

# 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.*

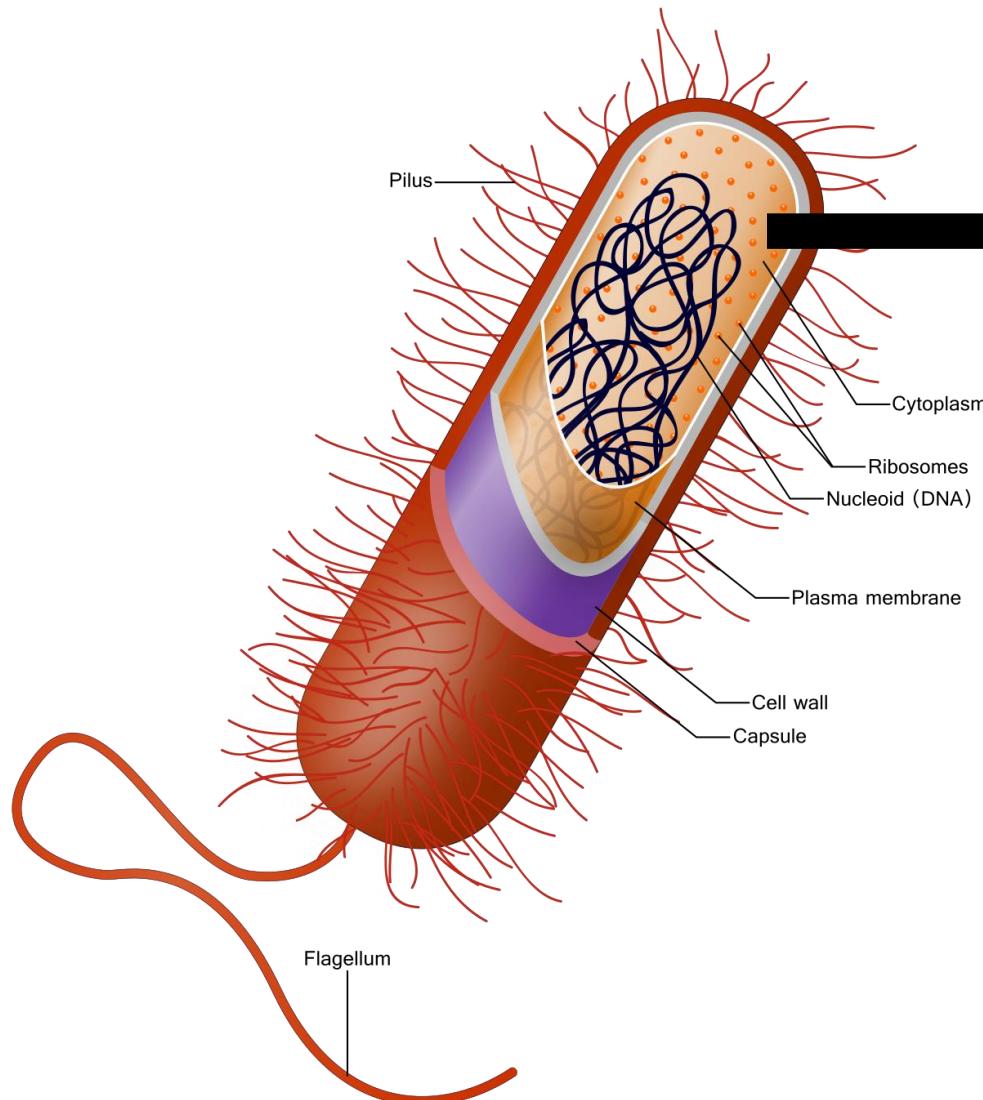
# LEARNING OBJECTIVES

- 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

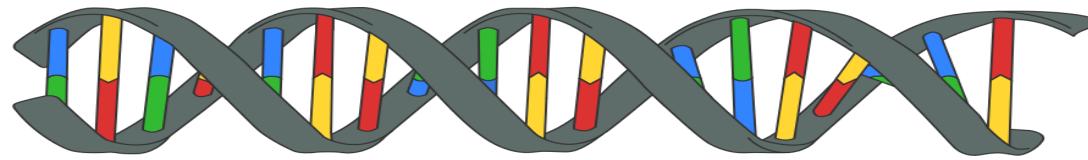
T

C

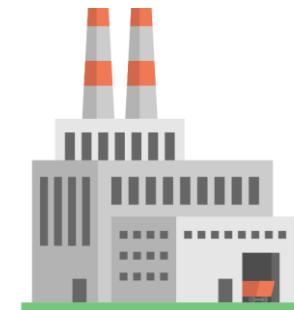
G

# Microbial whole-genome sequencing

DNA



Sequencing



Reads

A C A T

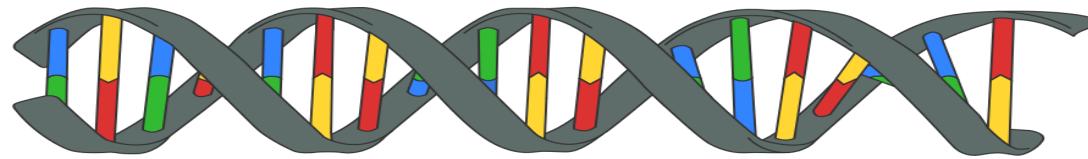
A G T T

T T A G

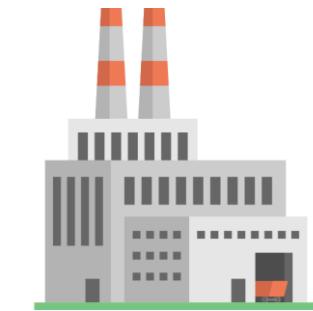
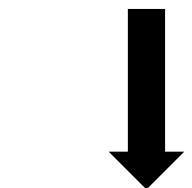
G A A T

# Microbial whole-genome sequencing

DNA

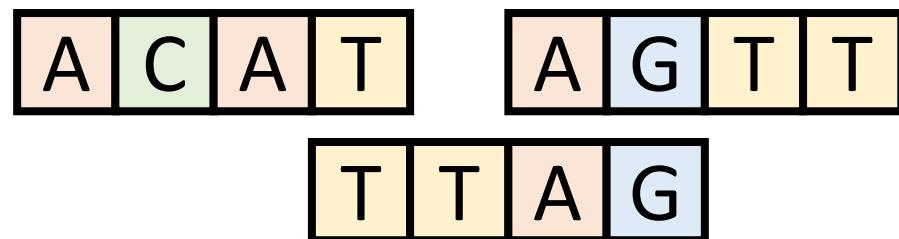


Fragment &  
Prepare DNA  
Library

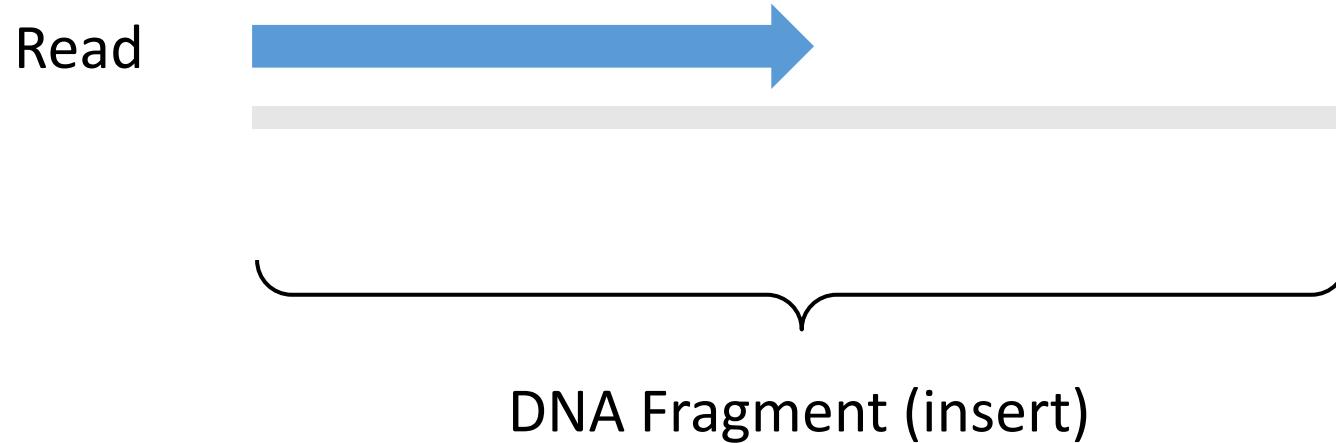


Sequencing  
(Illumina)

Reads

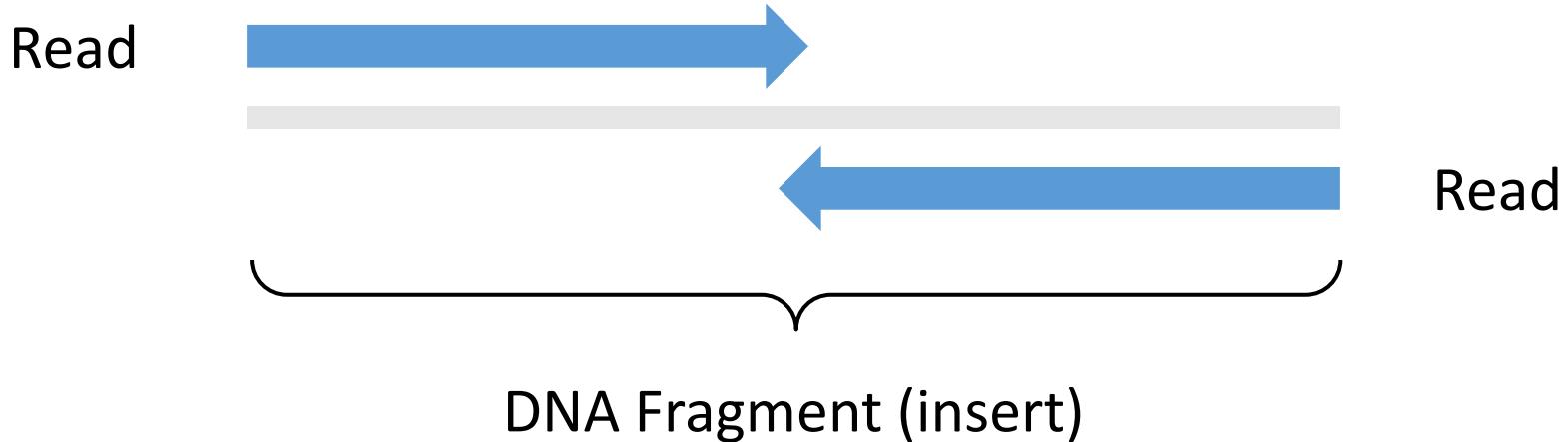


# Single-end sequencing



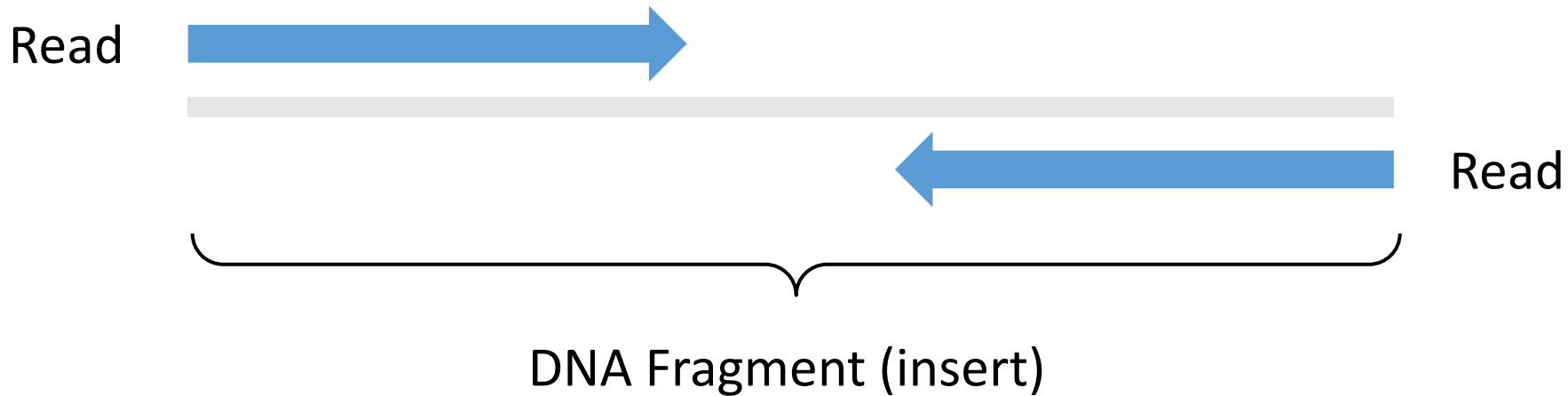
A DNA fragment is read once  
from one single end.

# Paired-end sequencing



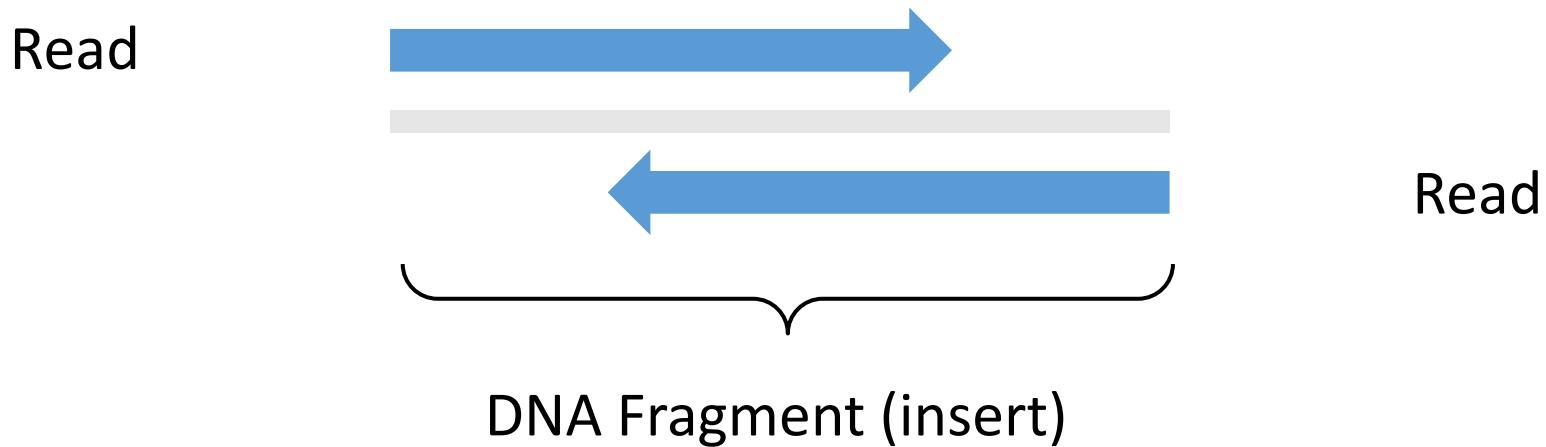
A DNA fragment is read twice,  
from each end.

# 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***

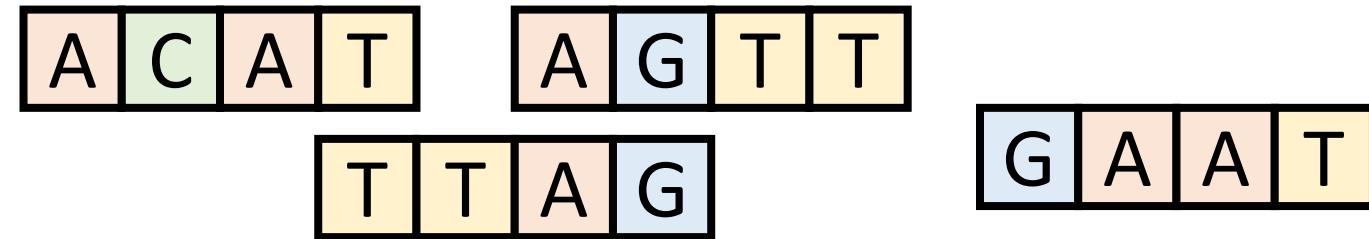
# Genome assembly



[https://en.wikipedia.org/wiki/Jigsaw\\_puzzle#/media/File:Sky\\_puzzle.jpg](https://en.wikipedia.org/wiki/Jigsaw_puzzle#/media/File:Sky_puzzle.jpg)

# Genome assembly

Reads



Transform reads to  
larger contiguous fragments  
("contigs")

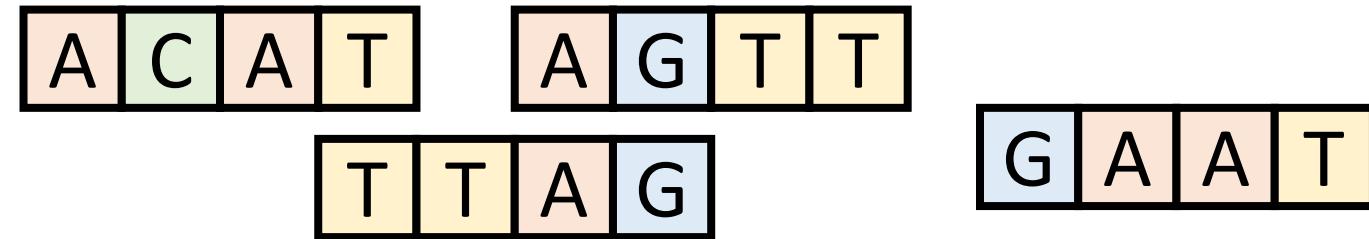


Assembly



# Genome assembly

Reads



Transform reads to  
larger contiguous fragments  
("contigs")



