ROYAL INSTITUTE OF TECHNOLOGY

Statistical genetics in and out of quasi-linkage equilibrium

Erik Aurell

BEvAS

EPFL/Lausanne, April 17-21, 2023

Google / DeepMind / AlphaFold
Andrew W. Senior *et al* ”Improved protein structure prediction using potentials from deep learning”, *Nature* **577**:706-710 (2020)

ROYAL INSTITUTE OF TECHNOLOGY

"Considerable progress has recently been made by leveraging genetic information. It is possible to infer which amino acid residues are in contact by analysing covariation in homologous sequences, which aids in the prediction of protein structures"

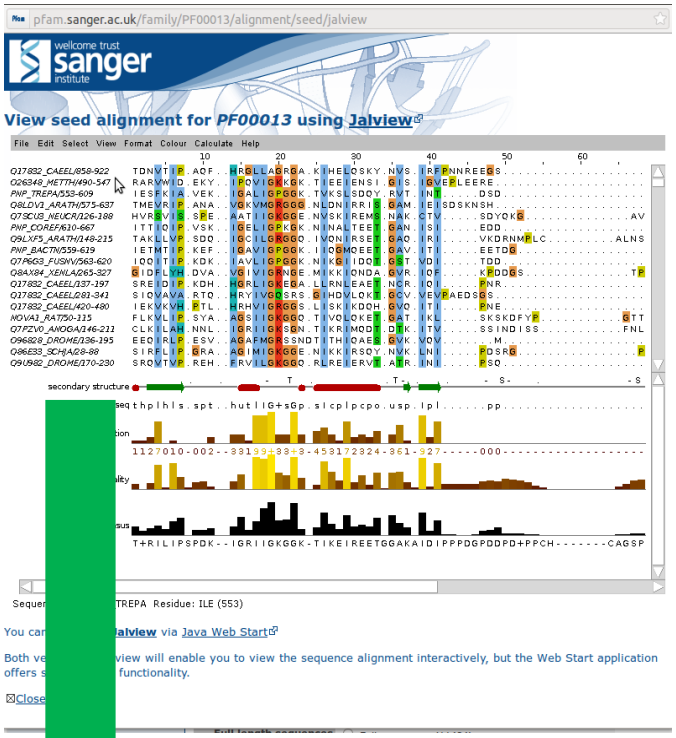
Andrew W. Senior *et al Nature* **577**:706-710 (2020) [abstract]

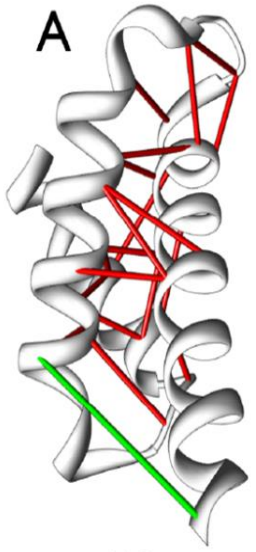
The talk is about this earlier class of methods.

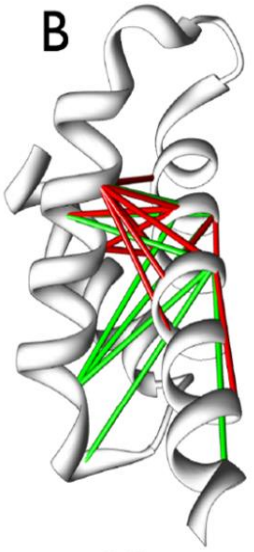
Collectively known as Direct Coupling Analysis (DCA).

In statistics one would say parameter inference in an exponential model family.

DCA in a nutshell







direct coupling analysis (DCA)

ranking by correlations

Weigt *et al*, *PNAS* 2009

⋮

Morcos *et al*, *PNAS* 2011

⋮

many others

⋮

bit.ly/3Mr8351

⋮

(*courtesy S. Ovchinnikov lab*)

$$P(\mathbf{x}) = \frac{1}{Z(h,J)} \exp\left(\sum_i h_i(x_i) + \sum_{ij} J_{ij}(x_i, x_j) \right)$$

1st main method: elements of *inverse* correlation matrix

$$E(s) = \sum_i h_i S_i + \sum_{ij} J_{ij} S_i S_j \qquad P^{\text{trial}}(s) = \Pi_i P_i(S_i)$$

$$F^{nMF} = \sum_i H\left(\tfrac{1+m_i}{2}\right) + H\left(\tfrac{1-m_i}{2}\right) + \sum_i h_i m_i + \sum_{ij} J_{ij} m_i m_j \qquad H(x) = -x \log x$$

$$\frac{\partial F^{nMF}}{\partial m_i} = 0 \quad \Longrightarrow \quad m_i = \tanh(h_i^{nMF} + \sum_j J_{ij} m_j)$$

$$\chi_{ij} = \frac{\partial m_i}{\partial h_j} = c_{ij} \qquad \text{An exact result, but used in } nMF \text{ approximation.}$$

$$\left(\chi^{nMF}\right)^{-1}_{ij} = \frac{\partial h_i^{nMF}}{\partial m_j} \approx \left(c^{-1}\right)_{ij} \quad \Longrightarrow \quad \left(c^{-1}\right)_{ij} \approx \tfrac{1}{1-m_i^2} 1_{ij} - J_{ij}$$

mean-field DCA: Morcos et al *PNAS* (2011) [M Weigt] + many later contributions theory in Kappen & Spanjers *Phys. Rev. E* (2001) and in Nguyen, Berg & Zecchina (2017)

2nd main method: pseudo-likelihood maximization

Maximum likelihood $P(\mathbf{S}) = \frac{1}{Z(\mathbf{h},\mathbf{J})} \exp\left(\sum_i h_i S_i + \sum_{ij} J_{ij} S_i S_j \right)$

$$\Pr(\mathbf{S}^{(1)},\ldots,\mathbf{S}^{(n)};\mathbf{h},\mathbf{J}) = P(\mathbf{S}^{(1)};\mathbf{h},\mathbf{J}) \cdots P(\mathbf{S}^{(n)};\mathbf{h},\mathbf{J})$$

$$\mathbf{h}^*, \mathbf{J}^* \in \arg\max \left[\sum_{ij} h_i \frac{1}{n} \sum_{s=1}^{n} x_i^{(s)} + \sum_{ij} J_{ij} \frac{1}{n} \sum_{s=1}^{n} x_i^{(s)} x_j^{(s)} - \log Z(\mathbf{h},\mathbf{J}) \right]$$

Pseudo-maximum likelihood (avoids computing Z):

$$P(S_r \mid S_{\setminus r}) = \exp\left(h_r S_r + \sum_l J_{rl} S_r S_l \right) \Big/ \sum_y \exp\left(h_r y + \sum_l J_{rl} y S_l \right)$$

$$h_r^{plm}, J_{rl}^{plm} \in \arg\max \left[\sum_{ij} h_i \frac{1}{n} \sum_{s=1}^{n} x_i^{(s)} + \sum_{ij} J_{ij} \frac{1}{n} \sum_{s=1}^{n} x_i^{(s)} x_j^{(s)} - f(h_r, J_{rl}, S_{\setminus r}) \right]$$

Julian Besag, *The Statistician* (1975); ***plmDCA,*** Ekeberg et al *Phys. Rev. E* (2013); ***GREMLIN***, Kamisetty et al *PNAS* (2014); ***CCMpred,*** Seemayer et al *Bioinformatics* (2014)

Why DCA today?

additive effects epistatic effects

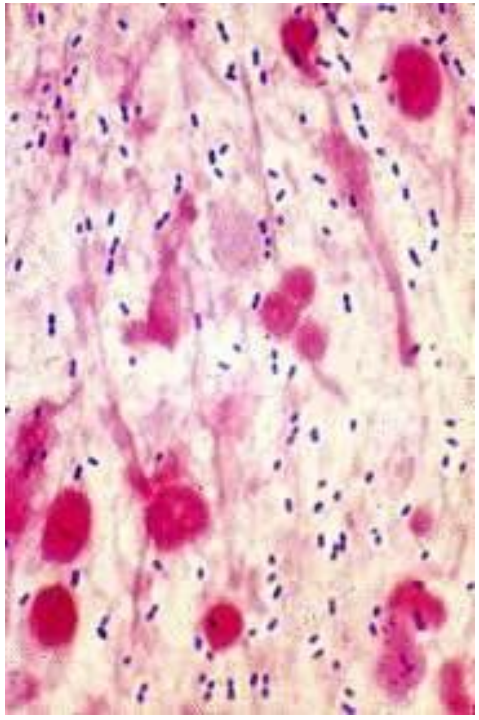
$$P(\mathbf{x}) = \frac{1}{Z(h,J)} \exp\left(\sum_i h_i(x_i) + \sum_{ij} J_{ij}(x_i, x_j) \right)$$

You may not (yet) have a large number of labelled examples on which to train a more complex AI method. **Examples:** RNA, protein-protein interactions, fitness landscapes….

Your model might be too big for deep learning. **Example:** genome scale models

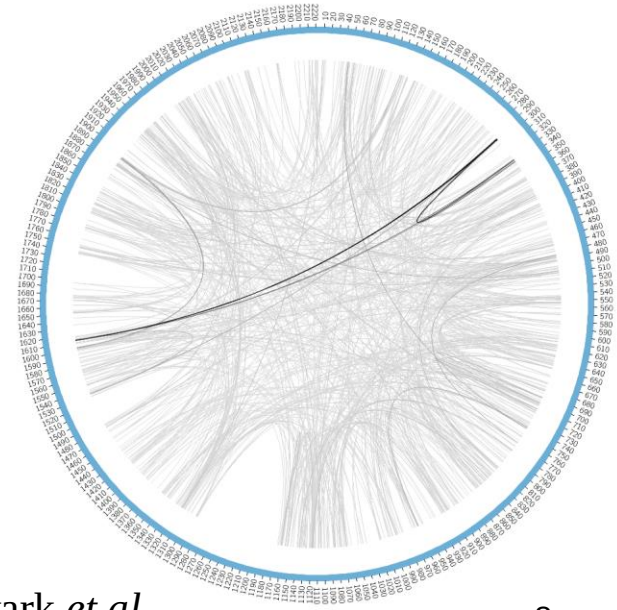
You may have a priori reasons to believe that the distribution actually is of the exponential type assumed in DCA.

A global-scale example



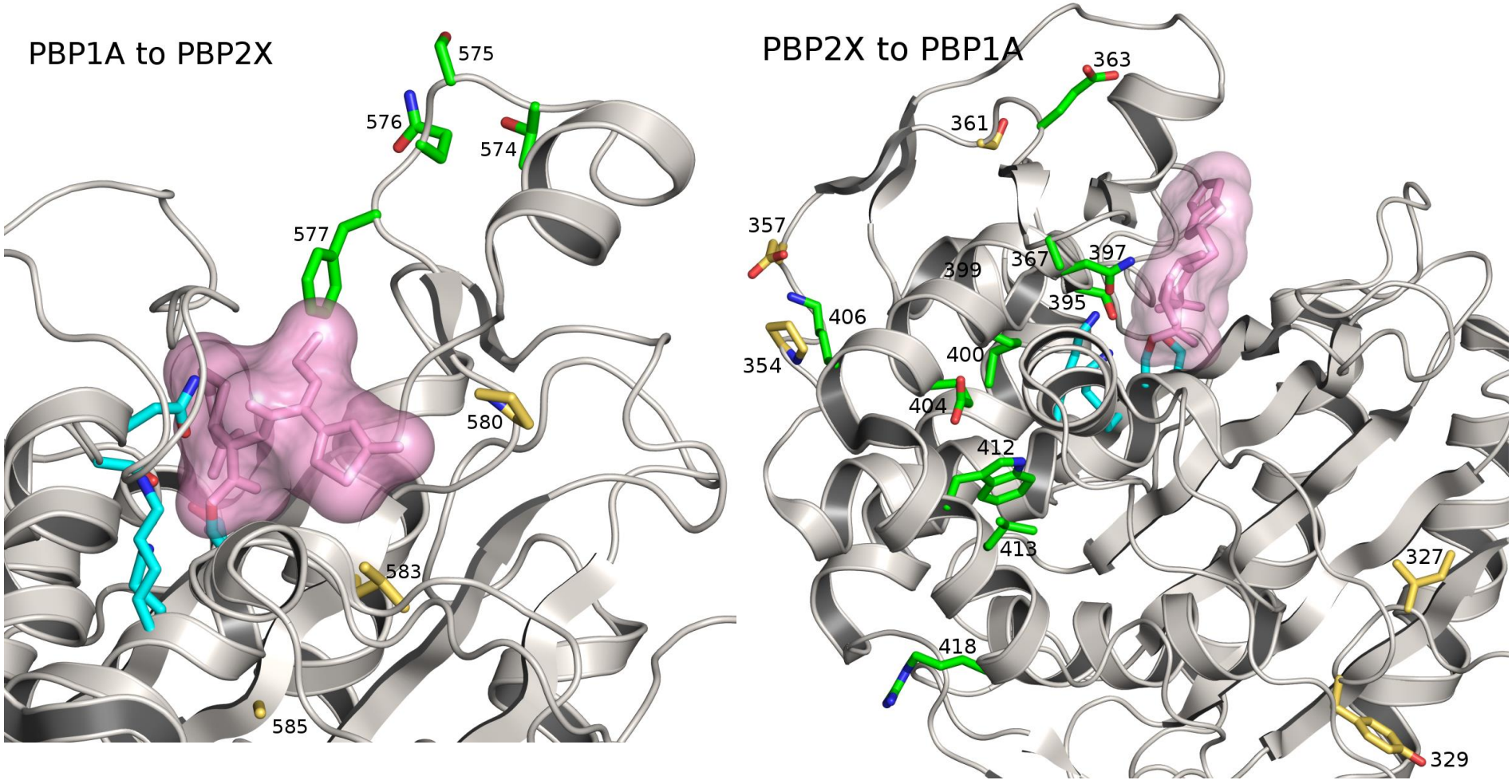
The "Maela" data set: ~3,000 genomes of *Streptococcus pneumoniae*, a bacterium with high rate of recombination.

