Big data in life sciences

from theory to applications

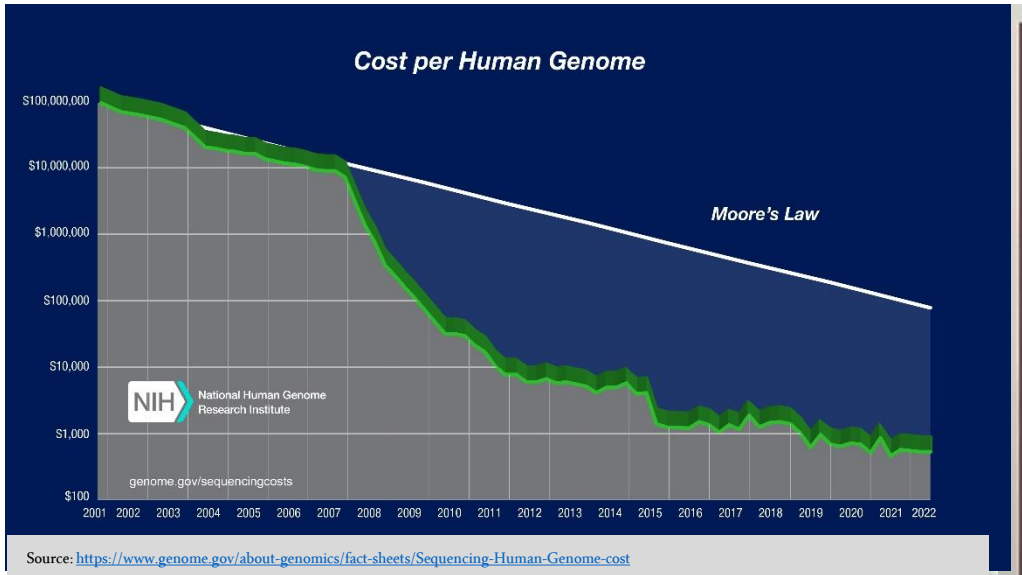
Stefano Ceri

DEIB | Dipartimento di Elettronica, Informazione e Bioingegneria

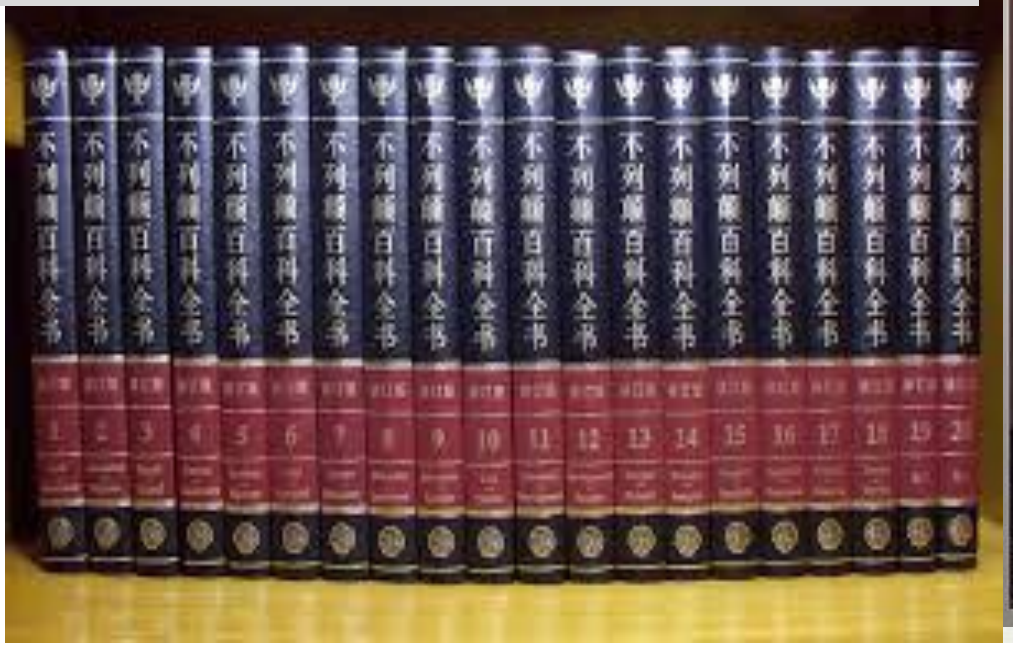
https://ceri.faculty.polimi.it/

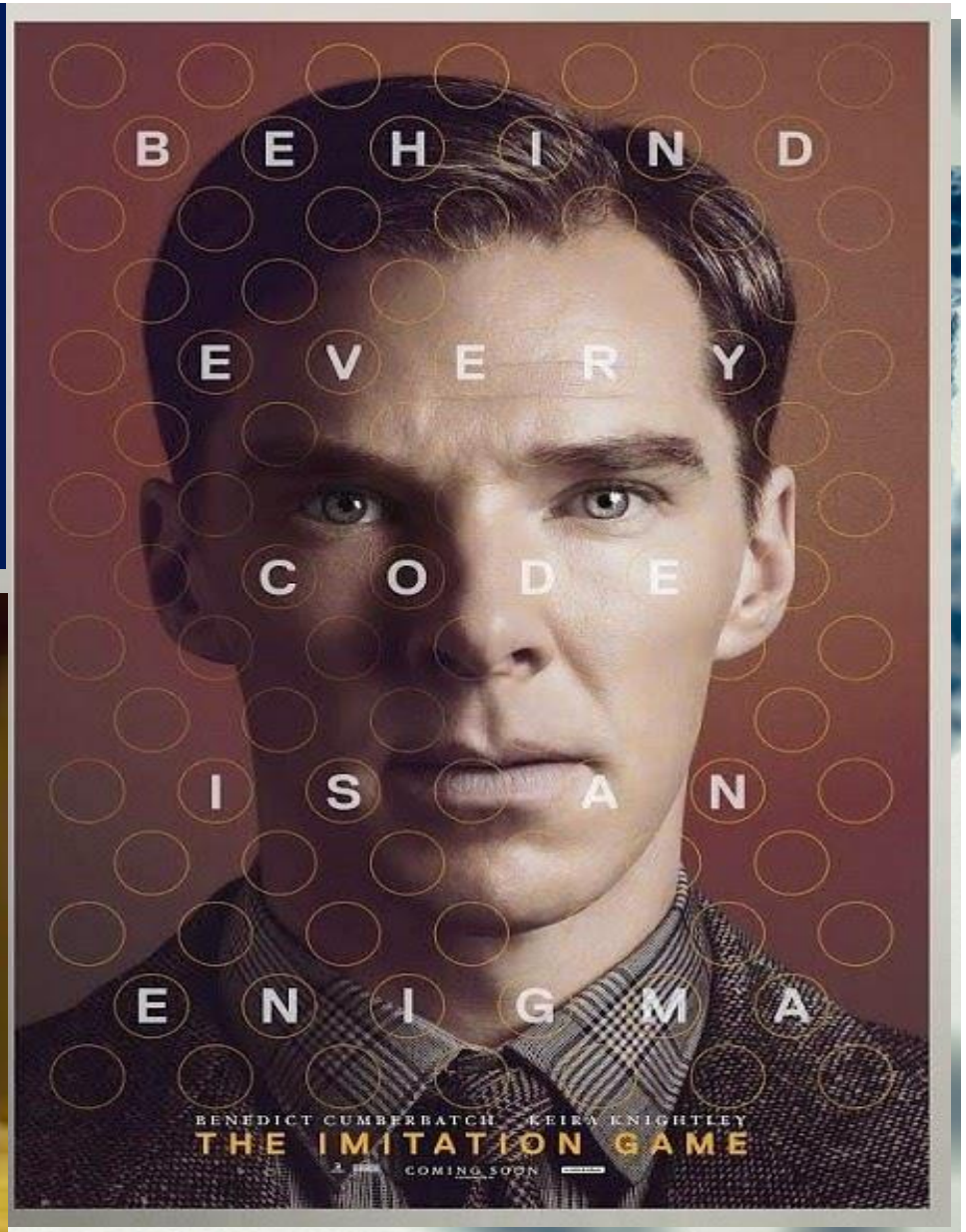
POLITECNICO
MILANO 1863

The big data challenge --- human genomics



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





Data-Driven Genomic Computing (GeCo)

ERC Advanced Grant, 2016-2021

Focus: Data design, integration, extraction and analysis for genomic data

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

Objectives: Open-source systems for genomic data management

Results: Big data management systems + integrated repositories, both in human and viral genomics, demonstrated through biological and clinical research.

Human vs viral genomic research

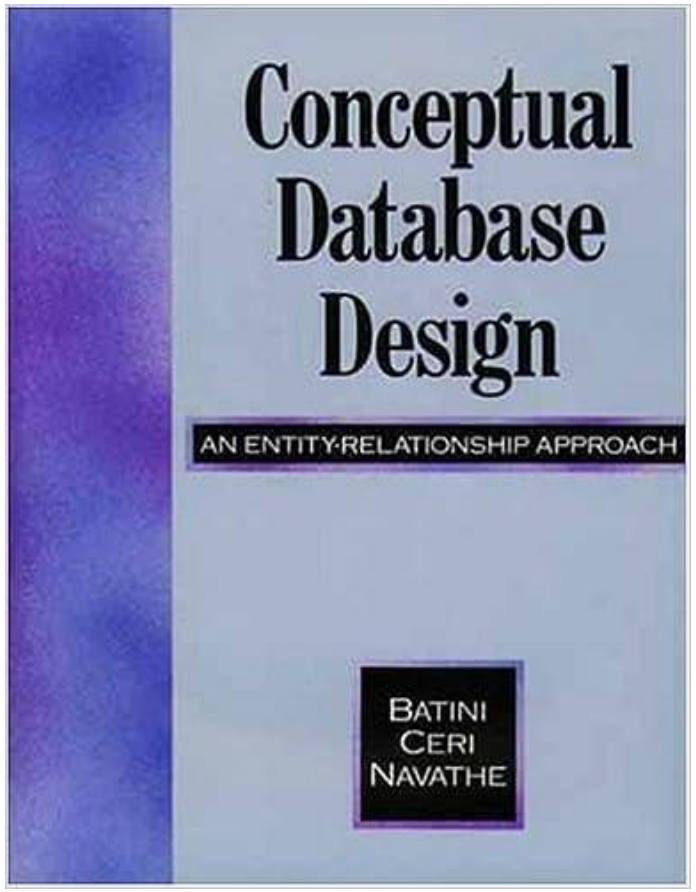
- **HUMAN GENOME: 3 Billlion DNA base pairs, about 20K genes**
  - Data: mutations, expression, methylation, 3D contacts, …
  - Biga data approach: design, integrate, search, analyze
  - Most relevant clinical applications: rare diseases and cancer – but genetics produces a huge number of human traits
  - Research highlights: basic science for genomics, disease prevention & treatment, personalization, drug repurposing.
- **SARS-CoV-2 VIRAL GENOME: 3OK RNA nucleotides, 12 genes**
  - Data: viral sequences and their nucleotide mutations
  - Reuse our big-data approach, applied to a "smaller" problem
  - Research driven by pandemic emergency: Understand genome evolution (mutations, variants), their mechanisms (co-evolution, recombination – now reassortment for influenza) and threats to humans, e.g. effects on COVID-19 disease spreading and severity.

DATA DESIGN

Design objectives & challenges

- TOP-DOWN: Clean design, with nice abstractions
- BOTTOM-UP: Matching with existing data, providing content

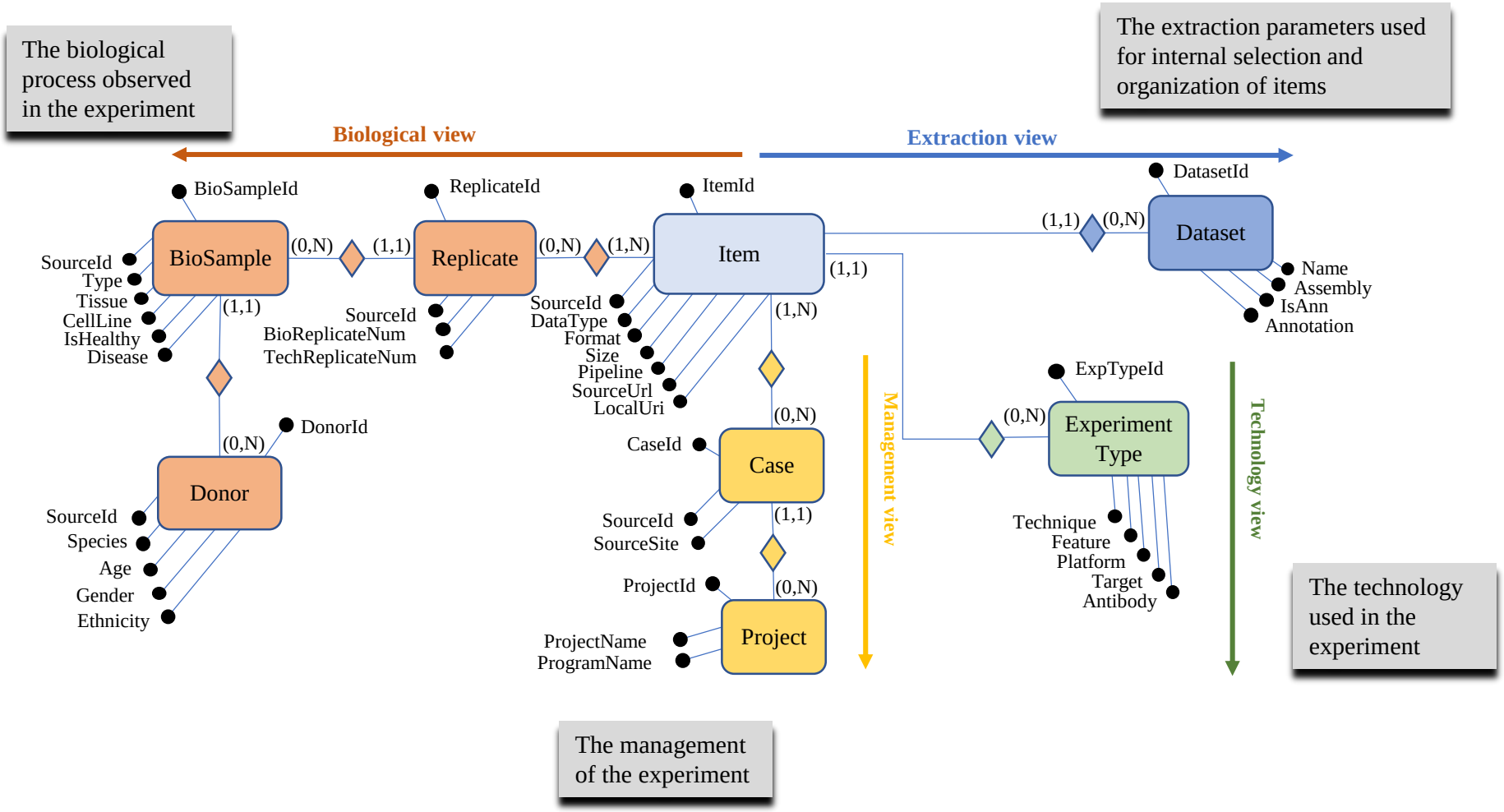
- Model: Entity-Relationship (P. Chen)



HUMAN GENOMICS:

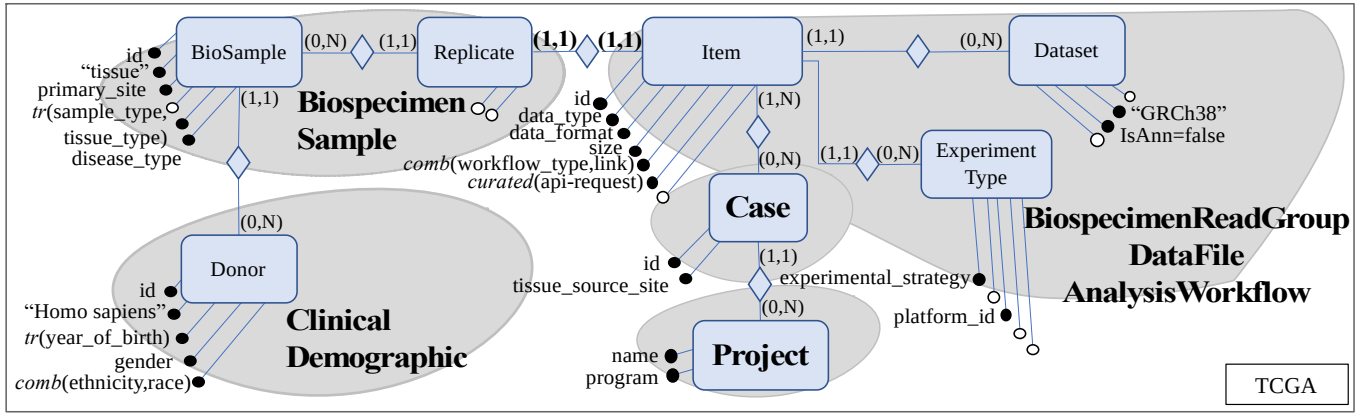
DATA DESIGN
DATA INTEGRATION
TOOLS

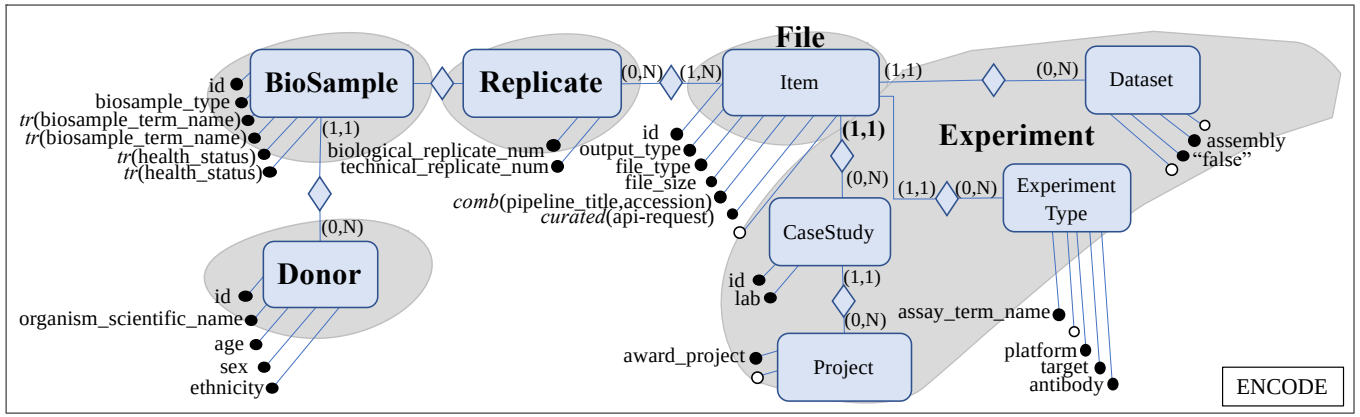
Modeling Genomic Metadata: The Genomic Conceptual Model

