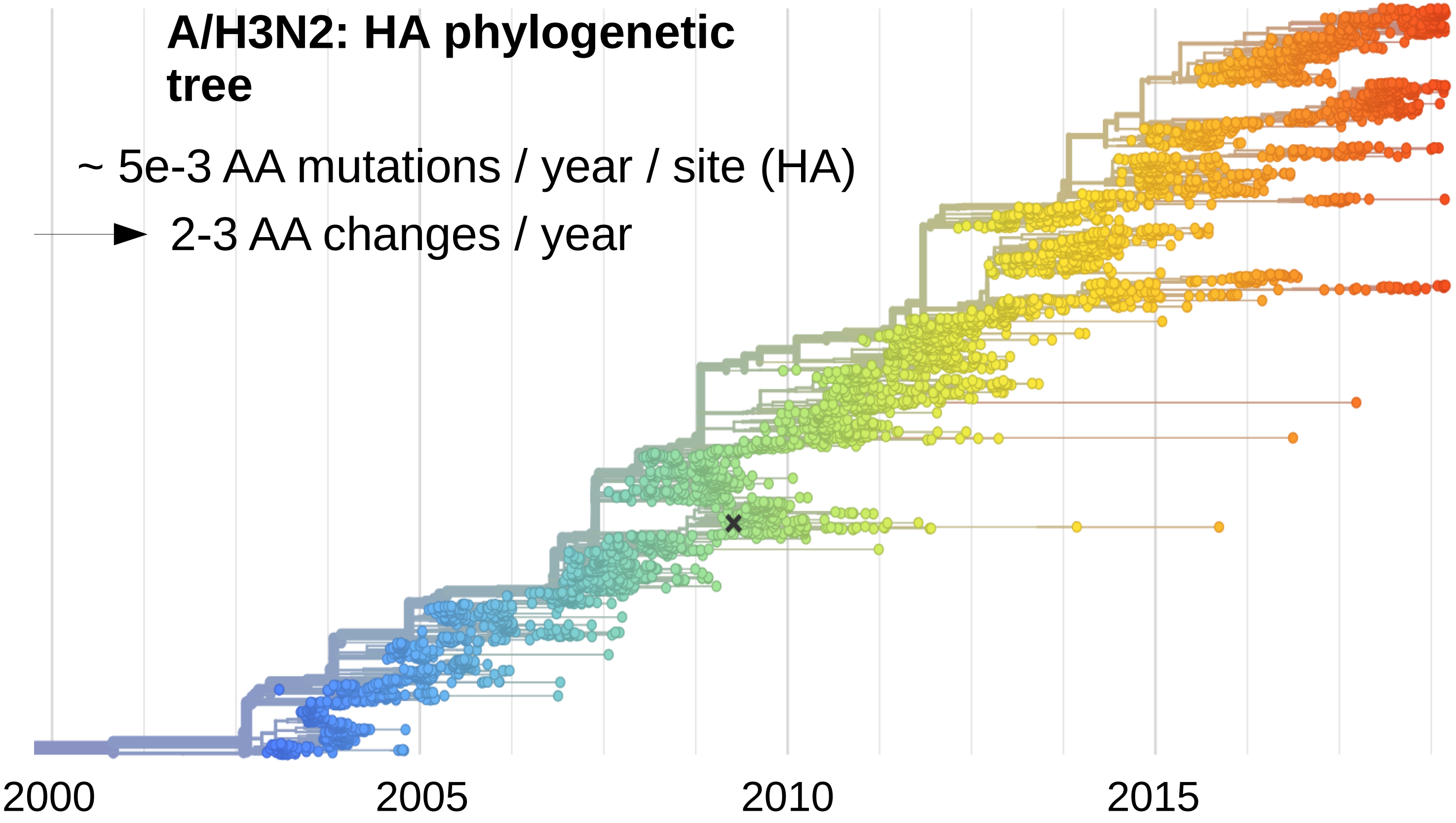
Seasonal influenza viruses:
Limited predictability of evolution
& Inference of reassortment networks

Pierre Barrat-Charlaix
Biozentrum, University of Basel

Human seasonal influenza virus

~ hundreds of million cases per year → 5-10 % of humans

In constant evolution (especially surface proteins HA & NA)


A/H3N2: HA phylogenetic tree

~ 5e-3 AA mutations / year / site (HA)

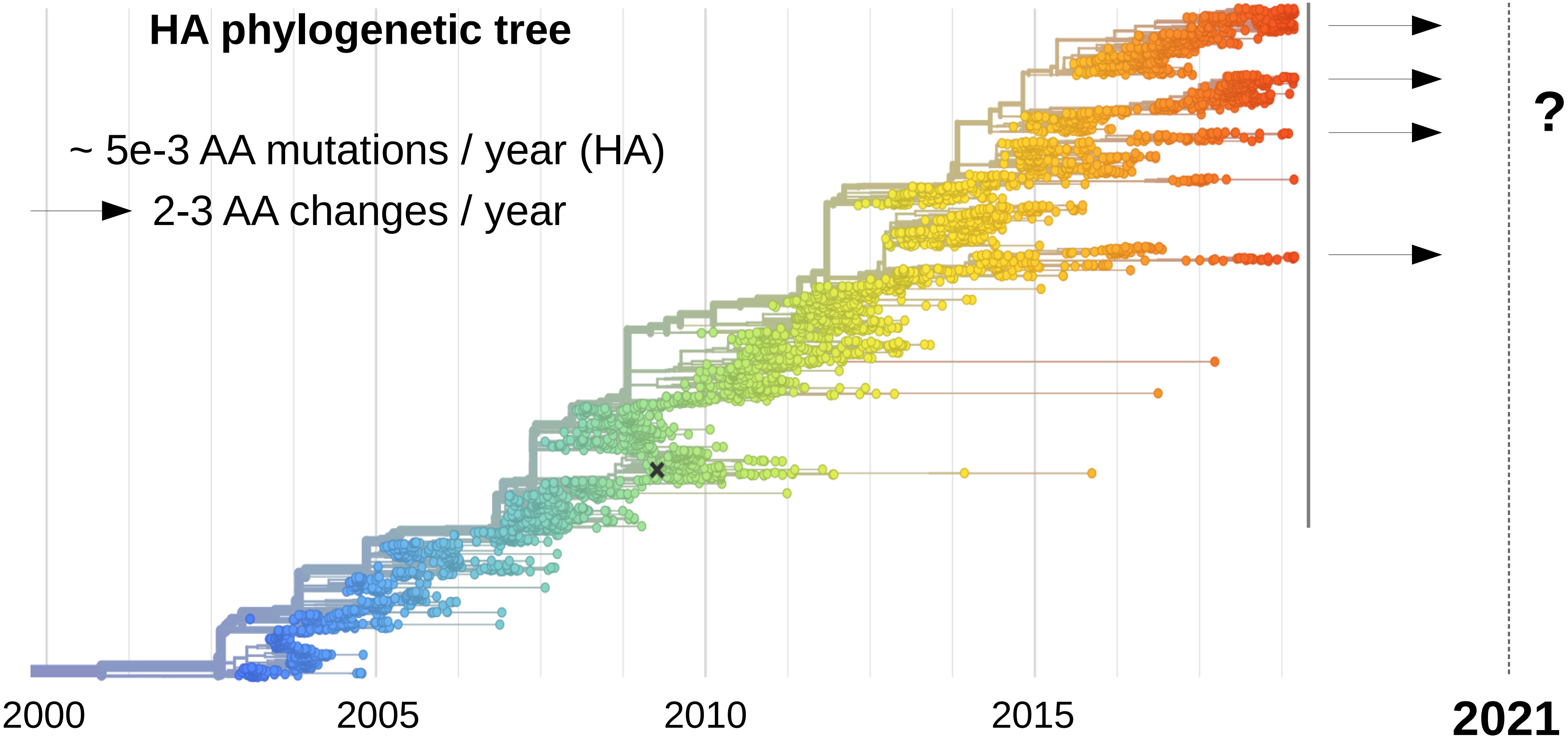
→ 2-3 AA changes / year

Variability in the present population

Human seasonal influenza virus

Can we understand/predict its evolution?

Which present clade will take over ?



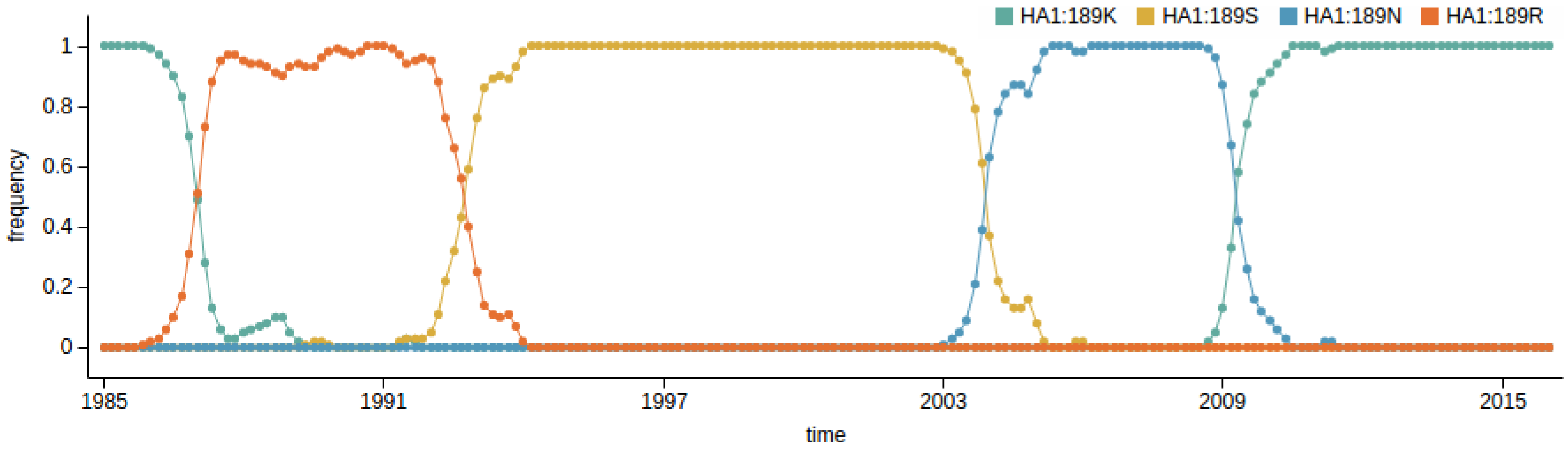
HA phylogenetic tree

~ 5e-3 AA mutations / year (HA)

► 2-3 AA changes / year

Selection in viral proteins

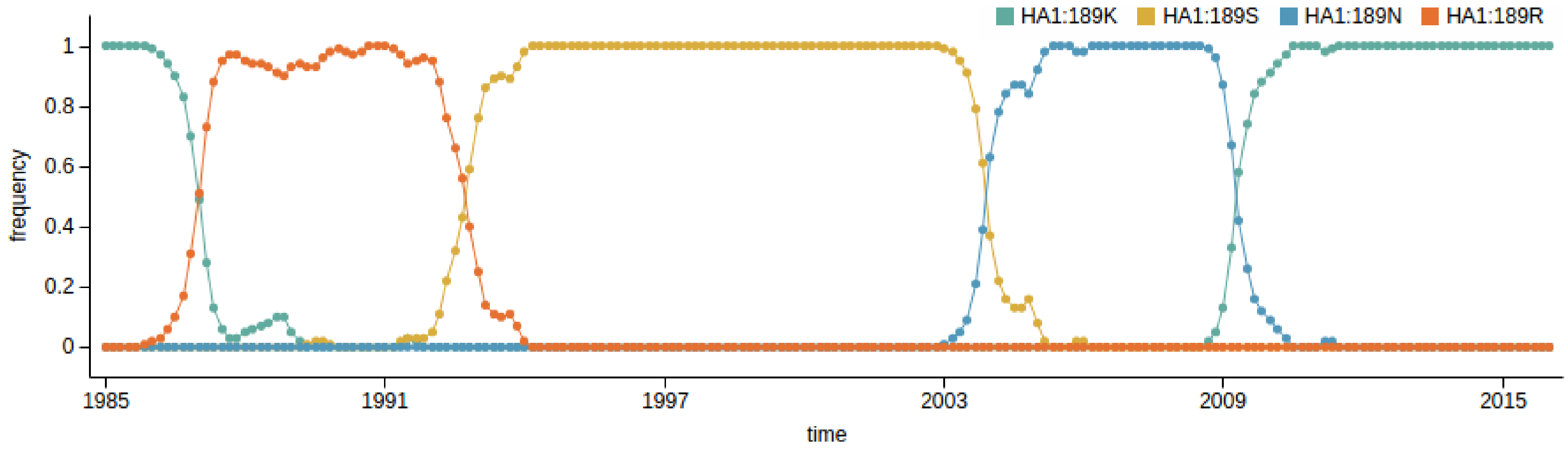
Frequency of amino acid mutations at position HA1:189



Selective pressure to avoid **human immunity** → Adaptive mutations with a **fitness advantage**

Selection in viral proteins

Frequency of amino acid mutations at position HA1:189

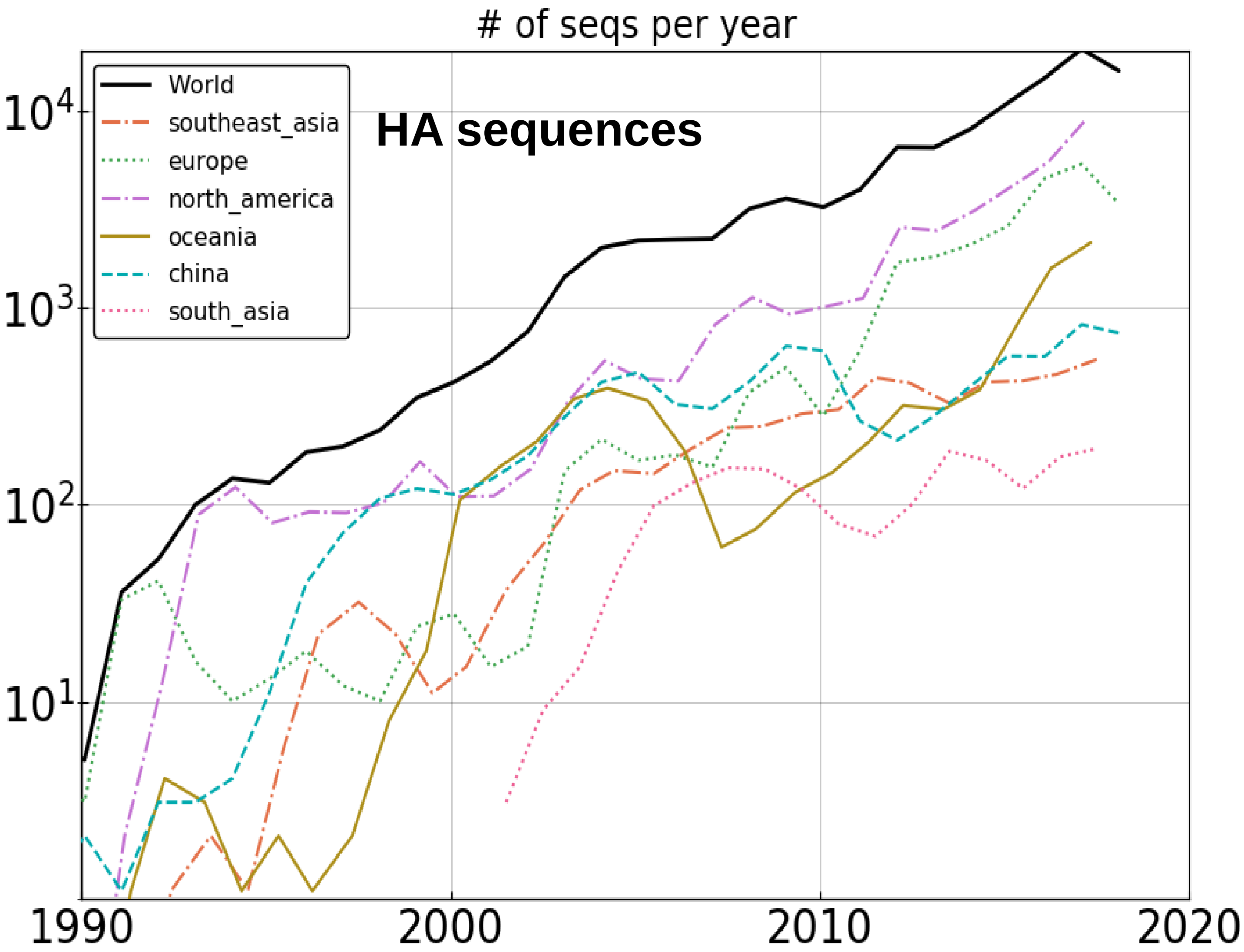


Selective pressure to avoid **human immunity** → Adaptive mutations with a **fitness advantage**

Questions

- Is this **way of viewing the data** correct ?
- Can we use it to **predict future evolution** ?

Influenza pandemic: a retrospective view



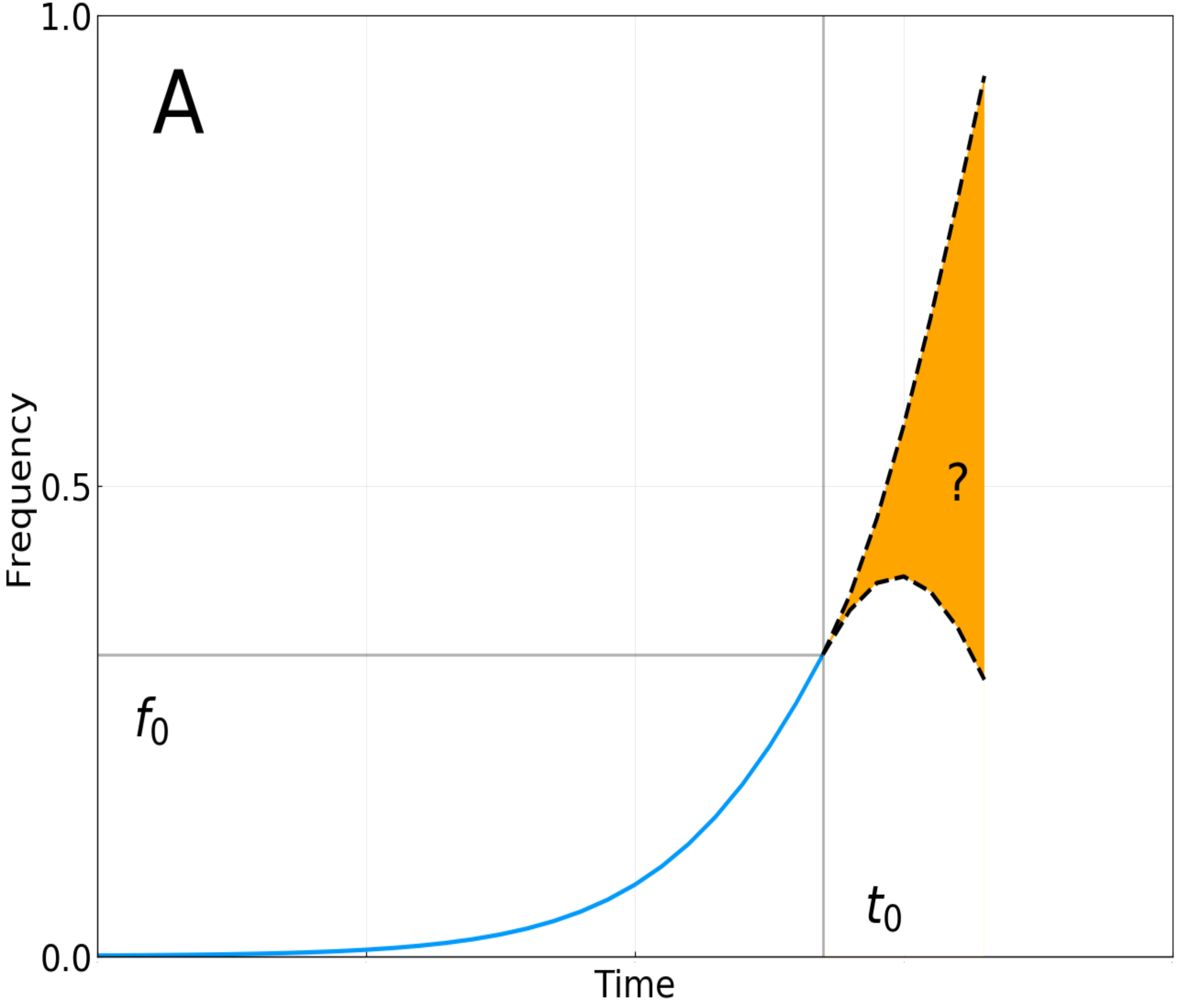
What can we learn from these past sequences?

