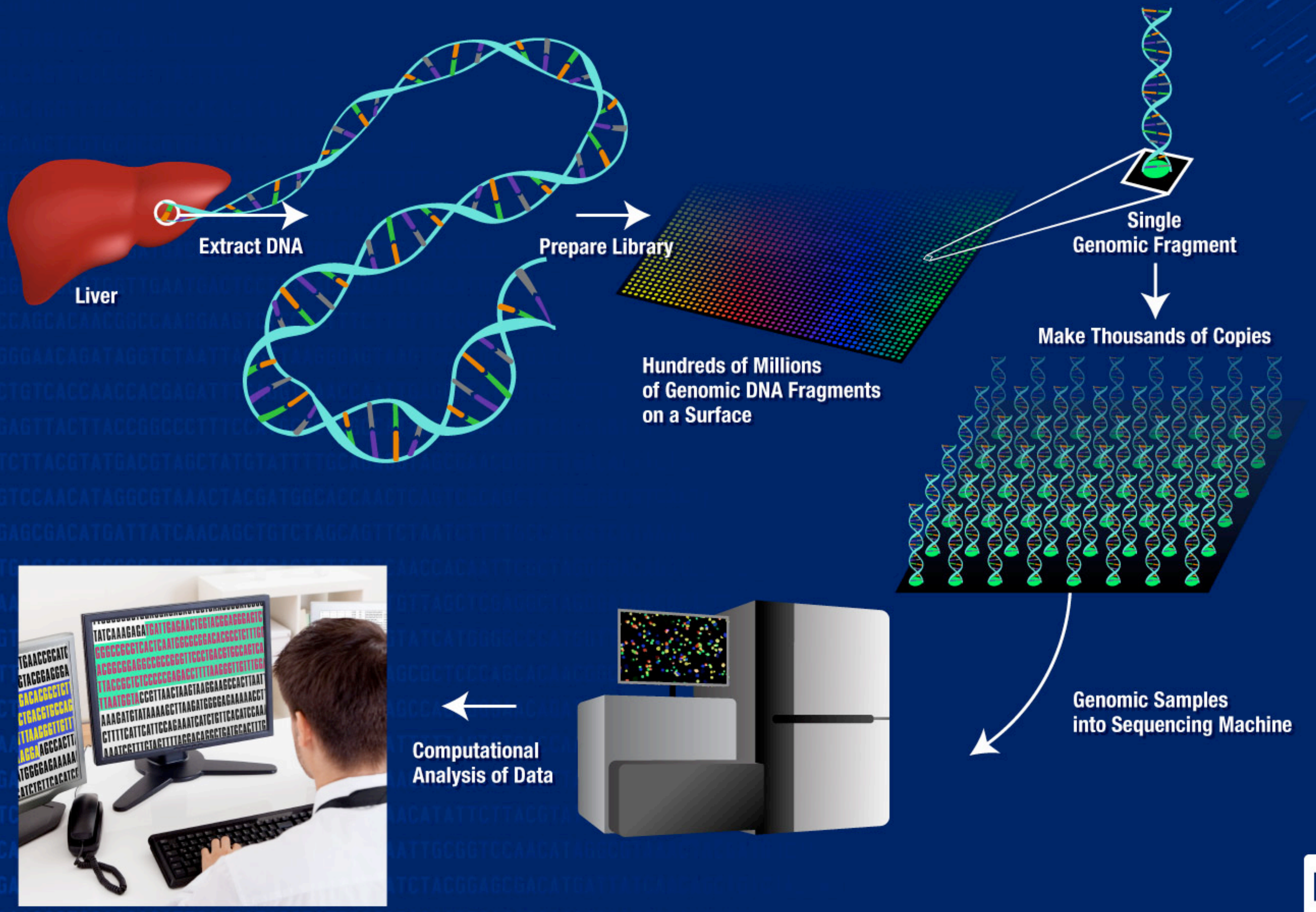
Sequencing Types and Sequence Quality

STAMPS 2022

Dna Sequencing

NHGRI FACT SHEETS
genome.gov



Liver

Extract DNA

Prepare Library

Hundreds of Millions of Genomic DNA Fragments on a Surface

Single Genomic Fragment

Make Thousands of Copies

Genomic Samples into Sequencing Machine

Computational Analysis of Data

NIH National Human Genome Research Institute

An (incomplete) sequencing survey

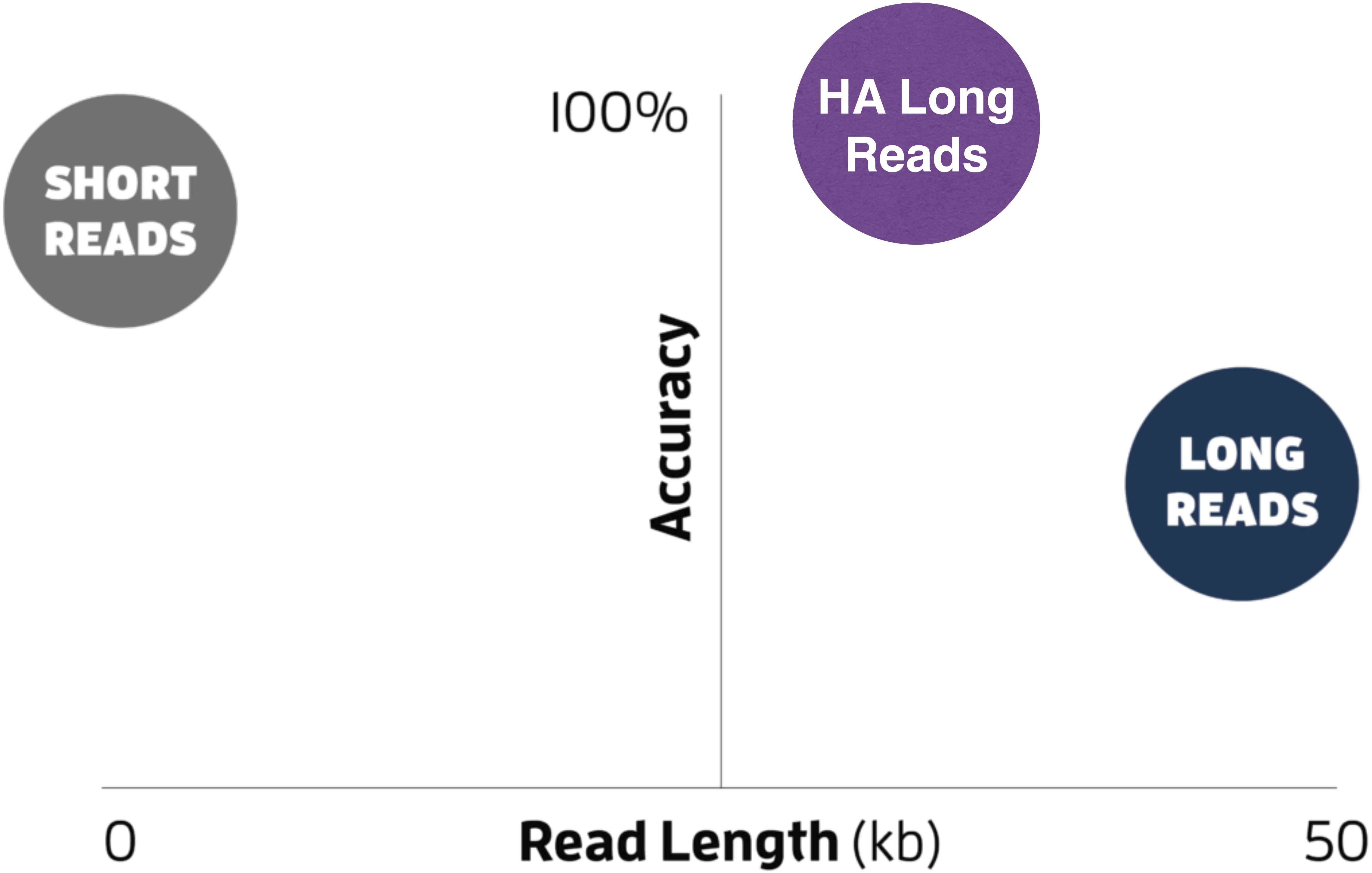


Image modified from: pacb.com

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%**

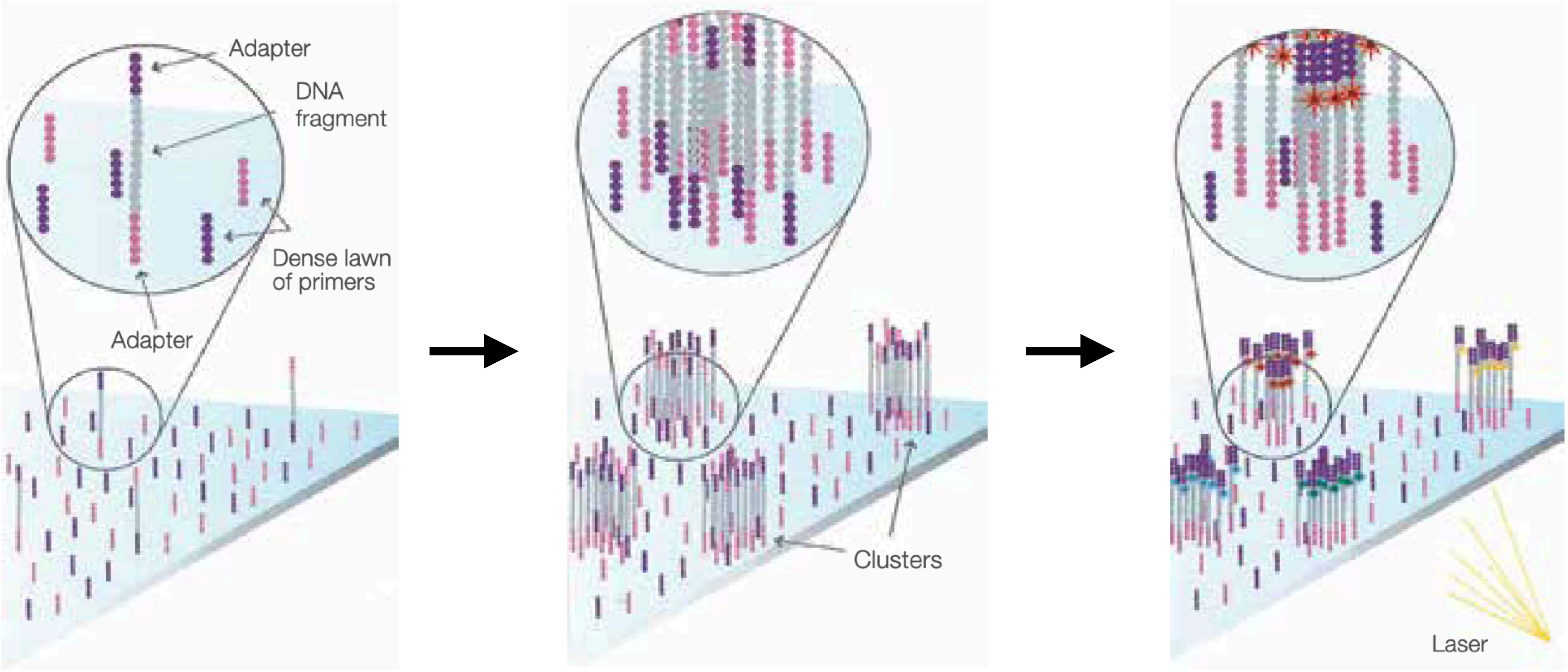
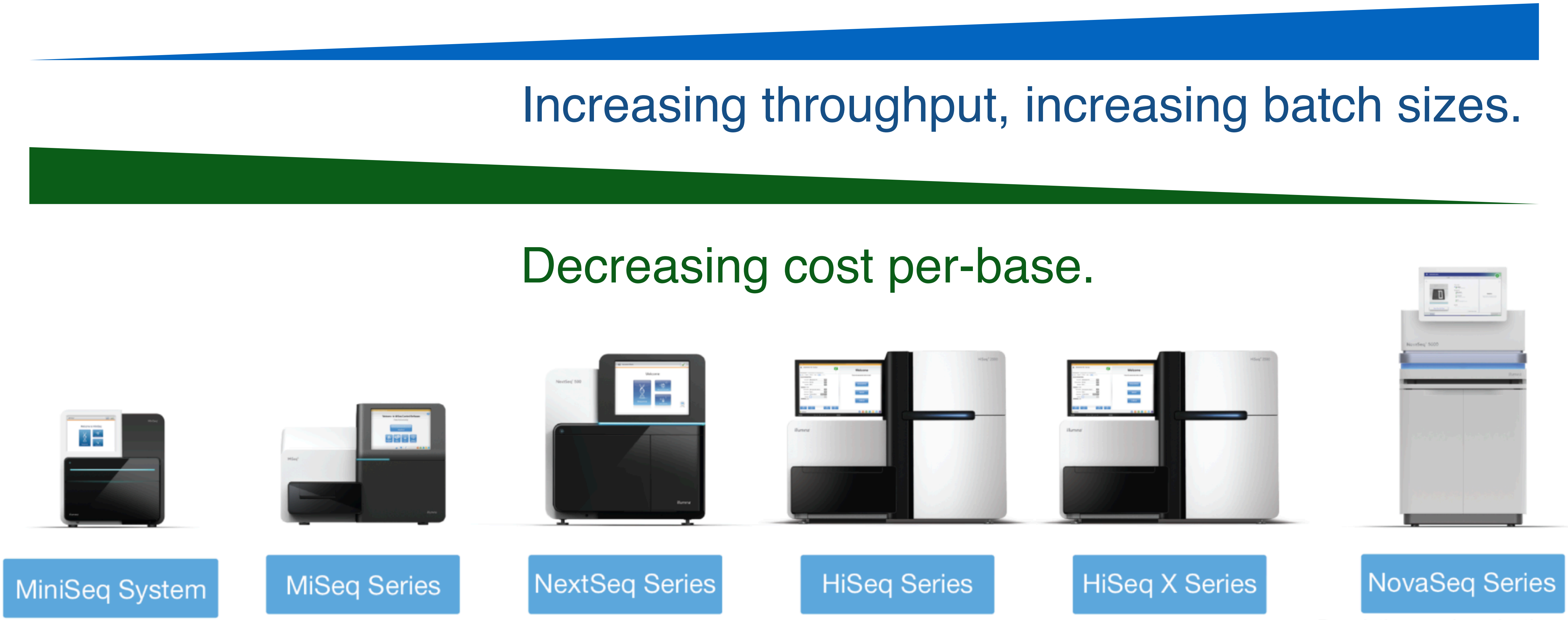


Image: Illumina

Illumina Short-read sequencing

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

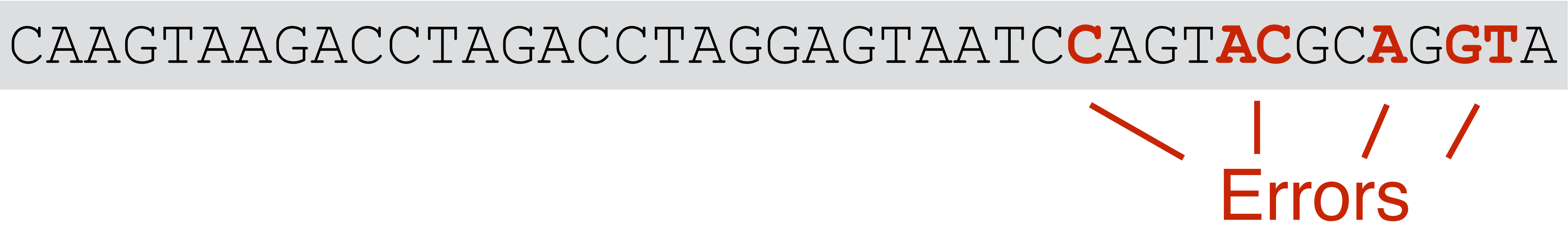
2 color

Image: Illumina

Illumina-specific Error Modes

Illumina-specific Error Modes

Declining accuracy towards end of reads:

