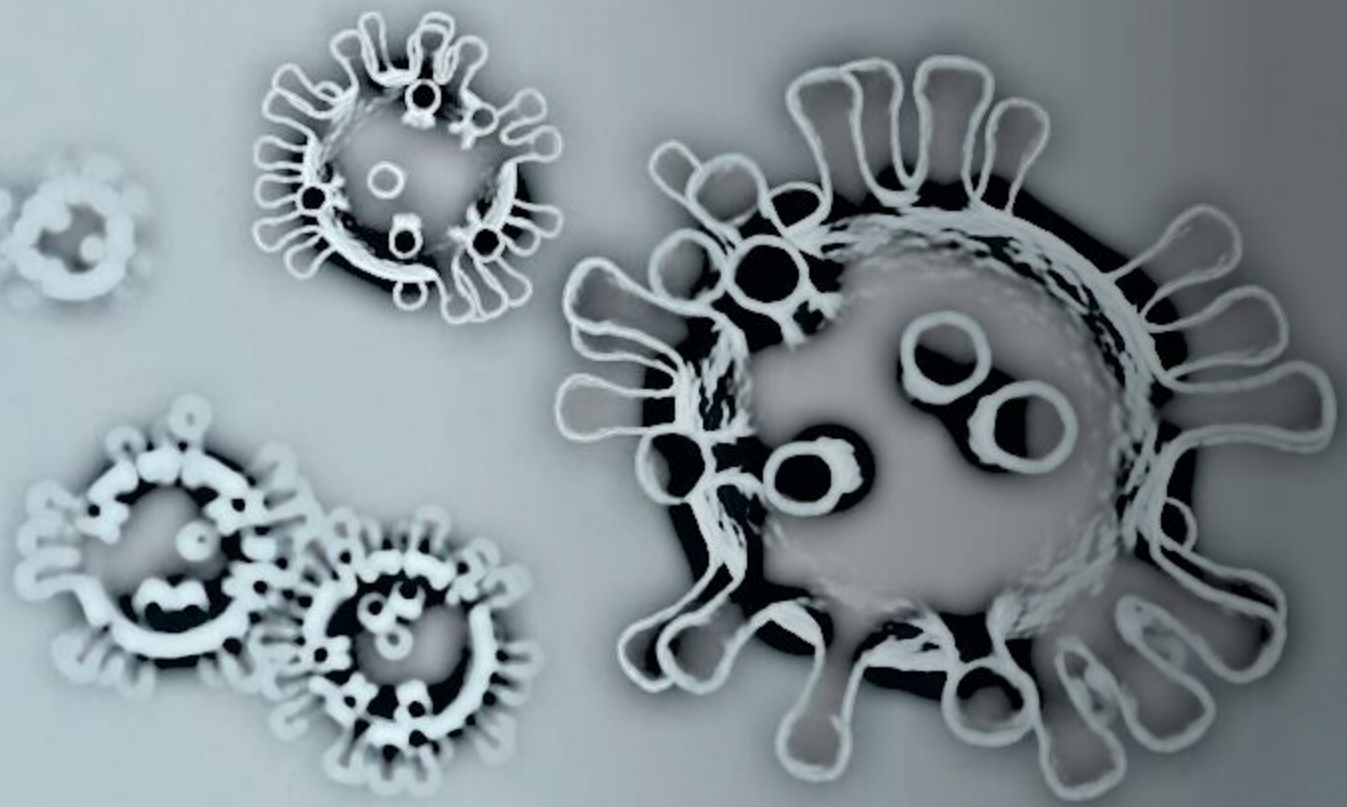
EMPOWERING VIRUS SEQUENCE RESEARCH THROUGH CONCEPTUAL MODELING

ANNA BERNASCONI, ARIF CANAKOGLU,
PIETRO PINOLI, STEFANO CERI

DEIB, POLITECNICO DI MILANO



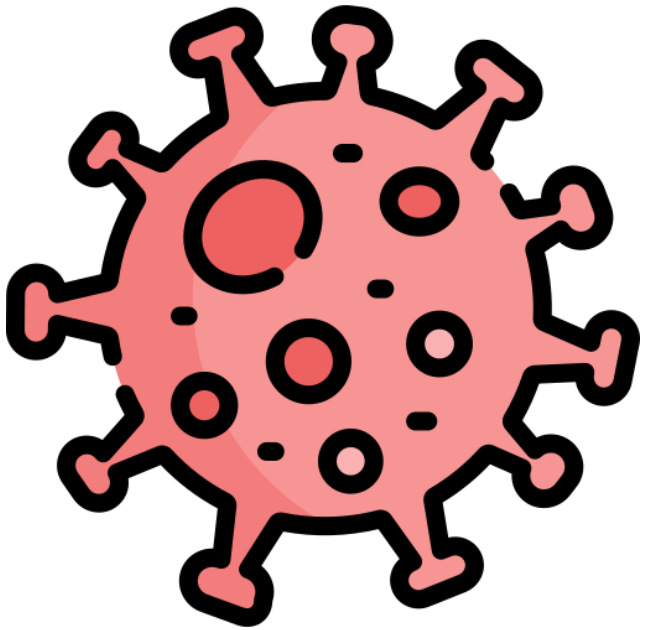
ER 2020 – ONLINE EVENT

WHAT NEEDS ARE WE RESPONDING TO?

UNPRECEDENTED ATTENTION TOWARDS THE GENETIC MECHANISMS OF VIRUSES
(caused by the pandemic outbreak of the coronavirus disease COVID-19)

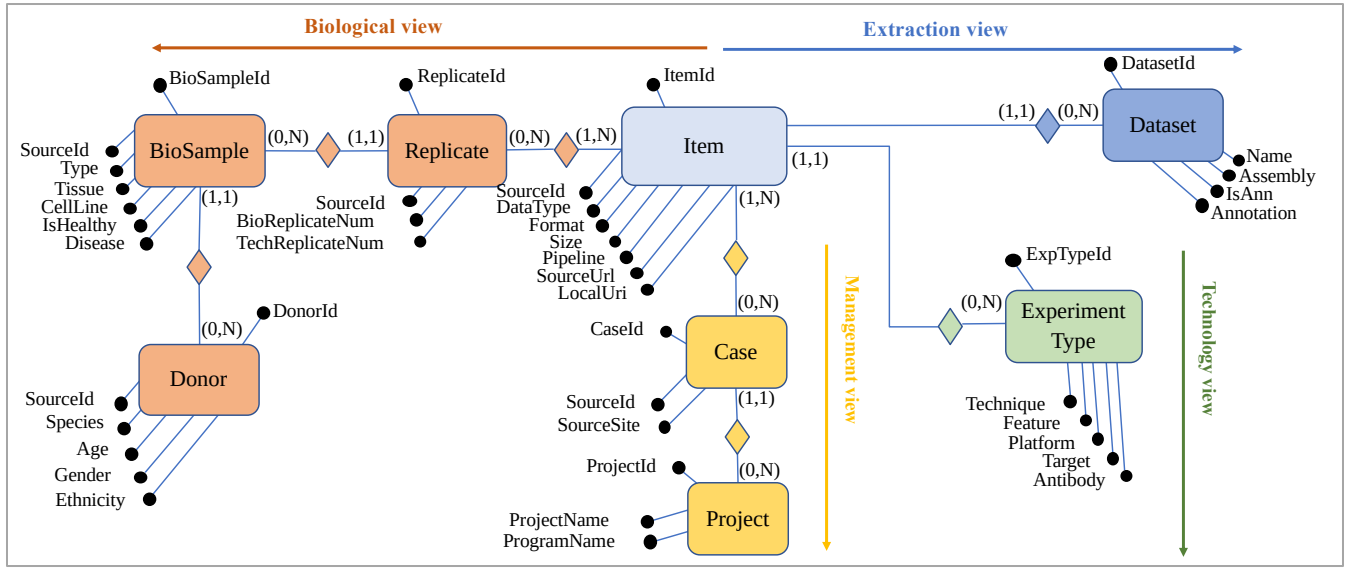
LACK OF PREPARATION OF THE RESEARCH COMMUNITY TO FACE PANDEMIC CRISES
(e.g., lack of **well-organized databases and search systems**)

NEED FOR FACILITATING CURRENT AND FUTURE RESEARCH STUDIES
(we provide a novel **conceptual model,** repository and search system collecting virus sequences and their properties)