Time-binning of the past sequences by 1 month intervals

→ **Snapshots of the population**

→ **Frequency trajectories**

Short term prediction



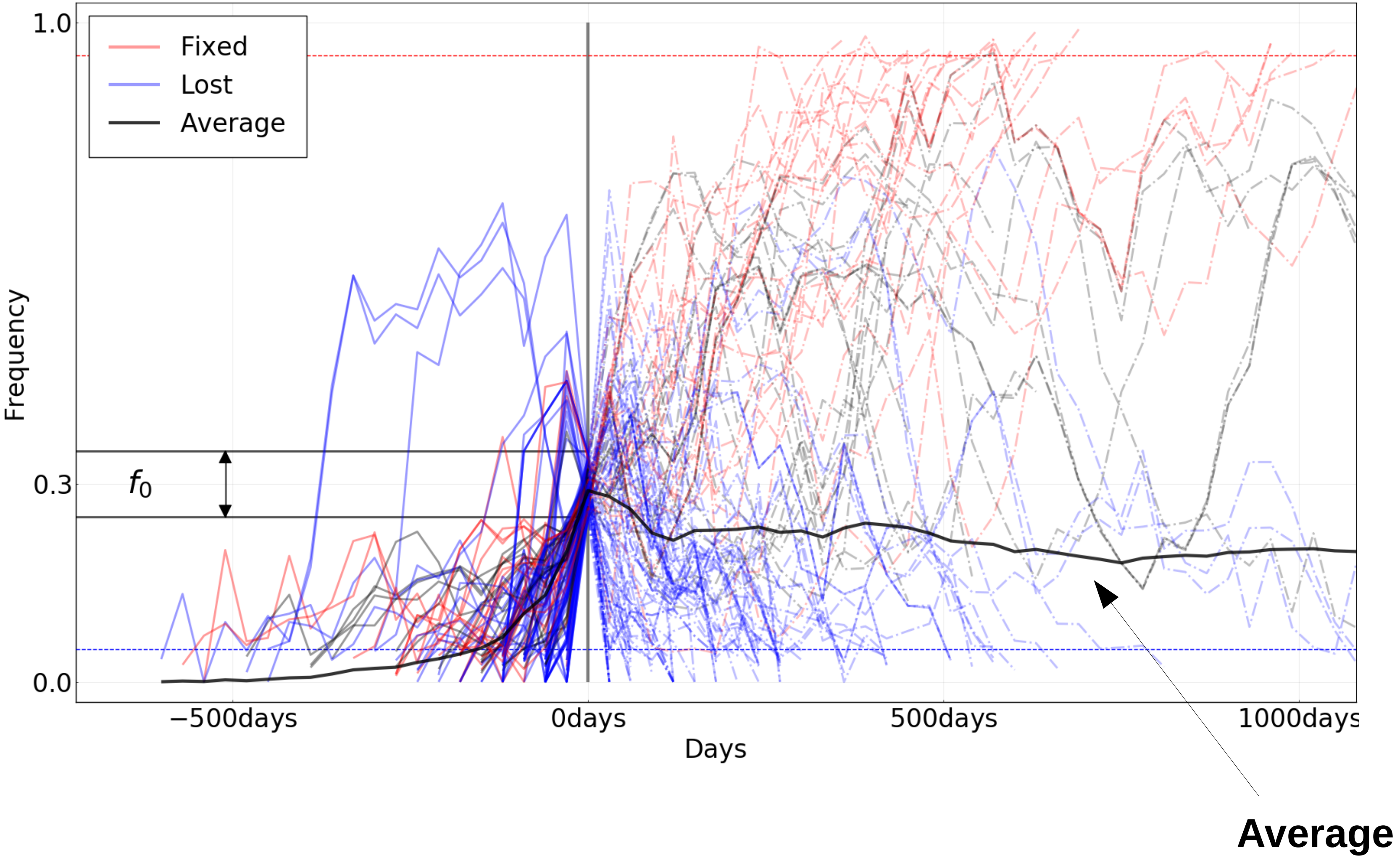
Frequency trajectories of amino acid mutations

Frequency distribution at t0+dt?

Statistics from 460 **rising** frequency trajectories from year 2000

Short term prediction

Influenza h3n2, HA protein

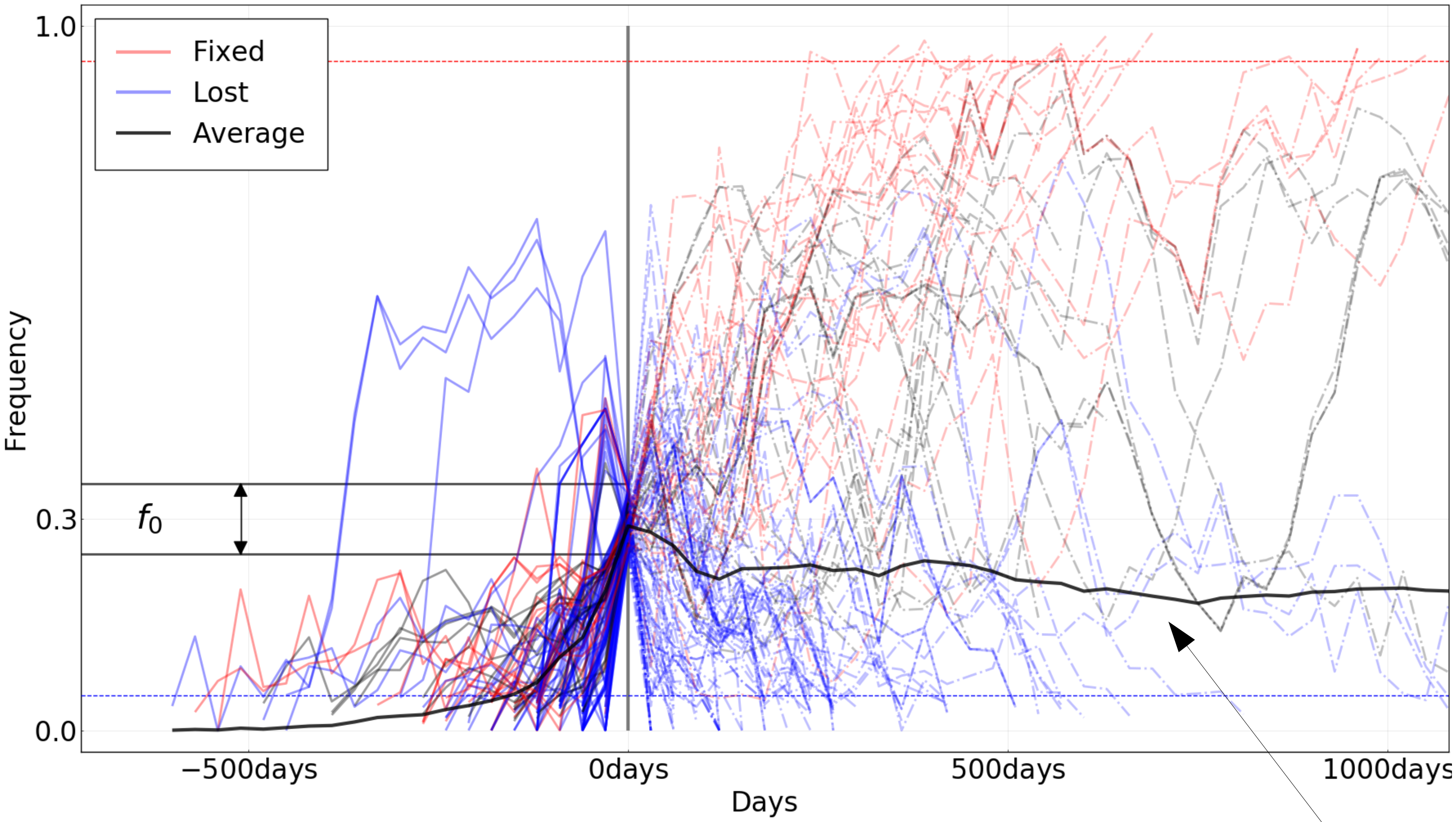


Mutations:

- Absent in the past
- Seen around f0=30%

Short term prediction

Influenza h3n2, HA protein



Mutations:

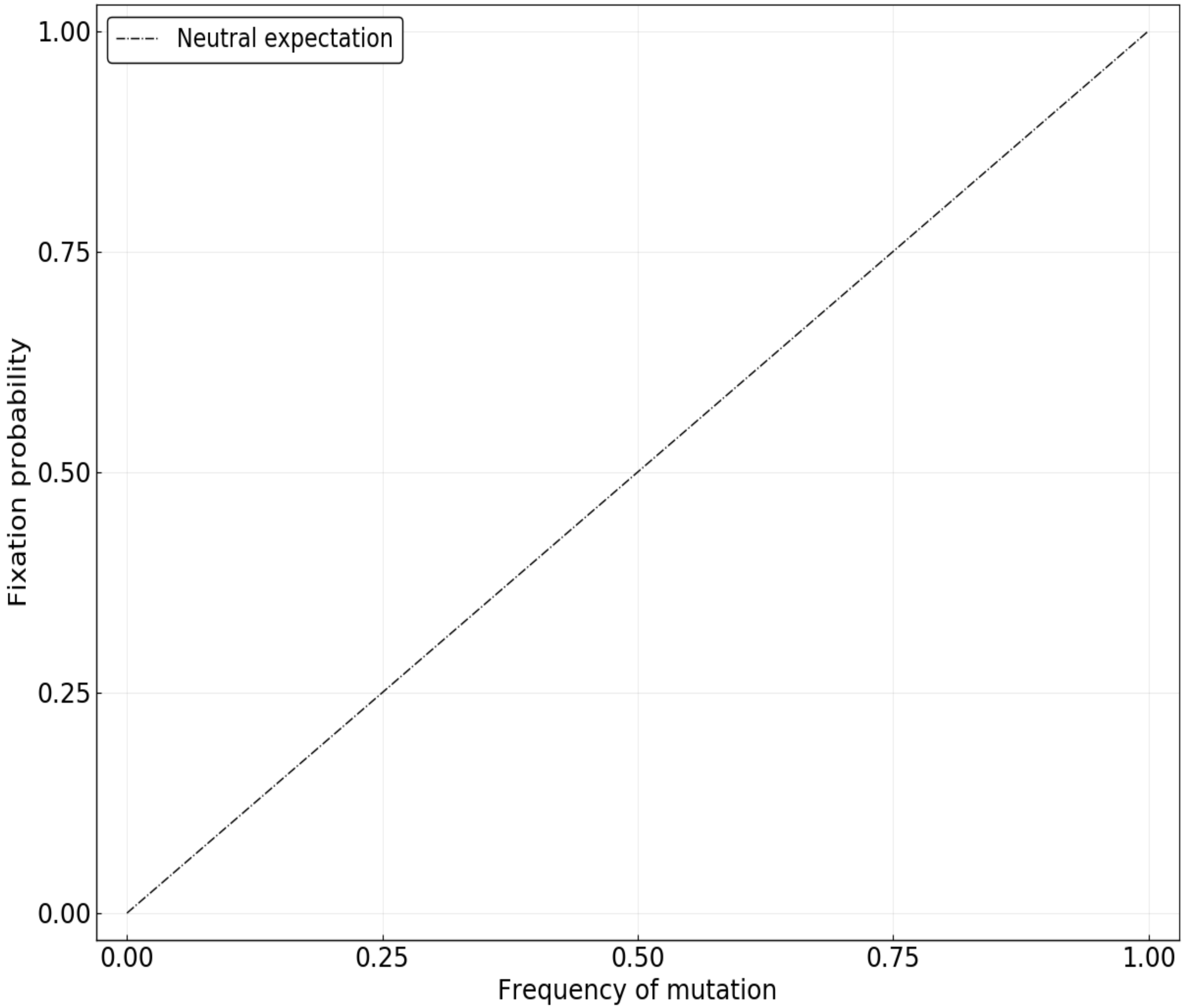
- Absent in the past
- Seen around f0=30%

Average

Average trajectory is **flat** after the conditioning

→ **No inertia**

Fixation probability



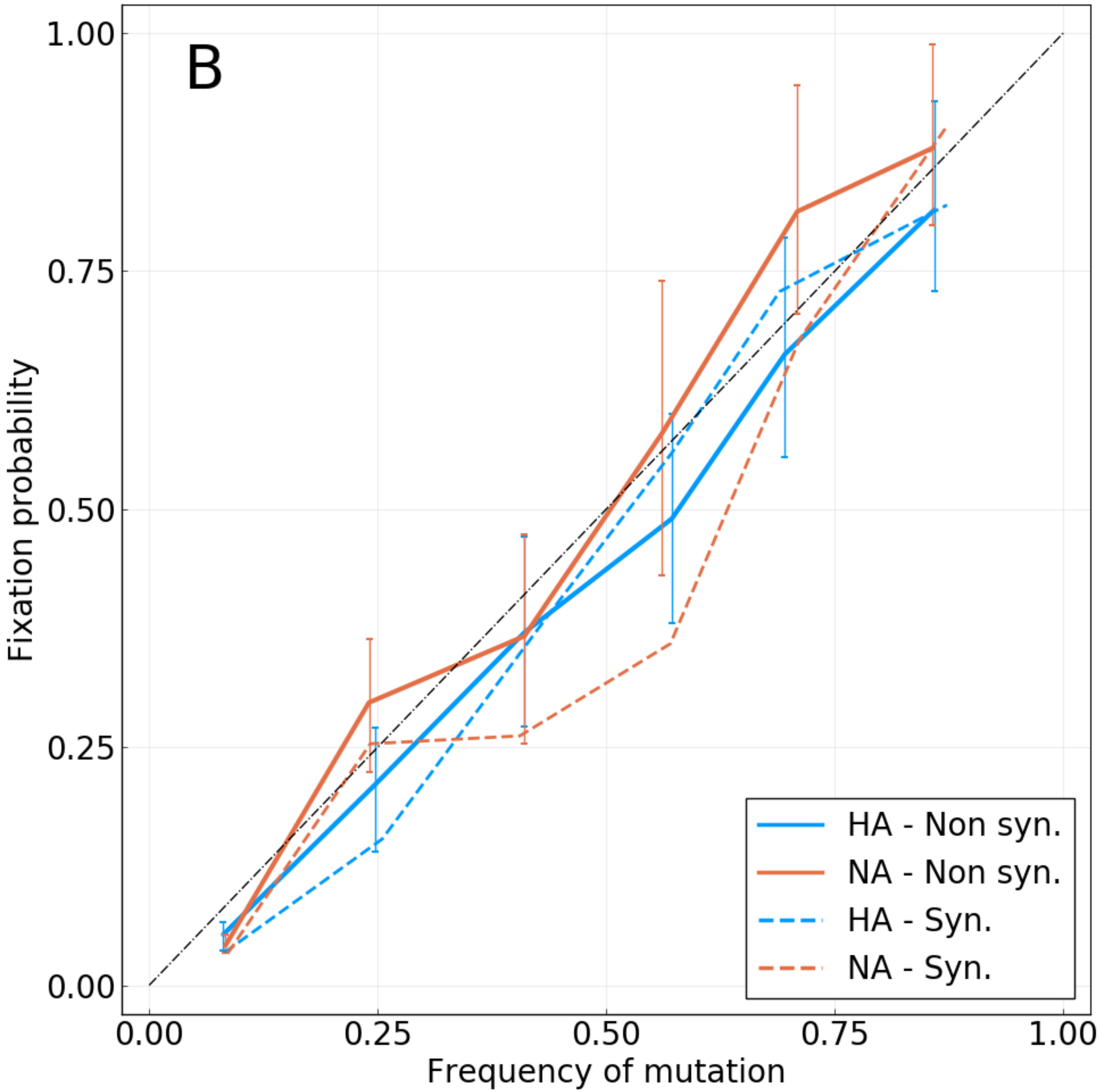
Neutral evolution, *e.g.* Wright-Fisher model

No selective advantage

↓

Probability of fixation is equal to frequency in the population

Fixation probability

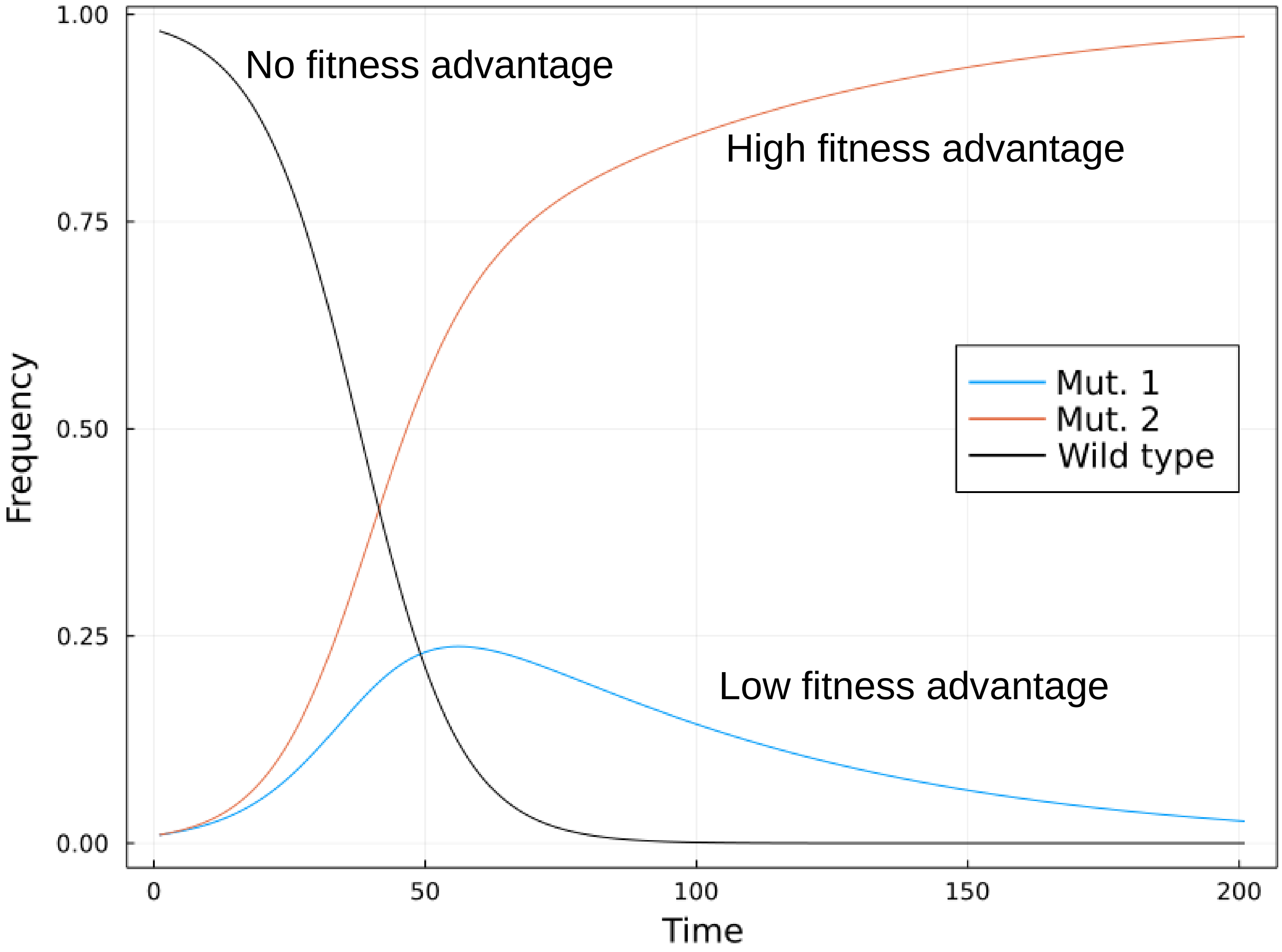


For rising trajectories

No signs of selection !