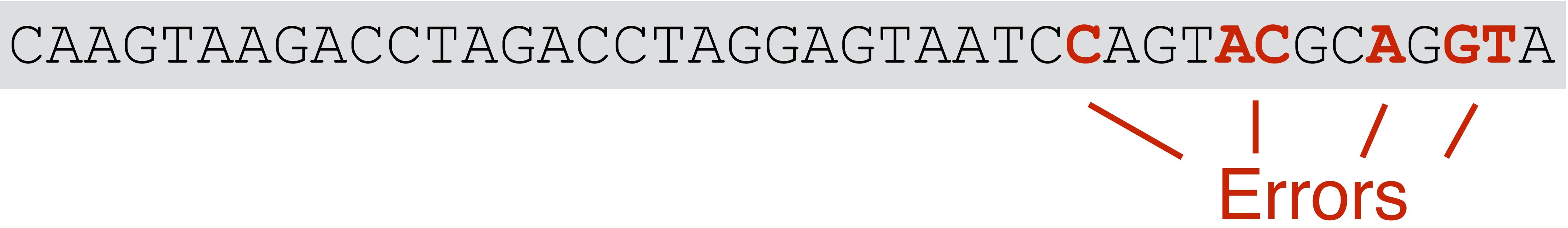


CAAGTAAGACCTAGACCTAGGAGTAATC**C**AGT**AC**GC**A**G**GT**A

Errors

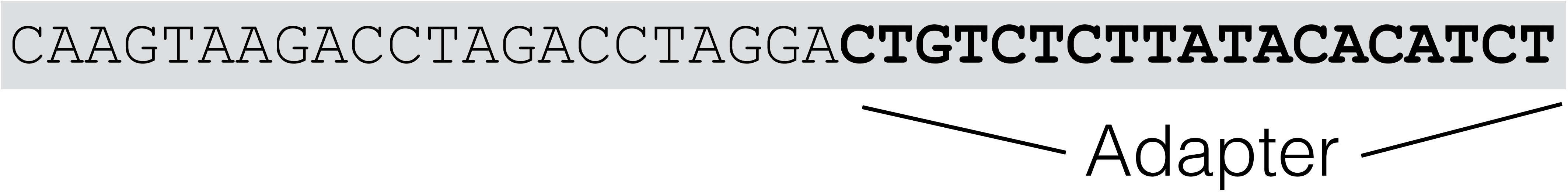
Illumina-specific Error Modes

Declining accuracy towards end of reads:



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

Read-through into adapter sequences:



CAAGTAAGACCTAGACCTAGGA**CTGTCTCTTATACACATCT**

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**GGGGGGGGGGGGGGGGGGGGGGG**

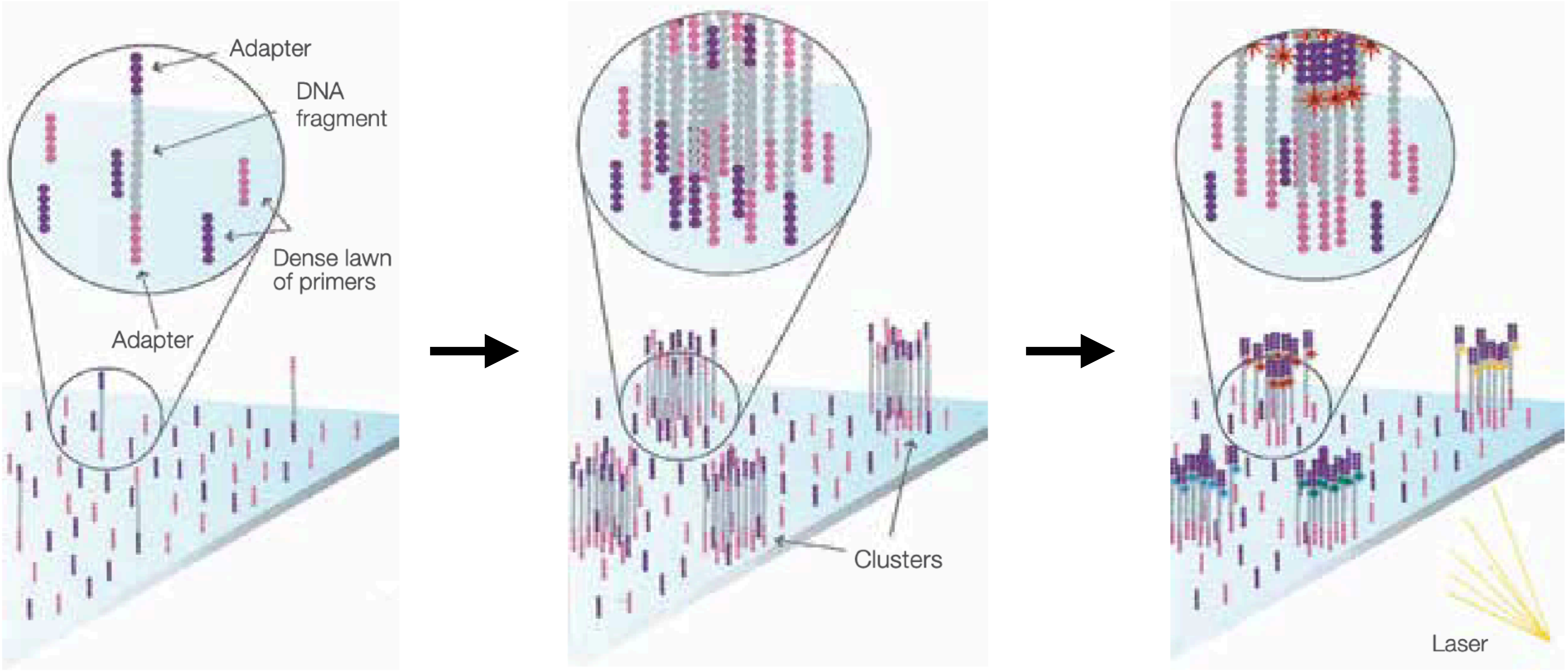
polyGs

Illumina-specific Error Modes

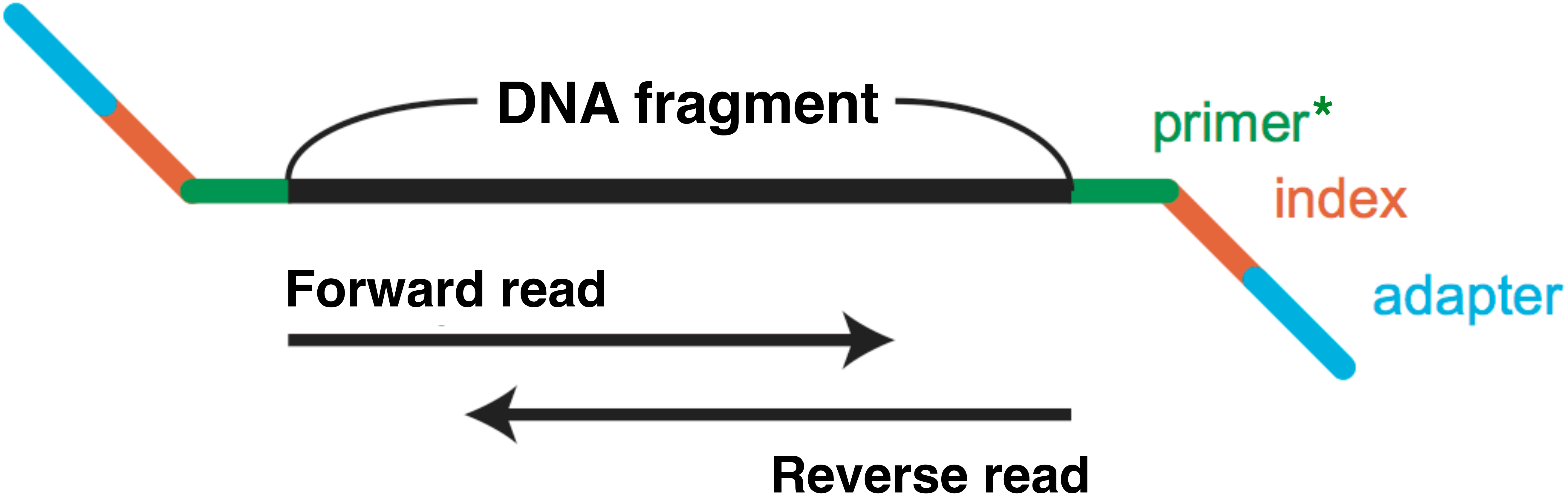
Declining accuracy towards end of reads: dephasing.

Read-through into adapter sequences: see below.

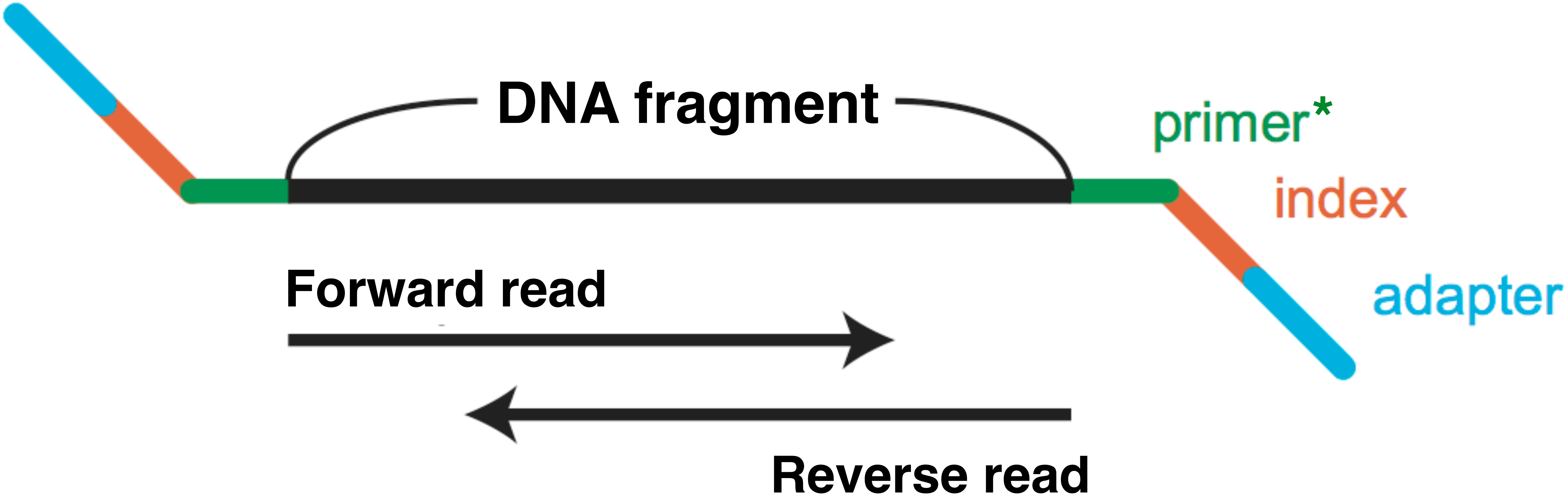
polyG tails in 2-color chemistries: G = no signal.



