UNIVERSITÀ DEGLI STUDI DI MILANO
DIPARTIMENTO DI BIOSCIENZE

DBS

SARS-CoV-2 e VOC: aspetti virologici di interesse clinico-epidemiologico



Matteo Chiara
(matteo.chiara@unimi.it)
Bioinformatics Evolution & Comparative Genomics

Outline

- Evolution and viruses
- “Genomic surveillance” and the hunt for SARS-CoV-2 variants
- How we identify novel variants of SARS-CoV-2
- Limitations/issues and mitigation strategies

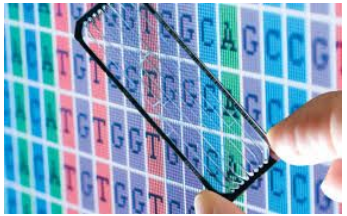
Evolution in viruses

- Viruses can **evolve** (and "improve") just like cell-based living organisms,
- Evolution is a trial and error process: most of the times changes have a little impact and/or are deleterious
- However, sometimes viruses might be lucky and get **better features**
- In the case of pathogens getting better might mean:
  - **evade** the immune system of the host
  - become more **infective**
  - become more **virulent**



Genomic surveillance of pathogens

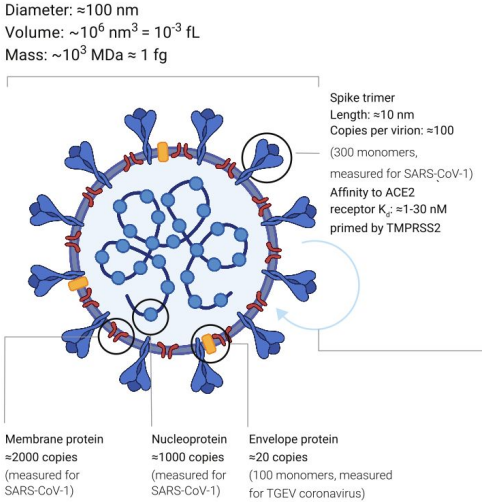
- Since we can not observe viruses directly, to understand what they are up to our best option is to sequence their **genome**
- Advantageous changes become very **frequent** and very **rapidly**
- During the pandemic we have **sequenced** more than **10M** different isolates of SARS-CoV-2
- To make of these data we use GENOMICS



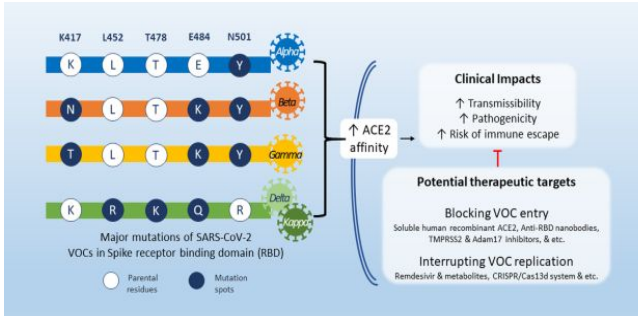


SARS-CoV-2 “genomic surveillance”

- Understand SARS-CoV-2 evolution
- **Compare** genomes at different times/in different countries to identify “novel” strains
- Find differences (genetic variants) and **define viral variants**
- Are novel variants associated with novel/advantageous **<<traits>>**?
- From Jan 2020 more than 10M genomic sequences!



USE Case: tracking dangerous variants





- International health authorities (WHO,ECDC,CDC) compile a list of dangerous variants (see ECDC for example)
- **Genomic surveillance=** sequence at least 5% of naso-pharingeal swabs from SARS-CoV-2 positive individuals
- Tracking: analyse and share the data, annotate and follow the variants

Dangerous variants: WHO

- VOC: variant of **Concern**
  - Increase in transmissibility or detrimental change in COVID-19 epidemiology; OR
  - Increase in virulence or change in clinical disease presentation; OR
  - Decrease in effectiveness of public health and social measures or available diagnostics, vaccines, therapeutics.

Currently circulating variants of concern (VOCs):

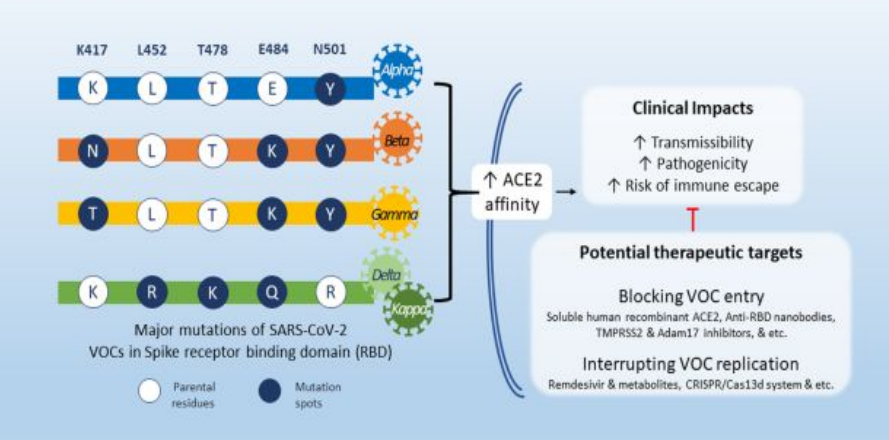
WHO label	Pango lineage•	GISAID clade	Nextstrain clade	Additional amino acid changes monitored°	Earliest documented samples	Date of designation
Delta	B.1.617.2	G/478K.V1	21A, 21I, 21J	+S:K417N +S:E484K	India, Oct-2020	VOI: 4-Apr-2021 VOC: 11-May-2021
Omicron*	B.1.1.529	GR/484A	21K, 21L, 21M	+S:R346K +S:L452R/Q +S:F486V	Multiple countries, Nov-2021	VUM: 24-Nov-2021 VOC: 26-Nov-2021

