

A top-down view of a petri dish containing various microbial cultures. The cultures are arranged in a grid pattern, with some showing distinct colors like yellow, white, and blue. The petri dish has a grid and numbers 1, 2, 3, 4 along the top edge. The background is dark, and the text is overlaid on the left side of the dish.

Introduction to metagenomics

13.1.2022, FindingPheno workshop

Dr. Katariina Pärnänen

Department of Computation

University of Turku

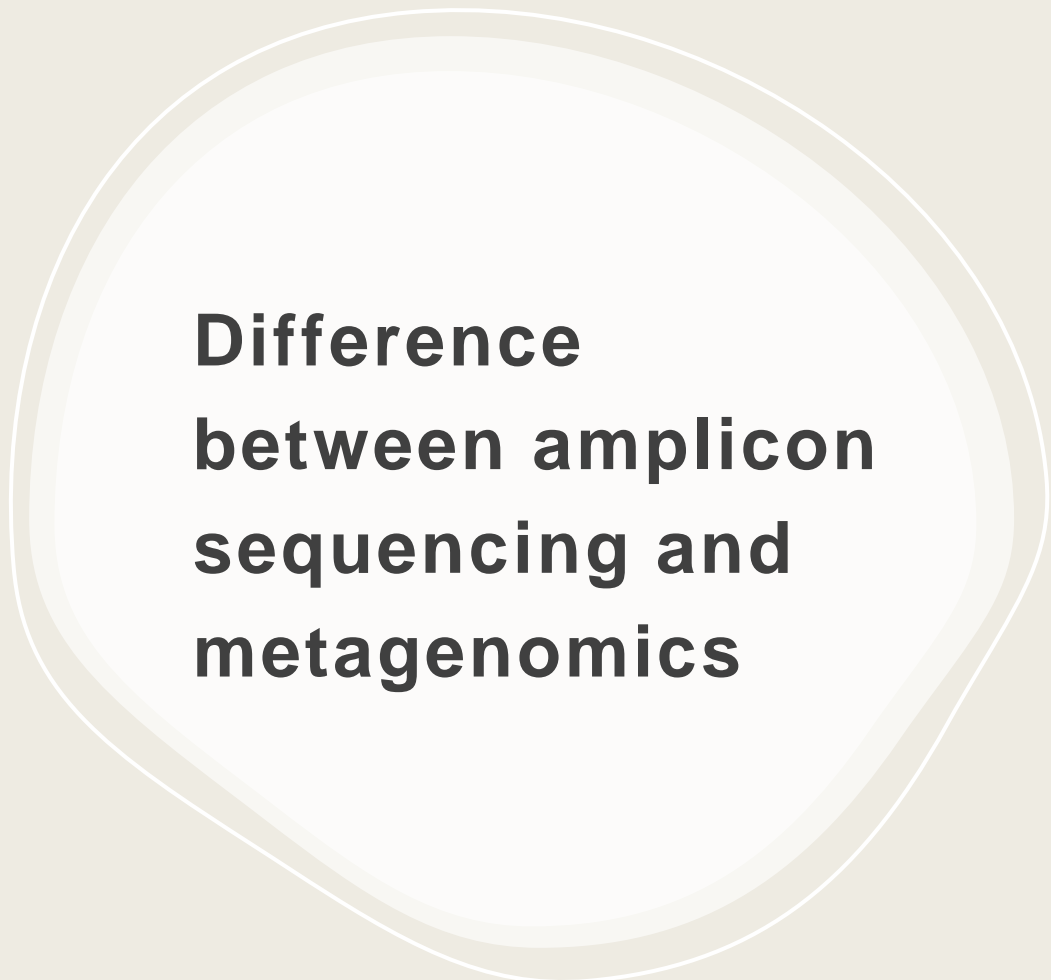
Division between culture dependent and culture independent

- Traditionally, bacteria and other microbes are studied using culture-based techniques where a single cell is isolated from a sample (**Robert Koch, Antony van Leeuwenhoek Julius Richard Petri**) and then pheno- (and genotyped).
- Culturing of microbes is still useful when investigating and confirming genotypic predictions of the functional capabilities of organisms, for example virulence or capacity to degrade a substrate



CULTURE INDEPENDENT TECHNIQUES – HISTORICAL AND MODERN

- **PCR**, or polymerase chain reaction, is used to amplify DNA in a sequence specific manner
- Uses short nucleotide oligos which bind to the target DNA.
 - Identifying samples or microbes containing specific genes (+/- results)
 - **PCR** requires *a priori* knowledge of the target sequence
- Terminal restriction fragment length polymorphism (**TRFLP**) is an old method used for characterizing microbial communities
 - Based on the position of a restriction sites in an amplified gene
- First and next generation sequencing methods also use PCR based techniques (amplicon sequencing)

