



Consiglio Nazionale delle Ricerche

Laboratorio di bioinformatica

Filippo Geraci

filippo.geraci@iit.cnr.it



ISTITUTO
DI INFORMATICA
E TELEMATICA

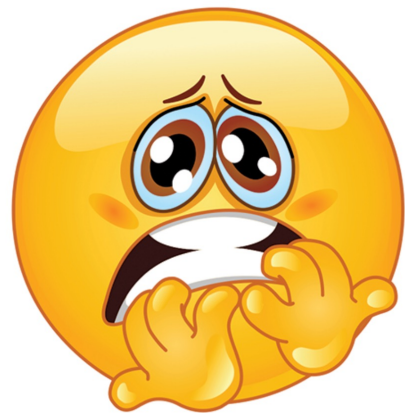
Perché la bioinformatica è difficile?



- Professionisti con competenze molto diverse
- Protocolli complessi anche per attività «semplici»
 - richiedono molte tecnologie contemporaneamente

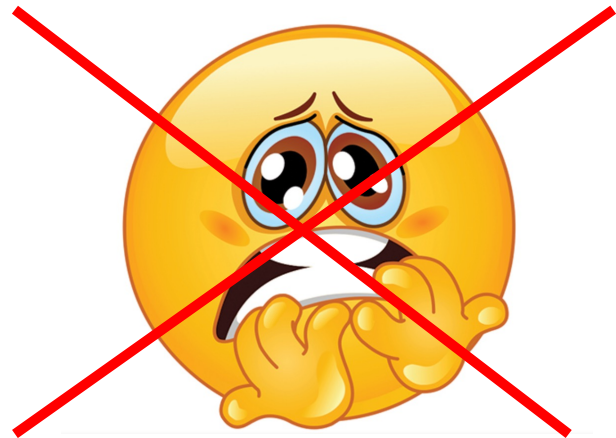
Cosa facciamo oggi

- Scarichiamo il genoma di riferimento del SARS-COV-2 e il sequenziamento di alcuni campioni, poi cerchiamo delle mutazioni

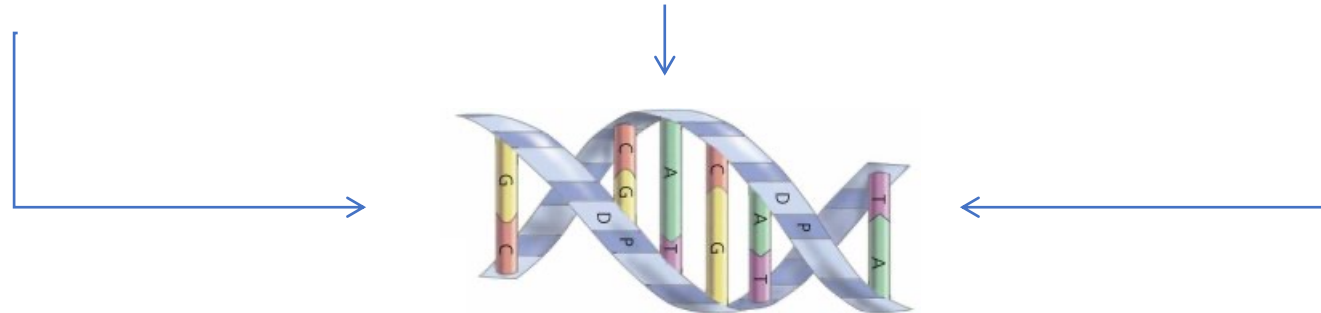


Cosa facciamo oggi

1. Comprendiamo come sono fatti i dati genomici e come vengono analizzati
2. Ci cimentiamo nello scaricare e visualizzare un genoma
3. Scarichiamo un sequenziamento e visualizziamo il suo allineamento

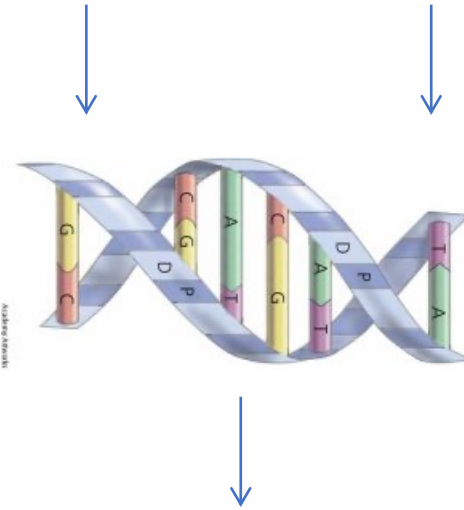
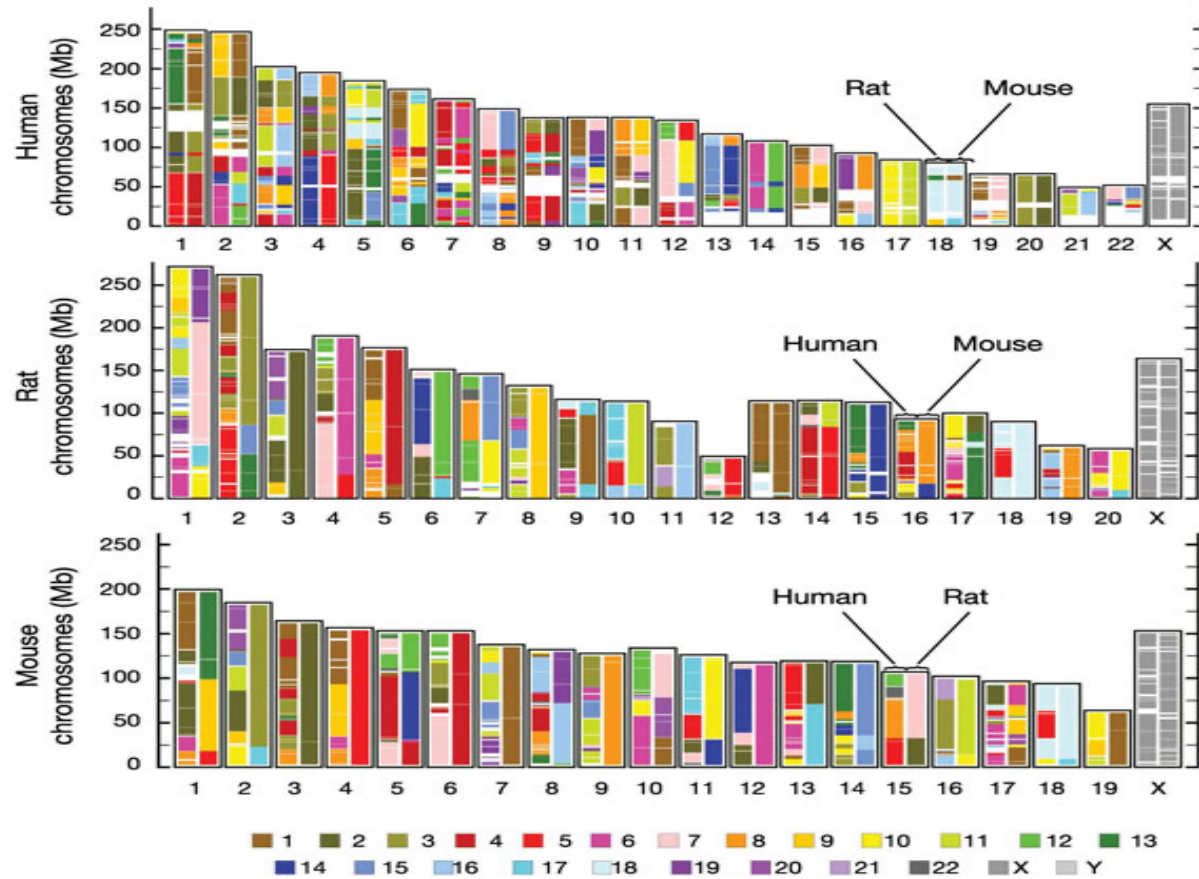


