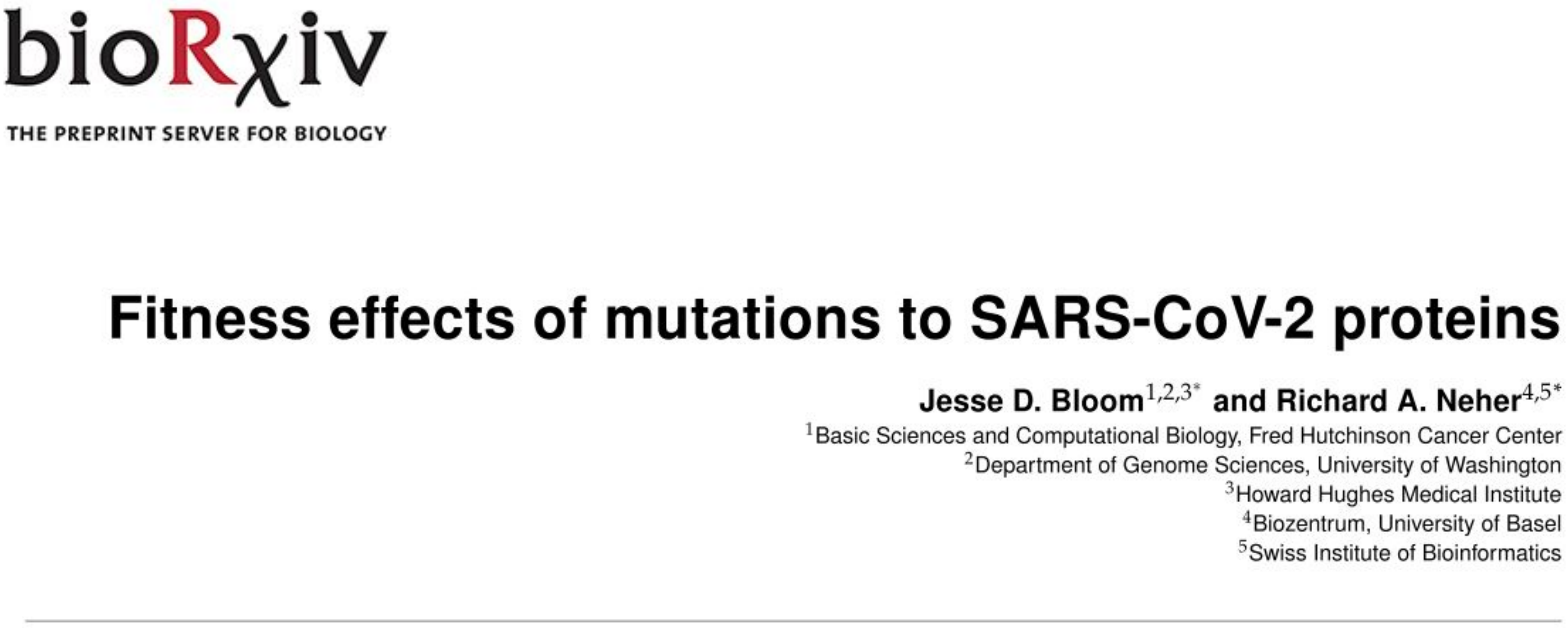


Fitness effects of mutations to SARS-CoV-2 proteins

Jesse D. Bloom[1,2,3*] **and Richard A. Neher**[4,5*]
[1]Basic Sciences and Computational Biology, Fred Hutchinson Cancer Center
[2]Department of Genome Sciences, University of Washington
[3]Howard Hughes Medical Institute
[4]Biozentrum, University of Basel
[5]Swiss Institute of Bioinformatics

doi: https://doi.org/10.1101/2023.01.30.526314

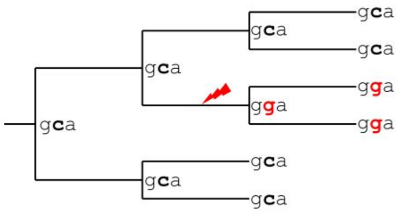
BASIC IDEA

- experimentally measuring single mutational effects is hopeless
  - **deep mutational scanning** data only available **for two SARS-CoV-2 proteins**
- **idea:**
  - there are now so many SARS-CoV-2 sequences, that all non-deleterious single-nucleotide mutations are **expected to independently occur many times**

