MEDICAL MICROBIOLOGY AND INFECTIOUS DISEASES CODING WORKSHOP

Presents

Downloading and assembling microbial sequence data

INSTRUCTED BY

Aaron Petkau, MSc Student

INFORMATION FOR PARTICIPANTS

All workshops are being recorded and posted to the MMID Coding Workshop - YouTube

Please hold your questions until Q & A session

Question and Answer period will not be recorded.

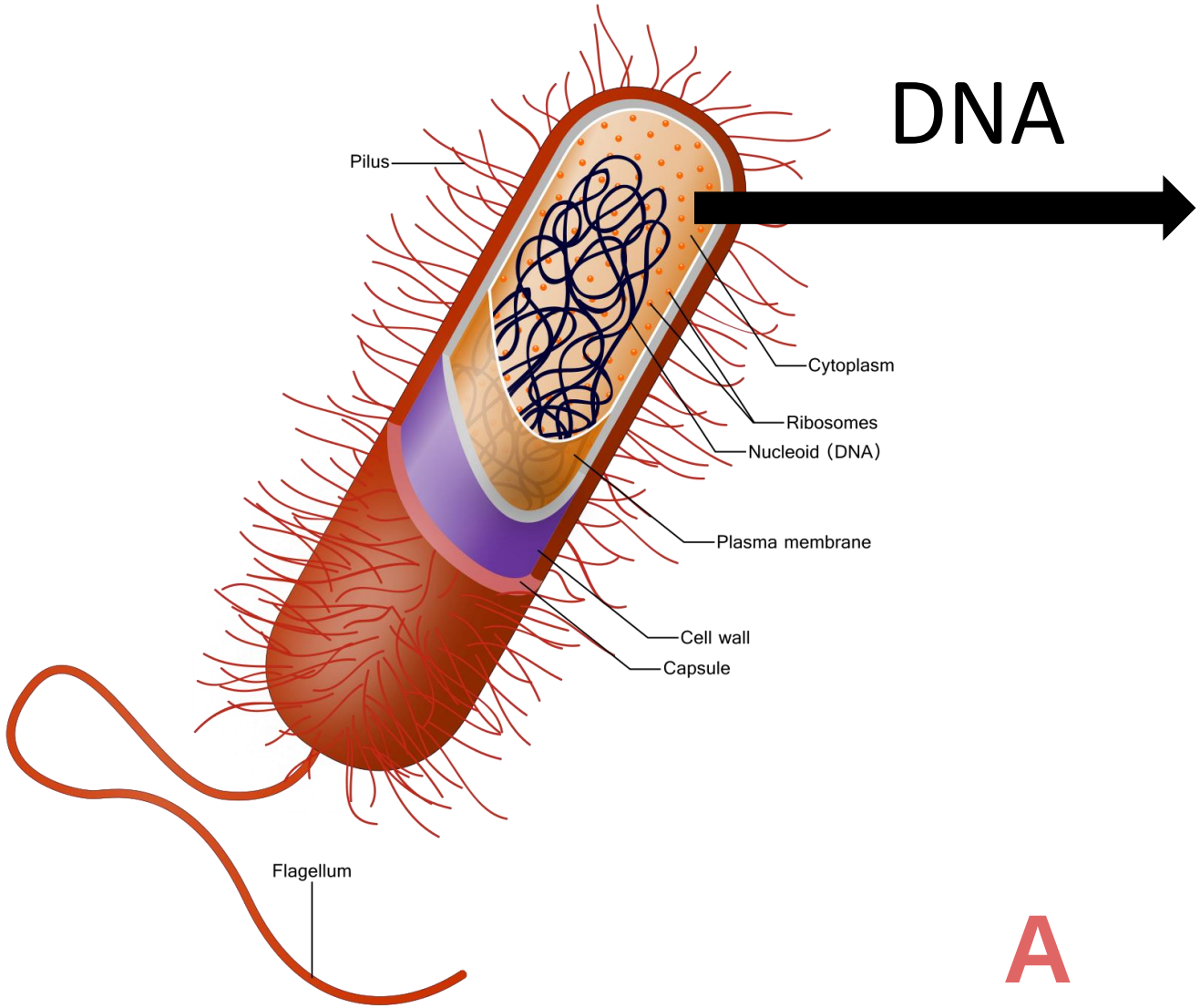
LEARNING OBJECTIVES

1. *What is microbial sequencing*
2. *What is genome assembly*
3. *Where to find sequence data*
4. *How to download*
5. *Quality of reads*
6. *How to assemble data*
7. *How to evaluate quality of the assembly*
8. *Data analysis tutorial*

LEARNING OBJECTIVES

1. *What is microbial sequencing*
2. *What is genome assembly*
3. *Where to find sequence data*
4. *How to download*
5. *Quality of reads*
6. *How to assemble data*
7. *How to evaluate quality of the assembly*
8. *Data analysis tutorial*

Microbial whole-genome sequencing



DNA

A T C G

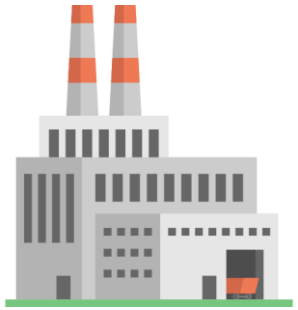
https://en.wikipedia.org/wiki/Prokaryote#/media/File:Prokaryote_cell.svg

https://en.wikipedia.org/wiki/DNA#/media/File:Eukaryote_DNA-en.svg

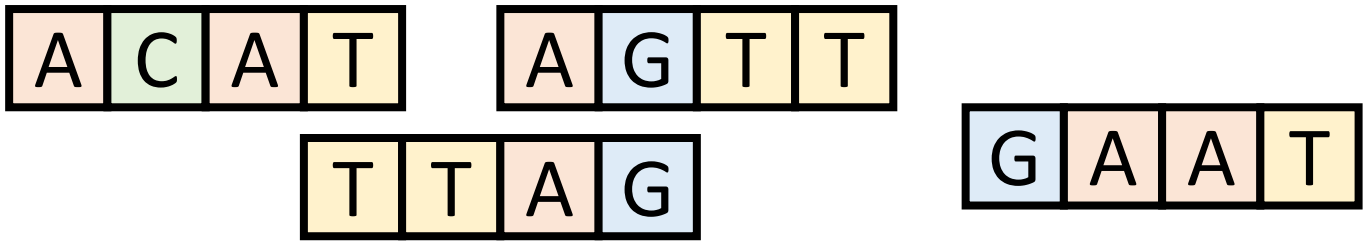
Microbial whole-genome sequencing

DNA

Sequencing



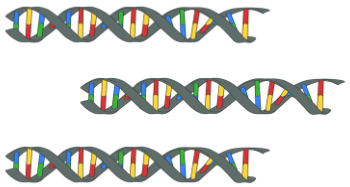
Reads

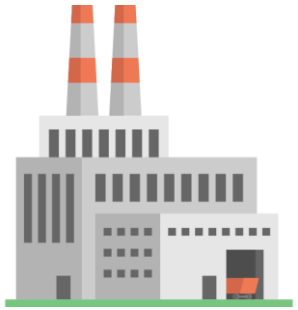


Microbial whole-genome sequencing

DNA

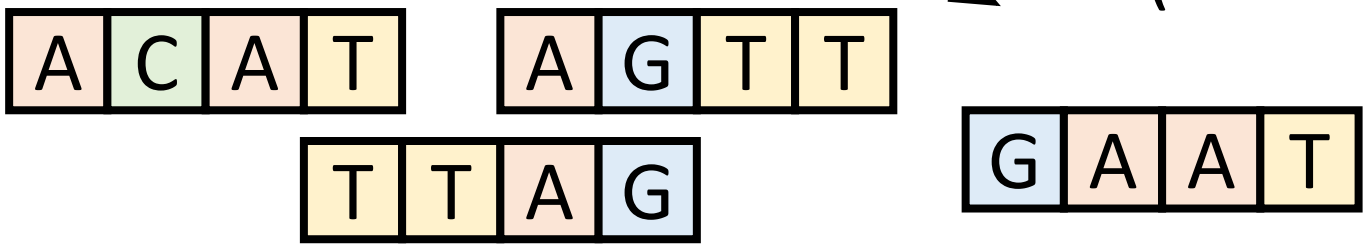
Fragment & Prepare DNA Library



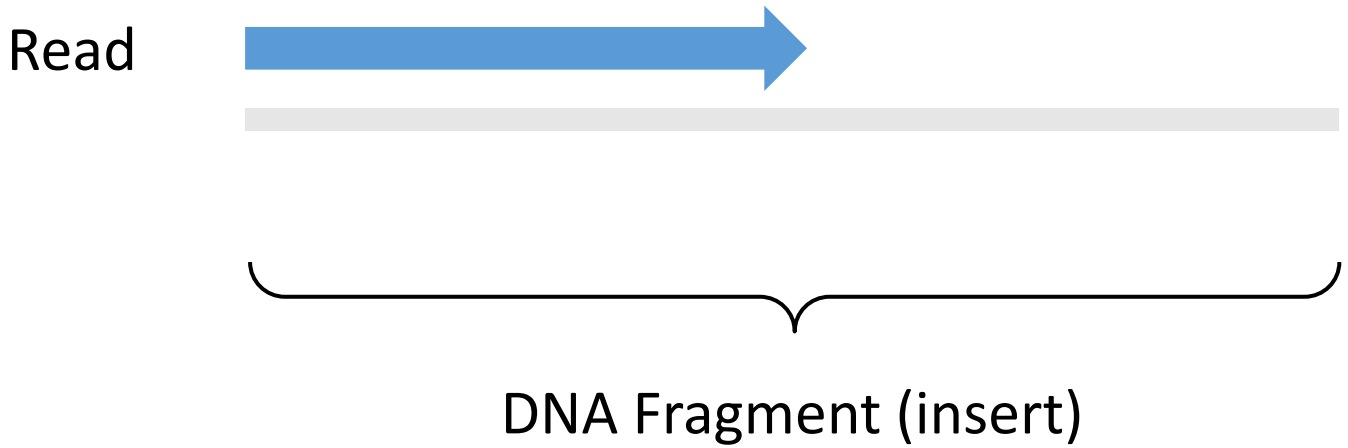


Sequencing (Illumina)

Reads

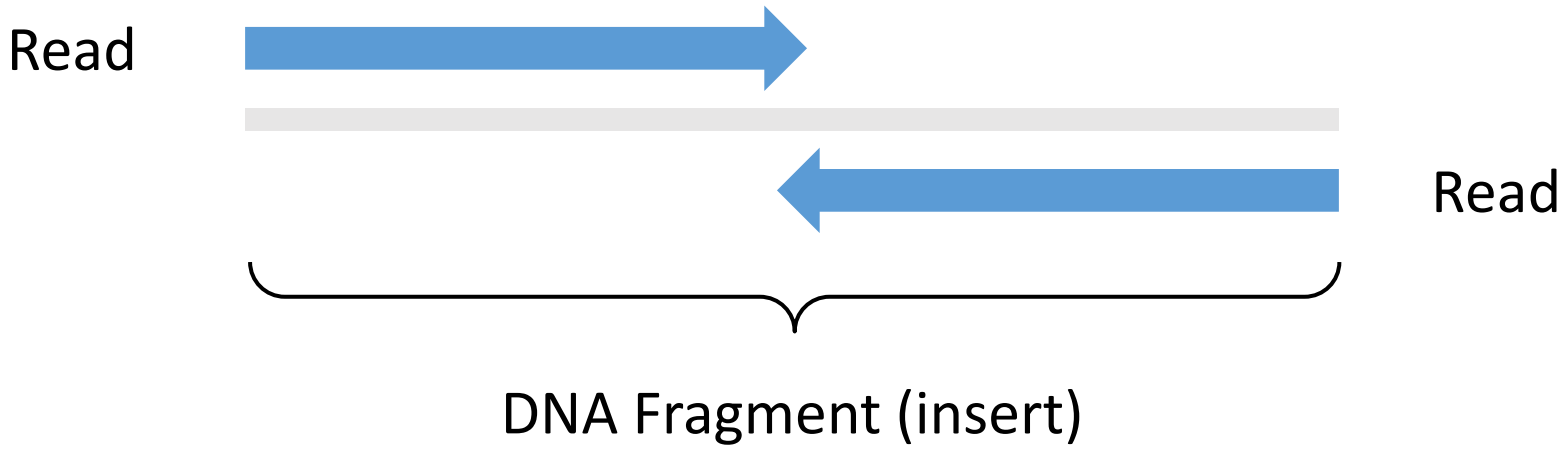


Single-end sequencing



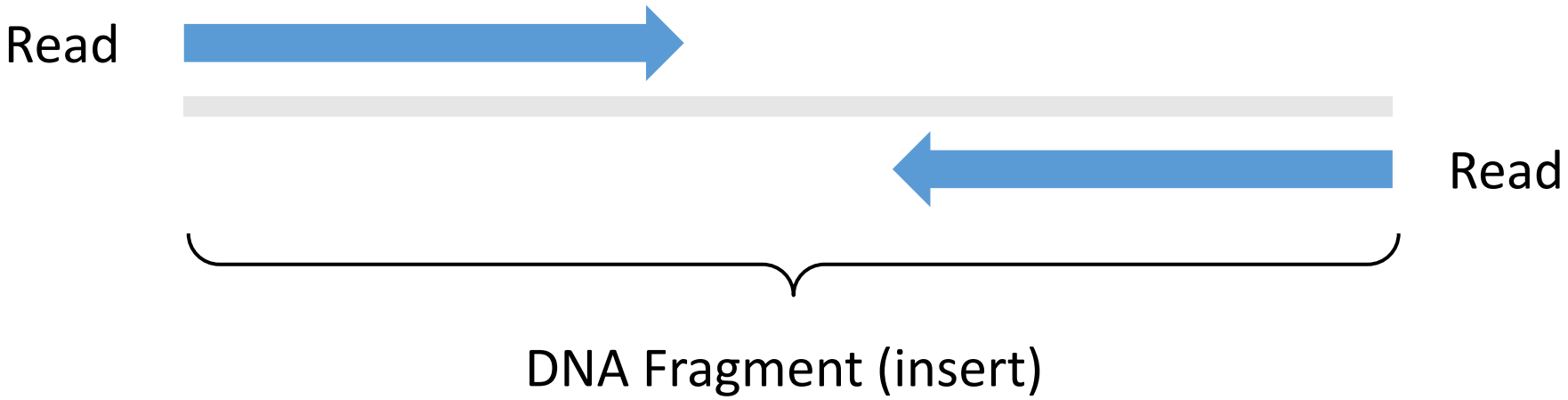
A DNA fragment is read once
from one single end.

Paired-end sequencing



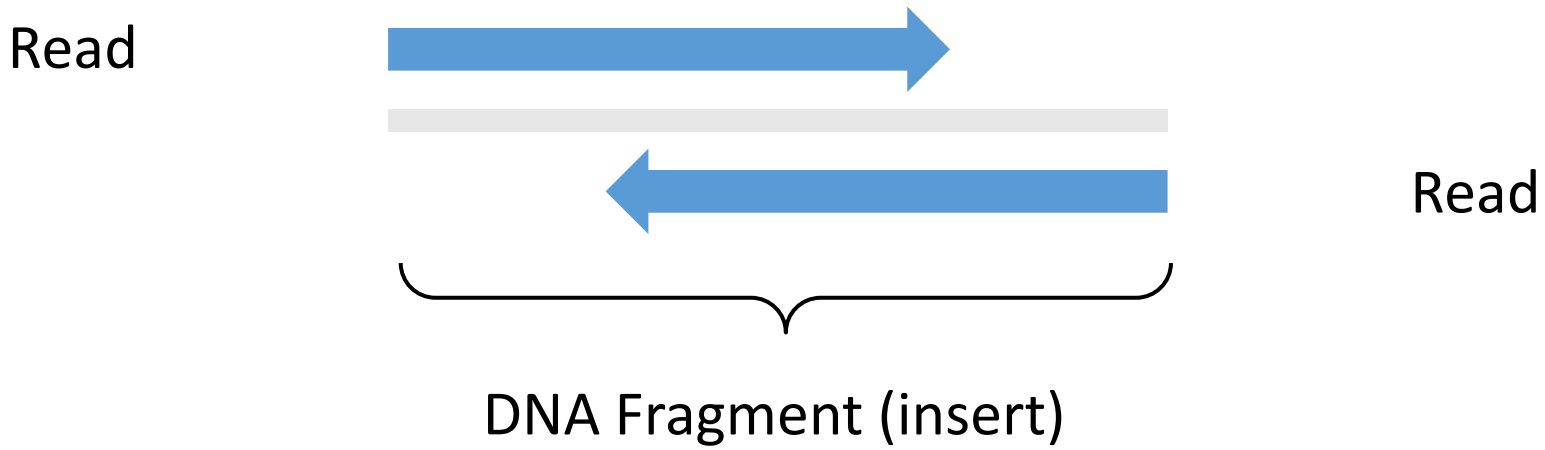
A DNA fragment is read twice,
from each end.

Paired-end sequencing



Fragment lengths can be long...

Paired-end sequencing



… or short.

LEARNING OBJECTIVES

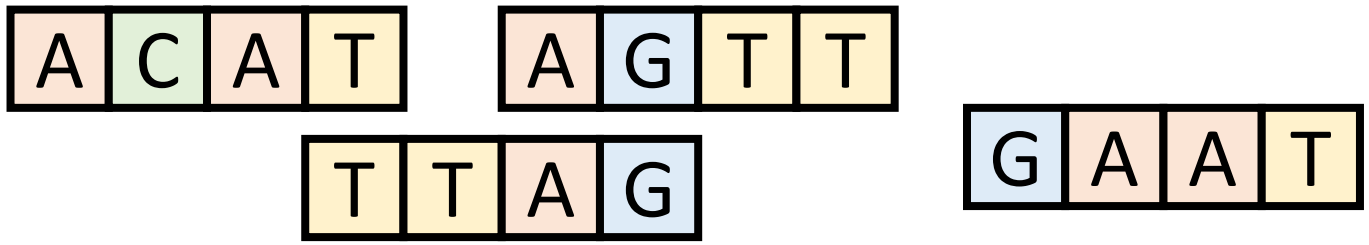
1. *What is microbial sequencing*
2. ***What is genome assembly***
3. *Where to find sequence data*
4. *How to download*
5. *Quality of reads*
6. *How to assemble data*
7. *How to evaluate quality of the assembly*
8. *Data analysis tutorial*

Genome assembly

https://en.wikipedia.org/wiki/Jigsaw_puzzle#/media/File:Sky_puzzle.jpg

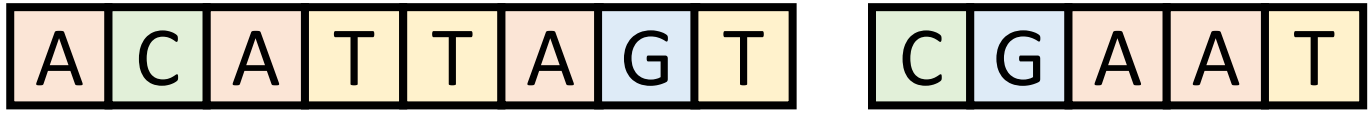
Genome assembly

Reads



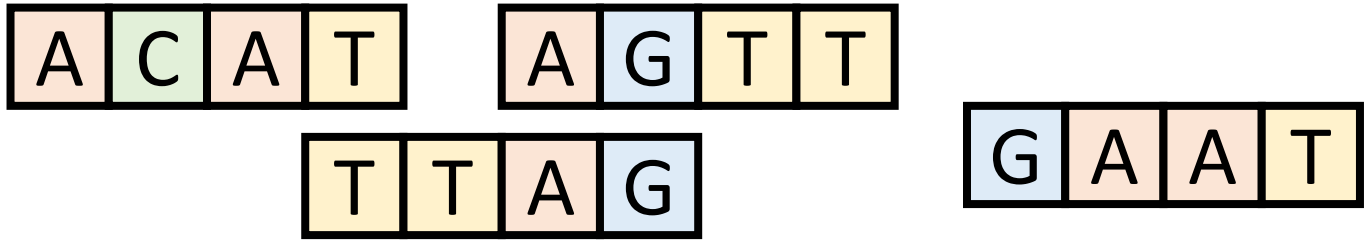
Transform reads to
larger contiguous fragments
(“contigs”)

Assembly



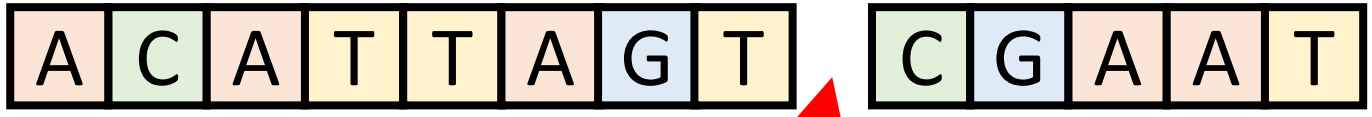
Genome assembly

Reads



Transform reads to
larger contiguous fragments
(“contigs”)

Assembly

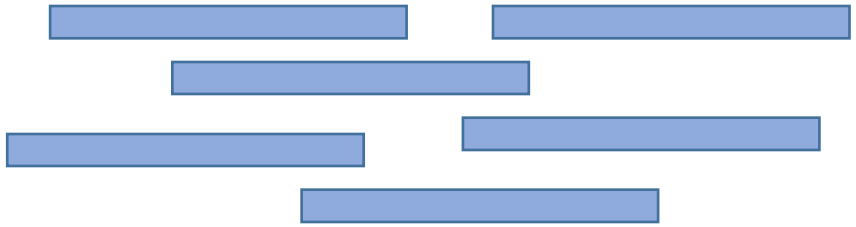


Breaks in contigs due to
repeats or missing data

Types of genome assembly

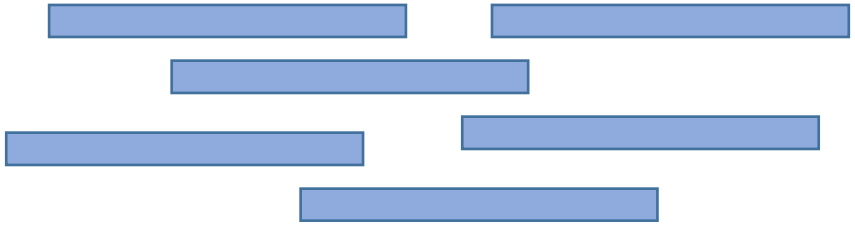
De novo assembly

Reads

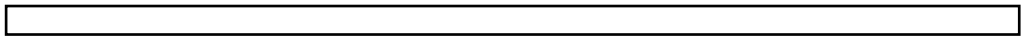


Contigs

Reference-guided assembly



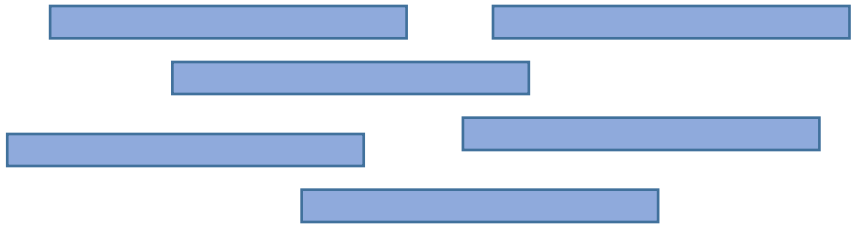
+



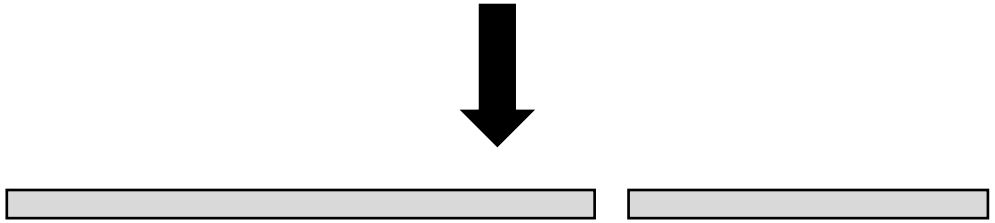
Reference genome

Types of genome assembly

De novo assembly

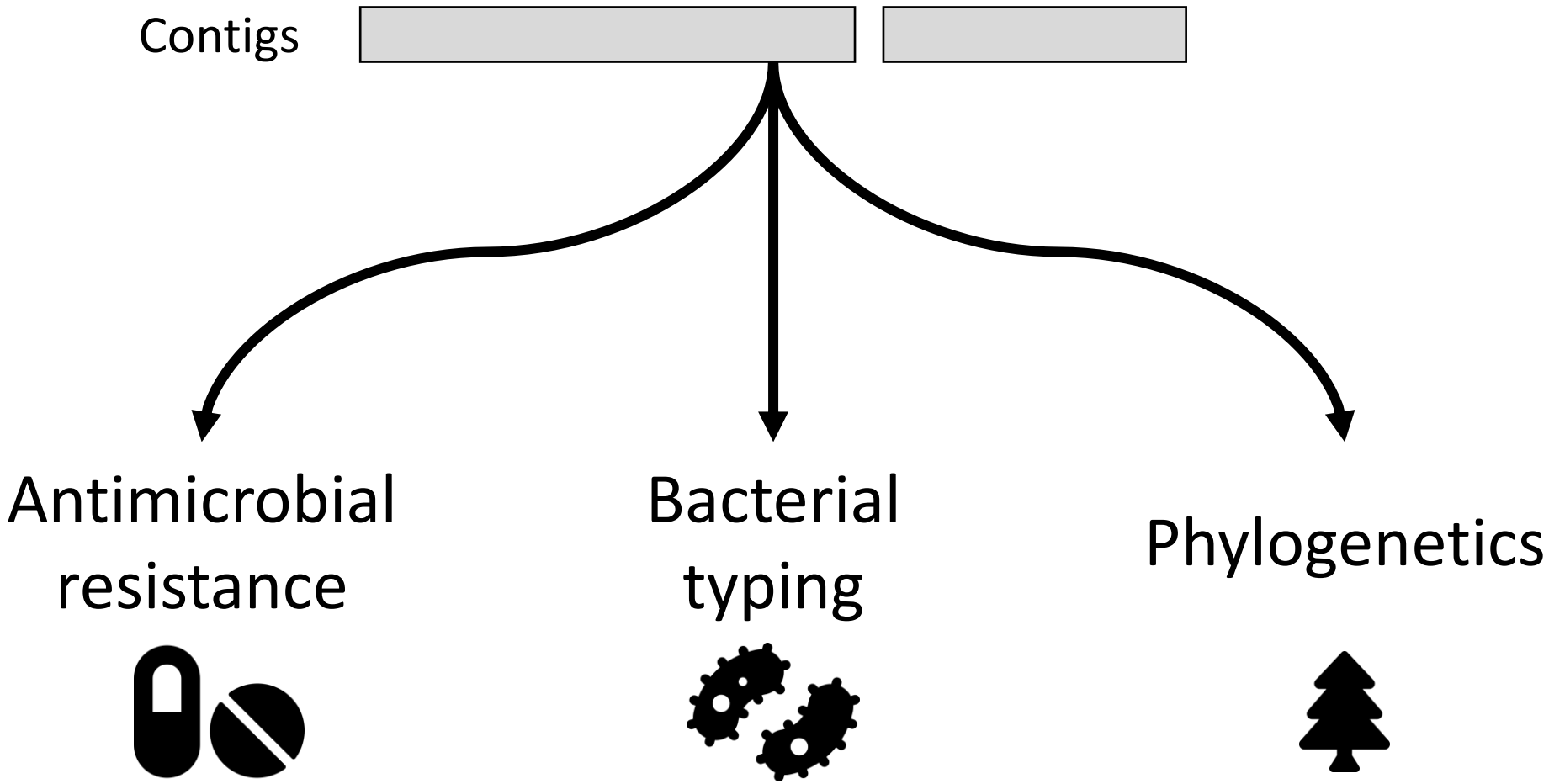


Focus on
de novo assembly



Reference-guided assembly

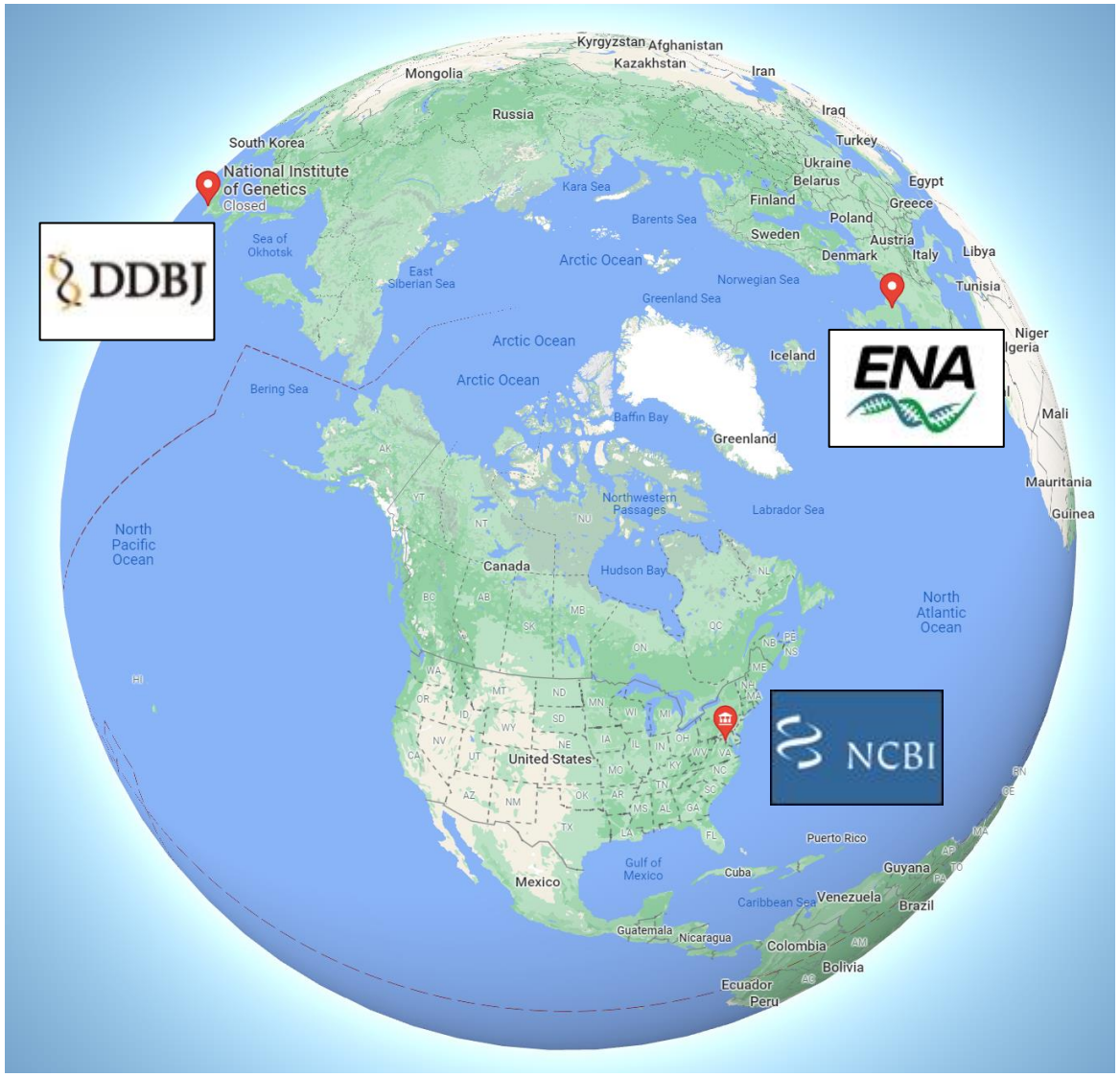
Purpose of assembly



LEARNING OBJECTIVES

1. *What is microbial sequencing*
2. *What is genome assembly*
3. ***Where to find sequence data***
4. *How to download*
5. *Quality of reads*
6. *How to assemble data*
7. *How to evaluate quality of the assembly*
8. *Data analysis tutorial*

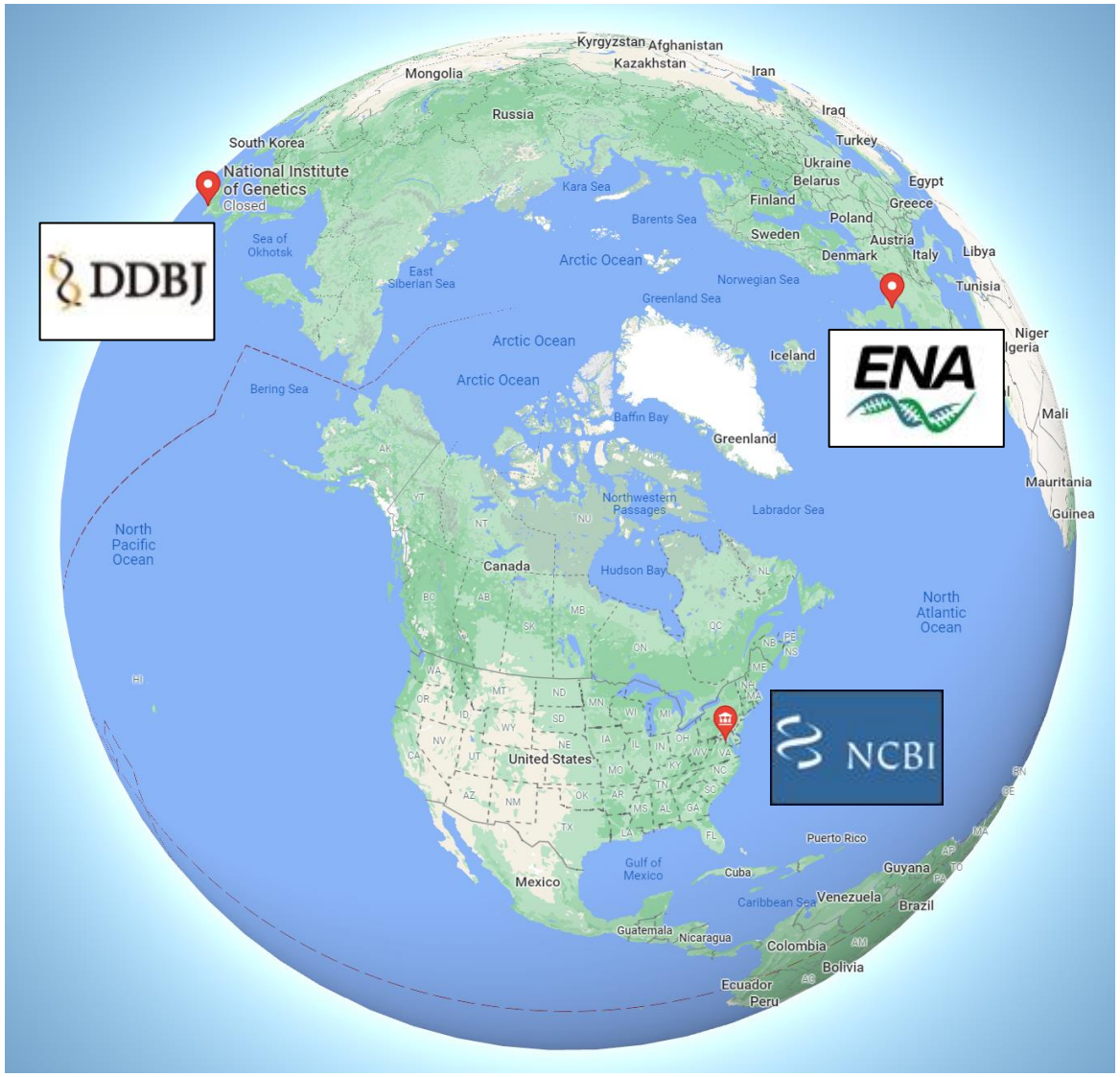
International Nucleotide Sequence Database Collaboration (INSDC)



