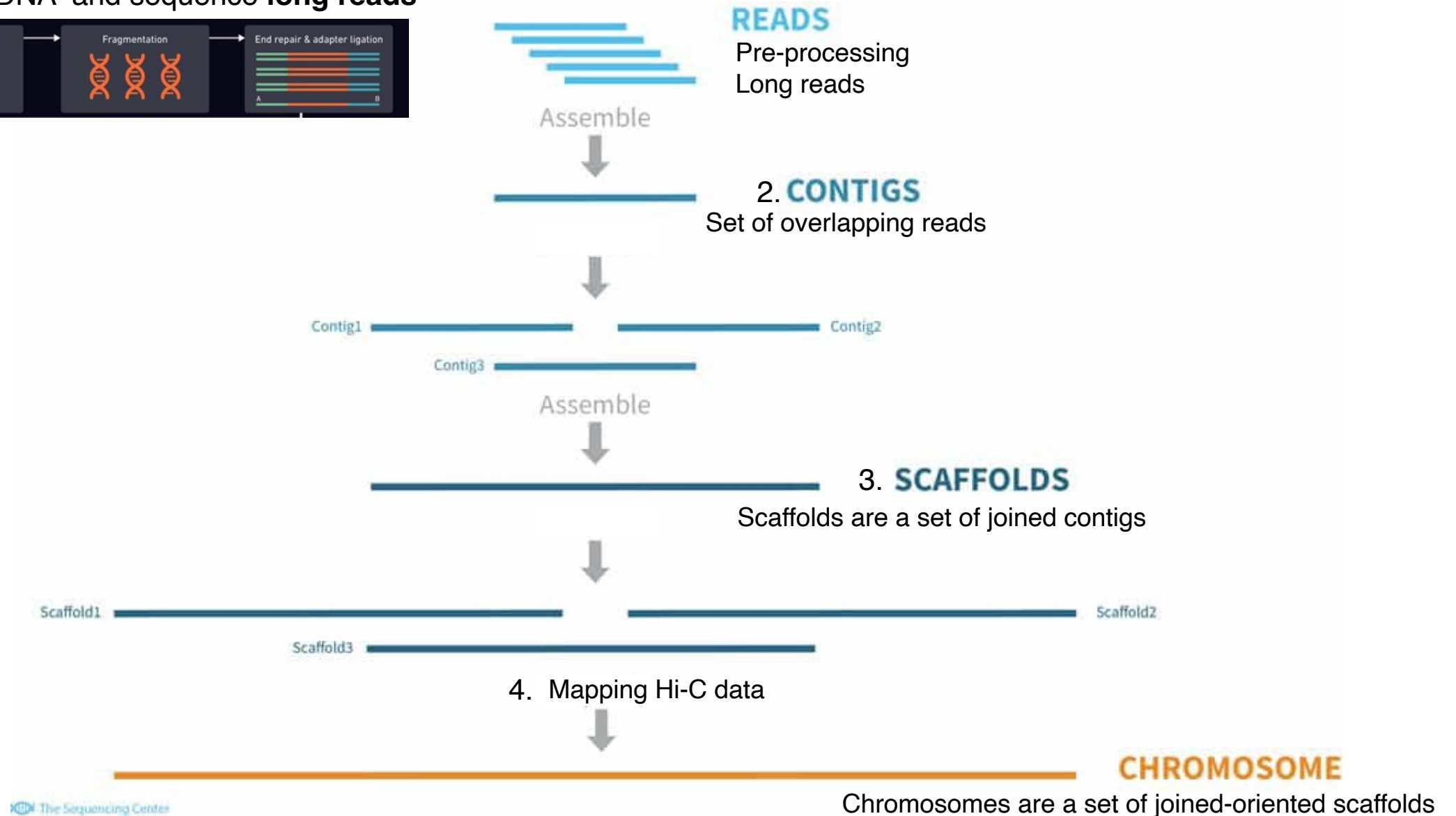
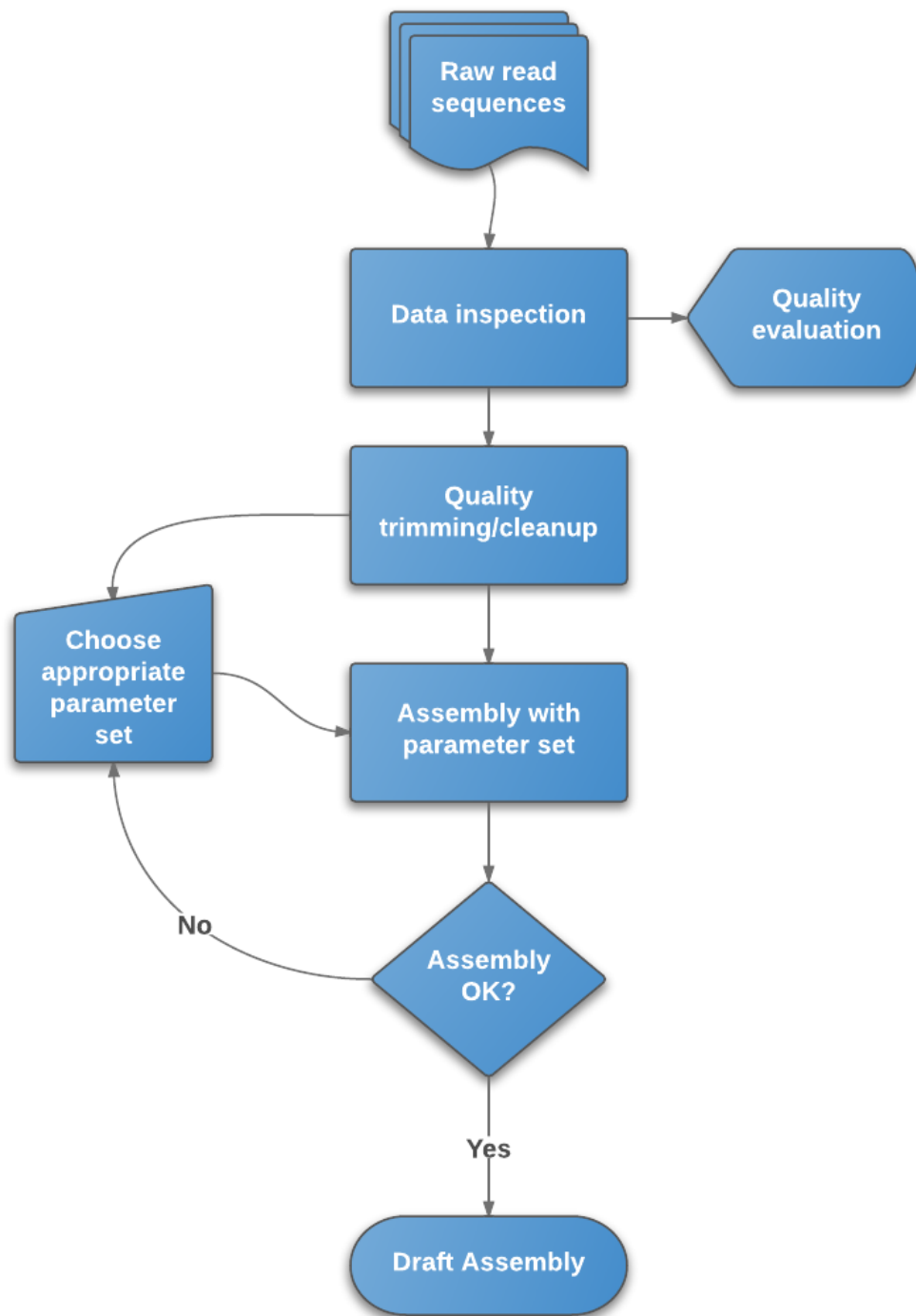


