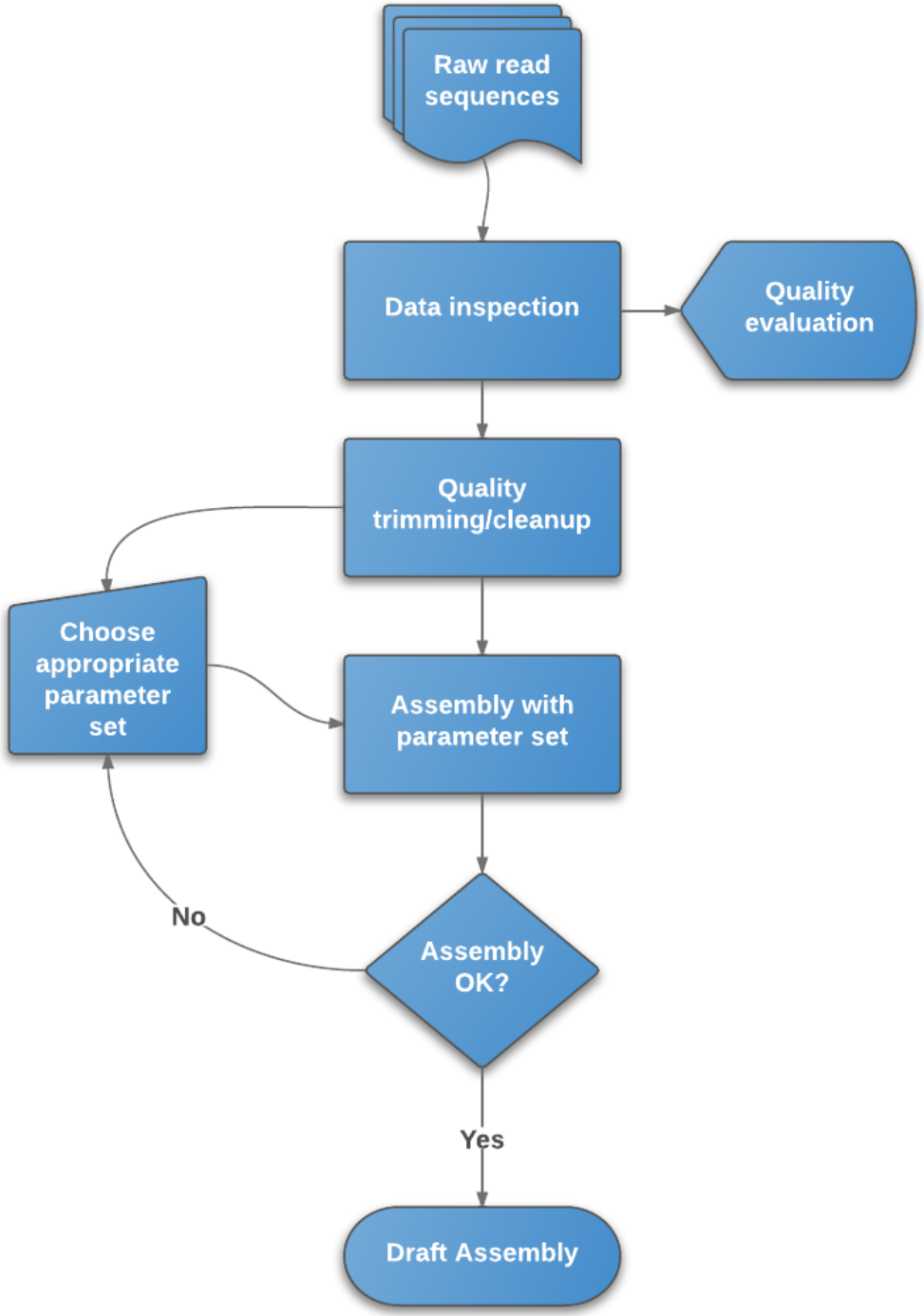
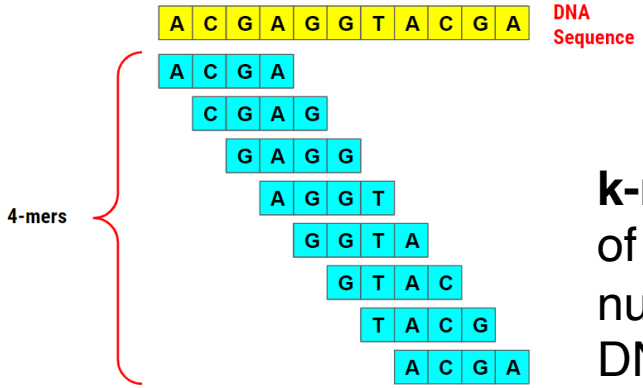
Genome assembly Biodiversity genomics course Tena-Ecuador 2024

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.



k-mers: Sequence of k consecutive nucleotides from a DNA sequence.