Le varianti genomiche (mutazioni): dal fenotipo al genotipo



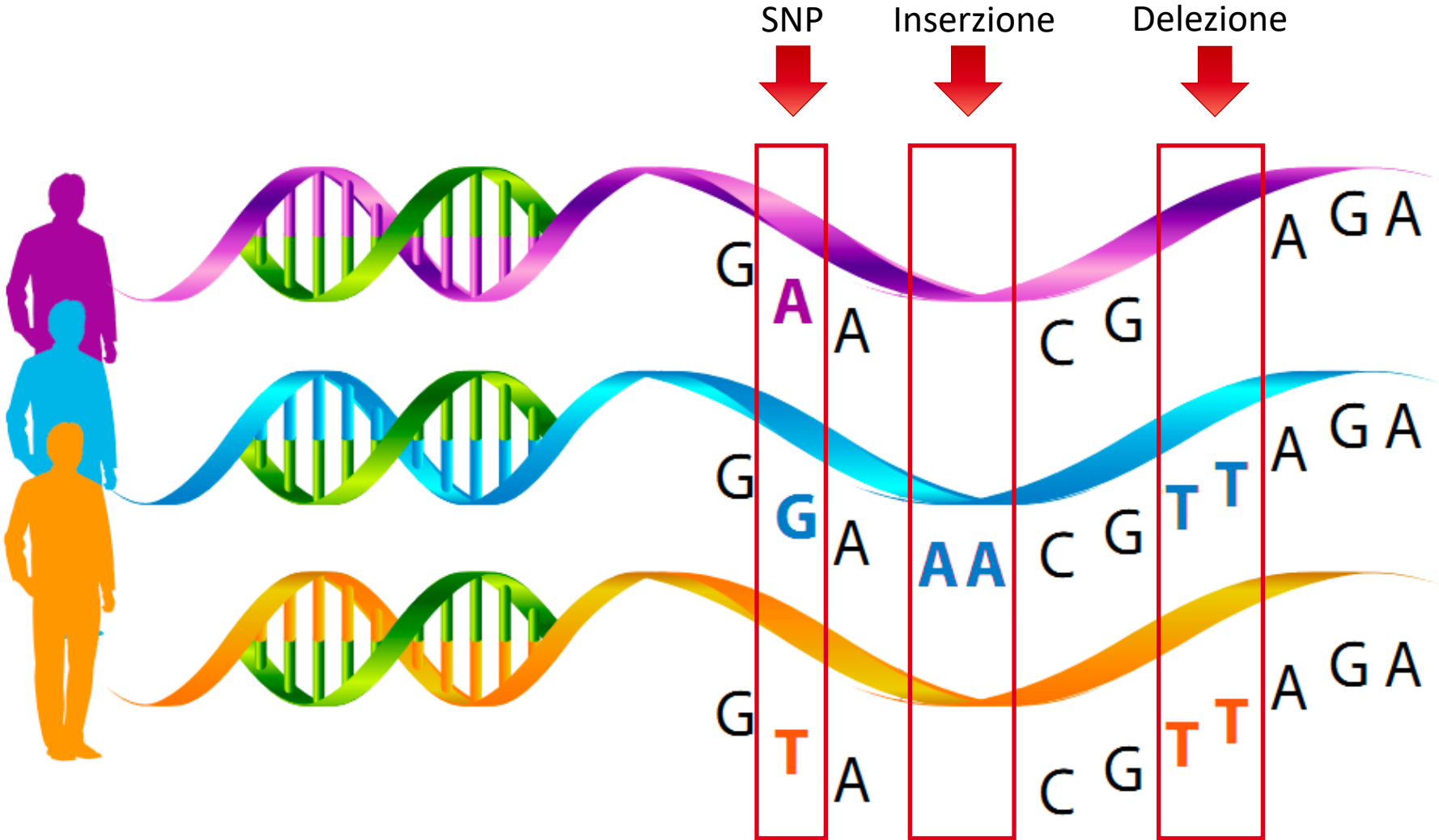
↓
Simili al 99%

Le varianti genomiche: dal fenotipo al genotipo



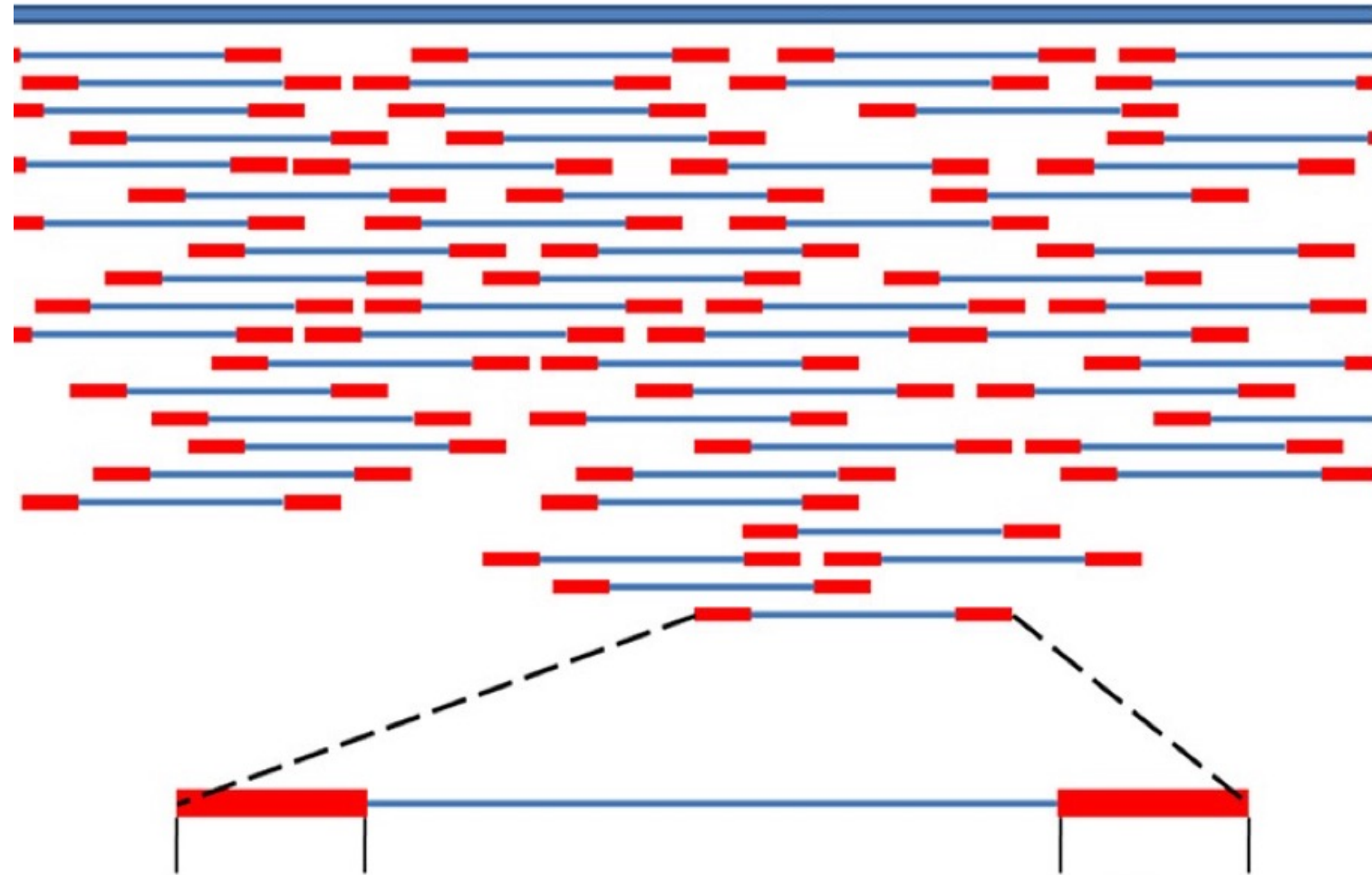
Simili al 98%

Le mutazioni

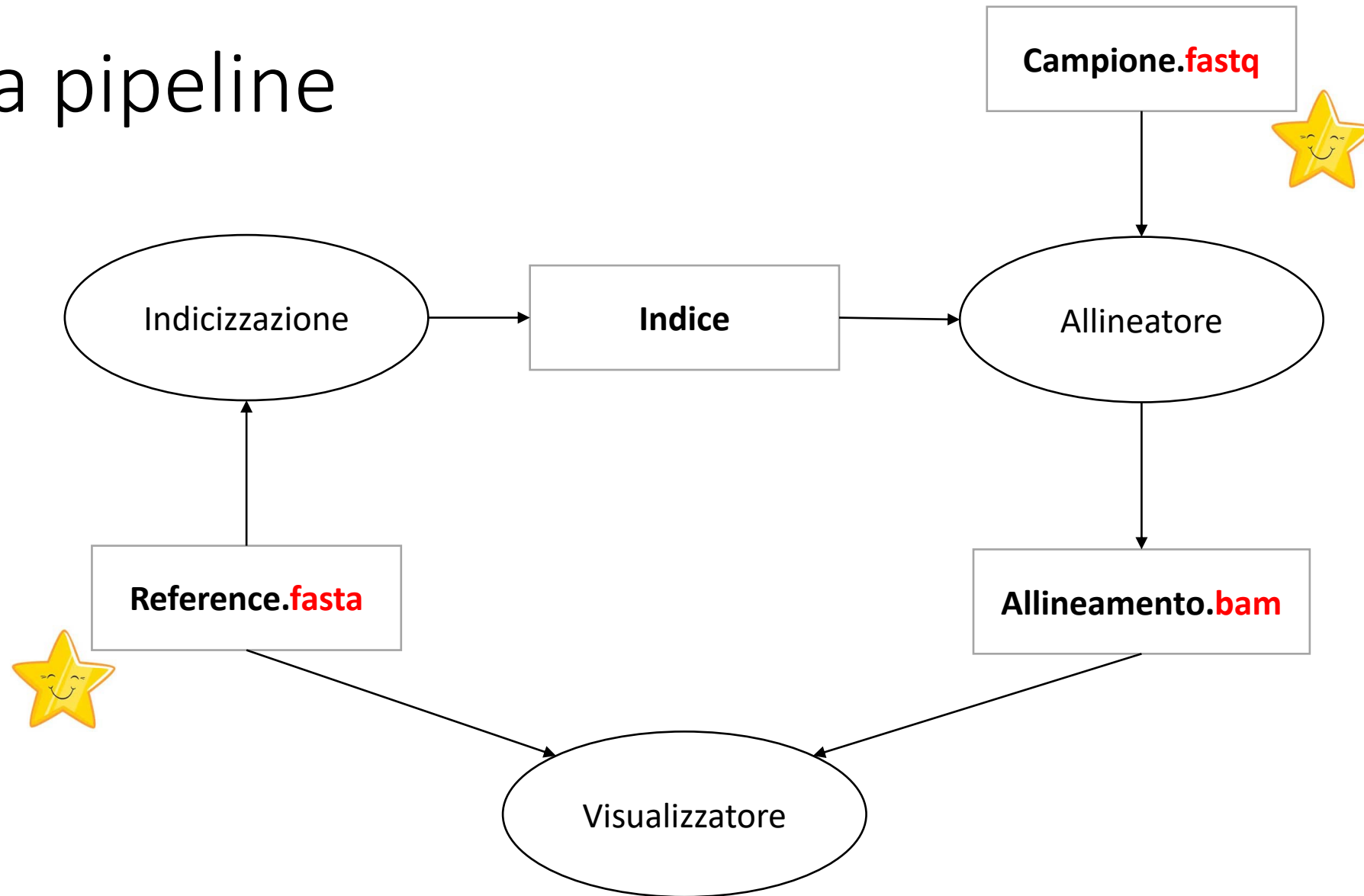


Genoma di riferimento e sequenziamento

Genoma di riferimento



La pipeline



I formati dei file che ci servono

- Fasta: memorizzano una o più sequenze genomiche
- Fastq: memorizzano i frammenti provenienti dal sequenziamento
 - I sequenziatori assegnano un punteggio ad ogni nucleotide in base alla probabilità di errore nella lettura della sequenza
- SAM/BAM: memorizzano i frammenti provenienti dal sequenziamento oltre informazioni sull'allineamento
 - Coordinate genomiche, etc.

fasta

```
>gi|511344723|gb|KE141102.1| Homo sapiens CAGGAGAATCGCTTGAACCCAGGAGGTGGAGGTTGCGGTGA  
GCCAAGATCACACCATTGCAGTCCAGCCTGAGCAACAGAGC  
AAGACTCTCTCTCGAGACAATAAAAACACACAAAAAATTAA  
CTCGCCATGATGGCACAGCCACGTCAGGCGGCACAGACTCC  
GCTCCCCAAGCCTGGCTCCCTTAGAGTTTGGGCTCAAAGGA  
TGTCG  
>gi|511344722|gb|KE141103.1| Homo sapiens GGCTCACTGCATCATGCGCCTCCCTGGTTCAAGCAATTCTG  
ATGCCTGAGCCTCCCAAGTAGGAGATTACAGGTGCACGCC  
ACCACACCCAGCTAATTCTTATATTTTCAGGAGAAATGGGG  
TTTCACCATGTTGCCTACATGCTCTCATTTGAGAAGAGCAT  
GGAGTTAAAGGTGAATGAAAATATTTGTCTCCCTCTCCCTC  
TCCCTCTCCCTCTCCCTCTCCCTCTCCCCACGGTCTCCCTC  
TCATGC
```

FastQ

```
@SRR21643281.1
GCCTCAACTTTGTCAAGCCGTGAAAGGATATCATTTAAAACGTTTTAAATGATATCCTTACACGTCTTGACAAAGTTGAGGC
+
AA6/AEAEAAAE6/E/EAEEE/AEE6A//AE EEEEE/EAAAAA6EEAEAE EEEEEAE/EEAE E//EE/AEE//E<A/E
@SRR21643281.2
CCAGCCCTTGAGACAACTACAGCAACTGGTNNNNNACCAGTTGCTGTAGTTGTCTCAAGGGCTGTNNNNN
+
/AAAAEEAAEAE E/AEEEEEEEEEEEE#####AAAAAE EEEEE EAAEEEE EEA/AEEEE E#####
@SRR21643281.3
ACTGTTGTCCAGCATATCGTAAANNNNNNNNNNNTTTACGATATGCTGGACAACAGTNNNNNNNNNNNN
+
AAAAEEEEEEEEEEEEEEEE#####AAAAAE EEEEE EAAEEEE EEA#####
@SRR21643281.4
GTAGGGCTGTTCAAGTTGAGGCAAACGCCTTTTTCAACTTCTAC
+
/A6AAA/EEEEEEEEEEEEEEEE//EEE/EEEEEEAAEEAAEEE
```

