ISTITVTO SVPERIORE DI SANITA'

Pipeline open-source per la ricostruzione e analisi dei genomi SARS-CoV-2

I laboratori di sanità pubblica veterinaria e la ricerca nella genomica di SARS Cov2: Esperienze a confronto

26 Novembre, 2020

Il database GISAID

Obiettivo GISAID:
raccolta dati per favorire condivisione e rapida valutazione dell'evoluzione e diffusione dei virus durante pandemie ed epidemie
Metodo: condivisione dati virus influenzali e del coronavirus che causa COVID-19.
Dati disponibili: la sequenza genetica, dati clinici ed epidemiologici associati ai virus umani, dati geografici e specie-specifici associati ai virus aviari e ad altri virus animali,

© 2008 - 2020 | Terms of Use | Privacy Notice | Contact

You are logged in as **Luca De Sabato** - logout

Primo genoma completo:

Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
NCBI Reference Sequence: NC_045512.2

Numero di sequenze disponibili:
217,673 totali
215,862 genomi completi

Koyama T, Platt D, Parida L. Variant analysis of SARS-CoV-2 genomes.
Bull World Health Organ. 2020;98(7):495-504. doi:10.2471/BLT.20.253591

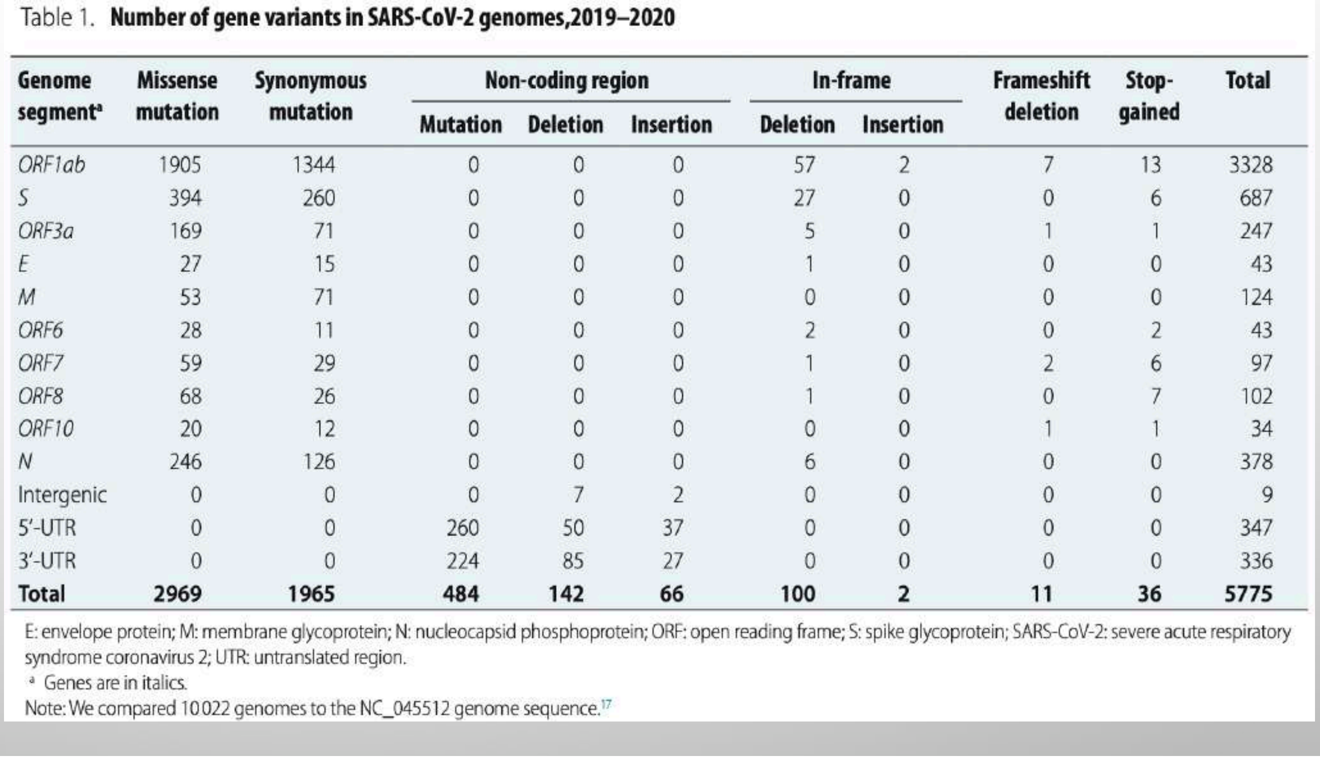


Table 1. **Number of gene variants in SARS-CoV-2 genomes,2019–2020**

Genome segment[a]	Missense mutation	Synonymous mutation	Non-coding region			In-frame		Frameshift deletion	Stop-gained	Total
			Mutation	Deletion	Insertion	Deletion	Insertion			
ORF1ab	1905	1344	0	0	0	57	2	7	13	3328
S	394	260	0	0	0	27	0	0	6	687
ORF3a	169	71	0	0	0	5	0	1	1	247
E	27	15	0	0	0	1	0	0	0	43
M	53	71	0	0	0	0	0	0	0	124
ORF6	28	11	0	0	0	2	0	0	2	43
ORF7	59	29	0	0	0	1	0	2	6	97
ORF8	68	26	0	0	0	1	0	0	7	102
ORF10	20	12	0	0	0	0	0	1	1	34
N	246	126	0	0	0	6	0	0	0	378
Intergenic	0	0	0	7	2	0	0	0	0	9
5′-UTR	0	0	260	50	37	0	0	0	0	347
3′-UTR	0	0	224	85	27	0	0	0	0	336
Total	**2969**	**1965**	**484**	**142**	**66**	**100**	**2**	**11**	**36**	**5775**

E: envelope protein; M: membrane glycoprotein; N: nucleocapsid phosphoprotein; ORF: open reading frame; S: spike glycoprotein; SARS-CoV-2: severe acute respiratory syndrome coronavirus 2; UTR: untranslated region.

[a] Genes are in italics.

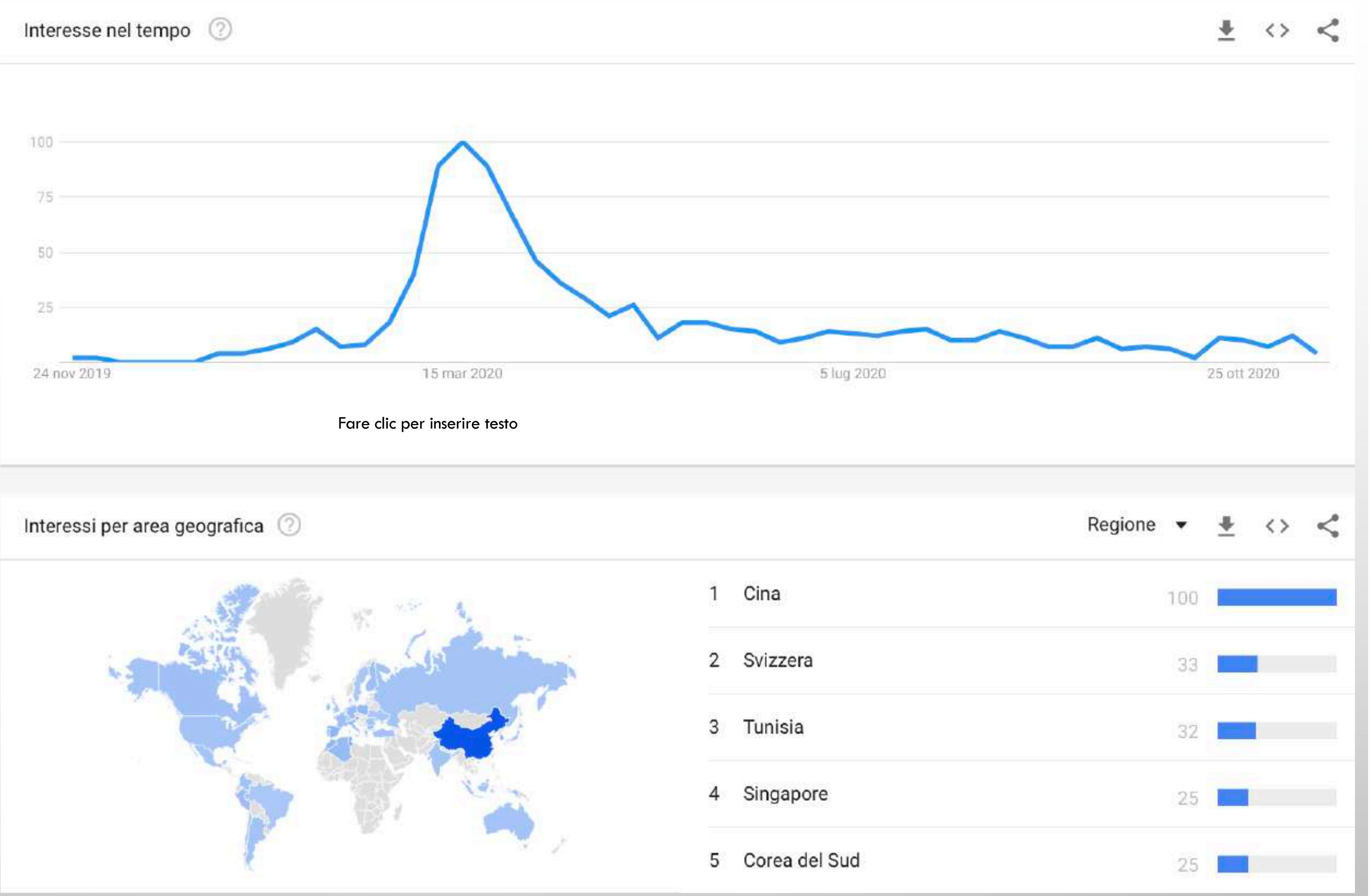
Note: We compared 10 022 genomes to the NC_045512 genome sequence.[17]