Genome assembly
Biodiversity
genomics
course
Tena-Ecuador 2024



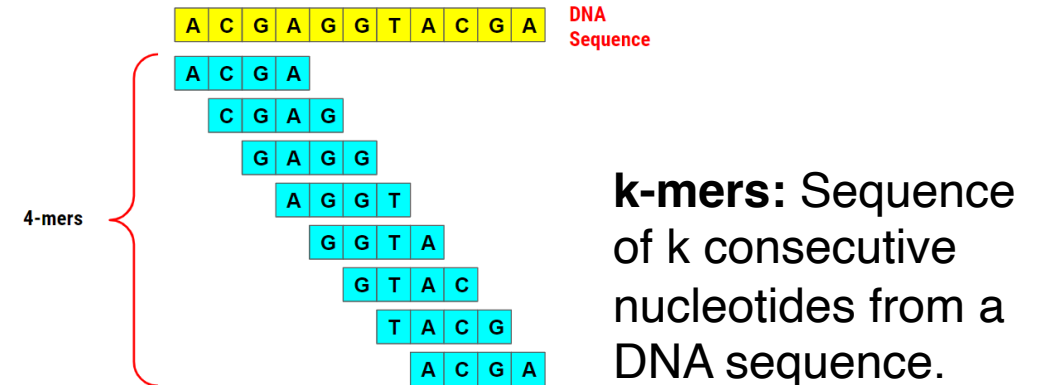
1. Extract DNA and sequence long reads





1. Sequencing of the reads

- Min. length is **500-1000 bp** for long reads (e.g., PacBio, Oxford Nanopore).
- Raw read sequences is usually stored in a **FastQ** file.
- Reads must overlap by a minimum number of base pairs, or **k-mers**, before they can be mapped together.