Canu
Trinity
SPAdes
Flye
…

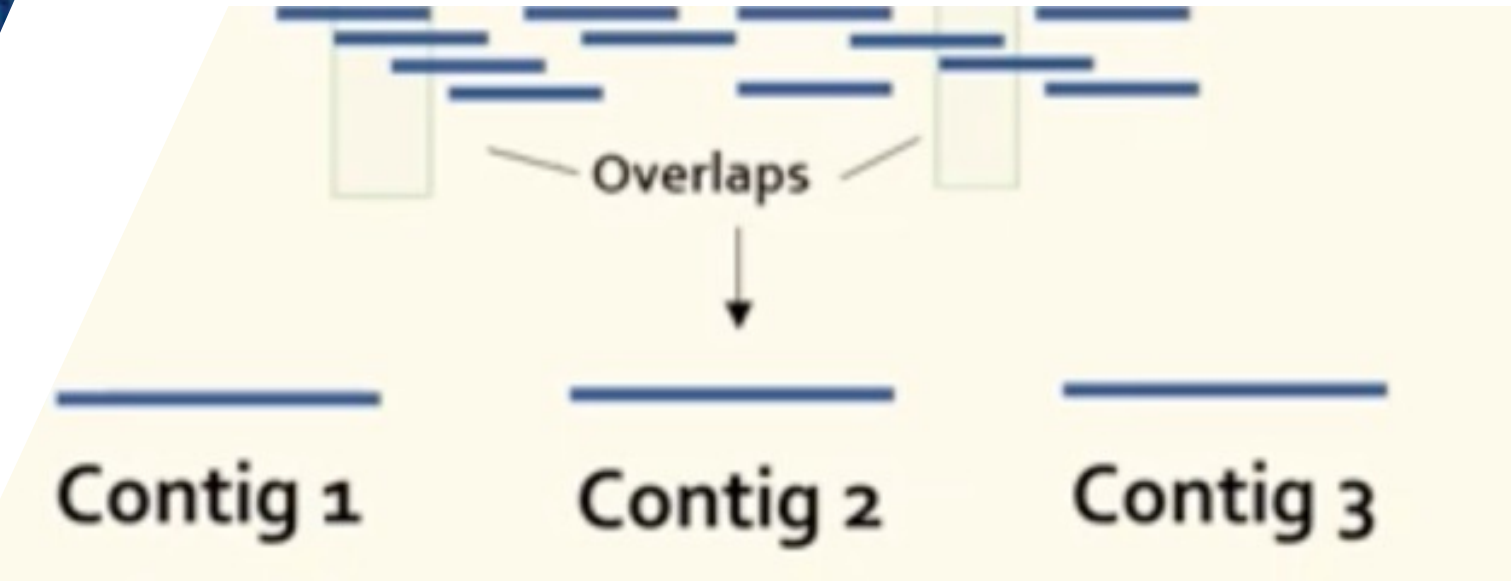
2. Assembling the reads

Aligned reads

TGAAGTCCTACAGTCATAGTC
AAGTCCTACAGTCATAGTCGA
GTCCTACAGTCATAGTCGATA
CCTACAGTCATAGTCGATATT
TACAGTCATAGTCGATATTT

Consensus contig TGAAGTCCTACAGTCATAGTCGATATTT

Although the contigs are longer than reads and do not have gaps, they are no large enough to cover or represent entire chromosomes.