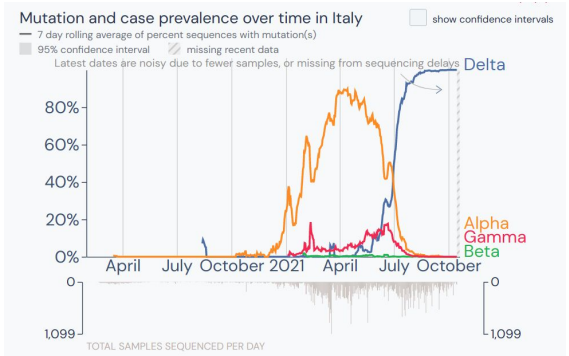
Previously circulating VOCs:

WHO label	Pango lineage•	GISAID clade	Nextstrain clade	Earliest documented samples	Date of designation
Alpha	B.1.1.7	GRY	20I (V1)	United Kingdom, Sep-2020	VOC: 18-Dec-2020 Previous VOC: 09-Mar-2022
Beta	B.1.351	GH/501Y.V2	20H (V2)	South Africa, May-2020	VOC: 18-Dec-2020 Previous VOC: 09-Mar-2022
Gamma	P.1	GR/501Y.V3	20J (V3)	Brazil, Nov-2020	VOC: 11-Jan-2021 Previous VOC: 09-Mar-2022

How do WE Identify "dangerous" variants?

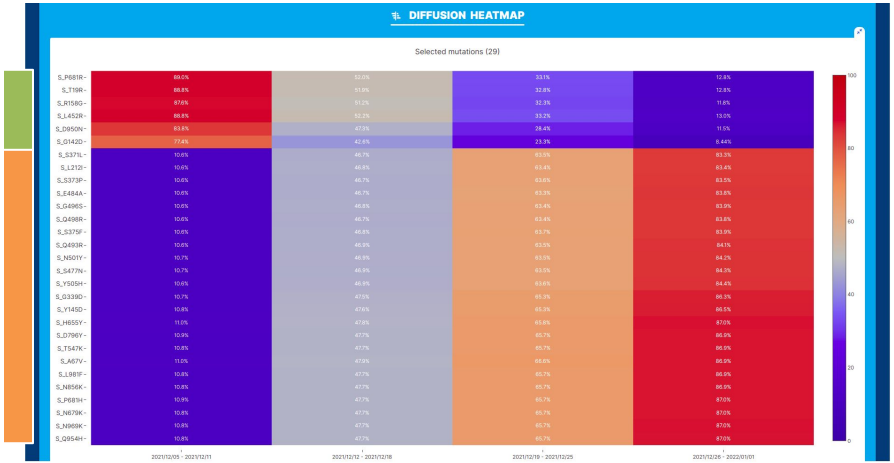
Mutations

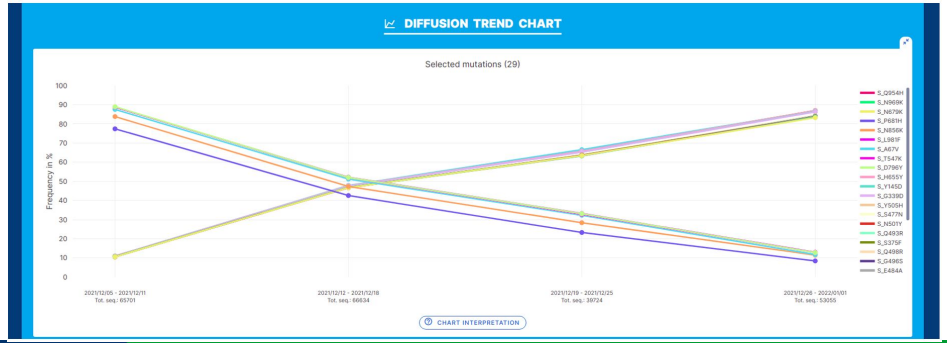




- Dangerous variants have an advantage over other variants, **hence they spread more rapidly**
- **This happened repeatedly for the 5* current variants of concern (VOC)**
- We identify dangerous variants retrospectively: i.e track the variant, see what happens

Tracking novel "dangerous" variants of SARS-COV-2





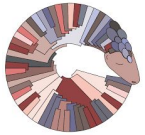
- Novel variants= Novel "mutations" in the genome
- If "growth advantage", novel mutations will eventually replace the old ones!
- VOC= novel viral alleles

Pinoli et al, submitted. See: https://virological.org/t/using-varianthunter-for-detecting-emerging-sars-cov-2-variants/797

Variant "hunting" starts with classification

- Normally a **single** mutation **does not change** the property of a virus significantly
- To track novel variants of the virus we need to observe and **track "co-occuring" of mutations**
- i.e. Do viruses that have specific combinations in their genome get better?

Different tools/methods to name/track variants

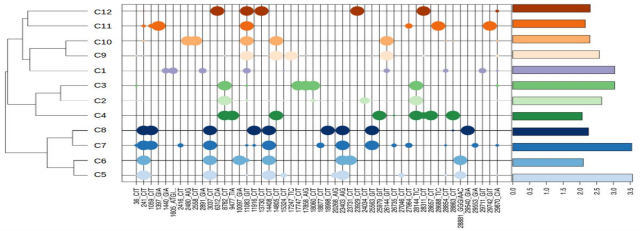


Pango. Rambaut et al

Nextstrain

Nextstrain. Hadfield et al

Again we need computational tools for this task: we currently have thousands of variants of SARS-CoV-2. Only **a few are considered dangerous**



HaploCoV. Chiara et al

Classification. How?

