Introduction to Next-generation sequencing

10.11.2021

Melanie Lang & Guillem Salazar

Origins of DNA sequencing

- 1953: Discovery of the structure of DNA


James Watson


Francis Crick


Maurice Wilkins


Rosalind Franklin

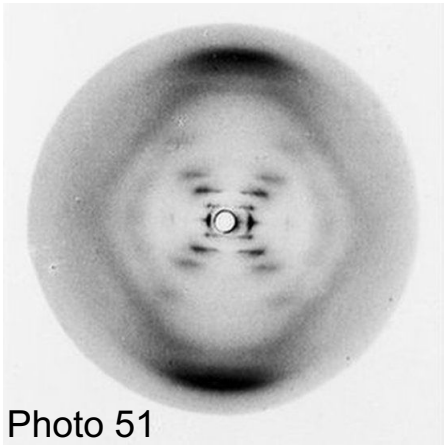
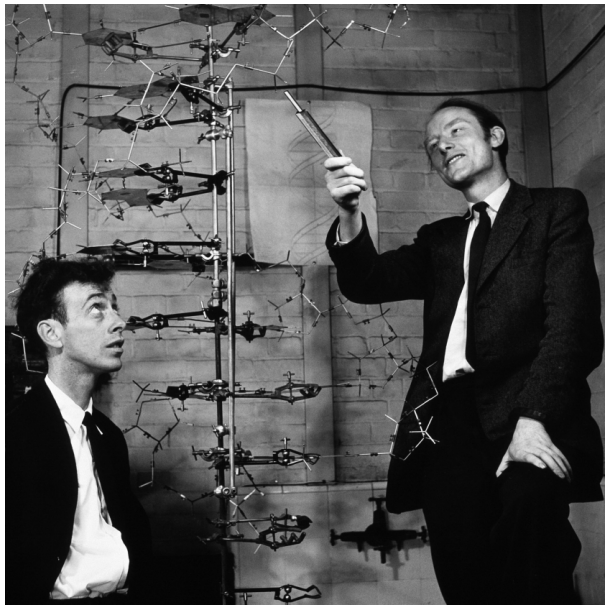
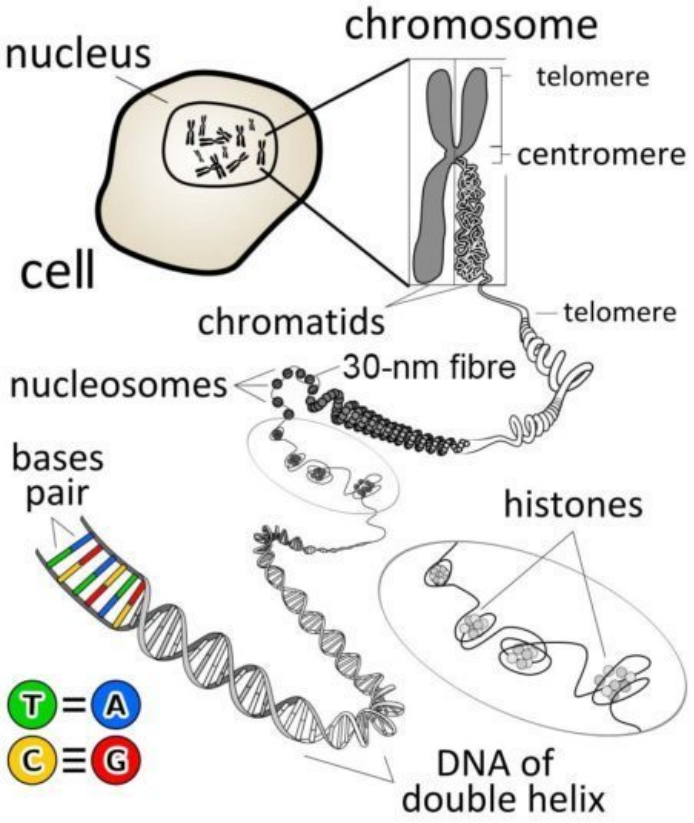


Photo 51

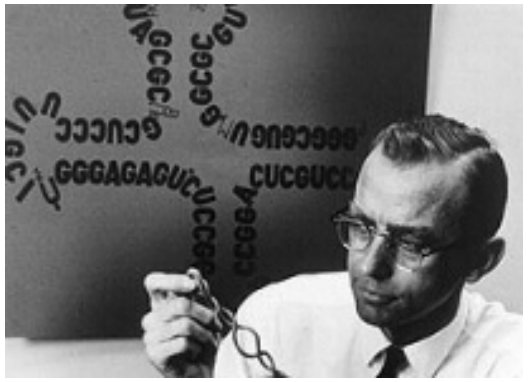




https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/

Origins of DNA sequencing

- 1953: Discovery of the structure of DNA
- 1965: “Sequencing” of the first tRNA
  → use of ribonucleases with cleaving sites at specific nucleotides
  → reconstruction of the original nucleotide sequence by determining the order in which small fragments occurred in the tRNA molecule

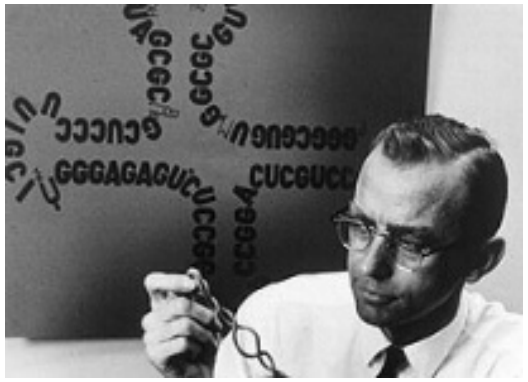


Robert W. Holley

https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/

Origins of DNA sequencing

- 1953: Discovery of the structure of DNA
- 1965: “Sequencing” of the first tRNA
  → use of ribonucleases with cleaving sites at specific nucleotides
  → reconstruction of the original nucleotide sequence by determining the order in which small fragments occurred in the tRNA molecule
- 1972: Sequencing of first complete gene (coat protein of bacteriophage MS2) via RNAse digestion and isolation of oligonucleotides



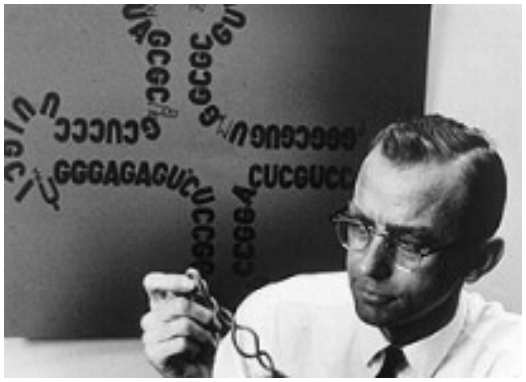
Robert W. Holley

Walter Fiers

https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/

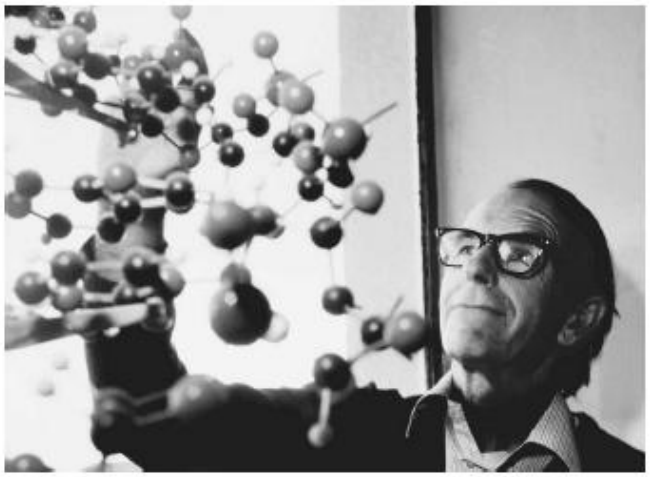
Origins of DNA sequencing

- 1953: Discovery of the structure of DNA
- 1965: "Sequencing" of the first tRNA
  → use of ribonucleases with cleaving sites at specific nucleotides
  → reconstruction of the original nucleotide sequence by determining the order in which small fragments occurred in the tRNA molecule
- 1972: Sequencing of first complete gene (coat protein of bacteriophage MS2) via RNAse digestion and isolation of oligonucleotides
- 1977: Release of "**chain termination method**" utilizing radiolabeled partially digested fragments → **FIRST GENERATION SEQUENCING**

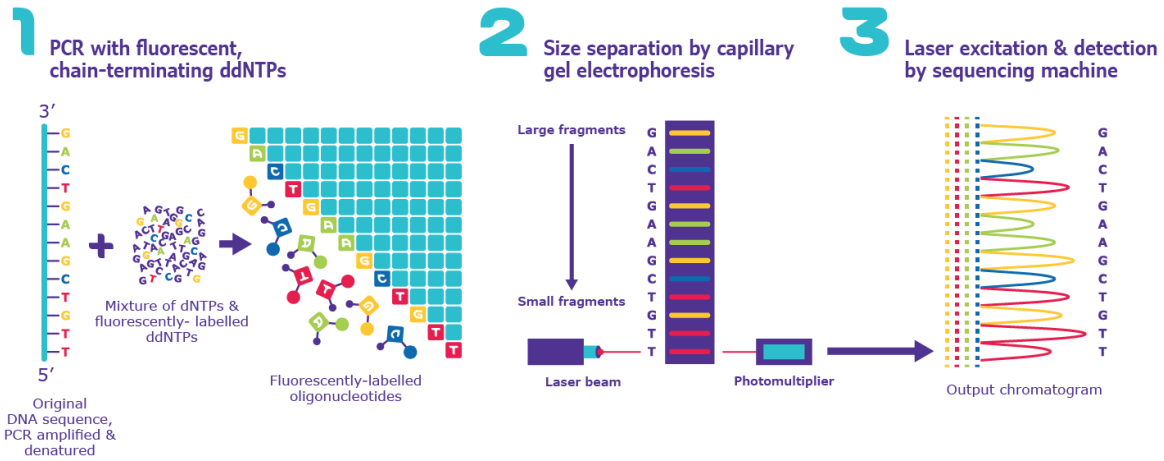


Robert W. Holley

Walter Fiers



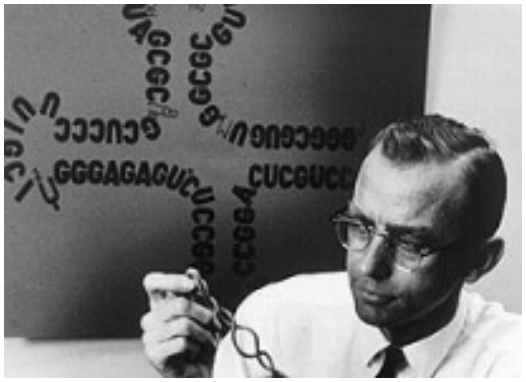
Frederick Sanger



https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/

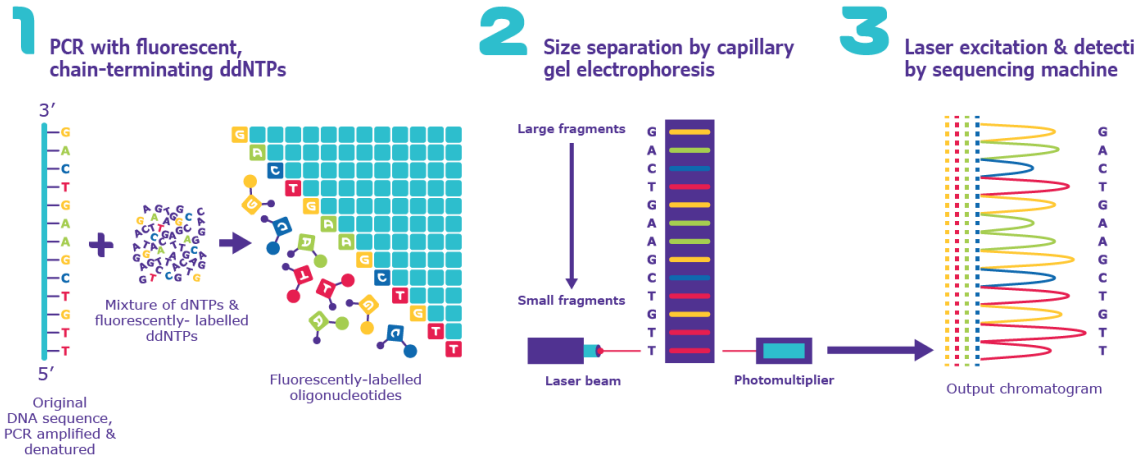
Origins of DNA sequencing

- 1953: Discovery of the structure of DNA
- 1965: “Sequencing” of the first tRNA
  → use of ribonucleases with cleaving sites at specific nucleotides
  → reconstruction of the original nucleotide sequence by determining the order in which small fragments occurred in the tRNA molecule
- 1972: Sequencing of first complete gene (coat protein of bacteriophage MS2) via RNAse digestion and isolation of oligonucleotides
- 1977: Release of “**chain termination method**” utilizing radiolabeled partially digested fragments → **FIRST GENERATION SEQUENCING**

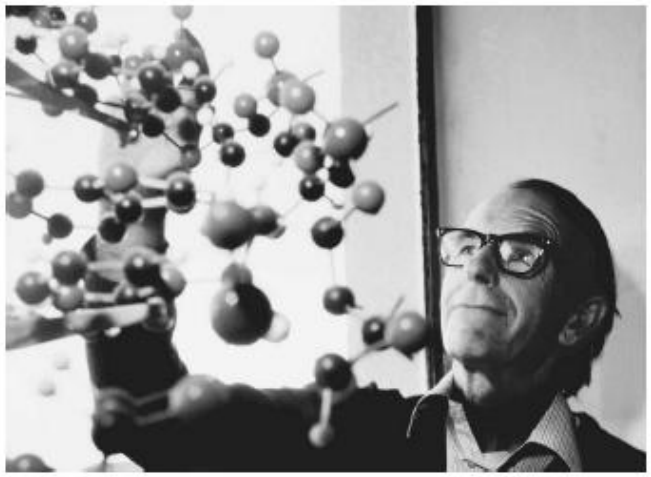


Robert W. Holley

Walter Fiers



→ Main sequencing technology for next 25 years
→ Key innovations mainly in automation of wet-lab and data analysis pipelines

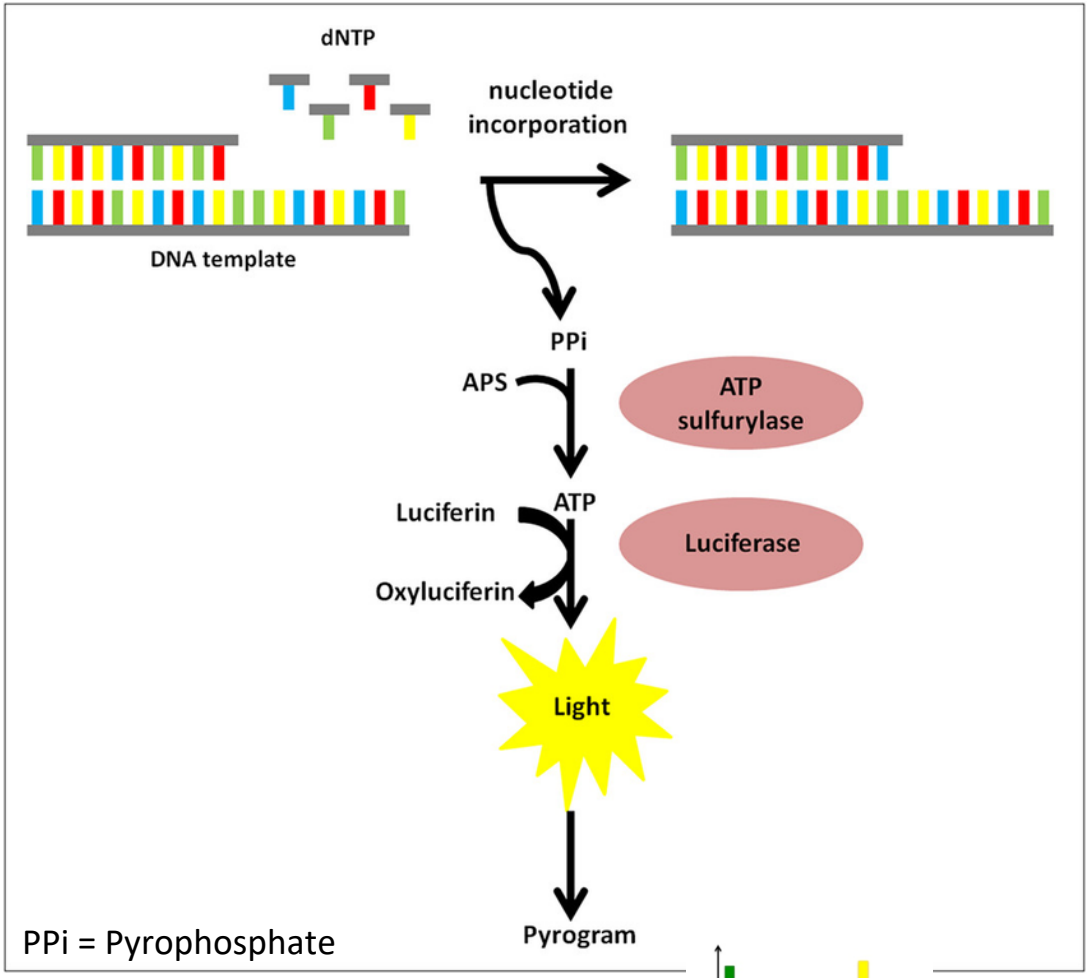


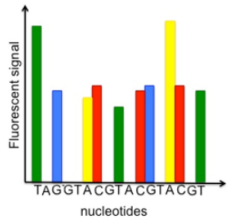
Frederick Sanger

https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/

Origins of DNA sequencing

- 1996: Beginning of **NEXT-GENERATION SEQUENCING**
  → Pyrosequencing

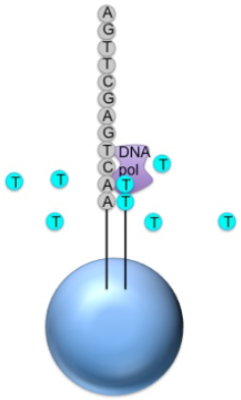


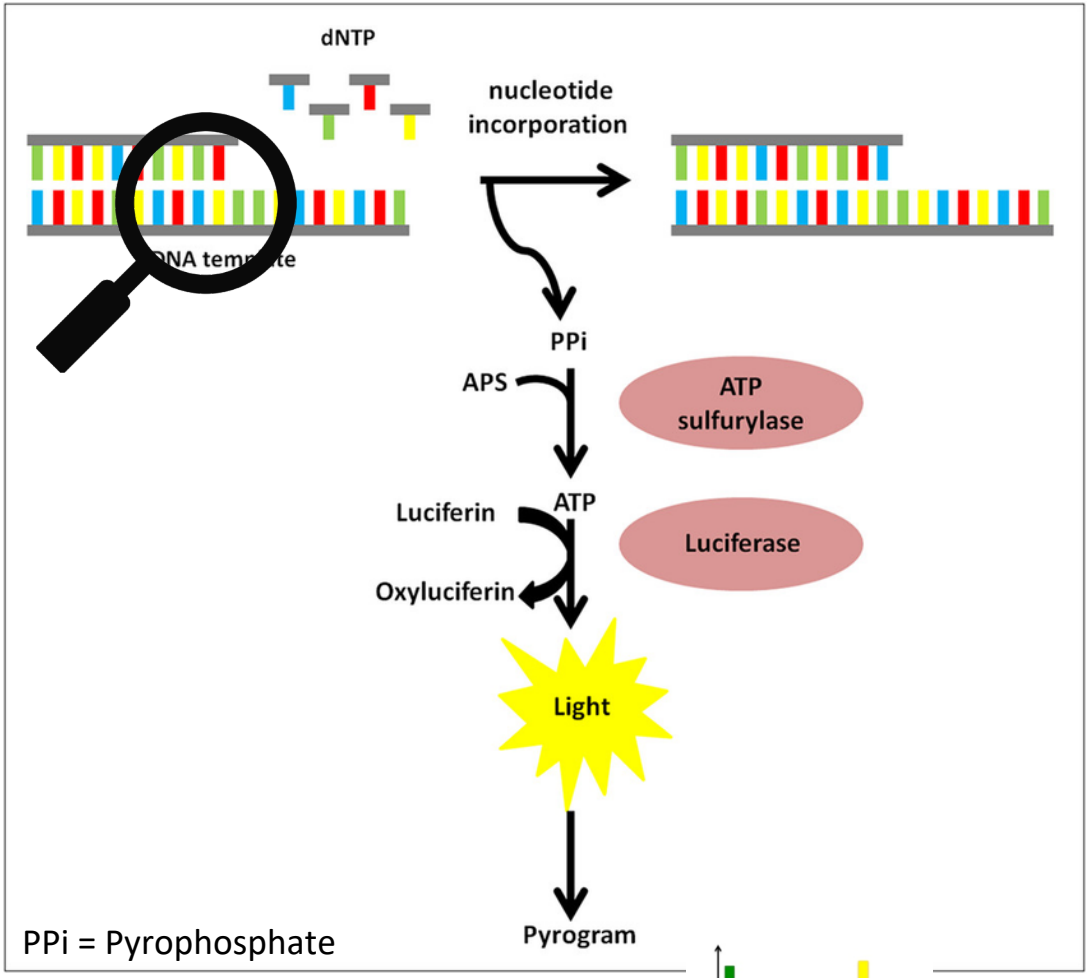


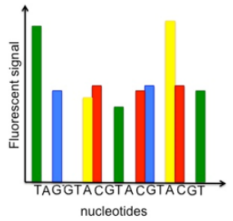
https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/