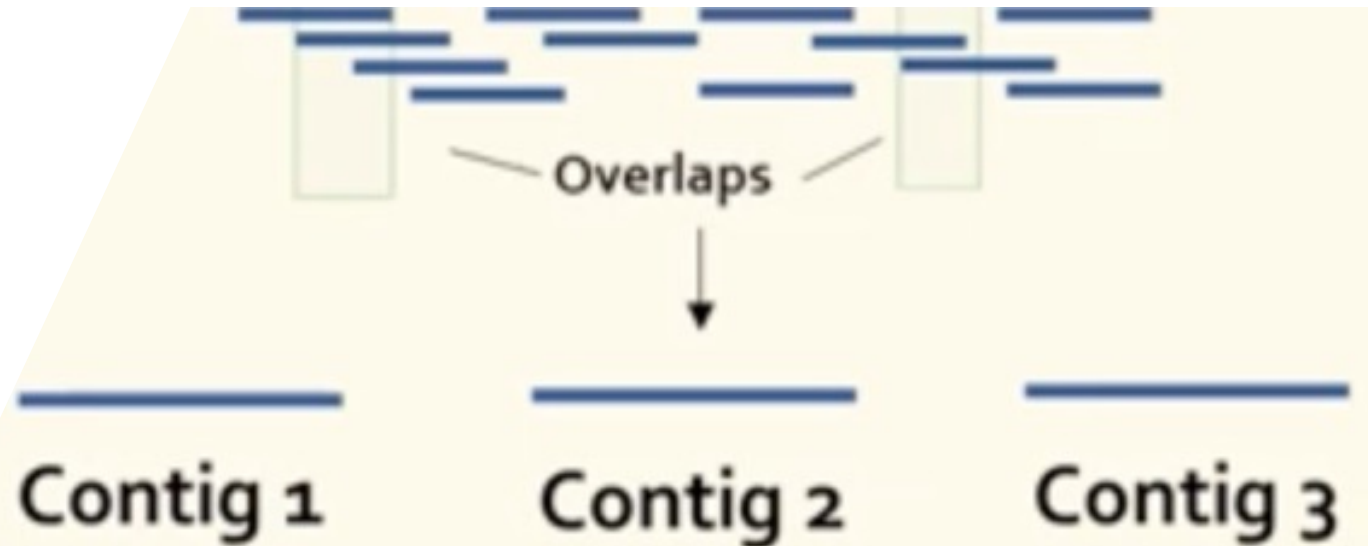
2. Assembling the reads

Aligned reads

```
TGAAGTCCTACAGTCATAGTC
AAGTCCTACAGTCATAGTCGA
GTCCTACAGTCATAGTCGATA
CCTACAGTCATAGTCGATATT
TACAGTCATAGTCGATATT
```

Consensus contig TGAAGTCCTACAGTCATAGTCGATATT

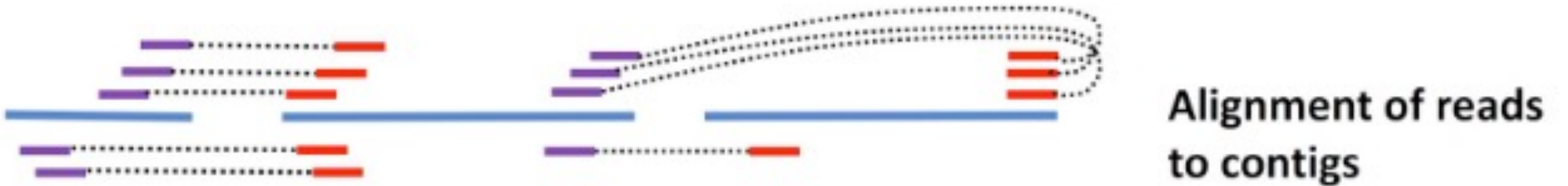
Although the contigs are longer than reads and do not have gaps, they are not large enough to cover or represent entire chromosomes.