The **rise in frequency** of a mutation does **not** inform us about its future **fixation**

Is this expected? Clonal interference

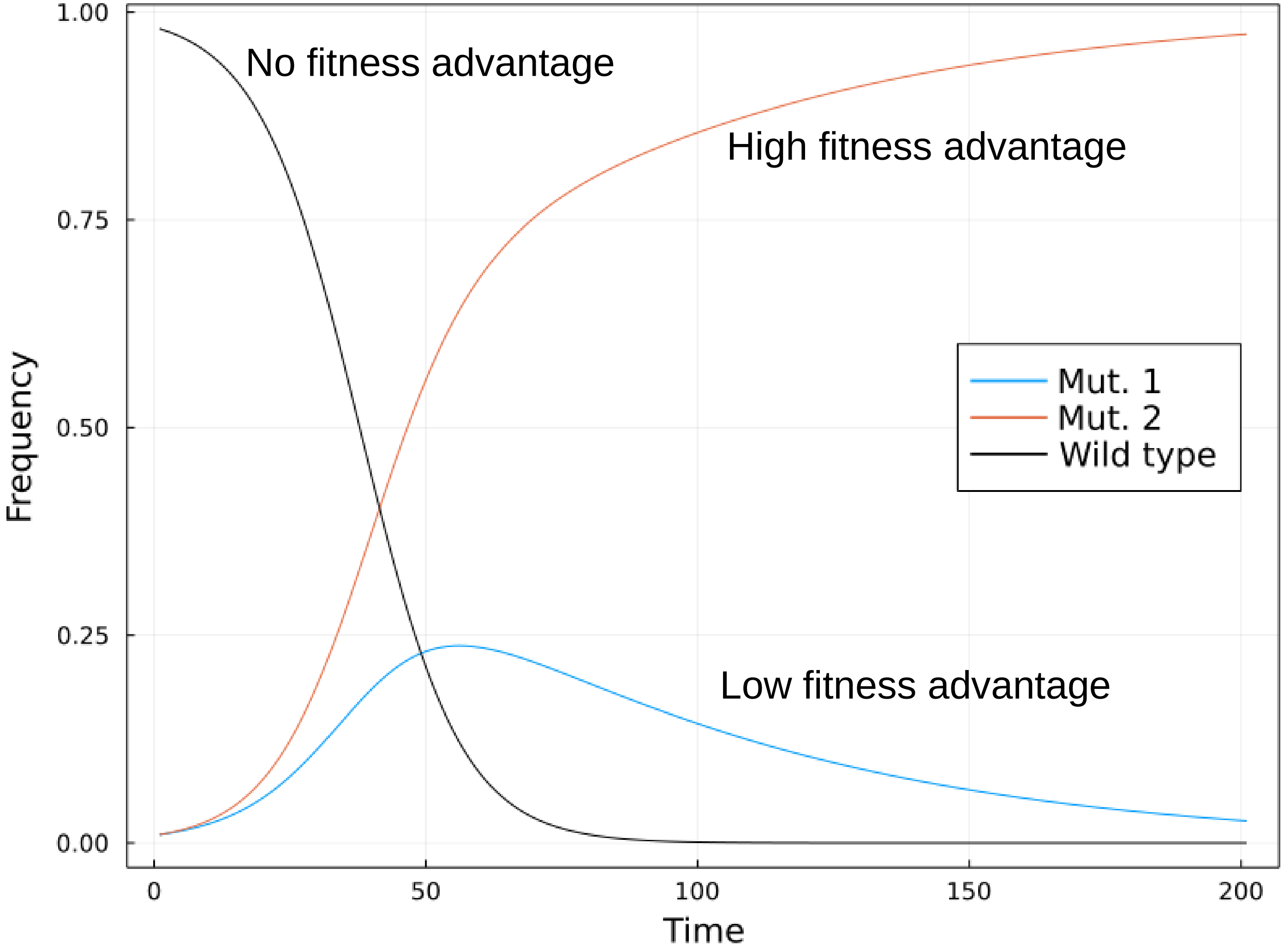


Adaptive mutations appearing on different individuals

↓

Competition

Is this expected? Clonal interference



Adaptive mutations appearing on different individuals

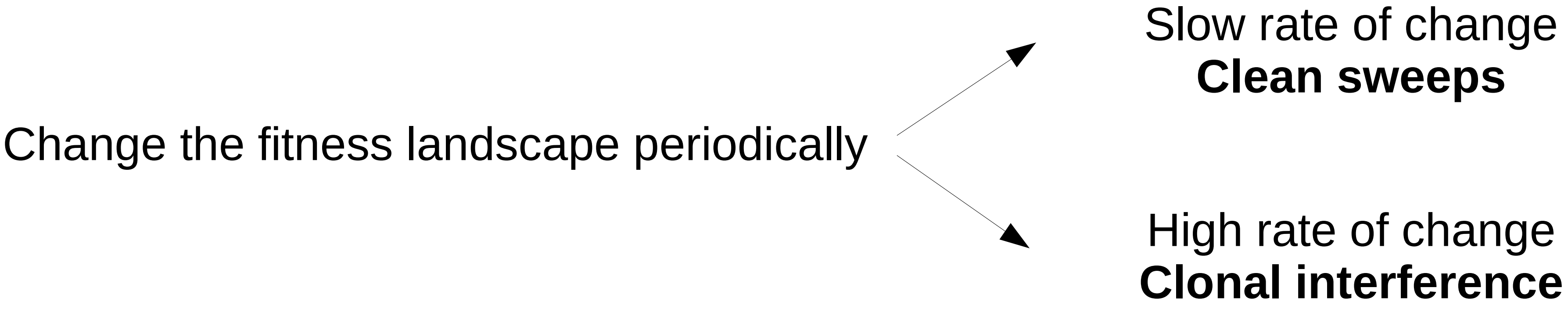
↓

Competition

→ **Simulation of toy model with clonal interference**

Genetic linkage: toy model

Simple fitness lanscape $$f(\vec{s}) = \sum_{i=1}^{L} h_i s_i$$



Change the fitness landscape periodically

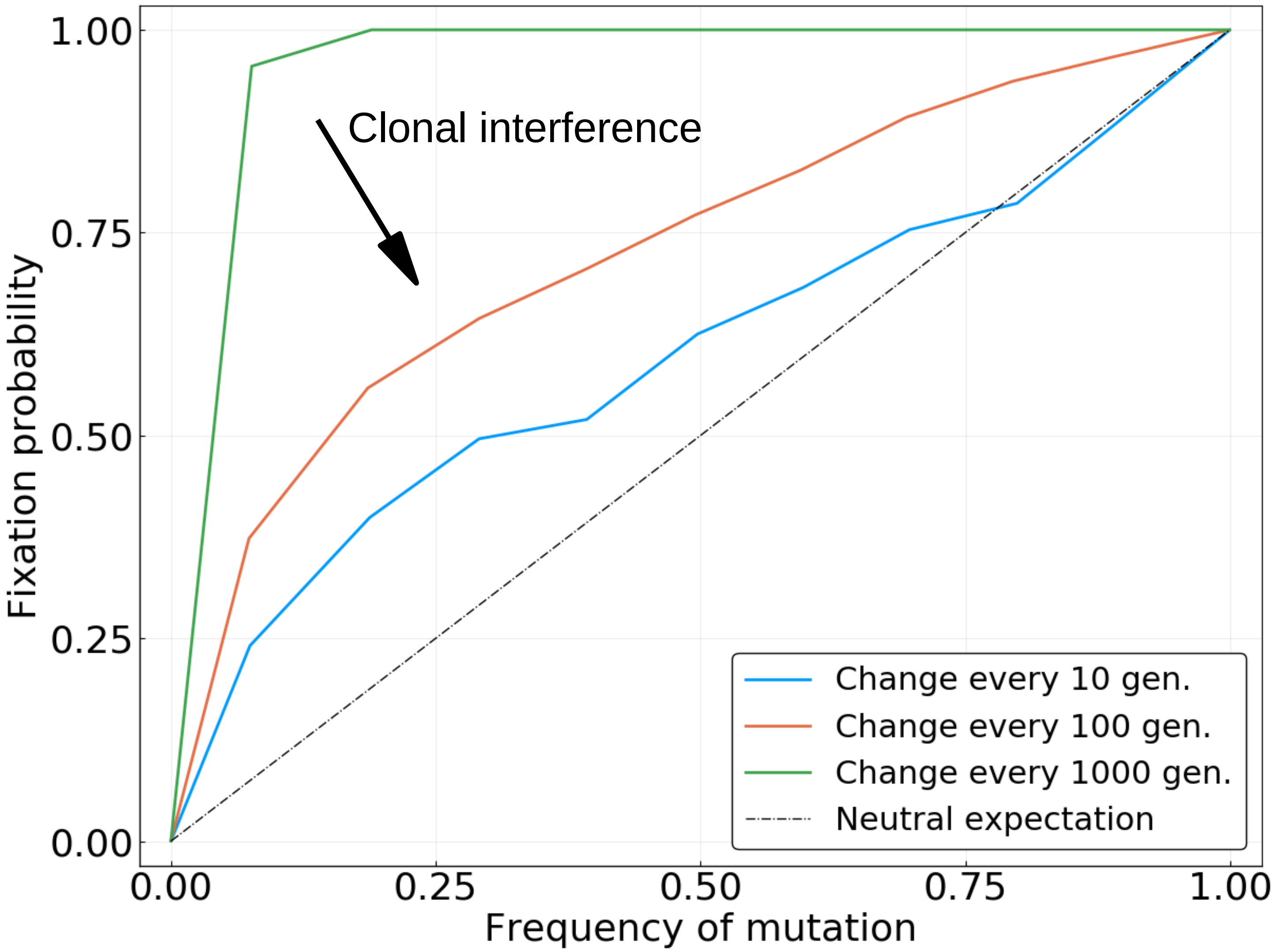
Slow rate of change
Clean sweeps

High rate of change
Clonal interference

Genetic linkage: toy model

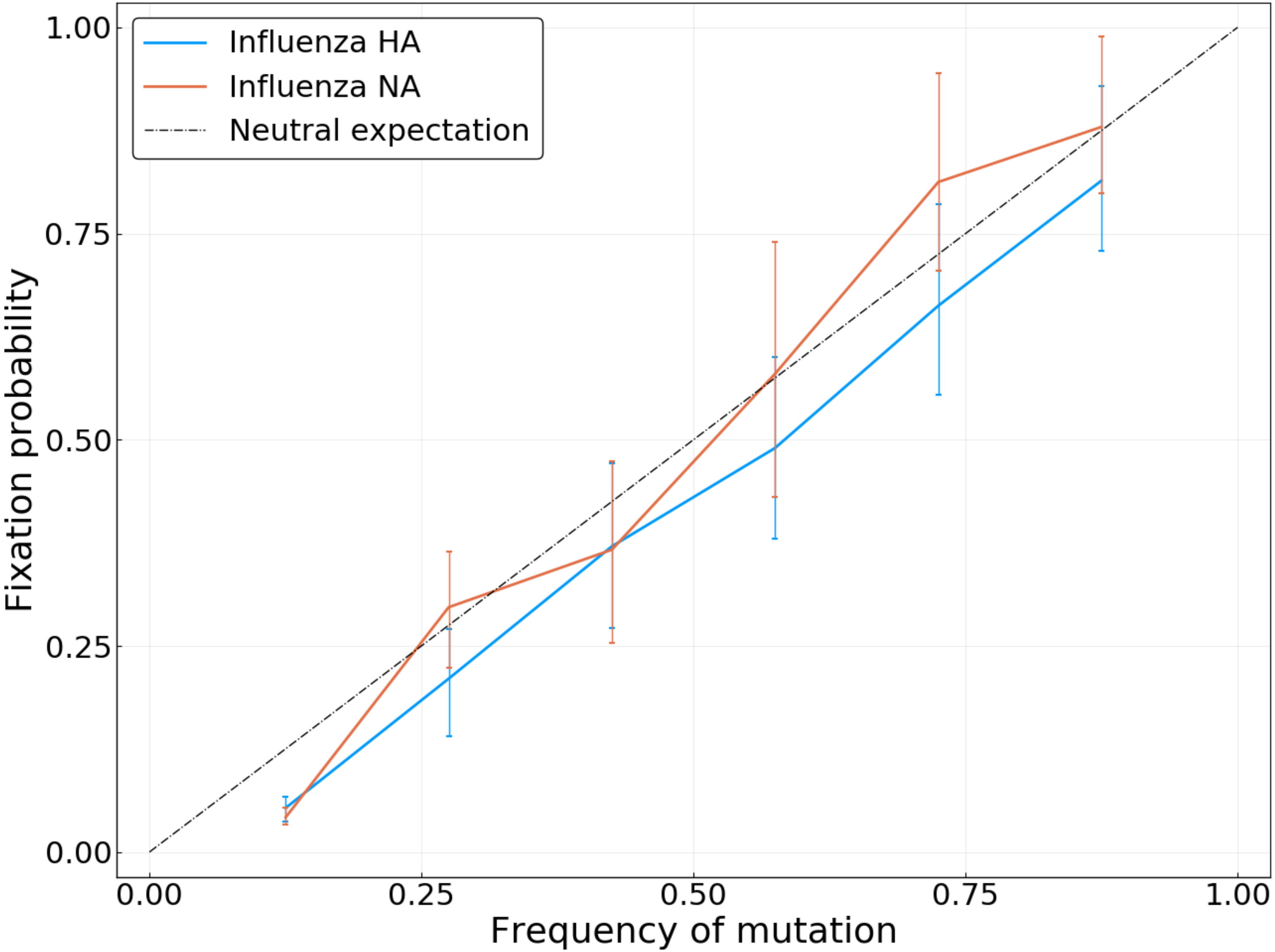
Sweep time ~400 generations

(vs ~3 years for flu)



It’s hard to mimic neutrality!

Fixation probability



Predictors of fixation?

Predictors of fixation?

Epitope positions

- Targeted by human immune system
- Expected to be under strong selection
- Used in models of selection in influenza

But often ascertained *post-hoc*

Shih *et. al.* 2007
Koel *et. al.* 2013
Luksza & Lässig 2014

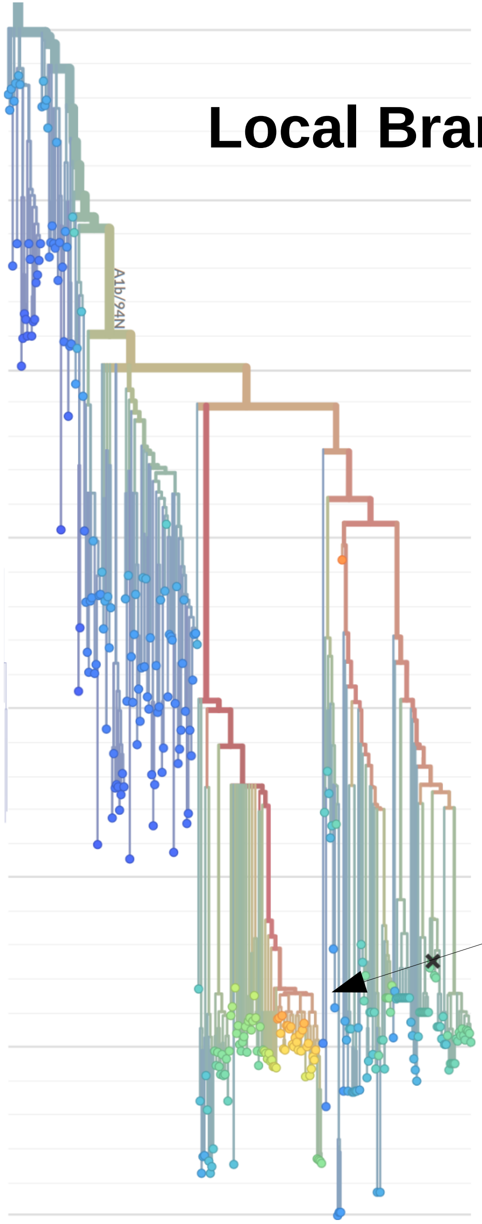
Predictors of fixation?

Epitope positions

- Targeted by human immune system
- Expected to be under strong selection
- Used in models of selection in influenza

But often ascertained *post-hoc*

Shih *et. al.* 2007
Koel *et. al.* 2013
Luksza & Lässig 2014



Local Branching Index (LBI)

Strains with high fitness have more offsprings

↓

Dense branching in the tree indicates high fitness

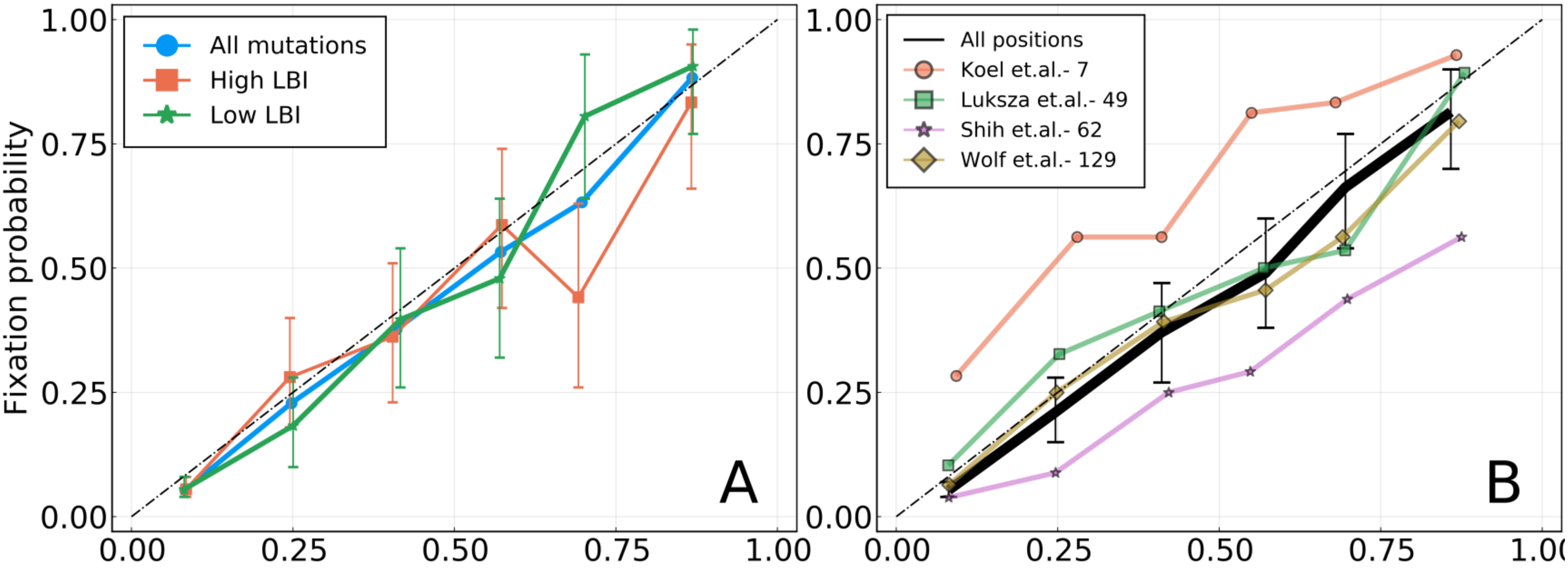
Dense branching

Neher et. al. 2014

Fixation probability: for specific mutations?

Local Branching Index (LBI)
Measure of fitness

Epitope positions
Potential targets of ABs



→ **Current models do not predict fixation!**

Summary

A/H3N2 influenza is under adaptive selection, but …

- Predictibility of frequency trajectories is low
- Fixation probability is equal to present frequency

→ « Apparent neutrality »

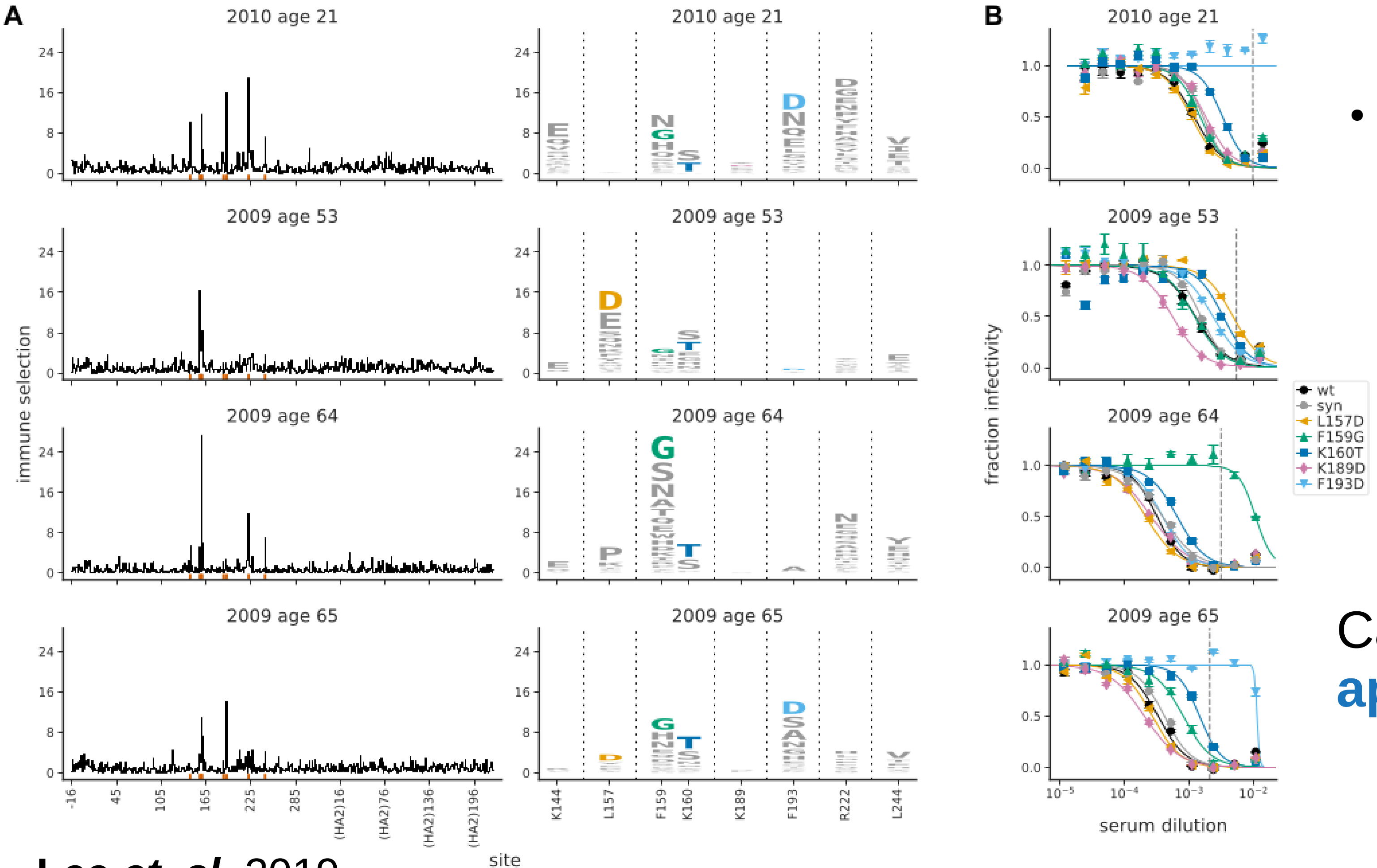
- Hard to find predictors of fixation / fitness
  - ➔ LBI
  - ➔ epitopes

→ **Influenza does not behave like models suggest !**

Open question: what could explain these results?