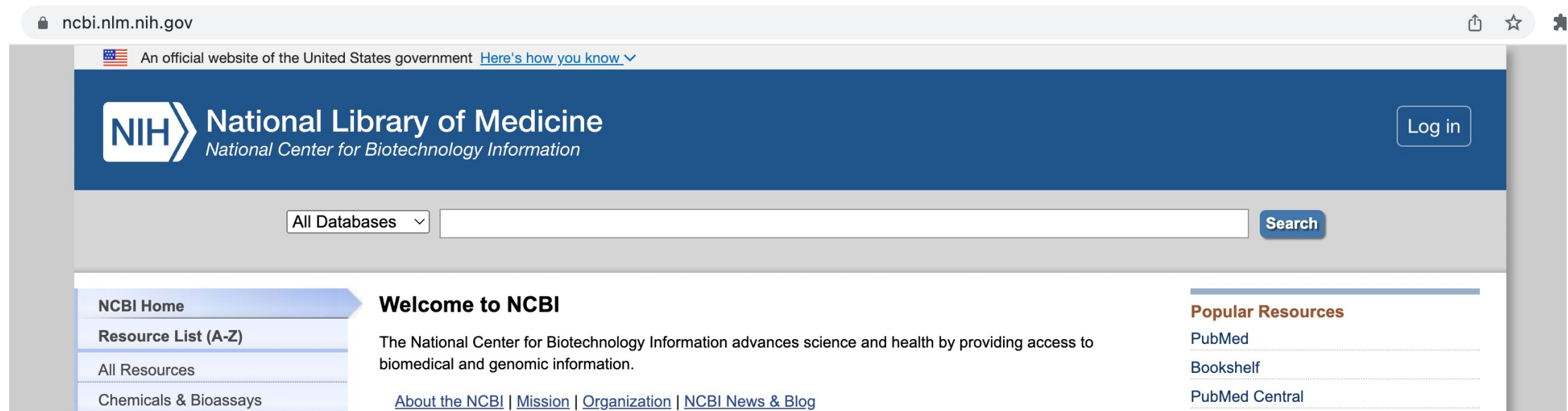
Qualità

- Phred score:

$$Q_{\text{sanger}} = -10 \log_{10} p$$

- dove p è la probabilità che la chiamata sia corretta
- Il Phred score viene codificato in una scala da 0 a 93 usando i caratteri ASCII da 33 a 126
 - !"#\$%&'()*+,-
./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~

Il sito di NCBI... la sorgente di tutti i nostri dati



The image shows a screenshot of the National Center for Biotechnology Information (NCBI) website. The browser address bar shows 'ncbi.nlm.nih.gov'. The page header includes the NIH logo and the text 'National Library of Medicine National Center for Biotechnology Information'. A search bar is visible with a dropdown menu set to 'All Databases' and a 'Search' button. The main content area features a 'Welcome to NCBI' message, a navigation menu on the left, and a 'Popular Resources' section on the right.

ncbi.nlm.nih.gov

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

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

All Databases Search

NCBI Home

- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Popular Resources

- PubMed
- Bookshelf
- PubMed Central

Il sito di NCBI... la sorgente di tutti i nostri dati



Le risorse sono suddivise in database tematici ed hanno un numero identificativo chiamato **Accession number**. La pagina «All Resources» contiene la descrizione di tutti i database.

NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures

All Resources

All Databases Downloads Submissions Tools How To

Databases

[Assembly](#)
A database providing information on the structure of assembled genomes, assembly names and other meta-data, statistical reports, and links to genomic sequence data.

I database di cui abbiamo bisogno

- [Assembly](#) A database providing information on the structure of assembled genomes, assembly names and other meta-data, statistical reports, and links to genomic sequence data.
- [Sequence Read Archive \(SRA\)](#) The Sequence Read Archive (SRA) stores sequencing data from the next generation of sequencing platforms including Roche 454 GS System[®], Illumina Genome Analyzer[®], Life Technologies AB SOLiD System[®], Helicos Biosciences Heliscope[®], Complete Genomics[®], and Pacific Biosciences SMRT[®].

I database di cui abbiamo bisogno

- [Assembly](#)

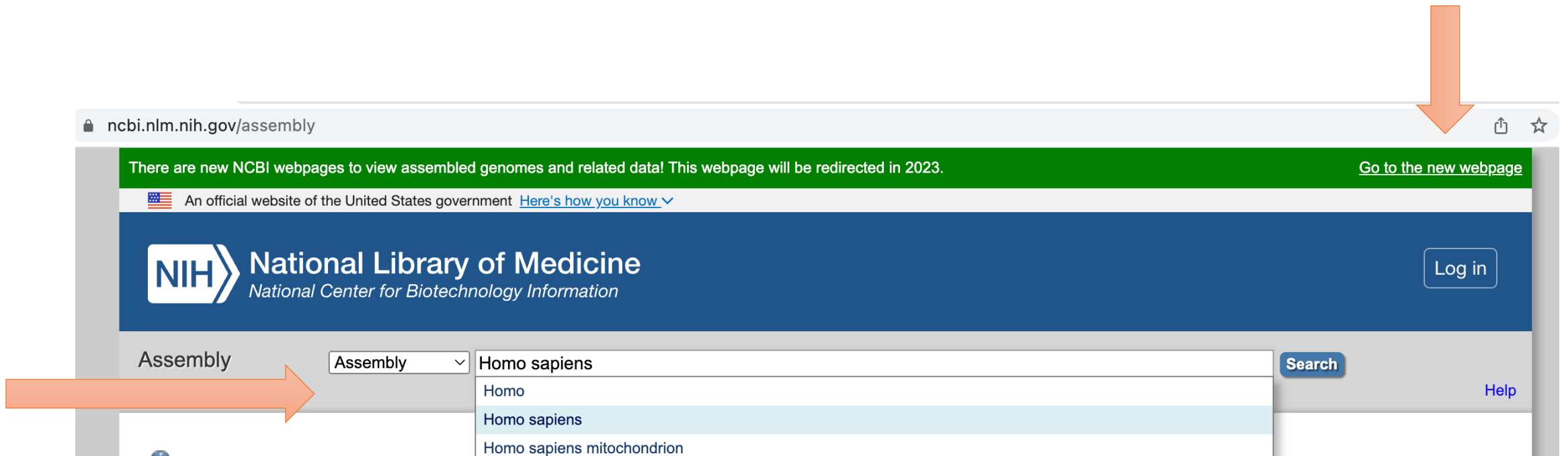
Il database da cui si scarica un genoma **GENERICO ASSEMBLATO** da usare come mappa per ricostruire il genoma dell'individuo che vogliamo studiare

- [Sequence Read Archive \(SRA\)](#)

Il database da cui si scaricano i genomi degli individui che vogliamo studiare (In questo DB i dati provengono da sequenziamenti NGS quindi sono ~~frullati~~ frammentati)

Esempio: il reference del genoma umano

- Il sito di NCBI si sta rinnovando
 - Possiamo ancora fare la ricerca sul vecchio, ma conviene imparare direttamente ad usare il nuovo



The screenshot shows a web browser window with the address bar displaying `ncbi.nlm.nih.gov/assembly`. A green banner at the top contains the text: "There are new NCBI webpages to view assembled genomes and related data! This webpage will be redirected in 2023." with a link "Go to the new webpage". Below the banner is a blue header for the "National Library of Medicine" with the NIH logo and a "Log in" button. The main content area features a search interface with a dropdown menu set to "Assembly" and a search button. The dropdown menu is open, showing a list of options: "Homo sapiens", "Homo", "Homo sapiens", and "Homo sapiens mitochondrion". An orange arrow points to the search button, and another orange arrow points to the "Homo sapiens" option in the dropdown menu.

ncbi.nlm.nih.gov/assembly

There are new NCBI webpages to view assembled genomes and related data! This webpage will be redirected in 2023. [Go to the new webpage](#)

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

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

Assembly Assembly Homo sapiens

Homo

Homo sapiens

Homo sapiens mitochondrion

Search

Help

Esempio: il reference del genoma umano

The image shows a screenshot of the National Library of Medicine (NIH) website, specifically the NCBI Genome section. The header features the NIH logo and the text "National Library of Medicine National Center for Biotechnology Information". Below the header is a search bar with the placeholder text "Search NCBI ..." and a "Log in" button. The main navigation menu includes "Datasets", "Taxonomy", "Genome", "Gene", "Command-line tools", and "Documentation". The "Genome" tab is currently selected. The main content area displays the word "Genome" in a large font, followed by a description: "Download a genome data package including genome, transcript and protein sequence, annotation and a data report". A search input field contains the text "homo", and a dropdown menu is open below it, showing two options: "Homo sapiens (human)" and "Homo (humans)". A yellow "BETA" badge is visible in the top right corner of the content area.

NIH National Library of Medicine
National Center for Biotechnology Information

Search NCBI ... Log in

Datasets Taxonomy **Genome** Gene Command-line tools Documentation

Genome

BETA

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

Selected taxa

homo

- Homo sapiens (human)
- Homo (humans)

Esempio: il reference del genoma umano




Download ▾ [Select columns](#) 1.083 genomes Rows per page 20 ▾ 1-20 of 1.083 < >

| <input type="checkbox"/> | Assembly | GenBank | Scientific name | Modifier | Size (Mb) | Level | Date | Action |
|--------------------------|----------------------------------|------------------|--------------------------------------|----------------|-----------|------------|-----------|--------|
| <input type="checkbox"/> | GRCh38.p14 ✓ | GCA_000001405.29 | Homo sapiens (human) | | 3,099 | Chromosome | Mar, 2022 | ⋮ |
| <input type="checkbox"/> | HuRef | GCA_000002125.2 | Homo sapiens (human) | | 2,844 | Chromosome | Oct, 2007 | ⋮ |
| <input type="checkbox"/> | CHM1_1.1 | GCA_000306695.2 | Homo sapiens (human) | CHM1 (isolate) | 3,038 | Chromosome | Jul, 2013 | ⋮ |
| <input type="checkbox"/> | T2T-CHM13v2.0 | GCA_009914755.4 | Homo sapiens (human) | | 3,117 | Complete | Feb, 2022 | ⋮ |
| <input type="checkbox"/> | WGS ⚠ | GCA_000002115.2 | Homo sapiens (human) | | 2,864 | Chromosome | Mar, 2004 | ⋮ |
| <input type="checkbox"/> | CRA_TCAGchr7v2 ⚠ | GCA_000002135.3 | Homo sapiens (human) | | 158.3 | Chromosome | Oct, 2004 | ⋮ |

Esempio: il reference del genoma umano

Reference

Download ▾ Select columns 1.083 genomes Rows per page 20 ▾ 1-20 of 1.083 < >

| <input type="checkbox"/> | Assembly | GenBank | Scientific name | Modifier | Size (Mb) | Level | Date | Action |
|--------------------------|--|------------------|--------------------------------------|----------------|-----------|------------|-----------|--------|
| <input type="checkbox"/> | GRCh38.p14  | GCA_000001405.29 | Homo sapiens (human) | | 3,099 | Chromosome | Mar, 2022 | ⋮ |
| <input type="checkbox"/> | HuRef | GCA_000002125.2 | Homo sapiens (human) | | 2,844 | Chromosome | Oct, 2007 | ⋮ |
| <input type="checkbox"/> | CHM1_1.1 | GCA_000306695.2 | Homo sapiens (human) | CHM1 (isolate) | 3,038 | Chromosome | Jul, 2013 | ⋮ |
| <input type="checkbox"/> | T2T-CHM13v2.0 | GCA_009914755.4 | Homo sapiens (human) | | 3,117 | Complete | Feb, 2022 | ⋮ |
| <input type="checkbox"/> | WGS  | GCA_000002115.2 | Homo sapiens (human) | | 2,864 | Chromosome | Mar, 2004 | ⋮ |
| <input type="checkbox"/> | CRA_TCAGchr7v2  | GCA_000002135.3 | Homo sapiens (human) | | 158.3 | Chromosome | Oct, 2004 | ⋮ |

Altri genomi non reference

- La disponibilità di altri genomi assemblati consente di fare studi di popolazione o di usare come riferimento individui piu' simili a quello che voglio studiare (esempio della stessa etnia)
 - **HuRef** – J. Craig Venter Institute (Genoma di J. C. Venter)
 - **YH2** - Beijing Genomics Institute (Cinese)
 - **CHM1_1.1** - Washington University School of Medicine
 - **BGIAF** - Beijing Genomics Institute (Africano Yoruba)
 - **Watson-partial** - Baylor College of Medicine (Genoma di J. Watson)
 - **WGSA** - Celera Genomics
 - **KOREF 1.0** – Genome research foundation (Coreano)
 - **MSB_human_1b** - University of Luebeck (Egyptian)

