

Storing and analyzing viral sequences through data-driven Genomic Computing

What you always wanted to know on viral sequences
(and never dared to ask)

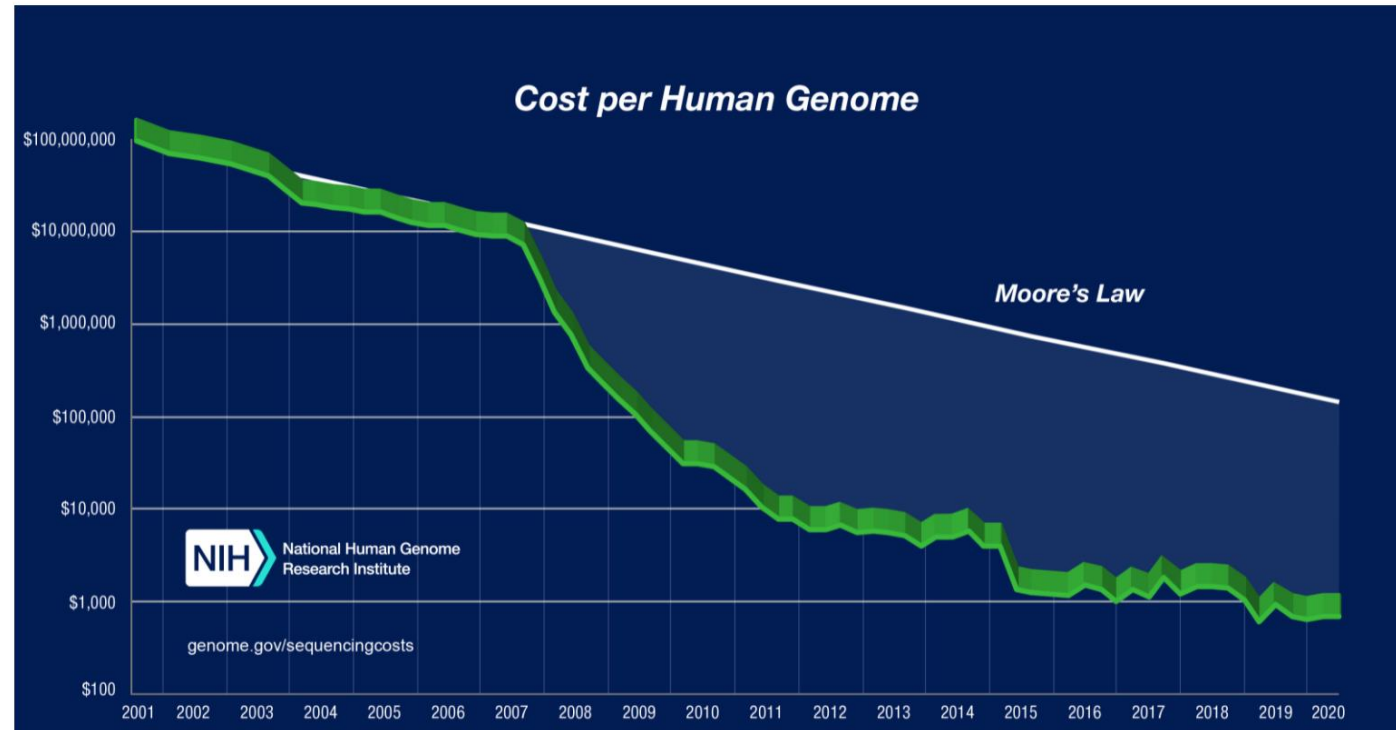
Stefano Ceri - GeCo Team

DEIB | Dipartimento di Elettronica,
Informazione e Bioingegneria

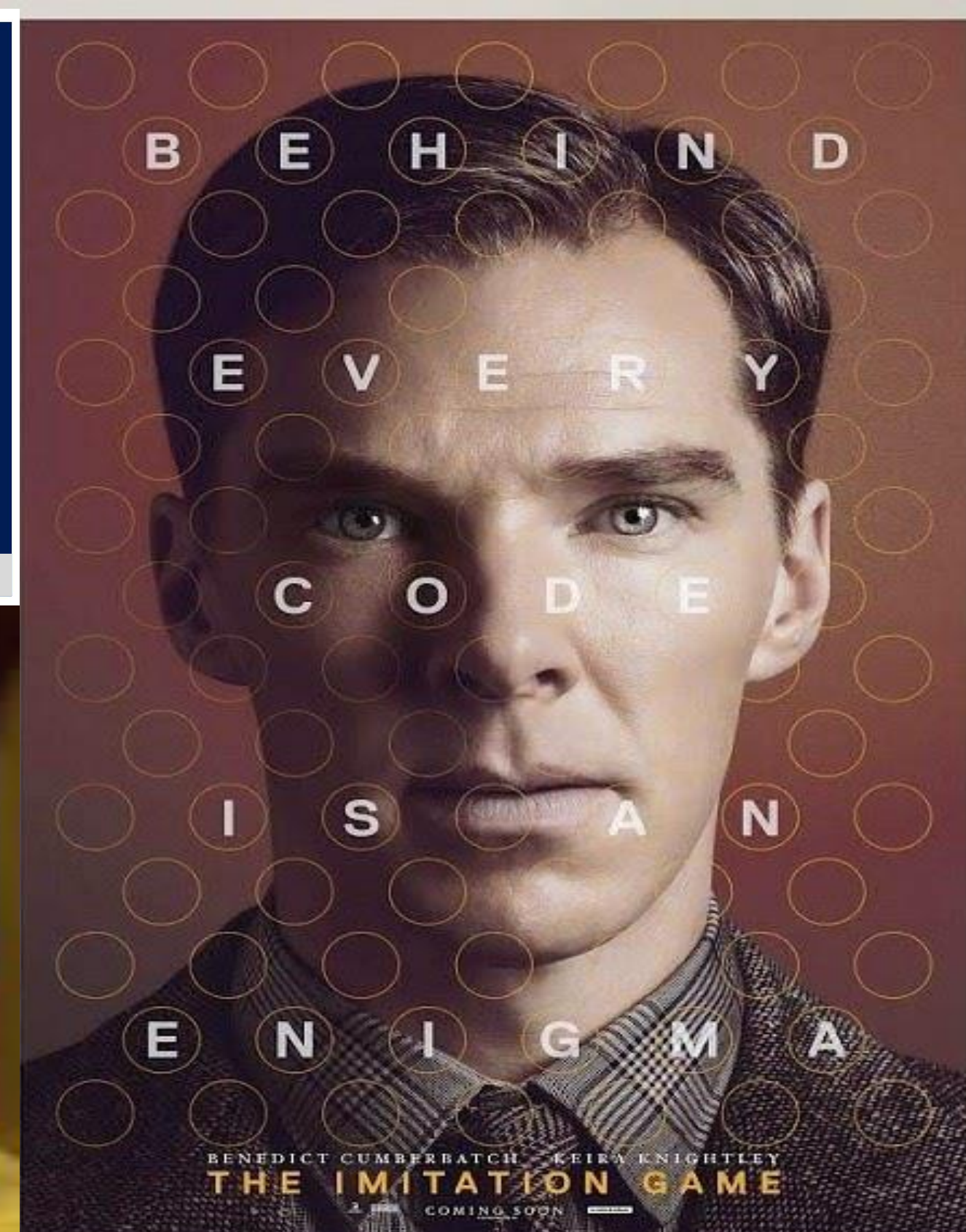
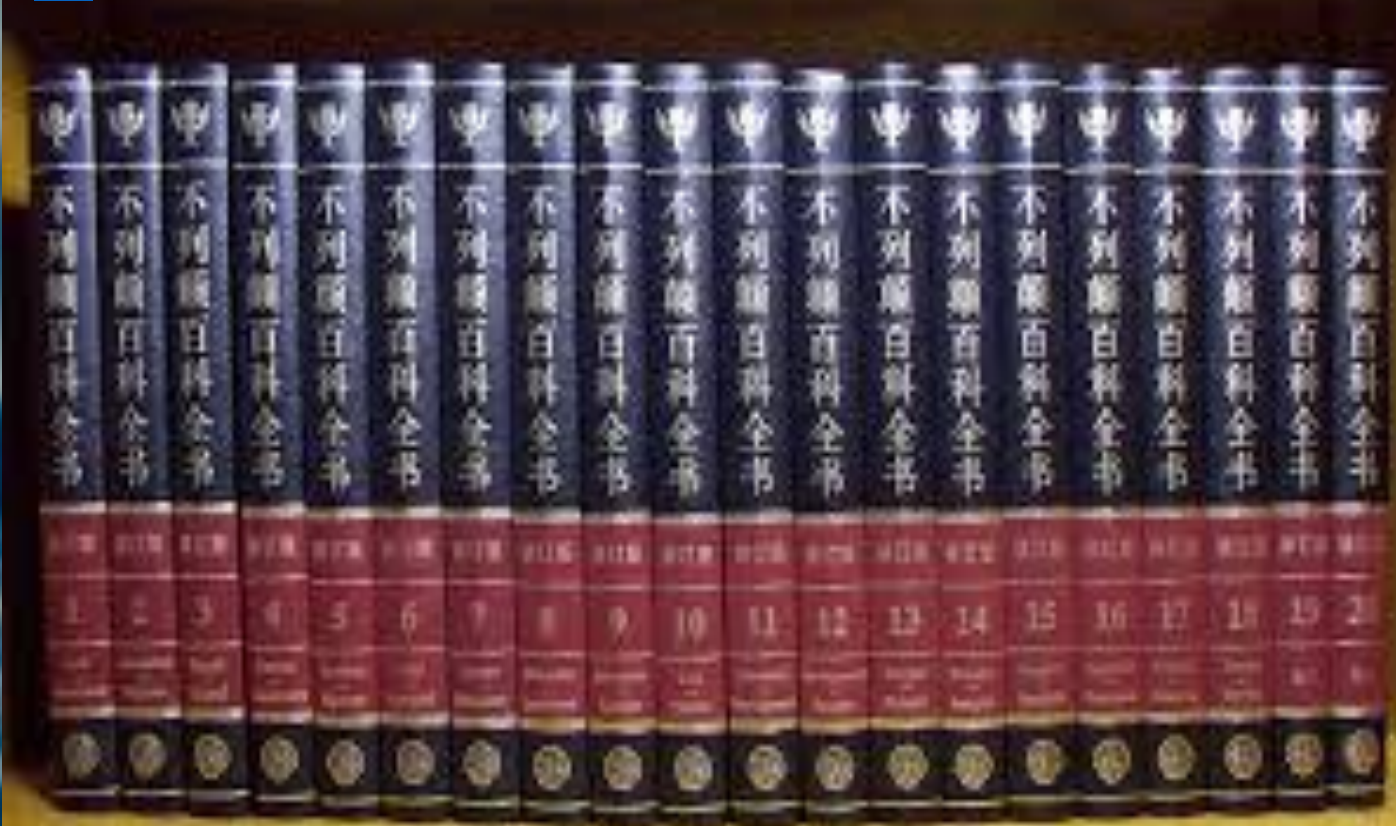


POLITECNICO
MILANO 1863

The challenge



Source: <https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost>



Data-Driven Genomic Computing (GeCo)

ERC Advanced Grant, Sept. 1, 2016 – August 31, 2021

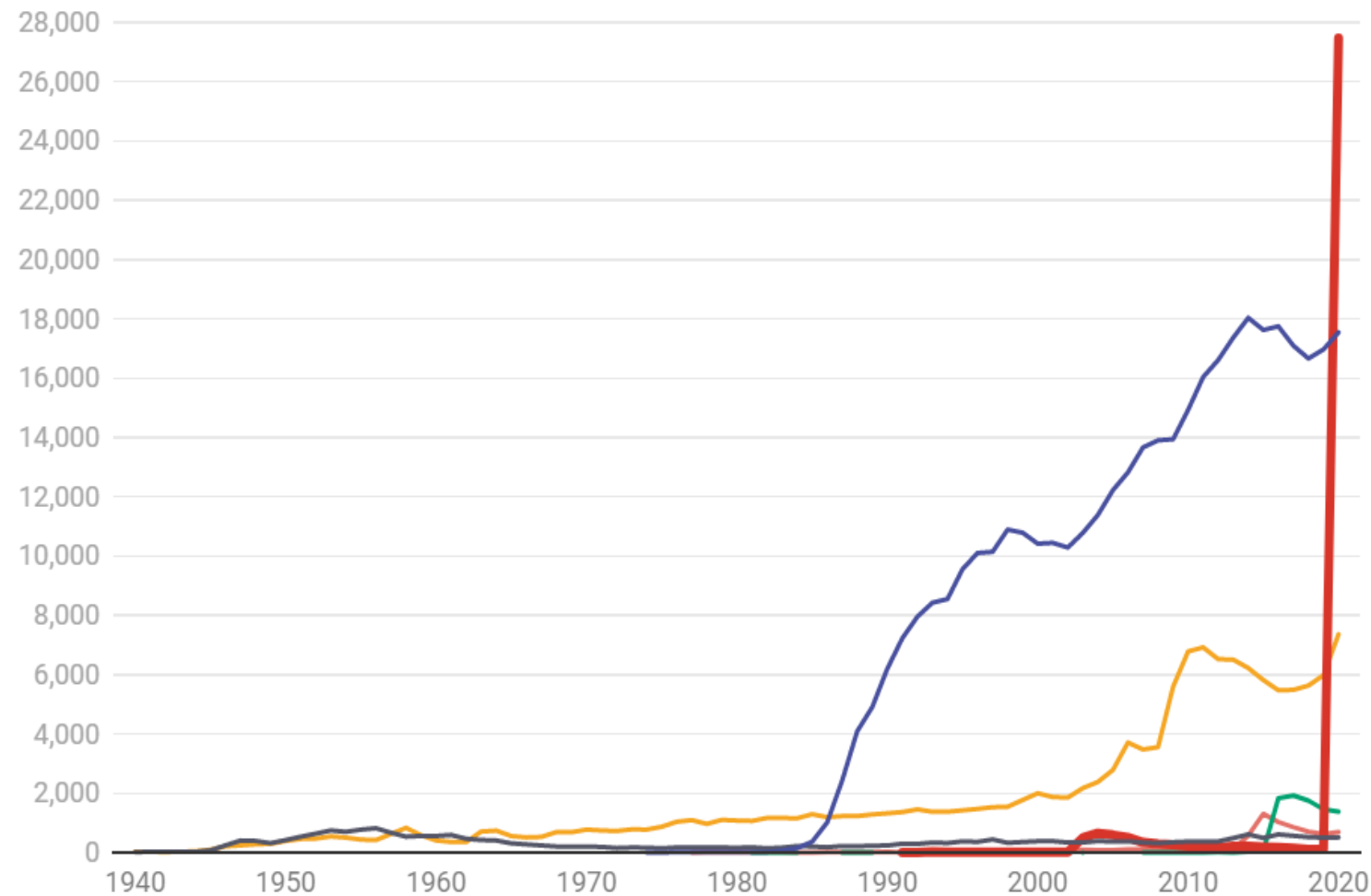
Focus: Data modeling, integration, extraction and search for integrative genomic analysis

Approach: Radical change in data management abstractions (broader and simpler)

Results so far: Big data management system (GMQL) + integrated repository (Metabase) + search system (GenoSurf), demonstrated through biological and clinical research

Ongoing work: GeCoAgent (user-friendly platform for biologists and clinicians)

The pandemics changes everything



Number of publications about:
Blue: HIV
Yellow: Influenza
Red: SARS-CoV-2

Source: <https://theconversation.com/as-scientists-turn-their-attention-to-covid-19-other-research-is-not-getting-done-and-that-can-have-lasting-consequences-154040>

Today's COVID-19 Open Research Dataset (CORD-19): over 500,000 scholarly articles

Focus change: from multi-omics for the human genome to RNA sequencing for the viral genome

- Human genome: 3×10^9 (billions) DNA nucleotides
- Viral genome: 3×10^4 (thousands) RNA nucleotides

Common aspects in human and viral genomics (GeCo approach)

One goal

Producing data-driven abstractions and systems

One systematic approach

Model

Select data sources
Propose conceptual models to unify relevant data sources

Integrate and build

Build data integration pipelines (including cleaning, normalization and semantic annotation)
Consolidate and maintain repositories

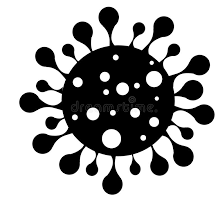
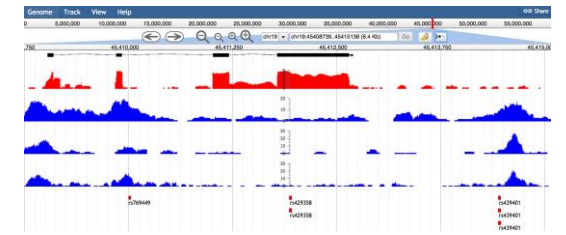
Search

Understand needs of end-users
Produce search interfaces over repositories

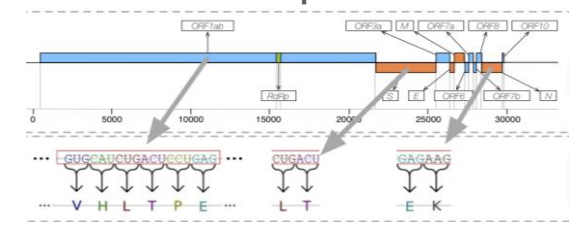
Two domains



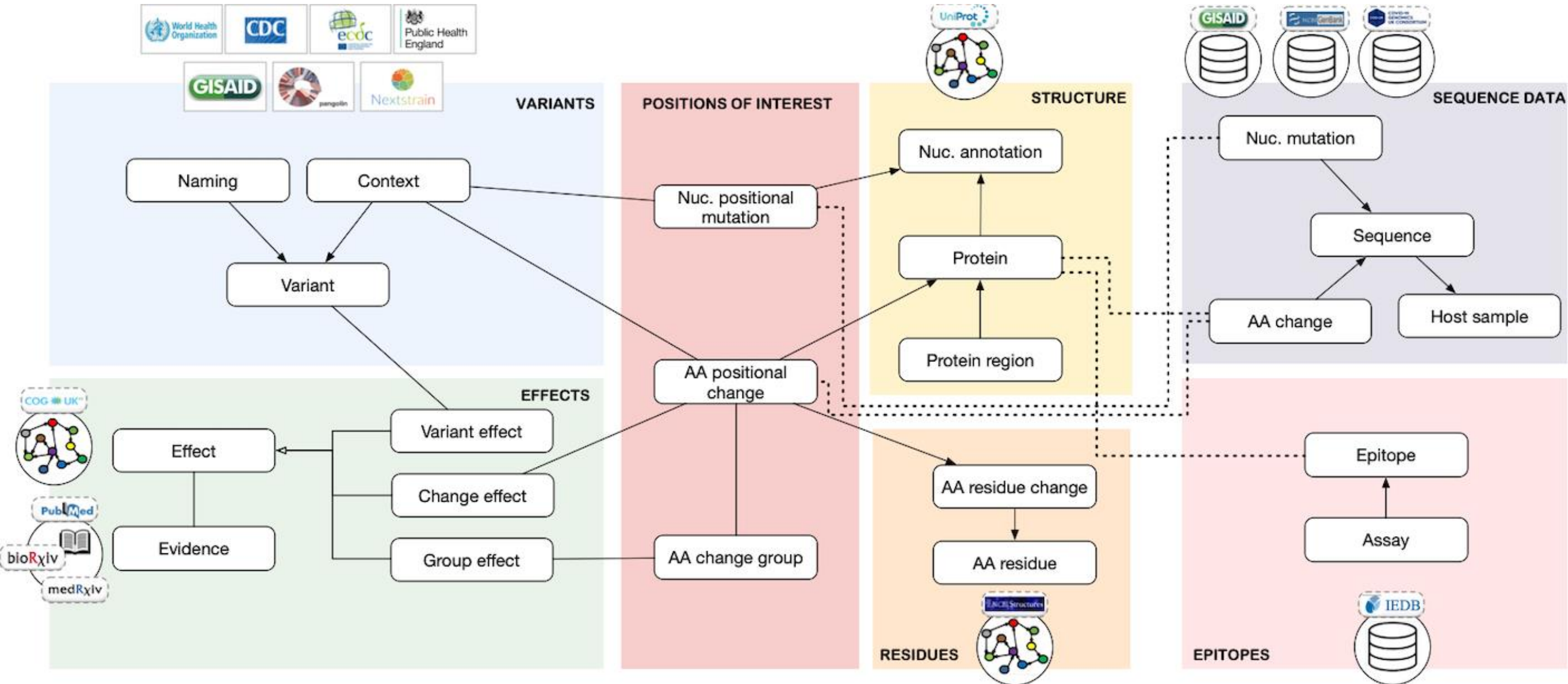
Human genomes



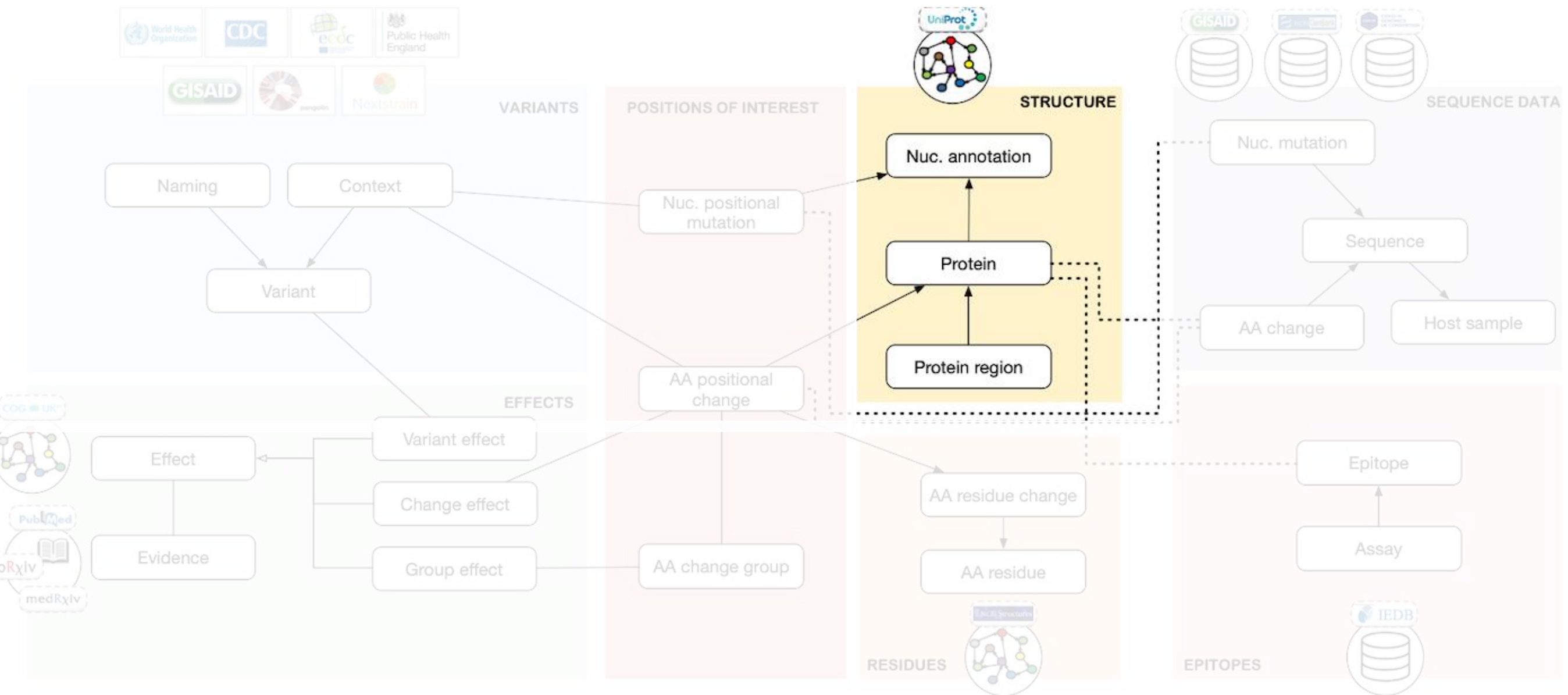
Virus sequences



PART 1 – SARS-CoV-2 Concepts (CoV2K)



The Knowledge Model: STRUCTURE



Basics

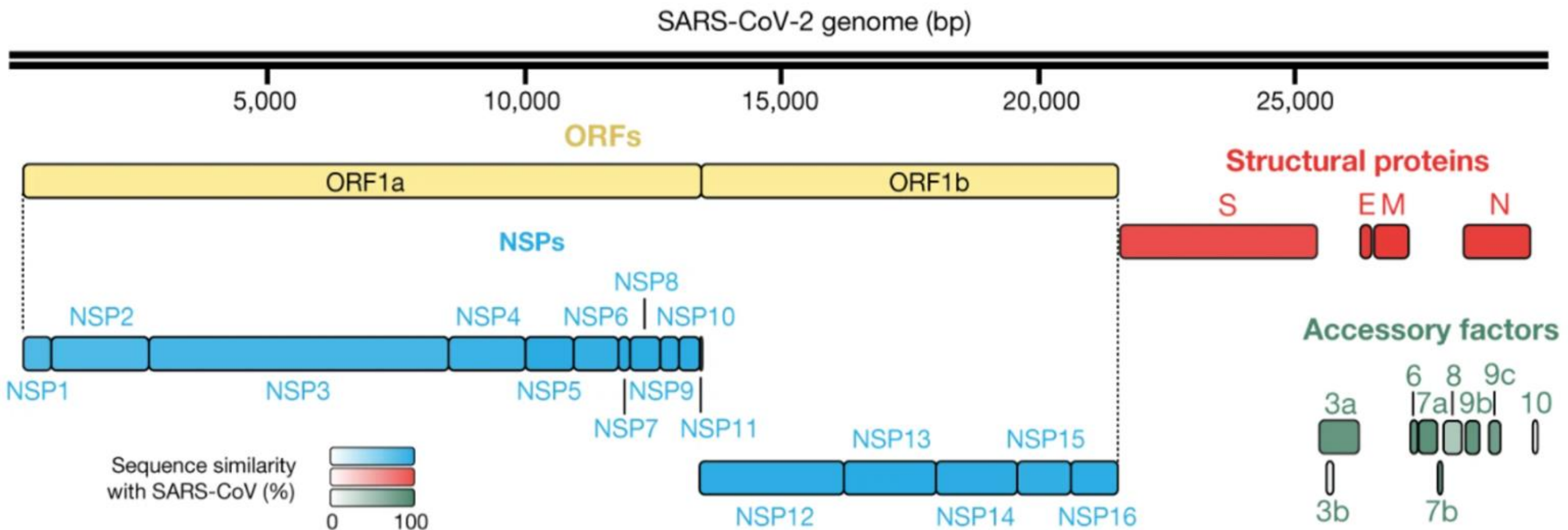
RNA viruses have ribonucleic acid (RNA) as their genetic material. They are usually single-stranded, and include the common cold, influenza, SARS, MERS, Covid-19, Dengue, Ebola, hepatitis C, hepatitis E, polio and measles.

SARS epidemics (2003) had no incubation and immediate symptoms, and was contained in Hong Kong. Quammen wrote about SARS in his book “Spillover” (2012); he anticipated the Grand Pandemics of the future (the “big one” expected by virologist) as carried by a “modified SARS virus, similar to influenza, highly infective before symptoms, moving from one city to the next on planes, acting as the angels of death”.

A close look to the genome of SARS-CoV-2

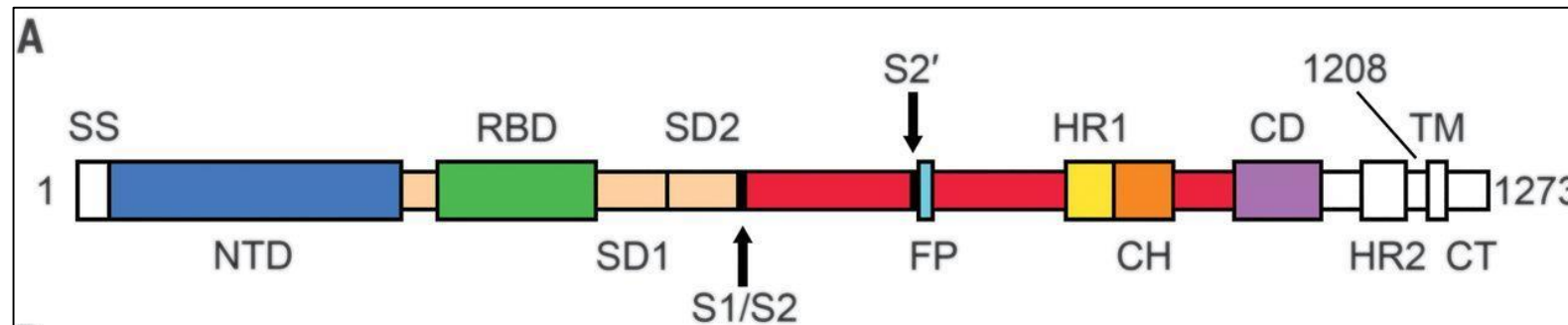
SARS-CoV-2 genome has about 30K bases (each base encodes guanine, uracil, adenine and cytosine, denoted by letters G, U, A and C) that direct the synthesis of proteins. It has strong sequence similarity with SARS-CoV responsible for SARS.

The genome of SARS-CoV-2 includes 4 structural proteins (Spike, E, M, N), 16 non structural proteins (NSP1-NSP16) and a variable number of “accessory proteins”. The 16 NSP proteins are encoded by two large, overlapping ORFs (“open reading frames”).