3. Scaffolding (Assembling the contigs)

1. Mapping HiC data to Contigs:

Alignment of reads to contigs

This mapping process aims to:

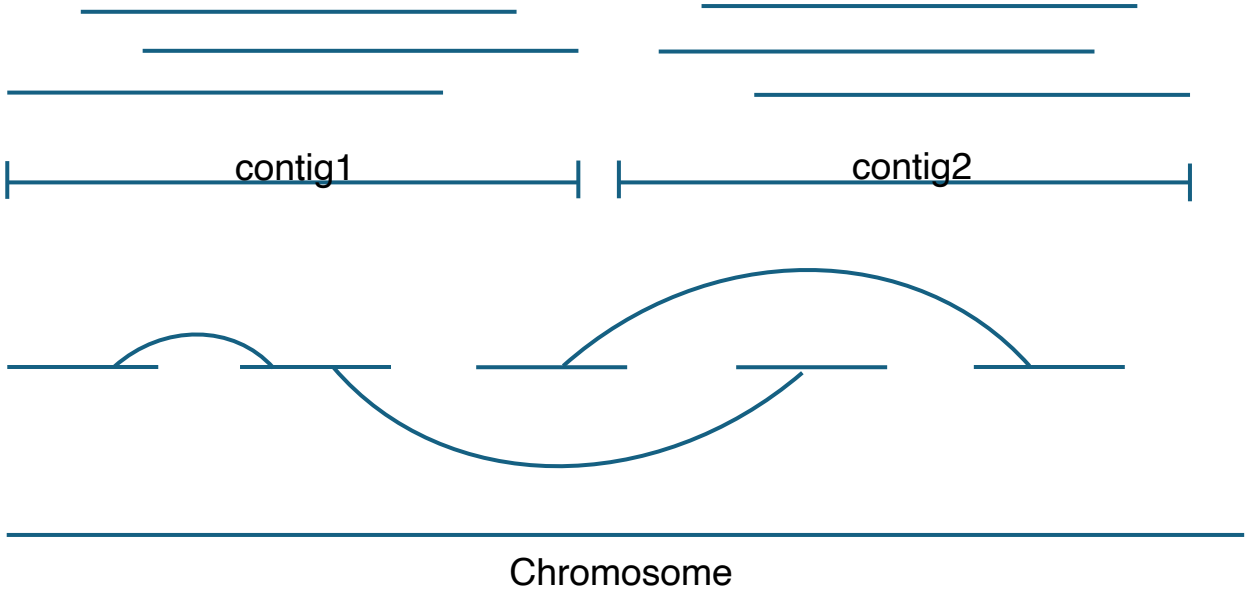
- Validate the contigs by confirming that the reads used to assemble them align correctly.
- Identify any errors in the contig sequences.
- Calculate coverage depth (how many reads cover each position of the contig).

Hi-C data

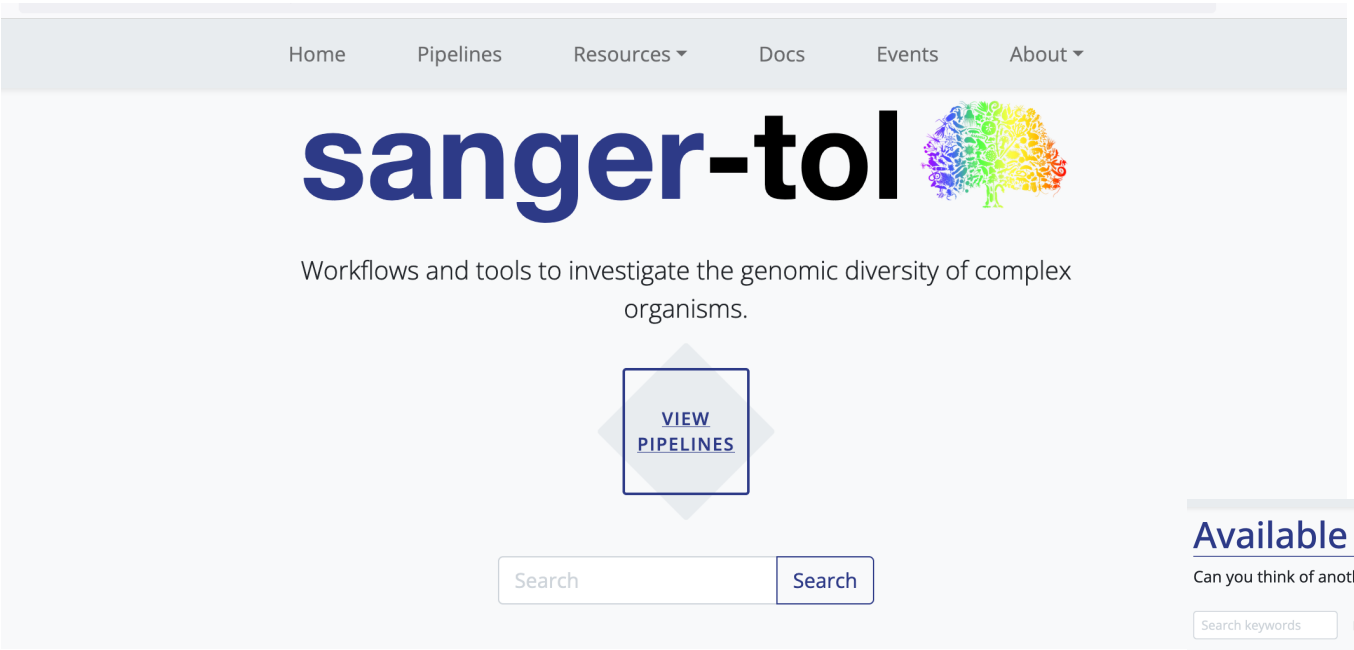
Long reads (PacBio /Nanopore)

