INSTITUTE FOR MATHEMATICS
STRUCTURES CLUSTER OF EXCELLENCE
UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

Michael Bleher
Institute for Mathematics, Heidelberg University

– Dioscuri Seminar – 13 Jan 2026 –

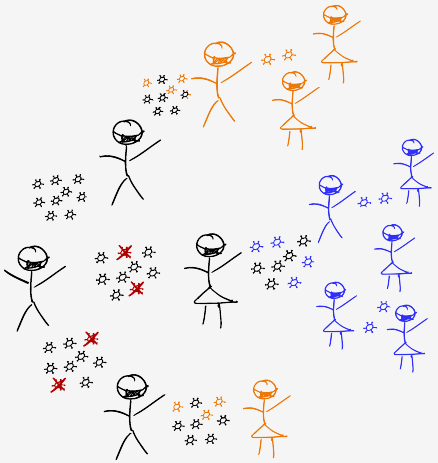
Topological Signatures of Convergence in Viral Evolution

based on

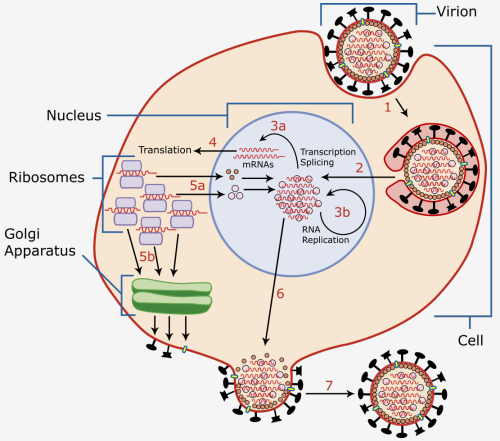
arXiv:2106.07292
arXiv:2207.03394
& ongoing work

Joint w/

Andreas Ott, Maximilian Neumann (Karlsruhe)
Lukas Hahn (Heidelberg)
Juan Patiño-Galindo (Mount Sinai)
Mathieu Carrière (Inria Sophia-Antopolis)
Raul Rabadan (Columbia)
Ulrich Bauer (Munich)
Samuel Braun, Holger Obermaier, Mehmet Soysal, René Caspart (Karlsruhe)



A Brief Introduction to Genomics and Epidemiology



Author: YK Times, Wikimedia Commons (CC BY-SA 3.0)

Viral Genome

Encodes instructions for host cell.

Sequence of nucleotides A, C, T, G.

```
>seq-id|date|location
ATGAAGAGCTTAGTCCTAG
```

Viral Life Cycle

1. Virus binds to host cell
2. Viral genome enters cell & nucleus
3. Replication and Transcription of viral RNA
4. Translation *(production of viral proteins)*

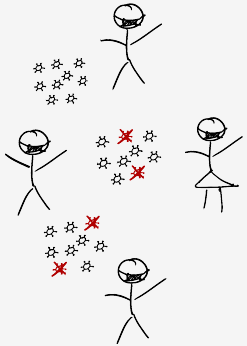
5. & 6. Assembly

7. Release

A Brief Introduction to Genomics and Epidemiology

Transmission modulates frequencies

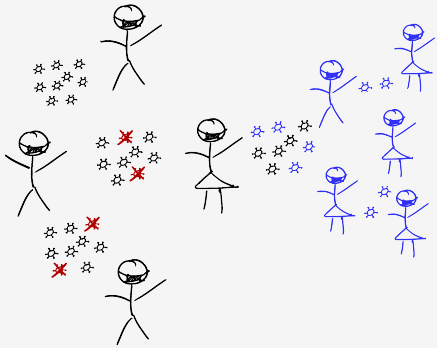
- not every mutation is beneficial



A Brief Introduction to Genomics and Epidemiology

Transmission modulates frequencies

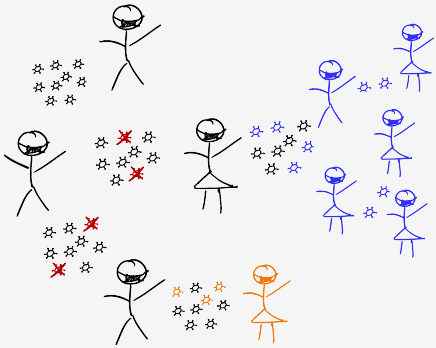
- not every mutation is beneficial
- mutations that spread widely are not necessarily beneficial (founder effects)



