

# Molecular & Genomic Epidemiology of Infectious Diseases

---

## Lessons and Directions After COVID-19

Brendan J. Kelly, MD, MS

Infectious Diseases, Epidemiology & Microbiology

University of Pennsylvania

14 November 2023

# Disclosures

- No conflicts of interest.
- Research supported by:
  - NIAID K23 AI121485
  - CDC BAA 200-2021-10986

COVID-19  
Genomic  
Epidemiology

COVID-19  
Genomic  
Epidemiology

Biases  
Limitations  
& Unknowns

COVID-19  
Genomic  
Epidemiology

Biases  
Limitations  
& Unknowns

COVID-19  
Aftermath

COVID-19  
Genomic  
Epidemiology

Biases  
Limitations  
& Unknowns

COVID-19  
Aftermath

# Genomic Epidemiology

- **Epidemiology:**
  - study of the distribution and determinants of health-related outcomes in a specified population
- **Molecular Epidemiology:**
  - joins understanding of disease at the molecular level with population-based study designs and approaches; links epidemiology with laboratory sciences
- **Genomic Epidemiology:**
  - use of pathogen genomic data to determine the distribution and spread of an infectious disease in a specified population

# SARS-CoV-2 Genomic Epidemiology

- **SARS-CoV-2 Genomes:**
  - SARS-CoV-2 genome: ~ 30,000 nucleotides (~ 10,000 amino acids)
  - infer relatedness from phylogenetic distance
- **RNA Genome Sequencing:**
  - extract RNA, reverse transcription, multiplex PCR, end repair & ligation indexing
  - short-read or long-read sequencing



# National & Local Applications

- **National:**
  - monitor emergence and movement of new strains
  - monitor trends after intervention (e.g., vaccination)
- **Local:**
  - investigate clusters for transmission (workplace, healthcare, etc)
  - reveal unexpected clusters