UTILIZZO, CONSULTAZIONE E CONTRIBUTO DI GISAID

Interesse nel tempo

I numeri rappresentano l'interesse di ricerca rispetto al punto più alto del grafico in relazione alla regione e al periodo indicati. Il valore 100 indica la maggiore frequenza di ricerca del termine, 50 indica la metà delle ricerche. Un punteggio pari a 0, invece, indica che non sono stati rilevati dati sufficienti per il termine.

Google Trends



Interesse nel tempo

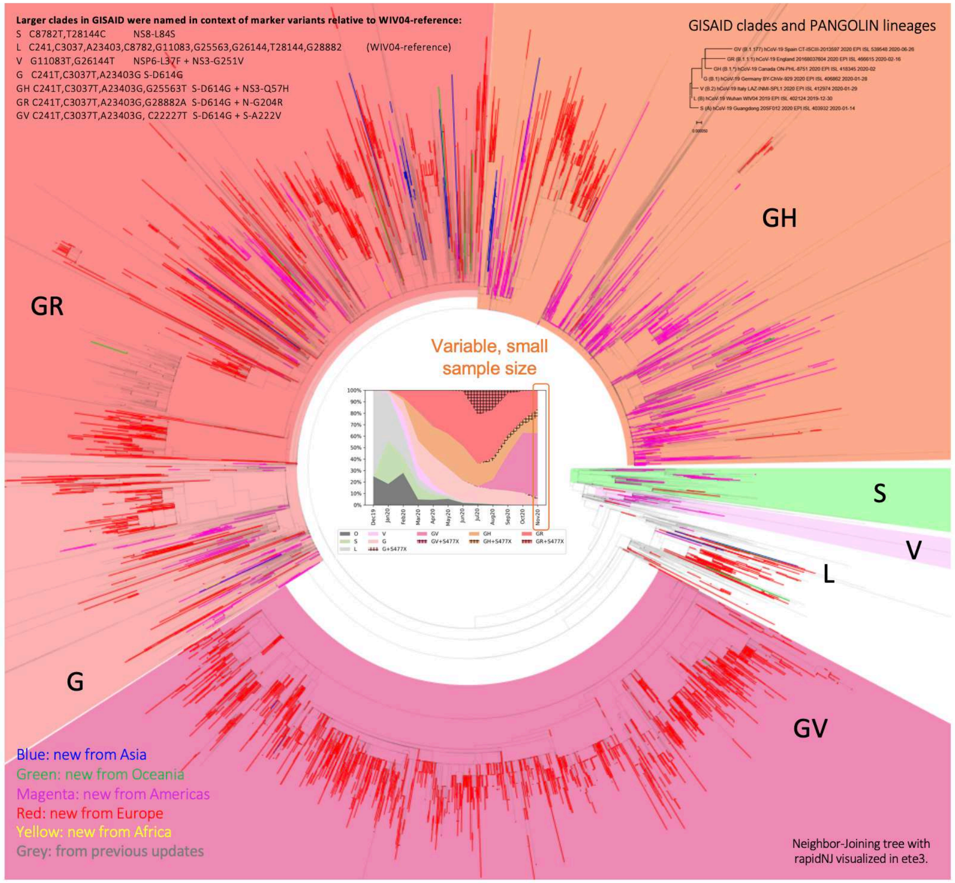
Fare clic per inserire testo

Interessi per area geografica

1	Cina	100
2	Svizzera	33
3	Tunisia	32
4	Singapore	25
5	Corea del Sud	25

Classificazione

Larger clades in GISAID were named in context of marker variants relative to WIV04-reference:
S   C8782T,T28144C   NS8-L84S
L   C241,C3037,A23403,C8782,G11083,G25563,G26144,T28144,G28882   (WIV04-reference)
V   G11083T,G26144T   NSP6-L37F + NS3-G251V
G   C241T,C3037T,A23403G S-D614G
GH C241T,C3037T,A23403G,G25563T S-D614G + NS3-Q57H
GR C241T,C3037T,A23403G,G28882A S-D614G + N-G204R
GV C241T,C3037T,A23403G, C22227T S-D614G + S-A222V



Larger clades in GISAID were named in context of marker variants relative to WIV04-reference:
S   C8782T,T28144C   NS8-L84S
L   C241,C3037,A23403,C8782,G11083,G25563,G26144,T28144,G28882   (WIV04-reference)
V   G11083T,G26144T   NSP6-L37F + NS3-G251V
G   C241T,C3037T,A23403G S-D614G
GH C241T,C3037T,A23403G,G25563T S-D614G + NS3-Q57H
GR C241T,C3037T,A23403G,G28882A S-D614G + N-G204R
GV C241T,C3037T,A23403G, C22227T S-D614G + S-A222V

GISAID clades and PANGOLIN lineages

GV (B.1.177) hCoV-19 Spain CT-ISCIII-2013597 2020 EPI ISL 539548 2020-06-26
GR (B.1.1.1) hCoV-19 England 20168037604 2020 EPI ISL 466615 2020-02-16
GH (B.1.*) hCoV-19 Canada ON-PHL-8751 2020 EPI ISL 418345 2020-02
G (B.1) hCoV-19 Germany BY-ChVir-929 2020 EPI ISL 406862 2020-01-28
V (B.2) hCoV-19 Italy LAZ-INMI-SPL1 2020 EPI ISL 412974 2020-01-29
L (B) hCoV-19 Wuhan WIV04 2019 EPI ISL 402124 2019-12-30
S (A) hCoV-19 Guangdong 20SF012 2020 EPI ISL 403932 2020-01-14
0.000050
GH
GR
S
V
L
G
GV
Variable, small sample size

Blue: new from Asia
Green: new from Oceania
Magenta: new from Americas
Red: new from Europe
Yellow: new from Africa
Grey: from previous updates

Neighbor-Joining tree with rapidNJ visualized in ete3.

Full genome tree derived from all outbreak sequences 2020-11-13

Notable changes:

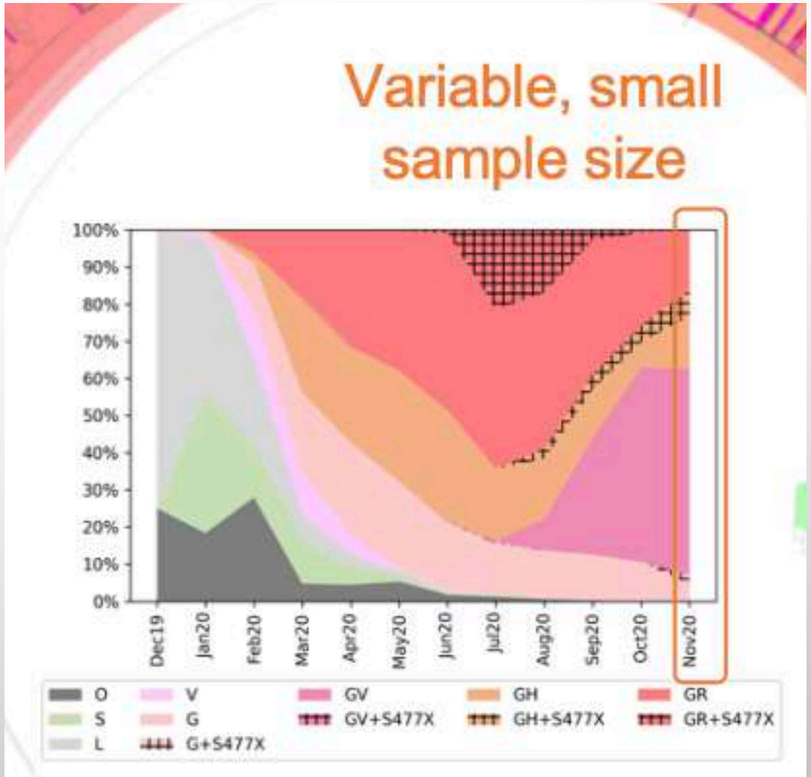
184,320 full genomes (+5,751) (excluding low coverage, out of 196,293 entries)

Updated clades:
S clade 6,913 (+38)
L clade 4,387 (+6)
V clade 5,434 (+20)
G clade [#S477X] 32,042 [116] (+582 [+0])
GR clade [#S477X] 68,267 [8,975] (+1,126 [+24])
GH clade [#S477X] 41,326 [2,723] (+1323 [+143])
GV clade [#S477X] 21,883 [6] (+2,641 [+1])
Other clades 4,068 (+15)

We gratefully acknowledge the Authors from Originating and Submitting laboratories of sequence data on which the analysis is based.