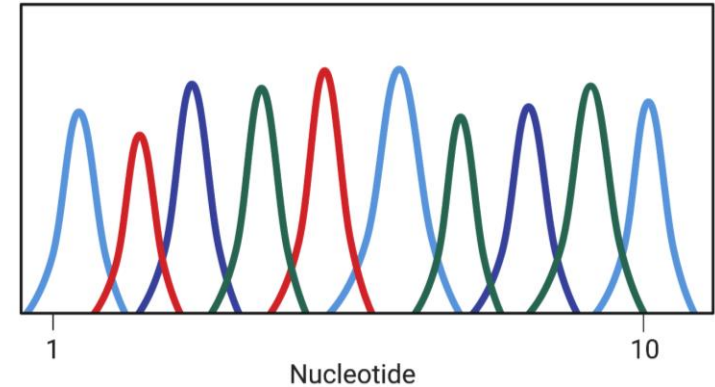


Difference between amplicon sequencing and metagenomics

- Amplicon sequencing = sequencing of PCR products amplified with target specific DNA primers, often refers to the amplification of the bacterial phylogenetic marker gene 16S rRNA gene
- Metagenomics or shotgun metagenomics = sequencing of random DNA from a sample without any preselection

Sequencing

- **High-throughput Sanger sequencing**
- Dye-terminators and capillary electrophoresis
 - 48/96 –capillars
 - Read length < 900 bp
 - < 90 Kbp / run
 - Sanger sequencing and improvements in 16 S rRNA gene PCR primers started the era of rapid expansion of the understanding of microbial diversity
 - Cultured isolates could not keep up
 - Only 1% to 0.001% of microbes in an environment can be isolated to pure cultures



life
technologies™

WHAT THEY DID BEFORE ILLUMINA?

- [Stein et al. 1996](#) recognized the limitations of amplicon sequence and pushed the field forward with the first attempt of metagenomic sequencing in Hawaiian ocean water
- [Stein et al. 1996](#) could not just perform a simple Illumina library preparation to get shotgun libraries.
- Instead, they had to use an E. coli Fosmid cloning vector to generate their library of large DNA fragments from the marine environment (30 liters of seawater were filtered!).

META|GEN|OME

“beyond the genome”

- First used to describe an approach for biosynthetic gene cluster (BGC) research by Handelsman et al. 1998
- Carl Woese introduced the idea of using 16S rRNA gene as a phylogenetic marker (Woese and Fox 1977) and said that phylogenetic research without the ability to “read” the genome is a “fruitless search”

What are the possible pros and cons for amplicon and shotgun metagenomic sequencing?

NGS sequencing

- Illumina systems currently provide 1 Gbp to 1 Tbp / run
- Run time 1–2 days
- Short read lengths

	Sequel II System	Sequel (I) System
SMRT Cell	SMRT Cell 8M	SMRT Cell 1M
Average Data Output*	~100Gb	~15Gb
Number of HiFi Reads >99% Accuracy*	Up to 4,000,000	Up to 500,000
Sequencing Run Time per SMRT Cell	Up to 30hrs	Up to 20hrs
Recommended species / genome size	Human (3Gb), Plant, or animal with more than 3Gb of Genome size	Plant, or animal with less than 3Gb of Genome size

*Number of HiFi reads is dependent upon the insert size and sample quality

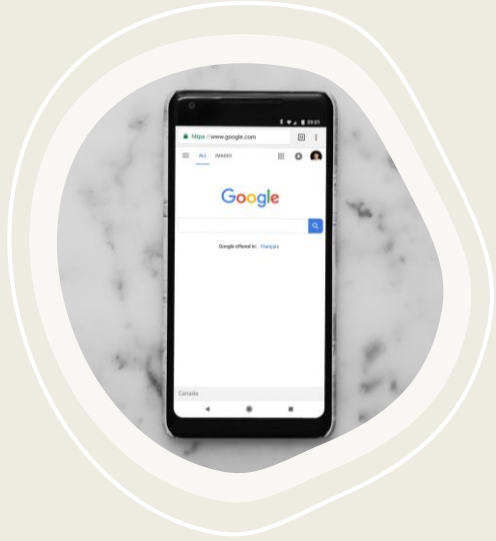
*Data Output is dependent upon the insert size and sample quality

Sequencing systems for every lab



Long read sequencing

- Third generation PacBio tens of kb length reads avg data output 15Gb to 100Gb
- “The PacBio RS II is ideal for whole-genome sequencing of small genomes, targeted sequencing, complex population analysis, RNA sequencing of targeted transcripts, and microbial epigenetics”



**Long-read sequencing -
From giant expensive machines to cheap and
small**

PacBio Sequel II HiFi reads:

- Read length 20 kb
- **30 Gbp /run**
- (99.92 % acc.)