# International Collaboration

## In Focus

### Submission Tracker

hCoV-19 Global  
hCoV-19 USA  
hMpxV  
RSV

### Phylodynamics

hCoV-19  
hMpxV  
Influenza  
RSV

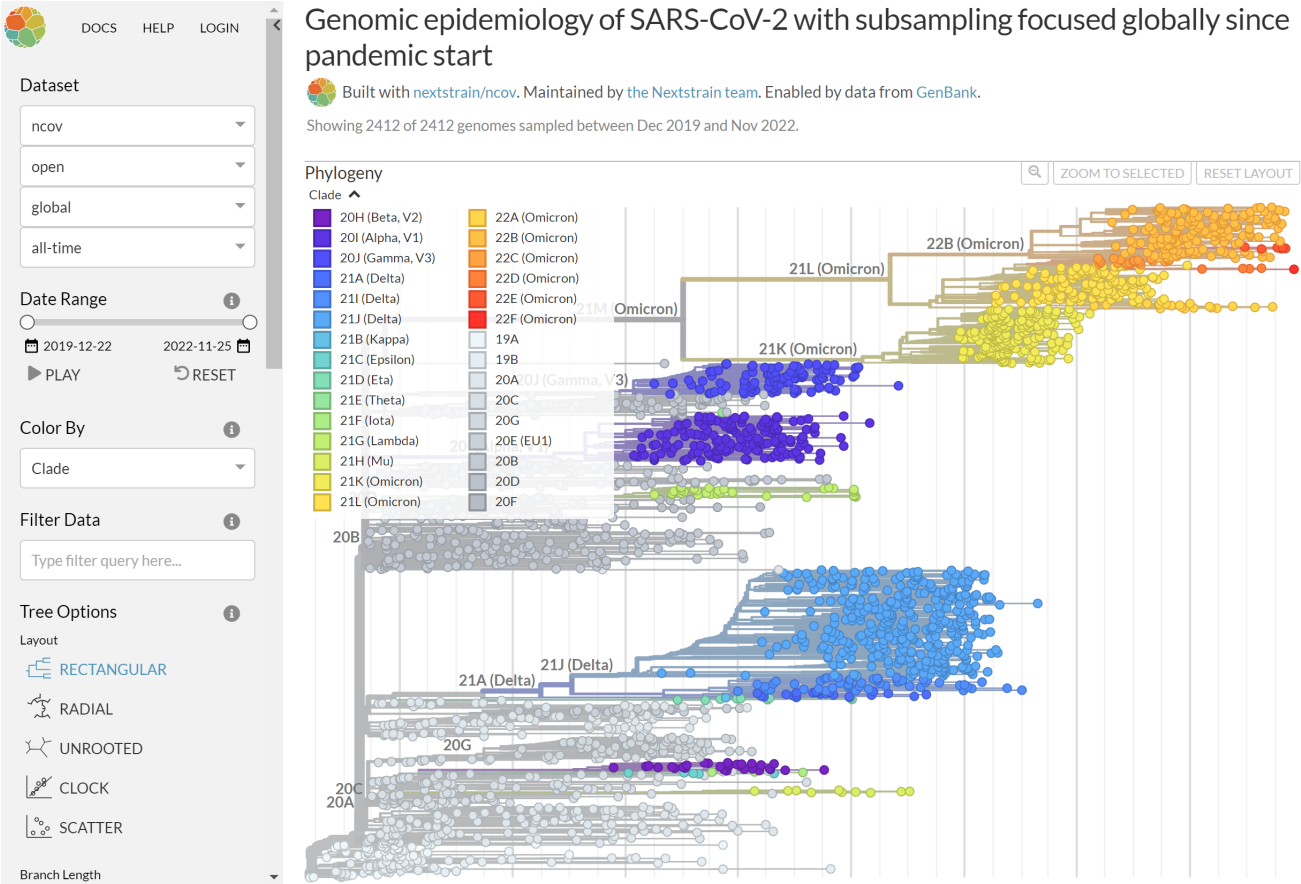
### Tracking Variants

hCoV-19 Variants  
hMpxV Variants  
Influenza Subtypes  
RSV Subtypes

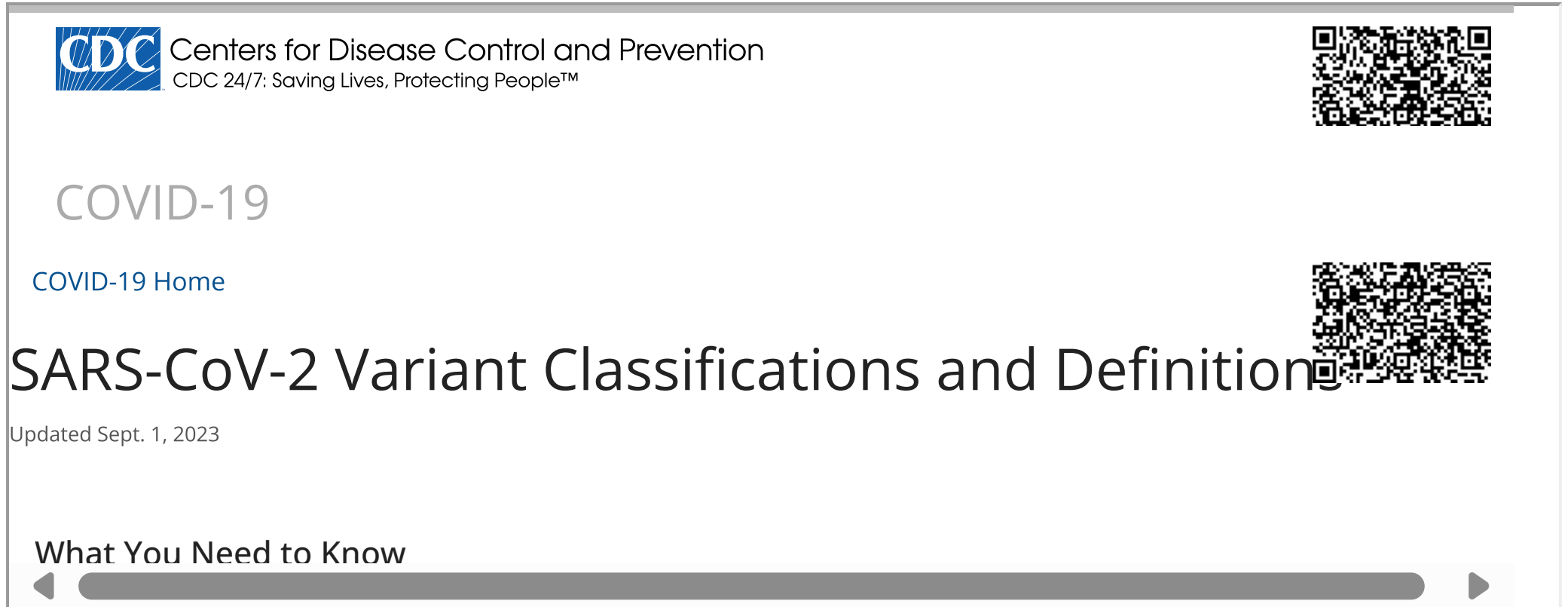
### Frequency Dashboards

hCoV-19  
hMpxV  
Influenza  
RSV

# International Collaboration



# Variants of Concern



The screenshot shows the top portion of a CDC webpage. At the top left is the CDC logo with the text 'Centers for Disease Control and Prevention' and the slogan 'CDC 24/7: Saving Lives, Protecting People™'. To the right of the logo is a QR code. Below the logo, the text 'COVID-19' is displayed in a large, light grey font, followed by a blue link 'COVID-19 Home'. The main title of the page is 'SARS-CoV-2 Variant Classifications and Definition' in a large black font, with another QR code to its right. Below the title, it says 'Updated Sept. 1, 2023'. At the bottom left, there is a section header 'What You Need to Know' above a grey progress bar with left and right navigation arrows.

**CDC** Centers for Disease Control and Prevention  
CDC 24/7: Saving Lives, Protecting People™

COVID-19

[COVID-19 Home](#)

## SARS-CoV-2 Variant Classifications and Definition

Updated Sept. 1, 2023

What You Need to Know

# Local Tracking

## SARS-CoV-2 Variants Circulating River Valley Tracked by Surveillance

As SARS-CoV-2 grows, the virus occasionally makes mistakes copying its genetic material. As it spreads in people, chance changes that increase replication in humans may lead to new variants. We are tracking these variants in order to understand their possible influences on the effectiveness of vaccines. Collaborators at the City of Philadelphia Department of Public Health, Jefferson Hospital and the University of Pennsylvania are sequencing to track the nature and spread of viral variants in the Delaware Valley. Samples are being collected from southwestern New Jersey, providing an overview of the dynamics in the Delaware Valley. Blog posts and reports on each genome individually.

The output of the sequencing effort to date is:

Most recent sequencing run: **2023-05-16**

Total number of sequenced samples: **7,977**

Sequenced samples with  $\geq 95\%$  genome coverage ( $\geq 5$  reads per position): **7,538**

Sequenced lineages (summaries/genomes/lineagesPlot.pdf) | Mutation tables (summaries/genomes/positionalMutationFreqTable.xlsx) | Run stats (summaries/seqRunSummary.zip) | Reports (summaries/SARS-CoV-2\_reports.zip) | Code base (<https://github.com/helixscript/SARS-CoV-2>)

