Estimating effects of mutations to SARS-CoV-2 proteins from natural sequences

Jesse Bloom & Richard Neher

Fred Hutch Cancer Center / HHMI

@jbloom_lab

Slides: https://slides.com/jbloom/sars2-mut-fitness

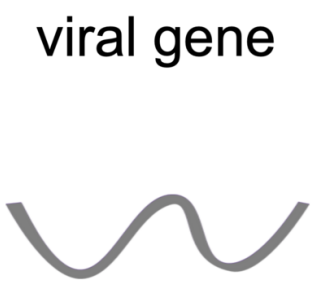
Richard Neher

Determining effects of viral mutations is important

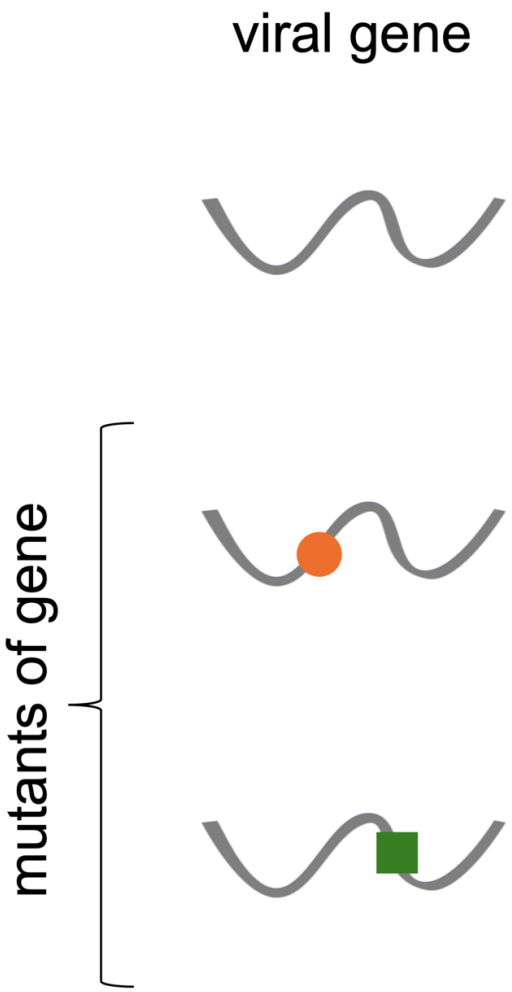
1. Interpret consequences of mutations seen during viral surveillance.
2. Inform design of drugs to target constrained regions.
3. Understand function and mechanisms of viral proteins.