A Brief Introduction to Genomics and Epidemiology

Transmission modulates frequencies

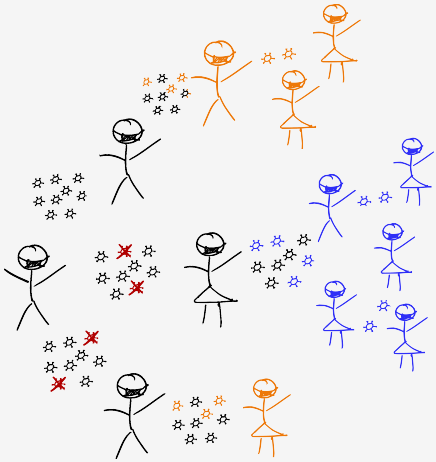
- not every mutation is beneficial
- mutations that spread widely are not necessarily beneficial (founder effects)
- not every beneficial mutation catches on



A Brief Introduction to Genomics and Epidemiology

Transmission modulates frequencies

- not every mutation is beneficial
- mutations that spread widely are not necessarily beneficial (founder effects)
- not every beneficial mutation catches on
- BUT: beneficial mutations tend to *appear repeatedly* (and may then spread more widely)



Recurrence is a hallmark of increased fitness.

Example: evolution of wings (birds, bats, insects)

Geometry of Viral Evolution

Viral genome data X

Goal

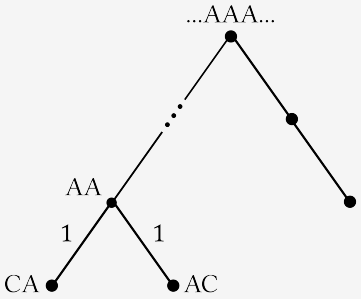
Monitor evolution of virus and determine influence of (single or groups of) mutations on its fitness.

Key idea

Reconstruct **phylogenetic tree** from sequences

$$\text{Hamming distance} = \text{Tree distance}$$

Minimum spanning tree reconstructs ancestral relations.

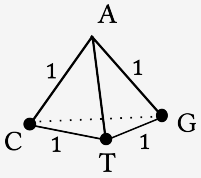


Hamming Geometry

Σ = finite alphabet
Σ^n = sequences of length n over Σ

RNA/DNA: $\Sigma = \{A, C, T, G\}$

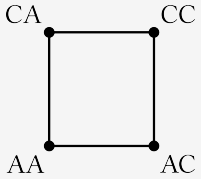
```
>seq 0
ATGAAGAGCTTAGTCCTAG
>seq 1
ATGAAGAGCTAAGTCCTAG
```



Hamming distance

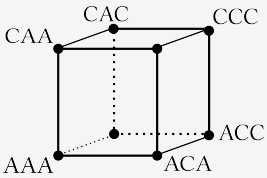
= number of differing positions between two sequences

$$d_H(x, y) := \#\{i \mid x_i \neq y_i\}$$



Hamming Space (Σ^n, d_H)

- Discrete metric space, highly symmetric
- Geodesic (shortest path = sequence of point mutations)



Geometry of Viral Evolution – Revisited

Viral genome data X

Goal

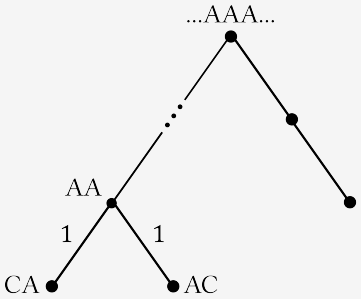
Monitor evolution of virus and determine influence of (single or groups of) mutations on its fitness.

Key idea

Reconstruct **phylogenetic tree** from sequences

$$\text{Hamming distance} = \text{Tree distance}$$

Minimum spanning tree reconstructs ancestral relations



Geometry of Viral Evolution – Revisited

Viral genome data $X \subset \Sigma^n$

Goal

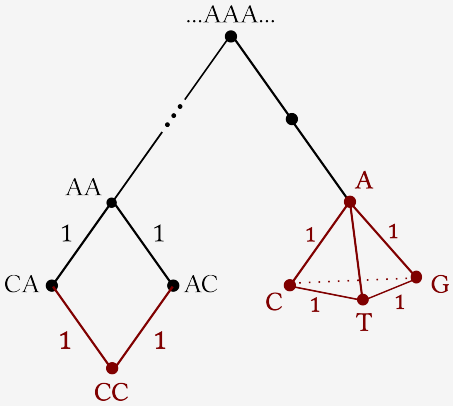
Monitor evolution of virus and determine influence of (single or groups of) mutations on its fitness.

Key idea

Reconstruct **phylogenetic network** from sequences

Hamming distance $\neq$ Tree distance

Minimum spanning tree reconstructs ancestral relations, but is not unique.
Use this to detect interesting phenomena.



Contractibility Lemma(s)

Rips, Gromov (60's & 80's)

(X, d) a δ-hyperbolic geodesic metric space $\implies$ $\mathrm{VR}_r(X)$ is contractible, $r \geq 4\delta$.