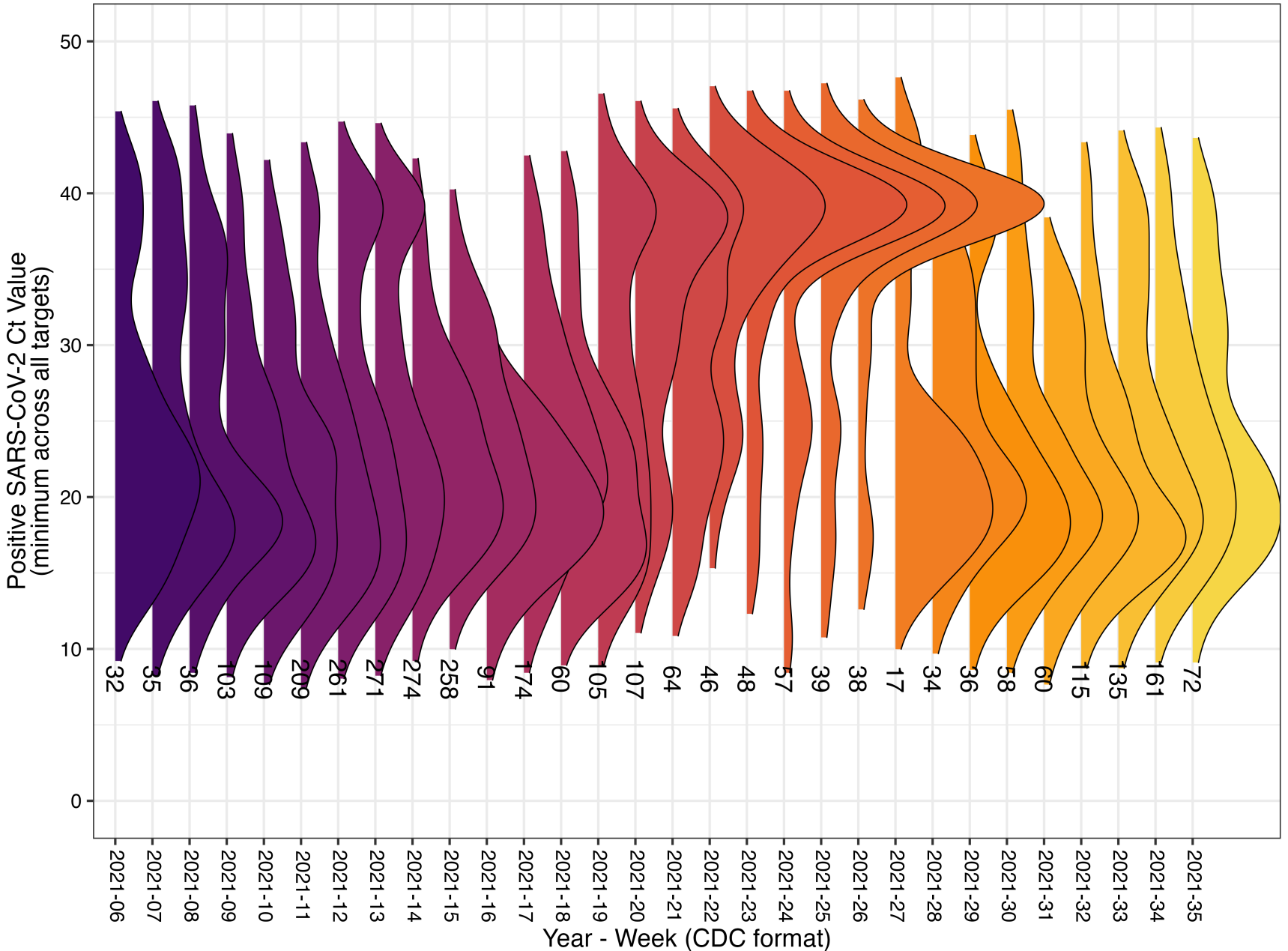
COVID-19  
Genomic  
Epidemiology

Biases  
Limitations  
& Unknowns

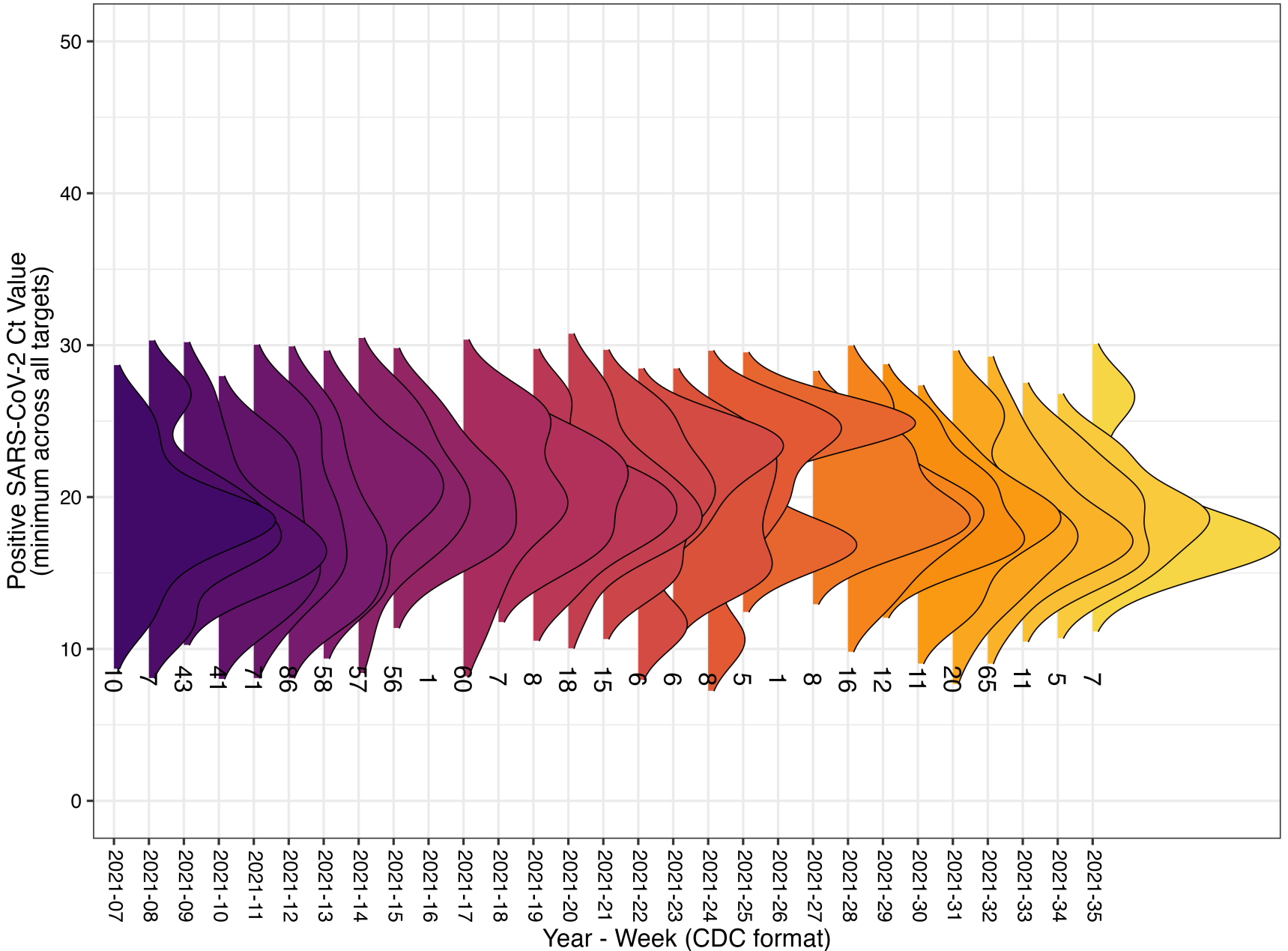
COVID-19  
Aftermath

# Biases: SARS-CoV-2 "Viral Load"

- The number of virions detectable in respiratory secretions varies between individuals and over the course of infection:
  - high viral loads are required for most sequencing-preparation pipelines
  - coverage/quality may be poor if viral loads are low
- What are the consequences of excluding low-viral-load samples from genomic epidemiology surveillance?







# Limitations: SARS-CoV-2 vs Population Immunity

- Molecular evolution occurs in the context of increasingly complex and heterogeneous population immunity:
  - waves of SARS-CoV-2 variants with different immune epitopes
  - the introduction of vaccines
- How best to measure SARS-CoV-2 molecular evolution in this context? (e.g., how best to perform "antigenic cartography"?)

# SARS-CoV-2 Molecular Evolution

- How do we measure SARS-CoV-2 genome change?
  - daily, weekly, monthly?
  - nucleotide, dN/dS, AA, gene?
  - how to pool within codons, genes?
  - covariance across genome?
- How does genomic change relate to changes in incidence?

# Genomic Positional Diversity

- Shannon diversity:

$$H' = - \sum p_i * \log_b (p_i)$$

(note: typically natural log or base 2 are used)

- richness: how many nucleotide variants are there?
- evenness: how are variants distributed?

# Aside on Information Theory

- Shannon diversity:

$$H' = - \sum p_i * \log_b (p_i)$$

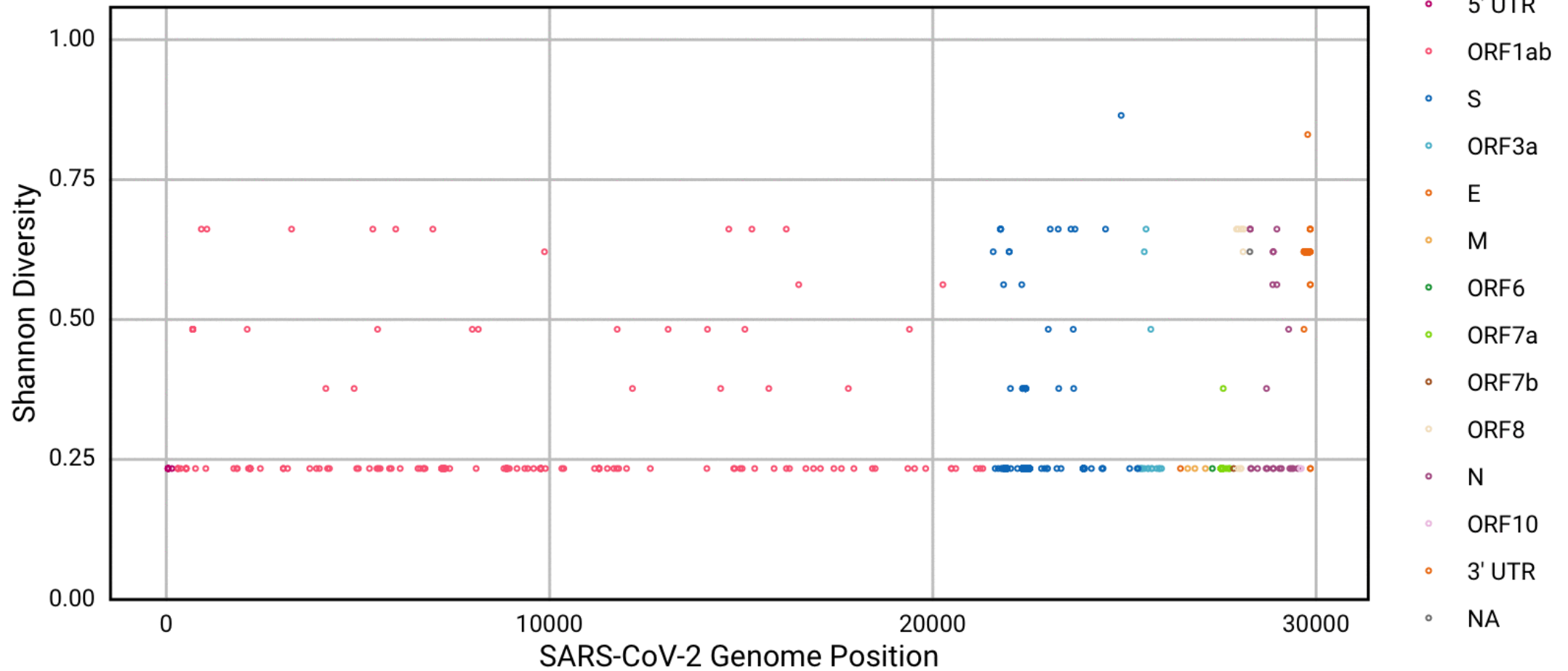
- Claude Shannon & information entropy:

$$H(p) = - \sum p_i * \log_b (p_i)$$

- "The uncertainty contained in a probability distribution is the average log-probability of an event."