The data had about 100,000 loci of variability, out of a genome 2.1Mbp (w/ some threshold).



Skwark *et al*
PLoS Genetics (2017)

Epistatically coupled loci in proteins in the PBP family



[purple] β-lactam; [cyan] active site; [green and yellow] groups of predictions

Some DCA on genome scale in bacteria and viruses

M. Skwark *et al*, ”Interacting networks of resistance, virulence and core machinery genes identified by genome-wide epistasis analysis” *PLoS Genetics* 2017 (*Streptococcus pneumoniae*, ”Maela” data set) (*Streptococcus pyogenes M1*)

B. Schubert, R. Maddamsetti, J. Nyman, M. R. Farhat & D. S. Marks, *Nature Microbiology* 2019 (*Neisseria gonorrhoeae*)

Cui et al. [Daniel Falush] *eLife* 2020 (*Vibrio parahaemolyticus*)

C. Chewapreecha et al [Jukka Corander], *Molecular Biology and Evolution* 2022 (*Burkholderia pseudomallei*, not quite DCA but by a similar method)

L Boeck et al [Julian Parkhill & R. Andres Floto], *Nature Microbiology* (2022) (*Mycobacterium abscessus*)

H-L Zeng et al [Erik Aurell] *PNAS* 2020 (*SARS-CoV-2*)

E Cresswell-Clay & V Periwal, *Mathematical biosciences* 2021 (*SARS-CoV-2*)

J Rodriguez-Rivas et al [Martin Weigt] *PNAS* 2022 (*SARS-CoV-2*)

ROYAL INSTITUTE OF TECHNOLOGY

Why does DCA work and when does it not?

Those are the questions of today's talk.

One can ask them for AI / learning as well. But then they are more difficult. And you have world-class experts here at EPFL, whom you can ask instead.

Note that I am not asking *if* DCA (or AI / learning) works, in many cases. It is well known by now that it does. But that is another question.

Statistical genetics

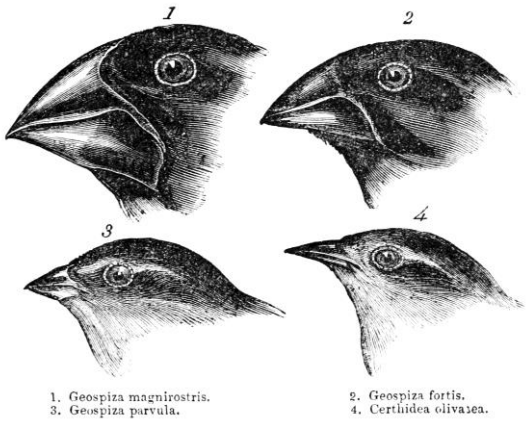
A general understanding of population genetics in analogy with statistical physics. This has a long history starting with Hardy and Weinberg (1908) and Fisher and Wright in the 1920ies and 1930ies.

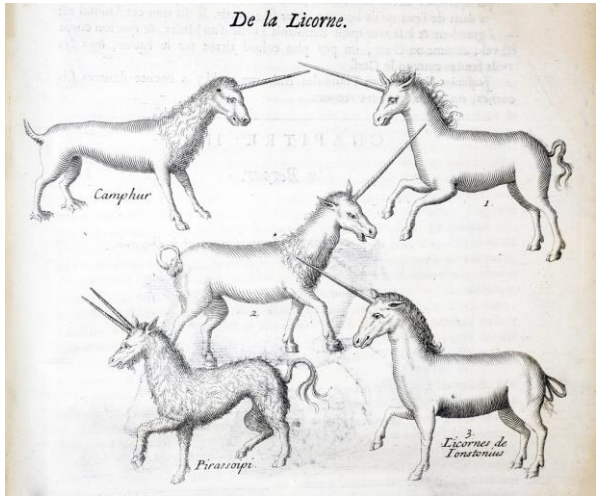
In statistical physics a goal is to deduce macroscopic properties of a body (thermodynamics) from underlying physical laws.

In statistical genetics the goal is analogously to deduce macroscopic properties of a population from the laws of evolution.

Understanding why and when DCA works from evolutionary models falls into this category of questions.

In other words: justify DCA from known laws of evolution



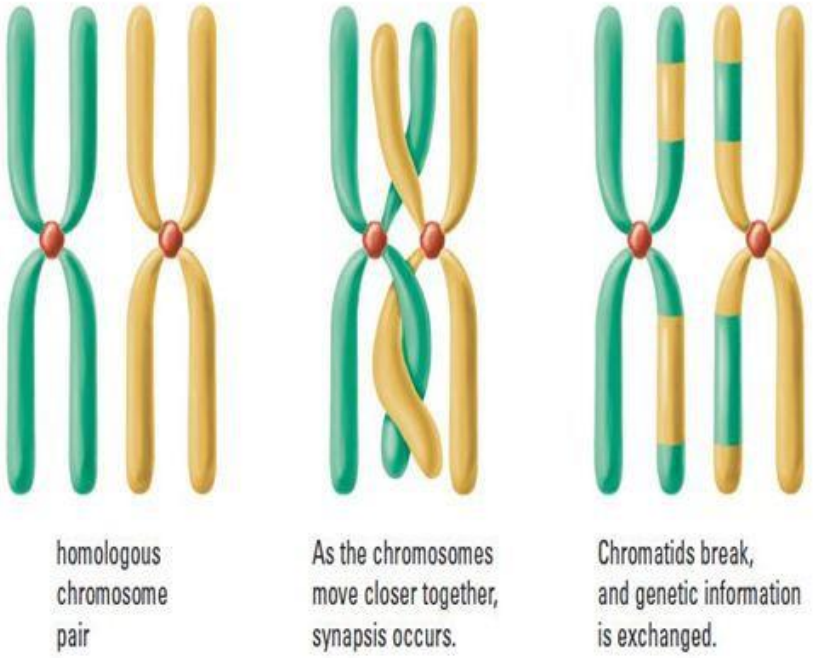


The distribution of genotypes in a population is shaped by the forces of evolution which are: (**1**) Darwinian selection (tendency to maximize fitness), (**2**) recombination, (**3**) mutations, and (**4**) genetic drift (finite-*N* effects)…

Unicorns are imagined instances of organisms which do not evolve due to effects (**1**), (**2**) or (**3**). The extinct two-horn Italian unicorn (the *Pirassoipi*) had a dense pelt. In unicorns these properties therefore disappeared together.

In recombination alleles are mixed between chromosomes

Cross-over happens 50-80 times in human during meiosis.



J Weaver, *Biotechniques* (2016)

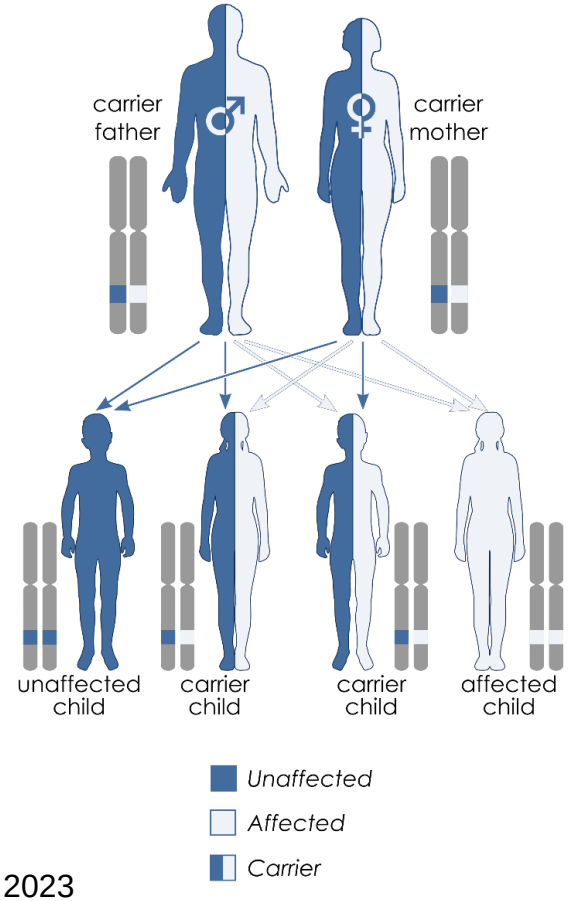
Recombination is sometimes used in the restricted sense of mixing of genetic material between two chromosomes in the same parent (cross-over).

Can also be used in the more general sense of any mixing of genetic material.

In bacteria recombination and sex are often used to mean the same thing.

In sex chromosomes are mixed between individuals

Allows completely healthy offspring from not completely healthy parents



Transformation, *transduction and conjugation* are the main forms of bacterial sex (or recombination)

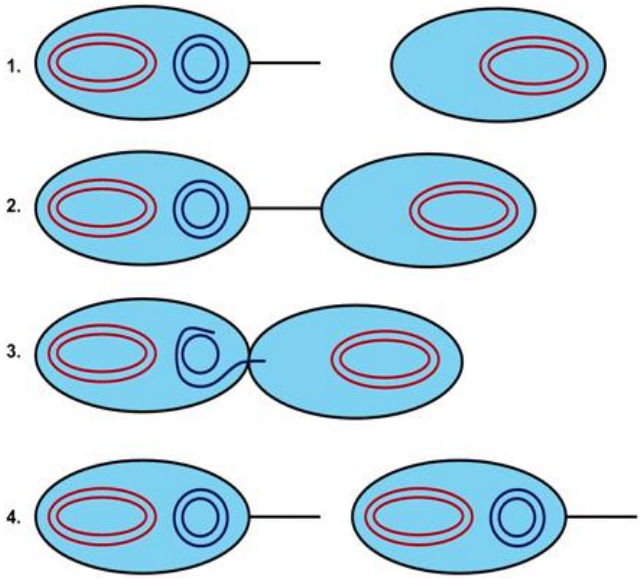


Illustration of conjugation
Raz and Tannenbaum, *Genetics* (2010)

Definitions

Linkage equilibrium: the distributions of alleles over loci are independent. Happens when recombination mix up genomes.

Linkage disequilibrium (LD): distributions at alleles are not independent. Can be due to fitness or inheritance (or both).

Formal: A population is said to be in a *quasi-linkage equilibrium* (QLE) phase if (1) multi-genome distributions factorize and (2) single-genome distributions lie in an exponential family with no higher terms than in the fitness function. Which for quadratic fitness means

$$P(x) = \frac{1}{Z(h,J)} \exp\left(\sum_i h_i(x_i) + \sum_{ij} J_{ij}(x_i, x_j) \right)$$

Kimura *Genetics* **52**:875–890 (1965)
Neher & Shraiman *PNAS* **106**:6866 (2009); *Rev Mod Phys* **83**:1283 (2011)
formal definition in Dichio, Zeng, EA (2023)

The Kimura-Neher-Shraiman theory (Neher-Shraiman version)

The distribution of genotypes in a population changes according to **selection**, **mutation**, **genetic drif**t (finite-*N*) and **recombination**.

$$\boldsymbol{g} = (s_1, s_1, \cdots, s_L) \quad s_r = \pm 1$$ "Ising genome"

$$P(\boldsymbol{g}, t+\Delta t) = \frac{e^{\Delta t F(\boldsymbol{g})}}{\left\langle e^{\Delta t F(\boldsymbol{g})} \right\rangle} P(\boldsymbol{g}, t) \quad F(g) = \sum f_i s_i + \sum f_{ij} s_i s_j$$ Fitness

$$P(\boldsymbol{g}, t+\Delta t) = P(\boldsymbol{g}, t) + \Delta t \mu \sum_i [P(M_i \boldsymbol{g}, t) - P(\boldsymbol{g}, t)]$$ Mutations

$$P(\boldsymbol{g}, t+\Delta t) = (1 - r\Delta t) P(\boldsymbol{g}, t) + \Delta t r \sum_{\boldsymbol{g}_m, \boldsymbol{g}_f} C(\boldsymbol{g}, \boldsymbol{g}_m, \boldsymbol{g}_f) P(\boldsymbol{g}_m, t) P(\boldsymbol{g}_f, t)$$

Two haploid parents copy themselves, produce a child, and the rest of both genomes is discarded. Directly appropriate for some yeasts. One can modify the above to also cover bacterial recombination.