Chan, Carlsson, Rabadan (2013)

If (X, d) is a tree, then $H_n(\mathrm{VR}_\bullet(X, d)\,) = 0$, $n \geq 1$.

Bauer, Roll (2022)

(X, d) a δ-hyperbolic ν-geodesic finite metric space $\implies$ $\exists$ discrete gradient collapse:

$$\mathrm{VR}_s(X) \searrow \mathrm{VR}_r(X) \searrow \{*\}\ ,\ s > r \geq 4\delta + 2\nu$$

$\implies$ **Persistent homology detects deviations from tree-like data**

Contractibility Lemma(s)

Rips, Gromov (60's & 80's)

(X, d) a δ-hyperbolic geodesic metric space $\implies$ $\mathrm{VR}_r(X)$ is contractible, $r \geq 4\delta$.

Chan, Carlsson, Rabadan (2013)

If (X, d) is a tree, then $H_n(\mathrm{VR}_\bullet(X, d)\,) = 0$, $n \geq 1$.

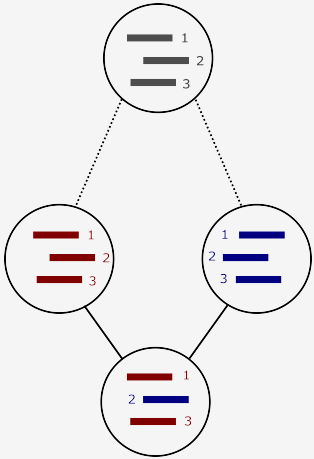
Bauer, Roll (2022)

(X, d) a δ-hyperbolic ν-geodesic finite metric space $\implies$ $\exists$ discrete gradient collapse:

$$\mathrm{VR}_s(X) \searrow \mathrm{VR}_r(X) \searrow \{*\}\ ,\ s > r \geq 4\delta + 2\nu$$

$\implies$ **Persistent homology detects deviations from tree-like data (and thus evolutionary relevant phenomena!)**

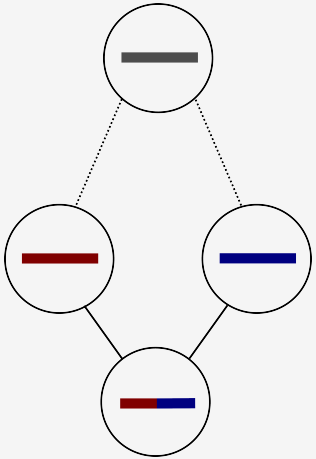
Topology of Viral Evolution



Reassortment

Some viruses have disconnected genome, e.g. Flu (HxNy).
Co-infection can lead to “reassortment” during assembly.

Topology of Viral Evolution



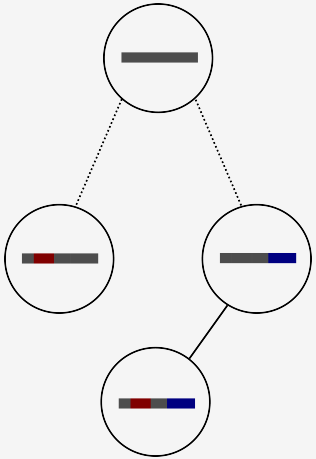
Reassortment

Some viruses have disconnected genome, e.g. Flu (HxNy).
Co-infection can lead to “reassortment” during assembly.

Recombination

Replication apparatus can “switch template”.
Co-infection can lead to recombination into a hybrid genome.

Topology of Viral Evolution



Reassortment

Some viruses have disconnected genome, e.g. Flu (HxNy).
Co-infection can lead to “reassortment” during assembly.

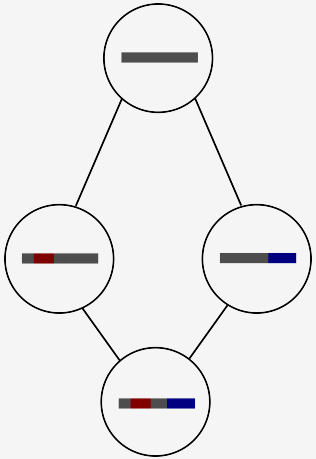
Recombination

Replication apparatus can “switch template”.
Co-infection can lead to recombination into a hybrid genome.

Convergence / Homoplasy

independent emergence of similar traits.
example: evolution of flight (mammals, insects, bats)

Topology of Viral Evolution



Reassortment

Some viruses have disconnected genome, e.g. Flu (HxNy).
Co-infection can lead to “reassortment” during assembly.

Recombination

Replication apparatus can “switch template”.
Co-infection can lead to recombination into a hybrid genome.