Canu
Trinity
SPAdes
Flye
...

3. Scaffolding (Assembling the contigs)

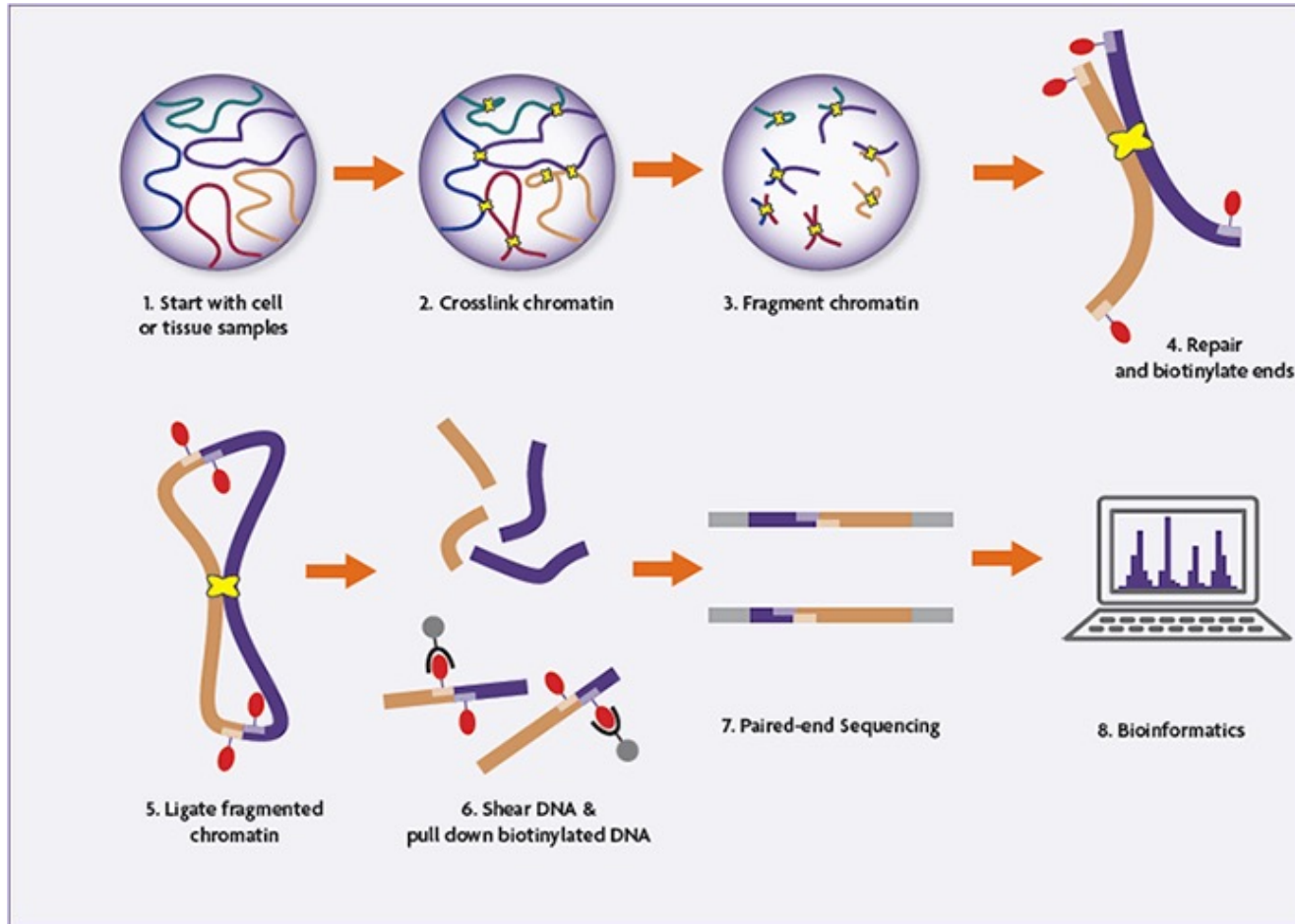
1. Mapping HiC data to Contigs:



This mapping process aims to:

- Validate the contigs by confirming that the reads used to assemble them align correctly.
- Identify any errors in the contig sequences.
- Calculate coverage depth (how many reads cover each position of the contig).