  → **frequent mutations are beneficial while rare ones are deleterious**
- what does “frequent” mean?
  - compare the **expected number of mutations given no selection**
  - with the **actual number** of observed mutations

CALCULATING FITNESS EFFECTS

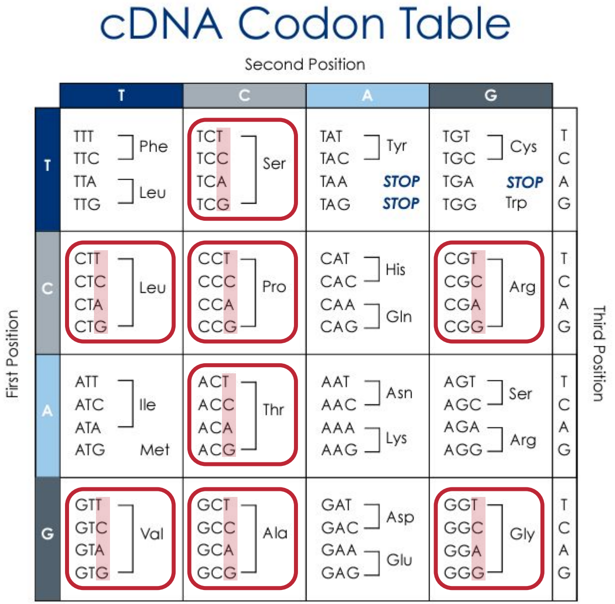
- use the **phylogenetic tree** of (~ 7 million) public SARS-CoV-2 sequences
  - only count **individual occurrences** of mutations



+ quality-control steps

- **expected mutation counts**: from **four-fold degenerate sites**
  - **no protein-level selection**

1. take all four-fold degenerate sites along the genome
2. choose the ones with original nucleotide *x*
3. count the number of individual mutations with nucleotide *y* at these sites
4. divide by the number of relevant sites



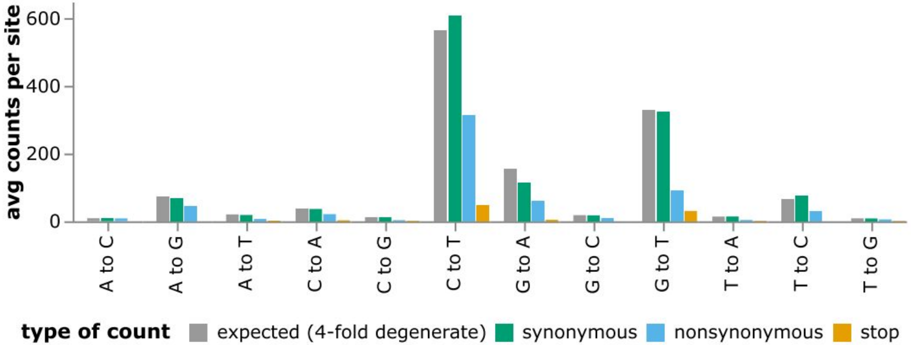
cDNA Codon Table

Second Position

First Position	T		C		A		G		Third Position
T	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys	T
	TTC		TCC		TAC		TGC		C
	TTA	Leu	TCA		TAA	STOP	TGA	STOP	A
	TTG		TCG		TAG	STOP	TGG	Trp	G
C	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	Gln	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	Lys	AGA	Arg	A
	ATG	Met	ACG		AAG		AGG		G
G	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	Glu	GGA		A
	GTG		GCG		GAG		GGG		G