- To **understand**/study SARS-CoV-2 we need to develop systems that can identify relevant **"viral types" or "lineages"**

  - **Nextstrain*:** Only clades that reach a prevalence of 20% or higher. Based on phylogeny. Currently: 26 "clades". Each clade > 20.000 genomes

    * Hadfield et al, Bioinformatics (2018)

  - **Pangolin: Rambault** et al 2020.** Lineage: group of 5 or more genomes with 1 or more differences with their "ancestor". 2210 lineages in total. Each lineage has 80 genomes on average

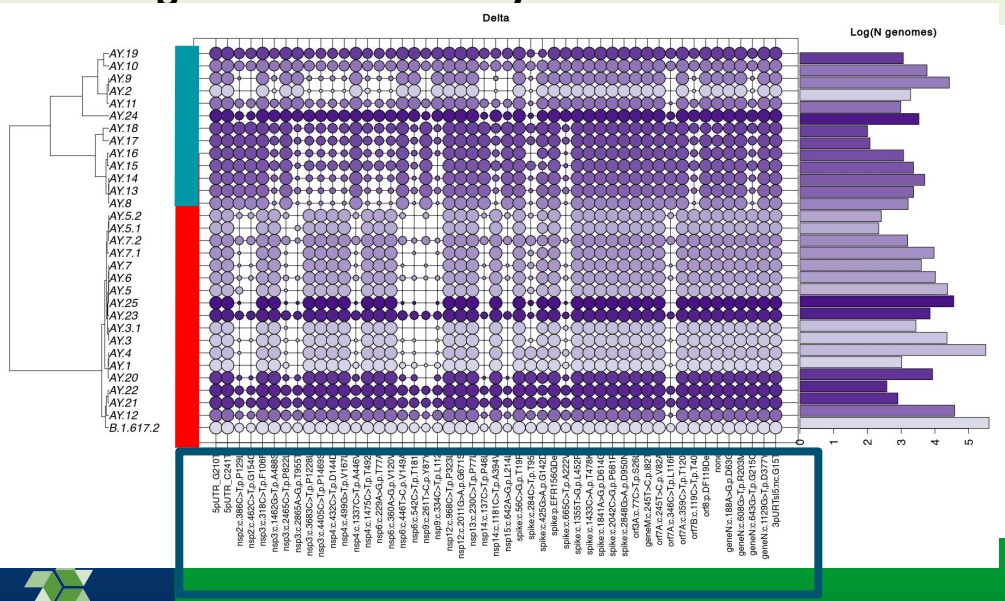
    ** Rambault et al, Nat Microb (2020)

  - **HaploCov: Chiara et al 2021.** Only mutations that reach a local frequency above 1%. Each group has at least 2 characteristic mutations. 1670 groups. Each group has 400 genomes on average

HaploCoV: a novel approach to classification

A simple method for tracking genome diversity in SARS-CoV-2:

- Variants with **high prevalence** (AF≥1%), for a long interval of time (>30 days)
- **Integrates with other systems**



Prioritization?

- **predict dangerous variants?**

ACCEPTED MANUSCRIPT

Comparative genomics reveals early emergence and biased spatio-temporal distribution of SARS-CoV-2

Matteo Chiara, David S Horner, Carmela Gissi, Graziano Pesole Author Notes

Molecular Biology and Evolution, msab049, https://doi.org/10.1093/molbev/msab049
Published: 19 February 2021

δ+ variant

doi:10.1093/molbev/msab049

Prioritization: how do we identify hints of adaptation?

- A specific **variant** can become "highly prevalent" for different reasons, including drift, bottlenecks and founder effects

- Functional annotation of **Mutations**:
  - What is the "predicted" effect?
  - Is the mutation recurrent (different lineages)?
  - Is it associated with known functional elements?
  - **Experimental Validation:** can we associate a real effect?

How do we Identify "Interesting" Mutations?

- Computer algorithms can **predict**/evaluate the effect of mutations *in silico*
- Since we know the genetic code, we can predict the effect of a mutation on the protein sequence
- We also know the biology of the virus, and we know which proteins are more important
- CorGAT, additional annotations. Regulatory elements. Epitopes. Prevalence worldwide (and regional)

Using this knowledge we can **build computer systems** that evaluate, flag and identify "interesting" mutations

How do we Identify “Interesting” Variants?

- Different lines of evidence/resources are integrated

HyPhy

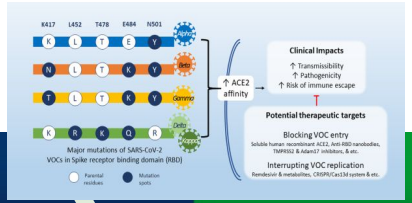
- Evolution/ selection

IMMUNE EPITOPE DATABASE AND ANALYSIS RESOURCE

- Viral Epitopes

UniProt

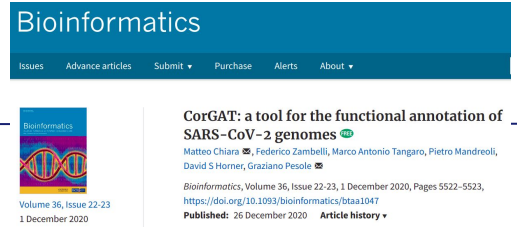
- Functional domains



- VOC and prevalence

CorGAT

evolut. patterns


Bioinformatics

Issues Advance articles Submit Purchase Alerts About

Volume 36, Issue 22-23
1 December 2020

CorGAT: a tool for the functional annotation of SARS-CoV-2 genomes

Matteo Chiara, Federico Zambelli, Marco Antonio Tangaro, Pietro Mandreoli, David S Horner, Graziano Pesole

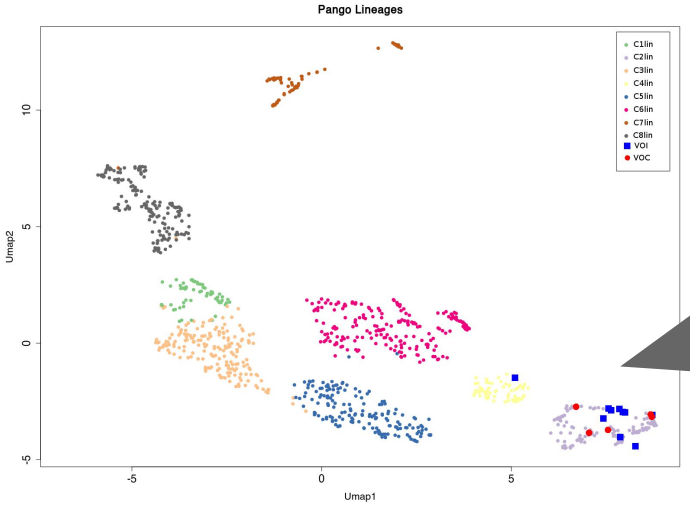
Bioinformatics, Volume 36, Issue 22-23, 1 December 2020, Pages 5522–5523, https://doi.org/10.1093/bioinformatics/btaa1047

Published: 26 December 2020 **Article history**

doi:10.1093/bioinformatics/btaa1047

CorGAT+HaploCoV: how do dangerous variants differ?