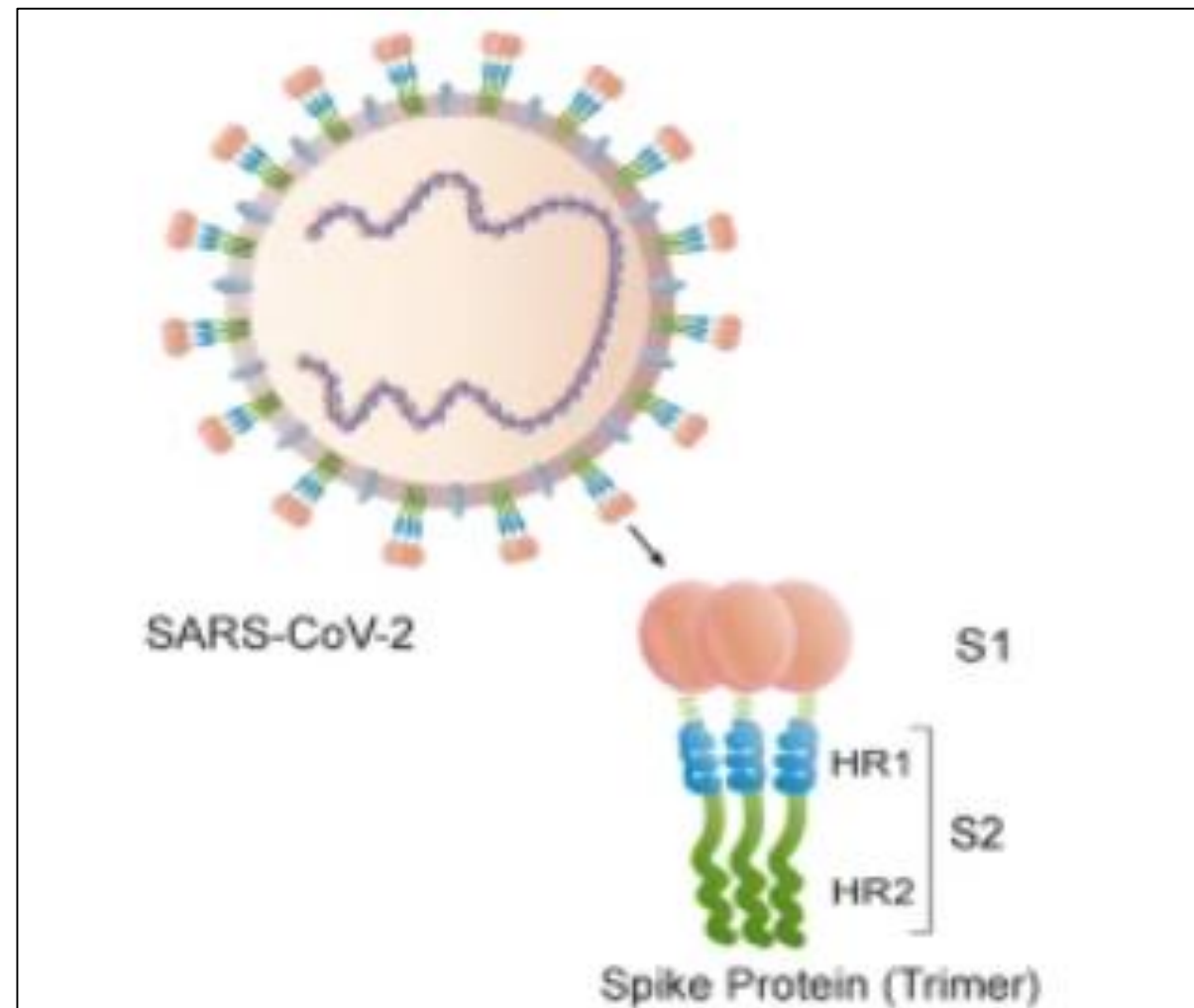


The Spike protein

Relevant regions: RBD (receptor binding domain) and NTD (N-terminal domain)

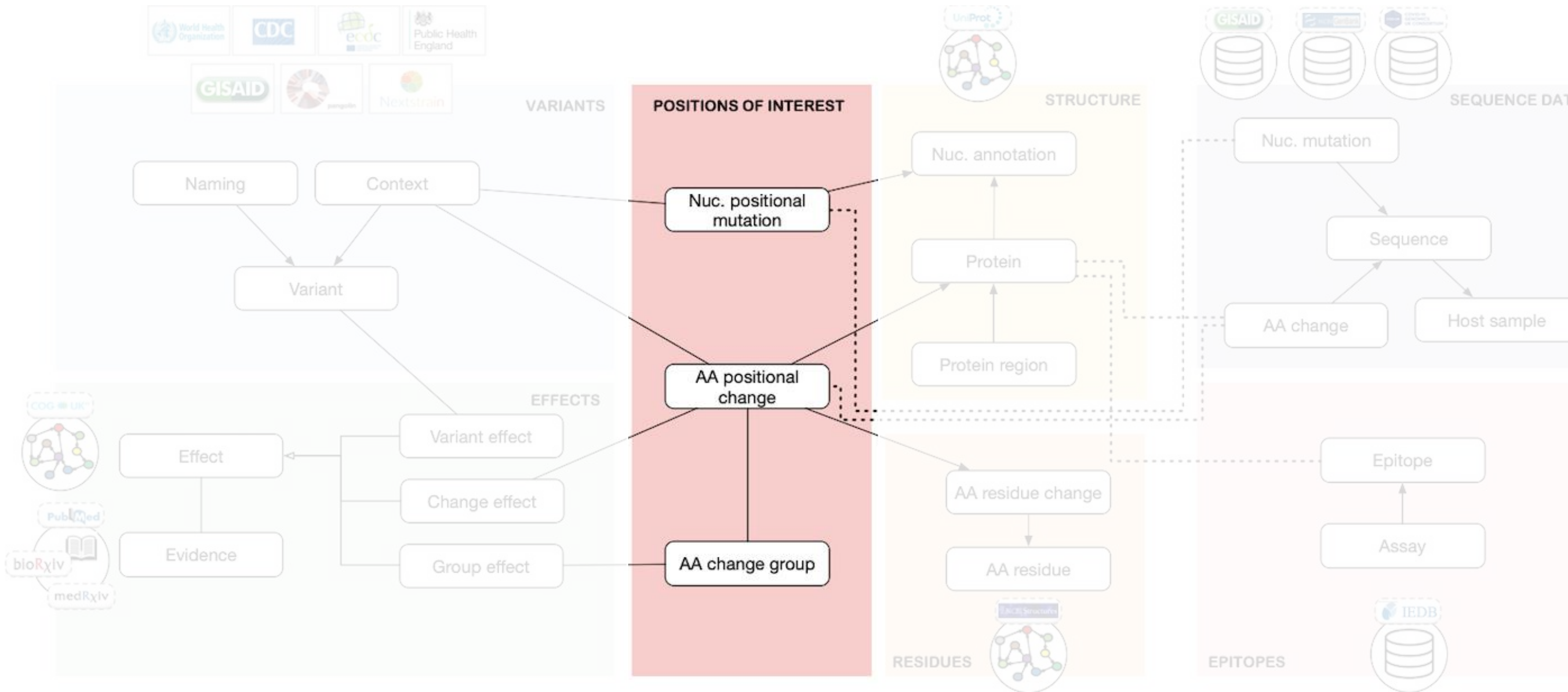


Source: Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, Graham BS, McLellan JS. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*. 2020 Mar 13;367(6483):1260-3. <https://doi.org/10.1126/science.abb2507>



Source: Huang Y, Yang C, Xu XF, Xu W, Liu SW. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacologica Sinica*. 2020 Sep;41(9):1141-9. <https://doi.org/10.1038/s41401-020-0485-4>

The Knowledge Model: POSITIONS OF INTEREST



How to describe mutational processes

- Nucleotide Mutation:

ReferenceNuc/GenomicCoordinates/AlternativeNuc

e.g. A23403G

- Amino Acid Change:

Protein: ReferenceAA/PositionInProtein/AlternativeAA

e.g. spike:D614G (this is the AA change that occurs as result of nucleotide A23403G mutation)

Amino acid changes are more relevant – they can impact the protein function. Their effect depends on:

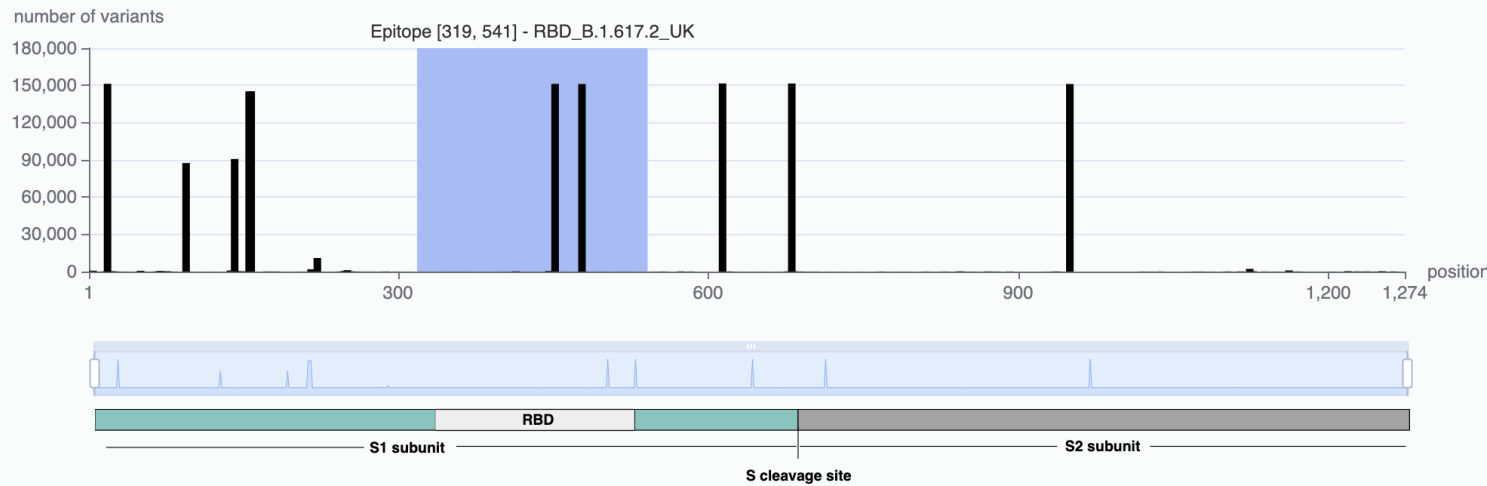
- The substitution (regardless of the position)
- The position in the protein
- The co-occurrence with other changes

Preview: Delta mutations over important ranges

Spike RBD mutations and immune escape

Of all RBD residues for which substitutions affected recognition by convalescent sera, DMS identified E484 as being of principal importance, with amino acid changes to K, Q or P reducing neutralization titres by more than an order of magnitude³⁹.

151,559 sequences



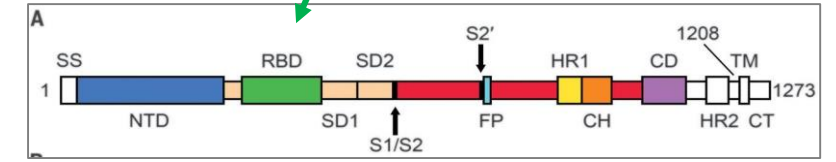
position: 452

- L→R: 151223 (100%)
- L→W: 1 (0%)

position: 478

- T→K: 151127 (100%)
- T→I: 1 (0%)

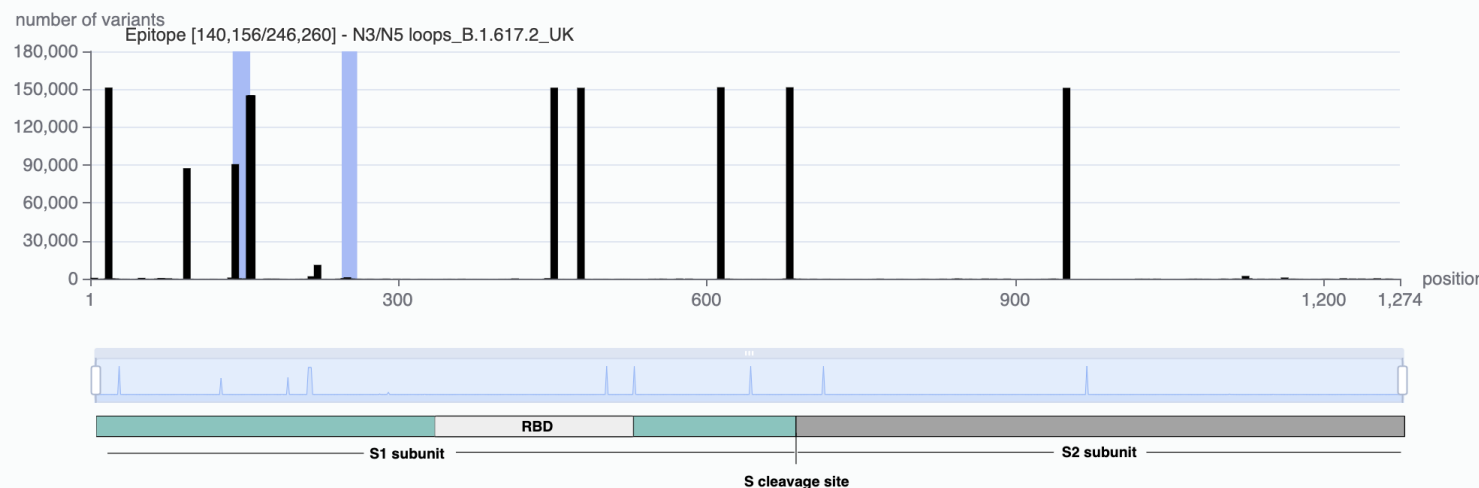
Regionⁱ 319 - 541 Receptor-binding domain (RBD)



Spike NTD mutations and immune escape

In the NTD, most of the evidence for immune evasion focuses on a region centred at a conformational epitope consisting of residues 140–156 (N3 loop) and 246–260 (N5 loop)

151,559 sequences



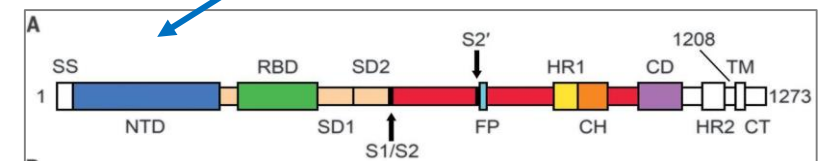
position: 142

- G→V: 1 (0%)
- G→-: 4 (0%)
- G→D: 90676 (60%)
- G→I: 1 (0%)
- G→A: 2 (0%)
- G→Y: 3 (0%)
- G→N: 4 (0%)

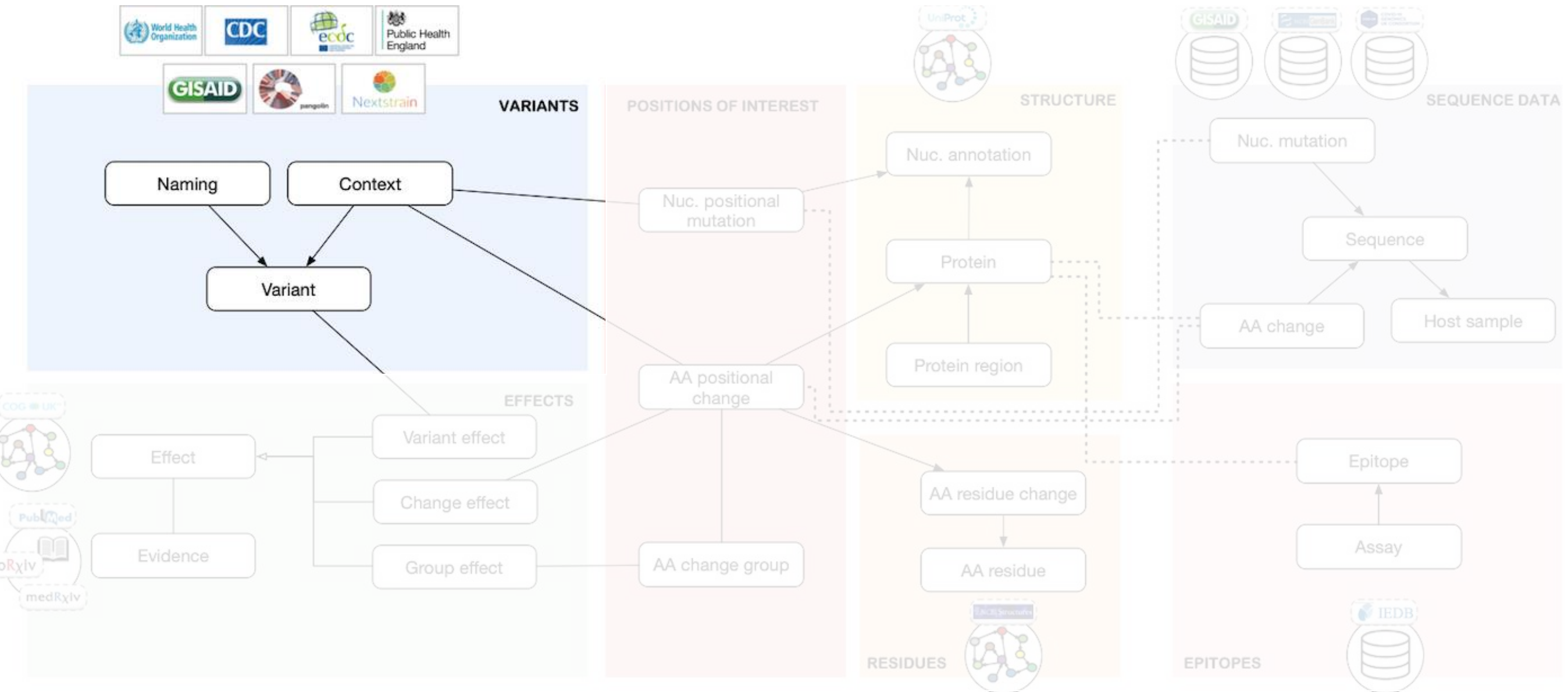
position: 156

- E→-: 145034 (96%)
- E→G: 4 (0%)

Domainⁱ 13 - 303 BetaCoV S1-NTD



The Knowledge Model: VARIANTS



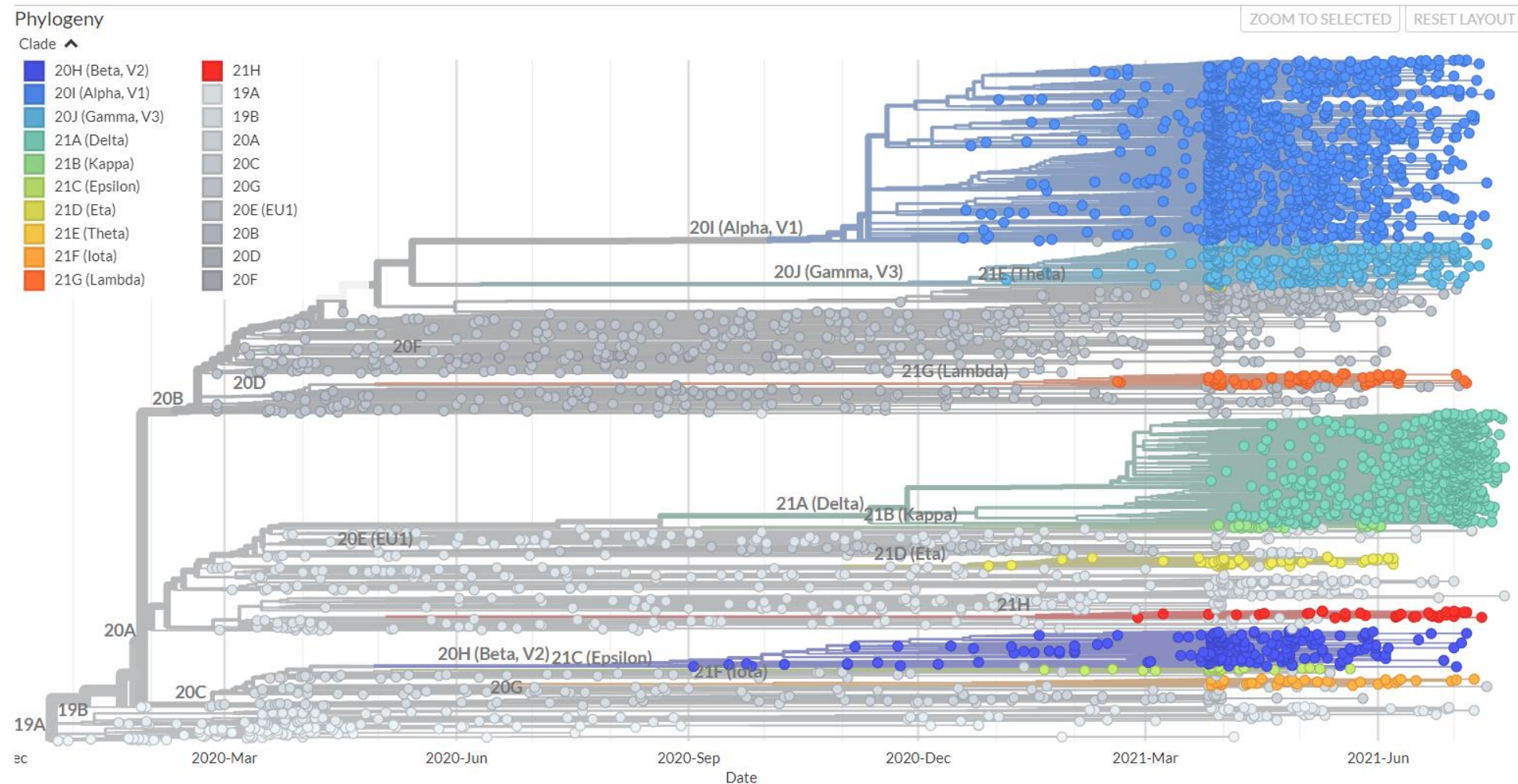
Variants

Characterized by several co-occurring amino acid changes. The cumulative effect of characterizing amino acid changes gives to “variants” significant advantages, e.g. alpha and delta variants – which have become dominant.

Determination:

Phylogenetic analysis provides an in-depth understanding of how SARS-CoV-2 sequences evolve through genetic changes.

Different phylogenetic trees are captured and named by different organizations (lineages, clades..)



Issues in Variant Characterization

- Variants associated to an increased risk to global public health prompted WHO (World Health Organization) to characterize specific Variants of Interest (VOIs) and Variants of Concern (VOCs).
- Many other sources provide different classification methods, e.g. CDC (US Center for Disease Control) classifies them as Variants of Interest (VOIs), Variants of Concern (VOCs), and Variant of High Consequence (VOHCs).

Issues in Variant Naming

- Names are primarily assigned by GISAID, Pangolin, and Nextstrain. Pangolin's "B.1.1.7 lineage" was then named as the "UK variant".
- The emergence of VOIs and VOCs prompts WHO to consider easy-to-pronounce and non-stigmatising labels for those VOIs and VOCs using letters of the Greek Alphabet e.g. Alpha, Beta, and Gamma.

Currently designated Variants of Concern:

WHO label	Pango lineage*	GISAID clade	Nextstrain clade	Additional amino acid changes monitored ^o	Earliest documented samples	Date of designation
Alpha	B.1.1.7 [#]	GRY	20I (V1)	+S:484K +S:452R	United Kingdom, Sep-2020	18-Dec-2020
Beta	B.1.351	GH/501Y.V2	20H (V2)	+S:L18F	South Africa, May-2020	18-Dec-2020
Gamma	P.1	GR/501Y.V3	20J (V3)	+S:681H	Brazil, Nov-2020	11-Jan-2021
Delta	B.1.617.2 [§]	G/478K.V1	21A	+S:417N	India, Oct-2020	VOI: 4-Apr-2021 VOC: 11-May-2021

Alpha variant: context

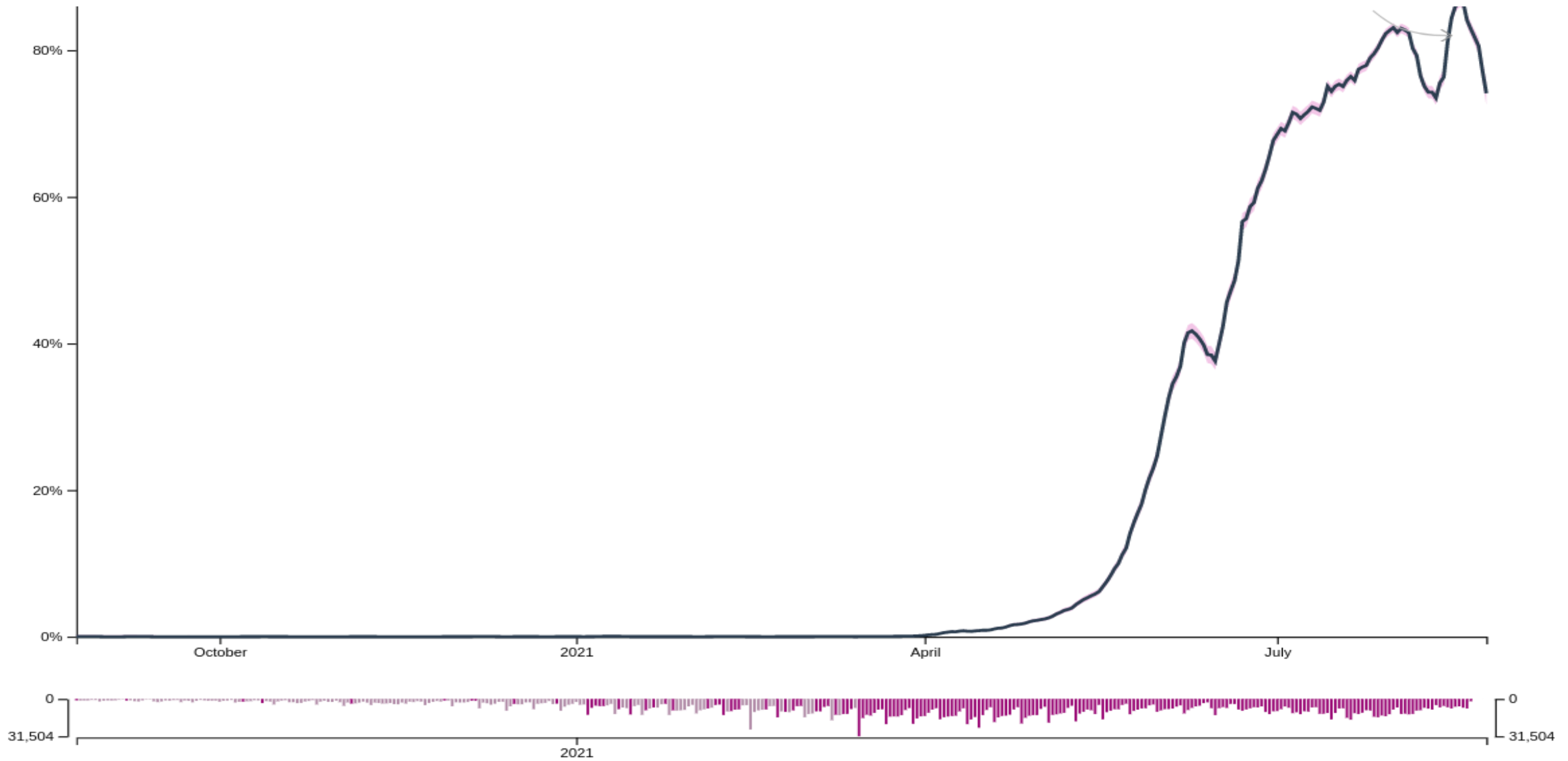
Characteristic mutations for a lineage: those amino acid changes or deletions that occur in more than 75% of lineage sequences.

Accordingly, the B.1.1.7 lineage (named Alpha variant by WHO) has 22 characteristic mutations. Eight of them are on the spike protein.