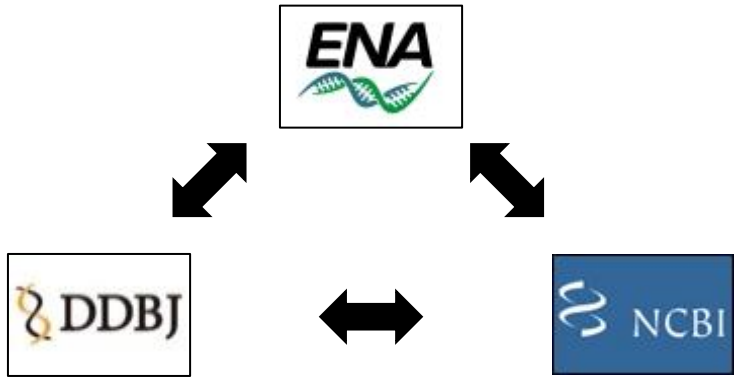
- Free and unrestricted access to nucleotide sequence and other data

https://www.insdc.org/

International Nucleotide Sequence Database Collaboration (INSDC)

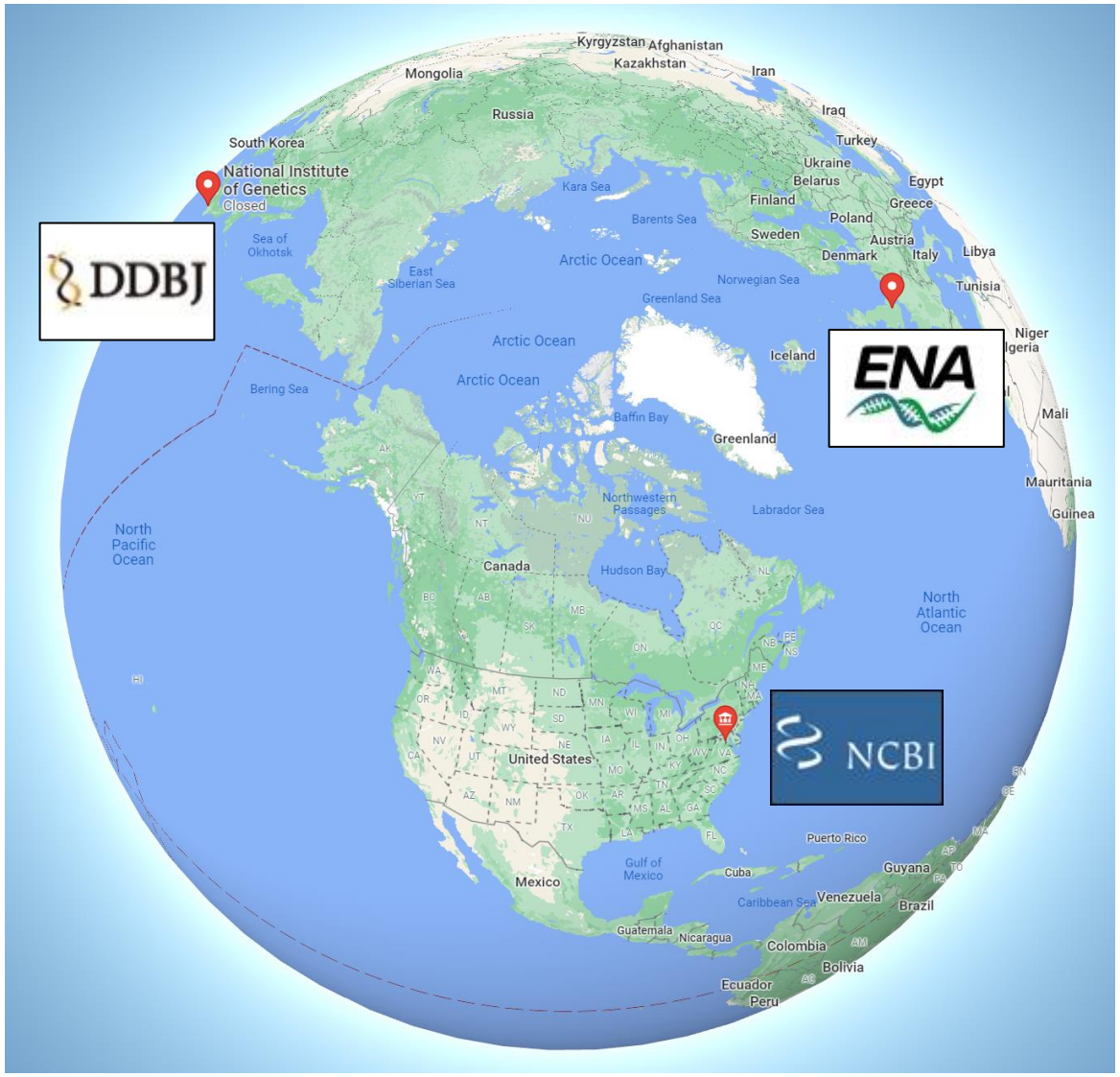


- Free and unrestricted access to nucleotide sequence and other data
- Data mirrored across three institutions

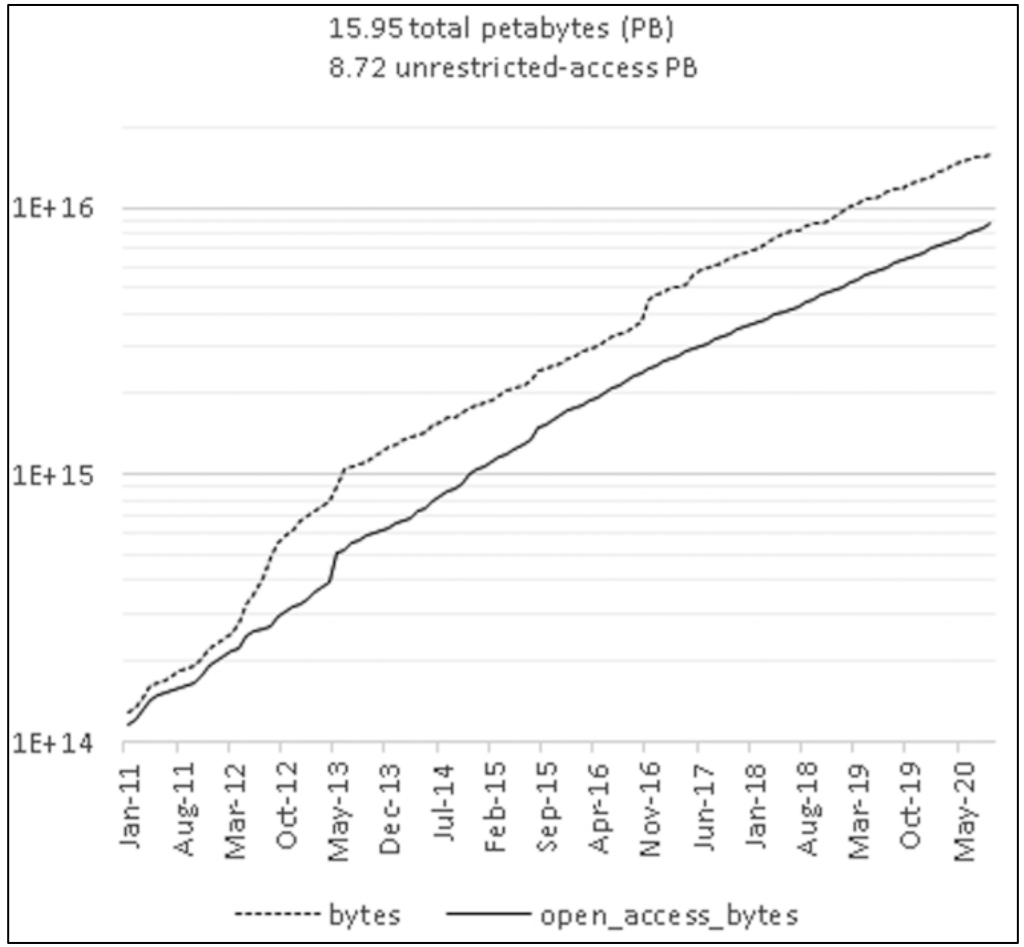


https://www.insdc.org/

International Nucleotide Sequence Database Collaboration (INSDC)



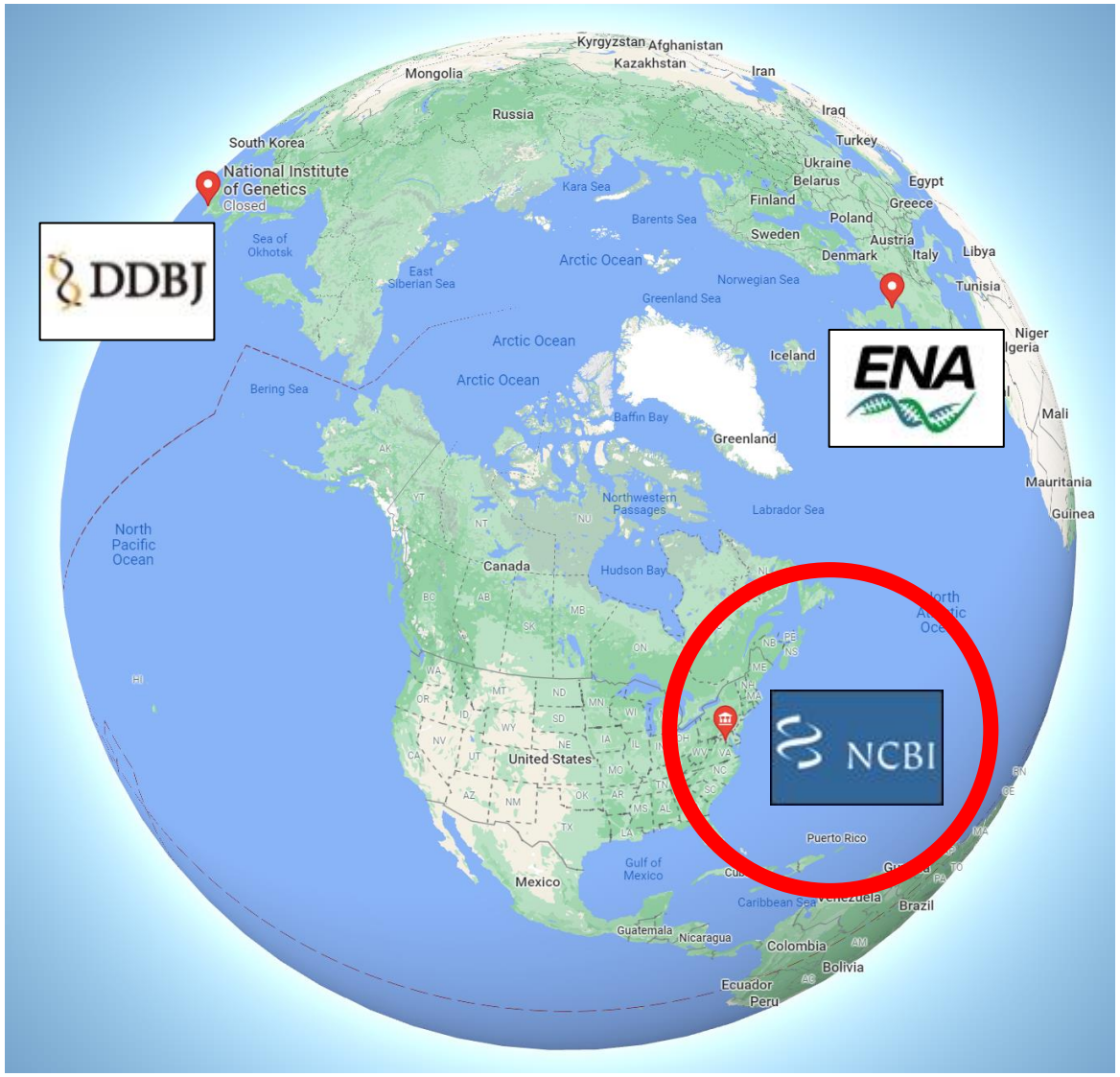
Petabytes (PB) of data



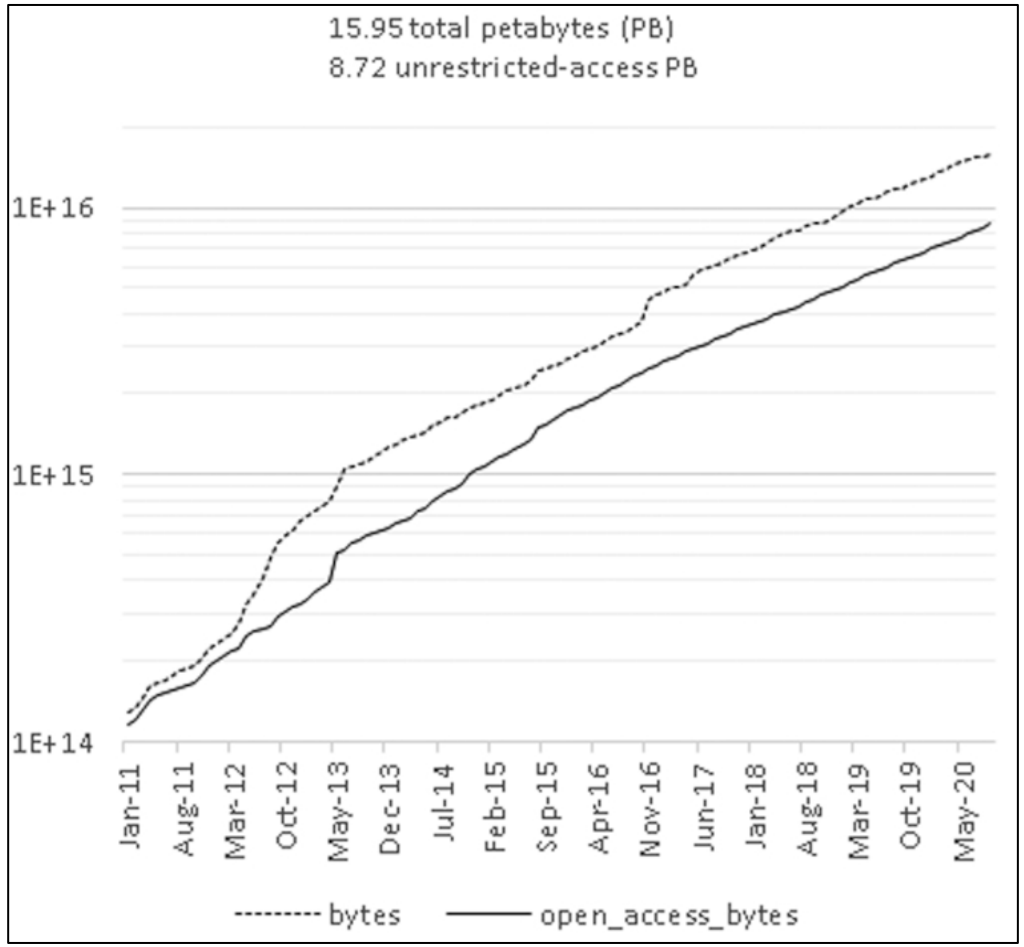
https://www.insdc.org/

Masanori Arita, et. Al. 2021. https://doi.org/10.1093/nar/gkaa967

International Nucleotide Sequence Database Collaboration (INSDC)



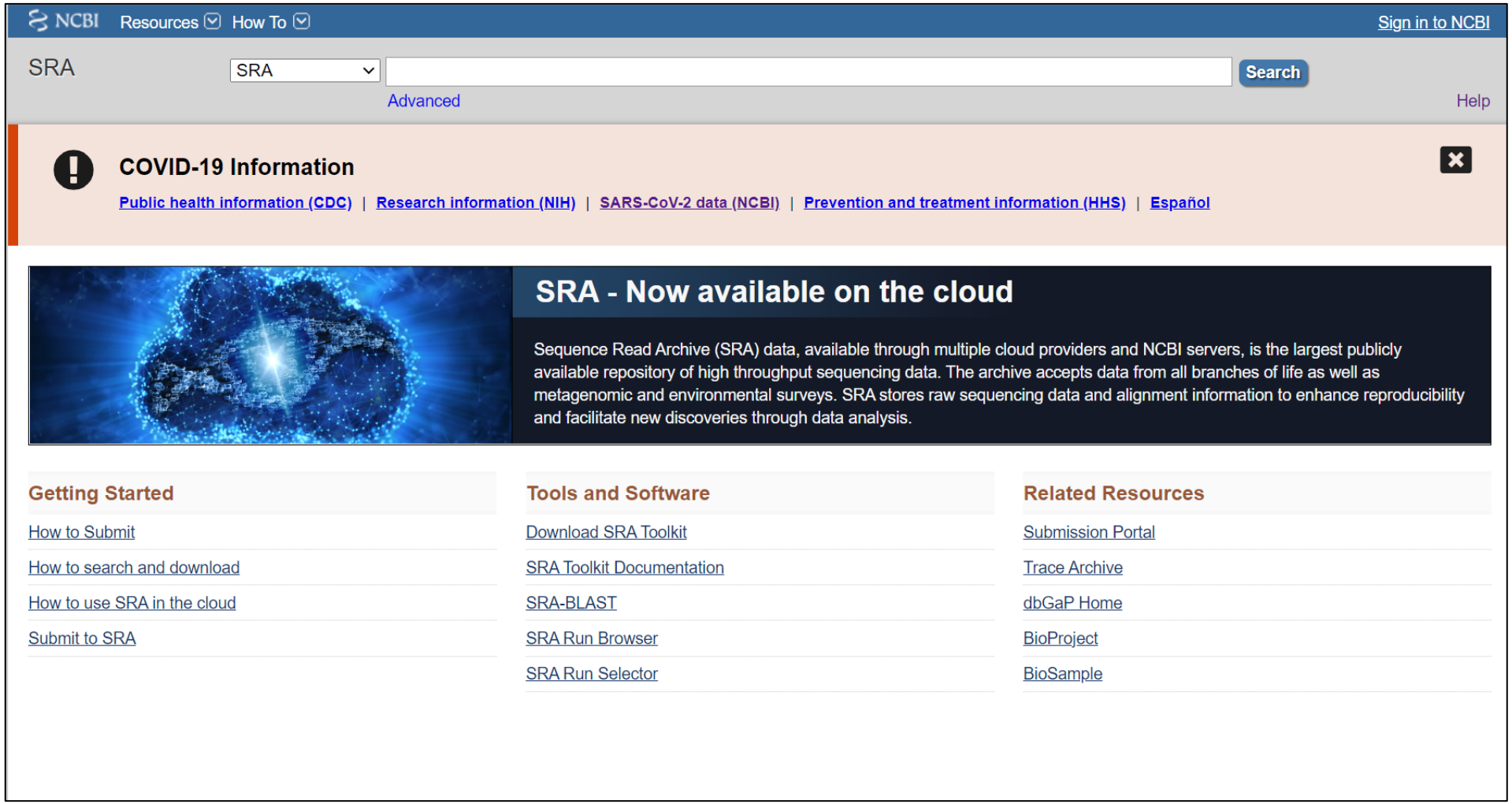
Petabytes (PB) of data



https://www.insdc.org/

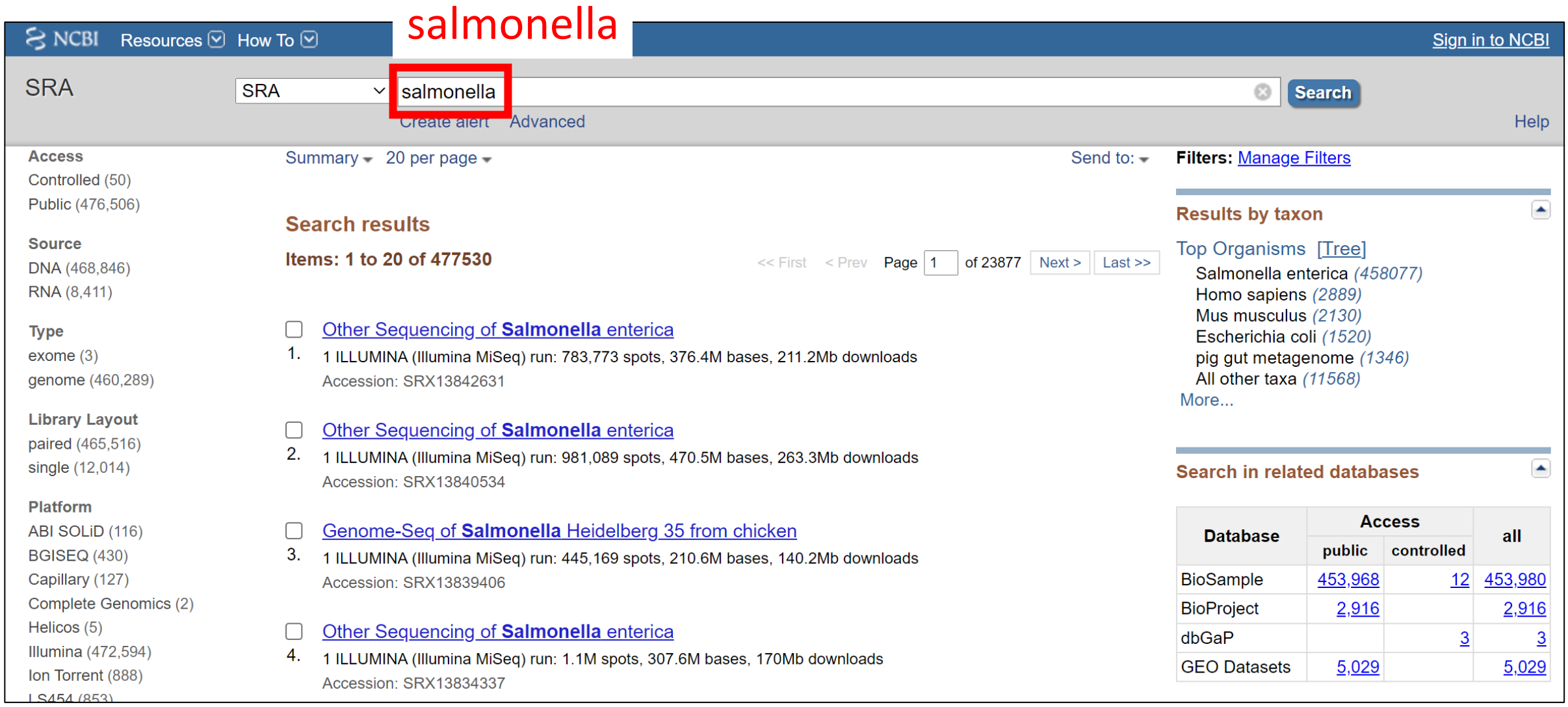
Masanori Arita, et. Al. 2021. https://doi.org/10.1093/nar/gkaa967

NCBI's Sequence Read Archive (SRA)



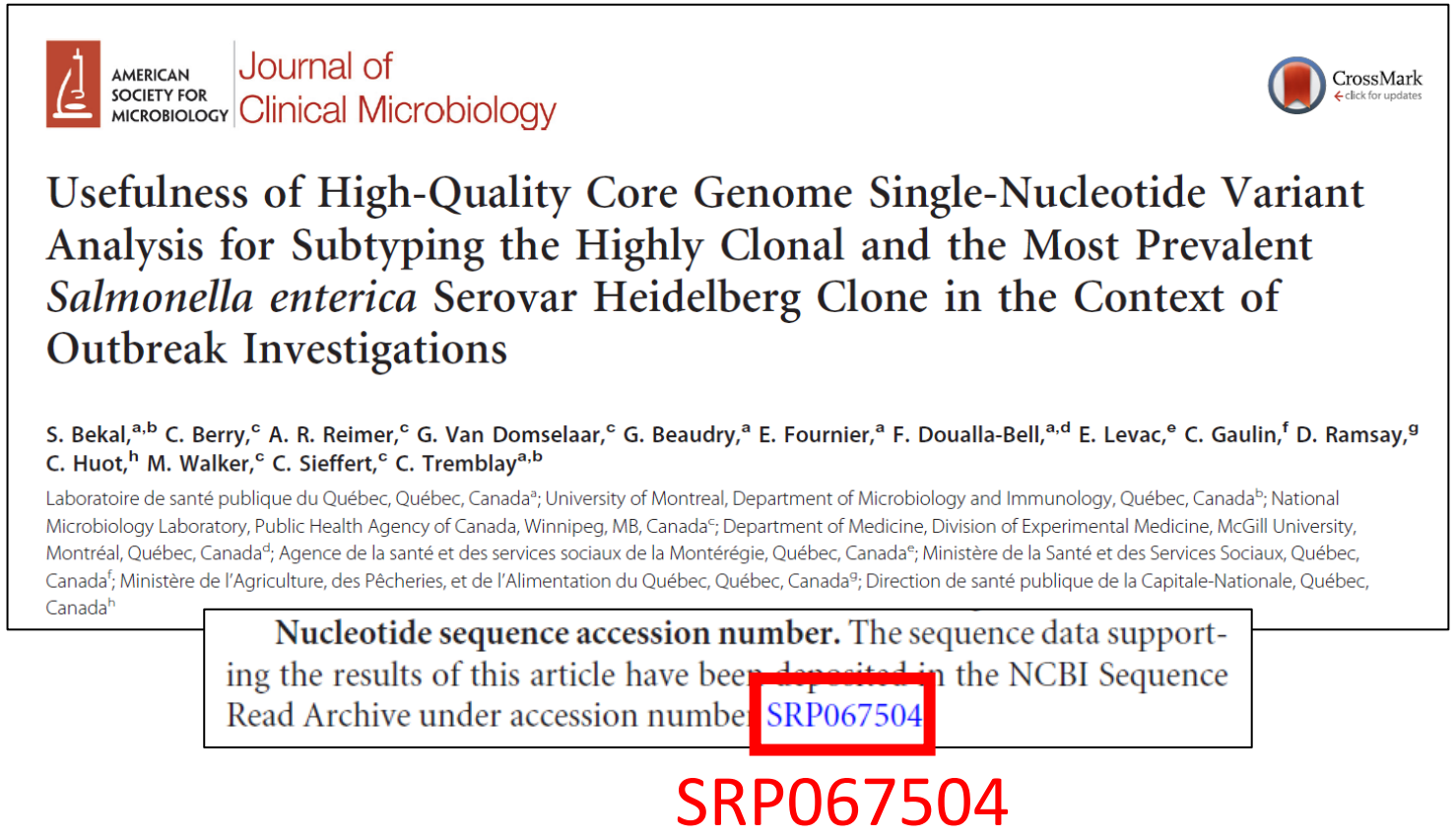
https://www.ncbi.nlm.nih.gov/sra/

Search by organism



Search by identifiers

Data identifiers in paper



AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Clinical Microbiology

CrossMark click for updates

Usefulness of High-Quality Core Genome Single-Nucleotide Variant Analysis for Subtyping the Highly Clonal and the Most Prevalent *Salmonella enterica* Serovar Heidelberg Clone in the Context of Outbreak Investigations

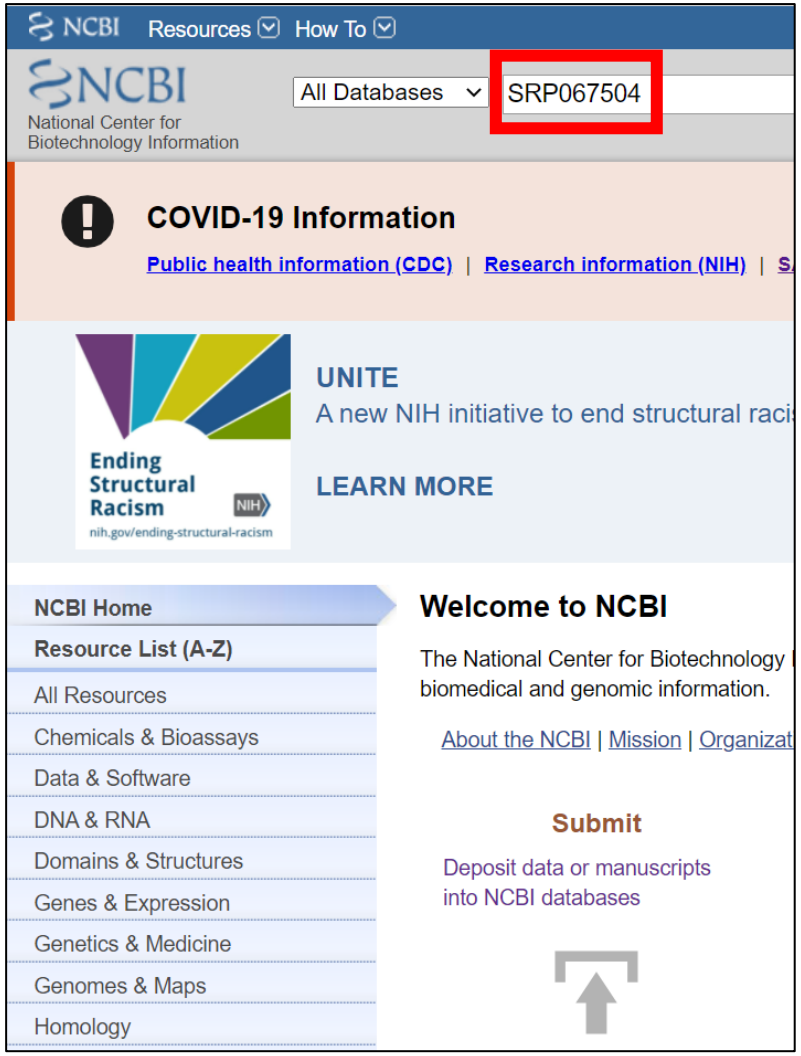
S. Bekal,[a,b] C. Berry,[c] A. R. Reimer,[c] G. Van Domselaar,[c] G. Beaudry,[a] E. Fournier,[a] F. Doualla-Bell,[a,d] E. Levac,[e] C. Gaulin,[f] D. Ramsay,[g] C. Huot,[h] M. Walker,[c] C. Sieffert,[c] C. Tremblay[a,b]

Laboratoire de santé publique du Québec, Québec, Canada[a]; University of Montreal, Department of Microbiology and Immunology, Québec, Canada[b]; National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, MB, Canada[c]; Department of Medicine, Division of Experimental Medicine, McGill University, Montréal, Québec, Canada[d]; Agence de la santé et des services sociaux de la Montérégie, Québec, Canada[e]; Ministère de la Santé et des Services Sociaux, Québec, Canada[f]; Ministère de l'Agriculture, des Pêcheries, et de l'Alimentation du Québec, Québec, Canada[g]; Direction de santé publique de la Capitale-Nationale, Québec, Canada[h]

Nucleotide sequence accession number. The sequence data supporting the results of this article have been deposited in the NCBI Sequence Read Archive under accession number SRP067504

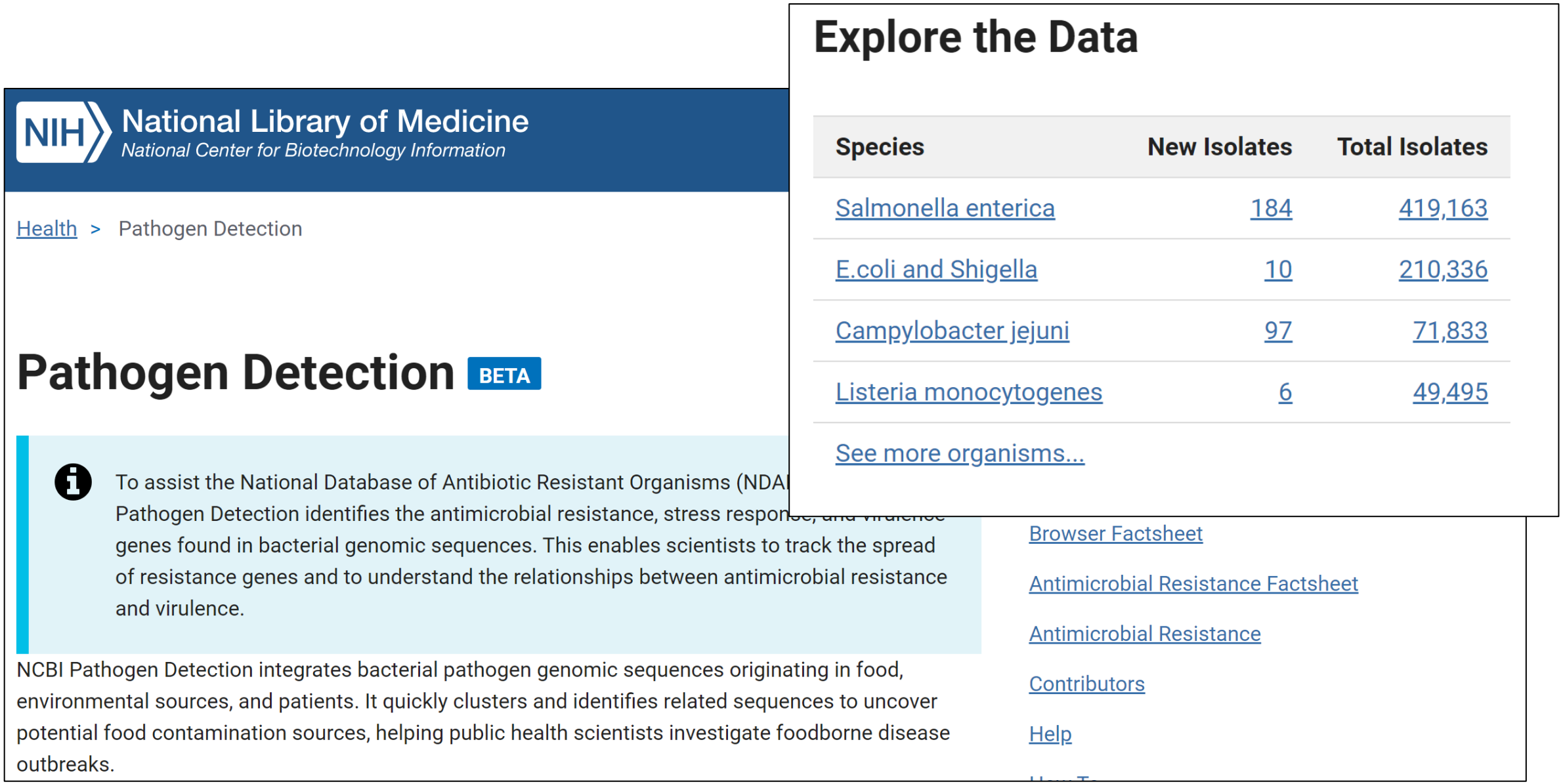
SRP067504

Search by identifiers



Targeted repositories

www.ncbi.nlm.nih.gov/pathogens



National Library of Medicine
National Center for Biotechnology Information

Health > Pathogen Detection

Pathogen Detection BETA

To assist the National Database of Antibiotic Resistant Organisms (NDA
Pathogen Detection identifies the antimicrobial resistance, stress respons
genes found in bacterial genomic sequences. This enables scientists to track the spread of resistance genes and to understand the relationships between antimicrobial resistance and virulence.

