Bart Keagy

The UCSC SARS-CoV-2 Genome Browser: One-stop Shopping for the Latest Molecular Details of SARS-CoV-2

David Haussler

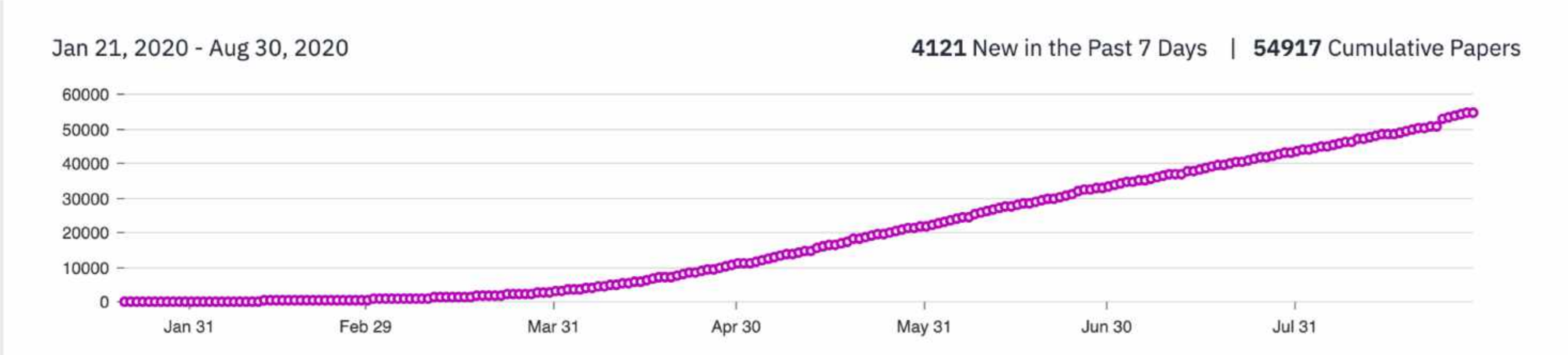
UC Santa Cruz Genomics Institute

https://genome.ucsc.edu/covid19.html

https://genome.ucsc.edu/cgi-bin/hgTracks?db=wuhCor1

University of California Santa Cruz Genomics Institute
hhmi Howard Hughes Medical Institute
University of California Office of the President
CITRIS Center for Information Technology Research in the Interest of Society
Schmidt Futures
NIH NHGRI

SARS-CoV-2 Research is Generating Data at an Astonishing Pace



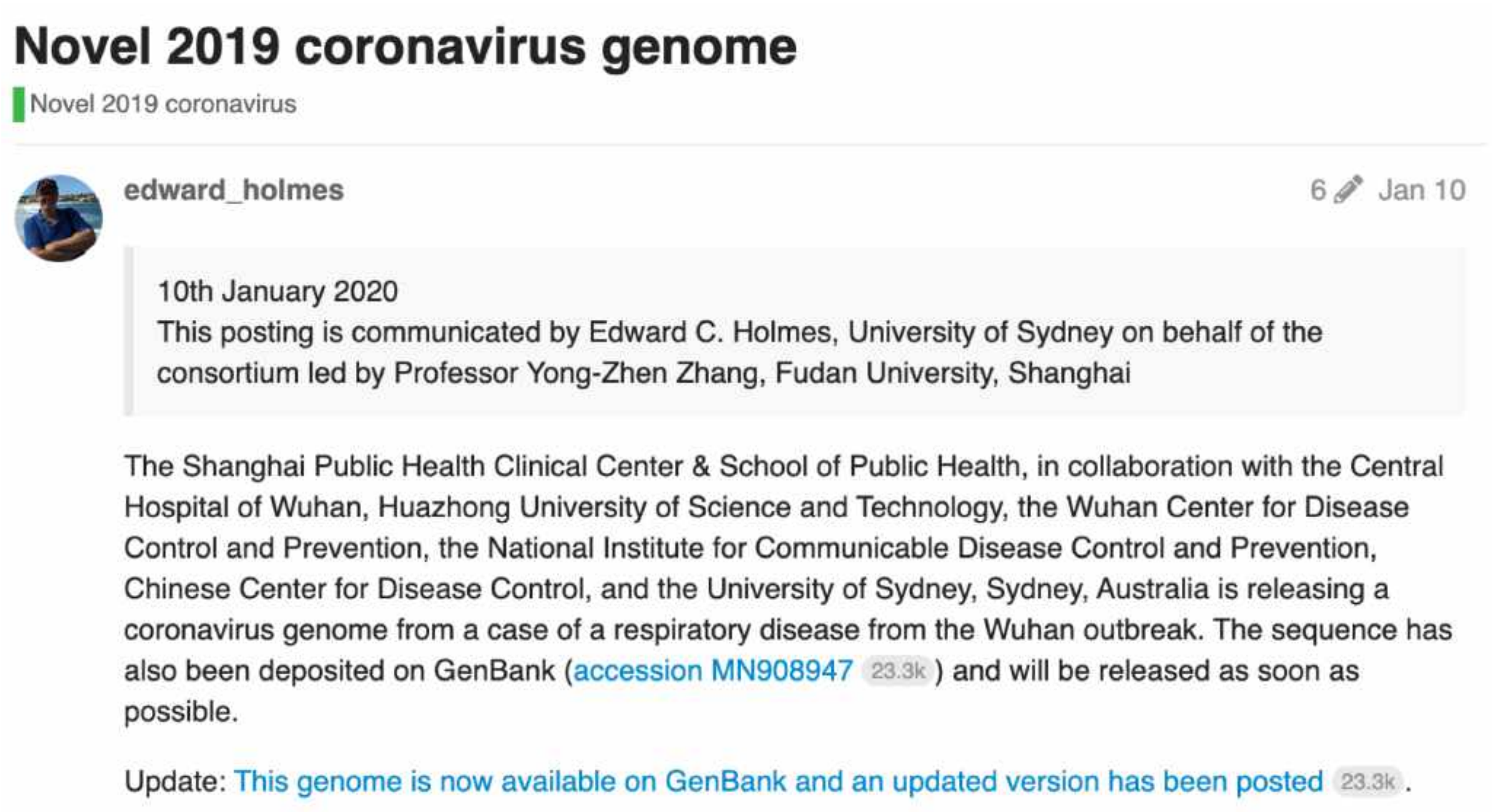
In April, SARS CoV-2 papers had a doubling time of ~14.5 days.

(The virus doubling time in April was ~7 days)

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

Source: primer.ai

Genomic Data has also grown at an exponential rate



Novel 2019 coronavirus genome

Novel 2019 coronavirus

edward_holmes

6 ✎ Jan 10

> 10th January 2020
> This posting is communicated by Edward C. Holmes, University of Sydney on behalf of the consortium led by Professor Yong-Zhen Zhang, Fudan University, Shanghai

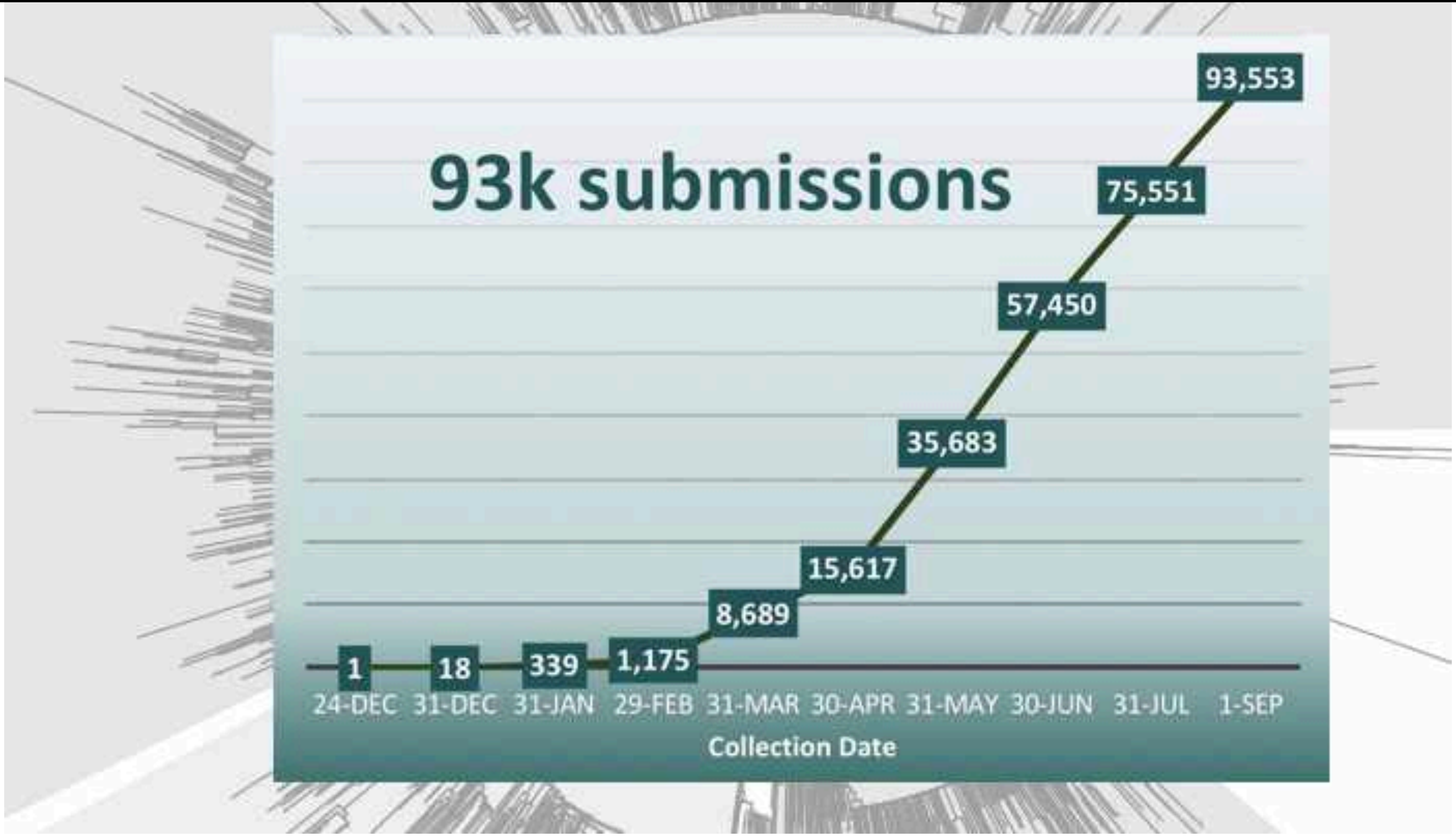
The Shanghai Public Health Clinical Center & School of Public Health, in collaboration with the Central Hospital of Wuhan, Huazhong University of Science and Technology, the Wuhan Center for Disease Control and Prevention, the National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control, and the University of Sydney, Sydney, Australia is releasing a coronavirus genome from a case of a respiratory disease from the Wuhan outbreak. The sequence has also been deposited on GenBank (accession MN908947 23.3k) and will be released as soon as possible.