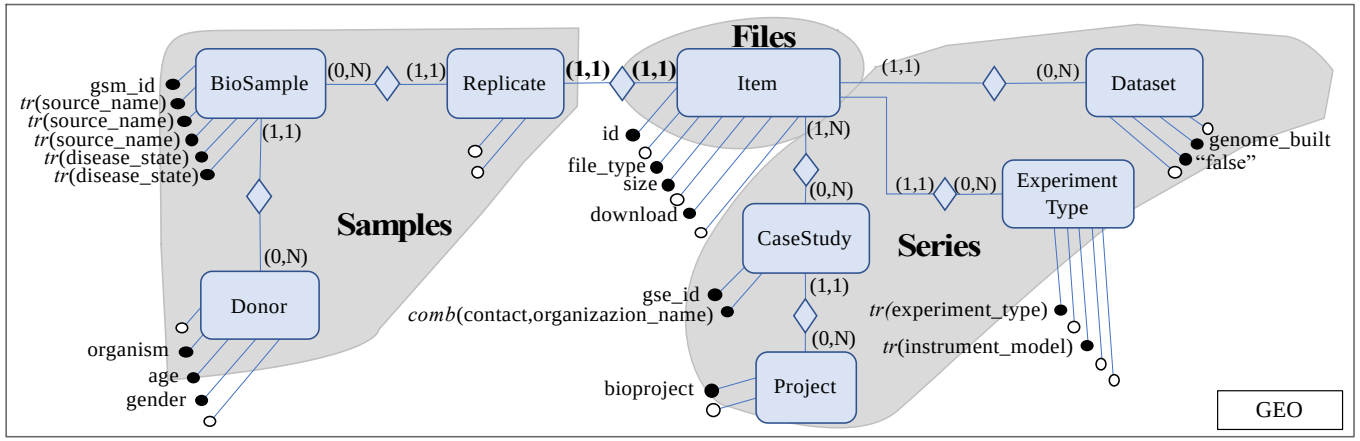


Bernasconi, Ceri, Campi, Masseroli (2017) Conceptual Modeling for Genomics: Building an Integrated Repository of Open Data. In: Mayr H., Guizzardi G., Ma H., Pastor O. (eds) Conceptual Modeling. ER 2017. Lecture Notes in Computer Science, vol 10650. Springer, Cham. https://doi.org/10.1007/978-3-319-69904-2_26

Sources elements are mapped into the global schema





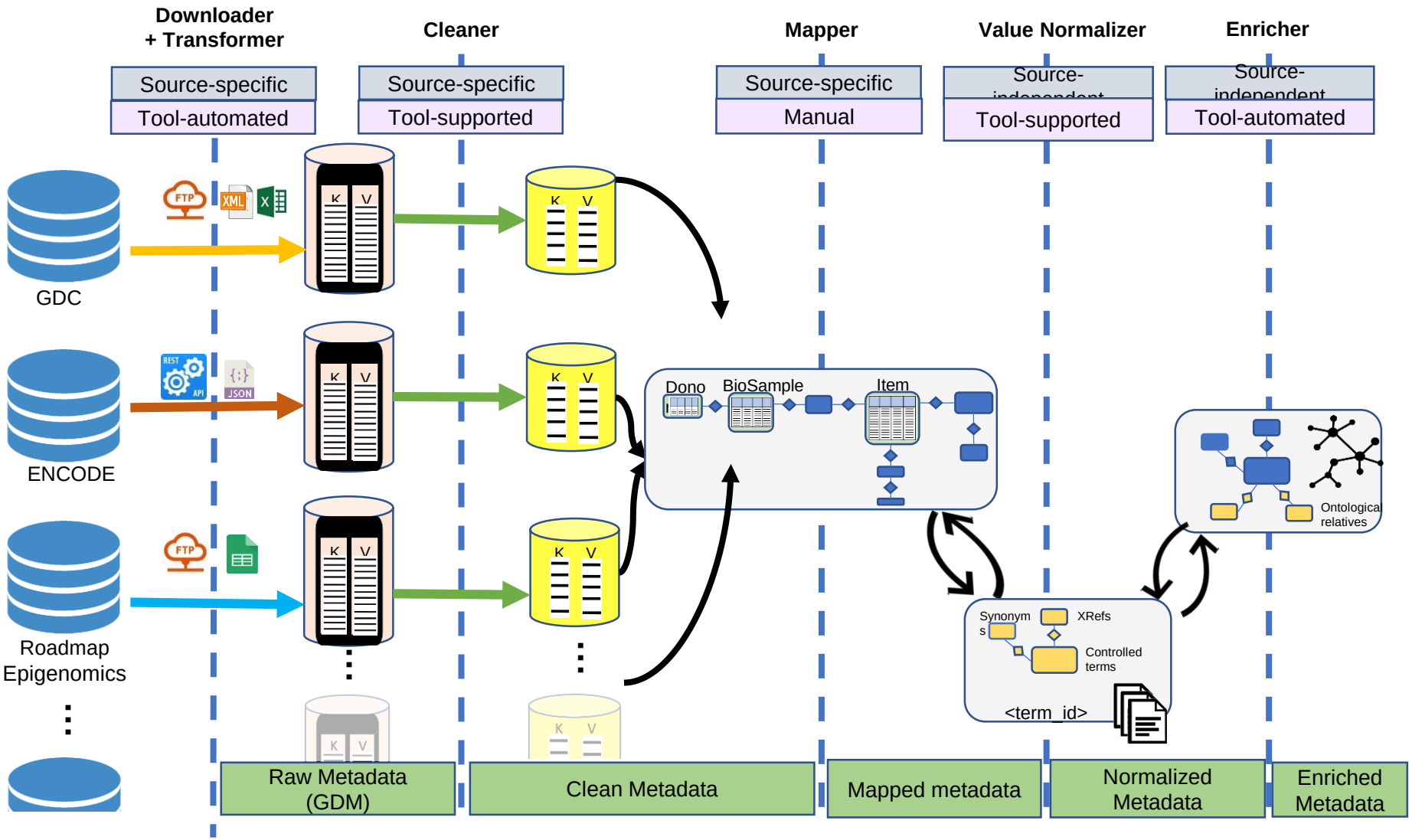


THE CANCER GENOME ATLAS

ENCODE

GEO
Gene Expression Omnibus

META-BASE: phases of integration



Bernasconi et al. (2020) META-BASE: a Novel Architecture for Large-Scale Genomic Metadata Integratio. In *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, https://doi.org/10.1109/TCBB.2020.2998954

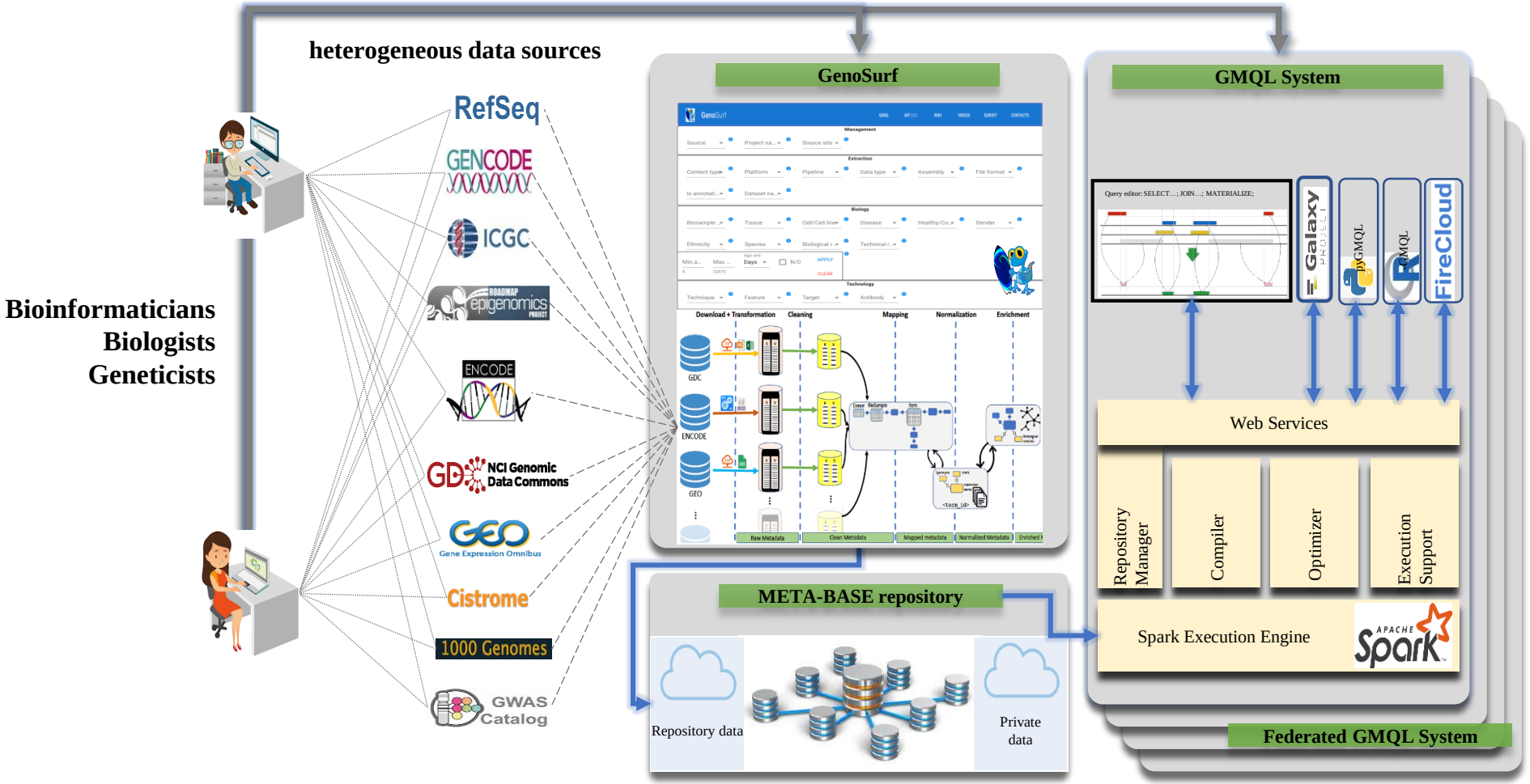
REPOSITORY

Consortium	Imported datasets	# of samples	File size (MB)
ENCODE	GRCh38_ENCODE_BROAD	850	6,869
	GRCh38_ENCODE_NARROW	11,573	128,316
	HG19_ENCODE_BROAD	844	18,382
	HG19_ENCODE_NARROW	10,342	111,925
ROADMAP EPIGENOMICS	HG19_ROADMAP_EPIGENOMICS_BED	156	968
	HG19_ROADMAP_EPIGENOMICS_BROAD	979	24,332
	HG19_ROADMAP_EPIGENOMICS_DMR	66	3,060
	HG19_ROADMAP_EPIGENOMICS_GAPPED	979	6,875
	HG19_ROADMAP_EPIGENOMICS_NARROW	1,032	11,788
	HG19_ROADMAP_EPIGENOMICS_RNA_expression	399	2,453
TCGA	HG19_TCGA_cnv	22,632	797
	HG19_TCGA_dnamethylation	12,860	247,742
	HG19_TCGA_dnaseq	6,914	286
	HG19_TCGA_mirnaseq_isoform	9,909	4,207
	HG19_TCGA_mirnaseq_mirna	9,909	746
	HG19_TCGA_rnaseq_exon	3,675	47,668
	HG19_TCGA_rnaseq_gene	3,675	5,327
	HG19_TCGA_rnaseq_spljxn	3,675	44,377
	HG19_TCGA_rnaseqv2_exon	9,825	124,343
	HG19_TCGA_rnaseqv2_gene	9,825	21,862
	HG19_TCGA_rnaseqv2_isoform	9,825	53,082
	HG19_TCGA_rnaseqv2_spljxn	9,825	115,088
GDC - TCGA	GRCh38_TCGA_copy_number	22,374	686
	GRCh38_TCGA_copy_number_masked	22,375	337
	GRCh38_TCGA_gene_expression	11,091	56,542
	GRCh38_TCGA_methylation	12,218	1,348,516
	GRCh38_TCGA_miRNA_expression	10,947	1,502
	GRCh38_TCGA_miRNA_isoform_expression	10,999	5,004
	GRCh38_TCGA_somatic_mutation_masked	10,188	2,280
GENCODE	GRCh38_ANNOTATION_GENCODE	24	1,798
	HG19_ANNOTATION_GENCODE	20	1,324
REFSEQ	GRCh38_ANNOTATION_REFSEQ	31	740
	HG19_ANNOTATION_REFSEQ	30	275
…	…	…	…
Grand total	**67 datasets**	**564,422**	**9.23 TB**

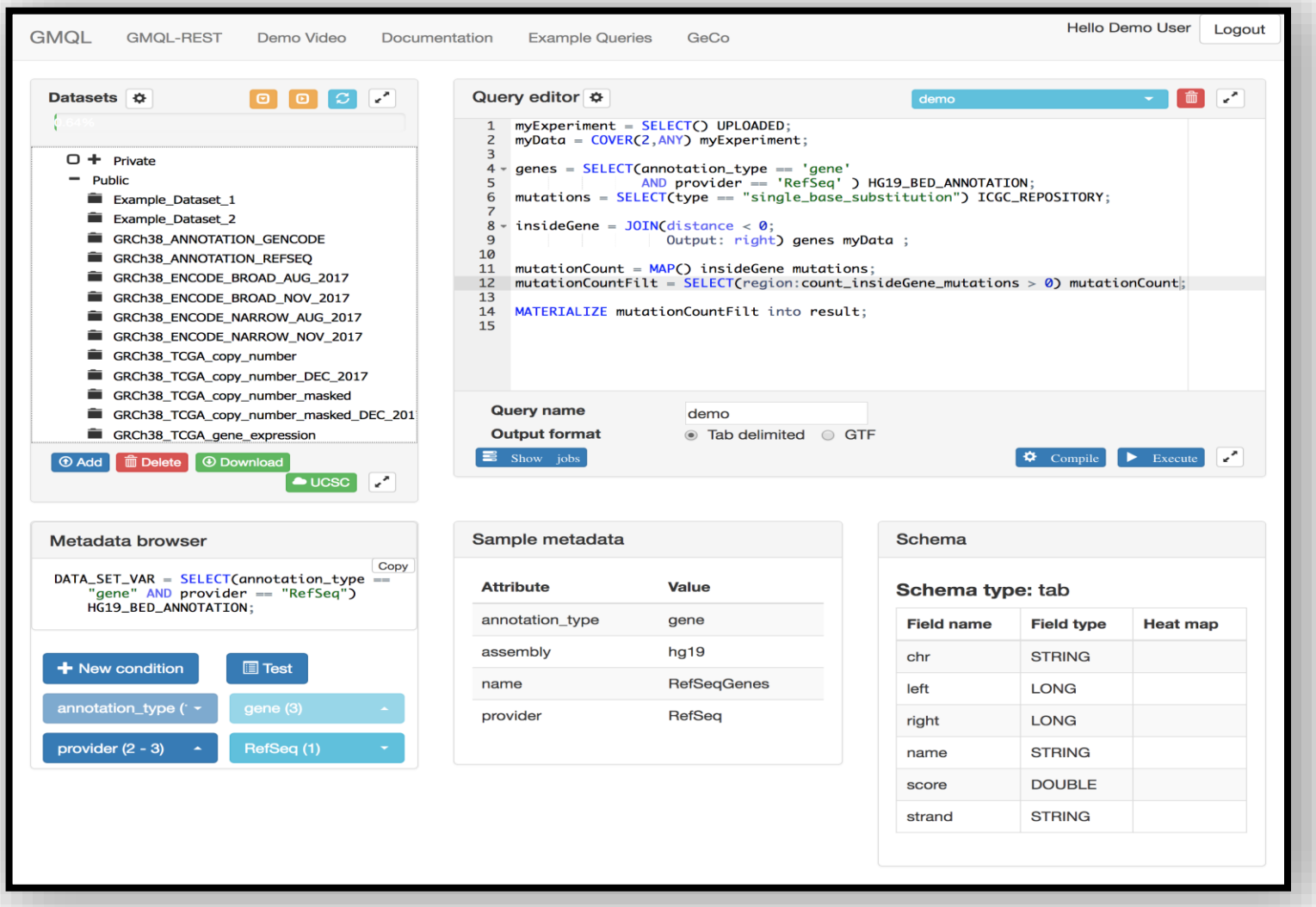
Experimental datasets and annotations from external databases

67 datasets
564k samples
9.23 TB

GECO Technology – bird eye's view

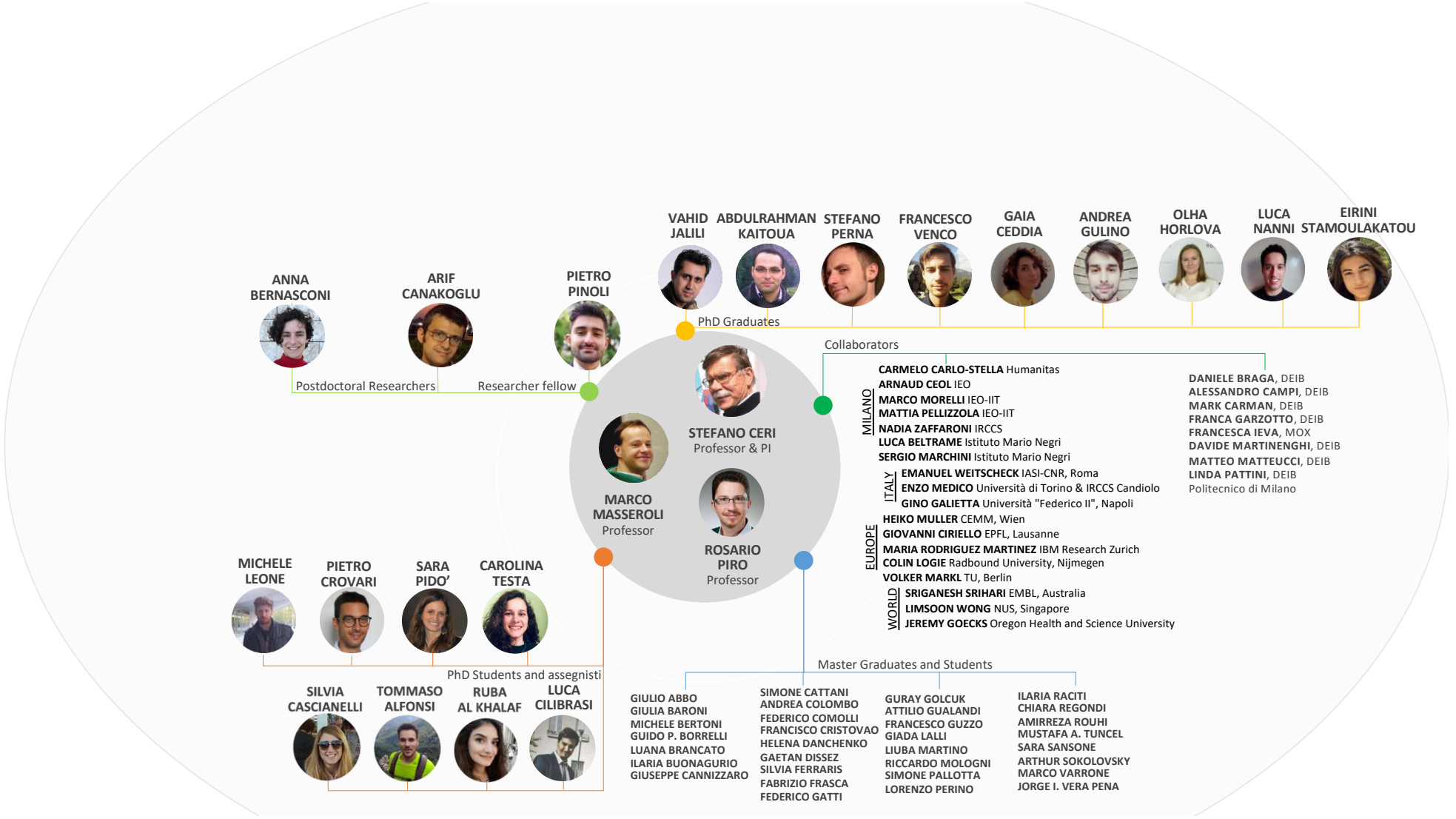


GMQL user Interface



Masseroli, M., Pinoli, P., Venco, F., Kaitoua, A., Jalili, V., Palluzzi, F., Muller, H. and Ceri, S., 2015. GenoMetric Query Language: a novel approach to large-scale genomic data management. Bioinformatics, 31(12), pp.1881-1888.

