COLLABORATIVE RESEARCH CENTER 1310

Predictability in Evolution

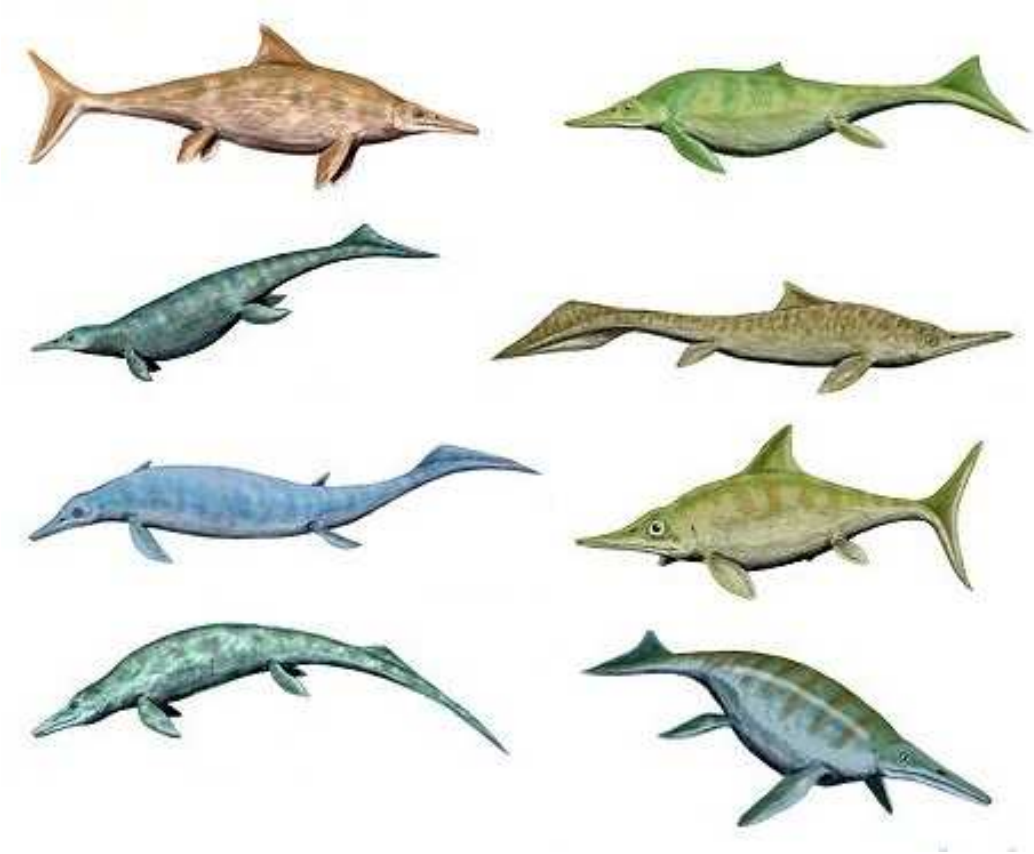
How predictable is evolution?

Joachim Krug
Institute for Biological Physics, University of Cologne

Physikalisches Kolloquium, Carl von Ossietzky Universität Oldenburg, 21.11.2022

How predictable is evolution?

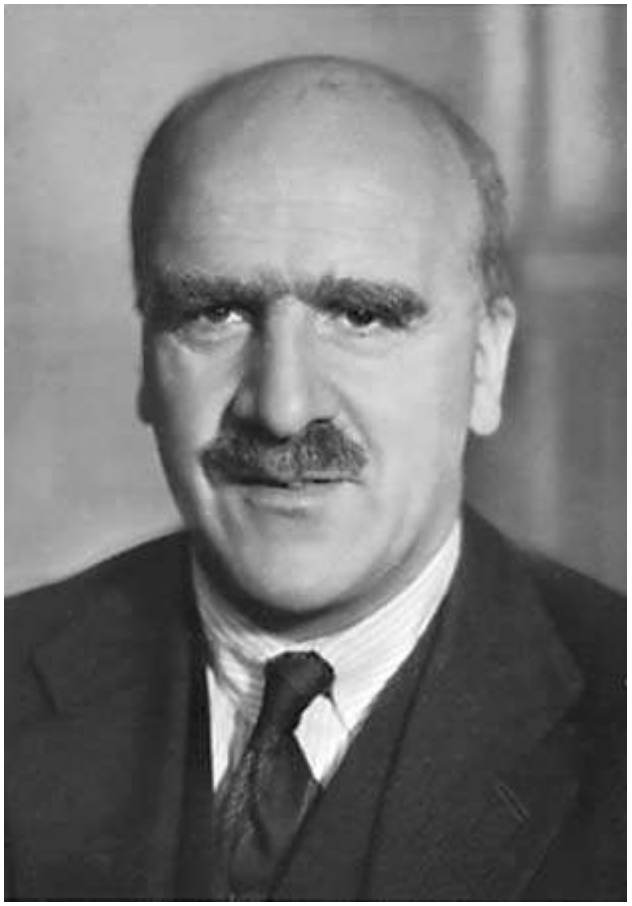
- If we could replay the ‘tape of life’, would the outcome be similar to the current biosphere or something completely different? S.J. Gould (1989)



- “The evolutionary routes are many, but the destinations are limited.”

S. Conway Morris (2003)

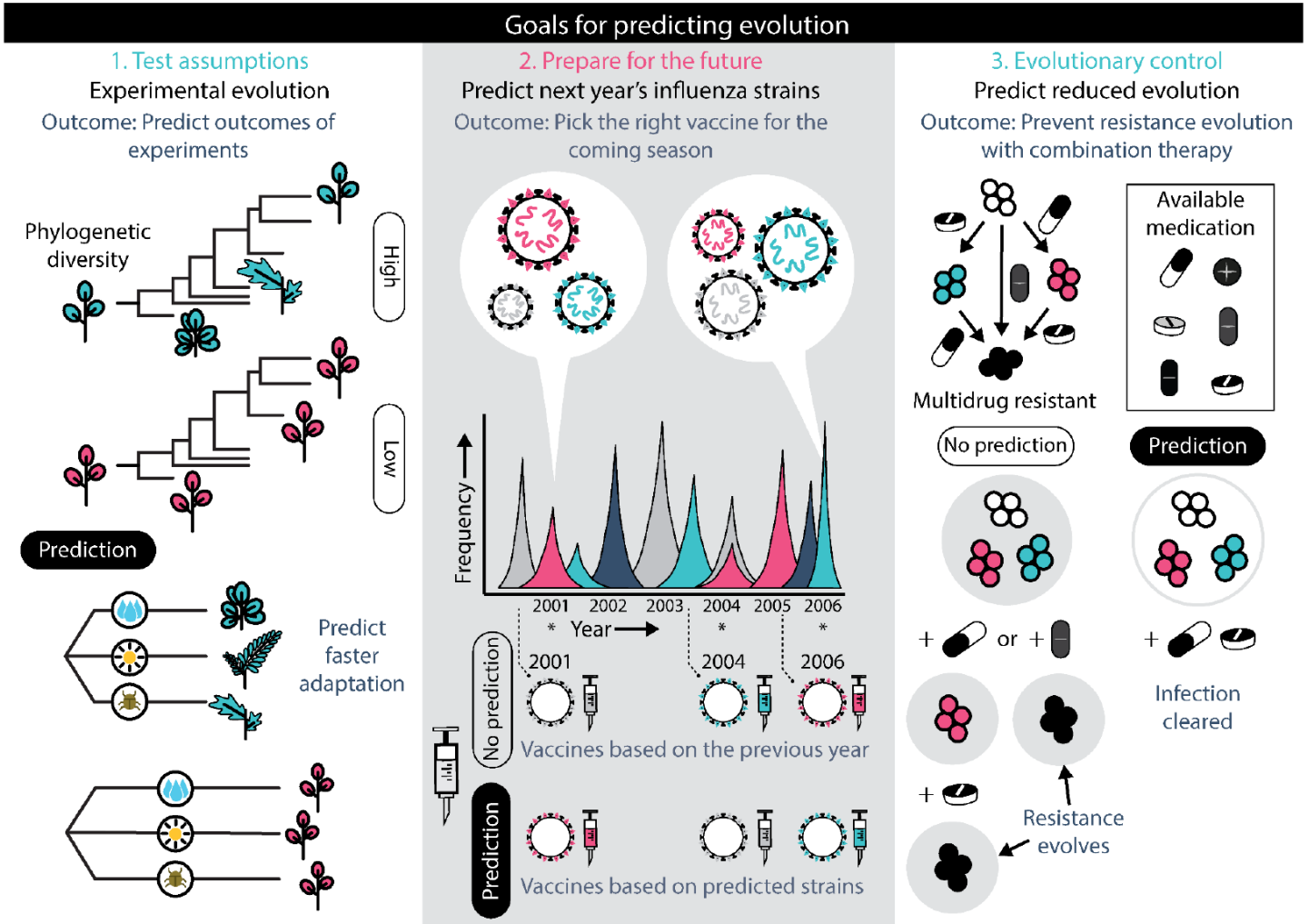
Why predict evolution?