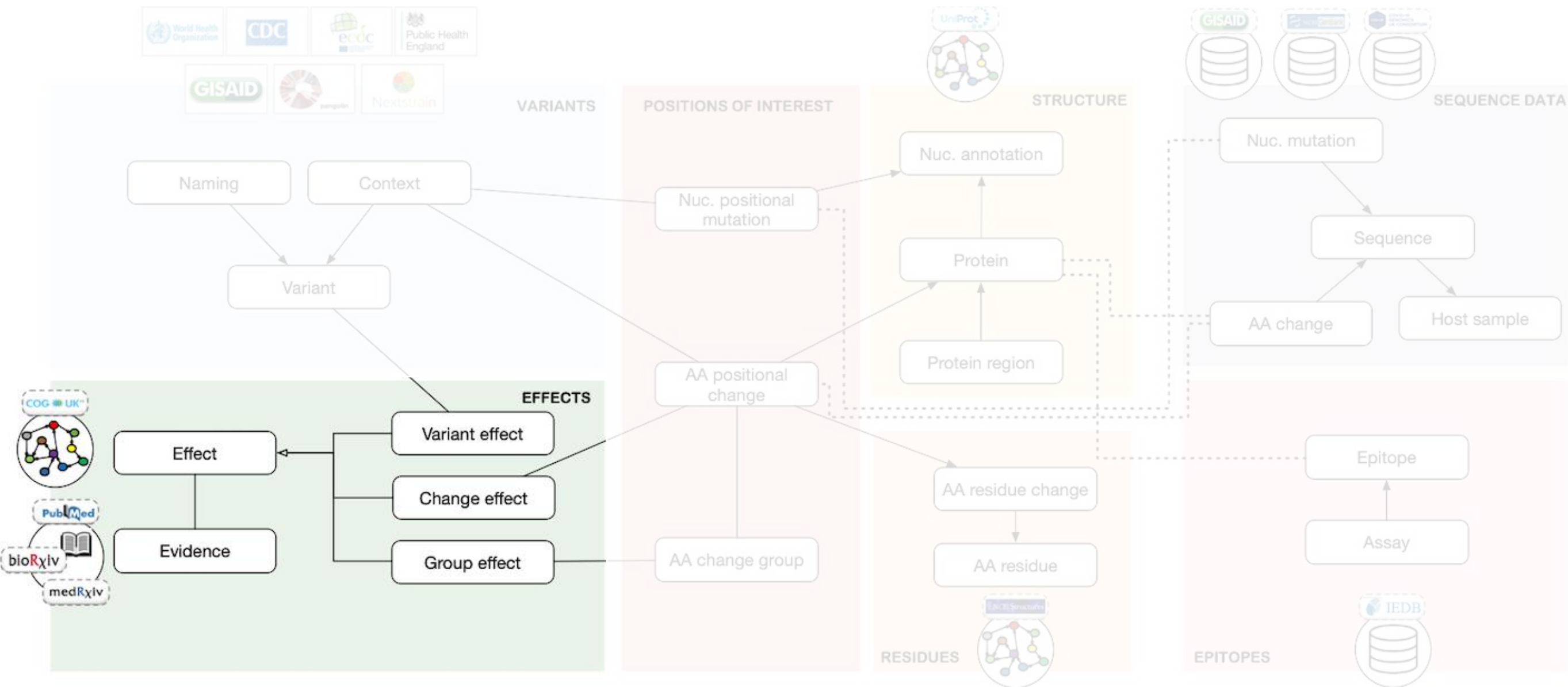
Variants are characterized in different ways by different organizations, resulting in different “contexts”

Protein	Change
NSP3	T183I
NSP3	A890D
NSP3	I1412T
NSP6	S106K
NSP6	del107/108
NSP12	P314L
Spike	del69/70
Spike	del144/145
Spike	N501Y
Spike	A570D
Spike	D614G
Spike	P681H
Spike	T716I
Spike	S982A
Spike	D1118H
ORF8	Q27*
ORF8	R52I
ORF8	Y73C
N	D3L
N	R203K
N	G204R
N	S235F

The Delta variant (VOC): prevalence worldwide



The Knowledge Model: EFFECTS



Effects of amino acid changes linked to genomic positions

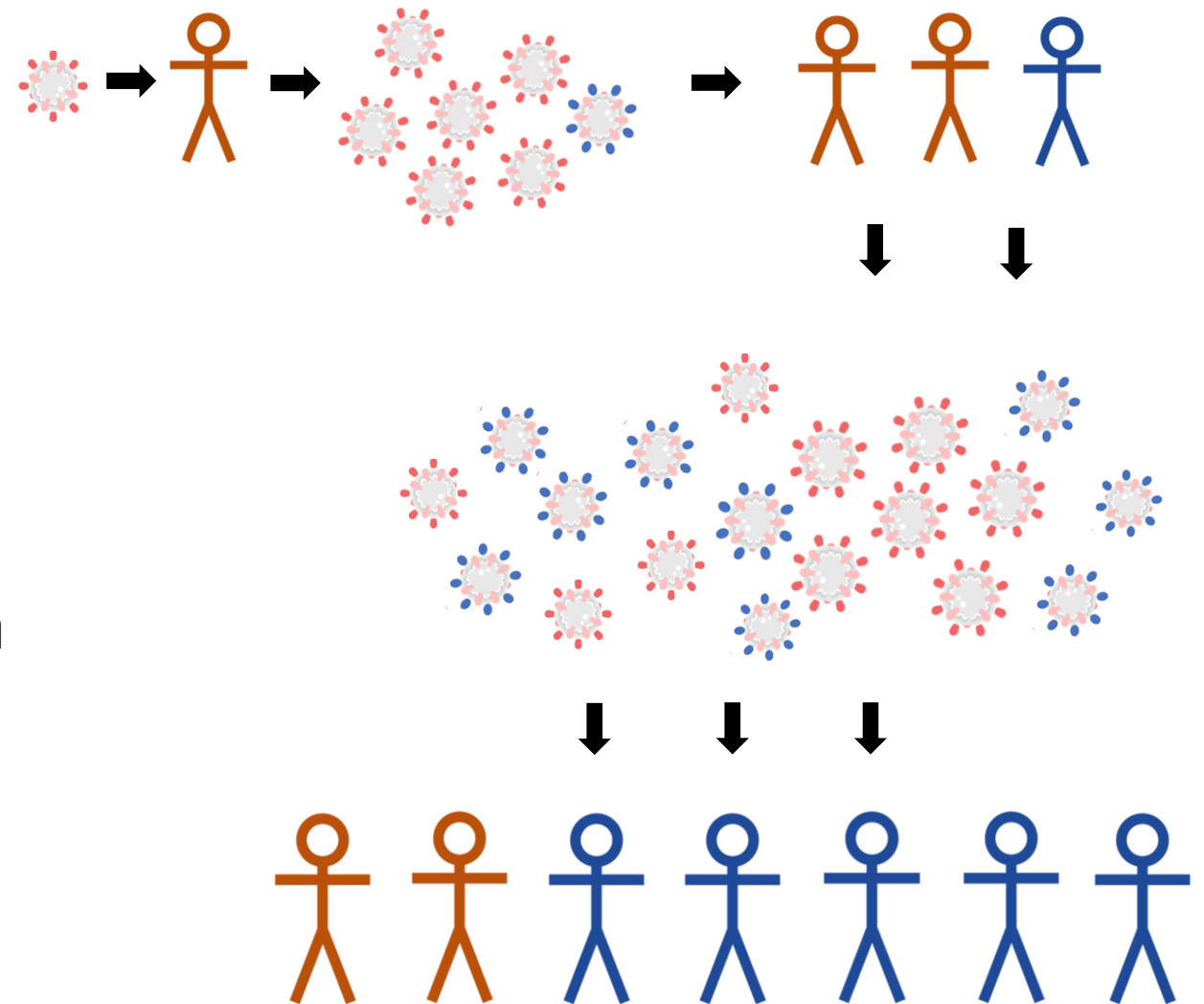
Associated to single changes or to groups of changes

Depend on the proximity to critical positions, e.g.

- Spike protein
- Inside Spike protein, Relational Binding Domain (RDB) or N-Terminal Domain (NTD)

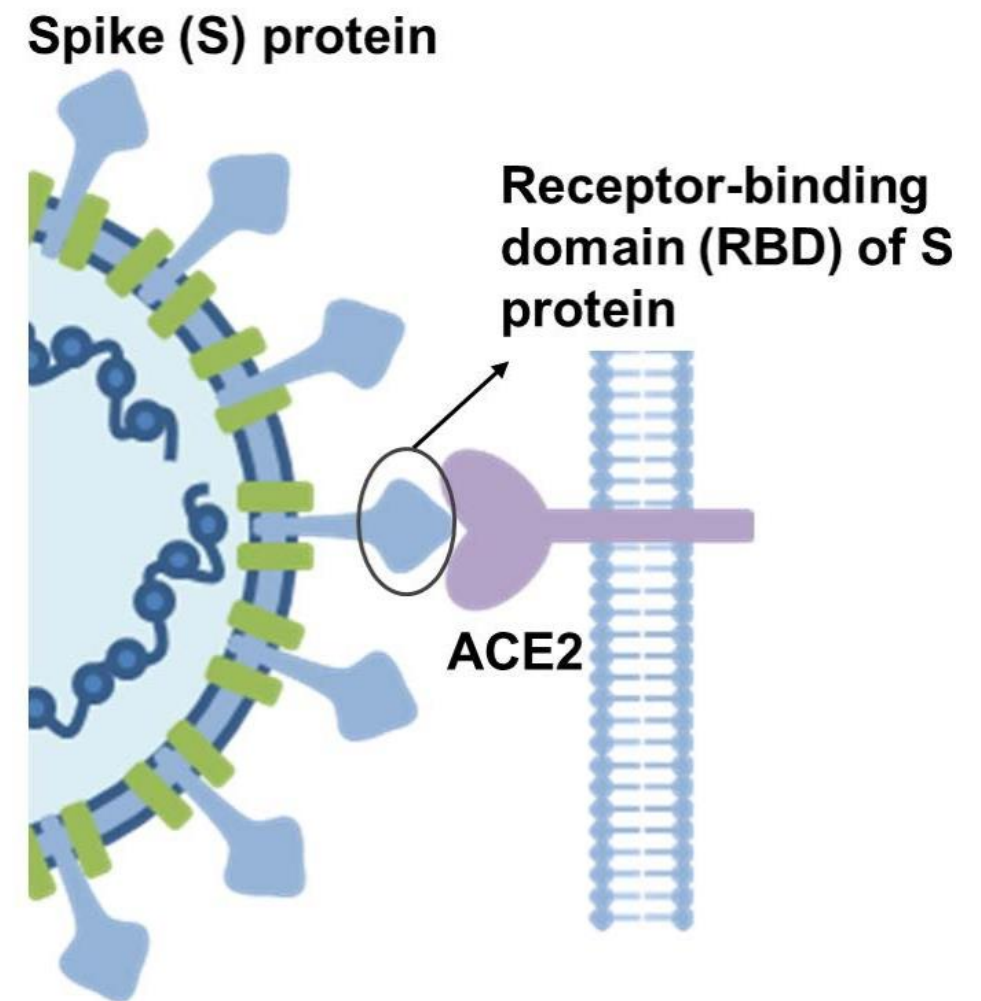
Effects of amino acid changes:epidemiology

- **Viral transmission:** Viral capability to pass from a host to another host.
- **Infectivity:** Capability of a transmitted virus to establish an infection.
- **Disease severity:** Associates with more severe symptoms caused by the virus.
- **Fatality rate:** Proportion of persons who died after the viral infection over the total confirmed cases.



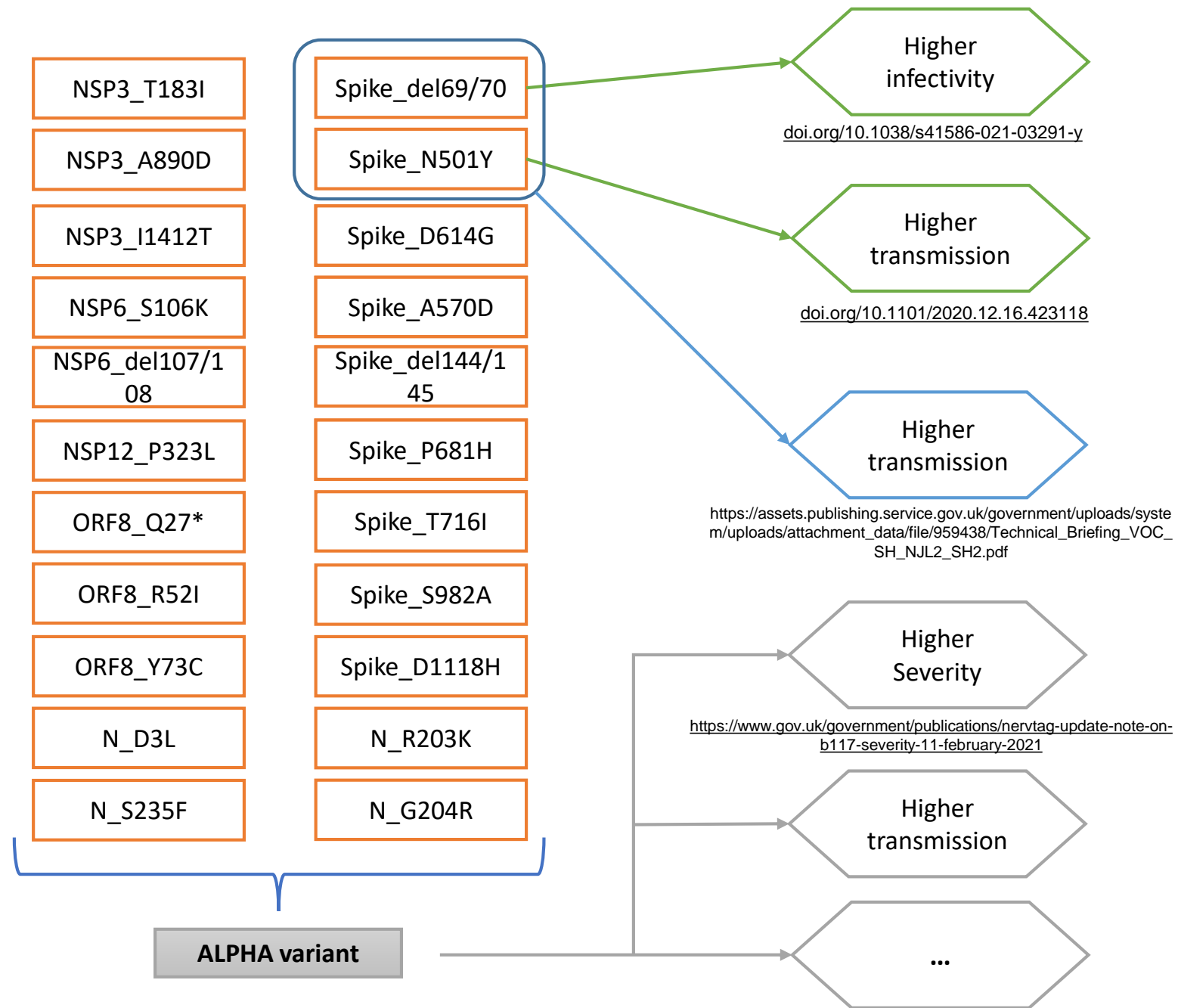
Effects of amino acid changes: Immunology

- **Sensitivity to convalescent sera (or vaccinated sera):** obtained from recovered (or vaccinated) individuals, used for prevention and treatment.
- **Sensitivity to neutralizing monoclonal antibodies:** currently the target of biomedical research for COVID-19 treatment.
- **Binding affinity to host receptor:** alteration of the interaction between receptor binding domain (RBD) of the Spike protein and the host's ACE2 receptor - changes infectivity.

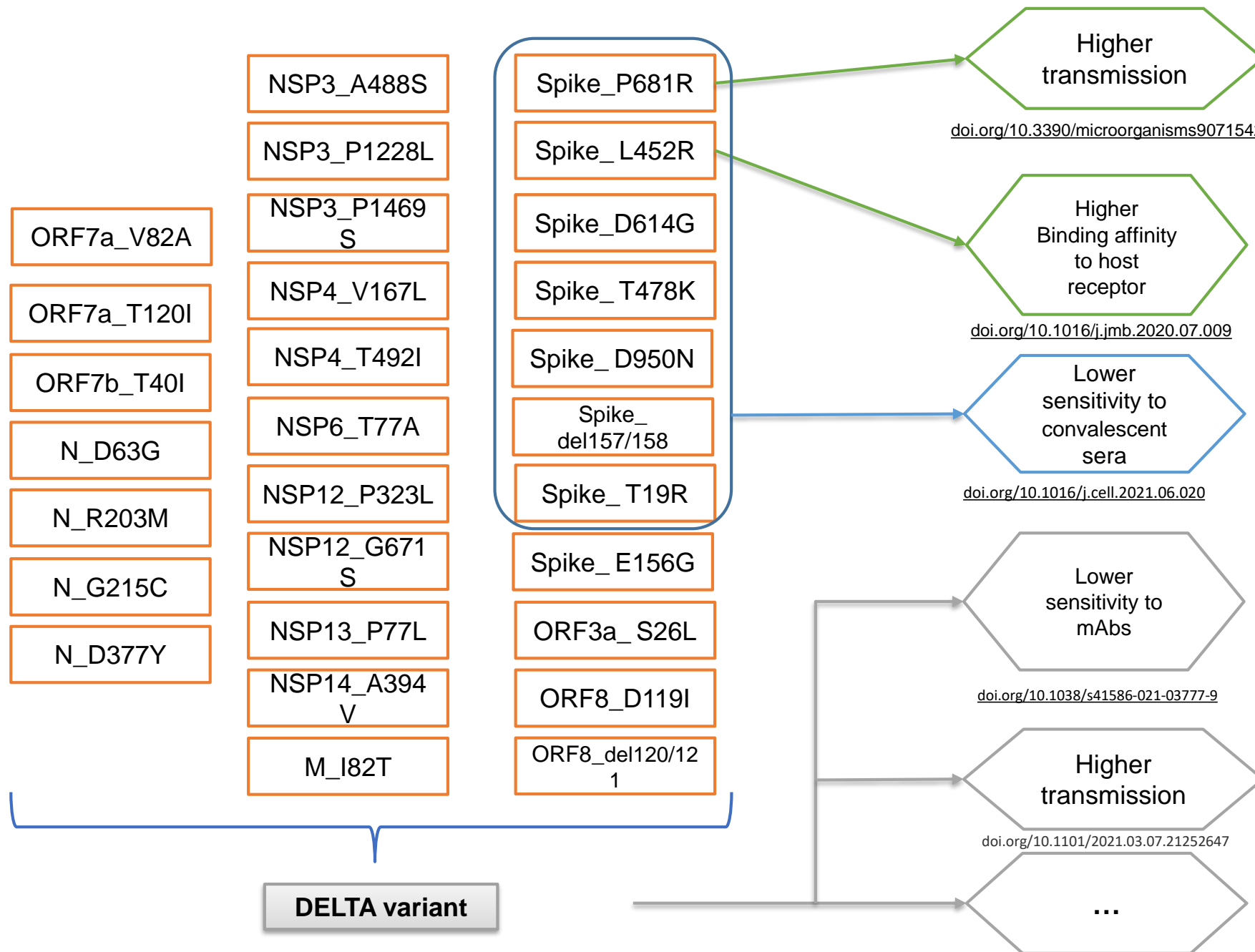


Other effects on protein dynamics and kinetics

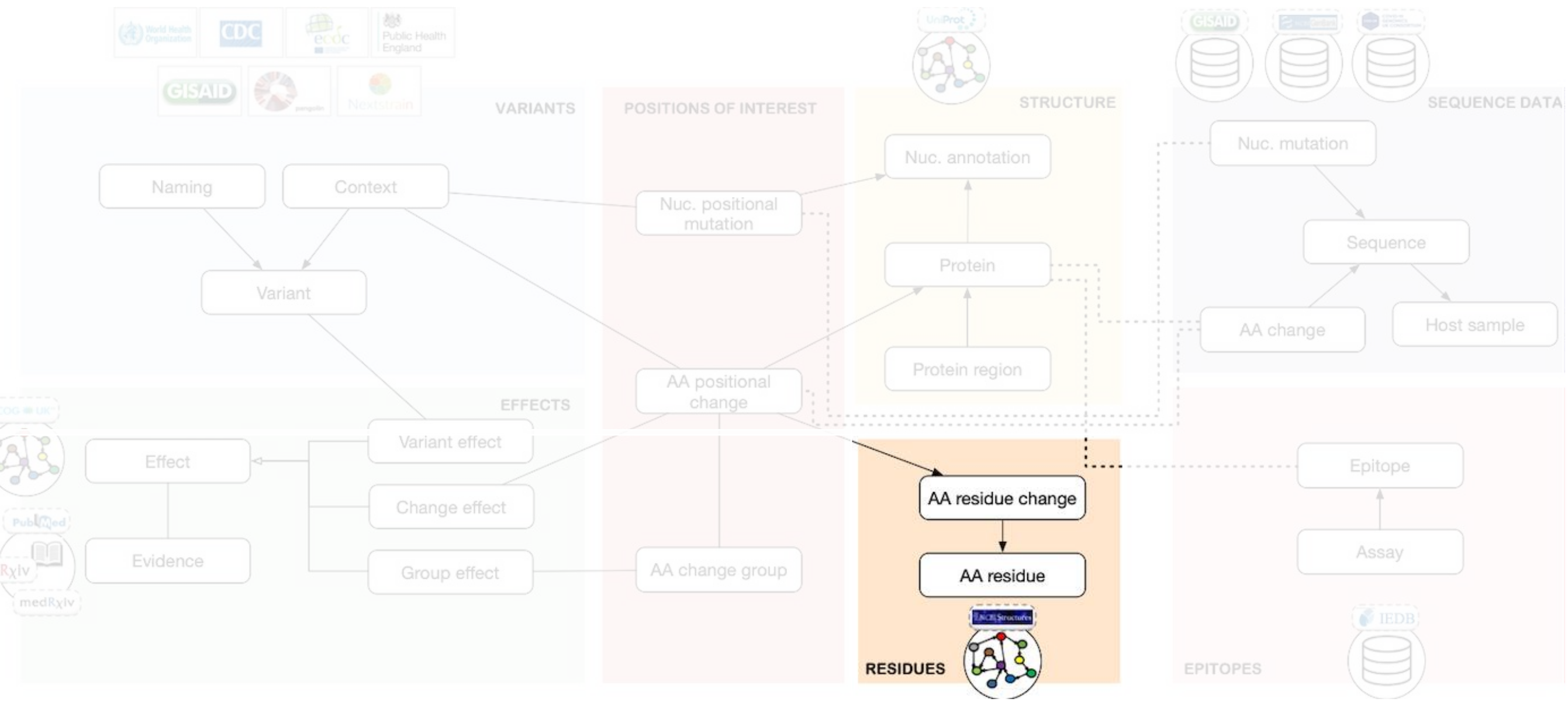
Alpha Variant: Characteristic changes and their effect



Delta Variant: Characteristic changes and their effect



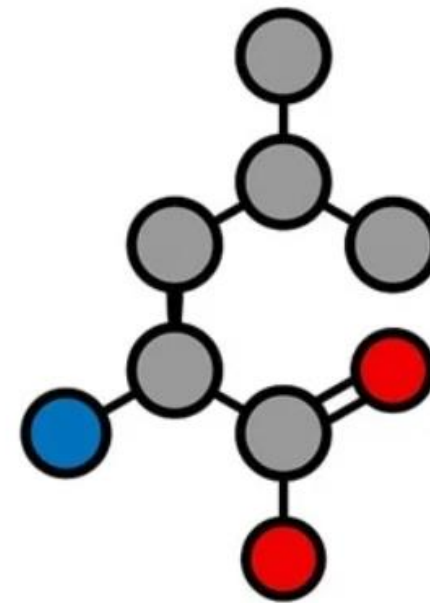
The Knowledge Model: RESIDUES



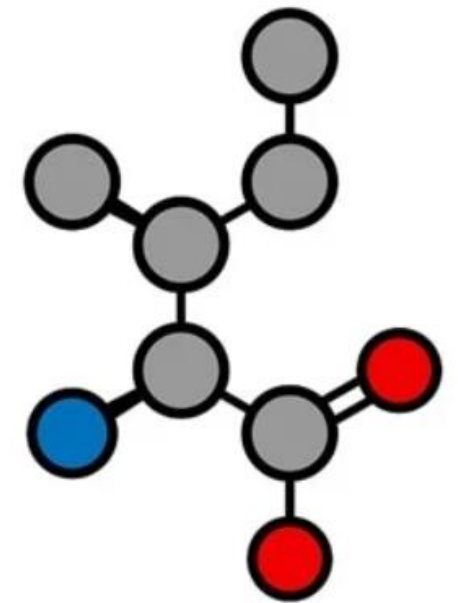
Amino acid change effects (regardless of position)

Each amino acid has its own properties depending on its chemical structure, e.g.

- **Molecular weight:** The mass of a given molecule.
 - **Isoelectric point:** pH at which a molecule carries no net electrical charge.
 - **Hydrophobicity:** A physical property of a molecule that measures affinity for water.
-
- The more distant (e.g. R. Grantham 1974) two amino acids are, the more damaging is their substitution.
 - E.g. Ile and Leu are very close amino acids, therefore, their substitutions probably will not lead to change in the overall phenotype.

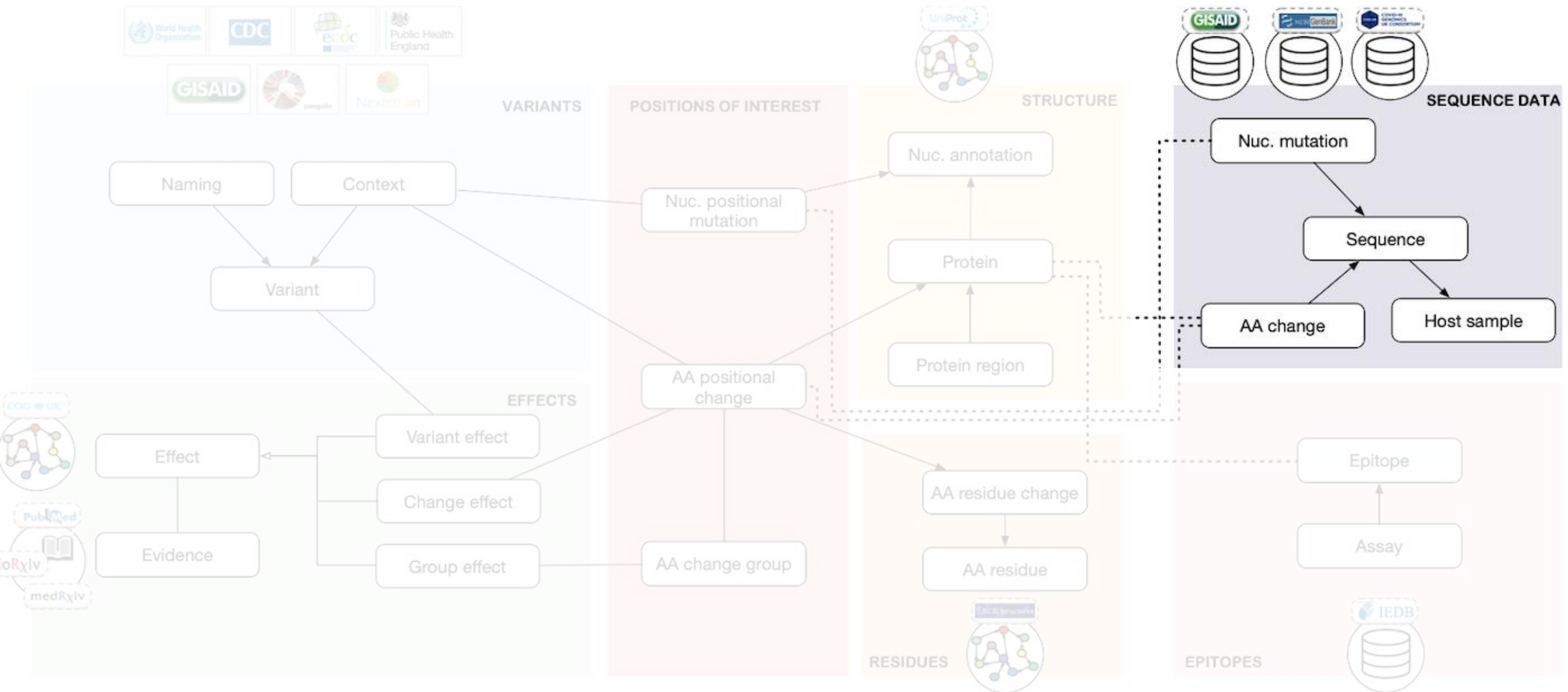


L-leucine (Leu, L)



L-isoleucine (Ile, I)

The Knowledge Model: SEQUENCE DATA



The process of sequence deposition

Occurs from laboratories worldwide

Main worldwide resource (databases)

- GenBank (fully open, worldwide)
- CogUK (fully open, UK)
- GISAID (open but protected, authorized access given to registered users)

Deposited data: the full RNA sequence (FASTA format, 30K bases)

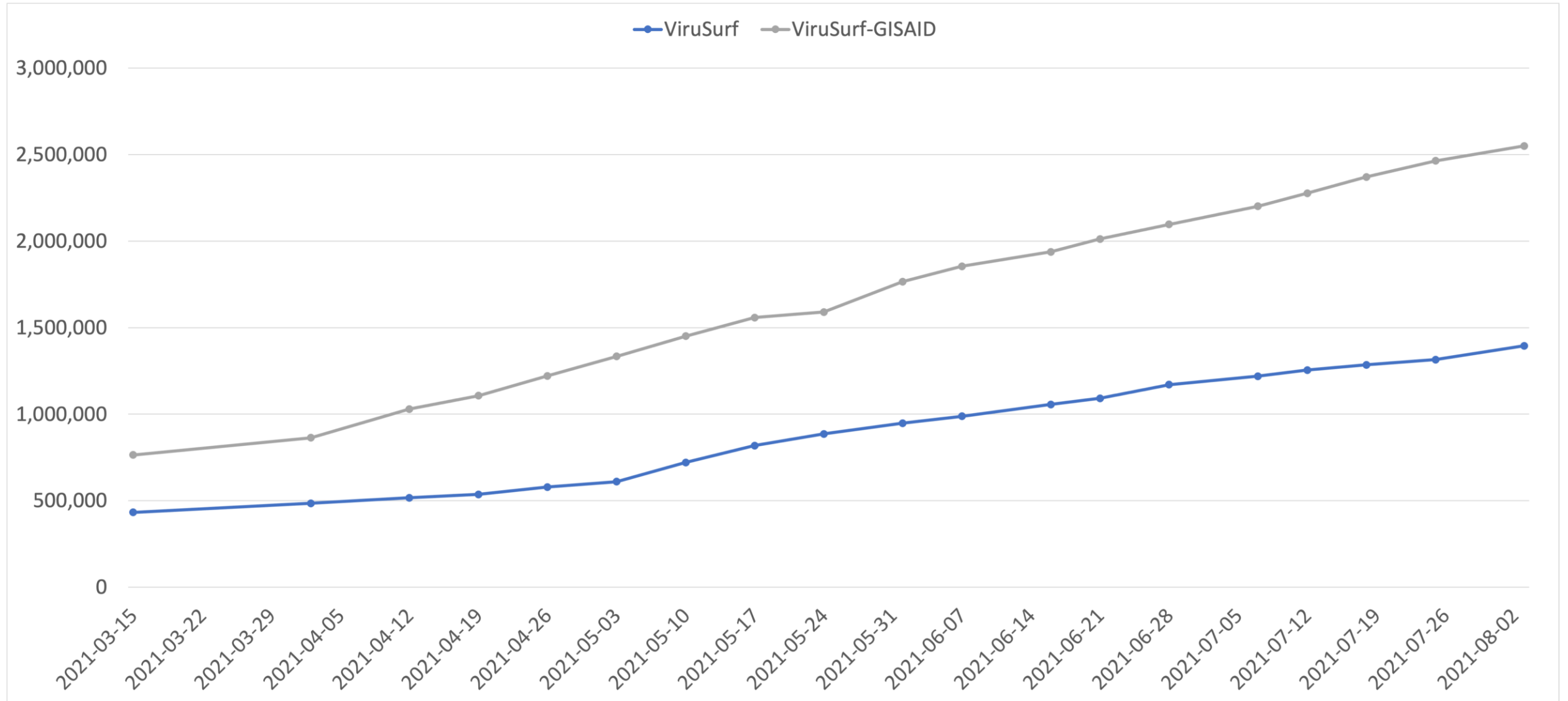
Relevant metadata:

- Collection date (of the host's sample)
- Deposition date (to the database)
- Depositing lab
- Geo-location (continent, country, region, sub-region)
- Lineage (according to some method, e.g. Pangolin)

