Introduction to

Microbiome Sequencing

STAMPS 2025

What is a Chicken?

- **If you keep feeding them, they keep laying eggs**
- **They don’t fly away**
- **They are hardy and self-sufficient**

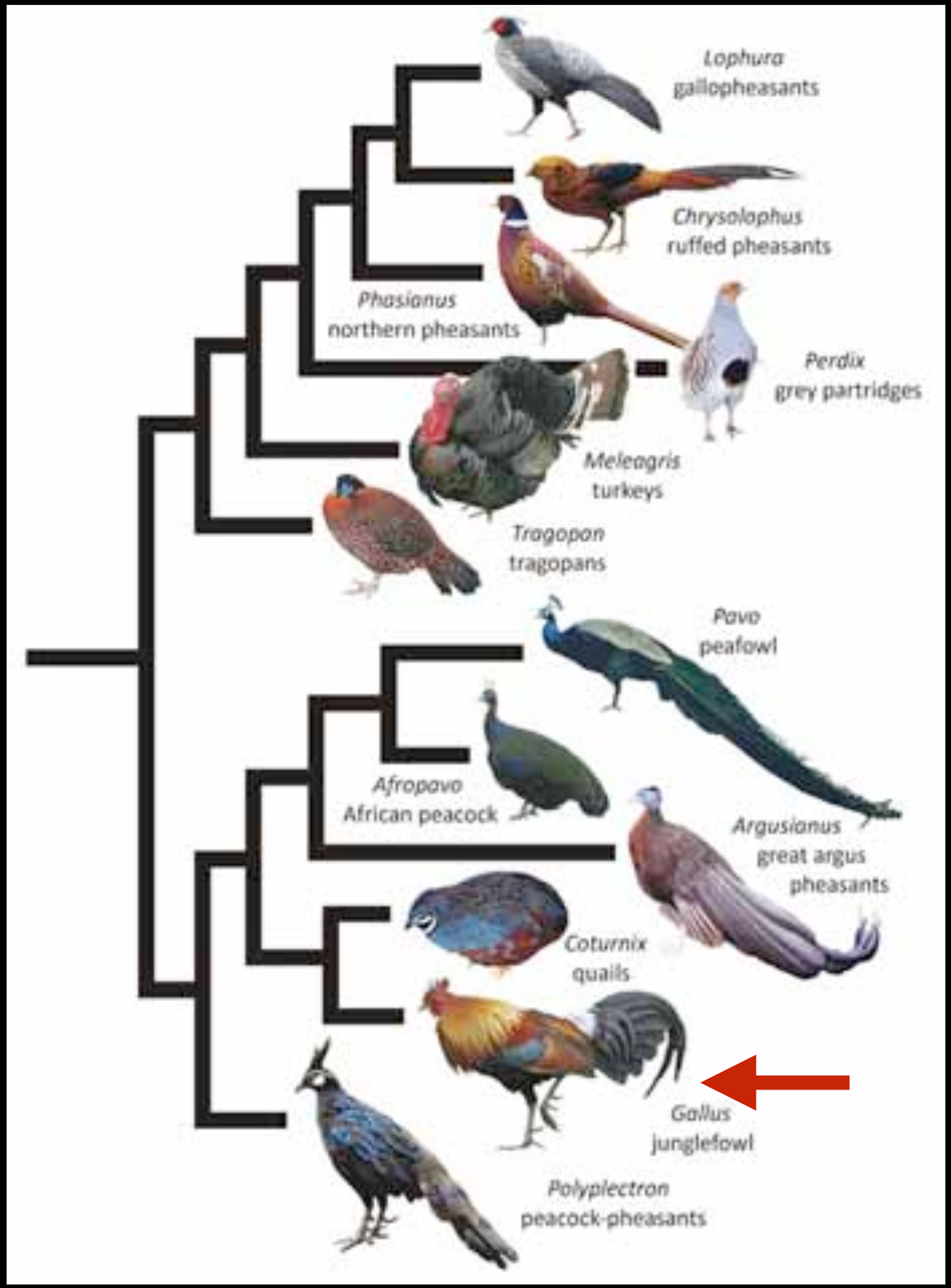
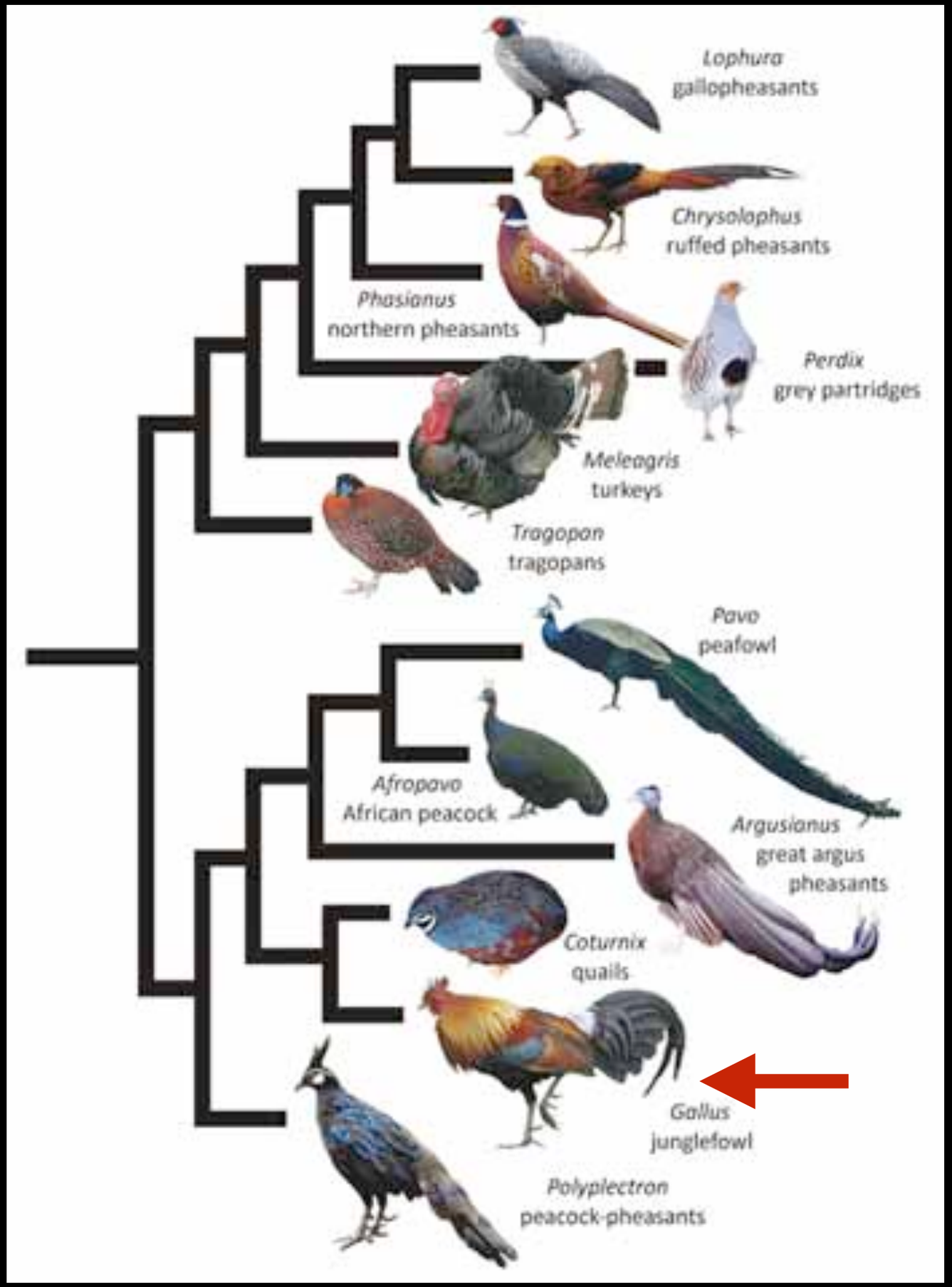


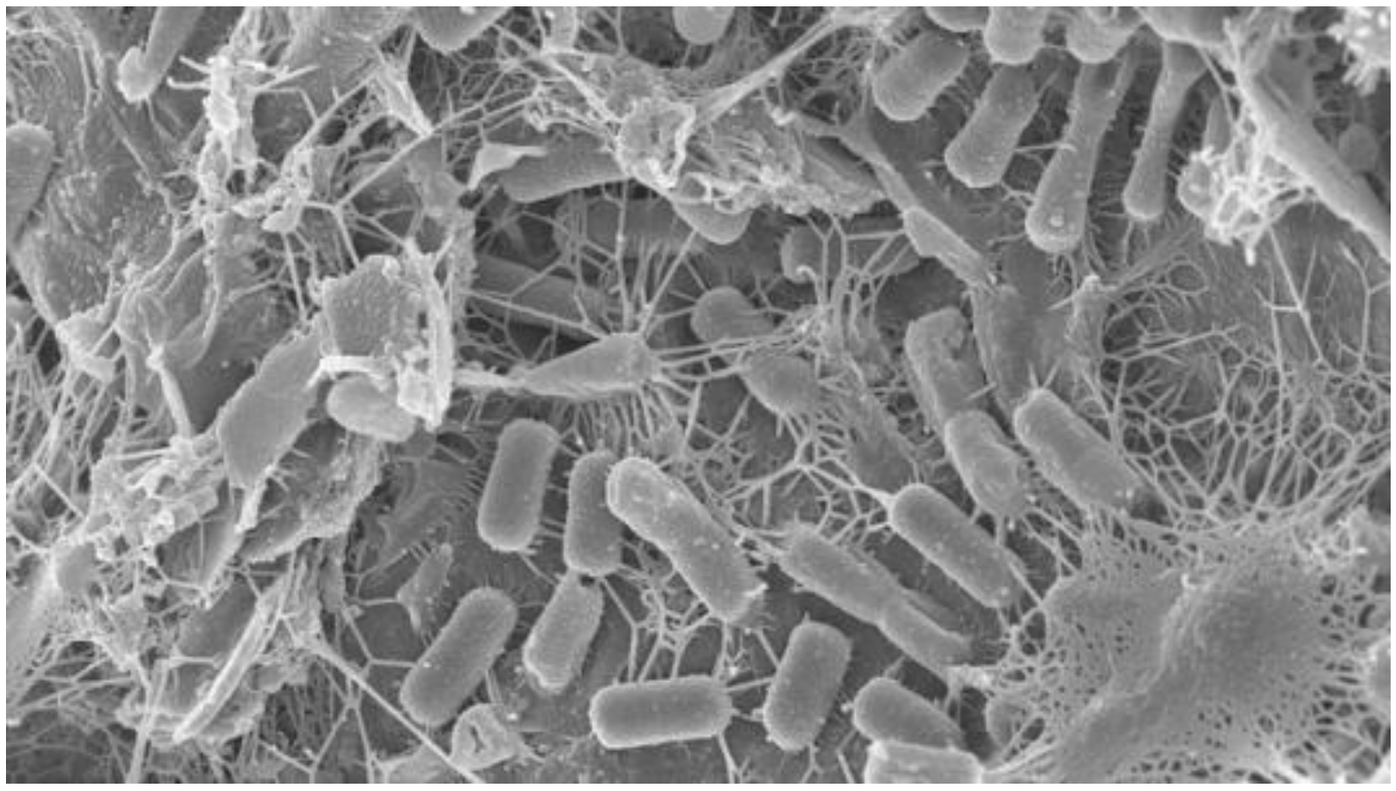
Image: Nash, Scientific American, 2013



Biological species concept.

Phylogenetic organization by relatedness.

Image: Nash, Scientific American, 2013



Images: Wikipedia, Argonne National Lab

What is an *E. coli*?