GECO Team --- 2016-2021



VIRAL GENOMICS:

DATA DESIGN
DATA INTEGRATION

INTERWIEWS (2020)




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)]
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)]

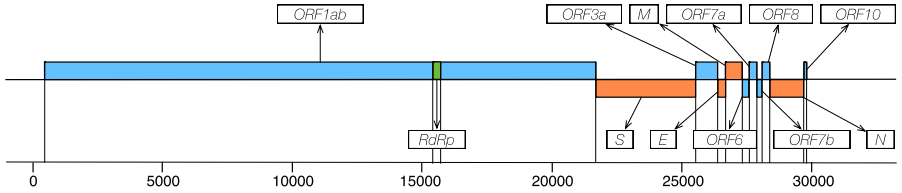
Private sector

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

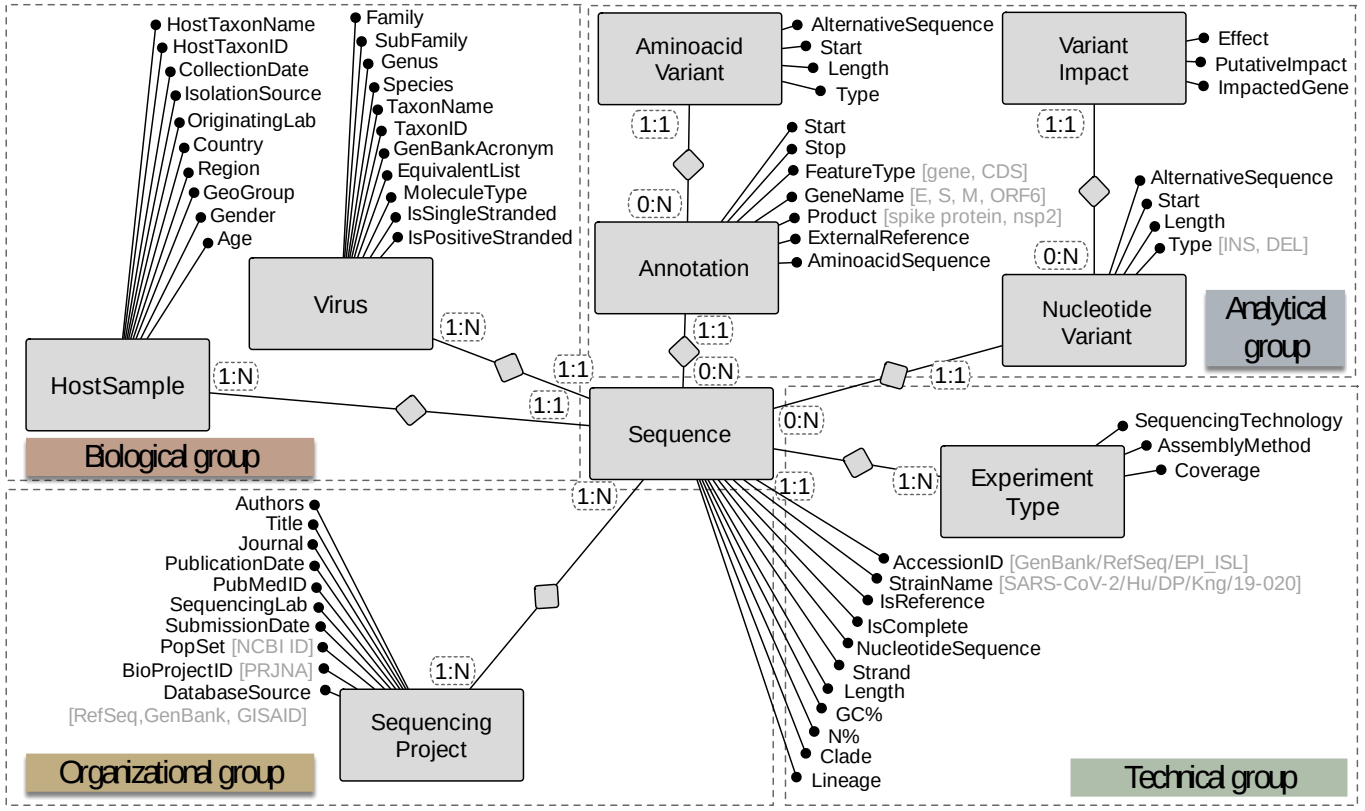
- Viral Conceptual Model
- COVID-19 Host Genetics Initiative Data Dictionary
- Genetic COVID-19 study (Siena)
- Sequence Integrated Database
- ViruSurf query engine
- VirusLab private annotation pipeline
- VirusViz
- EpiSurf
- ViruClust
- CoV2K
- ...

Viral Conceptual Model



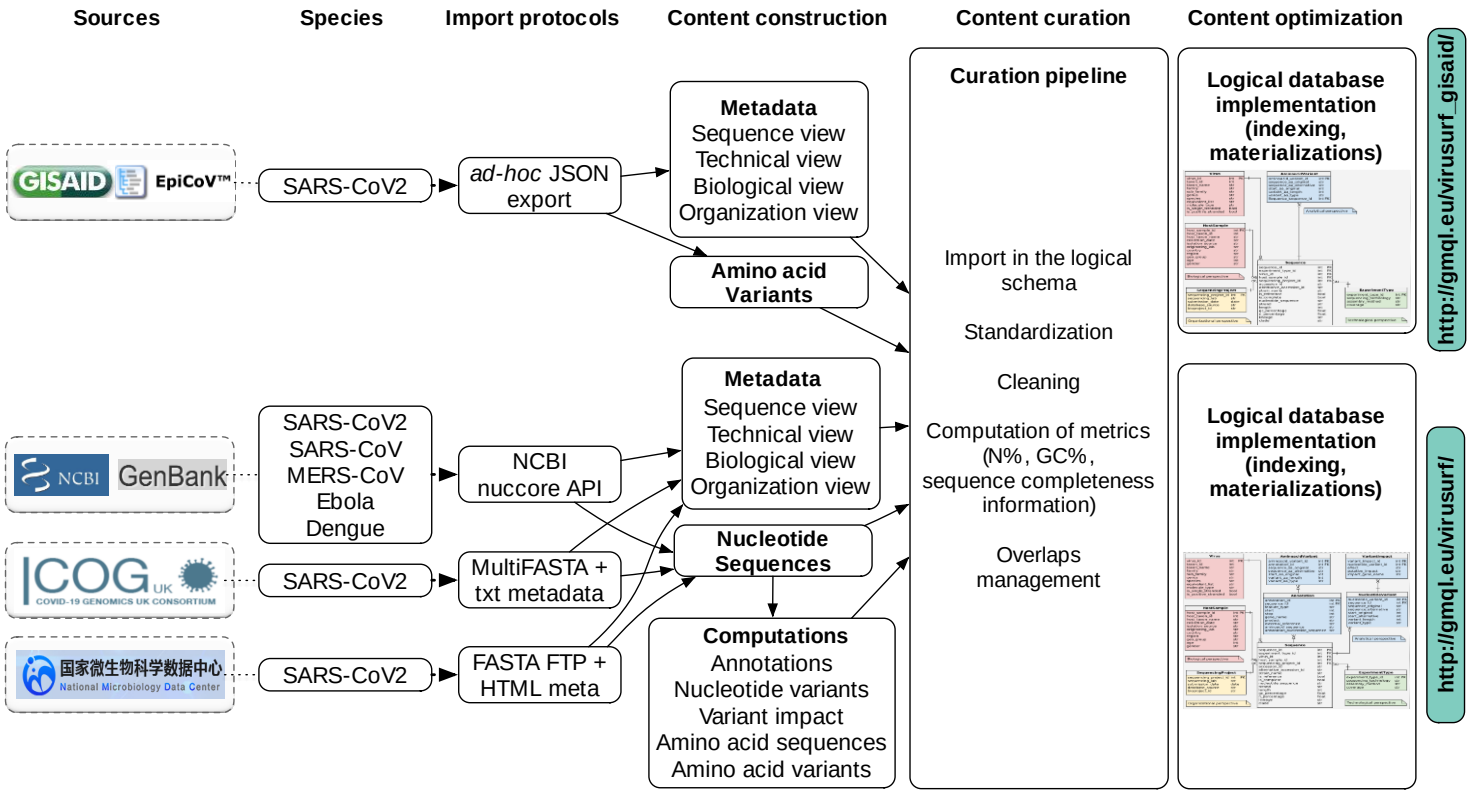
The **Viral Conceptual Model (VCM)**, centered on the virus **sequence** described from four perspectives:

- ***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, Canakoglu, Pinoli, Ceri. Empowering virus sequence research through conceptual modeling. In *Proceedings of the International Conference on Conceptual Modeling* (ER 2020)

Data integration pipelines for viral genomes data



January 2022

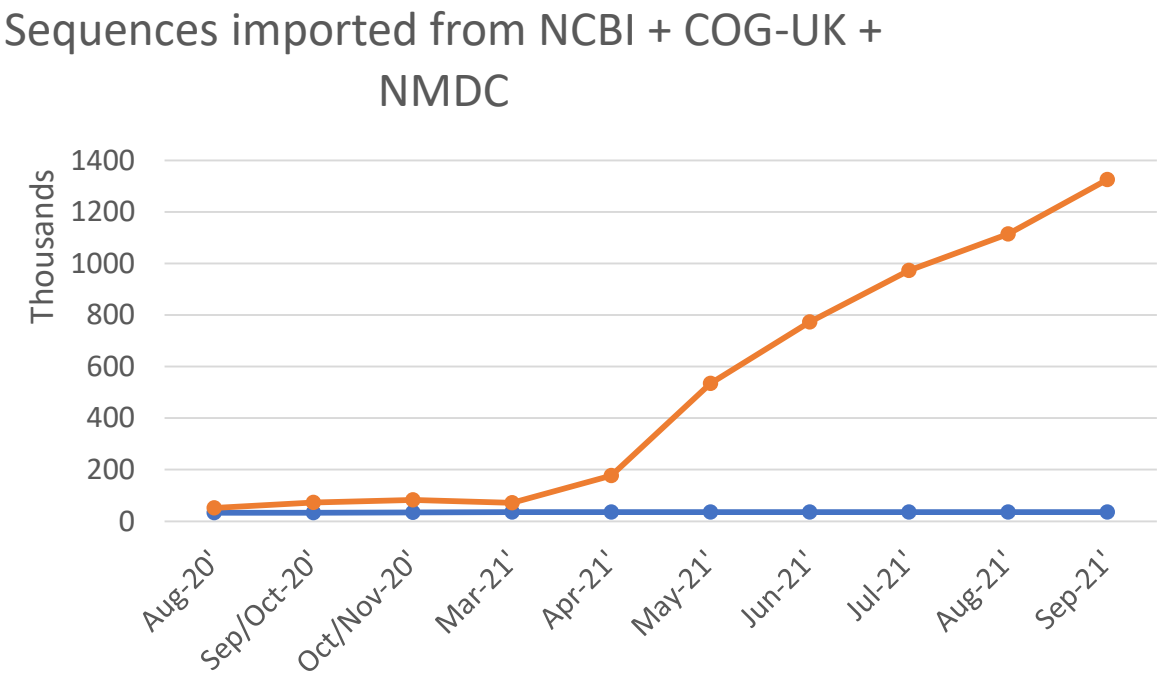
Sources for ViruSurf-GISAID

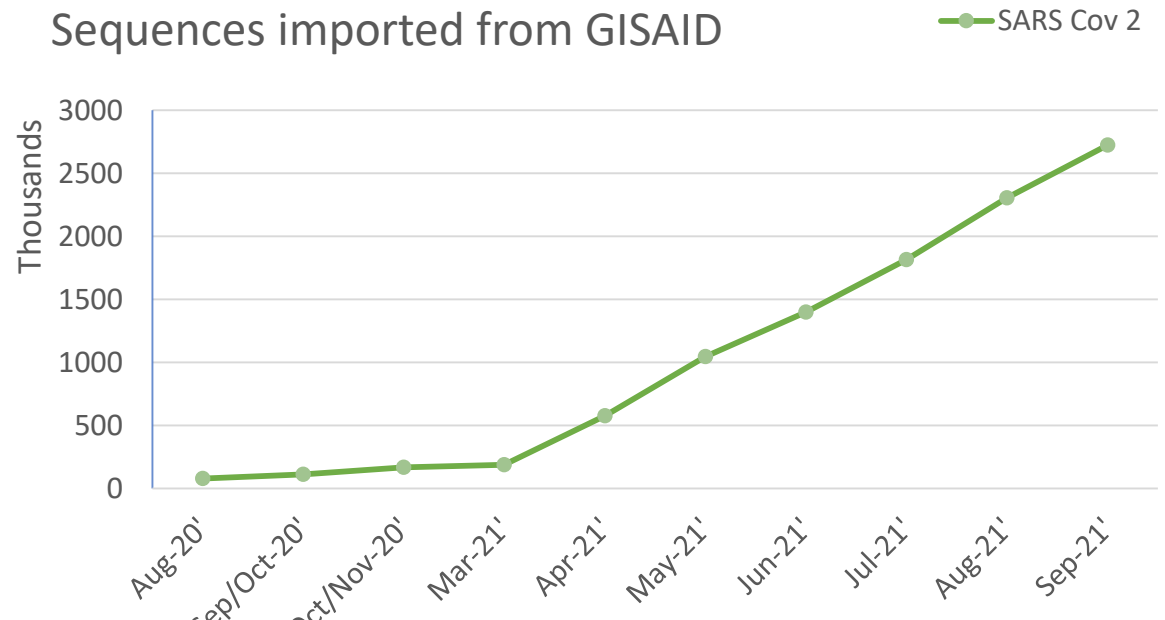
- GISAID EpiCoV™ db
  ~ 7M sequences

Sources for ViruSurf

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

Numbers of the import process



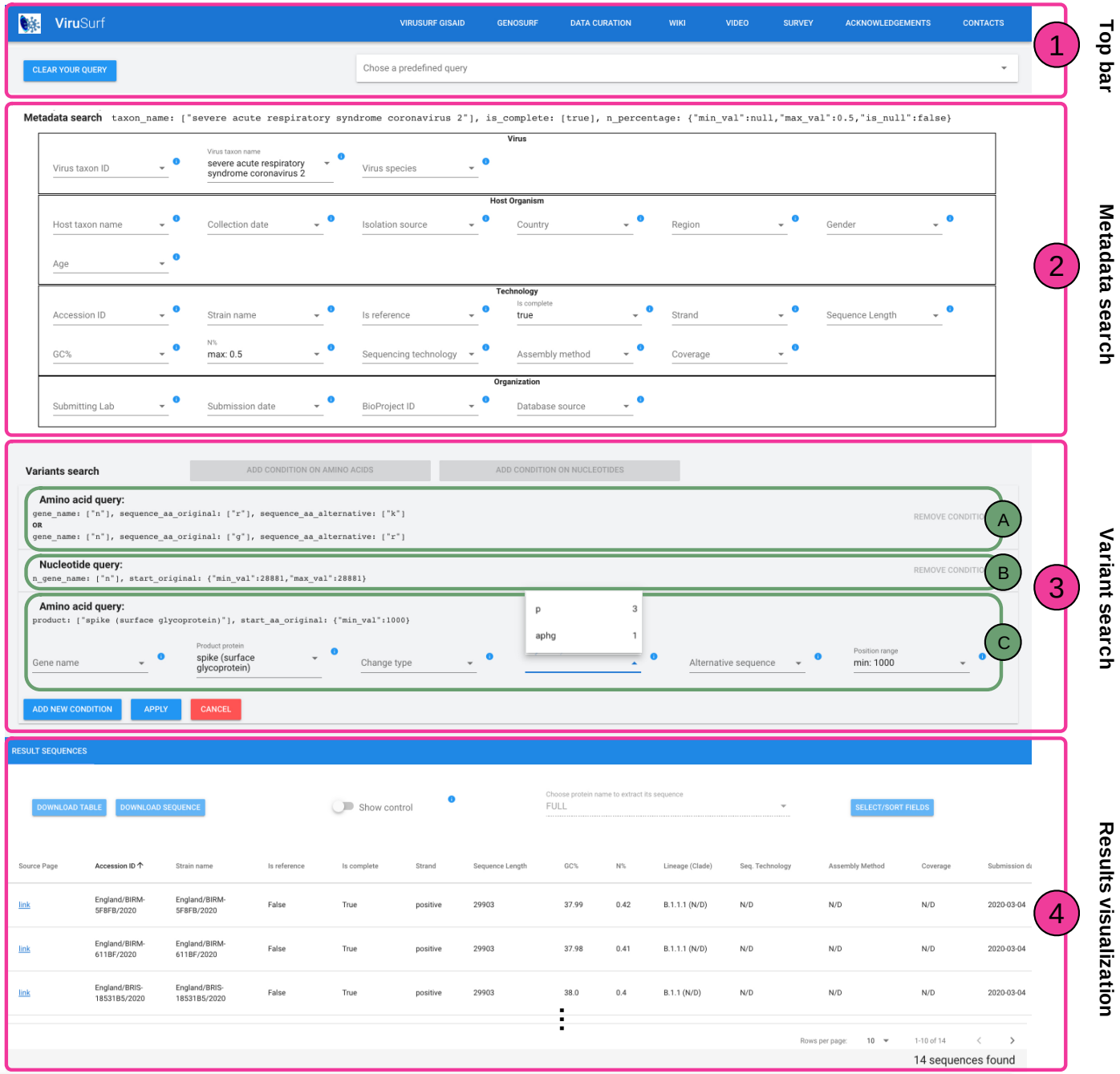


... ~16 million sequences July 2023!