Open question: what could explain these results?

Diversity in human immune response



- Adaptive mutations only allow escape to a fraction of the population
- Fitness advantage expires before fixation

Can this result in **apparent neutrality** ?

Lee *et. al.* 2019

Epidemiological considerations

Influenza is a seasonal virus

⟶ In temperate regions : exponential increase (winter) followed by bottleneck

Geographical structure

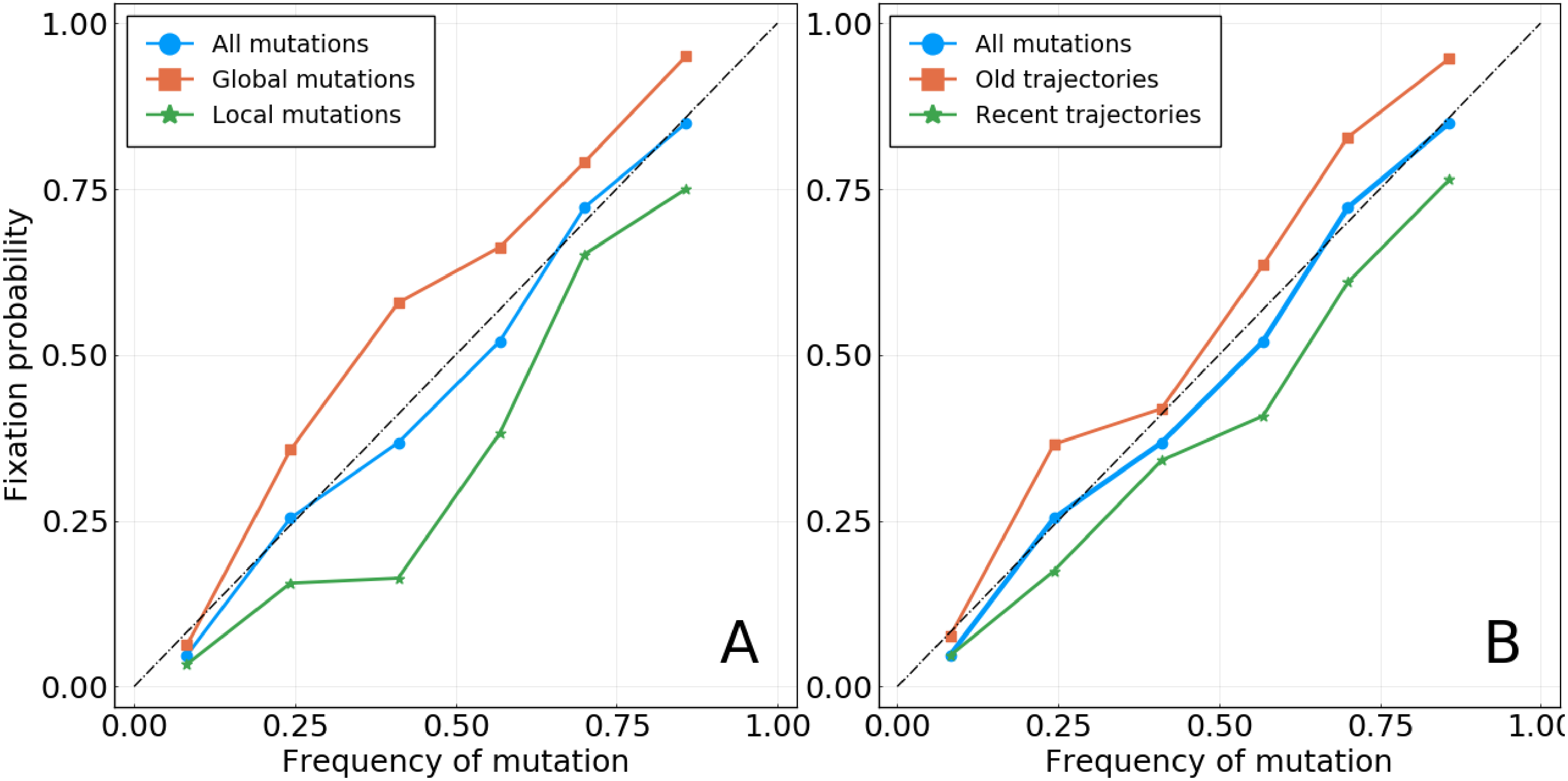
⟶ Frequency of variants varies in different regions

⟶ **What does this mean for frequency trajectories ?**

To be investigated...

Epidemiological considerations

For A/H3N2 - HA

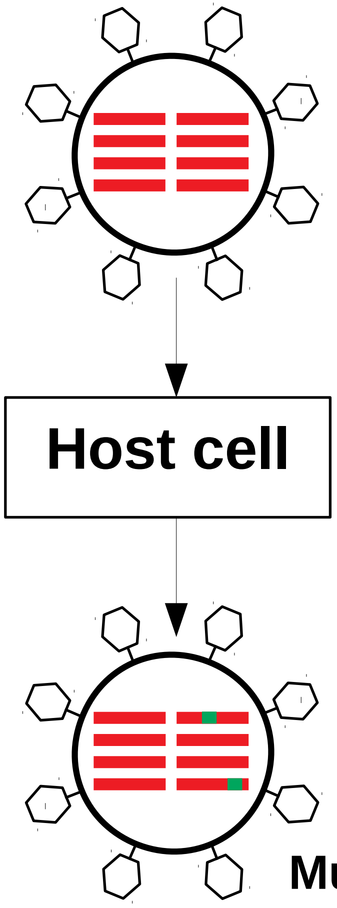


Seasonal influenza viruses:
Limited predictability of evolution
& Inference of reassortment networks

Pierre Barrat-Charlaix
Biozentrum, University of Basel

Evolution of influenza: Mutations and reassortment

Mutation



Flu virus

🟥 Gene : 8 in total

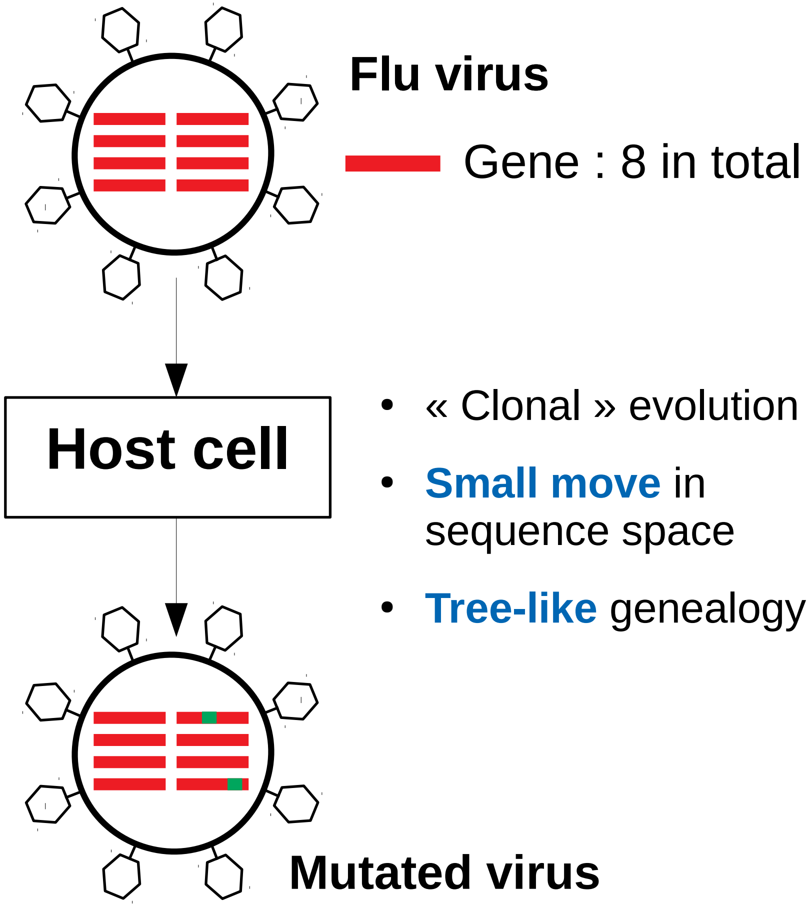
Host cell

Mutated virus

- « Clonal » evolution
- **Small move** in sequence space
- **Tree-like** genealogy

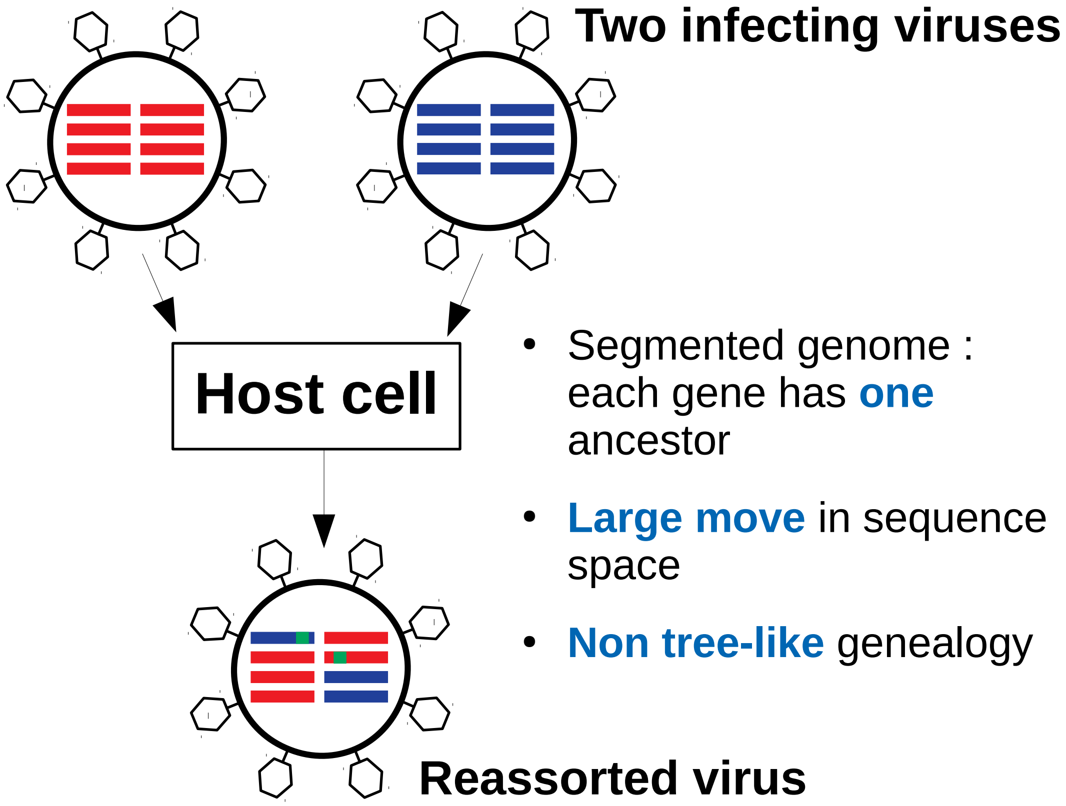
Evolution of influenza: Mutations and reassortment

Mutation



- « Clonal » evolution
- **Small move** in sequence space
- **Tree-like** genealogy

Reassortment



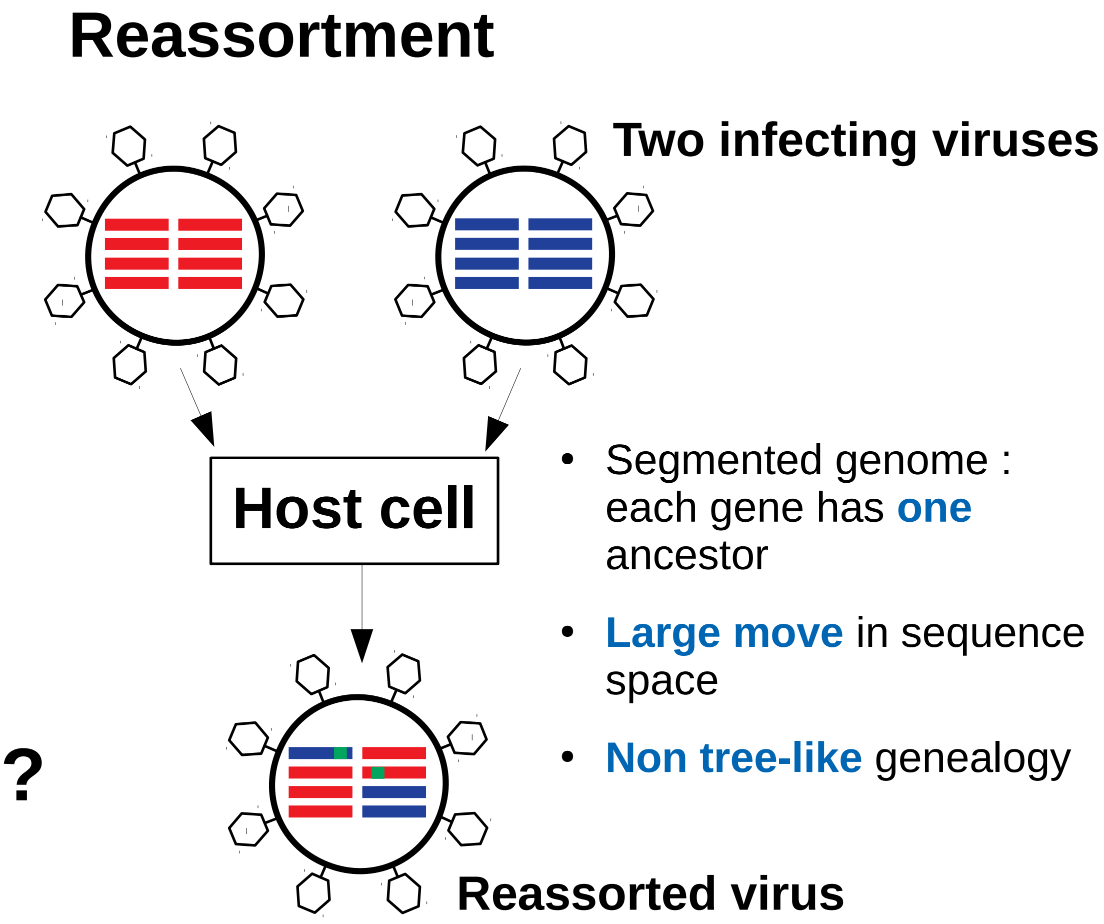
- Segmented genome : each gene has **one** ancestor
- **Large move** in sequence space
- **Non tree-like** genealogy

Reassortment in influenza

- Combines strains from **different subtypes**, or from **human/animal** hosts.
- Origin of many **pandemics**
  - Asian flu – 1957
  - Hong Kong flu – 1968
  - H1N1 pandemic – 2009
- Also happens at “smaller” scale: within a subtype.

- How often does it happen?
- Contribution to immune escape and adaptation?

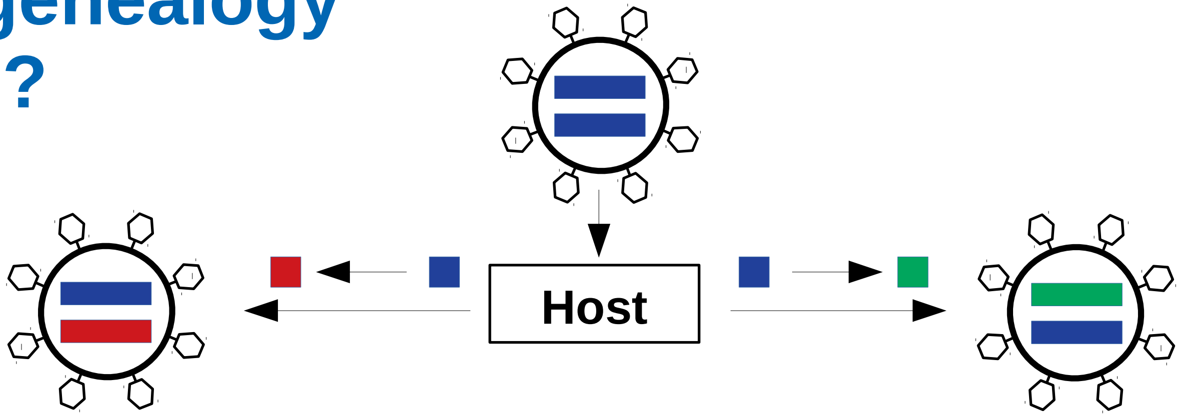
→ **?**



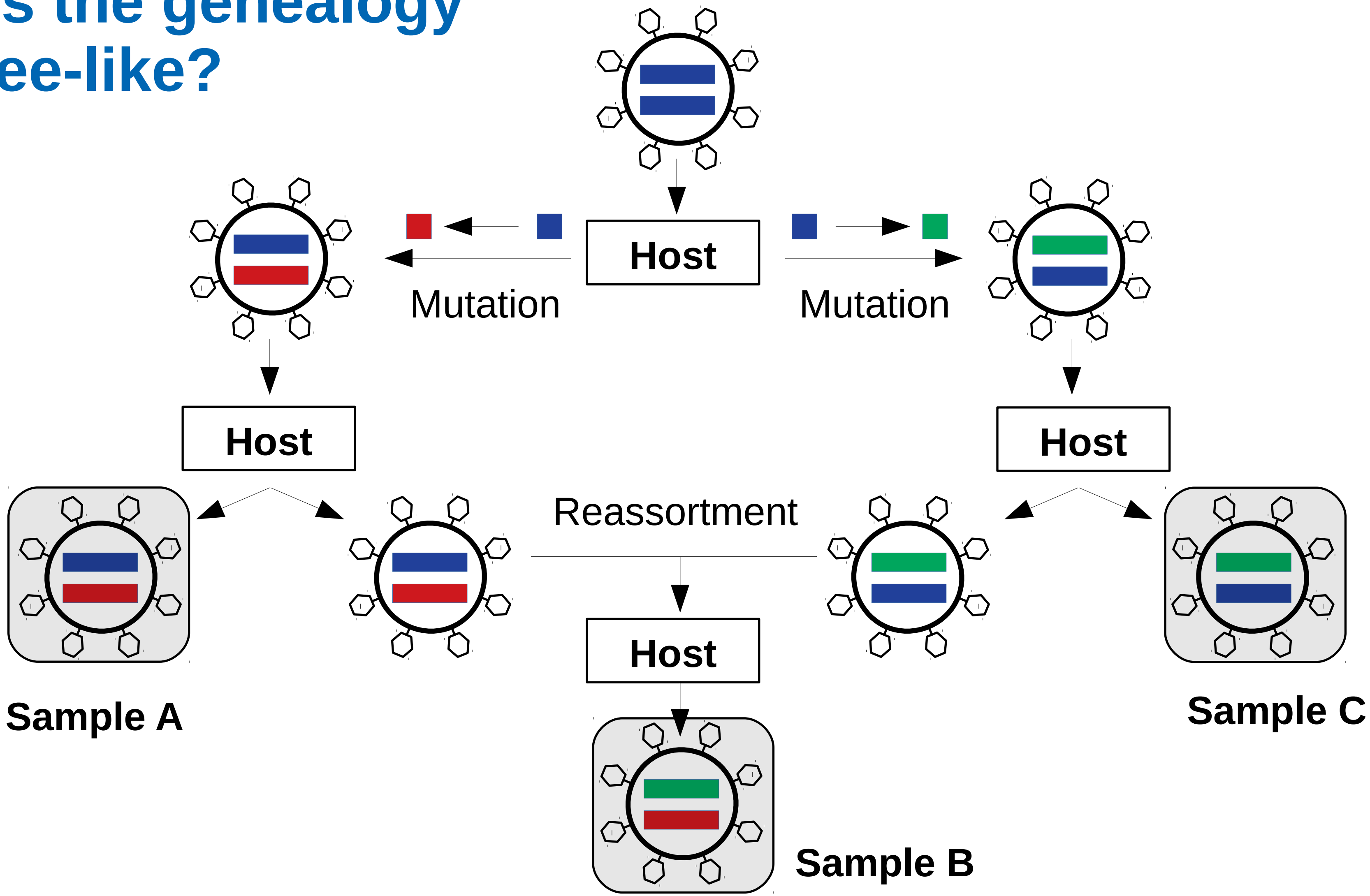
- Segmented genome : each gene has **one** ancestor
- **Large move** in sequence space
- **Non tree-like** genealogy

Reassortments are hard to infer from sequences!