Diversity of Life

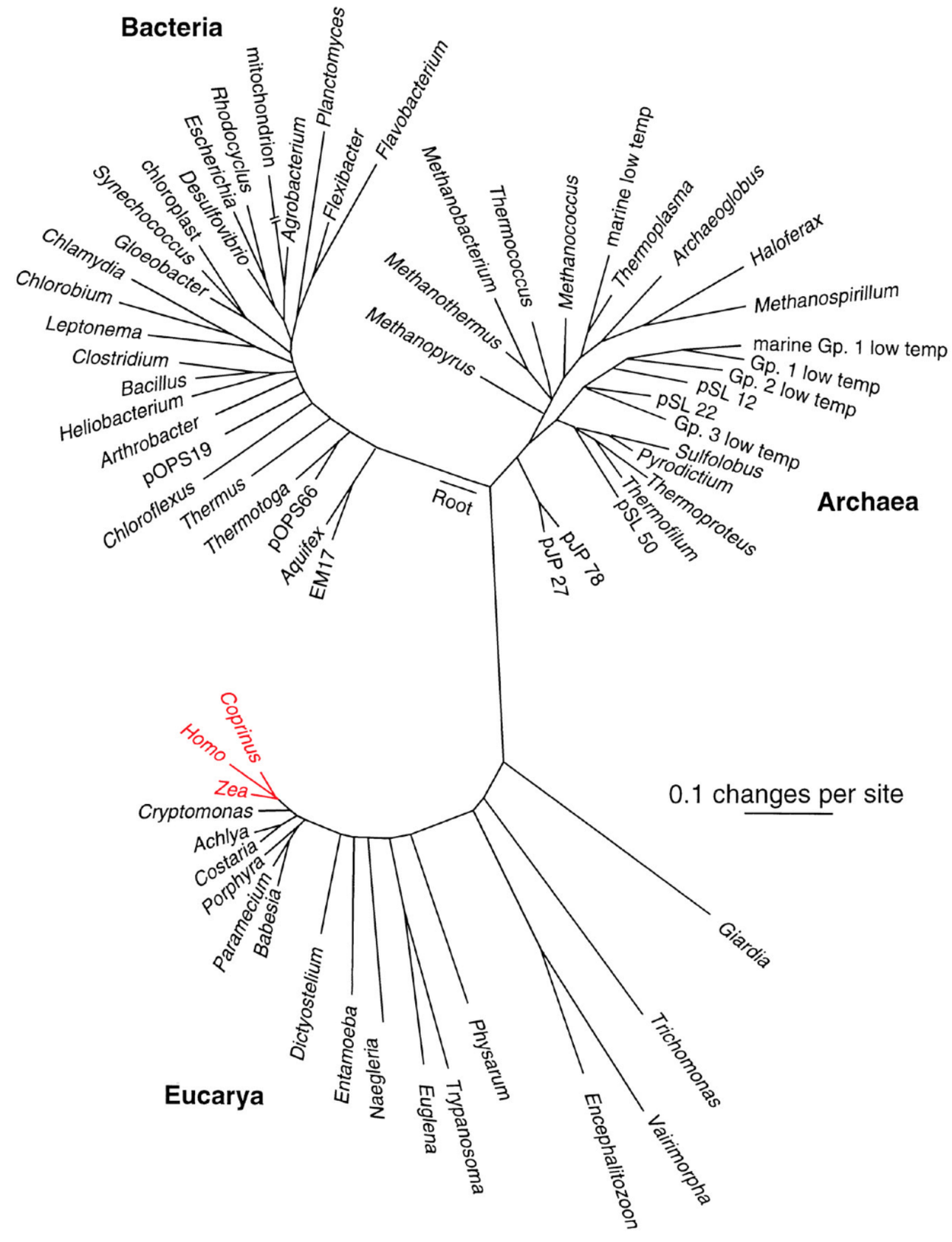


Image: Norman Pace

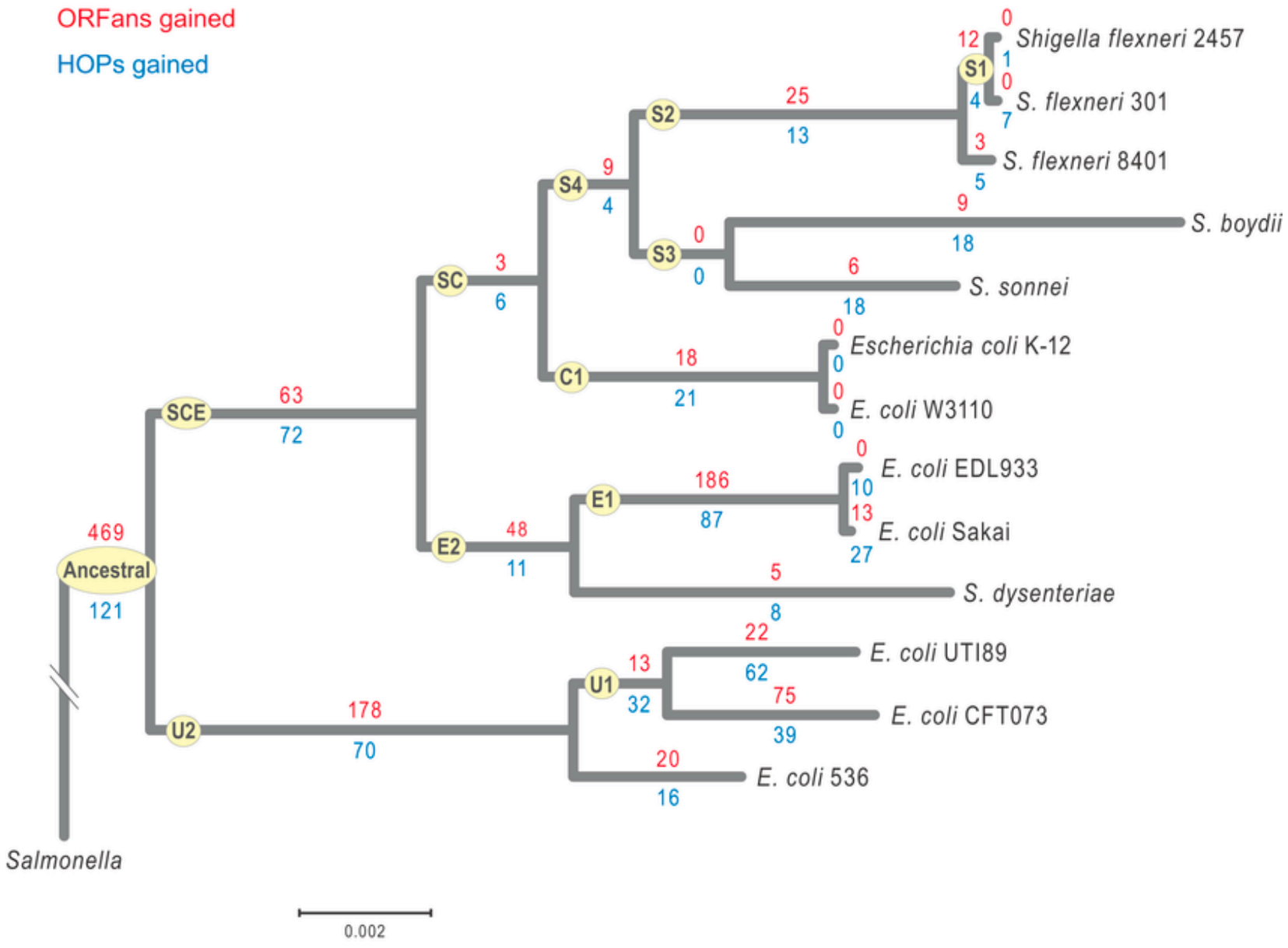
Organizing the Microbes

- **Early Linnaen (1872-1900)**: pathogenic potential, chemical reactions, requirements for growth, and morphology ("haphazard and non-scientific")

Reference: "Reconciling Microbial Systematics & Genomics", ASM 2006.

Organizing the Microbes

- **Early Linnaen (1872-1900)**: pathogenic potential, chemical reactions, requirements for growth, and morphology ("haphazard and non-scientific")



Reference: "Reconciling Microbial Systematics & Genomics", ASM 2006.

Organizing the Microbes

- **Early Linnaen (1872-1900)**: pathogenic potential, chemical reactions, requirements for growth, and morphology ("haphazard and non-scientific")

- **Botanical (1900-1950s)**: morphology first, then physiology to discriminate among the more closely aligned organisms. Bacterial species defined: "the type culture together with such other cultures or strains of bacteria that are accepted by bacteriologists as sufficiently closely related"

Reference: "Reconciling Microbial Systematics & Genomics", ASM 2006.

Organizing the Microbes

- **Early Linnaen (1872-1900)**: pathogenic potential, chemical reactions, requirements for growth, and morphology ("haphazard and non-scientific")

- **Botanical (1900-1950s)**: morphology first, then physiology to discriminate among the more closely aligned organisms. Bacterial species defined: "the type culture together with such other cultures or strains of bacteria that are accepted by bacteriologists as sufficiently closely related"

- **Cellular and molecular (1950s-1980s):** Chemotaxonomy, DNA-DNA hybridization, numerical phenotyping, polyphasic taxonomy.

Reference: "Reconciling Microbial Systematics & Genomics", ASM 2006.

Organizing the Microbes

- **Early Linnaen (1872-1900)**: pathogenic potential, chemical reactions, requirements for growth, and morphology ("haphazard and non-scientific")

- **Botanical (1900-1950s)**: morphology first, then physiology to discriminate among the more closely aligned organisms. Bacterial species defined: "the type culture together with such other cultures or strains of bacteria that are accepted by bacteriologists as sufficiently closely related"

- **Cellular and molecular (1950s-1980s):** Chemotaxonomy, DNA-DNA hybridization, numerical phenotyping, polyphasic taxonomy.

- **DNA and phylogeny (1980s-present):** 16S rRNA gene, microbial genomics, phylogenetics, sequence similarity thresholds.

Reference: "Reconciling Microbial Systematics & Genomics", ASM 2006.

Organizing the Microbes

- **Early Linnaen (1872-1900)**: pathogenic potential, chemical reactions, requirements for growth, and morphology ("haphazard and non-scientific")

- **Botanical (1900-1950s)**: morphology first, then physiology to discriminate among the more closely aligned organisms. Bacterial species defined: "the type culture together with such other cultures or strains of bacteria that are accepted by bacteriologists as sufficiently closely related"

- **Cellular and molecular (1950s-1980s):** Chemotaxonomy, DNA-DNA hybridization, numerical phenotyping, polyphasic taxonomy.

- **DNA and phylogeny (1980s-present):** 16S rRNA gene, microbial genomics, phylogenetics, sequence similarity thresholds.

Increasing role for genetics and phylogeny

Reference: "Reconciling Microbial Systematics & Genomics", ASM 2006.

A current controversy

Resource | Open access | Published: 19 September 2022

SeqCode: a nomenclatural code for prokaryotes described from sequence data

Brian P. Hedlund, Maria Chuvochina, Philip Hugenholtz, Konstantinos T. Konstantinidis, Alison E. Murray, Marike Palmer, Donovan H. Parks, Alexander J. Probst, Anna-Louise Reysenbach, Luis M. Rodriguez-R, Ramon Rossello-Mora, Iain C. Sutcliffe, Stephanus N. Venter & William B. Whitman

Nature Microbiology **7**, 1702–1708 (2022) | Cite this article

Most prokaryotes are not available as pure cultures and therefore ineligible for naming under the rules and recommendations of the International Code of Nomenclature of Prokaryotes (ICNP). Here we summarize the development of the SeqCode, a code of nomenclature under which genome sequences serve as nomenclatural types…

Describing the Microbiome

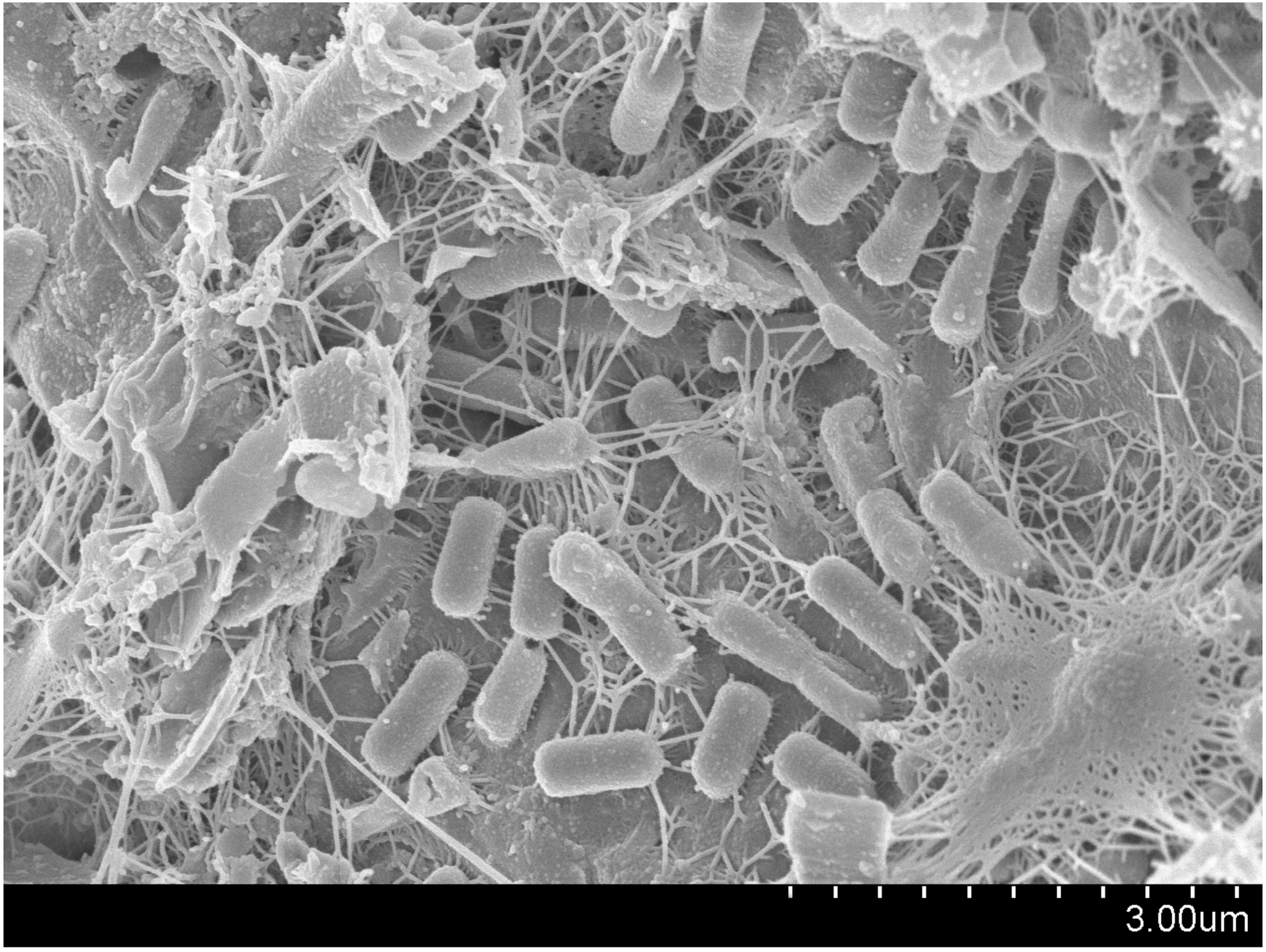


Image: Anthony D'Onofrio, Northeastern University, 2009

A Microbial Census

Marker-gene and Metagenomic Sequencing

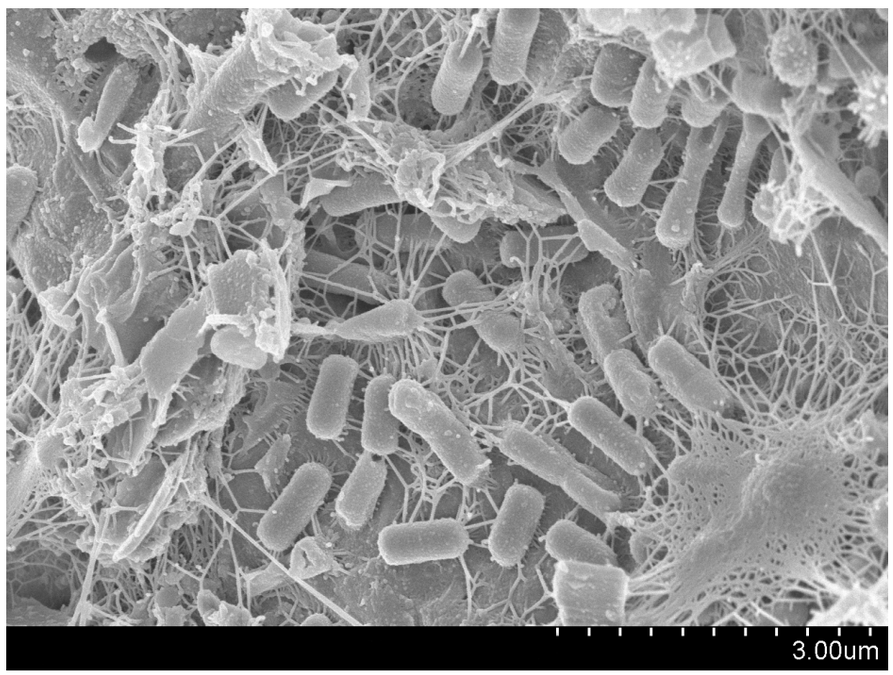


Image: Anthony D'Onofrio, Northeastern University, 2009

A Microbial Census

Marker-gene and Metagenomic Sequencing

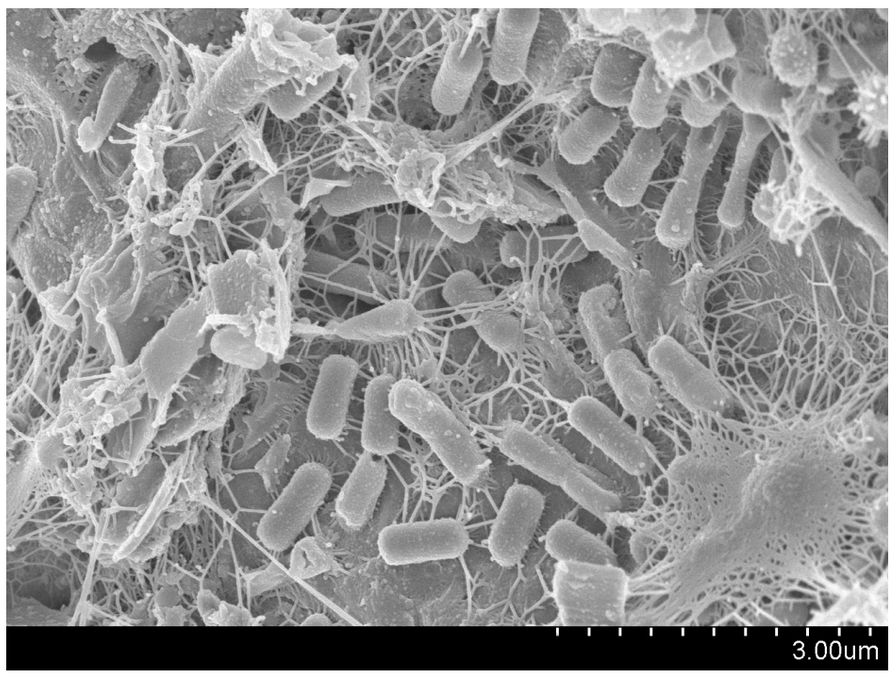
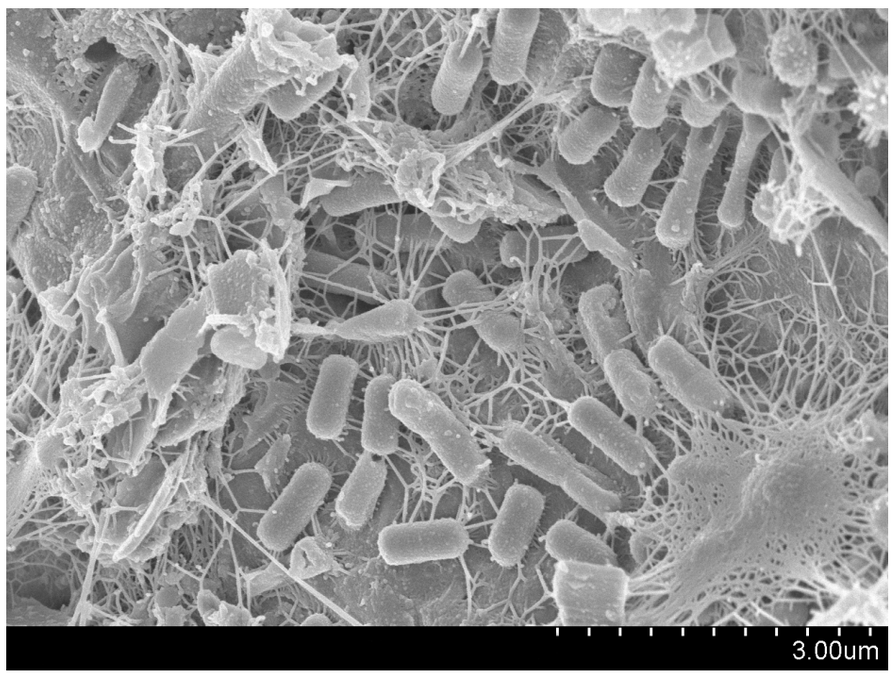


Image: Anthony D'Onofrio, Northeastern University, 2009

A Microbial Census

Marker-gene and Metagenomic Sequencing

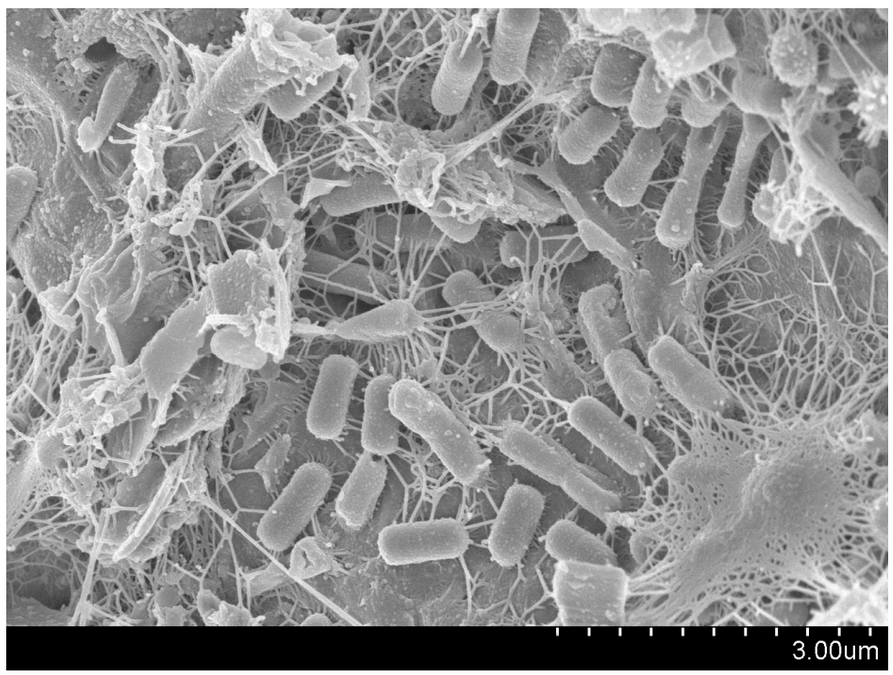


Lactobacillus crispatus	1300	5	0	882	596
Ureaplasma urealytica	15	0	220	0	0
Gardnerella vaginalis	22	0	1	0	412
Prevotella intermedia	0	0	8	12	0
...	...	...	...	...	...