GISAID

by BII/GIS, A*STAR Singapore



Variable, small sample size

Next Generation Sequencing (NGS)

GISAID
NumeroSARS-CoV-2 genomi completi:

Illumina 147,238
Nanopore 46,186
Ion torrent 3,488

Tecnologia	Metodo	Lunghezza delle reads	Tasso di errore (%)	Output (GB/run)
Illumina	Synthesis	100-300 bp	0.1	200-600
Oxford Nanopore MinION	Nanopore	up to 1000 kb	5-20	5-10
Ion Torrent	Synthesis	200-600 bp	1	8-80

Vikas Bansal et al., Sequencing Technologies and Analyses: Where Have We Been and Where Are We Going? iScience 2019

Criticità del sequenziamento tramite NGS

-le potenzialità e la capacità di archiviazione del computer

-la competenza necessaria per analizzare e interpretare in modo completo i dati

-il volume dei dati e la loro gestione

-il costo effettivo del sequenziamento di NGS è trascurabile

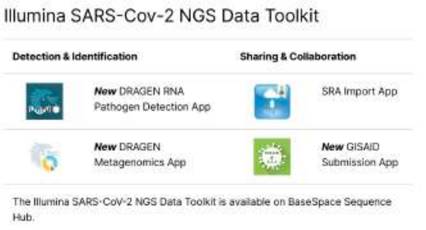
Behjati S, Tarpey PS. What is next generation sequencing?. *Arch Dis Child Educ Pract Ed.* 2013;98(6):236-238. doi:10.1136/archdischild-2013-304340

Software commerciali vs open-source

Software commerciali

Download and install QIAGEN CLC Genomics Workbench

Enjoy a FREE full-feature trial for 14 days



Illumina SARS-Cov-2 NGS Data Toolkit

Detection & Identification	Sharing & Collaboration
New DRAGEN RNA Pathogen Detection App	SRA Import App
New DRAGEN Metagenomics App	***New*** GISAID Submission App

The Illumina SARS-CoV-2 NGS Data Toolkit is available on BaseSpace Sequence Hub.

A fast, easy-to-use platform for microbiome sequencing and analysis

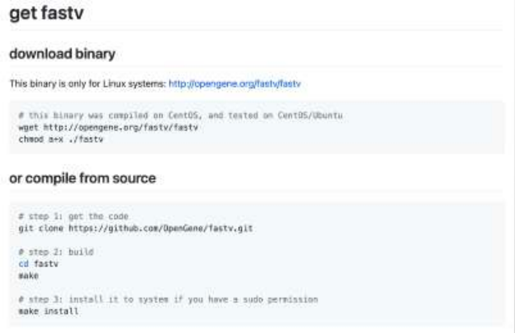
Get started quickly with an intuitive interface and rapid, accurate analysis. One Codex is the world's largest microbial reference database and can support millions of microbiome and infectious disease samples.

Analyze your WGS sequencing data automatically.

BugSeq uses evidence-based, pathogen-specific pipelines to produce actionable reports.*

Open-source

Installazione



get fastv

download binary

This binary is only for Linux systems: http://opengene.org/fastv/fastv

```
# this binary was compiled on CentOS, and tested on CentOS/Ubuntu
wget http://opengene.org/fastv/fastv
chmod a+x ./fastv
```

or compile from source

```
# step 1: get the code
git clone https://github.com/OpenGene/fastv.git

# step 2: build
cd fastv
make

# step 3: install it to system if you have a sudo permission
make install
```

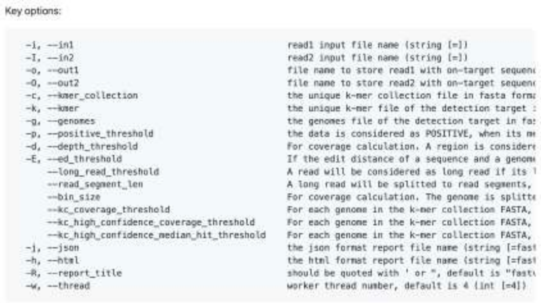
fastv
IRMA
MiCall
StaPH-B ToolKit

Pacchetti aggiuntivi

Contents

- Python
- Docker
- Singularity
- Java
- Installing the Toolkit

Comandi aggiuntivi



Key options:

```
  -i, --in1                                    read1 input file name (string [=])
  -I, --in2                                    read2 input file name (string [=])
  -o, --out1                                   file name to store read1 with on-target sequen
  -O, --out2                                   file name to store read2 with on-target sequen
  -c, --kmer_collection                        the unique k-mer collection file in fasta form
  -k, --kmer                                   the unique k-mer file of the detection target 
  -g, --genomes                                the genomes file of the detection target in fa
  -p, --positive_threshold                     the data is considered as POSITIVE, when its m
  -d, --depth_threshold                        For coverage calculation. A region is consider
  -E, --ed_threshold                           If the edit distance of a sequence and a genom
      --long_read_threshold                    A read will be considered as long read if its 
      --read_segment_len                       A long read will be splitted to read segments,
      --bin_size                               For coverage calculation. The genome is splitt
      --kc_coverage_threshold                  For each genome in the k-mer collection FASTA,
      --kc_high_confidence_coverage_threshold  For each genome in the k-mer collection FASTA,
      --kc_high_confidence_median_hit_threshold For each genome in the k-mer collection FASTA,
  -j, --json                                   the json format report file name (string [=fas
  -h, --html                                   the html format report file name (string [=fas
  -R, --report_title                           should be quoted with ' or ", default is "fast
  -w, --thread                                 worker thread number, default is 4 (int [=4])
```

https://github.com/CDCgov/SARS-CoV-2_Sequencing#bioinformatics

Galaxy / ARIES

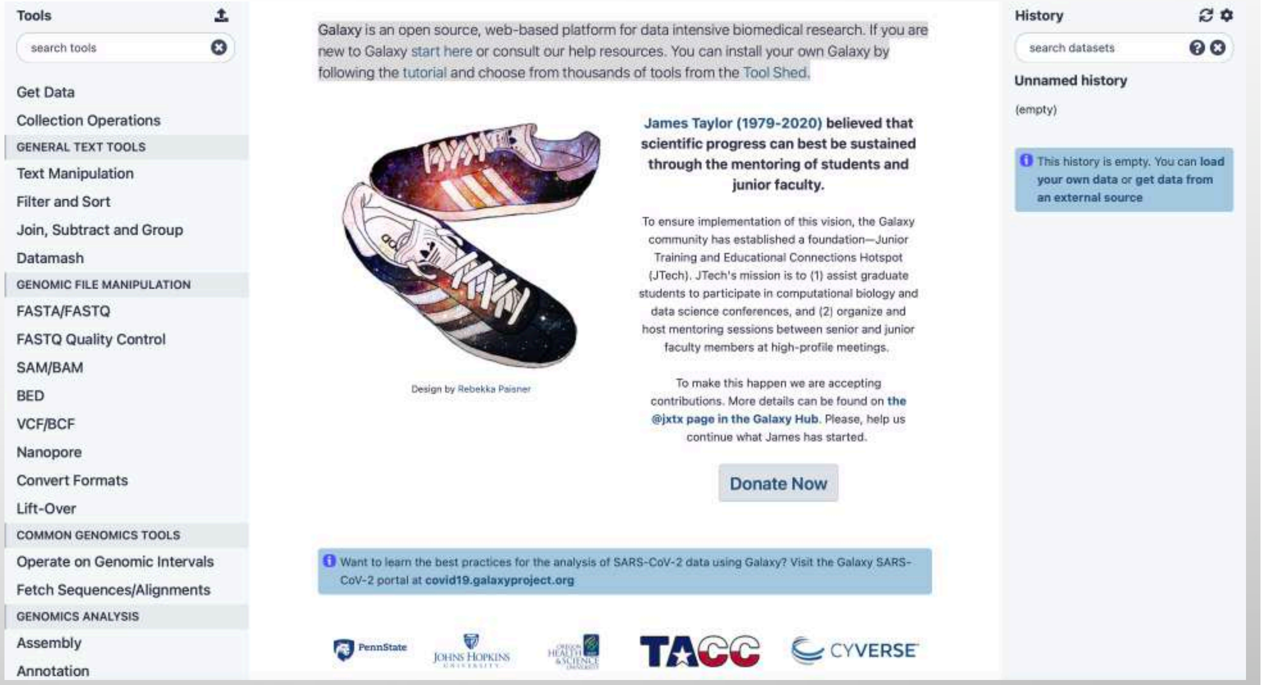
Studio:

Galaxy è una piattaforma web open source per la ricerca biomedica ad alta intensità di dati con migliaia di strumenti dal Tool Shed.