OUR BACKGROUND

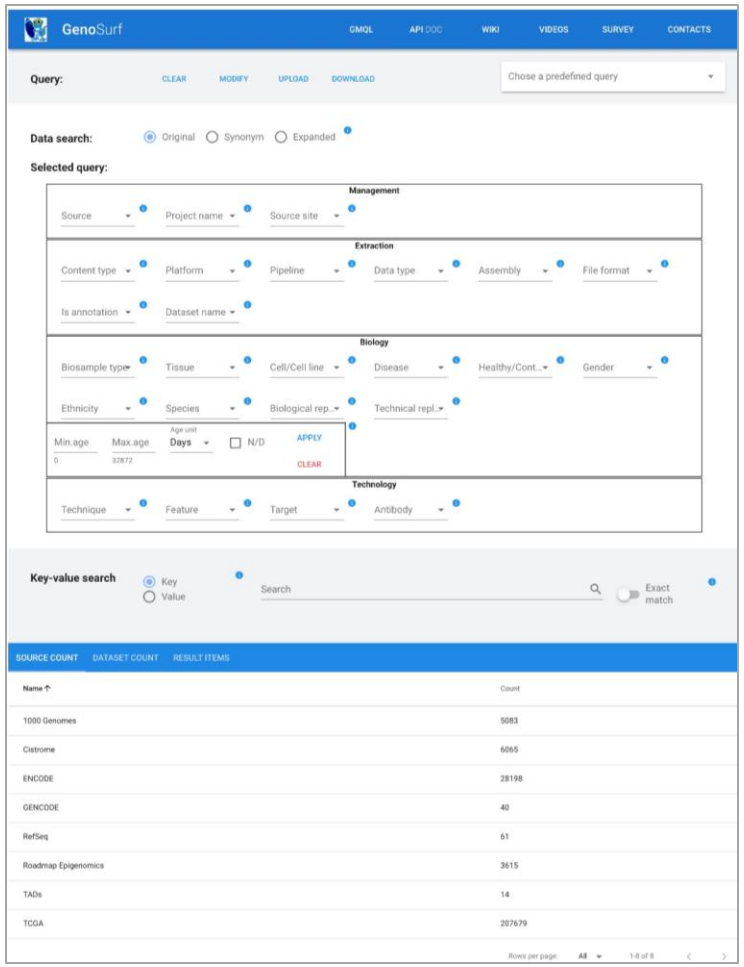
Genomic Conceptual Model



Bernasconi et al. «Conceptual Modeling for Genomics: Building an Integrated Repository of Open Data». ER 2017. https://doi.org/10.1007/978-3-319-69904-2_26

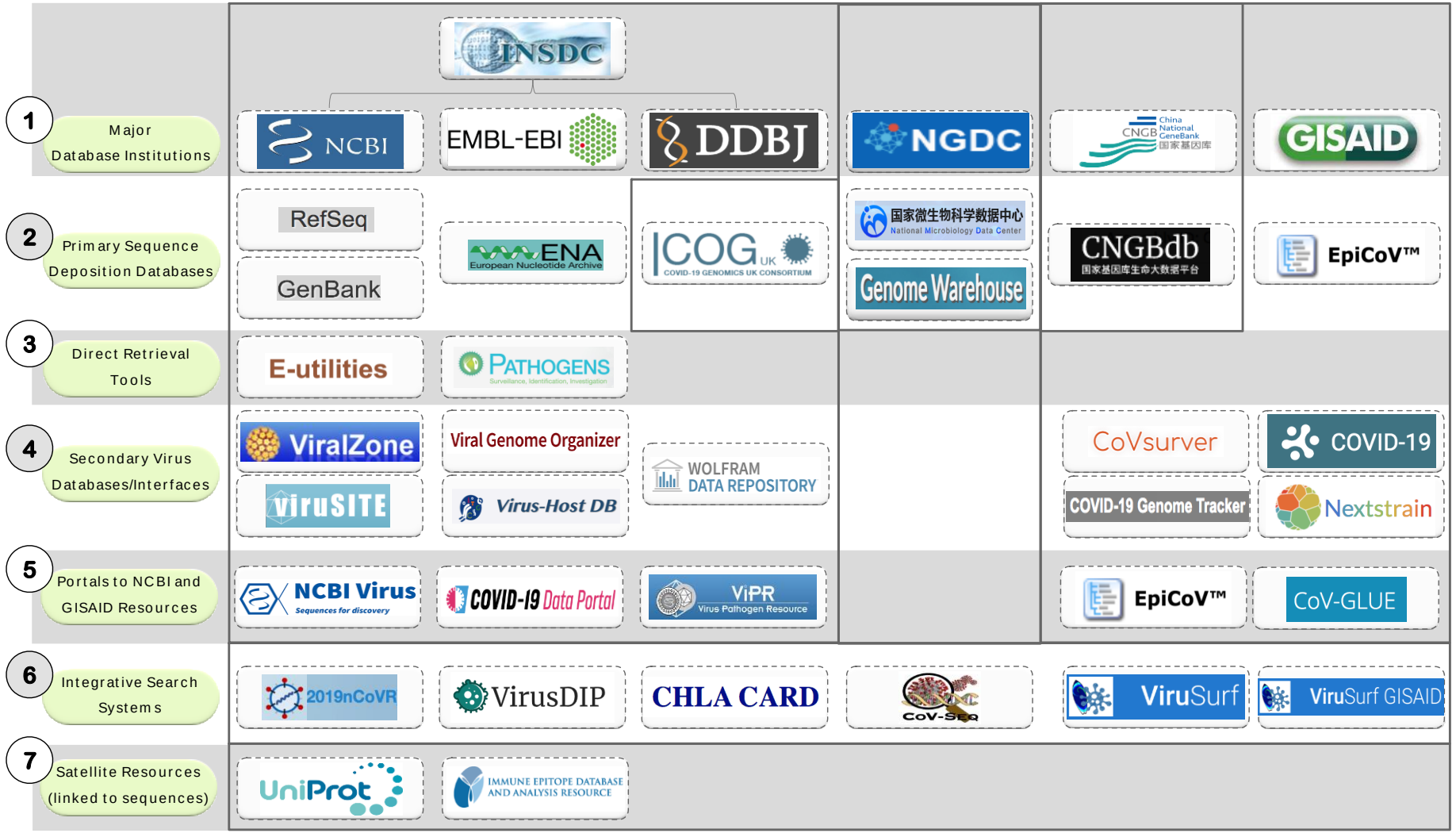
GenoSurf interface

http://gmql.eu/genosurf/



Canakoglu et al. «GenoSurf: metadata driven semantic search system for integrated genomic datasets». Database, Volume 2019, 2019, baz132, https://doi.org/10.1093/database/baz132

BACKGROUND ANALYSIS: VIRUS RESOURCES SCENARIO



BACKGROUND ANALYSIS: AVAILABLE METADATA

in SARS-CoV2 search engines

GISAID	Searchable	Table	Individual	CoV-GLUE	Table	2019nCoVR	Searchable	Table	Individual	ViruSurf	Searchable
Virus name	X	X	X	Virus name	X	Virus Strain Name		X	X	StrainName	X
Accession ID	X	X	X	GISAID ID	X	Accession ID		X	X	AccessionId	X
Type			X							Genus	X
Passage details/history		X	X								
Collection date	X	X	X	Collection date	X	Sample Collection Date	X	X	X	CollectionDate	X
Location	X	X	X	Country/Location	X	Country/Region/Province/City/Locatio	X	X	X	Country, Region, GeoGroup	X
Host	X	X	X			Host	X	X	X	Species	X
Additional location information			X								
Gender			X								
Patient age			X								
Patient status			X								
Specimen source			X							IsolationSource	X
Additional host information			X								
Outbreak			X								
Last vaccinated			X								
Treatment			X								
Sequencing Technology			X							SequencingTechnology	X
Assembly method			X							AssemblyMethod	X
Coverage	X		X							Coverage	X
Comment		X	X								
Originating lab		X	X	Originating lab	X	Originating Lab		X	X	OriginatingLab	X
Address			X								
Sample ID given by the sample provider			X								
Submitting lab		X	X	Submitting lab	X	Submitting lab		X	X	SequencingLab	X
Address			X								
Sample ID given by the submitting laboratory			X								
Authors			X	Authors	X						
Submitter			X								
Submission date	X	X	X			SubmissionDate		X		SubmissionDate	X
Address			X								
Complete	X					Nuc.Completeness	X	X	X	IsComplete	X
Length		X				Sequence Length	X	X		Length	X
						Data Source	X	X	X	DatabaseSource	X
						Sequence Quality	X	X			
						Quality Assessment	X	X	X		
										IsReference	X
										GC%	X
										BioprojectId	X

BACKGROUND ANALYSIS: REQUIREMENTS COLLECTION

Extensive **interviews to groups of virologists** of various specializations:

Ilaria Capua - One Health Center of Excellence (University of Florida, US)
Matteo Chiara - Università degli Studi di Milano Statale (IT)
Ana Conesa - University of Florida (US)
Luca Ferretti - Oxford Big Data Institute (UK)
Alice Fusaro - Istituto Zooprofilattico Sperimentale delle Venezie (IT)
Ruba Al Khalaf - Politecnico di Milano (IT)
Susanna Lamers - BioInfoExperts (Louisiana, US)
Stefania Leopardi - Istituto Zooprofilattico Sperimentale delle Venezie (IT)
Alessio Lorusso - Istituto Zooprofilattico Sperimentale Abruzzo Molise (IT)
Francesca Mari - Università di Siena (IT)
Carla Mavian - Department of Pathology, College of Medicine (University of Florida, US)
Graziano Pesole - Università di Bari (IT)
Alessandra Renieri - Università di Siena (IT)
Anna Sandionigi - Università degli Studi di Milano-Bicocca (IT)
Stephen Tsui - The Chinese University of Hong Kong (HK)
Limsoon Wong - National University of Singapore (SGP)
Federico Zambelli - Università degli Studi di Milano Statale (IT)

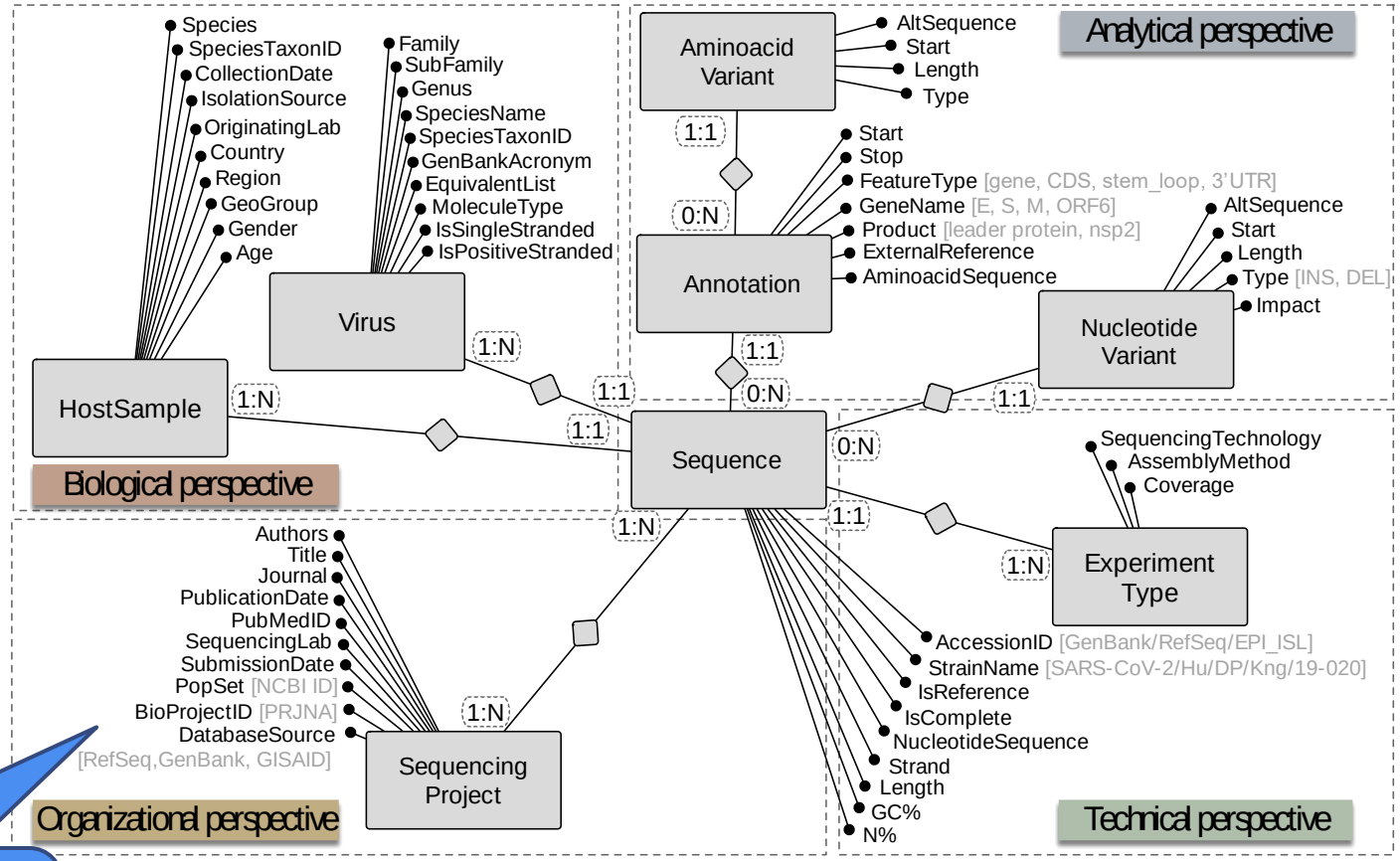
Each researcher provided us with **a viewpoint on applications of virology** that serve as **requirements** for **progressively adding** relevant **features** to our database as well as relevant search services to comply with their needs:

- Diagnosis
- Vaccine development
- Drug-resistance and drug-resistance associated mutations

PROPOSED CONCEPTUAL MODEL

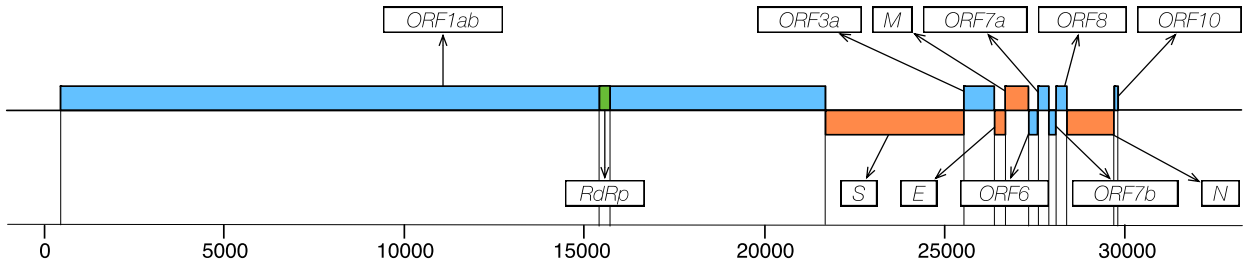
The **Viral Conceptual Model (VCM),** centered on the virus **sequence** described from four perspectives:

- ***biological perspective*** (virus species and host environment)
- ***technological perspective*** (sequencing technology)
- ***organizational perspective*** (project responsible for producing the sequence)
- ***analytical perspective*** (properties of the sequence, such as known annotations and variants)



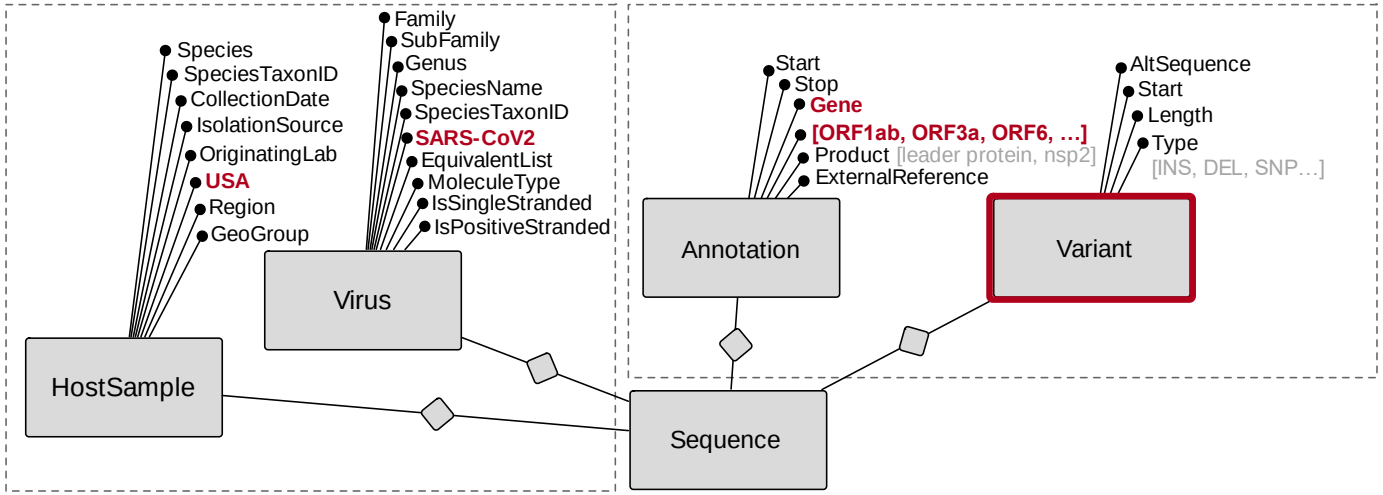
The schema is general and applies to any virus.