ViruSurf

search system

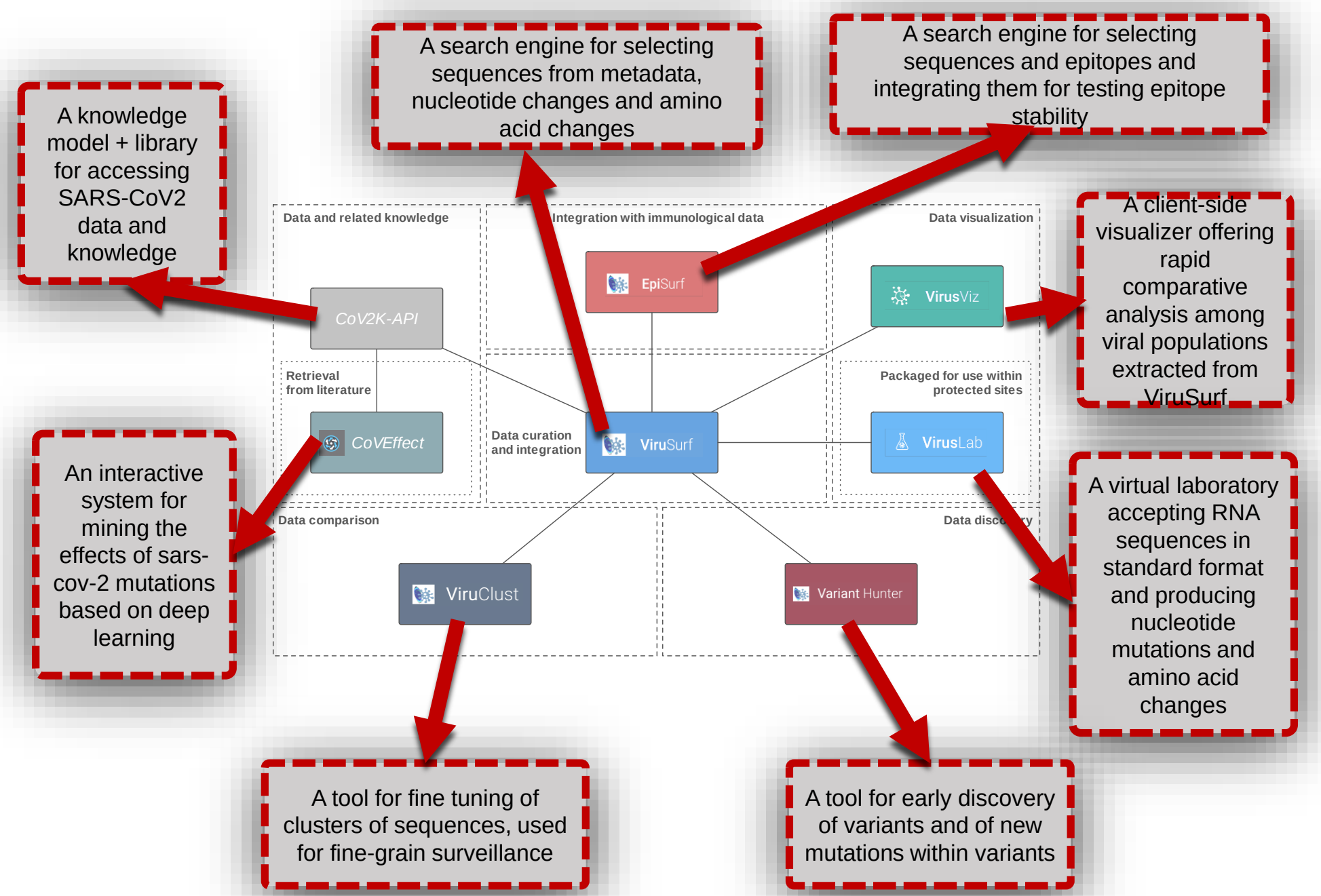
http://gmql.eu/virusurf/



A. Canakoglu, P. Pinoli, A. Bernasconi, et al. Virusurf: an integrated database to investigate viral sequences. *Nucleic Acids Research*, 49(D1):D817–D824, 2020.

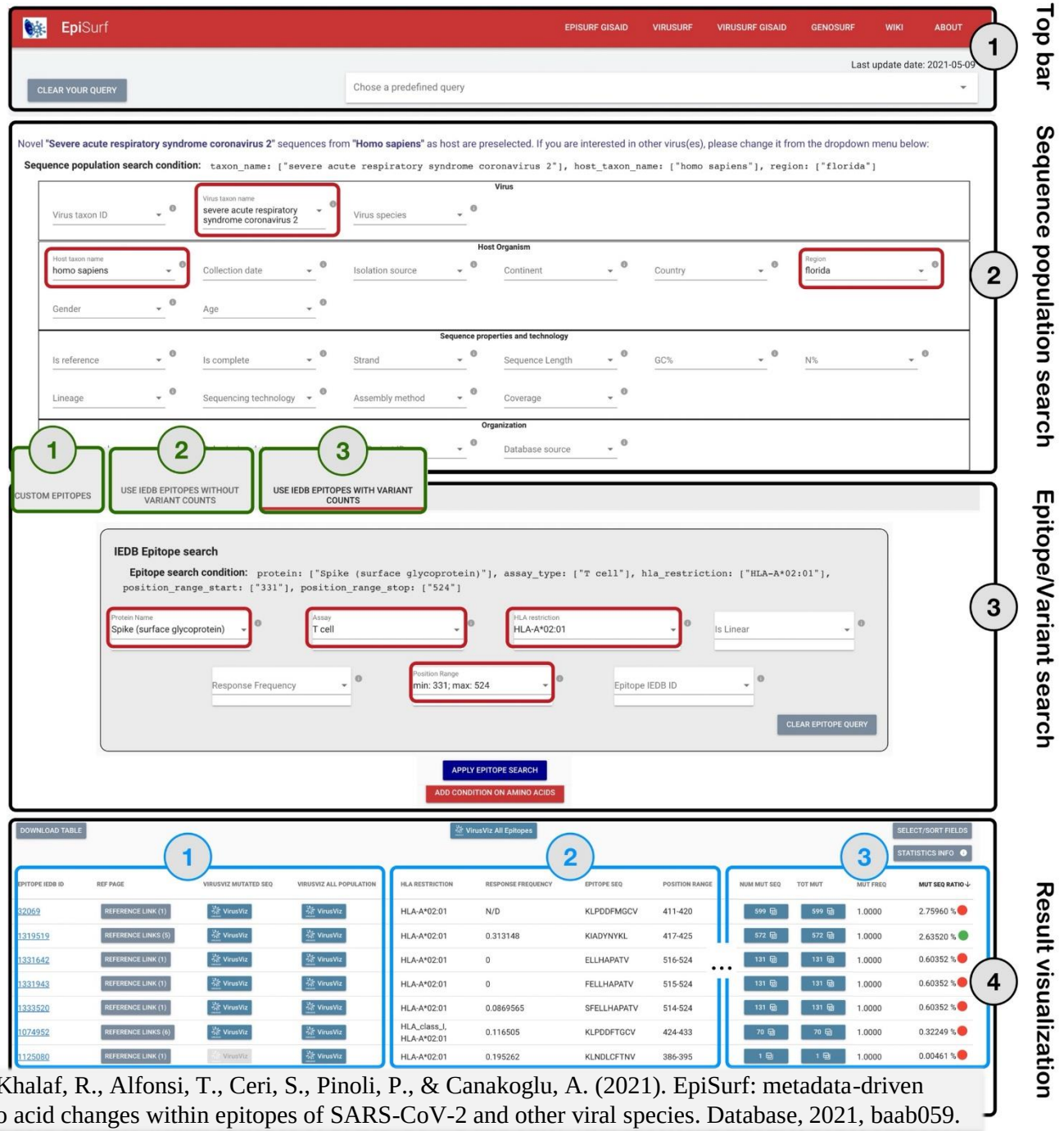
TOOLS

Tools for viral genomics



EpiViruSurf

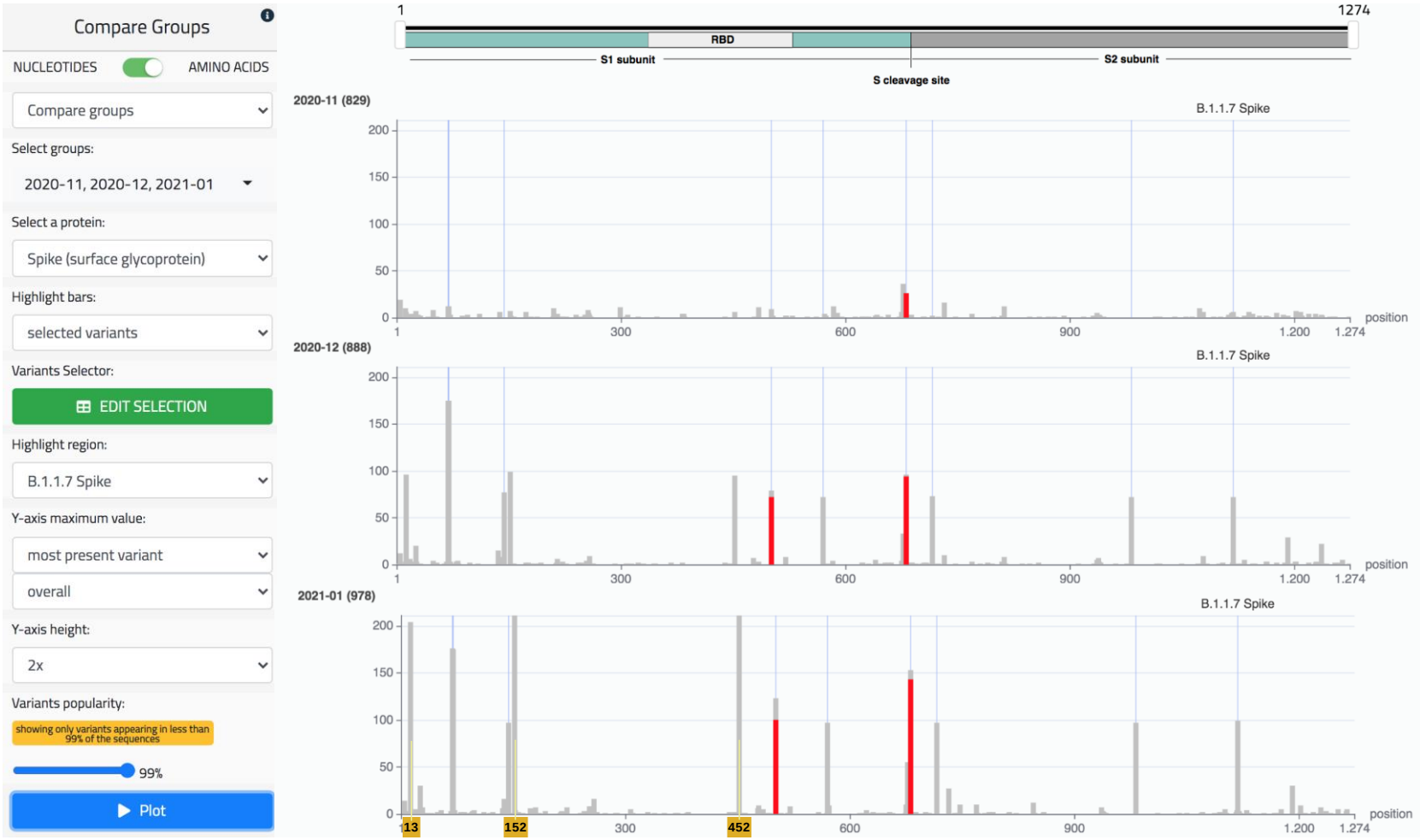
https://gmql.eu/epivirusurf



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 within epitopes of SARS-CoV-2 and other viral species. Database, 2021, baab059.

VirusViz

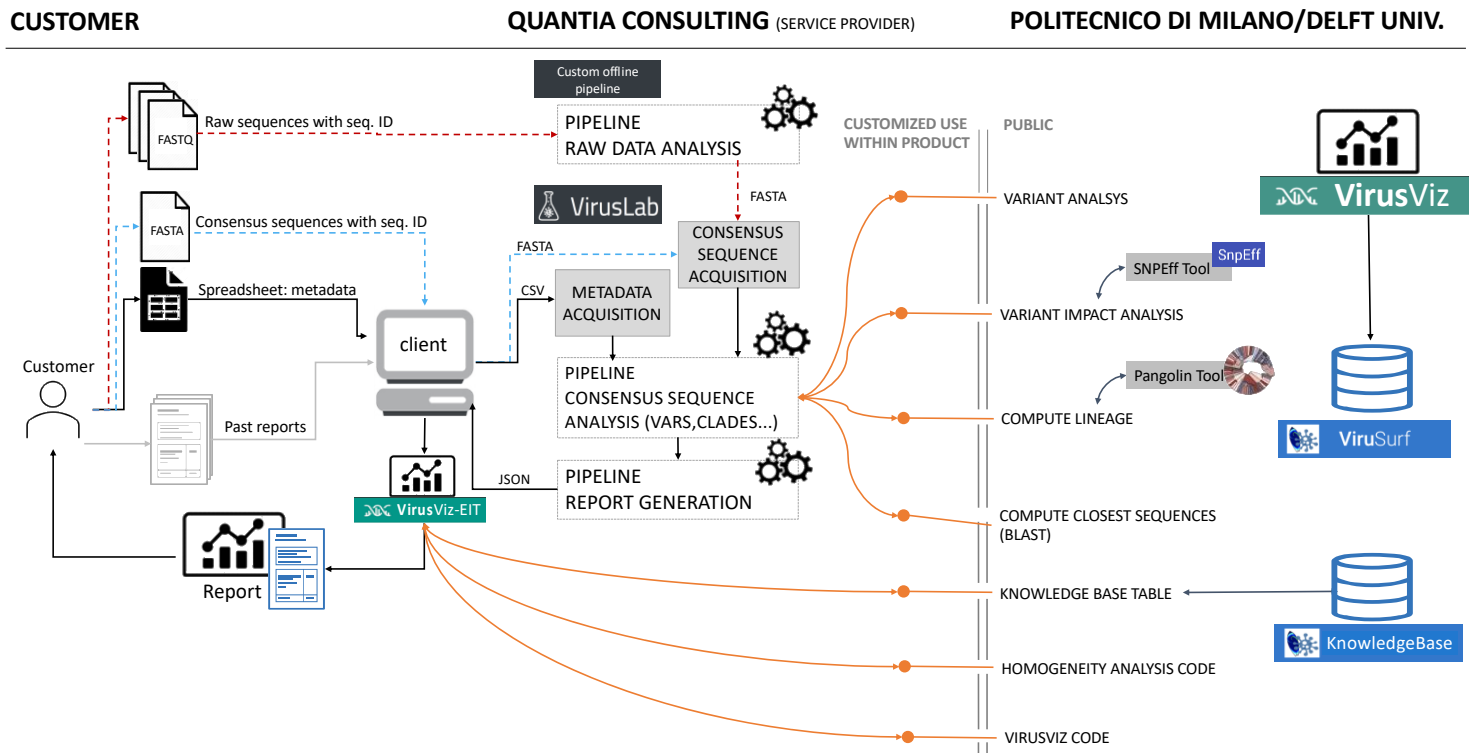
https://gmql.eu/virusviz



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-e90.