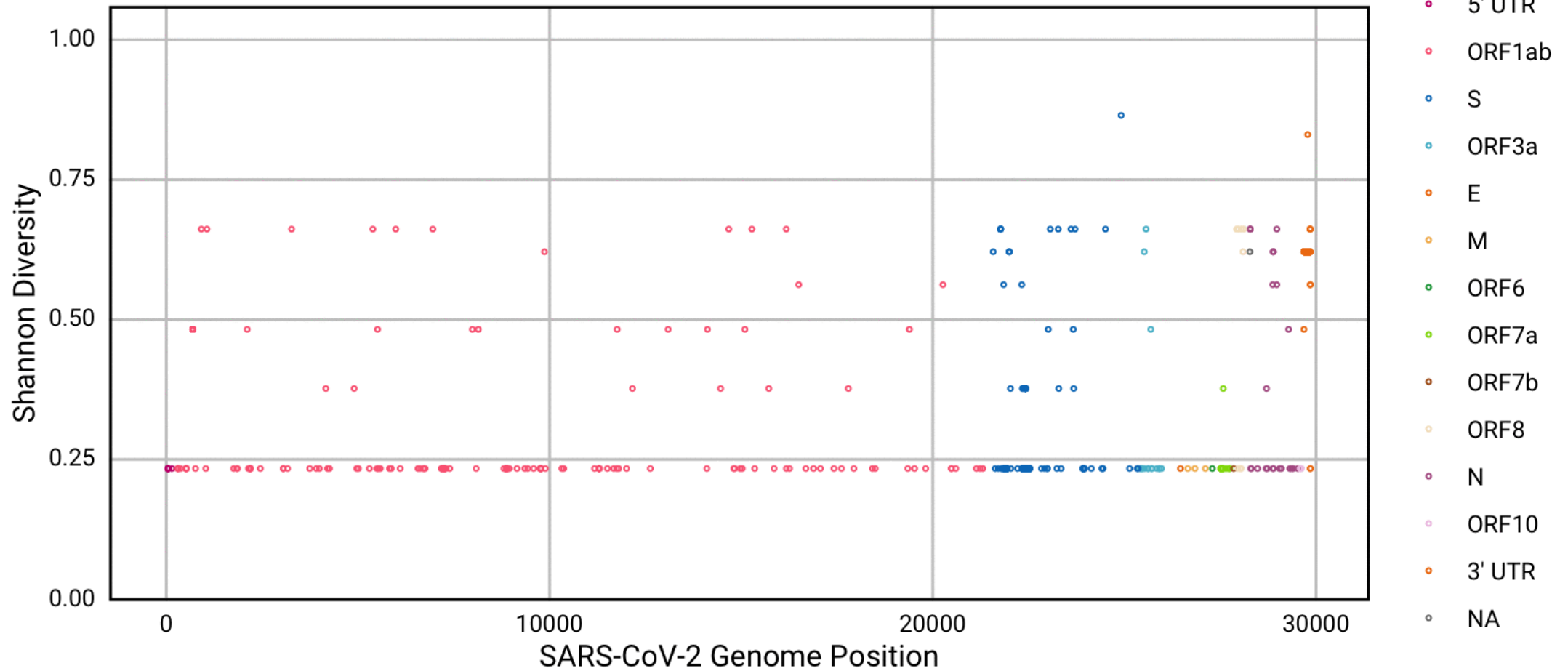
# Philadelphia Region COVID-19 Genomic Surveillance (CDC BAA 200-2021-10986)