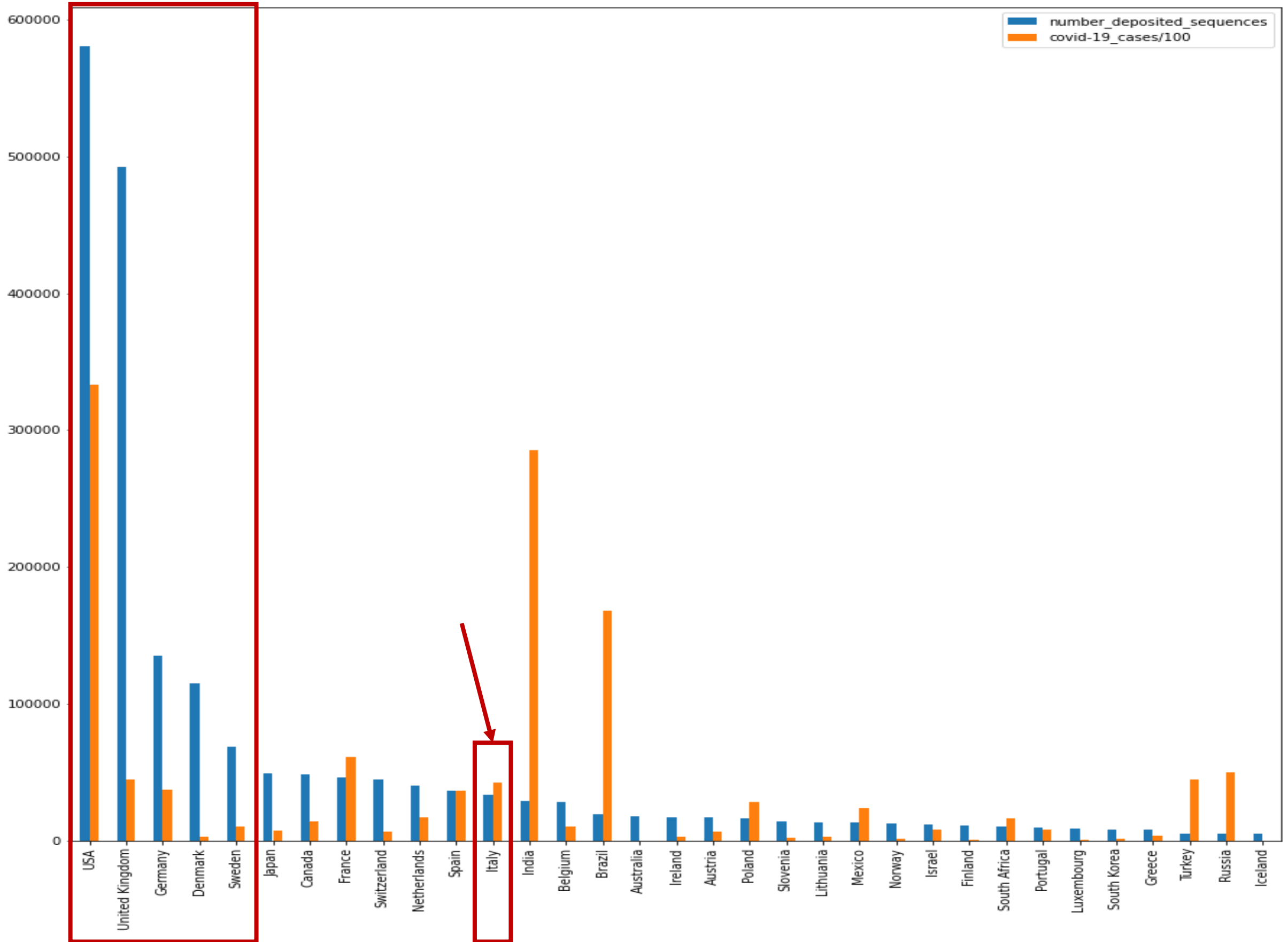
Relevant processing:

- Extract nucleotide mutations
- Extract amino acid changes

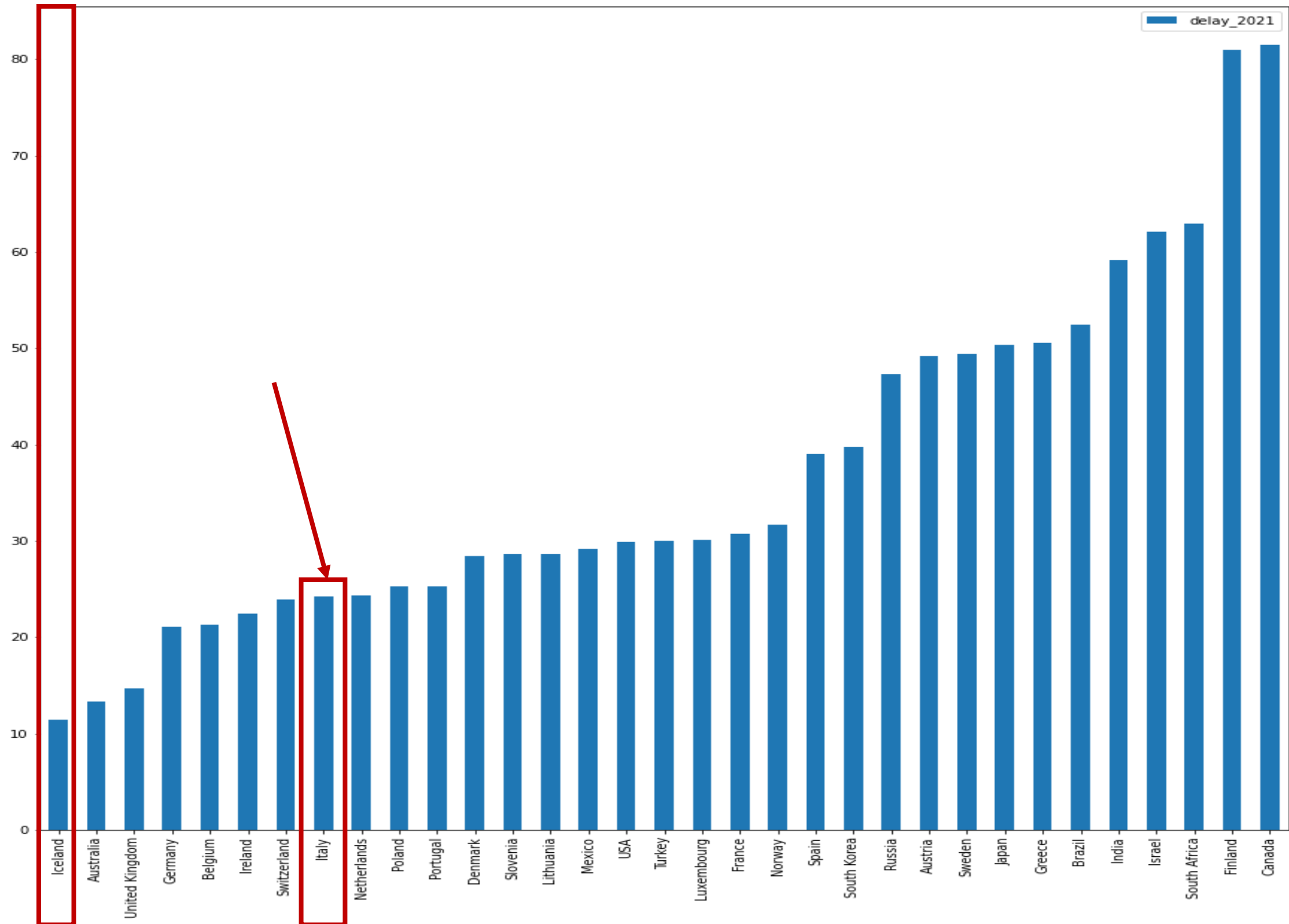
Growth of deposited sequences (observed in ViruSurf)



Depositions vs. COVID-19 cases

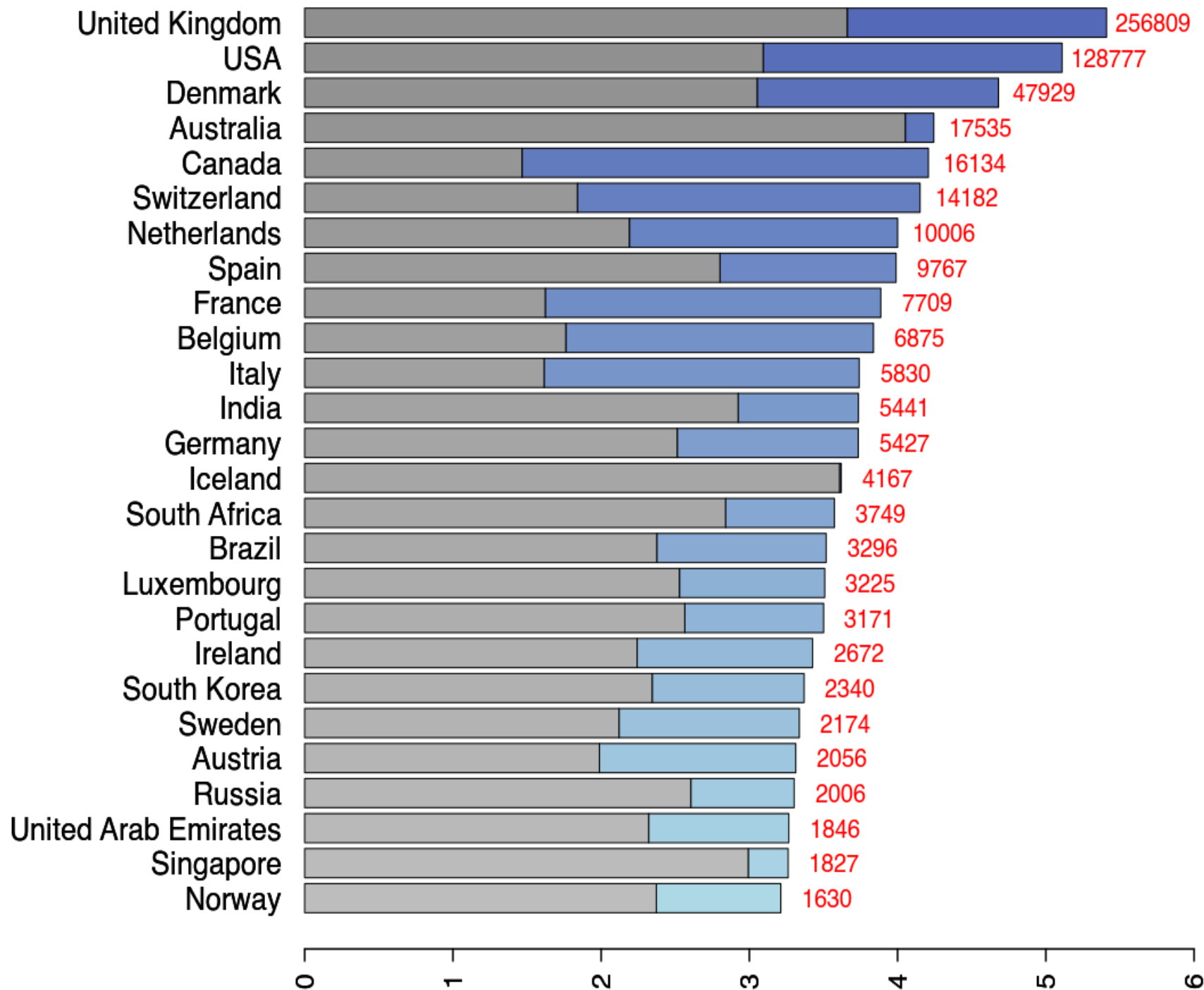


Delay between collection date and sequence deposition (2021)



Uneven sampling in time

SARS-CoV-2 Genomes in GISAID, by country(March 1st 2021)



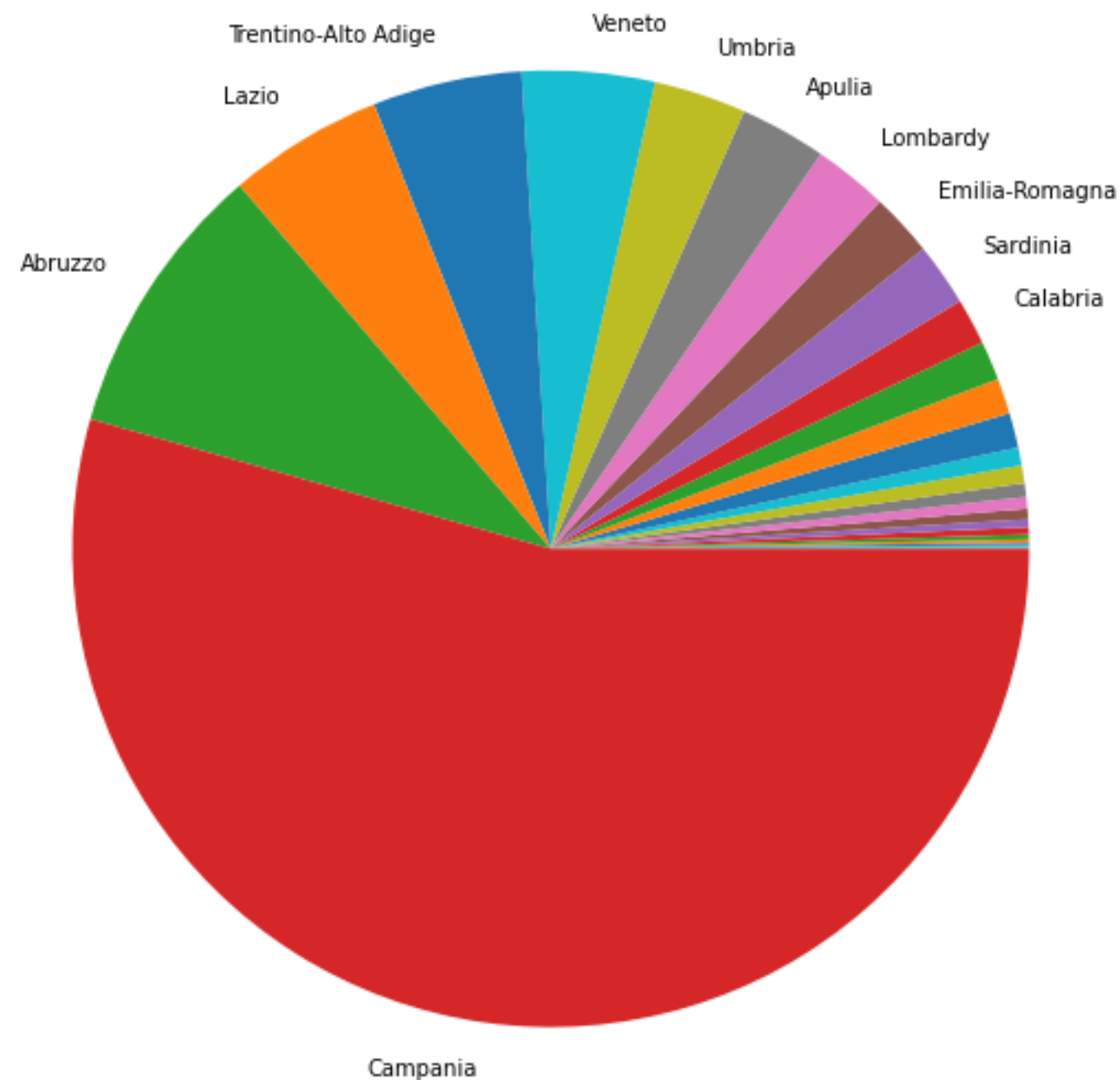
Number of SARS-CoV-2 genomic sequences available from the GISAID portal. Only countries for which 1000 or more genomes are available are represented.

Blue: “novel” genomes, submitted between Jan 24th and Feb 26th 2021. **Gray:** genomes available in GISAID before Jan 24th 2021.

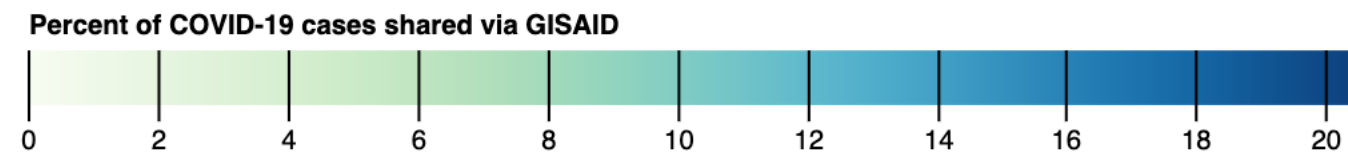
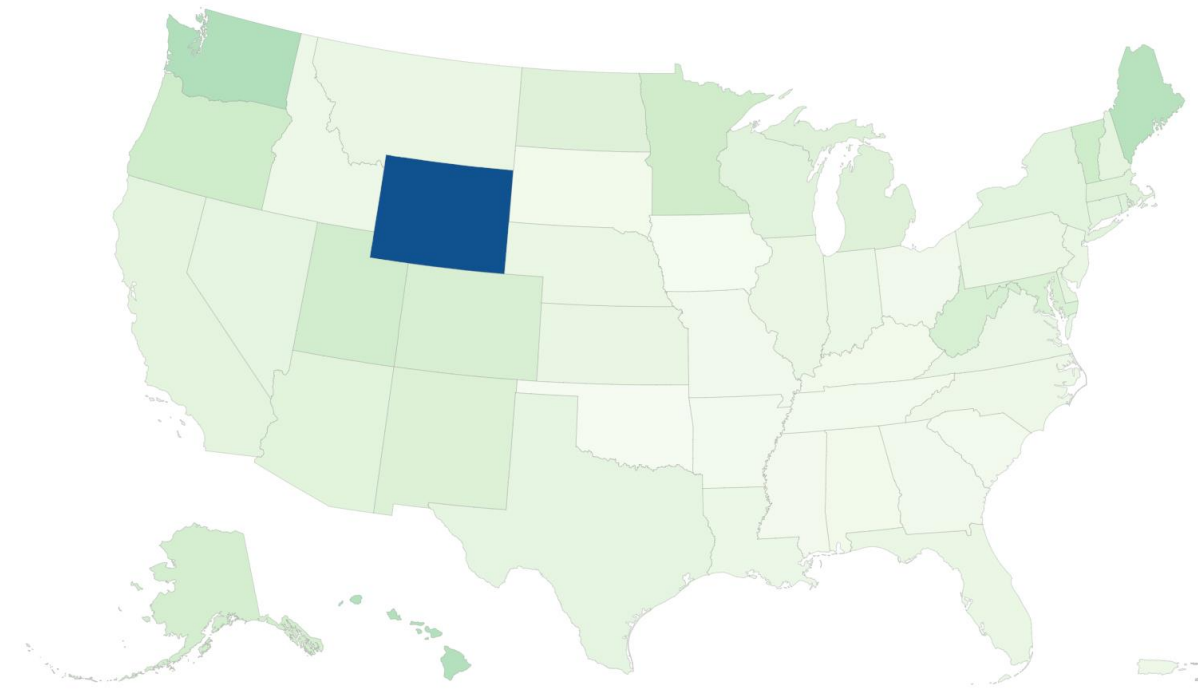
Uneven geographical sampling

As of July 1st, 2021, Italy has submitted 36888 sequences to GISAID.

More than 50% comes only from 1 of the 20 Italian regions, Campania (Naples region)



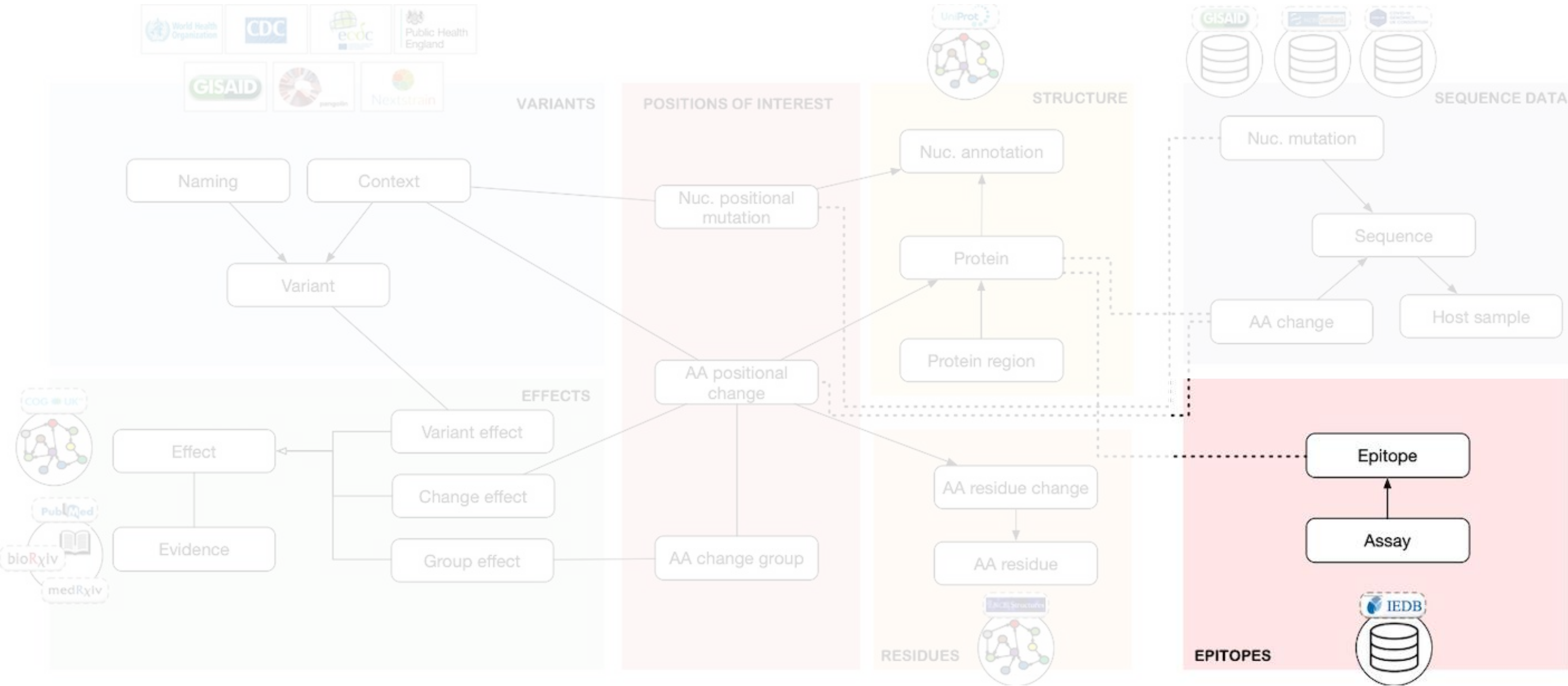
Uneven sampling in the US



Correct as of 02 August 2021 3:31:hrs UTC+2

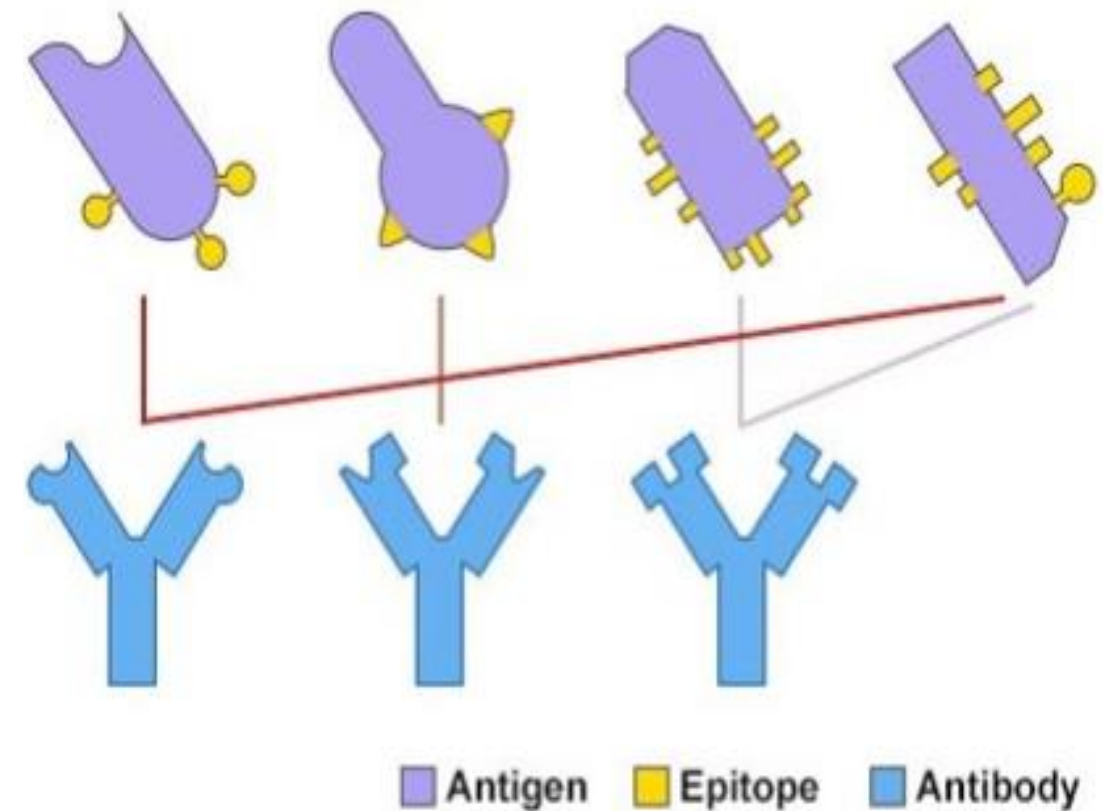
State	Genomes shared	Reported COVID-19 cases	% of cases sequenced and shared	Median days to deposition
<i>Wyoming</i>	12,598	65,127	19.34	137
<i>Washington</i>	33,574	473,076	7.1	25
<i>Hawaii</i>	2,946	42,862	6.87	36
<i>Maine</i>	4,611	70,463	6.54	22
<i>Vermont</i>	1,149	24,889	4.62	19
<i>Oregon</i>	10,110	219,755	4.6	46
<i>Minnesota</i>	27,675	612,701	4.52	24
<i>Utah</i>	18,610	432,467	4.3	69
<i>Alaska</i>	3,005	75,486	3.98	18
<i>West Virginia</i>	6,268	167,016	3.75	26
<i>Colorado</i>	20,935	575,082	3.64	28
<i>Maryland</i>	15,676	469,095	3.34	26
<i>District of Columbia</i>	1,644	50,398	3.26	32
<i>New Mexico</i>	6,798	210,416	3.23	30
<i>Michigan</i>	28,922	1,011,106	2.86	26
<i>North Dakota</i>	3,186	111,674	2.85	35

The Knowledge Model: EPITOPES



Other biological concepts: Epitopes

- Epitopes are strings of amino acids from an antigen (e.g. derived from virus protein) that can be recognized by antibodies or B/T-cell receptors to provoke an immune response.



Source: <https://vaccsbook.com/>

The Immune Epitope Database (IEDB)

- The largest open-source collection of epitopes for many species. IEDB has over a 1M epitopes.
- Currently includes about six thousands epitopes for SARS-CoV-2; almost 3 thousands refer to the Spike protein.

Vita, R., Mahajan, S., Overton, J.A., Dhanda, S.K., Martini, S., Cantrell, J.R., Wheeler, D.K., Sette, A. and Peters, B., 2019. The immune epitope database (IEDB): 2018 update. *Nucleic acids research*, 47(D1), pp.D339-D343.

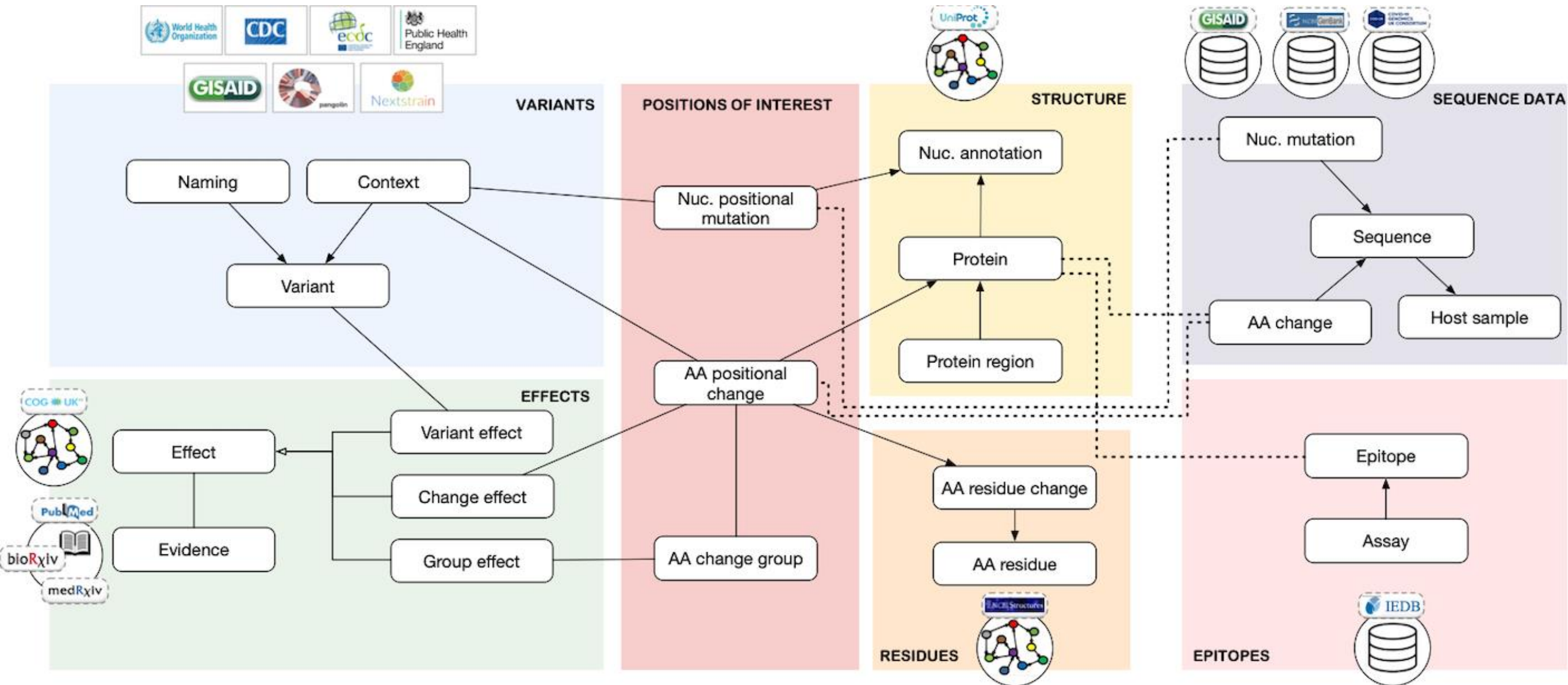
- From what is known about Pfizer and Moderna and vaccines, they use all available epitopes on the Spike protein at the time of their

Dae Eun Jeong, Matthew McCoy, Karen Artilles, Orkan Ilbay, Andrew Fire*, Kari Nadeau, Helen Park, Brooke Betts, Scott Boyd, Ramona Hoh, and Massa Shoura. Assemblies of putative SARS-CoV2-spike-encoding mRNA sequences for vaccines BNT-162b2 and mRNA-1273. <https://virological.org/t/assemblies-of-putative-sars-cov2-spike-encoding-mrna-sequences-for-vaccines-bnt-162b2-and-mrna-1273/663>

- A mutation in an epitope might compromise the epitope's recognition from the immune system. E.g., E484K, E484Q, E484P are associated with the reduction of neutralization titres, possibly generating an immune escape.