Origins of DNA sequencing

- 1996: Beginning of **NEXT-GENERATION SEQUENCING**
  → Pyrosequencing



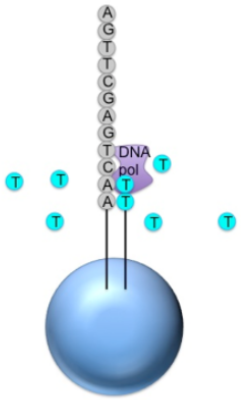


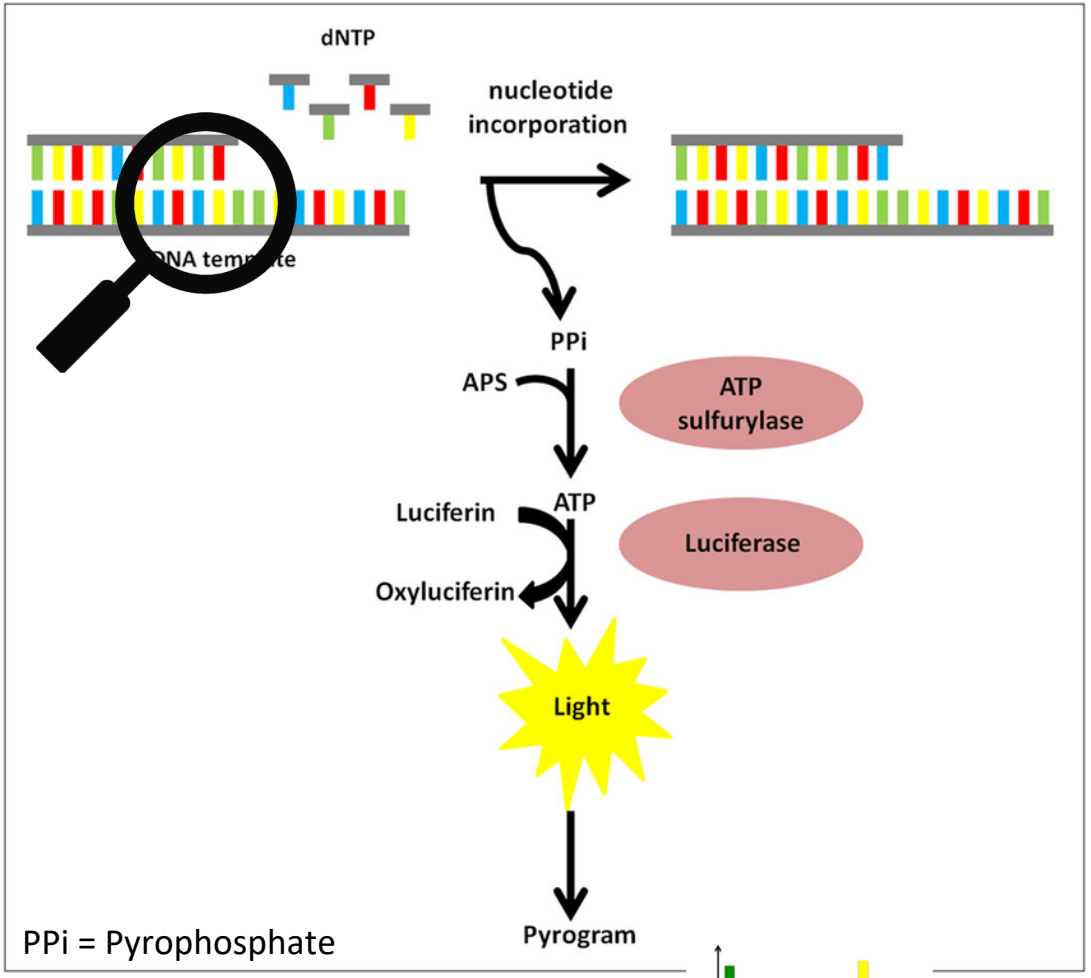


https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/
https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/next-generation-sequencing/454-sequencing/

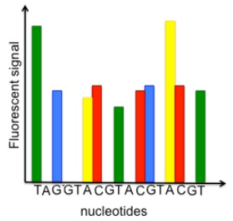
Origins of DNA sequencing

- 1996: Beginning of **NEXT-GENERATION SEQUENCING**
  → Pyrosequencing





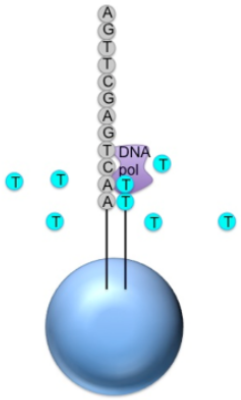
→Each read will have a different length because different numbers of nucleotides will be added during each wash

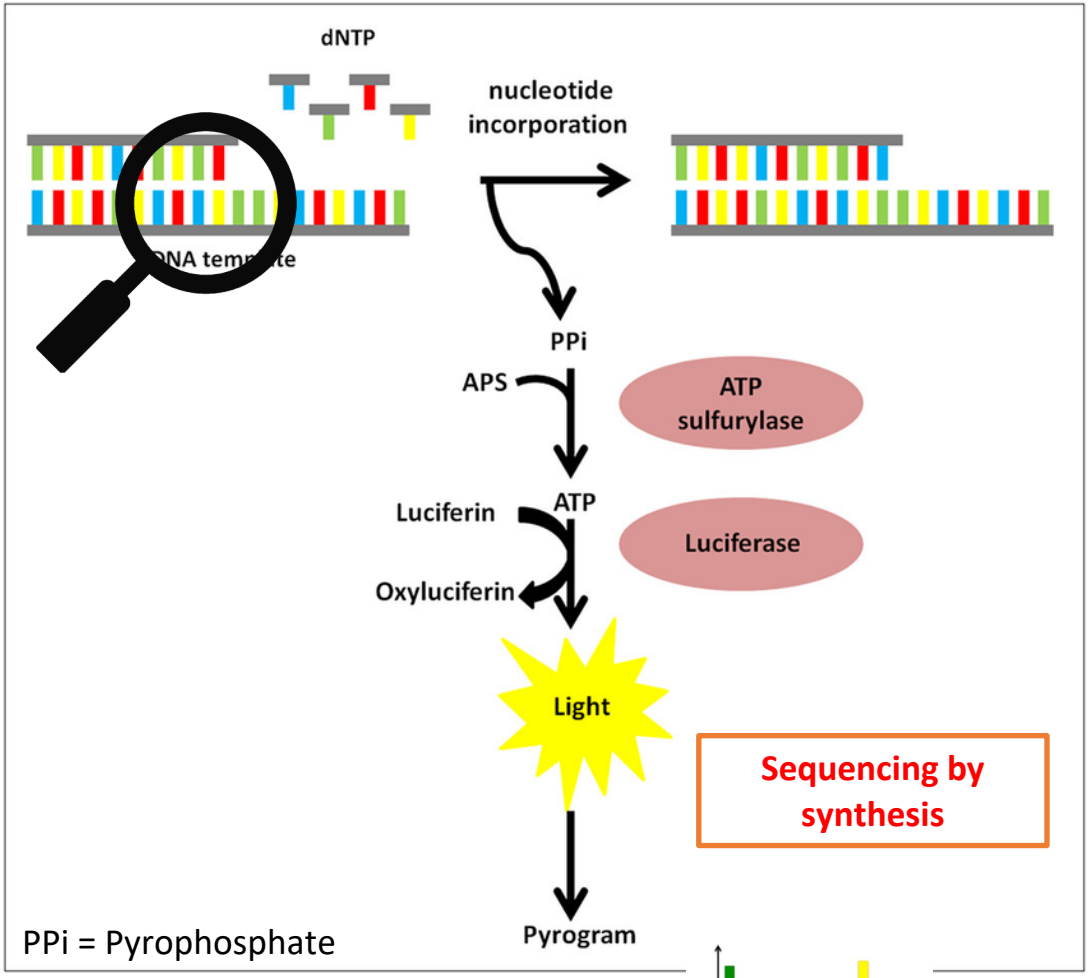


https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/
https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/next-generation-sequencing/454-sequencing/

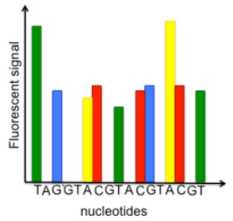
Origins of DNA sequencing

- 1996: Beginning of **NEXT-GENERATION SEQUENCING**
  → Pyrosequencing





→Each read will have a different length because different numbers of nucleotides will be added during each wash



https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/
https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/next-generation-sequencing/454-sequencing/

Origins of DNA sequencing

- 1996: Beginning of **NEXT-GENERATION SEQUENCING**
  → Pyrosequencing
- 2005: Implementation of pyrosequencing in automated system
  → 454 sequencing platform

Roche 454 Sequencing System

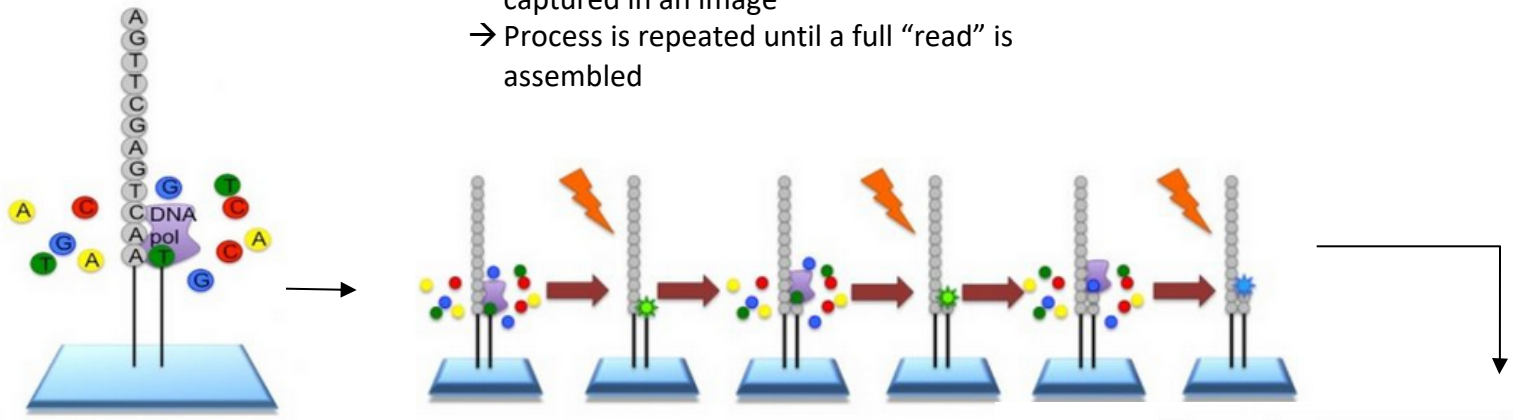
https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/

Origins of DNA sequencing

- 1996: Beginning of **NEXT-GENERATION SEQUENCING**
  → Pyrosequencing
- 2005: Implementation of pyrosequencing in automated system
  → 454 sequencing platform
- 2007: Illumina acquires Solexa
  → Advanced sequencing technology
  → Improved throughput

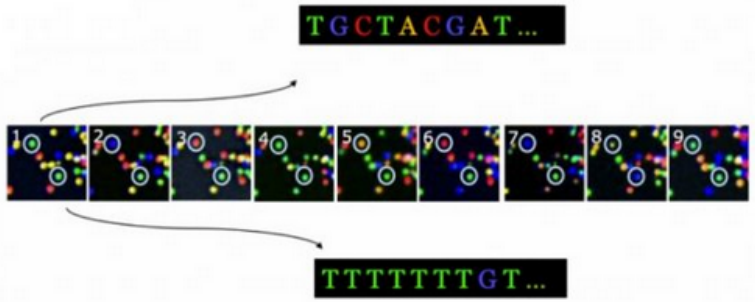
Illumina MiSeq Sequencing platform

→ In each cycle, one dNTP is incorporated into the reaction and it's fluorescent signal captured in an image
→ Process is repeated until a full "read" is assembled



→ Single-stranded short DNA sequences are bound to a glass flow cell
→ Flow cell is flooded with fluorescently labelled dNTPs and polymerase

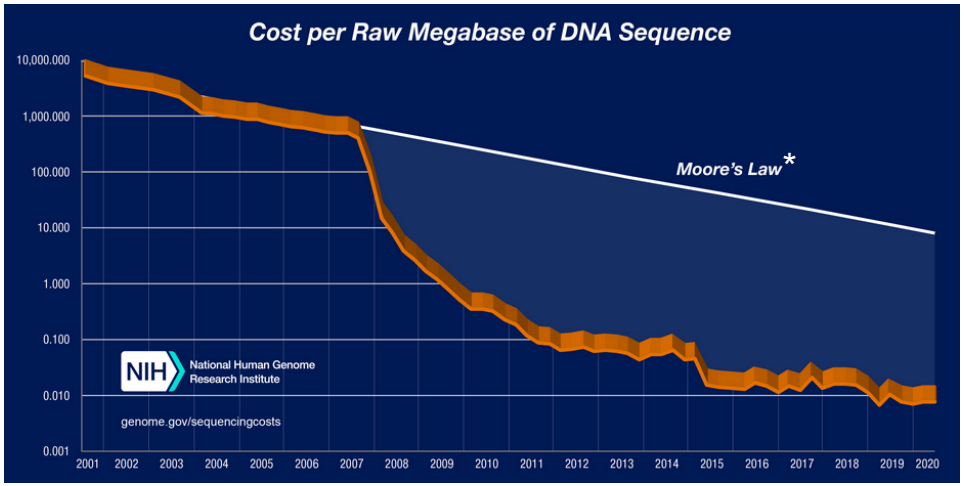
→ Translation of thousands of fluorescent images into a sequence read

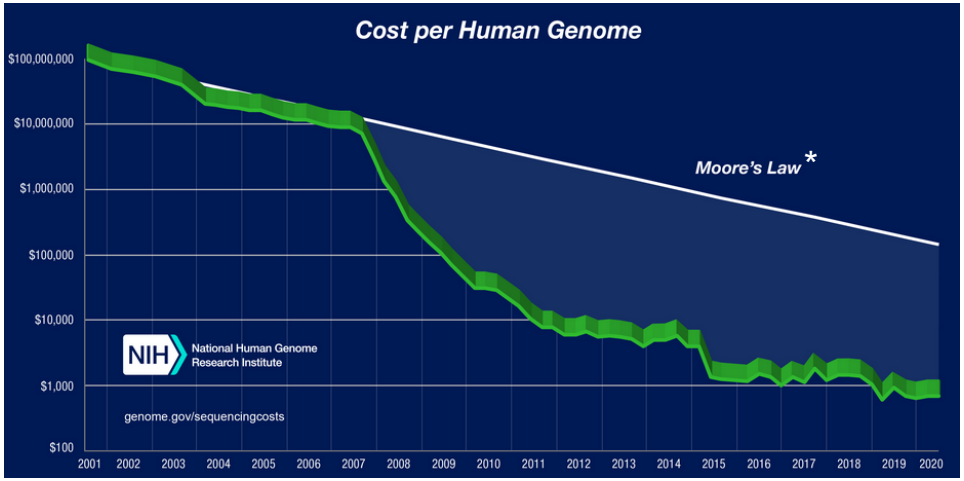


https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/
https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/next-generation-sequencing/454-sequencing/

Origins of DNA sequencing

Improvements in DNA sequencing: Some numbers…





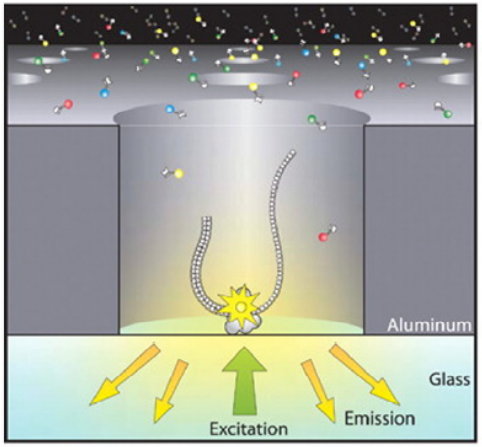
*Moore's law is an observation and projection of a historical trend. Rather than a law of physics, it is an empirical relationship linked to gains from experience in production

https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

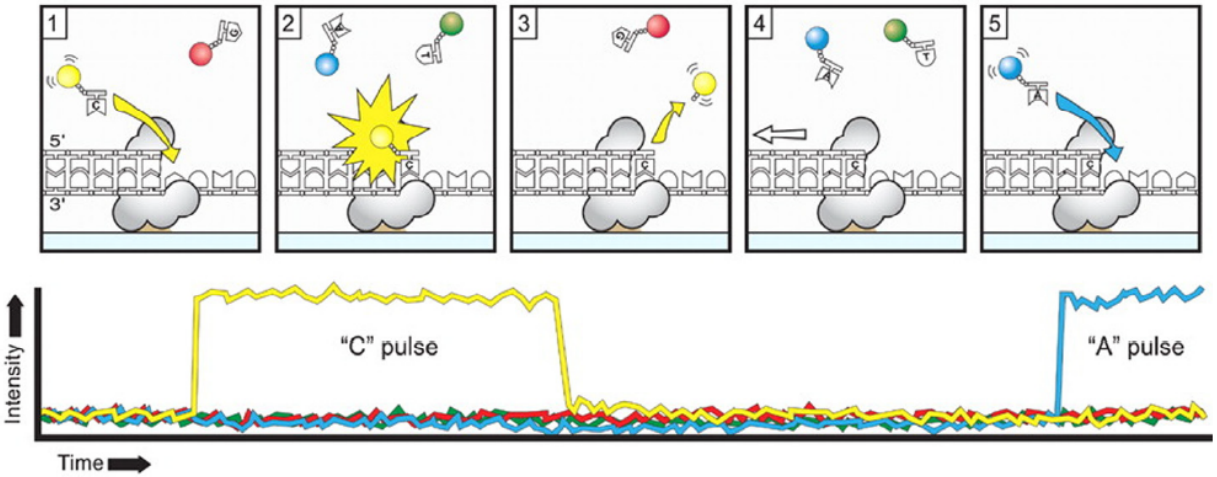
Origins of DNA sequencing

- 2010: Beginning of **THIRD-GENERATION SEQUENCING**
  → PacBio sequencing (Pacific Biosciences, Inc.)