Assembly

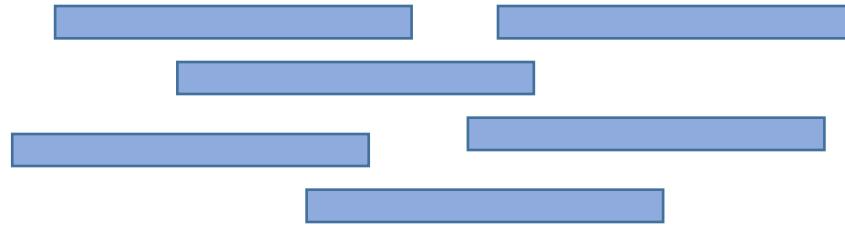


Breaks in contigs due to  
repeats or missing data

# Types of genome assembly

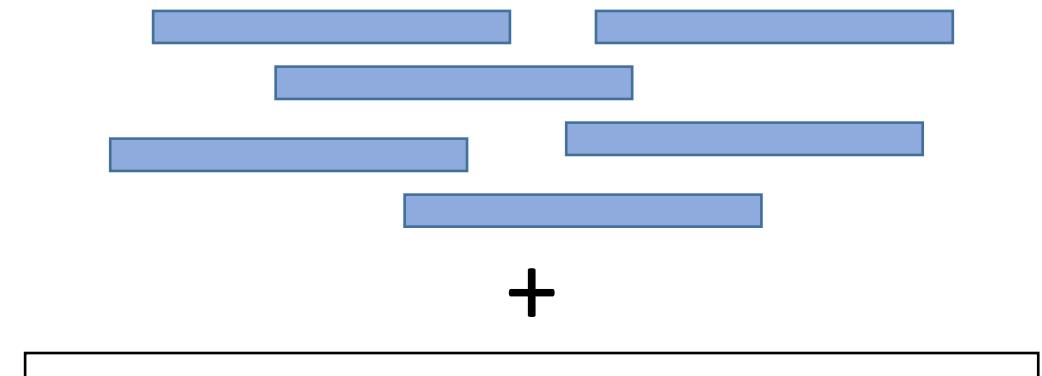
*De novo*  
assembly

Reads



Contigs

Reference-guided  
assembly

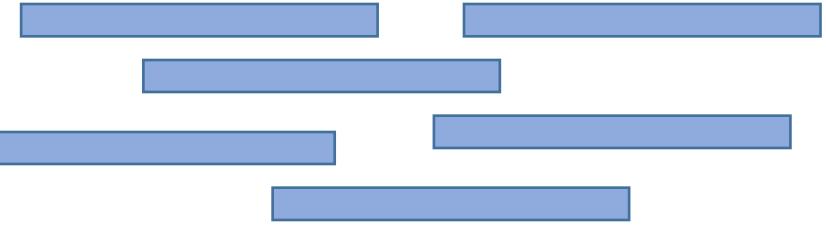


Reference genome

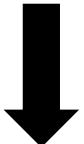
# Types of genome assembly

*De novo*  
assembly

Reads

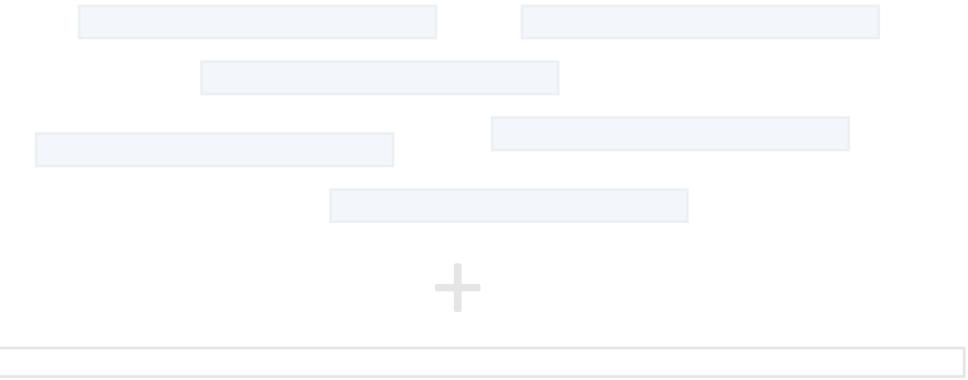


Focus on  
*de novo* assembly



Contigs

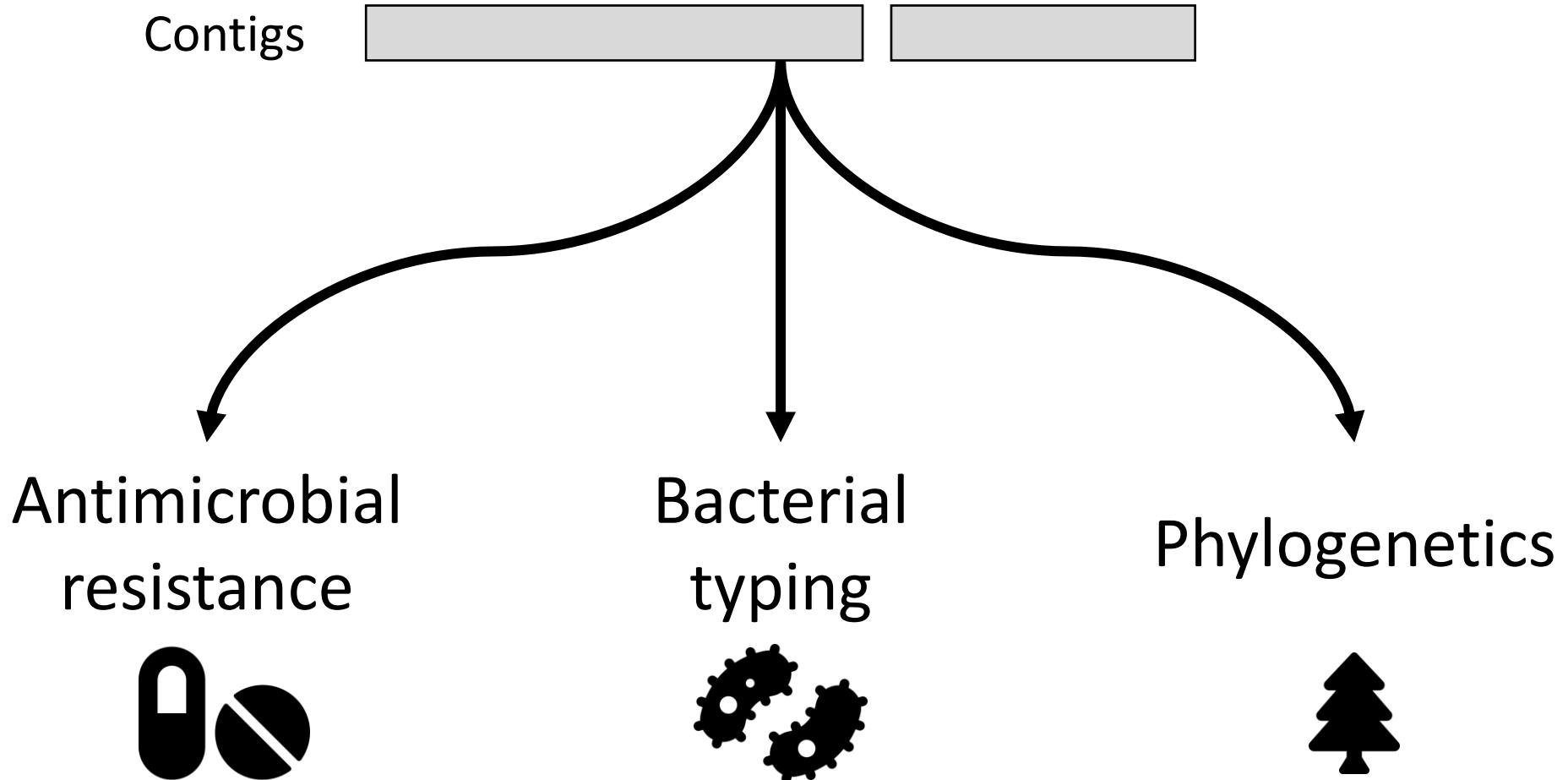
Reference-guided  
assembly



Reference genome



# 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)

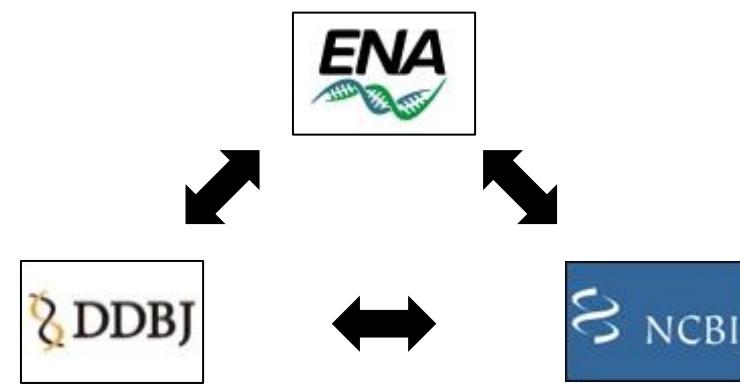


- Free and unrestricted access to nucleotide sequence and other data

# International Nucleotide Sequence Database Collaboration (INSDC)



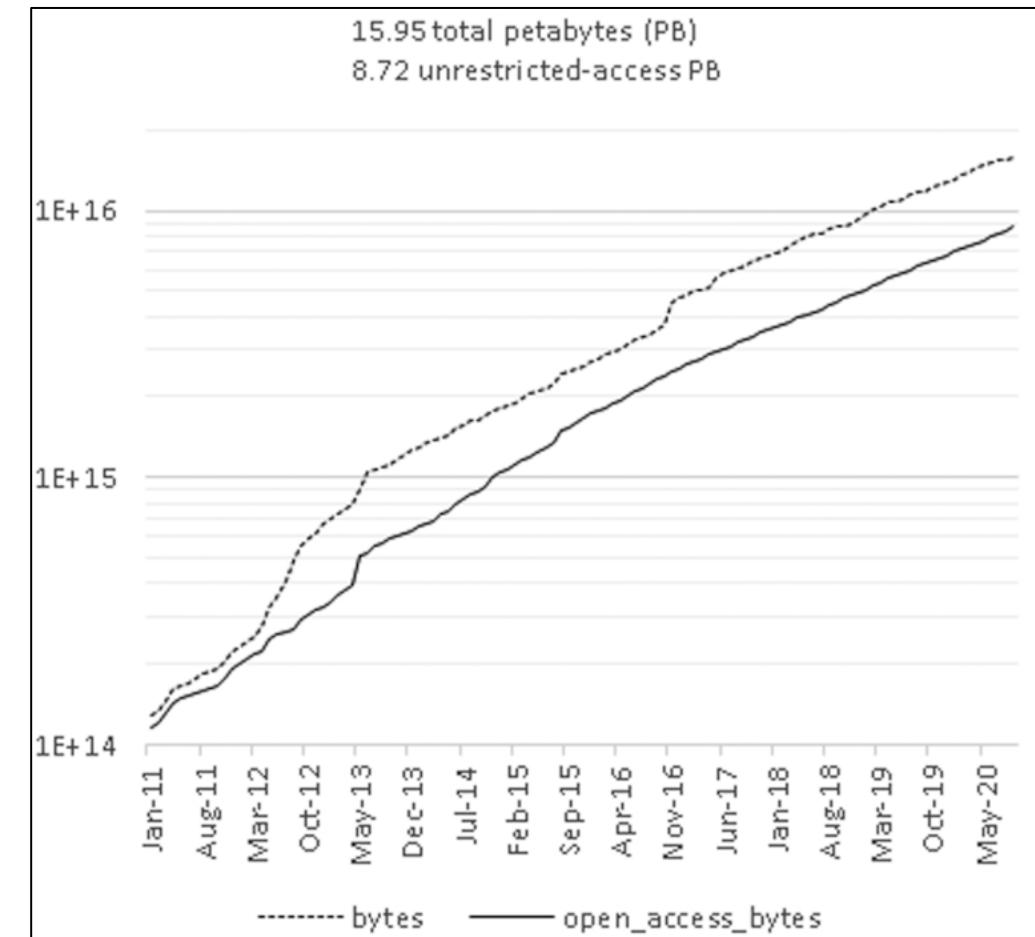
- Free and unrestricted access to nucleotide sequence and other data
- Data mirrored across three institutions



# International Nucleotide Sequence Database Collaboration (INSDC)



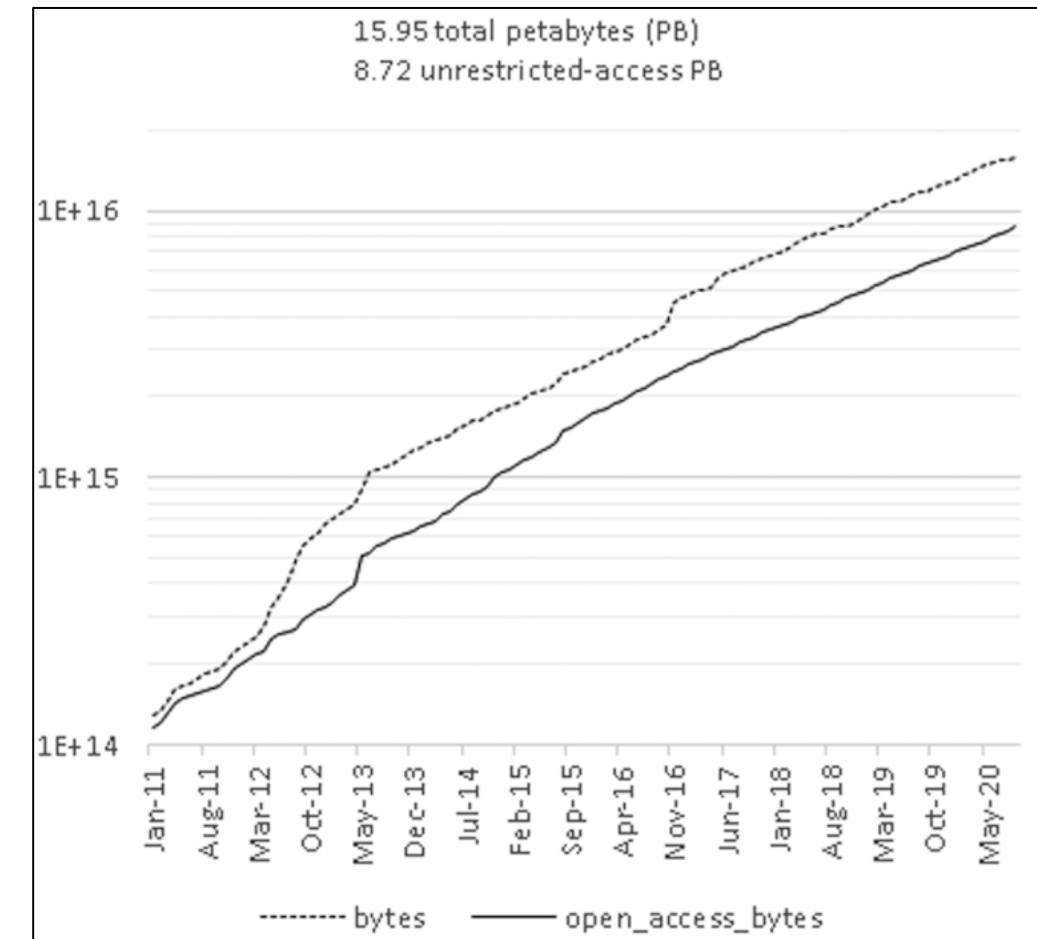
## Petabytes (PB) of data



# International Nucleotide Sequence Database Collaboration (INSDC)



## Petabytes (PB) of data



# NCBI's Sequence Read Archive (SRA)

NCBI Resources How To

SRA SRA Advanced Search Help

**COVID-19 Information** X

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)



**SRA - Now available on the cloud**

Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.

Getting Started	Tools and Software	Related Resources
<a href="#">How to Submit</a>	<a href="#">Download SRA Toolkit</a>	<a href="#">Submission Portal</a>
<a href="#">How to search and download</a>	<a href="#">SRA Toolkit Documentation</a>	<a href="#">Trace Archive</a>
<a href="#">How to use SRA in the cloud</a>	<a href="#">SRA-BLAST</a>	<a href="#">dbGaP Home</a>
<a href="#">Submit to SRA</a>	<a href="#">SRA Run Browser</a>	<a href="#">BioProject</a>
	<a href="#">SRA Run Selector</a>	<a href="#">BioSample</a>

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

# Search by organism

NCBI Resources ▾ How To ▾ Sign in to NCBI

SRA SRA **salmonella** Search Create alert Advanced Help

Access Summary ▾ 20 per page ▾ Send to: ▾ Filters: Manage Filters

Controlled (50) Public (476,506)

Source DNA (468,846) RNA (8,411)

Type exome (3) genome (460,289)

Library Layout paired (465,516) single (12,014)

Platform ABI SOLiD (116) BGISEQ (430) Capillary (127) Complete Genomics (2) Helicos (5) Illumina (472,594) Ion Torrent (888) LS454 (853)

**Search results**  
Items: 1 to 20 of 477530

<< First < Prev Page 1 of 23877 Next > Last >>

[Other Sequencing of \*\*Salmonella enterica\*\*](#)  
1. 1 ILLUMINA (Illumina MiSeq) run: 783,773 spots, 376.4M bases, 211.2Mb downloads  
Accession: SRX13842631

[Other Sequencing of \*\*Salmonella enterica\*\*](#)  
2. 1 ILLUMINA (Illumina MiSeq) run: 981,089 spots, 470.5M bases, 263.3Mb downloads  
Accession: SRX13840534

[Genome-Seq of \*\*Salmonella Heidelberg\*\* 35 from chicken](#)  
3. 1 ILLUMINA (Illumina MiSeq) run: 445,169 spots, 210.6M bases, 140.2Mb downloads  
Accession: SRX13839406

[Other Sequencing of \*\*Salmonella enterica\*\*](#)  
4. 1 ILLUMINA (Illumina MiSeq) run: 1.1M spots, 307.6M bases, 170Mb downloads  
Accession: SRX13834337

**Results by taxon**

Top Organisms [Tree]  
[Salmonella enterica](#) (458077)  
[Homo sapiens](#) (2889)  
[Mus musculus](#) (2130)  
[Escherichia coli](#) (1520)  
[pig gut metagenome](#) (1346)  
[All other taxa](#) (11568)  
[More...](#)

**Search in related databases**

Database	Access		all
	public	controlled	
BioSample	<a href="#">453,968</a>	<a href="#">12</a>	<a href="#">453,980</a>
BioProject	<a href="#">2,916</a>		<a href="#">2,916</a>
dbGaP		<a href="#">3</a>	<a href="#">3</a>
GEO Datasets	<a href="#">5,029</a>		<a href="#">5,029</a>

# Search by identifiers

## Data identifiers in paper

The image shows a journal article from the Journal of Clinical Microbiology. At the top left is the American Society for Microbiology logo and the journal title "Journal of Clinical Microbiology". On the right side of the header is a CrossMark logo. The main title of the article is "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". Below the title is a list of authors: S. Bekal, C. Berry, A. R. Reimer, G. Van Domselaar, G. Beaudry, E. Fournier, F. Doualla-Bell, E. Levac, C. Gaulin, D. Ramsay, C. Huot, M. Walker, C. Sieffert, and C. Tremblay. A paragraph of text follows, mentioning various Canadian institutions. At the bottom of the page, a box contains the text: "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**". The word "SRP067504" is highlighted with a red box.

**SRP067504**

## Search by identifiers

The image shows the NCBI homepage. At the top, there is a search bar with the identifier "SRP067504" entered. This input field is highlighted with a red box. Below the search bar, there is a section titled "COVID-19 Information" with links to "Public health information (CDC)" and "Research information (NIH)". To the right of this is a "UNITE" initiative logo for ending structural racism, with a link to "nih.gov/ending-structural-racism" and a "LEARN MORE" button. On the left side of the page, there is a sidebar with a "Welcome to NCBI" message and a "Resource List (A-Z)" menu. The menu items include "All Resources", "Chemicals & Bioassays", "Data & Software", "DNA & RNA", "Domains & Structures", "Genes & Expression", "Genetics & Medicine", "Genomes & Maps", and "Homology". At the bottom right, there is a "Submit" button with the text "Deposit data or manuscripts into NCBI databases" and an upward-pointing arrow icon.