"No scientific theory is worth anything unless it enables us to predict something which is actually going on. Until that is done, theories are a mere game of words, and not such a good game as poetry."

J.B.S. Haldane (1937)

Why predict evolution?



M.T. Wortel et al., EcoEvoRxiv 2022 (to appear in Evol. Appl.)

Notions of predictability

- The evolutionary process is an intricate interplay of deterministic selection and stochastic mutational and reproductive events

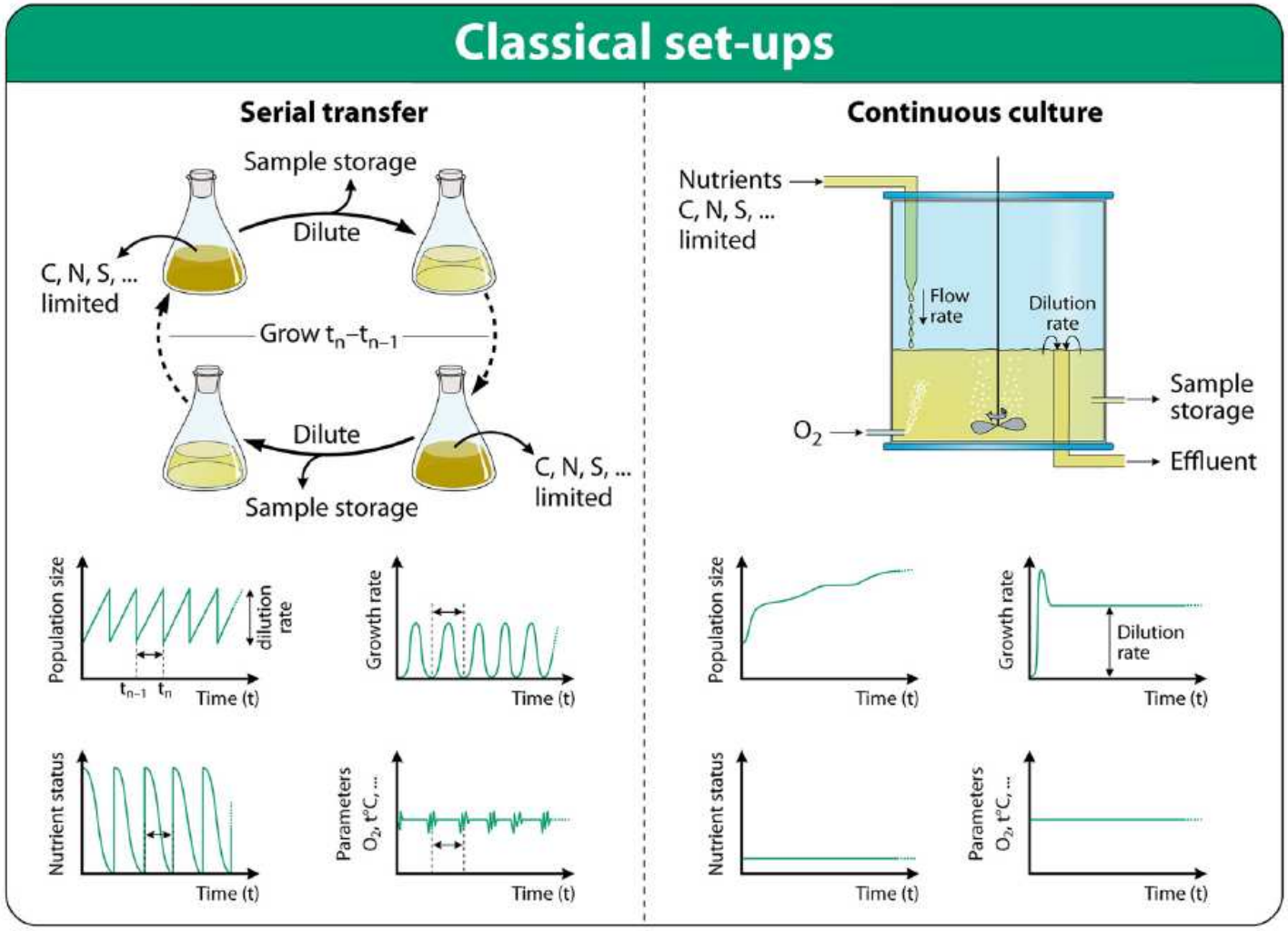
- Strong predictability implies the ability to forecast evolution forward in time (e.g., to predict the genetic evolution of SARS-CoV2 or the emergence of antibiotic resistance)

- Weak (*a posteriori*) predictability implies repeatability in replicate realizations of the process