Traditional way to determine effect of mutations is experiments

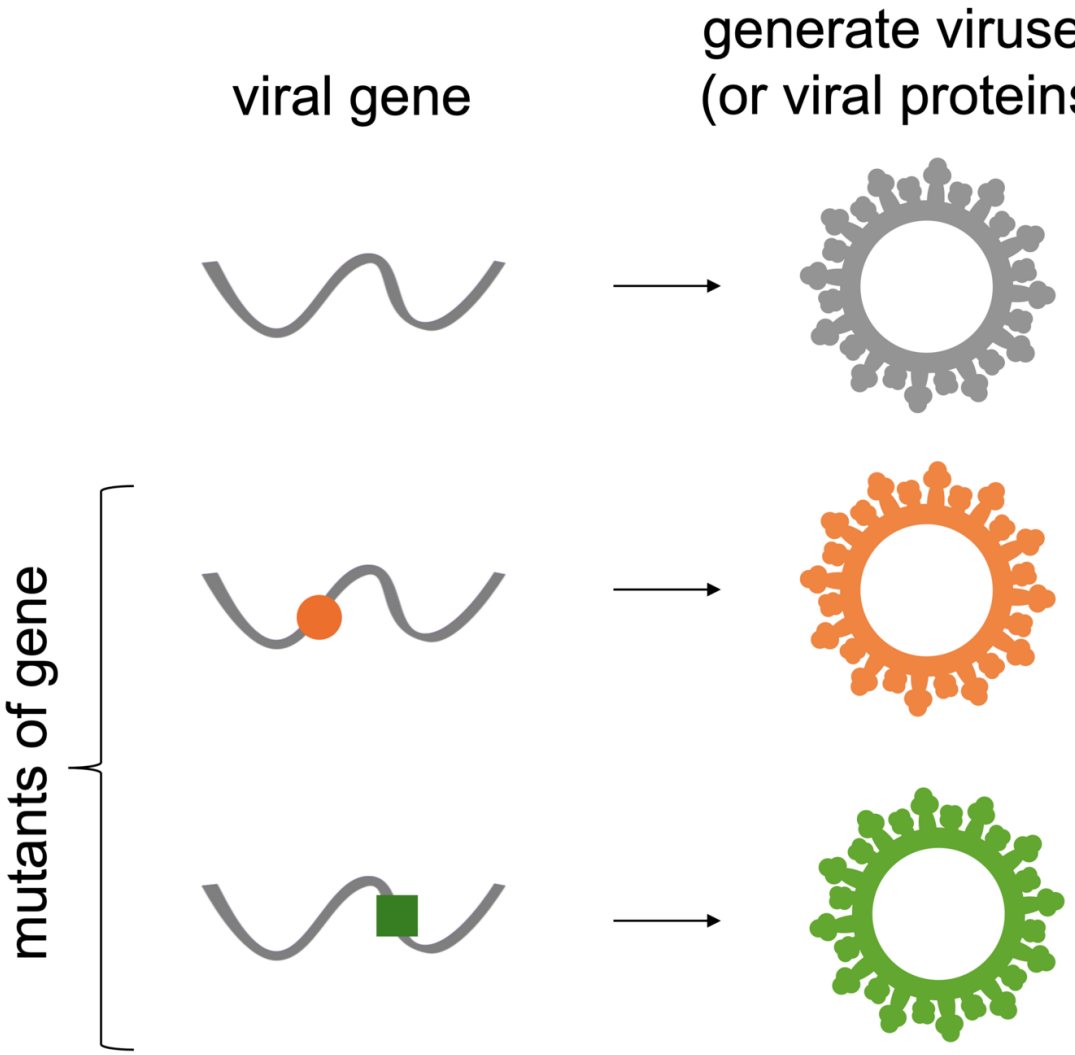
Traditional way to determine effect of mutations is experiments



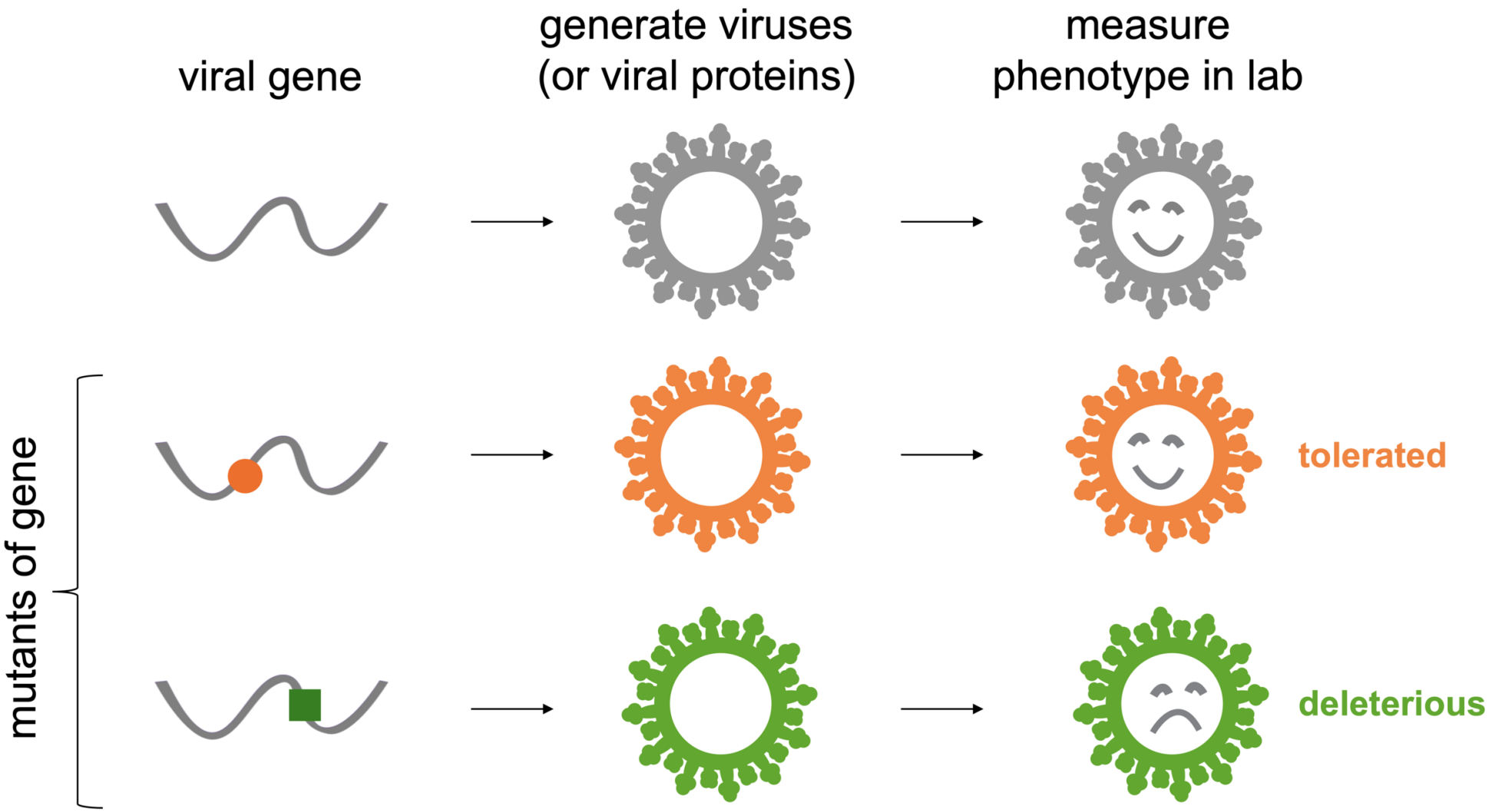
Traditional way to determine effect of mutations is experiments



