מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

Thesis for the degree

Master of Science

Submitted to the Scientific Council of the

Weizmann Institute of Science

עבודת גמר (תזה) לתואר

מוסמך למדעים

מוגשת למועצה המדעית של

מכון ויצמן למדע

By

Jonathan Grozovski

מאת

יונתן גרוזובסקי

בחינה חישובית של טעויות תרגום, מוטציות ב-SARS-Cov-2; והפוטנציאל החיסוני שבהם

In-silico investigation of translation errors, SARS-Cov-2 mutations; and their immunization potential

Advisor:

Prof. Yitzhak Pilpel

מנחה:

פרופ' יצחק פלפל

February 2022

אדר תשפ"ב

Table of Contents

Abstract

Translation errors, a type of phenotypic mutation, affect all domains of life at any given time, yet little is known on their contribution to the evolution of early life, and specifically to the evolution of the Standard Genetic Code (SGC). In this work we evaluate the cost both genetic and translation errors have on a genetic code. We use computational analyses to compare the SGC to hypothetical genetic codes and - by means of their error costs - to explore the fitness landscape of genetic codes.
Affected by the outbreak of the COVID-19 pandemic, our work turns to explore the correlation between emerging viral variants and their clinical properties. We suggest a method to detect the effects of individual Single-Nucleotide Polymorphisms in the viral genome on the phenotypes exhibited by the infected host, akin to methods from the field of Genome Wide Association Studies.
Lastly, our work brings these seemingly unrelated fields together by proposing a new methodology for vaccine design. This approach relies on translation errors to deliver a multiple variant vaccine, based on a single mRNA molecule. We suggest designing the mRNA template according to the SNPs with the highest potential to damage the host.

On the Potential of Evolutionary Effects of Translation Fidelity on the Genetic Code

Introduction

All throughout life, survival and reproduction of an organism are dependent on its ability to carry out evolutionarily developed plans and responses with temporal, spatial and structural precision. At an almost contradictory trajectory, the potential for improvement and adaptation to new or changing environments depend on the organism's "failures". Errors made by the molecular machinery of the central dogma are essential for flexibility, adaptation and evolution. These errors - specifically transient ones - are at the focus of this work.

Errors in the DNA, colloquially known as mutations, have long-lasting effects on the organism and its offspring. They affect all copies of the proteins or functions encoded by the mutated locus in the organism, and are consequently inherited to any offspring that may inherit that mutation. In contrast

to genetic mutations, errors in downstream steps of the central dogma of molecular genetics, tend to be more transient. For example, a mistranslated protein has a limited impact in space and in time. If the protein has a small number of interactions or is quickly degraded its impact will be small. A prion or a protein transferred to a daughter cell during mitosis will have a longer effect, yet still limited in comparison to DNA mutations.

A key component of the central dogma, and hence in the prevention or contribution to errors in it, is the genetic code. While scientists conveniently described it as a compact 4x4x4 table, the genetic code consists at the molecular level of a complex network of dozens of molecules. All of these work in unison to ensure the faithful transition of information from the genetic to the protein level. The number of known genetic codes utilized by different forms of life is exceedingly low and all are very similar (Watanabe and Suzuki 2008), especially considering the vast number of possible codon - amino-acid combinations conceivable by combinatorics. The Standard Genetic Code (SGC) is used in all domains of life, with the most notable divergence of the SGC appearing in mitochondria (Watanabe and Suzuki 2008). The complexity of the molecular network and the uniqueness of the genetic code suggest that it has either been under tremendous evolutionary pressure, was subject to a founder's effect at the early stages of life on earth, or both.

One way to quantitatively assess how well the genetic code has adapted through evolution is to computationally compare it to other hypothetical genetic codes, by giving each one a score. It has been previously shown that the genetic code is robust to genetic mutations by using a similar strategy (Freeland and Hurst 1998). More recently the genetic code was assessed and was shown to be optimally robust to the effect of mutation on the nitrogen consumption of the encoded amino acid (Shenhav and Zeevi 2020). The genetic code could have also been optimized with respect to "phenotypic mutations", namely the errors that the transcription and translation machineries are prone to make (Mordret et al. 2019). A hypothetical genetic code could be considered robust to phenotypic mutations if amino acids, correctly and incorrectly translated by that genetic code, would tend to be chemically similar.

In our attempt to evaluate the robustness of the genetic code to phenotypic mutations we rely on previous work done in our lab. Namely, a systematic mapping of translation errors (Mordret et al.

2019). In brief, in this work our lab unbiasedly extracted proteins from bacteria and yeast, degraded them to short peptides and determined their amino-acid sequence, using Mass Spectrometry. The identified peptides were aligned to the appropriate genomes and sorted into two groups – sequences that could be perfectly aligned to a genome encoded peptide, and sequences that could be aligned perfectly, except for one mis-aligned amino-acid. Peptides that did not fall into either category were excluded. A careful comparison of these two groups revealed the identity (genomic location and type) and frequency of translation errors in each of the model species. This provided the comprehensive and unbiased data required to conduct the evaluation of the role these errors played in the evolution of the genetic code.

In this chapter we employ this data to explore by computational analysis the fitness landscape of the genetic code, with respect to mutation and translation error costs, and the possible forces behind its evolution. Our data suggest that the SGC is more optimized for robustness to genetic mutations than to translation errors, but give strong indications of evolutionary pressure exerted on the SGC towards avoidance of translation errors as well.

Results and Discussion

Before exploring possible alternatives of the Standard Genetic Code (SGC, Fig 1A left table) we considered some of its key features. As a conversion mechanism, the genetic code adheres to three basic conditions: (1) all amino-acids and the STOP codon must be assigned to at least one codon, (2) no codon can encode more than one amino-acid, (3) no codon remains unused. Conditions (2) and (3) may be eased at the expense of ambiguity and restriction of the genomic codon usage, respectively. Therefore, in order for genetic codes to be useful as accurate conversion mechanisms at a similar degree to the SGC, they must fully comply with these conditions.

One of the most noticeable features of the SGC is its ”box structure”. It is degenerate in an organized fashion, such that most codons that encode for the same amino acid are grouped into “boxes” and share similarity in nucleotide sequence, often in the first two codon positions. Utilizing the wobble capability, which allows a mismatch between a codon’s 3rd position and the anti-codon’s 1st position, this feature is a prominent one in making the genetic code robust to errors. Any hypothetical genetic code that would not feature this mode of organization will likely be significantly less robust than the SGC. A higher level of organization acquires the SGC further robustness to mutations since

chemically similar amino acids tend to be encoded by codons with similar nucleotide sequence. The implication of these two properties ("boxiness" and chemical proximity) is that relative to a totally random code, that would still fulfill the above three requirements, the SGC often ensures that when mutations occur, they have a higher chance to still encode the same, or at most a chemically similar, amino acid.

Hence, to conduct a proper assessment of the robustness of the SGC to mutations, that is done in comparison to a null model of alternative - e.g., randomly generated GCs – we needed nonetheless to consider only GCs that preserve additional properties of the SGC, i.e., its organized degeneracy. Such genetic codes are easily generated by randomly relabeling the codon boxes (see Methods). With this method, the number of possible genetic codes is the number of all possible unique re-assignments between blank codon boxes (Fig 1A, right table) and amino-acids. Since 20 amino-acids exist, and none are repeated, we get 20! (factorial) possible such combinations, amounting to ~10^{18} hypothetical genetic codes.

Having established the boundaries for hypothetical genetic code alternatives, we then turned to focus on the method for comparing them to one another. We required a method that would give a single score to an entire genetic code, based on its robustness to errors.

Regardless of their origin, being phenotypic or (non-silent) genetic, errors to coding regions have consequences at the protein level. Be it structure integrity or reactivity, the incorporation of a wrong amino-acid can harm the functionality of a protein. Based on this property, we used the misincorporation of an amino-acid as our basic unit of error. Summing up all possible amino-acid misincorporations per genetic code would yield a single score of its robustness to error.

The comparison between any two amino-acids can be done by a variety of quantitative measures. The question we are addressing requires a measure that does not depend on mutual tendency to mutate into one another, as these already reflect the existing genetic code. The simplest measures, and those that necessarily predated the genetic code - and biology in general - are physical and chemical ones. Hence, we chose to use an amino-acid similarity matrix that is based on their physico-chemical properties. In particular, we used the Miyata distance matrix that focuses on the volume and polarity of the amino-acids as a genetic code free measure of similarity between amino acid pairs (Miyata, Miyazawa, and Yasunaga 1979).

The next step in our calculation was to find the weight each possible amino-acid replacement will have on the overall robustness score of a genetic code. This could easily be done by calculating the probability of that error to occur. The probability for each amino-acid replacement to occur consists of three error components - the codon, the reading frame, and the specific nucleotide replacement. The values these components take differ based on the molecular level at which the error occurs – the genetic or the ribosomal: At the genetic level, since the DNA Pol I is oblivious to both codons and reading frames, the probability for a codon to have an error is the probability that the codon was used in the genome, and hence it is given by the codon usage bias; the frame probability is equal to 1/3 in all frames; the probability of the nucleotide in that frame to be misread as another one is given by the error propensities of DNA Pol I (Fig 1B). At the ribosomal level, the codon probability can be obtained from the error spectrum (Fig 1D and Methods); both the probability that an error will occur at a certain frame, and the probability of a specific error type are given by the ribosome error propensity, as it appears in the ribosome confusion matrix (Fig 1C).

To illustrate this point, let us consider the probability that the ACT codon will be misread as the CCT codon. This probability is computed by the multiplication of three terms (see Eq. 1): the probability that the ACT codon had an error, the probability that this error occurred at the first position of the codon, and the probability that A was misread as C in that position. An error of this type can occur in both molecular levels. At the proteomic level, this error can occur due to the ribosome misreading the codon and erroneously incorporating a GGA t-RNA at an ACT codon. Multiplying these three terms we can obtain the probability of this phenotypic error. Separately, we can also calculate the probability of the equivalent genetic mutation - in which the A would be erroneously copied as C by the DNA Pol I, creating a codon change that will result in the incorporation of the wrong amino-acid.

Eq.1

$$P(ACT \rightarrow CCT) = P(ACT) * P(error\ in\ frame\ 1) * P(A \rightarrow C\ in\ frame\ 1)$$

To calculate a single score for the robustness to errors of a whole genetic code, the real one or a hypothetical one, we used the above probability computation, together with the amino-acid similarity metric. Firstly, we calculated the probability for each codon-to-codon error. Different genetic codes, computationally generated or the real one, map these codon-to-codon errors to different amino-acid to amino-acid errors. Hence, the codon-to-codon errors are translated to different amino-acid to

amino-acid errors by each GC. Following the same example we used above, an ACT->CCT codon error is mapped by the SGC to a Threonine->Proline amino acid replacement error. In a different, hypothetical, GC it may be an Alanine->Arginine error, or any other pair.

The Miyata matrix allowed us to assign a similarity score to every error in amino-acid space. The Miyata distances between amino-acids increase for pairs that are farther apart in their volume and polarity. Hence, it is useful to think of this scoring method as the overall error cost for a GC. The farther apart the distance between erring amino-acids, the bigger the cost of the error. Multiplying one error's probability by its individual similarity cost gives the partial contribution of that error type to the overall cost of errors in a particular GC (see Eq. 2). Summing all possible errors together yields a weighted average of all errors and their effect in a single cost value (see Eq. 3). This procedure is blind to the origin of the codon-to-codon error (be it genetic or proteomic), and therefore we could effortlessly repeat it in both levels, with the only difference between the two being the error probabilities.

Eq. 2

$$Cost(ACT \rightarrow CCT) \Rightarrow$$

Translation by SGC:

$$\Rightarrow P(ACT \rightarrow CCT) * dissimilarity(Threonine \rightarrow Proline)$$

Eq. 3

$$Total\ Cost = \sum_{i}^{codon-list} \sum_{j}^{codon-list} cost\ (codon_i \rightarrow codon_j)$$

(where 'codon-list' consists of all 61 non-STOP codons)

Applying this scoring system to the SGC, or any hypothetical GC, can be seen as an exploration of the "fitness" landscape of genetic codes. Instead of genes or proteins, the genetic codes that are assessed here differ by their codon to amino-acid assignments. In this exploration of the landscape, we assess the fitness of an organism by its ability to accurately produce all proteins, based on the entire genetic code. We obtained these scores twice for every GC - with regard to genetic mutations and translation errors.

As with fitness landscapes of genes, there are two useful approaches to mapping them. The first, by randomly sampling the domain - in our case, genetic codes - and calculating their score. For the randomized sampling of the landscape, we used random GCs, following the previously discussed "organized degeneracy" scheme.
The second, by making small perturbations to a point of interest. In our case of the SGC, the small perturbations were single (1st order neighbours) or double (2nd order neighbour) amino-acid assignment swaps. An example for such a GC is a code in which the two codons for Phe (UUU and UUC) and two codons for Leu (UUG and UUC) are swapped such that in the minimally swapped GC UUU and UUC encode for Leu, and UUG and UUC encode for Phe. Since in our terminology this is a single swap, similarly swapped GCs were named 1st order neighbours, or 1-neighbours. We also explore the broader vicinity of the SGC by applying two such swaps in a single GC, relative to the SGC. following the same convention, these are named 2nd order neighbours, or 2-neighbours of the SGC.

When comparing the SGC to random codes we could clearly see it is robust to genetic mutation (Fig 2A), as expected. It was placed second out of 10,000 randomized codes in robustness to mutations. This observation resonates very well with literature of the SGC that indeed have shown in the past that it is optimal in minimizing the effect of genetic mutations on amino acid properties (Freeland and Hurst 1998). However, in its robustness to translation errors, while still above average, the SGC didn't fare as well, compared to randomly generated hypothetical genetic codes. It was placed within the top 1st percentile of the 10,000 hypothetical GCs (Fig 2A). So, while the SGC is "one in a 100" in robustness to translation errors, it is "one in 10,000" in robustness to genetic mutations. Note that the robustness of the randomly generated codes to genetic mutations and translation errors are highly correlated (Fig 2A-C), namely that codes that are robust to genetic mutations tend to be robust also to translation errors, and codes that are not robust to genetic mutations would typically not be robust to translation errors. This correlation likely stems from the nature of errors in question. Both DNA mutations and translation errors that are considered in this analysis manifest as a single nucleotide change in every codon. Hence, the types of errors possible within each GC table are limited. Specifically, errors could cause a misincorporation of amino-acids on the same row or column as the original acid, but not one that is in a diagonal direction (since these are two or three nucleotide changes away). Being an inherent, global attribute of a GC, this restriction at the level of amino-acid

errors is identical for our mutation and translation error cost calculations and therefore a likely a substantial contributor to the observed correlation. Another possible contributor to the observed correlation is a correlation at the codon error probability level. Genomic codon bias determines which codons are more translated, and are therefore more likely to also encounter a translation error event. Thus, the genetic codon bias propagates into the translation error spectrum. Given this correlation, the mere fact that the SGC is robust to genetic mutations guarantees a certain level of robustness to translation errors, which is not necessarily above its observed robustness to translation errors. Hence, based on this analysis we could conclude by suggesting that the SGC may have evolved to minimize costs of genetic mutations, and the extent of its robustness to translation errors was obtained as a by-product.

Exploring the SGC's one- and two-neighbours revealed a similar pattern (Fig 2B, 2C) to the one we saw with the random GCs. Even in its local "neighbourhood" the SGC is among the best in terms of robustness to genetic mutation (Fig 2D, 2F). In its robustness to translation errors, it attains a score similar, or a bit higher, than the majority of the minimally perturbed codes (Fig 2E, 2G). Comparing the random GCs with the neighbours as a group, we observed a stark difference between the mutation and translation error scores. None of the one-neighbours, and only a small minority of two-neighbours, had a mutation error cost higher than the average error cost of the random codes (Fig 2D, 2F). Contrarily, on the translation error axis, a noticeable number of one- and two-neighbours had higher error costs than the average of the randomly generated GCs. This would suggest that on the robustness to DNA mutations fitness landscape the SGC is located on a broad peak, where similar codes are likely to be robust to errors. On the robustness to translation errors fitness landscape, the SGC seems to be placed on a steep cliff or a rough landscape, with even small perturbations holding the potential for considerable drops in performance.

Perturbing the model to explore the fitness landscape of genetic codes is not limited to the GCs we use. One further way to enhance our understanding of the genetic code was to modify in a controlled fashion the underlying parameters used by the model. Thus, we were able to examine their contribution to the results we observed so far. Scrolling back to Eq. 1, we can see that the immediate contenders for such perturbations are the different components that make up the individual codon-to-codon error probabilities. Firstly, we used uniform codon probabilities to investigate their effect on

overall error costs in both error types. Replacing the real codon probabilities by a uniform probability as the first term in Eq 1., one at a time (Fig 3B, 3C), or both together (Fig 3D), we repeated the procedure described above to calculate the error costs of the real and hypothetical GCs. In this approach, we relied on the hypothetical GCs to provide context for the changes we observed in the error score of the SGC. For instance, comparing the average and standard deviation of error scores on the DNA mutation axis (compare Figs 3A vs. 3C; Figs 3F vs. 3H) we could see that almost none of the GCs – random or neighbours of the SGC – were heavily affected by a different codon usage bias at the genetic level. This came as no surprise to us, as examples of this phenomenon exist in nature - different species have different codon biases, yet preserve the same genetic code. This could suggest that the "boxy" structure of the genetic code is a contributing factor to its robustness to DNA mutations under varying codon biases. Contrary to that, replacing the codon error occurrence by a uniform probability of codons causes a major shift in both the average and standard deviation of translation errors costs (compare Figs 3A vs. 3B; Figs 3E vs. 3G). Even amongst the one-neighbours of the SGC we can see a consolidation of the higher cost tail of one-neighbours towards their mean. When comparing the SGC to itself across conditions, its translation error cost seems to be less affected by the uniform codon error occurrence, then by the uniform codon usage bias. Together, we take these findings to imply that codon error probability has the potential to significantly affect translation errors' cost, even among GCs that are extremely similar to the SGC. Furthermore, under the tested parameters, the SGC and a small sub-group of its neighbours remain uniquely resilient to translation errors. Additional exploration of GCs, under more codon usage conditions, may be required to establish this hypothesis and to elucidate the factors that grant the SGC this unique property, if indeed it exists.

Moving on to the next terms of Eq. 1, we repeated the parameter perturbation analysis for the nucleotide replacement matrices. This time, replacing each of the ribosome confusion matrix and the DNA pol error propensity by a uniform distribution, and once more calculating the error costs of hypothetical GCs (Fig 4A-D). At the DNA mutations level, plugging in a uniform DNA-pol error propensity matrix caused the costs of all simulated GCs – random or 1- or 2-neighbours – to slightly increase (compare Figs 4A vs. 4C; Figs 4F vs. 4H). In contrast, at the translation error level, using a uniform confusion matrix produced a dramatic reduction in both average cost per GC and the standard deviation among the generated codes (compare Figs 4A vs. 4B; Figs 4E vs. 4G). At first

glance, this result seems to suggest that, contrary to evolutionary logic, the SGC (or any GC, for that matter) would fare better in a world where errors occur completely at random without preference for some replacements. However, it is vital to remember that this is a highly hypothetical condition. In reality, the errors of the ribosome are dictated by the physical and chemical properties of the molecules involved in the process of translation. Meaning that a completely uniform error pattern could not exist, and that the SGC has evolved in light of the true error propensity of the ribosome. Hence, we can understand that in real life conditions, most GCs have much higher costs than they would have had if the ribosome was to err in a random fashion. Therefore, our analysis illuminates the fact that the SGC has such a low overall translation error cost, relative to the many hypothetical alternatives.