MinION

-
- Read length: >4 Mb reads
 - 1-50Gb/flow cell
 - (~97-99% acc. depending on algorithm)

PromethION (available in early 2022)

- 280 Gbp / PromethION flow cell (~ 99.1 % acc.)
- PromethION devices have two to 48 flow cells
- **Real-time!**
- **Portable!**

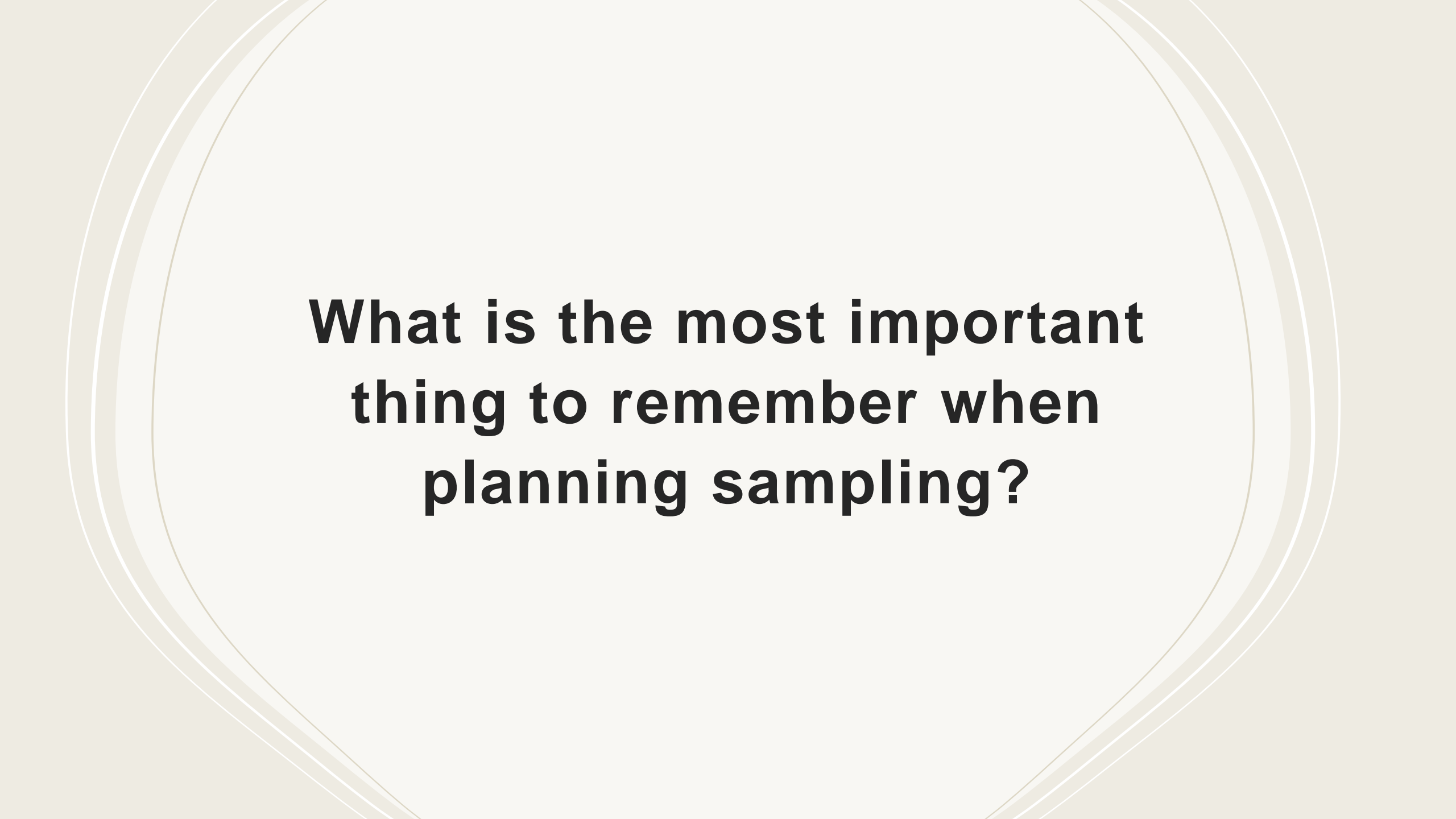


From samples to
sequences – the
most important
part happens
before you have
sequences

35000 samples were collected from all the world's oceans during the 2009–2013 expeditions.

IMAGE: Plankton: Noe and Christian Sardet / [Plankton Chronicles](#); Boat: F.Latreille / [Tara Expéditions](#)





**What is the most important
thing to remember when
planning sampling?**

Sampling

RESEARCH QUESTION

Sample so that you can investigate your hypotheses

How many samples?

Power estimations often impossible

Cross-sectional or time series?

Adequate metadata

Controls



Technical aspects to consider: DNA extraction

DNA extraction

Short fragments for Illumina or long read sequencing

Concurrent RNA sequencing

DNA concentration and total DNA mass requirements are different for different library preparation kits

Nextera XT kit is good for low concentration samples, can take anywhere from 1 to 100 ng of DNA depending on the lab doing the library prep

Short read sequencing or long read or both?

Will you assemble genomes or not?

Technical aspects to consider: Choosing sequencing technology and sequencing depth

Are you planning to assemble genomes?