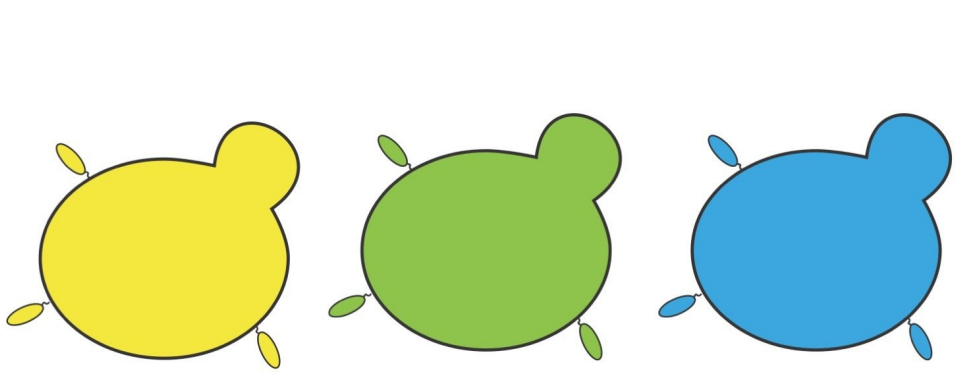
Traditional way to determine effect of mutations is experiments

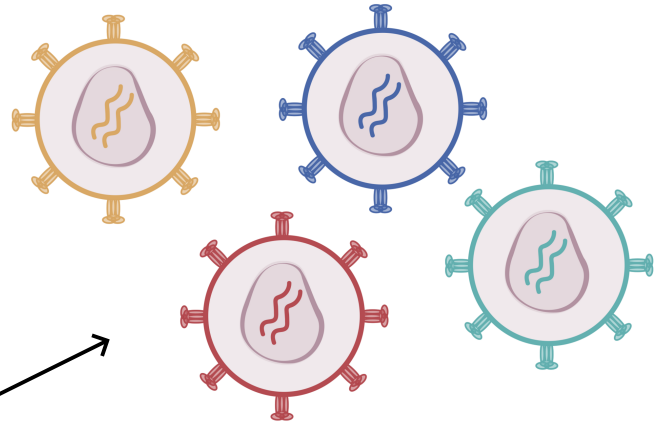


Traditional way to determine effect of mutations is experiments



My group tries to do such experiments at large scale via **deep mutational scanning**