This conclusion was further emphasised by a more complex type of perturbation of the ribosome confusion matrix. In this analysis, we shuffled the order of frames in the ribosome confusion matrix (Fig 5A), to produce 1 real and 5 hypothetical confusion matrices. When comparing the SGC to itself, under these shuffled options, the real confusion matrix caused it to result in almost the highest error cost of all 6 options (Fig 5B). However, when comparing the SGC to other GCs, under similar conditions, it becomes evident that it has an error cost that is at worst better than average, if not superior (Fig 5C-H).

Taken together, our results suggest that between DNA mutations and translation errors, robustness to mutations was the stronger evolutionary force in the shaping of the genetic code at the early stages of life on earth. Nonetheless, we can also observe that translation errors are not a negligible force in this respect either. The genetic code, it seems, was shifted by evolution on both fitness landscapes, either in parallel or in consecutive order, with mutations setting the primary direction and translation errors making the finer adjustments.

Methods

Genetic codes randomization

Hypothetical genetic codes were created while preserving the “box structure” of the SGC (Fig 1A). Amino-acids were severed from their codon boxes and compound groups and re-assigned in a random fashion. The total number of combinatorically possible GCs of this structure is 10^{18}, and hence cannot be covered in a feasible computational time.

Codon error occurrence

was calculated by summing up all the errors originating from each codon, and dividing buy the total number of errors recorded in the error spectrum, thus producing the probability for errors to appear in every codon (see Fig 1D).

Figure legends

Figure 1 - Table permutations

A: The Standard Genetic Code (SGC) (**left**), and a blank codon table (**right**) used to create random GCs. The table contains 20 blank, uniquely coloured boxes and compound groups that were randomly populated by the 20 AAs for each hypothetical GC such that every colour is populated by a different AA. All codons are grouped as they are in the SGC. Stop codons were not randomized or reassigned.

Confusion matrices of the DNA-pol-2 (**B**) and the ribosome (**C**).

D: A table containing all codon to AA translation errors (**left**) detected by (Mordret et al. 2019). By summing the rows, we can attain the probability an error will occur at each codon (**right**).

Figure 2 - Genetic code landscape exploration

Scatter plots of error robustness scores of GCs. The scores on the X-axis measure robustness to DNA mutations, while the Y-axis denotes robustness scores to translation errors. The curves on the axes (top and right of each plot) show the marginal distribution of the scores on the X- and Y-axis, respectively. The SGC is marked by a red dot on the scatter plots and by a red line on the marginal distribution curves, along with (**A**) 10K random GCs (Blue, in all plots); (**B**) 100K random GCs and ~200 1st order neighbours of the SGC (orange); (**C**) 100K random GCs and {~12K} 1st and 2nd order neighbours of the SGC (orange).

Distribution plots of error robustness scores – a focus on the marginal distributions in (**B**) and (**C**). Plots (**D**) and (**E**) correspond to the X- and Y-axis in plot (**B**), respectively. Plots (**F**) and (**G**) correspond to the X- and Y-axis in plot (**C**), respectively. The curves are normalized such that the area under each curve is 1. The mean (mu) and standard deviation (sigma) in each plot describe the distribution of the random codes exclusively (SCG neighbouring GCs were excluded). The captioned box in each plot denotes the order of magnitude of the ratio between GCs that outperformed and the GCs that under-performed, relative to the SGC.

Figure 3 - Uniform codon probabilities

Scatter plots of error robustness scores, as described in Fig **2B** – 100K random GCs and 1st order neighbours of the SGC - with (**A**) no changes made, or by using uniform probabilities as input for (**B**)

codon error occurrence; (**C**) codon usage bias; or (**D**) both codon error occurrence and codon usage bias.
Histograms of error robustness scores of GCs on the translation errors or mutations axes, separately. An unnormalized count of the data appearing as the marginal distributions in (E) the y-axis in (**A**); (**F**) the x-axis in (**A**); (**G**) the y-axis in (**B**); (**H**) the x-axis in (**C**).

Figure 4 - Uniform NT substitution matrices

Scatter plots of error robustness scores, as described in Fig **2B** – 100K random GCs and 1st order neighbours of the SGC - with (**A**) no changes made, or by using uniform nucleotide replacement probabilities as input for (**B**) the ribosome confusion matrix; (**C**) DNA-pol error propensity; or (**D**) both the ribosome confusion matrix and the DNA-pol error propensity.
Histograms of error robustness scores of GCs on the translation errors or mutations axes, separately. An unnormalized count of the data appearing as the marginal distributions in (**E**) the y-axis in (**A**); (**F**) the x-axis in (**A**); (**G**) the y-axis in (**B**); (**H**) the x-axis in (**C**).

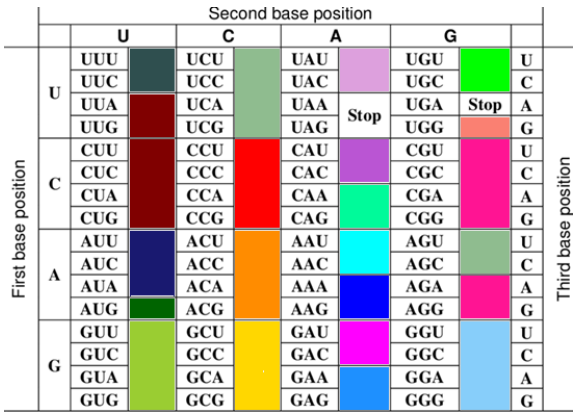
Figure 5 - Ribosome confusion matrix shuffling

A: The index annotation (0,1,2) at the bottom of each frame of the ribosome confusion matrix serve to clarify the convention used throughout this figure. **B:** Robustness to translation errors of the SGC under different permutations of the ribosome confusion matrix, as denoted by the order of frames at the bottom of each column.
C-H: Histograms of translation robustness scores under different hypothetical ribosome confusion matrices. The order of permutation of the confusion matrix is denoted in the title of the plot, corresponding to the order introduced in (**A**). The SGC is marked by the red line, random GCs are coloured blue and 1st order neighbours of the SGC are in orange. The mean (mu) and standard deviation (sigma) in each plot describe the distribution of the random codes exclusively (SCG neighbouring GCs were excluded).

Figure 1 - Table permutations

A

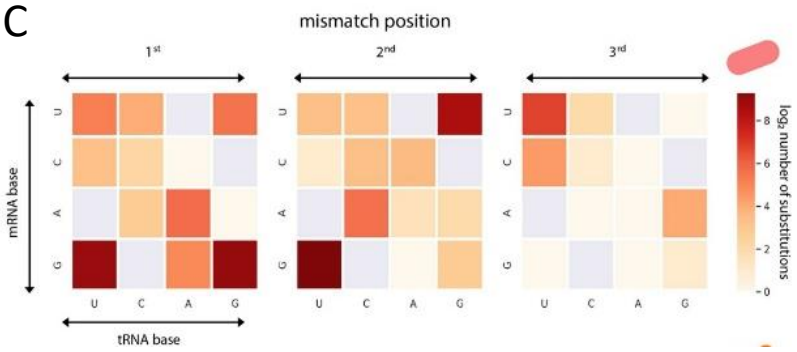
First base position	Second base position: U				Second base position: C				Second base position: A				Second base position: G				Third base position
U	UUU	F			UCU	S			UAU	Y			UGU	C			U
	UUC				UCC				UAC				UGC				C
	UUA	L			UCA				UAA	Stop			UGA	Stop			A
	UUG				UCG				UAG				UGG	W			G
C	CUU	L			CCU	P			CAU	H			CGU	R			U
	CUC				CCC				CAC				CGC				C
	CUA				CCA				CAA	Q			CGA				A
	CUG				CCG				CAG				CGG				G
A	AUU	I			ACU	T			AAU	N			AGU	S			U
	AUC				ACC				AAC				AGC				C
	AUA				ACA				AAA	K			AGA	R			A
	AUG	M			ACG				AAG				AGG				G
G	GUU	V			GCU	A			GAU	D			GGU	G			U
	GUC				GCC				GAC				GGC				C
	GUA				GCA				GAA	E			GGA				A
	GUG				GCG				GAG				GGG				G



B

From	To	Relative abundance
A	C	0.04
A	G	0.13
A	T	0.09
C	A	0.05
C	G	0.02
C	T	0.17
G	A	0.14
G	C	0.02
G	T	0.05
T	A	0.09
T	C	0.16
T	G	0.04

C



D

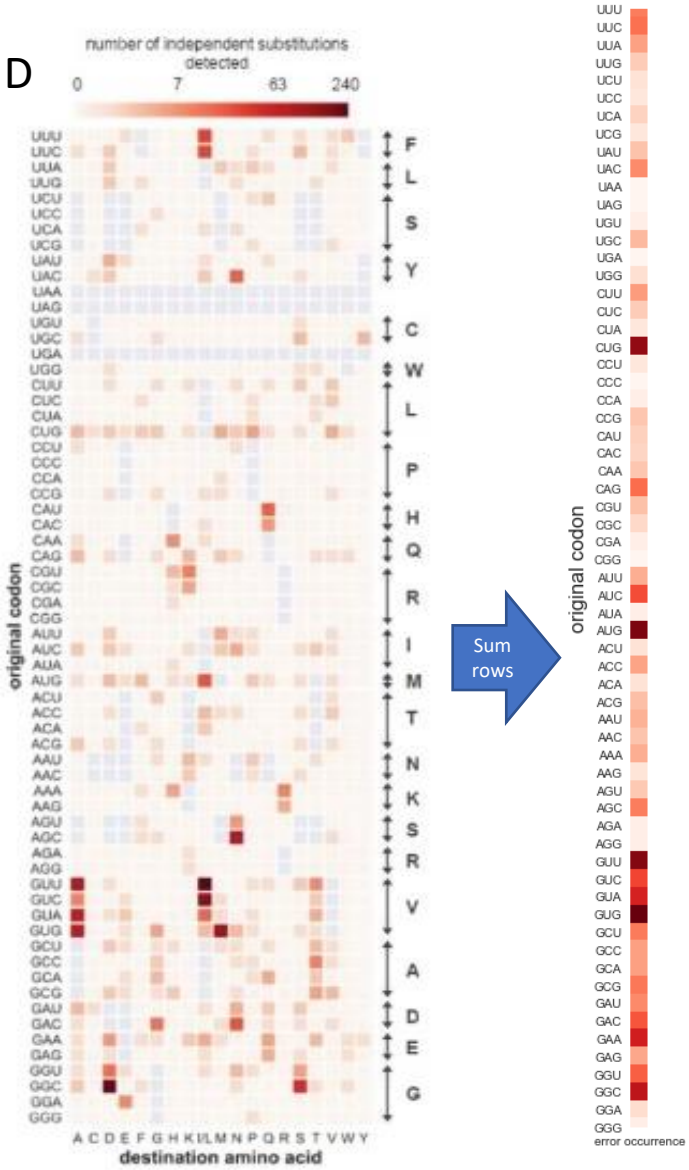
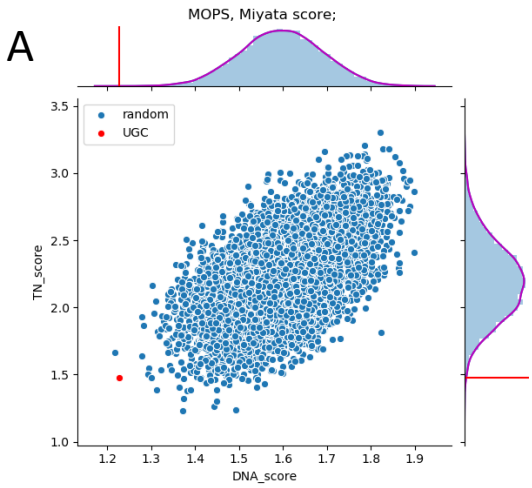
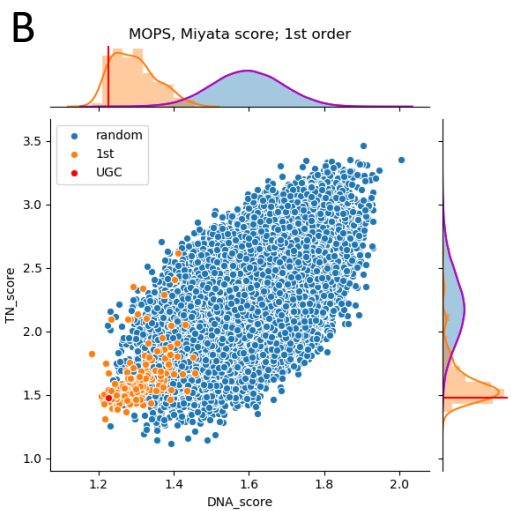
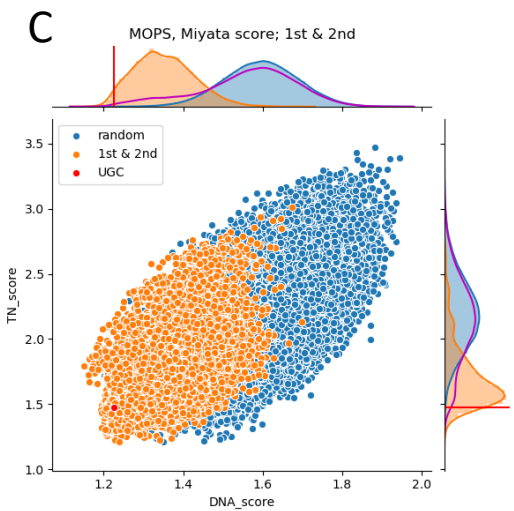
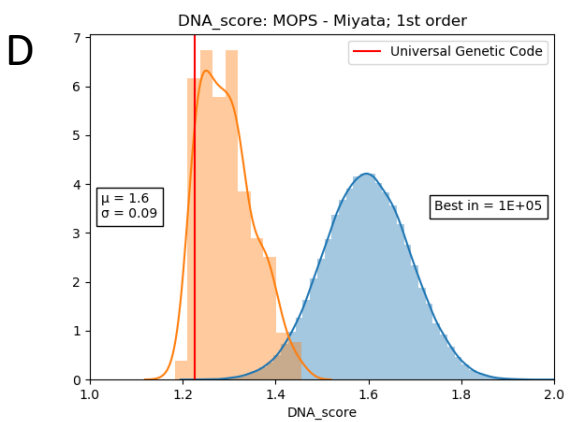


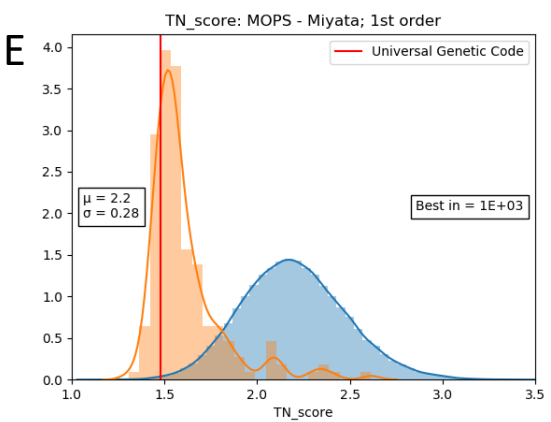
Figure 2 - Genetic code landscape exploration

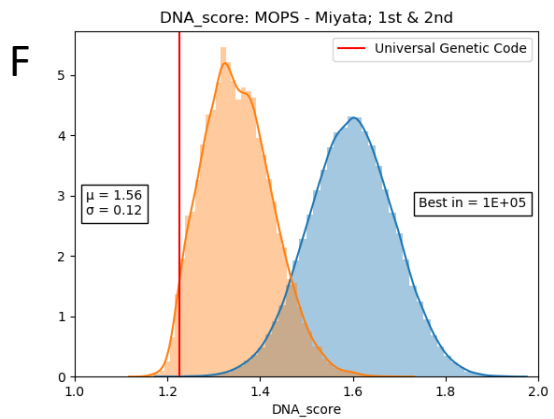












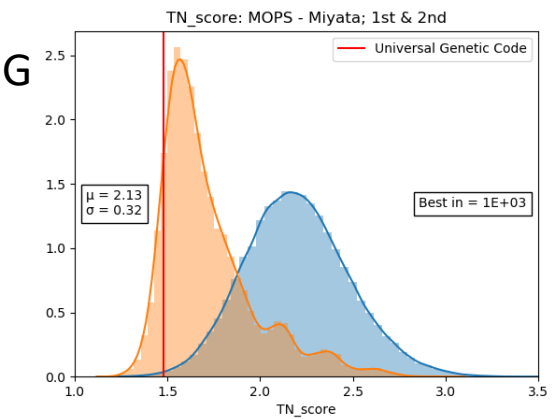
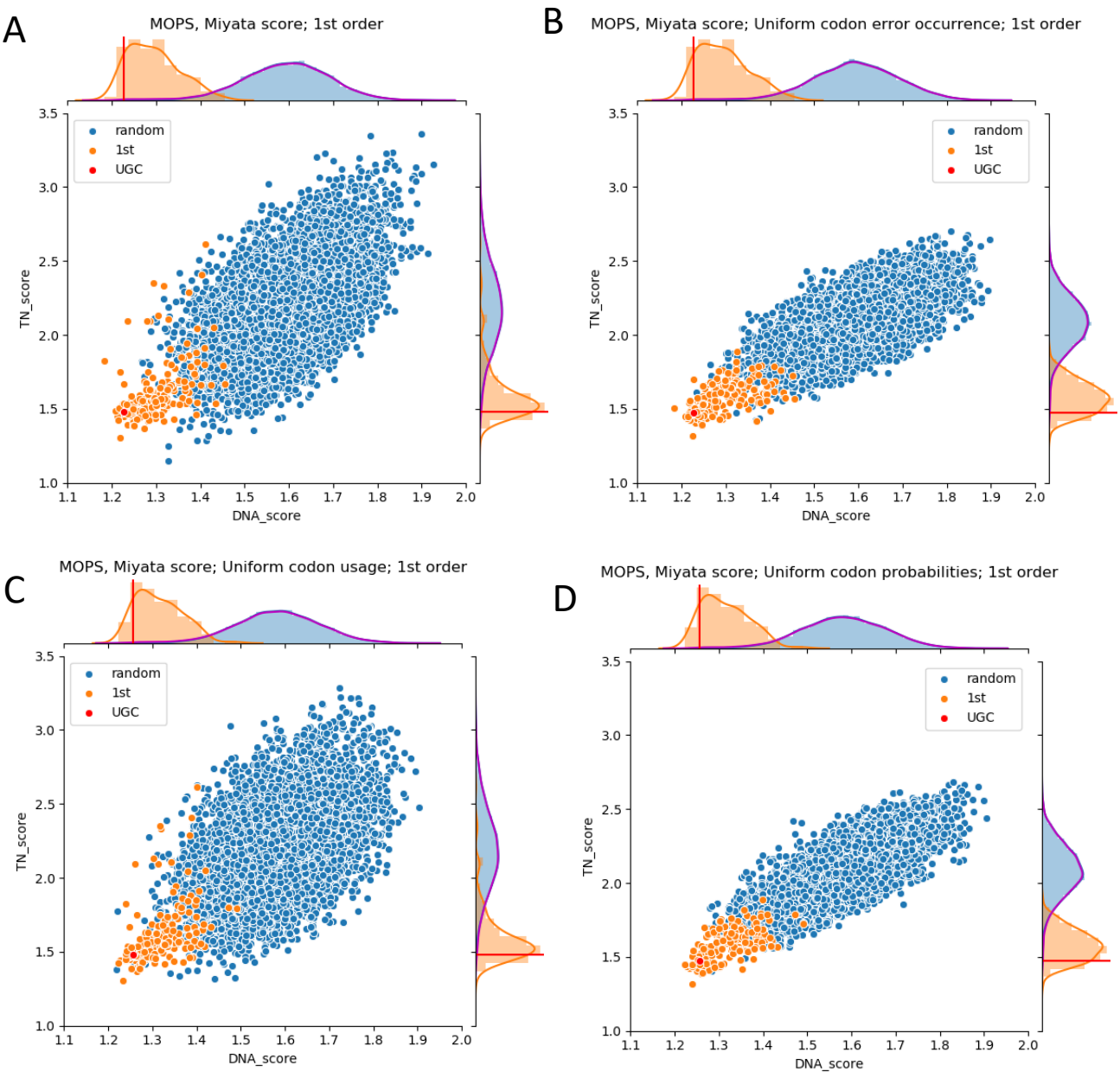