How abundant are the pathways/taxa/genes you are looking for

How deep sequencing you need to answer your research questions

Shotgun sequencing depths can range anywhere from 1 Gb to 1 Tb of data per sample

Typical Illumina libraries (short reads) for fecal samples are from 4 Gb to 10 Gbs for read based approaches (and even assembly-based studies)

Soil and sediment samples up to 1 TB

Estimating the diversity of the sample before-hand with 16S rRNA amplicon sequencing or based on literature!

Note that diversity can differ greatly

– *For example, infants and adult individuals*



NextSeq 550* †



NextSeq 1000 & 2000*

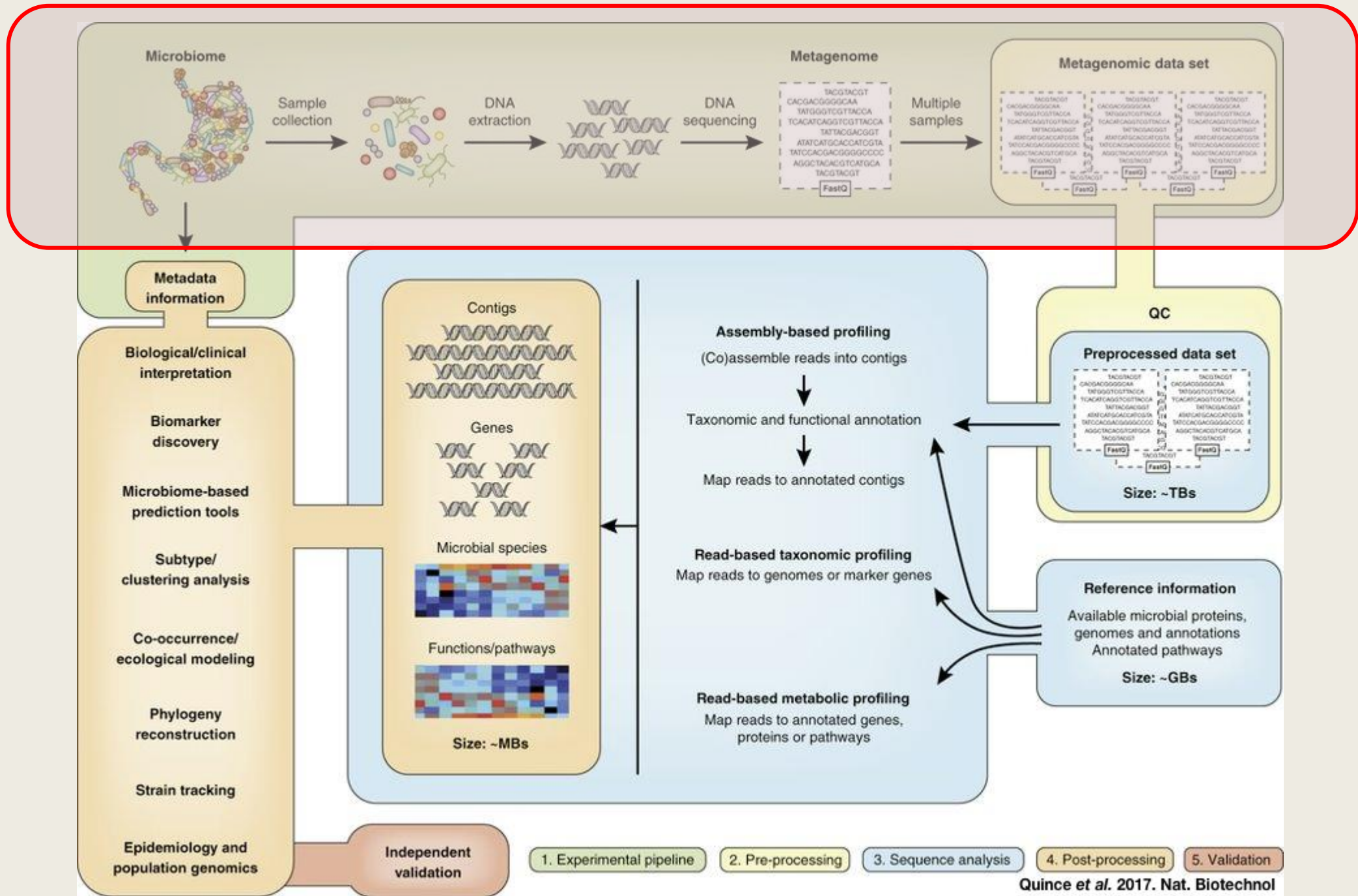


NovaSeq 6000* ††

Output Range	20–120 Gb	40–330 Gb***	65–6000 Gb
Run Time	11–29 hours	11–48 hours	13–44 hours
Reads Per Run	130–400 million	0.4–1.1 billion***	Up to 20 billion
Max Read Length	2 × 150 bp	2 × 150 bp	2 × 250 bp†††
Samples Per Run[§]	8	8–20	12–400
Relative Price Per Sample[§]	Higher Cost	Mid Cost	Low Cost