Illumina paired-end sequencing



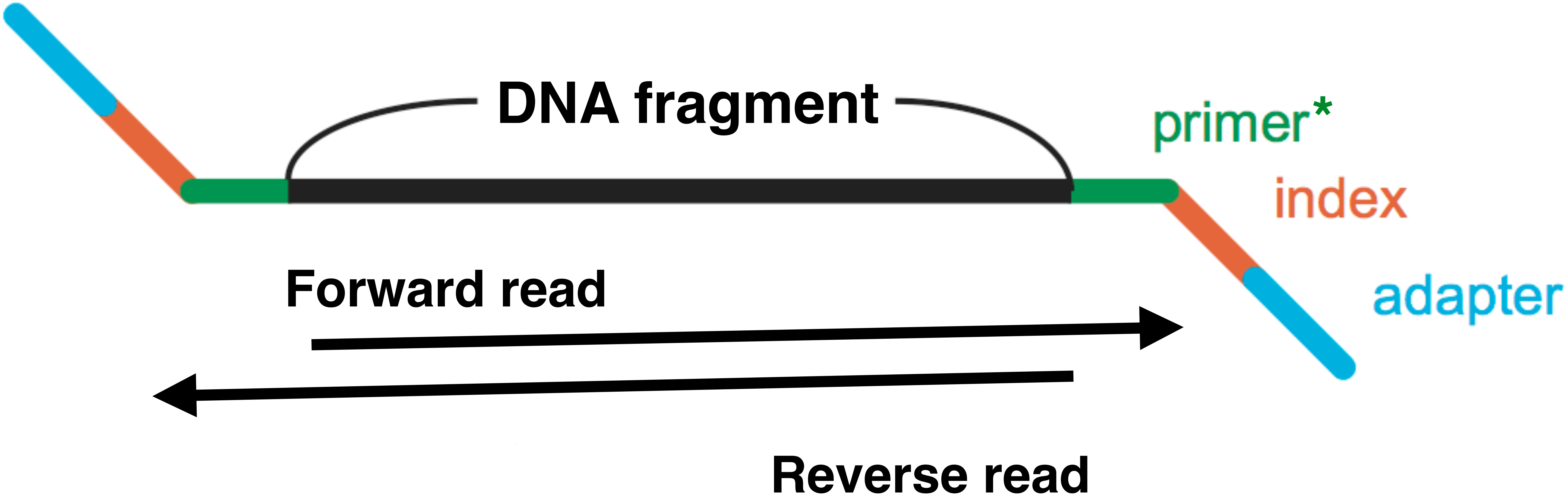
Illumina paired-end sequencing



Overlapping

Read-length < DNA length < 2 x Read-length

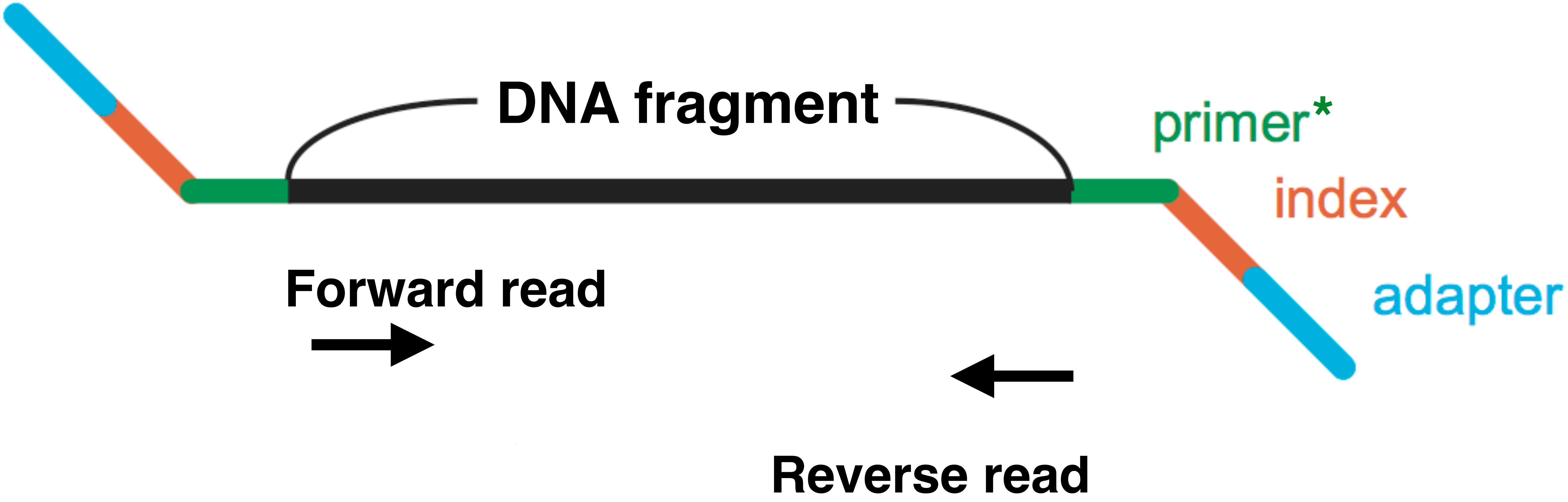
Illumina paired-end sequencing



Overhang

Read-length > DNA length

Illumina paired-end sequencing



Separated

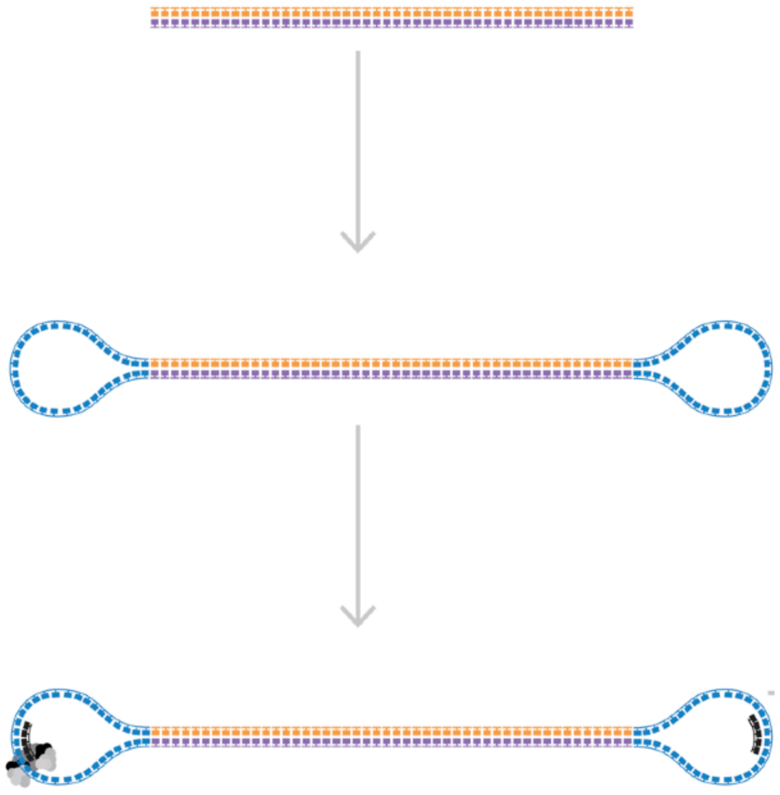
2 x Read-length < DNA length

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%**



Start with high-quality double stranded DNA

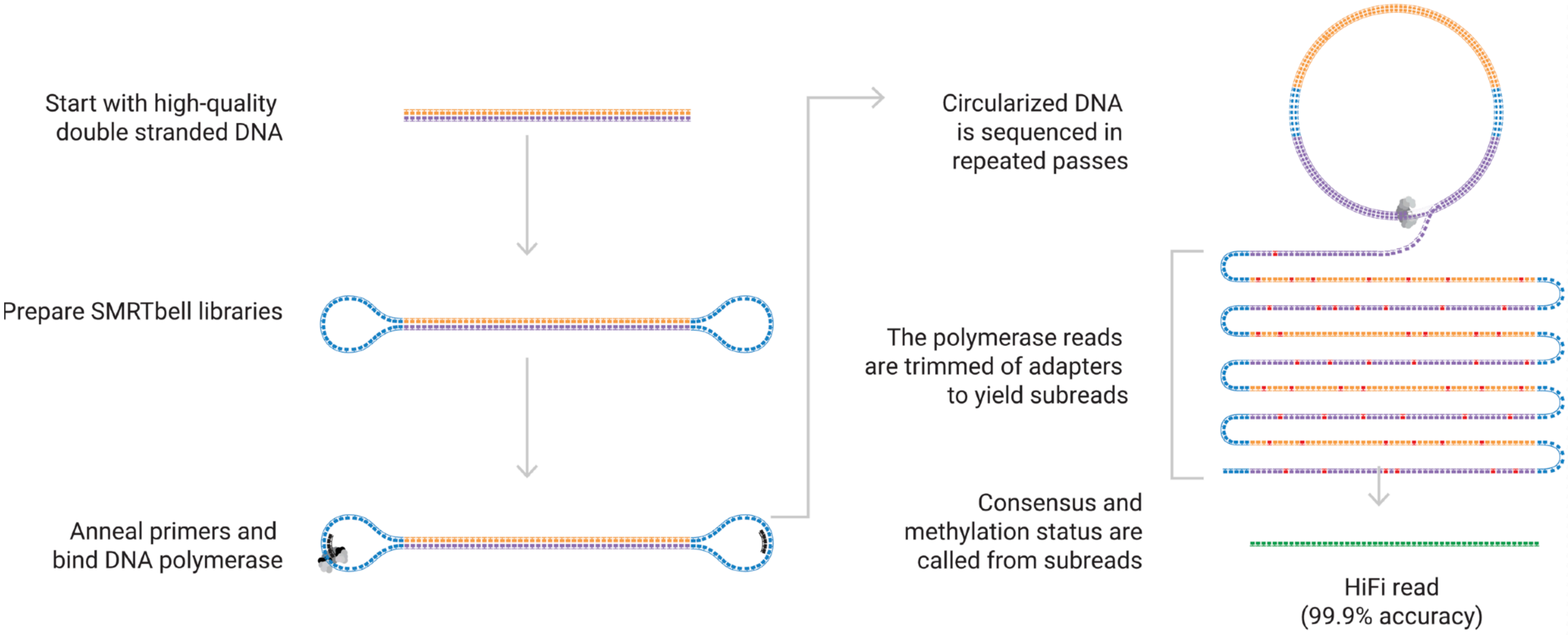
Prepare SMRTbell libraries

Anneal primers and bind DNA polymerase

Image: Pacbio

PacBio HiFi sequencing

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



Start with high-quality double stranded DNA

Prepare SMRTbell libraries

Anneal primers and bind DNA polymerase

Circularized DNA is sequenced in repeated passes

The polymerase reads are trimmed of adapters to yield subreads

Consensus and methylation status are called from subreads

HiFi read (99.9% accuracy)

Image: Pacbio

Pacbio HiFi specific Error Modes

Pacbio HiFi specific Error Modes

None.

Pacbio HiFi specific Error Modes

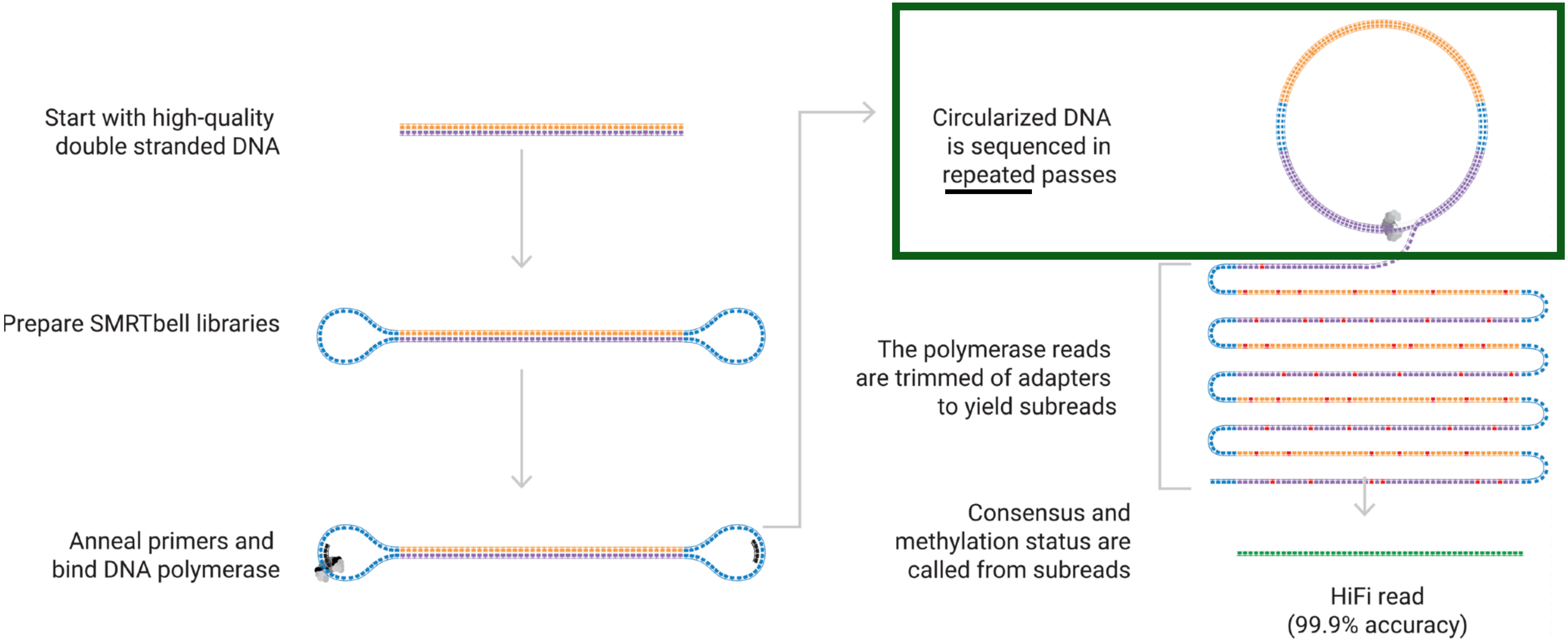
None.

* that the speaker has been able to identify

Pacbio HiFi specific Error Modes

None.

$$: Higher per-base costs.



Start with high-quality double stranded DNA

Prepare SMRTbell libraries

Anneal primers and bind DNA polymerase

Circularized DNA is sequenced in repeated passes

The polymerase reads are trimmed of adapters to yield subreads

Consensus and methylation status are called from subreads

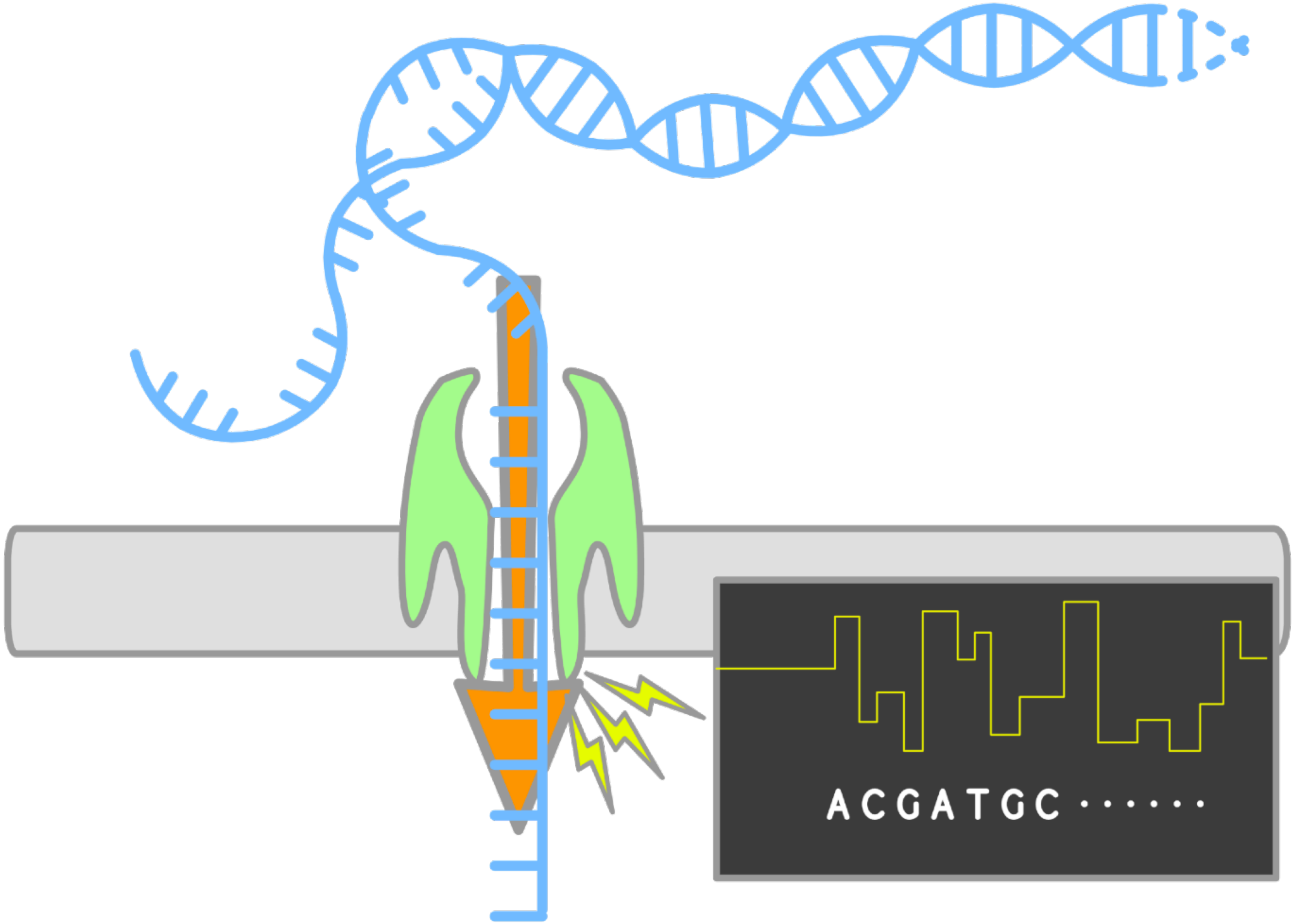
HiFi read (99.9% accuracy)

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%**

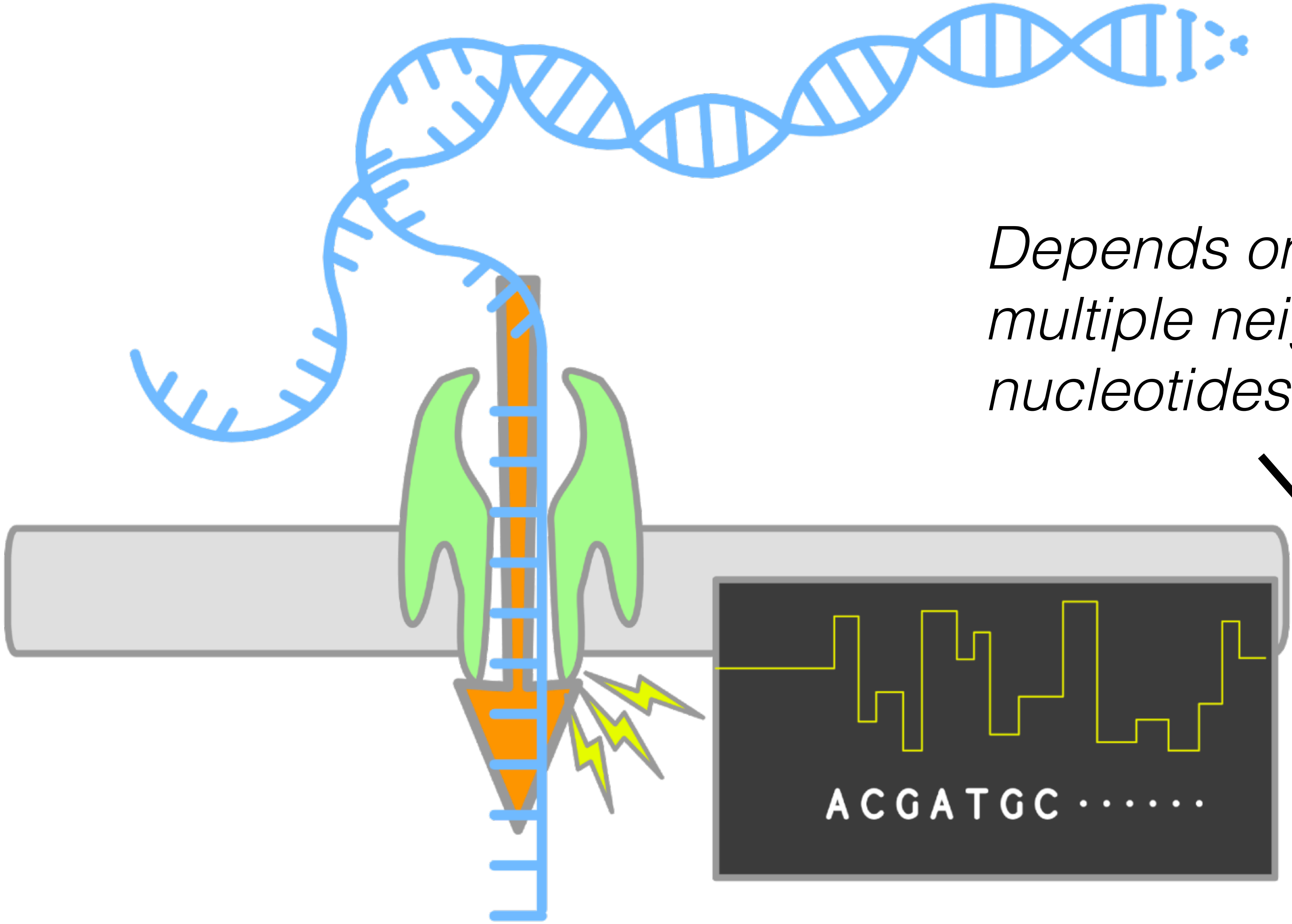


Electric current density

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

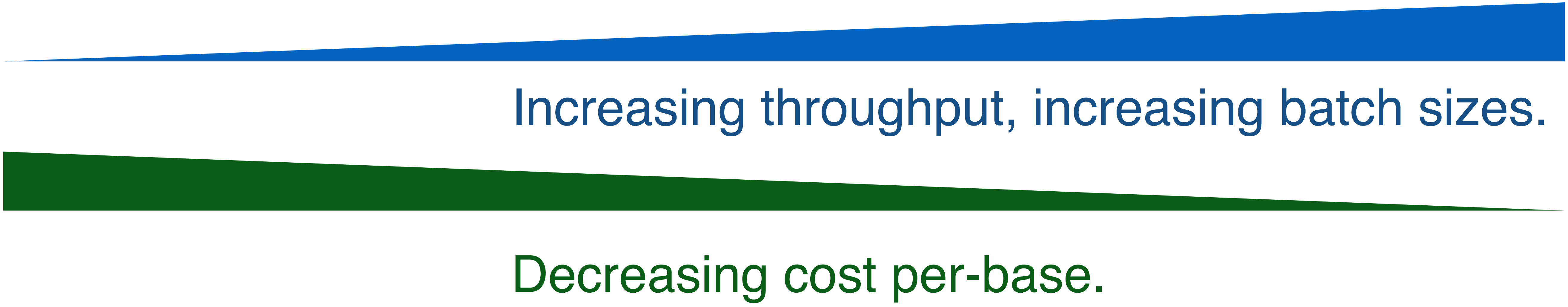
Oxford Nanopore

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`**`CCCCCCCCCCCCCCC`**`TTATA`