Figure 3 - Uniform codon probabilities



Uniform codon error occurrence

Uniform codon usage

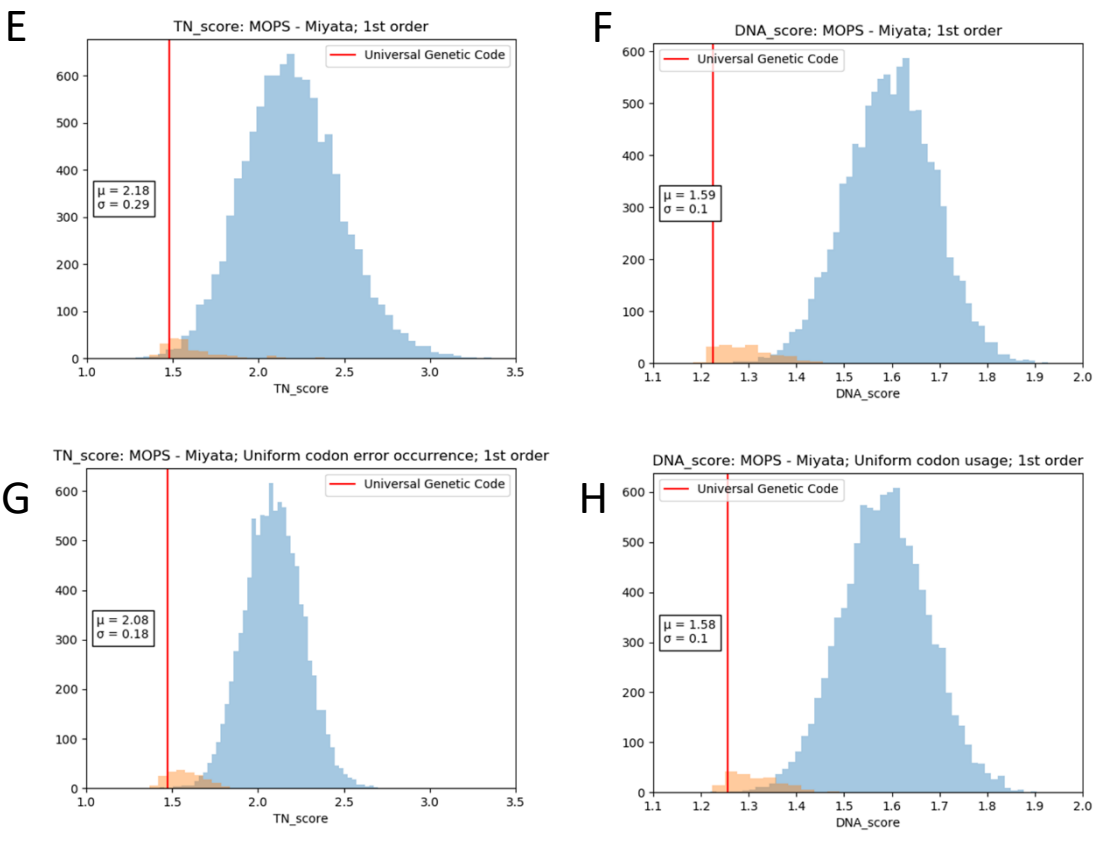
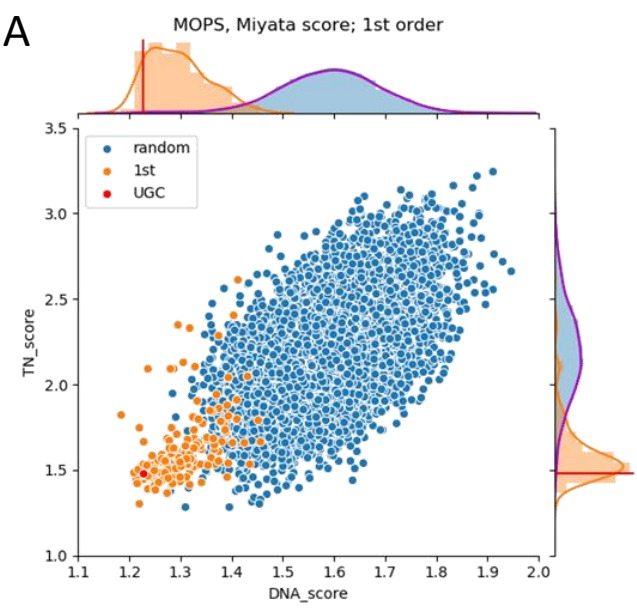
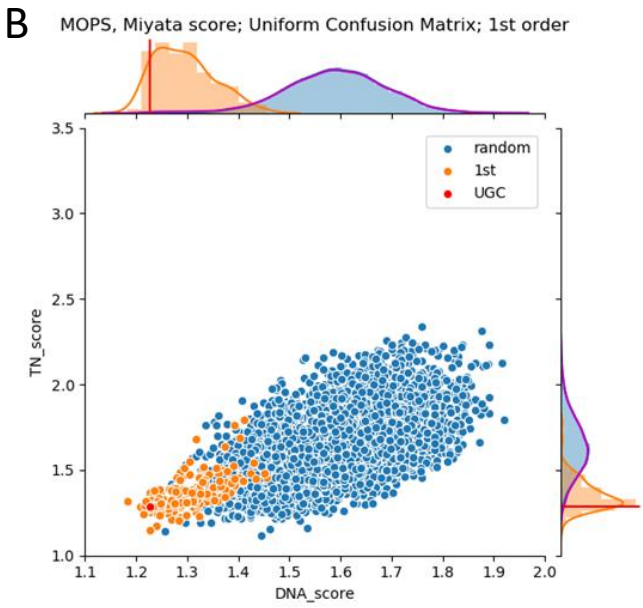
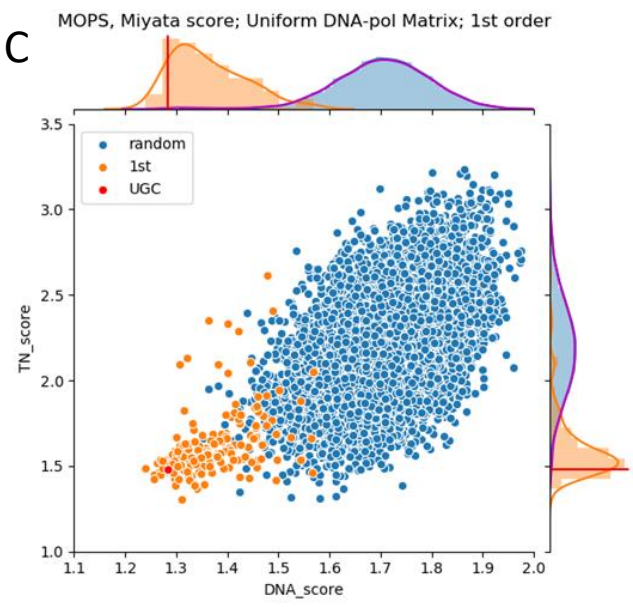
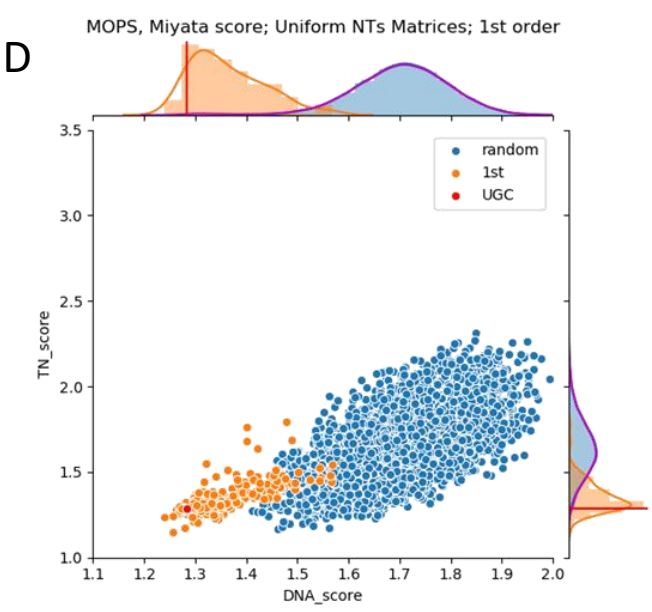


Figure 4 - Uniform NT substitution matrices



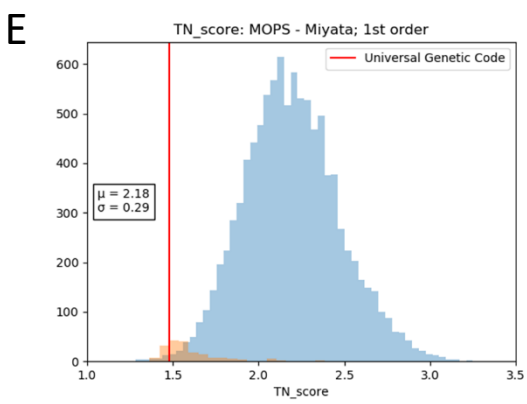


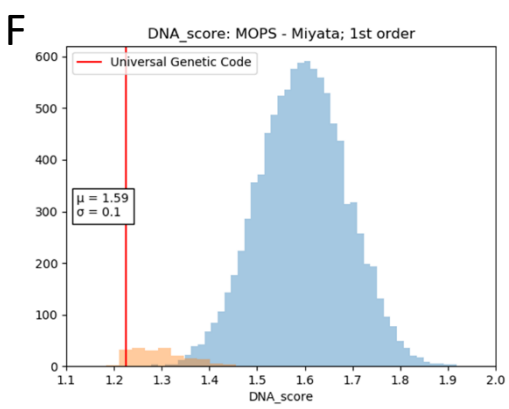


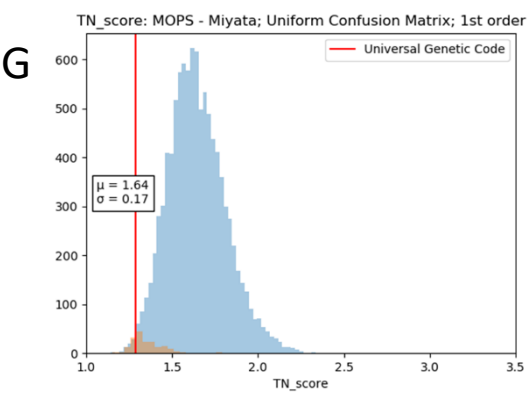


Uniform ribosome confusion matrix

Uniform DNA-pol matrix







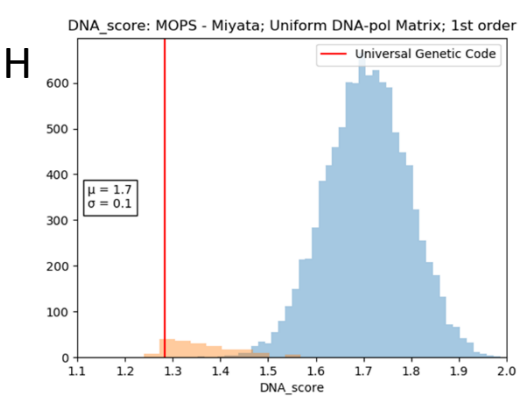
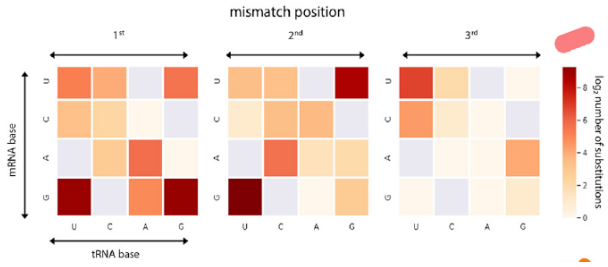


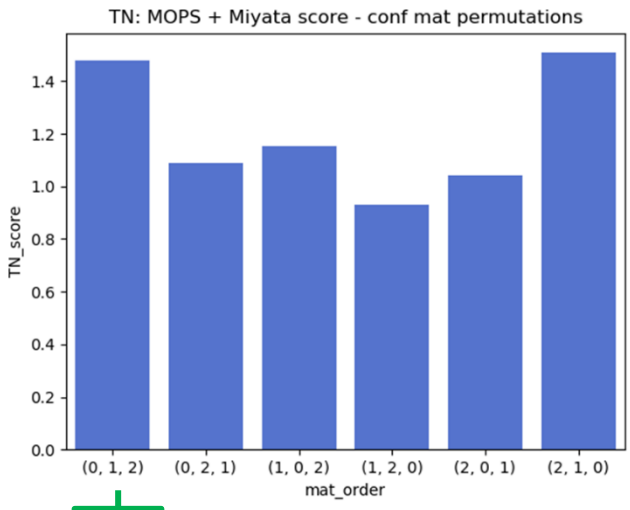
Figure 5 - Ribosome confusion matrix shuffling

A

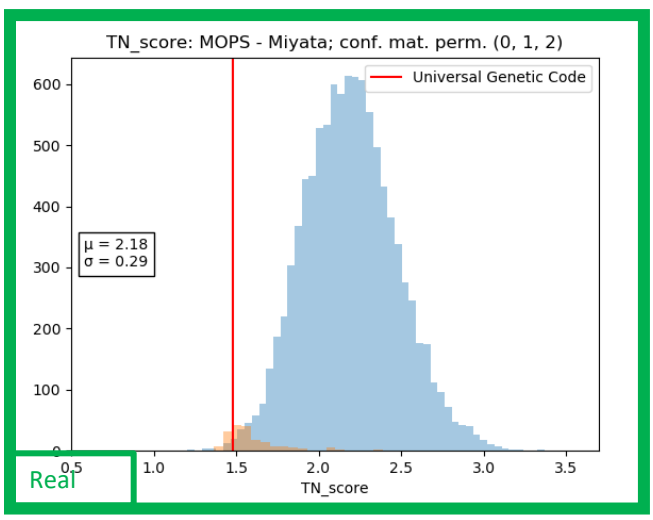


Index: 0 1 2

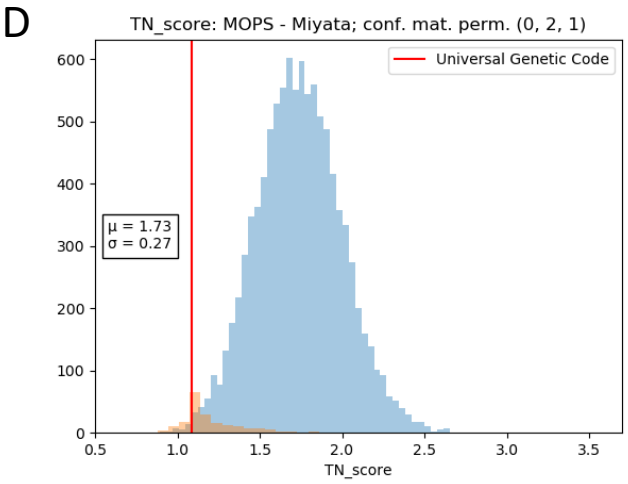
B



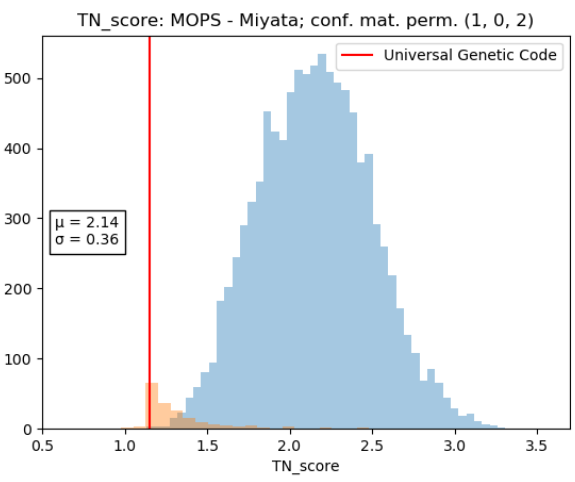
C



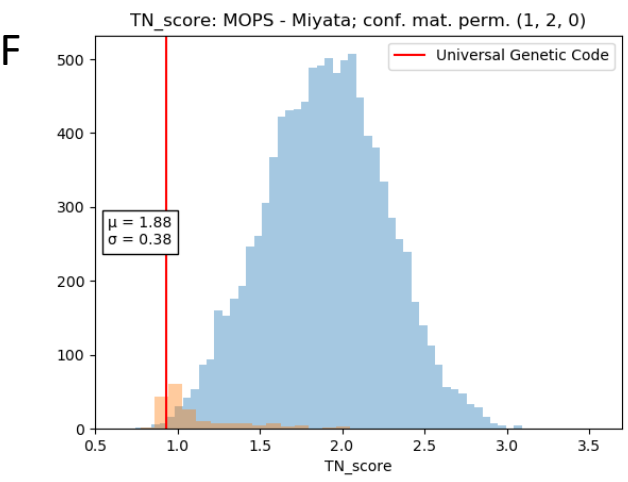
D



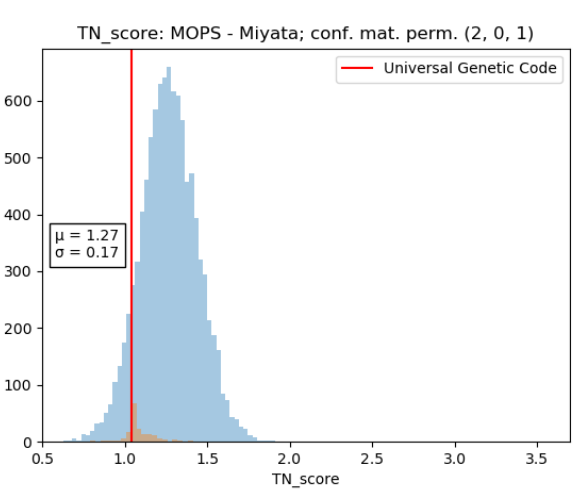
E



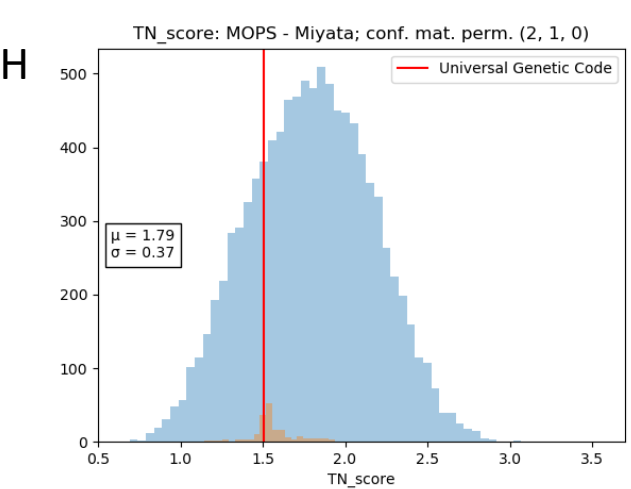
F



G



H



GWAS-like Analysis of SARS-CoV-2

Note: the data used in the following section was downloaded on 7/3/21. Figures and results refer to the data that was available at the time. Some conclusions and references benefit from the elapsed time since.

