ROYAL INSTITUTE OF TECHNOLOGY

Lessons learned on SARS-CoV-2 genomics from GISAID

Erik Aurell

NITheCS Colloquium

August 23, 2021

ROYAL INSTITUTE OF TECHNOLOGY

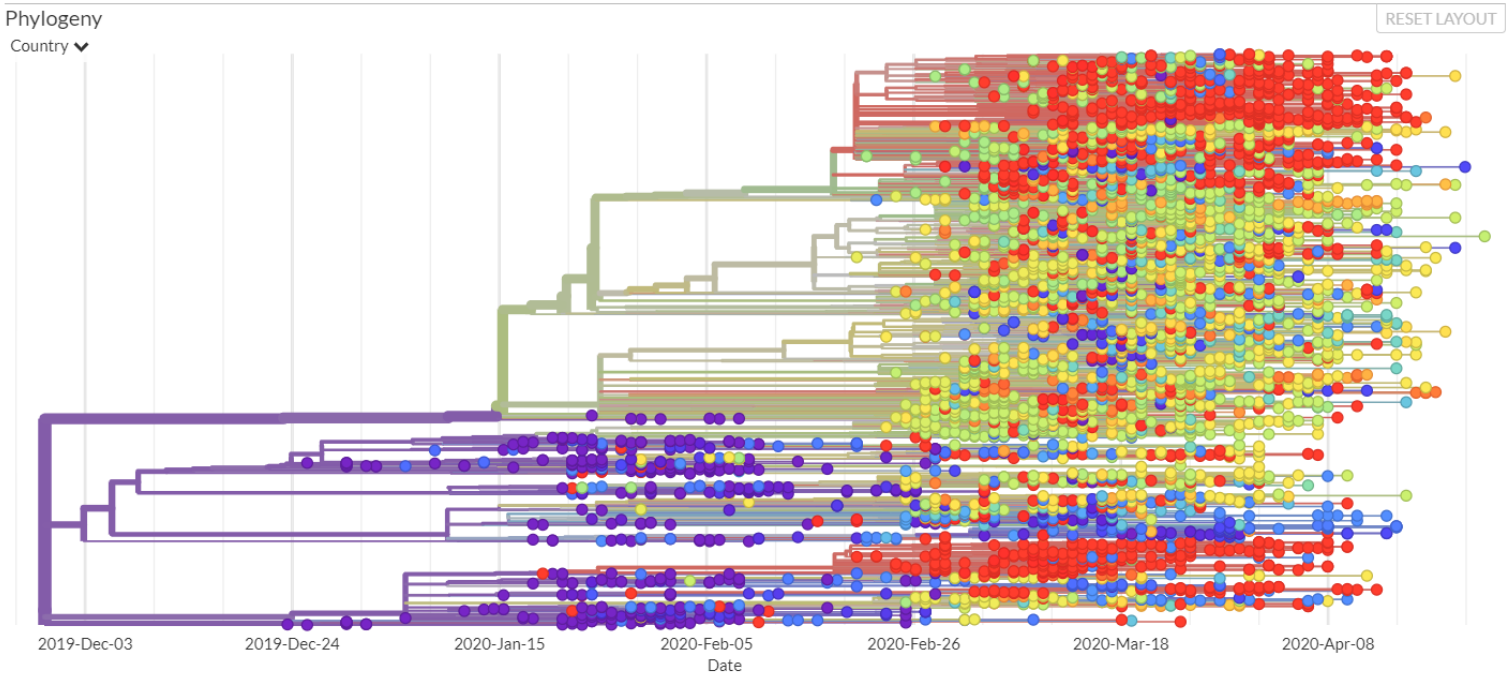
GISAID ("global initiative on sharing avian influenza data") **is a global science initiative established in 2008. It has become the leading repository for SARS-CoV-2 genomic data.**

GISAID holds 2,916,063 SARS-CoV-2 submissions (8/20 2021)**. Though not all "high-quality, full-length genomes", a very large data set.**

ROYAL INSTITUTE OF TECHNOLOGY

“The use of pathogen genomes on this scale to track the spread of the virus internationally, study local outbreaks and inform public health policy signifies a new age in virus genomic investigations.”

Harvey et al, “SARS-CoV-2 variants, spike mutations and immune escape”
Nature Reviews Microbiology **19:**409–424 (2021)



A visualization of genomes accumulated on GISAID up to April 2020 with inferred phylogenetic trees. Color coding represents mutations. A data cloud!

ROYAL INSTITUTE OF TECHNOLOGY

I will tell two stories

a rather simple time series analysis

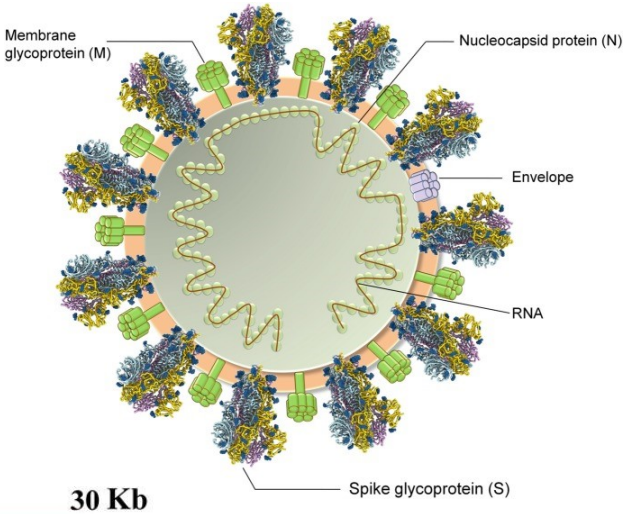
H-L Zeng, Y Liu, K Thorell, R Nordén, EA
"Uneven growth of SARS-CoV-2 clones evidenced by more than 500,000 whole-genome sequences", bioRxiv 2021.04.06.437914 (April 06, 2021)

a more ambitious / speculative theory

H-L Zeng, V Dichio, E Rodríguez Horta, K Thorell, EA
"Global analysis of more than 50,000 SARS-CoV-2 genomes reveals epistasis between eight viral genes"
PNAS **117** (49) 31519-31526 (December 8, 2020)

SARS-CoV-2 overview

Similar to other coronaviruses, SARS-CoV-2 encodes two large multi-protein genes (ORF1ab) and several single-protein genes (Spike, M, ORF3a, etc.)



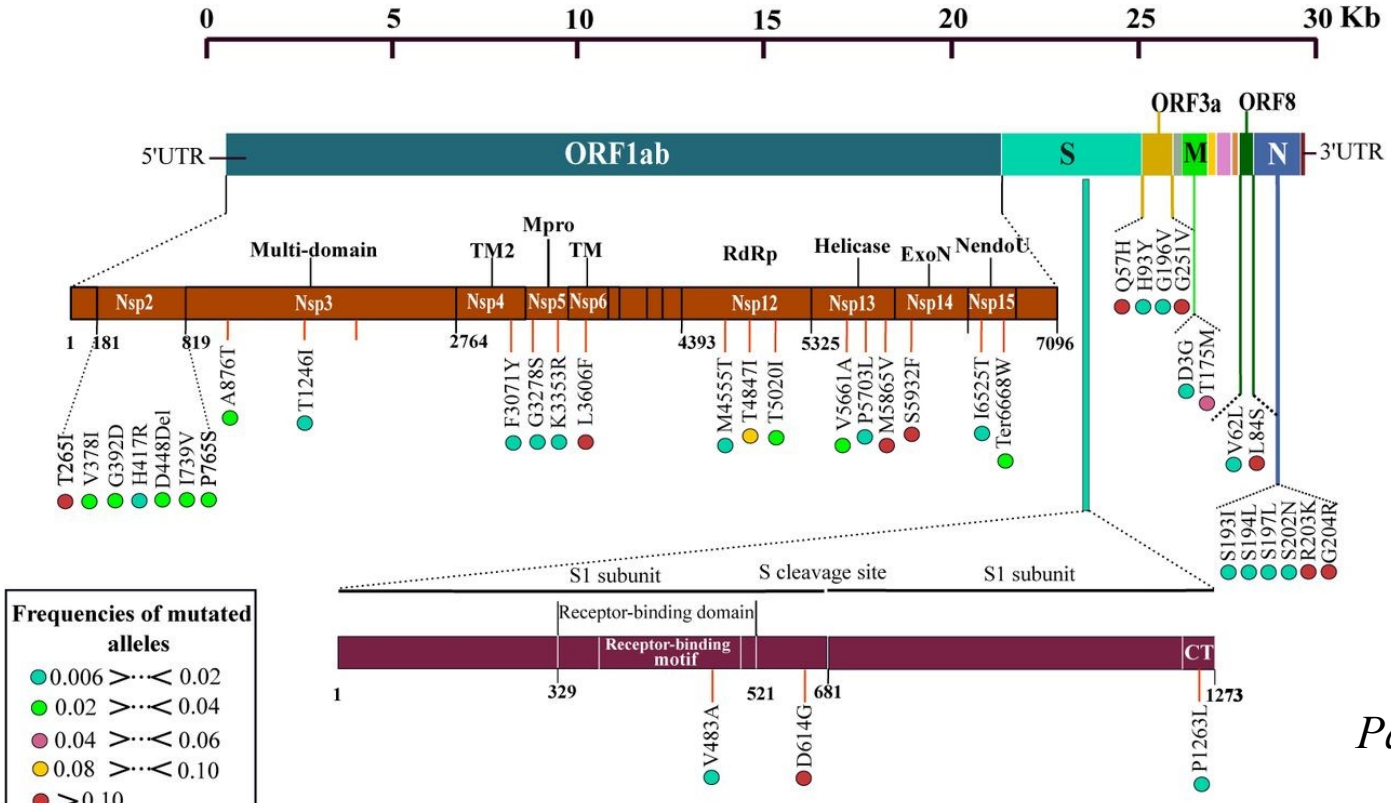


Illustration from Alouane et al, *Pathogens* **9**:829 (2020)