PacBio RSII sequencer



→ polymerase immobilized at the bottom of a "well" (zero-mode waveguide ZMW) in a SMRTcell



→ Incorporation of fluorescent dNTPs produces a base-specific light pulse
→ Replication process in all ZMWs is recorded as a "movie" in real-time

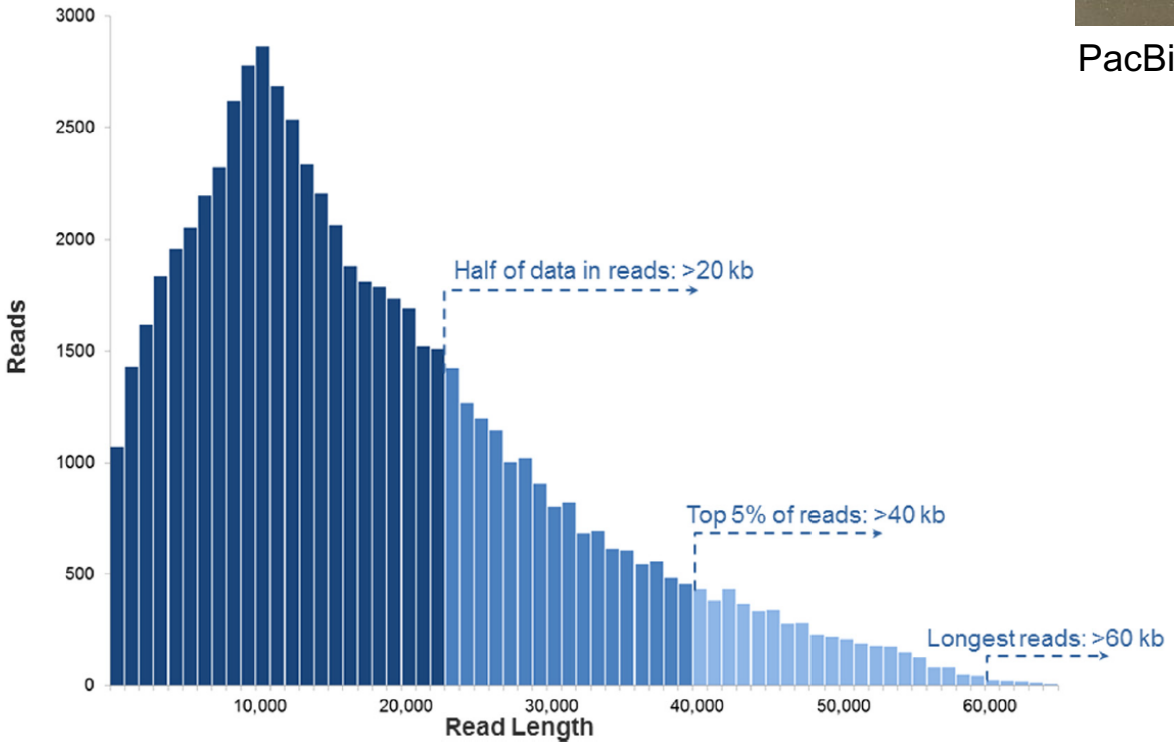
https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/
Rhoads & Au (2015) PacBio sequencing and its applications *Genomics, Proteomics & Bioinformatics* 13(5): 278-289

Origins of DNA sequencing

- 2010: Beginning of **THIRD-GENERATION SEQUENCING**
  → PacBio sequencing (Pacific Biosciences, Inc.)

PacBio RSII sequencer

→ Generation of long-reads!!

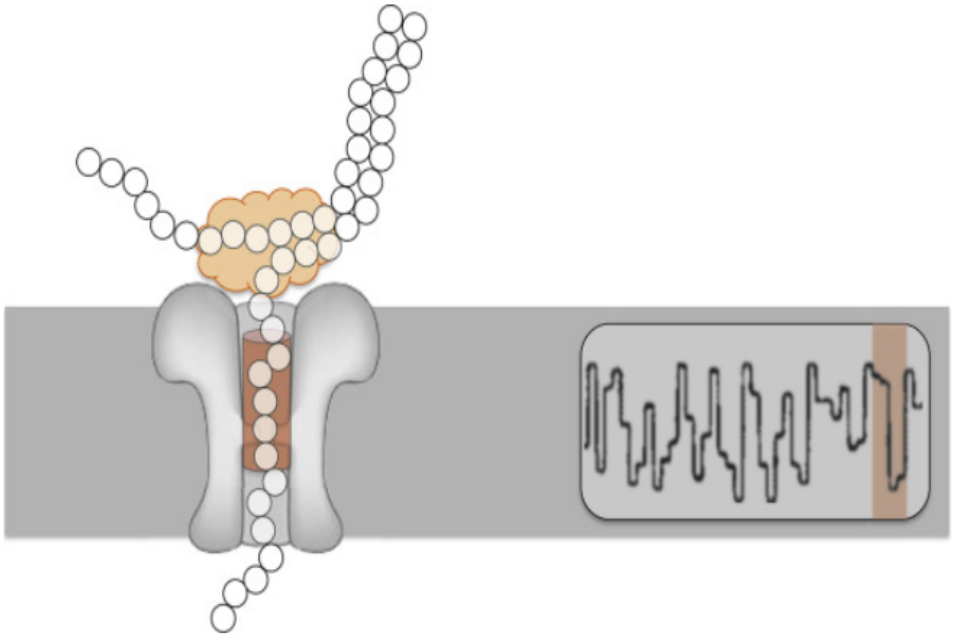


https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/
Rhoads & Au (2015) PacBio sequencing and its applications *Genomics, Proteomics & Bioinformatics* 13(5): 278-289

Origins of DNA sequencing

- 2010: Beginning of **THIRD-GENERATION SEQUENCING**
  - → PacBio sequencing (Pacific Biosciences, Inc.)
  - → Nanopore sequencing (Oxford Nanopore Technologies)

Nanopore MinION



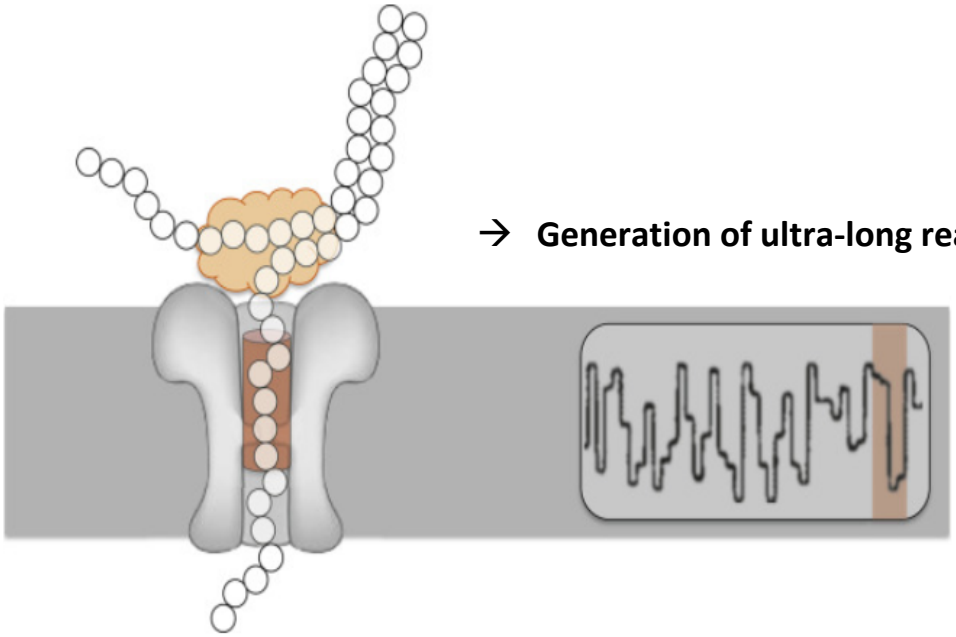
- → single-stranded DNA/RNA molecules pass through protein nanopore
- → Each nucleotide that passes the pore leads to a different change in electrical current across pore
- → Resulting signal is decoded to provide sequence information

https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/
https://www.sciencedirect.com/topics/neuroscience/nanopore-sequencing
https://nanoporetech.com/applications/dna-nanopore-sequencing

Origins of DNA sequencing

- 2010: Beginning of **THIRD-GENERATION SEQUENCING**
  - → PacBio sequencing (Pacific Biosciences, Inc.)
  - → Nanopore sequencing (Oxford Nanopore Technologies)

Nanopore MinION



→ **Generation of ultra-long reads (2Mb)!!**

→ single-stranded DNA/RNA molecules pass through protein nanopore
→ Each nucleotide that passes the pore leads to a different change in electrical current across pore
→ Resulting signal is decoded to provide sequence information

https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/
https://www.sciencedirect.com/topics/neuroscience/nanopore-sequencing
https://nanoporetech.com/applications/dna-nanopore-sequencing

See also: https://www.nature.com/immersive/d42859-020-00099-0/index.html
for milestones of genome sequencing…

https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

Origins of DNA sequencing: Platform comparison

Table 1 Performance comparison of sequencing platforms of various generations

Method	Generation	Read length (bp)	Single pass error rate (%)	No. of reads per run	Time per run	Cost per million bases (USD)	Refs.
Sanger ABI 3730×1	1st	600–1000	0.001	96	0.5–3 h	500	[14,18–21]
Ion Torrent	2nd	200	1	8.2×10^7	2–4 h	0.1	[15,25]
454 (Roche) GS FLX+	2nd	700	1	1×10^6	23 h	8.57	[14,17,27]
Illumina HiSeq 2500 (High Output)	2nd	2×125	0.1	8×10^9 (paired)	7–60 h	0.03	[9,16,26]
Illumina HiSeq 2500 (Rapid Run)	2nd	2×250	0.1	1.2×10^9 (paired)	1–6 days	0.04	[9,16,26]
SOLiD 5500×1	2nd	2×60	5	8×10^8	6 days	0.11	[14,24]
PacBio RS II: P6-C4	3rd	$1.0–1.5 \times 10^4$ on average	13	$3.5–7.5 \times 10^4$	0.5–4 h	0.40–0.80	[5,12,15]
Oxford Nanopore MinION	3rd	$2–5 \times 10^3$ on average	38	$1.1–4.7 \times 10^4$	50 h	6.44–17.90	[22,23]

→ Numbers outdated, main features still remain!

Rhoads & Au (2015) PacBio sequencing and its applications *Genomics, Proteomics & Bioinformatics* 13(5): 278-289

Origins of DNA sequencing: Platform comparison

Table 1 Performance comparison of sequencing platforms of various generations

Method	Generation	Read length (bp)	Single pass error rate (%)	No. of reads per run	Time per run	Cost per million bases (USD)	Refs.
Sanger ABI 3730×1	1st	600–1000	0.001	96	0.5–3 h	500	[14,18–21]
Ion Torrent	2nd	200	1	8.2×10^7	2–4 h	0.1	[15,25]
454 (Roche) GS FLX+	2nd	700	1	1×10^6	23 h	8.57	[14,17,27]
Illumina HiSeq 2500 (High Output)	2nd	2×125	0.1	8×10^9 (paired)	7–60 h	0.03	[9,16,26]
Illumina HiSeq 2500 (Rapid Run)	2nd	2×250	0.1	1.2×10^9 (paired)	1–6 days	0.04	[9,16,26]
SOLiD 5500×1	2nd	2×60	5	8×10^8	6 days	0.11	[14,24]
PacBio RS II: P6-C4	3rd	$1.0–1.5 \times 10^4$ on average	13	$3.5–7.5 \times 10^4$	0.5–4 h	0.40–0.80	[5,12,15]
Oxford Nanopore MinION	3rd	$2–5 \times 10^3$ on average	38	$1.1–4.7 \times 10^4$	50 h	6.44–17.90	[22,23]

Variable (medium) read length, ultra-accurate, low-medium throughput

→ Numbers are outdated, main features still remain!

Rhoads & Au (2015) PacBio sequencing and its applications *Genomics, Proteomics & Bioinformatics* 13(5): 278-289

Origins of DNA sequencing: Platform comparison

Table 1 Performance comparison of sequencing platforms of various generations

Method	Generation	Read length (bp)	Single pass error rate (%)	No. of reads per run	Time per run	Cost per million bases (USD)	Refs.
Sanger ABI 3730×1	1st	600–1000	0.001	96	0.5–3 h	500	[14,18–21]
Ion Torrent	2nd	200	1	8.2×10^7	2–4 h	0.1	[15,25]
454 (Roche) GS FLX+	2nd	700	1	1×10^6	23 h	8.57	[14,17,27]
Illumina HiSeq 2500 (High Output)	2nd	2×125	0.1	8×10^9 (paired)	7–60 h	0.03	[9,16,26]
Illumina HiSeq 2500 (Rapid Run)	2nd	2×250	0.1	1.2×10^9 (paired)	1–6 days	0.04	[9,16,26]
SOLiD 5500×1	2nd	2×60	5	8×10^8	6 days	0.11	[14,24]
PacBio RS II: P6-C4	3rd	$1.0–1.5 \times 10^4$ on average	13	$3.5–7.5 \times 10^4$	0.5–4 h	0.40–0.80	[5,12,15]
Oxford Nanopore MinION	3rd	$2–5 \times 10^3$ on average	38	$1.1–4.7 \times 10^4$	50 h	6.44–17.90	[22,23]

Variable (medium) read length, ultra-accurate, low-medium throughput

Fixed short read length, high accuracy, high/ultra-high throughput

→ Numbers are outdated, main features still remain!

Rhoads & Au (2015) PacBio sequencing and its applications *Genomics, Proteomics & Bioinformatics* 13(5): 278-289

Origins of DNA sequencing: Platform comparison

Table 1 Performance comparison of sequencing platforms of various generations

Method	Generation	Read length (bp)	Single pass error rate (%)	No. of reads per run	Time per run	Cost per million bases (USD)	Refs.
Sanger ABI 3730×1	1st	600–1000	0.001	96	0.5–3 h	500	[14,18–21]
Ion Torrent	2nd	200	1	8.2×10^7	2–4 h	0.1	[15,25]
454 (Roche) GS FLX+	2nd	700	1	1×10^6	23 h	8.57	[14,17,27]
Illumina HiSeq 2500 (High Output)	2nd	2×125	0.1	8×10^9 (paired)	7–60 h	0.03	[9,16,26]
Illumina HiSeq 2500 (Rapid Run)	2nd	2×250	0.1	1.2×10^9 (paired)	1–6 days	0.04	[9,16,26]
SOLiD 5500×1	2nd	2×60	5	8×10^8	6 days	0.11	[14,24]
PacBio RS II: P6-C4	3rd	$1.0–1.5 \times 10^4$ on average	13	$3.5–7.5 \times 10^4$	0.5–4 h	0.40–0.80	[5,12,15]
Oxford Nanopore MinION	3rd	$2–5 \times 10^3$ on average	38	$1.1–4.7 \times 10^4$	50 h	6.44–17.90	[22,23]

Variable (medium) read length, ultra-accurate, low-medium throughput
Fixed short read length, high accuracy, high/ultra-high throughput
Variable long read length, low/medium accuracy, medium/high throughput

→ Numbers are outdated, main features still remain!

Rhoads & Au (2015) PacBio sequencing and its applications *Genomics, Proteomics & Bioinformatics* 13(5): 278-289

Origins of DNA sequencing: Platform comparison

Table 1 Performance comparison of sequencing platforms of various generations

Method	Generation	Read length (bp)	Single pass error rate (%)	No. of reads per run	Time per run	Cost per million bases (USD)	Refs.
Sanger ABI 3730×1	1st	600–1000	0.001	96	0.5–3 h	500	[14,18–21]
Ion Torrent	2nd	200	1	8.2×10^7	2–4 h	0.1	[15,25]
454 (Roche) GS FLX+	2nd	700	1	1×10^6	23 h	8.57	[14,17,27]
Illumina HiSeq 2500 (High Output)	2nd	2×125	0.1	8×10^9 (paired)	7–60 h	0.03	[9,16,26]
Illumina HiSeq 2500 (Rapid Run)	2nd	2×250	0.1	1.2×10^9 (paired)	1–6 days	0.04	[9,16,26]
SOLiD 5500×1	2nd	2×60	5	8×10^8	6 days	0.11	[14,24]
PacBio RS II: P6-C4	3rd	$1.0–1.5 \times 10^4$ on average	13	$3.5–7.5 \times 10^4$	0.5–4 h	0.40–0.80	[5,12,15]
Oxford Nanopore MinION	3rd	$2–5 \times 10^3$ on average	38	$1.1–4.7 \times 10^4$	50 h	6.44–17.90	[22,23]

Variable (medium) read length, ultra-accurate, low-medium throughput
Fixed short read length, high accuracy, high/ultra-high throughput
Variable long read length, low/medium accuracy, medium/high throughput

→ Numbers are outdated, main features still remain!
→ Platforms of all generations still in use today…

Rhoads & Au (2015) PacBio sequencing and its applications *Genomics, Proteomics & Bioinformatics* 13(5): 278-289

Origins of DNA sequencing: Platform comparison

Table 1 Performance comparison of sequencing platforms of various generations

Method	Generation	Read length (bp)	Single pass error rate (%)	No. of reads per run	Time per run	Cost per million bases (USD)	Refs.
Sanger ABI 3730×1	1st	600–1000	0.001	96	0.5–3 h	500	[14,18–21]
Ion Torrent	2nd	200	1	8.2×10^7	2–4 h	0.1	[15,25]
454 (Roche) GS FLX+	2nd	700	1	1×10^6	23 h	8.57	[14,17,27]
Illumina HiSeq 2500 (High Output)	2nd	2×125	0.1	8×10^9 (paired)	7–60 h	0.03	[9,16,26]
Illumina HiSeq 2500 (Rapid Run)	2nd	2×250	0.1	1.2×10^9 (paired)	1–6 days	0.04	[9,16,26]
SOLiD 5500×1	2nd	2×60	5	8×10^8	6 days	0.11	[14,24]
PacBio RS II: P6-C4	3rd	$1.0–1.5 \times 10^4$ on average	13	$3.5–7.5 \times 10^4$	0.5–4 h	0.40–0.80	[5,12,15]
Oxford Nanopore MinION	3rd	$2–5 \times 10^3$ on average	38	$1.1–4.7 \times 10^4$	50 h	6.44–17.90	[22,23]

Variable (medium) read length, ultra-accurate, low-medium throughput
Fixed short read length, high accuracy, high/ultra-high throughput
Variable long read length, low/medium accuracy, medium/high throughput

→ Numbers are outdated, main features still remain!
→ Platforms of all generations still in use today…

Rhoads & Au (2015) PacBio sequencing and its applications *Genomics, Proteomics & Bioinformatics* 13(5): 278-289

Brainstorm: (NGS) sequencing platform applications

Brainstorm: (NGS) sequencing platform applications

Different data types for different applications/questions!