CALCULATING FITNESS EFFECTS

- **actual mutation counts**: same technique for **all possible genomic sites**

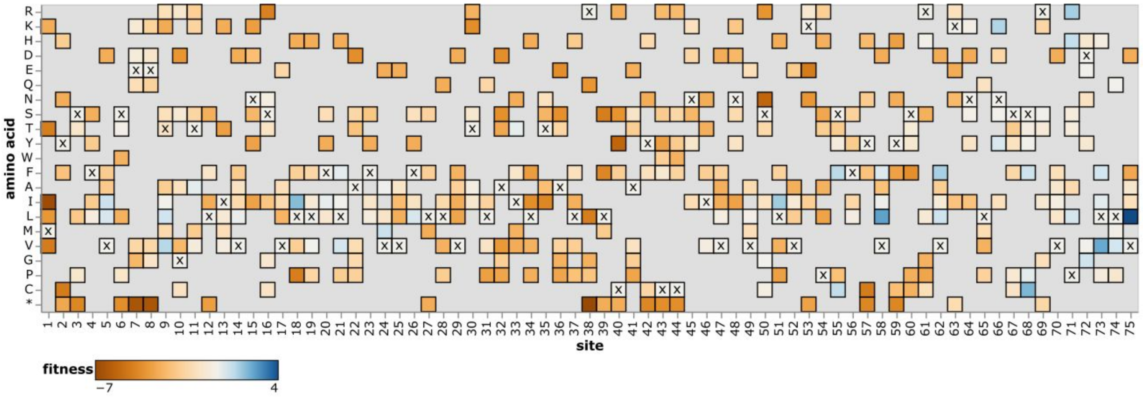


- synonymous (including 4-fold deg.) ~ expected
- nonsynonymous mutations are rare
- stop-codon mutations are even rarer

purifying selection

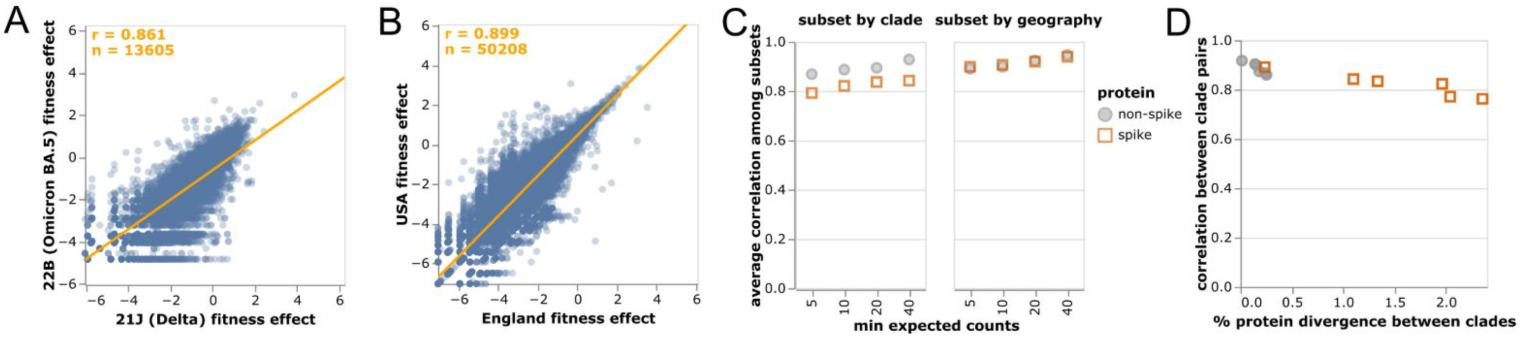
CALCULATING FITNESS EFFECTS

- **converting to AA counts** from nucleotide counts
  - sum all nucleotide mutation counts that encode the same AA mutation
  - exclude any mutations that are not from the clade-founder codon identity
- **overall estimate**: sum for all possible clades
- **estimated fitness:** $\Delta f = \log\left(\frac{n_{actual}+0.5}{n_{expected}+0.5}\right)$