Harvey, W.T., Carabelli, A.M., Jackson, B., Gupta, R.K., Thomson, E.C., Harrison, E.M., Ludden, C., Reeve, R., Rambaut, A., Peacock, S.J. and Robertson, D.L., 2021. SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology*, 19(7), pp.409-424.

Putting all together again



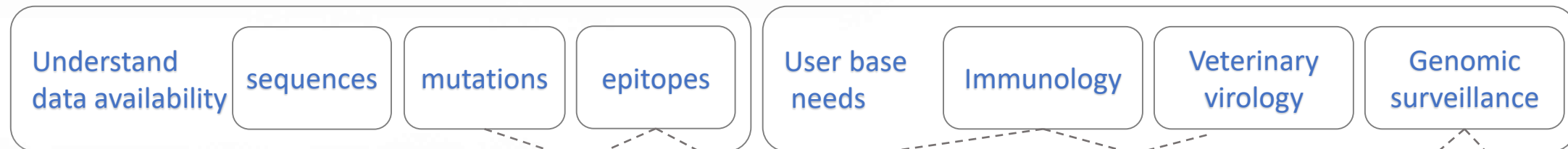
Part 2: Our Contributions

1. Understanding data --- covered
2. Modeling data
3. Building repositories
4. Building tools
5. Using tools

From requirements elicitation to action



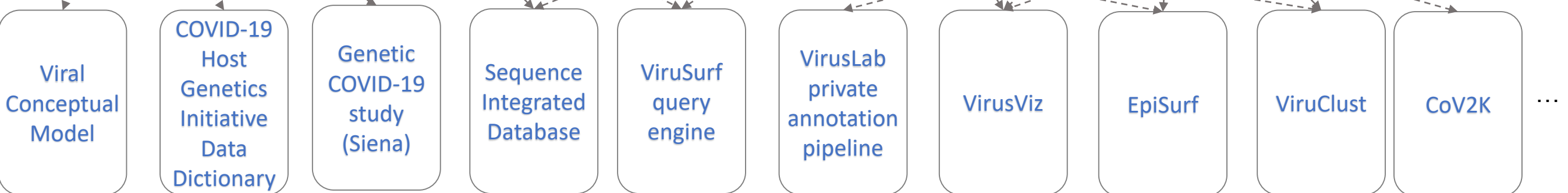
Identification of area of interest



Requirements analysis



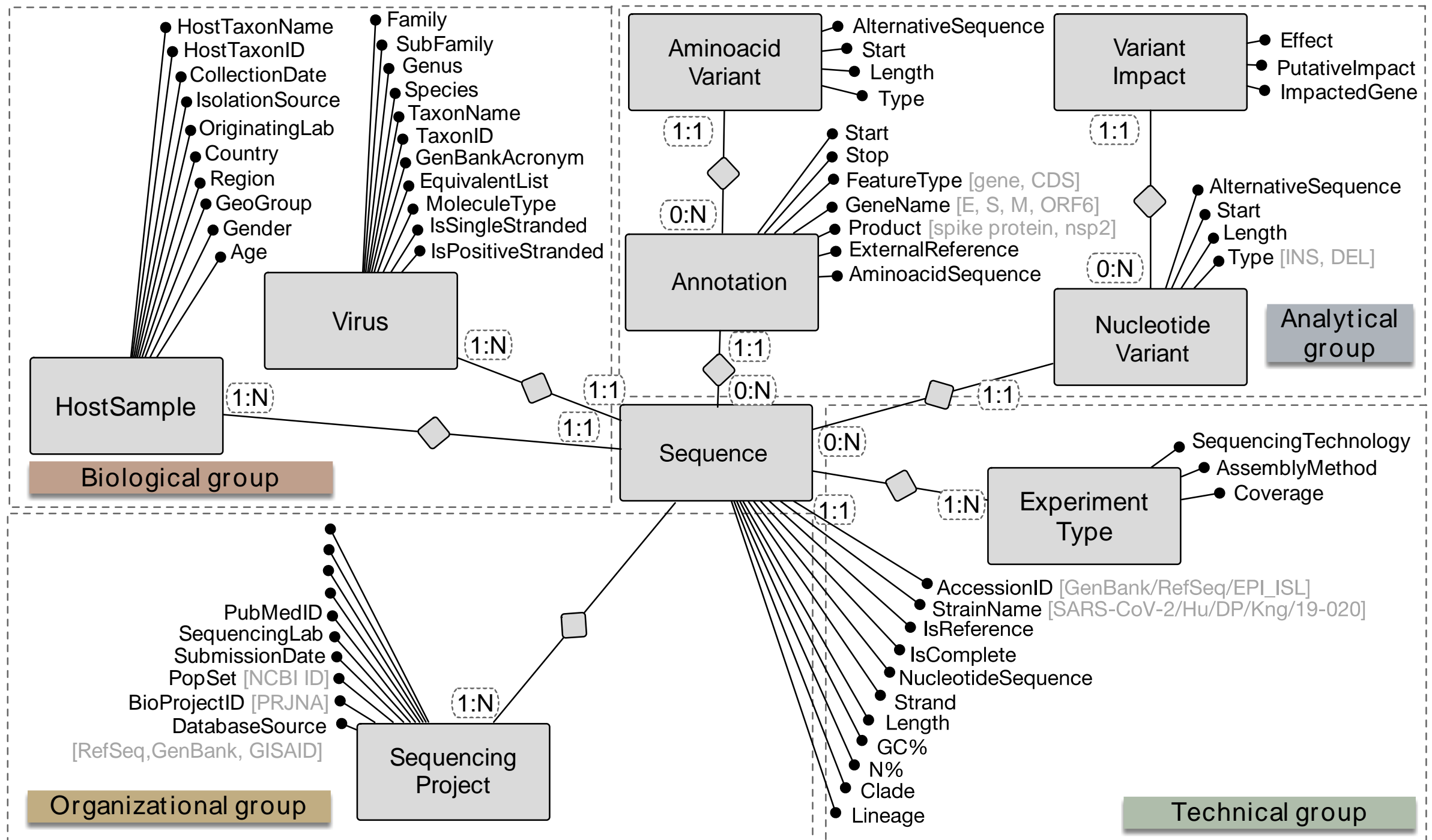
Systems and studies design



Viral 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)



Phenotype Data Dictionary

We coordinated the efforts to produce the patient phenotype definition that will be used as a standard to collect and harmonize data from studies.

Focus on Patient and his/her data at:

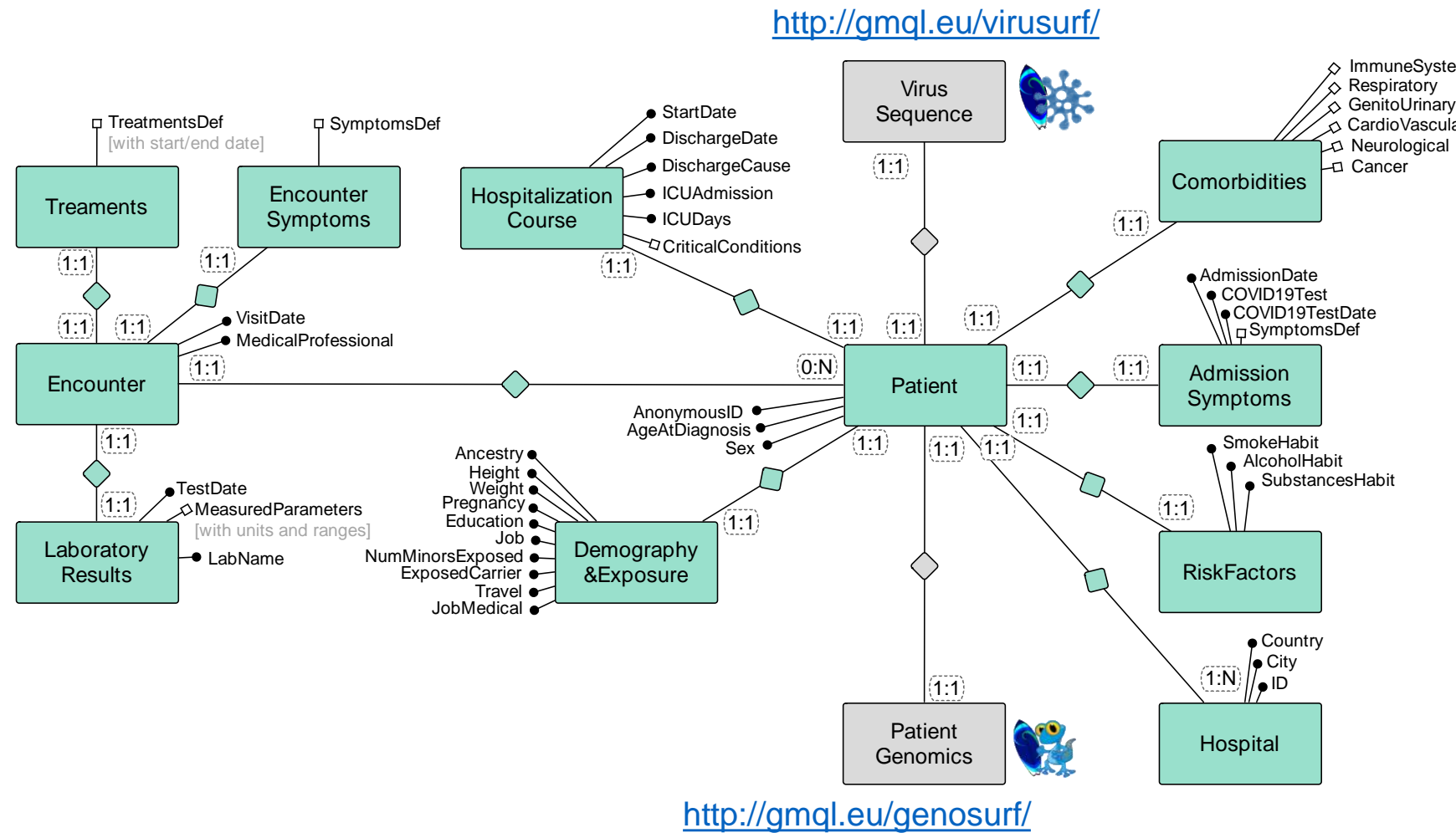
- Admission
- Course of hospitalization
- Discharge

Collected data will be hosted by **EGA** (European Genome-phenome Archive) of EMBL-EBI

[FREEZE-1 DATA DICTIONARY](#) released on April 2020

[FREEZE-2 DATA DICTIONARY](#) released on August 2020 with additions and improvements

Used for studies reported in the following publication:



COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. Nature (2021). <https://doi.org/10.1038/s41586-021-03767-x>

Working group

Stefano Ceri and Anna Bernasconi (Politecnico di Milano)
Alessandra Renieri and Francesca Mari (Università degli Studi di Siena)

Tools at a glance

VirusSurf: search engine for selecting sequences from metadata, nucleotide changes and amino acid changes

- Endpoints: VirusSurf for GenBank/CogUK, VirusSurf_GISAID for GISAID

VirusViz: client-side visualizer offering rapid comparative analysis among viral populations extracted from VirusSurf

EpiSurf: searching engine for selecting sequences and epitopes and integrating them for testing epitope stability

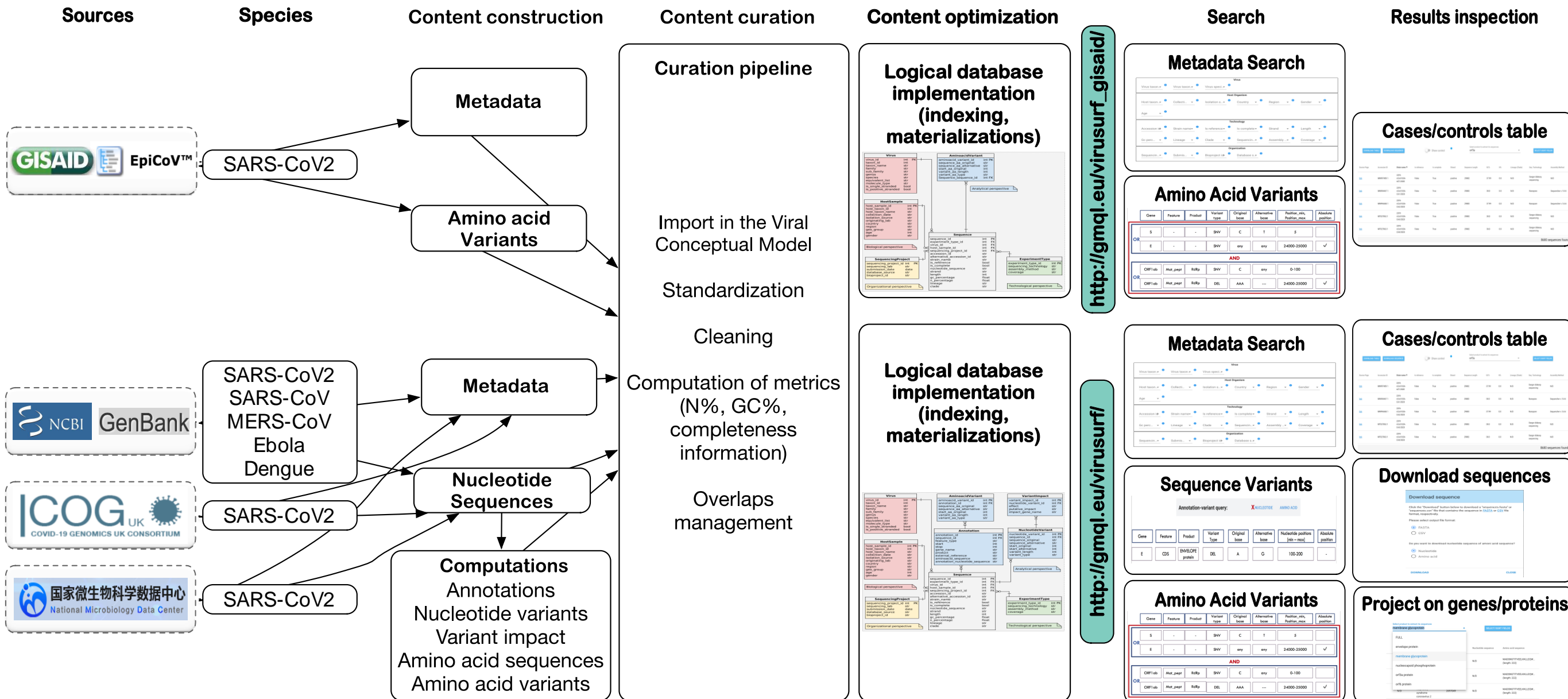
- Endpoints: EpiSurf for GenBank/CogUK, EpiSurf_GISAID for GISAID

VirusLab: virtual laboratory accepting RNA sequences in standard format and producing nucleotide mutations and amino acid changes

VirusClust: tool for fine tuning of clusters of sequences, used for fine-grain surveillance

CoV2K: knowledge model + library for accessing SARS-CoV2 data and knowledge

Data integration pipelines from data sources (July 30th, 2021)



Virusurf-GISAID content

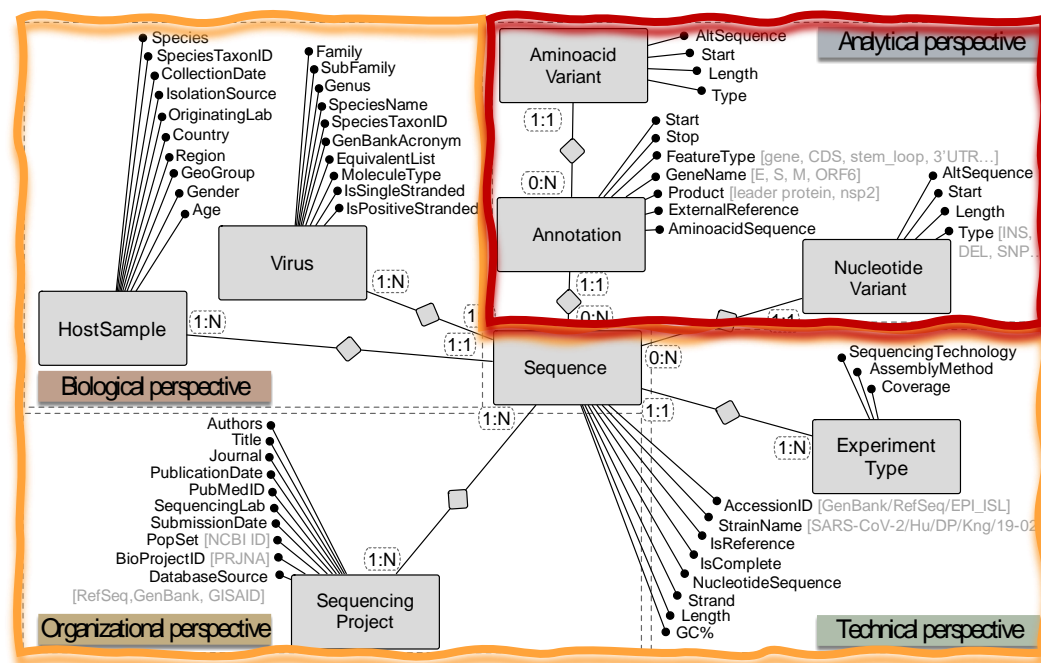
- GISAID EpiCoV™ db ~ 2.5M sequences (~ 1.9M are GISAID specific)

Virusurf content

- GenBank ~ 717K sequences (SARS-CoV-2)
- GenBank ~ 35K sequences (other viruses)
- COG-UK ~ 598K sequences
- NMDC ~ 300 sequences

Virusurf search system

<http://gmql.eu/virusurf/>



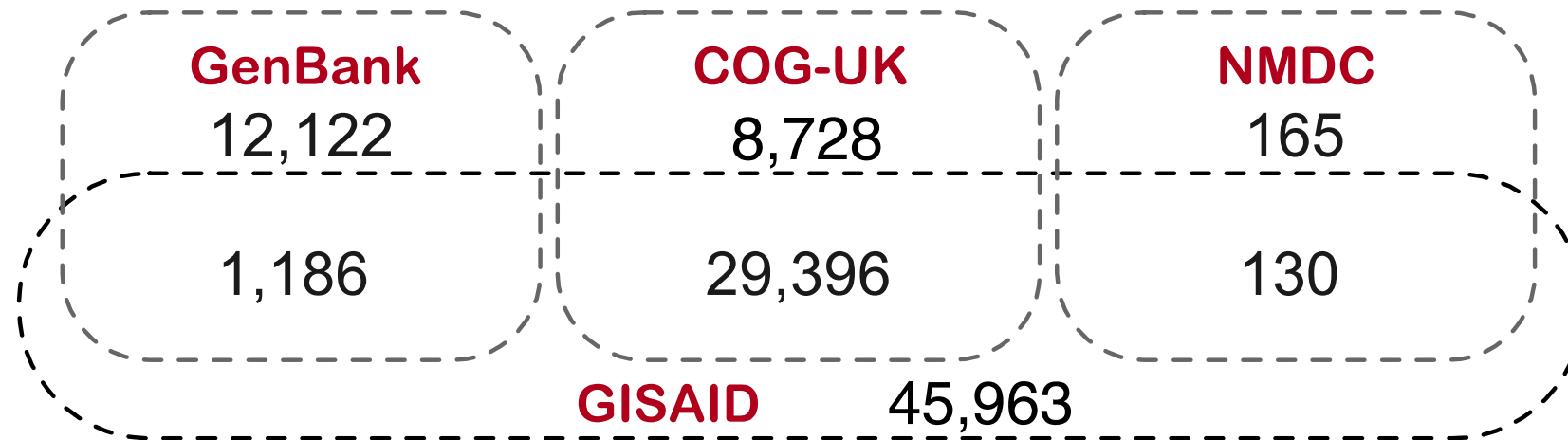
The screenshot shows the Virusurf web interface with the following sections:

- Top bar**: Navigation links for VIRUSURF GISAID, GENOSURF, DATA CURATION, WIKI, VIDEO, SURVEY, ACKNOWLEDGEMENTS, and CONTACTS.
- Metadata search**: A search interface with a "CLEAR YOUR QUERY" button and a dropdown for "Choose a predefined query". It includes filters for Virus (Virus taxon ID, Virus species), Host Organism (Host taxon name, Collection date, Isolation source, Country, Region, Gender), Technology (Accession ID, Strain name, Is reference, Strand, Sequence Length), and Organization (Submitting Lab, Submission date, BioProject ID, Database source).
- Variant search**: A section for searching variants with filters for Amino acid query, Nucleotide query, and Amino acid query. It includes a dropdown for "Change type" and a "Position range" filter.
- Results visualization**: A table showing search results with columns for Source Page, Accession ID, Strain name, Is reference, Is complete, Strand, Sequence Length, GC%, N%, Lineage (Clade), Seq. Technology, Assembly Method, Coverage, and Submission date. The table shows 14 sequences found.

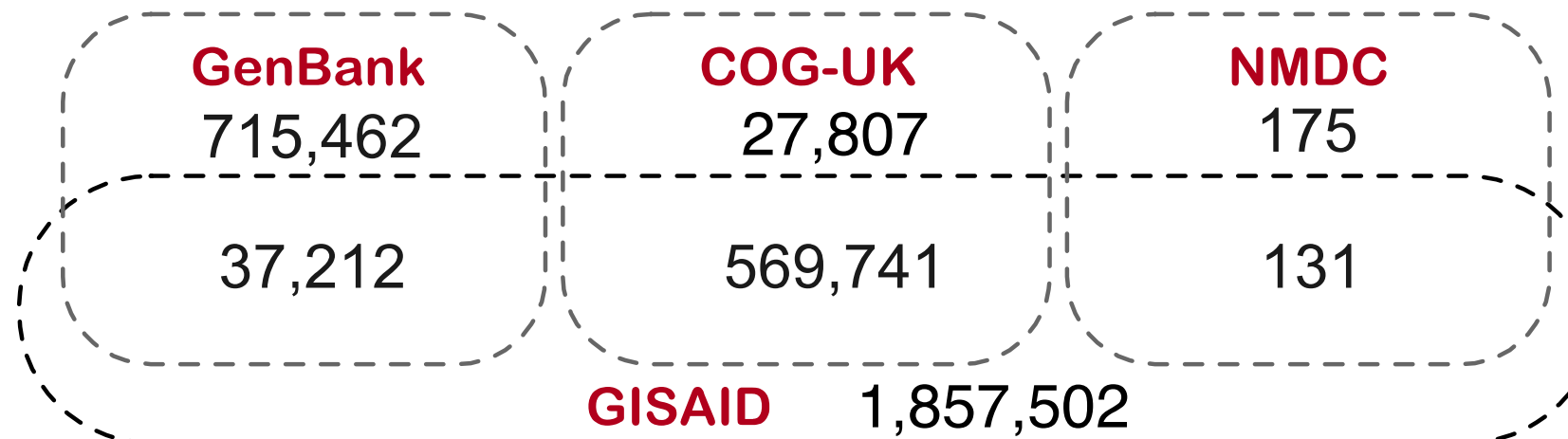
Four sections:

- 1) a menu bar to access the different services;
- 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.

Sequencing numbers: one year ago and today



Aug. 4th,
2020



July 25th,
2021

Top-runners in GISAID depositions (July 25th, 2021)



VirusSurf GISAID

enabled by data from



Last update date: 2021-07-25

CLEAR YOUR QUERY

APPLY YOUR SEARCH

APPLY GISAID SPECIFIC

Chose a predefined query

Metadata search

Sequences of the virus responsible for COVID-19 (hCoV-19)

Host Organism

Host (Host taxon ... Collection date Specimen Source ... Location (Geo gro... Location (Country) Location (Region)

Originating lab

Sequence properties

Is reference Is complete Strand Sequence Len...

Lineage Clade

Organization

Submitting Lab Submission da...

usa	654152
united kingdom	602166
germany	143924
denmark	125281
canada	85690
sweden	76444

Virusurf example of use

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.

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

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

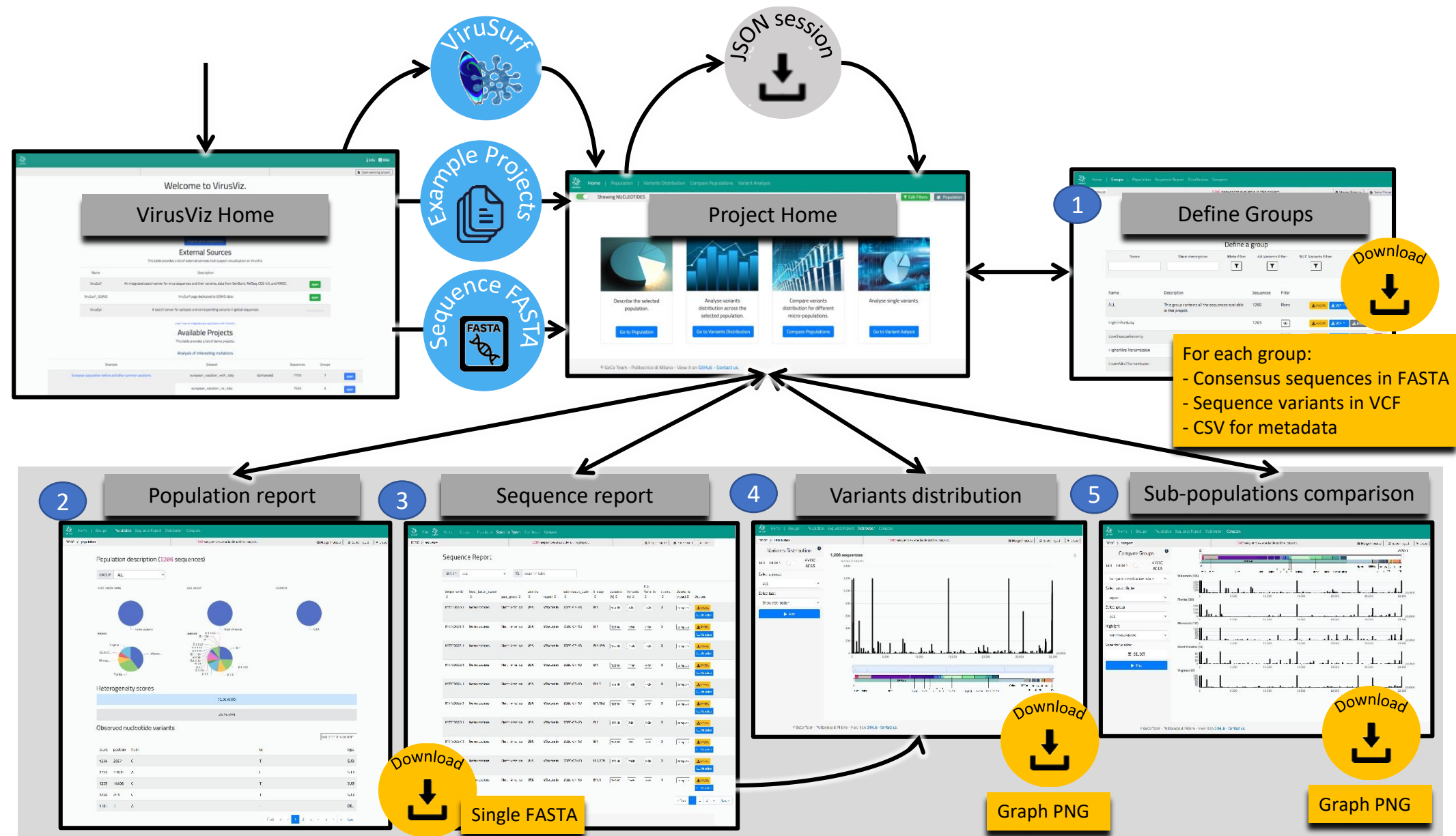
G614 genotype:

- not detected in February 2020
- found with low frequency in March 2020
- increased rapidly from April onward