Neher-Shraiman theory of QLE

Neher & Shraiman, *Rev Mod Phys* **83**:1283 (2011)

[for Potts not Ising] Gao, Cecconi, Vulpiani, Zhou, EA, *Phys. Biol*. **16** 026002 (2019)

Recombination is parametrized by a cross-over indicator variable $\boldsymbol{\xi}$

$$g^{(i)} = \xi_i g_m^{(i)} + (1-\xi_i) g_f^{(i)} \qquad C(\boldsymbol{g}, \boldsymbol{g}_m, \boldsymbol{g}_f) = \mathrm{C}(\boldsymbol{\xi})$$

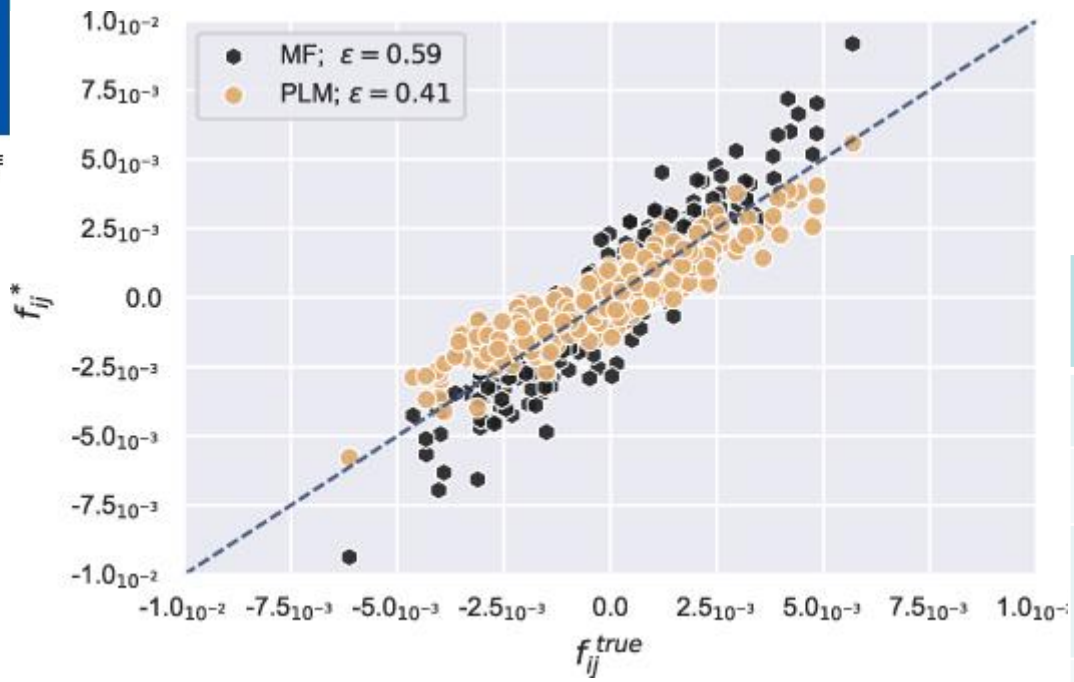
Recombination acts on pairwise dependencies through

$$c_{ij} = \sum_{\xi} C(\xi) \left[\xi_i (1-\xi_j) + \xi_j (1-\xi_i)\right]$$

Assume that $P(\boldsymbol{g})$ is initially close to a Gibbs distribution of an Ising energy function (h_i, J_{ij}) and recombination rate r is large:

$$\partial_t P(\boldsymbol{g}, t) = \cdots \Rightarrow \dot{J}_{ij} = f_{ij} - r c_{ij} J_{ij} \Rightarrow J_{ij} = \frac{f_{ij}}{r c_{ij}}$$

In steady-state QLE the Ising parameters J_{ij} are proportional to pairwise fitness parameters f_{ij}, the proportionality being $(r c_{ij})^{-1}$.



Example of a scatter plot for the reconstructed epistatic fitness components f_{ij}^{*} (y-axis) versus true underlying parameters f_{ij} (x-axis).

MF (mean-field) and PLM (pseudo-likelihood maximization) versions of DCA give similar reconstruction performance.

Simulation parameters of **FFPopsim** [Zanini and Neher *Bioinformatics* **28** 3332–3 (2012)]

	Value	Description
N	200	n. individuals
L	25	n. of loci
T	2.5 x 10^3	n. of generations
ω	0.5	crossover rate
r	[0.0:1.0]	rate of recombination
μ	[0.005 :0.1]	rate of mutation
σ_e	[0.001 :0.02]	$f_{ij} \sim \mathcal{N}(0, \sigma_e)$

$$f_{ij}^{*} = r \cdot c_{ij} \cdot J_{ij}^{*} \qquad c_{ij} \approx \frac{1}{2}$$

Mauri-Zeng-Dichio-Aurell-Cocco-Monasson revised theor(ies)

Derived by a Gaussian closure on moments, but can also be done similarly to the Neher-Shraiman analysis. Several levels of inference formulae were found, out of which I will here only use the simplest (which NB bi-passes the need for DCA)

$$f_{ij}^{*} = \chi_{ij} \cdot \frac{4\mu + r c_{ij}}{\left(1-\chi_i^2\right)\left(1-\chi_j^2\right)} \quad \chi_i = \langle s_i \rangle \quad \chi_{ij} = \langle s_i s_j \rangle - \chi_i \chi_j$$

Note the presence of mutation rate μ. The formula reduces to Kimura-Neher-Shraiman in the small-coupling regime and in the limit when μ tends to zero.

Mauri, Cocco, Monasson, *Europhys Lett* **132** 56001 (2021)
Zeng, Mauri, Dichio, Cocco, Monasson, EA *JSTAT* 2021 083501 (2021)

KNS vs MZDACM

Regression of inferred epistasis (f^*_{ij}) on underlying "true" epistasis (f_{ij}).

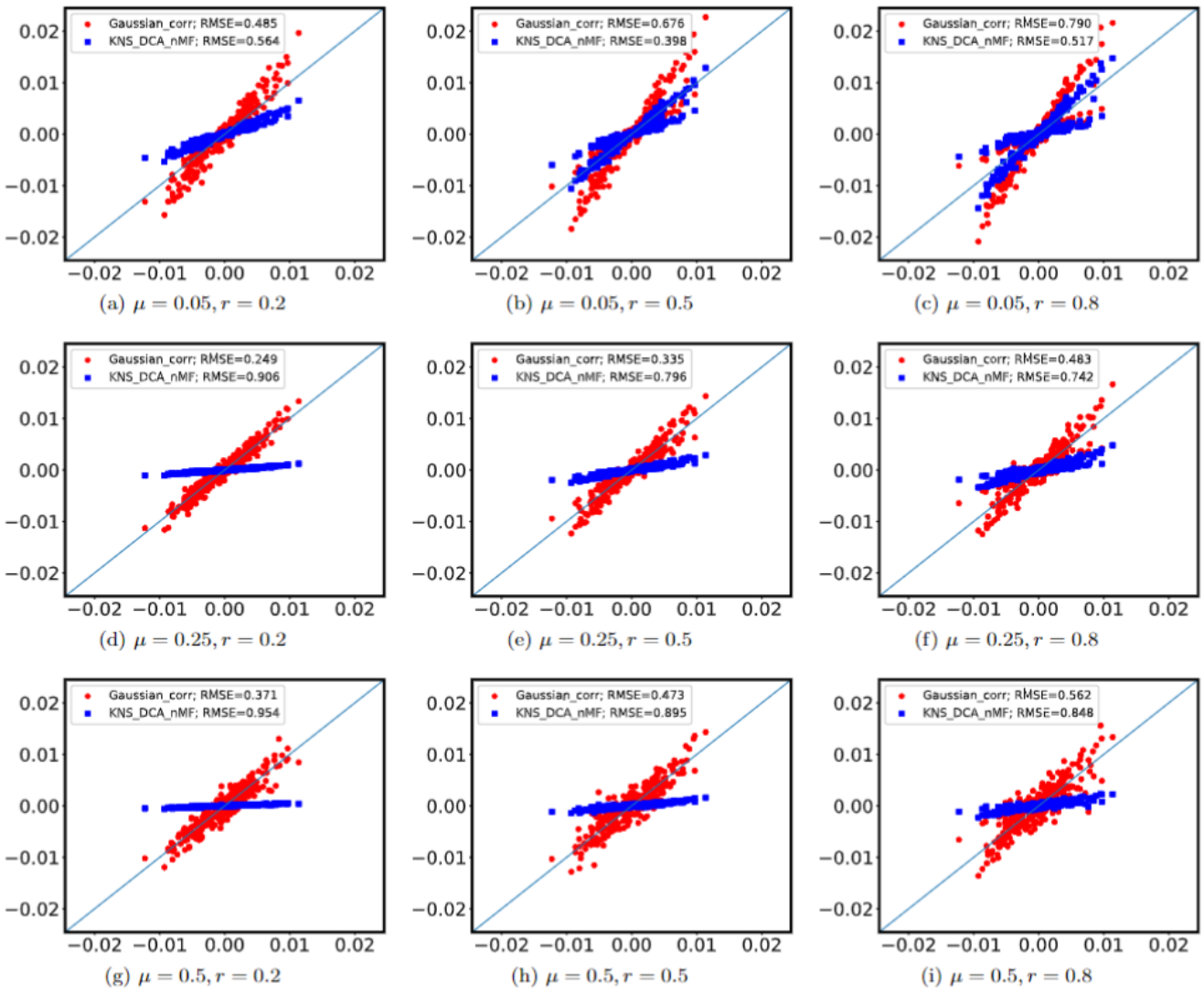
Comparison of the **KNS** formula:

$$f^*_{ij} = r \cdot c_{ij} \cdot J^*_{ij},$$

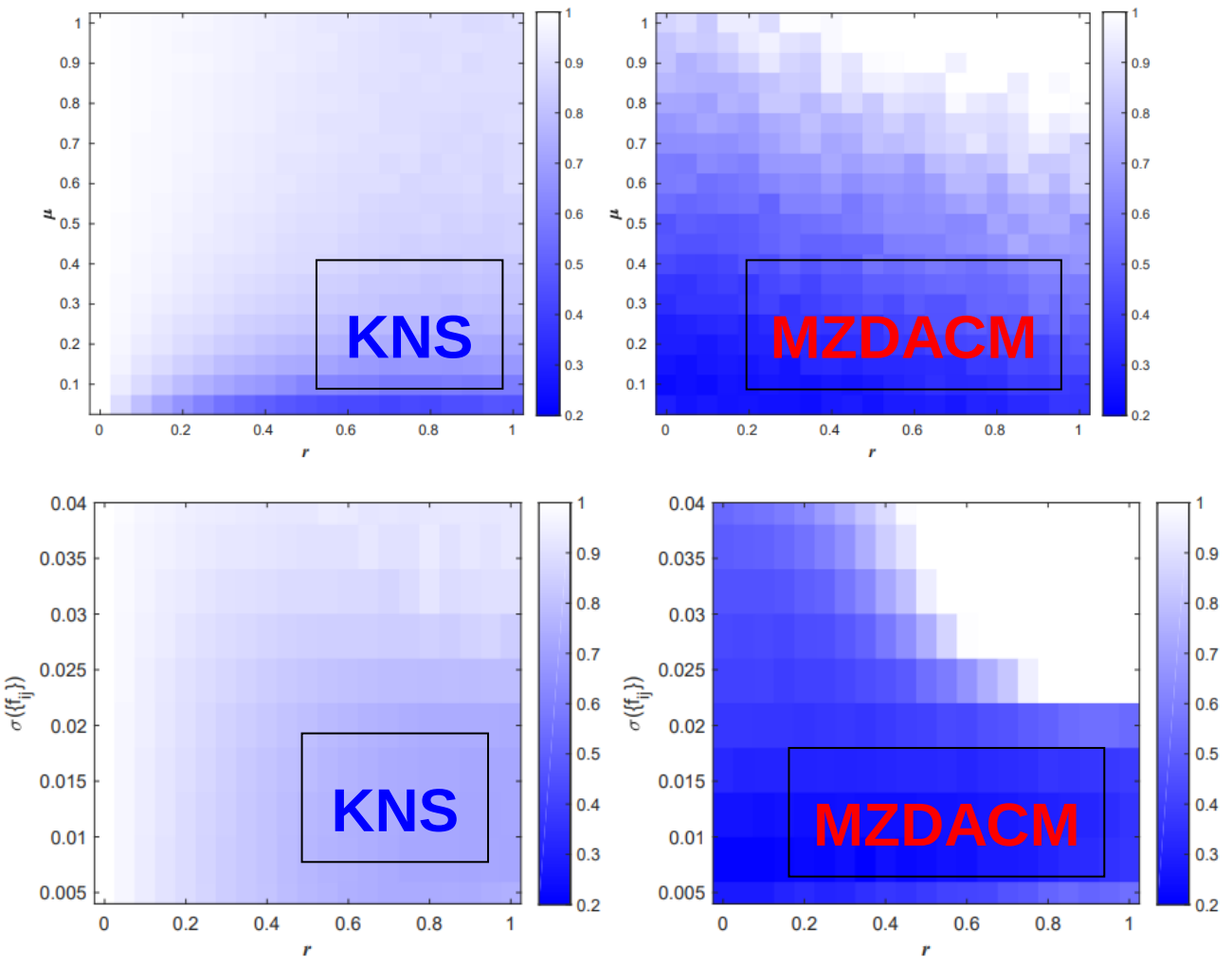
and the **MZDACM** formula;

$$f^*_{ij} = \frac{x_{ij} \cdot (4\mu + r c_{ij})}{(1 - x_i^2)(1 - x_j^2)}.$$

Zeng *et al JSTAT* 083501 (2021)



Performance phase diagrams



μ vs r at random additive fitness $\sigma_a = 0.05$ and random epistatic fitness $\sigma_e = 0.004$. One realization for each parameter.