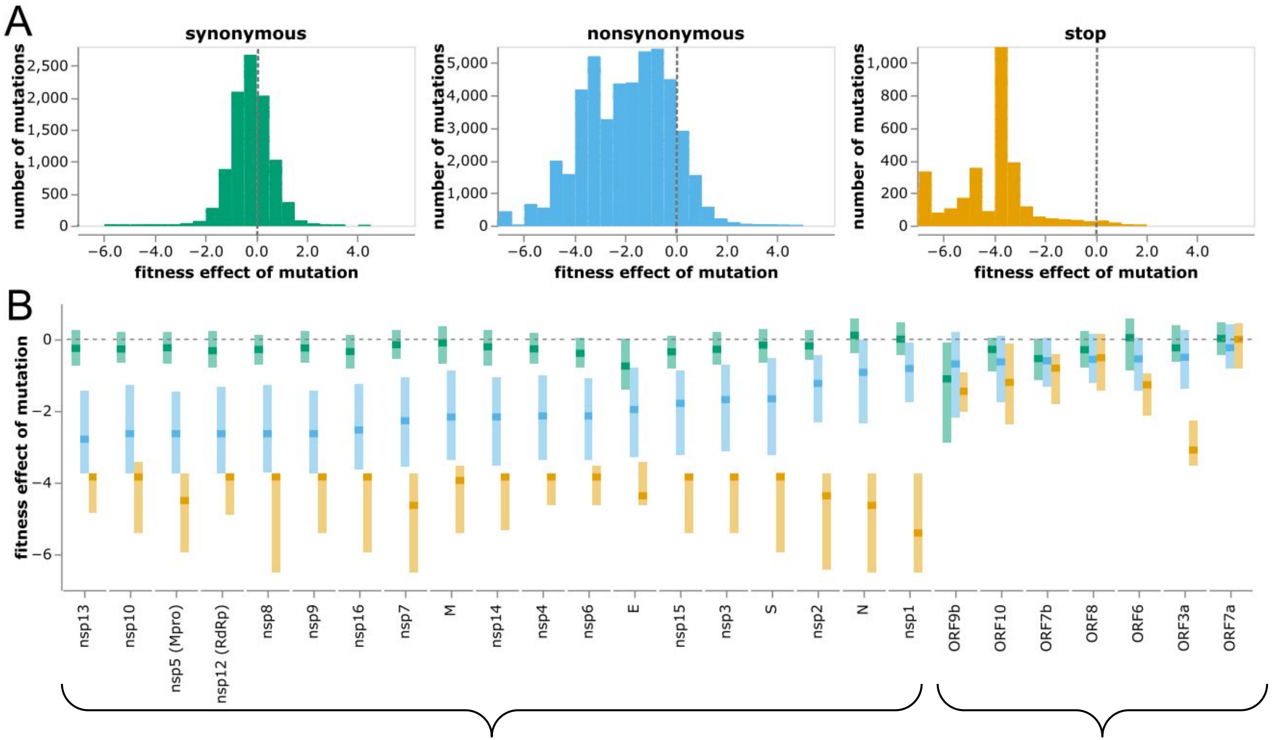
https://jbloomlab.github.io/SARS2-mut-fitness/

ROBUSTNESS



- **correlations** between **subsampled datasets** are reasonably **high**
  - differences due to **statistical noise**? → limiting data to **high-confidence mutations**
    - subsetting by geography → correlation consistently increases
    - subsetting by clade → correlation increases for non-spike, but **remains lower for spike**
  - correlations decline for clades with higher protein divergence
    - **epistasis**? **changes in the selective landscape**?