Incorrect length

Nanopore specific Error Modes

Homopolymers:

CAAGTAAGACCTAGACCTAGGA**CCCCCCCCCCCCCCC**TTATA

Incorrect length

Indels:

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

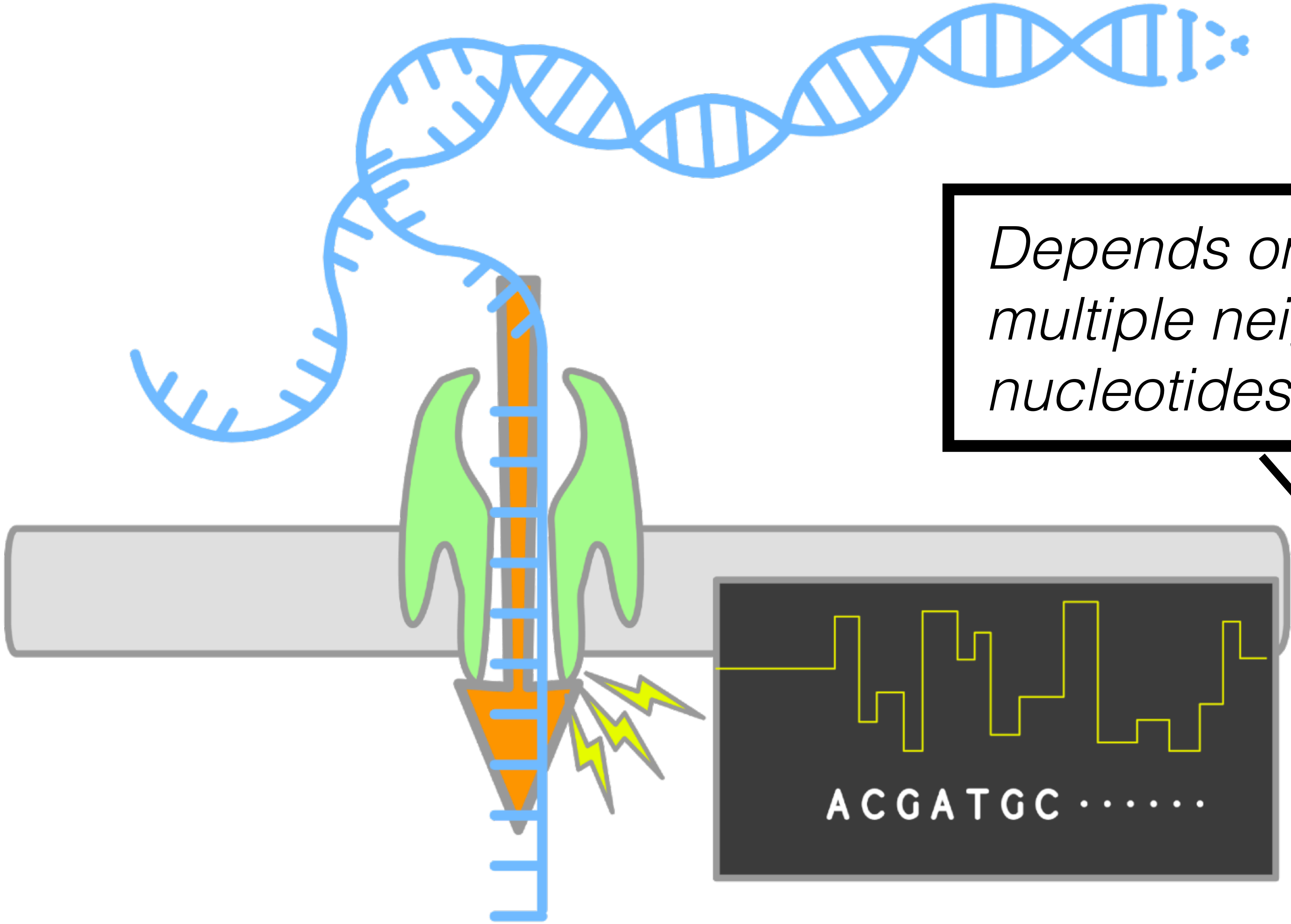
Insertions

Deletion

Nanopore specific Error Modes

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

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

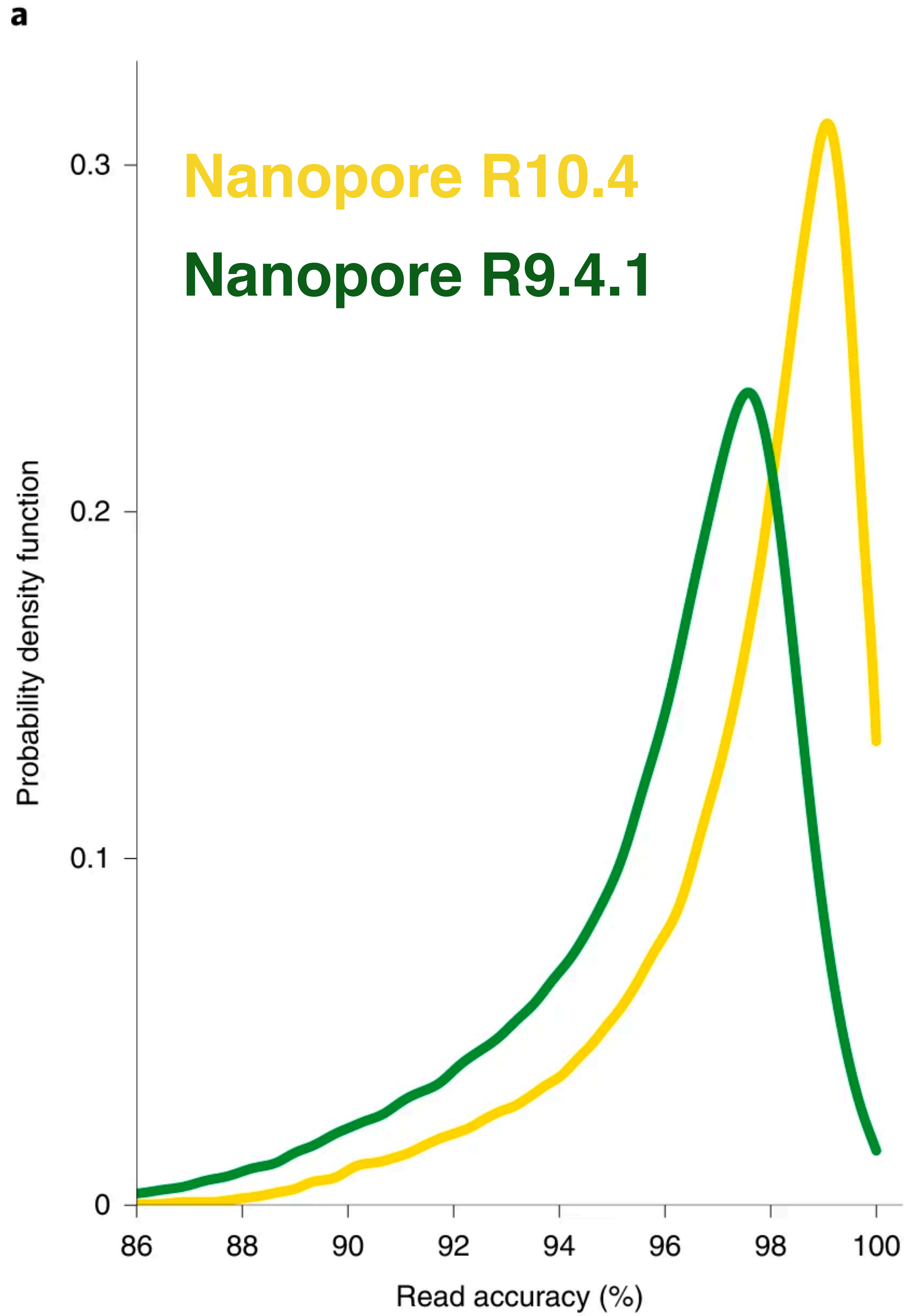


Depends on multiple neighboring nucleotides

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

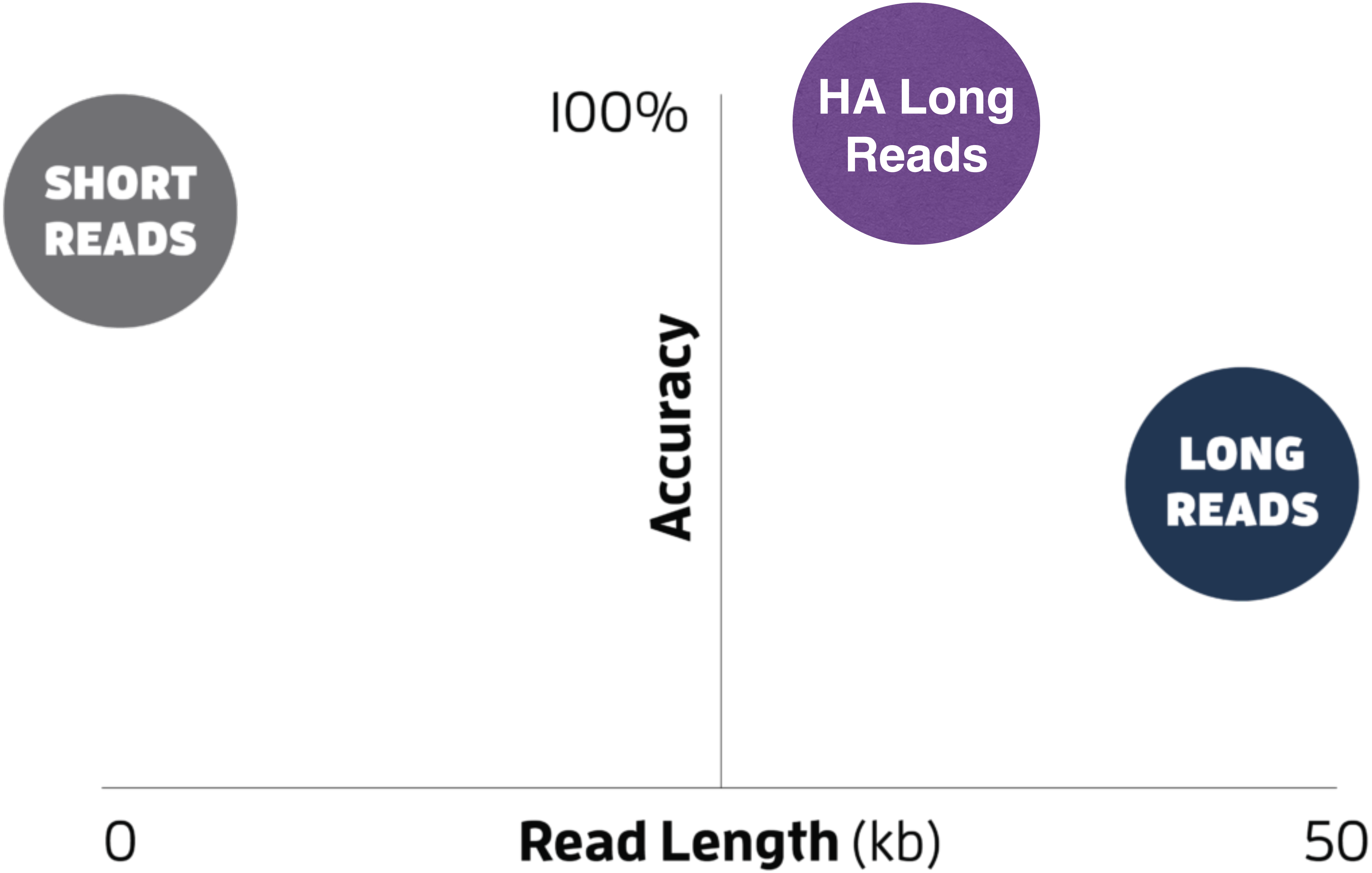
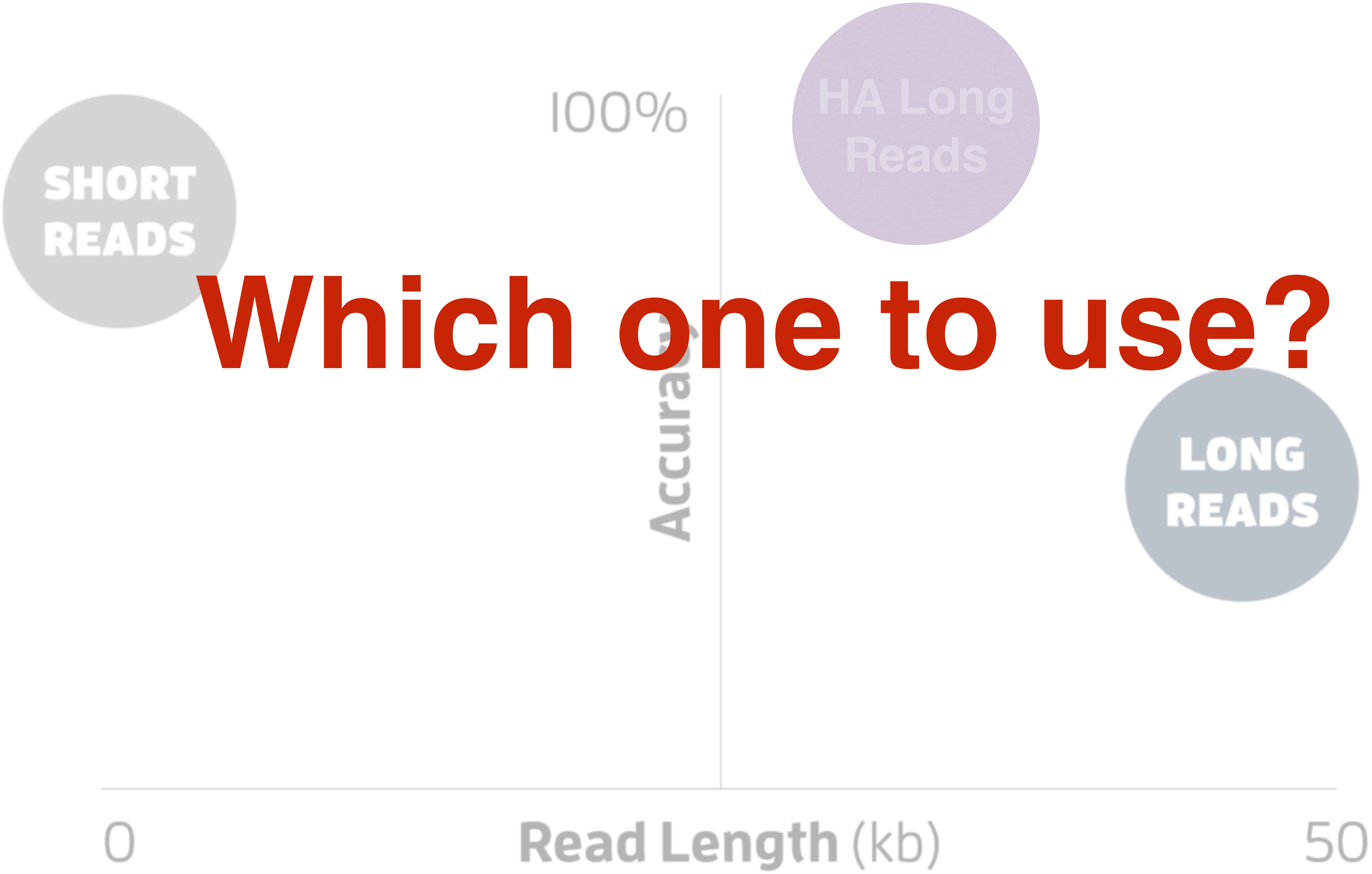