EXAMPLE QUERY

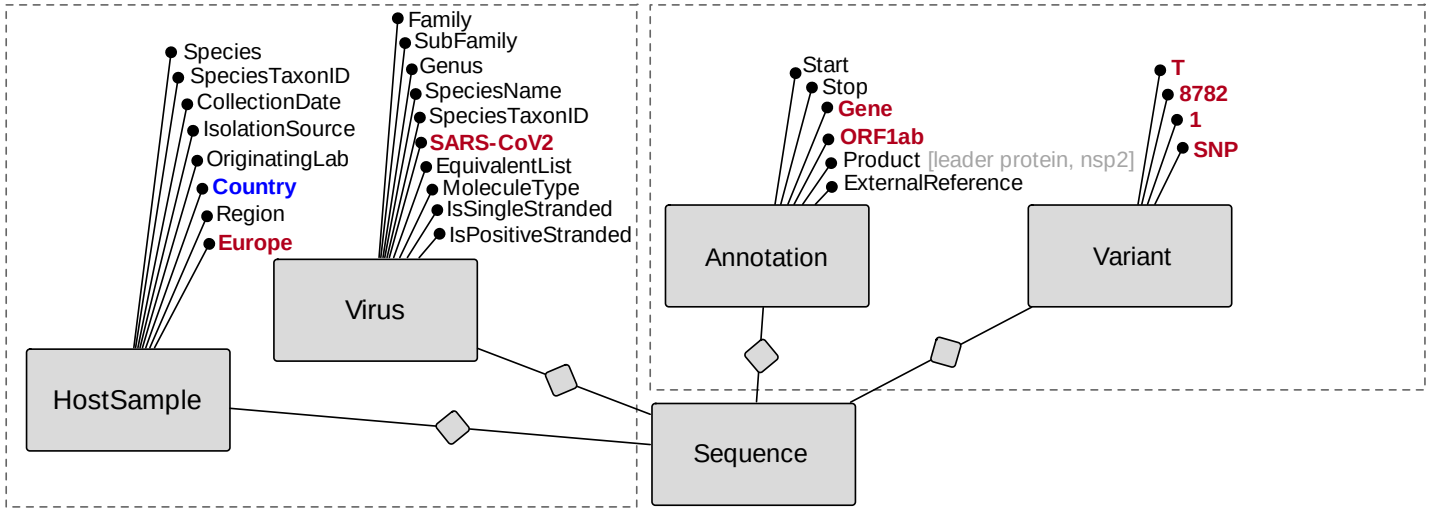


Application on SARS-CoV2 virus: complex conceptual queries upon VCM are able to replicate the search results of recent articles, hence demonstrating huge potential in supporting research upon viruses

Extract SARS-CoV2 sequences from samples of US patients that present nucleotide variants in genes that codify for open reading frames.



EXAMPLE QUERY



Select sequences from European patients affected by a SARS-CoV2 virus, only if they do not have a specific variant on the first gene (ORF1ab), selected by using the triple <*position, alternative_sequence, type*> (e.g., 8,782 SNP from C to T).

```
ANNOTATION: GeneName in [ORF1ab]
VARIANT: Start=8782, AltSequence=T, Type=SNP
count=0
```

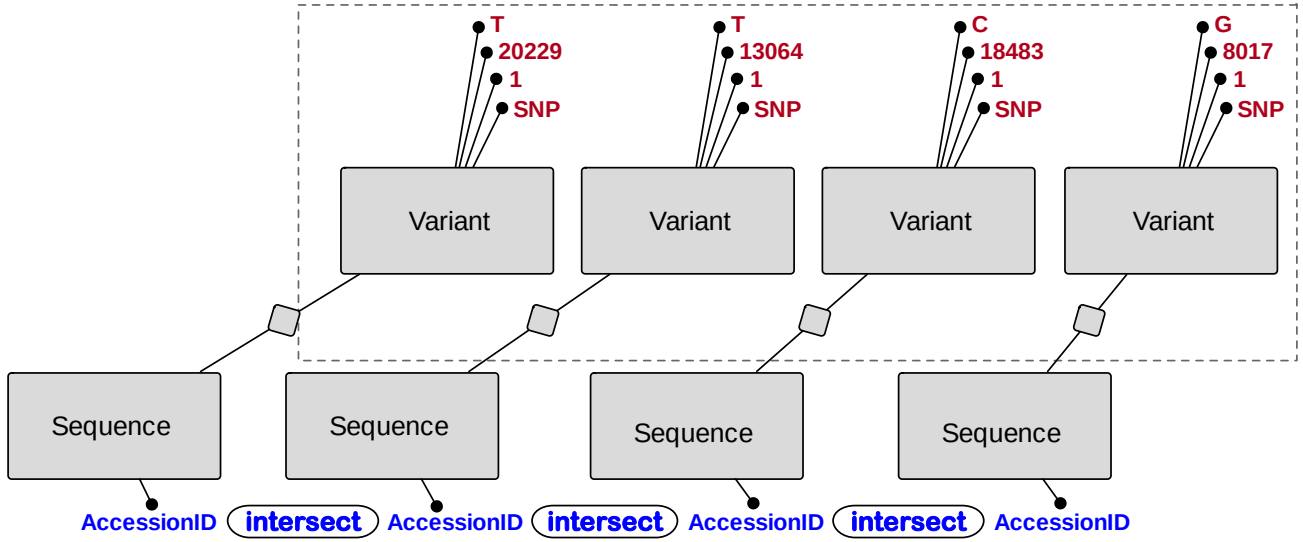
and

```
HOST_SAMPLE: GeoGroup in [Europe]
```

EXAMPLE QUERY

In Gudbjartsson *et al.* (2020), specific sequence variants are used to define clades/haplogroups (e.g., the A group is characterized by the 20,229 and 13,064 nucleotides, originally C mutated to T, by the 18,483 nucleotide T mutated to C, and by the 8,017, from A to G).

Select sequences with all four variants corresponding to the A clade group defined in Gudbjartsson *et al.* (2020).



Clade	Pos	Ref	Alt
B	28144	T	C
B1	18060	C	T
B1a	17858	A	G
B1a1	17747	C	T
B1a1a	24694	A	T
B1a1a1	9445	T	C
B1a1a1a	17531	T	C
B1a1a1a	18756	G	T
B1a1a1b	29140	G	T
B4	28878	G	A
B4	29742	G	A
B2	29095	C	T
A	20229	C	T
A	13064	C	T
A	18483	T	C
A	8017	A	G
⋮	⋮	⋮	⋮

VARIANT: *Start*=8017, *AltSequence*=G, *Type*=SNP

and

VARIANT: *Start*=13064, *AltSequence*=T, *Type*=SNP

and

VARIANT: *Start*=18483, *AltSequence*=C, *Type*=SNP

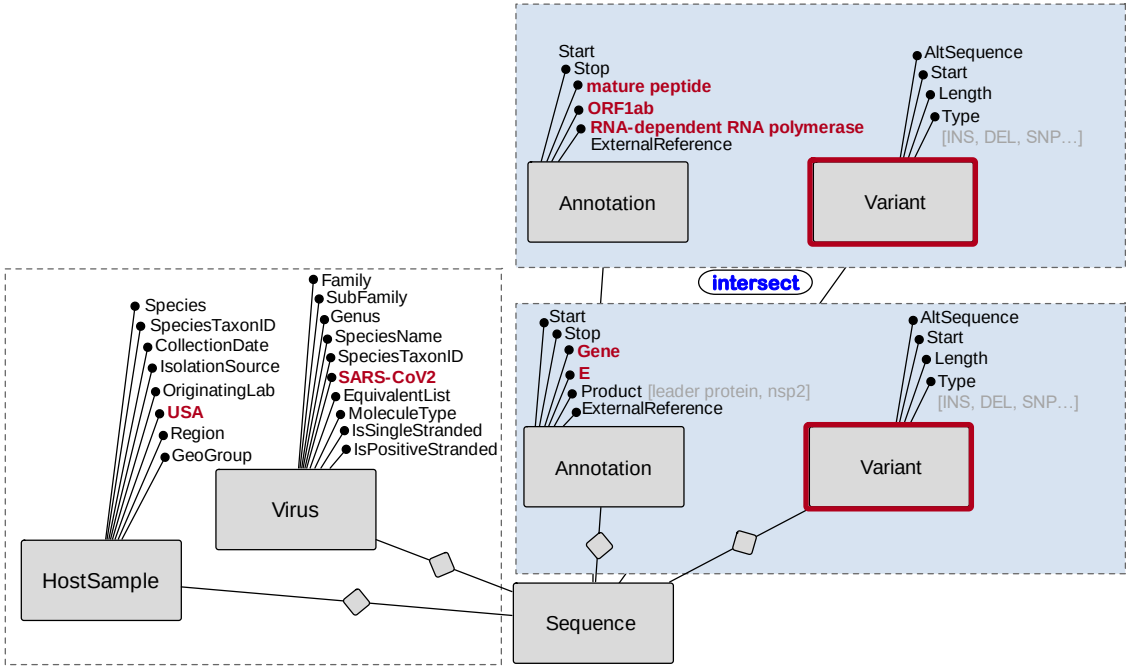
and

VARIANT: *Start*=20229, *AltSequence*=T, *Type*=SNP

EXAMPLE QUERY

According to Corman et al. (2020), E and RdRp genes are highly mutated and thus crucial in diagnosing COVID-19 disease; first-line screening tools of 2019-nCoV should perform an E gene assay, followed by confirmatory testing with the RdRp gene assay.