- Dimension-reduction analyses identify distinct clusters associate with "**dangerous**" variants (VOC,VOI,VUM)

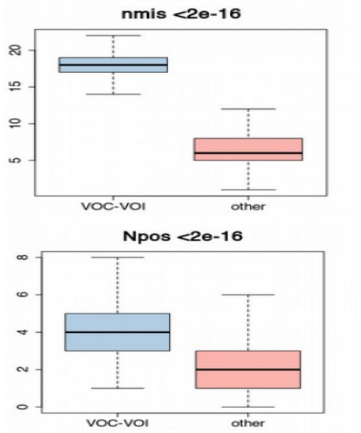


All the "dangerous" variants have some features in common.

and "cluster" in the same group



doi.org/10.1101/2021.06.23.449558



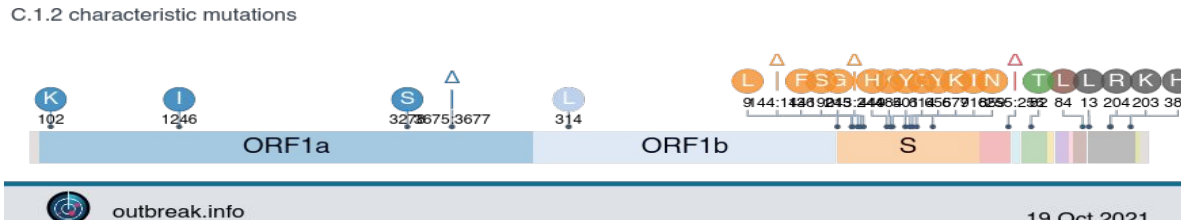
Prioritization of “dangerous” variants

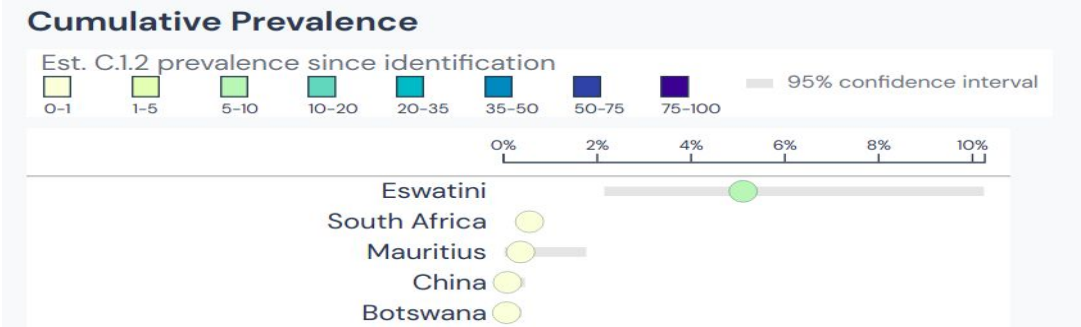
- Features integrated into a linear score
- With a simple model we identify **>93%** of dangerous variants (flagged by international health authorities)
- Plus c.a 20 additional lineages, with similar mutational profiles (not flagged by health authorities)
- The score can also identify/pinpoint novel sub-variants with increased “VOC-potential”
  - currently we identify 5 within Omicron





C.1.2

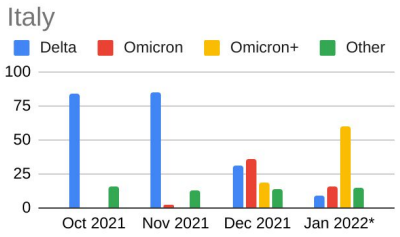




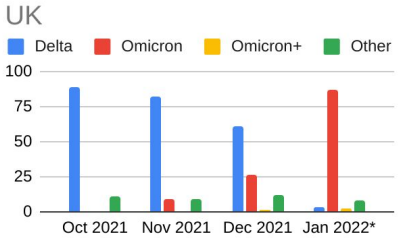
- First observed in May 2021
- VUM in Sept 2021
- Prioritized by scoring system in c.a 1 second

Omicron Variant, Evolution

Italy (12418)	Oct 2021	Nov 2021	Dec 2021	Jan 2022*
Delta	84	85	31	9
Omicron	0	2	36	16
Omicron+	0	0	19	60
Other	16	13	14	15



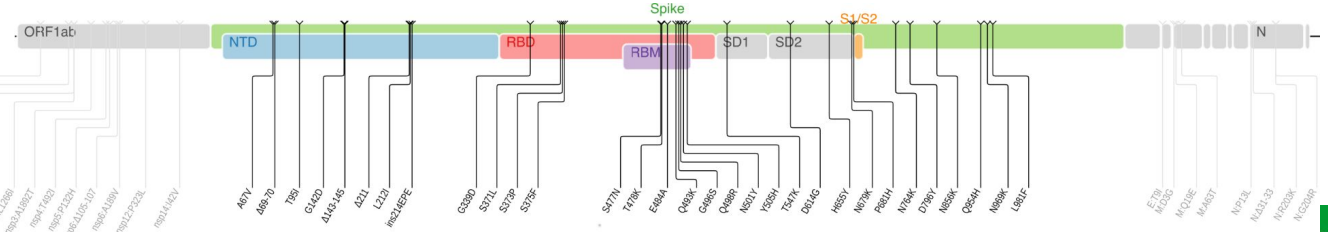
UK (183112)	Oct 2021	Nov 2021	Dec 2021	Jan 2022*
Delta	89	82	61	3
Omicron	0	9	26	87
Omicron+	0	0	1	2
Other	11	9	12	8



- I.e lack of mutations is due to errors in sequencing/assembly
- ARCTIC V4 particularly popular in the UK (apparently)

Sequencing artifact

S:K417N
S:N440K
S:G446S



Omicron sub-variants

- 5 sub-variants of Omicron predicted to be than parental “more dangerous”:
  - BA.2.12 and BA.2.12.1
  - BA.2.16
  - BA.4
  - BA.5