VirusLab: 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

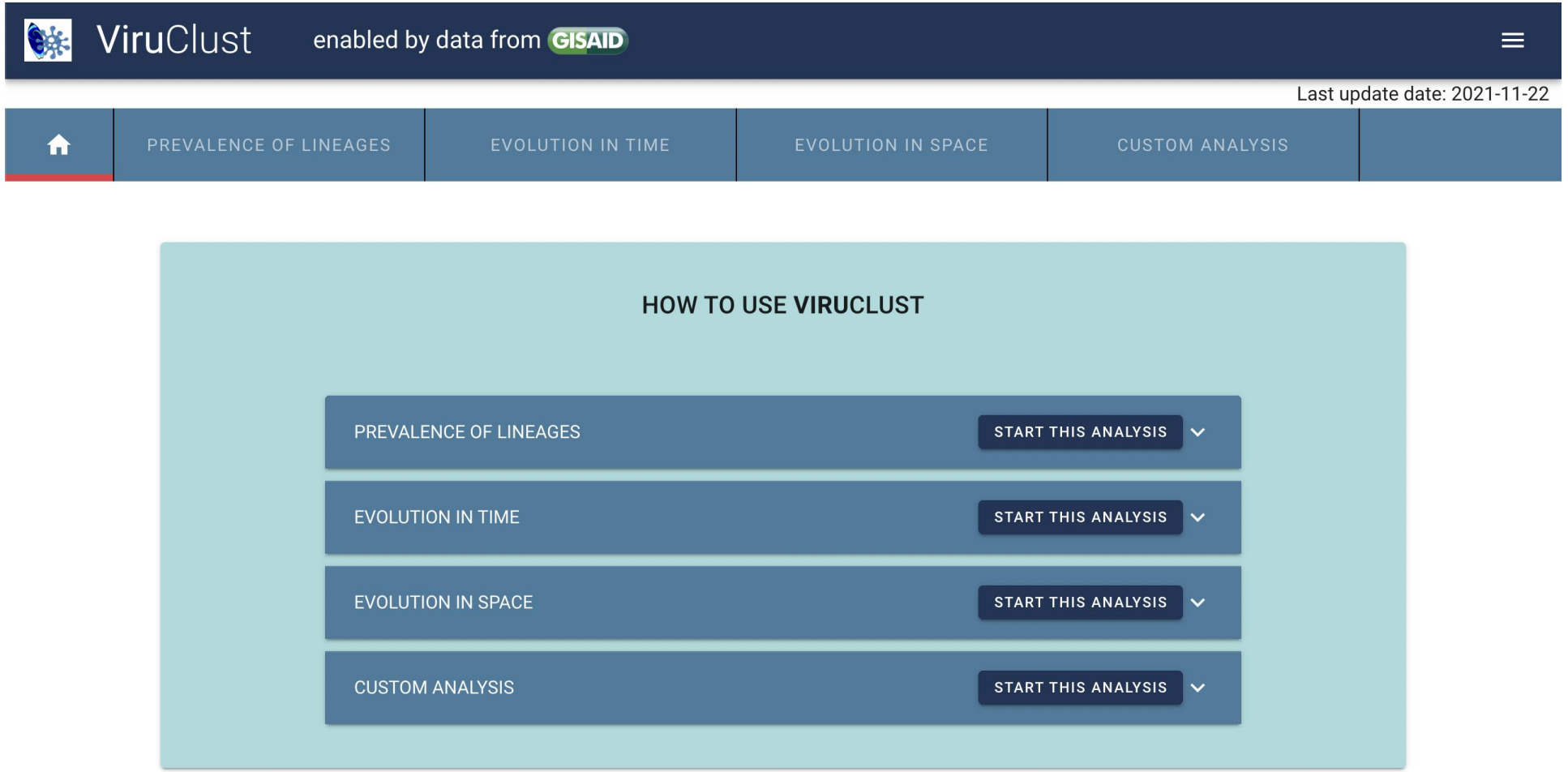


Project supported by:
EIT "DATA against COVID-19" Innovation Activity, Project 20663 "ViruSurf"

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

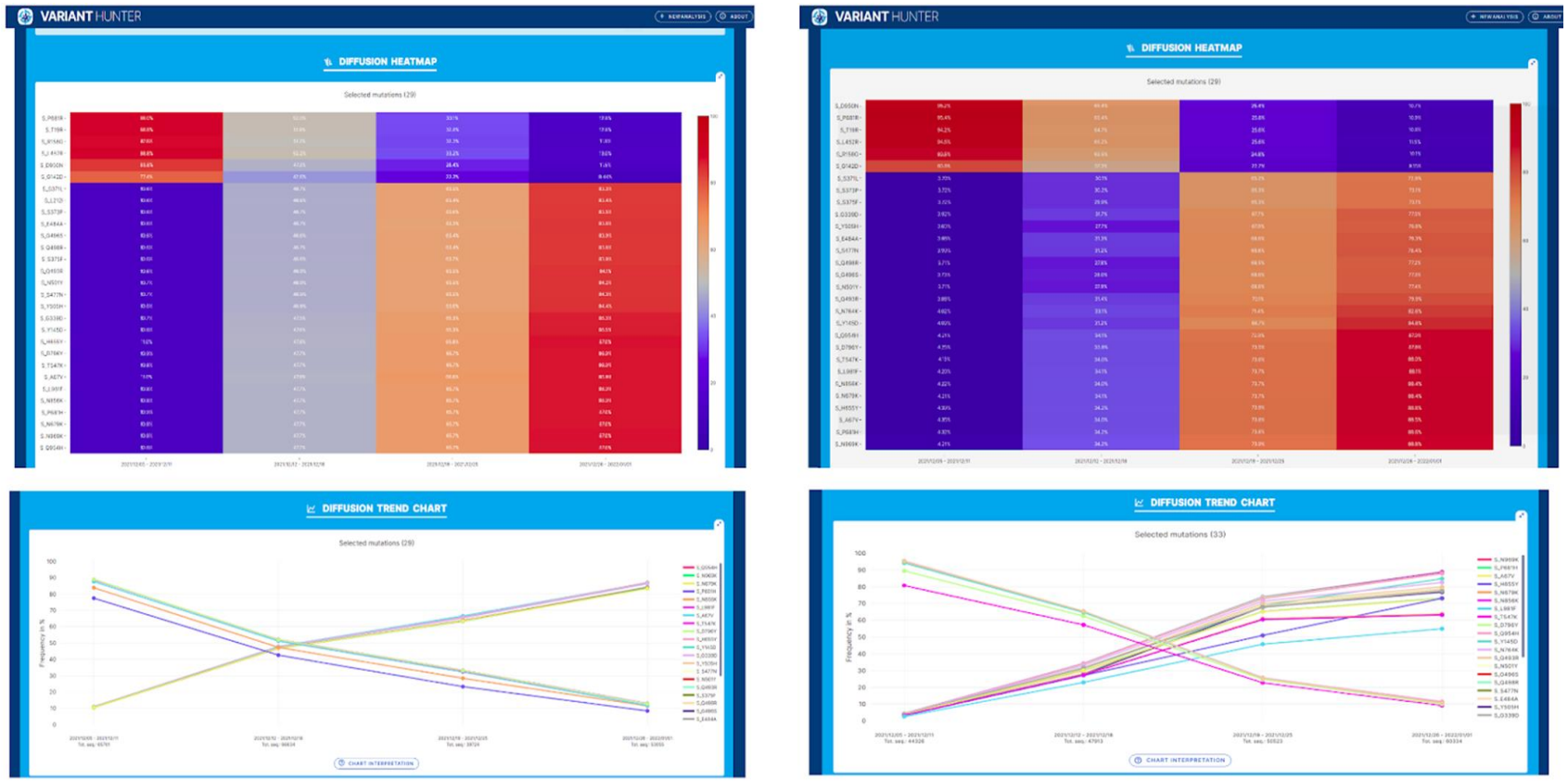
https://gmql.eu/viruclust



Cilibrasi, L., Pinoli, P., Bernasconi, A., Canakoglu, A., Chiara, M., & Ceri, S. (2022). ViruClust: direct comparison of SARS-CoV-2 genomes and genetic variants in space and time. Bioinformatics, 38(7), 1988-1994.

https://gmql.eu/variant_hunter

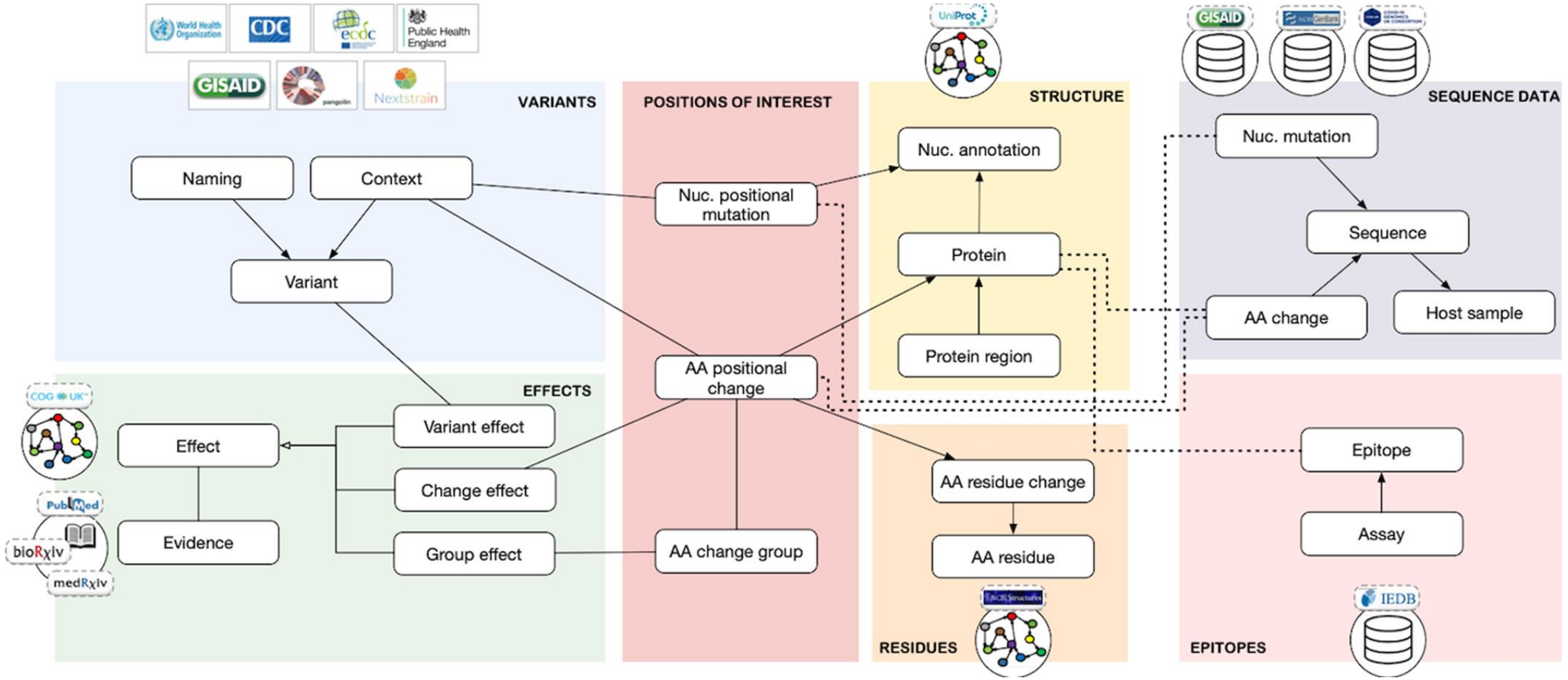
In the example: substitution of Delta variant by Omicron variant in Europe and North America



Pinoli, P., Canakoglu, A., Ceri, S., Chiara, M., Ferrandi, E., Minotti, L., & Bernasconi, A. (2023). VariantHunter: a method and tool for fast detection of emerging SARS-CoV-2 variants. Database, 2023, baad044.

CoV2K – Knowledge and data interplay

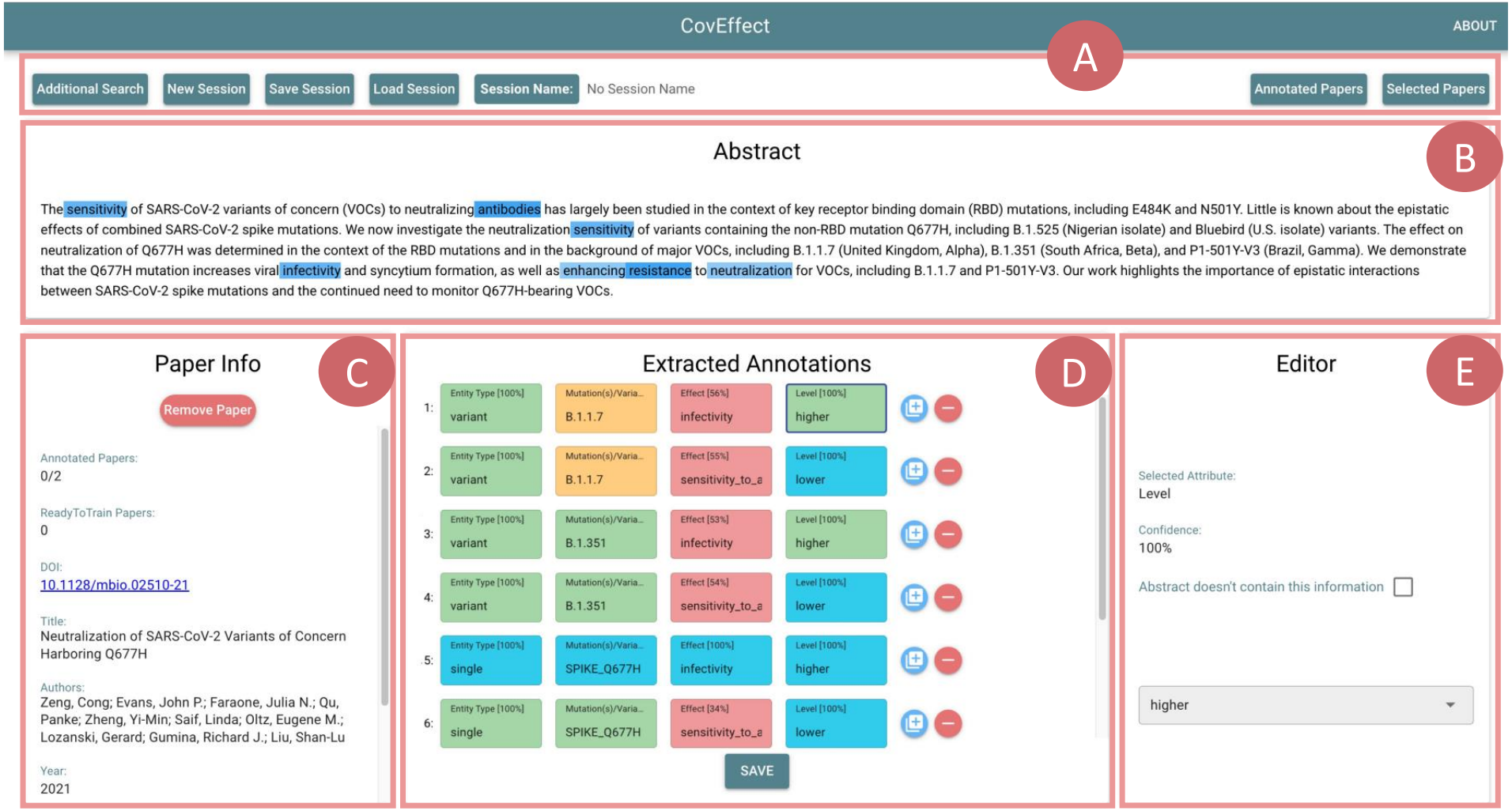
http://gmql.eu/cov2k/api/



Alfonsi, T., Al Khalaf, R., Ceri, S., & Bernasconi, A. (2022). CoV2K model, a comprehensive representation of SARS-CoV-2 knowledge and data interplay. Scientific Data, 9, 260.

CoVEffect

https://gmql.eu/coveffect

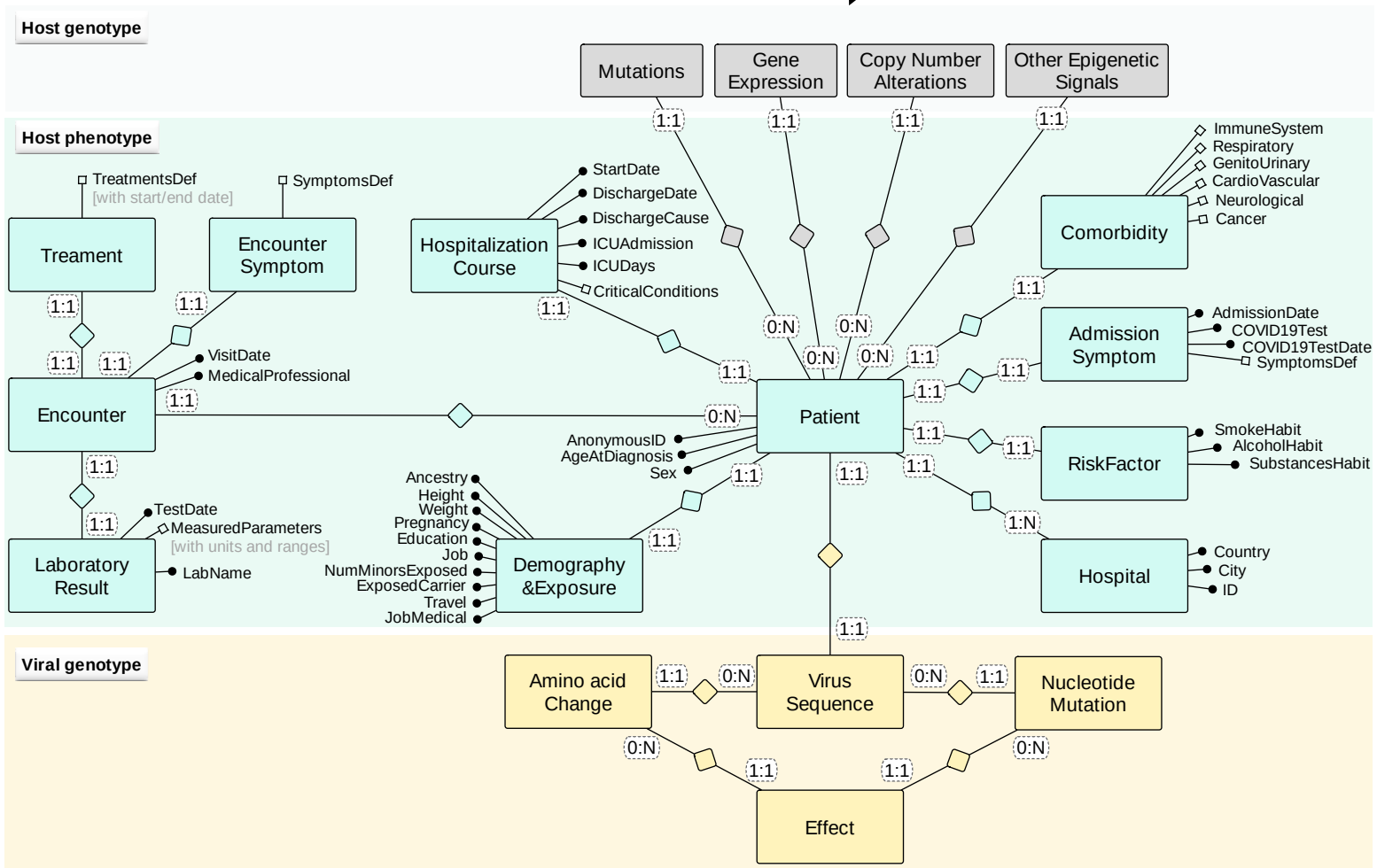


Serna García, G., Al Khalaf, R., Invernici, F., Ceri, S., & Bernasconi, A. (2023). CoVEffect: interactive system for mining the effects of SARS-CoV-2 mutations and variants based on deep learning. GigaScience, 12, giad036.

