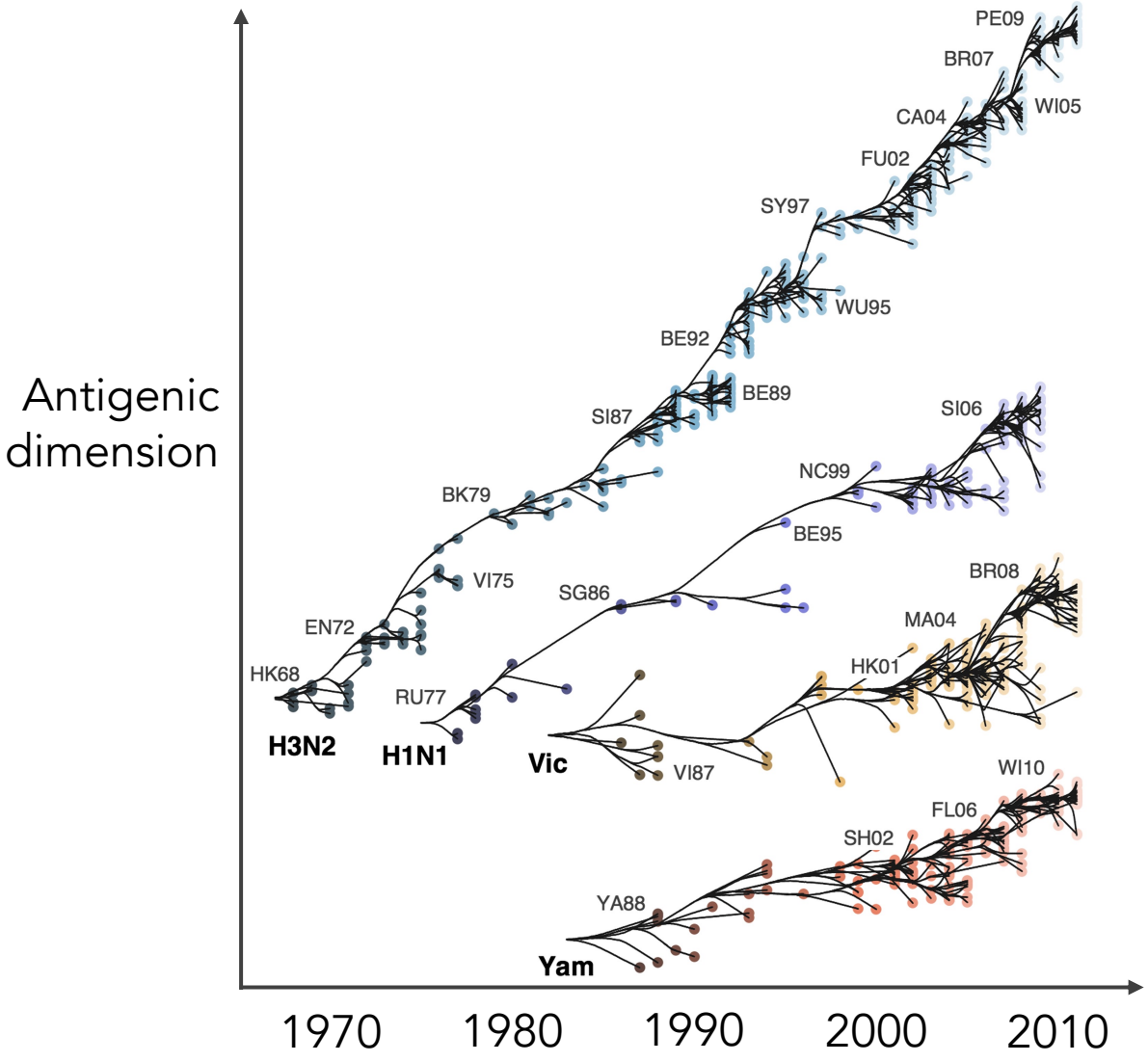
Learning about evolution from time-series sequence data, from the lab to the globe

Biological Evolution Across Scales Workshop
Bernoulli Center, EPFL

John Barton
Dept of Computational and Systems Biology @ Pitt
@_jpbarton
www bartonlab.github.io

2023-04-18

Pathogens can evolve to escape immunity and increase replication or transmissibility



Influenza undergoes antigenic drift, escaping past immune responses

SARS-CoV-2 variants drive new waves of infection

Model evolution quantitatively to identify critical mutations and predict future dynamics

Figure: Bedford et al, eLife 2014

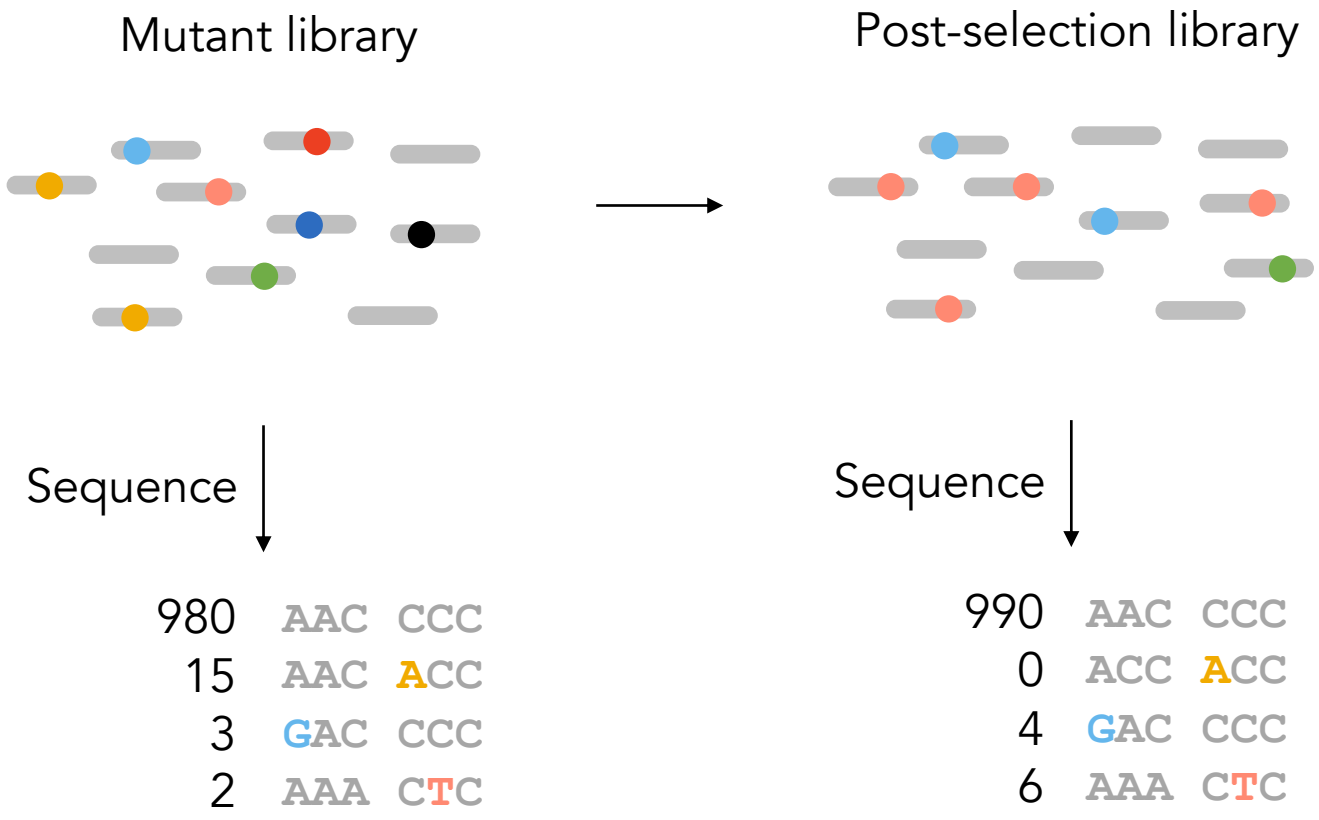
Inferring the functional effects of mutations from deep mutational scanning experiments