Introduction

A few months after the beginning of the COVID-19 pandemic, as clinical reports and data began to accumulate, it became possible to start to follow the inevitable evolution of the SARS-CoV-2 virus, as it made its way through the global human population. The first variant to get significant attention from the scientific community contained the D614G mutation. It emerged early in 2020 and quickly became the most prevalent virus strain (Korber et al. 2020). By March 2021, when this analysis was underway, the variant that soon became known as the "Alpha" [1] variant was nearing its peak relative frequency, in terms of global infections (https://nextstrain.org/ncov/gisaid/global,(Korber et al. 2020; Hadfield et al. 2018)). With every new strain and mutation, infecting the then unvaccinated population, there was an urgent need to quickly and accurately assess the expected clinical prognosis of patients.

From the onset of the pandemic much of the global sequencing and data collection efforts became publicly available through GISAID (https://www.gisaid.org/) - a global consortium originally established to share flu sequences amongst the research community (Pearson 2003). As the virus spread, sequencing capabilities were ramped up and sequences accumulated exponentially, becoming a go to tool for evolutionary and comparative research of the virus.

The SARS-CoV-2 viral genome is 30Kb long, it consists of ~14 ORFs encoding for ~27 proteins, including RNA-transcriptase and Exoribonuclease (Naqvi et al. 2020). The Spike protein (colloquially known as the S protein) is the protein that is used by the virus to dock to human cells and enter them (Naqvi et al. 2020). So, while the S protein is the usual suspect when it comes to clinical manifestation of the infection, an unbiased systematic approach may detect mutations throughout the entire viral genome that affect the progression of the disease in patients.

[1] For context, the Greek alphabet system for naming COVID-19 variants was introduced at the end of May 2021 ("Website," n.d.).

One commonly used method for unbiased genome wide studies are, as their name implies, Genome Wide Association Studies (GWAS). These tools were developed to find a correlation between gene loci, and/or specific SNPs in a group of individuals and the phenotypes or traits they exhibit. While most are phenotypes of the individual, such as morphology or disease, some GWAS analyses have been also used to show correlation between genetics and mental characteristics, or even socioeconomic ones (Marioni et al. 2014). This approach implies that the GWAS concept is generalizable outside its “classical” role. Following this line of thought, we show that openly available data may be of use to conduct an inter-species GWAS, with the genome taken from a pathogen (e.g., SARS-CoV-2) against the phenotype of the host. We introduce the conceptual approach, as well as some of its obstacles and ways to overcome them.

Results and Discussion

When the project was first considered, in April 2020, a few thousands viral genomic sequences have been submitted to GISAID. Relying on public data, it was necessary to filter these raw numbers to ensure quality and usefulness of data. By the time we conducted most of our analyses, in early March 2021, although the number of submitted sequences grew to ~700K, useful sequences amounted to 18,594. These consisted of 27,549 sequences of quality deemed high enough to be included in our Multiple Sequence Alignment (MSA), and 18,594 sequences with rich and useful metadata (see Methods). The intersection between these groups built the database upon which our analyses were performed.

The available metadata, to a great part consisting of free text annotation and clinical diagnoses, was curated into several categorical or numerical parameters: severity (multi-category), condition (severity condensed into two categories), symptoms, hospitalization, alive, gender, age, time (date sample was collected), continent and country (see Methods). With samples coming from a large assortment of sources and no unifying clinical procedure for data collection, metadata was only partially available for each sample.

Most samples have no data regarding each category (Fig 1A). It is also important to note that the distribution of some observed parameters is not the one expected from what is known about the epidemiology of the pandemic. In the symptoms distribution (Fig 1A, top left plot), a small fraction of our samples was marked as asymptomatic (~7%) whereas several studies have found the numbers

to be much higher – 20%-75% (18% on the Diamond Princess cruise ship, for example (Mizumoto et al. 2020)). Similarly, most samples for which we have information regarding severity came from critically ill patients, while most infected with COVID-19 only experience mild symptoms.

As a first inspection of the sequences, we calculated the entropy at each position of the viral genome (Fig 1B). This analysis revealed to us that the observed mutations are not restricted to specific parts of the genome, and changes relative to the Wuhan WT strain have been observed in all viral genes and non-coding regions.

We then proceeded to analyze positions on the genome and individually determine their association with one or more of the metadata parameters we had. For each position, we grouped the samples by the nucleotide identity in their genome at that position. We then compared these groups by a statistical test to find a significant correlation between nucleotide identity and the parameter at hand. We tested categorical data with the chi-squared test and numeric data with ANOVA. To account for multiple testing, we corrected the calculated p-values using the Benjamini-Hochberg procedure.

Our analysis of all sequences from around the world, showing the significance of correlation between genomic positions and the severity of the disease recorded in the patients, is given by a Manhattan plot in Fig 2A. As was our observation in the global entropy plot, the 252 positions significantly associated with severity we detected were not limited to a specific gene, but rather scattered throughout the genome. We sorted the corrected p-values (from here on referred to as q-values) from smallest to largest (Fig 2B). The left plot highlights the stark difference between the significant positions, at the left bottom corner of the plot, and the rest of the positions, most of which had a q-value nearing 1.

For positions we found to be significantly associated with some parameter, we evaluated the direction of their effect – whether patients infected by one variant exhibited worse of better symptoms, relative to patients infected by the WT strain. This analysis was only carried out for binary parameters. The result of such an analysis is depicted in Fig 2C, highlighting the positions that were significantly associated with the appearance of symptoms, or lack thereof, in the infected patients. These 140 significant positions were also assigned a marker to indicate the direction of the detected correlation.

The procedure we carried out in these analyses is further illustrated by data in the tables in Fig 2D. The tables contain hypothetical counts of sequences for a single genomic position of interest. Rows indicate the examined parameter, with one considered to be a “good” condition (e.g., asymptomatic / mild condition / patient alive) and one a “bad” condition (e.g., symptomatic / severe condition / patient deceased). Columns indicate if the nucleotide in the corresponding box appeared in the WT strain, or in the mutated one. The left-most table examines a theoretical genomic position in which 1 viral WT sequence came from a patient with a “good condition” and 100 viral WT sequences came from patients that exhibited a “bad” condition (‘ref’ column). In the ‘mut’ column, on the other hand, 1 viral sequence with a mutation in the position of interest came from a patient with a “good” condition and 20 viral sequences with the same mutation came from patients who exhibited a “bad” condition. This is a situation in which both the WT and the mutated nucleotide at the position of interest cause, more often than not, a “bad” condition. However, in the mutant viruses, the “bad” condition appears less often, relative to the WT. Hence, in this position the mutation has caused a slightly better condition of the patients who contracted the virus harboring that mutation. This is indicated by an upwards pointing three-point star marker. Conversely, the middle table shows a situation in which a larger fraction of patients who contracted the mutant strain displayed a slightly worse condition than the patients who were infected by the WT. Positions that display this type of data distribution were marked by a down pointing three-point star marker. Lastly, the right-most table indicates a situation in which the WT nucleotide at the position of interest was found in a larger fraction of patients with the “bad” condition, but the mutant nucleotide at the same position was found in a larger fraction of patients with the “good” condition. Hence, it is marked by an upwards pointing triangle, indicating an improved prognosis associated with the mutant virus, relative to the WT strain.

Among the parameters for which we repeated the analysis, was also the gender of the patients (Fig 3A). The 96 positions in this plot, that were found to be significantly associated with gender, seem to suggest a correlation between the genomic content of the virus and the gender of the infected individual. Without a known biological mechanism that would explain this result, we suspected this, and perhaps previous results, are due to spurious correlations. As a negative control, we randomly shuffled the values assigned to each sequence in the following parameters: gender, outcome (dead/alive), age, condition, symptoms and continent. When we re-ran our analysis on these shuffled parameters, we found 0 positions significantly associated with any of the parameters (data not

shown), as expected. We next turned to examine the overlap between significant positions in different parameters (Fig 3B). E.g., in the top table we see that 140 positions were found to be significantly associated with Symptoms in total. Out of those 140, 53 positions were also found to be significantly correlated with the Severity of the disease in patients. Only 18 out of the 140 were uniquely associated with Symptoms, and no other parameter. The bottom table lets us understand that 37% of the significant positions with regard to Symptoms were shared with Severity, while 12% of the positions associated with Symptoms were associated only with Symptoms and no other parameter. This approach has proved to be effective, as it clearly shows a considerable overlap between the different parameters. Specifically, 100% of the positions associated with gender were also associated with the continent the sequences originated from (bottom table), leaving 0 positions uniquely associated with gender (top table). This suggested a bias in our data, specifically with regard to gender, arising from the geographical distribution of our samples. And indeed, when comparing the gender distribution between different continents (Fig 3C) we can clearly see a gender gap – e.g., while samples from Africa came mostly from female patients, samples from Asia came mostly from male patients. As data from Europe was the most abundant, and lacked the above-mentioned bias, we decided to perform the rest of the analysis on European sequences exclusively.

For the 7,717 European sequences, we repeated the analyses for our parameters of interest, their overlap and the negative controls. From the overlap tables of the European sequences (Fig 4A) it becomes apparent that our approach was successful in eliminating the bias in the data that resulted in the genetic association with the gender of the patients. From the previously 96 positions significantly associated with gender (Fig 3A,3B), we remained with a single one (Fig 4A, 4B). Conversely, in other parameters, such as the condition of the patients (Fig 4C) or the outcome of their disease (Fig 4D), the analysis still shows a large number of positions significantly associated with each one – 39 and 78 positions, respectively. Nonetheless, the right table in Fig 4A underscores some additional concerns about the data. Most parameters have either very little, or no positions at all, that are unique only for them, suggesting that the parameters are still strongly intertwined. A new possible contributor of bias that emerges from these tables is the date at which the samples were taken. Most parameters have either a large or a complete overlap between the positions associated with them and the positions associated with time. Following the success of this approach, we could theoretically continue to disentangle the parameters, by analysing smaller and cleaner subsets of

the data, each time factoring out different parameters. However, given the scarcity of data at the time the research was conducted, this strategy would quickly leave us without enough samples to make statistically significant and/or generalizable observations.

Bearing the above-mentioned caveats in mind, we were curious as to which of the genomic positions would manifest at the proteomic level. Here we assume that genomic mutations are more likely to have a biologically functional and clinical effect if they are translated into a different amino-acid (as opposed to silent mutations). We performed the same analyses as above, but rather than focusing on nucleotides, we looked into the association of amino-acid mutation with our parameters of interest. Building on the previous conclusion, this analysis was limited to the European samples within our data. Furthermore, since the S protein is known as a very significant protein in the interaction of the virus with human tissues (Huang et al. 2020), we wanted to focus on it with the limited deduction power that we had. From the global overview on these analyses (Fig 5A), we could see that, unsurprisingly, the phenomena we previously observed also appear at the proteomic level. Namely, the intertwining of significant positions between parameters, and the seemingly high overall overlap of all parameters with the date at which the samples were taken.

Our analysis of amino-acid positions associated with symptoms (Fig 5B) strongly points at mutation D614G as significantly associated with a slightly higher chance of showing symptoms, relative to the WT variant. While drawing immediate attention from the research community, the contribution of this mutation to the presentation of symptoms (or lack thereof) remained undetermined, with mixed indications at different time points (Huang et al. 2020; Leung et al. 2021). Of note, the same mutation appeared in our analysis as significantly associated (albeit to a lesser degree) with the condition of the patient (Fig 5C). This stands in contrast with reports that no such association was detected in clinical studies (Huang et al. 2020; Leung et al. 2021; Dao et al. 2021). Other positions, more strongly associated with the condition of the patients are del69-70 – both associated with a milder prognosis in patients harboring the mutant variant. These two deletions were also significantly associated with a lower chance of death (Fig 5D). This again contradicts reports of increased hospitalization, ICU admission, and mortality increase in the Alpha variant, which carries these deletions (Lin et al. 2021). This contradiction is repeated in two additional mutations - N501Y and P681H, also found in the Alpha variant – that are the most strongly associated with the outcome of the disease

(death/survival). Our analysis found both to be associated with a slightly lower chance of mortality, while clinical data pointed to the contrary (Lin et al. 2021). All the positions mentioned in this paragraph were also found by us to be significantly associated with the date of sample collection (data not shown).

Taken together, our results portray a method for the detection of mutations' association with clinical properties of the disease caused by the studied pathogen. Unfortunately, due to scarcity in data at the time we did our analysis, we were unable to eliminate some of the biases we detected. Repeating this analysis today, with millions of genomes publicly available, might yield different and more actionable results altogether.

Methods

Sequences acquisition procedure

Data was acquired on 2021-03-07 from gisaid.com, using the web app search engine with the following filters:

- "high coverage" checkbox - checked. [Definition: Only entries with <1% Ns and <0.05% unique amino acid mutations (not seen in other sequences in database) and no insertion/deletion unless verified by submitter]
- "w/Patient status" checkbox - checked. [Definition: Only entries with Patient status available]
- Host: Human

Download is limited to 10K sequences at a time. Hence, data was split and downloaded in batches.

MSA

Downloaded sequences were scanned for mutations relative to the SARS-CoV-2 reference genome using Nextclade alignment tool (https://clades.nextstrain.org, ("Website," n.d.)). The detected nucleotide changes were used to construct a Multiple Sequence Alignment (MSA) from all sequences, focused on substitutions and deletions. Insertions were discarded at this stage of the analysis.

Specifically, downloaded batches (see above) were further split to chunks containing ~1500 sequences each. Each chunk was processed using the "neherlab/nextclade nextclade.js" docker

image with '--jobs=10' option and fasta file as input and json file as output. Processed json files of chunks were combined to create the full MSA using an in-house written Python script.

Metadata processing

Metadata corresponding to the downloaded sequences was filtered and curated using an in-house written Python script. In brief, the process included:

- Verification that the data is exclusively from Human hosts.
- Exclusion of erroneously filled fields.
- Curation of free text fields into binary / limited categories fields.
- Standardization of all categories' entries.
- Exclusion of metadata poor sequences, defined as sequences that had no age data, no gender data, and only indication regarding the appearance of symptoms.
- Oceania (17 samples) dropped.

In total, processing of metadata reduced the number of sequences considered for the analysis from an initial 27,550 to the 18,594 mentioned in the main text.

The categories and their values after curation: severity [mild, moderate, severe, critical, asymptomatic], symptoms [yes, no], hospitalization [yes, no], alive [yes, no], gender [male, female], age, time (date sample was collected), continent, country. Severity was further condensed into a 'condition' category with two bins: [mild, severe]. The first, consisting of sequences with the 'mild' or 'asymptomatic' severity, and the latter consisting of the rest of the severity values.

Consensus filtering

16,647 positions in the NT analysis showed 100% conservation and were dropped from further calculations.

585 positions in the AA analysis showed 100% conservation and were dropped from further calculations.

Figure legends

Figure 1 – Data overview

A: a quantitative description of available metadata. Plots show the data distribution of various categories used throughout the analysis. Note that for many parameters the most abundant value is N/A = Not Available.

B: Entropy per nucleotide position of the SARS-CoV-2 genome. X axis indicates the position of the nucleotide along the viral genome. Viral genes are coloured as indicated in the legend to the right. Grey dots represent NTs that fall outside any gene, and are marked “non-coding”.

Figure 2 – Results of global analysis

A: Manhattan plot of viral genomic positions’ significance with regard to their effect on Severity in all samples. Y axis values are in negative log(q-value). Positions are colored by their gene association, as indicated by the X axis labels. Color scheme follows the palette introduced in figure **1B**.

B: Significance distribution of nucleotide positions in the viral genome, with regard to their effect on disease Severity in all samples.

Left – all q-values (p-values corrected for multiple testing) sorted from smallest to largest.

Right – only significant q-values, with a logarithmic y axis, similarly sorted from smallest to largest.

C: Manhattan plot of viral genomic positions’ significance with regard to their effect on Symptoms [in global samples]. Non-significant positions are indicated by a dot marker. Significant positions were assigned a marker (see legend) indicating the direction the mutated nucleotide had on infected patients’ prognosis, relative to WT (see **D**).

D: Mock data to illustrate the assignment of marker signs in figure **C**, and subsequent figures throughout this section (See main text).

Figure 3 – Global analysis; summary and issues

A: Manhattan plot of viral genomic positions’ significance with regard to their association with Gender in all samples.

B: Overlap between positions detected as significant with regard to various parameters in all samples.

Top - The bottom row (marked blue) gives the total number of positions found to be significant with regard to the parameters in the columns. Numbers on the main diagonal (marked red) are counts of

positions that were uniquely significant with regard to the parameter in the column – i.e., not shared with any other parameter.

Bottom – overlap table (above) normalized to the total number of significant positions associated with the parameter in the column.

Note that the percentages do not sum to 100, as significant positions can be shared between multiple parameters.

C: Number of sequences that came from patients by gender in different continents

Figure 4 – Europe analysis; summary and benefits

A: Overlap between positions detected as significant with regard to various parameters in European samples. See Fig **3B**.

B: Manhattan plot of viral genomic positions' significance with regard to their association with Gender in European samples.

C: Manhattan plot of viral genomic positions' significance with regard to their effect on Condition in European samples.

D: Manhattan plot of viral genomic positions' significance with regard to their effect on Outcome in European samples.

Figure 5 – Protein level analysis summary and key results

Note: the analyses depicted in this figure focus on amino-acid positions, as opposed to nucleotide positions in previous figures.

A: Overlap between proteomic positions detected as significant with regard to various parameters in European samples. See Fig **3B**.

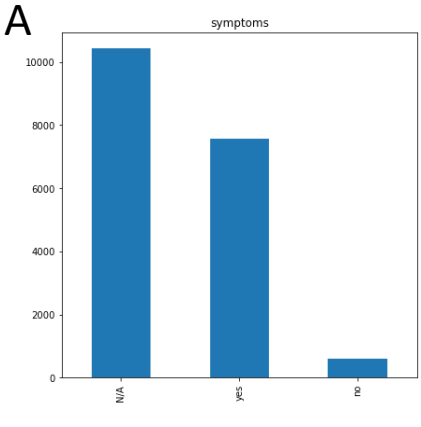
B: Manhattan plot of viral proteomic positions' significance with regard to their effect on Symptoms in European samples.

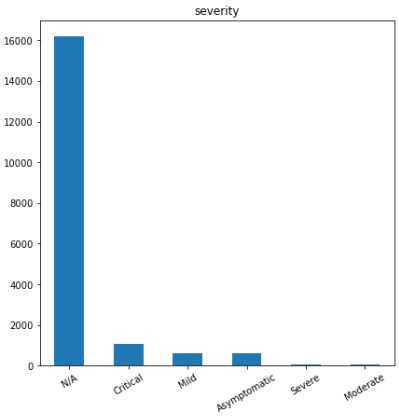
C: Manhattan plot of viral proteomic positions' significance with regard to their effect on Condition in European samples.