Esempio: il reference del genoma umano

Livello di assemblaggio

Download ▾ Select columns

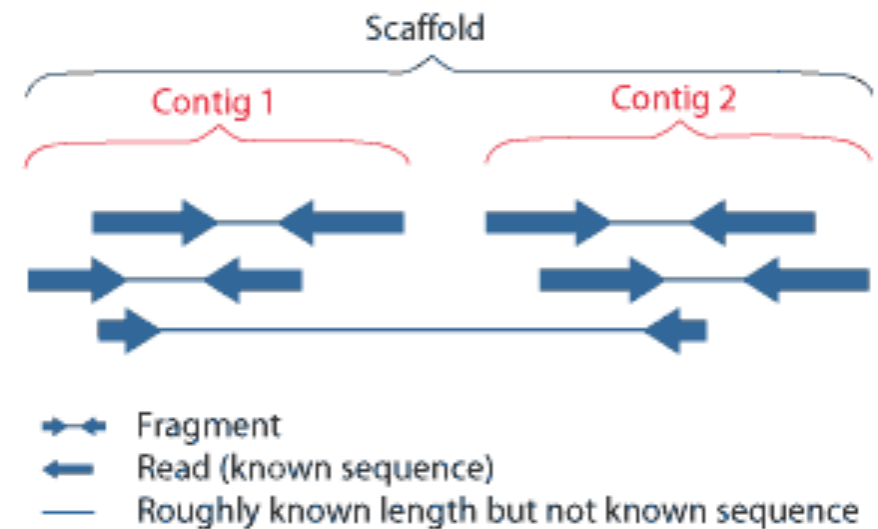
1.083 genomes

Rows per page 20 ▾ 1-20 of 1.083 < >

| <input type="checkbox"/> | Assembly | GenBank | Scientific name | Modifier | Size (Mb) | Level | Date | Action |
|--------------------------|----------------------------------|------------------|--------------------------------------|----------------|-----------|------------|-----------|--------|
| <input type="checkbox"/> | GRCh38.p14 ✓ | GCA_000001405.29 | Homo sapiens (human) | | 3,099 | Chromosome | Mar, 2022 | ⋮ |
| <input type="checkbox"/> | HuRef | GCA_000002125.2 | Homo sapiens (human) | | 2,844 | Chromosome | Oct, 2007 | ⋮ |
| <input type="checkbox"/> | CHM1_1.1 | GCA_000306695.2 | Homo sapiens (human) | CHM1 (isolate) | 3,038 | Chromosome | Jul, 2013 | ⋮ |
| <input type="checkbox"/> | T2T-CHM13v2.0 | GCA_009914755.4 | Homo sapiens (human) | | 3,117 | Complete | Feb, 2022 | ⋮ |
| <input type="checkbox"/> | WGS ⚠ | GCA_000002115.2 | Homo sapiens (human) | | 2,864 | Chromosome | Mar, 2004 | ⋮ |
| <input type="checkbox"/> | CRA_TCAGchr7v2 ⚠ | GCA_000002135.3 | Homo sapiens (human) | | 158.3 | Chromosome | Oct, 2004 | ⋮ |

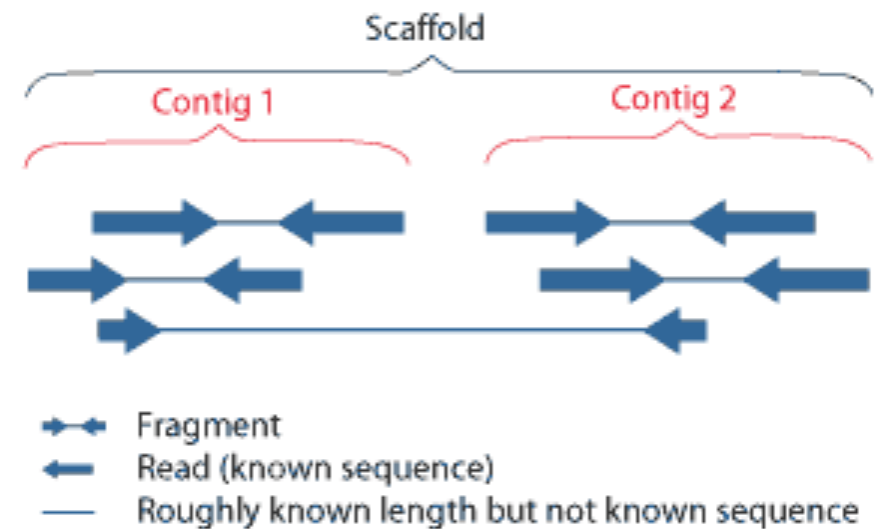
Assembly level

- **Complete genome** - tutti i cromosomi sono senza buchi e non hanno sequenze ambigue più lunghe di 10 nucleotidi.
- **Chromosome** - è presente la sequenza di uno o più cromosomi, ma possono esserci dei buchi.
- **Scaffold** - l'assemblato contiene per lo più frammenti di «grandi» dimensioni
- **Contig** - l'assemblato contiene solo frammenti di piccole dimensioni



Assembly level

- ➔ • **Complete genome** - tutti i cromosomi sono senza buchi e non hanno sequenze ambigue più lunghe di 10 nucleotidi.
- ➔ • **Chromosome** - è presente la sequenza di uno o più cromosomi, ma possono esserci dei buchi.
- **Scaffold** - l'assemblato contiene per lo più frammenti di «grandi» dimensioni
- **Contig** - l'assemblato contiene solo frammenti di piccole dimensioni



Esempio: il reference del genoma umano

[Datasets](#) / [Genome](#) / GRCh38.p14



Genome assembly GRCh38.p14

reference

Download

datasets

curl

| | | |
|--------------------|-----------------------------|---|
| Reference sequence | RefSeq GCF_000001405.40 | ⋮ |
| Submitted sequence | GenBank GCA_000001405.29 | ⋮ |
| Taxon | <i>Homo sapiens</i> (human) | |
| Synonym | hg38 | |
| Assembly type | haploid-with-alt-loci | |
| Submitter | Genome Reference Consortium | |
| Date | Feb 3, 2022 | |

View the [legacy Assembly page](#)

Esempio: il reference del genoma umano

Datasets / Genome / GRCh38.p14



Genome assembly GRCh38.p14

reference

Opzioni di download

Download

datasets

curl

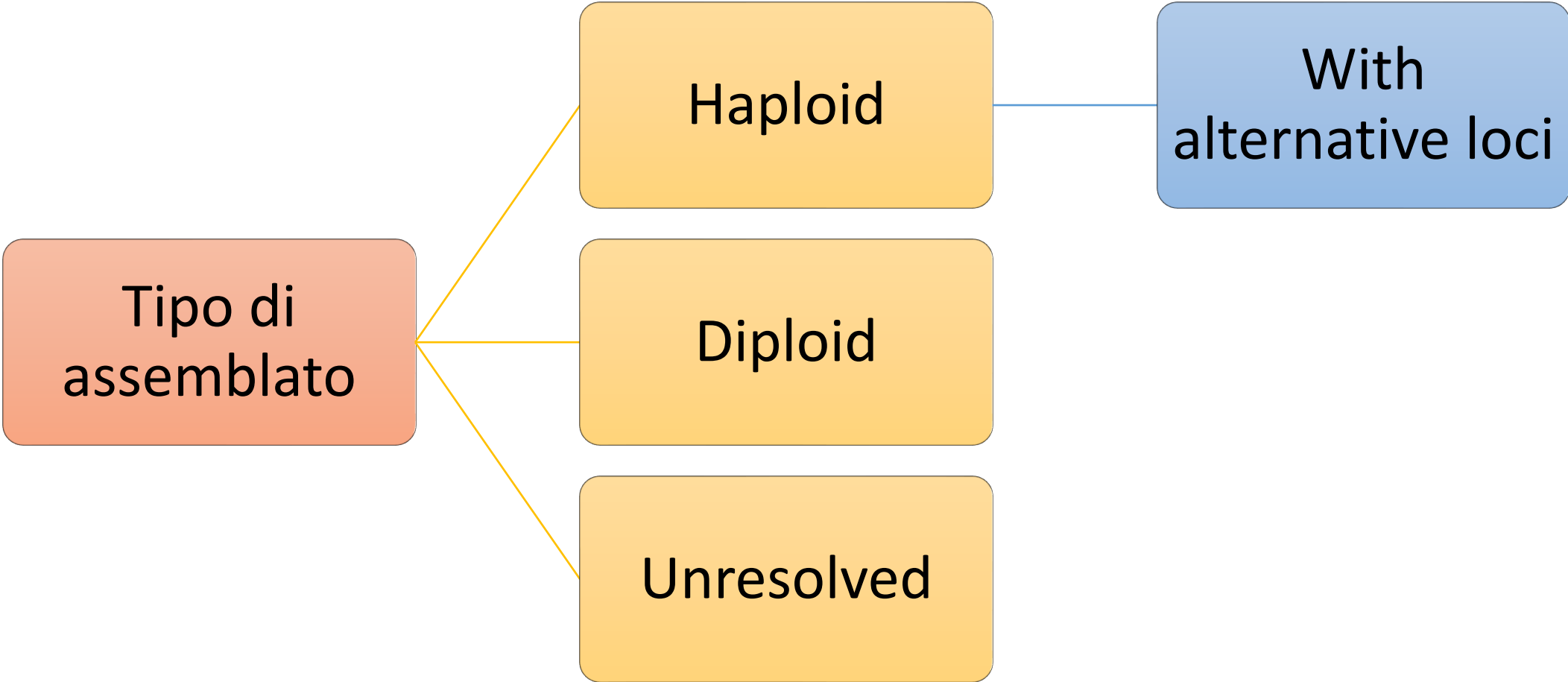
| | |
|--------------------|-----------------------------|
| Reference sequence | RefSeq GCF_000001405.40 |
| Submitted sequence | GenBank GCA_000001405.29 |
| Taxon | <i>Homo sapiens</i> (human) |
| Synonym | hg38 |
| Assembly type | haploid-with-alt-loci |
| Submitter | Genome Reference Consortium |
| Date | Feb 3, 2022 |

View the [legacy Assembly page](#)

Accession ID

Tipo di
assemblato

Tipo di assemblato



Esempio: il reference del genoma umano

Assembly statistics



These statistics describe the RefSeq genome sequence GCF_000001405.40

| | |
|-----------------------|------------|
| Genome size | 3.1 Gb |
| Number of chromosomes | 24 |
| Number of scaffolds | 470 |
| Scaffold N50 | 67.8 Mb |
| Scaffold L50 | 16 |
| Number of contigs | 996 |
| Contig N50 | 57.9 Mb |
| Contig L50 | 18 |
| GC percent | 40.5 |
| Assembly level | Chromosome |

Esempio: il reference del genoma umano

Assembly statistics

These statistics describe the RefSeq genome sequence GCF_000001405.40

| | | |
|-----------------------|------------|---|
| Genome size | 3.1 Gb |  |
| Number of chromosomes | 24 | |
| Number of scaffolds | 470 | |
| Scaffold N50 | 67.8 Mb | |
| Scaffold L50 | 16 | |
| Number of contigs | 996 | |
| Contig N50 | 57.9 Mb | |
| Contig L50 | 18 | |
| GC percent | 40.5 | |
| Assembly level | Chromosome |  |