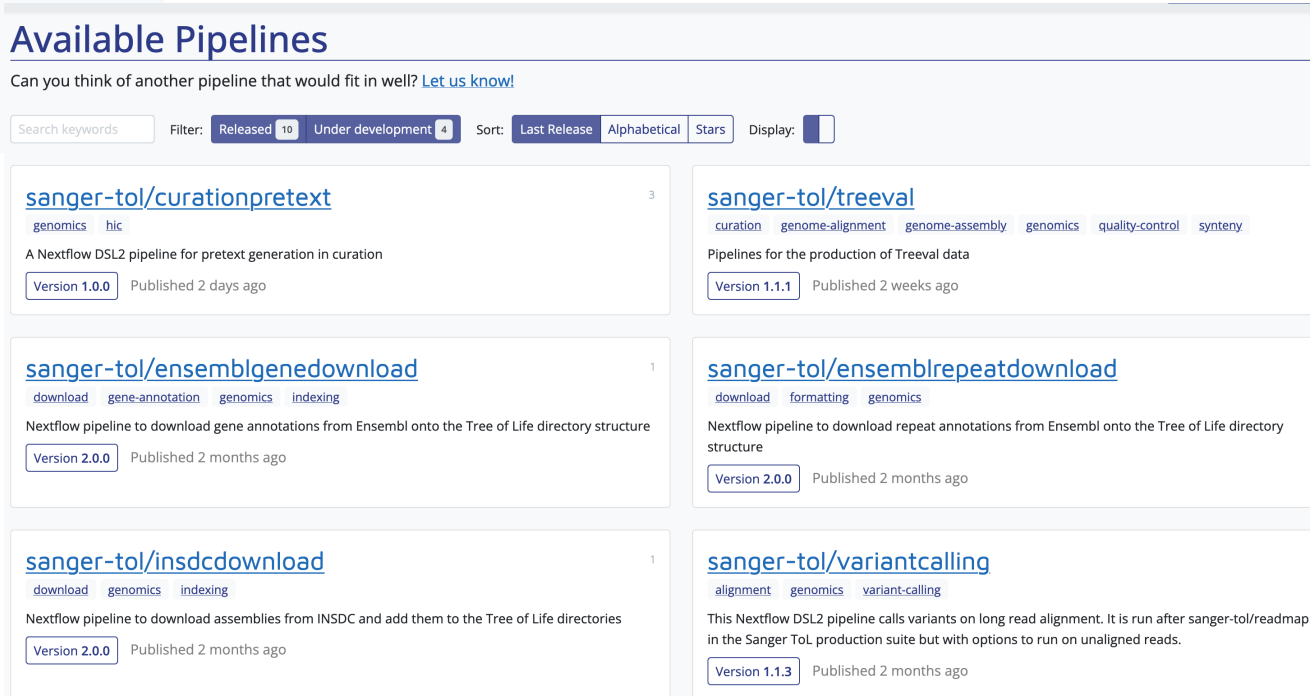
+

Hi-C data



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

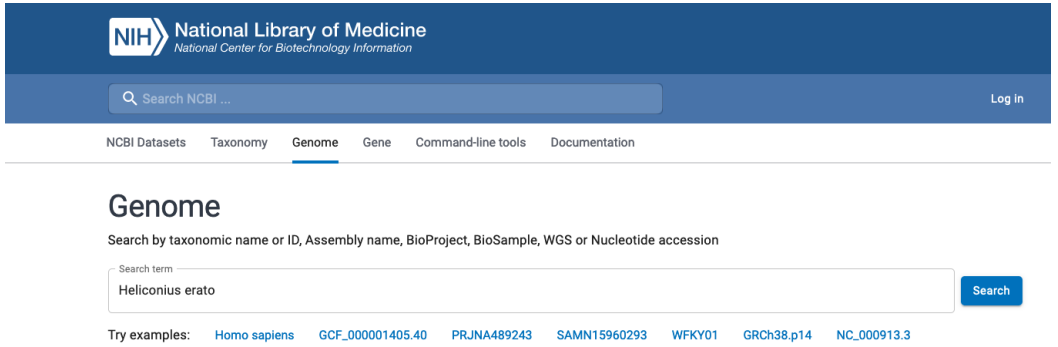