NCBI Pathogen Detection integrates bacterial pathogen genomic sequences originating in food, environmental sources, and patients. It quickly clusters and identifies related sequences to uncover potential food contamination sources, helping public health scientists investigate foodborne disease outbreaks.

Explore the Data

Species	New Isolates	Total Isolates
Salmonella enterica	184	419,163
E.coli and Shigella	10	210,336
Campylobacter jejuni	97	71,833
Listeria monocytogenes	6	49,495

See more organisms...

Browser Factsheet

Antimicrobial Resistance Factsheet

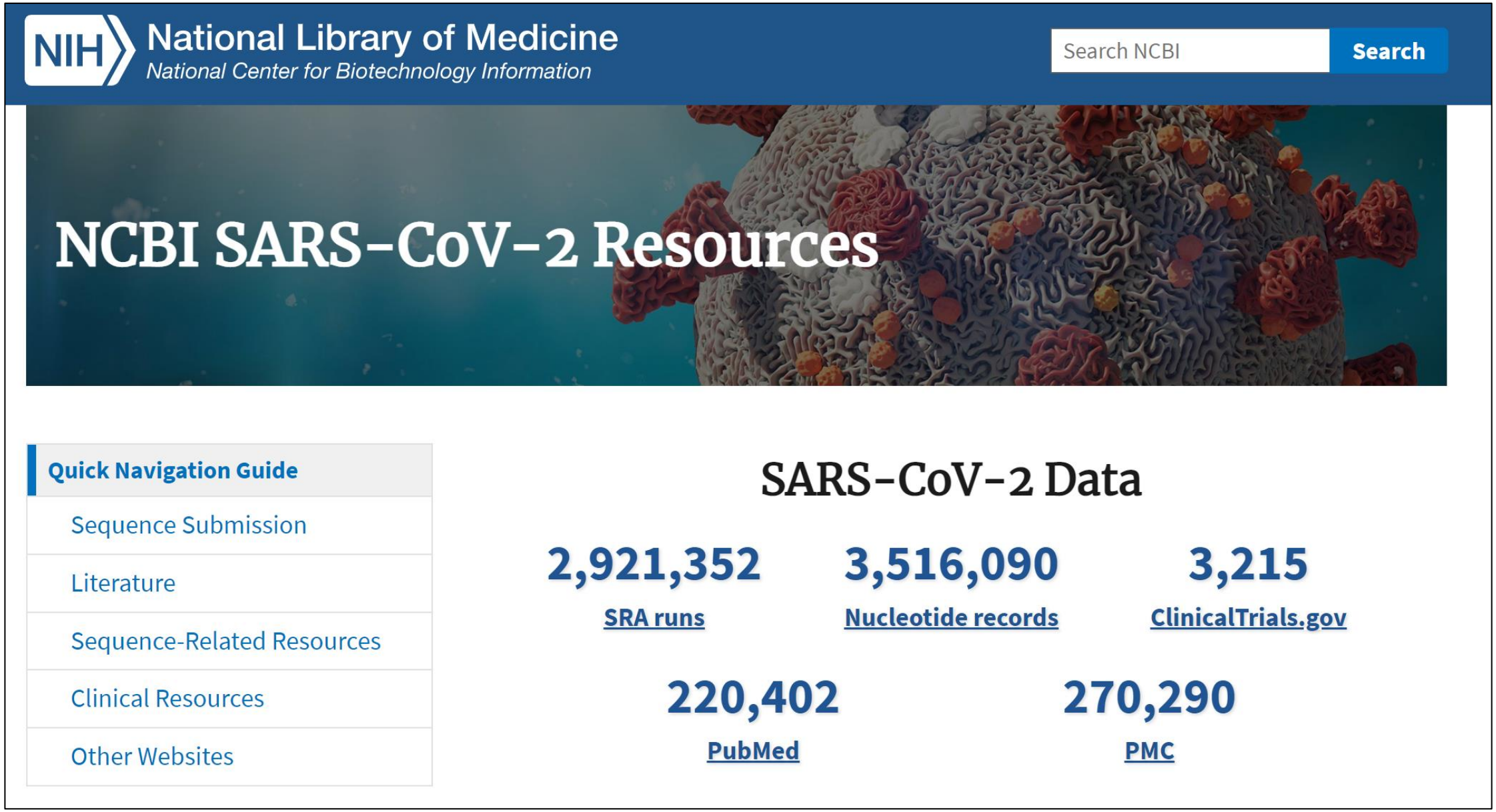
Antimicrobial Resistance

Contributors

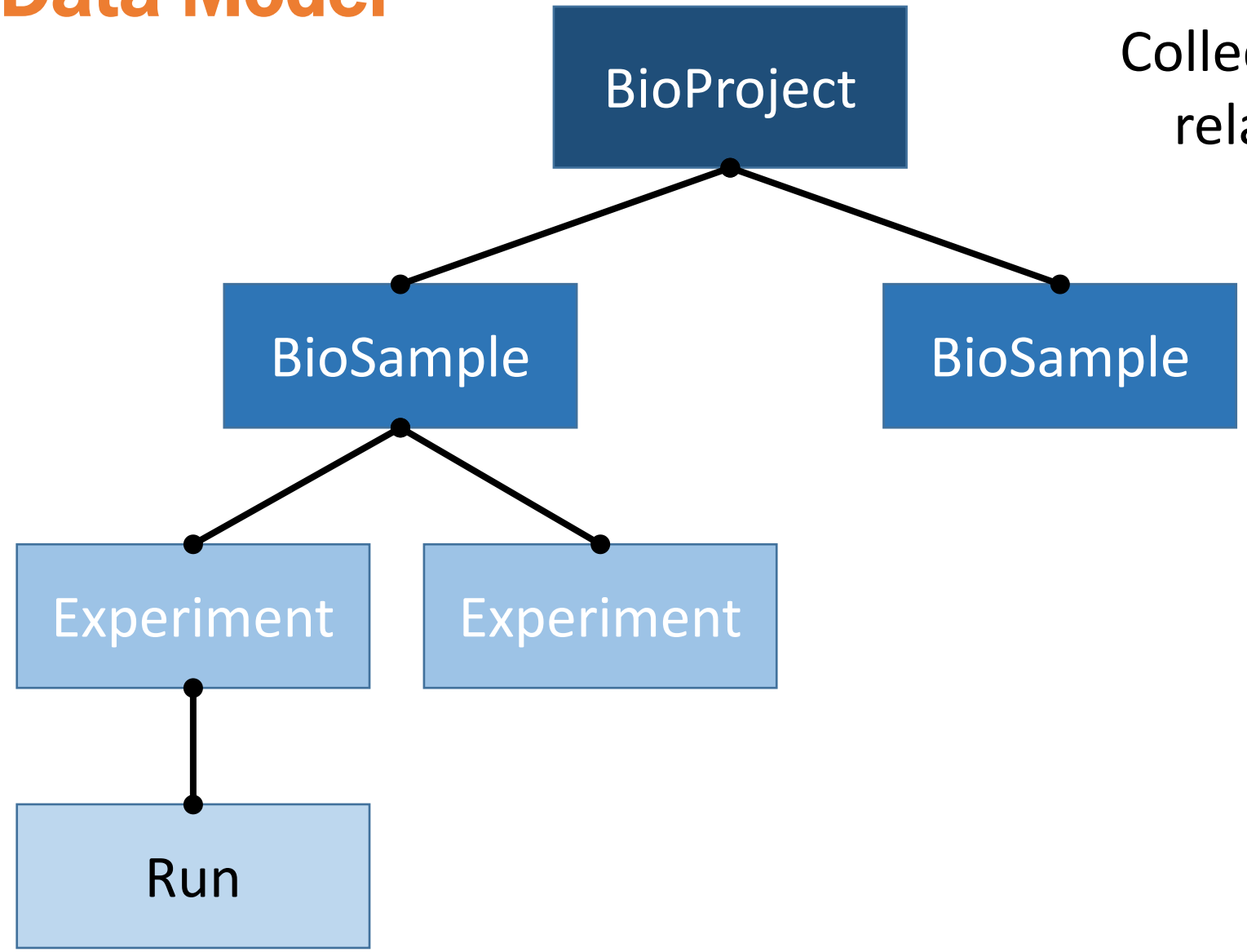
Help

Targeted repositories

www.ncbi.nlm.nih.gov/sars-cov-2/



SRA Data Model



Collection of biological data related to a single project

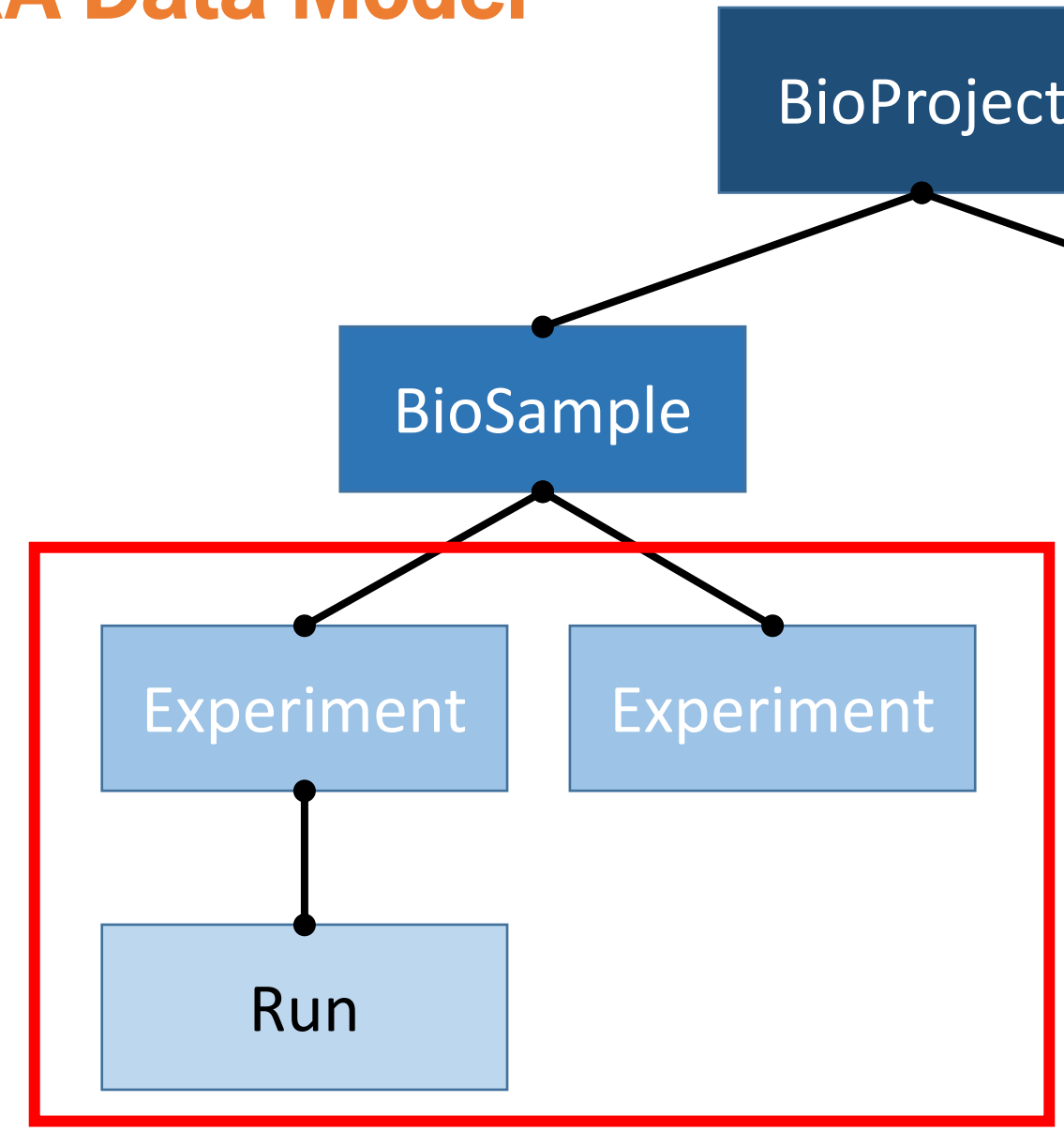
Record of a biological isolate

Unique sequencing result

Sequencing data

www.ncbi.nlm.nih.gov/sra/docs/submitmeta/

SRA Data Model



Collection of biological data related to a single project

Record of a biological isolate

Unique sequencing result

Sequencing data

www.ncbi.nlm.nih.gov/sra/docs/submitmeta/

SRA Experiment Record

SRX1489277: WGS of Salmonella enterica subsp. enterica serovar Heidelberg: SH08-001
1 ILLUMINA (Illumina MiSeq) run: 824,262 spots, 354.1M bases, 188.3Mb downloads

Design: "Genomic DNA was extracted using the Epicentre Metagenomic DNA Isolation Kit for Water on bacteria cultured overnight in BHI broth and sample libraries were prepared using MiSeq Nextera XT library preparation kit (Illumina, Inc., San Diego, CA, USA)."

Submitted by: McGill University

Study: Salmonella serovar Heidelberg genome sequencing
PRJNA305824 • SRP067504 • All experiments • All runs
show Abstract

Sample:
SAMN04334683 • SRS1211375 • All experiments • All runs
Organism: Salmonella enterica subsp. enterica serovar Heidelberg

Library:
Name: SH08-001
Instrument: Illumina MiSeq
Strategy: WGS
Source: GENOMIC
Selection: RANDOM
Layout: PAIRED

Spot descriptor:

Runs: 1 run, 824,262 spots, 354.1M bases, 188.3Mb

Run	# of Spots	# of Bases	Size	Published
SRR3028792	824,262	354.1M	188.3Mb	2015-12-19

www.ncbi.nlm.nih.gov/sra/SRX1489277

SRA Experiment Record

Library information (Illumina, paired-end)

www.ncbi.nlm.nih.gov/sra/SRX1489277

SRX1489277: WGS of Salmonella enterica subsp. enterica serovar Heidelberg: SH08-001
1 ILLUMINA (Illumina MiSeq) run: 824,262 spots, 354.1M bases, 188.3Mb downloads

Design: "Genomic DNA was extracted using the Epicentre Metagenomic DNA Isolation Kit for Water on bacteria cultured overnight in BHI broth and sample libraries were prepared using MiSeq Nextera XT library preparation kit (Illumina, Inc., San Diego, CA, USA)."

Submitted by: McGill University

Study: Salmonella serovar Heidelberg genome sequencing
PRJNA305824 • SRP067504 • All experiments • All runs
show Abstract

Sample:
SAMN04334683 • SRS1211375 • All experiments • All runs
Organism: Salmonella enterica subsp. enterica serovar Heidelberg

Library:
Name: SH08-001
Instrument: Illumina MiSeq
Strategy: WGS
Source: GENOMIC
Selection: RANDOM
Layout: PAIRED

Spot descriptor:

Runs: 1 run, 824,262 spots, 354.1M bases, 188.3Mb

Run	# of Spots	# of Bases	Size	Published
SRR3028792	824,262	354.1M	188.3Mb	2015-12-19

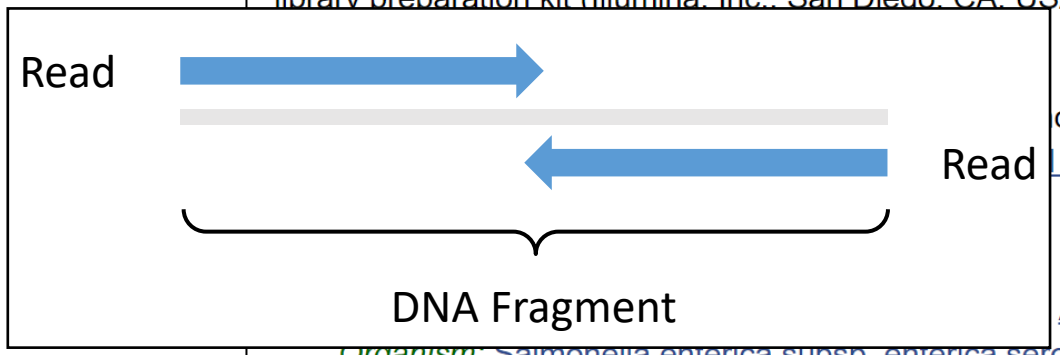
SRA Experiment Record

Library information (Illumina, paired-end)

www.ncbi.nlm.nih.gov/sra/SRX1489277

SRX1489277: WGS of Salmonella enterica subsp. enterica serovar Heidelberg: SH08-001
1 ILLUMINA (Illumina MiSeq) run: 824,262 spots, 354.1M bases, 188.3Mb downloads

Design: "Genomic DNA was extracted using the Epicentre Metagenomic DNA Isolation Kit for Water on bacteria cultured overnight in BHI broth and sample libraries were prepared using MiSeq Nextera XT library preparation kit (Illumina, Inc., San Diego, CA, USA)."



cing
l runs

All runs

Organism: Salmonella enterica subsp. enterica serovar Heidelberg

Library:
Name: SH08-001
Instrument: Illumina MiSeq
Strategy: WGS
Source: GENOMIC
Selection: RANDOM
Layout: PAIRED

Spot descriptor:

Runs: 1 run, 824,262 spots, 354.1M bases, 188.3Mb

Run	# of Spots	# of Bases	Size	Published
SRR3028792	824,262	354.1M	188.3Mb	2015-12-19

SRA Experiment Record

Experiment

BioProject

BioSample

Run

SRX1489277: WG of Salmonella enterica subsp. enterica serovar Heidelberg: SH08-001
1 ILLUMINA (Illumina MiSeq) run: 824,262 spots, 354.1M bases, 188.3Mb downloads

Design: "Genomic DNA was extracted using the Epicentre Metagenomic DNA Isolation Kit for Water on bacteria cultured overnight in BHI broth and sample libraries were prepared using MiSeq Nextera XT library preparation kit (Illumina, Inc., San Diego, CA, USA)."

Submitted by: McGill University

Study: Salmonella erovar Heidelberg genome sequencing
PRJNA305824 • SRP067504 • All experiments • All runs
show Abstract

Sample:
SAMN04334683 • SRS1211375 • All experiments • All runs
Organism: Salmonella enterica subsp. enterica serovar Heidelberg

Library:
Name: SH08-001
Instrument: Illumina MiSeq
Strategy: WGS
Source: GENOMIC
Selection: RANDOM
Layout: PAIRED

Spot descriptor:

Runs: 1 run, 824,262 spots, 354.1M bases, 188.3Mb

Run	# of Spots	# of Bases	Size	Published
SRR3028792	824,262	354.1M	188.3Mb	2015-12-19

SRA Experiment Record

Experiment

BioProject

BioSample

Run

SRX1489277: WG of Salmonella enterica subsp. enterica serovar Heidelberg: SH08-001
1 ILLUMINA (Illumina MiSeq) run: 824,262 spots, 354.1M bases, 188.3Mb downloads

Design: "Genomic DNA was extracted using the Epicentre Metagenomic DNA Isolation Kit for Water on bacteria cultured overnight in BHI broth and sample libraries were prepared using MiSeq Nextera XT library preparation kit (Illumina, Inc., San Diego, CA, USA)."

Submitted by: McGill University

Study: Salmonella erovar Heidelberg genome sequencing
PRJNA305824 • SRP067504 • All experiments • All runs
show Abstract

Sample:
SAMN04334683 • SRS1211375 • All experiments • All runs
Organism: Salmonella enterica subsp. enterica serovar Heidelberg

Library:
Name: SH08-001
Instrument: Illumina MiSeq
Strategy: WGS
Source: GENOMIC
Selection: RANDOM
Layout: PAIRED

Spot descriptor:

Runs: 1 run, 824,262 spots, 354.1M bases, 188.3Mb

Run	# of Spots	# of Bases	Size	Published
SRR3028792	824,262	354.1M	188.3Mb	2015-12-19

SRA Run

trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR3028792

WGS of Salmonella enterica subsp. enterica serovar Heidelberg: SH08-001 (SRR3028792)

Metadata | Analysis | Reads | Data access

Run	Spots	Bases	Size	GC content	Published	Access Type
SRR3028792	824.3k	354.1Mbp	197.5M	51.8%	2015-12-19	public

Quality graph (bigger)

This run has 2 reads per spot:

$\bar{L}$=214, σ=89.4, 100%	$\bar{L}$=215, σ=90.0, 100%

Legend

Experiment	Library Name	Platform	Strategy	Source	Selection	Layout	Action
SRX1489277	SH08-001	Illumina	WGS	GENOMIC	RANDOM	PAIRED	BLAST

Show design

Biosample	Sample Description	Organism	Links
SAMN04334683 (SRS1211375)		Salmonella enterica subsp. enterica serovar Heidelberg	PRJNA305824

Bioproject	SRA Study	Title
PRJNA305824	SRP067504	Salmonella serovar Heidelberg genome sequencing

Show abstract

SRA Run: Reads

trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR3028792

WGS of Salmonella enterica subsp. enterica serovar Heidelberg: SH08-001 (SRR3028792)

Metadata | Analysis | **Reads** | Data access

Filter: [ ] Find | Filtered Download | What does it do?

What can the filter be applied to?

< 1 [1] [82427] >

View: ☑ biological reads ☐ technical reads ☐ quality scores

1. SRR3028792.1 SRS1211375
name: 1, member: 8
2. SRR3028792.2 SRS1211375
name: 2, member: 8
3. SRR3028792.3 SRS1211375
name: 3, member: 8
4. SRR3028792.4 SRS1211375
name: 4, member: 8
5. SRR3028792.5 SRS1211375
name: 5, member: 8
6. SRR3028792.6 SRS1211375
name: 6, member: 8
7. SRR3028792.7 SRS1211375
name: 7, member: 8
8. SRR3028792.8 SRS1211375
name: 8, member: 8

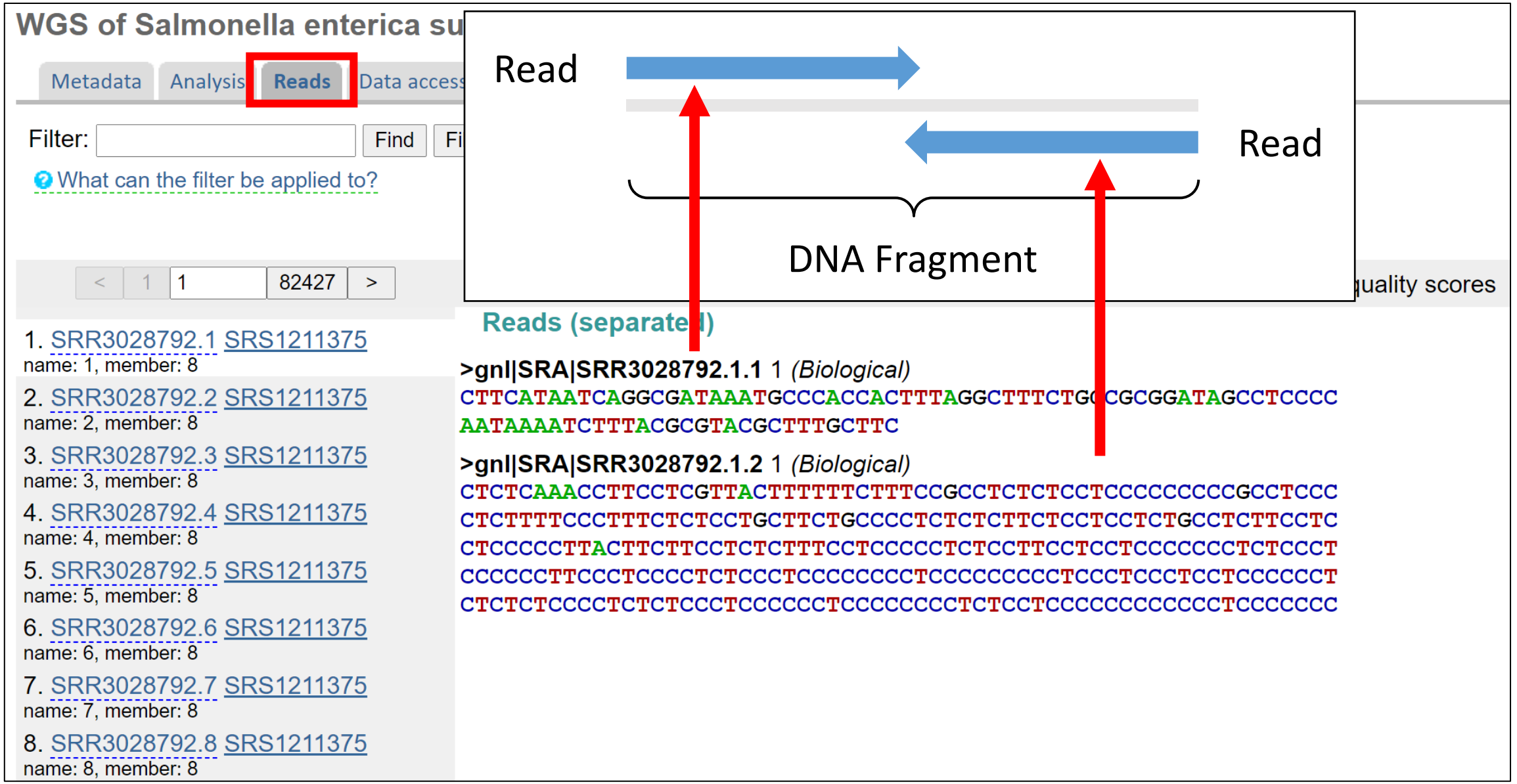
Reads (separated)

>gnl|SRA|SRR3028792.1.1 1 *(Biological)*
CTTCATAATCAGGCGATAAATGCCCACCACTTTAGGCTTTCTGGCGCGGATAGCCTCCCC
AATAAAATCTTTACGCGTACGCTTTGCTTC

>gnl|SRA|SRR3028792.1.2 1 *(Biological)*
CTCTCAAACCTTCCTCGTTACTTTTTTCTTTCCGCCTCTCTCCTCCCCCCCCCGCCTCCC
CTCTTTTCCCTTTCTCTCCTGCTTCTGCCCCTCTCTCTTCTCCTCCTCTGCCTCTTCCTC
CTCCCCCTTACTTCTTCCTCTCTTTCCTCCCCCTCTCCTTCCTCCTCCCCCCCTCTCCCT
CCCCCCTTCCCTCCCCTCTCCCTCCCCCCCCTCCCCCCCCCTCCCTCCCTCCTCCCCCCT
CTCTCTCCCCTCTCTCCCTCCCCCCTCCCCCCCCTCTCCTCCCCCCCCCCCCTCCCCCCC

SRA Run: Reads

trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR3028792



LEARNING OBJECTIVES

1. *What is microbial sequencing*
2. *What is genome assembly*
3. *Where to find sequence data*
4. ***How to download***
5. *Quality of reads*
6. *How to assemble data*
7. *How to evaluate quality of the assembly*
8. *Data analysis tutorial*

Finding an SRA run

WGS of Salmonella enterica subsp. enterica serovar Heidelberg: SH08-001

Metadata | Analysis | Reads | **Data access**

SRA archive data

SRA archive data is normalized by the SRA load process and used by the SRA Toolkit to read and produce formats like FASTQ, SAM, etc. The default toolkit configuration enables it to find and retrieve SRA runs by accession.

Public SRA files are now available from GCP and AWS cloud platforms as well as from NCBI. Access to most data in the cloud requires a user account with the cloud service provider. The user's account will incur costs for cloud compute or to copy data outside of the specified cloud service region.

