Data-driven SARS-CoV-2 understanding and hunting
(searching for the new Omicron)

Anna Bernasconi
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano

POLITECNICO
MILANO 1863

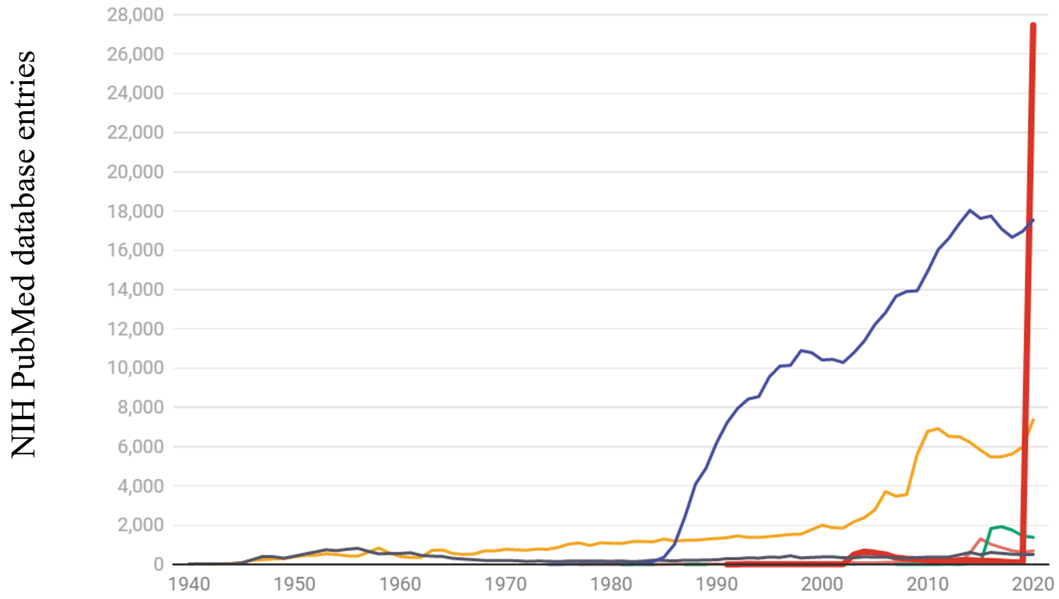
online,
14 January, 2022

Seminar outline

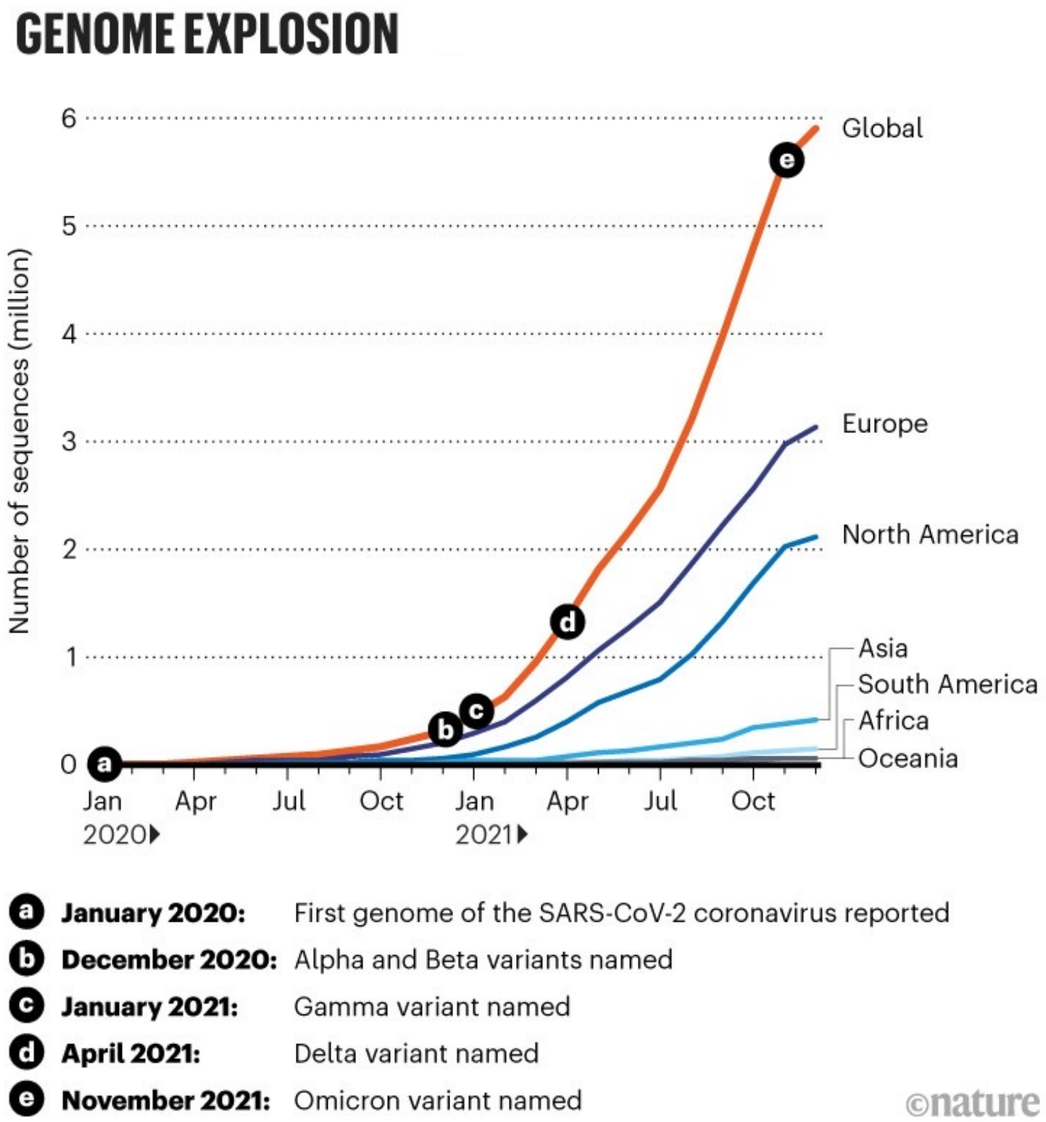
- Crash course on viral biology
- **Data sources**
- **Models**
  - Viral Conceptual Model
  - OntoVCM: an ontologically unpacked conceptual model
  - CoV2K: an abstract model for data and knowledge about SARS-CoV-2
  - Clinical phenotype data dictionary
- **Databases**
  - Data integration pipeline
  - Knowlegde integration pipeline
- **Data tools**
  - ViruSurf search engine
  - Searching mutations in epitopes with EpiSurf
  - Mutation distribution visualization with VirusViz and VirusLab
  - Variant analysis in time and space with ViruClust
- **Knowledge tools**
  - CoV2K-API
  - MutEffStage
- **Analyses**
  - Time series analysis of amino acid changes
  - Omicron mutations effects and impact on epitopes
  - Omicron and “Cameroon” analysis with the VariantHunter tool
  - Omicron as other variants recombination

The pandemics changes everything

Blue: HIV
Yellow: Influenza
Red: SARS-CoV/SARS-CoV2

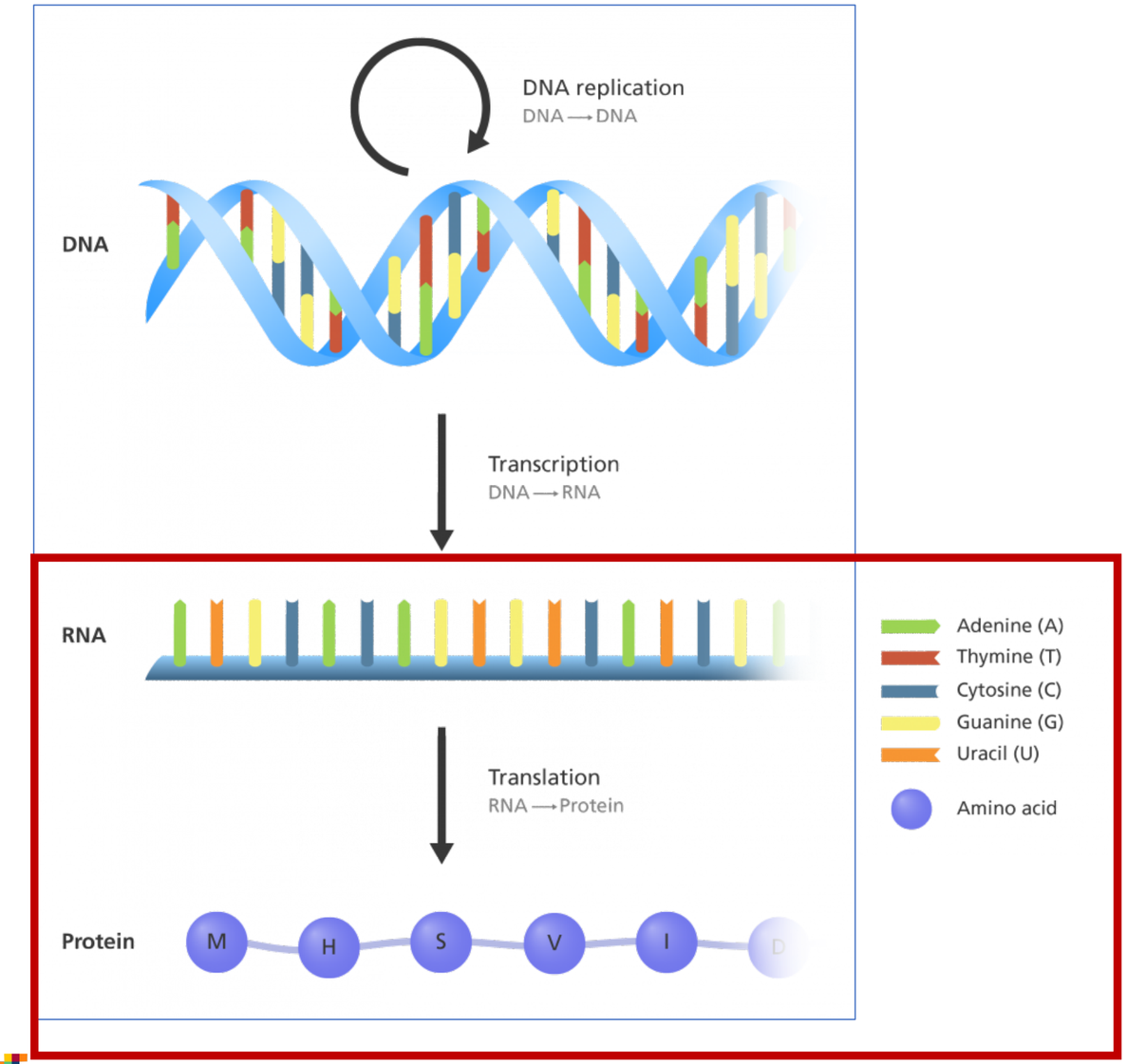


Source: https://theconversation.com/as-scientists-turn-their-attention-to-covid-19-other-research-is-not-getting-done-and-that-can-have-lasting-consequences-154040



Source: https://www.nature.com/articles/d41586-021-03698-7

RNA viruses crash course



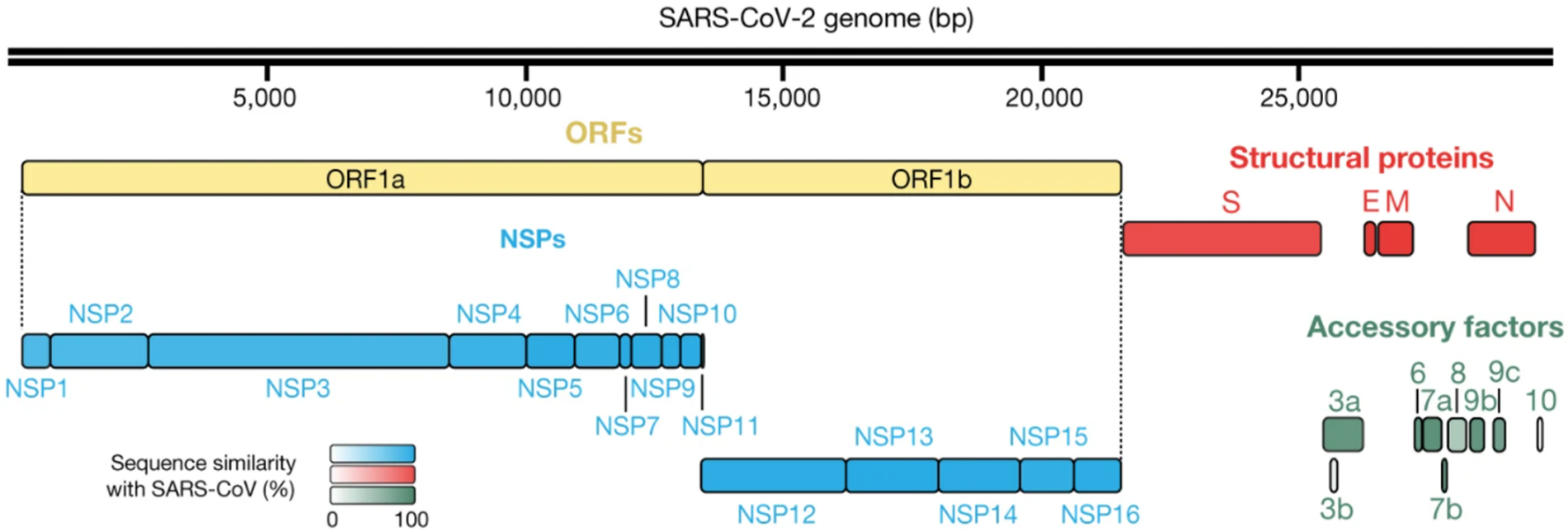
RNA viruses:

- ribonucleic acid (RNA) is the genetic material
- usually single-stranded
- include the common cold, influenza, SARS, MERS, Covid-19, Dengue, Ebola, hepatitis C, hepatitis E, polio and measles

Source: https://www.yourgenome.org/facts/what-is-the-central-dogma. Credits: Genome Research Limited

SARS-CoV-2 genome crash course

- ~30K bases (A, C, G, U) directing the synthesis of proteins
- strong sequence similarity with SARS-CoV responsible for SARS
- Protein structure of SARS-CoV-2: 4 structural proteins (Spike, E, M, N), 16 non structural proteins (NSP1-NSP16) and other accessory factor regions, included within two open reading frames



Source: Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, O'Meara MJ, Rezelj VV, Guo JZ, Swaney DL, Tummino TA. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature. 2020 Jul;583(7816):459-68. https://doi.org/10.1038/s41586-020-2286-9

How to describe mutational processes

Nucleotide Mutation:

<reference_nucleotide, coordinate_on_genome, alternative_nucleotide>

e.g. A23403G

Amino Acid Change

<protein, reference_amino_acid_residue, coordinate_on_protein, alternative_amino_acid_residue>

e.g. Spike:D614G (occurring as a result of nucleotide A23403G mutation)

Amino acid changes alter the protein function. Their effect depends on:

- The residue substitution (different residues have different physical/chemical properties)
- The position in the protein (e.g., in the Spike, the Receptor Binding Domain is the most critical)
- The co-occurrence with other changes

From requirements elicitation to action

Identification of area of interest

Understand data availability: sequences, mutations, epitopes

User base needs: Immunology, Veterinary virology, Genomic surveillance

Requirements analysis

Clinicians

Alessio Agheno [Humanitas Hospital, Milan (IT)]
Simone Furini, Francesca Mari, Nicola Picchiotti & **Alessandra Renieri** [Università di Siena (IT)]
Isabelle Migeotte [Erasme Hospital, Bruxellles (BE)]
COVID-19 Hoste Genetics Initiative

Pathologists

Carla Mavian, **Brittany Rife Magalis** & **Marco Salemi** [Department of Pathology, College of Medicine (University of Florida, US)]

Veterinary virologists

Ilaria Capua [One Health Center of Excellence (University of Florida, US)]
Alice Fusaro & Stefania Leopardi [Istituto Zooprofilattico Sperimentale delle Venezie (IT)]
Alessio Lorusso [Istituto Zooprofilattico Sperimentale Abruzzo Molise (IT)]

Molecular biologists

Ana Conesa [University of Florida (US)]
Shay Fleishon [Central Virology Laboratory, Israel Ministry of Health (IL)]
Matteo Chiara & **Federico Zambelli** [Università degli Studi di Milano Statale (IT)]
Graziano Pesole [Università di Bari (IT)]
Alba Grifoni [La Jolla Un (US)]

Epidemiologists

Renato Casagrandi & **Lorenzo Mari** [Politecnico di Milano (IT)]
Luca Ferretti [Oxford Big Data Institute (UK)]
Stephen Tsui [The Chinese University of Hong Kong (HK)]
Limsoon Wong [National University of Singapore (SGP)]

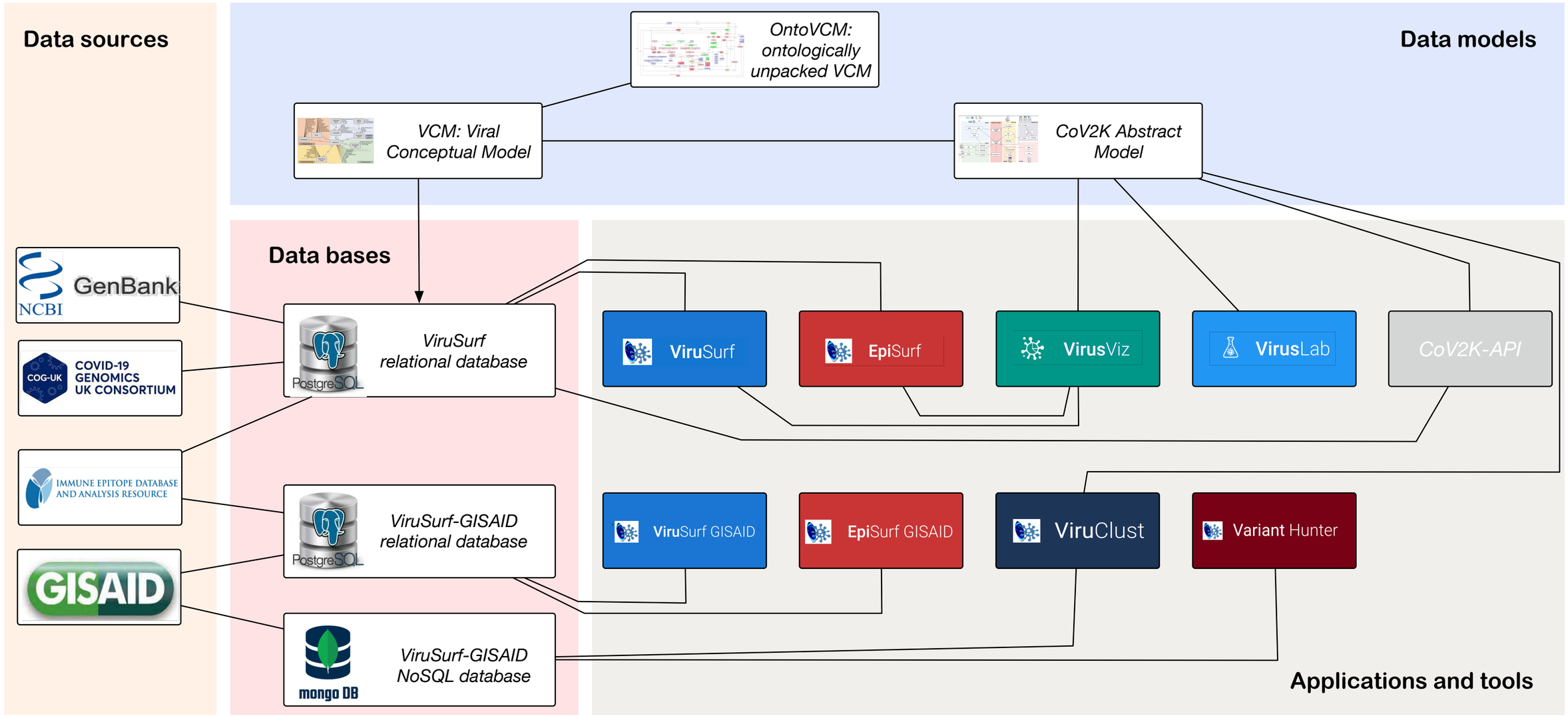
Private sector

Riccardo Colombo [BioRep, Milan (IT)]
Susanna Lamers [BioInfoExperts (Louisiana, US)]
Ilkka Lappalainen, Heikki Lehväslaiho & Francesca Morello [ELIXIR Converge]
Anna Sandionigi [Quantia Consulting S.R.L.]

Systems and studies design

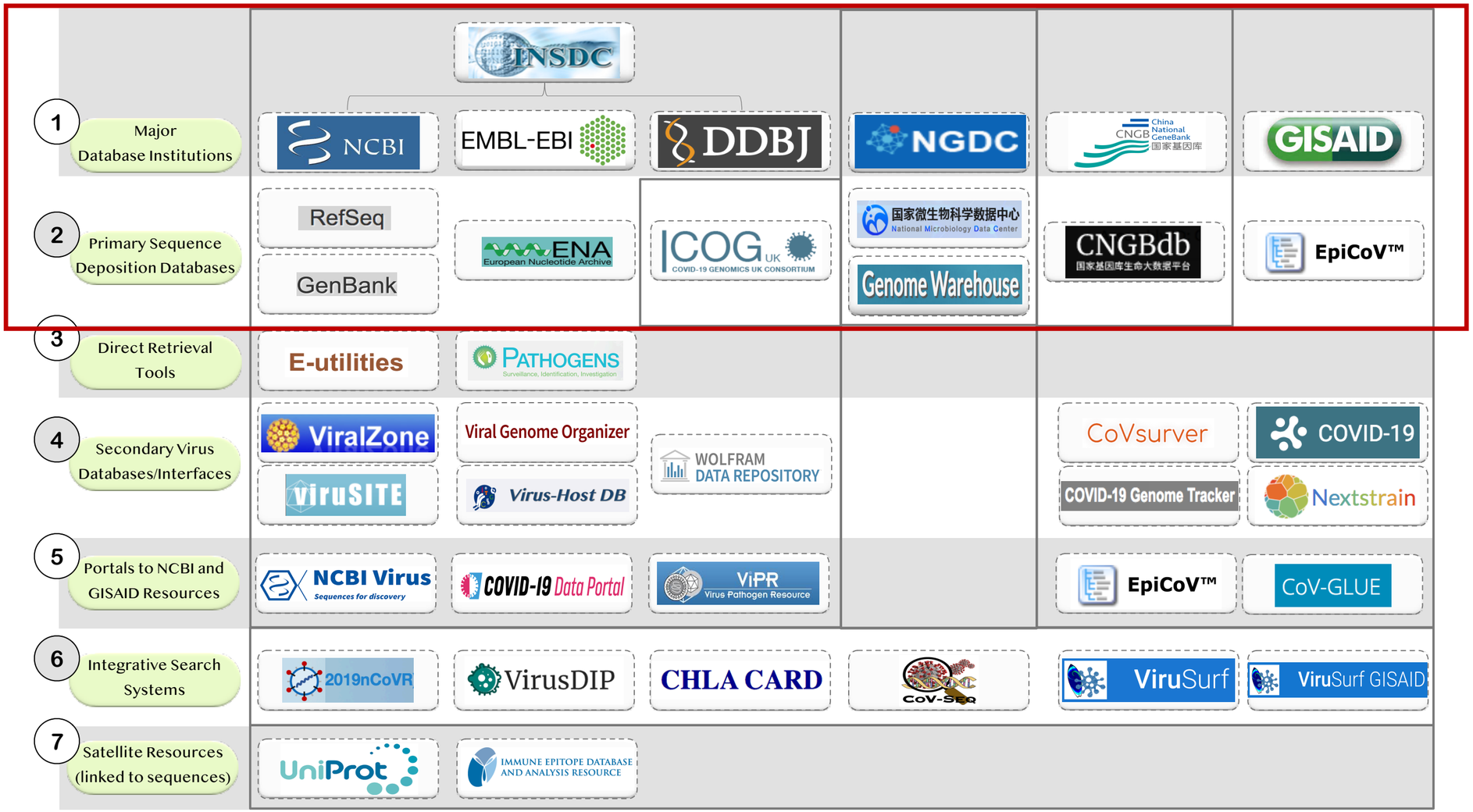
- Viral Conceptual Model
- COVID-19 Host Genetics Initiative Data Dictionary
- Genetic COVID-19 study (Siena)
- Sequence Integrated Database
- ViruSurf query engine
- VirusLab private annotation pipeline
- VirusViz
- EpiSurf
- ViruClust
- CoV2K
- ...