Evolution of SARS-CoV-2 for increased transmissibility

Inferring the functional effects of mutations from deep mutational scanning experiments

Evolution of SARS-CoV-2 for increased transmissibility

Deep mutational scanning (DMS) provides massively parallel measurements of the functional effects of mutations



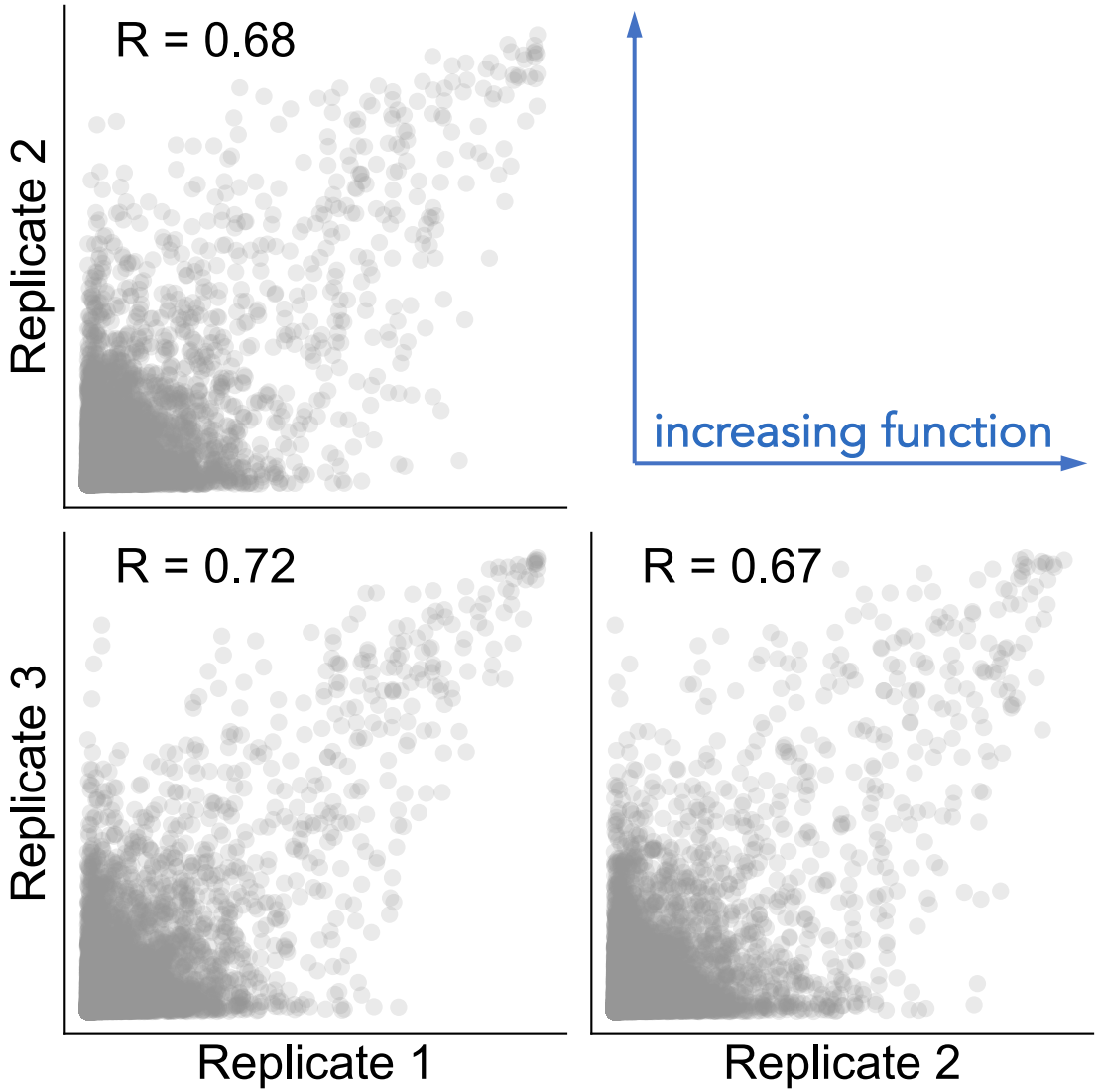
Experimentalists first synthesize a large, random library of mutants and sequence them

After passing through selection (sometimes multiple rounds), the library is sequenced again

Change in frequency gives insights into function

Examples: virus resistance to antibodies, binding affinity, …

However, reproducibility is surprisingly limited



Function is often quantified by enrichment ratios, x_{final} / $x_{initial}$, rescaled at each site to form preferences that fall between 0 and 1

In many DMS data sets, correlation between replicates is only ~0.7

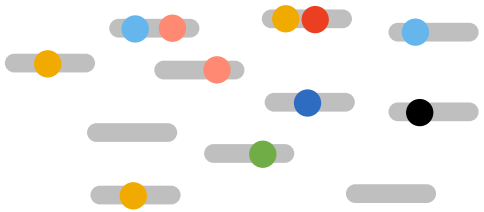
About half of variance in scores explained by other replicates, with some significant differences

We analyzed DMS data as if it came from an evolving population of individuals with different fitness values

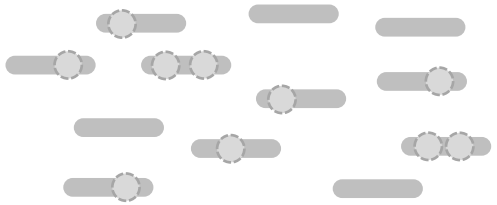
Data: Haddox et al, eLife 2018

Population evolution under the Wright-Fisher model

Population (t = 0)



Population (t = 1)



Sequences in the next generation are chosen w/ probability proportional to their **fitness f** and **frequency x**