Why is the genealogy not tree-like?



Why is the genealogy not tree-like?



Ancestral Reassortment Graph

Observed sequences

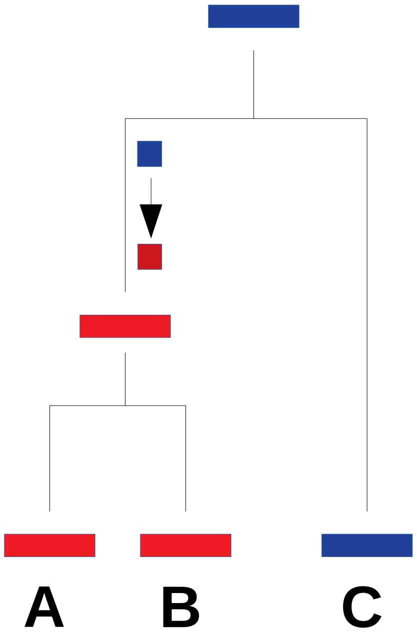
A	🟥	🟦
B	🟥	🟩
C	🟦	🟩

↓

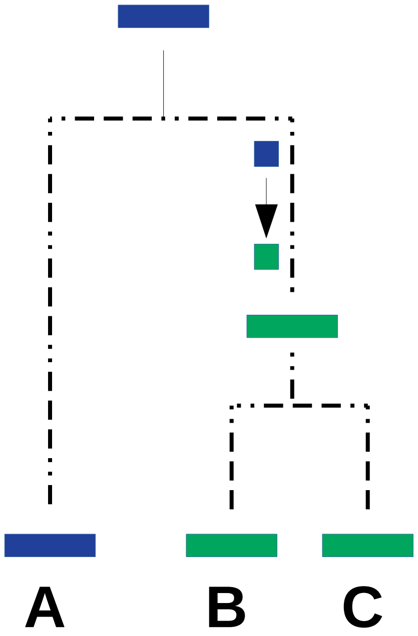
Reconstructed segment trees

↓

Topological differences

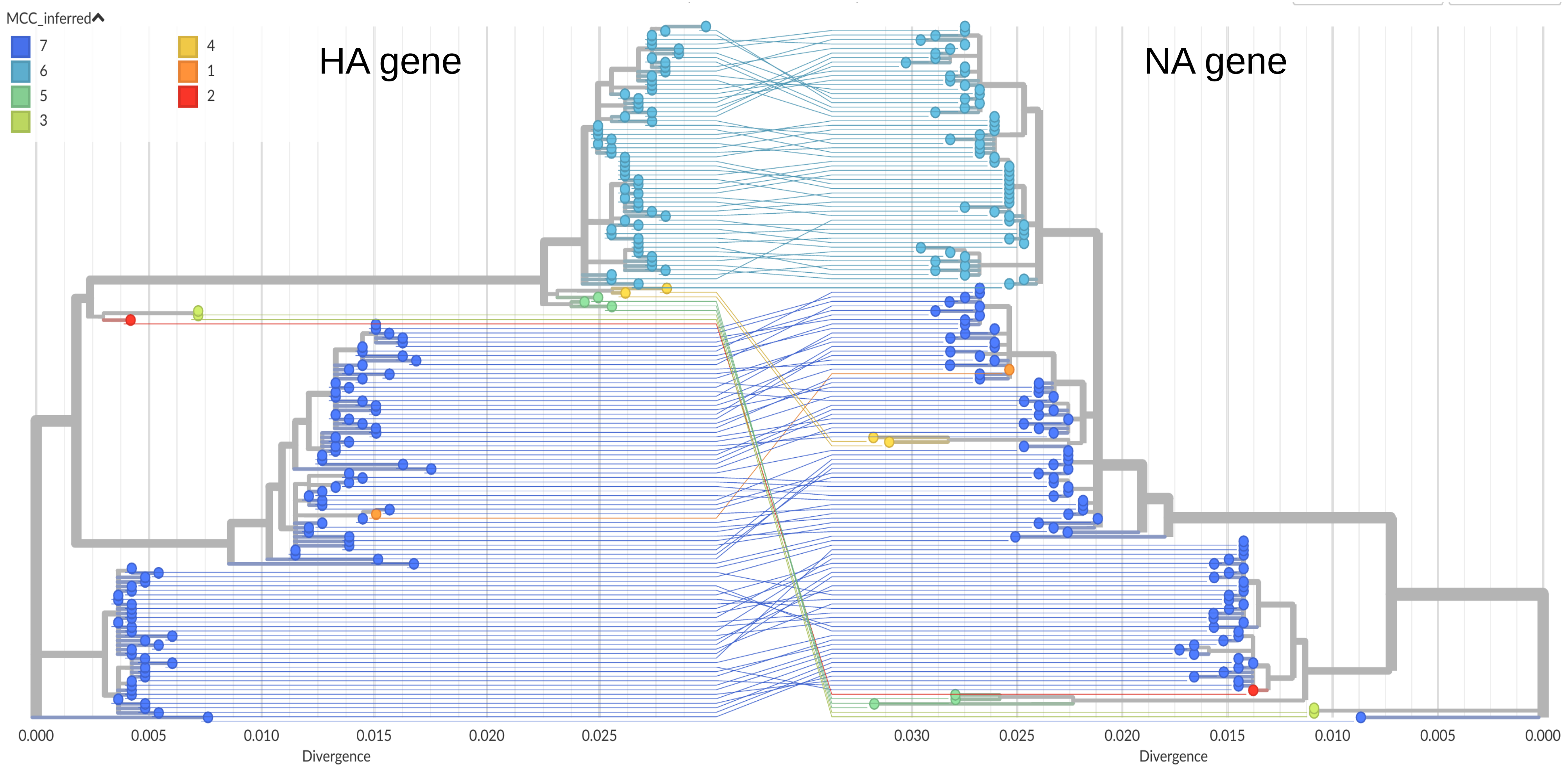


Genealogy of first gene

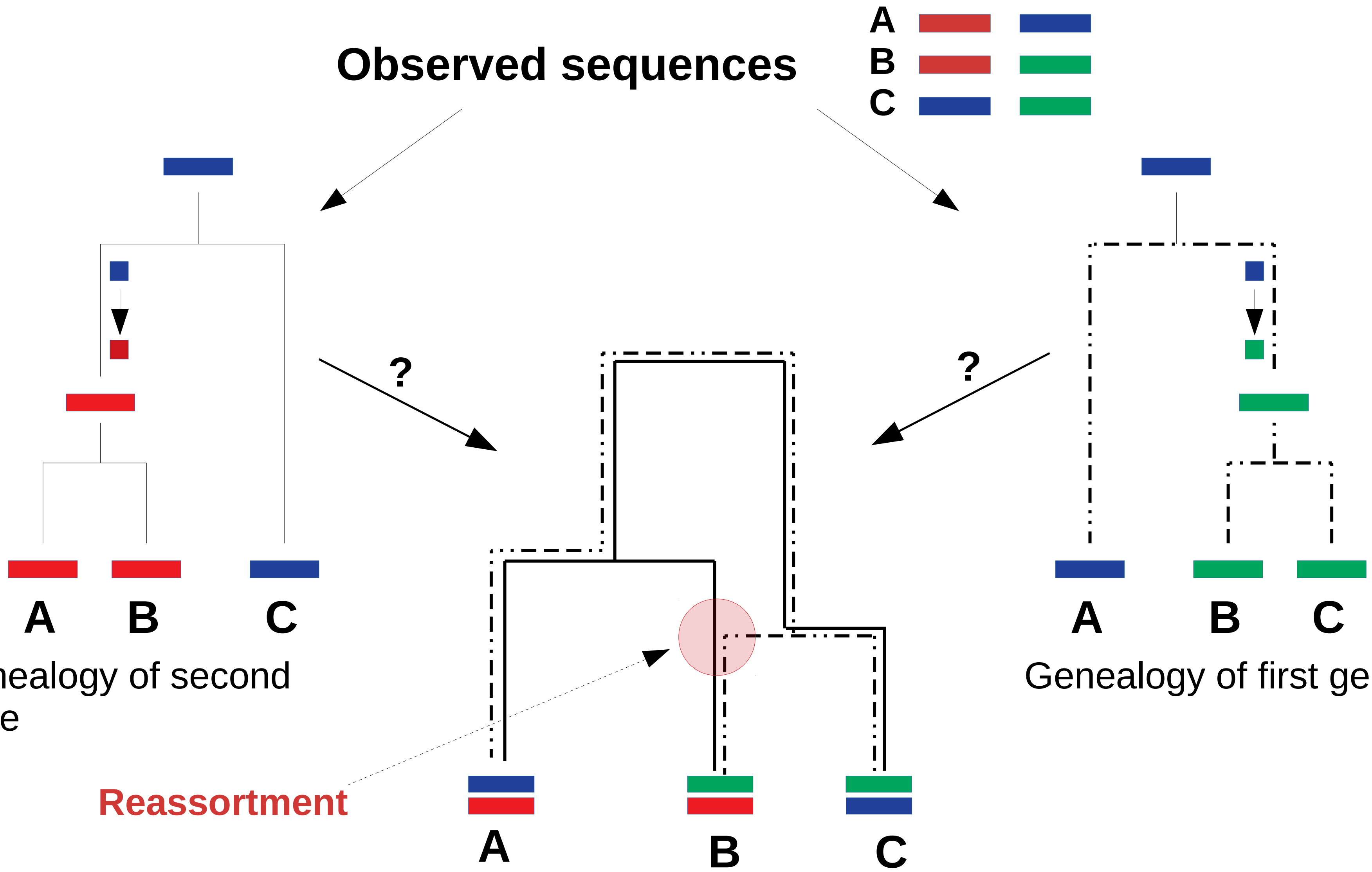


Genealogy of second gene

Example of flu trees



Why is the genealogy not tree-like?



Observed sequences

Genealogy of second gene

Genealogy of first gene

Reassortment

Ancestral Recombination Graph (ARG)

Inferring reassortments / Reconstructing the ARG

Existing methods

- Manual inspection of trees
  (*e.g.* **[Holmes et. al. 2005], [Boni et. al. 2010]**)
- Methods based on genetic distance **[Rabadan et. al. 2008]**
- Trees + mutation methods [**Villa & Lässig 2017]**
- Tree topology based methods **[Nagarajan & Kingsford 2011]**

→ Finds a subset of reassortment events

- Maximum likelihood methods **[Müller et. al. 2020]** → Accurate but slow

→ **No "reference" method**

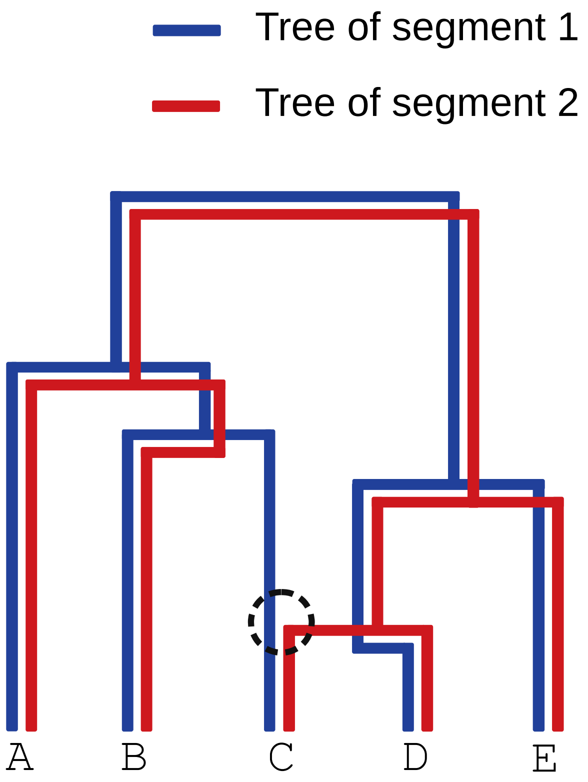
We want something that is

- **Fast** : can be easily applied to new sequences
- Finds **all reassortments**, and not only large obvious ones
- Works for the 2-genes case (simplicity)

Inferring the ARG: the Treeknit method

We want something that is

- **Fast** : can be easily applied to new sequences
- Finds **all reassortments**, and not only large obvious ones
- Works for the 2-genes case (simplicity)

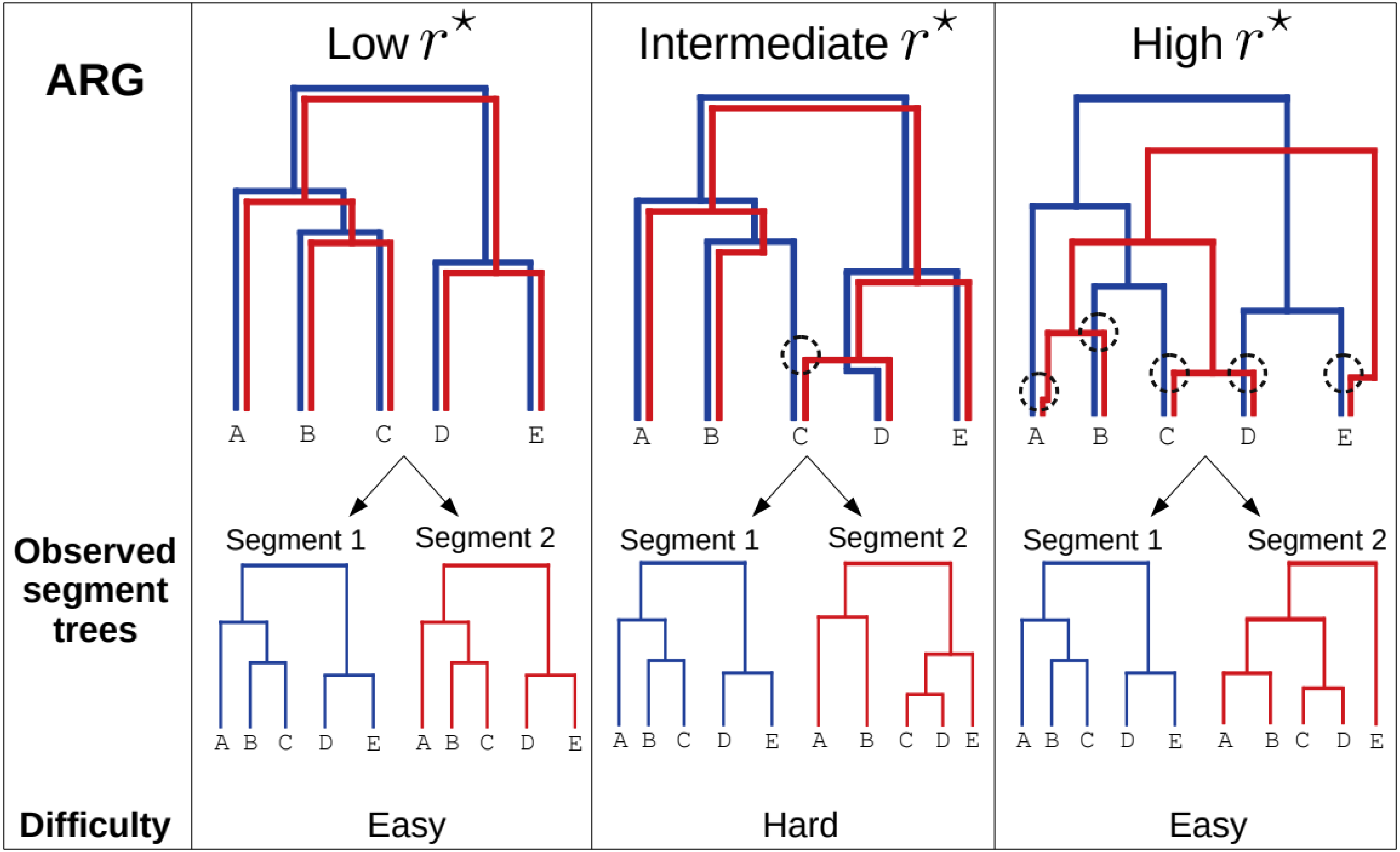


Main idea :

- The ARG is a **collage of gene trees**
- We can **infer each tree** from sequences (iqtree, RaxML, …)
- **Topological differences** between these trees are **due to reassortment**

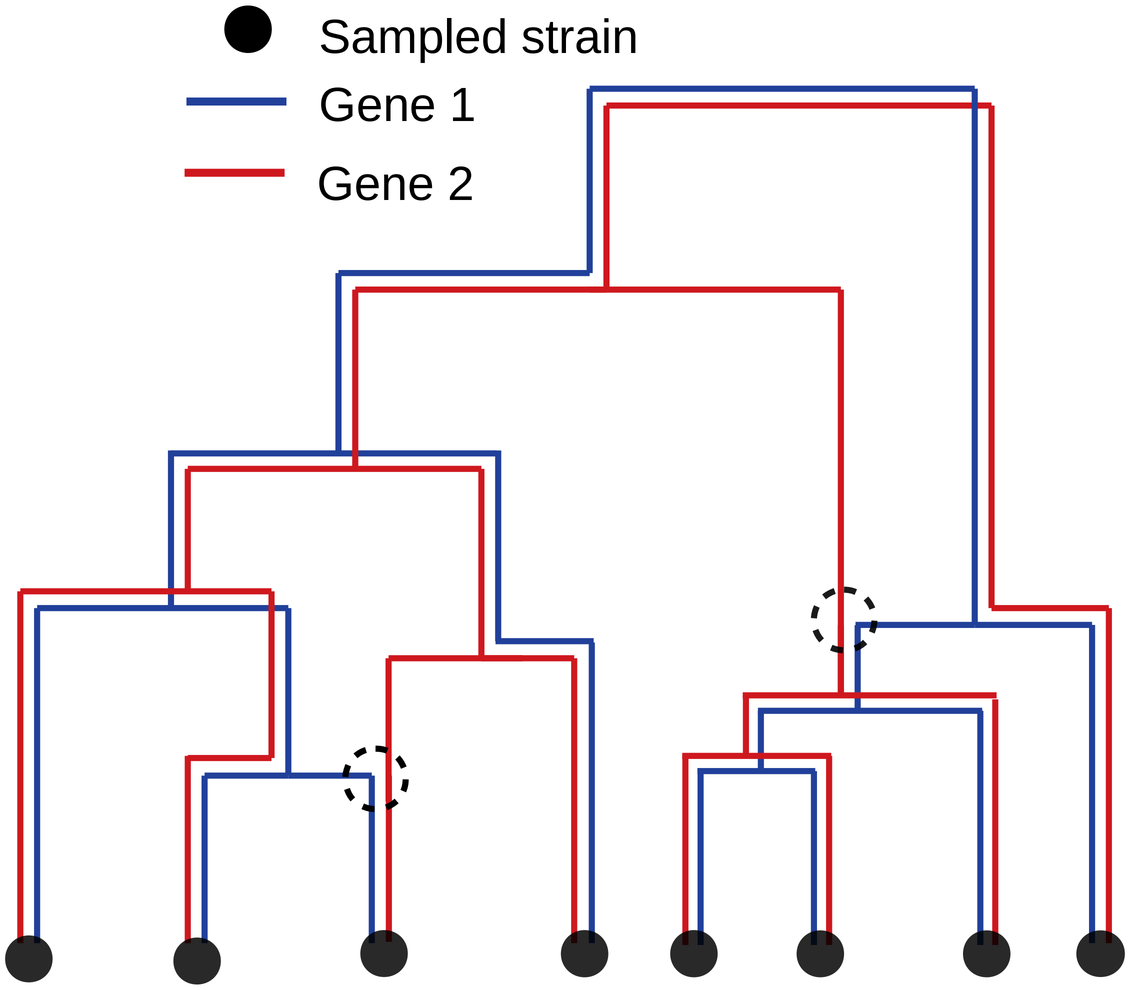
→ **Method based on topological differences between trees**

Inferring the ARG



Maximally compatible clades (MCCs)