Type	Size	Location	Name
run	192,858 Kb	GCP	gs://sra-pub-crun-2/SRR3028792/SRR3028792.1
		AWS	https://sra-pub-run-odp.s3.amazonaws.com/sra/SRR
		NCBI	https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos1/sr

SRA Toolkit

github.com/ncbi/sra-tools

- A suite of command-line tools for working with data from the sequence read archive

SRA Toolkit Documentation

SRA Toolkit Installation and Configuration Guide

Protected Data Usage Guide

Frequently Used Tools:

fastq-dump: Convert SRA data into fastq format

prefetch: Allows command-line downloading of SRA, dbGaP, and ADSP data

sam-dump: Convert SRA data to sam format

sra-pileup: Generate pileup statistics on aligned SRA data

vdb-config: Display and modify VDB configuration information

vdb-decrypt: Decrypt non-SRA dbGaP data ("phenotype data")

SRA Toolkit

github.com/ncbi/sra-tools

- A suite of command-line tools for working with data from the sequence read archive

SRA Toolkit Documentation

SRA Toolkit Installation and Configuration Guide

Protected Data Usage Guide

Frequently Used Tools:

fastq-dump: Convert SRA data into fastq format

prefetch: Allows command-line downloading of SRA, dbGaP, and ADSP data

sam-dump: Convert SRA data to sam format

sra-pileup: Generate pileup statistics on aligned SRA data

vdb-config: Display and modify VDB configuration information

vdb-decrypt: Decrypt non-SRA dbGaP data ("phenotype data")

SRA Toolkit

github.com/ncbi/sra-tools

- **Installation (conda)**

```
$ conda create -n sra-tools sra-tools
$ conda activate sra-tools
```

- **Verify tools are available**

```
$ prefetch --version

"prefetch" version 2.11.0
```

Prefetch

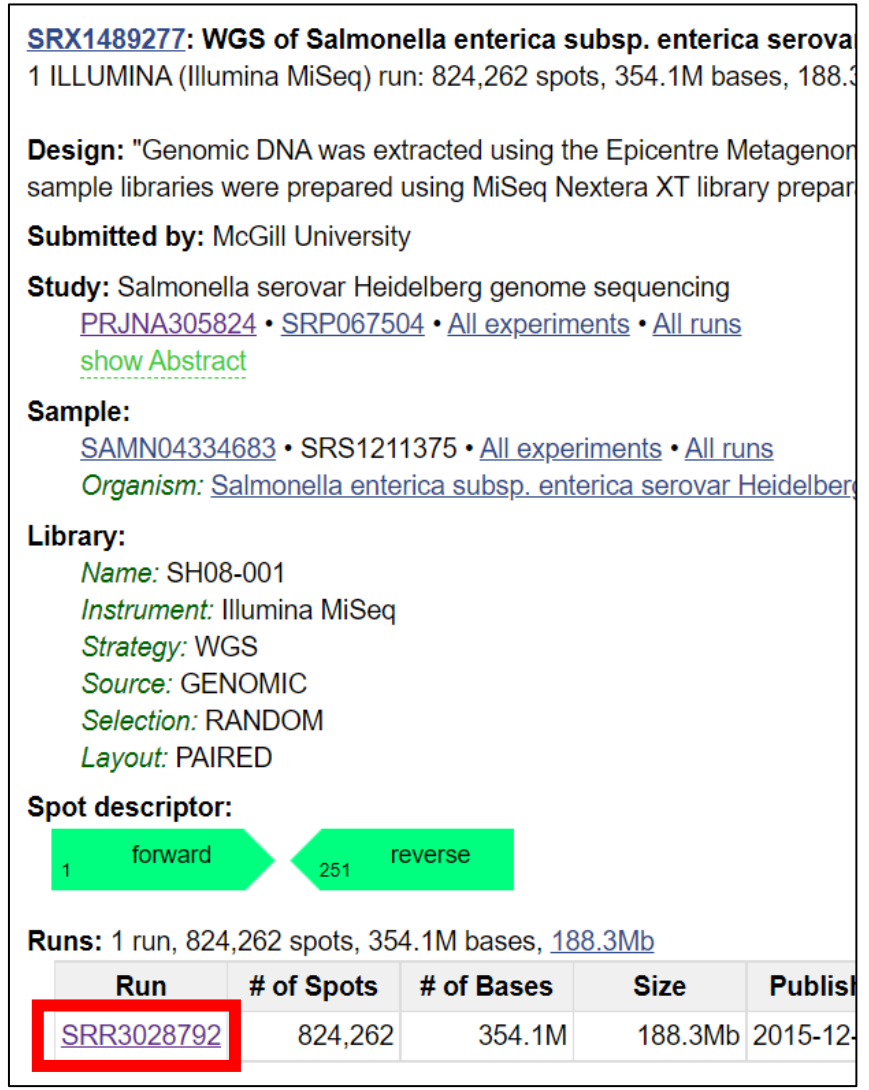
github.com/ncbi/sra-tools/wiki/08.-prefetch-and-fasterq-dump

- “prefetch” can be used to download SRA data

```
$ prefetch SRR3028792
```

- Once completed you can see where the file is stored with “srapath”

```
$ srapath SRR3028792
/path/to/sra/SRR3712208.sra
```



SRX1489277: WGS of Salmonella enterica subsp. enterica serova
1 ILLUMINA (Illumina MiSeq) run: 824,262 spots, 354.1M bases, 188.3

Design: "Genomic DNA was extracted using the Epicentre Metagenor sample libraries were prepared using MiSeq Nextera XT library prepar

Submitted by: McGill University

Study: Salmonella serovar Heidelberg genome sequencing
PRJNA305824 • SRP067504 • All experiments • All runs
show Abstract

Sample:
SAMN04334683 • SRS1211375 • All experiments • All runs
Organism: Salmonella enterica subsp. enterica serovar Heidelber

Library:
Name: SH08-001
Instrument: Illumina MiSeq
Strategy: WGS
Source: GENOMIC
Selection: RANDOM
Layout: PAIRED

Spot descriptor:
1 forward 251 reverse

Runs: 1 run, 824,262 spots, 354.1M bases, 188.3Mb

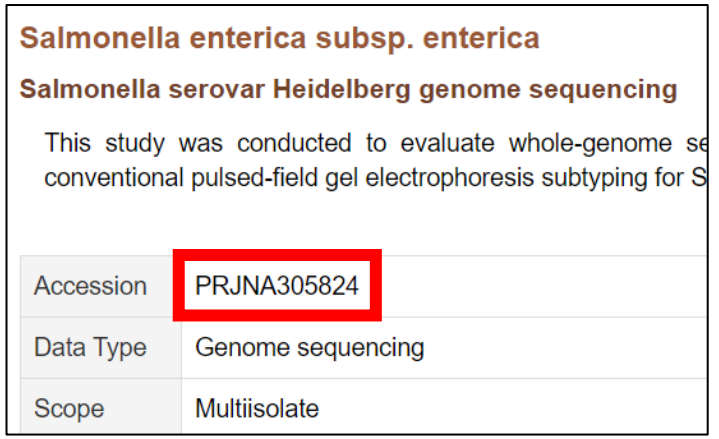
Run	# of Spots	# of Bases	Size	Publis
SRR3028792	824,262	354.1M	188.3Mb	2015-12-

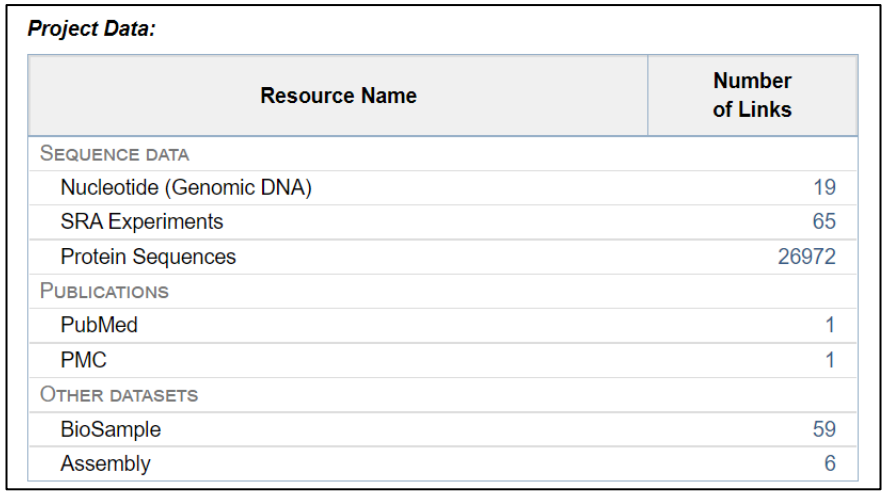
Prefetch

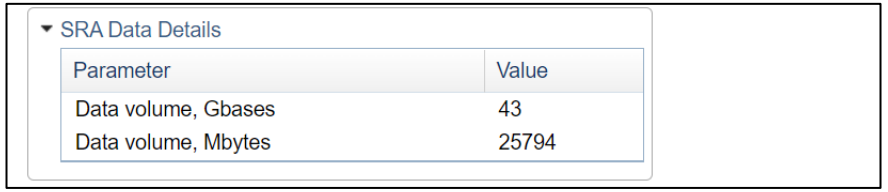
github.com/ncbi/sra-tools/wiki/08.-prefetch-and-fasterq-dump

- **You can use “prefetch” to download an entire BioProject**

```
$ prefetch PRJNA305824
```







www.ncbi.nlm.nih.gov/bioproject/PRJNA305824

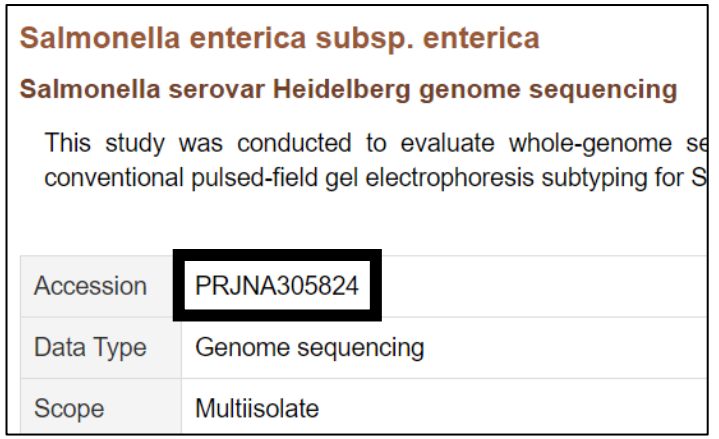
Prefetch

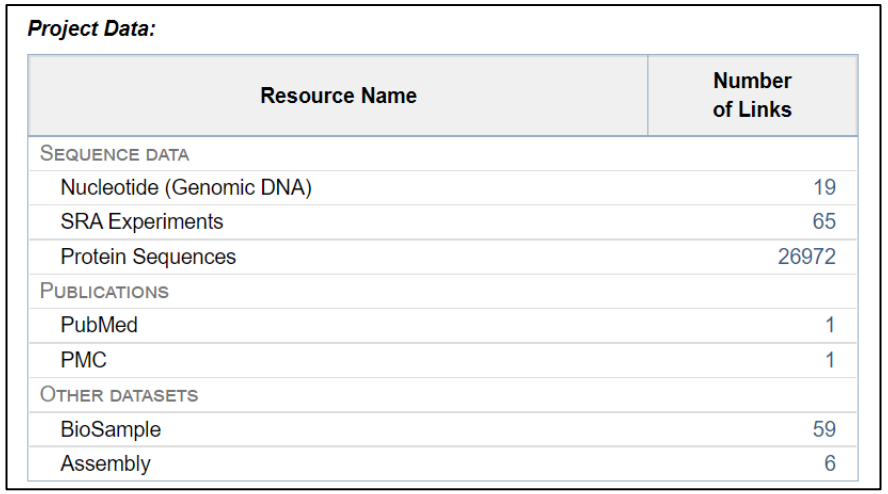
github.com/ncbi/sra-tools/wiki/08.-prefetch-and-fasterq-dump

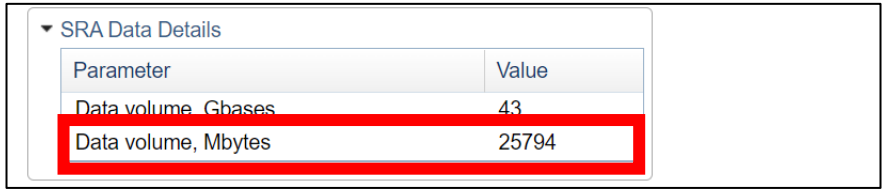
- **You can use "prefetch" to download an entire BioProject**

```
$ prefetch PRJNA305824
```

This might result in a large amount of data downloaded (e.g., ~25 GB here).







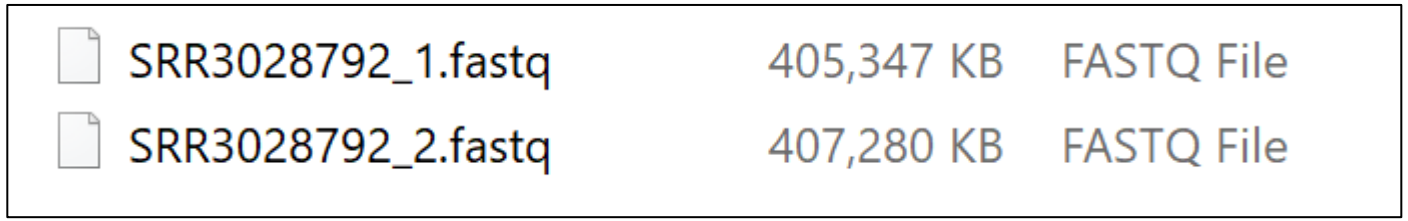
www.ncbi.nlm.nih.gov/bioproject/PRJNA305824

Fasterq-dump

github.com/ncbi/sra-tools/wiki/08.-prefetch-and-fasterq-dump

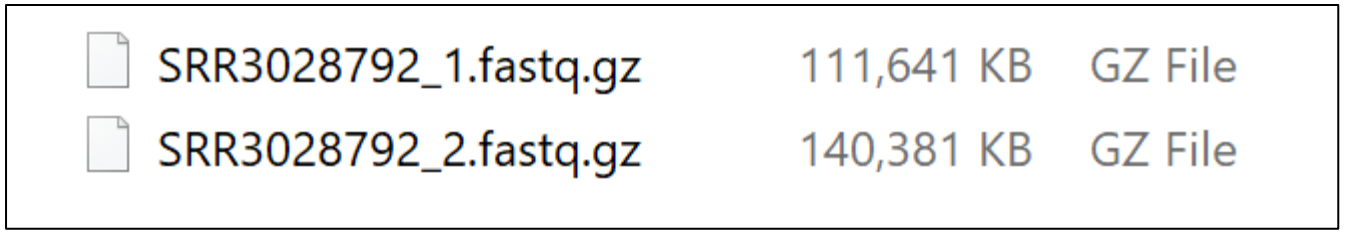
- **Convert to FASTQ with “fasterq-dump”**

```
$ fasterq-dump SRR3028792
```



- **You can compress the files with “gzip”**

```
$ gzip *.fastq
```

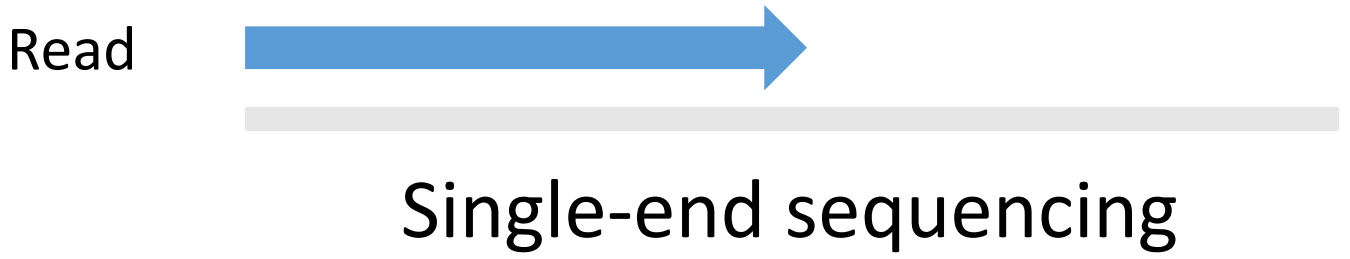


FASTQ

en.wikipedia.org/wiki/FASTQ_format

File: *reads.fastq*

```
@SRR3028792.1
CTTCATAATCAGGCGATAAATGCCCACCA
+SRR3028792.1
AABC-C--CE,-,C++@,,,<,CCC,BB,
```



Single-end sequencing

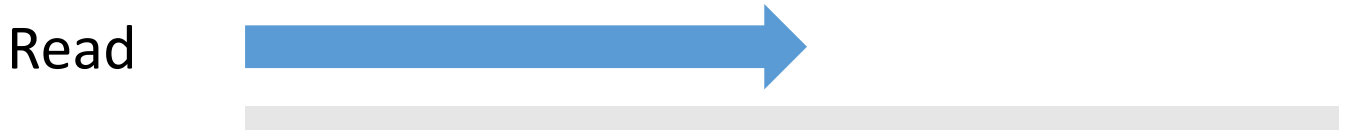
FASTQ

en.wikipedia.org/wiki/FASTQ_format

File: *reads.fastq*

Read ID

```
@SRR3028792.1
CTTCATAATCAGGCGATAAATGCCCACCA
+SRR3028792.1
AABC-C--CE,-,C++@,,,<,CCC,BB,
```



Single-end sequencing

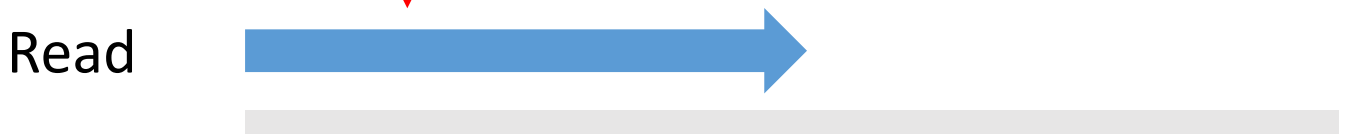
FASTQ

en.wikipedia.org/wiki/FASTQ_format

File: *reads.fastq*

```
@SRR3028792.1
CTTCATAATCAGGCGATAAATGCCCACCA
+SRR3028792.1
AABC-C--CE,-,C++@,,,<,CCC,BB,
```

Nucleotides



Single-end sequencing

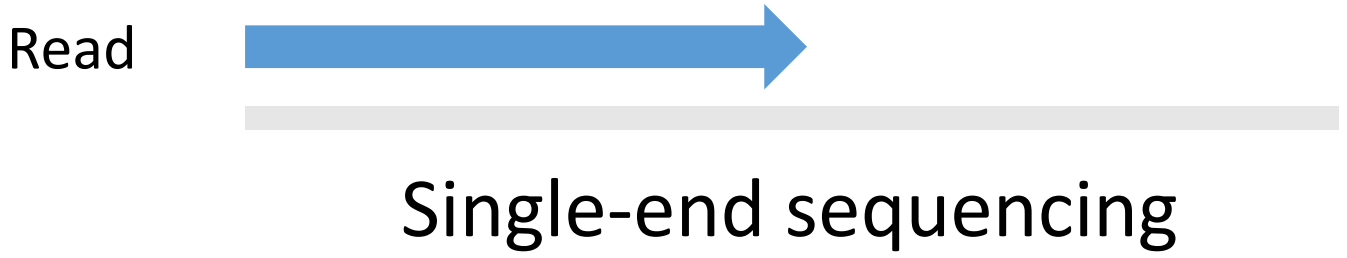
FASTQ

en.wikipedia.org/wiki/FASTQ_format

File: *reads.fastq*

```
@SRR3028792.1
CTTCATAATCAGGC
+SRR3028792.1
AABC-C--CE,-,C++@,,,<,CCC,BB,
```

Separator (read ID again)



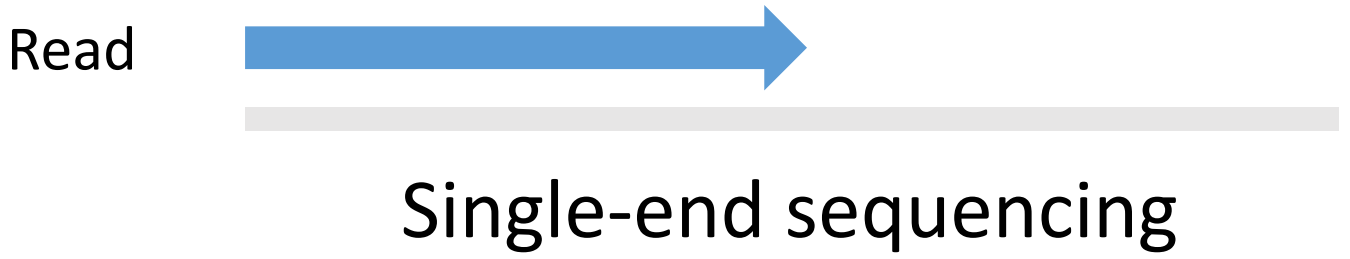
Single-end sequencing

FASTQ

Encoded quality of each nucleotide

File: *reads.fastq*

```
@SRR3028792.1
CTTCATAATCAGGCGATAAATGCCCACCA
+SRR3028792.1
AABC-C--CE,-,C++@,,,<,CCC,BB,
```



Single-end sequencing

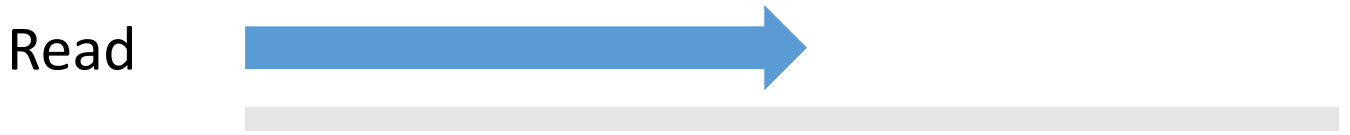
Quality score encoding

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
|                              |    |    |
0.2.....................26....32.....40
```

Character

Score

```
+SRR3028792.1
AABC-C--CE,-,C++@,,,<,CCC,BB,
```



Single-end sequencing

Quality score encoding

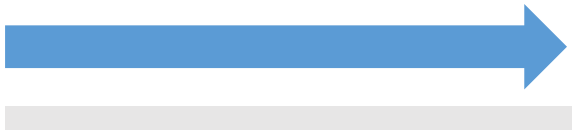
```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
|                              |    |    |
0.2.....................26....32.....40
```

Character

Score

```
+SRR3028792.1
AABC-C--CE,-,C++@,,,<,CCC,BB,
```

Score 10: P(error) = 10%
Score 20: P(error) = 1%
Score 30: P(error) = 0.1%



Single-end sequencing

Phred scores

en.wikipedia.org/wiki/Phred_quality_score

FASTQ

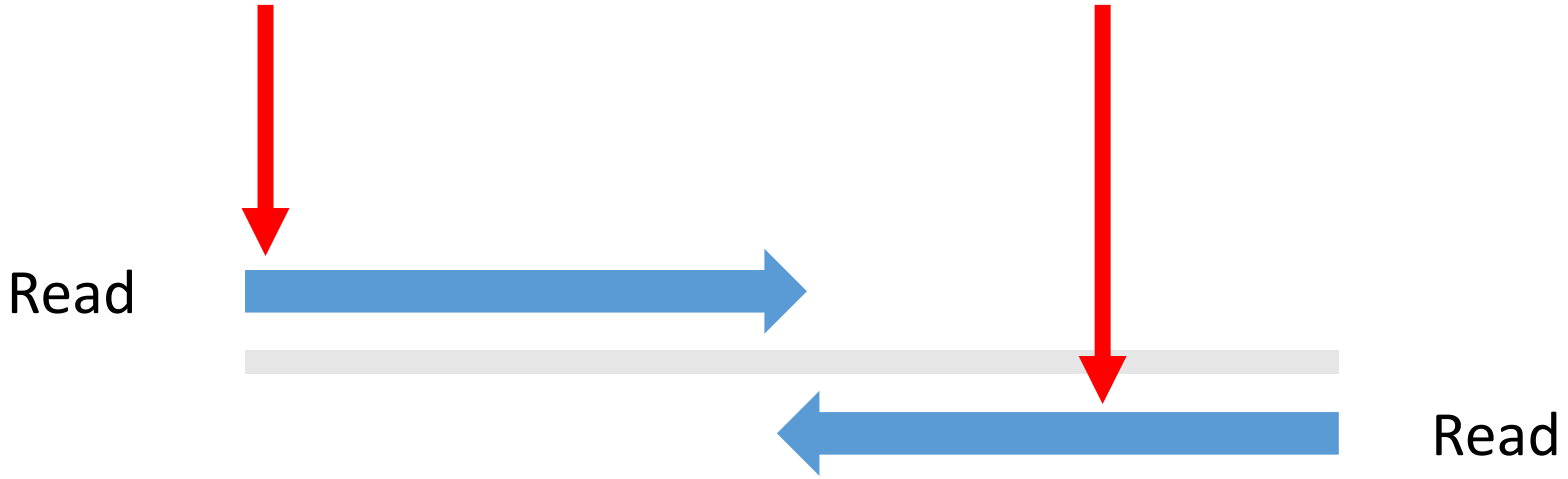
en.wikipedia.org/wiki/FASTQ_format

File: *reads_1.fastq*

```
@SRR3028792.1
CTTCATAATCAGGCGA
+SRR3028792.1
AABC-C--CE,-,C++
```

File: *reads_2.fastq*

```
@SRR3028792.1
CTCTCAAACCTTCCT
+SRR3028792.1
--,-8,-,;:,;,;,
```



Paired-end sequencing

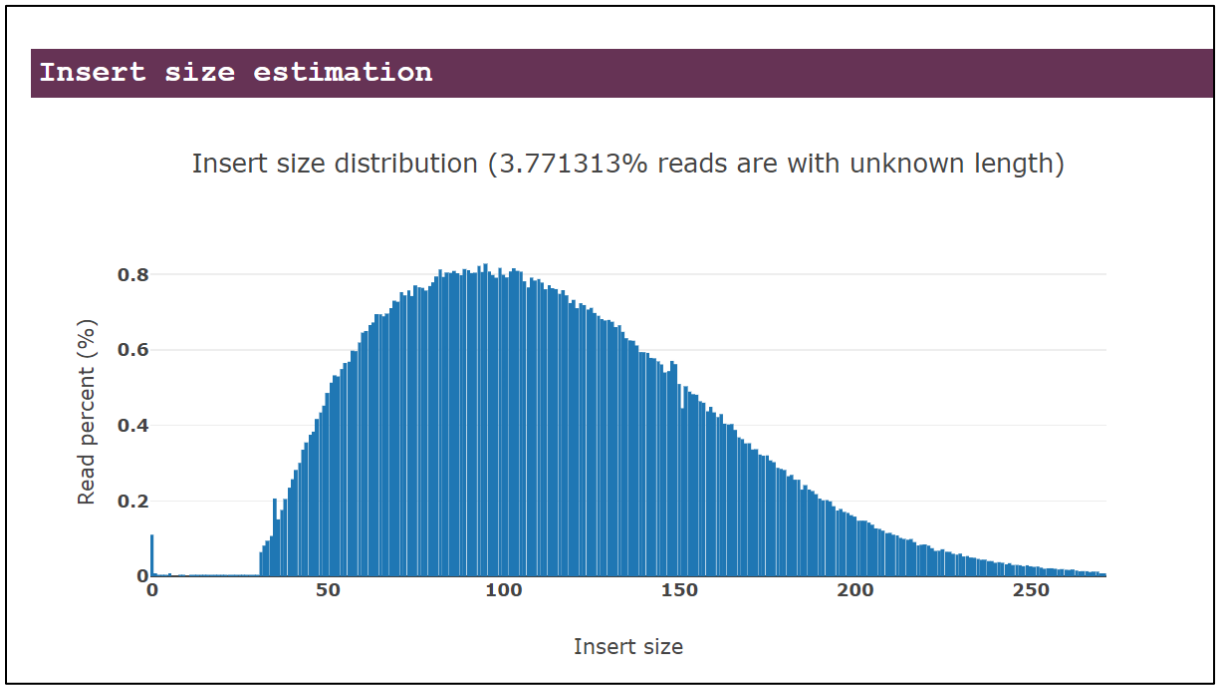
LEARNING OBJECTIVES

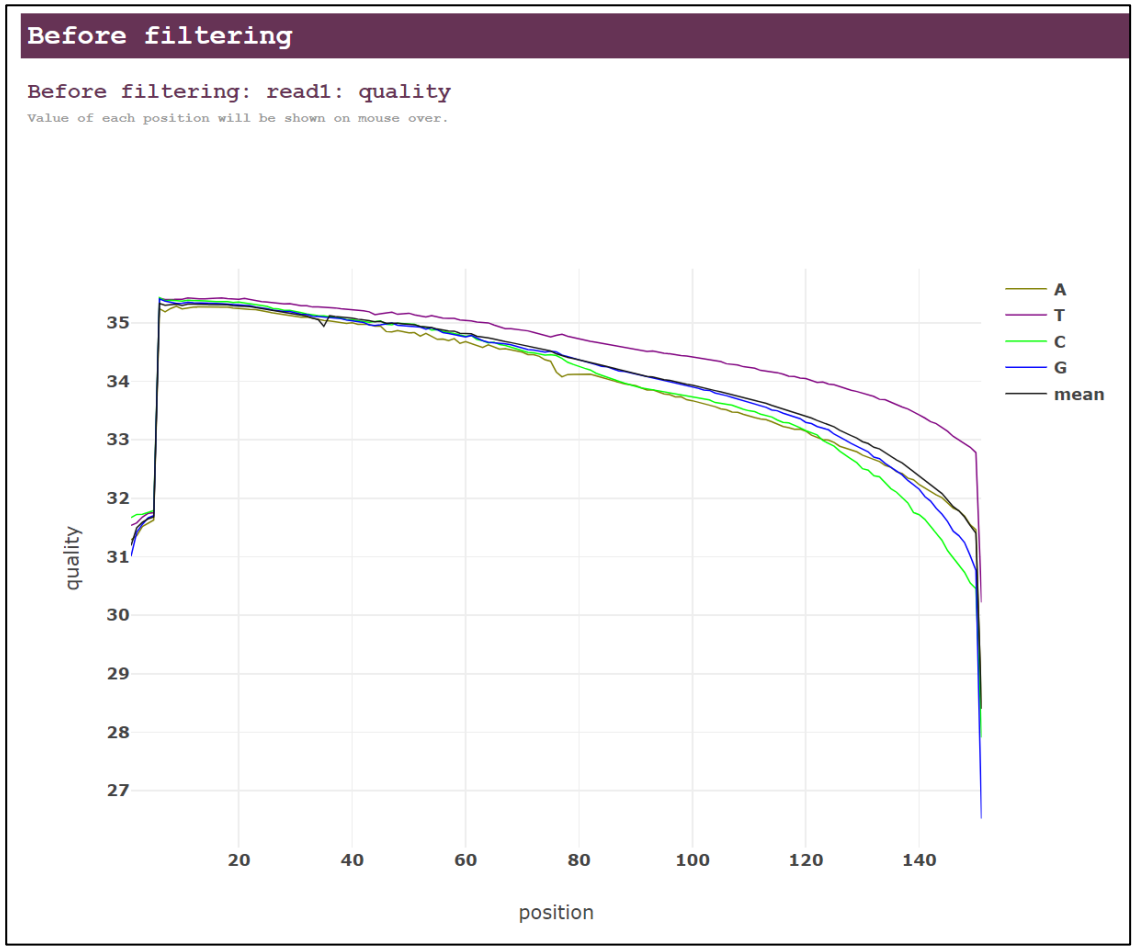
1. *What is microbial sequencing*
2. *What is genome assembly*
3. *Where to find sequence data*
4. *How to download*
5. ***Quality of reads***
6. *How to assemble data*
7. *How to evaluate quality of the assembly*
8. *Data analysis tutorial*

Evaluate quality of reads

github.com/OpenGene/fastp

- **Fastp** can be used to evaluate the quality of the reads and remove poor-quality data





Read quality with fastp

github.com/OpenGene/fastp

- **Install the tool with conda**

```
$ conda create -n fastp fastp
```

- **Process one pair of files**

```
$ fastp
    --in1 reads_1.fastq.gz --in2 reads_2.fastq.gz
    --out1 reads_1.fp.fastq.gz --out2 reads_2.fp.fastq.gz
    --html report.html
```

Read quality with fastp

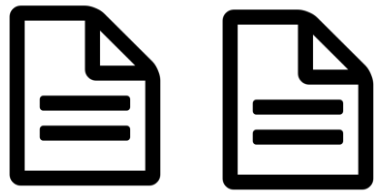
github.com/OpenGene/fastp

- **Install the tool with conda**

```
$ conda create -n fastp fastp
```

- **Process one pair of files**

```
$ fastp
    --in1 reads_1.fastq.gz --in2 reads_2.fastq.gz
    --out1 reads_1.fp.fastq.gz --out2 reads_2.fp.fastq.gz
    --html report.html
```



Filtered reads (fastq)

Downstream analysis

Read quality with fastp

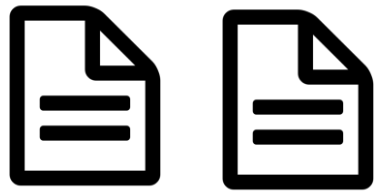
github.com/OpenGene/fastp

- **Install the tool with conda**

