A resource for SARS-Cov2 mutation tracking and analysis:

The LANL/GISAID COVID pipeline

Bette Korber

Will Fischer

Los Alamos National Laboratory
Los Alamos, New Mexico, USA

5 November 2020

powered by

GISAID

Los Alamos NATIONAL LABORATORY EST. 1943

Unprecedented *outbreak*: Unprecedented *data*

High case numbers meet high sequencing capacity and bioinformatic infrastructure

- The GISIAD database (initially tasked with archiving influenza sequences) collects and distributes SARS-Cov2 sequence data (January 2020).
- The Korber team at the LANL HIV database begins analyzing SARS-Cov2 data, processing the raw sequences for mutational analysis (February 2020).
- LANL goal: forward-looking threat assessment of mutational escape from vaccines/therapeutics

GISAID

Los Alamos NATIONAL LABORATORY EST. 1943

LANL COVID-19/ SARS-Cov2 website: **https://cov.lanl.gov**

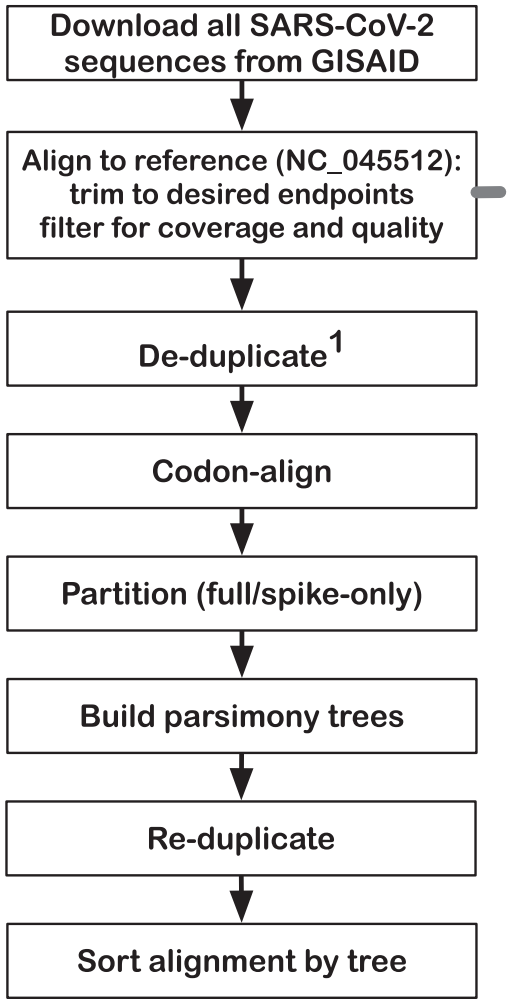
- Leverages infrastructure of LANL HIV-DB
- Provides analytical tools and links to computed analyses
- Emphasis on geographical distribution and temporal evolution of non-synonymous mutations in coding regions, especially the Spike protein
- Sequence data and analytical results updated daily.

Sequence Processing Pipeline

Quality control and processing

- Data received at LANL in custom JSON feed
- Processing:
  - mapping to reference strain
  - quality filtering
  - alignment
  - phylogenetic inference

A

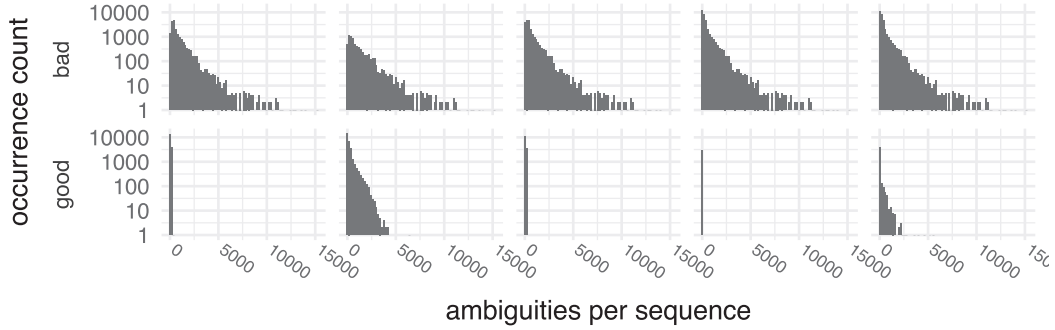


B

Sequence filtering results

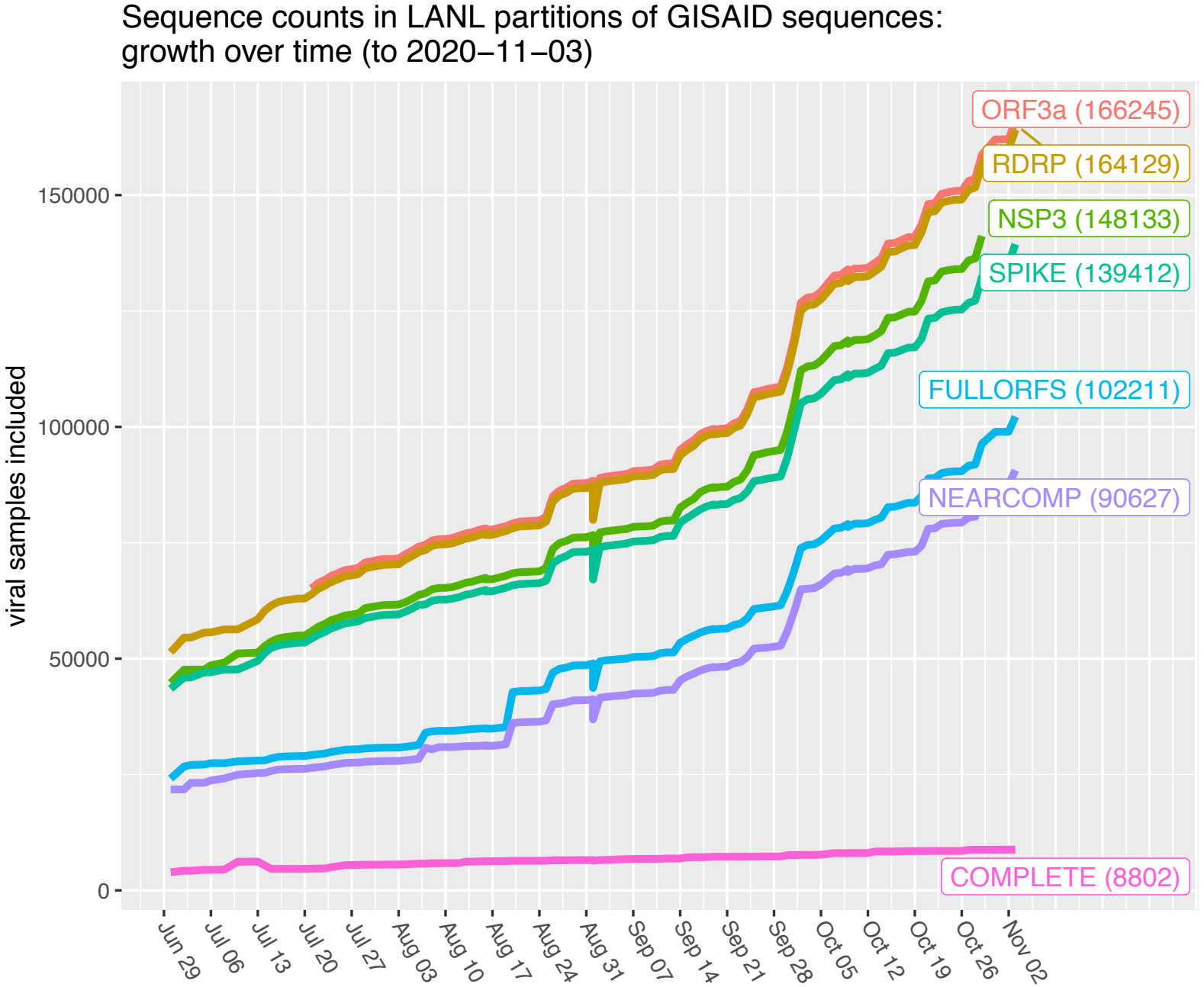
category	CODING REGIONS[2]	SPIKE	NEAR-COMPLETE	COMPLETE	5' UTR
"good"	17222	28576	14105	3019	4228
"bad"	17172	6382	19607	28984	27886
5' end gaps	73	160	437	2321	26733
3' end gaps	142	35	1546	624	9
5' end gaps + 3' end gaps	148	12	733	9391	125
fragmented matching	15091	4541	15115	15132	6
Excess "N" stretches	323	400	270	9	6
Excess mismatches (>30)	1323	1065	1159	57	10
5' end gaps + mismatches	48	33	114	1229	72
3' end gaps + mismatches	21	11	174	51	21
5' + 3' end gaps + mismatches	0	0	57	0	5
No matching detected	3	122	2	2	920

Ambiguous base calls (primarily N)



Continued growth in sequence submission

- Sequence counts tripled through July-October 2020
- LANL partitions sequences into regions of interest, assembling aligned datasets containing only high-quality bases throughout a given region
  - FULLORFS: start codon of first reading frame (ORF1ab) to the stop codon of the last reading frame (ORF10): 29,409 nt
  - NEARCOMP: includes the most-commonly-used primer sites: 29,782 nt
  - SPIKE and other coding regions
  - Sequences are excluded from a given partition if they have many uncalled bases, long indels, or extensive mismatches



The O, S, L, and V clade are rarely sampled after June 1, through the summer of 2020