Retrieve all sequences with mutations within genes E and RdRp of humans affected in China.




```
ANNOTATION: GeneName in [E], FeatureType=gene
VARIANT: count>0
```

and

```
ANNOTATION: GeneName in [ORF1ab], FeatureType=mat_peptide,
      Product=RNA-dependent RNA polymerase
VARIANT: count>0
```

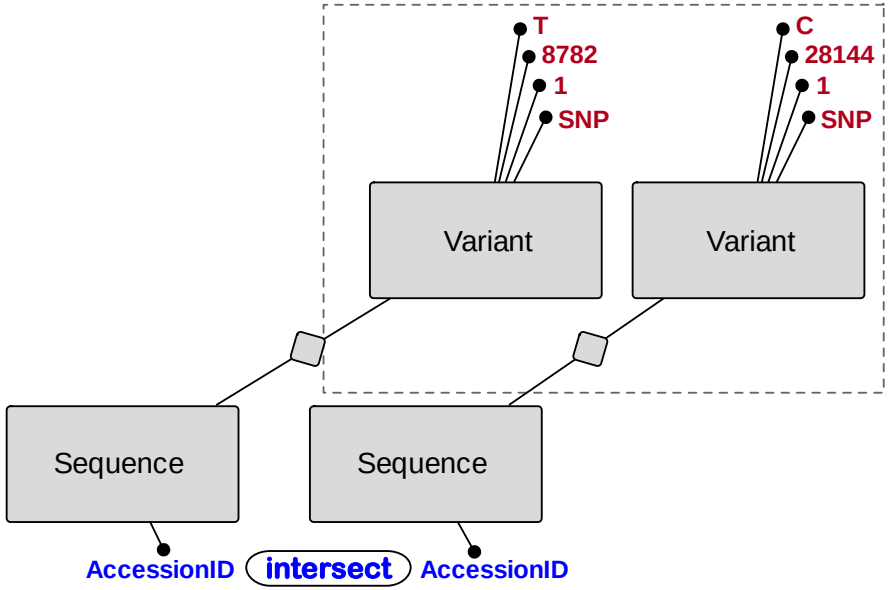
and

```
HOST_SAMPLE: Country in [China]
```

EXAMPLE QUERY

Tang *et al.* (2020) claim that there are two clearly definable "major types" (S and L) of SARS-CoV2 in this outbreak, that can be differentiated by transmission rates. S and L types can be distinguished by two SNPs at positions 8,782 (within the ORF1ab gene from C to T) and 28,144 (within ORF8 from T to C).

Retrieve all sequences with these two SNPs.



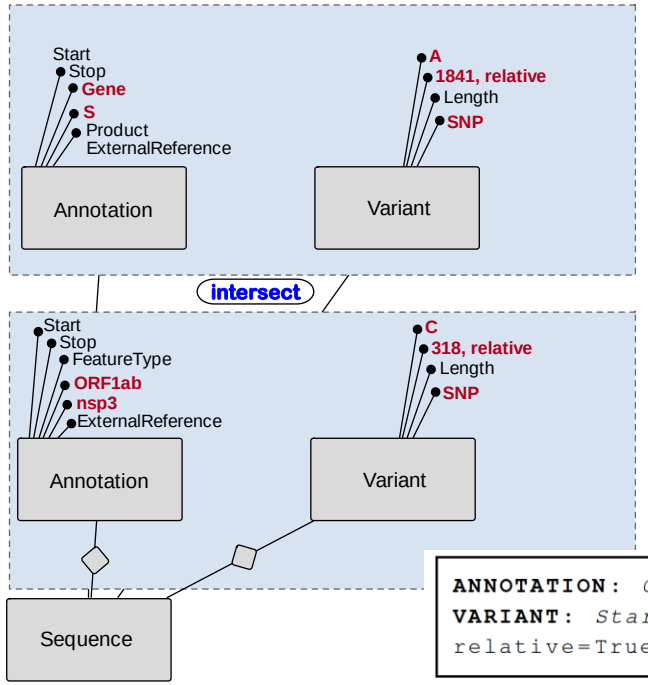

```
VARIANT: Start=8782, AltSequence=T, Type=SNP
```

and

```
VARIANT: Start=28144, AltSequence=C, Type=SNP
```

EXAMPLE QUERY

Morais Junior *at al.* (2020) propose a subdivision of the global SARS-CoV2 population into sixteen subtypes, defined using "widely shared polymorphisms" identified in nonstructural (nsp3, nsp4, nsp6, 27 nsp12, nsp13 and nsp14) cistrons, structural (spike and nucleocapsid), and accessory (ORF8) genes. Extract sequences from subtype I.




```
ANNOTATION: GeneName in [S], FeatureType=gene
VARIANT: Start=1841, AltSequence=A, Type=SNP
relative=True
```

and

```
ANNOTATION: GeneName in [ORF1ab], Product in [nsp3]
VARIANT: Start=318, AltSequence=C, Type=SNP
relative=True
```

and

...

SUBTYPE	N*	WSP POSITION											
		nsp3	*nsp4*	*nsp6*	*nsp12*	*nsp13*		*nsp14*	*S*	*ORF8*	*N*		
		#318	#228	#111	#967	#1511	#1622	#21	#1,841	#251	#608	#609	#610
I	132	C [Phe]†	C [Ser]	G [Leu]	C [Pro]	C [Pro]	A [Tyr]	C [Leu]	A [Asp]	U [Leu]	G [Arg]	G [Arg]	G [Arg]
II	122	U [Phe]	C [Ser]	G [Leu]	U [Leu]	C [Pro]	A [Tyr]	C [Leu]	G [Gly]	U [Leu]	G [Arg]	G [Arg]	G [Arg]
III	101	C [Phe]	U [Ser]	G [Leu]	C [Pro]	U [Leu]	G [Cys]	U [Leu]	A [Asp]	C [Ser]	G [Arg]	G [Arg]	G [Arg]

IMPLEMENTATION

We integrate public data from different DNA/RNA sequences with their annotation. We enrich it with variation data (i.e., mutations) computed with a sequence alignment algorithm.

Sources for http://gmql.eu/virusurf/:
- SARS-CoV2 and SARS-CoV from GenBank/RefSeq ~ 8K sequences
(available through E-utilities API NCBI nuccore db)