Bird's eye view of the models&tools suite



Data Sources

Main actors in the virus data landscape



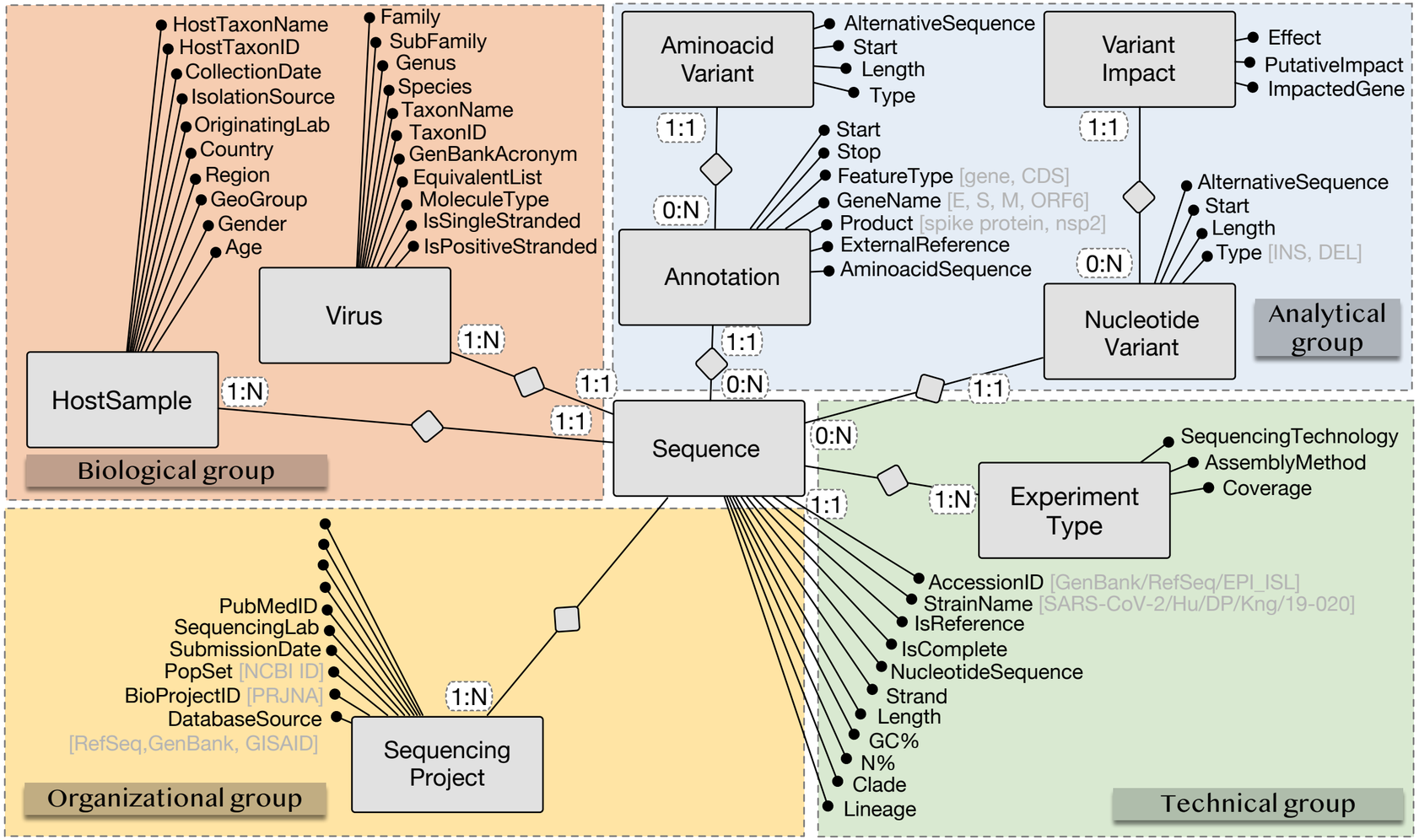
Bernasconi, A., Canakoglu, A., Masseroli, M., Pinoli, P., & Ceri, S. (2021). A review on viral data sources and search systems for perspective mitigation of COVID-19. Briefings in Bioinformatics, 22(2), 664-675. https://doi.org/10.1093/bib/bbaa359 - IF 11.6, SJR: Q1

Models

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)

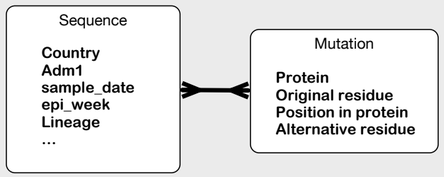


Bernasconi, A., Canakoglu, A., Pinoli, P., & Ceri, S. (2020, November). Empowering Virus Sequence Research Through Conceptual Modeling. In International Conference on Conceptual Modeling (pp. 388-402). Springer, Cham. https://doi.org/10.1007/978-3-030-62522-1_29

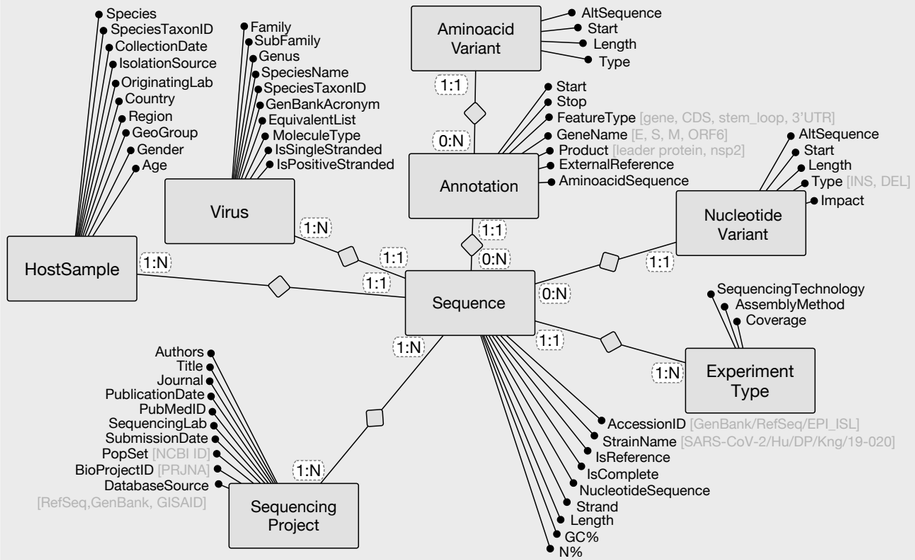
Using a foundational ontology to achieve VCM conceptual clarification

UFO foundational ontology + OntoUML modeling language

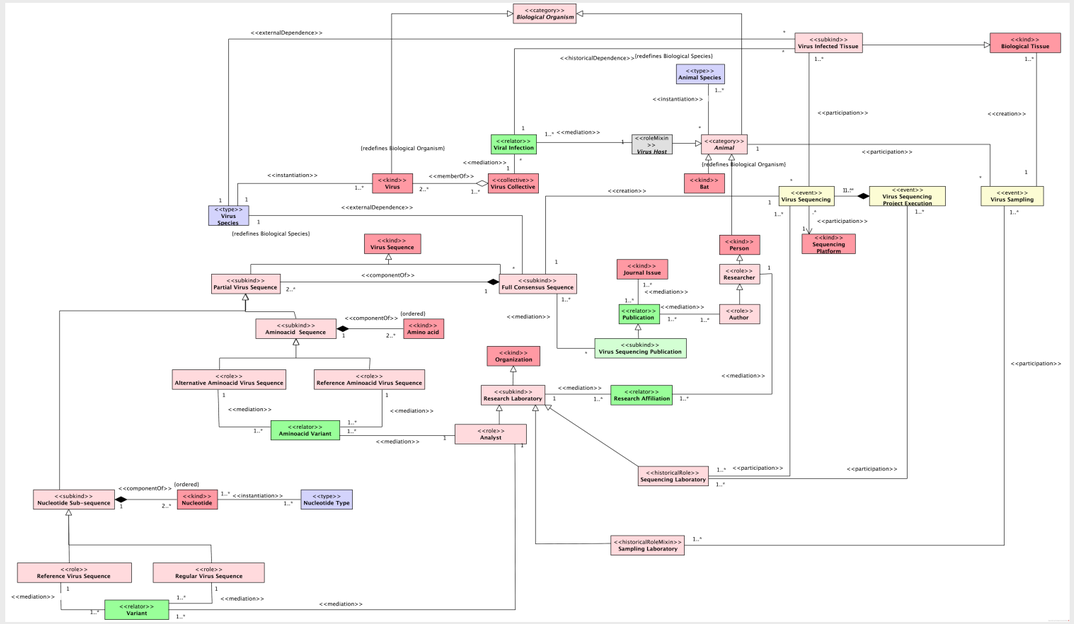
The basic data sources schema



The Viral Conceptual Model (VCM)



VCM ontologically unpacked version



Guizzardi, G., **Bernasconi, A.**, Pastor, O., & Storey, V. C. (2021, October). Ontological Unpacking as Explanation: The Case of the Viral Conceptual Model. In International Conference on Conceptual Modeling (pp. 356-366). Springer, Cham. https://doi.org/10.1007/978-3-030-89022-3_28

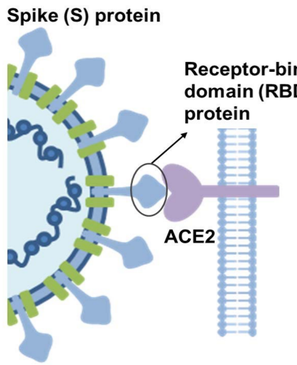
CoV2K: understanding impact of virus mutations

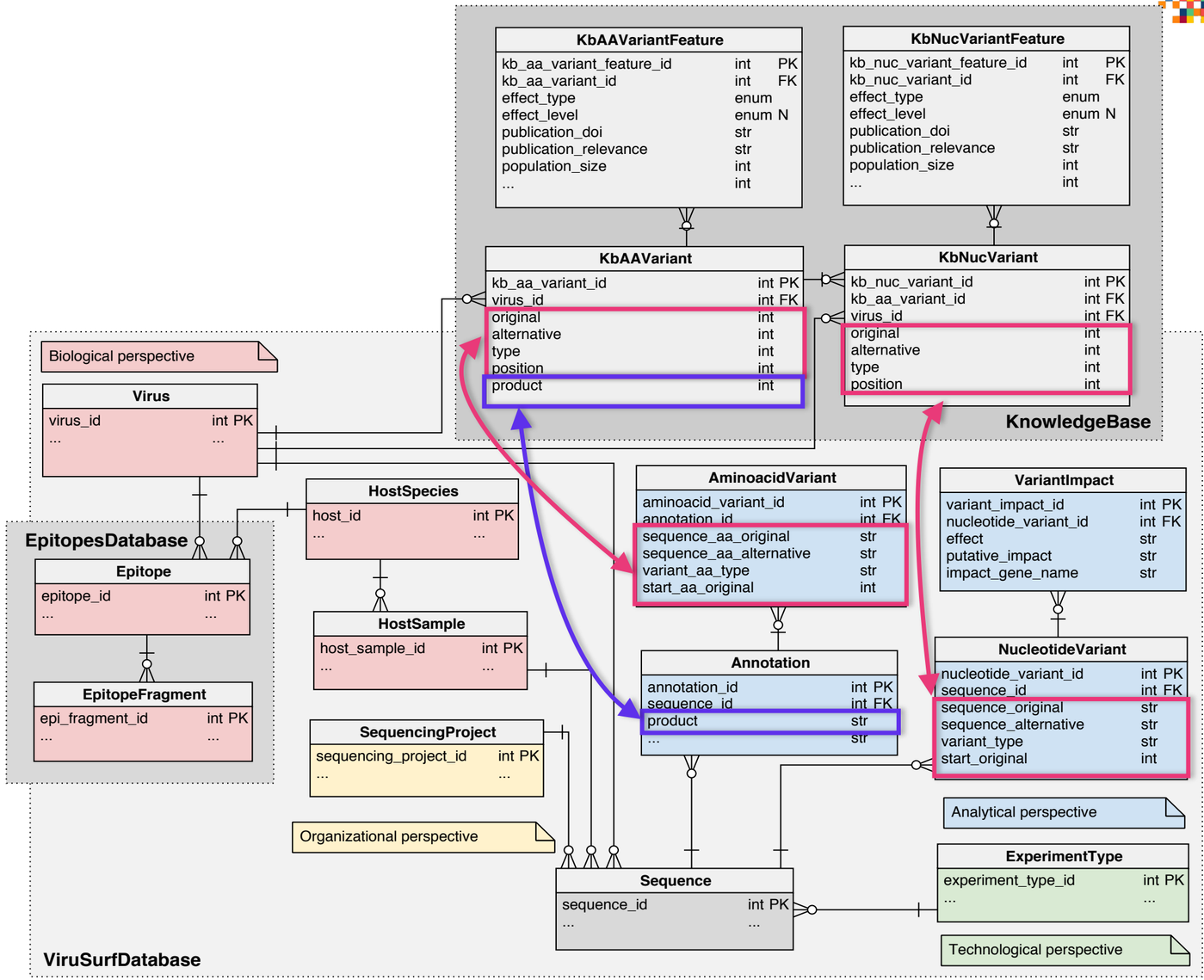
Organized collection of SARS-CoV-2 variants information, manually extracted from scientific literature

Taxonomy of variant impacts (increased/decreased by a given variant's presence)

- Protein stability
- Epidemiology
  - Viral transmission,
  - Infectivity,
  - Disease severity
  - Fatality rate
- Immunology
  - Sensitivity to convalescent sera
  - Sensitivity to neutralizing mAbs
  - Binding affinity to host receptor

…





Al Khalaf R., Alfonsi T., Ceri S., & **Bernasconi A.** (2021, May). CoV2K: a Knowledge Base of SARS-CoV-2 Variant Impacts.In International Conference on Research Challenges in Information Science (pp. 274-282). Springer, Cham. https://doi.org/10.1007/978-3-030-75018-3_18

Variants (or lineages)

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

Phylogenetic analysis:
in-depth understanding of how SARS-CoV-2 sequences evolves though genetic changes

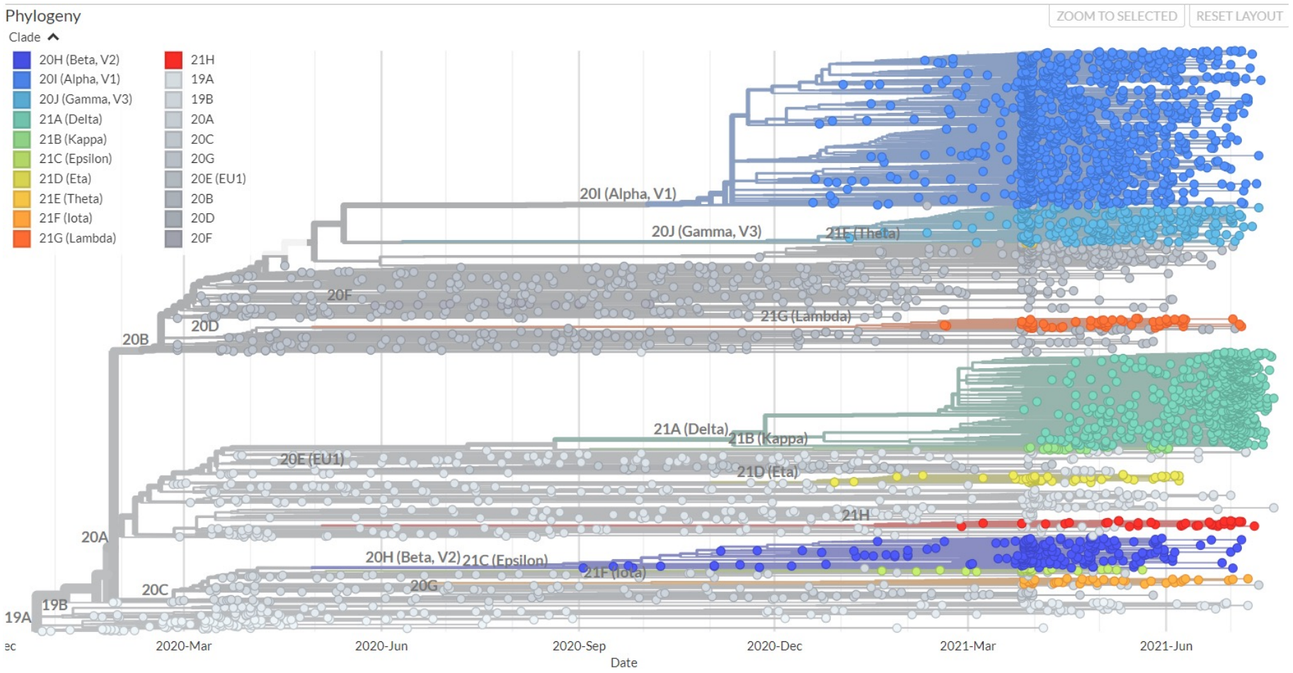


Image generated with: https://nextstrain.org/ncov/gisaid/global on Aug. 2, 2021

Lineage characteristic mutations: those amino acid changes that occur in more than 75% of lineage sequences.

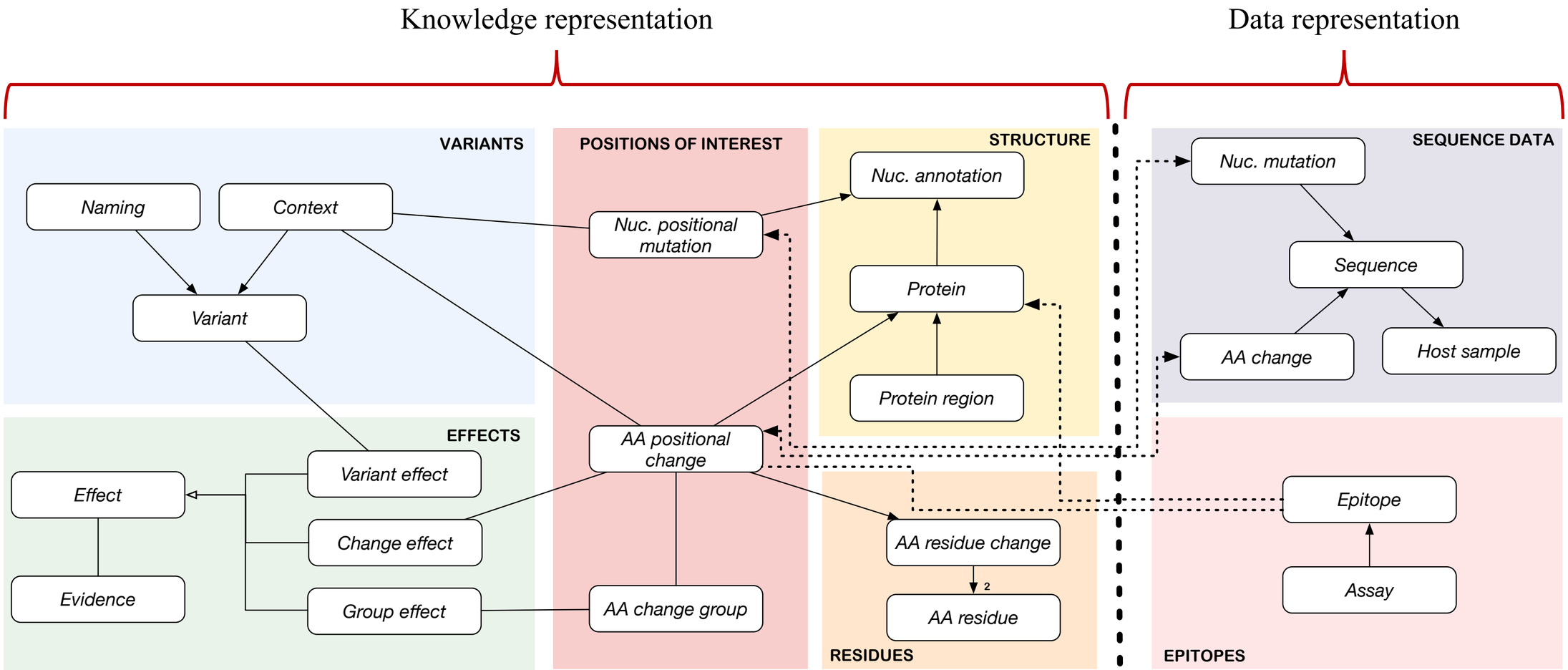
The B.1.1.7 lineage (named Alpha variant by WHO) has 22 characteristic mutations (9 on the Spike protein)

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

Information quality issues

Namings						*classes*				*Contexts*				
WHO	**Pangolin**	**GISAID (reported by WHO)**	**Nextstrain (reported by WHO)**	**Nextstrain (reported by CDC)**	**PHE**	**WHO**	**PHE**	**ECDC**	**CDC**	**AA CHANGES (ECDC)**	**AA CHANGES (CDC)**	**AA CHANGES (COVARIANTS)**	**NUC CHANGES (COVARIANTS)**	**defining SNPs (Pangolin)**
Alpha	B.1.1.7	GRY (formerly GR/501Y.V1)	20I (V1)	20I/501Y.V1	VOC-20DEC-01	VOC	VOC	VOC	VOC	S:N501Y, S:D614G, S:P681H	S:69del, S:70del, S:144del, S:(E484K*), S:(S494P*), S:N501Y, S:A570D, S:D614G, S:P681H, S:T716I, S:S982A, S:D1118H, S:(K1191N*)	S:H69-, S:V70-, S:Y144, S:N501Y, S:A570D, S:D614G, S:P681H, S:T716I, S:S982A, S:D1118H, ORF1a:T1001I, ORF1a:A1708D, ORF1a:I2230T, ORF1a:S3675- ,ORF1a:G3676-, ORF1a:F3677-, N:D3L, N:R203K, N:G204R, N:S235F, ORF1b:P314L, ORF8:Q27*, ORF8:R52I, ORF8:Y73C	C241T, C913T, C3037T, C5986T, C14676T, C15279T, T16176C	orf1ab:T1001I, orf1ab:A1708D, orf1ab:I2230T, del:11288:9, del:21765:6, del:21991:3, S:N501Y, S:A570D, S:P681H, S:T716I, S:S982A, S:D1118H, Orf8:Q27*, Orf8:R52I, Orf8:Y73C, N:D3L, N:S235F
Beta	B.1.351	GH/501Y.V2	20H (V2)	20H/501.V2	VOC-20DEC-02	VOC	VOC	VOC	VOC	S:K417N, S:E484K, S:N501Y, S:D614G, S:A701V	S:D80A, S:D215G, S:241del, S:242del, S:243del, S:K417N, S:E484K, S:N501Y, S:D614G, S:A701V	S:D80A, S:D215G, S:L241-, S:L242-, S:A243-, S:K417N, S:E484K, S:N501Y, S:D614G, S:A701V, ORF3a:Q57H, ORF1a:T265I, ORF1a:K1655N, ORF1a:K3353R, ORF1a:S3675- , ORF1a:G3676-, ORF1a:F3677-, N:T205I, ORF1b:P314L, E:P71L	G174T, C241T, C3037T, C28253T	E:P71L, N:T205I, orf1a:K1655N, S:D80A, S:D215G, S:K417N, S:A701V, S:N501Y, S:E484K
Gamma	P.1	GR/501Y.V3	20J (V3)	20J/501Y.V3	VOC-21JAN-02	VOC	VOC	VOC	VOC	S:K417T, S:E484K, S:N501Y, S:D614G, S:H655Y	S:L18F, S:T20N, S:P26S, S:D138Y, S:R190S, S:K417T, S:E484K, S:N501Y, S:D614G, S:H655Y, S:T1027I	S:L18F, S:T20N, S:P26S, S:D138Y, S:R190S, S:K417T, S:E484K, S:N501Y, S:D614G, S:H655Y, S:T1027I, S:V1176F, ORF3a:S253P, ORF1a:S1188L, ORF1a:K1795Q, ORF1a:S3675-, ORF1a:G3676-, ORF1a:F3677-, N:P80R, N:R203K, N:G204R, ORF1b:P314L, ORF1b:E1264D, ORF8:E92K	C241T, T733C, C2749T, C3037T, A6319G, A6613G, C12778T, C13860T, A28877T, G28878C	orf1ab:S1188L, orf1ab:K1795Q, del:11288:9, S:L18F, S:T20N, S:P26S, S:D138Y, S:R190S, S:K417T, S:E484K, S:N501Y, S:H655Y, S:T1027I, orf3a:G174C, orf8:E92K, N:P80R
Delta	B.1.617.2	G/478K.V1	21A	20A/S:478K	VOC-21APR-02	VOC	VOC	VOC	VOC	S:L452R, S:T478K, S:D614G, S:P681R	S:T19R, S:(G142D), S:156del, S:157del, S:R158G, S:L452R, S:T478K, S:D614G, S:P681R, S:D950N	S:T19R, S:E156-, S:F157-, S:R158G, S:L452R, S:T478K, S:D614G, S:P681R, S:D950N, ORF1b:P314L, ORF1b:P1000L, M:I82T, N:D63G, N:R203M, N:D377Y, ORF3a:S26L, ORF7a:V82A, ORF7a:T120I	G210T, C241T, C3037T, A28271-, G29742T	S:T19R, S:L452R, S:T478K, S:P681R, S:D950N, ORF3a:S26L, M:I82T, ORF7a:V82A, ORF7a:T120I, N:D63G, N:R203M, N:D377Y
Epsilon	B.1.427/B.1.429	GH/452R.V1	21C			VOI		VOI		S:L452R, S:D614G				
Epsilon	B.1.427	GH/452R.V1	21C	20C/S:452R					VOC		S:L452R, S:D614G			
Epsilon	B.1.429	GH/452R.V1	21C	20C/S:452R					VOC		S:S13I, S:W152C, S:L452R, S:D614G			
Zeta	P.2	GR/484K.V2	20B/S.484K	20J	VUI-21JAN-01	VOI	VUI	VUM	VOI	S:E484K, S:D614G	S:E484K, S:(F565L*), S:D614G, S:V1176F			
Eta	B.1.525	G/484K.V3	21D	20A/S:484K	VUI-21FEB-03	VOI	VUI	VOI	VOI	S:E484K, S:D614G, S:Q677H	S:A67V, S:69del, S:70del, S:144del, S:E484K, S:D614G, S:Q677H, S:F888L	S:Q52R, S:A67V, S:H69-, S:V70-, S:Y144-, S:E484K, S:D614G, S:Q677H, S:F888L, ORF1b:P314F, N:S2-, N:D3Y, N:A12G, N:T205I, M:I82T, ORF1a:T2007I- ORF1a:S3675-, ORF1a:G3676-, ORF1a:F3677-, E:L21F, ORF6:F2-	C241T, C1498T, A1807G, G2659A, C3037T, T8593C, C9593T, C18171T, A20724G, C24748T, A28699G, G29543T	orf1ab:L4715F, S:Q52R, S:E484K, S:Q677H, S:F888L, E:L21F, E:I82T, del:11288:9, del:21765:6, del:28278:3
Theta	P.3	GR/1092K.V1	21E	20J	VUI-21MAR-02	VOI	VUI	VOI		S:E484K, S:N501Y, S:D614G, S:P681H				
Iota	B.1.526	GH/253G.V1	21F	20C/S:484K		VOI		VUM	VOI	S:E484K, S:D614G, S:A701V	S:(L5F*), S:T95I, S:D253G, S:(S477N*), S:(E484K*), S:D614G, S:(A701V*)			
Kappa	B.1.617.1	G/452R.V3	21B	20A/S:154K	VUI-21APR-01	VOI	VUI	VOI	VOI	S:L452R, S:E484Q, S:D614G, S:P681R	S:(T95I), S:G142D, S:E154K, S:L452R, S:E484Q, S:D614G, S:P681R, S:Q1071H			
Lambda	C.37	GR/452Q.V1	20D			VOI		VUM		S:L452Q, S:F490S, S:D614G				