Convergence / Homoplasy

independent emergence of similar traits.
example: evolution of flight (mammals, insects, bats)

Persistent Homology of SARS-CoV-2

Consider genomic data with Hamming distance as finite metric space (X, d_H).

```
>seq 0
ATGAAGAGCTTAGTCCTAG
>seq 1
ATGAAGAGCTAAGTCCTAG
>seq 2
ATGAACAGCTAAGTCCTAG
```

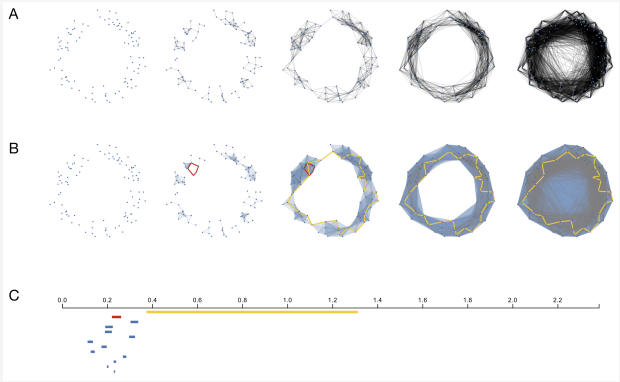
$$d_H = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{pmatrix}$$

Construct Vietoris-Rips complex

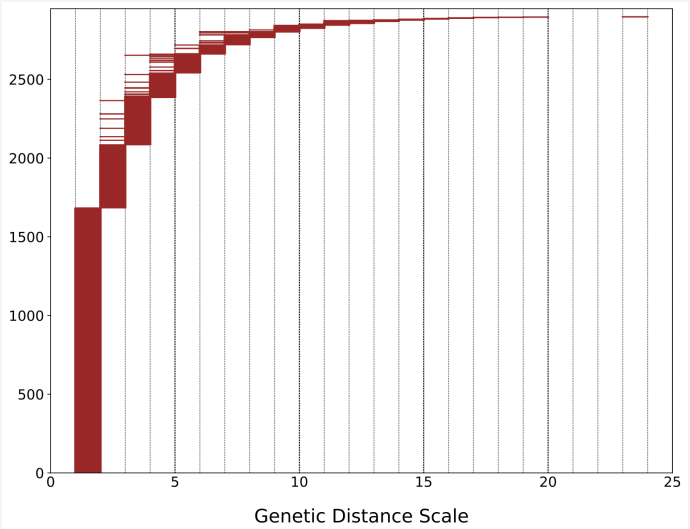
$$VR_\bullet(X, d_H)$$

Calculate homology

$$H_k(VR_\bullet(X, d_H))$$



Persistent Homology of SARS-CoV-2



February 28th, 2021

$\sim 450,000$ isolates

$\sim 160,000$ unique sequences

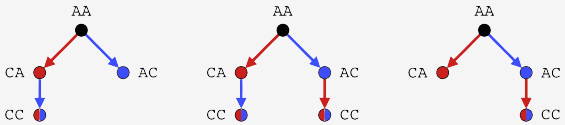
$\Rightarrow |H_1| \sim 2,900$

Signal or Noise?

Back-of-the-envelope

$p \simeq 1/30,000 \simeq \mathcal{O}(10^{-4})$
$\#\text{unique sequences} = \mathcal{O}(10^{6})$

$\Rightarrow$ expect $\mathcal{O}(100)$ cycles are noise

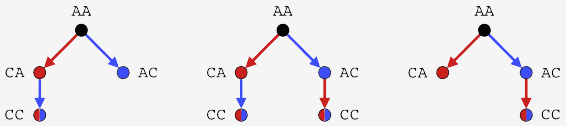


Signal or Noise?

Back-of-the-envelope

$p \simeq 1/30,000 \simeq \mathcal{O}(10^{-4})$
$\#\text{unique sequences} = \mathcal{O}(10^6)$

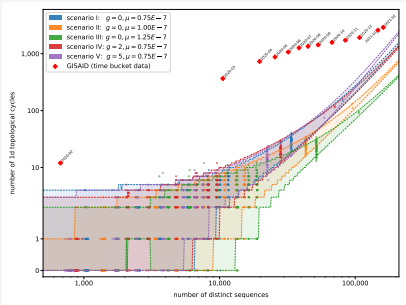
$\Rightarrow$ expect $\mathcal{O}(100)$ cycles are noise



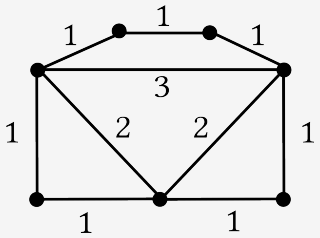
Simulations of neutral evolution

- uniform mutation probability
- no fitness advantages
- no recombinations

$\Rightarrow$ expect 350-400
(at worst: $1,200 \sim 50\%$)



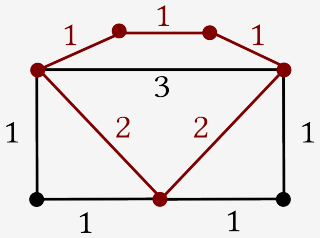
Extracting the Signal: From homology classes to mutations.



example: $[1, 3)$-persistent class

Which mutations are responsible for homology?

