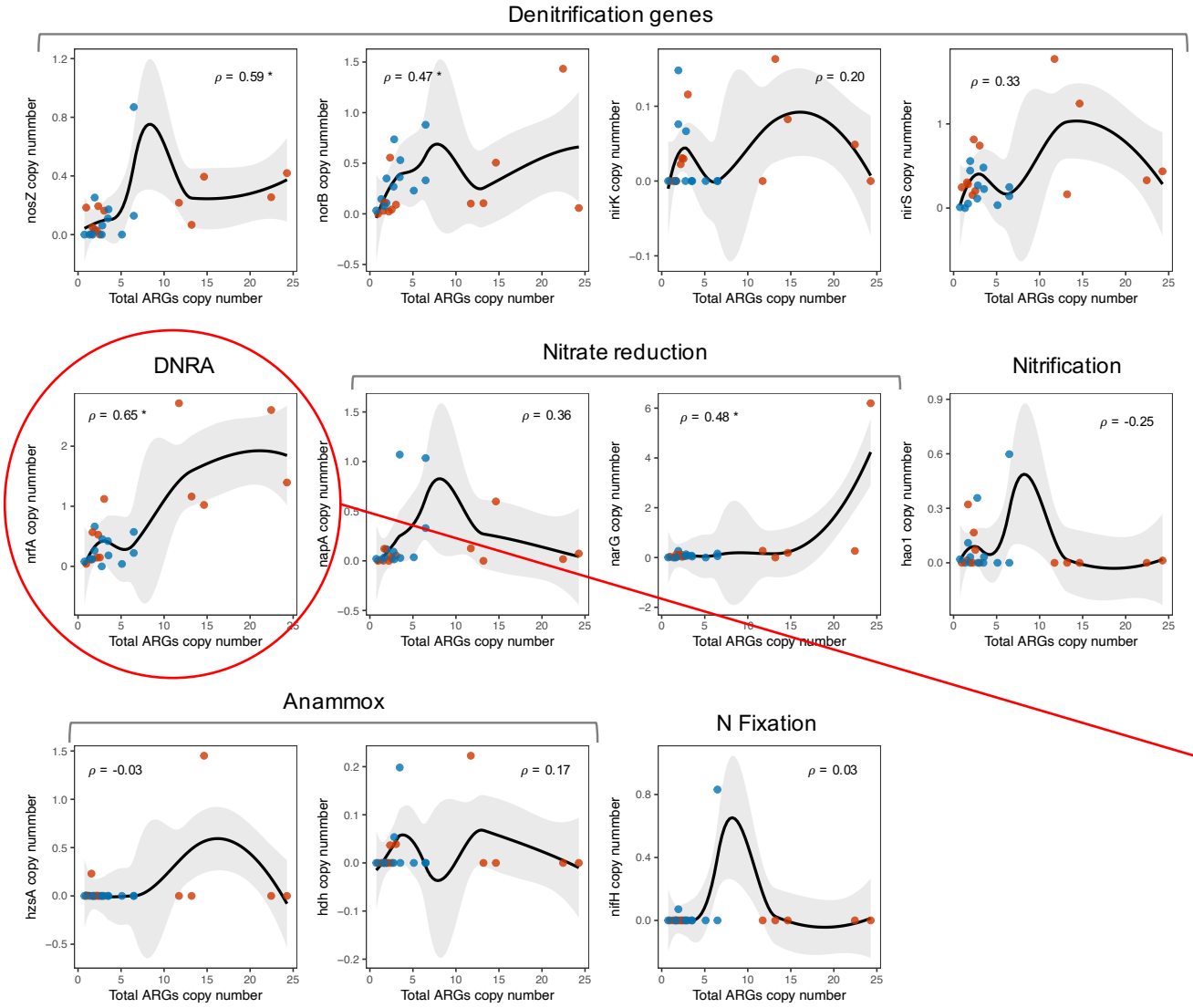


Semedo and Song, 2023. STOTEN

What to do with metagenome sequences

Gene estimated abundances

Gene or metabolic pathway relationships



Driven by co-selection or parallel responses to similar factors?

Semedo and Song, 2023. STOTEN

What to do with metagenome sequences

At the individual genome level!