```
$ conda create -n fastp fastp
```

- **Process one pair of files**

```
$ fastp
    --in1 reads_1.fastq.gz --in2 reads_2.fastq.gz
    --out1 reads_1.fp.fastq.gz --out2 reads_2.fp.fastq.gz
    --html report.html
```



Filtered reads (fastq)

Downstream analysis

Quality report (html)

Fastp quality report

github.com/OpenGene/fastp

fastp report

Summary

General

fastp version:	0.23.2 (https://github.com/OpenGene/fastp)
sequencing:	paired end (301 cycles + 301 cycles)
mean length before filtering:	213bp, 215bp
mean length after filtering:	211bp, 211bp
duplication rate:	0.337593%
Insert size peak:	35

Filtering result

reads passed filters:	158.086000 K (95.642757%)
reads with low quality:	7.202000 K (4.357243%)
reads with too many N:	0 (0.000000%)
reads too short:	0 (0.000000%)

Fastp quality report

github.com/OpenGene/fastp

fastp report

Summary

General

fastp version:	0.23.2 (https://github.com/OpenGene/fastp)
sequencing:	paired end (301 cycles + 301 cycles)
mean length before filtering:	213bp, 215bp
mean length after filtering:	211bp, 211bp
duplication rate:	0.337593%
Insert size peak:	35

Filtering result

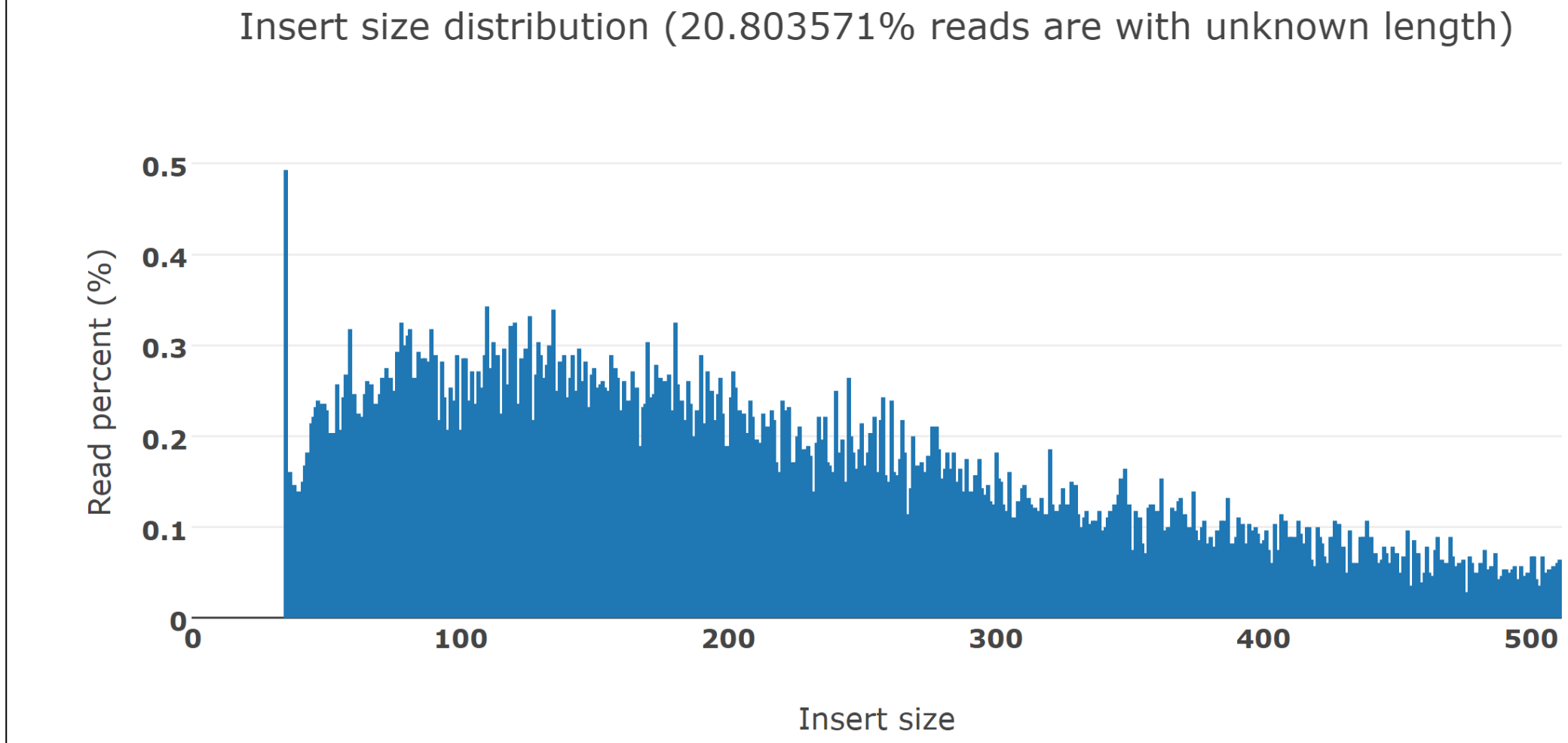
reads passed filters:	158.086000 K (95.6
reads with low quality:	7.202000 K (4.357
reads with too many N:	0 (0.000000%)
reads too short:	0 (0.000000%)

96% passed filters
4% removed

Fastp quality report

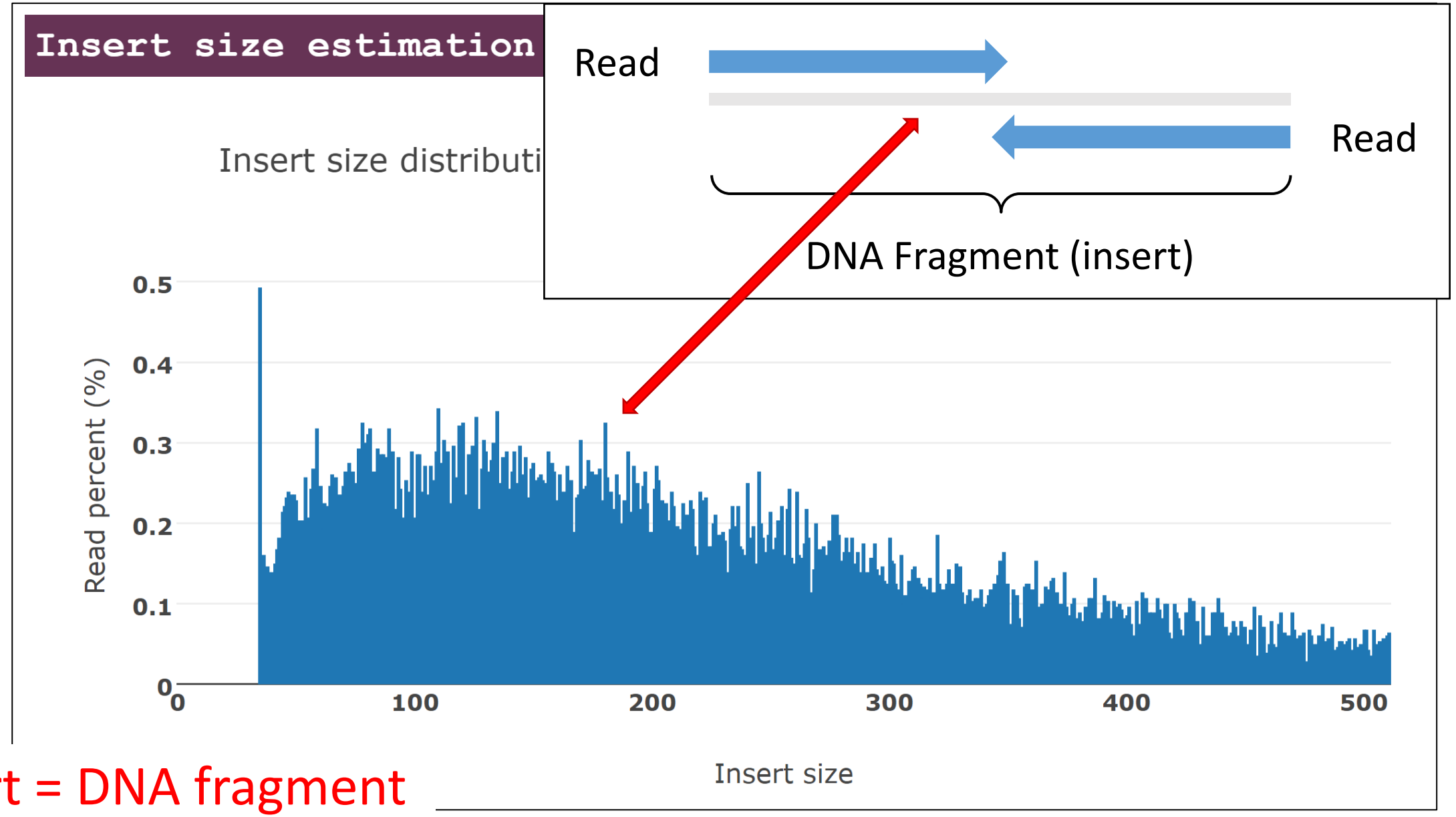
github.com/OpenGene/fastp

Insert size estimation



Fastp quality report

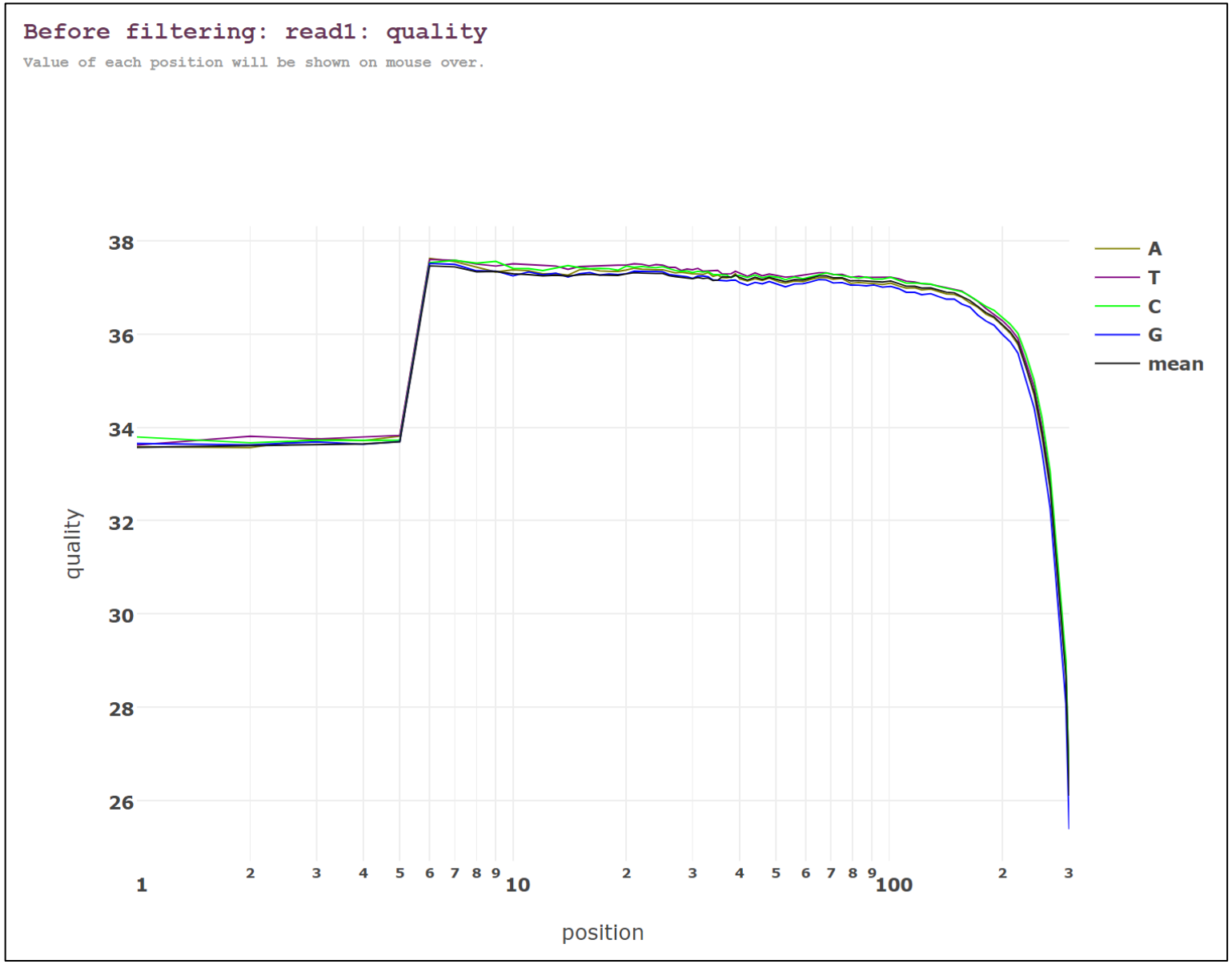
github.com/OpenGene/fastp



Insert = DNA fragment

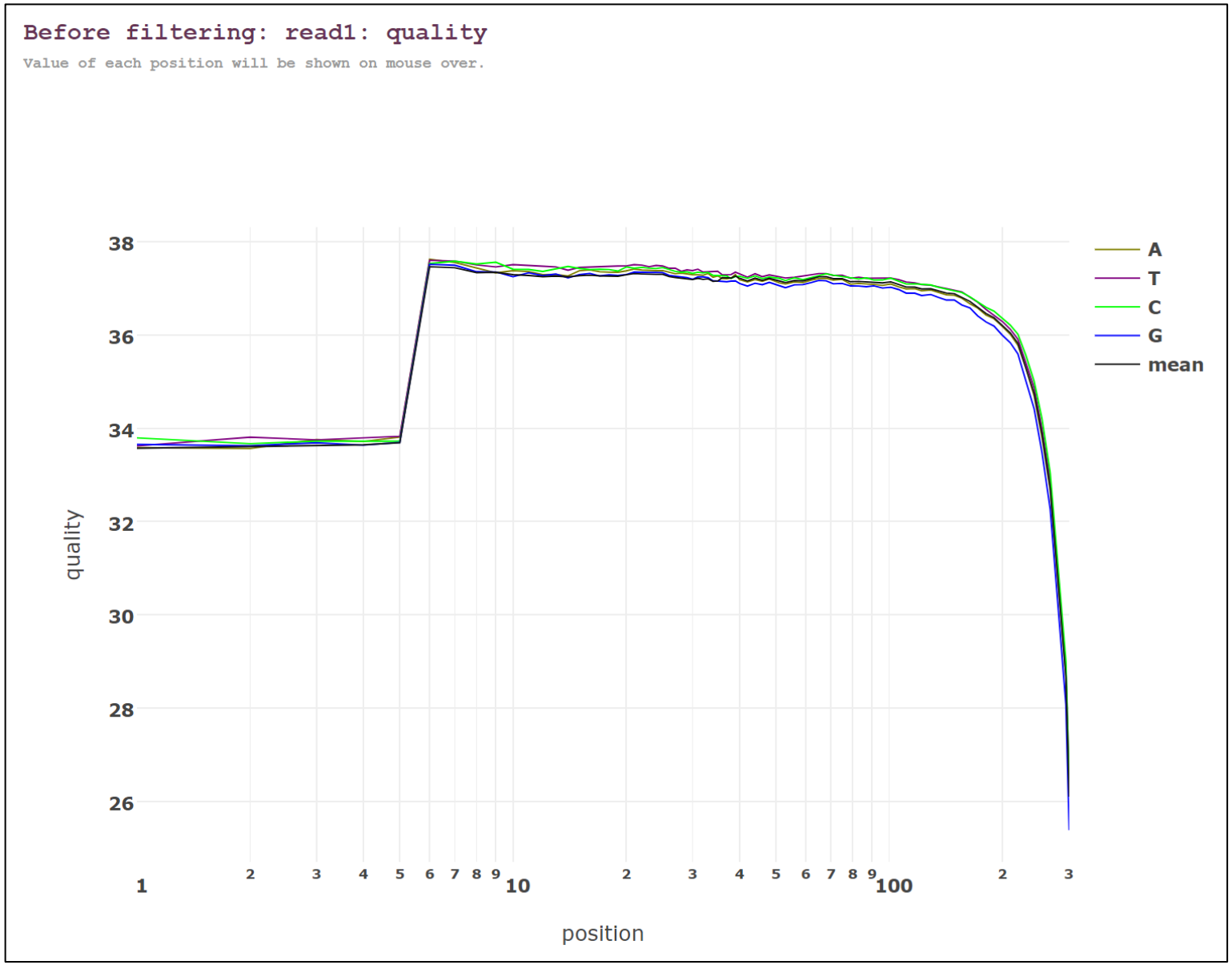
Fastp quality report

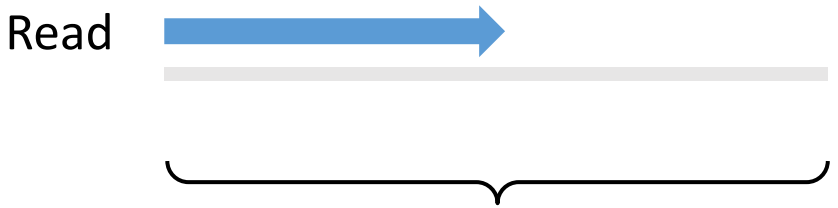
github.com/OpenGene/fastp



Fastp quality report

github.com/OpenGene/fastp





Mean score over all reads

LEARNING OBJECTIVES

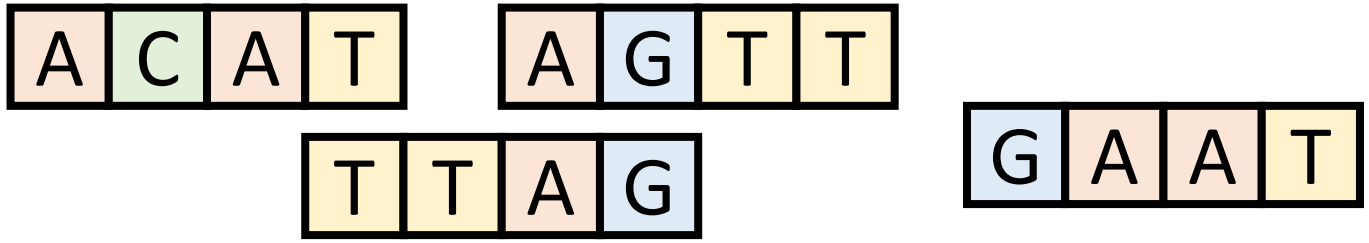
1. *What is microbial sequencing*
2. *What is genome assembly*
3. *Where to find sequence data*
4. *How to download*
5. *Quality of reads*
6. ***How to assemble data***
7. *How to evaluate quality of the assembly*
8. *Data analysis tutorial*

Genome assembly

https://en.wikipedia.org/wiki/Jigsaw_puzzle#/media/File:Sky_puzzle.jpg

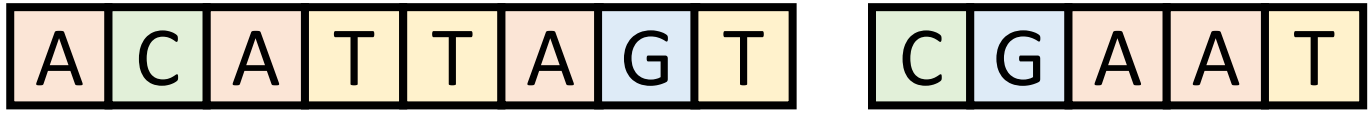
Genome assembly

Reads



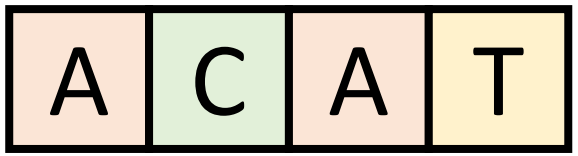
Transform reads to
larger contiguous fragments
(“contigs”)

Assembly



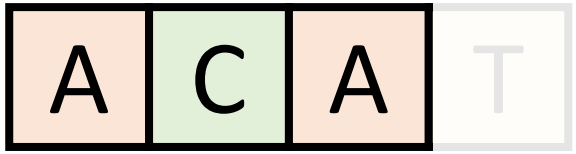
Genome assembly

Read

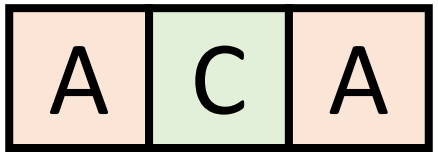


Genome assembly

Read

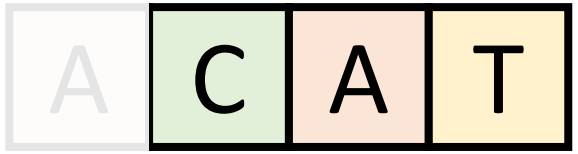


k-mer
(k=3)

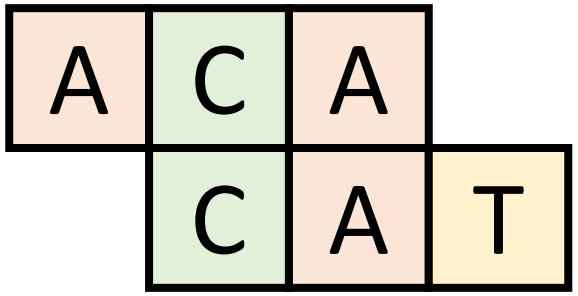


Genome assembly

Read

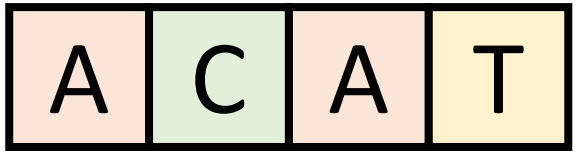


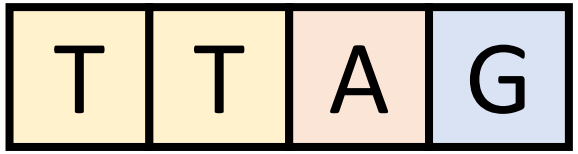
k-mer
(k=3)



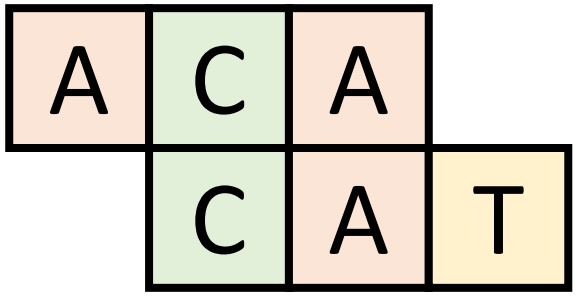
Genome assembly

Read





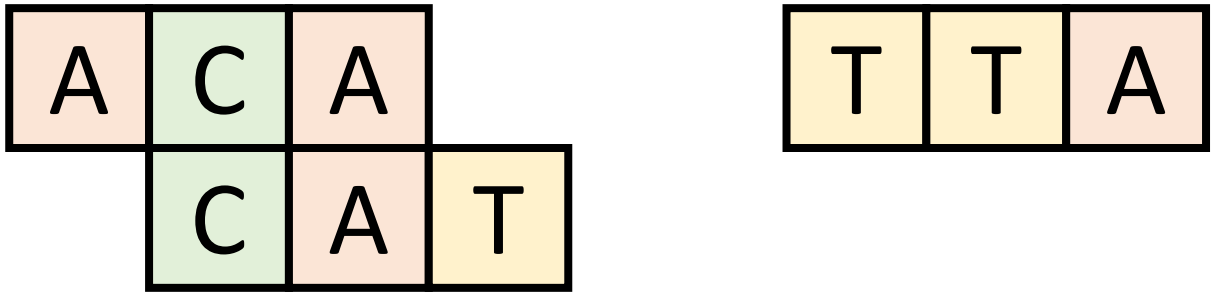
k-mer
(k=3)



Genome assembly

Read

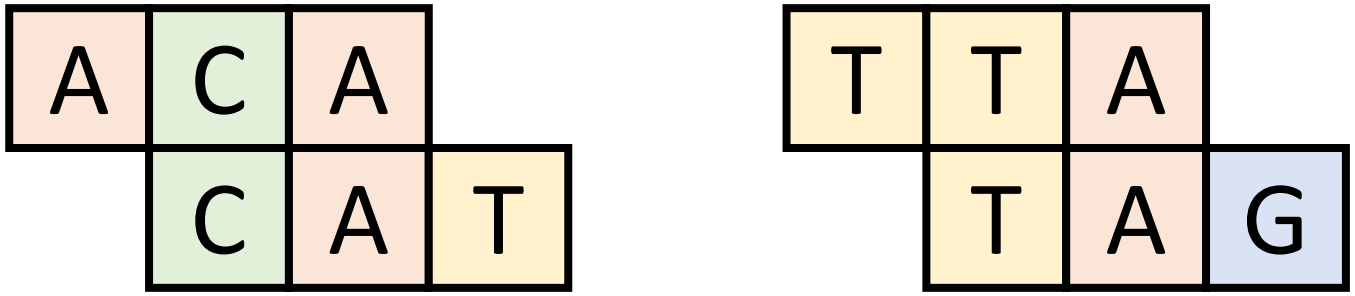
k-mer
(k=3)



Genome assembly

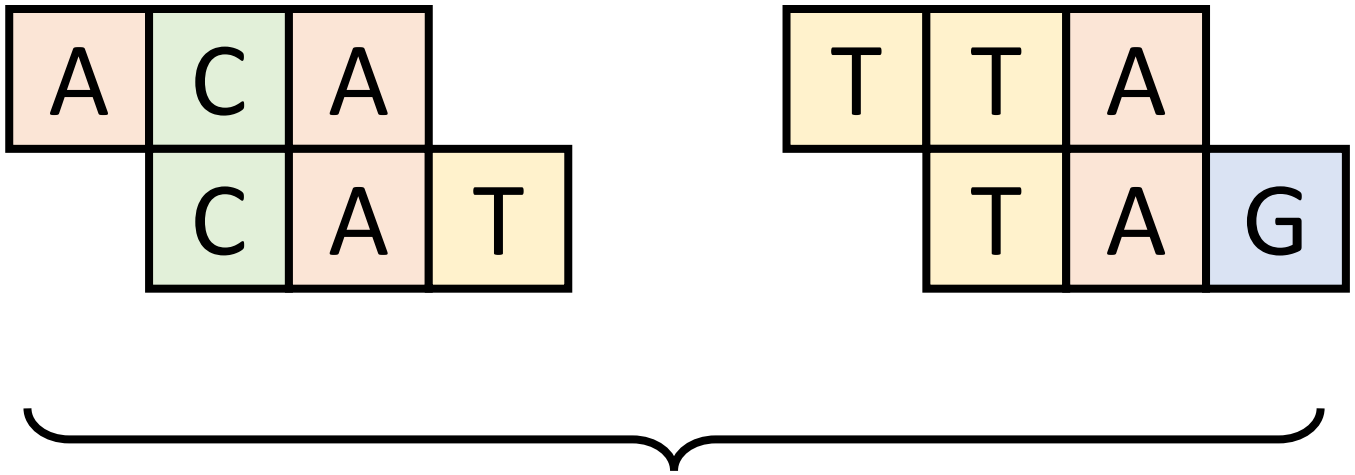
Read

k-mer
(k=3)



Genome assembly

k-mer
(k=3)

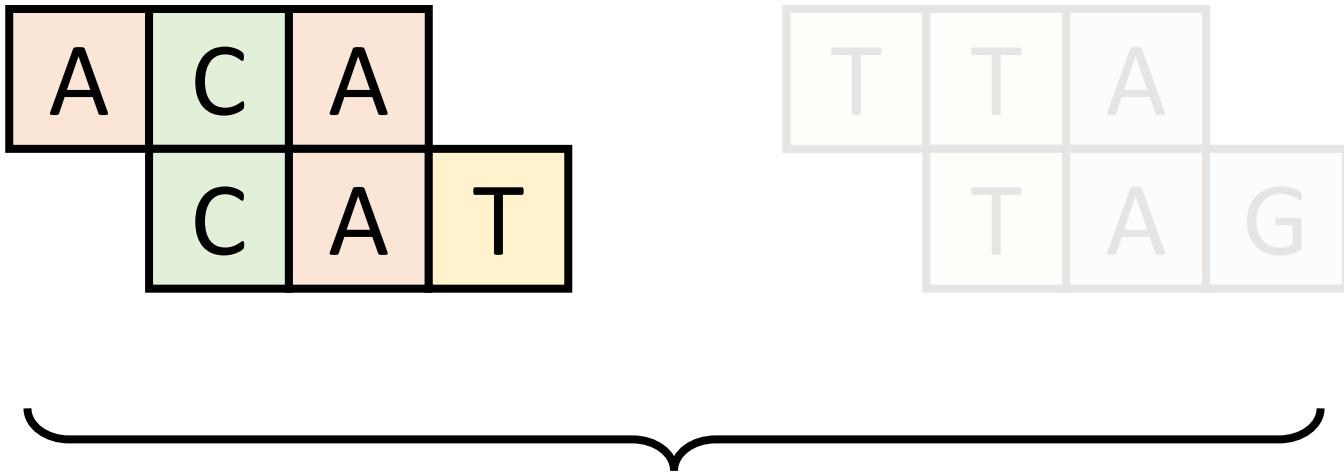


De bruijn
graph

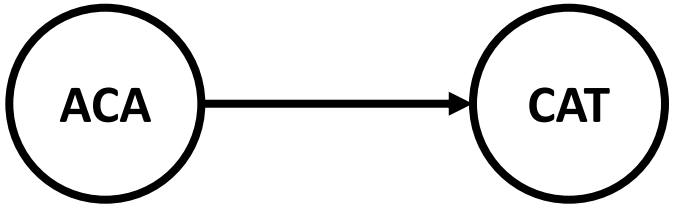
Construct a graph from k-mers

Genome assembly

k-mer
(k=3)



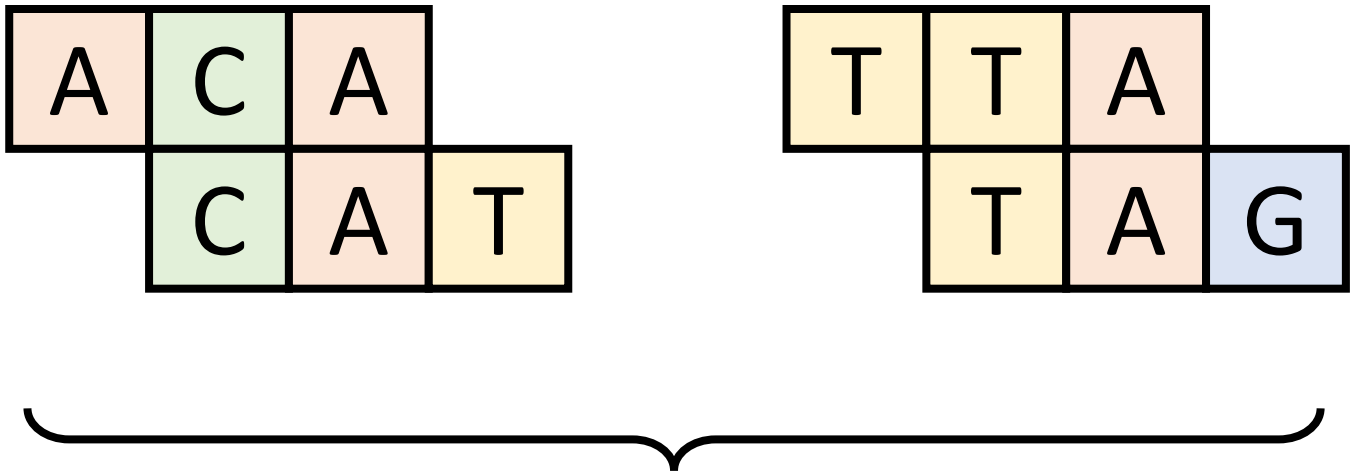
De bruijn
graph



Construct a graph from k-mers

Genome assembly

k-mer
(k=3)

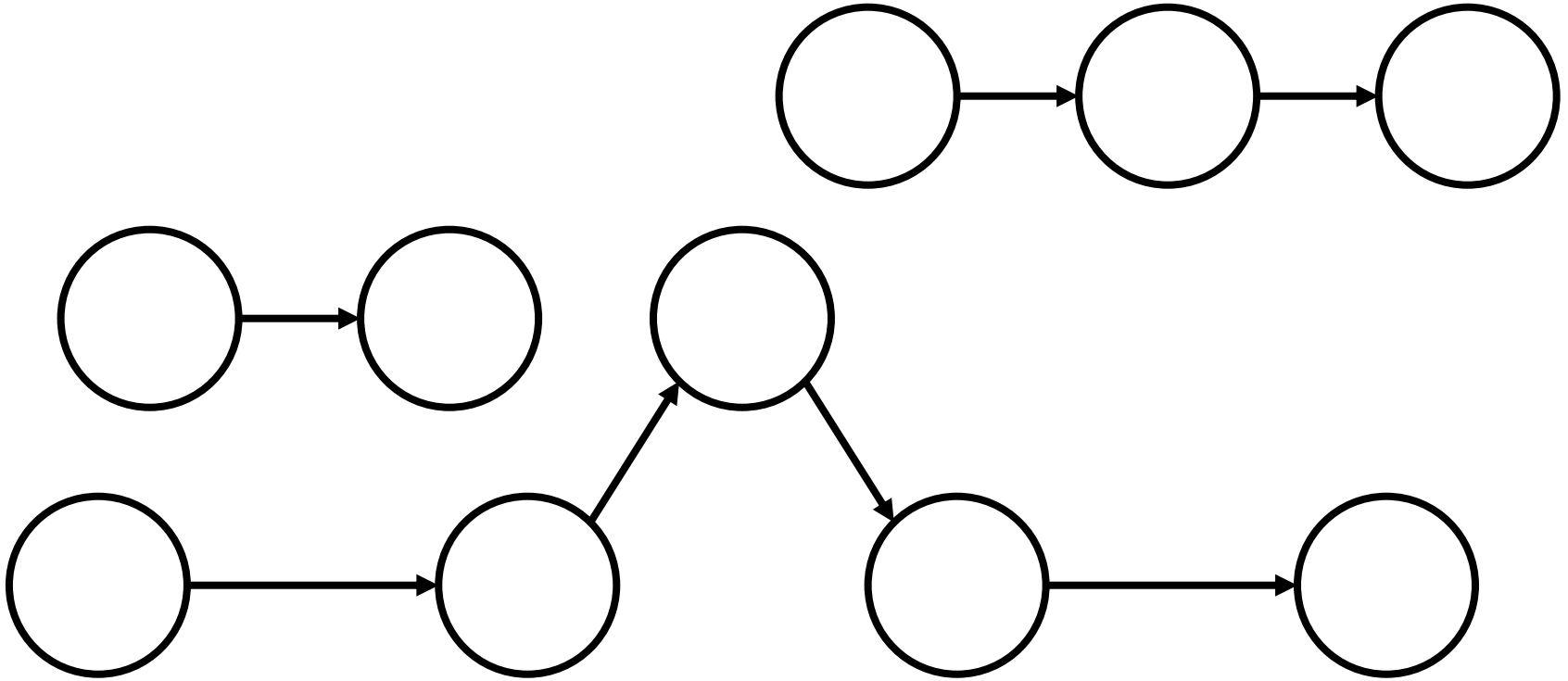


De bruijn
graph

Construct a graph from k-mers

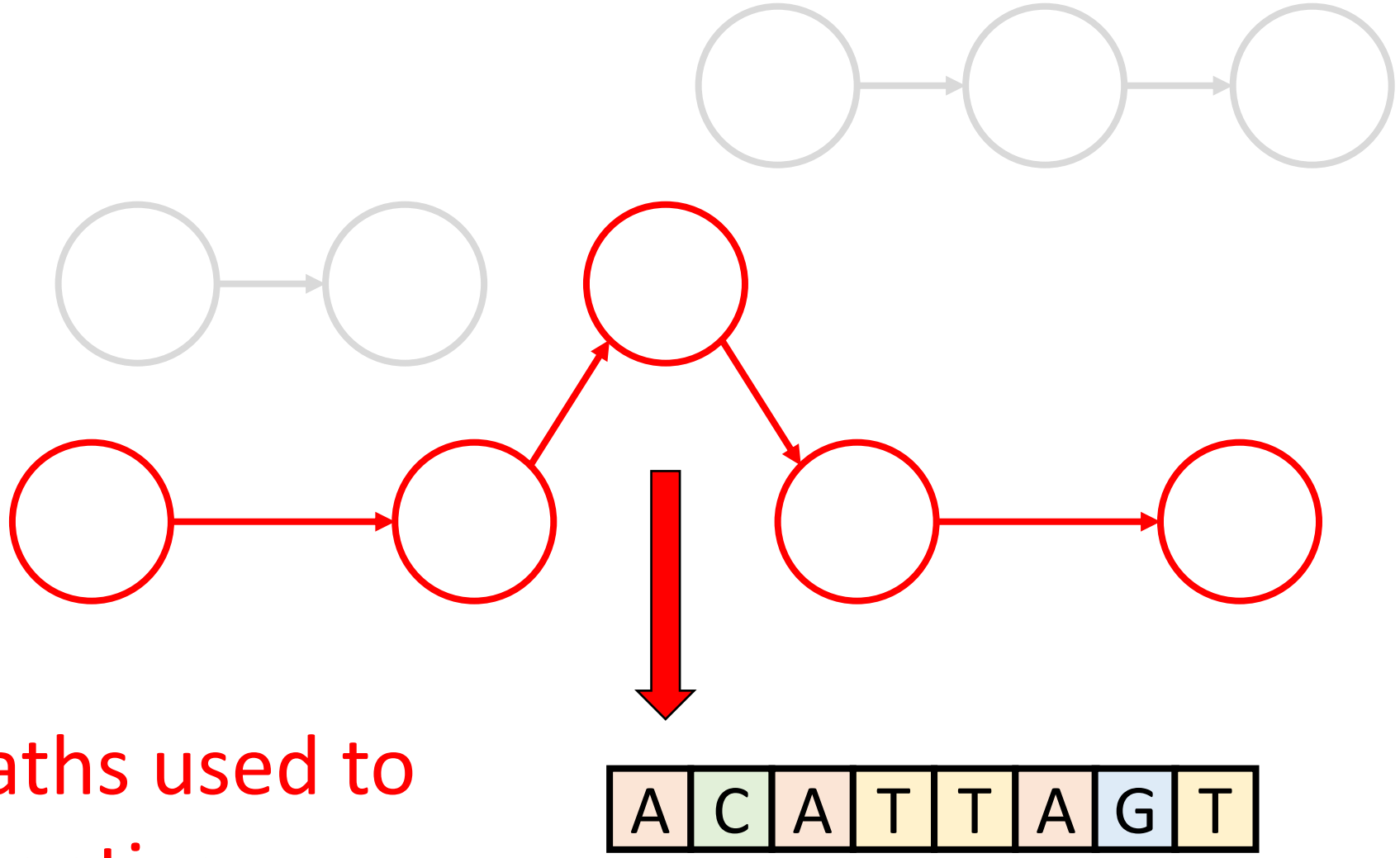
Genome assembly

Completed graph



Genome assembly

Completed graph



Unambiguous paths used to construct contigs

SKESA

github.com/ncbi/SKESA

Souvorov *et al. Genome Biology* (2018) 19:153
https://doi.org/10.1186/s13059-018-1540-z

Genome Biology

SOFTWARE Open Access

SKESA: strategic k-mer extension for scrupulous assemblies

Alexandre Souvorov[1], Richa Agarwala[1*] and David J. Lipman[1,2]

Abstract

SKESA is a DeBruijn graph-based de-novo assembler designed for assembling reads of microbial genomes sequenced using Illumina. Comparison with SPAdes and MegaHit shows that SKESA produces assemblies that have high sequence quality and contiguity, handles low-level contamination in reads, is fast, and produces an identical assembly for the same input when assembled multiple times with the same or different compute resources. SKESA has been used for assembling over 272,000 read sets in the Sequence Read Archive at NCBI and for real-time pathogen detection. Source code for SKESA is freely available at https://github.com/ncbi/SKESA/releases.

Keywords: Illumina reads, De-novo assembly, DeBruijn graphs, Sequence quality, Contamination

- Microbial genomes
- Illumina sequence data

SKESA

github.com/ncbi/SKESA

- **Install with conda**