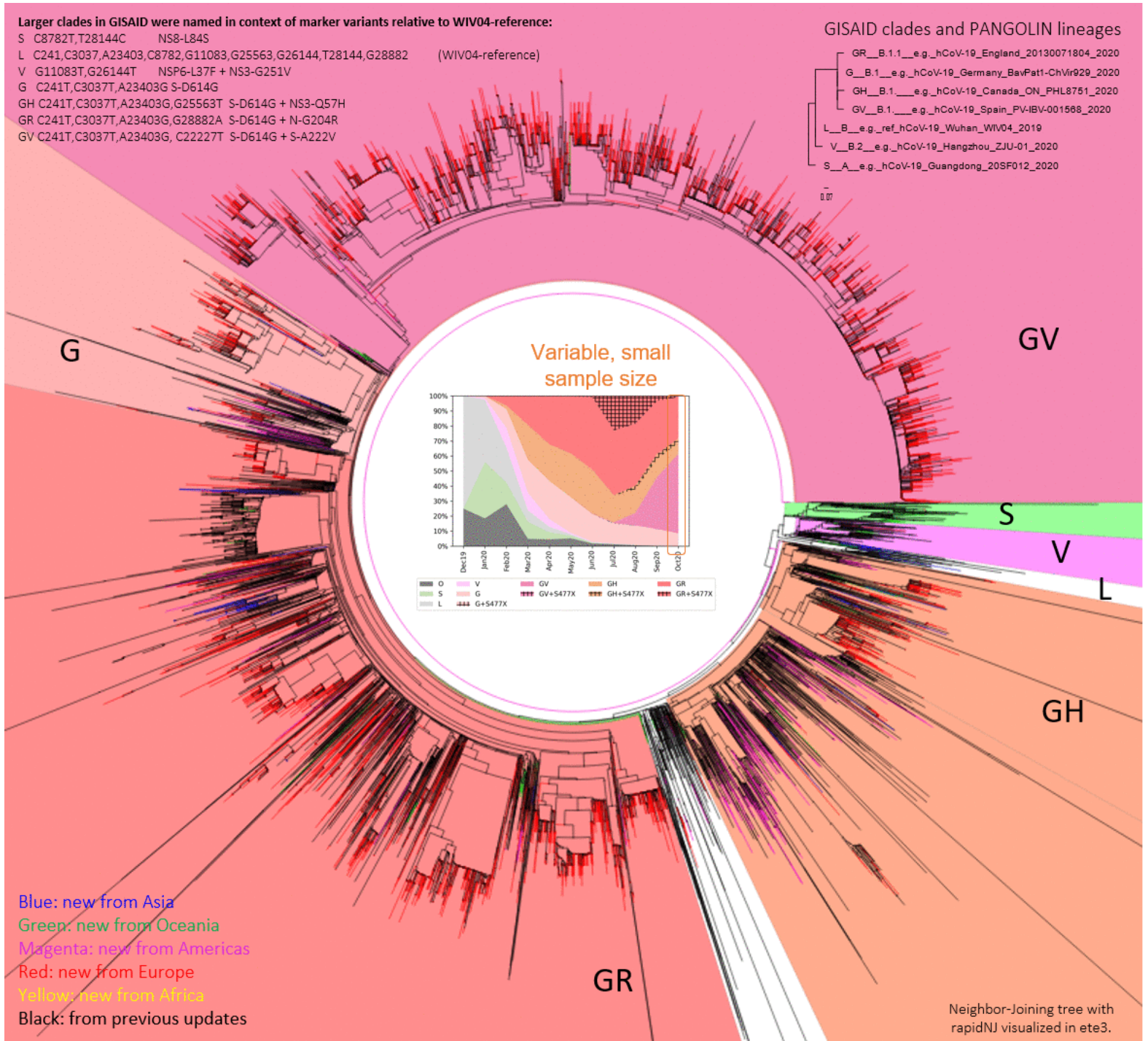
G is Spike D614G:

G has 3 sub-lineages
GR,
GH,
and GV (Spike A222V)

GR is the most frequently sampled, but is very common in the UK which is highly sampled.

GH is also frequently sampled, and is common in the US.

L is complex, may include recombinants?



Full genome tree derived from all outbreak sequences 2020-10-30

Notable changes:

155,961 full genomes (+7,632) (excluding low coverage, out of 167,272 entries)

Updated clades:
S clade 6,495 (+39)
L clade 4,236 (+1)
V clade 5,343 (+21)
G clade [#S477X] 28,633 [105] (+769 [+3])
GR clade [#S477X] 60,971 [8,868] (+2,130 [+33])
GH clade [#S477X] 34,922 [1,098] (+932 [+204])
GV clade [#S477X] 11,431 [5] (+3,719 [+0])
Other clades 3,930 (+21)

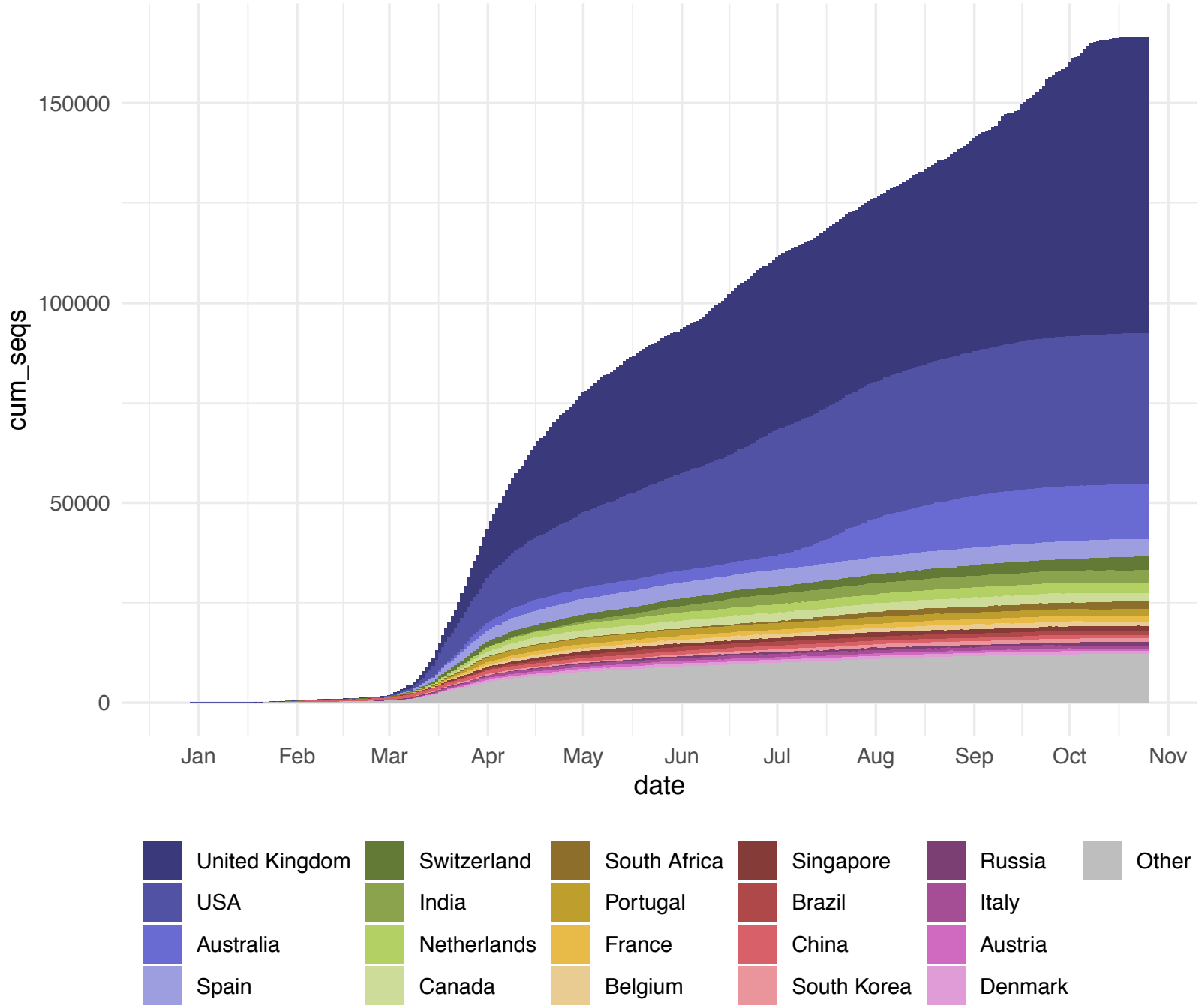
We gratefully acknowledge the Authors from Originating and Submitting laboratories of sequence data on which the analysis is based.

GISAID

by BII/GIS, A*STAR Singapore

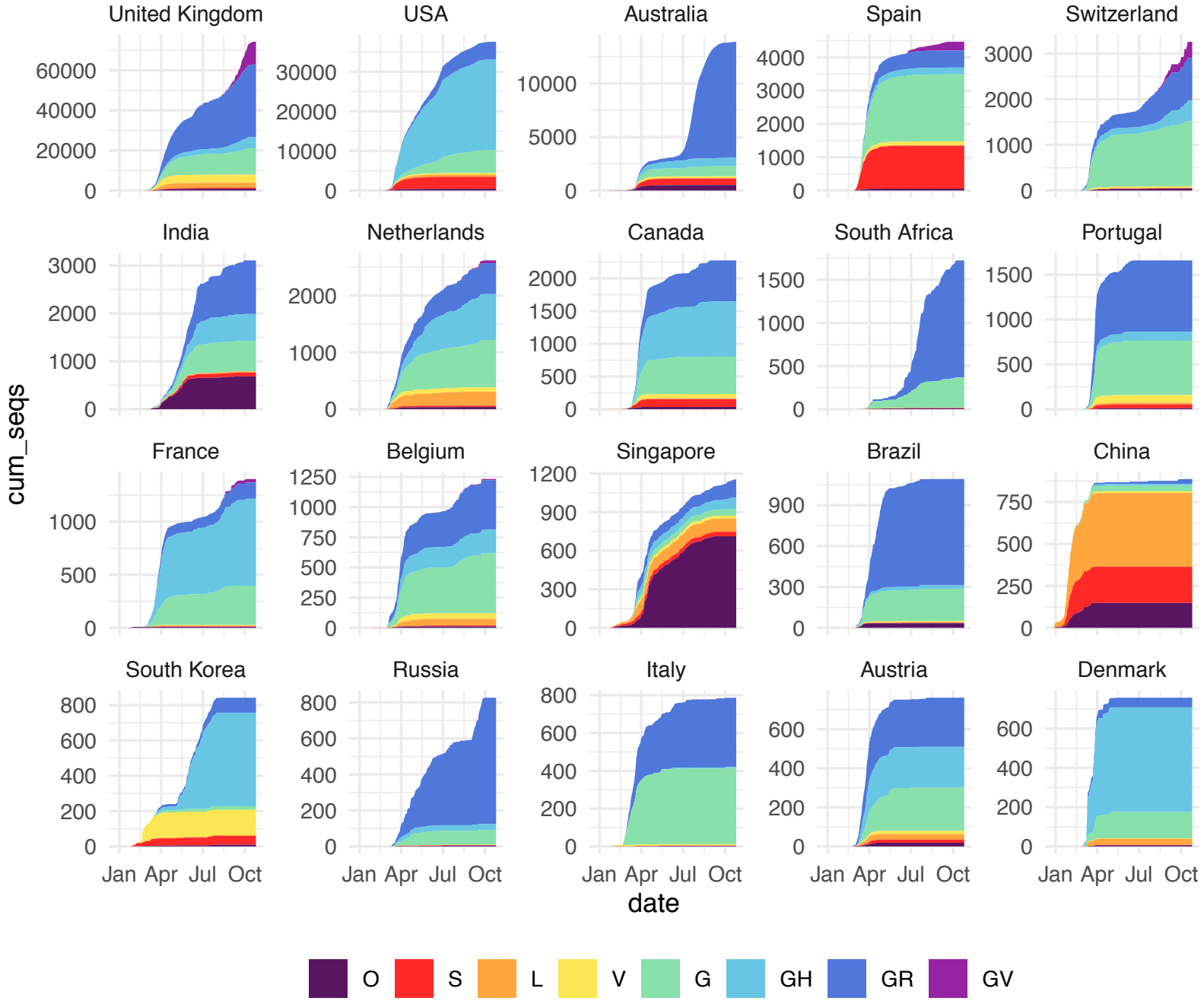
Sequencing effort is highly biased

- Geographically:
  - The top 20 countries account for 92% of all sequences submitted.
  - 78% of GISAID SARS-Cov2 sequence data come from the top 4 countries (U.K., USA, Australia, Spain); >44% from the U.K. alone.
- Temporally
  - many countries employed concentrated bursts of sequencing effort, with extensive sampling gaps