- 3 “validated” by recent reports



Publication Preprint

BA.2.12.1, BA.4 and BA.5 escape antibodies elicited by Omicron infection

Yunlong Richard Cao, Ayijiang Yisimayi, Fanchong Jian, Weiliang Song, Tianhe Xiao, Lei Wang, Shuo Du, Jing Wang, Qianqian Li, Xiaosu Chen, Peng Wang, Zhiying Zhang, Pulan Liu, Ran An, Xiaohua Hao, Yao Wang, Jing Wang, Rui Feng, Haiyan Sun, Lijuan Zhao, Wen Zhang, Dong Zhao, Jiang Zheng, Lingling Yu, Can Li, Na Zhang, Rui Wang, Xiao Niu, Sijie Yang, Xuetao Song, Linlin Zheng, Zhiqiang Li, Qingqing Gu, Fei Shao, Weijin Huang, Ronghua Jin, Zhongyang Shen, Youchun Wang, Xiangxi Wang, Junyu Xiao, and Xiaoliang Sunney Xie view affiliations

bioRxiv

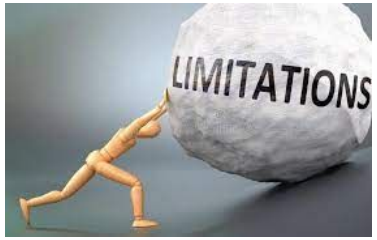
published 02 May 2022 • accessed 10 May 2022

Record provided by bioRxiv bioRxiv Learn more

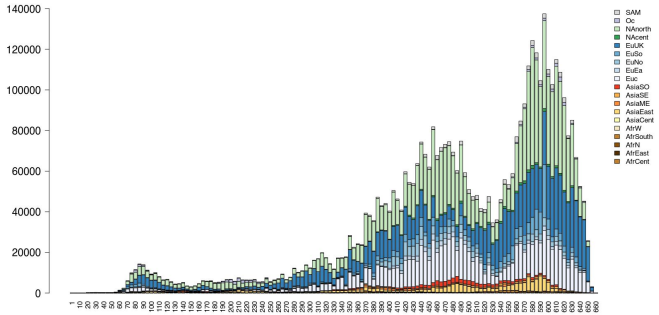
The recently emerged SARS-CoV-2 Omicron sublineages BA.2.12.1, BA.2.13, BA.4 and BA.5 all contain L452 mutations and show potential higher transmissibility over BA.2.1. The new variants receptor binding and immune evasion capability require immediate investigation, especially on the role of L452 substitutions. Herein, coupled with structural comparisons, we show that BA.2 sublineages, including BA.2.12.1 and BA.2.13, exhibit increased ACE2-binding affinities compared to BA.1; while BA.4/BA.5 displays the weakest receptor-binding activity due to F486V and R493Q reversion. Importantly, compared to BA.2, BA.2.12.1 and BA.4/BA.5 exhibit stronger neutralization evasion against the plasma of 3-dose vaccinees and, most strikingly, of vaccinated BA.1 convalescents. To delineate the underlying evasion mechanism, we

Genomic surveillance: limitations

- WHO and other international health authorities recommended to sequence **between 5 and 10%** of all positive samples
- Currently most countries do not follow these recommendations
- **Data sharing is still an issue! Only c.a. 1 out of 4 genome sequences are available from open access repositories**



Missing data == Missing VOC/VOI/VUM?



Sampling bias:

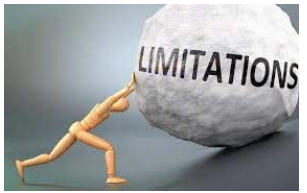
- >50% data UK or USA
- less 5% data Africa
- >70% data EU



Genomic surveillance: limitations

- **only c.a 1 out of 4 genome sequences are available from open access repositories**
- **~90% of the “open access sequences” are from the U.S.A**

Missing data == Missing VOC/VOI/VUM?



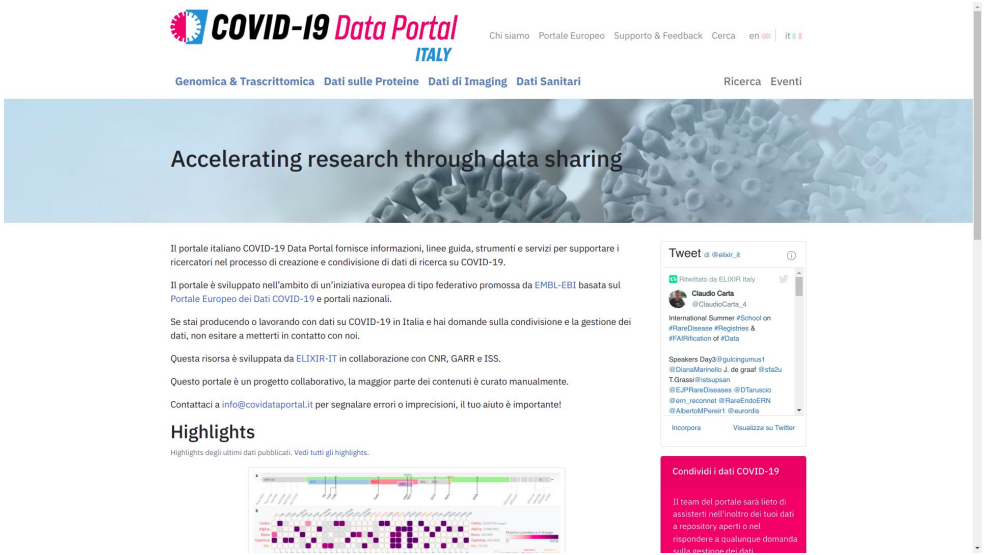
Getting data to Scientists: the COVID-19 data-portal

- Unprecedented amounts of **data** were produced during the COVID-19 pandemic
- Making this data **available** and accessible is a fundamental prerequisite to advance our knowledge
- The EU has launched and international initiative to pro promote best practices for data sharing and curation: COVID-19 data portal
- We currently run the Italian instance