Esempio: il reference del genoma umano

Chromosomes

| Chromosome | GenBank | RefSeq | Size (bp) | GC content (%) |
|------------|----------------------------|------------------------------|-------------|----------------|
| 1 | CM000663.2 | NC_000001.11 | 248.956.422 | 41,5 |
| 2 | CM000664.2 | NC_000002.12 | 242.193.529 | 40 |
| 3 | CM000665.2 | NC_000003.12 | 198.295.559 | 39,5 |
| 4 | CM000666.2 | NC_000004.12 | 190.214.555 | 38 |
| 5 | CM000667.2 | NC_000005.10 | 181.538.259 | 39 |
| 6 | CM000668.2 | NC_000006.12 | 170.805.979 | 39,5 |
| 7 | CM000669.2 | NC_000007.14 | 159.345.973 | 40,5 |

Esempio: il reference del genoma umano

| Chromosomes | Accession ID | | Dimensione | |
|-------------|----------------------------|------------------------------|-------------|-----------|
| | Chromosome | GenBank | RefSeq | Size (bp) |
| 1 | CM000663.2 | NC_000001.11 | 248.956.422 | 41,5 |
| 2 | CM000664.2 | NC_000002.12 | 242.193.529 | 40 |
| 3 | CM000665.2 | NC_000003.12 | 198.295.559 | 39,5 |
| 4 | CM000666.2 | NC_000004.12 | 190.214.555 | 38 |
| 5 | CM000667.2 | NC_000005.10 | 181.538.259 | 39 |
| 6 | CM000668.2 | NC_000006.12 | 170.805.979 | 39,5 |
| 7 | CM000669.2 | NC_000007.14 | 159.345.973 | 40,5 |

Esempio: il reference del genoma umano

Revision History

| GenBank | RefSeq | Name | Level | Date | Action |
|----------------------------------|----------------------------------|------------|------------|------------|--------|
| GCA_000001405.29 | GCF_000001405.40 | GRCh38.p14 | Chromosome | 2022-02-03 | ⋮ |
| GCA_000001405.28 | GCF_000001405.39 | GRCh38.p13 | Chromosome | 2019-02-28 | ⋮ |
| GCA_000001405.27 | GCF_000001405.38 | GRCh38.p12 | Chromosome | 2017-12-21 | ⋮ |
| GCA_000001405.26 | GCF_000001405.37 | GRCh38.p11 | Chromosome | 2017-06-14 | ⋮ |

Esempio: il reference del genoma umano

Download Package

Download a data package for GCF_000001405.40

Select file types

- Genomic sequence, (FASTA)
- Annotated features (GTF)
- Annotated features (GFF3)
- Sequence and annotation (GBFF)
- Transcripts (FASTA)
- Genomic CDS (FASTA)
- Proteins (FASTA)


Your selected data will be downloaded as a ZIP archive
Estimated file size is 794 MB

Name your file

[Cancel](#) [Download](#)

Esempio: il reference del genoma umano

```
[geraci]:viola [pc-2022] -> ls
GCF_000001405.40.zip
[geraci]:viola [pc-2022] -> unzip GCF_000001405.40.zip
Archive:  GCF_000001405.40.zip
  inflating: README.md
  inflating: ncbi_dataset/data/data_summary.tsv
  inflating: ncbi_dataset/data/assembly_data_report.jsonl
  inflating: ncbi_dataset/data/GCF_000001405.40/GCF_000001405.40_GRCh38.p14_genomic.fna
  inflating: ncbi_dataset/data/GCF_000001405.40/sequence_report.jsonl
  inflating: ncbi_dataset/data/dataset_catalog.json
[geraci]:viola [pc-2022] -> █
```



Domande

- In che formato sarà il file?
- Come lo apro?

Domande

- In che formato sarà il file?
 - Trattandosi di sequenze di DNA assemblate mi aspetto un file in formato (multi)fasta
- Come lo apro?
 - Mi serve un editor di testi, non posso usare un word processor



- Editor grafico adeguato
- Prompt dei comandi
 - Il comando «more» è disponibile sia su Windows che su MacOS che su Linux

Visualizzare il reference

ncbi.nlm.nih.gov/labs/data-hub/genome/GCF_000001405.40/

[Datasets](#) / [Genome](#) / GRCh38.p14



Genome assembly GRCh38.p14

reference

Download

datasets

curl

| | |
|--------------------|--------------------------------------|
| Reference sequence | RefSeq GCF_000001405.40 |
| Submitted sequence | GenBank GCA_000001405.29 |
| Taxon | Homo sapiens (human) |
| Synonym | hg38 |
| Assembly type | haploid-with-alt-loci |
| Submitter | Genome Reference Consortium |
| Date | Feb 3, 2022 |

View the [leacv Assembly page](#)

Additional

[Browse all Homo sa](#)

BioProject

[PRJNA31257](#)

The Human Genom

the
(RC)

ior

Genome Biol 2008

[Finishing the finish
22 sequence](#)

CG Cole, et al.

- Download RefSeq
- BLAST against this genome
- [See in Genome Data Viewer](#)
- See more files on FTP

Visualizzare il reference

The image shows the National Library of Medicine (NIH) Genome Data Viewer interface. The header includes the NIH logo and the text "National Library of Medicine National Center for Biotechnology Information". Below this, the page is titled "Genome Data Viewer".

The main content area is for "Homo sapiens (human)". It shows the assembly "GRCh38.p14 (GCF_000001405.40)" and the chromosome "Chr 1 (NC_000001.11)". A search bar contains "BRCA1".

The interface is divided into several sections:

- Assemblies:** A dropdown menu shows "GCF_000001405.40 (GRCh38.p14)".
- Ideogram View:** A sidebar shows a karyotype of human chromosomes, with chromosome 1 highlighted in green.
- Genomic Track:** The main area displays a genomic track for chromosome 1. It includes:
 - NC_000001.11:** A track showing the reference sequence with a scale from 0 to 140 Mb.
 - Genes, NCBI Homo sapiens Annotation Release 110, 2022-04-08:** A track showing gene models for CAMTA1, KAZN, A6BL4, DAB1, NEGR1, and DPYD.
 - Genes, Ensembl release 108:** A track showing Ensembl gene models with IDs like ENSG00000171735, ENSG00000189337, ENSG00000186094, ENSG00000172260, ENSG00000173406, ENSG00000188641, and ENSG00000237505.
 - Cited Variations, dbSNP b155 v2:** A track showing SNP positions.
 - Live RefSNPs, dbSNP b155 v2:** A track showing live reference SNP positions.
 - RNA-seq exon coverage, aggregate (filtered), NCBI Homo sapiens Annotation Release 110 - log base:** A track showing exon coverage with a scale from 0 to 8192.
 - RNA-seq intron-spanning reads, aggregate (filtered), NCBI Homo sapiens Annotation Release 110 - log base:** A track showing intron-spanning reads with a scale from 0 to 2048.

Visualizzare il reference

The screenshot displays the National Library of Medicine (NIH) National Center for Biotechnology Information (NCBI) Genome Data Viewer. The interface is for **Homo sapiens** (human) and shows the assembly **GRCh38.p14 (GCF_000001405.40)** for **Chr 1 (NC_000001.11)**. The search assembly is **BRCA1**. The main view shows the **NC_000001.11: 1 - 248,956,422** region. The interface includes a search bar, assembly selection, ideogram view, and various genomic tracks such as Genes (NCBI and Ensembl), Cited Variations (dbSNP), and RNA-seq data. A warning message states: "Exon Navigator: There are too many (5485) genes in this region. Please narrow the region to enable exon navigation." The ideogram view shows chromosomes 1 through 14, with chromosome 1 highlighted.

E ora?



Esercizio

- Proviamo a scaricare il genoma del SARS-COV-2 (severe acute respiratory syndrome coronavirus 2)

Suggerimento

ncbi.nlm.nih.gov/data-hub/genome/GCF_009858895.2/

Datasets Taxonomy **Genome** Gene Command-line tools Documentation

Datasets / Genome / ASM985889v3

Genome assembly ASM985889v3 reference

[Download](#) [datasets](#) [curl](#)

| | | |
|--------------------|---|---|
| Reference sequence | RefSeq GCF_009858895.2 | ⋮ |
| Submitted sequence | GenBank GCA_009858895.3 | ⋮ |
| Taxon | Severe acute respiratory syndrome coronavirus 2 | |
| Assembly type | haploid | |

Scarichiamo un campione da SRA

SRA SARS-CoV-2 wuhan
[Create alert](#) [Advanced](#)

Access
Public (5,596)

Source
DNA (271)
RNA (374)

Type
genome (25)

Library Layout
paired (308)
single (5,288)

Platform
Illumina (5,487)
Oxford Nanopore (101)
PacBio SMRT (8)

Strategy
Exome (58)
Genome (25)
other (5,513)

Summary ▾ 20 per page ▾ [Send to:](#)

View results as an expanded interactive table using the RunSelector. [Send results to Run selector](#)

Search results
Items: 1 to 20 of 5596 [<< First](#) [< Prev](#) Page of 280 [Next >](#) [Last >](#)

- [Wuhan-Hu-1_3](#)
1. 1 ILLUMINA (NextSeq 500) run: 771,903 spots, 78.9M bases, 33.9Mb downloads
Accession: SRX17643256
- [Wuhan-Hu-1_2](#)
2. 1 ILLUMINA (NextSeq 500) run: 946,777 spots, 102.1M bases, 45.8Mb downloads
Accession: SRX17643255
- [Wuhan-Hu-1_1](#)
3. 1 ILLUMINA (NextSeq 500) run: 744,388 spots, 82.6M bases, 36.8Mb downloads
Accession: SRX17643254

Scarichiamo un campione da SRA

Full ▾

Send to: ▾

[SRX17643254](#): Wuhan-Hu-1_1

1 ILLUMINA (NextSeq 500) run: 744,388 spots, 82.6M bases, 36.8Mb downloads

Design: RNA was converted to cDNA using the NEBNextARTIC SARS-CoV-2 FSLibrary prep kit. The cDNA was amplified using the VarSkip short express protocol (NEB). The library was sequenced using the Illumina platform. Five hundred MB of data (2x 150 bp, paired ends) was produced per sample.

Submitted by: University of the Witwatersrand

Study: Whole genome sequences of SARS-CoV-2 variants

[PRJNA882477](#) • [SRP398285](#) • [All experiments](#) • [All runs](#)

[show Abstract](#)

Sample: Wuhan-Hu-1_1

[SAMN30937604](#) • [SRS15177384](#) • [All experiments](#) • [All runs](#)

Organism: [Severe acute respiratory syndrome coronavirus 2](#)

Library:

Name: CE-A1_S15

Instrument: NextSeq 500

Strategy: WGS

Source: VIRAL RNA

Selection: RT-PCR

Layout: PAIRED

Runs: 1 run, 744,388 spots, 82.6M bases, [36.8Mb](#)

| Run | # of Spots | # of Bases | Size | Published |
|-----------------------------|------------|------------|--------|------------|
| SRR21643281 | 744,388 | 82.6M | 36.8Mb | 2022-09-20 |

Scarichiamo un campione da SRA

Wuhan-Hu-1_1 (SRR21643281)

Metadata Analysis Reads Data access FASTA/FASTQ download

Run

| Run | Spots | Bases | Size | GC Content | Published | Access Type |
|-------------|--------|-------|-------|------------|------------|-------------|
| SRR21643281 | 744.4k | 82.6M | 36.8M | 38.5% | 2022-09-20 | public |

Quality graph [\(bigger\)](#)

This run has 2 reads per spot:

$\bar{L}=45, \sigma=22.8, 100\%$ $\bar{L}=66, \sigma=46.6, 100\%$

[? Legend](#)

Scarichiamo un campione da SRA

Wuhan-Hu-1_1 (SRR21643281)

Metadata

Analysis

Reads

Data access

FASTA/FASTQ download

Download for Experiment SRX17643254

| <input type="checkbox"/> Accession | Total Bases | Spots | |
|---|-------------|--------|----------|
| | | Total | Filtered |
| <input checked="" type="checkbox"/> SRR21643281 | 82.6M | 744.4k | |

Filter Runs

Search by sub-sequence,



Filter

[What can the filter be applied to?](#)

Download

Filtered Clipped

FASTA

or

FASTQ

Procuriamoci il materiale per l'allineamento



bowtie alignment

<https://bowtie-bio.sourceforge.net> · [Traduci questa pagina](#) ⋮

Bowtie: An ultrafast, memory-efficient short read aligner

Bowtie is an ultrafast, memory-efficient short read **aligner**. It aligns short DNA sequences (reads) to the human genome at a rate of over 25 million 35-bp reads ...