Image modified from: pacb.com

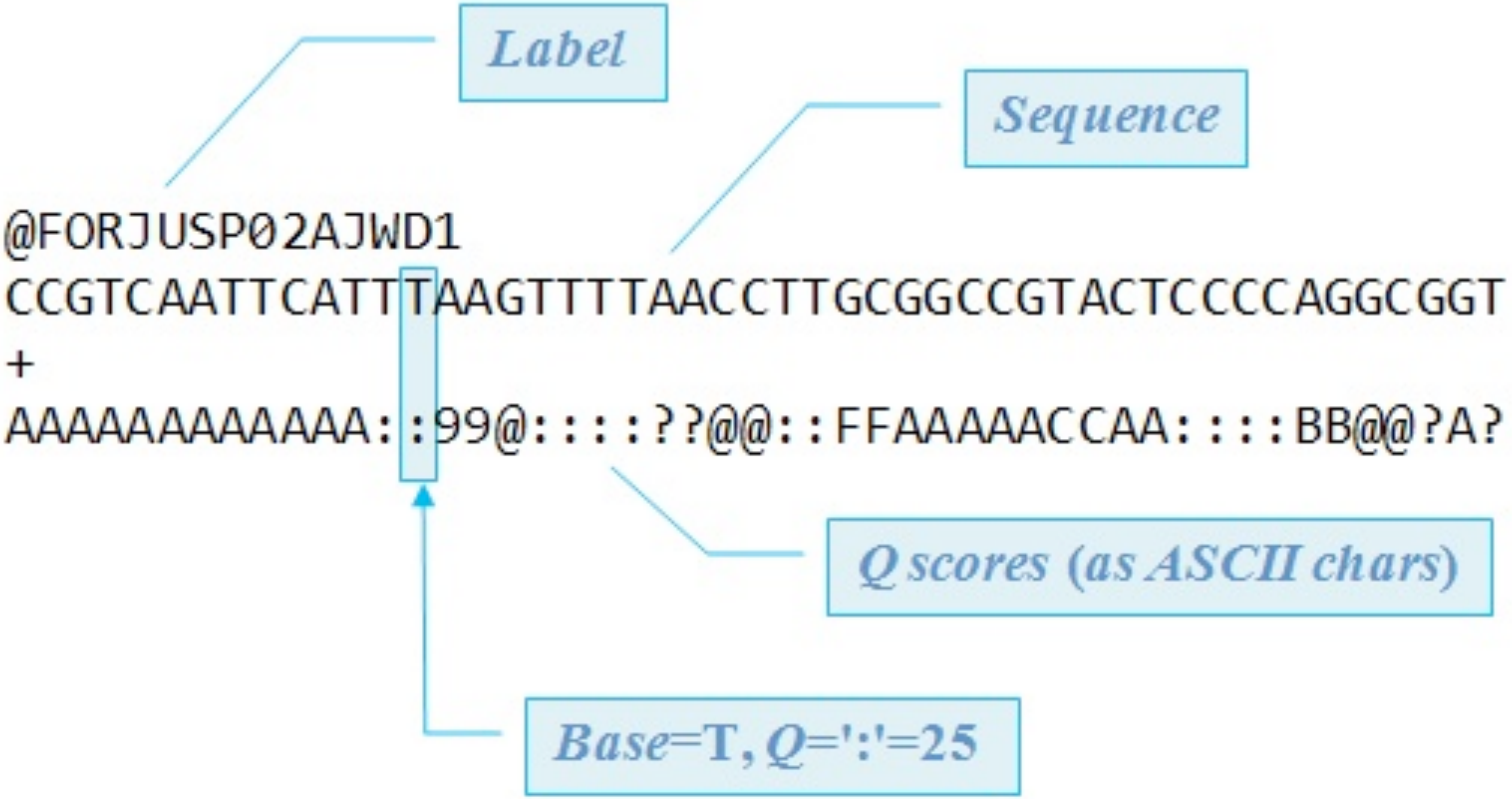
An (incomplete) sequencing survey



Which one to use?

Image modified from: pacb.com

Fastq files and Quality scores



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

Label

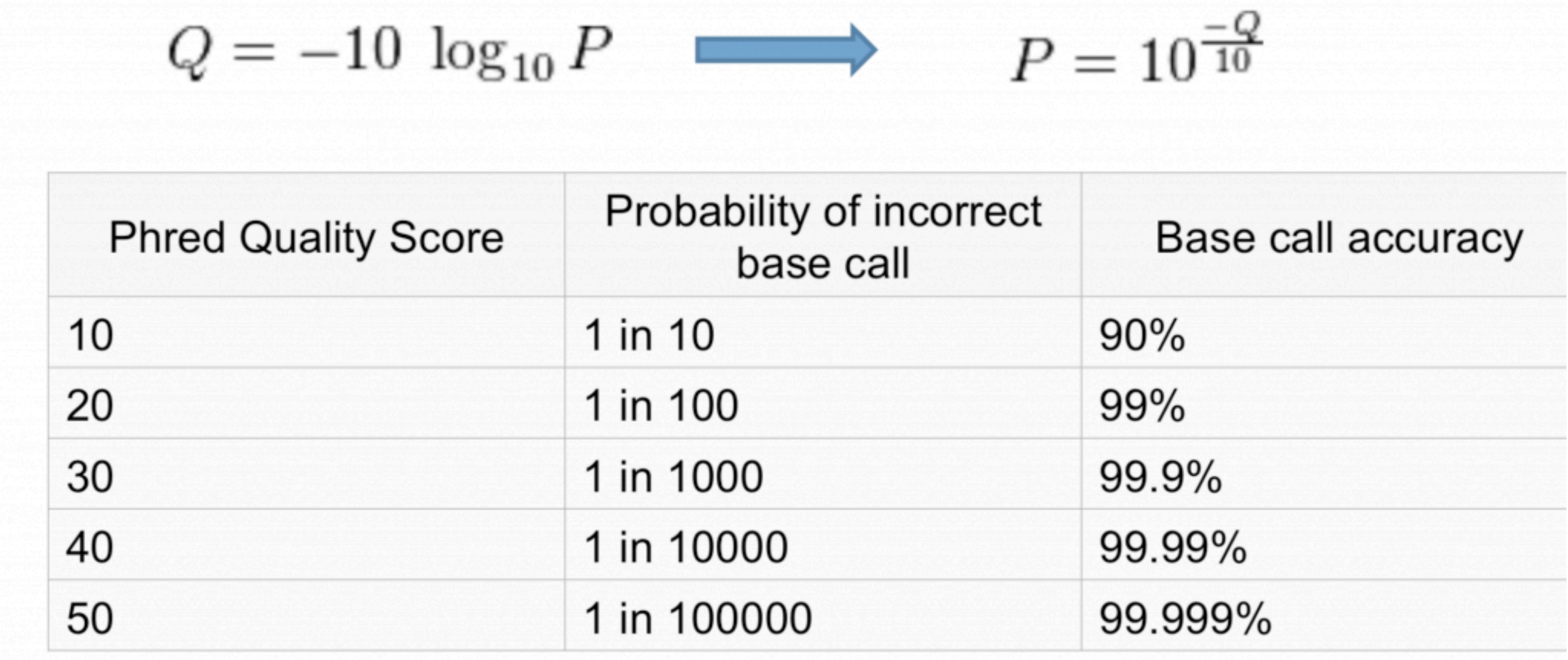
Sequence

Q scores (as ASCII chars)

Base=T, Q=':'=25

Image: Robert Edgar, drive5.com

Fastq files and Quality scores

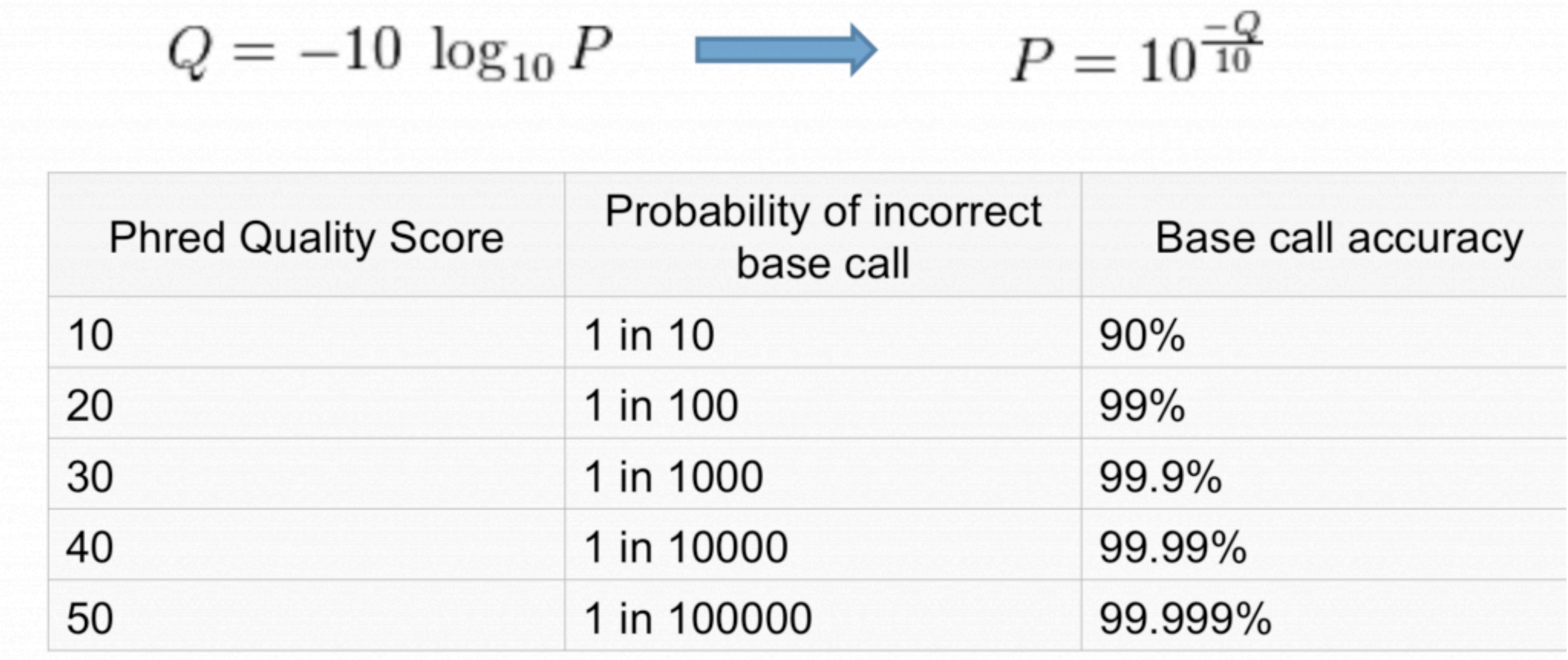


$$Q = -10\ \log_{10} P \longrightarrow P = 10^{\frac{-Q}{10}}$$

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

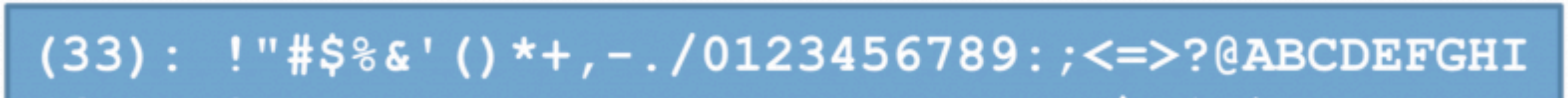
Fastq files and Quality scores



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

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:

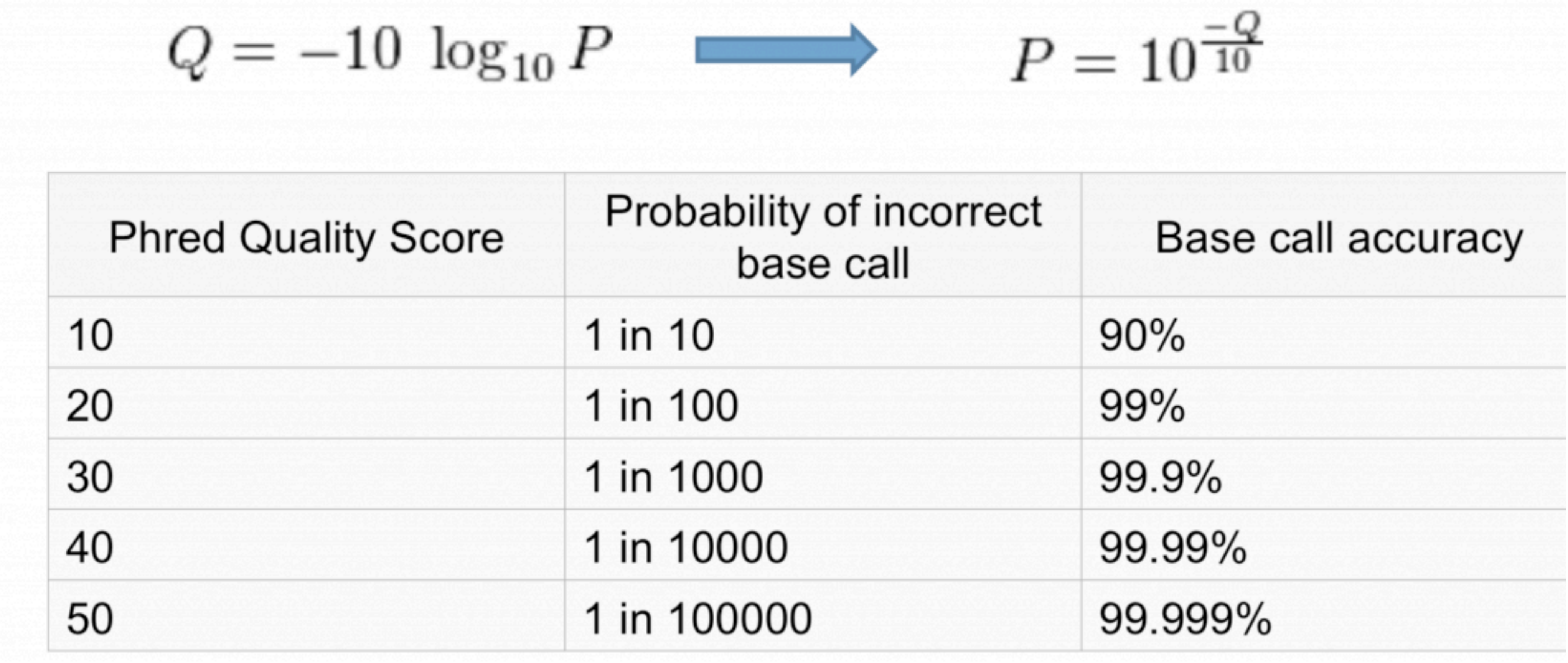


```
(33): !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
```