CoV2K abstract model



Alfonsi, T., Al Khalaf, R., Ceri, S., & **Bernasconi, A.** CoV2K model, a comprehensive representation of SARS-CoV-2 knowledge and data interplay. Submitted to Scientific Data Journal – 2-year IF 2019: 5.5, SJR: Q1

Connection with clinical data

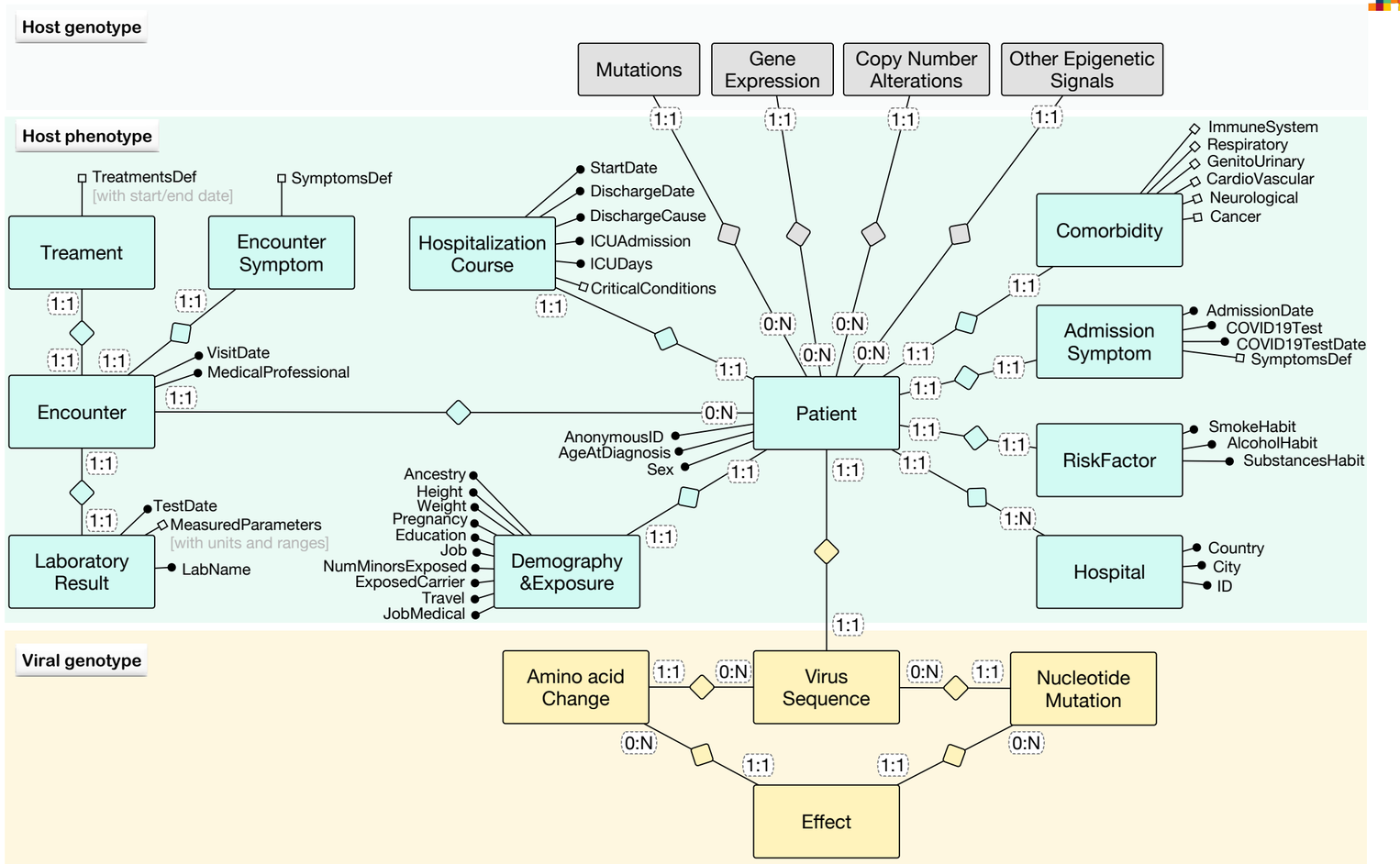
http://gmql.eu/phenotype/

The patient phenotype definition (~150 attributes) was used as a standard to collect and harmonize data from studies in the COVID-19 Host Genetics Initiative (https://www.covid19hg.org/)

Focus on Patient and his/her data at:
- Admission
- Course of hospitalization
- Discharge

Data collected through this are hosted by the **European Genome-phenome Archive** of EMBL-EBI

COVID-19 Host Genetics Initiative Coordination, **Data dictionary** working group:
Stefano Ceri and **Anna Bernasconi** (Politecnico di Milano), Alessandra Renieri and Francesca Mari (Università degli Studi di Siena)



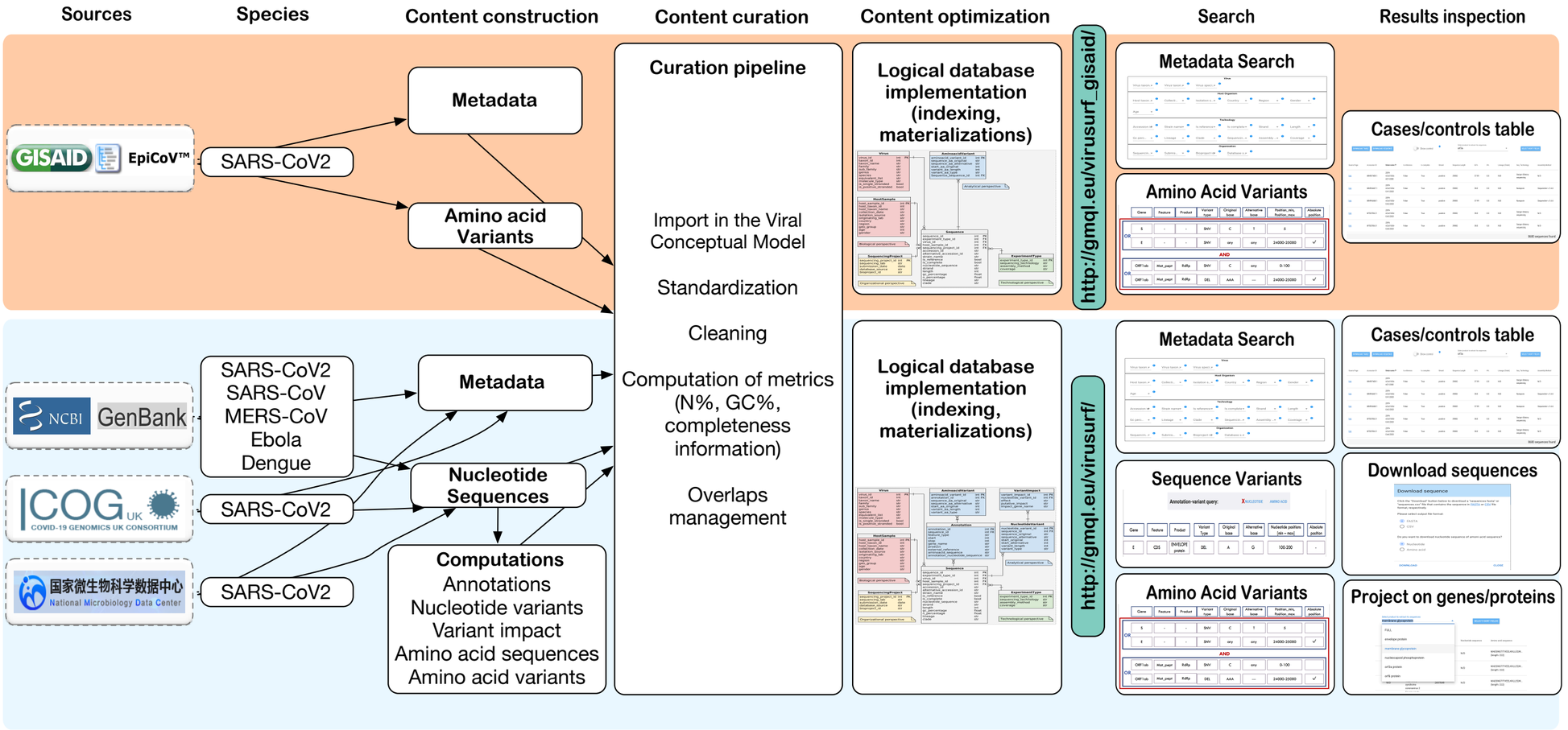
Used for studies reported in the following publications:

COVID-19 Host Genetics Initiative (2021). Mapping the human genetic architecture of COVID-19. Nature 600, 472–477 (2021). https://doi.org/10.1038/s41586-021-03767-x

Van Blokland, I.V., …, Lifelines COVID-19 cohort study, The COVID-19 Host Genetics Initiative, et al. (2021) Using symptom-based case predictions to identify host genetic factors that contribute to COVID-19 susceptibility. PloS one, 16(8), p.e0255402. https://doi.org/10.1371/journal.pone.0255402

Databases

Data integration pipelines from data sources



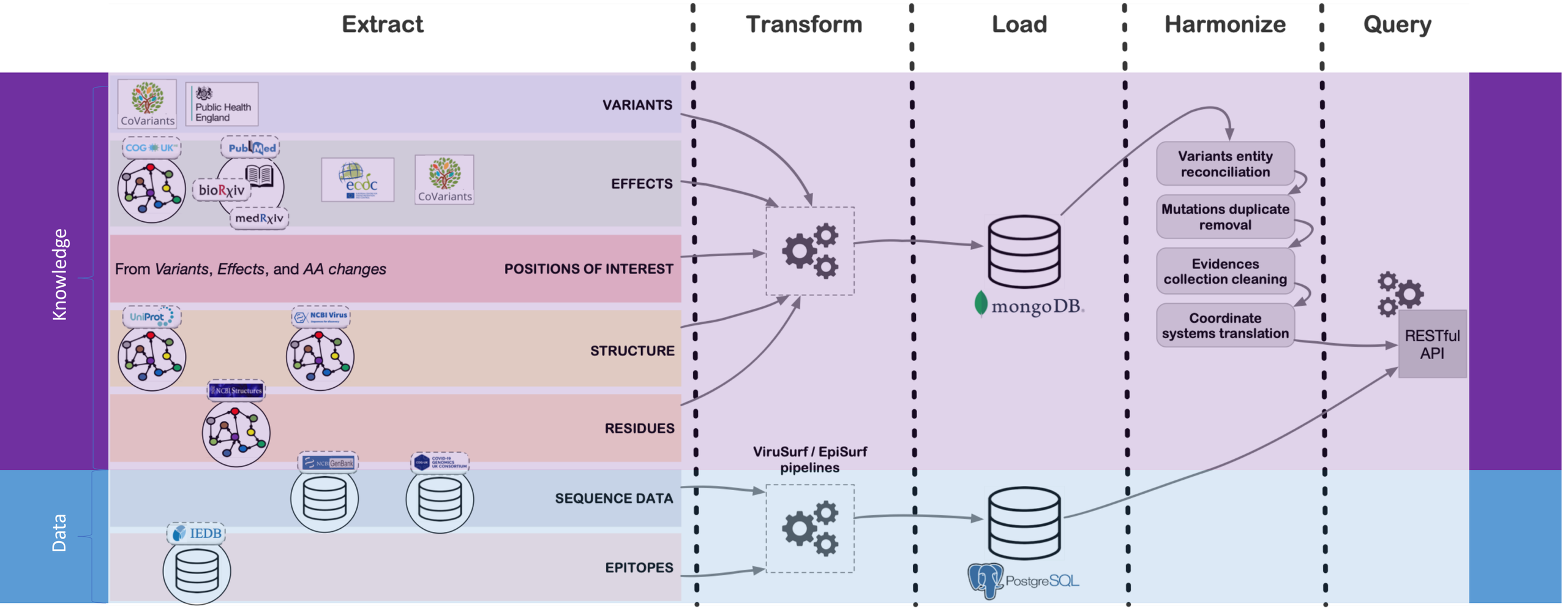
ViruSurf-GISAID content (Jan. 2021)

- GISAID EpiCoV™ db ~ 7M sequences

ViruSurf content (Jan. 2021)

- GenBank ~ 3.2M sequences (SARS-CoV-2)
- GenBank ~ 35K sequences (other viruses)
- COG-UK ~ 800K sequences
- NMDC ~ 300 sequences

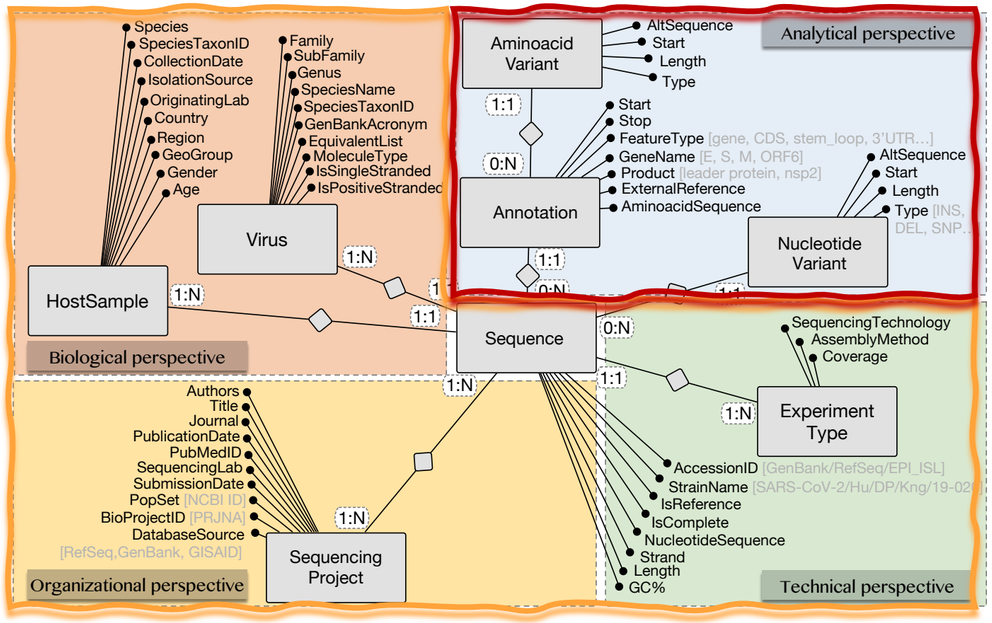
CoV2K knowledge and data integration pipeline

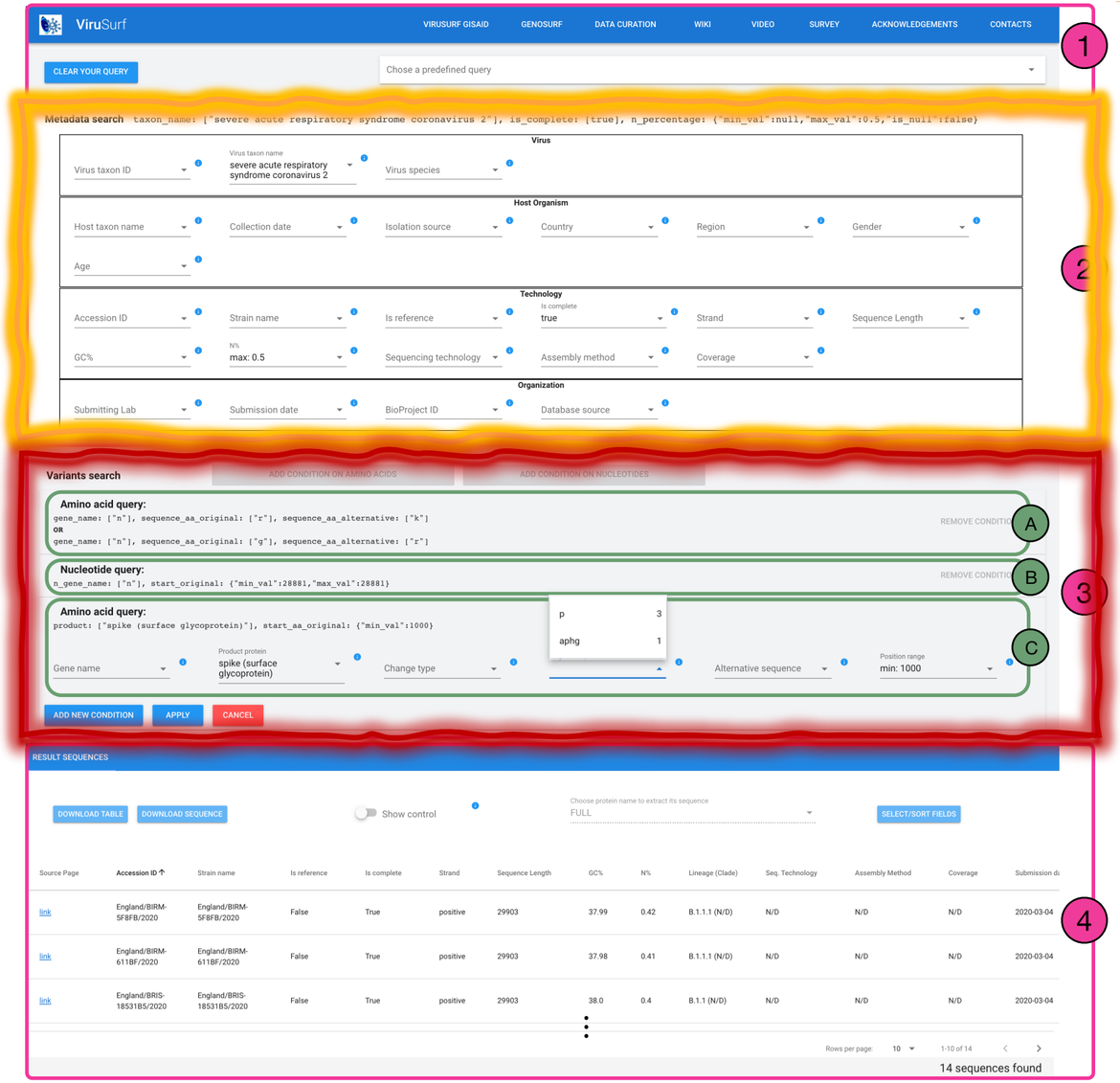


Data Tools

ViruSurf search system

http://gmql.eu/virusurf/





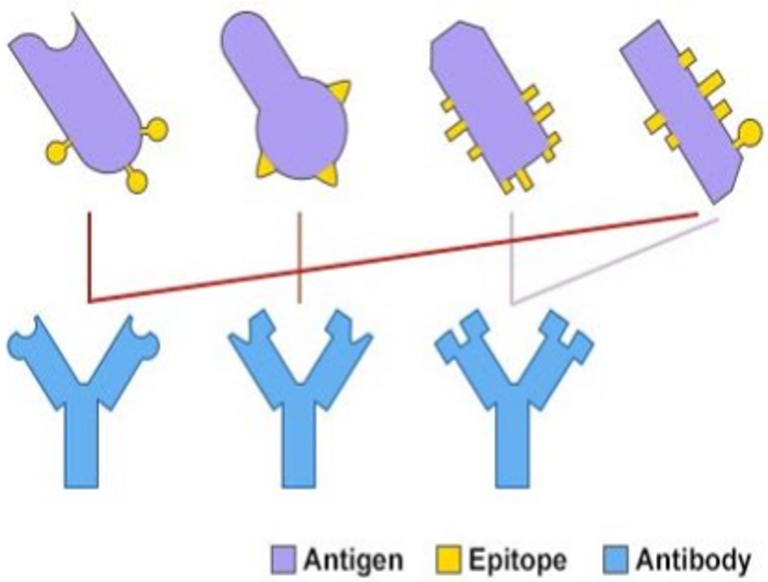
Four sections:

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.