Hi-C data



1. Cross-linking: To fix the DNA and preserve interactions between regions that are close together in the 3D.

2. Fragment: To cut the cross-linked DNA into smaller pieces.

3. Proximity Ligation: To join DNA fragments that were close together in the 3D space, capturing their interactions

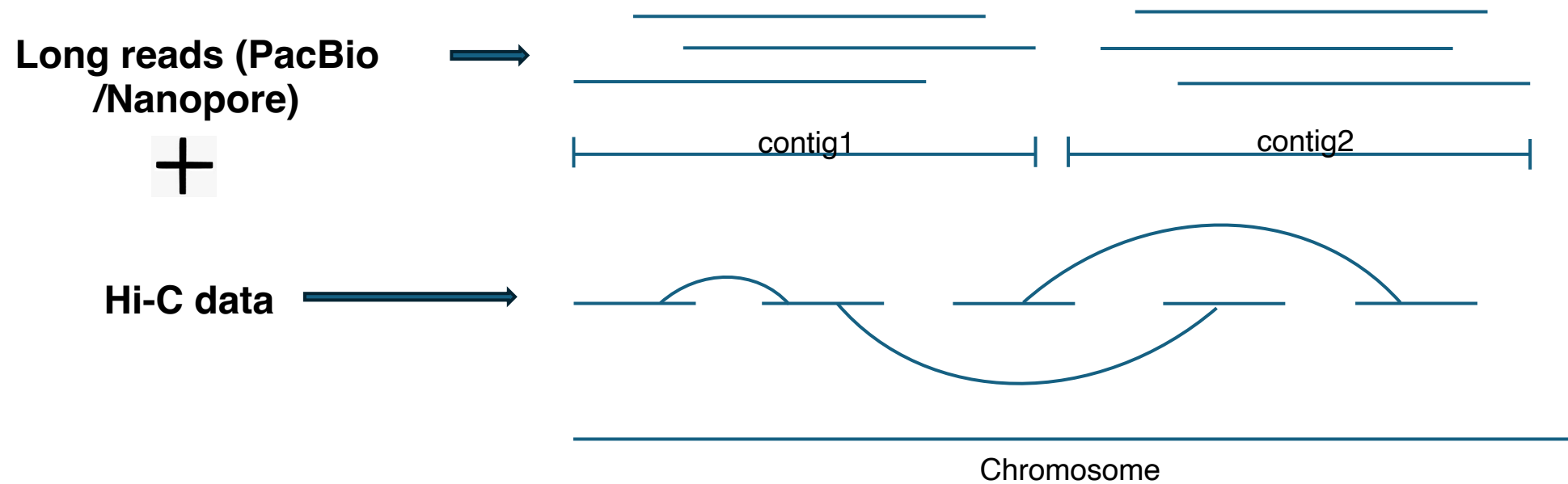
4. Reversal of Cross-links: To revert the cross-linking and purify the DNA.

5. DNA Fragmentation and Size Selection

6. Library Preparation: add adapters

7. Sequencing


Chromosome assembly



https://pipelines.tol.sanger.ac.uk/

[Home](#)[Pipelines](#)[Resources](#)[Docs](#)[Events](#)[About](#)

sanger-tol



Workflows and tools to investigate the genomic diversity of complex organisms.

VIEW PIPELINES

Search

Search


Available Pipelines

Can you think of another pipeline that would fit in well? [Let us know!](#)

Search keywords

Filter: Released 10 Under development 4

Sort: Last Release Alphabetical Stars

Display: 

[sanger-tol/curationpretext](#)

genomics hic

A Nextflow DSL2 pipeline for pretext generation in curation

Version 1.0.0 Published 2 days ago

[sanger-tol/treeval](#)

curation genome-alignment genome-assembly genomics quality-control synteny

Pipelines for the production of Treeval data

Version 1.1.1 Published 2 weeks ago

[sanger-tol/ensemblgenedownload](#)

download gene-annotation genomics indexing

Nextflow pipeline to download gene annotations from Ensembl onto the Tree of Life directory structure

Version 2.0.0 Published 2 months ago

[sanger-tol/ensemblrepeatdownload](#)

download formatting genomics

Nextflow pipeline to download repeat annotations from Ensembl onto the Tree of Life directory structure

Version 2.0.0 Published 2 months ago

[sanger-tol/insdcdownload](#)

download genomics indexing

Nextflow pipeline to download assemblies from INSDC and add them to the Tree of Life directories