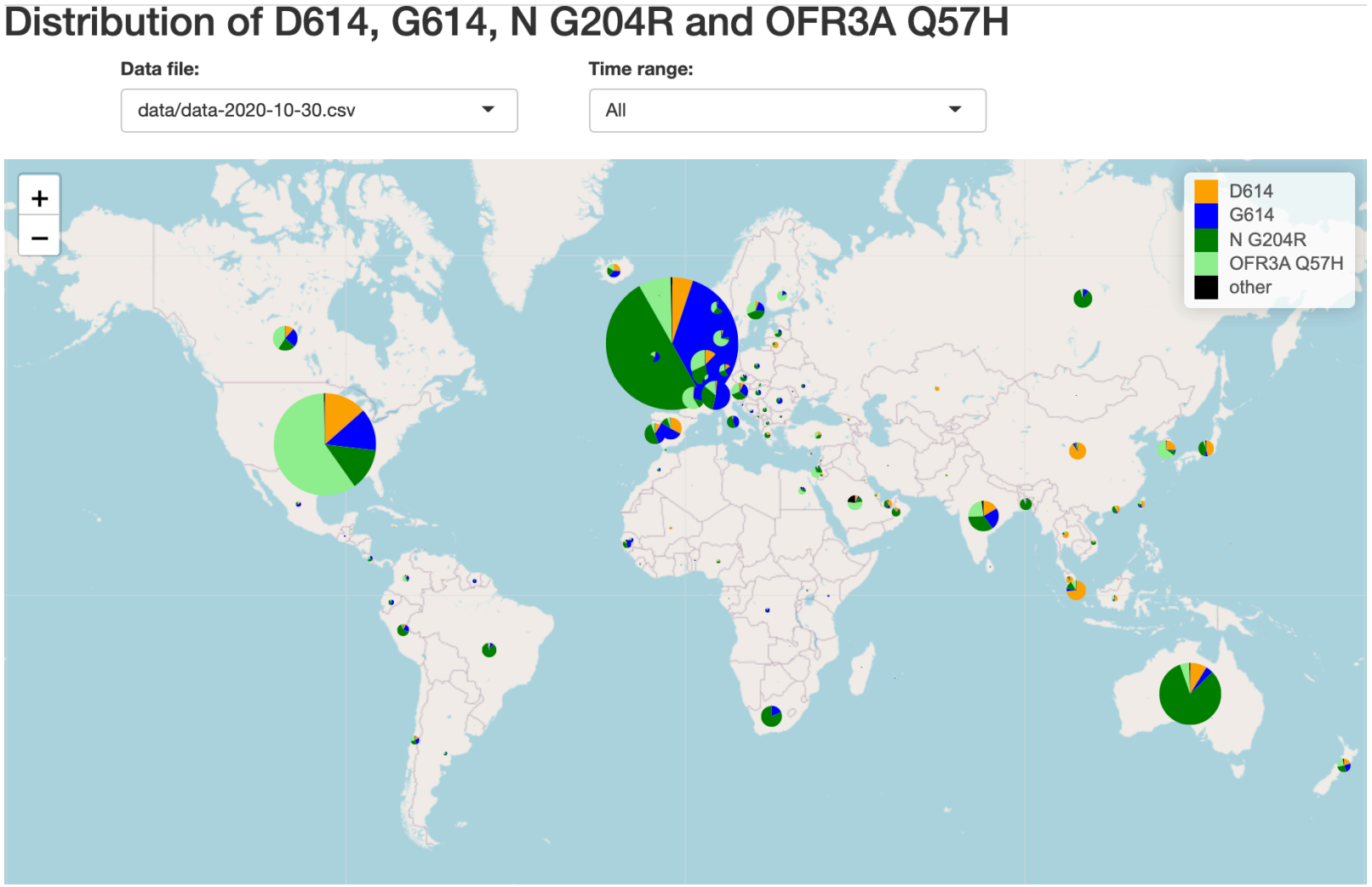
Clade distribution varies between countries

- Cumulative sequence counts (vertical change shows added sequences)
- In most countries, early clades (O,S,L,V) are replaced by G clade and its descendants
- Examples:
  - arrival and expansion of GR clade in Australia in July 2020
  - appearance of GV in mid-August in the U.K., France, and Switzerland; earlier in Spain
  - continued detection of the original "O" form in Singapore
  - "frozen" early frequency distributions where few later sequences reported (e.g., China, Portugal, Denmark)



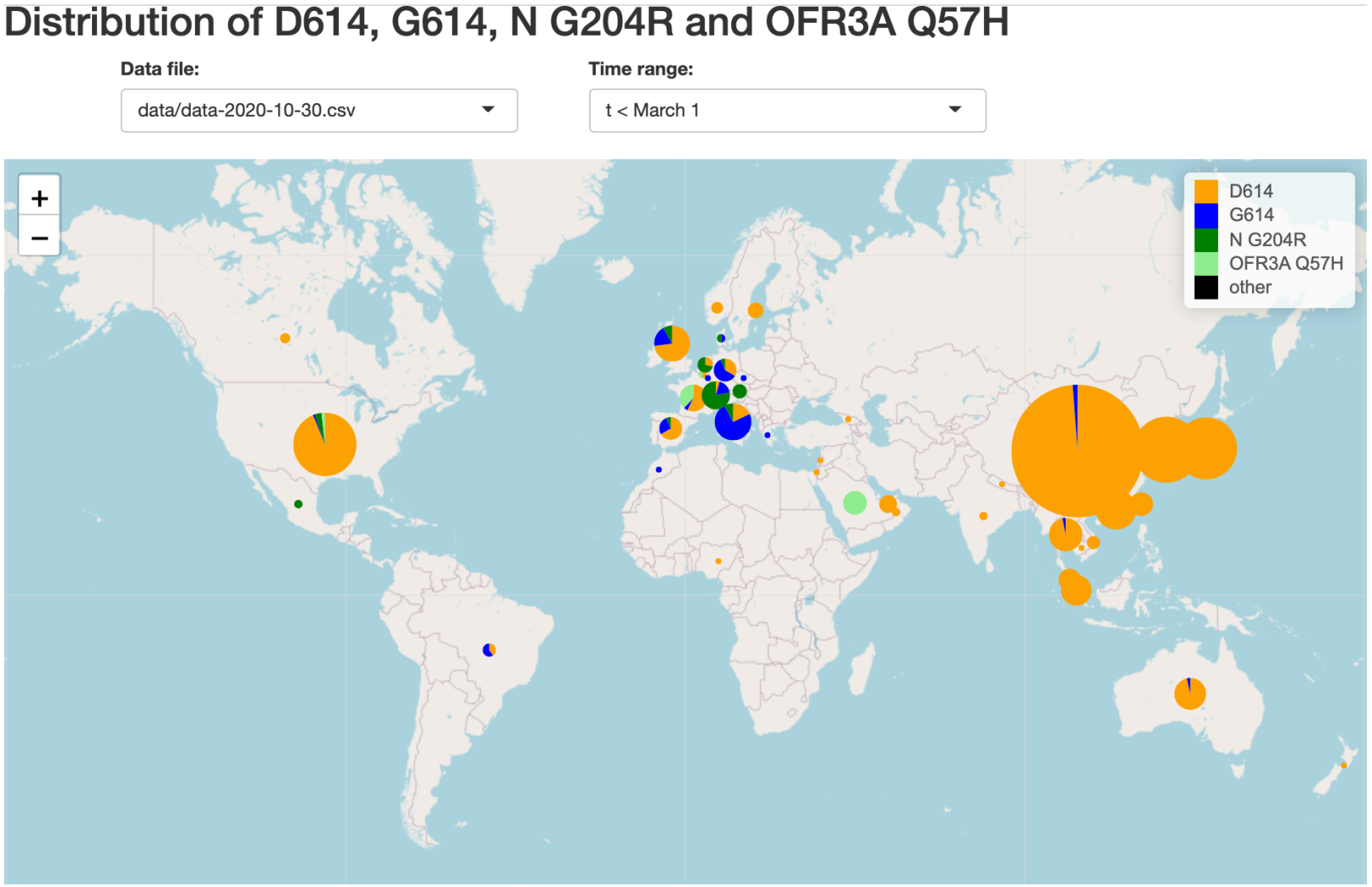
Mutational clade distributions

- Distributions sized by total sequence count (not case count!)
- Two-week intervals from March 1, 2020
- After March, "G" and descendent clades predominate worldwide



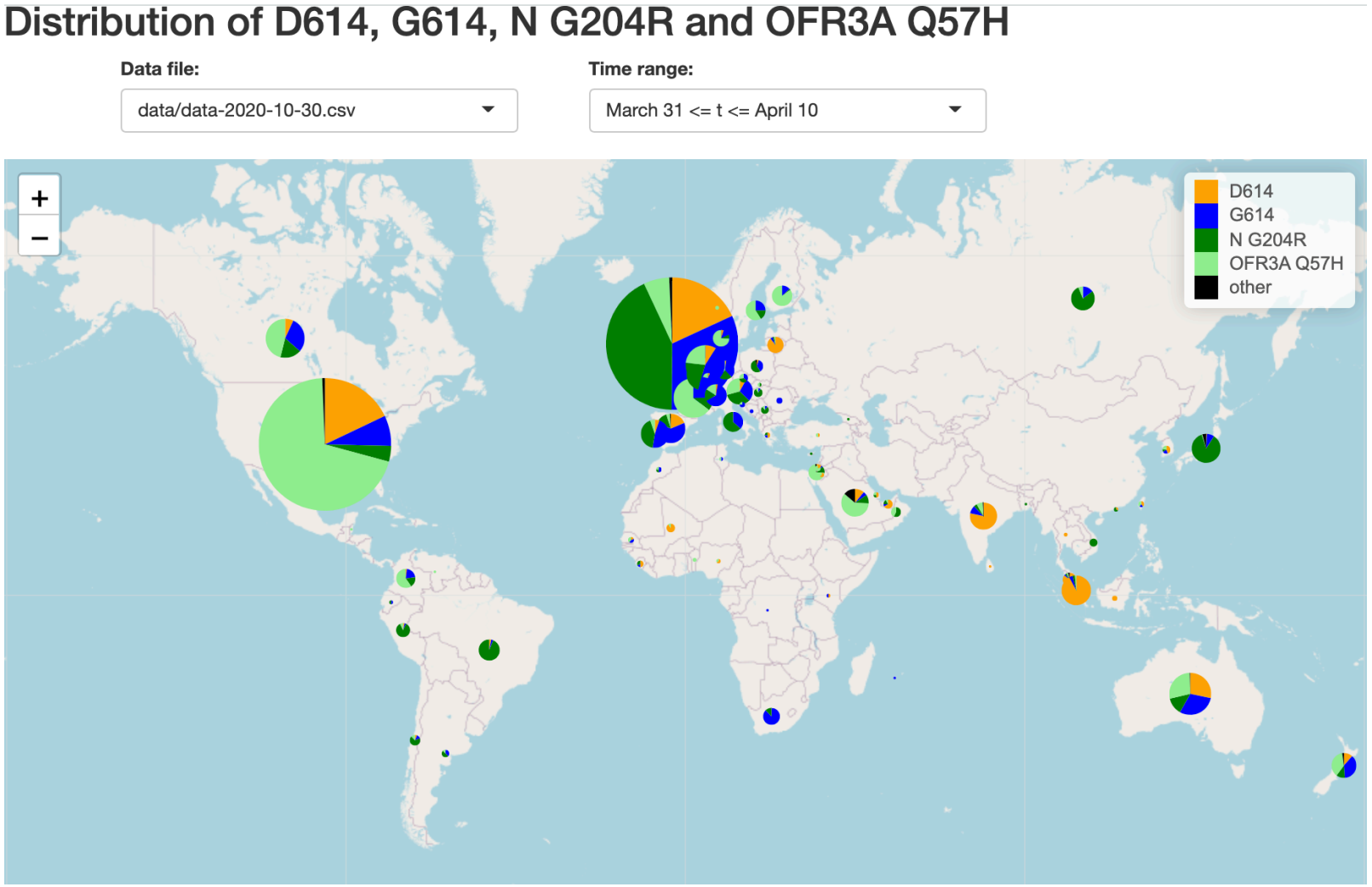
Mutational clade distributions

- Distributions sized by total sequence count (not case count!)
- Two-week intervals from March 1, 2020
- After March, "G" and descendent clades predominate worldwide