CLINICAL RESEARCH

COVID-19 Conceptual Model

Described by the Viral Conceptual Model



Described by the Genomic Conceptual Model

http://gmql.eu/genosurf/

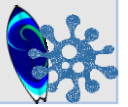
GenoSurf

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

PhenotypeDB

http://gmql.eu/virusurf/

ViruSurf



Research empowered by availability of clinical + genetic data

Article | Open Access | Published: 08 July 2021

Mapping the human genetic architecture of COVID-19

COVID-19 Host Genetics Initiative

Nature **600**, 472–477 (2021) | Cite this article

284k Accesses | **370** Citations | **1900** Altmetric | Metrics

Article | Open Access | Published: 07 March 2022

Whole-genome sequencing reveals host factors underlying critical COVID-19

Athanasios Kousathanas, Erola Pairo-Castineira, Konrad Rawlik, Alex Stuckey, Christopher A. Odhams, Susan Walker, Clark D. Russell, Tomas Malinauskas, Yang Wu, Jonathan Millar, Xia Shen, Katherine S. Elliott, Fiona Griffiths, Wilna Oosthuyzen, Kirstie Morrice, Sean Keating, Bo Wang, Daniel Rhodes, Lucija Klaric, Marie Zechner, Nick Parkinson, Afshan Siddiq, Peter Goddard, Sally Donovan, GenOMICC investigators, 23andMe investigators, COVID-19 Human Genetics Initiative, ... J. Kenneth Baillie

+ Show authors

Nature **607**, 97–103 (2022) | Cite this article

135k Accesses | **94** Citations | **1560** Altmetric | Metrics

Matters Arising | Open Access | Published: 03 August 2022

A first update on mapping the human genetic architecture of COVID-19

COVID-19 Host Genetics Initiative

Nature **608**, E1–E10 (2022) | Cite this article

30k Accesses | **57** Citations | **217** Altmetric | Metrics

Research empowered by availability of clinical + genetic data

Article | Open Access | Published: 17 January 2021

Employing a systematic approach to biobanking and analyzing clinical and genetic data for advancing COVID-19 research

Sergio Daga, Chiara Fallerini, Margherita Baldassarri, Francesca Fava, Floriana Valentino, Gabriella Doddato, Elisa Benetti, Simone Furini, Annarita Giliberti, Rossella Tita, Sara Amitrano, Mirella Bruttini, Ilaria Meloni, Anna Maria Pinto, Francesco Raimondi, Alessandra Stella, Filippo Biscarini, Nicola Picchiotti, Marco Gori, Pietro Pinoli, Stefano Ceri, Maurizio Sanarico, Francis P. Crawley, Giovanni Birolo, GEN-COVID Multicenter Study, ... Elisa Frullanti + Show authors

European Journal of Human Genetics **29**, 745–759 (2021) | Cite this article

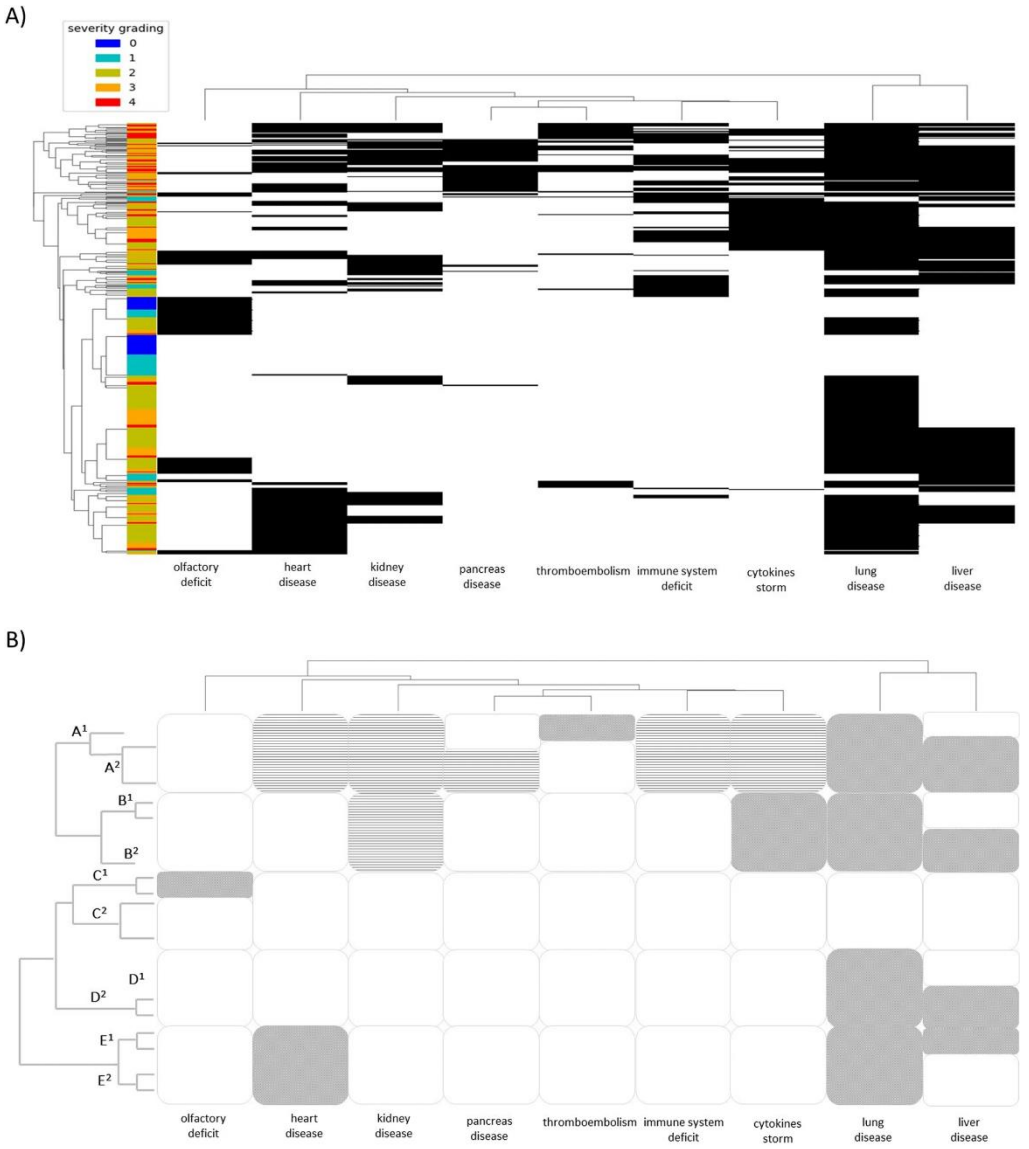
Cardiol Cardiovasc Med 2021; 5 (6): 673-694 DOI: 10.26502/fccm.92920232

Research Article

Post-Mendelian Genetic Model in COVID-19

Nicola Picchiotti[1,2#], Elisa Benetti[3#], Chiara Fallerini[3,4#], Sergio Daga[3,4], Margherita Baldassarri[3,4], Francesca Fava[3,4,5], Kristina Zguro[3], Floriana Valentino[3,4], Gabriella Doddato[3,4], Annarita Giliberti[3,4], Rossella Tita[5], Sara Amitrano[5], Mirella Bruttini[3,4,5], Laura Di Sarno[3,4], Diana Alaverdian[3,4], Giada Beligni[3,4], Maria Palmieri[3,4], Susanna Croci[3,4], Mirjam Lista[3,4], Ilaria Meloni[3,4], Anna Maria Pinto[5], Chiara Gabbi[6], Stefano Ceri[7], Antonio Esposito[7], Pietro Pinoli[7], Francis P. Crawley[8], Elisa Frullanti [3,4], Francesca Mari[3,4,5], GEN-COVID Multicenter Study, Marco Gori[1,9], Alessandra Renieri[3,4,5*], Simone Furini[3*]

[1]University of Siena, DIISM-SAILAB, Siena, Italy
[2]Department of Mathematics, University of Pavia, Pavia, Italy
[3]Med Biotech Hub and Competence Center, Department of Medical Biotechnologies, University of Siena, Italy
[4]Medical Genetics, University of Siena, Italy
[5]Genetica Medica, Azienda Ospedaliero-Universitaria Senese, Italy
[6]Department of Biosciences and Nutrition, Karolinska Institutet, Stockholm, Sweden
[7]Politecnico di Milano, DEIB, Milano, Italy
[8]Good Clinical Practice Alliance-Europe (GCPA) and Strategic Initiative for Developing Capacity in Ethical Review-Europe (SIDCER), Leuven, Belgium.
[9]Universite Côte d'Azur, Inria, CNRS, I3S, Maasai



Association between SARS-CoV-2 viremia and COVID-19 mortality

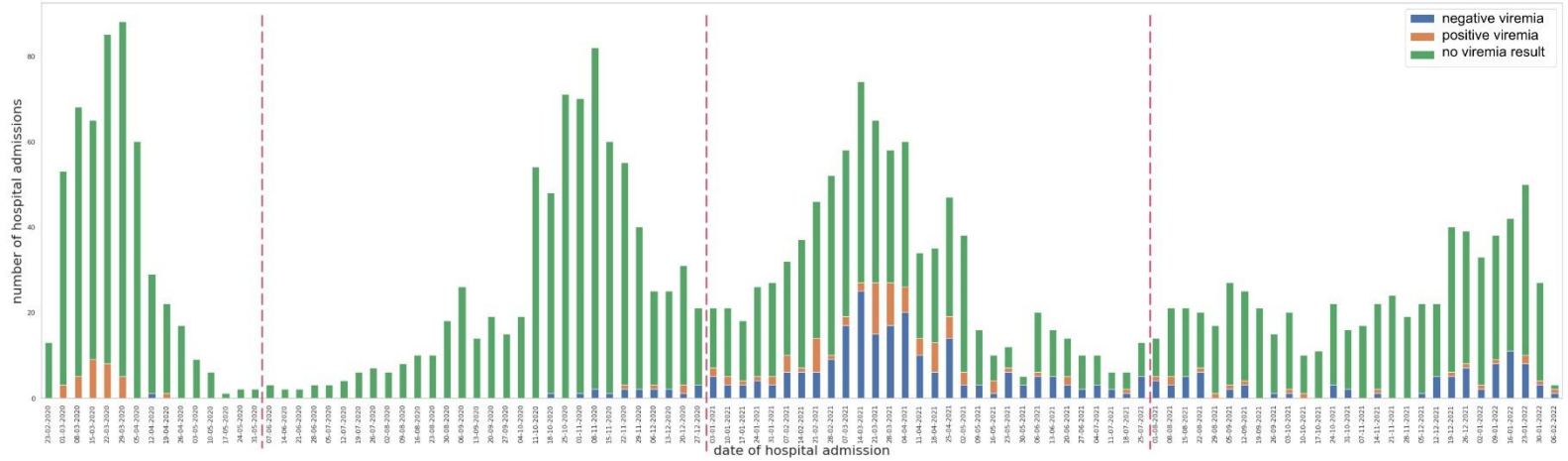


Table 2. Logistic regression analysis of factors associated with death.

Characteristics	OR	95%CI	aOR	95%CI
Females *vs* Males	0.57	0.35–0.94	0.63	0.38–1.03
Age (per 1 year more)	1.01	1.01–1.02	1.09	1.06–1.11
CCI (per one point more)	0.99	0.98–1.02	–	–
SARS-CoV-2 epidemic waves				
1 *vs* 2	1.71	0.54–5.41	–	–
1 *vs* 3	1.18	0.57–2.42	–	–
1 *vs* 4	0.72	0.31–1.67	–	–
Days from symptoms onset (per one day more)	1.00	1.00–1.01	–	–
Disease severity (Mild/moderate vs severe/critical)	1.85	1.12–3.06	1.55	0.95–2.54
Doses of vaccine				
0 *vs* 1	0.97	0.35–2.66	–	–
0 *vs* 2	1.02	0.45–2.30	–	–
0 *vs* 3	1.16	0.24–5.71	–	–
Positive SARS-CoV-2 viremia *vs* negative	5.16	3.15–8.45	6.48	4.05–10.55

List of abbreviations: OR, odds ratio; CI, confidence interval; aOR, adjusted odds ratio; CCI, Charlson comorbidity index.

Giacomelli, A., Righini, E., Micheli, V., Pinoli, P., Bernasconi, A., Rizzo, A., ... & Rizzardini, G. (2023). SARS-CoV-2 viremia and COVID-19 mortality: A prospective observational study. Plos one, 18(4), e0281052.

METHODS FOR VIRAL GENOMICS

Time-series analysis of viral amino acid changes

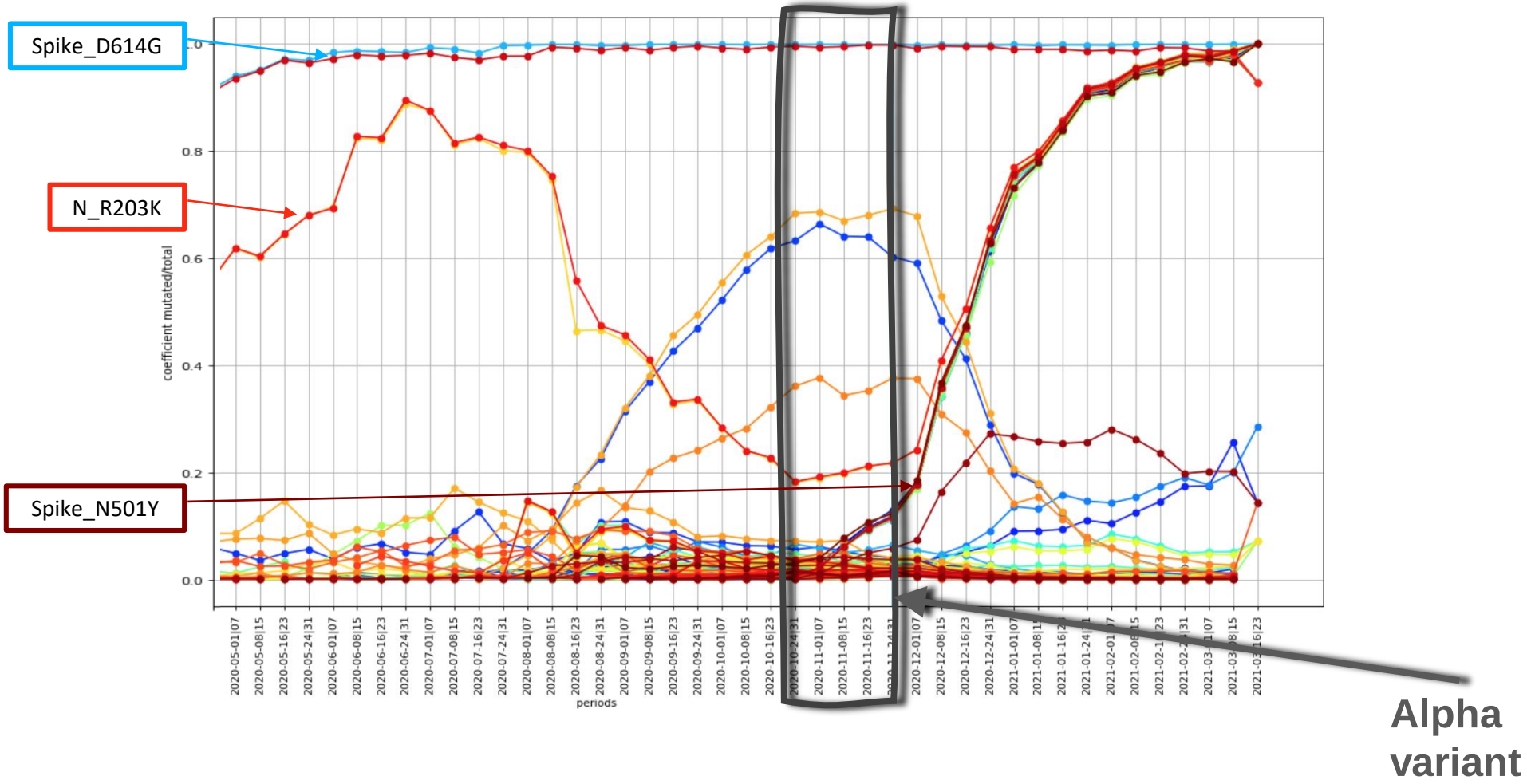
Focus
Co-occurrence of mutations (→ amino acid changes) on the virus (→ variant)

Idea
A variant can be identified by observing the dynamics of its amino acid changes (weekly prevalences in a geo-location)

Different changes could indicate the birth of a variant when:

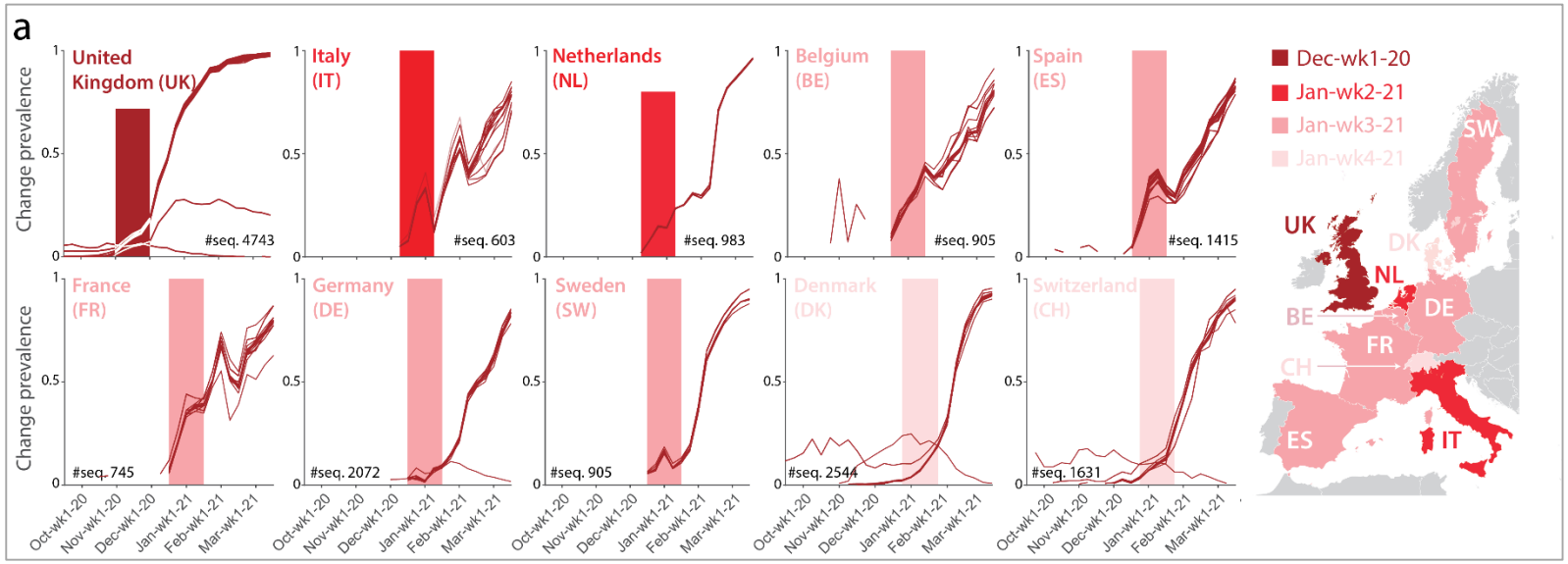
- ❖ their time-series are similar
- ❖ they are all growing