Where can we find genomes?

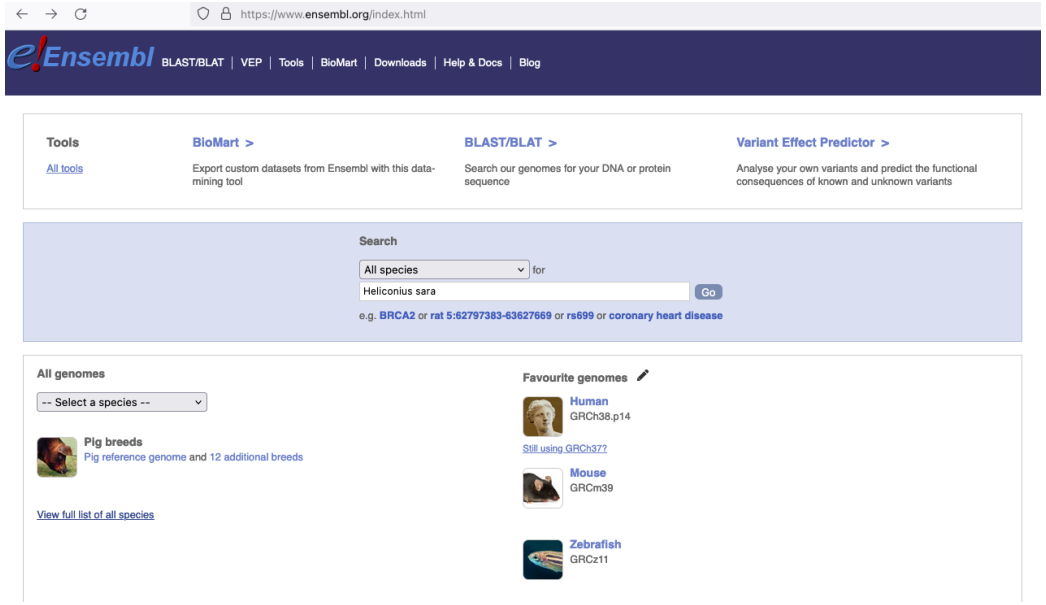
- NCBI (National Center for Biotechnology Information):