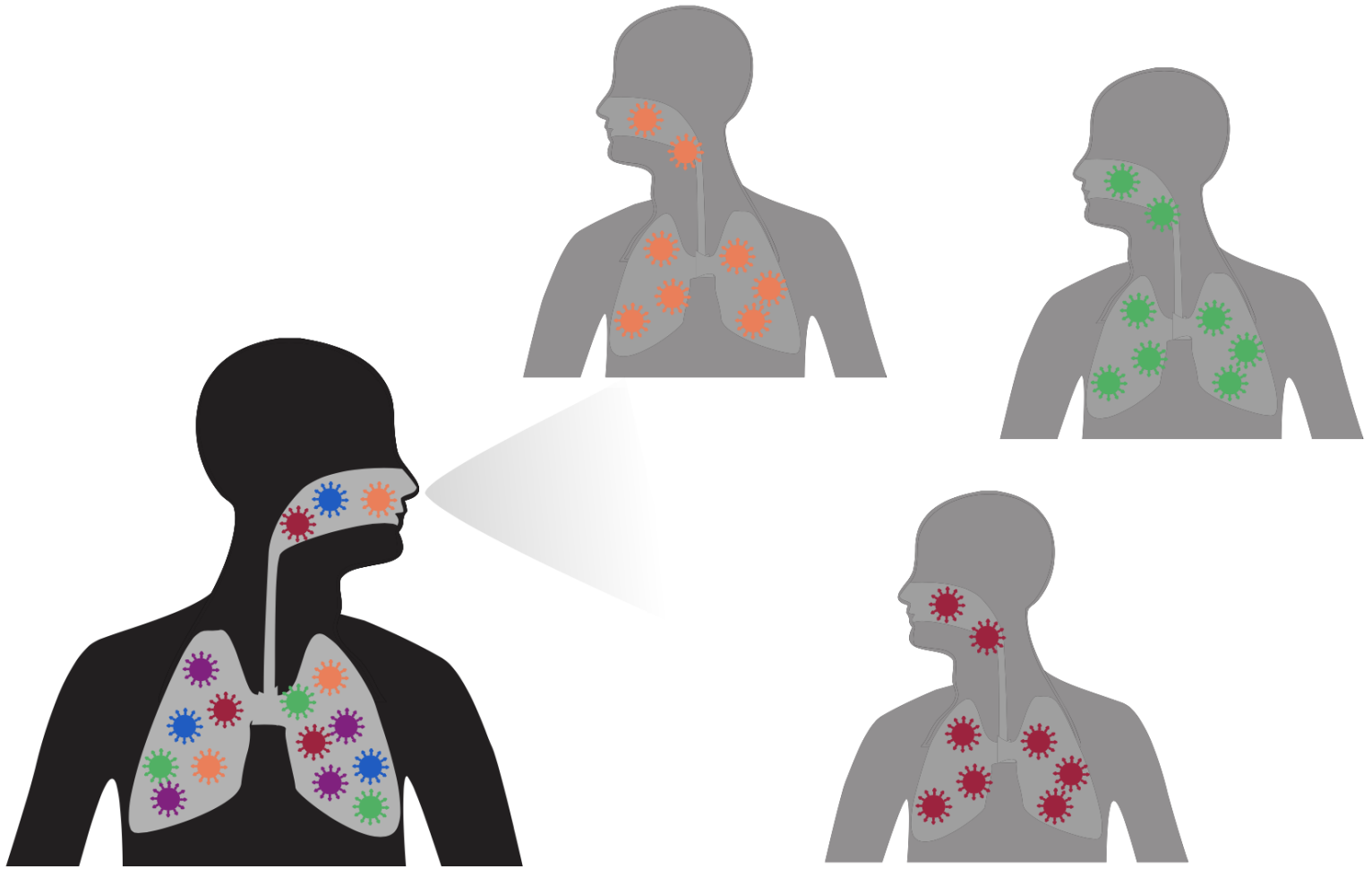


Yeast display or **lentiviral pseudotype** libraries allow us to measure many mutants at once by pooling them all together and reading out effects of mutations by deep sequencing (Starr et al, 2020; Dadonaite et al, 2022)