- Repeatability can be studied in evolution experiments with microbes

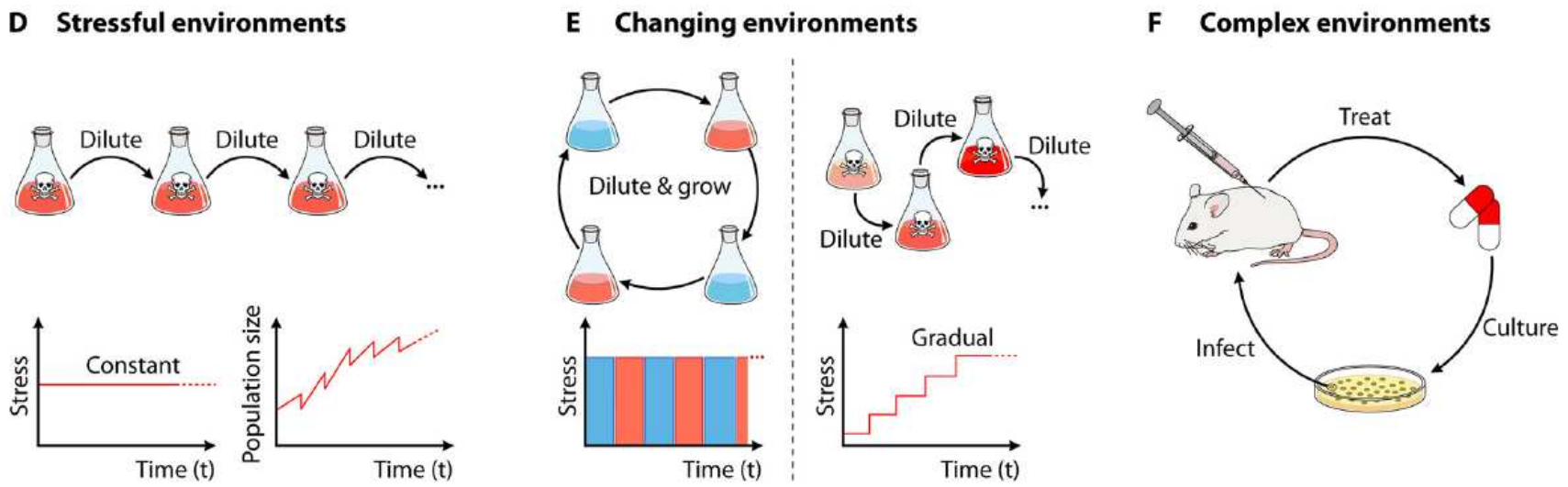
Experimental evolution

Experimental evolution with microbes



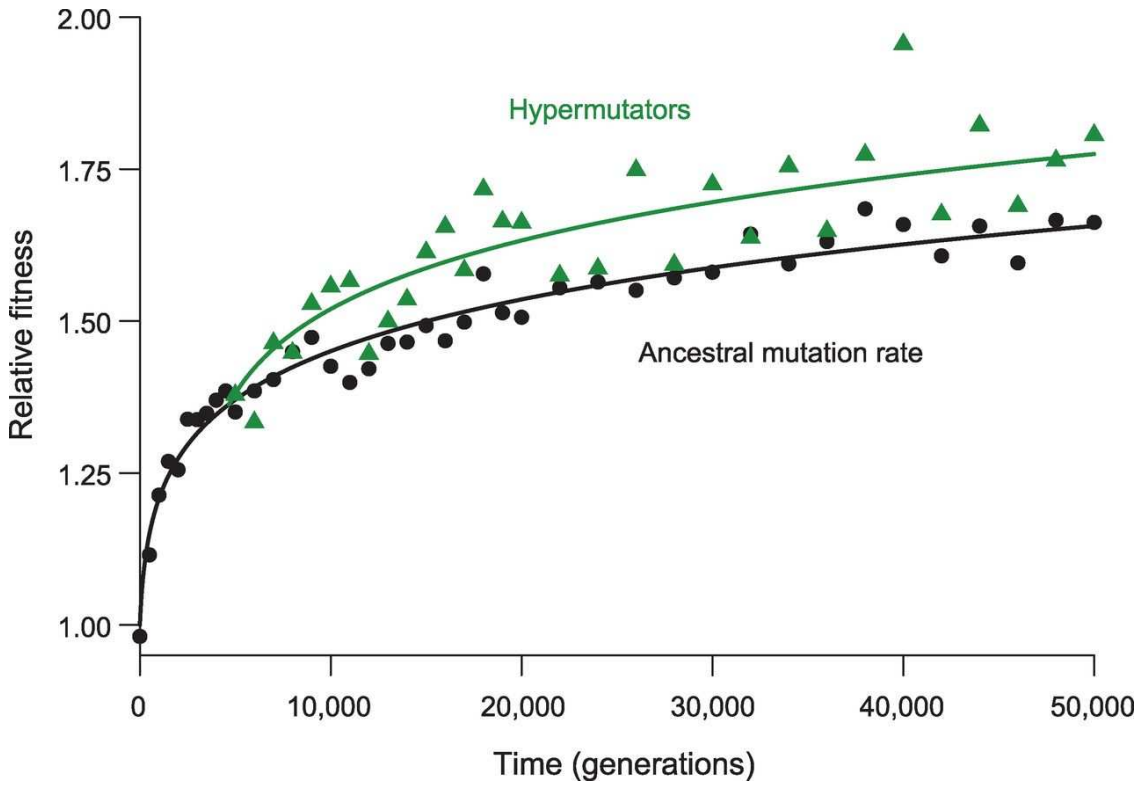
Van den Bergh et al., Microbiology and Molecular Biology Reviews 2018

Experimental evolution with microbes



Van den Bergh et al., Microbiology and Molecular Biology Reviews 2018

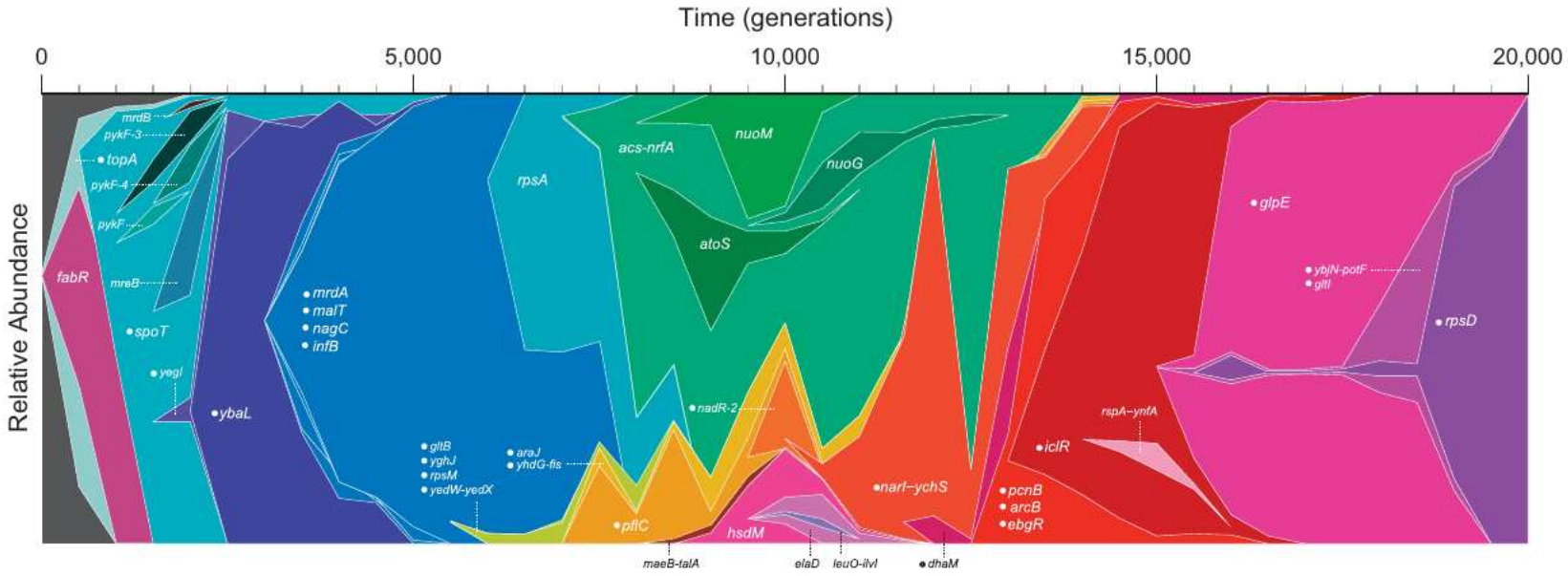
The long-term evolution experiment with *E. coli*



Wiser et al., Science 2013

- Started in 1988 with 12 populations in a poor nutrient environment
- Fitness (= growth rate relative to ancestor) increases consistently, but at a slowing pace

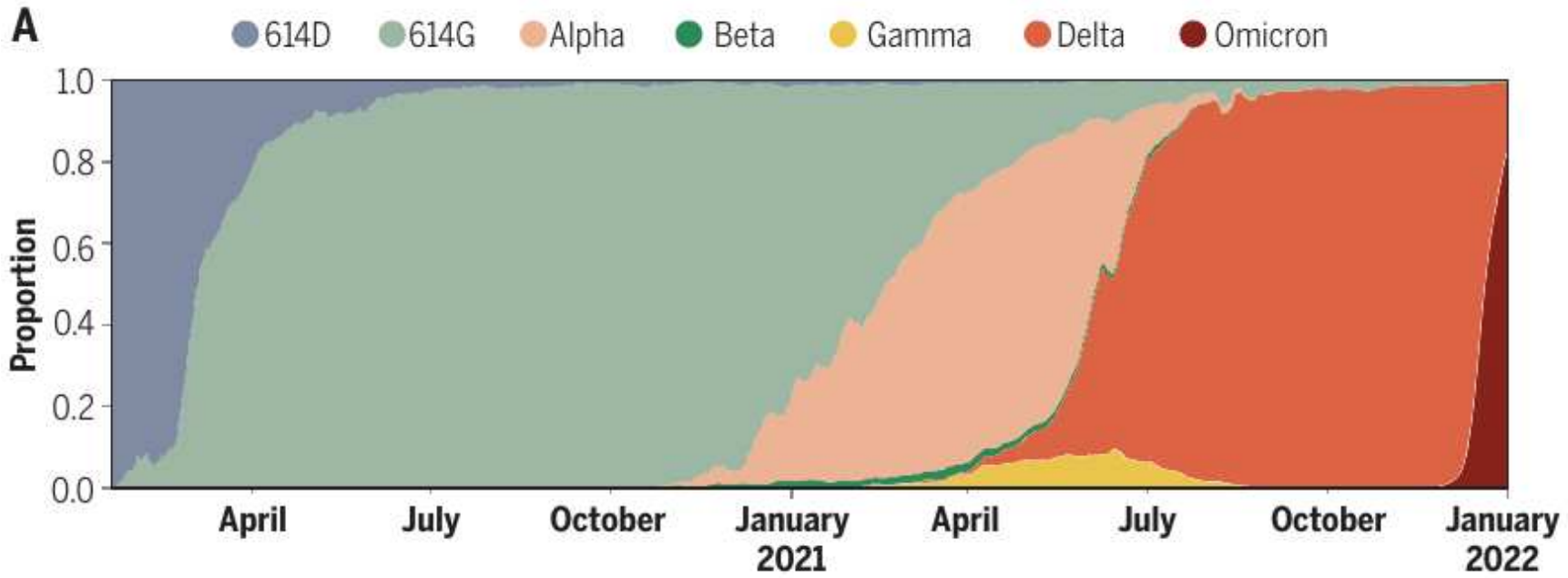
Complex genetic evolution



R.E. Lenski, The ISME Journal 2017

- Muller plot shows abundances of 42 mutations in one of the 12 populations
- Labels indicate names of mutated genes, and dots mark mutations that eventually take over the population (= fix)
- Clonal interference between subpopulations carrying different mutations