European Commission

https://www.covid19dataportal.it/





EMBL-EBI

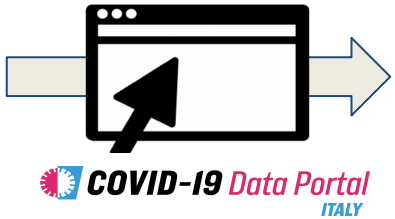
elixir

Why do we need a data portal in Italy https://www.covid19dataportal.it/

Research coordination issues within the Country

- Clinical research coordinated within region/institution
- Lack of national research facilities
- Limited Open Science, DM/DS practices awareness
- Inefficient efforts duplications
- Lack of dedicated funding







- Stimulate coordination between institutions
- Increase best practice implementation
- Rise Open Science, DM/DS practices awareness
- Increase use of national available resources
- Support joint grant applications

Increase coordination

Data portal, **training** and dissemination









Conclusions

- Genomics is a very powerful tool for studying/understanding evolution
- By sequencing the genomes of SARS-CoV-2 we can identify and monitor "dangerous" or more aggressive forms of the virus
- This requires bioinformatics tools and methods
  - Currently available methods are very good for tracking.
    But **Limited data**
  - Currently available methods are very good for tracking.
  - But **Biased Sampling**
- "Predictive" tools. Can identify relevant variants
  - Sampling however remains an issue: **we need to see variants!**

Credits

Publications

Cilibrasi L, et al. . ViruClust: direct comparison of SARS-CoV-2 genomes and genetic variants in space and time . *Bioinformatics*. 2022;btac030. doi:10.1093/bioinformatics/btac030

Tangaro MA,, et al. Laniakea@ReCaS: exploring the potential of customisable Galaxy on-demand instances as a cloud-based service. *BMC Bioinformatics*. 2021;22(Suppl 15):544. Published 2021 Nov 8. doi:10.1186/s12859-021-04401-3

Oranger, et al. Accurate detection and quantification of SARS-CoV-2 genomic and subgenomic mRNAs by ddPCR and meta-transcriptomics analysis. *Commun Biol*. 2021;4(1):1215. Published 2021 Oct 22. doi:10.1038/s42003-021-02748-0

Chiara M, et al. Comparative Genomics Reveals Early Emergence and Biased Spatiotemporal Distribution of SARS-CoV-2. *Mol Biol Evol*. 2021;38(6):2547-2565. doi:10.1093/molbev/msab049

Chiara M, et al. G. CorGAT: a tool for the functional annotation of SARS-CoV-2 genomes [published online ahead of print, 2020 Dec 21]. *Bioinformatics*. 2020;btaa1047. doi:10.1093/bioinformatics/btaa1047

Chiara M,et al. Next generation sequencing of SARS-CoV-2 genomes: challenges, applications and opportunities. Brief Bioinform. 2021 Mar 22;22(2):616-630. doi: 10.1093/bib/bbaa297. PMID: 33279989; PMCID: PMC7799330.

Chiara et al. Unsupervised classification of SARS-CoV-2 genomic sequences uncovers hidden genetic diversity and suggests an efficient strategy for genomic surveillance. biorXiv

Funding

(H2020) ELIXIR-CONVERGE GA: 871075

(H2020) B1MG GA:951724

(HE) ByCOVID GA: 101046203

European Commission

Dissemination

https://sibbm.zanichelli.it

https://www.covid19dataportal.it/it/highlights/

CREDITS

(Department of Biosciences, University of Milan)



Matteo Chiara



David S. Horner



Giulio Pavesi



Federico Zambelli



Daniele Traversa



Pietro Mandreoli

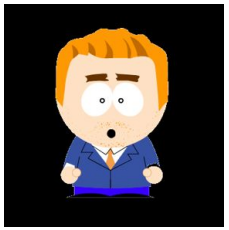


Erika Ferrandi

(Department of Biosciences, Biotechnology and Biopharmaceutics, University of Bari “A. Moro”)



Carmela Gissi



Graziano Pesole



Anna Maria D’Erchia



Caterina Manzari



Ernesto Picardi



Marco A. Tangaro

Credits (data portal)

II Team dell' Italian COVID-19 Data Portal:

Head of Node ELIXIR-IT
Prof. Graziano Pesole - Università di Bari e CNR-IBIOM

Coordinatore Tecnico ELIXIR-IT
Dott. Federico Zambelli - Università di Milano e CNR-IBIOM

Coordinatore di Progetto
Dott. Marco Carraro - Università di Padova

Genomic and Transcriptomic Data
Responsabile: Dott. Matteo Chiara - Università di Milano e CNR-IBIOM
Dott. Federico Zambelli - Università di Milano e CNR-IBIOM
Dott. Flavio Vito Licciulli - CNR-ITB
Dott. Marco Beccuti - Università di Torino
Dott.sa Tiziana Castrignanò - Università della Tuscia

Protein Data
Responsabile: Dott. Marco Carraro - Università di Padova
Dott.sa Luana Licata - Università di Roma Tor Vergata
Dott.sa Letizia Bernardo - CNR-ITB

Imaging Data
Responsabile: Dott. Giuseppe D'Avenio - Istituto Superiore di Sanità
Dott. Evaristo Cisbani - Istituto Superiore di Sanità
Dott. Marco Beccuti - Università di Torino

Health Data
Responsabile: Dott. Giuseppe D'Avenio - Istituto Superiore di Sanità
Dott. Evaristo Cisbani - Istituto Superiore di Sanità
Dott. Marco Beccuti - Università di Torino

Implementazione Software
Responsabile: Dott. Flavio Vito Licciulli - CNR-ITB
Dott. Giorgio Grillo - CNR-ITB

Infrastruttura IT
Responsabile: Dott. Fulvio Galeazzi - Consortium GARR
Dott. Alberto Colla - Consortium GARR
Dott. Matteo Fazio - Consortium GARR

Content Manager
Erika Ferrandi - CNR-IBIOM

Comunicazione
Responsabile: Dott.sa Francesca de Leo - CNR-IBIOM
Dott.sa Diana Battistella - Università di Padova
Dott.sa Federica Tanlongo - Consortium GARR
Dott.sa Sara Di Giorgio - Consortium GARR



IBIOM
Istituto di Biomembrane, Bioenergetica e Biotecnologie Molecolari
CNR-IBIOM

CNR-ITB

Università di Milano

Università di Padova

Università di Roma Tor Vergata

Università di Torino

Università della Tuscia

Istituto Superiore di Sanità

Consortium GARR

Thanks!