Obiettivo:

tool per la ricostruzione dei genomi del SARS-CoV-2 e l'analisi di risultati comparabili.

Caratteristiche: user friendly; non è necessario essere programmatori



Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

Design by Rebekka Paisner

James Taylor (1979-2020) believed that scientific progress can best be sustained through the mentoring of students and junior faculty.

To ensure implementation of this vision, the Galaxy community has established a foundation—Junior Training and Educational Connections Hotspot (JTech). JTech's mission is to (1) assist graduate students to participate in computational biology and data science conferences, and (2) organize and host mentoring sessions between senior and junior faculty members at high-profile meetings.

To make this happen we are accepting contributions. More details can be found on **the @jxtx page in the Galaxy Hub**. Please, help us continue what James has started.

This history is empty. You can **load your own data** or **get data from an external source**

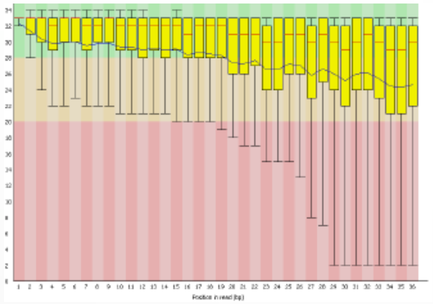
Want to learn the best practices for the analysis of SARS-CoV-2 data using Galaxy? Visit the Galaxy SARS-CoV-2 portal at **covid19.galaxyproject.org**

https://aries.iss.it

Metodo: Costruzione della pipeline

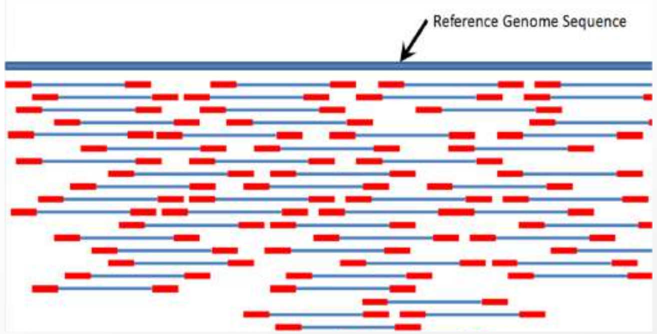
Reads totali

Trimming



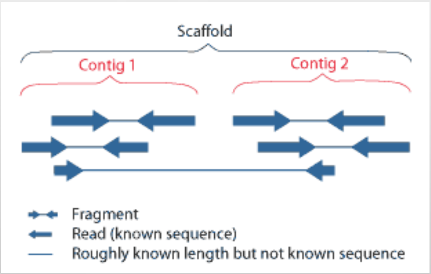
Trimmomatic tool
Quality base
Filter shorter reads

Rimozione sequenze dell’ospite



Bowtie tool
Human genome reference sequence

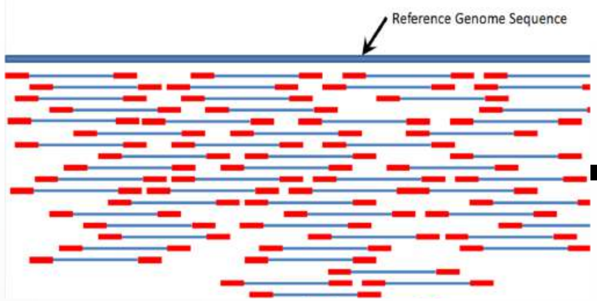
De novo assembling



Spades 3.12

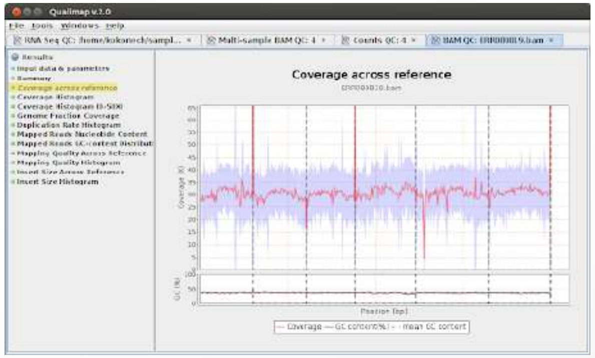
Sequenze non filtrate

SARS-CoV-2 reference genome



Bowtie tool

Analisi del coverage


Qualimap v.2.0
File Tools Windows Help
RNA Seq QC: /home/kokonech/sampl...
Multi-sample BAM QC: 4
Counts QC: 4
BAM QC: ERR000819.bam
Results
Input data & parameters
Summary
Coverage across reference
Coverage Histogram
Coverage Histogram (0-50X)
Genome Fraction Coverage
Duplication Rate Histogram
Mapped Reads Nucleotide Content
Mapped Reads GC-content Distribution
Mapping Quality Across Reference
Mapping Quality Histogram
Insert Size Across Reference
Insert Size Histogram
Coverage across reference
Coverage (X)
GC (%)
Position (bp)
Coverage — GC content(%) - - mean GC content

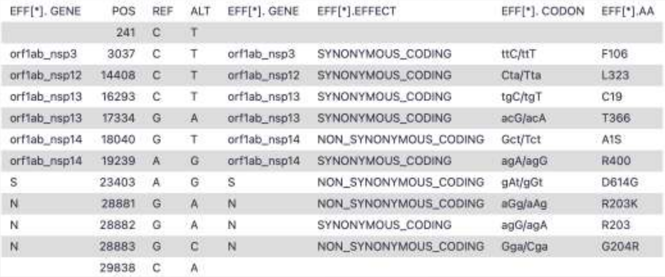
Qualimap 2

Ricostruzione della consensus

```
ACCCTCTTTTCTTATCATTGACATTTAAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCCCGGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCGGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCCAAGTGGCCCCGGGGCTTGATTTTTGCTTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCTT
TGCCGAGTGTGCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACTTCCAGTCCGGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA
GGAGGTGGGGACGCACTTTGCATCCAGACCTCCTCTGCATCGCAGTTCACGACATCCACGCTTGGGAAAG
TCCGTACCCGCGCCTGGAGCGCTTAAAGACACCCTGCCGCGGGTCGGGCGAGGTGCAGCAGAAGTTTCCC
GCGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC
```

SAMtools: mpileup tool, freebayes, vcf tool

Analisi delle varianti



EFF[*]. GENE	POS	REF	ALT	EFF[*]. GENE	EFF[*].EFFECT	EFF[*]. CODON	EFF[*].AA
	241	C	T				
orf1ab_nsp3	3037	C	T	orf1ab_nsp3	SYNONYMOUS_CODING	ttC/ttT	F106
orf1ab_nsp12	14408	C	T	orf1ab_nsp12	SYNONYMOUS_CODING	Cta/Tta	L323
orf1ab_nsp13	16293	C	T	orf1ab_nsp13	SYNONYMOUS_CODING	tgC/tgT	C19
orf1ab_nsp13	17334	G	A	orf1ab_nsp13	SYNONYMOUS_CODING	acG/acA	T366
orf1ab_nsp14	18040	G	T	orf1ab_nsp14	NON_SYNONYMOUS_CODING	Gct/Tct	A1S
orf1ab_nsp14	19239	A	G	orf1ab_nsp14	SYNONYMOUS_CODING	agA/agG	R400
S	23403	A	G	S	NON_SYNONYMOUS_CODING	gAt/gGt	D614G
N	28881	G	A	N	NON_SYNONYMOUS_CODING	aGg/aAg	R203K
N	28882	G	A	N	SYNONYMOUS_CODING	agG/agA	R203
N	28883	G	C	N	NON_SYNONYMOUS_CODING	Gga/Cga	G204R
	29838	C	A				

SnpEff tool

Annotazione ORFs

```
>ORF10
[illegible]
>N
[illegible]
>ORF8
[illegible]
>ORF7b
[illegible]
>ORF7a
[illegible]
```