Update: This genome is now available on GenBank and an updated version has been posted 23.3k .

First virus genome released on Jan 10, 2020

Source: virological.org

Genomic Data has also grown at an exponential rate

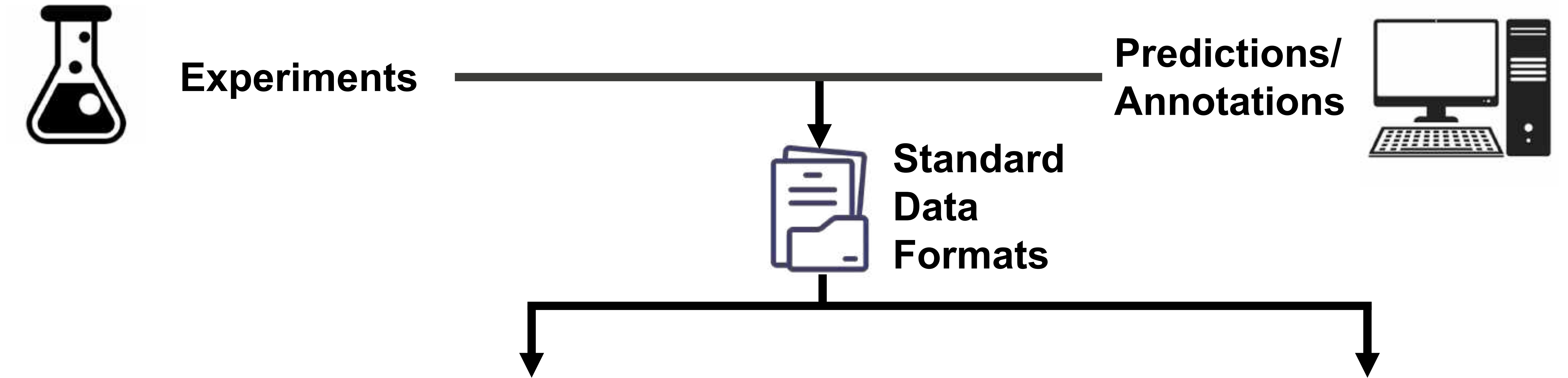


UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

More than 100,000 genomes now sequenced!

Source: GISAID

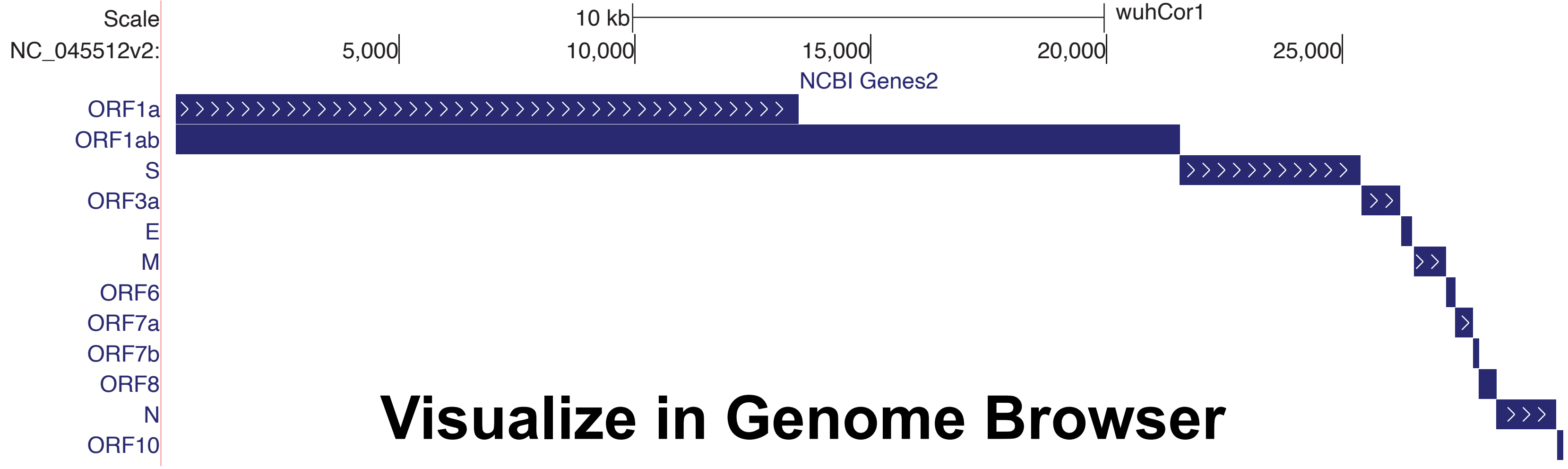
How do we make use of all this genomic and molecular data for analysis?



Experiments

Predictions/ Annotations

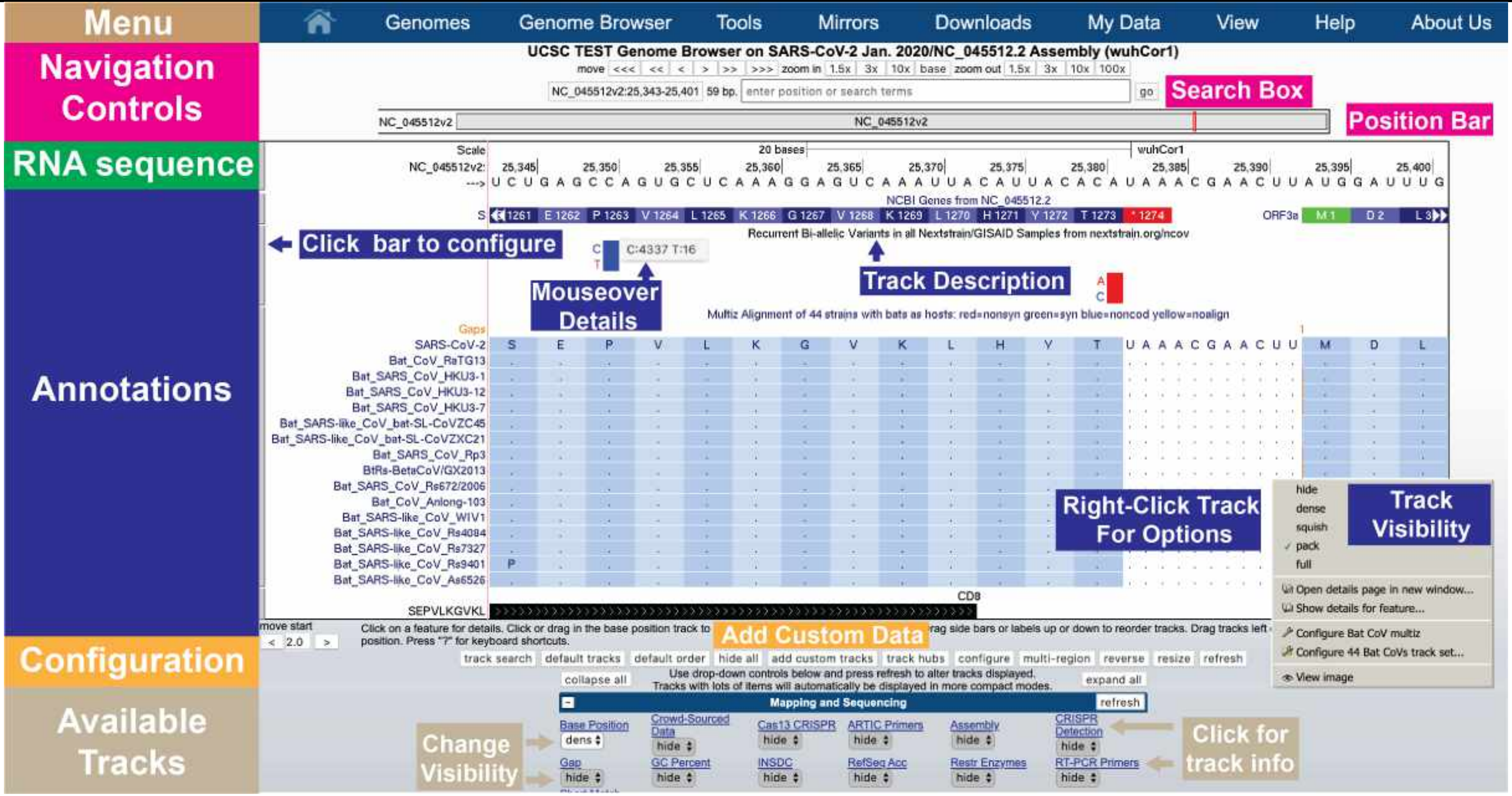
Standard Data Formats



Visualize in Genome Browser

Compare Datasets

The Genome Browser annotates nucleotides with information stored in tracks



UCSC TEST Genome Browser on SARS-CoV-2 Jan. 2020/NC_045512.2 Assembly (wuhCor1)

Click on a feature for details. Click or drag in the base position track to … position. Press "?" for keyboard shortcuts. … rag side bars or labels up or down to reorder tracks. Drag tracks left …