ROYAL INSTITUTE OF TECHNOLOGY

Three natural experiments

growths of Variants of Concern

H-L Zeng, Y Liu, K Thorell, R Nordén, EA
"Uneven growth of SARS-CoV-2 clones evidenced by more than 500,000 whole-genome sequences", bioRxiv 2021.04.06.437914 (April 06, 2021)

+ later results up to July 2021

ROYAL INSTITUTE OF TECHNOLOGY

“UK variant” B.1.1.7 a.k.a alpha

First defined in
Meera Chand et al
Public Health England
“Investigation of novel SARS-COV-2 variant
Variant of Concern 202012/01”

Mutations listed for UK variant

gene	nucleotide	amino acid
ORF1ab	C3267T	T1001I
	C5388A	A1708D
	T6954C	I2230T
	11288-11296 del	SGF 3675-3677 del
Spike	21765-21770 del	HV 69-70 del
	21991-21993 del	Y144 del
	A23063T	N501Y
	C23271A	A570D
	C23604A	P681H
	C23709T	T716I
	T24506G	S982A
	G24914C	D1118H
ORF8	C27972T	Q27stop
	G28048T	R52I
	A28111G	Y73C
N	28280 GAT->CTA	D3L
	C28977T	S235F

17 non-synonymous mutations
Meera Chand et al op cit
Table 1

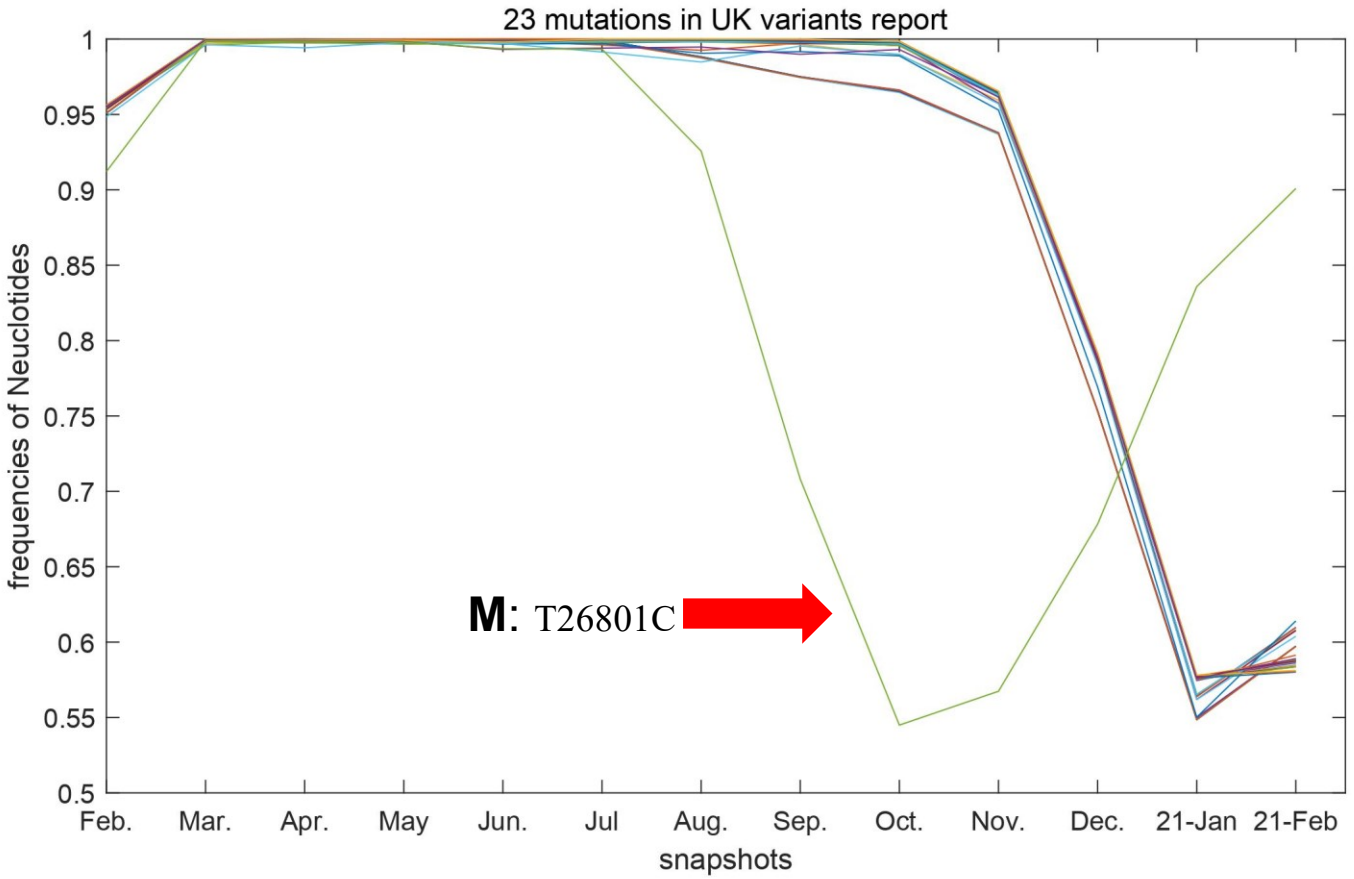
6 synonymous mutations
Meera Chand et al op cit
in text above Table 1

gene	nucleotide
ORF1ab	C913T
	C5986T
	C14676T
	C15279T
	C16176T
M	T26801C

Spike D614G is not listed by Chand et al, though often in other lists.

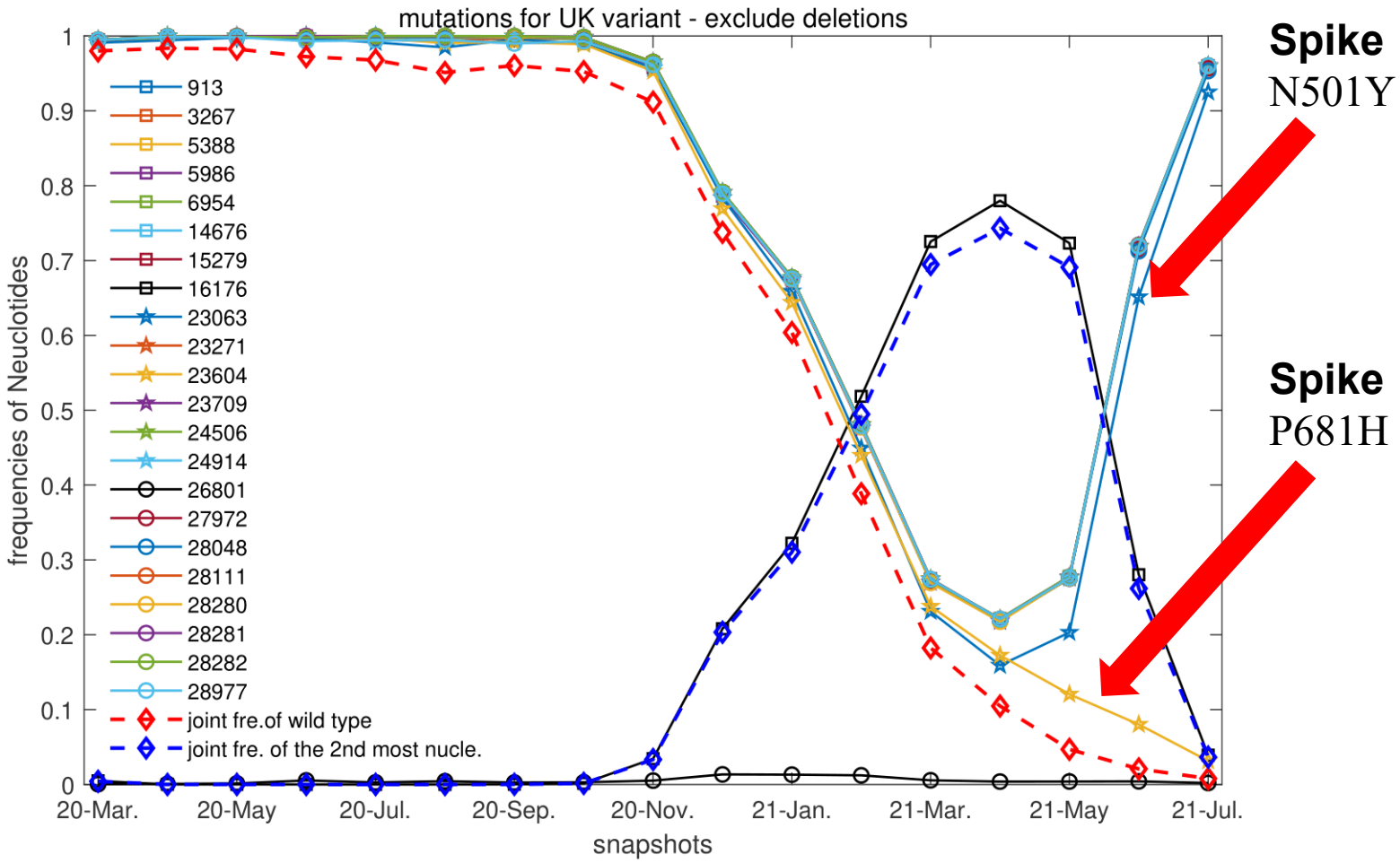
wt frequencies vs time

Frequencies of wild-type allele per month until February 2021



Mutation T26801C has a dynamics unrelated to the rest. C16176T (not shown) has the precise opposite dynamics. Likely that mutation was in fact T16176C!