Megablast vs SARS-CoV-2 ORFs

RISULTATI

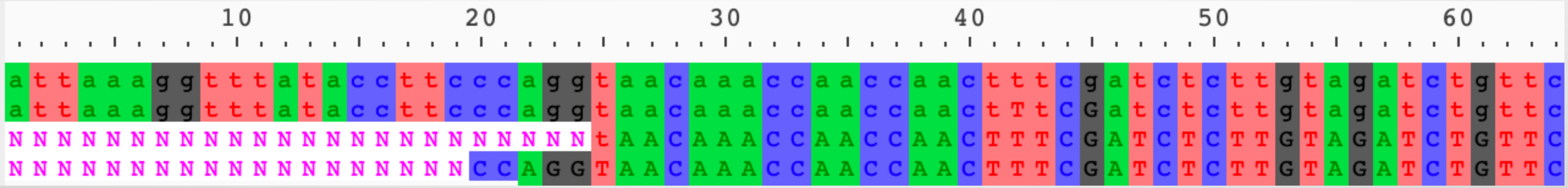
SARS-CoV-2 RECoVERY

REconstruction of COronaVirus gEnomes & Rapid analYsis

Metodo: comparazione risultati ottenuti con la pipeline RECoVERY utilizzando sequenze di genomi già disponibili online

Genomi completi ricostruiti (database Sequence Read Archive, SRA) :

- 100 raw data Illumina
- 100 raw data Nanopore
- 50 raw data IonTorrent



```
attaaaggtttataccttcccaggtaacaaaccaaccaactttcgatctcttgtagatctgttc
attaaaggtttataccttcccaggtaacaaaccaaccaactTtCGatctcttgtagatctgttc
NNNNNNNNNNNNNNNNNNNNNNNNtAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTC
NNNNNNNNNNNNNNNNNNNCCAGGTAACAAACCAACCAACTTTCGATCTCTTGTAGATCTGTTC
```

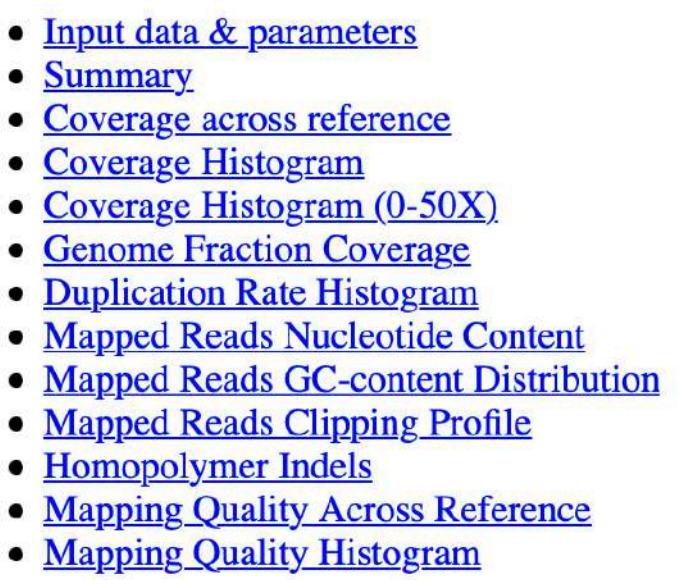
Software utilizzati:

- CLC Genomics Workbench Ver. 9.5 (Qiagen)
- online tool Genome Detective Virus Tool
- SARS-CoV-2 RECoVERY

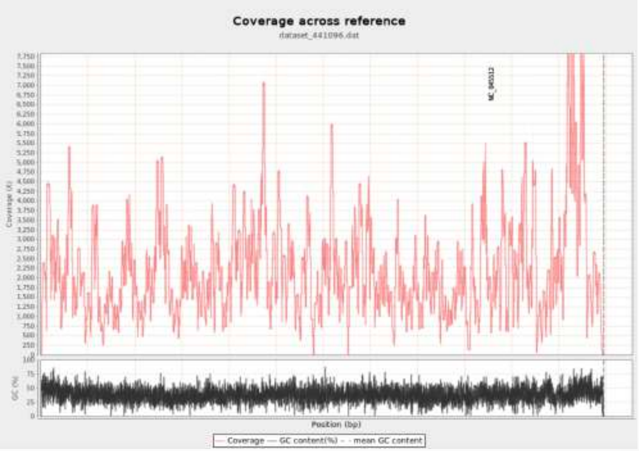
Differenze tra genomi ricostruiti in termini di differenza di lunghezza rispetto alle sequenze di riferimento GISAID e numero di nucleotidi diversi chiamati

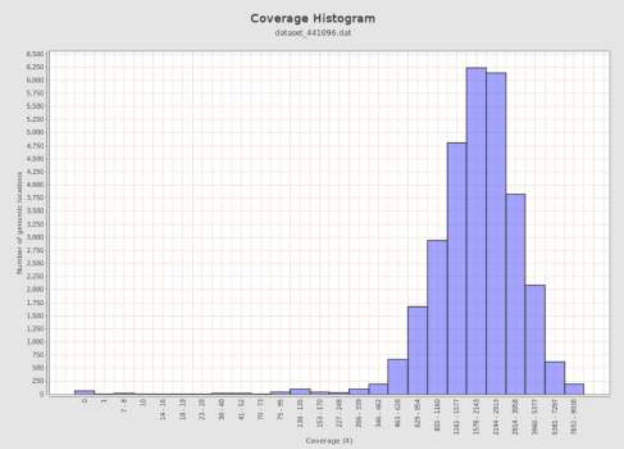
RISULTATI: GLI OUTPUT

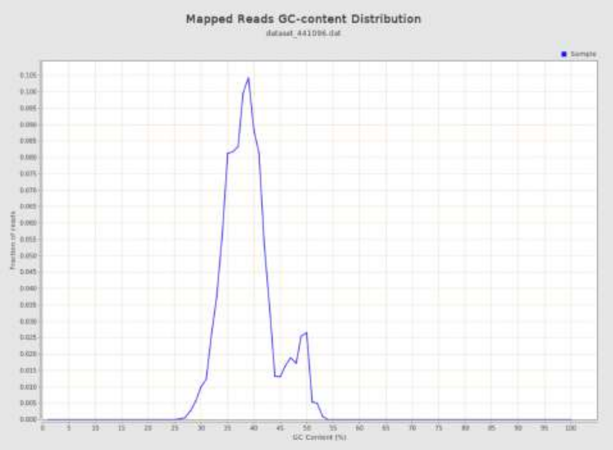
Report Qualimap

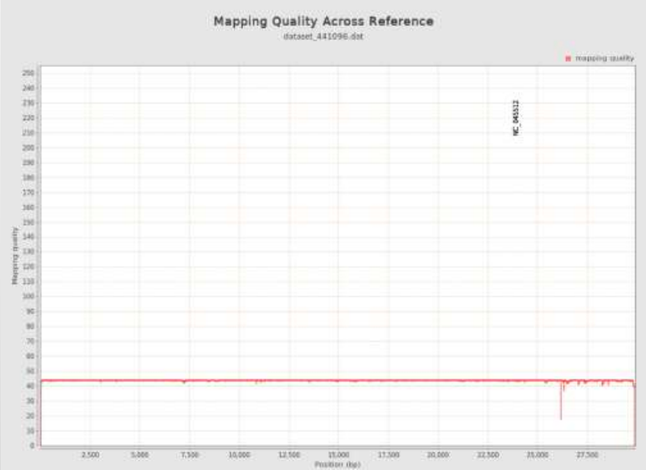


- Input data & parameters
- Summary
- Coverage across reference
- Coverage Histogram
- Coverage Histogram (0-50X)
- Genome Fraction Coverage
- Duplication Rate Histogram
- Mapped Reads Nucleotide Content
- Mapped Reads GC-content Distribution
- Mapped Reads Clipping Profile
- Homopolymer Indels
- Mapping Quality Across Reference
- Mapping Quality Histogram