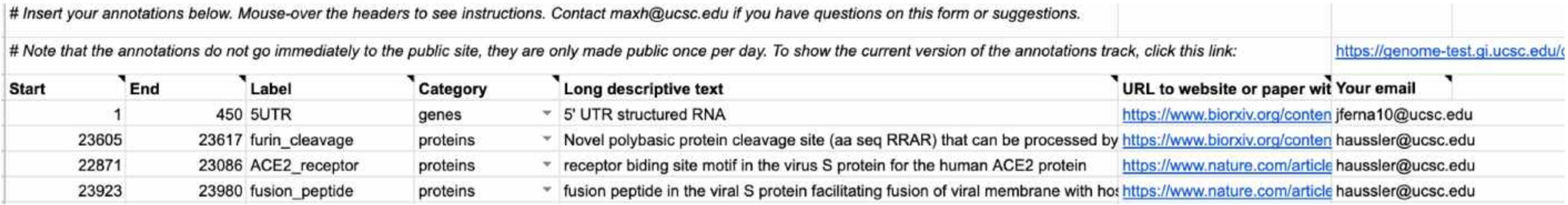
Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

http://genome.ucsc.edu/s/SARS_CoV2/Figure1

Users add annotations via “crowd-sourced” annotations

Users add annotations to spreadsheet at: http://bit.ly/cov2annots

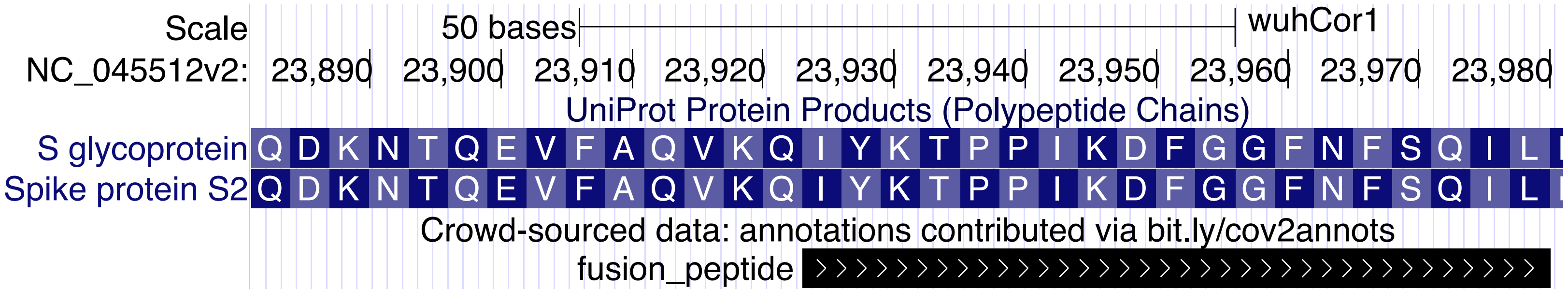


# Insert your annotations below. Mouse-over the headers to see instructions. Contact maxh@ucsc.edu if you have questions on this form or suggestions.

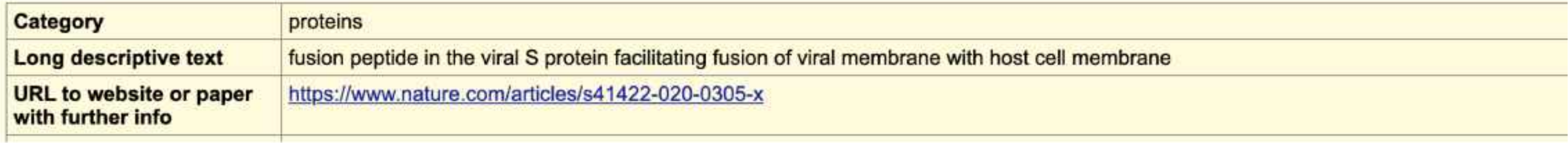
# Note that the annotations do not go immediately to the public site, they are only made public once per day. To show the current version of the annotations track, click this link: https://genome-test.gi.ucsc.edu/c

Start	End	Label	Category	Long descriptive text	URL to website or paper wit	Your email
1	450	5UTR	genes	5' UTR structured RNA	https://www.biorxiv.org/conten	jferna10@ucsc.edu
23605	23617	furin_cleavage	proteins	Novel polybasic protein cleavage site (aa seq RRAR) that can be processed by	https://www.biorxiv.org/conten	haussler@ucsc.edu
22871	23086	ACE2_receptor	proteins	receptor biding site motif in the virus S protein for the human ACE2 protein	https://www.nature.com/article	haussler@ucsc.edu
23923	23980	fusion_peptide	proteins	fusion peptide in the viral S protein facilitating fusion of viral membrane with ho:	https://www.nature.com/article	haussler@ucsc.edu

Annotation appears on genome.ucsc.edu after approval:


Scale
50 bases
wuhCor1
NC_045512v2: 23,890 23,900 23,910 23,920 23,930 23,940 23,950 23,960 23,970 23,980
UniProt Protein Products (Polypeptide Chains)
S glycoprotein Q D K N T Q E V F A Q V K Q I Y K T P P I K D F G G F N F S Q I L
Spike protein S2 Q D K N T Q E V F A Q V K Q I Y K T P P I K D F G G F N F S Q I L
Crowd-sourced data: annotations contributed via bit.ly/cov2annots
fusion_peptide

Click element for information



Category	proteins
Long descriptive text	fusion peptide in the viral S protein facilitating fusion of viral membrane with host cell membrane
URL to website or paper with further info	https://www.nature.com/articles/s41422-020-0305-x

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

How can we leverage genomic data and the UCSC Genome Browser to understand a pandemic?

1. **Genome sequencing to trace transmission events:**
   - Are we accurately reconstructing evolutionary relationships?

2. **Ensuring testing remains accurate:**
   - Are PCR primers detecting all isolates accurately?

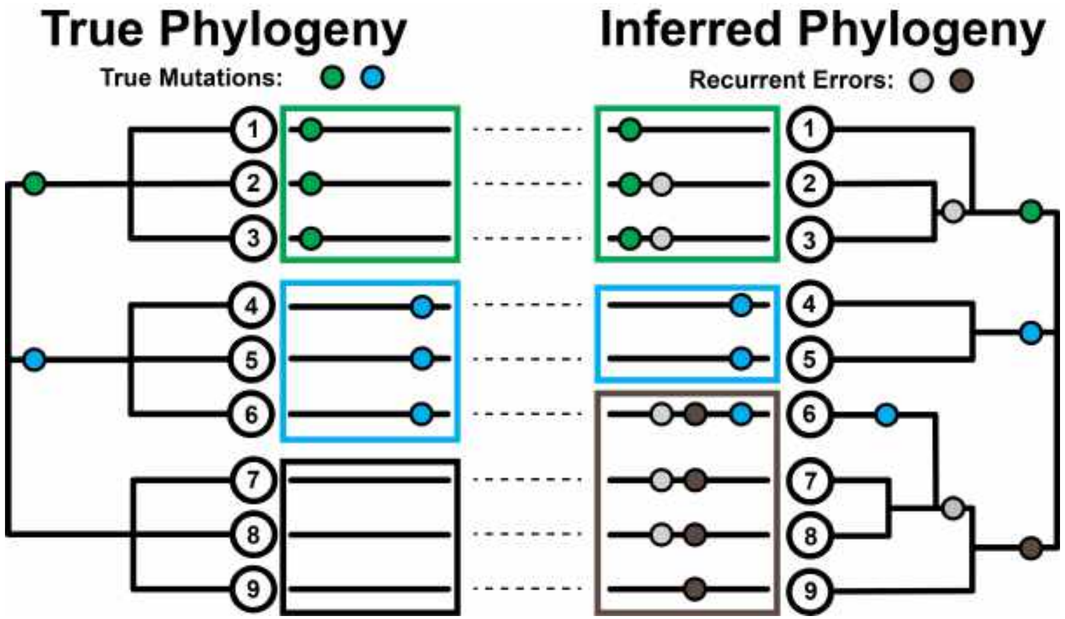
3. **Aid in rational vaccine design:**
   - Can the virus escape specific antibodies?

4. **Aid basic science research to prevent the next pandemic:**
   - What mutations led to this pandemic? Can we predict the next one?

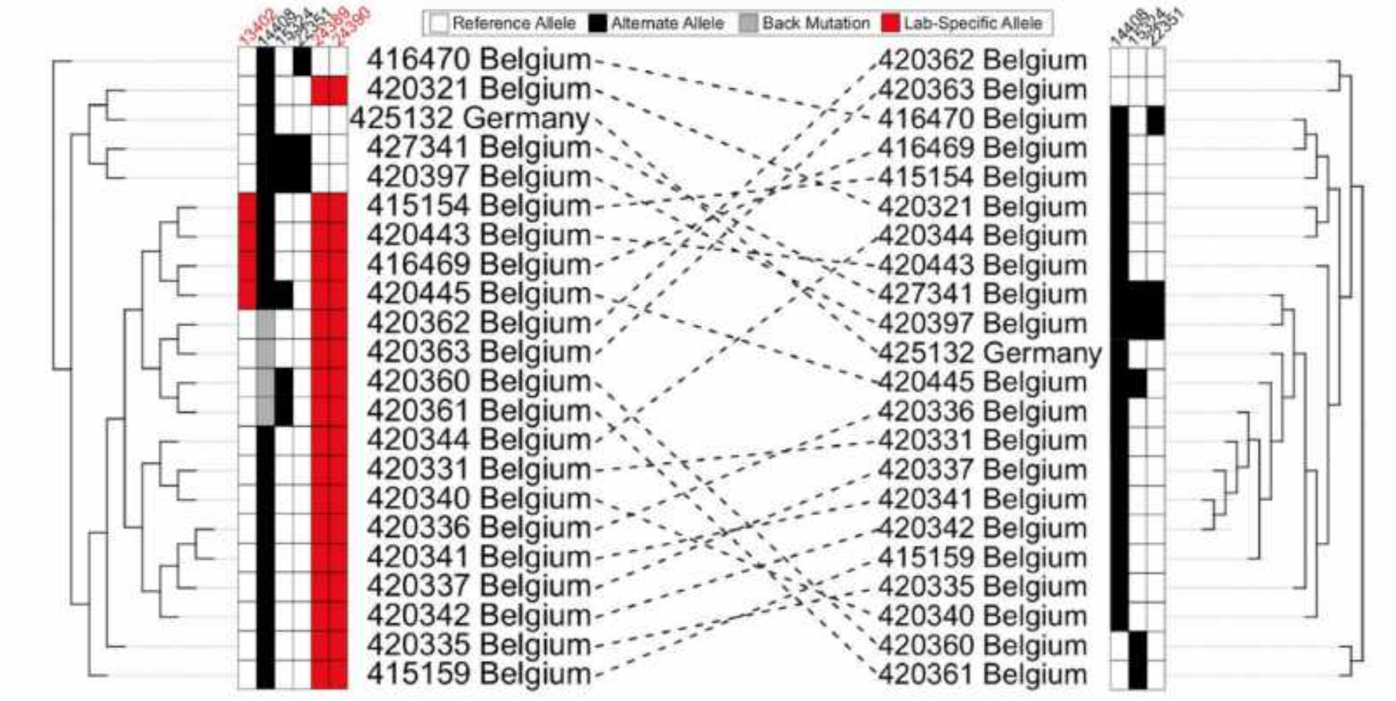
5. **Develop technologies to do better next time:**
   - What are the obstacles to true real-time viral genomics?