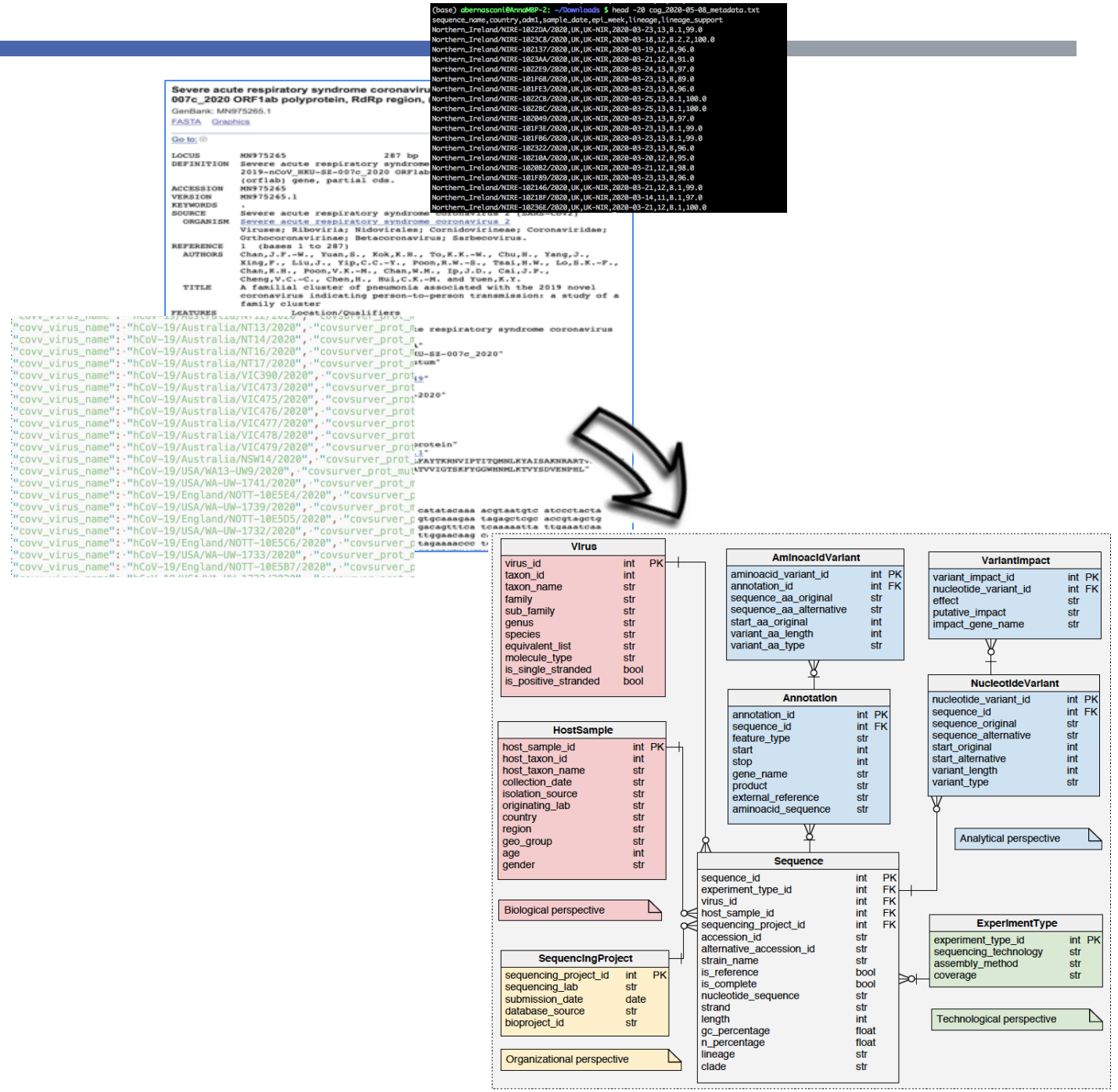
- COG-UK ~ 16K sequences

Sources for http://gmql.eu/virusurf_gisaid/:

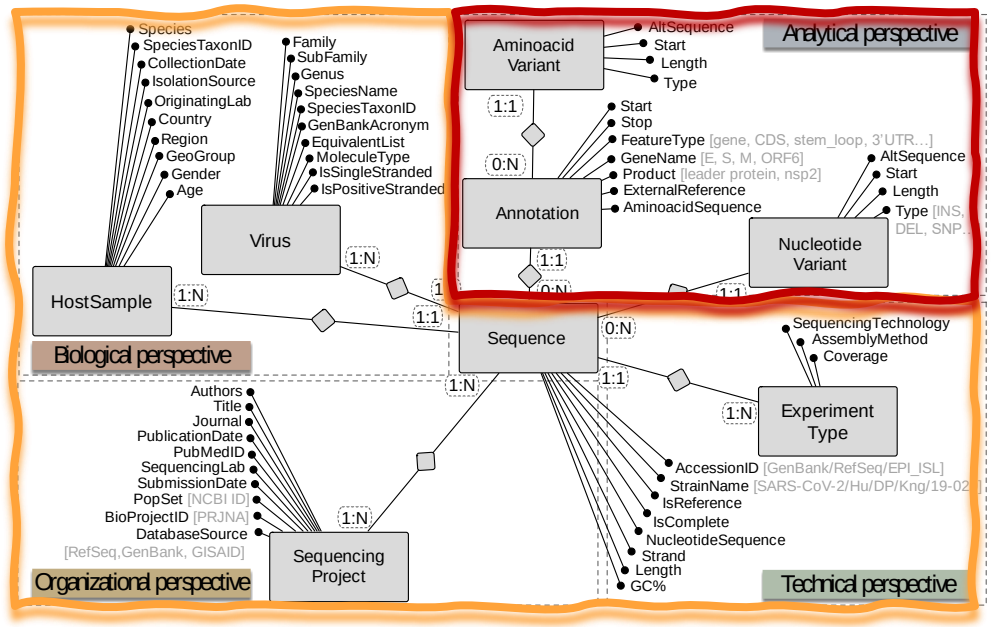
- GISAID EpiCoV™ db ~ 57K sequences (available through special agreement)

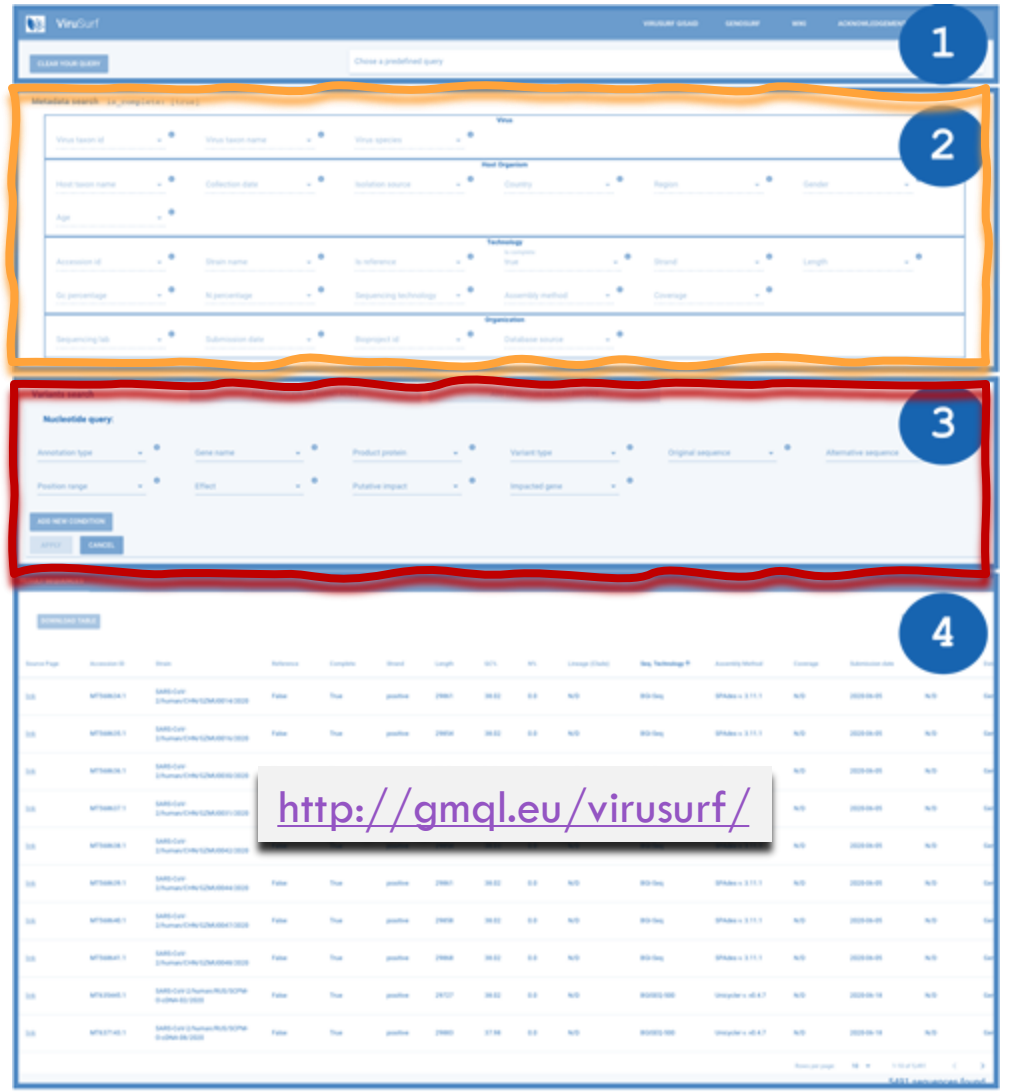
Input formats:
XML, JSON, TSV

Output format:
Relational database



FROM THE CM TO THE SEARCH SYSTEM





Arif Canakoglu, Pietro Pinoli, Anna Bernasconi, Tommaso Alfonsi, Damianos P Melidis, Stefano Ceri. "ViruSurf: an integrated database to investigate viral sequences". *Nucleic Acids Research, gkaa846,* https://doi.org/10.1093/nar/gkaa846