Image: Anthony D'Onofrio, Northeastern University, 2009

A Microbial Census

Marker-gene and Metagenomic Sequencing



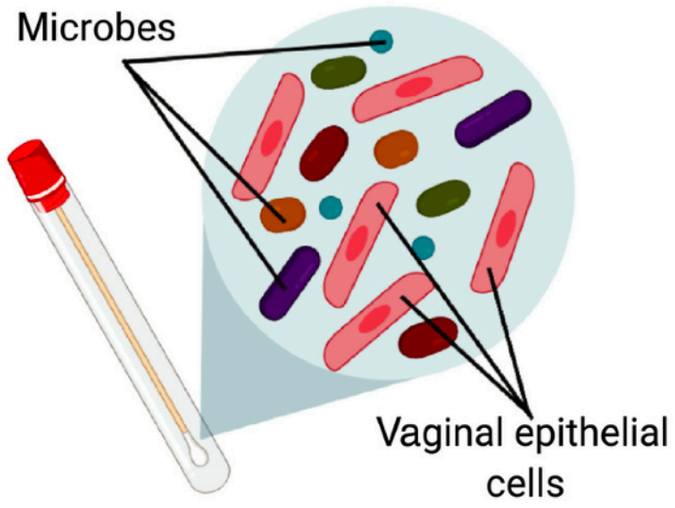
Lactobacillus crispatus	1300	5	0	882	596
Ureaplasma urealytica	15	0	220	0	0
Gardnerella vaginalis	22	0	1	0	412
Prevotella intermedia	0	0	8	12	0
...	...	...	...	...	...

Inference

Visualization

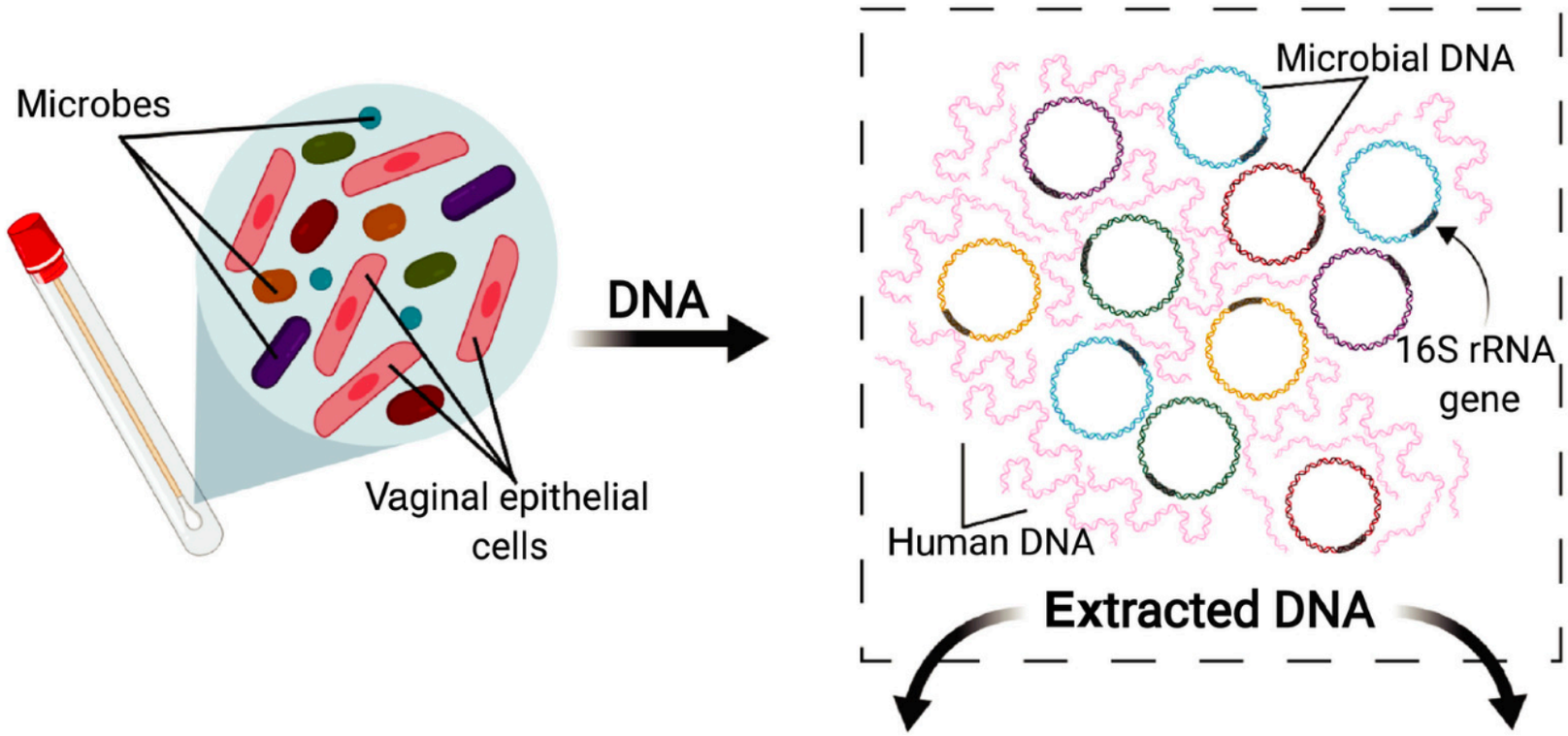
Exploration

Community Sequencing



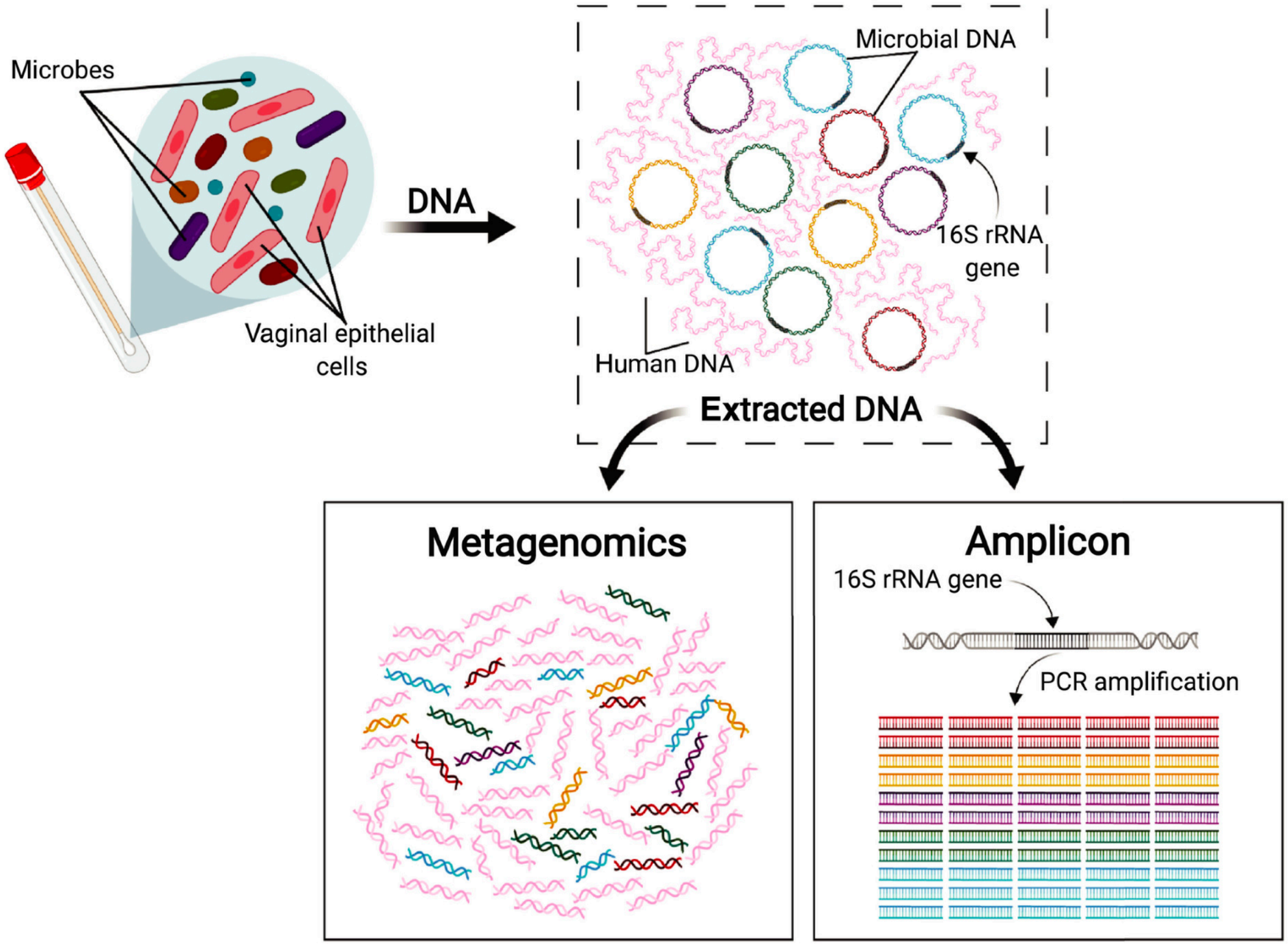
Berman, BJOG 2020.

Community Sequencing



Berman, BJOG 2020.

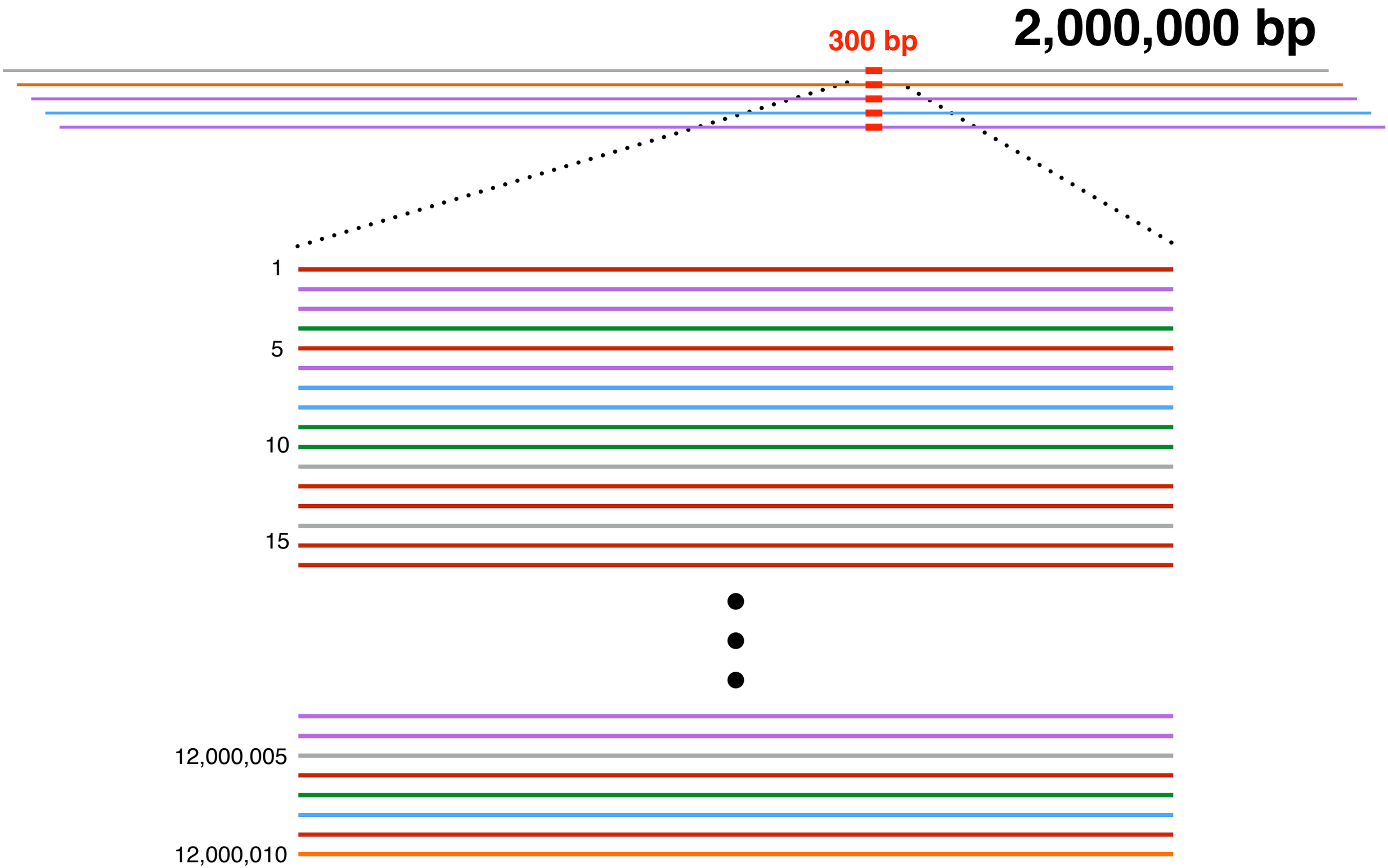
Community Sequencing



Berman, BJOG 2020.