Bioinformatics

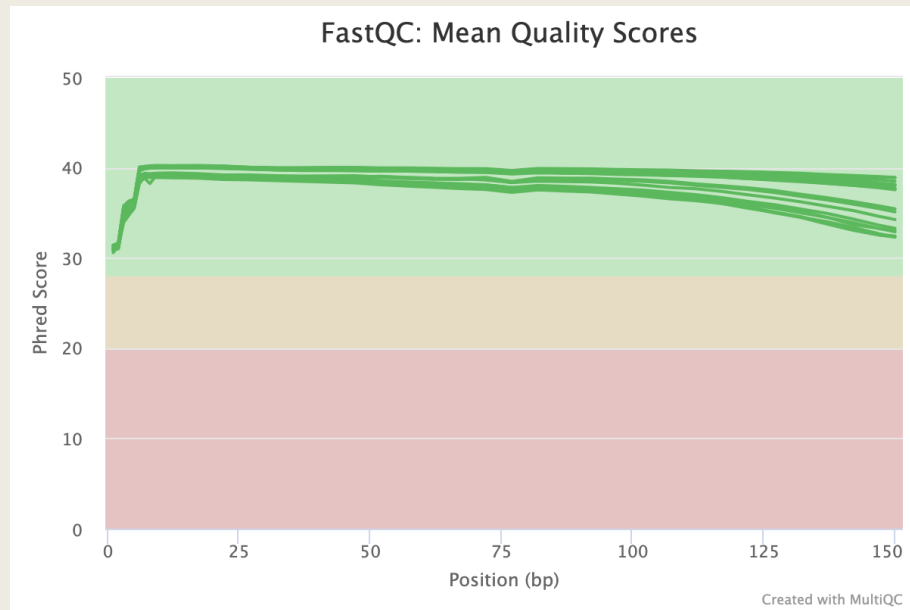


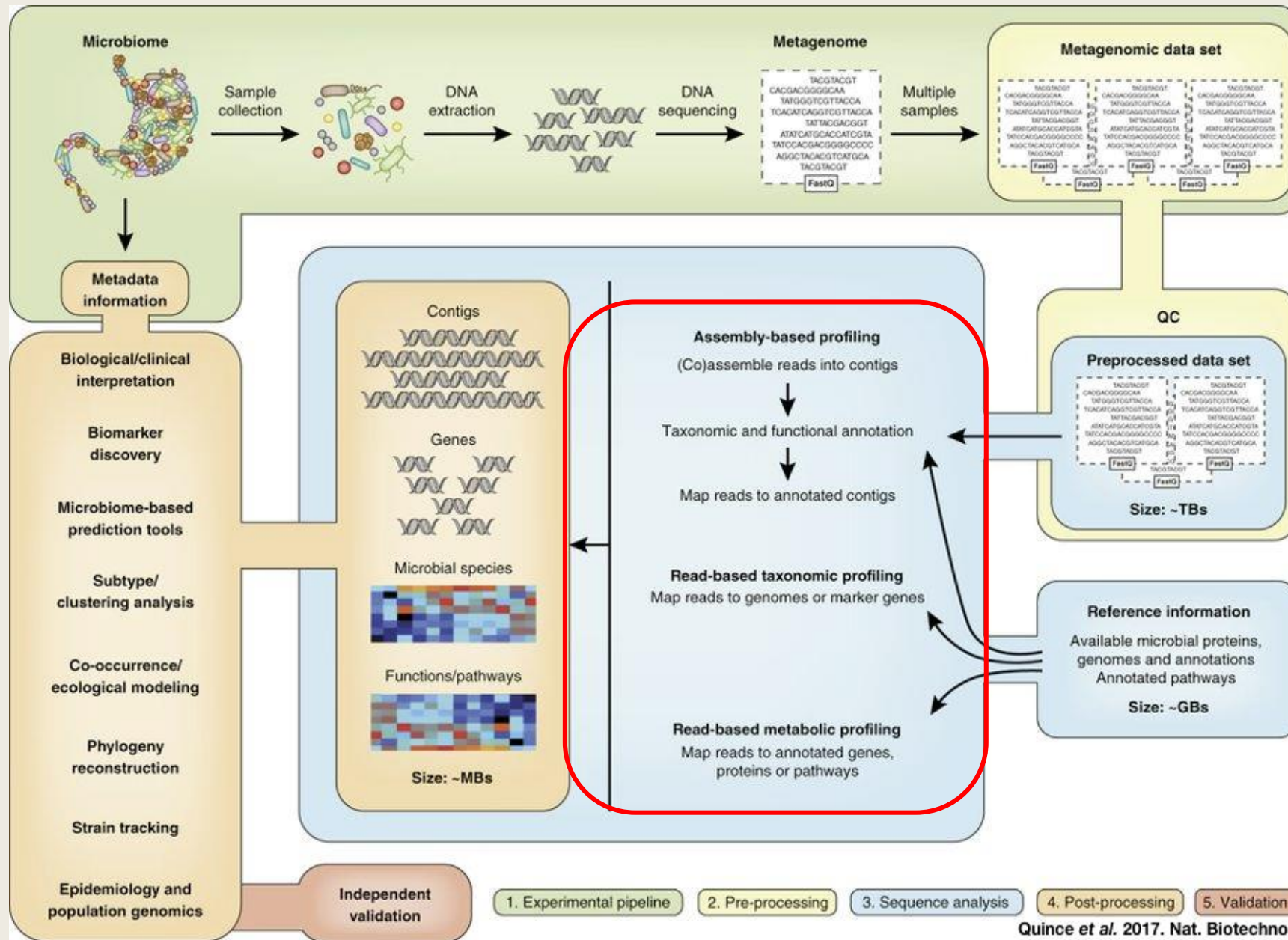


Bioinformatics