```
$ conda create -n skesa skesa
```

- **Assemble genome**

```
$ skesa
   --reads reads_1.fp.fastq.gz,reads_2.fp.fastq.gz
   --contigs_out contigs.fasta
```

SKESA

github.com/ncbi/SKESA

- **Install with conda**

```
$ conda create -n skesa skesa
```

- **Assemble genome**

```
$ skesa
   --reads reads_1.fp.fastq.gz,reads_2.fp.fastq.gz
   --contigs_out contigs.fasta
```

- **Output**

Contigs
(fasta)

SKESA

github.com/ncbi/SKESA

- **Install with conda**

```
$ conda create -n skesa skesa
```

- **Assemble genome**

```
$ skesa
   --reads reads_1.fp.fastq.gz,reads_2.fp.fastq.gz
   --contigs_out contigs.fasta
```

- **Output**

Contigs
(fasta)

Information on
assembly

SKESA assembly information

github.com/ncbi/SKESA

```
Kmer: 21 Graph size: 4403385 Contigs in: 0
Valley: 0

Mark used kmers in  0.000003s wall, 0.000000s user + 0.000000s system = 0.000000s CPU (n/a%)
Kmers in multiple/single contigs: 0 0
Fragments before: 8949 4280997
Fragments after: 8786 4280997
New seeds: 7533
New seeds in  1.205928s wall, 51.980000s user + 0.010000s system = 51.990000s CPU (4311.2%)
Fragments before: 15066 316386
Fragments after: 15066 316386
Connectors: 0 Extenders: 15066
Connections and extensions in  0.205358s wall, 4.870000s user + 0.150000s system = 5.020000s CPU (2444.5%)
Contigs out: 7533 Genome: 3918836 N50: 858 L50: 1373
Assembled in  1.428887s wall, 56.870000s user + 0.160000s system = 57.030000s CPU (3991.2%)
```

SKESA assembly information

github.com/ncbi/SKESA

Kmer size (k=21)

```
Kmer: 21 Graph size: 4403385 Contigs in: 0
Valley: 0

Mark used kmers in  0.000003s wall, 0.000000s user + 0.000000s system = 0.000000s CPU (n/a%)
Kmers in multiple/single contigs: 0 0
Fragments before: 8949 4280997
Fragments after: 8786 4280997
New seeds: 7533
New seeds in  1.205928s wall, 51.980000s user + 0.010000s system = 51.990000s CPU (4311.2%)
Fragments before: 15066 316386
Fragments after: 15066 316386
Connectors: 0 Extenders: 15066
Connections and extensions in  0.205358s wall, 4.870000s user + 0.150000s system = 5.020000s CPU (2444.5%)
Contigs out: 7533 Genome: 3918836 N50: 858 L50: 1373
Assembled in  1.428887s wall, 56.870000s user + 0.160000s system = 57.030000s CPU (3991.2%)
```

SKESA assembly information

github.com/ncbi/SKESA

Kmer size (k=21)

Graph size (4.4 million kmers)

```
Kmer: 21 Graph size: 4403385
Valley: 0

Mark used kmers in  0.000003s wall, 0.000000s user + 0.000000s system = 0.000000s CPU (n/a%)
Kmers in multiple/single contigs: 0 0
Fragments before: 8949 4280997
Fragments after: 8786 4280997
New seeds: 7533
New seeds in  1.205928s wall, 51.980000s user + 0.010000s system = 51.990000s CPU (4311.2%)
Fragments before: 15066 316386
Fragments after: 15066 316386
Connectors: 0 Extenders: 15066
Connections and extensions in  0.205358s wall, 4.870000s user + 0.150000s system = 5.020000s CPU (2444.5%)
Contigs out: 7533 Genome: 3918836 N50: 858 L50: 1373
Assembled in  1.428887s wall, 56.870000s user + 0.160000s system = 57.030000s CPU (3991.2%)
```

SKESA assembly information

github.com/ncbi/SKESA

Kmer size (k=21)

```
Kmer: 21 Graph size: 4403385
Valley: 0

Mark used kmers in  0.000003s wall, 0.000000s user + 0.000000s system = 0.000000s CPU (n/a%)
Kmers in multiple/single contigs: 0 0
Fragments before: 8949 4280997
Fragments after: 8786 4280997
New seeds: 7533
New seeds in  1.205928s wall, 51.980000s user + 0.010000s system = 51.990000s CPU (4311.2%)
Fragments before: 15066 316386
```

Graph size (4.4 million kmers)

Contigs produced (~7,500)

```
870000s user + 0.150000s system = 5.020000s CPU (2444.5%)
Contigs out: 7533 Genome: 3918836 N50: 858 L50: 1373
Assembled in  1.428887s wall, 56.870000s user + 0.160000s system = 57.030000s CPU (3991.2%)
```

SKESA assembly information

github.com/ncbi/SKESA

Kmer size (k=21)

```
Kmer: 21 Graph size: 4403385
Valley: 0

Mark used kmers in  0.000003s wall, 0.000000s user + 0.000000s system = 0.000000s CPU (n/a%)
Kmers in multiple/single contigs: 0 0
Fragments before: 8949 4280997
Fragments after: 8786 4280997
New seeds: 7533
New seeds in  1.205928s wall, 51.980000s user + 0.010000s system = 51.990000s CPU (4311.2%)
Fragments before: 15066 316386
```

Graph size (4.4 million kmers)

Contigs produced (~7,500)