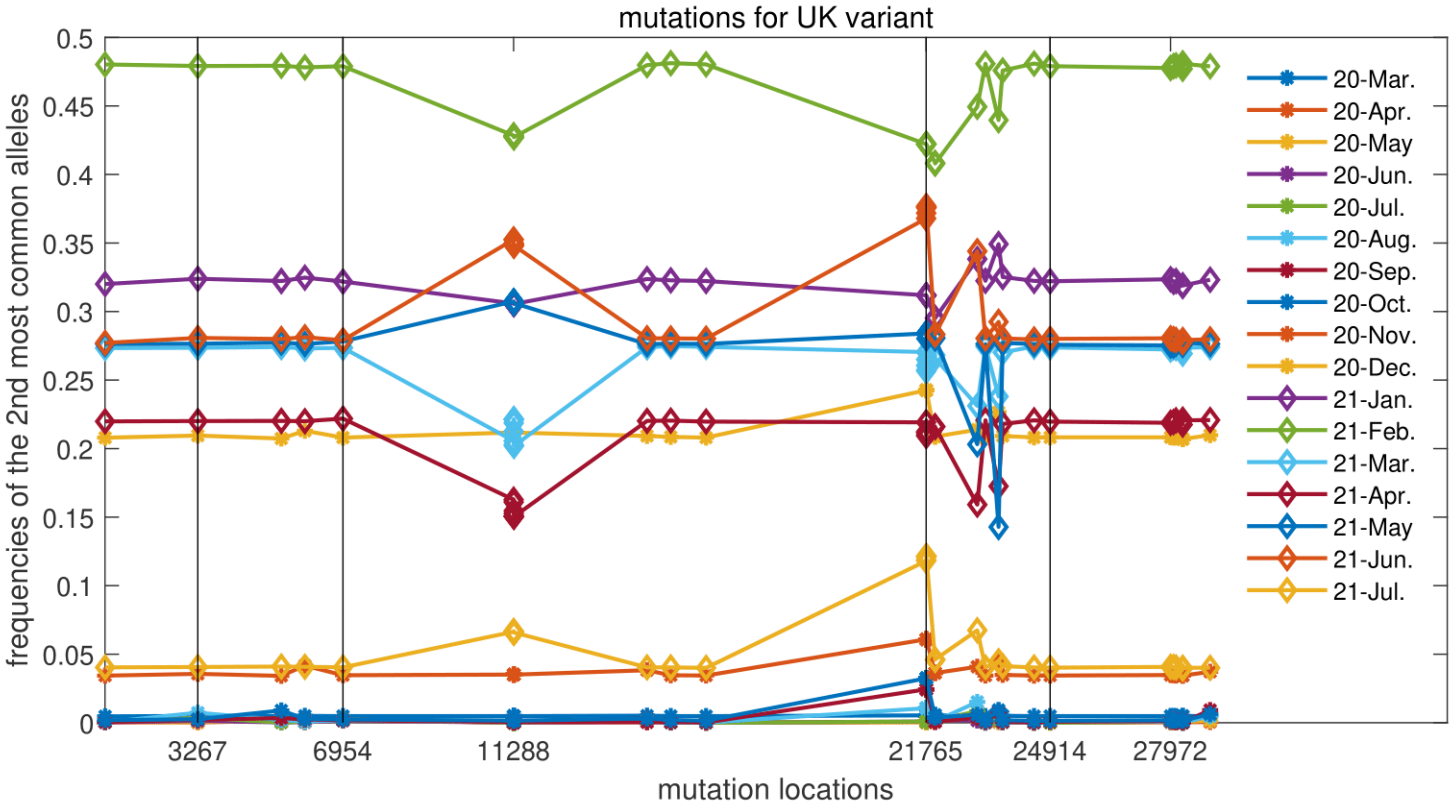
wt frequencies vs time



Single-allele wt frequencies except T26801C, and joint frequencies, until July-21.

ROYAL INSTITUTE OF TECHNOLOGY

+frequencies vs position



Frequencies for mutant alleles at different loci except T26801C and C23604A (Spike P681H). Highest frequencies are in February 2021 (green, open rhombi). Initially frequencies grow in synchrony, but later some variations appear.

ROYAL INSTITUTE OF TECHNOLOGY

“South Africa variant”
B.1.351
a.k.a. beta

Maybe not absolutely first definition….but at least as in Europe was defined in

M. Chand,et al.,
“Investigation of SARS-CoV-2 variants of concern in England”
Public Health England
Technical briefing 6, Table 4a

Mutations listed for SA variant

gene	nucleotide	amino acid
NSP2	C1059T	T265I
NSP3	G5230T	K1655N
NSP5	A10323G	K3353R
NSP6	11288-96 del	3675-3677 del
Spike	C21614T	L18F
	A21801C	D80A
	A22206G	D215G
	—	242-244del
	G22299T	R246I
	G22813T	K417N
	G23012A	SGF E484K
	A23063T	N501Y
	C23664T	A701V
ORF3a	G25563T	Q57H
	C25904T	S171L
E	C26456	P71L
N	C28887T	T205I

17 non- synonymous mutations are listed by Public Health England

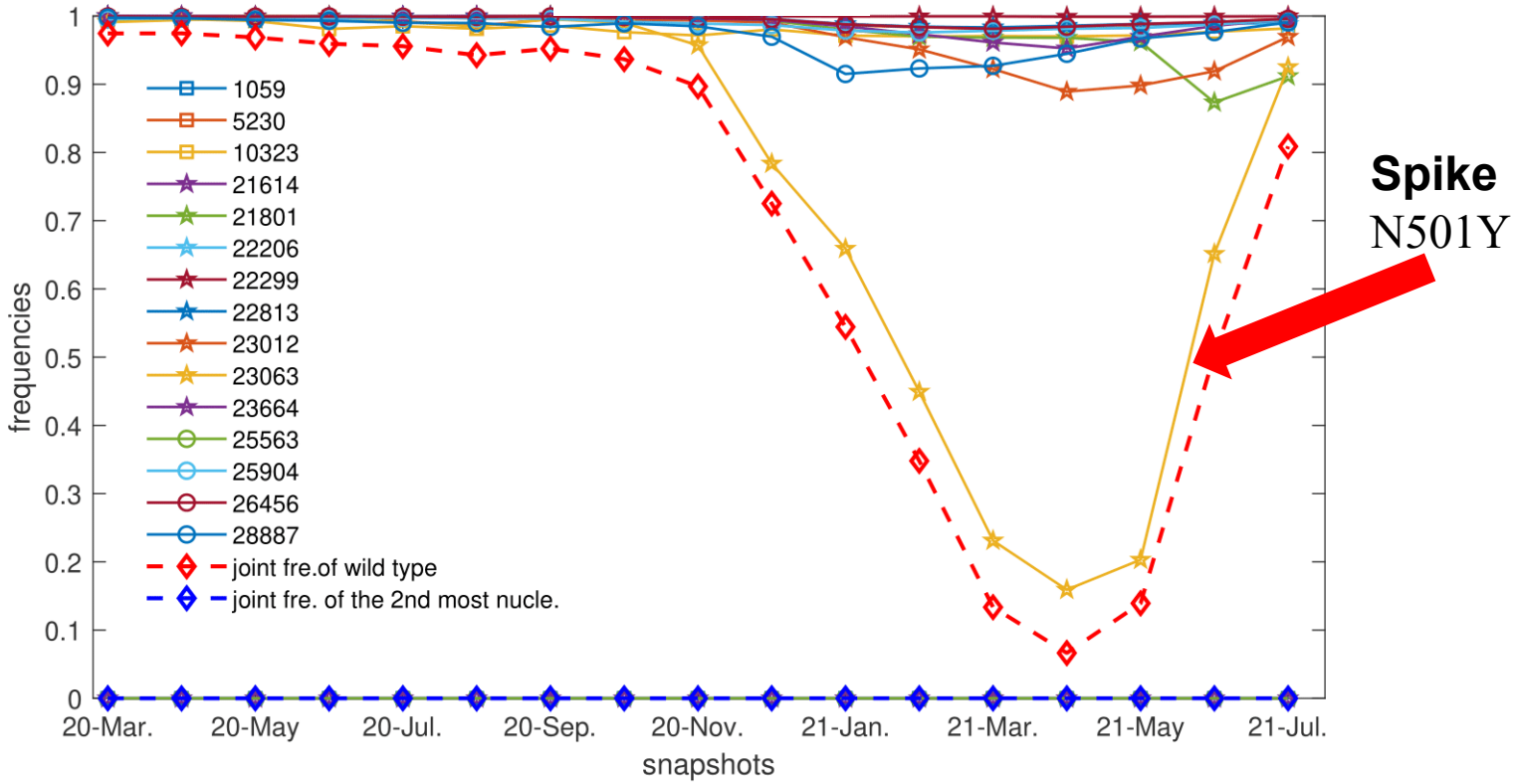
Of these 17 mutations only 7 were retained in the Pangolin B.1.351.

These 7 are marked by green arrows from light.

Overlap with UK variant marked by blue arrows from right.

wt frequencies vs time

(except deletions, so only one overlap with UK variant)



South Africa variant never reached as high prevalence world-wide as UK variant. Excepting for Spike N501Y relative frequencies still vary more….