# Targeted repositories

[www.ncbi.nlm.nih.gov/pathogens](http://www.ncbi.nlm.nih.gov/pathogens)

 **National Library of Medicine**  
*National Center for Biotechnology Information*

[Health](#) > Pathogen Detection

## Pathogen Detection BETA

**i** To assist the National Database of Antibiotic Resistant Organisms (NDAR), Pathogen Detection identifies the antimicrobial resistance, stress response, and virulence 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
<a href="#">Salmonella enterica</a>	<a href="#">184</a>	<a href="#">419,163</a>
<a href="#">E.coli and Shigella</a>	<a href="#">10</a>	<a href="#">210,336</a>
<a href="#">Campylobacter jejuni</a>	<a href="#">97</a>	<a href="#">71,833</a>
<a href="#">Listeria monocytogenes</a>	<a href="#">6</a>	<a href="#">49,495</a>

[See more organisms...](#)

[Browser Factsheet](#)

[Antimicrobial Resistance Factsheet](#)

[Antimicrobial Resistance](#)

[Contributors](#)

[Help](#)

# Targeted repositories

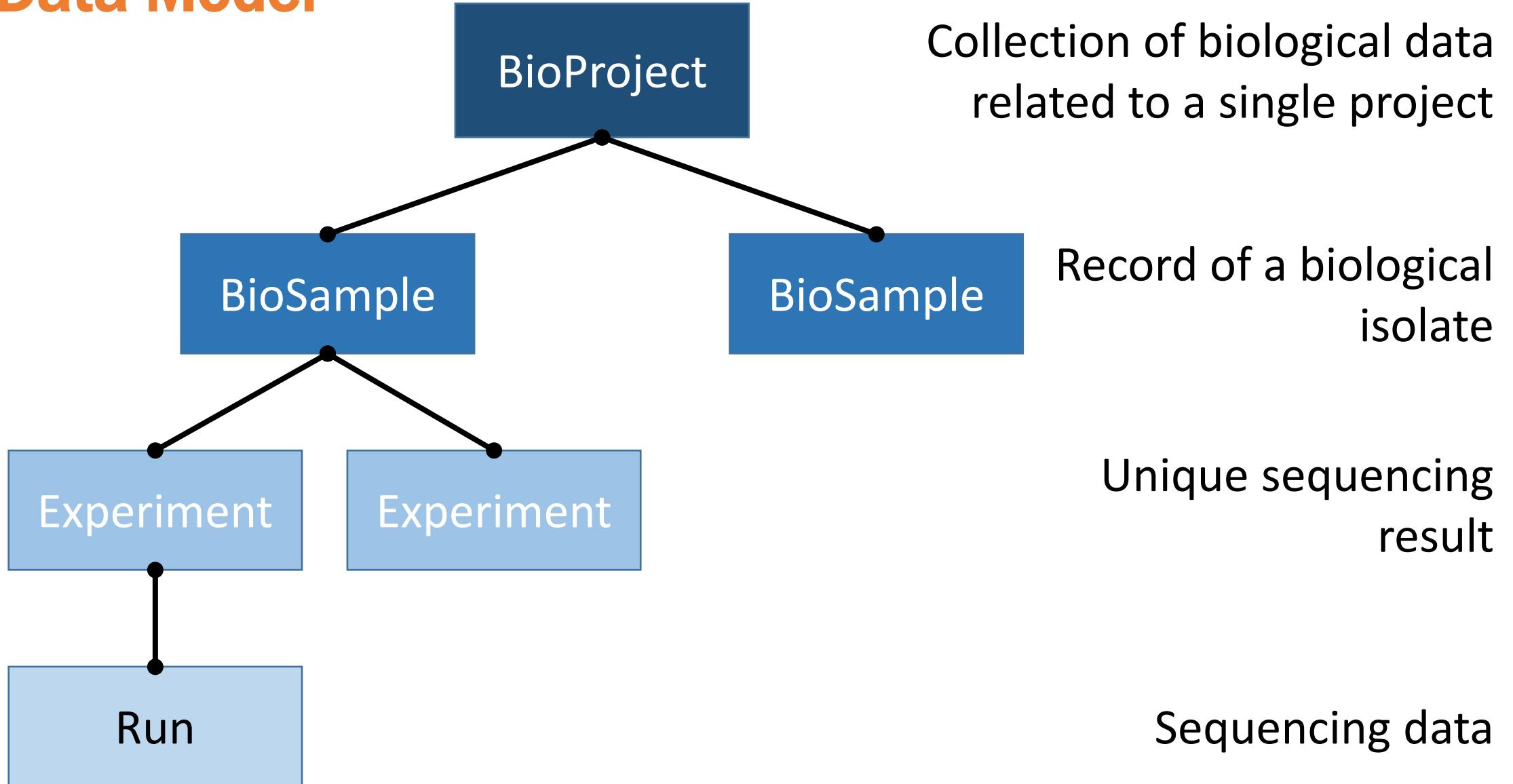
[www.ncbi.nlm.nih.gov/sars-cov-2/](http://www.ncbi.nlm.nih.gov/sars-cov-2/)



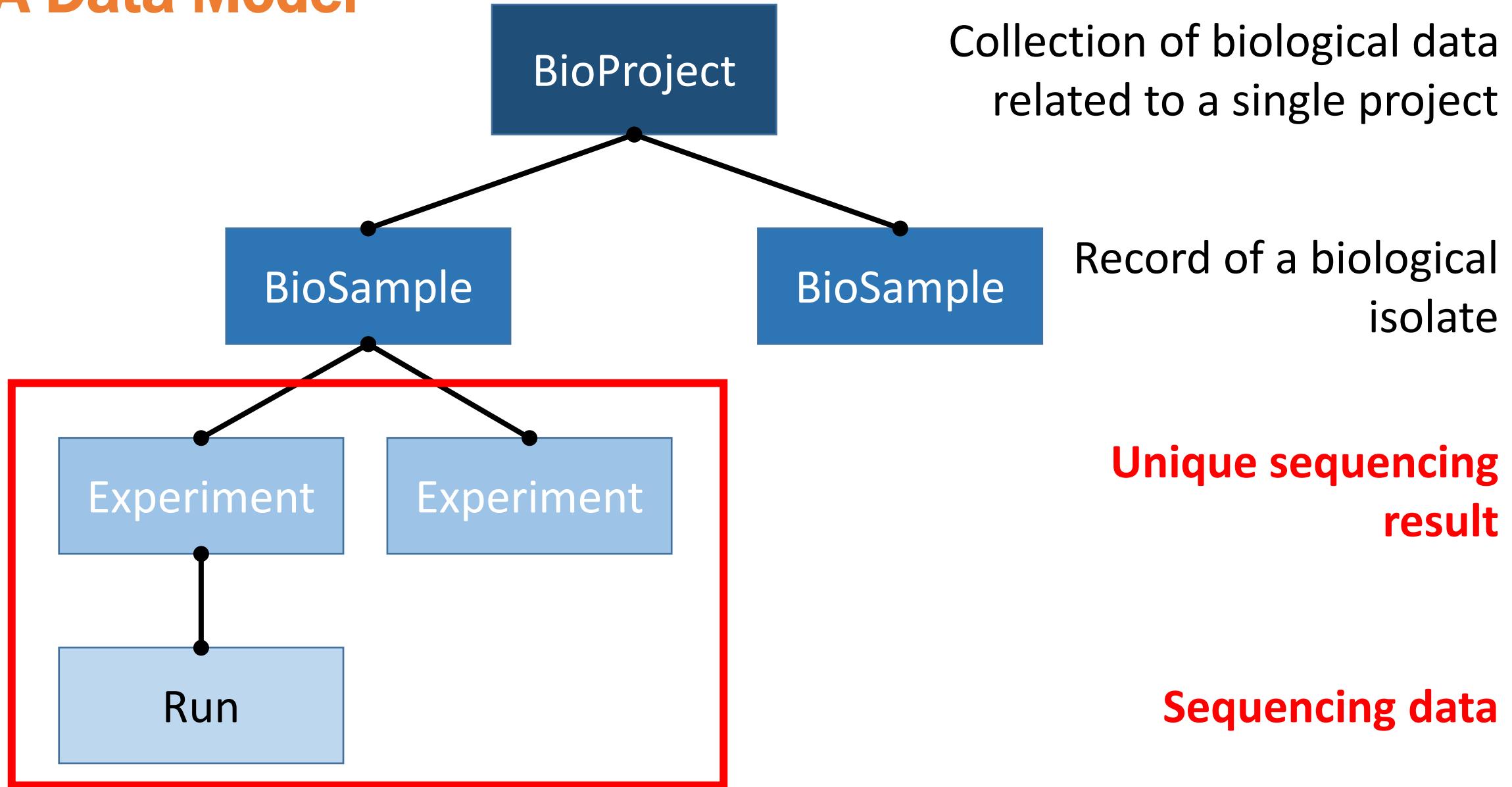
The image shows the homepage of the NCBI SARS-CoV-2 Resources. At the top left is the NIH/National Library of Medicine logo. A search bar with the placeholder "Search NCBI" and a blue "Search" button are on the right. The main title "NCBI SARS-CoV-2 Resources" is displayed prominently in white text against a dark background. Below the title is a large, detailed 3D rendering of a SARS-CoV-2 virus particle, showing its characteristic spike proteins and internal structure. To the left, a sidebar titled "Quick Navigation Guide" lists links to Sequence Submission, Literature, Sequence-Related Resources, Clinical Resources, and Other Websites. To the right, a section titled "SARS-CoV-2 Data" displays several large numbers representing different types of resources: 2,921,352 SRA runs, 3,516,090 Nucleotide records, 3,215 ClinicalTrials.gov entries, 220,402 PubMed articles, and 270,290 PMC articles.

SARS-CoV-2 Data	
<b>2,921,352</b>	<b>SRA runs</b>
<b>3,516,090</b>	<b>Nucleotide records</b>
<b>3,215</b>	<b>ClinicalTrials.gov</b>
<b>220,402</b>	<b>PubMed</b>
<b>270,290</b>	<b>PMC</b>

# SRA Data Model



# SRA Data Model



# SRA Experiment Record

[www.ncbi.nlm.nih.gov/sra/SRX1489277](http://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
<a href="#">SRR3028792</a>	824,262	354.1M	188.3Mb	2015-12-19

# SRA Experiment Record

## Library information (Illumina, paired-end)

[www.ncbi.nlm.nih.gov/sra/SRX1489277](http://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](#)

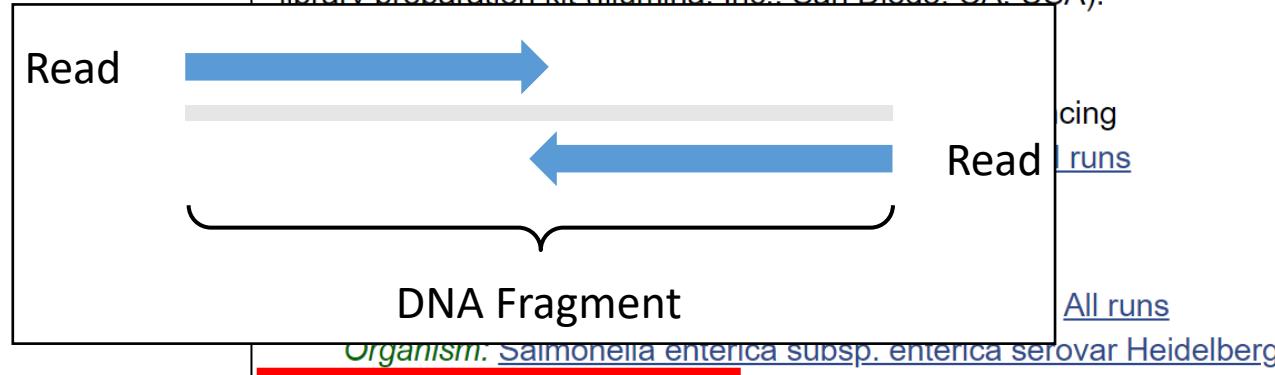
Run	# of Spots	# of Bases	Size	Published
<a href="#">SRR3028792</a>	824,262	354.1M	188.3Mb	2015-12-19

# SRA Experiment Record

## Library information (Illumina, paired-end)

[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)."



*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
<a href="#">SRR3028792</a>	824,262	354.1M	188.3Mb	2015-12-19

# SRA Experiment

## Record

Experiment

BioProject

BioSample

Run

[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
<a href="#">SRR3028792</a>	824,262	354.1M	188.3Mb	2015-12-19

# SRA Experiment

## Record

Experiment

BioProject

BioSample

Run

[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
<a href="#">SRR3028792</a>	824,262	354.1M	188.3Mb	2015-12-19

# SRA Run

[trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR3028792](http://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:

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

[Legend](#)

Experiment	Library Name	Platform	Strategy	Source	Selection	Layout	Action
<a href="#">SRX1489277</a>	SH08-001	Illumina	WGS	GENOMIC	RANDOM	PAIRED	<a href="#">BLAST</a>

[Show design](#)

Biosample	Sample Description	Organism	Links
<a href="#">SAMN04334683 (SRS1211375)</a>		<a href="#">Salmonella enterica</a> subsp. <i>enterica</i> serovar Heidelberg	<a href="#">PRJNA305824</a>

Bioproject	SRA Study	Title
<a href="#">PRJNA305824</a>	<a href="#">SRP067504</a>	Salmonella serovar Heidelberg genome sequencing

[Show abstract](#)

# SRA Run: Reads

[trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR3028792](http://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

**Reads (separated)**

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

>gnl|SRA|SRR3028792.1.1 1 (Biological)  
**CTTCATAATCAGGCGATAAATGCCACCACTTAGGCTTTCTGGCGCGGATAGCCTCCCC  
AATAAAATCTTACGCGTACGCTTGCTTC**

>gnl|SRA|SRR3028792.1.2 1 (Biological)  
**CTCTCAAACCTTCGTACTTTTTCTTCCGCCTCTCTCCTCCCCCCCCGCCTCCCC  
CTCTTTCCCTTCTCTCCTGCTCTGCCCTCTCTCTCTCCTCCTCTGCCTCTTCCCT  
CTCCCCCTACTCTCCCTCTTCCCTCCCCCTCTCCTCCCTCTCCCCCTCT  
CCCCCCCCTTCCCTCCCCCTCTCCCTCCCCCCCCCTCCCCCTCCCCCT  
CTCTCTCCCCCTCTCCCTCCCCCTCCCCCCCCCTCTCCTCCCCCCCCCTCCCCCT**

# SRA Run: Reads

[trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR3028792](http://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR3028792)

WGS of *Salmonella enterica* subsp.

Metadata Analysis **Reads** Data access

Filter:  Find Filter

What can the filter be applied to?

< 1 1 82427 >

1. SRR3028792.1 [SRS1211375](#)  
name: 1, member: 8

2. SRR3028792.2 [SRS1211375](#)  
name: 2, member: 8

3. SRR3028792.3 [SRS1211375](#)  
name: 3, member: 8

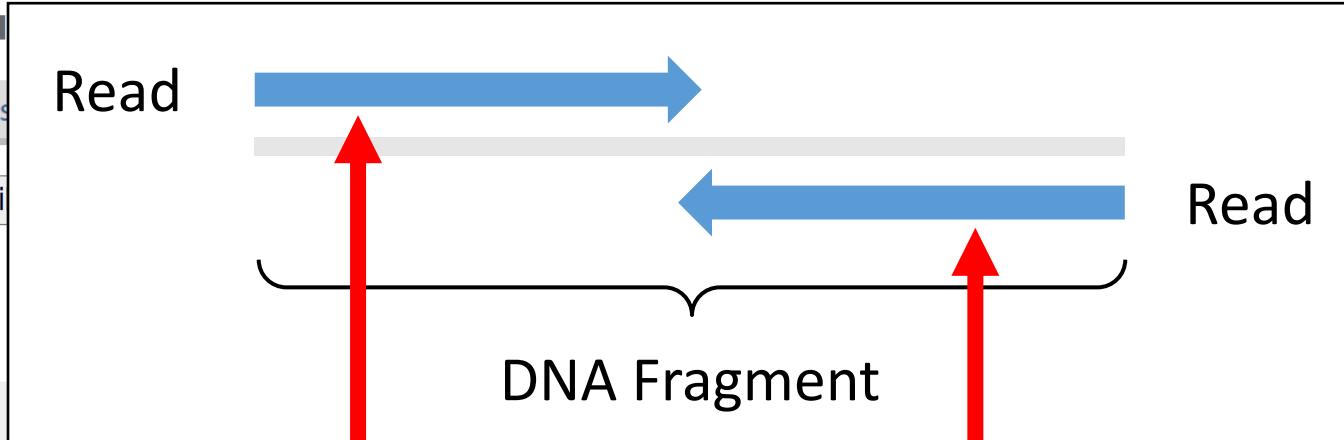
4. SRR3028792.4 [SRS1211375](#)  
name: 4, member: 8

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



Read

Read

DNA Fragment

Reads (separate)

>gnl|SRA|SRR3028792.1.1 1 (Biological)  
CTTCATAATCAGGCGATAAATGCCAACCACTTAGGCTTTCTGC CGCGGATAGCCTCCCC  
AATAAAATCTTACGCGTACGCTTTGCTTC

>gnl|SRA|SRR3028792.1.2 1 (Biological)  
CTCTCAAACCTTCCCGTTACTTTTTCTTCCGCTCTCTCCTCCCCCCCCGCCTCCCC  
CTCTTTCCCTTCTCTCCTGCTCTGCCCTCTCTCTCTCCTCCTCTGCCTCTTCCCT  
CTCCCCCTACTTCTCCCTCTTCCCTCCCCCTCTCCTTCCCTCTCCCCCTCT  
CCCCCCCCTTCCCTCCCCCTCTCCCTCCCCCCCCCTCCCCCTCCCCCTCCCC  
CTCTCTCCCCCTCTCTCCCTCCCCCTCTCCCCCCCCCTCTCCTCCCCCCCCCTCCCC

# 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-crunch-2/SRR3028792/SRR3028792.1
		AWS	<a href="https://sra-pub-run-odp.s3.amazonaws.com/sra/SRR3028792.1">https://sra-pub-run-odp.s3.amazonaws.com/sra/SRR3028792.1</a>
		NCBI	<a href="https://sra-downloaddb.be-md.ncbi.nlm.nih.gov/sos1/sra/SRR3028792.1">https://sra-downloaddb.be-md.ncbi.nlm.nih.gov/sos1/sra/SRR3028792.1</a>

- 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")

- 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")

- **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](https://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* serovar 1 ILLUMINA (Illumina MiSeq) run: 824,262 spots, 354.1M bases, 188.3

**Design:** "Genomic DNA was extracted using the Epicentre Metagenomic 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 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-

# Prefetch

[github.com/ncbi/sra-tools/wiki/08.-prefetch-and-fasterq-dump](https://github.com/ncbi/sra-tools/wiki/08.-prefetch-and-fasterq-dump)

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

```
$ prefetch PRJNA305824
```

## Salmonella enterica subsp. enterica

### Salmonella serovar Heidelberg genome sequencing

This study was conducted to evaluate whole-genome sequencing versus conventional pulsed-field gel electrophoresis subtyping for S.

Accession	PRJNA305824
Data Type	Genome sequencing
Scope	Multiisolate

#### Project Data:

Resource Name	Number of Links
<b>SEQUENCE DATA</b>	
Nucleotide (Genomic DNA)	19
SRA Experiments	65
Protein Sequences	26972
<b>PUBLICATIONS</b>	
PubMed	1
PMC	1
<b>OTHER DATASETS</b>	
BioSample	59
Assembly	6

#### ▼ SRA Data Details

Parameter	Value
Data volume, Gbases	43
Data volume, Mbytes	25794

[www.ncbi.nlm.nih.gov/bioproject/PRJNA305824](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA305824)

# Prefetch

[github.com/ncbi/sra-tools/wiki/08.-prefetch-and-fasterq-dump](https://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).*

**Salmonella enterica subsp. enterica**

**Salmonella serovar Heidelberg genome sequencing**

This study was conducted to evaluate whole-genome sequencing versus conventional pulsed-field gel electrophoresis subtyping for S.