The ARG is a **collage of gene trees**



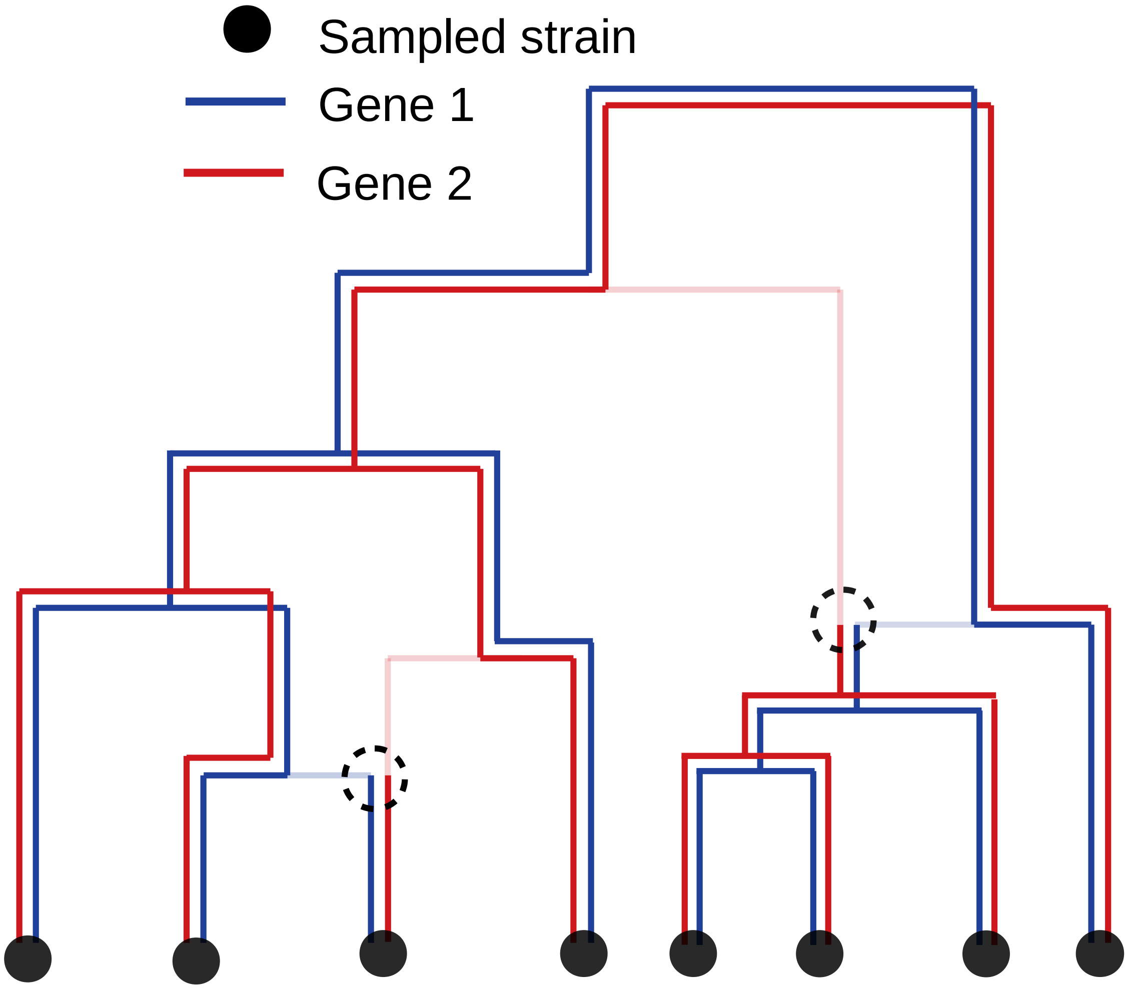
Maximally compatible clades (MCCs)

The ARG is a **collage of gene trees**

Restricting to branches that **belong to both trees**

↓

Maximally compatible clades



Maximally compatible clades (MCCs)

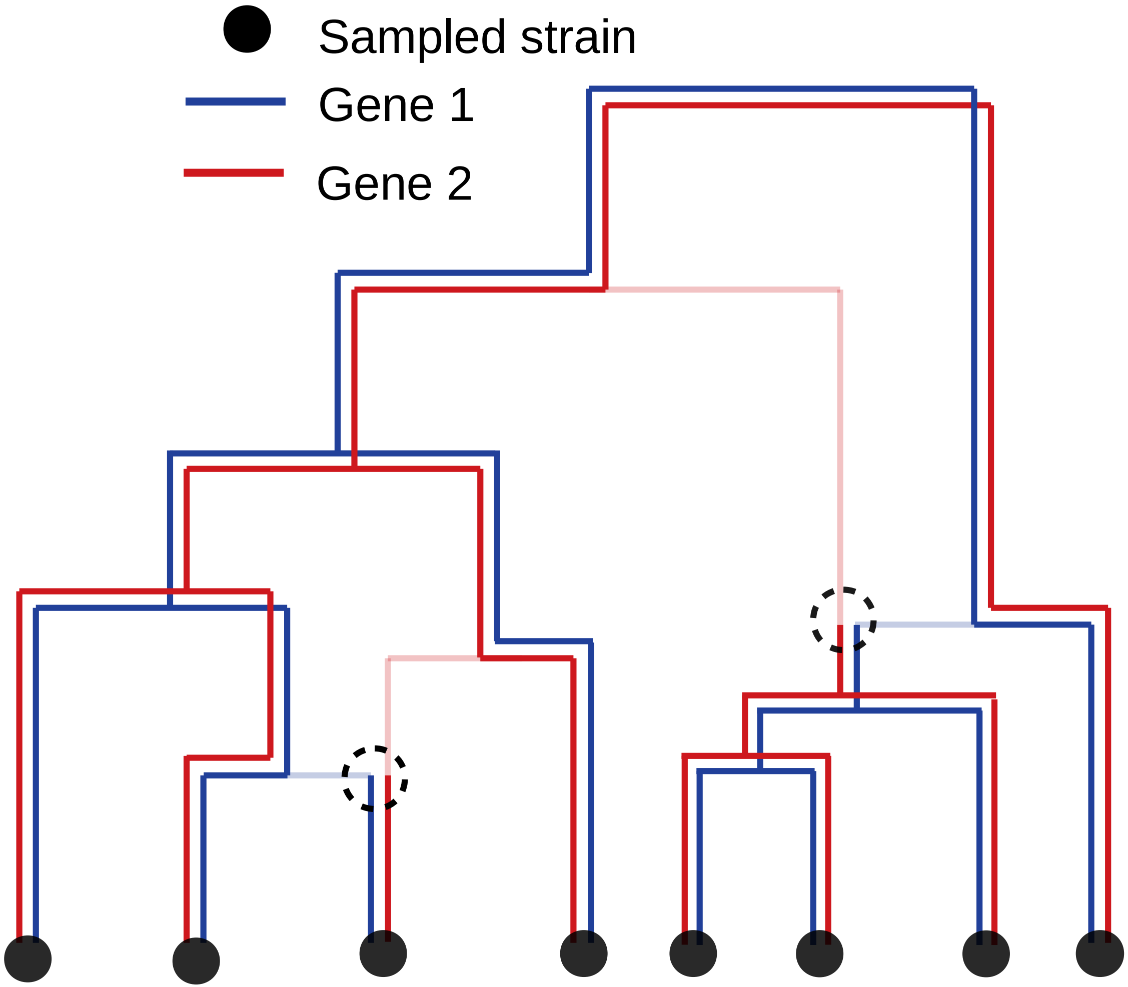
The ARG is a **collage of gene trees**

Restricting to branches that **belong to both trees**

↓

Maximally compatible clades

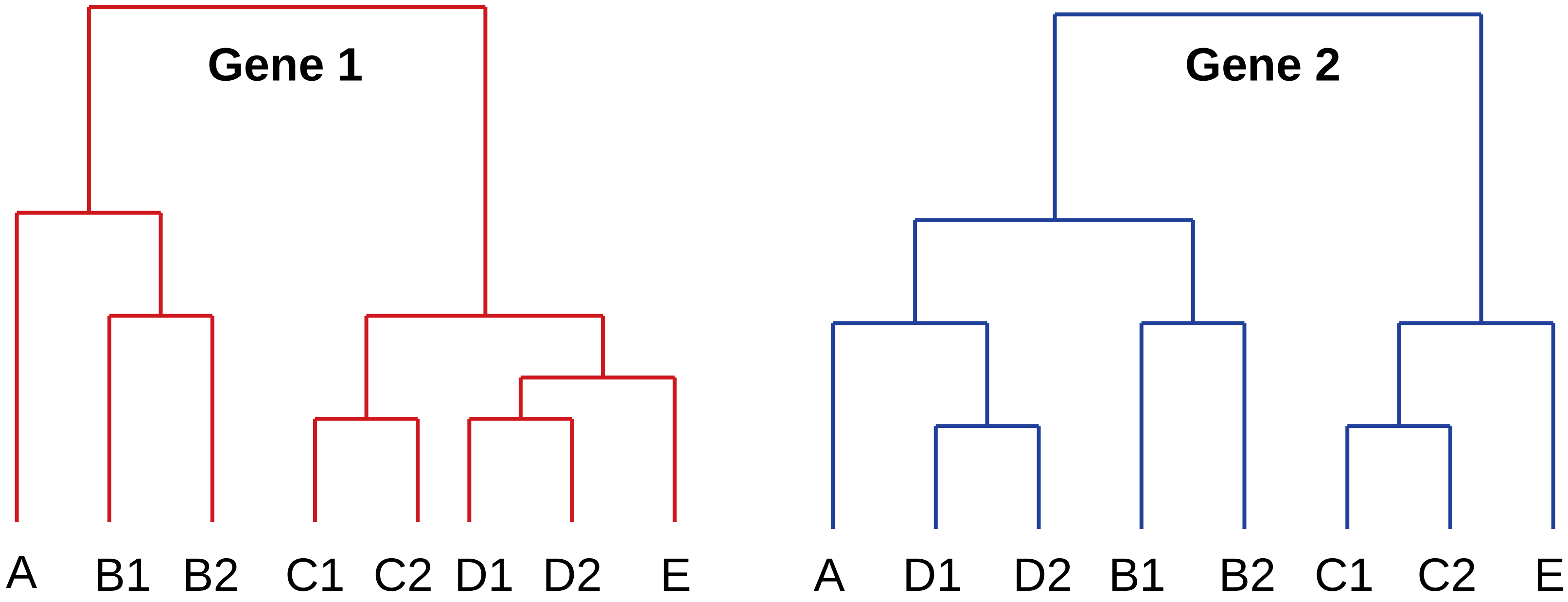
- The **root of an MCC** is either
  - A reassortment
  - The root of both trees
- If **both trees** and **all MCCs** are known, then the **ARG** is known



Inferring the ARG ⟶ Inferring MCCs

First step: naive estimation of MCCs

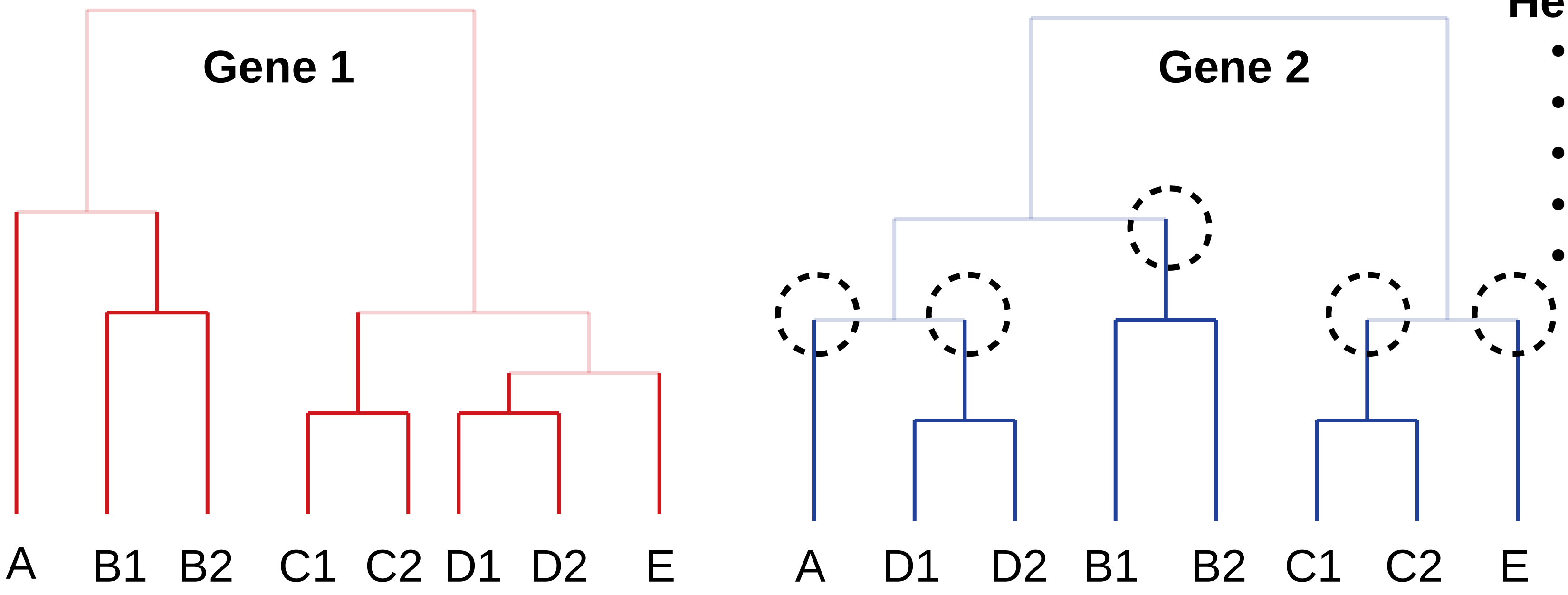
⟶ Take clades that have exactly matching topologies



Inferring the ARG → Inferring MCCs

First step: naive estimation of MCCs

→ Take clades that have exactly matching topologies



Here : 5 naive MCCs

- A
- B1, B2
- C1, C2
- D1, D2
- E

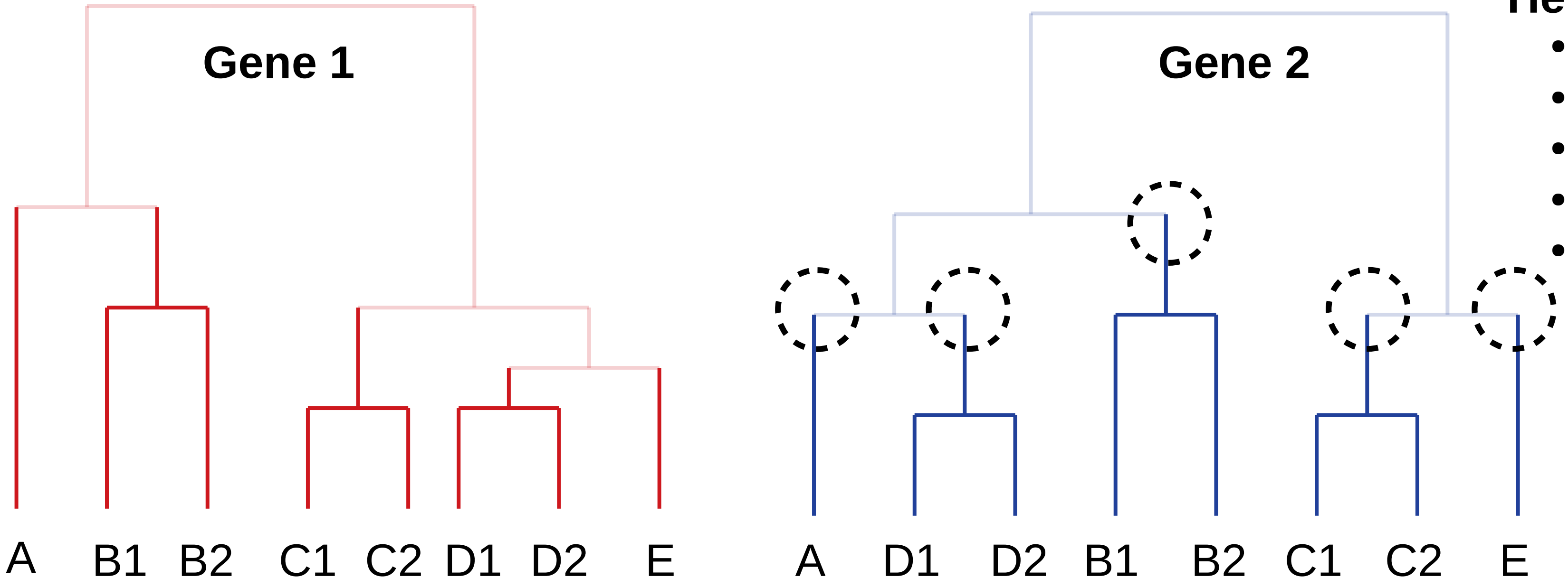
↓

5 reassortments !

Inferring the ARG → Inferring MCCs

First step: naive estimation of MCCs

→ Take clades that have exactly matching topologies



Here : 5 naive MCCs

- A
- B1, B2
- C1, C2
- D1, D2
- E

↓

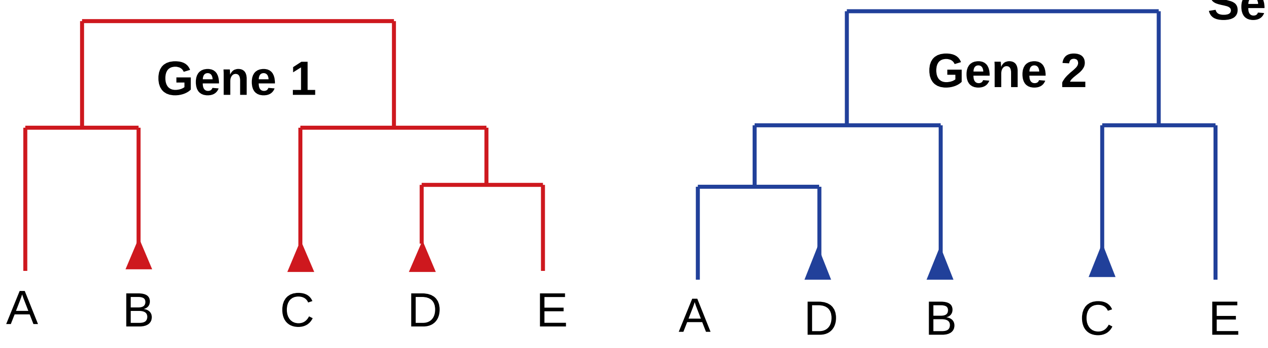
5 reassortments !

Naive estimation :

Finds too many MCCs → Too many reassortments

Conservative approach → Does not overextend MCCs

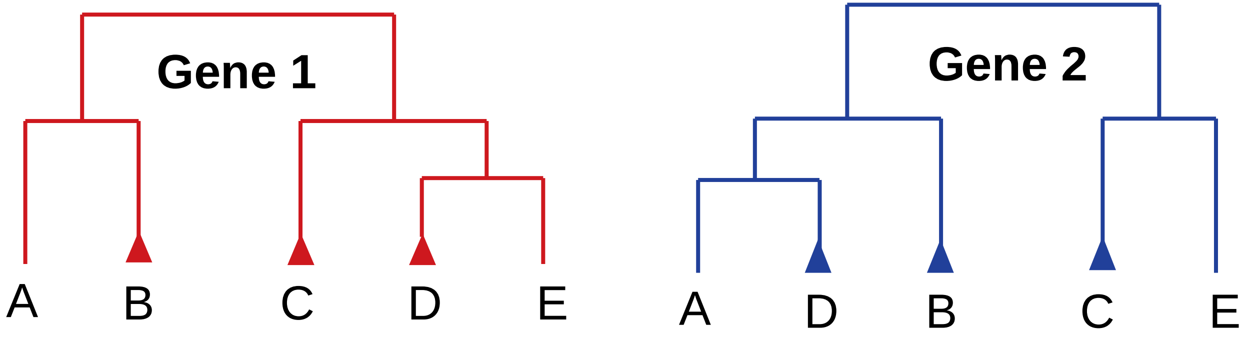
Inferring MCCs



Second step: “reduce” to naive MCCs

- (B1, B2) → B
- (C1, C2) → C
- (D1, D2) → D

Inferring MCCs: Parsimonious approach



First step: “reduce” to naive MCCs

- (B1, B2) → B
- (C1, C2) → C
- (D1, D2) → D

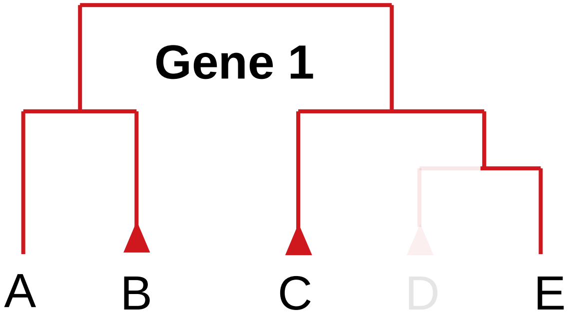
By eye:
D is the reassorted clade.
How can we **formalize** this?