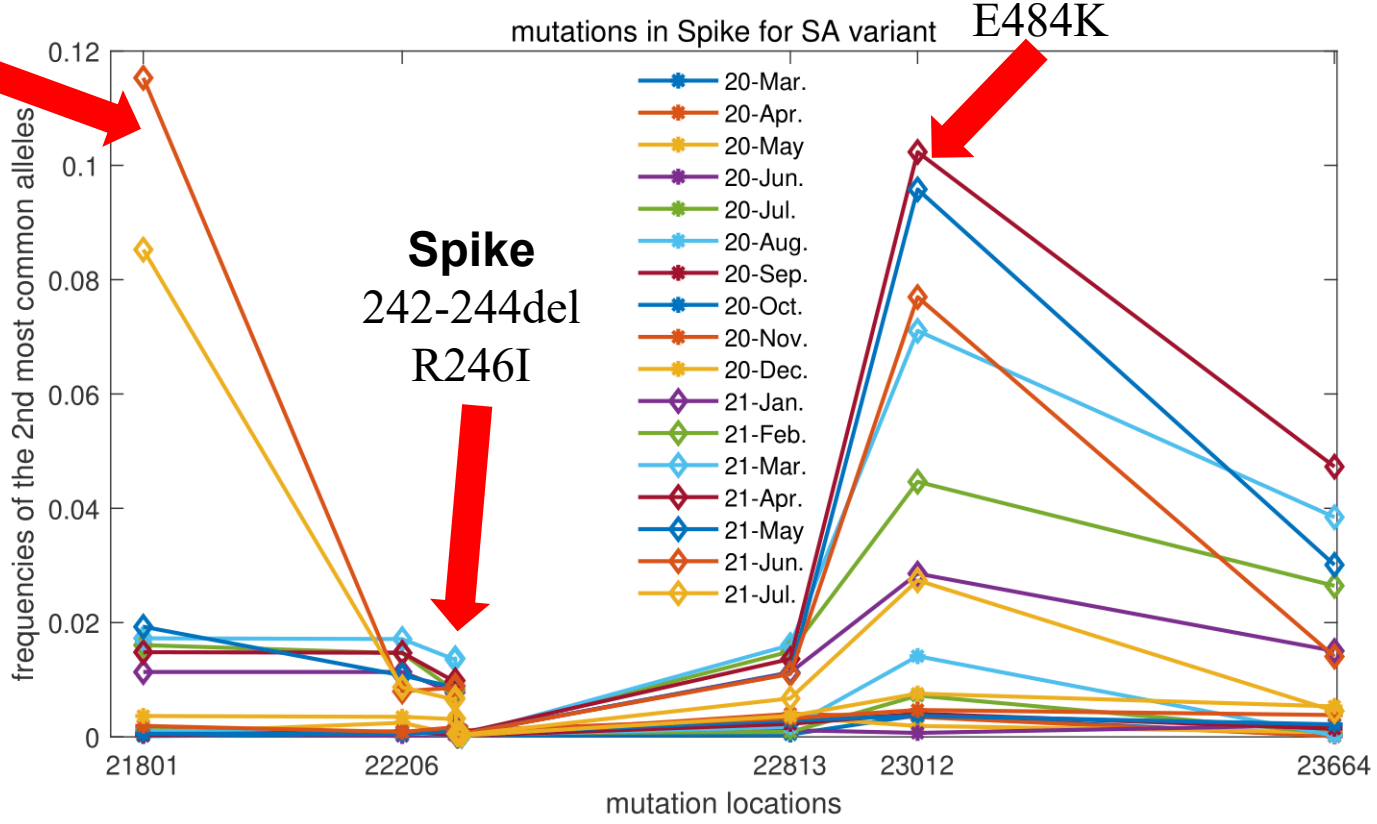
+ frequencies vs position

Spike
D80A

Spike
E484K

Spike
242-244del
R246I

N501Y (also in UK) not shown.



Frequencies for mutant alleles at different loci in Spike except A23063T (N501Y). Highest frequencies in different months at different loci. Some listed mutations grow very little.

ROYAL INSTITUTE OF TECHNOLOGY

“India variant” B.1.617.2 a.k.a. delta

Variant definitions from Nextstrain and Pangolin

We have not been able to find a list of mutations in Technical Briefings from Public Health England for this variant (only a link to a github page)

Mutations listed for India variant

gene	Nucleotide	amino acid
ORF1b	C14408T	P314L
	C16466T	P1000L
Spike	C21618G	T19R
	22028-22030	E156 del
	22031-22033	F157 del
	A22034G	R158G
	T22917G	L452R
	C22995A	T478K
	A23403G	D614G
	C23604G	P681R
	G24410A	D950N
M	T26767C	I82T
N	A28461G	D63G
	G28881T	R203M
	G29402T	D377Y
ORF3a	C25469T	S26L
ORF7a	T27638C	V82A
	C27752T	T120I

Nextstrain lists 18 non-synonymous mutations as 21A/delta. Those are the ones listed in the table.

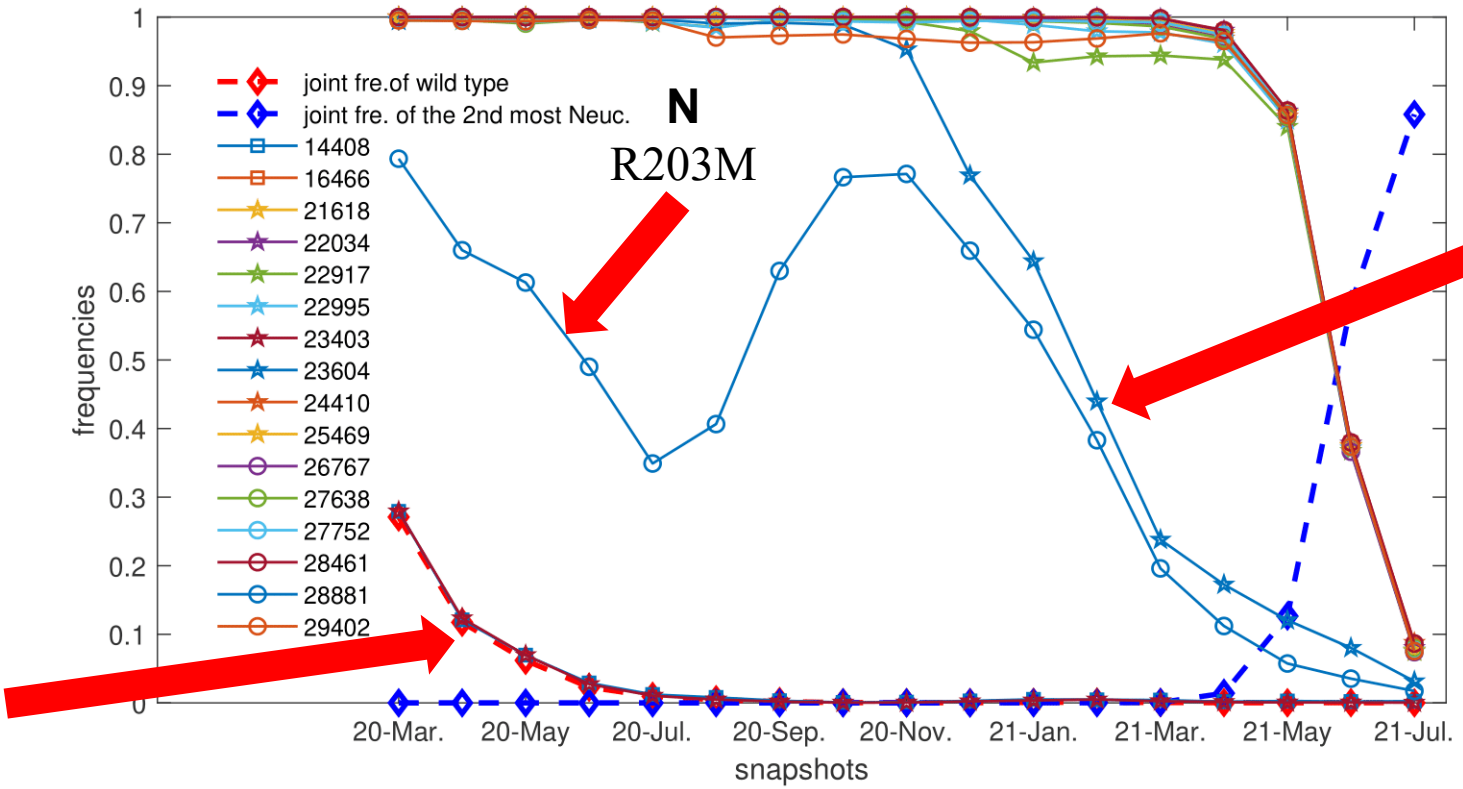
Of these 18 mutations 6 are missing in the definition in Pangolin B.1.617.2. The missing ones are marked by red arrows from left.

The black arrow is D614G. The blue arrow is a wt residue listed in UK variant (P681H/R).

Source Public Health England
phe-genomics/variant_definitions VOC-21APR-02
curators Natalie Groves, Jeff Barrett