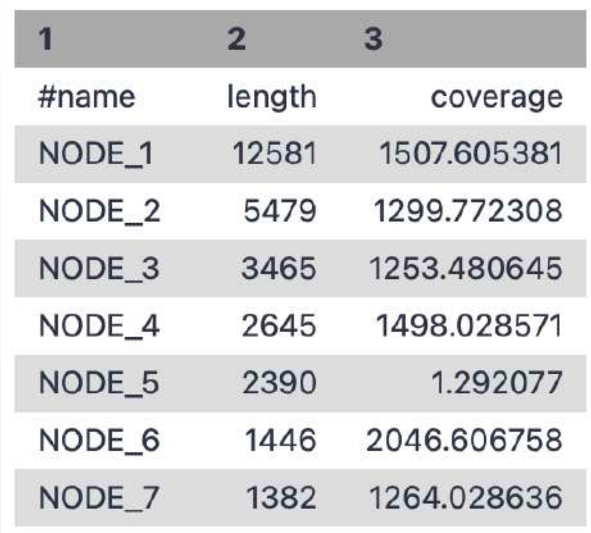






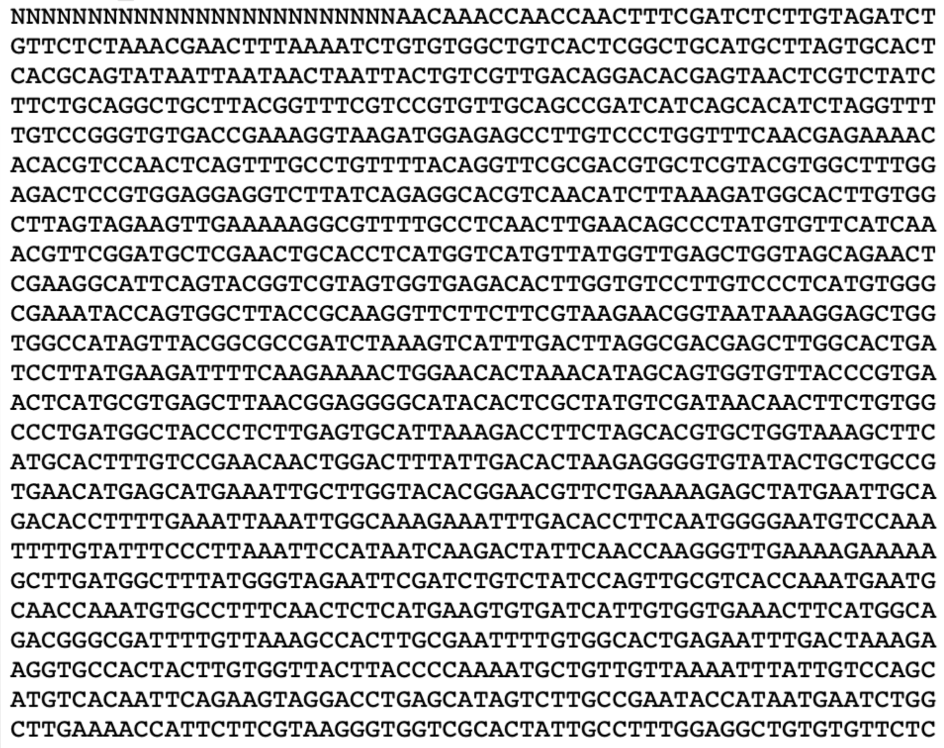


Scaffold: statistiche e fasta



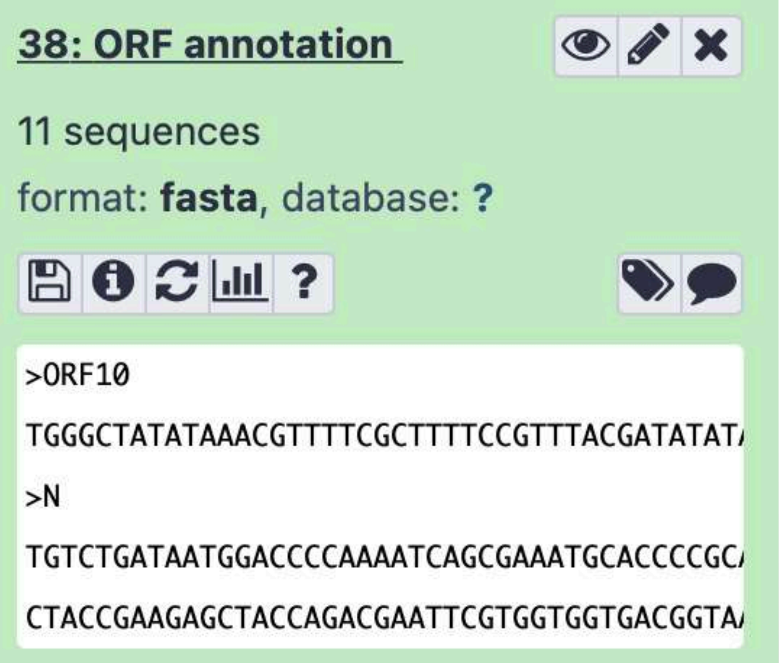
1	2	3
#name	length	coverage
NODE_1	12581	1507.605381
NODE_2	5479	1299.772308
NODE_3	3465	1253.480645
NODE_4	2645	1498.028571
NODE_5	2390	1.292077
NODE_6	1446	2046.606758
NODE_7	1382	1264.028636

Genoma Completo fasta



```
NNNNNNNNNNNNNNNNNNNNNNNNNAACAAACCAACCAACTTTCGATCTCTTGTAGATCT
GTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGTGCACT
CACGCAGTATAATTAATAACTAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATC
TTCTGCAGGCTGCTTACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTT
TGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTCCCTGGTTTCAACGAGAAAAC
ACACGTCCAACTCAGTTTGCCTGTTTTACAGGTTCGCGACGTGCTCGTACGTGGCTTTGG
AGACTCCGTGGAGGAGGTCTTATCAGAGGCACGTCAACATCTTAAAGATGGCACTTGTGG
CTTAGTAGAAGTTGAAAAAGGCGTTTTGCCTCAACTTGAACAGCCCTATGTGTTCATCAA
ACGTTCGGATGCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACT
CGAAGGCATTCAGTACGGTCGTAGTGGTGAGACACTTGGTGTCCTTGTCCCTCATGTGGG
CGAAATACCAGTGGCTTACCGCAAGGTTCTTCTTCGTAAGAACGGTAATAAAGGAGCTGG
TGGCCATAGTTACGGCGCCGATCTAAAGTCATTTGACTTAGGCGACGAGCTTGGCACTGA
TCCTTATGAAGATTTTCAAGAAAACTGGAACACTAAACATAGCAGTGGTGTTACCCGTGA
ACTCATGCGTGAGCTTAACGGAGGGGCATACACTCGCTATGTCGATAACAACTTCTGTGG
CCCTGATGGCTACCCTCTTGAGTGCATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTC
ATGCACTTTGTCCGAACAACTGGACTTTATTGACACTAAGAGGGGTGTATACTGCTGCCG
TGAACATGAGCATGAAATTGCTTGGTACACGGAACGTTCTGAAAAGAGCTATGAATTGCA
GACACCTTTTGAAATTAAATTGGCAAAGAAATTTGACACCTTCAATGGGGAATGTCCAAA
TTTTGTATTTCCCTTAAATTCCATAATCAAGACTATTCAACCAAGGGTTGAAAAGAAAAA
GCTTGATGGCTTTATGGGTAGAATTCGATCTGTCTATCCAGTTGCGTCACCAAATGAATG
CAACCAAATGTGCCTTTCAACTCTCATGAAGTGTGATCATTGTGGTGAAACTTCATGGCA
GACGGGCGATTTTGTTAAAGCCACTTGCGAATTTTGTGGCACTGAGAATTTGACTAAAGA
AGGTGCCACTACTTGTGGTTACTTACCCCAAAATGCTGTTGTTAAAATTTATTGTCCAGC
ATGTCACAATTCAGAAGTAGGACCTGAGCATAGTCTTGCCGAATACCATAATGAATCTGG
CTTGAAAACCATTCTTCGTAAGGGTGGTCGCACTATTGCCTTTGGAGGCTGTGTGTTCTC
```