Limitations of using experiments to understand mutation effects

- Laborious in three years, entire field has only made large-scale measurements for two proteins:
  - spike and its RBD (Starr et al, 2022; Dadonaite et al, 2022)
  - Mpro (Flynn et al, 2022; Iketani et al, 2022)
- Lab assays measure effects of mutations in cells or mice, not humans.
- Some viral proteins have poorly understood functions that lack good lab assays

Nature is "testing" effects of viral mutations in humans all the time



Average neutral single-nucleotide mutation has occurred ~15,000 independent times in human transmitted SARS-CoV-2

- Viral substitution rate at synonymous sites: ~**7.5e-4 substitutions/year** (Neher, 2022)
- Typical infection duration: ~5 days = **0.01 years/infection**
- Total human infections with SARS-CoV-2: **~6e9 infections** (as of early 2023)
- So total synonymous substitutions per site: 7.5e-4 x 0.01 x 6e9 = **45,000**
- There are three possible mutations per site: 45,000 / 3 = **15,000**
- Mutation spectrum uneven, so some mutations have occurred more than others:
  - C->T mutations have occurred ~50,000 times
  - A->C mutations have occurred ~1,000 times

We can use publicly available human SARS-CoV-2 sequences to "read out" effects of viral mutations on human transmission

- We use the ~6.5 million public sequences in the UShER mutation-annotated tree
- These sequences represent ~0.1% of all human SARS-CoV-2 infections as of early 2023

First calculate how often each mutation **expected** to be observed without selection by analyzing 4-fold degenerate sites

UUU, UUC } Phe UUA, UUG } Leu	UCU, UCC, UCA, UCG } Ser	UAU, UAC } Tyr UAA, UAG } Stop	UGU, UGC } Cys UGA Stop UGG Trp
CUU, CUC, CUA, CUG } Leu	CCU, CCC, CCA, CCG } Pro	CAU, CAC } His CAA, CAG } Gln	CGU, CGC, CGA, CGG } Arg
AUU, AUC, AUA } Ile AUG Met*	ACU, ACC, ACA, ACG } Thr	AAU, AAC } Asn AAA, AAG } Lys	AGU, AGC } Ser AGA, AGG } Arg
GUU, GUC, GUA, GUG } Val	GCU, GCC, GCA, GCG } Ala	GAU, GAC } Asp GAA, GAG } Glu	GGU, GGC, GGA, GGG } Gly

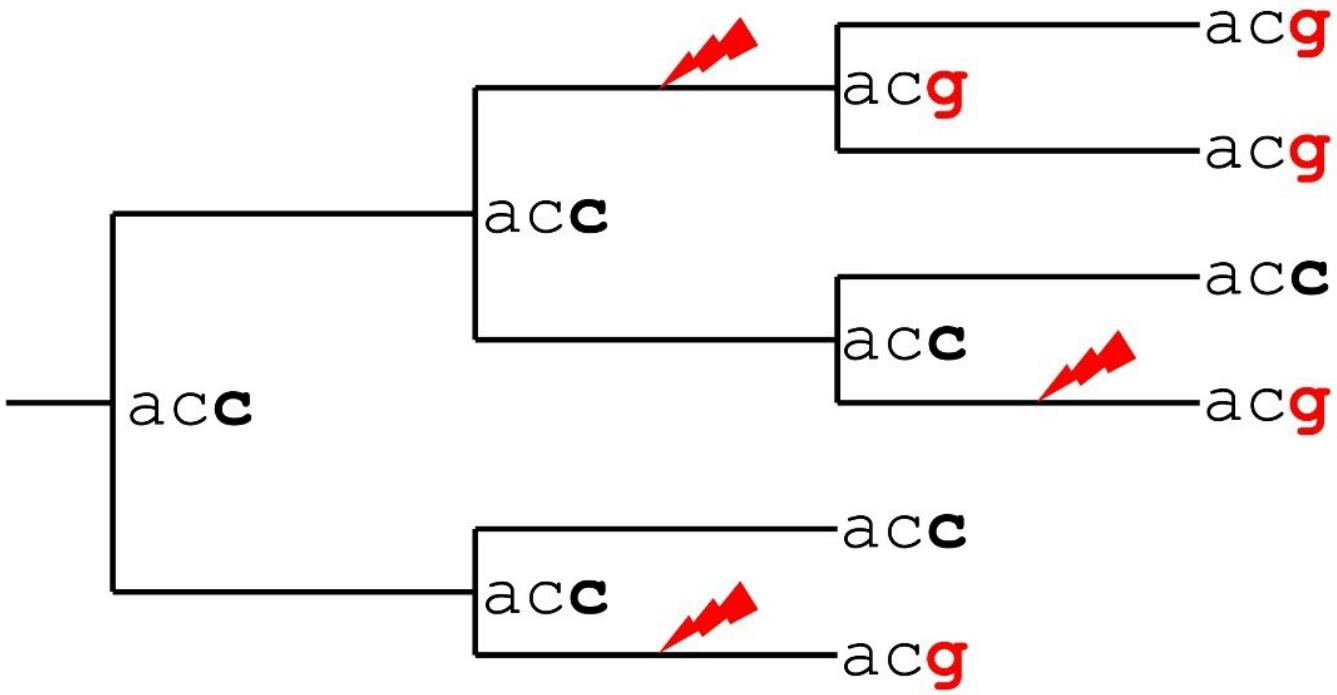
At **four-fold degenerate sites**, all nucleotide mutations are synonymous (do not change protein sequence).