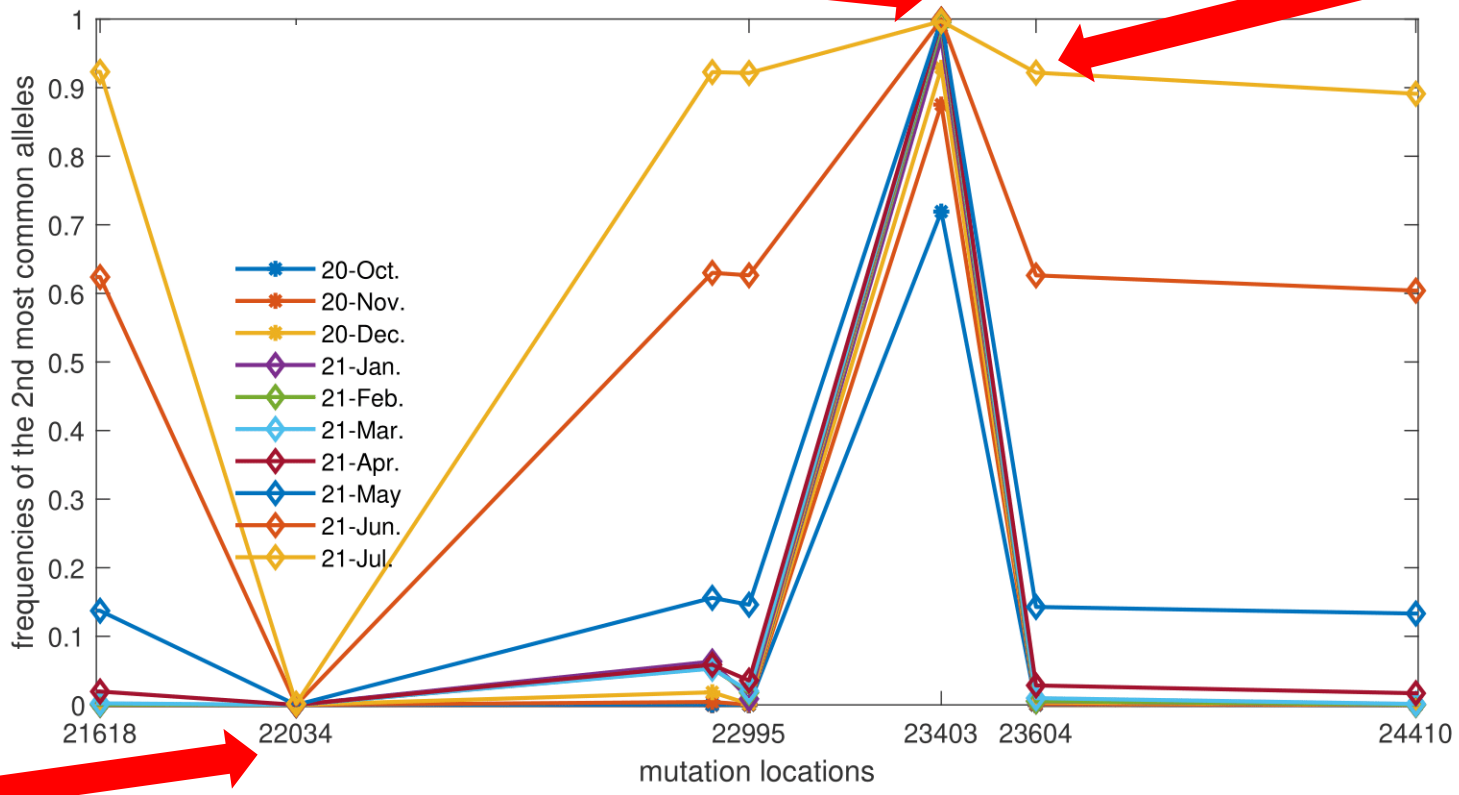
Public Health England lists the same variants as Pangolin, but separated into ”variants” (green) and ”additional mutations” (yellow)

wt frequencies vs time



India variant has reached high prevalence world-wide, and is now dominating. Except for Spike P681H/R (also in UK), listed mutations G28881T (R203M in N) and A22034G (R158G in Spike) differ.

+ frequencies vs position



Frequencies for mutant alleles at different loci in Spike.
With two exceptions, approximately uniform growth.

ROYAL INSTITUTE OF TECHNOLOGY

Observations

simple time series analysis

impossible before GISAID

some listed mutations must be wrong

the rest do not grow evenly
mixtures of clones?

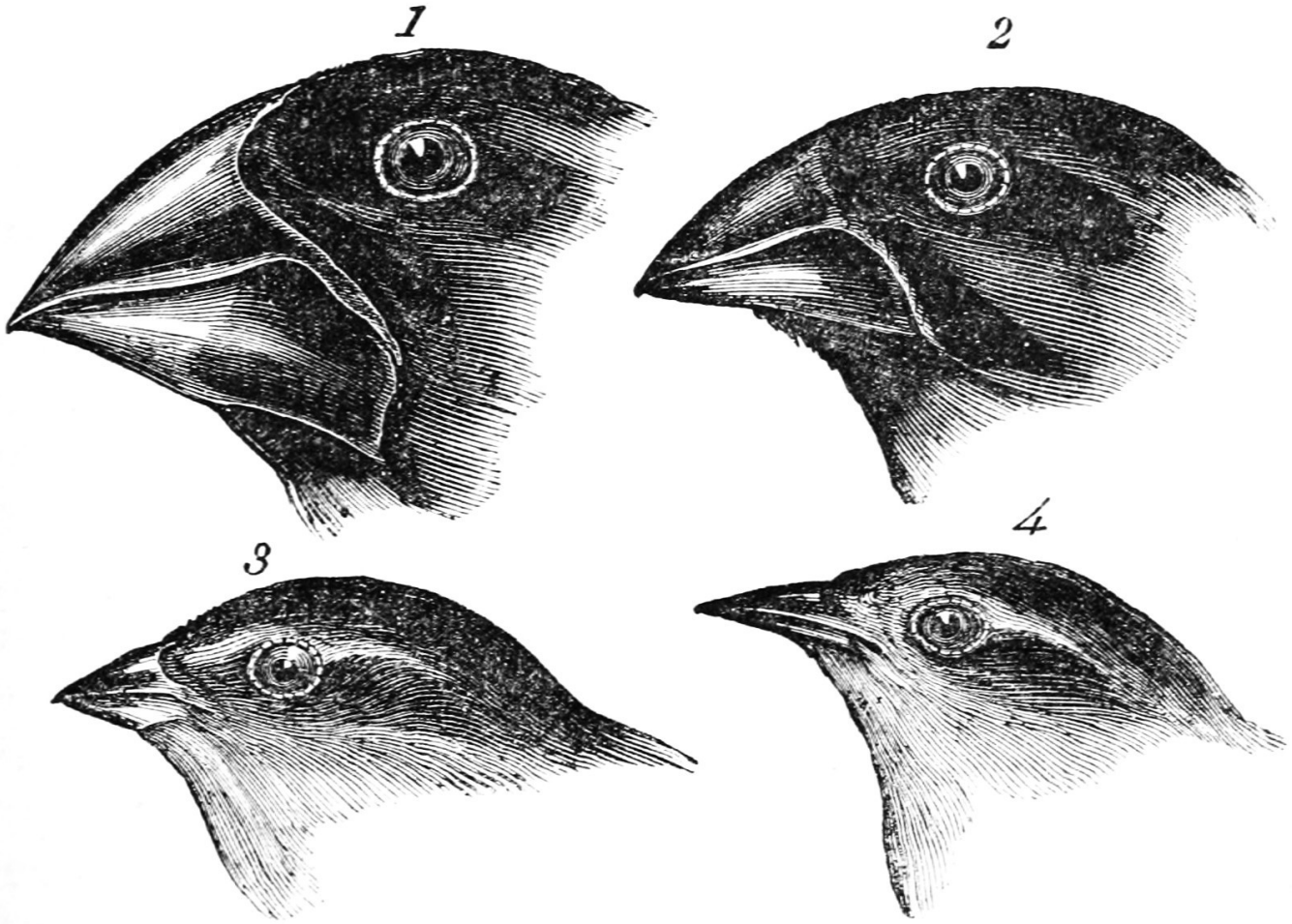
ROYAL INSTITUTE OF TECHNOLOGY

Reflexions of viral recombination

H-L Zeng, V Dichio, E Rodríguez Horta, K Thorell, EA
”Global analysis of more than 50,000 SARS-CoV-2 genomes reveals epistasis between eight viral genes”
PNAS **117** (49) 31519-31526 (December 8, 2020)

+ some later data

Force of evolution: natural selection



1. Geospiza magnirostris.
2. Geospiza fortis.
3. Geospiza parvula.
4. Certhidea olivasea.

Source: wikipedia

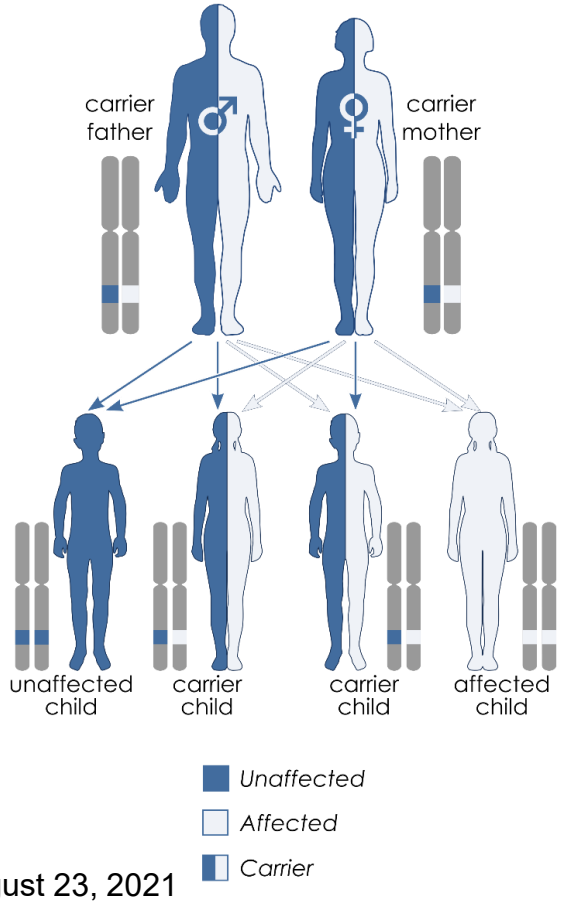
Force of evolution: mutations

Every row is here the amino acid sequence of a protein in a protein family. Some conservation, some changes (mutations and deletions).