Sequencing artifacts can impact phylogenetic inferences

Mutations can trace transmission BUT artifacts can confound analysis



Some “mutations” that influence tree topology are lab-specific artifacts



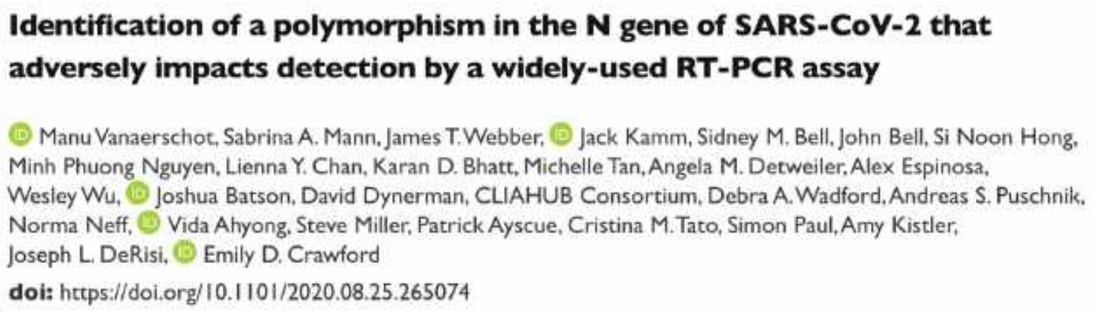
List of “problematic sites” to mask from analysis available on genome.ucsc.edu

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

Turakhia et al, 2020 [in revision] https://www.biorxiv.org/content/10.1101/2020.06.08.141127v1

Variation can affect the ability to accurately detect virus

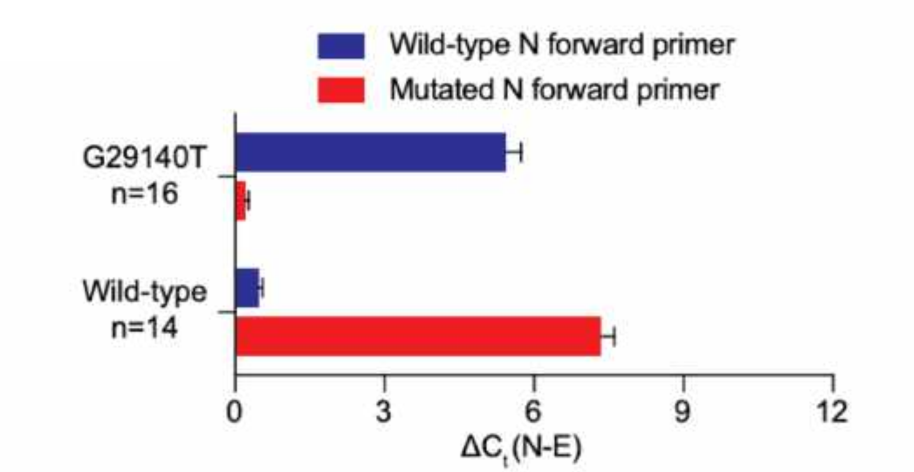
CZI Biohub recently reported variant that affected detection



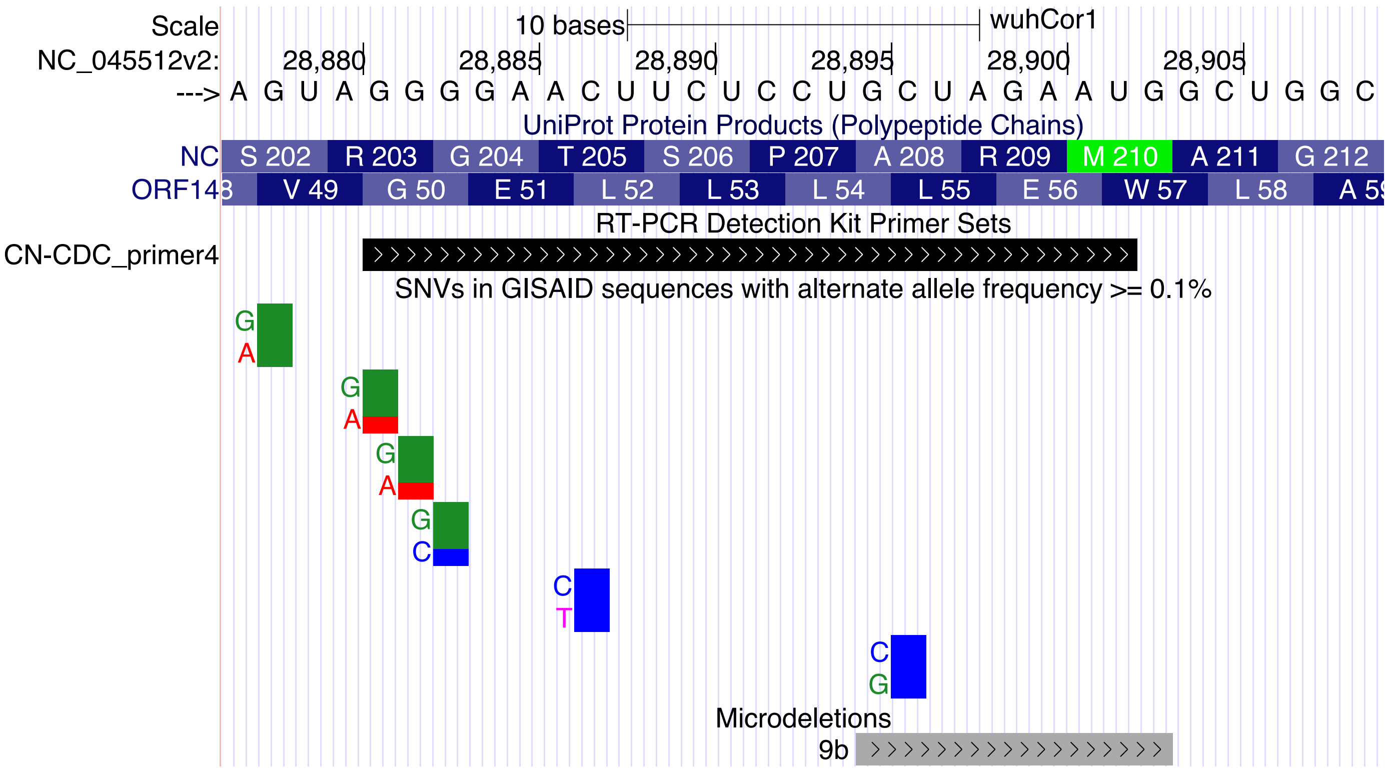
Identification of a polymorphism in the N gene of SARS-CoV-2 that adversely impacts detection by a widely-used RT-PCR assay

Manu Vanaerschot, Sabrina A. Mann, James T. Webber, Jack Kamm, Sidney M. Bell, John Bell, Si Noon Hong, Minh Phuong Nguyen, Lienna Y. Chan, Karan D. Bhatt, Michelle Tan, Angela M. Detweiler, Alex Espinosa, Wesley Wu, Joshua Batson, David Dynerman, CLIAHUB Consortium, Debra A. Wadford, Andreas S. Puschnik, Norma Neff, Vida Ahyong, Steve Miller, Patrick Ayscue, Cristina M. Tato, Simon Paul, Amy Kistler, Joseph L. DeRisi, Emily D. Crawford