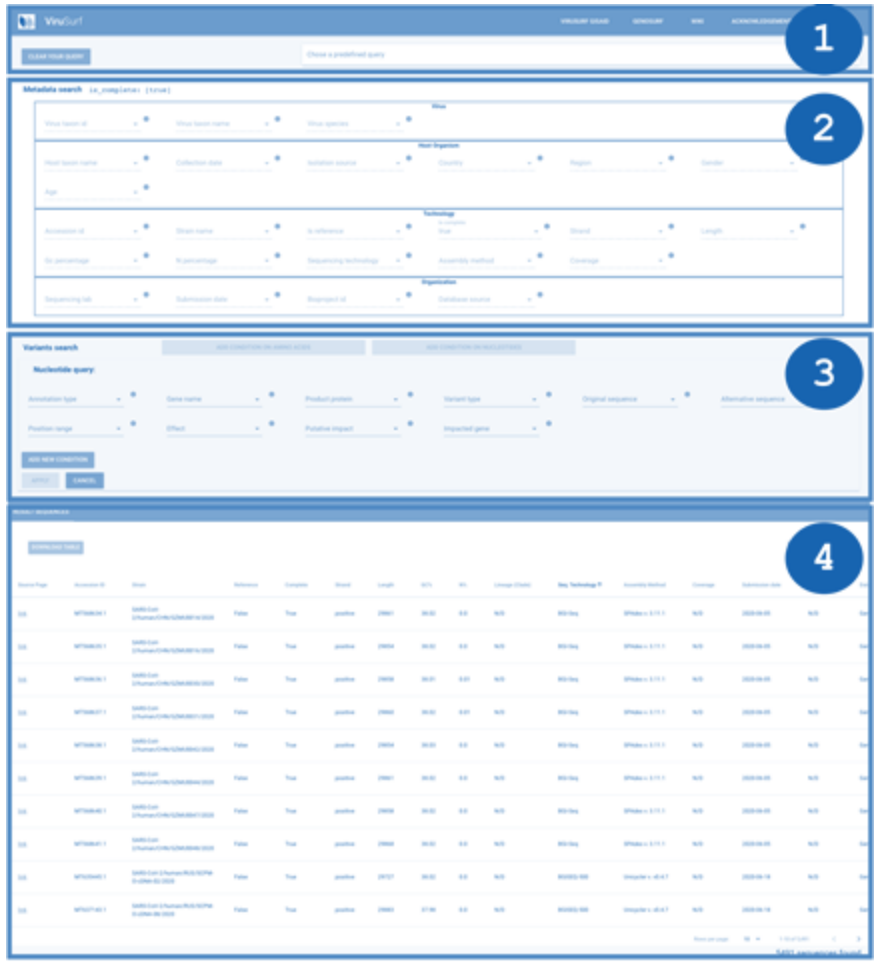
VIRUSURF INTERFACE

http://gmql.eu/virusurf/

The interface is composed of 4 sections:

1) a menu bar to access the different services, documentation and query utilities;
2) the search interface over the metadata attributes;
3) the search interface over annotations and nucleotide/amino acid variant information;
4) a result visualization section, showing a flexible table with the resulting sequences, described by their metadata.

The interface enables an interplay between search performed within parts (2) and (3), thereby allowing to build complex queries given as the logical conjunction - of arbitrary length - of filters set in (2) and in (3).



Arif Canakoglu, Pietro Pinoli, Anna Bernasconi, Tommaso Alfonsi, Damianos P Melidis, Stefano Ceri. "ViruSurf: an integrated database to investigate viral sequences". *Nucleic Acids Research, gkaa846, https://doi.org/10.1093/nar/gkaa846*

EXAMPLE CASE ON VIRUSURF

(Pachetti et al., 2020)

mutation located in SARS-CoV2 gene N at position 28881

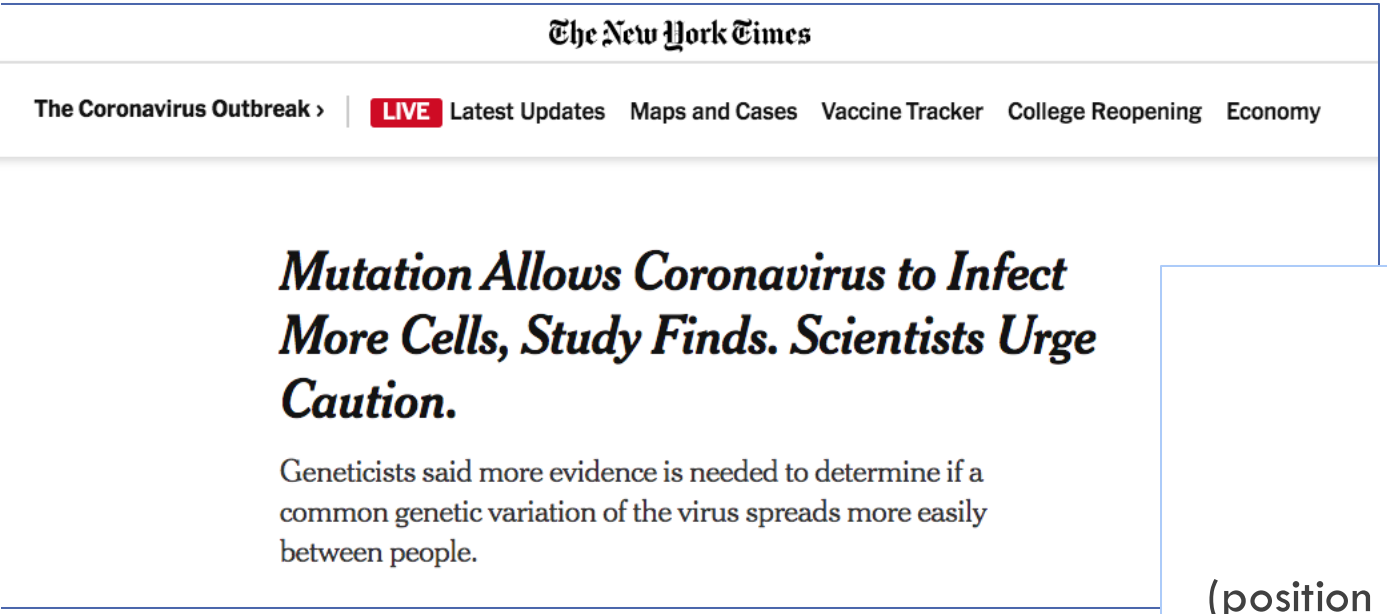
related to a double codon mutation

inducing the substitution of two amino acids: 28881 (R to K) and 28881 (G to R)

https://youtu.be/_jjwK04eE6s

Pachetti, M., Marini, B., Benedetti, F., Giudici, F., Mauro, E., Storici, P., Masciovecchio, C., Angeletti, S., Ciccozzi, M., Gallo, R.C. and Zella, D., 2020. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *Journal of Translational Medicine*, *18*, pp.1-9.

EXAMPLE CASE ON VIRUSURF



The New York Times

The Coronavirus Outbreak › | LIVE Latest Updates Maps and Cases Vaccine Tracker College Reopening Economy

Mutation Allows Coronavirus to Infect More Cells, Study Finds. Scientists Urge Caution.

Geneticists said more evidence is needed to determine if a common genetic variation of the virus spreads more easily between people.

Source:
https://www.nytimes.com/2020/06/12/science/coronavirus-mutation-genetics-spike.html

(Zhang et al., 2020)

SARS-CoV-2 viruses

with D614G mutation in Spike protein

(position 614 from D (Aspartic acid) to G (Glycine) amino acids)

seem to infect a cell more likely

than viruses without that mutation

G614 genotype:
- not detected in February
- found with low frequency in March
- increased rapidly from April onward

→ indicating a transmission advantage over viruses with D614

https://youtu.be/IJcflefxzzM

Zhang, L., Jackson, C. B., Mou, H., Ojha, A., Rangarajan, E. S., Izard, T., Farzan, M., & Choe, H. (2020). **The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity.** *bioRxiv : the preprint server for biology*, 2020.06.12.148726.

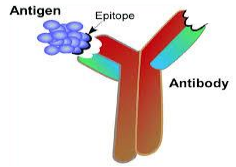
RESULTS FROM QUERIES ON VIRUSURF

	ViruSurf	ViruSurf-GISAID	ViruSurf	ViruSurf-GISAID
	≤ 31/03/2020		≥ 01/04/2020	
With D614G	6,592	15034	23,649	18,421
Without D614G	4,664	8821	3,331	3369
D614%	58.56%	63.02%	87.65%	84.54%
total	61.59%		86.26%	

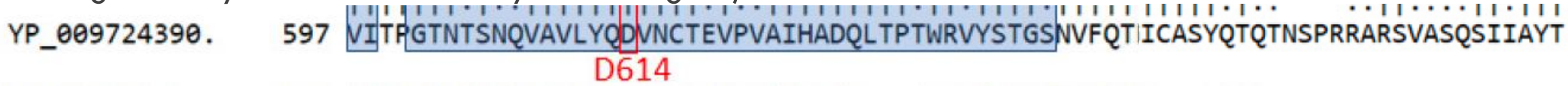
Zhang, L., Jackson, C. B., Mou, H., Ojha, A., Rangarajan, E. S., Izard, T., Farzan, M., & Choe, H. (2020). **The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity.** *bioRxiv : the preprint server for biology*, 2020.06.12.148726.