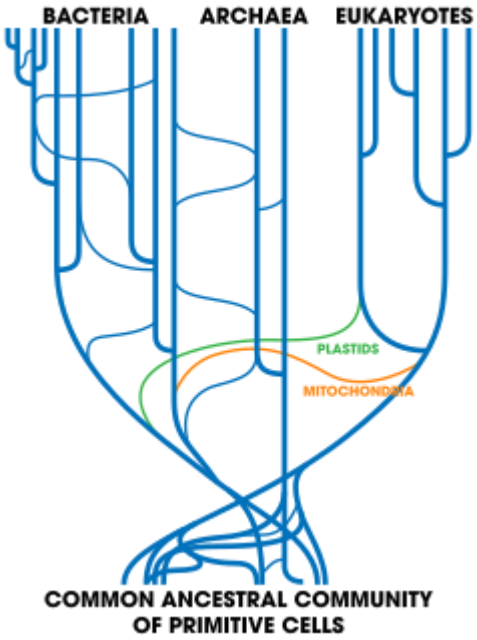
```
LLLDGSSSLPESYFDMMKSFAKAFISKANIGPHLTQVSVLQYGSINTID
LLLDGSSSLPASYFEEMKSFAKAFISKANIGPHHTQVSVLQYGSITTID
LLLDGSSGFPASEFDEMKSFAKAFISKANIGPQLTQVSVLQYGSITTID
FVLDGSSSVRASQFEEMKTFVKAFIKKVNIGVGATQVSVLQYGWRNILE
VLLDGSTNIMEPQFEEMKTFVKELIKKVDIGNNGTQISVVQYGKTNTLE
FILDTSSSVGKDNFEKIRKWVADLVDSFDVSPDKTRVAVVLYSDRPTIE
LAVDTSQSMEIQDLTVIKSVVDDFISHRK-N---DRIGLILFGTQAYLQ
FLVDTSGSLQKNGFDDEKVFVNSLLSHIRVSYKSTYVSVVLFGTSATID
LALDTSATTGETILDHITRGAQIGLAALS---DRSKVGVWLYGEDHRVV
YVIDTSGSMHGAKIEQTRESMVAILQDLH---EEDHFGILLFERKISYW
FLIDTSRSLGLRAYQKELQFVERVLEGYEIGTNRTKVAVITFSAGSRLE
ILLDTSSSIKINNFDLIRKFVANIINQFEVGRNGLMVGMATYS--RSVQ
FILDTSGSVGSYNFEKMKTFVKNVVDFFNIGPKGTHVAVITYSTWA--Q
FALDTSTSIGSQNFEREKQFVLAFVTDMDIGRSDVQVSVGTFSDNARRY
LLLDTSGSMQGAAIEALLSLKDEL-VKNSIAARRVEIAIVTFDSHINVV
LLLDTSGSMKGEPLDALRTFQQEL-DRDSLAKKRVEVAIVTFNSDVEIV
LSVDVSLSMLARRISALRDIAIRFVQKRK----NDRVGLVTYSGEALAR
LAMDVSGSMQANRLEAAKDVAISFINNRNIG-------MVTFAGESFTQ
MSVDVSLSMLARRLTALKNIAKKFVDKRP----GDRIGLVTYSGEAFTK
VLADVSGSMQGEPIAA-AAFTRYL-QNEV-ASKRVEVAVVTFGTVATVL
```

Force of evolution: recombination

Allows completely healthy offspring from not completely healthy parents



According to a well-known hypothesis by Woese it was very common between all organisms in early life on Earth.



Carl Woese
"The universal ancestor"
PNAS **95**:6854-6859 (1998)

Source: wikipedia

Recombination is very common in coronaviruses due to its mode of RNA replication

ROYAL INSTITUTE OF TECHNOLOGY

Recombination in vivo

An instability of clones is supported by recent observations pointing towards the emergence of multiple lineages of SARS-CoV-2 within the same individual. In all cases the patients had prolonged viremia and received convalescent plasma treatment and/or monoclonal antibody therapy.

Three selected references (there are more):

”Case Study: Prolonged Infectious SARS-CoV-2 Shedding from an Asymptomatic Immunocompromised Individual with Cancer”
Victoria A. Avanzato et al, *Cell* **183:**1901–1912 (2020)

”Prolonged Severe Acute Respiratory Syndrome Coronavirus 2 Replication in an Immunocompromised Patient”
Baang et al *The Journal of Infectious Diseases* **223**:23–27 (2021)

“SARS-CoV-2 evolution during treatment of chronic infection”
Kemp et al, *Nature* **592**:277–282 (2021)

Statistical genetics

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

Linkage equilibrium: sex and recombination mix up genetic variants (alleles) at different loci so they are independent.

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

Quasi-linkage equilibrium: small correlations remain because sex and recombination is not infinitely fast.

Kimura *Genetics* **52**:875–890 (1965)
Neher & Shraiman *PNAS* **106**:6866 (2009);
Rev Mod Phys **83**:1283 (2011)
Gao, Cecconi, Vulpiani, Zhou & E.A.,
Phys. Biol. **16** 026002 (2019)

The QLE distribution is like stat mech

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

additive effects

synergistic effects or "epistasis"

The parameters of the QLE distribution are determined by the forces of evolution, as we will see shortly. These parameters can also be determined from data by many methods developed over the last decade ("direct coupling analysis", DCA).

Jordan & Wainwright (2008)
Roudi, Aurell, & Hertz (2009)

Nguyen, Berg & Zecchina (2017)
Cocco, Feinauer, Figliuzzi, Monasson & Weigt (2018)

ROYAL INSTITUTE OF TECHNOLOGY

Main results from DCA

Hong-Li Zeng et al PNAS 2020 (Table III)

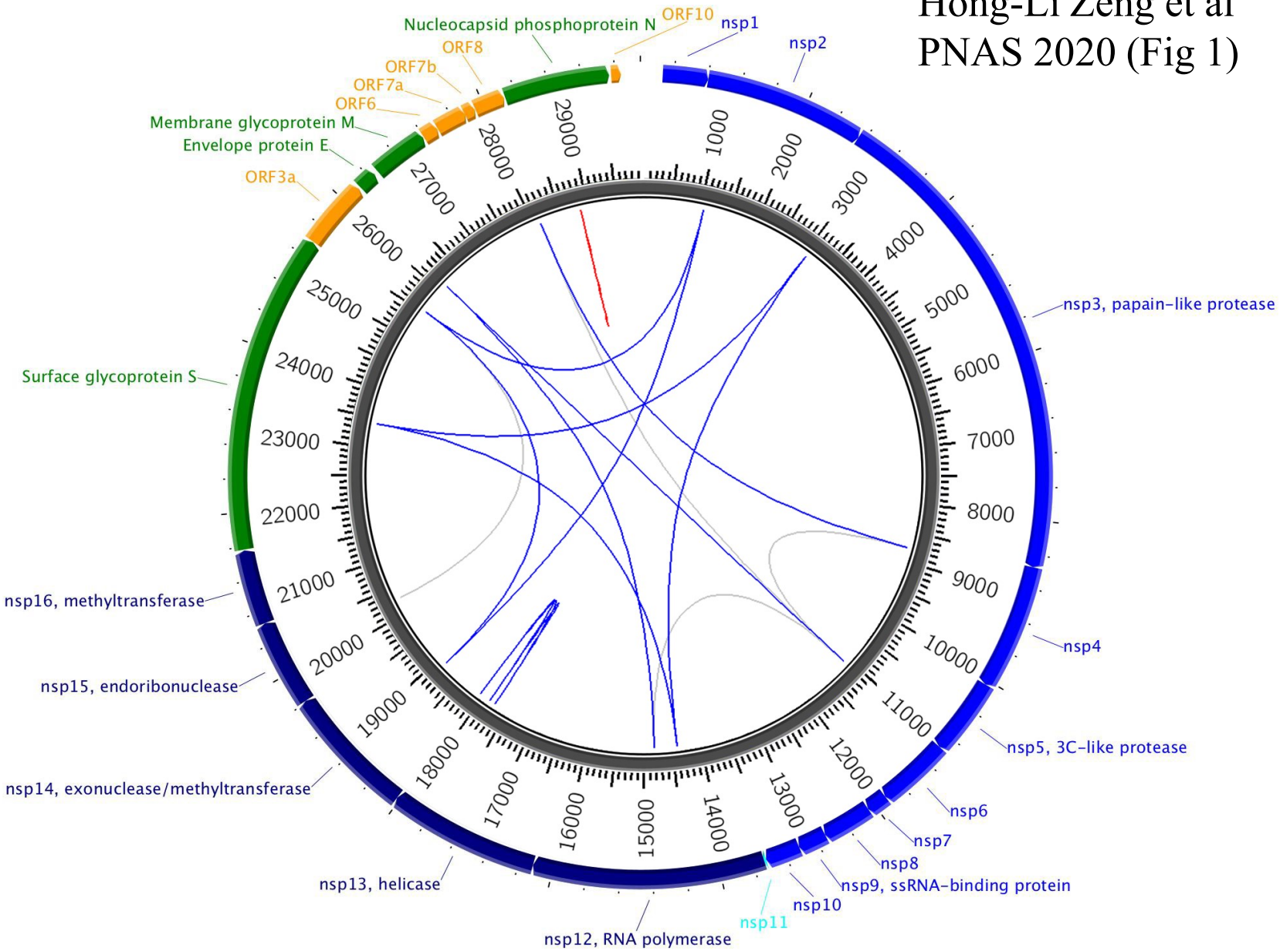
Rank	Locus 1	Mutation 1	Locus 2	Mutation 2
1	1059-nsp2	T85I	25563-ORF3a	Q57H
5	8782-nsp4	S76S	28144-ORF8	L84S
9	14805-nsp12	T455I	26144-ORF3a	G251V
21	1059-nsp2	T85I	18877-nsp14	L280L
26	17858-nsp13	T541C	18060-nsp14	L7L
27	17747-nsp13	P504L	17858-nsp13	T541C
36	17747-nsp13	P504L	18060-nsp14	L7L
47	11083-nsp6	L37F	26144-ORF3a	G251V

Large data boil down to short list (typical for DCA)

Viral gene ORF3a appears in several interactions

ROYAL INSTITUTE OF TECHNOLOGY

Hong-Li Zeng et al
PNAS 2020 (Fig 1)