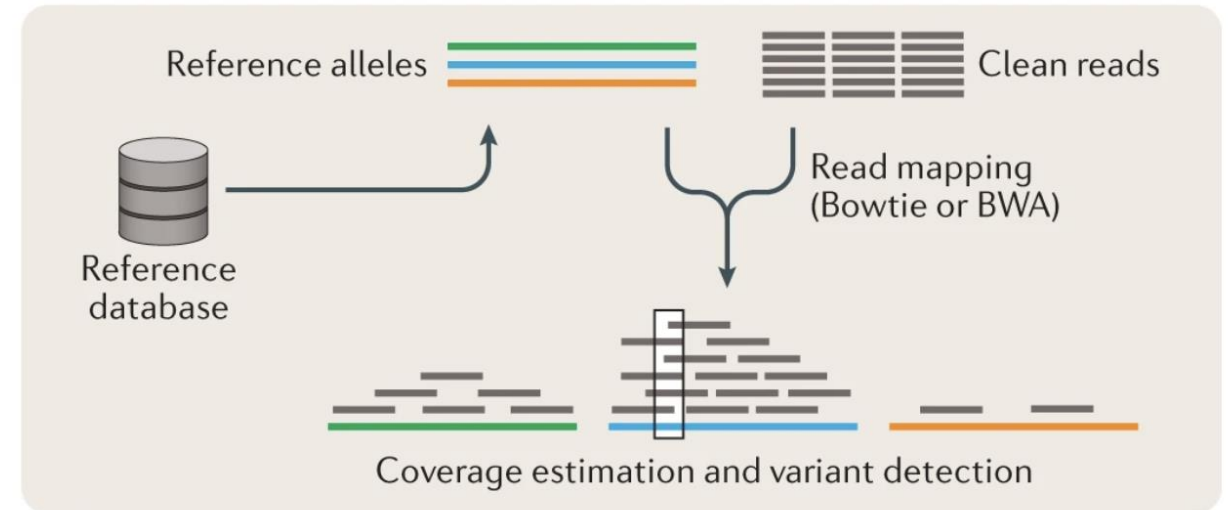
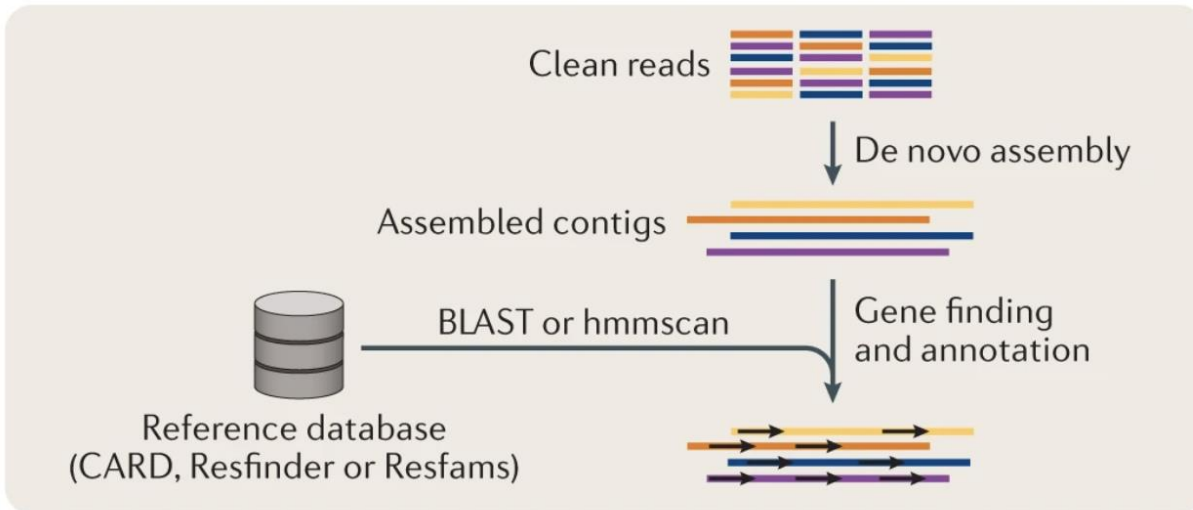
Quality control (QC)