Extracting the Signal: From homology classes to mutations.

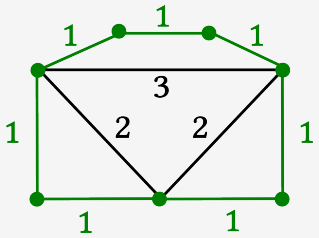


example: $[1, 3)$-persistent class

Which mutations are responsible for homology?

use cycle representatives

Extracting the Signal: From homology classes to mutations.



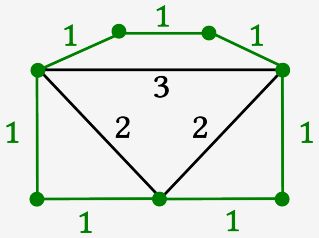
example: $[1, 3)$-persistent class

Which mutations are responsible for homology?

use cycle representatives
from **exhaustive** reduction

Every edge of length 1 corresponds to a unique single neucleotide variation (SNV).

Extracting the Signal: From homology classes to mutations.



example: $[1, 3)$-persistent class

Which mutations are responsible for homology?

use cycle representatives
from **exhaustive** reduction

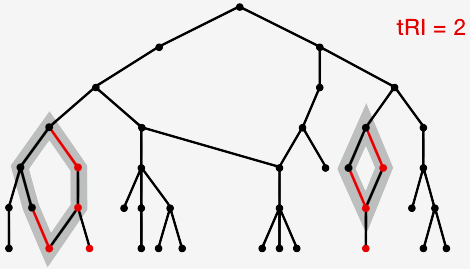
Every edge of length 1 corresponds to a unique single neucleotide variation (SNV).

SNV-cycles := Exhaustive representatives of $[1, d)$ classes

The topological Recurrence Index (tRI)

Z_{SNV} – set of SNV-cycles in H_1

μ – mutation of interest
(notation: `RefPosAlt`, e.g. `A614C`)



Definition

$$\mathrm{tRI}(\mu) := \#\{\gamma \in Z_{\mathrm{SNV}} \mid \mu \in \gamma\}$$

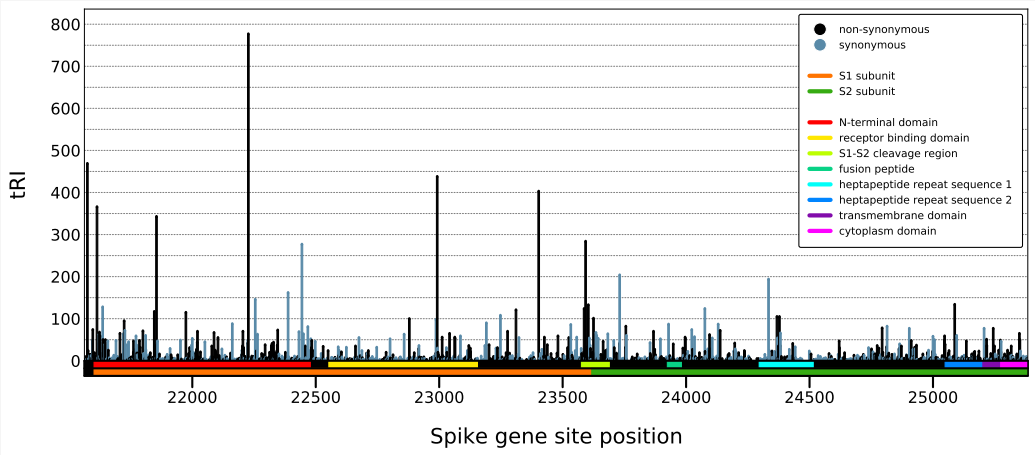
Proposition

$\mathrm{tRI}(\mu) =$ minimal number of independent occurences of μ in X.