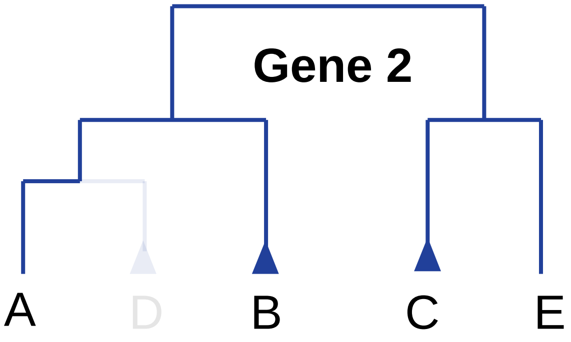
Surrounding of each leaf: **clade** defined by parent:

- A → (A,B) / (A,D)
- B → (A,B) / (A,D,B)
- C → (C,D,E) / (C,E)
- D → (D,E) / (A,D)
- E → (D,E) / (C,E)

→ **5 incompatibilities**

Inferring MCCs: Parsimonious approach





First step: “reduce” to naive MCCs

- (B1, B2) → B
- (C1, C2) → C
- (D1, D2) → D

By eye:
D is the reassorted clade.
How can we **formalize** this?

Surrounding of each leaf: **clade** defined by parent:

- A → (A,B) / (A,D)
- B → (A,B) / (A,D,B)
- C → (C,D,E) / (C,E)
- D → (D,E) / (A,D)
- E → (D,E) / (C,E)

→ **5 incompatibilities**

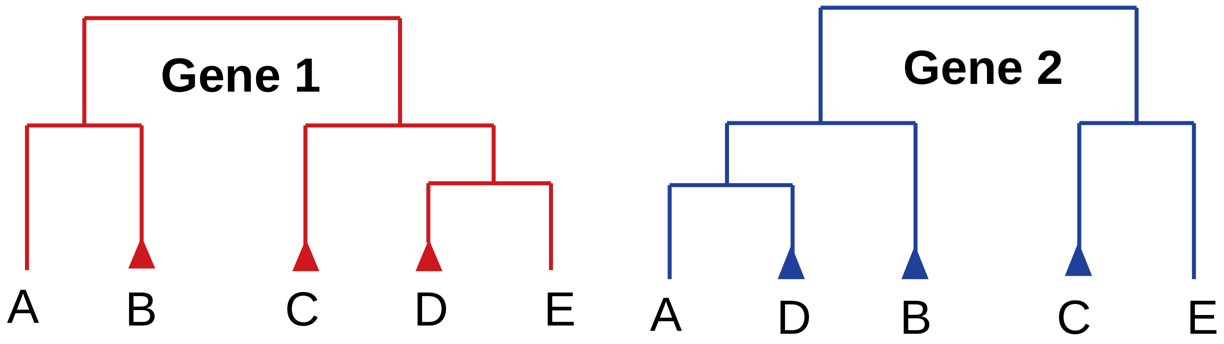
Hypothesis: D is a reassortant → Remove it from the trees

- A → (A,B) / (A,B)
- B → (A,B) / (A,B)
- C → (C,E) / (C,E)
- ~~D → (D,E) / (A,D)~~
- E → (C,E) / (C,E)

→ **0 incompatibilities**

0 remaining reassortments!

Inferring MCCs



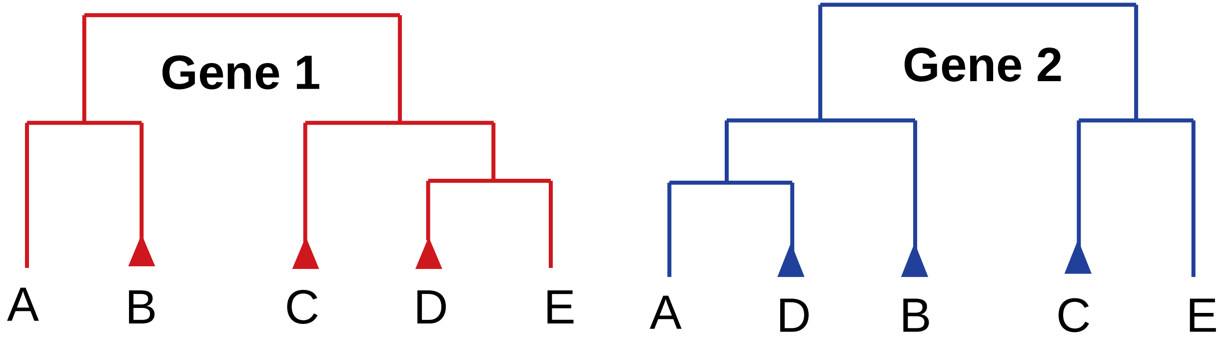
For each leaf ***n***

→ σ_n <
1 if we **remove** ***n***
0 otherwise

→ $\vec{\sigma} = (\sigma_1 \ldots \sigma_L)$: “configuration” vector

→ $\Delta(n, \vec{\sigma})$ <
1 if **incompatibility** above ***n***
0 otherwise

Inferring MCCs



For each leaf ***n***

→ σ_n
- 1 if we **remove** ***n***
- 0 otherwise

→ $\vec{\sigma} = (\sigma_1 \ldots \sigma_L)$: “configuration” vector

→ $\Delta(n, \vec{\sigma})$
- 1 if **incompatibility** above ***n***
- 0 otherwise

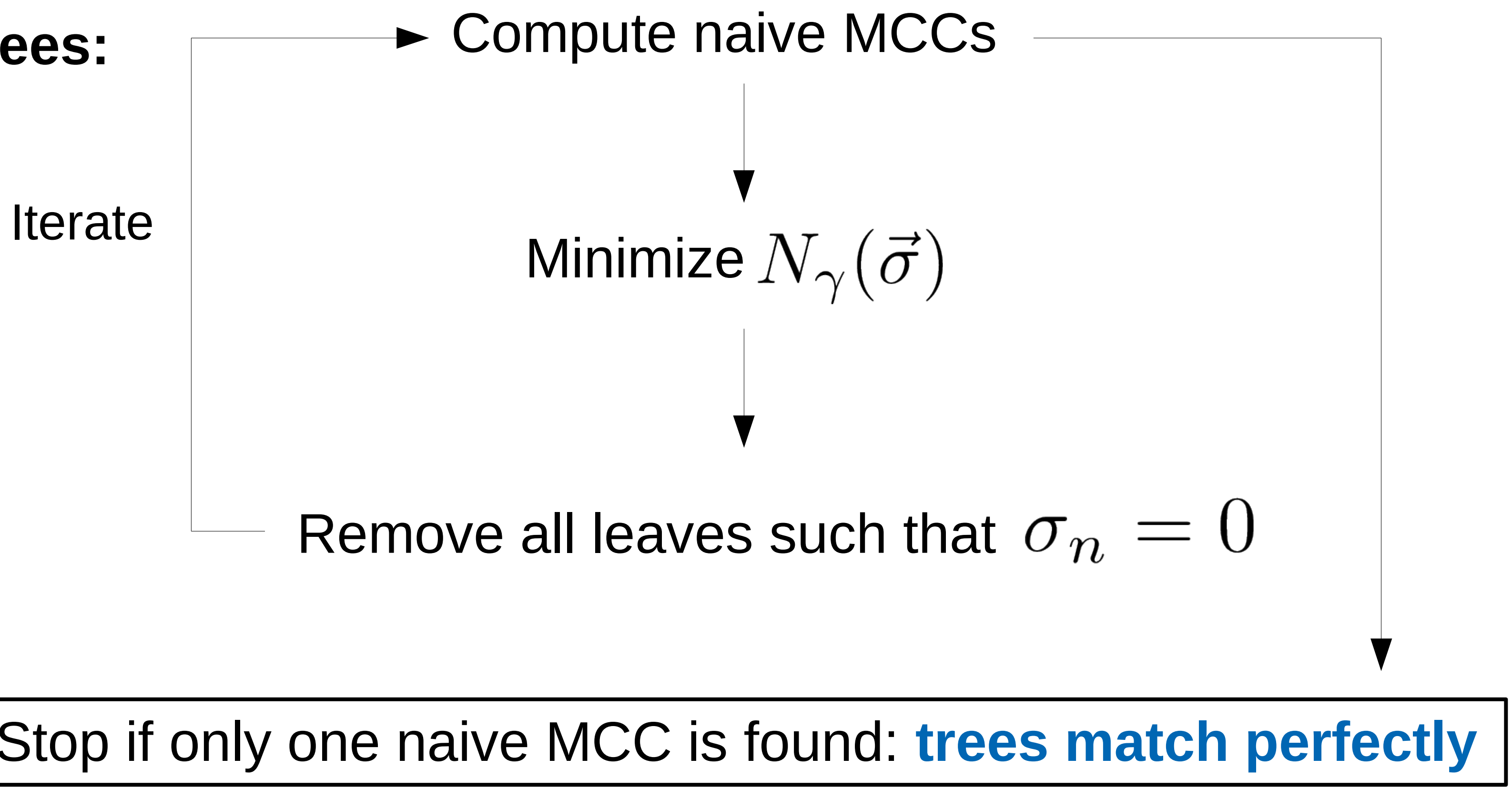
Minimize $$N_\gamma(\vec{\sigma}) = \overbrace{\sum_{n \in leaves} \Delta(n, \vec{\sigma})\sigma_n}^{\text{\# of incompatibilties}} + \gamma\overbrace{(L - |\vec{\sigma}|)}^{\text{\# of removed leaves}}$$

(Simulated annealing)

Minimize incompatibilities with a minimal number of reassortments

Inferring MCCs: summary

Given two trees:



Iterate

Compute naive MCCs

Minimize $N_\gamma(\vec{\sigma})$

Remove all leaves such that $\sigma_n = 0$

Stop if only one naive MCC is found: **trees match perfectly**

Interpretation of gamma

$$N_\gamma(\vec{\sigma}) = \sum_{n \in leaves} \Delta(n, \vec{\sigma})\sigma_n + \gamma(L - |\vec{\sigma}|)$$

- $\gamma \to \infty$ **Infinite cost** for removing leaves $\longrightarrow$ **Naive approach**

Interpretation of gamma

$$N_\gamma(\vec{\sigma}) = \sum_{n \in leaves} \Delta(n, \vec{\sigma})\sigma_n + \gamma(L - |\vec{\sigma}|)$$

- $\gamma \to \infty$ **Infinite cost** for removing leaves → **Naive approach**

- $\gamma = 1$ $N(\vec{\sigma})$ = **# incompatibilities** + **# removed leaves**

  Reassortments w. naive approach ↑ (# incompatibilities)

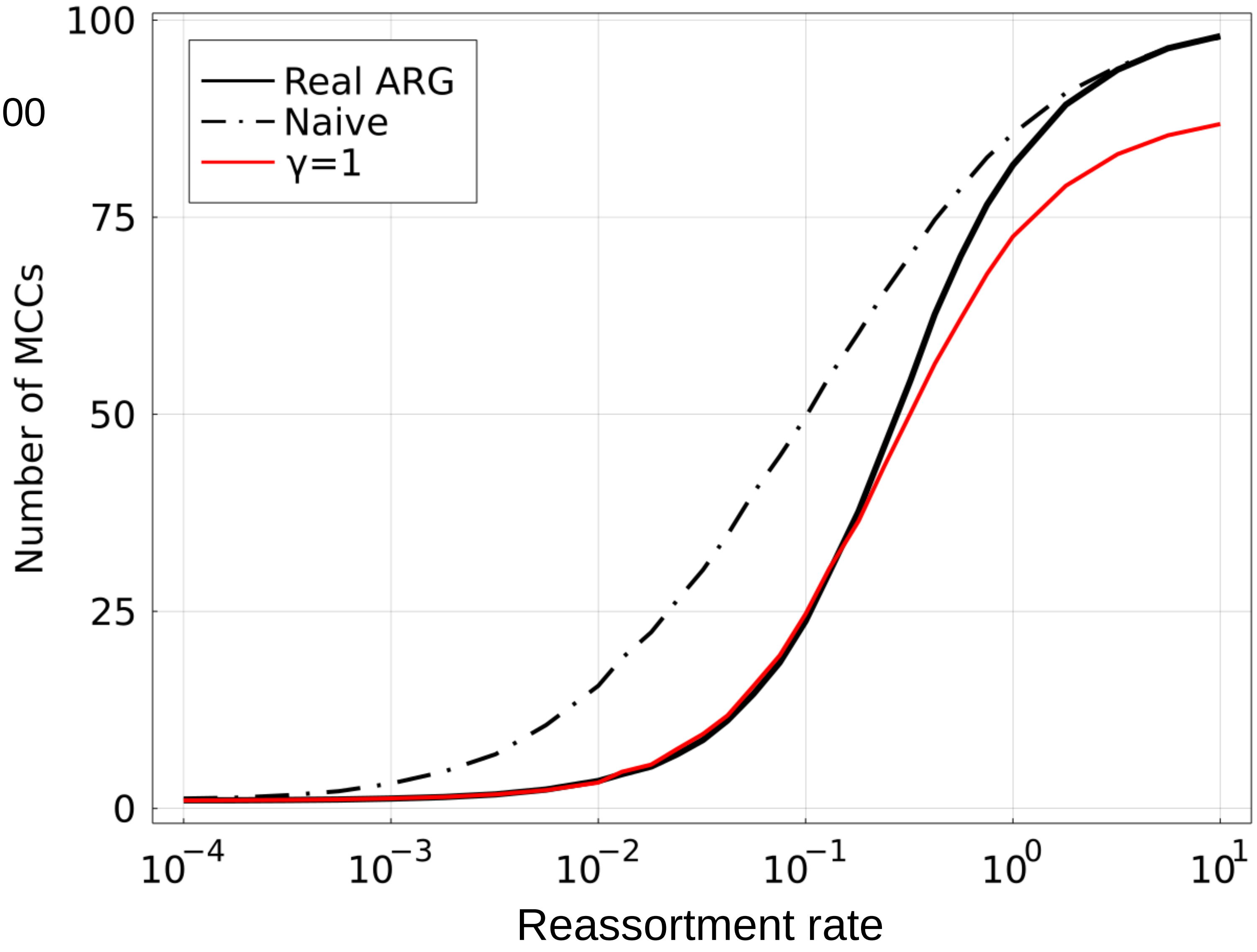
  Enforced reassortments ↑ (# removed leaves)

  $N(\vec{\sigma})$ = **Total number of reassortments** → **“Parsimonious” approach**

- Intermediate γ → **Interpolate** between **naive** and **“parsimonious”**

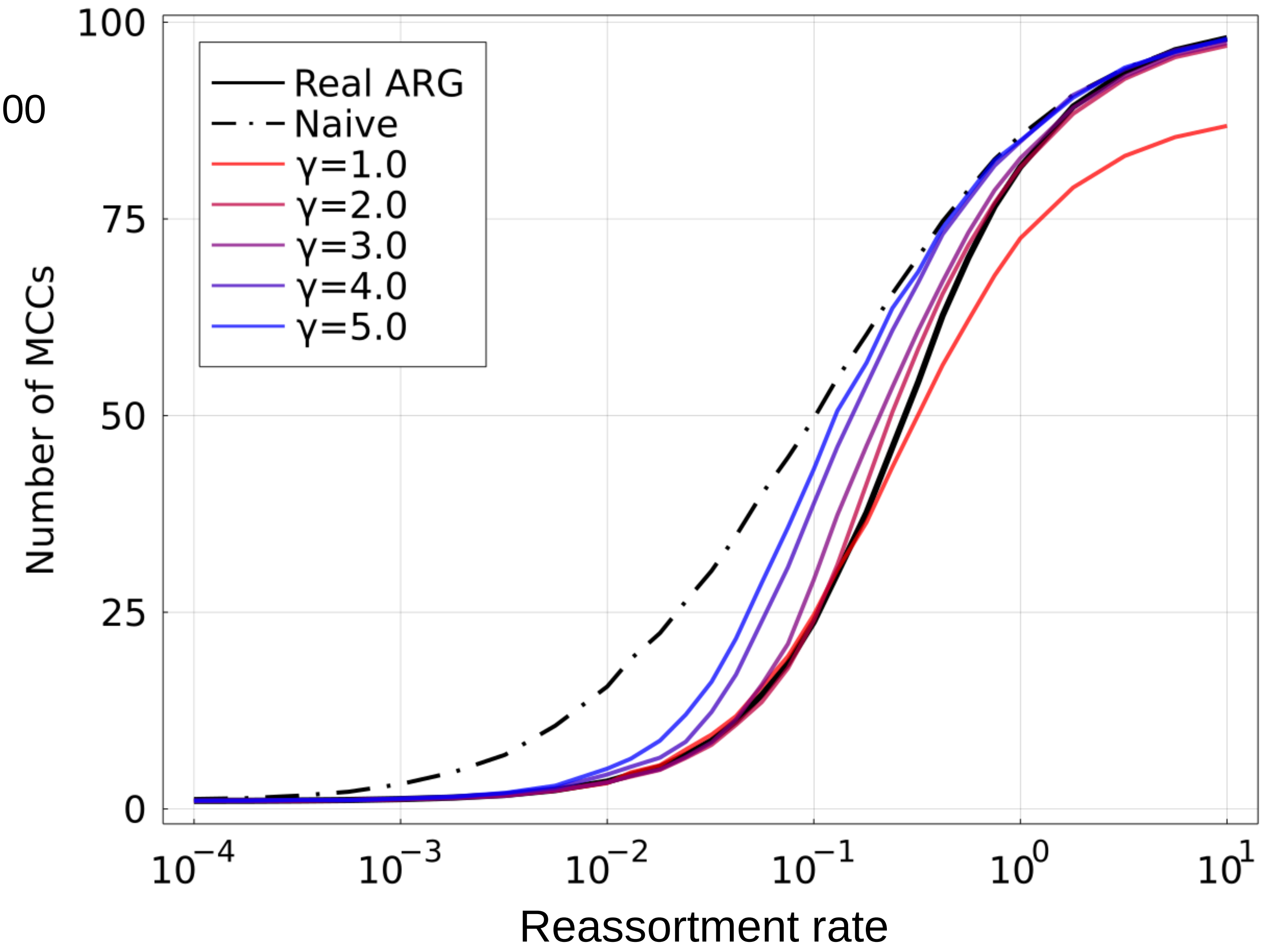
Interpretation of gamma

ARG of 100
leaves



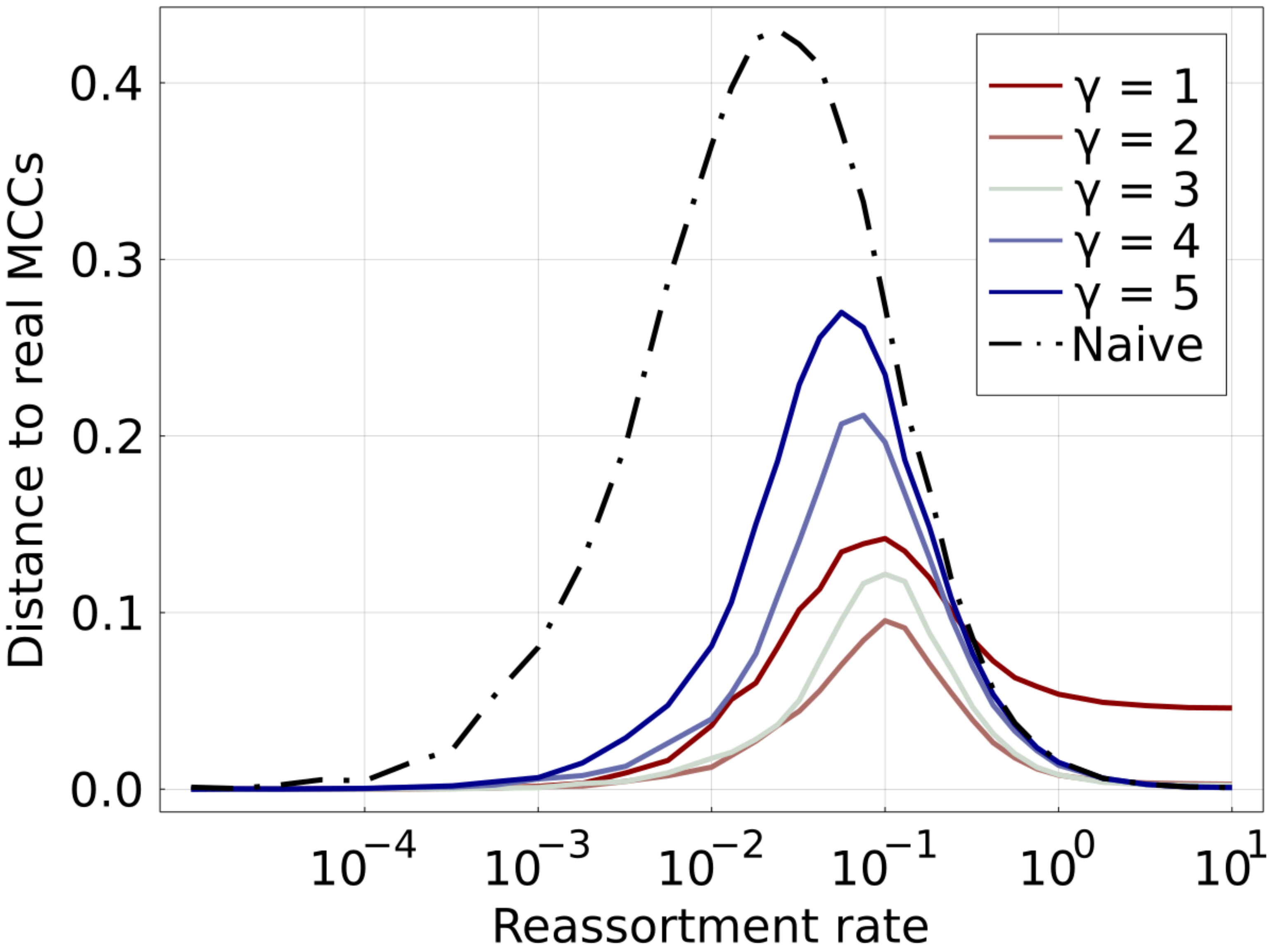
Interpretation of gamma

ARG of 100
leaves



Choosing gamma

Distance: Variation of Information (VI)