Accession	PRJNA305824
Data Type	Genome sequencing
Scope	Multiisolate

**Project Data:**

Resource Name	Number of Links
<b>SEQUENCE DATA</b>	
Nucleotide (Genomic DNA)	19
SRA Experiments	65
Protein Sequences	26972
<b>PUBLICATIONS</b>	
PubMed	1
PMC	1
<b>OTHER DATASETS</b>	
BioSample	59
Assembly	6

▼ SRA Data Details

Parameter	Value
Data volume, Gbases	43
Data volume, Mbytes	25794

[www.ncbi.nlm.nih.gov/bioproject/PRJNA305824](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA305824)

# Fasterq-dump

[github.com/ncbi/sra-tools/wiki/08.-prefetch-and-fasterq-dump](https://github.com/ncbi/sra-tools/wiki/08.-prefetch-and-fasterq-dump)

- Convert to FASTQ with “fasterq-dump”

```
$ fasterq-dump SRR3028792
```

 SRR3028792_1.fastq	405,347 KB	FASTQ File
 SRR3028792_2.fastq	407,280 KB	FASTQ File

- You can compress the files with “gzip”

```
$ gzip *.fastq
```

 SRR3028792_1.fastq.gz	111,641 KB	GZ File
 SRR3028792_2.fastq.gz	140,381 KB	GZ File

# FASTQ

[en.wikipedia.org/wiki/FASTQ\\_format](https://en.wikipedia.org/wiki/FASTQ_format)

**File:** *reads.fastq*

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

Read



Single-end sequencing

# FASTQ

[en.wikipedia.org/wiki/FASTQ\\_format](https://en.wikipedia.org/wiki/FASTQ_format)

File: *reads.fastq*

Read ID  
@SRR3028792.1  
CTTCATAATCAGGCGATAAATGCCACCA  
+SRR3028792.1  
AABC-C--CE,-,C++@,,,,<,CCC,BB,

Read 

Single-end sequencing

# FASTQ

[en.wikipedia.org/wiki/FASTQ\\_format](https://en.wikipedia.org/wiki/FASTQ_format)