Pathway co-ocurrence

Taxonomic identification

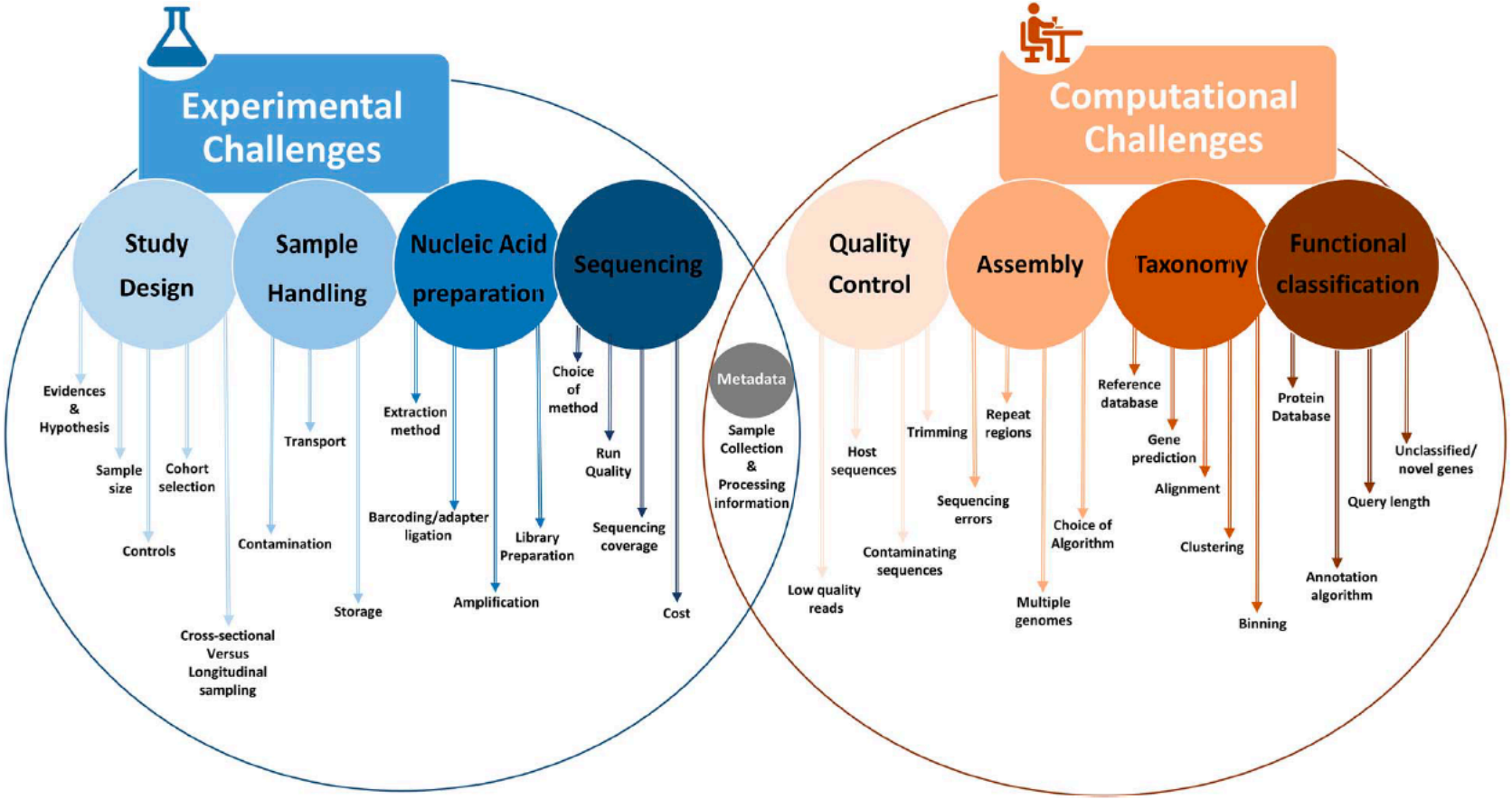
Functional profiling

Metagenome assembled genomes (MAGs)

Station	Creek	MAG ID	Genus (class/phylum)	# ARG	Nitrate red. narG	Nitrate red. napA	DNRA nrfA	Denitrification nosZ	Denitrification norB	Denitrification nirK	Denitrification nirS	Nitrif. hao1	Anammox hzsA	Anammox hdh	N fix. nifH	% comp.	% cont.
Upstream	Impacted	bin.010	Piscinibacter (Gamma-proteob.)	6		yes	yes				yes					95,6	5,7
		bin.013	UBA3961 (Bacteroidia)	6				yes								87,1	2,0
		bin.057	unclass. (Verrucomicrobiae)	6		yes										99,3	7,4
		bin.017	PHOS-HE28 (Bacteroidia)	5				yes								95,1	3,6
		bin.018	Aestuariivirga (Alpha-proteobacteria)	5	yes	yes	yes			yes						93,2	4,2
		bin.025	unclass. (Acidimicrobiia)	5												95,9	1,8
		bin.037	JAAZBK01 (Acidimicrobiia)	5		yes										94,0	6,3
		bin.042	JJ008 (Bacteroidia)	4				yes								77,6	1,1
		bin.055	JACADZ01 (Alpha-proteobacteria)	4	yes		yes									94,4	2,1
		bin.061	unclass. (Thermoanaerobaculia)	4	yes	yes										66,9	8,4
		bin.016	Methylomirabilis (Methylomirabilia)	3	yes	yes					yes			yes		90,2	3,5
		bin.038	UBA7227 (Anaerolineae)	3	yes	yes	yes				yes					70,4	4,2
		bin.046	JAAUPO01 (Cyanobacteriia)	3												84,1	1,0
		bin.047	unclass. (Kapabacteria)	3		yes		yes								86,9	0,1
		bin.050	Terricaulis (Alpha-proteobacteria)	3	yes											81,6	4,9
		bin.051	unclass. (Ignavibacteria)	3		yes	yes									86,5	1,9
		bin.056	unclass. (Myxococcota)	3	yes	yes					yes					60,7	4,4
		bin.006	unclass. (Gamma-proteob.)	2		yes		yes			yes					95,7	1,9
		bin.019	UBA8403 (Bacteoidia)	2												68,9	1,0
	Ref.	bin.009	Sulfuricella (Gamma-proteob.)	2		yes	yes	yes	yes		yes					85,7	6,9
		bin.004	SM1-50 (Thermoplasmatota)	0												93,2	3,2
		bin.013	PALSA-986 (Thermoproteota)	0												55,5	1,9

Semedo and Song, 2023. STOTEN

Challenges associated with NGS-based research



Bharti and Grimm, 2021

Questions?