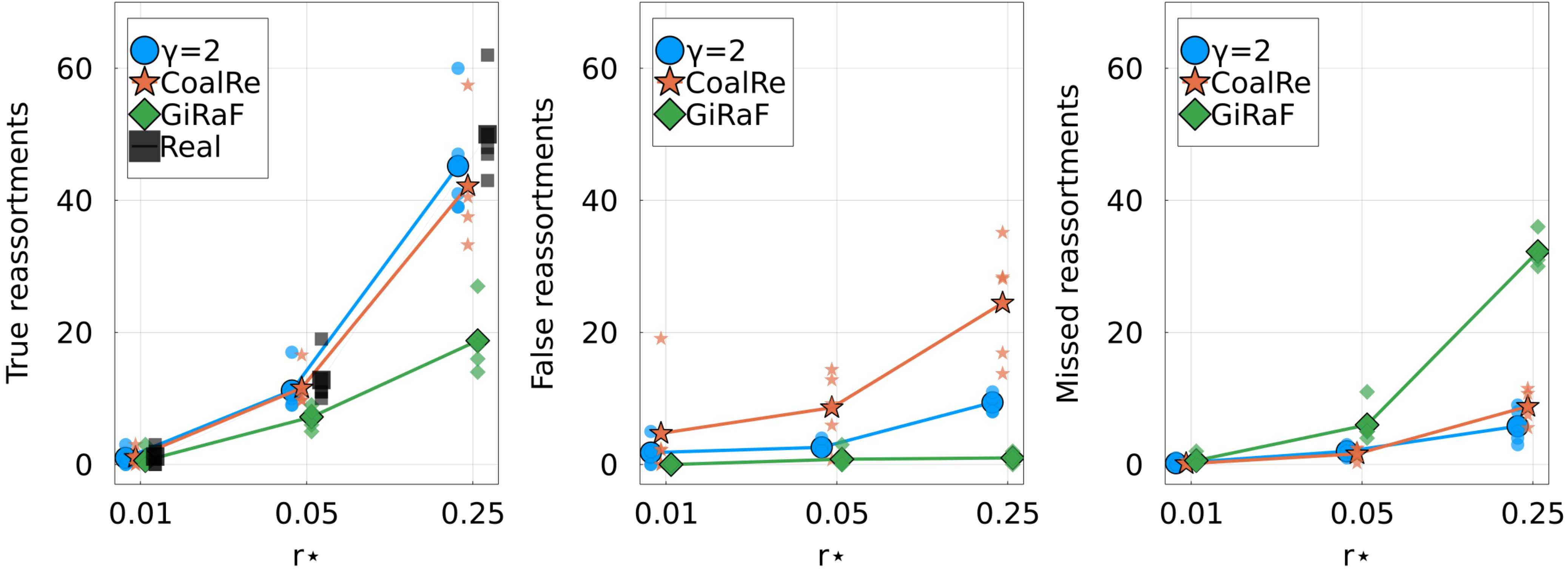


$$\boxed{\gamma = 2}$$

Comparison w. other methods

CoalRe: ML based **[Müller et. al. 2020]**

GiRaF: topology based **[Nagarajan & Kingsford 2011]**



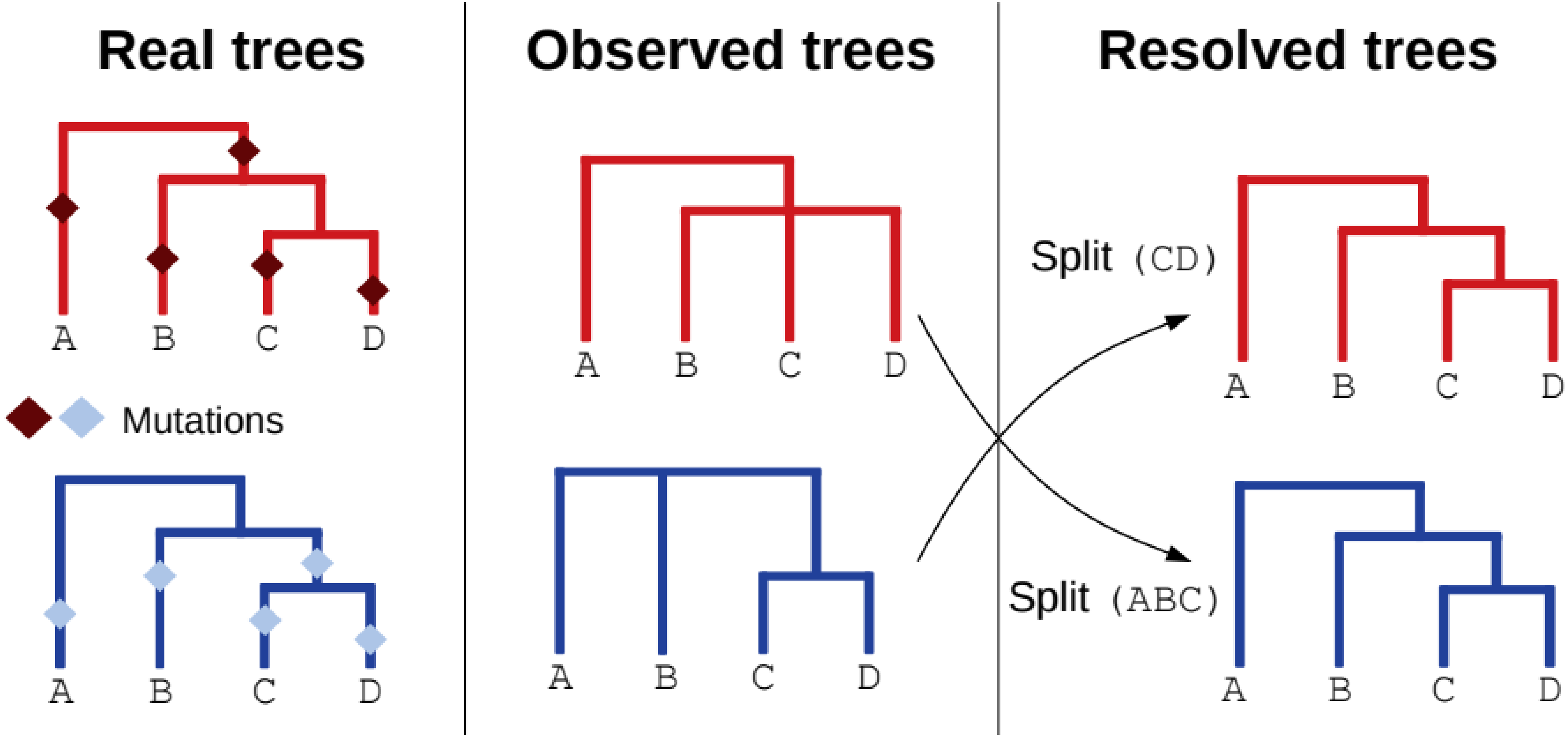
Runtime

	CoalRe	GiRaF	Treeknit
Inferring trees		20min	30s
Inferring the ARG	~days	40s	40ms

for 100 leaves

Application: better resolved trees

Using sequences of both segments

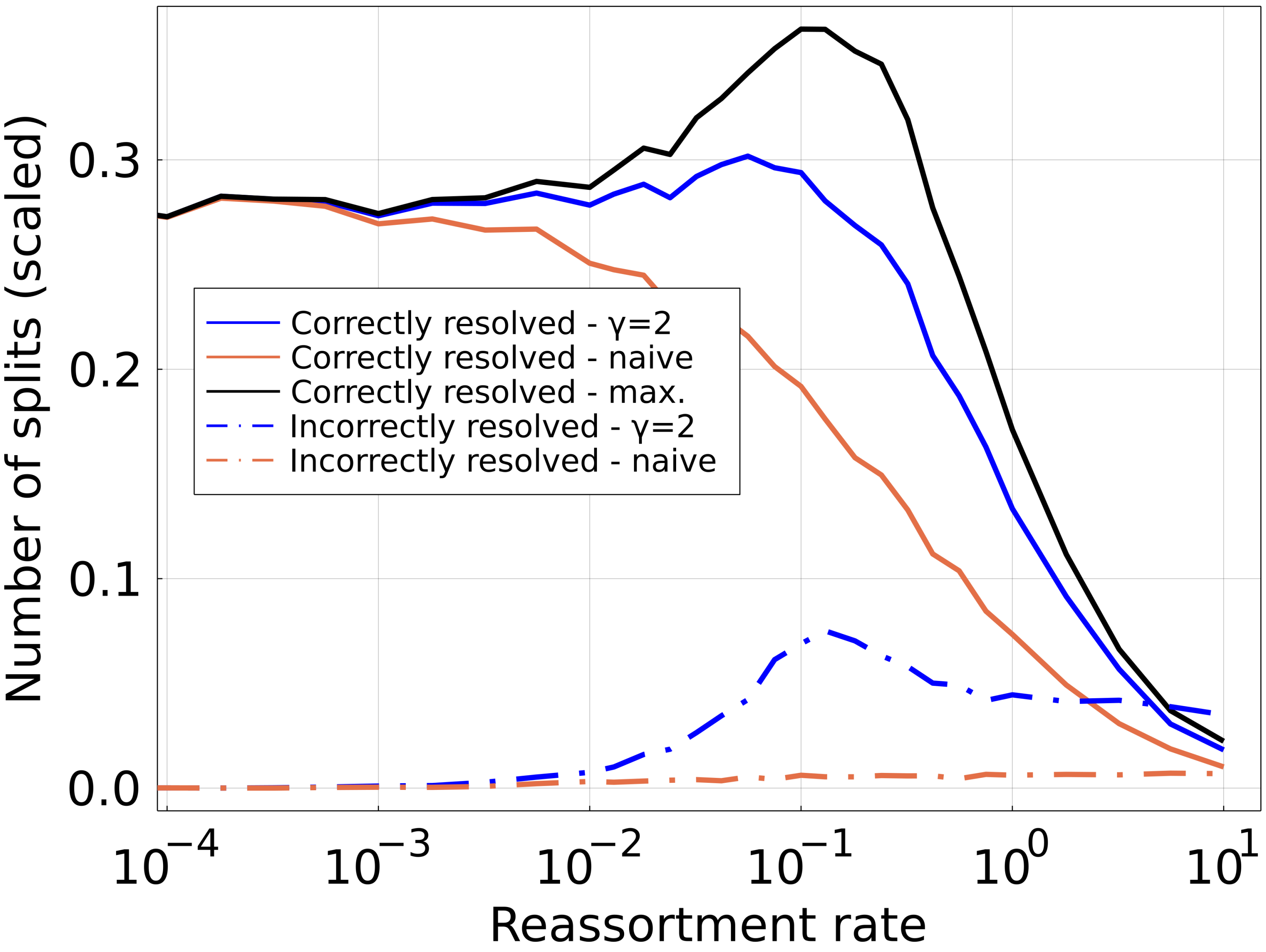


Real trees

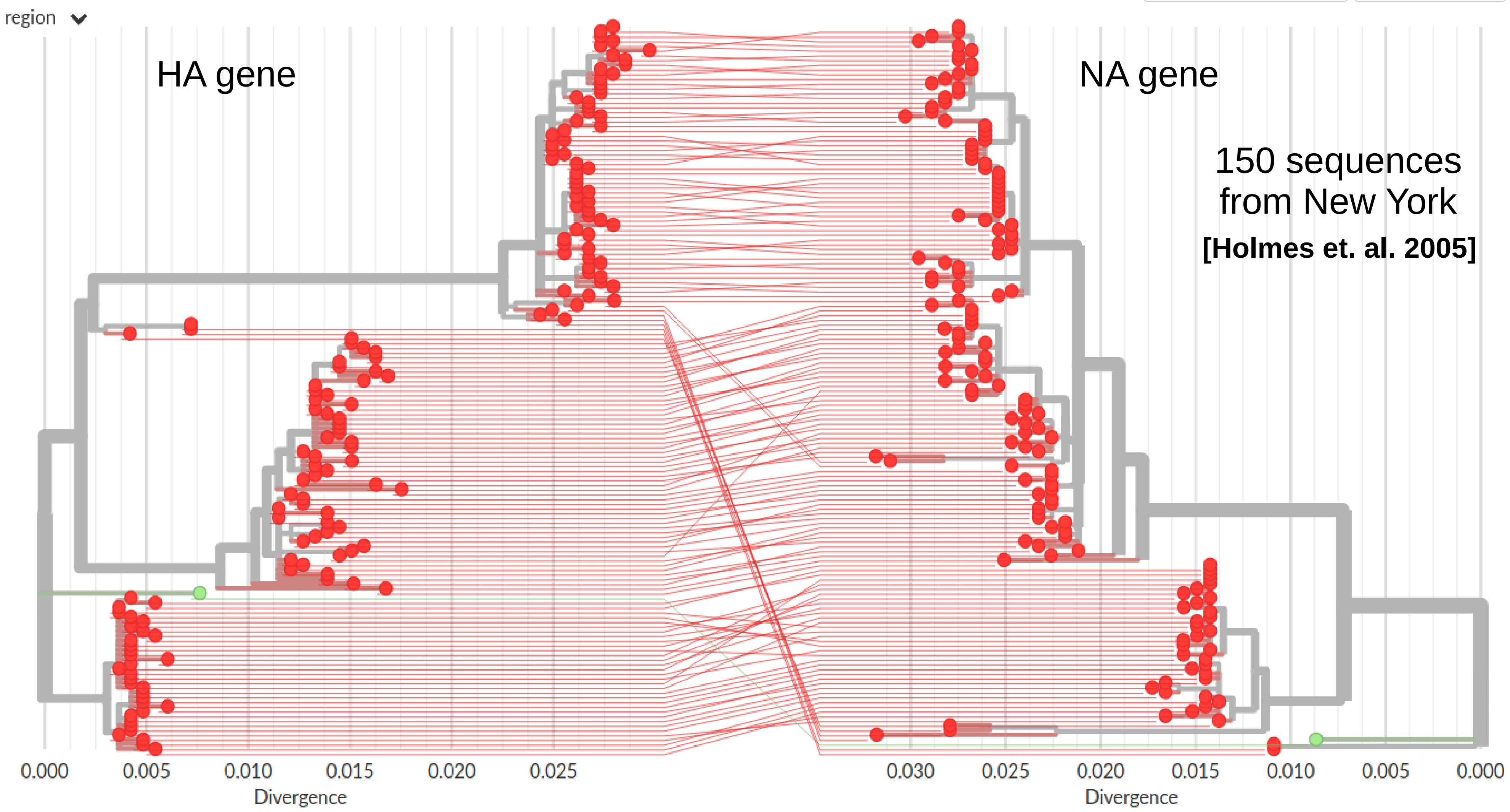
Observed trees

Resolved trees

Application: better resolved trees



Application: disentangling tanglegrams

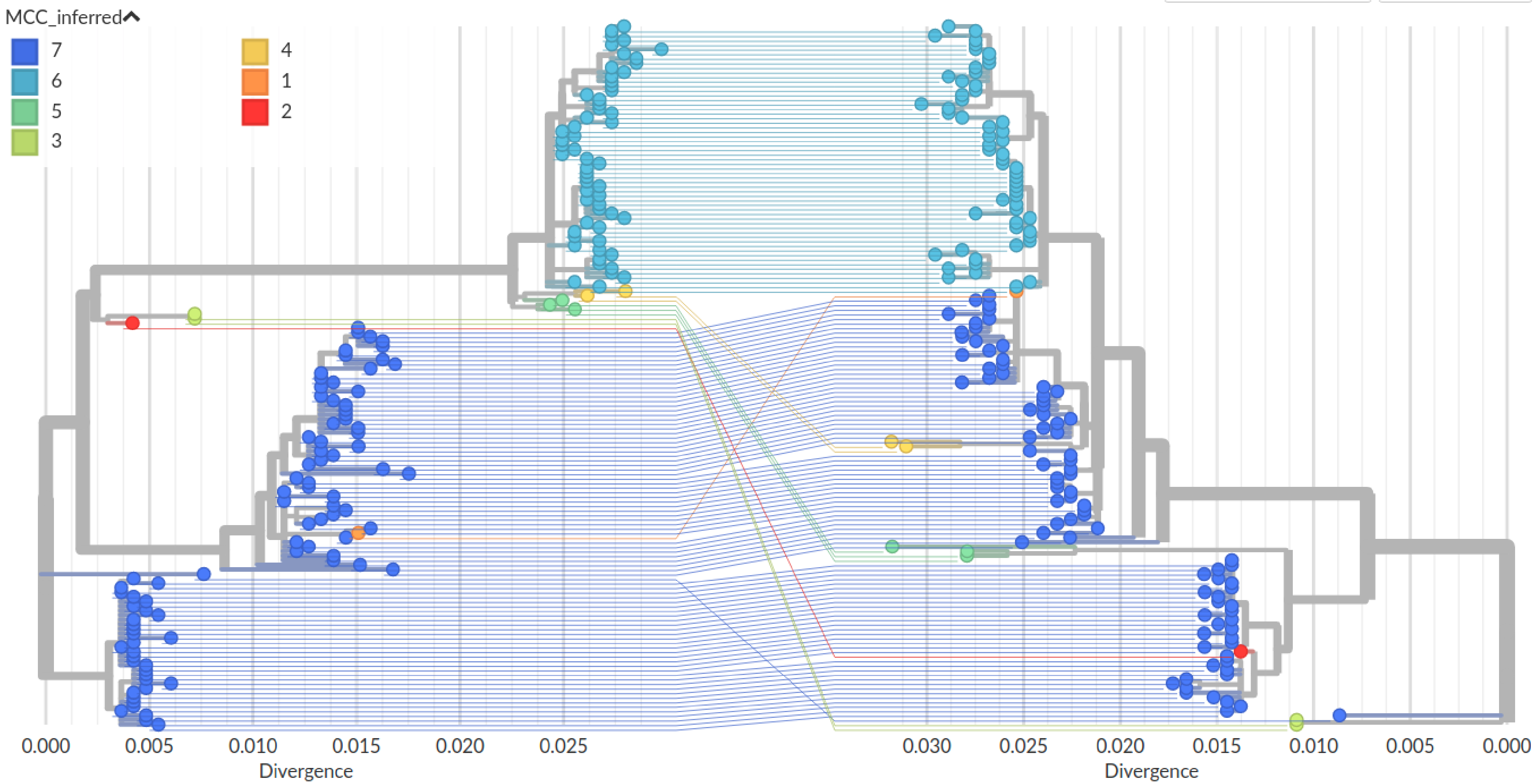


150 sequences
from New York

[Holmes et. al. 2005]

Without the knowledge of reassortments: hard problem

Application: disentangling tanglegrams



With the knowledge of reassortments: easy

Summary

Available at **github.com/PierreBarrat/RecombTools**

Results

- **Treeknit:** Heuristic to infer ARGs from two trees
- Depends on **one parameter**, **interpolating** between naive and parsimonious inference
- Very **fast** runtime
- **Good performance** on **simulated data** for all reassortment rates

Applications

- Resolve trees
- Visualisation: disentangle tanglegrams
- Knowledge of the ARG → Effect of reassortment on influenza evolution

Thank you for listening!