CoVolution

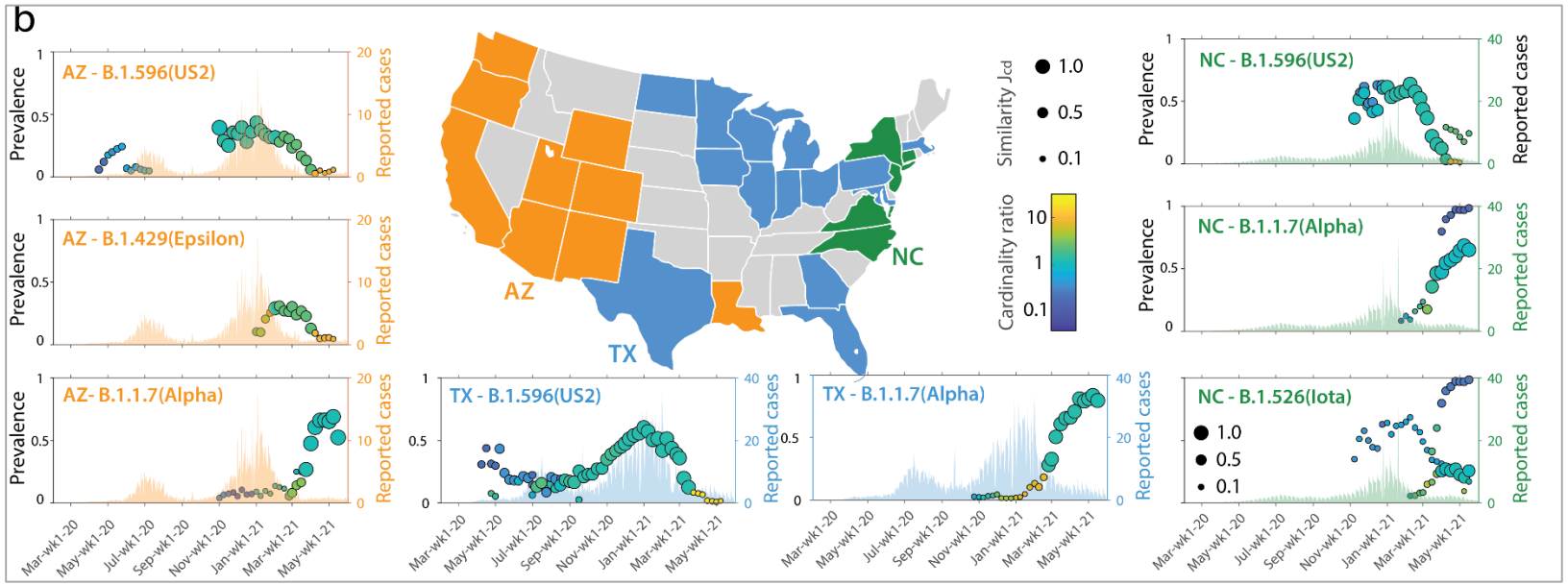

Spike_D614G
N_R203K
Spike_N501Y
coefficient mutated/total
periods
Alpha variant

Europe and US variants

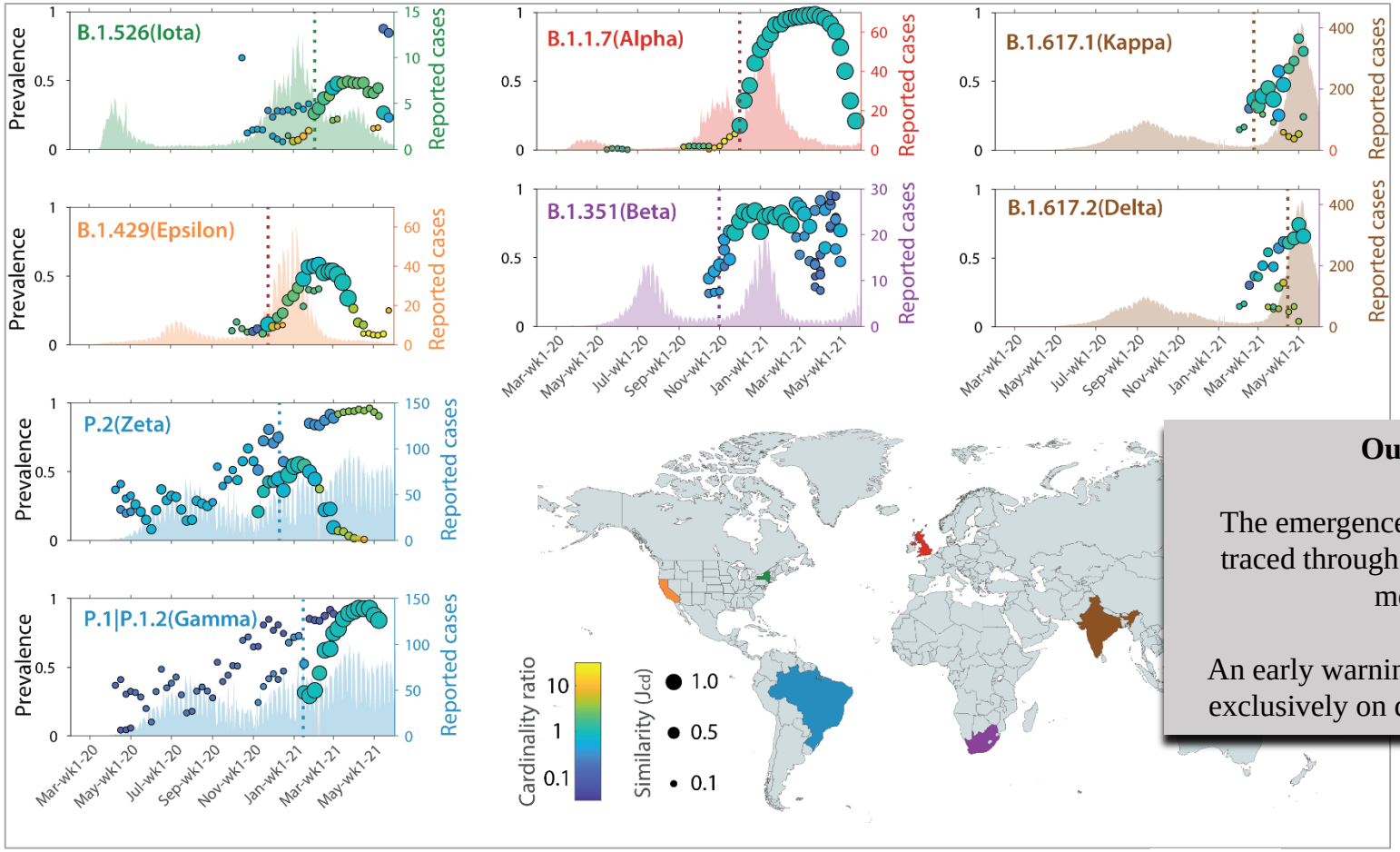
Temporal dynamics of the Alpha variant in European



Temporal dynamics of notable variants in the US



Assessment of the early warning system



Outcomes

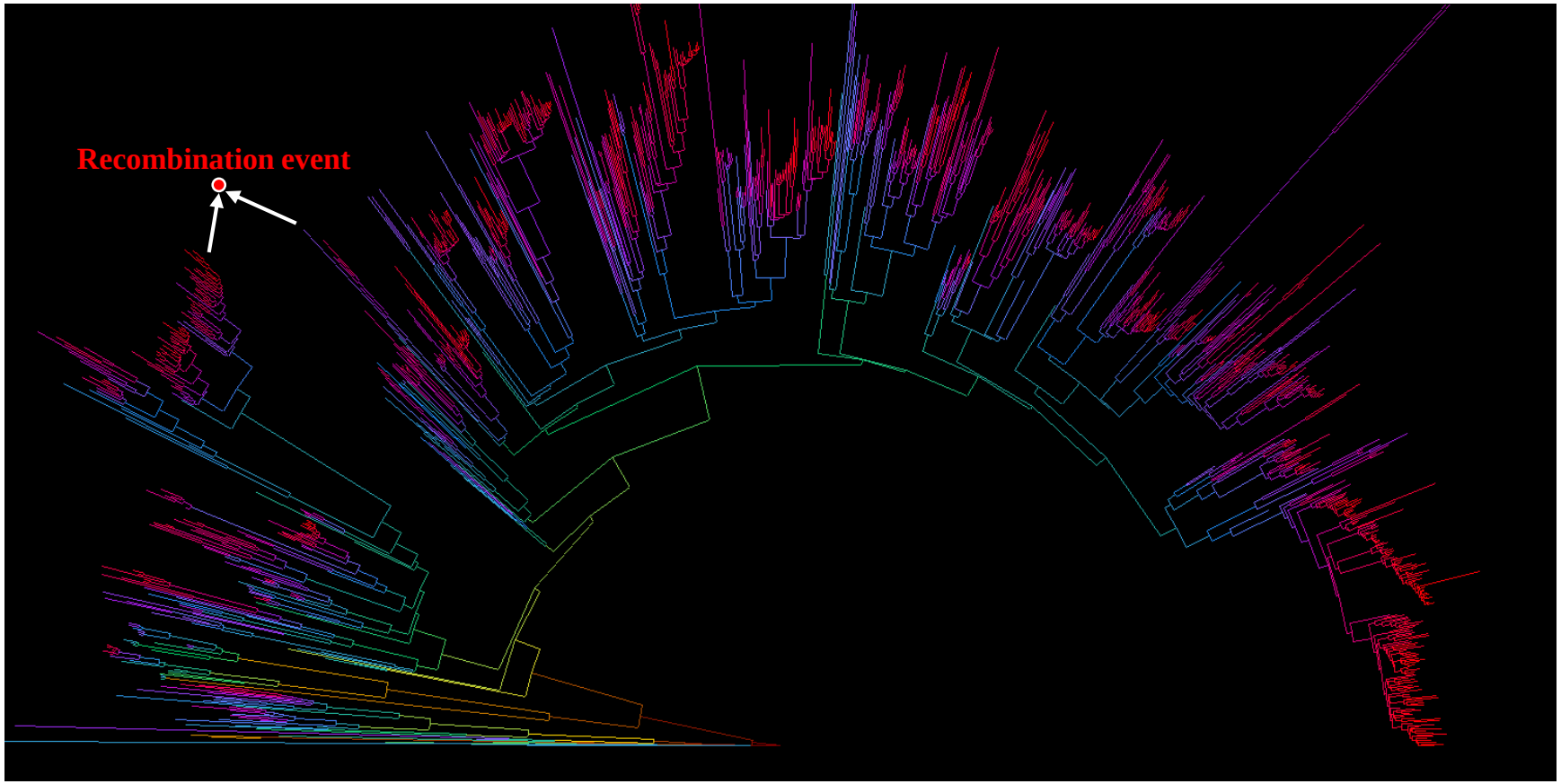
The emergence of variants can be traced through purely data-driven methods

An early warning system could rely exclusively on deposited sequences

Emergence of variants in their country of origin.

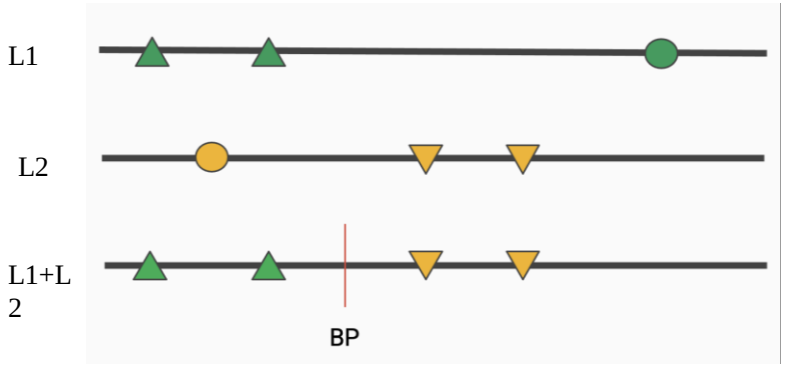
Bernasconi, A., Mari, L., Casagrandi, R., Ceri, S. **Data-driven analysis of amino acid change dynamics timely reveals SARS-CoV-2 variant emergence.** Scientific Reports 11, 21068 (2021). DOI: 10.1038/s41598

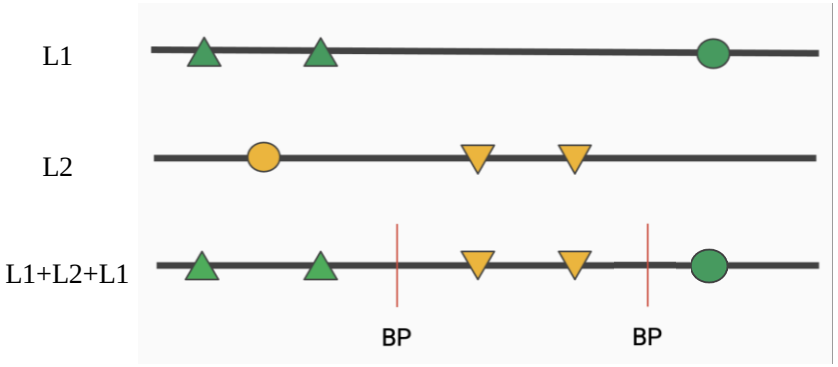
Viral recombination



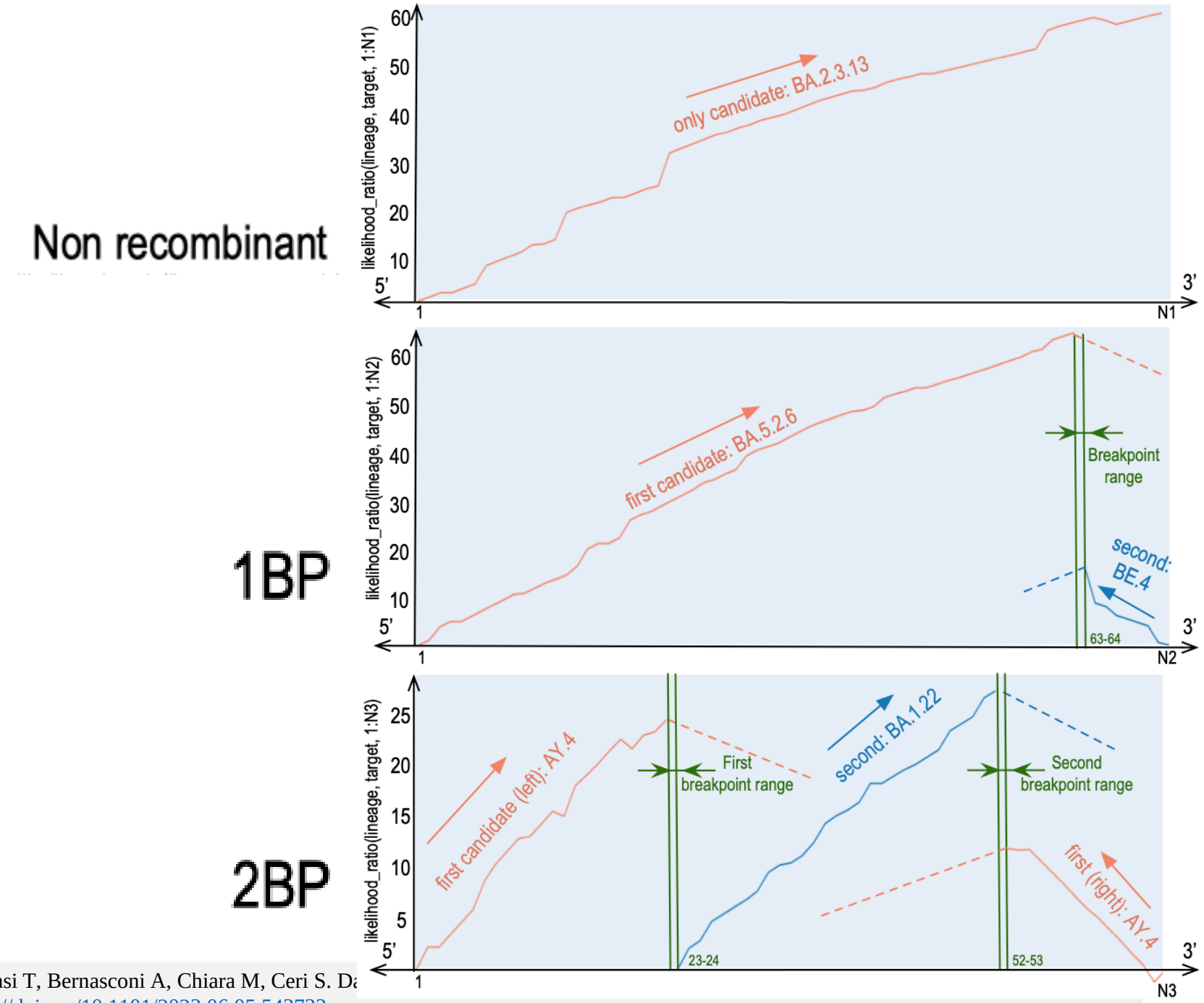
Recombination in SARS-CoV-2

- 76 recombination events recognized in the SARS-CoV-2 virus evolution history (as of July 2023)
- Lineages XBB.1.5 and XBB.1.16 (descendants of the recombinant XBB) are considered **Variants Of Interest** (WHO, August 2023) leading to:
  - https://www.who.int/activities/tracking-SARS-CoV-2-variants tion
  - reduced efficacy of treatments
  - predicted increase in transmissibility or disease severity
- Typical recombination patterns:





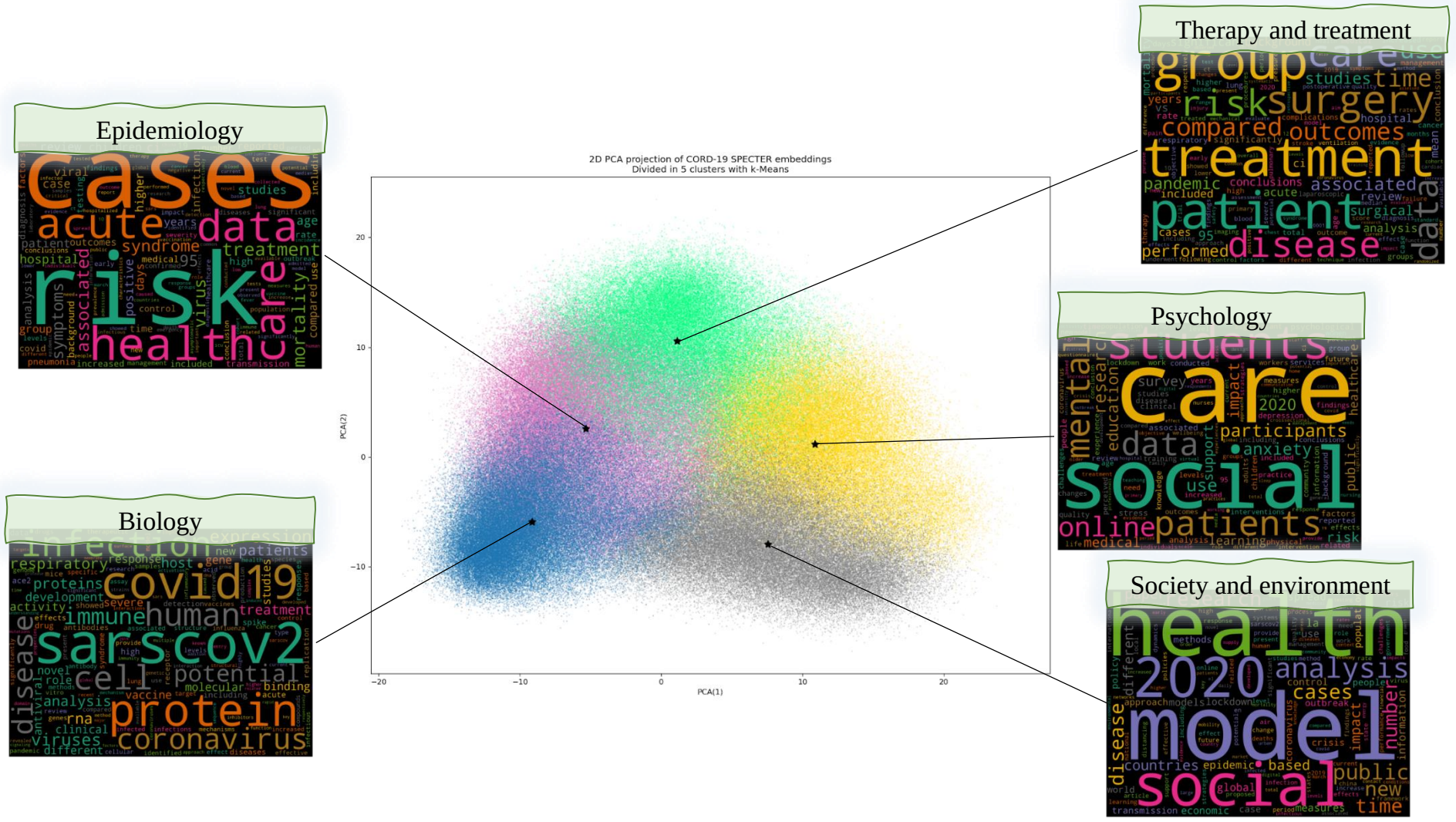
The RecombinHunt method



Alfonsi T, Bernasconi A, Chiara M, Ceri S. Da
https://doi.org/10.1101/2023.06.05.543733.