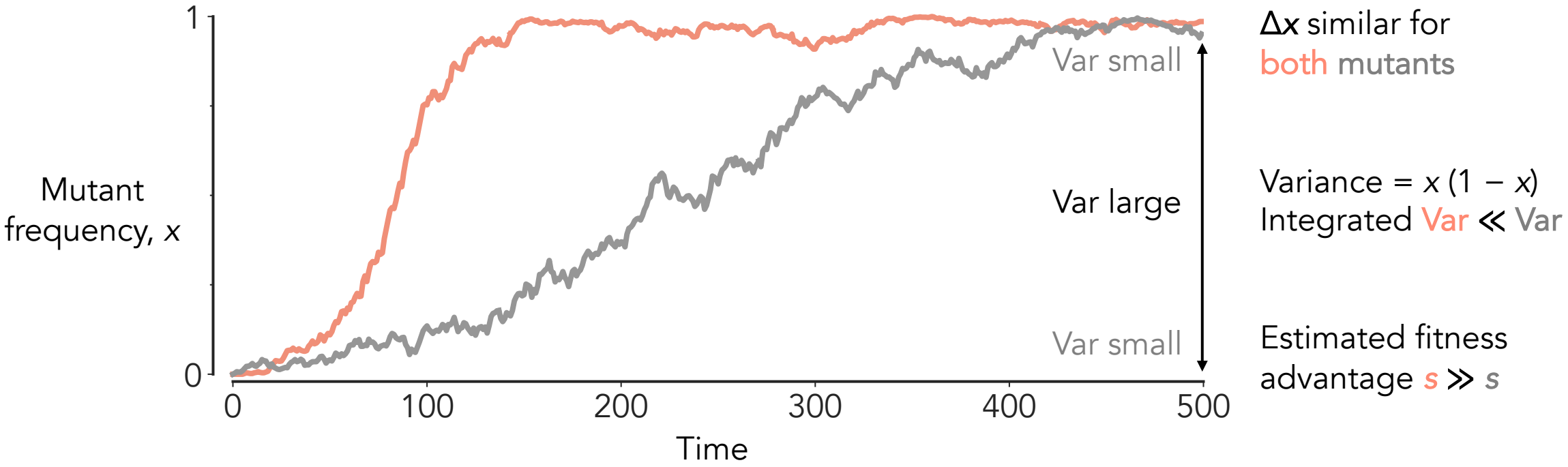
$f(\text{—}) = 1$

$f(\text{—}_{i}\text{—}_{j}) = 1 + s_i + s_j$

Probabilities are normalized by the average fitness of the population

To simplify dynamics, consider a limit where number of sequences $N \to \infty$ and fitness effects **s** ~ O(1/N)

An intuitive expression for estimating the fitness effects of mutations



To estimate fitness advantage (selection coefficient s), multiply change in frequency (Δx) with inverse of the frequency covariance matrix (C) integrated over time (plus reg.)

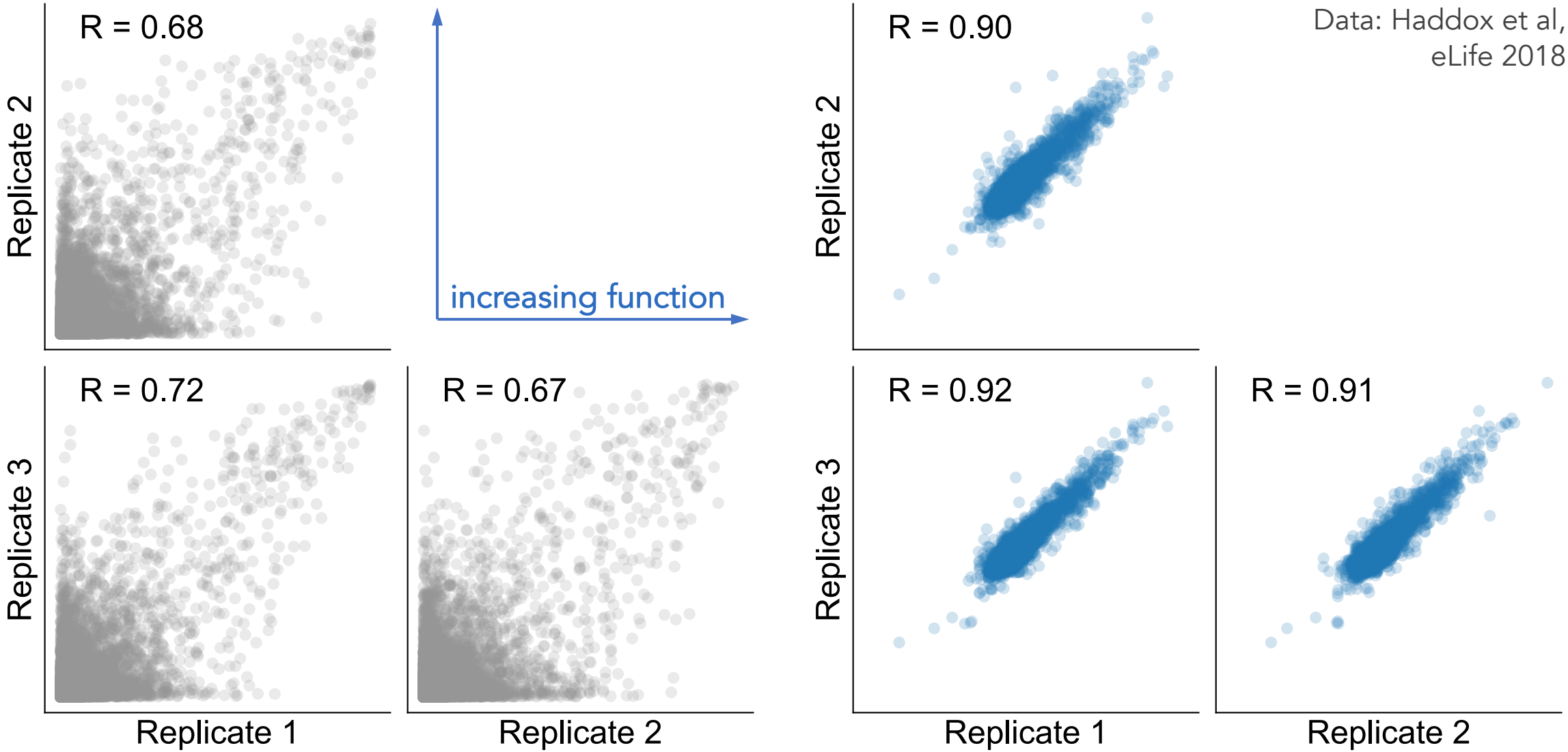
$$\hat{s} = \left(C_{\text{int}} + \gamma\right)^{-1} \Delta x$$

Off-diagonal covariance terms account for genetic linkage

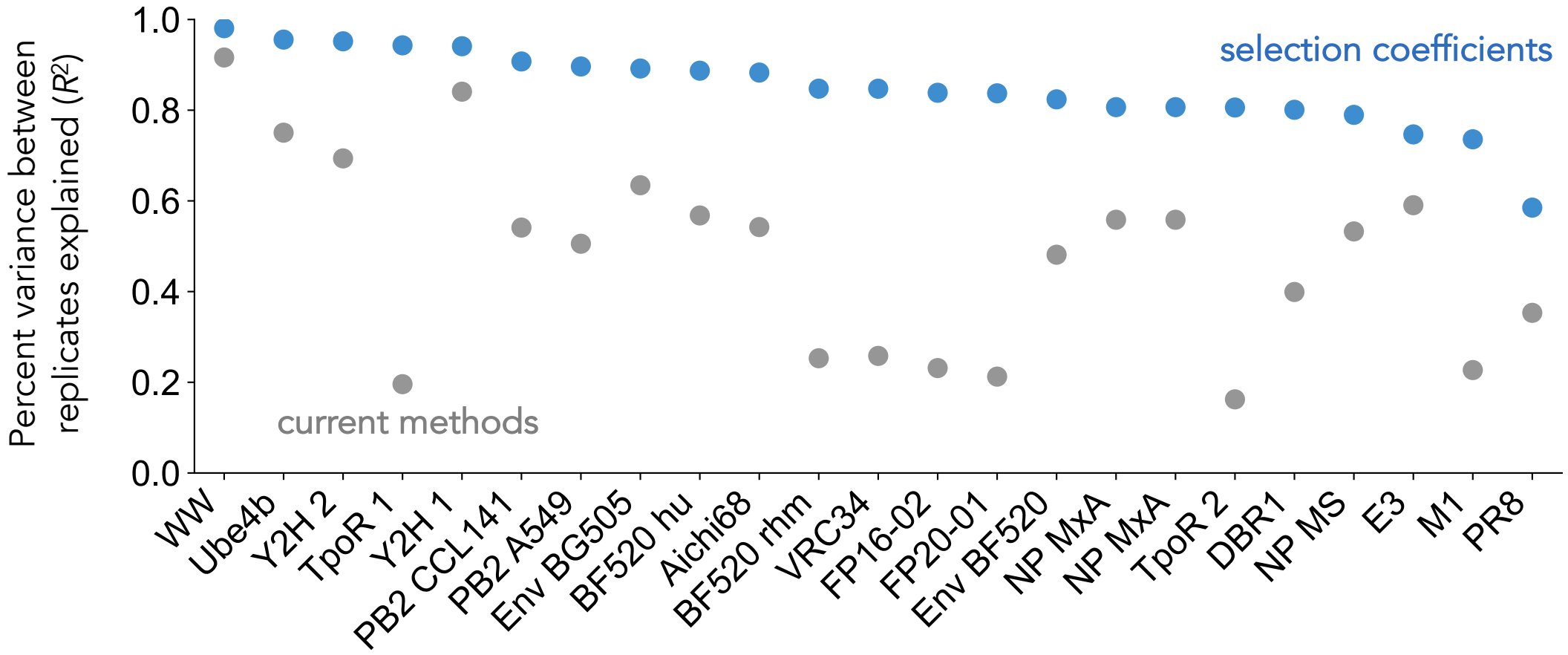
Additional adjustments for DMS data (error rates, etc.)

