INSTITUTE FOR MATHEMATICS

STRUCTURES
CLUSTER OF
EXCELLENCE

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Michael Bleher

Institute for Mathematics, Heidelberg University

4TH WORKSHOP ON COMPUTATIONAL PERSISTENCE, GRAZ 2024

FAST COMPUTATION OF PATHWISE PERSISTENCE

IN PANDEMIC-SCALE SARS-COV-2 GENOME DATA

based on

arXiv:2106.07292

arXiv:2207.03394

& ongoing work

Joint w/

Andreas Ott, Maximilian Neumann (Karlsruhe)

Lukas Hahn (Heidelberg)

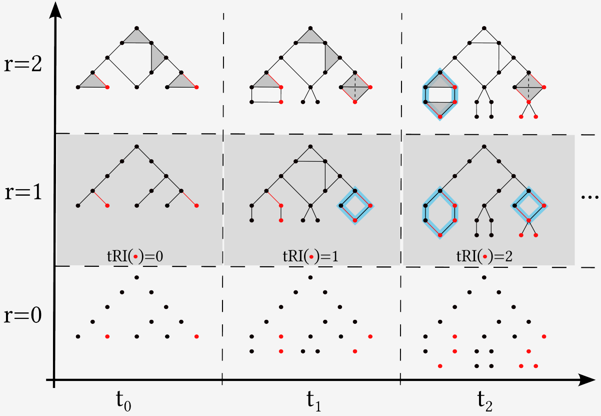
Juan Patiño-Galindo (Mount Sinai)

Mathieu Carrière (Inria Sophia-Antopolis)

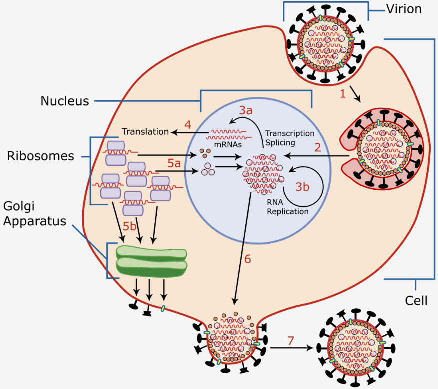
Raul Rabadan (Columbia)

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)*

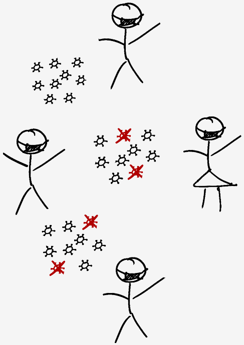
5. & 6. Assembly

7. Release

A Brief Introduction to Genomics and Epidemiology

Transmission modulates frequencies

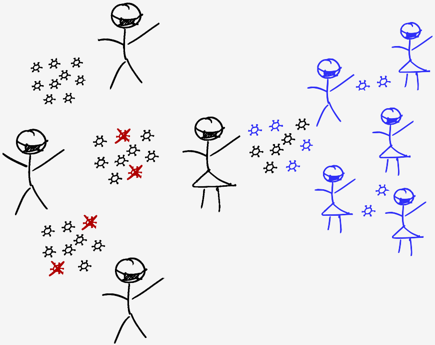
- not every mutation is beneficial



A Brief Introduction to Genomics and Epidemiology

Transmission modulates frequencies

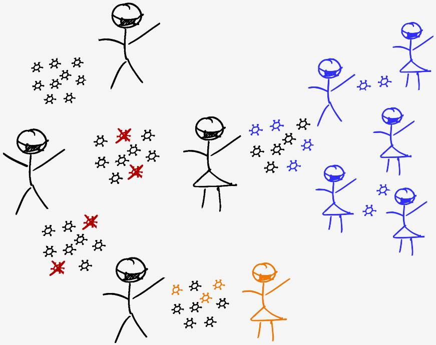
- not every mutation is beneficial
- mutations that spread widely are not necessarily beneficial (founder effects)



A Brief Introduction to Genomics and Epidemiology

Transmission modulates frequencies

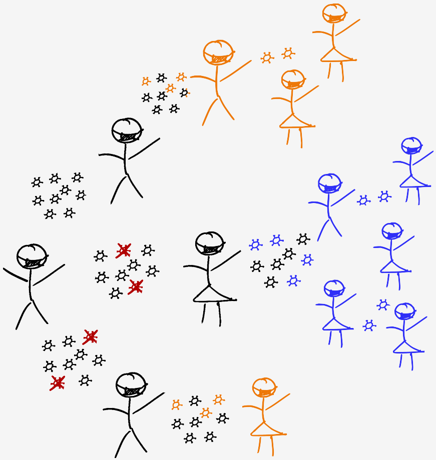
- not every mutation is beneficial
- mutations that spread widely are not necessarily beneficial (founder effects)
- not every beneficial mutation catches on