Open Reading Frames: nucleotidi



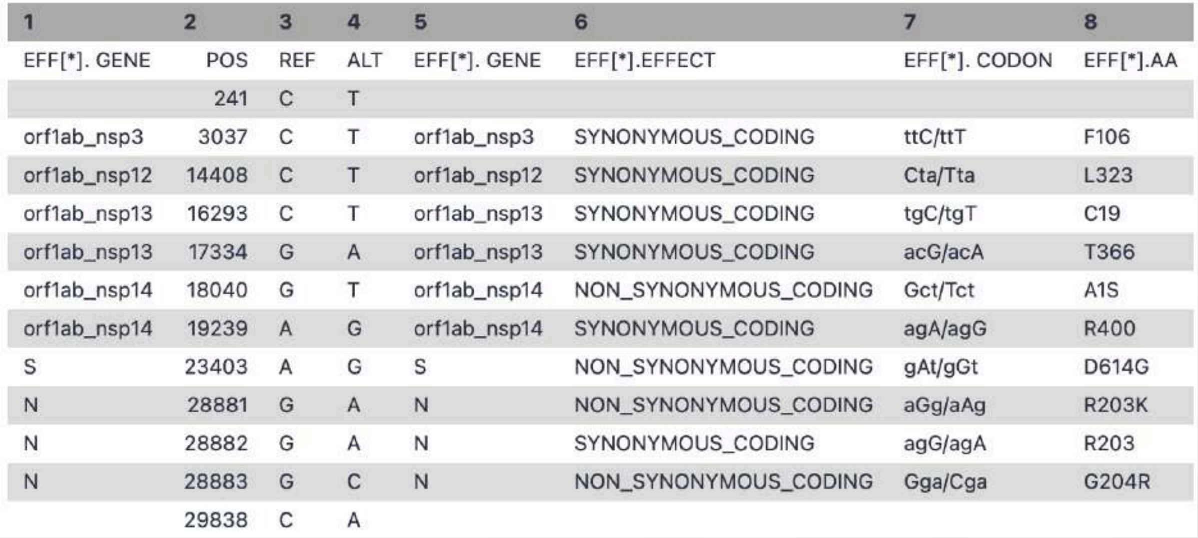
38: ORF annotation

11 sequences

format: **fasta**, database: **?**

```
>ORF10
TGGGCTATATAAACGTTTTCGCTTTTCCGTTTACGATATAT
>N
TGTCTGATAATGGACCCCAAAATCAGCGAAATGCACCCCGC
CTACCGAAGAGCTACCAGACGAATTCGTGGTGGTGACGGTA
```

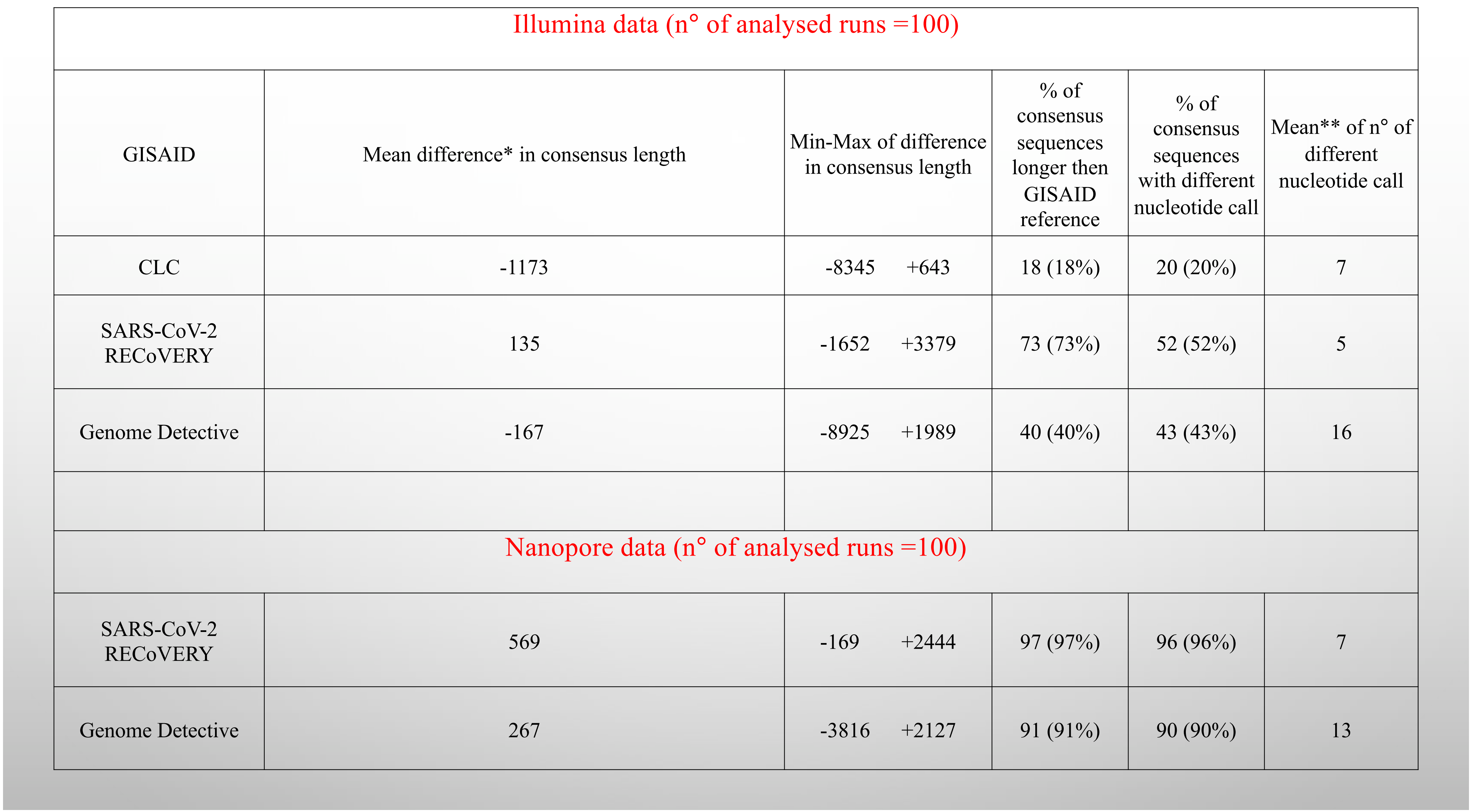
Chiamata delle varianti



1	2	3	4	5	6	7	8
EFF[*]. GENE	POS	REF	ALT	EFF[*]. GENE	EFF[*].EFFECT	EFF[*]. CODON	EFF[*].AA
	241	C	T				
orf1ab_nsp3	3037	C	T	orf1ab_nsp3	SYNONYMOUS_CODING	ttC/ttT	F106
orf1ab_nsp12	14408	C	T	orf1ab_nsp12	SYNONYMOUS_CODING	Cta/Tta	L323
orf1ab_nsp13	16293	C	T	orf1ab_nsp13	SYNONYMOUS_CODING	tgC/tgT	C19
orf1ab_nsp13	17334	G	A	orf1ab_nsp13	SYNONYMOUS_CODING	acG/acA	T366
orf1ab_nsp14	18040	G	T	orf1ab_nsp14	NON_SYNONYMOUS_CODING	Gct/Tct	A1S
orf1ab_nsp14	19239	A	G	orf1ab_nsp14	SYNONYMOUS_CODING	agA/agG	R400
S	23403	A	G	S	NON_SYNONYMOUS_CODING	gAt/gGt	D614G
N	28881	G	A	N	NON_SYNONYMOUS_CODING	aGg/aAg	R203K
N	28882	G	A	N	SYNONYMOUS_CODING	agG/agA	R203
N	28883	G	C	N	NON_SYNONYMOUS_CODING	Gga/Cga	G204R
	29838	C	A				

Comparazione dei risultati

Ion Torrent data (n° of analysed runs =50)					
GISAID	*Mean difference* in consensus length*	*Min-Max of difference in consensus length*	*% of consensus sequences longer then GISAID reference*	*% of consensus sequences with different nucleotide call*	*Mean** of n° of different nucleotide call*
CLC	*-137*	*-544 +47*	*2 (4%)*	*28 (56%)*	*4*
SARS-CoV-2 RECoVERY	*54*	*-2 +106*	*48 (96%)*	*48 (93.7%)*	*7*
Genome Detective	*-5172*	*-18454 +11*	*0 (0%)*	*49 (99.9%)*	*41*



Illumina data (n° of analysed runs =100)					
GISAID	Mean difference* in consensus length	Min-Max of difference in consensus length	% of consensus sequences longer then GISAID reference	% of consensus sequences with different nucleotide call	Mean** of n° of different nucleotide call
CLC	-1173	-8345 +643	18 (18%)	20 (20%)	7
SARS-CoV-2 RECoVERY	135	-1652 +3379	73 (73%)	52 (52%)	5
Genome Detective	-167	-8925 +1989	40 (40%)	43 (43%)	16
Nanopore data (n° of analysed runs =100)					
SARS-CoV-2 RECoVERY	569	-169 +2444	97 (97%)	96 (96%)	7
Genome Detective	267	-3816 +2127	91 (91%)	90 (90%)	13

Conclusioni

1. Pipeline con un'interfaccia user-friendly

2. Rapidità di analisi: 10-60 minuti in base al numero di reads della corsa (50mila - 6 milioni di sequenze).

3. Indipendenza delle analisi dalla piattaforma di sequenziamento

4. Prestazioni comparabili se non migliori dei software disponibili.

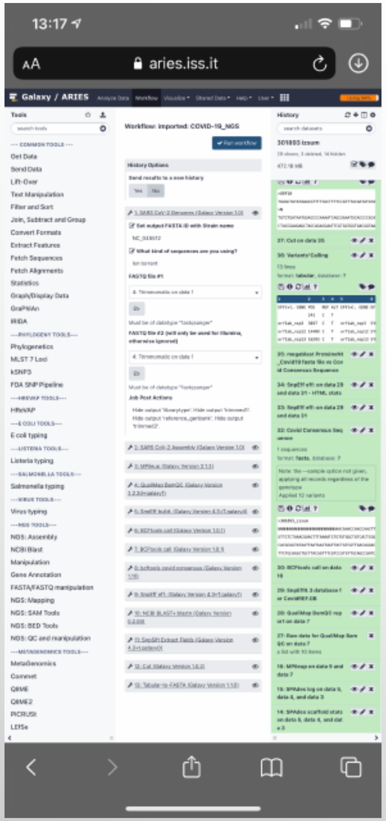
5. Risultati immediatamente utilizzabili per ulteriori analisi