Sohail, Louie, McKay, & JPB
Nature Biotechnology 2021

Selection coefficients that we infer are much more reliable than past estimates



Our estimated functional effects were more consistent than existing approaches across all 23 data sets that we tested



Intuitive advantages compared to current methods

For beneficial mutations, frequency gain is slow when mutations are at very low or very high frequencies, not accounted for by enrichment ratios/regression

High frequency estimates are particularly important for WT residues, often used to normalize other estimates of functional effects

Effect magnitudes meaningful across sites

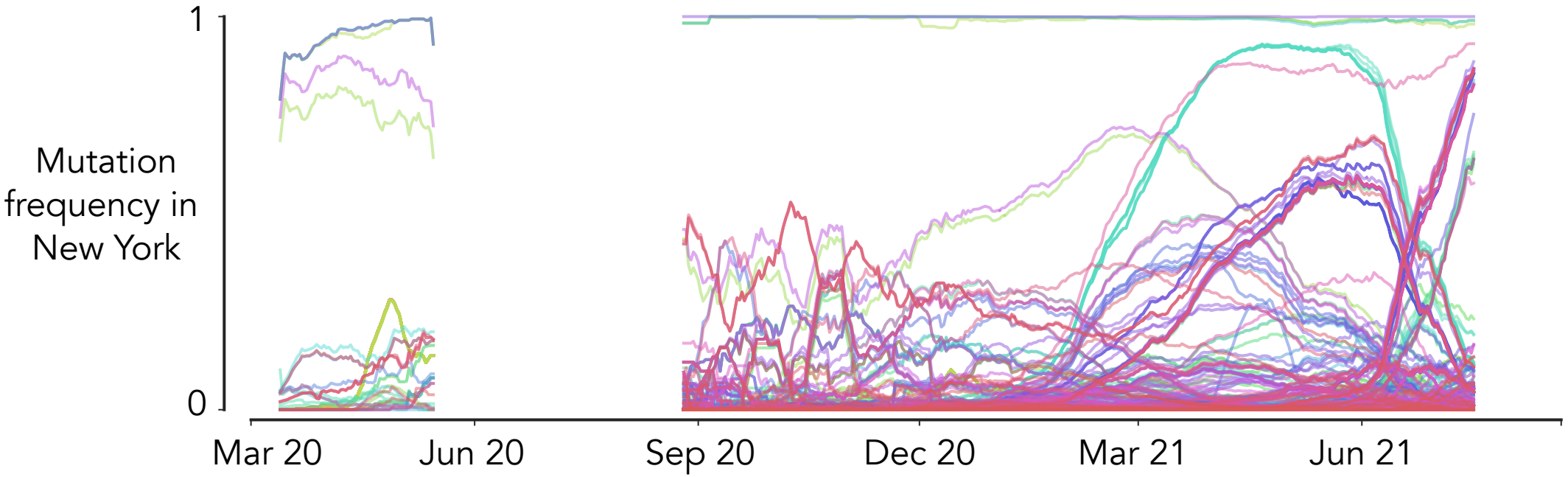
Regularization controls spurious inferences when data is weak

Preprint coming soon!

Inferring the functional effects of mutations from deep mutational scanning experiments

Evolution of SARS-CoV-2 for increased transmissibility

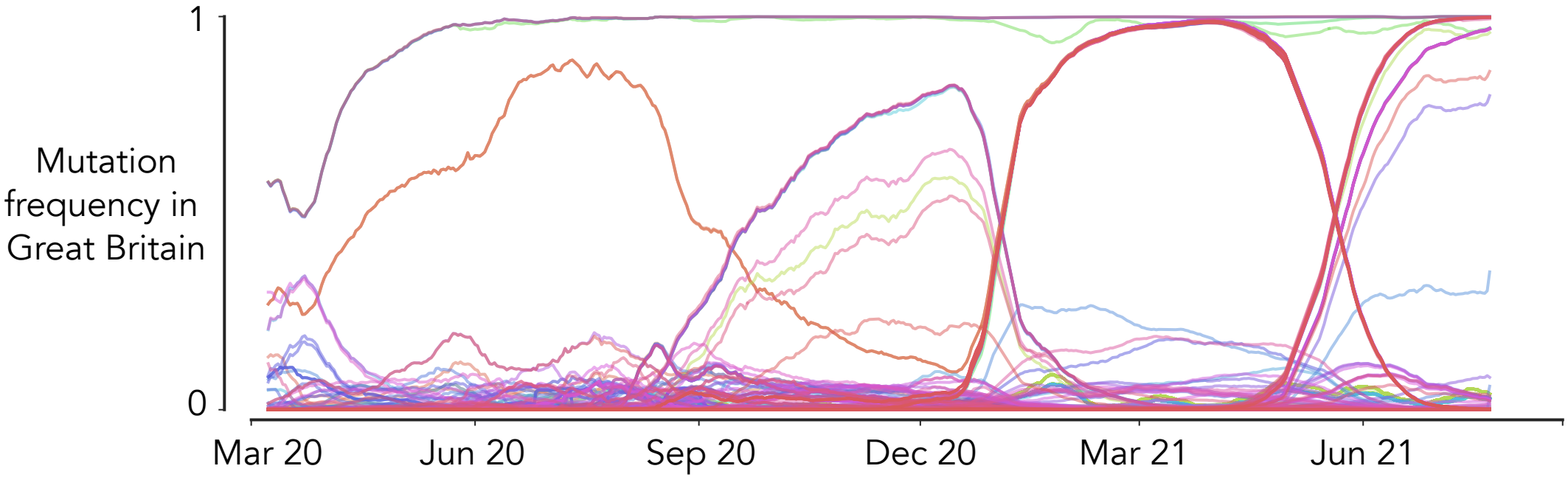
Evolutionary history of SARS-CoV-2 is complex



Not just
ACTG
GCTG
but
ACTG
GCCG
ATTA
GCCA
...