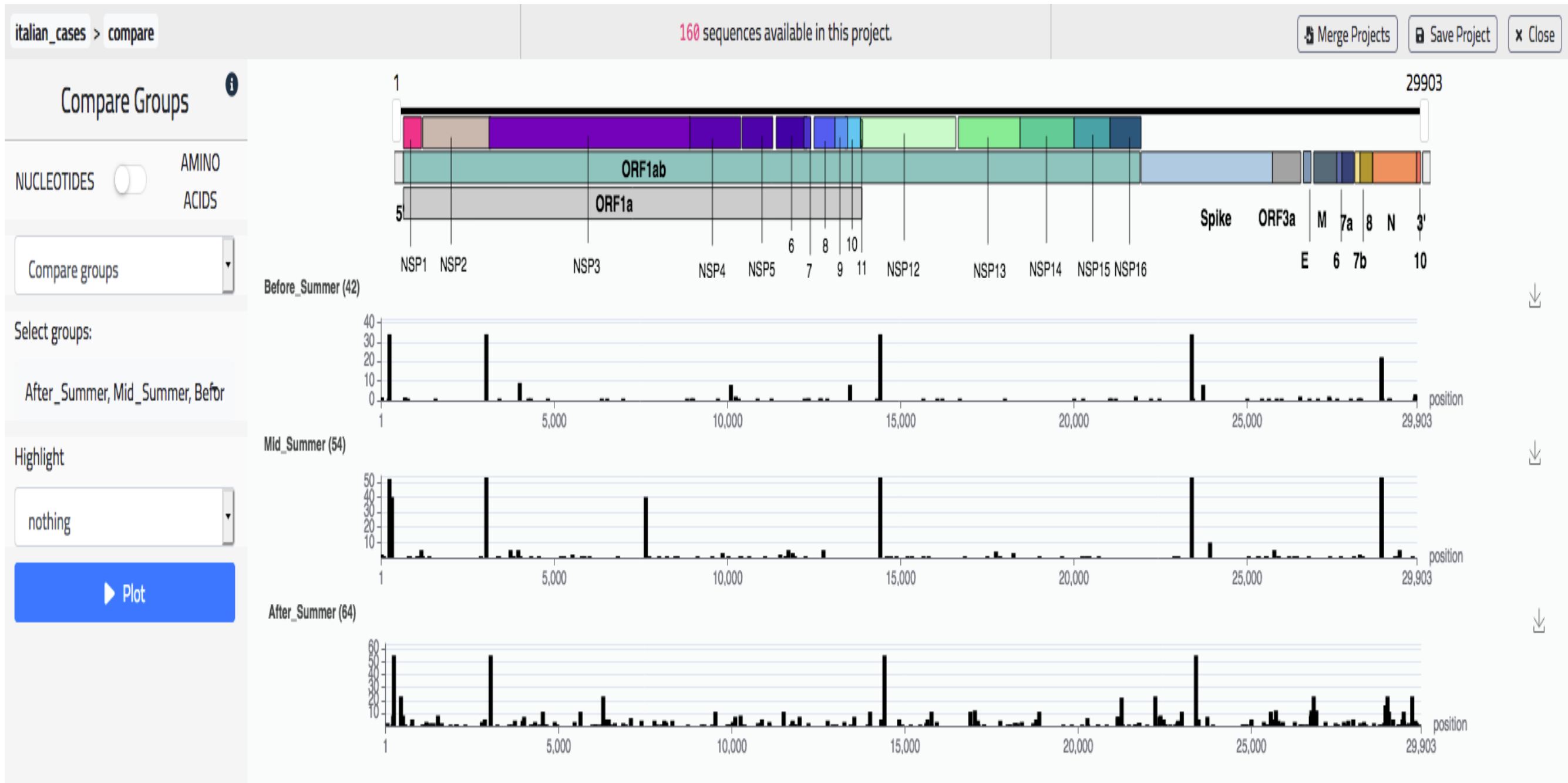
→ indicating a transmission advantage over viruses with D614G

	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
D614G%	58.56%	63.02%	87.65%	84.54%
total	61.59%		86.26%	

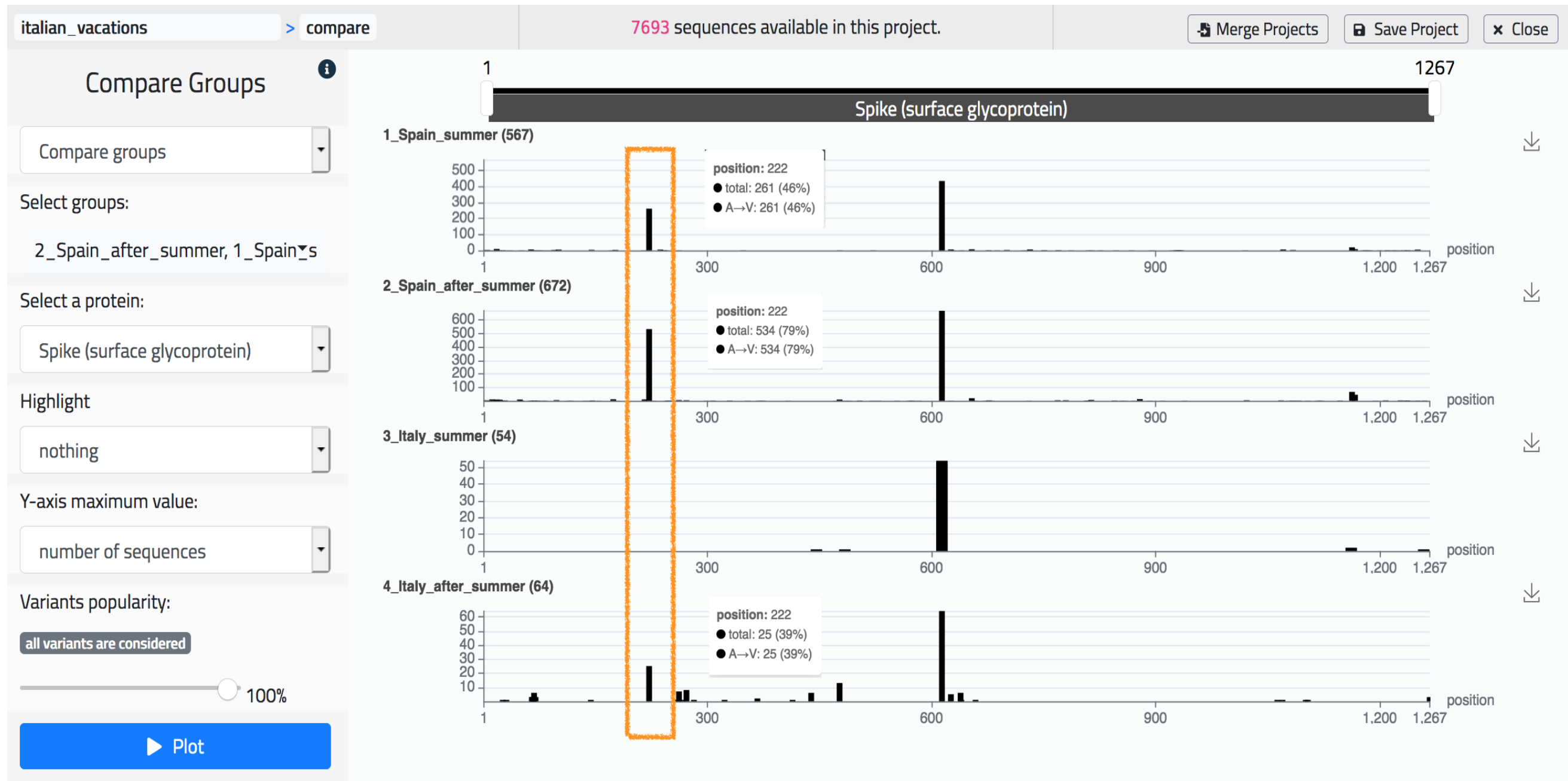
VirusViz: comparative visualization of nucleotide/amino acid variants



Italians returning from summer 2020 vacations present many mutations



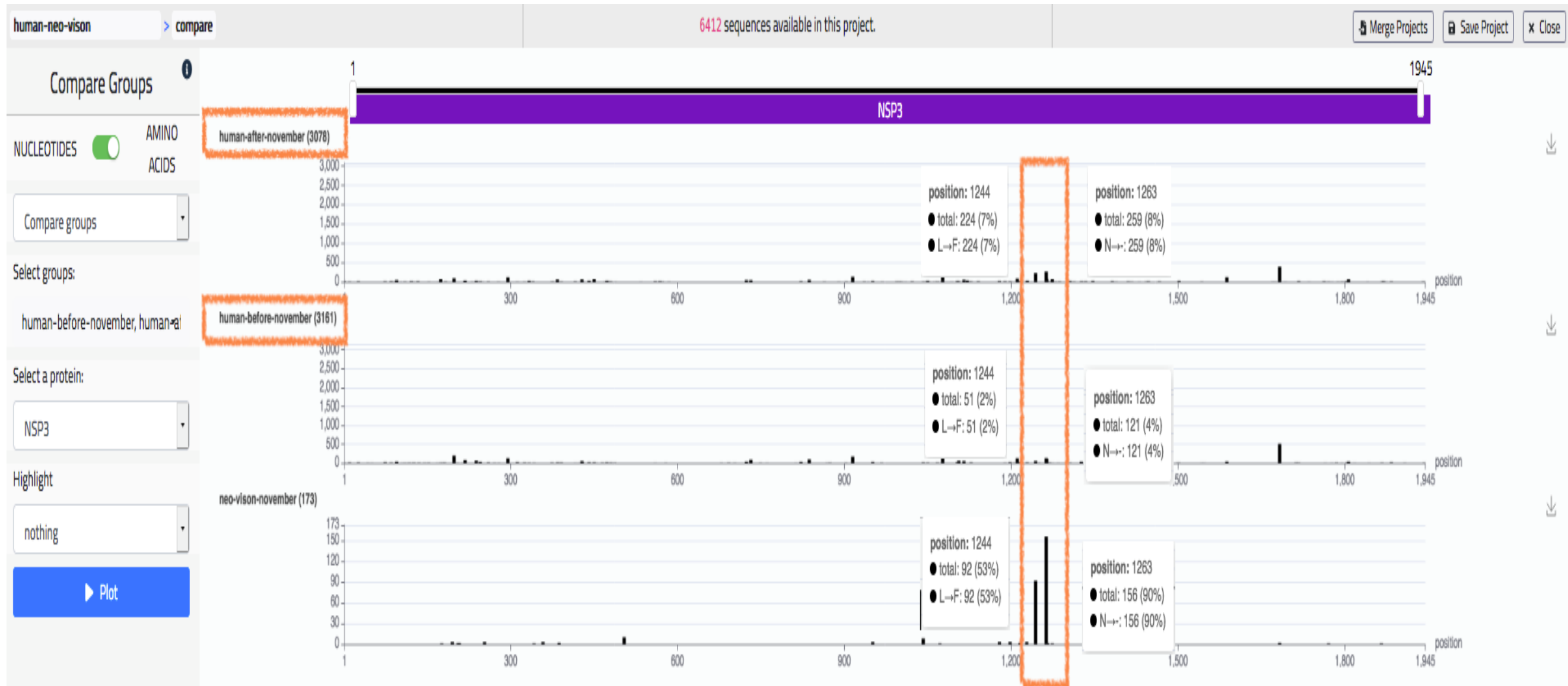
The Spike mutation A222V arrived from Spain after Summer



Hodcroft, E.B., Zuber, M., Nadeau, S., Vaughan, T.G., Crawford, K.H., Althaus, C.L., Reichmuth, M.L., Bowen, J.E., Walls, A.C., Corti, D. and Bloom, J.D., 2021. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. Nature, pp.1-9.

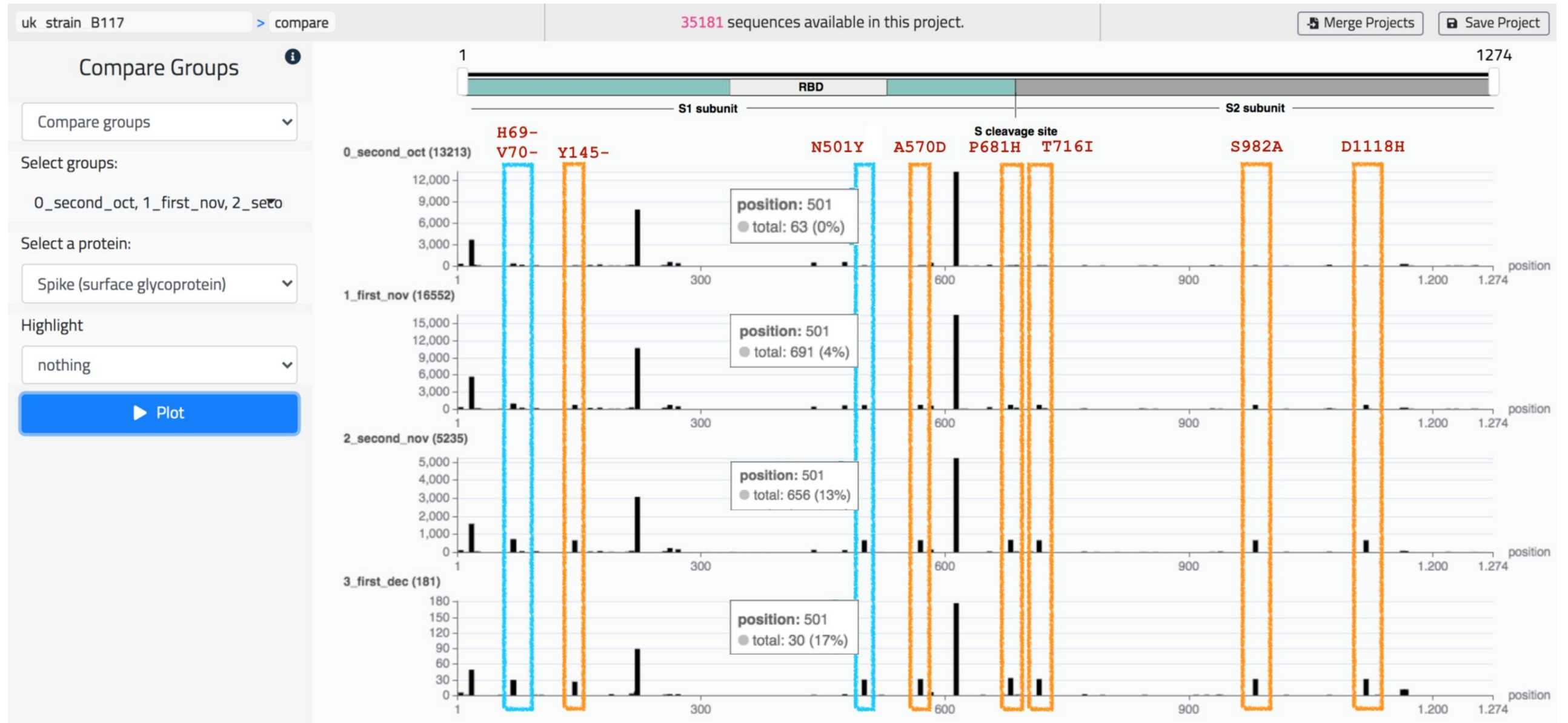
Minks mutation in Denmark

An NSP3 mutation may have been transmitted from minks to humans after the first spillover in Denmark (Nov. 2020)



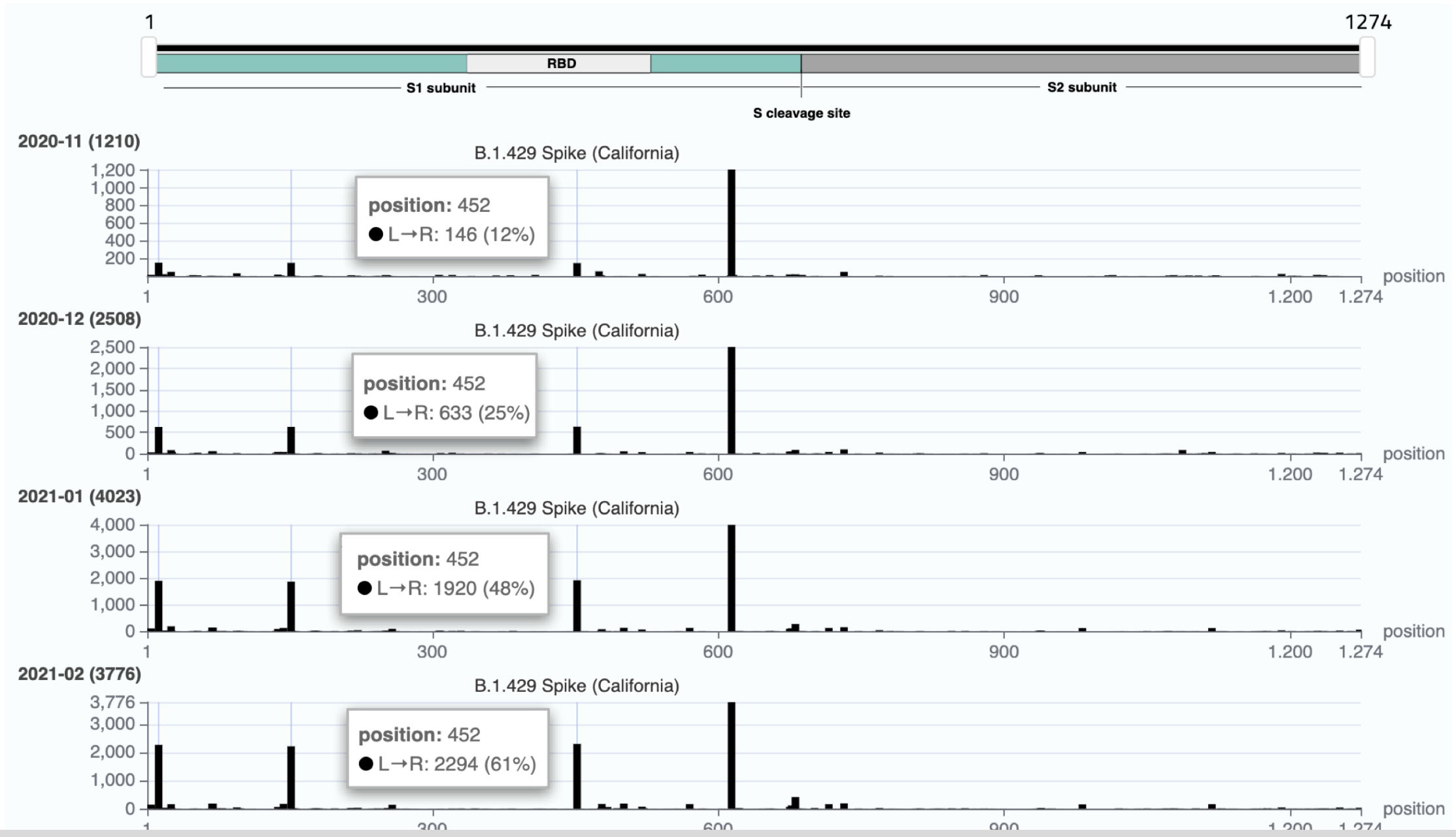
UK variant (Alpha)

The first alert of the “Alpha variant” was in a post of Dec. 2020 (indicating N501Y and N69-/V70- on Spike). Shortly later more amino acid changes were added.

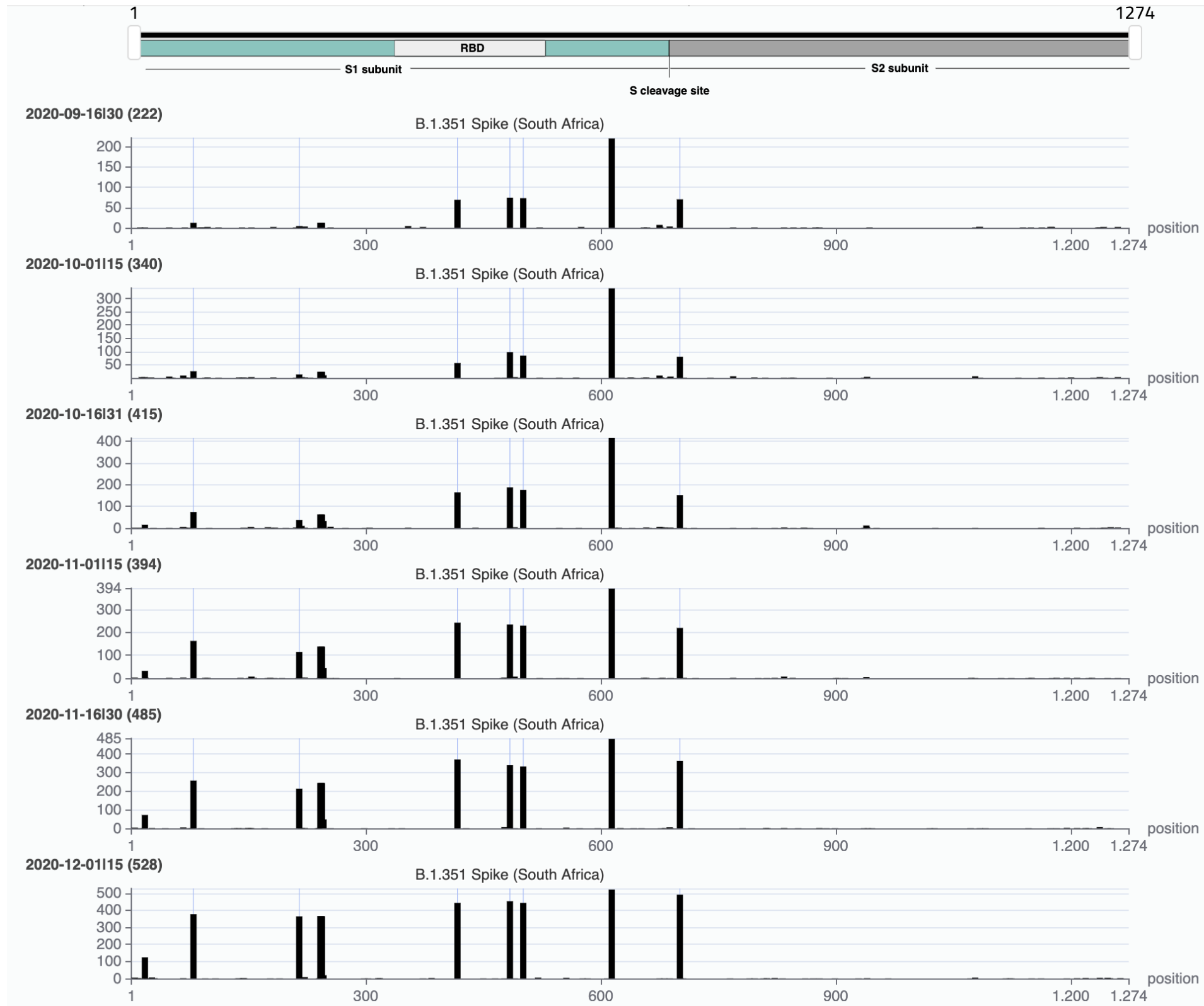


Californian variant (Epsilon)

The first alert of the “Epsilon variant” was in a letter dated Feb. 11, 2021, indicating 3 amino acid changes on Spike: S13I, W152C, L452R.

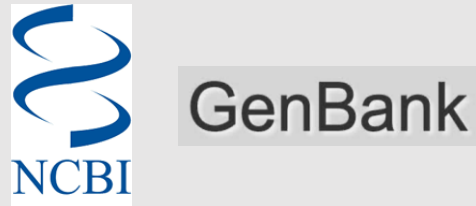


South African variant (Beta)



EpiSurf

Sequence population search



Novel "Severe acute respiratory syndrome coronavirus 2" sequences from "Homo sapiens" as host are preselected. If you are interested in other virus(es), please change it from the dropdown menu below:

Sequence population search condition: taxon_name: ["severe acute respiratory syndrome coronavirus 2"], host_taxon_name: ["homo sapiens"]

Virus	
Virus taxon ID	severe acute respiratory syndrome coronavirus 2
Virus species	

Host Organism					
Host taxon name	Collection date	Isolation source	Continent	Country	Region
homo sapiens					
Gender	Age				

Sequence properties and technology					
Is reference	Is complete	Strand	Sequence Len...	GC%	N%
Lineage	Sequencing techn...	Assembly method	Coverage		

Organization			
Submitting Lab	Submission date	BioProject ID	Database source

Epitope search

CUSTOM EPITOPES USE IEDB EPITOPES WITHOUT VARIANT COUNTS USE IEDB EPITOPES WITH VARIANT COUNTS

IEDB Epitope search

Epitope search condition: protein: ["Spike (surface glycoprotein)"], assay_type: ["B cell"]

Protein Name	Assay	HLA restriction	Is Linear	
Spike (surface glycoprotei	B cell		true	724
			false	290

Response Frequency Position Range Epitope IEDB ID

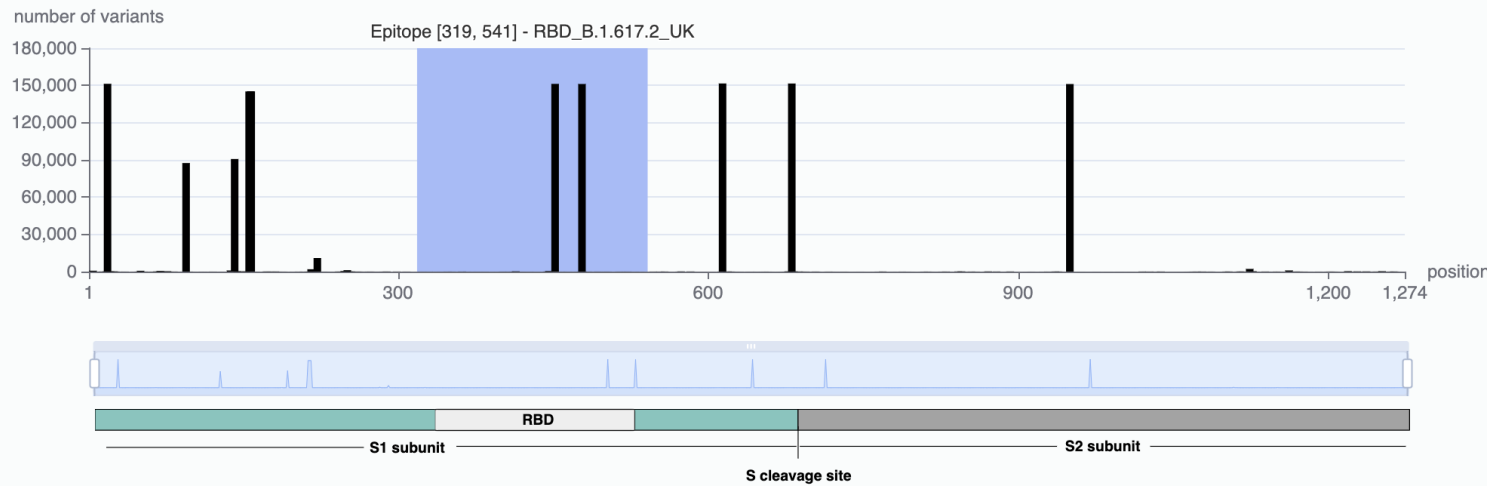
CLEAR EPITOPE QUERY

Checking the Delta mutations in UK over important epitope ranges

Spike RBD mutations and immune escape

Of all RBD residues for which substitutions affected recognition by convalescent sera, DMS identified E484 as being of principal importance, with amino acid changes to K, Q or P reducing neutralization titres by more than an order of magnitude³⁹.

151,559 sequences



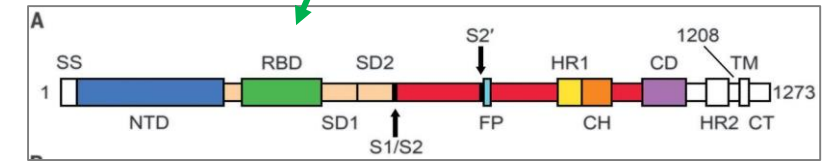
position: 452

- L→R: 151223 (100%)
- L→W: 1 (0%)

position: 478

- T→K: 151127 (100%)
- T→I: 1 (0%)

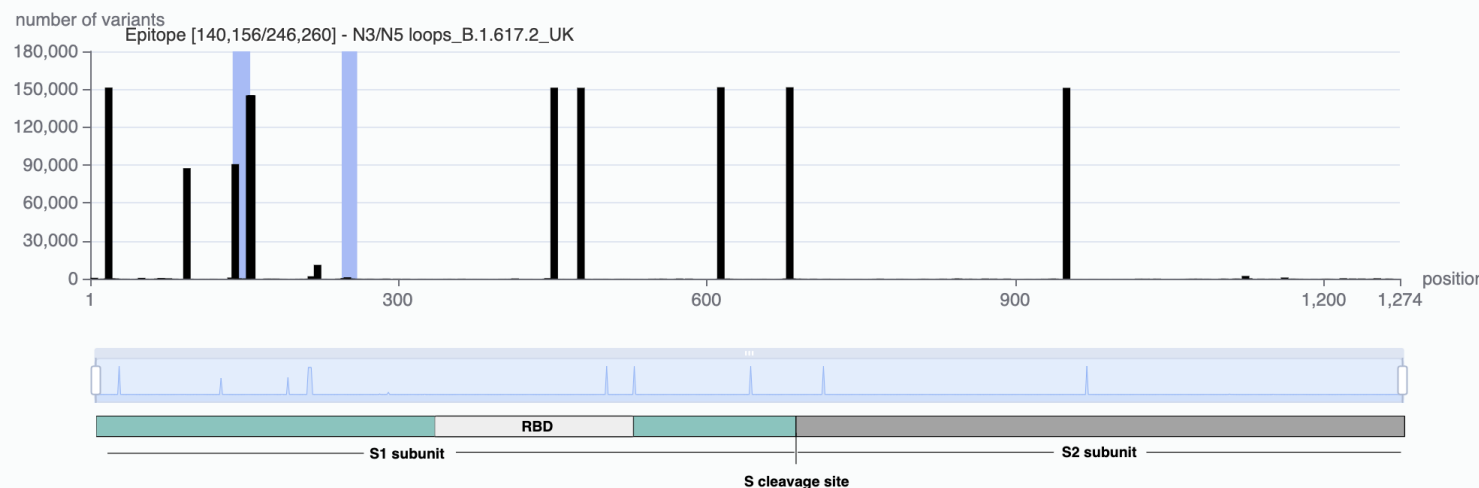
Regionⁱ 319 - 541 Receptor-binding domain (RBD)



Spike NTD mutations and immune escape

In the NTD, most of the evidence for immune evasion focuses on a region centred at a conformational epitope consisting of residues 140–156 (N3 loop) and 246–260 (N5 loop)

151,559 sequences



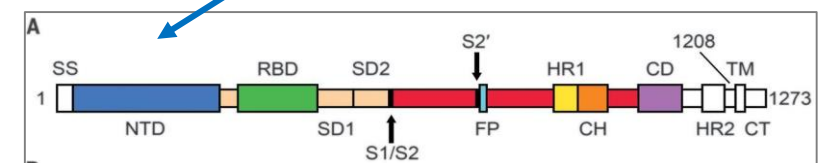
position: 142

- G→V: 1 (0%)
- G→-: 4 (0%)
- G→D: 90676 (60%)
- G→I: 1 (0%)
- G→A: 2 (0%)
- G→Y: 3 (0%)
- G→N: 4 (0%)

position: 156

- E→-: 145034 (96%)
- E→G: 4 (0%)

Domainⁱ 13 - 303 BetaCoV S1-NTD



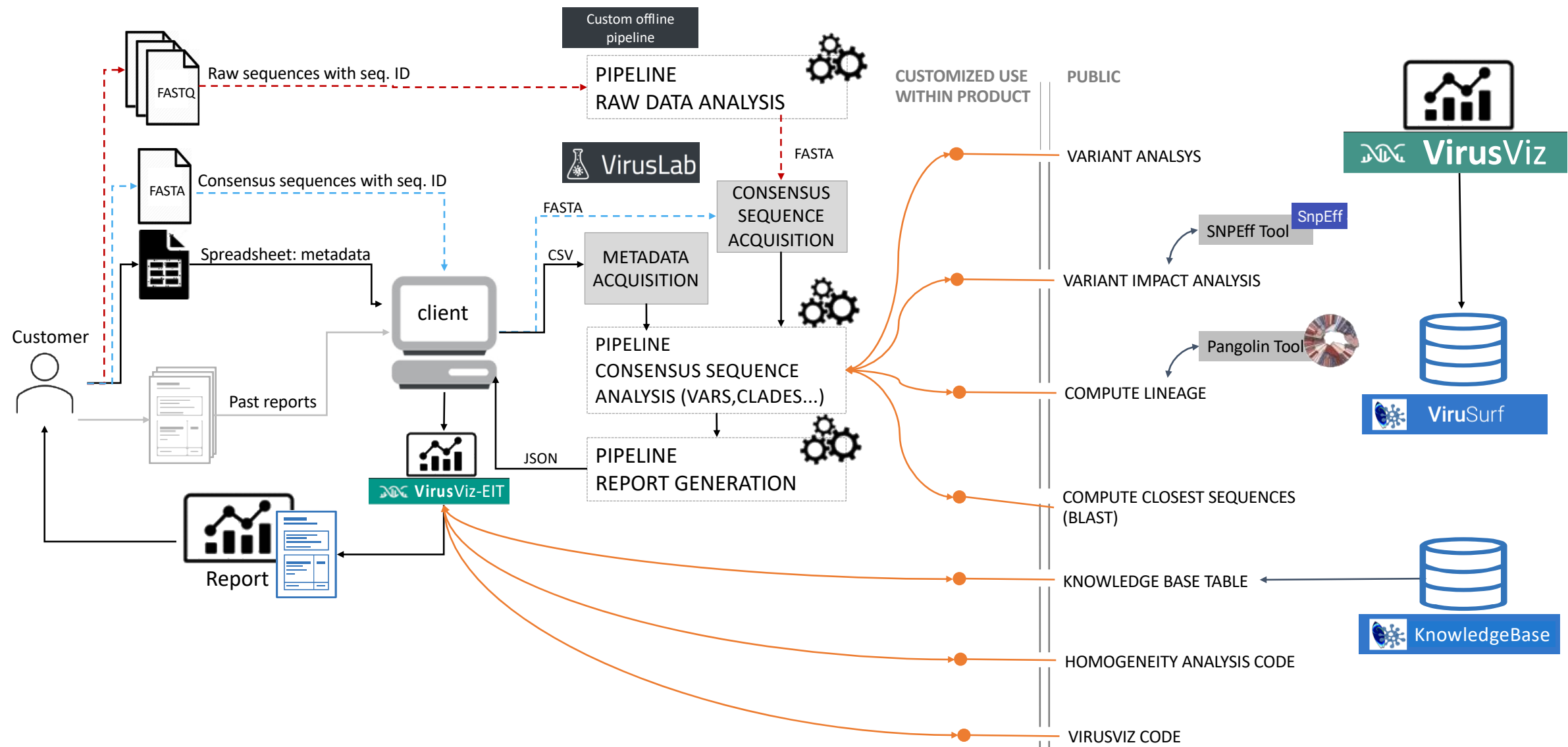
VirusLab: Packaging of customizable services for use within protected sites

For supporting a laboratory wishing to perform secondary (raw data) analysis and to add sensible metadata (e.g. clinical), still using our data and knowledge bases and visualization tools

CUSTOMER

QUANTIA CONSULTING (SERVICE PROVIDER)

POLITECNICO DI MILANO/DELFT UNIV.