Canakoglu, A., Pinoli, P., **Bernasconi, A.**, Alfonsi, T., Melidis, D. P., & Ceri, S. (2021). ViruSurf: an integrated database to investigate viral sequences. Nucleic Acids Research, 49(D1), D817-D824. https://doi.org/10.1093/nar/gkaa846 - IF 16.9, SJR: Q1

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) is the largest open-source collection of epitopes for many species (over 1 million epitopes, >6K for SARS-CoV-2, ~3K refer to the Spike protein)

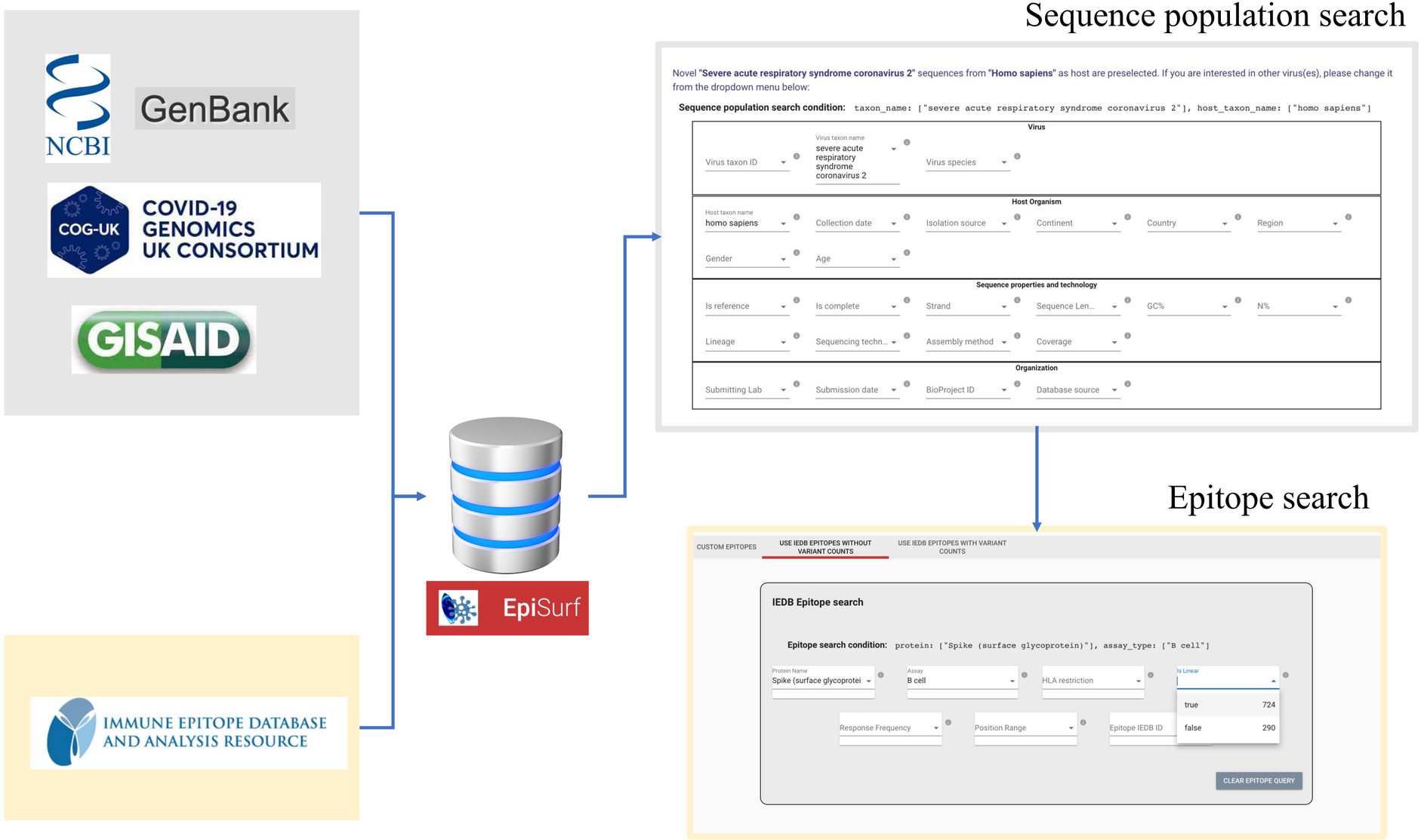
Vita, R., et al., 2019. The immune epitope database (IEDB): 2018 update. Nucleic acids research, 47(D1), pp.D339-D343.

- From what is known about Pfizer and Moderna vaccines, they use all available epitopes on the Spike protein in their design.

Jeong, D.E., et al. 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.

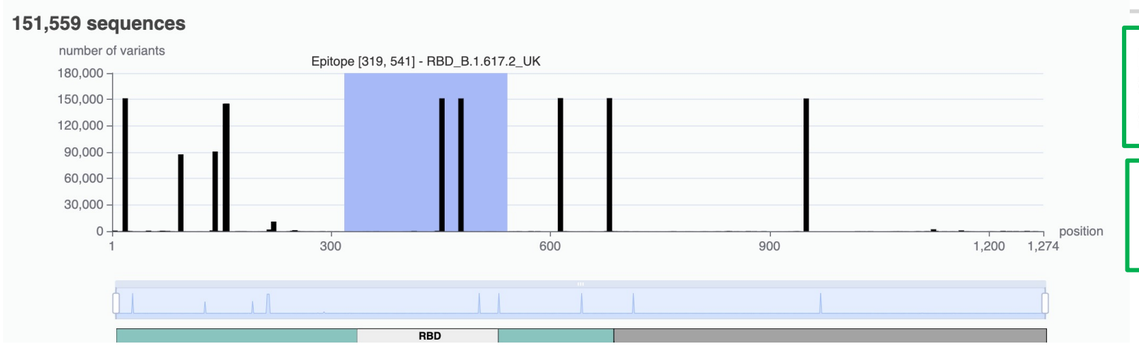
Harvey, W.T., et al. 2021. SARS-CoV-2 variants, spike mutations and immune escape. Nature Reviews Microbiology, 19(7), pp.409-424.



Bernasconi, A., Cilibrasi, L., Al Khalaf, R., Alfonsi, T., Ceri, S., Pinoli, P., & Canakoglu, A. (2021) EpiSurf: metadata-driven search server for analyzing amino acid changes on epitopes of SARS-CoV-2 and other viral species. Database, 2021. https://doi.org/10.1093/database/baab059 – IF: 3.5, SJR: Q1

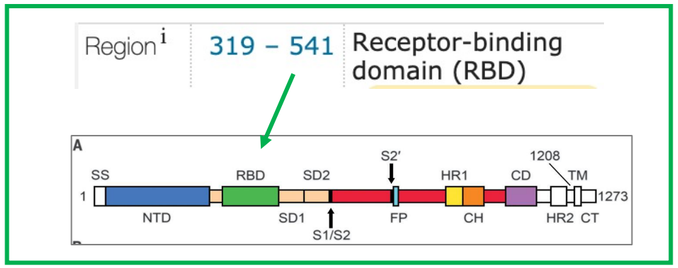
Checking the Delta mutations in UK over important epitope ranges

Spike RBD mutations and immune escape

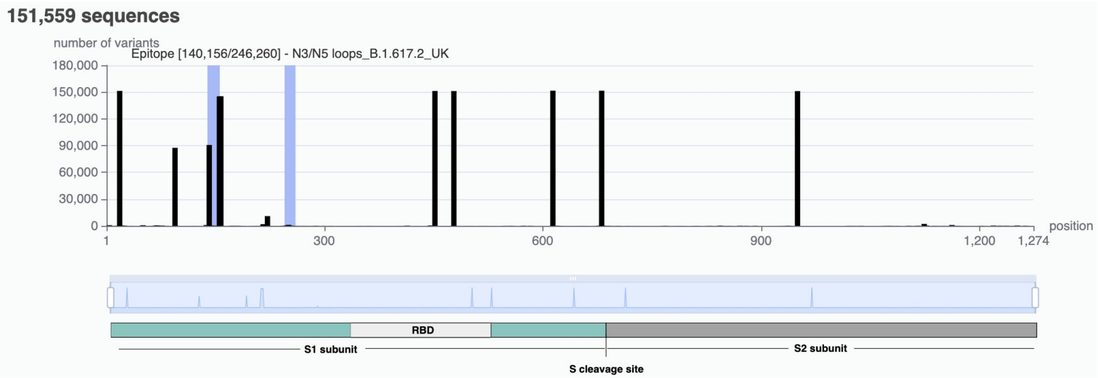


position: 452
- L→R: 151223 (100%)
- L→W: 1 (0%)

position: 478
- T→K: 151127 (100%)
- T→I: 1 (0%)

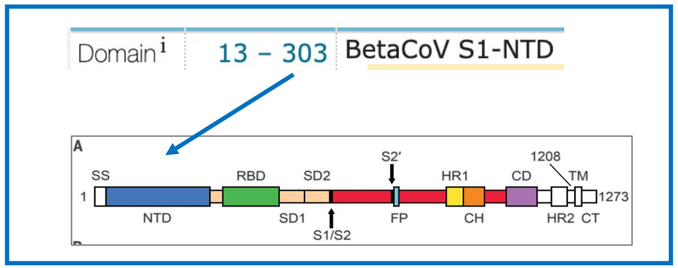


Spike NTD mutations and immune escape



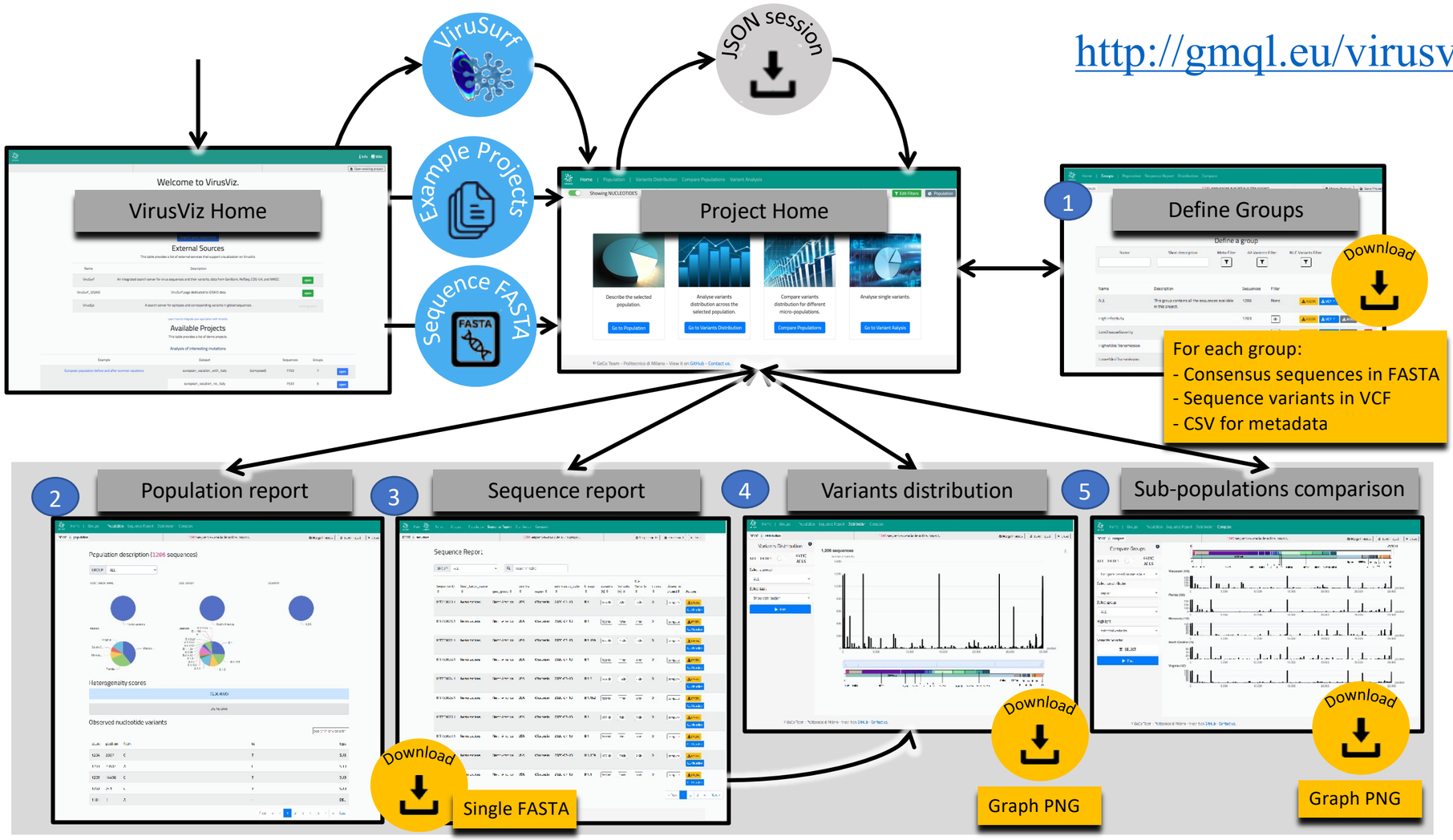
position: 142
- G→V: 1 (0%)
- G→-: 4 (0%)
- G→D: 90676 (60%)
- G→I: 1 (0%)
- G→A: 2 (0%)
- G→Y: 3 (0%)
- G→N: 4 (0%)

position: 156
- E→-: 145034 (96%)
- E→G: 4 (0%)



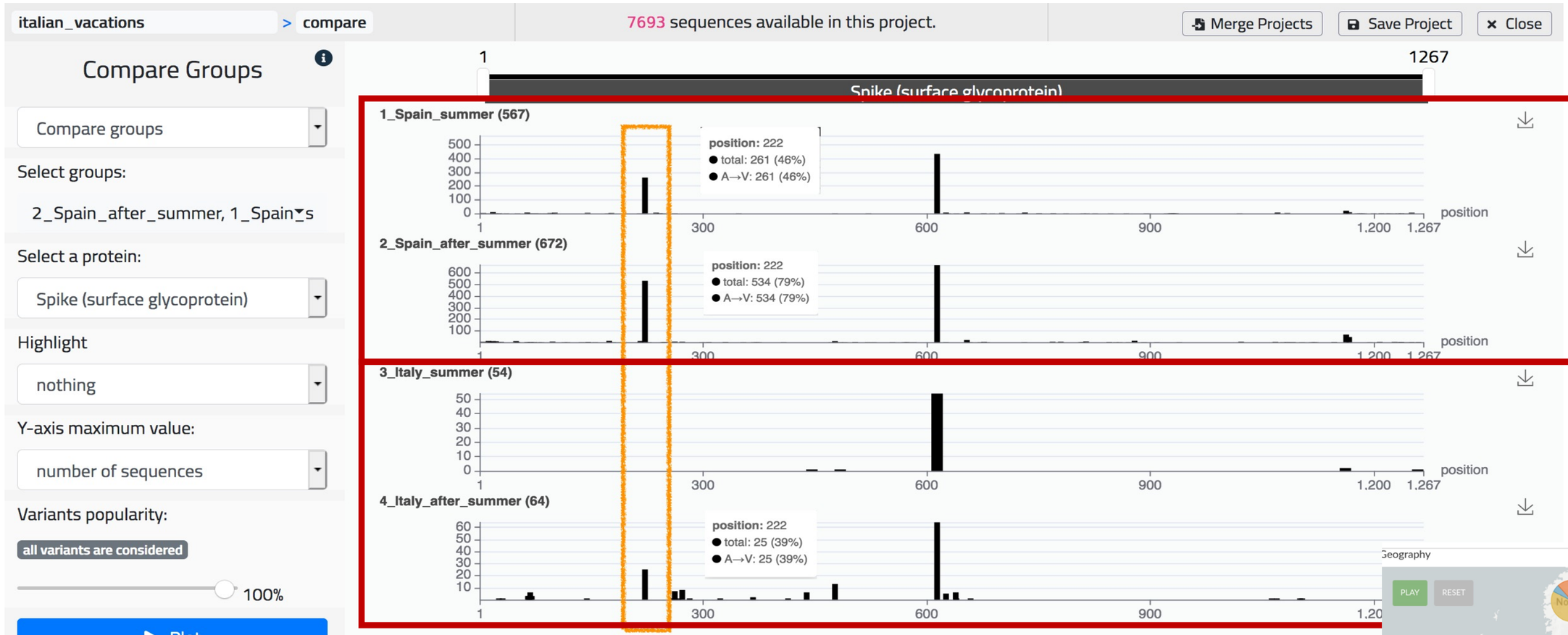
Harvey, W.T., Carabelli, A.M., Jackson, B., Gupta, R.K., Thomson, E.C., Harrison, E.M., Ludden, C., Reeve, R., Rambaut, A., Peacock, S.J. and Robertson, D.L., 2021. SARS-CoV-2 variants, spike mutations and immune escape. Nature Reviews Microbiology, 19(7), pp.409-424.

VirusViz: comparative visualization of nucleotide/amino acid mutations



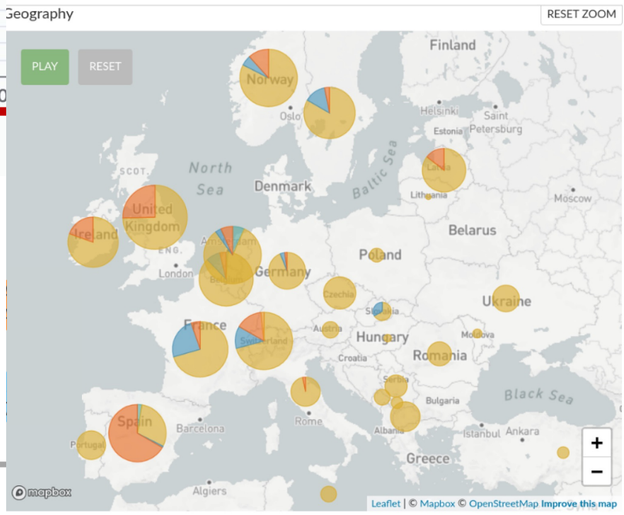
Bernasconi, A., Gulino, A., Alfonsi, T., Canakoglu, A., Pinoli, P., Sandionigi, A., & Ceri, S. (2021) VirusViz: Comparative analysis and effective visualization of viral nucleotide and amino acid variants. Nucleic Acids Research, 49(15), e90. https://doi.org/10.1093/nar/gkab478 - IF 16.9, SJR: Q1

The Spike mutation A222V arrived from Spain after Summer 2020



SPAIN
46%→ 79%

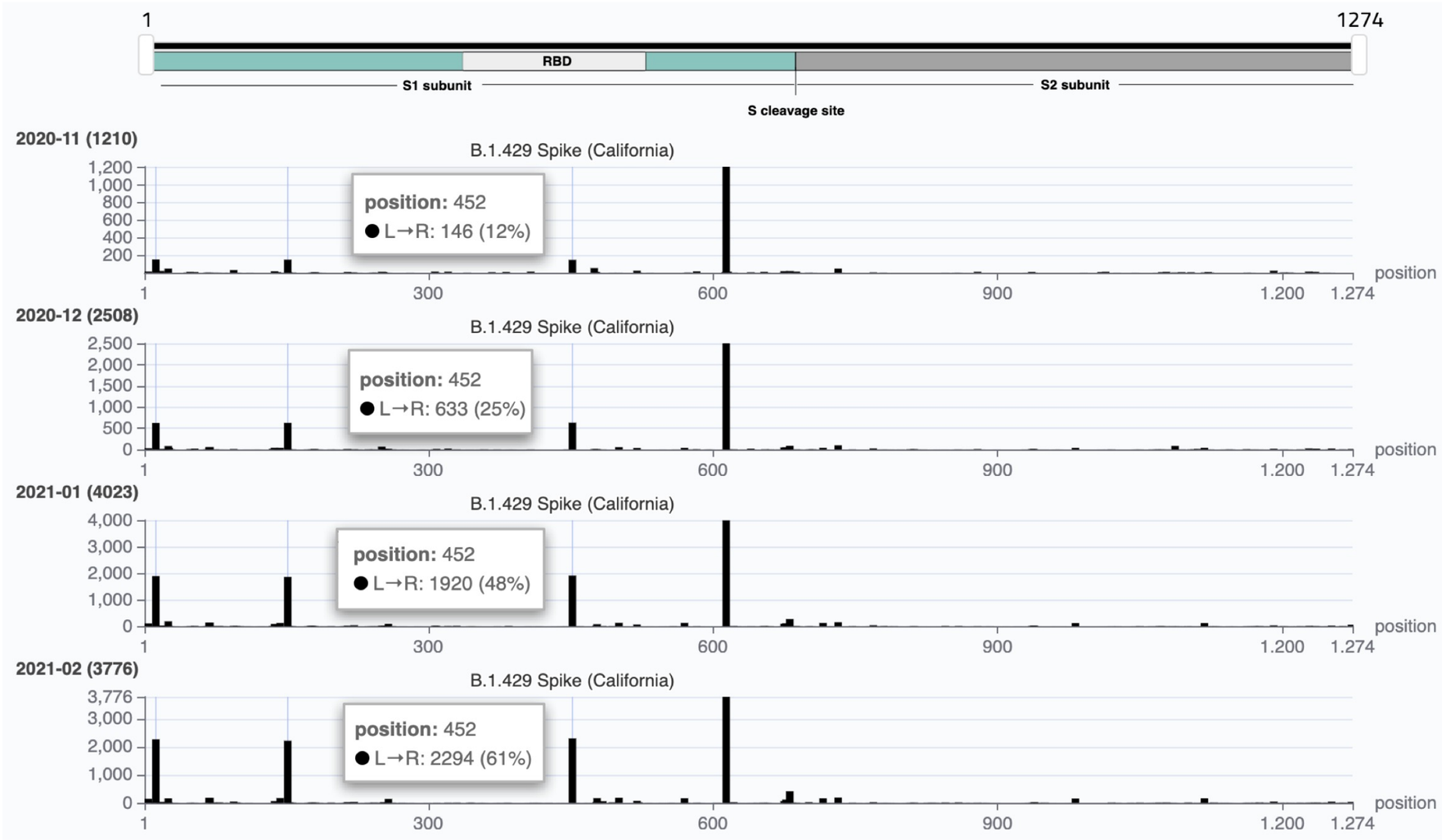
ITALY
0%→ 39%



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.

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.

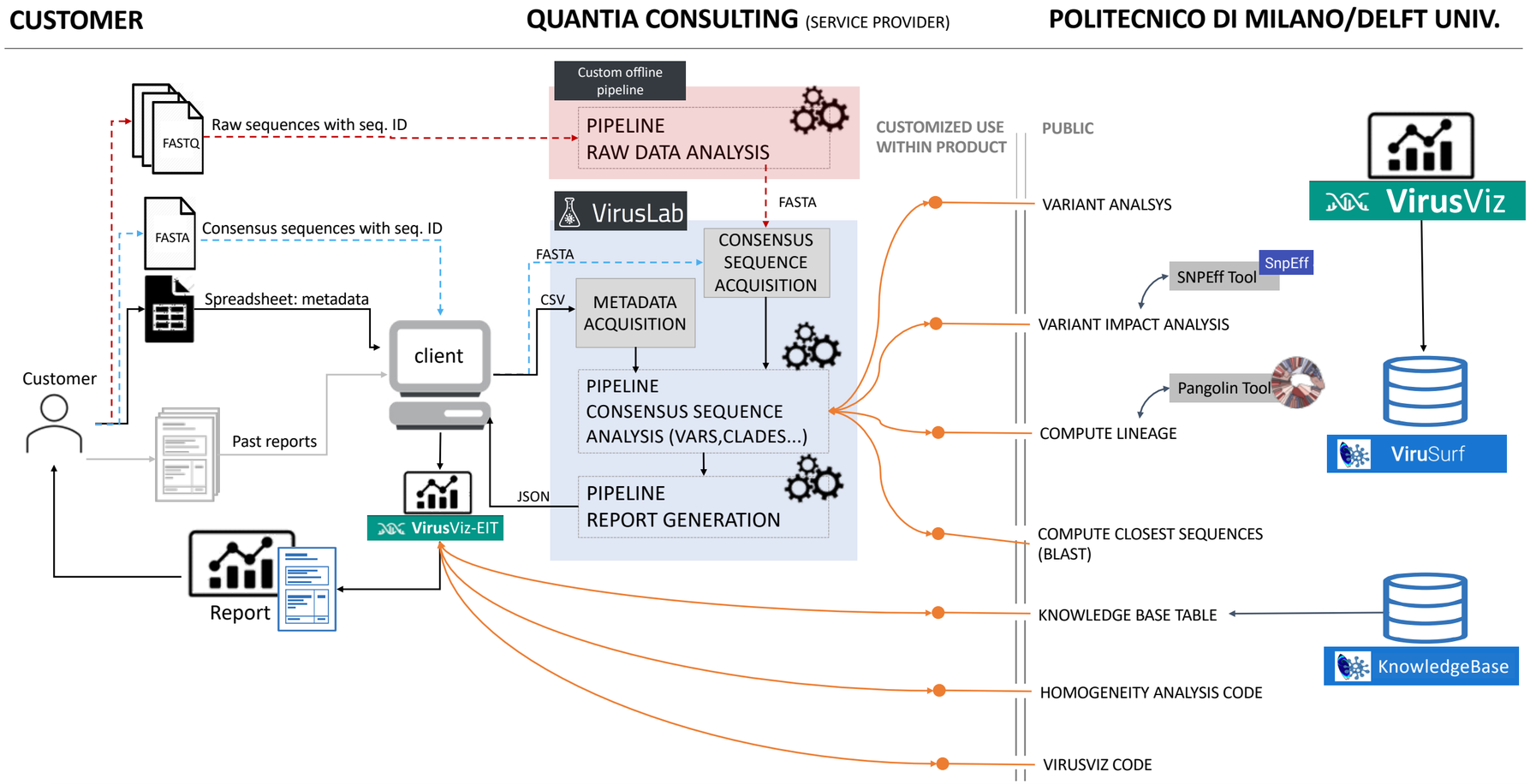


Zhang, W., Davis, B.D., Chen, S.S., Martinez, J.M.S., Plummer, J.T. and Vail, E., 2021. Emergence of a novel SARS-CoV-2 variant in Southern California. Jama, 325(13), pp.1324-1326.