Nucleotide Diversity Observed within One Week (single-week plot): 2021.38



# Philadelphia Region COVID-19 Genomic Surveillance (CDC BAA 200-2021-10986)

Nucleotide Diversity Observed within One Week (cumulative plot): 2021.38

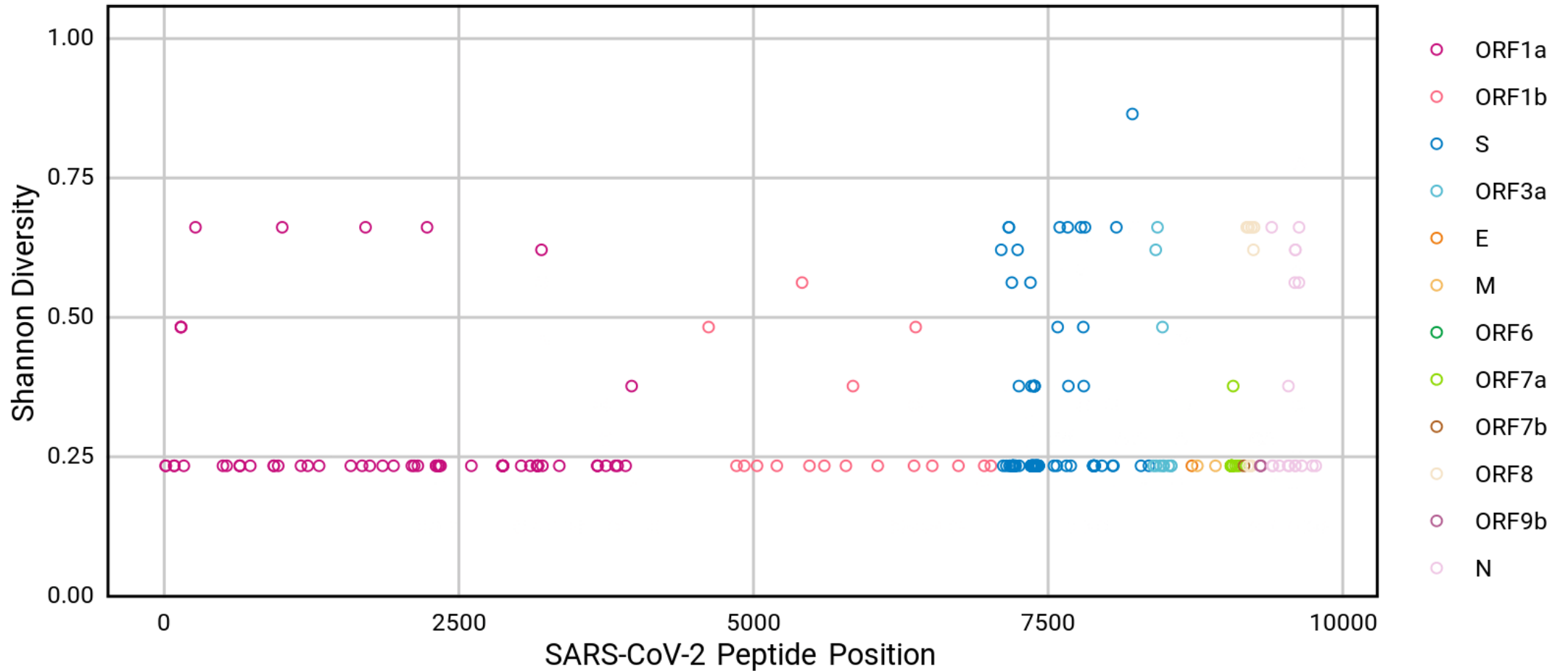






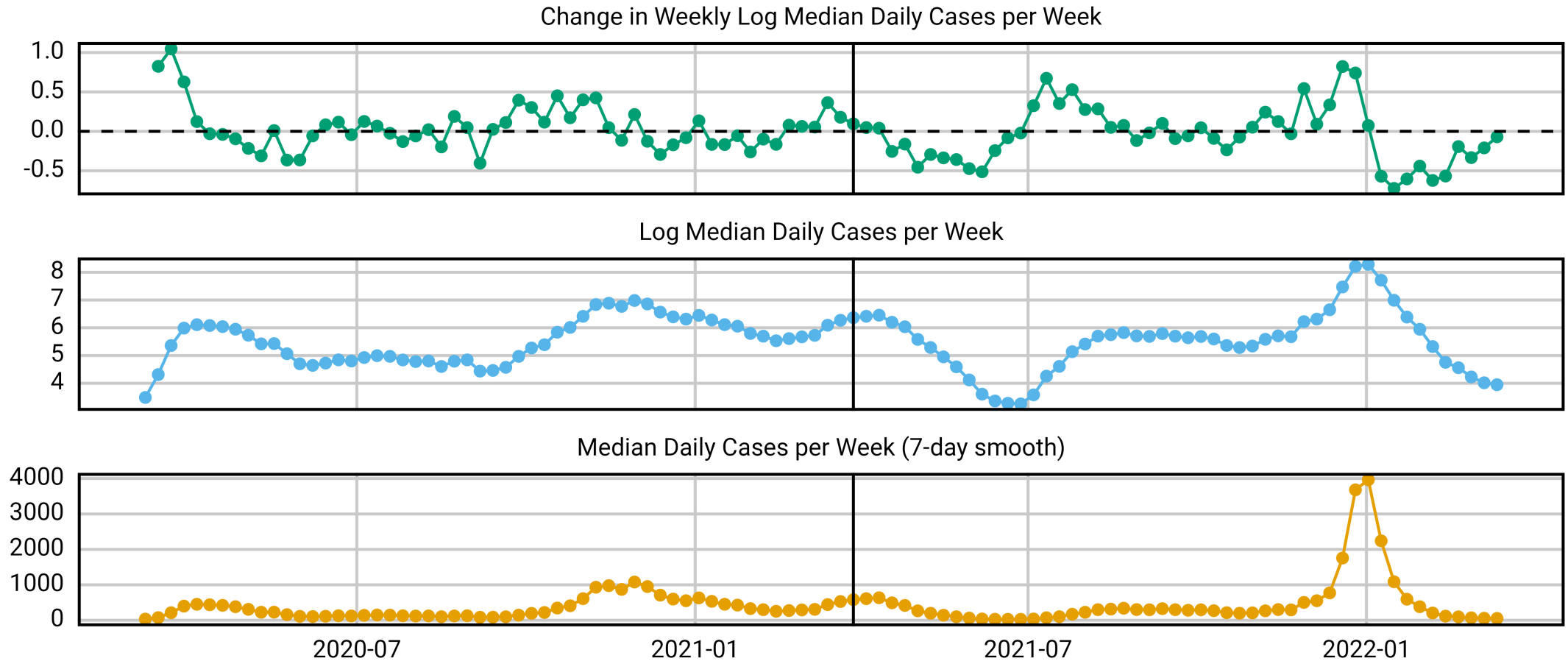
# PhiladelphiaRegion COVID-19 Genomic Surveillance (CDC BAA 200-2021-10986)

Peptide Diversity Observed within One Week (single-week plot): 2021.38



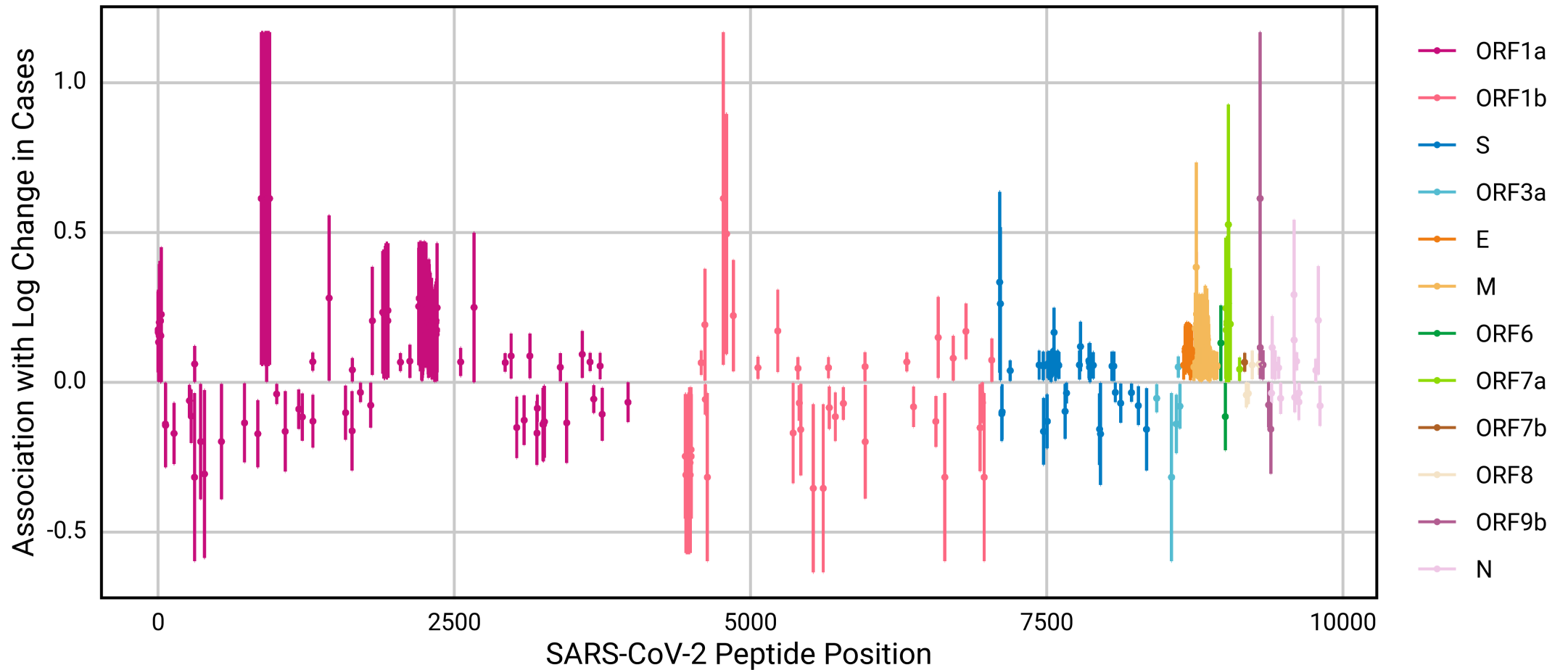
# Philadelphia Region COVID-19 Genomic Surveillance (CDC BAA 200-2021-10986)