doi: https://doi.org/10.1101/2020.08.25.265074



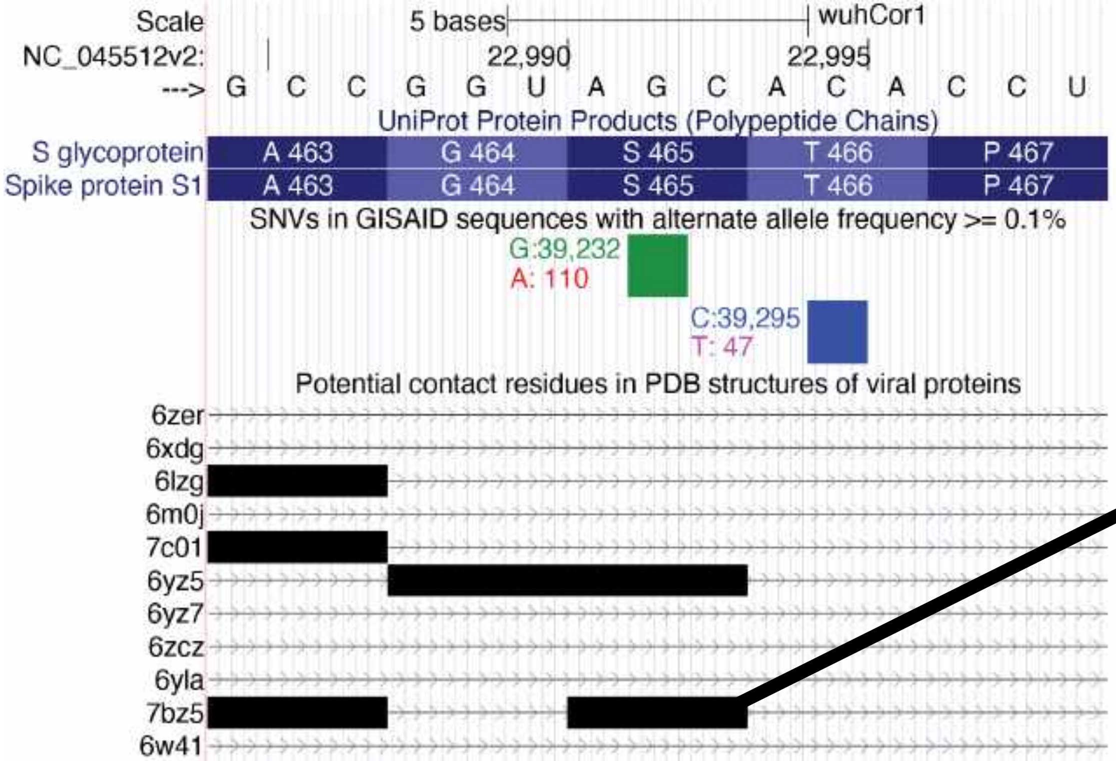
Genome Browser overlays standard detection primers with emerging variants



Deletions and variants alter primer choice as pandemic progresses.

“PDB Ligand Contacts” track allows visualization of mutational patterns at antibody-antigen interface

Genome browser view of variants and predicted antibody-spike contacts



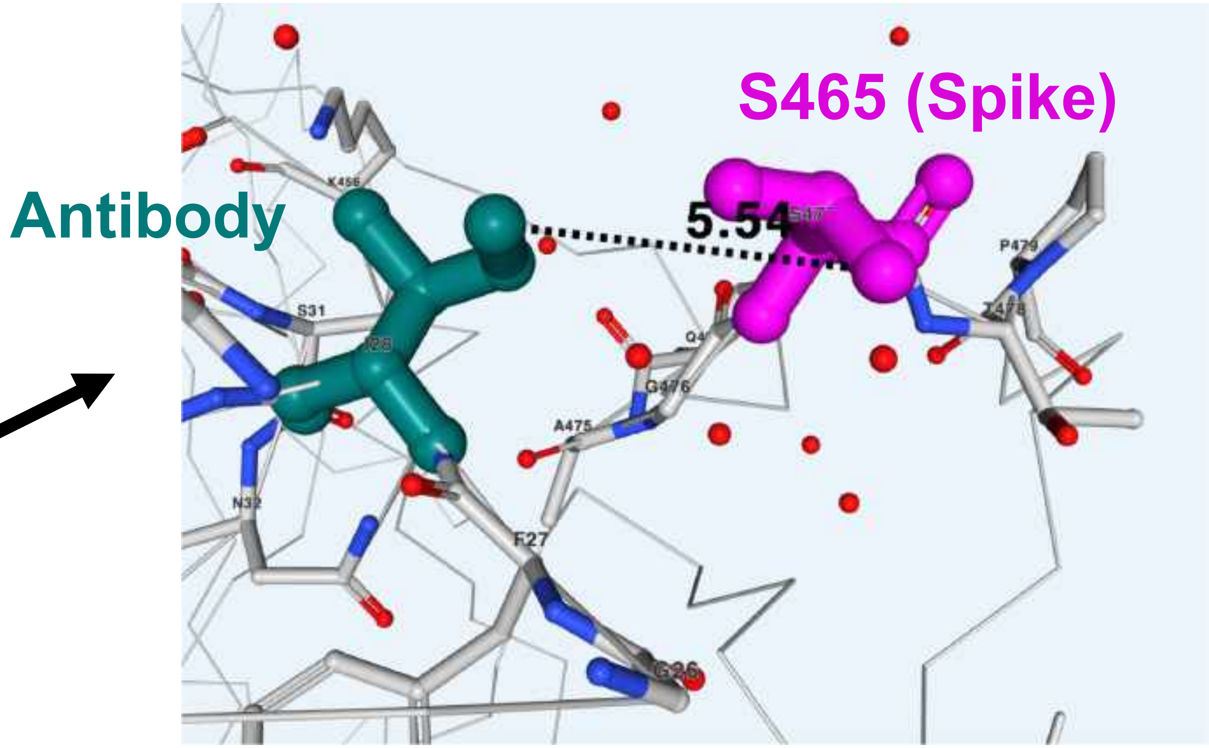
Scale 5 bases wuhCor1
NC_045512v2: 22,990 22,995
---> G C C G G U A G C A C A C C U
UniProt Protein Products (Polypeptide Chains)

S glycoprotein	A 463	G 464	S 465	T 466	P 467
Spike protein S1	A 463	G 464	S 465	T 466	P 467

SNVs in GISAID sequences with alternate allele frequency >= 0.1%
G:39,232
A: 110
C:39,295
T: 47

Potential contact residues in PDB structures of viral proteins
6zer
6xdg
6lzg
6m0j
7c01
6yz5
6yz7
6zcz
6yla
7bz5
6w41

Interactive viewer on click



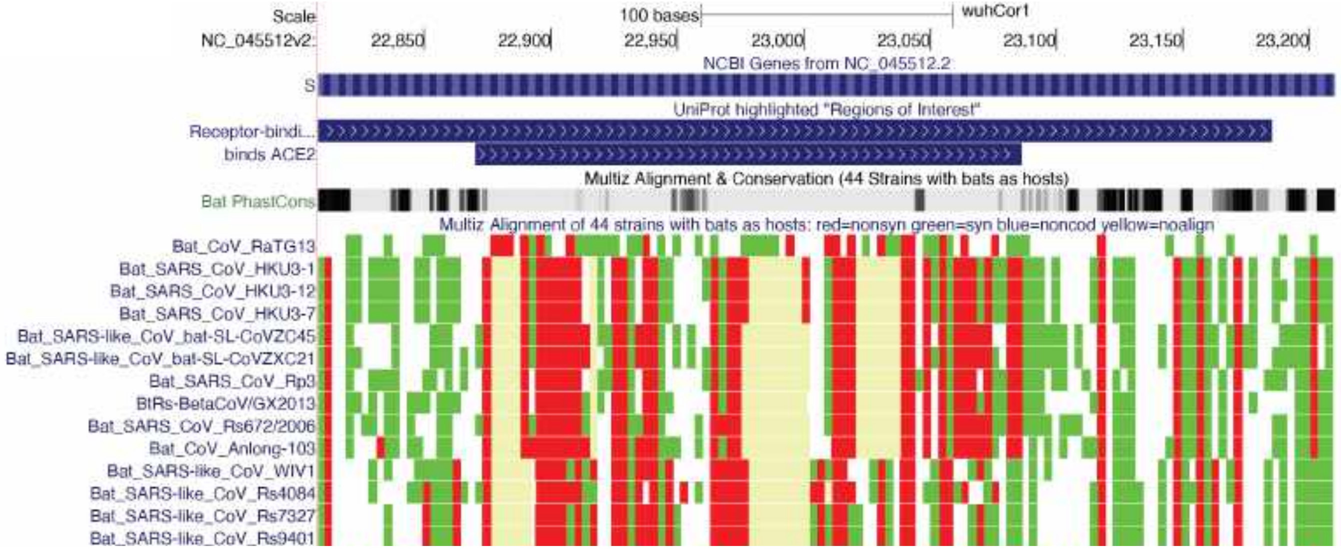
Neutralizing antibodies could contact mutable residues.
However, no contact sites for S antibodies have variants >1% frequency.
Nothing to worry about so far!

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

PDB from Wu et al., *Science*, 2020

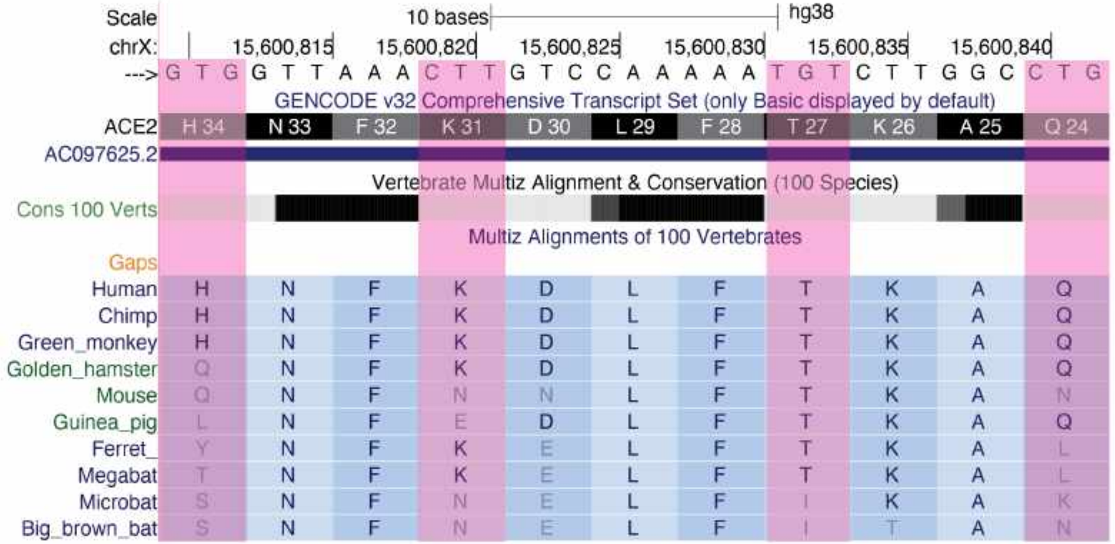
Virus and host receptor interfaces rapidly evolve