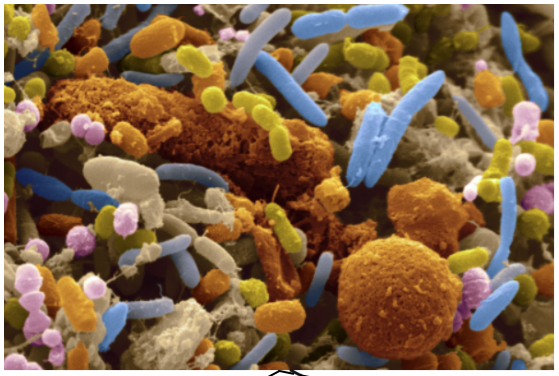
Break…

NGS short-read sequencing: Different data types

NGS short-read sequencing: Different data types

Given a community of bacteria in any given habitat (soil, gut, ocean, …) we want to know:

NGS short-read sequencing: Different data types



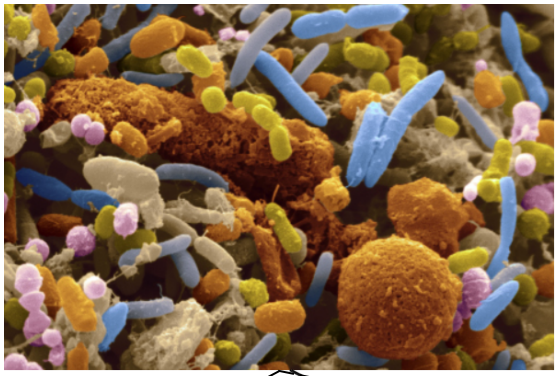
Meta-barcoding (metaB)

Meta-genomics (metaG)

Meta-transcriptomics (metaT)

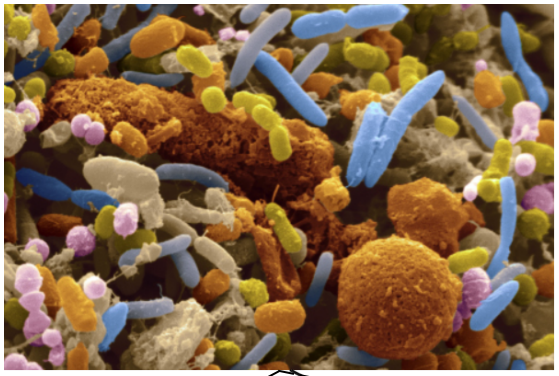
Seq approach
Seq material
Target
Taxonomic precision
Resolution

NGS short-read sequencing: Different data types



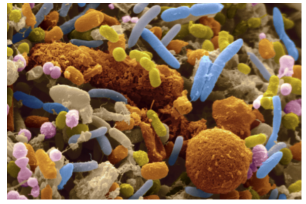
	Meta-barcoding (metaB)	Meta-genomics (metaG)	Meta-transcriptomics (metaT)
Seq approach	Targeted	Non-targeted	Non-targeted
Seq material	Amplicon DNA	Whole genomic DNA	Transcribed RNA
Target	Bacterial genomes	All genomes	All active genomes
Taxonomic precision	Genus/Species	Strain/Genome	Strain/Genome
Resolution	Higher	Lower	Lower

NGS short-read sequencing: Different data types

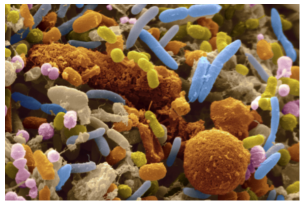


	Meta-barcoding (metaB)	Meta-genomics (metaG)	Meta-transcriptomics (metaT)
Seq approach	Targeted	Non-targeted	Non-targeted
Seq material	Amplicon DNA	Whole genomic DNA	Transcribed RNA
Target	Bacterial genomes	All genomes	All active genomes
Taxonomic precision	Genus/Species	Strain/Genome	Strain/Genome
Resolution	Higher	Lower	Lower

NGS short-read sequencing: MetaB



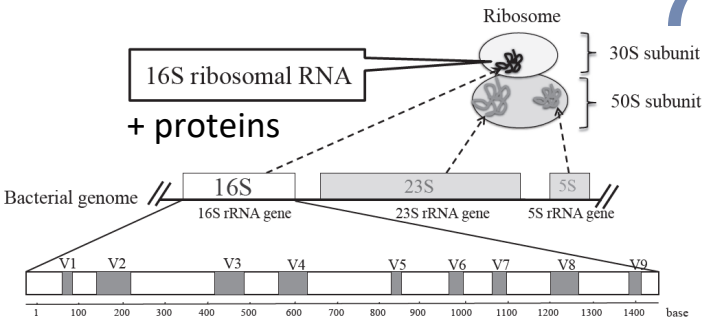
Goal: Identify the members of a bacterial community and its composition
Method: …

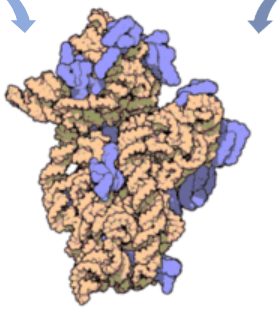


NGS short-read sequencing: MetaB

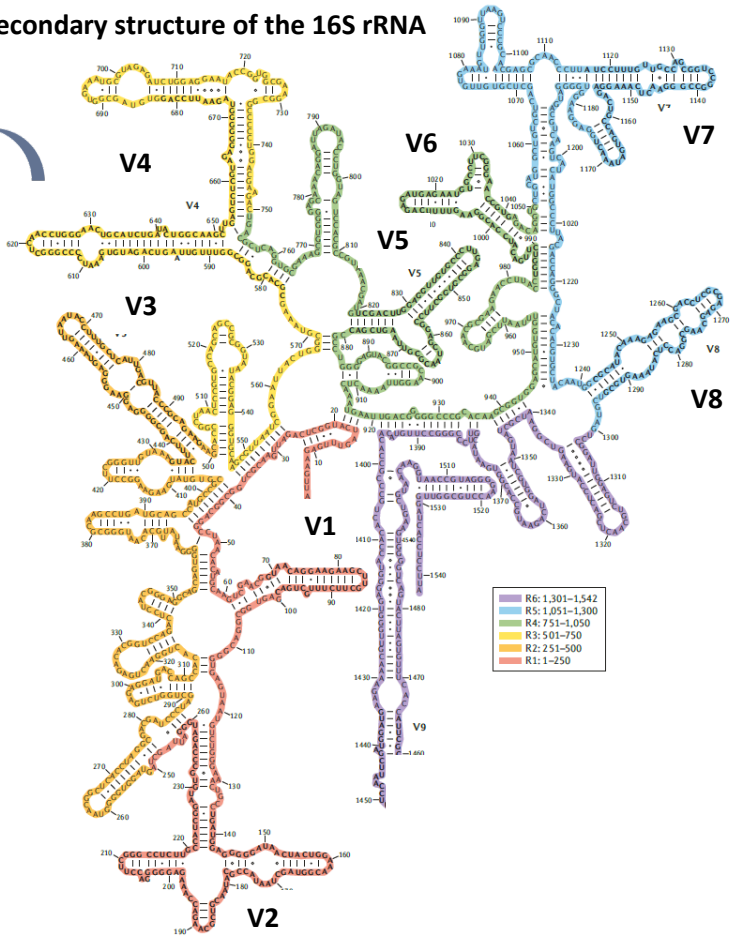
Goal: Identify the members of a bacterial community and its composition
Method: PCR amplification of (part of) bacterial universal marker gene

The 16S rRNA gene (part of prokaryotic ribosome)





Secondary structure of the 16S rRNA

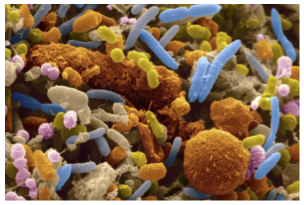


Conserved region
Hypervariable region*

ribosomal proteins
16S rRNA

*Form helical structures, which allow for considerable sequence variation

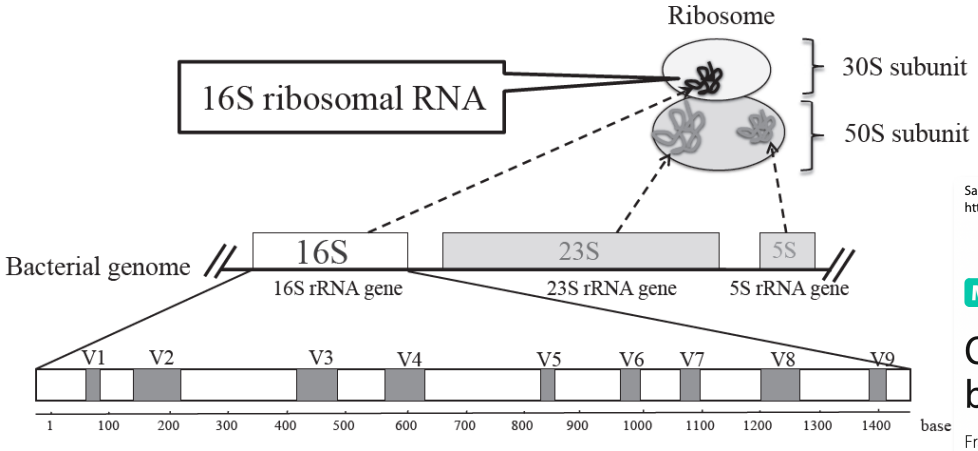
Fukuda et al. (2016) Molecular approaches to studying microbial communities: Targeting the 16S ribosomal RNA gene *Journal of UOEH* 38(3): 223-232
Yarza et al. (2014) Uniting the classification of cultured and uncultured bacteria and arachae using 16S rRNA gene sequences *Nature Reviews Microbiology* 12: 635-645



NGS short-read sequencing: MetaB

Goal: Identify the members of a bacterial community and its composition
Method: PCR amplification of (part of) bacterial universal marker gene

The 16S rRNA gene (part of prokaryotic ribosome)



Conserved region → **ideal as primer binding sites!**
Hypervariable region* → **ideal to resolve sequence variation in bacterial population**

*Form helical structures, which allow for considerable sequence variation

SCIENTIFIC DATA

OPEN **Data Descriptor: The effect of 16S rRNA region choice on bacterial community metabarcoding results**

Sambo *et al. BMC Bioinformatics* (2018) 19:343
https://doi.org/10.1186/s12859-018-2360-6

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Optimizing PCR primers targeting the bacterial 16S ribosomal RNA gene

Francesco Sambo[1], Francesca Finotello[2], Enrico Lavezzo[3], Giacomo Baruzzo[1], Giulia Masi[3], Elektra Peta[3], Marco Falda[3], Stefanc

frontiers in Microbiology

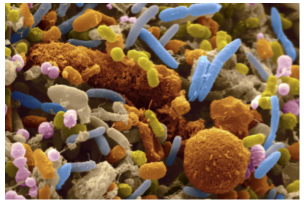
ORIGINAL RESEARCH
published: 04 August 2015
doi: 10.3389/fmicb.2015.00771

Primer and platform effects on 16S rRNA tag sequencing

Julien Tremblay[1,2], Kanwar Singh[1], Alison Fern[1], Edward S. Kirton[1], Shaomei He[1], Tanja Woyke[1], Janey Lee[1], Feng Chen[3], Jeffery L. Dangl[4] and Susannah G. Tringe[1]

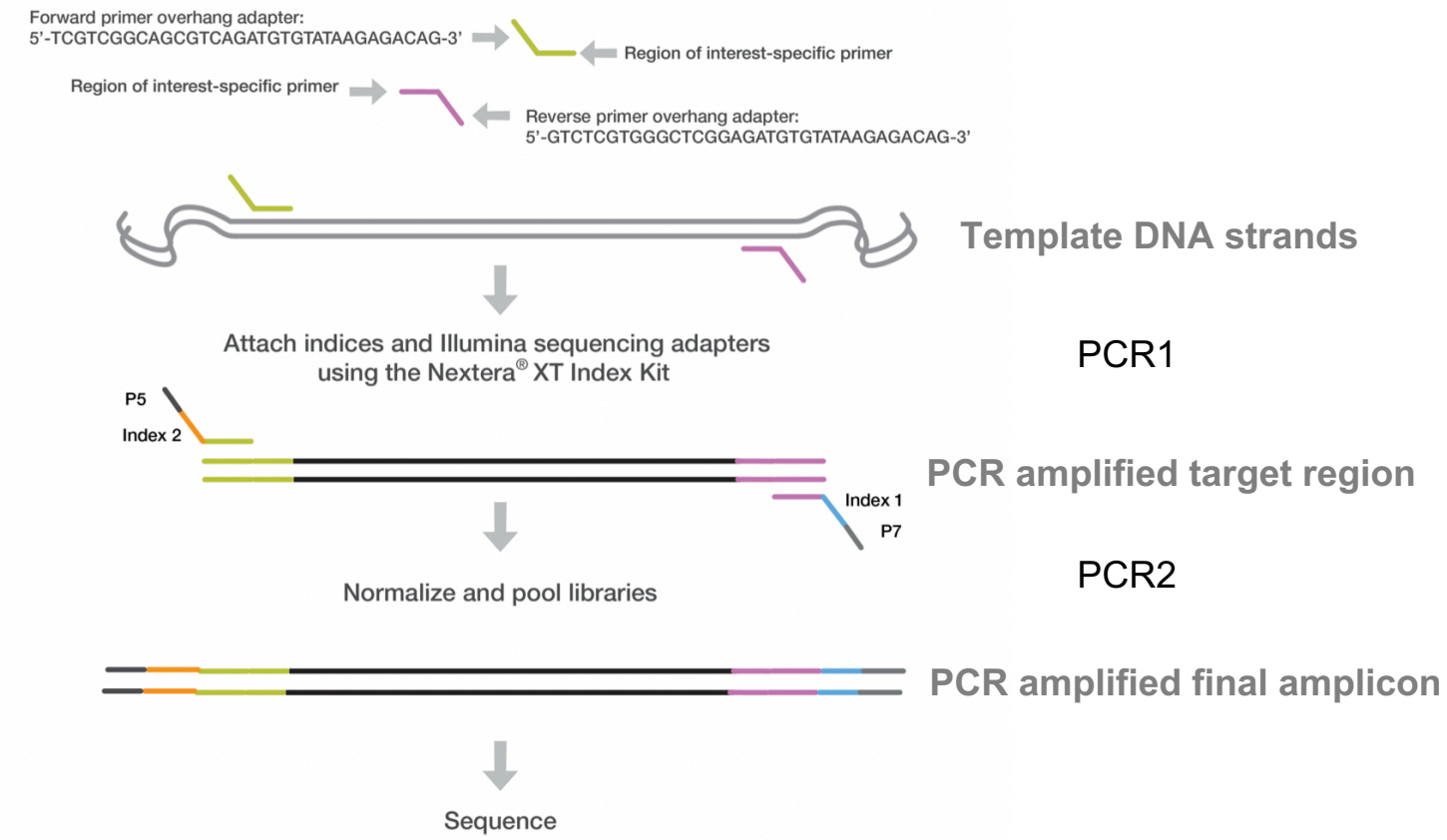
[1] Department of Energy Joint Genome Institute, Walnut Creek, CA, USA, [2] National Research Council Canada, Montreal, QC, Canada, [3] Illumina, Inc., San Francisco, CA, USA, [4] Department of Biology and Howard Hughes Medical Institute, Curriculum in Genetics and Molecular Biology, Department of Microbiology and Immunology, Carolina Center for Genome Sciences, University of North Carolina, Chapel Hill, NC, USA

Fukuda et al. (2016) Molecular approaches to studying microbial communities: Targeting the 16S ribosomal RNA gene *Journal of UOEH* 38(3): 223-232
Yarza et al. (2014) Uniting the classification of cultured and uncultured bacteria and arachae using 16S rRNA gene sequences *Nature Reviews Microbiology* 12: 635-645

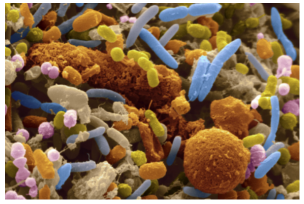


NGS short-read sequencing: MetaB

Goal: Identify the members of a bacterial community and its composition
Method: PCR amplification of (part of) bacterial universal marker gene



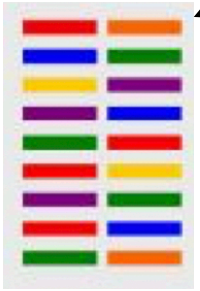
https://support.illumina.com/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf



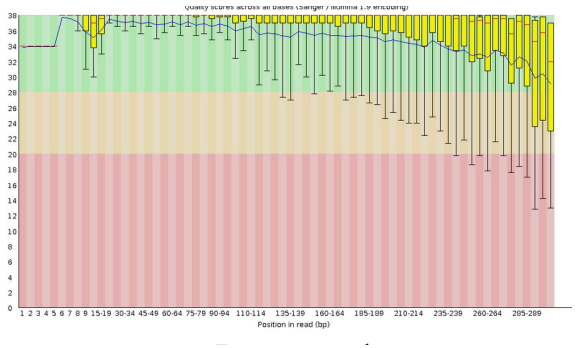
NGS short-read sequencing: MetaB

Data analysis pipeline

Individual reads



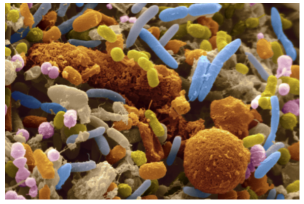
Read QC



Read clustering

Error correction

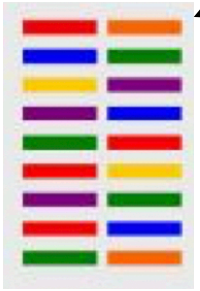
http://web.cbio.uct.ac.za/~gerrit/slides/data_analysis_of_16S_rRNA_gene_amplicons_uct.pdf



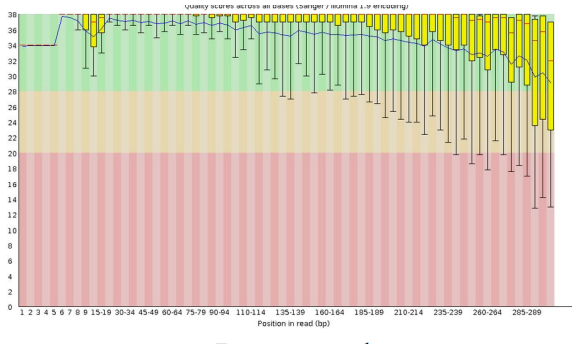
NGS short-read sequencing: MetaB

Data analysis pipeline

Individual reads



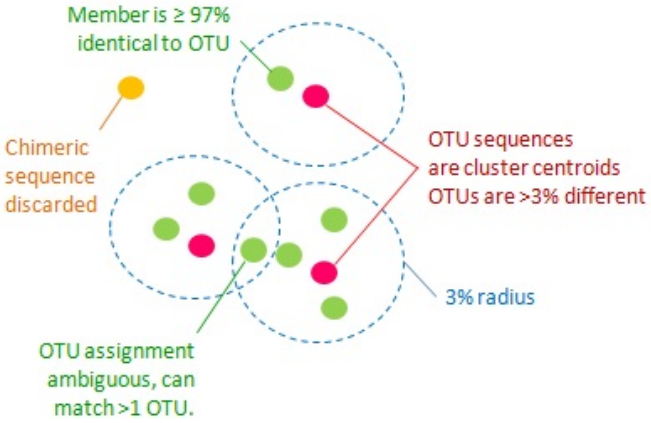
Read QC



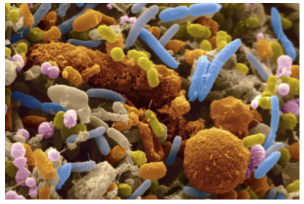
Read clustering

Error correction

Generation of operational taxonomic units (OTUs)



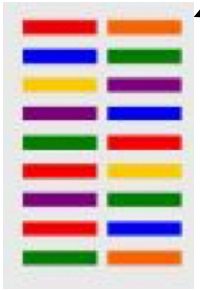
https://drive5.com/usearch/manual/uparseotu_algo.html



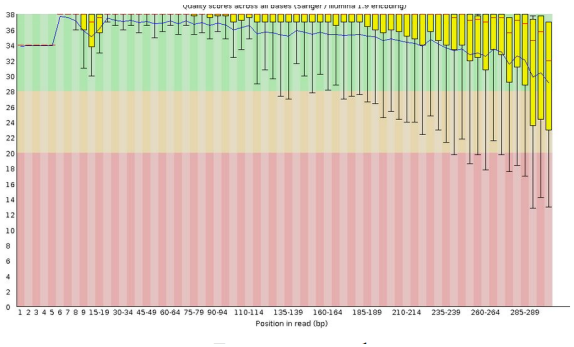
NGS short-read sequencing: MetaB

Data analysis pipeline

Individual reads



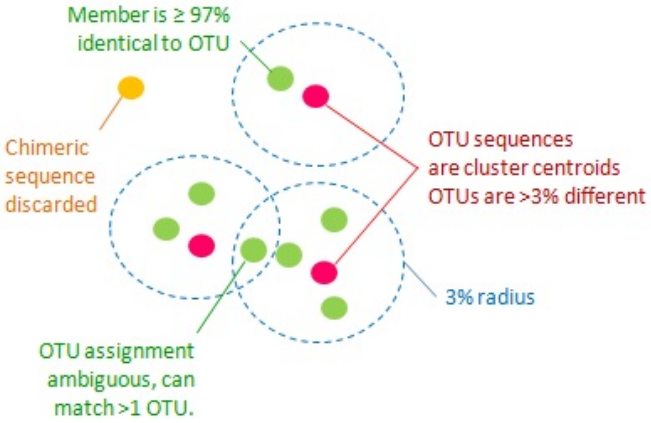
Read QC



Read clustering

Error correction

Generation of operational taxonomic units (OTUs)



https://drive5.com/usearch/manual/uparseotu_algo.html

Popular 16S rRNA gene analysis tools

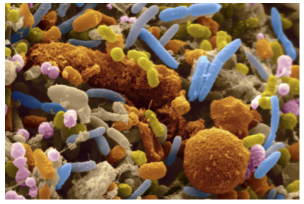
Mothur, University of Michigan
https://www.mothur.org/

USEARCH
Ultra-fast sequence analysis

Robert Edgar
http://www.drive5.com/usearch/

Quantitative Insights Into Microbial Ecology
qiime

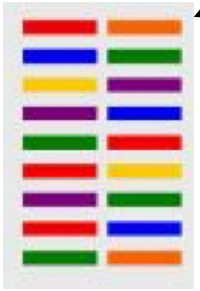
University of Colorado
http://qiime.org/