FROM THE SEARCH SYSTEM TO USEFUL APPLICATIONS

QUERIES USEFUL FOR IDENTIFYING VACCINE PROPERTIES



→ Building an extension of VCM/ViruSurf that includes **EPITOPES** (short amino acid sequences that are recognized by the host immune system antigens)

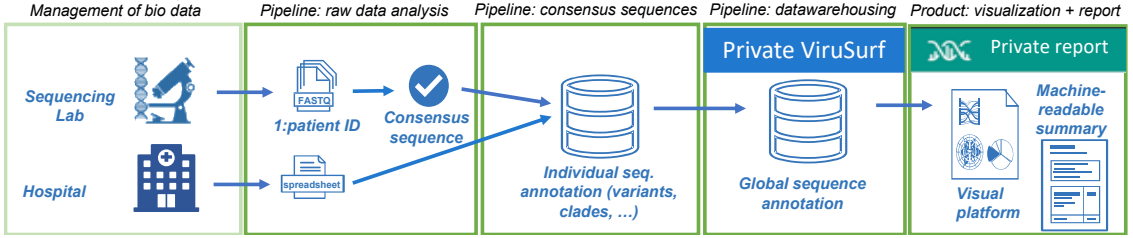


QUERIES USEFUL FOR UNDERSTANDING IMPACT OF VIRUS MUTATIONS

→ Building a "knowledge base" of variants linked to their correlation with clinical and epidemiological impacts (e.g., disease severity, virus transmissibility, antigenicity, protein stability…)

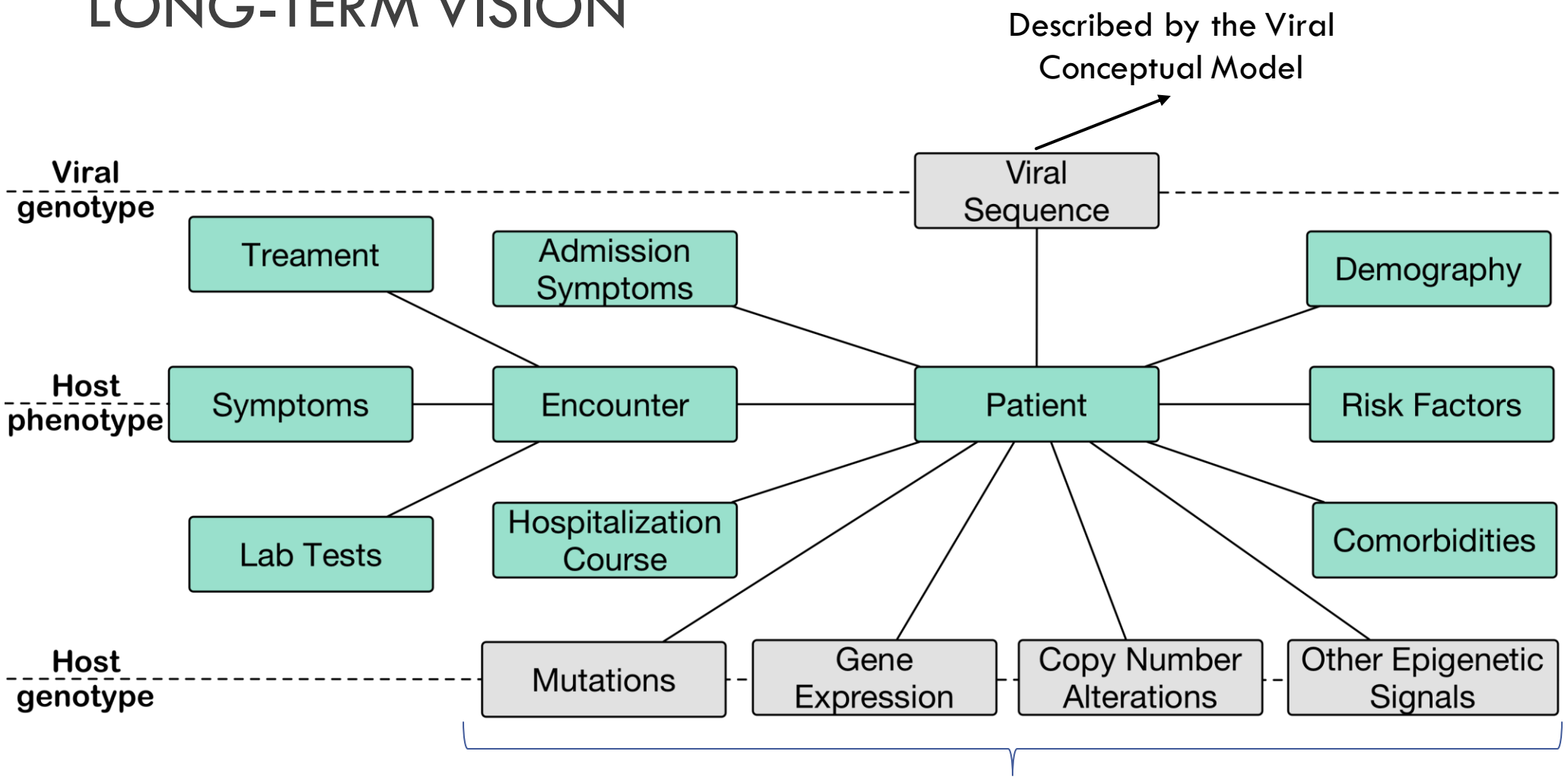
PACKAGING OF SERVICES FOR CONFIDENTIAL USE BY HOSPITALS (and others)

→ We allow users that cannot share their data to use our database and knowledge base functionalities



Ongoing project funded by EIT Digital innovation activity "DATA against COVID-19"

LONG-TERM VISION



http://gmql.eu/virusurf/

ViruSurf

e.g., data dictionary of https://www.covid19hg.org/

PhenotypeDB

http://gmql.eu/genosurf/

GenoSurf

BIBLIOGRAPHY

- Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DK, Bleicker T, Brünink S, Schneider J, Schmidt ML, Mulders DG. "Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR." *Eurosurveillance* 25.3 (2020): 2000045. https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000045
- Gudbjartsson DF, Helgason A, Jonsson H, Magnusson OT, Melsted P, Norddahl GL, Saemundsdottir J, Sigurdsson A, Sulem P, Agustsdottir AB, Eiriksdottir B. "Spread of SARS-CoV-2 in the Icelandic population." *New England Journal of Medicine* (2020). https://doi.org/10.1056/NEJMoa2006100
- Junior IJ, Polveiro RC, Souza GM, Bortolin DI, Sassaki FT, Lima AT. "The global population of SARS-CoV-2 is composed of six major subtypes." *bioRxiv* (2020). https://doi.org/10.1101/2020.04.14.040782
- Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, Cui J. "On the origin and continuing evolution of SARS-CoV-2." *National Science Review* (2020). https://doi.org/10.1093/nsr/nwaa036
- Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, Masciovecchio C, Angeletti S, Ciccozzi M, Gallo RC, Zella D. "Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant." *Journal of Translational Medicine* (2020). https://doi.org/10.1186/s12967-020-02344-6
- Zhang L, Jackson CB, Mou H, Ojha A, Rangarajan ES, Izard T, Farzan M, Choe H. "The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity." *bioRxiv* preprint manuscript, 2020.06.12.148726. https://doi.org/10.1101/2020.06.12.148726

- Bernasconi A, Ceri S, Campi A, Masseroli M. "Conceptual modeling for genomics: building an integrated repository of open data." *International Conference on Conceptual Modeling*. Springer, Cham, 2017. https://doi.org/10.1007/978-3-319-69904-2_26
- Canakoglu A, Bernasconi A, Colombo A, Masseroli M, Ceri S. "GenoSurf: metadata driven semantic search system for integrated genomic datasets." *Database* (2019). https://doi.org/10.1093/database/baz132
- Bernasconi A, Canakoglu A, Pinoli P, Ceri S. "Empowering Virus Sequence Research through Conceptual Modeling." International Conference on Conceptual Modeling (ER 2020). (https://doi.org/10.1101/2020.04.29.067637 preprint version)
- Canakoglu A, Pinoli P, Bernasconi A, Alfonsi T, Melidis DP, Ceri S. "ViruSurf: an integrated database to investigate viral sequences." Nucleic Acids Research, gkaa846, https://doi.org/10.1093/nar/gkaa846

EMPOWERING VIRUS SEQUENCE RESEARCH THROUGH CONCEPTUAL MODELING

THANK YOU FOR YOUR INTEREST IN OUR PRESENTATION

ANNA BERNASCONI, ARIF CANAKOGLU,
PIETRO PINOLI, STEFANO CERI

DEIB, POLITECNICO DI MILANO

ER 2020 – ONLINE EVENT