File: *reads.fastq*

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

```

Read 

Single-end sequencing

# FASTQ

[en.wikipedia.org/wiki/FASTQ\\_format](https://en.wikipedia.org/wiki/FASTQ_format)

File: *reads.fastq*

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

Separator (read ID again)

Read 

Single-end sequencing

# FASTQ

Encoded quality of each nucleotide

File: *reads.fastq*

@SRR3028792.1

CTTCATAATCAGGCGATAAATGCCACCA

+SRR3028792.1

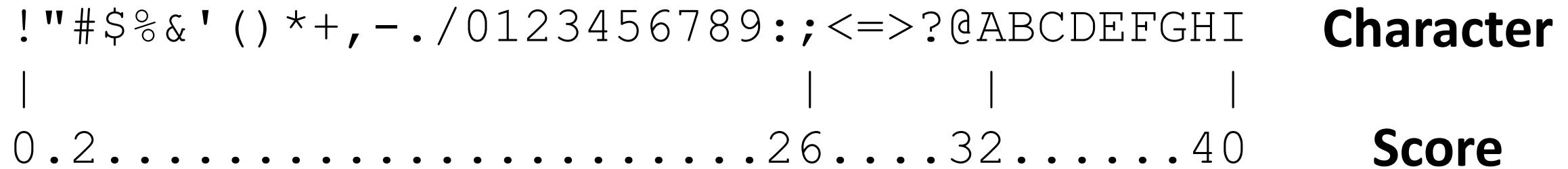
AABC-C--CE,-,C++@,,,,<,CCC,BB,

Read



Single-end sequencing

# Quality score encoding



+SRR3028792.1

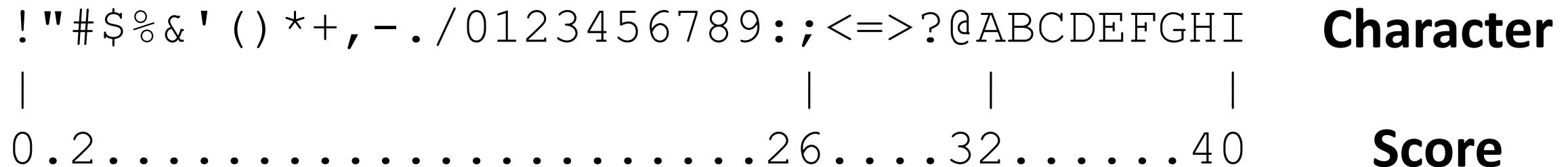
AABC-C--CE,-,C++@,,,<,CCC,BB,

Read



Single-end sequencing

# Quality score encoding



+SRR3028792.1

AABC-C--CE,-,C++@,,,,<,CCC,BB,

Read



Single-end sequencing

Phred scores

[en.wikipedia.org/wiki/Phred\\_quality\\_score](https://en.wikipedia.org/wiki/Phred_quality_score)

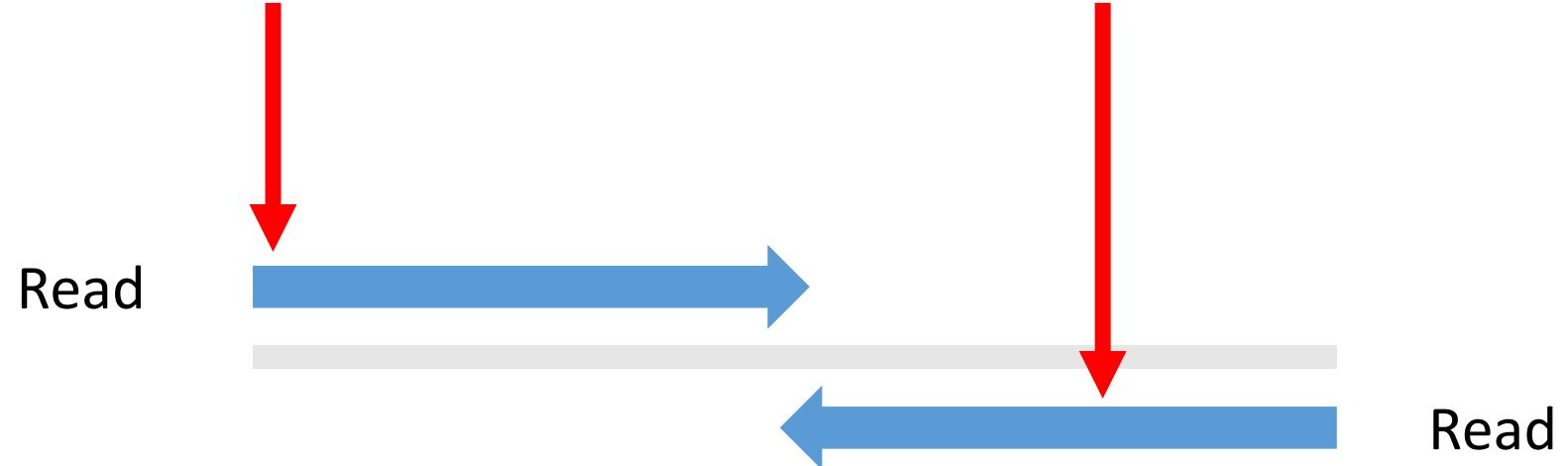
Score 10:  $P(\text{error}) = 10\%$   
Score 20:  $P(\text{error}) = 1\%$   
Score 30:  $P(\text{error}) = 0.1\%$

# FASTQ

[en.wikipedia.org/wiki/FASTQ\\_format](https://en.wikipedia.org/wiki/FASTQ_format)

## File: *reads\_1.fasta*

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



## File: *reads\_2.fasta*

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

Paired-end sequencing

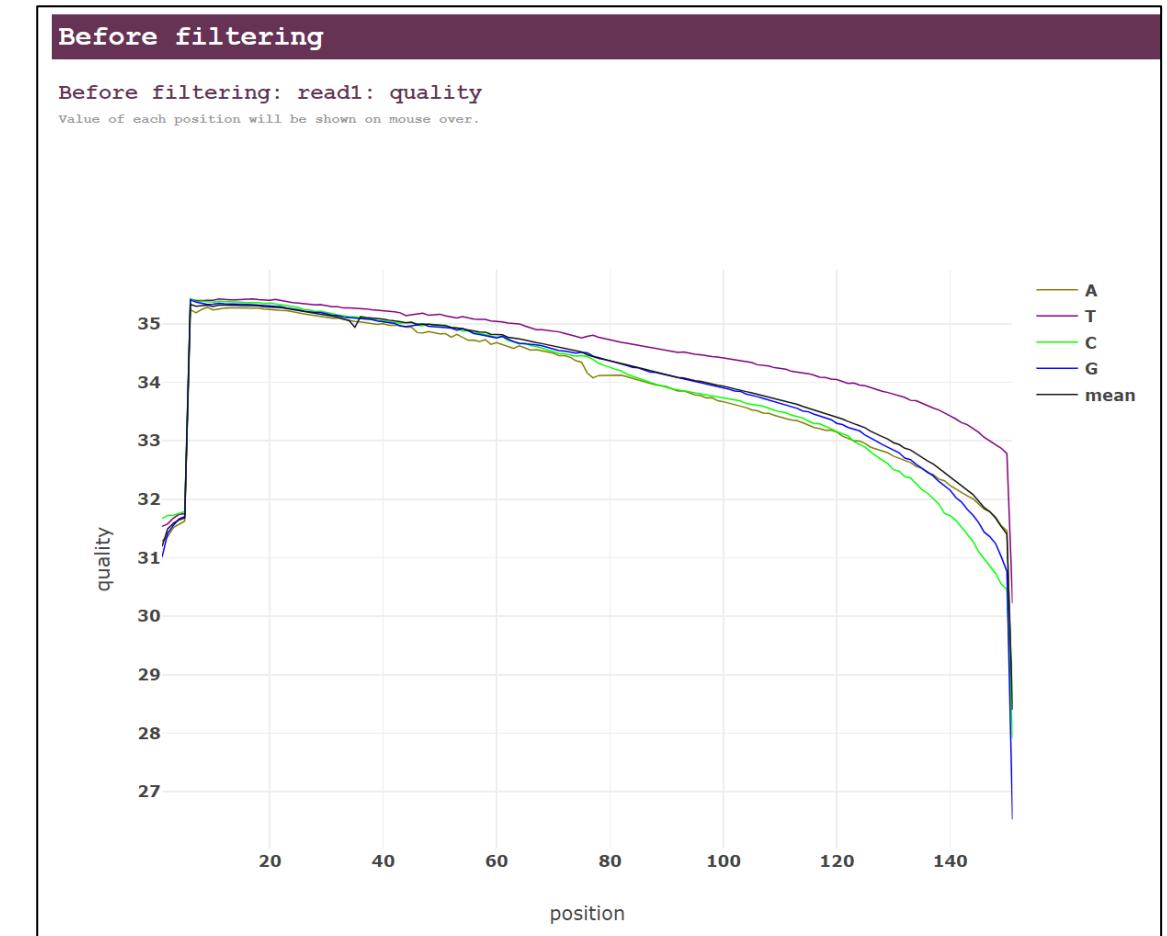
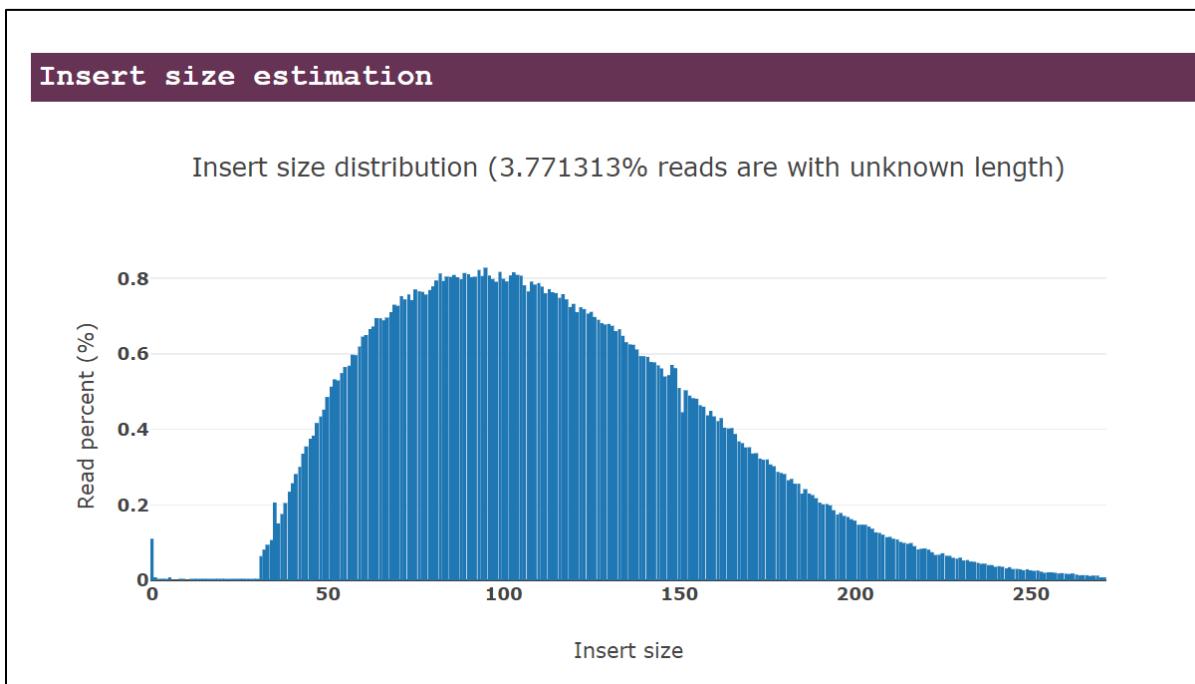
# LEARNING OBJECTIVES

- 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](https://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](https://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](https://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](https://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](https://github.com/OpenGene/fastp)

## fastp report

### Summary

#### General

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

### Filtering result

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

# Fastp quality report

[github.com/OpenGene/fastp](https://github.com/OpenGene/fastp)

## fastp report

### Summary

#### General

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

### Filtering result

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

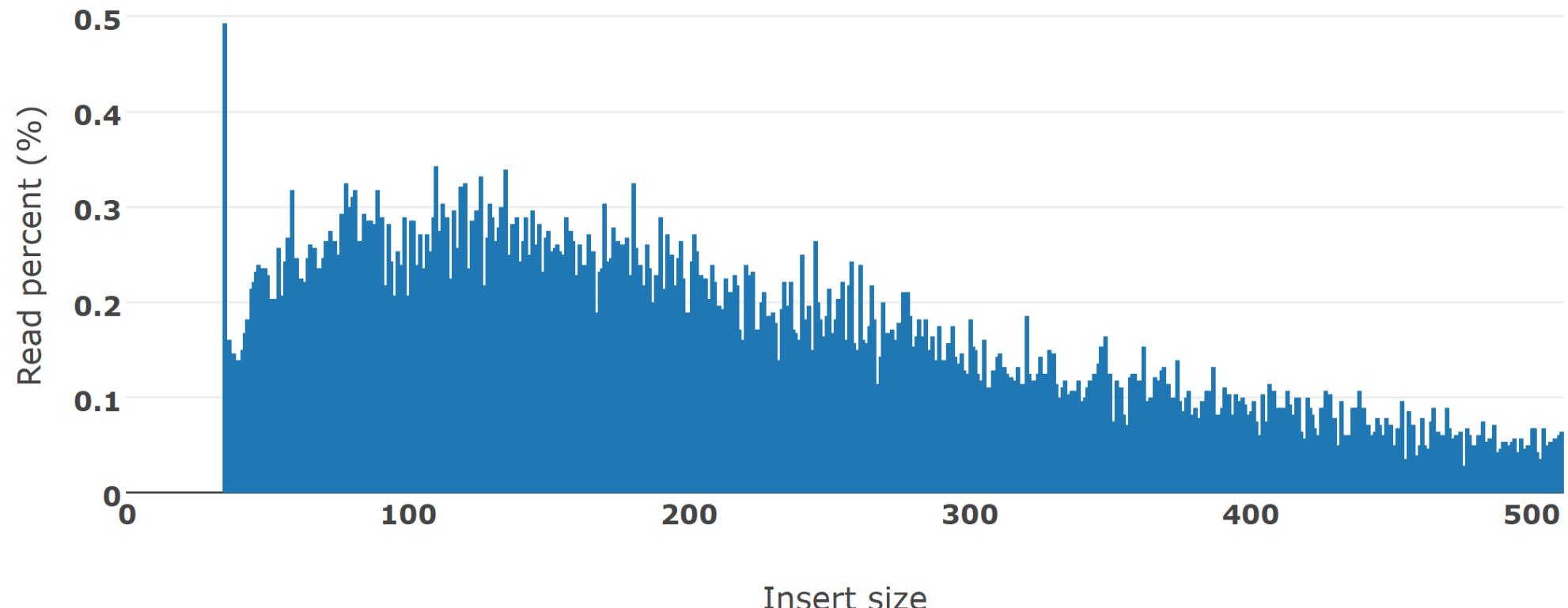
96% passed filters  
4% removed

# Fastp quality report

[github.com/OpenGene/fastp](https://github.com/OpenGene/fastp)

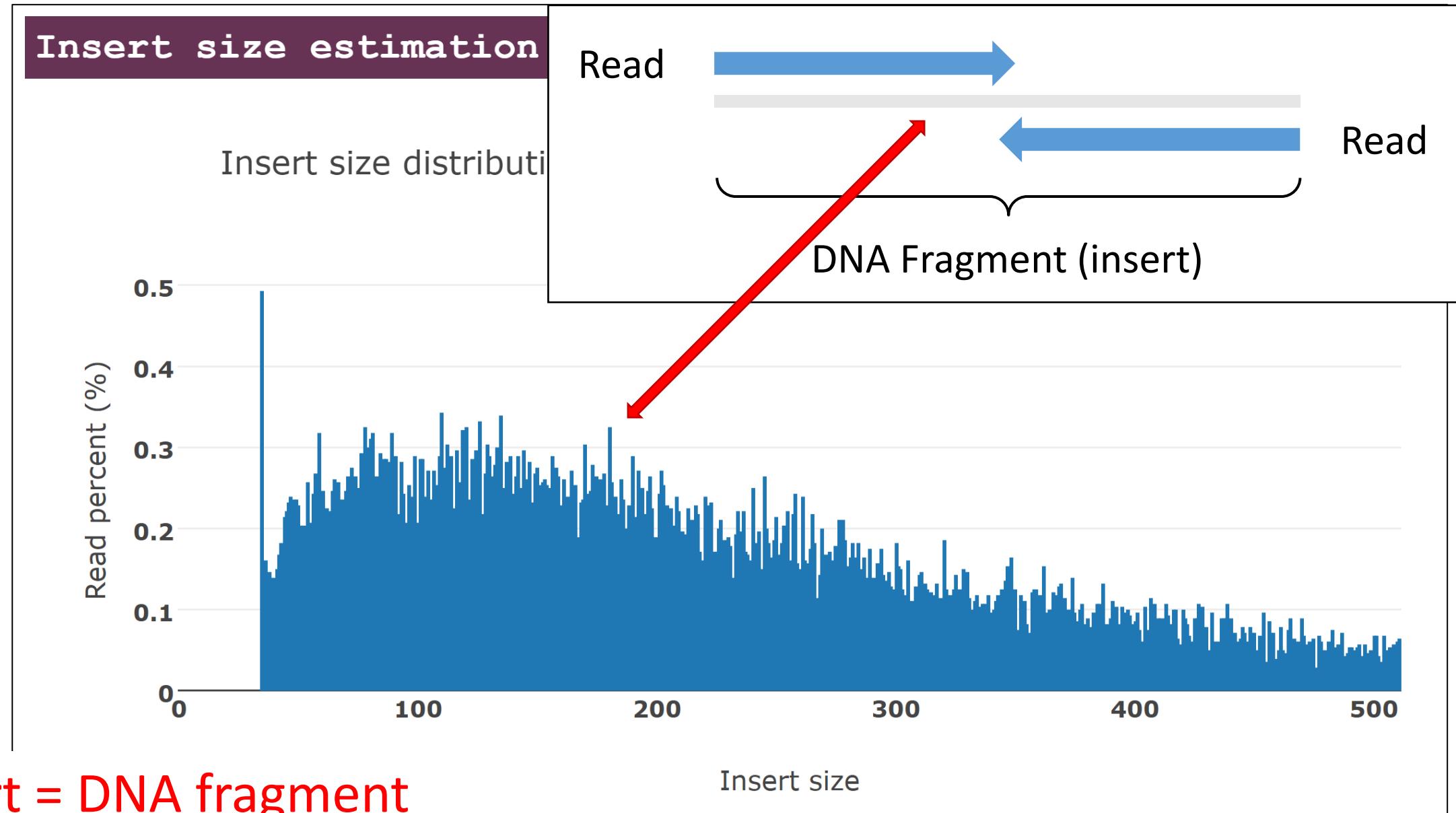