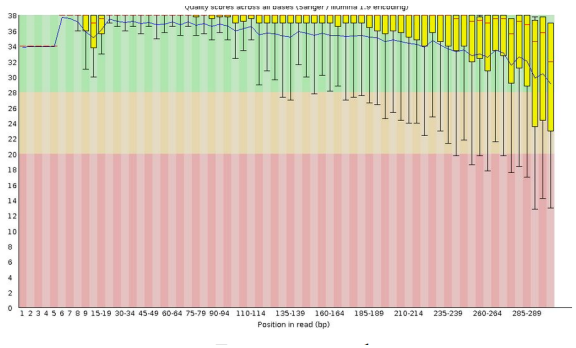
NGS short-read sequencing: MetaB

Data analysis pipeline

Individual reads



Read QC

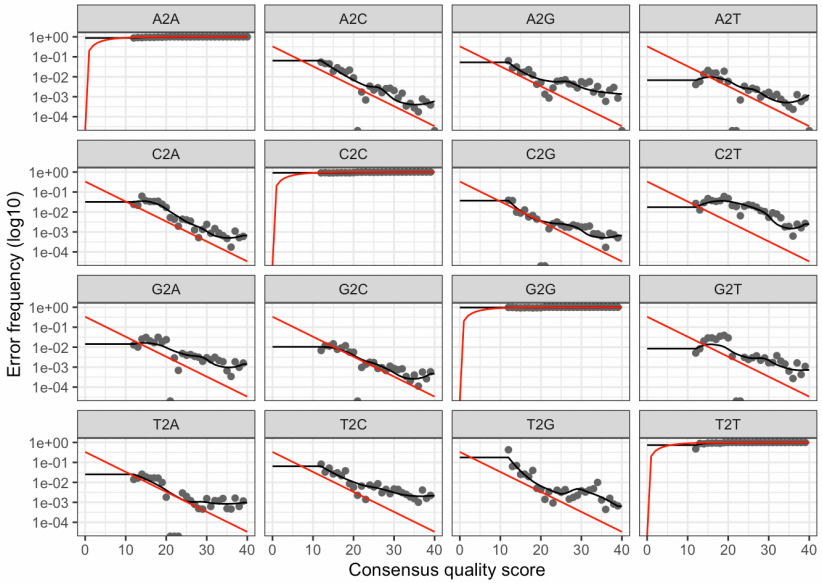


Read clustering

Error correction

DADA2

Amplicon Sequencing. **Exactly.** *Version 1.18*

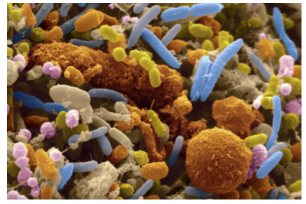


Learning the error model from the sequencing data:

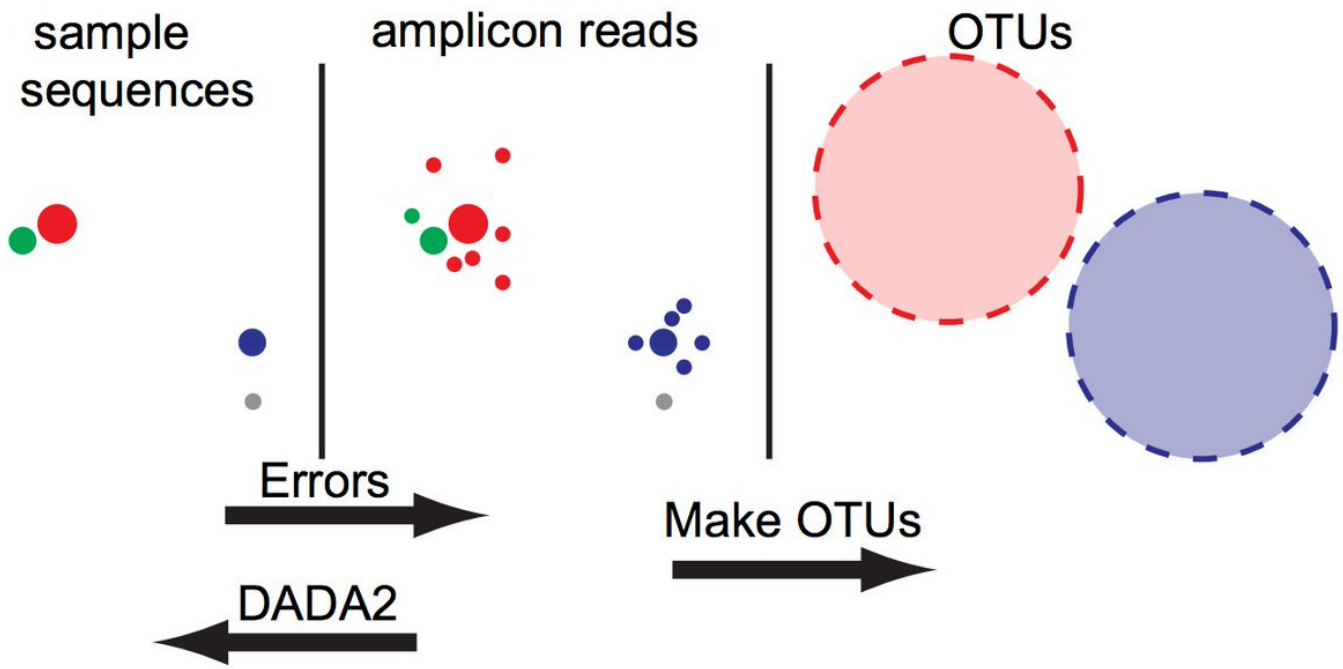
- Infer error rates for all possible nucleotide transitions per consensus quality score
- black line represents the estimated error rates after convergence of the machine-learning algorithm
- Inferred error model is used to correct individual reads (separately for forward and reverse read)
- Merging of forward and reverse reads
- Only identical consensus reads are grouped into **Amplicon Sequence variants (ASVs)**

https://benjjneb.github.io/dada2/tutorial.html

NGS short-read sequencing: MetaB

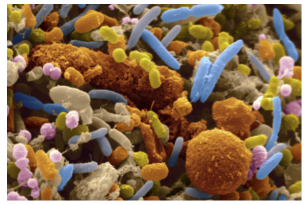


OTUs vs ASVs



Callahan, et al. Nature Methods, 2016.

NGS short-read sequencing: MetaB



Taxonomic annotation

Popular 16S rRNA gene databases

Taxonomy:
Kingdom
Phylum
Class
Order
Family
Genus
Species

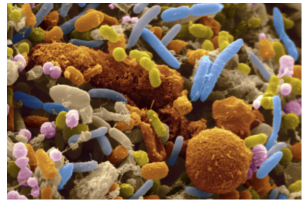
University of Michigan
https://rdp.cme.msu.edu/

LBNL Berkeley,
now second genomes
http://greengenes.lbl.gov

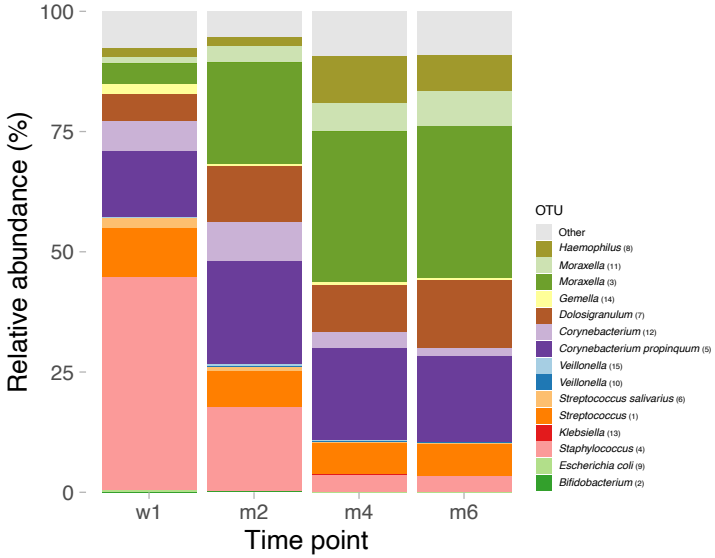
MPI Bremen
https://www.arb-silva.de/

https://drive5.com/usearch/manual/uparseotu_algo.html

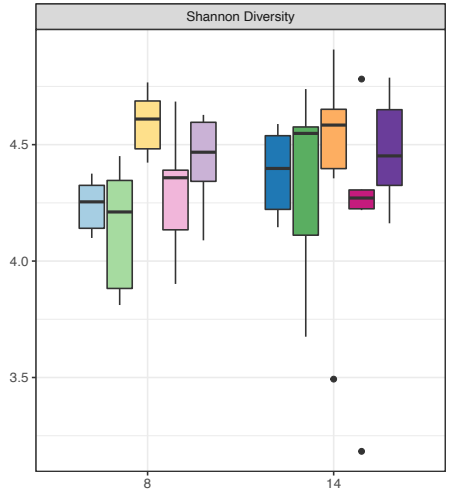
NGS short-read sequencing: MetaB



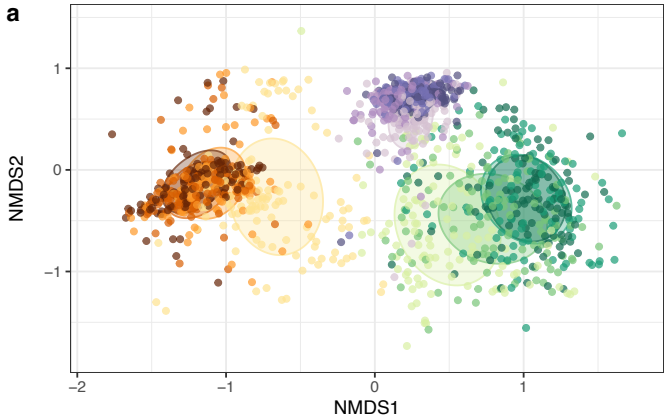
Community structure analysis



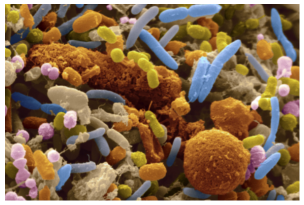
Relative abundances
note: suffers from
problem of compositionality



Alpha-diversity
ecologically defined as
a function of the number of
species (richness) and the
evenness of species distribution

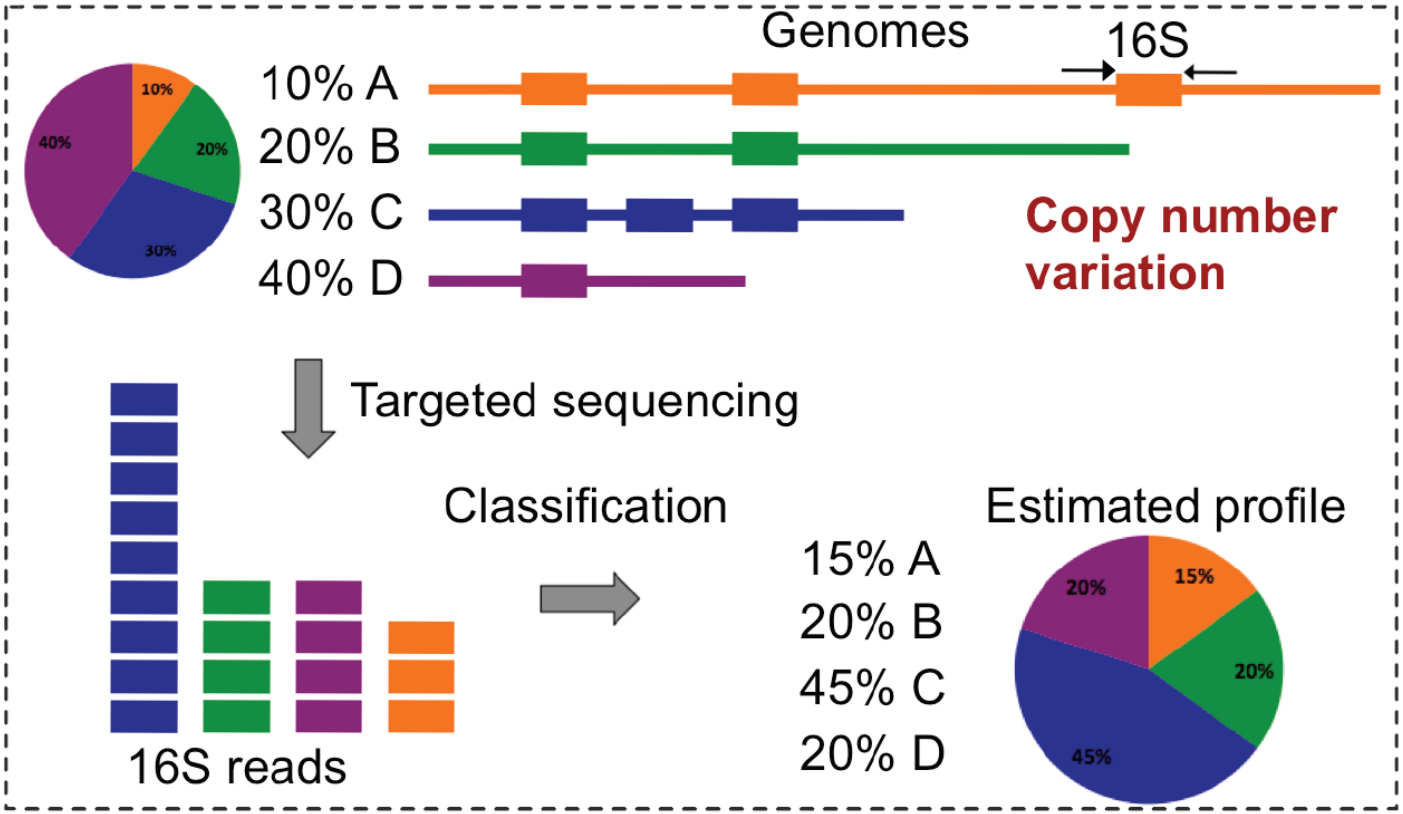


Beta-diversity
analysis of between-
community dissimilarity



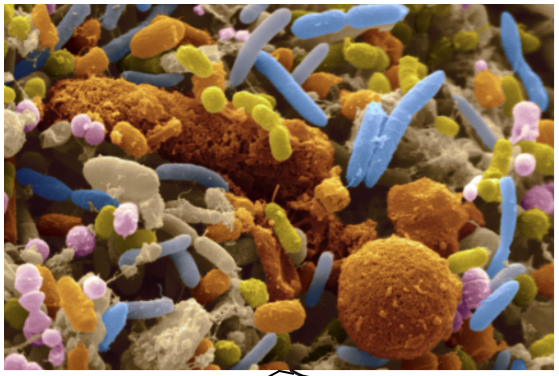
NGS short-read sequencing: MetaB

Caveat of multiple 16S rRNA gene copies



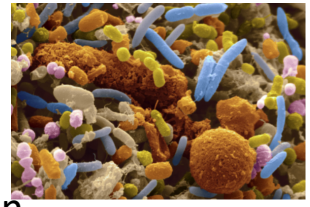
(a) Targeted sequencing of 16S rRNA

NGS short-read sequencing: Different data types

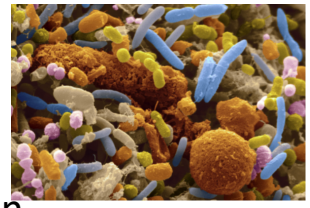


	Meta-barcoding (metaB)	Meta-genomics (metaG)	Meta-transcriptomics (metaT)
Seq approach	Targeted	Non-targeted	Non-targeted
Seq material	Amplicon DNA	Whole genomic DNA	Transcribed RNA
Target	Bacterial genomes	All genomes	All active genomes
Taxonomic precision	Genus/Species	Strain/Genome	Strain/Genome
Resolution	Higher	Lower	Lower

NGS short-read sequencing: MetaG

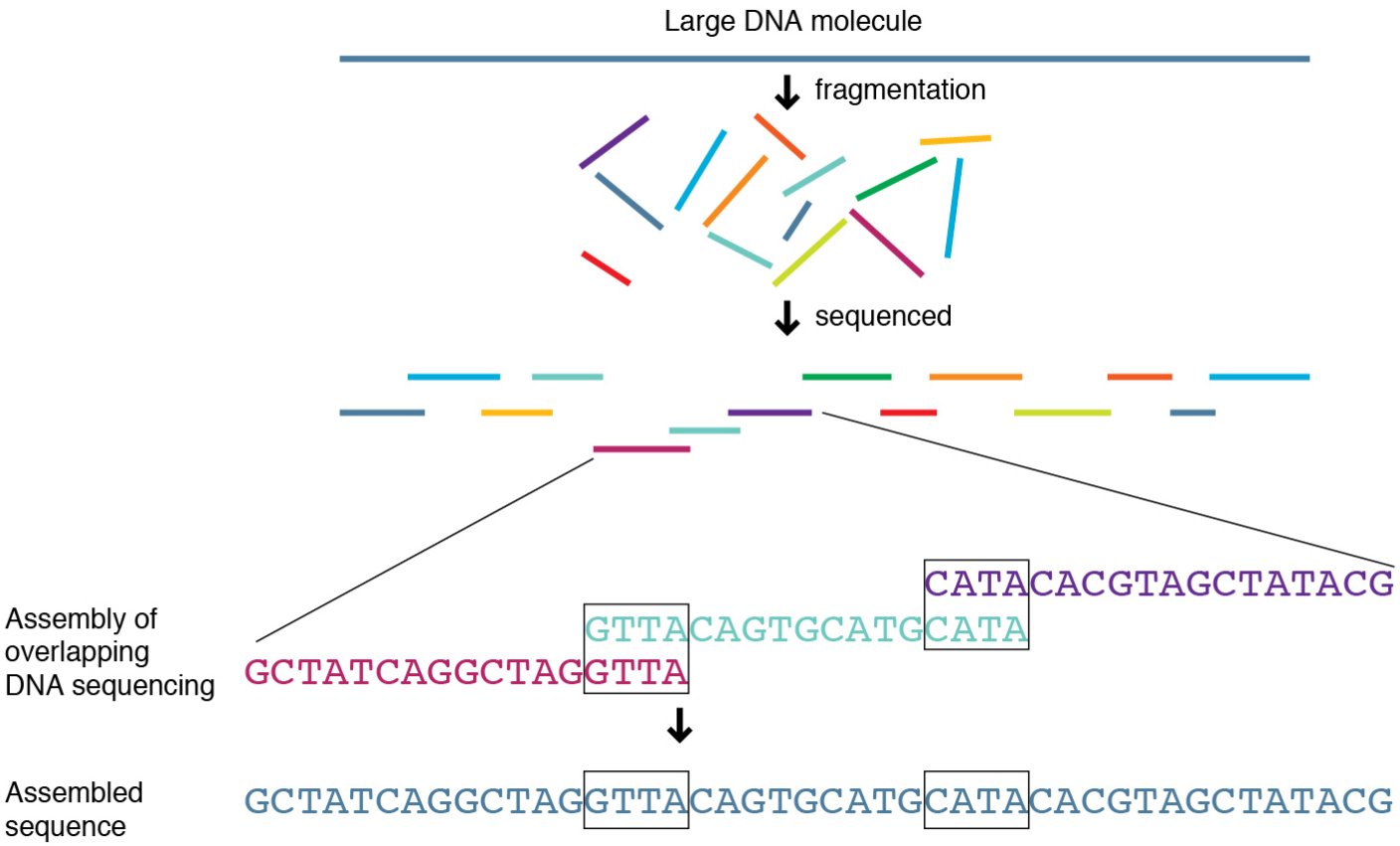


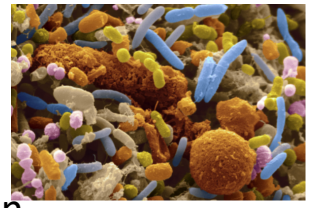
Goal: Identify the genomic content of a community, its composition and function
Method: …



NGS short-read sequencing: MetaG

Goal: Identify the genomic content of a community, its composition and function
Method: Shotgun sequencing