σ_e vs r at mutation rate $\mu = 0.2$.

For other parameters, see paper.

Zeng *et al JSTAT* 083501 (2021)

ROYAL INSTITUTE OF TECHNOLOGY

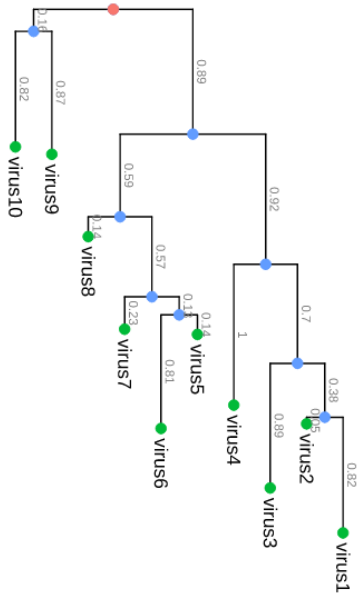
Loss of QLE

Rep. Prog. Phys. **86** 052601 (2023) [arXiv:2105.01428]

and a brief review of earlier work

QLE vs clonal competition

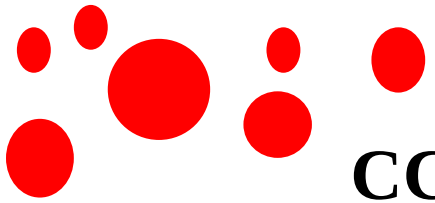
Neher & Shraiman *PNAS* **106**:6866 (2009); *Rev Mod Phys* **83**:1283 (2011);
Neher, Vucelja, Mézard, Shraiman *JSTAT* 01008 (2013)



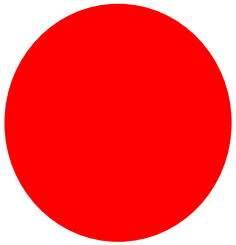
Mixture models
Multi-genome distributions are complex

$\mu \approx 0$

Exponential models
Multi-genome distributions factorize



$r^* \sim \sigma\sqrt{\log N}$



CC **QLE** r

recombination strength in units of fitness variation

Clonal competition *Quasi-linkage equilibrium*

$$P(\boldsymbol{g}) \approx \sum \mu_c P_c(\boldsymbol{g})$$

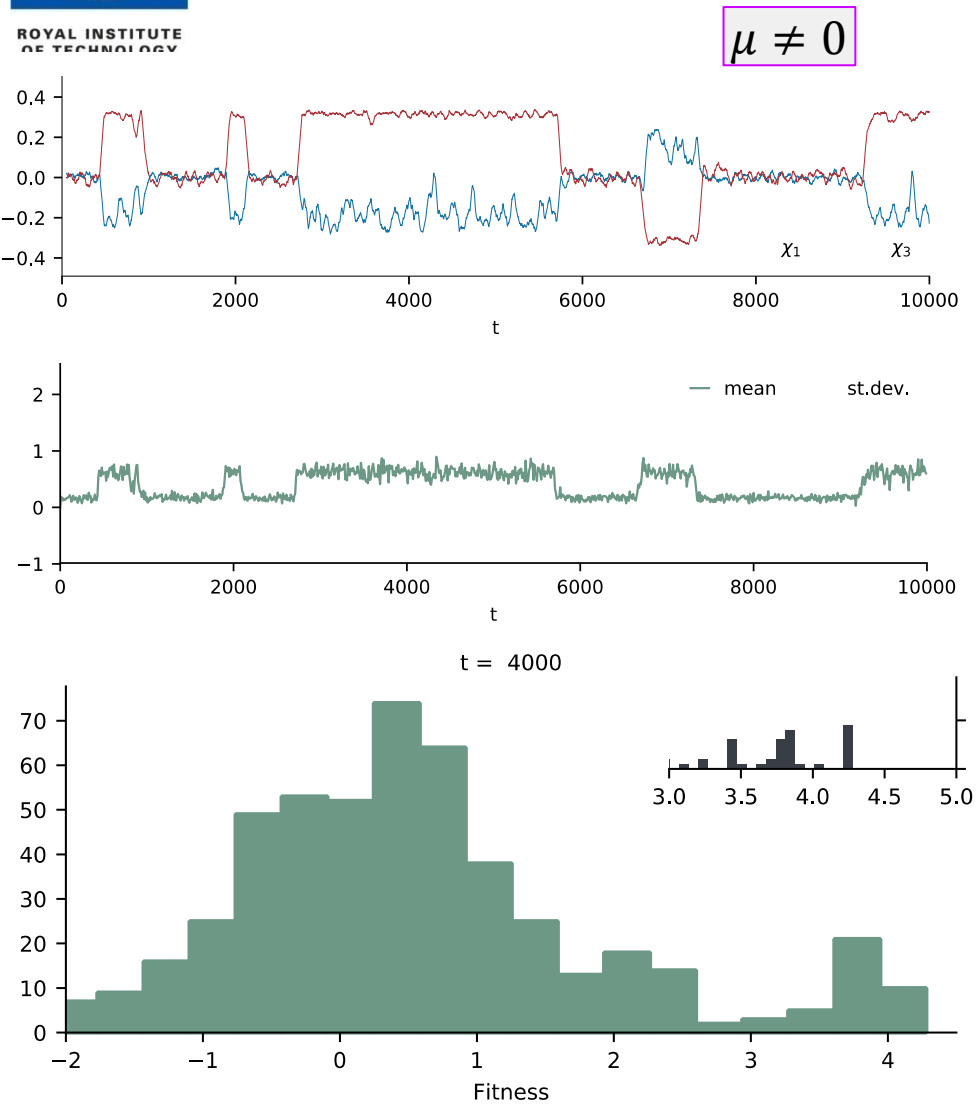
$$P(\boldsymbol{g}_1, \boldsymbol{g}_2) \approx P(\boldsymbol{g}_1) \cdot P(\boldsymbol{g}_2)$$

$$P(\boldsymbol{g}) \sim e^{\sum h_i(g_i) + \sum J_{ij}(g_i, g_j)}$$

$$P(\boldsymbol{g}_1, \boldsymbol{g}_2) \approx P(\boldsymbol{g}_1) \cdot P(\boldsymbol{g}_2)$$

At $N = \infty$ there is no QLE! However, $\sqrt{\log \mathcal{N}_{avo}} \approx 7{,}4\ldots$

Non-random coexistence



At finite mutation rate the loss of QLE manifests itself differently. For finite populations appears an intermittent regime fluctuating between QLE and Non-Random Coexistence (NRC).

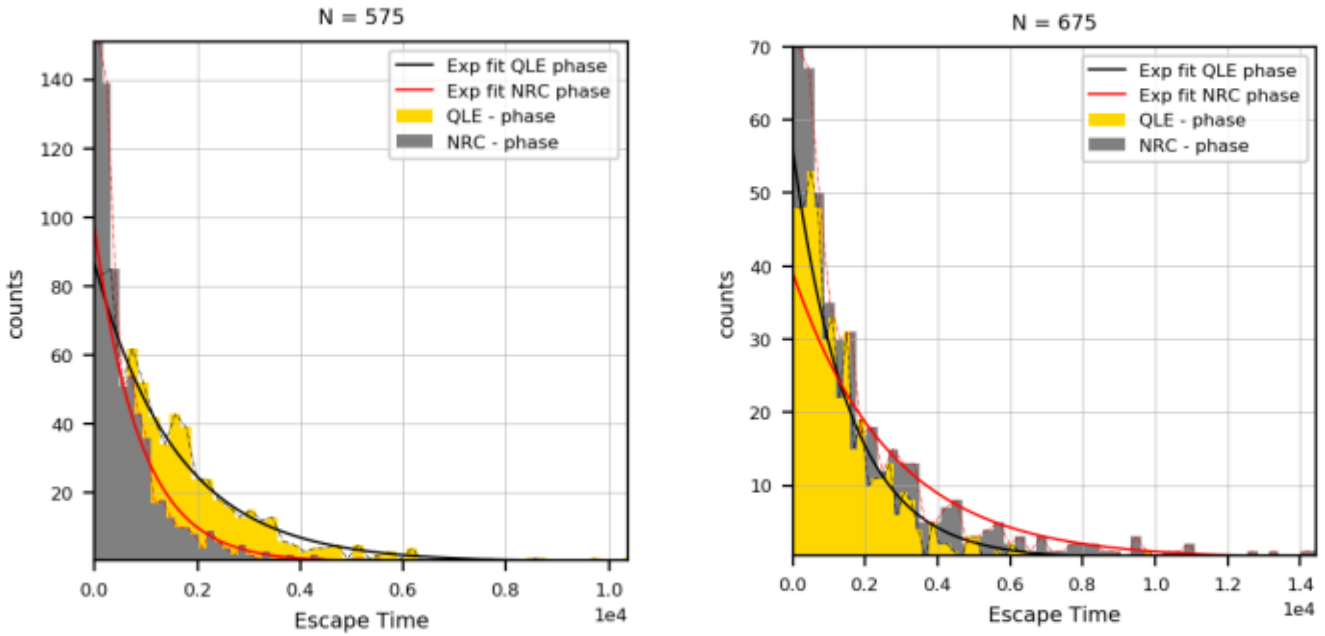
Total mean fitness in the population fluctuates, and is higher in NRC.

Snapshot of the fitness distribution at $t = 4000$ in the above (NRC interval). Differently to QLE, the distribution is bimodal with a group of individuals at high fitness.

Similar to predictions in CC, though here no exact clones, due to mutations.

ROYAL INSTITUTE OF TECHNOLOGY

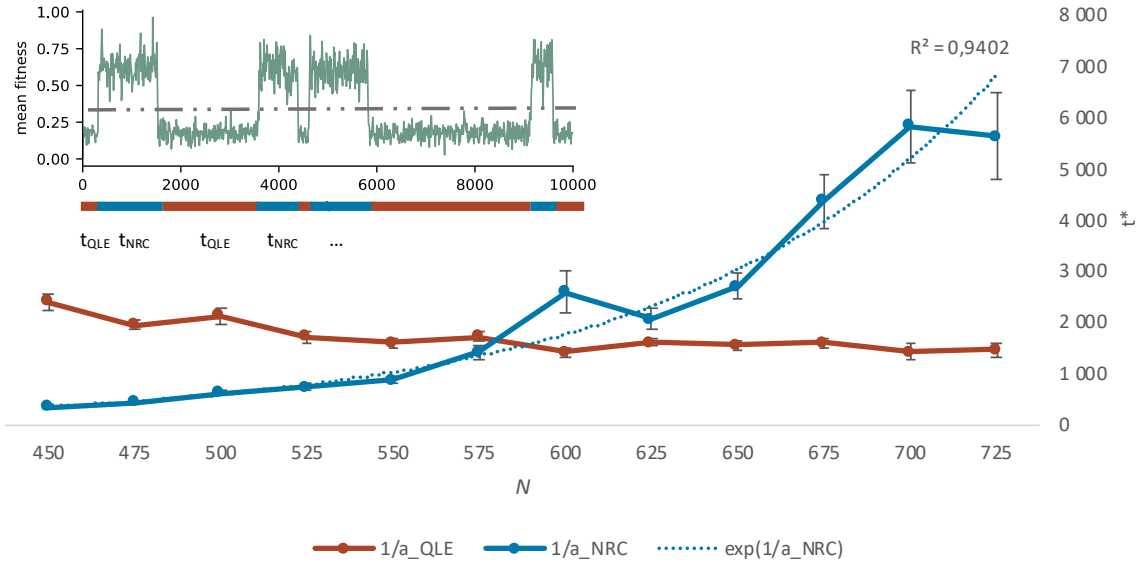
Escape time distributions



Empirical distribution of escape times from respectively QLE and NRC. Simulations are run in a region of the parameter space (including N, here 575 and 675) where the systems dynamics visually jumps back and forth between QLE and NRC. Both distributions are well fitted as exponentials. The inverse rate is the mean escape time, in either direction. Other parameters: $L = 25$, $T = 1.5 \cdot 10^6$, $\mu = r = \omega = 0.5$, $\sigma_e = 0.029$.