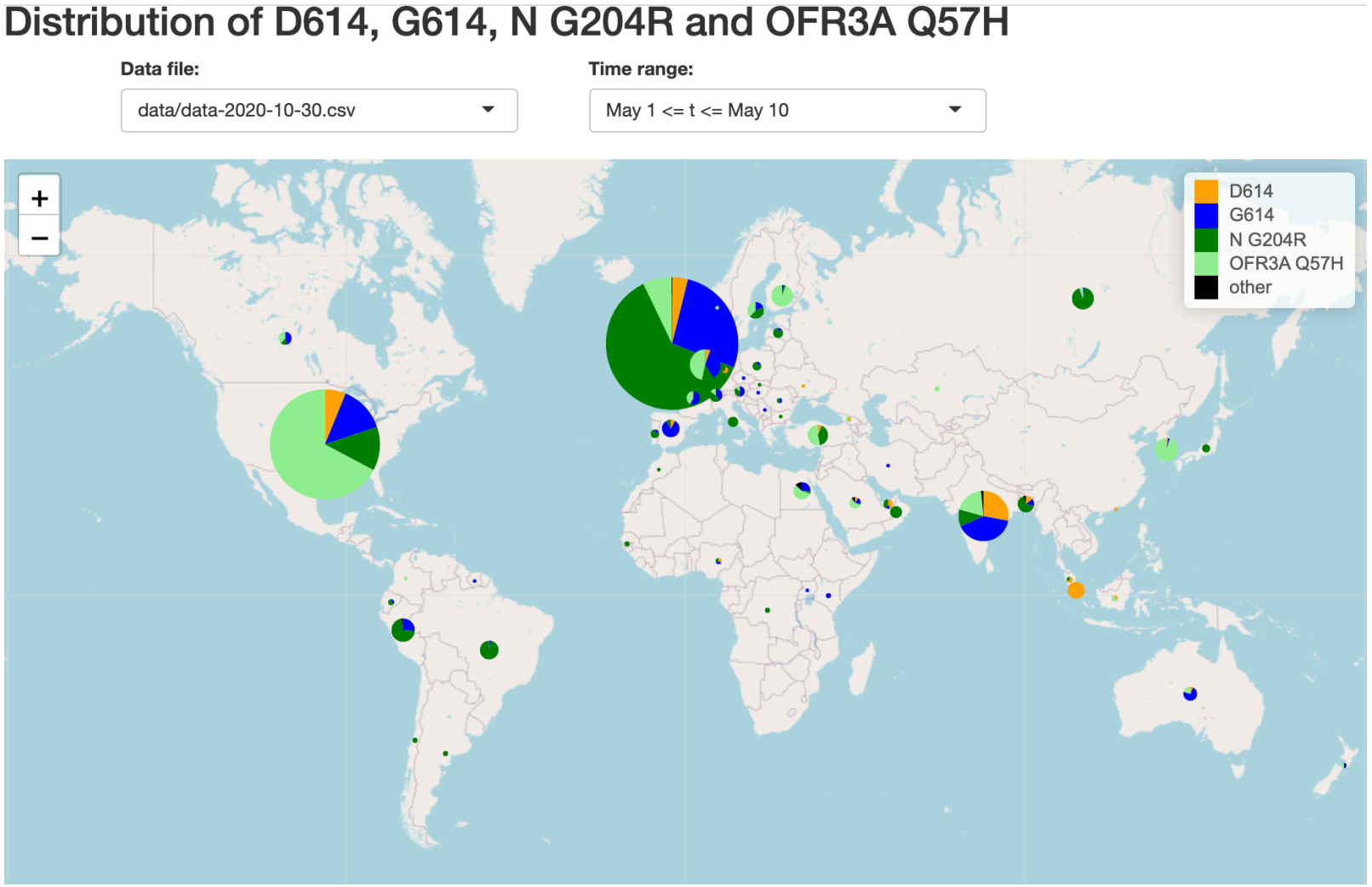
Mutational clade distributions

- Distributions sized by total sequence count (not case count!)
- Two-week intervals from March 1, 2020
- After March, "G" and descendent clades predominate worldwide



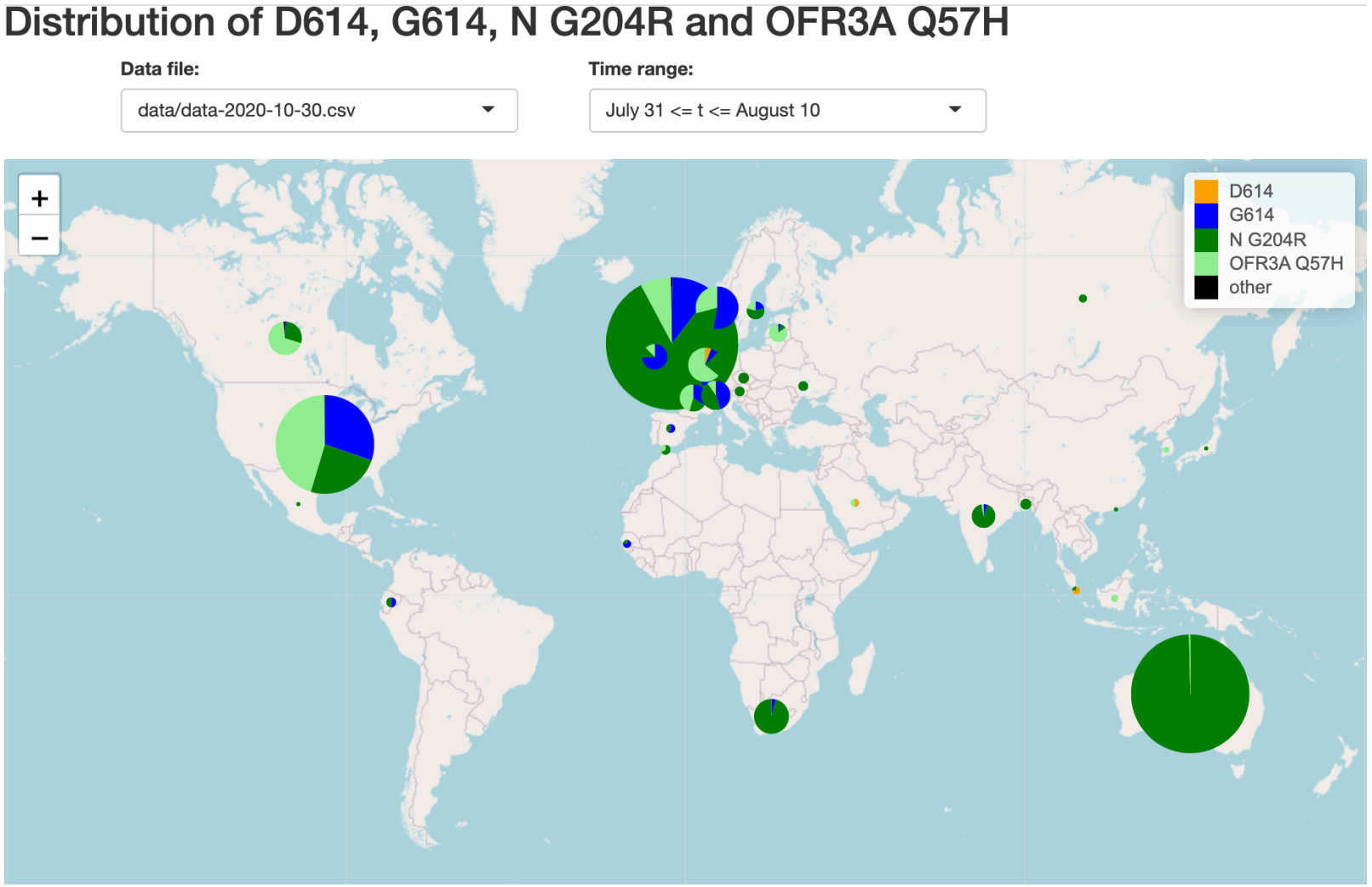
Mutational clade distributions

- Distributions sized by total sequence count (not case count!)
- Two-week intervals from March 1, 2020
- After March, "G" and descendent clades predominate worldwide