Evolution at such sites is primarily driven by underlying mutation spectrum due to absence of protein-level selection.

(There could still be some selection due to factors such as RNA structure.)

Bloom, Beichman, Neher, Harris (2022)

We count unique occurrences of mutation, not number of sequences with mutation



We count the number of ***mutation occurrences*** on branches of the phylogenetic tree, not the mutated sequences in final alignment. So in above tree, we count three c->g mutations (indicated by) even though more than three sequences have mutation.

Bloom and Neher (2023)

Mutations expected to be observed ~8 to ~500 times in absence of selection

https://jbloomlab.github.io/SARS2-mut-fitness/avg_counts.html

Bloom and Neher (2023)

There are enough sequences to calculate effects on a per-mutation basis

- We calculate effect as log of actual versus expected mutation counts
- Effects of zero indicate neutral mutation, negative indicates deleterious mutation
- Estimates are more accurate (less noise) for mutations with larger expected counts

Bloom and Neher (2023)

Distribution of effects of all mutations

https://jbloomlab.github.io/SARS2-mut-fitness/effects_histogram.html

Bloom and Neher (2023)

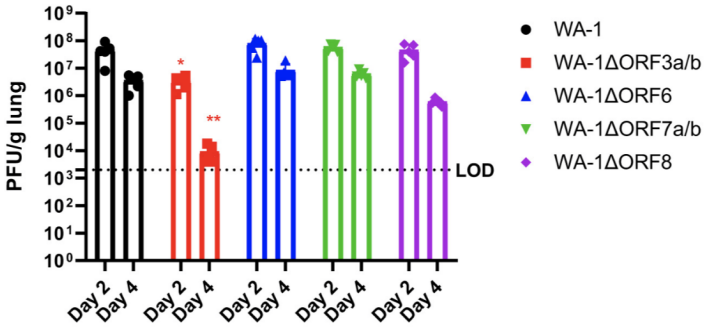
We can see which genes are under strong purifying selection

Bloom and Neher (2023)

Among accessory genes, ORF3a is under strongest selection against stop codons

https://jbloomlab.github.io/SARS2-mut-fitness/effects_dist.html

Experiments show that only accessory gene deletion that strongly attenuates virus in animal models is ORF3 (McGrath et al, 2022)



We can also look in detail at mutation level

https://jbloomlab.github.io/SARS2-mut-fitness/ORF3a.html

These maps can identify constrained sites

https://jbloomlab.github.io/SARS2-mut-fitness/nsp6.html

Estimated mutation effects are robust to sequence sampling location

Estimated mutation effects are robust to viral clade identity

Estimated mutation effects correlate well with deep mutational scanning

Two spike deep mutational scans using different underlying methodologies: lentiviral pseudotyping of spike or yeast display of RBD

Maps of mutation effects to all viral proteins

https://jbloomlab.github.io/SARS2-mut-fitness/