PURIFYING SELECTION ON PROTEINS

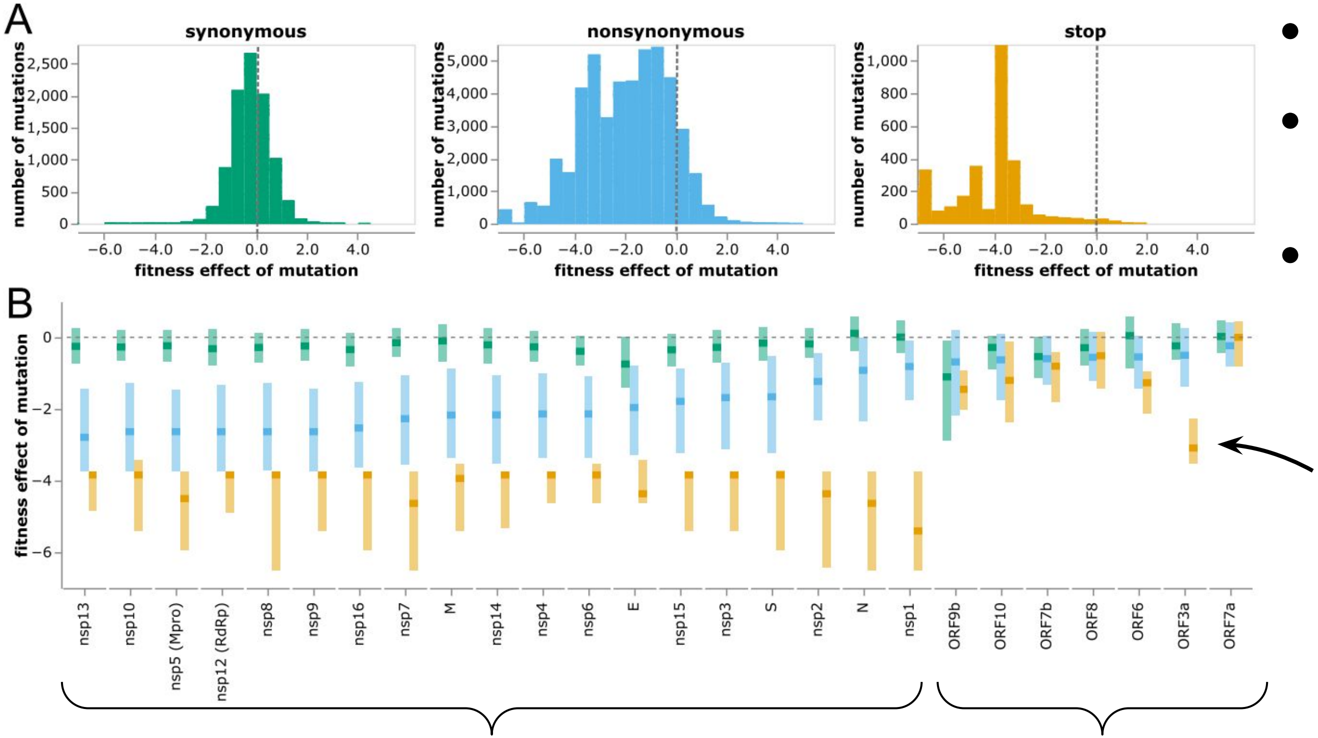


structural and non-structural proteins are under strong purifying selection

accessory proteins are under little constraint

- synonymous mutations are usually neutral
- nonsynonymous mutations have varied effects
- stop-codon mutations are deleterious