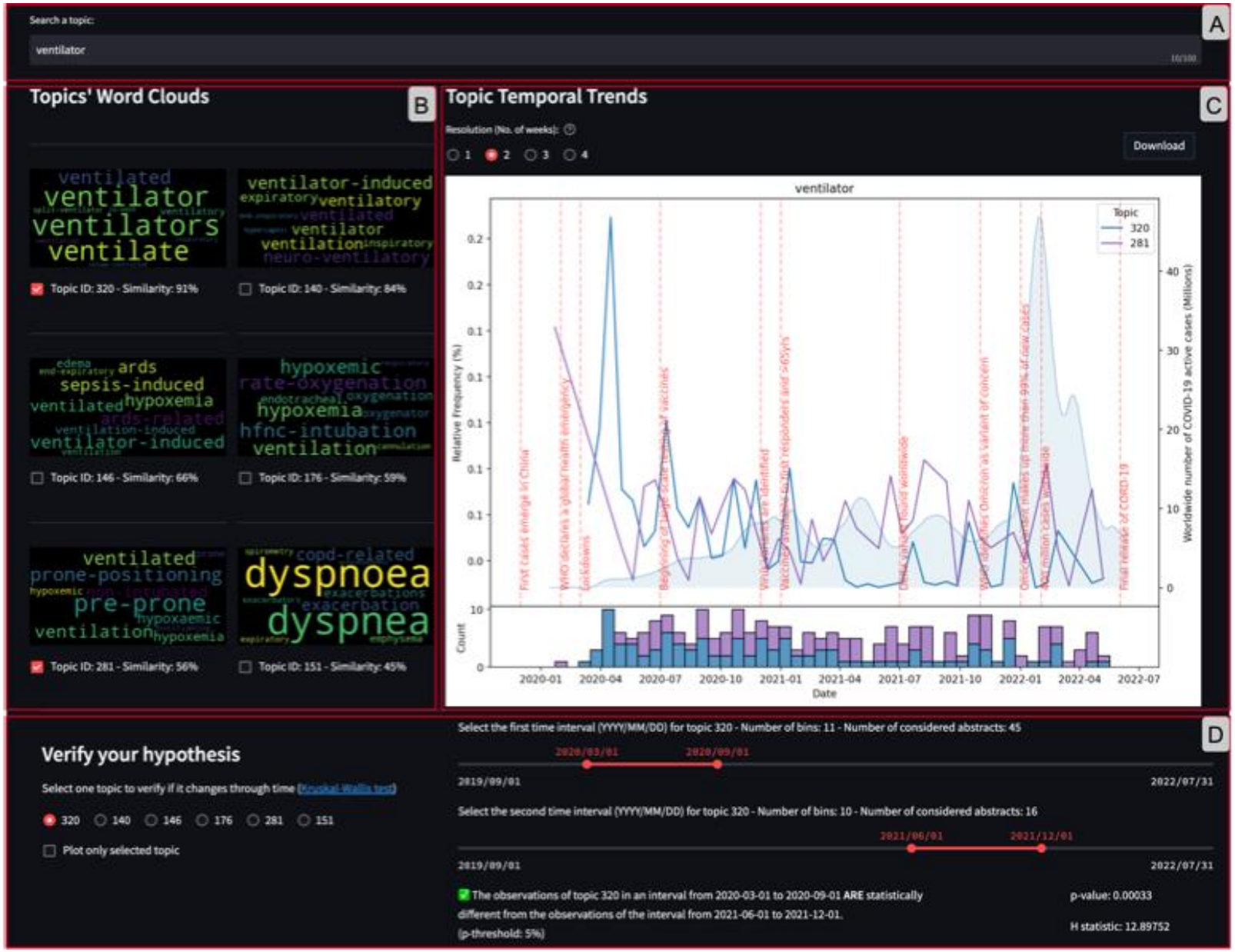
COVID-19 RELOADED

Five topic clusters in CORD-19



Our tool for reading the history of COVID-19: CORToViz dashboard

http://gmql.eu/cortoviz



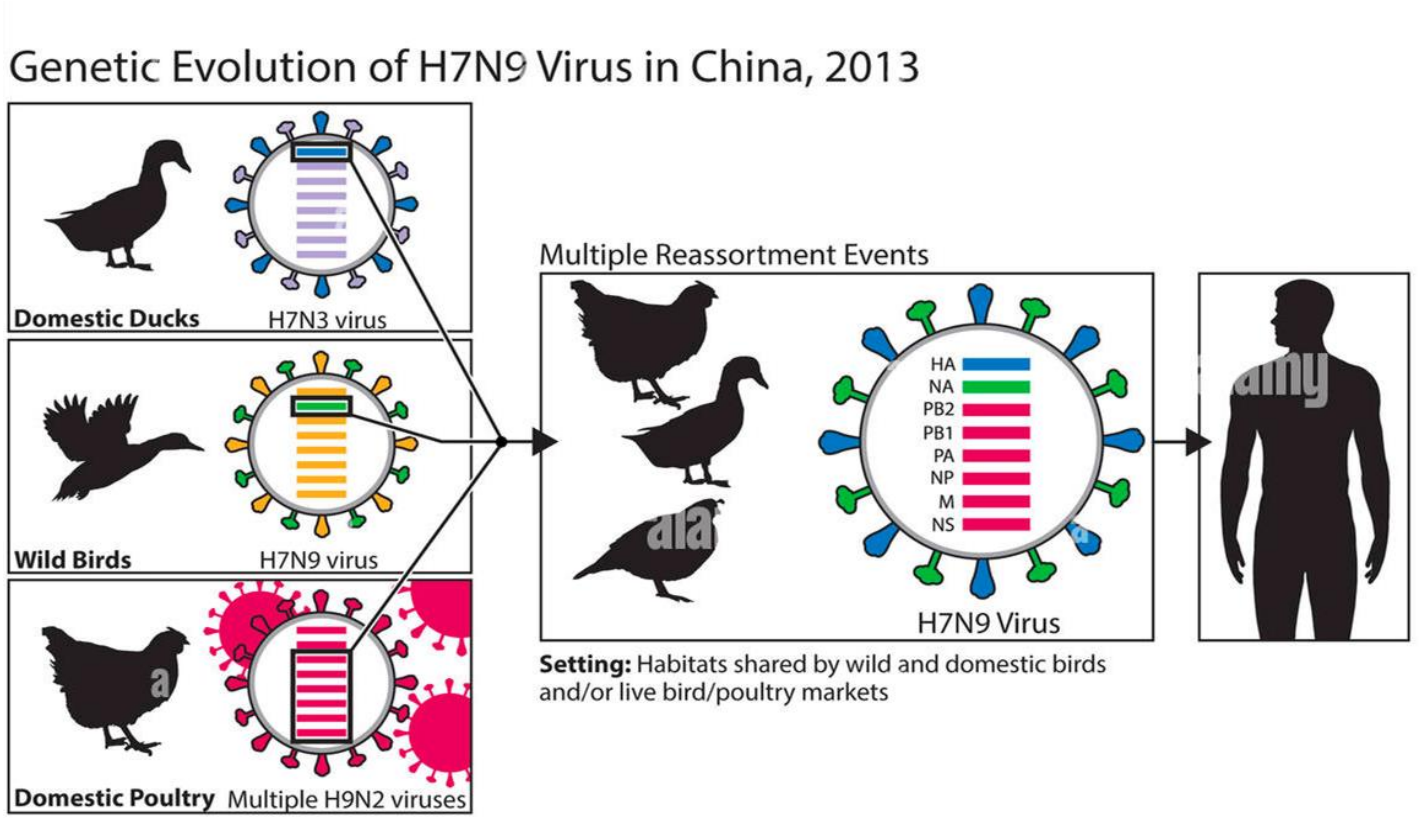
FUTURE RESEARCH HIGHLIGHTS

Recombination & reassortment in the Influenza virus

Source of variation in Influenza

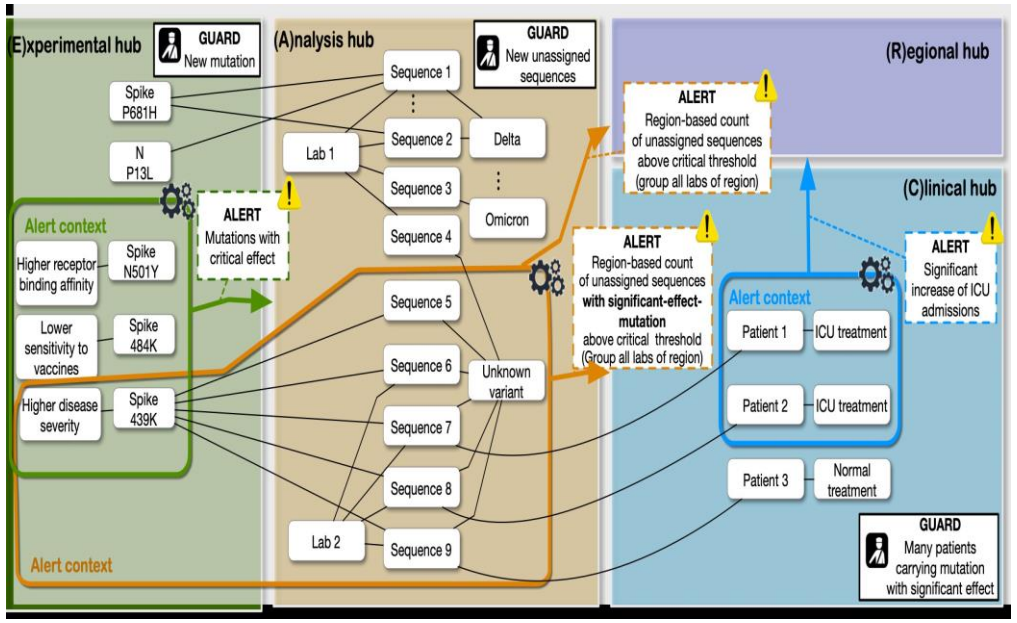
- Reassortment
  (exchange of genome segments)
- Intra-segment recombination

- Joint work with Ilaria Capua (virologist)



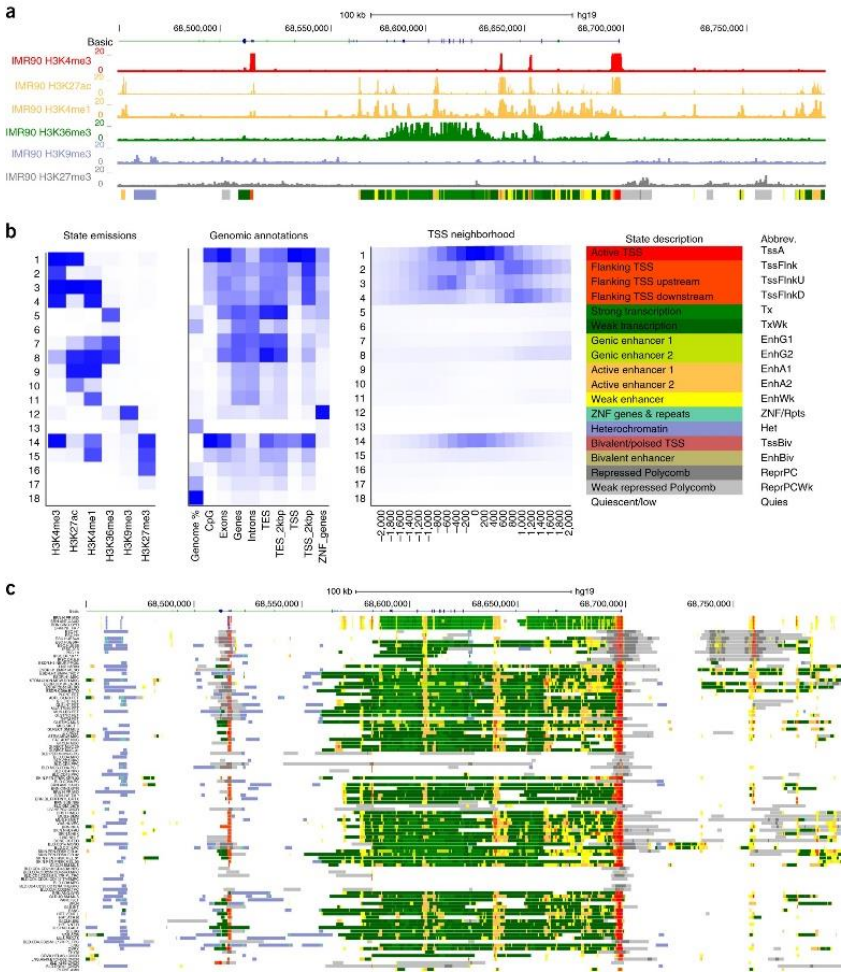
- Reactive Knowledge Management for COVID and Beyond

Highlights



- Partitioned knowledge graphs, partitions assigned to knowledge hub representing scientific communities/domains
- Reactive computations across domains make them aware of critical conditions
- Core work in knowledge management technology

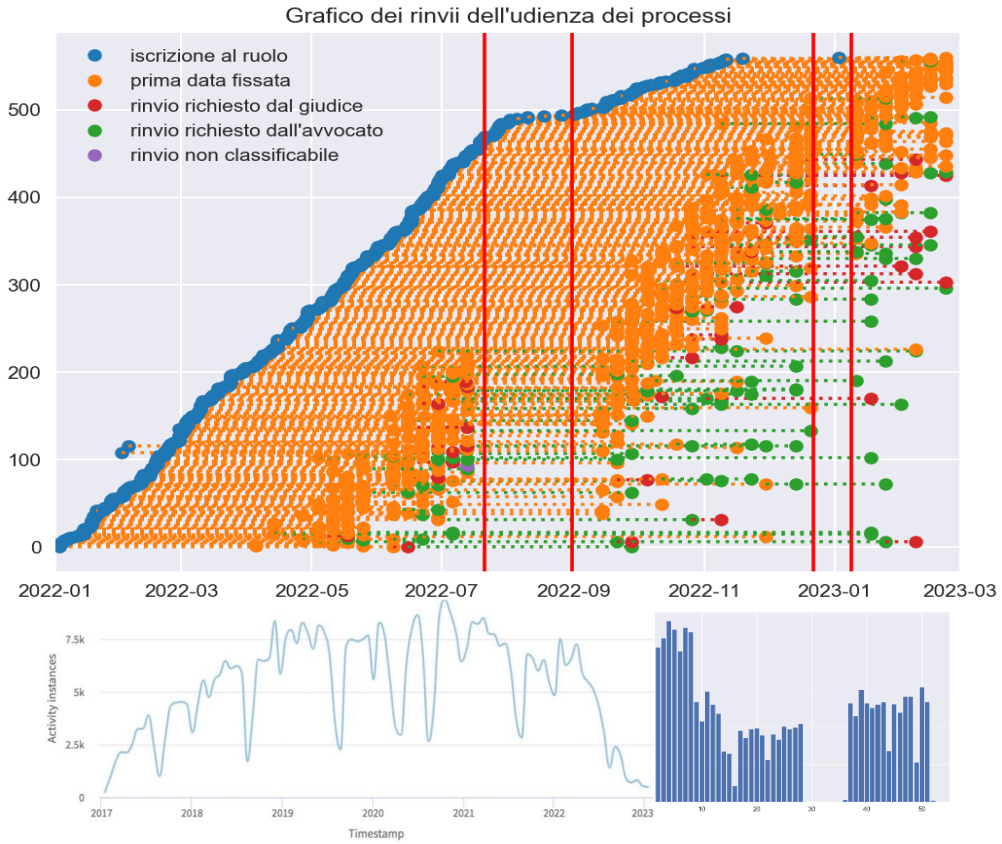
Sound of Genome



Highlights

- Regular structure of the genome (chromosomes, compartments, topological domains, loops)
- Already sonified (starting from the Integrated Genome Browser)
- Will the sound be useful, e.g. for recognizing cancer? Will it be nice to hear (e.g. Xenakis)?
- Project with: G. Haus (Musical informatics), A. Sarti (sound engineering), F. Avanzini, L. Nanni, P.Pinoli (genomics), F. Invernici, A. Bernasconi (Data Scientists) L. Francesconi (composer)

Time of Justice



Highlights

- Within Next Generation UPP, understand the times of justice (here: recent events in “Corte d’Appello”, Milano, courtesy of Barbara Pernici, after extremely complicated process analysis)
- We now have a large data warehouse upon which we will do data mart / data cube / machine learning for duration prediction
- Project with: B. Pernici, A. Campi, and several students
- No relationship to life science (but it is funded by NextGenEU and super-intriguing…)

GeCo – Viral Team