Evolution of SARS CoV-2



K. Koelle et al., Science 2022

- Sequential replacement of variants with increasing transmissibility

Goal of this talk

- Describe two case studies where the effect of different factors on evolutionary repeatability could be quantified using mathematical models

- Both studies are based on experiments addressing the evolution of antibiotic resistance

Goal of this talk

- Describe two case studies where the effect of different factors on evolutionary repeatability could be quantified using mathematical models

- Both studies are based on experiments addressing the evolution of antibiotic resistance

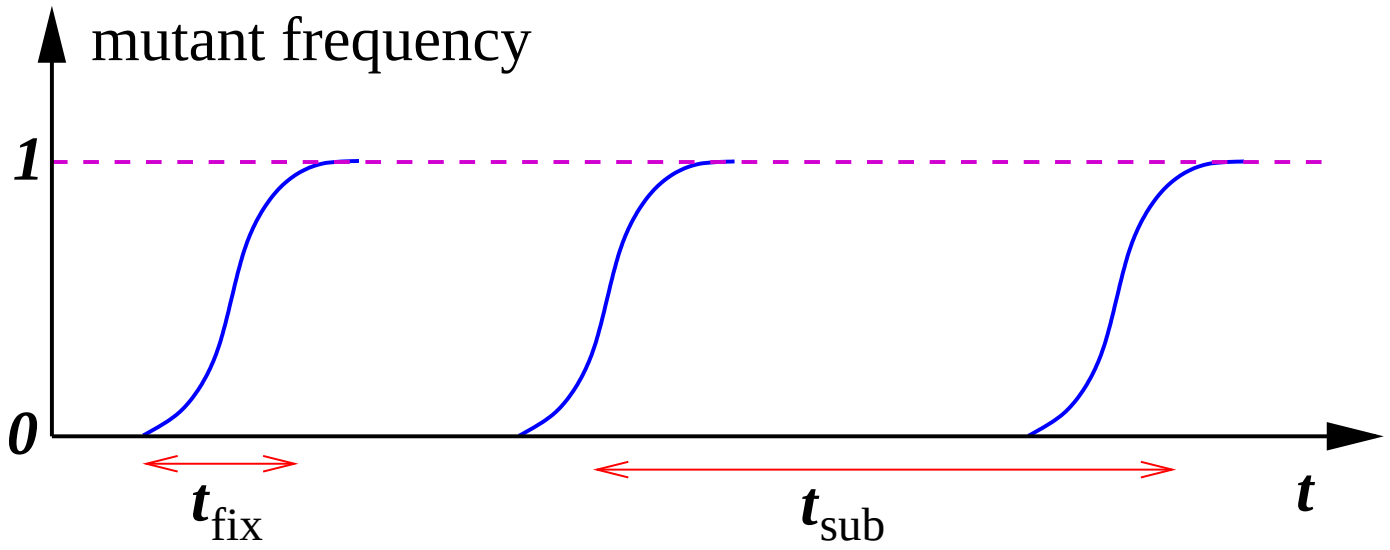
Outline

- Unpredictable repeatability of a single step of evolution
  S.G. Das, JK, PNAS 119:e2209373119 (2022)

- Repeatability of evolutionary pathways in large vs. small populations
  Schenk, Zwart et al., Nature Ecology & Evolution 6:439 (2022)

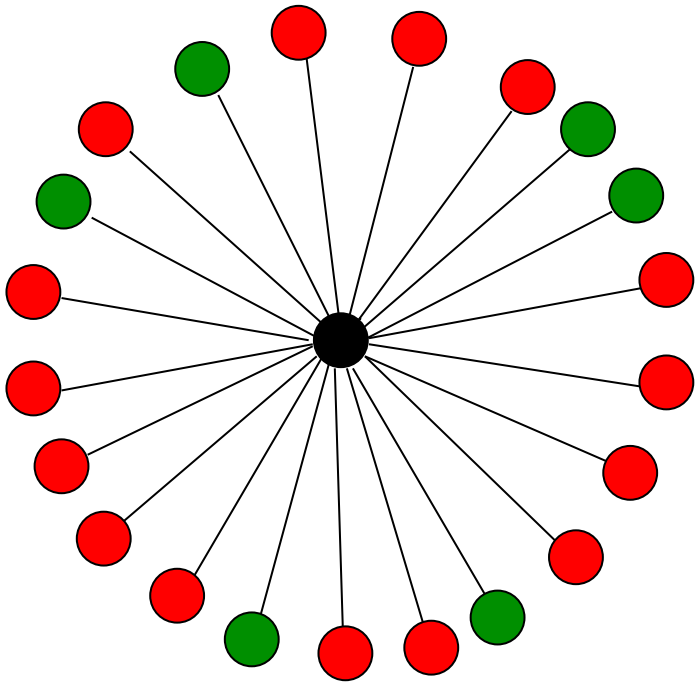
A single step of evolution

Sequential evolution



- Beneficial mutations that increase fitness arise sequentially and fix
- The time between fixation events is longer than the duration of an event

A single step of evolution



- The current type has access to a set of deleterious and beneficial mutations
- A step of evolution occurs by fixation of one of the beneficial mutations
- What is the probability that the same mutation is fixed in two replicate populations?

An analogy

- What is the probability that two fair dice show the same number of dots?

An analogy

- What is the probability that two fair dice show the same number of dots?
- What happens to this probability if the dice are loaded?

The probability of parallel evolution

H.A. Orr, Evolution **59**, 216 (2005)

- n beneficial single step mutations are available
- Each mutant is characterized by its fitness advantage $s_i > 0$
- The fixation probability for the i'th mutant is $2s_i$ (Haldane 1927), hence the probability that the i'th mutant is the first to fix is given by

$$\pi_i = \frac{s_i}{\sum_{j=1}^{n} s_j}$$

  and the same mutation is fixed in k replicate populations with probability

$$P_k = \sum_{i=1}^{n} \pi_i^k$$

- This is a random variable determined by the distribution of beneficial fitness effects (DBFE)

The extreme value hypothesis

- Gillespie 1983, Orr 2002: Because viable organisms are already very well adapted the DBFE can be described by extreme value theory (EVT)

- Any distribution falls into one of three EVT classes:
  - Weibull with bounded tails
  - Gumbel with exponential-like unbounded tails (also normal distribution)
  - Fréchet with power-law like heavy tails: $\mathrm{Prob}[s > x] \sim x^{-\alpha}$

- For the Weibull and Gumbel classes all moments of the DBFE exist and

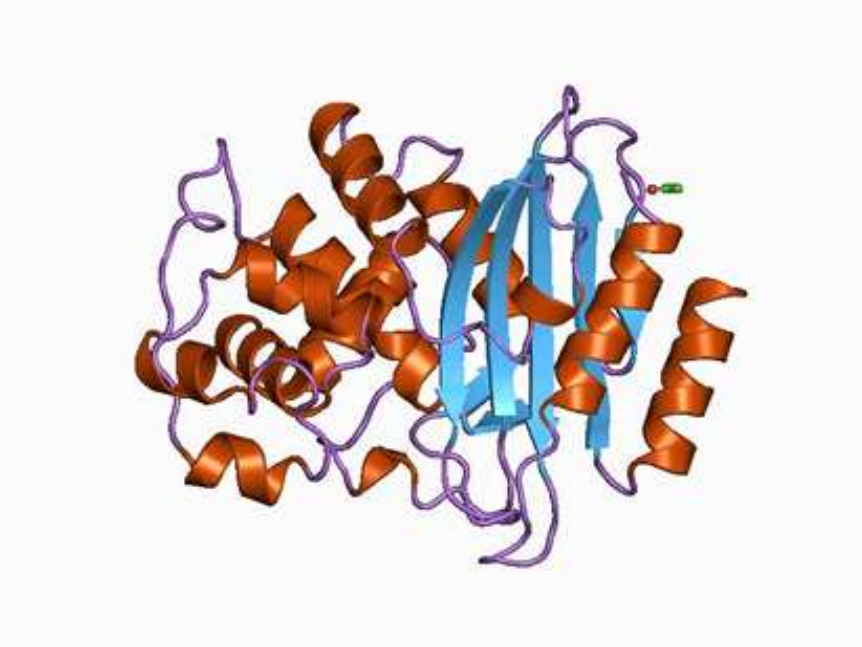
$$P_k \to \frac{n\langle s_i^k \rangle}{(n\langle s_i \rangle)^k} \sim \frac{1}{n^{k-1}}$$

  for large n, which is fully determined by the DBFE

- For the Fréchet class moments of order $k > \alpha$ do not exist

Empirical example: An antibiotic resistance enzyme

M.F. Schenk, I.G. Szendro, JK, J.A.G.M. de Visser, PLoS Genet. (2012)