PURIFYING SELECTION ON PROTEINS



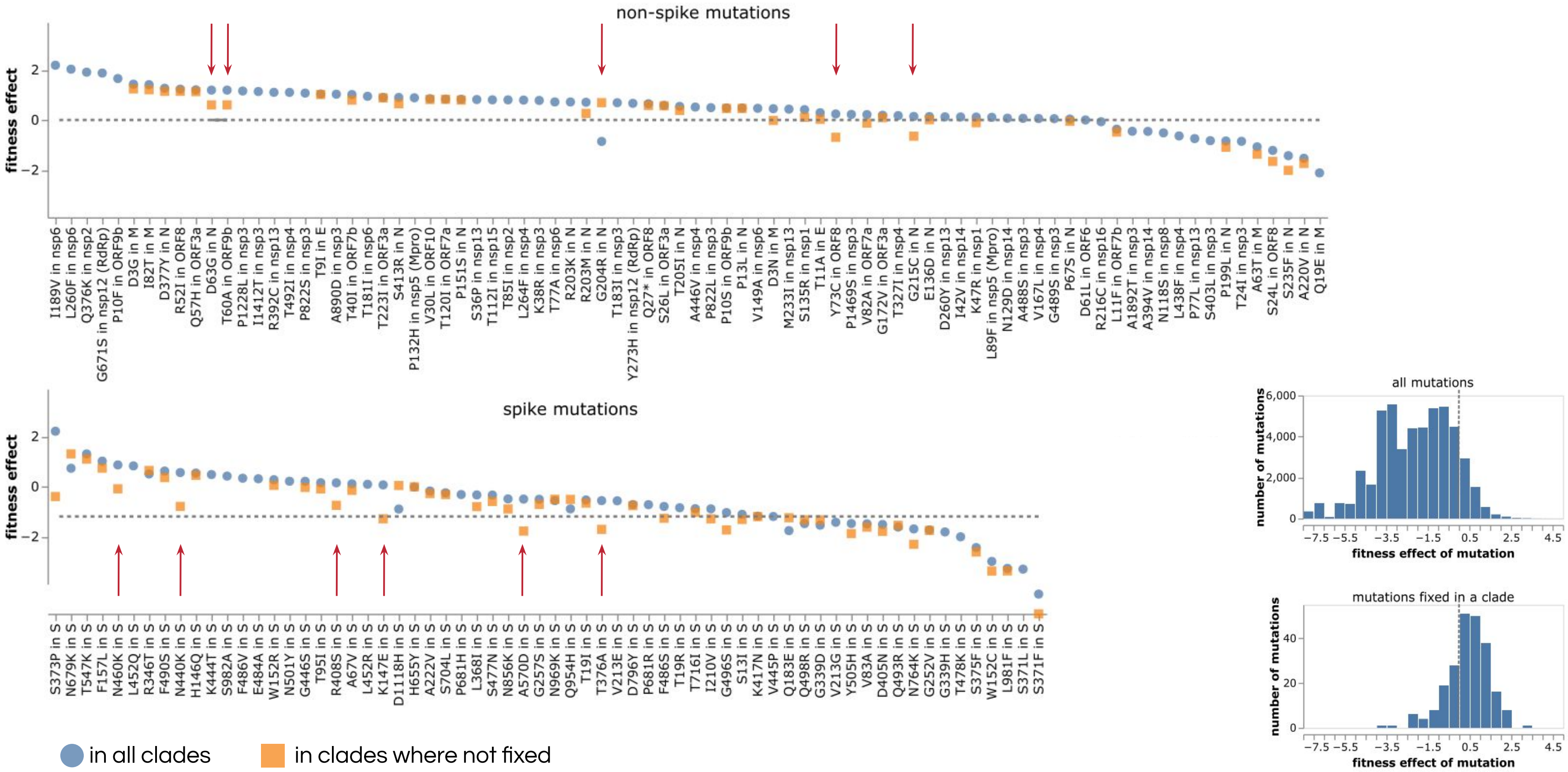
- synonymous mutations are usually neutral
- nonsynonymous mutations have varied effects
- stop-codon mutations are deleterious

with the exception of ORF3a
(consistent with experiments, but otherwise unexplained)