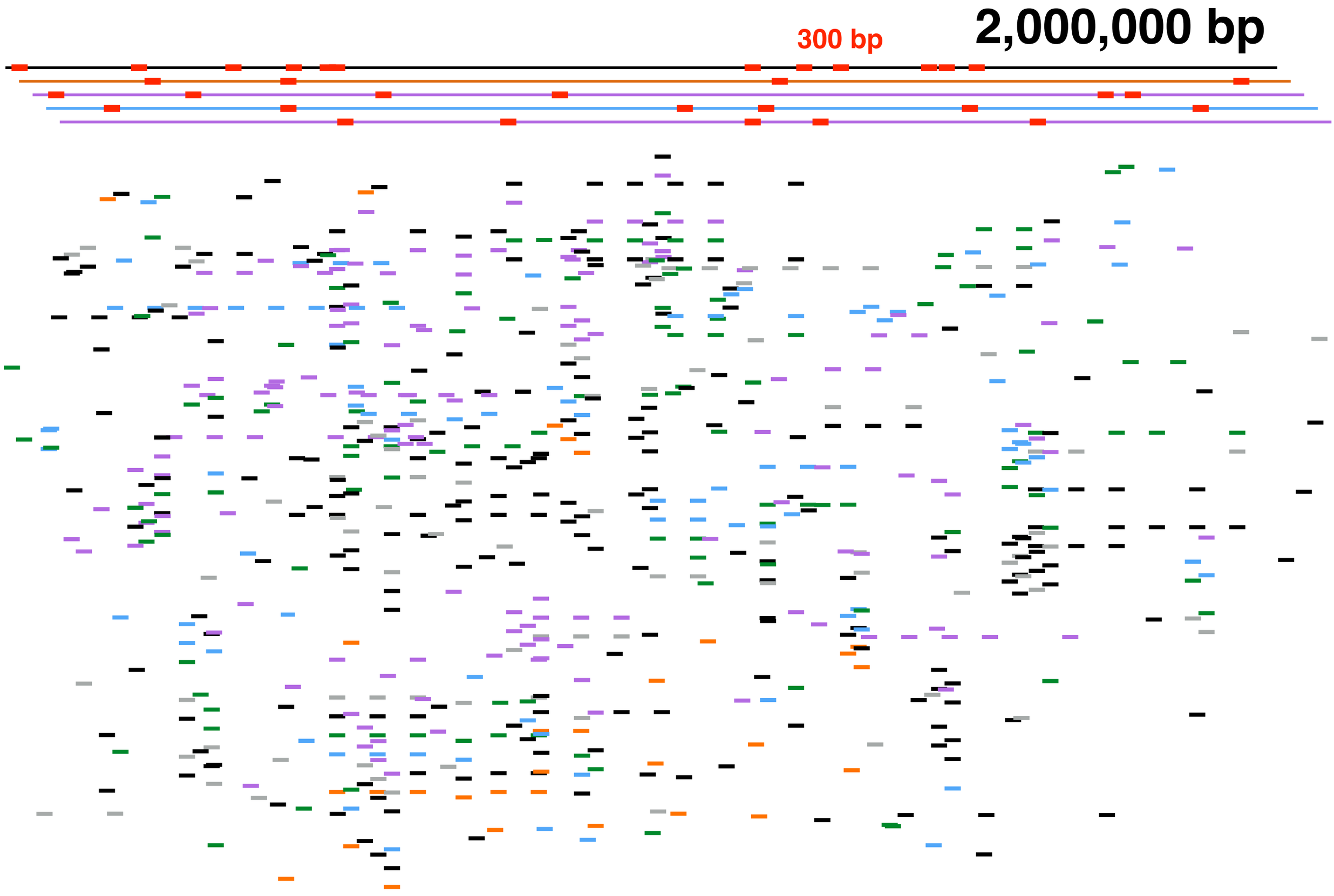
Marker-gene Sequencing

Marker-gene Sequencing

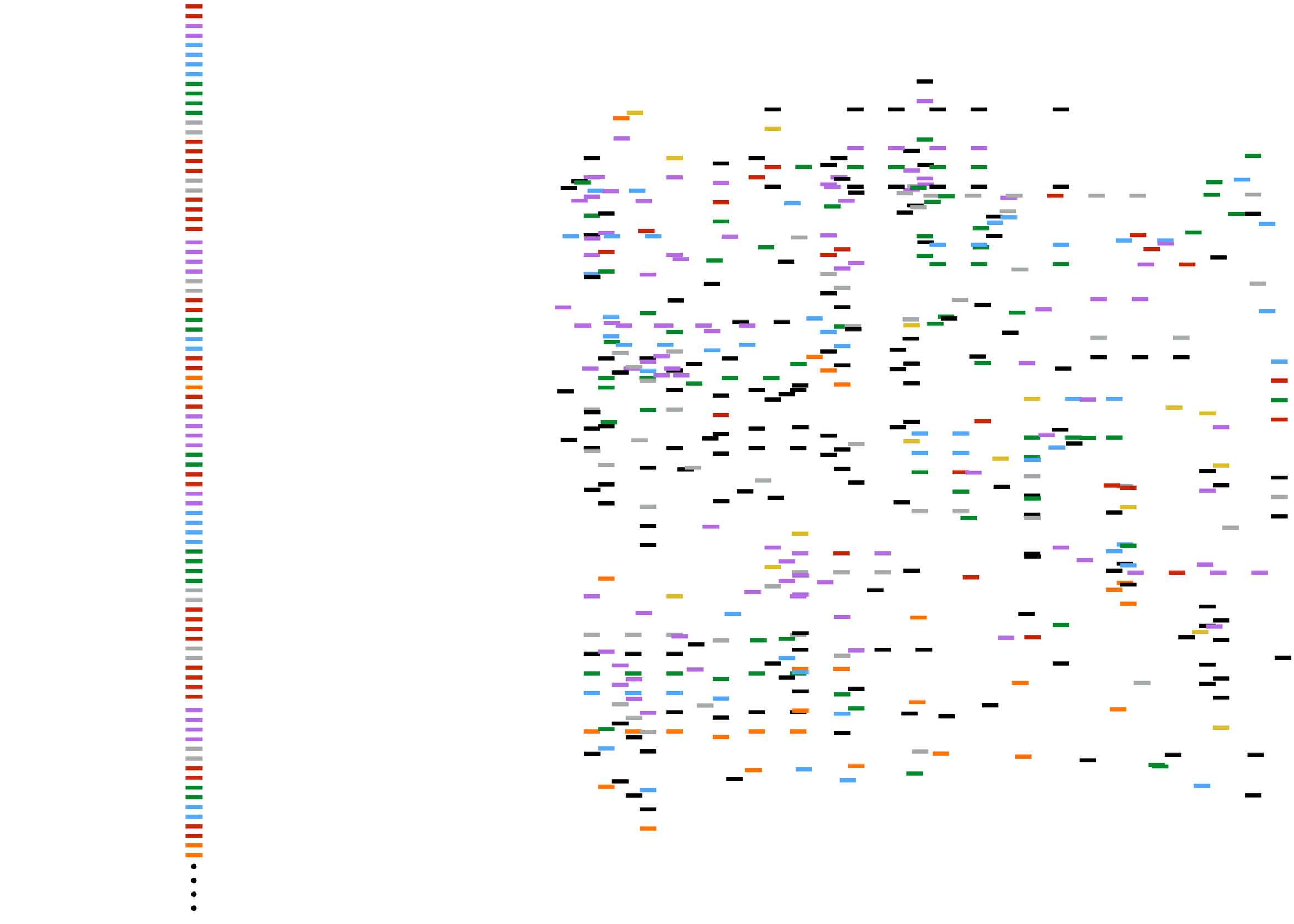


Shotgun Sequencing

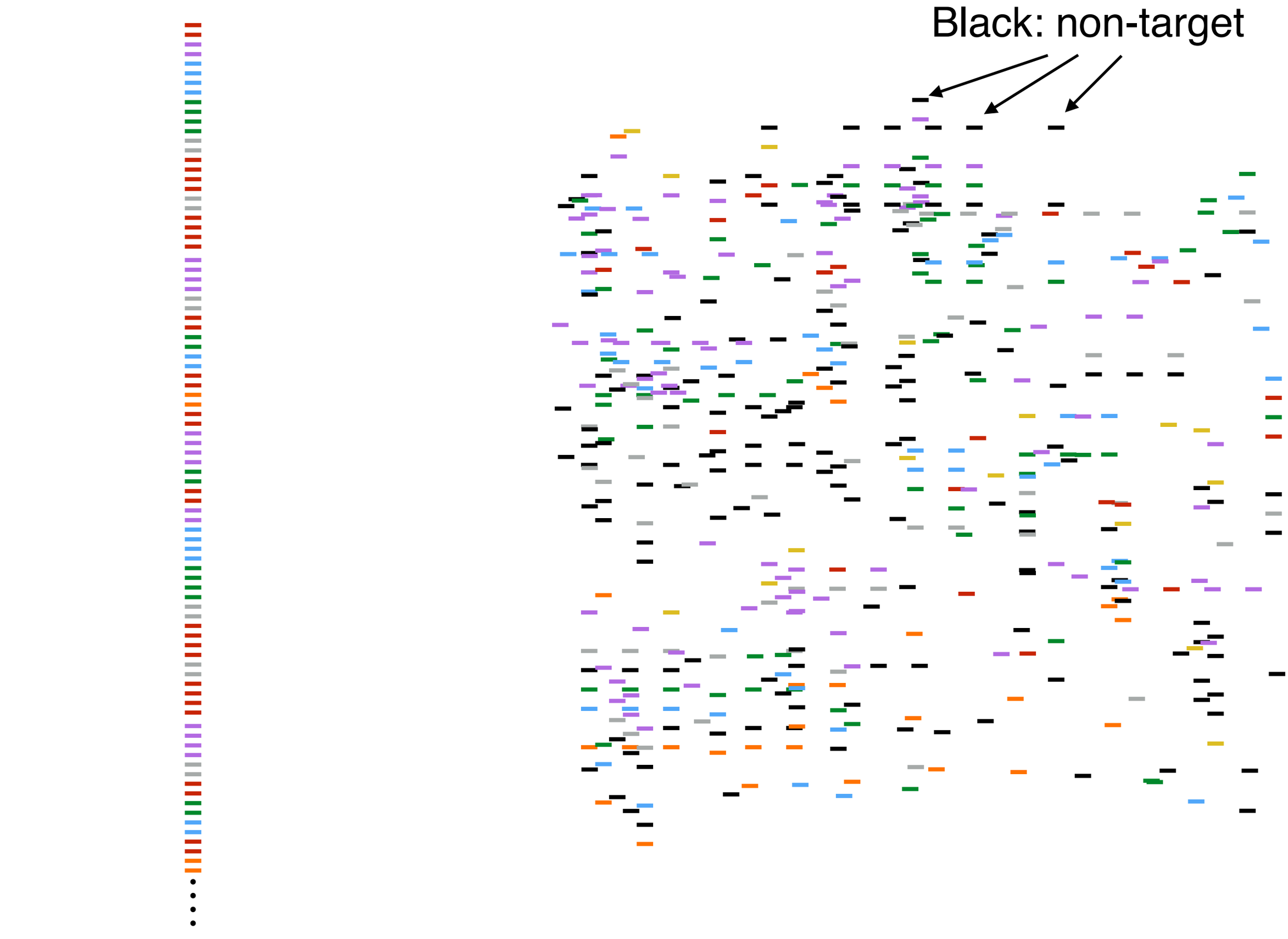
Shotgun Sequencing



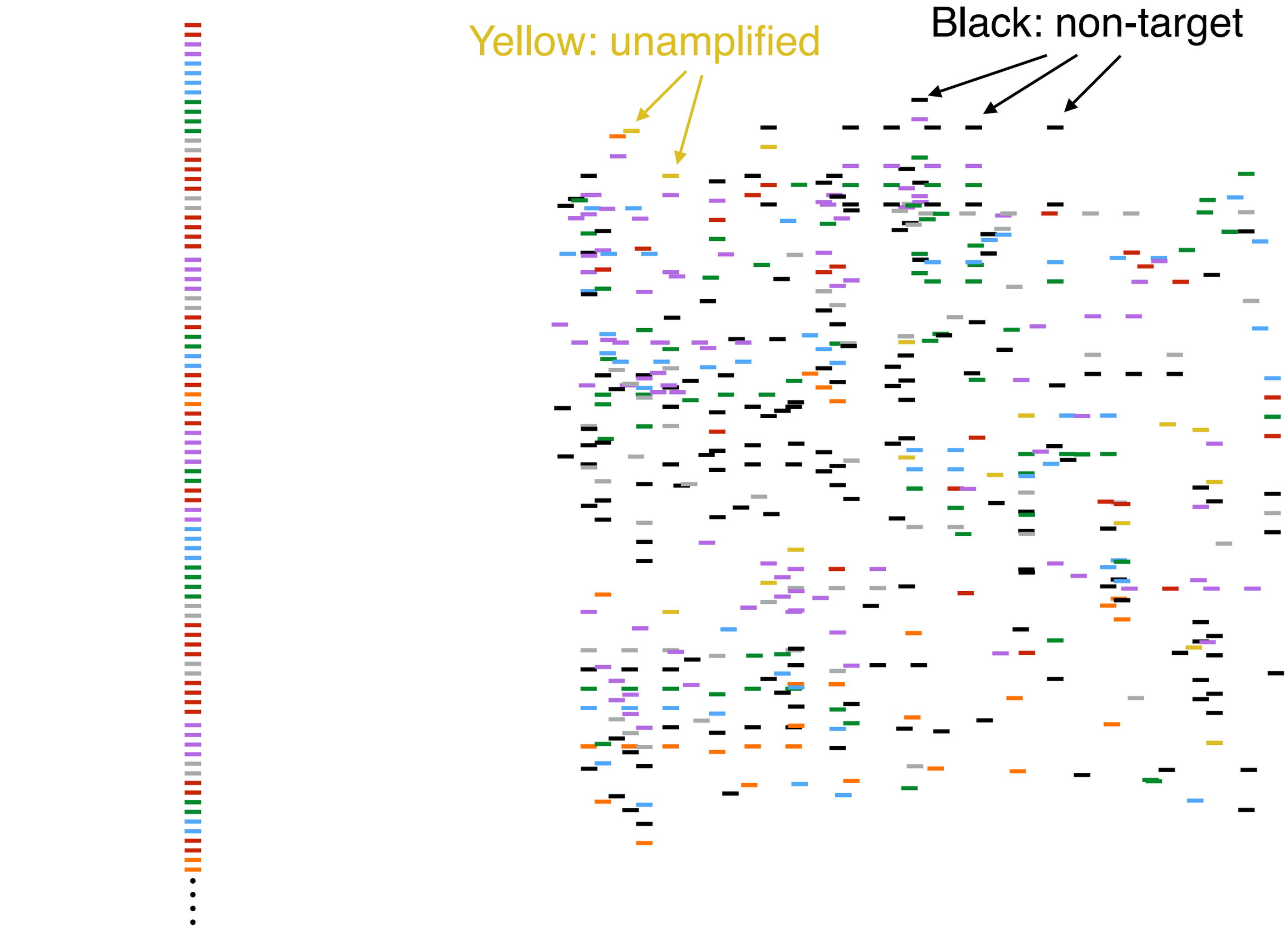
Marker-gene vs. Metagenomics



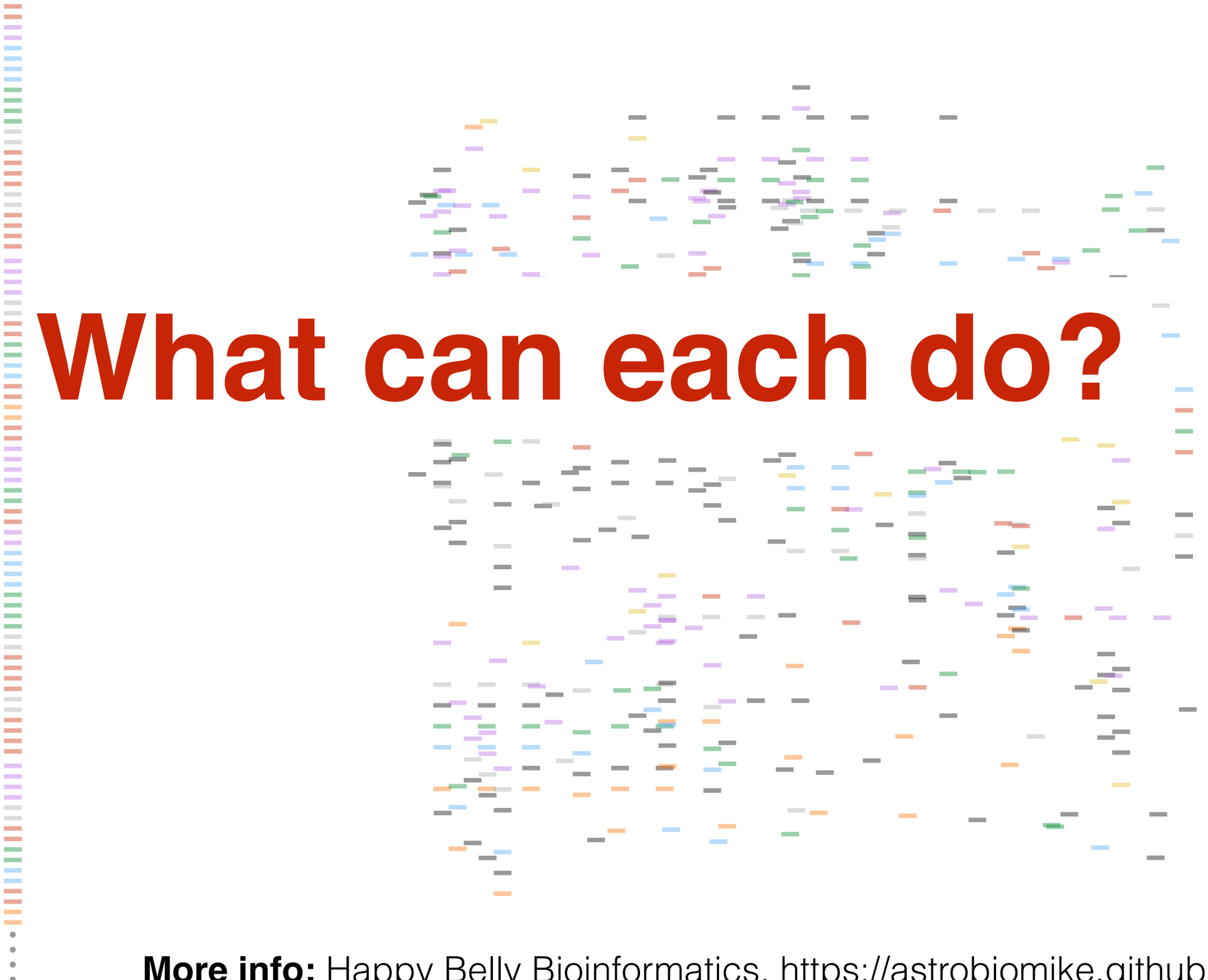
Marker-gene vs. Metagenomics



Marker-gene vs. Metagenomics



Marker-gene vs. Metagenomics



What can each do?