Alignment of Coronaviruses (SARS-CoV-2 Browser)



Green = synonymous mutations Red = non-synonymous mutations Yellow = alignment gap

Alignment of 100 vertebrates (Human Genome Browser: hg38)

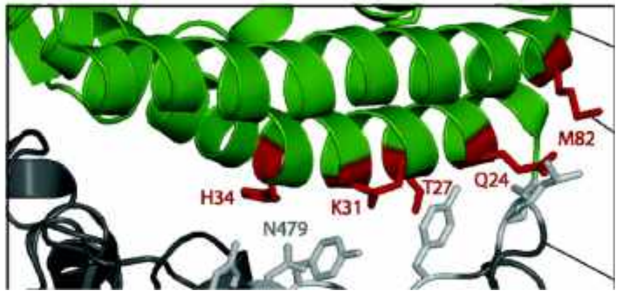

Scale
10 bases
hg38
chrX:
15,600,815
15,600,820
15,600,825
15,600,830
15,600,835
15,600,840
---> G T G G T T A A A C T T G T C C A A A A A T G T C T T G G C C T G
GENCODE v32 Comprehensive Transcript Set (only Basic displayed by default)
AC097625.2
Vertebrate Multiz Alignment & Conservation (100 Species)
Cons 100 Verts
Multiz Alignments of 100 Vertebrates
Gaps

ACE2	H 34	N 33	F 32	K 31	D 30	L 29	F 28	T 27	K 26	A 25	Q 24
Human	H	N	F	K	D	L	F	T	K	A	Q
Chimp	H	N	F	K	D	L	F	T	K	A	Q
Green_monkey	H	N	F	K	D	L	F	T	K	A	Q
Golden_hamster	Q	N	F	K	D	L	F	T	K	A	Q
Mouse	Q	N	F	N	N	L	F	T	K	A	N
Guinea_pig	L	N	F	E	D	L	F	T	K	A	Q
Ferret_	Y	N	F	K	E	L	F	T	K	A	L
Megabat	T	N	F	K	E	L	F	T	K	A	L
Microbat	S	N	F	N	E	L	F	I	K	A	K
Big_brown_bat	S	N	F	N	E	L	F	I	T	A	N

ACE2 residues that contact S

Residues in S-ACE2 interface are rapidly evolving in both virus and host.

SARS-CoV-2 is evolved to be successful in humans.



ACE2

S

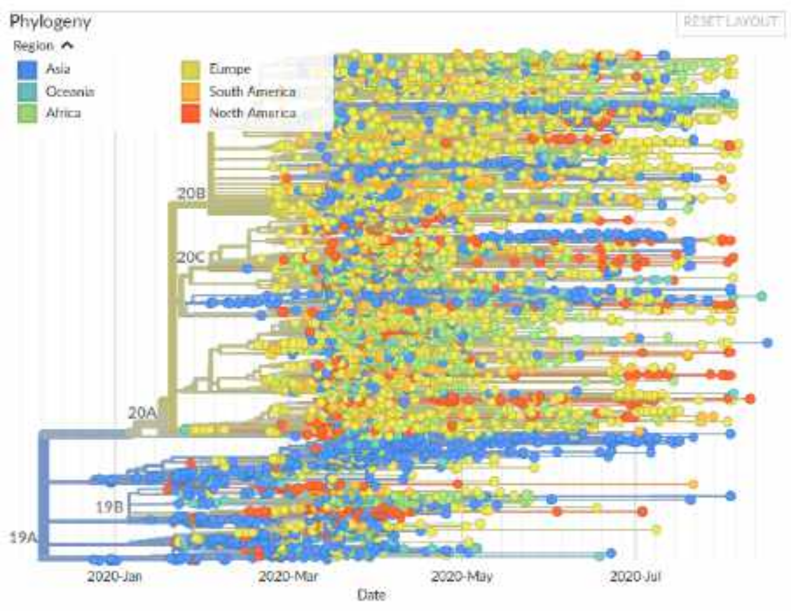
UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

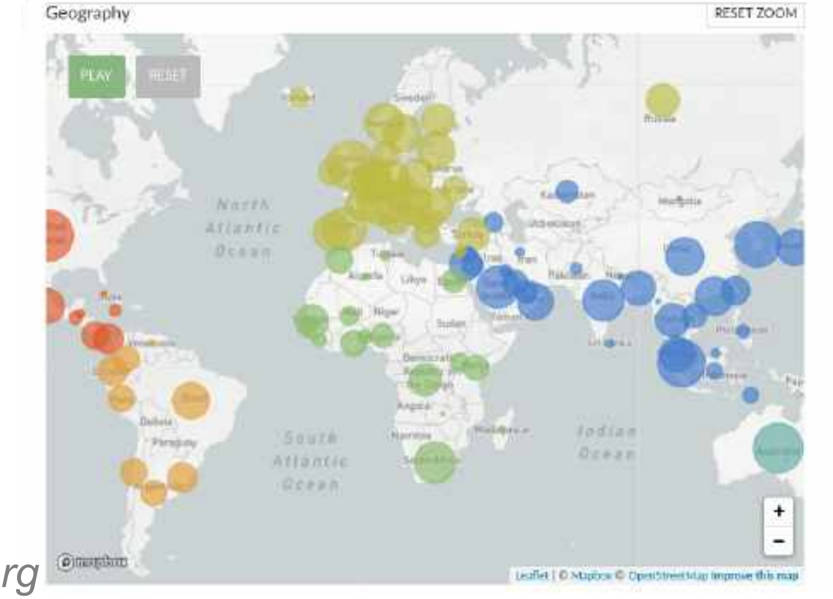
Improving genomic analyses during an outbreak

Workflow:

1. Sequence & Assemble Genome
2. Upload sequences to database.
3. Place new sequences in context of existing global phylogenetic tree.
4. Trace spread via genomic epidemiology.

Although vastly better than previous efforts, each step is not truly ”real-time”. Specifically, current phylogenetics software is not built to scale to 100,000 genomes!





Nexstrain.org

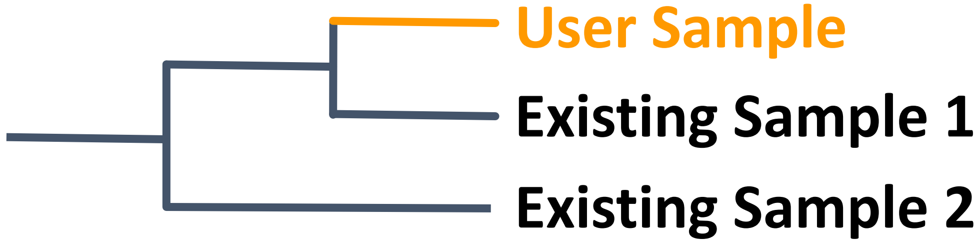
UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

Ultrafast Sample placement on Existing tRees (UShER) is a step towards real-time viral genomics

Workflow:

1. Sequence & Assemble Genome
2. Upload sequences to database.
3. **Place new sequences in context of existing global phylogenetic tree.**
4. Trace spread via genomics.

Although vastly better than previous efforts, each step is not truly ”real-time”. Current methods not built to scale for 100,000 genomes!



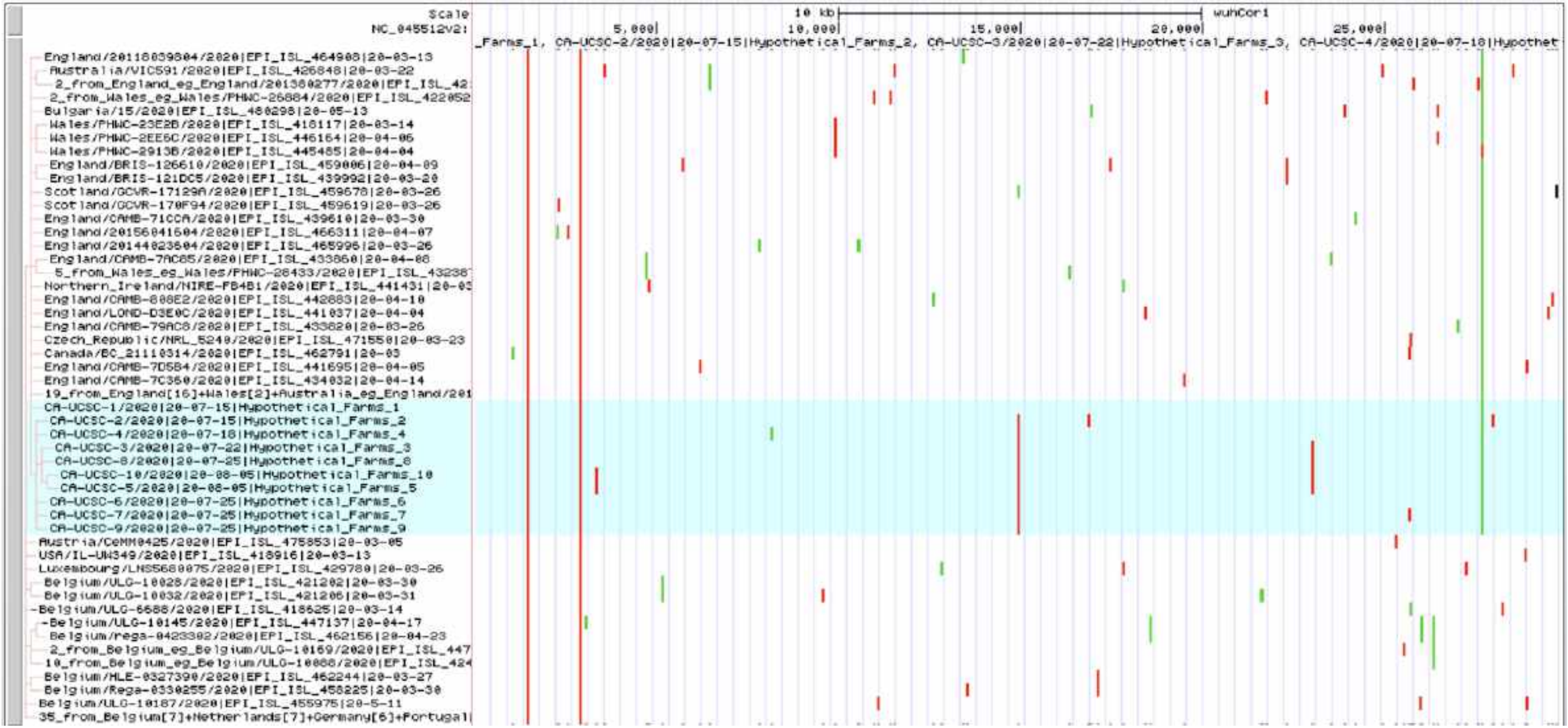
Method	Time to Place 1000 Sequences
PAGAN2	24+ Hours
IQ-TREE2	24+ Hours
TreeBeST	24+ Hours
RAxML epa	24+ Hours
UShER	**43.2 SECONDS**