Version 2.0.0 Published 2 months ago

[sanger-tol/variantcalling](#)

alignment genomics variant-calling

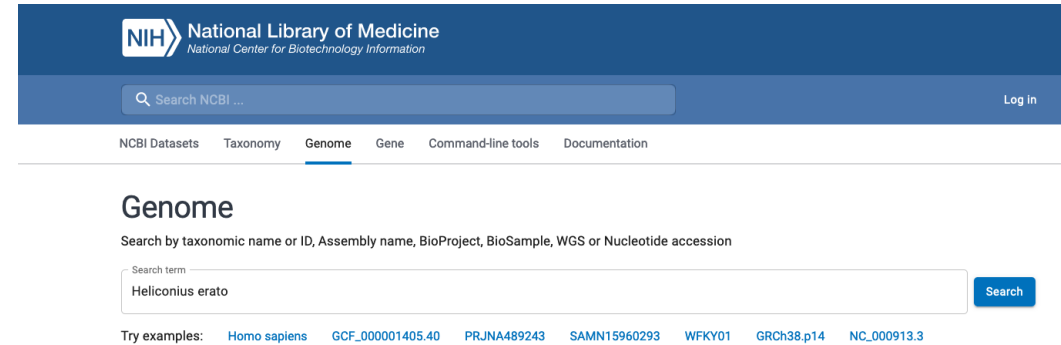
This Nextflow DSL2 pipeline calls variants on long read alignment. It is run after sanger-tol/readmap in the Sanger ToL production suite but with options to run on unaligned reads.

Version 1.1.3 Published 2 months ago

Where can we find genomes?

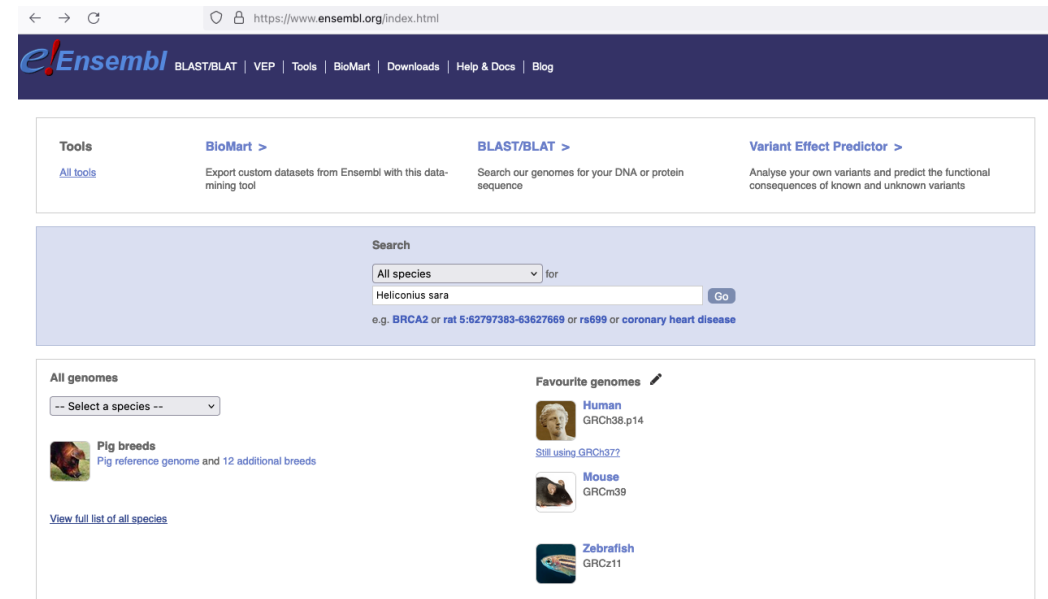
- NCBI (National Center for Biotechnology Information):