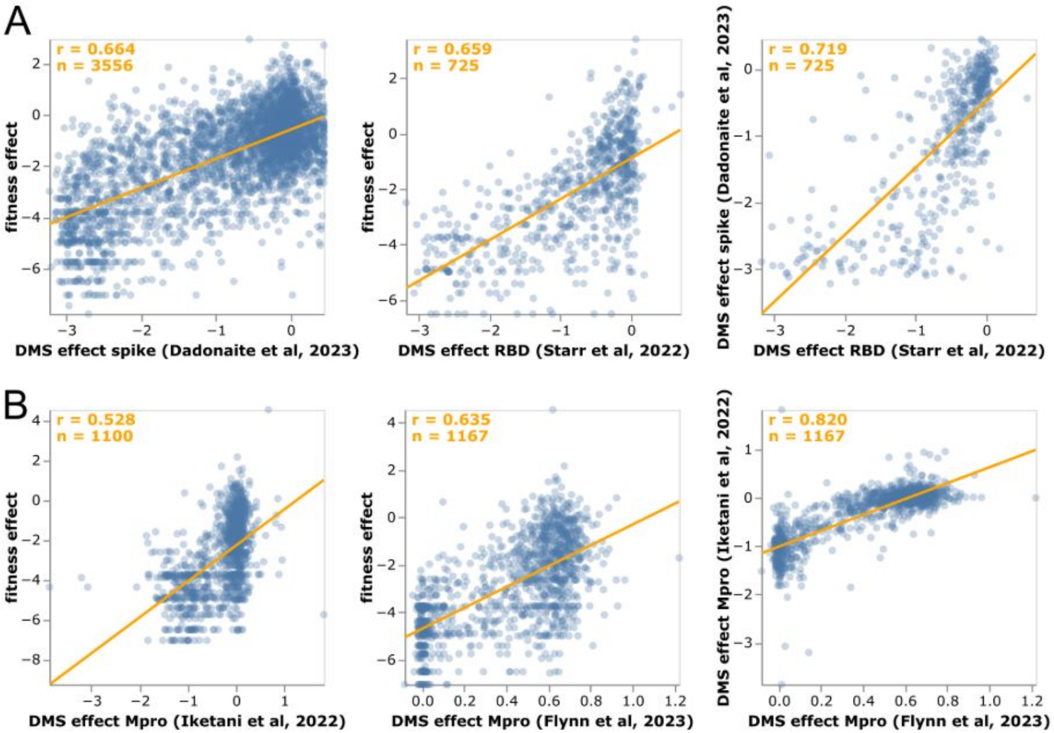
structural and non-structural proteins are under strong purifying selection