- β-lactam antibiotics such as penicillin target cell wall synthesis
- TEM-1 β-lactamase confers resistance by hydrolyzing the enzyme, but has low activity against the novel antibiotic cefotaxime
- At least 48 out of 2538 point mutations increase resistance against cefotaxime, and the effect sizes fall into the Fréchet EVT class with $\alpha \approx 1$

Unpredictable repeatability

S.G Das & JK, PNAS 2022

- For $k > \alpha$ the probability of parallel evolution P_k is non-self-averaging

  Niwa 2022

- Case I: Moderately heavy tailed distributions ($\alpha > 1$)
  - Fluctuations in P_k remain of the same order as the mean even for $n \to \infty$
  - Mean and typical value of P_k show different scaling with n:

$$\langle P_k \rangle \sim n^{-(\alpha-1)}, \quad P_k^{\text{typ}} \sim n^{-k(1-\alpha^{-1})}$$

- Case II: Severely heavy tailed distributions ($\alpha < 1$)
  - P_k converges to a non-degenerate random variable with Poisson-Dirichlet distribution for $n \to \infty$, with mean value

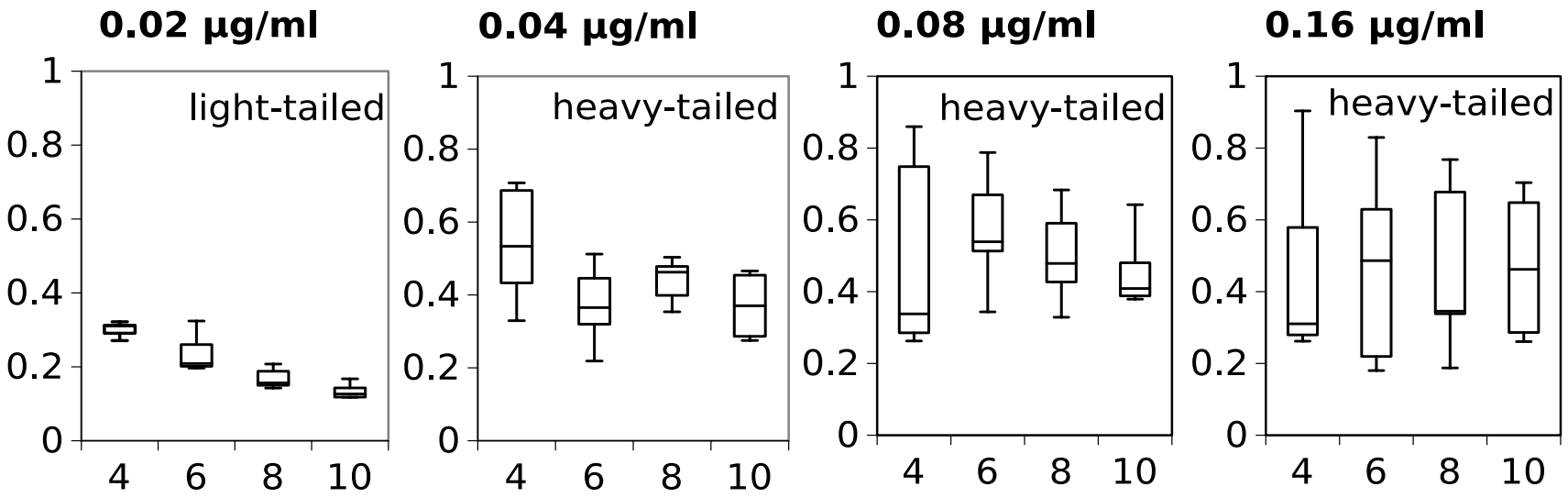
$$\langle P_k \rangle = \frac{\Gamma(k-\alpha)}{\Gamma(k)\Gamma(1-\alpha)}$$

Derrida 1994; Pitman & Yor 1997

  - In particular, $\langle P_2 \rangle = 1 - \alpha$

Application to TEM-1 β-lactamase

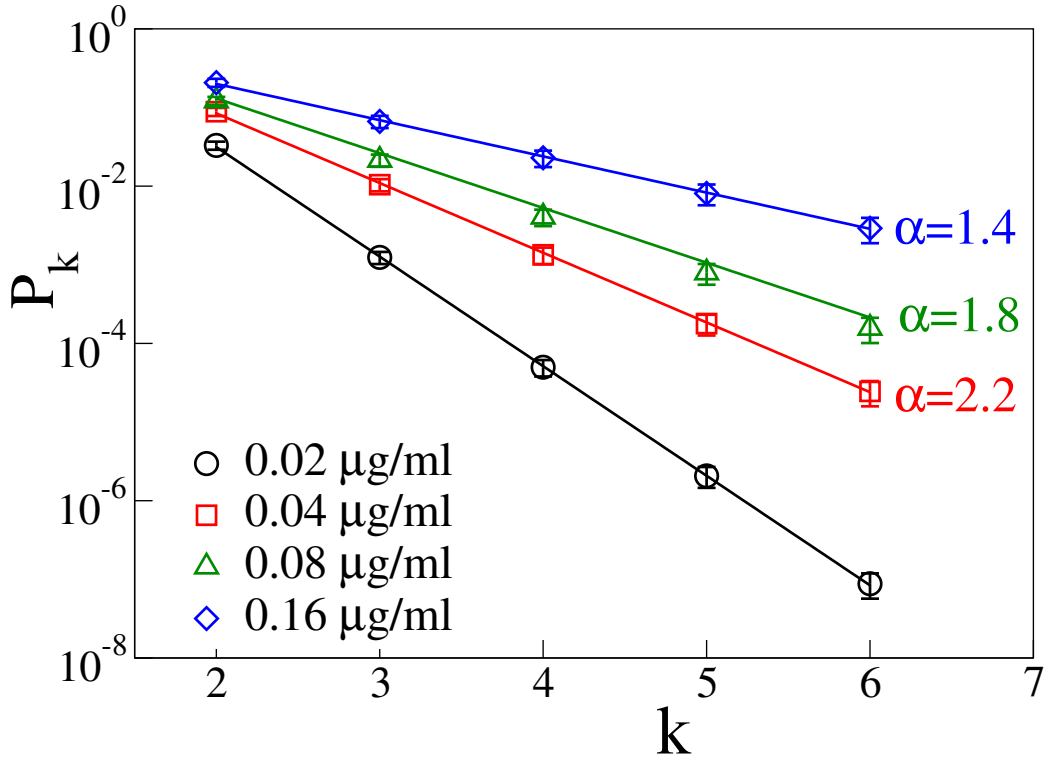
- P_2 vs. n for subsamples of the data set