VirusLab: Packaging of customizable services for use within protected sites

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

http://viruslab.quantiaconsulting.com/viruslab/



Project supported by:
EIT “DATA against COVID-19” Innovation Activity, Project 20663 “ViruSurf”

Pinoli, P., **Bernasconi, A.**, Sandionigi, A., & Ceri, S. VirusLab: a tool for customized SARS-CoV-2 data analysis. BioTech 2021, 10(4), 27. https://doi.org/10.3390/biotech10040027

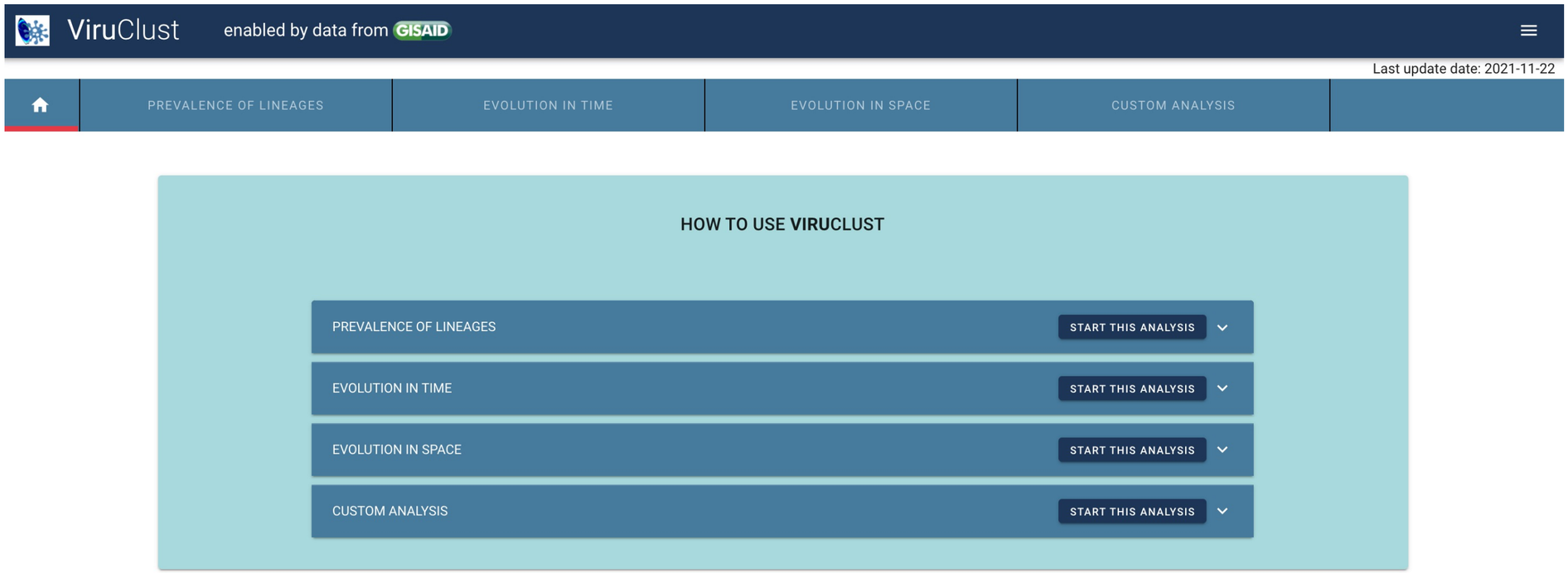
ViruClust

http://gmql.eu/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)

Integration of different types of analyses allows the generation of "testable hypotheses" that can be used to pinpoint interesting evolutionary patterns



Cilibrasi, L., Pinoli, P., **Bernasconi, A.**, Canakoglu, A., Chiara, M., & Ceri, S. ViruClust: direct comparison of SARS-CoV-2 genomes and genetic variants in space and time. Accepted by the Bioinformatics Journal – IF: 6.9, SJR: Q1 *[in print]*

ViruClust application: Early evolution of Delta

Are there mutations not observed in India (place of origin) that are instead observed in the rest of Asia?

Spatial analysis [India (target) vs. rest of Asia (background)]

- Results highlight 5 mutations observed in the spike protein of the Delta variant in India but with a much higher frequency outside of India

Temporal analysis [3 months (May to July 2021)]

- In India: the 5 changes never reach >90% prevalence, with different, increasing, profiles of frequency over time
- In the rest of Asia: the 5 changes are fixed at >90% prevalence; only one G142D show a detectable increase

If the "novel" mutations are "selected" and confer some advantages to Delta, they should not be observed in other, closely related, variants (e.g., Kappa)

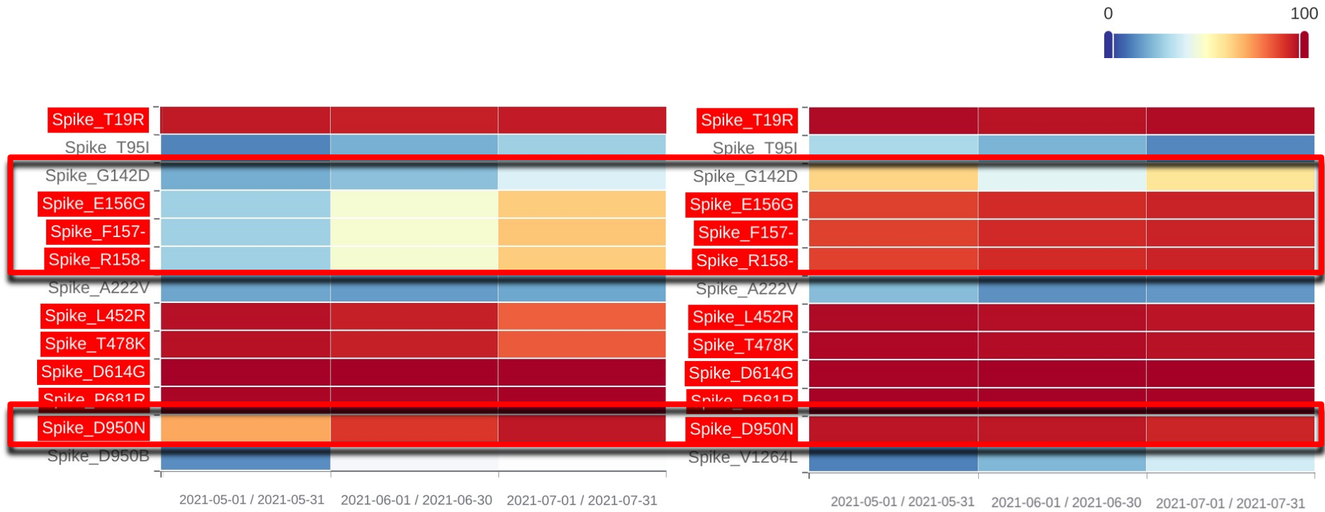
Custom analysis [Delta vs Kappa]

- Delta shows Spike:E156G, F157-, R158-, and D950N
- Kappa did not have these (only G142D)

… suggesting that the acquisition of novel/improved mutations can have epidemiological implications

TABLE

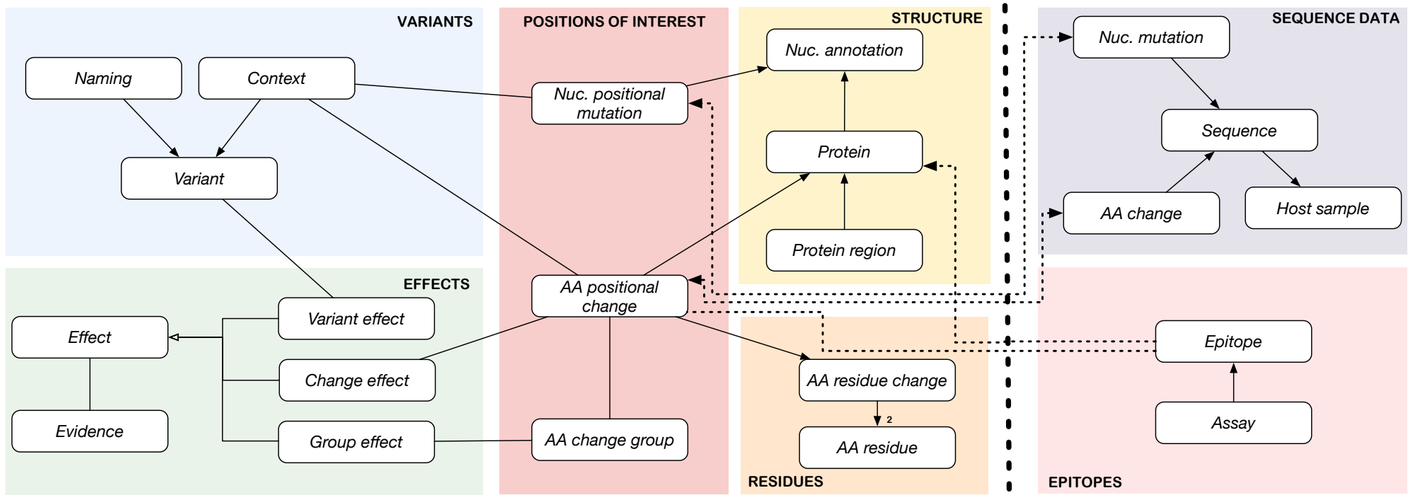
mutation	p_value ↓ 3	odds_ratio	%_target ↓ 2	%_background ↓ 1
Spike_D614G	0.13824	0.99970	99.87255 % (15673)	99.90291 % (9261)
Spike_P681R	0.00000	0.98656	98.40056 % (15442)	99.74110 % (9246)
Spike_T478K	0.00000	0.95328	94.52622 % (14834)	99.15858 % (9192)
Spike_L452R	0.00000	0.95888	95.02963 % (14913)	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)



Knowledge Tools

CoV2K-API

http://gmql.eu/cov2k/api/



One endpoint for each entity of CoV2K:

- Without parameters;
- With a *same entity identifier* as a path parameter (returning only one instance);
- With an *attribute-value pair* as a query parameter (filter on the entity);
- With *another entity identifier* query parameter (returning the set of instances of the first entity that are linked to the instances of the identified second entity)

Possibility to traverse paths of the graph (with the «combine» endpoints)

Example queries:

- What are the characteristics (Grantham distance and type) of the residue changes of the Alpha variant?
- Which amino acid changes of VOC-20DEC-02 fall within the Receptor Binding Domain (RBD)?
- Which are the effects of the variants that include the Spike amino acid change D614G?

Alfonsi, T., Al Khalaf, R., Ceri, S., & **Bernasconi, A.** CoV2K model, a comprehensive representation of SARS-CoV-2 knowledge and data interplay. Submitted to Scientific Data Journal – 2-year IF 2019: 5.5, SJR: Q1

Mutations' Effects Extraction from Abstracts

SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate

Manuel Becerra-Flores | **Timothy Cardozo**

Department of Biochemistry and Molecular Pharmacology, NYU Langone Health, New York, NY, USA

Correspondence
Timothy Cardozo, NYU Langone Health, 550 First Avenue, New York, NY 10016, USA.
Email: cardot01@nyumc.org

Funding information
No funding sources encumber this work.

Abstract

Aim: The COVID-19 pandemic is caused by infection with the SARS-CoV-2 virus. The major mutation detected to date in the SARS-CoV-2 viral envelope spike protein, which is responsible for virus attachment to the host and is also the main target for host antibodies, is a mutation of an aspartate (D) at position 614 found frequently in Chinese strains to a glycine (G). We sought to infer health impact of this mutation.

Result: Increased case fatality rate correlated strongly with the proportion of viruses bearing G614 on a country by country basis. The amino acid at position 614 occurs at an internal protein interface of the viral spike, and the presence of G at this position was calculated to destabilise a specific conformation of the viral spike, within which the key host receptor binding site is more accessible.

Conclusion: These results imply that G614 is a more pathogenic strain of SARS-CoV-2, which may influence vaccine design. The prevalence of this form of the virus should also be included in epidemiologic models predicting the COVID-19 health burden and fatality over time in specific regions. Physicians should be aware of this characteristic of the virus to anticipate the clinical course of infection.

Spike:D614G → higher fatality rate

Article

The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity

Qianqian Li,[1,2,5] Jiajing Wu,[1,5] Jianhui Nie,[1,5] Li Zhang,[1,5] Huan Hao,[1] Shuo Liu,[1] Chenyan Zhao,[1] Qi Zhang,[3] Huan Liu,[1] Lingling Nie,[1] Haiyang Qin,[1] Meng Wang,[1] Qiong Lu,[1] Xiaoyu Li,[1] Qiyu Sun,[1] Junkai Liu,[1] Linqi Zhang,[3] Xuguang Li,[4] Weijin Huang,[1,*] and Youchun Wang[1,2,6,*]

[1]Division of HIV/AIDS and Sex-Transmitted Virus Vaccines, Institute for Biological Product Control, National Institutes for Food and Drug Control (NIFDC) and WHO Collaborating Center for Standardization and Evaluation of Biologicals, No. 31 Huatuo Street, Daxing District, Beijing 102629, China
[2]Graduate School of Peking Union Medical College, No. 9 Dongdan Santiao, Dongcheng District, Beijing 100730, China
[3]Center for Global Health and Infectious Diseases, Comprehensive AIDS Research Center, and Beijing Advanced Innovation Center for Structural Biology, School of Medicine, Tsinghua University, Beijing 100084, China
[4]Centre for Vaccine Evaluation, Biologics and Genetic Therapies Directorate, HPFB, Health Canada and WHO Collaborating Center for Standardization and Evaluation of Biologicals, Ottawa, ON K1A 0K9, Canada
[5]These authors contributed equally
[6]Lead Contact
*Correspondence: huangweijin@nifdc.org.cn (W.H.), wangyc@nifdc.org.cn (Y.W.)
https://doi.org/10.1016/j.cell.2020.07.012

SUMMARY

The spike protein of SARS-CoV-2 has been undergoing mutations and is highly glycosylated. It is critically important to investigate the biological significance of these mutations. Here, we investigated 80 variants and 26 glycosylation site modifications for the infectivity and reactivity to a panel of neutralizing antibodies and sera from convalescent patients. D614G, along with several variants containing both D614G and another amino acid change, were significantly more infectious. Most variants with amino acid change at receptor binding domain were less infectious, but variants including A475V, L452R, V483A, and F490L became resistant to some neutralizing antibodies. Moreover, the majority of glycosylation deletions were less infectious, whereas deletion of both N331 and N343 glycosylation drastically reduced infectivity, revealing the importance of glycosylation for viral infectivity. Interestingly, N234Q was markedly resistant to neutralizing antibodies, whereas N165Q became more sensitive. These findings could be of value in the development of vaccine and therapeutic antibodies.

Spike:A475V → lower sensitivity to neutralizing monoclonal antibodies

Becerra-Flores, Manuel, and Timothy Cardozo. "SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate." International journal of clinical practice 74.8 (2020): e13525.

Li, Qianqian, et al. "The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity." Cell 182.5 (2020): 1284-1294

Mutations' Effects Extraction from Abstracts http://geco.deib.polimi.it/muteffstage/

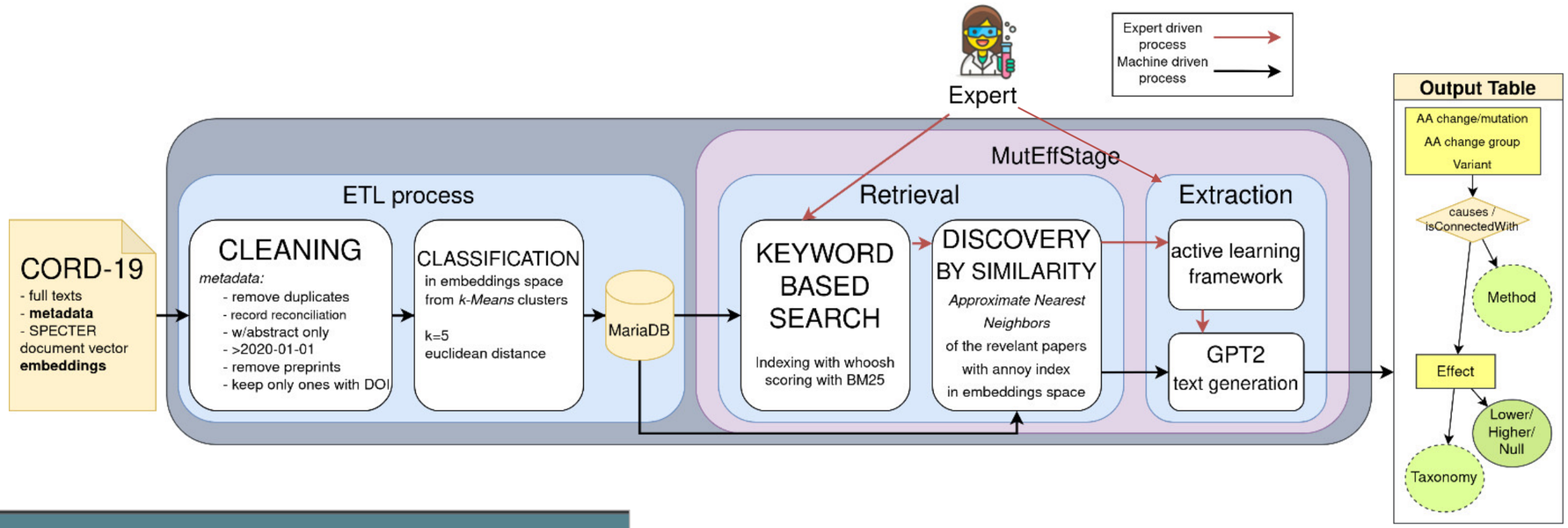
AI2 Allen Institute for AI

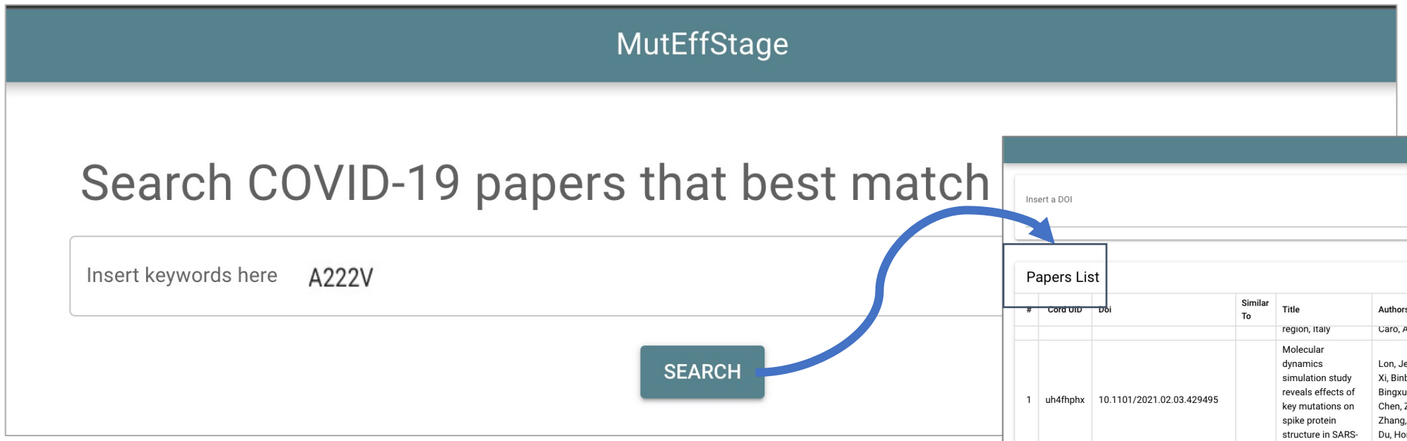
Research

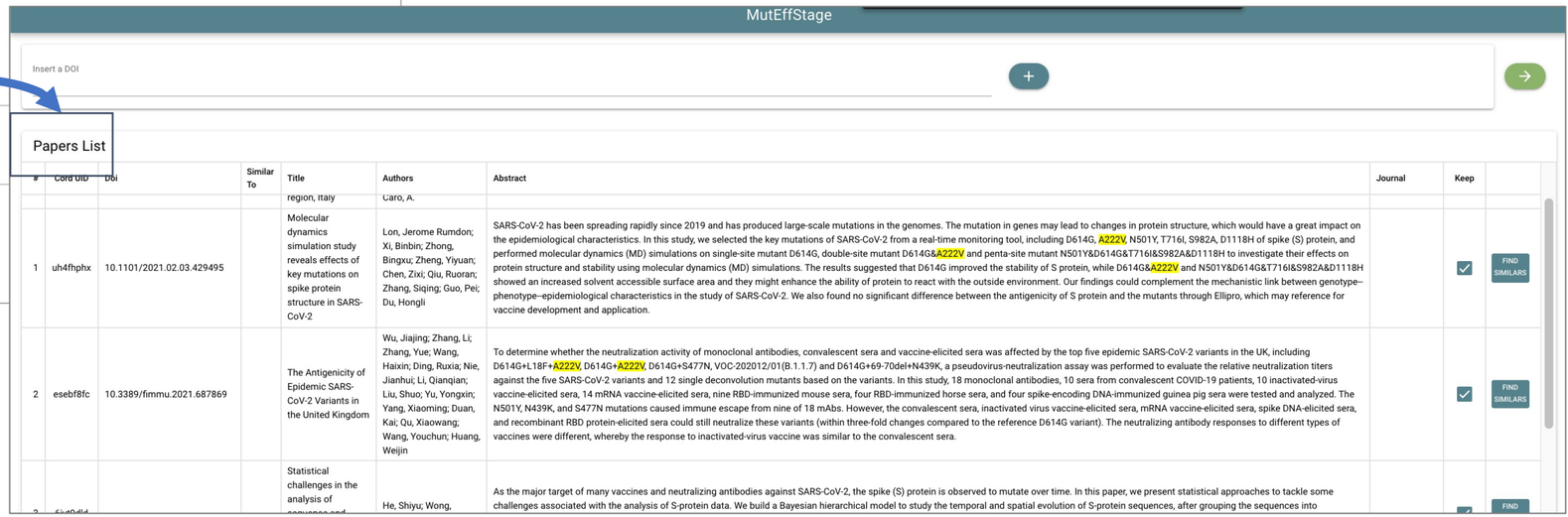
CORD-19: COVID-19 Open Research Dataset

Semantic Scholar • 2020

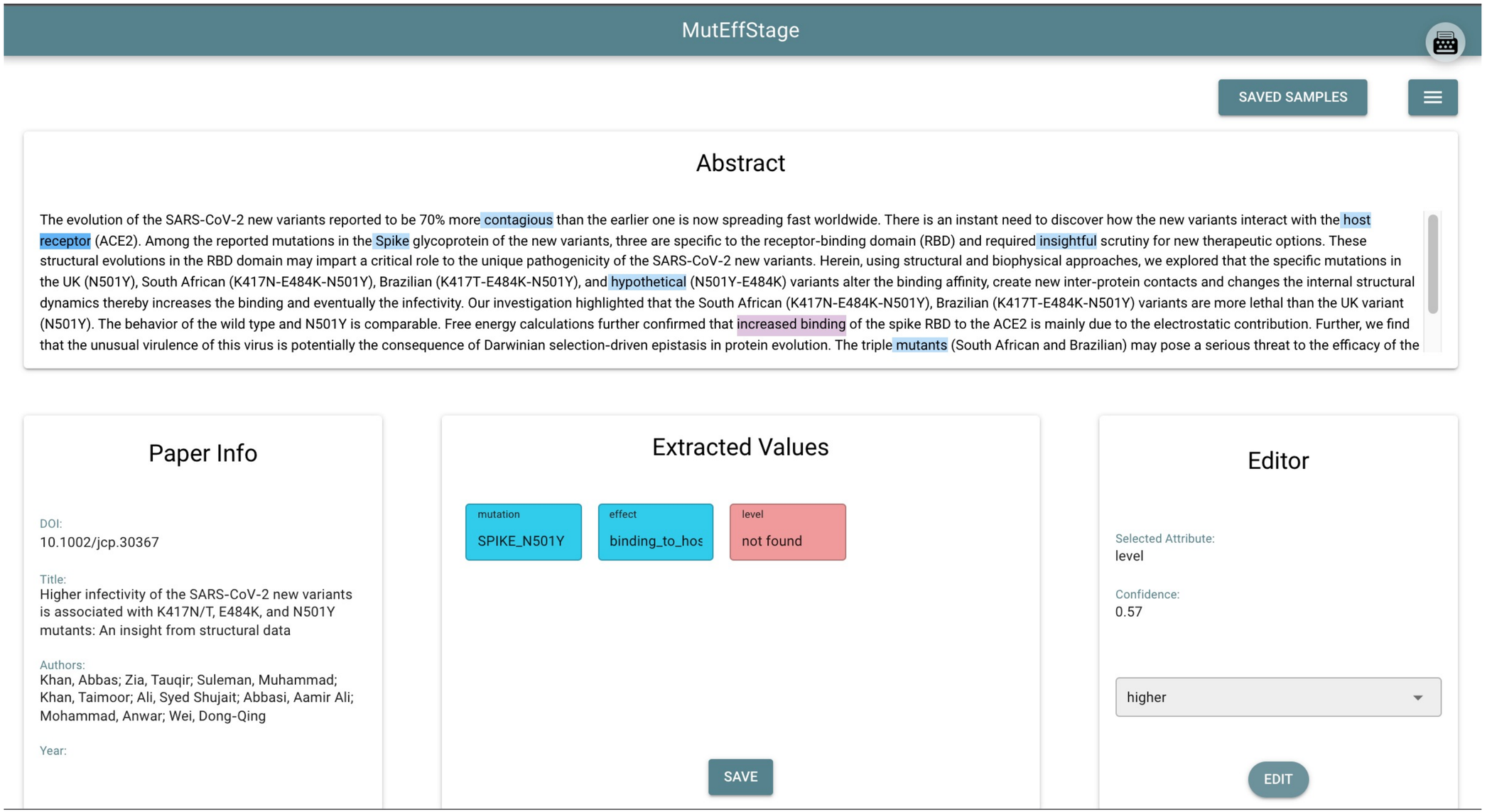
CORD-19 is a free resource of tens of thousands of scholarly articles about COVID-19, SARS-CoV-2, and related coronaviruses for use by the global research community.