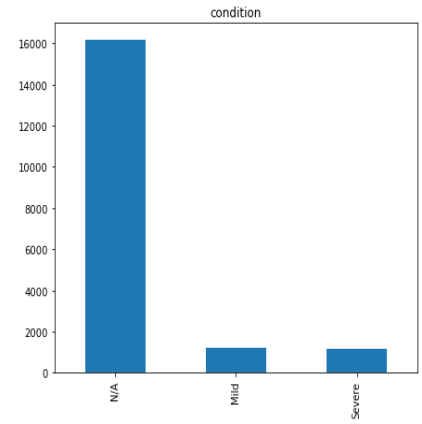
D: Manhattan plot of viral proteomic positions' significance with regard to their effect on Outcome in European samples.

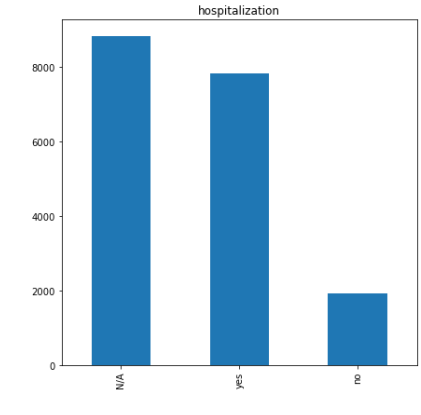
Figure 1 – Data overview

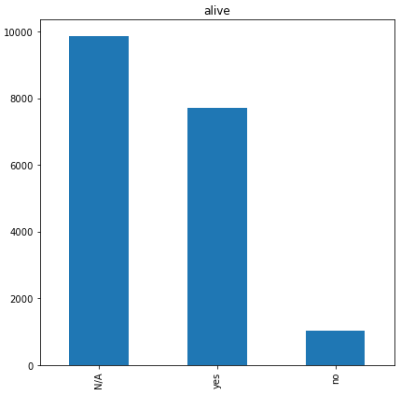
A

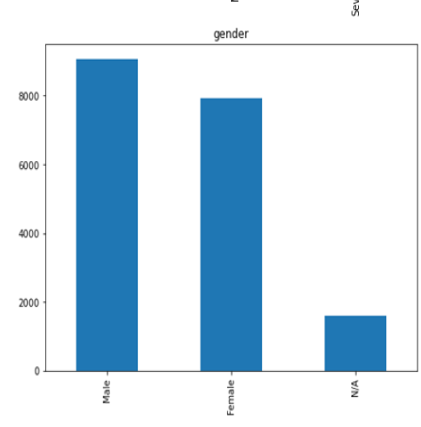


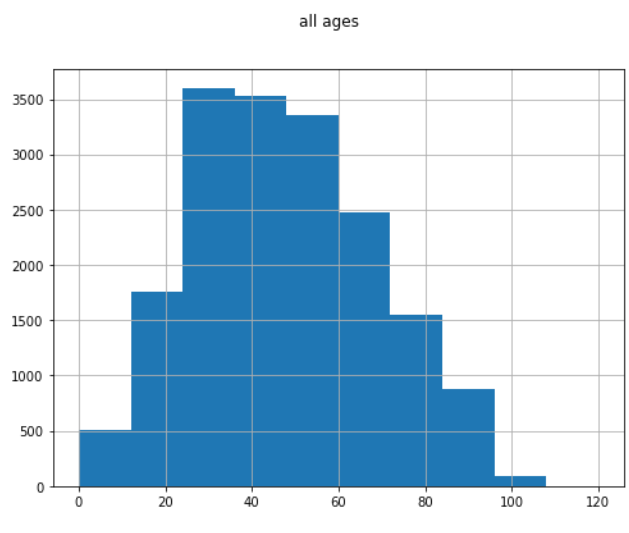


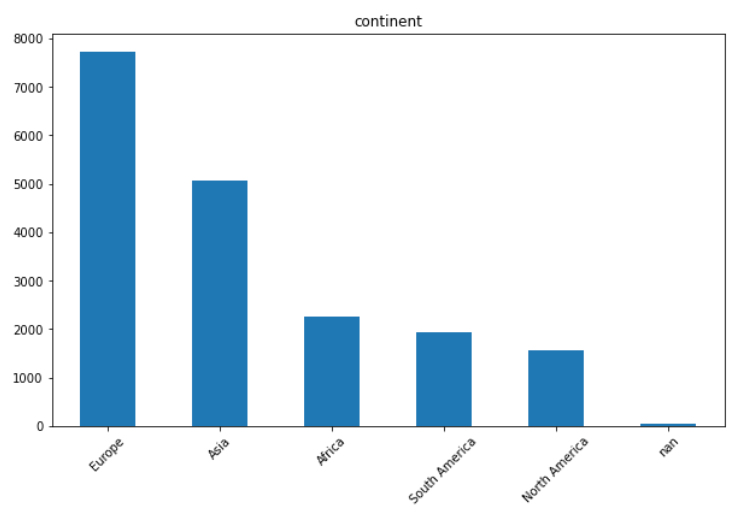


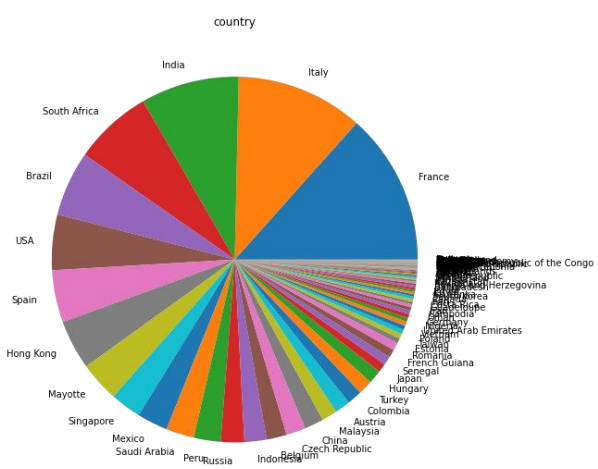


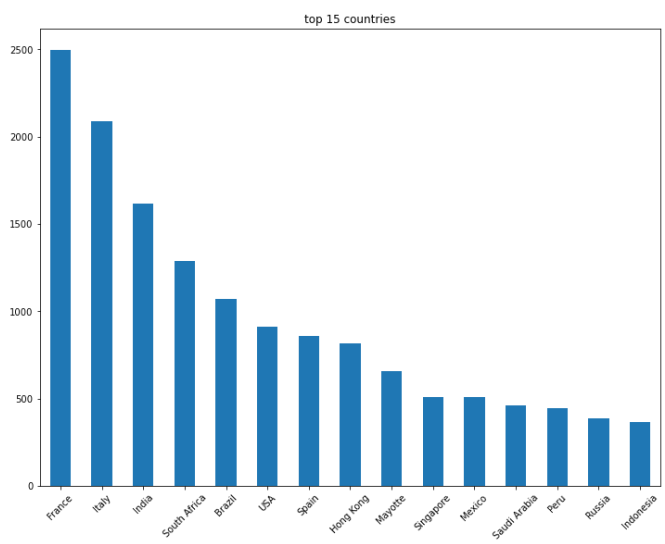












B

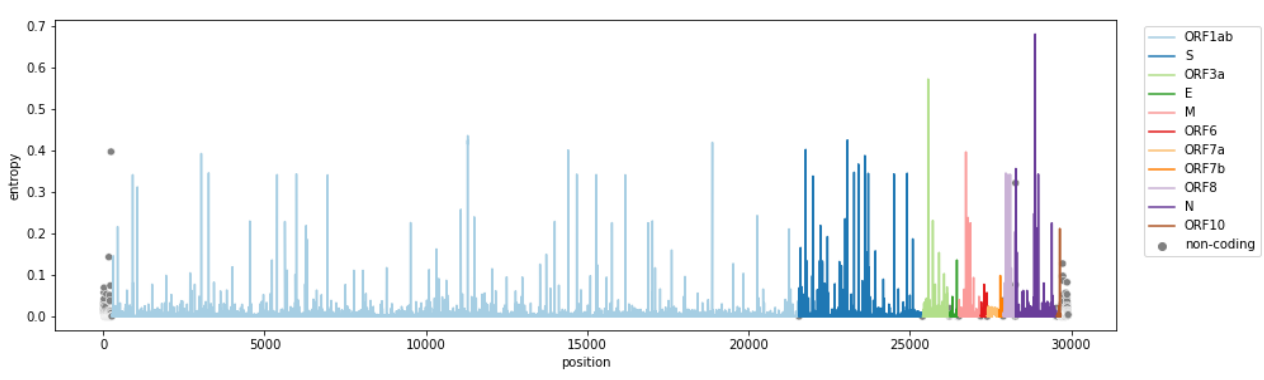
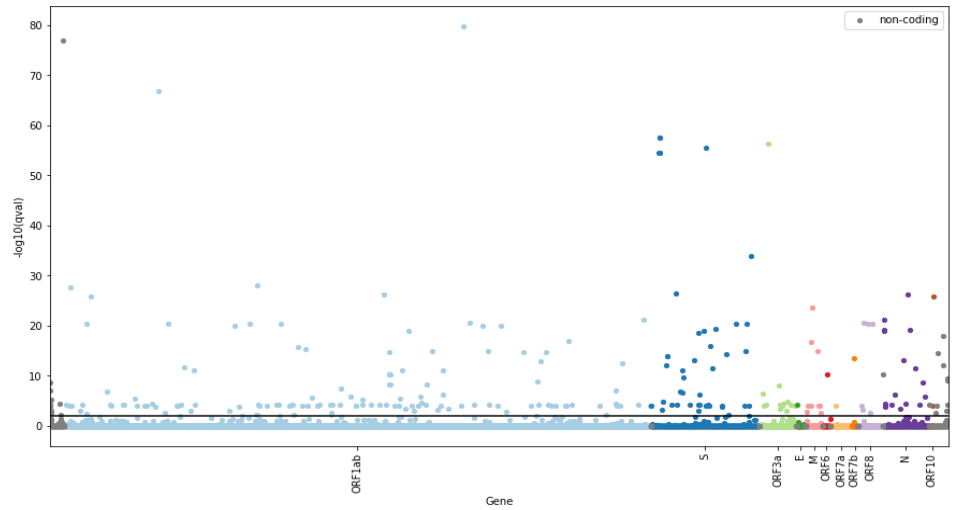
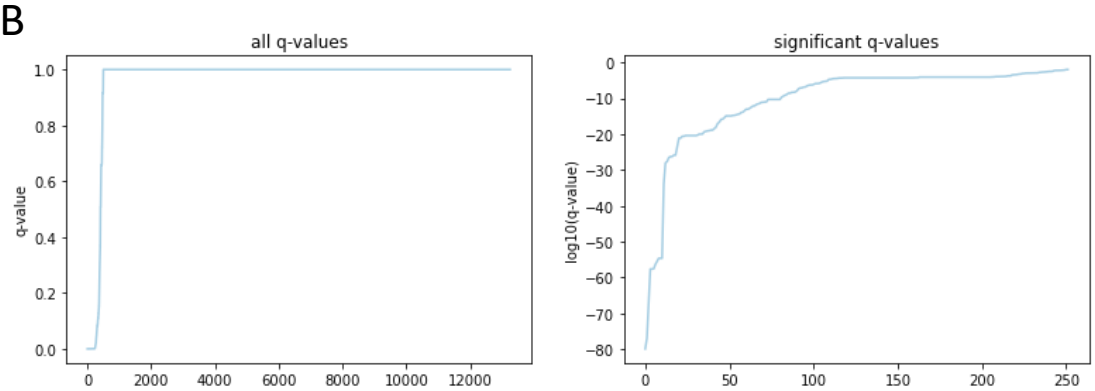


Figure 2 – Results of global analysis

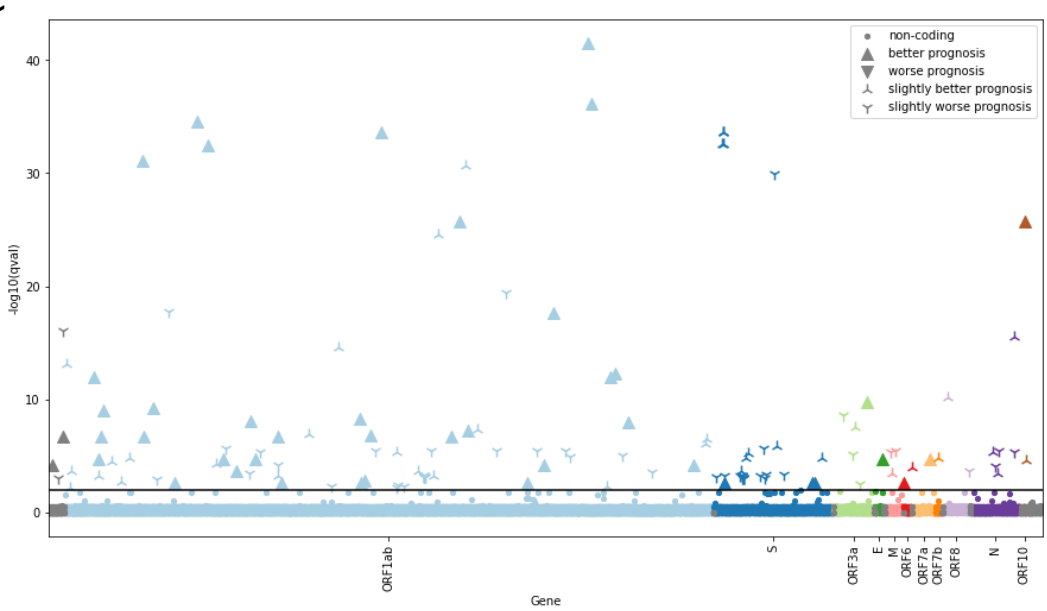
A



B



C



D

bad -> less bad

	ref	mut
“good”	1	1
“bad”	100	20

bad -> worse

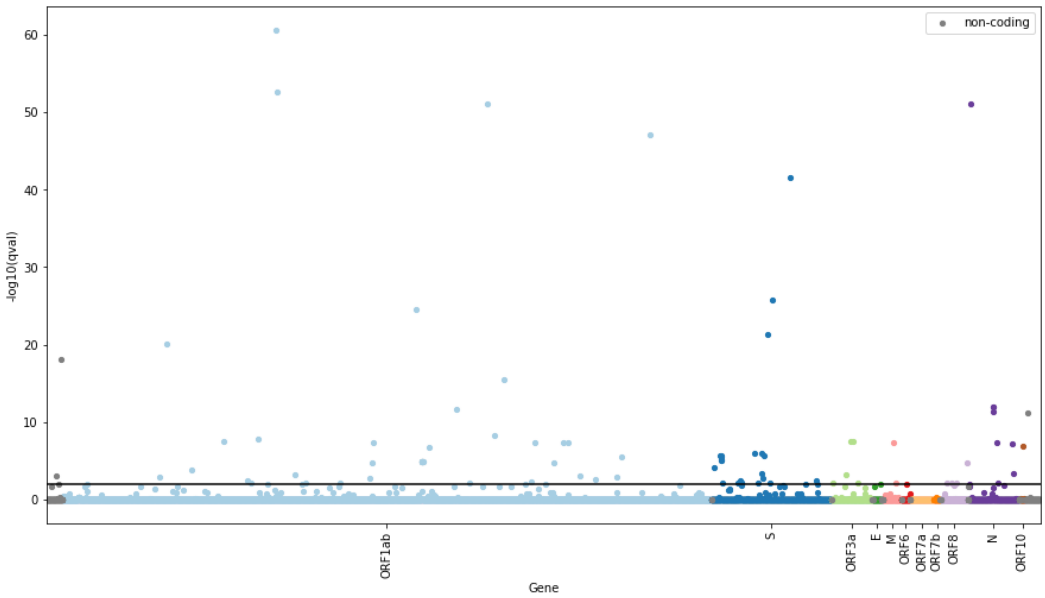
	ref	mut
“good”	20	1
“bad”	100	20

bad -> good

	ref	mut
“good”	1	20
“bad”	100	1

Figure 3 – Global analysis; summary and issues

A



B

	age	alive	condition	continent	gender	severity	symptoms	time
age	21	76	68	206	73	71	64	195
alive	76	14	70	159	44	72	33	114
condition	68	70	0	110	40	109	43	103
continent	206	159	110	1058	96	153	120	831
gender	73	44	40	96	0	43	53	83
severity	71	72	109	153	43	93	53	127
symptoms	64	33	43	120	53	53	18	88
time	195	114	103	831	83	127	88	218
total	246	175	111	2008	96	252	140	1074

	age	alive	condition	continent	gender	severity	symptoms	time
age	8%	43%	61%	10%	76%	28%	45%	18%
alive	30%	8%	63%	7%	45%	28%	23%	10%
condition	27%	40%	0%	5%	41%	43%	30%	9%
continent	83%	90%	99%	52%	100%	60%	85%	77%
gender	29%	25%	36%	4%	0%	17%	37%	7%
severity	28%	41%	98%	7%	44%	36%	37%	11%
symptoms	26%	18%	38%	5%	55%	21%	12%	8%
time	79%	65%	92%	41%	86%	50%	62%	20%

C

		count
continent	**gender**	
Africa	**Female**	1278
	Male	843
Asia	**Female**	1756
	Male	3209
Europe	**Female**	3496
	Male	3384
North America	**Female**	661
	Male	860
South America	**Female**	724
	Male	725

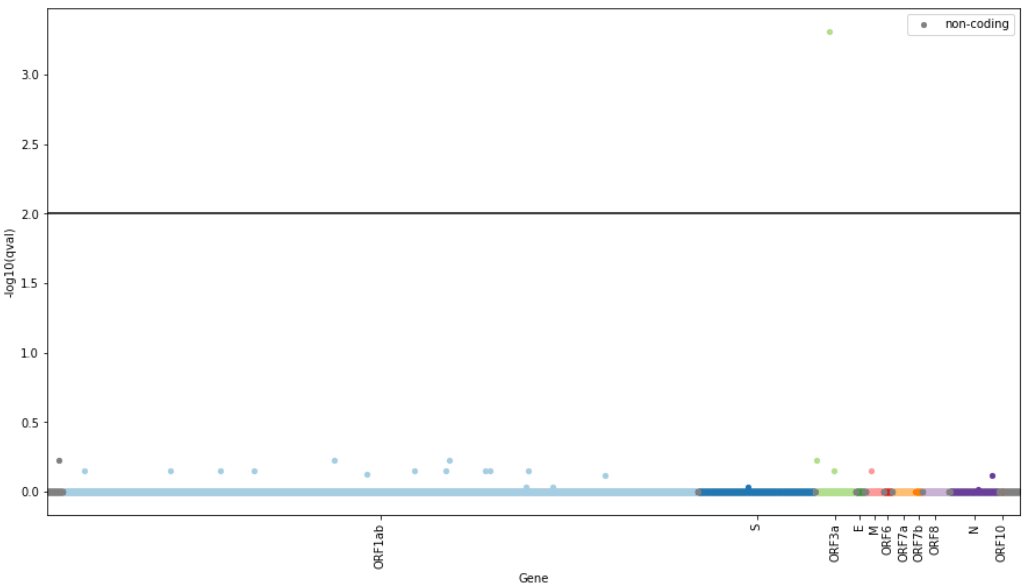
Figure 4 – Europe analysis; summary and benefits