[Manual](#) · [Tools that use Bowtie](#) · [Bowtie 1.3.1](#)

Bowtie2



- bowtie2-2.5.0-linux-aarch64.zip
- bowtie2-2.5.0-source.zip
- bowtie2-2.5.0-mingw-x86_64.zip
- bowtie2-2.5.0-macos-arm64.zip
- bowtie2-2.5.0-linux-x86_64.zip
- bowtie2-2.5.0-macos-x86_64.zip

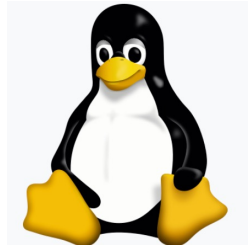


```
$ sudo apt-get install bowtie2
```



```
-> brew install bowtie2
```

Samtools



```
~$ sudo apt-get install samtools
```



```
[~] -> brew install samtools
```



- Scaricare da <http://bioinformatics.iit.cnr.it/pc/samtools.zip>
 - (Il link rimarrà attivo per qualche giorno)
- Scompattare il file zip
- samtools.exe va eseguito dalla PowerShell di windows

Creazione dell'indice

```
[lab] -> mkdir index
```

```
[lab] -> bowtie2-build GCF_009858895.2_ASM985889v3_genomic.fna index/GCF_009858895.2
```

```
[geraci]:viola [lab] -> ls index/
```

```
GCF_009858895.2.1.bt2      GCF_009858895.2.3.bt2      GCF_009858895.2.rev.1.bt2  
GCF_009858895.2.2.bt2      GCF_009858895.2.4.bt2      GCF_009858895.2.rev.2.bt2
```

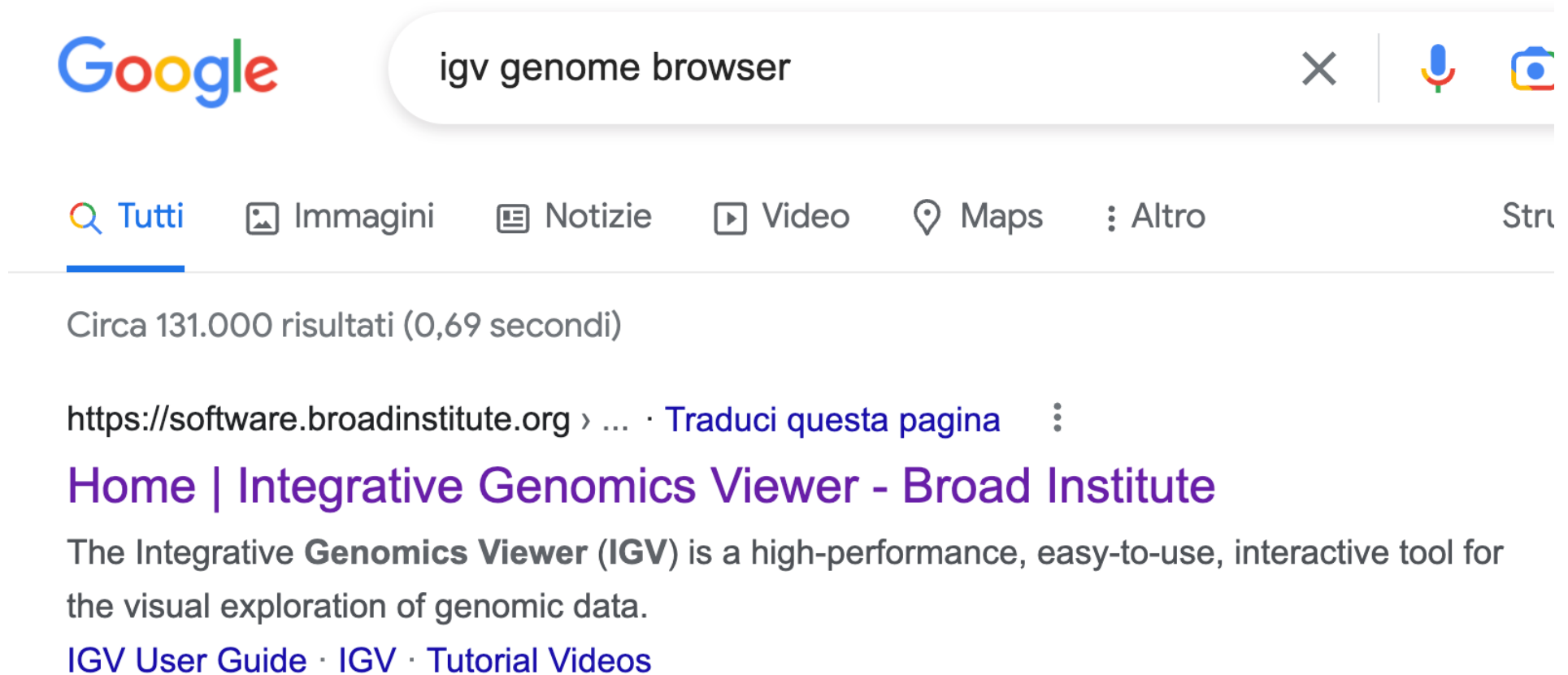
- Per poter allineare le sequenze di un campione dobbiamo creare un indice del genoma di riferimento.
- L'indice funziona come l'indice analitico di un libro: data una sequenza permette di trovare la sua posizione

Allineamento

```
[geraci]:viola [lab] -> bowtie2 -x index/GCF_009858895.2 -U SRR21643281.fastq -S SRR21643281.sam
1488776 reads; of these:
  1488776 (100.00%) were unpaired; of these:
    986049 (66.23%) aligned 0 times
    502727 (33.77%) aligned exactly 1 time
    0 (0.00%) aligned >1 times
33.77% overall alignment rate
```

```
[geraci]:viola [lab] -> samtools sort SRR21643281.sam -o SRR21643281.sort.bam
[geraci]:viola [lab] -> samtools index SRR21643281.sort.bam
[geraci]:viola [lab] -> ls SRR21643281*
SRR21643281.fastq          SRR21643281.sort.bam
SRR21643281.sam           SRR21643281.sort.bam.bai
```

Visualizzare l'allineamento (IGV)



The image shows a Google search interface. The search bar contains the text "igv genome browser". Below the search bar, there are navigation options: "Tutti" (selected), "Immagini", "Notizie", "Video", "Maps", "Altro", and "Stru". The search results show approximately 131,000 results in 0.69 seconds. The top result is from "https://software.broadinstitute.org" and is titled "Home | Integrative Genomics Viewer - Broad Institute". The description states: "The Integrative **Genomics Viewer (IGV)** is a high-performance, easy-to-use, interactive tool for the visual exploration of genomic data." Below the description are links for "IGV User Guide", "IGV", and "Tutorial Videos".

Google

igv genome browser

Tutti Immagini Notizie Video Maps Altro Stru

Circa 131.000 risultati (0,69 secondi)

<https://software.broadinstitute.org> > ... · [Traduci questa pagina](#)

Home | Integrative Genomics Viewer - Broad Institute

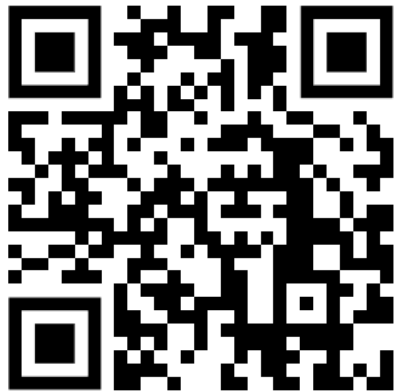
The Integrative **Genomics Viewer (IGV)** is a high-performance, easy-to-use, interactive tool for the visual exploration of genomic data.

[IGV User Guide](#) · [IGV](#) · [Tutorial Videos](#)







Visualizzare l'allineamento (IGV)



Materiale

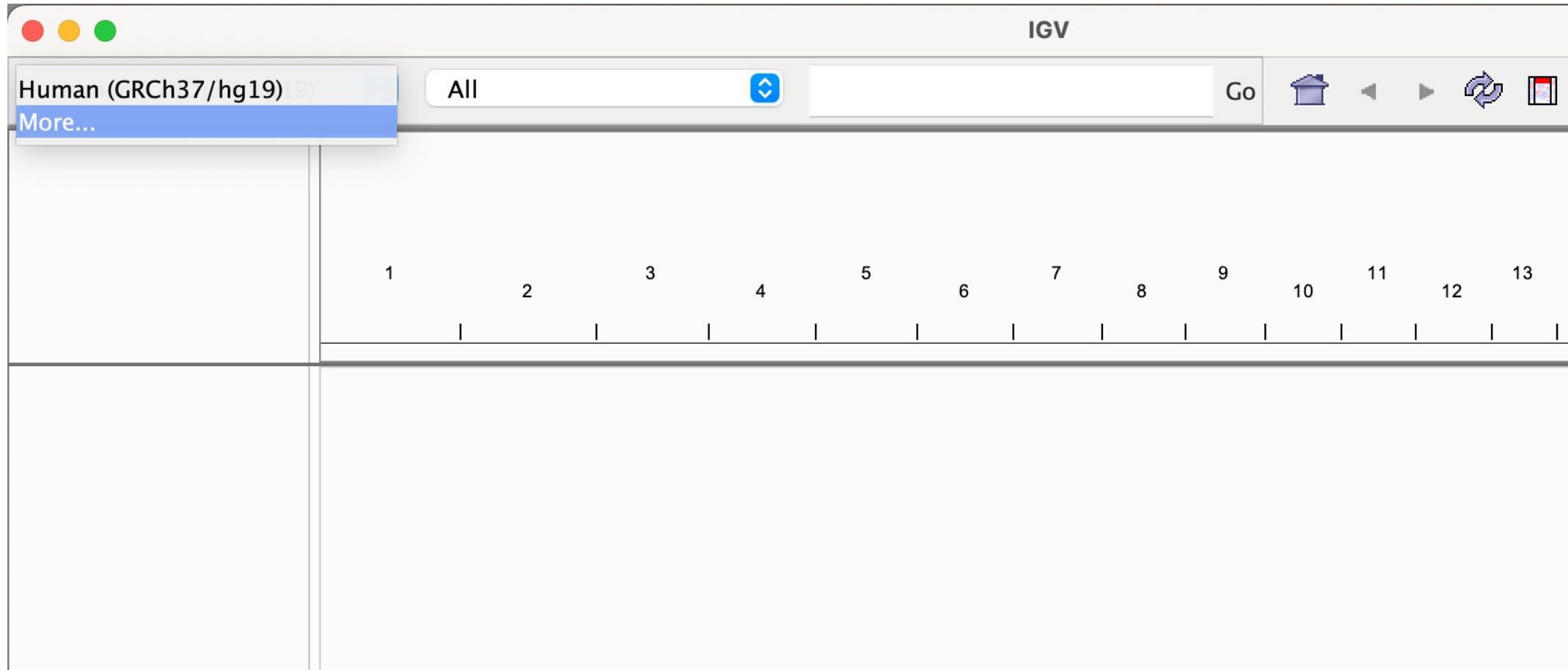


- Potete scaricare i dati da questo link
 - <http://bioinformatics.iit.cnr.it/pc/>
- Il link rimane attivo per qualche giorno