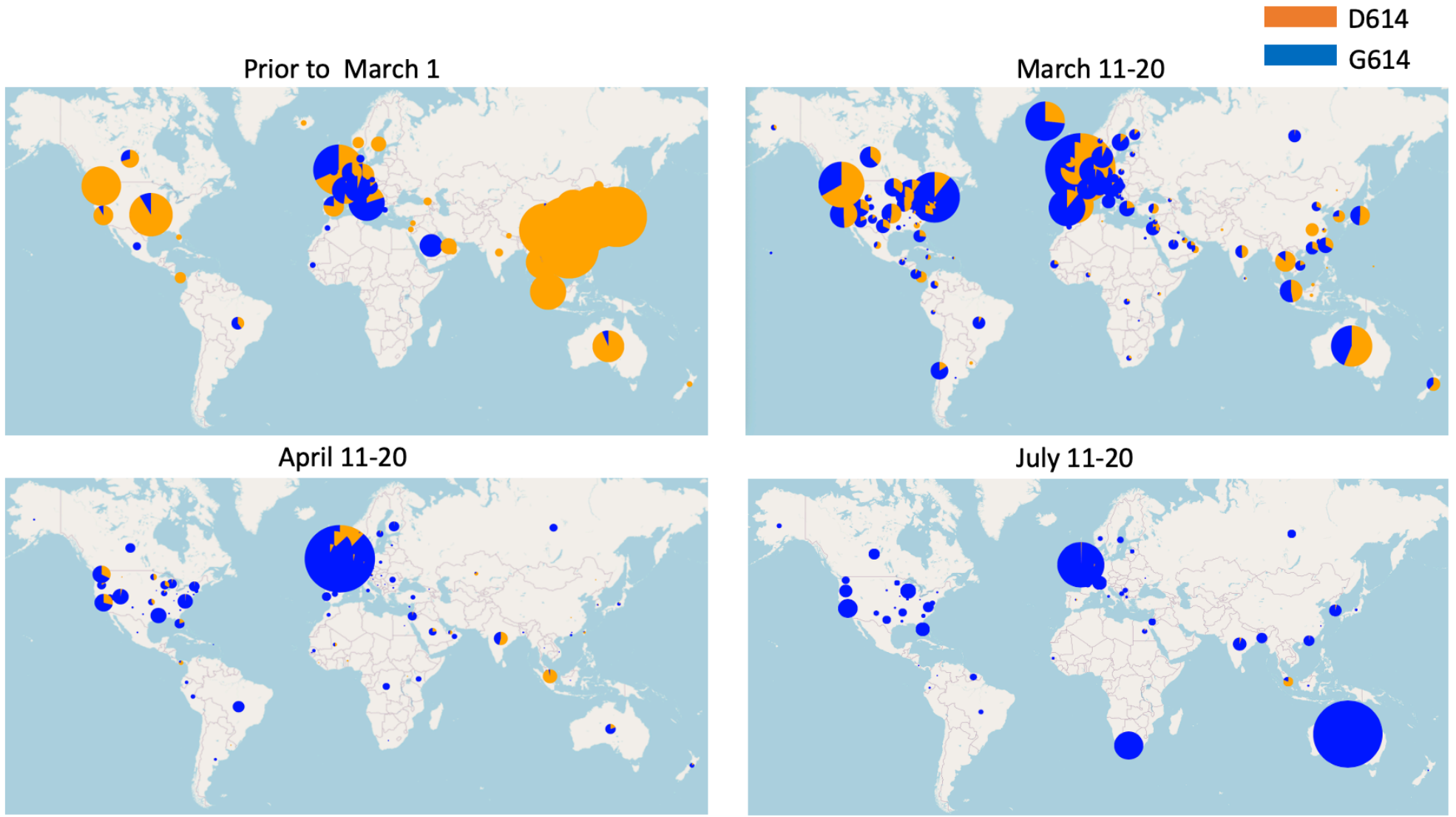


Mutational clade distributions

- Distributions sized by total sequence count (not case count!)
- Two-week intervals from March 1, 2020
- After March, "G" and descendent clades predominate worldwide



A rapid change in frequency of the G614 form was apparent early in 2020



Geographic/temporal mutation tracking in the COVID-19 pandemic:
https://cov.lanl.gov

Fueled by GISAID

The D614G mutation

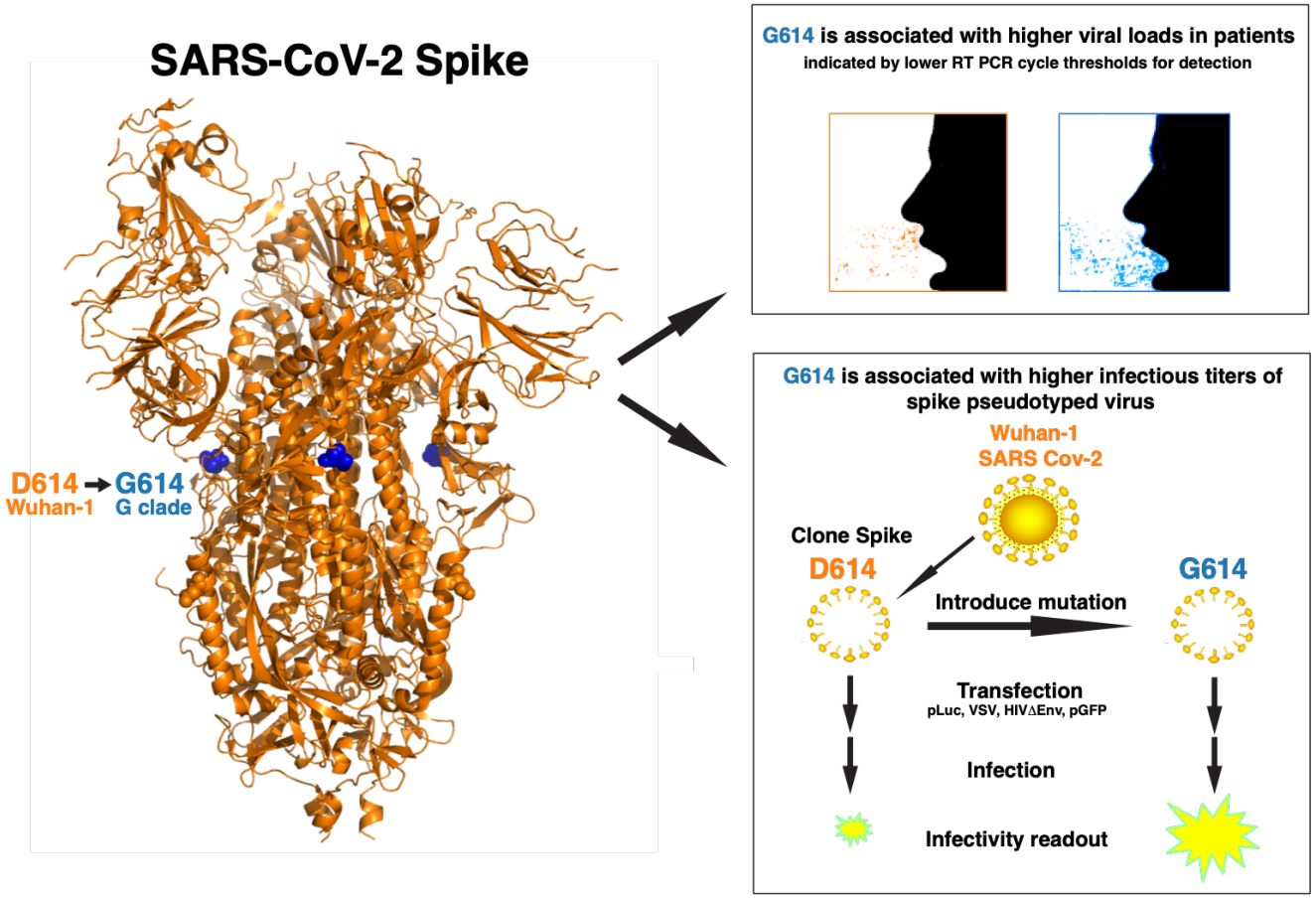
SARS-CoV-2 Spike is the protein that mediates virus entry into cells and it's the prime target of COVID vaccines.

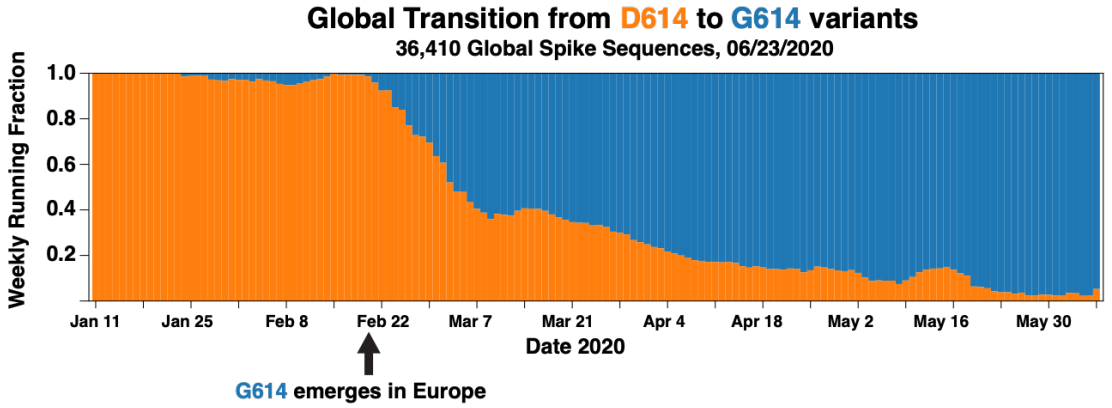
A single mutation in Spike, D614 to G614, has become the dominant form of the virus in the pandemic.

The G614 form is more infectious.

The G614 form is associated with higher levels of viral RNA in the upper respiratory tract of infected people.

The G614 was *not* associated with increased hospitalization.

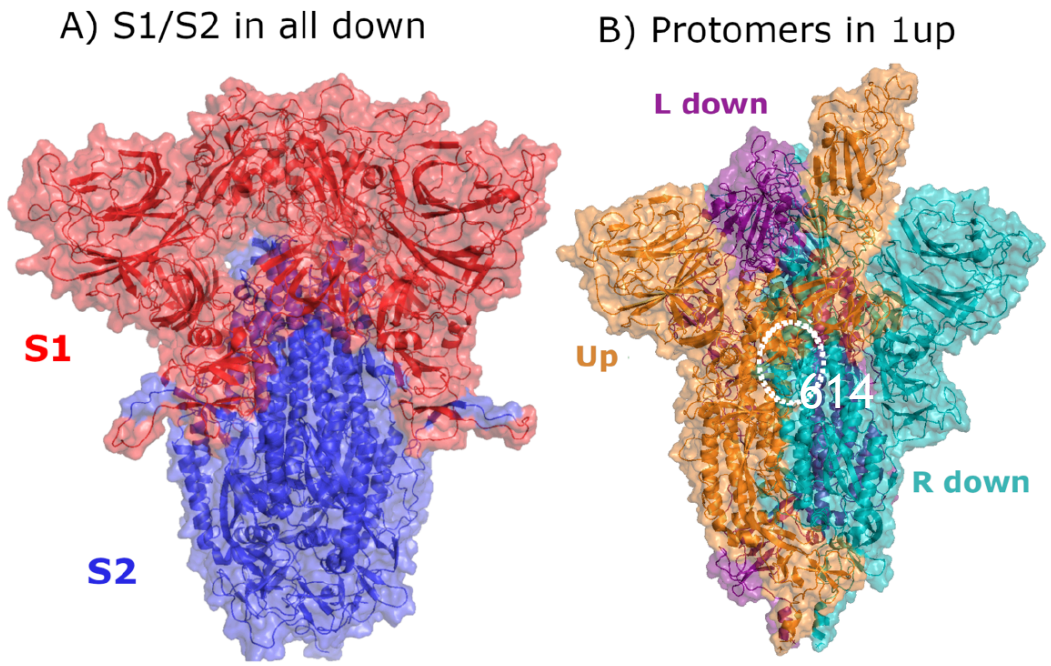




Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus.
Korber B, Cell. 2020 Aug 20;182(4):812-827.e19.

GISAID

G614 is likely to be more infectious because it favors an open "one-up" conformation that makes its ACE2 receptor binding site more accessible.

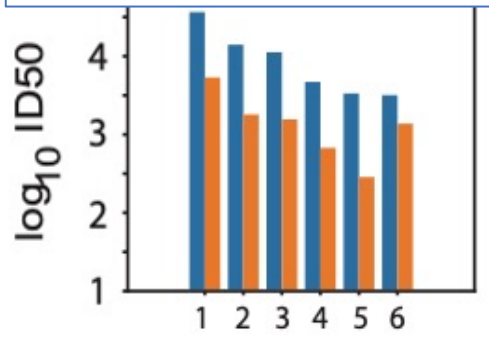


The SARS-CoV-2 Spike Variant D614G Favors an Open Conformational State
Mansbach et al. submitted and under review,
also in: bioRxiv. 2020 doi: 10.1101/2020.07.26.219741.
LANL/Duke collaboration

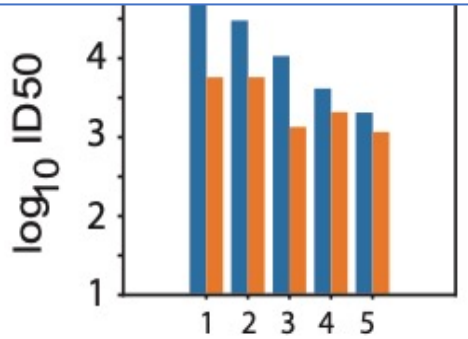
We were concerned that the more infectious G614 mutated virus might be more resistant to antibodies.