Finite-N dependence

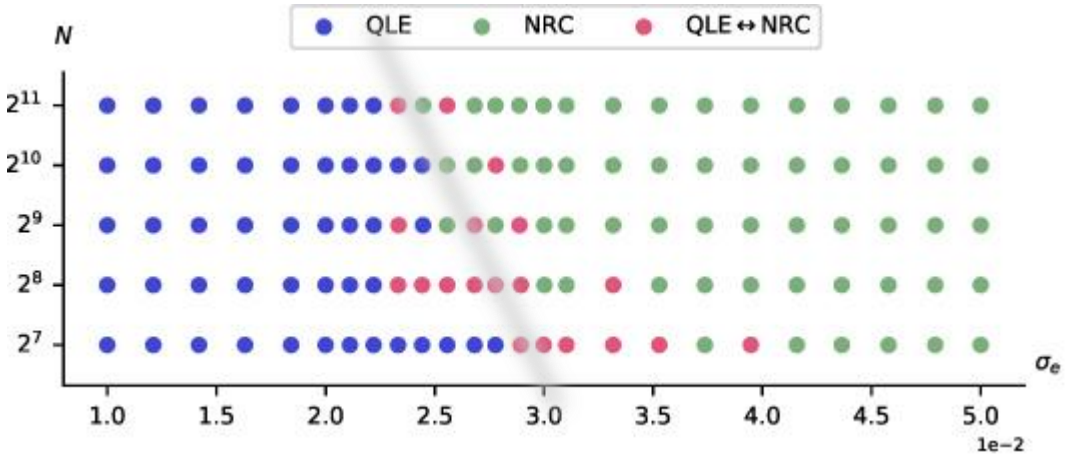
Estimated mean escape times from QLE and NRC. Inset: The dynamics undergoes multiple transitions QLE ↔ NRC ($T = 1.0 \cdot 10^4$).



The QLE → NRC transition happens when an individual in a finite population finds a high-fitness state. Analogous to the biophysical problem of transcription factors finding a binding site. Expected waiting time N^{-1}.

The NRC → QLE transition happens when a group of high-fitness individuals is lost from the population. Analogous to Muller ratchet. Expected waiting time exponential in N.

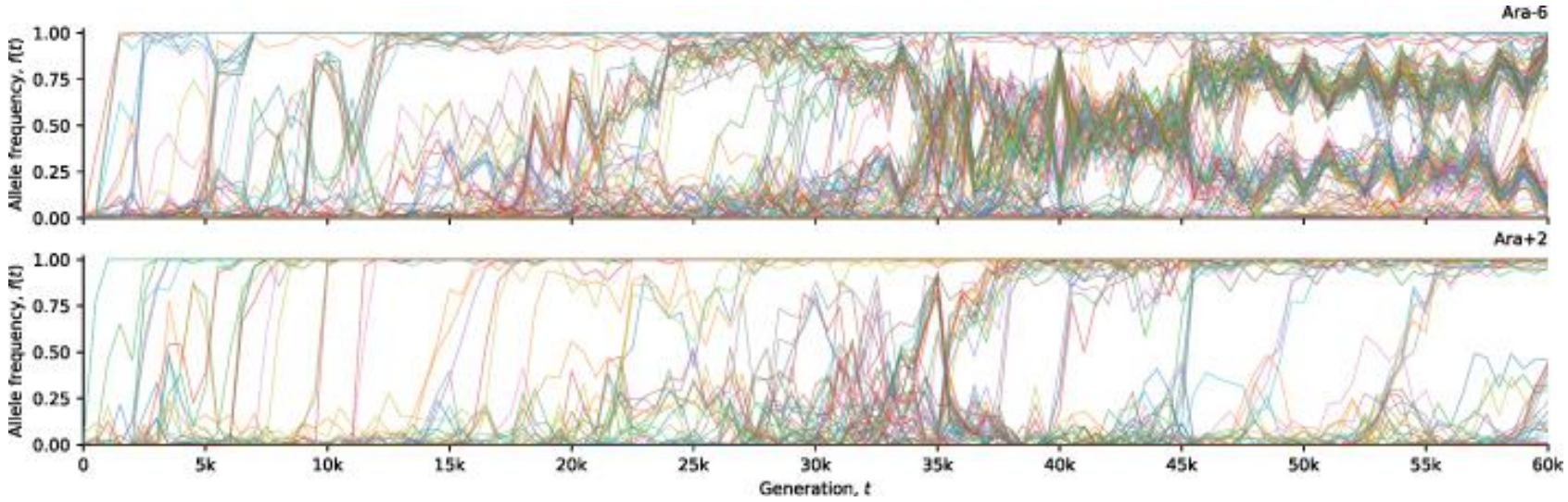
“Phase diagram” in (N, σ_e)



A number of simulations are run for the same time ($2.0 \cdot 10^4$). If the population remains in the QLE (NRC) the point is marked as **blue** (**green**). If at any point a transition QLE↔NRC is observed, the corresponding point is marked as **red**.

The previous heuristic theory predicts that for high N we the population should *always* be in NRC (same as in the Clonal Competition loss channel). This seems to be in agreement with the simulations (provided there is at least one transition).

Long-term evolution exps.



Allele frequency trajectories of all de novo mutations detected in 2 of the 12 LTEE populations, labelled respectively Ara-6 and Ara+2. Population Ara-6 (top row) shows quasi-stable coexistence of clades while Ara+2 (bottom row) shows mutations that fix rapidly. Quasi-stable coexistence was reported in 9 out of 12 LTEE populations [Good, McDonald, Barrick, Lenski, Desai 2017 *Nature* **551** 45–50 (2017)].

Figure previously unpublished, private communication from Profs B H Good and M M Desai, reproduced with permission.

ROYAL INSTITUTE OF TECHNOLOGY

Outlook & loose ends

SARS-CoV-2

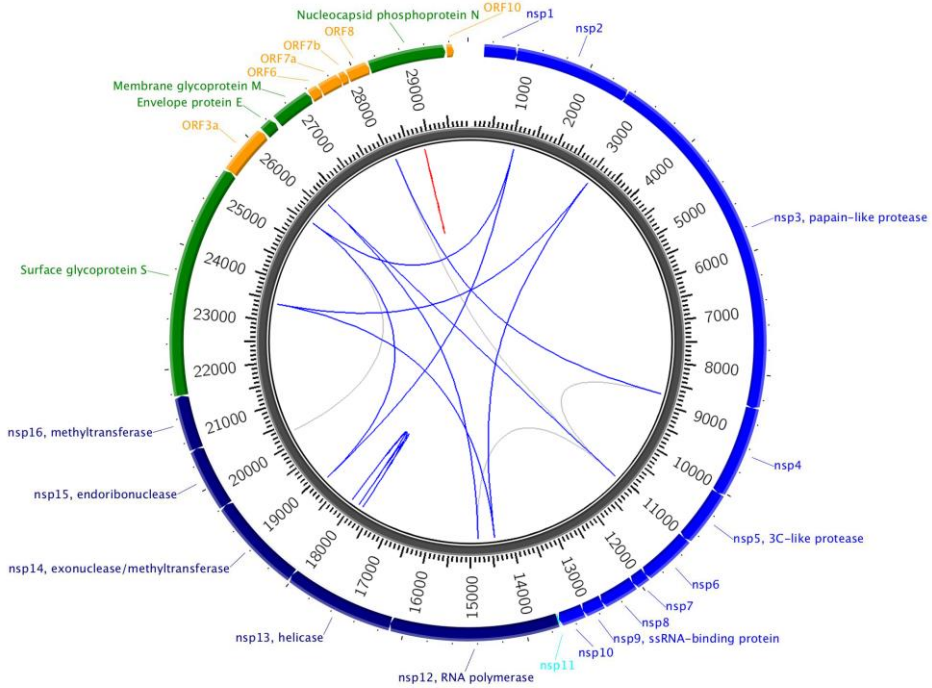
H-L Zeng et al [Erik Aurell] *PNAS* 2020 (*SARS-CoV-2*)

E Cresswell-Clay & V Periwal, *Mathematical biosciences* 2021 (*SARS-CoV-2*)

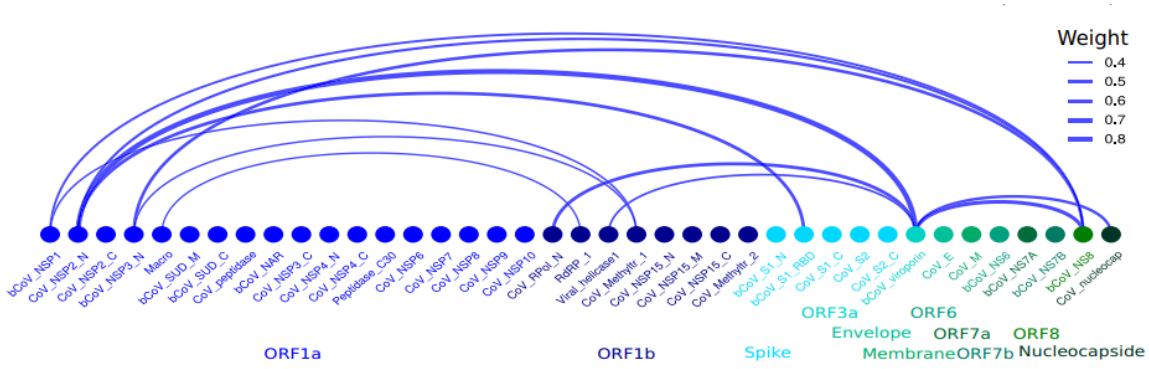
J Rodriguez-Rivas et al [Martin Weigt] *PNAS* 2022 (*SARS-CoV-2*)

Human (not yet attempted)

Global-scale data difficulties



Zeng et al (2020) and Cresswell-Clay & Periwal (2021) predicted many of the same interactions, from SARS-CoV-2 sequences on GISAID in 2020.



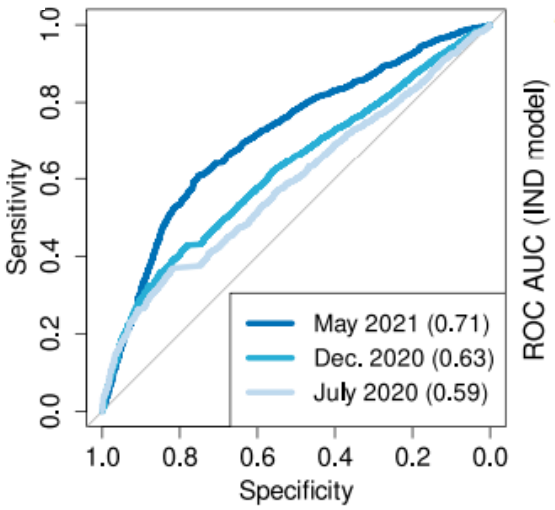
Rodriguez-Rivas et al (2022) predicted inter-actions from other coronaviruses.

Why Rodriguez-Rivas "better"?

These authors used DCA to predict mutation scores, which were then evaluated on GISAID variability.

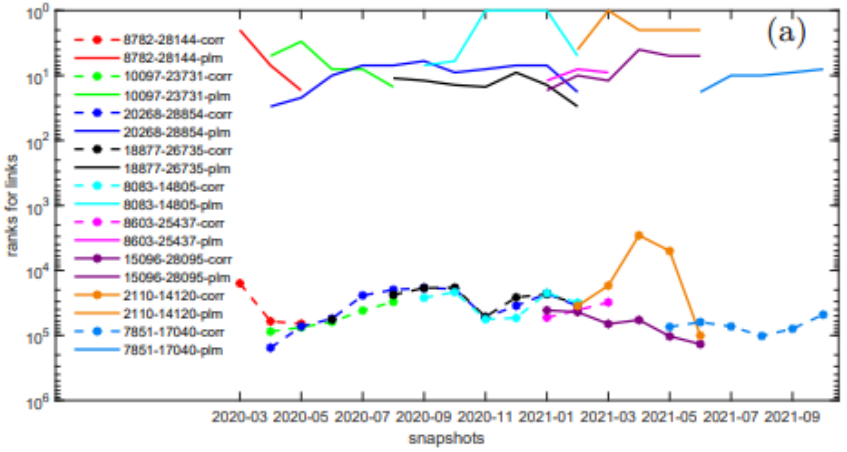
$$\Delta E_{DCA}(i,b) = \log P_{DCA}(a_1,\ldots,a_i,\ldots,a_L) - \log P_{DCA}(a_1,\ldots,b,\ldots,a_L)$$

$$S_{IND/DCA}(i) = \frac{1}{q}\sum_{k=1}^{q} \Delta E_{IND/DCA}(i,b_k)$$



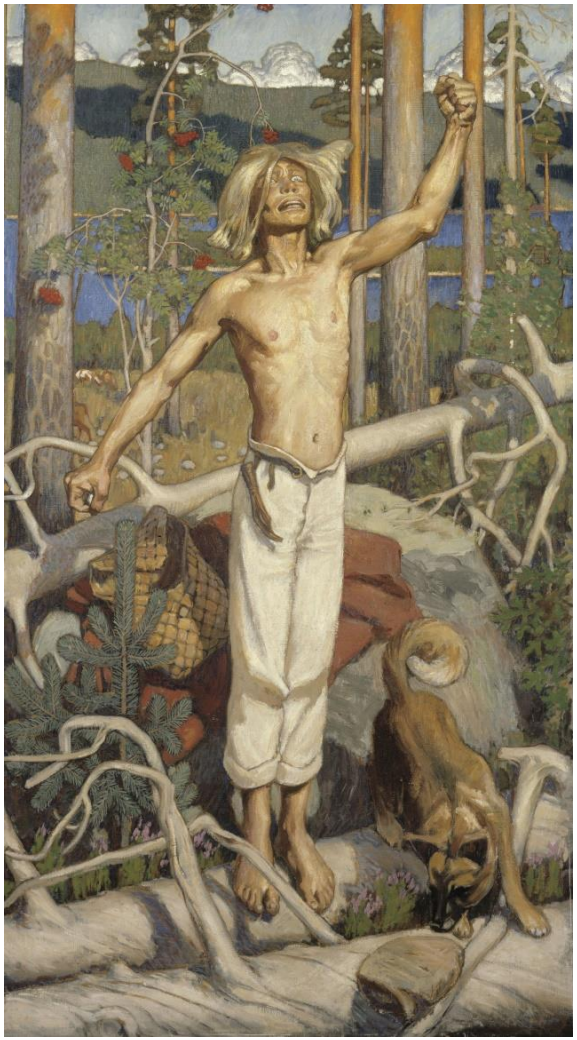
Rodriguez-Rivas et al *PNAS* 2022 Fig 4.B

Zeng et al and Cresswell-Clay & Periwal hit the problem that many variable loci in GISAID in 2020 later became fixed. Other interactions popped up.