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



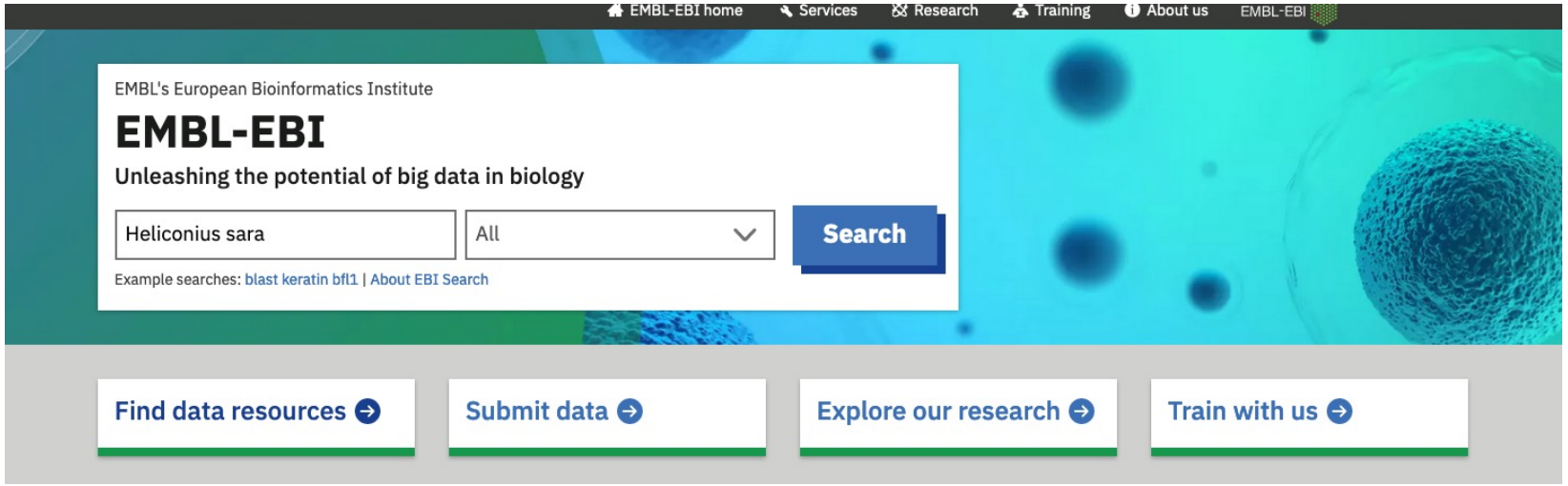
- ensembl

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

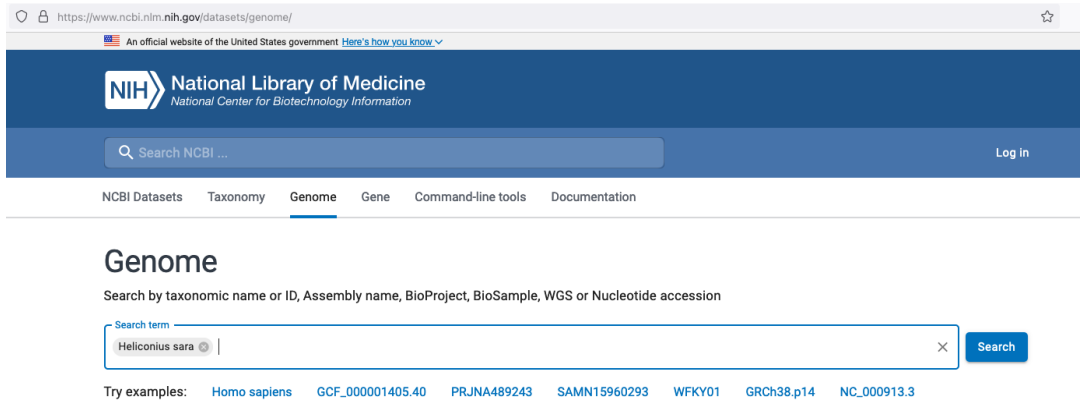


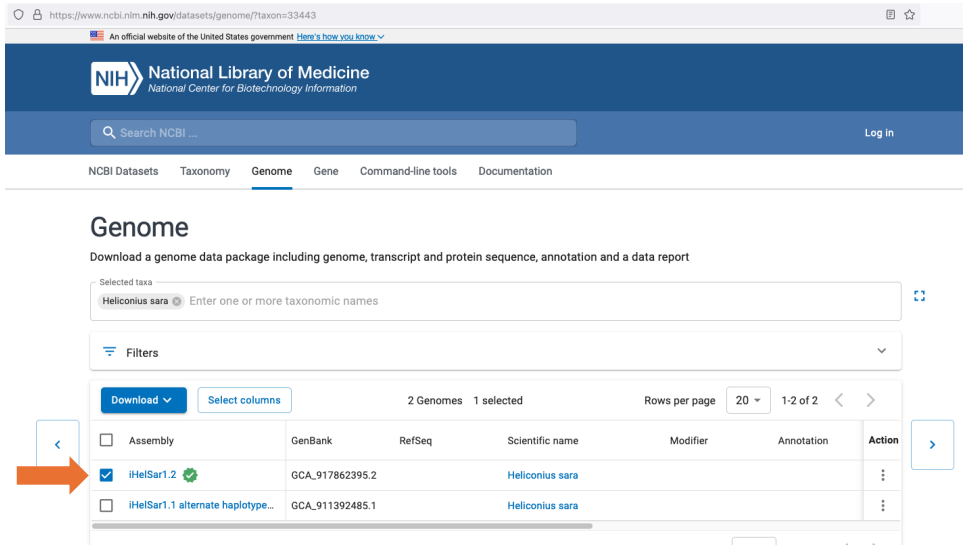
- EMBL-EBI (European Molecular Biology Laboratory - European Bioinformatics Institute):