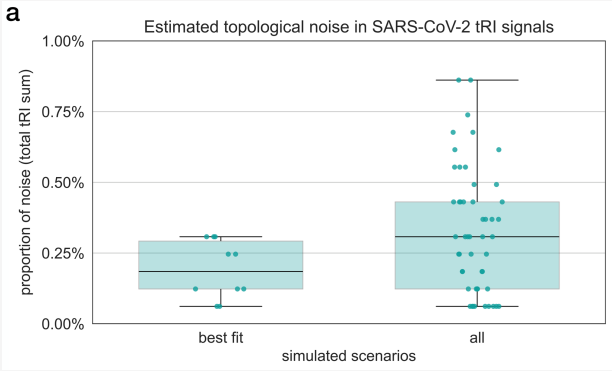
$\Longrightarrow$ **tRI is a measure for convergence**

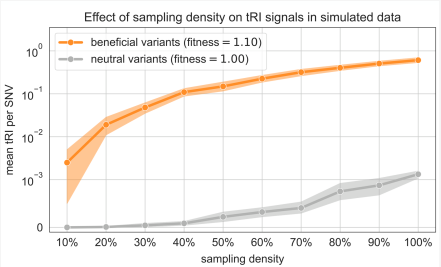
(and thus fitness)

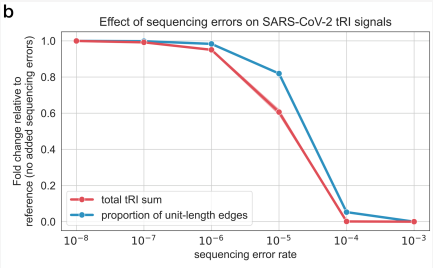
Topological Recurrence of Spike mutations



Robustness of tRI

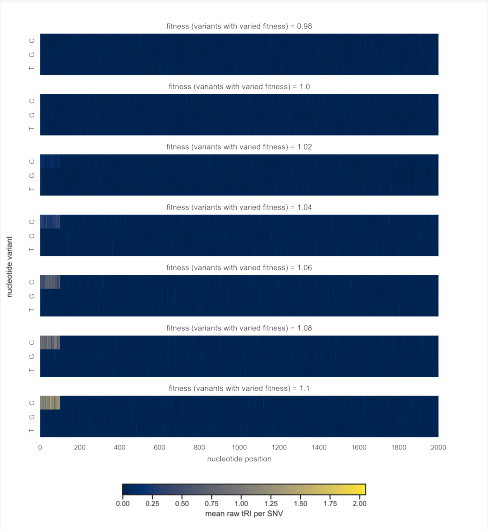


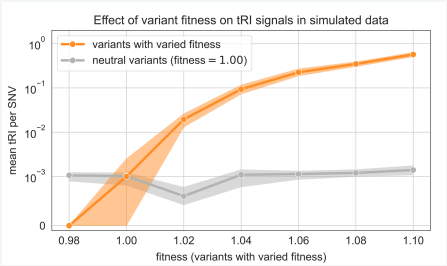




tRI is robust to noise, sequencing errors, and subsampling.

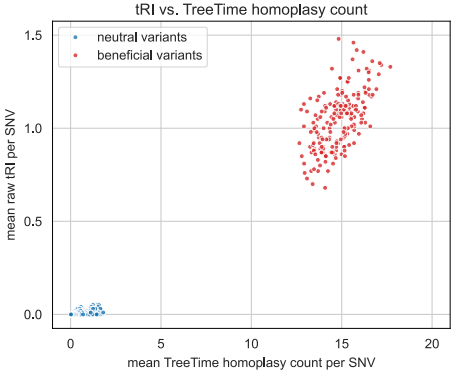
Comparison with Fitness in Simulations





tRI is sensitive to fitness increase

Comparison with Established Fitness Measures – Recurrence counts (tree-based, simulations)

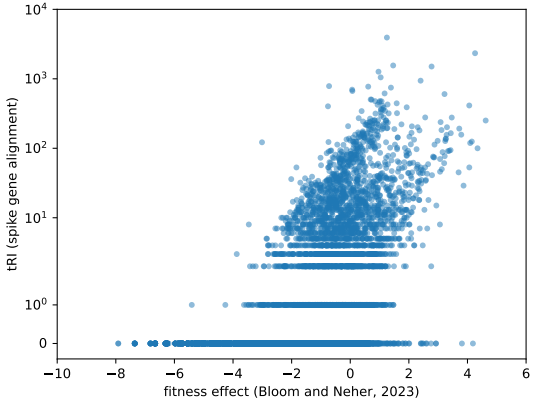


Spearman correlation (mean ± 95% CI) between raw tRI profiles and TreeTime homoplasy counts

Neutral variants	Beneficial variants
$\rho = 0.04 \pm 0.007$	$\rho = 0.59 \pm 0.01$
$p = 0.32 \pm 0.06$	$p < 1e\text{-}09$