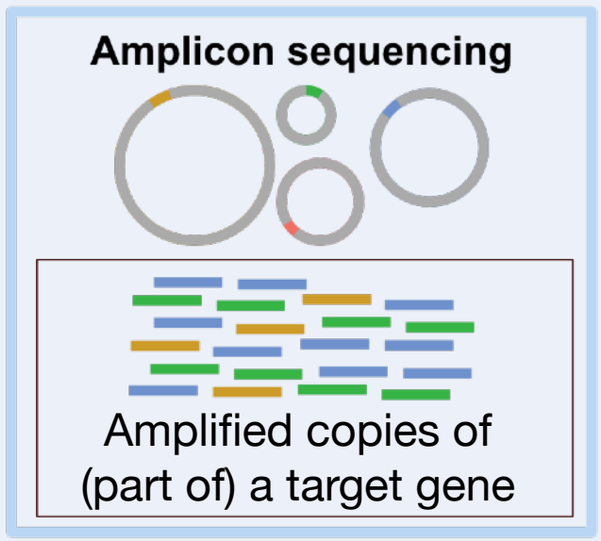
More info: Happy Belly Bioinformatics, https://astrobiomike.github.io

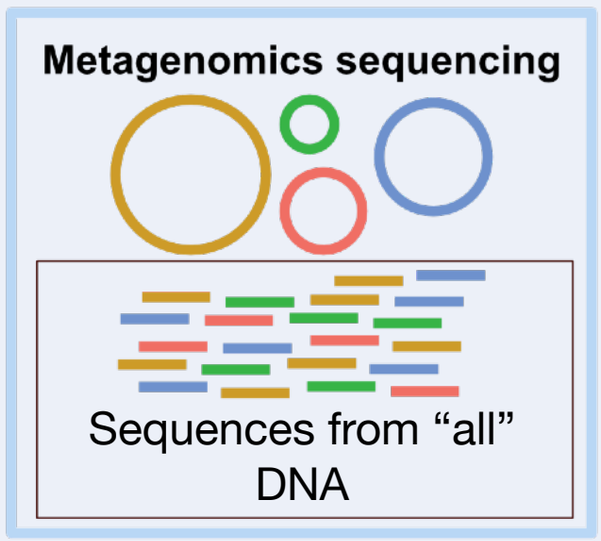
Marker-gene vs. Metagenomics

What can each <u>not</u> do?

More info: Happy Belly Bioinformatics, https://astrobiomike.github.io

Two of the main tools in the microbial ecologist's toolkit

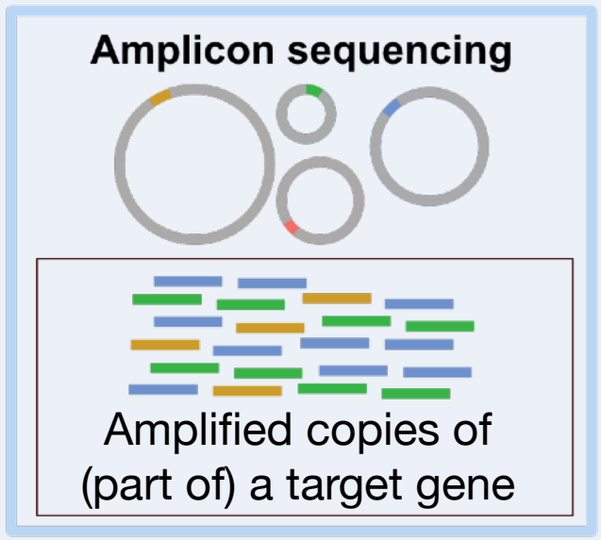




Pros and cons

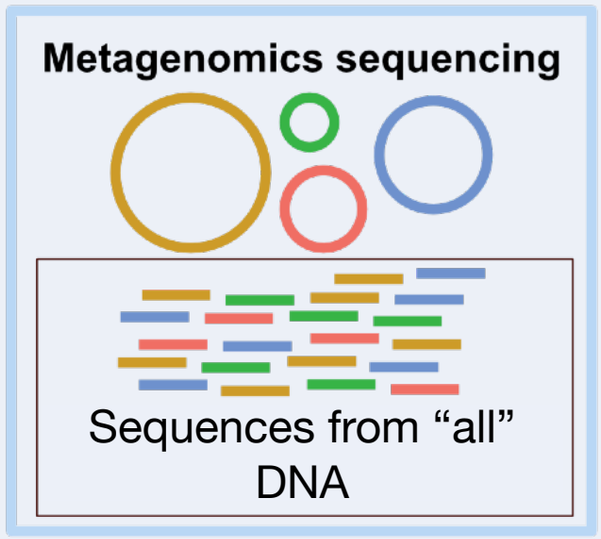
Slide courtesy of: Mike Lee, STAMPS 2024 course faculty

Two of the main tools in the microbial ecologist's toolkit



Pros and cons

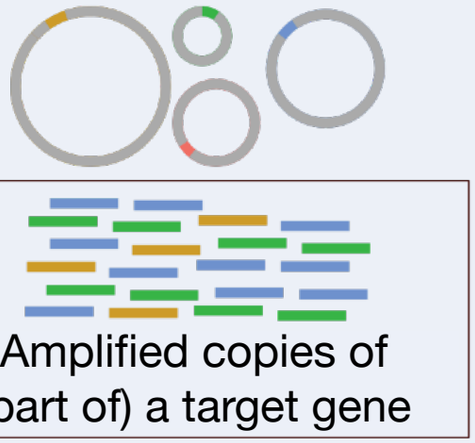
- ⊕ can be more affordable, particularly if there are many samples
- ⊕ Analysis is less computationally intensive, more straightforward
- ⊕ Focuses sequencing effort on the targeted taxonomic group
- ⊖ "universal" primers don't capture all sequences that have the target gene
- ⊖ many ecologically relevant biological entities are not captured by this approach (e.g., viruses and plasmids)
- ⊖ does not directly provide any information on potential function



Slide courtesy of: Mike Lee, STAMPS 2024 course faculty

Two of the main tools in the microbial ecologist's toolkit

Amplicon sequencing



- ⊕ can be more affordable, particularly if there are many samples
- ⊕ Analysis is less computationally intensive, more straightforward
- ⊕ Focuses sequencing effort on the targeted taxonomic group
- ⊖ "universal" primers don't capture all sequences that have the target gene
- ⊖ many ecologically relevant biological entities are not captured by this approach (e.g., viruses and plasmids)
- ⊖ does not directly provide any information on potential function

Metagenomics sequencing

- ⊕ grants access to additional biological entities such as viruses and plasmids
- ⊕ provides information on the functional potential of the microbial community
- ⊕ can enable the recovery of metagenome-assembled genomes (MAGs)
- ⊖ can be more expensive, especially in the presence of large amounts of non-target DNA
- ⊖ requires greater computational resources to process and can require more expertise to analyze and interpret

Pros and cons

Slide courtesy of: Mike Lee, STAMPS 2024 course faculty

What is the right tool for your question?

An (incomplete) sequencing survey

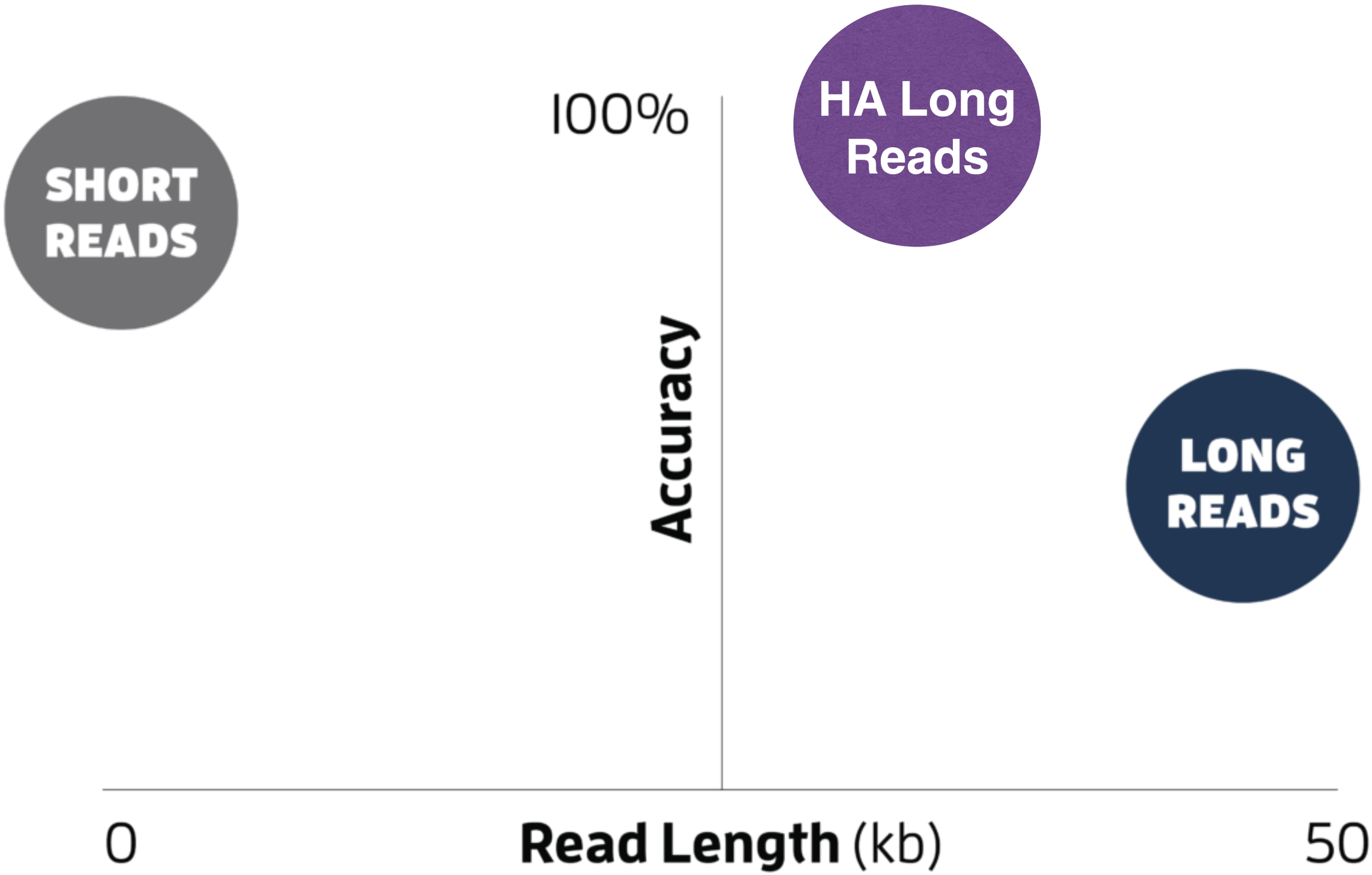


Image modified from: pacb.com

Illumina Short-read sequencing

Read length: **100 - 300 nts**, Per-base error-rate: **0.1 - 0.5%**

Illumina Short-read sequencing

Read length: **100 - 300 nts**, Per-base error-rate: **0.1 - 0.5%**

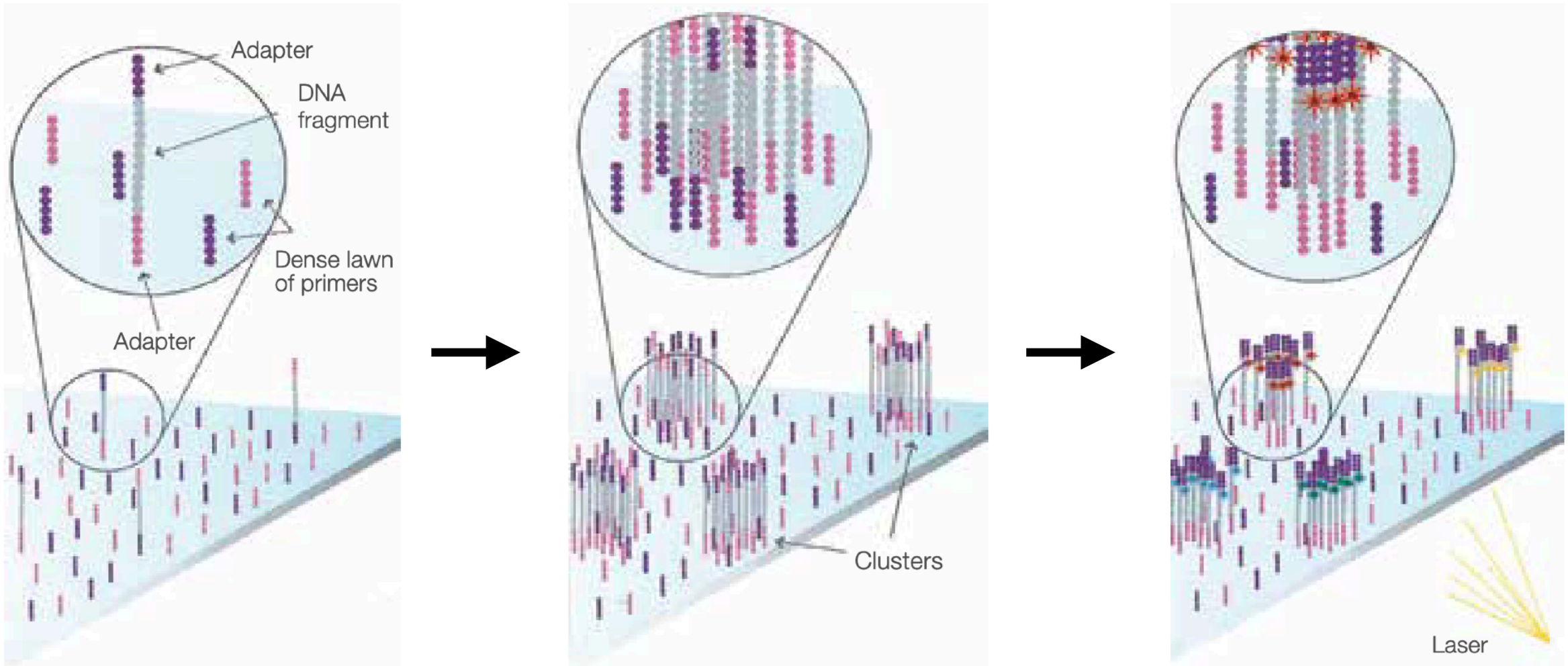


Image: Illumina

Illumina Short-read sequencing

MiniSeq System

MiSeq Series

NextSeq Series

HiSeq Series

HiSeq X Series

NovaSeq Series

Image: Illumina

Illumina Short-read sequencing

Increasing throughput, increasing batch sizes.

Decreasing cost per-base.

MiniSeq System

MiSeq Series

NextSeq Series

HiSeq Series

HiSeq X Series

NovaSeq Series

Image: Illumina

Illumina Short-read sequencing

Increasing throughput, increasing batch sizes.

Decreasing cost per-base.

MiniSeq System

2 color

MiSeq Series

4 color

NextSeq Series

2 color

HiSeq Series

4 color

HiSeq X Series

4 color(?)