Data: GISAID

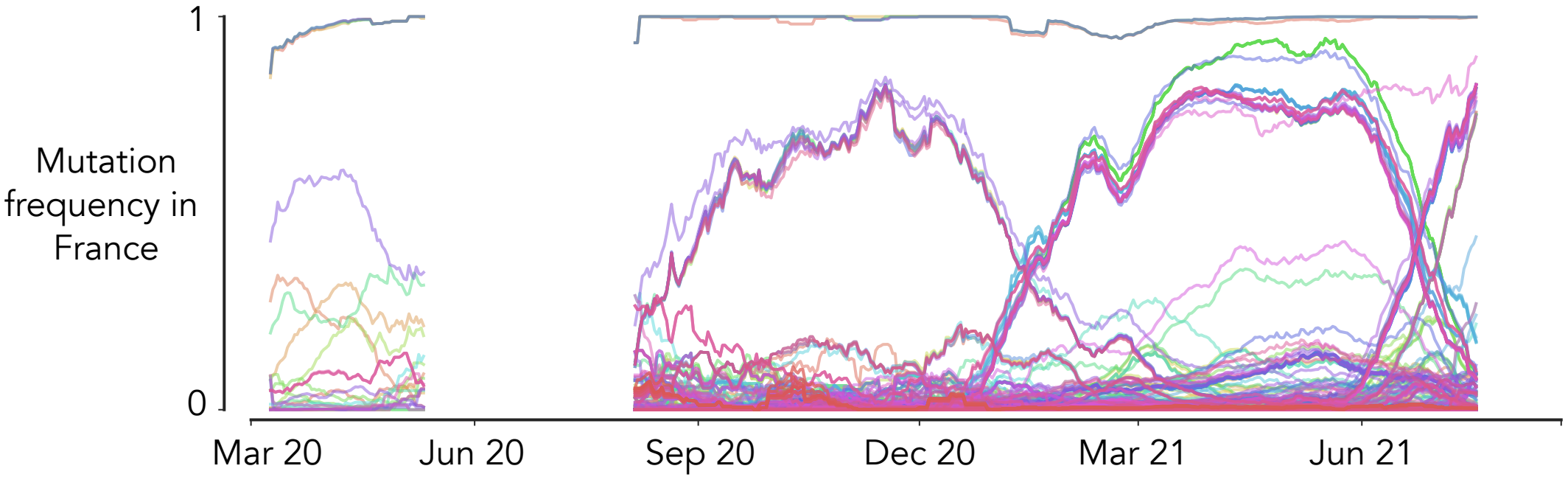
Evolutionary history of SARS-CoV-2 is complex



Not just ACTG / GCTG but ACTG / GCCG / ATTA / GCCA / ...

Data: GISAID

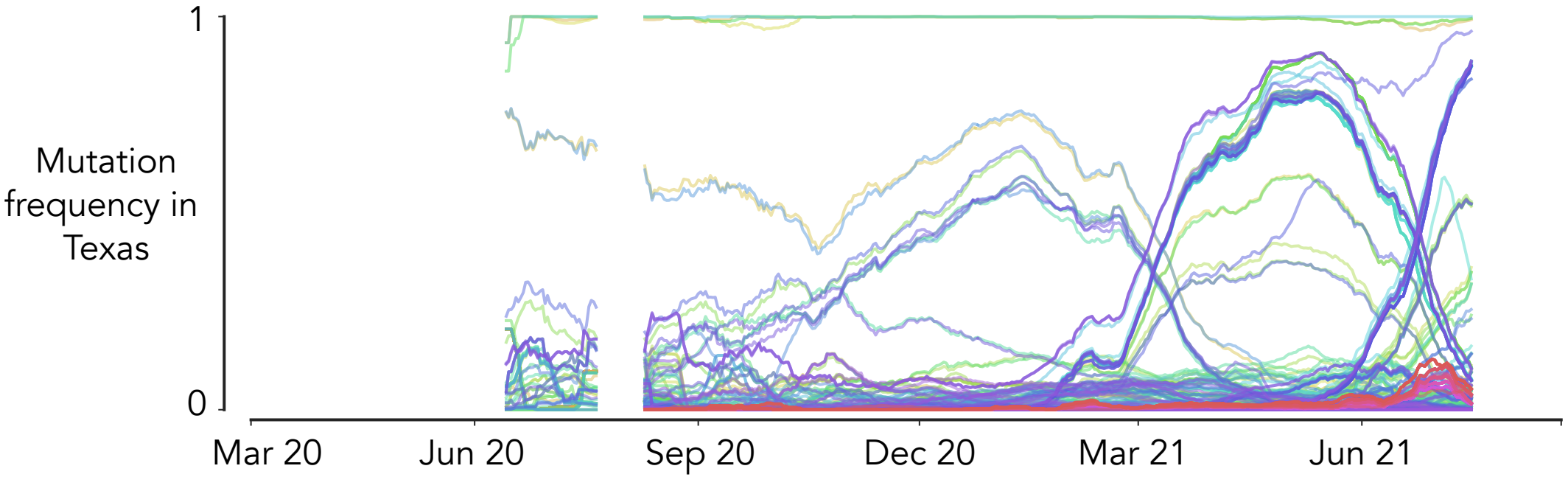
Evolutionary history of SARS-CoV-2 is complex



Not just
ACTG
GCTG
but
ACTG
GCCG
ATTA
GCCA
...

Data: GISAID

Evolutionary history of SARS-CoV-2 is complex

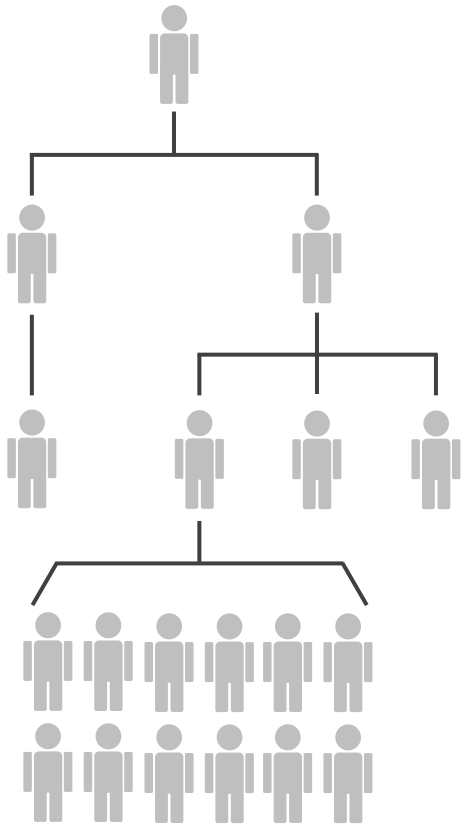


Not just
ACTG
GCTG
but
ACTG
GCCG
ATTA
GCCA
...