Zeng et al *PRE* 2022 Fig 4(a)

Human-scale DCA?



WHY? Why not?

Perhaps a way to address the shortcomings of GWAS studies, that many traits are not well explained by variability of single genes.

Example: human obesity (BMI). In a cohort of 250k individuals and 2.8M genetic differences (SNPs) only 18 new loci explaining <4% of variability of BMI were found.

Speliotes *et al. Nat Genet.* 2010

Human-scale DCA

Problems and challenges

(***Population biology***:) Are large sets of human genomes described by exponential distributions? Are human populations in quasi-linkage equilibrium? (of course, this cannot be exactly so, but in some approximate sense?)

(***Algorithmic***:) How to effectively compute the largest J_{ij} when the number of loci is in the millions or billions? NB, this is not a totally trivial problem even for correlations. In computer science it's the "light bulb problem".

> L. Valiant. "Functionality in neural nets", In First Workshop on Computational Learning Theory, pages 28–39, 1988; G. Valiant. "Finding correlations in sub-quadratic time, with applications to learning parities and juntas", FOCS 2012

(***Usability and validation***:) the success of DCA and more recently AI methods such as Alpha-fold are to a large measure built on that many protein structures are known. There is a (partial) ground truth. This is (usually) not so on the genome scale. Better approaches to use and validate predictions would be advantageous.

Thanks

Hong-Li Zeng
Vito Dichio

Yue Liu

Eugenio Mauri
Simona Cocco
Rémi Monasson

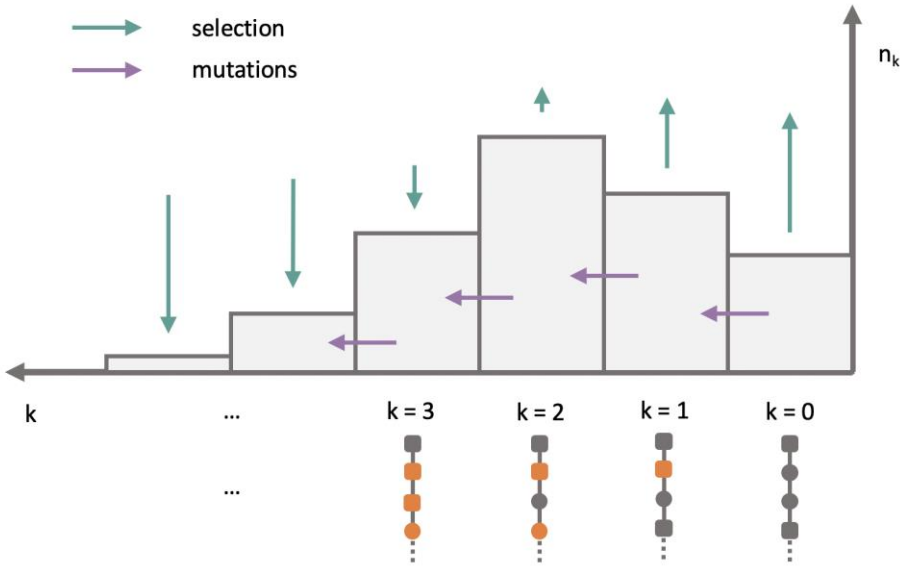
Fabbio Cecconi
Chen-Yi Gao
Angelo Vulpiani
Hai-Jun Zhou

Boris Shraiman
Richard Neher

Benjamin Good
Michael Desai

National Natural Science Foundation of China (11705097), Natural Science Foundation of Nanjing University of Posts and Telecommunications (Grant Nos. 221101 and 222134), Swedish Research Council (Grant 2020-04980).

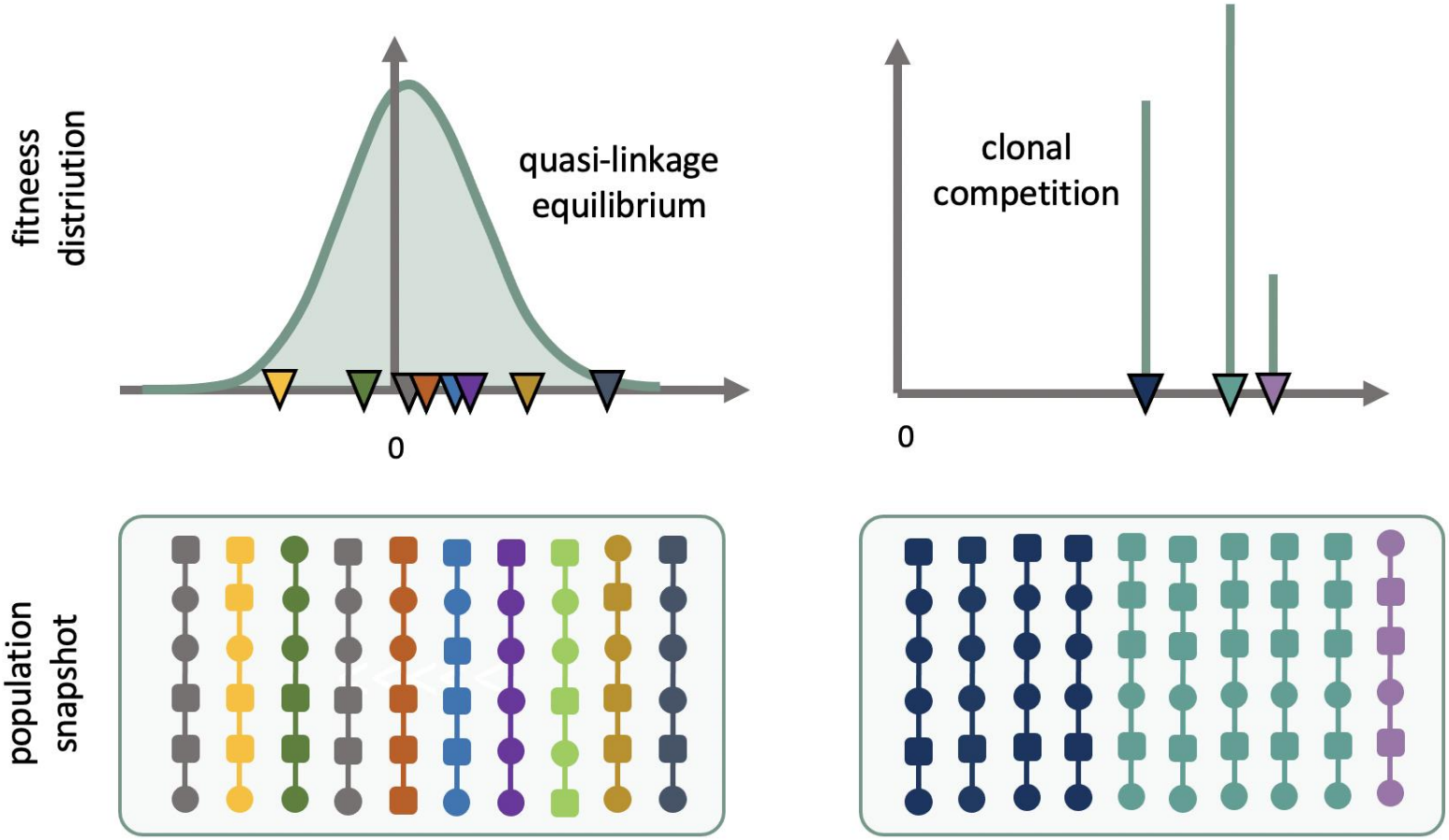
The role of genetic drift



Deleterious mutation–selection balance. The population is distributed among classes of individuals carrying *k* deleterious mutations. Classes with few mutations grow due to selection (green arrows), but lose individuals through mutations (violet arrows).

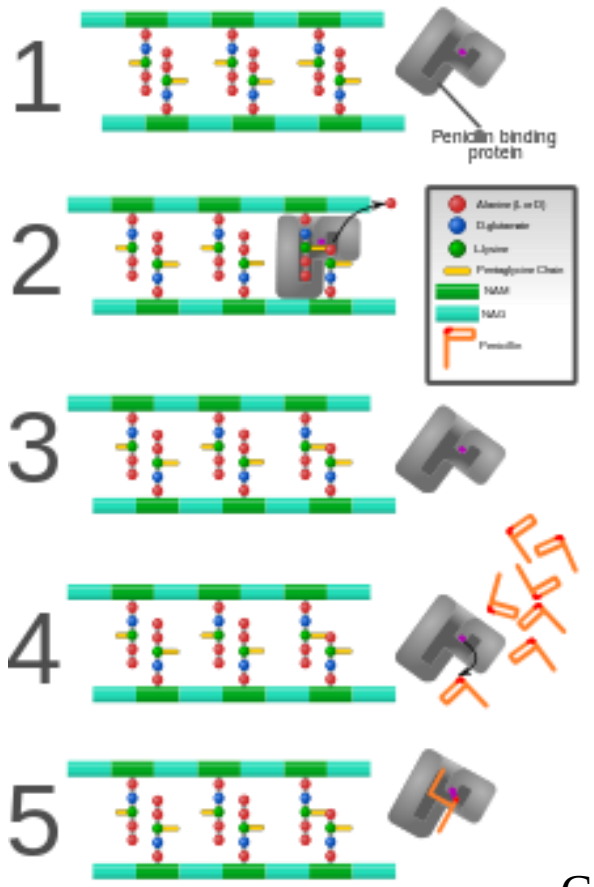
This can be done in Kimura-Neher-Shraiman theory (and was done by Neher & Shraiman). The master equations become stochastic differential equations.

This is (was) a serious issue in numerical simulations of QLE. Without mutations, eventually allele diversity at any locus is lost in a finite population. Correlations and Potts terms vanish, without any change in underlying fitness.



Quasi-linkage (QLE) equilibrium vs clonal competition (right). In a QLE state (left), individuals with the same genotype are rare and the fitness distribution is broad. In a clonal competition regime (right), few different genotypes are present in the population, each of them characterizing a number of individuals (a clone).

β-lactam (penicillin) resistance



@Mcstrother
Wikimedia Commons

PBPs (Penicillin-binding proteins)

B. Spratt, *Eur. J. Biochem.* (1977)

PASTA (PF03793)

Penicillin-binding protein **a**nd **s**erine/**t**hreonine kinase **a**ssociated domain [..] binds beta-lactam antibiotics and their peptidoglycan analogues […] describe this previously uncharacterized domain and infer that it binds beta-lactam antibiotics and their peptidoglycan analogues.

C. Yeats, RD Finn, A. Bateman, *Trends Biochem Sci.* (2002)
"The PASTA domain: a beta-lactam-binding domain".

Spike-spike

Zeng et al *PRE* 2022 Table I

August 2021					September 2021					October 2021				
rank	locus 1	AA-m.	locus 2	AA-m.	rank	locus 1	AA-m.	locus 2	AA-m.	rank	locus 1	AA-m.	locus 2	AA-m.
7	23284	D574D	25339	D1259D	7	23284	D574D	25339	D1259D	9	23284	D574D	25339	D1259D
16	21987	G142D	24410	D950N	15	21987	G142D	24410	D950N	11	21995	T145H	22227	A222V
67	22093	M177I	22104	G181V	45	21995	T145H	22227	A222V	15	21987	G142D	24410	D950N
70	22917	R452L	22995	K478T						135	21846	T95I	24208	I882I
71	22082	P174S	22093	M177I										
74	22081	Q173H	22093	M177I										
190	22082	P174S	22104	G181V										
195	22081	Q173H	22104	G181V										

TABLE I. Largest DCA terms with both terminals in Spike coding region, August-October 2021. Top-200 couplings computed as plmDCA scores are considered. For each of them in the three months displayed, there's the indication of the rank, the two loci involved and the corresponding amino acid (AA) mutations. Green color indicates that this mutation is found in delta variant. Red color indicates that this mutation is found in omicron variant. Couplings with one or both terminals colored green are attributed to a phylogenetic effect. The single pair with one terminal colored red is not attributed to a phylogenetic effect, the growth of omicron being later than October-2021.