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)

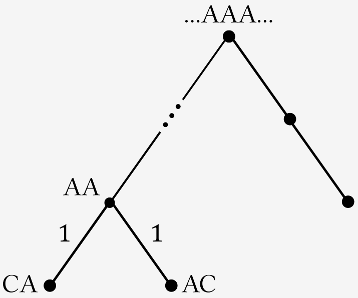
Geometry of Viral Evolution

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

Construct **phylogenetic tree** from sequences.

Hamming distance = Tree distance

Minimum spanning tree reconstructs ancestral relations



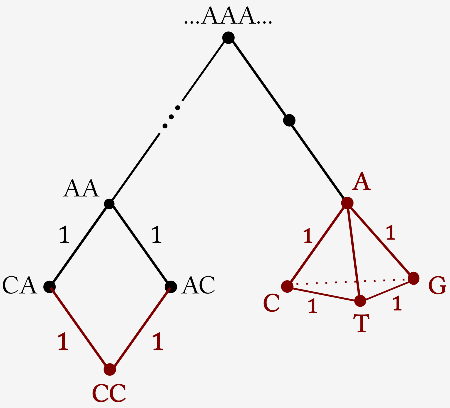
Geometry of Viral Evolution

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

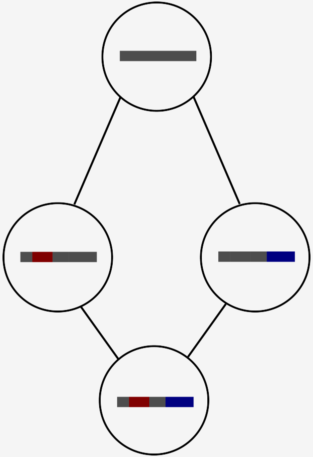
Construct **phylogenetic network** from sequences.

Hamming distance $\neq$ Tree distance

Minimum spanning tree reconstructs ancestral relations, but is not unique.



Topology of Viral Evolution



Reassortment

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

Recombination

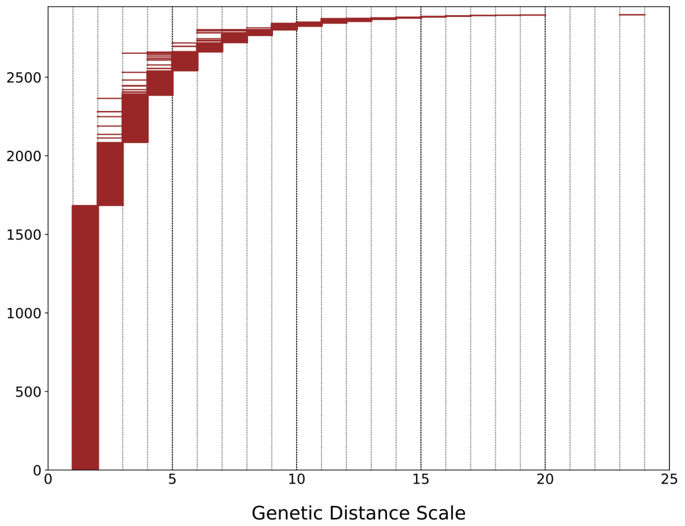
Replication apparatus can switch RNA template.
Co-infection can lead to recombination into a hybrid genome.

Convergence

independent emergence of similar traits.