@elixir_ita

www.elixir-europe.org

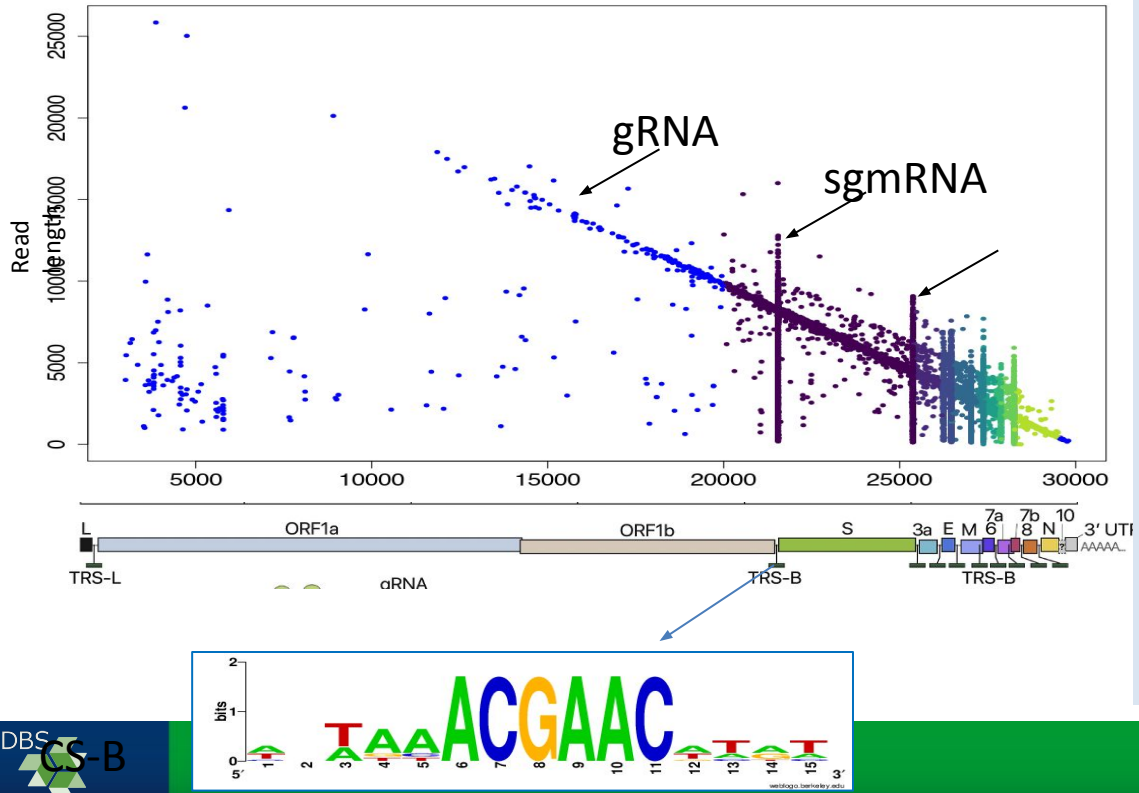

Any Questions?

Pango system in depth

- Pango/pangolin is currently considered the standard
- Each **new** lineage/group is identified by a letter (from A to Z)
- Each time a novel group is formed within an existing group a number from 1 to N is appended to name of the group.
  - i.e B is father to B.1
  - B.1 is father to B.1.1
  - and so on
- Only 4 levels of hierarchy are supported. I.e up to 1 letter and 3 numbers. When the 4th level is reached, the name gets “rebooted” and a new letter is used
  - B.1.617 is father to B.1.617.2
  - B.1.616.2 has a child. Hierarchy saturated! New name is rebooted to AY (and so on)

SARS-CoV-2 transcription

ONT RNA-seq data in Kim et al. (Cell, 2020) consisting of >250000 split reads (average L of 2794 nt) plotting read starts (x-axis) and lengths (y-axis)



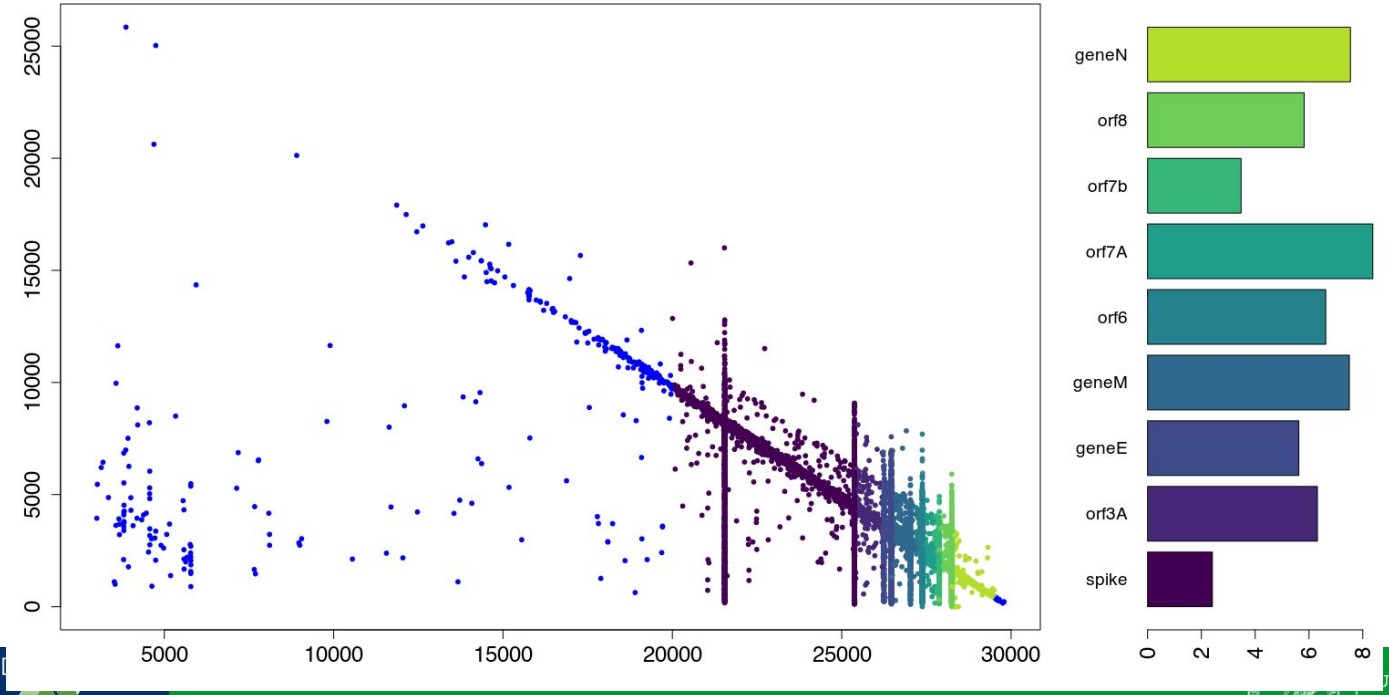
Alignment of sgRNA sequences identifies highly conserved repetitive sequence motifs (CS-B). CS-B has 9 occurrences in the genome: one in TRS-L, 8 (TRS-B) corresponding very well with the ATG of sgRNAs genes