<https://www.ncbi.nlm.nih.gov/datasets/genome/>



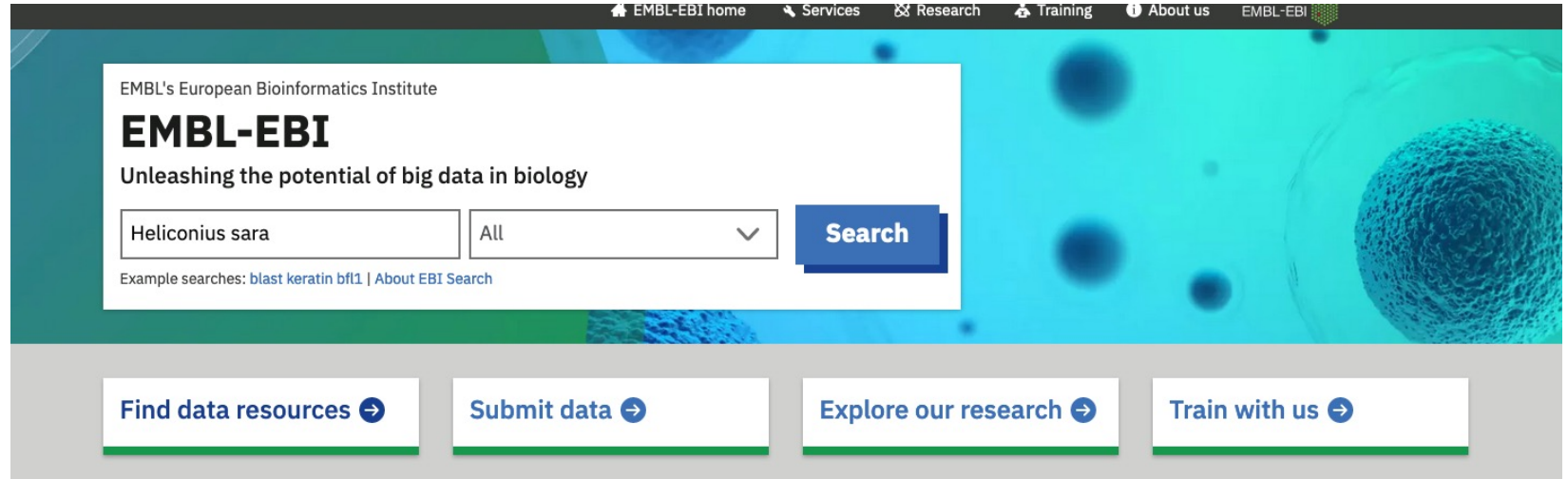
- ensembl

<https://www.ensembl.org/index.html>



➤ EMBL-EBI (European Molecular Biology Laboratory - European Bioinformatics Institute):

<https://www.ebi.ac.uk/>



Example:

https://www.ncbi.nlm.nih.gov/datasets/genome/

An official website of the United States government [Here's how you know](#)

NIH

National Library of Medicine

National Center for Biotechnology Information

Search NCBI ...

Log in

NCBI Datasets

Taxonomy

Genome

Gene

Command-line tools

Documentation

Genome

Search by taxonomic name or ID, Assembly name, BioProject, BioSample, WGS or Nucleotide accession

Search term

Heliconius sara

×

Search

Try examples: [Homo sapiens](#) [GCF_000001405.40](#) [PRJNA489243](#) [SAMN15960293](#) [WFKY01](#) [GRCh38.p14](#) [NC_000913.3](#)

https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=33443

An official website of the United States government [Here's how you know](#)

NIH

National Library of Medicine

National Center for Biotechnology Information

Search NCBI ...

Log in

NCBI Datasets

Taxonomy

Genome

Gene

Command-line tools

Documentation

Genome

Download a genome data package including genome, transcript and protein sequence, annotation and a data report

Selected taxa

Heliconius sara

Enter one or more taxonomic names

⌵

Filters

⌵

Download

Select columns

2 Genomes 1 selected

Rows per page

20

1-2 of 2

<

>

☐

Assembly

GenBank

RefSeq

Scientific name

Modifier

Annotation

Action

☒

iHelSar1.2

GCA_917862395.2

Heliconius sara

⋮

☐

iHelSar1.1 alternate haplotype...

GCA_911392485.1

Heliconius sara

⋮

https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_917862395.2/

An official website of the United States government [Here's how you know](#)

NIH

National Library of Medicine

National Center for Biotechnology Information

Search NCBI ...

Log in

NCBI Datasets

Taxonomy

Genome

Gene

Command-line tools

Documentation

Genome assembly iHelSar1.2

Download

datasets

URL

FTP

Submitted GenBank assembly

GCA_917862395.2

⋮

Taxon

Heliconius sara

WGS project

CAKJTV02

Assembly type

haploid

Submitter

WELLCOME SANGER INSTITUTE

Date

Apr 6, 2023

View the [legacy Assembly page](#)

BLAST the reference genome

https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/917/862/395/GCA_917862395.2_iHelSar1.2/