NovaSeq Series

2 color

Image: Illumina

Illumina-specific Error Modes

Illumina-specific Error Modes

Declining accuracy towards end of reads:

CAAGTAAGACCTAGACCTAGGAGTAATC**C**AGT**AC**GC**A**G**GT**A

Errors

Illumina-specific Error Modes

Declining accuracy towards end of reads:

CAAGTAAGACCTAGACCTAGGAGTAATC**C**AGT**AC**GC**A**G**GT**A

Errors

Read-through into adapter sequences:

CAAGTAAGACCTAGACCTAGGA**CTGTCTCTTATACACATCT**

Adapter

Illumina-specific Error Modes

Declining accuracy towards end of reads:

CAAGTAAGACCTAGACCTAGGAGTAATC**C**AGT**AC**GC**A**G**GT**A

Errors

Read-through into adapter sequences:

CAAGTAAGACCTAGACCTAGGA**CTGTCTCTTATACACATCT**

Adapter

polyG tails in 2-color chemistries:

CAAGTAAGACCTAGACCT**GGGGGGGGGGGGGGGGGGGGGGGGG**

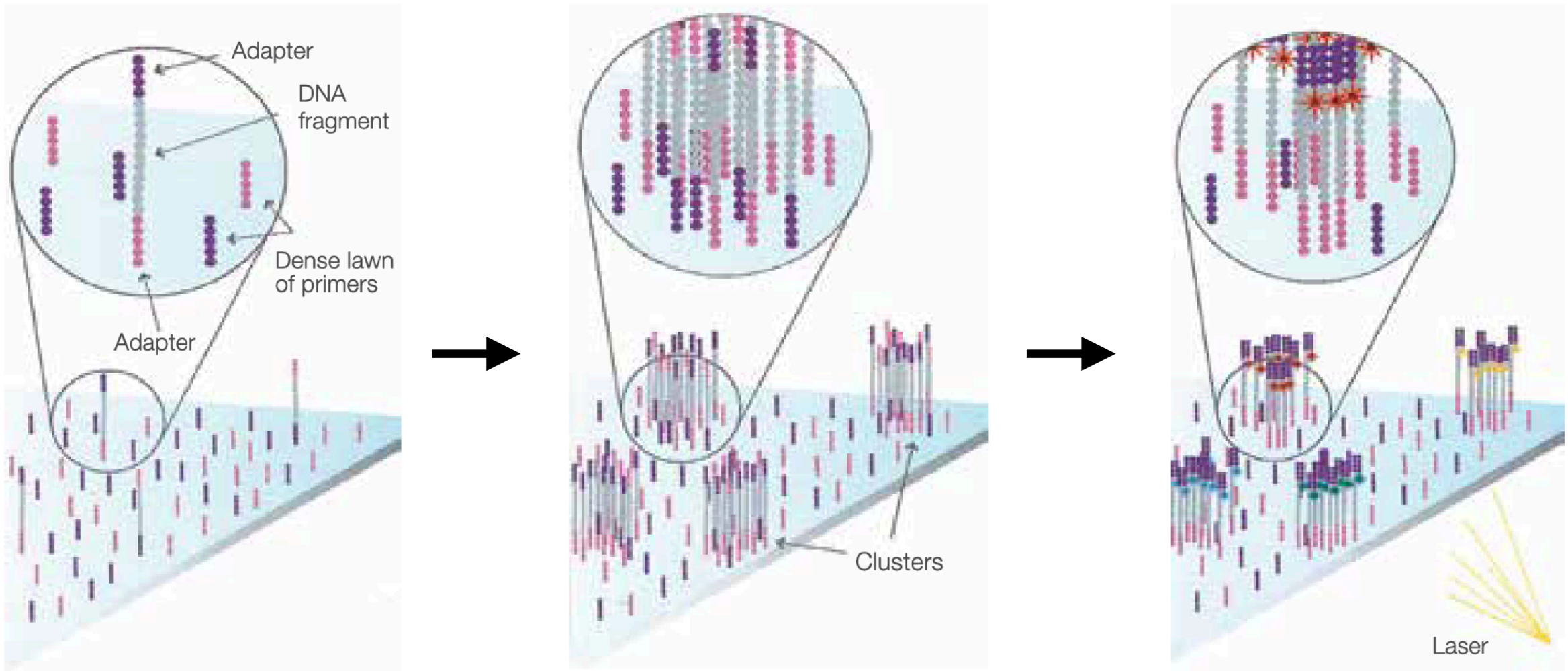
polyGs

Illumina-specific Error Modes

Declining accuracy towards end of reads: dephasing.

Read-through into adapter sequences: see below.

polyG tails in 2-color chemistries: G = no signal.



PacBio HiFi sequencing

Read length: **1 - 50 kbases,** Per-base error-rate: **< 0.1%**

PacBio HiFi sequencing

Read length: **1 - 50 kbases,** Per-base error-rate: **< 0.1%**

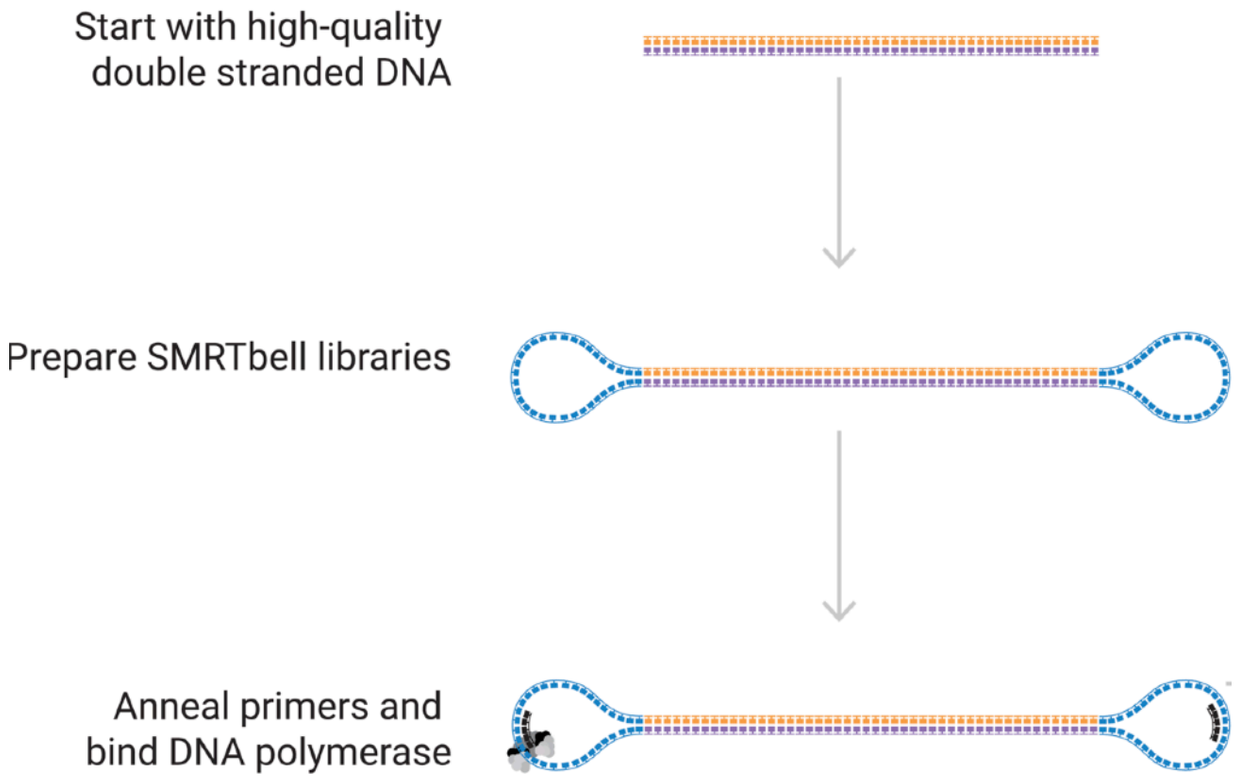


Image: Pacbio

PacBio HiFi sequencing

Read length: **1 - 50 kbases,** Per-base error-rate: **< 0.1%**

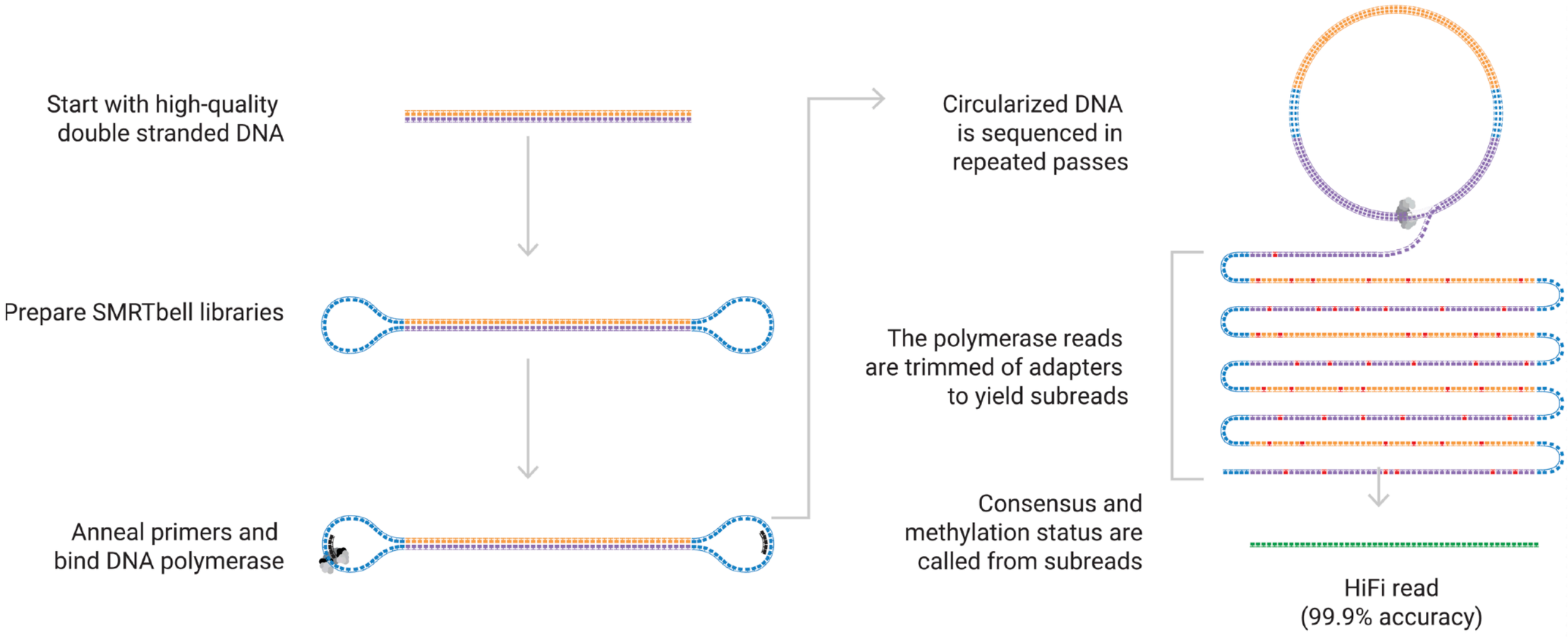


Image: Pacbio

Pacbio HiFi specific Error Modes

Pacbio HiFi specific Error Modes

None.

Pacbio HiFi specific Error Modes

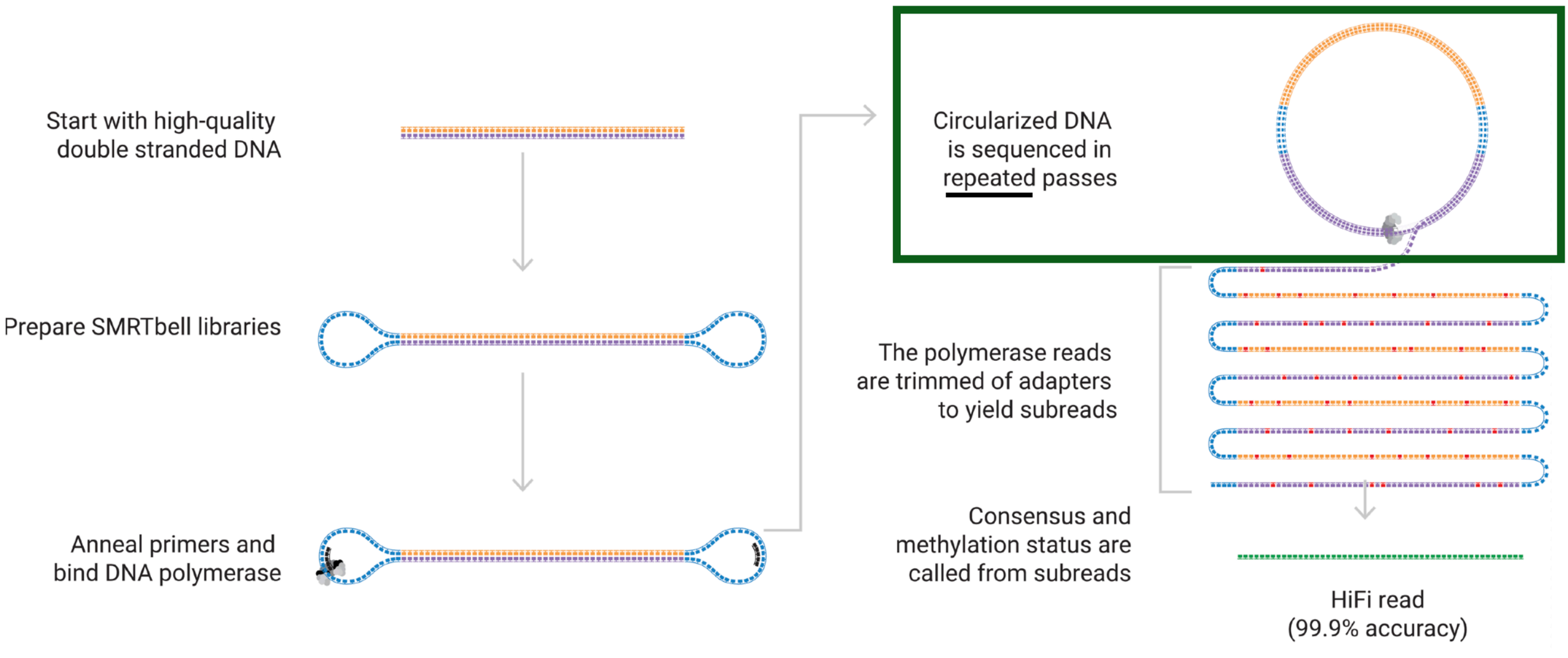
None.

*** that the speaker has been able to identify**

Pacbio HiFi specific Error Modes

None.

$$: Higher per-base costs.