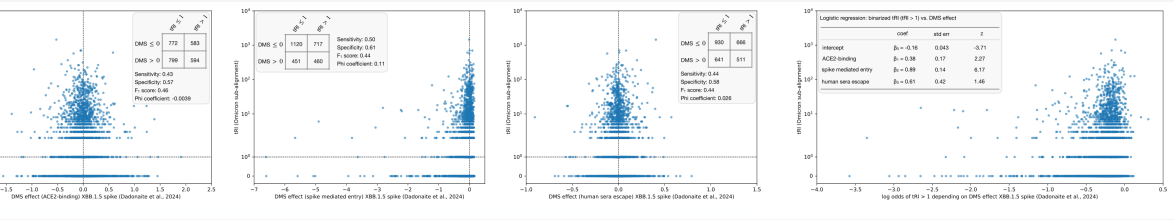
tRI is correlated with tree-based recurrence counts (HomoplasyFinder, Crispell et al., 2019)

Comparison with Established Fitness Measures – Fitness Index (tree-based, SARS-CoV-2)



tRI is correlated with tree-based fitness index (Bloom & Neher, 2022)

Comparison with Established Fitness Measures – Deep Mutational Scanning (experimental, SARS-CoV-2)



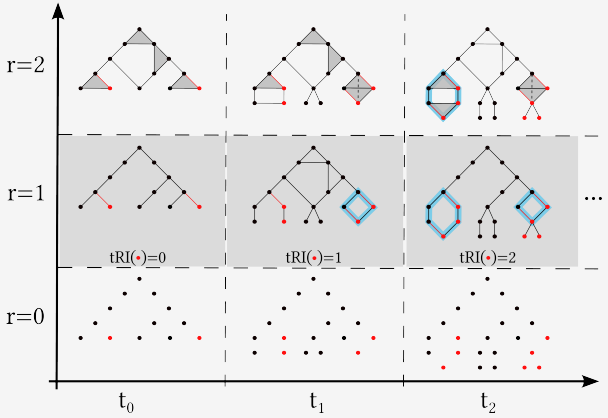
tRI is correlated with experimental measures of fitness increase.
(Starr et al., 2022)

Time, Multipersistence, and a Computational Trick

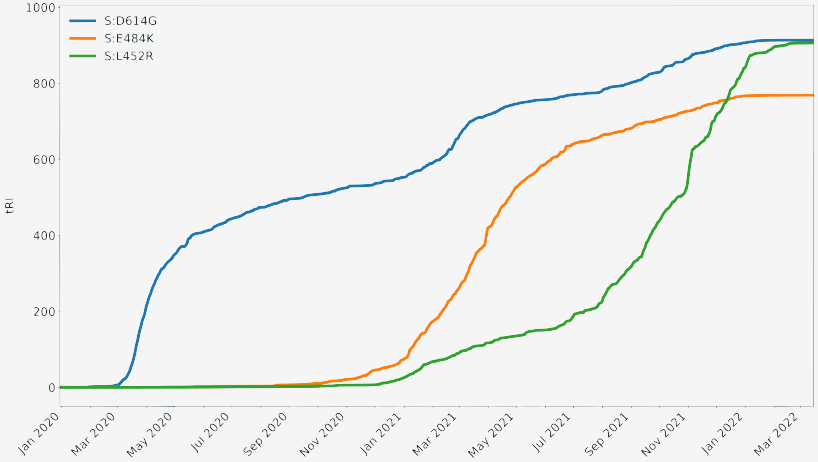
Include time series information
→ **2-parameter persistence**

Good News: Get all SNV-cycles from restriction to 1d subfiltration @ $r = 1$.

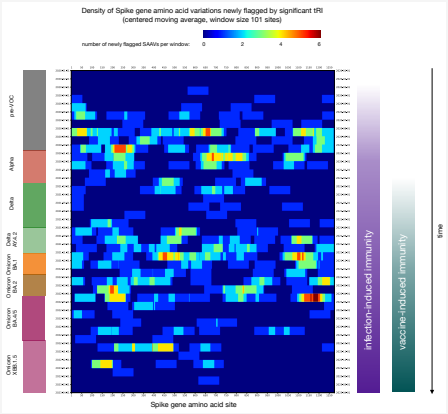
Trick: Equivalent to deformation of metric
→ **Ripser "Add-on": `MuRiT`**
Multipersistence through Rips Transformations
calculates pathwise persistence from
distance matrix + additional filtration