$\Rightarrow$ **cycles in phylogenetic network at different scales.**

Persistent Homology of SARS-CoV-2



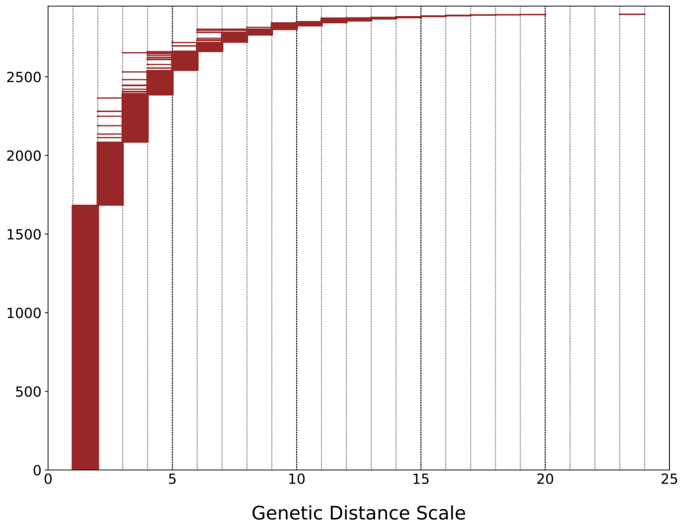
February 28th, 2021

$\sim 450,000$ isolates

$\sim 160,000$ unique sequences

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

Persistent Homology of SARS-CoV-2



February 28th, 2021

$\sim 450,000$ isolates
$\sim 160,000$ unique sequences

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

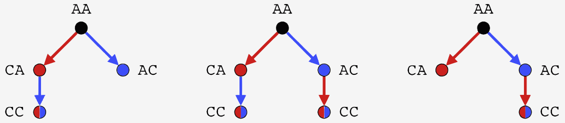
How? Luck and patience.

Signal or Noise?

Back-of-the-envelope

$p \simeq 2/30,000 \simeq \mathcal{O}(10^{-4})$
#unique sequences $= \mathcal{O}(10^6)$

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



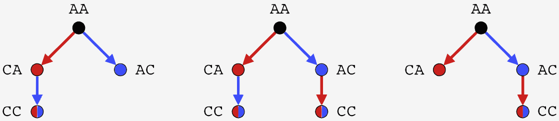
Signal or Noise?

Back-of-the-envelope

$p \simeq 2/30,000 \simeq \mathcal{O}(10^{-4})$

#unique sequences $= \mathcal{O}(10^{6})$

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

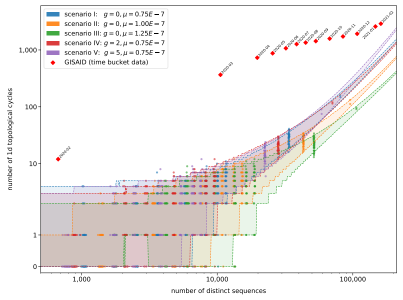


Simulations of neutral evolution

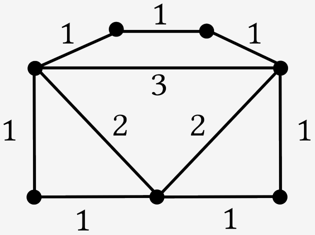
- uniform mutation probability
- no fitness advantages
- no recombinations

$\Rightarrow$ expect 350-400

(at worst: $1,200 \sim 50\%$)



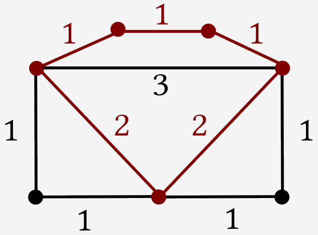
The topological Recurrence Index (tRI)



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

Which mutations are responsible for homology?

The topological Recurrence Index (tRI)

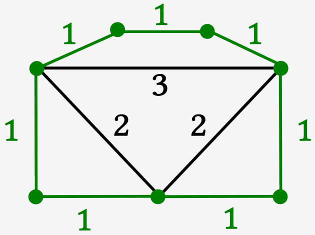


example: $[1, 3)$-persistent class

Which mutations are responsible for homology?

use cycle representatives

The topological Recurrence Index (tRI)



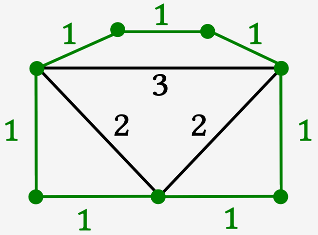
example: $[1, 3)$-persistent class

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).