```
870000s user + 0.150000s system = 5.020000s CPU (2444.5%)
Contigs out: 7533 Genome: 3918836 N50: 858 L50: 1373
Assembled in  1.428887s wall, 56.870000s user + 0.160000s system = 57.030000s CPU (3991.2%)
```

SKESA builds graphs using multiple kmer sizes

Output contigs

github.com/ncbi/SKESA

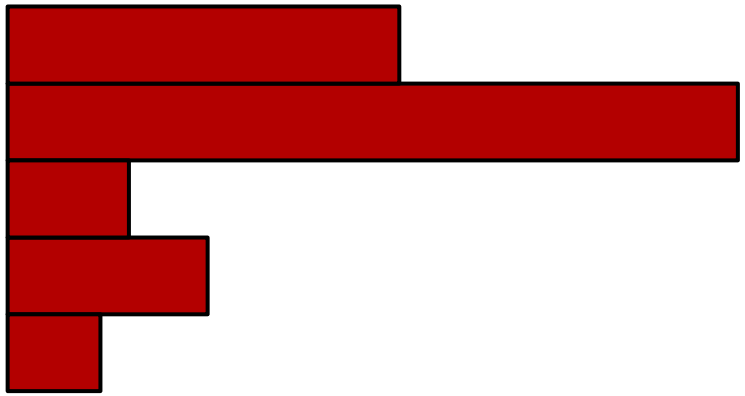
```
>Contig_3_5.5619
AAAAAACCGCCCATCGCCAGATTCTCTTCTACCGTCATCCGCGAGAAGACGCGCCGCCCTTCCGGCACTATCGCC
CTTTACCGTCAAACACCACCCGTCCGCTGGAGGCGCGCGGATCGCCGCACAAGGTGCCAAGCAGCGTCGTTTTCC
GATGTGCAGACTGACGTCGTGCAGCGCCTGAATCTTGCCGTAGTGGGCGCTGACTTTTTCAAACGTTAACATCGT
CGGGTTGTTGCGGATCTCTTCCGGCGTGCCGTTCGCCAGCGGCGTGCCCTGGTTCACCACGTAAATACGGTCGGA
>Contig_4_8.89207
AAAAAACTGGTACGTCGGCATTGGGCGCTTTTCCGCCGACGAGCACATACAGTCTGATGAGATTCGTTTCGAGAC
GACTCGGCAGTGCGGGCCACGCACCTGGTGAGCGGGATTAGCGTGAAGGTTCAGTCGGAGCGCAGTCAACATGCG
AGCAACGGCAAAATGAGTGTGCGGCGTTGAAGTCTGAAAGGCGTCTGTTCCATCATCAGATTGAGCGCGGCAATC
CGGTGCTACCAGCGCTGTTATAGGGCTGTCTGGCTATTCGCGTGTAGCCGCAGCGCTATTCAACGCATCCAGGAG
CGTCCAGAAATGGGACGGGCATTTTAGCGCTTTCTTACTTCGCAGGAATCGCTTTCAGCAATTCGGTCAGCAGAG
AGAGTGAGGTCCGGTAATGGTAGGCCCAATAGAAACCATGTCCATATCCGGATAGGGTTTCTTAAACAGACCGCA
AACAGACGCTGATAGGTTTCCCGCACCAGGTGCATGACCGGCGAGTTCGCATCGGGCTGCCAGCCAGGATAGCTG
CCAGCATACTCACCACATAATCTTTACCGCTGTCGATAAGAGAGCGAATCAGGCAGTGAATTTCGACATTCGCAT
GCCTTTCGCCACGTCTGAATTGCGGATCACGCCGTTCGGCGTTGCGTTCAGCAGGCGCACAAAGGTATCGCGTGA
TGCAGCGTCAGGTTTTTCTCTTTTTCCGCCAGTTCGTTTTTCAGAATATCCTGGTAAGCGTTCACTAATGTTTTC
CTTCGCGCGGAATCGCGTTACGCAGC
>Contig_5_4.58495
AAAAAAGCGAACATTTTTACGTAACGCCCGCTTCGAGCGACAAGAAGGGCGGCAAAAAACGCCGGTGGTTCGCAC
ACCCGCAGCCAGCATTCATCATCGGTAAGTAACGCTTTTTTCATCATAGACTCCTTCATTAAGCATAGTGCTTGC
TCTTGCGTTTTAATTCTTATCGCTCACCAGGAACGTTTTGAACTGCGGCGACATCCAGGTTTTAAAGCCTTCACC
ACCTCTTGCGCCTTTTGCGTACCAAGCACCATCAGGCCGGTGTCCCAGCCGTCCGCTTCCAACGCCGTTGGCGCG
```

LEARNING OBJECTIVES

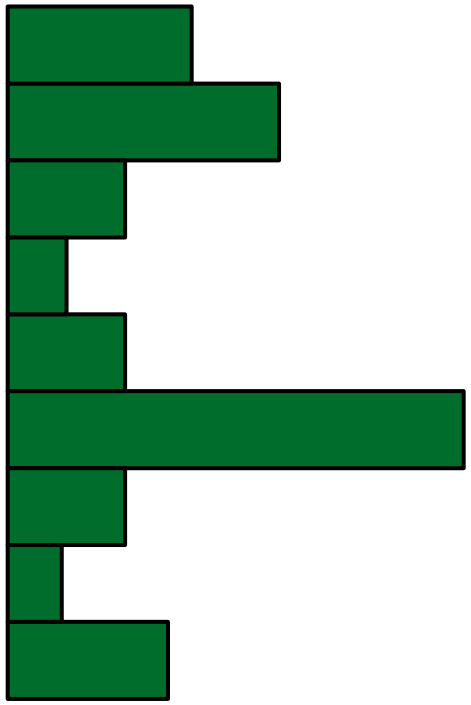
1. *What is microbial sequencing*
2. *What is genome assembly*
3. *Where to find sequence data*
4. *How to download*
5. *Quality of reads*
6. *How to assemble data*
7. ***How to evaluate quality of the assembly***
8. *Data analysis tutorial*

Scenario:
Two assembled genomes (contigs)

A

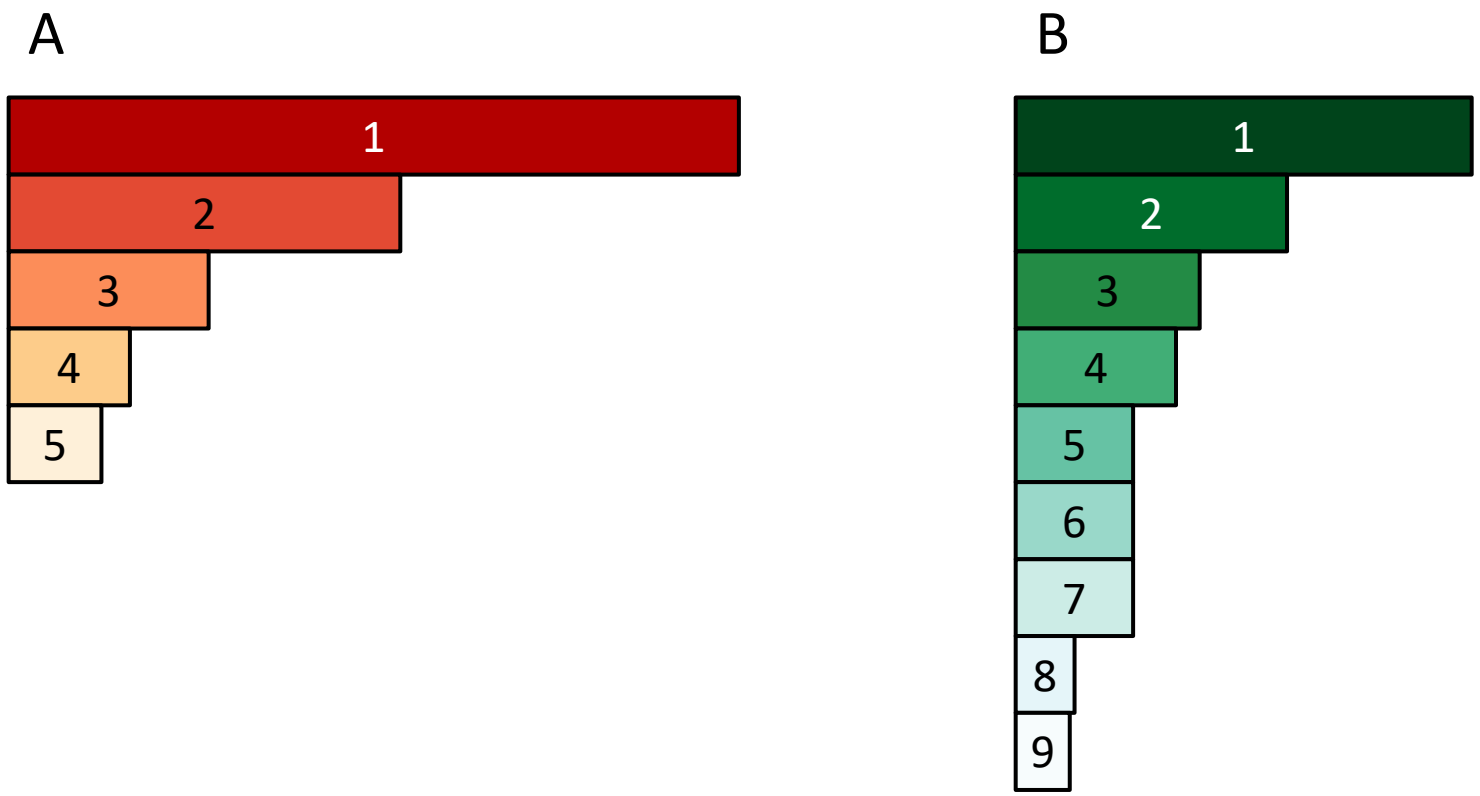


B



Which is more “complete”
or “contiguous”?

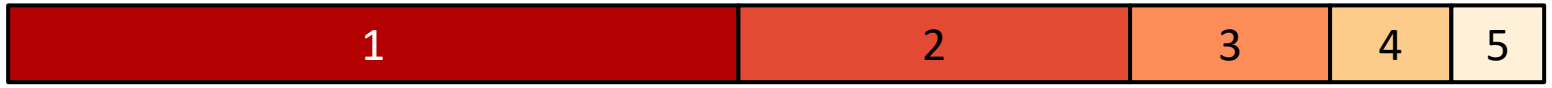
Order and count contigs



Assembly **A** has less contigs. Is this better?

Find cumulative length

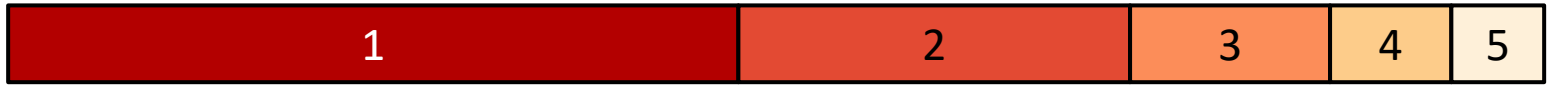
A



B

Find cumulative length

A



1 Mbp

B

1 Mbp

Both same length. Assembly **A** is looking good (same length, less contigs).

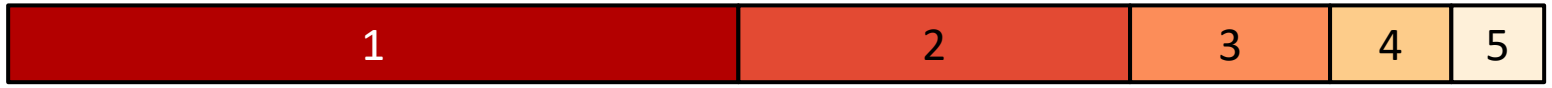
N50

The contig length such that at least half of the nucleotides in the assembly belong to contigs of this length or greater.

L50

The number of contigs with N50 length or greater.

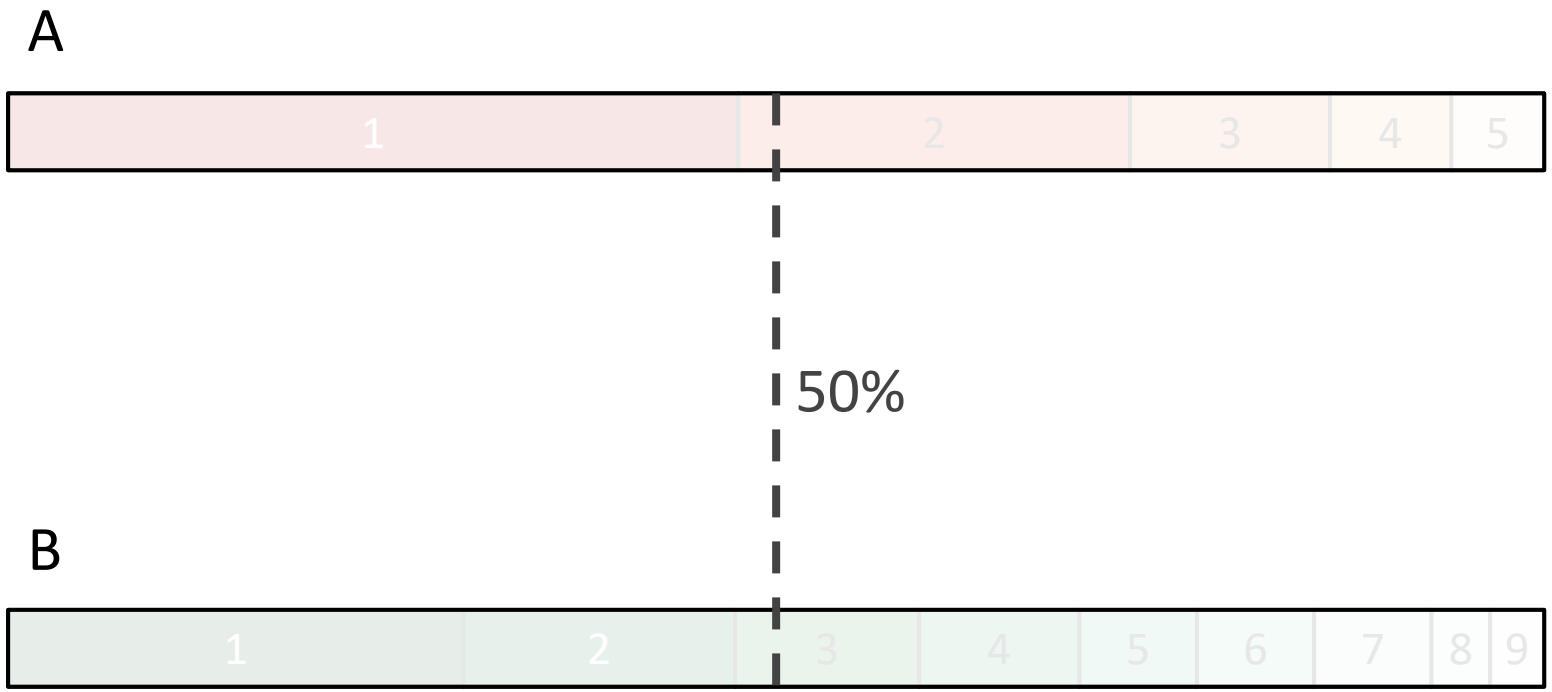
A

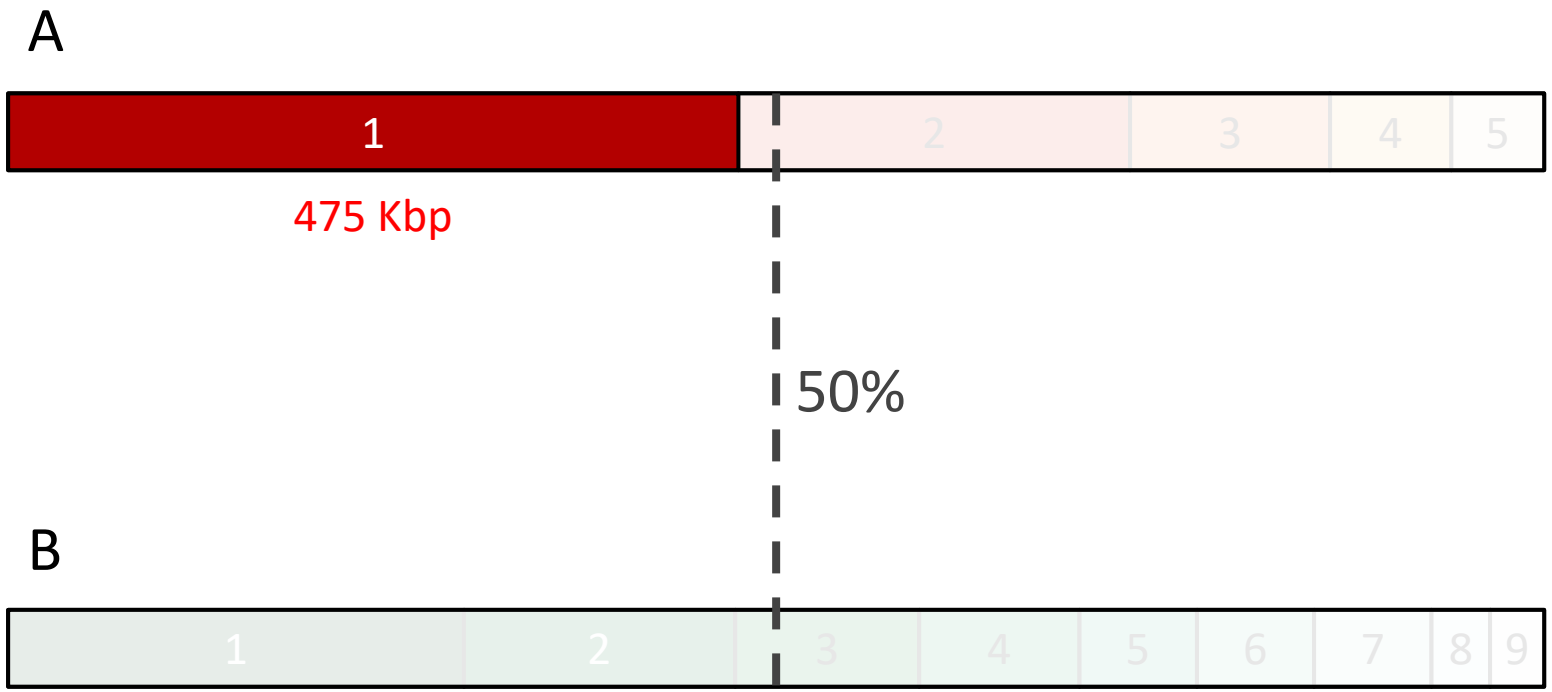


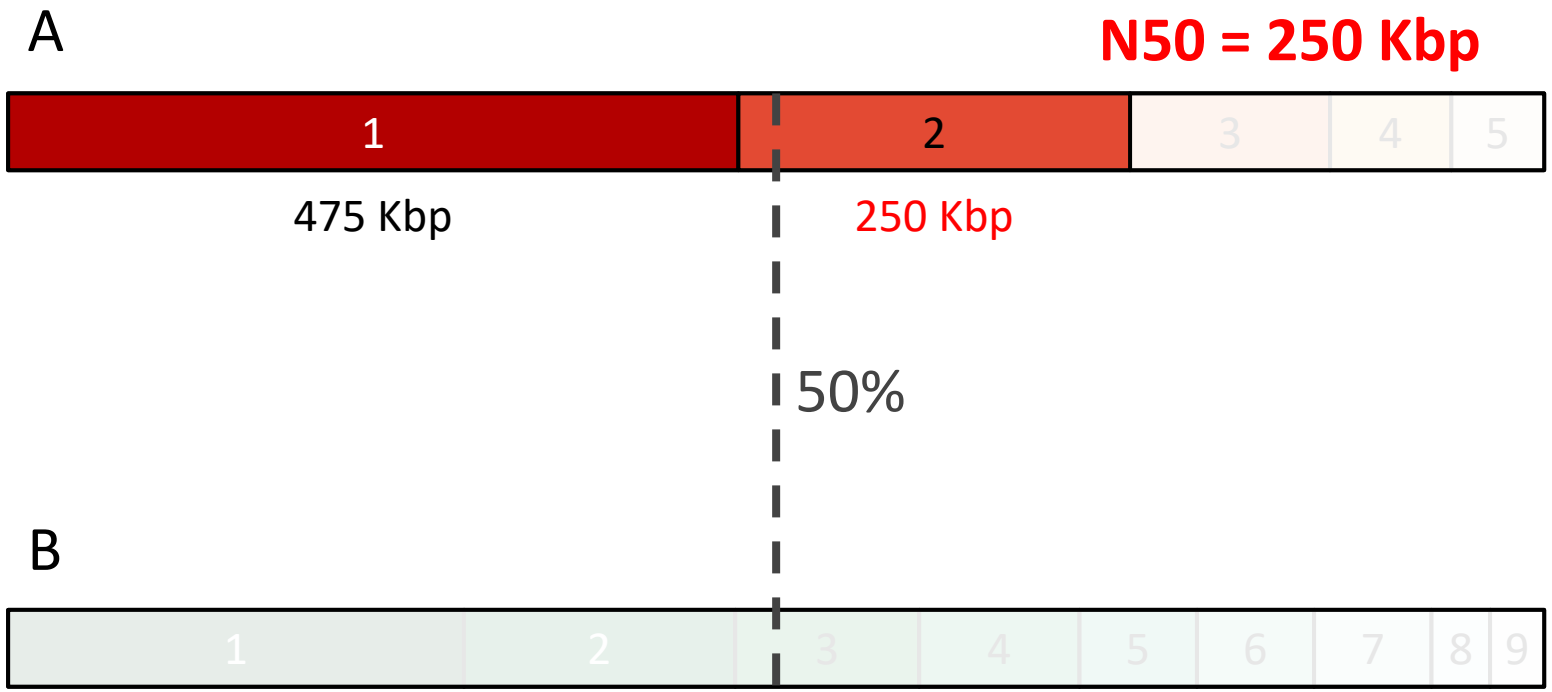
B

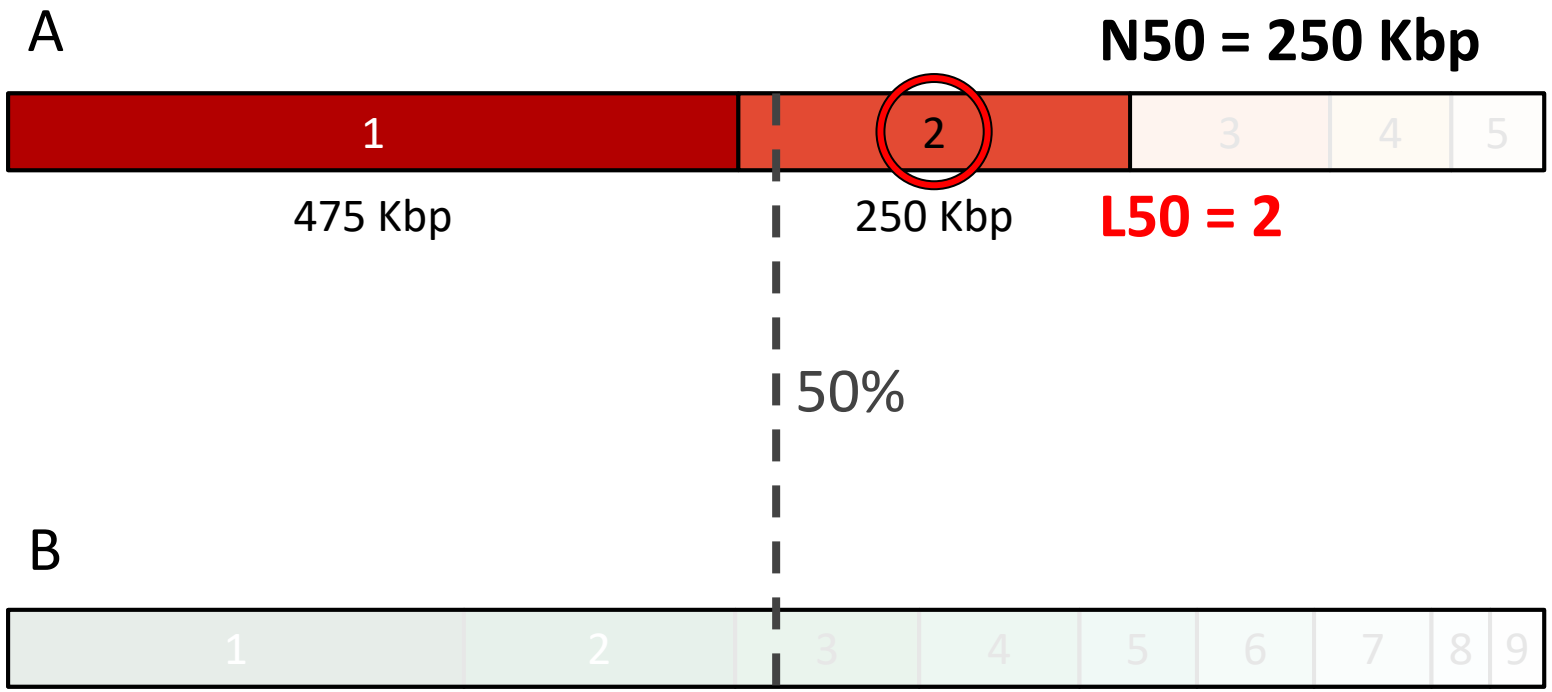
A

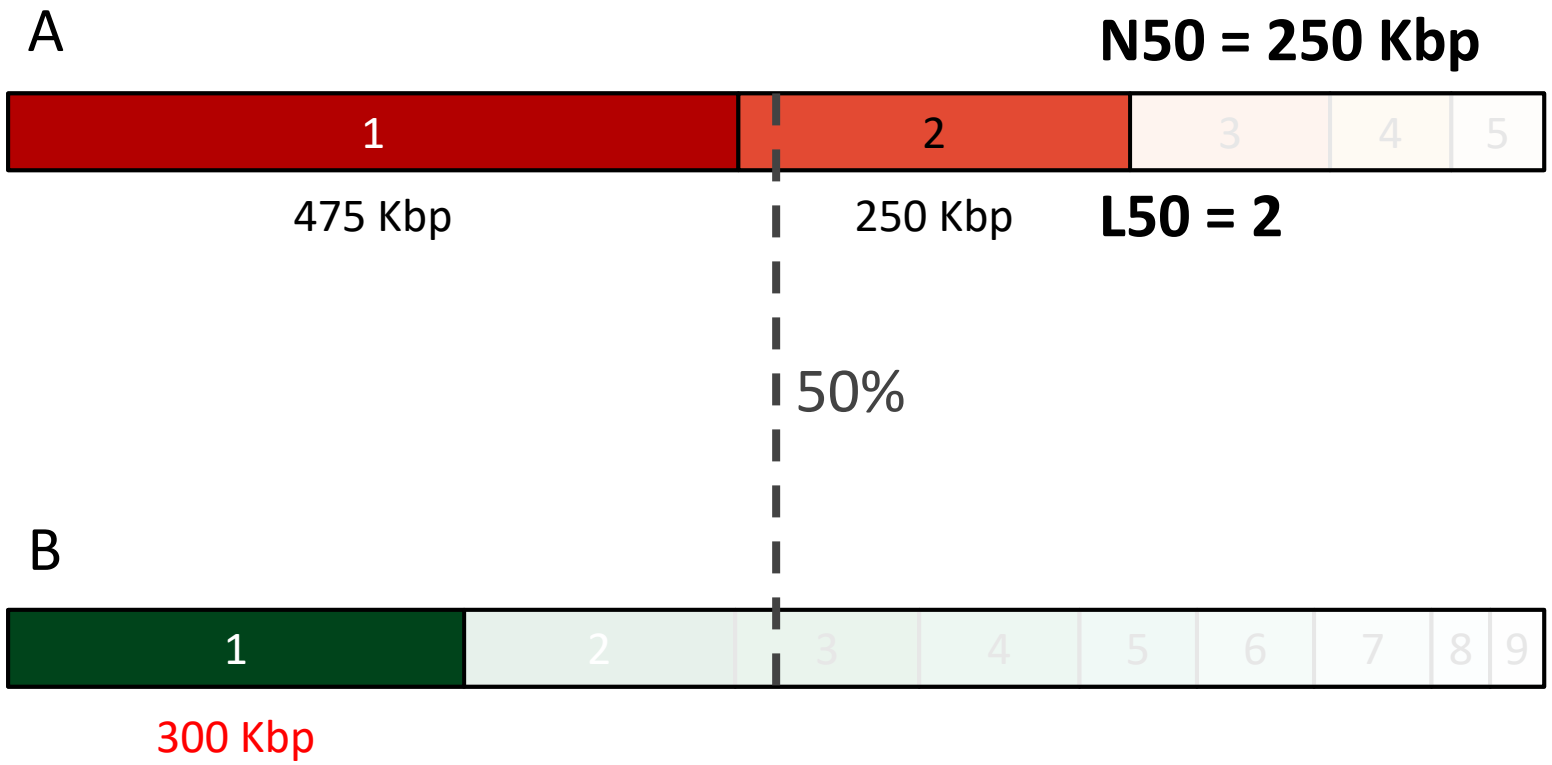
B

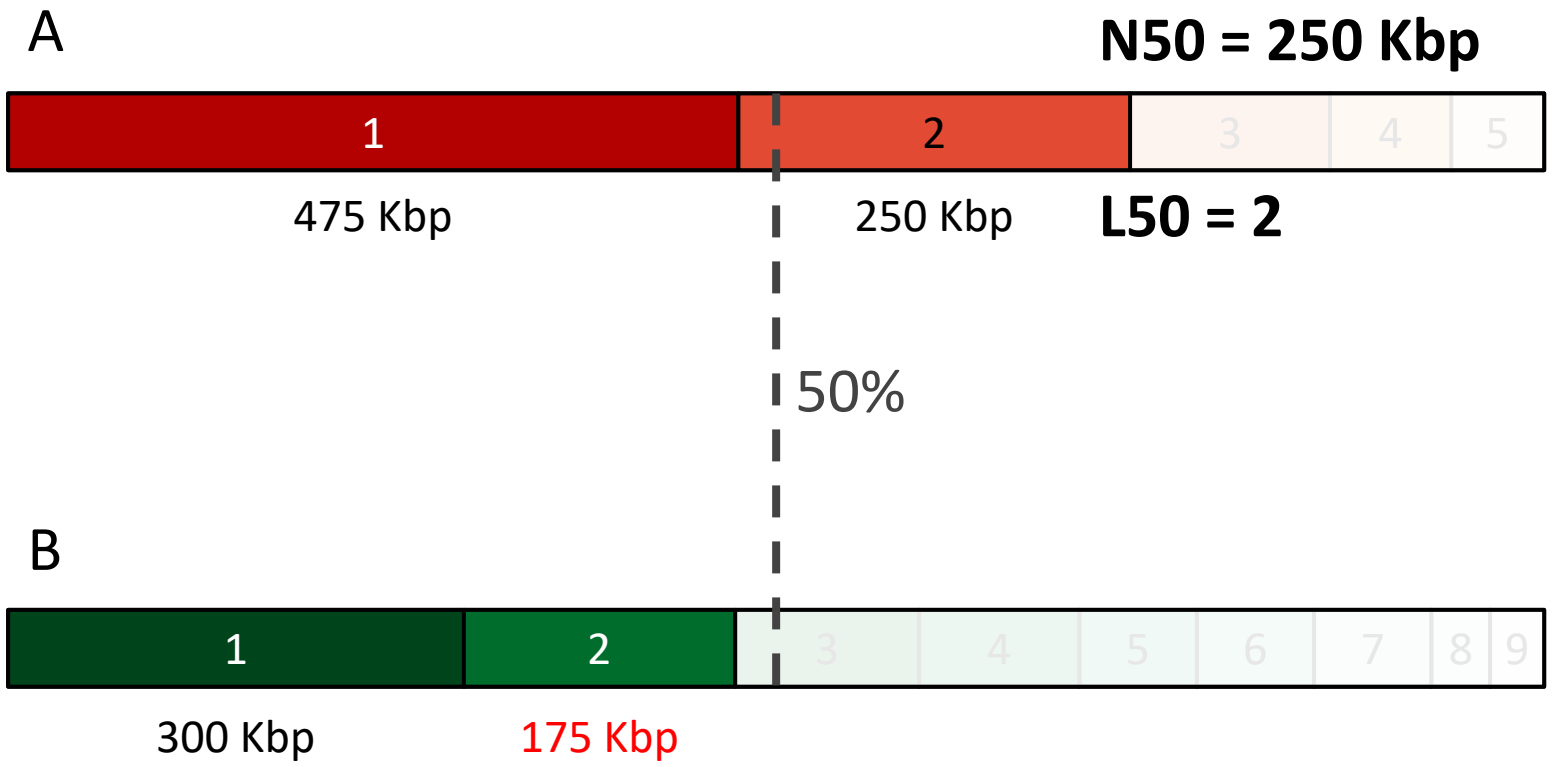

A
1
2
3
4
5
50%
B
1
2
3
4
5
6
7
8
9

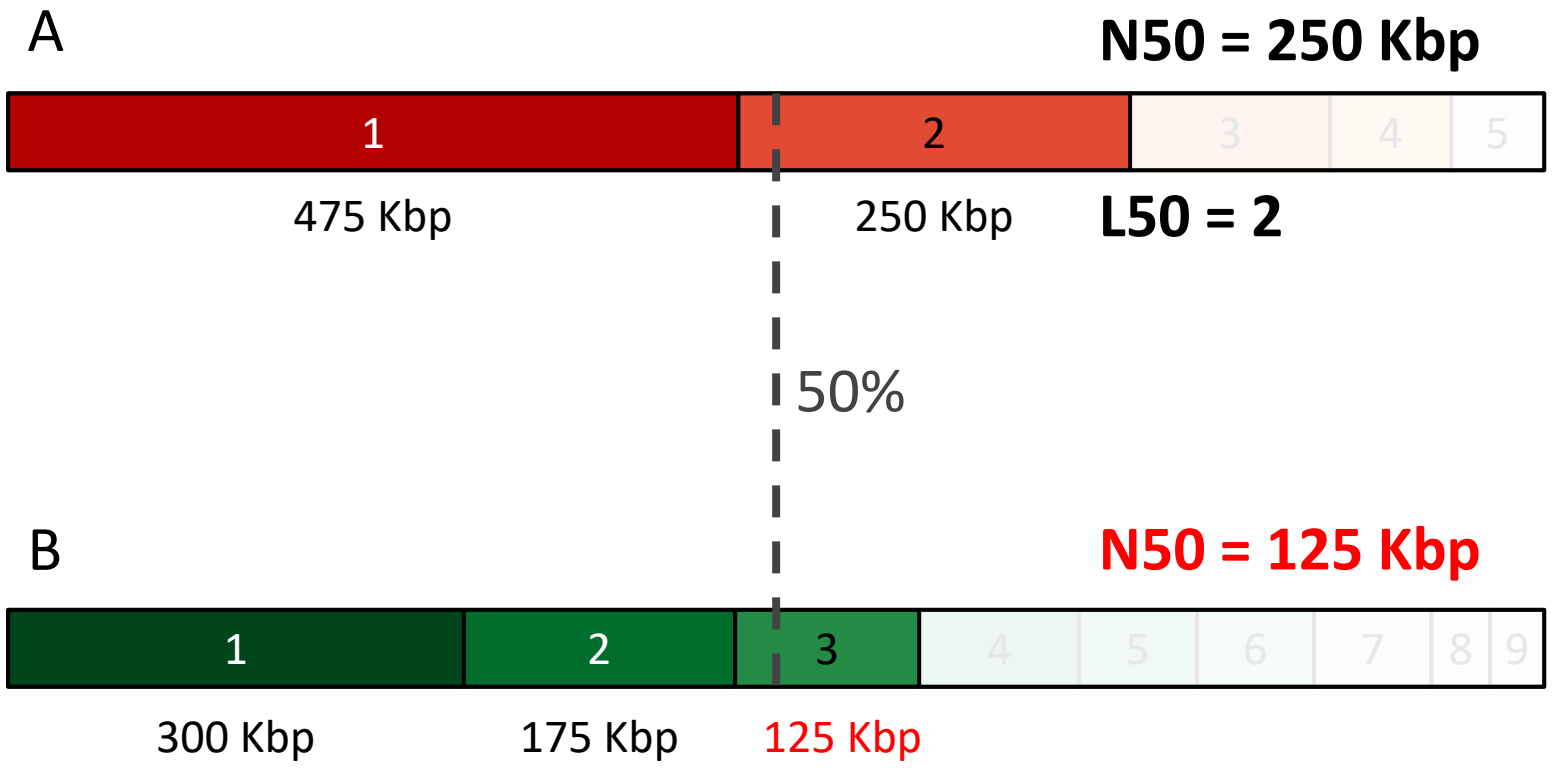

A
1
2
3
4
5
475 Kbp
50%
B
1
2
3
4
5
6
7
8
9

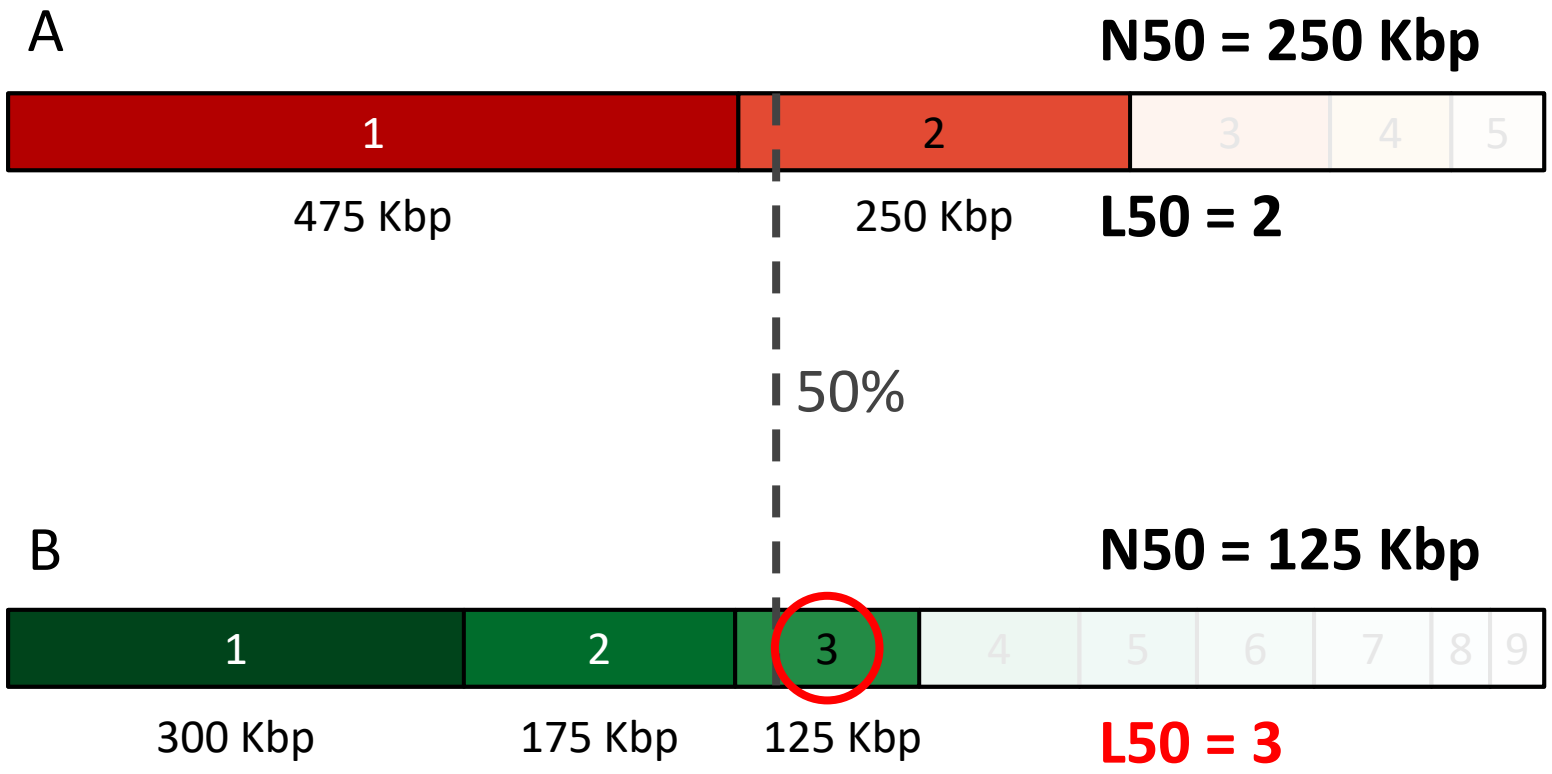

A
N50 = 250 Kbp
1
2
3
4
5
475 Kbp
250 Kbp
50%
B
1
2
3
4
5
6
7
8
9

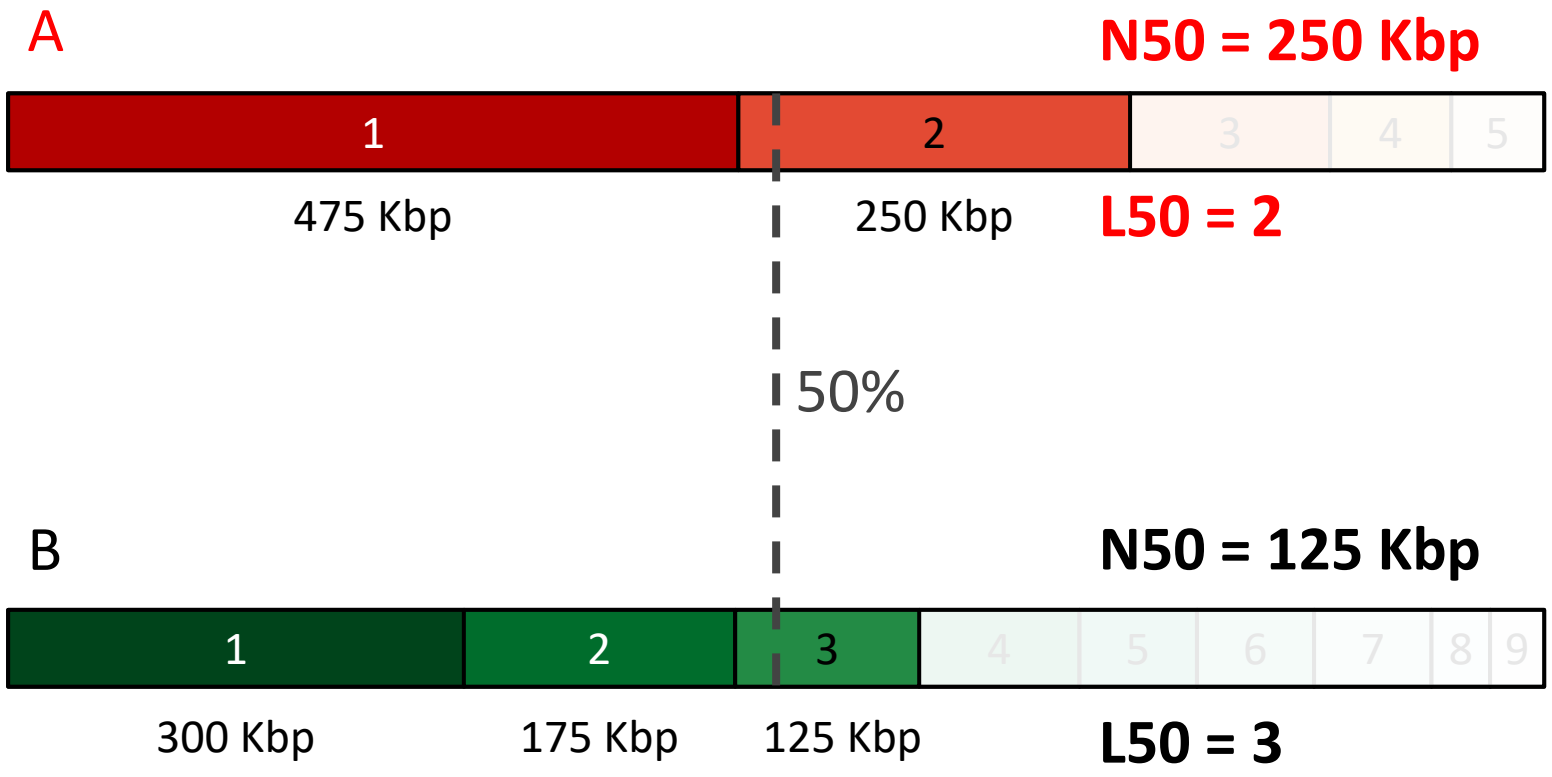

A
N50 = 250 Kbp
1
2
3
4
5
475 Kbp
250 Kbp
L50 = 2
50%
B
1
2
3
4
5
6
7
8
9


A
N50 = 250 Kbp
1
2
3
4
5
475 Kbp
250 Kbp
L50 = 2
50%
B
1
2
3
4
5
6
7
8
9
300 Kbp


A
N50 = 250 Kbp
1
2
3
4
5
475 Kbp
250 Kbp
L50 = 2
50%
B
1
2
3
4
5
6
7
8
9
300 Kbp
175 Kbp


A
N50 = 250 Kbp
1
2
3
4
5
475 Kbp
250 Kbp
L50 = 2
50%
B
N50 = 125 Kbp
1
2
3
4
5
6
7
8
9
300 Kbp
175 Kbp
125 Kbp

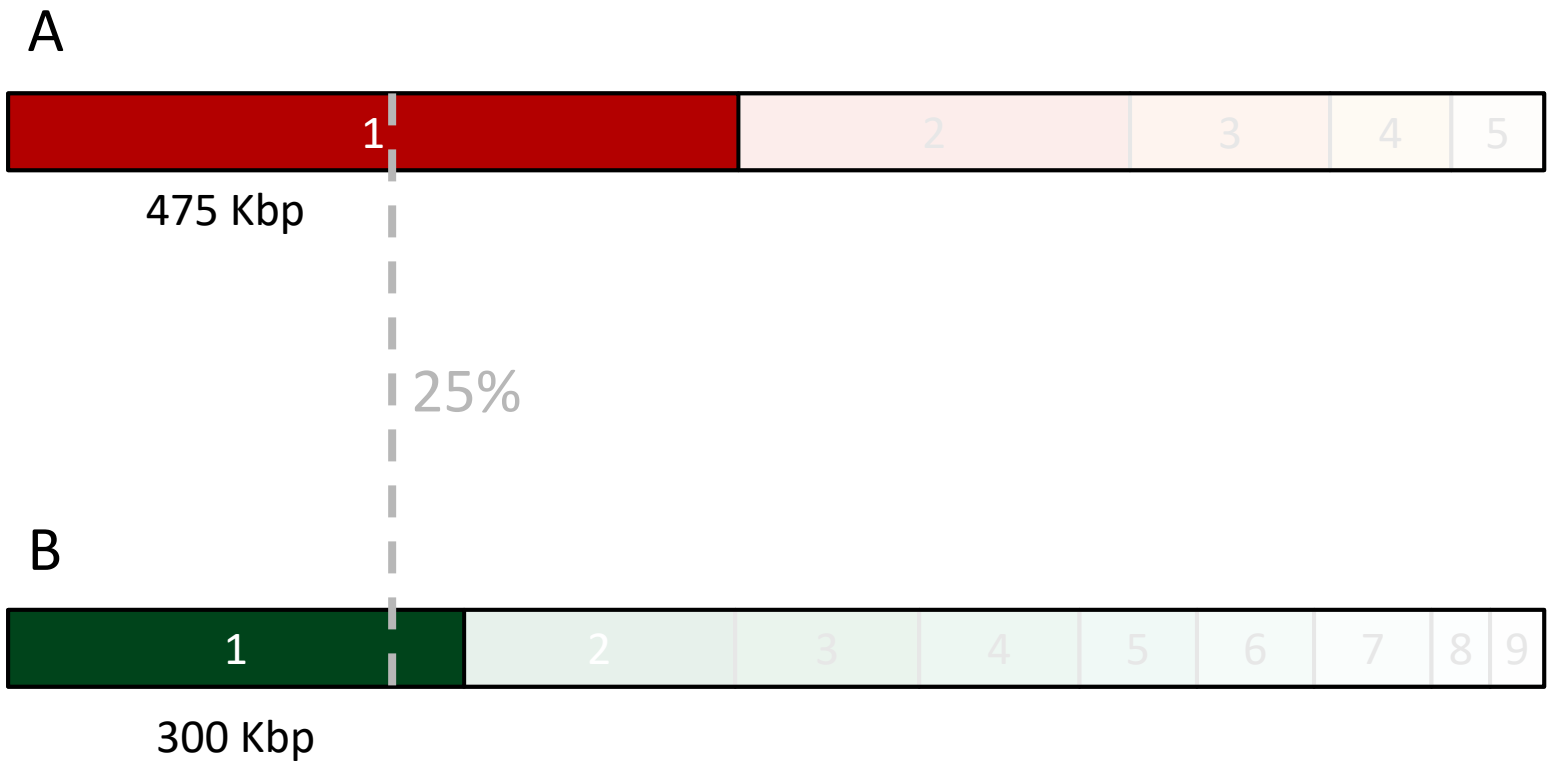

A
N50 = 250 Kbp
1
2
3
4
5
475 Kbp
250 Kbp
L50 = 2
50%
B
N50 = 125 Kbp
1
2
3
4
5
6
7
8
9
300 Kbp
175 Kbp
125 Kbp
L50 = 3



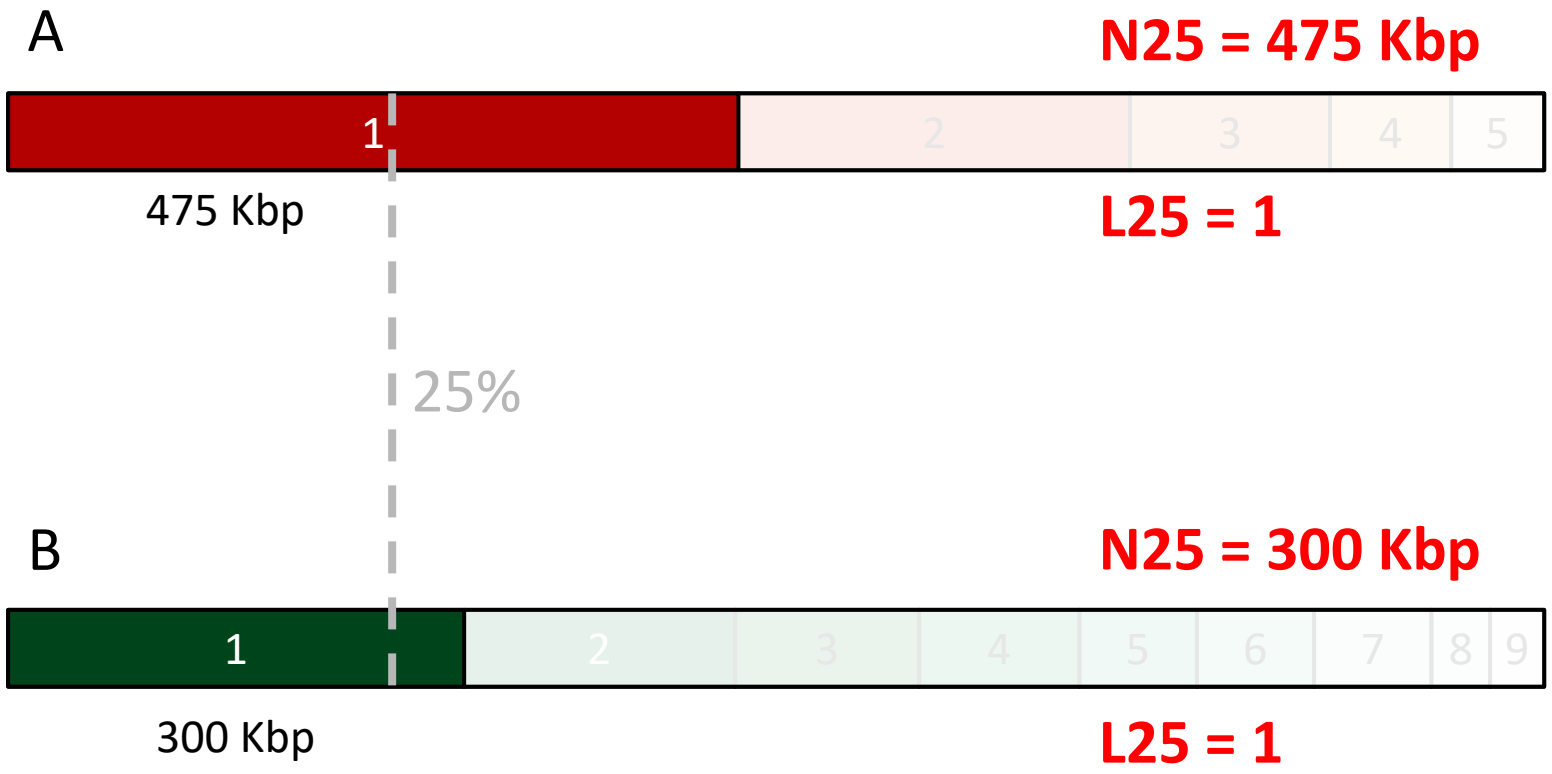
Which assembly is more complete?

A

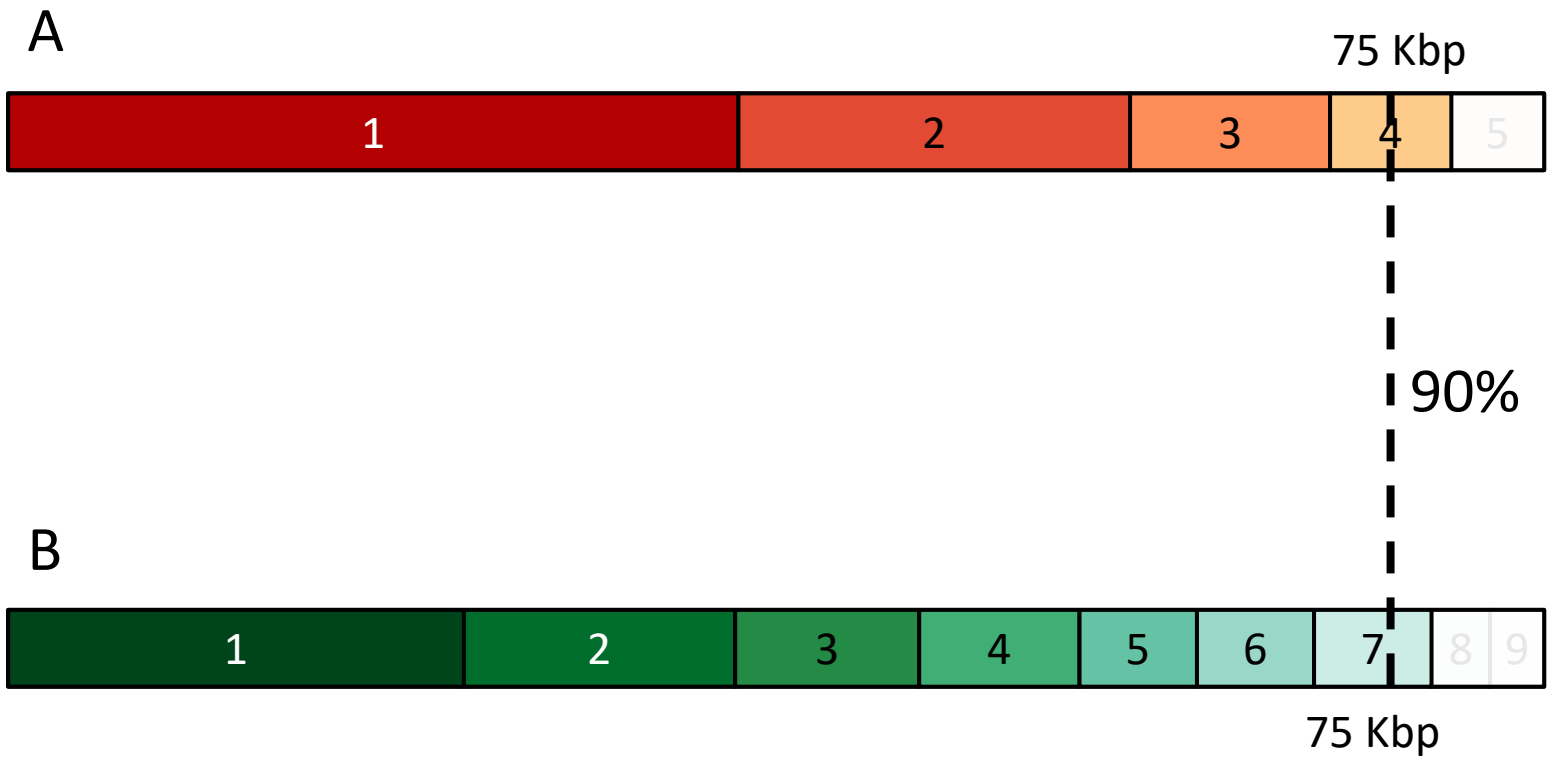
N25 & L25



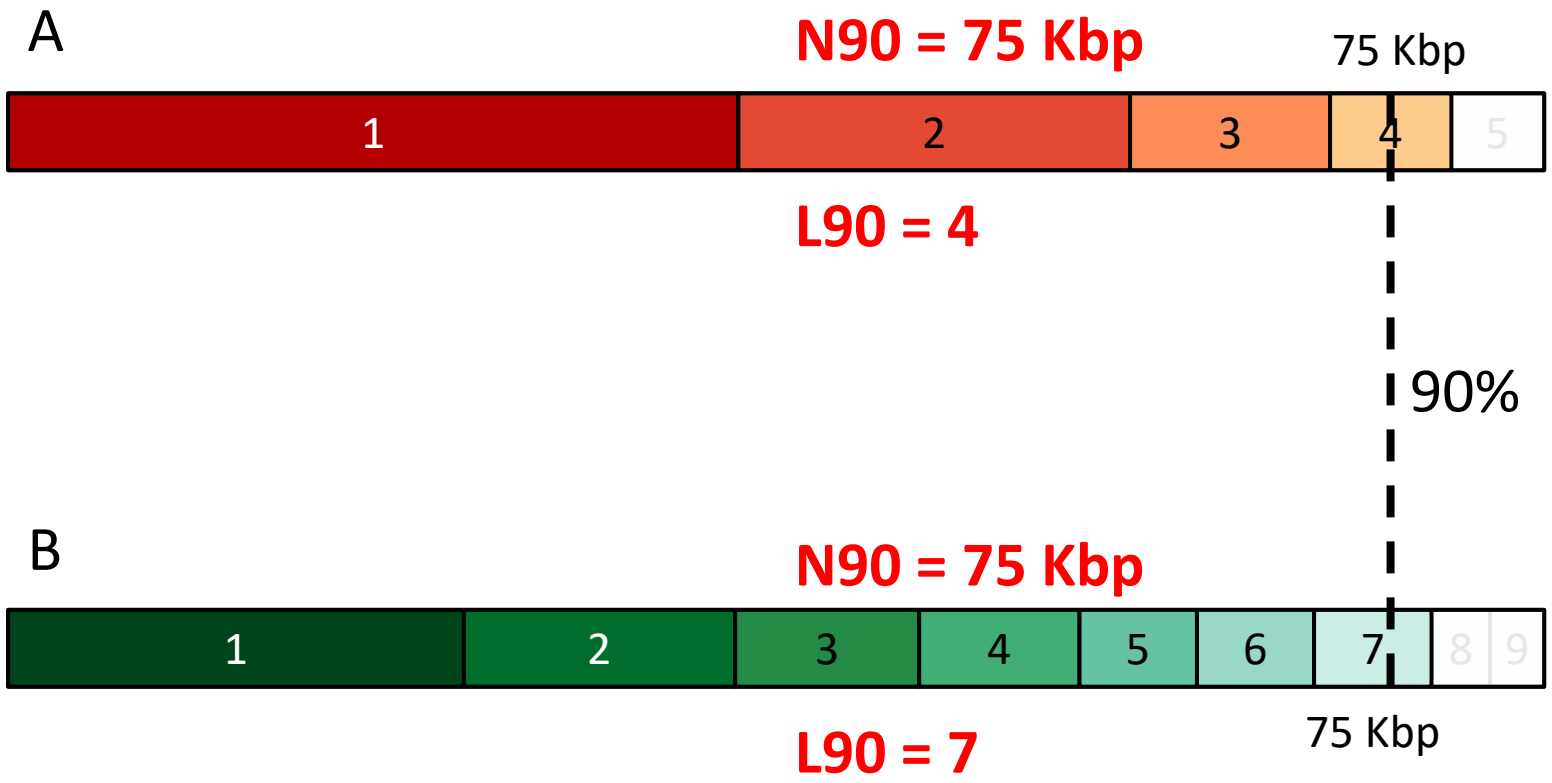
N25 & L25


A
N25 = 475 Kbp
1
2
3
4
5
475 Kbp
L25 = 1
25%
B
N25 = 300 Kbp
1
2
3
4
5
6
7
8
9
300 Kbp
L25 = 1

N90 & L90



N90 & L90


A
N90 = 75 Kbp
75 Kbp
1
2
3
4
5
L90 = 4
90%
B
N90 = 75 Kbp
1
2
3
4
5
6
7
8
9
75 Kbp
L90 = 7

QUAST

github.com/ablab/quast

QUAST

Quality Assessment Tool for Genome Assemblies by CAB

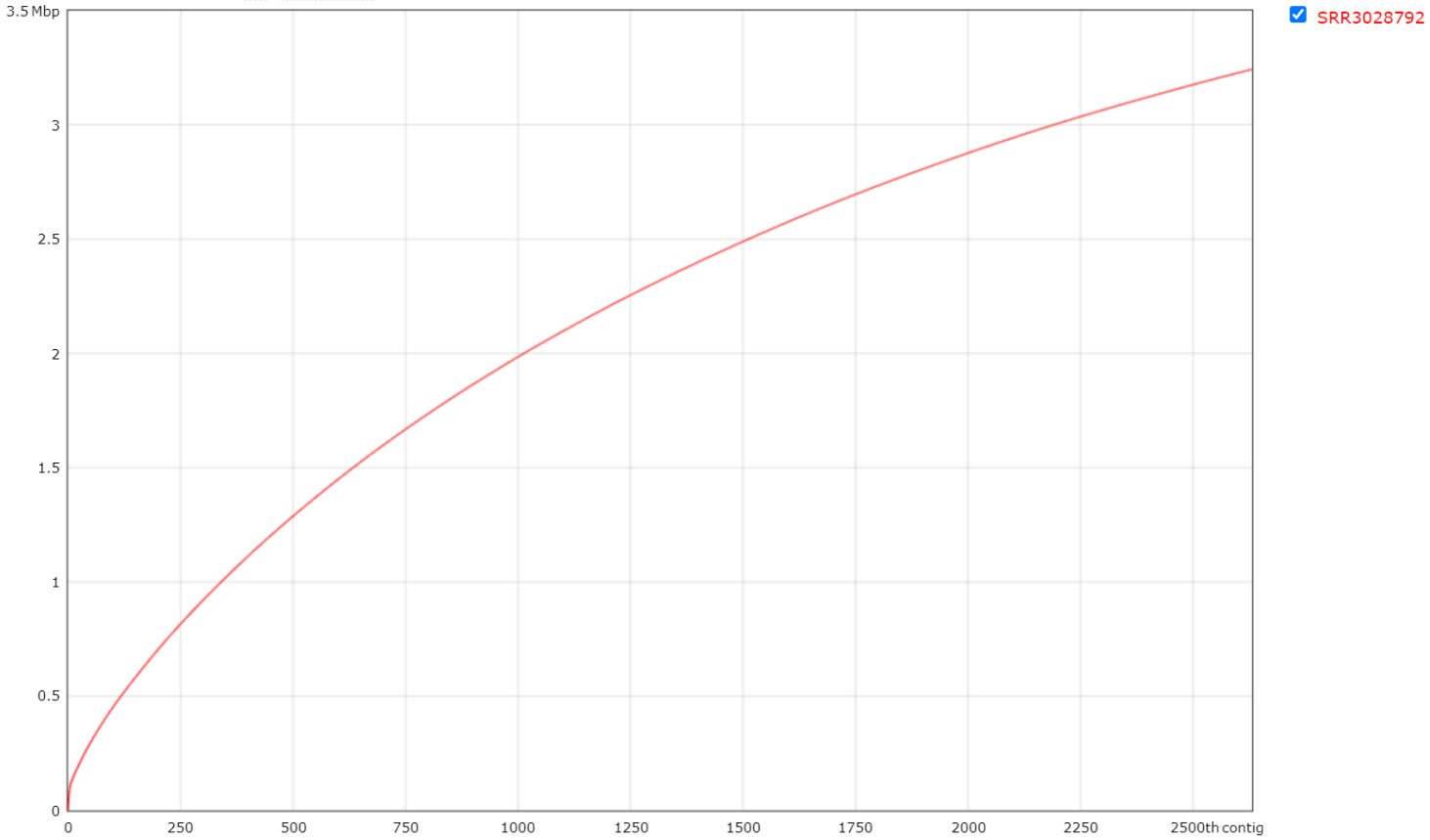
24 January 2022, Monday, 10:26:20

View in Icarus contig browser

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Statistics without reference	SRR3028792
# contigs	2631
# contigs (>= 0 bp)	3463
# contigs (>= 1000 bp)	1254
# contigs (>= 5000 bp)	10
# contigs (>= 10000 bp)	4
# contigs (>= 25000 bp)	2
# contigs (>= 50000 bp)	0
Largest contig	33 114
Total length	3 243 061
Total length (>= 0 bp)	3 579 102
Total length (>= 1000 bp)	2 258 968
Total length (>= 5000 bp)	137 834
Total length (>= 10000 bp)	102 283
Total length (>= 25000 bp)	60 964
Total length (>= 50000 bp)	0
N50	1415
N75	907
L50	717
L75	1437
GC (%)	51.83
Mismatches	
# N's	0
# N's per 100 kbp	0

Plots: Cumulative length Nx GC content



Contigs are ordered from largest (contig #1) to smallest.

QUAST

github.com/ablab/quast

- **Install with conda**