Q=0 → Q=40

Image: NYU Center For Genomics and Systems Biology

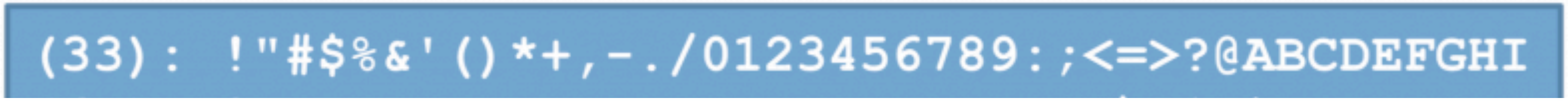
Fastq files and Quality scores



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

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
```

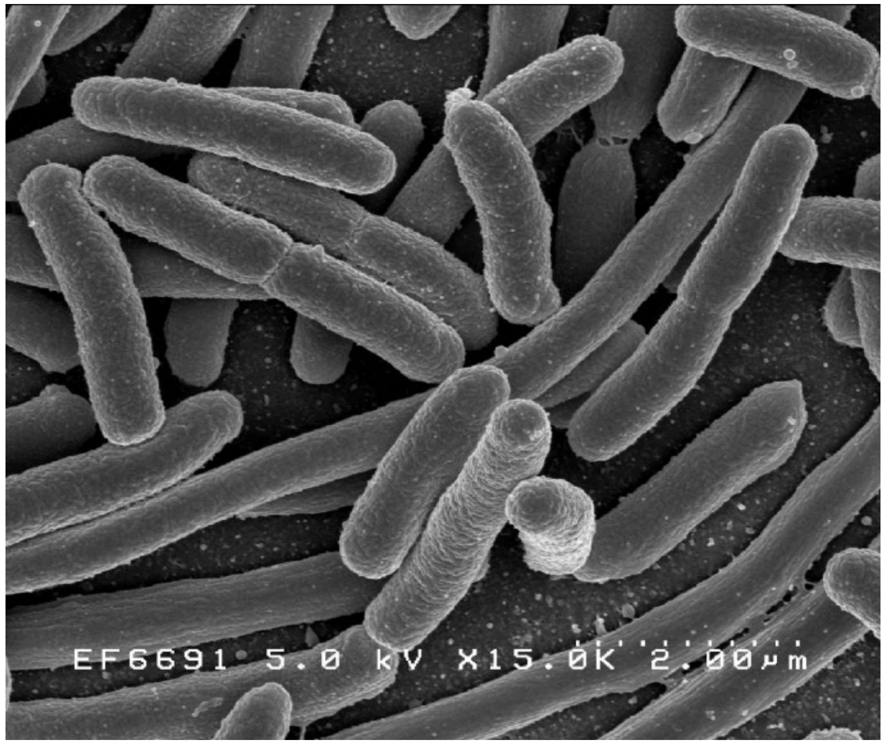
Q=0 → Q=40

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

Image: NYU Center For Genomics and Systems Biology

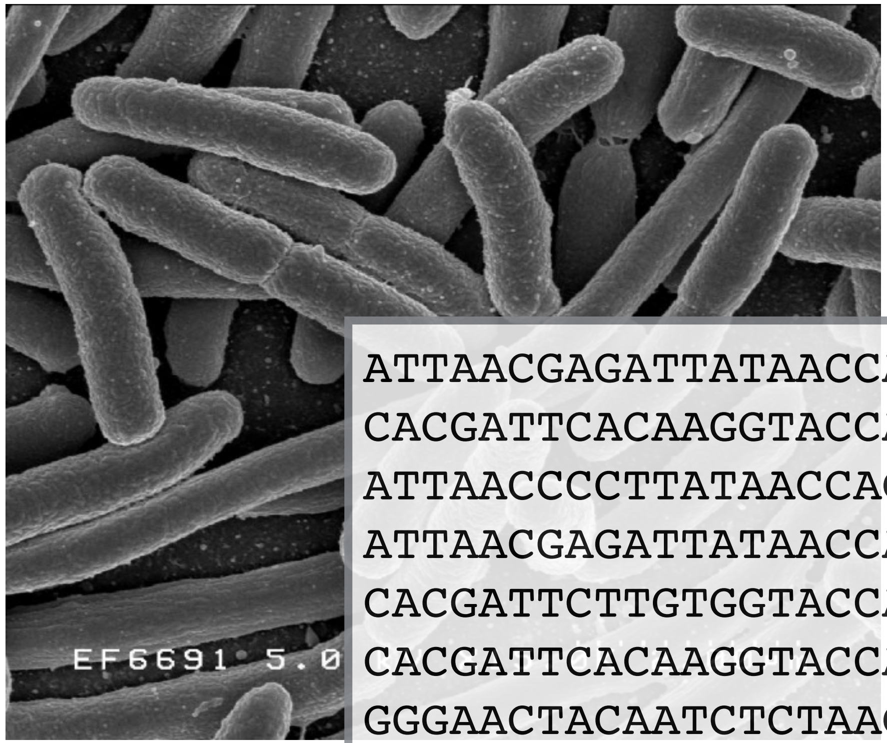
A Microbial Census

Marker-gene or Metagenomics Sequencing ***(MGS)***



A Microbial Census

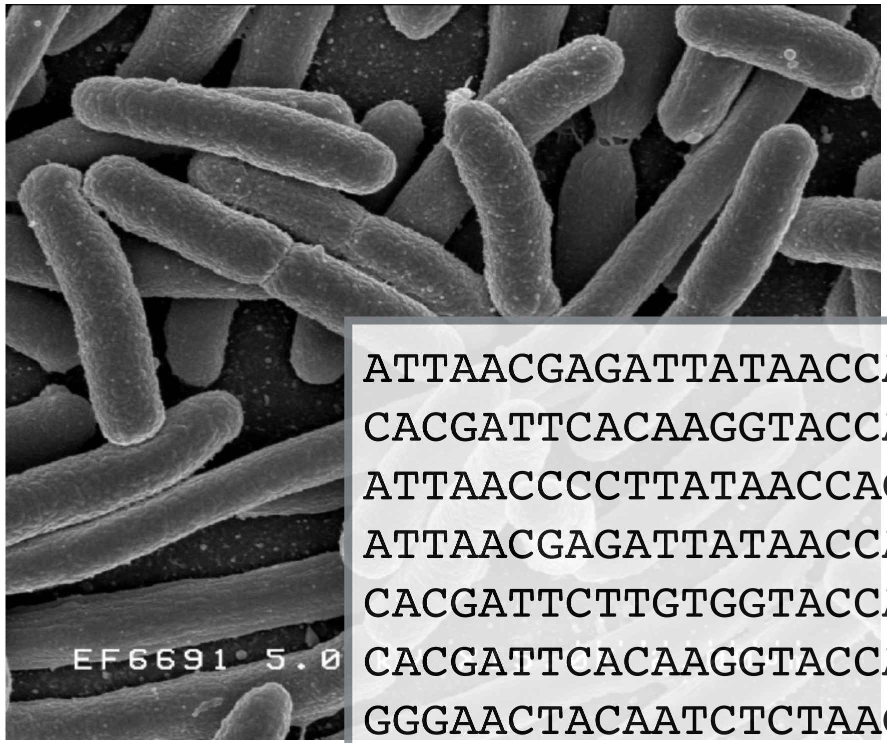
Marker-gene or Metagenomics Sequencing ***(MGS)***



```
ATTAACGAGATTATAACCAGAGTACGAATACCGAAC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
ATTAACCCCTTATAACCAGAGTACGAATACCGAACA
ATTAACGAGATTATAACCAGAGAGAGAATACCGAAC
CACGATTCTTGTGGTACCACAAGGTAACATAGCTCC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
GGGAACTACAATCTCTAAGGTGAAGTCTCAGTCTAT
ATTAACGAGATTATAACCAGAGTACGAATACCGAAC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
ATTAACGAGATTATAACCAGAGTACGAATACCGAAC
```

A Microbial Census

Marker-gene or Metagenomics Sequencing ***(MGS)***

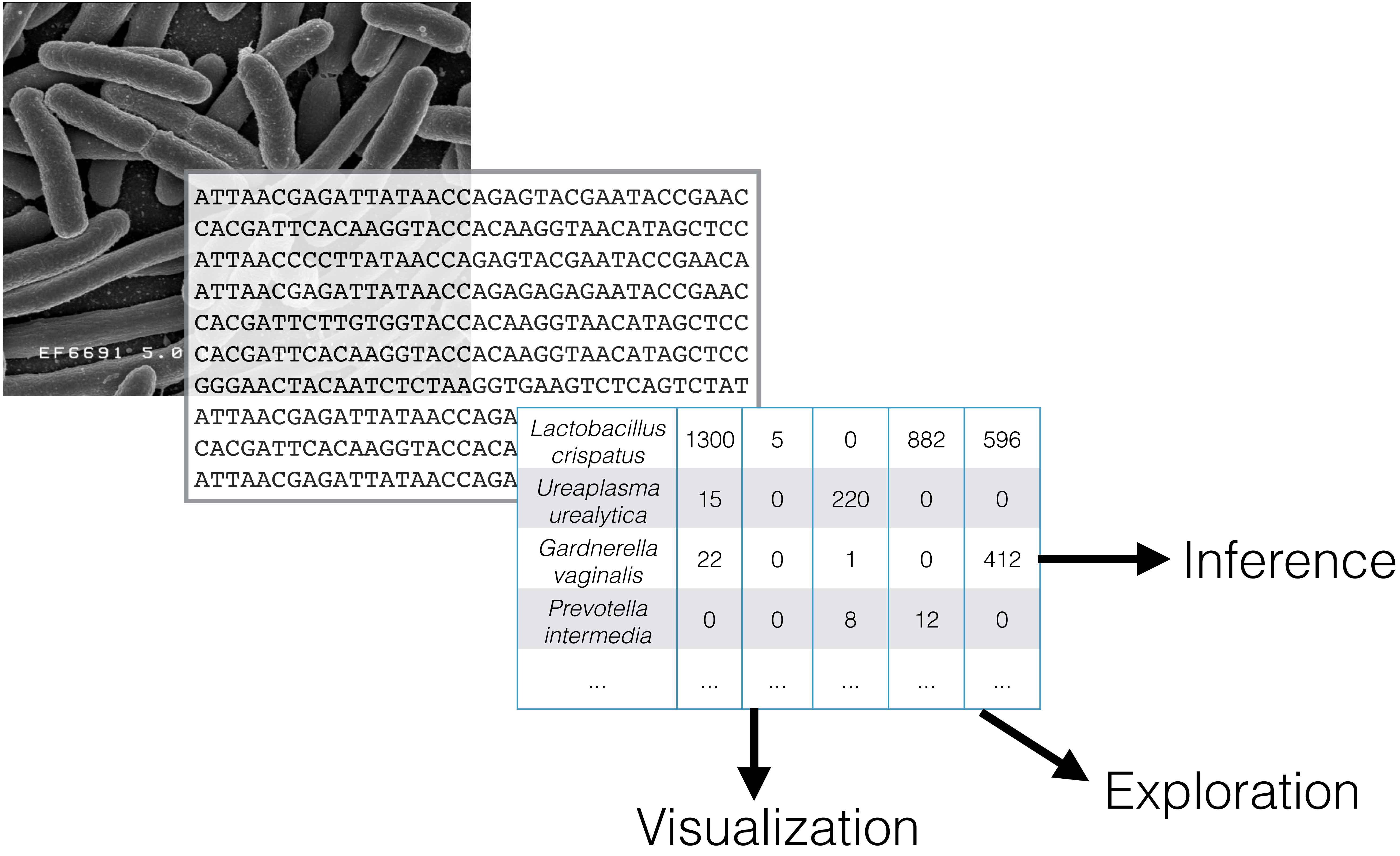


```
ATTAACGAGATTATAACCAGAGTACGAATACCGAAC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
ATTAACCCCTTATAACCAGAGTACGAATACCGAACA
ATTAACGAGATTATAACCAGAGAGAGAATACCGAAC
CACGATTCTTGTGGTACCACAAGGTAACATAGCTCC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
GGGAACTACAATCTCTAAGGTGAAGTCTCAGTCTAT
ATTAACGAGATTATAACCAGA
CACGATTCACAAGGTACCACA
ATTAACGAGATTATAACCAGA
```

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
…	…	…	…	…	…

A Microbial Census

Marker-gene or Metagenomics Sequencing ***(MGS)***