- Non-self-averaging behavior for drug concentrations $\geq 0.04\,\mu$g/ml

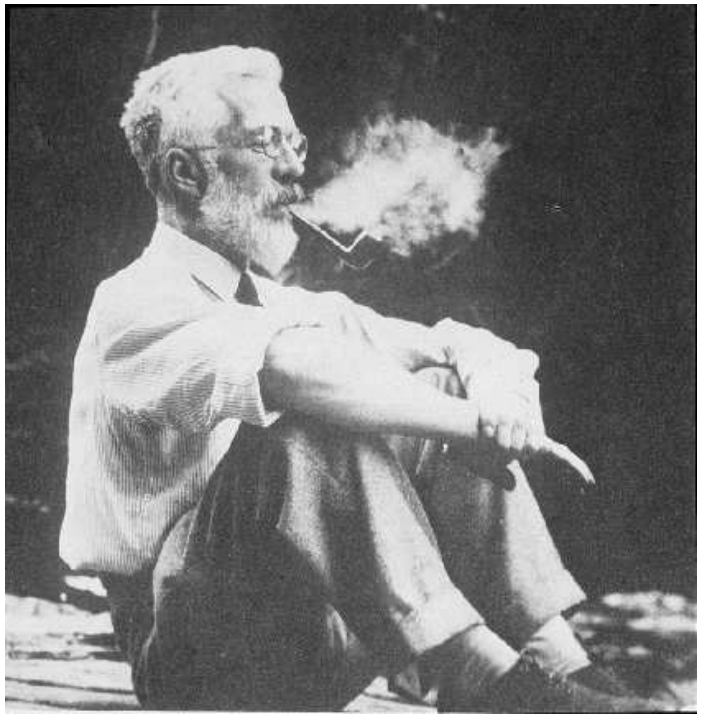
Application to TEM-1 β-lactamase

- Repeatability increases with increasing drug concentration



- Inference of α from the scaling $P_k^{\text{typ}} \sim n^{-k(1-\alpha^{-1})}$

Repeatability and population size

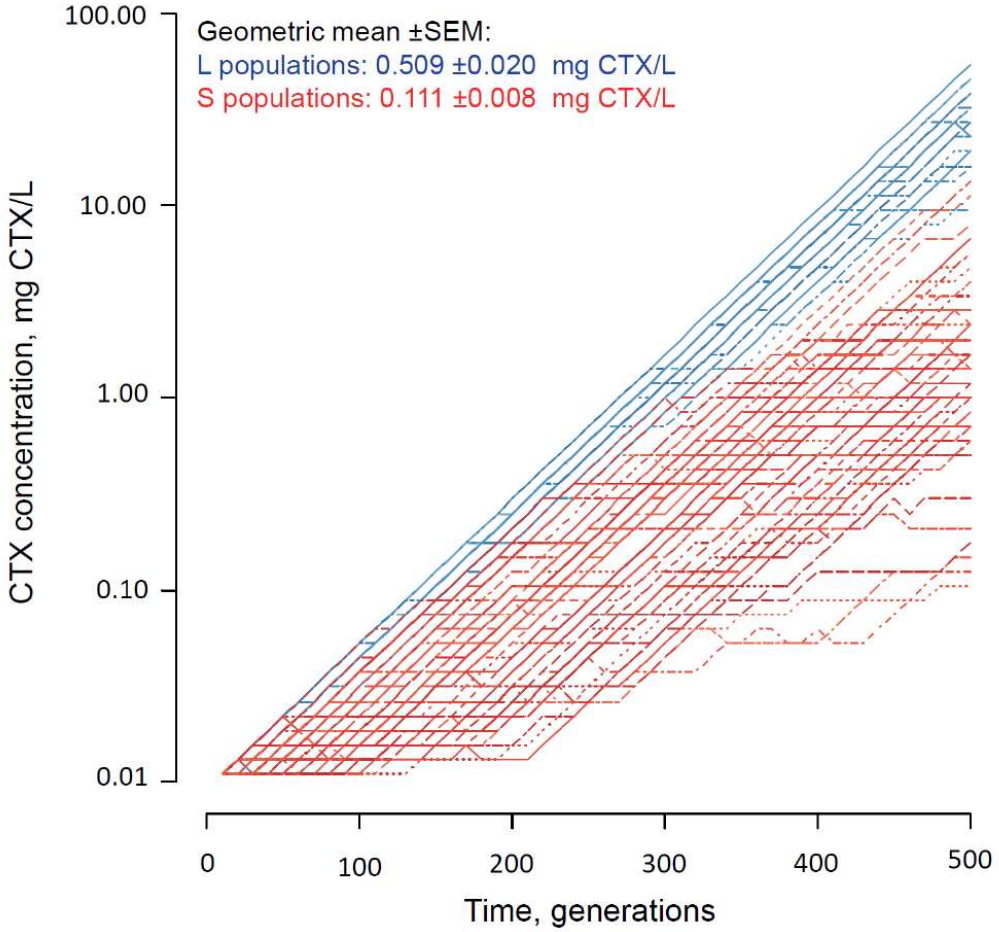


"The regularity of the [rate of adaptation] is in fact guaranteed by the same circumstance which makes a statistical assemblage of particles, such as a bubble of gas obey, without appreciable deviation, the law of gases. A visible bubble will indeed contain several billions of molecules, and this would be a comparatively large number for an organic population, but the principle ensuring regularity is the same." Ronald A. Fisher (1958)

Evolution experiment

Schenk, Zwart et al., Nat. Ecol. & Evol. 2022

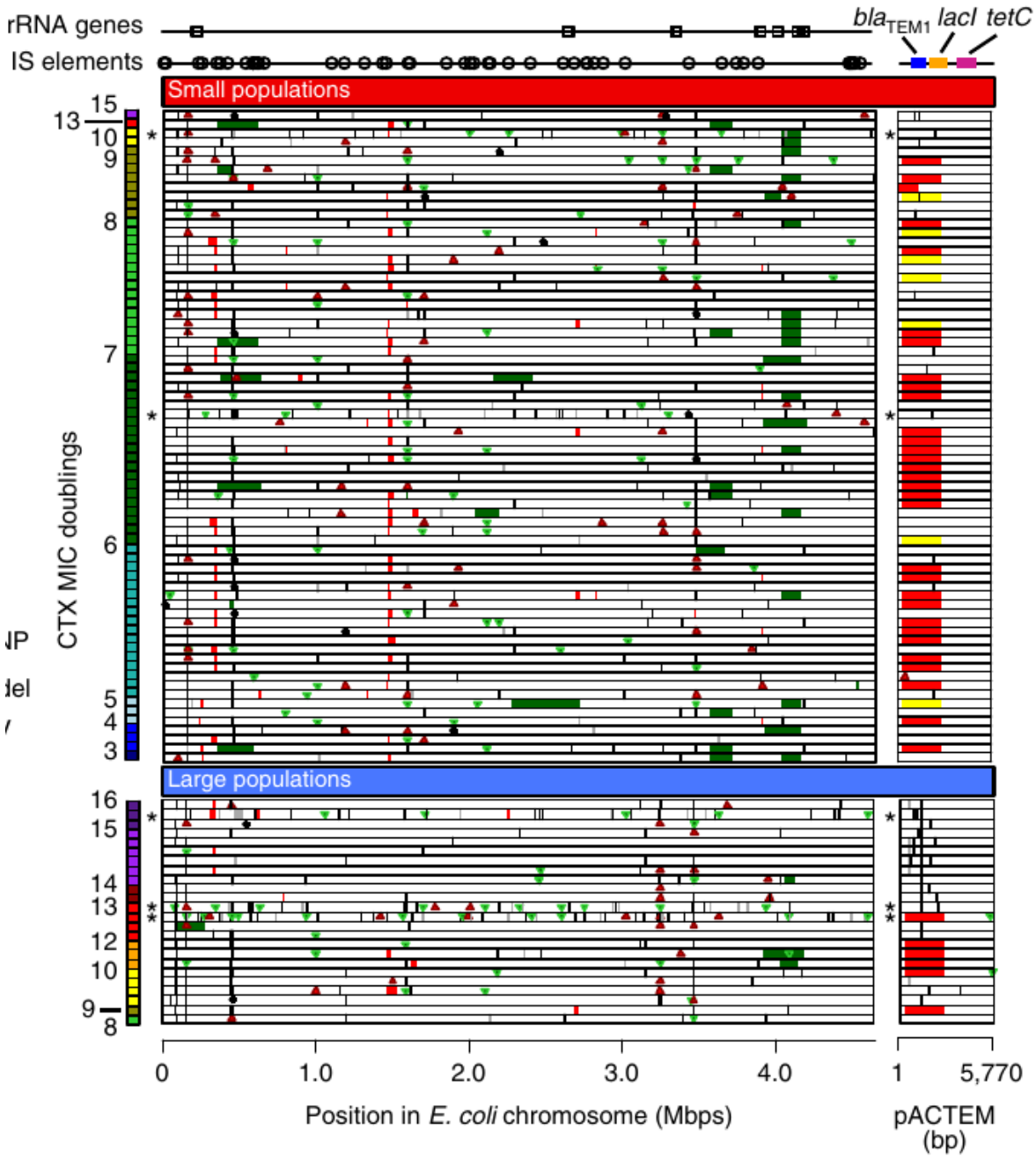
- Adaptation of *E. coli* to increasing levels of cefotaxime
- Population size $N \approx 2 \times 10^6$ (72 lines) and $N \approx 2 \times 10^8$ (24 lines)



Resistance mutations

- Bacteria carry a low activity TEM-1 β-lactamase gene on a plasmid, but resistance mutations can occur everywhere in the genome
- Large populations evolve higher levels of resistance
- Sequencing of the endpoint populations reveals 1194 mutations in plasmid and chromosome:
  - 706 point mutations (Single Nucleotide Polymorphisms, SNP's)
  - 275 small scale insertions and deletions of less than 1000 base pairs (indels)
  - 213 large scale duplications and deletions of more than 1000 base pairs (Structural Variants, SV's)
- TEM-1 is often deleted from the plasmid unless rescued by a point mutation