accessory proteins are under little constraint

MUTATIONS FIXED IN CLADES



MUTATION EFFECT VS. DMS

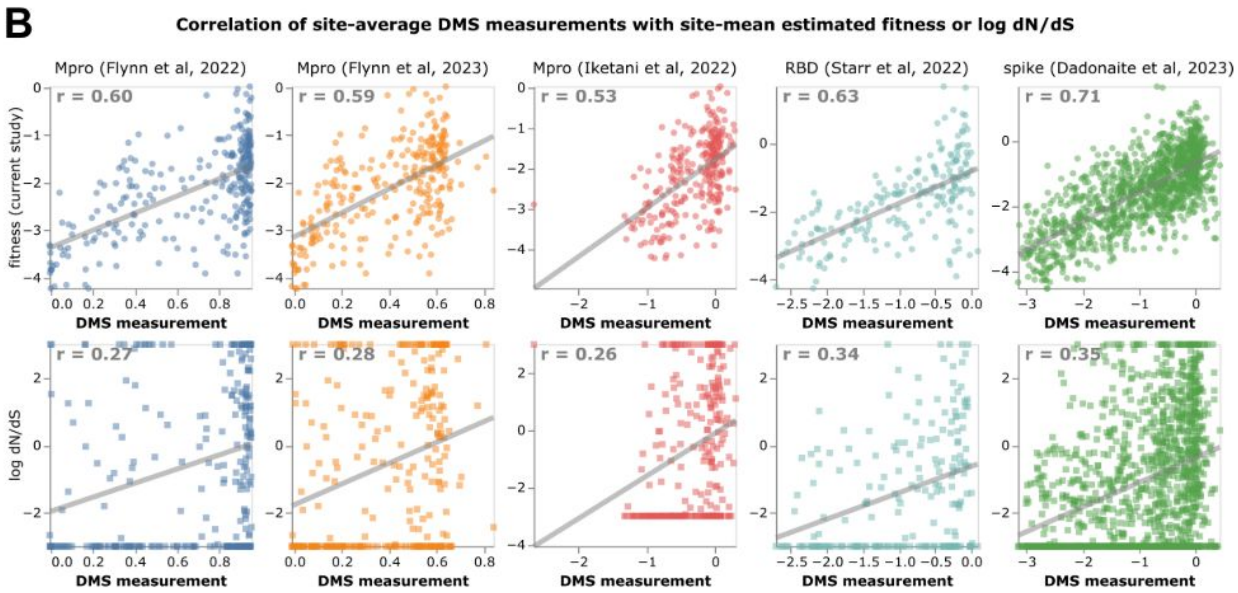


- for Spike: **correlation between fitness effect and experiments is similar to that of between different experiments**

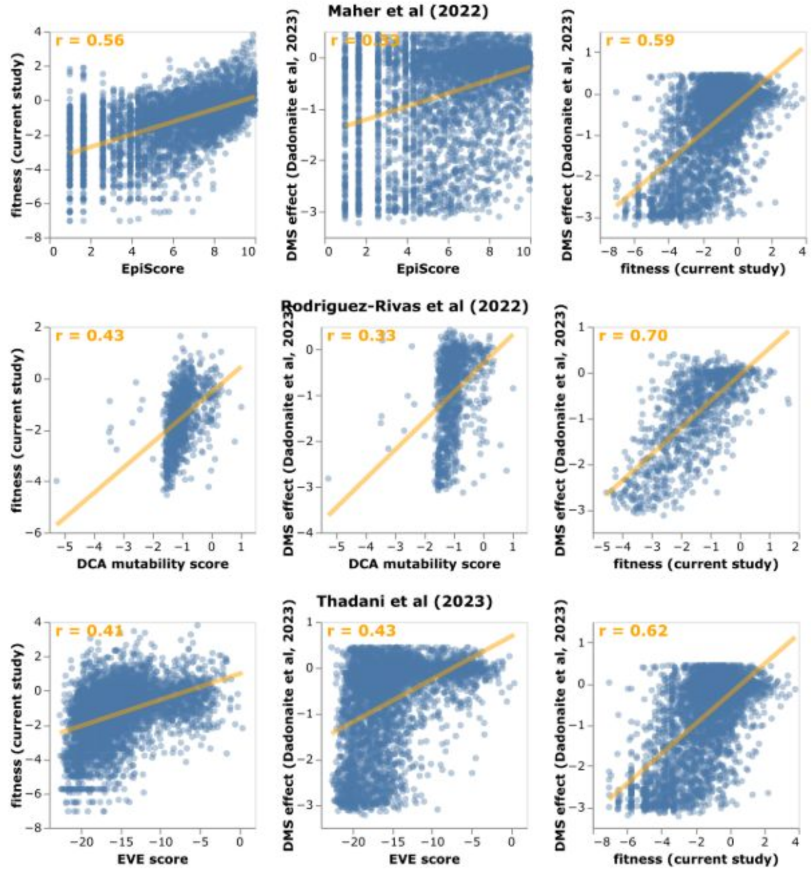
- for Mpro: correlation between experiments is higher than between fitness effect and experimental results
  ← systematic experimental artefacts?

MUTATION EFFECT VS. OTHER PREDICTORS

average site-level fitness effect **correlates better with experimental results** than traditional log dN/dS values



MUTATION EFFECT VS. OTHER PREDICTORS



Maher et al (2022)

Rodriguez-Rivas et al (2022)

Thadani et al (2023)

- fitness effect **moderately correlates** with other predictors of mutational effect
- fitness effect **outperforms all other predictors** when correlated to experimental DMS results

Maher et al:

(already discussed) LINK

no epistasis

Rodriguez-Rivas et al:

considers epistasis

Thadani et al: LINK

EVEscape deep learning model

trained on sequences available before 2020

supposedly “captures” epistasis