```
ATTAACGAGATTATAACCAGAGTACGAATACCGAAC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
ATTAACCCCTTATAACCAGAGTACGAATACCGAACA
ATTAACGAGATTATAACCAGAGAGAGAATACCGAAC
CACGATTCTTGTGGTACCACAAGGTAACATAGCTCC
CACGATTCACAAGGTACCACAAGGTAACATAGCTCC
GGGAACTACAATCTCTAAGGTGAAGTCTCAGTCTAT
ATTAACGAGATTATAACCAGA
CACGATTCACAAGGTACCACA
ATTAACGAGATTATAACCAGA
```

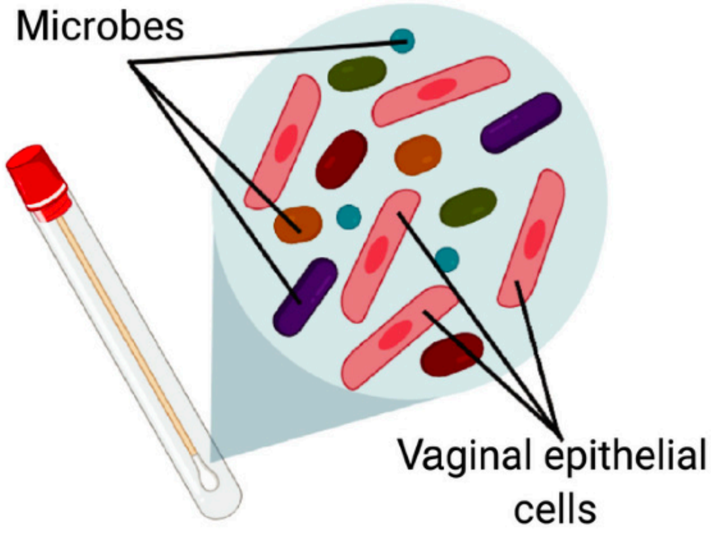
Lactobacillus crispatus	1300	5	0	882	596
Ureaplasma urealytica	15	0	220	0	0
Gardnerella vaginalis	22	0	1	0	412
Prevotella intermedia	0	0	8	12	0
...	...	...	...	...	...

Inference

Visualization

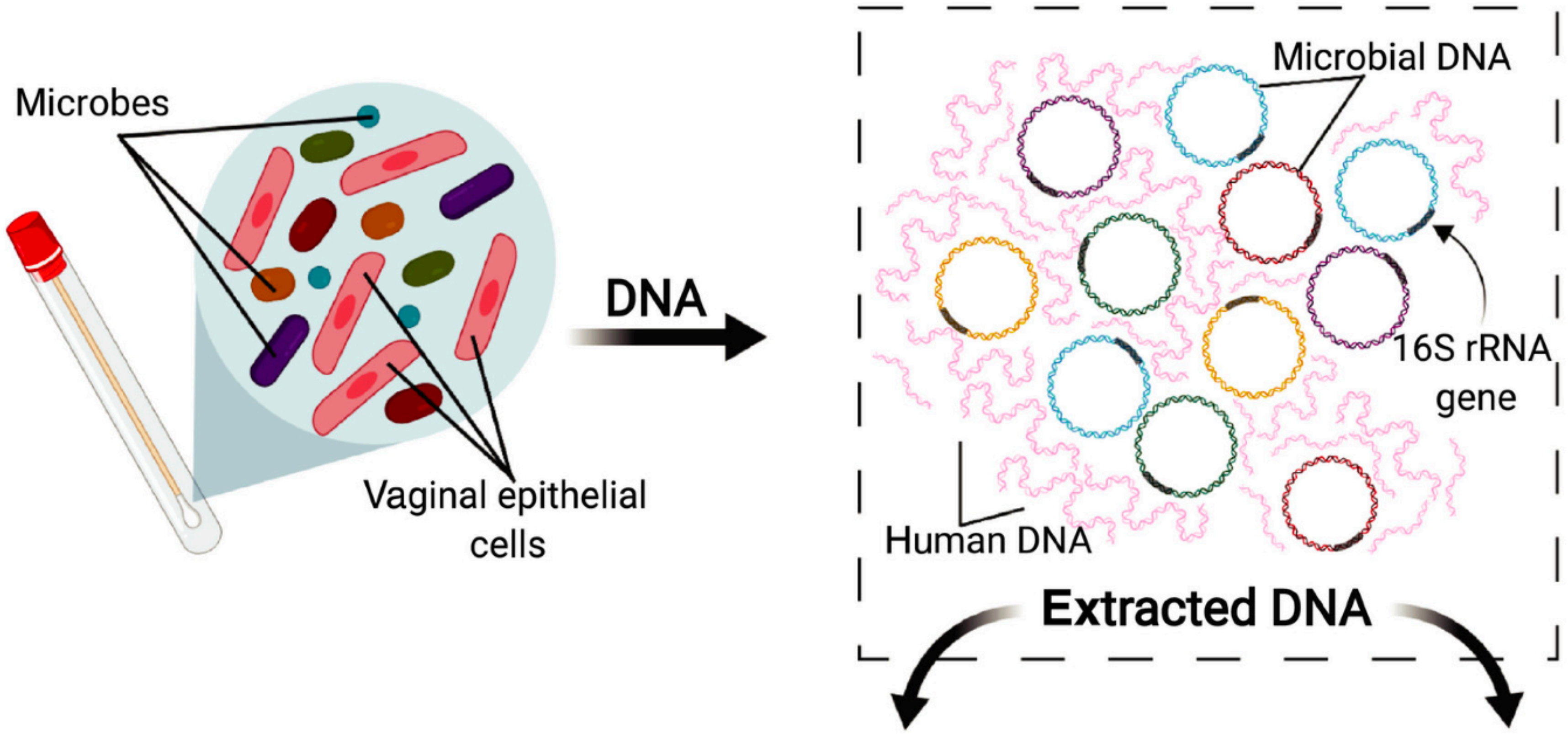
Exploration

Community Sequencing



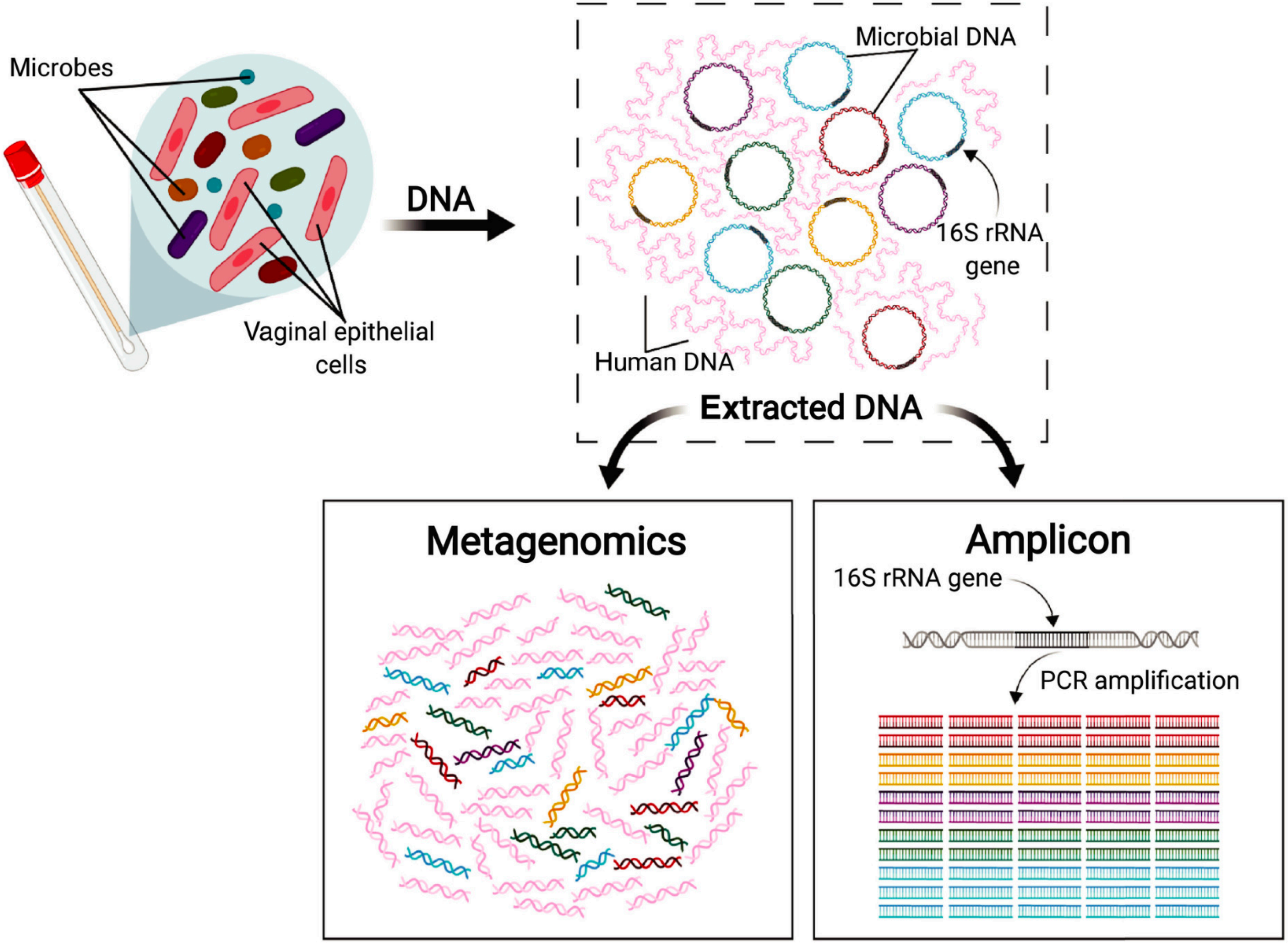
Berman, BJOG 2020.

Community Sequencing



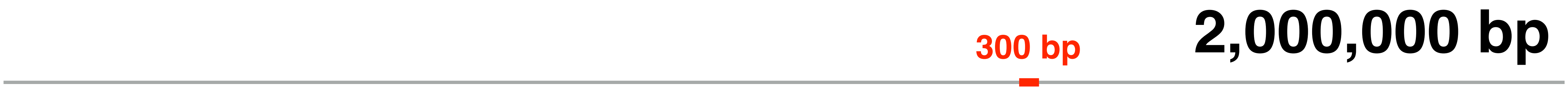
Berman, BJOG 2020.

Community Sequencing



Berman, BJOG 2020.