Estimate transmission effects of mutations that best explain observed evolutionary histories, including linkage (correlation)

Data: GISAID

Modeling epidemiological/evolutionary dynamics as a branching process



Infected individuals transmit a random number of new infections

The distribution is heavy-tailed (negative binomial) to model superspreading

Expected number of new infections R_a depends on the variant a

$$R_a = R(1 + w_a)$$

with $g_i^a = 1$ if variant a has mutation i and 0 otherwise

selection coefficient for variant a

selection coefficient for mutation i

Ultimately, expressions for relative frequencies of variants are almost identical to the previous model

Lee et al, medRxiv 2022
Lloyd-Smith et al, Nature 2005

Epidemiological dynamics of SARS-CoV-2

We analyzed GISAID data through June 2022 (5.6 million sequences across 126 regions) to estimate effects of mutations on SARS-CoV-2 transmission

Evolutionary histories in different regions are considered independent, a good approximation when most transmission is local rather than due to travel

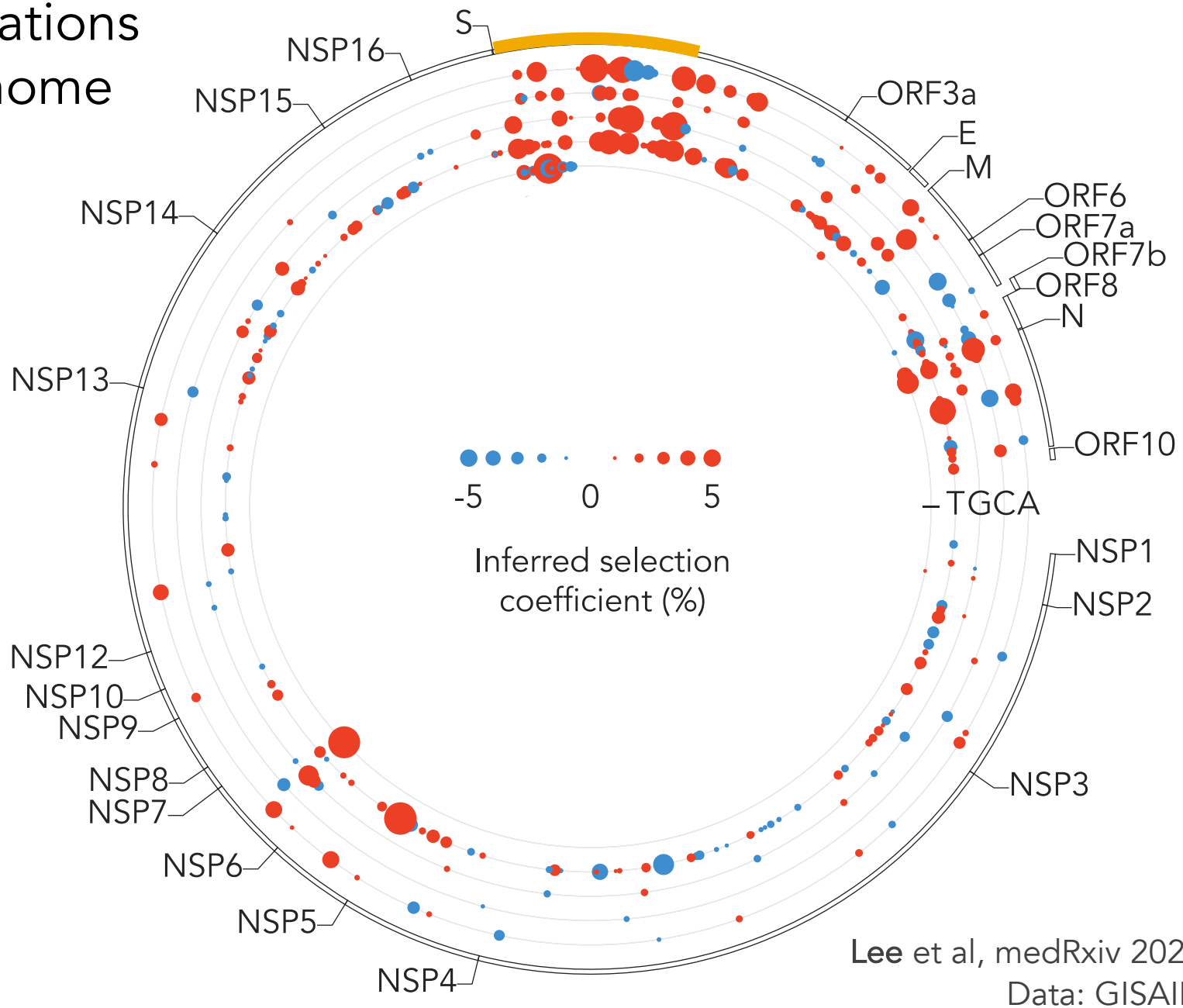
To obtain joint estimates of selection coefficients, sum contributions from all regions (computed exactly from posterior!)

Lee et al, medRxiv 2022

Transmission effects of mutations across the SARS-CoV-2 genome

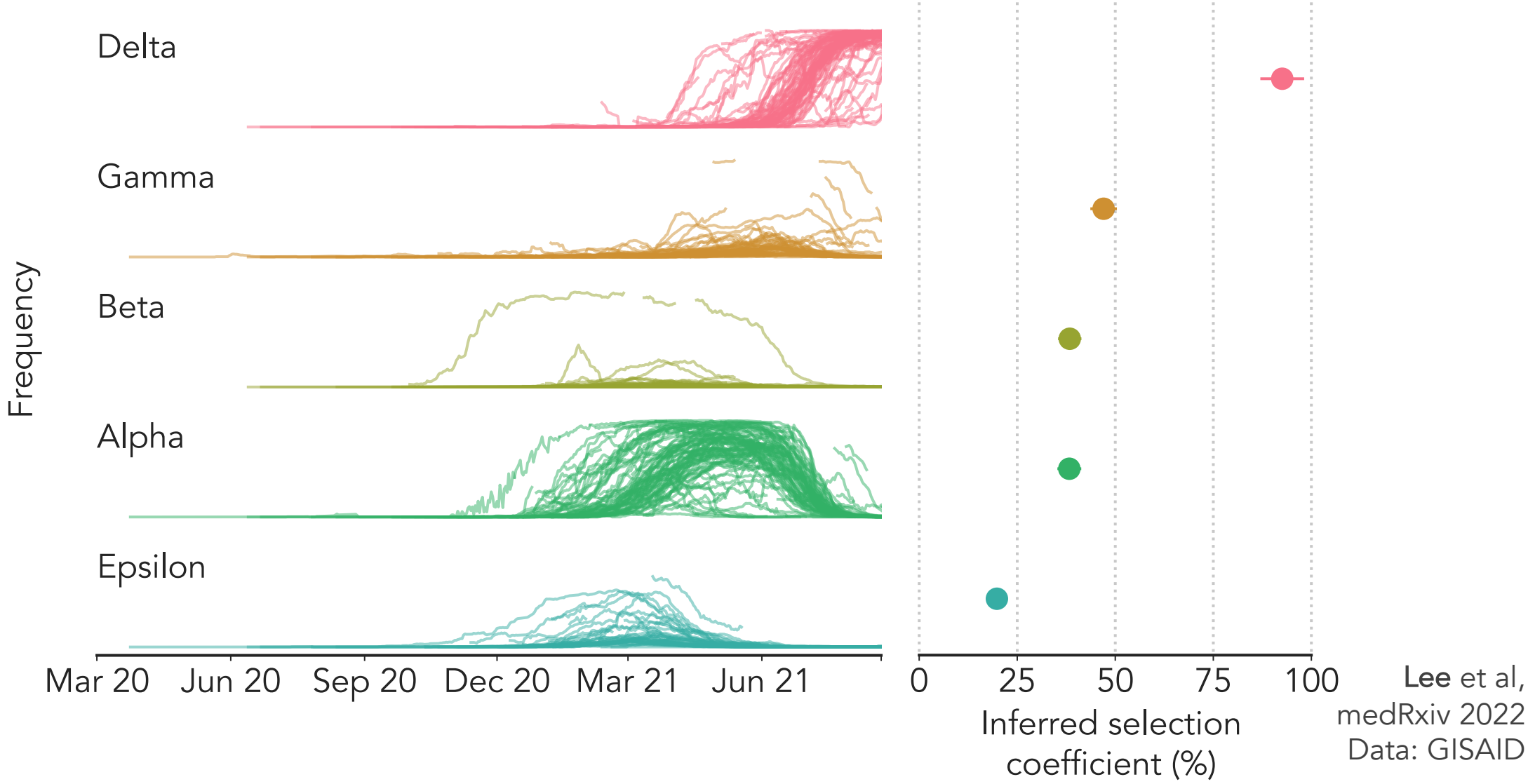
Strong selection in Spike S1 subunit, including mutations
Δ142 (NTD, Ab evasion),
L452Q/R and Q498R (RBD),
P681R/H (FCS)