## Insert size estimation

Insert size distribution (20.803571% reads are with unknown length)



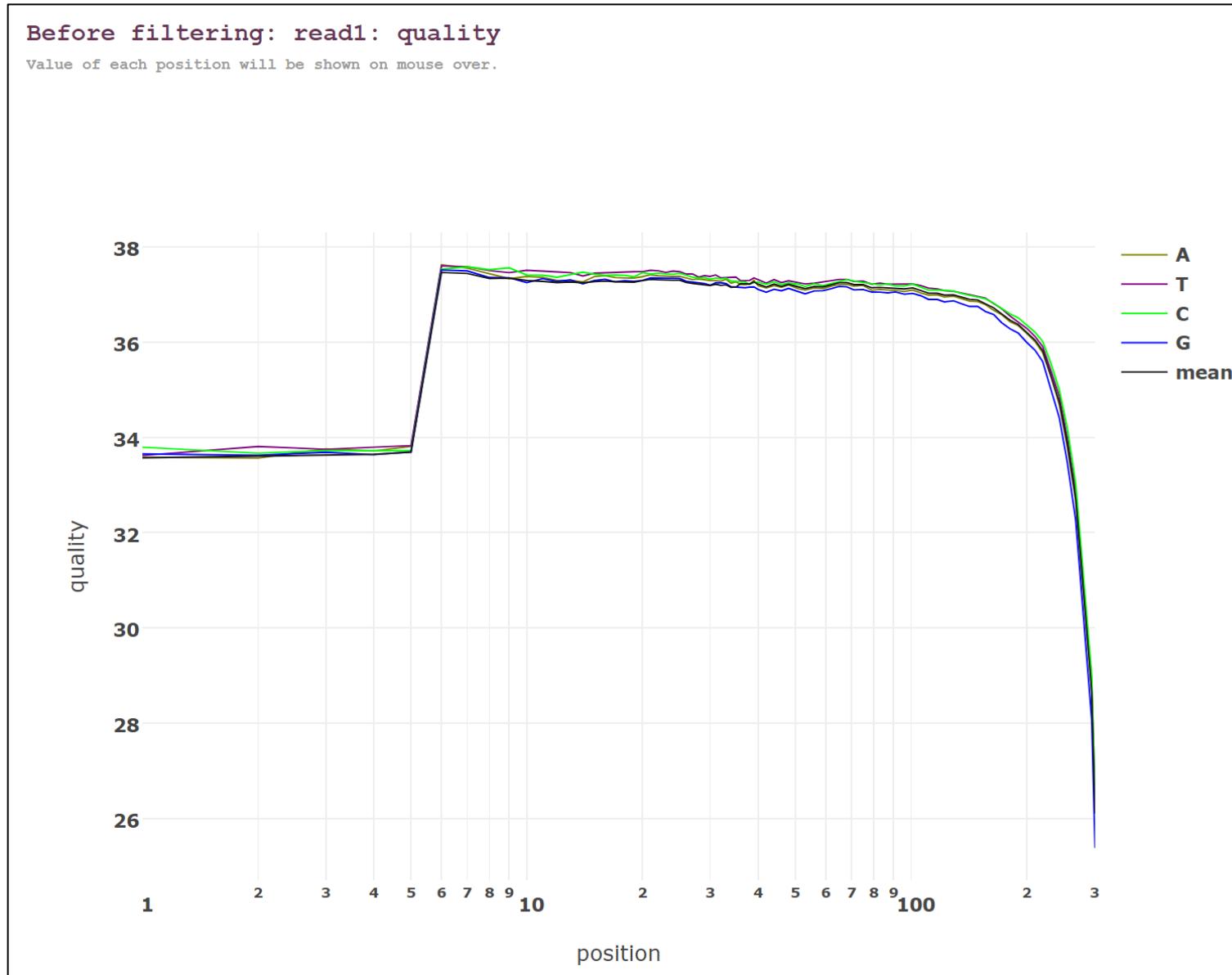
# Fastp quality report

[github.com/OpenGene/fastp](https://github.com/OpenGene/fastp)



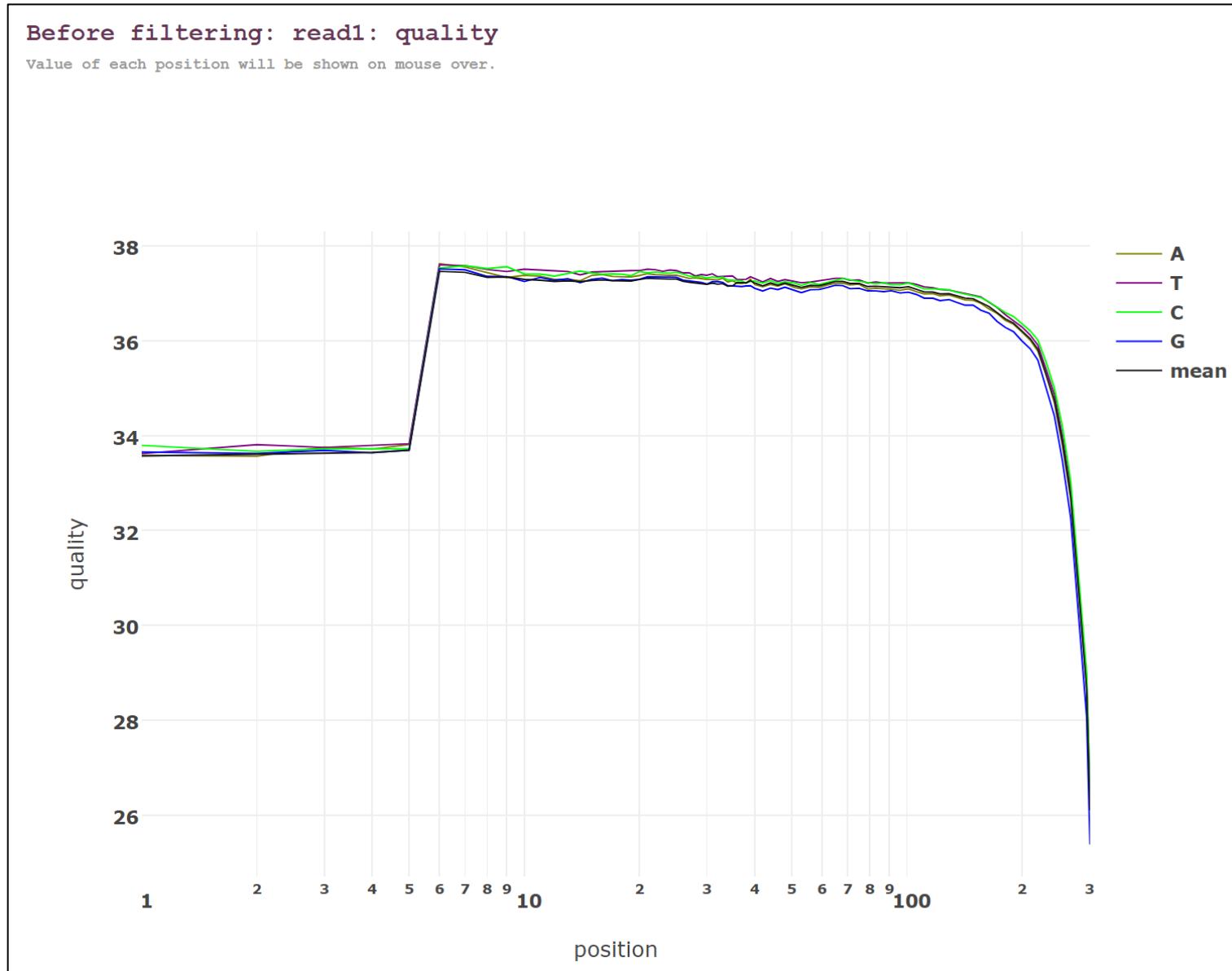
# Fastp quality report

[github.com/OpenGene/fastp](https://github.com/OpenGene/fastp)



# Fastp quality report

[github.com/OpenGene/fastp](https://github.com/OpenGene/fastp)



# 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*

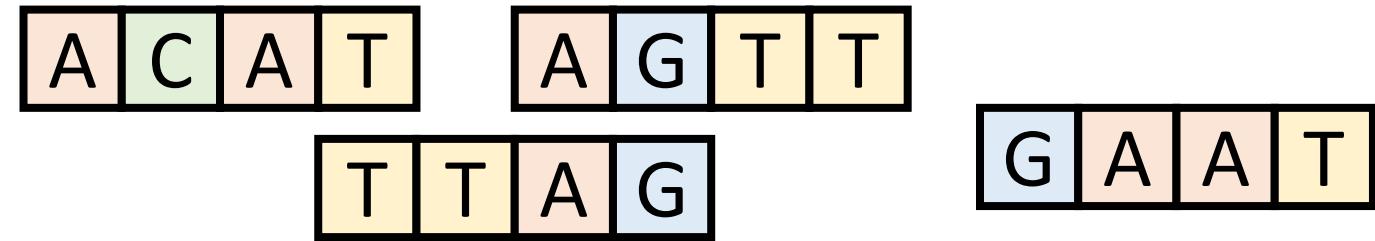
# Genome assembly



[https://en.wikipedia.org/wiki/Jigsaw\\_puzzle#/media/File:Sky\\_puzzle.jpg](https://en.wikipedia.org/wiki/Jigsaw_puzzle#/media/File:Sky_puzzle.jpg)

# Genome assembly

Reads



Transform reads to  
larger contiguous fragments  
("contigs")

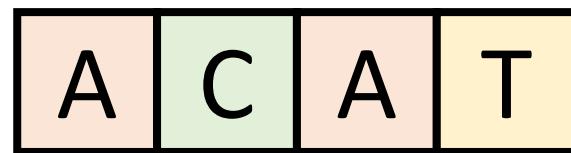


Assembly



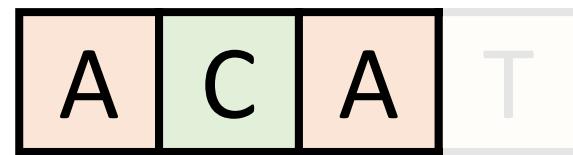
# Genome assembly

Read

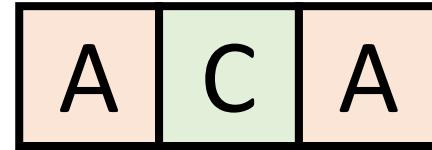


# Genome assembly

Read

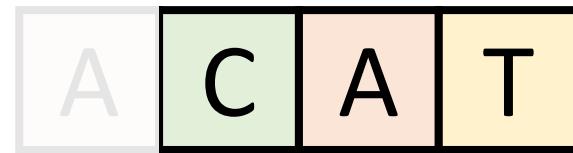


*k*-mer  
(*k*=3)

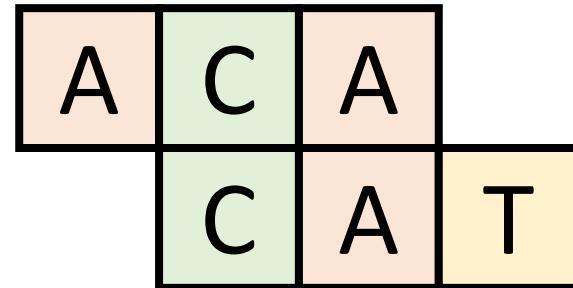


# Genome assembly

Read



$k$ -mer  
 $(k=3)$

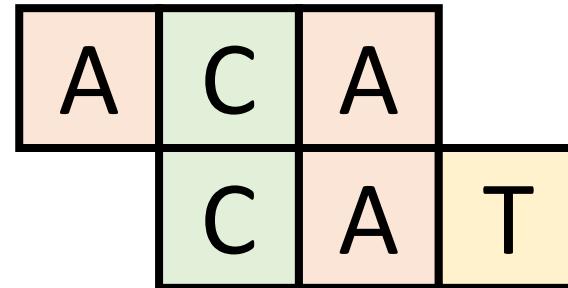


# Genome assembly

Read



$k$ -mer  
( $k=3$ )

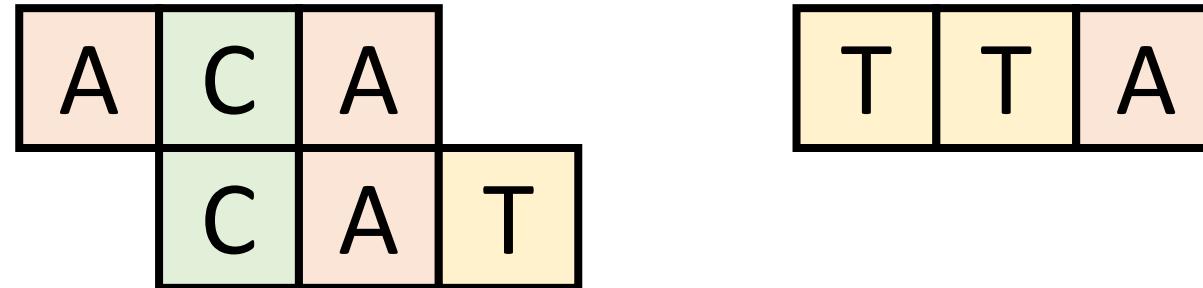


# Genome assembly

Read

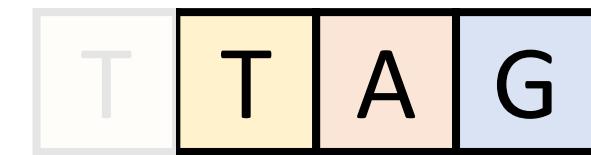
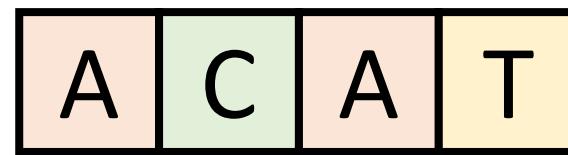


$k$ -mer  
 $(k=3)$

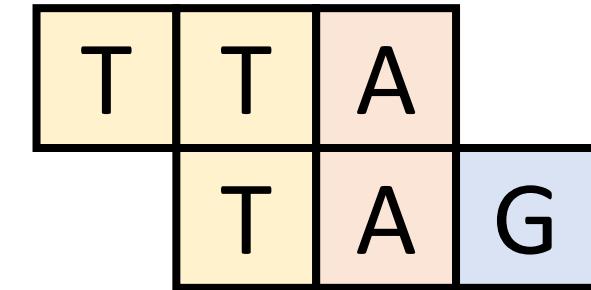
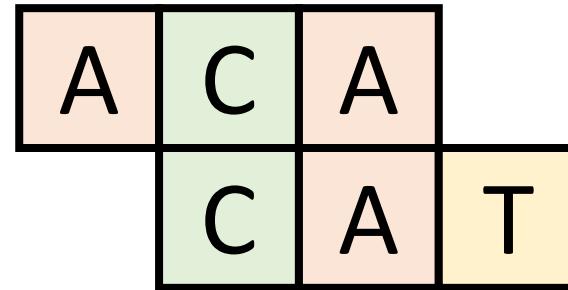


# Genome assembly

Read

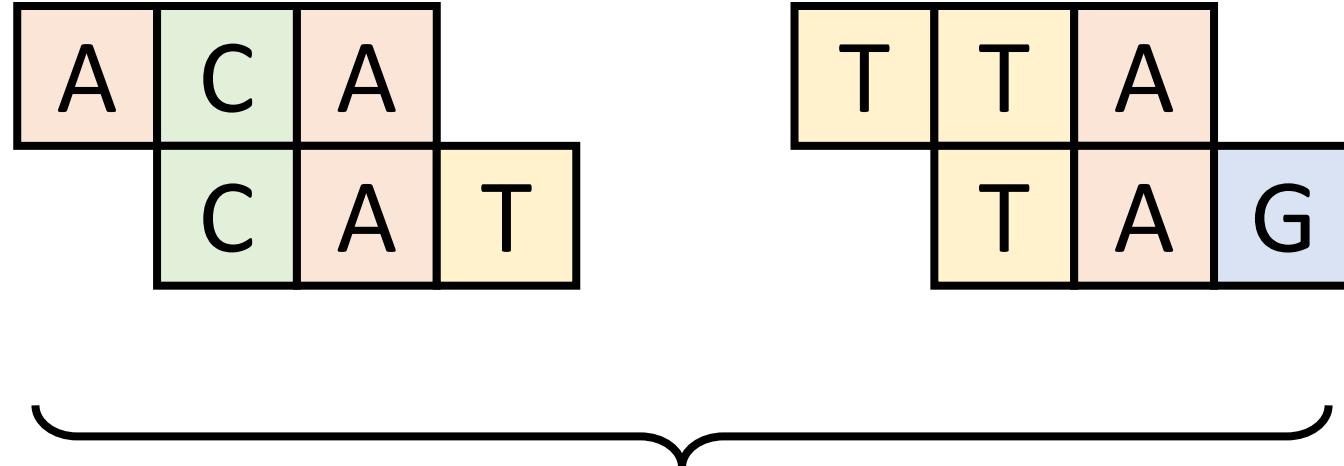


$k$ -mer  
 $(k=3)$



# Genome assembly

$k$ -mer  
( $k=3$ )

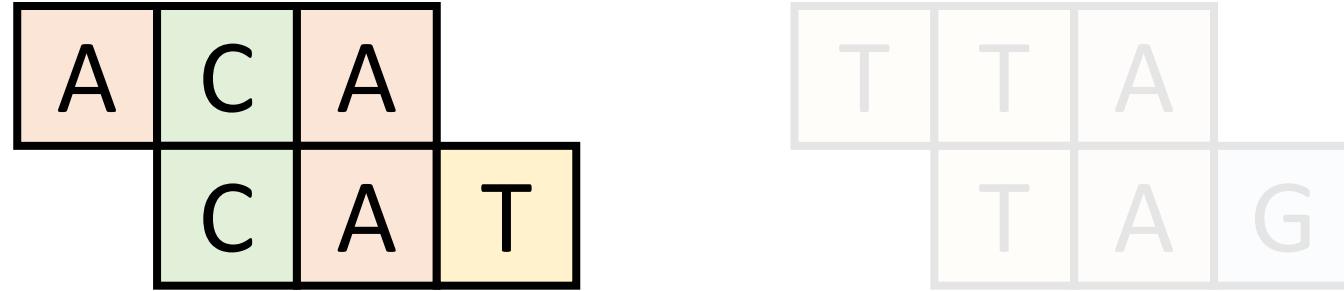


*De bruijn*  
graph

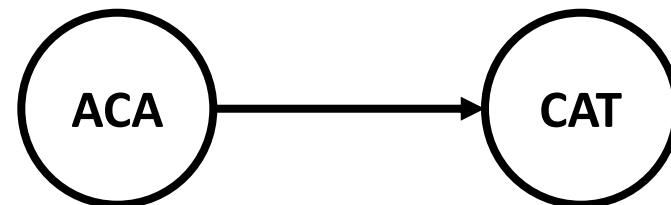
Construct a graph from  $k$ -mers

# Genome assembly

$k$ -mer  
( $k=3$ )



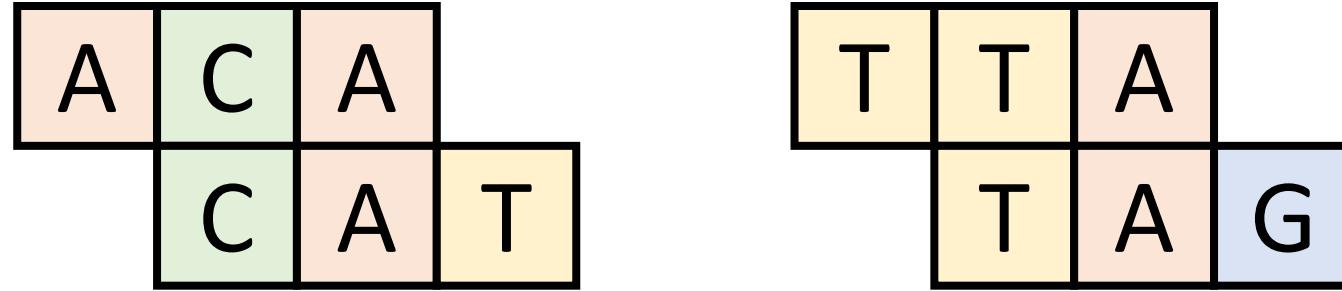
*De bruijn*  
graph



Construct a graph from k-mers

# Genome assembly

$k$ -mer  
( $k=3$ )



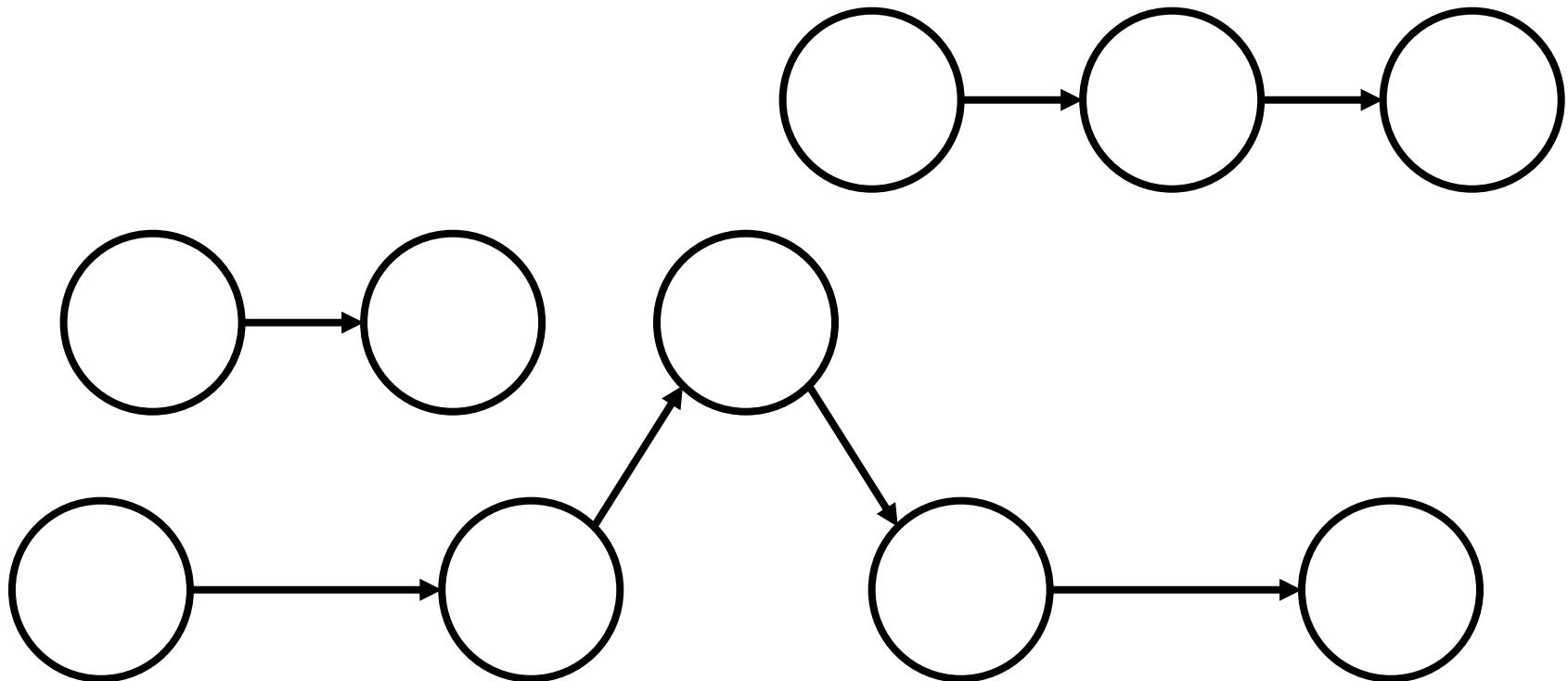
*De bruijn*  
graph



Construct a graph from  $k$ -mers

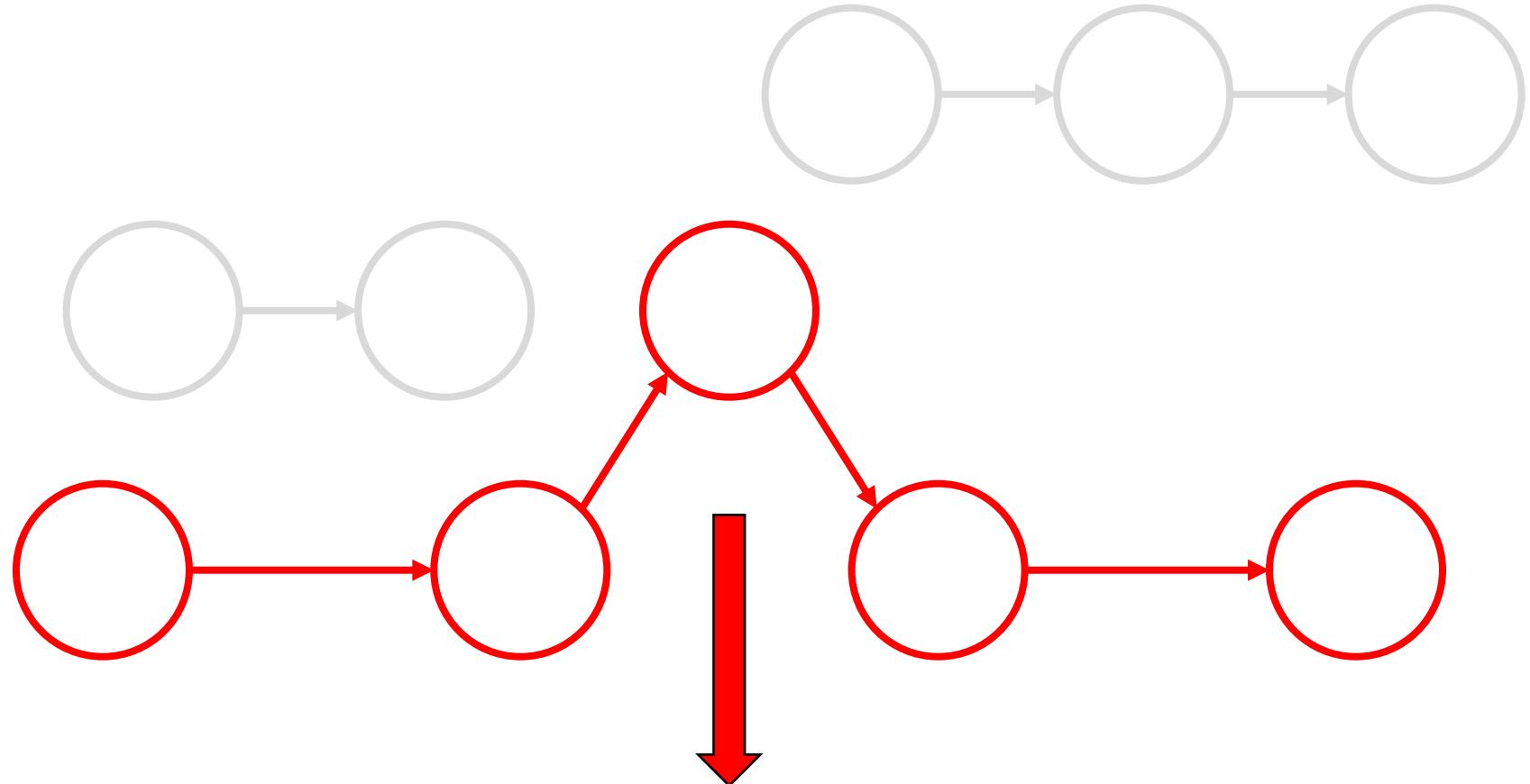
# Genome assembly

Completed  
graph



# Genome assembly

Completed  
graph



Unambiguous paths used to  
construct contigs

A C A T T A G T

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<sup>1</sup>, Richa Agarwala<sup>1\*</sup> and David J. Lipman<sup>1,2</sup>

### 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

- **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
```