UsHER uses parsimony annotations of tree branches & an optimized binary file

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

UShER is now available and integrated into the SARS-CoV-2 Genome Browser

New samples added in blue in interactive environment with alignment and mutation calls



```
                                                     Scale              10 kb |------------------------------| wuhCor1
                                              NC_045512v2:      5,000|      10,000|      15,000|      20,000|      25,000|
                                   _Farms_1, CA-UCSC-2/2020|20-07-15|Hypothetical_Farms_2, CA-UCSC-3/2020|20-07-22|Hypothetical_Farms_3, CA-UCSC-4/2020|20-07-18|Hypothet
England/20118039804/2020|EPI_ISL_464908|20-03-13
 Australia/VIC591/2020|EPI_ISL_426848|20-03-22
  2_from_England_eg_England/201360277/2020|EPI_ISL_42
 2_from_Wales_eg_Wales/PHWC-26884/2020|EPI_ISL_422052
Bulgaria/15/2020|EPI_ISL_480298|20-05-13
 Wales/PHWC-23E2B/2020|EPI_ISL_418117|20-03-14
 Wales/PHWC-2EE6C/2020|EPI_ISL_446164|20-04-06
 Wales/PHWC-2913B/2020|EPI_ISL_445485|20-04-04
 England/BRIS-126610/2020|EPI_ISL_459006|20-04-09
 England/BRIS-121DC5/2020|EPI_ISL_439992|20-03-20
Scotland/GCVR-17129A/2020|EPI_ISL_459678|20-03-26
Scotland/GCVR-170F94/2020|EPI_ISL_459619|20-03-26
England/CAMB-71CCA/2020|EPI_ISL_439610|20-03-30
England/20156041604/2020|EPI_ISL_466311|20-04-07
England/20144023604/2020|EPI_ISL_465996|20-03-26
 England/CAMB-7AC85/2020|EPI_ISL_433860|20-04-08
  5_from_Wales_eg_Wales/PHWC-28433/2020|EPI_ISL_43238
Northern_Ireland/NIRE-FB4B1/2020|EPI_ISL_441431|20-03
England/CAMB-808E2/2020|EPI_ISL_442883|20-04-10
England/LOND-D3E0C/2020|EPI_ISL_441037|20-04-04
England/CAMB-79AC8/2020|EPI_ISL_433820|20-03-26
Czech_Republic/NRL_5240/2020|EPI_ISL_471550|20-03-23
Canada/BC_21110314/2020|EPI_ISL_462791|20-03
England/CAMB-7D5B4/2020|EPI_ISL_441695|20-04-05
England/CAMB-7C360/2020|EPI_ISL_434032|20-04-14
19_from_England[16]+Wales[2]+Australia_eg_England/201
CA-UCSC-1/2020|20-07-15|Hypothetical_Farms_1
 CA-UCSC-2/2020|20-07-15|Hypothetical_Farms_2
 CA-UCSC-4/2020|20-07-18|Hypothetical_Farms_4
  CA-UCSC-3/2020|20-07-22|Hypothetical_Farms_3
  CA-UCSC-8/2020|20-07-25|Hypothetical_Farms_8
   CA-UCSC-10/2020|20-08-05|Hypothetical_Farms_10
   CA-UCSC-5/2020|20-08-05|Hypothetical_Farms_5
 CA-UCSC-6/2020|20-07-25|Hypothetical_Farms_6
 CA-UCSC-7/2020|20-07-25|Hypothetical_Farms_7
 CA-UCSC-9/2020|20-07-25|Hypothetical_Farms_9
Austria/CeMM0425/2020|EPI_ISL_475853|20-03-05
USA/IL-UW349/2020|EPI_ISL_418916|20-03-13
Luxembourg/LNS5680075/2020|EPI_ISL_429780|20-03-26
 Belgium/ULG-10028/2020|EPI_ISL_421202|20-03-30
 Belgium/ULG-10032/2020|EPI_ISL_421206|20-03-31
-Belgium/ULG-6688/2020|EPI_ISL_418625|20-03-14
 -Belgium/ULG-10145/2020|EPI_ISL_447137|20-04-17
  Belgium/rega-0423302/2020|EPI_ISL_462156|20-04-23
  2_from_Belgium_eg_Belgium/ULG-10169/2020|EPI_ISL_447
 10_from_Belgium_eg_Belgium/ULG-10088/2020|EPI_ISL_424
 Belgium/MLE-0327390/2020|EPI_ISL_462244|20-03-27
 Belgium/Rega-0330255/2020|EPI_ISL_458225|20-03-30
Belgium/ULG-10187/2020|EPI_ISL_455975|20-5-11
35_from_Belgium[7]+Netherlands[7]+Germany[6]+Portugal|
```

Link & Demonstration: https://genome.ucsc.edu/cgi-bin/hgPhyloPlace

www.github.com/russcd/USHER_DEMO/

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

An idealized roadmap for how future outbreaks might be traced in true real-time

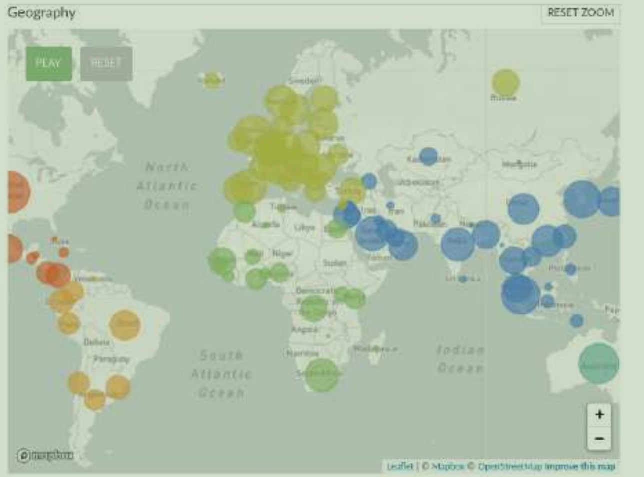
1. Sequence & assemble genome using nanopore and laptop in the field

2. Upload sequences to database automatically.

3. Place new sequences in context of existing global phylogenetic tree and get analysis immediately.

UShER

4. True real time genomic contact tracing!



UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

All data is easily accessible via the SARS-CoV-2 Browser

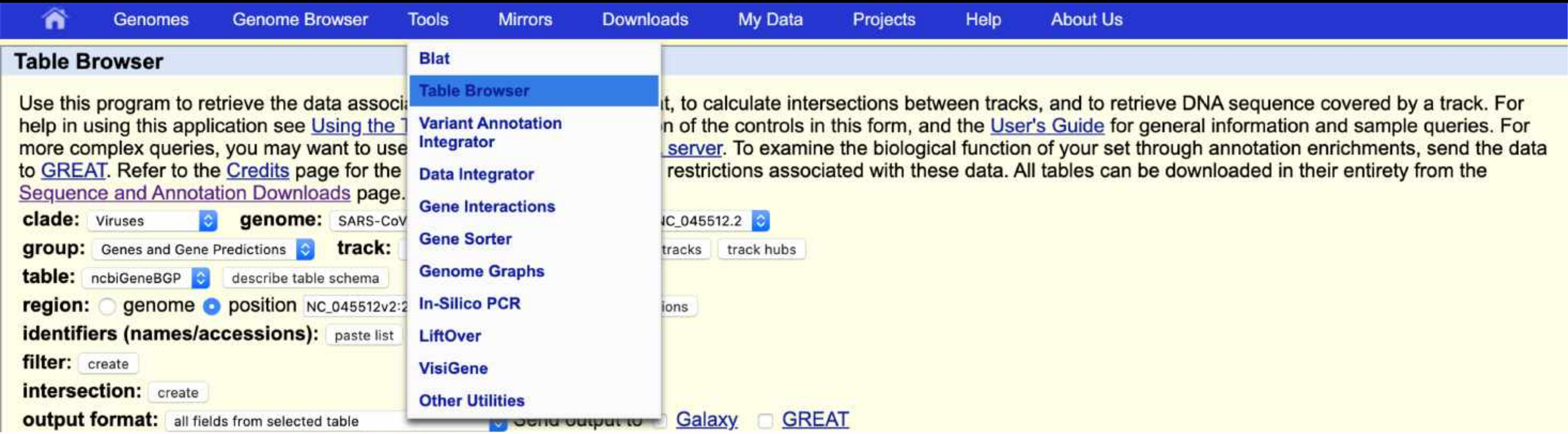


Table Browser

Use this program to retrieve the data associ it, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the T n of the controls in this form, and the User's Guide for general information and sample queries. For more complex queries, you may want to use server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Refer to the Credits page for the restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Viruses **genome:** SARS-CoV IC_045512.2