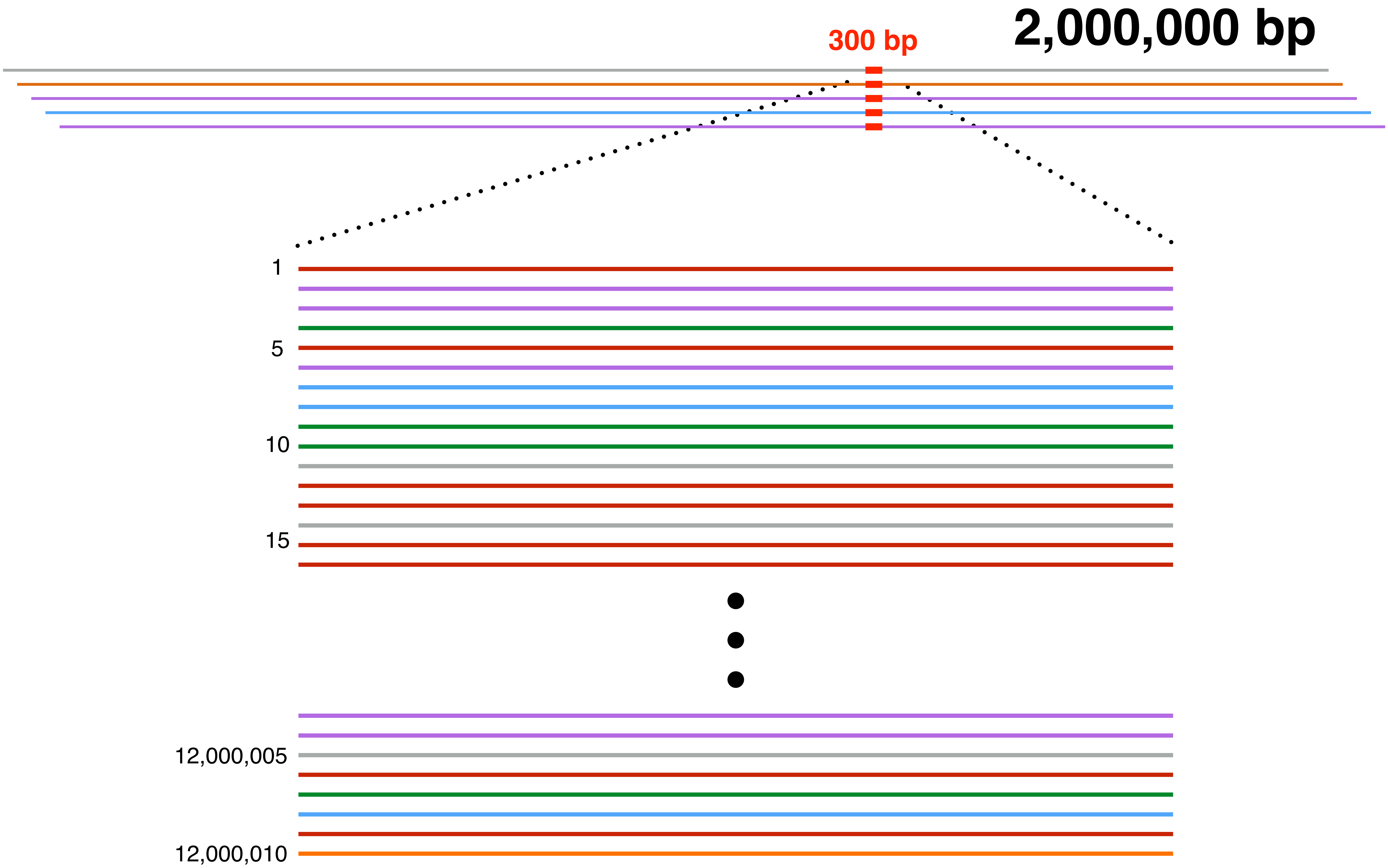
Marker-gene Sequencing



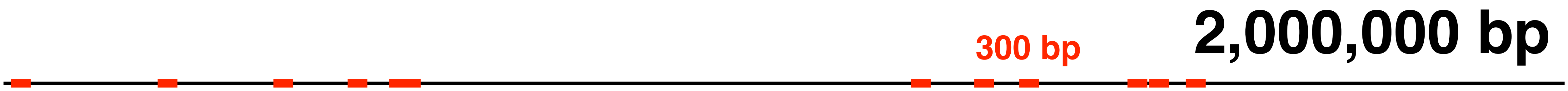
300 bp

2,000,000 bp

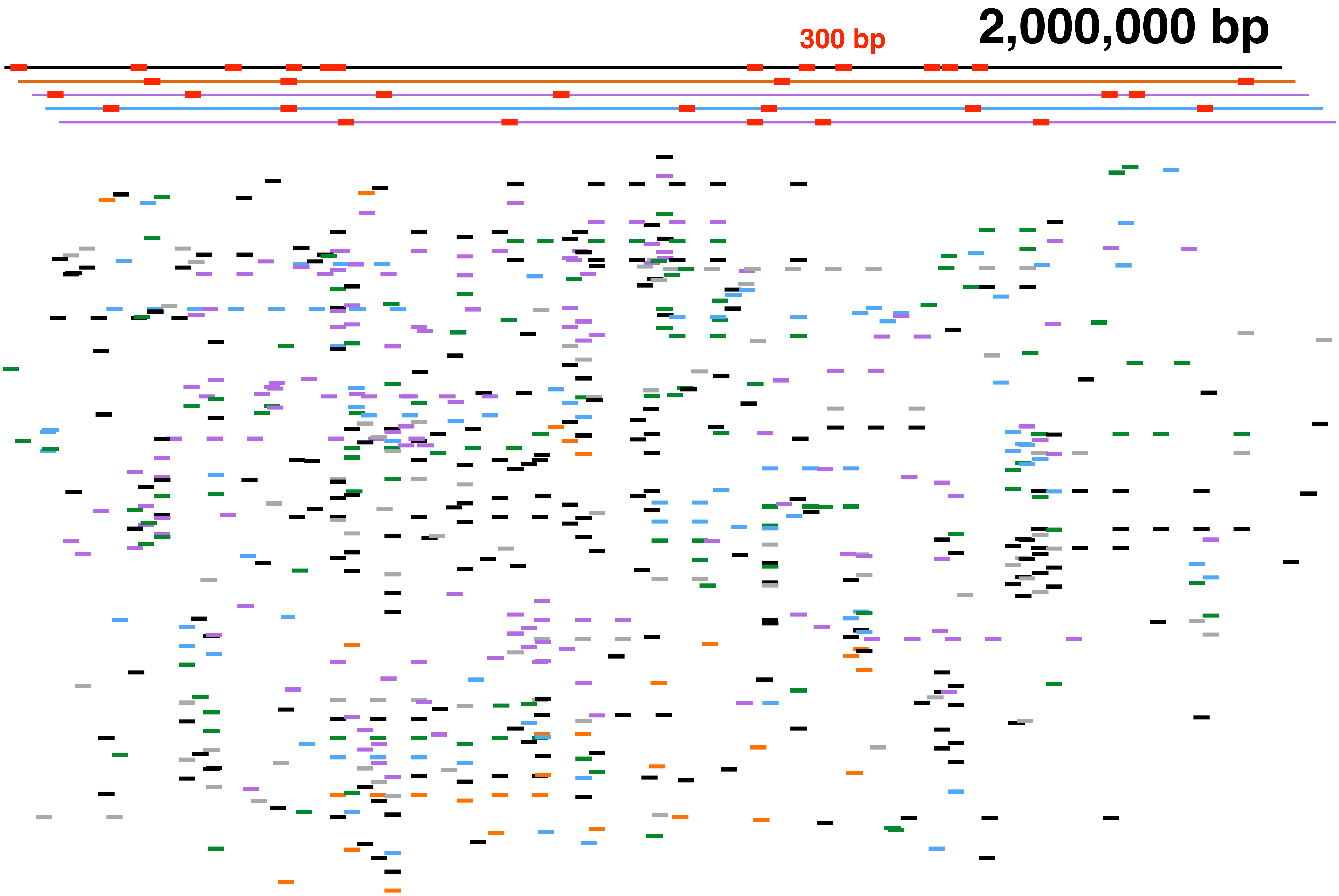
Marker-gene Sequencing



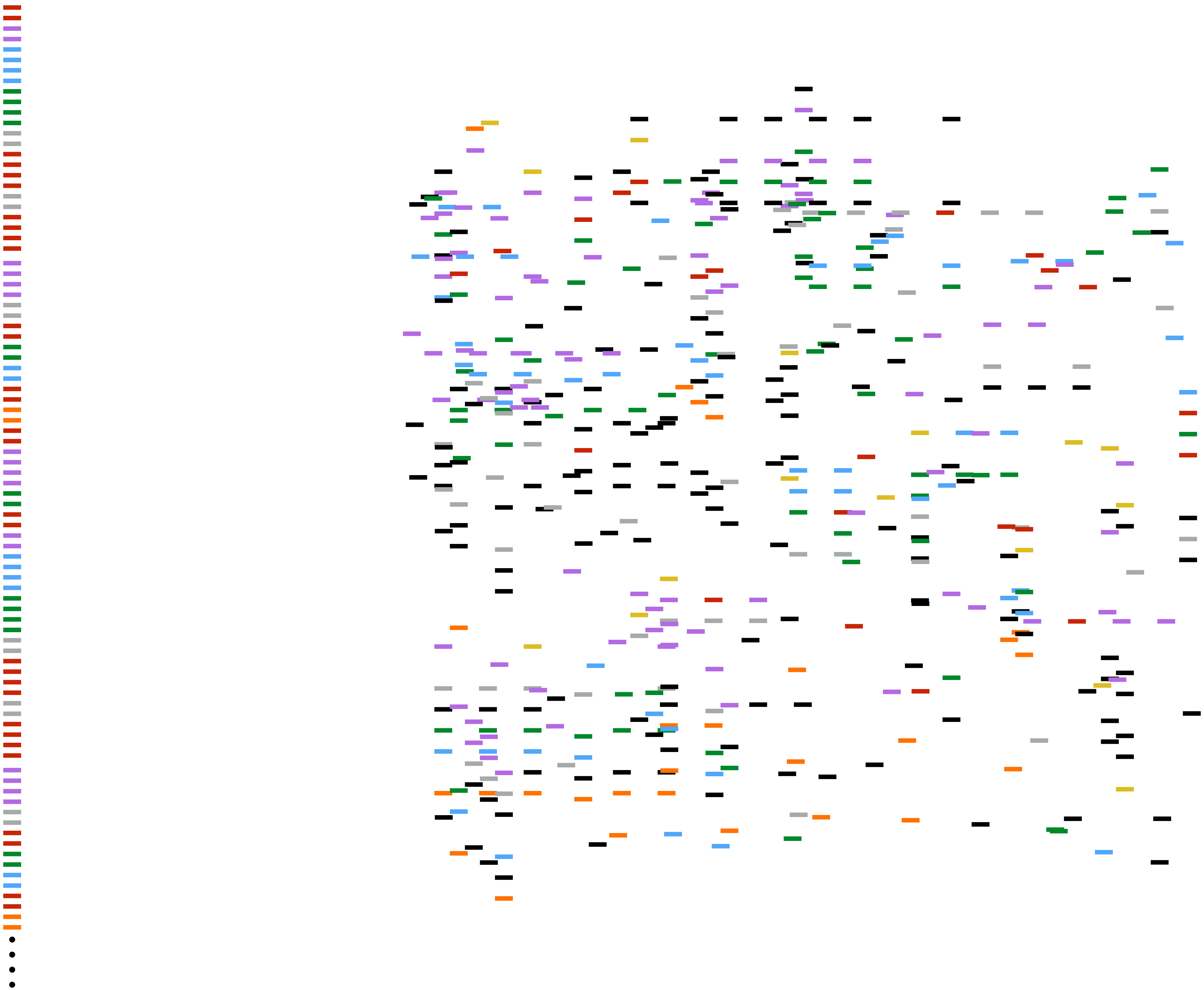
Shotgun Sequencing



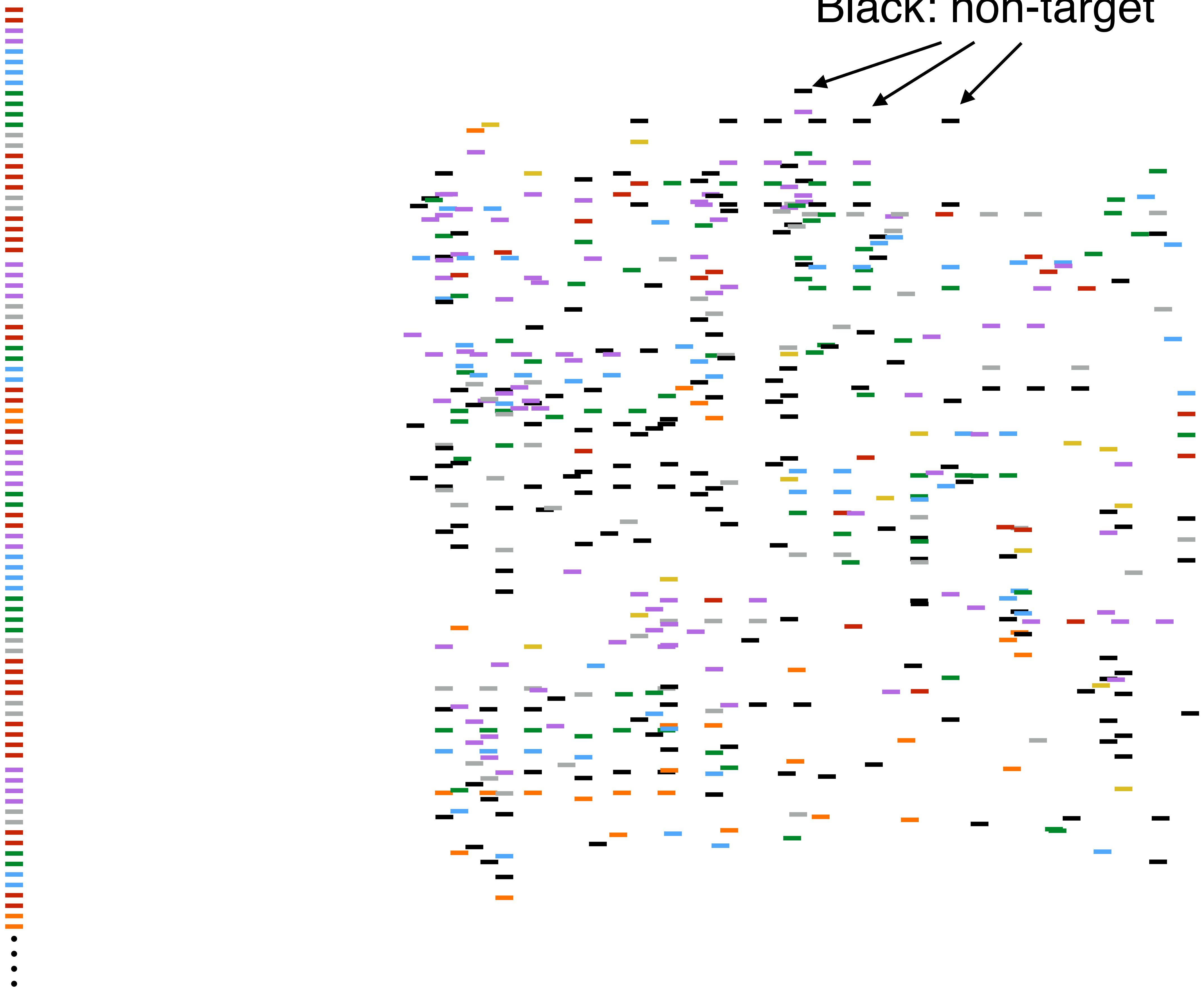
Shotgun Sequencing



Marker-gene vs. Metagenomics

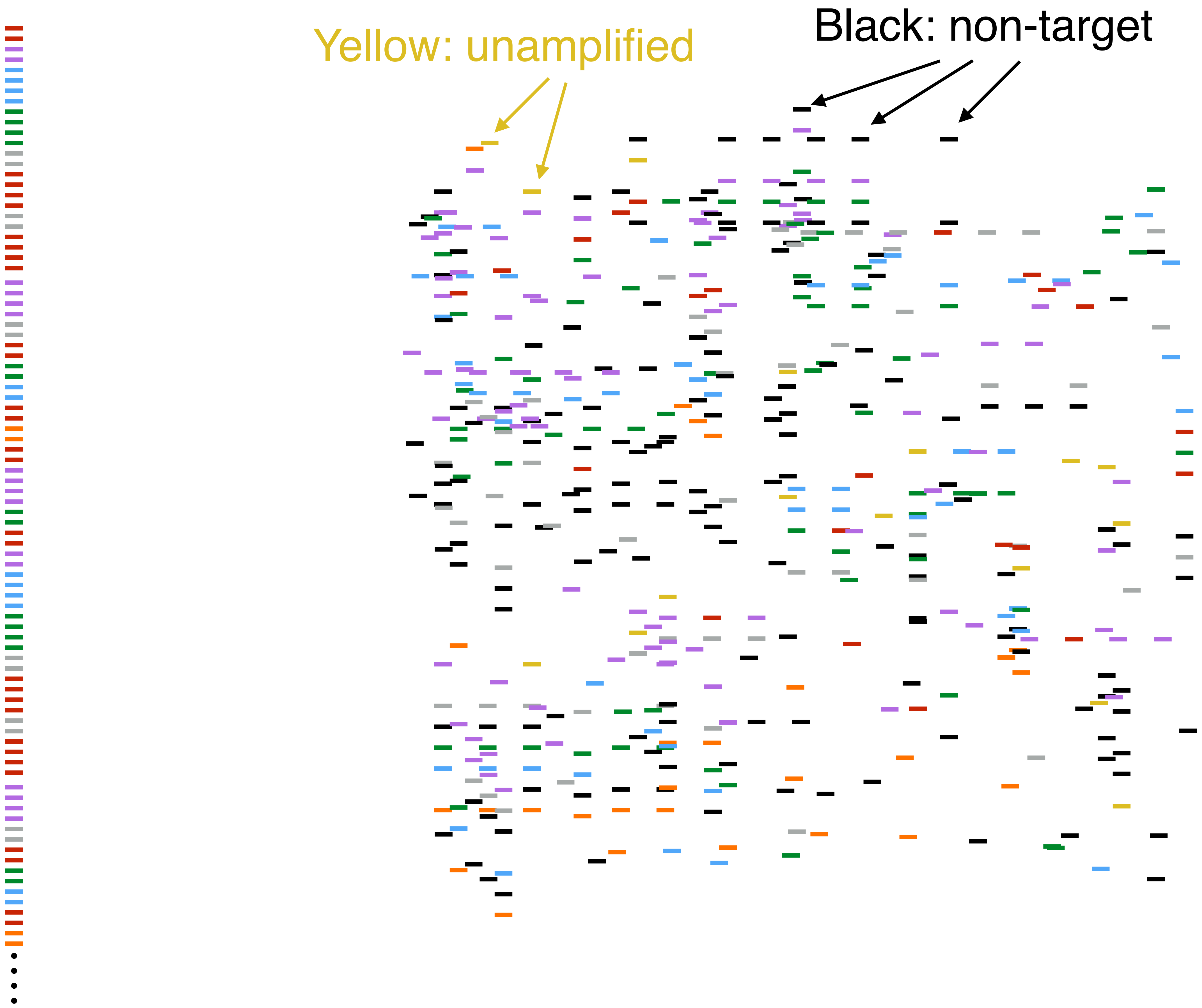


Marker-gene vs. Metagenomics



Black: non-target

Marker-gene vs. Metagenomics



Yellow: unamplified

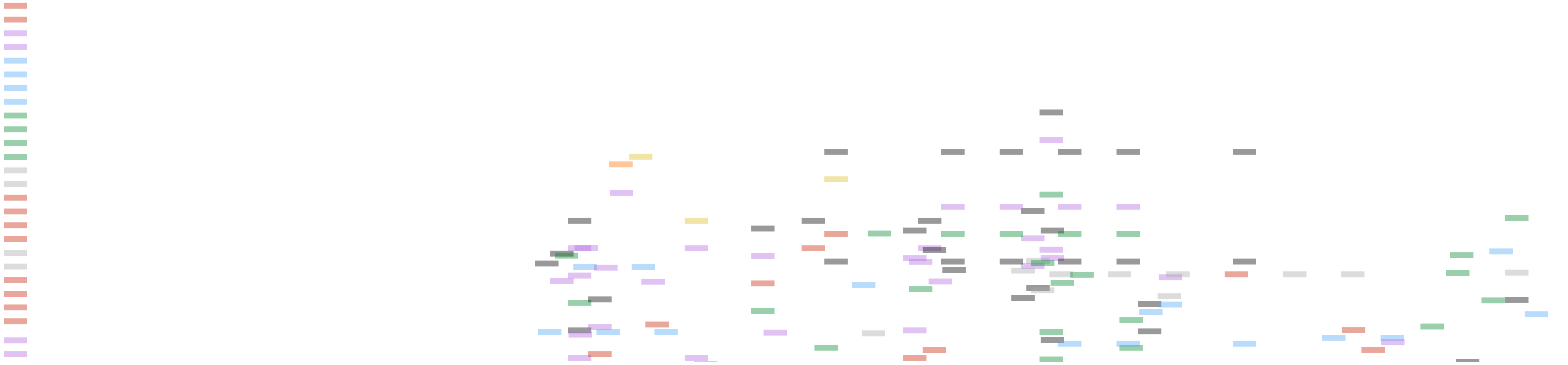
Black: non-target

Marker-gene vs. Metagenomics

What can each do?

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

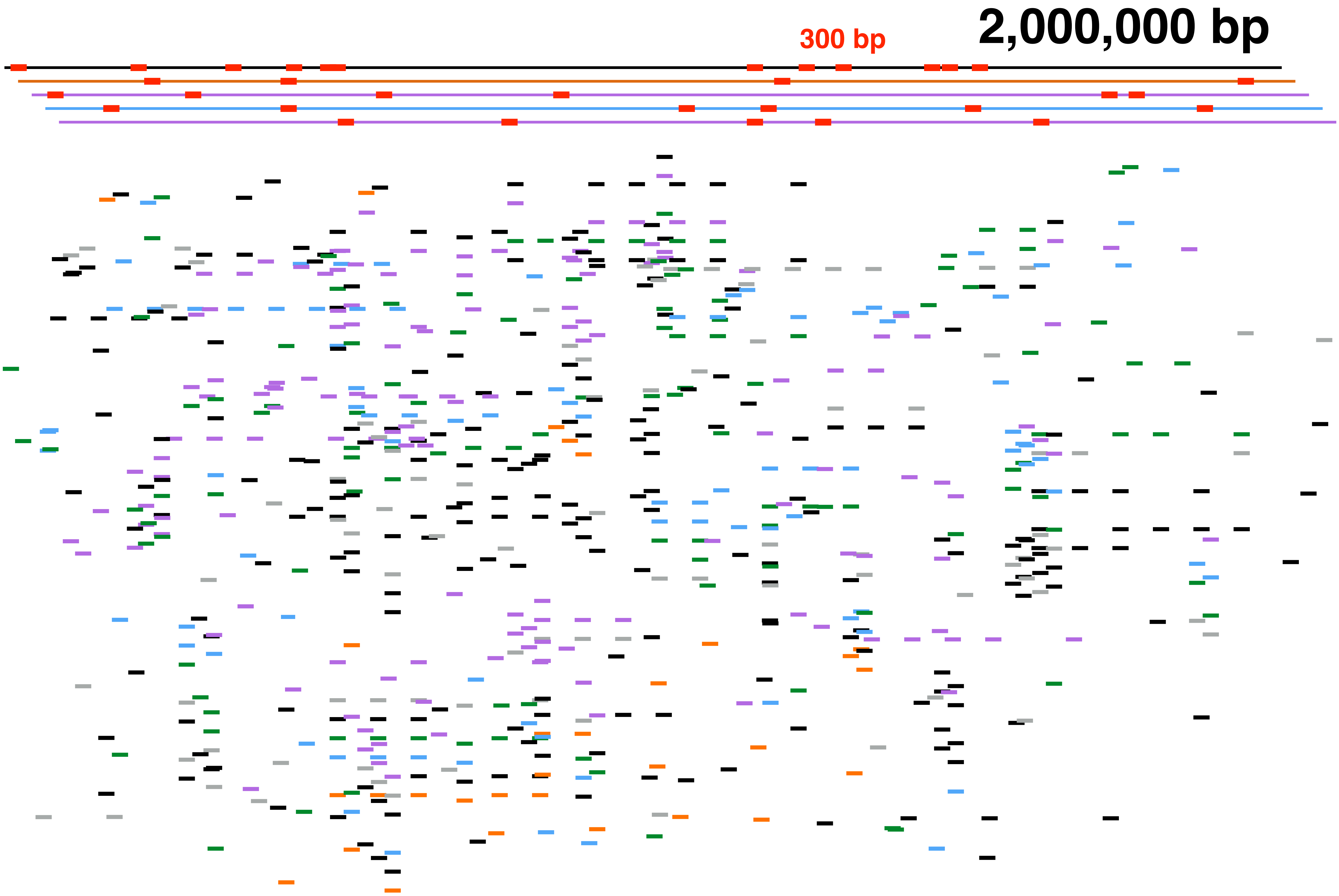
Marker-gene vs. Metagenomics



What can each not do?

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

Shotgun Metagenomics


300 bp
2,000,000 bp