A

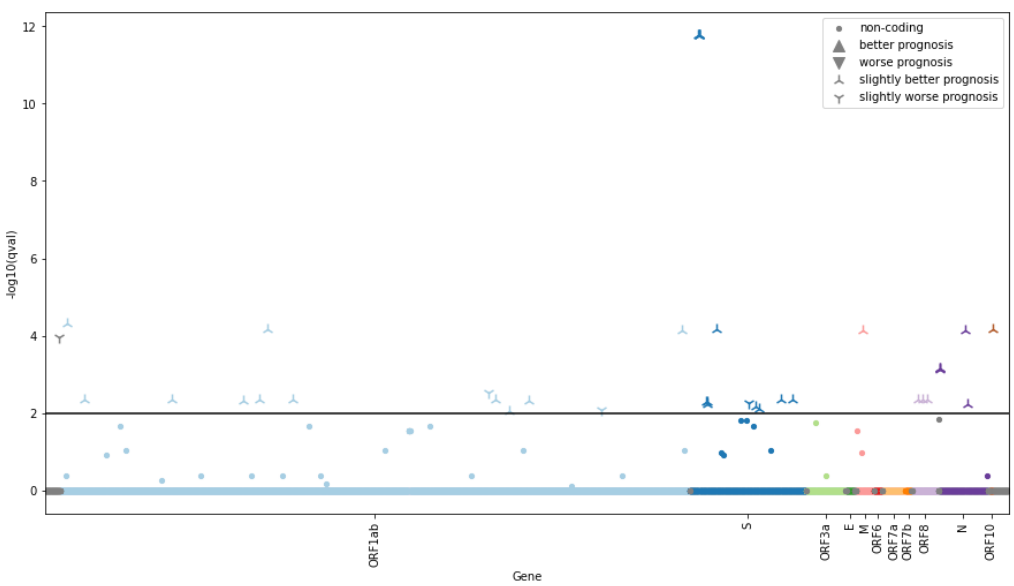
	age	alive	condition	gender	severity	symptoms	time
age	27	56	28	0	35	8	101
alive	56	2	23	0	32	2	67
condition	28	23	0	0	30	12	39
gender	0	0	0	0	0	0	1
severity	35	32	30	0	5	33	62
symptoms	8	2	12	0	33	11	34
time	101	67	39	1	62	34	249
total	133	78	39	1	78	52	399

	age	alive	condition	gender	severity	symptoms	time
age	20%	71%	71%	0%	44%	15%	25%
alive	42%	2%	58%	0%	41%	3%	16%
condition	21%	29%	0%	0%	38%	23%	9%
gender	0%	0%	0%	0%	0%	0%	0%
severity	26%	41%	76%	0%	6%	63%	15%
symptoms	6%	2%	30%	0%	42%	21%	8%
time	75%	85%	100%	100%	79%	65%	62%

B

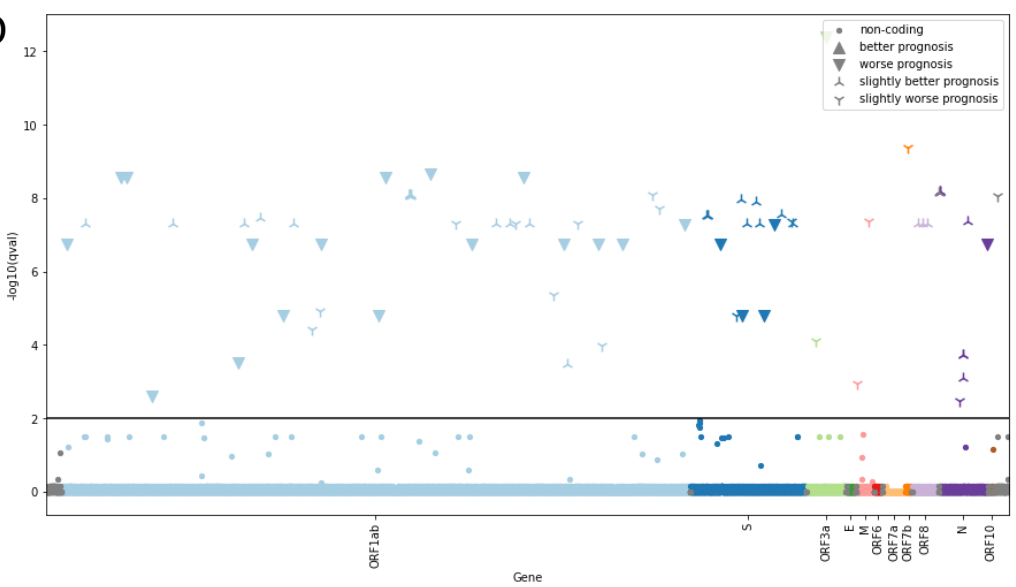


C




```
Mild      923
Severe    281
```

D




```
yes     1889
no       150
```

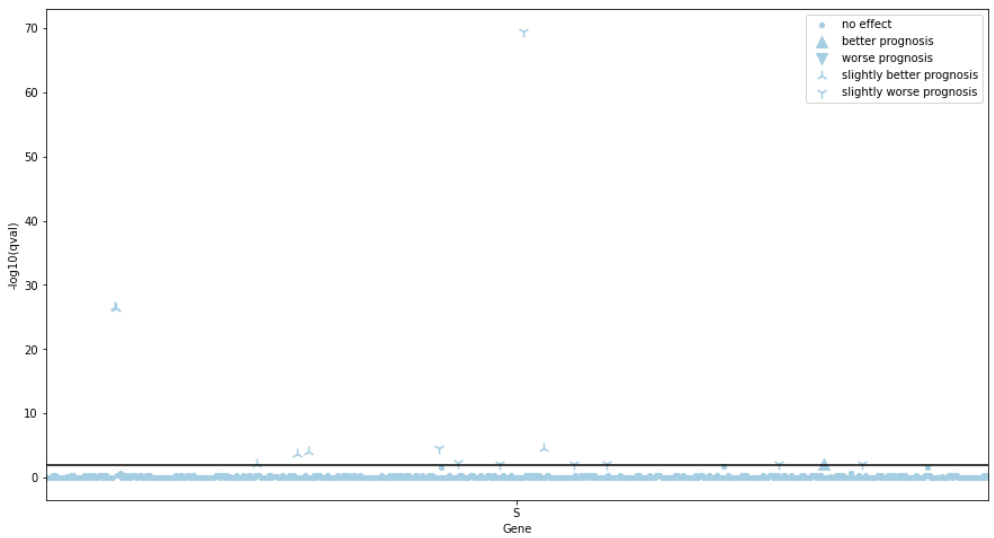
Figure 5 – Protein level analysis summary and key results

A

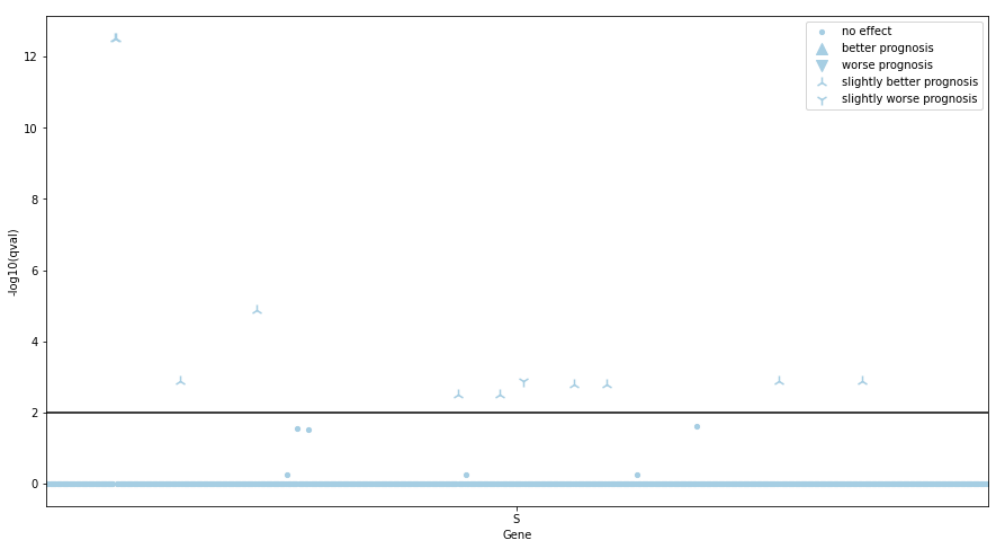
	age	alive	condition	gender	severity	symptoms	time
age	1	12	9	0	7	8	20
alive	12	3	9	0	8	8	13
condition	9	9	0	0	9	10	11
gender	0	0	0	0	0	0	0
severity	7	8	9	0	1	10	11
symptoms	8	8	10	0	10	2	13
time	20	13	11	0	11	13	26
total	22	17	11	0	12	15	53

	age	alive	condition	severity	symptoms	time
age	4%	70%	81%	58%	53%	37%
alive	54%	17%	81%	66%	53%	24%
condition	40%	52%	0%	75%	66%	20%
gender	0%	0%	0%	0%	0%	0%
severity	31%	47%	81%	8%	66%	20%
symptoms	36%	47%	90%	83%	13%	24%
time	90%	76%	100%	91%	86%	49%

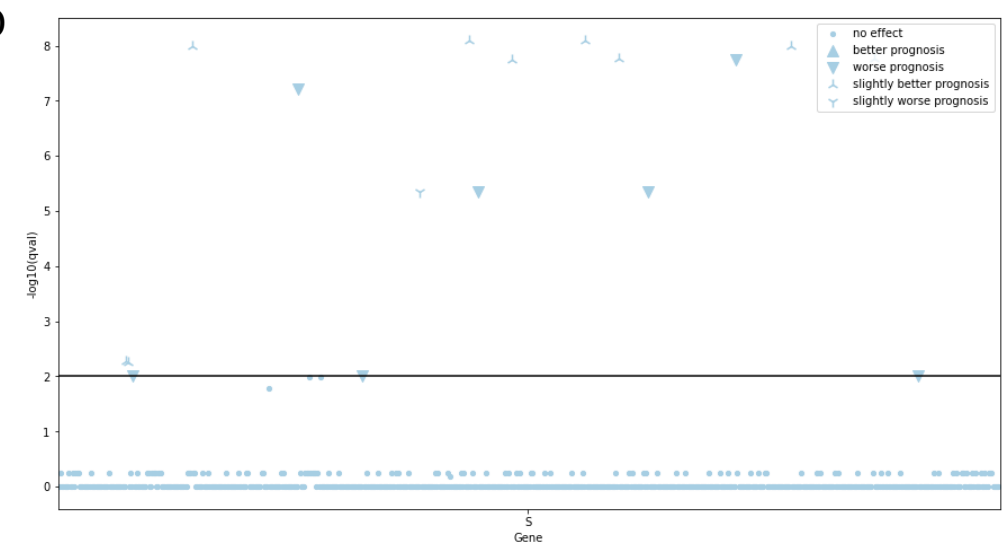
B



C



D



Multi-variant Vaccine Design

Abstract

Vaccination of the human population against SARS-Cov2 yields promising results in reducing infection, severe disease, mortality and transmission. Yet, emergence of Variants of Concern (VoC), viral mutants with enhanced infectivity, or with ability to evade immunity, is a major public-health concern. As the number of variants increases rapidly, it may become impractical to design, assay, pass through clinical trials, approve, deliver and administer multiple tailored-made vaccines, one against each VoC. Thus, we tackle here the challenge of redesigning a single mRNA sequence of the vaccine that may collaterally target several VoCs.

Codon optimization of vaccine sequence is a common practice in which among all the synonymous codons of each amino acid along the antigen protein sequence, an optimal codon is chosen. So-far codon optimization was mainly used for the enhancement of protein expression level. Here we propose to optimize codon choice of vaccines to allow collateral targeting of multiple VoCs. We aim to achieve the goal by incorporating our recent knowledge of the codon-level patterns and statistics of translation errors made by translating ribosomes.

Our proposal rests on the recently gained knowledge that evolution appears to have “designed” gene sequences to govern also the extent and type of error made upon their translation. Our new methodology (Mordret et al. *Molecular Cell* 2019) has detected and quantified translation errors proteome-wide, and it allowed us to discover that: (i) amino acid mis-incorporation events tend to recur in multiple proteins, and their prevalence may be as high as 1% or even higher; (ii) errors tend to predictably occur depending on choice of synonymous codon used to encode the original amino acid, and they often dictate the identity of the amino acid destination of a mis-translation event; (iii) our preliminary experiments (Samuels et al., unpublished) indicate that in mammalian cells peptides with mistranslation tended to be presented in the MHC - class I in frequencies that appear at time comparable to the original correctly-translated peptide, thus revealing substantial potential for immunogenic effect of translation error products.

We thus examined here bioinformatically if codon re-design of current mRNA vaccines of SARS-Cov2 could control the rate and patterns of translation errors and if it has a potential to create translation errors that will mimic genetic mutations that appear in VoCs. Our preliminary

investigations of the Pfizer and Moderna vaccines suggests a potential for codon substitutions that through translation errors might collaterally target multiple VoCs.

Introduction

The predictable and potentially programmable nature of translation errors made in cells

We have recently developed a new proteomic and informatics methodology to detect and quantify most errors made by the translation machinery of cells (Mordret et al. *Molecular Cell* 2019). We have mapped the translation errors made by ribosomes and other translation factors within cells and have made several important discoveries: (i) some errors may occur at very high rate (as high as 0.01 - 0.1 translated peptides have an amino acid mis-incorporation); (ii) errors appear to be predictable and even "programmable", namely for example, some of the synonymous codons for the same amino acid may be translated with more errors than others, and in addition, the selection of a codon can dictate the amino acid destination of a translation error event. Figure 1 presents the proteome-wide amino acid mis-incorporation data matrix in *E. coli* and yeast. The matrix depicts by a color code the number of unique peptides in the proteome in which an amino acid encoded by each of the 61 codons that appears on each row was found to be replaced, likely due to a translation error, with each of the other 19 amino acids which appear each on each column in the matrix. Most substitutions are consistent with mispairing between a codon and the wrong anti-codon (although others are ascribed to mischarging of an amino acid on the wrong tRNA). Different codons for the same amino acid show different patterns and rates of errors. For example, consider the four bottom rows that correspond to the four synonymous codons of the amino acid Gly, which have the form GGN. As can be seen, the most common mis-incorporation instead of Gly are Asp, Glu, and Ser. Focusing on Asp and Glu reveals that the codons GGU and GGC tend to lead to a Gly->Asp translation error, the codon GGA leads to Gly->Glu, while the 4th codon for Gly, GGG is often translated with less errors. Thus, if a wild-type antigen has a Gly amino acid, but a VoC has an Asp at that position, a recommended codon at that position should be GGU or GGC as these have a higher chance to give rise to a Gly-Asp translation error and thus generate the alternative antigen too. The converse holds too (in the *E. coli* based data) – from the two codons for Asp, namely GAU, GAC, only the latter has a high propensity to be replaced by Gly. The wide spread D614G mutation in SARS-Cov2 (Korber et al.

Cell 2020) presents exactly this said substitution, and it is thus tempting to speculate that a vaccine that would encode the Asp with GAC, or the Gly with GGU or GGC might provide better cross immunity for both variants. Figure 1B shows an example of a very frequent mispairing event between a codon and an anti-codon that occurs in the 2nd position of the codon where G in the codon pairs against U (rather than C) in the anti-codon. The two sets of three 4*4 matrices below show the prevalence of each such deduced mispairing in the prokaryote and the eukaryote, when mispairing was deduced to occur in each of the three codon positions, pairing each possible nucleotide in the codon against each of the non- Watson Crick pair in the anticodon. The matrices reveal universality across the species, obeying (often known) chemical tendencies, and they provide the basis for our design and control of desired translation errors in vaccines. We have recently begun to obtain such error propensities data in human cells too (see Figure 1C for preliminary results).

Further, errors are by definition rare. We thus aimed to examine how often translation errors occur in specific peptides in human cells. Not only have we mapped errors in the proteome, we have also isolated the MHC Class I from the cell surface (done by Yardena Samuels's lab) and detected and quantified peptides generated with particular amino acid mis-incorporation events. We found multiple such peptides and were further impressed by the high intensity - often proportional to high level of abundance - of these peptides. In fact, we often observed (Figure 2) that the correctly translated peptide, and the error-bearing counterpart, appear in comparable intensities on the MHC Class I. This could indicate that although error products are not as prevalent as the correctly translated peptides, due to a putative potential to destabilize their protein, they might be presented to the immune system at similar amounts.

Methodology and database

A compilation of SARS-Cov2 VoCs

We have constructed a set of potential VoCs for SARS-Cov2. Official VoC are characterized here by the mutations that they carry in specific amino acid positions along their genome. Since most current vaccines target the Spike protein, we focused on this protein only. Future work can expand the same effort to other proteins of the virus, and to other viruses or alternative immunological targets as well. Our compilation consists of two subsets of SARS-Cov2 VoCs. The "retrospective" set consists of

variants of the virus that already appeared in the human population, while the “prospective” sub-set consists of variants that were found by *in-vitro* molecular screens and due to their properties that are predicted to have a potential to invade the population and/or evade immunity. Some overlap does exists between the two subsets.

Note: our current compilation might not correspond precisely to the official definition of VoC, we use the term here rather loosely to refer to mutation that may or may not raise to the level of being a concern.

For the retrospective set we downloaded and aligned the Spike protein nucleic acid sequence from ~366,000 infected human individuals from GISAID (Elbe and Buckland-Merrett 2017). In each sequence position along the Spike gene we detect the most frequent amino acid substitutions relative to the consensus amino acid at that position. In a further analysis that will not be presented in this document we further prioritized these substitutions according to their GWAS-like association with severity of symptoms of the infected individuals. Thus future work could in particular aim to target existing mutations of especially higher public-health and medical concern.

The prospective sub-set of VoCs was constructed from experimental molecular screens of mutated versions of the Spike protein. In particular the selected mutations were those that were either found (i) when yeast cells expressing Spike were evolved to increase affinity to the ACE2 receptor (Zahradník et al. 2021); (ii) when a library of Spike mutations derived from deep mutagenesis scan (Starr et al. 2020, 2021; Greaney, J., Starr, et al. 2021; Greaney, J., Loes, et al. 2021), of the Receptor Binding Domain were measured for increased expression, increased affinity to the ACE2 receptor or for reduced affinity to neutralizing antibodies. The combined list from this retrospective and prospective set contained respectively 1,133 and 5,705 non-unique mutations that might emerge as potential VoC (Tables S1 & S2, respectively). Of those, 1,389 mutations were unique - consisting of non-repeating origin-position-target triplets - appearing in 169 unique positions in Spike.

The Pfizer and Moderna vaccine sequences

We obtained from the public domain the BNT162b2 (Pfizer-BioNTech) and mRNA-1273 (Moderna) nucleotide sequence of Spike vaccine sequence (NAalytics n.d.).

Results

A computational screen for codon changes that could elicit desired translation errors and collateral immunity

Our computational screen consists of three components. The first is the data on translation error propensities of codons in the genetic table, the second is and the above lists of SARS-Cov2 VoCs, and the third are the nucleotide sequences of the two current SARS-Cov2 mRNA vaccines.

We developed a simple algorithm that scans the existing nucleotide sequences of the Pfizer and the Moderna vaccines, and at each codon position examines (i) if there exists a VoC whose amino acid sequence differs from the sequence of the vaccine at that position, and (ii) if there is an alternative synonymous codon for the amino acid at that position that, upon prone-to-occur translation error, could mis-incorporate the VoC amino acid at the position instead of the original amino acid at that position.