Spike-non-spike

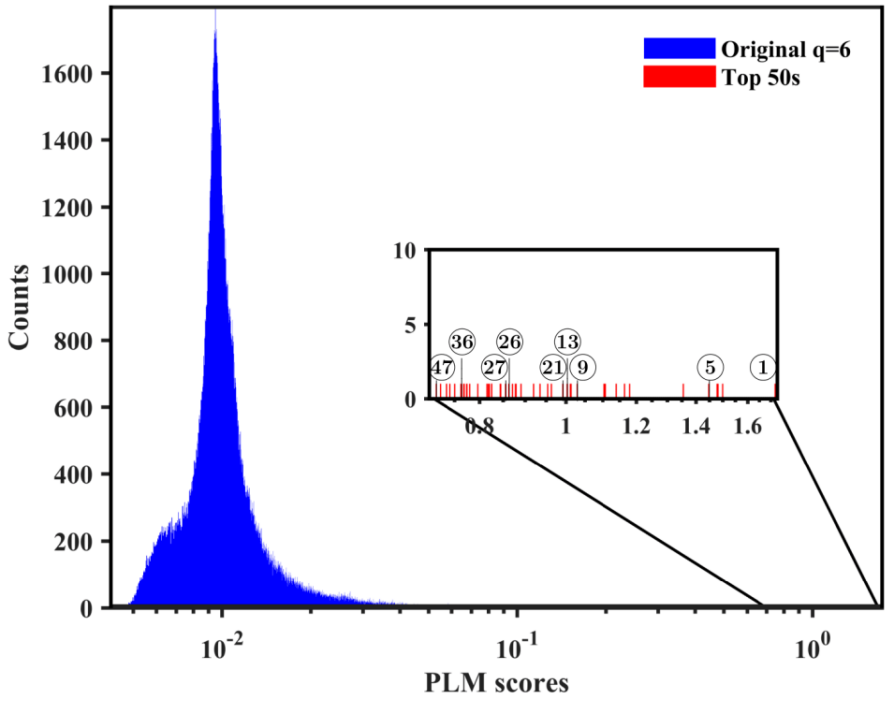
ROYAL INSTITUTE OF TECHNOLOGY

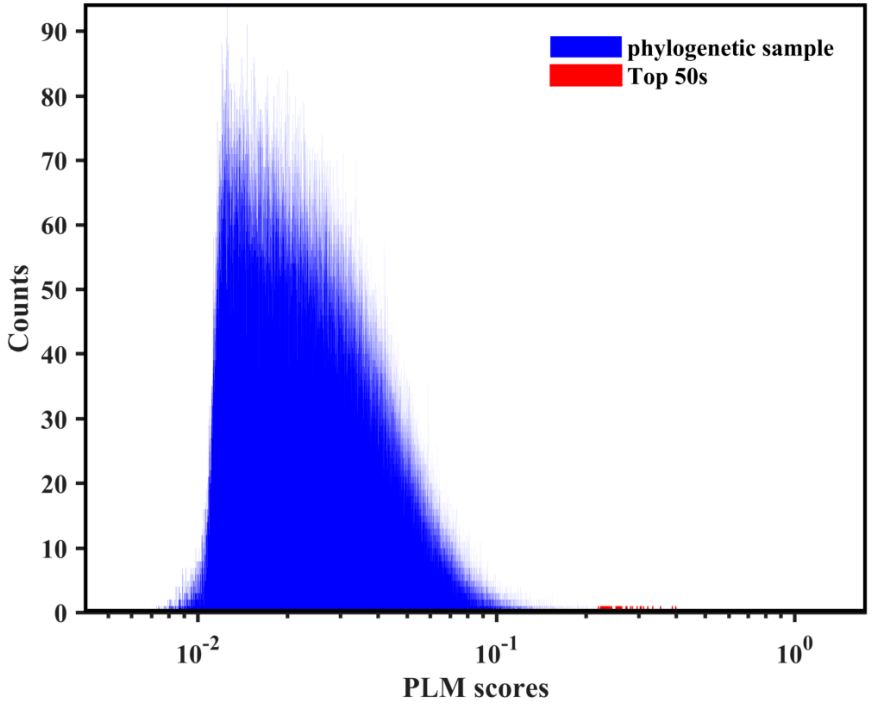
Zeng et al *PRE* 2022 Table II

August 2021					September 2021					October 2021				
rank	Partner		Spike		rank	Partner		Spike		rank	Partner		Spike	
	locus	AA-m.	locus	AA-m.		locus	AA-m.	locus	AA-m.		locus	AA-m.	locus	AA-m.
1	17236	nsp13:I334V	24208	I882I	1	17236	nsp13:I334V	24208	I882I	1	17236	nsp13:I334V	24208	I882I
14	7851	nsp3:A1711V	21846	T95I	13	7851	nsp3:A1711V	21846	T95I	10	7851	nsp3:A1711V	21846	T95I
20	28461	N:G63D	24410	D950N	16	28461	N: D63G	24410	D950N	17	28461	N:D63G	24410	D950N
27	1048	nsp2:K81N	21846	T95I	36	1048	nsp2:K81N	21846	T95I	20	25614	ORF3a:S74S	21995	T145H
52	26107	ORF3a:E239Q	21897	S112L	52	25614	ORF3a:S74S	21995	T145H	21	25614	ORF3a: S74S	22227	A222V
57	27507	ORF7a:G38G	21897	S112L	57	26107	ORF3a:E239Q	21897	S112L	30	1048	nsp2:K81N	21846	T95I
62	18086	nsp14:T16I	22792	I410I	58	25614	ORF3a:S74S	22227	A222V	51	10977	nsp6:A2V	21846	T95I
76	27291	ORF6:D30D	24208	I882I	71	27507	ORF7a:G38G	21897	S112L	56	27291	ORF6:D30D	24208	I882I
79	1729	nsp2:V308V	22792	I410I	82	27291	ORF6:G30G	24208	I882I	60	26107	ORF3a:E239Q	21897	S112L
151	28007	ORF8:P38P	21846	T95I	83	11514	nsp6:T181I	22227	A222V	63	29253	N:S327L	21846	T95I
168	27604	ORF7a:V71I	21846	T95I	128	17236	nsp13:I334V	21846	T95I	64	18744	nsp14:T235T	24130	N856N
174	17236	nsp13:I334V	21846	T95I	151	18744	nsp14:T235T	24130	N856N	74	27507	ORF7a:G38G	21897	S112L
197	11514	nsp6:T181I	22227	A222V	190	5584	nsp3:T955T	22227	A222V	80	17236	nsp13:I334V	21846	T95I
					195	13019	nsp9:L112L	22227	A222V	124	15952	nsp12:S837S	21846	T95I
										153	26107	ORF3a:E239	21846	T95I
										163	28299	N:Q9L	21846	T95I
										190	27507	ORF7a:G38G	21846	T95I
										194	11562	nsp6:C197F	21897	S112L
										197	11514	nsp6:T181I	22227	A222V

TABLE II. Largest DCA terms with only one terminal in Spike coding region, August-October 2021. Top-200 couplings computed as plmDCA scores are considered. For each of them in the three months displayed, there's the indication of the rank, the locus in the Spike coding region and corresponding amino acid (AA) mutation, the locus in the partner coding region and corresponding amino acid (AA) mutation. Green color indicates that this mutation is found in delta variant. Red color indicates that this mutation is found in omicron variant. Pairs with one or both terminals colored green are attributed to a phylogenetic effect, while the several pairs with one terminal colored red are not, the growth of omicron being later than October-2021. Omicron mutations used here are taken from [67] on page 18, deletions not considered.

Phylogeny (inheritance) a confounder?

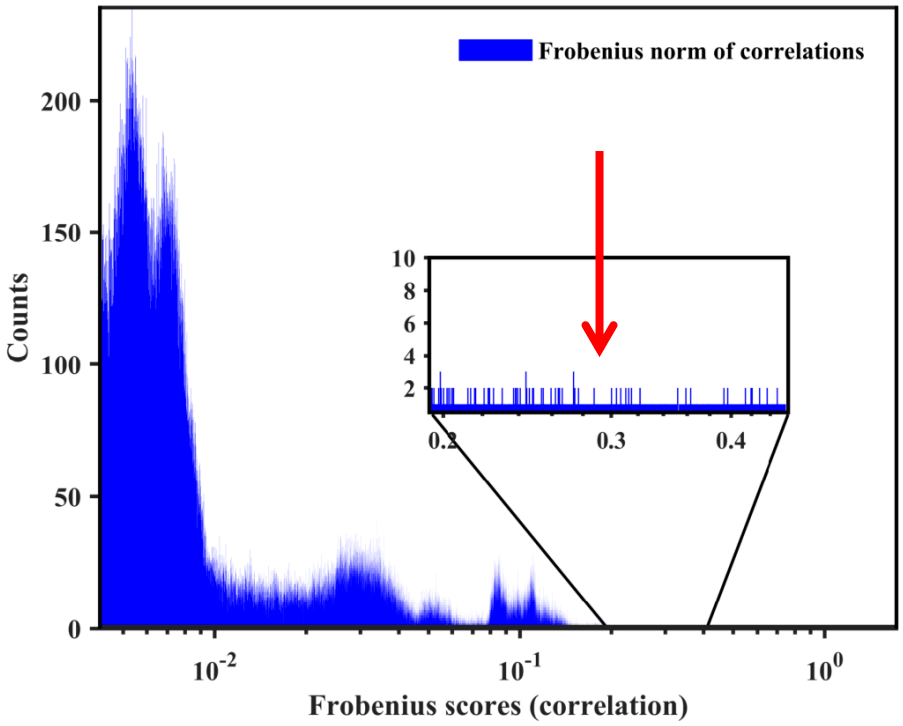


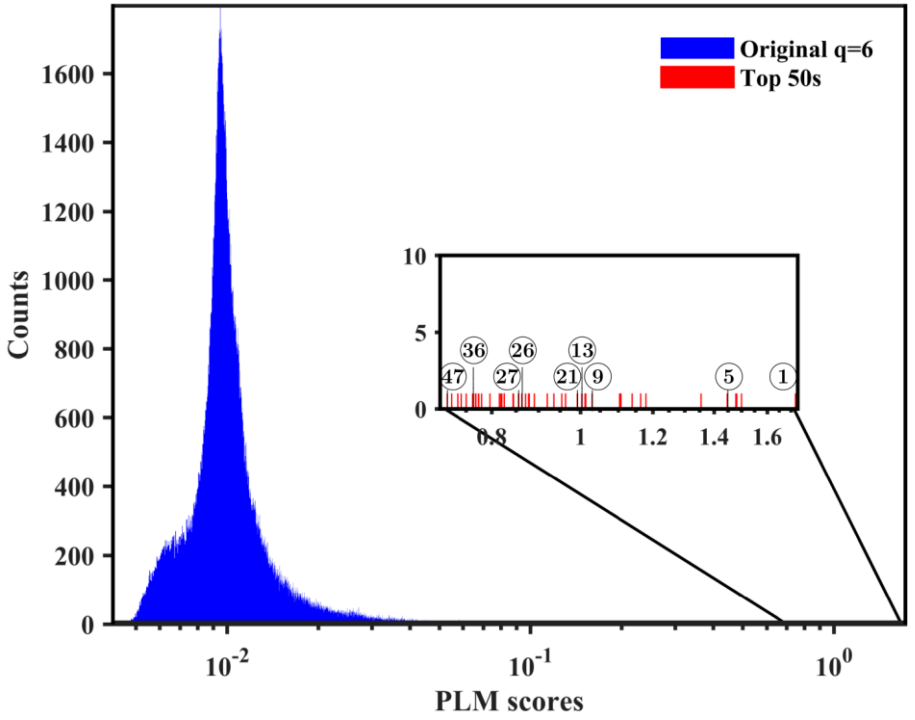


Is the effect due to inherited variation? We tested by scrambling MSA while preserving inter-sequence distances.

Edwin Rodriguez Horta, Martin Weigt
bioRxiv 2020.08.12.247577

Is it just correlation analysis?