A negative correlation is observed between the position of CS-B and the length of the associated transcript: most transcripts are a complete read-through from CS-B to the end of the genome.

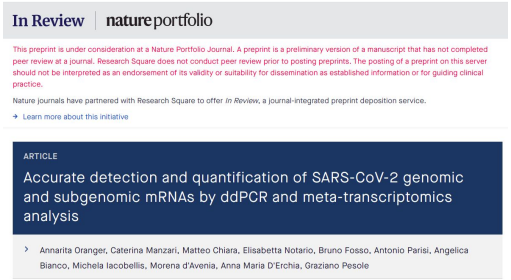
A pervasive transcription along the genome is clearly evident.

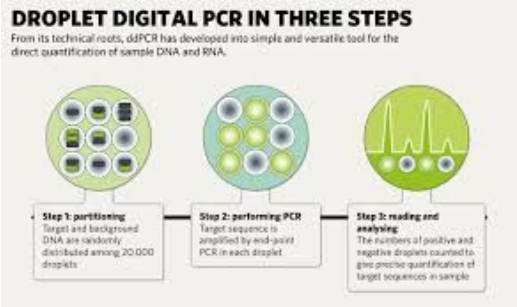
SARS-Cov-2 transcription

expression levels (log2 TPM) are indicated in the barplot. Each gene has a different color: the **spike** protein has the lowest “expression level” whereas **geneN** and **ORF7A** are the most expressed genes (almost 70x more than spike)

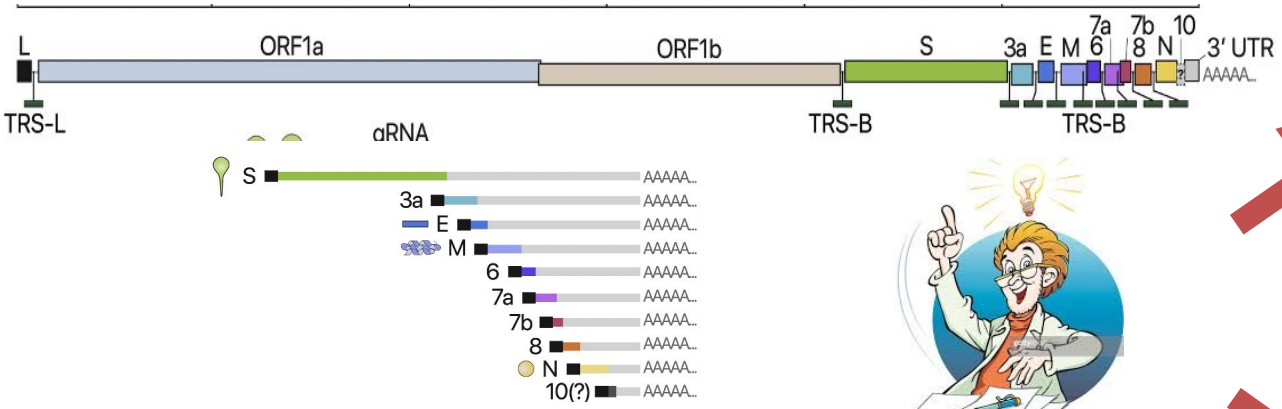


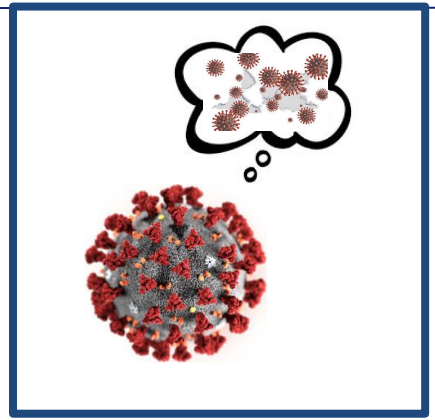
- Current methods for the detection of SARS-CoV-2 use PCR to detect the genome of the virus
- To <actively replicate> the virus need to use the genome to make proteins (**transcription**)
- We recently implemented a lab procedure than detect/quantify the transcripts of SARS-CoV-2
- Potentially improving the diagnosis
  - i.e not just **present/absent** but aso **active/not active** viruses

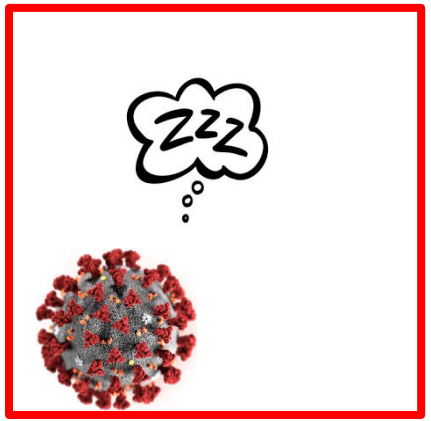




From https://www.nature.com/articles/d42473-020-00123-x







Assay based on transcripts. You know if virus is “alive” and replicating