Nanopore Long-read sequencing

Read length: **up to 100s of kb,** Per-base error-rate: **2-10%**

Nanopore Long-read sequencing

Read length: **up to 100s of kb,** Per-base error-rate: **2-10%**

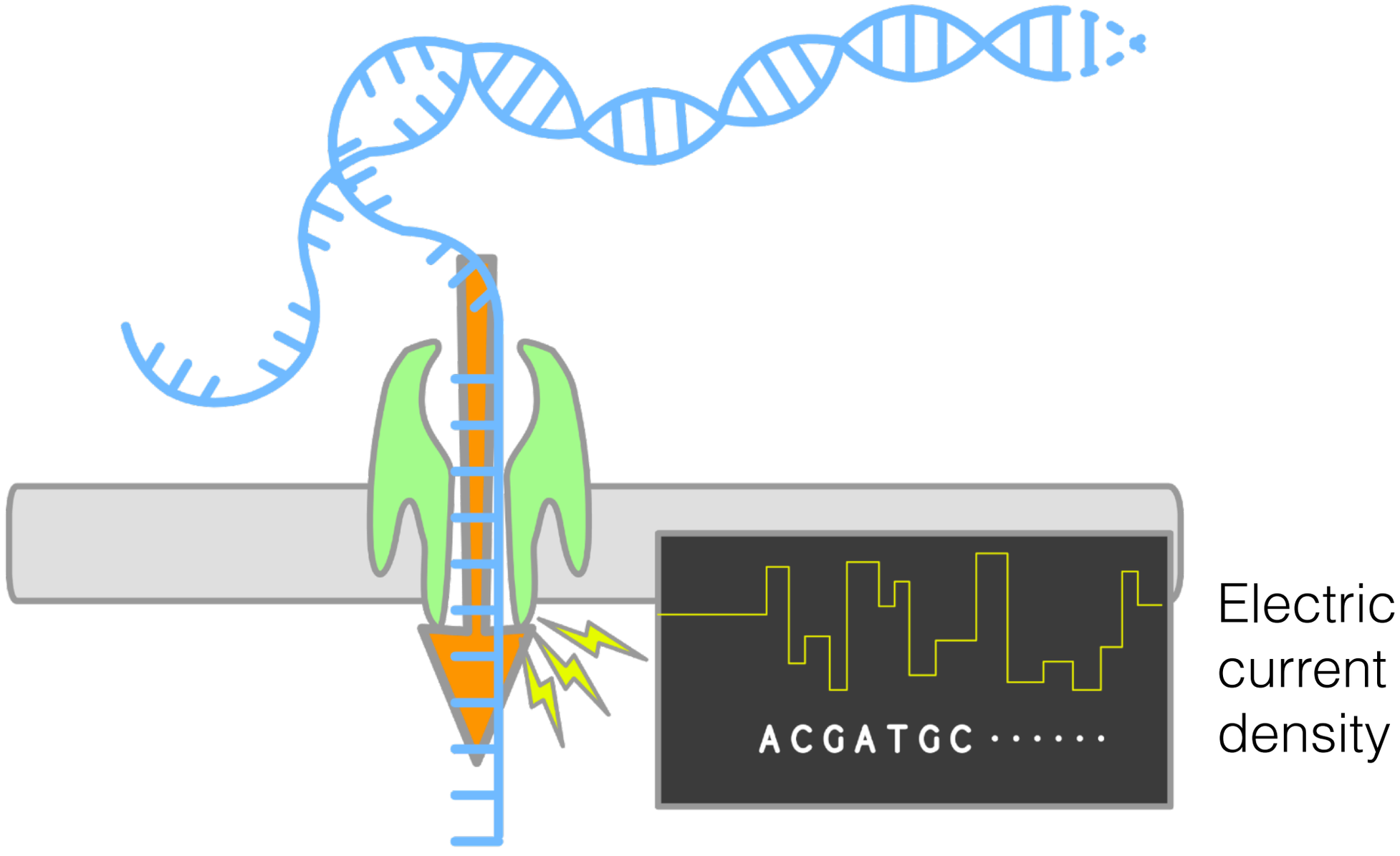
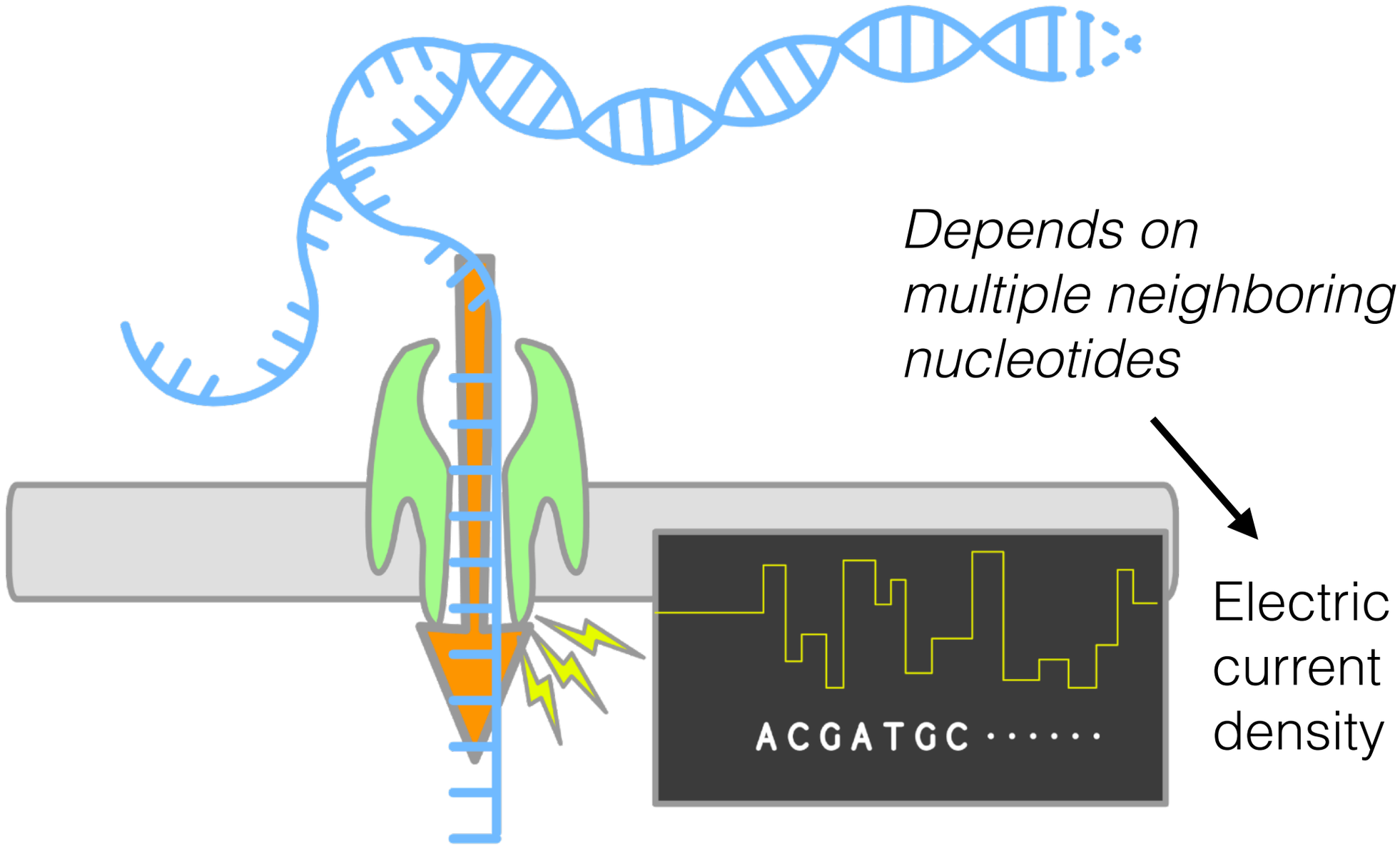


Image: Wikipedia

Nanopore Long-read sequencing

Read length: **up to 100s of kb,** Per-base error-rate: **2-10%**



Depends on multiple neighboring nucleotides

Electric current density

Image: Wikipedia

Oxford Nanopore

MinION

GridION

PromethION

Oxford Nanopore

Increasing throughput, increasing batch sizes.

Decreasing cost per-base.

MinION

GridION

PromethION

Oxford Nanopore

Increasing throughput, increasing batch sizes.

Decreasing cost per-base.

MinION

GridION

PromethION

Highly portable,
Fits in hand.

Nanopore specific Error Modes

Homopolymers:

CAAGTAAGACCTAGACCTAGGA**CCCCCCCCCCCCCCCCC**TTATA

Incorrect length

Nanopore specific Error Modes

Homopolymers:

CAAGTAAGACCTAGACCTAGGA**CCCCCCCCCCCCCCCCC**TTATA

Incorrect length

Indels:

CAAGTAAGACCT**T**AGACCTAGGAGTAATCG**C**AGT-GCAGGTA

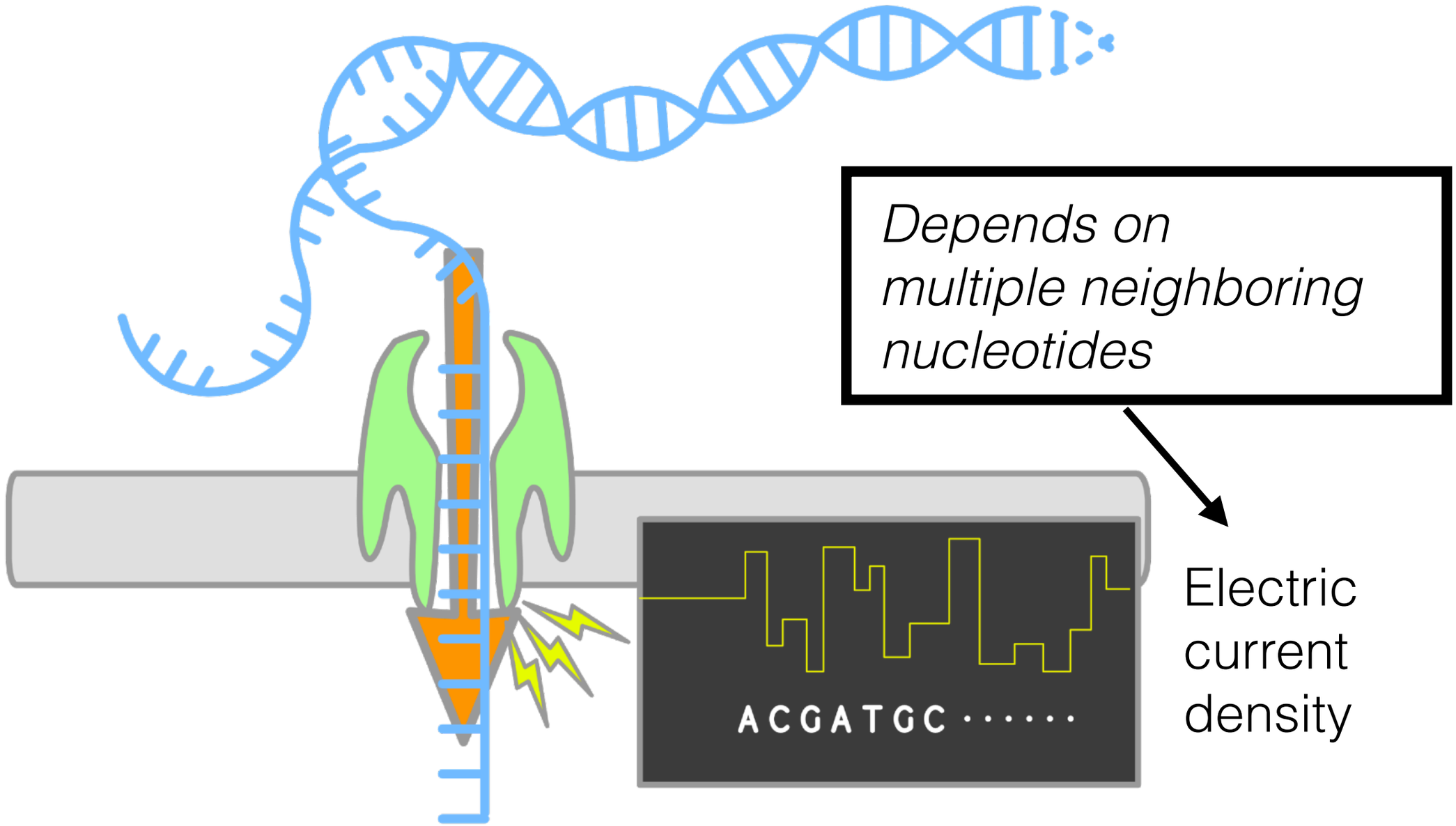
Insertions

Deletion

Nanopore specific Error Modes

Homopolymers: Signal not 1-1 with nucleotide, see below.

Indels: Signal not 1-1 with nucleotide, see below.

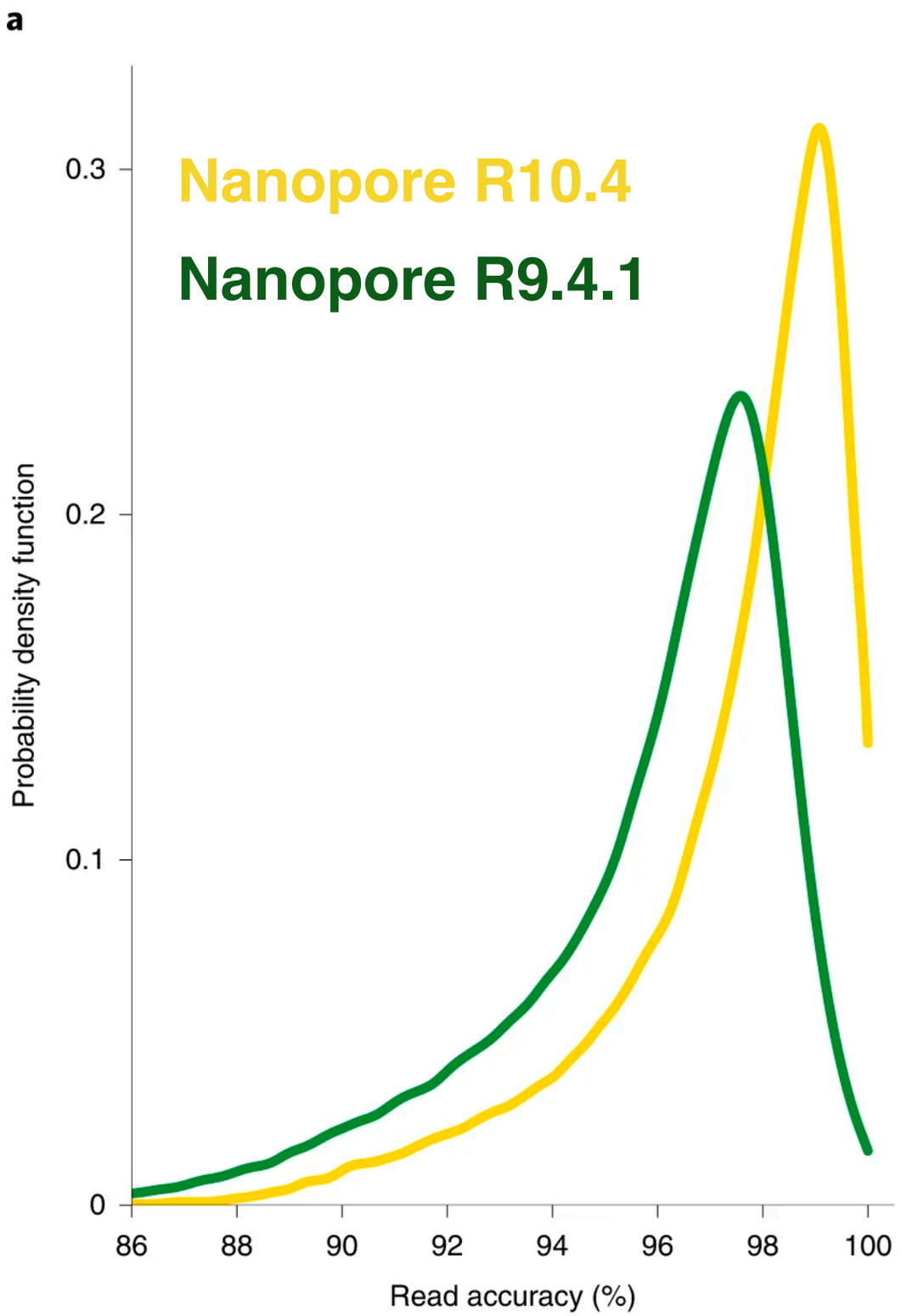


Depends on multiple neighboring nucleotides

Electric current density

Image: Wikipedia

Improving ONT Error Rates



Base-calling and chemistry has substantially improved. Error rates are down to ~2% in the latest versions.

Sereika, Nature Methods, 2022.

An (incomplete) sequencing survey

SHORT READS

Expanding short read options!