Index of /genomes/all/GCA/917/862/395/GCA_917862395.2_iHelSar1.2

Name	Last modified	Size
Parent Directory		-
GCA_917862395.2_iHelSar1.2_assembly_report.txt	2023-04-18 17:05	34K
GCA_917862395.2_iHelSar1.2_assembly_stats.txt	2023-04-18 17:05	31K
GCA_917862395.2_iHelSar1.2_feature_count.txt	2024-02-15 01:58	168
GCA_917862395.2_iHelSar1.2_genomic.fna.gz	2023-04-18 17:05	107M
GCA_917862395.2_iHelSar1.2_genomic.gbff.gz	2023-04-18 17:05	138M
GCA_917862395.2_iHelSar1.2_genomic_gaps.txt.gz	2023-04-18 17:05	1.4K
GCA_917862395.2_iHelSar1.2_wgsmaster.gbff.gz	2023-04-18 17:05	1.4K
README.txt	2024-04-11 16:11	54K
annotation_hashes.txt	2024-02-17 10:02	410
assembly_status.txt	2024-07-15 13:53	14
md5checksums.txt	2024-02-17 10:02	625
uncompressed_checksums.txt	2024-06-14 01:21	167

[HHS Vulnerability Disclosure](#)

wget

https://ftp.ncbi.nlm.nih.gov/genomes/all/GC
A/917/862/395/GCA_917862395.2_iHelSar1.2/GC
A_917862395.2_iHelSar1.2_genomic.fna.gz

FASTA Format

It is a text-based format for representing nucleotide sequences (DNA or RNA)

```
>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6) ← Description Line  
ACCTCTTTTCTTATCATTGACATTTAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGCTGTCAGATCT  
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC  
CCTCCGCTCCCAGGTAACCGCCCCGGGCTCCGGCCCCGGCCCCGGCTCGGGGCCCGCGGGGCCTCTCCGCTG  
CCAGCGACTGCTGTCCCCAAATCAAAGCCCCGCCCAAGTGGCCCCGGGGCTTGATTTTTTGCTTTTAAAAG  
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCTT  
TGCCGAGTGTGCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACTTCCAGTCCGGCCCT ← Sequence Lines  
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA  
GGAGGTGGGGACGCACTTTGCATCCAGACCTCCTCTGCATCGCAGTTCACGACATCCACGCTTGGGAAAG  
TCCGTACCCGCGCCTGGAGCGCTTAAAGACACCCTGCCGCGGGTCGGGCGAGGTGCAGCAGAAGTTTCCC  
GCGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC
```

Description Line: Begins with a > character followed by an identifier and optional description.

Sequence Lines: The actual sequence data, typically represented over multiple lines for readability.

Assembly statistics

	GenBank
Genome size	562.2 Mb
Total ungapped length	562.2 Mb
Number of contigs	203,084
Contig N50	9.9 kb
Contig L50	14,088
GC percent	33.5
Genome coverage	1.9x
Assembly level	Contig

It is the median of the contig lengths.

It is the number of contigs (or scaffolds) that together contain at least 50% of the total assembly length.

Significance

N50: A higher N50 value indicates that the assembly contains longer contigs, suggesting better assembly quality and continuity.

L50: A lower L50 value indicates that fewer contigs are needed to cover 50% of the genome, suggesting higher assembly contiguity.

★ **Quick exercise:** Imagine you have contigs of the following lengths (in kb): 10, 15, 25, 30, 50, 70.

1. Sort the contigs in descending order
2. Calculate the total length
3. Find the N50
4. Find the L50

1. Sort the contigs in descending order: 70, 50, 30, 25, 15, 10.
2. Calculate the total length: $70 + 50 + 30 + 25 + 15 + 10 = 200$ kb.
3. Find the N50:
 - 50% of 200 kb is 100 kb.
 - Cumulative lengths: 70 (70 kb), $70 + 50 = 120$ kb (reaches 100 kb threshold).
 - The contig length where we first exceed 100 kb cumulative is 50 kb.
 - **N50 = 50 kb.**
4. Find the L50:
 - Cumulative lengths: 70 (70 kb), $70 + 50 = 120$ kb (reaches 100 kb threshold).
 - Number of contigs used to reach this point: 2.
 - **L50 = 2 contigs.**

References

- Jay Ghurye and Mihai Popl. 2019. Modern technologies and algorithms for scaffolding assembled genomes. Plos computational biology
- Edward S. Rice and Richard E. Green. 2018. New Approaches for Genome Assembly and Scaffolding. Annual Review of Animal Biosciences.
- https://ucdavis-bioinformatics-training.github.io/2020-Genome_Assembly_Workshop/kmers/kmers
- <https://www.youtube.com/watch?v=8w-CbyGV-pgv>
- <https://star-protocols.cell.com/protocols/1799>
- <https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html#per-sequence-quality-scores>