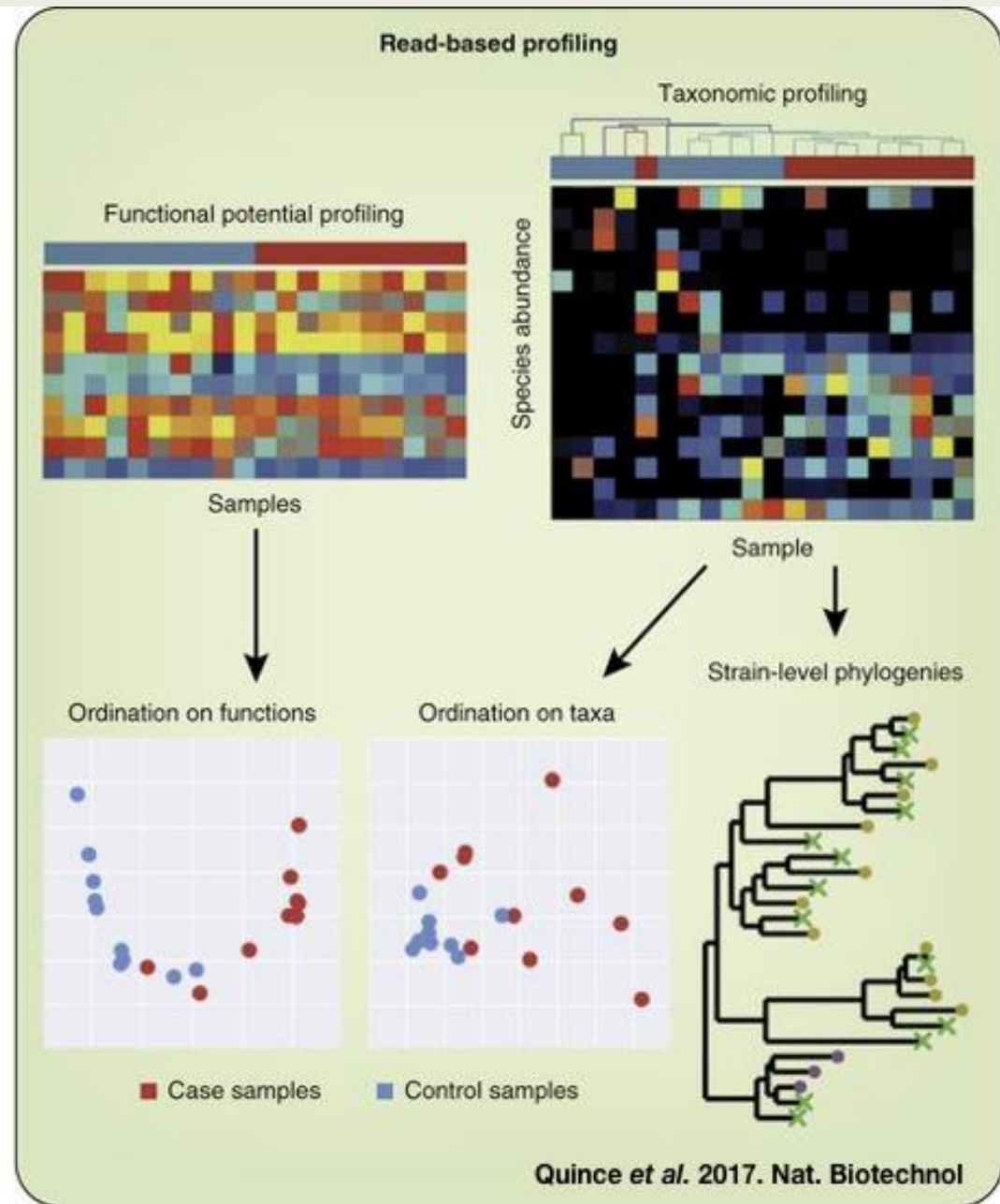
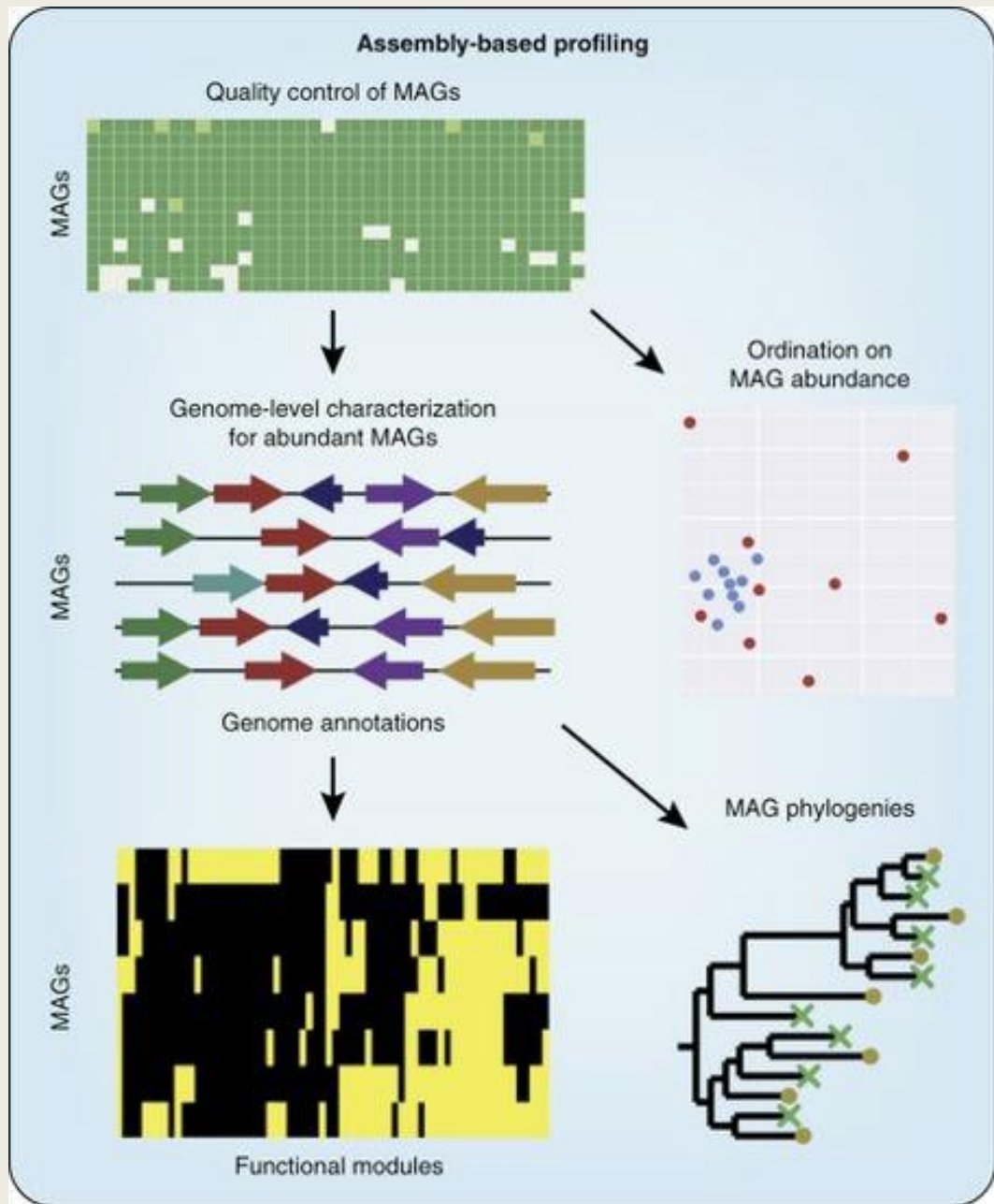
Phred score:
 $Q = -10 \log E$
 $E = 10^{-Q/10}$