6. Report con tutte le varianti caratterizzate, fanno di questa pipeline uno strumento prezioso soprattutto per scienziati con poca o nessuna competenza in bioinformatica

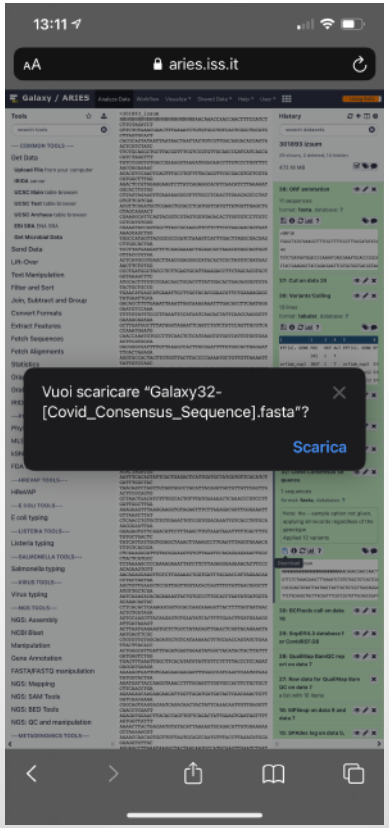
7. **Analisi eseguibili indipendentemente dall'hardware degli utenti utilizzando qualsiasi browser da desktop**

8. **Fornire un servizio alla comunità scientifica per aumentare la conoscenza sull'evoluzione della SARS-CoV-2**

CURIOSITÀ: SMARTPHONE



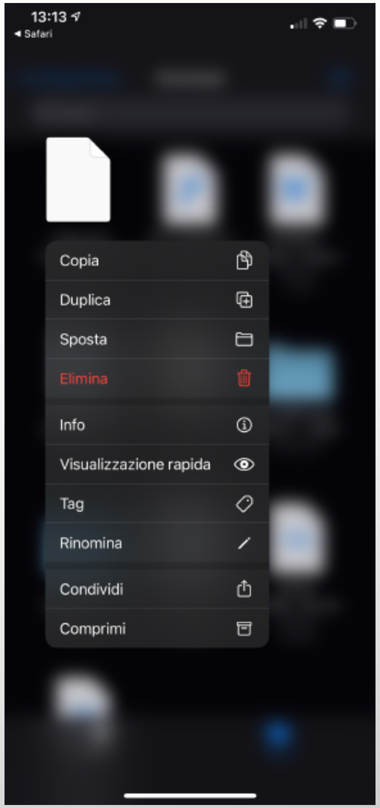
Run workflow



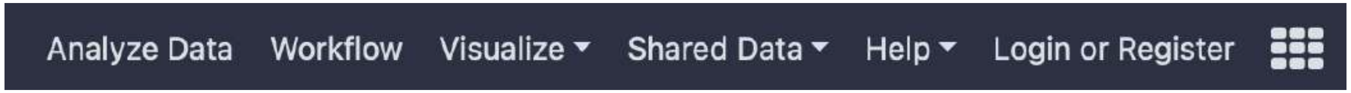
Vuoi scaricare "Galaxy32-[Covid_Consensus_Sequence].fasta"?

Scarica

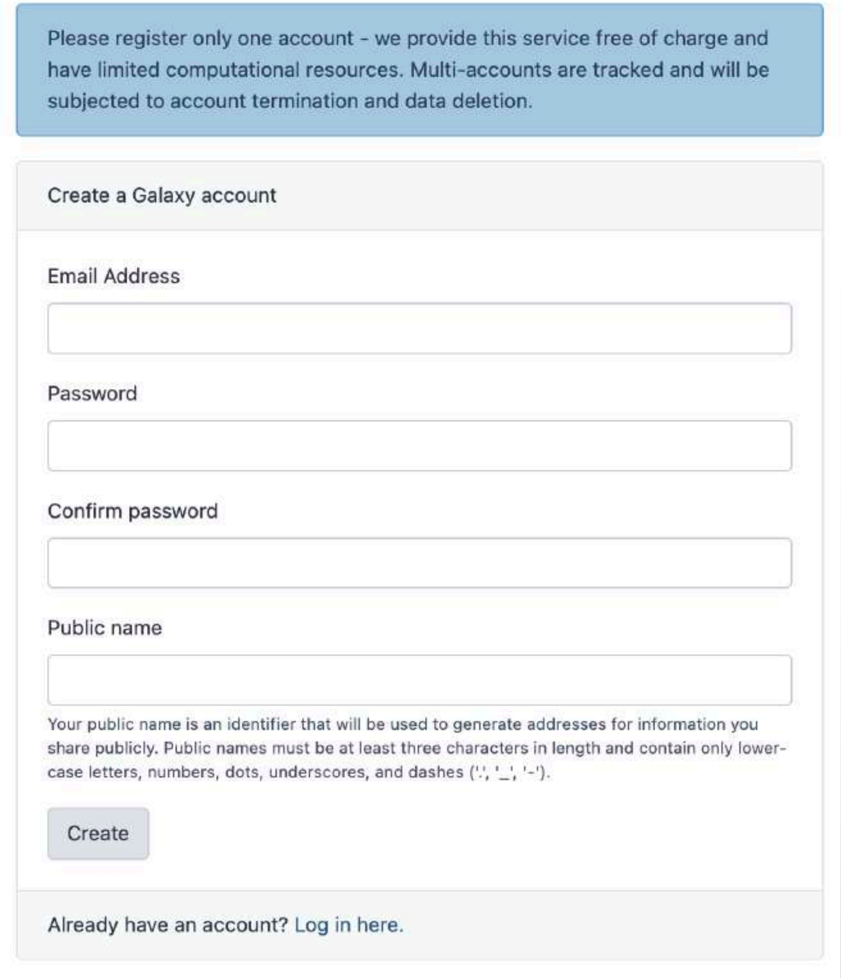
Download file



Send results



https://aries.iss.it



Please register only one account - we provide this service free of charge and have limited computational resources. Multi-accounts are tracked and will be subjected to account termination and data deletion.

Create a Galaxy account

Email Address

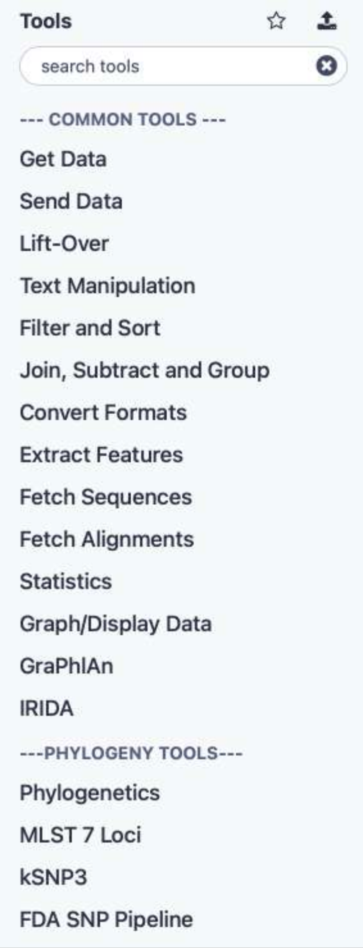
Password

Confirm password

Public name

Your public name is an identifier that will be used to generate addresses for information you share publicly. Public names must be at least three characters in length and contain only lower-case letters, numbers, dots, underscores, and dashes ('.', '_', '-').

Already have an account? Log in here.



Tools

search tools

--- COMMON TOOLS ---

- Get Data
- Send Data
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Statistics
- Graph/Display Data
- GraPhlAn
- IRIDA

---PHYLOGENY TOOLS---

- Phylogenetics
- MLST 7 Loci
- kSNP3
- FDA SNP Pipeline

Grazie dell'attenzione

a voi………

ISTITVTO SVPERIORE DI SANITA

e colleghi

Laboratorio Europeo di Riferimento per *E. coli:*
Stefano Morabito
Knijn Arnold

Reparto Zoonosi Emergenti:
Gabriele Vaccari
Ilaria Di Bartolo
Giovanni Ianiro