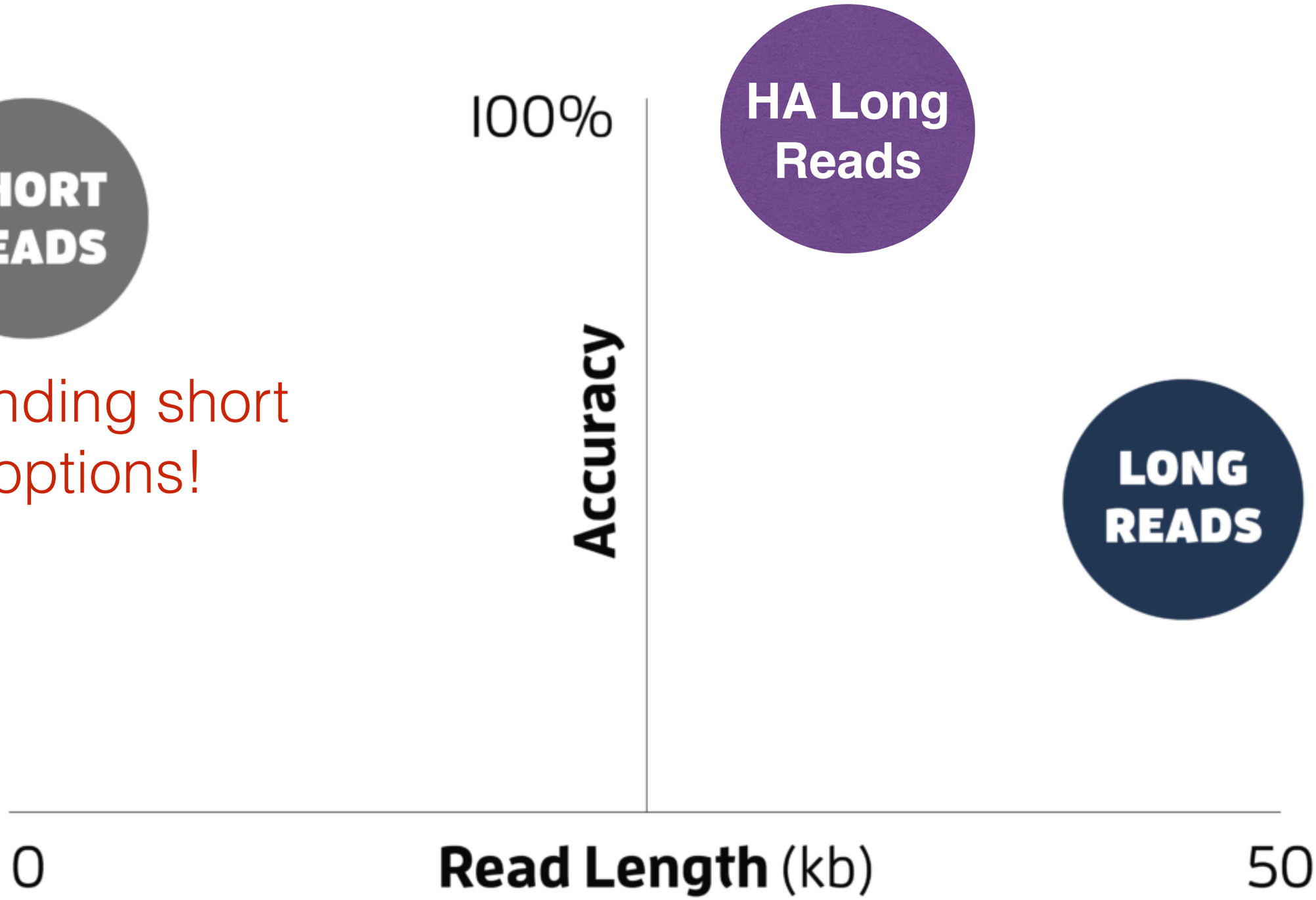


Image modified from: pacb.com

What is the right tool for your question?

Fastq files and Quality scores

Image: Robert Edgar, drive5.com

Fastq files and Quality scores

$$Q = -10 \log_{10} P \quad \Longrightarrow \quad P = 10^{\frac{-Q}{10}}$$

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 10000	99.99%
50	1 in 100000	99.999%

Image: NYU Center For Genomics and Systems Biology

Fastq files and Quality scores

$$Q = -10 \log_{10} P \quad \Longrightarrow \quad P = 10^{\frac{-Q}{10}}$$

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 10000	99.99%
50	1 in 100000	99.999%

Q is encoded as ASCII characters:

```
(33):  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
```

Q=0 ⟶ Q=40

Image: NYU Center For Genomics and Systems Biology

Fastq files and Quality scores

$$Q = -10 \log_{10} P \quad \Longrightarrow \quad P = 10^{\frac{-Q}{10}}$$

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 10000	99.99%
50	1 in 100000	99.999%

Q is encoded as ASCII characters:

```
(33):  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
```

Q=0 → Q=40

If it looks like a swear word — #$!!%& — it's bad quality!

Image: NYU Center For Genomics and Systems Biology

Binned Quality scores

RTA2

Sequence data

CAGAACCTGACCCGAACCTGACC
TTGGCATTCCATTGGCATTTCCA
TAGCATCATGGATTAGCATCATGGAT
GAGTCAACATCAGAGTCAACAGTCA

▼

Q-table

Metric 1	Metric 2	Metric 3	Metric 4	Metric 5
0	1	3	3.2	0
21	74	2	2.2	0
32	85	2	2.2	0
46	99	2	2.2	0
49	102	2	2.2	0
52	105	2	2.2	0
60	113	2	2.2	0
78	131	1	1.2	0
89	142	1	1.2	0
100	153	1	1.2	0

▼

Q-scores

8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81

RTA3

Sequence data

CAGAACCTGACCCGAACCTGACC
TTGGCATTCCATTGGCATTTCCA
TAGCATCATGGATTAGCATCATGGAT
GAGTCAACATCAGAGTCAACAGTCA

▼

Q-table

Metric 1	Metric 2	Metric 3	Metric 4	Metric 5
0	1	3	3.2	0
862	915	0.5	0.9	0
2125	2178	0.05	0.06	1
3256	3309	0.05	0.07	1

▼

Q-scores

2 | 12 | 23 | 37

Image: Illumina

Binned Quality scores

RTA2

Sequence data

CAGAACCTGACCCGAACCTGACC
TTGGCATTCCATTGGCATTTCCA
TAGCATCATGGATTAGCATCATGGAT
GAGTCAACATCAGAGTCAACAGTCA

▼

Q-table

Metric 1	Metric 2	Metric 3	Metric 4	Metric 5
0	1	3	3.2	0
21	74	2	2.2	0
32	85	2	2.2	0
46	99	2	2.2	0
49	102	2	2.2	0
52	105	2	2.2	0
60	113	2	2.2	0
78	131	1	1.2	0
89	142	1	1.2	0
100	153	1	1.2	0

▼

Q-scores

8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81

RTA3

Sequence data

CAGAACCTGACCCGAACCTGACC
TTGGCATTCCATTGGCATTTCCA
TAGCATCATGGATTAGCATCATGGAT
GAGTCAACATCAGAGTCAACAGTCA

▼

Q-table

Metric 1	Metric 2	Metric 3	Metric 4	Metric 5
0	1	3	3.2	0
862	915	0.5	0.9	0
2125	2178	0.05	0.06	1
3256	3309	0.05	0.07	1

▼

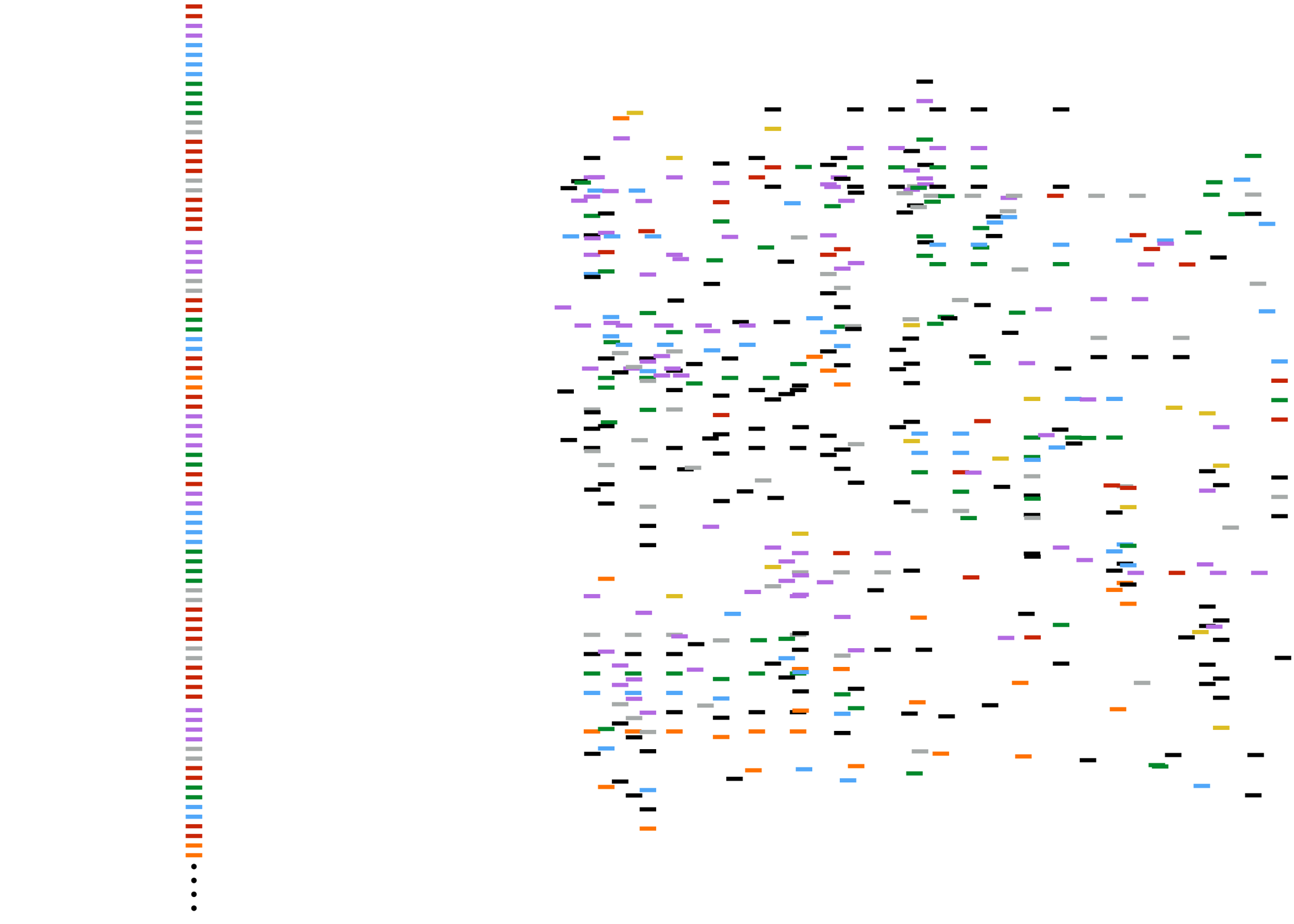
Q-scores

2 | 12 | 23 | 37

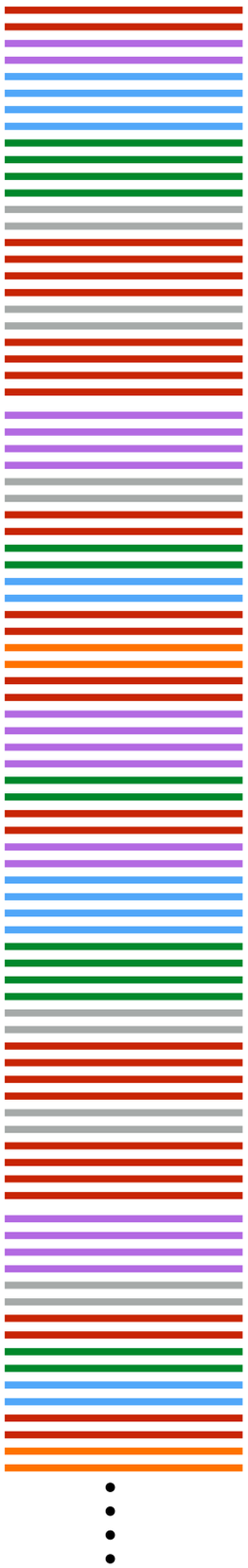
High throughput sequencers (NovaSeq, Revio, ...) increasingly use binned quality scores to save space.

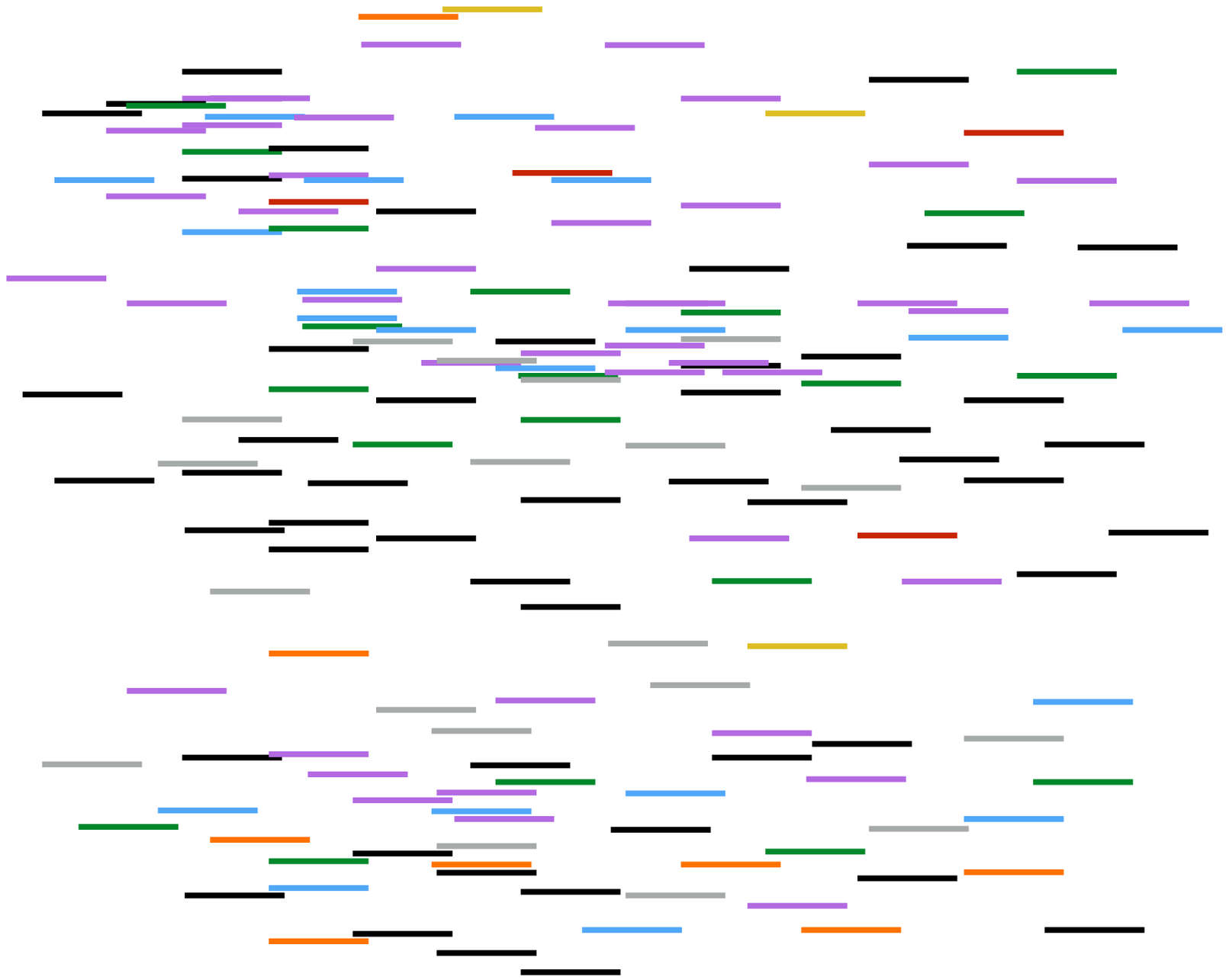
Image: Illumina NovaSeq and RTA3 white paper

Marker-gene vs. Metagenomics

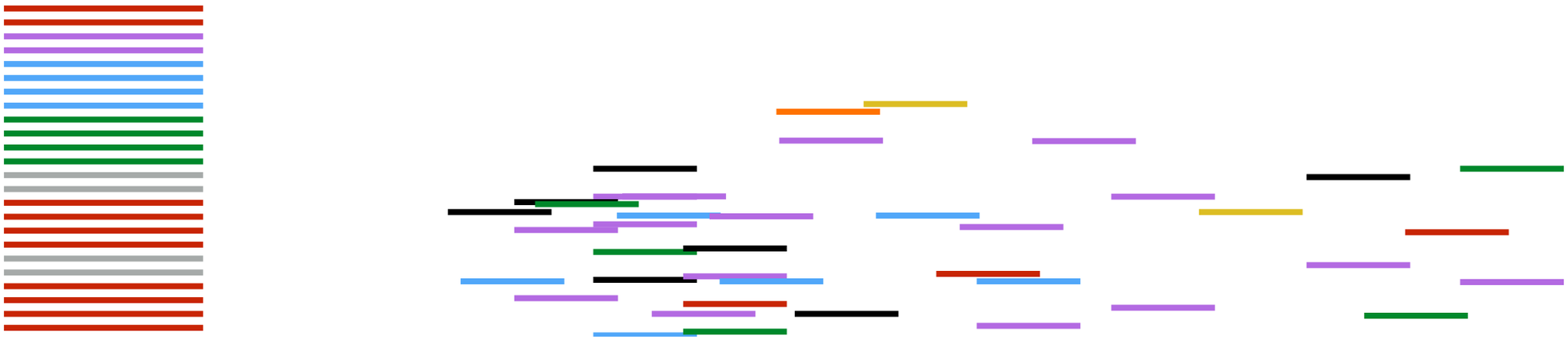


Marker-gene vs. Metagenomics





Marker-gene vs. Metagenomics



How is each affected by reads 10x as long?