The topological Recurrence Index (tRI)



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

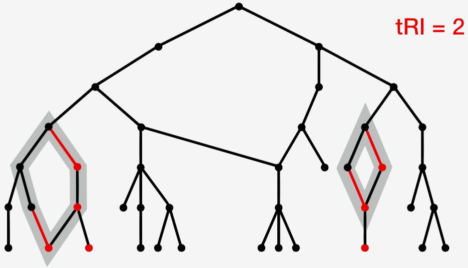
The topological Recurrence Index (tRI)

Z_{SNV} – set of all SNV-cycles in H_1

μ – mutation of interest

(notation: **RefPosAlt**, e.g. **D614G**)

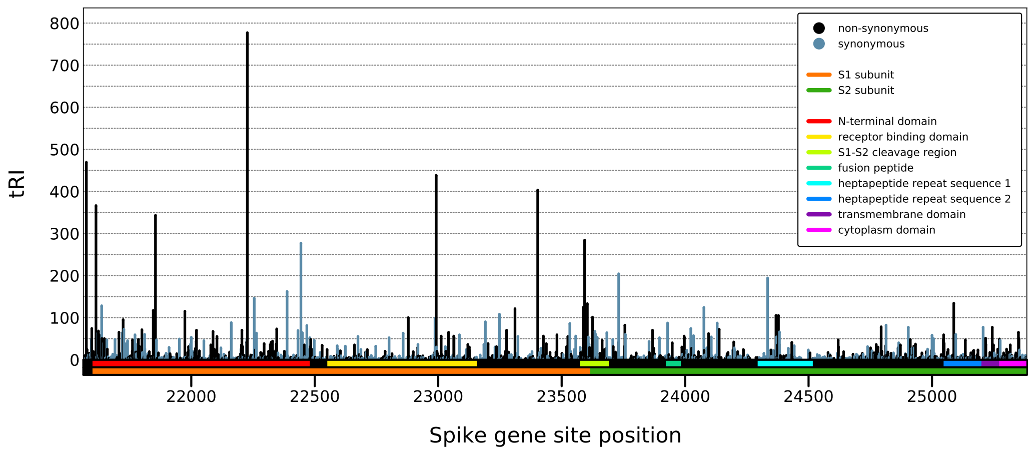
$$\mathrm{tRI}(\mu) := \#\{\gamma \in Z_{\mathrm{SNV}} \mid \mu \in \gamma\}$$



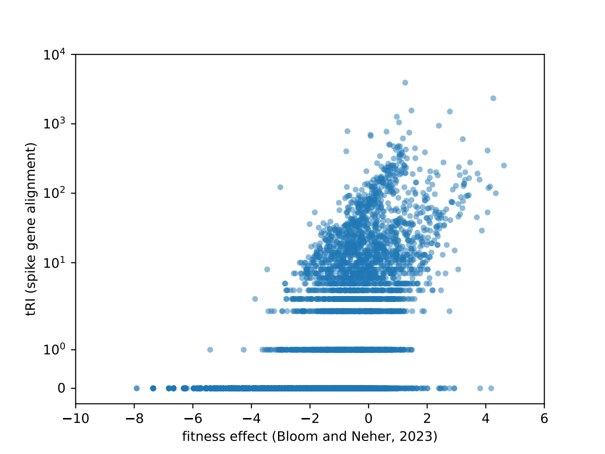
$\Longrightarrow$ **tRI is a measure for convergence**

(and thus fitness)