## COVID-19 Case Incidence in the City of Philadelphia



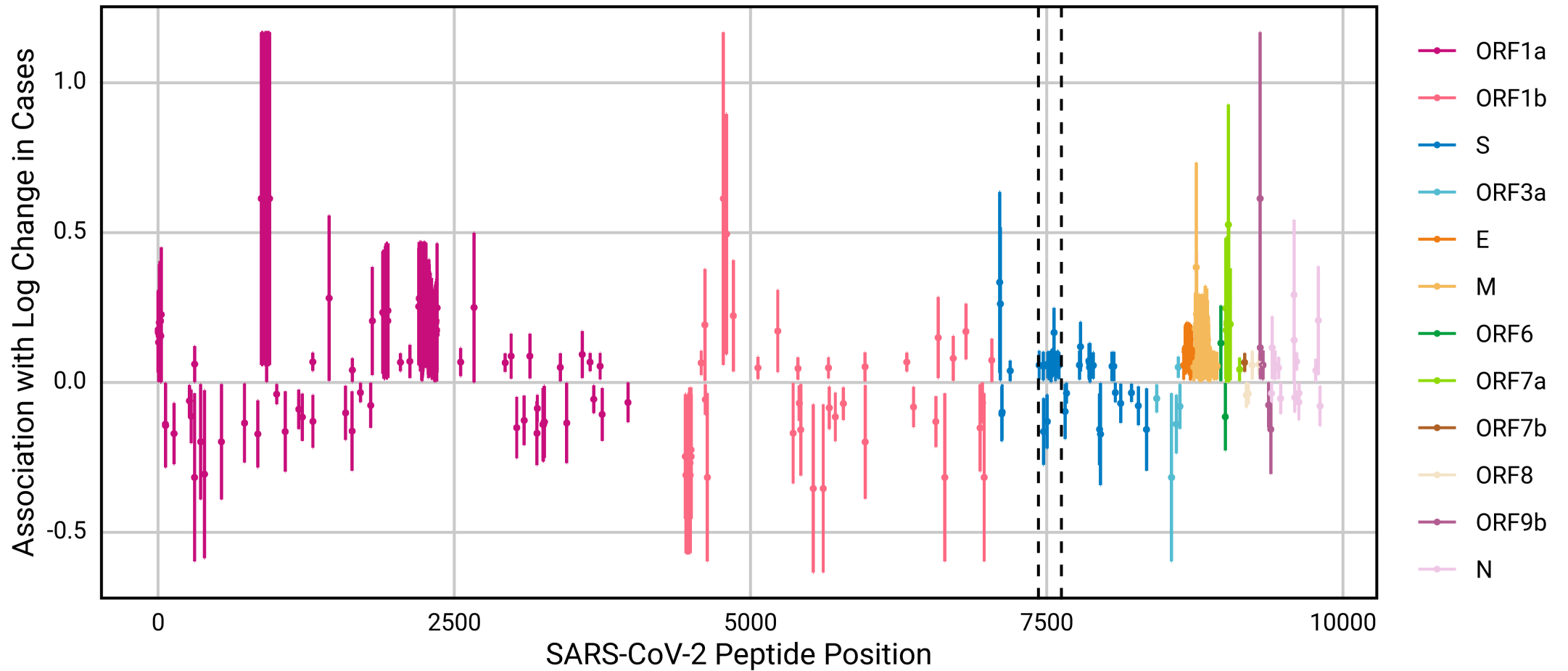
# Philadelphia Region COVID-19 Genomic Surveillance (CDC BAA 200-2021-10986)

## Relationship Between Peptide Diversity & Log Change in Cases



# Philadelphia Region COVID-19 Genomic Surveillance (CDC BAA 200-2021-10986)

## Relationship Between Peptide Diversity & Log Change in Cases



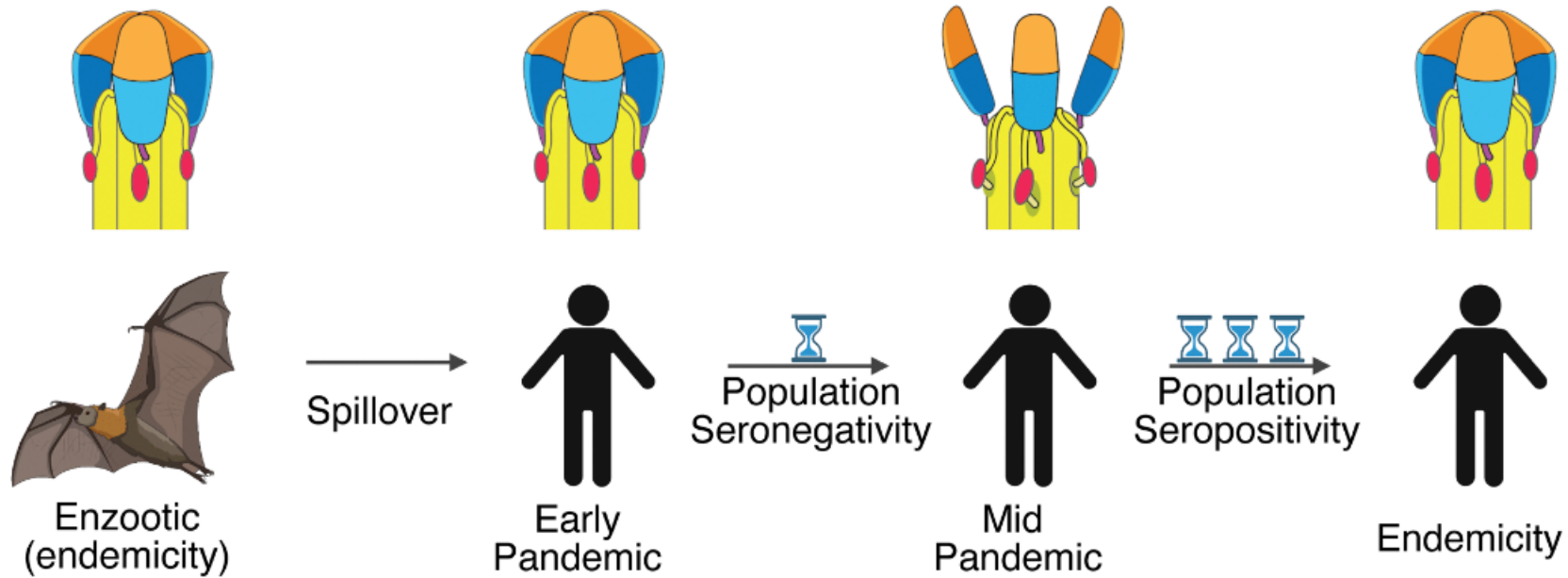
# The Canyon Hypothesis

- SARS-CoV-2 molecular evolution & the "Canyon Hypothesis":
  - highly conserved cell entry mechanism mediated by the spike protein (S gene)
  - spike engages angiotensin converting enzyme 2 (ACE2) & host proteases enable efficient membrane fusion between virions and target cells
  - a model by which sarbecoviruses are activated for fusion competency and interplay between humoral immunity and the molecular evolution

# The Canyon Hypothesis

- during circulation in populations with high rates of humoral immunity, viral entry proteins favor predominantly closed RBD configurations
- immediately after spillover into a population that lacks immunity, the newly emergent virus remains closely related to its ancestor
- during sustained transmission between seronegative individuals, wide transmission bottlenecks facilitate rapid emergence of variants that favor open RBD configurations to spread rapidly
- as humoral immunity expands, it gradually leads to a return to closed RBDs as repeat exposures facilitate the affinity maturation of expansive antibody repertoires that are disproportionately costly to open RBD configurations