```
$ conda create -n quast quast
```

- **Run quast to generate quality report**

```
$ conda run -n quast quast contigs.fasta [contigs2.fasta]…
```

QUAST

github.com/ablab/quast

- **Install with conda**

```
$ conda create -n quast quast
```

- **Run quast to generate quality report**

```
$ conda run -n quast quast contigs.fasta [contigs2.fasta]…
```

- **Output**

quast_results/

QUAST

github.com/ablab/quast

- **Install with conda**

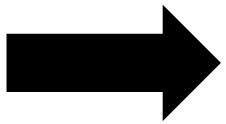
```
$ conda create -n quast quast
```

- **Run quast to generate quality report**

```
$ conda run -n quast quast contigs.fasta [contigs2.fasta]…
```

- **Output**

quast_results/



report.html

Interested primarily in the report

Quality report

github.com/ablab/quast

Worst Median Best

☑ Show heatmap

Statistics without reference	SRR3028792	SRR3028792_full
# contigs	2631	32
# contigs (>= 0 bp)	3463	38
# contigs (>= 1000 bp)	1254	27
# contigs (>= 5000 bp)	10	24
# contigs (>= 10000 bp)	4	24
# contigs (>= 25000 bp)	2	22
# contigs (>= 50000 bp)	0	18
Largest contig	33 114	1 022 474
Total length	3 243 061	4 839 502
Total length (>= 0 bp)	3 579 102	4 842 245
Total length (>= 1000 bp)	2 258 968	4 836 193
Total length (>= 5000 bp)	137 834	4 825 833
Total length (>= 10000 bp)	102 283	4 825 833
Total length (>= 25000 bp)	60 964	4 797 030
Total length (>= 50000 bp)	0	4 631 893
N50	1415	298 538
N75	907	184 096
L50	717	4
L75	1437	10
GC (%)	51.83	52.04
Mismatches		
# N's	0	0
# N's per 100 kbp	0	0

Quality report

github.com/ablab/quast

Worst Median Best

Show heatmap

Statistics without reference	SRR3028792	SRR3028792_full
# contigs	2631	32
# contigs (>= 0 bp)	3463	38
# contigs (>= 1000 bp)	1254	27
# contigs (>= 5000 bp)	10	24
# contigs (>= 10000 bp)	4	24
# contigs (>= 25000 bp)	2	22
# contigs (>= 50000 bp)	0	18
Largest contig	33 114	1 022 474
Total length	3 243 061	4 839 502
Total length (>= 0 bp)	3 579 102	4 842 245
Total length (>= 1000 bp)	2 258 968	4 836 193
Total length (>= 5000 bp)	137 834	4 825 833
Total length (>= 10000 bp)	102 283	4 825 833
Total length (>= 25000 bp)	60 964	4 797 030
Total length (>= 50000 bp)	0	4 631 893
N50	1415	298 538
N75	907	184 096
L50	717	4
L75	1437	10
GC (%)	51.83	52.04
Mismatches		
# N's	0	0
# N's per 100 kbp	0	0

Length: 3.2 Mbp vs. 4.8 Mbp

Quality report

github.com/ablab/quast

Worst Median Best

☑ Show heatmap

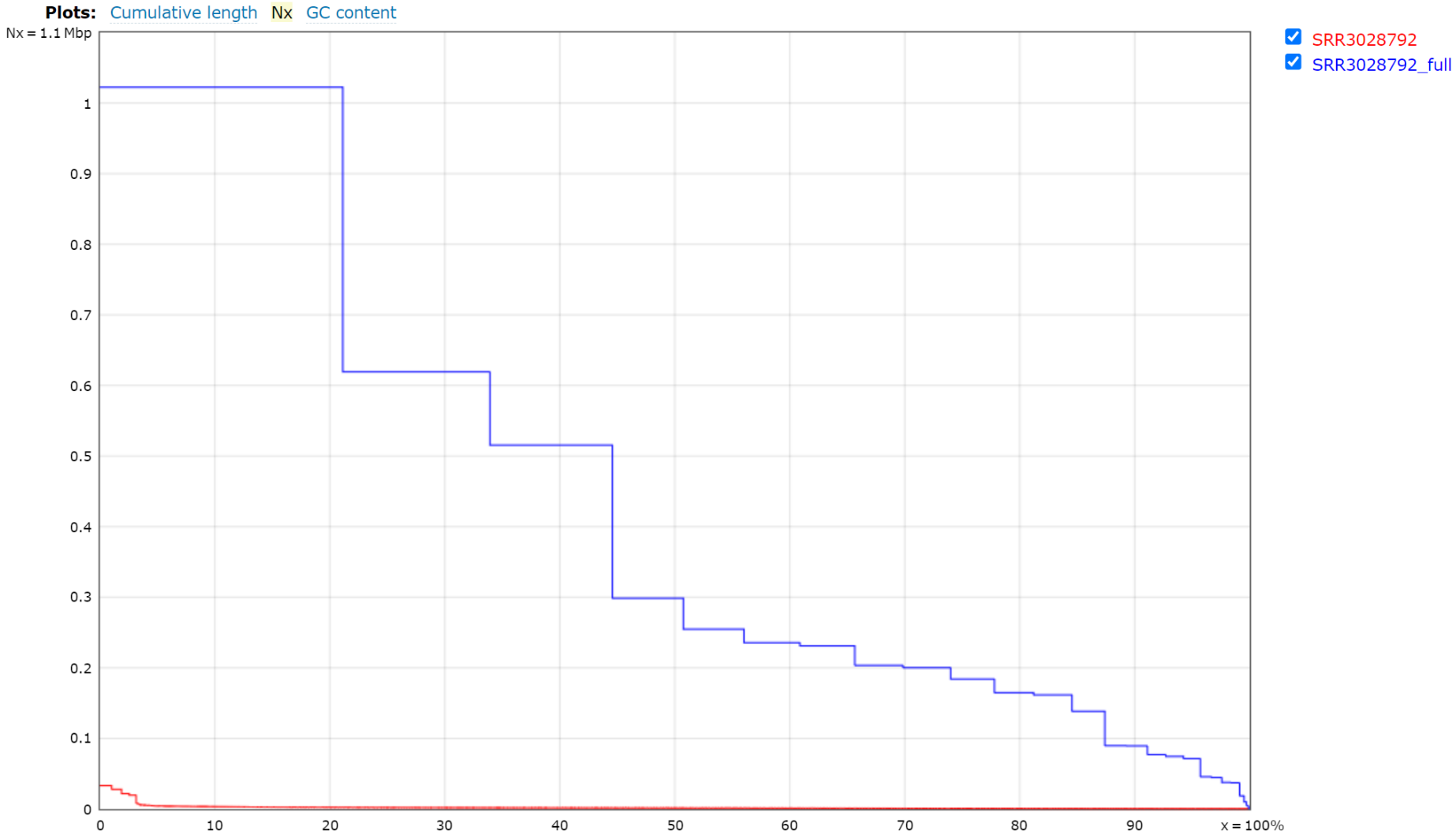
Statistics without reference	SRR3028792	SRR3028792_full
# contigs	2631	32
# contigs (>= 0 bp)	3463	38
# contigs (>= 1000 bp)	1254	27
# contigs (>= 5000 bp)	10	24
# contigs (>= 10000 bp)	4	24
# contigs (>= 25000 bp)	2	22
# contigs (>= 50000 bp)	0	18
Largest contig	33 114	1 022 474
Total length	3 243 061	4 839 502
Total length (>= 0 bp)	3 579 102	4 842 245
Total length (>= 1000 bp)	2 258 968	4 836 193
Total length (>= 5000 bp)	137 834	4 825 833
Total length (>= 10000 bp)	102 283	4 825 833
Total length (>= 25000 bp)	60 964	4 797 030
Total length (>= 50000 bp)	0	4 631 893
N50	1415	298 538
N75	907	184 096
L50	717	4
L75	1437	10
GC (%)	51.83	52.04
Mismatches		
# N's	0	0
# N's per 100 kbp	0	0

Length: 3.2 Mbp vs. 4.8 Mbp

N50: 1.4 Kbp vs. 300 Kbp

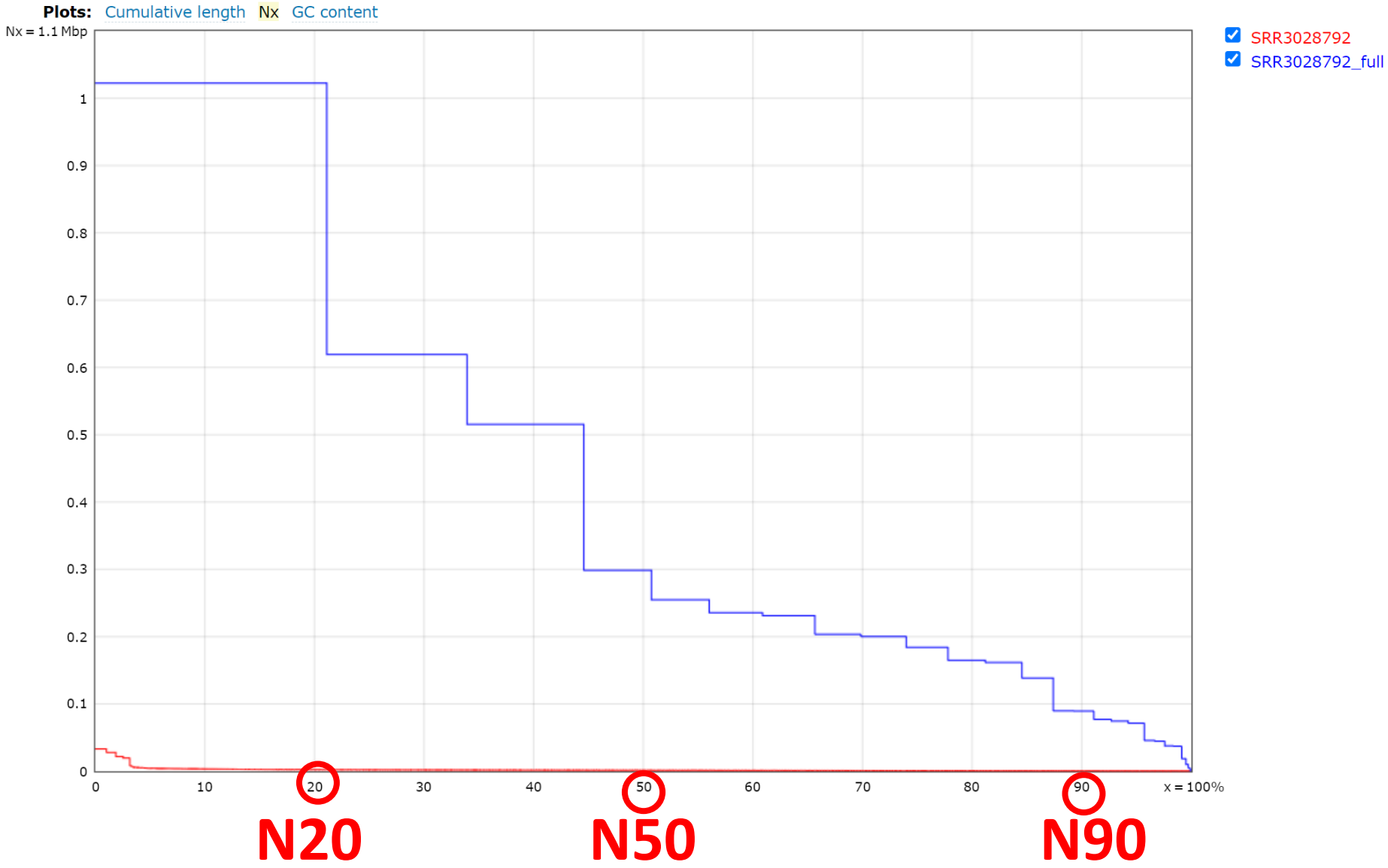
Quality report

github.com/ablab/quast



Quality report

github.com/ablab/quast



LEARNING OBJECTIVES

1. *What is microbial sequencing*
2. *What is genome assembly*
3. *Where to find sequence data*
4. *How to download*
5. *Quality of reads*
6. *How to assemble data*
7. *How to evaluate quality of the assembly*
8. ***Data analysis tutorial***

Full tutorial on genome assembly

github.com/MMID-coding-workshop/2022-01-26-Downloading-and-assembling-microbial-sequence-data

Downloading and assembling microbial sequence data

launch binder

This repository contains the slides and additional materials for the **Downloading and assembling microbial sequence data** MMID Coding workshop for **January 26, 2022**.

1. Interactive Jupyter notebook

The commands used for downloading and assembling microbial genomes is provided as an interactive Jupyter notebook:

- microbial-genome-assembly.ipynb
  - If you wish to launch this notebook in a cloud-based environment to follow along please click the launch binder link.

Full tutorial on genome assembly

github.com/MMID-coding-workshop/2022-01-26-Downloading-and-assembling-microbial-sequence-data

Downloading and assembling microbial sequence data

launch binder

This repository contains the slides and additional materials for the **Downloading and assembling microbial sequence data** MMID Coding workshop for **January 26, 2022**.

1. Interactive Jupyter notebook

The commands used for downloading and ed as an interactive Jupyter notebook:

View workshop tutorial (command-line)

- microbial-genome-assembly.ipynb
  - If you wish to launch this notebook in a cloud-based environment to follow along please click the launch binder link.

Run using command-line

Copy this command to terminal

Expected output

2.2. Install `sra-tools` using conda

Now we can install `sra-tools`. The below command will create a new conda environment `sra-tools` and install the package `sra-tools` in this environment.

The `-y` option here means to automatically answer `yes` to any prompts for input by conda. This is important when running using Jupyter (but can be left out if you are running via the command-line).

```
[2]: conda create -y -n sra-tools sra-tools
```

```
Collecting package metadata (current_repodata.json): done
Solving environment: done

## Package Plan ##

  environment location: /home/CSCScience.ca/apetkau/miniconda3/envs/sra-tools

  added / updated specs:
    - sra-tools


The following NEW packages will be INSTALLED:
```

Full tutorial on genome assembly

github.com/MMID-coding-workshop/2022-01-26-Downloading-and-assembling-microbial-sequence-data

Downloading and assembling microbial sequence data

launch binder

This repository contains the slides and additional materials for the **Downloading and assembling microbial sequence data** MMID Coding workshop for **January 26, 2022**.

1. Interactive Jupyter notebook

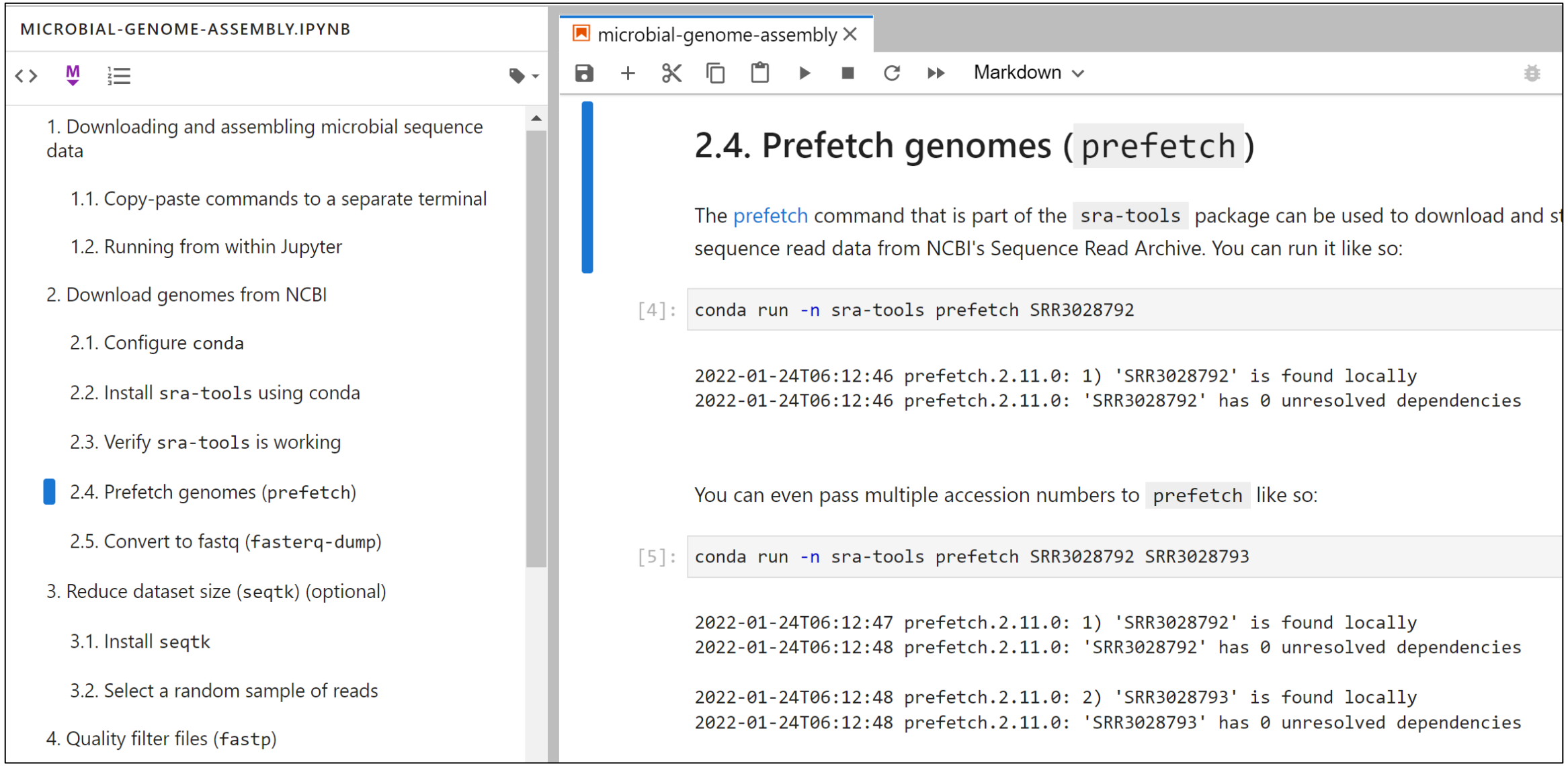
The commands used for downloading and assembling microbial genomes is provided as an interactive Jupyter notebook:

- microbial-genome-assembly.ipynb
  - If you wish to lau to follow along please click the launch binder

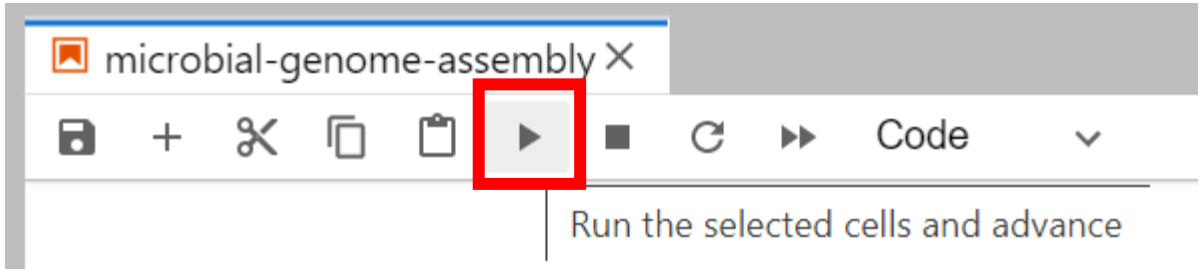
Launch tutorial in a cloud environment

Run tutorial using cloud

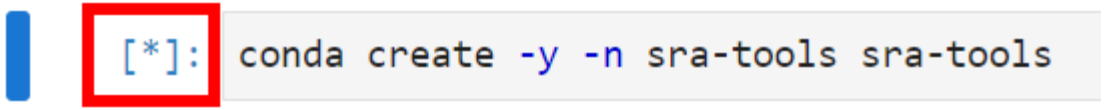
Uses Jupyter to provide instructions + BASH commands



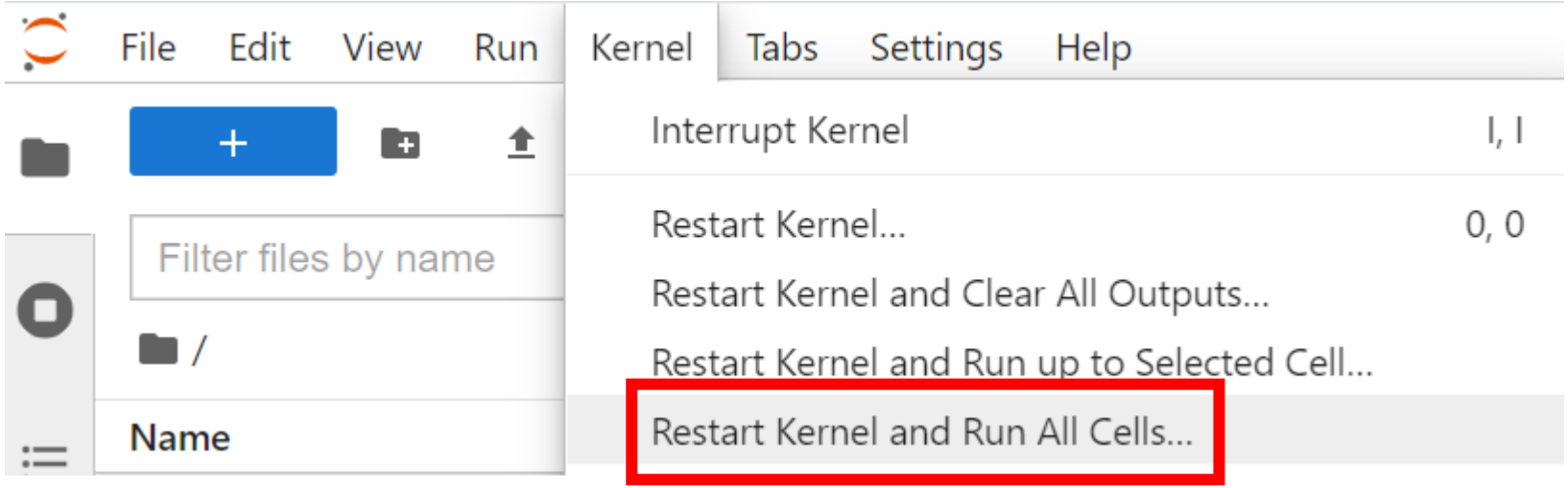
Run tutorial using cloud



Run by selecting "**Play**" button



Run by selecting "**Run All Cells**"



HELPFUL RESOURCES

Illumina Sequencing: youtu.be/fCd6B5HRaZ8

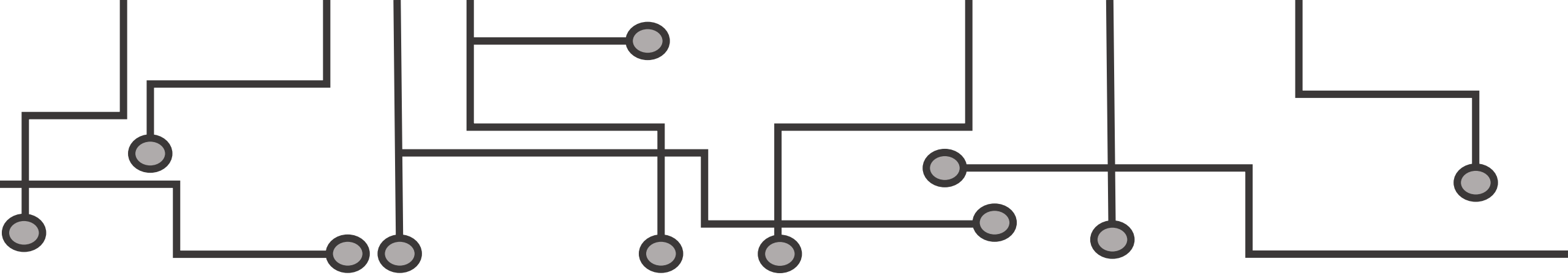
Nanopore Sequencing:
nanoporetech.com/resource-centre/introduction-nanopore-sequencing

Downloading from SRA: www.ncbi.nlm.nih.gov/sra/docs/sradownload/

Galaxy Bioinformatics Training: https://training.galaxyproject.org/
(Galaxy is a web-based bioinformatics analysis platform)

OTHER RESOURCES

Some icons from Font Awesome and used under a Creative Commons Attribution 4.0 International license



THANK YOU FOR ATTENDING!

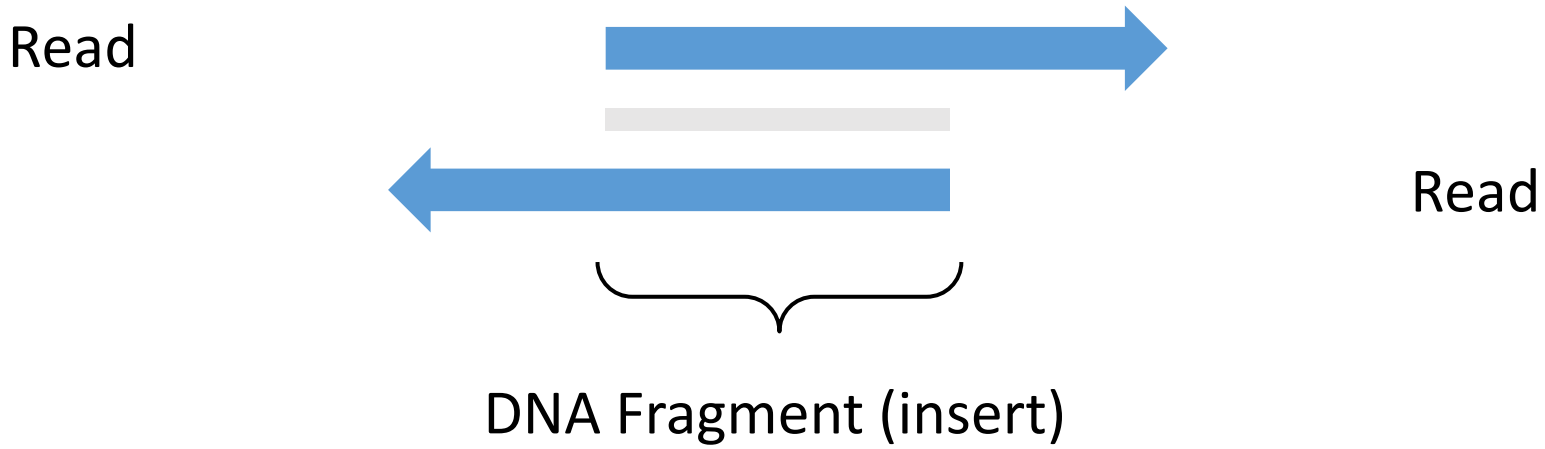
The Q&A Session will now begin.

Please make sure to fill out the Exit Survey

We value your feedback!

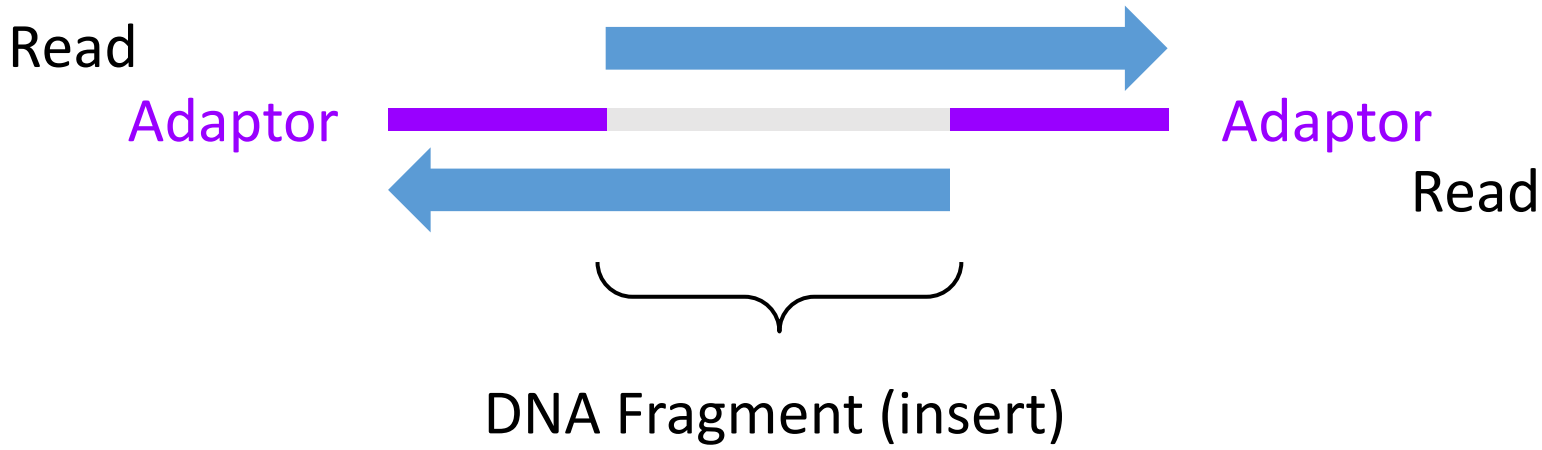
More questions? Please email us at
mmid.coding.workshop@gmail.com or post them to the workshop slack channel

Paired-end sequencing



...or very short.

Paired-end sequencing



...or very short.