In proposing to replace the original codon for an amino acid in the vaccine by a synonymous one, we compute the “mis-incorporation fraction” of each of the synonymous codons of the original amino acid at a position. The mis-incorporation fraction is the normalized number of translation error events from a particular codon to a desired amino acid destination that can replace the original one by a translation error event. We typically propose to use the synonymous codon with a maximal mis-incorporation fraction as it might maximize the tendency to mis-incorporate the original amino acid by that of the VoC at the position.

To determine whether there exists a synonymous codon of the original amino acid that upon error can be replaced by a desired one, we must put a threshold lower-bound on the mis-incorporation fraction. If there exists a desired codon with higher than threshold mis-incorporation fraction score we declare that the substitution is possible. Figure 3 shows a summary statistics of the number of amino acid positions along the vaccine in which we can suggest a potentially useful synonymous codon replacement. We gradually increase along the x-axis the minimal cut-off score of the mis-incorporation fraction, using data derived from each of the three species (Figure 1), on each of the two vaccines, done each, on the retrospective and prospective data (Figure 3 A-D). For example, for the Pfizer vaccine, in the retrospective compilation, given the human data of translation errors, at a

cut mis-incorporation fraction = 0.03, we can propose about 25 potentially useful codon substitution events to cover VoCs. In Figure 3E we plot (for Pfizer and below, Moderna vaccine) the number of positions (left), or unique mutations within such positions (right), in which we can propose a codon mutation as a function of the fraction of the VoC mutation in the human population. A handful of mutations can be proposed that would collaterally cover Spike mutations that are currently represented in >=1% of the human population.

As for the prospective compilation of VoCs: the data in Figure 4 shows that we can propose 52 and 54 codon substitutions for the Pfizer and Moderna vaccines respectively that will allow improved collateral targeting of VoCs.

Comment: Since the completion of my thesis work, the lab, together with the labs of Yardena Samuels and Tami Geiger, have joined forces to examine the predictions of this algorithm. They have finished the redesign of the Moderna and Pfizer vaccines, have synthesized new versions of these vaccines, introduced them into antigen presenting B cells and they are about to perform mass spec on the proteome and MHC peptides to reveal whether predicted mutations changed the antigens made and presented by the vaccine.

Methods

Translation error matrices

Translation error matrices in the form of 64 codons by 19 amino acids were obtained for *E. Coli* and *S. Cerevisiae* from previously published data (Mordret et al. 2019). Matching proteomic data for *H. Sapiens* was collected using the protocol described by (Mordret et al. 2019), from A375 and SKMEL30 cell lines, by the Samuels lab. Data collected from untreated SKMEL30 was summed with both untreated and PUNCH-P treated A375 cells to create a single, joined table for *H. Sapiens*. All tables were normalized by the sum of the table (all error types detected) to obtain the codon to amino acid replacement fraction.

Translation error based codon selection

Amino acid substitutions in the sequence of the SARS-CoV-2 Spike protein were evaluated for potential codon replacement. A substitution is defined here as the replacement of one amino acid

(denoted the ‘origin’) by another (denoted the ‘destination’). See vaccine design strategies below for the process of identifying substitutions of interest. All codons encoding for the origin amino acid were compared to one another in terms of their replacement fraction to the target amino acid, using the 64*19 normalized translation error matrices. The codon with the highest replacement fraction was selected. The process was repeated for both vaccines with a publicly available sequence (NAalytics n.d.). If equally good codon options existed (identical replacement fractions), one was picked at random. However, if one of those equally good options was the original codon, or if no data was available for any of the codons, the codon originally used in vaccine design was kept unchanged.

Retrospective VoC compilation

SARS-CoV-2 genomic sequencing data was downloaded in fasta file format from GISAID (https://www.gisaid.org/) via the website’s ‘Downloads’ tab, selecting the ‘unmasked MSA’ option. This data included all the sequences available at the time of downloading (2021-02-07). The downloaded fasta file was then analysed with the local Nextclade Docker tool (https://github.com/nextstrain/nextclade) using the command “neherlab/nextclade nextclade.js --jobs=10 --input-fasta 'input_file.fasta' --output-json 'results_file.json' ”. The tool was used to filter out sequences of low quality, as defined by the default settings of Nextclade, and to map mutations and deletions in sequences, relative to the reference genome (NCBI Reference Sequence: NC_045512.2). The resulting json files were used as input for an in-house written script to reconstruct a multiple sequence alignment of all remaining sequences (post filtering) with only mutations and deletions. Insertions were disregarded from further analysis. Two amino acid positions were excluded, namely 2 prolines in positions 986, 987 (pre-fusion structure stabilizing Prolines - SARS-CoV-2 S-2P) (Wrapp et al. 2020).

Prospective VoC compilation

Experimental RBD evolution and DMS data was obtained from publications and filtered for enhanced expression, increased ACE2 binding and immune evasion fraction. In detail: RBD evolution data (Zahradník et al. 2021) - all mutations associated with increased affinity to ACE2; DMS expression and ACE2 binding mutations (Starr et al. 2020) - filtered such that only mutations that appear to have a positive effect in both repeats were selected; Immune evasion data (Starr et al. 2020; Greaney, J., Starr, et al. 2021; Greaney, J., Loes, et al. 2021; Starr et al. 2021) was filtered as follows - standard

deviation (SD) was calculated for all variants. If the escape fraction was over 0.1 SD above the average, the highest 10% mutations were selected.
For each mutation, if translation error data existed, a codon was selected as previously described. The process was repeated for all species with available data. Frequency in real-life sequences was assigned to mutations for reference, based on the previously described MSA.

Figure Legends

Figure 1 Amino acid mistranslation patterns at codon resolution in *E. coli* and *S. cerevisiae* and human

A: the substitutions identification matrices of *S. cerevisiae* (green channel, left) and *E. coli* (red channel, right) are compared and overlaid (middle). The intensity of the color is proportional to the logarithm of the number of independent identifications, with one pseudo-count. Values are normalised by the highest entry in the matrix for each of the two organisms. The blue box highlights the recently described property of eukaryotic AlaRS to mischarge $tRNA^{Cys}$.

B: Upper panel: an example for a substitution resulting from mis-match between codon and an anti-codon. Lower panels: NeCE are classified by the mismatch most likely to generate them. The shade intensity reflects the ratio of independent substitution to the number of substitution types associated with the corresponding mismatch. Gray boxes are either correct base-pairings, or mismatches to which no substitutions could be unambiguously mapped.

C: the substitution matrix in the proteome of human cells (SKMEL30).

Figure 2 Same error-bearing peptides are detected repeatedly on melanoma cancerous MHC-I

Presented are MHC-I presented peptides with translation errors. Each dot represents a pair of correct and error-bearing peptide counterparts, detected on MHC-I, marked with original amino acid, positions of error, and destination amino acid upon error. Peptides are positioned on x- and y-axis according to their mass-spec intensity (naturally, typically higher for the correct peptide). The color code for each peptide signifies the number of samples, out of 8 examined, in which the error-bearing peptide in the pair has been identified.

Figure 3 - Retrospective analysis (all spike positions) & prospective analysis – Pfizer vaccine, the Moderna vaccine

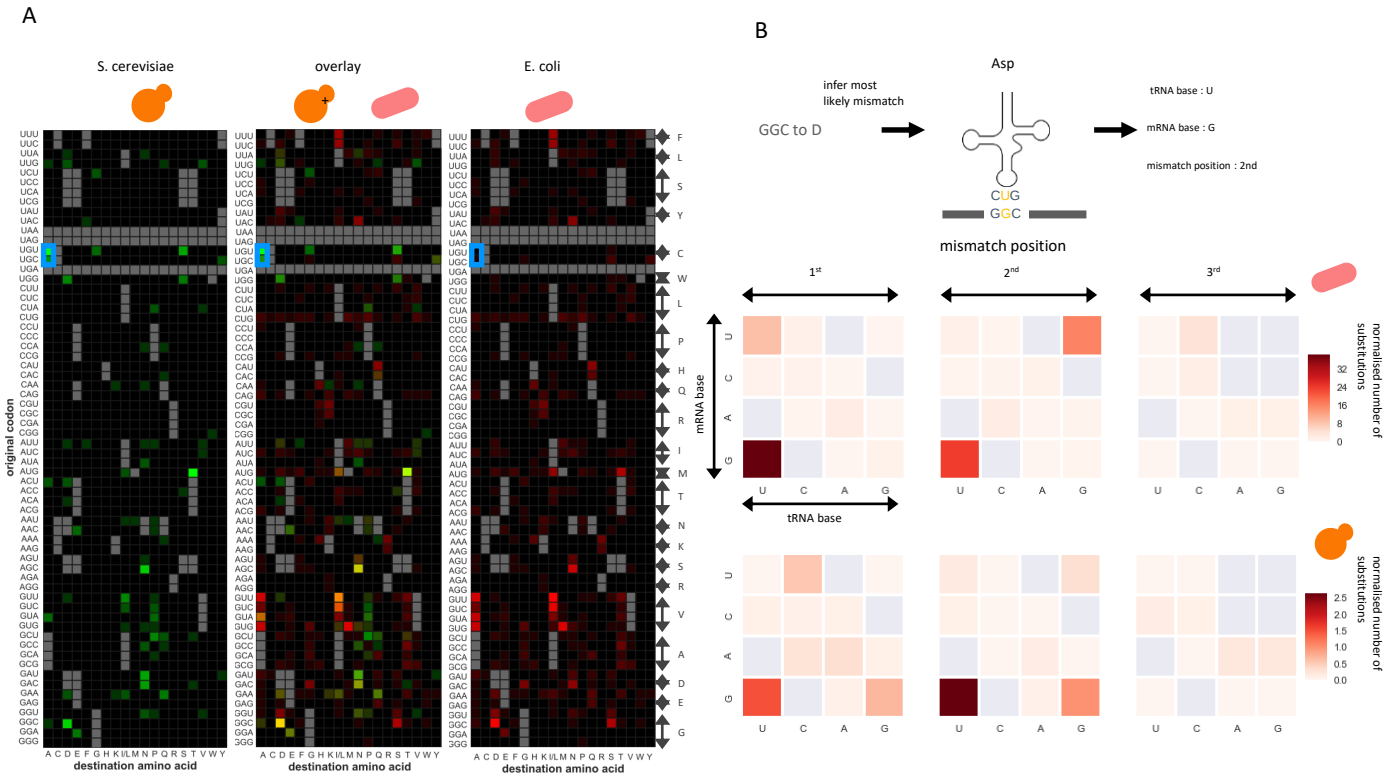
Mutations along the Spike protein for which we could suggest a beneficial codon swap. Number of mutations is given for a certain cutoff of replacement fraction or higher. Colored lines represent repeated analysis based on replacement fraction derived from the error data derived from the *e. coli,* yeast, or human data. Mutations with 0 frequency were excluded. **A** and **B** are for the retrospective and **C** and **D** are for the prospective compilations. (**E, F**) Number of proposed codon replacement mutations (left), of positions within the sequence (right) as a function of the frequency of the mutation in the human population in the retrospective compilation. **E** - Pfizer vaccine, **F** -Moderna

Figure 4 – Number of proposed codon replacement mutations relative to the Prospective compilation given the Pfizer (A) and the Moderna (B) vaccines

The Vann diagrams of analyzed mutations, taken from evolutionary and DMS studies on the Spike protein RBD, with a legend for the subsets (right). Values indicate the number of mutations/positions in the corresponding subset.

Table 1 Sample of selected proposed codon swap mutations in retrospective and prospective VoC compilation

Figure 1 Amino acid mistranslation patterns at codon resolution in *E. coli* and *S. cerevisiae* and human



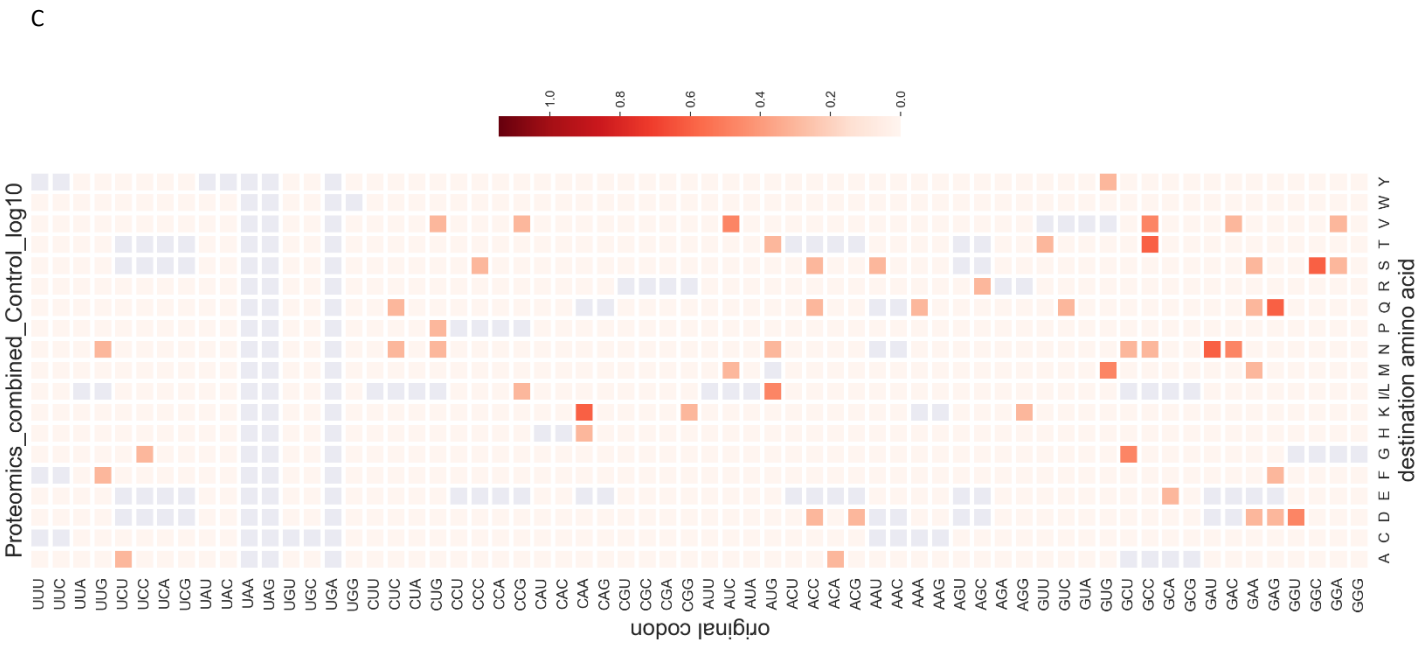


Figure 2 Same error-bearing peptides are detected repeatedly on melanoma cancerous MHC-I

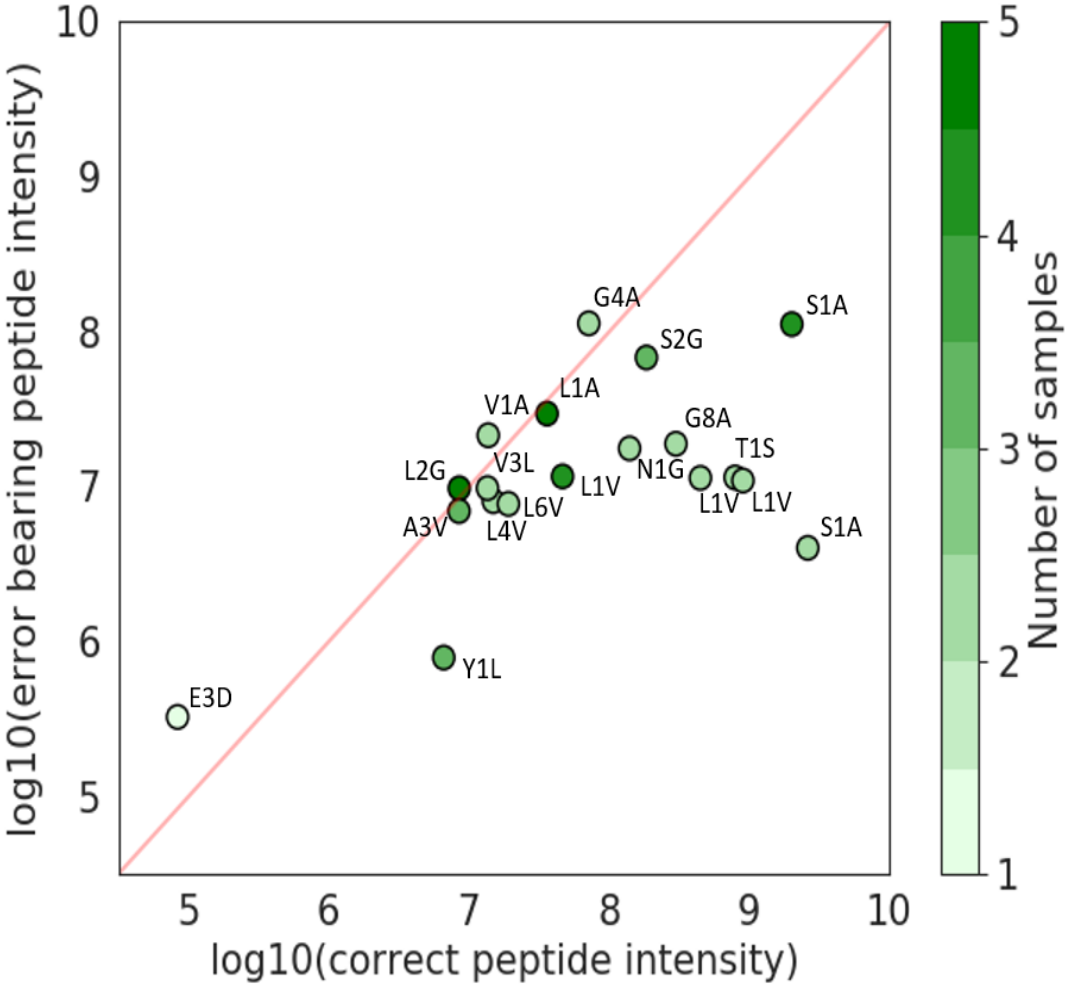
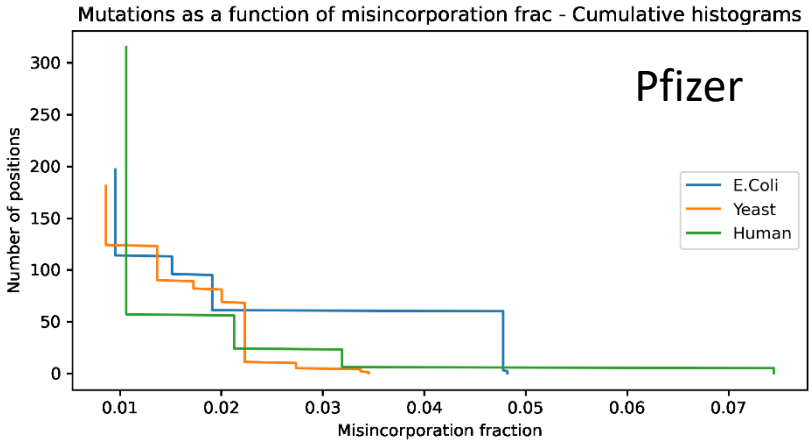
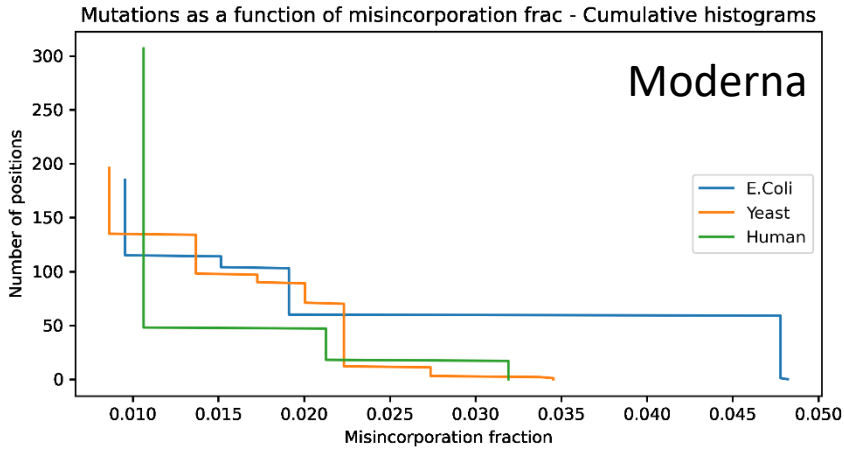