Finally: data analysis - A Study on the Delta (VOC) variant

- First identified in the state of Maharashtra (India) in late 2020
- Currently the most widespread variant of SARS-CoV-2 worldwide
- 50 to 60% more efficient in infecting human cells than any other variant of the virus, with enhanced ability to escape immune response (see Otto et Al. 2021, PM: 34314723)

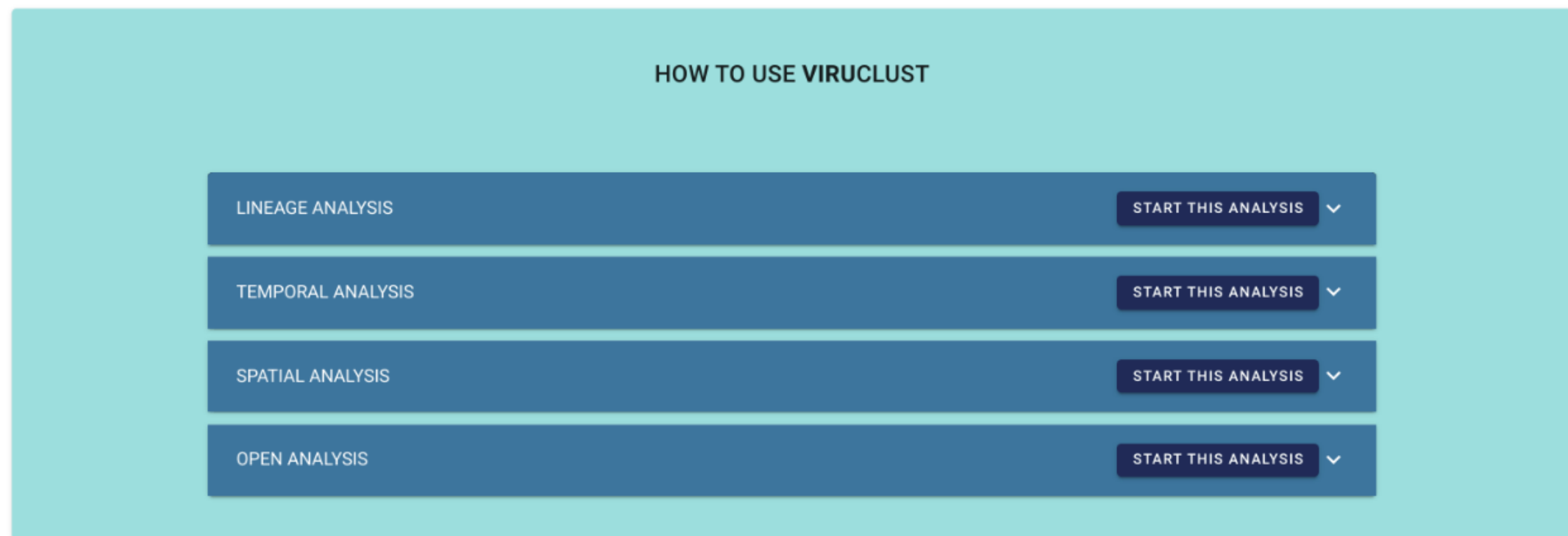
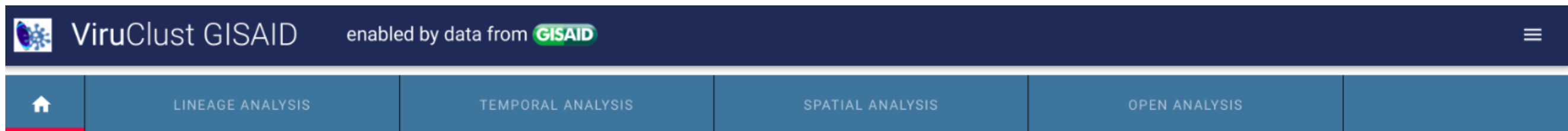
How can we monitor the properties/evolution of Delta?

- Identify mutations in the genome unique to Delta (w.r.t. other variants)
- Study evolution in time (when were all the mutations acquired)
- study evolution in space (novel mutations acquired in different countries)
- focus on the “Spike” protein:
 - determines the <infectivity> of the virus
 - main target of the immune system (hence used by all the vaccines)

Introducing our last tool: ViruClust

Supporting pairwise comparison (target/background) of viral populations in

- **lineages** (with different mutations)
- **time** (time intervals vs monthly/weekly periods)
- **space** (different continents, countries, regions)
- **any of the above** (custom analyses)



http://geco.deib.polimi.it/viruclust_gisaid/

What happened in the “early” evolution of Delta? are there mutations in India (the likely place of origin) that are not observed elsewhere? **Spatial analysis**

- India VS Rest of Asia
- India VS Rest of the World

LINEAGE ANALYSIS TEMPORAL ANALYSIS **SPATIAL ANALYSIS** OPEN ANALYSIS

PICK A LINEAGE, PLACE AND INTERVAL OF TIME

B.1.617.2 (Delta) X

Please select a region to pick a province
(multiple values may be selected here)

Asia X India X region province

Comparison: India vs Asia Exclude one or more places from the background: country

Overview of the comparison

Home icon

LINEAGE ANALYSIS TEMPORAL ANALYSIS **SPATIAL ANALYSIS** OPEN ANALYSIS

India

TARGET:	NUM SEQUENCES TARGET:	BACKGROUND:	NUM SEQUENCES BACKGROUND:
India	15360	Asia	8898

TIME INTERVAL

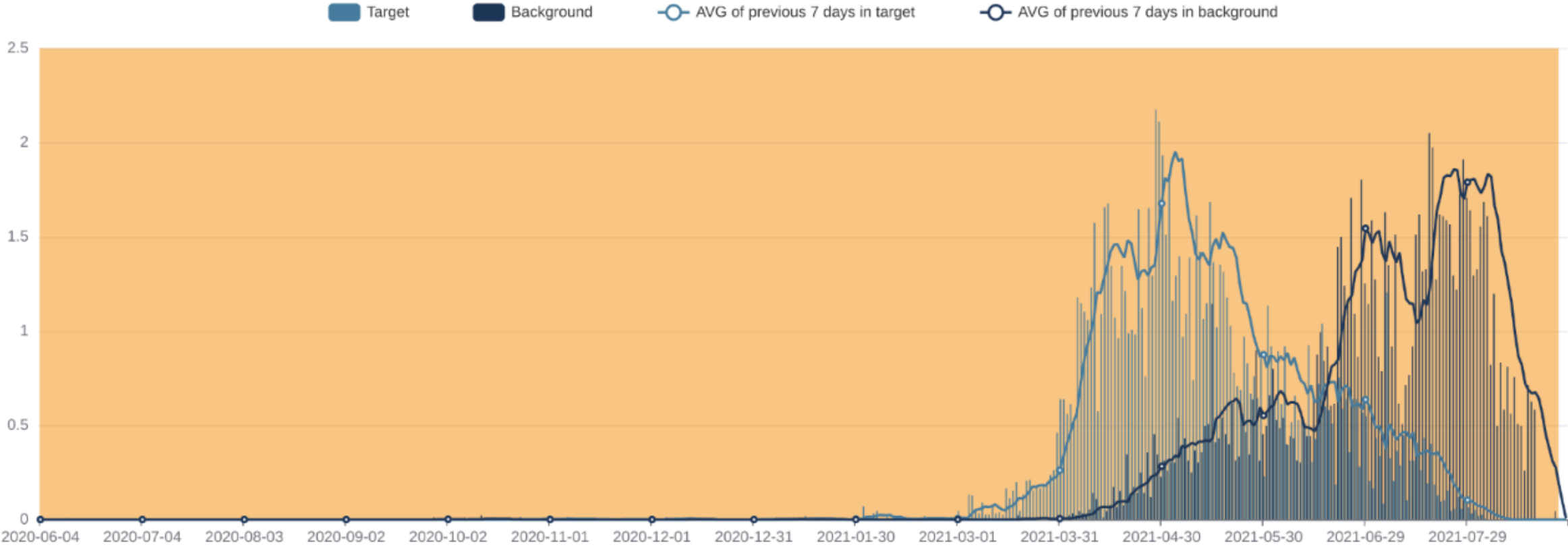
START:	END:
2020-06-04 ✓	2021-08-25 ✓
<small>(input date using the YYYY-MM-DD format)</small>	<small>(input date using the YYYY-MM-DD format)</small>

SELECT THE PROTEINS TO COMPARE
(selecting none is equivalent to "all proteins")

Protein

APPLY

India (target) VS Asia (background) timeframe



How to select, mutations of interest (advanced filters)

We focus on mutations with a high prevalence (observed in at least 10 genomes and/or 1% of the total)

ADVANCED FILTERS

% BACKGROUND: ⓘ MIN: 0 MAX: 100	# SEQUENCES IN BACKGROUND: ⓘ MIN: 0 MAX: 9270	% TARGET: ⓘ MIN: 5 MAX: 100	# SEQUENCES IN TARGET: ⓘ MIN: 10 MAX: 15693
P-VALUE: ⓘ MIN: 0 MAX: 1	ODDS RATIO: ⓘ MIN: 0 MAX: 73 INF: <input checked="" type="checkbox"/>	FILTER PROTEIN: ⓘ Protein ▼	


APPLY

Results

TABLE 

mutation	p_value ↓ 3	odds_ratio	%_target ↓ 2	%_background ↓ 1
Spike_D614G	0.13824	0.99970	99.87255 % (15673)	99.90291 % (9261)
Spike_P681R	0.00000	0.98656	98.40056 % (15442)	99.74110 % (9246)
Spike_T478K	0.00000	0.95328	94.52622 % (14834)	99.15858 % (9192)
Spike_L452R	0.00000	0.95888	95.02963 % (14913)	99.10464 % (9187)
Spike_T19R	0.00000	0.94584	93.56401 % (14683)	98.92125 % (9170)
Spike_D950N	0.00000	0.77111	73.24285 % (11494)	94.98382 % (8805)
Spike_E156G	0.00000	0.33527	30.88638 % (4847)	92.12513 % (8540)
Spike_F157-	0.00000	0.33648	30.99471 % (4864)	92.11435 % (8539)
Spike_R158-	0.00000	0.33797	31.10304 % (4881)	92.02805 % (8531)
Spike_G142D	0.00000	0.33431	25.67387 % (4029)	76.79612 % (7119)

Increase of frequency outside India

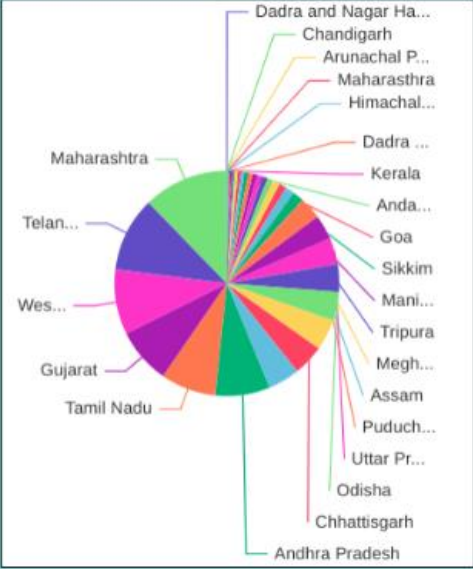
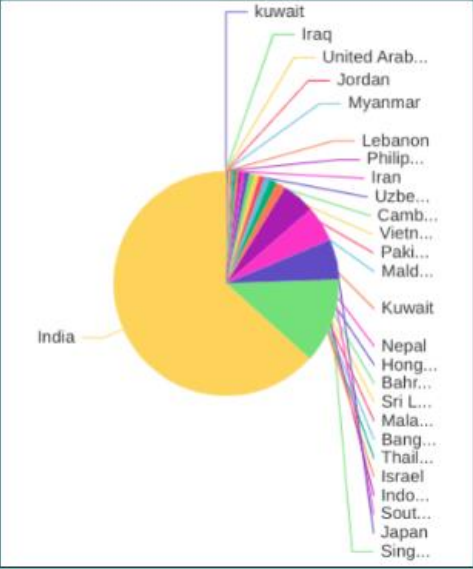
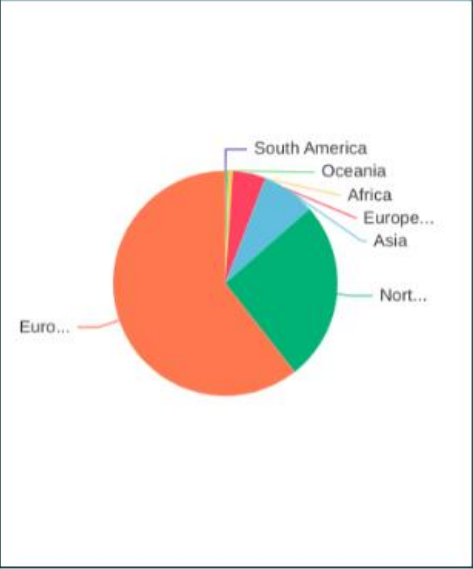


WHY?

India VS world ... repeat ...

Home | LINEAGE ANALYSIS | TEMPORAL ANALYSIS | **SPATIAL ANALYSIS** | OPEN ANALYSIS

B.1.617.2 (Delta) X



Please select a region to pick a province
(multiple values may be selected here)

Asia X | India X | region | province

Comparison
India vs World

Exclude one or more places from the background
continent

Results

3 alleles (5 mutations) are observed in the spike protein of the Delta variant in India but have a much higher frequency outside of India.

Mut	% India	% Asia (no India)	% World (no Asia)	Allele
Spike_D950N	72.66%	94.85%	97.55%	A1
Spike_E156G	29.43%	91.91%	94.17%	A2
Spike_F157-	29.54%	91.89%	93.55%	A2
Spike_R158-	29.65%	91.81%	94.20%	A2
Spike_G142D	24.41%	76.90%	89.33%	A3

Several possible alternative explanations:

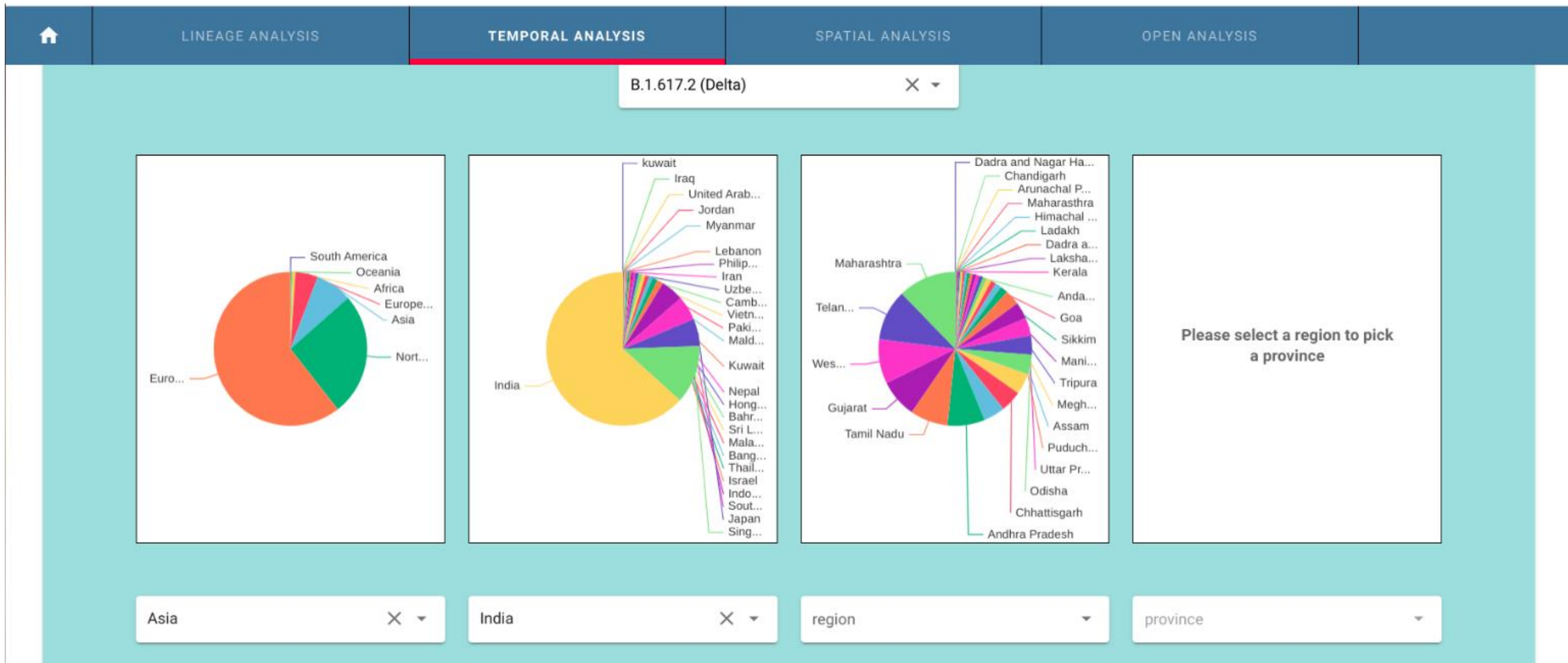
- 1) founder effect
- 2) selection outside of India
- 3) selection in time
- 4) selection in time and outside of India

Can ViruClust help?

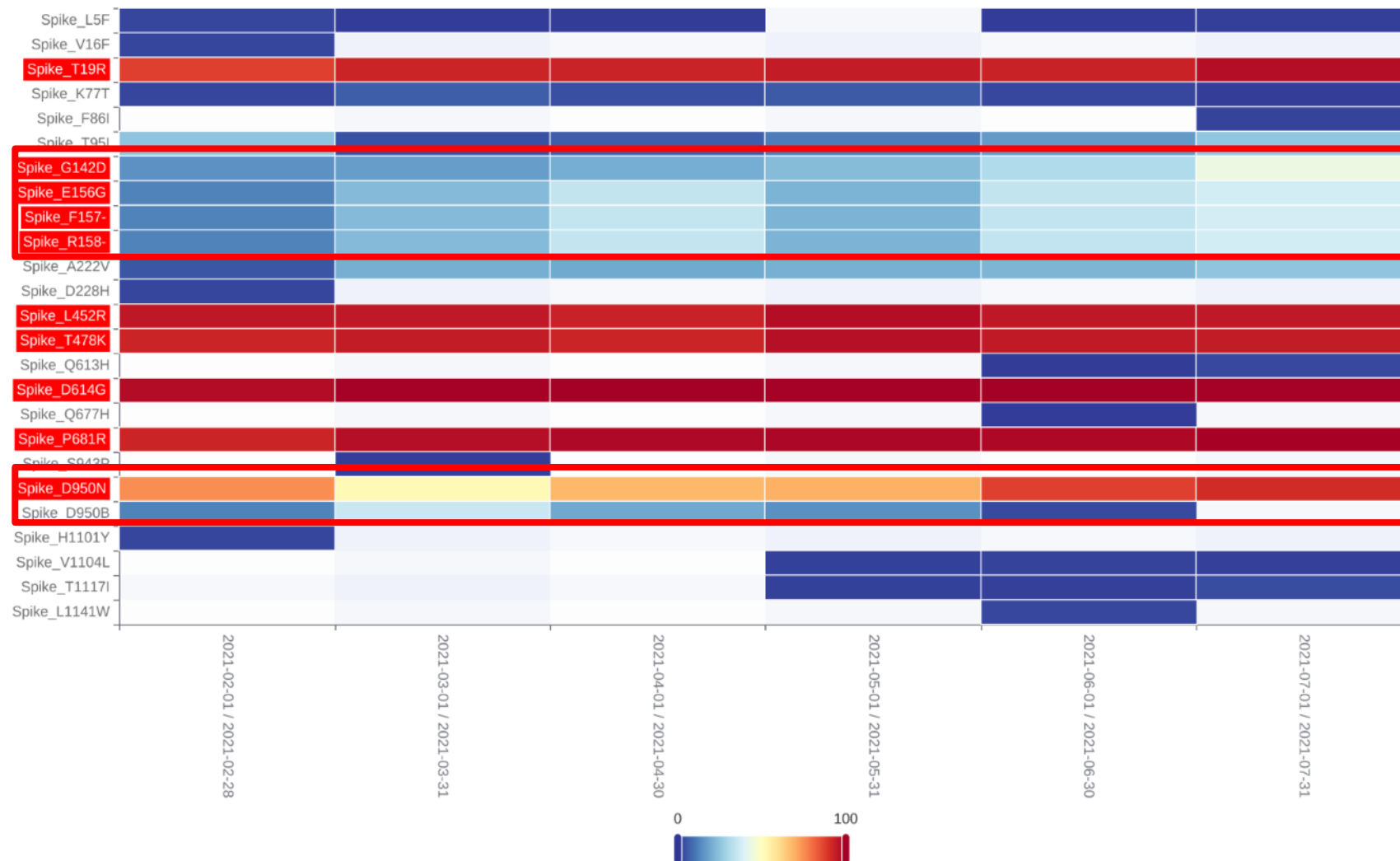
Temporal analysis

What is the pattern in time? are the alleles fixed over time? in:

- India
- Asia not India
- Rest of the world



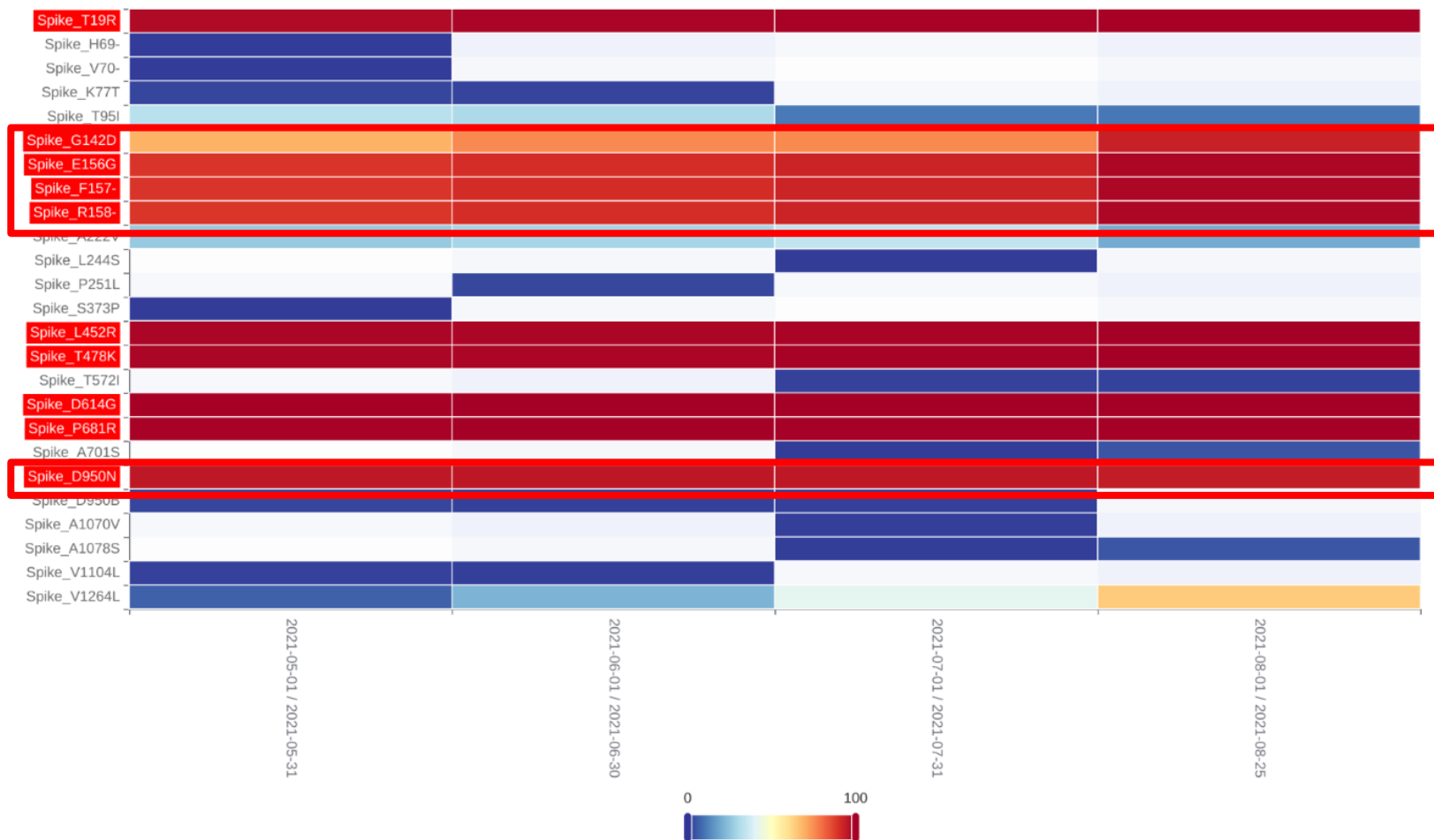
Temporal analysis (X3: **India**, Asia, World)



India:

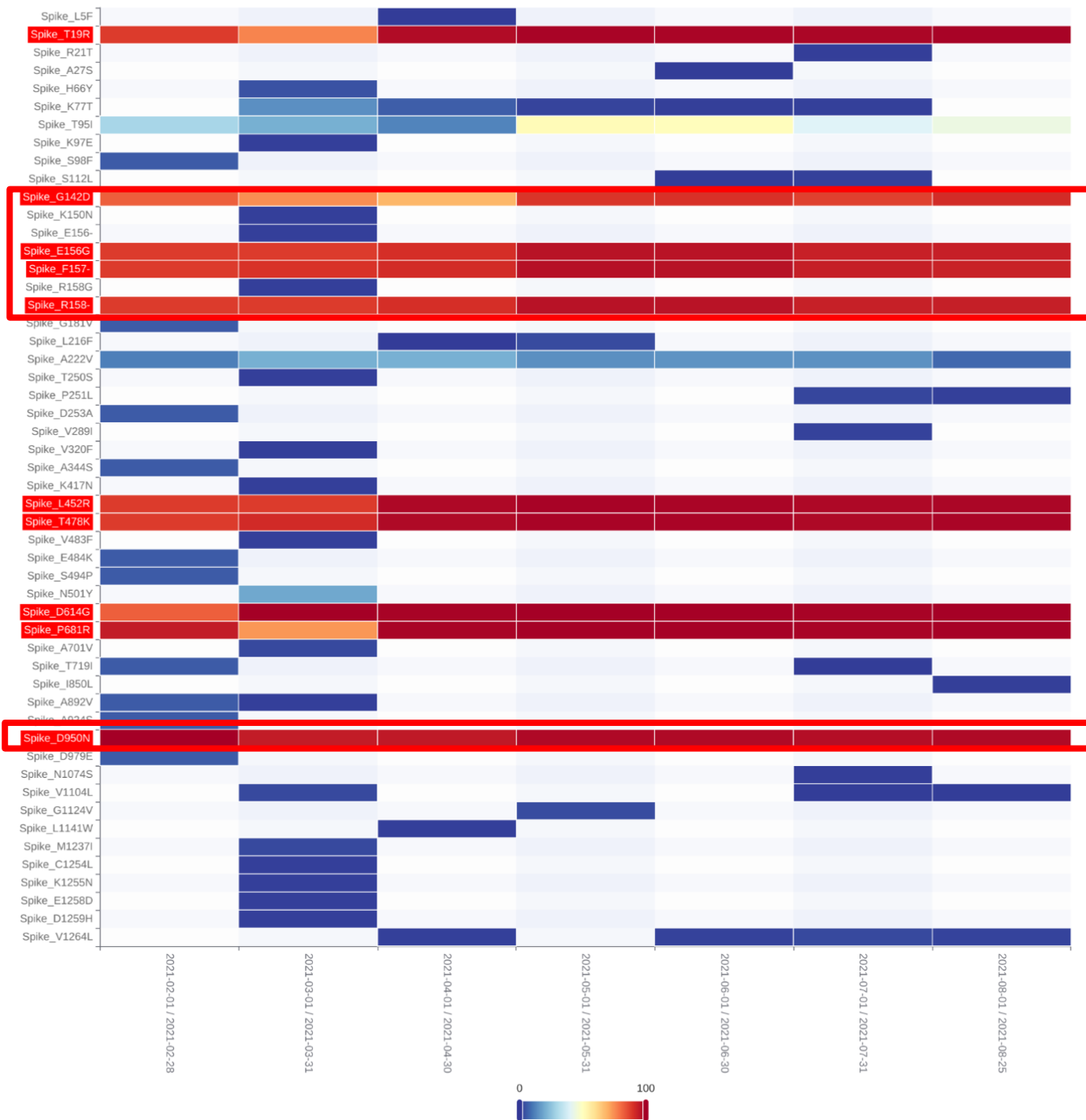
- the 3 alleles are never fixed at >90% prevalence
- all have a different profile of frequency
- all increase in frequency over time

Temporal analysis (X3: India, **Asia**, World)



Asia:

- the 3 alleles are fixed at >90% prevalence
- only one G142D show a detectable change (increase) in prevalence



World:

- the 3 alleles are fixed at >90% prevalence
- we see additional mutations that are not seen in Asia (Delta expanded considerably outside of Asia)

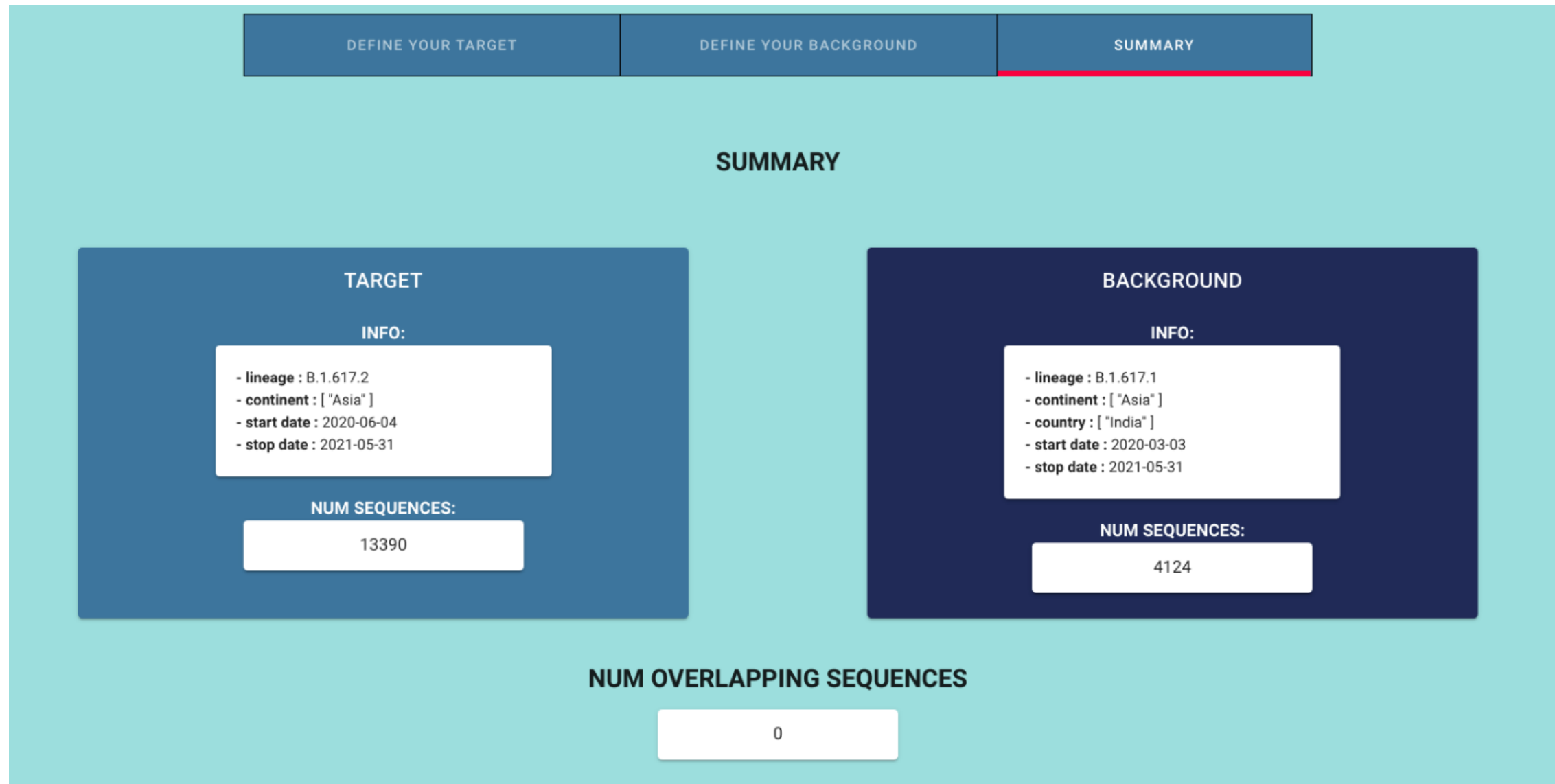
But then...

If the “novel” mutations are “selected” and confer some advantages to Delta, they should not be observed in other, closely related, variants that did not become a VOC

Delta has a very close (VOI) relative, the Kappa variant. Kappa emerged (more or less) at same point in time and space than Delta. But never reached the same level of prevalence. **WHY? What are the differences?**

In ViruClust we can use the “open” analysis to compare virtually anything! thus we can compare Delta and Kappa ...

Open analysis: compare lineages



How do the Spike of Delta (VOC) and Kappa (VOI) differ?

CHANGES IN TARGET vs BACKGROUND LOCATION

TABLE 

mutation	p_value	odds_ratio	%_target ↓ 1	%_background
Spike_D614G	0.00000	1.00430	99.82076 % (13366)	99.39379 % (4099)
Spike_P681R	0.09024	1.00156	98.43167 % (13180)	98.27837 % (4053)
Spike_L452R	0.00000	1.01352	95.42942 % (12778)	94.15616 % (3883)
Spike_T478K	0.00000	489.59023	94.97386 % (12717)	0.19399 % (8)
Spike_T19R	0.00000	552.44789	93.77147 % (12556)	0.16974 % (7)
Spike_D950N	0.00000	366.43232	71.08290 % (9518)	0.19399 % (8)
Spike_R158-	0.00000	242.69692	35.30993 % (4728)	0.14549 % (6)
Spike_F157-	0.00000	241.77295	35.17550 % (4710)	0.14549 % (6)
Spike_E156G	0.00000	241.36229	35.11576 % (4702)	0.14549 % (6)
Spike_G142D	0.00000	0.61431	27.46826 % (3678)	44.71387 % (1844)
Spike_A222V	0.00000	84.85153	20.57506 % (2755)	0.24248 % (10)

Moreover: we found evidence that two of these Spike changes S_158- and S_G142D are implicated in immune escape (McCallum et Al., <https://doi.org/10.1016/j.cell.2021.03.028>)

Some conclusions on ViruClust

- Allows the execution of (otherwise) complex comparison of SARS-CoV-2 genomes
- Comparisons can be performed at different levels (geographic, temporal, and lineage)
- Integration of different types of analyses allows the generation of “testable hypotheses” that can be used to pinpoint interesting evolutionary patterns

For example in the case of Delta...

We identify a set of mutations in the Spike glycoprotein:

1. Not universally present at the “beginning” but fixed outside the place of origin
2. Which constantly increase in frequency (inside and outside of India)
3. Reach complete fixation almost everywhere
4. Are potentially associated with immune escape
5. Are not observed in the closely related Kappa VOI

Taken all-together these analyses highlight a potentially important event

Some more general conclusions on “the COVID-19 Virus”

“Variants” are like football teams

- The more amino acid changes, the more players (Delta: 29 players)
- Each player brings together its effects, the stronger variants have the best players (Ronaldo=S:N501Y)
- Some players team together effectively (Ronaldo & Dybala Pogba = S:N501Y+S:del69/70).

“Variant effects” depend on genomic positions

- Most dangerous ones are on Spike, close to RBD or NTD
- Changes falling upon Spike “epitopes” affect their stability hence vaccine efficacy (vaccines use spike epitopes known at the time of their design)

Data is powerful for explaining SARS-CoV-2 evolution

- Should be deposited continuously and quickly on public sources as result of government regulations
- Thanks to data, it could be possible to quickly determine dangerous amino acid changes within local clusters, and possibly restrict their diffusion (active surveillance)

Today’s virus is no longer the same

- E.g. in Italy/EU we went from Wuhan to Alpha to Delta
- Other steps are very likely, perhaps virus evolution can even be predicted.

Some relevant publications, SARS-CoV-2

Bernasconi A., Canakoglu A., Masseroli M., Pinoli P., Ceri S. **A review on viral data sources and search systems for perspective mitigation of COVID-19** Briefings in Bioinformatics (IF 8.990), 2021.

Bernasconi A., Canakoglu A., Pinoli P., Ceri S. **Empowering Virus Sequence Research through Conceptual Modeling**, International Conference on Conceptual Modeling (ER), 2021.

Canakoglu A., Pinoli P., Bernasconi A., Alfonsi T., Melidis D.P., Ceri S. **VirusSurf: an integrated database to investigate viral sequences** Nucleic Acids Research, Database Issue (IF 16.971), 2021.

Bernasconi A., Gulino A., Alfonsi T., Canakoglu A., Pinoli P., Sandionigi A., Ceri S. **VirusViz: Comparative analysis and effective visualization of viral nucleotide and amino acid variants** Nucleic Acids Research (IF 16.971), 2021.

Bernasconi A., Cilibrasi L., Al Khalaf R., Alfonsi T., Ceri S., Pinoli P., A., Canakoglu A. **EpiSurf: metadata-driven search server for analyzing amino acid changes on epitopes of SARS-CoV-2 and other viral species**, to appear on Database, Oxford Journals (IF 3.4).

Al Khalaf R., Alfonsi T., Ceri S., Bernasconi A. **CoV2K: a Knowledge Base of SARS-CoV-2 Variant Impacts**. The 15th International Conference on Research Challenges in Information Science (RCIS), 2021.

Bernasconi A., Al Khalaf R., Alfonsi T., Ceri S. **Knowledge model about SARS-CoV-2 sequences and their mutations/variants**, submitted as commentary to Scientific Data (Nature Group, IF 5.9).

Cilibrasi L., Pinoli, P., Bernasconi, A., Canakoglu A., Chiara, M., Ceri S. **VirusClust: direct comparison of SARS-CoV-2 genomes and genetic variants in space and time**, submitted to Bioinformatics (IF 5.6).

Pinoli, P., Bernasconi, A., Sandionigi, A., Ceri S. **VirusLab, a tool for customized SARS-CoV-2 data analysis**, submitted to BioTech (MDPI), special issue "Bioinformatics and High-Performance Computing Methods for Deciphering and Fighting COVID-19".

TRY OUR SEARCH SYSTEMS UPON VIRAL DATABASES

VirusSurf

Includes:

- Web interface for browsing an integrated database to investigate viral sequences.
- Processed data from [GenBank](#) and [COG-UK](#)

EpiSurf

Includes:

- Web interface for browsing an integrated database to investigate epitopes and viral sequences.
- Processed data from [IEDB](#), [GenBank](#) and [COG-UK](#)

GitHub

Includes:

- [GeCo group GitHub repository](#) (all the open source files of the GeCo project)
- [Front-end project](#)
- [Back-end project](#)
- [EpiSurf dedicated project](#)

VirusSurf GISAID

Includes:

- Web interface for browsing an integrated database to investigate viral sequences from GISAID
- Processed data from [GISAID](#)


EpiSurf GISAID

Includes:

- Web interface for browsing an integrated database to investigate epitopes and viral sequences from GISAID
- Processed data from [IEDB](#) and [GISAID](#)

Documentation

Includes:

- [VirusSurf wiki documentation](#)
- [VirusSurf video tutorials](#) 
- [EpiSurf wiki documentation](#)
- [VirusViz wiki documentation](#)

VirusViz

Includes:

- Web tool for analyzing viral sequences and visualizing their variants and characteristics.
- Visualize data from [VirusSurf](#) and [EpiSurf](#)

VirusClust GISAID

Includes:

- Web tool for comparison of SARS-CoV-2 genomes and genetic variants in space and time.
- Processed data from [GISAID](#)

GeCo Team goes Viral, since March 2020

**ANNA
BERNASCONI**



Postdoctoral Researchers

**ARIF
CANAKOGLU**



Researcher fellow

**PIETRO
PINOLI**



STEFANO CERI
Professor & PI

Collaborators

ITALY

Matteo Chiara [Università di Milano]
Simone Furini [Università di Siena]
Francesca Mari [Università di Siena]
Nicola Picchiotti [Università di Siena]
Alessandra Renieri [Università di Siena]
Anna Sandionigi [Quantia Consulting S.R.L.]

Renato Casagrandi [DEIB]
Lorenzo Mari [DEIB]
Politecnico di Milano

EUROPE

Giancarlo Guizzardi [University of Twente, NL]
Damianos P. Melidis [L3S Hannover University, DE]
Oscar Pastor [Universitat Politecnica de Valencia, ES]

WORLD

Ilaria Capua [University of Florida, US]
Brittany Rife Magalis [University of Florida, US]
Marco Salemi [University of Florida, US]
Veda Storey [Georgia State University, US]

Master Graduates and Students

ANTONIO ESPOSITO
ELISABETTA FEDELE
SILVIA FRANZINI
FRANCESCO INVERNICI

**TOMMASO
ALFONSI**



**RUBA
AL KHALAF**



**LUCA
CILIBRASI**



PhD Students and assegnisti

More work on COVID-19 (the beauty of data science)

Mapping the human genetic architecture of COVID-19, Covid-19 Host Genetics Initiative, (with A. Bernasconi, A. Canakoglu, P. Pinoli and over one thousand contributors), Nature, July 8, 2021.

Employing a systematic approach to biobanking and analyzing clinical and genetic data for advancing COVID-19 research, S. Daga et Al. (with P. Pinoli) European Journal of Human Genetics, 2021.

Post-Mendelian genetic model in COVID-19, N. Picchiotti et Al., (with P. Pinoli), submitted to BMC-Bioinformatics.

Data-driven analysis of amino acid change dynamics timely reveals SARS-CoV-2 variant emergence, A. Bernasconi et Al., minor rev., Scientific Reports.

VaccinItaly: monitoring Italian conversations around vaccines on Twitter and Facebook, F. Pierri et Al., Proceedings of the International AAAI Conference on Web and Social Media (ICWSM'2021).

Socioeconomic differences and persistent segregation of Italian territories during COVID-19 pandemic, G. Bonaccorsi et Al. (with F. Pierri), minor rev., Scientific Reports.

Reasoning on company takeovers during the COVID-19 crisis with knowledge graphs Bellomarini et Al. RuleML+ RR Conference, 2011 (with D. Magnanimiti)

Reasoning on Company Takeovers: From Tactic to Strategy, Luigi Bellomarini et Al. (with D. Magnanimiti), submitted to Data and Knowledge Engineering.

A few words about GeCo

Completed on August 30 (5years, 2.5 MEuro)

Main results

- Language paradigm: GMQL
 - High level, optimized, somehow cited and used
 - Available at DEIB (Polimi), CINECA and on TERRA(Broad Int.) and GALAXY
 - Accessible through Python and R
- Human Genomics Repository
 - Conceptually well defined
 - Connected to major repositories: Encode, TCGA-CDC, Epigenomic Roadmap, Cistrome, GenCode, Geo, GWAS, ...
 - Over 300K files with uniform format
 - Searchable through GenoSurf
- Several biological and clinical problems considered
 - Three-dimensional structure of the genome
 - Drug repurposing
 - Precision medicine for cancer

Over 100 papers, 12PhD graduated, 6 in process, a new LM on “Computational biology for genomics” (joint Polimi-University of Milan)

TRY GMQL

WEB Interface

Includes:

- Web interface for browsing datasets and building GMQL queries
- Processed data from ENCODE, Roadmap Epigenomic and TCGA public sources

Downloads

Includes:

- Local mode or MapReduce mode (over Hadoop, or Hadoop YARN) for GNU/Linux systems
- Quick start - Install GMQL and get started

RGMQL

Includes:

- GMQL operations in R
- Bioconductor interface




REST APIs

Includes:

- REST APIs for programmatic access to GMQL repository and query execution engines

Documentation

Includes:

- Introduction to the language [pdf](#) 
- Example queries [pdf](#) 
- Biological examples [pdf](#) 

FireCloud Workspace @ Broad Institute

Includes:

- Integration of GMQL in WDL workflows
- Examples of usage

GitHub Site

Includes:

- GeCo group GitHub repository (contains all the open source files of the GeCo project)

PyGMQL

Includes:

- GMQL operations in Python
- Pandas interface

Federated GMQL

Includes:

- Introduction to the language [pdf](#) 
- Example queries [pdf](#) 

TRY GENOSURF

GenoSurf is an integrated metadata-based semantic search server for genomic datasets. Enjoy the genomic data interoperability!

We gathered metadata from multiple heterogeneous sources (e.g., ENCODE, TCGA, Roadmap Epigenomics) and integrated them in a unified view using the [Genomic Conceptual Model \(GCM\)](#), i.e., a conceptual schema for metadata, that summarizes the most important metadata information shared between different genomic data sources. The interface allows to specify filters, i.e., specific values for attributes such as “technique” or “data type” and retrieve a list of available genomic items that match the search. Items are typically files of genomic regions containing coordinates and their properties.

For all details, please refer to our recently published paper on Database journal, [here](#).

WEB Interface

Includes:

- Web interface for browsing genomic datasets and samples
- Interactive use of drop-down menus
- Integrated and original metadata search

Downloads

Includes:

- GenoSurf Database Dumps


REST APIs

Includes:

- REST APIs for programmatic access to GenoSurf database

Documentation

Includes:

- [Wiki documentation](#)
- [Video tutorials](#) 
- [Questionnaire](#)

GitHub

Includes:

- [GeCo group GitHub repository](#)(contains all the open source files of the GeCo project)
- [Front-end project](#)
- [Back-end project](#)

Applications

Includes:

- Simple example Jupyter Notebooks to exploit the GenoSurf APIs

Some Recent GeCo Work (journal papers only, 2020-21)

GECO TECHNOLOGY

The road towards data integration in human genomics: players, steps and interactions, A Bernasconi et Al., Briefings in Bioinformatics (with A Canakoglu and M Masseroli) (IF 8.9)

Federated sharing and processing of genomic datasets for tertiary data analysis, A Canakoglu, P Pinoli et Al., Briefings in Bioinformatics (with L Nanni, M Masseroli) (IF 8.9)

OpenGDC: Unifying, Modeling, Integrating Cancer Genomic Data and Clinical Metadata E Cappelli et Al, Applied Sciences (with A Bernasconi, A Canakoglu, M Masseroli) (IF 2.6)

DRUG PREDICTION THROUGH MATRIX FACTORIZATION

Matrix factorization-based technique for drug repurposing, IEEE journal of biomedical and health informatics, 2020, G. Ceddia et Al. (with P. Pinoli and M. Masseroli) (IF 5.2)

Predicting Drug Synergism by Means of Non-Negative Matrix Tri-Factorization, P. Pinoli et Al., IEEE/ACM Transactions on Computational Biology and Bioinformatics (with M. Masseroli) (IF 2.4)

Some Recent GeCo Work (journal papers only, 2020-21)

3D STRUCTURE

Systematic inference and comparison of multi-scale chromatin sub-compartments connects spatial organization to cell phenotypes, Y. Liu et Al., Nature Communications (with L. Nanni) (IF 12.1)

Spatial patterns of CTCF sites define the anatomy of TADs and their boundaries, L. Nanni et Al., Genome Biology (IF 10.8)

Exploring chromatin conformation and gene co-expression through graph embedding, L. Nanni et Al., Bioinformatics (IF 7.4)

CANCER

Pan-cancer analysis of somatic mutations and epigenetic alterations in insulated neighbourhood boundaries, P Pinoli et Al., Plos-One (IF 3.2)

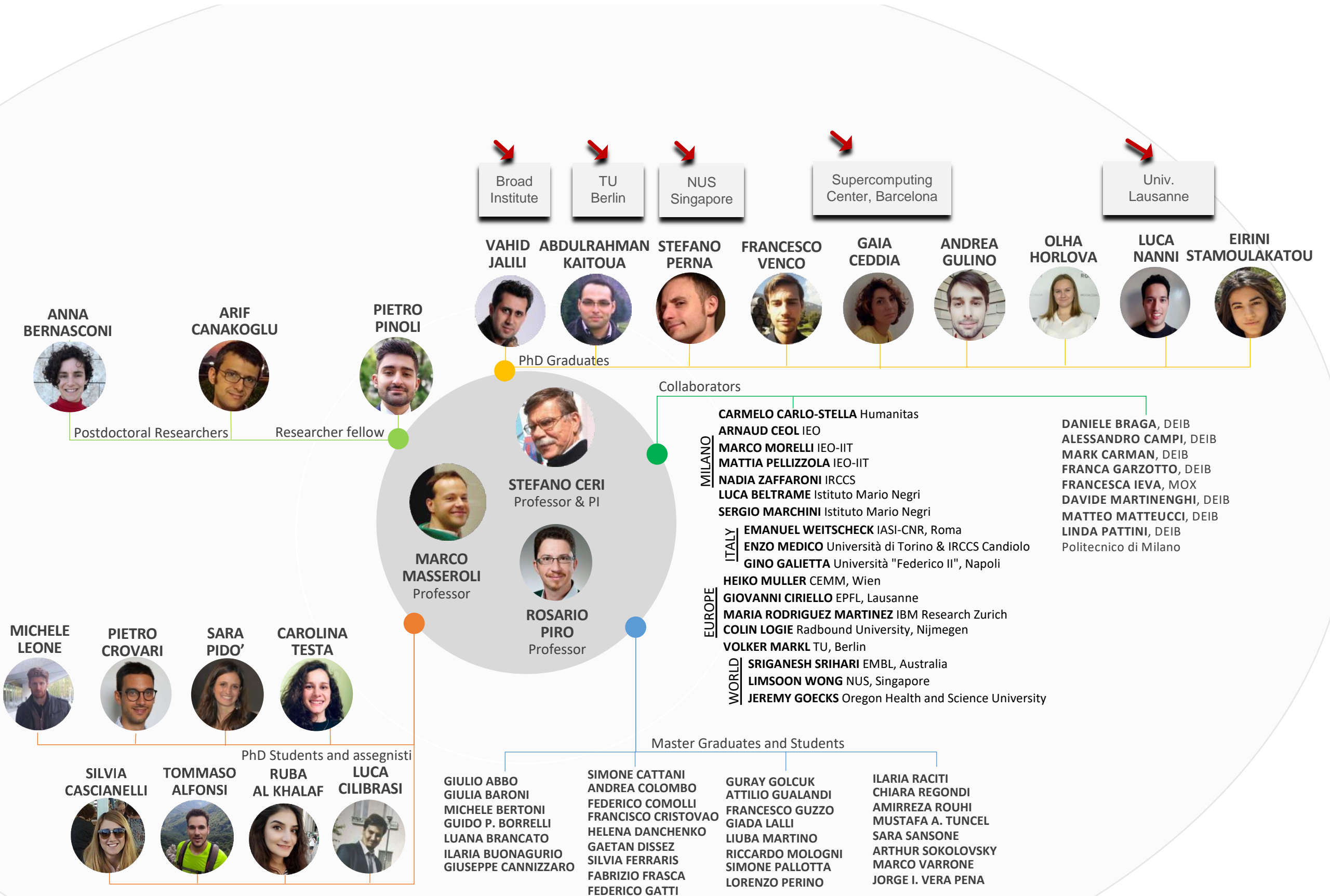
Identifying collateral and synthetic lethal vulnerabilities within the DNA-damage response, P. Pinoli et Al., BMC bioinformatics (IF 3.2)

EPIGENETICS

NAUTICA: classifying transcription factor interactions by positional and protein-protein interaction information S Perna et Al., Biology Direct (with P. Pinoli) (IF 4.5)

Search and comparison of (epi) genomic feature patterns in multiple genome browser tracks A Ceol et Al., BMC Bioinformatics (with M. Masseroli) (IF 3.2)

GeCo Team, Sept 2021



The second GeCo spinoff: GeCoAgent

Multi-modal tool for facilitating the use of genomic and data science.

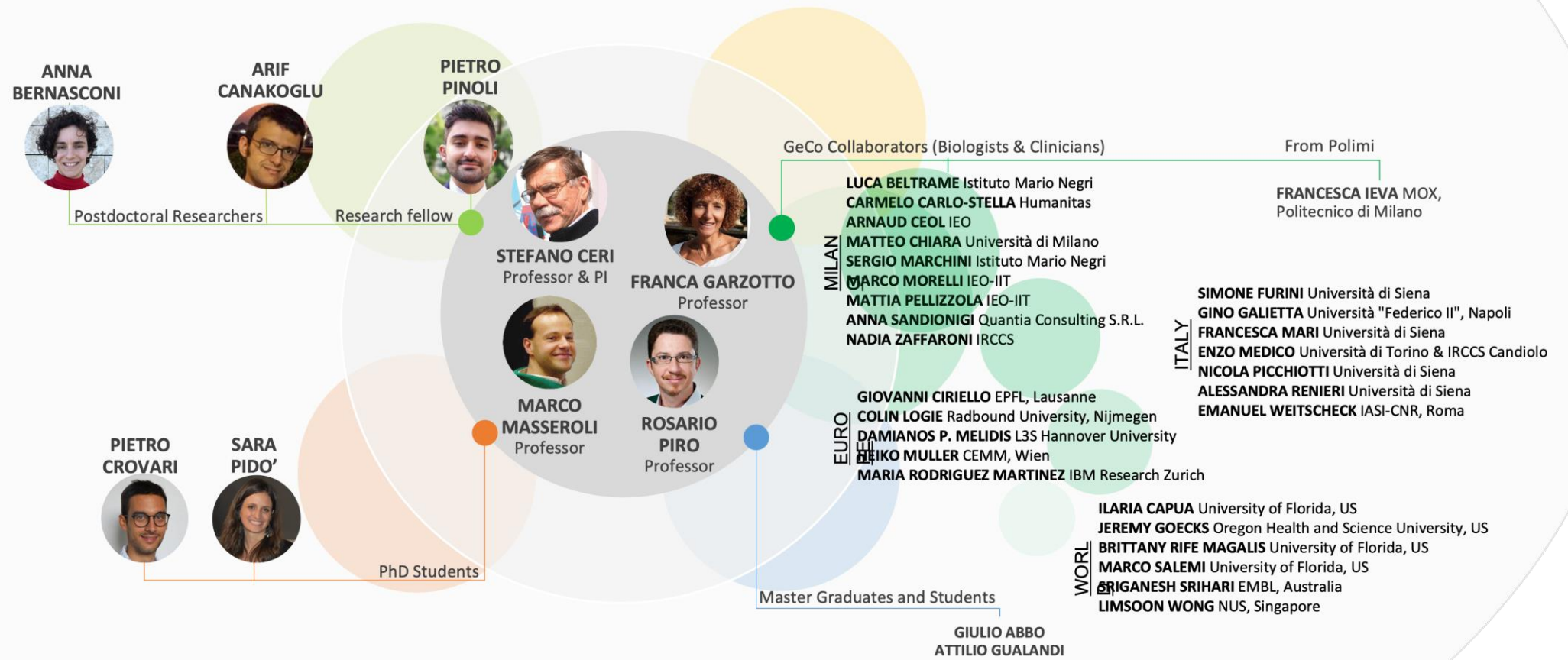
- Addresses data extraction (genomic –specific) and data analysis (general data science)
- Uses a conversational agent in the form of a chat box and a dashboard with several objects, progressively built and shown to the user as the interaction goes on.
- Directly addresses genomics experts
- Directly produces streamlined, cleaner and error-free solutions from requirements.

GeCoAgent: a Conversational Agent for Empowering Genomic Data Extraction and Analysis. P. Crovari et Al, Accepted for publication on ACM Transactions on Computing for Healthcare (with A. Bernasconi, P. Pinoli, F. Garzotto)

.Show, Don't Tell. Reflections on the Design of Multi-modal Conversational Interfaces. P. Crovari et Al. International Workshop on Chatbot Research and Design, 2020 (with F. Garzotto)

OK, DNA! A Conversational Interface to Explore Genomic Data. P. Crovari et Al., Conference on Conversational User Interfaces, 2020 (with P. Pinoli and F. Garzotto)

GeCoAgent Team



Special Thanks



ANNA BERNASCONI



ARIF CANAKOGLU



PIETRO PINOLI