MutEffStage interface using Active Learning and Saliency maps



Analysis

CoVolution

Focus: co-occurrence of specific amino acid changes, collectively named ‘virus variant’

Intuition: a variant can be identified by observing the time series dynamics of their amino acid changes. Different changes could indicate the birth of a variant if:

- Their time series (of weekly prevalences in a geo-location) are similar
- They are all growing

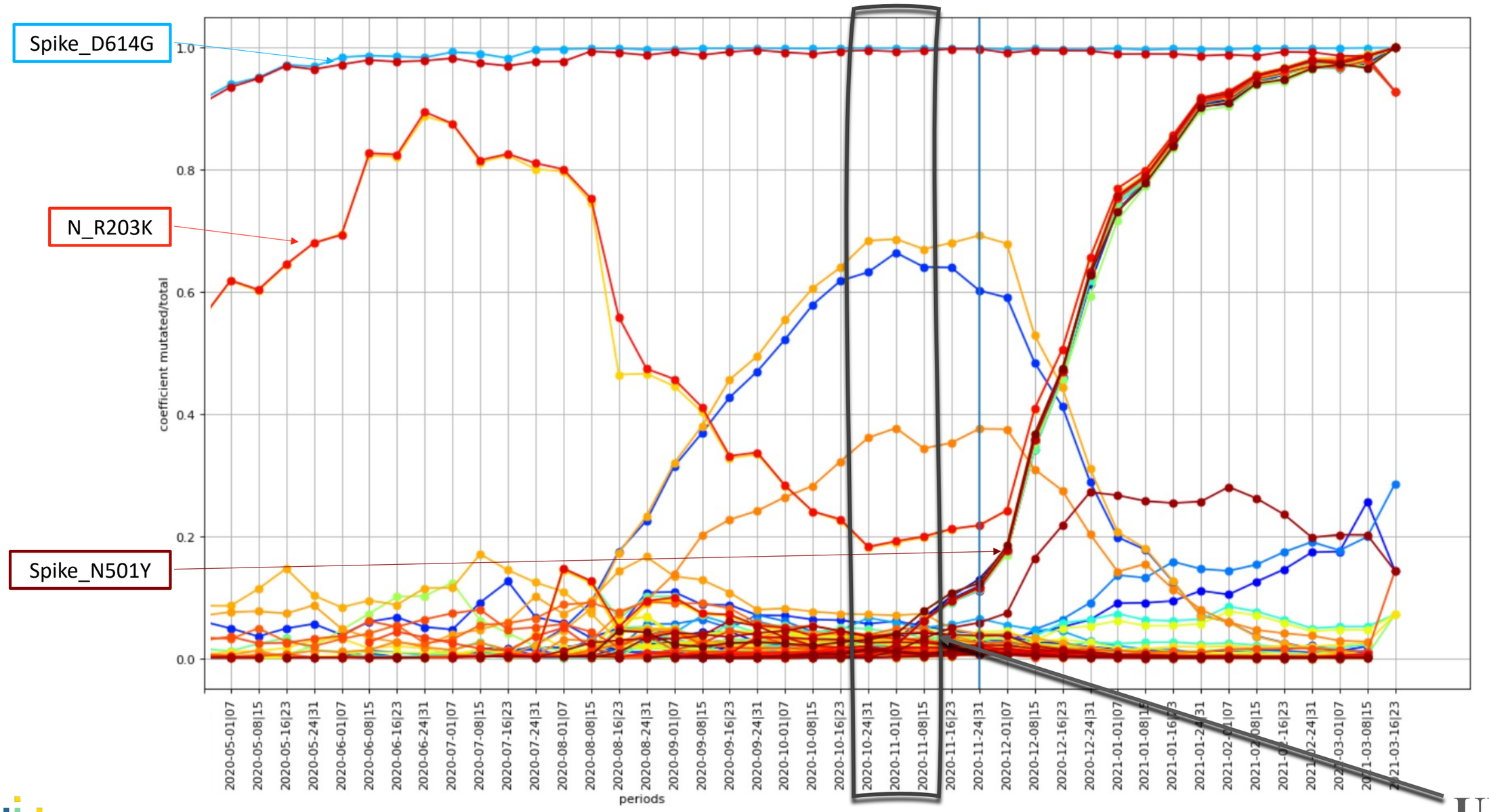
As little «soldiers»!

Outcomes:

- The emergence of variants can be traced through purely data-driven methods
- An early warning system could rely exclusively on deposited sequences

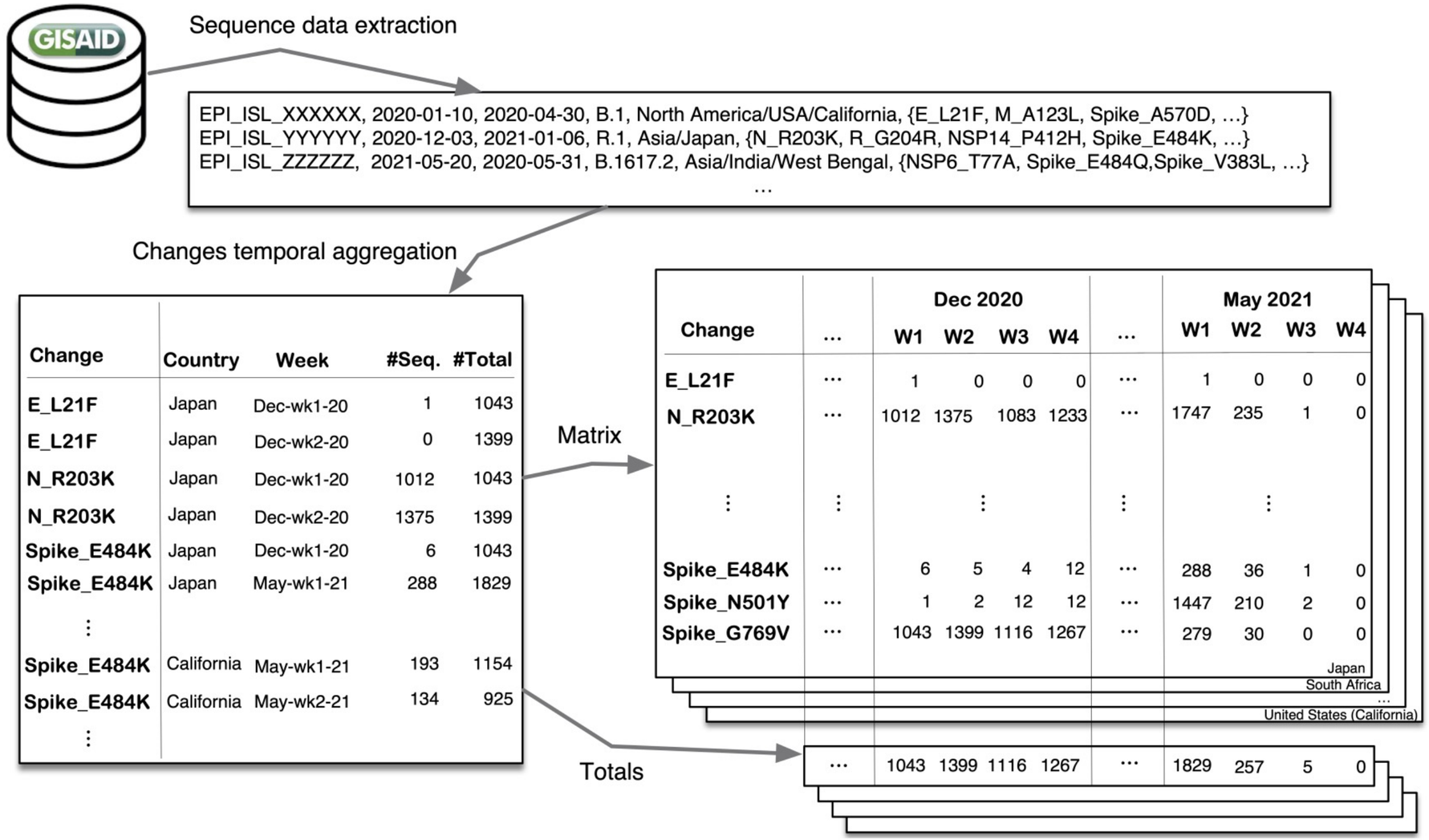
Bernasconi, A., Mari, L., Casagrandi, R., & Ceri, S. (2021) Data-driven analysis of amino acid change dynamics timely reveals SARS-CoV-2 variant emergence. Scientific Reports 11, 21068(2021). https://doi.org/10.1038/s41598-021-00496-z – 2-year IF: 4.4, SJR: Q1

Intuition



UK variant

Data extraction and temporal aggregation



Lineage Dictionary

For all relevant known lineages, we computed the set of characterizing changes, pragmatically defined as those that appear in at least 75% of the sequences annotated with that lineage in GISAID.

Label	Dictionary: changes in > 75% sequences
B.1.1.214(JP1)	N_M234I, NSP14_P43L, NSP16_R287I
B.1.1.284(JP2)	N_P151L, NSP12_A423V, NSP3_S543P, NSP5_P108S
B.1.1.519	NSP3_P141S, NSP4_T492I, NSP6_I49V, NSP9_T35I, Spike_P681H, Spike_T478K, Spike_T732A
B.1.1.7(Alpha)	N_D3L, N_S235F, NS8_Q27*, NS8_R52I, NS8_Y73C, NSP3_A890D, NSP3_I1412T, NSP3_T183I, NSP6_F108-, NSP6_G107-, NSP6_S106-, Spike_A570D, Spike_D1118H, Spike_H69-, Spike_N501Y, Spike_P681H, Spike_S982A, Spike_T716I, Spike_V70-, Spike_Y144-
B.1.160.[|20]	N_A376T, N_M234I, NS3_Q57H, NSP12_A185S, NSP12_V776L, NSP13_E261D, NSP13_K218R, NSP4_M324I, Spike_S477N
B.1.177.12	N_A220V, N_P365S, NS3_R122I, Spike_A222V
B.1.177.[7|8|48|59|65]	N_A220V, Spike_A222V, Spike_L18F
B.1.177.86	N_A220V, NS7b_E39*, Spike_A222V, Spike_L18F
B.1.177.[|21|50|75]|Z.1	N_A220V, Spike_A222V
B.1.2(US1)	N_P199L, N_P67S, NS3_G172V, NS3_Q57H, NS8_S24L, NSP14_N129D, NSP16_R216C, NSP2_T85I, NSP5_L89F
B.1.214.2	N_D3L, NSP12_R583G, NSP3_I580V, NSP3_T1063I, NSP8_A74V, N_T205I, Spike_N450K, Spike_Q414K, Spike_T716I
B.1.221	N_P199L, NS3_G172R, NS3_Q38R, NS3_V202L, NSP3_H295Y, Spike_S98F
B.1.234	N_S194L, NSP2_V485I, NSP3_K1241R, NSP4_D217G, N_T391I
B.1.258.17	NS3_Q185H, NS8_E64*, NSP12_V720I, NSP13_A598S, NSP13_H290Y, NSP13_P53L, NSP14_E453D, NSP14_P46L, NSP3_I1683T, NSP9_M101I, Spike_H69-, Spike_L189F, Spike_N439K, Spike_V70-, Spike_V772I
B.1.258.[|24]	NSP13_H290Y, NSP3_I1683T, Spike_N439K
B.1.351(Beta)	E_P71L, NS3_Q57H, NS3_S171L, NSP2_T85I, NSP3_K837N, NSP5_K90R, NSP6_F108-, NSP6_G107-, NSP6_S106-, N_T205I, Spike_A243-, Spike_A701V, Spike_D215G, Spike_D80A, Spike_E484K, Spike_K417N, Spike_L242-, Spike_L244-, Spike_N501Y
B.1.36.[|8|16|24]	N_S194L, NS3_Q57H
B.1.427(Epsilon)	NS3_Q57H, NSP13_D260Y, NSP13_P53L, NSP2_T85I, NSP4_S395T, N_T205I, Spike_L452R, Spike_S13I, Spike_W152C

⋮

Rambaut, Andrew, et al. "A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology." Nature microbiology 5.11 (2020): 1403-1407.
Shu, Yuelong, and John McCauley. "GISAID: Global initiative on sharing all influenza data–from vision to reality." *Eurosurveillance* 22.13 (2017): 30494.

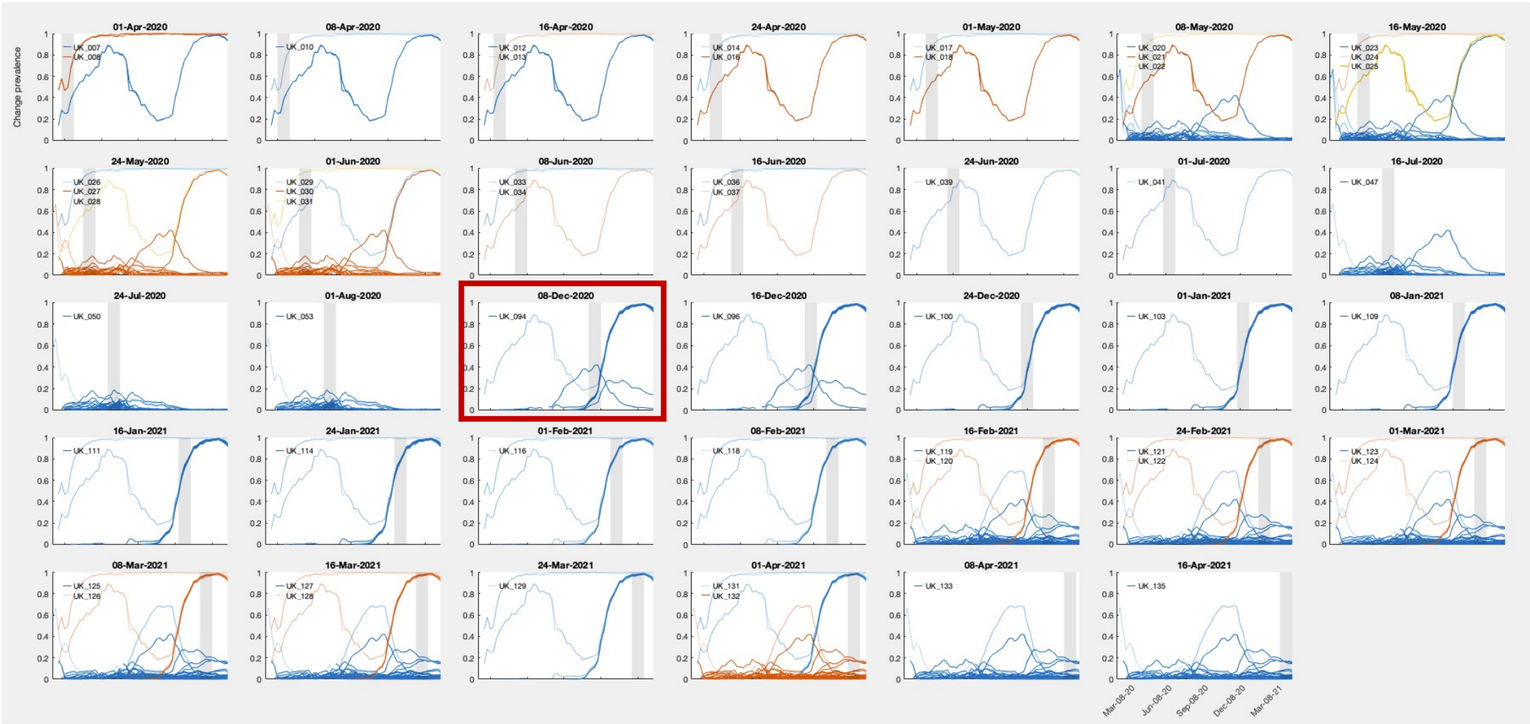
Data analysis method 1

Do temporal patterns of changes' prevalence occur in a coherent manner?

Time-series clustering algorithm

Method:

- For each t, consider the *n(t)* time-series of a change prevalence (# seq. w change / # total seq.) observed continuously over the 4 weeks prior to t
- Partition time-series via k-medoids clustering (PAM algorithm, pairwise distances between time-series evaluated via dynamic time warping)
- The optimal value of *k* is the one that maximizes the average silhouette score evaluated over the set of the *n(t)* change time-series



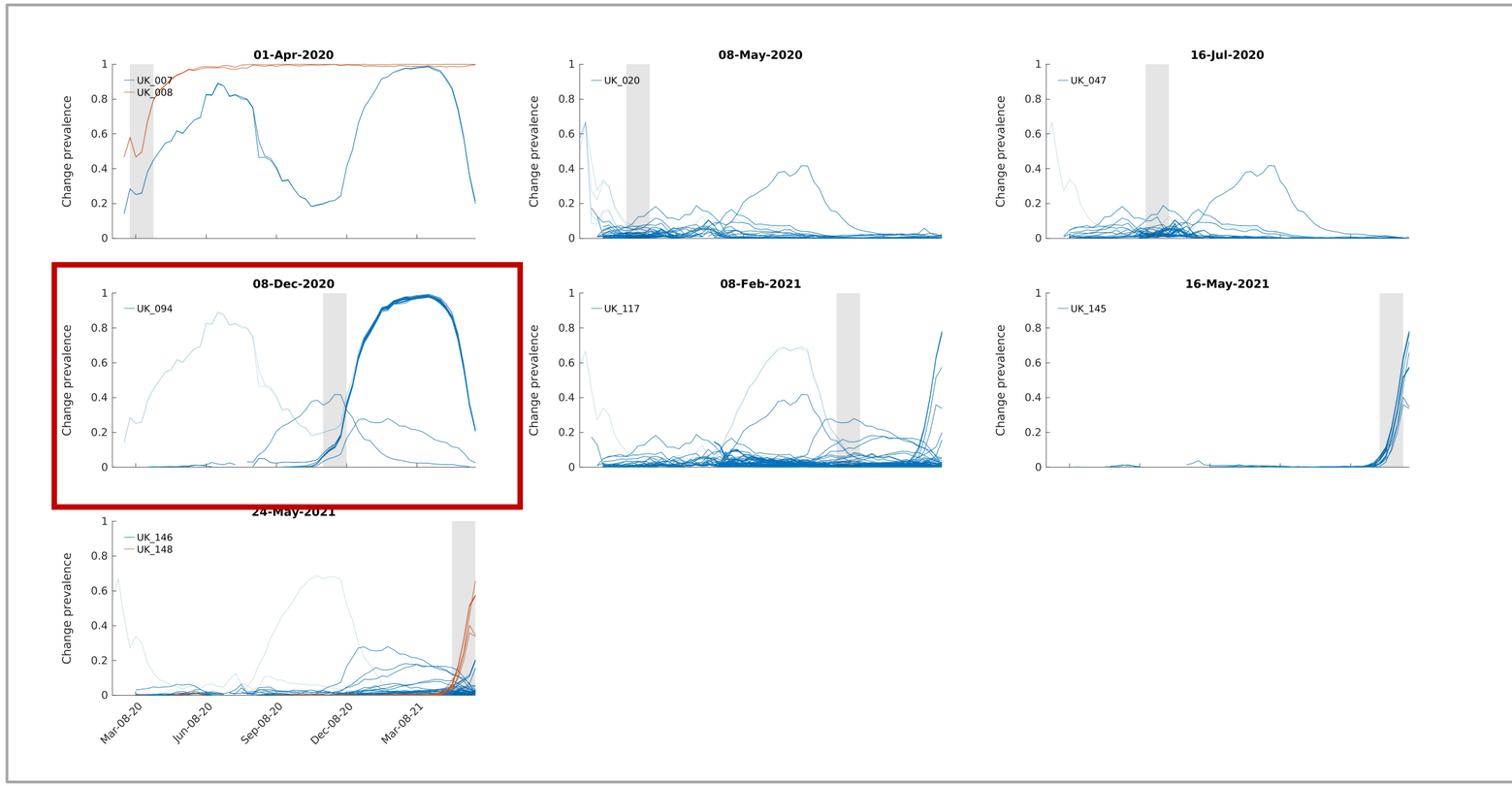
Data analysis method 2

Which clusters could signal the emergence of a new virus variant?