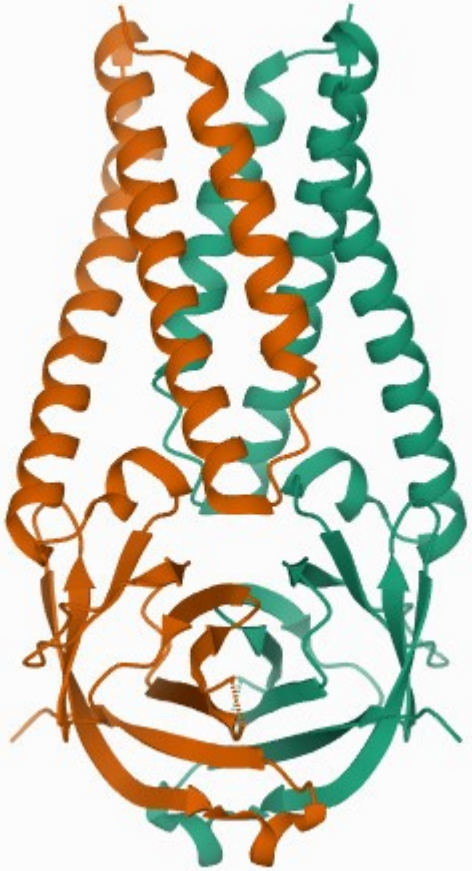
SARS-CoV-2 ORF3a

ROYAL INSTITUTE OF TECHNOLOGY

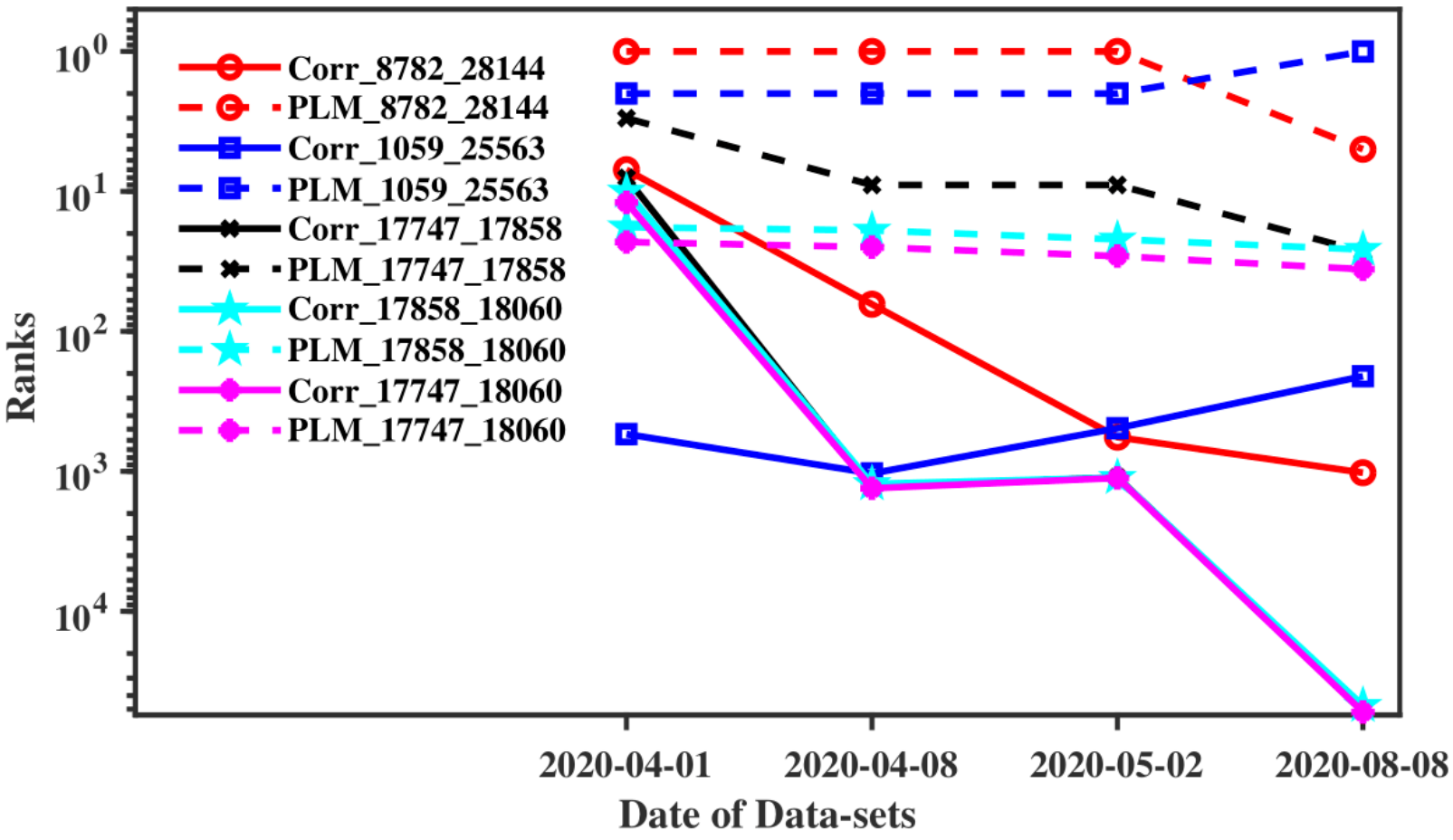
6XDC (Kern at et, bioRxiv)
Cryo-EM structure deposited
2020-06-10

In SARS-CoV, ORF3a up-regulates expression of fibrinogen subunits FGA, FGB and FGG in host lung epithelial cells and activates the NLRP3 inflammasome.

Unfortunately there are for the moment no listed approved drugs expected to target ORF3a or its human interactors (Gordon et al, Nature 2020)

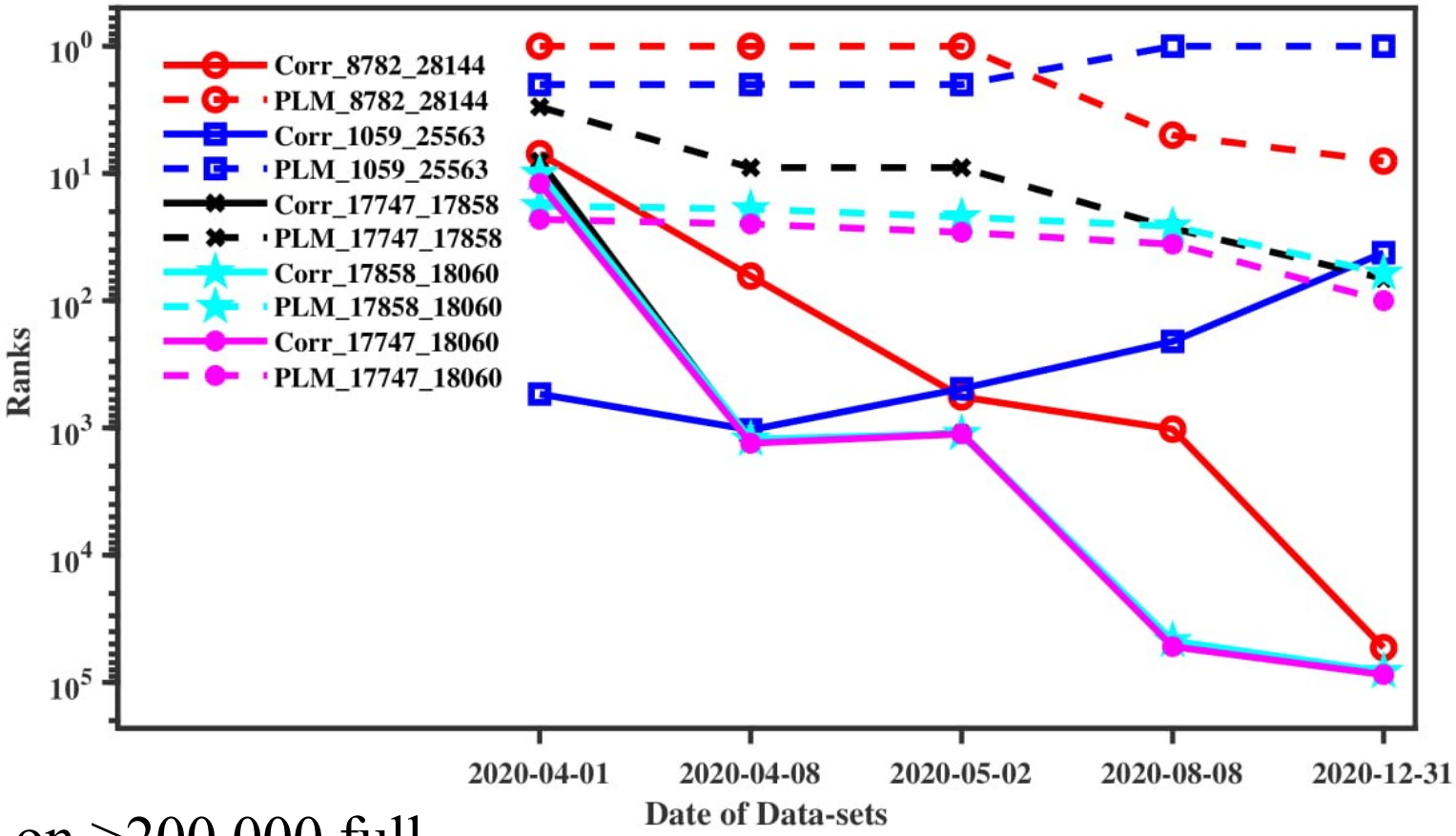


DCA predictions are stable over time



Hong-Li Zeng et al PNAS 2020 (Fig 4)

And have largely remained so



Analysis on >200,000 full-length genomes on GISAID by end of 2020

Hong-Li Zeng et al [2021, in preparation]

ROYAL INSTITUTE OF TECHNOLOGY

Morals of the story

generative models are more stable than correlations as measures of statistical dependency in GISAID data

consistent with recombination having an effect in the global viral population

there is more to SARS-CoV-2 than Spike

ROYAL INSTITUTE OF TECHNOLOGY

Thanks

Hong-Li Zeng
Vito Dichio
Edwin Rodríguez Horta
Kaisa Thorell

National Science Foundation of China (NSFC); National Science Foundation of Jiangsu Province

EU MSCARISE-2016 ”Infernet”

Yue Liu
Rickard Nordén

Swedish Science Council (VR)

Martin Weigt

GISAID

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

The distribution of genotypes in a population changes according to **selection**, **mutation**, **genetic drift** (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 quasi-linkage equilibrium

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 $(r c_{ij})^{-1}$.