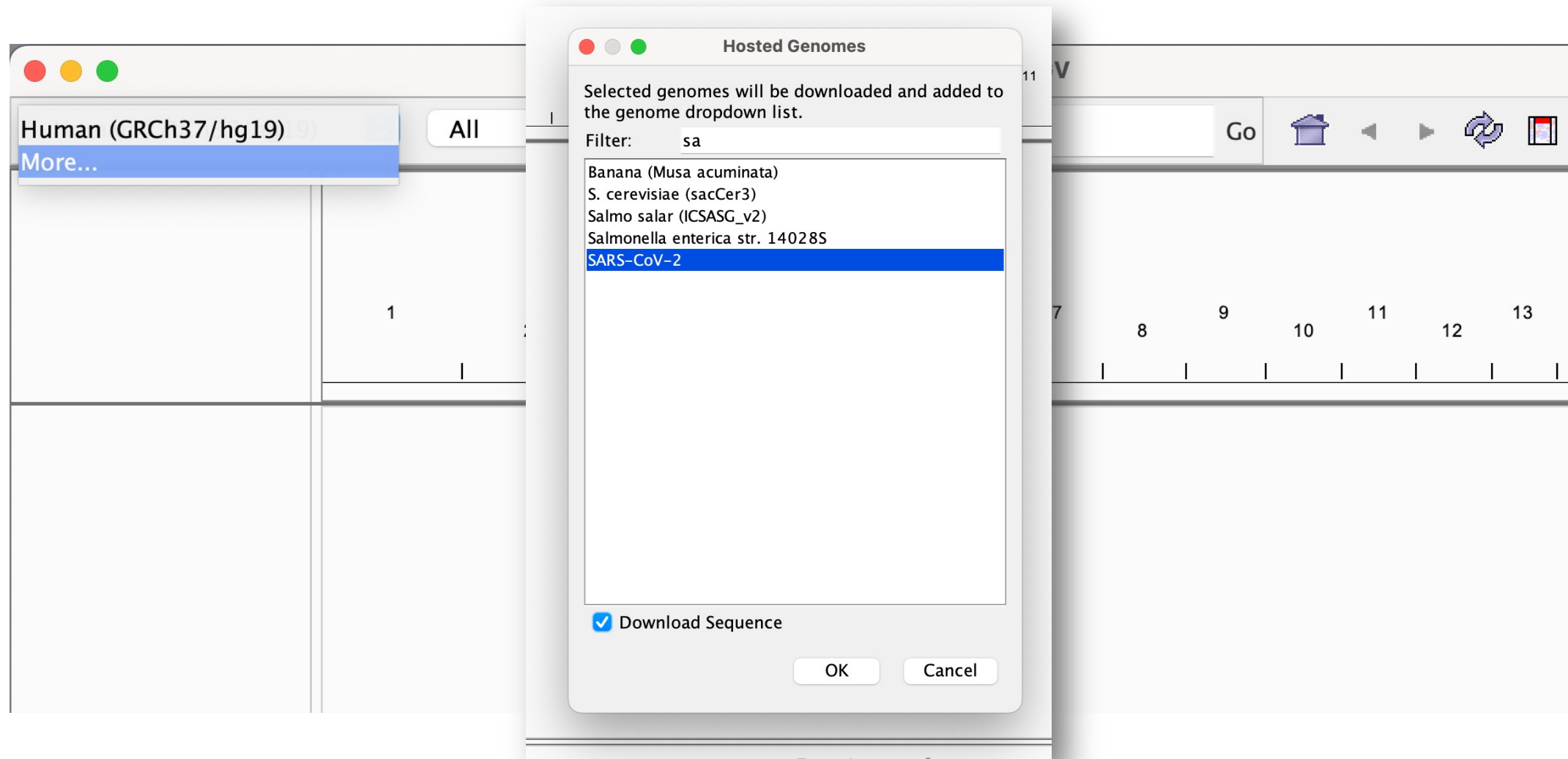
| <u>Name</u> | <u>Last modified</u> | <u>Size</u> | <u>Description</u> |
|--|----------------------|-------------|--------------------|
|  Parent Directory | | - | |
|  SRR21643281-sampled.sort.bam | 2022-11-11 09:06 | 5.3M | Computer lento |
|  SRR21643281-sampled.sort.bam.bai | 2022-11-11 09:06 | 184 | |
|  SRR21643281.sort.bam | 2022-11-11 10:00 | 24M | Computer veloce |
|  SRR21643281.sort.bam.bai | 2022-11-11 10:00 | 280 | |
|  samtools.zip | 2022-11-11 08:14 | 4.0M | |

Apache/2.4.41 (Ubuntu) Server at bioinformatics.iit.cnr.it Port 80

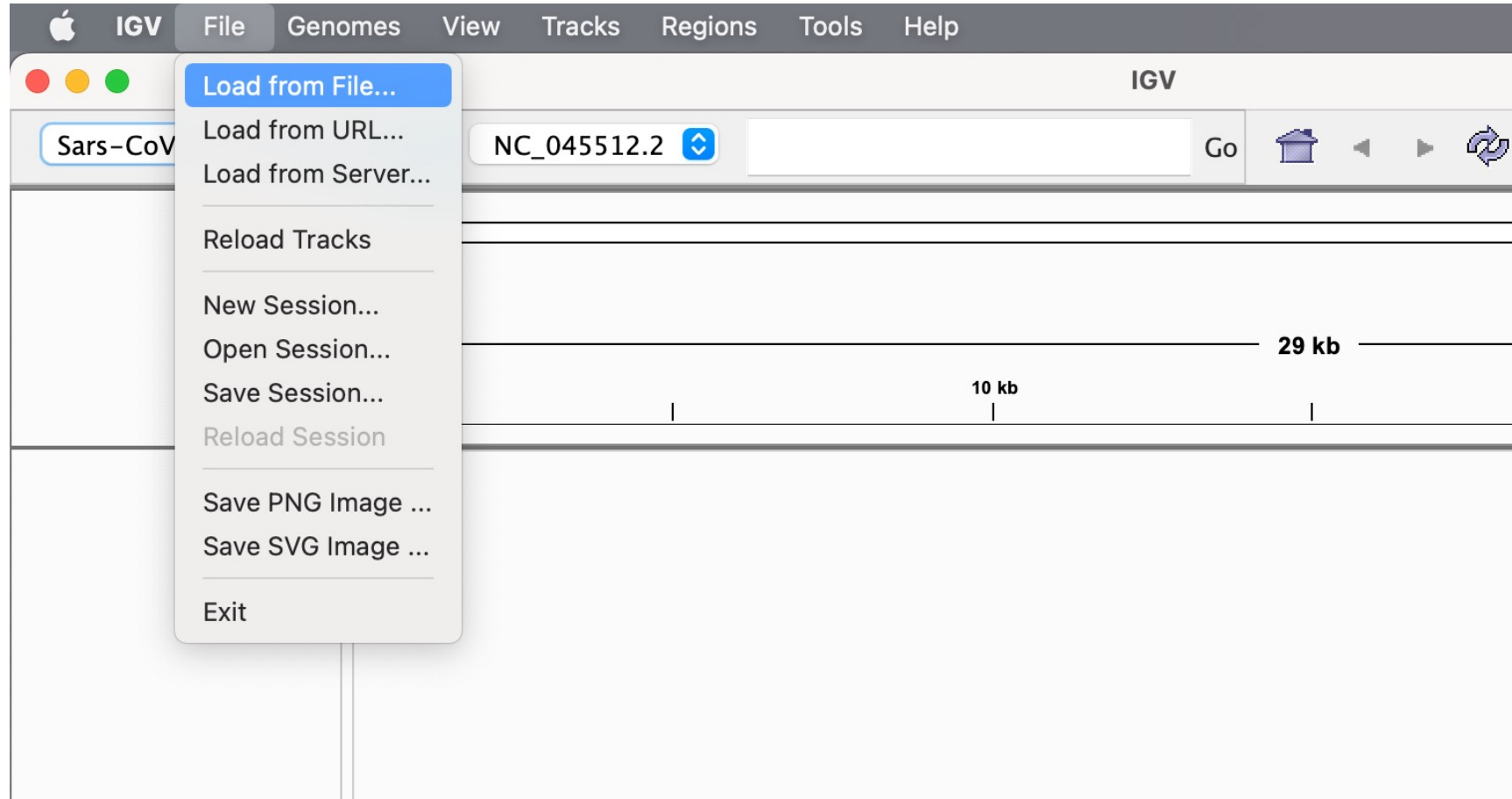
Caricare il genoma di riferimento in IGV



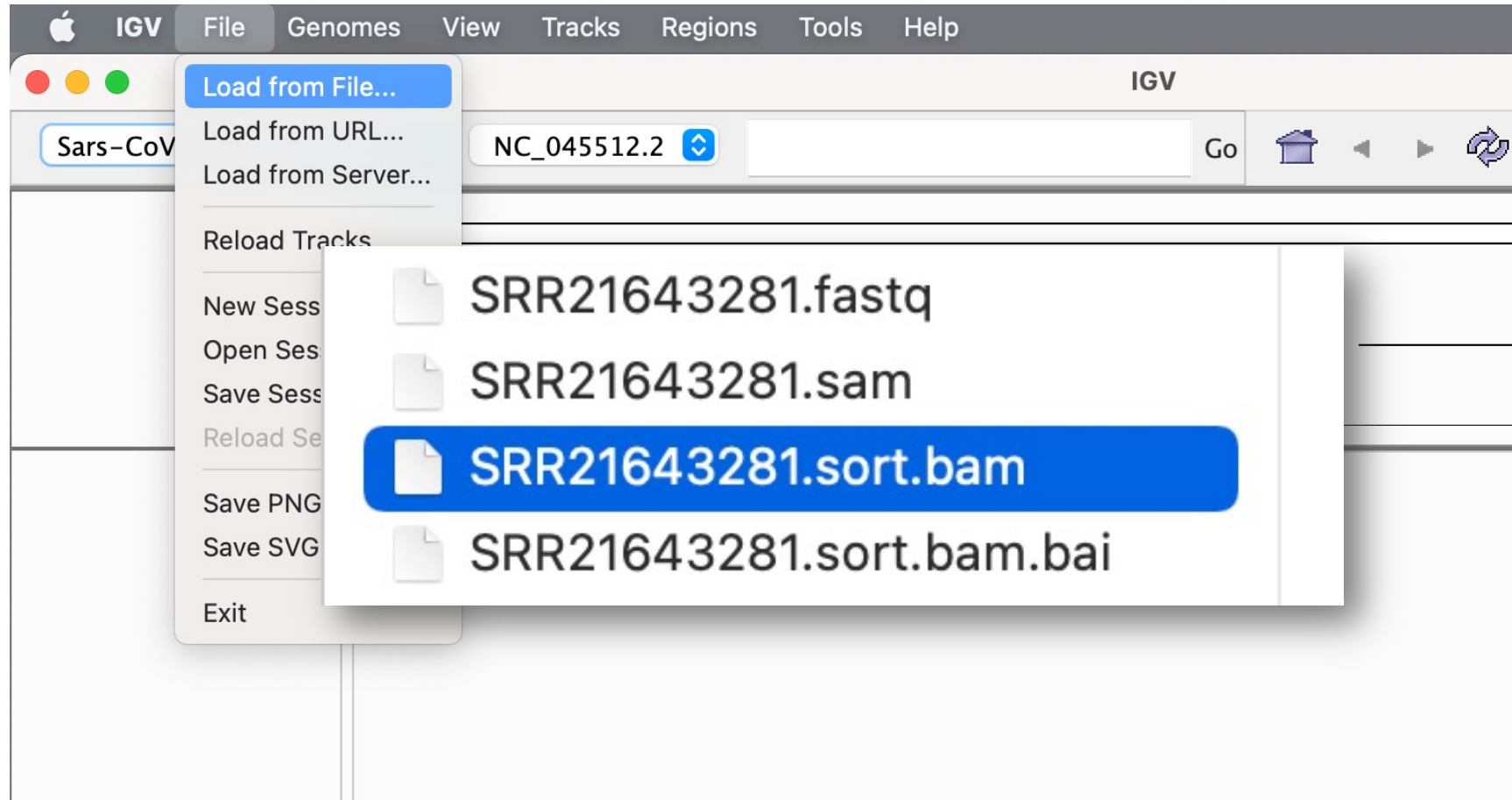
Caricare il genoma di riferimento in IGV