Topological Recurrence of Spike mutations

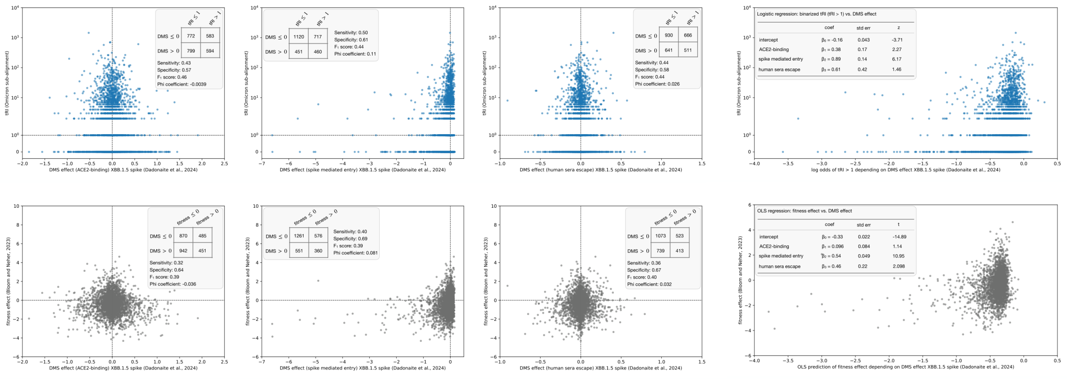


Comparison with Established Fitness Measures



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

Comparison with Established Fitness Measures



	tRI ≤ 1	tRI > 1
DMS ≤ 0	772	583
DMS > 0	799	594

Sensitivity: 0.43
Specificity: 0.57
F_1 score: 0.46
Phi coefficient: -0.0039

	tRI ≤ 1	tRI > 1
DMS ≤ 0	1120	717
DMS > 0	451	460

Sensitivity: 0.50
Specificity: 0.61
F_1 score: 0.44
Phi coefficient: 0.11

	tRI ≤ 1	tRI > 1
DMS ≤ 0	930	666
DMS > 0	641	511

Sensitivity: 0.44
Specificity: 0.58
F_1 score: 0.44
Phi coefficient: 0.026

Logistic regression: binarized tRI (tRI > 1) vs. DMS effect

	coef	std err	z
intercept	$\beta_0 = -0.16$	0.043	-3.71
ACE2-binding	$\beta_1 = 0.38$	0.17	2.27
spike mediated entry	$\beta_2 = 0.89$	0.14	6.17
human sera escape	$\beta_3 = 0.61$	0.42	1.46

	fitness ≤ 0	fitness > 0
DMS ≤ 0	870	485
DMS > 0	942	451

Sensitivity: 0.32
Specificity: 0.64
F_1 score: 0.39
Phi coefficient: -0.036

	fitness ≤ 0	fitness > 0
DMS ≤ 0	1261	576
DMS > 0	551	360

Sensitivity: 0.40
Specificity: 0.69
F_1 score: 0.39
Phi coefficient: 0.081

	fitness ≤ 0	fitness > 0
DMS ≤ 0	1073	523
DMS > 0	739	413

Sensitivity: 0.36
Specificity: 0.67
F_1 score: 0.40
Phi coefficient: 0.032

OLS regression: fitness effect vs. DMS effect

	coef	std err	t
intercept	$\beta_0 = -0.33$	0.022	-14.89
ACE2-binding	$\beta_1 = 0.096$	0.084	1.14
spike mediated entry	$\beta_2 = 0.54$	0.049	10.95
human sera escape	$\beta_3 = 0.46$	0.22	2.098

tRI is correlated with experimental measures of fitness increase.

Time, Multipersistence, and a Computational Trick

Include time series information

→ **2-parameter persistence**

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

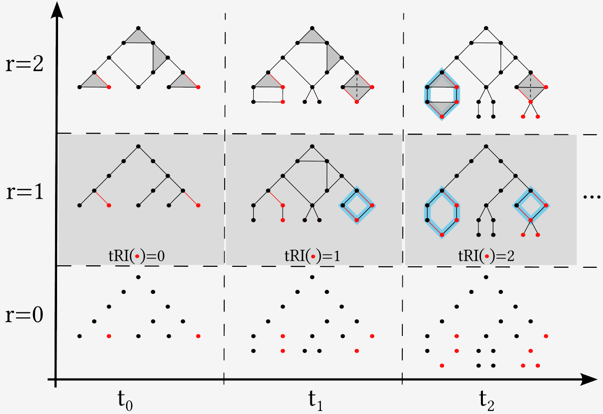
Trick: Equivalent to deformation of metric

→ **Ripser "Add-on":** `MuRiT`

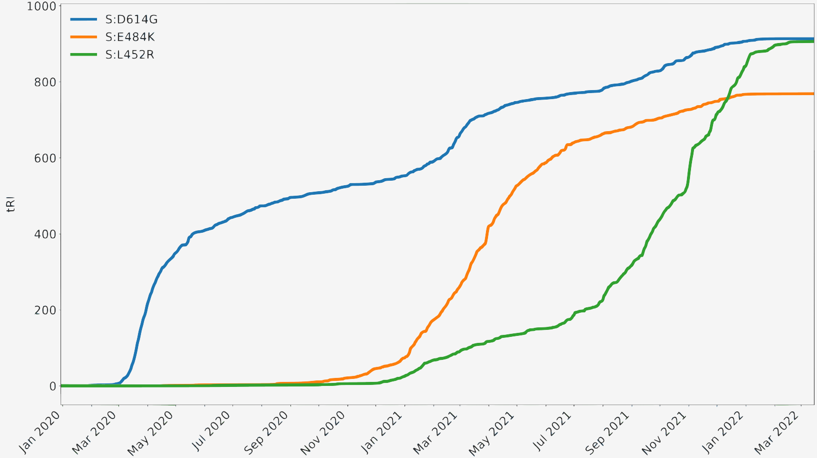
Multipersistence through Rips Transformations