Instead, it is even *more* sensitive to vaccine-induced antibodies, COVID-19 convalescent sera, and Spike antibodies. This is also likely a result of favoring the 1-up conformation.

G614 Spike is more sensitive to sera from Spike vaccinated monkeys than D614



G614 Spike is more sensitive to sera from Spike vaccinated people than D614



D614G Spike Mutation Increases SARS CoV-2 Susceptibility to Neutralization.
Weissman et al. in press, Cell Host and Microbe, Oct. 2020
Also in: medRxiv: doi.org/10.1101/2020.07.22.20159905
Duke/U. Penn./LANL collaboration

Sites of particular interest in Spike, Nov. 2, 2020

- The original viruses that carried **D614** are very rarely now sampled globally. Among the **D614G** G clade viruses, GH and GR clades have emerged. The GR clade is now globally the most common, and is tending to increase in frequency relative to the G and GH clade.
  - The defining amino acid substitutions for the GR clade are outside of Spike
- The Spike **S477N** mutation virus started to dominant the sample in Australia in the summer, particularly in sequences from Victoria, and is now 6.3% of the global GISAID sample. It is increasing in the UK and several countries in Europe as well.
  - It is always found in a D614G context, but arises independently within the G, GR and GH clades
- The Spike **A222V** mutation has overtaken S477N as the second most common mutation in GISAID it is now at 7.9% of the global sample.
  - It is always found in a D614G context
  - It is most common found in the UK, but in increasing now in several countries.
- The Spike **N439K** mutation is becoming more common in the UK. It is currently present in 0.78% of the global sample. It is of particular interest because it is in the most common mutation in the RBM
  - The UK is by are the most heavily sampled country in GISAID, so this biases the global sample. It is not very common outside the UK at the point.
  - It is always found in a D614G context.
  - Embedded in several known neutralizing antibody epitope
    - Starr TN, et al. … Bloom JD. Cell. 2020 Aug 11:S0092-8674(20)31003-5. doi: 10.1016/j.cell.2020.08.012. PMID: 32841599
  - It can confer escape and is a selected resistance mutation of Nab C135
    - Weisblum Y et al. …Bieniasz PD. bioRxiv. 2020 Jul 22:2020.07.21.214759. doi: 10.1101/2020.07.21.214759. Preprint. PMID: 32743579

Example: tracking variation in site Spike 477

Summary of Spike Mutations: Spreadsheet

- A unannotated version of both of the spreadsheet tables that summarize variation are provided with a daily updates from GISAID at *cov.lanl.gov*

- The tab labeled "Spike Variation" contains a row for each position in Spike that includes:
  - The number of each variant, the entropy of each site, the local entropy of each 10 amino acid stretch
  - Spikes sites with > 0.3% variation (or within 4 Å of ACES, 0.1%) in GISAID are highlighted in red
  - Sites and local linear regions that have relatively high entropy are highlighted in yellow
  - Sites are annotated with Spike regions and mAb features are informed by f the following sources:
    - Starr TN, et al. … Bloom JD. Cell. 2020 Aug 11:S0092-8674(20)31003-5. doi: 10.1016/j.cell.2020.08.012. PMID: 32841599
      Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding
      Annotation is based on: https://jbloomlab.github.io/SARS-CoV-2-RBD_DMS/
    - Weisblum Y et al. …Bieniasz PD. bioRxiv. 2020 Jul 22:2020.07.21.214759. doi: 10.1101/2020.07.21.214759. Preprint. PMID: 32743579
      Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants
    - Barnes CO, et al. … Bjorkman PJ. Cell. 2020 Aug 20;182(4):828-842.e16. doi: 10.1016/j.cell.2020.06.025 .PMID: 32645326
      Structures of Human Antibodies Bound to SARS-CoV-2 Spike Reveal Common Epitopes and Recurrent Features of Antibodies

- The tab labeled "Common Spike mutations", summarizes amino acids that on Nov 2. had >0.3% variants
  - We excluded D614G because otherwise it overwhelms the information as it is globally dominant now
  - For other varying sites we provide counts, codons, amino acid variants, and geographic regions
  - A version of this table is provided with a daily update is available at cov.lanl.gov

Example of data in the spread sheet

The variation summary in this table is updated daily with new GISAID entries at: https://cov.lanl.gov/components/sequence/COV/int_sites_tbls.comp

Spike Regions	Barnes... Bjorkman	Ref seq posit	AA from NC_045512	10 mer, includes alignment gaps	011/02/2020 common form and countt	11/02/2020 Variants out of 135,316 Spikes in the alighment	11/02/2020 Entropy	11/02/2020 Avg Entropy 10-mer	<4 Angstoms from bound ACE2	Starr... Bloom RBD mutations ACE2 binding, from: jbloomlab.github.io	Starr... Bloom RBD mutations impact on ACE2 binding, focusing on sites with natural variantss: enhance, reduce, from: jbloomlab.github.io	Starr... Bloom RBD mutations impact on expression, with a focus natural variants in variable positions: enhance, reduce, from: jbloomlab.github.io	lst seq position	Barnes... Bjorkman. 2020 mAb C105 contacts	Pinto... Corti, 2020. <4 A from the bound S309 Fab, strucutre PDB 6WPT	Per. Comm. Gnanakaran & Chakraborty <4 A from the S309 epitope when bound to a one up-RBD	<4 A from the S309 epitope when bound to a down-RBD	Starr... Bloom mAb B38 epitope jbloomlab.github.io	Starr... Bloom mAb S309 epitope jbloomlab.github.io	Starr... Bloom mAb CR3022 epitope jbloomlab.github.io	Starr... Bloom mAb VHH-72 epitope jbloomlab.github.io	Starr... Bloom mAb 80R epitope jbloomlab.github.io	Starr... Bloom mAb S230 epitope jbloomlab.github.io	Starr... Bloom mAb F26G19 epitope jbloomlab.github.io
RBM	Domain B	472	I	IYQAGSTPCN	I(135305)	V(7) X(4)	0.001	0.0257					472										
RBM	Domain B	473	Y	YQAGSTPCN	Y(135311)	F(1) X(4)	0	0.0256					473	<4Å									
RBM	Domain B	474	Q	QAGSTPCNG	Q(135309)	H(1) X(6)	0	0.0262					474	<4Å									
RBM	Domain B	475	A	AGSTPCNGV	A(135296)	V(14) X(6)	0.001	0.0266	ACE2 <4 A	ACE2-binding	V -0.14	V -0.21	475	<4Å									
RBM	Domain B	476	G	GSTPCNGVE	G(135274)	A(1) F(1) S(26) X(14)	0.002	0.0266	ACE2 <4 A		S -0.02	S -0.06	476	<4Å									
RBM	Domain B	477	S	STPCNGVEG	S(126710)	G(2) I(54) N(8481) R(17) T(2) X(50)	0.239	0.0269			N 0.06	N 0.06	477	<4Å									
RBM	Domain B	478	T	TPCNGVEGF	T(135185)	A(2) I(103) K(9) R(1) X(16)	0.007	0.003			I -0.04	I -0.16	478										
RBM	Domain B	479	P	PCNGVEGFN	P(135215)	L(3) S(91) X(7)	0.006	0.0023			S -0.03	S -0.20	479										
RBM	Domain B	480	C	CNGVEGFNC	C(135311)	F(1) W(1) X(3)	0	0.0017					480										
RBM	Domain B	481	N	NGVEGFNCY	N(135296)	D(12) K(3) X(5)	0.001	0.0018					481										
RBM	Domain B	482	G	GVEGFNCYF	G(135308)	S(4) X(4)	0	0.0017					482										
RBM	Domain B	483	V	VEGFNCYFP	V(135213)	A(49) F(38) I(3) L(3) X(10)	0.006	0.0017			A -0.03	A 0.09	483										
RBM	Domain B	484	E	EGFNCYFPL	E(135260)	A(4) D(2) K(16) Q(24) R(1) X(9)	0.004	0.0012			Q 0.03, K 0.06	Q -0.08, K 0.1	484										
RBM	Domain B	485	G	GFNCYFPLQ	G(135276)	R(8) S(2) X(30)	0.001	0.0013					485	6.5-8.5Å									
RBM	Domain B	486	F	FNCYFPLQSY	F(135229)	I(1) L(84) X(2)	0.005	0.0012	ACE2 <4 A	ACE2-binding	L -0.47	L 0.03	486	4-6.5Å									
RBM	Domain B	487	N	NCYFPLQSY	N(135310)	H(1) I(1) K(1) X(3)	0	0.0007	ACE2 <4 A	ACE2-binding			487	4-6.5Å									
RBM	Domain B	488	C	CYFPLQSYG	C(135310)	R(1) W(1) X(4)	0	0.0007					488	6.5-8.5Å									

Steps:

1) Tables of variation in Spike
    1) Summaries of sites that have >0.3% variation, updated daily: under 477, you can see mutations in this site are by far the most common in Australia
    2) Variation in all positions in Spike are be tracked

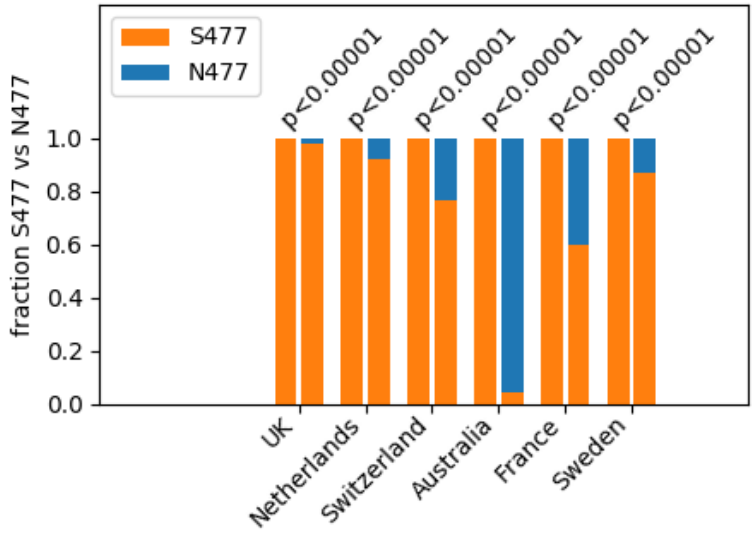
2) Tracking places where variation is accruing:
    1) New mutations are accruing in a G614 background, so exclude D614 the original form
    2) Using Isotonic Regression
    3) Using Relative Frequency Change By Geographical Region

3) Tracking mutations: Follow variation in a position over time in a given place: e.g. 477 Australia