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

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

$$F^{nMF} = \sum_i H\left(\tfrac{1+m_2}{s}\right) + H\left(\tfrac{1-m_2}{s}\right) + \sum_i h_i m_i + \sum_{ij} J_{ij} m_i m_j \qquad H(x) = -x \log x$$

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

$$\chi_{ij} = \frac{\partial m_i}{\partial h_j} = c_{ij}$$

Exact, a fluctuation-dissipation relation. An immediate result for pairwise exponential models.

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

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):

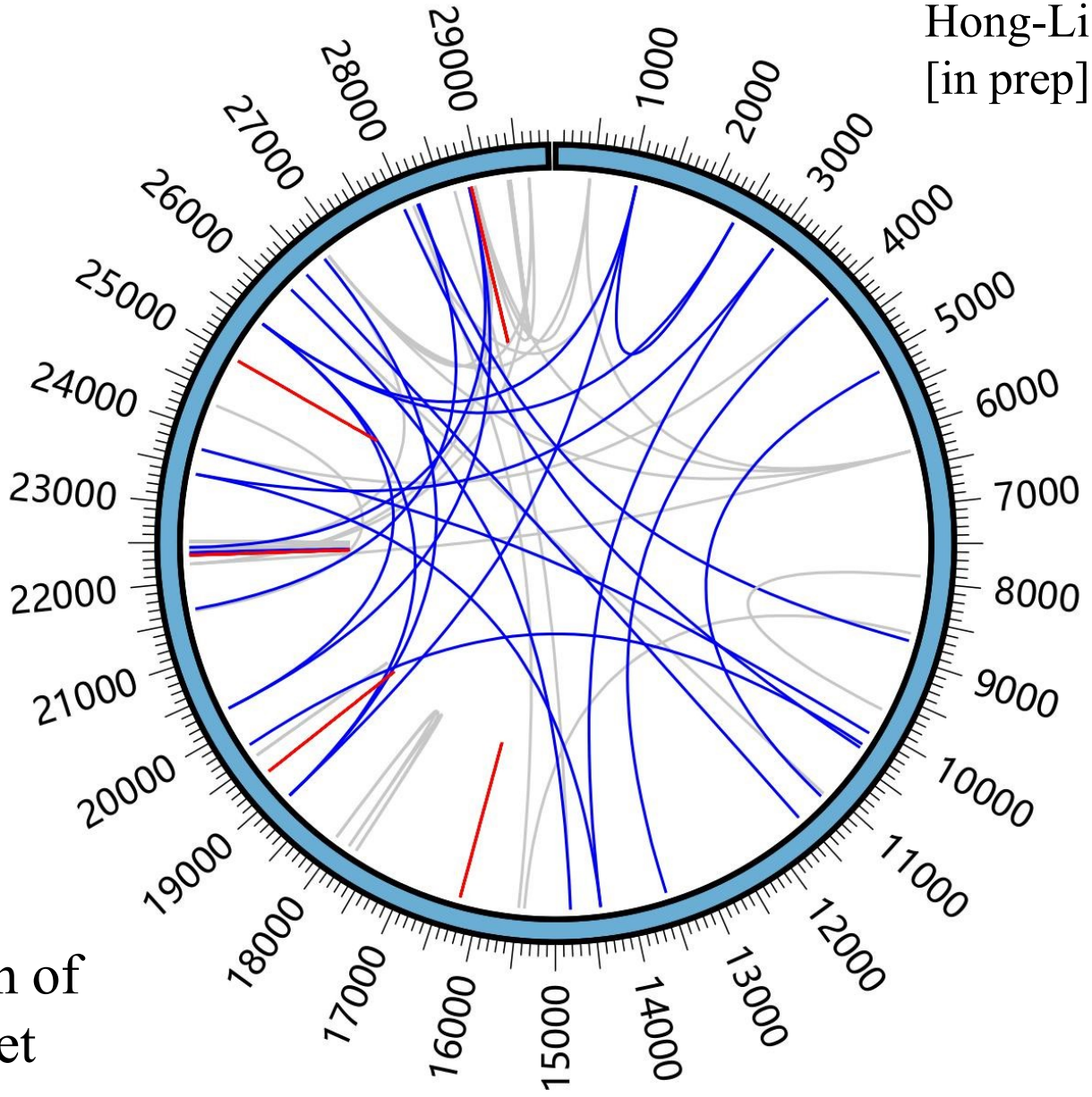
$$P(S_r \mid S_{\setminus r}) = \exp\left(h_r S_r + \sum_l J_{rl} S_r S_l \right) \Big/ \sum_y \exp\left(h_r y + \sum_l J_{rl} y S_l \right)$$

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

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)

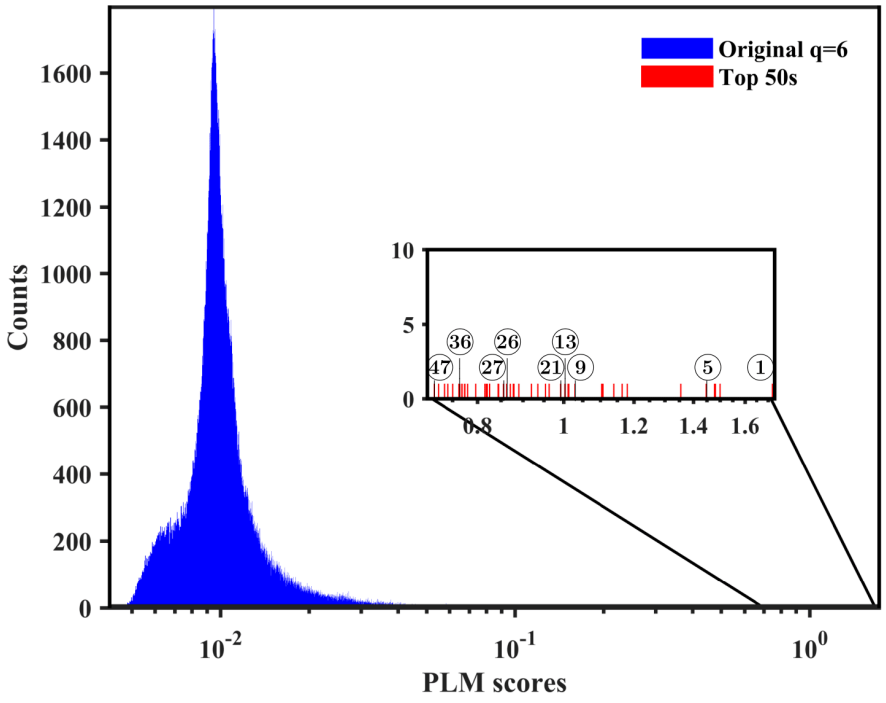
ROYAL INSTITUTE OF TECHNOLOGY

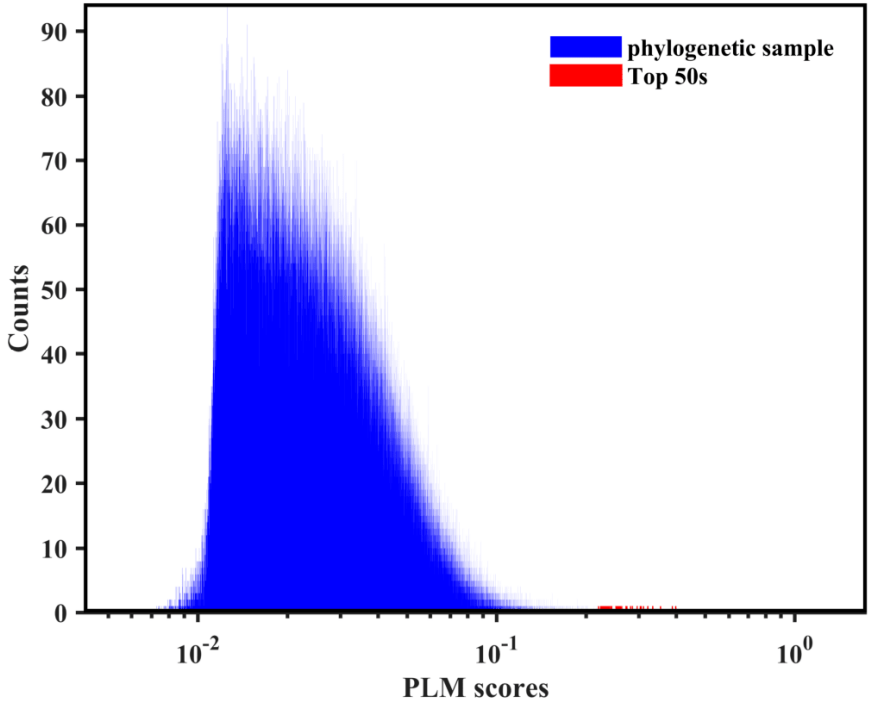
Hong-Li Zeng et al
[in prep] 2021



Visualization of larger data set

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

QLE vs clonal competition

Neher & Shraiman *PNAS* **106**:6866 (2009); *Rev Mod Phys* **83**:1283 (2011);
Neher, Kessinger & Shraiman *PNAS* **110**:15836-15841 (2013)

Mixture models
Multi-genome distributions are complex

Exponential models
Multi-genome distributions factorize

Standard DCA makes no sense

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

$$P(\boldsymbol{g}_1, \boldsymbol{g}_2) \not\approx P(\boldsymbol{g}_1) \cdot P(\boldsymbol{g}_2)$$

DCA may work

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

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