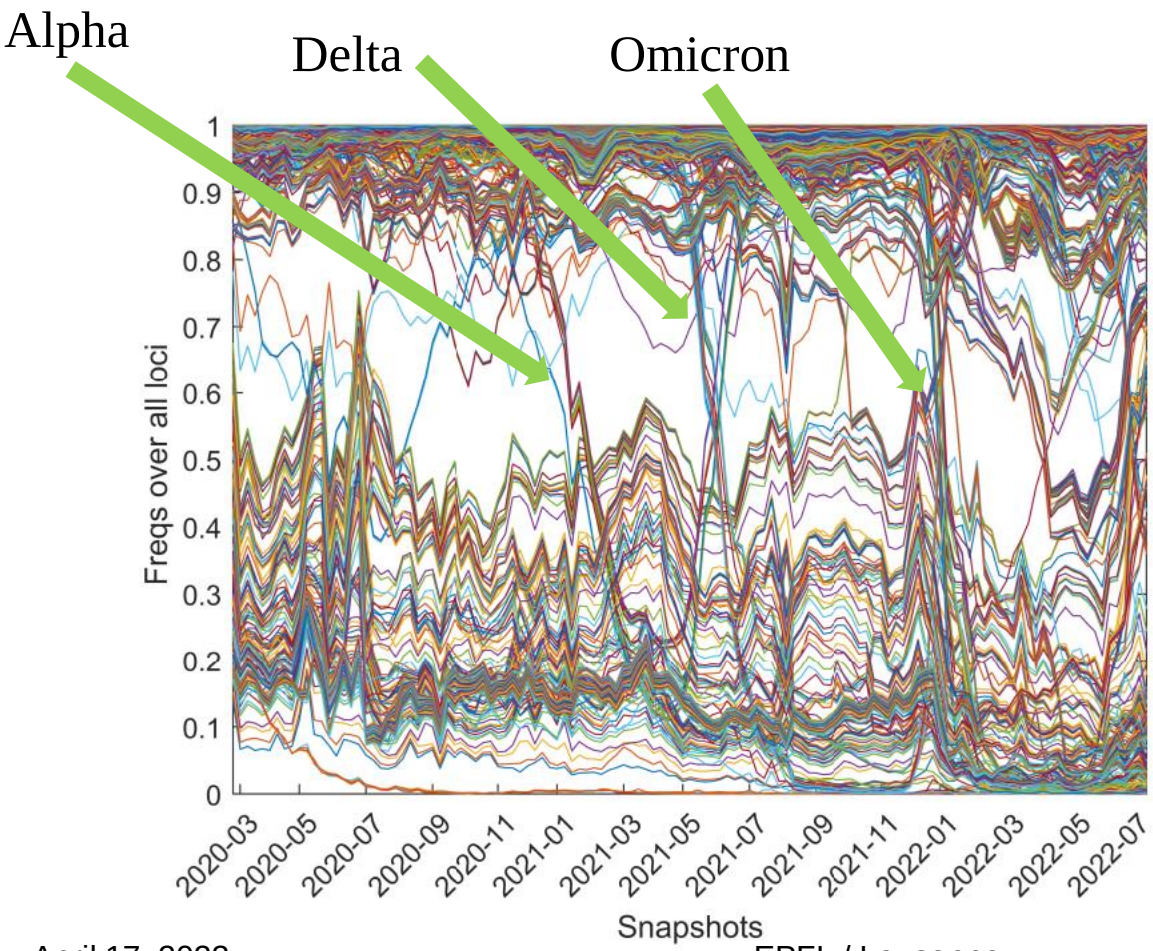




There is very little overlap between the leading predictions from DCA and most correlated pairs.

SARS-CoV-2 perhaps also NRC?

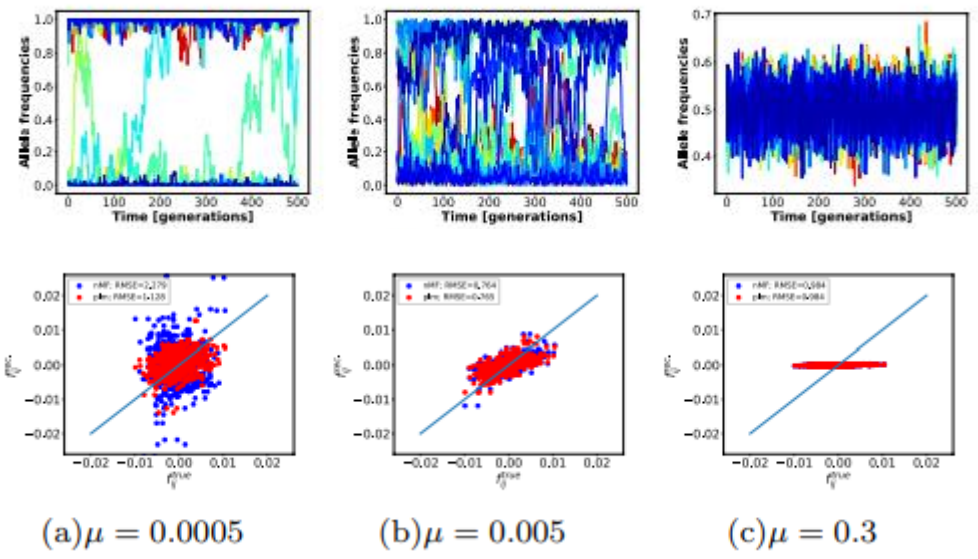
Coronaviruses recombine. This has been observed in SARS-CoV-2, *in vivo*. Plots of allele frequencies at *all* loci show the well-known VoCs Alpha, Beta, Delta, Omicron…but also a bit more.



frozen loci

Frequencies of all alleles on all positions per week from GISAID up to August 2022 [Zeng & Liu, unpublished]
[see also arXiv:2109.02962]

An NRC phase? Most of these intermittently fluctuating loci lie in the 5' or 3' end of the SARS-CoV-2 genome.



(a)$\mu = 0.0005$ (b)$\mu = 0.005$ (c)$\mu = 0.3$

Dynamics and fitness inference.

N = 200 (population size)

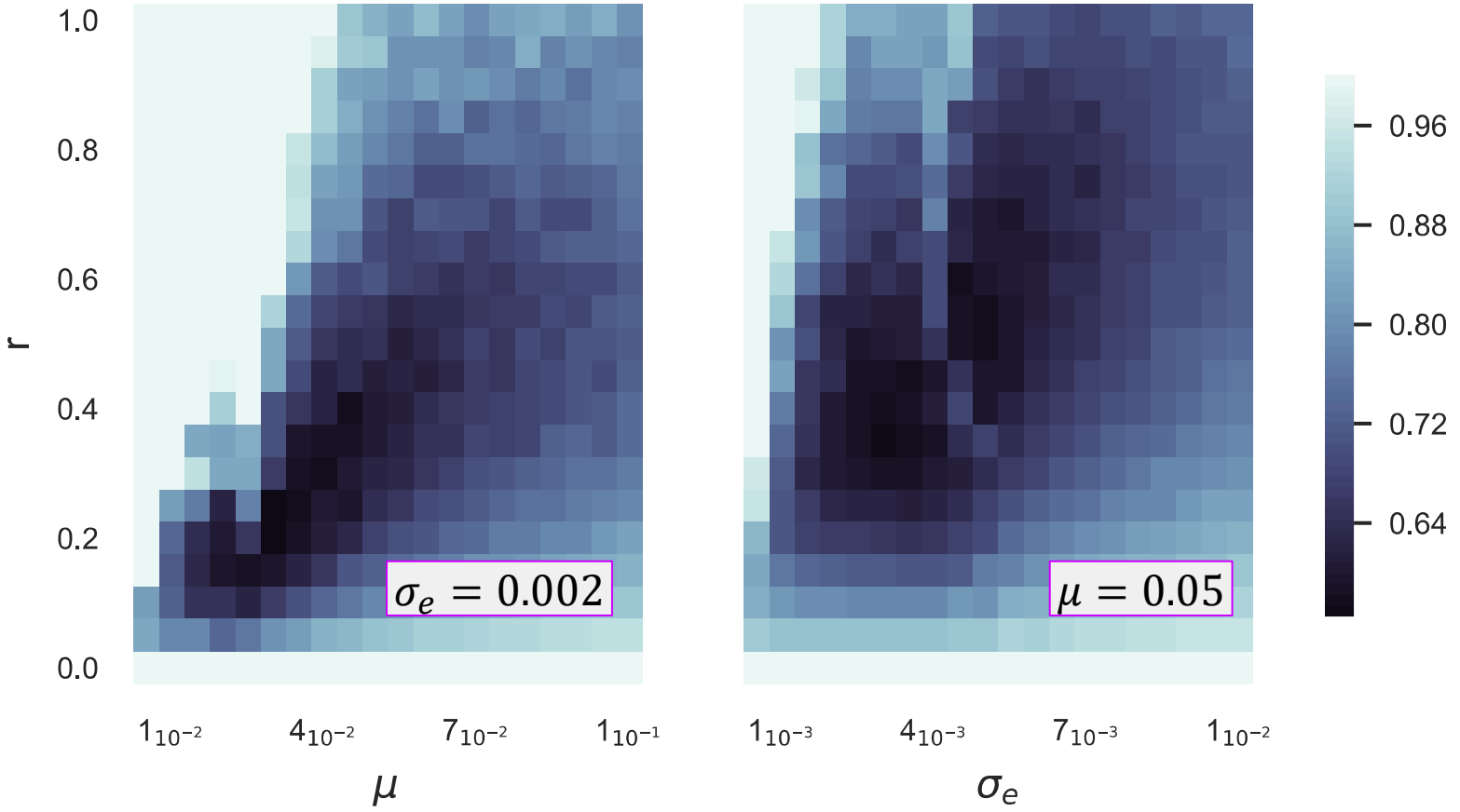
L = 25 (# of loci)

$(r, \rho, \sigma_e) = (0.05, 0.5, 0.002)$

T = 5 × 500 (simulation time)

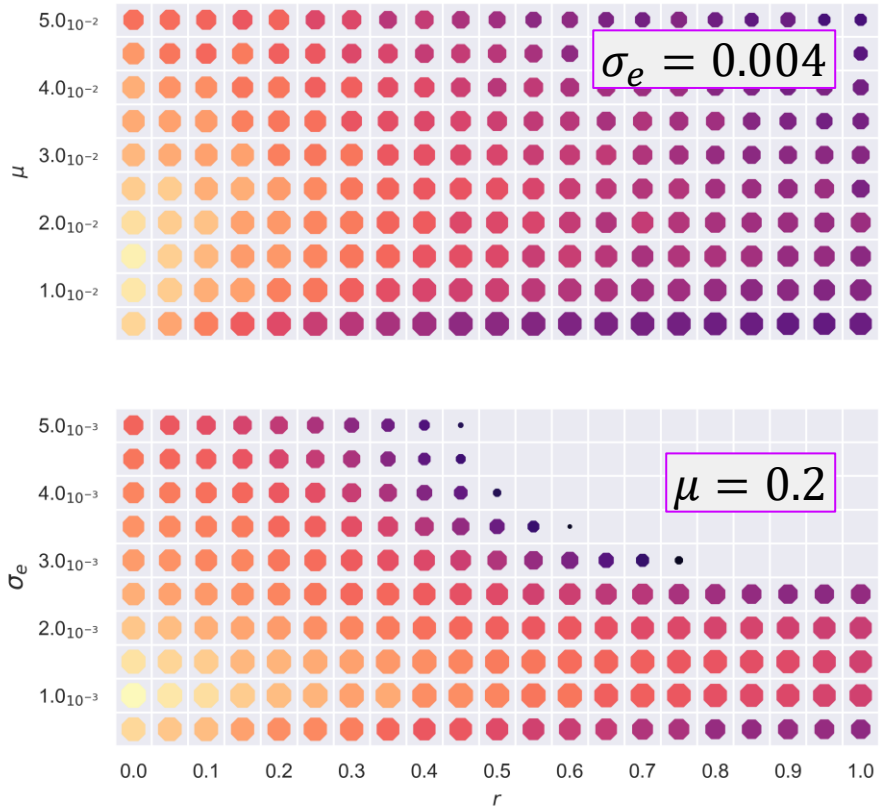
In (**a**) [low mutation] there is not enough variability in the data over this duration.

Zeng, EA, *Phys Rev E* **101**(5) 052409 (2020)



Phase diagrams. Parameters as in previous slide. ***Low*** r does not work because Kimura-Neher-Shraiman does not apply.

Zeng, EA, *Phys Rev E* **101**(5) 052409 (2020)



	Value	Description
N	200	n. individuals
L	25	n. of loci
T	10,000	n. of generations
ω	0.5	crossover rate
r	[0.0:1.0]	rate of recombination
μ	[0.05:0.5]	rate of mutation
σ_e	[0.004:0.04]	$f_{ij} \sim \mathcal{N}(0, \sigma_e)$
σ_a	0.05	$f_i \sim \mathcal{N}(0, \sigma_a)$

Reconstruction of the epistatic fitness components in the phase spaces $r \leftrightarrow \mu$ and $r \leftrightarrow \mu$. Bigger dots means higher accuracy (lower reconstruction errors) according to MZDACM theory (simplest version, previous slide). Colors indicate the difference to Kimura-Neher-Shraiman theory. MZDACM accurate Both formu recombination nor mutation high compared to epistatic fitness). Replotted after Zeng et al JSTAT 2021.