Early-warning clusters detection

Method:

- Identify clusters of changes with an increasing trend in their prevalence time-series (non-parametric Kendall's τB statistic, select those with a positive trend $\tau B>0$ at significance level $\alpha=0.05$)
- Evaluate the similarity for each pair of clusters with Jaccard index Jcc (cardinality of intersection / cardinality of the union of the two clusters)
- Retain clusters sufficiently different from previously observed trending clusters ($Jcc<0.5$)



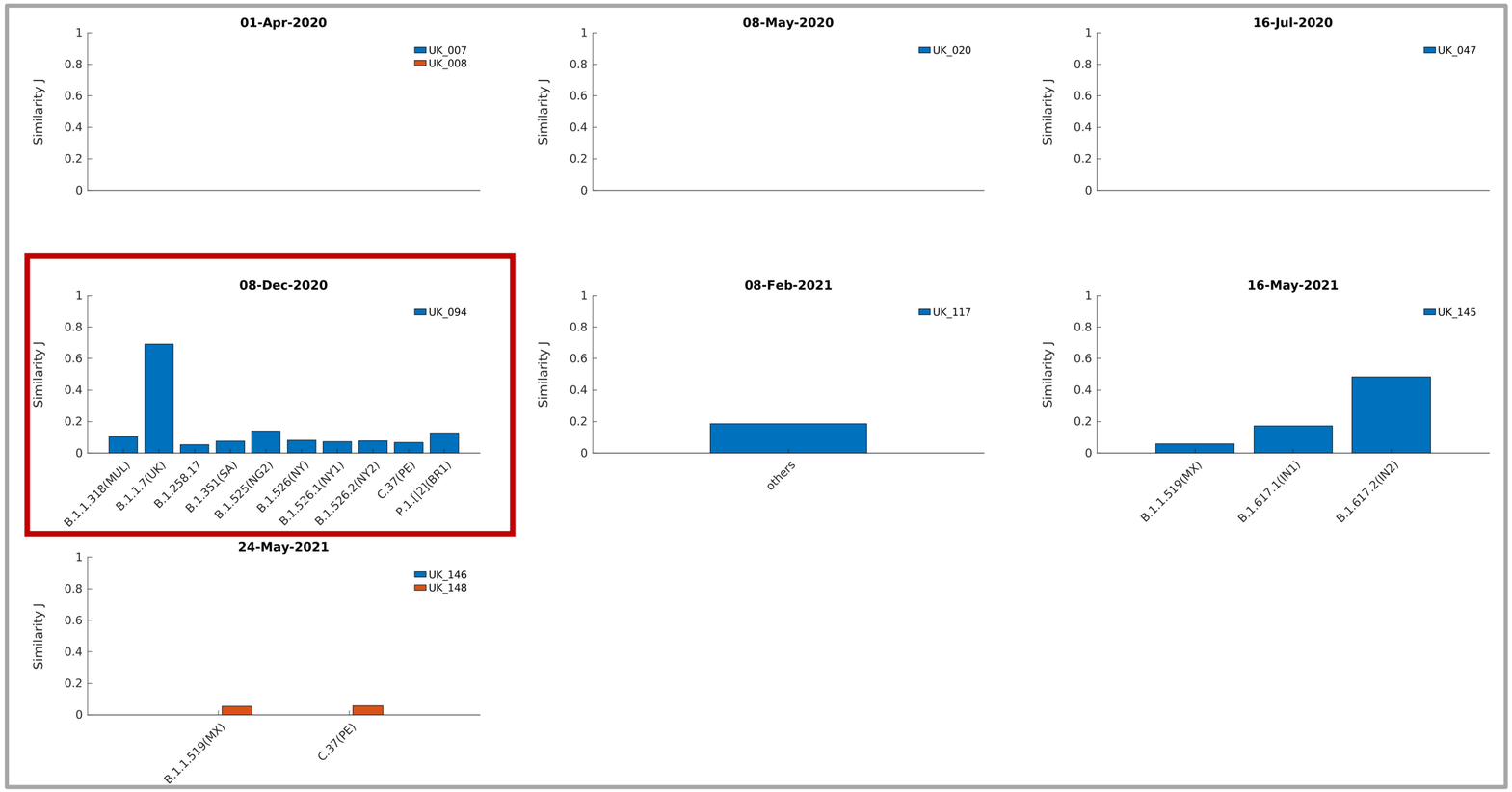
Data analysis method 3

Which clusters resemble (a posteriori) known lineages?

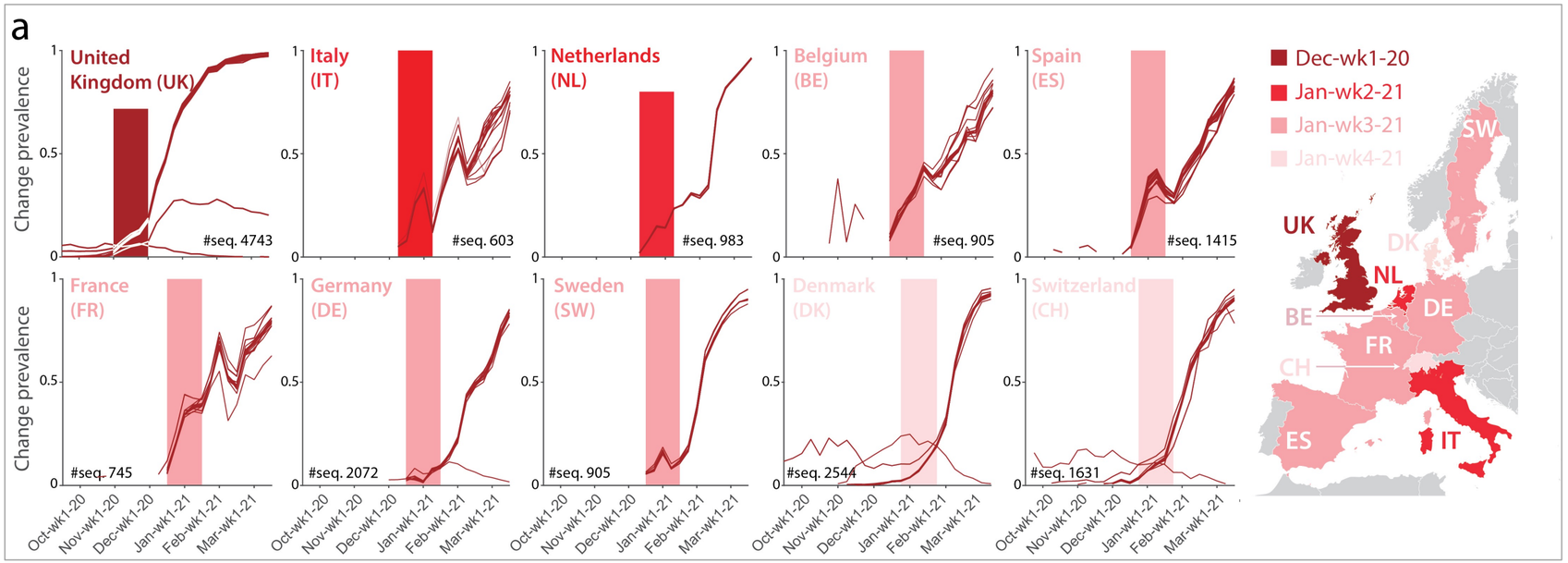
Cluster-dictionary comparison

Method:

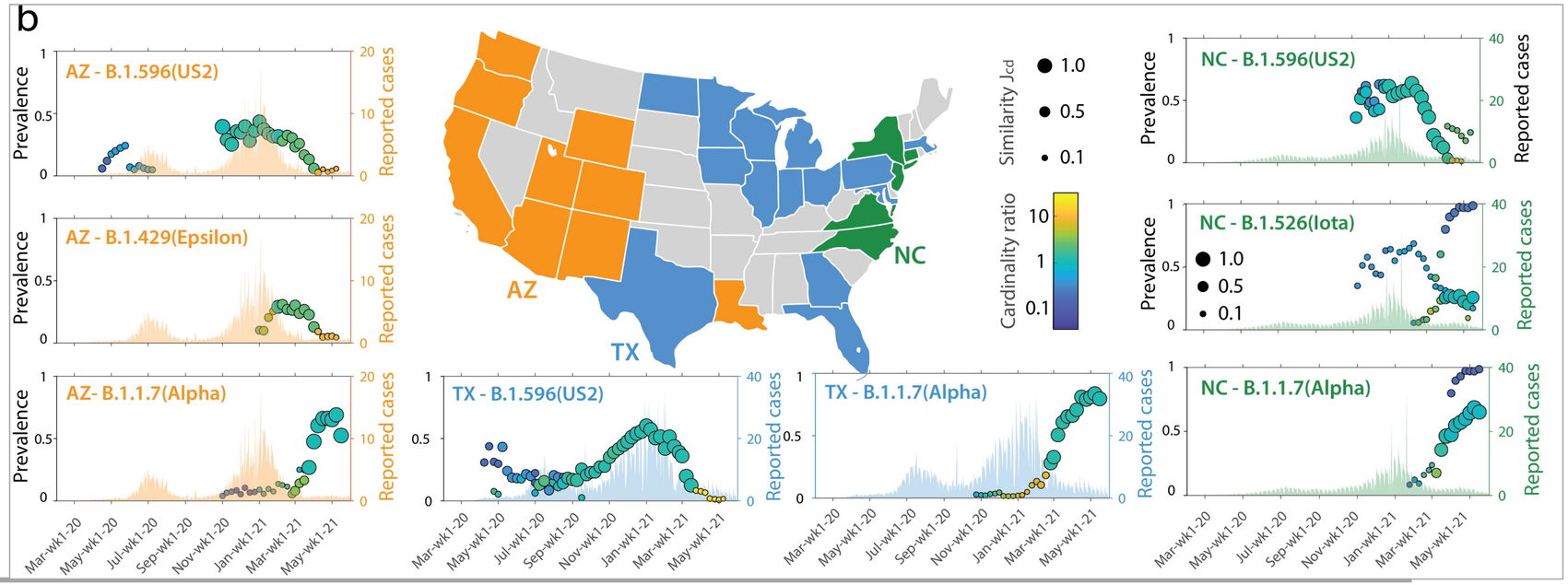
- Assign *a posteriori* the observed change time-series to known lineages, using similarity between early warning clusters and the lineage dictionaries (Jaccard similarity index Jcd, cluster vs. dictionary change composition)
- Apply a threshold Jcd>0.5 to identify clusters for which there is a close compositional match to a known lineage



Europe and US variants

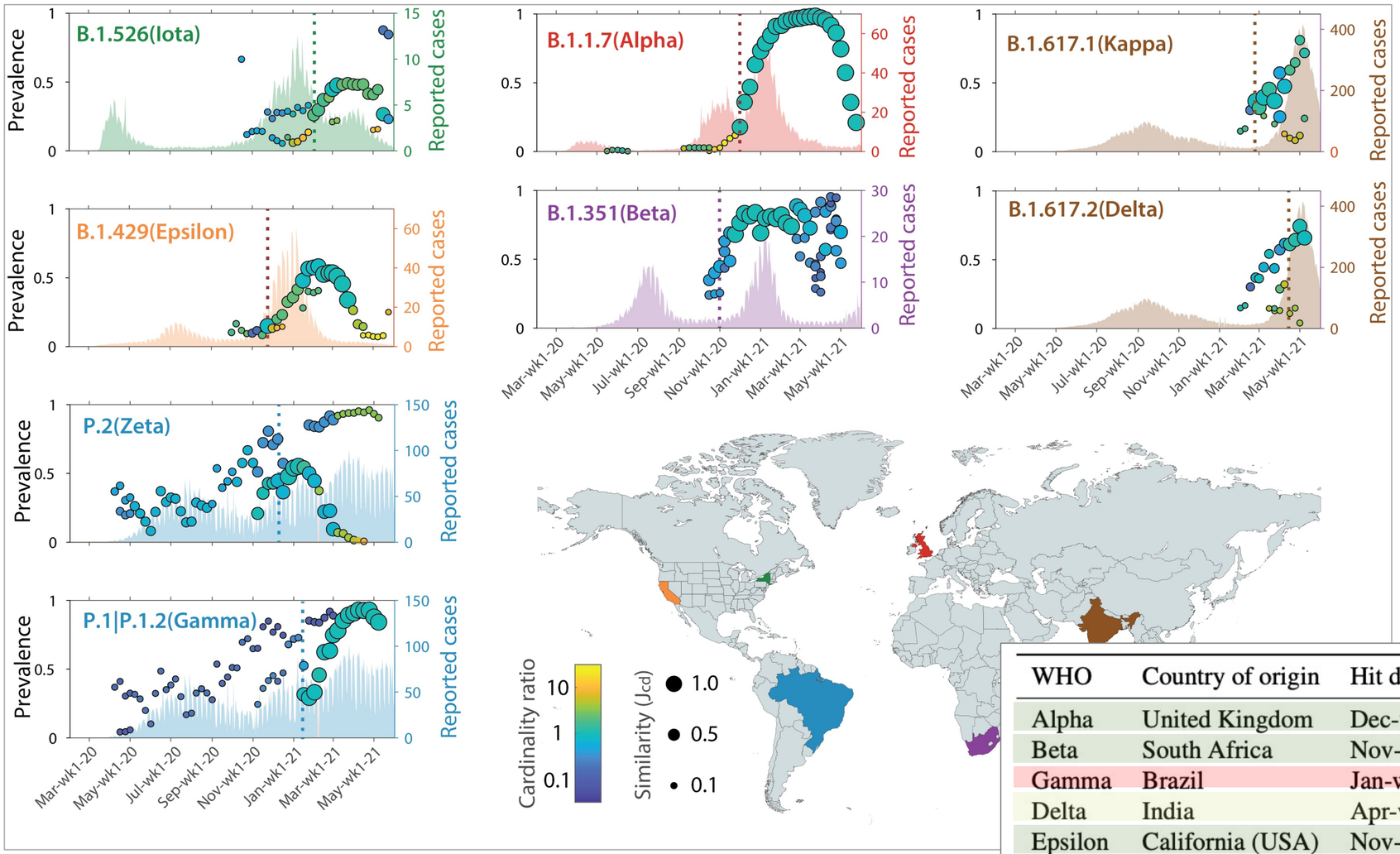


Temporal dynamics of the Alpha variant in European countries



Temporal dynamics of notable variants in the US

Assessment of the early warning system

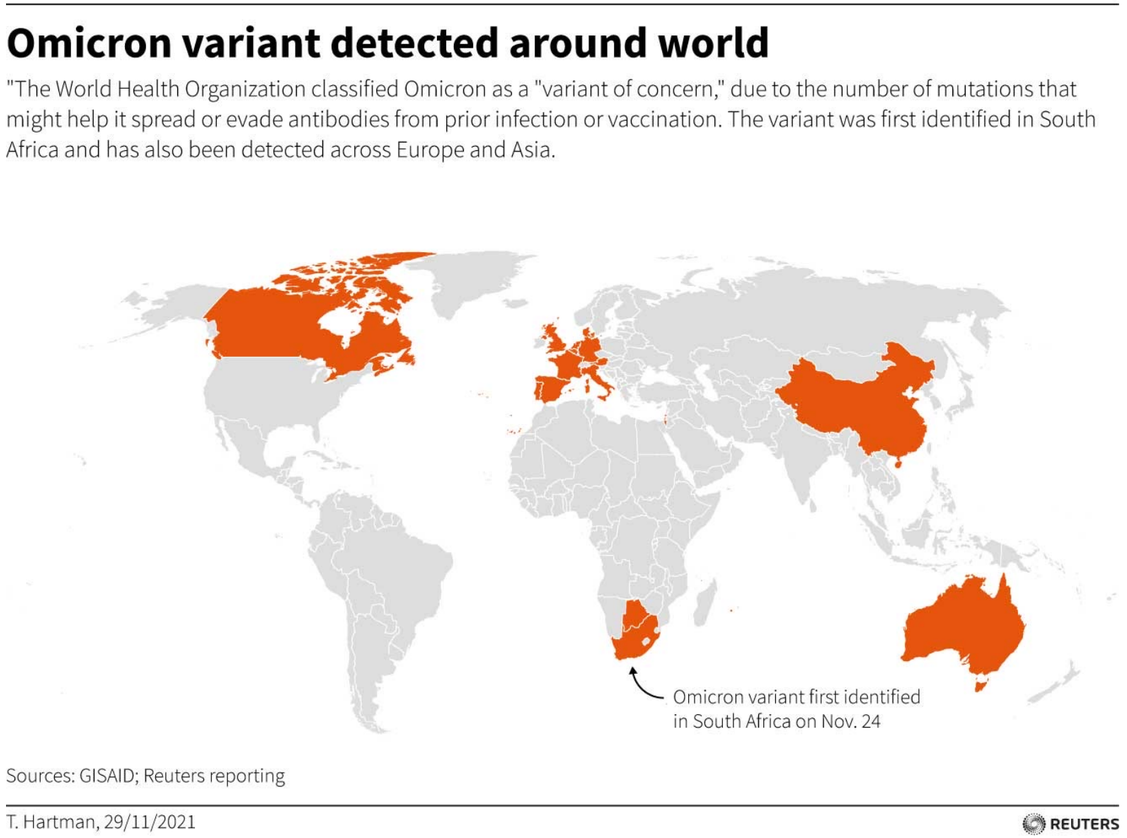


Emergence of variants in their country of origin.

WHO	Country of origin	Hit date	Early hit	1st communication	Comparison (weeks)
Alpha	United Kingdom	Dec-wk1-20	✓	Dec-wk3-20	2 weeks earlier
Beta	South Africa	Nov-wk1-20	× (2 weeks)	Dec-wk3-20	6 weeks earlier
Gamma	Brazil	Jan-wk3-21	✓	Jan-wk1-21	2 weeks later
Delta	India	Apr-wk3-21	✓	Apr-wk3-21	same time
Epsilon	California (USA)	Nov-wk4-20	✓	Jan-wk3-21	7 weeks earlier
Zeta	Brazil	Dec-wk2-20	✓	Dec-wk4-20	2 weeks earlier
Iota	New York (USA)	Feb-wk2-21	× (1 week)	Feb-wk2-21	same time
Kappa	India	Feb-wk4-21	✓	Mar-wk4-21	4 weeks earlier

Ongoing work

The Omicron case



Omicron (21K or B.1.1.529) is primarily of concern due to the large number of mutations on the Spike gene. Many of these variants are in the receptor binding domain and N-terminal domain, and thus may play key roles in ACE2 binding and antibody recognition.

Omicron mutations impacts

Report on Omicron Spike mutations on epitopes and immunological/epidemiological/kinetics effects from literature

SARS-CoV-2 coronavirus | nCoV-2019 Genomic Epidemiology

sunbrn 26d

Report on Omicron Spike mutations on epitopes and immunological/epidemiological/kinetics effects from literature

Authors: Anna Bernasconi*, Pietro Pinoli*, Ruba Al Khalaf, Tommaso Alfonsi, Arif Canakoglu, Luca Cilibrasi, and Stefano Ceri
Affiliation: Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Milan, Italy
Notes: * The first two authors contributed equally to the work

The B.1.1.529 variant has entered abruptly the landscape of SARS-CoV-2 variants at the end of November 2021; it was reported to WHO on November 24th, named Omicron and considered a Variant of Concern, following the advice of the Technical Advisory Group on SARS-CoV-2 Virus Evolution (TAG-VE). We report here the collection of our findings, following a long-standing effort of our group dedicated to SARS-CoV-2 sequences (involving data integration pipelines, search and knowledge management systems, see http://www.bioinformatics.deib.polimi.it/geco/?virus 118.

We collected Omicron Spike mutations of interest from ECDC (https://www.ecdc.europa.eu/en/covid-19/variants-concern 124) and compared them to CoVariants ones (https://covariants.org/variants/21K.Omicron 104). The provided lists only differ for the representation of the following deletions/insertions:

- G142D, Δ143-145 (ECDC) becomes G142-, V143-, Y144-, Y145D (CoVariants);
- Δ211-212, ins214EPE (ECDC) becomes N211- (CoVariants).

Nov 30

1 / 1
Nov 30

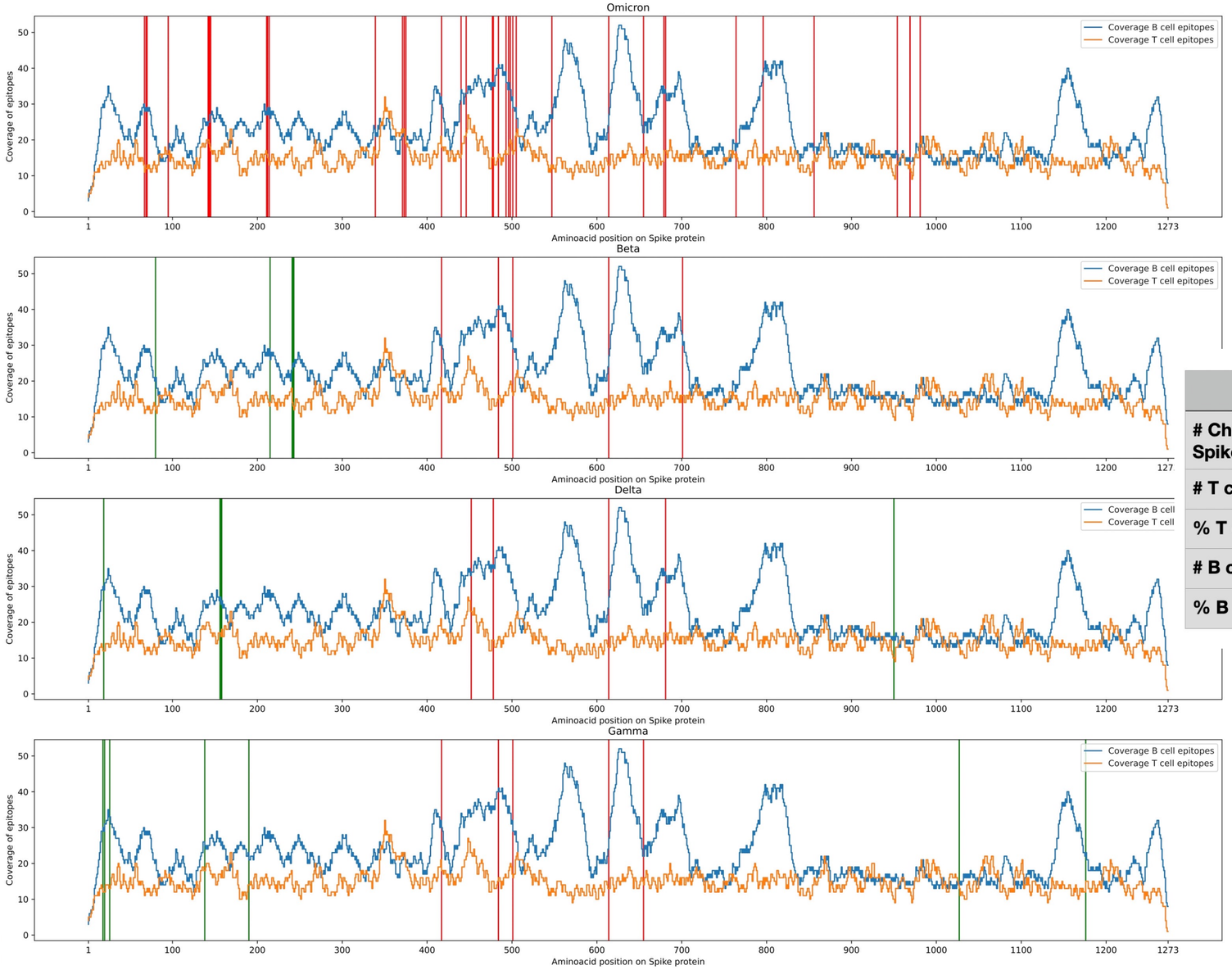
26d ago

The WHO declared Omicron a VoC on Nov 26th

On Nov 30th, we made the 1st post on Omicron on the Virological.org web blog, reaching 13.8k views (as of Jan 14th, 2022)

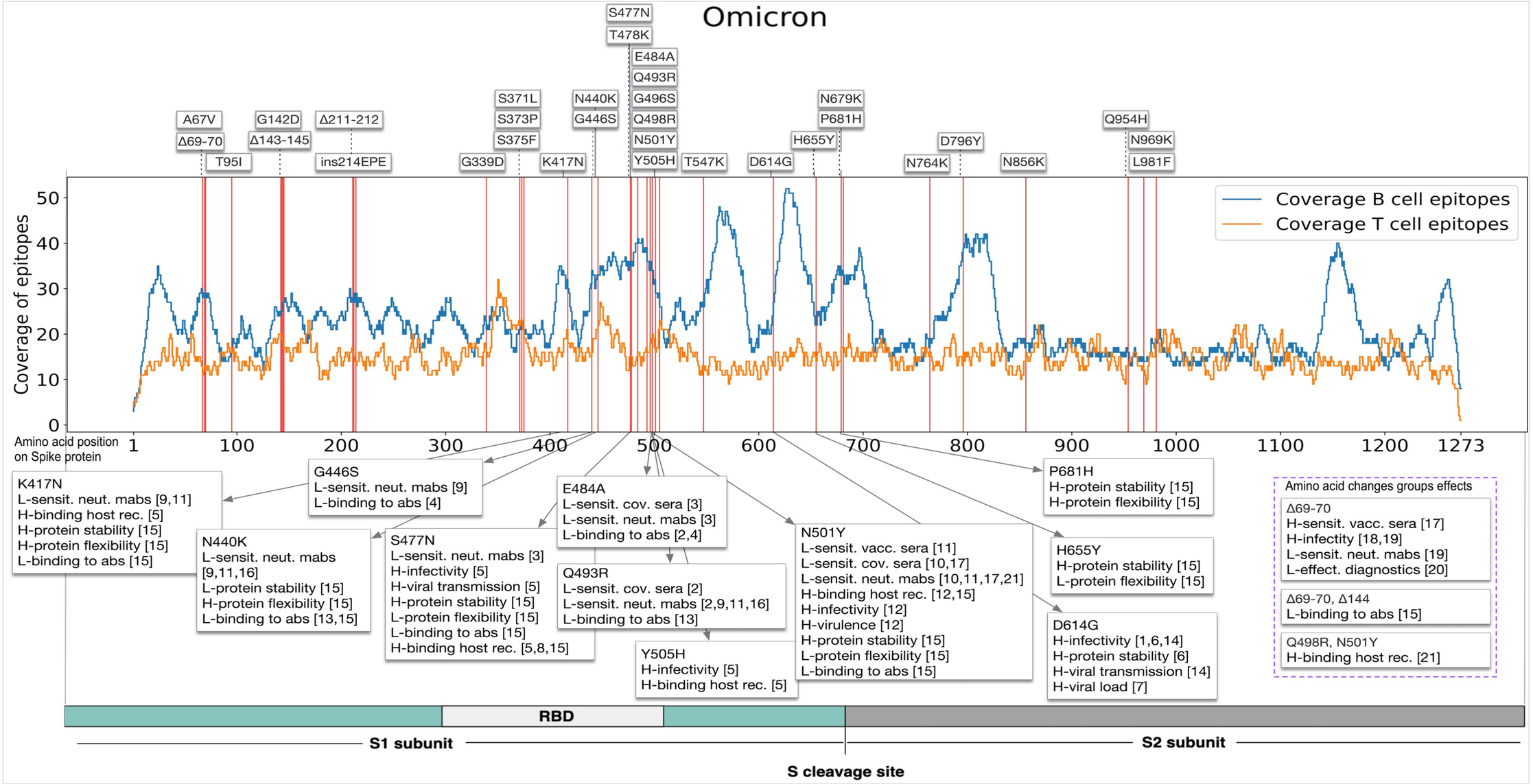
https://virological.org/t/report-on-omicron-spike-mutations-on-epitopes-and-immunological-epidemiological-kinetics-effects-from-literature/770