group: Genes and Gene Predictions **track:** tracks track hubs

table: ncbiGeneBGP describe table schema

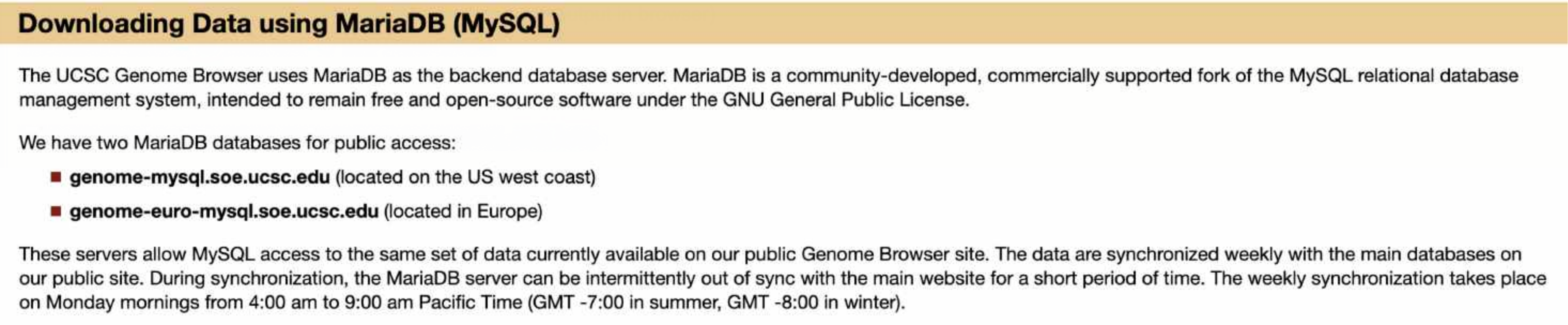
region: ○ genome ● position NC_045512v2:2 ions

identifiers (names/accessions): paste list

filter: create

intersection: create

output format: all fields from selected table Send output to □ Galaxy □ GREAT



Downloading Data using MariaDB (MySQL)

The UCSC Genome Browser uses MariaDB as the backend database server. MariaDB is a community-developed, commercially supported fork of the MySQL relational database management system, intended to remain free and open-source software under the GNU General Public License.

We have two MariaDB databases for public access:

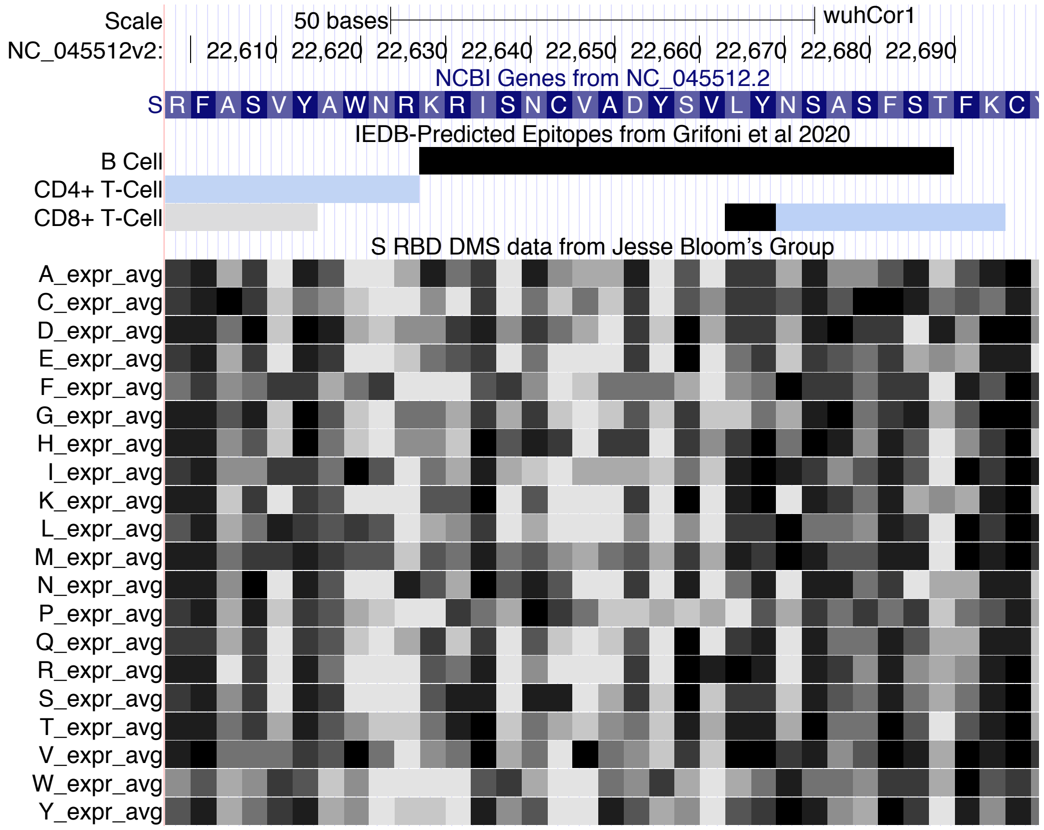
- **genome-mysql.soe.ucsc.edu** (located on the US west coast)
- **genome-euro-mysql.soe.ucsc.edu** (located in Europe)

These servers allow MySQL access to the same set of data currently available on our public Genome Browser site. The data are synchronized weekly with the main databases on our public site. During synchronization, the MariaDB server can be intermittently out of sync with the main website for a short period of time. The weekly synchronization takes place on Monday mornings from 4:00 am to 9:00 am Pacific Time (GMT -7:00 in summer, GMT -8:00 in winter).

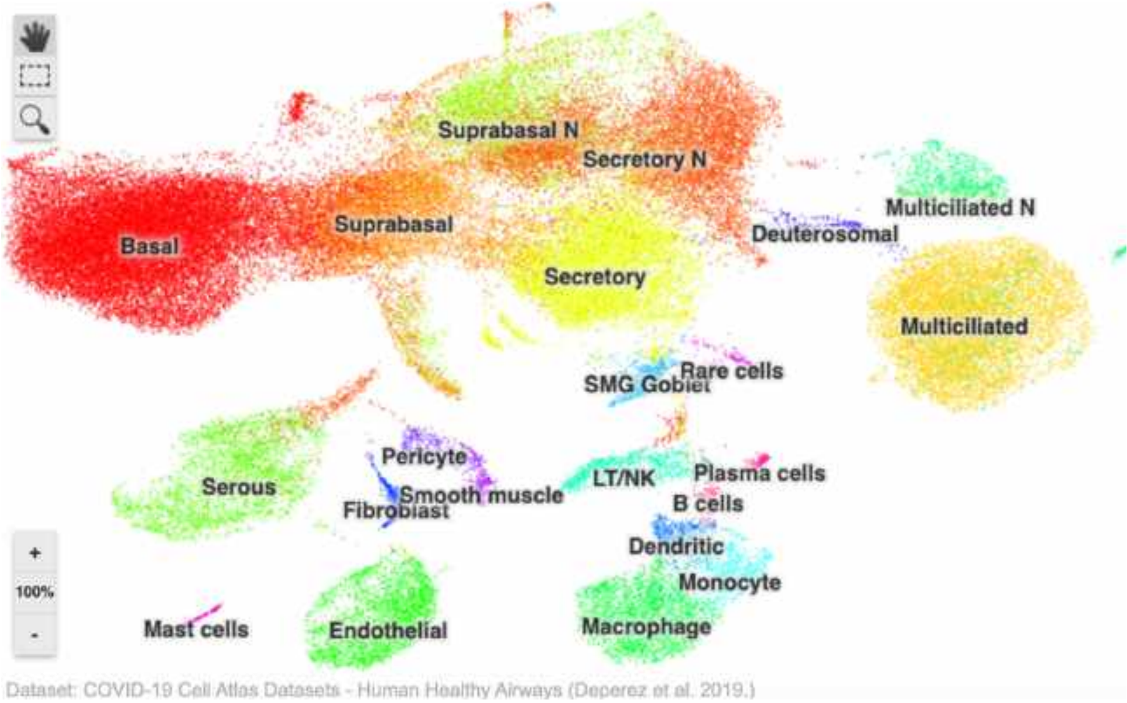
Please consider adding your genomic data!

Data from many HHMI colleagues already online:

S protein Deep Mutational Scanning Data *(Jesse Bloom Lab)* available at genome.ucsc.edu:



COVID19 Cell Atlas *(Mark Krasnow Lab)* available in interactive scRNA-seq browser at **cells.ucsc.edu:**



Dataset: COVID-19 Cell Atlas Datasets - Human Healthy Airways (Deperez et al. 2019.)

Acknowledgements

UCSC SARS-CoV-2 Browser

Jason Fernandes
Hiram Clawson
Angie Hinrichs
Jairo Navarro Gonzalez
Brian T . Lee
Luis R. Nassar
Brian J. Raney
Kate R. Rosenbloom
Santrupti Nerli
Arjun A. Rao
Daniel Schmelter
Alastair Fyfe
Nathan Maulding
Ann S. Zweig
Todd M. Lowe
Manuel Ares Jr
Jim Kent
Max Haeussler

UNIVERSITY OF CALIFORNIA SANTA CRUZ | Genomics Institute

Recurrent Errors & UShER

Russ Corbett Lab (UCSC)
Bryan Thornlow
Landen Gozashti

Yatish Turakhia (UCSC now, starting lab at UCSD, 2021)

Rob Lanfear (Australian National Univ)

Nick Goldman Lab (EBI)
Nicola De Maio
Conor R. Walker
Lukas Weilguny

Rui Borges (Institut für Populationsgenetik)
Greg Slodkowicz (MRC)

FUNDING

UCSC Human Genome Browser: NIH National Human Genome Research Institute

SARS-CoV-2 genome browser and data annotation tracks:

- Pat & Rowland Rebele
- Eric and Wendy Schmidt by recommendation of the Schmidt Futures program
- Center for Information Technology Research in the Interest of Society (CITRIS)
- University of California Office of the President (UCOP)

Funding for open access charge: National Human Genome Research Institute (NHGRI)