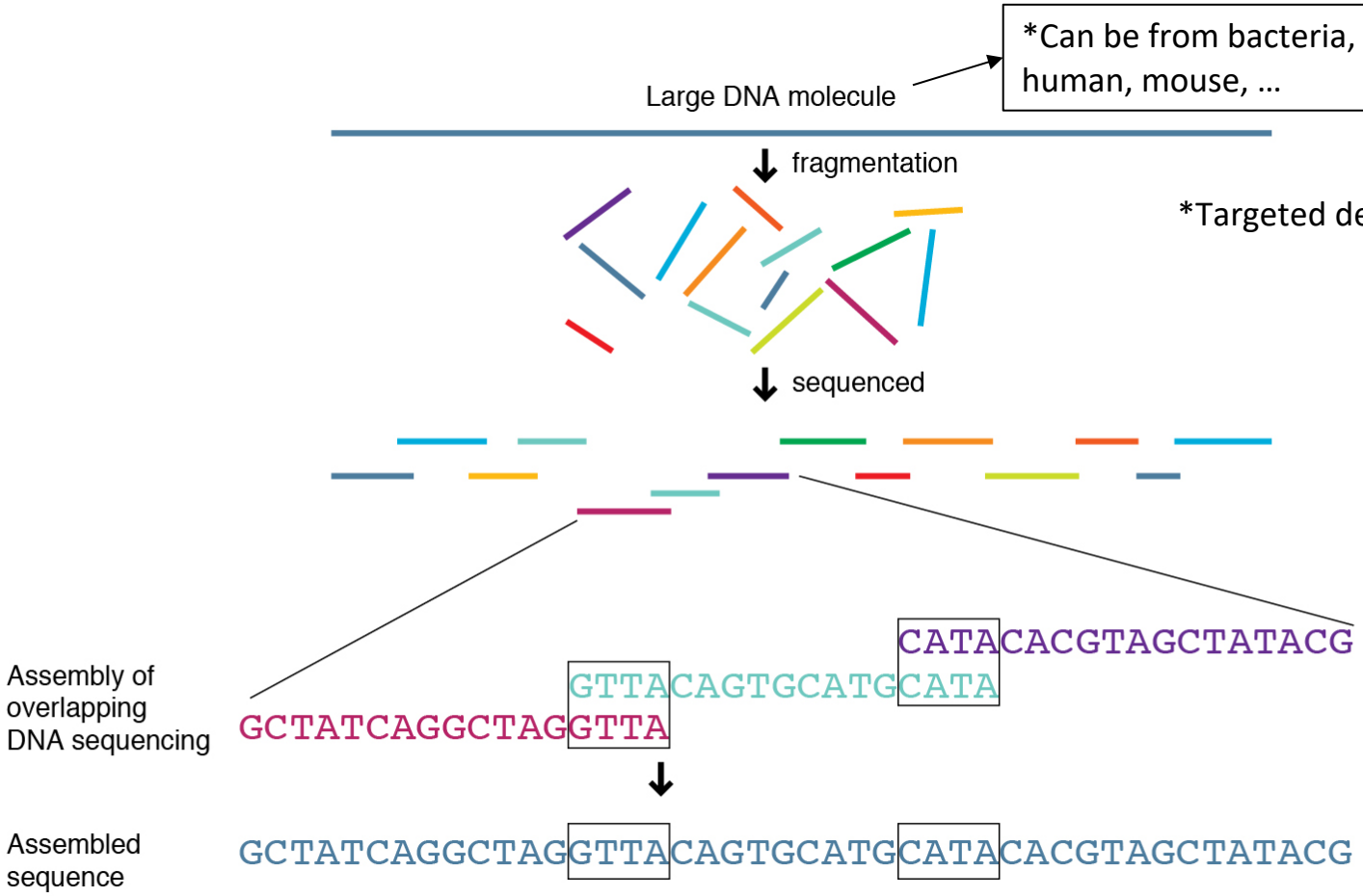


NGS short-read sequencing: MetaG

Goal: Identify the genomic content of a community, its composition and function

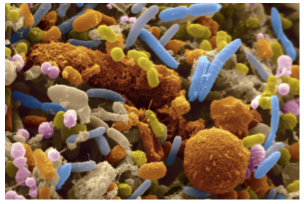
Method: Shotgun sequencing

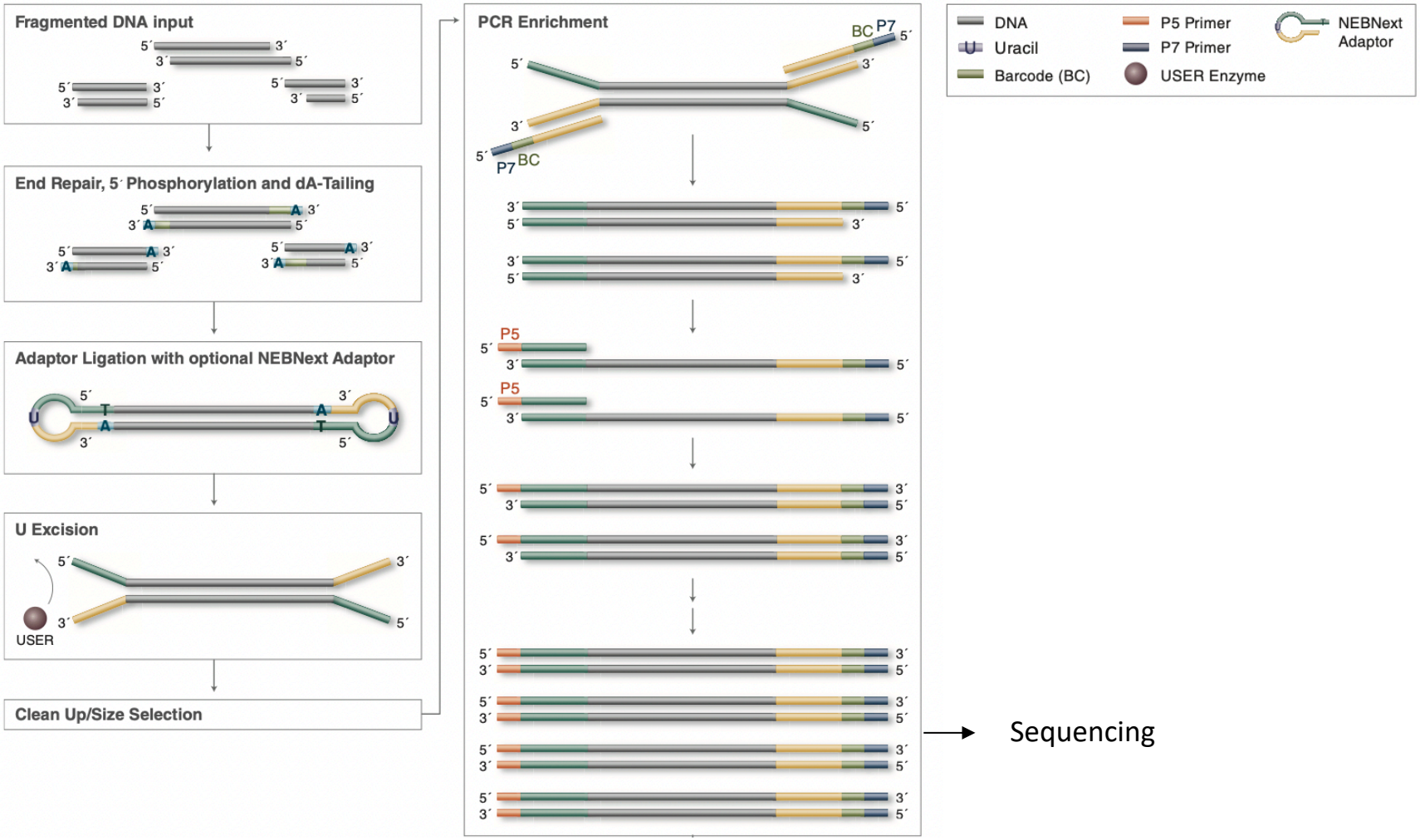
*Can be from bacteria, virus, fungus, human, mouse, …



https://www.genome.gov/genetics-glossary/Shotgun-Sequencing

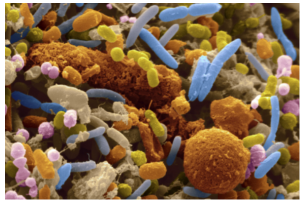
NGS short-read sequencing: MetaG

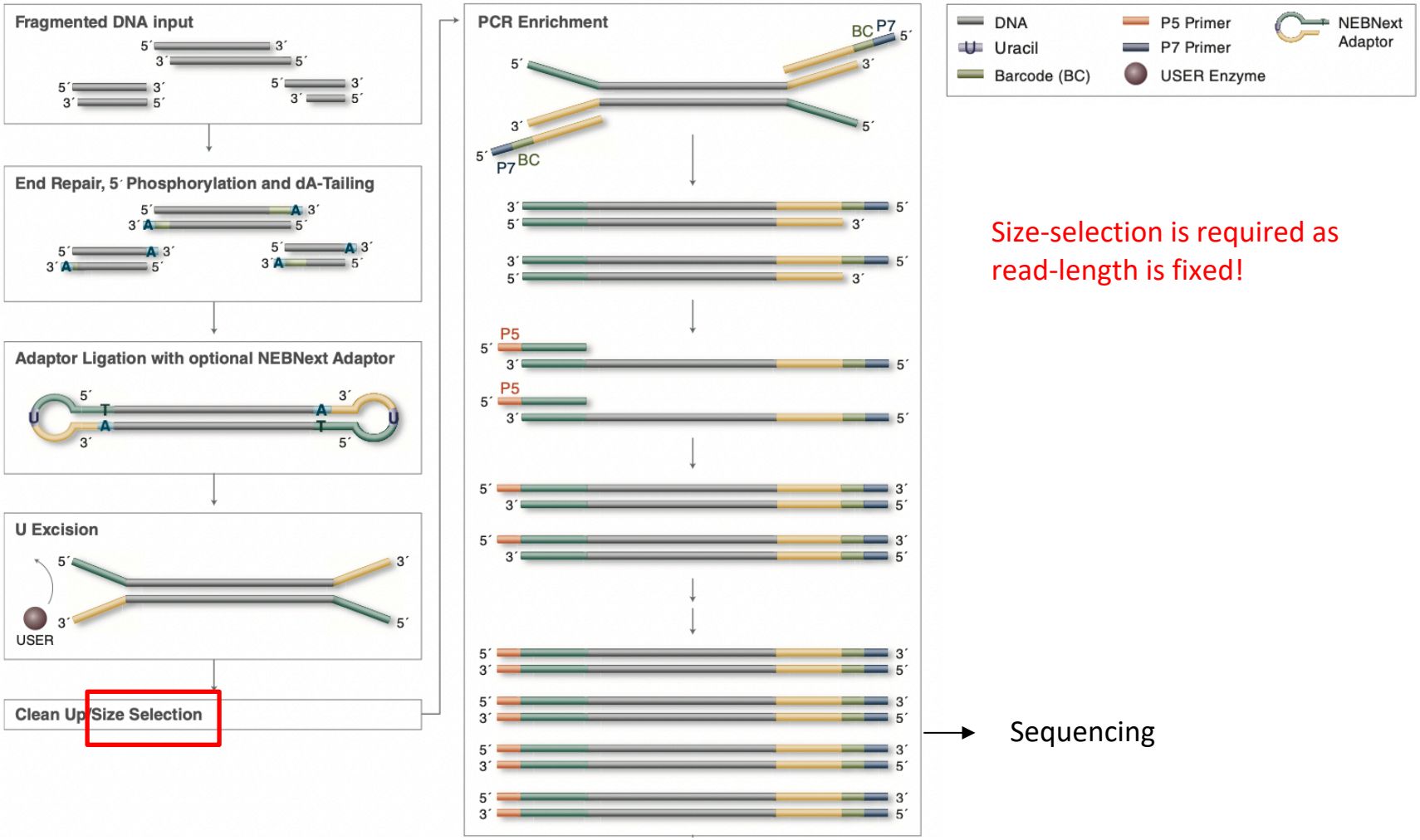




https://international.neb.com/-/media/nebus/files/manuals/manuale7103-e7645.pdf?rev=339d6c65a9314c9e988851a9d671fd9a&hash=2AF765847CD54F1B7464205F7920A50F

NGS short-read sequencing: MetaG

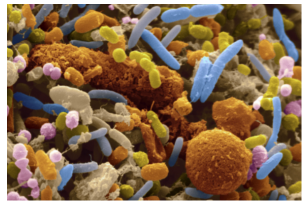




Size-selection is required as read-length is fixed!

→ Sequencing

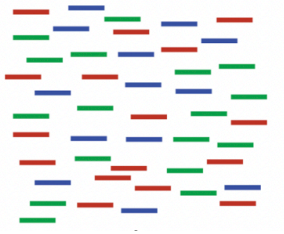
https://international.neb.com/-/media/nebus/files/manuals/manuale7103-e7645.pdf?rev=339d6c65a9314c9e988851a9d671fd9a&hash=2AF765847CD54F1B7464205F7920A50F



NGS short-read sequencing: MetaG

Data analysis pipeline

Individual reads

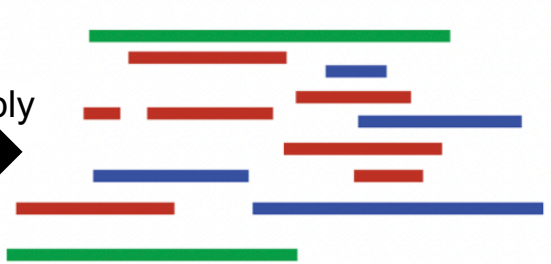


Read QC

1. Remove sequencing adaptors
2. Quality trimming of reads
3. Remove unwanted reads (human, mouse, etc.)

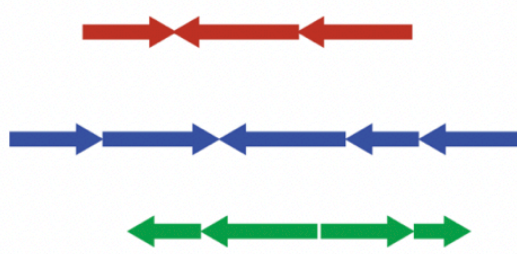
Assembly

Assembled reads



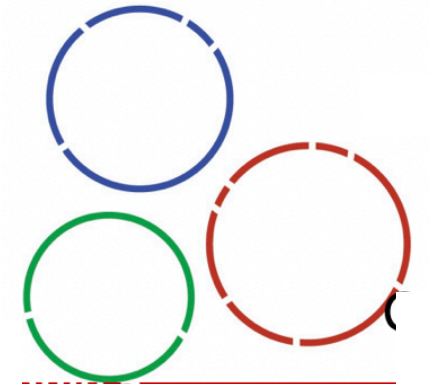
Gene calling

Annotated scaffolds



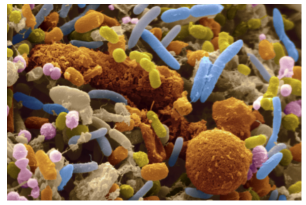
Binning

MAGs*



***Metagenome-assembled genomes**
→ More this afternoon from Lucas

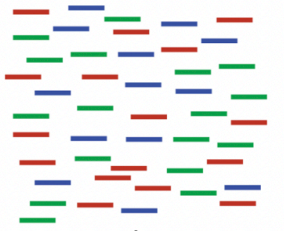
https://bioinformatics.ca/workshops/2018-analysis-of-metagenomic-data/



NGS short-read sequencing: MetaG

Data analysis pipeline

Individual reads

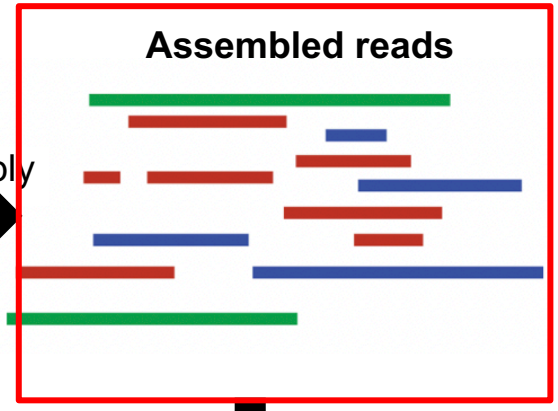


Read QC

1. Remove sequencing adaptors
2. Quality trimming of reads
3. Remove unwanted reads (human, mouse, etc.)

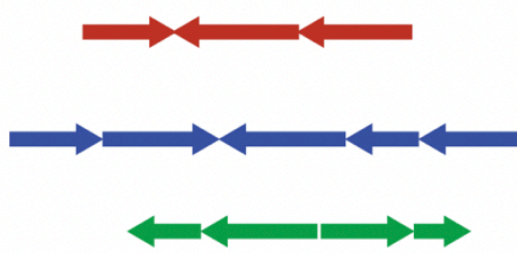
Assembly

Assembled reads



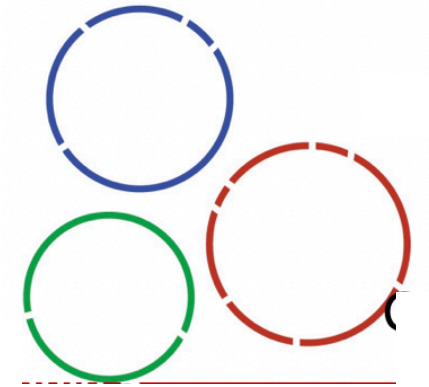
Gene calling

Annotated scaffolds



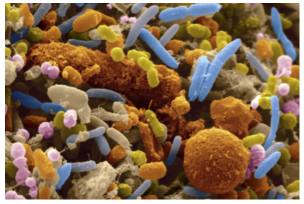
Binning

MAGs*



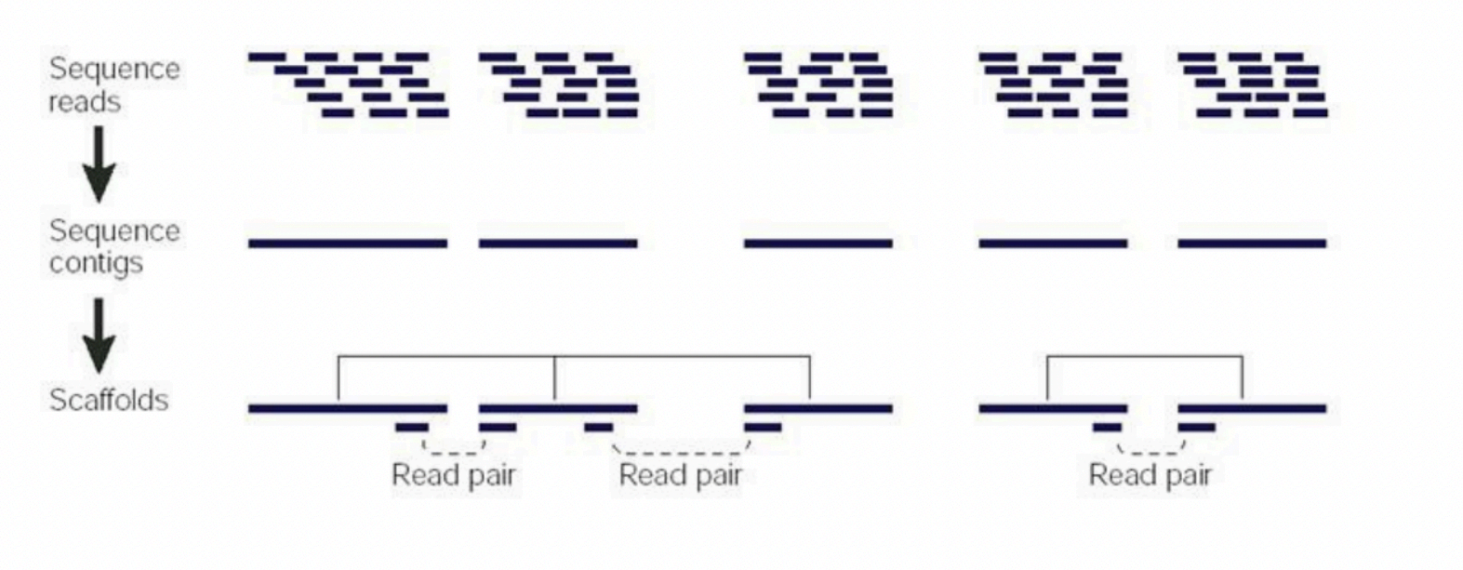
***Metagenome-assembled genomes**
→ More this afternoon from Lucas

https://bioinformatics.ca/workshops/2018-analysis-of-metagenomic-data/



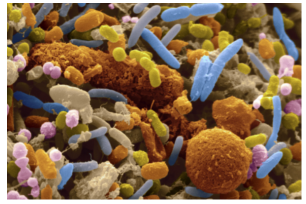
NGS short-read sequencing: MetaG

Data analysis pipeline



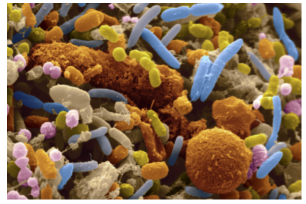
https://bioinformatics.ca/workshops/2018-analysis-of-metagenomic-data/