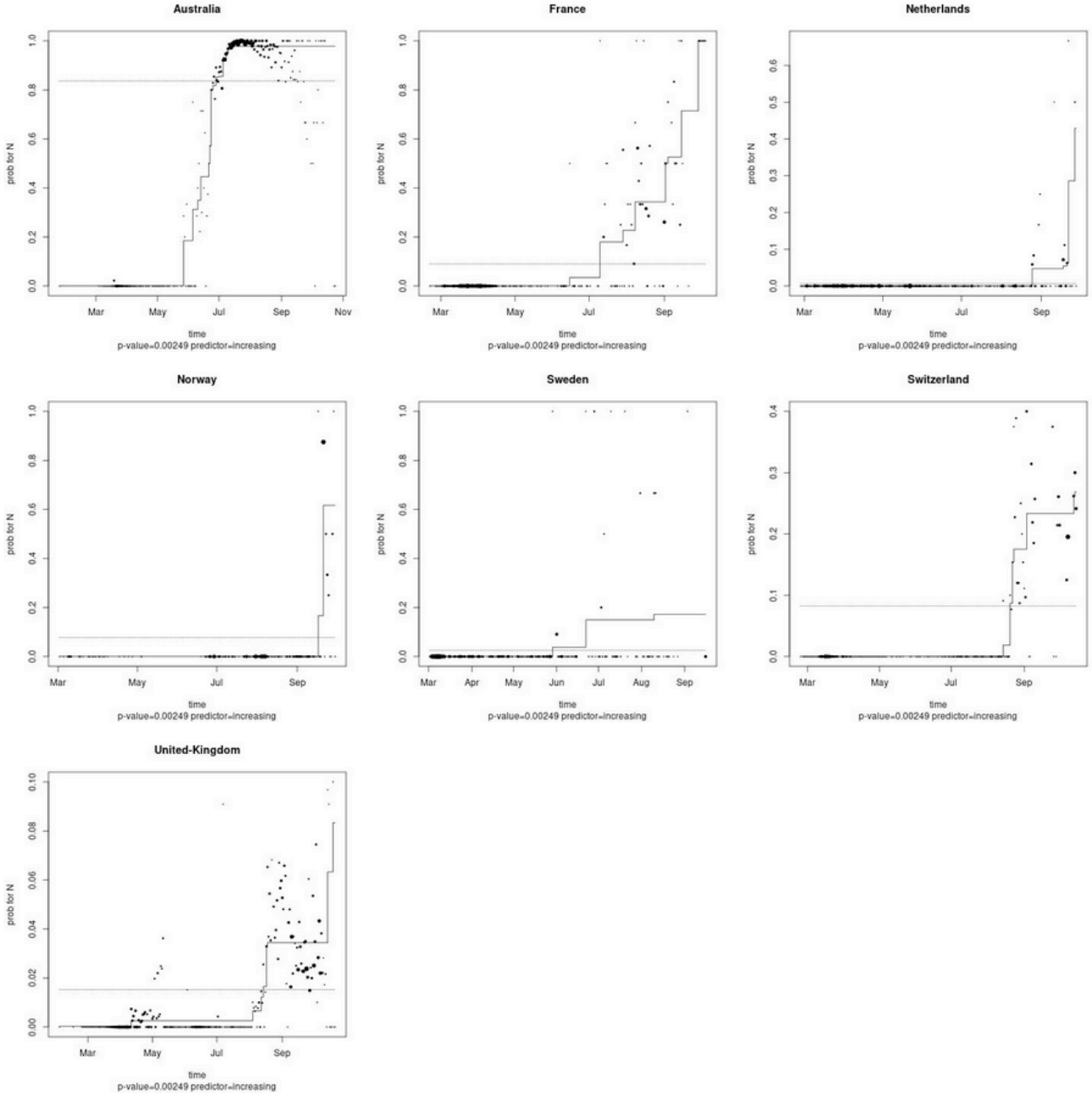
4) Analyze Align: e.g. How often are mutations in 614, 222, 439, 477 tracking together?

The mutation S477N has recently increased to 6.3% of GISAID, and is beginning to significantly increase in a number of countries

Relative Frequency Change By Geographical Region



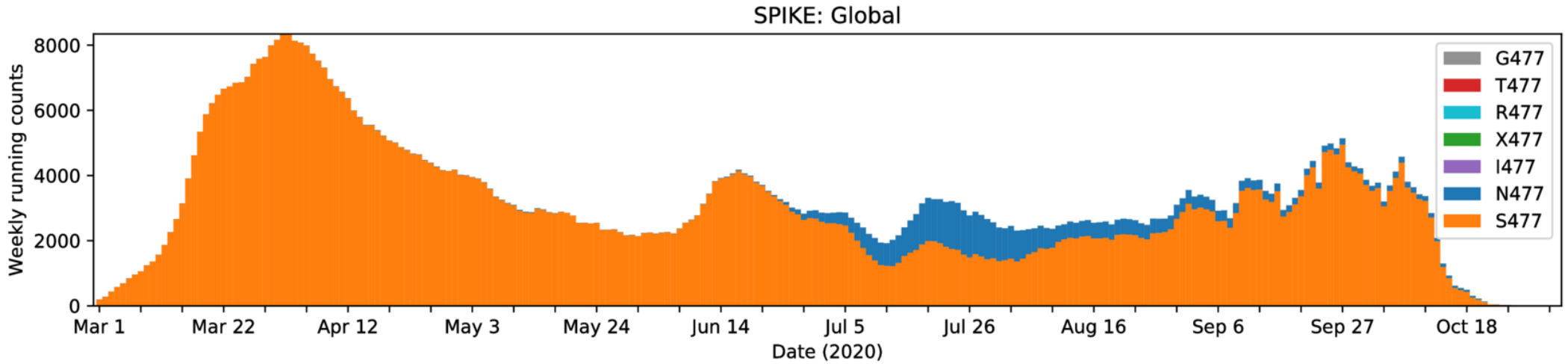
Isotonic Regression



Tracking mutations:
The frequency S477N in the global sample is dominated by Australia

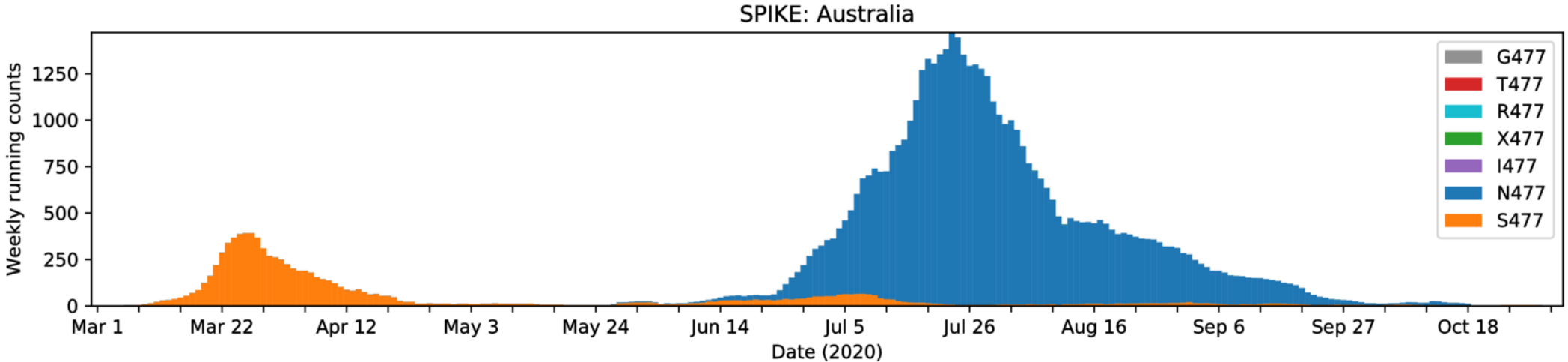
SPIKE 477 Global: 116454 good entries: 7% S477N total, 6% from Australia, 1% from other places

S477: 108094, N477: 8240, I477: 54, X477: 48, R477: 16, T477: 2, G477: 0



SPIKE 477 Australia: 8391 good entries. S477N is 83.6% of the Australian sample.

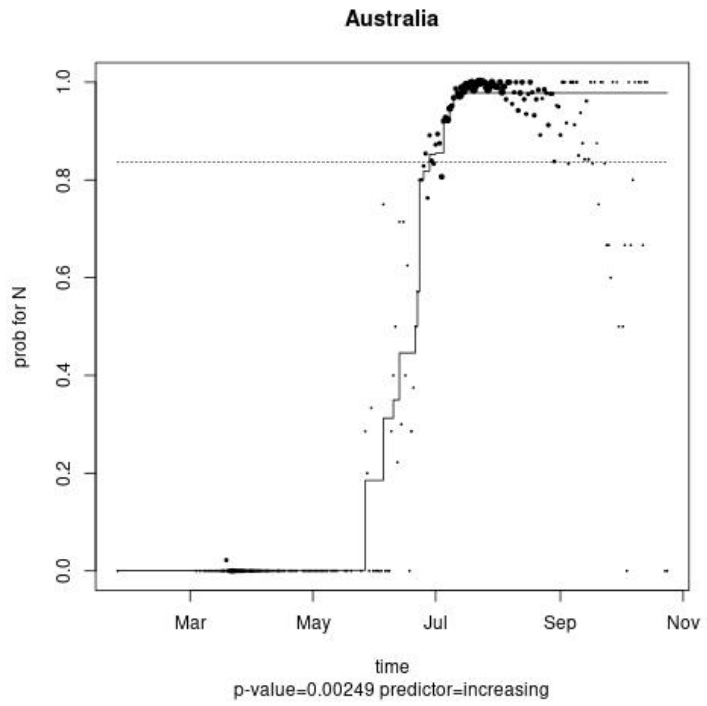
S477: 1373, N477: 7017, I477: 0, X477: 1, R477: 0, G477: 0, T477: 0