b Assembly-based versus read-based approach





Bioinformatic tools

- Read based: **MetaPhlAn 3.0**, **HUMAnN 3.0**, **StrainPhlAn**, **Kraken**, **Centrifuge**
- MetaPhlAn: 17,000 reference genomes (~13,500 bacterial and archaeal, ~3,500 viral, and ~110 eukaryotic)
- Kraken & Centrifuge: kmer based taxonomic annotation against taxonomic databases
- HUMAnN: Functional profiling
- StrainPhlan: Strain-level identification of taxa
- Mapping: **Bowtie**, **BWA**

Assembly based

- Assemblers: Megahit, MetaSpades, IDBA-UD, SOAPdenovo, MetaVelvet
- Computational binning: MetaBat, Groopm, Autometa, MetaWrap
- DAS Tool is an automated method that integrates the results of a flexible number of binning algorithms to calculate an optimized, non-redundant set of bins from a single assembly
- CheckM: Assessing the quality of metagenome bins
- Manual binning: Anvi'o
- Gene prediction: Prodigal
- Gene annotation: HMMER, Blast, Prokka,...

Manual binning of contigs into MAGs versus automated binning

- MAGs = metagenome assembled genomes
- Contigs are binned into bins based on coverage (mapping short reads back to contigs, variation in coverage in different samples), GC content, kmer-usage and sometimes the taxonomic annotation
- Automated binning is fast and relatively easy i.e. **Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life Parks et al. 2017**
- However, genomes are mosaic and contain repetitive elements (mobile genetic elements)
- Results from computational binning should always be checked manually before making any major biological conclusions



A Huge Chunk of a Tardigrade's Genome Comes From Foreign DNA

- Press Release - Source:
University of North Carolina at
Chapel Hill

- Posted November 23, 2015 7:44
PM, www.astrobiology.com