NGS short-read sequencing: MetaG



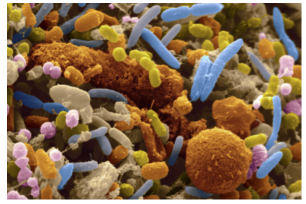
Data applications

NGS short-read sequencing: MetaG



Data applications

1. Accurate microbial abundance estimation using marker genes
2. Increased taxonomic resolution
3. Linking function to phenotype

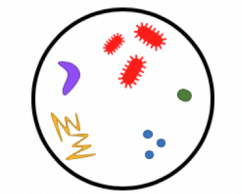


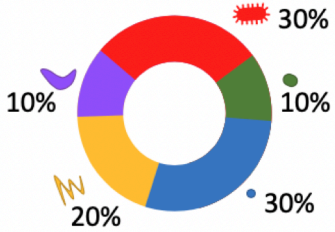
NGS short-read sequencing: MetaG

1. Accurate microbial abundance estimation using marker genes

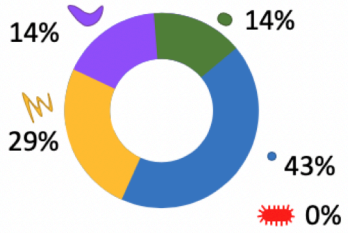
Common problems when using metaB data:

- Variation in 16S copy number
- Taxonomic annotation is database-dependent





If the red species is not in the database

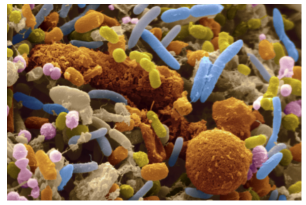


Further:

- Genomes from different species can be up to 95% identical[1]
  → hard to map reads of length 100-150 to the original genome
- Genomes have different length

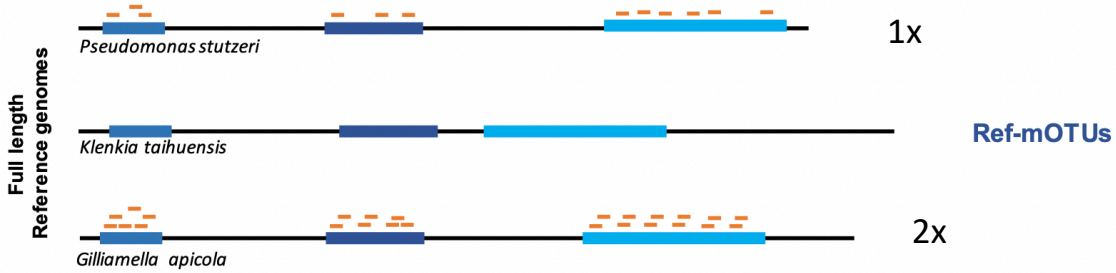
[1]Jain et al. Nature Comm. (2018)

NGS short-read sequencing: MetaG



1. Accurate microbial abundance estimation using marker genes

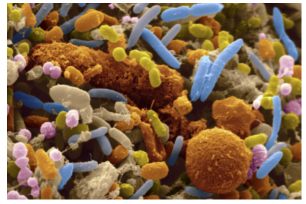
Solution: Single-copy universal marker genes
→ Present in almost all known organisms
→ Only one copy within each genome



Uses 10 universal single-copy marker genes (here 3 are represented)

Map metagenomic reads to marker genes using
a) Reference genomes

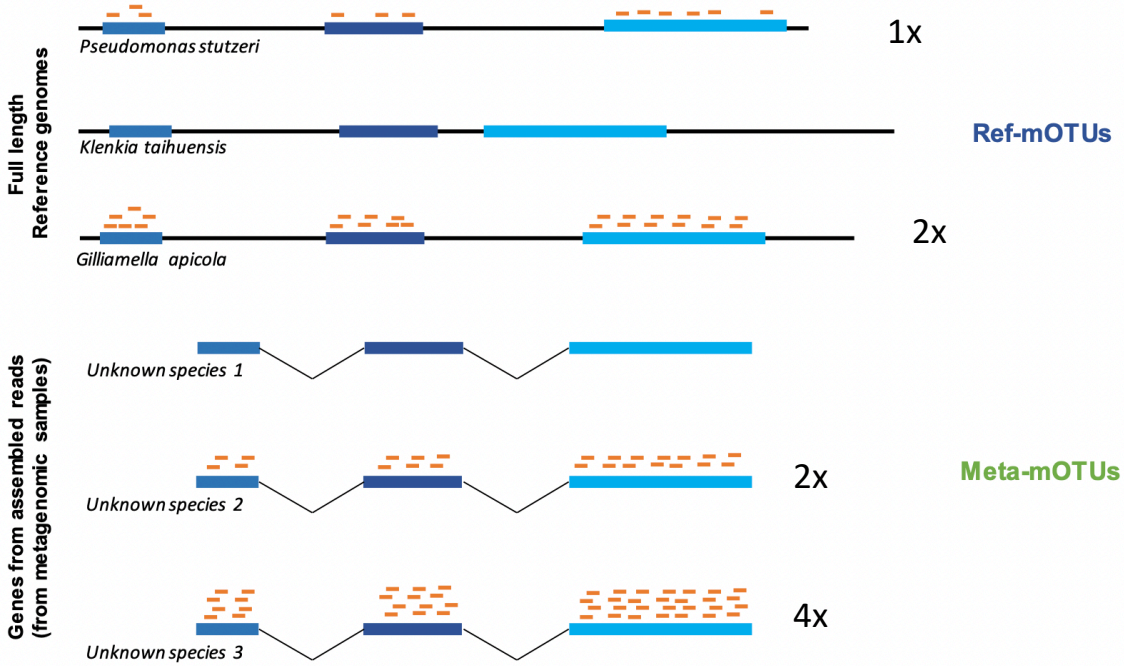
Sunagawa et al. Nature methods (2013); Milanese et al. Nature comm (2019)



NGS short-read sequencing: MetaG

1. Accurate microbial abundance estimation using marker genes

Solution: Single-copy universal marker genes
→ Present in almost all known organisms
→ Only one copy within each genome

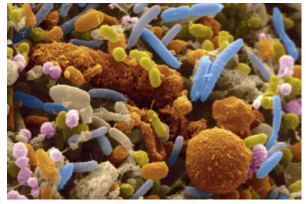


Uses 10 universal single-copy marker genes (here 3 are represented)

Map metagenomic reads to marker genes using
a) Reference genomes
b) Assembled and linked contigs

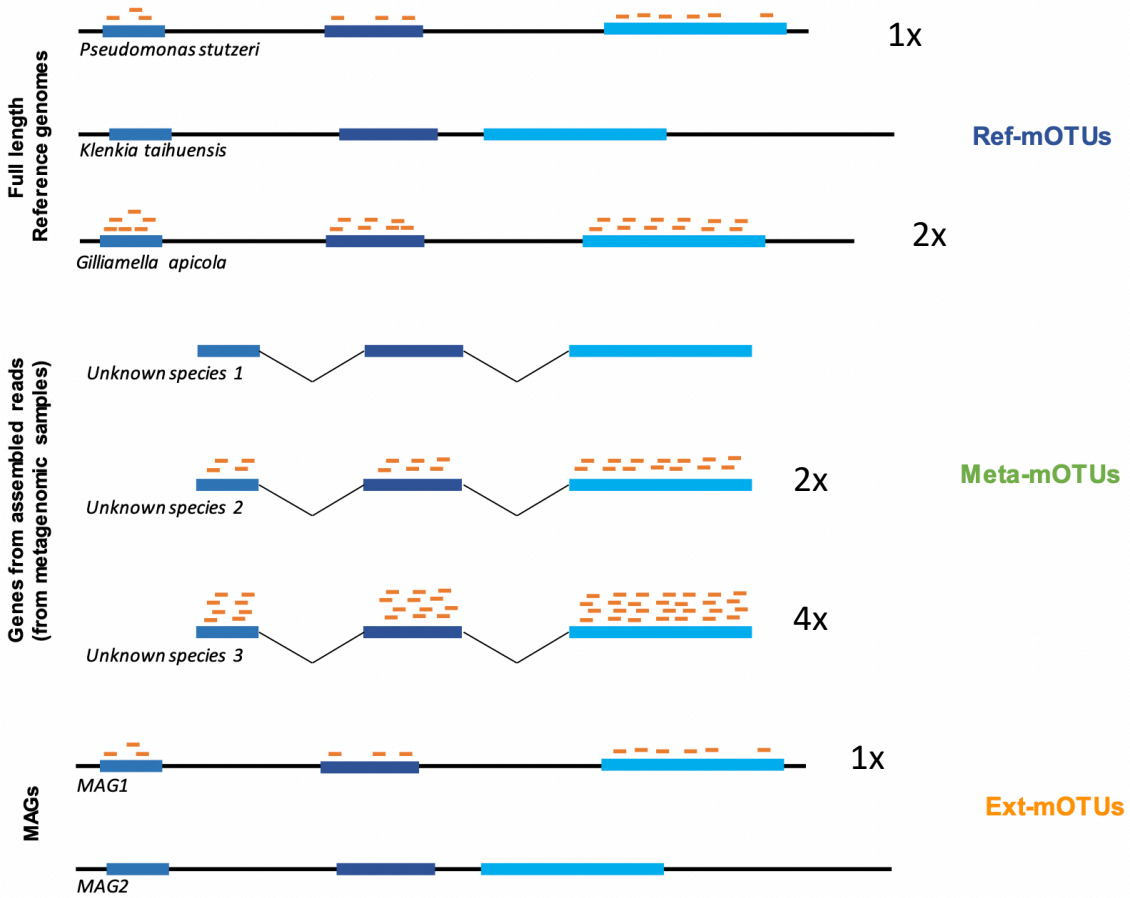
Sunagawa et al. Nature methods (2013); Milanese et al. Nature comm (2019)

NGS short-read sequencing: MetaG



1. Accurate microbial abundance estimation using marker genes

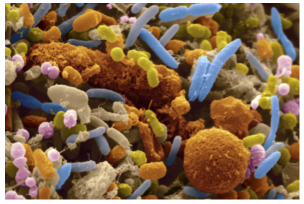
Solution: Single-copy universal marker genes
→ Present in almost all known organisms
→ Only one copy within each genome



Uses 10 universal single-copy marker genes (here 3 are represented)

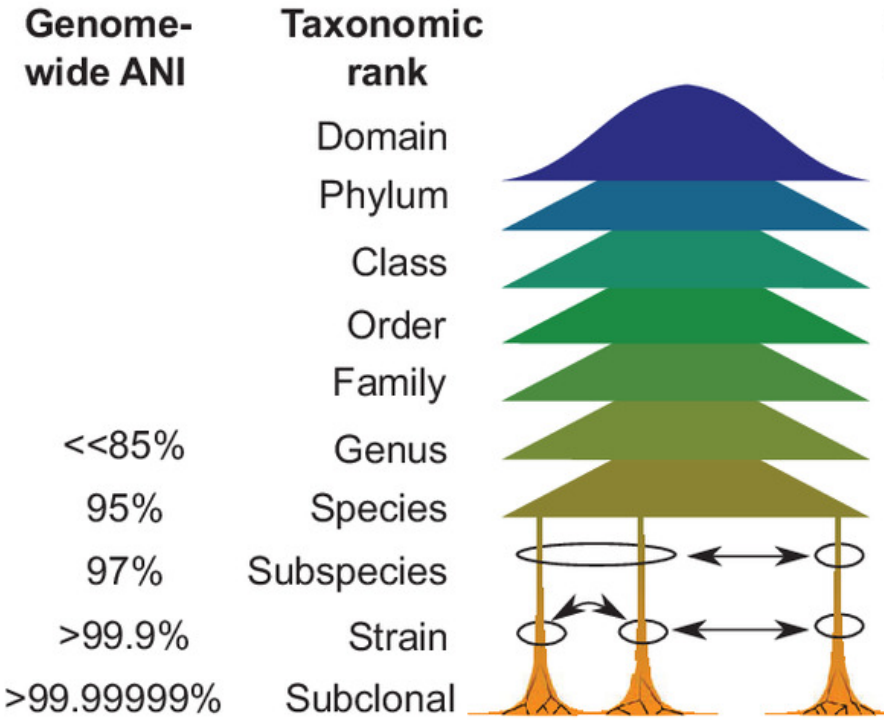
Map metagenomic reads to marker genes using
a) Reference genomes
b) Assembled and linked contigs
c) MAGs

Sunagawa et al. Nature methods (2013); Milanese et al. Nature comm (2019)

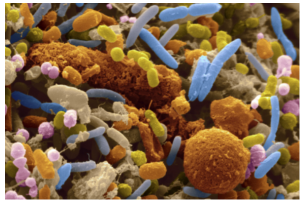


NGS short-read sequencing: MetaG

2. Increased taxonomic resolution

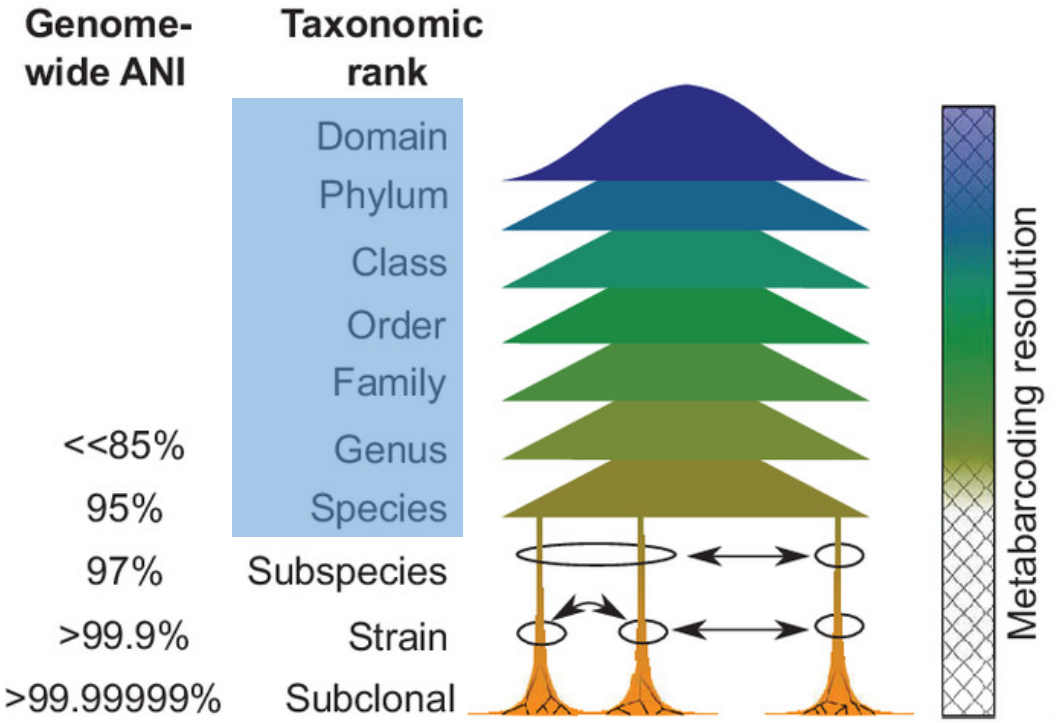


Hildebrand (2021) mSystems 6(4) e00881-21

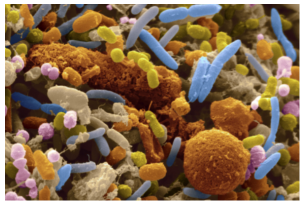


NGS short-read sequencing: MetaG

2. Increased taxonomic resolution

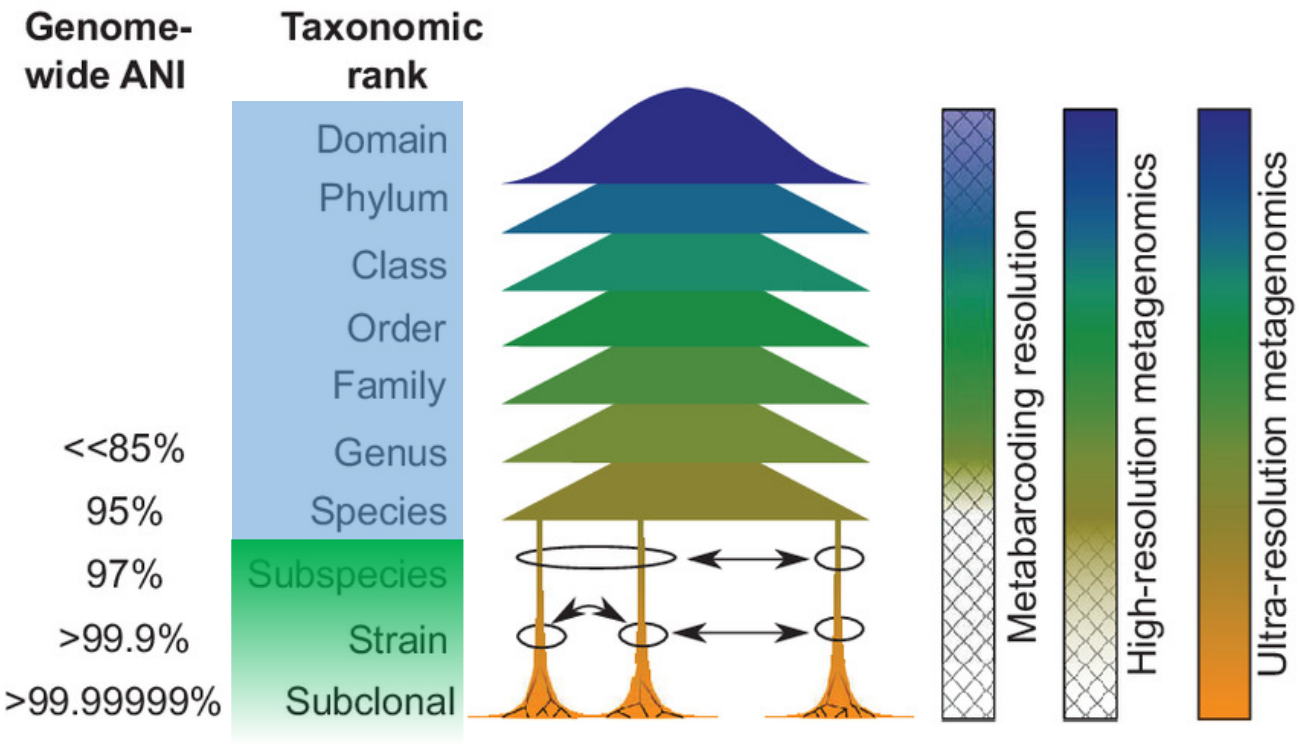


Hildebrand (2021) mSystems 6(4) e00881-21



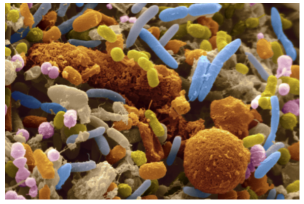
NGS short-read sequencing: MetaG

2. Increased taxonomic resolution



Hildebrand (2021) mSystems 6(4) e00881-21

NGS short-read sequencing: MetaG



3. Linking function to phenotype

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE
Host-Microbe Biology

A Metagenomic Meta-analysis Reveals Functional Signatures of Health and Disease in the Human Gut Microbiome

Courtney R. Armour,[a,b] Stephen Nayfach,[c,d] Katherine S. Pollard,[d,e,f] Thomas J. Sharpton[b,g]

[a]Molecular and Cellular Biology Program, Oregon State University, Corvallis, Oregon, USA
[b]Department of Microbiology, Oregon State University, Corvallis, Oregon, USA
[c]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA
[d]Gladstone Institutes, San Francisco, California, USA
[e]Department of Epidemiology & Biostatistics, Institute for Human Genetics, Quantitative Biology Institute, and Institute for Computational Health Sciences, University of California, San Francisco, California, USA
[f]Chan-Zuckerberg Biohub, San Francisco, California, USA
[g]Department of Statistics, Oregon State University, Corvallis, Oregon, USA