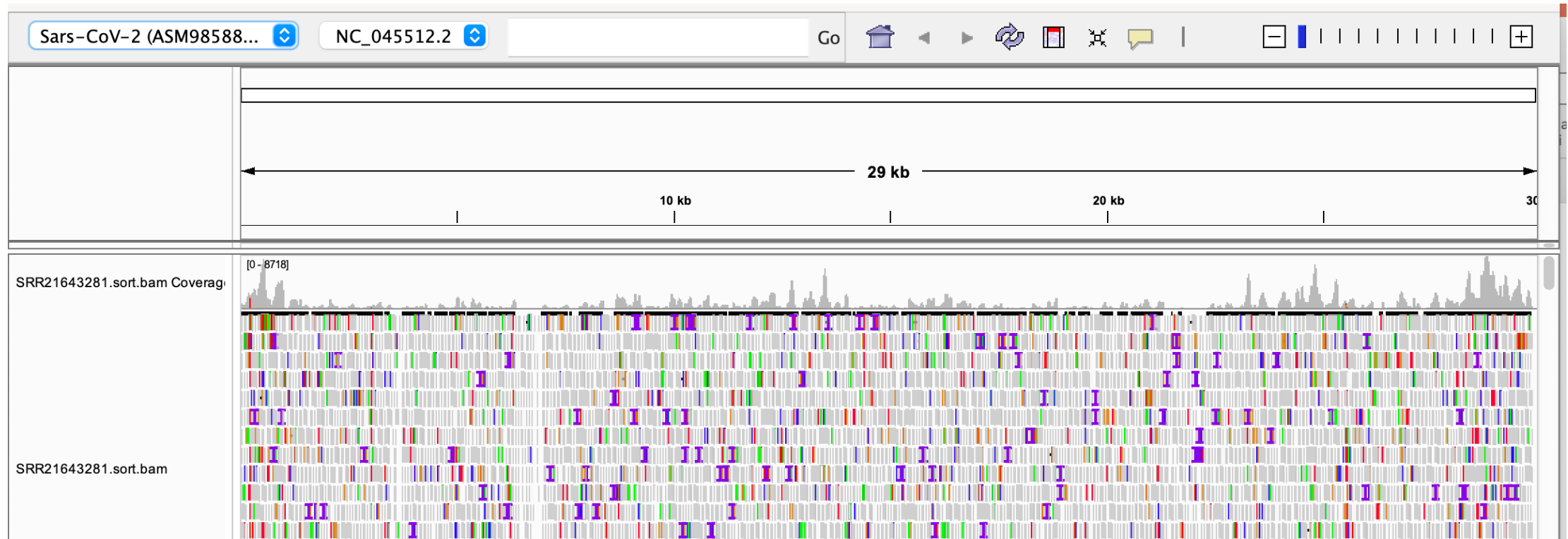
Carichiamo l'allineamento



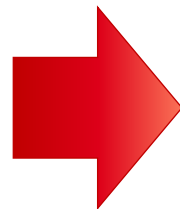
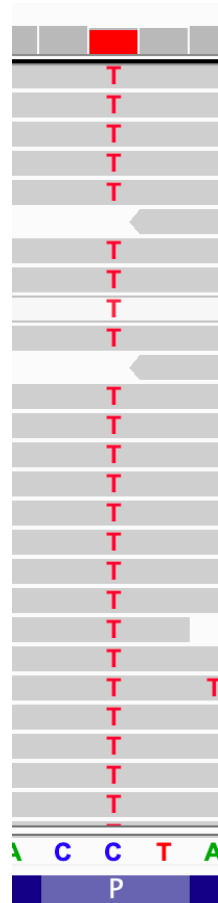
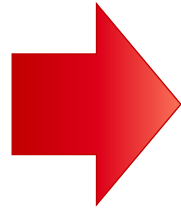
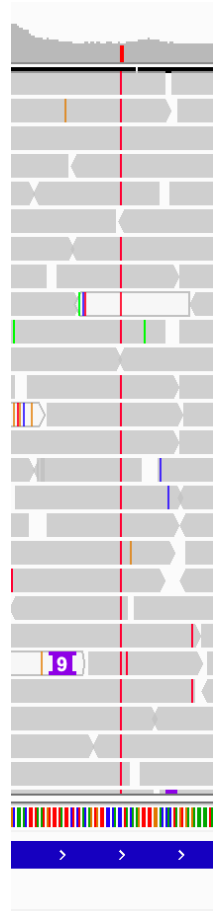
Carichiamo l'allineamento



Visualizziamo l'allineamento



Una variazione...finalmente



apolare polare basico acido codone di stop

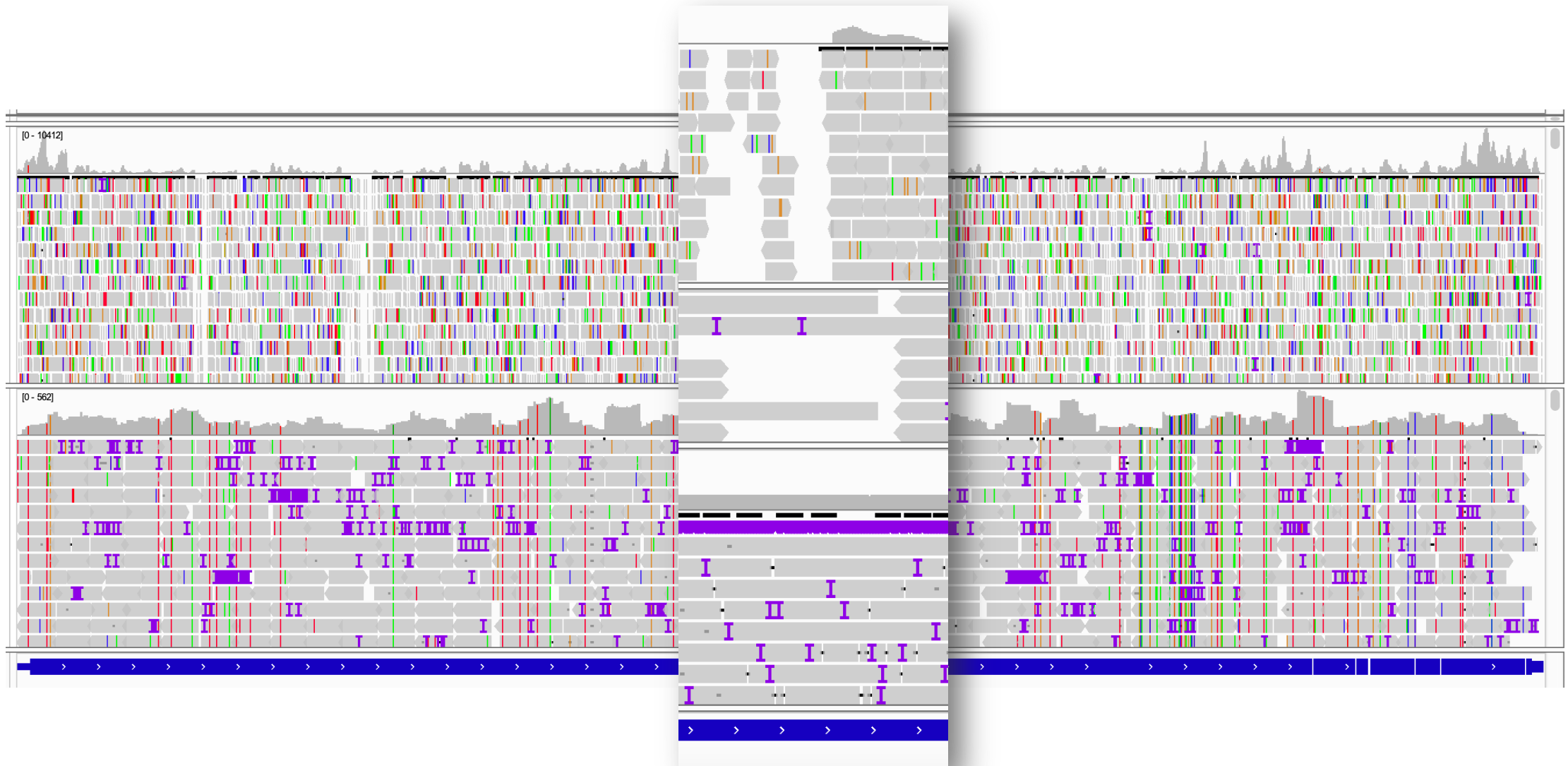
Codice genetico standard

| Prima base | Seconda base | | | | | | | | Terza base | |
|------------|--------------|----------------------|-----|------------------|-----|--------------------------|-----|--------------------|------------|---|
| | T | | C | | A | | G | | | |
| T | TTT | (Phe/F) Fenilalanina | TCT | (Ser/S) Serina | TAT | (Tyr/Y) Tirosina | TGT | (Cys/C) Cisteina | T | |
| | TTC | | TCC | | TAC | | TGC | | C | |
| | TTA | | TCA | | TAA | Stop (Ocra) | TGA | Stop (Opale) | A | |
| | TTG | | TCG | | TAG | Stop (Ambra) | TGG | (Trp/W) Triptofano | G | |
| C | CTT | (Leu/L) Leucina | CCT | (Pro/P) Prolina | CAT | (His/H) Istidina | CGT | (Arg/R) Arginina | T | |
| | CTC | | CCC | | CAC | | CGC | | | C |
| | CTA | | CCA | | CAA | (Gln/Q) Glutamina | CGA | | | A |
| | CTG | | CCG | | CAG | | CGG | | | G |
| A | ATT | (Ile/I) Isoleucina | ACT | (Thr/T) Treonina | AAT | (Asn/N) Asparagina | AGT | (Ser/S) Serina | T | |
| | ATC | | ACC | | AAC | | AGC | | | C |
| | ATA | | ACA | | AAA | (Lys/K) Lisina | AGA | (Arg/R) Arginina | A | |
| | ATG | (Met/M) Metionina | ACG | | AAG | | AGG | | G | |
| G | GTT | (Val/V) Valina | GCT | (Ala/A) Alanina | GAT | (Asp/D) Acido aspartico | GGT | (Gly/G) Glicina | T | |
| | GTC | | GCC | | GAC | | GGC | | | C |
| | GTA | | GCA | | GAA | (Glu/E) Acido glutammico | GGA | | | A |
| | GTG | | GCG | | GAG | | GGG | | | G |

La ricerca scientifica al tempo della pandemia



La ricerca scientifica al tempo della pandemia



Grazie



Ogni diritto è l'effetto
collaterale di un dovere

F. Geraci