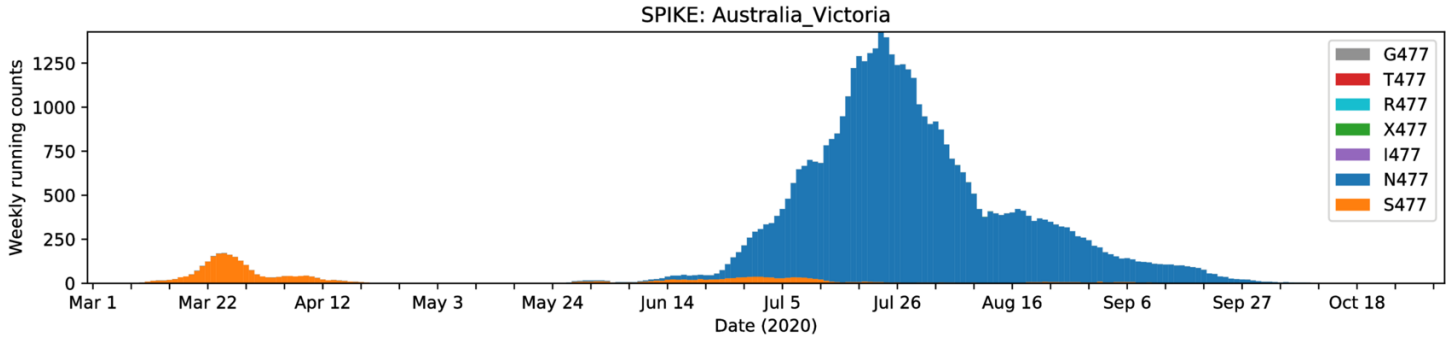


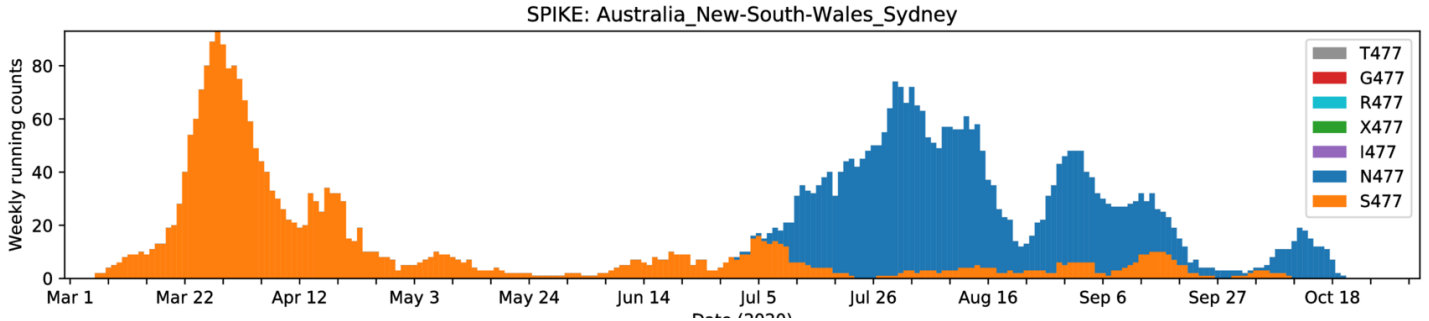
Australia 477

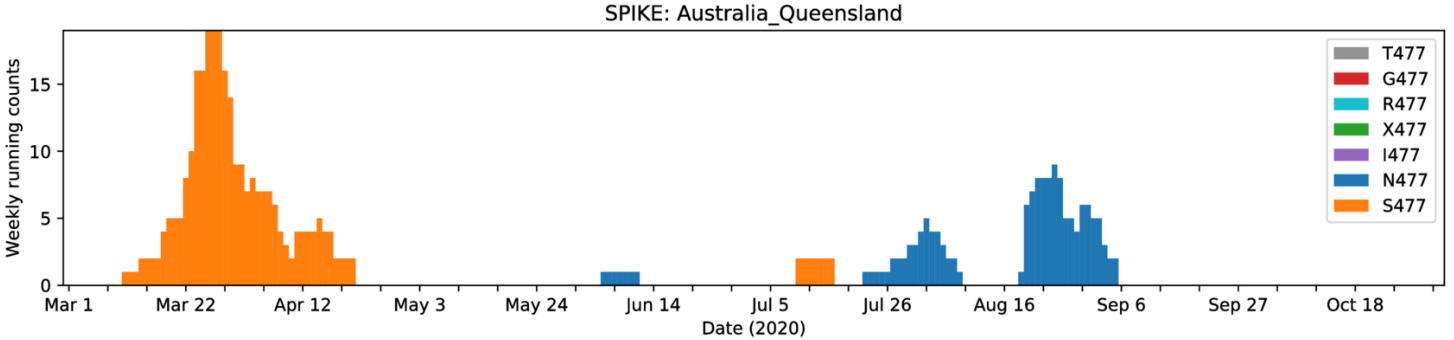
First became common in GISAID due to over 6000 sequences from Victoria sampled and provided to GISAID during the late summer. Victoria had completely switched to S477N.

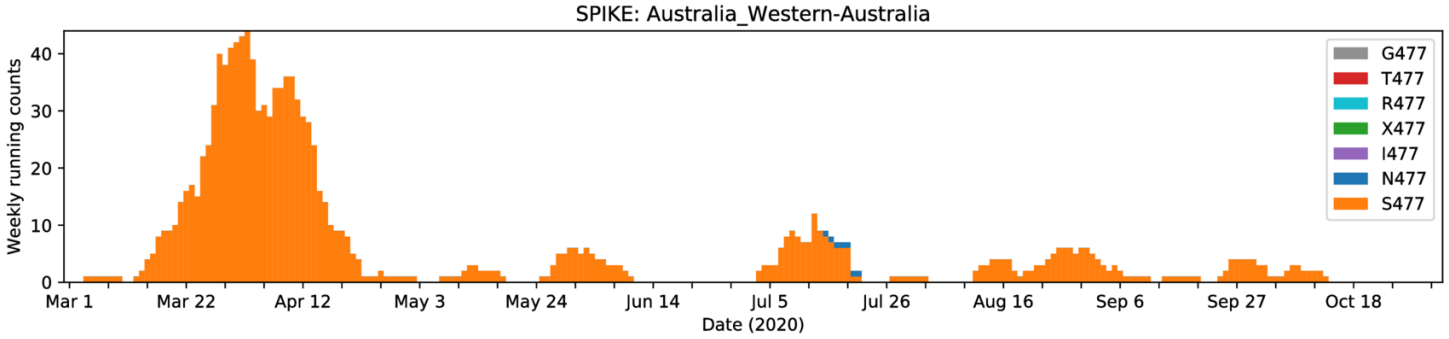
Most Australian samples reflect this switch (Sydney and Queensland), but not all (Western Australia)



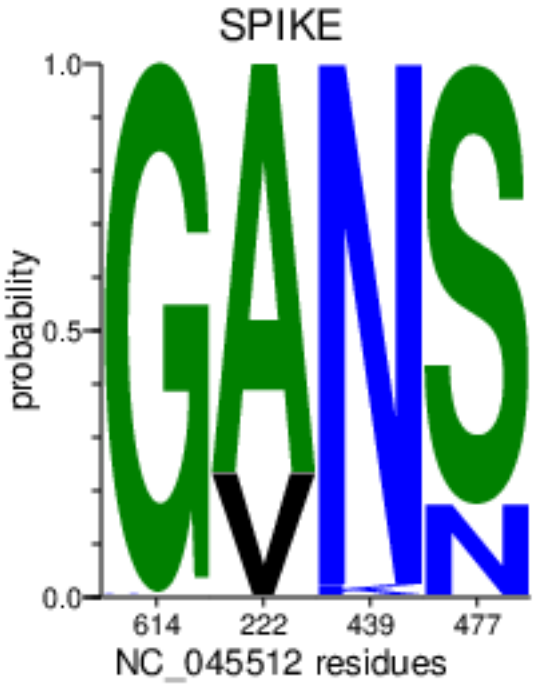








Analyze Align 7/1/2020 - 11/3/2020



Form	N muts	Count	Percent
GANS	0	25964	56.85
-V--	1	10510	23.01
---N	1	7749	16.97
--K-	1	977	2.14
D---	1	202	0.44
---I	1	41	0.09
-V-N	2	16	0.04

121 were found in Singapore

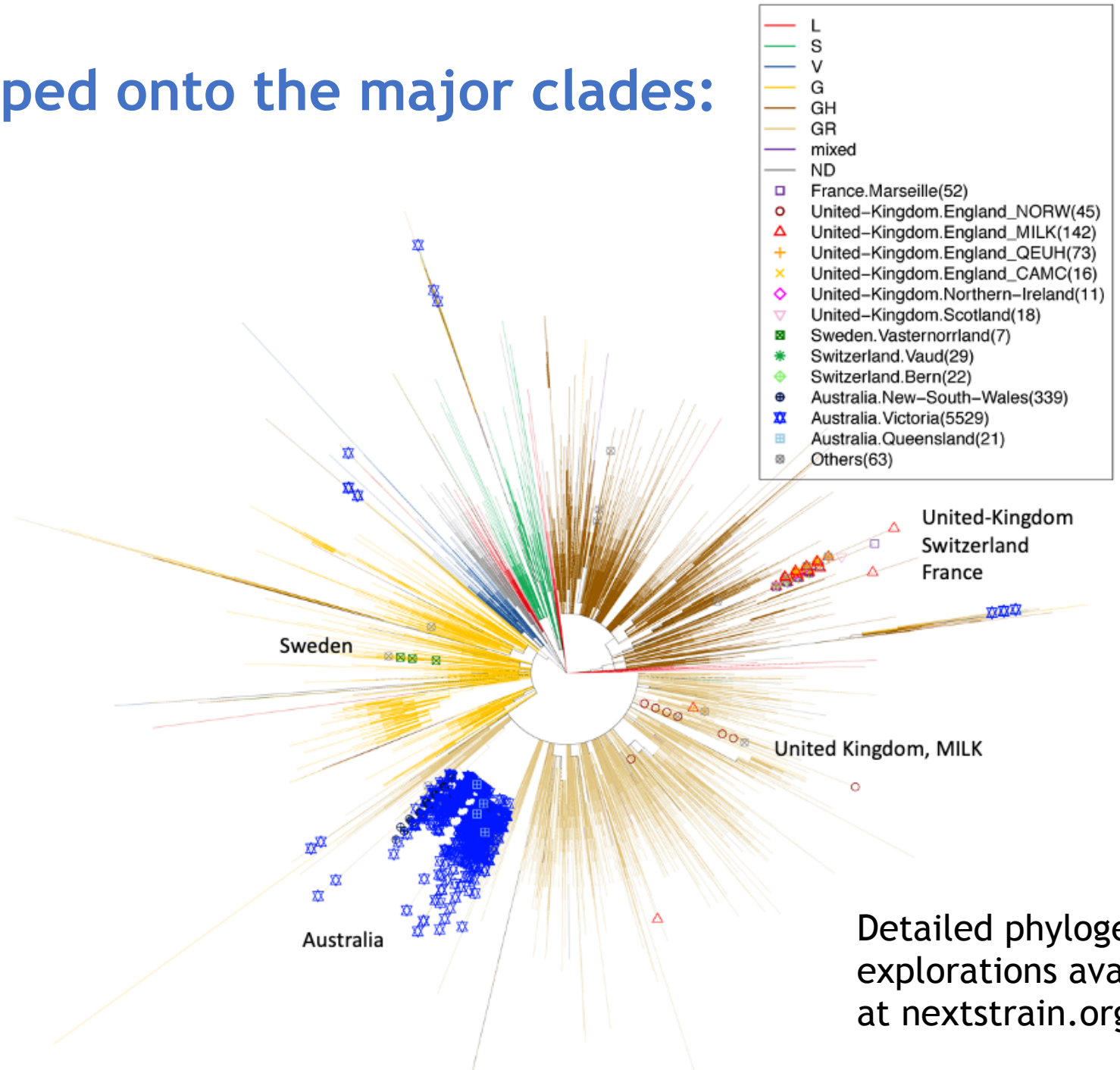
Combinations are very rare

The 477 mutation mapped onto the major clades:

Four clades
Australia: GR
UK, Switzerland, France: GH
UK, MILK: GR
Sweden: G

The scattered long branch
Australia, Victoria:

We hypothesize these are likely to be recombinants



Detailed phylogenetic explorations available at nextstrain.org