Figure 3 - Retrospective analysis (all spike positions) & prospective analysis – Pfizer vaccine, the Moderna vaccine

A

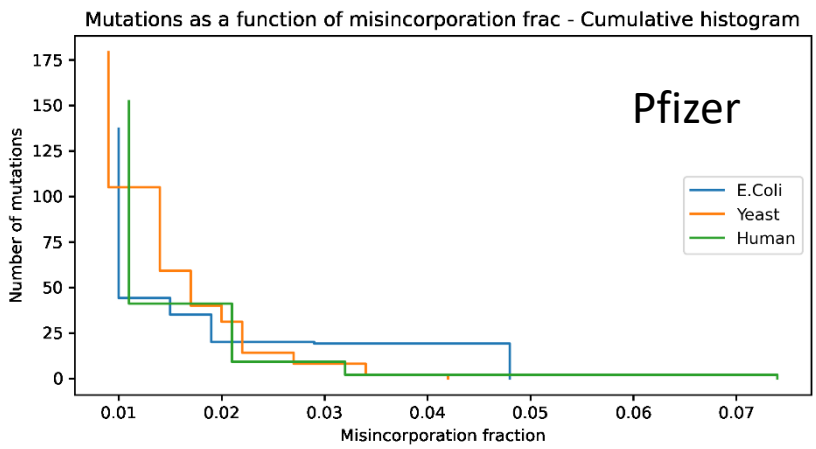


B

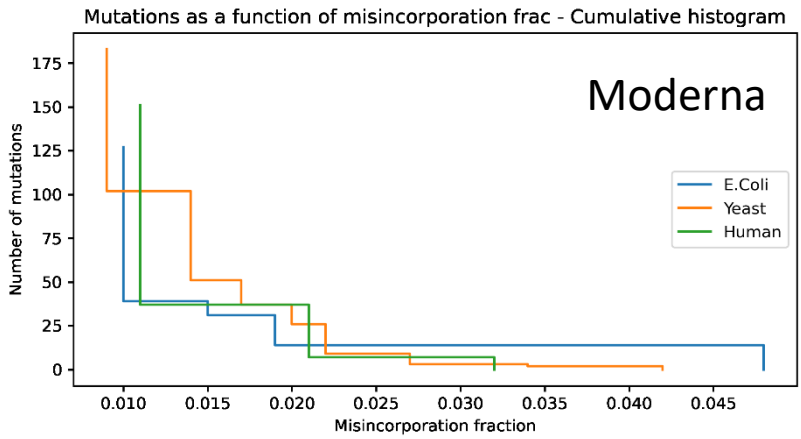


Retrospective VoC

C



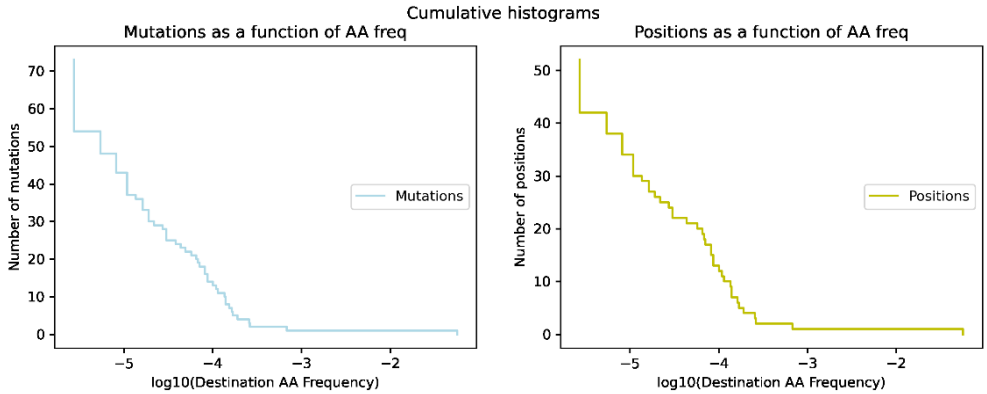
D



Prospective VoC

Figure 3

E


Cumulative histograms
Mutations as a function of AA freq
Positions as a function of AA freq
Number of mutations
Number of positions
Mutations
Positions
log10(Destination AA Frequency)
log10(Destination AA Frequency)

F

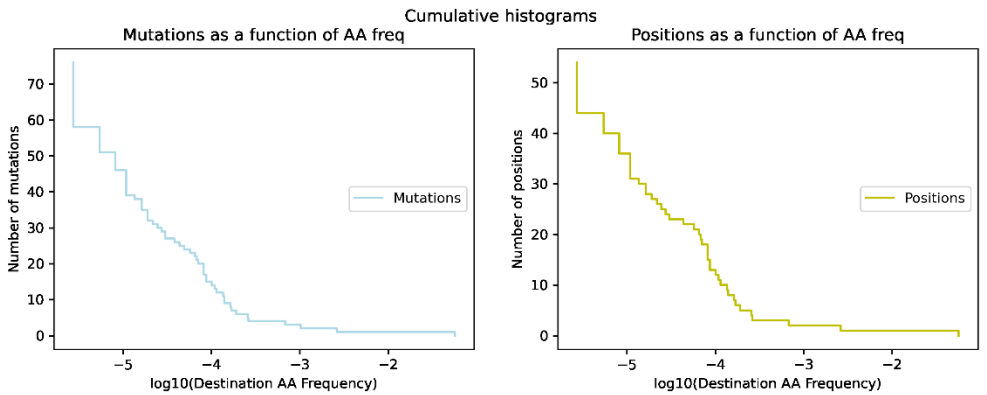
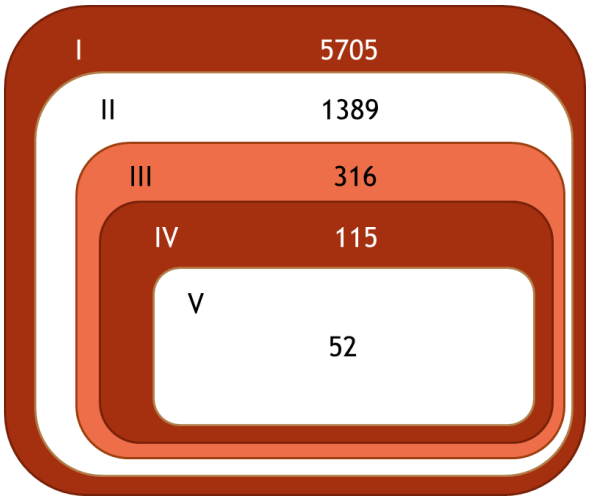
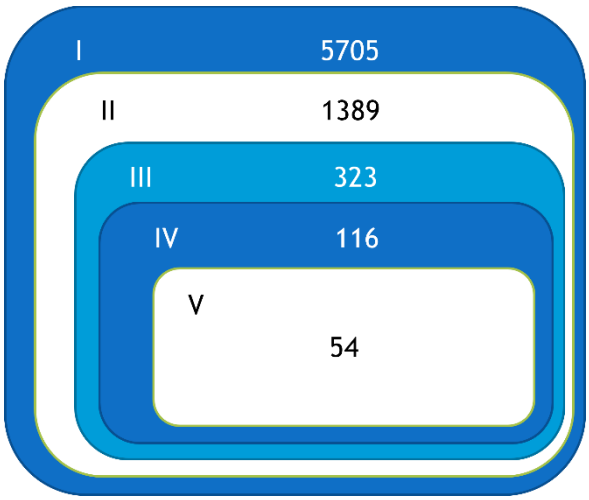

Cumulative histograms
Mutations as a function of AA freq
Positions as a function of AA freq
Number of mutations
Number of positions
Mutations
Positions
log10(Destination AA Frequency)
log10(Destination AA Frequency)

Figure 4 – Number of proposed codon replacement mutations relative to the Prospective compilation given the Pfizer (A) and the Moderna (B) vaccines



I – all aggregated mutations (includes duplicates)
II – unique mutations (e.g. I358F, E484K, E484L)
III – mutations for which we found at least one beneficial codon swap , based on at least one species' data.
IV – unique *positions* of III (e.g. 358, 484)
V – positions (in IV) for which the target AA has been observed in a naturally occurring isolate



I – all aggregated mutations (includes duplicates)
II – unique mutations (e.g. I358F, E484K, E484L)
III – mutations for which we found at least one beneficial codon swap, based on at least one species' data.
IV – unique *positions* of III (e.g. 358, 484)
V – positions (in IV) for which the target AA has been observed in a naturally occurring isolate

Table 1 Sample of selected proposed codon swap mutations in retrospective and prospective VoC compilation

Position in Spike	Original AA	Destination AA	Destination AA frequency in human population	Pfizer codon	Moderna Codon	Proposed codon (based on human error data)	Proposed codon error rate
222	A	V	0.225163	GCT	GCC	GCC	0.021277
982	S	A	0.109632	AGC	AGC	TCT	0.010638
18	L	F	0.101683	CTG	CTG	TTG	0.010638
477	S	N	0.056579	AGC	TCC	TCA	0.010638
614	G	D	0.047158	GAC	GAC	GGT	0.021277
1202	E	Q	0.000543	GAA	GAG	GAG	0.074468
314	Q	K	0.000494	CAG	CAG	CAA	0.031915
1223	G	S	0.000450	GGA	GGC	GGC	0.031915
111	D	N	0.000363	GAC	GAC	GAT	0.031915
239	Q	K	0.000041	CAG	CAG	CAA	0.031915

Mutation	Original AA	AA position	Destination AA	Destination AA frequency in human population	Moderna codon	Pfizer Codon	Proposed codon (based on yeast error data)	Proposed codon error rate	Experimental Evidence for VoC
S477N	S	477	N	0.056579	TCC	AGC	AGC	0.034	evolutionary (expression & ACE2 binding); DMS expression; DMS ACE2 binding
E484K	E	484	K	0.002625	GAG	GAA	GAA	0.009	evolutionary (expression & ACE2 binding); DMS ACE2 binding; 12 ABs escapes
G446V	G	446	V	0.000262	GGC	GGC	GGT	0.009	14 ABs escapes
V483A	V	483	A	0.000164	GTG	GTG	GTA	0.017	DMS expression
E484Q	E	484	Q	0.000139	GAG	GAA	GAA	0.020	DMS ACE2 binding; 10 ABs escapes

Literature

Dao, Thi Loi, Van Thuan Hoang, Philippe Colson, Jean Christophe Lagier, Matthieu Million, Didier Raoult, Anthony Levasseur, and Philippe Gautret. 2021. “SARS-CoV-2 Infectivity and Severity of COVID-19 According to SARS-CoV-2 Variants: Current Evidence.” *Journal of Clinical Medicine Research* 10 (12). https://doi.org/10.3390/jcm10122635.

Elbe, Stefan, and Gemma Buckland-Merrett. 2017. “Data, Disease and Diplomacy: GISAID’s Innovative Contribution to Global Health.” *Global Challenges (Hoboken, NJ)* 1 (1): 33–46.

Freeland, Stephen J., and Laurence D. Hurst. 1998. “The Genetic Code Is One in a Million.” *Journal of Molecular Evolution*. https://doi.org/10.1007/pl00006381.

Greaney, Allison J., Andrea N. Loes, Katharine H. D. Crawford, Tyler N. Starr, Keara D. Malone, Helen Y. Chu, and Jesse D. Bloom. 2021. “Comprehensive Mapping of Mutations in the SARS-CoV-2 Receptor-Binding Domain That Affect Recognition by Polyclonal Human Plasma Antibodies.” *Cell Host & Microbe* 29 (3): 463–76.e6.

Greaney, Allison J., Tyler N. Starr, Pavlo Gilchuk, Seth J. Zost, Elad Binshtein, Andrea N. Loes, and Sarah Hilton et al. 2021. “Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain That Escape Antibody Recognition.” *Cell Host & Microbe* 29 (1): 44–57.e9.

Hadfield, James, Colin Megill, Sidney M. Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A. Neher. 2018. “Nextstrain: Real-Time Tracking of Pathogen Evolution.” *Bioinformatics* 34 (23): 4121–23.

Huang, Yuan, Chan Yang, Xin-Feng Xu, Wei Xu, and Shu-Wen Liu. 2020. “Structural and Functional Properties of SARS-CoV-2 Spike Protein: Potential Antivirus Drug Development for COVID-19.” *Acta Pharmacologica Sinica* 41 (9): 1141–49.

Korber, Bette, Will M. Fischer, Sandrasegaram Gnanakaran, Hyejin Yoon, James Theiler, Werner Abfalterer, Nick Hengartner, et al. 2020. “Tracking Changes in SARS-CoV-2 Spike: Evidence That D614G Increases Infectivity of the COVID-19 Virus.” *Cell*. https://doi.org/10.1016/j.cell.2020.06.043.

Leung, Kathy, Yao Pei, Gabriel M. Leung, Tommy T. Y. Lam, and Joseph T. Wu. 2021. “Estimating the Transmission Advantage of the D614G Mutant Strain of SARS-CoV-2, December 2019 to June 2020.” *Eurosurveillance*. https://doi.org/10.2807/1560-7917.es.2021.26.49.2002005.

Lin, Lixin, Ying Liu, Xiujuan Tang, and Daihai He. 2021. “The Disease Severity and Clinical Outcomes of the SARS-CoV-2 Variants of Concern.” *Frontiers in Public Health*. https://doi.org/10.3389/fpubh.2021.775224.

Marioni, Riccardo E., Gail Davies, Caroline Hayward, Dave Liewald, Shona M. Kerr, Archie Campbell, Michelle Luciano, et al. 2014. “Molecular Genetic Contributions to Socioeconomic Status and Intelligence.” *Intelligence* 44 (100): 26–32.

Miyata, Takashi, Sanzo Miyazawa, and Teruo Yasunaga. 1979. “Two Types of Amino Acid Substitutions in Protein Evolution.” *Journal of Molecular Evolution*. https://doi.org/10.1007/bf01732340.

Mizumoto, Kenji, Katsushi Kagaya, Alexander Zarebski, and Gerardo Chowell. 2020. “Estimating the Asymptomatic Proportion of Coronavirus Disease 2019 (COVID-19) Cases on Board the Diamond Princess Cruise Ship, Yokohama, Japan, 2020.” *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin* 25 (10). https://doi.org/10.2807/1560-7917.ES.2020.25.10.2000180.

Mordret, Ernest, Orna Dahan, Omer Asraf, Roni Rak, Avia Yehonadav, Georgina D. Barnabas, Jürgen Cox, Tamar Geiger, Ariel B. Lindner, and Yitzhak Pilpel. 2019. “Systematic Detection of Amino Acid Substitutions in Proteomes Reveals Mechanistic Basis of Ribosome Errors and Selection for Translation Fidelity.” *Molecular Cell* 75 (3): 427–41.e5.

NAalytics. n.d. “NAalytics/Assemblies-of-Putative-SARS-CoV2-Spike-Encoding-mRNA-Sequences-for-Vaccines-BNT-162b2-and-mRNA-1273.” Accessed May 1, 2021. https://github.com/NAalytics/Assemblies-of-putative-SARS-CoV2-spike-encoding-mRNA-sequences-for-vaccines-BNT-162b2-and-mRNA-1273.

Naqvi, Ahmad Abu Turab, Kisa Fatima, Taj Mohammad, Urooj Fatima, Indrakant K. Singh, Archana Singh, Shaikh Muhammad Atif, Gururao Hariprasad, Gulam Mustafa Hasan, and Md Imtaiyaz Hassan. 2020. “Insights into SARS-CoV-2 Genome, Structure, Evolution, Pathogenesis and Therapies: Structural Genomics Approach.” *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*. https://doi.org/10.1016/j.bbadis.2020.165878.

Pearson, Helen. 2003. “Competition in Biology: It’s a Scoop!” *Nature*. https://doi.org/10.1038/news031124-9.

Shenhav, Liat, and David Zeevi. 2020. “Resource Conservation Manifests in the Genetic Code.” *Science* 370 (6517): 683–87.

Starr, Tyler N., Allison J. Greaney, Amin Addetia, William W. Hannon, Manish C. Choudhary, Adam S. Dingens, Jonathan Z. Li, and Jesse D. Bloom. 2021. “Prospective Mapping of Viral Mutations That Escape Antibodies Used to Treat COVID-19.” *Science* 371 (6531): 850–54.

Starr, Tyler N., Allison J. Greaney, Sarah K. Hilton, Daniel Ellis, Katharine H. D. Crawford, Adam S. Dingens, and Mary Jane Navarro al. 2020. “Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding.” *Cell* 182 (5): 1295–1310.e20.

Watanabe, Kimitsuna, and Tsutomu Suzuki. 2008. “Universal Genetic Code and Its Natural Variations.” *eLS*. https://doi.org/10.1002/9780470015902.a0000810.pub2.

“Website.” n.d. https://www.who.int/news/item/31-05-2021-who-announces-simple-easy-to-say-labels-for-sars-cov-2-variants-of-interest-and-concern.

———. n.d. Aksamentov, I., Roemer, C., Hodcroft, E. B., & Neher, R. A., (2021). Nextclade: clade assignment, mutation calling and quality control for viral genomes. Journal of Open Source Software, 6(67), 3773, https://doi.org/10.21105/joss.03773.

Wrapp, Daniel, Nianshuang Wang, Kizzmekia S. Corbett, Jory A. Goldsmith, Ching-Lin Hsieh, Olubukola Abiona, Barney S. Graham, and Jason S. McLellan. 2020. “Cryo-EM Structure of the 2019-nCoV Spike in the Prefusion Conformation.” *Science* 367 (6483): 1260–63.

Zahradník, Jiří, Shir Marciano, Maya Shemesh, Eyal Zoler, Jeanne Chiaravalli, Björn Meyer, Yinon Rudich, Orly Dym, Nadav Elad, and Gideon Schreiber. 2021. “SARS-CoV-2 RBD in Vitro Evolution Follows Contagious Mutation Spread, yet Generates an Able Infection Inhibitor.” *bioRxiv*. https://doi.org/10.1101/2021.01.06.425392.

Acknowledgements

I thank my advisor Tzachi Pilpel, first and foremost, for his guidance, support, encouragement and unwavering patience. For his catching curiosity and enthusiasm, his enlightening observations and hours of fascinating conversations, both on this work and other’s. I thank Orna Dahan for her keen eye for catching elusive details and for the hours of discussions that followed, trying to figure out their meaning. I thank Omer Asraf for teaching me the ins and outs of past work in the lab. I want to thank my lab members for creating a welcoming and collaborative environment. Lastly, I thank my friends and family for their resilient belief in me, and I thank E.Z. who’s help was pivotal in the last stretch of this work.