Other clusters of beneficial mutations including in
NSP6 (Δ106-108),
N (R203M, S202N)

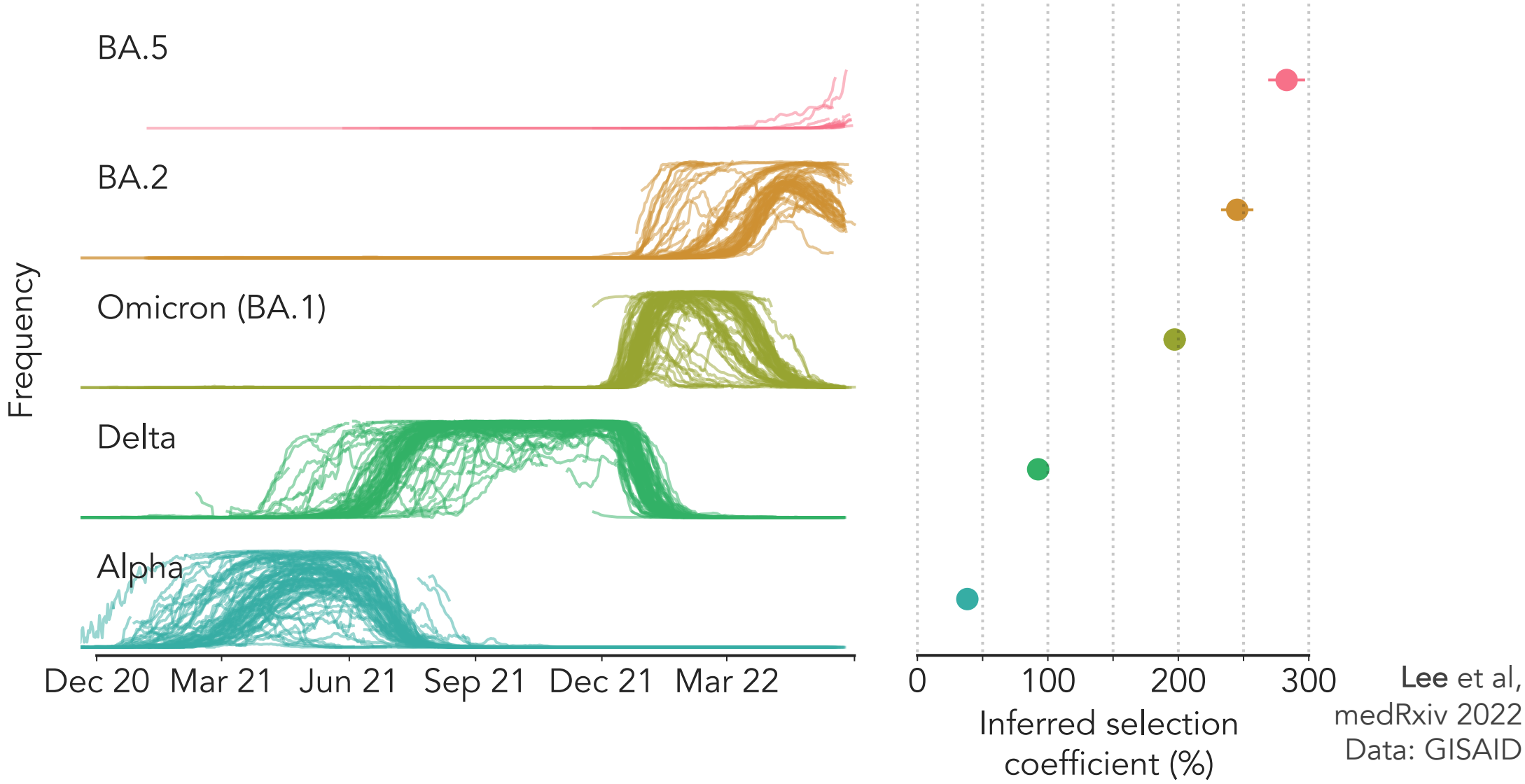


Lee et al, medRxiv 2022
Data: GISAID

Collectively, we obtain estimates for the net transmission advantage of well-known variants



Collectively, we obtain estimates for the net transmission advantage of well-known variants, showing clear dominance of Omicron



Lee et al, medRxiv 2022
Data: GISAID

Rapid detection of new variants with significant transmission advantage: separating fluctuations from signal

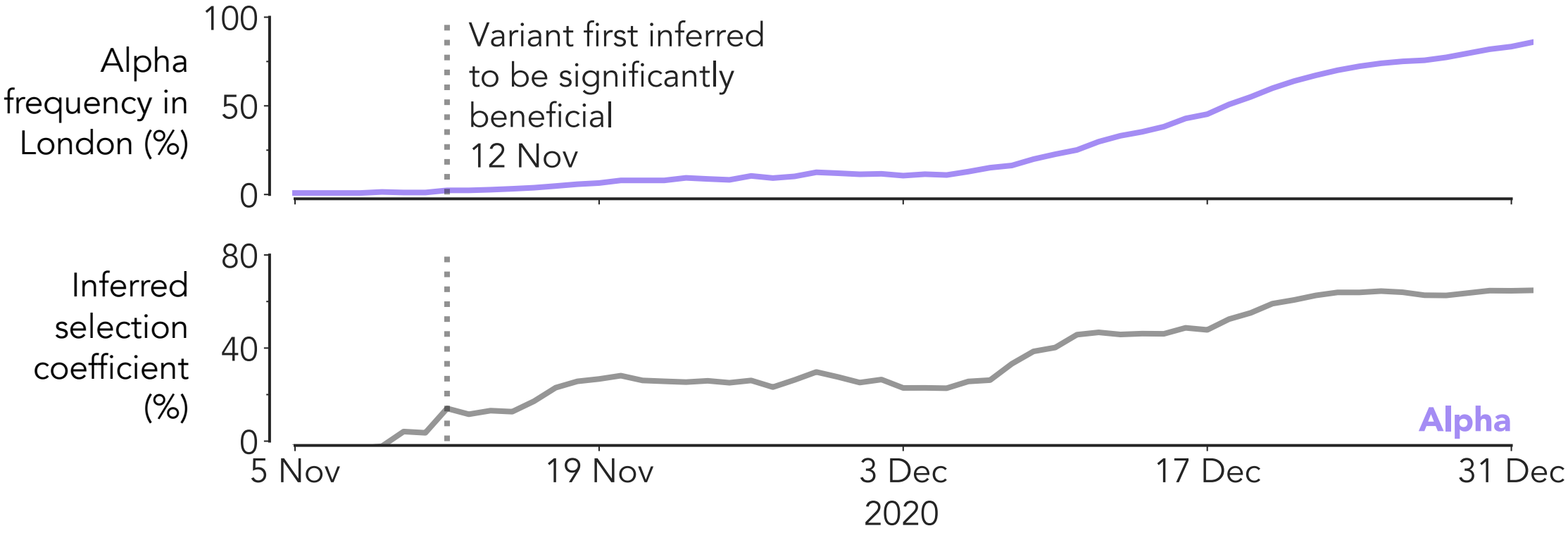
Measure selection for all linked groups of mutations globally and in all individual regions, at all intermediate times (until Feb 2021)