# The Canyon Hypothesis



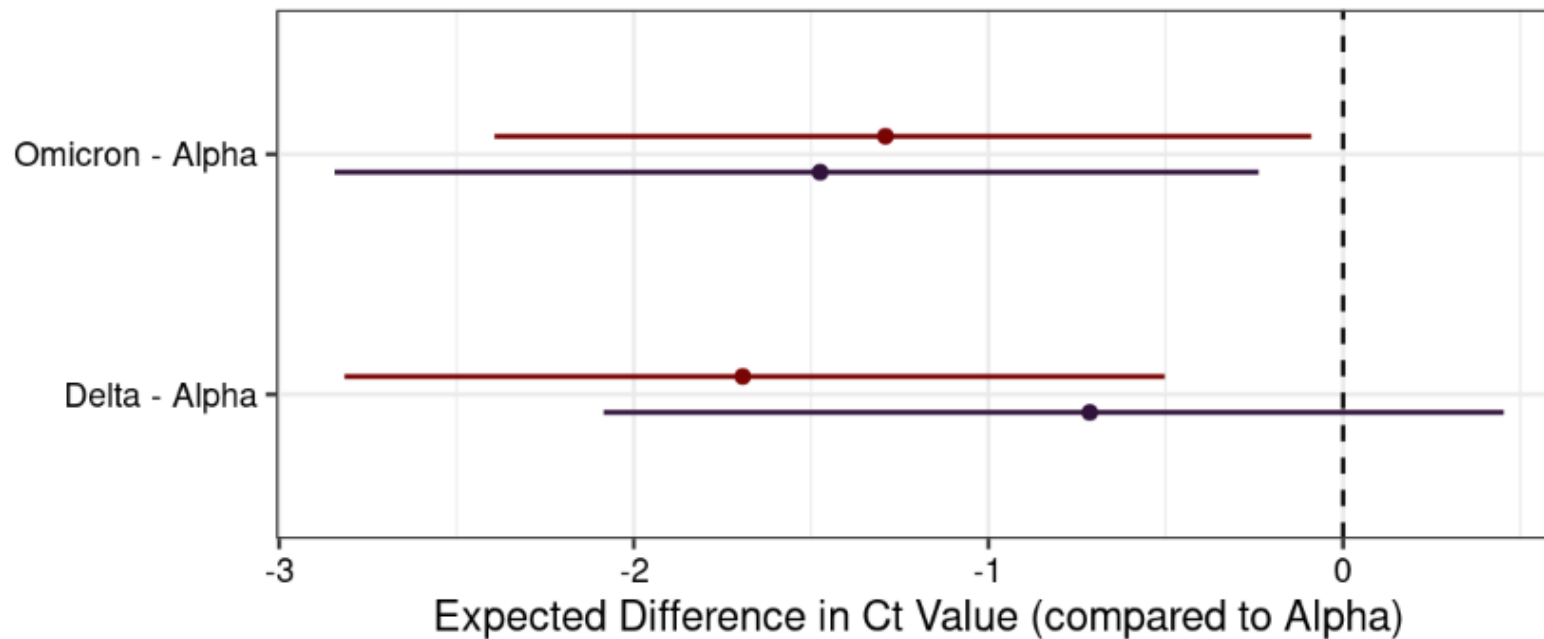
Wolf et al *mBio* 2022; Rossman et al *J Biol Chem* 1989

# The Canyon Hypothesis

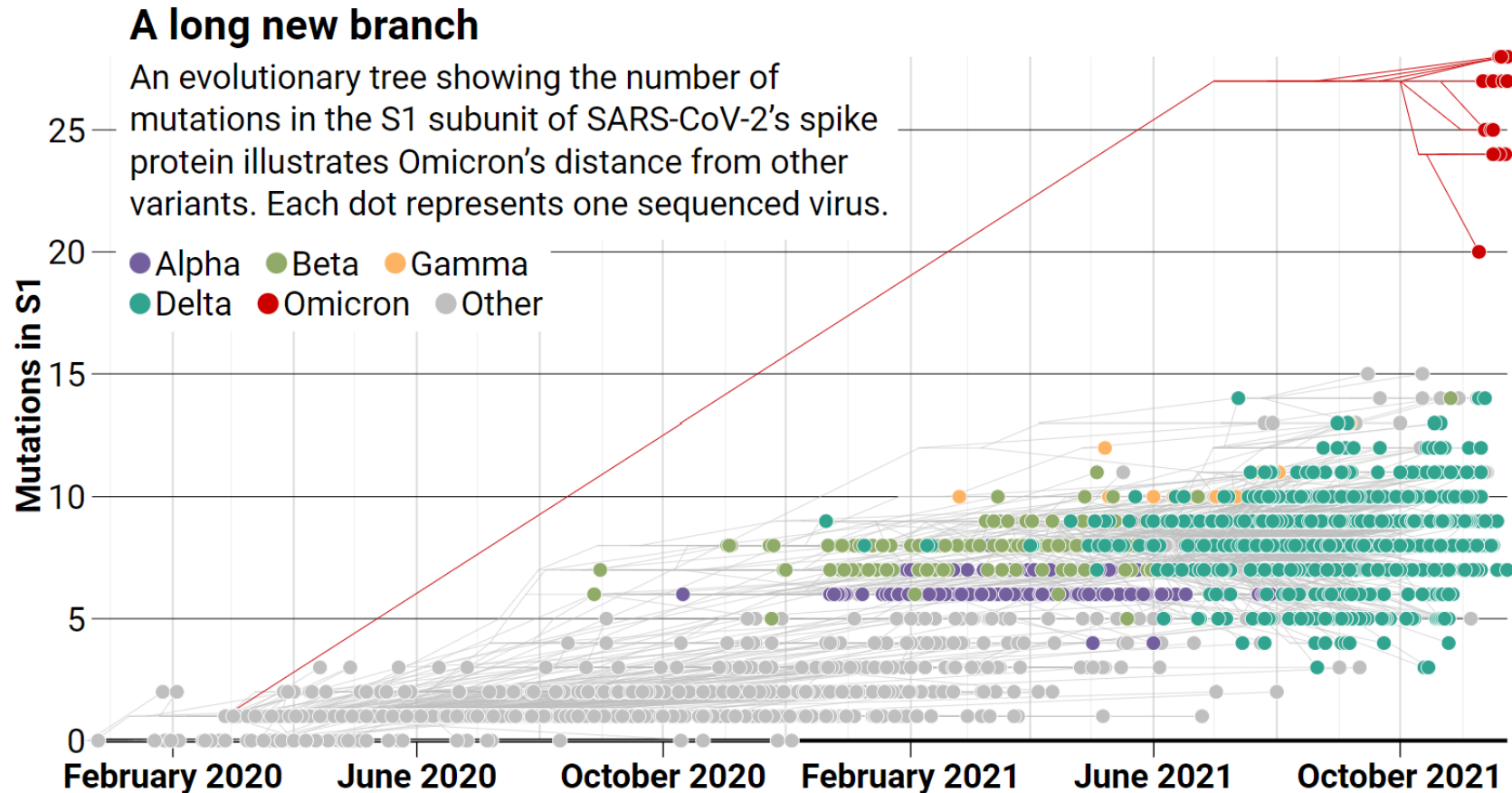
- most changes in S prior to the emergence of the Omicron variant appear to have been driven by selection for improved transmission between immunologically naive hosts
- S:D614G spike mutation improved ACE2 affinity but also made RBDs more likely to hold the open conformation
- cryo-EM studies of the Omicron S protein suggest that while the Delta Spike predominantly occupies conformations with 1 or more RBDs open simultaneously, the Omicron Spike appears to prefer conformations with 0 or 1 open RBD
- antigenicity of stabilizing elements & tendency of primary immune responses to generate a limited repertoire of antibodies may explain the selection for open RBDs early in the pandemic & shift in the selective landscape that led to the Omicron variant's emergence and rapid sweep



Model	Contrast	Difference in Ct Value (median & 95%CrI)
Adjusted	Delta - Alpha	-1.69 (-2.82 to -0.5)
Adjusted	Omicron - Alpha	-1.29 (-2.39 to -0.09)
Unadjusted	Delta - Alpha	-0.71 (-2.09 to 0.45)
Unadjusted	Omicron - Alpha	-1.47 (-2.84 to -0.24)