- **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)

- **Install with conda**

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

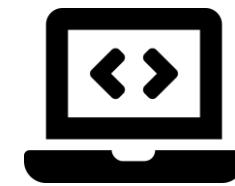
- **Assemble genome**

```
$ 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](https://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](https://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](https://github.com/ncbi/SKESA)

## Kmer size (k=21)

```
Kmer: 21 Graph size: 4403385
```

```
Valley: 0
```

Graph size (4.4 million kmers)

```
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](https://github.com/ncbi/SKESA)

## Kmer size (k=21)

Kmer: 21	Graph size: 4403385
Valley: 0	

### Graph size (4.4 million kmers)

```
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
```

### Contigs produced (~7,500)

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](https://github.com/ncbi/SKESA)

## Kmer size (k=21)

Kmer: 21	Graph size: 4403385
Valley: 0	

### Graph size (4.4 million kmers)

```
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
```

### Contigs produced (~7,500)

```
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](https://github.com/ncbi/SKESA)

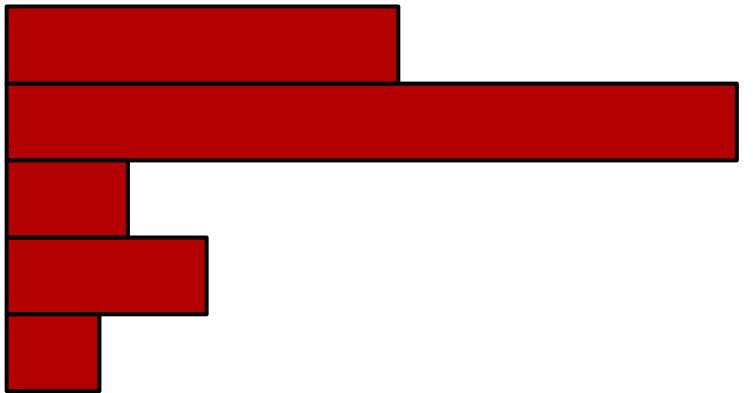
```
>Contig_3_5.5619
AAAAAAACCGCCCATGCCAGATTCTCTTCTACCGTCATCCGCAGAGAACGCGCCGCCCTCCGGCACTATGCC
CTTACCGTCAAACACCACCCGTCCGCTGGAGGCGCGGGATGCCGCACAAGGTGCCAAGCAGCGTCGTTTCC
GATGTGCAGACTGACGTGTCAGCGCCTGAATCTGCCGTAGTGGCGCTGACTTTCAAACGTTAACATCGT
CGGGTTGCGGATCTCTCCGGCGTGCCAGCGCGTGCCCTGGTTACCACGTAAATACGGTCGGA^
>Contig_4_8.89207
AAAAAAACTGGTACGTCGGCATTGGCGCTTTCCGCCGACGAGCACATACAGTCTGATGAGATTGTTCGAGAC
GACTCGGCAGTGCAGGCCACGCACCTGGTGAGCGGGATTAGCGTAAGGTTAGTCGGAGCGCAGTCAACATGCC
AGCAACGGAAAATGAGTGTGCGCGTTGAAGTCTGAAAGGCGTCTGTTCCATCATCAGATTGAGCGCGGCAATC
CGGTGCTACCAGCGCTGTTAGGGCTGTCTGGCTATTCGCGTAGCCGAGCGCTATTCAACGCATCCAGGAC
CGTCCAGAAATGGGACGGGCATTTAGCGCTTCTTACTTCGAGGAATCGCTTCAGCAATTGGTCAGCAGAC
AGAGTGAGGTCCGGTAATGGTAGGCCAATAGAAACCATGTCCATATCCGGATAGGGTTCTAAACAGACCGCA
AACAGACGCTGATAGGTTCCCGCACCGGTGCATGACCGGGAGTTGCATCGGCTGCCAGCCAGGATAGCTC
CCAGCATACTCACCACATAATCTTACCGCTGTCGATAAGAGAGCGAATCAGGCAGTGAATTGACATTGCA
GCCTTCGCCACGTCTGAATTGCGGATCACGCCGTTGGCGTTGCCTAGCAGGCGACAAAGGTATCGCTGA
TGCAGCGTCAGGTTTCTCTTCCGCCAGTCGTTTCAAGAATATCCTGGTAAGCGTTACTAATGTTTC
CTTCGCGCGGAATCGCGTTACGCAGC
>Contig_5_4.58495
AAAAAAAGCGAACATTTACGTAACGCCGCTCGAGCGACAAGAAGGGCGGAAAAAACGCCGGTGGTCGCAC
ACCCGCAGCCAGCATTGATCATCGGTAAAGTAACGCTTTTCACTAGACTCCTCATTAGCATAGTGCCTGC
TCTTGCCTTAAATTCTATCGCTACCAGGAACGTTTGAAGTGCAGGCGACATCCAGGTTAAAGCCTCACC
ACCTCTTGCCTTGCCTACCAAGCACCATCAGGCCGGTGTCCCAGCCGTCCGCTCCAACGCCGTTGGCGC^
```

# LEARNING OBJECTIVES

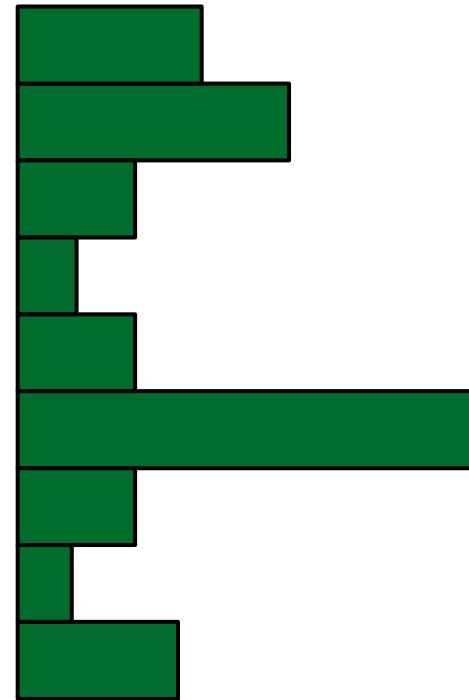
1. *What is microbial sequencing*
2. *What is genome assembly*
3. *Where to find sequence data*
4. *How to download*
5. *Quality of reads*
6. *How to assemble data*
7. ***How to evaluate quality of the assembly***
8. *Data analysis tutorial*

# Scenario: Two assembled genomes (contigs)

A



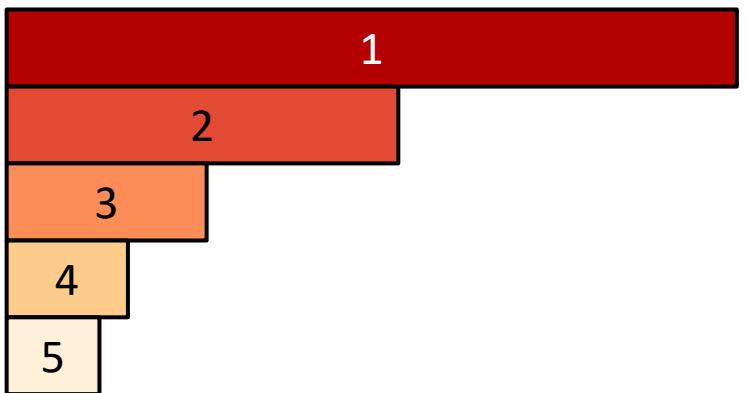
B



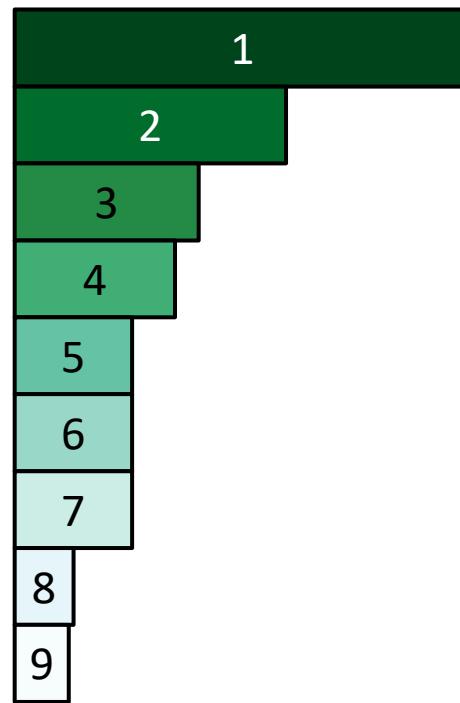
Which is more “complete”  
or “contiguous”?

# Order and count contigs

A



B



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



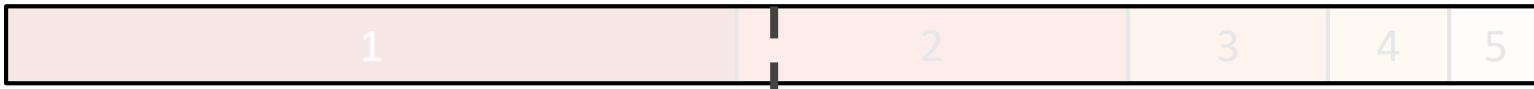
A



B

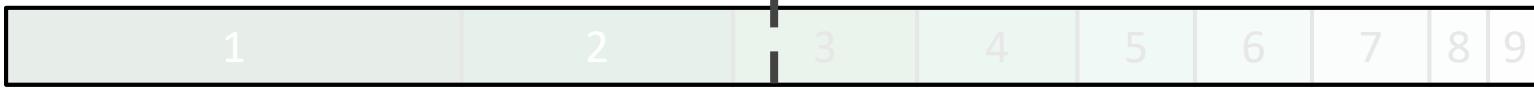


A

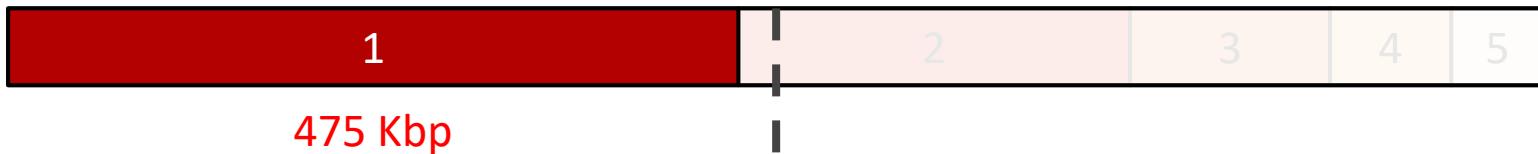


50%

B

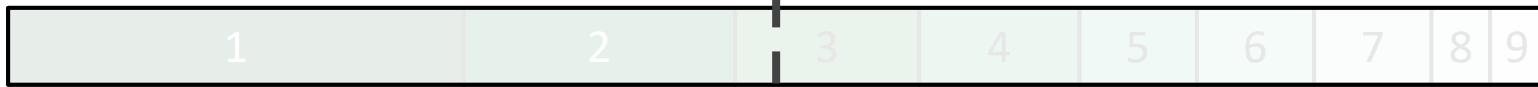


A



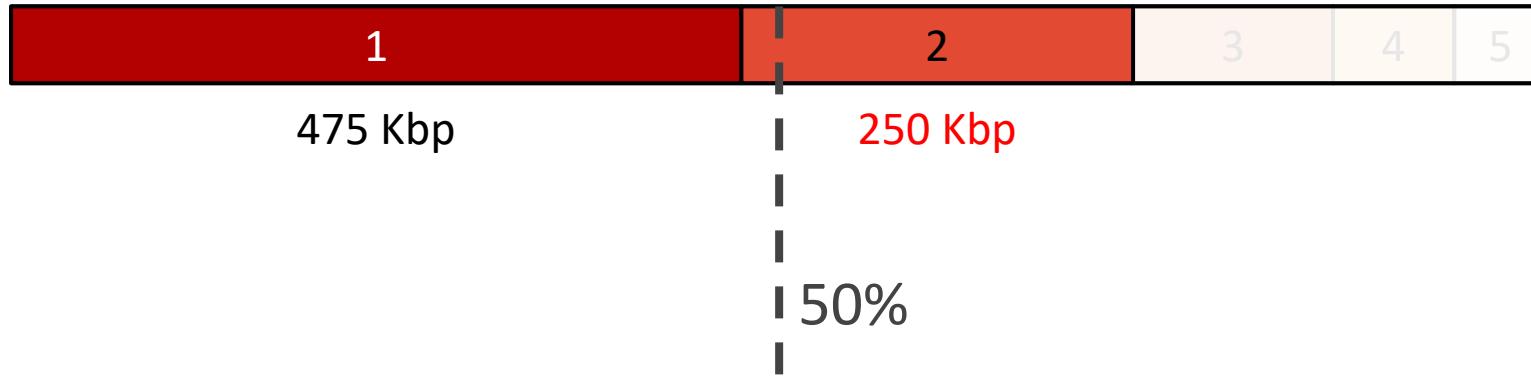
50%

B



A

**N50 = 250 Kbp**

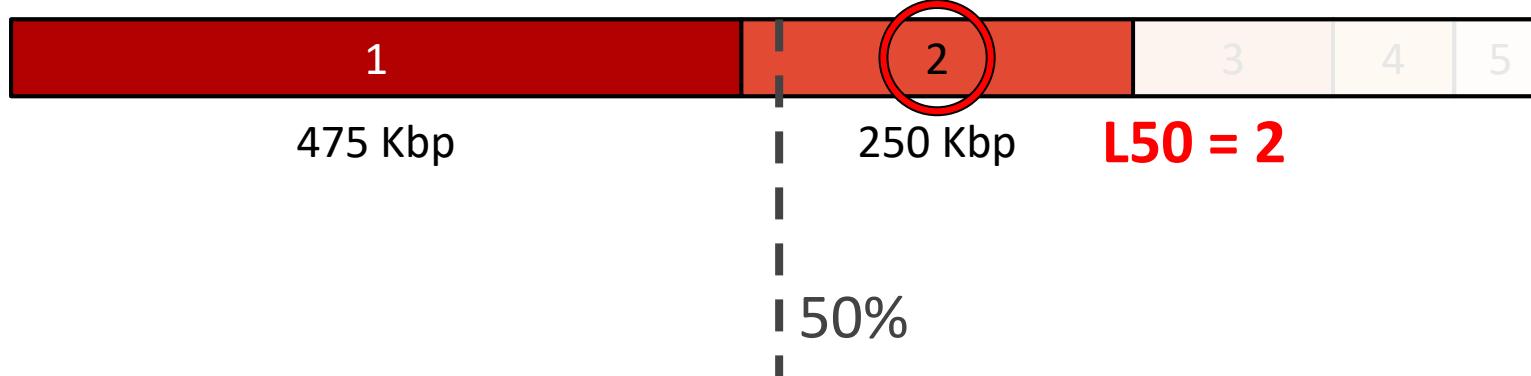


B

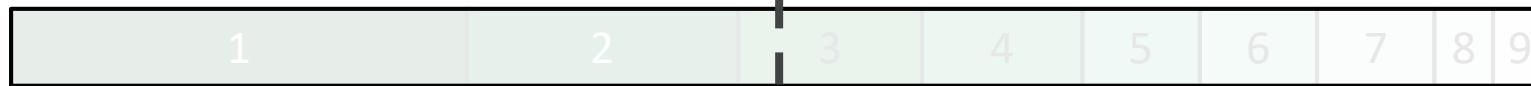


A

**N50 = 250 Kbp**

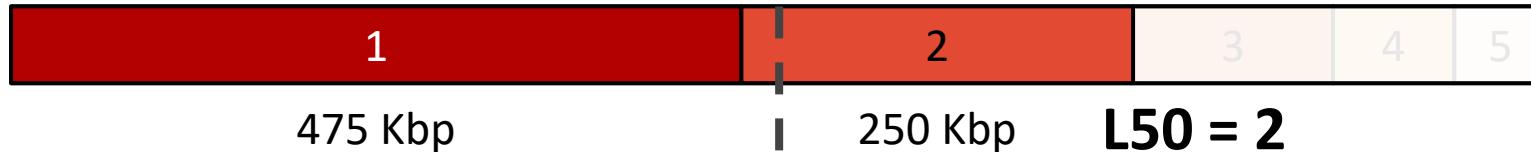


B



A

**N50 = 250 Kbp**



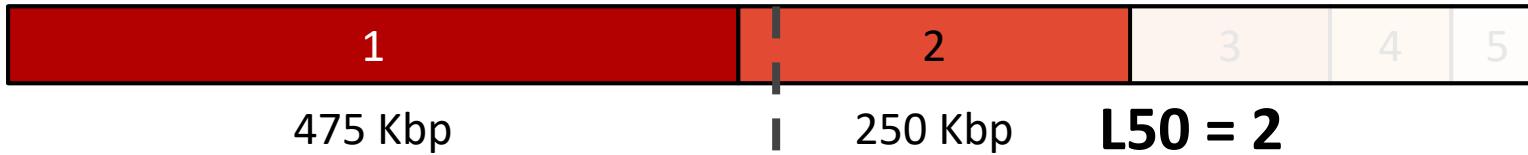
50%

B



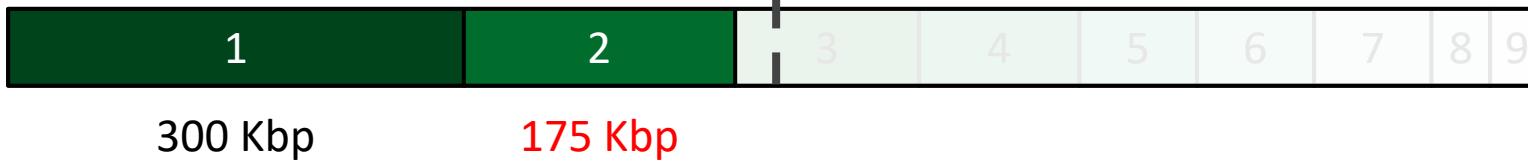
A

**N50 = 250 Kbp**



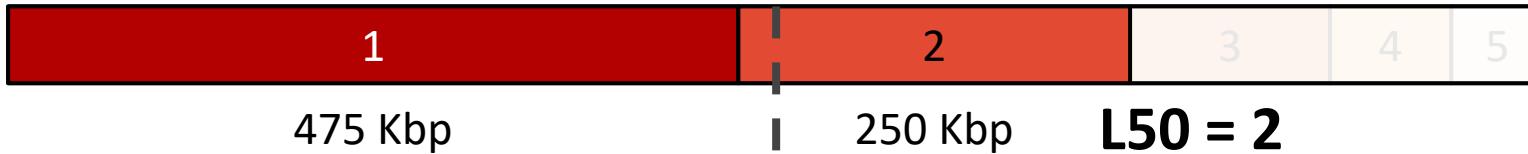
50%

B



A

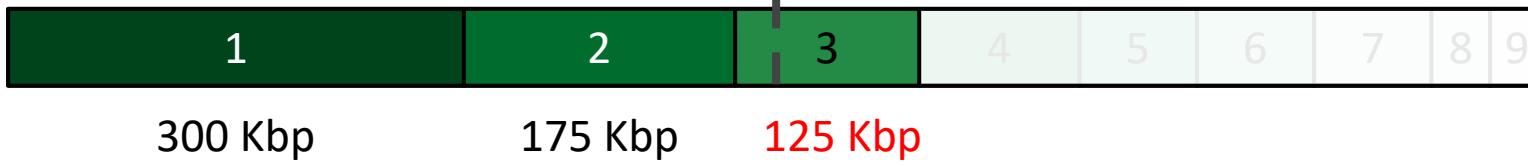
**N50 = 250 Kbp**



50%

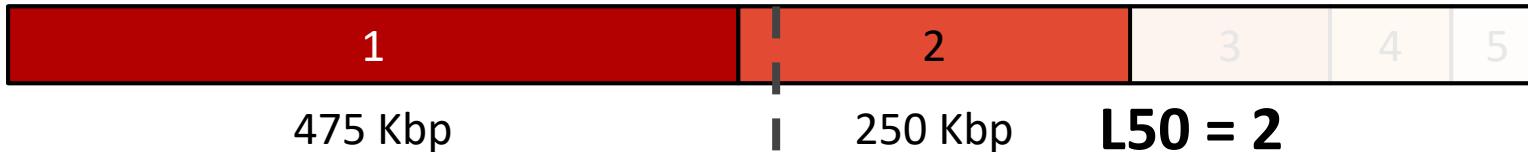
B

**N50 = 125 Kbp**



A

**N50 = 250 Kbp**

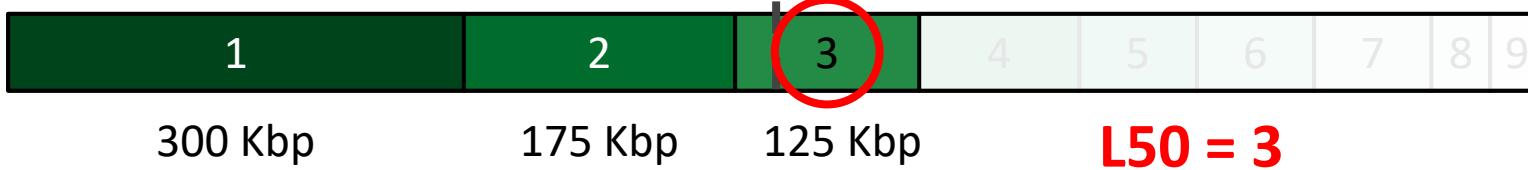


**L50 = 2**

50%

B

**N50 = 125 Kbp**



**L50 = 3**

A

**N50 = 250 Kbp**



475 Kbp

2

250 Kbp

**L50 = 2**

3

4

5

50%

B

**N50 = 125 Kbp**



300 Kbp

2

175 Kbp

3

125 Kbp

**L50 = 3**

4

5

6

7

8

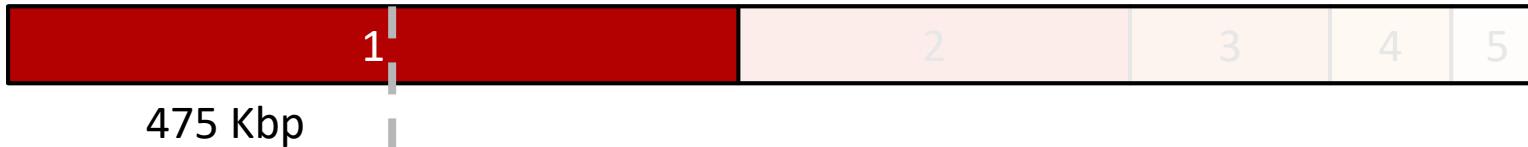
9

Which assembly is more  
complete?

A

# N25 & L25

A



25%

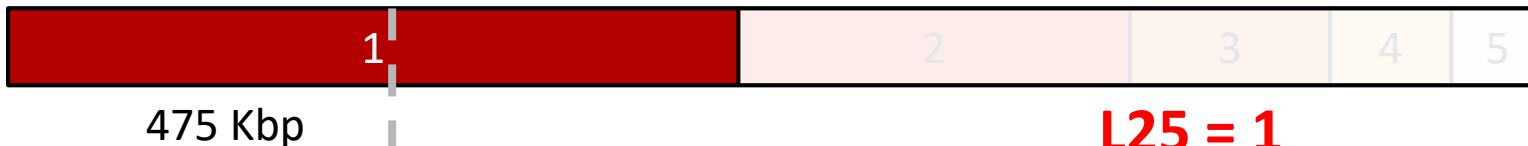
B



# N25 & L25

A

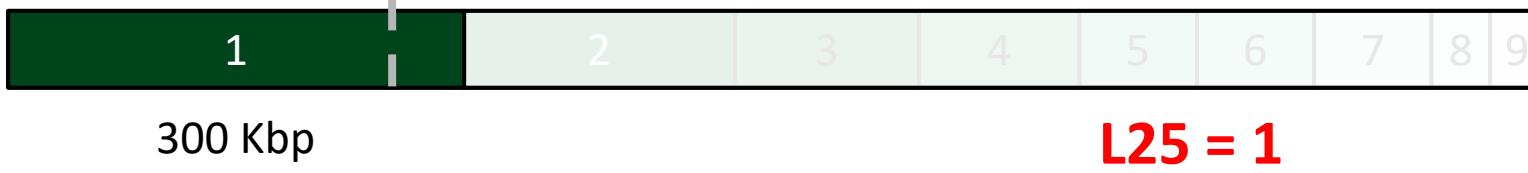
**N25 = 475 Kbp**



25%

B

**N25 = 300 Kbp**



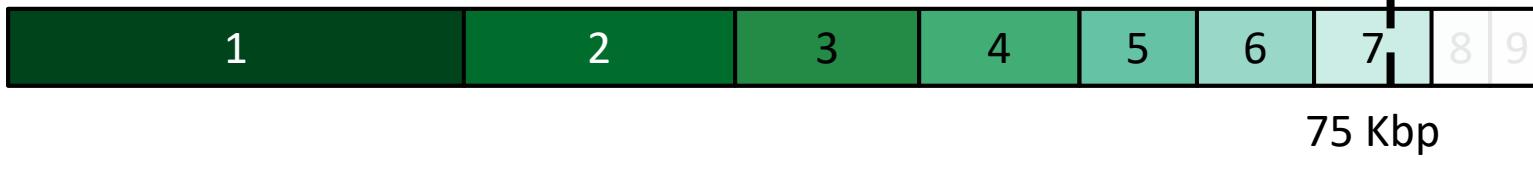
# N90 & L90

A



90%

B

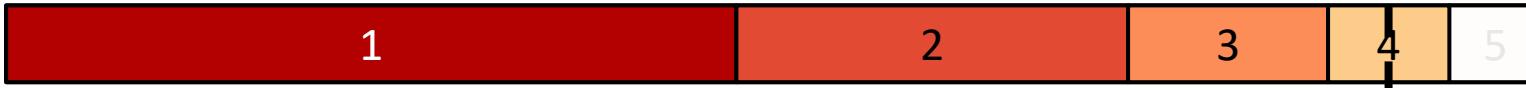


# N90 & L90

A

**N90 = 75 Kbp**

75 Kbp



**L90 = 4**

90%

B

**N90 = 75 Kbp**

75 Kbp



**L90 = 7**

**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  $\geq 500$  bp, unless otherwise noted (e.g., "# contigs ( $\geq 0$  bp)" and "Total length ( $\geq 0$  bp)" include all contigs).

**Statistics without reference** SRR3028792

# contigs	2631
# contigs ( $\geq 0$ bp)	3463
# contigs ( $\geq 1000$ bp)	1254
# contigs ( $\geq 5000$ bp)	10
# contigs ( $\geq 10000$ bp)	4
# contigs ( $\geq 25000$ bp)	2
# contigs ( $\geq 50000$ bp)	0
Largest contig	33 114
Total length	3 243 061
Total length ( $\geq 0$ bp)	3 579 102
Total length ( $\geq 1000$ bp)	2 258 968
Total length ( $\geq 5000$ bp)	137 834
Total length ( $\geq 10000$ bp)	102 283
Total length ( $\geq 25000$ bp)	60 964
Total length ( $\geq 50000$ 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

The plot shows the cumulative length of contigs ordered from largest to smallest. The x-axis is labeled "Contigs" and ranges from 0 to 2500, with major ticks every 250 units. The y-axis is labeled "Length" and ranges from 0 to 3.5 Mbp, with major ticks every 0.5 units. A red curve starts at the origin (0,0) and rises monotonically, indicating that the total length increases as more contigs are included. The curve appears to be concave down, suggesting that larger contigs contribute significantly to the total length.

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

- **Install with conda**

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

- **Run quast to generate quality report**

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

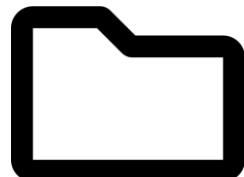
- **Install with conda**

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

- **Run quast to generate quality report**

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

- **Output**



quast\_results/

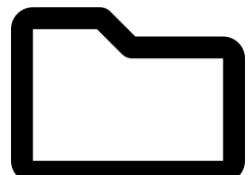
- **Install with conda**

```
$ 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](https://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](https://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](https://github.com/ablab/quast)

Worst Median Best  Show heatmap

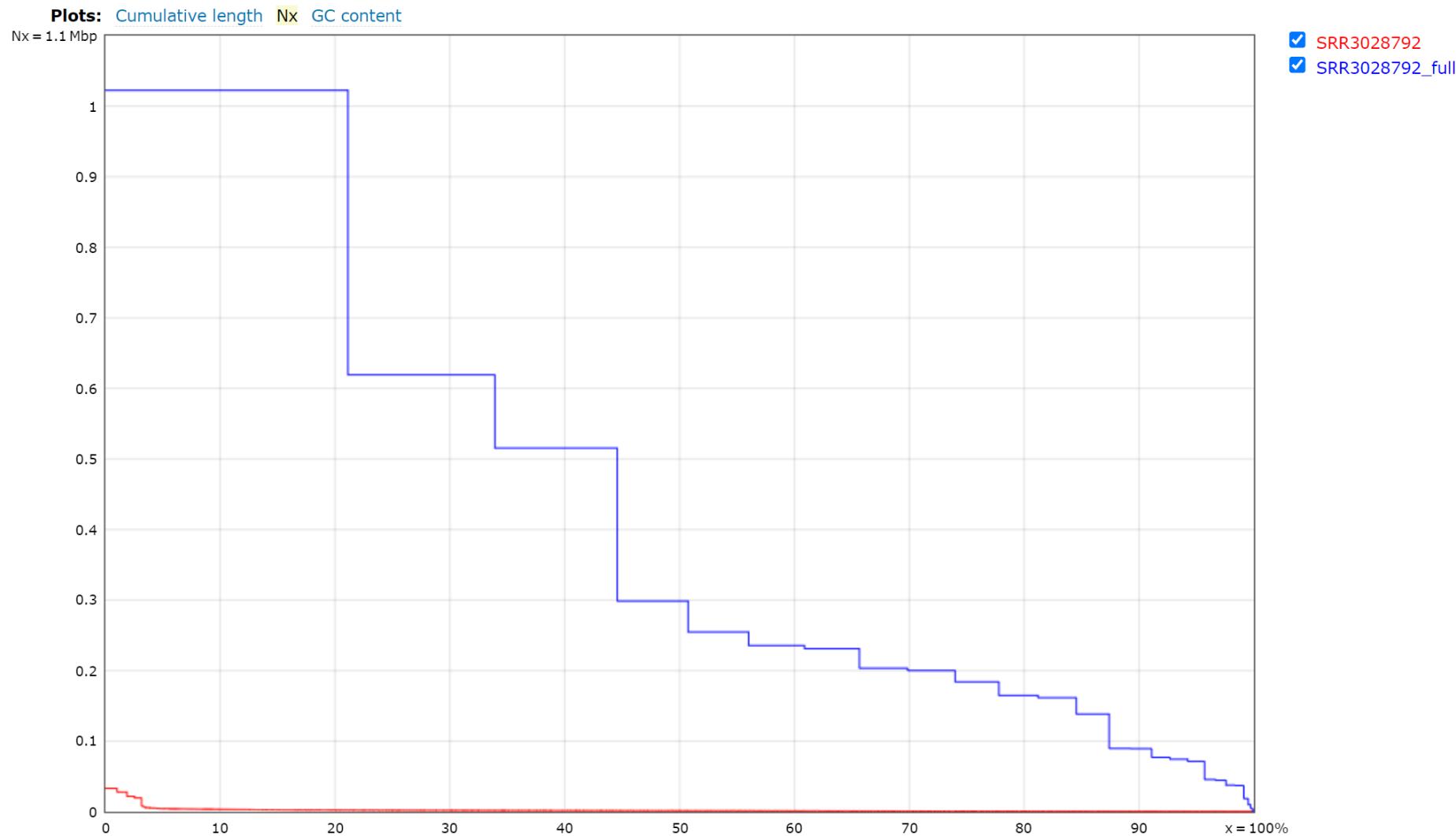
Statistics without reference			SRR3028792	SRR3028792_full
# contigs	2631	32		
# contigs (>= 0 bp)	3463	38		
# contigs (>= 1000 bp)	1254	27		
# contigs (>= 5000 bp)	10	24		
# contigs (>= 10000 bp)	4	24		
# contigs (>= 25000 bp)	2	22		
# contigs (>= 50000 bp)	0	18		
Largest contig	33 114	1 022 474		
Total 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		
Matches				
# 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