Linked groups with selection coefficients <10% in global data are never inferred >13% in any region, at any time

Variants >13% in any region at any time thus have reliably have higher transmission, could be considered "concerning"

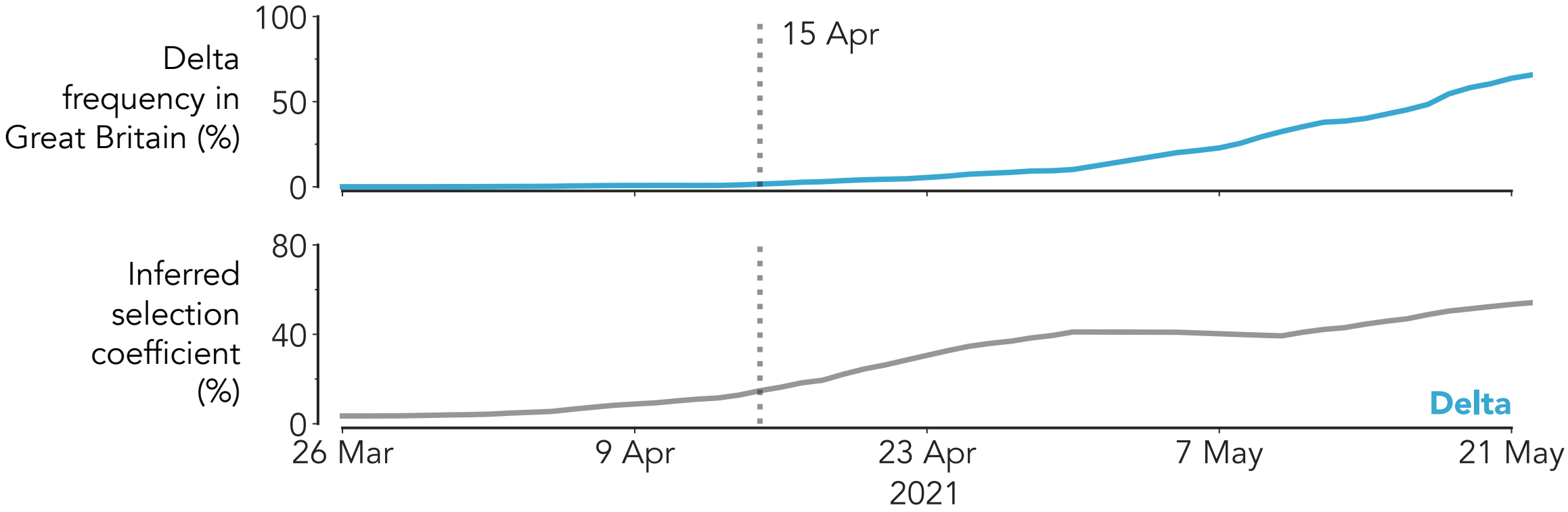
Lee et al, medRxiv 2022
Data: GISAID

Rapid detection of new variants with significant transmission advantage: separating fluctuations from signal



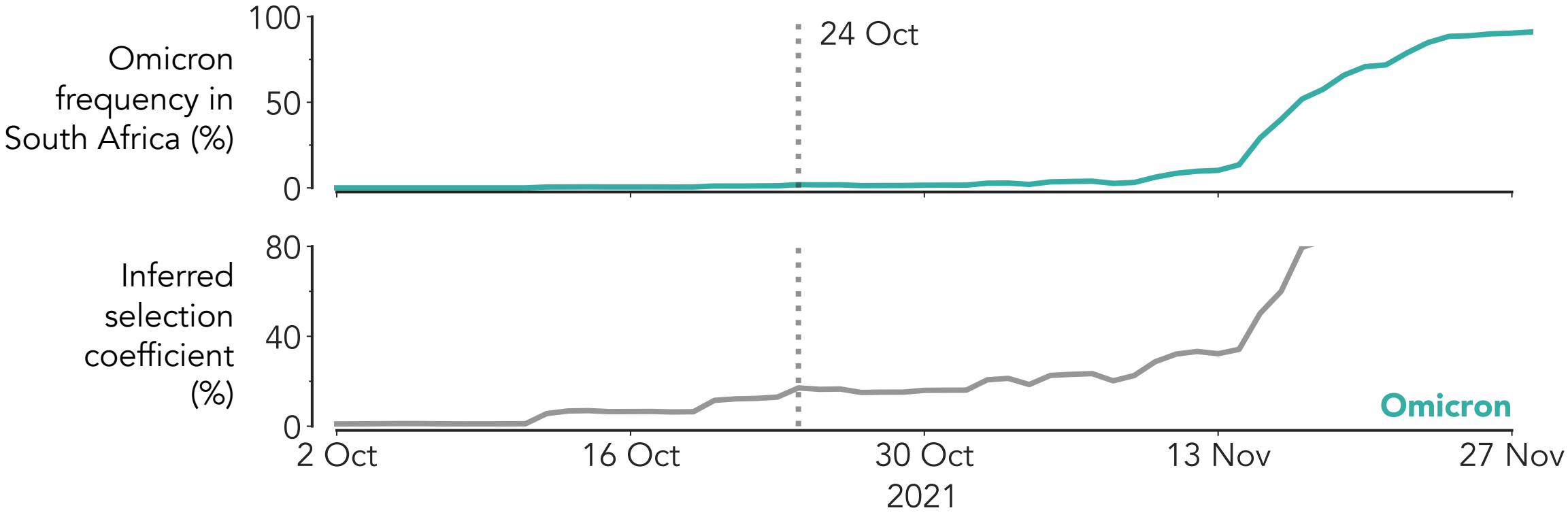
Lee et al, medRxiv 2022
Data: GISAID

Rapid detection of new variants with significant transmission advantage: separating fluctuations from signal



Lee et al, medRxiv 2022
Data: GISAID

Rapid detection of new variants with significant transmission advantage: separating fluctuations from signal



Lee et al, medRxiv 2022
Data: GISAID

Summary

- Developed an analytical method to infer the fitness effects from time-series sequence data
- Inferred fitness effects from deep mutational scanning data are much more consistent than with current approaches
- Application to SARS-CoV-2 finds clusters of beneficial mutations and net transmission benefit for variants with multiple mutations
- Model is sensitive enough to provide rapid detection of more transmissible variants
- Extensible to epistasis (Sohail et al, MBE 2022), etc.

Acknowledgments

Barton lab members*


Brian Lee*


Liz Finney*


Faraz Ahmed


Matt Mckay


Ray Louie


Saqib Sohail


Ahmed Quadeer

NIH MIRA R35GM138233