# Unknowns: Where did 'weird' Omicron come from?



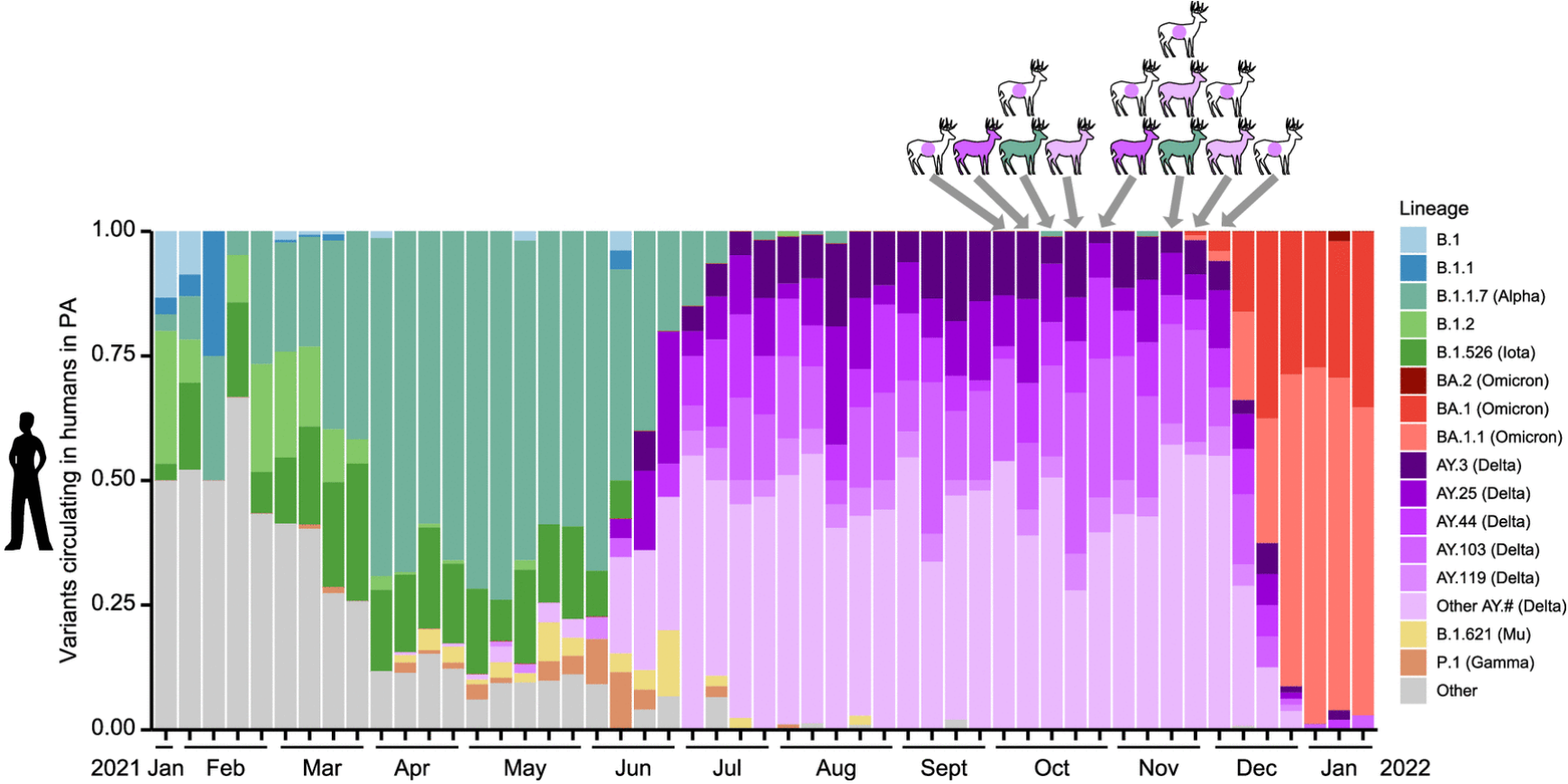
NEXTSTRAIN.ORG, ADAPTED BY N. DESAI/SCIENCE

# Unknowns: Where did 'weird' Omicron come from?

Omicron clearly did not develop out of one of the earlier variants of concern, such as Alpha or Delta. Instead, it appears to have evolved in parallel—and in the dark. Omicron is so different from the millions of SARS-CoV-2 genomes that have been shared publicly that pinpointing its closest relative is difficult, says Emma Hodcroft, a virologist at the University of Bern. It likely diverged early from other strains, she says. “I would say it goes back to mid-2020.”

That raises the question of where Omicron’s predecessors lurked for more than a year. Scientists see essentially three possible explanations: The virus could have circulated and evolved in a population with little surveillance and sequencing. It could have gestated in a chronically infected COVID-19 patient. Or it might have evolved in a nonhuman species, from which it recently spilled back into people.

# Pennsylvania Deer



COVID-19  
Genomic  
Epidemiology

Biases  
Limitations  
& Unknowns

**COVID-19  
Aftermath**

# Lessons & Directions After COVID-19

- What has come of this unprecedented investment in molecular/genomic epidemiology?
  - unprecedented pathogen sequencing capacity
  - novel methods of population surveillance: wastewater surveillance
  - political animus against CDC & resulting budget cuts

# Ongoing SARS-CoV-2 Variant Tracking



Centers for Disease Control and Prevention  
CDC 24/7: Saving Lives, Protecting People™

## COVID Data Tracker

Maps, charts, and data provided by CDC, updates Mondays and Fridays by 8 p.m. ET

United States  
At a Glance

Trend in % Test Positivity **-0.5%** in most recent week



Trend in % Emergency Department Visits **-8.1%** in most recent week

[Data Tracker Home](#)  
<https://covid.cdc.gov/covid-data-tracker/#variant-proportions>

[< Back to Variants & Genomic Surveillance](#)

### Variant Proportions

Trends

# Monitoring Variant Proportions

# Wastewater SARS-CoV-2 Surveillance



Centers for Disease Control and Prevention  
CDC 24/7: Saving Lives, Protecting People™

## COVID Data Tracker

Maps, charts, and data provided by CDC, updates Mondays and Fridays by 8 p.m. ET



Note: As of September 15, 2023, testing data is temporarily unavailable from about 400 wastewater testing sites nationwide. A new contract

United States  
At a Glance

Trend in % Test Positivity **-0.5%** in most recent week

Trend in % Emergency Department Visits **-8.1%** in most recent week

<https://covid.cdc.gov/covid-data-tracker/#wastewater-surveillance>



# Centers for Pathogen Genomics



Centers for Disease Control and Prevention  
CDC 24/7: Saving Lives, Protecting People™

## CDC Newsroom

[CDC Newsroom Home](#)

# CDC announces \$90M funding to support Pathogen Genomics

Centers will foster and improve innovation to better prevent, control and respond to microbial threats

## Press Release

For Immediate Release: Tuesday, September 20, 2022

Contact: [Media Relations](#)

(404) 639-3286

<https://www.cdc.gov/media/releases/2022/p0920-PGCoE-network.html>

Today, CDC announced 5-year awards to five state public health departments. The awards will establish the Pathogen Genomics Centers of Excellence, increasing capacity in pathogen genomics, molecular epidemiology, and bioinformatics to better prevent, control, and respond to microbial threats.

# NWSS: National Wastewater Surveillance System



Centers for Disease Control and Prevention  
CDC 24/7: Saving Lives, Protecting People™

National Wastewater Surveillance System (NWSS)

National Wastewater Surveillance System (NWSS)



<https://www.cdc.gov/nwss/index.html>

# CDC Budget Cuts

- CDC and NIH have become political targets:
  - \$1.3 billion (~18%) cut from CDC as part of debt ceiling agreement
  - some budget proposals for FY2024 include a total \$11.5 billion (\$2.8 billion increase from FY2023)
  - other budget proposals for FY2024 include additional 20-30% cuts

# Lessons & Directions After COVID-19

- What has come of this unprecedented investment in molecular/genomic epidemiology?
  - unprecedented pathogen sequencing capacity
  - novel methods of population surveillance: wastewater surveillance
  - political animus against CDC & resulting budget cuts
- **Will re-investment come before or after the next pandemic?**

COVID-19  
Genomic  
Epidemiology

Biases  
Limitations  
& Unknowns

COVID-19  
Aftermath

A 3D visualization of a protein structure. The protein is shown in two colors: red and grey. The red parts are more prominent and appear to be the surface of the protein, while the grey parts are more internal and form a mesh-like structure. There are several small orange and yellow spheres scattered throughout the structure, possibly representing specific atoms or ligands. The background is dark grey.

Questions?