Impact on in-use epitopes – comparison with other variants



	Omicron	Beta	Gamma	Delta
# Changes on Spike	37	10	12	9
# T cell Epitopes	348	125	159	108
% T cell Epitopes	27.29%	9.80%	12.47%	8.47%
# B cell Epitopes	550	231	273	198
% B cell Epitopes	30.91%	12.98%	15.34%	11.12%

Single and group mutations effects (according to CoV2K)



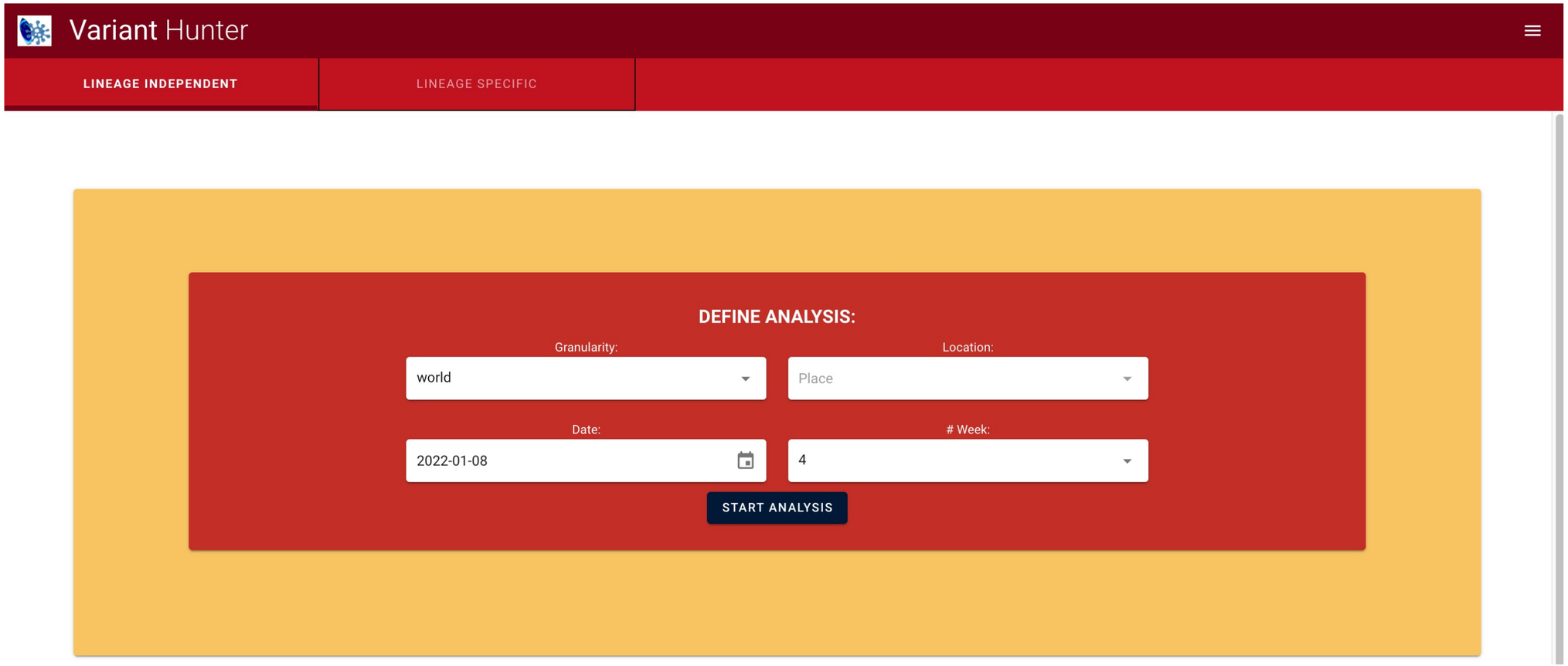
Effects definitions: https://github.com/DEIB-GECO/cov2k_data_collector/blob/master/CoV2K_Effects_Taxonomy.pdf

VariantHunter

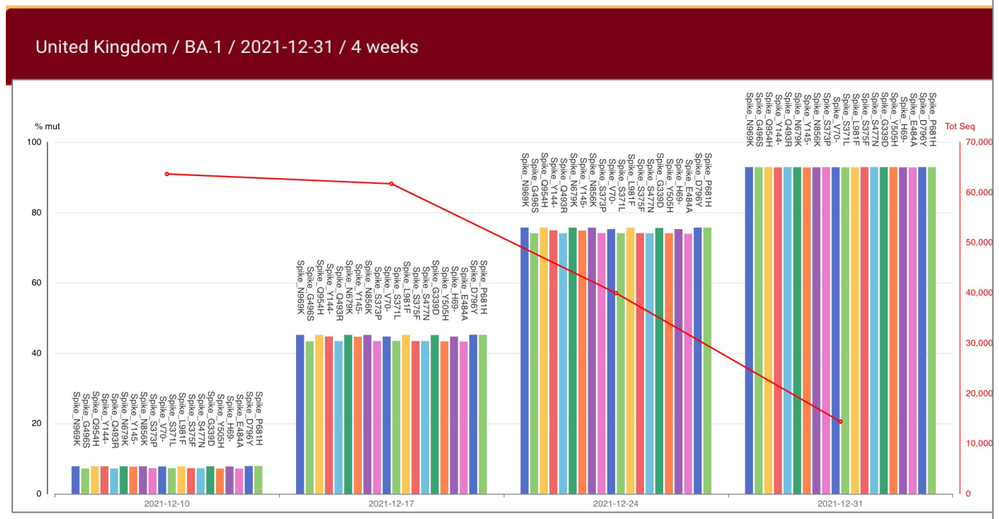
https://github.com/DEIB-GECO/Docker_VariantHunter

Supporting two working modes:

- Lineage **independent** mutation growth
- Lineage **dependent** mutation growth



Relevant Spike mutations in United Kingdom at the end of 2021



	2021-12-10	2021-12-17	2021-12-24	2021-12-31
Spike_G339D	7.80	45.0	76.0	93.0
Spike_N969K	7.80	45.0	76.0	93.0
Spike_G496S	7.20	43.0	74.0	93.0
Spike_Q954H	7.80	45.0	76.0	93.0
Spike_Y144-	7.80	45.0	75.0	93.0
Spike_Q493R	7.20	43.0	74.0	93.0
Spike_N679K	7.80	45.0	76.0	93.0
Spike_Y145-	7.70	45.0	75.0	93.0
Spike_N856K	7.80	45.0	76.0	93.0
Spike_S373P	7.30	43.0	74.0	93.0
Spike_V70-	7.80	45.0	75.0	93.0
Spike_S371L	7.30	43.0	74.0	93.0
Spike_L981F	7.80	45.0	76.0	93.0
Spike_S375F	7.30	43.0	74.0	93.0
Spike_S477N	7.20	43.0	74.0	93.0
Spike_Y505H	7.20	43.0	74.0	93.0
Spike_H69-	7.80	45.0	75.0	93.0
Spike_E484A	7.20	43.0	74.0	93.0
Spike_D796Y	7.90	45.0	76.0	93.0
Spike_P681H	7.90	45.0	76.0	93.0
Spike_T547K	7.80	45.0	76.0	93.0
Spike_L212I	7.20	43.0	74.0	93.0
Spike_Q498R	7.20	43.0	74.0	93.0
Spike_N501Y	7.20	43.0	74.0	93.0
Spike_N211-	7.20	43.0	74.0	93.0
Spike_H655Y	8.10	45.0	76.0	93.0
Spike_A67V	8.10	45.0	76.0	93.0
Spike_V143-	7.70	45.0	74.0	93.0
Spike_N764K	5.30	18.0	32.0	56.0
Spike_A701V	2.40	12.0	19.0	24.0
Spike_T95I	79.0	88.0	95.0	99.0
Spike_R346K		4.00	8.00	14.0
Spike_I1081V		4.80	7.40	8.20
Spike_G142D	94.0	96.0	97.0	100
Spike_T478K	99.0	98.0	98.0	100
Spike_D614G	100	100	100	100
Spike_V36F	4.30	2.60	1.20	
Spike_V1264L	7.10	4.90	2.50	
Spike_Y145H	25.0	16.0	7.80	2.30
Spike_A222V	25.0	16.0	8.00	2.30
Spike_E156G	90.0	53.0	24.0	7.00
Spike_F157-	90.0	53.0	24.0	7.00
Spike_L452R	92.0	55.0	24.0	7.20
Spike_T19R	92.0	54.0	24.0	7.00
Spike_D950N	91.0	54.0	24.0	7.00
Spike_R158-	90.0	53.0	24.0	7.00
Spike_P681R	92.0	55.0	24.0	7.00

Spike mutations in United Kingdom compared to Omicron (BA.1)

United Kingdom / BA.1 / 2021-12-31 / 4 weeks

TABLE

Location	Lineage	Protein	Mut	Slope	Y-intercept	P-value with mut	P-value without mut	P-value comparative ↑ 1
United Kingdom	BA.1	Spike	N764K	-1.9	55	0.0	0.0	0.0
United Kingdom	BA.1	Spike	R346K	2.6	6.3	0.0	0.0	1.2e-92
United Kingdom	BA.1	Spike	I1081V	-0.47	11	0.0	0.0	5.2e-12
United Kingdom	BA.1	Spike	A701V	-0.67	27	0.0	0.0	0.000019

Rows per page: 10 1-4 of 4

The observations of the third p-value and of the heatmap indicate that the Omicron lineage with the additional Spike mutation R346K is growing significantly faster than Omicron alone

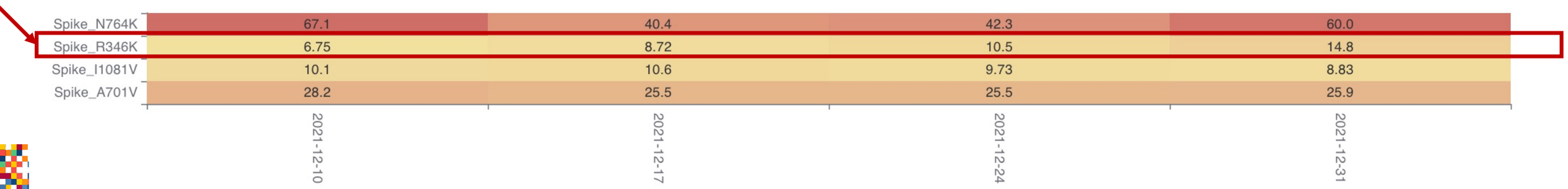
Spike_R346K mutation is currently under scrutiny.. … but so far its addition to Omicron did not affect the neutralization susceptibility

> Clin Infect Dis. 2021 Dec 16;ciab1041. doi: 10.1093/cid/ciab1041. Online ahead of print.

Neutralization of SARS-CoV-2 Omicron variant by sera from BNT162b2 or Coronavac vaccine recipients

Lu Lu [1], Bobo Wing-Yee Mok [1], Lin-Lei Chen [1], Jacky Man-Chun Chan [2], Owen Tak-Yin Tsang [2], Bosco Hoi-Shiu Lam [3], Vivien Wai-Man Chuang [4], Allen Wing-Ho Chu [1], Wan-Mui Chan [1], Jonathan Daniel Ip [1], Brian Pui-Chun Chan [1], Ruiqi Zhang [5], Cyril Chik-Yan Yip [1 6], Vincent Chi-Chung Cheng [1 6], Kwok-Hung Chan [1], Dong-Yan Jin [7], Ivan Fan-Ngai Hung [5], Kwok-Yung Yuen [1 6], Honglin Chen [1], Kelvin Kai-Wang To [1 6]

P value with mut: shows if the population «lineage + mutation» is growing differently compared to everything else
P value without mut: shows if the population «lineage without mutation» is growing differently compared to everything else
P value comparative: shows if the population «lineage + mutation» is growing differently compared to the population «lineage without mutation»



The IHU variant from Cameroon

AFRICA TIMES

POLITICS & POLICY • LEADERS & COMPANIES • FINANCE & ECON
SECURITY • HEALTH • CUTTING EDGE • YOUNG VOICES • AFR
News Articles Commentary Blogs In Depth

HOME > HEALTH > **FRENCH SCIENTISTS REPORT NEW COVID VARIANT TRACED TO CAMEROON**

French scientists report new COVID variant traced to Cameroon

By AT editor - 3 January 2022 at 9:07 pm

06 Jan

Covid-19: France detects new variant called IHU in man who travelled to Cameroon

news24 Lenin Ndebele SHARE

Comments (5)

Emergence in Southern France of a new SARS-CoV-2 variant of probably Cameroonian origin harbouring both substitutions N501Y and E484K in the spike protein

Philippe Colson, Jérémy Delerce, Emilie Burel, Jordan Dahan, Agnès Jouffret, Florence Fenollar, Nouara Yahi, Jacques Fantini, Bernard La Scola, Didier Raoult

doi: https://doi.org/10.1101/2021.12.24.21268174

This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should *not* be used to guide clinical practice.

Abstract Full Text Info/History Metrics Preview PDF

ABSTRACT

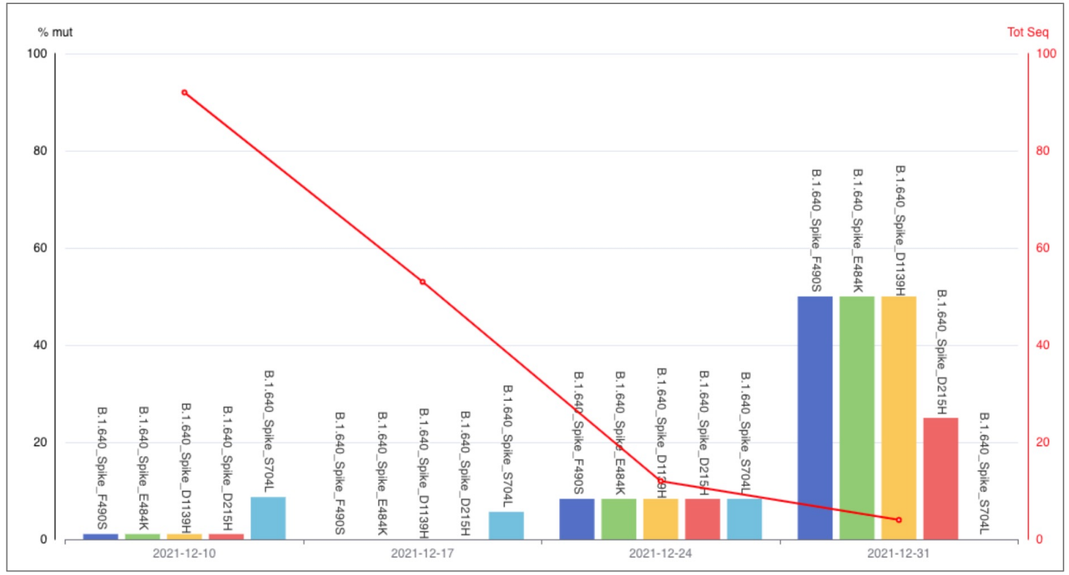
SARS-CoV-2 variants have become a major virological, epidemiological and clinical concern, particularly with regard to the risk of escape from vaccine-induced immunity. Here we describe the emergence of a new variant. For twelve SARS-CoV-positive patients living in the same geographical area of southeastern France, qPCR testing that screen for variant-associated mutations showed an atypical combination. The index case returned from a travel in Cameroon. The genomes were obtained by next-generation sequencing with Oxford Nanopore Technologies on GridION instruments within ≈8 h. Their analysis revealed 46 mutations and 37 deletions resulting in 30 amino acid substitutions and 12 deletions. Fourteen amino acid substitutions, including N501Y and E484K, and 9 deletions are located in the spike protein. This genotype pattern led to create a new Pangolin lineage named B.1.640.2, which is a phylogenetic sister group to the old B.1.640 lineage renamed B.1.640.1. Both lineages differ by

B.1.640 mutations «run faster» and create the B.1.640.2 lineage

/ B.1.640 / 2021-12-31 / 4 weeks

TABLE

Location	Lineage	Protein	Mut	Slope	Y-intercept	P-value with mut	P-value without mut	P-value comparative ↑ 1
World	B.1.640	Spike	F490S	15	-8.0	0.064	3.6e-9	0.000072
World	B.1.640	Spike	E484K	15	-8.0	0.064	3.6e-9	0.000072
World	B.1.640	Spike	D1139H	15	-8.0	0.064	3.6e-9	0.000072
World	B.1.640	Spike	D215H	7.9	-3.0	0.41	7.8e-9	0.047
World	B.1.640	Spike	S704L	-2.0	9.0	0.27	9.2e-7	0.93



B.1.640.1 and B.1.640.2 lineages

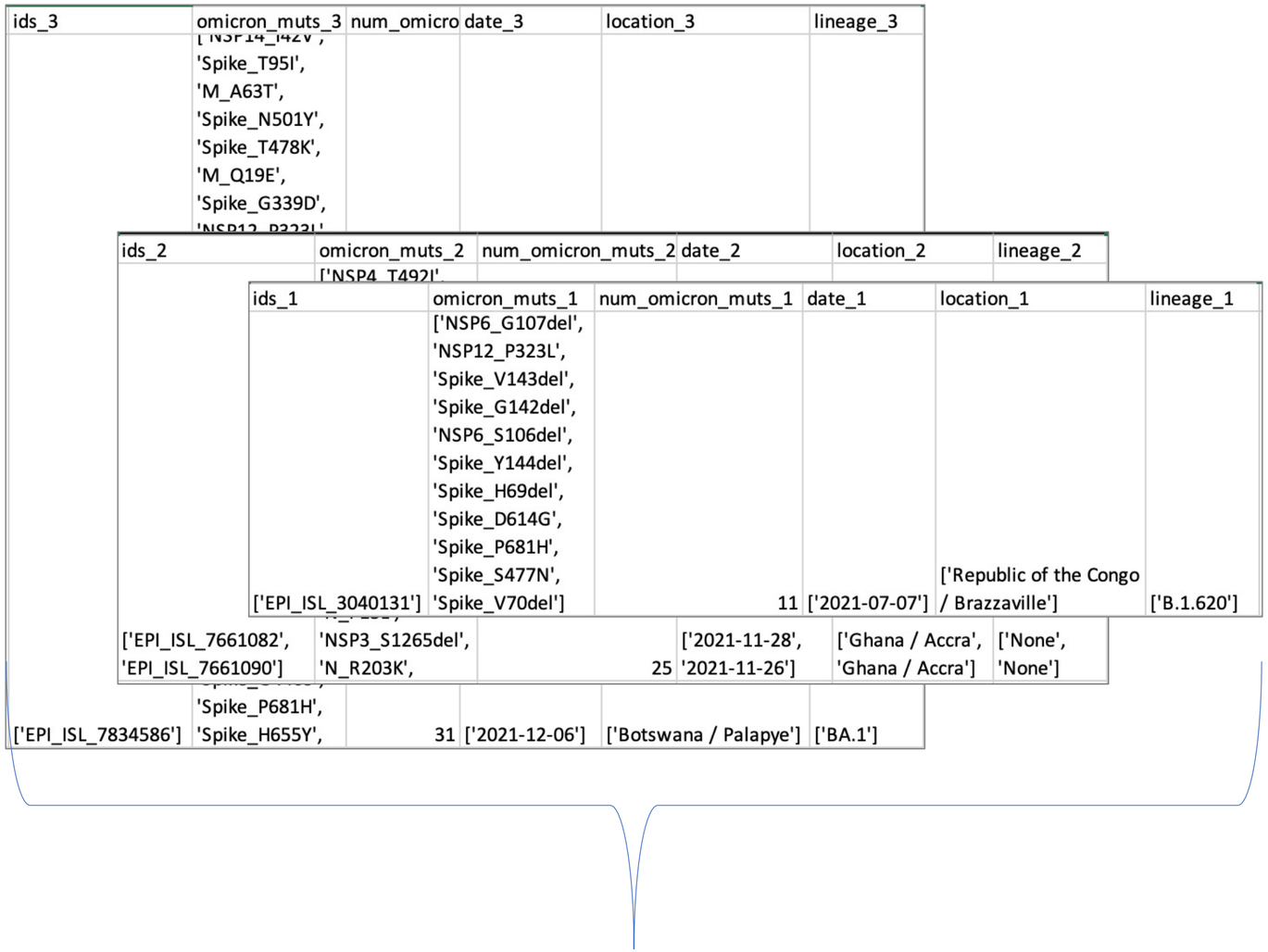
B.1.640.1	B.1.640.2 (IHU variant)
S:P9L	S:P9L
S:E96Q	S:E96Q
Deletions S:C136-, S:N137-, S:D138-, S:P139-, S:F140-, S:L141-, S:G142-, S:V143-, S:Y144-	Deletions S:C136-, S:N137-, S:D138-, S:P139-, S:F140-, S:L141-, S:G142-, S:V143-, S:Y144-
S:R190S	S:R190S
S:I210T	-
-	S:D215H
S:R346S	S:R346S
S:N394S	S:N394S
S:Y449N	S:Y449N
-	S:E484K
S:F490R	S:F490S
S:N501Y	S:N501Y
S:D614G	S:D614G
S:P681H	S:P681H
S:T859N	S:T859N
S:D936H	-
-	S:D1139H

Omicron as a result of other lineages' recombination

Hypothesis:
Omicron is a recombinant set of variants that have evolved over many months

Data extraction method:
Compute table of triplets of sequence groups (identified by a set of amino acid changes) that are «candidate recombinants»
Desirable properties of triplets:
- Minimal intersection
- Cardinality of their union ~ cardinality of Omicron characterizing changes (61 changes)
- Tentative: balanced cardinalities (i.e., three groups with 20/20/20 changes are preferrable than 50/5/5)

To be verified by means of:
- Phylogenetic analysis
- Alignment checks
- Location/Date coherency



Example three groups of sequences whose union reaches 57 Omicron characterizing mutations (their intersection has only 8)

Summary

A data-driven approach has been successful for analyzing SARS-CoV-2 virus characteristics and for gaining insights on its evolution.

The objectives of our group are:
- Developing methods and tools dedicated to generic users
- Using them for our own analyses (variants birth identification, Omicron mutations impact, …)

Ongoing work:
- A curated knowledge graph of SARS-CoV-2 and its interactions with the host
- An early warning system for variants
- Lineages' evolution prediction based on established co-occurring pairs of mutations
- Conceptual model-driven comparison between the genomics of humans and of other species, including viruses (on site collaboration with UPV)

High-level goals:
- Making the pandemic explainable by exposing viral data and knowledge
- Defining and resolve new computational challenges in viral genomics, taking advantage of Data Science methods

Special thanks

ARIF CANAKOGLU

RENATO CASAGRANDI

STEFANO CERI

MATTEO CHIARA
(UniMi)

LORENZO MARI

PIETRO PINOLI

TOMMASO ALFONSI

RUBA AL KHALAF

LUCA CILIBRASI

ANDREA GULINO

FRANCESCO INVERNICI

GIUSEPPE SERNA