Mutations in endpoint populations



Repeatability index

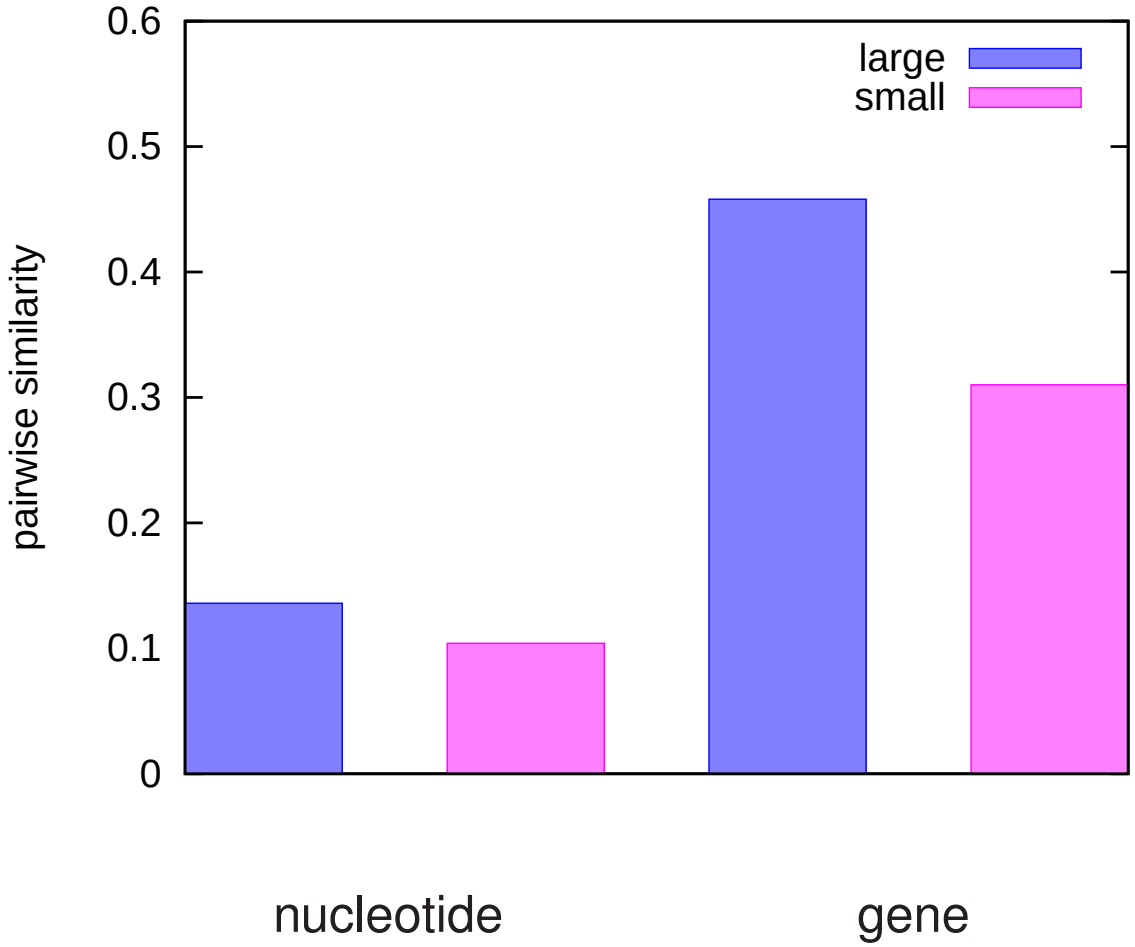
- Pairwise comparison between genotypes A and B with m and n mutations

$$H_{\mathrm{A,B}} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} \frac{|A_i \cap B_j|}{|A_i|}}{\sum_{k=1}^{m} \sum_{l=1}^{m} \frac{|A_k \cap A_l|}{|A_k|}}$$

- Symmetrized similarity measure

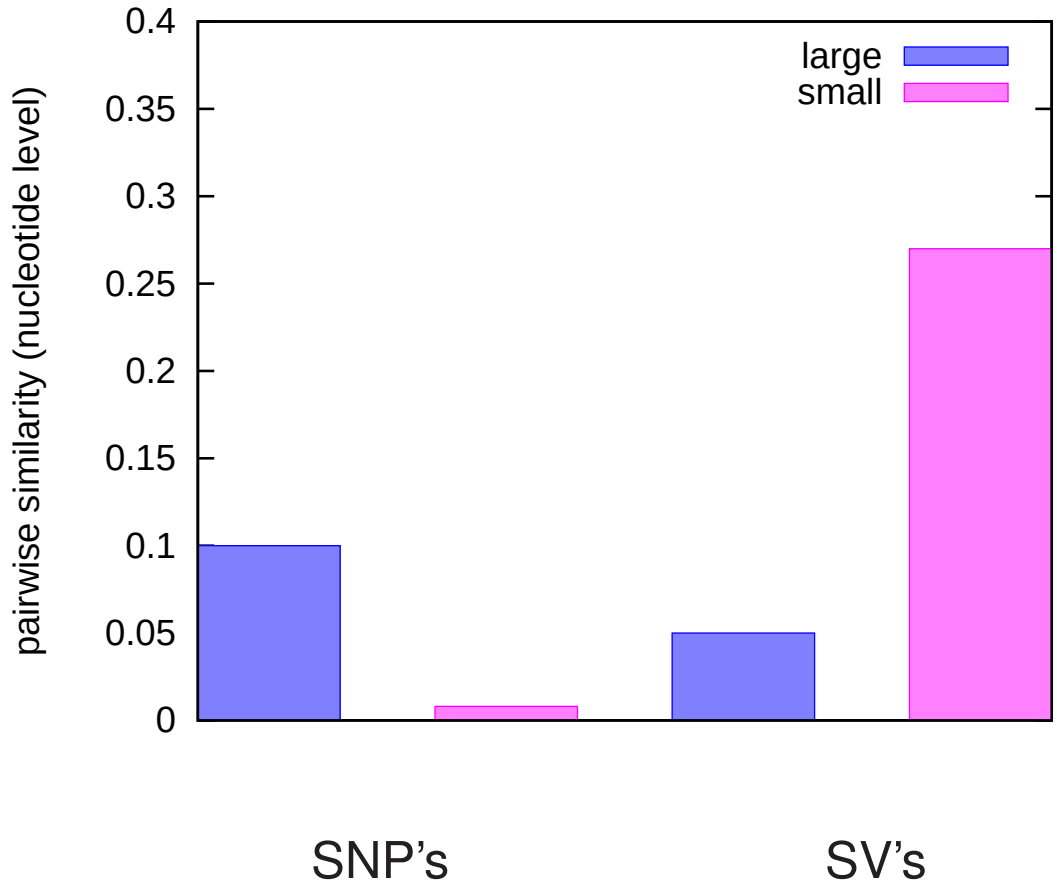
$$H = \frac{H_{\mathrm{A,B}} + H_{\mathrm{B,A}}}{2}$$

Patterns of repeatability: Population size



- Higher repeatability on gene vs. nucleotide level, and in large vs. small populations

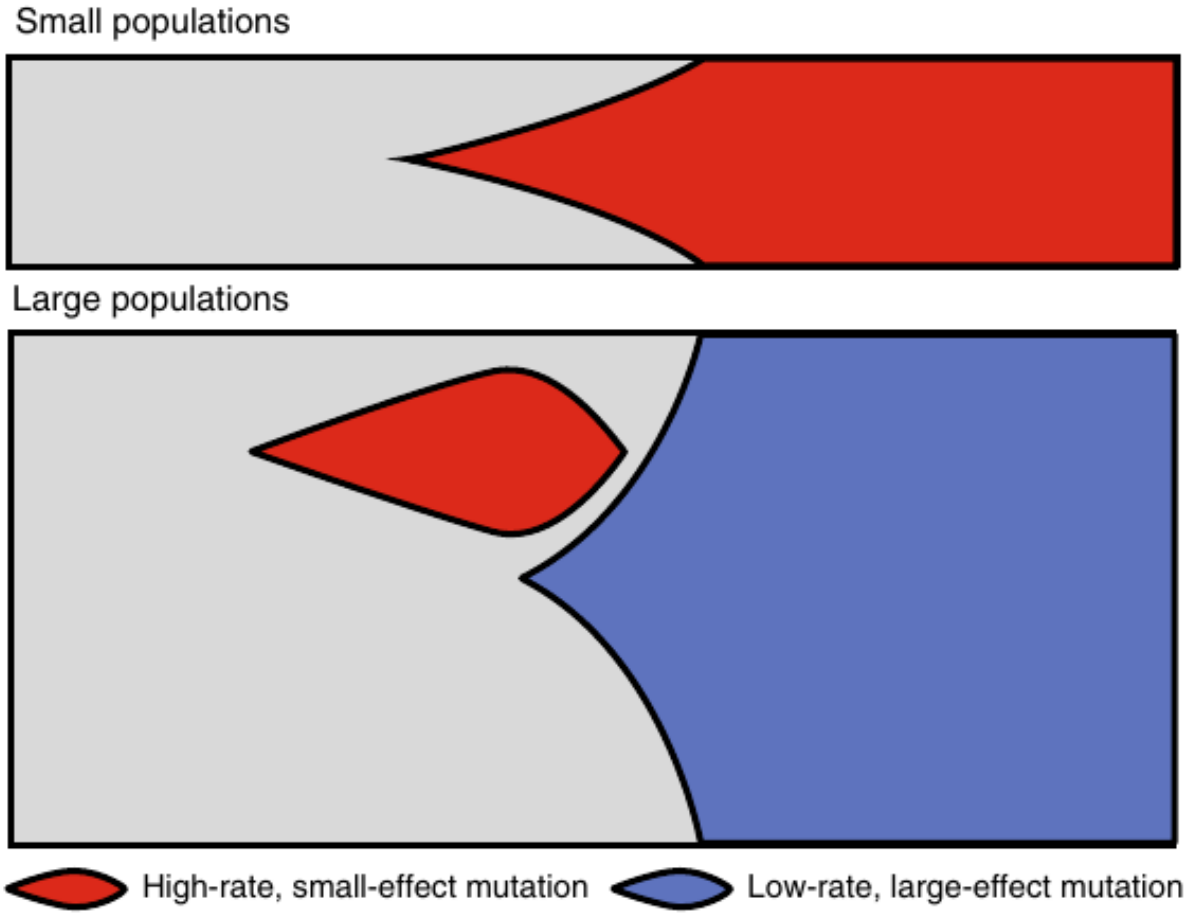
Patterns of repeatability: Mutation classes