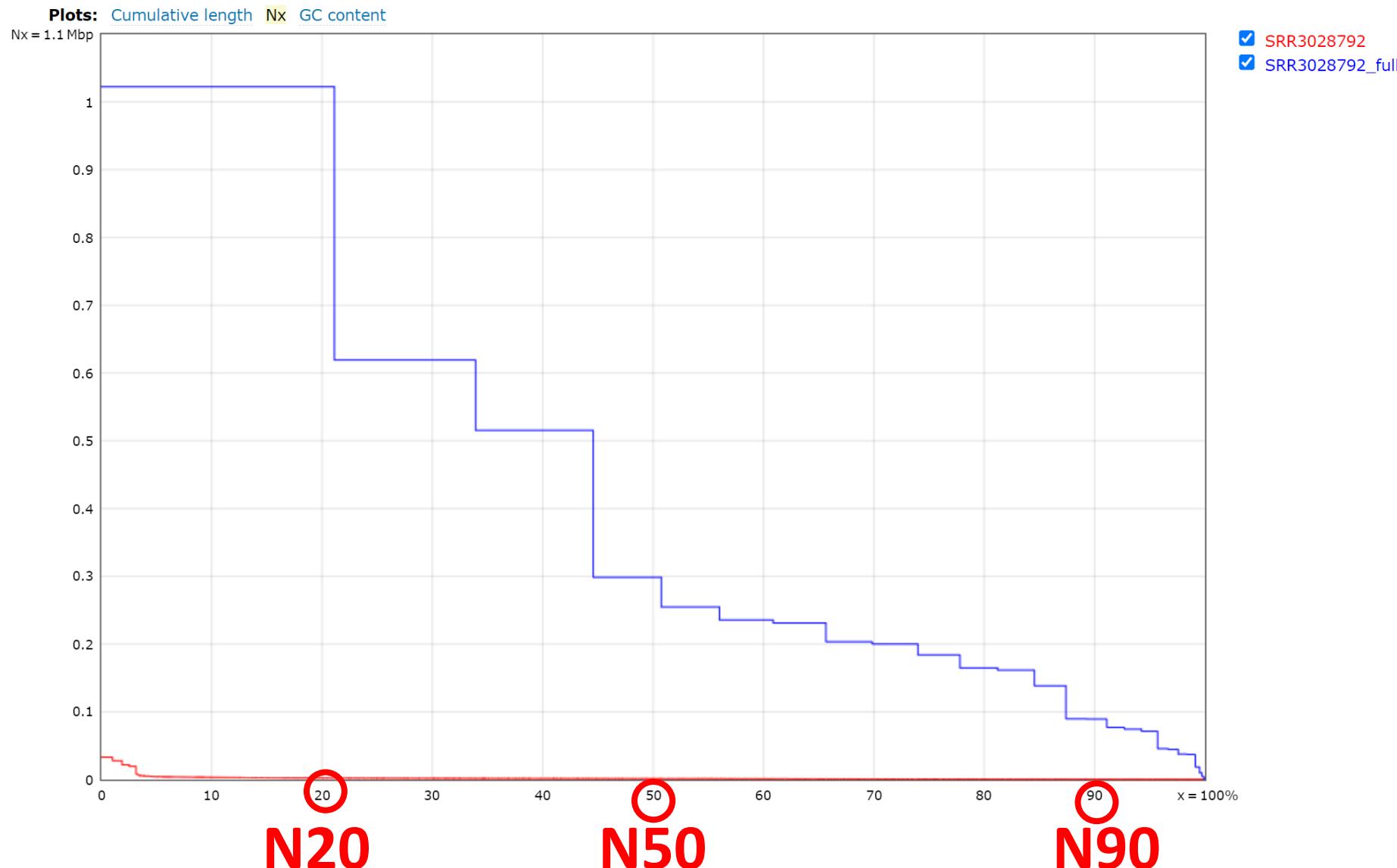
# Quality report

[github.com/ablab/quast](https://github.com/ablab/quast)



# Quality report

[github.com/ablab/quast](https://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](https://github.com/MMID-coding-workshop/2022-01-26-Downloading-and-assembling-microbial-sequence-data)

## Downloading and assembling microbial sequence data



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 link.

# Full tutorial on genome assembly

[github.com/MMID-coding-workshop/2022-01-26-Downloading-and-assembling-microbial-sequence-data](https://github.com/MMID-coding-workshop/2022-01-26-Downloading-and-assembling-microbial-sequence-data)

## Downloading and assembling microbial sequence data



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 notebook:

- [microbial-genome-assembly.ipynb](#)

[View workshop tutorial  
\(command-line\)](#)

◦ If you wish to launch this notebook in a cloud-based environment to follow along please click the



link.

# Run using command-line

## 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:
```

Copy this  
command to  
terminal

Expected  
output

# Full tutorial on genome assembly

[github.com/MMID-coding-workshop/2022-01-26-Downloading-and-assembling-microbial-sequence-data](https://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

 launch binder

Launch tutorial in a cloud environment

to follow along please click the

# Run tutorial using cloud

Uses [Jupyter](#) to provide instructions + BASH commands

The screenshot shows a Jupyter Notebook interface. On the left, there is a sidebar titled "MICROBIAL-GENOME-ASSEMBLY.IPYNB" containing a table of contents:

- 1. Downloading and assembling microbial sequence data
  - 1.1. Copy-paste commands to a separate terminal
  - 1.2. Running from within Jupyter
- 2. Download genomes from NCBI
  - 2.1. Configure conda
  - 2.2. Install sra-tools using conda
  - 2.3. Verify sra-tools is working
  - 2.4. Prefetch genomes (prefetch) [selected]
  - 2.5. Convert to fastq (fasterq-dump)
- 3. Reduce dataset size (seqtk) (optional)
  - 3.1. Install seqtk
  - 3.2. Select a random sample of reads
- 4. Quality filter files (fastp)

The main content area has a title bar "microbial-genome-assembly" and a toolbar with various icons. The main content is a Markdown cell:

## 2.4. Prefetch genomes (prefetch)

The `prefetch` command that is part of the `sra-tools` package can be used to download and store sequence read data from NCBI's Sequence Read Archive. You can run it like so:

```
[4]: conda run -n sra-tools prefetch SRR3028792
```

2022-01-24T06:12:46 prefetch.2.11.0: 1) 'SRR3028792' is found locally  
2022-01-24T06:12:46 prefetch.2.11.0: 'SRR3028792' has 0 unresolved dependencies

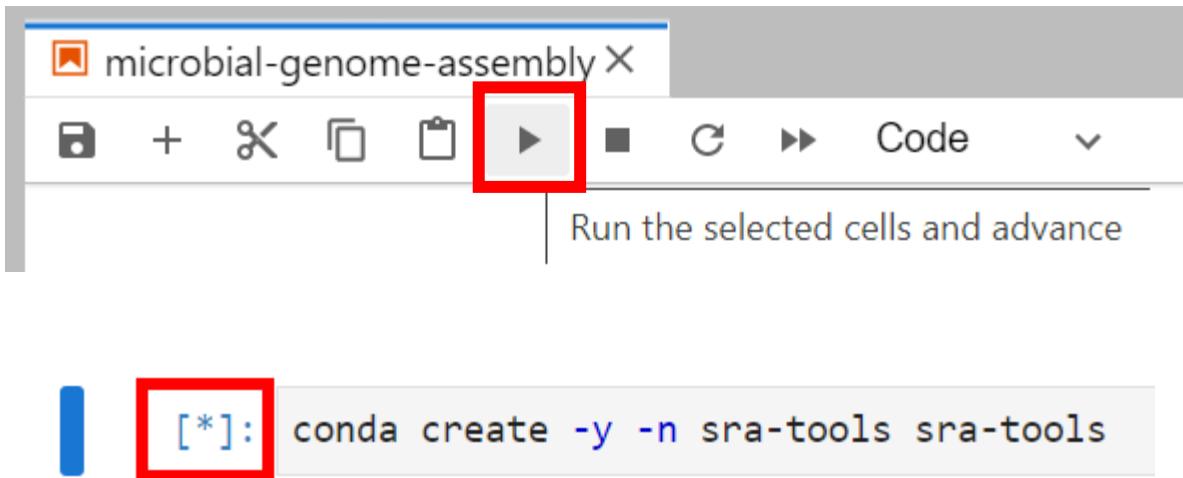
You can even pass multiple accession numbers to `prefetch` like so:

```
[5]: conda run -n sra-tools prefetch SRR3028792 SRR3028793
```

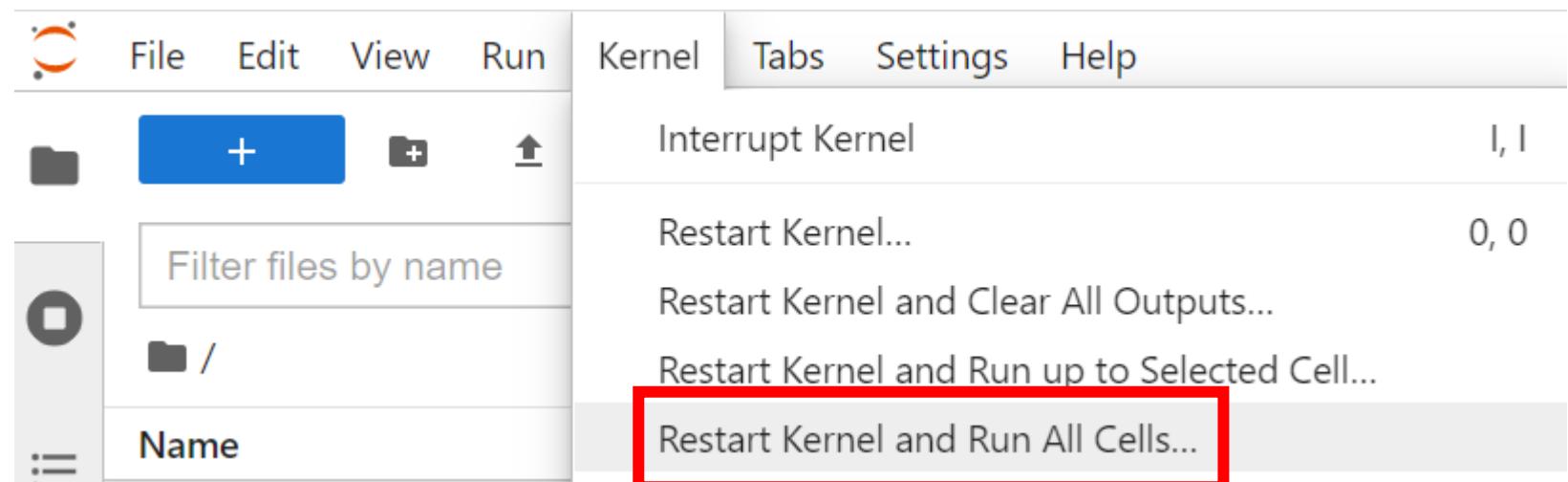
2022-01-24T06:12:47 prefetch.2.11.0: 1) 'SRR3028792' is found locally  
2022-01-24T06:12:48 prefetch.2.11.0: 'SRR3028792' has 0 unresolved dependencies

2022-01-24T06:12:48 prefetch.2.11.0: 2) 'SRR3028793' is found locally  
2022-01-24T06:12:48 prefetch.2.11.0: 'SRR3028793' has 0 unresolved dependencies

# Run tutorial using cloud



Run by selecting  
“Run All Cells”



Run by selecting  
“Play” button

# HELPFUL RESOURCES

**Illumina Sequencing:** [youtu.be/fCd6B5HRaZ8](https://youtu.be/fCd6B5HRaZ8)

**Nanopore Sequencing:**

[nanoporetech.com/resource-centre/introduction-nanopore-sequencing](https://nanoporetech.com/resource-centre/introduction-nanopore-sequencing)

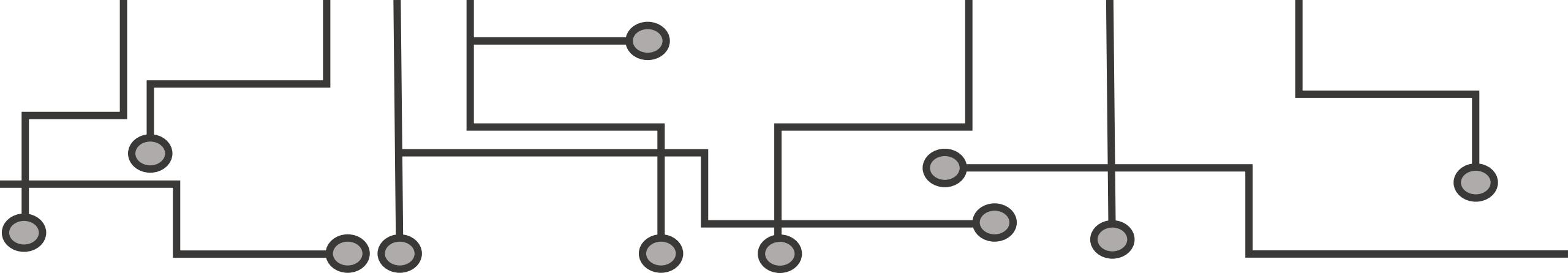
**Downloading from SRA:** [www.ncbi.nlm.nih.gov/sra/docs/srownload/](https://www.ncbi.nlm.nih.gov/sra/docs/srownload/)

**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*

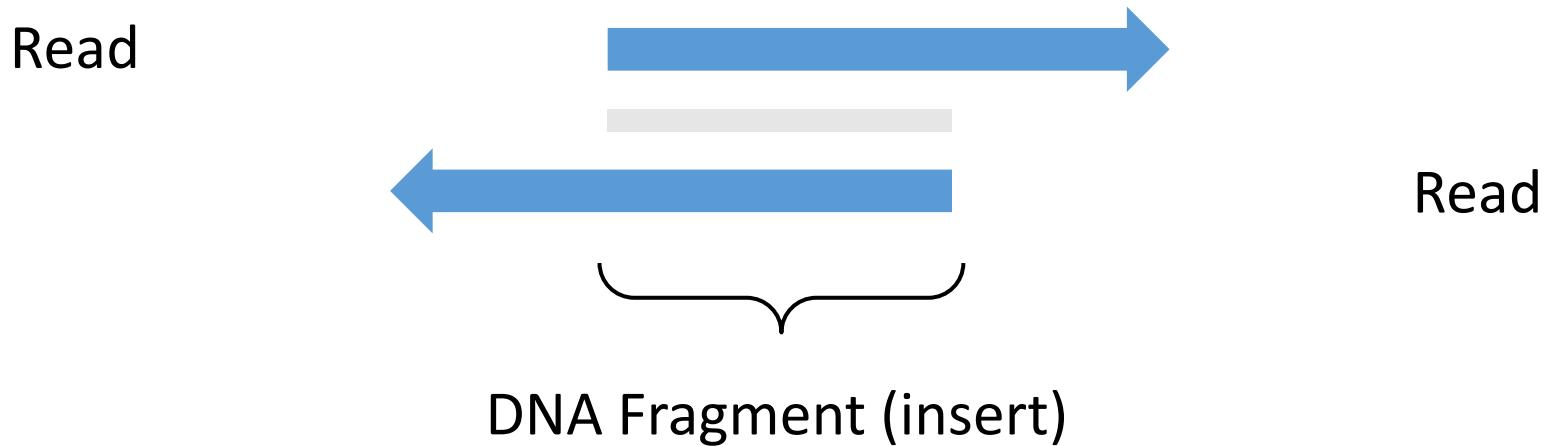


**THANK YOU FOR ATTENDING!**  
*The Q&A Session will now begin.*

**Please make sure to fill out the [Exit Survey](#)**  
**We value your feedback!**

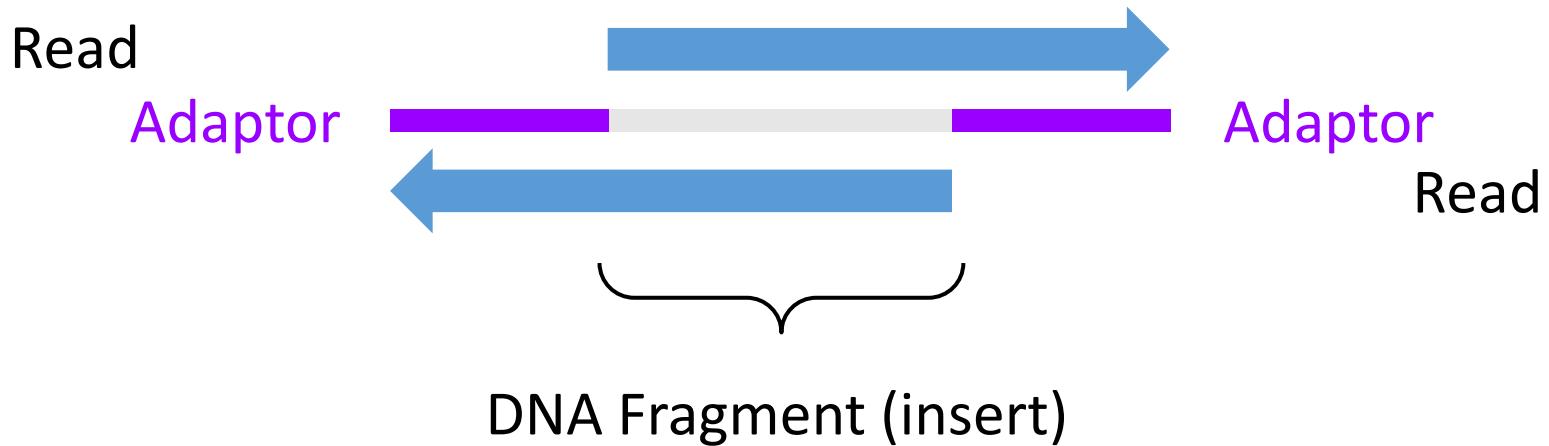
**More questions? Please email us at  
[mmid.coding.workshop@gmail.com](mailto:mmid.coding.workshop@gmail.com) or post them to the workshop slack channel**

# Paired-end sequencing



...or very short.

# Paired-end sequencing



...or very short.