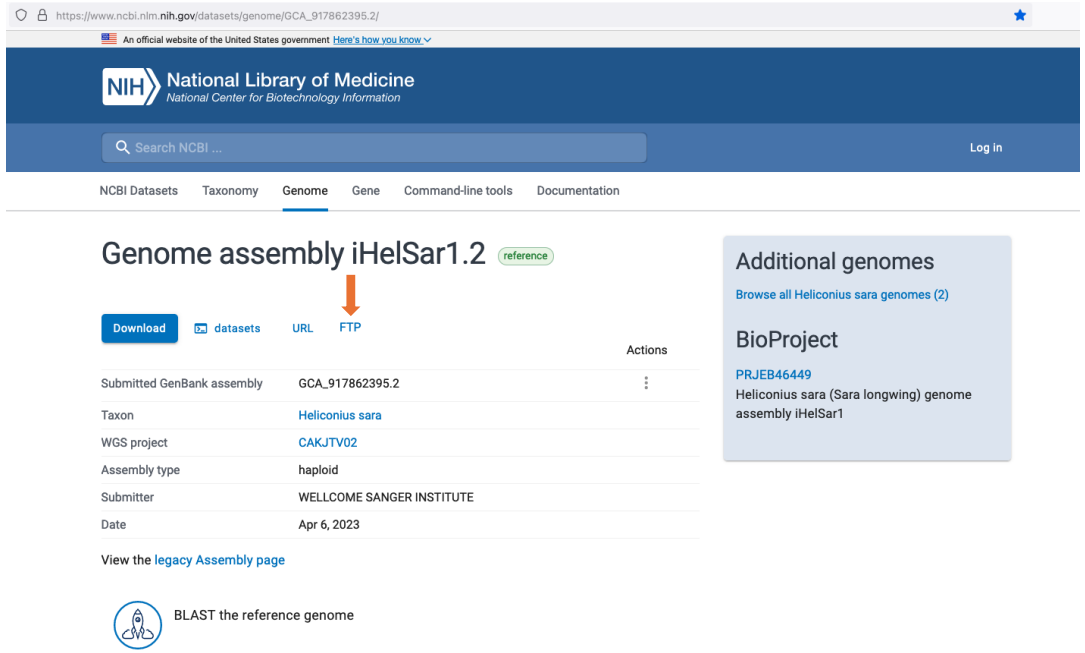
https://www.ebi.ac.uk/

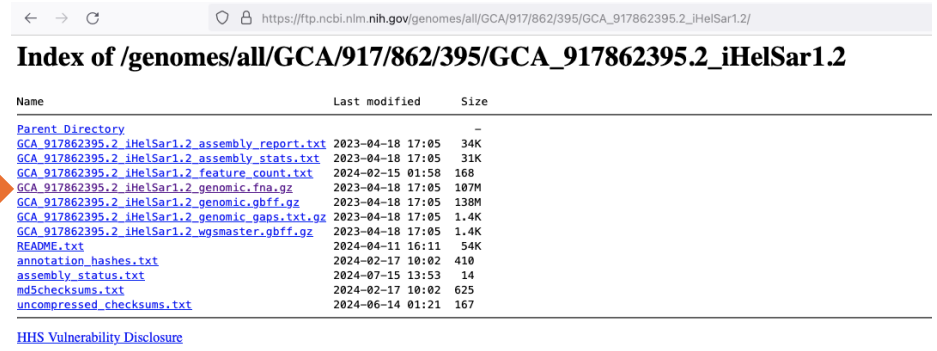


Example:










```
wget
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/917/862/395/GCA_917862395.2_iHelSar1.2/GCA_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)
ACCCTCTTTTCTTATCATTGACATTTAAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCCCGGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCGGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCCAAGTGGCCCCGGGGCTTGATTTTTGCTTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCTT
TGCCGAGTGTGCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACTTCCAGTCCGGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA
GGAGGTGGGGACGCACTTTGCATCCAGACCTCCTCTGCATCGCAGTTCACGACATCCACGCTTGGGAAAG
TCCGTACCCGCGCCTGGAGCGCTTAAAGACACCCTGCCGCGGGTCGGGCGAGGTGCAGCAGAAGTTTCCC
GCGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC
```

Description Line ← Description Line

Sequence Lines ← Sequence Lines

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