`EvotRec.py` – Evolution of topological Recurrence



Dynamic Fitness Landscape and Epistasis

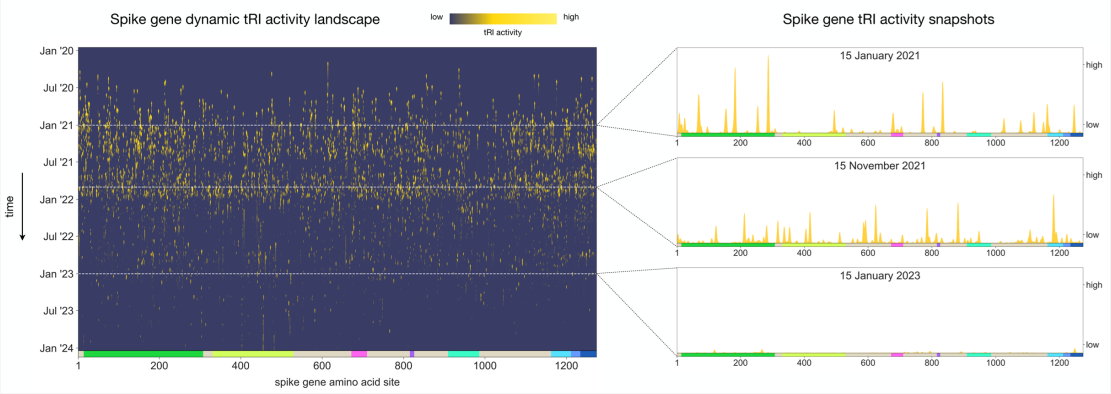


Time-resolved tRI activity along the genome shows surprising amount of time-dependence.

Looks like tRI measures *epistasis*:
influence of current mutational background on fitness of newly acquired mutations.

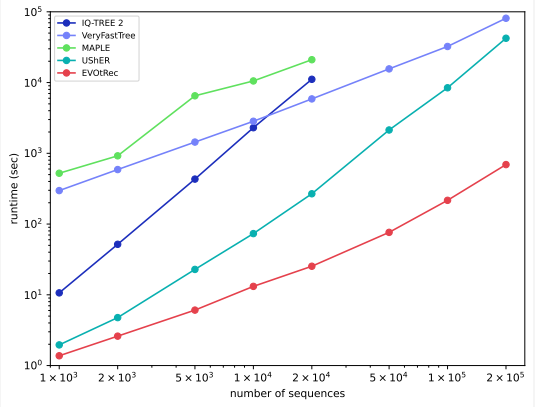
This is possible because SNV-cycles are *localized* in a particular genetic background.

Dynamic Fitness Landscape and Epistasis

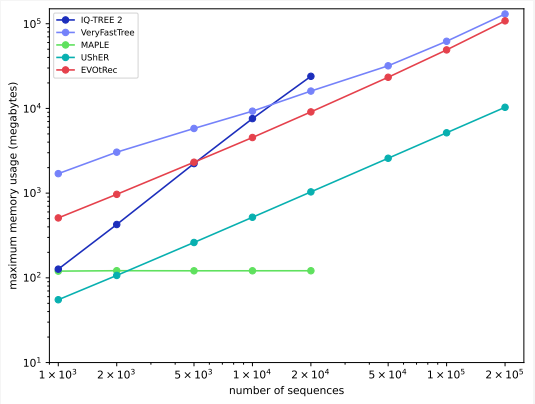


Computational Benchmarks

Runtime

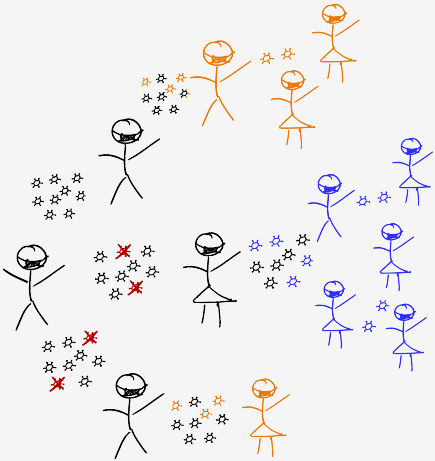


Memory



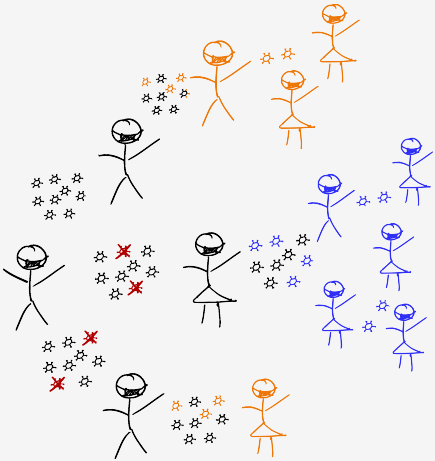
Summary

- Persistent homology measures evolutionary relevant phenomena
- topological Recurrence Index (tRI) is sensitive to fitness effects
- `EvotRec` computations are fast and efficient
- tRI activity might allow study of epistasis
- Differentiation between beneficial and deleterious mutations must rely on experiments, but persistent homology can tell us where to look



Summary

- Persistent homology measures evolutionary relevant phenomena
- topological Recurrence Index (tRI) is sensitive to fitness effects
- `EvotRec` computations are fast and efficient
- tRI activity might allow study of epistasis
- Differentiation between beneficial and deleterious mutations must rely on experiments, but persistent homology can tell us where to look



Thank you!