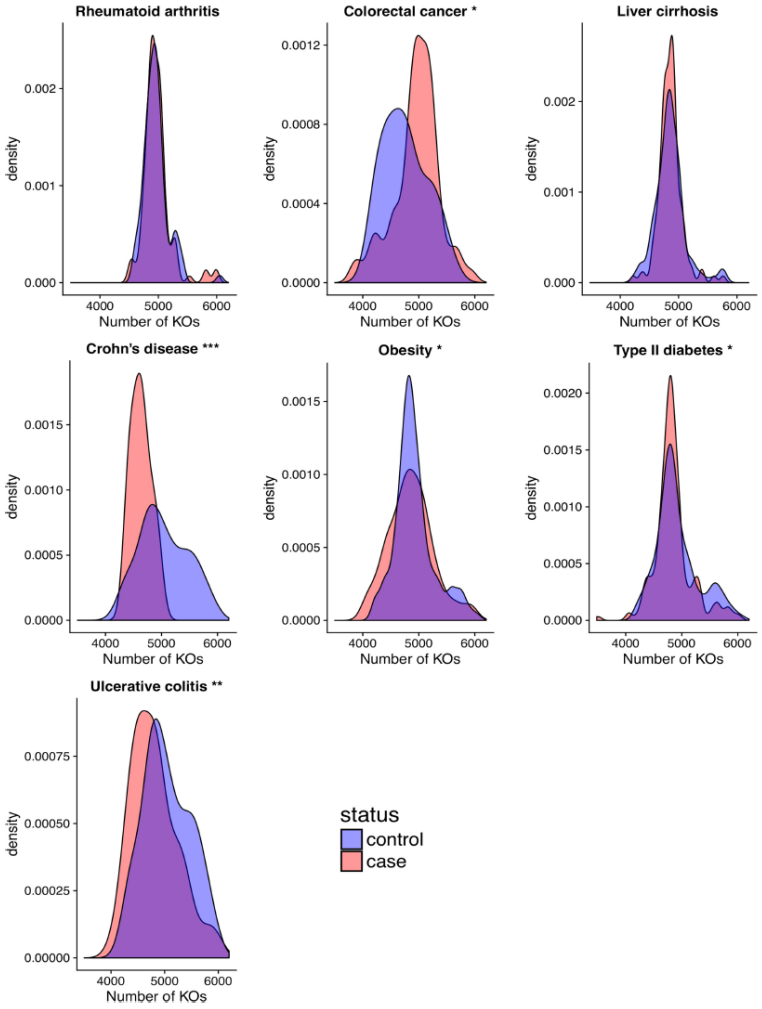
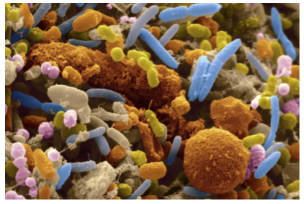


FIG 1 Protein family richness associates with disease. Density plots of the distribution of protein family richness across case and control populations for seven diseases. Asterisks beside plot titles indicate significance from Student's *t* test (*, *P*0.05; **, *P*0.01; ***, *P*0.001)

Armour, R., Nayfach, S., Pollard, K.S., Sharpton, T.J. (2019) mSystems 4(4) e00332-18



NGS short-read sequencing: MetaG

3. Linking function to phenotype

- Shown are MDS plots based on Bray-Curtis dissimilarity between all samples based on their KO abundances
- functional composition of the gut microbiome differs between case and control populations for 6 out of 7 diseases

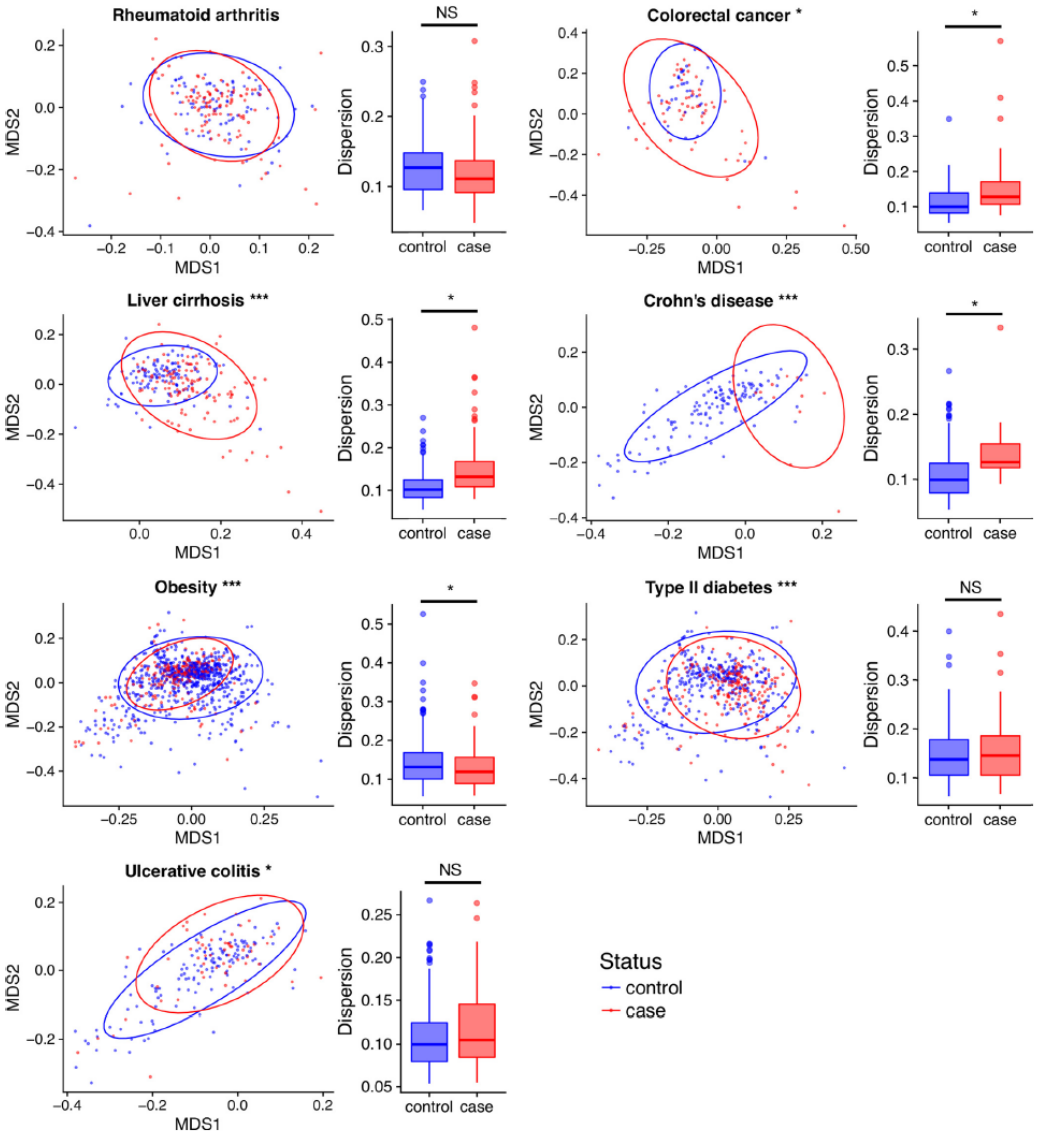
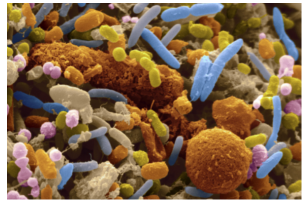


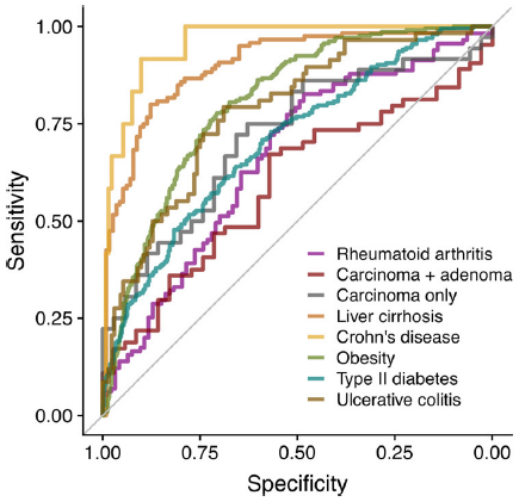
FIG 2 Changes in functional composition associate with disease. NMDS plots of Bray-Curtis dissimilarity between cases and controls across diseases; ellipses represent 95% confidence level. Asterisks in NMDS plot titles indicate significance from PERMANOVA (***, *P*0.001; Table S6). Box plots represent dispersion in beta-diversity within groups. Asterisks in box plots denote significance from *P* test and ANOVA (*, *P* 0.05).

Armour, R., Nayfach, S., Pollard, K.S., Sharpton, T.J. (2019) mSystems 4(4) e00332-19



NGS short-read sequencing: MetaG

3. Linking function to phenotype



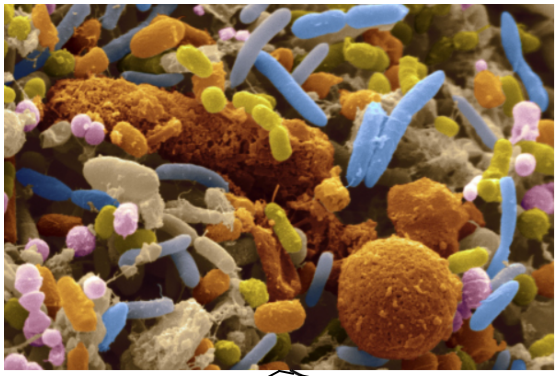
Random Forest by Disease				
Level	**Disease**	**Color**	**OOB error**	**AUC**
Module	Crohn's Disease	Yellow	5.56%	0.954
Module	Liver cirrhosis	Orange	17.09%	0.902
Module	Obesity	Green	22.57%	0.803
Module	Ulcerative colitis	Brown	25.26%	0.783
Module	Type II diabetes	Blue	31.58%	0.708
Module	Rheumatoid arthritis	Purple	35.58%	0.664
Module	Colorectal carcinoma	Red	36.36%	0.596
Module	Carcinoma (without adenoma)	Grey	35.21%	0.715

FIG 4 Classifying disease status based on the functional composition of the microbiome. ROC curves from random forest classifiers for cases and controls in each disease. The table shows OOB error and AUC values.

"These results suggest that the potential for use of the functional composition of the gut microbiome in disease diagnosis varies by the type and severity of disease"

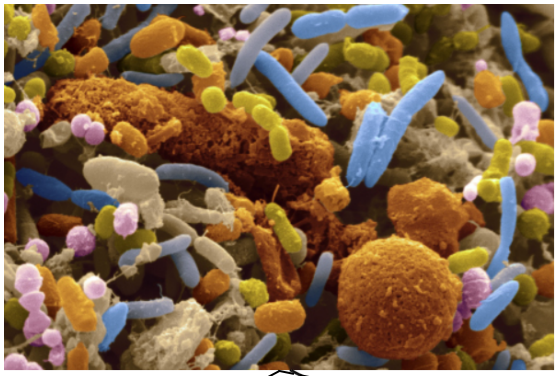
Armour, R., Nayfach, S., Pollard, K.S., Sharpton, T.J. (2019) mSystems 4(4) e00332-18

NGS short-read sequencing: Different data types



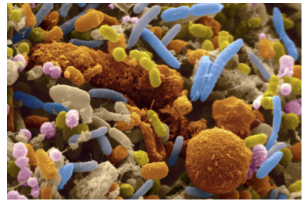
	Meta-barcoding (metaB)	Meta-genomics (metaG)	Meta-transcriptomics (metaT)
Seq approach	Targeted	Non-targeted	Non-targeted
Seq material	Amplicon DNA	Whole genomic DNA	Transcribed RNA
Target	Bacterial genomes	All genomes	All active genomes
Taxonomic precision	Genus/Species	Strain/Genome	Strain/Genome
Resolution	Higher	Lower	Lower
Who is there?			
At what proportions?			
What are they doing?			

NGS short-read sequencing: Different data types



	Meta-barcoding (metaB)	Meta-genomics (metaG)	Meta-transcriptomics (metaT)
Seq approach	Targeted	Non-targeted	Non-targeted
Seq material	Amplicon DNA	Whole genomic DNA	Transcribed RNA
Target	Bacterial genomes	All genomes	All active genomes
Taxonomic precision	Genus/Species	Strain/Genome	Strain/Genome
Resolution	Higher	Lower	Lower
Who is there?	Yes	Yes	Yes, if active
At what proportions?	Yes (with limitations)	Yes	Yes if normalized with metaG
What are they doing?	No	Yes (metabolic potential)	Yes

SUMMARY



Different DNA sequencing technology have been developed over the last 30 years

Different sequencing technologies still in use today for specific applications/questions

Usually there is a trade-off between throughput and accuracy (but still improving) → needs to be tailored to research question

Different technologies generate different data types with individual characteristics (Pros and Cons) → needs to be tailored to research question

Meta-barcoding: Cheap, abundance-independent, limited taxonomic resolution, no functional information

Meta-genomics: expensive, abundance-dependent, high taxonomic resolution, functional information

Break…

Block-course study data: **The gut microbiome in acute myeloid leukemia (AML)**

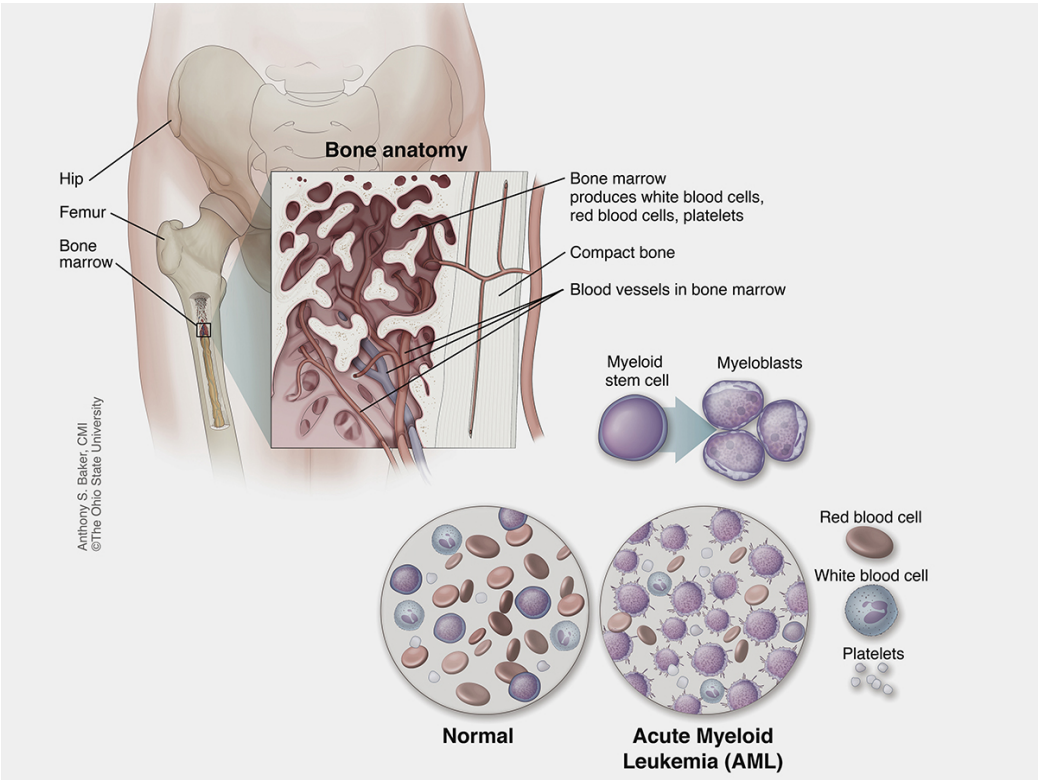
AML = Acute myeloid leukemia

→ Cancer of the blood and bone marrow that progresses quickly and always ends in death if untreated

→ Increased incidence with age

→ Different genetic variants known to affect treatment outcome

→ Current best treatment approach: **Intensive chemotherapy**



Less room for healthy cells
→ Frequent infections, anemia, bleeding…

Block-course study data: **Impact of intestinal microbiota on systemic infections, response to chemotherapy and overall outcomes in patients with acute myeloid leukemia – a prospective, non-interventional, single-center study**

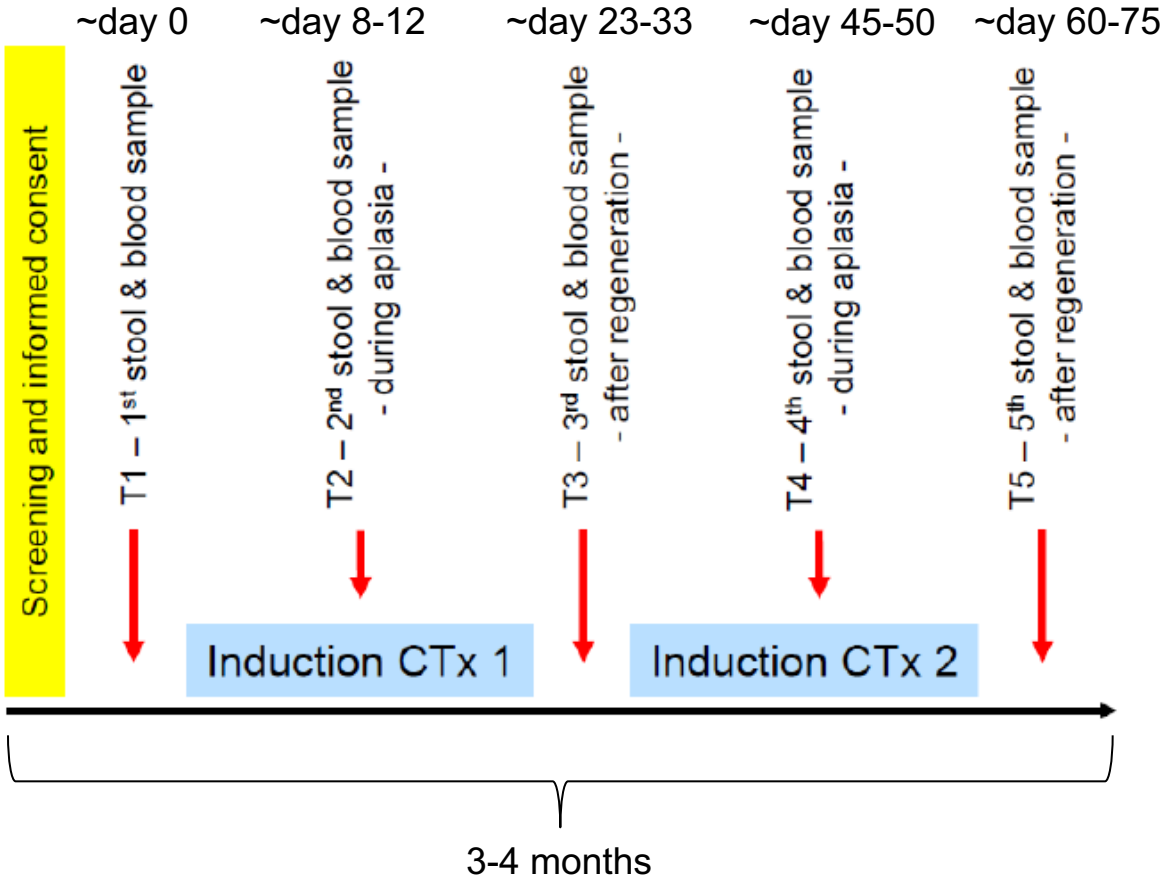
Intensive chemotherapy

- highly toxic (Gastrointestinal mucositis with enterocolitis extremely common)
- high risk of life-threatening infections during neutropenia
- Gut is main source of bacteria causing infections → use of antibiotics/gut decontamination
- Overall benefit of gut decontamination unknown
- Dysbiosis of the gut microbiome caused by gut decontamination might aggravate patient susceptibility to infection

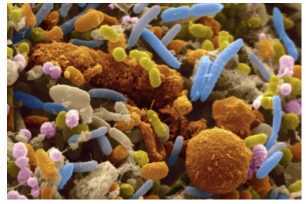
Bottom line: **"The impact of intensive chemotherapy with/without prophylactic gut decontamination on the microbiota, systemic infections and leukemia response in AML patients has not been clarified"**

Block-course study data: **Impact of intestinal microbiota on systemic infections, response to chemotherapy and overall outcomes in patients with acute myeloid leukemia – a prospective, non-interventional, single-center study**

Study design



NGS long-read sequencing: MetaB and MetaG



…