- Different mutation classes drive repeatability in large vs. small populations

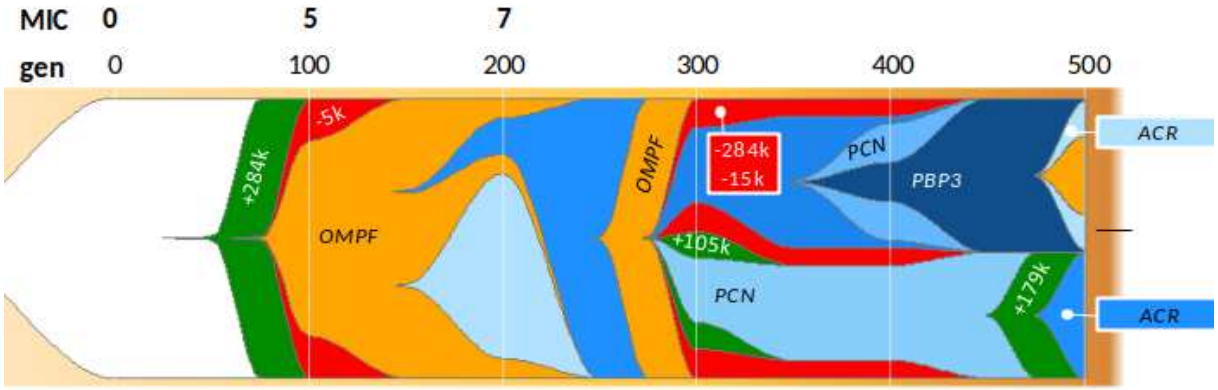
Mutation bias and population size

- Hypothesis: Clonal interference mediates a transition from SV's of high rate and small effect to SNP's of low rate and large effect

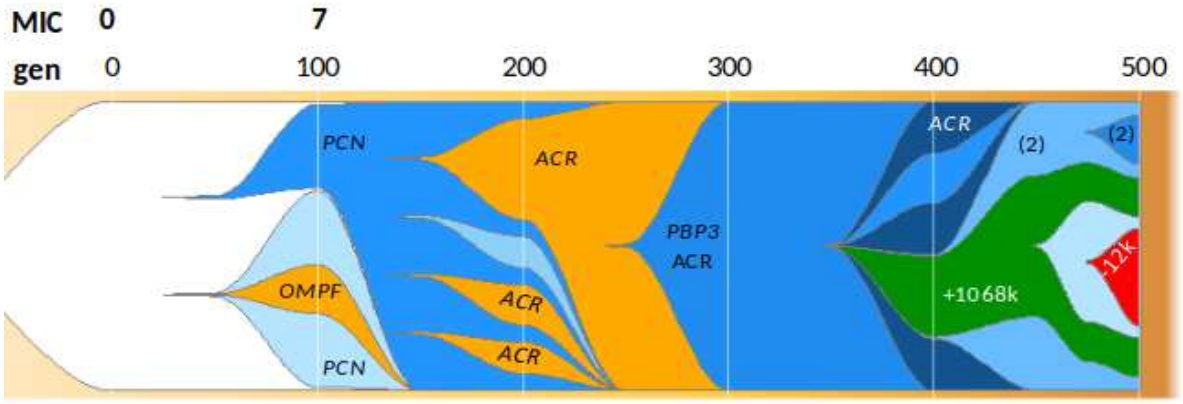


Evidence I: Time course data

- Sequencing of five small and five large populations at multiple time points
- Small populations: Large scale duplications and deletions fix first



- Large populations: SNP's fix first

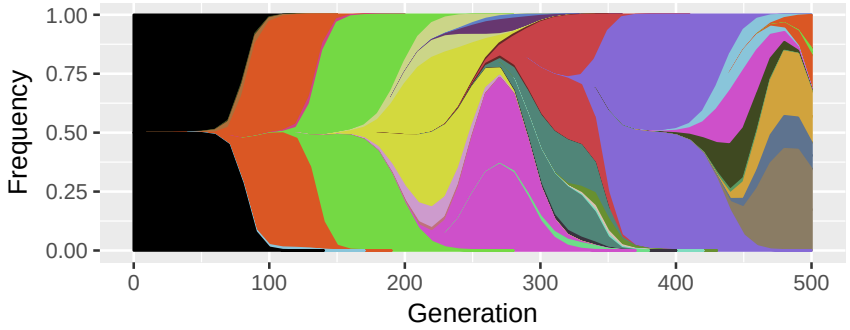


Evidence II: Computational inference

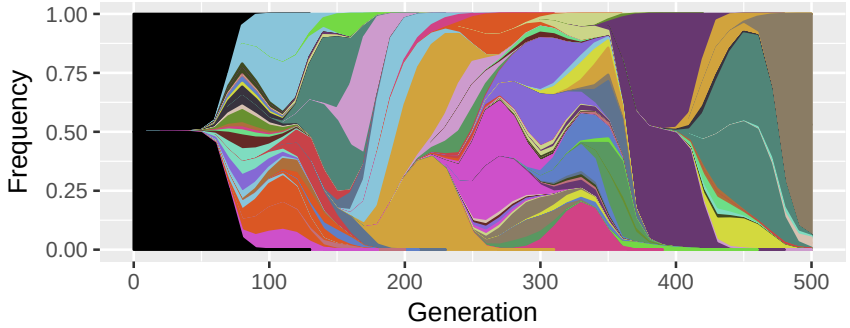
Sungmin Hwang

- Stochastic simulations with three classes of mutations with exponentially distributed fitness effects

- small population



- large population



Evidence II: Computational inference

Sungmin Hwang

- Stochastic simulations with three classes of mutations with exponentially distributed fitness effects

- Use a neural network to learn the functional relation

$$\{\mu_i, s_i\}_{i=1,2,3} \rightarrow \left\{m_i^{\text{small}}, m_i^{\text{large}}, \sigma_i^{\text{small}}, \sigma_i^{\text{large}}\right\}_{i=1,2,3}$$

  where m_i and σ_i is the mean and the standard deviation of the number of mutations of class i in the endpoint populations

- Ordering of selection coefficients and mutation rates supports the hypothesis:

$$s_{\text{SNP}} \approx 0.41 > s_{\text{indel}} \approx 0.25 > s_{\text{SV}} \approx 0.14$$

$$\mu_{\text{SNP}} \approx 2.2 \times 10^{-8} < \mu_{\text{indel}} \approx 1.8 \times 10^{-7} < \mu_{\text{SV}} \approx 7.1 \times 10^{-6}$$

Summary

Factors contributing to evolutionary predictability:

- Distribution of beneficial fitness effects
- Mutation supply, as determined by population size and mutation rate
- Mutation bias between different mutation classes
- Genetic interactions between mutations (epistasis)

Thanks to

- Suman Das (Cologne) and Sungmin Hwang (Paris)
- Arjan de Visser, Martijn Schenk, Mark Zwart (Wageningen)