calculates pathwise persistence from

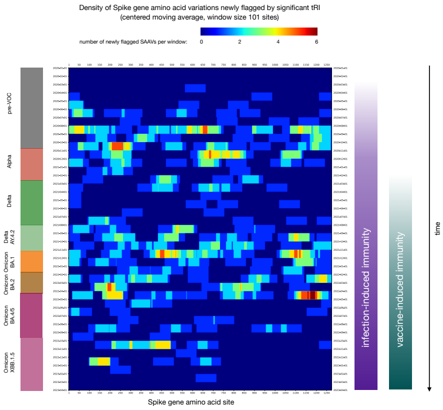
distance matrix + additional filtration



EvotRec.py – Evolution of topological Recurrence



Dynamic Fitness Landscape and Epistasis


Density of Spike gene amino acid variations newly flagged by significant tRI
(centered moving average, window size 101 sites)

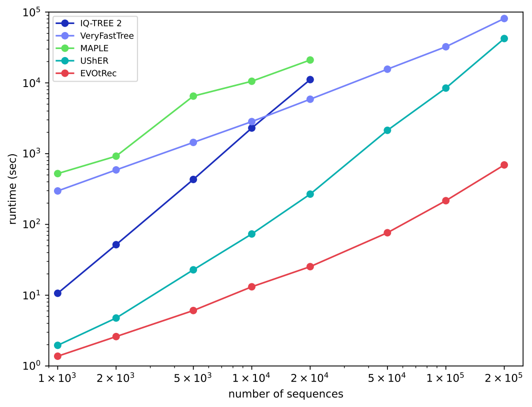
time-resolved tRI activity along the genome shows surprising amount of time-dependence.

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

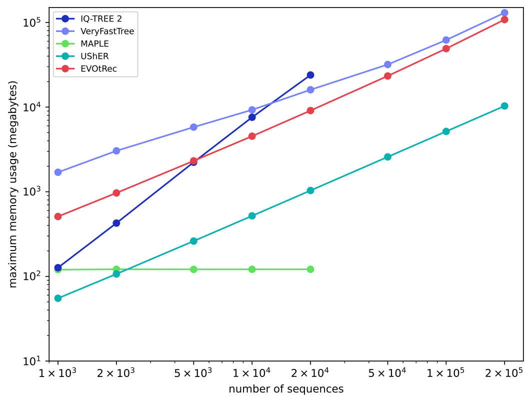
This is possible because SNV-cycles are *localized* in a particular genetic background.

Computational Benchmarks

Runtime

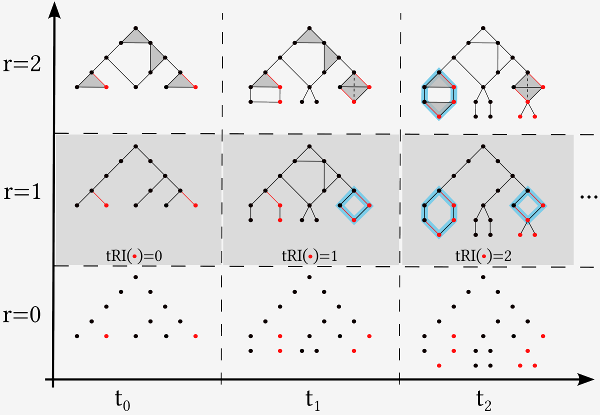


Memory



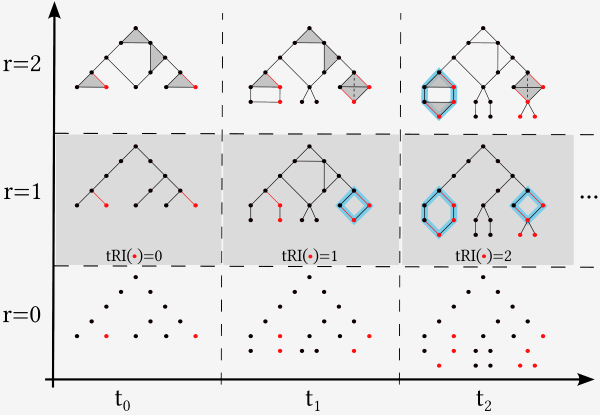
Summary

- Persistent homology measures evolutionarily 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 evolutionarily 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!